Title: CELL FITNESS IS AN OMNIPHENOTYPE

Authors: Nicholas C. Jacobs[1,2], Jiwoong Park[1,3], Nancy L. Saccone[4-5], Timothy R. Peterson[1,4,6-8]*

Affiliations: [1] Department of Medicine, Division of Basic and Biomedical Sciences: [2] Computational and Systems Biology and [3] Molecular and Cellular Biology programs, [4] Department of Genetics, [5] Division of Biostatistics, [6] Institute for Public Health, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110, USA. [7] BIOIO, 4340 Duncan Ave., Suite 236, St. Louis, MO 63110, USA. [8] Healthspan Technologies, Inc., 4340 Duncan Ave. Suite 265, St. Louis, MO 63110, USA.

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

Short summary: Cell fitness is a biological hash function that enables interoperability of biomedical data.

ABSTRACT

Moore's law states that computers get faster and less expensive over time. In contrast in biopharma, there is the reverse spelling, Eroom's law, which states that drug discovery is getting slower and costing more money every year. At the current pace, we estimate it costing $9.9T and it taking to the year 2574 to find drugs for less than 1% of all potentially important protein-protein interactions. Herein, we propose a solution to this problem. Borrowing from how Bitcoin works, we put forth a consensus algorithm for inexpensively and rapidly prioritizing new factors of interest (e.g., a gene or drug) in human disease research. Specifically, we argue for synthetic interaction testing in mammalian cells using cell fitness – which reflect changes in cell number that could be due many effects – as a readout to judge the potential of the new factor. That is, if we combine perturbing a known factor with perturbing the unknown factor and they produce a synergistic, i.e., multiplicative rather than additive cell fitness phenotype, this justifies proceeding with the unknown gene/drug in more complex models where the known perturbation is already validated. This recommendation is backed by the following evidence we demonstrate herein: 1) human genes currently known to be important to cell fitness 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 common 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, disease, 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

Despite or maybe because of technology improvements there is currently no solution for Eroom's law. Eroom's law is the observation made by Scannell et al. that drug development gets slower and more expensive over time (1, 2). It might be said that this problem is one of lack of consensus: There are so many approaches to solve biomedical problems, and a growing number of approaches, that the lack of coordination around the most high-yield approaches increases the time and money spent. To address this, perhaps we need to look to lessons on how other fields think about problems of coordination. Consider how the longstanding computer science problem known as the Byzantine Generals Problem was solved. The Byzantine Generals Problem was articulated by computer scientists working for NASA back in the 1970's to think through air traffic control and the realization that computers were soon going to be flying planes (3, 4). The general form of the Byzantine Generals Problem describes a leaderless situation where involved parties must agree on a single strategy in order to avoid failure, but where some of the involved parties are unreliable or are disseminating unreliable information (5). Interestingly, like with important problems in biology that were solved using knowledge from other disciplines, such as physics, key insights to the Byzantine Generals Problem came from outside of computer science. Namely, in economics it was solved by a consensus algorithm based on "proof-of-work" that is now used to run the Bitcoin peer-to-peer monetary system (6). Developing such a consensus experimental approach to optimize biomedical research does not exist currently. Needless to say, such an approach would be extremely valuable if one were to identify one.

Consensus in biomedical research has historically taken the form of researchers agreeing on what biological model systems they will focus their labs around. Interestingly, some of the most low-overhead model systems such as yeast and worms have made a disproportionate

contribution to human health (7-9). This is notable because often they were used without regards for their eventual applications for human health. For example, the broadly medically important mechanistic target of rapamycin (mTOR) pathway was first discovered in yeast using a cell fitness-based, i.e., cell counting-based, screen (10, 11). Still, many argue that even mice aren't representative models of human diseases. Some researchers are nevertheless pushing forward and focusing on creating new ways to accelerate human disease research using model organisms (12-14). 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, must be taken into an account. There is therefore a need for models that strike a balance between the ease and robustness of model organisms while preserving the signaling connectivity important to human disease.

Modeling disease has gained new relevance as the number of human population genetic studies has exploded (15). 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 (16), 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 proliferation that PTEN inactivation causes. However, it is clearer when one considers that patients with autism often have larger brains and more cortical cells (17). Nevertheless, the extent to which the various functions of genes like PTEN are relevant to disparate diseases begs the questions of how we molecularly classify and model disease and whether we truly understand the function of these genes beyond cancer.

As with 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, cholesterol-lowering statins were initially held back from the clinic because they were shown to be toxic in animal models. Those early findings scared off pharmaceutical companies because they suggested statins might not be safe for humans (18, 19). Merck eventually realized this toxicity was on-target related to the drug mechanism of action and not generic off-target activity, and the statins went on to be one of the biggest success stories in drug development history (20). The successes of drug “repositioning” also reveal to us that drugs can function in multiple “unrelated” contexts. For example, the malarial drug, hydroxychloroquine (Plaquenil), is now routinely used to treat rheumatoid arthritis and autoimmune disorders (21). There is also chemotherapeutics being considered for depression (22). Drug repositioning makes sense from a molecular perspective because many drug side effects could actually be on-target effects (23). To take it further, now with so much knowledge in the “-omic” era, it might be the rule 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 (24), and regulates gut motility and pulmonary blood flow (25, 26) 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. First, we quantify costs incurred for studying proteins for eventual drug development and demonstrate that continuing with the status quo will require prohibitive amounts of time and money. Then, we present evidence that synthetic interaction testing using cell fitness provides a novel solution to this challenge. Specifically, by leveraging diverse data types: gene-inactivation screens in cells, PubMed, and human genome-wide association studies (GWAS), we demonstrate that synthetic interaction testing in cells

would be predicted to yield important insight on a new factor of interest (e.g., gene or drug) irrespective of the ultimate phenotype of interest for that factor. This predictive power comes from two related findings. One, that cell fitness is altered upon perturbation of most, if not all, biological pathways. Two, that a gene's or drug's relevance to a disease is correlated with its relevance to an in vivo proxy of cell fitness, cancer.

Based on these findings, 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 restricted to the context of cancer (27, 28). This makes sense considering cancer has become defined by alterations individual phenotypes that contribute to cell number/fitness, e.g., proliferation and apoptosis (29, 30). However, we argue fitness-based synthetic interaction testing has a much more general purpose. Our evidence indicates that cell fitness is a generalizable phenotype because it is an aggregation of individual phenotypes. To the extent that it might be an aggregation of all possible phenotypes – an *omni*phenotype – suggests its potential as a pan-disease model for biological discovery and drug development.

RESULTS

Problem: The current approach to studying human disease biology is time-consuming and costly. Proteins are key building blocks of life. They are also what most drugs are designed to target. Proteins interact with each other to perform complex functions, such as DNA replication, ATP production, and the trafficking and degradation of various molecules. The human protein interactome is vast, and by one estimate 8.8M of them could be essential for cell viability (31). We asked how much time and money has been spent on studying these

interactions. When we look at a high-confidence list of 21,729 protein-protein interactions (PPIs) (32), NIH funded grants have studied 3.3% of these pairs (717 pairs) and $3.5B has been spent on them. At the current rate, it can be estimated to take until the year 2574 and cost upwards of $150B to study all 21,729 interactions. Notably, this does not include investment to develop drugs based on this information, which is estimated to push the numbers up 66X, totaling $9.9T and likely taking much longer than the year 2574 (33) (Fig. 1A). Also, these numbers are for the 21,729 pairs, which is less than 1% of the total 8.8M potentially important ones. Clearly, our current strategy is insufficient to understand human genes in a timely and cost-effective manner.

We considered the biomedical industry as a poorly coordinating network. To make the network more efficient, we considered lessons from computer science and the study of distributed systems. A famous problem in distributed systems that seemed relevant to Eroom's law is the Byzantine Generals Problem. The Byzantine Generals Problem is a hypothetical situation in which a group of generals lead armies that are geographically separated from each other (such as on other sides of two hills), which makes their communication difficult (Fig. 1B). They must choose either to attack the enemy (which is in the valley between them) or retreat. A half-hearted attempt at either will result in a rout, and so the generals must come to a consensus decision on which approach to take. To extend the analogy to the biomedical industry, we considered how the lack of consensus by researchers on what experiments should be prioritized has led to an inability to win the "war on cancer" as well as other 'battles' against other diseases (Fig. 1B) (34).

