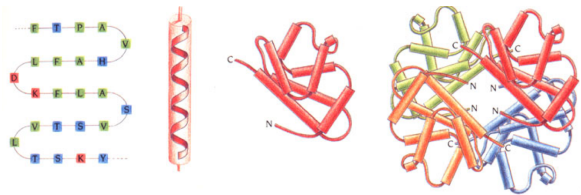
Scooped! Estimating Rewards for Priority in Science

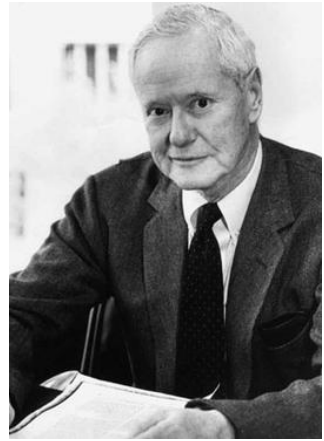
Ryan Hill
Northwestern University
ryan.hill@kellogg.northwestern.edu

Carolyn Stein
Massachusetts Institute of Technology
cstein@mit.edu



Priority in Scientific Discovery

- Priority: Credit given to the individual who first makes a scientific discovery.
- Many notable scientific races (e.g. Newton vs. Leibniz). We all worry about getting scooped!
- There is very little empirical evidence about priority rewards and how racing affects science.



Robert K. Merton

Research Questions

1. What is the causal effect of getting scooped?
   - Short-run effect on project: Publication, journal placement, and citations
   - Long-run effect on career: Future productivity of scientists
2. Does the priority reward system reinforce inequality in science? (Matthew Effect)
   - What drives citations: being first or being famous?

Preview of Findings

- We analyze ~1,600 priority races in structural biology using the Protein Data Bank (PDB).
  - Priority paper gets 54% of total citations and scooped paper gets 46%.
  - Scooped projects are less likely to be published, and less likely to appear in a top-10 journal.
  - In the next five years, scooped scientists have the same number of publications, but fewer citations.
  - Citation penalty is larger for low-ranked teams than it is for high-ranked teams.

A "Race" through the Literature

- Structure of races and distribution of rewards is policy-relevant:
  - Pace, quality, and direction of science
    - Merton (1957), Dasgupta and David (1994), Lerner (1997), Bikard (2013)
  - Strategic behavior of scientists
    - Loury (1979), Lee and Wilde (1980), Dasgupta and Stiglitz (1980), Fudenberg et al. (1983), Bobtcheff et al. (2017)
- Priority and the "Matthew Effect:"
  - Merton (1968), Stephan (1996), Azoulay, Stuart, and Wang (2013), Jin et al. (2019), Hill (2019)

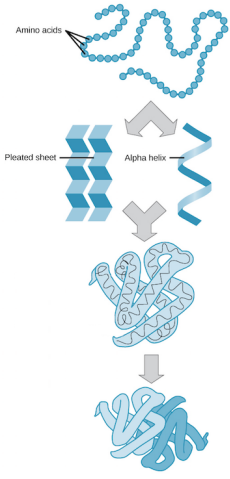
Agenda

Key Empirical Challenges

1. Need a setting with well-defined problems and "one right answer."
2. Need an objective measure of scientific proximity.
3. Need a view of potential abandonments prior to publication.

What is Structural Biology?

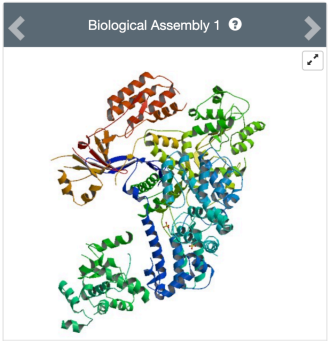
- Structural biologists determine the molecular structure of proteins, DNA, and RNA.
- Proteins carry out most of the functions within cells, and often "form determines function."
- Structures are solved by X-ray crystallography. Successful experiments result in diffraction data and a model that describes the protein shape.



The Protein Data Bank

- The Protein Data Bank (PDB) contains structural data of 100,000+ proteins and meta-data about projects.
- Major scientific journals require scientists to submit their structure data to the PDB before publication.
- All structures are deposited confidentially a few months before article publication.
- Bioinformatics algorithm links projects with identical biological features.

