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@bioio.tech (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. Herein, we propose a solution to this problem. 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 an 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. We provide proof of concept of the Omniphenotype model using the widely used osteoporosis drug, bisphosphonates, implicating its mechanism of action (MoA) genes, ATRAID, SLC37A3, and FDPS, as potential gerotargets for neurodegenerative conditions. 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 risks considering 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 with physics in helping solve the structure of DNA, 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 could be identified.

Consensus in biomedical research has historically taken the form of researchers agreeing on what biological model systems they will focus their labs around. Interestingly, relatively simple model systems such as yeast and worms have made a disproportionate contribution to human

health (7-9). 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). Moreover, evidence shows that even more complex models such as mice aren't representative models of human diseases (12, 13). Some researchers are nevertheless pushing forward and focusing on creating new ways to accelerate human disease research using model organisms (14-16). 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 (17). 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 (18), 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 (19). 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,

105 cholesterol-lowering statins were initially held back from the clinic because they were shown to
106 be toxic in animal models. Those early findings scared off pharmaceutical companies because
107 they suggested statins might not be safe for humans (20, 21). Merck eventually realized this
108 toxicity was on-target related to the drug mechanism of action and not generic off-target activity,
109 and the statins went on to be one of the biggest success stories in drug development history (22).
110 The successes of drug "repositioning" also reveal to us that drugs can function in multiple
111 "unrelated" contexts. For example, the malarial drug, hydroxychloroquine (Plaquenil), is now
112 routinely used to treat rheumatoid arthritis and autoimmune disorders (23). There is also
113 chemotherapeutics being considered for depression (24). Drug repositioning makes sense from
114 a molecular perspective because many drug side effects could actually be on-target effects (25).
115 To take it further, now with so much knowledge in the "-omic" era, it might be the rule rather than
116 the exception that a mechanistic target for a drug has many functions in the body. For example,
117 the serotonin transporter, SLC6A4 (a.k.a. SERT), is the target of the widely prescribed
118 antidepressants, Prozac and Zoloft. SLC6A4 is expressed at high levels in the intestines and
119 lungs (26), and regulates gut motility and pulmonary blood flow (27, 28) in addition to mood. This
120 demonstrates the need to find a better way to understand the diverse roles of drugs and their
121 targets.
122
123 Herein, we address these aforementioned issues. First, we quantify costs incurred for studying
124 proteins for eventual drug development and demonstrate that continuing with the status quo will
125 require prohibitive amounts of time and money. Then, we present evidence that synthetic
126 interaction testing using cell fitness provides a novel solution to this challenge. Specifically, by
127 leveraging diverse data types: gene-inactivation screens in cells, PubMed, and human genome-
128 wide association studies (GWAS), we demonstrate that cell fitness-based synthetic interaction
129 testing would be predicted to yield important insight on a new factor of interest (e.g., gene or drug)
130 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 (29, 30). This makes sense considering cancer has become defined by alterations in individual phenotypes that contribute to cell number/fitness, e.g., proliferation and apoptosis (31, 32). 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 omniphenotype – 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 viability (33). We asked how much time and money has been spent on studying these interactions to date.

When we look at a high-confidence list of 21,729 protein-protein interactions (PPIs) (34), NIH funded grants have studied 3.3% of these pairs (717 pairs) over the past 20 years and $3.5B has been spent on them. At the current rate, it can be estimated to take many more decades even centuries and costing hundreds of billions or even trillions 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 potentially up to 66X more (35) (Fig. 1A). Also, these numbers are for the 21,729 pairs, which is less than 1% of the total 8.8M potentially important ones. While these estimates do not factor in improvements in data analysis, e.g., artificial intelligence (AI), and PPIs are only a subset of the problem space, clearly, we could make improvements to understand human genes in a more time- 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) (36).

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 (37-39). 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 (40, 41). 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 (42). 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 explain 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.

208 A simple way to understand synthetic interactions is to consider two genes. It is said there is a
209 synthetic interaction between two genes if the combined effect of mutating both on cell fitness,
210 hereafter denoted by "ρ" (as in relative cell fitness or phenotype), is much greater than the effects
211 of the individual mutations (Fig. 2). If there is an interaction, it can be either be a negative synergy
212 (Fig. 2, 5th from left panel), where the double mutant cells grow much worse (a.k.a., synthetic
213 lethality), or a positive synergy, where the cells grow much better (a.k.a., synthetic buffering),
214 respectively, than the individual mutant cells (Fig. 2, 6th from left panel). Both phenotypes
215 demonstrate a synergy on a genetic level between these two genes).
216
217 Amongst the ~400 million possible synthetic interactions between protein coding human genes,
218 less than 0.1% have been mapped to date (43). It is challenging to scale synthetic interaction
219 testing in human cells both because of the size of the experiments needed and because
220 sometimes neither interactor has a known function in the ultimate biological context of interest
221 (44). Therefore, we sought to develop a framework that would allow us to infer synthetic
222 interactions at scale as well as leverage interactors of known significance to explain interactors
223 of unknown significance. The convention of characterizing an unknown in the context of a known
224 is a typical approach taken by biologists for making discoveries. To formalize our approach, we
225 hereafter refer to the known interactor in a synthetic interaction pair as *K,* and the unknown as *U.*
226 With this convention, *U* is what's new and interests the researcher, whereas *K* is a factor that's
227 known in the field. *U* and *K* could each represent a gene, disease, or drug. Throughout this work,
228 we provide three pieces of evidence that taken together provide a generalizable framework for
229 using cell fitness-based synthetic interaction testing to solve for *U* using *K*.
230
231 ***Evidence #1: Nearly all cell processes affect cell fitness***. Recent works by Hart et al., Blomen
232 et al., and Wang et al. identified "cell essential" genes – genes that affect cell fitness to the degree
233 that they are required for cell viability (45-47) – and we wondered what GO pathways they

234 participate in. In this case, we took the above published lists of cell essential genes and
235 considered the synthetic interaction testing as follows: the known, *K,* is the gene that is essential
236 in at least one cell type and the unknown, *U,* is the synthetically-interacting GO pathway of
237 interest. GO pathways are collections of genes. As precedence for this approach of considering
238 *U* as a collection of genes rather than an individual gene we applied the principal of the minimal
239 cut set (MCS). MCS is an engineering term used by biologists to mean the minimal number of
240 deficiencies that would prevent a signaling network from functioning (48, 49). MCS fits well with
241 the concept of synthetic lethality (29), and this was recognized by Francisco Planes and
242 colleagues to identify RRM1 as an essential gene in certain multiple myeloma cell lines (48). MCS
243 also relates to the concept of co-dependency where two genes are said to be co-dependent if
244 their gene deficiency cell fitness profiles correlate with each other across cell lines (50).
245
246 We generalized the idea of MCS to hypothesize that different cell types might be relatively
247 defective in components of specific GO pathways such that if gene(s) in one of these pathways
248 were knocked out, they would produce a synthetic lethal interaction with those components (For
249 clarity throughout this study, by “genes”, we refer to human protein coding genes). Applying this
250 MCS approach, we found the essential genes from the Hart, Blomen, and Wang studies are
251 involved in diverse GO processes and the number of GO processes they participate in scales with
252 the number of cell lines examined (Fig. 3A, Supp. Table 1). For example, if one cell line is
253 assessed, at most 2054 out of 21246 total genes assessed were essential and these essential
254 genes involve at most 36% of all possible GO categories (6444 out of 17793. 17793 represents
255 the number of unique – i.e., irrespective of the GO hierarchy - human gene-associated GO
256 categories). In examining five cell lines, 3916 essential genes were identified that involve 54% of
257 all GO categories (9577/17793). With ten cell lines, 6332 essential genes were identified that
258 involve 68% of all GO categories (12067/17793). Therefore, essential genes involve the majority
259 of GO processes. Because the Fig. 3A graph appears to approach 1, this suggests if enough cell

260 types are tested, potentially all genes and GO processes would be essential in at least some cell
261 context. Consistent with this, when we relaxed the significance on the cell fitness scores from $p <$
262 0.05 to $p < 0.10$, this relaxed threshold encompasses nearly all GO processes and genes (Fig.
263 3A(, 97.2%, 17291/17793 GO terms; 17473/21246 genes). Taken together, this suggests the
264 majority if not all GO processes and their associated genes affect cell fitness.
265
266 Recently, genome-scale gene inactivation screening has been performed on hundreds of human
267 cell lines (40). Using this data we reasoned that the single gene mutant cell fitness profiles for
268 genes in the same GO processes would strongly correlate across cell lines. We selected gene
269 pairs from diverse GO categories (immune system process (GO:0002376), behavior
270 (GO:0007610), metabolic process (GO:0008152), and developmental process (GO:0032502) and
271 asked how the fitness profiles of other members of the same GO category would rank compared
272 with all 17,634 fitness profiles that were assessed. These categories were chosen because they
273 are direct children of the top-level "Biological Process" GO category (GO:0008150, with a "is_a"
274 relationship) that cover broad areas of human health and have similar numbers of genes
275 associated with them. Impressively, highly co-cited genes from each top-level GO category were
276 highly correlated in their fitness profiles ($p < 1e-8$, Fig. 3B, Supp. Table 1). Gene co-citation is a
277 human determined connection between two genes whereas cell fitness correlations are machine
278 determined. Therefore, that all significantly correlating cell fitness profiles were also enriched as
279 being co-cited (compare light and dark blue in Fig. 3B, hypergeometric distribution model
280 enrichment p-values: immune system process $p = 1.4E-4$; metabolic process $p = 6.9E-3$;
281 development $p = 2.9E-3$; behavior $p = 2.8E-2$) suggests purely computational approaches using
282 cell fitness can accurately predict gene function for disparate biological processes.
283
284 ***Evidence #2: Nearly all common disease genes are cancer genes.*** The omnigenic model
285 states that all genes expressed in disease relevant tissues are important to the disease (51). 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). For context, the non-cancer literature includes conditions like the top ten most common causes of death according to the World Health Organization (52) 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 (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 a gene’s physiological function or those that focus on rare diseases could undermine our goal of making an apples-to-apples comparison of non-cancer diseases 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 (52) 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 diseases, 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 (31, 32).

PubMed can be considered biased because it is curated by humans and because co-citation does not necessarily tell us the reason for a gene and a disease being mentioned together. Therefore we considered the orthogonal approach of using genome-wide association studies (GWAS) to further characterize the cancer/non-cancer correlations 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 broad acceptance and large sample size (46), and specifically the widely-used summary GWAS results of this dataset from the Ben Neale lab (53). 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 (see Methods for more details) 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 citations 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.

