Data-Driven Search and Innovation [*]

Matteo Tranchero
UC Berkeley-Haas

First Version: June 12th, 2021
This Version: March 6th, 2023

Abstract

In a growing variety of contexts, correlational data are used to find successful innovations even when inventors do not fully grasp why those innovations should work. This method of data-driven recombinant search starkly contrasts with traditional approaches that rely on cause-effect knowledge to identify breakthroughs. In this paper, I develop a novel theoretical framework to examine the implications of a data-driven search process on innovative outcomes. Next, I empirically study this phenomenon in the context of human genomics, where the advent of genome-wide association studies (GWASs) has enabled a form of data-driven search for the genetic roots of diseases divorced from theoretical considerations. By comparing gene-disease combinations introduced by GWASs with those from theory-driven studies, I provide unique evidence of how the search process shapes innovation outcomes. My results show that discoveries introduced by GWASs span a wider portion of the genetic landscape, are more likely to involve neglected human genes, and are of higher scientific value than comparable combinations introduced by theory-driven studies. However, heterogeneity analyses reveal that data-driven search performs poorly with interdependent components because correlational data neglect complex interactions that only a theoretical understanding can capture. This paper contributes to exploring how data shape innovation and the conditions under which they might unlock breakthrough discoveries.

[*] E-mail: m.tranchero@berkeley.edu.

1 Introduction

Innovation is generated by recombining technological components in new ways. The universe of potential combinations is often depicted as a landscape over which innovators search (Aharonson and Schilling, 2016; Fleming and Sorenson, 2001; Nelson and Winter, 1982). Such a landscape has peaks of value in correspondence with the best opportunities but also large valleys of low-value recombinations. Previous research has documented a proclivity for individuals and organizations to overly focus on incremental modifications of previously successful combinations (Audia and Goncalo, 2007; Denrell and March, 2001; Rzhetsky et al., 2015; Stuart and Podolny, 1996). But as search in familiar domains necessarily runs into technological exhaustion, exploring uncharted spaces becomes necessary to discover new peaks (March, 1991; Levinthal, 1997; Fleming, 2001). This realization leads to a central question in management research: how can innovators find breakthrough recombinations in unknown technological spaces?

A growing body of research has suggested that innovators should be guided by cause-effect reasoning (Arora and Gambardella, 1994; Ehrig and Schmidt, 2022; Fleming and Sorenson, 2004). By developing new combinations based on theoretically-motivated hypotheses, inventors can find the best innovations more efficiently than haphazard trial-and-error (Camuffo et al., 2022; Gavetti and Levinthal, 2000; Felin and Zenger, 2017; Kneeland et al., 2020). However, a theory-driven search process seems at odds with the search practices of many successful firms nowadays (Allen, 2022; Agrawal et al., 2022; Thomke, 2020). For instance, scientists in pharmaceutical firms gather information on millions of compounds without knowing ex ante which ones might function as drugs (Evans and Rzhetsky, 2010; Jayaraj and Gittelman, 2018). Entrepreneurs use A/B testing to triage ideas at scale, iteratively testing configurations without assuming why some should be better (Koning et al., 2022). In an increasing variety of fields, innovators leverage large quantities of data instead of their theoretical priors to find peaks in novel domains (Nagaraj, 2022).

In this paper, I theorize that the ability to generate and analyze "big data" offers a new way to innovate in unknown technological spaces. Large-scale data allow the extraction of signals on the value of technological combinations, an approach that I call *data-driven search*. Inventors can use data to search through vast combinatorial spaces and to triage potential combinations at scale, without relying on theoretical priors or costly experiments (Agrawal et al., 2022; Choudhury et al., 2021; Christin, 2020; Shrestha et al., 2021). While theories operate by reducing the dimensionality of search problems (Gavetti and Levinthal, 2000), narrowing down efforts to areas that seem promising (Fleming and Sorenson, 2004), data-driven search enlarges the scope of combinations that can be assessed without actually testing them. Relative to theory-driven search, the result should be an increase in search breadth that might reduce the tendency to focus only on well-understood areas of the combinatorial landscape (Denrell and March, 2001).

While data enable a novel strategy for recombinant search, the consequences for innovation quality are not straightforward. If agents are already searching in the most fruitful areas of the landscape, then increasing the breadth of search will not improve the value of innovations. However, given that the tendency to focus on known components leads to technological exhaustion (Fleming, 2001), I argue that increasing search breadth will generally raise the likelihood of introducing breakthroughs (Katila and Ahuja, 2002; Leiponen and Helfat, 2010; Schilling and Green, 2011). Nevertheless, I also theorize that the beneficial effects of data-driven search will depend on the area of the landscape explored. Data-driven search facilitates recombinations in uncharted technological domains because, unlike theory-driven approaches, its efficacy is not bound to components theoretically understood (Anderson, 2008). But when knowledge of cause-effect mechanisms is available, this should allow theory-driven processes to recombine interdependent technological elements more effectively than data-driven search (Arora and Gambardella, 1994; Arts and Fleming, 2018; Fleming and Sorenson, 2004).

An empirical investigation of these ideas requires comparing the nature of discoveries resulting from theory-driven and data-driven search strategies. This is challenging for two reasons. First, one must find a well-defined search problem involving observable technological components. In particular, a meaningful comparison necessitates observing both search strategies used in the same landscape, on the same combinatorial task, and by comparable actors. Second, 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 knowing what kind of search process generated them (Schilling and Green, 2011; Kneeland et al., 2020; Maggitti et al., 2013). Investigating the consequences of data-driven search requires finding a setting where both these conditions are met at the same time.

This paper addresses these challenges with an empirical study of how scientists search for the genetic roots of human diseases. First, genes and diseases constitute the relevant components for this search problem. Any gene could, in principle, be tied to any condition, generating a landscape of millions of potential gene-disease combinations. The key objective of researchers is finding which combinations can serve as targets for drug development and enable new therapies. Second, searching for gene-disease associations can happen in two ways. Scientists interested in a disease can carry out *candidate gene studies* using their theoretical priors to target specific genes, or they can perform atheoretical *genome-wide association studies* (GWASs) that scan the whole genome to locate gene variants correlated to the disease. Importantly, I can infer what search process led to a specific gene-disease combination by coding the method used in the scientific article that introduced it. Comparing combinations established by GWASs to those discovered by candidate gene studies allows me to descriptively explore the characteristics and impact of data-driven search in this task so crucial for drug development.

I assemble a new dataset that includes the characteristics of gene-disease associations (GDAs) introduced in the period 1980-2016. The raw data are taken from DisGeNET, the most comprehensive aggregator of information on the genes associated with human diseases. DisGeNET collects the list of PubMed articles that studied each GDA, including the publication that first reported it and the list of subsequent papers that investigated it. I employ the count of follow-on studies that directly explored a gene-disease association as a GDA-level measure of scientific importance, regardless of whether such studies cite the paper that first introduced the association. Next, I use data from the European Bioinformatics Institute to identify papers that use a genome-wide approach, thus being able to code which associations were established with a GWAS. Using these data, the goal of this paper is not to estimate the causal effects of search processes on innovation outcomes but rather to provide a comparative analysis of how a data-driven approach changes the processes and results of recombinant search, and then point to potential channels through which these differences arise.

To fix ideas, consider the following two studies investigating the genetic origins of type 1 diabetes. Qu et al. (2007) leveraged their theoretical understanding of gene IRF5's function to explore for the first time its role in diabetes. In contrast, Hakonarson et al. (2007) chose a data-driven approach and documented a correlation between the same disease and mutations of KIAA0350.[1] Which of these two approaches can find the most scientifically valuable gene-disease combinations that might lead to a therapy? On the one hand, the theory-driven approach could be better because it leverages cause-effect knowledge of biological mechanisms. On the other hand, data-driven search might uncover promising drug targets if theoretical knowledge leads to search too narrowly. Indeed, in the case of type 1 diabetes, the GWAS by Hakonarson et al. (2007) associated a gene whose function was unknown at the time and thus never targeted by theory-driven studies. Moreover, my data show that twenty-two studies subsequently explored and validated the role of KIAA0350 in type 1 diabetes, placing this finding in the top tail of scientific importance, while only one study pursued the IRF5-type 1 diabetes combination. Since both papers were carried out by the same principal investigator, in the same year, and for the same disease, one can suspect that the nature of the search process itself is shaping these outcomes.

My empirical analysis suggests that this example is not isolated. Results show that data-driven search diversifies innovation by enabling the exploration of a wider portion of the genetic landscape. In baseline estimates, gene-disease combinations discovered by GWASs are 114-155% and 86-108% more likely to involve the least studied or more recently discovered among human genes, respectively. Additional tests suggest that GWASs achieve this by removing the ex ante choice of which genes to target, hence overcoming the path dependency that characterizes theory-driven search. Further, a comparison between data-driven and theory-driven search shows that the

[1] The current name for this gene is CLEC16A, but at the time when Hakonarson et al. (2007) published their study it was still known as KIAA0350. For the purposes of this paper, I will keep referencing it as KIAA0350.

former enables the discovery of more breakthrough gene-disease combinations, even holding the characteristics of genes and diseases constant. One potential concern is that this pattern reflects the superior research ability of the scientists that carry out GWASs. Like in the example above, I rule out this possibility by including principal investigator fixed effects, which control for time-invariant researcher characteristics that might lead to more breakthrough discoveries. Even among gene-disease associations introduced by the same researcher, those uncovered with a genome-wide association study are 77.4% more likely to be in the top tail of scientific importance.

Heterogeneity analyses confirm the hypothesis that the beneficial effects of data-driven search are contingent on the area of the landscape explored. I find that data are helpful in locating the best opportunities in uncharted technological areas where there is little theoretical guidance on how genes and diseases are related. However, candidate gene studies largely outperform GWAS in areas with a better theoretical understanding of genetic biology. To address the concern that the distribution of theoretical knowledge on the landscape is endogenous (Gittelman, 2016; Nelson, 2003), I exploit the fact that certain genes have historically been neglected simply because they cannot be studied using the lab mouse (Stoeger et al., 2018). Exploiting this variation in theory availability unrelated to the therapeutic potential of the genes, I confirm that data-driven search can effectively locate breakthroughs among technological components that are less theoretically known.

Finally, I investigate a potential mechanism behind this heterogeneity result. Previous literature has shown that the complexity of recombining interdependent technological elements is a key moderator of inventive success (Fleming and Sorenson, 2001). I operationalize this idea by coding which genes regulate the function of multiple downstream genes, thus being part of interrelated biological processes (Hermosilla and Lemus, 2019). My results show that gene-disease combinations uncovered by GWAS are less valuable when they include genes involved in larger regulatory networks. This suggests a fundamental limitation of data-driven search: data signals report correlations that are unable to consider complex interactions between coupled components (Ghosh, 2021). Instead, in the presence of a solid-enough theoretical understanding, researchers can locate the best technological combinations because they can take into account how complex elements interact (Fleming and Sorenson, 2004). Additional robustness tests reassure that my results are not due to the characteristics of scientists adopting data-driven search, the choice of diseases studied, or the definition of breakthrough employed.

This paper speaks to the burgeoning strand of research suggesting that data analytics is profoundly reshaping how invention happens (Agrawal et al., 2022; Choudhury et al., 2021; Cockburn et al., 2019). I contribute to this literature by providing new empirical evidence on the mechanisms that underlie this transformation (Allen, 2022; Bessen et al., 2022; Nagaraj, 2022; Lou and Wu, 2021;

Wu et al., 2020). I also add to the work on recombinant search by theorizing that data enable an alternative way to search for valuable recombinations in technological landscapes (Denrell and March, 2001; Fleming and Sorenson, 2001; Gavetti and Levinthal, 2000; Katila and Ahuja, 2002; Levinthal, 1997; March, 1991). My empirical analysis illustrates the boundary conditions of theory-driven search, informing the growing literature that warns against the perils of exclusive reliance on data (Cao et al., 2021; Hoelzemann et al., 2023). In particular, this paper highlights the role of complexity in moderating the value of data: when dealing with interdependent components, data signals can be a misleading guide, raising the need for theoretical knowledge (Choudhury et al., 2020; Tranchero, 2023).

The paper proceeds as follows. Section 2 describes the construct of data-driven search and discusses its implications for discovering breakthrough innovation. Section 3 provides an overview of how scientists search for gene-disease associations and explains the characteristics of genome-wide association studies. Section 4 describes the data and research design, while Section 5 presents empirical estimates of the role of data in shaping search patterns and breakthrough innovations. Section 6 concludes.

2 Theoretical Framework

2.1 Data-Driven Search

Recent years have seen the emergence and diffusion of large-scale datasets that offer complete "maps" of technological spaces (Nagaraj and Stern, 2020). Researchers have started investigating their impact on corporate decision-making (Brynjolfsson and McElheran, 2016), organizational structure (Wu et al., 2019a), startup growth (Bessen et al., 2022), and scientific production (Nagaraj et al., 2020; Nagaraj and Tranchero, 2023). However, when focusing on innovation, the results are ambiguous. Reliance on data seems to harm the generation of breakthrough innovation (Deniz, 2020; Ghosh, 2021; Lou and Wu, 2021; Wu et al., 2020), and only a few firms seem able to benefit from data technologies (Brynjolfsson et al., 2021; Nagaraj, 2022). These mixed findings appear at odds with the hype surrounding big data and underscore the need for a better understanding of how data are actually used in technological innovation.

In this paper, I propose that the availability of large-scale data unlocks a new strategy to search for promising technological combinations. Instead of relying on knowledge from past attempts or theoretical priors, inventors can use data to extract signals of what combinations seem more fruitful for follow-on experimentation, a process that can be described as *data-driven search*. In practice, data can be used to triage and rank potential combinations in silico, thus focusing on the most promising ones without the need for costly experimentation (Christin, 2020; Hoelzemann et al., 2023). The power of this approach is that data can help assess a vast number of combinations that would be impossible to individually test or theoretically model (Choudhury et al., 2021; Evans and Rzhetsky, 2010; Shrestha et al., 2021). Moreover, innovators no longer need to pre-select a subset of components to recombine based on their prior knowledge, a choice that leads to path dependency in search (Denrell and March, 2001; Gavetti and Levinthal, 2000; Rzhetsky et al., 2015; Stuart and Podolny, 1996).

