Academic Bubbles

Richard Lowery[1], Michael Sockin[1], and Matteo Tranchero[2]
University of Texas at Austin[1], University of Pennsylvania[2]

NBER Summer Institute Science of Science Funding — July 2026

Motivation

- Science incentivizes progress by awarding credit to researchers for their discoveries (Merton, 1957; Hill and Stein, 2025)
  - important for career advancement (Tuckman and Leahey, 1975, Diamond Jr, 1986)

- Credit primarily assigned via citations (Dasgupta and David, 1994)
  - **in principle**, should align with social value of scientific inquiry (Merton, 1968)
  - **in practice**, may lead to overcrowding in established fields (Wang et al., 2017, Packalen and Bhattacharya, 2017, 2020)

- **This paper:** study science as a (collaborative) industry with its incentives

Conceptual Question: Does researcher compensation based on publications and citations incentivize them to pursue most socially valuable topics?

- If researcher and society's interests properly aligned, decline in dynamism is a result of ideas becoming harder to find (Bloom et al. 2020, Park et al., 2023)

- if not, decline in new ideas from misaligned incentives $\rightarrow$ can be remedied!

This Paper

A dynamic general equilibrium **principal-agent model** of coordination among strategic researchers:

- researchers exert hidden effort to find new topics or investigate active ones
- rewarded both for contributions to active topic and when it matures
- noisy public signal about a topic's social value

Researchers compensated for both citations and prestige (career advancement)

- principal observes publications, citations, and social value upon maturity
- principal (society) designs (team) incentives to satisfy participation and incentive compatibility constraints

Test predictions using research on **genetic determinants of disease**

- clear topic assignment (genes) & ex-ante measures of social value (GWAS)
- researcher heterogeneity to map career concerns into crowding incentives

Key Results

1. **Strength of career incentives determines if researchers coordinate on only (ex-ante) promising topics**
   - researching an active topic garners citations regardless of its ultimate import (c.f. *searching for a new topic is uncompensated*)
   - strong career incentives $\rightarrow$ continue researching promising topics even after negative signal **(Academic Bubbles)**
   - incentives coordination role $\downarrow$ citation informativeness, research productivity
2. **Optimal contract may reward all research to get promising research**
   - **first-best:** all researchers participate and study only promising topics
   - **second-best (w/o career incentives):** subset of researchers participate / potential for no research (*coordination failure*)
   - **second-best (w/ career incentives):** can get academic bubbles but avoid inefficiencies of contract without them
3. **Publications on more studied genes have $\uparrow$ citation rates and crowded areas disproportionately attract researchers**
   - results hold even after controlling for ex-ante likely scientific importance (GWAS, mouse-ortholog instrument)
   - crowding weaker for senior, NIH intramural, lower ranked-school scholars
   - more-studied genes $\rightarrow$ more clinical but less patent citations

Related Literature

1. **Rate and direction of scientific progress** e.g., Nelson, 1962, Dasgupta and David, 1994, Aghion et al., 2008, Azoulay et al., 2011, Bryan and Lemus, 2017, Wang et al., 2017, Akcigit et al., 2021, Hopenhayn and Squintani, 2021, Carnehl and Schneider, 2025, Hill and Stein, 2025a, Hoelzemann et al., 2025
    - (to our knowledge) provide first GE characterization of how citation-based incentives coordinate researchers and shape direction of scientific progress

2. **Agency problems in encouraging innovation** e.g., Holmstrom, 1989, Bolton and Harris, 1999, Manso, 2011, Aghion et al., 2013, Bhattacharya and Packalen (2020)
    - we focus on the contract externalities associated with the implicit insurance provided by career incentives (i.e., citations)

3. **Resource misallocation** e.g., Hsieh and Klenow, 2009, Eisfeldt and Rampini, 2006, Glode and Lowery, 2016, Sims, 2003, Myers, 2020, Acemoglu, 2023
    - we focus on misallocation of academic scientific and innovative effort from a principal-agent perspective