A Potential Solution: Synthetic interaction testing using cell fitness. At the heart of the Bitcoin's solution to the Byzantine General's Problem is its blockchain. A blockchain is a publicly viewable and editable database (Fig. 1C). It has comparisons to a collection of Google or Excel spreadsheets in that everyone anywhere can read and write to them. However, a key distinction

is that on a blockchain one can't overwrite other people's additions. This cryptographically secured step maintains the fidelity of all the data, which thus leads to a universally shared understanding of it. Consensus in a blockchain is achieved using hash functions, which transform arbitrary sized data into fixed sized hashes (Fig. 1C). The point of hashes is that they are much easier to analyze relative to the data that comprises them. This allows people to relatively easily trust that the data as a whole is accurate without having to verify each individual piece of it themselves. Researchers are starting to articulate how biomedical research could be aided by the concept of blockchains (35-37). The volume of data in biomedical research is exploding, yet it is growing rapidly in different silos. We wondered what kind of biological 'hash' function might be well suited to faithfully 'chain' together disparate 'blocks' of biomedical data. Recently, the Cancer Dependency Map, a.k.a., DepMap, shed important light on PPIs across the human proteome (38, 39). This is interesting because the DepMap only required cell fitness as a phenotype to identify the PPIs. Another study, PRISM, used a similar approach as DepMap, but instead of mutating genes to perturb cell fitness, it used drugs (40). Both DepMap and PRISM relied on many of the same cell lines to make their predictions. Thus, we asked if one were to 'chain' together the DepMap and PRISM 'blocks' of cell fitness data, how many drugs targeting the aforementioned 21,729 PPIs might be identified. We estimate that these two datasets alone could provide a first pass at drugging 27% of these PPIs, representing $2.67T and 149 years of future money and time spent if they weren't used (Supp. Table 1). The underlying hash function-like, consensus algorithm we hypothesize one would use to 'chain' together the DepMap and PRISM involves imputing synthetic interactions between genes and drugs from their cell fitness data (Fig. 1C). Therefore, to understand how we arrived at this 27% number and how we might eventually get to 100% coverage using cell fitness data, it is first important to explain synthetic interaction testing.

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, hereafter denoted by "ρ" (as in relative cell fitness or phenotype), is much greater than the effects of the individual mutations (Fig. 2). If there is an interaction, it can be either be a negative synergy (Fig. 2, 5th from left panel), 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. 2, 6th from left panel). Both phenotypes demonstrate a synergy on a genetic level between these two genes.

Amongst the ~400 million possible synthetic interactions between protein coding human genes, less than 0.1% have been mapped to date (41). 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 (42). Therefore, we sought to develop a framework that would allow us to infer synthetic interactions at scale as well as leverage interactors of known significance 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 hereafter 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 each represent a gene, disease, or drug. Throughout this work, we provide three pieces of evidence that taken together provide a generalizable framework for solving for *U* using *K* via cell fitness-based synthetic interaction testing.

Evidence #1: Nearly all cell processes affect cell fitness. Recent works by Hart et al., Blomen et al., and Wang et al. identified "cell essential" genes – genes that are required for cell

fitness (43-45) – and we wondered what GO pathways they participate in. In this case, we took the above published lists of cell essential genes and considered the synthetic interaction testing as follows: the known, *K,* is the gene that is essential in at least one cell type and the unknown, *U,* is the synthetically-interacting GO pathway of interest. As precedence for this approach of considering *U* as a collection of genes rather than an individual gene we applied the principal of the minimal cut set (MCS). MCS is an engineering term used by biologists to mean the minimal number of deficiencies that would prevent a signaling network from functioning (46, 47). MCS fits well with the concept of synthetic lethality (27), and this was recognized by Francisco Planes and colleagues to identify RRM1 as an essential gene in certain multiple myeloma cell lines (46).

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 gene(s) in one of these pathways were knocked out, they would produce a synthetic lethal interaction with those components (for clarity throughout this study, by “genes”, we refer to human protein coding genes).We found the essential genes from the Hart, Blomen, and Wang studies 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, at most 2054 out of 21246 total genes assessed were essential and these essential genes involve at most 36% of all possible GO categories (6444 out of 17793. 17793 represents the number of unique – i.e., irrespective of the GO hierarchy - human gene-associated 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. 3A graph appears to approach 1.0, this suggests if enough cell types are tested, potentially all genes and GO processes would be essential in at least some cell context.

To extrapolate beyond the ten mammalian cell lines where cell essential genes were systematically characterized, we relaxed the significance on the cell fitness scores from $p < 0.05$ to $p < 0.10$. This relaxed threshold encompasses nearly all GO processes and genes (Fig. 3A, 97.2%, 17291/17793 GO terms; 17473/21246 genes). Taken together, this suggests the majority if not all GO processes and their associated genes affect cell fitness.

Recently, genome-scale gene inactivation screening has been performed on hundreds of human cell lines (38). Using this data 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 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. These categories were chosen because they are direct children of the top-level “Biological Process” GO category (GO:0008150, with a “is_a” relationship) that cover broad areas of human health and have similar numbers of genes associated with them. Impressively, highly co-cited genes from each top-level GO category were highly correlated in their fitness profiles ($p < 1e-8$, Fig. 3B, Supp. Table 1). Gene co-citation is a human determined connection between two genes whereas cell fitness correlations are machine determined. Therefore, that all significantly correlating cell fitness profiles were also enriched as being co-cited (compare light and dark blue in Fig. 3B, hypergeometric distribution model enrichment p-values: immune system process $p = 1.4E-4$; metabolic process $p = 6.9E-3$; development $p = 2.9E-3$; behavior $p = 2.8E-2$) suggests machine learning approaches using cell fitness can accurately predict gene function for disparate biological processes.

***Evidence #2: Nearly all common disease genes are cancer genes*.** The omnigenic model states that all genes are important to a common disease (48). 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 way to inform us on which genes are core genes for a given common 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.

Here we defined a synthetic interaction as follows: the gene(s) mentioned in a citation is the unknown, *U*. In this convention we make the simplifying assumption that this gene(s) is not known in whatever context was studied, e.g., diabetes, which is why it warrants a publication. Whereas all other already published genes in that context at the time of publication we consider as 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 common disease defined by altered cell number. Specifically, we assessed all citations linked to the MeSH term "neoplasm" vs. other MeSH terms in the Disease category (we hereafter refer to these as "cancer" vs. "non-cancer" citations, see Methods for more details). In analyzing all currently available PubMed citations (~30M), we identified a compelling relationship (Correlation coefficient, r = 0.624): most genes are cited in the cancer literature to a similar extent as with the non-cancer literature (Fig. 4A, Supp. Table 1).

Some non-cancer citations such as those that focus on normal gene function as well as rare disease, could undermine our interest in making an apples-to-apples comparison with cancer. To more strictly determine whether the cancer vs. non-cancer correlation pertains to common diseases, we restricted the non-cancer set to common conditions that include the top causes of death according to the World Health Organization (49) and other sources: Alzheimer's,

cardiovascular disease, diabetes, obesity, depression, infection, osteoporosis, hypertension, stroke, and inflammation (~10M citations). This more stringent analysis surprisingly produced a stronger correlation (r = 0.69) than the more all-encompassing non-cancer citations list (Fig. S1, Supp. Table 1). These strong correlations suggest for many non-cancer disease states, genes important to them that are not yet known can be identified using synthetic interaction testing. Also, 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. 4A in more detail. To understand what GO categories led to better cancer/non-cancer citation correlations, we binned the cancer and non-cancer citations for all genes by their top-level GO categories we used in Fig. 3: immune system process, behavior, metabolic process, and developmental process. We also included cell population proliferation as that is directly related to cancer. We found no clear pattern of overtly cancer-related GO sub-categories that contribute most vs. least to the high cancer/non-cancer citation correlation (Fig. 4B, Supp. Table 1). For example, cell population proliferation (GO:0008283) sub-categories are interspersed amongst the other four top-level GO categories when sorted by correlation strength (Fig. 4B, Supp. Table 1). This lack of a pattern suggests that many more GO processes than might be expected can 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 (29, 30).