Data-driven search requires a few conditions to be feasible. First, the relevant characteristics of the technological components must be measurable, ensuring that the space of possible combinations is well-defined. This means that data-driven search might 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 a metric of technological potential on which the promise of each potential combination can be assessed. Such metric constitutes the objective function that data-driven search tries to maximize by finding the candidate combinations that score highest (Agrawal et al., 2022). Third, and relatedly, it must be possible to foresee the effect of novel combinations on the objective of interest. Said otherwise, 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 of data-driven search.

2.2 How Do Data Change Recombinant Search?

Comparing data-driven search with alternative approaches to uncover breakthroughs can help to discern how data reshape innovation. In particular, a rich strand of work suggests that innovators should leverage their theoretical priors to focus technological search on the portions of the combinatorial space expected to yield the highest returns (Csaszar and Levinthal, 2016; Gavetti and Levinthal, 2000; Nelson, 1982). The source of inventors' priors can range from scientific information (Fleming and Sorenson, 2004; Gambardella, 1995; Kneeland et al., 2020) to abstract knowledge of cause-effect relationships among components (Arora and Gambardella, 1994; Ehrig and Schmidt, 2022; Felin and Zenger, 2017). In practice, the gist of this literature is that inventors achieve superior performance by using their current beliefs to inform their recombination attempts, following a *theory-driven search* (Camuffo et al., 2022; Gittelman, 2016).

Gavetti and Levinthal (2000) note that search strategies can be characterized by three properties: how alternatives are evaluated, the extensiveness of other options considered, and how distant they are from the domain currently being searched. Starting with the mode of evaluation, both theory-driven and data-driven search operate *off-line*, meaning that they assess the merit of alternative combinations without the need to actually invest in experimentally validating them. However, they diverge in the scope of alternatives considered (Agrawal et al., 2022). Any theory, however accurate, can only account for some dimensions of the reality represented (Gavetti and Levinthal, 2000).

Reducing the dimensionality of search problems will necessarily come at the cost of narrowing search on some dimensions. In contrast, big data allow to obtain signals about a potentially much larger set of alternative combinations.

In turn, the crucial difference in the extensiveness of alternatives considered will likely shape the portion of the technological landscape explored. Theory-driven search exploits known cause-effect relationships to funnel experimentation efforts in areas where inventors theoretically understand how components could be recombined (Arora and Gambardella, 1994; Denrell and March, 2001). The result is usually a persistent search in the same narrow portion of the landscape, even in the face of decreasing returns (Fleming, 2001; Helfat, 1994). Instead, data-driven search breaks the path dependency that characterizes theory-driven search (Gavetti and Levinthal, 2000; Rzhetsky et al., 2015). Extracting signals from data should not be biased by beliefs or capabilities that steer innovators away from explorative attempts, unless the data themselves are fundamentally biased (Cao et al., 2021). Assuming that data provide a complete map of the landscape, innovators will be more likely to recombine components further away from domains previously explored (Lou and Wu, 2021; Wu et al., 2020). These arguments imply the following baseline hypothesis:

Hypothesis 1: *Data-driven search leads to greater search breadth than theory-driven search.*

2.3 The Consequences of Data-Driven and Theory-Driven Search

Does data-driven search enable the discovery of more breakthrough innovations relative to theory-driven search? Existing research does not offer a clear-cut answer. Scientific theories tend to focus efforts where returns are expected to be highest (Arora and Gambardella, 1994; Fleming and Sorenson, 2004). If the theories are correct and innovators are already searching in the most fruitful areas of the landscape, then increasing the breadth of search will likely not increase breakthroughs. However, there are reasons to believe that this might not be the case (Leiponen and Helfat, 2010). On the one hand, theory-driven search is very focused, possibly at the cost of neglecting areas that might harbor breakthroughs (Denrell and March, 2001; Gavetti and Levinthal, 2000; Langley et al., 1987). Novel discoveries often require going against established beliefs to recombine diverse knowledge (Audia and Goncalo, 2007; Katila and Ahuja, 2002). On the other hand, cumulative search can suffer from decreasing returns and eventually reach technological exhaustion (Fleming, 2001; March, 1991; Schilling and Green, 2011). Insofar as breakthrough innovations are the result of bridging distant technological domains, we should then expect data-driven search to yield more breakthroughs than theory-driven approaches:

Hypothesis 2: *Data-driven search is more likely to introduce breakthrough combinations than theory-driven search.*

In addition to the hypothesis above, one can further hypothesize that the effectiveness of either search strategy will be contingent on the area of the landscape explored. Theory-driven search should be especially effective in areas where inventors possess deep knowledge to build upon (Kaplan and Vakili, 2015; Kneeland et al., 2020). But, almost by definition, this approach becomes weaker when there is less knowledge to inform recombinant attempts (Fleming and Sorenson, 2004). Instead, data-driven search could be of greater help in such uncharted domains because data signals help triage recombinations even if the mechanisms of what makes a combination valuable are unknown (Anderson, 2008; Christin, 2020; Wu et al., 2020). As long as correlational signals are somewhat informative, inventors need not know *why* data indicate a certain area as promising to leverage that information (Nagaraj, 2022; Tranchero, 2023). This line of reasoning can be summarized in the following hypothesis:

Hypothesis 3: *Data-driven search is more effective to recombine unknown components than theory-driven search.*

However, research on recombinant search suggests that the complexity of the combinatorial task is a key moderator of search effectiveness (Aharonson and Schilling, 2016; Fleming and Sorenson, 2001). Recombining highly interdependent elements is difficult because slight modifications can result in major functional changes. In such cases, a deep understanding of an area can provide the necessary capabilities to unlock breakthroughs recombinations (Arts and Fleming, 2018; Nelson and Winter, 1982; Rzhetsky et al., 2015). Theories should be especially useful to foresee the value of combinations between interrelated elements and avoid fruitless avenues (Fleming and Sorenson, 2004). On the contrary, atheoretical exploration of complex domains could impede value identification since one needs sufficient knowledge to assess how data signals map into combinations of coupled elements (Kaplan and Vakili, 2015; Lou and Wu, 2021). Data signals might offer poor guidance in those instances since they can, at best, provide correlations that are unable to account for complex interactions between components (Deniz, 2020; Ghosh, 2021). Therefore, I hypothesize the following:

Hypothesis 4: *Data-driven search is less effective in recombining interdependent components than theory-driven search.*

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 organism's functioning. Knowing the genetic roots of diseases has significant practical consequences since genes causative of a disease can serve as drug targets to treat the condition (Nelson et al., 2015). Most common diseases such as diabetes, Alzheimer's, or hypertension are *polygenic*: they are not due to a single genetic factor but rather to multiple genes and their interaction with the environment during human life (Bush and Moore, 2012). Discovering the genes involved in each of the thousands of polygenic diseases requires searching through the space of $\sim 20,000$ known human genes. How do scientists look for new gene-disease combinations in this huge combinatorial space?

Scientists traditionally followed a *candidate gene approach* consisting of three main steps (Tabor et al., 2002). First, the scientist decides on the disease to study, likely motivated by its prevalence or funding availability. Second, she hypothesizes what genes might have a role in its etiology. Finally, she focuses the analysis on those genes, typically employing family linkage studies, case-control studies, or experiments with lab animals. Importantly, selecting the target genes reflects scientists' biological understanding of which genes might be important and why. For instance, after the IL12B gene was associated with psoriasis, some scientists relied on their knowledge of IL12B's role in the metabolic pathway of the IL23R gene to hypothesize a connection between IL23R and psoriasis. Indeed, this reasoning led to the candidate gene study that first documented the role of IL23R in psoriasis (Cargill et al., 2007), a finding that led to the development of several FDA-approved drugs.

Despite several successful discoveries, candidate gene studies lead scientists to consider only genes for which it is possible to formulate functional hypotheses (Haynes et al., 2018). As a result, there has been an extreme concentration of attention on a small number of theoretically well-known genes (Oprea et al., 2018; Stoeger et al., 2018). Gates et al. (2021) report that 1% of genes has received 22% of all gene-related publications. The emphasis on a handful of "superstar" genes might be due to the high risks involved in exploring the remaining gene pool, which is likely to harbor mostly dead ends. However, this situation is probably suboptimal since our chances of finding a cure for polygenic diseases would benefit from exploring a larger number of genes (Edwards et al., 2011; Stoeger et al., 2018). As a consequence, currently approved treatments exploit only around 10% of the potential drug targets highlighted by the Human Genome Project, 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 International HapMap Project in 2005. The HapMap was designed to supply a detailed reference genome that could be used to relate genetic mutations with phenotype changes (Bush and Moore, 2012). The second, and related, was the diffusion of commercial genotyping microarrays. Unlike whole genome sequencing, which processes every DNA basis, microarrays only collect data about specific genetic loci. The HapMap enabled to design microarrays targeting loci that can be extrapolated to capture the characteristics of their genetic surroundings, thus allowing to parsimoniously infer the characteristics of most of the genome (Bush and Moore, 2012). The result was a steep decrease in the cost of collecting data on genomes, prompting the emergence of *genome-wide association studies* (Visscher et al., 2017).

Genome-wide association studies (or GWASs) are case-control studies where researchers genotype a large number of genomes and look to see if any genetic mutation is more likely to appear in subjects showing a specific condition rather than in the control group (Pearson and Manolio, 2008; Uffelmann et al., 2021). Figure 1 schematically depicts how a typical GWAS unfolds. Researchers start by collecting DNA samples from both cases and controls. All DNA samples are genotyped using DNA microarrays and imputed through reference genomes to reconstruct complete genotypes. Finally, researchers test for statistically significant differences between the genotypes of cases and controls. The genes harboring variants associated with a disease can be suspected to play a role in its etiology, hence being potential targets for pharmaceutical intervention.

For instance, consider once more the genome-wide association study by Hakonarson et al. (2007). These researchers performed their analysis on a study population of 563 patients with type 1 diabetes and 1,146 healthy controls. Using genotyping microarrays, Hakonarson and colleagues reconstructed the genotype of all their subjects. A comparison between cases and controls highlighted several significant differences across the entire genome, some pertaining to genes already known to be related to diabetes (e.g., the insulin gene INS), but some located in the gene KIAA0350. This gene is responsible for coding a protein whose function was not known at the time, which explains why it had been previously neglected by candidate gene studies. The KIAA0350-type I diabetes association proved very impactful and has been further investigated by several studies and clinical trials. Appendix B presents additional details on this case study.

The example above clarifies that unlike candidate gene studies, where researchers decide which subset of genes to target, genome-wide association studies look for genetic variants across the whole genome (Visscher et al., 2017; Uffelmann et al., 2021). Conditional on the choice of disease, a genome-wide search permits scanning the entire set of possible combinations, pointing directly to the most promising genes (Panel (a) of Figure 2). In practice, this search strategy removes one degree of freedom from the researcher, who is no longer required to specify genetic targets ex ante. This ensures that GWASs are not biased by prior biological knowledge and beliefs, thus avoiding researchers' tendency to focus on familiar genes. Genome-wide association studies generate discoveries thanks to what directly emerges from the data, making them a prime example of data-driven search (Evans and Rzhetsky, 2010).

Genome-wide association studies have been harshly criticized for their shortcomings. On the one hand, these studies are inherently correlational, which means that any finding could be a false positive (Marigorta et al., 2018). On the other hand, even when the associations discovered by GWASs are robust, scholars have suggested that this approach neglects 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 in them could be quite small (Goldstein et al., 2009). As a result, besides a few success stories, many gene-disease associations uncovered by GWAS have not stood up to the expectations (Boyle et al., 2017; Visscher et al., 2017). These criticisms explain why candidate gene approaches remain popular among researchers, but it must be noted that the debate on whether GWASs discover scientifically impactful gene-disease associations is still unsettled.

4 Data and Methods

4.1 Research Design

The goal of this paper is to understand the potential channels through which data shape recombinant search. Accordingly, the analysis does not provide causal evidence on the effects of adopting either search process. Instead, I adopt a comparative approach and descriptively explore how gene-disease combinations introduced by a data-driven process differ from those introduced through a more theoretical process. I use regression analysis to juxtapose the outcomes of data-driven and theory-driven search after taking into account the characteristics of genes, diseases, and scientists. While my results are not causal in nature, I leverage rich and detailed data to rule out several obvious confounders and isolate plausible mechanisms.

More specifically, I compare the properties of gene-disease associations introduced by genome-wide association studies with gene-disease associations that involve the same disease but are uncovered by candidate gene studies. While the empirics are at the discovery level, I use information about the paper where the gene-disease association first appeared to take into account a number of factors. First, I include year dummies to limit comparisons between discoveries made in the same time period. Second, I control for the prestige of the journal where the study was published since it might influence the visibility and diffusion of its results. Finally, in light of evidence that larger research teams seem to be less disruptive (Wu et al., 2019b), I also control for the number of co-authors listed on the paper introducing the gene-disease associations.

An additional concern could be that scientists carrying out GWASs are systematically different from those who do not, potentially confounding my estimates. Said otherwise, I need to ensure that the search outcomes are the reflection of the process used and not of the scientists' ability or preferences. One way to do so is by limiting the comparison between gene-disease associations introduced by the same researcher, exploiting the fact that many scientists use both approaches during the course of their careers. The addition of principal investigator fixed effects ensures that the estimates reflect the change in outcomes due to a change in search methods, holding constant all time-invariant characteristics of the scientist.

Finally, when comparing the scientific importance of discoveries resulting from data-driven and theory-driven search, one might fear that genes differ in terms of discovery opportunities. For instance, the gene TP53 is crucial to regulate cell growth, and mutations of this gene are found in over half of tumor sequences (Gates et al., 2021). Given its crucial role in one of the most important cellular processes, it is reasonable to expect that gene-disease associations involving this gene are more likely to be scientifically important. More in general, if the genes recombined by GWASs are intrinsically more likely to yield high-value associations, then an increase in breakthrough innovation might mechanically derive from that. To control for this possibility, I also estimate models that include gene fixed effects when assessing the scientific quality of discoveries made by genome-wide association studies.