Model

N active researchers that retire and are replaced with Poisson arrival rate η

- at each t: can search for new topic or research active one at effort cost κ

Researchers have preferences over intermediate compensation and effort

$$u\left(c_{nt}, \beta_{nt}, C_{nt}, \boldsymbol{r}^n, \boldsymbol{e}^n\right) = -\beta_{nt}^{-\gamma\phi_X} e^{-\gamma(\phi_A \Delta Y_t + \phi_X C_{nt}) + \kappa(\boldsymbol{r}^n + \boldsymbol{e}^n)' \iota_L}$$

where Y_t is output, ϕ_A loading on Y_t, ϕ_X loading on citations C_t & prestige β_{nt}

- $r_{lt}^n, e_{lt}^n \in \{0, 1\}$: indicate active research vs exploration in field l

Citation count C_t evolves according to

$$dC_{nt} = \sum_l r_{lt} r_{lt}^n dt - C_{nt-} dQ_{nt},$$

while prestige β_{nt} evolves according to

$$d\beta_{nt} = \beta_{nt-} \sum_l \pi_{lj} r_{lt-}^n dB_t^l + (1 - \beta_{nt-})\, dQ_{nt},$$

Participation constraint: initial expected utility must weakly exceed $-\frac{e^{-\delta}}{\rho+\eta}$

Model

L fields and field l at time t has either 0 or 1 active topic

- a new topic arrives by Poisson process with intensity λe_{lt}: e_{lt} is number of researchers investigating new topic
- an active topic matures by Poisson process with intensity λr_{lt}: r_{lt} is number of researchers researching active topic

A new topic j in field l is either significant $\pi_{jl} = 1$ or not $\pi_{jl} = 0$

- significant topics that mature raise aggregate productivity
- insignificant topics that mature reduce productivity (*misallocation*)

When a new topic j is discovered, there is a **public signal** S_{lj}

- ex-ante probability a topic is significant is p
- $\Pr(\pi_{jl} = 1 \mid S_{lj} = 1) = p_1$ and $\Pr(\pi_{jl} = 1 \mid S_{lj} = 0) = p_0$ with $p_0 < p_1$
- promising topics $S_{lj} = 1$ turn speculative $S_{lj} = 0$ with (arbitrarily) small Poisson arrival rate ξ

Markov Perfect Equilibrium of Researcher Mean-field Game

Assume $p_0 \leq \frac{1}{2}$, so socially optimal to abandon speculative topics $S_{lj} = 0$, then

1. **Active promising topic** $S_{lj} = 1 \rightarrow$ all researchers work on it until it matures or negative news arrives
2. **Active speculative topic** $\rightarrow$ if career concerns strong ($\phi_X \geq \phi_X^*(\phi_A)$), researchers work on speculative topic until it matures
3. **No active topic** $\rightarrow$ all researchers search for a new topic

An equilibrium if effort cost κ is sufficiently small (else, no research conducted)

If $\phi_X = 0$ (no career concerns), researchers always abandon speculative topics

Intuition: Paying only when agenda matures exposes researchers to substantial risk and encourages free-riding - citations insure and (excessively) coordinate

Testable Predictions: Researchers with weaker career concerns should crowd less $\rightarrow$ tenured (Tripodi et. al., 2025), non-academic, outside top institutions

Optimal Contracting by Principal (Society / Universities)

Suppose researchers are compensated by affine contract

$$w_{nt} = \phi_0 + \phi_A Y_{nt} + \phi_X X_{nt},$$

where $X_{nt} = \log \beta_{nt} + C_{nt}$ is status

Principal maximizes net output with subjective discount rate ρ

$$U_P = \sup_{\{\phi_0, \phi_A, \phi_X\}} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(Y_t - \sum_{n=1}^{N} w_{nt}\right) dt\right]$$

First-best: All N researchers active and research only promising topics

Second-best with $\phi_X \equiv 0$: $M < N$ researchers active and potential for zero active researchers (*coordination failure*)

Second-best: ϕ_X set so that researchers research all topics when κ sufficiently large (*relaxes constraint on M*) and to eliminate coordination failure

Empirical Analysis: Genetic Determinants of Disease

Key issues with testing model:

- need a setting with a clear definition of a topic / field
- need (ex-ante) measure of which topics are promising
- need observable heterogeneity in career incentives