PubMed can be considered biased because it is curated by humans, therefore we considered the orthogonal approach of using genome-wide association studies (GWAS) to further

characterize the cancer/non-cancer correlation we identified using PubMed. Synthetic interaction testing is a genetic concept, therefore it made sense to consider GWAS because it's a mainstay for human genetics. We used the UKBioBank dataset because of its wide acceptance and large sample size (46), and specifically the widely-used summary GWAS results of this dataset from the Ben Neale lab (50). We selected a set of 100 cancer and non-cancer conditions (Supp. Table 1) that were most co-cited with genes (as opposed to medical contexts where genes are not reported to be influential). To determine the interaction strength of a gene in cancer vs. non-cancer, we examined the proportion of diseases in each category for which there was an associated SNP in that gene. Our results show a high correlation using either p-value ($r = .819$, Fig. 4C) or beta value thresholds ($r = .812$). Because the thresholds we used for p-values ($p < 0.001$) and beta values ($|beta| >= 0.0031$) can be considered modest it is possible that the strong correlations could be confounded by gene length. That is, the longer the gene, the more likely it is that it will contain a significant SNP. However, when we control for gene length, our correlations remained statistically significant ($p < 0.001$). Combined with the results of our PubMed analysis, these findings suggest a strong relationship between the effects of a gene perturbation on cell fitness – represented by cancer – and all other phenotypes – represented by non-cancer conditions.

***Evidence #3: Drug effects on patient cell fitness are reflective of the patient condition response to drug*.** Like with variation 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. We performed a systematic PubMed analysis by identifying 25 studies where patient derived cells were generated and treated with drugs and where cell fitness was measured alongside the patients themselves being treated and their drug responses monitored. What was

notable in all the studies we identified is that cell fitness was predictive of patient response for widely divergent medications (Fig. 5A). These results might be surprising outside the context of cancer therapy, but they arguably 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. 5A, Supp. Table 1).

Because we found a relatively small number of studies that met our strict criteria in the drug effects PubMed analysis, we considered the cancer vs. non-cancer paradigm we used in Figure 4 to give us more comprehensive insight on the relationship between drug response and cell fitness. 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 all human protein coding genes and all FDA-approved drugs being cited both in cancer and non-cancer to be increasing over time. Indeed, human genes are increasingly being cited with similar frequency in cancer and non-cancer disease contexts (Fig. 5B, 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. 5B, Supp. Table 1). This finding suggests drug repositioning opportunities for non-cancer conditions that are underexplored with the current dichotomous cancer gene vs. non-cancer gene mindset.

Lastly, to make an analogous analysis with drugs as we did with diseases in Figure 4, we binned the top cancer and non-cancer drugs taken by participants in UKBioBank. Figure 5C shows the proportion of drugs with an associated SNP, where associated is again defined by a

p-value < .001. The cancer vs. non-cancer drug correlation is strong (r = .66) and like with cancer vs. non-cancer diseases remains statistically significant when accounting for gene length (Fig. 5C). Combined with the results of our PubMed analysis, these studies lend support to the increased interest in repositioning many non-cancer drugs for cancer (51) (e.g., immunosuppressant, rapamycin and anti-diabetic, metformin), and vice versa (e.g., chemotherapeutics for autism, senolytics for aging (52).

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. 6). 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 of interest (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) (41, 53, 54). 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 (41, 53, 54). In this case, N-BPs and FDPS were the known, *K,* and ATRAID and SLC37A3 were the unknown, *U,* we discovered. Besides the N-BPs and the statins, there are numerous other examples in the literature such as with the mTOR pathway (55) 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. 3). Our Figure 3 results suggest one can study the majority of cell and molecular processes using cell fitness-based, synthetic interaction testing. Specifically, Figure 3 shows that to the extent that cell and molecular processes are well represented by GO terms, essential genes as determined by cell fitness (cell count) are involved in the vast majority of such processes. 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 (56). Similar percentages are obtained at the population level when comparing cell types (57, 58). 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 or scalable. Whereas cell fitness is robust and readily approximated by inexpensive assays, such as those to measure ATP levels.

Relevance to diseases (Fig. 4). Our PubMed and GWAS analyses lend growing support to using dry-lab approaches as first-class tools for discovering medically relevant biology. The costs to perform the wet-lab experiments required to publish in the "-omic" era are not small and favor well-resourced teams. Gleaning insights from publicly available data, which is large and growing rapidly, and accessible by anyone with programming skills, is a way to level the playing field.

One counterintuitive insight from Figure 4 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 (42), and approaches to aggregate data are known to favor the "discovery" of already well-studied genes over less studied genes (59). When applied in unbiased methods such as with genome-wide CRISPR-based screening (41), 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. 5). 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 (60). Though more work needs to be done to test the generalizability and "real-world" application of the studies we highlight in Figure 5A, they do provide important initial evidence for exploring 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 (61). 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 (62) can potentially be the best of both worlds because it extrapolates well to the phenotype of interest and provides the gene target in the same screen.

Still, we acknowledge limitations to 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. PubMed-based analyses are subject to biases. Many genes are not published on due to human factors. Even among those genes that are published, their representation is affected by publication bias towards positive results. Additionally, our citation analysis relies on curation by NCBI, which due to its high stringency and the current limitations of natural language processing strategies, likely leaves out a large number of studies that should be included. However, this is why we also analyzed GWAS, which provides a more unbiased look. 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.

FIGURE LEGENDS

Figure 1: The problems of and a potential solution for the lack of consensus-building mechanisms in biomedical research. (A) Current and projected time and money spending estimates to develop drugs for high-confidence protein-protein interaction (PPI) pairs obtained from Huttlin et al. (32, 63). The projected budget is a line of best fit based on the cumulative NIH

grant money spent on studying gene pairs multiplied by how much is needed to convert this funding to FDA approved drugs, i.e., 44,200/670 = 66X (33). The inset graph represents the current spending estimate that is extrapolated to make the projected spending and time estimates. The intersection of the light blue and tan lines represent the projected date and costs, respectively, when all drugs will be identified for each PPI. **(B)** A schematic of the Byzantine Generals Problem and its application to the biomedical industry. In a real-world scenario, achieving consensus becomes much more difficult when there are many parties involved (greater than the two depicted) that are not working together. **(C)** Data blocks are chained together using a robust algorithm called a hash function to create consensus understanding. In the case of bitcoin, the consensus determines who owns what bitcoin. To make an analogy on how the concept of a blockchain can address Eroom's law, the hypothesis is that cell fitness is a biological hash function that enables interoperability of biomedical data from different research efforts.

Figure 2: 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 significantly differs from the expected effect based on each effect alone; typically, an additive effect is expected. Synthetic interactions can cause lethality or hypersensitivity, or can be buffering or confer resistance. Both are depicted in the figure where ρ – the phenotype – is relative cell number. Synthetic effects can be thought of as multiplicative or exponential rather than linear or additive, as is the case in the condition (4th column from left) where $1 - \rho_x (0.75) + 1 - \rho_y (0.75) = \rho_{xy} = 0.5$. Note: In some scenarios, a perturbation is lethal on its own, which precludes an investigation of its synthetic lethal but not its buffering interactions. On the other hand, the lack of a perturbation's effect on its own on cell fitness does not preclude one from determining if it has a synthetic interaction with another perturbation.

Figure 3. Nearly all GO processes are essential for cell fitness. (A) The y-axis is the ratio of 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 (17,793) for all tested human genes (21,246). The x-axis is the number of essential genes determined for the indicated number of cell lines. “Relaxed” means $p < 0.10$ (as opposed to $p < 0.05$) 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. P-values were false discovery rate (FDR) corrected using the Benjamini and Hochberg method (64).

Figure 4: 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 MeSH terms in the Disease category 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). Categories are sorted from dark blue representing a stronger correlating GO category to lighter blue representing a weaker correlating category. **(C)** The proportion of non-cancer diseases vs. cancer diseases that have an associated SNP (p-value < .001) in UKBioBank for each protein-coding human gene. These proportions are out of the top 100 conditions, 26 cancer diseases and 74 non-cancer diseases, based on co-citation with genes. Correlation coefficient, $r = .819$.