364 We identified 25 studies on PubMed where patient derived cells were generated and treated with
365 drugs and where cell fitness was measured alongside the patients themselves being treated and
366 their drug responses monitored. What was notable in all the studies is that cell fitness was
367 predictive of patient response for widely divergent medications (Fig. 5A). These results might be
368 surprising outside the context of cancer therapy, but they arguably should not be because the
369 level of growth inhibition (inhibitory concentration - IC50) for these drugs varies widely in different
370 cell lines, which have different genetic/epigenetic backgrounds (Fig. 5A, Supp. Table 1).
371
372 Because we found a relatively small number of studies on PubMed that met our strict criteria, we
373 considered the cancer vs. non-cancer paradigm we used in Figure 4 to give us more
374 comprehensive insight on the relationship between drug response and cell fitness. Like with PTEN
375 as previously discussed, our understanding of the role of many genes in cancer has preceded
376 that of our understanding of their roles in other diseases. We reasoned that with several advances,
377 such as various new –omic analyses, our understanding of the molecular basis of many diseases
378 is “catching up” to cancer more recently. Thus, we expect the correlation between all human
379 protein coding genes and all FDA-approved drugs being cited both in and non-cancer to be
380 increasing over time. Indeed, human genes are increasingly being cited with similar frequency in
381 cancer and non-cancer disease contexts (Fig. 5B, Supp. Table 1). This supports the idea that one
382 should be able to increasingly make use of the literature rather than perform experiments to
383 understand an unknown gene’s function. Like with human genes, we found that clinical drugs are
384 commonly cited in both the cancer and non-cancer literature (correlation = 0.599), albeit this
385 correlation hasn’t increased over time like with genes (Fig. 5B, Supp. Table 1). This finding
386 suggests drug repositioning opportunities for non-cancer conditions that are underexplored given
387 the current mindset that identifying drug targets for cancer would require separate efforts from
388 those for non-cancer conditions.
389

390 Lastly, to make an analogous analysis with drugs as we did with diseases in Figure 4, we binned
391 the top cancer and non-cancer drugs taken by participants in UKBioBank. Figure 5C shows the
392 proportion of drugs with an associated SNP for both cancer and non-cancer drugs. The cancer
393 vs. non-cancer drug correlation is strong (r = .66) and like with cancer vs. non-cancer diseases in
394 Figure 4 and remains statistically significant when accounting for gene length (Fig. 5C). Combined
395 with the results of our PubMed analysis, these studies lend support to the increased interest in
396 repositioning many non-cancer drugs for cancer (54) (e.g., immunosuppressant, rapamycin and
397 anti-diabetic, metformin), and vice versa (e.g., chemotherapeutics for autism, senolytics for aging
398 (55).
399
400 ***Proof of concept of the Omniphenotype model using bisphosphonates and its MoA target***
401 ***genes, ATRAID, SLC37A3, and FDPS****.* We recently published a proof-of-concept of the
402 Omniphenotype model – which can be thought of as an expanded definition of synthetic
403 interaction testing – with the osteoporosis drugs, nitrogen-containing bisphosphonates (43, 56,
404 57). In this work, we identified a gene network, ATRAID-SLC37A3-FDPS, using cell-fitness-
405 based, drug interaction screening with the widely used osteoporosis drugs, bisphosphonates, and
406 subsequently demonstrated this network controls bisphosphonate responses in cells, in mice, and
407 in humans (43, 56, 57). In this case, bisphosphonates and FDPS were the known, *K,* and ATRAID
408 and SLC37A3 were the unknown, *U,* that we discovered were involved in the bisphosphonates
409 MoA.
410
411 In addition to elucidating drug MoA, another strength of the Omniphenotype model is its potential
412 to make new insights into disease biology and therapeutics applications. To take our
413 bisphosphonates findings a step further, we were interested in two questions: 1) How can the
414 bisphosphonates MoA be distinguished from the statins MoA, which also targets the mevalonate
415 pathway but at a different gene, HMGCR, upstream in the pathway (58)? 2) Where in the body

416 might bisphosphonates be therapeutic besides bone? To answer these questions and extrapolate
417 from our cell fitness data to in vivo phenotypes we made use of International Mouse Phenotyping
418 Consortium (IMPC) (59), which to date has performed whole body phenotyping on 6,440 knockout
419 mice.
420
421 We were interested to determine how the statins and bisphosphonates MoAs might be
422 distinguished by their in vivo effects. Because ATRAID, SLC37A3, FDPS, and HMGCR deficient
423 mice have not yet been generated by the IMPC, we sought a systematic, well-powered approach
424 that would enable us to indirectly measure the in vivo effects of their loss of function. Using the
425 MCS concept, we wondered to what extent might ATRAID, SLC37A3, FDPS, and HMGCR target
426 genes deficiency, in aggregate, reflect single gene ATRAID, SLC37A3, FDPS, and HMGCR
427 deficiency. By “target genes” we mean genes whose expression levels were affected by ATRAID,
428 SLC37A3, FDPS, and HMGCR deficiency.
429
430 To test the validity of this approach we started with bone density because we had an expectation
431 of the sign conventions of the effects we might see based on our previous work and that from
432 others where ATRAID and SLC37A3 inhibition negatively affected bone density and HMGCR and
433 FDPS inhibition positively affected it (57, 60-62). Indeed, based on a comparison of hundreds to
434 thousands of mice deficient in ATRAID, SLC37A3, FDPS, or HMGCR target genes with over
435 16,000 control mice, ATRAID and SLC37A3 target gene deficiency decreased bone density and
436 FDPS and HMGCR target gene deficiency increased bone density (Fig. 6A).
437
438 Like with bone density, our cell fitness data predicted that ATRAID and SLC37A3 inhibition should
439 negatively regulate neurological behaviors and FDPS inhibition should positively regulate it.
440 Moreover, FDPS inhibition should have stronger effects than HMGCR inhibition. Consistent with
441 these predictions, ATRAID and SLC37A3 target gene deficient mice had decreased cognitive

442 function and motor activities (Fig. 6B). For both activity tests, the beneficial effects of FDPS target
443 gene deficiency were statistically significantly greater than those of HMGCR target gene
444 deficiency (Fig. 6A, B). For the other phenotypes including bone density, FDPS target gene
445 deficiency trended to being more beneficial than HMGCR target gene deficiency. These results
446 provide a preliminary indication that ATRAID and SLC37A3 inhibition might promote
447 neurodegenerative symptoms and FDPS inhibition, e.g., with bisphosphonates treatment, would
448 potentially alleviate them and do so more than HMGCR inhibition.
449
450 **DISCUSSION**
451
452 We propose cell fitness as a simple, yet comprehensive approach to rank less-studied genes, cell
453 processes, diseases, and therapies in importance relative to better-studied ones (Fig. 7). Rather
454 than needing ever more refined experiments, this work suggests the antithetical approach. That
455 is, one can use fitness in mammalian cells to estimate the importance of a poorly characterized
456 factor of interest (herein referred to as *U*) in a biological context of interest by comparing the new
457 factor's individual effect on cell fitness vs. its combined with a factor with a known role in the
458 biological context of interest (herein referred to as *K*) (Fig. 7). Besides the bisphosphonates and
459 the statins, there are numerous other examples in the literature such as with the mTOR pathway
460 and rapamycin (63) where the steps from cell fitness to phenotypes relevant in people can be
461 readily traced.
462
463 Cell fitness is a phenotype that has long had traction in the yeast and bacteria communities, but
464 not yet in mammalian systems outside of cancer. To increase awareness of its utility, below we
465 address its relevance as a phenotype to the different levels at which mammalian biologists tackle
466 their work: at the level of cell and molecular processes, at the level of disease, and at the level of
467 environmental factors such as drugs.

468

469 *Relevance to cell and molecular processes (Fig. 3).* Our Figure 3 results suggest one can study
470 the majority of cell and molecular processes using cell fitness-based, synthetic interaction testing.
471 Specifically, Figure 3 shows that to the extent that cell and molecular processes are well
472 represented by GO terms, essential genes as determined by cell fitness (cell counting) are
473 involved in the vast majority of such processes. This is useful information for molecular biologists
474 because we often spend considerable time developing experimental models. For example, large
475 efforts are often spent upfront (before getting to the question of interest) developing technology,
476 such as cell types to model various tissues. This is done because of the often-unquestioned
477 assumption that the cell type being studied matters. While this can be true, it is worth considering
478 whether it's the rule or the exception. In support of the latter, in single cell RNAseq analysis, 60%
479 of all genes are expressed in single cells and the percentage jumps to 90% when considering 50
480 cells of the same type (64). Similar percentages are obtained at the population level when
481 comparing cell types (65, 66). This argues that the majority of the time, there should be a generic
482 cell context to do at least initial "litmus" synthetic interaction testing of a hypothesis concerning a
483 new factor of interest, *U,* before proceeding with a more complex set-up. Like with cellular
484 phenotypes, many molecular phenotypes, such as reporter assays require domain expertise, can
485 be hard to elicit, and might not be easily reproducible or scalable. Whereas cell fitness is robust
486 and readily approximated by inexpensive assays, such as those to measure ATP levels.

487

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

493

494 One counterintuitive insight from Figure 4 is that core genes on the diagonal might not be as good
495 drug targets as genes off the diagonal because the core genes are involved in more diseases.
496 This might mean that because of their importance, core genes might be too networked with other
497 genes to be targeted by therapies without the therapies causing off-target effects. Fortunately,
498 this might then also mean those genes off the diagonal might make more appealing drug targets
499 because they are less networked. Thus, in this case we would use cell fitness as a filter to prioritize
500 genes for further study. This is important because like with income inequality where the "rich get
501 richer" certain genes get cited over and over (44), and approaches to aggregate data are known
502 to favor the "discovery" of already well-studied genes over less studied genes (67). When applied
503 in unbiased methods such as with genome-wide CRISPR-based screening (43), we expect our
504 approach to be particularly impactful in addressing this issue of identifying important genes that
505 historically are ignored.
506
507 *Relevance to environmental factors, e.g., drugs* (Fig. 5). With the exception of oncology, it could
508 be argued pharmacogenomics is only slowly going mainstream in many fields. One reason for
509 this is the poor feasibility and high cost-to-benefit ratio of current genetic testing strategies to
510 predict drug response (68). Though more work needs to be done to test the generalizability and
511 "real-world" application of the studies we highlight in Figure 5A, they do provide important initial
512 evidence for exploring cell fitness as a diagnostic assay for precision medicine. In addition to drug
513 diagnostics, these studies have implications for drug discovery. Both target-based and
514 phenotypic-based drug discovery have drawbacks (69). Target-based approaches don't
515 necessarily tell you about the phenotype, and phenotypic screens don't tell you about the target.
516 On the other hand, cell fitness-based, drug response genetic screening (70) can potentially be
517 the best of both worlds because it extrapolates well to the phenotype of interest and provides the
518 gene target in the same screen.
519