PDB Example: Cas-9



Macromolecule Content

- Total Structure Weight: 318476.84
- Atom Count: 18888
- Residue Count: 2744
- Unique protein chains: 1

Display Files | Download Files

4CMP unique structure ID

Crystal structure of S. pyogenes Cas9

DOI: 10.2210/pdb4CMP/pdb

Classification: HYDROLASE
Organism(s): Streptococcus pyogenes serotype M1
Expression System: Escherichia coli BL21(DE3)

key dates

Deposited: 2014-01-16 **Released:** 2014-02-12
Deposition Author(s): Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A.

Literature | Download Primary Citation

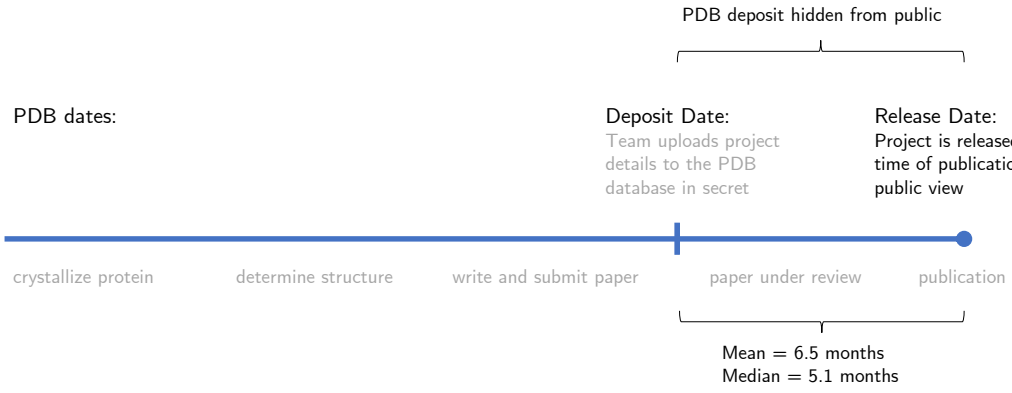
Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A.
(2014) Science **343**: 47997

PubMed: 24505130 Search on PubMed | Search on PubMed Central
DOI: 10.1126/science.1247997
Primary Citation of Related Structures:
4OGE, 4OGC, 4CMQ

PubMed Abstract:
Type II CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated) systems use an RNA-guided DNA endonuclease, Cas9, to generate double-strand breaks in invasive DNA during an adaptive bacterial immune response. Cas9 h ...

Project Timeline

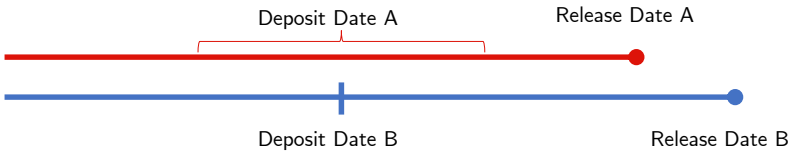


*If project goes unpublished, data is released publicly after one year

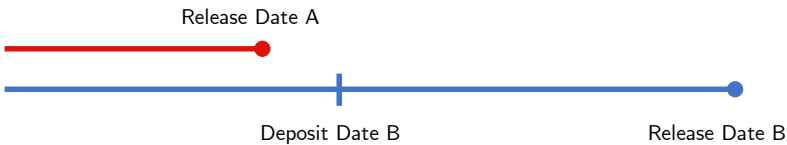
Scoop Definition

Rules: **1.** Take two projects that have identical sequence, different authors.
2. Assert that both projects are deposited before the first project is released.
3. Call the first to release the winner, call the second project "scooped."

Scenario 1: Project A scoops Project B

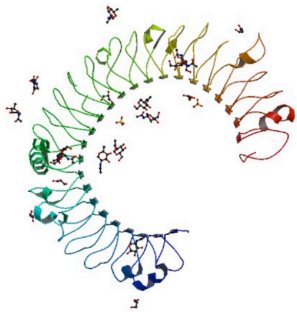


Scenario 2: Project A and Project B are excluded from racing sample



Example Race: Toll-like Receptor 3

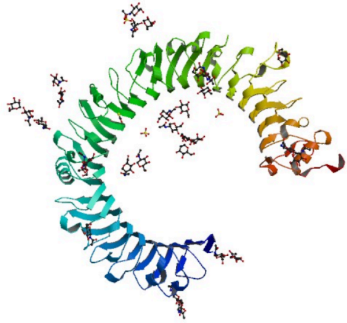
Winning Deposit: 1ZIW



Affiliation: Scripps Research Institute
Deposit Date: April 27, 2005
Release Date: June 28, 2005

