# Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies *

Matteo Tranchero
The Wharton School

May, 2025

**Abstract**

How does big data change the search for innovation? Data-driven predictions can identify promising technological combinations even when their underlying mechanisms are unknown. This has raised concerns that decoupling innovation from theoretical understanding weakens incentives to develop new theory and results in discoveries whose consequences are poorly understood. Using an evolutionary framework of variation and selection, I argue that data technologies can, instead, reinforce theory generation. Data-driven search broadens the space of combinations explored and increases the variability in outcomes compared to theory-driven approaches. As a result, this search strategy uncovers more surprising findings that stimulate, rather than substitute, new theorizing. I test these ideas in the domain of human genetics, where genome-wide association studies (GWAS) represent a data-driven search for the genetic roots of disease. Compared to traditional theory-based approaches, GWAS introduce gene-disease combinations that span a wider portion of the genetic landscape, more frequently fall at both extremes of scientific quality, and often defy expectations from existing knowledge. Rather than crowding out theory, GWAS findings trigger a surge of follow-on work aimed at elucidating their causal mechanisms. Together, the results reveal a complementarity between theory and data in search, suggesting that big data technologies can fuel virtuous cycles of theorizing sparked by empirical anomalies.

*E-mail: mtranc@wharton.upenn.edu. A previous version of this paper was circulated with the title "Data-Driven Search and Innovation". I gratefully acknowledge the financial support of Panmure House's Emergent Thought Award.

# 1 Introduction

Constantin Polychronakos is a leading researcher on the genetic factors behind type 1 diabetes. Pinpointing the genes that carry mutations responsible for the disease is critical, because they offer potential targets for therapeutic intervention. Yet, the task is daunting: the human genome is vast, while the resources to investigate it are necessarily limited. Faced with thousands of potential gene-disease combinations, how does Polychronakos decide where to search?

Understanding how Polychronakos searches in the genetic space can offer broader insight into the nature of innovation—not only for individual scientists, but also for organizations and firms. Prior research suggests that rather than searching randomly, actors typically focus on the combinations they expect to be most valuable (Kneeland et al., 2020). At its core, this is a prediction problem: actors draw on their knowledge, embodied in mental models and known cause-effect relationships between technological components, to predict which directions are promising (Fleming and Sorenson, 2004; Gavetti and Levinthal, 2000; Sorenson, 2024). This theory-driven process is exactly what Polychronakos followed in a 2007 study. Building on existing evidence that the IRF5 gene is implicated in lupus, he hypothesized that IRF5 might also play a role in other autoimmune diseases, leading him to investigate its connection with diabetes (Qu et al., 2007).

However, theory-driven search appears to be increasingly supplanted by innovation practices powered by machine learning and artificial intelligence (AI) (Allen and McDonald, 2025; Agrawal et al., 2024). In pharmaceutical research, for example, scientists routinely screen data on millions of compounds without relying on any theoretical knowledge (Jayaraj and Gittelman, 2018). Similarly, researchers in materials science employ computational models to evaluate countless unfamiliar material configurations (Merchant et al., 2023). Across many domains, innovators are moving away from theory-based reasoning and toward data-driven prediction to identify promising technological combinations (Kim, 2023). Polychronakos himself adopted this approach in a second 2007 study. Rather than drawing on prior theory, he collected data from thousands of individuals to find statistical associations between genetic mutations and type 1 diabetes. This data-driven strategy led to the identification of KIAA0350, a gene of unknown function, as a predicted culprit of the disease (Hakonarson et al., 2007).

The use of data technologies in search reveals a fundamental asymmetry: while theoretical knowledge supports the search for new ideas (Fleming and Sorenson, 2004), it is not always required to apply them in practice (Evans, 2010). Even in drug development, the U.S. Food and Drug Administration requires that treatments be safe and effective, but not that their biological mechanisms be understood. The power of data-driven search lies precisely in its ability to uncover valuable innovations even when there

is no prior theory. However, as machine learning and AI become increasingly central to the innovation process, the motivation to develop explanatory theories may erode (Anderson, 2008; Balasubramanian et al., 2022). The result is a potential "intellectual debt" (Zittrain, 2019): a growing dependence on innovations whose inner workings remain opaque and may carry unforeseen consequences. For instance, the drugs identified by AI models often have poorly understood properties, making it harder to anticipate how they will perform or what side effects they may cause (Heaven, 2023). Given the potential dangers of black-boxed innovation, it is crucial to assess whether the rise of big data technologies is undermining the development of new theory.

In this paper, I conceptualize how data technologies change innovation by drawing on a simple evolutionary framework of variation and selection. I argue that prior theory, while helping actors identify opportunities (Felin et al., 2024), can also blind them to alternative possibilities (Chai, 2017). This dynamic constrains the generation of variation by funneling attention to what is theoretically justified and then filtering out ideas that fall outside the dominant paradigm (Dosi, 1982). Data-driven search, by contrast, operates with fewer a priori constraints, enabling broader combinatorial exploration. While this likely increases quality variability, leading to a thicker-tailed distribution of outcomes, the absence of theory-based selection can surface unexpected patterns that existing knowledge would not predict. These surprising findings, in turn, prompt the need for cause-effect theories to explain them, triggering new cycles of theorizing (Kuhn, 1962). Therefore, this reasoning suggests the opposite of a crowding-out effect: the diffusion of big data increases the frequency and speed at which empirical anomalies are uncovered, leading to a net increase in theory development.

Empirically examining how data reshape innovation presents a challenge: one must observe both theory-driven and data-driven search processes and connect them to their resulting findings. This is difficult in most settings, where researchers see only the final innovation outcomes without knowing the search process that produced them (Kneeland et al., 2020; Schilling and Green, 2011). In this paper, I address this challenge by studying how scientists investigate the genetic roots of human diseases. The search for therapeutically useful gene-disease combinations usually happens with two distinct approaches. In *candidate gene studies*, researchers use theoretical priors to target genes predicted to be relevant for the disease. In contrast, *genome-wide association studies* (GWAS) take an atheoretical approach, scanning the entire genome to identify correlations between genetic mutations and diseases. Crucially, I can infer the search process that led to each gene-disease combinations from the methods described in the publication that introduced it. By comparing combinations introduced by GWAS to those from candidate gene studies, I analyze how data-driven search reshapes innovation in a domain critical to drug development.

I assemble a novel dataset capturing the characteristics of gene-disease associations introduced between 2005 and 2016. The raw data come from DisGeNET, a comprehensive aggregator of information on genes linked to human diseases. For each gene-disease pair, DisGeNET identifies the original publication reporting the association, along with all subsequent articles that further investigate it. I augment these data in two ways. First, I merge the list of articles introducing new associations with the GWAS Catalog, which records studies employing genome-wide methods. This allows me to identify which combinations resulted from a data-driven search process. Second, I merge the follow-on articles studying each association with NIH's PubTator3, which uses advanced AI models to classify the epistemic nature of scientific claims. This enables me to separately count subsequent studies that seek to elucidate how the gene is *causally* involved in the disease from those merely adding correlational evidence about the gene-disease combination. The former serves as my proxy for the development of new theoretical understanding.

To ground the discussion, consider again the two studies on the genetic basis of type 1 diabetes led by Polychronakos. In one, Qu et al. (2007), he leveraged theoretical knowledge about the IRF5 gene to target it with a candidate gene study. In the other, Hakonarson et al. (2007), he used an atheoretical data-driven approach that uncovered an association with the gene KIAA0350, whose function was unknown at the time and thus ignored by theory-driven studies. Unlike the IRF5-diabetes pair, the KIAA0350-diabetes combination involved an understudied gene, fell in the upper tail of scientific impact, and was highly unexpected given the lack of knowledge on the gene. It also triggered a flurry of follow-on research: 24 articles, mostly to elucidate how KIAA0350 is causally linked to diabetes, compared to just two for the IRF5-diabetes combination. That both studies were conducted by the same researcher, in the same year, and for the same disease underscores the central point: it is the nature of the search process that shapes the scope, quality, and unexpectedness of new gene-disease combinations, as well as the amount of downstream theorizing they generate.

My empirical analyses confirm the patterns illustrated in the case study. Compared to candidate-gene studies, gene-disease associations introduced by GWAS are 105–204% more likely to involve the least-studied human genes. GWAS findings skew more heavily toward both extremes of scientific quality, with a notable asymmetry: GWAS leads to a significantly higher rate of breakthroughs. Even among combinations by the same researcher, those identified via GWAS are 36% more likely to fall in the top tail of scientific importance. Moreover, GWAS discoveries are unexpected: they are less predictable based on prior knowledge, such as co-occurrence patterns of genes in other diseases, and more likely to be described as novel by researchers. Finally, difference-in-differences regressions show that GWAS-identified associations generate substantially more follow-on research, most of it aimed at

understanding causal mechanisms. Additional evidence suggests these findings stem from the removal of researcher choice in gene targeting, counteracting path dependency in research choices that prevent the discovery of new empirical anomalies.

This paper makes three contributions. First, it advances our understanding of how big data reshapes innovation search (Fleming, 2001; March, 1991; Schilling and Green, 2011). While prior work has highlighted the role of data in decision-making and investment selection (Brynjolfsson and McElheran, 2016; Kao, 2024; Nagaraj, 2022), I provide empirical evidence on the micro-mechanisms through which big data fuels the generation of novel ideas (Agrawal et al., 2024; Allen and McDonald, 2025; Lou and Wu, 2021). Second, the paper speaks to the emerging theory-based view in management research (Camuffo et al., 2024; Felin and Zenger, 2017). I extend this perspective by showing how theories function as focusing devices, enabling actors to recognize certain opportunities (Felin et al., 2024; Kaplan and Vakili, 2015) while also potentially blinding them to alternative paths (Chai, 2017). Third, my findings contribute to the epistemology of innovation by examining the interplay between theory and data (Dosi, 1982; Gittelman, 2016; Kuhn, 1962). Rather than displacing causal reasoning, the diffusion of new data technologies can complement theory development by uncovering empirical anomalies that stimulate higher-order search in the space of theories (Conti and Messinese, 2024; Mullainathan and Rambachan, 2024; Ott and Hannah, 2024).

The paper proceeds as follows. Section 2 outlines the conceptual framework. Section 3 illustrates the empirical setting. Section 4 describes the data and novel measurement approach, while Section 5 presents the empirical results. Section 6 discusses the findings.

# 2 Conceptual Framework

## 2.1 How Theory Shapes Variation and Selection in Search

The universe of potential innovations can be imagined as a vast combinatorial landscape across which innovators search for value peaks (Fleming, 2001; Levinthal, 1997; Schilling and Green, 2011). This search unfolds as an evolutionary process, typically involving two stages: variation and selection (Nelson and Winter, 1982; Knudsen and Levinthal, 2007). Innovators begin by generating or sourcing ideas for new technological combinations (Girotra et al., 2010; Piezunka and Dahlander, 2015; Singh and Fleming, 2010). Yet given constraints on time and resources, they do not pursue ideas at random; rather, they concentrate on those believed to offer the highest potential returns (Arora and Gambardella, 1994; Kneeland et al., 2020). Recombinant innovation thus involves making predictions about where to search, followed by a selection to prioritize the most promising options (Agrawal et al.,

2024; Schliesmann, 2025). Ultimately, the ability to find high-value combinations rests on an actor's capacity to generate and to recognize better predictions (Cohen and Levinthal, 1994; Tranchero, 2024).

Innovators who can make more accurate predictions are better positioned to prioritize valuable combinations and achieve superior performance. Recent research suggests that predictive success in search problems stems from the ability to formulate causal relationships between antecedents and outcomes (Felin et al., 2024; Sorenson, 2024). These causal structures—whether understood as mental models (Walsh, 1995), problem representations (Csaszar and Levinthal, 2016), abstract scientific principles (Arora and Gambardella, 1994), or more broadly, theories (Felin and Zenger, 2017)—enable actors to infer the potential value of a combination based on its ex ante observable features. Indeed, growing empirical evidence finds that the breadth, depth, and accuracy of such theories enhance foresight and improve decision-making in complex search problems (Camuffo et al., 2020; Gary and Wood, 2011; Heshmati and Csaszar, 2024; Kapoor and Wilde, 2024).

Theories relevant to a search problem shape both stages of the innovation process–variation and selection–by influencing how innovators generate and evaluate combinations to explore. On the one hand, cause-and-effect relationships enable forward-looking inference to formulate novel combinations (Sorenson, 2024). Theories guide combinatorial creativity by highlighting the components most likely to solve a given problem (Ehrig and Zenger, 2024). On the other hand, theoretical knowledge acts as a filter for evaluating ideas. This filtering can occur "online," by directing experimentation toward the most informative tests (Camuffo et al., 2024), or "offline," by enabling the assessment of combinations without the need for direct trial (Gavetti and Levinthal, 2000). As remarked by Popper (1963), "theory is a prohibition": by its very nature, theory defines boundaries around what is seen as plausible or acceptable. As a result, theory constrains the combinatorial space considered during both the generation and selection of combinations, ultimately lowering outcome variability while increasing efficiency (Fleming and Sorenson, 2004).

## 2.2 Data-Driven Search

In many domains, innovation takes place within technological spaces that have been increasingly well-charted through mapping efforts (Kao, 2024; Nagaraj, 2022). In drug discovery, for example, the boundaries of the chemical landscape are known (Jayaraj and Gittelman, 2018). However, the sheer size of this space means that existing theories can only cover portions of it, limiting the ability to find valuable combinations based solely on known principles (Gittelman, 2016). In such contexts, the growing availability of big data and predictive technologies provides an alternative approach to search (Agrawal et al., 2024). Innovators can now leverage large datasets to computationally predict

which technological combinations appear most promising, a process that can be described as *data-driven search*. Provided that the underlying data are not systematically biased or incomplete (Cao et al., 2024; Hoelzemann et al., 2025), data-driven methods can scan entire combinatorial landscapes in silico (Tranchero, 2024). In contrast to deductive reasoning, data-driven search automates a large-scale inductive approach (Kim, 2023; Wang et al., 2023).

In turn, the choice of search strategy likely shapes which regions of the technological landscape innovators explore. Theory-driven search directs exploration along well-established trajectories, where known cause-and-effect relationships constrain attention to areas defined by prevailing paradigms (Dosi, 1982; Kuhn, 1962). Variation is generated by recombining components that are more theoretically understood, while selection among potential options relies on established conceptual frameworks (Arora and Gambardella, 1994; Fleming and Sorenson, 2004; Katila and Ahuja, 2002). Data-driven search, by contrast, transforms both stages of the search process. In the variation phase, it enables innovators to generate combinations unconstrained by existing theory because signals are extracted directly from data without the need for causal understanding. In the selection phase, theoretical filtering is muted, since predictions emerge from statistical patterns rather than deductive reasoning. Barring systematic bias in the data, this atheoretical approach should broaden the scope of recombination efforts far from familiar or well-understood domains (Lou and Wu, 2021). These considerations lead to the following baseline hypothesis:

***Hypothesis 1:*** *Data-driven search leads to increased search breadth compared to theory-driven search*

Adopting a data-driven approach is likely to affect not just the direction of search, but also its yield. Prior research has found that expanding search breadth does not inherently guarantee better outcomes (Kaplan and Vakili, 2015). Rather, it is likely to affect both tails of the invention quality distribution (March, 1991). On one hand, increasing scope introduces more random variation, which can help escape competency traps and local optima (Knudsen and Levinthal, 2007; Levinthal and March, 1993). Breadth in variation has also been empirically linked to a higher chance of uncovering rare breakthroughs (Girotra et al., 2010; Singh and Fleming, 2010). On the other hand, data-driven predictions lack the quality filtering provided by theory because they are not grounded in causal reasoning (Felin et al., 2024). The flip side of increased random variation without selection is a higher likelihood of false positive leads (Tranchero, 2024). Which of the two effects dominates is unclear ex ante and likely depends on contextual factors, such as the accuracy of existing theories (Sorenson, 2024). Nevertheless, the overall effect of data-driven search should be to amplify the variability of outcomes by inflating on both tails of the quality distribution:

***Hypothesis 2:*** *Data-driven search increases the share of discoveries in both the bottom tail and the top tail of quality compared to theory-driven search*

A defining feature of data-driven search is that it functions as large-scale induction (Choudhury et al., 2021; Shrestha et al., 2021): it identifies patterns from the data without requiring prior theoretical assumptions. By mining vast datasets, innovators can uncover patterns that would be difficult or impossible to infer from existing knowledge alone (Mullainathan and Rambachan, 2024; Kim, 2023). Moreover, innovators working within a theoretical paradigm often tend to ignore or dismiss observations that contradict it (Kuhn, 1962), sometimes causing them to miss out on breakthrough discoveries (Chai, 2017). Inductive approaches are more open to novel patterns that defy expectations and would likely remain hidden when data are collected solely to test pre-specified theories (Gittelman, 2016; Ott and Hannah, 2024). By bypassing early theoretical filtering, data-driven search increases the likelihood of long jumps across the technological landscape—suggesting unconventional yet potentially high-impact recombinations (Levinthal, 1997; Shi and Evans, 2023). This line of reasoning can be synthesized as follows:

***Hypothesis 3:*** *Data-driven search increases the share of unexpected discoveries compared to theory-driven search*

## 2.3 From Empirical Anomalies to New Theory Generation

A growing strand of research explores the role of theories in entrepreneurship, innovation, and strategic decision-making (Felin et al., 2024). For instance, randomized controlled trials have shown that entrepreneurs perform better when they adopt a deductive, theory-based approach to experimentation (Camuffo et al., 2020). However, most of this work focuses on the performance implications of theories or mental models, leaving open the question of how such theories are developed in the first place. Said differently, where do cause-effect theories come from? The few existing perspectives tend to frame the emergence of new theory as a purely cognitive process. Felin and Zenger (2017) suggest that novel theories emerge when economic actors formulate contrarian conjectures about how to solve a problem. This can occur through analogical reasoning, borrowing causal structures from other domains with shared features (Gavetti et al., 2005; Kim, 2023). Alternatively, theories may stem from more idiosyncratic beliefs, emerging as initial theses later refined through reasoning and experimentation (Ott and Hannah, 2024).

Notably, these perspectives tend to downplay the role of data in generating new theoretical understanding. This stands in contrast to insights from the philosophy of science, which highlight how empirical

observations can fundamentally reshape paradigms (Chai, 2017; Kuhn, 1962). Theories impose boundaries that not only guide where to search but also define what outcomes are considered possible by logically narrowing the range of possibilities (Dosi, 1982; Popper, 1963). During phases of normal search, most empirical observations conform to these paradigm-induced expectations. However, the progressive discovery of empirical anomalies—observations that defy anticipated patterns—begins to expose the limits of the existing theory (Kuhn, 1962). Eventually, the accumulation of anomalies sparks a need for new explanations that can rationalize and explain them. The result is a period of theoretical ferment that may culminate in a paradigmatic shift and the formulation of a new theory.

One can conjecture that data-driven search sets in motion dynamics similar to those described in the philosophy of science. As argued in the previous section, actors who adopt a data-driven approach expand the breadth of their search and increase the variance in outcomes. Compared to findings derived through deductive reasoning, data-driven results are more likely to be unexpected and thus anomalous (Chai, 2017; Lou and Wu, 2021; Mullainathan and Rambachan, 2024). Empirical anomalies have a stronger "surprise effect" and greater potential to disrupt entrenched assumptions (Harrison and March, 1984; Ott and Hannah, 2024). When the evidence departs substantially from prior expectations, innovators will eventually update the prior belief that the existing theories still hold (Camuffo et al., 2024). The theory-free patterns uncovered in the data signal both a need and an opportunity to generate new causal understanding to account for them (Mullainathan and Rambachan, 2024). Taken together, this reasoning leads to the following hypothesis about the dynamic effects of data-driven search:

***Hypothesis 4:*** *Data-driven discoveries lead to more subsequent cause-and-effect theorizing compared to theory-driven discoveries*

Overall, this discussion sheds light on how big data technologies are changing the search for innovation. The ability to mine data in a theory-free manner does not necessarily displace theory, as some have suggested (Anderson, 2008). Rather, it could foster a virtuous cycle in which empirical anomalies give rise to new theoretical insights. The remainder of the paper tests whether the empirical evidence supports these hypotheses in the context of genetic research. In particular, I investigate how the shift from a theory-driven to a data-driven search approach alters both the discovery process and the development of new theories about the genetic roots of human disease.

# 3 Empirical Setting

## 3.1 Genetic Research Before GWAS

Genes are sequences of DNA bases that encode the "instructions" for synthesizing molecules fundamental for the organism's functioning. Genes often harbor DNA mutations, some of which can drive the emergence of diseases. Most common conditions, such as diabetes or hypertension, are polygenic: they result from mutations across multiple genes and their interaction with environmental factors (Boyle et al., 2017). Understanding the genetic roots of such diseases is critical because causative genes can become targets for therapeutic intervention. However, identifying the genes involved in thousands of polygenic diseases requires searching a massive combinatorial space of over $\sim 19,000$ known human genes.

