

# Intellectual Property Rights and Innovation Evidence From the Human Genome

by Heidi Williams

Konstantin Golyaev

University of Minnesota

February 9, 2010

# Motivation

- Competitive markets may fail to provide incentives for innovation
- Intellectual Property (IP) rights regulations aim to correct this market failure
- Effect of IP on subsequent innovation is ambiguous, especially in markets with cumulative technological process
  - trade-off between paying too much for access to innovation and losing all profits to high competition
- Empirically hard to measure: can expect selection into IP “treatment”

# Human Genome Sequencing

- The Human Genome Project (public) and Celera (private firm) sequenced human genome independently
  - HGP made all their results public
  - Celera kept its results private and obtained IP on them
- Once HGP re-sequenced Celera's results, they became public
- Can compare genes that ever had Celera's IP with those that were sequenced by HGP
  - and look at subsequent innovations across two groups

# Results Preview

Celera IP on genes has strong negative impact on future research and product development

- 35% fewer publications since 2001
- 16% points reduction in chance of gene having known uncertain genotype-phenotype link
- 2% points reduction in chance of gene having known and certain genotype-phenotype link
- 1.5% points less likely to be used in genetic tests

Also, Celera genes have not “caught up” with ex-ante similar genes sequenced by HGP as of 2009

# Outline

- 1 Introduction
- 2 Industry Background
- 3 Data and Preliminary Evidence
- 4 Results

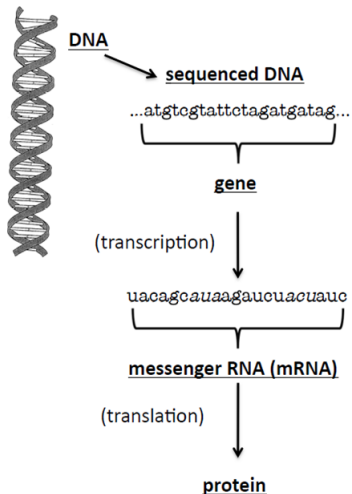
# A Primer on Human Genome

- **Genome** is like a source code for creating an organism
- Human genome is composed of **DNA**, which is a sequence of 3 bln nucleotide bases – A, C, G, and T
  - Sequencing genome means determining the order of the above bases in the DNA
- DNA are organized into 46 chromosomes, which carry **genes** (about 28, 000 in humans)
- Genes encode instructions to generate **proteins**, which carry out important body functions
- Two-step process
  - 1 Transcription – messenger RNA (**mRNA**) gets generated
  - 2 Translation – mRNA is used to generate a protein

mRNA encodes exactly one protein, and hence is our unit of tracking sequencing efforts

# Sequencing: Illustration

Figure 1: Overview of Scientific Background on the Sequencing of the Human Genome



# Chronology of Events

- 1990 – DOE and NIH launched the HGP, aimed to finish by 2005
- May 1998 – Celera founded, claimed to finish sequencing in 3 years
- Sep 1998 – HGP announced they will finish by 2003
- 2001 – both sides publish their results so far, Celera ceases operation
- 2003 – HGP completes the sequencing

# Sequencing and IP Strategies

- HGP made all their results public within 24 hours of discovery
- Genes sequenced by Celera only were kept as its IP
  - Sold data on sequenced DNAs to pharmaceutical companies and academia
  - Former paid \$5 – \$15 mln a year, latter – \$7500 – \$15,000 per lab
  - Some paid even though all data was about to be public at some point
- Quality of sequencing seems to be comparable and high

# Data

- Use mRNA as unit of sequencing efforts
- Aggregate mRNAs to genes, use genotype-phenotype links as economically meaningful outcome variables
  - genotype-phenotype links are links between certain gene and observable body function
- Use NIH RefSeq database for all mRNA data
- Use paper by Istrail et. al. (2004) for information on which genes where sequenced by Celera
- Use OMIM database to measure amount of “scientific knowledge” by gene
- Use GeneTests.org database for tracing availability of gene-based tests

Table 1: Summary Statistics for Gene-Level Data

	mean	standard deviation	minimum	maximum
<b>Panel A: Celera intellectual property (IP)</b>				
0/1, Celera gene	0.060	0.238	0	1
<b>Panel B: Outcome variables</b>				
publications in 2001-2009	2.197	9.133	0	231
0/1, known, uncertain phenotype	0.453	0.498	0	1
0/1, known, certain phenotype	0.081	0.273	0	1
0/1, used in any diagnostic test	0.060	0.238	0	1
<b>Panel C: Main covariates</b>				
year first mRNA disclosed	2002.962	3.551	1999	2009
publications in 1970	0.032	0.323	0	18
publications in 1971	0.027	0.262	0	18
publications in 1972	0.036	0.349	0	26
publications in 1973	0.029	0.301	0	26
publications in 1974	0.037	0.362	0	25
publications in 1975	0.039	0.412	0	35
publications in 1990	0.139	0.936	0	57
publications in 1991	0.158	0.968	0	46
publications in 1992	0.189	1.177	0	57
publications in 1993	0.176	0.990	0	32
publications in 1994	0.190	0.962	0	31
publications in 1995	0.232	1.125	0	31
publications in 1996	0.244	1.119	0	34
publications in 1997	0.258	1.158	0	33
publications in 1998	0.283	1.157	0	35
publications in 1999	0.289	1.188	0	32
<b>Panel D: Additional covariates</b>				
0/1, missing cytogenetic location	0.370	0.483	0	1
0/1, missing molecular location	0.059	0.235	0	1
$N = 27,882$				