520 Still, we acknowledge limitations to our findings. Our work focuses on genes and drugs as the
521 genetic and environmental variables of interest. Future work must determine to what extent our
522 work would apply to other variables, such as epigenetic signatures. PubMed-based analyses are
523 subject to biases. Many genes are not published on due to human factors. Even among those
524 genes that are published, their representation is affected by publication bias towards positive
525 results. Additionally, our citation analysis relies on curation by NCBI, which due to its high
526 stringency and the current limitations of natural language processing strategies, likely leaves out
527 a large number of studies that should be included. However, this is why we also analyzed GWAS,
528 which provides a more unbiased look. As more methods get developed and more knowledge
529 accumulates, a scientist's job of figuring out where to focus their attention gets harder. Our view
530 is that cell fitness could be used more often to help scientists triage their opportunities.
531
532 **FIGURE LEGENDS**
533
534 **Figure 1 | The problems of and a potential solution for the lack of consensus-building**
535 **mechanisms in biomedical research. (A)** Current and projected time and money spending
536 estimates to develop drugs for high-confidence protein-protein interaction (PPI) pairs obtained
537 from Huttlin et al. (34, 71). The projected budget is a line of best fit based on the cumulative NIH
538 grant money spent on studying gene pairs multiplied by how much is needed to convert this
539 funding to FDA approved drugs, i.e., 44,200/670 = 66X (35). The inset graph represents the
540 current spending estimate that is extrapolated to make the projected spending and time estimates.
541 The intersection of the light blue and tan lines represent the projected date and costs, respectively,
542 when all drugs will be identified for each PPI. **(B)** A schematic of the Byzantine Generals Problem
543 and its application to the biomedical industry. In a real-world scenario, achieving consensus
544 becomes much more difficult when there are many parties involved (greater than the two
545 depicted) that are not working together. **(C)** Data blocks are chained together using a robust

546 algorithm called a hash function to create consensus understanding. In the case of bitcoin, the
547 consensus determines who owns what bitcoin. To make an analogy on how the concept of a
548 blockchain can address Eroom's law, we hypothesize that cell fitness is a biological hash function
549 that enables interoperability of biomedical data from different research efforts.
550
551 **Figure 2 | A current, commonly used framework for synthetic interaction testing.** A synthetic
552 interaction between two perturbations, *X* and *Y*, means the effect of *X* combined with *Y* on cell
553 fitness significantly differs from the expected effect based on each effect alone; typically, an
554 additive effect is expected. Synthetic interactions can cause lethality or hypersensitivity, or can
555 be buffering or confer resistance. Both are depicted in the figure where ρ – the phenotype – is
556 relative cell number. Synthetic effects can be thought of as multiplicative or exponential rather
557 than linear or additive, as is the case in the condition (4th column from left) where 1 - ρ_x (0.75) +
558 1 - ρ_y (0.75) = ρ_{xy} = 0.5. Note: In some scenarios, a perturbation is lethal on its own, which
559 precludes an investigation of its synthetic lethal but not its buffering interactions. On the other
560 hand, the lack of a perturbation's effect on its own on cell fitness does not preclude one from
561 determining if it has a synthetic interaction with another perturbation.
562
563 **Figure 3 | Nearly all GO processes are essential for cell fitness. (A)** The y-axis is the ratio of
564 the number of distinct GO terms linked to the essential genes in the tested cell lines divided by
565 the total number of distinct GO terms (17,793) for all tested human genes (21,246). The x-axis is
566 the number of essential genes determined for the indicated number of cell lines. "Relaxed" means
567 $p < 0.10$ (as opposed to $p < 0.05$) cell fitness scoring. **(B)** Fitness profile correlations across 563
568 human cell lines of mutant cells of genes from the same GO categories. The statistically significant
569 genes that have been most highly co-cited with the indicated genes are color-coded light blue. P-
570 values were false discovery rate (FDR) corrected using the Benjamini and Hochberg method (72).
571

572 **Figure 4 | Nearly all common disease genes are cancer genes. (A)** The co-occurrences of
573 each human gene with cancer (x-axis) and non-cancer (y-axis) conditions in all currently existing
574 ~30 million PubMed citations were counted. The non-cancer conditions include all MeSH terms
575 in the Disease category not classified with the "neoplasm" MeSH classifier. It includes common
576 conditions like infection, Alzheimer's, cardiovascular disease, diabetes, obesity, depression,
577 inflammation, osteoporosis, hypertension, and stroke amongst many other MeSH classifiers. Line
578 of best fit - Slope, 0.67; Correlation coefficient, r = 0.624. **(B)** Gene ontology (GO) classifications
579 that comprise the non-cancer/cancer correlation in (A). Categories are sorted from dark blue
580 representing a stronger correlating GO category to lighter blue representing a weaker correlating
581 category. **(C)** The proportion of non-cancer diseases vs. cancer diseases that have an associated
582 SNP (p-value < .001) in UKBioBank for each protein-coding human gene. These proportions are
583 out of the top 100 conditions, 26 cancer diseases and 74 non-cancer diseases, based on co-
584 citation with genes. Correlation coefficient, r = .819.
585
586 **Figure 5 | Establishing the efficacy of FDA-approved drugs using cell fitness. (A)** Drugs
587 where their effects on a patient's condition and the fitness of the same patient's cells correlate.
588 Listed drugs met two criteria: 1) Their effect on a patient's condition and the same patient's cell
589 fitness were significantly correlated; 2) The drug had differing IC50 values – the concentration at
590 which cell numbers are 50% of their untreated values – in multiple, distinct human cell types or
591 cell lines. **(B)** Correlation (r) over time of human protein-coding genes or FDA-approved drugs co-
592 cited with cancer or non-cancer conditions analyzed as in Fig. 3A. The most recent data point of
593 genes corresponds to the same data that was used in Figure 4A. (**C**) The proportion of non-cancer
594 drugs vs. cancer drugs that have a significant SNP (p-value < .001) in UKBioBank for each
595 protein-coding human gene. These proportions are out of a set of 15 drugs used in UKBioBank
596 patients who had cancer vs. 15 other drugs used in non-cancer conditions that had the highest
597 number of cases in the UKBioBank. Correlation coefficient, r = .671.

598

599 **Figure 6 | The Omniphenotype model proof of concept using bisphosphonates target**
600 **genes.** In vivo skeletal and neural phenotyping of bisphosphonates MoA target gene output. **(A)**
601 The phenotypes shown are those of aggregated knockout lines of genes of correlating fitness
602 profiles with the indicated bisphosphonates MoA target genes. Descriptive statistics (mean,
603 standard deviation, number of samples) are shown. Red line indicates mean. P-values were
604 obtained using one-way ANOVA corrected for multiple comparisons using the Tukey Method.
605 Number of gene knockout mouse lines considered for each bisphosphonate MoA genotype: bone
606 density | ATRAID/SLC37A3: 23, HMGCR: 66, FDPS, 25. **(B)** Data was analyzed similar to A.
607 Number of gene knockout mouse lines considered for each bisphosphonate MoA genotype:
608 cognitive function | ATRAID/SLC37A3: 23, HMGCR: 9, FDPS, 5. Locomotor Activity |
609 ATRAID/SLC37A3: 11, HMGCR: 18, FDPS, 7. Grip Strength | ATRAID/SLC37A3: 8, HMGCR:
610 12, FDPS, 3.

611

612 **Figure 7 | Current vs. expanded conceptual frameworks for synthetic interaction testing.** A
613 currently, commonly used framework for synthetic interaction testing is as follows: 1) Genetic
614 interactions are tested between two genes, e.g., *X* and *Y*; 2) ρ, the phenotype being measured,
615 is relative cell fitness; 3) It isn't known how the interaction strength between *X* and *Y* might be
616 relevant to physiology or disease. The expanded framework described herein is as follows: 1)
617 Genetic interactions can be tested between aggregated factors such as multiple genes and/or
618 pathways (related to Figure 3); 2) ρ can be proxied by the relative involvement of the interactors
619 in cancer (related to Figure 4); 3) One interactor, *K*, has a known effect on a physiology or disease
620 of interest. The effect of the other, *U*, in that context can therefore be measured relative to *K*
621 (related to Figure 5). The implications of mammalian cell fitness-based synthetic interaction
622 testing as described here is that the results using cell fitness can be readily extrapolated to the
623 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 grants from the NIH: NIH/NIA K99AG047255, R00AG047255, and NIH/NIAMS R01AR073017 (Peterson PI), and AWS Research Credits to T.R.P. BIOIO was also supported by grants from the NIH: NIH/NIDDK R42DK121652, NIH/NIGMS R41GM137625 (Peterson PI).

CONFLICTS OF INTEREST

T.R.P is the founder of Bioio, Inc. 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 (34, 71). 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 (40).This resulted in a list of 5,824 interactions, which represent ~27% of the total. To calculate the cost of studying these pairs, we

676 used the NIH grant funding database, Federal RePorter (73). Abstracts and project terms for NIH-
677 funded grants were downloaded by year from 2000 to 2020. These datasets were queried to
678 obtain grants that contained the names of both genes in each pair (of the 5,824) and the funding
679 amount (in US dollars) for that grant. The cumulative number of studied pairs was taken for each
680 year and extrapolated using linear regression to find the date and cumulative dollar amount when
681 all pairs would be studied. The dollar amounts were multiplied by how much is estimated to be
682 needed to convert this funding to FDA approved drugs, i.e., 44,200/670 = 66X (35). From 2000
683 to 2020, $44,200 million is estimated to have been needed in total private investment to result in
684 the approval of 18 drugs where $670 million in total public funding was used to develop the
685 science that led to those 18 drugs.
686
687 **Gene ontology and cell essential gene analysis (Fig. 3A).** A list of all human genes was
688 obtained from NCBI, ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/. Gene ontology
689 (GO) information for each human gene was obtained from BioMart,
690 https://useast.ensembl.org/info/data/index.html. By "essential", we refer to cell essential –
691 required for viability of a cell – as opposed to organism essential – required for viability of an
692 organism. Cell essential gene lists were obtained through CRISPR screens performed by Hart et
693 al. and Wang et al. and haploid screens performed by Blomen et al (45-47). If knockdown of a
694 gene resulted in a statistically significantly decreased cell count ($p < .05$ or FDR of 5%) then the
695 gene was determined to be cell essential.(45-47). There were ten cell lines tested: K562, KBM7,
696 Jiyoye CS, Raji CS, HAP1, HCT116, DLD1, HeLa, GBM, RPE1. For the relaxed filter data point,
697 we used cell-count based fitness scores at $p < 0.1$ for both the Wang and Hart studies. For the
698 Blomen study, p-values > 0.05 weren't given, therefore we took all genes with fitness scores less
699 than 0.5 standard deviations above the mean.
700

701 **Gene inactivation fitness profile correlation (Fig. 3B).** Single gene inactivation cell count-
702 based cell fitness profiles from the Broad Institute and Sanger Institute were downloaded from
703 the DepMap website, https://depmap.org/portal/download/ using the 2019q2 dataset. Fitness
704 profiles for single genes from the same GO category, which were chosen because they are highly
705 co-cited. Pair-wise Pearson correlations were then calculated using the python pearsonr()
706 function. P-values were false discovery rate (FDR) corrected using Benjamini and Hochberg
707 methods. In python, this is set using SciPy as shown here: https://github.com/tim-
708 peterson/omniphenotype/blob/master/figure3B.py#L96. Co-citation gene-gene cell fitness
709 enrichment analysis was performed by counting all genes in the top 20 with greater than one
710 citation compared with the total number of genes with greater than one citation.
711
712 **PubMed analysis (Fig. 4 and 5).** In both Fig. 4 and 5, PubMed was programmatically accessed
713 using the E-utilities (https://www.ncbi.nlm.nih.gov/books/NBK25497/) API endpoint:
714 https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi. From this endpoint, PubMed IDs
715 (PMIDs), were counted (Fig. 4A-B, 5B), or abstracts were analyzed for the relevant information
716 (Fig. 5A).
717
718 *Cancer vs. non-cancer analysis (Fig. 4A).* All PMIDs for each human gene and their homologs
719 and all PMIDs for each Medical Subject Headings (MeSH,
720 https://www.nlm.nih.gov/mesh/meshhome.html) were collected into tab-delineated files. A citation
721 was determined to be associated with a particular MeSH term if it came up from searching
722 PubMed with "<MeSH term> [mesh]". A citation was determined to be associated with a particular
723 gene if it has been labeled in "Related articles in PubMed" within the Gene resource of NCBI. See
724 Omniphenotype Github repository for how these application programming interface (API) calls
725 were made. The gene-PMID file was intersected with the MeSH-PMID file, and the resulting gene-
726 MeSH term list was separated by the MeSH term "neoplasm", which incorporates the terms