Scientists have traditionally navigated this landscape using a *candidate gene approach*, which involves three main steps. First, they choose a disease to study, often guided by its prevalence or the availability of funding. Second, they hypothesize which genes might play a role in the disease's etiology. Third, they test these hypotheses using a case-control design: namely, by examining whether subjects with the disease are more likely to carry mutations in the candidate genes than control subjects. Importantly, the selection of candidate genes reflects scientists' theoretical understanding of which genes are likely to matter for the disease, and why. For example, after the IRF5 gene was associated with lupus, Qu et al. (2007) drew on their knowledge of autoimmune diseases to hypothesize that the same gene could also play a role in type 1 diabetes. This led to a candidate gene study that first explored the IRF5-type 1 diabetes combination (Panel (a) of Figure 1).

Candidate gene studies exemplify well the deductive model of scientific inquiry. Yet despite notable successes, these studies inherently steer scientists toward genes for which functional hypotheses can be formulated based on existing knowledge (Hoelzemann et al., 2025). This has led to an extreme concentration of attention on a small subset of theoretically well-understood genes (Stoeger et al., 2018). The focus on exploiting a handful of "superstar" genes might be justified by the risks involved in exploring the remaining gene pool, which likely contains many dead ends. Still, the lack of broader exploration across genetic space is arguably suboptimal since the chances of finding treatments for polygenic diseases would improve by casting a wider net (Uffelmann et al., 2021).

## 3.2 The Emergence of GWAS

In the early 2000s, two major developments converged to offer an alternative to candidate gene studies. First, the completion of reference genomes such as the Human Genome Project and the HapMap

Project provided a comprehensive map of the human genetic landscape. Second, the cost of collecting genetic data dropped dramatically (Appendix Figure F.1). Together, these advances paved the way for the emergence of *genome-wide association studies* (GWAS) (Visscher et al., 2017). As the name suggests, GWAS are case-control studies that use genetic data to identify mutations associated with disease across the entire genome.

Like candidate gene studies, researchers conducting a GWAS begin with the selection of a disease to study and then look for mutations that appear more frequently in individuals with the condition than in those without it. However, the two approaches diverge sharply in how genetic targets are selected (Uffelmann et al., 2021). Rather than focusing on specific gene-disease pairs informed by existing theory or prior evidence, GWAS rely on a fully data-driven approach. Researchers collect DNA samples from large numbers of cases and controls and systematically test for differences between their genetic make-up in every genetic location. Genes found to harbor mutations more common in cases than in controls are flagged as potentially implicated in the disease's etiology, regardless of whether their functional role is understood. As such, they become candidates for further investigation or pharmaceutical intervention, even in the absence of a clear causal explanation (Tranchero, 2024).

For example, consider the GWAS conducted by Hakonarson et al. (2007). The researchers analyzed a study population of 563 patients with type 1 diabetes and 1,146 controls without the condition. Using genotyping microarrays, they collected data about the genetic features of their study subjects. The comparison between cases and controls highlighted several significant differences across the genome: some in genes already known to be linked to diabetes (such as the insulin gene INS), but others in less familiar regions, including the gene KIAA0350 (Panel (b) of Figure 1). At the time, the function of KIAA0350 was unknown, which likely explains why it had been overlooked in candidate gene studies. Since the publication of Hakonarson et al. (2007), the KIAA0350–type 1 diabetes association has proven highly impactful and has been the focus of multiple follow-up studies and clinical trials. Additional details on this case are provided in Appendix B.

GWAS represent a prime example of data-driven search (Visscher et al., 2017). A genome-wide search scans the entire set of human genes for associations with a given disease, directly highlighting promising gene-disease combinations (Panel (c) of Figure 1). Unlike candidate gene studies, where researchers select targets based on theoretical considerations, GWAS are atheoretical and generate discoveries purely from patterns that emerge in the data. Yet, this very feature also makes GWAS a source of ongoing controversy within the scientific community, even years after their diffusion (Callaway, 2017). GWAS produce high rates of false positives (Tranchero, 2024), and they cannot

explain the underlying biological mechanisms of the associations uncovered (Boyle et al., 2017). Moreover, achieving sufficient statistical power across thousands of genetic loci requires large sample sizes, making GWAS much more expensive than targeted gene-candidate studies. The large cost of data collection posed a significant barrier to their diffusion,[1] leading some skeptics to dismiss GWAS as a costly detour in the quest for genetic insight (Loos, 2020). This is why candidate gene approaches remain in use and questions persist about the true value of data-driven science in genetics.

# 4 Measurement Strategy and Data

## 4.1 Defining the Search Landscape: Gene-Disease Combinations

While search is often conceptualized as occurring within a technological landscape, such landscapes are typically difficult to study empirically (Fleming and Sorenson, 2004; Schilling and Green, 2011). This paper leverages the setting of genetic research to address that challenge. A central goal in genetics is to identify which human genes could serve as drug targets for a given disease. In this context, genes and diseases constitute the relevant components of the search problem. Any gene can, in principle, be tied to any disease, creating a well-defined space of millions of potential gene-disease combinations (see Appendix C.1 for an illustration). Importantly, with the completion of the Human Genome Project in the early 2000s, the universe of over 19,000 protein-coding genes has been mapped. This allows the entire landscape of possible gene-disease combinations to be characterized separately from how search unfolds over it.

I obtain data on links between human diseases and their genetic causes from DisGeNET v7.0 (Piñero et al., 2020), an aggregator of scientific sources considered one of the most comprehensive repositories of genetic knowledge (Hermosilla and Lemus, 2019). The database compiles gene-disease associations (GDAs) from specialized sources, including curated datasets and publications indexed in PubMed. The units of observation are gene-disease pairs first introduced in scientific articles published between 2005 and 2016. For each association, I assemble both the first publication that reported it and all subsequent articles that investigated it.[2] I focus on associations that link a protein-coding gene to a disease, syndrome, or abnormality with clear health implications. The final dataset includes 369,302 newly introduced gene-disease combinations, encompassing a total of 14,136 genes and 9,863 disease

[1] Panel (b) of Appendix Figure F.1 shows how the progressive diffusion of GWAS is strongly correlated with the decrease of sequencing costs. For a discussion of early GWAS costs, see: *www.blog.goldenhelix.com/have-we-wasted-7-years-and-100-million-dollars-on-gwas*.

[2] Information on the bibliographic characteristics of papers introducing at least one novel gene-disease combination is taken from NIH's iCite data. To identify the principal investigator (PI), I obtain information on the last author of each publication from the Author-ity database (Torvik and Smalheiser, 2021).

categories.[3]

## 4.2 Data-Driven and Theory-Driven Search Processes

Empirically investigating how alternative search strategies shape innovation requires linking the search process to its resulting outputs. But this poses a fundamental challenge: researchers typically observe only the outcomes, without direct information about the search process that produced them (Kneeland et al., 2020; Maggitti et al., 2013). A common approach has been to infer search strategies from observable features of outputs. For example, the breadth of a paper's references or the diversity of a patent's USPC classes have been used as indicators of how inventors searched (Fleming, 2001; Schilling and Green, 2011). The problem is that these proxies are inherently ex-post and prone to measurement error. Therefore, the goal is to find a setting where the search strategy adopted is known without conditioning on observing successful outcomes.

In the context of genetics, I can address this issue because the search for therapeutically valuable combinations happens in two distinct ways. Scientists studying a disease can carry out candidate gene studies using their theoretical priors to target specific genes. Alternatively, they can use atheoretical GWAS, which scan the entire genome to identify genetic mutations correlated with the disease. Candidate gene studies vary in how targeted they are, ranging from testing a single gene-disease pair (as in the paper of Qu et al., 2007) to examining broader regions that may contain hundreds of genes. GWAS, by contrast, involve no genetic targeting at all. Crucially, I can infer which search process led to a given gene-disease combination by identifying the method used in the paper that first introduced it. This information comes from the GWAS Catalog, a manually curated database of all studies that adopt a genome-wide approach (MacArthur et al., 2017). Studies are eligible for inclusion in the GWAS Catalog only if they use a DNA microarray to scan the entire genome without targeting any specific gene. In total, the DisGeNET data include 8,655 gene-disease combinations introduced by 1,375 distinct GWAS, while the remainder of the sample comes from studies involving genetic targeting.[4]

## 4.3 Characteristics of Gene-Disease Combinations

I begin by examining the cross-sectional features of new combinations introduced through a data-driven process. Specifically, I ask: how do gene-disease associations established by GWAS differ

[3] Since the analysis focuses on comparing GWAS and candidate gene studies, the sample is restricted to new gene-disease combinations introduced on or after 2005 (the year of the first GWAS). As a result, the number of genes in the dataset is lower than the total number of human genes, as some were not implicated in any disease during the sample period.

[4] Following best practices, the curators of the GWAS Catalog use the Bonferroni correction and report as significant only associations with a high statistical significance (p-value $< 1.0 \times 10^{-5}$).

from those established by candidate gene studies? I address this question by focusing on dependent variables that correspond to the first three hypotheses outlined in Section 2.2.

**A. Search scope:** I use two alternative dependent variables to capture discoveries involving genes that had received scant attention before the emergence of GWAS. The first is a dummy variable that equals one if the gene-disease association involves a gene that had never been linked to any disease before. Returning to the example from the introduction, the KIAA0350 gene is coded as underexplored because it had not been associated with any disease prior to the study by Hakonarson et al. (2007). The second proxy is the gene's discovery date, since many genes mapped more recently are still relatively less understood (Stoeger et al., 2018). Accordingly, I test whether GWAS are more likely to introduce combinations involving genes discovered after 2000, the year the first draft of the human genome was released. For example, the KIAA0350 gene was discovered in 2002, but remained ignored until Hakonarson et al. (2007) linked it to type 1 diabetes in 2007. Both dependent variables are coded as dichotomous to allow for straightforward interpretation of the OLS coefficients as linear probability models. I also replicate all analyses using continuous versions of these variables in the Appendix.

**B. Quality of combination:** Researchers typically rely on paper-to-paper citation counts to measure impact, but that approach would be inappropriate in my setting since scientific articles often report multiple gene-disease associations. For example, the GWAS by Hakonarson et al. (2007) reported associations between type 1 diabetes and five genes (KIAA0350, INS, COL1A2, LPHN2, and PTPN22), making it impossible to assign a clear share of the article's citations to each individual combination. To avoid this limitation, I use a unique metric available at the combination level: the DisGeNET Score (Piñero et al., 2020). The DisGeNET Score is a composite index that reflects the ground truth scientific value of each gene-disease pair (see Appendix C.2 for details). Supporting its validity, Appendix Table C.1 shows that articles introducing combinations with a higher Score receive a larger number of citations. Furthermore, gene-disease combinations with higher DisGeNET Scores have substantially greater clinical relevance and downstream innovation impact (Appendix Table C.2). This metric allows me to assess how search methods shape the average quality of gene-disease combinations, as well as the likelihood of generating discoveries in both the top and bottom tails of value.

**C. Unexpectedness of combination:** I define empirical anomalies as gene-disease combinations that are surprising given the state of existing knowledge (Shi and Evans, 2023). Measuring how unexpected a finding is becomes feasible in this context because genes do not operate in isolation in the human body, but rather through chains of activation and inhibition. As a result, functionally linked genes tend to co-occur as associated with human diseases, since a mutation in any one of them is enough

to disrupt a disease-relevant biological process (Visscher et al., 2017). For any given disease, the associations up to a point in time can thus be used to forecast the likelihood of future ones based on gene co-occurrences in other diseases. I use the inverse of the predicted probability of a gene being the next one associated with a disease to build an Unexpectedness Index (see Appendix D for details). To validate this index, I draw on lexical cues from the abstracts of scientific papers (Mishra et al., 2023). Appendix Table D.1 shows that papers introducing unexpected combinations are more likely to describe their findings as "novel," but not more likely to use other words denoting generic hype. Appendix Table D.2 further confirms that these papers are rated as presenting new findings by expert reviewers on the Faculty Opinions platform.

## 4.4 Measuring New Theory Generation

Next, I examine the dynamic effects of combinations introduced through a data-driven process. Specifically, I ask: do gene-disease associations established by GWAS spur more follow-on work than those established by candidate gene studies? To answer this question, I need a metric that captures downstream developments building on a specific finding. Traditionally, researchers have relied on citations to the article where the finding is reported to measure its subsequent impact. However, this would be misleading in my setting, since citations are a proxy of attention and not necessarily of scientific quality (Bikard, 2018). GWAS papers are, on average, highly cited ($Cites_{GWAS}$ = 170 vs. $Cites_{Candidate\,Gene\,Study}$ = 41), but much of that attention comes from factors unrelated to the quality of the findings—such as reviews, critiques, or debate on the merits of the genome-wide approach. Instead, I construct a more targeted measure of scientific importance: the number of papers that *directly* build on the newly introduced gene-disease combination. This includes any follow-on study that investigates the association itself, regardless of whether it cites the original paper. As such, it provides a direct measure of impact at the level of the individual combination without the known shortcomings of bibliometrics (Arts et al., 2025).

Using the data from DisGeNET, I construct a panel dataset at the gene-disease-year level. Each observation captures the number of articles investigating a given gene-disease combination in a particular year, excluding the paper that originally introduced it. For example, Appendix Figure B.2 shows that the KIAA0350–type 1 diabetes association was investigated by 24 articles following the GWAS by Hakonarson et al. (2007). While many of these studies cite the original GWAS, some do not and would be missed by traditional citation-based measures of impact. By contrast, the same figure shows that the candidate gene study by Qu et al. (2007) generated little follow-up work on the IRF5–type 1 diabetes pair.

Finally, I leverage new data from NIH's PubTator3 to further characterize the nature of follow-on research (Wei et al., 2024). PubTator3 uses an AI-powered engine to automatically classify the type of gene-disease relationship studied in a paper, based on the article's text. In particular, PubTator3 can identify whether a study investigates a causal relationship between the gene and the disease (see details in Appendix E). I use this information to separately count follow-on articles that explore causal mechanisms—thereby contributing to new theoretical understanding—versus those that merely add correlational evidence. Supporting the face validity of the AI-based classification, Appendix Table E.1 shows that articles reporting clinical trial results are more likely to be classified as studying causal rather than correlational relationships in the PubTator3 data. Returning to the example of type 1 diabetes, Appendix Figure B.2 shows that a large share of the papers examining the KIAA0350–type 1 diabetes link did so in order to investigate its causal basis. While prior work in management has largely inferred recombination from the co-occurrence of components in a publication or patent, PubTator3 allows me to go further by identifying the *type* of relationship under study, and specifically whether the work is aimed at building causal understanding.

### 4.5 Summary Statistics

Table 1 lists the key variables along with summary statistics for the sample used in the analysis. Panel A presents statistics at the publication level, focusing on papers that introduce new gene-disease combinations during the study period. Compared to candidate gene papers, GWAS tend to introduce more associations and cover a larger number of genes on average, consistent with their untargeted nature. Panel B provides summary statistics at the gene-disease combination level. Previewing the following analysis, GWAS and candidate gene studies introduce new discoveries with similar average DisGeNET Scores. However, GWAS combinations are more likely to appear in both the bottom and top deciles of the Score distribution, indicating greater variability in their scientific value. The Unexpectedness Index for GWAS combinations is nearly five times higher than for those introduced by candidate gene studies. Panel C provides summary statistics at the panel level, where the unit of analysis is the gene-disease-year. Combinations introduced by GWAS attract more follow-on research, much of which focuses on clarifying the causal relationship between the gene and the disease.

## 5 Results

### 5.1 The Impact of Data on Genetic Search

In this section, I begin by examining how gene-disease combinations introduced through a data-driven search process differ from those introduced via a theory-driven one. I use regression analysis to

compare the outcomes of the two search strategies, controlling for the characteristics of both diseases and scientists. More specifically, I estimate the following cross-sectional OLS model using gene-disease level data:

$$(Features\ of\ gene\text{-}disease\ combination)_{i,j} = \alpha + \beta GWAS_{i,j} + \gamma Disease_i + \omega Scientist_j + \delta Year_{i,j} + \epsilon_{i,j} \quad (1)$$

where various features of the gene-disease pair are regressed on the indicator variable $GWAS_{i,j}$, which equals one for combinations introduced by a GWAS and zero for those introduced by candidate gene studies. While cross-sectional and descriptive, this specification accounts for several potential confounding factors by including fixed effects for targeted disease $i$, principal investigator $j$,[5] and year of first appearance. If data-driven search systematically leads to different outcomes, I should observe a statistically significant estimate of $\beta$. All specifications cluster standard errors at the disease level.

Figure 2 offers an intuitive visualization of the combinatorial space of all potential gene-disease associations. Candidate gene studies tend to build on existing knowledge, with the result of replicating familiar research patterns. In contrast, Panel (b) shows that for each disease investigated, GWAS explore a much broader range of genes, spanning the entire genome.[6] Regressions in Table 2 quantify these differences. The estimate of $\beta$ in Column 1 shows that GWAS increase by 26 percentage points the probability of combining a gene never associated with any disease before 2005, compared to a baseline of around 13 percentage points. Column 3 confirms the result using the gene's date of discovery as an alternative proxy. Even after controlling for scientist fixed effects, GWAS are 105–147% more likely to recombine historically understudied genes.[7] Taken together, the evidence supports Hypothesis 1: data-driven search expands the breadth of combinatorial exploration.

While data-driven search appears to broaden the scope of search, there is no guarantee that the resulting combinations are valuable. If anything, one might expect the opposite: scientists should already be focusing on what they consider the most promising areas of the technological landscape,

[5] My estimates would be biased if scientists conducting GWAS were systematically different from those who do not. In other words, I need to ensure that observed differences in outcomes reflect the search process itself, not the ability or resources of the scientist. One way to address this is by comparing gene-disease associations introduced by the same researcher, leveraging the fact that many scientists have used both approaches over the course of their careers. Including principal investigator fixed effects allows me to isolate the effect of the search method, holding constant all time-invariant characteristics of the scientist's laboratory. Furthermore, Appendix Figure F.3 shows that scientists adopting the GWAS approach are not statistically different in the average quality of their preceding discoveries, allaying concerns that sorting might be biasing my estimates.

[6] Figure 2 also shows that GWAS continue to focus on historically well-studied diseases (see also Figure F.2 in the Appendix). This is because the choice of disease remains in the hands of the principal investigator, regardless of the search strategy. Importantly, this further confirms that the diversification in gene space is due to the search strategy itself, which removes the researchers' discretion over which genes to target, rather than by a shift in disease focus.

[7] Appendix Table F.1 shows consistent results using continuous versions of the dependent variables. GWAS are more likely to recombine genes that had received 29-51% fewer publications before 2005, and that were discovered 2.1-3 years later.

making it less likely that breakthrough discoveries have been overlooked. I examine this question in Table 2. Columns 1 and 2 show that gene-disease combinations introduced by GWAS have, on average, a DisGeNET Score that is 19–20% higher. Interestingly, this average effect is driven by a thickening of both tails: GWAS combinations are more likely to be either failures or breakthroughs. This pattern is consistent with the idea that data-driven search increases variation generation while reducing selection, with the consequence of increasing variability in outcomes. Notably, the increase in top-tail combinations is twice as large, which explains the overall rise in average DisGeNET Score for GWAS-introduced combinations. These results hold under alternative thresholds for defining the tails of the DisGeNET Score distribution (Appendix Tables F.2 and F.3).[8] Taken together, the evidence is consistent with Hypothesis 2.

Finally, I test whether gene-disease combinations uncovered by GWAS are more likely to be unexpected empirical anomalies. Columns 1 and 2 of Table 4 show that associations introduced by GWAS are significantly more likely to be surprising given existing genetic knowledge. The effect is large relative to the sample mean and only slightly attenuated after controlling for principal investigator fixed effects. Unexpected findings could result either from the exploration of entirely new genes or from linking well-known genes to previously unassociated diseases. To distinguish between these alternatives, I re-estimate the model excluding combinations involving genes never previously recombined (such as the KIAA0350 gene before the GWAS of Hakonarson et al., 2007). Columns 3 and 4 show that the effect remains large and highly significant, though reduced in size. This suggests that most of the surprise stems from uncovering data signals tied to genes that had not been studied before. The results are robust when using an alternative text-based measure of anomaly, based on whether the findings are described as "novel" in paper abstracts (Appendix Table F.6). Taken together, the evidence offers strong support for Hypothesis 3.

