How much research is spurred when a gene is associated with a complex disease?

Travis J. Struck,[1] Brian K. Mannakee,[2] and Ryan N. Gutenkunst[1]

[1] *Department of Molecular and Cellular Biology, University of Arizona*

[2] *Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona*

The impact of genome-wide association studies (GWAS) on biomedical research output was quantitatively evaluated. GWAS have not changed the historical skew of publications toward genes involved in Mendelian diseases, but genes newly implicated by GWAS in complex disease do experience a modest increase in publication activity. The impact of GWAS on biomedical research into individual genes is declining, however, even as the rate of new GWAS associations increases.

A decade after the first publication of a successful GWAS [1], few results from GWAS have had clinical impact, because most associated variants have modest effect sizes or unclear functional consequences [2, 3]. Direct clinical impact is, however, not the only goal of GWAS. Another major goal is to identify and steer research toward novel genes involved in complex diseases [4]. For example, the first published GWAS unexpectedly identified Complement Factor H as associated with macular degeneration [1], spurring the development of complement-based therapeutics [5]. Similarly, most genes associated with multiple sclerosis through GWAS had not previously been considered candidates [6]. But what impact have GWAS had beyond these paradigm examples? GWAS are more highly cited than comparable candidate gene studies [7], but how much follow-up research do GWAS actually motivate? To answer this question, we quantified the effects of GWAS on biomedical research output.

We measured research output on genes using scientific publications, as collected in the NCBI Gene database [8]. We prefer this manually curated database to automatic text mining, because text mining may introduce false positives when a gene is mentioned in passing. We classified genes into those associated with Mendelian disease, complex disease, both, or no disease using the Online Mendelian Inheritance in Man (OMIM) database [9] and the EBI-NCBI GWAS catalog [10].

As expected [11, 12], we found that in the pre-GWAS era research was skewed toward a minority of human genes (Fig. 1A). The majority of highly-studied genes were involved in Mendelian disease, and many genes that would later be associated with complex disease received little attention (Fig. 1B). In the post-GWAS era, research output is even more skewed (Fig. 1C; coefficient of variation 3.4 vs 2.5). Most highly studied genes are still ones involved in Mendelian disease, and many genes associated with complex disease still receive little attention (Fig. 1D). By contrast, research output on yeast genes became much less skewed following the publication of the yeast genome (Fig. S1). The advent of GWAS has not reduced the bias of biomedical research toward Mendelian disease genes. But how has GWAS affected research on individual genes?

To quantify the immediate effect of GWAS on individual "hit" genes, we focused on the year a gene was associated with complex disease through GWAS and the following two years. For each new GWAS hit, we compared publications over this period with a control gene chosen to have as similar a prior publication history as possible. We found that association with complex disease through GWAS does indeed result in modestly more research on a gene (Fig. 2A). The median GWAS hit experienced 2 additional publications in the following three years, and the average number of additional publications per GWAS hit was 2.8.

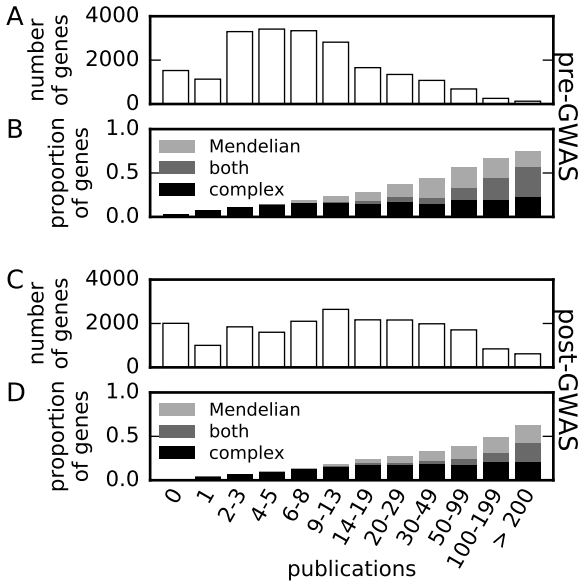

FIG. 1. Biomedical research output is skewed toward genes involved in Mendelian disease, even after the advent of GWAS. A: In the pre-GWAS era (before 2005), the distribution of publications among genes was highly skewed. B: Highly-studied genes tended to be involved in Mendelian disease. C: In the post-GWAS era (2005 and beyond), the distribution of research output is even more highly skewed. D: Highly-studied genes still tend to be those involved in Mendelian disease, with many genes involved in complex disease receiving little study.