Figure 5. Establishing the efficacy of FDA-approved drugs using cell fitness. (A) Drugs where their effects on a patient’s 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. The most recent data point of genes corresponds to the same data that was used in Figure 4A. (**C**) The proportion of non-cancer drugs vs. cancer drugs that have a significant SNP (p-value < .001) in UKBioBank for each protein-coding human gene. These proportions are out of a set of 15 drugs used in UKBioBank patients who had cancer vs. 15 other drugs used in non-cancer conditions that had the highest number of cases in the UKBioBank. Correlation coefficient, r = .671.

Figure 6. 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) *ρ,* 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 3); 2) *ρ* can be proxied by the relative involvement of the interactors in cancer (related to Figure 4); 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 5). 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.

Figure S1. Common disease-restricted cancer vs. non-cancer PubMed analysis. Nearly all common disease genes are cancer genes. The co-occurrences of each human gene with cancer (x-axis) and non-cancer (y-axis) conditions in ~27 million PubMed citations were counted. The following non-cancer conditions were considered: infection, Alzheimer's, cardiovascular disease, diabetes, obesity, depression, inflammation, osteoporosis, hypertension, stroke, and inflammation. Line of best fit - Slope, 0.75; Correlation coefficient, r = 0.69.

Table S1. Data related to all figures. Data to generate scatter plots as well as list of diseases and drugs analyzed are included in Supp. Table 1.

AUTHOR CONTRIBUTIONS

T.R.P, N.C.J, and J.P conceptualized the project. N.C.J and T.R.P performed the analysis. N.L.S. provided feedback on the GWAS analysis. T.R.P wrote the paper and N.L.S, N.C.J, and J.P edited it.

FUNDING

This work was funded by NIH grants, K99/R00AG047255, R01AR073017, R41GM137625 and R42DK121652, and AWS Research Credits to T.R.P. Though there are 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 companies, Bioio, LLC and Healthspan Technologies, Inc., which conduct research related to this work. The spouse of N.L.S. is listed as an inventor on Issued U.S. Patent 8,080,371

"Markers for Addiction" covering the use of certain single nucleotide polymorphisms in determining the diagnosis, prognosis, and treatment of addiction.

METHODS

Code. All code created for this work is available at: https://github.com/tim-peterson/omniphenotype. All unreported p-values for correlation coefficients are to be assumed to be so small that R rounds them down to 0. Our analyses throughout rely on gene symbol identification. We used NCBI official gene symbols in all cases to obtain data for each gene. The biases using this approach penalize against gene symbols that use common English words, such as “MICE”, or that are short in length, e.g., less than 4 characters. These criteria do not bias for any particular biological function, therefore they appeared to us to be a reasonable compromise between being comprehensive with our analysis while still maintaining a feasible workload.

Budget analysis (Fig. 1A). A list of high-confidence gene interaction pairs was created by incorporating protein-protein interaction and co-dependency, a.k.a., co-essentiality or imputed synthetic interaction, analyses. Protein-protein interaction data was obtained from BioPlex at https://bioplex.hms.harvard.edu/interactions.php (32, 63). Interactions detected in both HEK293T and HCT116 cell lines comprised our high-confidence protein-protein interactions list of 21,729 interactions. Imputed synthetic interaction analyses were performed by identifying the top 100 correlating genes in the DepMap with the list of 21,729 (38).This resulted in a list of 5,824 interactions, which represent ~27% of the total. To calculate the cost of studying these pairs, we used the NIH grant funding database, Federal RePorter (65). Abstracts and project terms for NIH-funded grants were downloaded by year from 2000 to 2020. These datasets were queried to obtain grants that contained the names of both genes in each pair (of the 5,824) and

the funding amount (in US dollars) for that grant. The cumulative number of studied pairs was taken for each year and extrapolated using linear regression to find the date and cumulative dollar amount when all pairs would be studied. The dollar amounts were multiplied by how much is estimated to be needed to convert this funding to FDA approved drugs, i.e., 44,200/670 = 66X (33). From 2000 to 2020, $44,200 million is estimated to have been needed in total private investment to result in the approval of 18 drugs where $670 million in total public funding was used to develop the science that led to those 18 drugs.

Gene ontology and cell essential gene analysis (Fig. 3A). 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. By "essential", we refer to cell essential – required for viability of a cell – as opposed to organism essential – required for viability of an organism. Cell essential gene lists were obtained through CRISPR screens performed by Hart et al. and Wang et al. and haploid screens performed by Blomen et al (43-45). If knockdown of a gene resulted in a statistically significantly decreased cell count ($p < .05$ or FDR of 5%) then the gene was determined to be cell essential. There were ten cell lines tested: K562, KBM7, Jiyoye CS, Raji CS, HAP1, HCT116, DLD1, HeLa, GBM, RPE1. For the relaxed filter data point, we used cell-count based 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. 3B). Single gene inactivation cell count-based 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 single genes from the same GO category, which were chosen because they are

highly co-cited. Pair-wise Pearson correlations were then calculated using the python pearsonr() function. P-values were false discovery rate (FDR) corrected using Benjamini and Hochberg methods. In python, this is set using SciPy as shown here: https://github.com/tim-peterson/omniphenotype/blob/master/figure2B.py#L96. Co-citation gene-gene cell fitness enrichment analysis was performed by counting all genes in the top 20 with greater than one citation compared with the total number of genes with greater than one citation.

PubMed analysis (Fig. 4 and 5). In both Fig. 4 and 5, PubMed was programmatically accessed using the E-utilities (https://www.ncbi.nlm.nih.gov/books/NBK25497/) API endpoint: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi. From this endpoint, PubMed IDs (PMIDs), were counted (Fig. 4A-B, 5B), or abstracts were analyzed for the relevant information (Fig. 5A).

Cancer vs. non-cancer analysis (Fig. 4A). 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", which incorporates the terms "cancer", "malignancy", "tumors", and other neoplasm-related terminology. 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 disease 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. 4B). 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 sorted by correlation strength with a secondary mapping of the higher order GO categories also displayed.

GWAS Analysis (Fig 4C and 5C). Beta-values and p-values for each SNP-disease and SNP-drug pairing were obtained through the Ben Neale lab analysis of UKBiobank data: http://www.nealelab.is/uk-biobank. We assigned SNPs to protein-coding genes according to the GENCODE release 19 (GRCh37.p13) by specifying if a SNP overlapped any portion of a gene (first to last exon), that it pertained to that gene. This means that a SNP could be assigned to multiple genes. Beta-values were included if they were greater than 0.0031 and p-values were included if they were less than 0.001. The proportion of conditions with an associated SNP was obtained by finding the proportion of cancer and non-cancer conditions that had any associated SNP for each gene.

Regarding our determination of p and beta values thresholds: A consequence of the omnigenic model that most genes affect most phenotypes is that there must be a large number of SNPs with small effect sizes that influence each phenotype (31). This has been borne out in the study of polygenic risk scores, which assess the contributions of many DNA variants to the heritability of complex diseases (66-68). This is particularly evident in the study of height, which has been shown to be influenced by a large number of variants with small effect size (69, 70). Even with UKBiobank's large sample size we lack statistical power to find these variants using Bonferroni-corrected p-values. To search for these small-effect size variants, we applied a low threshold to both p-values and beta values when performing our analyses.

Disease Selection (Fig. 4C). A set of the top 100 diseases were selected by ranking MeSH terms in the Diseases category by how often they were co-cited with genes in PubMed. This produced a list of diseases that were the most commonly studied in molecular biology contexts, i.e., in the context of genes. Within this ranking, we used the datasets corresponding to the self-reported illness code in UKBiobank. We then separated these diseases into cancer and non-cancer diseases.

Drug response in patients vs. patient cells analysis (Fig. 5A). To identify studies that reported findings on patient cell drug responses that correlated with the patient we used two search terms that are commonly associated with culturing patient cells: "human lymphoblastoid cell lines" and "peripheral mononuclear blood cell". 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_name] AND "cell viability" OR "IC50".

Co-citation time series analysis (Fig. 5B). 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 co-annotated with other disease MeSH terms.

Drug Selection (Fig. 5C). Drugs were defined as cancer-associated if greater than 10% of their PubMed citations were co-cited with the MeSH term “Neoplasms”. Of the drug codes in UKBiobank categorized in the Neale lab dataset, 15 fit this categorization. An additional 15 drugs were chosen in non-cancer-associated category by ranking drugs by the number of cases in UKBiobank. The names of each drug are listed in Supp. Table 1.