## 5.2 The Interplay Between Empirical Anomalies and Theory Generation

A key empirical challenge in studying the effect of new findings on subsequent theory generation is endogeneity. Since scientists do not search at random, new findings appearing in the scientific literature entail topics that would likely attract attention anyway, leading to upwardly biased estimates. In my setting, however, this endogeneity concern is minimized because GWAS scan for genetic variants across the entire genome (Uffelmann et al., 2021). By design, this method removes selection at the gene level, conditional on the choice of disease. As a result, GWAS discoveries are entirely unforeseen by

[8]Further supporting the results, the coefficients grow larger with more stringent definitions of failure or breakthrough, suggesting that GWAS are especially powerful at uncovering outlier combinations at both tails.

researchers–a claim further supported by the earlier cross-sectional findings using the Unexpectedness Index. The staggered appearance of these unexpected GWAS findings offers a natural identification strategy that allows me to identify their causal effect on subsequent scientific research (Tranchero, 2024).

Figure 3 shows descriptive evidence on the average number of publications following the introduction of a new gene-disease combination. Two patterns stand out. First, GWAS trigger roughly three times as many follow-on papers on the newly introduced combinations compared to candidate gene studies. Second, the majority of such papers are aimed at elucidating causal mechanisms, which is not the case for associations introduced with theory-based reasoning. To formally quantify these differences, I estimate the following panel difference-in-differences specification at the gene-disease-year level:

$$(Papers\ on\ gene\text{-}disease\ combination)_{i,j,t} = \alpha + \beta PostPublication_t \times GWAS_{i,j} + \lambda GD_{i,j} + \delta_t \times Disease_i + \omega_t \times Gene_j + \epsilon_{i,j,t} \quad (2)$$

where the count of articles exploring a given gene-disease pair $< i, j >$ is regressed on the interaction of two indicator variables: $PostPublication_t$, which equals one in the years following the initial publication of the gene-disease combination, and $GWAS_{i,j}$, which equals one if the association was first reported in a GWAS. $GD_{i,j}$ are fixed effects for the combination of gene $i$ and disease $j$, which account for pair-specific differentials in research potential. I include fixed effects for each gene-disease pair ($GD_{i,j}$) to account for inherent differences in the research potential of specific combinations. To control for time-varying disease-level shifts in market size, I include disease-year fixed effects ($\delta_t \times Disease_j$). I also account for changes in gene-specific interest over time by including gene-year fixed effects ($\omega_t \times Gene_i$). Standard errors are clustered at the disease level.

Column 1 of Table 5 shows that gene-disease combinations introduced by GWAS are followed by an average of 0.15 additional publications per year compared to those introduced by candidate gene studies, an effect that represents a +319% increase over the sample mean. Columns 2 and 3 break this down further by separating follow-on articles into those studying causal relationships and those contributing correlational evidence about the focal gene-disease pair. Strikingly, the increase in publications is driven almost entirely by work aiming to uncover causal mechanisms. Event study versions of Equation 2 support these findings and rule out the presence of pre-trends, confirming the validity of the identification strategy. Figure 4 plots the annual increase in follow-on research for GWAS-introduced associations relative to those from candidate gene studies. Panel (a) confirms the stronger post-discovery surge in scientific interest for GWAS combinations. When breaking down the nature of follow-on work, panel (b) shows that the increase is largely due to causal investigations.

Panel (c), by contrast, reveals no difference in the number of correlational studies following GWAS versus candidate gene studies. Overall, the results support Hypothesis 4 and confirm that empirical anomalies play a key role in generating new theoretical understanding of genetic mechanisms.

## 5.3 Why Do Scientists Underexplore?

The evidence presented in the previous section supports the conceptual framework introduced in Section 2. Still, the continued focus of candidate gene studies on familiar genes is puzzling given the large gains from exploring lesser-known ones. In this section, I provide suggestive evidence for two possible drivers of this underexploration. First, researchers may repeatedly focus on a narrow set of genes simply because they have better tools to study them (Furman and Teodoridis, 2020). This aligns with prior work showing that both lab equipment (Baruffaldi and Gaessler, 2025) and the availability of data (Hoelzemann et al., 2025) can shape the direction of research. In particular, gene-specific tools may lower the cost of investigation, encouraging search in areas of diminishing returns. Second, theory-driven learning is inherently cumulative and path-dependent, as captured by the Newtonian notion of "standing on the shoulders of giants". Over time, this can lead scientists to develop specialized human capital that locks them into familiar search trajectories, leading them to neglect exploration of new targets (Arts and Fleming, 2018; Levinthal and March, 1993).

To investigate the first possibility, I identify which human genes have a homolog in the laboratory mouse. Homologs are genes shared by two species through a common ancestor that retain similar biological functions. This allows researchers to study the role of homologous human genes through experiments on their animal counterparts. Because the mouse is the most widely used lab model, genes without mouse homologs are more difficult to study experimentally and may be overlooked for practical rather than scientific reasons (Stoeger et al., 2018).[9] If limited research tools were the main reason scientists focus on a narrow set of genes, GWAS findings should be more likely to involve genes without mouse homologs than candidate gene studies. However, Panel (a) of Figure 5 and Appendix Table F.7 show this is not the case, suggesting that tool availability is not the primary barrier to exploration.

To examine the second conjecture, I collect data on whether each gene belongs to a gene family. Gene families arise through duplication of a common ancestral gene and typically share similar biochemical functions (Daugherty et al., 2012). These genes share a root name and are numbered in the order of discovery (e.g., BRCA1 and BRCA2, discovered in 1994 and 1995, respectively). Differences in sci-

[9]This intuition is supported by my data, since genes lacking a mouse homolog have 22% fewer publications linking them to human disease on average.

entific attention between members of the same family often reflect the path dependence of researchers focusing on genes identified earlier, rather than any inherent difference in scientific potential (Stoeger et al., 2018).[10] Panel (b) of Figure 5 shows that candidate gene studies disproportionately recombine the first-discovered member of a gene family. In contrast, this bias disappears in combinations introduced by GWAS. This suggests that one way GWAS advances discovery is by breaking the inertia of established search trajectories that result in low exploration.

## 5.4 Robustness Checks

The findings from the previous sections suggest that it is *how* scientists search that shapes both the characteristics of the gene-disease combinations introduced and their downstream impact. Still, it remains possible that other channels contribute to the results. The aim of this section is to rule out a few alternative explanations.

**A. Scientists Publishing GWAS:** The inclusion of principal investigator fixed effects allows for the estimation of within-researcher coefficients, capturing changes in outcomes for the same scientist after switching search strategies. However, it is still possible that researchers who are systematically better or worse at identifying breakthroughs sort into adopting GWAS, potentially influencing the estimates.[11] Appendix Figure F.3 helps address this concern by showing that, prior to their first GWAS, principal investigators who eventually adopt a data-driven approach do not introduce gene-disease combinations of systematically different quality.

**B. Alternative Fixed Effect Structures:** Appendix Table F.4 presents a robustness check of the main cross-sectional analyses using more stringent fixed effect specifications. In Panel A, I include indicators for each disease $\times$ year pair, which isolates variation across gene-disease combinations linked to the same disease and introduced in the same year. This controls for time-varying, disease-specific factors that could confound the estimates. Panel B adds tight disease $\times$ PI $\times$ year fixed effects, thus isolating variation across gene-disease combinations introduced by the same laboratory, for the same disease, in the same year. The results remain robust, with the partial exception of a single coefficient, likely due to lower statistical power from the very restrictive specification.

[10] Supporting this idea, DisGeNET data show that the first gene in a family receives, on average, 49% more publications than the second. Yet, associations involving the second gene tend to have higher average DisGeNET Scores ($DisGeNET\ Score_{First\ Member}$=0.059 vs. $DisGeNET\ Score_{Second\ Member}$=0.063) and are more likely to fall in the top 10% of all associations ($Top\ 10\%\ Score_{First\ Member}$=0.098 vs. $Top\ 10\%\ Score_{Second\ Member}$=0.112).

[11] The direction of this bias is difficult to determine ex ante. If scientists adopt GWAS because they struggle to find valuable combinations through theory-driven approaches, the bias would be downward. If, instead, it is the most capable scientists who are early adopters of GWAS, the bias would point upward. In either case, note that establishing the causal impact of adopting GWAS on individual researcher performance is beyond the scope of this paper.

**C. Controlling for Gene Fixed Effects:** When comparing discoveries from data-driven versus theory-driven search, one concern is that genes may differ in their ex ante potential.[12] If the genes recombined by GWAS are inherently more likely to yield high-value or surprising associations, then the effects documented under Hypotheses 2 and 3 could be a mechanical consequence of the broader search scope discussed in Hypothesis 1. However, Appendix Figure F.4 shows that GWAS-introduced combinations are more likely to be both high-quality and unexpected even after controlling for gene fixed effects. The only partial discrepancy is that GWAS are no longer more likely to introduce combinations in the bottom tail of quality, suggesting that the increase in low-quality discoveries is due to between-gene variation (and that theory-driven search correctly avoids some of the least promising genes). This result suggests that data-driven search contributes to innovation not only by including previously short-changed components, but also by uncovering overlooked patterns within well-known components.

**D. Publication Bias:** It is well documented that studies reporting negative results face lower chances of publication. This raises a concern for my comparative analyses: if the "file drawer problem" is more severe for candidate gene studies relative to GWAS (as some evidence suggests, see Duncan and Keller, 2011), then my findings could partly reflect selective reporting rather than true differences. To address this, I draw on findings from Fanelli (2010), who shows that studies testing multiple hypotheses are more likely to report at least one null result. This pattern suggests that failures are more likely to appear in the literature when they can be bundled alongside positive results. Building on this logic, I replicate the main analyses limiting the sample to studies that introduce more than one gene-disease association, hence being less likely to suffer from selective reporting.[13] Appendix Table F.5 shows that the results remain largely robust, though somewhat noisier due to the smaller sample.

**E. Heterogeneity by Genetic Knowledge:** Prior research suggests that the impact of empirical anomalies on theory generation is likely shaped by the state of existing knowledge. In domains where theoretical understanding is more developed, the stronger filtering of variation may lead researchers to become stuck on local optima (Knudsen and Levinthal, 2007). At the same time, the stronger the belief in prevailing theories, the greater the surprise when data reveal anomalous patterns (Camuffo et al., 2024). Appendix Figure F.5 presents evidence consistent with these ideas: GWAS findings

[12] For instance, TP53 plays a critical role in regulating cell growth, and mutations in this gene are found in over half of tumor sequences (Stoeger et al., 2018). Given its central biological function, associations involving TP53 are more likely to be scientifically important and less likely to be unexpected, whereas the opposite might hold for neglected genes like KIAA0350.

[13] Consistent with Fanelli (2010), publication bias appears more pronounced among candidate gene studies, which typically test fewer hypotheses. Among studies reporting only one gene-disease association, 15.6% of candidate gene studies show findings in the bottom tail of the DisGeNET score, compared to 20.4% for GWAS. When multiple associations are introduced, these shares rise to 36.2% and 49.7%, respectively.

stimulate more follow-on research in disease areas with greater the pre-existing genetic knowledge. Notably, these effects are driven by studies investigating causal mechanisms, providing further support for the main results of the paper.

# 6 Discussion

This paper examines how data reshapes search in the context of genetics research. By comparing GWAS with candidate gene studies, I exploit a natural contrast in search strategies to assess their impact on innovation outcomes. The empirical results show that GWAS are more likely to introduce gene-disease combinations in underexplored areas of the genetic landscape. While theory-driven search tends to filter out low-quality outcomes, GWAS increases the incidence of outliers at both ends of the quality spectrum, ultimately raising average quality. Notably, GWAS findings often break from established patterns of genetic co-occurrence, uncovering surprising empirical anomalies. Difference-in-differences regressions show a sharp rise in scientific attention following these discoveries, with follow-on research largely focused on investigating their causal mechanisms.

My results underscore the distinct nature of inductive and deductive search, along with their nuanced implications. On the one hand, data-driven search can generate false positives and low-value leads. In strategic decision-making, where resources are limited and choices often irreversible, pursuing the wrong lead can seriously undermine performance (Tranchero, 2024). Yet on the other hand, those same leads may serve a different function: redirecting exploration toward more open-ended and potentially fruitful directions. While a high rate of false positives poses clear risks in strategy, it may be tolerable or even desirable in innovation. To paraphrase Cohen and Levinthal (1994), "fortune favors the curious firm," and data-driven leads can be exactly what sparks that curiosity. In contrast, false negatives are particularly harmful in innovation contexts, where prematurely dismissing a promising path can halt exploration altogether (Chai, 2017; Fleming and Sorenson, 2004; Hoelzemann et al., 2025). This underscores the asymmetric consequences of false positives and false negatives when using data to explore versus using data to make strategic decisions.

An important contribution of this paper is methodological. While the modeling of innovation search has a storied tradition (Levinthal, 1997; March, 1991; Nelson and Winter, 1982), empirical progress has lagged somewhat behind. Part of the challenge lies in the difficulty of studying a largely unobservable process through the features of its realized outcomes (Fleming, 2001; Kneeland et al., 2020). This paper addresses that challenge by showing how deep institutional knowledge enables the identification of the search processes agents actually follow. In turn, this ex-ante characterization can be used to explain how different search strategies shape outcomes on the empirical landscape—linking the search

process to the innovations it produces. Broader adoption of this approach could meaningfully advance the empirical study of search.

While the evidence supports the theoretical framework, three boundary conditions are worth noting. First, although empirical anomalies can spark theoretical innovation, the value of such anomalies depends on whether innovators pay attention to them. In an age of data deluge, the challenge shifts from identifying leads to prioritizing them, making attention a scarce and strategic resource (Bikard, 2018). If not managed well, an excess of data-driven variation can overwhelm researchers with noise and paradoxically hinder exploration (Piezunka and Dahlander, 2015). Second, data-driven signals can mislead if the data available are biased (Cao et al., 2024) or incomplete (Hoelzemann et al., 2025). Still, this constraint is gradually easing as more fields gain access to cheap and comprehensive datasets. Just as the Human Genome Project mapped the full genome, large-scale efforts are now charting innovation spaces in areas as diverse as chemistry, proteomics, earth sciences, and astronomy (Agrawal et al., 2024; Kao, 2024; Nagaraj, 2022). Third, theory plays an essential role in structuring the search space itself by defining what is worth measuring, tracking, or exploring (Dosi, 1982). Data can only guide innovation when grounded in clearly framed problems, something ultimately made possible by theory. This further underscores the central argument of the paper: the inherent complementarity between theory and data in driving innovation.

In many ways, the history of science and technology is shaped by episodes similar to those documented in this paper. For instance, Galileo's telescope enabled the observations that fueled the Copernican Revolution, and Rosalind Franklin's X-ray diffraction images prompted the discovery of the structure of DNA. In both cases, new data revealed anomalies that challenged prevailing theories and spurred conceptual breakthroughs. These examples reflect the enduring complementarity between theory and data. What makes the findings of the present paper especially timely is the rapid advancement and diffusion of big data technologies, particularly AI. These tools can dramatically accelerate the identification of empirical anomalies (Conti and Messinese, 2024), and with them, the pace of theoretical innovation. The true promise of AI, however, lies not only in its ability to exhaustively search existing landscapes but also in its potential to redefine those landscapes, opening up new ways of framing problems that were previously ill-posed. In this light, big data may not mark the "end of theory," as Anderson (2008) once proclaimed, but rather the beginning of an era where big data generates big theory, too.

# References


Agrawal, A., J. McHale, and A. Oettl (2024): "Artificial intelligence and scientific discovery: A model of prioritized search," *Research Policy*, 53, 104989.

Allen, R. and R. McDonald (2025): "Methodological pluralism and innovation in data-driven organizations," *Administrative Science Quarterly*, 70, 403–443.

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.

Arts, S., N. Melluso, and R. Veugelers (2025): "Beyond citations: Measuring novel scientific ideas and their impact in publication text," *Review of Economics and Statistics*, 1–33.

Balasubramanian, N., Y. Ye, and M. Xu (2022): "Substituting human decision-making with machine learning: Implications for organizational learning," *Academy of Management Review*, 47, 448–465.

Baruffaldi, S. and F. Gaessler (2025): "The returns to physical capital in knowledge production: Evidence from lab disasters," *American Economic Journal: Applied Economics*.

Bikard, M. (2018): "Made in academia: The effect of institutional origin on inventors' attention to science," *Organization Science*, 29, 818–836.

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

Callaway, E. (2017): "New concerns raised over value of genome-wide disease studies." *Nature*, 546, 463–464.

Camuffo, A., A. Cordova, A. Gambardella, and C. Spina (2020): "A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial," *Management Science*, 66, 564–586.

Camuffo, A., A. Gambardella, and A. Pignataro (2024): "Theory-driven strategic management decisions," *Strategy Science*, 9, 382–396.

Cao, R., R. Koning, and R. Nanda (2024): "Sampling bias in entrepreneurial experiments," *Management Science*, 70, 7283–7307.

Chai, S. (2017): "Near misses in the breakthrough discovery process," *Organization Science*, 28, 411–428.

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

Cohen, W. M. and D. A. Levinthal (1994): "Fortune favors the prepared firm," *Management Science*, 40, 227–251.

Conti, A. and D. Messinese (2024): "The Selective Tailwind Effect of Artificial Intelligence," *Available at SSRN*.

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.

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

Duncan, L. E. and M. C. Keller (2011): "A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry," *American Journal of Psychiatry*, 168, 1041–1049.

Ehrig, T. and T. Zenger (2024): "Competing with theories: Using awareness and confidence to secure resources and rents," *Strategy Science*, 9, 416–432.

Evans, J. A. (2010): "Industry induces academic science to know less about more," *American Journal of Sociology*, 116, 389–452.

Fanelli, D. (2010): ""Positive" results increase down the hierarchy of the sciences," *PloS One*, 5, e10068.

Felin, T., A. Gambardella, and T. Zenger (2024): "Theory-Based Decisions: Foundations and Introduction," *Strategy Science*, 9, 297–310.

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

Furman, J. L. and F. Teodoridis (2020): "Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering," *Organization Science*, 31, 330–354.

Gary, M. S. and R. E. Wood (2011): "Mental models, decision rules, and performance heterogeneity," *Strategic Management Journal*, 32, 569–594.

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

Gavetti, G., D. A. Levinthal, and J. W. Rivkin (2005): "Strategy making in novel and complex worlds: The power of analogy," *Strategic Management Journal*, 26, 691–712.

Girotra, K., C. Terwiesch, and K. T. Ulrich (2010): "Idea generation and the quality of the best idea," *Management Science*, 56, 591–605.

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

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.

Harrison, J. R. and J. G. March (1984): "Decision making and postdecision surprises," *Administrative Science Quarterly*, 26–42.

Heaven, W. D. (2023): "AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work," *MIT Technology Review*.

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

Heshmati, M. and F. A. Csaszar (2024): "Learning strategic representations: Exploring the effects of taking a strategy course," *Organization Science*, 35, 453–473.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2025): "The streetlight effect in data-driven exploration," *NBER wp 32401*.

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

Kao, J. (2024): "Charted Territory: 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.

Kapoor, R. and D. Wilde (2024): "Forecasting as a Problem of Cognitive Search: Experimental Evidence from Forecasting Tournaments in the Context of the Automotive Industry," *The Wharton School and University of Indiana*.

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.

Kim, S. (2023): "Shortcuts to Innovation: The Use of Analogies in Knowledge Production," *Columbia Business School*.

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.

Knudsen, T. and D. A. Levinthal (2007): "Two faces of search: Alternative generation and alternative evaluation," *Organization Science*, 18, 39–54.

Kuhn, T. S. (1962): *The structure of scientific revolutions*, Chicago University Press.

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

Levinthal, D. A. and J. G. March (1993): "The myopia of learning," *Strategic Management Journal*, 14, 95–112.

Loos, R. J. (2020): "15 years of genome-wide association studies and no signs of slowing down," *Nature communications*, 11, 5900.

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.

Merchant, A., S. Batzner, S. S. Schoenholz, et al. (2023): "Scaling deep learning for materials discovery," *Nature*, 624, 80–85.

Mishra, A., J. Diesner, and V. I. Torvik (2023): "A probabilistic model of 'Hype' in scientific abstracts," *International Society of Scientometrics and Informetrics Conference 2023 (ISSI)*.

Mullainathan, S. and A. Rambachan (2024): "From predictive algorithms to automatic generation of anomalies," *NBER wp 32422*.

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

Nelson, R. R. and S. Winter (1982): *An evolutionary theory of economic change*, Belknap Press.

Ott, T. E. and D. P. Hannah (2024): "On the origin of entrepreneurial theories: How entrepreneurs craft complex causal models with theorizing and data," *Strategy Science*, 9, 461–482.

Piezunka, H. and L. Dahlander (2015): "Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing," *Academy of Management Journal*, 58, 856–880.

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.

Popper, K. (1963): *Conjectures and refutations: The growth of scientific knowledge*, London: Routledge.

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.

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.

Schliesmann, D. (2025): "The Where of Search," *The Wharton School*.

Shi, F. and J. Evans (2023): "Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines," *Nature Communications*, 14, 1641.

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.

Singh, J. and L. Fleming (2010): "Lone inventors as sources of breakthroughs: Myth or reality?" *Management Science*, 56, 41–56.