727 “cancer”, “malignancy”, “tumors”, and other neoplasm-related terminology. Meaning, all citations
728 that were annotated as referring to a gene that were co-annotated with the MeSH term “neoplasm”
729 were considered a “cancer” citation and all other citations that were co-annotated as referring to
730 a gene and any other disease MeSH term besides “neoplasm” were considered “non-cancer”
731 citations. Those citations which mapped to both neoplasm as well as other conditions were
732 considered “cancer” citations. Genes with less than 10 citations were excluded from the analysis.
733 The correlation co-efficient of cancer vs. non-cancer citation counts for all genes was calculated
734 using R using the lm() function on log-log data.
735
736 *Gene Ontology-subsetted, cancer vs. non-cancer citation analysis (Fig. 4B).* To determine the
737 diversity of correlations for cancer vs. other conditions among GO categories, we took a subset
738 of major GO categories (immune system process, metabolic process, behavior, developmental
739 process, and cell population proliferation) as a representation of a diverse subsampling of
740 biological processes. We then iterated through the sub-category/child terms of each of these
741 major GO categories and found the correlation of all cancer vs. non-cancer citations for genes
742 associated with that term. The results are displayed as a 1-Dimensional heatmap sorted by
743 correlation strength with a secondary mapping of the higher order GO categories also displayed.
744
745 *GWAS Analysis (Fig. 4C and 5C).* Beta-values and p-values for each SNP-disease and SNP-drug
746 pairing were generated using the Ben Neale lab analysis of UKBiobank data:
747 http://www.nealelab.is/uk-biobank. SNPs were assigned to protein-coding genes according to the
748 GENCODE release 19 (GRCh37.p13) by specifying if a SNP overlapped any portion of a gene
749 (first to last exon), then it pertained to that gene. This means that a SNP could be assigned to
750 multiple genes. The proportion of conditions with an associated SNP was obtained by finding the
751 proportion of cancer and non-cancer conditions that had any associated SNP for each gene. The
752 phrase “associated SNP” as used throughout refers to the following thresholds: a p-value less

753 than 0.001 (10^{-3}) and a beta-value greater than 0.0031. The logic for these thresholds derive from
754 the omnigenic model which posits that a large number of genetic variants with small effect size
755 influence each phenotype (31). This model has been borne out in many studies including those
756 computing polygenic risk scores, which assess the contributions of many DNA variants to the
757 heritability of complex diseases and traits (74-76). For example, height is influenced by a large
758 number of variants with small effect size (77, 78). Consistent with the omnigenic model, using
759 traditional GWAS thresholds such as 10^{-8} for p-values we found few genes that met this threshold
760 per UKBiobank dataset, which precluded our analysis of cancer vs. non-cancer correlations. To
761 ensure robust results, we tested a range of p-values and beta-values and established 10^{-3} and
762 0.0031, respectively, as producing high gene coverage while still being stringent compared to
763 thresholds such as p-value < 0.05 that are generally accepted with smaller datasets.
764
765 *Disease Selection (Fig. 4C).* A set of the top 100 diseases were selected by ranking MeSH terms
766 in the Diseases category by how often they were co-cited with genes in PubMed. This produced
767 a list of diseases that were the most commonly studied in molecular biology contexts, i.e., in the
768 context of genes. Within this ranking, we used the datasets corresponding to the self-reported
769 illness codes in UKBiobank. We then separated these diseases into cancer and non-cancer
770 diseases.
771
772 *Drug response in patients vs. patient cells analysis (Fig. 5A).* To identify studies that reported
773 findings on patient cell drug responses that correlated with the patient we used two search terms
774 that are commonly associated with culturing patient cells: “human lymphoblastoid cell lines” and
775 “peripheral mononuclear blood cell”. The phrase “human lymphoblastoid cell lines proliferation”
776 was queried in PubMed and 804 abstracts were returned. “PMBC” as in peripheral mononuclear
777 blood cells was substituted for lymphoblastoid cell lines (LCLs) to identify the simvastatin study.
778 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.

Bisphosphonate proof of concept (Fig. 6). The inclusion criteria for a gene knockout mouse line to be included in the analysis was as follows: ATRAID or SLC37A3 deficient cells showed < 0.75 expression levels of the gene relative to wild-type and < 1 for both ATRAID and SLC37A3. HMGCR or FDPS inhibited cells showed > 1.5-fold increased expression of the gene relative to wild-type, which means the phenotyping data for these knockout lines had to be ‘inverted’ by

normalizing to wild-type. Gene expression from a microglia proxy (K562) cells was used, but data from a neuronal proxy (HEK293T) showed a similar pattern (data not shown). Phenotyping data is from the International Mouse Phenotyping Consortium (IMPC).

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.

830 11. Saxton RA, Sabatini DM. mTOR Signaling in Growth, Metabolism, and Disease. Cell.
831 2017;169(2):361-71. doi: 10.1016/j.cell.2017.03.035. PubMed PMID: 28388417.
832 12. Seok J, Warren HS, Cuenca AG, Mindrinos MN, Baker HV, Xu W, Richards DR,
833 McDonald-Smith GP, Gao H, Hennessy L, Finnerty CC, Lopez CM, Honari S, Moore EE, Minei
834 JP, Cuschieri J, Bankey PE, Johnson JL, Sperry J, Nathens AB, Billiar TR, West MA, Jeschke
835 MG, Klein MB, Gamelli RL, Gibran NS, Brownstein BH, Miller-Graziano C, Calvano SE, Mason
836 PH, Cobb JP, Rahme LG, Lowry SF, Maier RV, Moldawer LL, Herndon DN, Davis RW, Xiao W,
837 Tompkins RG, Inflammation, Host Response to Injury LSCRP. Genomic responses in mouse
838 models poorly mimic human inflammatory diseases. Proc Natl Acad Sci U S A. 2013;110(9):3507-
839 12. Epub 2013/02/13. doi: 10.1073/pnas.1222878110. PubMed PMID: 23401516; PMCID:
840 PMC3587220.
841 13. Ingber DE. Human organs-on-chips for disease modelling, drug development and
842 personalized medicine. Nat Rev Genet. 2022;23(8):467-91. Epub 2022/03/27. doi:
843 10.1038/s41576-022-00466-9. PubMed PMID: 35338360; PMCID: PMC8951665 member of its
844 board of directors.
845 14. Rodriguez TP, Mast JD, Hartl T, Lee T, Sand P, Perlstein EO. Defects in the
846 Neuroendocrine Axis Contribute to Global Development Delay in a Drosophila Model of NGLY1
847 Deficiency. G3 (Bethesda). 2018. doi: 10.1534/g3.118.300578. PubMed PMID: 29735526.
848 15. Sirr A, Scott AC, Cromie GA, Ludlow CL, Ahyong V, Morgan TS, Gilbert T, Dudley AM.
849 Natural Variation in SER1 and ENA6 Underlie Condition-Specific Growth Defects in
850 Saccharomyces cerevisiae. G3 (Bethesda). 2018;8(1):239-51. Epub 2017/11/16. doi:
851 10.1534/g3.117.300392. PubMed PMID: 29138237; PMCID: PMC5765352.
852 16. Li D, March ME, Gutierrez-Uzquiza A, Kao C, Seiler C, Pinto E, Matsuoka LS, Battig MR,
853 Bhoj EJ, Wenger TL, Tian L, Robinson N, Wang T, Liu Y, Weinstein BM, Swift M, Jung HM,
854 Kaminski CN, Chiavacci R, Perkins JA, Levine MA, Sleiman PMA, Hicks PJ, Strausbaugh JT,
855 Belasco JB, Dori Y, Hakonarson H. ARAF recurrent mutation causes central conducting lymphatic

856 anomaly treatable with a MEK inhibitor. Nat Med. 2019;25(7):1116-22. Epub 2019/07/03. doi:
857 10.1038/s41591-019-0479-2. PubMed PMID: 31263281.
858 17. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture:
859 the shape of the genetic contribution to human traits and disease. Nat Rev Genet.
860 2018;19(2):110-24. doi: 10.1038/nrg.2017.101. PubMed PMID: 29225335.
861 18. Knafo S, Esteban JA. PTEN: Local and Global Modulation of Neuronal Function in Health
862 and Disease. Trends Neurosci. 2017;40(2):83-91. doi: 10.1016/j.tins.2016.11.008. PubMed
863 PMID: 28081942.
864 19. Courchesne E, Mouton PR, Calhoun ME, Semendeferi K, Ahrens-Barbeau C, Hallet MJ,
865 Barnes CC, Pierce K. Neuron number and size in prefrontal cortex of children with autism. JAMA.
866 2011;306(18):2001-10. doi: 10.1001/jama.2011.1638. PubMed PMID: 22068992.
867 20. Endo A. The discovery and development of HMG-CoA reductase inhibitors. 1992.
868 Atheroscler Suppl. 2004;5(3):67-80. doi: 10.1016/j.atherosclerosissup.2004.08.026. PubMed
869 PMID: 15531278.
870 21. Endo A. A historical perspective on the discovery of statins. Proc Jpn Acad Ser B Phys
871 Biol Sci. 2010;86(5):484-93. Epub 2010/05/15. doi: 10.2183/pjab.86.484. PubMed PMID:
872 20467214; PMCID: PMC3108295.
873 22. Stossel TP. The discovery of statins. Cell. 2008;134(6):903-5. Epub 2008/09/23. doi:
874 10.1016/j.cell.2008.09.008. PubMed PMID: 18805080.
875 23. Plantone D, Koudriavtseva T. Current and Future Use of Chloroquine and
876 Hydroxychloroquine in Infectious, Immune, Neoplastic, and Neurological Diseases: A Mini-
877 Review. Clin Drug Investig. 2018. doi: 10.1007/s40261-018-0656-y. PubMed PMID: 29737455.
878 24. Misztak P, Panczyszyn-Trzewik P, Sowa-Kucma M. Histone deacetylases (HDACs) as
879 therapeutic target for depressive disorders. Pharmacol Rep. 2018;70(2):398-408. doi:
880 10.1016/j.pharep.2017.08.001. PubMed PMID: 29456074.