REFERENCES

1. https://en.wikipedia.org/wiki/Eroom%27s_law.
2. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191-200. doi: 10.1038/nrd3681. PubMed PMID: 22378269.
3. https://ntrs.nasa.gov/api/citations/19820022093/downloads/19820022093.pdf.
4. https://web.archive.org/web/20181004193040/https://www.microsoft.com/en-us/research/publication/reaching-agreement-presence-faults/.
5. https://en.wikipedia.org/wiki/Byzantine_fault.
6. https://bitcoin.org/bitcoin.pdf.
7. 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.
8. 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.
9. 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.
10. 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.
11. Saxton RA, Sabatini DM. mTOR Signaling in Growth, Metabolism, and Disease. Cell. 2017;169(2):361-71. doi: 10.1016/j.cell.2017.03.035. PubMed PMID: 28388417.
12. 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.

13. 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.

14. Li D, March ME, Gutierrez-Uzquiza A, Kao C, Seiler C, Pinto E, Matsuoka LS, Battig MR, Bhoj EJ, Wenger TL, Tian L, Robinson N, Wang T, Liu Y, Weinstein BM, Swift M, Jung HM, Kaminski CN, Chiavacci R, Perkins JA, Levine MA, Sleiman PMA, Hicks PJ, Strausbaugh JT, Belasco JB, Dori Y, Hakonarson H. ARAF recurrent mutation causes central conducting lymphatic anomaly treatable with a MEK inhibitor. Nat Med. 2019;25(7):1116-22. Epub 2019/07/03. doi: 10.1038/s41591-019-0479-2. PubMed PMID: 31263281.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. Stossel TP. The discovery of statins. Cell. 2008;134(6):903-5. Epub 2008/09/23. doi: 10.1016/j.cell.2008.09.008. PubMed PMID: 18805080.

21. 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.

22. 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.

23. Rudmann DG. On-target and off-target-based toxicologic effects. Toxicol Pathol. 2013;41(2):310-4. doi: 10.1177/0192623312464311. PubMed PMID: 23085982.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70. PubMed PMID: 10647931.

30. 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.

31. 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;174(4):953-67 e22. Epub 2018/07/24. doi: 10.1016/j.cell.2018.06.010. PubMed PMID: 30033366; PMCID: PMC6426455.

32. Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock A, Zarraga G, Tam S, Szpyt J, Gassaway BM, Panov A, Parzen H, Fu S, Golbazi A, Maenpaa E, Stricker K, Guha Thakurta S, Zhang T, Rad R, Pan J, Nusinow DP, Paulo JA, Schweppe DK, Vaites LP, Harper JW, Gygi SP. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell. 2021;184(11):3022-40 e28. Epub 2021/05/08. doi: 10.1016/j.cell.2021.04.011. PubMed PMID: 33961781; PMCID: PMC8165030.

33. https://catalyst.phrma.org/new-report-demonstrates-industry-leadership-in-drug-development-in-partnership-with-nih.

34. Hanahan D. Rethinking the war on cancer. Lancet. 2014;383(9916):558-63. Epub 2013/12/20. doi: 10.1016/s0140-6736(13)62226-6. PubMed PMID: 24351321.

35. Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies in genomics. Genome Res. 2018;28(9):1255-63. Epub 2018/08/05. doi: 10.1101/gr.207464.116. PubMed PMID: 30076130; PMCID: PMC6120626.

36. Leeming G, Ainsworth J, Clifton DA. Blockchain in health care: hype, trust, and digital health. Lancet. 2019;393(10190):2476-7. Epub 2019/06/25. doi: 10.1016/s0140-6736(19)30948-1. PubMed PMID: 31232356.

37. Gürsoy G, Bjornson R, Green ME, Gerstein M. Using blockchain to log genome dataset access: efficient storage and query. BMC Med Genomics. 2020;13(Suppl 7):78. Epub 2020/07/23. doi: 10.1186/s12920-020-0716-z. PubMed PMID: 32693796; PMCID: PMC7372787.

38. 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.

39. Pan J, Meyers RM, Michel BC, Mashtalir N, Sizemore AE, Wells JN, Cassel SH, Vazquez F, Weir BA, Hahn WC, Marsh JA, Tsherniak A, Kadoch C. Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. Cell Syst. 2018;6(5):555-68 e7. doi: 10.1016/j.cels.2018.04.011. PubMed PMID: 29778836; PMCID: PMC6152908.

40. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, Humeidi R, Peck D, Wu X, Tang AA, Wang VM, Bender SA, Lemire E, Narayan R, Montgomery P, Ben-David U, Garvie CW, Chen Y, Rees MG, Lyons NJ, McFarland JM, Wong BT, Wang L, Dumont N, O'Hearn PJ, Stefan E, Doench JG, Harrington CN, Greulich H, Meyerson M, Vazquez F, Subramanian A, Roth JA, Bittker JA, Boehm JS, Mader CC, Tsherniak A, Golub TR. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat Cancer. 2020;1(2):235-48. Epub 2020/07/03. doi: 10.1038/s43018-019-0018-6. PubMed PMID: 32613204; PMCID: PMC7328899.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. 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.

49. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death.

50. http://www.nealelab.is/uk-biobank.

51. 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.

52. 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.

53. Surface LE, Burrow DT, Li J, Park J, Kumar S, Lyu C, Song N, Yu Z, Rajagopal A, Bae Y, Lee BH, Mumm S, Gu CC, Baker JC, Mohseni M, Sum M, Huskey M, Duan S, Bijanki VN, Civitelli R, Gardner MJ, McAndrew CM, Ricci WM, Gurnett CA, Diemer K, Wan F, Costantino CL, Shannon KM, Raje N, Dodson TB, Haber DA, Carette JE, Varadarajan M, Brummelkamp TR, Birsoy K, Sabatini DM, Haller G, Peterson TR. ATRAID regulates the action of nitrogen-containing bisphosphonates on bone. Sci Transl Med. 2020;12(544). Epub 2020/05/22. doi: 10.1126/scitranslmed.aav9166. PubMed PMID: 32434850.

54. 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.

55. 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.

56. Peterson TR, Head, R. Personal communication on single-cell RNA seq on mouse liver. 2017.

57. 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.

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F, Gygi MP, Parzen H, Szpyt J, Tam S, Zarraga G, Pontano-Vaites L, Swarup S, White AE, Schweppe DK, Rad R, Erickson BK, Obar RA, Guruharsha KG, Li K, Artavanis-Tsakonas S, Gygi SP, Harper JW. Architecture of the human interactome defines protein communities and

disease networks. Nature. 2017;545(7655):505-9. doi: 10.1038/nature22366. PubMed PMID: 28514442; PMCID: PMC5531611.

64. Hochberg Ba. Controlling the false positive rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995;57:289-300.

65. https://federalreporter.nih.gov/FileDownload.

66. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759-72. Epub 2020/07/28. doi: 10.1038/s41596-020-0353-1. PubMed PMID: 32709988.

67. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? Jama. 2019;321(18):1820-1. Epub 2019/04/09. doi: 10.1001/jama.2019.3893. PubMed PMID: 30958510.

68. Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010;42(7):558-60. Epub 2010/06/29. doi: 10.1038/ng0710-558. PubMed PMID: 20581876.

69. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76-82. Epub 2010/12/21. doi: 10.1016/j.ajhg.2010.11.011. PubMed PMID: 21167468; PMCID: PMC3014363.

70. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9. Epub 2010/06/22. doi: 10.1038/ng.608. PubMed PMID: 20562875; PMCID: PMC3232052.

