Data-Driven Search and Innovation *

Matteo Tranchero
UC Berkeley-Haas

Preliminary Draft – Please Do Not Circulate

First Version: June 12th, 2021

This Version: June 1st, 2022

Abstract

How does the growing availability of data shape innovation? I propose that data analytics enables a new way to search for novel combinations among technological components. Instead of being tethered by past attempts or priors, inventors can use data to extract signals of what combinations are more promising for follow-on experimentation. I investigate this emerging phenomenon in the context of human genomics, where genome-wide association studies (GWASs) approximate the ideal of a data-driven search for the genetic roots of diseases. My results show that novel gene-disease associations introduced by GWASs span a wide portion of the genetic landscape, are likely to involve neglected human genes, and on average are of higher scientific value than comparable associations introduced by targeted studies. The latter result is stronger for genes difficult to experimentally study, but becomes negative for genes already well known. Taken together, these findings point to the potential and boundary conditions of data-driven search strategies in technological innovation.

*E-mail: `m.tranchero@berkeley.edu`.

1 Introduction

Innovation is generated by recombining ideas and technological components in novel ways (Fleming, 2001; Schilling and Green, 2011). Each successful innovation can be further recombined, hence constantly expanding the technological frontier (Weitzman, 1998). Yet, locating the combinations that produce useful innovation becomes increasingly harder in the exponentially-growing combinatorial space (Jones, 2009; Agrawal et al., 2019). Researchers have to adopt a search strategy to decide which combinations to prioritize. Common heuristics include incremental modifications of past successful attempts (Stuart and Podolny, 1996), or reliance on scientific theories to direct experimentation (Arora and Gambardella, 1994; Fleming and Sorenson, 2004). Both these strategies operate by targeting familiar or well-understood subsets of the landscape, with the unintended consequence of curtailing exploration in uncharted domains and eventually reaching technological exhaustion (Fleming, 2001). Previous work has noted how this trade-off between exploration and exploitation might be fundamentally unsolvable (March, 1991).

However, the rapid diffusion of "big data" and computational tools could provide an alternative solution to this conundrum (Agrawal et al., 2019; Cockburn et al., 2019). For instance, pharmaceutical firms use combinatorial chemistry to gather data on millions of candidate compounds for prioritization or R&D efforts (Jayaraj and Gittelman, 2018). Startups and venture capitalists use A/B testing and data analytics to triage entrepreneurial ideas at scale, iteratively testing many configurations without being constrained by pre-specified designs (Azevedo et al., 2020; Koning et al., 2022). Exploration firms sift through satellite photos to spot geographical areas worth investing in across the entire world (Nagaraj, 2021). In all these cases, decision makers are using large amounts of data to sort through vast combinatorial spaces and guide experimental attempts (Wu et al., 2020; Agrawal et al., 2022). Data analytics allows to extract signals that locate the most promising technological combinations, an approach that in this paper I call *data-driven search*.

But while data analytics is enabling a novel strategy for recombinant search, its effects on the direction and impact of innovation are an open question. On the one hand, data-driven search could address some of the shortcomings of targeted search. Access to comprehensive data on the landscape might broaden the search scope and diversify knowledge production (Nagaraj et al., 2020). Instead of being bounded to a small number of familiar components, analytical techniques should facilitate recombinations from distant technological domains (Wu et al., 2020). On the other hand, empirical evidence on the use of data analytics is ambiguous and shows large variance in outcomes (Brynjolfsson et al., 2021). The possibility of cheaply generating data was found to encourage incremental innovation (Deniz, 2020; Ghosh, 2021). Moreover, reliance on data in the absence of theoretical understanding makes it impossible to tease apart good leads from false positives (Felin et al., 2021; Lou and Wu, 2021). Overall, the conditions under which data-driven search strategies will improve innovation outcomes are unclear and have not yet been fully explored.

Empirical investigation of these ideas is challenging for three reasons. First, one needs to find a setting in which the combinatorial space is well-defined and measurable. Next, there has to be some variation in the availability of data (or possibility of easily generating them) that enables data-driven search strategies. Finally, to empirically study data-driven search, one needs to characterize distinct search processes and be able to tie them with the resulting innovation outputs. This is especially difficult because the researcher can only see realized outcomes, without usually being able to observe what kind of search strategy generated them (Kneeland et al., 2020). Understanding the consequences of data-driven search requires finding a setting where all these conditions are met at the same time.

In this paper, I address these challenges with a quantitative case study based on research that aims to find the genetic roots of human diseases. First, genes and diseases constitute the relevant components for this innovation problem. Any gene could in principle be tied to any condition, hence generating a space of tens of millions of pairwise gene-disease combinations. The key objective of researchers and firms is to find which combinations are the most promising for follow-on clinical investigation and drug development. Second, sharp reduction in genotyping costs in the early 2000s enabled a new approach to searching for new gene-disease associations, called *genome-wide association studies* (GWASs). Traditionally, scientists interested in a disease would pre-select one or a handful of genes for targeted analyses. Instead, GWASs exploit data on the entire genome to provide signals on which genes are tied to the disease studied, without being limited to a subset of the genomic space. Finally, by linking gene-disease combinations to the methods used in the scientific articles that introduced them, I can infer what search process was used. Comparing combinations introduced by GWASs to those discovered by candidate-gene studies allows me to descriptively explore the characteristics and impact of data-driven search.

I assemble a new dataset that includes all gene-disease associations introduced in the period 1980-2016. The data raw data are taken from DisGeNET, the most comprehensive aggregator of information of human gene-disease associations (GDAs). For each GDA, DisGeNET collects the list of journal articles in PubMed that studied it. I consider the earliest of these papers as the one that discovered the association, and the amount of follow-on work as a measure of its scientific impact. Next, I classify each association by whether it was introduced by a GWAS or by traditional candidate-gene approaches using data from the European Bioinformatic Institute. I focus on the impact of data-driven search on two related dimensions: the direction of search efforts and the scientific potential of the gene-disease combinations uncovered.

The results of this empirical analysis suggest that data-driven search diversify innovation by exploring a wider portion of the genetic landscape. In baseline estimates, gene-disease combinations discovered by GWAS are 154% and and 108% more likely to involve genes understudied or recently discovered, respectively. This is due to the fact that genome-wide association studies remove the ex ante choice of what genes to target, hence overcoming risk aversion and path dependency in search. Further, this change of focus is consequential and leads to introducing GDAs of higher scientific impact on average. Additional tests show that this effect grows in the context of genes complex to experiment with, but it reverses and becomes negative for genes that are already well characterized theoretically. A battery of robustness checks further validate these descriptive findings, confirming that data-driven search helps to locate the best opportunities in rugged technological areas but it can be suboptimal in known domains.

This work presents a case study that contributes to our understanding of the role of data in recombinant search. Most directly, it shows empirically how data analytics are changing pharmaceutical research (Williams, 2013; Hermosilla and Lemus, 2019; Kao, 2022). However, the phenomenon analyzed in this paper is general in nature and it is diffusing rapidly (Zolas et al., 2021). Studying genome-wide association studies allows me to explore the precise dynamics through which data-driven search affects the generation of innovation, thus contributing to the emerging literature on the economic and business impacts of data (Jones and Tonetti, 2020; Farboodi and Veldkamp, 2020; Bessen et al., 2021; Nagaraj, 2021). Moreover, the construct of data-driven search constitutes an important addition to the theory of innovation search (March, 1991; Gavetti and Levinthal, 2000). To date, there is little theoretical understanding of how data analytics is reshaping individual and organizational search. Addressing this gap is of first-order importance in the age of big data, and my work is a first attempt in this direction.

2 Theoretical Framework

2.1 Strategies for Recombinant Search

Searching for innovation requires costly experimentation with novel technological combinations to generate knowledge about their value. Researchers and inventors do not usually make random attempts, rather focusing on the portions of the combinatorial space that promise them the highest returns. The choice of which combinations to explore is thus guided by pre-existing knowledge and beliefs about the technological landscape searched (Klahr and Dunbar, 1988; Gavetti and

Levinthal, 2000; Kneeland et al., 2020).

A few different strategies can be used to guide recombinant search. Inventors can start from known combinations of components, and incrementally change them in small steps (Fleming and Sorenson, 2001). Searching locally has the advantage of building upon existing capabilities, which makes search more efficient and reduces variability in outcomes (Fleming, 2001; Kaplan and Vakili, 2015; Gittelman, 2016; Arts and Fleming, 2018). Alternatively, inventors can also draw upon scientific information to select the most promising technological combinations. Scientific theories can change their mental representation of the technological landscape by both discouraging them to try certain combinations and directly indicating which ones are predicted to yield the best results (Nelson, 1982; Fleming and Sorenson, 2004). This form of theory-driven search exploits known cause-effect relationships to funnel experimentation efforts in areas where inventors hold a theoretical understanding of how components could be recombined (Arora and Gambardella, 1994; Felin and Zenger, 2017).

Both local search and theory-driven search operate by focusing attention on a restricted area of the technological landscape. Local search happens in the neighborhood of past successful attempts, hence neglecting a wide space of possibilities that require more radical departures from the status quo. Theory-driven search rules out ex ante all the combinations theoretically deemed as inferior, leading to never experimenting with them; however, any theory is bound to be imperfect and contextual, which means that novel discoveries might require going against established beliefs (Kuhn, 1962). The result is that with both strategies inventors have an incentive to focus on familiar components, either because they are close to those previously used or because their characteristics are theoretically well-understood (Arts and Fleming, 2018). The result is an artificial restriction of the potential combinations considered by the innovator. If only combinations that are familiar and easier to experiment with are explored, this could prevent the discovery of the most valuable innovations (Rzhetsky et al., 2015).

2.2 Data-Driven Search

Recent years have seen the emergence and diffusion of data analytics (Zolas et al., 2021). Researchers have started investigating its impact on corporate decision-making (Brynjolfsson and McElheran, 2016), organizational structure (Wu et al., 2019), and human capital (Rock, 2019). However, when focusing on innovation, the results are ambiguous and heterogeneous (Brynjolfsson et al., 2021). Data technologies appear to support mostly incremental process improvements (Wu et al., 2020), or even worse, lead to inferior outcomes in unknown domains (Lou and Wu, 2021).