Identification challenge: crowded topics, all else equal, likely important topics

Our setting: genetic determinants of human diseases

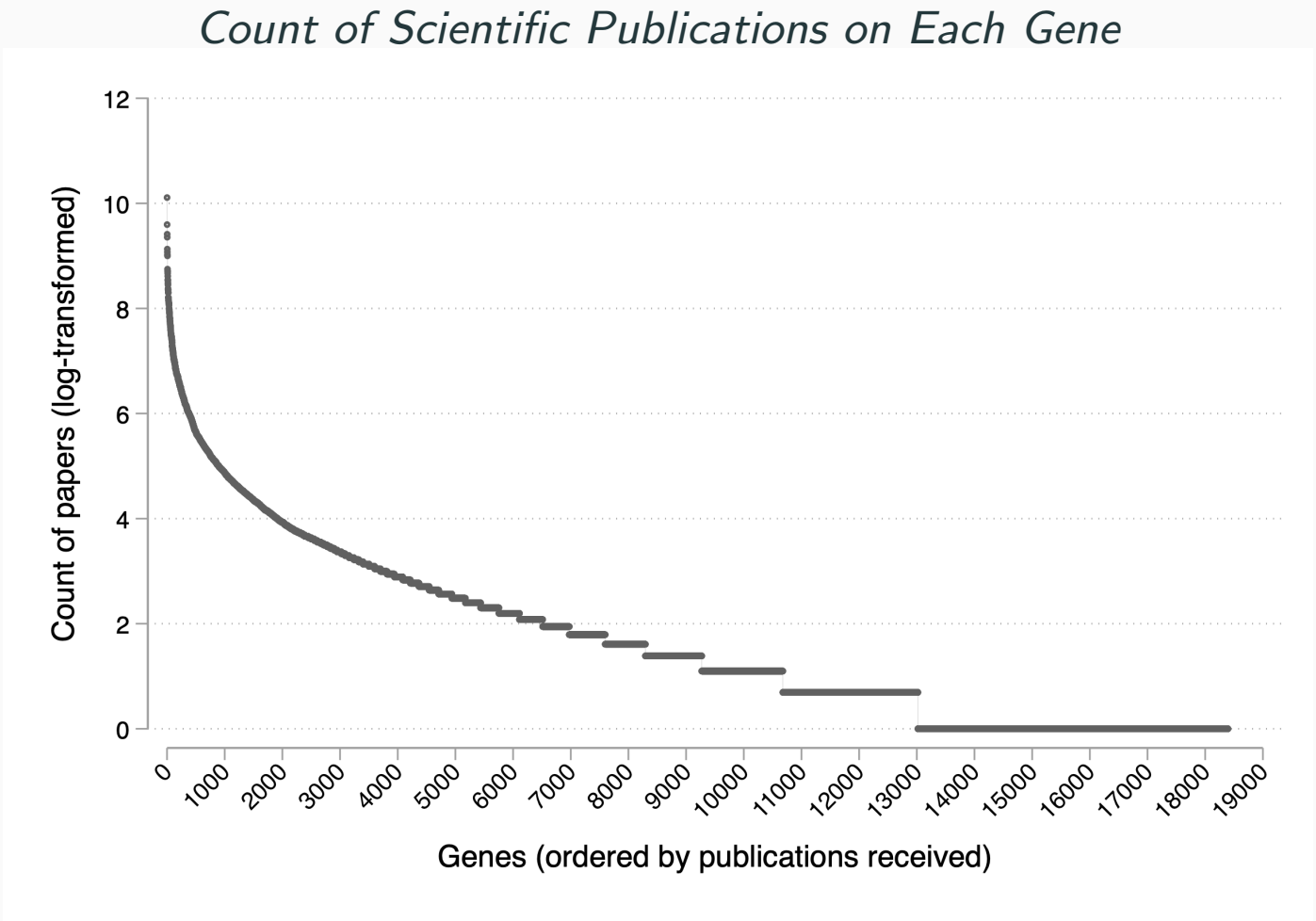
- large set of well-defined topics (human genes)
- data-driven proxy for which genes are more important (GWAS)
- instrument for crowding (mouse orthologs)
- heterogeneity in career incentives (e.g., NIH intramural researchers)

