RESEARCH EFFORT AND EVOLUTIONARY PROPERTIES OF GENES

by

Travis Struck

A Thesis Submitted to the Faculty of the

DEPARTMENT OF MOLECULAR AND CELLULAR BIOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016
STATEMENT BY AUTHOR

The thesis titled *Research Effort and Evolutionary Properties of Genes* prepared by *Travis Struck* has been submitted in partial fulfillment of requirements for a master’s degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: *Travis Struck*

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

__________________________________________  May 9th, 2016

Ryan Gutenkunst  
Assistant Professor of Molecular and Cellular Biology  
Date
Table of Contents

Abstract - Page 5
Research Effort and Evolutionary Properties of Genes - Page 6
List of Figures and Tables

Figure 1 - Page 8
Figure 2 - Page 10
Figure 3 - Page 12
Figure 4 - Page 13
Table 1 - Page 15
Abstract

Recent research effort (measured in number of publications) on genes is biased towards genes that have been studied heavily in the past. Some factors for why this occurs is that many of these historically studied genes are important for survival or there are more tools available that make genetic studies of them much more accessible. Studies of research effort on *Saccharomyces cerevisiae* genes characterized with genetic or protein interactions found that there is an aversion to studying lesser-known genes in networks. As well, in a study of three human protein families, many of the genes that have recently been discovered to have association with complex disease, through methods such as genome wide association studies (GWAS), are understudied in the present compared to the small number of historically heavily studied genes. In this study we explore possible causes of and diversion from this preferential bias with gene conservation and human genes being disease-associated.

We find there is some evidence of conservation driving biases in research effort for essential genes in *Saccharomyces cerevisiae*, but inconclusive evidence in other organisms. We look for effects of disease association through Mendelian and complex diseases in a historical, pre-GWAS, and contemporary, post-GWAS, context. Within both contexts we find that Mendelian disease genes may drive preferential study bias. For contemporary research effort we utilize a model of publication rates and find that there are individual GWAS genes that tend to be investigated more than predicted compared to non-GWAS genes. It appears that the proportion of GWAS genes that had highly unexpected increases in publication rate compared to model predictions rose fairly quickly but has been declining. Our analysis suggests that GWAS has had a small impact on what genes some scientists study despite preferential study biases. However GWAS gene-disease association’s impact on research effort appears to be declining, possibly due to scientists not being as interested in GWAS results as time goes on.
Research Effort and Evolutionary Properties of Genes

Travis J. Struck\textsuperscript{1} Brian K. Mannakee\textsuperscript{2} Ryan N. Gutenkunst\textsuperscript{1,*}

\textbf{1} Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, United States
\textbf{2} Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, Tucson, Arizona, United States

* rgutenk@email.arizona.edu

\textbf{Introduction}

Within biology there is a preferential study bias towards genes discovered early on [1, 2, 3], which may drive disproportionate knowledge among genes. This could lead to a situation in which highly studied genes are perceived as functionally important to an organism while others go understudied and their functional importance remains obscure. In molecular evolution it is a common belief that conserved genes are important [4, 5], despite the fact that conservation is largely uncorrelated with attempts at estimating functional importance through knockout screens [5]. Recent findings from Mannakee and Gutenkunst have suggested that there is a correlation between evolutionary conservation and genes with rate reactions that biochemical models are sensitive to perturbation of [6]. Because there is an aversion to studying less well known genes characterized to have interactions [1], it is possible that the conserved genes in the systems biology models that Mannakee and Gutenkunst found to be sensitive are sensitive due to the models being built around the well studied conserved genes in the model, because more is known about them compared to understudied genes.