Journal: *Science*
Journal Impact Factor: 30.9
5-year Citations: 196

Scooped Deposit: 2A0Z



Affiliation: National Institutes of Health
Deposit Date: June 27, 2005
Release Date: August 2, 2005

Journal: *PNAS*
Journal Impact Factor: 10.2
5-year Citations: 129

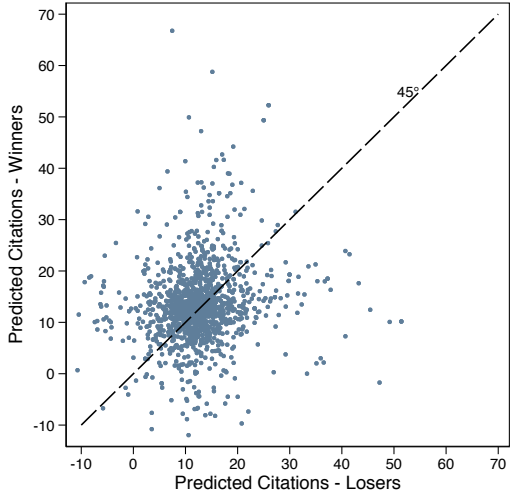
Predicted Citation Balance

Race winners are not randomly assigned, but seem highly unpredictable.

Lasso model of predicted citations:

- Team size and age
- Past deposits and publications
- University rank and location

Difference in predicted citations: 0.212 (p-value = 0.587)



Estimating the Scoop Penalty

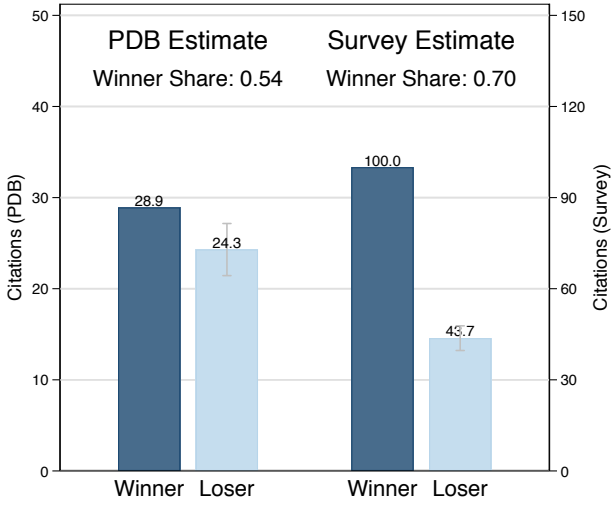
- Basic specification: For deposit i of protein (race) p:

$$Y_{ip} = \alpha + \beta Scooped_i + X_i'\delta + \gamma_p + \epsilon_{ip}$$

where

- $Scooped_i$ is a dummy for losing priority race.
- γ_p is the coefficient on a protein (i.e. race) fixed effect.
- X_i is a vector of individual and lab controls selected by PDS-Lasso method (Belloni et al. 2014).

Citation Penalty



Scoop Penalty - Results

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	asinh(Five-year citations) (4)	Top-10% five year citations (5)
Scooped	-0.025***	-0.178***	-0.060***	-0.197***	-0.035***
	(0.010)	(0.032)	(0.014)	(0.045)	(0.010)
Winner Y mean	0.880	-0.031	0.318	28.918	0.150
Observations	3,319	3,319	3,319	2,546	2,546

Note: All regressions include controls selected by PDS-Lasso as well as year fixed effects. Unpublished papers have impact factor imputed to minimum factor journal. Citation regressions restricted to papers published before 2014. Column 4 dependent variable is asinh(five-year citations) but mean citations is reported in levels.

The Long-Run Consequences of Being Scooped

- Long run outcomes (excluding winning/scooped paper):
  - Active in PDB five years later
  - Total publications - five years
  - Total citations - five years
- Estimate for scientist s, deposit i, for protein (race) p:

$$Y_{isp} = \alpha + \beta Scooped_{is} + X'_{is}\delta + \gamma_p + \epsilon_{isp}$$

- Estimate separately for novices (<1 year of PDB experience) and veterans.