881 25. Rudmann DG. On-target and off-target-based toxicologic effects. Toxicol Pathol.
882 2013;41(2):310-4. doi: 10.1177/0192623312464311. PubMed PMID: 23085982.
883 26. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J,
884 Huss JW, 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing
885 gene annotation resources. Genome Biol. 2009;10(11):R130. doi: 10.1186/gb-2009-10-11-r130.
886 PubMed PMID: 19919682; PMCID: PMC3091323.
887 27. Grover M, Camilleri M. Effects on gastrointestinal functions and symptoms of serotonergic
888 psychoactive agents used in functional gastrointestinal diseases. J Gastroenterol.
889 2013;48(2):177-81. doi: 10.1007/s00535-012-0726-5. PubMed PMID: 23254779; PMCID:
890 PMC3698430.
891 28. Berard A, Sheehy O, Zhao JP, Vinet E, Bernatsky S, Abrahamowicz M. SSRI and SNRI
892 use during pregnancy and the risk of persistent pulmonary hypertension of the newborn. Br J Clin
893 Pharmacol. 2017;83(5):1126-33. doi: 10.1111/bcp.13194. PubMed PMID: 27874994; PMCID:
894 PMC5401975.
895 29. O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. Nat Rev Genet.
896 2017;18(10):613-23. doi: 10.1038/nrg.2017.47. PubMed PMID: 28649135.
897 30. Nijman SM. Synthetic lethality: general principles, utility and detection using genetic
898 screens in human cells. FEBS Lett. 2011;585(1):1-6. doi: 10.1016/j.febslet.2010.11.024. PubMed
899 PMID: 21094158; PMCID: PMC3018572.
900 31. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70. PubMed
901 PMID: 10647931.
902 32. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell.
903 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.
904 33. Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED,
905 Kampmann M, Peterson TR, Nakamura K, Fischbach MA, Weissman JS, Gilbert LA. Mapping the

906 Genetic Landscape of Human Cells. Cell. 2018;174(4):953-67 e22. Epub 2018/07/24. doi:
907 10.1016/j.cell.2018.06.010. PubMed PMID: 30033366; PMCID: PMC6426455.
908 34. Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP,
909 Thornock A, Zarraga G, Tam S, Szpyt J, Gassaway BM, Panov A, Parzen H, Fu S, Golbazi A,
910 Maenpaa E, Stricker K, Guha Thakurta S, Zhang T, Rad R, Pan J, Nusinow DP, Paulo JA,
911 Schweppe DK, Vaites LP, Harper JW, Gygi SP. Dual proteome-scale networks reveal cell-specific
912 remodeling of the human interactome. Cell. 2021;184(11):3022-40 e28. Epub 2021/05/08. doi:
913 10.1016/j.cell.2021.04.011. PubMed PMID: 33961781; PMCID: PMC8165030.
914 35. https://catalyst.phrma.org/new-report-demonstrates-industry-leadership-in-drug-
915 development-in-partnership-with-nih.
916 36. Hanahan D. Rethinking the war on cancer. Lancet. 2014;383(9916):558-63. Epub
917 2013/12/20. doi: 10.1016/s0140-6736(13)62226-6. PubMed PMID: 24351321.
918 37. Ozercan HI, Ileri AM, Ayday E, Alkan C. Realizing the potential of blockchain technologies
919 in genomics. Genome Res. 2018;28(9):1255-63. Epub 2018/08/05. doi: 10.1101/gr.207464.116.
920 PubMed PMID: 30076130; PMCID: PMC6120626.
921 38. Leeming G, Ainsworth J, Clifton DA. Blockchain in health care: hype, trust, and digital
922 health. Lancet. 2019;393(10190):2476-7. Epub 2019/06/25. doi: 10.1016/s0140-6736(19)30948-
923 1. PubMed PMID: 31232356.
924 39. Gürsoy G, Bjornson R, Green ME, Gerstein M. Using blockchain to log genome dataset
925 access: efficient storage and query. BMC Med Genomics. 2020;13(Suppl 7):78. Epub 2020/07/23.
926 doi: 10.1186/s12920-020-0716-z. PubMed PMID: 32693796; PMCID: PMC7372787.
927 40. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S,
928 Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J,
929 Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn
930 WC. Defining a Cancer Dependency Map. Cell. 2017;170(3):564-76 e16. doi:
931 10.1016/j.cell.2017.06.010. PubMed PMID: 28753430; PMCID: PMC5667678.

932 41. Pan J, Meyers RM, Michel BC, Mashtalir N, Sizemore AE, Wells JN, Cassel SH, Vazquez
933 F, Weir BA, Hahn WC, Marsh JA, Tsherniak A, Kadoch C. Interrogation of Mammalian Protein
934 Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. Cell Syst.
935 2018;6(5):555-68 e7. doi: 10.1016/j.cels.2018.04.011. PubMed PMID: 29778836; PMCID:
936 PMC6152908.
937 42. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, Humeidi R, Peck
938 D, Wu X, Tang AA, Wang VM, Bender SA, Lemire E, Narayan R, Montgomery P, Ben-David U,
939 Garvie CW, Chen Y, Rees MG, Lyons NJ, McFarland JM, Wong BT, Wang L, Dumont N, O'Hearn
940 PJ, Stefan E, Doench JG, Harrington CN, Greulich H, Meyerson M, Vazquez F, Subramanian A,
941 Roth JA, Bittker JA, Boehm JS, Mader CC, Tsherniak A, Golub TR. Discovering the anti-cancer
942 potential of non-oncology drugs by systematic viability profiling. Nat Cancer. 2020;1(2):235-48.
943 Epub 2020/07/03. doi: 10.1038/s43018-019-0018-6. PubMed PMID: 32613204; PMCID:
944 PMC7328899.
945 43. Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED,
946 Kampmann M, Peterson TR, Nakamura K, Fischbach MA, Weissman JS, Gilbert LA. Mapping the
947 Genetic Landscape of Human Cells. Cell. 2018. doi: 10.1016/j.cell.2018.06.010. PubMed PMID:
948 30033366.
949 44. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the
950 reasons why potentially important genes are ignored. PLoS Biol. 2018;16(9):e2006643. doi:
951 10.1371/journal.pbio.2006643. PubMed PMID: 30226837; PMCID: PMC6143198.
952 45. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M,
953 Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS,
954 Durocher D, Angers S, Moffat J. High-Resolution CRISPR Screens Reveal Fitness Genes and
955 Genotype-Specific Cancer Liabilities. Cell. 2015;163(6):1515-26. doi: 10.1016/j.cell.2015.11.015.
956 PubMed PMID: 26627737.

957 46. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van
958 Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J,
959 Superti-Furga G, Brummelkamp TR. Gene essentiality and synthetic lethality in haploid human
960 cells. Science. 2015;350(6264):1092-6. doi: 10.1126/science.aac7557. PubMed PMID:
961 26472760.
962 47. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM.
963 Identification and characterization of essential genes in the human genome. Science.
964 2015;350(6264):1096-101. doi: 10.1126/science.aac7041. PubMed PMID: 26472758; PMCID:
965 PMC4662922.
966 48. Apaolaza I, José-Eneriz ES, Tobalina L, Miranda E, Garate L, Agirre X, Prósper F, Planes
967 FJ. An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. Nature
968 Communications. 2017;8(1):459. doi: 10.1038/s41467-017-00555-y.
969 49. Klamt S. Generalized concept of minimal cut sets in biochemical networks. Biosystems.
970 2006;83(2-3):233-47. doi: 10.1016/j.biosystems.2005.04.009. PubMed PMID: 16303240.
971 50. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV,
972 Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington
973 WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M,
974 Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. Computational
975 correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in
976 cancer cells. Nat Genet. 2017;49(12):1779-84. Epub 2017/10/31. doi: 10.1038/ng.3984. PubMed
977 PMID: 29083409; PMCID: PMC5709193.
978 51. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to
979 Omnigenic. Cell. 2017;169(7):1177-86. doi: 10.1016/j.cell.2017.05.038. PubMed PMID:
980 28622505.
981 52. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death.
982 53. http://www.nealelab.is/uk-biobank.

983 54. Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T,
984 Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M. Drug repurposing:
985 progress, challenges and recommendations. Nat Rev Drug Discov. 2018. doi:
986 10.1038/nrd.2018.168. PubMed PMID: 30310233.
987 55. Kirkland JL, Tchkonia T, Zhu Y, Niedernhofer LJ, Robbins PD. The Clinical Potential of
988 Senolytic Drugs. J Am Geriatr Soc. 2017;65(10):2297-301. doi: 10.1111/jgs.14969. PubMed
989 PMID: 28869295; PMCID: PMC5641223.
990 56. Surface LE, Burrow DT, Li J, Park J, Kumar S, Lyu C, Song N, Yu Z, Rajagopal A, Bae Y,
991 Lee BH, Mumm S, Gu CC, Baker JC, Mohseni M, Sum M, Huskey M, Duan S, Bijanki VN, Civitelli
992 R, Gardner MJ, McAndrew CM, Ricci WM, Gurnett CA, Diemer K, Wan F, Costantino CL,
993 Shannon KM, Raje N, Dodson TB, Haber DA, Carette JE, Varadarajan M, Brummelkamp TR,
994 Birsoy K, Sabatini DM, Haller G, Peterson TR. ATRAID regulates the action of nitrogen-containing
995 bisphosphonates on bone. Sci Transl Med. 2020;12(544). doi: 10.1126/scitranslmed.aav9166.
996 PubMed PMID: 32434850; PMCID: PMC7882121.
997 57. Yu Z, Surface LE, Park CY, Horlbeck MA, Wyant GA, Abu-Remaileh M, Peterson TR,
998 Sabatini DM, Weissman JS, O'Shea EK. Identification of a transporter complex responsible for
999 the cytosolic entry of nitrogen-containing-bisphosphonates. Elife. 2018;7. doi:
1000 10.7554/eLife.36620. PubMed PMID: 29745899.
1001 58. Pinal-Fernandez I, Casal-Dominguez M, Mammen AL. Statins: pros and cons. Med Clin
1002 (Barc). 2018;150(10):398-402. Epub 2018/01/03. doi: 10.1016/j.medcli.2017.11.030. PubMed
1003 PMID: 29292104; PMCID: PMC6019636.
1004 59. Meehan TF, Conte N, West DB, Jacobsen JO, Mason J, Warren J, Chen CK, Tudose I,
1005 Relac M, Matthews P, Karp N, Santos L, Fiegel T, Ring N, Westerberg H, Greenaway S, Sneddon
1006 D, Morgan H, Codner GF, Stewart ME, Brown J, Horner N, Haendel M, Washington N, Mungall
1007 CJ, Reynolds CL, Gallegos J, Gailus-Durner V, Sorg T, Pavlovic G, Bower LR, Moore M, Morse
1008 I, Gao X, Tocchini-Valentini GP, Obata Y, Cho SY, Seong JK, Seavitt J, Beaudet AL, Dickinson