4.2 Defining the Search Landscape: Gene-Disease Associations

I construct a dataset of all novel gene-disease associations (GDAs) introduced from 1980 to 2016 inclusive. I retrieve such information from DisGeNET (v7.0), an aggregator considered a 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 specialized sources, including curated datasets and publications indexed in PubMed. My data are at the GDA level, and for each association, I retrieve both the publication that introduced it and the list of all follow-up articles that investigated it. I focus on associations mapping a protein-coding gene to a disease, syndrome, or abnormality with clear health implications. My final dataset includes 352,162 gene-disease associations between 14,072 genes[2] and 9,740 narrow disease categories.[3]

To identify which of these 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 a comprehensive list of genome-wide association studies published in peer-reviewed journals. Studies are eligible for inclusion in the GWAS Catalog if they use a DNA microarray to scan the entire genome without targeting 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). In total, I identify in my data 8,440 GDAs that were introduced by 1,216 distinct genome-wide association studies, while the remainder of my sample was the result of theoretically-driven search processes. Panel (b) of Figure 2 shows the

[2] The number of protein-coding genes in my data is lower than the total number of human genes because some of them have never been implicated in a disease.

[3] Scientists routinely complain that associations proposed in academic publications often turn out not to be robust (Tabor et al., 2002). Therefore, there is the risk of considering scientifically important associations that receive attention just because of criticisms and not because of their value. I address this concern using the DisGeNET-provided *Evidence Index* to retain in my data only associations for which contradictory results represent less than 10% of the available publications about them. However, all my results are robust to either stricter thresholds of the Evidence Index or to keeping the whole DisGeNET data. See Appendix Figure C.4 for robustness checks.

rapid growth of GWASs since 2005, when the first such study was published.

Information on the bibliographic characteristics of papers introducing at least one novel gene-disease combination is taken from NIH's iCite data. Specifically, I record the number of authors of each article, the journal, and the number of citations received. I use the 2020 SCImago journal ranking to measure the relative prestige of the publication venue of each paper. To identify the principal investigator (PI), I extract information on the last author of each publication from the Author-ity database (Torvik and Smalheiser, 2021).[4] Author-ity is a highly accurate database that disambiguates the authors of PubMed papers leveraging information on names, coauthors, MeSH codes, affiliations, and paper keywords. I use these data to estimate fixed effect models that estimate the effect of carrying out a GWAS controlling for time-invariant characteristics of the principal investigator.

I also gather additional gene-level attributes to test for mechanisms. For each gene, I record if it is part of a gene family. Genes in a family are formed by duplication of a single ancestral gene and generally share similar biochemical properties (Daugherty et al., 2012). Such genes have the same name followed by a number reflecting the order in which they were discovered (e.g., BRCA1 and BRCA2, discovered in 1994 and 1995, respectively). Any discrepancy in attention between members of the same gene family is likely to reflect path dependence in studying the genes discovered earlier in time instead of differences in scientific potential (Stoeger et al., 2018). Confirming this type of bias, DisGeNET data show that the first member of a gene family receives, on average, 56% more publications than the second member of the same family.

Next, I code which human genes have a homolog gene in the lab mouse (Clarke, 2002). Homologs are genes inherited in two species from a common ancestor, thus retaining comparable functions and biology. This property allows scientists to carry out experiments on homolog genes in animals to learn about human biology. 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 less convenient to study experimentally and thus often neglected for reasons not related to their importance (Baba and Walsh, 2010; Stoeger et al., 2018). Finally, I code which genes regulate the function of more than one downstream gene (Türei et al., 2016). Studying the role of those genes in disease requires learning how they influence several interdependent biological pathways, thus offering a potentially more complex route to therapeutic development (Hermosilla and Lemus, 2019).

[4] Authorship norms in the life sciences prescribe that the principal investigator is placed in last position on the authorship roster of a paper. The focus on principal investigators is justified by the fact that they have agency in directing the methodological choice of each study.

4.3 Outcome Variables

My objective is to compare the characteristics of GWAS-established gene-disease associations vis-á-vis associations established with a candidate gene approach. To do so, I ask: how do gene-disease associations introduced by a data-driven process differ from those introduced through a more theoretical process? I address this question focusing on two outcomes:

Underexplored Gene: I use two alternative dependent variable to capture discoveries involving genes that received scant attention before the emergence of GWASs. The first proxy is a dummy that takes value one for gene-disease associations that include a gene never associated with a disease before 2005, the year of the first GWAS. Going back to the example of the introduction, the KIAA0350 gene is coded as underexplored because it had not been associated with any disease before the paper of Hakonarson et al. (2007). The second proxy is the gene's discovery date, 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 relatively more likely to introduce combinations involving genes discovered after the year 2000 (when the partial draft of the human genome was released). For instance, the KIAA0350 gene was discovered in 2002, but ignored until being linked to type 1 diabetes. Both dependent variables are coded as dummies to allow a straightforward interpretation of the OLS coefficients as linear probability models, but I repeat the same analyses using continuous versions of these variables in the Appendix.

Scientific Importance: The second dependent variable is a dummy that takes value one for gene-disease associations of large scientific importance. Usually, researchers rely on paper-to-paper citation counts to measure impact, but this would be inappropriate in my setting since scientific articles often study more than one gene-disease association. For instance, the GWAS by Hakonarson et al. (2007) reports the association between type 1 diabetes and three genes (KIAA0350, INS, and PTPN22), preventing a straightforward way to assign to each GDA its share of the citations received by the article introducing it.[5] Instead, I exploit DisGeNET to construct a novel measure of GDA scientific importance: the number of papers that *directly* build on the gene-disease combination. These include empirical and experimental work that investigates the proposed association, regardless of whether they cite the paper that first introduced it, hence being a direct measure of impact for each individual GDA. For each year, I code all new GDAs in the 95th percentile of follow-on work received as breakthrough discoveries.[6] As an example, the KIAA0350-type 1 diabetes association is coded as a breakthrough because it was later studied by

[5] Moreover, citation counts would be misleading in this context since GWASs are highly cited on average (μ_{GWAS}=169 vs. $\mu_{Candidate\,Gene}$=41) 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.

[6] This choice follows the approach of papers on breakthrough innovation that operationalize outlier performance as falling within the top 5% of the sample (Arts and Fleming, 2018; Kaplan and Vakili, 2015). However, my results are unchanged if I use alternative cut-offs (Appendix Table C.2) or directly the count of follow-on papers as a dependent variable (Appendix Table C.3).

more papers than 95% of the other discoveries made in the same year.

4.4 Summary Statistics

Table 1 lists the key variables together with the summary statistics for the sample used in the analysis. Panel A provides summary statistics about the publications that established new GDAs in the period considered. Besides being more cited than candidate gene papers, on average GWASs introduce more associations spanning a larger number of genes. Panel B provides summary statistics at the GDA level. Previewing the following analysis, the incidence of high-impact associations is higher for GWASs (8.3%) than for candidate gene papers (4.9%). It also appears that genes associated with a disease by GWASs are more likely to be less studied, to lack a mouse homolog, and be the second member of gene families. The share of breakthrough combinations that include an understudied gene is 13.8% among those established by candidate gene studies but grows to 37.6% among those found by genome-wide association studies.

5 Results

5.1 Data-Driven Search and the Direction of Innovation

In this section, I estimate the following linear probability model 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 and principal investigator fixed effect, as well as controls for the 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 for GDAs introduced by candidate gene studies. This specification estimates the difference between GDAs that first appeared in a genome-wide association study and GDAs that were introduced by candidate gene papers. If data-driven search leads to diversify search, then I should find that the OLS estimate β is positive and statistically significant.[7] All 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 implicate understudied genes with human diseases than candidate gene studies. Specifically, the estimate of β in Column 1 indicates an average increase of 20 percentage points on the probability of combining a gene never associated with a disease before 2005, a large increase given that the baseline is about 13 percentage points. Column 3 shows that this finding is robust to using the date of discovery as an alternative proxy for genes that have been historically less studied. However, if the characteristics of researchers that publish

[7] The choice of using a linear probability model is motivated by the large number of fixed effects employed, which prevents convergence in nonlinear models.

GWAS correlate with the likelihood of exploring less studied genes, then these results could reflect an upward bias. I account for this possibility by adding PI fixed effects that absorb time-invariant researcher attributes: Columns 2 and 4 of Table 2 show that principal investigators that carry out a GWAS are 86-114% more likely to recombine understudied genes relative to when they adopt a candidate gene approach.[8] These results confirm the baseline hypothesis that data-driven search increases search breadth.

Figure 3 presents an intuitive visualization of the combinatorial space of pairwise gene-disease combinations. Comparing the areas searched by candidate gene 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 theory-driven 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 historically well-studied diseases (see also Figure C.1 in the Appendix). This result shows how data-collection decisions remain crucial in determining the direction of search, but also shows that the diversification in gene space is due to the search strategy itself and not to a change in disease focus.

Note that focusing on a narrow subset of genes is not necessarily a problem if they are chosen because of their higher scientific promise. To explore this possibility, I consider new associations recombining genes that are part of a gene family. Since genes in the same family carry out similar functions, the main difference between the first and second family members is often just the order in which they were discovered and not their importance (Stoeger et al., 2018). Figure 4 shows that candidate gene papers tend to study the first gene of a family much more than the second member. Such discrepancy reflects how scientists continue targeting the gene discovered earlier, even holding therapeutic potential constant (see also Appendix Table C.4). However, the difference between the first and second genes in a family completely disappears for GDAs introduced by genome-wide association studies. This evidence suggests that one of the mechanisms through which GWASs help discovery is counteracting inertial forces in scientists' research paths (Denrell and March, 2001; Rzhetsky et al., 2015), which are especially damaging to our understanding of polygenic diseases (Haynes et al., 2018).

5.2 Does Data-Driven Search Lead to Better Innovations?

While data-driven search broadens the scope of search, nothing ensures that the new discoveries are scientifically interesting. Indeed, one might suspect that scientists are already exploring the

[8] Appendix Table C.1 reports similar results when considering continuous versions of the dependent variables instead of discrete codings. GWASs are more likely to recombine genes that received 31-34% fewer publications before 2005 and that were discovered 1.6-2.2 years later.

most fruitful areas of the technological landscape, reducing the chances that breakthrough findings have been missed. In what follows, I tackle this question by 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 scientifically important combinations discovered in a given year, which is the common definition for breakthrough innovation. All other variables and controls are identical to the previous section's specification.

Table 3 presents the results. In the basic specification reported in Column 1, new gene-disease associations that first appeared in a genome-wide association study are, on average, 24% more likely to be among those of high scientific impact. But what are the drivers of this result? Part of it is certainly due to the exploration of new high-value genes that were previously ignored, but it can also be that researchers are neglecting the study of well-known genes in relation to new diseases. To investigate this possibility, I estimate the same model by adding gene fixed effects, hence absorbing the cross-sectional variation linked to genes' scientific potential. Column 2 of Table 3 shows that the estimate's magnitude and statistical significance substantially grow when considering only variation within genes. This suggests that data-driven search permits to find combinations of high value that were ignored even in the context of well-known genes. Results are also robust to the inclusion of PI fixed effects (Column 3 of Table 3), further confirming Hypothesis 2.

In the Appendix, I further explore the robustness of my findings. Table C.2 shows that results do not change if I adopt alternative ways to define breakthrough discoveries, such as gene-disease combinations in the top 10% or top 1% of scientific importance. If anything, the magnitude of the coefficient grows larger for more stringent definitions of breakthrough, suggesting that GWASs are more powerful in uncovering outlier combinations. I also show that the results are unchanged if the dependent variable is defined as the count of follow-on papers received by each combination instead of a dichotomous definition of breakthroughs (Table C.3). Finally, I find the same results if I limit my analysis to GDAs involving members of gene families (Table C.4). In sum, my analysis documents that gene-disease associations introduced by GWASs have, on average, a higher scientific impact than comparable associations discovered with candidate gene approaches, even when holding the intrinsic characteristics of the genes and the attributes of the principal investigator constant.

5.3 Data, Theory, and Landscape Characteristics

The results reported in the previous sections capture the average effect of GWAS on discovery, but this could conceal large heterogeneity in the effectiveness of data-driven search. In particular, one might expect that theory-driven search becomes more effective in areas where scientists have more profound biological knowledge. In those domains, theoretical knowledge should lead researchers directly to the best technological combinations and outperform data-driven search (Arora and Gambardella, 1994; Fleming and Sorenson, 2004). Figure 5 explores this idea by plotting the likelihood of introducing high-impact GDAs in correspondence with each gene, separately by the search strategy that introduced them. Genes on the X-axis are sorted by the number of pre-2005 publications received, which can serve as a proxy for the depth of biological knowledge available. 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.[9]

A potential drawback of this analysis is that the distribution of theoretical expertise on the landscape is endogenous (Gittelman, 2016; Nelson, 2003). Genes more heavily studied are likely to be those considered most relevant by scientists, hence this might explain why theory-driven search performs better in that case. To allay this concern, I exploit the fact that some genes have historically been neglected because they cannot be studied with the lab mouse (Stoeger et al., 2018). This simple intuition is borne by my data, because I find that genes without a counterpart in the mouse genome have received, on average, 22% fewer publications on how they might be implicated in human diseases. Since these genes are less theoretically characterized due only to experimental convenience, this provides variation in theory availability that does not reflect their therapeutic potential. Figure 6 shows that, unlike GWAS, candidate gene studies are markedly less successful when exploring genes less theoretically known because they lack a mouse homolog. Regression coefficients reported in Table 4 confirm this pattern and provide support for Hypothesis 3.

The analyses reported above established that theory-driven search is more effective in areas where scientists have a deeper knowledge, while data-driven search is powerful in recombining little-known components. But what is the precise mechanism behind this finding? Hypothesis 4 suggests that the density of interrelationships between components moderates the relative effectiveness of the two search strategies (Gavetti and Levinthal, 2000; Fleming and Sorenson, 2001). I explore this possibility by estimating the efficacy of genome-wide association studies when they involve genes that regulate the function of more than one other gene. Intuitively, when a gene governs the function of many downstream genes, it becomes complex to figure out the precise mechanism through which it is related to a disease (Hermosilla and Lemus, 2019). The results in Table 5 confirm that the relative advantage of candidate gene studies in the context of highly-studied genes is largely due to their ability to recombine interdependent genes. This suggests a key limitation of data-driven search: data signals report correlations that are unable to discern complex interactions between coupled components (Fleming and Sorenson, 2004; Ghosh, 2021). Instead, in the presence of a

[9] The same pattern emerges if I use the date of discovery of the gene as a proxy for the availability of biological theory, see Appendix Figure C.2.

solid-enough theoretical understanding, researchers can locate the best gene-disease combinations because they can account for how genes interact (Boyle et al., 2017).