Sorenson, O. (2024): "Theory, search, and learning," *Strategy Science*, 9, 372–381.

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.

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

Tranchero, M. (2024): "Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation," *The Wharton School*.

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.

Walsh, J. P. (1995): "Managerial and organizational cognition: Notes from a trip down memory lane," *Organization Science*, 6, 280–321.

Wang, H., T. Fu, Y. Du, et al. (2023): "Scientific discovery in the age of artificial intelligence," *Nature*, 620, 47–60.

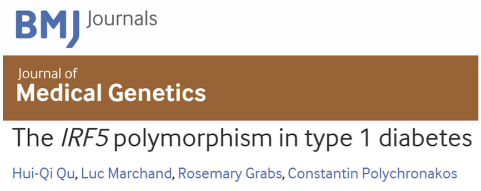
Wei, C.-H., A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, and Z. Lu (2024): "PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge," *Nucleic Acids Research*, 52, W540–W546.

Zittrain, J. (2019): "The hidden costs of automated thinking," *The New Yorker*.

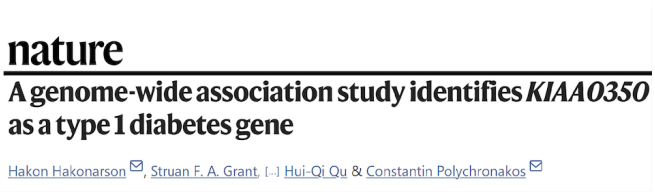
# 7 Figures and Tables

Figure 1: Scientists search for new gene-disease associations with candidate gene studies or with GWAS.

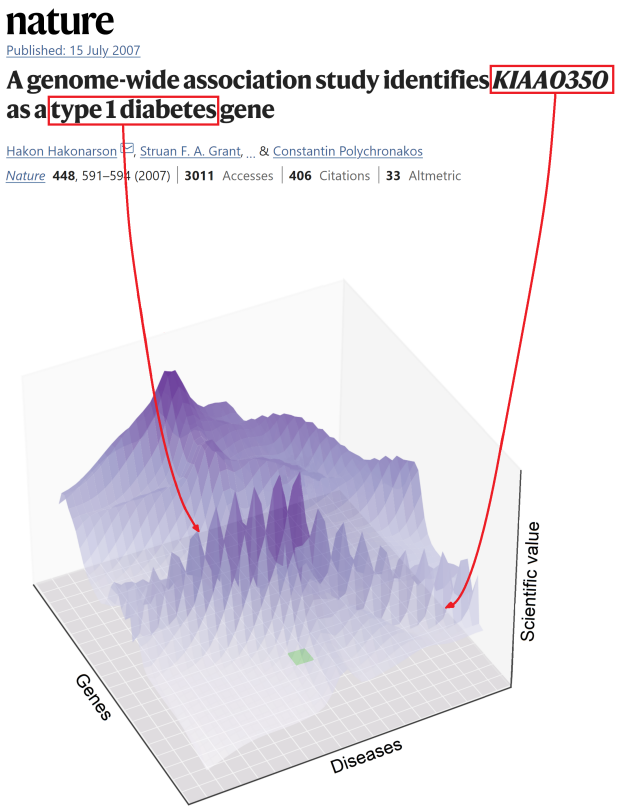
(a) *Example of candidate gene study*




(b) *Example of GWAS*



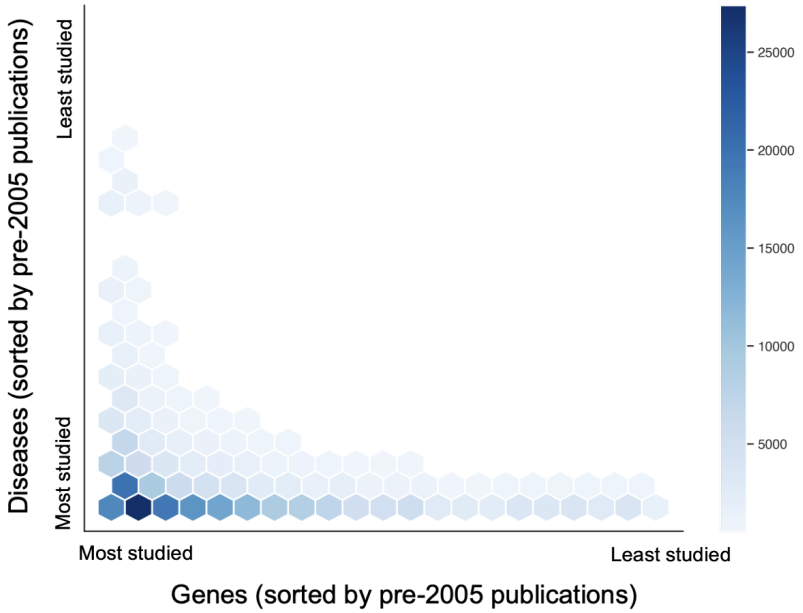

(c) *GWAS as data-driven search on a landscape of gene-disease combinations*



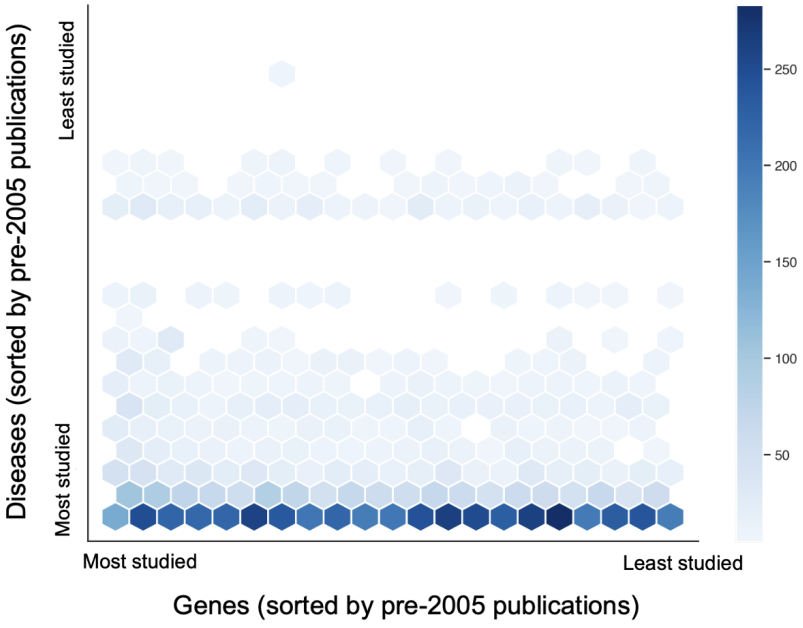

*Note:* Panel (a) shows the title of the candidate gene study by Qu et al. (2007) published in *The Journal of Medical Genetics* in 2007. Panel (b) shows the title of the GWAS by Hakonarson et al. (2007) published in *Nature* in 2007. Both studies searched for the genetic roots of type 1 diabetes and were carried out by the same principal investigator (PI) in the same year. More detail on these studies is provided in the case study in Appendix B. Panel (c) depicts how a typical GWAS introduces a new gene-disease association in the combinatorial landscape of all possible gene-disease pairs. Each combination of gene and disease has a specific scientific value, ideally captured by the elevation at that location. See text for details.

Figure 2: Conditional on the choice of disease, GWAS introduce new gene-disease associations spanning a larger portion of the genetic landscape relative to candidate gene studies.

(a) *GDAs introduced by candidate gene studies*



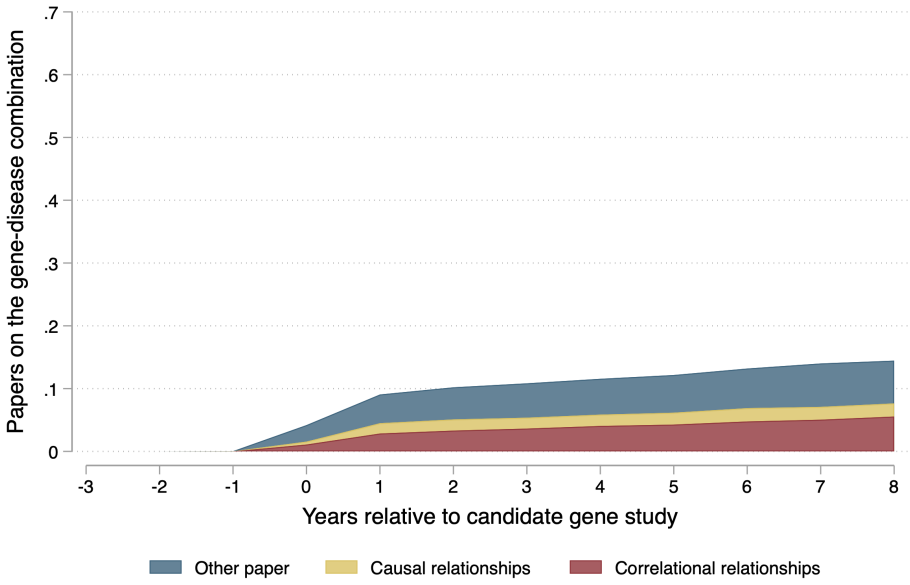

(b) *GDAs introduced by GWAS*



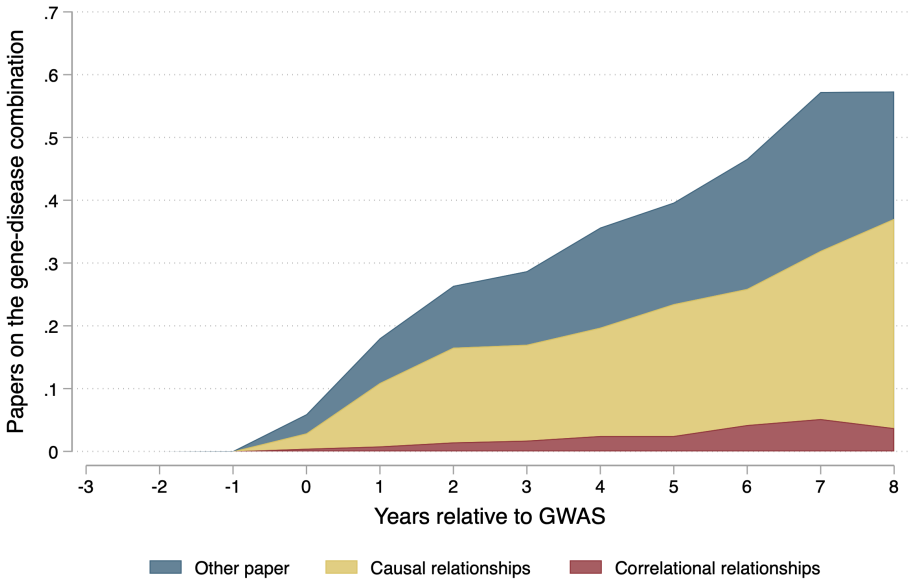

*Note:* Panel (a) shows a heatmap of new gene-disease associations introduced after 2005 by candidate gene studies. Panel (b) shows a heatmap of new gene-disease associations introduced after 2005 by genome-wide association studies. Both panels have 14,136 genes on the X axis, sorted from the most to the least studied in the pre-GWAS era, and 9,863 narrowly defined disease categories on the Y axis, sorted from the most to the least studied in the pre-GWAS era. Darker blue bins correspond to a higher number of associations involving those genes. See text for details.

Figure 3: GWAS introduce new gene-disease associations that receive more follow-on studies investigating causal mechanisms relative to candidate gene studies.

(a) *Follow-on articles on GDAs introduced by candidate gene studies*



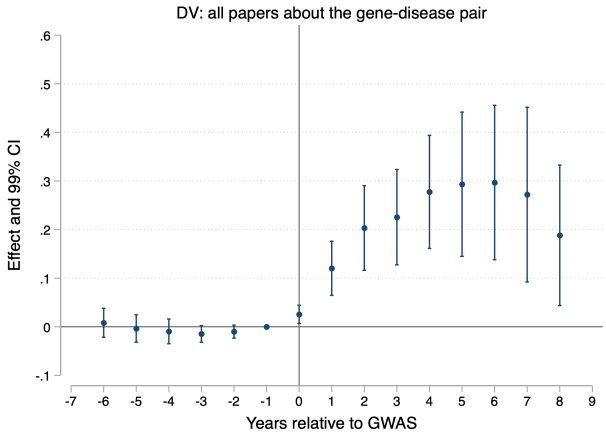

(b) *Follow-on articles on GDAs introduced by GWAS*



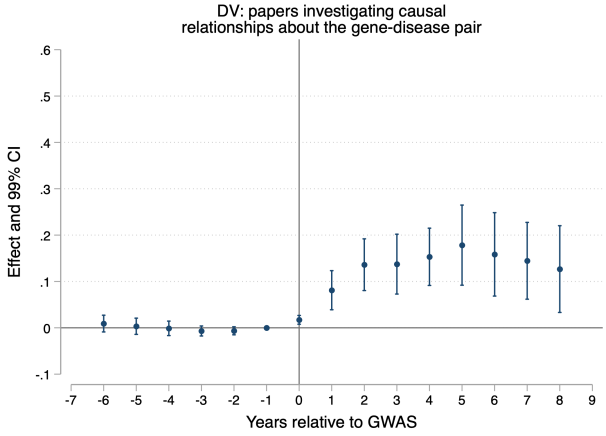

*Note:* Panel (a) shows the average yearly number of papers exploring gene-disease associations first introduced by a candidate gene study. Panel (b) shows the average yearly number of papers exploring gene-disease associations first introduced by a GWAS. Note that follow-on papers are those directly studying the same gene-disease pair, regardless of whether they cite the study that introduced it. In both figures, the count of papers is split out by whether the follow-on papers explore causal or correlational relationships according to the AI engine of PubTator3 (Wei et al., 2024). See text for details.

Figure 4: Time-varying estimates of the additional increase in follow-on studies for new gene-disease associations introduced by GWAS relative to candidate gene studies.

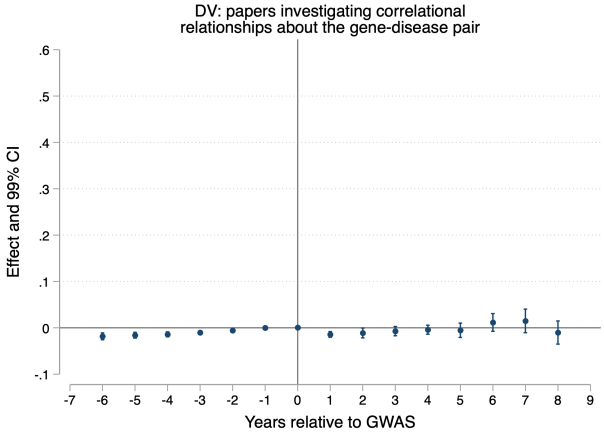
*(a) All papers about the gene-disease combination*




*(b) Papers investigating causal relationships about the gene-disease combination*



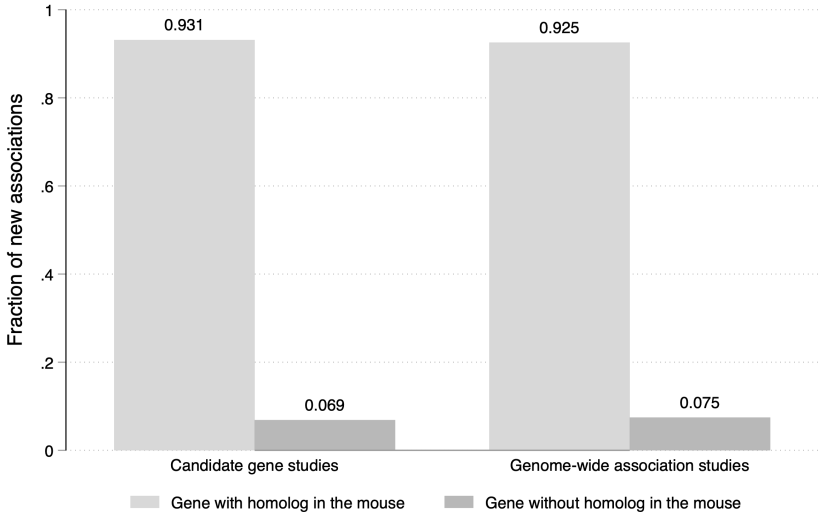

*(c) Papers investigating correlational relationships about the gene-disease combination*



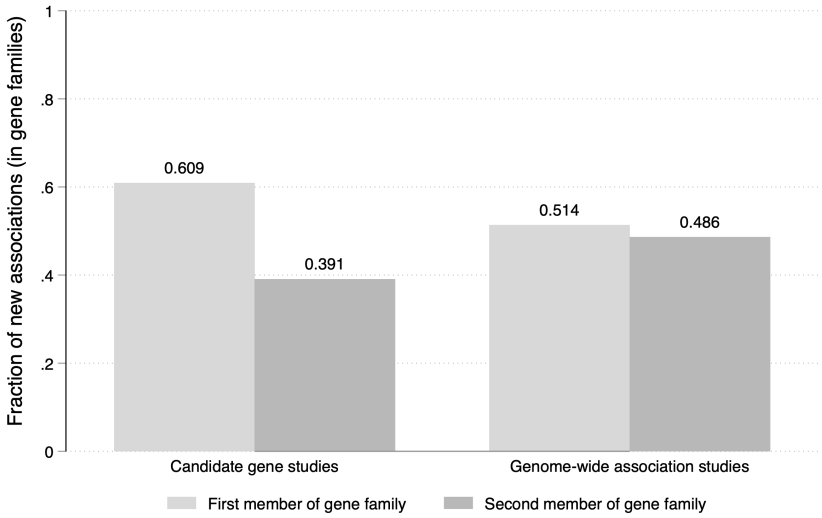

*Note:* This figure provides visual illustrations of the event study version of the difference-in-differences regressions evaluating the increase in follow-on publications for new gene-disease associations introduced by GWAS relative to new gene-disease associations introduced by candidate gene studies. Each panel shows the event study coefficients estimated from the following specification: $Papers_{i,j,t} = \alpha + \sum_z \gamma_z PostPublication_{i,j} \times 1(z) + \sum_z \beta_z PostPublication \times GWAS_{i,j} \times 1(z) + \lambda GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$. The charts plot values of $\beta_z$ for different lags $z$ before and after the gene-disease association is first reported by a GWAS with 99% confidence intervals. Standard errors are clustered at the disease level. The dependent variables are the yearly count of all papers about the gene-disease combination (panel (a)), the yearly count of papers investigating causal relationships about the gene-disease combination (panel (b)), and the yearly count of papers investigating correlational relationships about the gene-disease combination (panel (c)). Information on which papers explore causal or correlational relationships comes from the AI engine of PubTator3 (Wei et al., 2024). See text for details.

Figure 5: While GWAS are as likely as candidate gene studies to ignore genes without a homolog in the mouse, they are not biased towards the first member of gene families.

(a) *GDAs by whether gene has a mouse homolog*




(b) *GDAs by whether gene is first in a gene family*




*Note:* Panel (a) shows the share of new gene-disease associations on genes with and without a homolog gene in the lab mice, separately by methods used in the paper introducing them. Data used in panel (a) include all new gene-disease associations introduced in the period 2005-2016. Panel (b) shows the share of new gene-disease associations on the first vs. the second member of a gene family, separately by methods used in the paper introducing them. Data used in panel (b) include only new gene-disease associations involving members of a gene family introduced in the period 2005-2016. The associated regression estimates are in Appendix Table F.7. See text for details.