Both theoretical and empirical work shows that reliance on data might harm the generation of innovation (Cao et al., 2021; Hoelzemann et al., 2022). These results seem at odds with the hype surrounding big data analytics and underscore the need for a better understanding of how data are used in the generation of novelty.

In this paper, I propose that data analytics affects innovation by enabling a new strategy to search for promising technological combinations (Dougherty and Dunne, 2012). Instead of relying on knowledge from past attempts or theoretical priors, inventors can use data to extract signals of what combinations are more promising for follow-on experimentation. Data enables a global scan of the combinatorial landscape that can lead directly to the best technological combinations, a search strategy that I call *data-driven search*. Compared to targeted search, the innovator no longer needs to pre-select a subset of components to recombine, a choice that is usually plagued by path dependency, inertia, risk aversion, or even cognitive limitations. Instead, the key decision becomes where to direct the data-collection effort: while bias could still exist in this choice, decreases in data costs and increasing availability of technological maps suggest that data-driven search might be of broader scope than other search strategies (Jayaraj and Gittelman, 2018; Nagaraj et al., 2020).

Data-driven search requires a few conditions to be feasible. First, the relevant characteristics of the components to recombine must be measurable, ensuring that the space of possible combinations is well-defined. This means that data-driven search will be of little help when trying to invent entirely new technologies that do not emerge from old components (Wu et al., 2020). Second, there has to exist an agreed-upon metric of technological potential on which the promise of each combination can be assessed. This constitutes the objective function that data-driven search tries to maximize by finding the candidate combinations that score highest on it. Third, and relatedly, it must be possible to foresee the effect of novel combinations on the objective of interest. Taken together, data-driven search requires that it is possible to predict the value of potential recombinations from the data available on components. Appendix A presents an example from combinatorial chemistry that illustrates how these boundary conditions define the feasibility data-driven search.

But how could data-driven search affect innovation outcomes in setting where it is viable? Data-driven search allows to consider a wider range of combinatorial possibilities compared to targeted approaches. This should remove the proclivity of inventors toward exploitation of known components, possibly leading to diversify knowledge production. However, the effects on the value of the resulting combinations are more ambiguous and hinge on two factors.[1] First, the expected value

[1] I am grateful to Daniel P. Gross for suggesting me this way of conceptualizing data-driven search. Appendix A provides additional discussions.

from an expansion of the components searched will depend on how thick the right tail of outcomes is vis-á-vis the left tail of inferior alternatives (Azevedo et al., 2020). If most of the additional combinations considered are low value, data-driven search might lead to worse outcomes. However, research has shown that extreme outcomes are more likely in complex technological landscapes where components are difficult to recombine (Fleming and Sorenson, 2001), suggesting that data-driven search might yield better results in those cases. Second, it will also depend on how good targeted search is. For well-known components, data-driven search will likely fare poorly, since inventors are already able to locate and focus on the best combinations (Kaplan and Vakili, 2015).

3 Empirical Setting

3.1 Scientific Background

Genes are sequences of DNA bases that encode the "instructions" to synthesize gene products with a fundamental role in the functioning of the organism. Knowing the genetic roots of diseases has important practical consequences since many genes involved with diseases have been proven to be effective drug targets (Nelson et al., 2015). For decades, researchers have focused on genes that are individually responsible for diseases. However, this class of diseases (called *Mendelian*) is much rarer than so-called *complex diseases*, such as diabetes, Alzheimer's, or cancer. Complex diseases are not due to a single genetic factor, but rather by many genes and their interaction with the environment during the course of human life (Bush and Moore, 2012). Discovering all the genes involved in each of the thousands of polygenic diseases requires searching through the $\sim 19,000$ known human genes. How do scientists look for for new gene-disease associations in this huge combinatorial space?

Scientists traditionally followed a *candidate-gene approach* consisting in three main steps (Tabor et al., 2002). First, the scientist would decide the disease to study, likely motivated by its prevalence or funding availability. Second, she would hypothesize what genes might have a role in its aetiology. Finally, she would focus the analysis on those gene, typically by means of family linkage studies, case-control studies, or gene knockout in lab animals. Importantly, the selection of the target genes reflects the search strategy followed by the researcher. One approach consists in looking in the neighborhood of previously established gene-disease associations. For instance, once BRCA2 was tied to female breast cancer, it was reasonable to expect that it could also be tied to other neoplasms. This reasoning led Gudmundsson et al. (1995) to find that BRCA2 was associated to ovarian and prostate cancer too. Alternatively, a researcher could use a more deductive approach and rely on existing biological theory. This is how Cargill et al. (2007) linked the IL23R gene to psoriasis:

knowledge of the role of IL12B in the metabolic pathway of IL23R together with the fact that IL12B had been associated to psoriasis led them to correctly postulate the IL23R-psoriasis nexus.

Despite many successful examples like the ones discussed above, both these strategies led scientists to consider only a limited number of genes. The result has been an extreme concentration of attention towards previously established research patterns (Oprea et al., 2018; Stoeger et al., 2018). Gates et al. (2021) report that 22% of gene-related publications referenced just 1% of genes. This situation is suboptimal since our understanding of polygenic diseases would strongly benefit from exploring a larger pool of genes. The excessive research emphasis on a handful of "superstar" genes means that many potentially important genes are simply ignored (Edwards et al., 2011; Stoeger et al., 2018). As a consequence, only around 10% of the potential drug targets highlighted by the Human Genome Project have been targeted by approved drugs, leaving many therapeutic opportunities still untried (Gates et al., 2021)

3.2 Genome-Wide Association Studies as Data-Driven Search

Starting from the early 2000s, two events concurred in providing an alternative to candidate-gene studies. The first was the completion of the first phase of the International HapMap Project (2005). The HapMap was designed to provide a detailed reference genome that could be used as the basis to relate genetic mutations with phenotype changes (Bush and Moore, 2012). The second, and related, was the diffusion of commercial genotyping microarrays. Unlike full genome sequencing, DNA microarrays only detect the activity of specific genetic loci. The HapMap enabled to design optimal microarrays markers that can be extrapolated to capture the characteristics of their genetic surroundings, thus allowing to parsimoniously infer the characteristics of the whole genome (Bush and Moore, 2012). The result of was a steep decrease in the cost of collecting data on genomes that prompted the emergence of *genome-wide association studies* (Visscher et al., 2017).

Genome-wide association studies (or GWASs) are case-control studies where researchers sequence a large number of genomes and look to see if any genetic variation is more likely to appear in the group showing a specific trait rather than in the control group (Pearson and Manolio, 2008; Uffelmann et al., 2021). Researchers start by collecting DNA samples both from both cases and controls. All DNA samples are genotyped using DNA microarrays and imputed through reference genomes to reconstruct full genotypes. Finally, researchers test for statistically significant differences between the genotypes of cases and controls. The genes in which there are variants strongly associated with the presence of a disease can be suspected to play a role in its aetiology, hence being potential targets for pharmaceutical intervention. Appendix B presents additional details and an example of genome-wide association study.

Unlike candidate-gene studies, where researchers decide which subset of genes to target, GWAS is a type of observational study that looks for genetic variants across the whole genome (Visscher et al., 2017; Uffelmann et al., 2021). A genome-wide search approach permits to scan the entire set of possible combinations, pointing directly to the most promising ones (Figure 1). In practice, this search strategy removes one degree of freedom from the researcher, who is no longer required to specify a genetic target ex-ante. This ensures that GWASs are unbiased with respect to prior biological knowledge and beliefs, thus avoiding the tendency to focus on familiar genes. Genome-wide association studies generate discoveries thanks to what directly emerges from the data, which makes them a prime example of data-driven search (Evans and Rzhetsky, 2010).

Yet, genome-wide association studies have been harshly criticized for a number of shortcomings. On the one hand, these studies are inherently correlational in nature, which means that there is the risk that any GWAS finding could be a false positive (Marigorta et al., 2018). On the other hand, even if the associations discovered by GWASs are statistically significant and replicable, scholars have suggested that GWAS neglect more complex interaction structures between genes (Boyle et al., 2017). Moreover, most associations explain a small fraction of the genetic variation in disease susceptibility, which means that the therapeutic benefit from intervening on them would be quite small (Goldstein et al., 2009). These criticisms explain why candidate-gene approaches remain popular among many researchers, but no research to date has empirically explored whether and under which conditions GWASs discover scientifically impactful gene-disease associations

4 Data

4.1 Information on Gene-Disease Associations

I construct a dataset of all novel gene-disease associations (GDAs) introduced in the period 1980-2016. I retrieve such information from DisGeNET (v7.0), an aggregator that is considered the most complete repository of scientific results linking human diseases to their genetic causes (Hermosilla and Lemus, 2019; Piñero et al., 2020). This database collects GDAs harvested from an array of specialized sources, including curated datasets and raw publications. My data are at the GDA level, and for each association I retrieve the list of journal articles indexed in PubMed that studied it. I focus my attention on associations mapping a protein-coding gene to a disease, syndrome, or abnormality with clear health implications.[2] My final sample includes 358,390 gene-disease

[2] Scientists routinely complain that associations proposed in academic publications often turn out not to be robust. To restrict the sample to the most plausible ones, I use the DisGeNET-provided *Evidence Index* to retain in my data

associations between 14,112 genes and 15,039 narrow disease categories.

To identify which associations are introduced with a data-driven approach, I rely on the GWAS Catalog, a manually curated source managed by the European Bioinformatics Institute (MacArthur et al., 2017). The GWAS Catalog is the most reliable list available of genome-wide association studies published in peer-reviewd journals. Studies are eligible for inclusion in the GWAS Catalog if they include an array-based GWAS analysis that does not target any specific gene ex ante. The Catalog also collects the details of the specific gene-disease associations tested in the study. Following the best research practices, only associations with a high statistical significance (p-value $< 1.0 \times 10^{-5}$) are considered (Marigorta et al., 2018). My sample includes 8,661 GDAs introduced by 1,251 distinct genome-wide association studies. Panel (b) of Figure 1 shows the rapid growth of GWASs since 2005, when the first such study was published.