It has been suggested that disease association is a large driver of genetic studies in humans [7, 8]. However, preferential study biases in molecular biology [2, 3] have possibly led to genes important for understanding human disease being understudied [9, 10]. Most studies geared towards diseases have possibly been on Mendelian diseases due to their nature of a single gene mutation leading to the disease phenotype being more easily identified. Recent advances in genomic technology have allowed for the study of more complex diseases through methods such as genome-wide association studies (GWAS). Previous studies suggest that most genes associated with complex diseases through genomic methods, such as GWAS, are not studied more after new knowledge about disease-association is discovered for these genes [2, 3]. These understudied complex disease genes could be viable targets for treating or further understanding of complex diseases [9, 10].

Here we use research effort on genes, measured in number of publications for each individual gene, to explore possible biases in the study of conserved genes and disease association. We find that while high conservation in some genes, especially essential \textit{Saccharomyces cerevisiae} genes, is correlated with research effort, this correlation is not the case in all organisms. We further explore the idea that disease association may drive research effort, expanding on previous studies looking at kinases, hormone receptors, and nuclear receptors in humans [2, 3], by looking at the research effort of the whole protein coding genome for humans. We look at historical research effort on genes before GWAS would have had influence over research effort and find Mendelian diseases genes, as defined by the Online Mendelian Inheritance in Man (OMIM) database, were preferentially studied while complex disease genes did not receive any particular attention.
We then explore contemporary research effort after GWAS would have possibly affected research effort on complex disease genes. Utilizing a model of publication rates, we find there is a small but significant increase in the number of publications when a gene is disease associated through GWAS. These findings suggest the research effort on many genes in organisms may be geared towards Mendelian disease genes rather than conservation level. Also our findings suggest that new technology has a small effect on diverting this preferential study bias. However, the effect of GWAS itself seems to be in decline, possibly due to a loss of interest.

Separately from research effort and biases in biology, we will discuss a study building on Pandya and Gutenkunst’s work on positive selection against tyrosine residue content in organisms proteomes [11]. Our study and the Pandya and Gutenkunst study were based on finding that there is a negative correlation between tyrosine content in an organism’s proteome and the number of tyrosine kinases in that organism. We found negative results in agreement with Pandya and Gutenkunst using an alternative method for detecting exposure of tyrosine residues to possibly being spuriously phosphorylated. Our method here involved a pipeline for assembling homology models of 16 metazoans with structural information in the Protein Data Bank (PDB) estimated by the DSSP database.

Results/Discussion

Research Effort and Evolutionary Conservation

Findings by Mannakee and Gutenkunst have suggested that the more conserved a gene is, the more sensitive biochemical networks are to perturbation in rate constants of these conserved genes [6]. However, there is a preferential study bias within biology [1, 2, 3], which may be geared towards conserved genes due to theoretical importance [4, 5]. As such, findings from Mannakee and Gutenkunst may be an artifact of preferential study biases for conserved genes as it is possible that theses models may be so sensitive to these conserved genes because there is a disproportionate amount of knowledge about them compared to other gene species in the models.

In order to assess if gene research effort, quantified by the number of PubMed publications on a gene, was biased towards evolutionarily conserved genes, we explored the evolution rate of genes, determined by PAML (dN/dS), for four organisms: *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and humans (Figure 1). Research effort was fairly correlated with evolution conservation in *Saccharomyces cerevisiae* (Spearman rank $\rho = -0.33$) and *Drosophila melanogaster* (Spearman rank $\rho = -0.30$). There was a weaker correlation in *Caenorhabditis elegans* (Spearman rank $\rho = -0.15$) and humans (Spearman rank $\rho = -0.25$). Particularly, we find that essential genes in *Saccharomyces cerevisiae* have even more of a correlation between research effort and evolutionary conservation (Spearman rank $\rho = -0.40$).

Because there is a bias towards preferentially studying older genes [1, 2, 3] and many of *Saccharomyces cerevisiae* essential genes were discovered earlier on [1], our findings suggest that preferential study bias of conserved genes may have a role in the study of *Saccharomyces cerevisiae*. However, due to this conclusion not being as convincing in other organisms, a question remains: what would drive the study of genes in these organisms?