Data: Genetic Determinants of Disease

- 857,025 papers written between 1980 and 2018
- Publications focused on protein-coding genes from NIH's Pubtator3
- Uses NLP to extract the genes studied in papers indexed in PubMed
- List of citations for each publication from NIH's ICite, also provides journal
- Principal Investigator (PI) associated with each publication from Author-ity 2018 (Vetle, 2021)
- Performs robust author name disambiguation for all articles in PubMed
- Author affiliations for each publication from OpenAlex

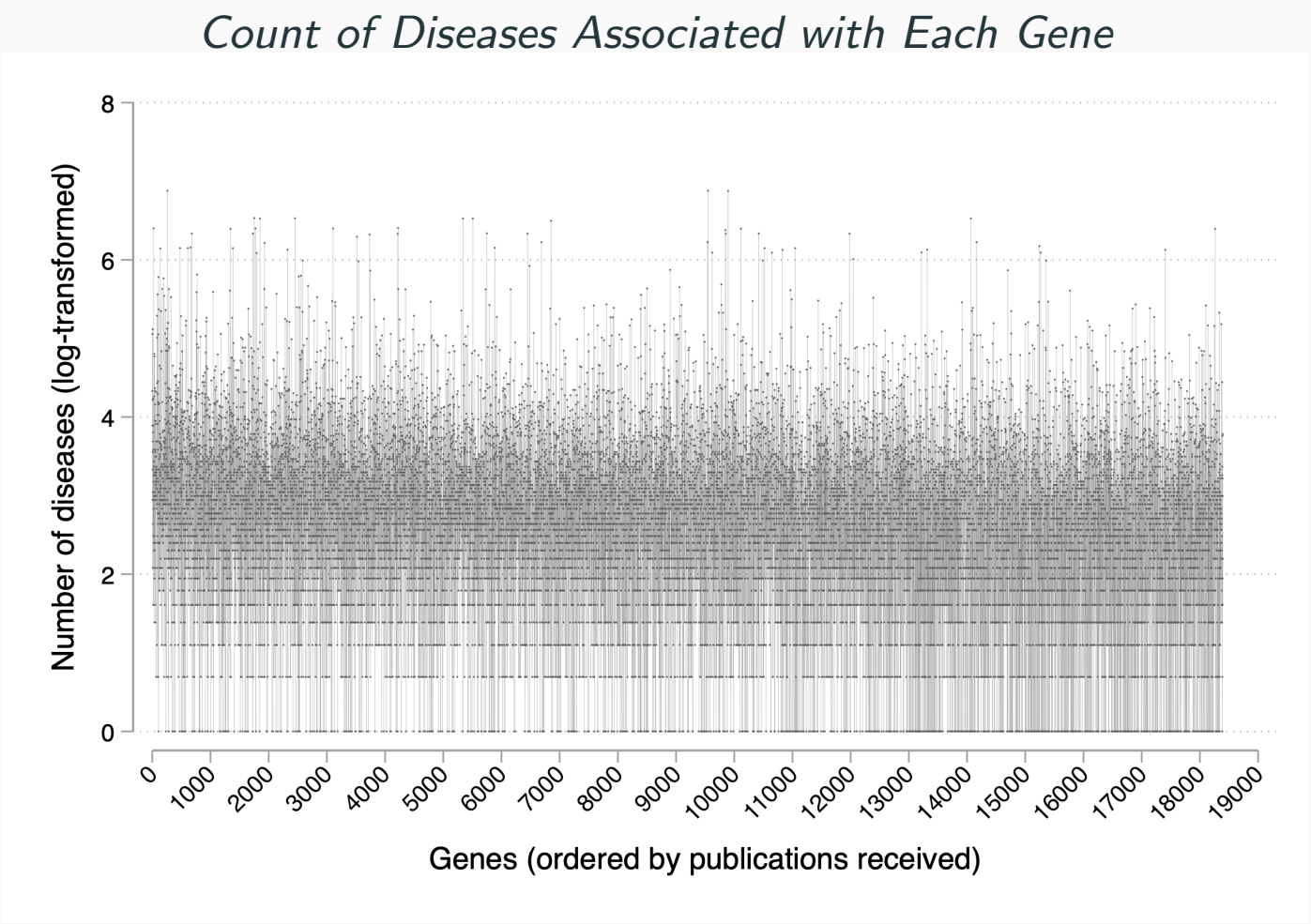
- Number of diseases with which a gene has been associated by unbiased, atheoretical **genome-wide association studies** (GWAS)
  - compiled by U of New Mexico's Target Illumination GWAS Analytics (TIGA)
  - the more a gene harbors mutations associated with diseases, the more likely to be of high scientific importance
  - independent of historical publication pattern