I also gather a few additional gene level attributes. For each gene I record whether it has a homolog gene in the lab mouse (Clarke, 2002; Murray et al., 2016). Homologs are genes inherited in two species by a common ancestor, thus retaining similar functions and biological features. Since the mouse is the most used scientific tool for gene knockouts (i.e., a lab technique to study the role of a gene by preventing its normal functioning), genes without a mouse homolog are usually more complex to experimentally study.[3] Similarly, I code a variable for genes that are systematically expressed in fewer tissues of the body. Gene expression is the process by which the information encoded in a gene is used in the creation of a gene product, such as proteins (Lopes et al., 2021). Certain genes are expressed only in select human tissues, which makes it less convenient to collect genetic material for experimental studies. Stoeger et al. (2018) documents that for this reason, scientists have historically focused on genes expressed in a large variety of body tissues.

4.2 Outcome Variables

My objective is comparing the characteristics of GWAS-established gene-disease associations vis-á-vis associations established with a candidate-gene approach. To do so, I ask if conditional on being introduced by a GWAS, gene-disease associations are more likely to present different attributes. In particular, I focus on two outcome variables:

only associations for which contradictory results represent less than 10% of the available publications about them. The results are robust to stricter thresholds, as well as to keeping the whole DisGeNET data. See the Appendix for robustness checks.

[3] This intuition is confirmed in the DisGeNET data, where I observe that genes without a mouse homolog received on average 22% less publications on how they might be implicated in human diseases. In general, reliance on model organisms and the availability of gene homologs profoundly shapes research choices (Baba and Walsh, 2010).

Underexplored Gene: The first dependent variable is a dummy that takes value one for gene-disease associations that include a gene never associated to a disease before 2005. In additional analyses, I use two alternative proxies to capture which genes received scant attention before the emergence of GWASs. The first is the date of discovery of the gene, since many of the genes mapped by the Human Genome Project are still overlooked due to path-dependent research choices (Stoeger et al., 2018). Accordingly, I explore if GWASs are more likely to implicate in a disease genes discovered after the year 2000. The second proxy is the presence of genetic annotations dated before 2005, as recorded by the Gene Ontology (GO Consortium, 2021). A Gene Ontology annotation is any statement about the function of a gene, which means that genes without annotations received very little study.

Scientific Impact: The second dependent variable is a dummy that takes value one for gene-disease associations that have a large scientific impact. Usually researchers rely on paper-to-paper citation counts to measure impact, but this would be misleading in this context since GWASs are highly cited on average (μ_{GWAS}=187 vs $\mu_{TargetedSearch}$=42). This is due to a variety of reasons unrelated with the scientific quality of the findings, such as reviews, criticisms, or commentaries that discuss the results of the genome-wide approach. Therefore, I exploit DisGeNET to build a cleaner measure of scientific impact: the number of papers that *directly* build upon the gene-disease combination. These include empirical and experimental work that investigates the proposed association, regardless of whether they cite the paper that introduced it, hence being a more truthful measure of impact. For each year, I code as high-impact all new GDAs in the 95th percentile of follow-on work received.

4.3 Summary Statistics

Table 1 lists the key variables used in the analysis with the summary statistics for the sample. Panel A provides summary statistics about the publications that introduced new GDAs in the period 2005-2016. Besides being more cited than candidate-gene papers, each GWAS introduces on average more associations spanning a larger number of genes. Panel B provides summary statistics at the GDA level. Previewing the analysis follows, it appears that genes associated to a disease by GWASs are more likely to be understudied and complex to experimentally study. The incidence of high-impact associations is also higher for GWASs than for candidate-gene papers.

5 Results

5.1 Data-Driven Search and the Scope of Innovation

I use OLS to estimate the following regression specification using gene-disease level data: $\mathbb{I}(GDA\ with\ understudied\ gene\ > 0)_i = \alpha + \beta\ \mathbb{I}(Introduced\ by\ GWAS\ > 0)_i + \gamma \boldsymbol{X}_i + \epsilon_i$, where $\boldsymbol{X}_i$ include disease fixed effect and controls for year, journal prestige, and number of authors of the paper that introduced GDA i. $\mathbb{I}(GDA\ with\ understudied\ gene\ > 0)_i$ is an indicator variable equal to one if GDA i includes an understudied gene. $\mathbb{I}(Introduced\ by\ GWAS\ > 0)_i$ takes value one for GDAs introduced by a GWAS, and zero otherwise. This specification compares the difference between GDAs that have first appeared in a genome-wide association study with GDAs that appeared in candidate-gene papers. If data-driven search leads to diversify search, then I should find that the OLS estimate β is positive. All my specifications cluster standard errors two-way at the gene and disease level.

Table 2 presents estimates from this regression. The main result is that genome-wide association studies are significantly more likely to associate understudied genes with human diseases. Specifically, the estimate of β in Column 1 indicates an average increase of 0.20 percentage points on the probability of combining a gene never associated to a disease before 2005, a significant increase given that the baseline is about 0.13 percentage points. This means the likelihood of introducing innovation encompassing little-studied genes is more than doubled, albeit on a low base-rate. Column 2 and 3 show that this result are robust to alternative definitions of gene popularity, including the date of discovery of the gene and the absence of biological annotations in the Gene Ontology before the emergence of GWASs. The results suggest that a change in search strategy forcefully affects the direction of innovation, leading to consideration of otherwise short-changed genes.

Figure 2 presents an intuitive visualization of the combinatorial space of pairwise gene-disease combinations. Comparing the areas searched by targeted studies with the findings of genome-wide association studies illustrates the difference between the two strategies. New combinations introduced by GWASs span a much wider area of the technological landscape, while targeted search tends to replicate existing research patterns. Panel (b) also validates the global nature of GWAS: for each disease investigated, the range of genes associated spans the entire genome. However, the figure points to the fact that GWASs keep focusing on diseases historically well studied. This confirms that the decision of where to direct the data-collection effort remains crucial in determining the direction of search, but also that the results above are due to the search strategy itself and not to a change in disease focus.

Why do scientists keep focusing on a narrow subset of genes? One possibility is that this choice is guided by specific gene functional characteristics that make them more meaningful to study. To explore this idea, I consider new associations that recombine genes that are part of a gene family. Such genes usually have the same name followed by a number that characterizes the order in which they were historically discovered (e.g., BRCA1 and BRCA2, discovered in 1994 and 1995 respectively). Genes in a family are formed by duplication of a single original gene, and generally share very similar biochemical functions (Daugherty et al., 2012). However, Stoeger et al. (2018) document that the first gene of a family tends to be much more studied than the second member. The discrepancy is large and cannot be explained by biological relevance (Gates et al., 2021). Figure 3 shows that GWAS lead to diversifying the direction of search even in this case, raising by 20% the probability that the second gene in the family is recombined. This suggests that one of the mechanisms through which GWASs help discovery is by counteracting inertial forces in scientists' research paths.

5.2 When Does Data-Driven Search Lead to Better Innovations?

While the previous section showed that data-driven search broadens the scope of search, its impact on the value of innovation is theoretically ambiguous. It could be that the diversification of genetic focus comes at the cost of lower-valued innovations (Arts and Fleming, 2018). Whether reliance on data outperforms targeted approaches, and the conditions under which this happens, is an empirical question. In this section, I answer this question estimating the following specification: $\mathbb{I}(GDA\ in\ top\ 5\%\ of\ impact > 0)_i = \alpha + \beta\ \mathbb{I}(Introduced\ by\ GWAS\ > 0)_i + \gamma \boldsymbol{X}_i + \epsilon_i$, where $\mathbb{I}(GDA\ in\ top\ 5\%\ of\ impact > 0)_i$ is an indicator variable equal to one if GDA i is among the top 5% most impactful combinations. All the other variables and controls are the same of the specification in the previous section. Column 1 of Table 3 presents the results. New gene-disease associations that first appeared in a genome-wide association study are on average 32% more likely to be among those of high scientific impact.

The positive effect of data-driven search on the value of innovation is thus large and statistically significant. Yet, it could be just a reflection of the change in genetic focus documented in the previous section. If the new genes recombined by GWASs are intrinsically more likely to yield high-value associations, then the result above would be mechanical. To investigate this possibility, I estimate the same model with the addition of gene fixed effects, hence absorbing the cross-sectional variation linked to a gene's scientific potential. Column 2 of Table 3 shows that the magnitude of the coefficient grows after the addition of gene fixed effects. Taken together, the estimates in Table 3 find that GDAs introduced by genome-wide association studies have on average higher scientific impact than comparable associations discovered with candidate-gene approaches.

In additional heterogeneity analyses, I use split sample regressions to explore how the results above change for different subsets of genes. First, I restrict my attention to genes more complex to experimentally study, either because they lack a homolog in the lab mouse or because they are expressed in fewer tissues of the body (Stoeger et al., 2018). Figure 4 graphically shows the estimates of β from the above specification. Compared to the baseline of Column 2 in Table 3, the effect size grows by 50% in magnitude. The estimates are more noisy, reflecting the higher variability of outcomes when recombining complex components, but it confirms that data-driven search yields better outcomes on rugged terrains where breakthroughs are more likely to happen (Fleming and Sorenson, 2001).

Second, I explore if targeted studies could be more effective in areas where scientists have better biological knowledge. Figure 5 plots the share of high-impact GDAs in correspondence of each gene, distinguishing by the search strategy that introduced them. Genes on the X axis are sorted by the number of pre-2005 publications received. A striking pattern emerges: while the ability of GWASs to introduce valuable gene-disease combinations is roughly constant across the genetic landscape, candidate gene studies are much more effective for genes that have received more study in the past. Auxiliary regressions in the Appendix confirm the pattern emerging from this graphical representation, underscoring how targeted search yields higher value innovations in known areas of the technological landscapes.

5.3 Robustness Checks

A. Considering Only Less Controversial Associations: Gene-disease associations require extensive follow-on work to be validated and contradictory results on their robustness are not infrequent. The main sample used in this paper considered associations for which DisGeNET reports less than 10% of contradictory papers on them. In Figure C.3 I test the robustness of the main results to different selections of the sample, ranging from all DisGeNET associations to the inclusion of only those for which no contrasting evidence exists. Results are robust and quantitatively similar regardless of the sample chosen.

B. Epistemic Uncertainty of Data-Driven Findings: One might worry that the effects showed in Table 3 are generally due to the fact that GWAS provide a less solid form of evidence, so that a higher level of follow-on work would only be confirming that more research is necessary to validate *any* data-driven finding. However, if this was the case, among GWAS-established gene-disease links we should observe that those with more convincing evidence need less confirmatory follow-on work. To test this conjecture, I managed to match a subset of GWAS-established gene-disease associations with their respective p-value. Table C.1 shows that more convincing gene-disease links (i.e. those with lower p-values, hence higher -$\log_{10}$(p-value)) have also more follow-on work, but just for understudied genes. It is indeed for those that a stronger p-value should help reduce uncertainty more, hence generating larger scientific interest.