1009 ME, Herault Y, Wurst W, de Angelis MH, Lloyd KCK, Flenniken AM, Nutter LMJ, Newbigging S,
1010 McKerlie C, Justice MJ, Murray SA, Svenson KL, Braun RE, White JK, Bradley A, Flicek P, Wells
1011 S, Skarnes WC, Adams DJ, Parkinson H, Mallon AM, Brown SDM, Smedley D. Disease model
1012 discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. Nat
1013 Genet. 2017;49(8):1231-8. Epub 2017/06/27. doi: 10.1038/ng.3901. PubMed PMID: 28650483;
1014 PMCID: PMC5546242.
1015 60. Wang Z, Li Y, Zhou F, Piao Z, Hao J. Effects of Statins on Bone Mineral Density and
1016 Fracture Risk: A PRISMA-compliant Systematic Review and Meta-Analysis. Medicine (Baltimore).
1017 2016;95(22):e3042. Epub 2016/06/04. doi: 10.1097/md.0000000000003042. PubMed PMID:
1018 27258488; PMCID: PMC4900696.
1019 61. Mundy G, Garrett R, Harris S, Chan J, Chen D, Rossini G, Boyce B, Zhao M, Gutierrez G.
1020 Stimulation of bone formation in vitro and in rodents by statins. Science. 1999;286(5446):1946-9.
1021 Epub 1999/12/03. doi: 10.1126/science.286.5446.1946. PubMed PMID: 10583956.
1022 62. L. E. Surface DTB, J. Li, J. Park, S. Kumar, C. Lyu, N. Song, Z. Yu, A. Rajagopal,Y. Bae,
1023 B. H. Lee, S. Mumm, C. C. Gu, J. C. Baker, M. Mohseni, M. Sum, M. Huskey, S. Duan,V. N.
1024 Bijanki, R. Civitelli, M. J. Gardner, C. M. McAndrew, W. M. Ricci, C. A. Gurnett, K. Diemer, F.
1025 Wan,C. L. Costantino, K. M. Shannon, N. Raje, T. B. Dodson, D. A. Haber, J. E. Carette, M.
1026 Varadarajan,T. R. Brummelkamp, K. Birsoy, D. M. Sabatini, G. Haller, T. R. Peterson. ATRAID
1027 regulates the action of nitrogen-containing bisphosphonates on bone. Sci Transl Med.
1028 2020;12(eaav9166). doi: 10.1126/scitranslmed.aav9166.
1029 63. Sabatini DM. Twenty-five years of mTOR: Uncovering the link from nutrients to growth.
1030 Proc Natl Acad Sci U S A. 2017;114(45):11818-25. doi: 10.1073/pnas.1716173114. PubMed
1031 PMID: 29078414; PMCID: PMC5692607.
1032 64. Peterson TR, Head, R. Personal communication on single-cell RNA seq on mouse liver.
1033 2017.

1034 65. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare
1035 MJ, Neville AM, Simpson AJ, Strausberg RL. Comprehensive sampling of gene expression in
1036 human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci U S A.
1037 2003;100(8):4702-5. Epub 2003/04/03. doi: 10.1073/pnas.0831040100. PubMed PMID:
1038 12671075; PMCID: PMC153619.
1039 66. Garcia-Ortega LF, Martinez O. How Many Genes Are Expressed in a Transcriptome?
1040 Estimation and Results for RNA-Seq. PLoS One. 2015;10(6):e0130262. Epub 2015/06/25. doi:
1041 10.1371/journal.pone.0130262. PubMed PMID: 26107654; PMCID: PMC4479379.
1042 67. Skinnider MA, Stacey RG, Foster LJ. Genomic data integration systematically biases
1043 interactome mapping. PLoS Comput Biol. 2018;14(10):e1006474. doi:
1044 10.1371/journal.pcbi.1006474. PubMed PMID: 30332399; PMCID: PMC6192561.
1045 68. Nelson MR, Johnson T, Warren L, Hughes AR, Chissoe SL, Xu CF, Waterworth DM. The
1046 genetics of drug efficacy: opportunities and challenges. Nat Rev Genet. 2016;17(4):197-206. doi:
1047 10.1038/nrg.2016.12. PubMed PMID: 26972588.
1048 69. Swinney DC. Phenotypic vs. target-based drug discovery for first-in-class medicines. Clin
1049 Pharmacol Ther. 2013;93(4):299-301. doi: 10.1038/clpt.2012.236. PubMed PMID: 23511784.
1050 70. Jost M, Weissman JS. CRISPR Approaches to Small Molecule Target Identification. ACS
1051 Chem Biol. 2018;13(2):366-75. doi: 10.1021/acschembio.7b00965. PubMed PMID: 29261286;
1052 PMCID: PMC5834945.
1053 71. Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F,
1054 Gygi MP, Parzen H, Szpyt J, Tam S, Zarraga G, Pontano-Vaites L, Swarup S, White AE,
1055 Schweppe DK, Rad R, Erickson BK, Obar RA, Guruharsha KG, Li K, Artavanis-Tsakonas S, Gygi
1056 SP, Harper JW. Architecture of the human interactome defines protein communities and disease
1057 networks. Nature. 2017;545(7655):505-9. doi: 10.1038/nature22366. PubMed PMID: 28514442;
1058 PMCID: PMC5531611.

1059 72. Hochberg Ba. Controlling the false positive rate: a practical and powerful approach to
1060 multiple testing. Journal of the Royal Statistical Society Series B. 1995;57:289-300.
1061 73. https://federalreporter.nih.gov/FileDownload.
1062 74. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score
1063 analyses. Nat Protoc. 2020;15(9):2759-72. Epub 2020/07/28. doi: 10.1038/s41596-020-0353-1.
1064 PubMed PMID: 32709988.
1065 75. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? Jama.
1066 2019;321(18):1820-1. Epub 2019/04/09. doi: 10.1001/jama.2019.3893. PubMed PMID:
1067 30958510.
1068 76. Gibson G. Hints of hidden heritability in GWAS. Nat Genet. 2010;42(7):558-60. Epub
1069 2010/06/29. doi: 10.1038/ng0710-558. PubMed PMID: 20581876.
1070 77. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
1071 analysis. Am J Hum Genet. 2011;88(1):76-82. Epub 2010/12/21. doi: 10.1016/j.ajhg.2010.11.011.
1072 PubMed PMID: 21167468; PMCID: PMC3014363.
1073 78. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath
1074 AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large
1075 proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9. Epub 2010/06/22.
1076 doi: 10.1038/ng.608. PubMed PMID: 20562875; PMCID: PMC3232052.
1077

Figure 1

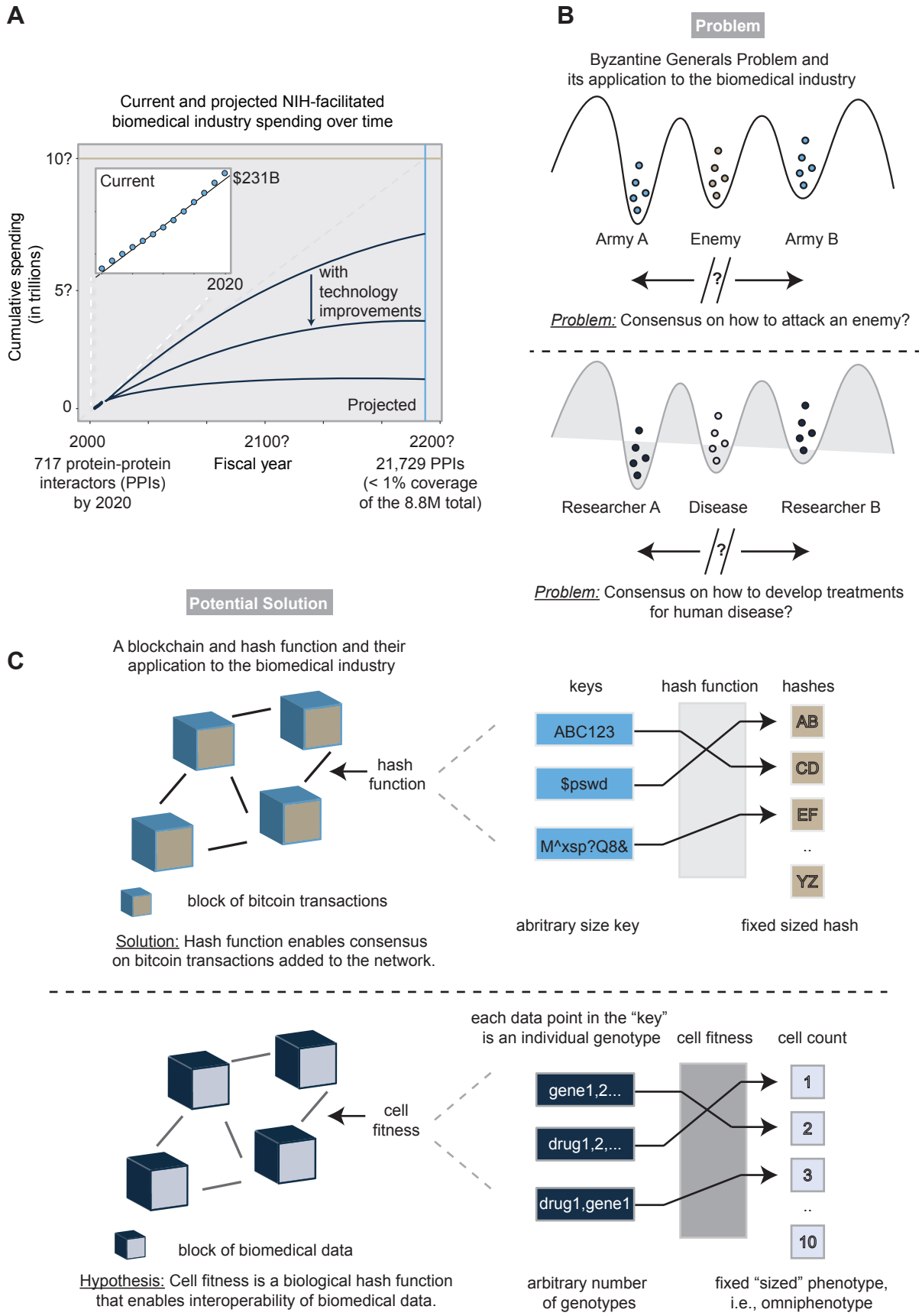

A
Current and projected NIH-facilitated biomedical industry spending over time
Cumulative spending (in trillions)
10?
5?
0
Current
$231B
2020
with technology improvements
Projected
2000
2100?
2200?
717 protein-protein interactors (PPIs) by 2020
Fiscal year
21,729 PPIs (< 1% coverage of the 8.8M total)
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