Genetics Research Concentrated on Few Genes...



Papers receive on avg. 2.85 academic citations / year, with a right tail extending to more than 700 citations annually

...But Many Disease-Relevant Genes Remain Neglected



Scientists often remark that many important genes remain overlooked (Soteger et al., 2018, Richardson et. al., 2024)

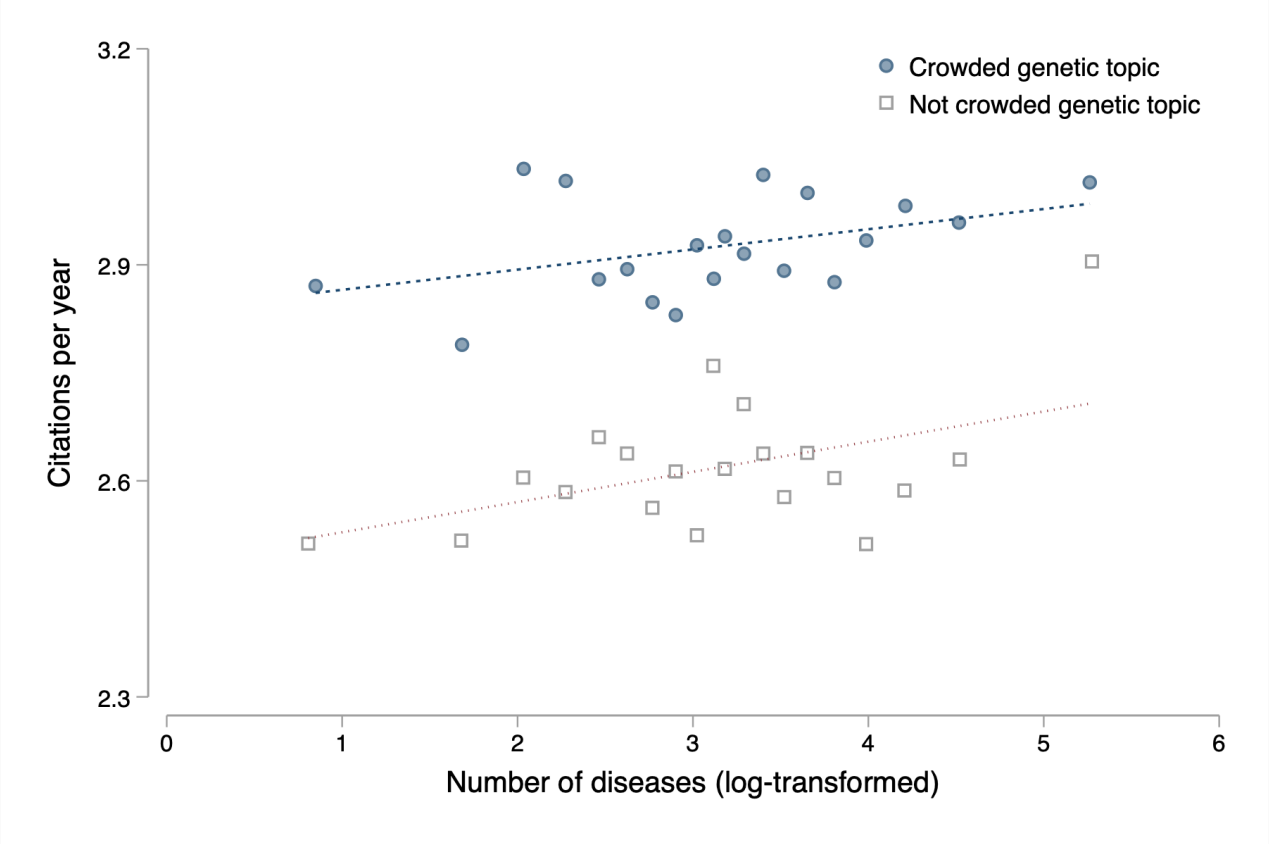
23% publications on understudied genes - cum. publication count below median at publication