C. Alternative Measure of Scientific Validity: Instead of relying on the number of subsequent publications as a measure of associations' scientific potential, I carry out a robustness check using DisGeNET's *GDA Score* (Piñero et al., 2020). The GDA Score synthetically captures the scientific reliability of the existing evidence on the gene-disease association. Table C.2 presents the coefficient of the OLS regressions for each of the subsamples of genes analyzed in Section 5.2. The results confirm the earlier findings on the effectiveness of data-driven search in introducing combinations of higher scientific value.

6 Conclusion

In this paper, I explore how data shapes the search for innovation. Unlike local or theory-driven search, I argue that data enables global search strategies that lead to more exploratory experimentation. Empirical results in the context of genome-wide association studies confirm this idea, showing that data-driven search is untethered by past choices. Data leads innovators to experiment with short-changed areas of the technological landscapes and helps them to uncover combinations of higher average value. The latter result is stronger in rugged areas of the landscape, but targeted approaches are more effective when theoretical knowledge can be used to guide search.

This paper has practical implications for scientists, managers, and governments. For individual researchers, my results show the conditions under which alternative search strategies are more or less effective, suggesting that data analytics should be used especially when venturing in uncharted domains. More in general, data analytics are diffusing in every sector of the economy, but the returns remain heterogeneous and concentrated among few companies (Brynjolfsson et al., 2021). My results provide an additional rationale for furthering investments in large-scale public data sources that might enable data-driven search (Nagaraj et al., 2020; Kao, 2022).

It is also important to note that despite its large potential, data-driven search is not a panacea for recombinant search. Exclusive reliance on data could itself end up hindering search, either because it can lead scholars to look only "where the light is" (Hoelzemann et al., 2022) or because blind reliance on available data might replicate their biases (Cao et al., 2021). Moreover, data will be of little help when trying to invent entirely new technologies that do not emerge from old components (Wu et al., 2020). As such, data-driven search constitutes within-paradigm search that might be subject to technological exhaustion unless new technological components are added over time (Fleming, 2001).

Finally, despite the contributions outlined above, a few limitations of this paper must be acknowledged. First, the present study is purely descriptive in nature. The results do seem robust to a battery of tests, but my work does not account for the endogenous choice of whether to use data-driven search or not. Second, the patterns documented are from a quantitative case study of a single domain. The task of locating the genetic roots of human diseases is crucial for drug discovery, but it has specificities that might not directly translate to other settings. While an increasing number of domains is receiving complete maps of the relevant technological landscapes, just like the Human Genome Project did for the genome, data-driven search might remain unfeasible in other contexts. More research will be needed to investigate the external validity of my findings. Finally, my work does not explore how data-driven findings affect the downstream generation of new drugs that build over them. This is an exciting avenue for follow-up work that is outside the scope of my paper.

References

Agrawal, A., J. McHale, and A. Oettl (2019): "Finding needles in haystacks: Artificial intelligence and recombinant growth," in *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press.

——— (2022): "Superhuman science: How artificial intelligence may impact innovation," *Brookings Working Paper*.

Arora, A. and A. Gambardella (1994): "The changing technology of technological change: general and abstract knowledge and the division of innovative labour," *Research policy*, 23, 523–532.

Arts, S. and L. Fleming (2018): "Paradise of novelty—or loss of human capital? Exploring new fields and inventive output," *Organization Science*, 29, 1074–1092.

Azevedo, E. M., A. Deng, J. L. Montiel Olea, J. Rao, and E. G. Weyl (2020): "A/b testing with fat tails," *Journal of Political Economy*, 128, 4614–000.

Baba, Y. and J. P. Walsh (2010): "Embeddedness, social epistemology and breakthrough innovation: The case of the development of statins," *Research Policy*, 39, 511–522.

Bessen, J. E., S. M. Impink, L. Reichensperger, and R. Seamans (2021): "The Role of Data for AI Startup Growth," *Available at SSRN 3896463*.

Boyle, E. A., Y. I. Li, and J. K. Pritchard (2017): "An expanded view of complex traits: from polygenic to omnigenic," *Cell*, 169, 1177–1186.

Brynjolfsson, E., W. Jin, and K. McElheran (2021): "The power of prediction: predictive analytics, workplace complements, and business performance," *Business Economics*, 56, 217–239.

Brynjolfsson, E. and K. McElheran (2016): "The rapid adoption of data-driven decision-making," *American Economic Review*, 106, 133–39.

Bush, W. S. and J. H. Moore (2012): "Genome-wide association studies," *PLoS Computational Biology*, 8, e1002822.

Cao, R., R. M. Koning, and R. Nanda (2021): "Biased sampling of early users and the direction of startup innovation," *NBER Working Paper 28882*.

Cargill, M., S. J. Schrodi, M. Chang, V. E. Garcia, R. Brandon, K. P. Callis, N. Matsunami, K. G. Ardlie, D. Civello, J. J. Catanese, et al. (2007): "A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes," *The American Journal of Human Genetics*, 80, 273–290.

Clarke, T. (2002): "Mice make medical history," *Nature*.

Cockburn, I. M., R. Henderson, and S. Stern (2019): "The impact of artificial intelligence on innovation," in *The Economics of Artificial Intelligence: An Agenda*, Chicago: Chicago University Press, 115–146.

Daugherty, L. C., R. L. Seal, M. W. Wright, and E. A. Bruford (2012): "Gene family matters: expanding the HGNC resource," *Human Genomics*, 6, 1–6.

Deniz, B. C. (2020): "Experimentation and incrementalism: The impact of the adoption of A/B Testing," *Stanford GSB Working Paper*.

Dougherty, D. and D. D. Dunne (2012): "Digital science and knowledge boundaries in complex innovation," *Organization Science*, 23, 1467–1484.

Edwards, A. M., R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and H. Y. Frank (2011): "Too many roads not taken," *Nature*, 470, 163–165.

Evans, J. and A. Rzhetsky (2010): "Machine science," *Science*, 329, 399–400.

Farboodi, M. and L. Veldkamp (2020): "Long-run growth of financial data technology," *American Economic Review*, 110, 2485–2523.

Felin, T., J. Koenderink, J. I. Krueger, D. Noble, and G. F. Ellis (2021): "Data bias," *Genome biology*, 22, 1–4.

FELIN, T. AND T. R. ZENGER (2017): "The theory-based view: Economic actors as theorists," *Strategy Science*, 2, 258–271.

FLEMING, L. (2001): "Recombinant uncertainty in technological search," *Management Science*, 47, 117–132.

FLEMING, L. AND O. SORENSON (2001): "Technology as a complex adaptive system: evidence from patent data," *Research Policy*, 30, 1019–1039.

——— (2004): "Science as a map in technological search," *Strategic Management Journal*, 25, 909–928.

GATES, A. J., D. M. GYSI, M. KELLIS, AND A.-L. BARABÁSI (2021): "A wealth of discovery built on the human genome project—by the numbers," *Nature*, 590, 212–215.

GAVETTI, G. AND D. LEVINTHAL (2000): "Looking forward and looking backward: Cognitive and experiential search," *Administrative Science Quarterly*, 45, 113–137.

GHOSH, S. (2021): "Experimental Approaches to Strategy and Innovation," Ph.D. thesis, Harvard University.

GITTELMAN, M. (2016): "The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery," *Research Policy*, 45, 1570–1585.

GO CONSORTIUM (2021): "The Gene Ontology resource: enriching a GOld mine," *Nucleic acids research*, 49, D325–D334.

GOLDSTEIN, D. B. ET AL. (2009): "Common genetic variation and human traits," *New England journal of medicine*, 360, 1696.

GUDMUNDSSON, J., G. JOHANNESDOTTIR, J. T. BERGTHORSSON, A. ARASON, S. INGVARSSON, V. EGILSSON, AND R. B. BARKARDOTTIR (1995): "Different tumor types from BRCA2 carriers show wild-type chromosome deletions on 13q12–q13," *Cancer research*, 55, 4830–4832.

HERMOSILLA, M. AND J. LEMUS (2019): "Therapeutic Translation of Genomic Science," in *Economic Dimensions of Personalized and Precision Medicine*, Chicago: Chicago University Press.

HOELZEMANN, J., G. MANSO, A. NAGARAJ, AND M. TRANCHERO (2022): "The Streetlight Effect in Data-Driven Exploration," *mimeo*.

INTERNATIONAL HAPMAP CONSORTIUM (2005): "A haplotype map of the human genome," *Nature*, 437, 1299.

JAYARAJ, S. AND M. GITTELMAN (2018): "Scientific Maps and Innovation: Impact of the Human Genome on Drug Discovery," *DRUID Society Conference Paper*, 1–56.

JONES, B. F. (2009): "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" *The Review of Economic Studies*, 76, 283–317.

JONES, C. I. AND C. TONETTI (2020): "Nonrivalry and the Economics of Data," *American Economic Review*, 110, 2819–58.

KAO, J. (2022): "Charted Territory: Evidence from Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry," *Mimeo*.

KAPLAN, S. AND K. VAKILI (2015): "The double-edged sword of recombination in breakthrough innovation," *Strategic Management Journal*, 36, 1435–1457.

KLAHR, D. AND K. DUNBAR (1988): "Dual space search during scientific reasoning," *Cognitive Science*, 12, 1–48.

KNEELAND, M. K., M. A. SCHILLING, AND B. S. AHARONSON (2020): "Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation," *Organization Science*, 31, 535–557.

KONING, R., S. HASAN, AND A. CHATTERJI (2022): "Experimentation and Start-up Performance: Evidence from A/B Testing," *Management Science*.

KUHN, T. S. (1962): *The Structure of Scientific Revolutions*, Chicago: Chicago University Press.

LE FANU, J. (2011): *The rise and fall of modern medicine*, London: Hachette.

LOPES, I., G. ALTAB, P. RAINA, AND J. P. DE MAGALHAES (2021): "Gene Size Matters: An Analysis of Gene Length in the Human Genome," *Frontiers in Genetics*, 12, 30.