# Table 1: Descriptive statistics.

| | Panel A: paper-level descriptives (cross sectional) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Candidate gene studies | | | | | | GWAS | | | | | |
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| Forward citations | 40.84 | 22 | 82.985 | 0 | 8084 | 200,173 | 169.77 | 77 | 295.411 | 0 | 2822 | 1,375 |
| Associations per paper | 3.011 | 2 | 5.010 | 1 | 926 | 200,173 | 8.815 | 4 | 18.556 | 1 | 301 | 1,375 |
| Genes per paper | 1.621 | 1 | 2.265 | 1 | 677 | 200,173 | 5.712 | 3 | 10.340 | 1 | 134 | 1,375 |
| Year of publication | 2011.52 | 2012 | 3.296 | 2004 | 2016 | 200,173 | 2012.27 | 2012 | 2.512 | 2005 | 2016 | 1,375 |
| | Panel B: gene-disease level descriptives (cross sectional) | | | | | | | | | | | |
| | Candidate gene studies | | | | | | GWAS | | | | | |
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| With never associated genes (%) | 0.121 | 0 | 0.326 | 0 | 1 | 360,647 | 0.404 | 0 | 0.491 | 0 | 1 | 8,655 |
| With recently discovered gene (%) | 0.098 | 0 | 0.298 | 0 | 1 | 360,647 | 0.256 | 0 | 0.437 | 0 | 1 | 8,655 |
| Average DisGeNET Score | 0.056 | 0.01 | 0.100 | 0.01 | 1 | 360,647 | 0.069 | 0.01 | 0.122 | 0.01 | 1 | 8,655 |
| In bottom 10% DisGeNET Score (%) | 0.100 | 0 | 0.299 | 0 | 1 | 360,647 | 0.119 | 0 | 0.324 | 0 | 1 | 8,655 |
| In top 10% DisGeNET Score (%) | 0.094 | 0 | 0.291 | 0 | 1 | 360,647 | 0.150 | 0 | 0.357 | 0 | 1 | 8,655 |
| Unexpectedness Index | 0.024 | 0 | 0.135 | 0 | 1 | 357,840 | 0.105 | 0 | 0.291 | 0 | 1 | 8,617 |
| Gene is 2$^{nd}$ member of gene family (%) | 0.390 | 1 | 0.488 | 0 | 1 | 91,732 | 0.488 | 0 | 0.4999 | 0 | 1 | 2,162 |
| Gene without mouse homolog gene (%) | 0.069 | 0 | 0.254 | 0 | 1 | 360,647 | 0.074 | 0 | 0.263 | 0 | 1 | 8,655 |
| Year of the association | 2011.04 | 2011 | 3.385 | 2005 | 2016 | 360,647 | 2012.64 | 2013 | 2.614 | 2005 | 2016 | 8,655 |
| | Panel C: gene-disease-year descriptives (panel) | | | | | | | | | | | |
| | Candidate gene studies | | | | | | GWAS | | | | | |
| | mean | median | st d | min | max | N | mean | median | st d | min | max | N |
| Papers about the gene-disease pair | 0.047 | 0 | 0.387 | 0 | 101 | 4,673,131 | 0.088 | 0 | 0.980 | 0 | 88 | 108,940 |
| . . . investigating causal relationships | 0.024 | 0 | 0.219 | 0 | 72 | 4,673,131 | 0.051 | 0 | 0.570 | 0 | 53 | 108,940 |
| . . . investigating correlational relationships | 0.016 | 0 | 0.172 | 0 | 56 | 4,673,131 | 0.006 | 0 | 0.092 | 0 | 7 | 108,940 |
| Post publication (0/1) | 0.458 | 0 | 0.498 | 0 | 1 | 4,673,131 | 0.337 | 0 | 0.473 | 0 | 1 | 108,940 |
| Year | 2010 | 2010 | 3.742 | 2004 | 2016 | 4,673,131 | 2010 | 2010 | 3.742 | 2004 | 2016 | 108,940 |

*Note:* Panel A presents descriptive statistics on papers that introduce new gene-disease associations (GDAs) after 2005. *Forward citations*= citations received by the focal article up to 2020 inclusive (data from NIH iCite); *Associations per paper*= number of new GDAs introduced by the focal article; *Genes per paper*= number of genes associated with a disease by the focal article; *Year of publication*= year in which the focal article is published. 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); *Average DisGeNET Score*= average DisGeNET Score of GDAs' scientific quality; *In bottom 10% DisGeNET Score (%)*= share of GDAs that fall in the bottom 10$^{th}$ percentile of DisGeNET Score; *In top 10% DisGeNET Score (%)*= share of GDAs that fall in the top 90$^{th}$ percentile of DisGeNET Score; *Unexpectedness Index*= a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before (more details in Appendix D); *Gene is 2$^{nd}$ member of 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); *Gene without mouse homolog gene (%)*= share of GDAs that include a gene lacking a homolog gene in the mouse (data from NIH's https://www.ncbi.nlm.nih.gov/datasets); *Year of the association*= year in which the article introducing the GDA is published. Panel C presents descriptive statistics for the gene-disease-year level panel. *Papers about the gene-disease pair*= count of yearly papers about a specific gene-disease combination; *Papers about the gene-disease pair investigating causal relationships*= count of yearly papers about a specific gene-disease combination that investigate causal relationships; *Papers about the gene-disease pair investigating correlational relationships*= count of yearly papers about a specific gene-disease combination that investigate correlational relationships; *Post publication (0/1)*= 0/1 = 1 in all years after a gene-disease pair is reported by a paper for the first time; *Year*= average year of the observations in the panel. 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: GWAS are more likely to introduce gene-disease associations involving less-studied genes relative to candidate gene studies.

| | I(GDA with never associated gene>0) | | I(GDA with recently discovered gene>0) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | 0.261*** | 0.188*** | 0.144*** | 0.108*** |
| | (0.00747) | (0.0105) | (0.00569) | (0.00850) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.128 | 0.128 | 0.103 | 0.103 |
| Percentage increase | **204%** | **147%** | **140%** | **105%** |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *I(GDA with never associated gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene never associated with a disease before 2005 (the year of the first GWAS); *I(GDA with recently discovered gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene discovered after the year 2000 (the year of the Human Genome Project's first draft completion); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. Percentage changes are reported in bold for significant coefficients. See text for details and Appendix Table F.1 for the corresponding results using continuous versions of the dependent variables.

Table 3: GWAS increase variability in the quality of gene-disease associations relative to candidate gene studies, but also lead to a proportionally larger increase of associations with high scientific importance.

| | DisGeNET Score of the GDA | | I(GDA in bottom 10% DisGeNET Score>0) | | I(GDA in top 10% DisGeNET Score>0) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.0117** | 0.0111* | 0.0199*** | 0.0128* | 0.0412** | 0.0355* |
| | (0.00436) | (0.00448) | (0.00461) | (0.00654) | (0.0135) | (0.0141) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.059 | 0.059 | 0.100 | 0.100 | 0.100 | 0.100 |
| Percentage increase | **20%** | **19%** | **20%** | **13%** | **41%** | **36%** |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2), (4) and (6) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *DisGeNET Score*: synthetic measure of scientific reliability of the gene-disease association provided by DisGeNET (see Appendix for details); *I(GDA in bottom 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 10$^{th}$ percentile of the sample; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 90$^{th}$ percentile of the sample; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. Percentage changes are reported in bold for significant coefficients. See text for details.

Table 4: GWAS are more likely to introduce unexpected gene-disease associations relative to candidate gene studies.

| | Unexpectedness Index of the GDA | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | 0.0805*** | 0.0713*** | 0.00698*** | 0.00757*** |
| | (0.00751) | (0.00779) | (0.000873) | (0.00163) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 365,340 | 345,040 | 359,303 | 338,984 |
| Number of diseases | 8,730 | 8,429 | 8,730 | 8,429 |
| Mean of the DV | 0.027 | 0.027 | 0.011 | 0.011 |
| Percentage increase | **299%** | **265%** | **65%** | **71%** |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. Columns (1) and (2) include the whole sample, while Columns (3) and (4) exclude new gene-disease associations involving a gene never studied previously (which mechanically have Unexpectedness Index equal to 1). *Unexpectedness Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before (more details in Appendix D); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. Percentage changes are reported in bold for significant coefficients. See text for details.

Table 5: Relative to candidate gene studies, GWAS introduce gene-disease associations that generate more studies focused on the causal mechanisms linking the same gene-disease pair.

| | All papers | Papers about causal relationships | Papers about correlational relationships |
|---|---|---|---|
| | (1) | (2) | (3) |
| Post Publication (0/1) | 0.0561*** | 0.0250*** | 0.0151*** |
| | (0.00114) | (0.000650) | (0.000613) |
| ... × GWAS (0/1) | 0.153*** | 0.0922*** | 0.00410 |
| | (0.0215) | (0.0131) | (0.00224) |
| Gene-disease FE | YES | YES | YES |
| Disease × year FE | YES | YES | YES |
| Gene × year FE | YES | YES | YES |
| Observations | 4,782,071 | 4,782,071 | 4,782,071 |
| Number of diseases | 9,853 | 9,853 | 9,853 |
| Mean of the DV | 0.048 | 0.024 | 0.016 |
| Percentage increase (GWAS) | **319%** | **384%** | 26% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Difference-in-differences panel regressions at the gene-disease-year level. Std. err. clustered at the disease level. All models include gene-disease pair, disease × year, and gene × year fixed effects. *All papers*: yearly count of all scientific papers investigating a gene-disease pair; *Papers about causal relationships*: yearly count of all scientific papers investigating causal relationships about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Papers about correlational relationships*: yearly count of all scientific papers investigating correlational relationships about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Post Publication*: 0/1 = 1 in all years after a gene-disease pair is first reported in a study; *GWAS*: 0/1 = 1 for gene-disease associations introduced by a GWAS. All dependent variables count papers directly working on the gene-disease combinations, regardless of whether they cite the study that first introduced it. Percentage increases are computed for the interaction relative to the sample mean and rounded to the closest integer. Percentage changes are reported in bold for significant coefficients. See text for details.

# Data-Driven Search and the Birth of Theory: Evidence from Genome-Wide Association Studies

## Appendix

# A Additional Details on Genomic Research

## A.1 Scientific Background

Genomics is the branch of biological science focused on the study of genomes—that is, the complete set of an organism's genes. Genes are sequences of DNA bases that encode the "instructions" for synthesizing gene products, most notably proteins. They play a fundamental role in the functioning of the human body, but their sequences can sometimes acquire mutations. When this happens, genes may alter their behavior and affect the phenotypic traits of the organism, sometimes with serious consequences and the emergence of health conditions. At the same time, the role of genes in the etiology of disease creates opportunities for therapeutic intervention: genes linked to a condition can often serve as drug targets (Nelson et al., 2015). When a drug molecule binds to its target, it can modify the gene's function, potentially improving the condition. As a result, understanding the genetic roots of disease has direct and significant implications for pharmaceutical drug design.

Genetic research has historically been very effective at identifying individual genes responsible for specific disease conditions. These are known as Mendelian disorders, usually visible from birth and traceable through family history. Understanding the causes of Mendelian disorders was one of the earliest and most important successes of genetic research (Bush and Moore, 2012). Take, for example, the case of cystic fibrosis. This rare disorder can be caused by multiple DNA mutations, which tend to cluster in a specific region of the genome: the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. The approach to make this discovery involved genotyping families affected by cystic fibrosis. Despite small sample sizes, due to the rarity of the disease, the strong effect and localization of the responsible mutations in a single gene made it possible to unambiguously identify the gene as causally linked to the condition.

However, Mendelian diseases are typically rare, as their severity tends to make them subject to negative evolutionary selection. Much more common are complex diseases, which are not caused by a single genetic factor but result from mutations in many genes. In these cases, a mutation may increase disease risk without being either necessary or sufficient, meaning that it usually accounts for only a small fraction of the condition's heritability. Although complex disorders also cluster within families, they do not follow predictable inheritance patterns, as they are shaped by intricate interactions between genetic and environmental factors. For this reason, family-based studies have proven less effective when applied to more common, complex diseases.

## A.2 A Scientific Primer on GWAS

To make progress in the study of complex diseases, researchers have shifted focus toward the idea that common disorders are likely influenced by genetic mutations that are also common in the population (Reich and Lander, 2001). Rather than searching for individual genes with strong phenotypic effects, the field has moved toward studying common genetic variants that individually have only a small effect on disease risk (Bush and Moore, 2012). But what exactly is a variant? At the most basic level, two genomes differ at a specific genetic locus if they contain different single nucleotides (adenine, thymine, cytosine, or guanine) at that position. When such a difference occurs in at least 1% of the population, it is referred to as a single-nucleotide polymorphism (SNP). One way to associate SNPs with disease is based on the idea that a causative variant should appear more frequently in individuals with the disease than in those without it. In practice, researchers look for statistical correlations between specific variants and disease outcomes in large population samples, without requiring any family relationship among participants.

Building on this logic, candidate gene studies are a hypothesis-driven approach used in genetic research to identify associations between specific genes and diseases or traits. Panel (a) of Figure A.1 shows a stylized representation. Researchers begin by selecting one or more genes based on prior biological knowledge, typically genes believed to be involved in the physiological pathways relevant to the disease under investigation. The process involves genotyping individuals in a case-control design to detect whether genetic variants in the candidate genes occur more frequently in individuals with the disease than in those without it. Statistical tests are then used to assess whether these variants are significantly associated with disease presence or severity. Note that this approach genotypes individuals only at the specific genetic locations hypothesized as important. While efficient in targeted hypothesis testing, candidate gene studies are limited by their reliance on existing theory, which can bias discovery toward well-known genes and overlook novel or unexpected genetic contributors.

In contrast, genome-wide association studies (GWAS) are hypothesis-free methods for identifying associations between genetic regions (Visscher et al., 2017). Panel (b) of Figure A.1 shows a stylized representation. Like most candidate gene studies, GWAS are case-control studies: researchers collect DNA from patients with the disease under study and individuals with similar demographics but without the disease. However, in this case, researchers rely on high-throughput microarrays to genotype the entire genome, capturing data on millions of genetic variants across the entire genome of the individual. Variants found significantly more often in affected individuals may be biologically important for the

disease and related to its etiology. Compared to candidate gene studies, GWAS usually need much larger sample sizes to achieve statistical significance with such a large number of multiple hypotheses tested, thus substantially increasing their costs. It is important to note that array-based GWAS do not sequence DNA base by base, since they only detect the presence or absence of specific SNPs. While microarrays can genotype millions of SNPs, they still cover less than 0.1% of the genome. However, because the co-occurrence of nearby genetic variants is not random (a phenomenon known as linkage) researchers can use reference genomes (such as HapMap) to efficiently infer the characteristics of the broader genome from the smaller subset of SNPs directly genotyped (Bush and Moore, 2012; Uffelmann et al., 2021).[14]

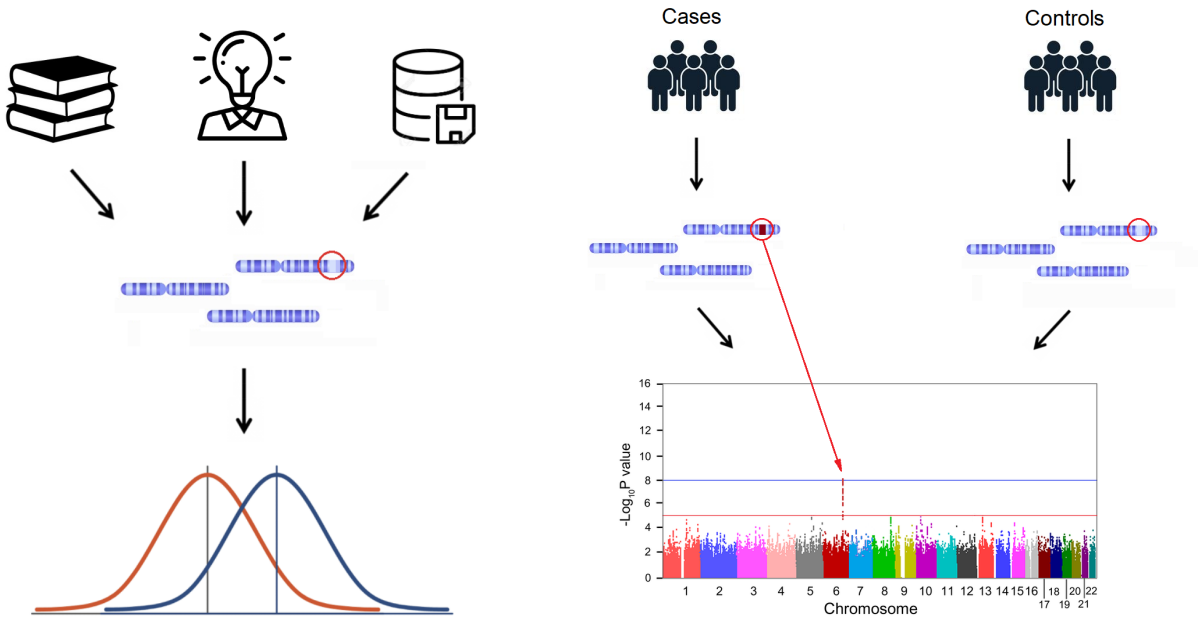
GWAS have led to many important scientific discoveries but have also come under scrutiny for several limitations. First, array-based genome-wide scans can only detect common variants that serve as markers for broader genomic regions likely to contain causal mutations. However, GWAS cannot pinpoint the exact causal SNPs with certainty, so additional analyses or follow-up studies are typically required to narrow down the association region. Second, even when the relevant mutation is identified, understanding the biological mechanism behind its role in human health requires further investigation. Third, most complex diseases are influenced by a large number of genes, meaning the variance explained by any single variant is usually small. This often limits the therapeutic value of GWAS findings, as variants with very small effects may not offer actionable drug targets (Goldstein, 2009). Finally, critics have noted that GWAS tend to overlook more complex genetic architectures, since they are primarily designed to test pairwise gene-disease associations (Boyle et al., 2017).

[14]Unlike whole-genome sequencing, which reads every DNA base, microarrays collect data only on selected genetic loci. This targeted approach is made possible by the HapMap project, which enabled the design of microarrays that focus on loci whose variation can be extrapolated to represent their surrounding genetic regions. As a result, researchers can efficiently infer the broader structure of the genome by sampling only a tiny fraction of DNA bases (Visscher et al., 2017).

Figure A.1: Stylized comparison of alternative types of genetic studies

(a) *Schema of a typical candidate gene study*

(b) *Schema of a typical GWAS*




*Note:* Panel (a) shows a schema of how a candidate gene study unfolds. First, researchers select the disease of interest and decide which gene(s) to investigate based on previous literature, existing evidence, or their own intuition. Then, the genome of people with and without the condition is genotyped *at that specific genetic location* in search of differences. Finally, statistical methods are used to test the association between mutations in the targeted genes and the disease of interest. Panel (b) shows a schema of how a GWAS unfolds. First, researchers select the disease of interest and assemble a group of cases (subjects showing the condition) and one of controls (subjects without the condition). Then, *the entire genome* of people with and without the condition is genotyped in search of differences at any genetic location. Finally, statistical methods are used to test the association between mutations in each gene and the disease of interest. The results are often graphically represented as a "Manhattan plot" and show the p-value of multiple statistical tests between DNAs in the case and control groups. The y-axis usually reports $-log_{10}(p - value)$, and hence higher values correspond to statistically stronger associations.

# B Case Study: Qu et al. (2007) vs. Hakonarson et al. (2007)

## B.1 Genetic Research on type 1 diabetes

It is estimated that 10% of Americans (around 37.3 million people) have diabetes.[15] The two main forms of diabetes are type 1 and type 2, with type 1 accounting for approximately 5–10% of cases. Type 1 diabetes is a chronic condition that typically begins in childhood. More specifically, it is an autoimmune disease caused by the immune system's destruction of pancreatic $\beta$-cells (DiMeglio et al., 2018). As a result, the pancreas is unable to produce sufficient insulin, the hormone essential for enabling sugar to enter cells, produce energy, and regulate blood glucose levels. To date, there is no known way to prevent type 1 diabetes, and lifelong insulin therapy remains necessary for patient survival.

Genetics plays a significant role in the onset of type 1 diabetes: children with a parent affected by the condition face a relative risk of 1–9% of developing it themselves (DiMeglio et al., 2018). Early candidate gene studies identified several genetic determinants, primarily within a tightly linked group of genes known as the major histocompatibility complex (MHC). MHC genes encode cell surface proteins that are essential for initiating and directing the immune response. When these genes malfunction, they can trigger autoimmune activity, such as the immune system attacking the body's own $\beta$-cells in type 1 diabetes. However, MHC genes account for only slightly more than half of the genetic risk associated with the disease, suggesting that other, still unidentified, loci are involved. A systematic effort to detect these remaining genes could reveal alternative therapeutic targets and shed light on the deeper causes of type 1 diabetes.

## B.2 The Candidate Gene Study of Qu et al. (2007)

In June 2007, Qu et al. (2007) published a candidate gene study based on a sample of 947 nuclear family trios with type 1 diabetes (i.e., one affected child and both parents). Using a targeted sequencing assay, the researchers tested the association between type 1 diabetes and a specific gene: interferon regulatory factor 5 (IRF5). The choice of IRF5 was informed by prior findings linking it to autoimmune diseases such as systemic lupus erythematosus. Based on these similarities, the authors hypothesized that IRF5 might also play a role in type 1 diabetes, another autoimmune condition with overlapping features. As they put it, the *"[...] association of IRF5 with other autoimmune diseases, such as T1D, has a*

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

*high prior probability"* based on existing evidence and genetic understanding. However, despite being grounded in theoretical reasoning and prior knowledge, the hypothesis turned out to be incorrect. The results showed no significant association between mutations in IRF5 and type 1 diabetes risk, leading Qu et al. (2007) to conclude that their proposed gene-disease combination had limited therapeutic potential.