Crowded Fields are Cited More Regardless of Ex Ante Disease Importance

Figure plots citations vs ex ante disease importance for above and below the median gene in cumulative publications at time of publication

$$y_{i,t} = \alpha + \theta \cdot \text{importance}_i + \mu_{j,t} + d_c + a_p + \varepsilon_{i,t},$$

(journal-by-year fixed effects ($\mu_{j,t}$), Medical Subject Headings (MeSH) disease-class fixed effects (d_c), and PI fixed effects (a_p))



Papers on below-median genes receive 0.2 fewer citations per year.

Using Human-Mouse Genetic Overlap as an Instrument for Crowdedness

Panel A: Percentile Rank

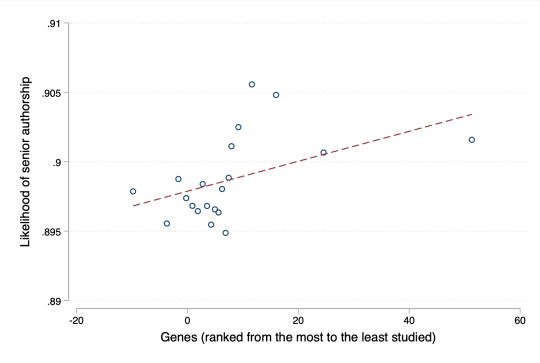
	(1) Inverse Study Rank	(2) Citations	(3) Inverse Study Rank	(4) Citations
Mouse Ortholog (0/1)	-0.356***		-0.361***	
	(0.0634)		(0.0925)	
Inverse Study Rank		-0.454***		-0.254*
		(0.0994)		(0.121)
F-Statistic (First Stage)	31.558		15.241	
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	790,650	790,650	604,906	604,906

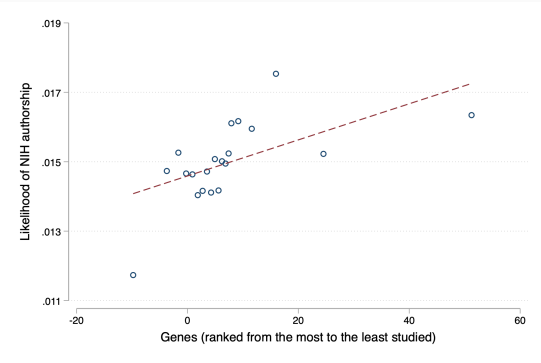
Panel B: Dummy Version

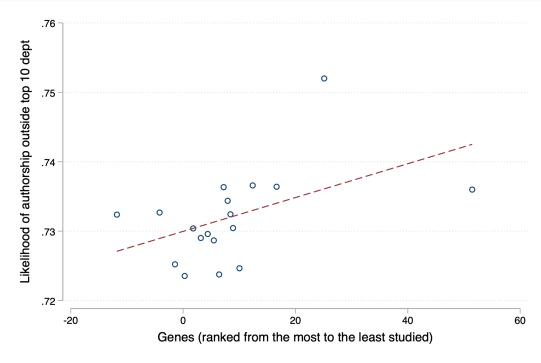
	(1) Understudied (0/1)	(2) Citations	(3) Understudied (0/1)	(4) Citations
Mouse Ortholog (0/1)	-0.0147***		-0.0133***	
	(0.00183)		(0.00263)	
Understudied (0/1)		-11.036***		-6.877*
		(1.979)		(3.098)
F-Statistic (First Stage)	63.947		25.806	
Scientific Importance	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Disease Class FE	Yes	Yes	Yes	Yes
Principal Investigator FE	No	No	Yes	Yes
N	790,650	790,650	604,906	604,906

Heterogeneity in Researchers' Career Concerns

Researchers who are senior, work for NIH as intramural researchers, and (suggestive) outside top 10 universities study less-crowded genes