## First Look

Table 2: Differences Across Celera and non-Celera Genes in Gene-Level Data

non-Celera genes sequenced in:	(1) -	(2) all	(3) all	(4) 2001	(5) 2001	(6) ≥2000	(7) ≥2000
	Celera mean	mean	<i>p</i> - value	mean	<i>p</i> - value	mean	<i>p</i> - value
<b><u>Panel A: Outcome variables</u></b>							
publications in 2001-2009	1.239	2.258	[0.000]	2.116	[0.000]	1.083	[0.250]
0/1, known, uncertain phenotype	0.401	0.456	[0.000]	0.563	[0.000]	0.301	[0.000]
0/1, known, certain phenotype	0.046	0.083	[0.000]	0.073	[0.000]	0.038	[0.126]
0/1, used in any diagnostic test	0.030	0.062	[0.000]	0.054	[0.000]	0.027	[0.430]

# Main Specification

$$outcome_g = \beta \cdot \mathbf{1} \{celera_g\} + \lambda' x_g + \varepsilon_g,$$

- $g$  indexes genes
- $x_g$  — observable covariates – number of prior publications and time dummies
- Poisson regression or OLS
- Possible concern: selection into “treatment” (i.e. Celera picked genes to sequence not at random)
  - panel specification to follow

## Main Specification: Results

Table 3: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes:  
Sample of Genes Sequenced in or after 2000

	(1)	(2)
<b>Panel A: publications in 2001-2009</b>		
mean = 1.095		
<i>celera</i>	-0.535 (0.117)***	-0.432 (0.112)***
<b>Panel B: 0/1, known, uncertain phenotype</b>		
mean = 0.309		
<i>celera</i>	-0.162 (0.015)***	-0.158 (0.015)***
<b>Panel C: 0/1, known, certain phenotype</b>		
mean = 0.039		
<i>celera</i>	-0.027 (0.007)***	-0.018 (0.006)***
<b>Panel D: 0/1, used in any diagnostic test</b>		
mean = 0.027		
<i>celera</i>	-0.023 (0.006)***	-0.015 (0.005)***
indicator variables for year of disclosure	yes	yes
number of publications in each year 1970-77	no	yes
<i>N</i>	21,824	21,824

# Panel Specification

$$outcome_{gy} = \alpha_g + \gamma_y + \beta \cdot \mathbf{1} \{celera_{gy}\} + \varepsilon_{gy},$$

- $y$  – indexes years
- $\alpha_g$  – gene fixed effect
- $\gamma_y$  – year fixed effect
- Results largely similar to main specification

# Panel Specification: Results

Table 5: Panel Estimates of the Impact of Celera IP on Innovation Outcomes:  
Sample of Genes Sequenced in or after 2000

	(1)	(2)	(3)
<b>Panel A: gene-year publications</b>			
mean = 0.122			
<i>celera</i>	-0.112 (0.017)***	-0.084 (0.014)***	-0.052 (0.010)***
<b>Panel B: 0/1, known, uncertain phenotype</b>			
mean = 0.223			
<i>celera</i>	-0.151 (0.009)***	-0.148 (0.009)***	-0.068 (0.008)***
year fixed effects	yes	yes	yes
indicator variables for year of disclosure	yes	yes	-
number of publications in each year 1970-77	no	yes	-
gene fixed effects	no	no	yes
<i>N</i>	196,416	196,416	196,416

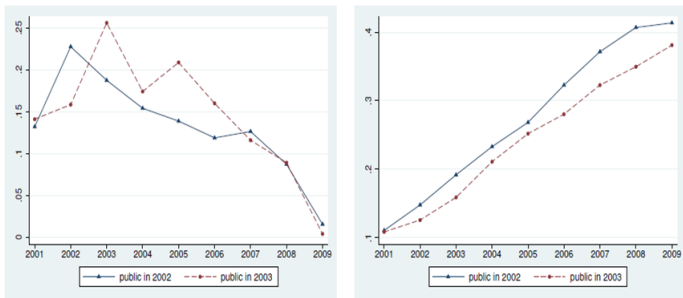
# Within-Celera Genes Variation

Only look at genes first sequenced by Celera here

- Some were re-sequenced by HGP in 2002, other in 2003
- Compare variation in outcome variables between these subgroups
- Find the non-catching-up effect

# Within-Celera Genes: Results

Figure 4: Average Innovation Outcomes for Celera Genes by Year, by Year of Re-sequencing by the Public Effort



- (a) Outcome variable: Gene-year publication count (b) Outcome variable: Indicator for a gene having any known/uncertain phenotype link in that year

# Conclusion / Comments

- IP on Celera genes has a significant negative effect on future innovations based on those genes
- Should not conclude IP was welfare decreasing
  - Sequencing might had been accelerated by competition
- Overall, good treatment effect paper
  - lots of data work, little econometrics