## B.3 The GWAS of Hakonarson et al. (2007)

In August 2007, Hakonarson et al. (2007) published a GWAS based on a study population of 563 patients with type 1 diabetes and 1,146 controls.[16] The analysis was conducted using a microarray capable of genotyping 550,000 single nucleotide polymorphisms (SNPs) across the whole genome. The study identified several SNPs significantly associated with type 1 diabetes. Figure B.1 presents the key results from Hakonarson et al. (2007). Some of the significant SNPs were found in genes already known to be related to diabetes (for example, the insulin gene INS), but three were located in KIAA0350. It is now understood that this gene helps regulate $\beta$-cell function and thus plays a role in preventing diabetes. However, KIAA0350 was among the least studied human genes at the time, and its function was largely unknown (Soleimanpour et al., 2014). The association between KIAA0350 and type 1 diabetes proved robust and has since been investigated in multiple follow-up

[16]The study also replicated its main analysis in a separate sample of 483 nuclear family trios, leveraging genetic differences between affected children and their parents.

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) | $1.32 \times 10^{-7}$ | *PTPN22* |
| 11 | rs1004446 | 0.62 (0.53, 0.73) | $4.38 \times 10^{-9}$ | *INS* |
| 16 | rs2903692 | 0.65 (0.56, 0.76) | $4.77 \times 10^{-8}$ | *KIAA0350* |
| 11 | rs6356 | 1.52 (1.31, 1.76) | $1.78 \times 10^{-8}$ | *INS* |
| 16 | rs725613 | 0.67 (0.58, 0.78) | $3.24 \times 10^{-7}$ | *KIAA0350* |
| 7 | rs10255021 | 0.58 (0.44, 0.77) | $1.16 \times 10^{-4}$ | *COL1A2* |
| 11 | rs10770141 | 0.65 (0.56, 0.76) | $7.20 \times 10^{-8}$ | *INS* |
| 1 | rs672797 | 1.54 (1.29, 1.85) | $2.67 \times 10^{-6}$ | *LPHN2* |
| 16 | rs17673553 | 0.66 (0.55, 0.78) | $1.30 \times 10^{-6}$ | *KIAA0350* |
| 11 | rs7111341 | 0.63 (0.53, 0.76) | $3.77 \times 10^{-7}$ | *INS* |
| 11 | rs10743152 | 0.67 (0.57, 0.78) | $4.73 \times 10^{-7}$ | *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.

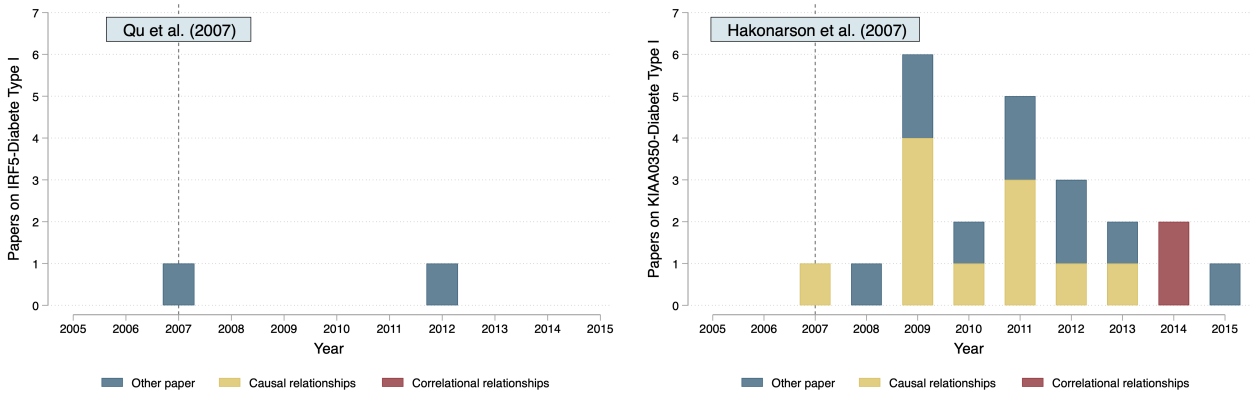
*Note:* Panel (a) shows an excerpt from Table 1 of Hakonarson et al. (2007). The genetic location of the single nucleotide polymorphisms significantly associated with type 1 diabetes is shown in the rightmost column. Highlighted in red are the three genetic variants located in the KIAA0350 gene. Panel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

studies, contributing to a deeper genetic understanding of the disease (Gingerich et al., 2020).

The studies by Hakonarson et al. (2007) and Qu et al. (2007) offer an interesting comparison: both were conducted in the same year, on the same disease, and by the same principal investigator. Figure B.2 shows the amount of follow-on research each of the two gene-disease combination received, based on DisGeNET data. Over the following ten years, 24 papers investigated the role of KIAA0350 in type 1 diabetes, with most of them aiming to understand the causal relationship between the gene and the disease. In contrast, the IRF5–type 1 diabetes combination received very limited attention, maybe not surprisingly given the weak findings. What is noteworthy, however, is that the hypothesis proposed by Qu et al. (2007) was guided by existing knowledge, which in this case led to targeting an already well-studied gene with low potential, while overlooking the opportunity presented by a neglected gene like KIAA0350.

Figure B.2: Follow-on papers investigating the association between type 1 diabetes and the IRF5 and KIAA0350 genes, respectively.

(a) *Papers on IRF5-T1D after Qu et al. (2007)* (b) *Papers on KIAA0350-T1D after Hakonarson et al. (2007)*




*Note:* Panel (a) shows the yearly count of publications exploring the relationship between IRF5 and type 1 diabetes following the candidate gene study by Qu et al. (2007). Panel (b) shows the yearly count of publications exploring the relationship between KIAA0350 and type 1 diabetes following the GWAS by Hakonarson et al. (2007). Data on publications and the gene-diseases studied come from DisGeNET, while information on the type of relationship studied is from PubTator3. In both panels, the focal paper introducing the gene-disease combination is excluded.

# C DisGeNET Data

This appendix describes the empirical approach used in the paper and then provides additional details on the DisGeNET data.

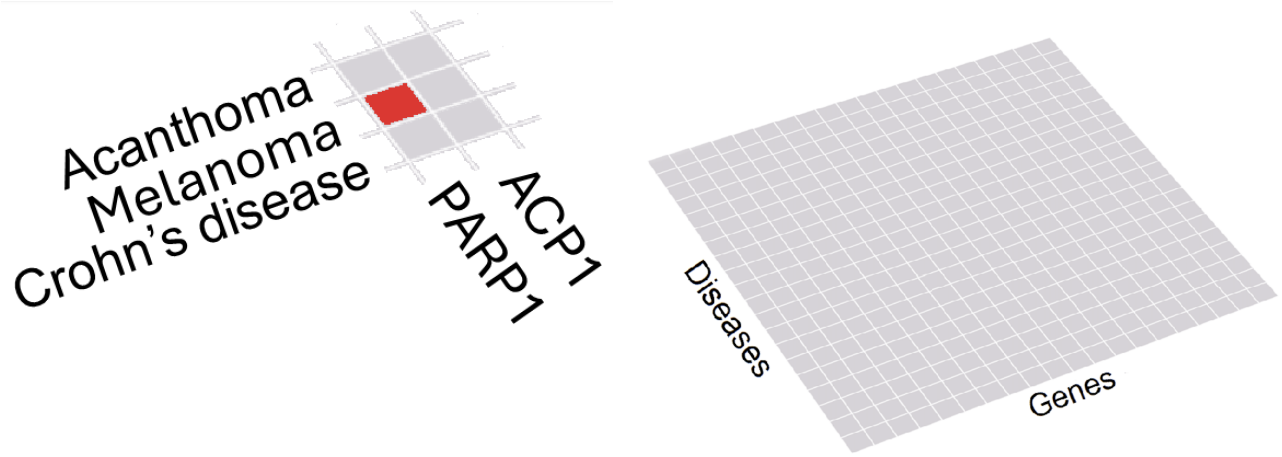
## C.1 The Gene-Disease Landscape

One advantage of focusing on a specific search problem, such as identifying the genetic roots of human diseases, is that the search landscape is well-defined. The relevant components in this combinatorial problem are genes and diseases. In principle, any gene could contain mutations that contribute to a given condition. This means the search task can, in first approximation, be reduced to identifying gene–disease pairs $< i, j >$ in which mutations in gene $i$ are causally linked to the emergence of disease $j$ (Panel (a) of Figure C.1). Another benefit of this setting is that all the relevant components of the problem have already been mapped and codified in exhaustive taxonomies. The Human Genome Project cataloged approximately 19,000 protein-coding genes and assigned them unique names and identifiers. Similarly, thousands of human diseases have been classified in systems such as the Unified Medical Language System (UMLS), which also provides unique disease identifiers. The Cartesian product of these entities defines the full combinatorial landscape in which search takes place, as illustrated in Panel (b) of Figure C.1.

In terms of empirical design, the combinatorial landscape represents the "ground truth," that is, the space of entities over which researchers conduct their search activities. Mapping this knowledge landscape allows for the empirical study of how actors search across it (Aharonson and Schilling, 2016). Specifically, one can locate research outputs (e.g., scientific publications) on the landscape by extracting the relevant entities (e.g., genes and diseases), thereby characterizing search in spatial terms. This approach offers two main advantages. First, it enables researchers to record the features of each combination relative to its position on the landscape. For example, Figure 2 illustrates how the spatial distribution of new combinations can be used to assess whether a search is "local" or "distant." Second, treating search as a landscape-based activity removes the need for citations to track knowledge evolution (Arts et al., 2025). Instead, one can observe changes in follow-on research that directly engages with the same entities before and after a given study. For instance, Figure 3 plots the yearly number of papers focused on a specific gene–disease combination, regardless of whether they cite the original study that introduced it.

Figure C.1: Empirically tracing the combinatorial landscape to study search.

(a) *Example of gene-disease combinations* (b) *The gene-disease combinatorial landscape*




Note: Panel (a) shows an example of gene-disease pairwise combinations. Genes and diseases can be sorted based on similarity or relatedness, such as melanoma and acanthoma, which are both skin tumors. The figure identifies in red the combination of PARP1 and melanoma, which was introduced by a GWAS. Panel (b) shows the idealized combinatorial landscape of all possible gene-disease pairs.

This empirical approach extends beyond the specific application in this paper. In practice, it can be used in any context where the relevant search landscape can be defined ex ante. Advances in science and technology have increasingly made this possible. From genetic atlases (Kao, 2024) to satellite imagery (Nagaraj, 2022), large-scale mapping efforts are turning a growing number of search problems into well-defined landscapes.

## C.2 The DisGeNET Score

The main source of data for this paper is DisGeNET (v7.0), a platform that integrates information from multiple sources to create a comprehensive repository of scientific findings linking human diseases to their genetic causes (Hermosilla and Lemus, 2019; Piñero et al., 2020). The database compiles gene–disease associations (GDAs) from curated datasets, experimental studies on animal models, and literature mining of publications indexed in PubMed. The version used in this paper includes over 628,000 gene-disease pairs involving 17,549 genes and 24,166 diseases. Genes are identified using their NCBI Gene ID (formerly EntrezGene ID), and diseases are coded using UMLS concept unique identifiers. In my paper, I focus on those introduced on or after 2005, the year of the first GWAS.

DisGeNET is designed to help researchers in both academia and industry prioritize promising genetic targets. To support this goal, it provides a synthetic DisGeNET Score for each gene–disease combi-

nation. The Score ranges from 0 to 1, with higher scores indicating associations that are scientifically more robust and therapeutically more promising. The DisGeNET Score incorporates both the number and type of sources supporting a given association, as well as the number of publications that have studied it. In the version used in this paper (v7.0), the Score is defined as follows:

$$\text{DisGeNET Score of gene-disease combination} < \text{i, j} > = C_{i,j} + M_{i,j} + I_{i,j} + L_{i,j}$$

The first component $C_{i,j}$ summarizes the evidence from curated sources reporting gene-disease combination $< i, j >$:

$$C_{i,j} = \begin{cases} 0.6 & \text{if } N_{sources_c} > 2 \\ 0.5 & \text{if } N_{sources_c} = 2 \\ 0.3 & \text{if } N_{sources_c} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.1}$$

where $N_{sources_c}$ is the number of curated sources supporting a gene-disease association, including CGI, ClinGen, Genomics England, CTD, PsyGeNET, Orphanet, and UniProt.

The second component $M_{i,j}$ summarizes the evidence from experiments using animal models reporting gene-disease combination $< i, j >$:

$$M_{i,j} = \begin{cases} 0.2 & \text{if } N_{sources_m} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{C.2}$$

where $N_{sources_m}$ is the number of sources using the lab rat or lab mouse from RGD, MGD, and CTD.

The third component $I_{i,j}$ summarizes the evidence inferred from experiments on gene-disease combination $< i, j >$:

$$I_{i,j} = \begin{cases} 0.1 & \text{if } N_{sources_i} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{C.3}$$

where $N_{sources_i}$ is the number of sources from HPO and CLINVAR.[17]

Finally, the component $L_{i,j}$ summarizes the evidence mined from the literature about gene-disease

[17] In the original version of the Score, papers listed in the GWAS Catalog where also included in this count. However, this would generate a mechanical upward bias in the DisGeNET Score of gene-disease pairs introduced by GWAS. Therefore, the current paper excludes those sources from the calculation of the DisGeNET Score, ensuring that the results in Table 3 are not due to a different weighting of GWAS papers.

combination $< i, j >$:

$$L_{i,j} = \begin{cases} 0.1 & \text{if } N_{publications} > 9 \\ \text{N}_{publications}\text{*0.01} & \text{if } N_{publications} <= 9 \end{cases} \tag{C.4}$$

where $N_{publications}$ is the number of publications supporting a gene-disease association as mined by LHGDN and BEFREE.

The DisGeNET Score has strong face validity and has been thoroughly validated (Piñero et al., 2020). Returning to the earlier example of type 1 diabetes, the data show that the IRF5–type 1 diabetes combination has a low DisGeNET Score of 0.03, compared to a much higher score of 0.46 for the KIAA0350–type 1 diabetes pair. These scores align well with the relative scientific and therapeutic impact of the two combinations (Gingerich et al., 2020). Table C.1 shows that publications introducing novel gene-disease combinations with higher DisGeNET Score receive a much larger number of citations. The result is robust to controlling for journal and scientist fixed effects. As an additional validation, Table C.2 compares the DisGeNET Score of each gene–disease combination with real-world innovation and therapeutic outcomes at the gene-disease level. If the Score is a good proxy for the underlying quality of a combination, it should correlate with successful downstream pharmaceutical developments. This is exactly what the data show: combinations with higher DisGeNET Scores are associated with significantly more follow-on applications in patents, clinical trials, and FDA-approved drugs. Importantly, none of these downstream outcomes are used in calculating the Score itself.

Table C.1: The DisGeNET Score of a given gene-disease combination is strongly associated with the citations accrued to the paper first introducing it.

| | Citations to the paper introducing the gene-disease pair | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| DisGeNET Score of the gene-disease pair | 28.90*** (7.525) | 60.91*** (7.863) | 63.39*** (6.214) |
| Journal FE | YES | YES | YES |
| Year of discovery FE | YES | YES | YES |
| Disease FE | NO | YES | YES |
| Principal investigator FE | NO | NO | YES |
| Observations | 374,292 | 368,995 | 348,742 |
| Number of diseases | 15,133 | 9,852 | 9,501 |
| Mean of the DV | 50.49 | 50.49 | 50.49 |
| Percentage increase | 57% | 121% | 126% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for journal and year fixed effects. Column 3 adds disease fixed effects, and Column 4 adds scientist fixed effects. *Citations to the paper introducing the gene-disease pair*: count of scientific citations received by the focal article introducing the new gene-disease combination (data from NIH's iCite); *DisGeNET Score of the gene-disease pair*: a synthetic measure of scientific reliability of the gene-disease association provided by DisGeNET. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table C.2: The DisGeNET Score of a given gene-disease combination is strongly associated with more USPTO granted patents, clinical impact, and FDA-approved drugs on the same combination.

| | Patents granted (1) | Clinical citations (2) | Drugs approved (3) |
|---|---|---|---|
| DisGeNET Score of the gene-disease pair | 1.313*** (0.156) | 27.31*** (2.455) | 0.0440*** (0.0121) |
| Disease FE | YES | YES | YES |
| Year of discovery FE | YES | YES | YES |
| Observations | 369,291 | 369,291 | 369,291 |
| Number of diseases | 9,862 | 9,862 | 9,862 |
| Mean of the DV | 0.193 | 2.128 | 0.008 |
| Percentage increase | 680% | 1,283% | 550% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects. *Patents granted*: count of USPTO patents granted from 2015 to 2023 that mention in their text a specific gene-disease combination. Data from EBI's SureChEMBL Workshop 2024 is available here: https://ftp.ebi.ac.uk/databases/SureChEMBL/SureChEMBLWorkshop2024/. *Clinical citations*: count of clinical articles up to 2024 citing the paper introducing a specific gene-disease combination. Data from NIH's iCite is available here: https://icite.od.nih.gov/. *Drugs approved*: count of FDA-approved drugs up to 2023 targeting a specific gene-disease combination. Data from DrugCentral 2023 is available here: https://drugcentral.org/. *DisGeNET Score of the gene-disease pair*: a synthetic measure of scientific reliability of the gene-disease association provided by DisGeNET. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

# D Unexpectedness Index

This appendix provides details on the construction of the Unexpectedness Index, which is computed for each novel gene–disease association based on established patterns of genetic co-occurrence. It begins with a stylized example to illustrate the logic of the measure, followed by a more detailed explanation and a set of robustness checks.

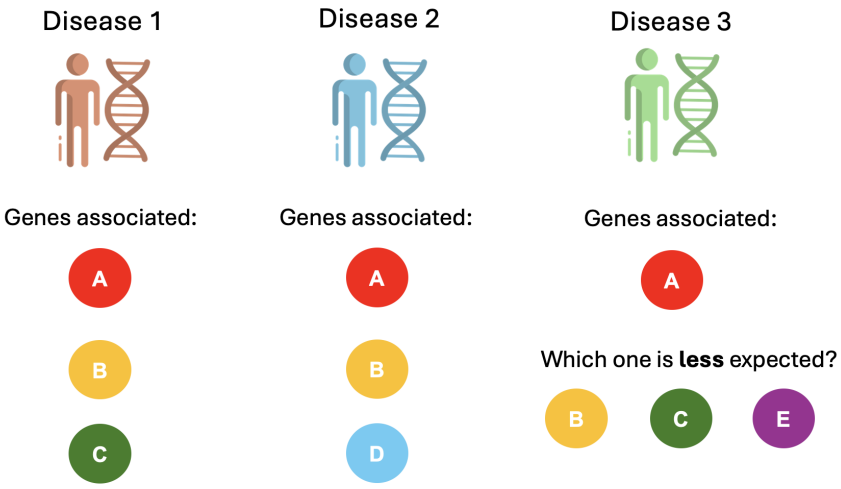
## D.1 Stylized Example

Consider the following example. There are three diseases:

- Disease 1, which has been associated with genes $\{A, B, C\}$ at time $t$;
- Disease 2, which has been associated with genes $\{A, B, D\}$ at time $t$;
- Disease 3, which has been associated with gene $\{A\}$ at time $t$.

Suppose now that Disease 3 can potentially be associated with genes $B$, $C$, and $E$ in time $t + 1$. How expected would be each potential combination between these gens and Disease 3?

Figure D.1: Example of how the Unexpectedness Index for each new gene-disease association is computed.




*Note:* See text for the step-by-step computation based on this simple example.

The first step involves computing the frequency of gene co-occurrences from past discoveries. In this example, we have the pairs of genes $\{(A, B), (A, C), (B, C)\}$ from Disease 1 and $\{(A, B), (A, D), (B, D)\}$ from Disease 2. Counting in how many diseases each pair occurs gives us the following frequencies:

- $\text{freq}(A, B) = 2$ (appears in Diseases 1 and 2)
- $\text{freq}(A, C) = 1$ (appears only in Disease 1)
- $\text{freq}(B, C) = 1$ (appears only in Disease 1)
- $\text{freq}(A, D) = 1$ (appears only in Disease 2)
- $\text{freq}(B, D) = 1$ (appears only in Disease 2)
- Any other potential pair (e.g. $(C, D), (A, E)$) has frequency 0.

Against this backdrop, Disease 3 has only been associated with the gene $\{A\}$ up to time $t$. Based on the pattern of genetic co-occurrences in Diseases 1 and 2, the association between Disease 3 and gene $\{A\}$ permits to compute how much "expected" each potential new gene association is for that disease in time $t + 1$. Hence, the *expectedness* of associating the generic gene $g$ to Disease 3 and its previous "gene set" $\{A\}$ is:

$\text{Expectedness}(g, \text{Disease 3}) = \sum_{x \in \{A\}} \text{freq}(g, x) = \text{freq}(g, A)$

Which then leads to the following:

- $\text{freq}(B, A) = 2 \implies \text{Expectedness}(B, \text{Disease 3}) = 2$
- $\text{freq}(C, A) = 1 \implies \text{Expectedness}(C, \text{Disease 3}) = 1$
- $\text{freq}(E, A) = 0 \implies \text{Expectedness}(E, \text{Disease 3}) = 0$

From this, we define the Unexpectedness Index as:

$\text{Unexpectedness}(g) = \frac{1}{1+\text{Expectedness}(g)}$,

thus leading to the following calculations:

$$\text{Unexpectedness}(B, \text{Disease 3}) = \frac{1}{1+2} = \frac{1}{3} \approx 0.33,$$
$$\text{Unexpectedness}(C, \text{Disease 3}) = \frac{1}{1+1} = \frac{1}{2} = 0.50,$$
$$\text{Unexpectedness}(E, \text{Disease 3}) = \frac{1}{1+0} = 1.0.$$

Intuitively, the Unexpectedness Index ranges from 0 to 1, with higher values indicating that a given gene–disease pair is more surprising based on prior discovery patterns in other diseases. The Unexpectedness Index is not defined for the first gene ever associated with a given disease, and it is mechanically equal to 1 if the gene has never been previously linked to any disease (as in the case of gene $E$ in the example above).