5.4 Robustness Checks

The findings of the previous sections suggest that it is *how* discoveries are made that drives the characteristics of gene-disease combinations. However, due to the descriptive nature of this paper, it could still be that other channels account for part of the results. My goal in this section is to investigate these alternative channels and see to what extent they might offer alternative explanations for my findings.

A. Characteristics of Scientists Publishing GWASs: The inclusion of principal investigator fixed effects allows estimating within researcher coefficients, capturing the change in outcomes experienced by the same researcher after switching search strategies. However, it could still be that researchers who are systematically more explorative or able to identify breakthroughs sort into adopting GWASs, leading to a potential upward bias in the estimates. Appendix Figure C.3 shows that before their first GWAS, principal investigators that will eventually adopt a genome-wide search are not more likely to target less studied genes or to introduce breakthrough gene-disease combinations. This robustness test ameliorates concerns that researchers sort into carrying out a GWAS based on endogenous characteristics related to my outcomes of interest.

B. Disease Selection Over Time: One additional concern is that my analyses could overstate the benefits of adopting GWAS if the search method is applied first to diseases where it is more likely to yield discoveries involving unknown genes. Note that Figures 3 and C.1 already show that GWASs target the same highly-studied diseases as the majority of candidate gene studies, and the inclusion of disease fixed effects should reduce the practical relevance of this issue. Appendix Table C.5 directly tests this idea, showing that diseases that receive their first GWAS earlier in my sample period are not more likely to be associated with short-changed genes. Looking at the list of diseases that received the most GWASs, one finds polygenic diseases with a large incidence in the population, suggesting that this was the primary criterion guiding disease choice.

C. Alternative Samples: Gene-disease associations require extensive follow-on work to be validated, and contradictory results are not infrequent (Uffelmann et al., 2021). The sample used in this paper considered associations for which DisGeNET reports less than 10% of contradictory papers about them to avoid confounding effects from false positive discoveries. In Figure C.4, 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 quantitatively similar regardless of the sample chosen, suggesting that the choice of sample is not driving my findings.

D. Different Measure of Scientific Value: Instead of relying on the number of subsequent publications to measure associations' scientific potential, I perform a robustness check using DisGeNET's *GDA Score* (Piñero et al., 2020). The GDA Score synthetically captures the scientific reliability of all the existing evidence on the gene-disease association. Table C.6 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.

E. Therapeutic Value of GWAS Findings: The analyses reported so far showed that GWAS uncover GDAs with a higher scientific value on average, but this does not guarantee that those gene-disease pairs will also prove more therapeutically useful (Gittelman and Kogut, 2003). To capture downstream medical translation, I code a dummy that captures whether papers introducing new gene-disease combinations are later cited by articles reporting clinical trials. Appendix Table C.7 reports that genome-wide association studies are more likely to be cited by clinical trial articles than comparable candidate gene studies. This finding reassures that findings scientifically impactful are also those with larger value to developing treatments.

6 Conclusion

This paper explores how a data-driven search strategy changes innovation outcomes. Unlike theory-driven search, I argue that data enable global search strategies that lead to more exploratory recombinations of technological elements. Empirical results in the context of genome-wide association studies confirm this idea, showing that path dependency does not tether data-driven search. Data lead innovators to experiment with short-changed areas of the technological landscapes and help them uncover combinations of higher average value. The latter result is stronger in uncharted landscape areas, but theory-driven approaches are more effective when deeper theoretical knowledge can be used to guide search. The key mechanism underlying this finding is the ability of theory to account for complex interdependencies among technological elements. Instead, data signals often report misleading correlations when dealing with highly coupled components.

These findings have practical implications for the innovation search strategy adopted by scientists, managers, and governments. For individual researchers, my results underscore when alternative approaches to search are more or less effective, suggesting to rely on data analytics when venturing into uncharted domains. More in general, data analytics is diffusing in every sector of the economy, but the returns remain heterogeneous and concentrated among few companies (Brynjolfsson et al., 2021). This paper highlights that managers should collect and rely on data to guide risky experimentation but also trust their knowledge base when working in well-trodden domains. Moreover, decision-makers must be wary that data signals could be misleading when dealing with interdependent components (Ghosh, 2021; Tranchero, 2023). My results also provide an additional rationale for furthering investments in large-scale public data sources that might enable data-driven search (Nagaraj, 2022; Kao, 2022).

It is also important to note that despite its considerable potential, data-driven search happens mainly within the boundaries of the prevailing technological and scientific paradigms (Dosi, 1982; Kuhn, 1962). In turn, this has three main implications. First, the paradigm can help determine the metric used to assess the value of alternative technological combinations from data signals. This means that meta-theoretical understandings of "what is important" will concur in shaping the ability to locate breakthrough innovations. Second, the decision of where to direct data collection efforts might itself reflect the priorities implicit in the current paradigm. Insofar as this choice is biased, it might end up reinforcing previously established research patterns (Cao et al., 2021; Hoelzemann et al., 2023). However, the increasing diffusion of complete maps of diverse technological landscapes, such as the Sloan Digital Sky Survey or the Human Genome Project, should reduce the practical relevance of this concern (Nagaraj and Stern, 2020). Third, data might 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 might be subject to technological exhaustion unless new elements are added over time (Fleming, 2001).

My work makes several contributions. First, I provide a theoretical and empirical framework to explore data-driven innovation. Data can drastically change the "technology of technical change" (Arora and Gambardella, 1994; Agrawal et al., 2022; Cockburn et al., 2019), but to date, there is little understanding of how data analytics is reshaping the generation of novelty (Allen, 2022; Bessen et al., 2022; Cao et al., 2021; Hoelzemann et al., 2023; Nagaraj, 2022; Wu et al., 2020). Addressing this gap is of first-order importance in the age of big data, and my work is a first step in this direction. Second, the construct of data-driven search constitutes an addition to the theory of recombinant search (March, 1991; Fleming and Sorenson, 2001; Gavetti and Levinthal, 2000; Katila and Ahuja, 2002; Levinthal, 1997; Nelson and Winter, 1982). By framing data as reshaping recombinant search, I leverage this important body of research to clarify the conditions under which theory-driven search will outperform data-driven search. Moreover, my study is one of the first to empirically measure how alternative search strategies map into discoveries (Kneeland et al., 2020; Maggitti et al., 2013). Finally, I provide new empirical evidence on how data reshape genomics and lead to discoveries relevant to pharmaceutical innovation (Cockburn, 2006; Gambardella, 1995; Hermosilla and Lemus, 2019; Kao, 2022; Krieger, 2021). My results are timely also given the lack of consensus about the scientific value of genome-wide association studies (Boyle et al., 2017; Pearson and Manolio, 2008; Visscher et al., 2017).

Finally, a few limitations of this paper must be acknowledged. First, the present study allows for meaningful comparisons of the consequences of engaging in theory-driven versus data-driven search, but the research design does not fully account for the endogenous choice of the form of search. Second, the patterns documented are from the quantitative case study of a single domain. 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 are 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 downstream investments in new drugs. This is an exciting avenue for follow-up work that is outside the scope of this paper.

References

Agrawal, A., J. McHale, and A. Oettl (2022): "Superhuman science: How artificial intelligence may impact innovation," *Brookings Working Paper*.

Aharonson, B. S. and M. A. Schilling (2016): "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution," *Research Policy*, 45, 81–96.

Allen, R. (2022): "Methodological pluralism and innovation in data-driven organizational cultures," *Harvard Business School*.

Anderson, C. (2008): "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine*, 16, 16–07.

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.

Audia, P. G. and J. A. Goncalo (2007): "Past success and creativity over time: A study of inventors in the hard disk drive industry," *Management Science*, 53, 1–15.

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., S. M. Impink, L. Reichensperger, and R. Seamans (2022): "The role of data for AI startup growth," *Research Policy*, 51, 104513.

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.

Camuffo, A., A. Gambardella, F. Maccheroni, M. Marinacci, and A. Pignataro (2022): "Microfoundations of low-frequency high-impact decisions," *CEPR Discussion Paper No. DP17392*.

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.

Choudhury, P., R. T. Allen, and M. G. Endres (2021): "Machine learning for pattern discovery in management research," *Strategic Management Journal*, 42, 30–57.

Choudhury, P., E. Starr, and R. Agarwal (2020): "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 41, 1381–1411.

Christin, A. (2020): "What data can do: A typology of mechanisms," *International Journal of Communication*, 14, 20.

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

Cockburn, I. M. (2006): "Is the pharmaceutical industry in a productivity crisis?" *Innovation Policy and the Economy*, 7, 1–32.

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

Csaszar, F. A. and D. A. Levinthal (2016): "Mental representation and the discovery of new strategies," *Strategic Management Journal*, 37, 2031–2049.

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*.

Denrell, J. and J. G. March (2001): "Adaptation as information restriction: The hot stove effect," *Organization Science*, 12, 523–538.

DiMeglio, L. A., C. Evans-Molina, and R. A. Oram (2018): "Type 1 diabetes," *The Lancet*, 391, 2449–2462.

Dosi, G. (1982): "Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change," *Research Policy*, 11, 147–162.

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.

Ehrig, T. and J. Schmidt (2022): "Theory-based learning and experimentation: How strategists can systematically generate knowledge at the edge between the known and the unknown," *Strategic Management Journal*.

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

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.

Gambardella, A. (1995): *Science and innovation: The US pharmaceutical industry during the 1980s*, Cambridge University Press.

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.

Gingerich, M. A., V. Sidarala, and S. A. Soleimanpour (2020): "Clarifying the function of genes at the chromosome 16p13 locus in type 1 diabetes: CLEC16A and DEXI," *Genes & Immunity*, 21, 79–82.

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

Gittelman, M. and B. Kogut (2003): "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns," *Management Science*, 49, 366–382.

Goldstein, D. B. et al. (2009): "Common genetic variation and human traits," *New England Journal of Medicine*, 360, 1696.

Hakonarson, H., S. F. Grant, J. P. Bradfield, L. Marchand, C. E. Kim, J. T. Glessner, R. Grabs, et al. (2007): "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene," *Nature*, 448, 591–594.

Haynes, W. A., A. Tomczak, and P. Khatri (2018): "Gene annotation bias impedes biomedical research," *Scientific Reports*, 8, 1–7.

Helfat, C. E. (1994): "Evolutionary trajectories in petroleum firm R&D," *Management Science*, 40, 1720–1747.

Hermosilla, M. and J. Lemus (2019): "Therapeutic translation of genomic science," in *Economic dimensions of personalized and precision medicine*, University of Chicago Press.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2023): "The streetlight effect in data-driven exploration," *UC Berkeley and University of Vienna*.

Jayaraj, S. and M. Gittelman (2018): "Scientific maps and innovation: Impact of the Human Genome on drug discovery," *DRUID Society Conference Paper*, 1–56.

Kao, J. (2022): "Charted territory: Evidence from mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

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

Katila, R. and G. Ahuja (2002): "Something old, something new: A longitudinal study of search behavior and new product introduction," *Academy of Management Journal*, 45, 1183–1194.

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*, Forthcoming.

Krieger, J. L. (2021): "Trials and terminations: Learning from competitors' R&D failures," *Management Science*, 67, 5525–5548.

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

Langley, P., H. A. Simon, G. L. Bradshaw, and J. M. Zytkow (1987): *Scientific discovery: Computational explorations of the creative processes*, MIT press.

Le Fanu, J. (2011): *The rise and fall of modern medicine*, Hachette.

Leiponen, A. and C. E. Helfat (2010): "Innovation objectives, knowledge sources, and the benefits of breadth," *Strategic Management Journal*, 31, 224–236.

Levinthal, D. A. (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

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, et al. (2017): "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," *Nucleic Acids Research*, 45, D896–D901.

Maggitti, P. G., K. G. Smith, and R. Katila (2013): "The complex search process of invention," *Research Policy*, 42, 90–100.

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.

Nagaraj, A. (2022): "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry," *Management Science*, 68, 564–582.

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.

NAGARAJ, A. AND S. STERN (2020): "The economics of maps," *Journal of Economic Perspectives*, 34, 196–221.

NAGARAJ, A. AND M. TRANCHERO (2023): "How Does Data Access Shape Science? Evidence from the Impact of U.S. Census's Research Data Centers on Economics Research," *UC Berkeley*.

NELSON, M. R., H. TIPNEY, J. L. PAINTER, J. SHEN, P. NICOLETTI, Y. SHEN, A. FLORATOS, P. C. SHAM, 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.

——— (2003): "On the uneven evolution of human know-how," *Research Policy*, 32, 909–922.

NELSON, R. R. AND S. WINTER (1982): *An evolutionary theory of economic change*, Belknap Press.

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," *Journal of the American Medical Association*, 299, 1335–1344.

PIÑERO, J., J. M. RAMÍREZ-ANGUITA, J. SAÜCH-PITARCH, RONZANO, ET AL. (2020): "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, 48, D845–D855.

QU, H.-Q., L. MARCHAND, R. GRABS, AND C. POLYCHRONAKOS (2007): "The IRF5 polymorphism in type 1 diabetes," *Journal of Medical Genetics*, 44, 670–672.

REICH, D. E. AND E. S. LANDER (2001): "On the allelic spectrum of human disease," *TRENDS in Genetics*, 17, 502–510.

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.

SHRESTHA, Y. R., V. F. HE, P. PURANAM, AND G. VON KROGH (2021): "Algorithm supported induction for building theory: How can we use prediction models to theorize?" *Organization Science*, 32, 856–880.

SOLEIMANPOUR, S. A., A. GUPTA, M. BAKAY, ET AL. (2014): "The diabetes susceptibility gene Clec16a regulates mitophagy," *Cell*, 157, 1577–1590.

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.

THOMKE, S. H. (2020): *Experimentation works: The surprising power of business experiments*, Harvard Business Press.

TORVIK, V. I. AND N. R. SMALHEISER (2021): "Author-ity 2018 - PubMed author name disambiguated dataset," *University of Illinois at Urbana-Champaign*.

TRANCHERO, M. (2023): "Finding diamonds in the rough: Data-driven decisions and pharmaceutical innovation," *UC Berkeley*.

TÜREI, D., T. KORCSMÁROS, AND J. SAEZ-RODRIGUEZ (2016): "OmniPath: Guidelines and gateway for literature-curated signaling pathway resources," *Nature Methods*, 13, 966–967.