Figure 1

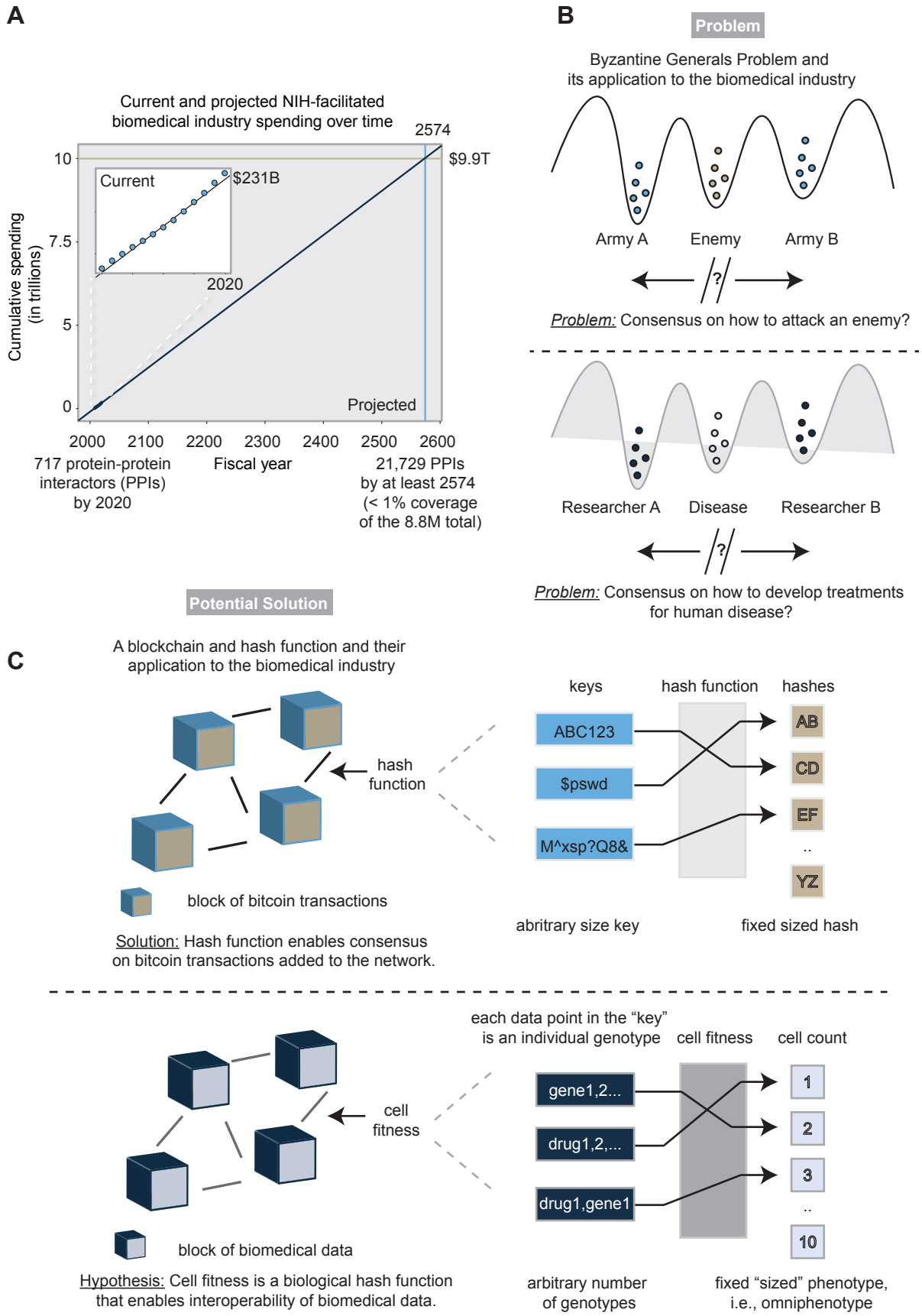

A
Current and projected NIH-facilitated biomedical industry spending over time
2574
$9.9T
Current
$231B
2020
Projected
Cumulative spending (in trillions)
10
7.5
5
0
2000
2100
2200
2300
2400
2500
2600
Fiscal year
717 protein-protein interactors (PPIs) by 2020
21,729 PPIs by at least 2574 (< 1% coverage of the 8.8M total)
B
Problem
Byzantine Generals Problem and its application to the biomedical industry
Army A
Enemy
Army B
?
Problem: Consensus on how to attack an enemy?
Researcher A
Disease
Researcher B
?
Problem: Consensus on how to develop treatments for human disease?
Potential Solution
C
A blockchain and hash function and their application to the biomedical industry
hash function
block of bitcoin transactions
Solution: Hash function enables consensus on bitcoin transactions added to the network.
keys
hash function
hashes
ABC123
$pswd
M^xsp?Q8&
AB
CD
EF
..
YZ
abritrary size key
fixed sized hash
each data point in the "key" is an individual genotype
cell fitness
cell count
cell fitness
block of biomedical data
Hypothesis: Cell fitness is a biological hash function that enables interoperability of biomedical data.
gene1,2...
drug1,2,...
drug1,gene1
1
2
3
..
10
arbitrary number of genotypes
fixed "sized" phenotype, i.e., omniphenotype

Figure 2

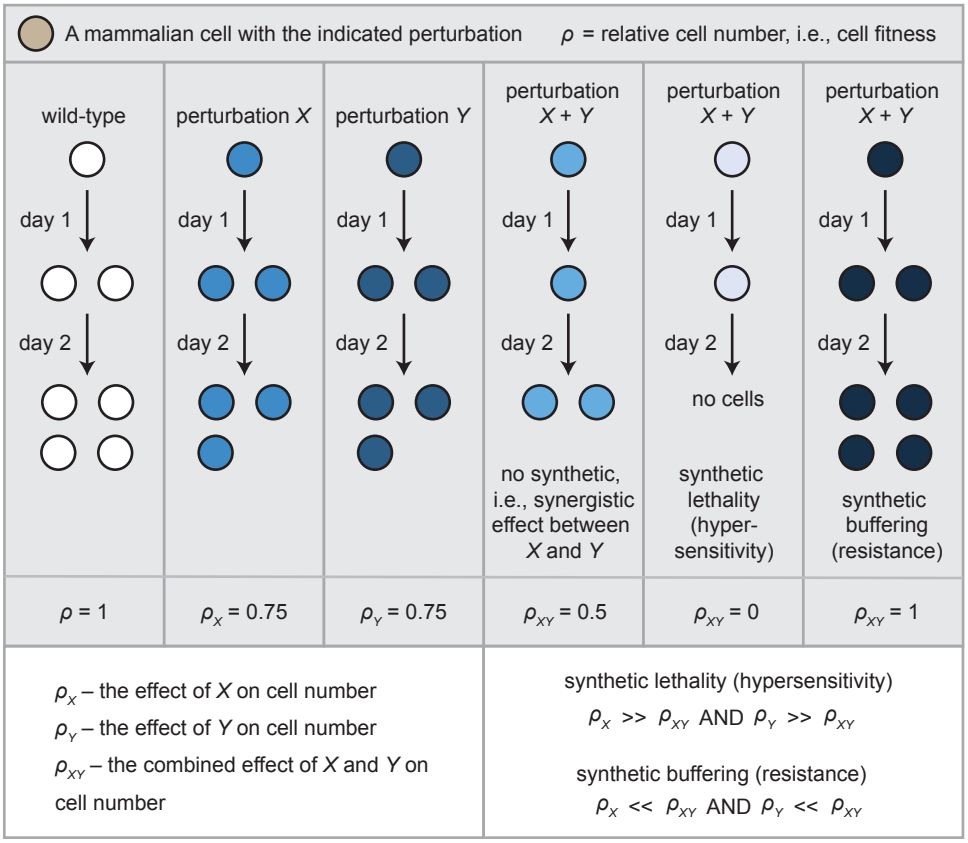

A mammalian cell with the indicated perturbation
ρ = relative cell number, i.e., cell fitness
wild-type
perturbation X
perturbation Y
perturbation X + Y
perturbation X + Y
perturbation X + Y
day 1
day 2
no cells
no synthetic, i.e., synergistic effect between X and Y
synthetic lethality (hyper-sensitivity)
synthetic buffering (resistance)
ρ = 1
ρX = 0.75
ρY = 0.75
ρXY = 0.5
ρXY = 0
ρXY = 1
ρX – the effect of X on cell number
ρY – the effect of Y on cell number
ρXY – the combined effect of X and Y on cell number
synthetic lethality (hypersensitivity)
ρX >> ρXY AND ρY >> ρXY
synthetic buffering (resistance)
ρX << ρXY AND ρY << ρXY