## D.2 Validation of the Unexpectedness Index

Following the procedure outlined above, I computed the Unexpectedness Index for every gene–disease combination in DisGeNET introduced after 2005. In total, I was able to calculate the Index for 367,586 gene–disease pairs (98.1% of the sample); the remaining 6,040 pairs involve the first gene ever associated with a given disease, for which the Index is not defined. The Index takes the maximum value of 1 for the 11,289 combinations that include a gene never previously linked to any disease. Consider again the studies by Hakonarson et al. (2007) and Qu et al. (2007), both published in 2007 and focused on type 1 diabetes—a disease that, by the end of 2006, had already been associated with 1,020 genes. The Unexpectedness Index for the KIAA0350–type 1 diabetes combination is 1, since it marked the first time KIAA0350 had been linked to any disease. In contrast, the Unexpectedness Index for the IRF5–type 1 diabetes combination is 0.00077, reflecting the fact that IRF5 had frequently co-occurred with genes already associated with diabetes up to that point.

Several additional analyses support the validity of the Unexpectedness Index. First, I compare the Index with the lexical choices used in the abstracts of scientific publications that first report each gene–disease combination. To do so, I draw on data from Mishra et al. (2023), who use a probabilistic model to classify adjectives in abstracts. Appendix Table D.1 shows that articles describing their findings as "novel" introduced gene–disease pairs that have an Unexpectedness Index 39–46% higher on average. By contrast, the same effect is not observed for more generic terms such as "major" or "critical." This suggests that the relationship is not simply driven by hype in the abstract, but rather reflects the Index's ability to capture truly surprising discoveries.

Second, I follow the approach of Shi and Evans (2023) and compare the Unexpectedness Index with expert ratings from the Faculty Opinions platform. Faculty Opinions curates a post-publication peer-review system in which experts evaluate and annotate published research using predefined tags, such as whether a paper presents a "new finding" or a "technical advance." These data are available for a subset of 11,465 papers that introduce new gene–disease combinations in my sample. Despite the limited coverage, the results in Appendix Table D.2 show that the Unexpectedness Index is significantly associated only with papers tagged as introducing a "new finding." The same pattern does not appear for other forms of novelty, such as the use of a new technique. This provides further support for the Index as a measure of surprising scientific discoveries.

Table D.1: The Unexpectedness Index is higher for gene-disease combinations described as "novel" in the paper abstract, but not for other words generically denoting hype

| | Unexpectedness Index | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (46 |
| I(Results described as "Novel"$>$0) | 0.0123*** | 0.0105*** | | | | |
| | (0.00142) | (0.00159) | | | | |
| I(Results described as "Major"$>$0) | | | -0.000851 | -0.000852 | | |
| | | | (0.00122) | (0.00177) | | |
| I(Results described as "Critical"$>$0) | | | | | 0.000171 | -0.000130 |
| | | | | | (0.00122) | (0.00170) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 |
| Percentage increase | 46% | 39% | -3% | -3% | 1% | 0% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Unexpectedness Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *I(Results described as "Novel"$>$0)*: 0/1 = 1 if the gene-disease association is described as "novel" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *I(Results described as "Major"$>$0)*: 0/1 = 1 if the gene-disease association is described as "major" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *I(Results described as "Critical"$>$0)*: 0/1 = 1 if the gene-disease association is described as "critical" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023). Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table D.2: The Unexpectedness Index is higher for gene-disease combinations rated as "new findings", but not for other type of ratings by scientists on Faculty Opinions.

| | Unexpectedness Index | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| I(Rated as "New Finding">0) | 0.0877** | | | |
| | (0.0266) | | | |
| I(Rated as "New Drug Target">0) | | 0.0209 | | |
| | | (0.0148) | | |
| I(Rated as "Controversial">0) | | | -0.0163 | |
| | | | (0.0267) | |
| I(Rated as "New Technique">0) | | | | -0.0119 |
| | | | | (0.0212) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES |
| Observations | 11,465 | 11,465 | 11,465 | 11,465 |
| Number of diseases | 1,471 | 1,471 | 1,471 | 1,471 |
| Mean of the DV | 0.057 | 0.057 | 0.057 | 0.057 |
| Percentage increase | 154% | 37% | -29% | -21% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease, year, and principal investigator (PI) fixed effects. *Unexpectedness Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before; *I(Results rated as "New Finding">0)*: 0/1 = 1 if the gene-disease association is rated as a "new finding" by scientists in the Faculty Opinions platform; *I(Results rated as "New Drug Target">0)*: 0/1 = 1 if the gene-disease association is rated as constituting a "new drug target" by scientists in the Faculty Opinions platform; *I(Results rated as "Controversial">0)*: 0/1 = 1 if the gene-disease association is rated as being "controversial" by scientists in the Faculty Opinions platform; *I(Results rated as "New Technique">0)*: 0/1 = 1 if the gene-disease association is rated as employing a "new technique" by scientists in the Faculty Opinions platform. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

# E PubTator3 Data

This appendix provides additional details on PubTator3 and includes examples illustrating how its AI engine captures the epistemic nature of the gene–disease relationships investigated in scientific articles.

## E.1 Details on PubTator3

PubTator3 is an advanced text-mining tool developed by the NIH's National Center for Biotechnology Information (NCBI) to help researchers extract biomedical concepts from scientific literature (Wei et al., 2024). It integrates natural language processing and state-of-the-art AI techniques to automatically recognize, annotate, and normalize bio-entities (such as genes and diseases) as well as the relationships between them (e.g., a gene causing a disease vs just co-occurring with it). PubTator3 currently provides entity and relation annotations across roughly 36 million PubMed abstracts and 6 million full-text articles from the PubMed Central open-access subset. The tool is freely accessible to the research community at: https://www.ncbi.nlm.nih.gov/research/pubtator3/ (see Wei et al. 2024, for details).

PubTator3 employs a high-performance entity search engine that normalizes different forms of the same biological entity into standardized names, ensuring consistent retrieval of relevant articles regardless of terminology variation. The system uses AIONER, a newly developed AI annotation tool, to identify six key bio-entities: genes, diseases, chemicals, variants, species, and cell lines. Each entity type is normalized using specialized natural language processing modules tailored to existing terminology standards, such as GNormPlus for genes (mapped to NCBI Gene identifiers) and TaggerOne for diseases (mapped to MeSH terms). This normalization process ensures accurate and consistent identification across the literature. In addition, PubTator3 integrates BioREx, a transformer-based relation extraction model capable of identifying multiple types of relationships between different bio-entities, including gene–disease interactions. BioREx significantly enhances relation extraction performance, achieving an F-score of 79.6% on standard test sets (Wei et al., 2024).

## E.2 Entity Relationships from PubTator3

PubTator3 extracts a total of twelve possible types of relationships among bio-entities using the transformer-based relation extraction method known as BioREx. For the purposes of this paper, I

classify the following PubTator3 relationship types as reflecting interactions of a *causal* nature:

- Gene A *causes* or *stimulates* disease B: when the status of one entity increase (or decrease) as the other increase (or decreases);
- Gene A *inhibits* or *prevents* disease B: when the status of one entity increase (or decrease) as the other decreases (or increase);
- Gene A *treats* or *co-treats* disease B: if a chemical or drug treats a disease, alone or in isolation, through a given gene.

Note that all results remain robust when restricting the analysis to the strict relationship in which Gene A *causes* Disease B. Similarly, I classify the following PubTator3 relationship types as reflecting interactions of a *correlational* nature:

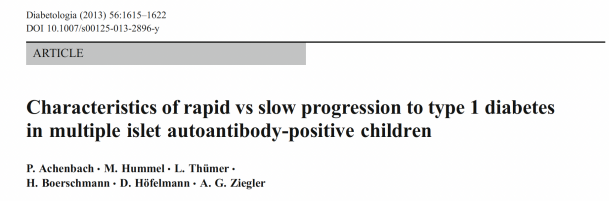
- Gene A *negatively correlate* with disease B: when the status of the two entities tends to be opposite;
- Gene A *positively correlate* with disease B: when the status of one entity tends to increase (or decrease) as the other increase (or decreases);

I use the unique PubMed ID of each paper in DisGeNET to merge it with the corresponding relationships extracted by PubTator3. I retain only those relationships that involve either a causal or correlational link between the relevant genes and diseases. As a result, any paper in DisGeNET can be classified as presenting either causal or correlational relationships, based on how PubTator3 categorizes the gene–disease interactions it contains. Out of the 201,548 unique articles in my dataset, 74,634 include at least one causal relationship, 46,233 include at least one correlational relationship, and 29,432 include both (but involving different gene–disease pairs).

As an example, consider the GWAS by Hakonarson et al. (2007). PubTator3 allows for a more detailed characterization of the follow-on studies investigating the KIAA0350–type 1 diabetes relationship. Panel (a) presents the title of a study by Achenbach et al. (2013), an observational study examining factors that influence the rate at which children with multiple islet autoantibodies develop type 1 diabetes. Among 1,650 children followed, 23 progressed to diabetes within 3 years (rapid progressors), while 24 remained non-diabetic for over 10 years (slow progressors). The study found that the presence

Figure E.1: Example of publications investigating correlational and causal relationships between the gene KIAA0350 and type 1 diabetes.

(a) *Papers on correlational relationship between KIAA0350-T1D*




(b) *Papers on causal relationship between KIAA0350-T1D*


*Note:* Panel (a) reports the abstract from Achenbach et al. (2013). Panel (b) reports the abstract from Soleimanpour et al. (2014). Clec16a is an alternative name for the KIAA0350 gene.

or absence of several mutations, including in the gene KIAA0350, was correlated with the speed of disease progression. Accordingly, this paper is tagged in PubTator3 with the relationship "gene KIAA0350 associates with disease diabetes mellitus".[18]

Panel (b) shows the title of the study by Soleimanpour et al. (2014), which investigates the mechanisms through which KIAA0350 influences the diseases it has been associated with, including type 1 diabetes. The researchers discovered that KIAA0350 encodes a protein that interacts with an enzyme called Nrdp1, protecting it from degradation. This interaction is critical, as Nrdp1 regulates another key protein involved in the clearance of damaged mitochondria. These findings reveal KIAA0350's role in preserving mitochondrial health via the regulation of mitophagy, shedding light on the biological mechanisms through which this gene could emerge as a therapeutic target for diabetes. In PubTator3, this paper is tagged with the relationship "gene KIAA0350 inhibits disease diabetes mellitus." Notably, one of the authors of Soleimanpour et al. (2014) also co-authored the earlier GWAS by Hakonarson et al. (2007).

I use a case–control approach to further validate PubTator3's ability to capture the nature of gene–disease relationships studied in scientific articles. Specifically, one would expect clinical trials to be more likely to investigate causal relationships, given their objective of causally testing the effect of therapies. To test this idea, I draw on data from NIH's iCite, which classifies an article as clinical if it is tagged with MeSH terms such as "Clinical Trial" or "Randomized Controlled Trial." Table E.1 confirms this intuition: articles reporting clinical trial results are significantly more likely to be classified as studying causal relationships in the PubTator3 data, and significantly less likely to present correlational

[18] Interestingly, Achenbach et al. (2013) does not cite Hakonarson et al. (2007), even if it investigates the KIAA0350-type 1 diabetes first introduced by it. This is one example of how my landscape-based approach can better capture the impact of combinations by not relying on paper-to-paper citations (Arts et al., 2025).

evidence. This provides additional support for the validity of the PubTator3 classification.

Table E.1: Clinical articles are more likely to explore causal relationships between a gene and a disease, and less likely to explore correlational relationships.

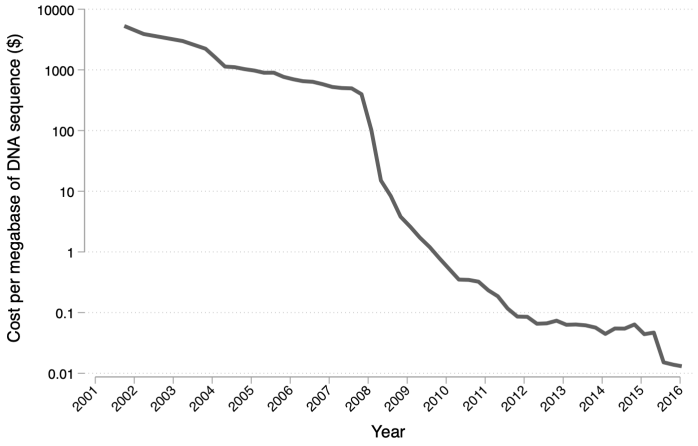
| | Causal relationship between gene and disease (0/1) | | Correlational relationship between gene and disease (0/1) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Clinical article (0/1) | 0.0437*** | 0.0269* | -0.0468*** | -0.0207* |
| | (0.0116) | (0.0121) | (0.00782) | (0.00858) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,305 | 348,930 | 369,305 | 348,930 |
| Number of diseases | 9,864 | 9,506 | 9,864 | 9,506 |
| Mean of the DV | 0.355 | 0.355 | 0.224 | 0.224 |
| Percentage increase | 12% | 8% | -21% | -9% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Causal relationship between gene and disease (0/1)*: 0/1 = 1 if the article introduces at least one causal relationship about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Correlational relationship between gene and disease (0/1)*: 0/1 = 1 if the article introduces at least one correlational relationship about a gene-disease pair according to the AI engine of PubTator3 (Wei et al., 2024); *Clinical article (0/1)*: 0/1 = 1 for new gene-disease associations introduced by an article reporting results of a clinical trial (data on clinical articles from NIH's iCite). Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

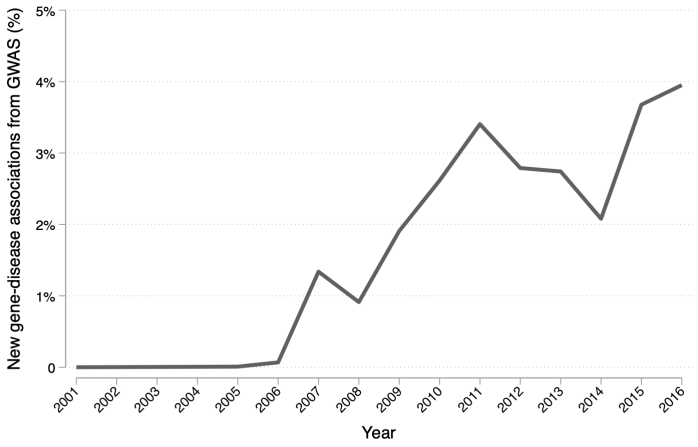
# F Additional Figures and Tables

Figure F.1: The emergence of GWAS coincided with a significant reduction in the cost of microarray-based DNA sequencing.

(a) *Cost of DNA sequencing*

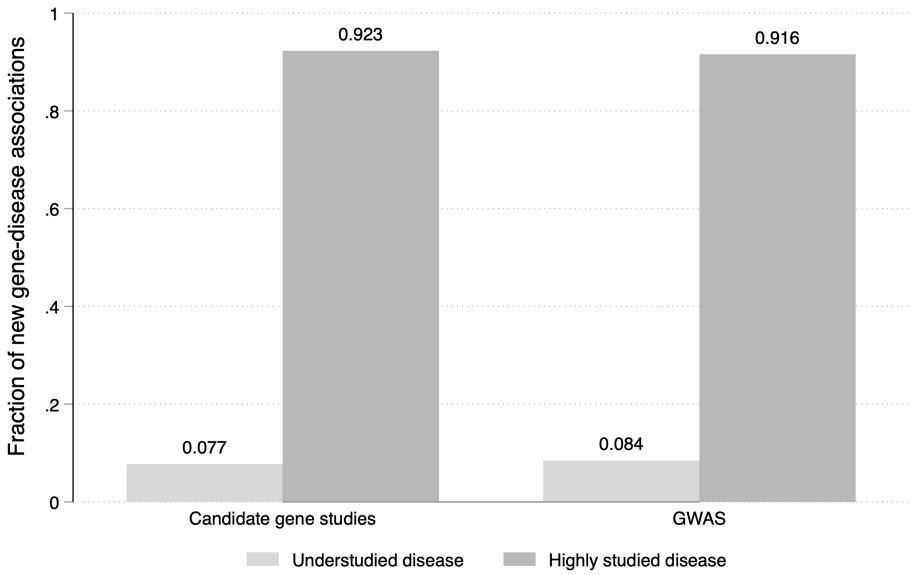



(b) *Share of new GDAs introduced by GWAS*




*Note:* Panel (a) shows the average cost of seqeuncing one megabase (i.e., a million bases) of DNA sequence over time. Data available at: `www.genome.gov/sequencingcostsdata`. Panel (b) displays the share of new gene-disease associations (GDAs) recorded in the DisGeNET database that were introduced by GWAS over time. See text for details.

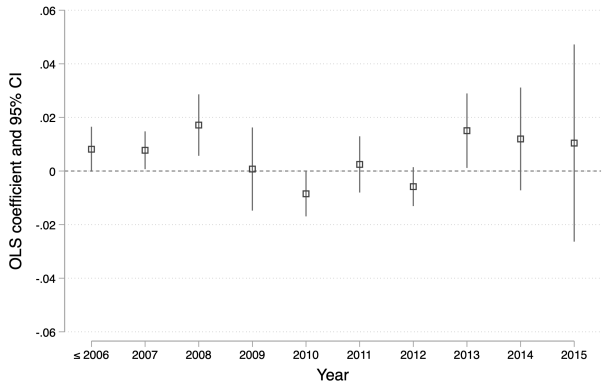
Figure F.2: GWAS induce diversification in gene space but do not change the disease focus, since the latter remains a choice of the scientist.



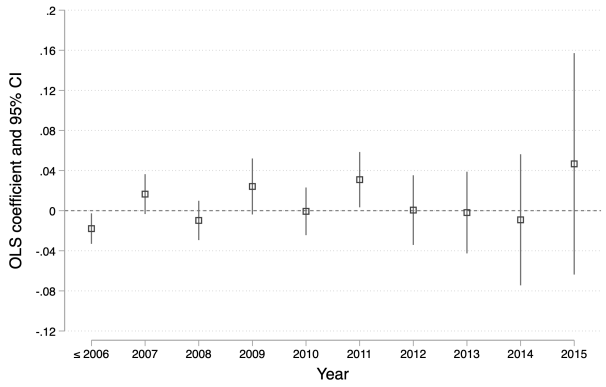

*Note:* The figure plots the share of new gene-disease associations that involve diseases without any genetic associations in the DisGeNET data as of 2005 vs well-studied diseases, separately by type of study. Data used in the graph include all new gene-disease associations introduced in the period 2005-2016.

Figure F.3: Robustness check: 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.