Historical Research Effort and Disease Association

Organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans* are model organisms of human diseases. As such, it is possible that disease association is a large driver of the study of genes, as has been previously suggested [7, 8]. To explore this, we looked at research effort of Mendelian disease, using Online
Figure 1: Evolutionary rate analysis. A: For *Saccharomyces cerevisiae* we see there is a Spearman rank $\rho$ of -0.33 between the research effort of genes and their conservation level. This is stronger ($\rho$ of -0.4) when considering only essential genes. B: For *Drosophila melanogaster* we find a Spearman rank $\rho$ of -0.3 between research effort and gene conservation. C: For *Caenorhabditis elegans* we find a Spearman rank $\rho$ of -0.15 between research effort and gene conservation. D: For human we find a Spearman rank $\rho$ of -0.25 between research effort and gene conservation.
Mendelian Inheritance in Man (OMIM) to identify Mendelian disease-associated genes, and complex diseases that have been identified by genome-wide association studies (GWAS), based on the National Human Genome Research Institute (NHGRI) - European Bioinformatics Institute (EBI) GWAS catalog. Because it has been suggested that new technology such as GWAS has little effect on preferential study biases, to begin understanding what, if any, effect GWAS would have on research effort, we first explored historical research effort on genes before GWAS would have had an effect on research effort. To do this we looked at research effort up to 2005, the year of the first GWAS, as there would have likely not be much influence on research effort from GWAS during this year [12].

We utilize a Lorenz curve, which is common in economics for analysis of inequalities in income, and find that there is heavy skewing of historical research effort, where only a few genes receive the majority of research effort (Figure 2A). The inequality of genes can be quantified with a Gini coefficient, which we find to be 0.909. When looking at the distribution of research effort given a disease association, we find that historical research effort was skewed towards Mendelian disease genes (Figure 2B, blue). While past the lowest levels of research effort complex disease-associated genes were equally distributed (Figure 2B, red). This suggests that historical research effort has been focused on Mendelian disease genes and that scientists did largely not attend to complex disease genes. Knowing how the historical research effort is distributed would allow us to now understand if contemporary research effort has been influenced by GWAS.

**Contemporary Effects of GWAS**

To find out if there was an effect on contemporary, post-GWAS, research effort we examined the research effort on genes after 2005. We found the distribution of post-GWAS research effort (Figure 2C) to be very similar to pre-GWAS (Figure 2B), with research effort for Mendelian disease gene skewed towards higher levels and research effort on complex disease genes mostly being equal throughout the distribution.

Because previous findings had shown that some genes can overcome preferential study biases [2, 3], we utilized a model that would allow us to estimate how research effort on individual genes would have been affected in each year if GWAS had never occurred and compare that to the actual research effort for genes that were found in a GWAS. The model we used came from Pfeiffer and Hoffman (Equation 1) [7], which assumes research effort is based largely on trends. As such, the model predicts the rates of publications for a gene in a given year in an organism, based on a gene’s own publication history up to a given year and from other genes in the organism being published. We fit this model’s five parameters to all non-GWAS genes using a maximum likelihood estimator, and these five parameters would be used to predict the rate of publications for all genes in years from 1950 to 2014 in humans.

We found that our model under predicts gene-year research effort for CFH, which was the first gene in a published GWAS in 2005 [12], and that the research effort on some years following the GWAS hit are statistically significantly different from the model’s predictions (Figure 3A). Compared to a gene, PSEN1, that has yet to be associated through GWAS and possessing a similar number of publications as CFH the model fits fairly well with no years being statistically significantly different from the model’s predictions (Figure 3B). Thus this model is viable for the purposes of detecting the effect of GWAS across all GWAS genes for post-GWAS years.