LOU, B. AND L. WU (2021): "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms," *MIS Quarterly*, 45.

MACARTHUR, J., E. BOWLER, M. CEREZO, L. GIL, P. HALL, E. HASTINGS, H. JUNKINS, A. MCMAHON, A. MILANO, J. MORALES, ET AL. (2017): "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic Acids Research*, 45, D896–D901.

MARCH, J. G. (1991): "Exploration and exploitation in organizational learning," *Organization Science*, 2, 71–87.

MARIGORTA, U. M., J. A. RODRÍGUEZ, G. GIBSON, AND A. NAVARRO (2018): "Replicability and prediction: lessons and challenges from GWAS," *Trends in Genetics*, 34, 504–517.

MURRAY, F., P. AGHION, M. DEWATRIPONT, J. KOLEV, AND S. STERN (2016): "Of mice and academics: Examining the effect of openness on innovation," *American Economic Journal: Economic Policy*, 8, 212–52.

NAGARAJ, A. (2021): "The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry," *Management Science*.

NAGARAJ, A., E. SHEARS, AND M. DE VAAN (2020): "Improving data access democratizes and diversifies science," *Proceedings of the National Academy of Sciences*, 117, 23490–23498.

NELSON, M. R., H. TIPNEY, J. L. PAINTER, J. SHEN, P. NICOLETTI, Y. SHEN, A. FLORATOS, P. C. SHAM, M. J. LI, J. WANG, ET AL. (2015): "The support of human genetic evidence for approved drug indications," *Nature Genetics*, 47, 856–860.

NELSON, R. R. (1982): "The role of knowledge in R&D efficiency," *Quarterly Journal of Economics*, 97, 453–470.

OPREA, T. I., C. G. BOLOGA, S. BRUNAK, A. CAMPBELL, GAN, ET AL. (2018): "Unexplored therapeutic opportunities in the human genome," *Nature reviews Drug discovery*, 17, 317–332.

PEARSON, T. A. AND T. A. MANOLIO (2008): "How to interpret a genome-wide association study," *Jama*, 299, 1335–1344.

PIÑERO, J., J. M. RAMÍREZ-ANGUITA, J. SAÜCH-PITARCH, F. RONZANO, E. CENTENO, F. SANZ, AND L. I. FURLONG (2020): "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, 48, D845–D855.

ROCK, D. (2019): "Engineering value: The returns to technological talent and investments in artificial intelligence," *Available at SSRN 3427412*.

RZHETSKY, A., J. G. FOSTER, I. T. FOSTER, AND J. A. EVANS (2015): "Choosing experiments to accelerate collective discovery," *Proceedings of the National Academy of Sciences*, 112, 14569–14574.

SCHILLING, M. A. AND E. GREEN (2011): "Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences," *Research Policy*, 40, 1321–1331.

SILVERBERG, G. AND B. VERSPAGEN (2007): "The size distribution of innovations revisited: an application of extreme value statistics to citation and value measures of patent significance," *Journal of Econometrics*, 139, 318–339.

STOEGER, T., M. GERLACH, R. I. MORIMOTO, AND L. A. NUNES AMARAL (2018): "Large-scale investigation of the reasons why potentially important genes are ignored," *PLoS Biology*, 16, e2006643.

STOKES, J. M., K. YANG, K. SWANSON, W. JIN, A. CUBILLOS-RUIZ, N. M. DONGHIA, C. R. MACNAIR, ET AL. (2020): "A deep learning approach to antibiotic discovery," *Cell*, 180, 688–702.

STUART, T. E. AND J. M. PODOLNY (1996): "Local search and the evolution of technological capabilities," *Strategic Management Journal*, 17, 21–38.

TABOR, H. K., N. J. RISCH, AND R. M. MYERS (2002): "Candidate-gene approaches for studying complex genetic traits: practical considerations," *Nature Reviews Genetics*, 3, 391–397.

UFFELMANN, E., Q. Q. HUANG, N. S. MUNUNG, J. DE VRIES, Y. OKADA, A. R. MARTIN, H. C. MARTIN, T. LAPPALAINEN, AND D. POSTHUMA (2021): "Genome-wide association studies," *Nature Reviews Methods Primers*, 1, 1–21.

VISSCHER, P. M., N. R. WRAY, Q. ZHANG, P. SKLAR, M. I. MCCARTHY, M. A. BROWN, AND J. YANG (2017): "10 years of GWAS discovery: biology, function, and translation," *The American Journal of Human Genetics*, 101, 5–22.

WEITZMAN, M. L. (1998): "Recombinant growth," *The Quarterly Journal of Economics*, 113, 331–360.

WILLIAMS, H. L. (2013): "Intellectual property rights and innovation: Evidence from the human genome," *Journal of Political Economy*, 121, 1–27.

WU, L., L. HITT, AND B. LOU (2020): "Data analytics, innovation, and firm productivity," *Management Science*, 66, 2017–2039.

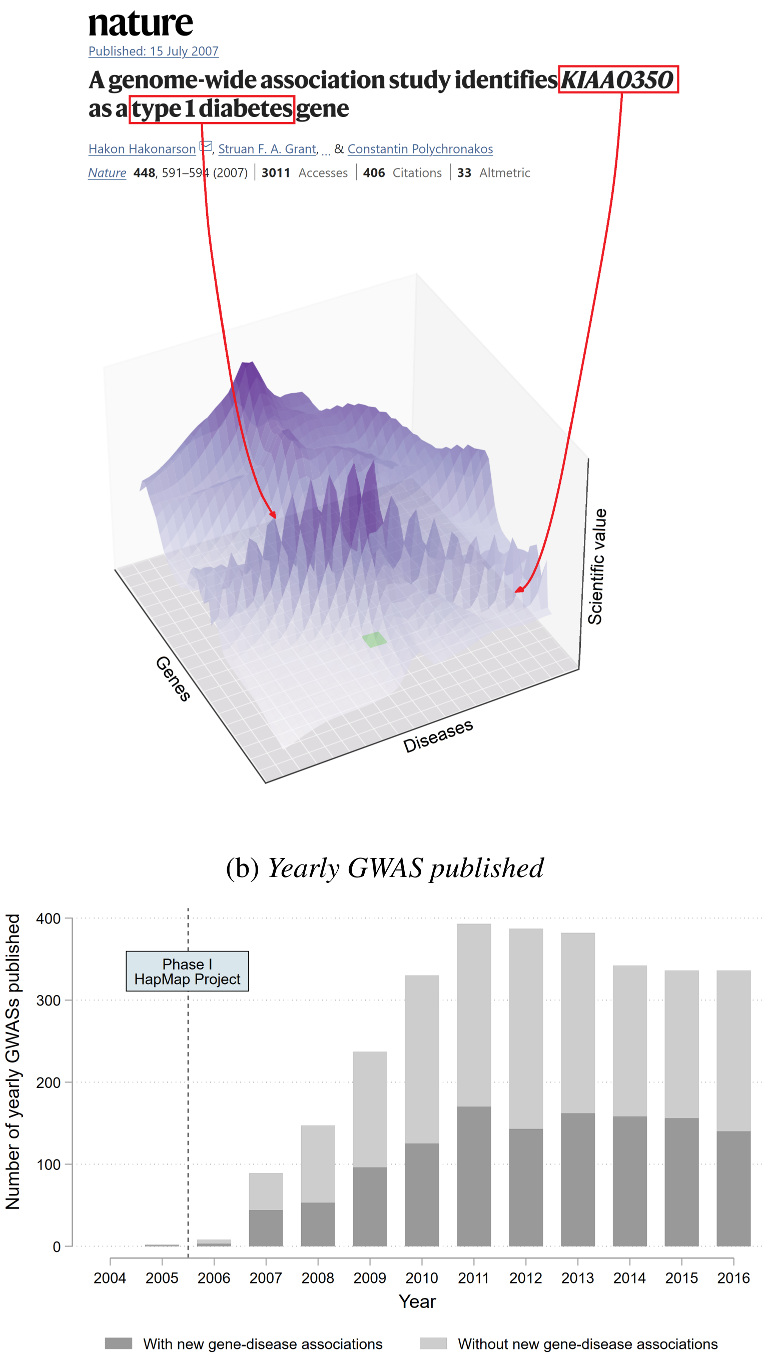
WU, L., B. LOU, AND L. HITT (2019): "Data analytics supports decentralized innovation," *Management Science*, 65, 4863–4877.

ZOLAS, N., Z. KROFF, E. BRYNJOLFSSON, K. MCELHERAN, D. N. BEEDE, C. BUFFINGTON, N. GOLDSCHLAG, L. FOSTER, AND E. DINLERSOZ (2021): "Advanced technologies adoption and use by US firms: Evidence from the annual business survey," *NBER Working Paper 28290*.