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.

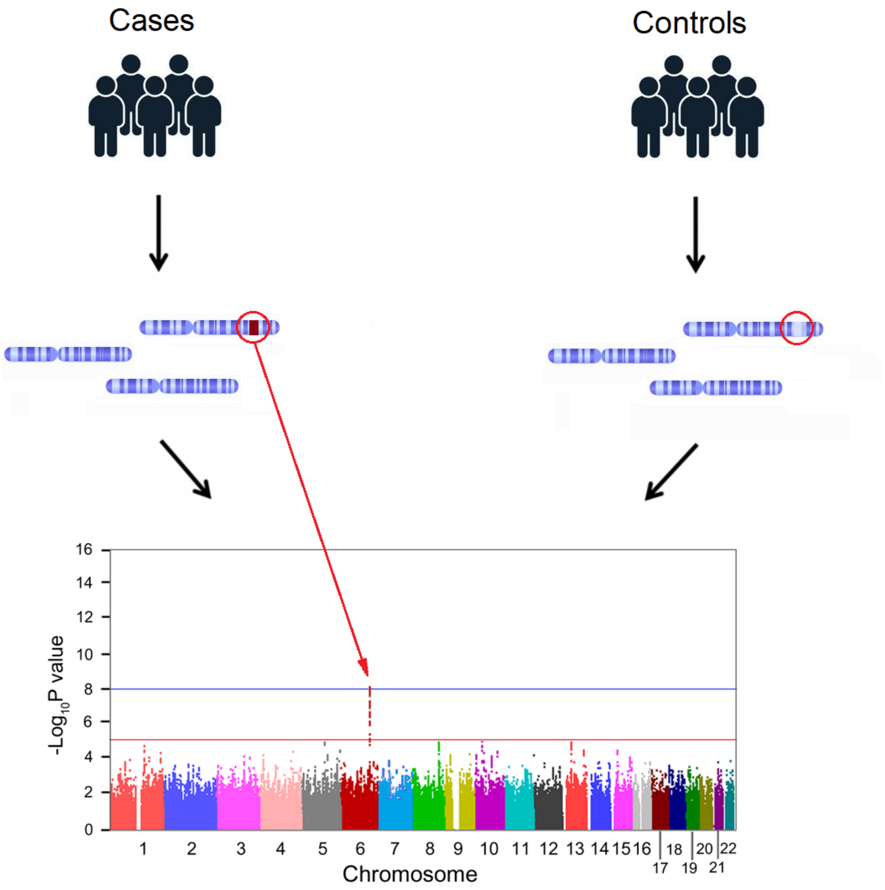
Wu, L., L. Hitt, and B. Lou (2020): "Data analytics, innovation, and firm productivity," *Management Science*, 66, 2017–2039.

Wu, L., B. Lou, and L. Hitt (2019a): "Data analytics supports decentralized innovation," *Management Science*, 65, 4863–4877.

Wu, L., D. Wang, and J. A. Evans (2019b): "Large teams develop and small teams disrupt science and technology," *Nature*, 566, 378–382.

7 Figures and Tables

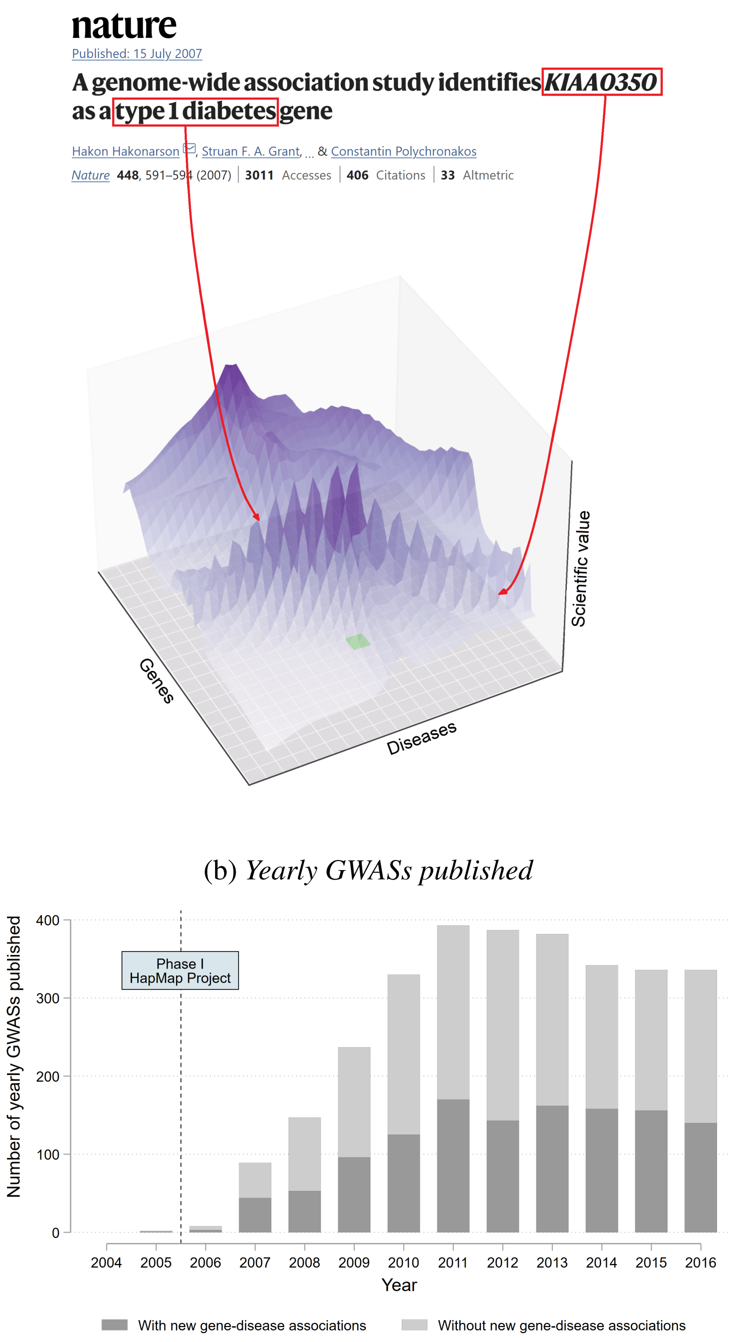
Figure 1: Schema of how a typical genome-wide association study unfolds.



Note: The figure depicts the main steps involved in a genome-wide association study. First, the researchers select the disease of interest and assemble a group of cases (subjects showing the condition) and one of controls (healthy subjects). Then, the genome of people with and without the condition are sequenced in search of significant genetic differences. Finally, statistical methods are used to test the association between any genetic variant and the disease of interest. The panel at the bottom is the characteristic "Manhattan plot" which indicates the location of the statistically significant genetic variants in the chromosome. On the Y axis there is the strength of the finding expressed as -$\log_{10}$(p-value), hence higher values correspond to stronger associations. For instance, the picture depicts a genetic variant in chromosome 6 that is significantly associated with the condition.

Figure 2: The emergence of genome-wide association studies in the search for useful gene-disease combinations.

(a) *GWAS as data-driven search*



nature

Published: 15 July 2007

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

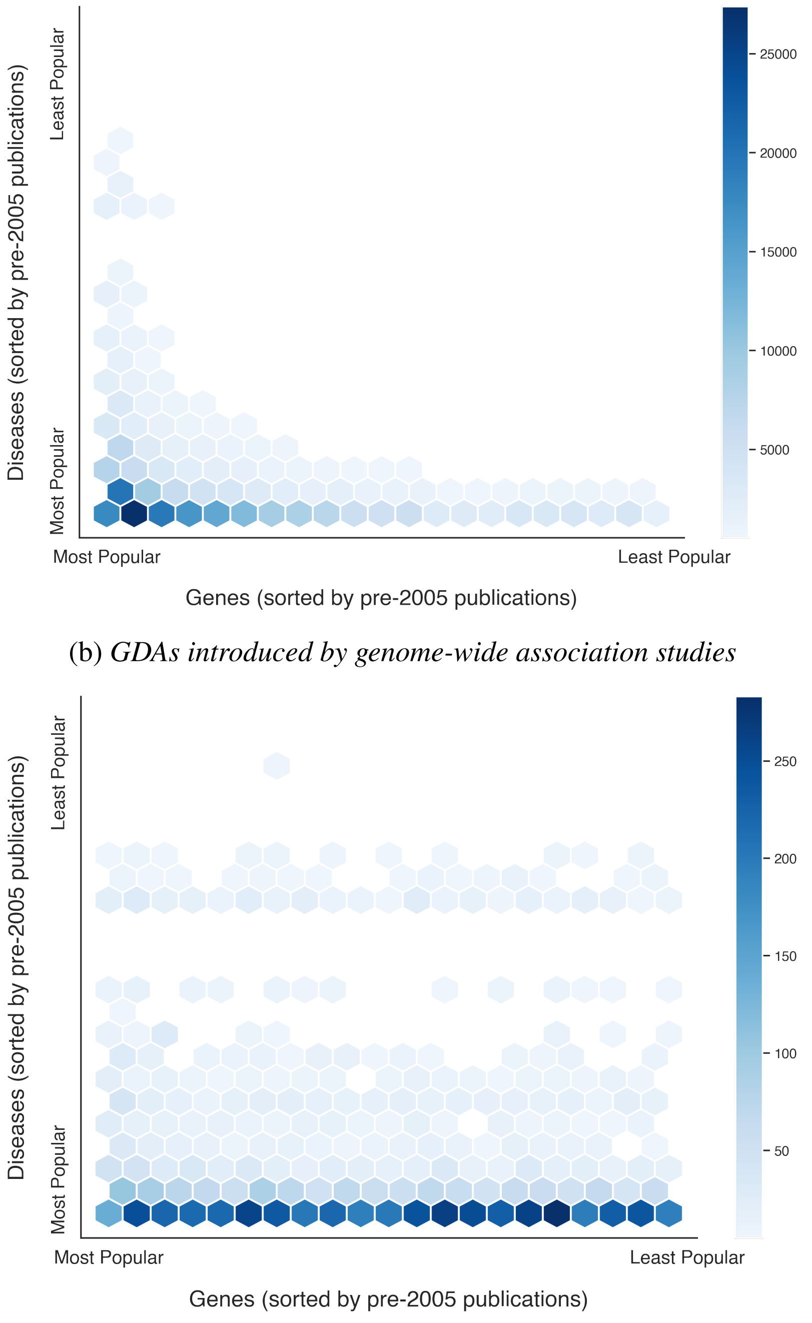
Hakon Hakonarson, Struan F. A. Grant, ... & Constantin Polychronakos

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

(b) *Yearly GWASs published*

Note: Panel (a) 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. Details on the GWAS by Hakonarson et al. (2007) are available in the Appendix B.3. Panel (b) shows the number of yearly genome-wide association studies published. The dark grey portion of the bars is the number of GWASs that introduced at least one novel gene-disease association, thus being the object of the present study. Data are from the GWAS Catalog. The vertical dashed line marks the completion of the Phase I of the HapMap project in 2005. See text for details.

Figure 3: Genome-wide associations studies span a larger portion of the genetic landscape relative to candidate gene studies.

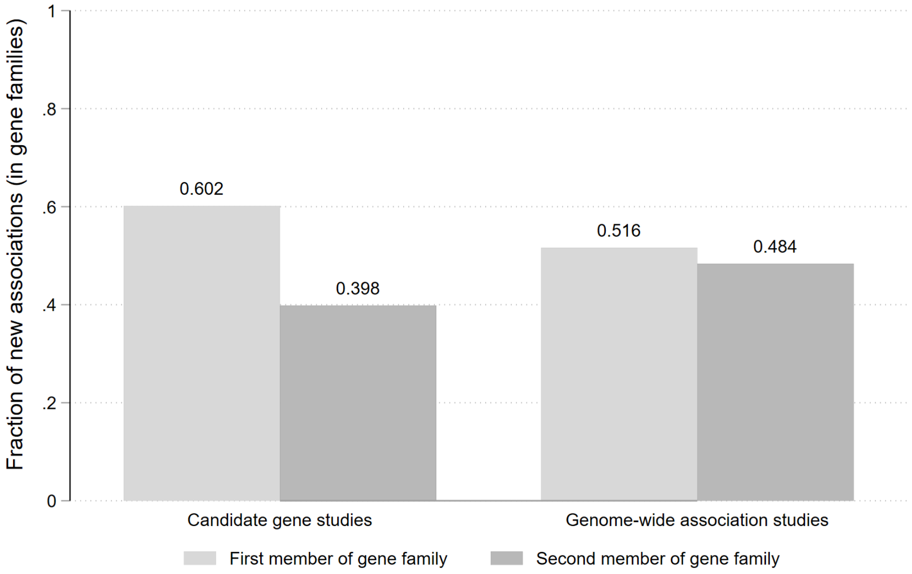


(a) *GDAs introduced by candidate gene studies*

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

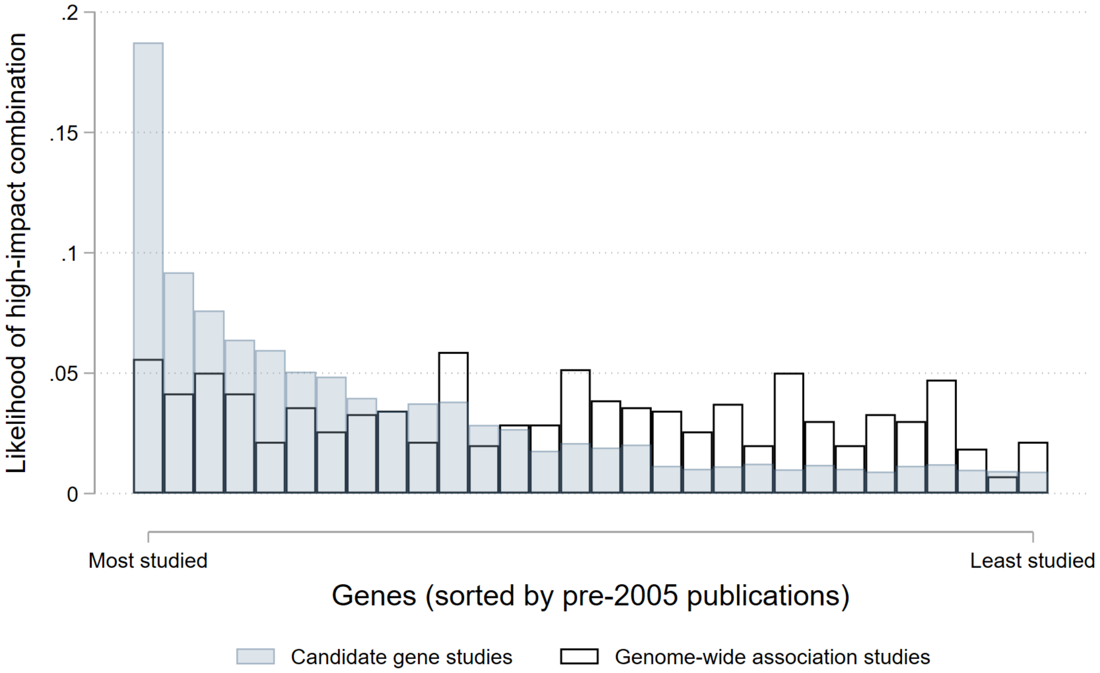
Note: Panel (a) shows a heatmap of new gene-disease associations introduced after 2005 by candidate gene studies. Panel (b) shows a heatmap of new gene-disease associations introduced after 2005 by genome-wide association studies. Both panels have 14,072 genes on the X axis, sorted from the most to the least studied in the pre-GWAS era, and 9,740 disease categories on the Y axis, sorted from the most to the least studied in the pre-GWAS era. Darker areas denote the introduction of a higher number of new gene-disease associations. Bins that include an equal number of genes. See text for details.