Figure 3

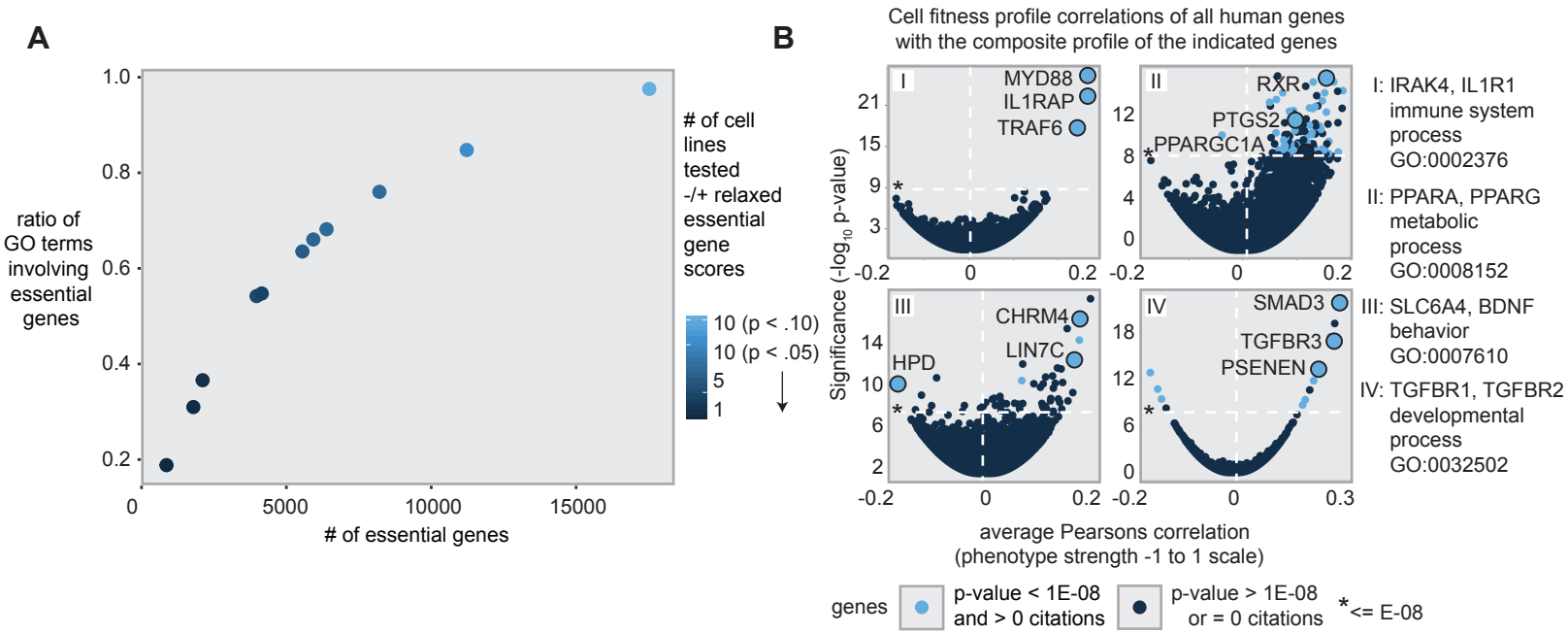

A
ratio of GO terms involving essential genes
1.0
0.8
0.6
0.4
0.2
0
5000
10000
15000
# of essential genes
# of cell lines tested -/+ relaxed essential gene scores
10 (p < .10)
10 (p < .05)
5
1
B
Cell fitness profile correlations of all human genes with the composite profile of the indicated genes
Significance (-log10 p-value)
I
MYD88
IL1RAP
TRAF6
II
RXR
PTGS2
PPARGC1A
III
CHRM4
LIN7C
HPD
IV
SMAD3
TGFBR3
PSENEN
average Pearsons correlation (phenotype strength -1 to 1 scale)
I: IRAK4, IL1R1 immune system process GO:0002376
II: PPARA, PPARG metabolic process GO:0008152
III: SLC6A4, BDNF behavior GO:0007610
IV: TGFBR1, TGFBR2 developmental process GO:0032502
genes
p-value < 1E-08 and > 0 citations
p-value > 1E-08 or = 0 citations
*<= E-08

Figure 4

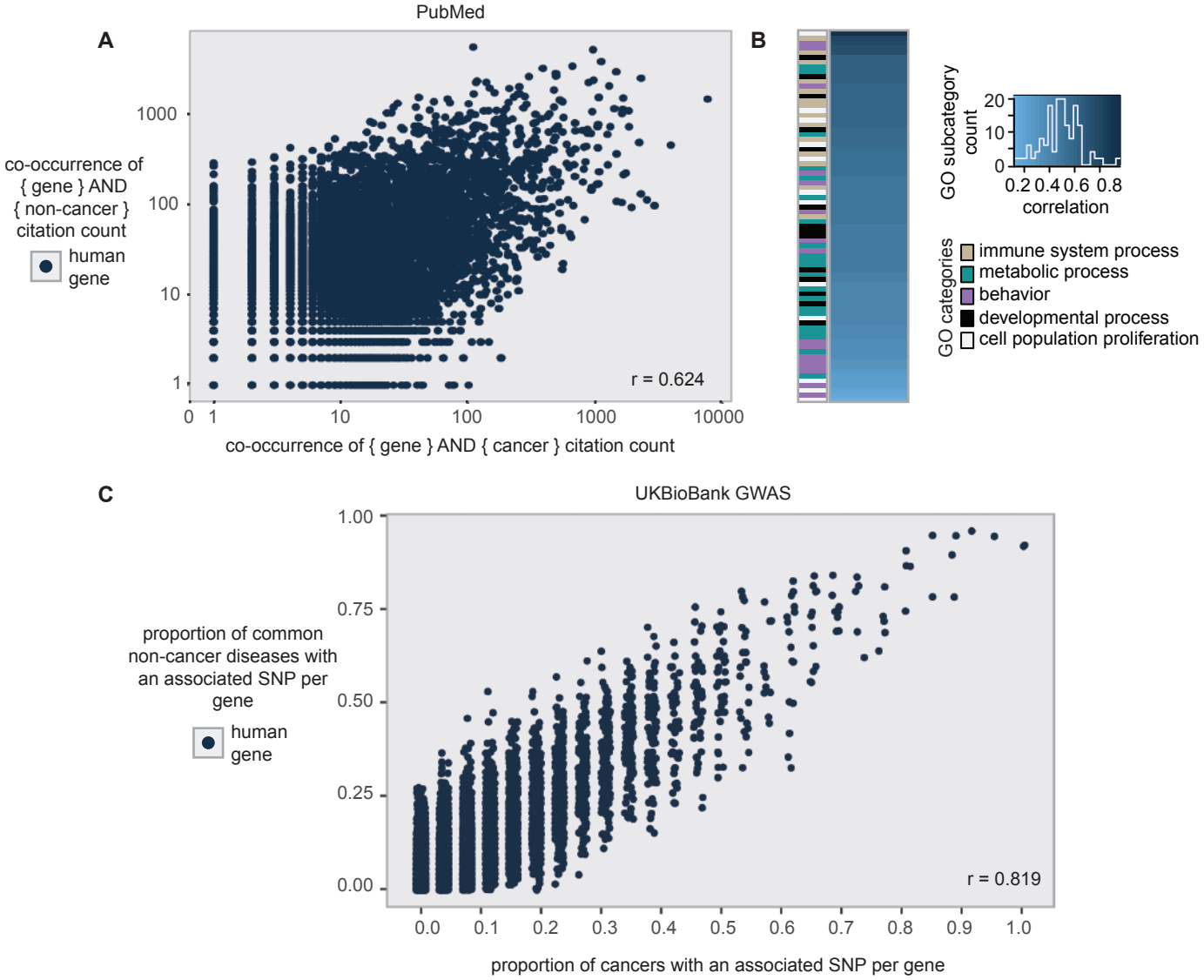
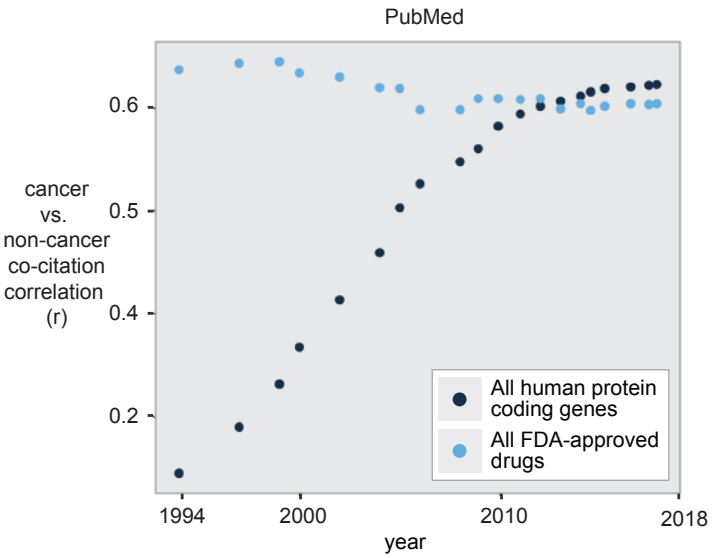