7 Figures and Tables

Figure 1: The emergence of genome-wide association studies in the search for novel gene-disease associations

(a) *GWAS as data-driven search*



nature

Published: 15 July 2007

A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene

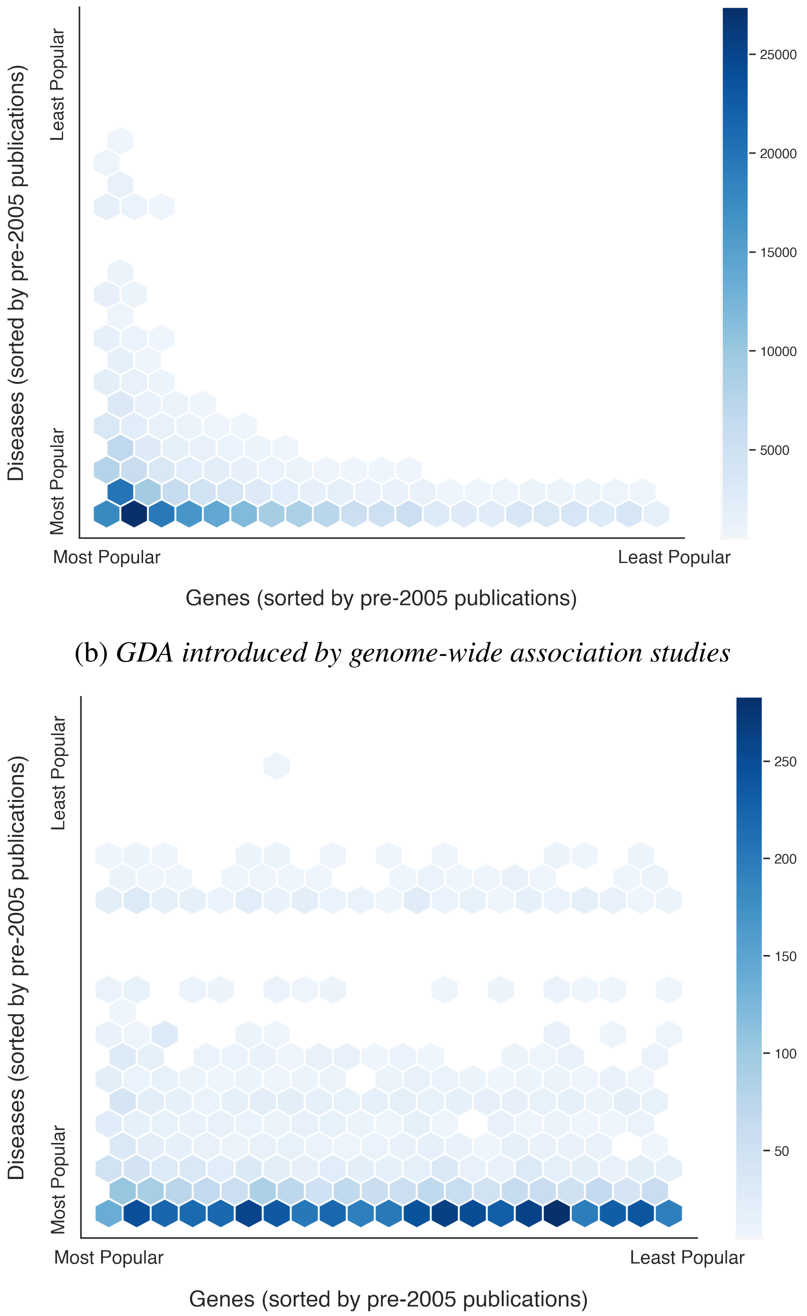
Hakon Hakonarson, Struan F. A. Grant, ... & Constantin Polychronakos

Nature **448**, 591–594 (2007) | **3011** Accesses | **406** Citations | **33** Altmetric

(b) *Yearly GWAS published*

Note: Panel (a) shows depicts how a typical GWAS introduces a new gene-disease associations in the combinatorial landscape. Each combination of gene and disease has a specific scientific value, captured by the elevation at that location. Panel (b) shows the number of yearly genome-wide association studies published. Data are from the GWAS Catalog (MacArthur et al., 2017). The vertical dashed line marks the completion of the Phase I of the HapMap project in October 2005.

Figure 2: Genome-wide associations studies span a larger portion of the genetic landscape relative to candidate-gene studies.

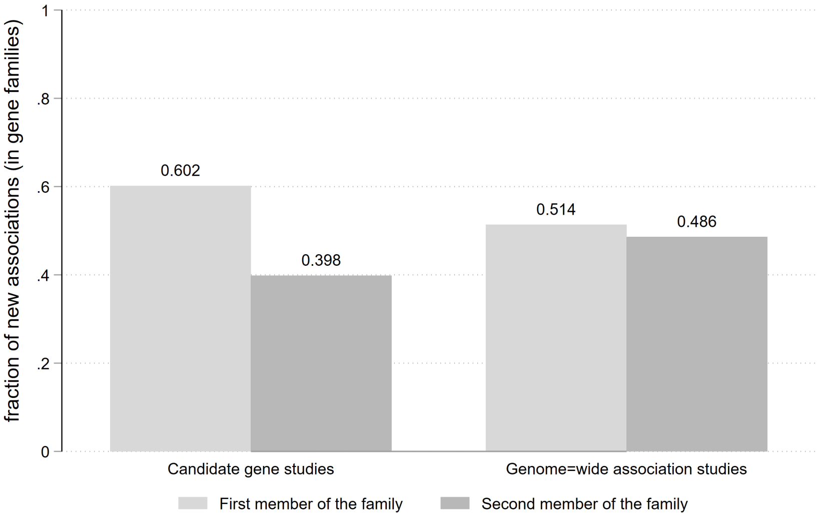


(a) *GDA introduced by candidate gene studies*

(b) *GDA introduced by genome-wide association studies*

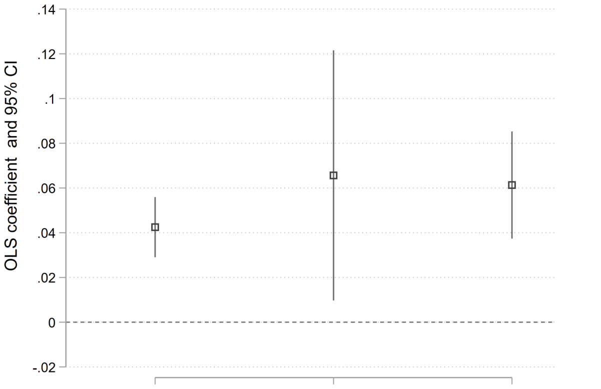
Note: Panel (a) shows a heatmap of new gene-disease associations introduced with targeted search strategies after 2005. Panel (b) shows a heatmap of new gene-disease associations introduced with data-driven search strategies after 2005. Darker areas correspond to a higher introduction of new GDAs. Both panels have 14,112 genes on the X axis, sorted from the most to the least studied in the pre-GWAS era, and 15,039 disease categories on the Y axis, sorted from the most to the least studied in the pre-GWAS era.

Figure 3: GWAS-introduced gene-disease associations are more likely to involve the second member of a gene family.



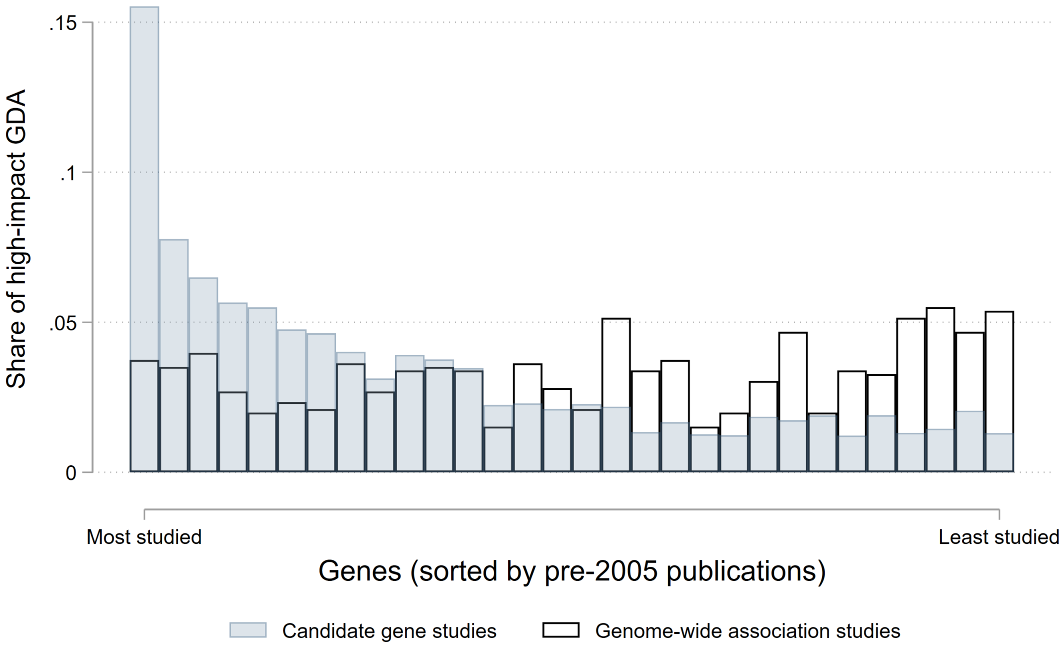
Note: The figure plots the share of new gene-disease associations involving a gene family, divided by whether the GDA involves the first or the second member of the family (e.g., BRCA1 vs. BRCA2). Data used in the graph are limited to all new gene-disease associations introduced in the period 2005-2016. Only diseases targeted by at least one GWAS are considered in this figure.

Figure 4: GWAS are especially effective to introduce high-impact gene-disease associations for genes complex to experimentally study.



Note: The graph plots OLS coefficients and 95% confidence intervals from split sample regressions. All models estimate the following specification: $\mathbb{I}(GDA\ in\ top\ 5\%\ of\ impact > 0)_{i,j} = \alpha + \beta\ \mathbb{I}(Introduced\ by\ GWAS\ > 0)_{i,j} + \delta Gene_j\ FE + \gamma \boldsymbol{X}_{i,j} + \epsilon_{i,j}$. The first coefficients is estimated on the full sample and provides the baseline estimate. The second coefficient is estimated only on GDAs involving a gene that does not have a homolog gene in the lab mouse. The third coefficient is estimated only on GDAs involving a gene expressed in few human body tissues. See text for details.

Figure 5: Gene-disease associations introduced by GWAS are less likely to be high-impact for highly-studied genes.



Note: The histogram plots the share of high-impact GDA for each gene distinguishing by the type of study that introduced them. The 14,112 genes on the X axis are sorted from the most to the least studied in the pre-GWAS era.

Table 1: Descriptive statistics.