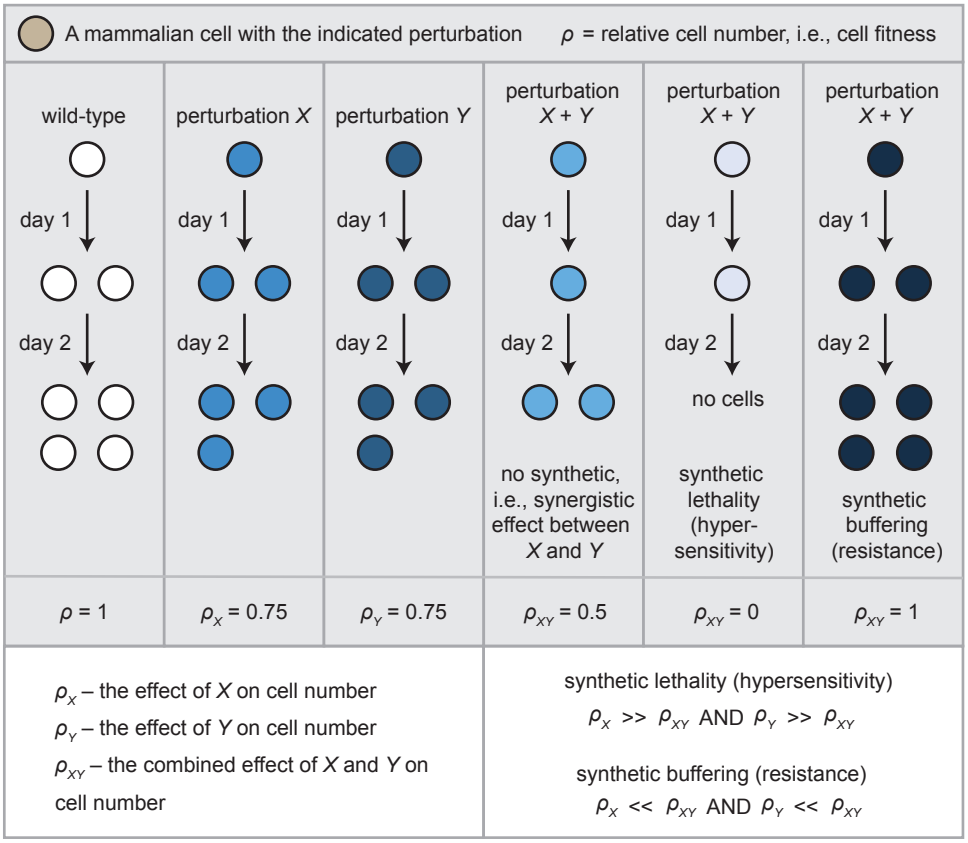

A mammalian cell with the indicated perturbation
ρ = relative cell number, i.e., cell fitness
wild-type
perturbation X
perturbation Y
perturbation X + Y
perturbation X + Y
perturbation X + Y
day 1
day 2
no cells
no synthetic, i.e., synergistic effect between X and Y
synthetic lethality (hyper-sensitivity)
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

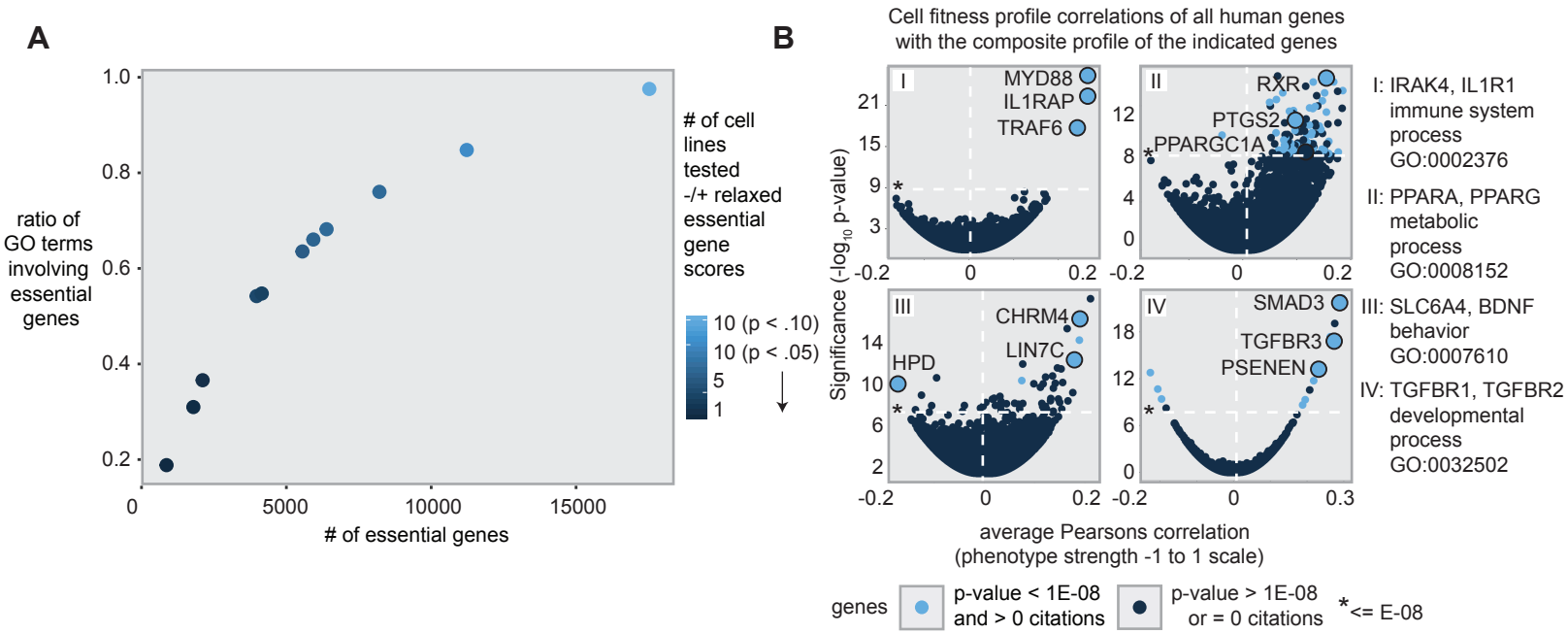

A
ratio of GO terms involving essential genes
1.0
0.8
0.6
0.4
0.2
0
5000
10000
15000
# of essential genes
# of cell lines tested -/+ relaxed essential gene scores
10 (p < .10)
10 (p < .05)
5
1
B
Cell fitness profile correlations of all human genes with the composite profile of the indicated genes
Significance (-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

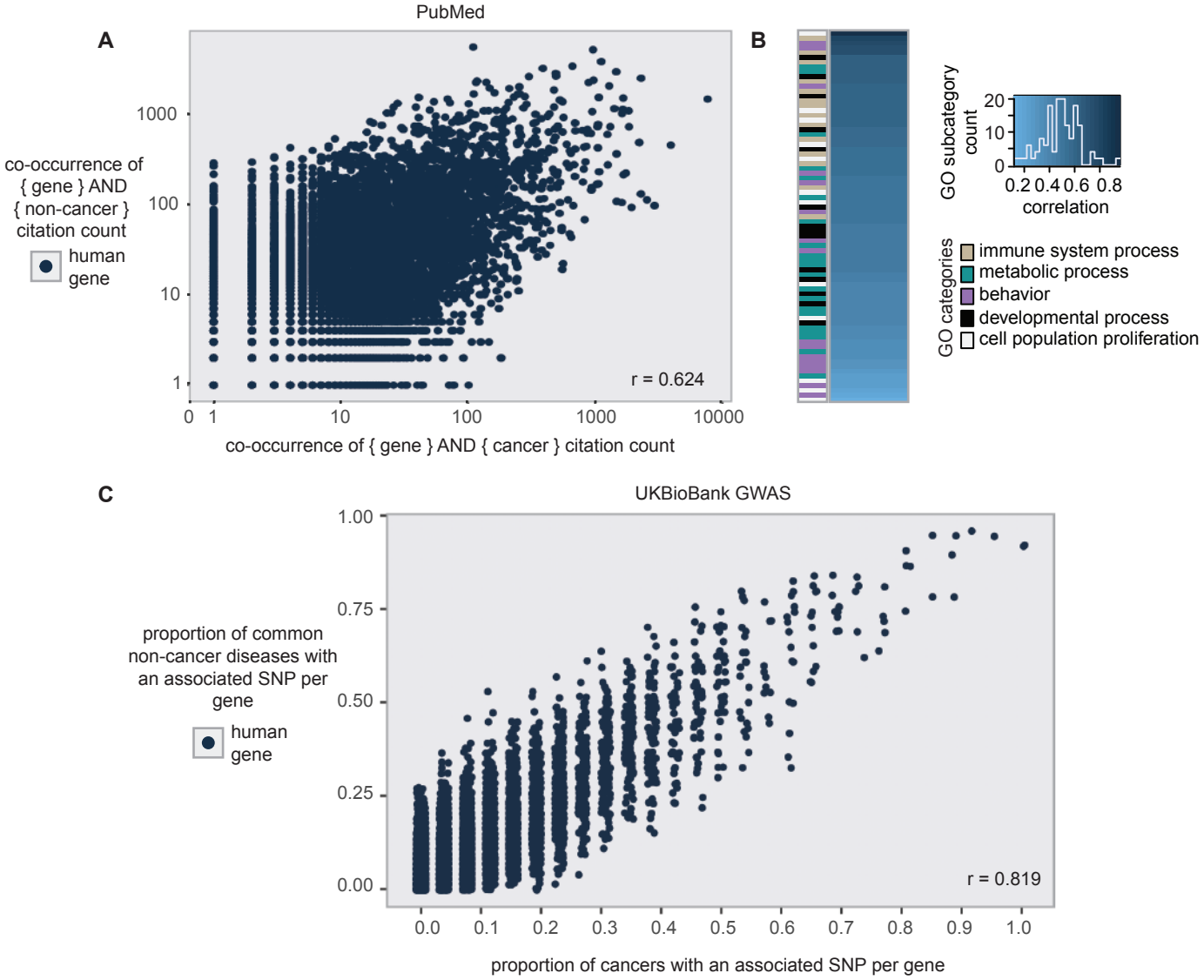
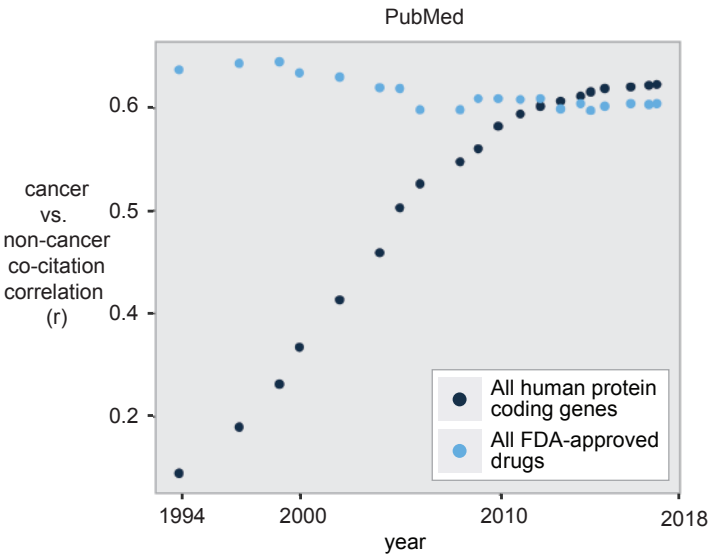

A
PubMed
co-occurrence of { gene } AND { non-cancer } citation count
human gene
1000
100
10
1
0
1
10
100
1000
10000
r = 0.624
co-occurrence of { gene } AND { cancer } citation count
B
GO subcategory count
20
10
0
0.2 0.4 0.6 0.8
correlation
GO categories
immune system process
metabolic process
behavior
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