Figure 4: Gene-disease associations introduced by a GWAS are not biased towards the first member of gene families.



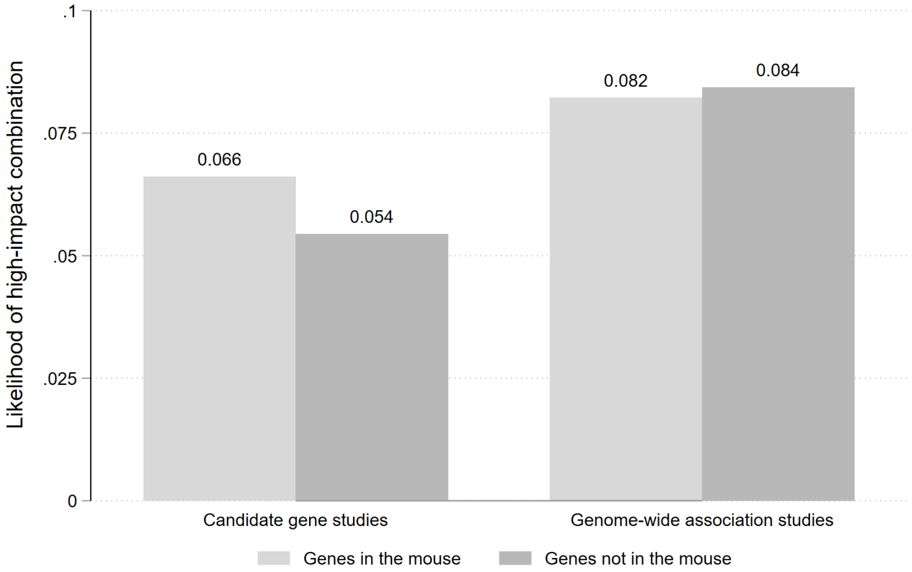
Note: The figure plots the share of new gene-disease associations on the first vs. the second member of a gene family, conditional on involving a gene family member. Data used in the graph are limited to all new gene-disease associations introduced in the period 2005-2016. Only diseases targeted by at least one GWAS are considered in this figure. See text for details.

Figure 5: Gene-disease associations introduced by GWAS are more likely to be high-importance for understudied genes, but candidate gene studies perform better when genes have been well studied.



Note: The histogram plots the share of high-importance GDAs for each gene distinguishing by the type of study that introduced them. Data used in the graph are limited to all new gene-disease associations introduced in the period 2005-2016. The 14,072 genes on the X axis are sorted from the most to the least studied in the pre-GWAS era. See text for details.

Figure 6: GWAS are especially effective to introduce high-importance gene-disease associations for genes that cannot be studied with the lab mouse.



Note: The figure plots the share of high-importance GDAs for genes with and without a mouse homolog distinguishing by the type of study that introduced them. Genes that lack a homolog gene in the lab mouse receive less attention and are more likely to be less known, but not because they are less biologically important. 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. See text for details.

Table 1: Descriptive statistics.

	Panel A: paper-level descriptives											
	Candidate gene studies						GWAS					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
Forward citations	41.47	23	84.107	0	8,084	134,670	169.11	81	287.289	0	2,822	1,216
Rank of the journal (ventile)	13.26	14	5.057	1	20	134,670	17.52	19	3.766	2	20	1,216
Associations per paper	2.55	2	5.231	1	916	134,670	6.86	3	14.049	1	241	1,216
Genes per paper	1.53	1	2.42	1	641	134,670	4.49	2	7.652	1	88	1,216
Number of authors	9.01	8	6.549	1	445	134,670	38.67	25	44.762	1	565	1,216
Year	2011.10	2011	3.392	2005	2016	134,670	2012.32	2012	2.509	2005	2016	1,216

	Panel B: association-level descriptives											
	Candidate gene studies						GWAS					
	mean	median	st d	min	max	N	mean	median	st d	min	max	N
With never associated genes (%)	0.122	0	0.327	0	1	343,722	0.406	0	0.491	0	1	8,440
With recently discovered gene (%)	0.099	0	0.298	0	1	343,722	0.258	0	0.438	0	1	8,440
With 2nd member of a gene family (%)	0.391	0	0.488	0	1	87,554	0.486	0	0.499	0	1	2,107
With genes lacking mouse homolog (%)	0.051	0	0.221	0	1	337,509	0.060	0	0.237	0	1	8,305
In top 5% of scientific importance (%)	0.049	0	0.215	0	1	343,722	0.083	0	0.276	0	1	8,440
With regulator genes (%)	0.502	0	0.500	0	1	343,722	0.235	0	0.424	0	1	8,440
Year of the association	2011.08	2011	3.381	2005	2016	343,722	2012.68	2013	2.603	2005	2016	8,440

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 GDAs introduced by the article; *Genes per paper*= number of genes associated with a disease by the article; *Number of authors*= number of researchers co-authoring the article. Panel B presents descriptive statistics on new gene-disease associations introduced after 2005. *With never associated genes (%)*= share of GDAs that include a gene never associated with 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 2nd member of a gene family (%)*= share of GDAs that include the second member of a gene family, conditional on being about a gene family (data on gene families from Stoeger et al. 2018); *With genes lacking mouse homolog (%)*= share of GDAs that include a gene lacking a homolog gene in the mouse (data from NIH); *In top 5% of scientific importance (%)*= share of GDAs that fall in the top 95th percentile of follow-on work (by year of discovery); *With regulator genes (%)*= share of GDAs that include a gene involved in controlling the function of more than one other gene (data from Türei et al. 2016); *Year of the association*= year in which the article introducing the GDA is published. The table reports only data that are effectively used in the empirical estimates, i.e., excluding observations that are dropped by the inclusion of fixed effects. See text for details.

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

Dependent Variable:	I(GDA for never associated gene>0)		I(GDA for recently discovered gene>0)	
GWAS	0.200*** (0.01399)	0.148*** (0.01563)	0.112*** (0.01108)	0.089*** (0.01406)
Disease FE	YES	YES	YES	YES
Principal Investigator FE	NO	YES	NO	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	352,162	331,825	352,162	331,825
Mean of the DV:	0.130	0.130	0.103	0.103
Number of diseases:	9,740	9,362	9,740	9,362
Number of genes:	14,072	13,910	14,072	13,910

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 with 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 2000; *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study. See text for details.

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

Dependent Variable:	I(GDA in the top 5% of scientific importance > 0)		
GWAS	0.012*	0.037***	0.044***
	(0.00663)	(0.00719)	(0.00918)
Disease FE	YES	YES	YES
Gene FE	NO	YES	YES
Principal Investigator FE	NO	NO	YES
Journal prestige FE	YES	YES	YES
Year of discovery FE	YES	YES	YES
Number of authors FE	YES	YES	YES
N	352,162	350,693	330,250
Mean of the DV:	0.05	0.05	0.05
Number of diseases:	9,740	9,726	9,347
Number of genes:	14,072	12,617	12,463

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 importance>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. See text for details.

Table 4: Genome-wide association studies are especially more likely to introduce gene-disease associations of high scientific importance for genes less studied because they lack a mouse homolog.

Dependent Variable:	I(GDA in the top 5% of scientific importance>0)			
Subsample:	Genes in the mouse		Genes not in the mouse	
GWAS	0.036*** (0.00738)	0.044*** (0.00942)	0.062* (0.03064)	0.236*** (0.07017)
Disease FE	YES	YES	YES	YES
Gene FE	YES	YES	YES	YES
Principal Investigator FE	NO	YES	NO	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	326,506	306,232	15,848	11,478
Mean of the DV:	0.05	0.05	0.04	0.04
Number of diseases:	9,480	9,069	1,952	1,467
Number of genes:	11,502	11,361	743	695

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 importance>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. The sample used in columns 1 and 2 includes only genes with a homolog gene in the lab mouse, while the sample used in columns 3 and 4 includes only genes without a homolog gene in the lab mouse. See text for details.

Table 5: Genome-wide association studies are less likely to introduce high-importance gene-disease associations than candidate gene studies for genes involved in complex regulatory networks, provided that there is enough theoretical knowledge about them.

Dependent Variable:	I(GDA in the top 5% of scientific importance>0)					
Subsample:	Genes in top 50% of most studied		Genes in top 75% of most studied		Genes in top 90% of most studied	
GWAS	0.038***	0.047***	0.031	0.072**	-0.032*	-0.001
	(0.01172)	(0.01590)	(0.01987)	(0.03251)	(0.01750)	(0.02411)
GWAS × Regulator Gene		-0.016		-0.072**		-0.057**
		(0.02071)		(0.03213)		(0.02501)
Disease FE	YES	YES	YES	YES	YES	YES
Gene FE	YES	YES	YES	YES	YES	YES
Journal prestige FE	YES	YES	YES	YES	YES	YES
Year of discovery FE	YES	YES	YES	YES	YES	YES
Number of authors FE	YES	YES	YES	YES	YES	YES
N	176,206	176,206	86,421	86,421	33,159	33,159
Number of diseases:	7,901	7,901	6,298	6,298	4,463	4,463
Number of genes:	2,189	2,189	642	642	175	175

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 importance>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; *Regulator Gene*:0/1=1 for genes that are involved in controlling the function of more than one other gene (data from Türei et al. 2016). Column 1 and 2 include only GDAs with genes that received an above median number of studies before 2005. Column 3 and 4 include only GDAs with genes that are in the top 75% of the distribution of studies received before 2005. Column 5 and 6 include only GDAs with genes that are in the top 90% of the distribution of studies received before 2005.

Data-Driven Search and Innovation

Appendix

Matteo Tranchero
UC Berkeley

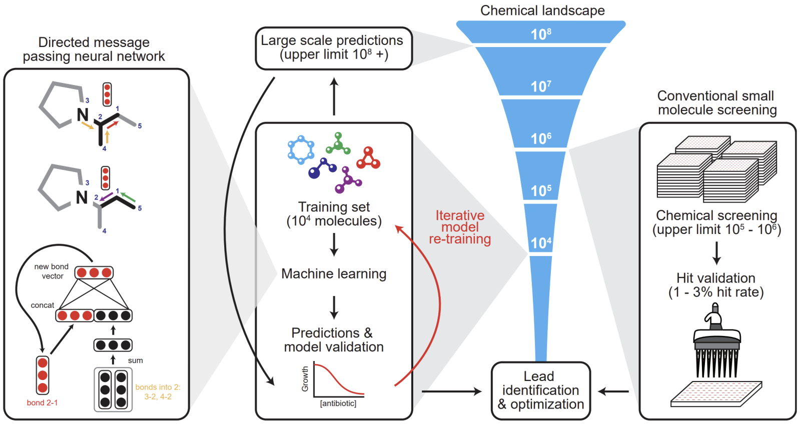
A An Example of Data-Driven Search from Combinatorial Chemistry

Consider the critical task of discovering new drugs. This is a difficult problem since the chemical space is complex and high-dimensional (Gambardella, 1995; Jayaraj and Gittelman, 2018; Rzhetsky et al., 2015). Historically, the most successful molecules were serendipitously identified with random search or by experimenting in the neighborhood of known ones. This process was very long, costly, and inefficient (Arora and Gambardella, 1994; 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 with drug-like characteristics (e.g., small molecules with a weight theoretically deemed to be suitable 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 study's authors discovered a new compound called halicin that has promising antibiotic properties. This result was achieved with a fraction of the time and costs involved in sequential assay screening, since it happened in silico without the need to undertake costly experimentation in the lab. Furthermore, this achievement 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

Figure A.1: Figure schematically representing the analysis of Stokes et al. (2020).


Directed message passing neural network
N
1
2
3
4
5
new bond vector
concat
sum
bond 2-1
bonds into 2: 3-2, 4-2
Large scale predictions (upper limit 10^8 +)
Training set (10^4 molecules)
Machine learning
Predictions & model validation
Growth
[antibiotic]
Iterative model re-training
Chemical landscape
10^8
10^7
10^6
10^5
10^4
Lead identification & optimization
Conventional small molecule screening
Chemical screening (upper limit 10^5 - 10^6)
Hit validation (1 - 3% hit rate)

Note: The figure shows Figure 1 from Stokes et al. (2020). Using a deep neural network model, Stokes et al. (2020) built a molecular representation based on a specific property, in this case the inhibition of the growth of E. coli. Then, they applied the model to multiple chemical libraries to identify potential lead compounds with activity against E. coli. Finally, Stokes et al. (2020) selected a list of promising candidates after ranking the candidates according to the score predicted by the model.

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 ex ante. 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 the highest on it. For Stokes et al. (2020) this was the growth inhibition of *Escherichia coli*, but it 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, it must be possible to foresee the effect of potential 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.