What factors determine how much additional research effort is expended on a new GWAS hit? The strength of an association is quantified by its statistical p-value and its effect size, most commonly an odds ratio. We found that the additional publications on a gene and the p-value of its association are weakly negatively correlated

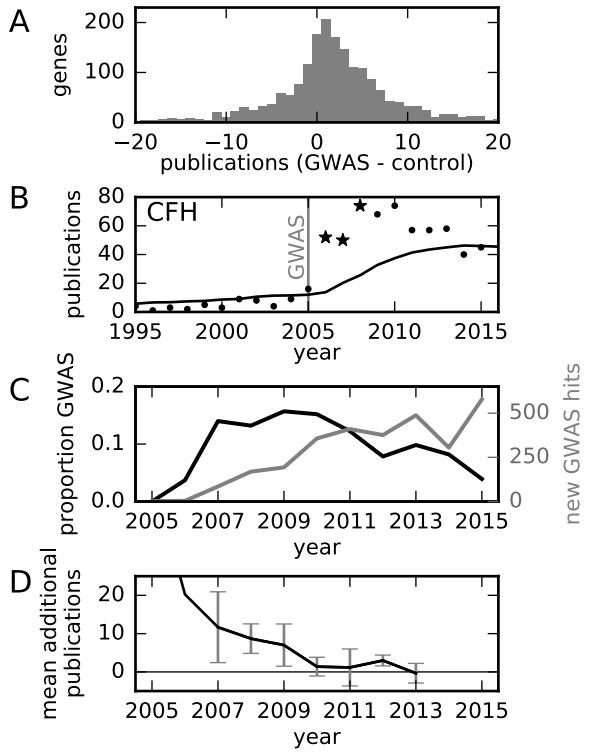

FIG. 2. The effect of GWAS on the output of research into associated genes is modest and declining. A: Genes newly implicated in complex disease via GWAS experience a modest increase in publications in the following three years. B: A significantly elevated number of studies were published on Complement factor H following its association with macular degeneration via GWAS. Solid line is the predicted publication history from the null model, points indicate actual publication counts, and starred points indicate years with a statistically significant excess of publications (one-sided Bonferroni-corrected $p < 0.05$). C: The proportion of genes exhibiting an unusual excess in publications that were recently identified in GWAS peaked at roughly 16% in 2009 and has since declined, even as the rate at which genes are newly associated via GWAS has increased. D: The mean number of additional publications (compared to control genes) for genes newly associated in GWAS has declined. Error bars denote 95% standard errors of the mean. Standard errors for 2005 and 2006 (not shown) are large due to the small number of genes in those years.

(Spearman rank correlation ρ = -0.069, N = 2,084, p = 0.0018). Additional publications and odds ratios are weakly positively correlated (Spearman rank correlation $\rho = 0.062$, N = 1,143, p = 0.035). Researchers are thus more likely to follow up on associations with higher statistical confidence and larger effect size, although both effects are weak.

The typical GWAS hit receives a modest increase in research effort, but some genes may receive large increases. To identify such genes, we used the model of Pfeiffer and Hoffmann [13] to predict the number of publications for each gene in each year, based on that gene's prior publication history. We trained the model on all genes never implicated in complex disease through GWAS. By comparing model predictions and publication data, we then identified particular years in which particular genes had unexpectedly large numbers of publications (Table S1). For example, Complement Factor H had a significant excess of publications in the three years after its association with macular degeneration (Fig 2B).

How has the impact of GWAS changed over time? Soon after the advent of GWAS, recent GWAS hits made up a substantial fraction of genes with excess publications in each year, but that fraction has declined dramatically, even as the number of new GWAS hits has increased (Fig. 2C). Moreover, the mean number of additional publications on a new GWAS hit is smaller for recent hits than early hits (Fig. 2D). The impact of GWAS on biomedical research into individual genes is thus declining.

By quantitatively analyzing the publication histories of genes associated with complex disease through GWAS, we have shown that the effect of GWAS on biomedical research into individual genes has been modest and is declining. Research output remains highly skewed toward Mendelian disease genes, with many complex disease genes receiving little attention (Fig. 1D). Early GWAS hits were subject to substantial additional study, but later GWAS hits have received little additional attention (Fig. 2D).