To assess the immediate impact that GWAS might have on the research effort of complex disease genes, we looked at a three-year window from when the gene was a GWAS hit and the following two years. We quantified the difference between the model and our data using the average difference between the actual research effort and the predicted research effort over the three-year window. We found a slight shift in the distribution of GWAS genes compared to control genes (Figure 3C, Mann-Whitney U $p < 10^{-10}$), which contained a large range of values. We quantified this distribution by the average difference between
Figure 2: The distribution of Research Effort is highly skewed towards Mendelian disease genes in humans. A: Lorenz plot of the entire distribution of gene publications for years up to 2005. The inequality of this distribution can be quantified by a Gini coefficient of 0.909. B: We binned our genes into sets of 2323 based on similar range of research effort for years up to 2005. In each bin we looked at what percentage of the genes were found to be a Mendelian disease gene through OMIM (blue), a complex disease gene through GWAS (red), or either type of disease gene (purple) to approximate a probability of a gene being associated with a disease at a given level of research effort. There is a high concentration of Mendelian disease genes at higher levels of research effort while GWAS past the lowest level is equally distributed. C: Same as B, but research effort for years 2006 to 2014. There are not any noticeable changes in the trends.
GWAS genes and control genes in this distribution, which resulted in 0.44 more publications for GWAS genes on average over this three-year window. This suggests that GWAS genes have received a small but statistically significant boost in research effort shortly after their association to a complex disease through GWAS compared to non-GWAS genes.

Some GWAS genes have very large differences between what their actual research effort is and what the model predicts. To assess the trend of the immediate impact that GWAS has on genes over time, we only looked at all the genes and years where research effort was statistically significantly different from predicted for the gene-year, after a Bonferroni correction for statistical significance. It appears that the immediate impact of GWAS on research effort increased quickly over time, but has been steadily declining (Figure 3D). It appears that this may be due to GWAS having less of an impact on the amount of publications that a gene will receive from being a GWAS hit, as the overall number of new genes that GWAS associates to complex diseases appears to be increasing each year since 2005 (Figure 3E). Overall, while Mendelian disease genes were highly focused on both historically (Figure 2B, blue) and presently (Figure 2C, blue), it is possible for some genes to overcome this preferential study bias in human genes due to technology that can discover complex disease associations (Figure 3A and 3C). However, this effect may decline over time (Figure 3D), possibly due to a decline in interest in new genes associated with complex diseases through GWAS (Figure 3E).

Structural biology and evolution

Tan et al. has suggested that because of the strong negative correlation between the number of tyrosine kinases and the amount of tyrosine in an organism’s proteome, there is possibly positive selection against tyrosine residue content in organism proteomes to reduce the chance of spurious phosphorylation that could be deleterious [13]. Previous work in our group by Pandya and Gutenkunst using SPINE-X [14] to assess the accessibility of tyrosine residues to tyrosine kinases led to negative results regarding Tan et al.’s spurious phosphorylation hypothesis [11]. SPINE-X results estimated that there is very little difference in how accessible tyrosine would be to tyrosine kinases across 16 metazoan organisms. This opposed the spurious phosphorylation hypothesis, which would have predicted that organisms with more tyrosine kinases would have less exposed tyrosine to avoid spurious phosphorylation. Because of the aversion to negative results in science [15], we utilize a pipeline to assemble homology models for all 16 metazoans with the sequences of proteins with solved structures from the Protein Data Bank (PDB) [16] in the DSSP [17] in order to obtain an alternate source of accessibility scores for tyrosine residues. Our results are in agreement with Pandya and Gutenkunst, finding that all metazoans seem to have very similar amounts of tyrosine accessibility across their proteomes (Figure 4).