B Genome-Wide Association Studies: Examples and Details

B.1 Genetic Research Before GWAS

Genomics is the branch of biological sciences concerned with the study of genomes, i.e., the entire collection of an organism's genes. In turn, genes are sequences of DNA bases that encode the "instructions" to synthesize gene products (e.g., proteins). Genes have a fundamental role in the functioning of the human body, but sometimes they can acquire mutations in their sequence. When this happens, genes might alter their behavior and affect phenotypic traits, sometimes with significant consequences and the emergence of severe health conditions. However, the role of genes in the etiology of diseases offers an avenue for therapeutic intervention since genes associated with a condition can often be used as drug targets (Nelson et al., 2015). When a drug molecule binds to the therapeutic target it can modify its functioning, favorably affecting the outcome of a disease. Therefore, knowing the genetic roots of diseases has important practical consequences in the design of pharmaceutical drugs.

Genetic research has historically been very effective in locating genes individually responsible for disease conditions. These are called *Mendelian disorders*, which can usually be seen since birth and be deduced based on family history. Consider, for instance, the case of cystic fibrosis. This rare disorder can be caused by multiple DNA mutations which however cluster in a well-defined area of the genome, namely in the *Cystic fibrosis transmembrane conductance regulator* (CFTR) gene. Studying this kind of Mendelian disorders was one of the earliest and more significant successes of family studies in genetics (Bush and Moore, 2012). The primary research strategy involved genotyping families affected by cystic fibrosis: even with very small sample sizes due to the rarity of the disease, the very strong effect of the CFTR gene mutations allowed to unambiguously identify the gene as causally connected with the disease.

However, Mendelian diseases are typically rare because they tend to be eliminated by evolutionary pressures due to their severity. Much more common are *complex diseases* that are not due to a single genetic factor, but rather by many genes. For complex diseases, any genetic mutation can increase the risk even without being neither necessary nor sufficient, which means that it is responsible for only a small proportion of the heritability of the disease. Although complex disorders often cluster in families too, they do not have a predictable inheritance pattern since they are influenced by convoluted interactions between genes and environmental factors. For this reason, family linkage studies have not fared well when applied to more common disorders.

B.2 A Scientific Primer on GWAS

To make progress in the study of complex diseases, researchers have started to focus on the idea that common disorders are likely influenced by genetic mutations that are also common in the population (Reich and Lander, 2001). Instead of looking for individual genes with strong effects on phenotypes, the field has moved toward the study of common, generic variants that individually have a small effect on the likelihood of having a disease (Bush and Moore, 2012). But what precisely is a variant? At the most fundamental level, two genomes differ in a specific genetic locus if they present an alternative single nucleotide (adenine, thymine, cytosine, or guanine) in that location. Such mutation in one DNA basis is called single-nucleotide polymorphism (SNP) when it appears in at least 1% of the population. One approach for associating SNPs with a disease relies on the fact that a causative variant should be found at a higher frequency in cases than in control subjects. In practice, researchers look for statistical correlations between specific genetic variants and diseases in large samples of people not necessarily related by family ties.

Building over this logic, genome-wide association studies are hypothesis-free methods for identifying associations between genetic regions and traits (Visscher et al., 2017). GWASs compare genetic differences between affected and unaffected individuals using genotyping technologies on large samples. In a typical GWAS project, researchers obtain DNA from two groups of participants, patients with the disease studied and healthy individuals with comparable demographics. Then, selected SNPs on the chromosome are scanned using high-throughput arrays that can genotype up to millions of SNPs for each individual. Variants significantly more likely to appear in the affected patients could be biologically important for the disease and are thus likely to be associated with its etiology. Such SNPs might affect gene expression and function, especially when located within a protein-coding gene. It is important to underscore that array-based genome-wide studies do not sequence the DNA base by base, since they only determine the presence or absence of a number of SNPs. While microarrays can genotype millions of SNPs, they only cover a tiny fraction (usually <0.1%) of the genome. Nevertheless, exploiting the fact that co-occurrence of variants in proximal genetic loci is not random (a phenomenon called linkage), researchers can use reference genomes (such as the HapMap) to parsimoniously infer the characteristics of the whole genome from the much smaller number of SNPs genotyped (Bush and Moore, 2012).

GWASs have unlocked many significant scientific findings, but have come under scrutiny due to several limitations. First, array-based genome-wide scans can only identify common variants which tag a region likely containing the causal variants of interest. GWAS cannot locate the causal SNPs with certainty, so additional analyses or follow-on studies are usually required to narrow the association region. Second, even after the causal variant has been pinned down, the biological mechanism underlying its role in human health requires additional study. Third, the majority of complex diseases are co-determined by a large number of genes, which means that the proportion of variance explained by any individual variant is small. This fact often hinders the therapeutic translation of GWAS findings, since variants with an effect that is too small do not provide actionable drug targets (Goldstein et al., 2009). Finally, scholars have suggested that GWASs neglect more complex interaction structures between genes, being limited to testing pairwise gene-disease associations (Boyle et al., 2017)

B.3 Hakonarson et al. (2007) and the KIAA0350-Type I Diabetes link

It is estimated that 10% of Americans (around 37.3 million people) have diabetes.[10] The two main forms of diabetes are type 1 diabetes and type 2 diabetes, with the former accounting for 5-10% of cases. Type 1 diabetes is a chronic condition that often begins during childhood. More in detail, type 1 diabetes is an autoimmune disease that originates from the destruction of β-cells by the immune system (DiMeglio et al., 2018). As a result, the pancreas fails to produce adequate insulin levels, the fundamental hormone needed to allow sugar to enter cells to produce energy and regulate normal glucose levels in the bloodstream. To date, there is no known way to prevent type 1 diabetes, and continuing treatment with insulin is required for patient survival.

Genetics has a sizable role in the emergence of type 1 diabetes: children who have a parent with this condition have a relative risk of 1-9% to present the same condition (DiMeglio et al., 2018). Early candidate gene studies on this disease identified a few genetic determinants, mostly within a set of closely linked genes known as the major histocompatibility complex (MHC). MHC genes code several cell surface proteins that are essential to trigger and target the immune system's response. Malfunctioning in such genes can lead to autoimmune responses, such as type 1 diabetes. Yet, MHC genes explain little more than half of the genetic risk for the disease, indicating that other unknown genetic loci exist. Systematic detection of all remaining genes involved could offer alternative therapeutic pathways and help clarify the root causes of diabetes.

In August 2007, Hakonarson et al. (2007) published an influential genome-wide association study performed on a study population of 563 patients with type 1 diabetes and 1,146 controls.[11] The analysis was done using a microarray capable of genotyping 550,000 single nucleotide polymorphisms (SNPs). The study identified several SNPs significantly associated with type 1 diabetes. Figure B.1 reports the main results of Hakonarson et al. (2007). Some significant SNPs were located in genes already known to be related to diabetes (e.g., the insulin gene *INS*), but three of them were located in the gene *KIAA0350*. KIAA0350 controls β-cell function and helps prevent diabetes, but notably it was one of the least studied human genes and its function was unknown at the time of the finding (Soleimanpour et al., 2014).

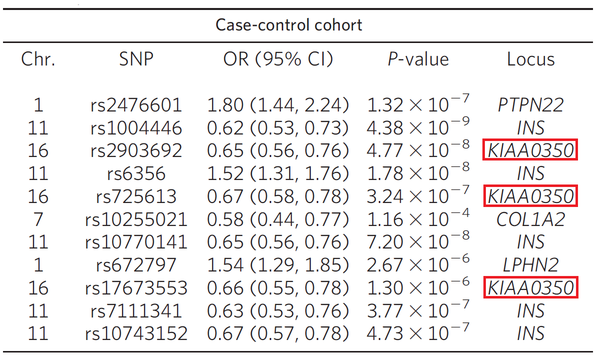
The KIAA0350-type I diabetes association is robust and has been further investigated in several studies (Gingerich et al., 2020). In my data, this association has a high value of the DisGeNET's GDA Score measure of scientific reliability and no contradictory evidence (Piñero et al., 2020). Twenty-two published papers have experimentally validated the role of KIAA0350 in type 1

[10] The figure is taken from: https://www.cdc.gov/diabetes

[11] The paper also replicated the main analysis on a different sample made of 483 complete family trios, i.e., exploiting differences in DNA between parents and their affected child.

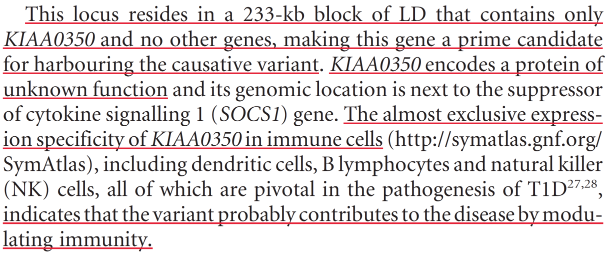
Figure B.1: Main results from the GWAS analysis of Hakonarson et al. (2007)

(a) *Results from main analysis*



		Case-control cohort		
Chr.	SNP	OR (95% CI)	*P*-value	Locus
1	rs2476601	1.80 (1.44, 2.24)	1.32×10^{-7}	*PTPN22*
11	rs1004446	0.62 (0.53, 0.73)	4.38×10^{-9}	*INS*
16	rs2903692	0.65 (0.56, 0.76)	4.77×10^{-8}	*KIAA0350*
11	rs6356	1.52 (1.31, 1.76)	1.78×10^{-8}	*INS*
16	rs725613	0.67 (0.58, 0.78)	3.24×10^{-7}	*KIAA0350*
7	rs10255021	0.58 (0.44, 0.77)	1.16×10^{-4}	*COL1A2*
11	rs10770141	0.65 (0.56, 0.76)	7.20×10^{-8}	*INS*
1	rs672797	1.54 (1.29, 1.85)	2.67×10^{-6}	*LPHN2*
16	rs17673553	0.66 (0.55, 0.78)	1.30×10^{-6}	*KIAA0350*
11	rs7111341	0.63 (0.53, 0.76)	3.77×10^{-7}	*INS*
11	rs10743152	0.67 (0.57, 0.78)	4.73×10^{-7}	*INS*

(b) *Text excerpt on KIAA0350 gene*



This locus resides in a 233-kb block of LD that contains only *KIAA0350* and no other genes, making this gene a prime candidate for harbouring the causative variant. *KIAA0350* encodes a protein of unknown function and its genomic location is next to the suppressor of cytokine signalling 1 (*SOCS1*) gene. The almost exclusive expression specificity of *KIAA0350* in immune cells (http://symatlas.gnf.org/SymAtlas), including dendritic cells, B lymphocytes and natural killer (NK) cells, all of which are pivotal in the pathogenesis of T1D[27,28], indicates that the variant probably contributes to the disease by modulating immunity.

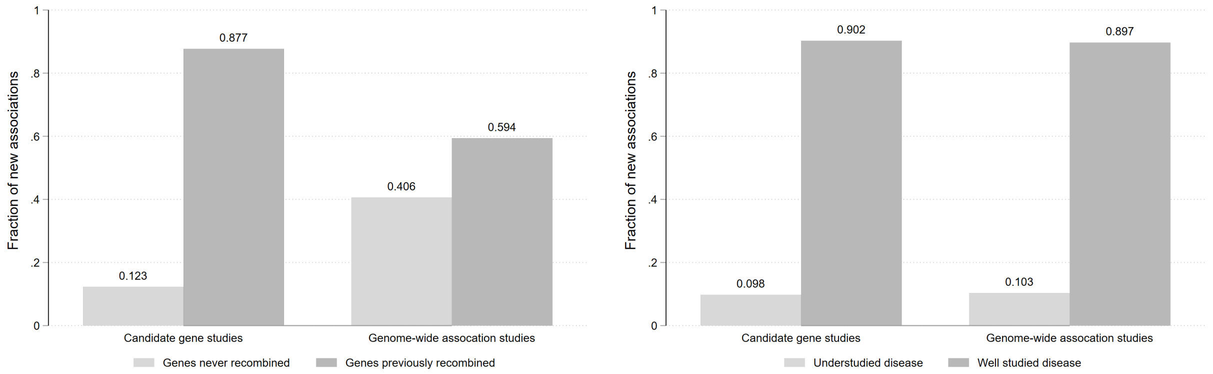
Note: Panel (a) shows an excerpt from Table 1 of Hakonarson et al. (2007). The genetic location of the single nucleotide polymorphisms significantly associated with type 1 diabetes are shown in the rightmost column. Highlighted are the three SNPs located in the KIAA0350 gene. Panel (b) shows the passage of Hakonarson et al. (2007) describing the inferred role of KIAA0350. T1D stands for type 1 diabetes.

diabetes, putting this discovery in the top 95th percentile of scientific importance in my data. Notably, since the study of Hakonarson et al. (2007), genome-wide association studies have discovered more than 60 additional genetic loci associated with the risk of type 1 diabetes (DiMeglio et al., 2018), setting the stage for an improved genetic understanding of the disease.

Interestingly, the same Authors of the GWAS also published a candidate gene study on the same disease a few months before. In this paper, Qu et al. (2007) leveraged the literature showing IRF5's role in immune response to hypothesize that the gene IRF5 might be related to type 1 diabetes as well. Unfortunately, despite being supported by reasoning based on available knowledge, the hypothesis proved wrong: the Authors report no strong association between IRF5 and type 1 diabetes. Not surprisingly, this gene-disease association was studied by only one subsequent paper according to the DisGeNET data. What is interesting to note, however, is that the knowledge that informs the hypothesis of Qu et al. (2007) is precisely what leads to (wrongly) target an already well-studied gene and to neglect the opportunity offered by genes like KIAA0350.

C Additional Figures and Tables

Figure C.1: GWAS induce diversification in gene space but do not change the disease focus.