The Long-run Consequences of Being Scooped

		Total count five years after race (excluding original paper)			
Dependent variable	Active in PDB 5 years later (1)	Publications (2)	Top-10 publications (3)	asinh of three-year citations (4)	Top-10% of three year citations (5)
Panel A. All scientists					
Scooped	-0.023**	0.325	0.047	-0.199***	-0.159***
	(0.010)	(0.264)	(0.105)	(0.053)	(0.052)
Winner Mean Y	0.797	13.541	4.441	187.129	1.529
Observations	6,642	12,488	12,488	9,297	9,297
Panel B. Novices					
Scooped	-0.013	-0.054	-0.076	-0.282**	-0.121***
	(0.022)	(0.183)	(0.065)	(0.112)	(0.039)
Winner Mean Y	0.587	2.667	0.929	50.692	0.440
Observations	2,273	3,554	3,554	2,868	2,868
Panel C. Veterans					
Scooped	-0.024**	0.504	0.118	-0.167***	-0.165**
	(0.009)	(0.373)	(0.145)	(0.051)	(0.071)
Winner Mean Y	0.930	19.042	6.221	263.894	2.143
Observations	4,027	8,251	8,251	5,913	5,913

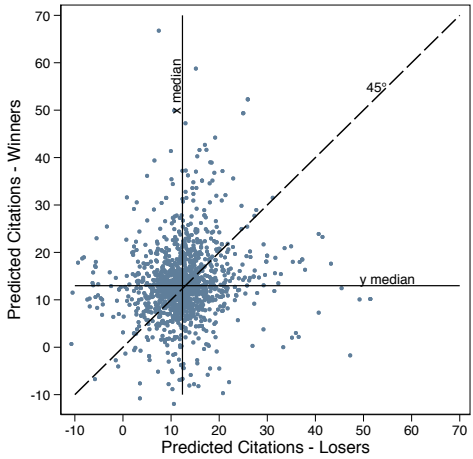
Agenda

Priority and Inequality

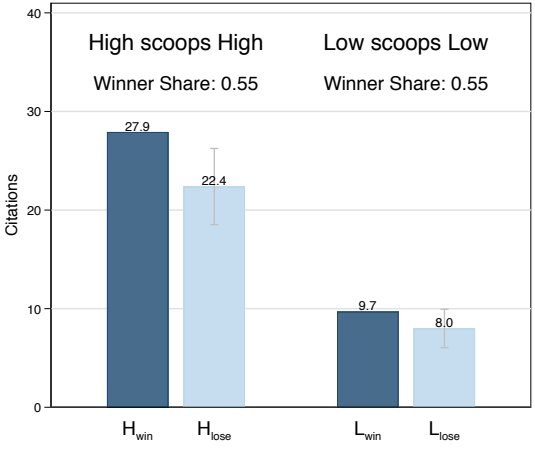
- Merton proposes two key drivers of academic attention:
  - Priority
  - Matthew Effect
- We test which of these effects dominates by comparing citations in races between high- and low-reputation teams.
- See the statistical discrimination model in the paper.

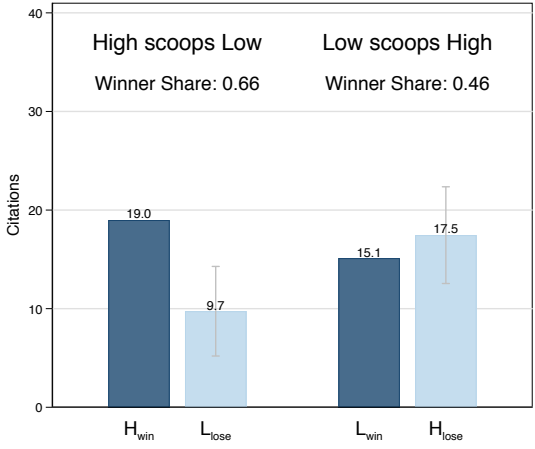
Defining Reputation

- Define pre-existing reputation using LASSO-generated predicted citations.
- Define H teams as those with above median predicted citations and L teams as those with below median.



Evenly-matched and Mismatched Races





Conclusion

Getting scooped lowers citations, but rewards are more evenly distributed than previously thought.

Normative implications: Is the premium for priority too large or too small?

- Priority may incentivize effort and timely disclosure.
- Racing may incentivize speed at the expense of quality and transparency.

Thank You!

- Ryan Hill: ryan.hill@kellogg.northwestern.edu
- Carolyn Stein: cstein@mit.edu