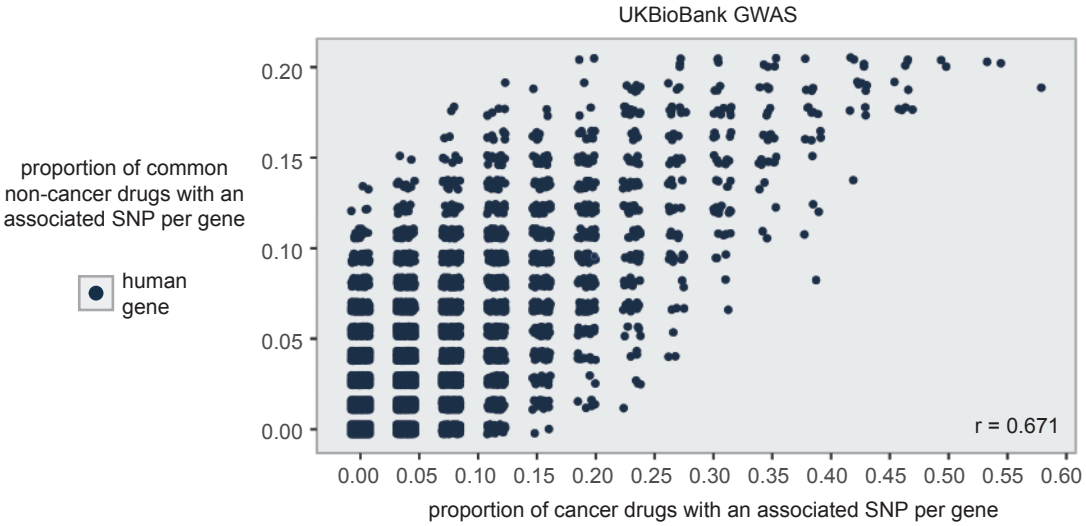


Figure 6

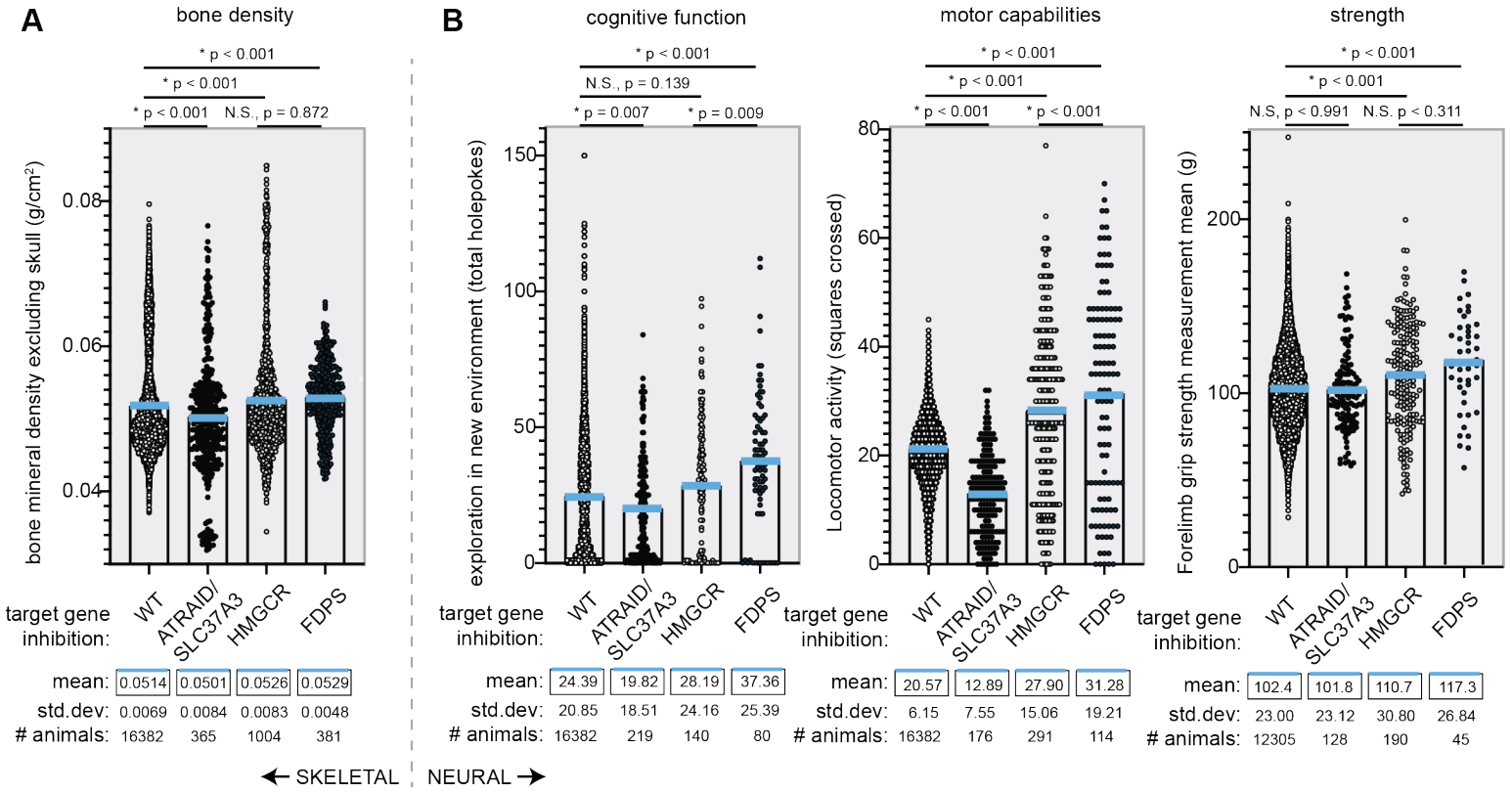

A
bone density
* p < 0.001
* p < 0.001
* p < 0.001
N.S., p = 0.872
bone mineral density excluding skull (g/cm²)
0.08
0.06
0.04
target gene inhibition:
WT
ATRAID/ SLC37A3
HMGCR
FDPS
mean: 0.0514 0.0501 0.0526 0.0529
std.dev: 0.0069 0.0084 0.0083 0.0048
animals: 16382 365 1004 381
B
cognitive function
* p < 0.001
N.S., p = 0.139
* p = 0.007
* p = 0.009
exploration in new environment (total holepokes)
150
100
50
0
target gene inhibition:
WT
ATRAID/ SLC37A3
HMGCR
FDPS
mean: 24.39 19.82 28.19 37.36
std.dev: 20.85 18.51 24.16 25.39
animals: 16382 219 140 80
motor capabilities
* p < 0.001
* p < 0.001
* p < 0.001
* p < 0.001
Locomotor activity (squares crossed)
80
60
40
20
0
target gene inhibition:
WT
ATRAID/ SLC37A3
HMGCR
FDPS
mean: 20.57 12.89 27.90 31.28
std.dev: 6.15 7.55 15.06 19.21
animals: 16382 176 291 114
strength
* p < 0.001
* p < 0.001
N.S, p < 0.991
N.S. p < 0.311
Forelimb grip strength measurement mean (g)
200
100
0
target gene inhibition:
WT
ATRAID/ SLC37A3
HMGCR
FDPS
mean: 102.4 101.8 110.7 117.3
std.dev: 23.00 23.12 30.80 26.84
animals: 12305 128 190 45
← SKELETAL
NEURAL →

Figure 7

Framework for synthetic interaction testing

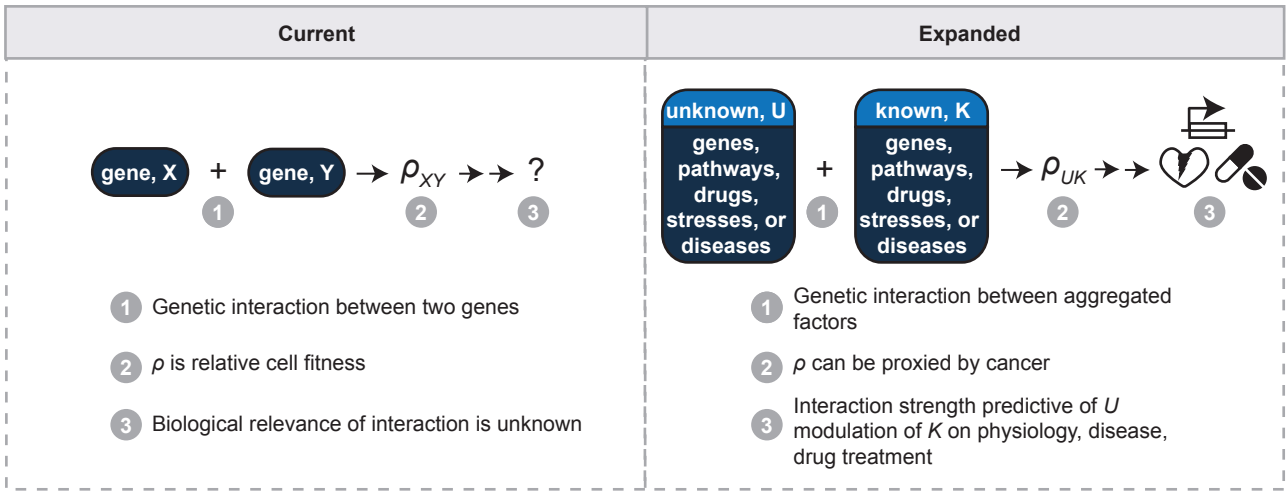
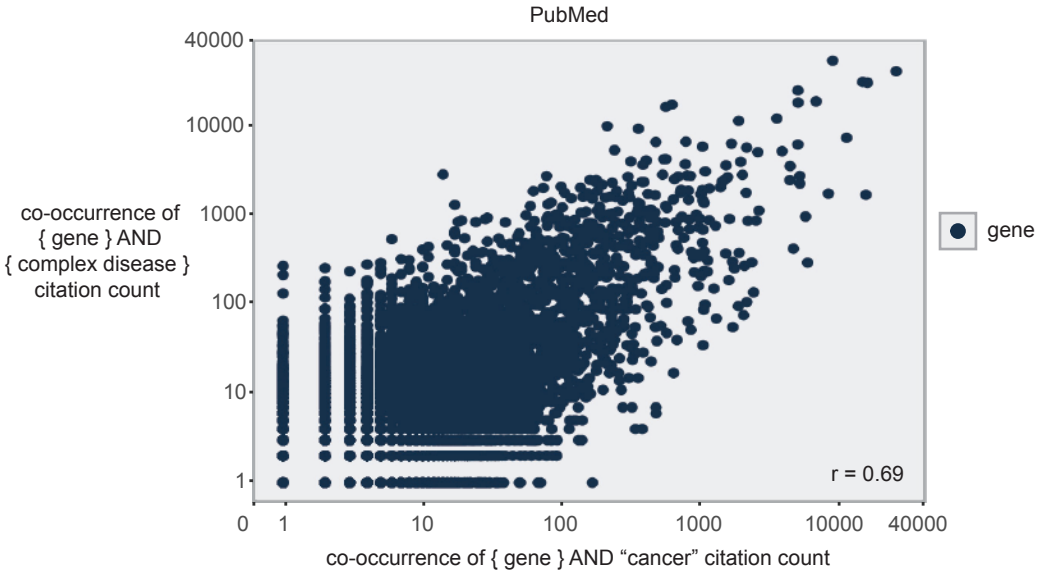


Figure S1


PubMed
40000
10000
1000
100
10
1
0
1
10
100
1000
10000
40000
co-occurrence of
{ gene } AND
{ complex disease }
citation count
r = 0.69
co-occurrence of { gene } AND “cancer” citation count
gene