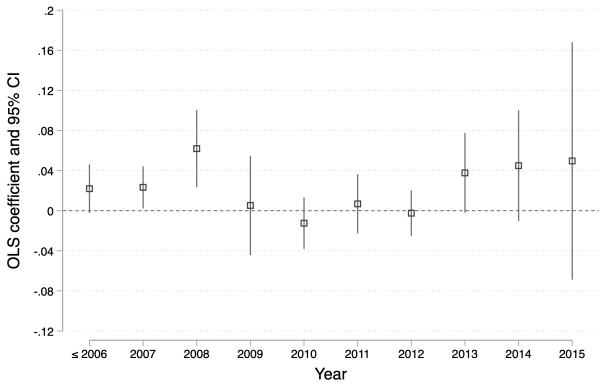
*(a) Average DisGeNET Score of gene-disease combinations*




*(b) Likelihood of gene-disease combinations in bottom 10% of DisGeNET Score*



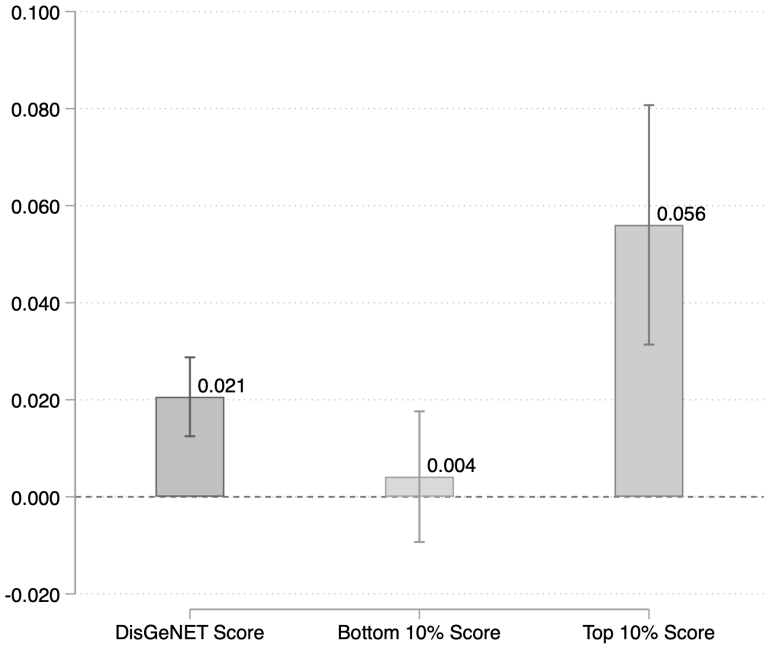

*(c) Likelihood of gene-disease combinations in top 10% of DisGeNET Score*



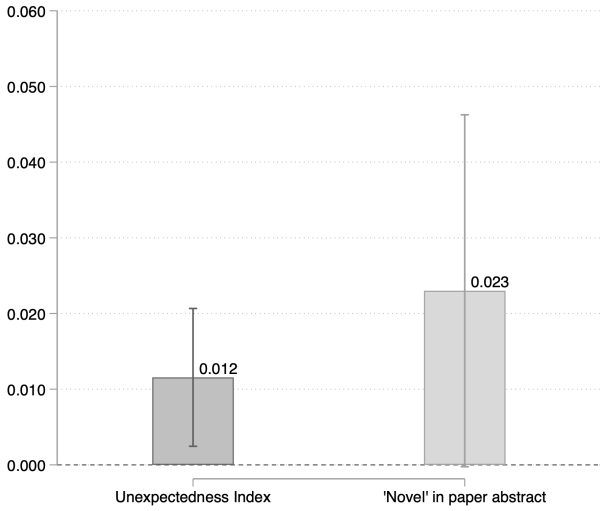

*Note:* This figure investigates if there are systematic differences in the scientific quality of gene-disease combinations introduced by scientists that self-select into publishing GWAS. Every coefficient is estimated from a separate regression for a given year, where I compare principal investigators who never publish a GWAS with principal investigators who will eventually publish one but have not done so already (that is, PIs are excluded from the sample after their first GWAS). In practice, each $\beta$ coefficient is estimated from the following specification: $(Quality\ of\ gene\text{-}disease\ combinations\ in\ year\ t)_{i,j} = \alpha + \beta(PI\ will\ publish\ GWAS)_{i,j,t} + \gamma Disease_i + \epsilon_{i,j,t}$. Panel (a) plots the coefficients and 95% confidence intervals from regressing the average DisGeNET score of gene-disease pairs introduced in a given year over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS in my sample period. Panel (b) plots the coefficients and 95% confidence intervals from regressing a dummy for gene-disease pairs in the bottom 10% of DisGeNET score introduced in a given year over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS in my sample period. Panel (c) plots the coefficients and 95% confidence intervals from regressing a dummy for gene-disease pairs in the top 10% of DisGeNET score introduced in a given year over a dummy for whether the principal investigator of the study will eventually publish at least one GWAS in my sample period. See text for details.

Figure F.4: Robustness check: Hypothesis 2 and Hypothesis 3 are confirmed even after the inclusion of gene fixed effects.

(a) *Scientific quality of gene-disease combinations*



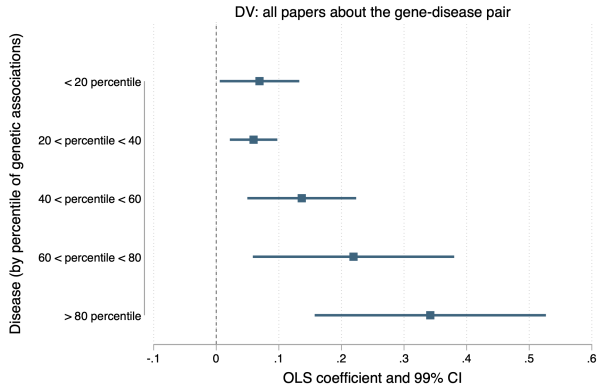

(b) *Unexpectedness of gene-disease combinations*



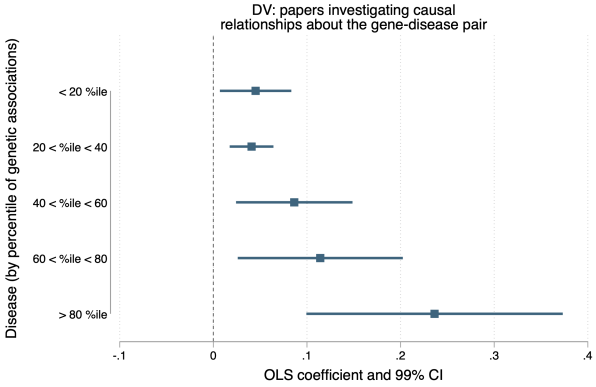

Note: Panel (a) shows the OLS coefficients of Columns (2), (4), and (6) of Table 3 with the inclusion of gene fixed effects. Panel (b) shows the OLS coefficients of Column (2) of Table 4 and Column (2) of Appendix Table F.6 with the inclusion of gene fixed effects. See respective tables for details.

Figure F.5: GWAS introduce new gene-disease associations that receive more follow-on scientific effort on the same gene-disease pair relative to candidate gene studies, especially in diseases where there is more pre-existing genetic knowledge.

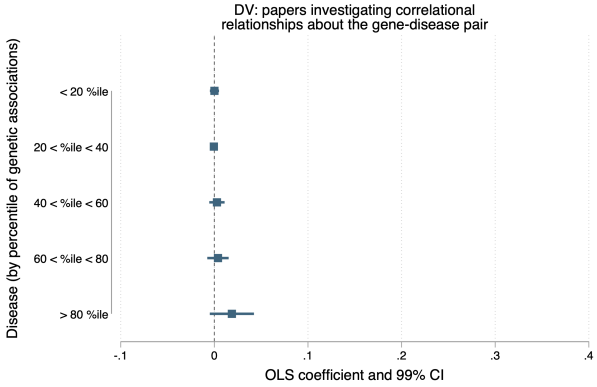
*(a) All papers about the gene-disease combination*




*(b) Papers investigating causal relationships about the gene-disease combination*




*(c) Papers investigating correlational relationships about the gene-disease combination*




*Note:* This figure shows estimates from OLS models evaluating the additional amount of follow-on publications received by new gene-disease associations introduced by GWAS relative to new gene-disease associations introduced by candidate gene studies. Each panel shows difference-in-differences coefficients estimated from the following specification: $Papers_{i,j,t} = \alpha + \gamma PostPublication_{i,j} + \beta PostPublication \times GWAS_{i,j} + \lambda GD_{i,j} + \delta_t \times Gene_i + \omega_t \times Disease_j + \epsilon_{i,j,t}$. The charts plot values of $\beta$ estimated with split-sample regressions for diseases with an increasing number of known genetic associations as of 2005 (the year of the first GWAS). Coefficients represent the additional number of publications received by gene-disease pairs introduced by GWAS with 99% confidence intervals. Standard errors are clustered at the disease level. The dependent variables are the yearly count of all papers about the gene-disease combination (panel (a)), the count of papers investigating causal relationships about the gene-disease combination (panel (b)), and the count of papers investigating correlational relationships about the gene-disease combination (panel (c)). Information on which papers explore causal or correlational relationships comes from the AI engine of PubTator3 (Wei et al., 2024). See text for details.

Table F.1: Robustness check: relative to candidate gene studies, GWAS introduce combinations involving genes that received fewer publications or were discovered later.

| | Publications on the gene (pre-2005) | | Years since the discovery of the gene (as of 2005) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | -50.84*** | -29.44*** | -2.978*** | -2.126*** |
| | (4.046) | (6.187) | (0.104) | (0.160) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 364,661 | 344,416 |
| Number of diseases | 9,863 | 9,504 | 9,786 | 9,422 |
| Mean of the DV | 100.57 | 100.57 | 12.81 | 12.81 |
| Percentage increase | -51% | -29% | -23% | -17% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *Publications on the gene (pre-2005)*: count of the publications received before 2005 by the gene associated with the disease in a new gene-disease combination; *Years since the discovery of the gene (as of 2005)*: years since the discovery of the gene associated with the disease in a new gene-disease combination; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.2: Robustness check: GWAS introduce more gene-disease associations in the bottom tail of scientific quality relative to candidate gene studies.

| GDA in... | ...bottom 5% DisGeNET Score (1) | ...bottom 10% DisGeNET Score (2) | ...bottom 15% DisGeNET Score (3) | ...bottom 20% DisGeNET Score (4) | ...bottom 25% DisGeNET Score (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.0108* | 0.0128* | 0.0224** | 0.0225* | 0.0219 |
| | (0.00437) | (0.00654) | (0.00783) | (0.00997) | (0.0113) |
| Disease FE | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES | YES |
| Observations | 348,921 | 348,921 | 348,921 | 348,921 | 348,921 |
| Number of diseases | 9,504 | 9,504 | 9,504 | 9,504 | 9,504 |
| Mean of the DV | 0.050 | 0.100 | 0.150 | 0.200 | 0.250 |
| Percentage increase | 22% | 13% | 15% | 11% | 9% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease, year, and principal investigator (PI) fixed effects. *GDA in bottom 5% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 5$^{th}$ percentile of the sample; *GDA in bottom 10% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 10$^{th}$ percentile of the sample; *GDA in bottom 15% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 15$^{th}$ percentile of the sample; *GDA in bottom 20% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 20$^{th}$ percentile of the sample; *GDA in bottom 25% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 25$^{th}$ percentile of the sample; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.3: Robustness check: GWAS introduce more gene-disease associations in the top tail of scientific quality relative to candidate gene studies.

| GDA in... | ...top 5% DisGeNET Score (1) | ...top 10% DisGeNET Score (2) | ...top 15% DisGeNET Score (3) | ...top 20% DisGeNET Score (4) | ...top 25% DisGeNET Score (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.0292* | 0.0355* | 0.0366* | 0.0196 | 0.00763 |
| | (0.0121) | (0.0141) | (0.0148) | (0.0146) | (0.0153) |
| Disease FE | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES |
| Principal investigator FE | YES | YES | YES | YES | YES |
| Observations | 348,921 | 348,921 | 348,921 | 348,921 | 348,921 |
| Number of diseases | 9,504 | 9,504 | 9,504 | 9,504 | 9,504 |
| Mean of the DV | 0.050 | 0.100 | 0.150 | 0.200 | 0.250 |
| Percentage increase | 58% | 36% | 24% | 10% | 3% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease, year, and principal investigator (PI) fixed effects. *GDA in top 5% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 95$^{th}$ percentile of the sample; *GDA in top 10% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 90$^{th}$ percentile of the sample; *GDA in top 15% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 85$^{th}$ percentile of the sample; *GDA in top 20% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 80$^{th}$ percentile of the sample; *GDA in top 25% DisGeNET Score*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 75$^{th}$ percentile of the sample; *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.4: Robustness check: Results are robust to the inclusion of more stringent fixed effect structures.

**Panel A: Including Disease × Year Fixed Effects**

| | I(GDA with never associated gene>0) (1) | I(GDA with recently discovered gene>0) (2) | I(GDA in bottom 10% DisGeNET Score>0) (3) | I(GDA in top 10% DisGeNET Score>0) (4) | Unexpectedness Index of GDA (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.261*** | 0.144*** | 0.0148** | 0.0601*** | 0.0848*** |
| | (0.00751) | (0.00626) | (0.00493) | (0.0116) | (0.00830) |
| Disease × year FE | YES | YES | YES | YES | YES |
| Observations | 351,013 | 351,013 | 351,013 | 351,013 | 348,807 |
| Number of diseases | 8,080 | 8,080 | 8,080 | 8,080 | 7,390 |
| Mean of the DV | 0.128 | 0.103 | 0.100 | 0.100 | 0.027 |
| Percentage increase | 204% | 140% | 15% | 60% | 314% |

**Panel B: Including Disease × PI × Year Fixed Effects**

| | I(GDA with never associated gene>0) (1) | I(GDA with recently discovered gene>0) (2) | I(GDA in bottom 10% DisGeNET Score>0) (3) | I(GDA in top 10% DisGeNET Score>0) (4) | Unexpectedness Index of GDA (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.141*** | 0.100** | 0.0139 | 0.0482† | 0.0468* |
| | (0.0381) | (0.0348) | (0.0228) | (0.0279) | (0.0230) |
| Disease × PI × year FE | YES | YES | YES | YES | YES |
| Observations | 175,728 | 175,728 | 175,728 | 175,728 | 173,842 |
| Number of diseases | 7,319 | 7,319 | 7,319 | 7,319 | 6,700 |
| Mean of the DV | 0.128 | 0.103 | 0.100 | 0.100 | 0.027 |
| Percentage increase | 110% | 97% | 14% | 48% | 173% |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. Panel A includes disease × year fixed effects, effectively comparing gene-disease pairs introduced in the same year *and* for the same disease. Panel B includes disease × PI × year fixed effects, effectively comparing gene-disease pairs introduced by the same principal investigator in the same year *and* for the same disease. *I(GDA with never associated gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene never associated with a disease before 2005 (the year of the first GWAS); *I(GDA with recently discovered gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene discovered after the year 2000 (the year of the Human Genome Project's first draft completion); *I(GDA in bottom 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 10$^{th}$ percentile of the sample; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 90$^{th}$ percentile of the sample; *Unexpectedness Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before (more details in Appendix D); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.5: Robustness check: Results are robust to the exclusion of studies that report only one new gene-disease association, hence being more likely to suffer from reporting bias.

**Panel A: Excluding PI Fixed Effects**

| | I(GDA with never associated gene>0) (1) | I(GDA with recently discovered gene>0) (2) | I(GDA in bottom 10% DisGeNET Score>0) (3) | I(GDA in top 10% DisGeNET Score>0) (4) | Unexpectedness Index of GDA (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.257*** | 0.140*** | 0.0170*** | 0.0399** | 0.0776*** |
| | (0.00782) | (0.00602) | (0.00482) | (0.0135) | (0.00752) |
| Disease FE | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | NO | NO | NO | NO |
| Observations | 308,483 | 308,483 | 308,483 | 308,483 | 305,511 |
| Number of diseases | 8,945 | 8,945 | 8,945 | 8,945 | 8,052 |
| Mean of the DV | 0.128 | 0.103 | 0.100 | 0.100 | 0.027 |
| Percentage increase | 201% | 136% | 17% | 40% | 287% |

**Panel B: Including PI Fixed Effects**

| | I(GDA with never associated gene>0) (1) | I(GDA with recently discovered gene>0) (2) | I(GDA in bottom 10% DisGeNET Score>0) (3) | I(GDA in top 10% DisGeNET Score>0) (4) | Unexpectedness Index of GDA (5) |
|---|---|---|---|---|---|
| GWAS (0/1) | 0.188*** | 0.112*** | 0.0101 | 0.0362* | 0.0676*** |
| | (0.0121) | (0.00923) | (0.00730) | (0.0159) | (0.00814) |
| Disease × PI × year FE | YES | YES | YES | YES | YES |
| Disease FE | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | NO | NO | NO | NO |
| Observations | 308,007 | 308,007 | 308,007 | 308,007 | 304,814 |
| Number of diseases | 8,912 | 8,912 | 8,912 | 8,912 | 8,009 |
| Mean of the DV | 0.128 | 0.103 | 0.100 | 0.100 | 0.027 |
| Percentage increase | 147% | 109% | 10% | 36% | 250% |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. Panel A includes disease and year fixed effects, while Panel B also includes dummies controlling for the principal investigator (PI) of the articles introducing a GDA. Both panels include only gene-disease associations introduced by articles with more than one discovery. *I(GDA with never associated gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene never associated with a disease before 2005 (the year of the first GWAS); *I(GDA with recently discovered gene>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene discovered after the year 2000 (the year of the Human Genome Project's first draft completion); *I(GDA in bottom 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score below the 10$^{th}$ percentile of the sample; *I(GDA in top 10% DisGeNET Score>0)*: 0/1 = 1 if the gene-disease association has a DisGeNET Score above the 90$^{th}$ percentile of the sample; *Unexpectedness Index*: a synthetic measure of how unlikely a given gene was to be associated with a disease given the pattern of genes previously associated with that disease up to the year before (more details in Appendix D); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.6: Robustness check: GWAS are more likely to introduce gene-disease associations described as "novel" relative to candidate gene studies, but not more likely to show generically hyped language.

| | I(Results described as "Novel">0) | | I(Results described as "Major">0) | | Hype Score | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| GWAS (0/1) | 0.0625*** | 0.0302* | -0.000212 | -0.00876 | -0.0394* | -0.0361* |
| | (0.0171) | (0.0124) | (0.00974) | (0.00804) | (0.0199) | (0.0169) |
| Disease FE | YES | YES | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES | NO | YES |
| Observations | 369,302 | 348,921 | 369,302 | 348,921 | 369,302 | 348,921 |
| Number of diseases | 9,863 | 9,504 | 9,863 | 9,504 | 9,863 | 9,504 |
| Mean of the DV | 0.134 | 0.134 | 0.060 | 0.060 | 0.249 | 0.249 |
| Percentage increase | 47% | 23% | 0% | -15% | -16% | -14% |

*Note:* *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2), (4) and (6) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *I(Results described as "Novel">0)*: 0/1 = 1 if the gene-disease association is described as "novel" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *I(Results described as "Major">0)*: 0/1 = 1 if the gene-disease association is described as "major" in the abstract of the paper reporting it, according to the data by Mishra et al. (2023); *Hype Score*: synthetic measure of the use of hyped language in the abstract of the paper reporting the gene-disease association, according to the data by Mishra et al. (2023); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

Table F.7: GWAS are more likely to associate the second member of gene families relative to candidate gene studies, but not genes without a mouse homolog.

| Subsample: | I(GDA with the second member of a gene family>0) Gene family members | | I(GDA with gene not present in the mouse>0) All genes | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| GWAS (0/1) | 0.0936*** | 0.0406† | -0.00315 | 0.00242 |
| | (0.0116) | (0.0232) | (0.00370) | (0.00568) |
| Disease FE | YES | YES | YES | YES |
| Year of discovery FE | YES | YES | YES | YES |
| Principal investigator FE | NO | YES | NO | YES |
| Observations | 91,581 | 79,285 | 369,302 | 348,921 |
| Number of diseases | 5,233 | 4,729 | 9,863 | 9,504 |
| Mean of the DV | 0.393 | 0.393 | 0.069 | 0.069 |
| Percentage increase | 24% | 10% | -5% | 4% |

*Note:* †, *, **,*** denote significance at the 10%, 5%, 1%, and 0.1% level, respectively. Cross-sectional regressions at the gene-disease association (GDA) level. Std. err. clustered at the disease level. All models include dummies controlling for disease and year fixed effects; Columns (2) and (4) also include dummies controlling for the principal investigator (PI) of the articles introducing a GDA. *I(GDA with the second member of a gene family>0)*: 0/1 = 1 if the new gene-disease association involves a gene that is the second member of a gene family (data on gene families from Stoeger et al. 2018); *I(GDA with gene not present in the mouse>0)*: 0/1 = 1 if the new gene-disease association encompasses a gene that lacks a homolog gene in the laboratory mouse (data from NIH's https://www.ncbi.nlm.nih.gov/datasets); *GWAS*: 0/1 = 1 for new gene-disease associations introduced by a GWAS. Percentage increases are computed relative to the sample mean and rounded to the closest integer. See text for details.

# Appendix References


Achenbach, P., M. Hummel, L. Thümer, H. Boerschmann, D. Höfelmann, and A. Ziegler (2013): "Characteristics of rapid vs slow progression to type 1 diabetes in multiple islet autoantibody-positive children," *Diabetologia*, 56, 1615–1622.

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.

Arts, S., N. Melluso, and R. Veugelers (2025): "Beyond citations: Measuring novel scientific ideas and their impact in publication text," *Review of Economics and Statistics*, 1–33.

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

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

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.

Goldstein, D. B. (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.

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.

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.

Soleimanpour, S. A., A. Gupta, M. Bakay, A. M. Ferrari, D. N. Groff, J. Fadista, L. A. Spruce, J. A. Kushner, L. Groop, S. H. Seeholzer, et al. (2014): "The diabetes susceptibility gene Clec16a regulates mitophagy," *Cell*, 157, 1577–1590.

Wei, C.-H., A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, and Z. Lu (2024): "PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge," *Nucleic Acids Research*, 52, W540–W546.