Many factors may be contributing to the declining impact of GWAS. For example, follow-up studies may be slower than in the early years of GWAS, due to increasing recognition of the potential complexity of the links between associated variants, functional variants, and target genes [14]. Availability of funding may also constrain follow-up studies, particularly because biomedical research spending has declined in both North America and Europe [15]. In any case, our results suggest that reforms may be needed for GWAS to reach the goal of understanding the mechanisms linking GWAS-identified genes to disease.

AUTHORS' CONTRIBUTIONS

RG, TS, and BM designed the analysis, which RG and TS carried out. RG and TS wrote the manuscript.

ACKNOWLEDGMENTS

We thank Yann Klimentidis and Tricia Serio for helpful comments.

FUNDING

This work was supported by DARPA contract WF911NF-14-1-0395 to RG. It was also supported by

the NSF, via Graduate Research Fellowship grant DGE-1143953 to BM. BM was also supported by an Achievement Rewards for College Scientists scholarship.

METHODS

a. Publication data. We obtained Entrez GeneIDs for all human protein-coding genes from NCBI Gene [8] on September 16, 2016. For all those genes, we collected PubMed identifiers of associated publications from NCBI Gene's gene2pubmed file, downloaded September 16, 2016. This file contains both associations created manually during the curation of Gene References Into Function (GeneRIFs) and associations collected from organism-specific databases, Gene Ontology, and other curated data sources. We then obtained date information for each publication from PubMed using BioPython [16]. We followed a similar procedure for yeast genes.

b. Disease data. To identify genes associated with Mendelian disease, we downloaded the OMIM Gene Map of connections from genes to traits [9] on December 8, 2016. We filtered to keep only entries with a confidence code of "confirmed" and to ignore entries indicating a potentially spurious mapping or association with a non-disease trait. We further considered only entries with Entrez GeneIDs, to avoid ambiguity among gene names and aliases. This procedure yielded 1,853 genes associated with disease traits. Of these, 1,517 genes were associated with Mendelian but not complex multifactorial disease, 163 were associated with complex multifactorial but not Mendelian disease, and 173 were associated with both Mendelian and complex multifactorial disease.

To further identify genes associated with complex disease and to gather GWAS data, we used the November 28th, 2016 release of NHGRI-EBI's GWAS Catalog [10]. We filtered the catalog to remove nondisease traits, by keeping only entries that were children of the term "disease" (`EFO_0000408`) in the Experimental Factor Ontology [17]. We considered a gene to be associated with a disease if an associated SNP was within that gene or within 500 basepairs of that gene, considering only genes with a reported Entrez GeneID. This procedure yielded 2,983 genes associated with complex disease.

Our analysis of OMIM and the GWAS catalog yielded 4,382 total disease-associated genes. Considering genes associated with only Mendelian disease in OMIM and not associated with disease through GWAS yielded 1,202 Mendelian disease genes. Considering genes associated with only complex multifactorial disease in OMIM or associated with disease through GWAS yielded 2,691 complex disease genes. The remaining 489 genes we associated with both Mendelian and complex disease.

Of the disease genes in the GWAS catalog, 2,084 were first associated prior to 2014, so we could analyze three full years of publication data. For those genes, we identified odds ratios as effect sizes without units for SNPs that had a reported frequency of the risk allele. For our odds ratio analysis, we analyzed the 1,143 genes for which an odds ratio was reported in the first year of GWAS association.

c. Control genes. For each of our 2,084 GWAS genes, we identified its control gene as the non-GWAS gene with the closest number of total publications prior to the year the gene was first associated with complex disease. If multiple genes were tied for closest, we compared the previous year as well, continuing either until there was no ambiguity or until we reached 1950. For the 224 GWAS genes with ambiguous control genes, we compared publications between the GWAS gene and the average of the control genes.

d. Publication rate model. We used the model of Pfeiffer and Hoffman to predict expected per-gene publication rates [13]:

$$\Delta P_{i,t+1} = \frac{k_1 P_t^* + k_2 P_{i,t} + k_3}{1 + (P_t^*/P_S)^\alpha}. \quad (1)$$

Here, $\Delta \mathrm{P}_{i,t+1}$ is the predicted number of publications for gene i in year $t+1$, and $P_{i,t}$ and P_t^* are the cumulative number of publications in previous years for the gene and the average cumulative number of publications for all genes in the organism, respectively. The term in the denominator models saturation of publication rates. The three rate parameters, k_1, k_2, and k_3, and the saturation parameters, P_S and α, were assumed to be identical for all genes. To fit the parameters to our data, we constructed a likelihood function by assuming that the number of publications each year for each gene was independently Poisson distributed with mean $\Delta P_{i,t+1}$ given by Eq. 1. We then maximized that likelihood with respect to the five model parameters, using publication data from 1950 to 2015 for all non-GWAS genes. The maximum-likelihood parameter values were $k_1 = 0.0230$, $k_2 = 0.235$, $k_3 = 0.00248$, $P_S = 23.3$, $\alpha = 1.37$. Five genes each had 1 publication prior to 1950 that was not included in the data fit.