	Panel A: paper level descriptives											
	Targeted search						GWAS					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Forward citations	42.19	23	89.016	0	8084	140,777	186.97	83	368.668	0	6244	1,251
Rank of the journal (ventile)	13.27	14	5.063	1	20	140,777	17.55	19	3.737	2	20	1,251
Associations per paper	4.75	3	6.848	1	927	140,777	10.53	5	20.531	1	307	1,251
Genes per paper	2.14	2	3.031	1	690	140,777	6.76	3	12.014	1	169	1,251
Number of authors	9.04	8	6.542	1	445	140,777	38.49	25	44.472	1	565	1,251
Year	2011.09	2011	3.394	2005	2016	140,777	2012.27	2012	2.527	2005	2016	1,251

	Panel B: association level descriptives											
	Targeted search						GWAS					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
With never recombined genes (%)	0.123	0	0.328	0	1	349,729	0.406	0	0.491	0	1	8,661
With recently discovered gene (%)	0.099	0	0.299	0	1	349,729	0.257	0	0.437	0	1	8,661
With never annotated genes (%)	0.480	0	0.500	0	1	349,729	0.675	1	0.468	0	1	8,661
With genes lacking mouse homolog (%)	0.052	0	0.221	0	1	343,446	0.060	0	0.238	0	1	8,524
With specific tissue expression genes (%)	0.247	0	0.431	0	1	349,729	0.271	0	0.444	0	1	8,661
In top 5% most impactful (%)	0.049	0	0.216	0	1	349,729	0.084	0	0.277	0	1	8,661
Year of the association	2011.08	2011	3.382	2005	2016	349,729	2012.68	2013	2.581	2005	2016	8,661

Note: Panel A presents descriptive statistics on papers that introduce new gene-disease associations after 2005. *Forward citations*= citations received by the focal article up to 2020 inclusive (data from NIH iCite); *Rank of the journal*= ventile of journal prestige (data from SCImago Journal Rank); *Associations per paper*= number of new GDA introduced; *Genes per paper*= number of genes associated to a disease. Panel B presents descriptive statistics on new gene-disease associations introduced after 2005. *With never recombined genes (%)*= share of GDAs that include a gene never associated to a disease before 2005; *With recently discovered genes (%)*= share of GDAs that include a gene discovered after the year 2000 (i.e., after the Human Genome Project); *With never annotated genes (%)*= share of GDAs that include a gene without any annotations in the Gene Ontology before 2005 (data from Gene Ontology); *With 2nd member of a gene family (%)*= share of GDAs that include the second member of a gene family, conditional on being about that gene family (data on gene families from stoeger2018large); *With genes lacking mouse homolog (%)*= share of GDAs that include a gene that does not have a gene homolog in the mouse (data from NIH); *With specific tissue expression genes (%)*= share of GDAs that include a gene is systematically expressed in fewer tissues of the body (data from stoeger2018large); *In top 5% most impactful (%)*= share of GDAs that fall in the top 95th percentile of follow-on work (by year of discovery); *Year of the association*= year in which the article introducing the GDA is published.

Table 2: Genome-wide association studies are more likely to introduce gene-disease associations involving less-studied genes.

Dependent Variable:	I(GDA for never associated gene>0)	I(GDA for recently discovered gene>0)	I(GDA for never annotated gene>0)
GWAS	0.201***	0.111***	0.140***
	(0.0139)	(0.0110)	(0.0134)
Disease FE	YES	YES	YES
Journal prestige FE	YES	YES	YES
Year of discovery FE	YES	YES	YES
Number of authors FE	YES	YES	YES
N	352,409	352,409	352,409
Mean of the DV:	0.130	0.103	0.485
Number of diseases:	9,740	9,740	9,740
Number of genes:	14,086	14,086	14,086

Note: *, **,*** denote significance at 10%, 5% and 1% level respectively. Observations at the gene-disease association (GDA) level. Std. err. clustered two-way at the gene and disease level. *I(GDA for never associated gene>0)*:0/1=1 if the gene-disease association involves a gene never associated to a disease before 2005; *I(GDA for recently discovered gene>0)*:0/1=1 if the gene-disease association involves a gene discovered after the year; *I(GDA for never annotated gene>0)*:0/1=1 if the gene-disease association involves a gene without any annotations in the Gene Ontology before 2005; *GWAS*=0/1=1 for GDAs introduced by a genome-wide association study.

Table 3: Genome-wide association studies are more likely to introduce gene-disease associations of high scientific impact.

Dependent Variable:	I(GDA in the top 5% of scientific impact > 0)	
GWAS	0.016**	0.042***
	(0.0063)	(0.0069)
Gene FE	NO	YES
Disease FE	YES	YES
Journal prestige FE	YES	YES
Year of discovery FE	YES	YES
Number of authors FE	YES	YES
N	352,409	350,932
Mean of the DV:	0.05	0.05
Number of diseases:	9,740	9,726
Number of genes:	14,186	12,623

Note: *, **,*** denote significance at 10%, 5% and 1% level respectively. Observations at the gene-disease association (GDA) level. Std. err. clustered two-way at the gene and disease level. *I(GDA in the top 5% of scientific impact>0)*:0/1=1 if the gene-disease association involves a gene in the top 95th percentile of follow-on work (by year of discovery); *GWAS*=0/1=1 for GDAs introduced by a genome-wide association study.

Data-Driven Search and Innovation

Appendix

Matteo Tranchero
UC Berkeley

A Data-Driven Search and Innovation Outcomes

A.1 An Example from Combinatorial Chemistry

Consider for instance the important task of discovering new drugs. This is a difficult problem since the chemical space is complex and high-dimensional (Rzhetsky et al., 2015; Jayaraj and Gittelman, 2018). Historically, most successful molecules were serendipitously identified with random search or in the neighborhood of known ones. This process was very long, costly, and inefficient (Gittelman, 2016). More recently, high-throughput screening (HTS) of large synthetic chemical libraries has opened up the possibility of large-scale rapid testing of millions of molecules. This approach usually involves pre-selecting a subset of the chemical space that has drug-like characteristics (e.g., small molecules with a weight theoretically deemed to be good for human-use drugs) and sequentially testing them in physical assay plates (Jayaraj and Gittelman, 2018). Yet, records of this approach have been mixed, since libraries are costly to maintain and the sequential screening has proved hard to scale to larger chemical spaces (Le Fanu, 2011).

However, new computational approaches and databases might provide an alternative for in silico testing that is not constrained by physical capacity or the compound libraries available. In a recent paper, Stokes et al. (2020) used a neural network approach to find molecules with antibacterial activity. Using data on known molecules to predict the bactericidal properties of structurally divergent molecules, the authors of the study discovered a new compound called halicin that has very promising antibiotic properties. This result was achieved with a fraction of the time and costs involved in sequential assay screening, and it is all the more remarkable considering that until then no clinical antibiotics had been discovered using targeted high-throughput screening (Stokes et al., 2020).

Besides being a consequential example, the discovery of halicin also highlights the conditions under which data-driven search is feasible. First, the relevant characteristics of all the potential components to recombine must be observable. In the case of drug discovery, for example, this translates into the need to measure the structural properties of chemical compounds screened. While almost tautological, this first condition restricts the scope of data-driven search to settings in which components are identifiable and measurable. Second, there has to exist an agreed-upon metric of technological potential on which the promise of each combination can be assessed. This constitutes the objective function that data-driven search tries to maximize by finding the candidate combinations that score highest on it. For Stokes et al. (2020) this was the growth inhibition of *Escherichia coli*, but is worth noting that a search guided only by data might not be feasible in fuzzier contexts where even the outcomes of the problem are ill-defined. Third, and relatedly, it must be possible to foresee the effect of novel combinations on the objective of interest. Going back to the example of antibiotics, this pertained to the prediction of whether a new compound could inhibit the growth of bacteria based on its structure.

A.2 When Does Data-Driven Search Yield Better Results?

Search strategies that operate by restricting the combinatorial space based on priors or experience face a clear trade-off. On the one hand, they decrease variability in outcomes and funnel experimentation on components expected to guarantee the highest payoffs. On the other hand, they limit the likelihood of new discoveries in underexplored areas of the technological landscape, potentially missing out on many innovations (Rzhetsky et al., 2015). Instead, data-driven search allows to consider a wider range of combinatorial possibilities, possibly leading to diversify knowledge production. Reliance on data to guide recombinant search removes the proclivity toward exploitation of known components. But will this lead to combinations of higher value? Existing research has found that more exploration often comes at the cost of lower-value inventions (Arts and Fleming, 2018). Therefore, whether data-driven search will ultimately improve the outcomes of search is an empirical question.

In this section, I present a simple statistical framework to explore the conditions under which data-driven search might overperform targeted search. Without loss of generality, we can represent search as starting from one component at a time, and then looking at which other component combine with it. More complex innovation involving several components can just be thought of as an iteration of this pairwise recombinant search. Each potential pairwise combination has a specific value, which gives rise to a distribution of outcomes. As in most innovation problems, we can safely assume that the distribution is very skewed: the majority of combinations are worth nothing, while a few are very valuable (Silverberg and Verspagen, 2007). The innovator can thus be imagined drawing from this distribution of outcomes, whose expected value and variance are synthetic measures of the average outcomes and the risk involved in search.

Against this backdrop, we can think how different search strategies might impact expected outcomes. Targeted search considers a subset of components expected to be of higher-mean, hence restricting the domain of the outcome distribution and resulting in a truncated distribution (Panel (a) of Figure C.1 shows an example). Truncated distributions have smaller variance – intuitively, they restrict the range of outcomes, thus limiting variability. In my context, this means that targeted search is less risky. However, the effect on truncation on the expected outcome depends on the interval over which the distribution has been truncated. For instance, if targeted search truncates the outcome distribution only from below, some probability mass has been shifted to higher values, hence increasing the yield of search. This would be the case if targeted search helps innovators to avoid dead ends and useless combinations (Fleming and Sorenson, 2004). The opposite would happen if search gets stuck on a suboptimal set of components (Fleming, 2001).