A
PubMed
co-occurrence of { gene } AND { non-cancer } citation count
human gene
1000
100
10
1
0
1
10
100
1000
10000
r = 0.624
co-occurrence of { gene } AND { cancer } citation count
B
GO subcategory count
20
10
0
0.2 0.4 0.6 0.8
correlation
GO categories
immune system process
metabolic process
behavior
developmental process
cell population proliferation
C
UKBioBank GWAS
proportion of common non-cancer diseases with an associated SNP per gene
human gene
1.00
0.75
0.50
0.25
0.00
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
r = 0.819
proportion of cancers with an associated SNP per gene

Figure 5

A

drug	human condition	citation (PMID)	IC50 range	citation(s) (PMID)
	patient drug response correlates with drug response on patient cells		drug IC50 value range on human cells or cell lines	
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

B



C

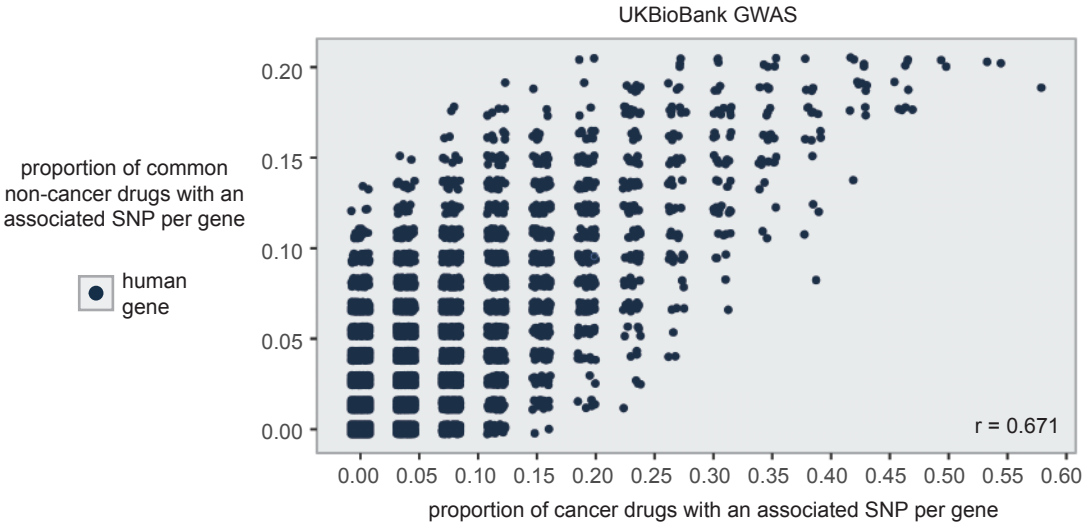


Figure 6

Framework for synthetic interaction testing

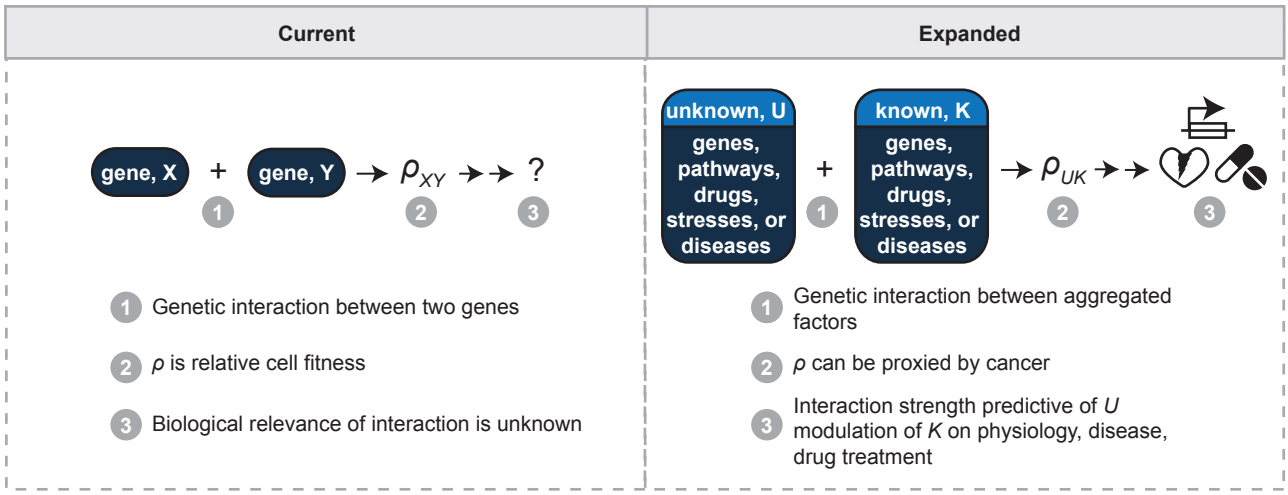


Figure S1

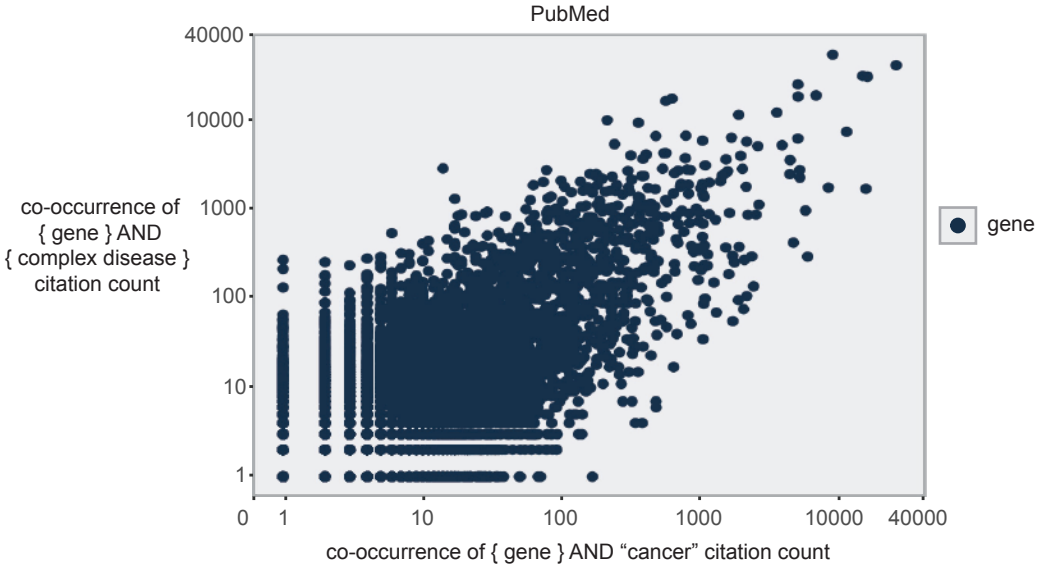


Figure S1: Nearly all common disease genes are cancer genes. The co-occurrences of each human gene with cancer (x-axis) and non-cancer (y-axis) conditions in all ~27 million PubMed citations were counted. The following non-cancer conditions were considered: infection, Alzheimer’s, cardiovascular disease, diabetes, obesity, depression, inflammation, osteoporosis, hypertension, stroke, and inflammation. Slope, 0.75; Correlation coefficient, r = 0.69.