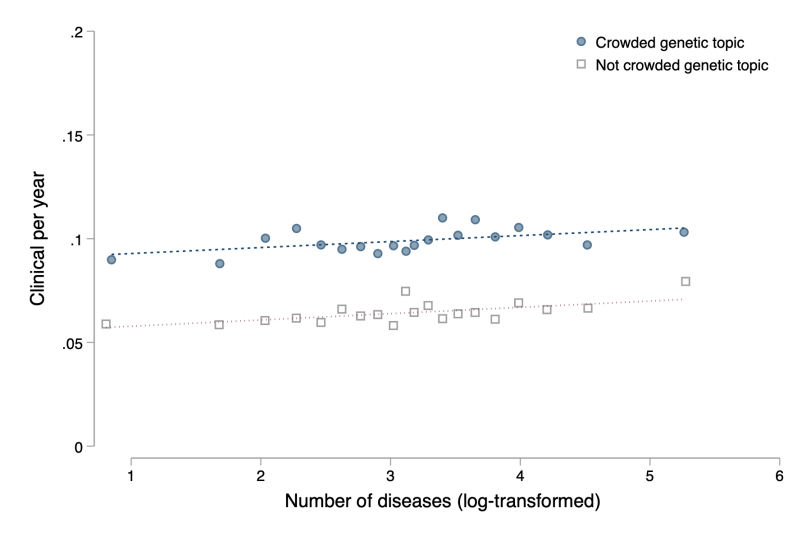


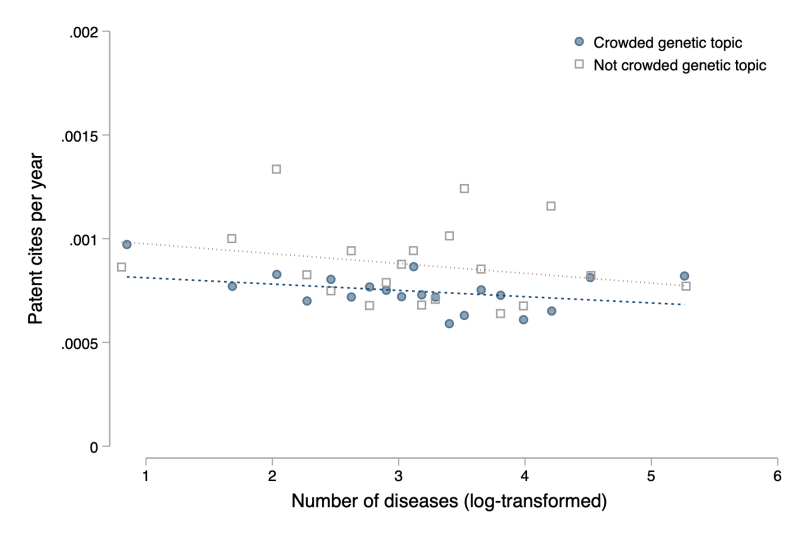




Crowded Genes get More Clinical but Less Firm Patent Citations

Crowding into genes by academic researchers has real effects from pass-through to clinical studies (crucial step in translating discoveries into therapeutic applications)





Crowded and Not crowded based on whether gene is above or below the median gene in cumulative publications at time of publication

(papers receive on avg. 0.09 clinical and 0.001 patent citations per year)

Summary

- Dynamic model of academic research as a coordination game in which rewarding based on citations / publications induces inefficient crowding
  - gives rise to **academic bubbles** in which researchers continue researching a topic even after negative news arrives
- Principal may optimally have to pay for all research to get any research when research effort costs are high or there is risk of coordination failure
  - prizes and awards may encourage exploration but concentrate risk
- Find evidence for predictions using data on genetics research: clearly defined topics and can measure ex ante significance (GWAS)
  - publications on less-crowded genes have a citation penalty
  - researchers with weaker career concerns crowd less
  - pass-through to clinical but not firm patent citations
- If ideas are becoming harder to find because of researcher incentives, then changing these incentives can restore the pace of discovery
- **Next Step:** Dynamic analysis of release of GWAS study on crowding