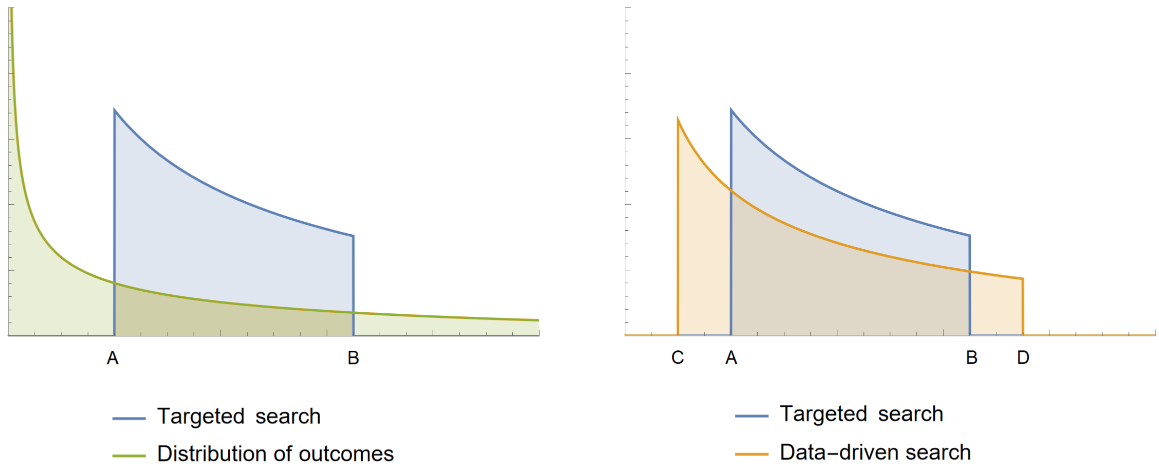
As argued in the previous section, data-driven search increases the breadth of components considered relative to targeted approaches. In the limit case when data on the entire landscape are available, the outcome distribution will be the true underlying distribution. Yet, whether this will lead to higher payoffs is ambiguous, and hinges on two factors. First, it will depend on how good targeted search is. If the outcome distribution is truncated from below and centered on the top tail, then any broadening of search will reduce the average payoff. This should be especially true when considering well-understood components, since inventors are already able to locate and focus on the best combinations (Kaplan and Vakili, 2015). Second, when targeted search truncates the outcome distribution from both sides (Panel (b) of Figure C.1), the expected value from an expansion of the domain will depend on how thick the right tail of outcomes is vis-á-vis the left tail of inferior alternatives (Azevedo et al., 2020). Research has shown that the right tail of extreme outcomes is thicker in complex technological landscapes where components are difficult to recombine (Fleming and Sorenson, 2001). In those cases, data-driven search might yield better results than targeted approaches.

B Genome-Wide Association Studies: Additional Details

Genome-wide association studies (GWAS) are case-control studies where researchers sequence the genomes from many people and look to see if any genetic variation in the DNA is more likely to appear in the group showing a specific trait rather than in the control group (Pearson and Manolio, 2008; Uffelmann et al., 2021). Figure C.2 provides a stylized depiction of a typical genome-wide study. Researchers start by collecting DNA samples both from people affected by the disease of interest and from healthy people. Then they use DNA microarrays to sequence genetic markers scattered throughout their genomes to reconstruct the genotype of the people in the sample. Finally, researchers test for statistically significant differences in their genotypes. Results are adjusted to account for multiple hypothesis testing and are graphically represented as a "Manhattan plot" showing the p-value of multiple statistical tests between DNAs in the case and control groups. The Y-axis is usually presented as -$\log_{10}$(p-value), hence higher values correspond to stronger associations.

C Additional Figures and Tables

Figure C.1: Consequences of alternative search strategies on the distribution of outcomes.

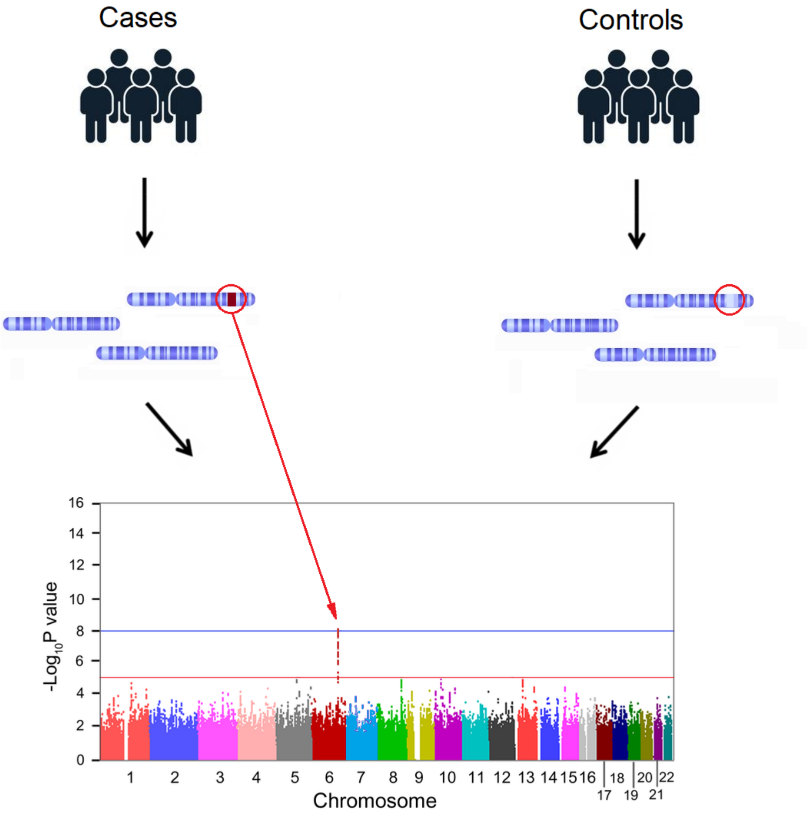


(a) *Targeted search as a truncated distribution*

(b) *Targeted search vs. data-driven search*

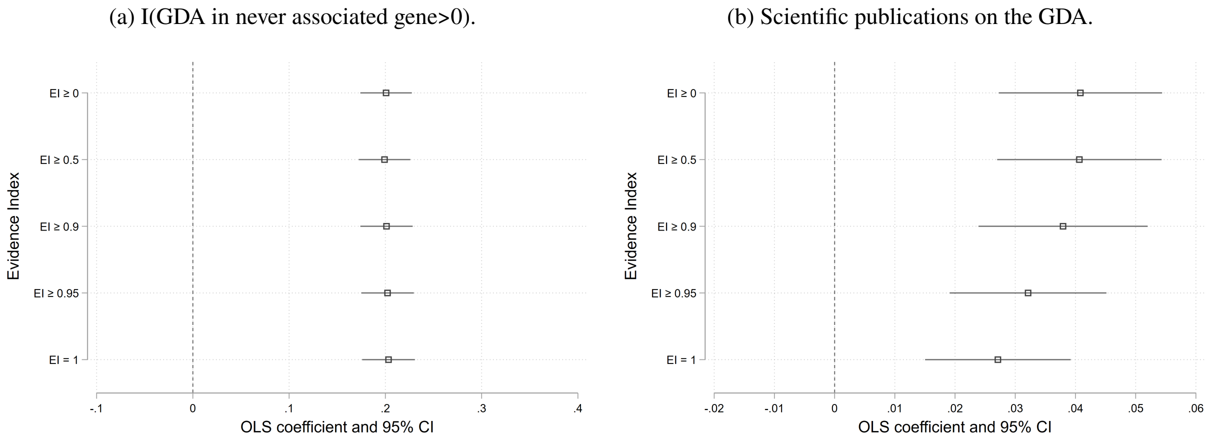
Note: Panel (a) exemplifies targeted search as a truncated distribution of outcomes. If for a given combinatorial problem there is a distribution of possible outcomes, then restricting the components considered will lead to over-sampling a high-mean subset of the population (in this case, the part on the [A,B] support). Panel (b) compares targeted search with data-driven search, which considers a wider range of potential combinations. While the variance of data-driven search will be higher, the comparison between expected values will depend on i) how good targeted search is; ii) how thick is the right tail of the outcome distribution compared to the left tail.

Figure C.2: Schema of how a typical genome-wide association study unfolds.



Note: The genome of people with and without a certain condition are sequence in search of significant differences; the panel at the bottom is the "Manhattan plot" which indicates the location in the chromosome of the statistically significant genetic variants. On the Y-axis there is the strength of the finding expressed as -log$_{10}$(p-value), hence higher values correspond to stronger associations.

Figure C.3: Robustness of the main results to the choice of sample.


(a) I(GDA in never associated gene>0).

(b) Scientific publications on the GDA.

Note: Panel (a) plots the coefficients and 95% confidence intervals of the regression of the likelihood that gene-disease associations discovered after 2005 involve an understudied gene on a dummy that indicates if the association was introduced by a GWAS. Panel (b) plots the coefficients and 95% confidence intervals of the regression of the number of follow-on scientific publications received by gene-disease associations discovered after 2005 on a dummy that indicates if the association was introduced by a GWAS. In each case the sample is restricted to associations with increasing values of the DisGeNET's *Evidence Index*, which captures the share of contradictory results on the association ($EI = \frac{N_{positive\ pubs}}{N_{total\ pubs}}$). The main analyses of the paper were done on the sample of $EI > 0.9$.

Table C.1: New gene-disease associations introduced by GWAS receive more follow-on work if the statistical association with the disease is stronger.

Dependent Variable:	Scientific publications on the association	
Subsample:	Genes previously associated	Genes never associated
-$\log_{10}$(p-value)	0.028	0.192**
	(0.0306)	(0.0922)
Gene FE	YES	YES
Disease FE	YES	YES
Journal prestige FE	YES	YES
Year of discovery FE	YES	YES
N	2,921	1,808
Mean of the DV:	1.013	0.971
Number of genes:	729	490
Number of diseases:	401	286

Note: *, **,*** denote significance at 10%, 5% and 1% level respectively. Observations at the gene-disease association (GDA) level. Std. err. clustered two-way at the gene and disease level. *Scientific publications on the association*= number of follow-on scientific articles on the new GDA identified. *-$\log_{10}$(p-value)*= statistical strength of the gene-disease association found by GWAS; higher values correspond to stronger associations. The first column considers all the GDA that involve a gene associated to a disease before 2005; the second column refers to all new GDA that involve a gene never associated to a disease before 2005.

Table C.2: Robustness of the main results to the use of DisGeNET's GDA Score as dependent variable.

Dependent Variable	DisGeNET *GDA Score* for gene-disease associations			
Subsample	All genes		Genes in mouse	Genes w/ specific expression
GWAS	0.037***	0.044***	0.048***	0.048***
	(0.0055)	(0.0054)	(0.0103)	(0.0066)
Gene FE	NO	YES	YES	YES
Disease FE	YES	YES	YES	YES
Journal prestige FE	YES	YES	YES	YES
Year of discovery FE	YES	YES	YES	YES
Number of authors FE	YES	YES	YES	YES
N	367,594	366,133	16,434	88,118
Number of diseases:	10,081	10,068	1,986	5,604
Number of genes:	14,089	12,641	745	3,450

Note: *, **,*** denote significance at 10%, 5% and 1% level respectively. Observations at the gene-disease association (GDA) level. Std. err. clustered two-way at the gene and disease level. *GDA Score*= synthetic measure of scientific reliability of the gene-disease association provded by DisGeNET; *GWAS*=0/1=1 for GDAs introduced by a genome-wide association study.