To identify years in which genes had significantly elevated publication rates, our null model was that publications were Poisson distributed with mean given by Eq. 1. Significant gene-years were defined as those in which the probability of generating at least the observed number of publications was less than the Bonferroni-corrected significance cutoff $0.05/(N_g N_y)$. Here $N_g = 20,681$ was the total number of genes considered and $N_y = 66$ was the total number of years.

SUPPLEMENTARY MATERIAL

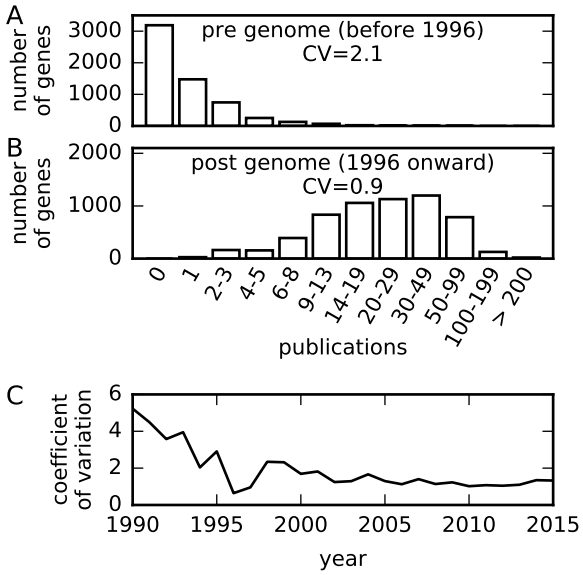

Figure S1: Biomedical research effort into yeast genes is less skewed than human genes. A: Before the release of the yeast genome sequence, a large number of yeast genes were unstudied. The overall distribution was, however, narrower for yeast genes than human genes (coefficient of variation 2.1 vs. 2.5). B: After the release of the genome, research effort was substantially more evenly distributed among yeast genes. C: The distribution of research effort among yeast genes has steadily become more uniform, as quantified by the coefficient of variation of publication counts in each year.

Table S1: Gene-years with a statistically significant excess of publications relative to the prediction of the Pfeiffer and Hoffman model. For GWAS disease genes, the date of the first GWAS to identify that gene is also recorded.

[1] Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–9 (2005).

[2] Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 17–19 (2009).

[3] Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14**, 549–58 (2013).

[4] Hirschhorn, J. N. Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).

[5] Gehrs, K. M., Jackson, J. R. & Brown, E. N. Complement, age-related macular degeneration and a vision of the future **128**, 349–358 (2010).

[6] Ricigliano, V. A. G. *et al.* Contribution of genome-wide association studies to scientific research: a pragmatic approach to evaluate their impact. *PLoS One* **8**, 1–5 (2013).

[7] Mansiaux, Y. & Carrat, F. Contribution of genome-wide association studies to scientific research: a bibliometric survey of the citation impacts of GWAS and candidate gene studies published during the same period and in the same journals. *PLoS One* **7**, 1–4 (2012).

[8] Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42 (2015).

[9] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).

[10] MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).

[11] Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470**, 163–165 (2011).

[12] Isserlin, R. *et al.* The human genome and drug discovery after a decade. Roads (still) not taken 14 (2011). arXiv:1102.0448.

[13] Pfeiffer, T. & Hoffmann, R. Temporal patterns of genes in scientific publications. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12052–6 (2007).

[14] Edwards, S. L., Beesley, J., French, J. D. & Dunning, M. Beyond GWASs: Illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).

[15] Chakma, J., Sun, G. H., Steinberg, J. D., Sammut, S. M. & Jagsi, R. Asia's ascent global trends in biomedical R&D expenditures. *N. Engl. J. Med.* **370**, 1–3 (2014).

[16] Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–3 (2009).

[17] Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).