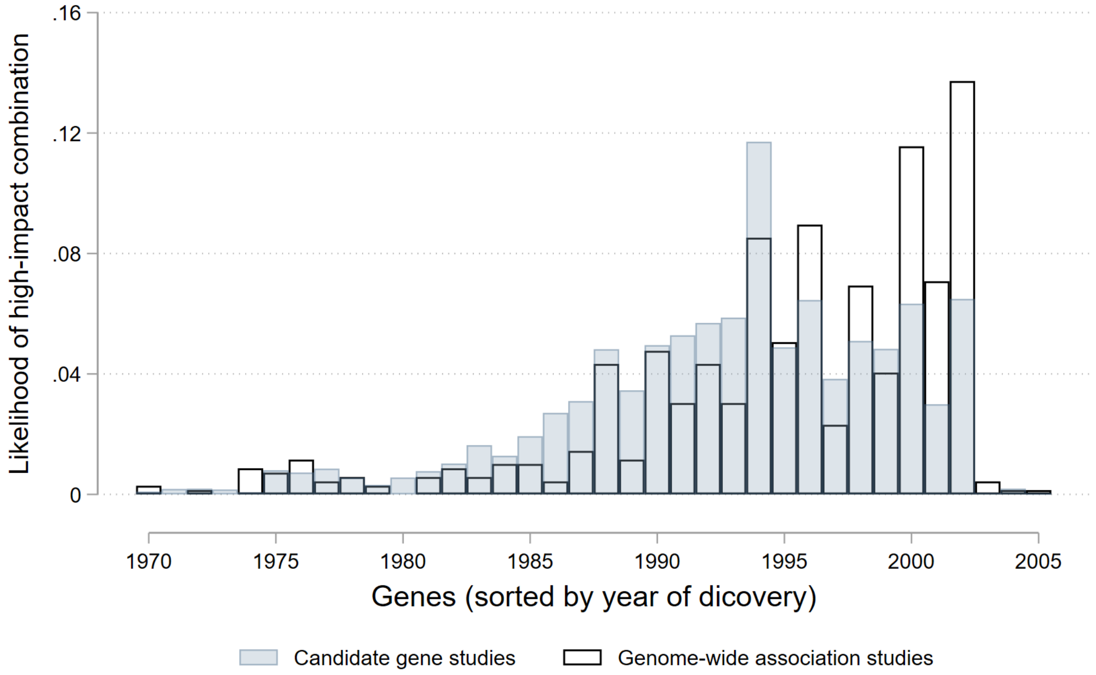


(a) Share of new GDAs by popularity of the gene involved.

(b) Share of new GDAs by popularity of the disease involved.

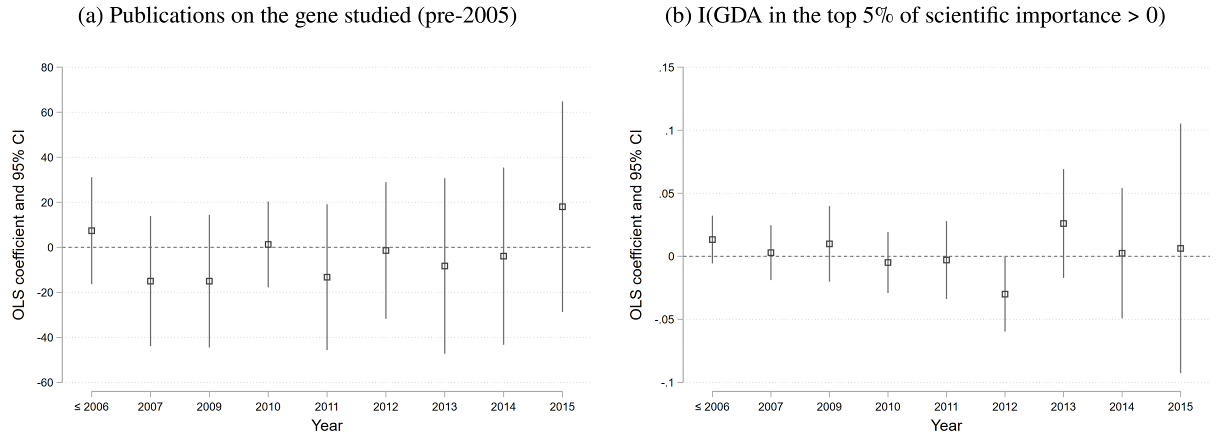
Note: Panel (a) plots the share of new gene-disease associations that involve genes never associated with a disease before vs well studied genes, separately by type of study. Panel (b) plots the share of new gene-disease associations that involve diseases with below median number of gene associations in the DisGeNET data as of 2005 (i.e., less than 2) vs well studied diseases, separately by type of study. Data used in the graph are limited to all new gene-disease associations introduced in the period 2005-2016.

Figure C.2: Gene-disease associations introduced by GWAS are less likely to be high-importance for genes discovered earlier, but more likely to be so for recently discovered genes.



Note: The histogram plots the share of high-importance GDAs for each gene distinguishing by the type of study that introduced them. Data used in the graph are limited to all new gene-disease associations introduced in the period 2005-2016. The 14,072 genes on the X axis are sorted by the year they were first reported in the literature.

Figure C.3: Before their first GWAS, principal investigators that adopt the genome-wide approach are not more likely to target less studied genes or to introduce breakthrough gene-disease associations.

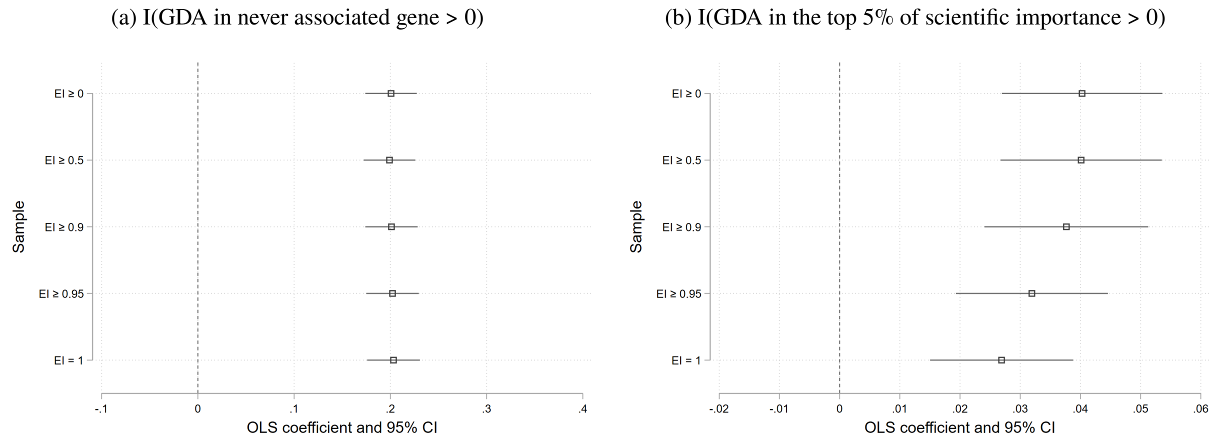


(a) Publications on the gene studied (pre-2005)

(b) I(GDA in the top 5% of scientific importance > 0)

Note: Every coefficient is estimated from a separate regression for a given year, where I compare principal investigators that never publish a GWAS with principal investigators that will eventually publish one but have not done so already (that is, PIs are excluded from the sample after their first GWAS). Panel (a) plots plots the coefficients and 95% confidence intervals from regressing the popularity of a gene studied over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS. Panel (b) plots the coefficients and 95% confidence intervals of the regression of the likelihood that a GDA is in the top 95th percentile of importance on a dummy that indicates if the principal investigator of the study will eventually publish at least one GWAS.

Figure C.4: Robustness of the main results to the choice of sample.



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

(b) I(GDA in the top 5% of scientific importance > 0)

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 likelihood that a GDA is in the top 95th percentile of importance 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 $EI \geq 0.9$.

Table C.1: Genome-wide association studies are more likely to introduce gene-disease associations involving genes that received less publications or were discovered later than candidate gene studies.

Dependent Variable:	Pre-2005 pubs on the gene		Year of discovery of the gene	
GWAS	-33.785***	-30.702***	2.159***	1.635***
	(6.14160)	(10.11909)	(0.20644)	(0.25135)
Disease FE	YES	YES	YES	YES
Principal Investigator FE	NO	YES	NO	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	352,162	331,825	347,736	327,532
Mean of the DV:	98.886	98.886	1992.208	1992.208
Number of diseases:	9,740	9,362	9,663	9,281
Number of genes:	14,072	13,910	13,770	13,615

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. *Pre-2005 pubs on the gene*: count of the publications received before 2005 by the gene associated to the disease; *Year of discovery of the gene*: year of discovery of the gene associated to the disease; *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study. See text for details.

Table C.2: Robustness of the main results to alternative definitions of breakthrough discoveries.

Dependent Variable:	I(GDA in the top 10% of scientific importance>0)		I(GDA in the top 2.5% of scientific importance>0)		I(GDA in the top 1% of scientific importance>0)	
GWAS	0.051***	0.053***	0.032***	0.033***	0.018***	0.017***
	(0.00940)	(0.01112)	(0.00557)	(0.00762)	(0.00389)	(0.00531)
Disease FE	YES	YES	YES	YES	YES	YES
Gene FE	YES	YES	YES	YES	YES	YES
Principal Investigator FE	NO	YES	NO	YES	NO	YES
Journal prestige FE	YES	YES	YES	YES	YES	YES
Year of discovery FE	YES	YES	YES	YES	YES	YES
Number of authors FE	YES	YES	YES	YES	YES	YES
N	350,693	330,250	350,693	330,250	350,693	330,250
Mean of the DV:	0.1	0.1	0.025	0.025	0.01	0.01
Number of diseases:	9,726	9,347	9,726	9,347	9,726	9,347
Number of genes:	12,617	12,463	12,617	12,463	12,617	12,463

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 10% of scientific importance>0)*:0/1=1 if the gene-disease association involves a gene in the top 90th percentile of discoveries with the most follow-on work (by year of discovery); *I(GDA in the top 2.5% of scientific importance>0)*:0/1=1 if the gene-disease association involves a gene in the top 97.5th percentile of discoveries with the most follow-on work (by year of discovery); *I(GDA in the top 1% of scientific importance>0)*:0/1=1 if the gene-disease association involves a gene in the top 99th percentile of discoveries with the most follow-on work (by year of discovery); *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study.

Table C.3: Genome-wide association studies introduce gene-disease associations that receive more follow-on scientific publications than candidate gene studies.

Dependent Variable:	Count of follow-on publications		
GWAS	0.545***	0.969***	1.211***
	(0.19668)	(0.24385)	(0.34976)
Disease FE	YES	YES	YES
Gene FE	NO	YES	YES
Principal Investigator FE	NO	NO	YES
Journal prestige FE	YES	YES	YES
Year of discovery FE	YES	YES	YES
Number of authors FE	YES	YES	YES
N	352,162	350,693	330,250
Mean of the DV:	0.56	0.56	0.56
Number of diseases:	9,740	9,726	9,347
Number of genes:	14,072	12,617	12,463

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. *Count of follow-on publications*: number of scientific articles studying the gene-disease association after it has been introduced the first time by either a GWAS or a candidate gene study; *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study.

Table C.4: Gene-disease association studies are more likely to involve the second member of gene families and to find breakthrough discoveries involving them.

Dependent Variable:	I(GDA with the second member of a gene family>0)		I(GDA in the top 5% of scientific importance>0)	
GWAS	0.083*** (0.01948)	0.065*** (0.03060)	0.036*** (0.01180)	0.070*** (0.02250)
Disease FE	NO	YES	YES	YES
Gene family FE	YES	YES	YES	YES
Principal Investigator FE	NO	YES	NO	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	87,360	75,254	87,360	75,254
Mean of the DV:	0.39	0.39	0.05	0.05
Number of diseases:	5,152	4,639	5,152	4,639
Number of genes:	3,009	2,918	3,009	2,918

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 with the second member of a gene family>0)*:0/1=1 if the gene-disease association involves the second member of a gene family; *I(GDA in the top 5% of scientific importance>0)*:0/1=1 if the gene-disease association involves a gene in the top 95th percentile of discoveries with the most follow-on work (by year of discovery); *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study. The sample is limited to genes that are members of a gene family.

Table C.5: Diseases targeted first by GWAS are not more likely to be associated to understudied genes.

Dependent Variable:	I(GDA for never associated gene>0)		
Subsample:	Early GWAS	Mid-period GWAS	Late GWAS
GWAS	0.133*** (0.02652)	0.108*** (0.02951)	0.137*** (0.04230)
Disease FE	YES	YES	YES
Principal Investigator 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	47,167	62,712	42,048
Number of diseases:	111	333	380
Number of genes:	9,586	9,914	8,830

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 with a disease before 2005; *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study. Column 1 includes only diseases that received their first GWAS before 2009. Column 2 includes only diseases that received their first GWAS between 2009 and 2011. Column 3 includes only diseases that received their first GWAS after 2011.

Table C.6: Robustness of the main results to the use of DisGeNET's GDA Score to measure scientific value of new gene-disease combinations.

Dependent Variable:	DisGeNET *GDA Score* for gene-disease associations				
Subsample:	All genes			Genes in mouse	Genes not in mouse
GWAS	0.031*** (0.00424)	0.037*** (0.00410)	0.038*** (0.00426)	0.037*** (0.00417)	0.049*** (0.00920)
Disease FE	YES	YES	YES	YES	YES
Gene FE	NO	YES	YES	YES	YES
Principal Investigator FE	YES	YES	YES	NO	NO
Journal prestige FE	YES	YES	YES	YES	YES
Year of discovery FE	YES	YES	YES	YES	YES
Number of authors FE	YES	YES	YES	YES	YES
N	352,162	350,693	330,250	326,506	15,848
Mean if the DV:	0.06	0.06	0.06	0.05	0.06
Number of diseases:	9,740	9,726	9,347	9,480	1,952
Number of genes:	14,072	12,617	12,463	11,502	743

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 provided by DisGeNET (Piñero et al., 2020); *GWAS*:0/1=1 for GDAs introduced by a genome-wide association study.

Table C.7: Genome-wide association studies that introduce new gene-disease associations are more likely to be cited by clinical trial articles than comparable candidate gene studies.

Dependent Variable:	Cited by a clinical article		
GWAS paper	0.188*** (0.01994)	0.193*** (0.01771)	0.086*** (0.01791)
Disease FE	YES	YES	YES
Gene FE	NO	YES	YES
Principal Investigator FE	NO	NO	YES
Journal prestige FE	YES	YES	YES
Year of discovery FE	YES	YES	YES
Number of authors FE	YES	YES	YES
N	352,162	350,693	330,250
Mean of the DV:	0.294	0.294	0.294
Number of diseases:	9,740	9,726	9,347
Number of genes:	14,072	12,617	12,463

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. *Cited by a clinical articl*:0/1=1 if the paper that introduced the gene-disease association is later cited by one article describing the outcomes of a clinical trial; *GWAS paper*:0/1=1 for papers reporting a genome-wide association study. See text for details.