An additional test that Pandya and Gutenkunst performed on the spurious phosphorylation hypothesis was to examine the tyrosine content in ordered and disorder regions between orthologous proteins in humans and *Saccharomyces cerevisiae*. If the spurious phosphorylation hypothesis held, it would be expected that there would be a greater loss of tyrosine content in disordered regions of human proteins compared to *Saccharomyces cerevisiae* due to the exposure of disordered regions to tyrosine kinases compared to ordered. However Pandya and Gutenkunst found a fairly similar loss of tyrosine residues in human orthologs compared to *Saccharomyces cerevisiae* orthologs in both ordered and disordered regions. We performed an additional control comparing *Drosophila melanogaster* orthologs to human orthologs to test for similar results to *Saccharomyces cerevisiae*. Indeed, we found similar results where the tyrosine content loss was fairly equal in ordered and disordered regions in human orthologs compared to *Drosophila melanogaster* orthologs (data not shown).

Overall, our results were in agreement with Pandya and Gutenkunst that the spurious phosphorylation
Figure 3: GWAS effect on contemporary research. A: The publication rate data (solid line) vs. the model’s predicted publication rate (dashed line) for CFH, the first GWAS hit. The data statistically significantly differs from the model after a Bonferroni correction in 2006 and 2008. B: The publication rate data (solid line) vs. the model’s predicted publication rate (dashed line) for PSEN1, a non-GWAS gene with very similar number of publications to CFH, which never statistically significantly differs from the model. C: Probability density distribution of differences between actual number of publications and model predictions for the year each GWAS gene (red) was associated with a disease in GWAS and the following two years. For each year a GWAS gene was a GWAS hit a random non-GWAS gene with a similar total number of publications up to the year the GWAS gene was a hit was used as a control taking the difference between the control gene’s actual publications and predicted publication. The distribution of GWAS genes is slightly shifted to the right of the control distribution. D: The distribution of the proportion of statistically significantly different GWAS genes within the three-year window of being a GWAS hit. This distribution increases quickly but is also declining quickly. E: Distribution of new GWAS hits over time between 2005 and 2014. There is a steady increase in the number of GWAS genes being discovered.
hypothesis was not the correct hypothesis for the decrease in tyrosine residues. It is possible that the loss of tyrosine was due to the increase in GC content [18]. Another possible explanation is that loss of tyrosine in proteins may be due to the evolution of metabolic pathways utilizing tyrosine for neurotransmitter and hormone synthesis made less tyrosine being taken up to synthesis proteins more favorable [19].

Materials and Methods

Evolutionary rate data and analysis

Saccharomyces cerevisiae evolution rate data was retrieved from Wall et al.’s ”Functional genomic analysis of the rates of protein evolution”, supplementary table 4 [20].

Drosophila melanogaster evolution rate data was retrieved from FlyBase’s PAML summary data from 12 species analysis from Clark Eisen et al.’s genome analysis [21].

Caenorhabditis elegans evolution data was retrieved from Castillo-Davis et al.’s ”The Functional Genomic Distribution of Protein Divergence in Two Animal Phyla: Coevolution, Genomic Conflict, and Constraint” supplementary file 2 [22].

Human evolution data was retrieved from Lindblad-Toh et al.’s ”A high-resolution map of human evolutionary constraint using 29 mammals” supplementary info [23].

Publication data

based off publications. Additional publications for genes on NCBI gene may be drawn from organism specific databases, Gene Ontology, and available curated systems biology data. Publication information was pulled from PubMed by BioPython [25].

GWAS & OMIM data

GWAS data was retrieved November 25th, 2015 from the NHGRI-EBI’s Catalog of Published Genome-Wide Association Studies [26]. We filtered for genes that were considered disease genes according to the Experimental Functional Ontology [27]. The GWAS catalog contains GWAS hits based on what the authors of the GWAS publication reports and based on a pipeline that maps the reported SNPs to gene through NCBI Gene and NCBI dbSNP. We pooled the genes identified by the authors and the NHGRI-EBI into one set. If the SNP was within 500 bases of a gene we considered that gene to be a GWAS hit.

OMIM data was retrieved July 8th, 2015 from OMIM FTP site [28]. We filtered out genes that were associated with complex disease, were not diseases, and did not have more than one confirmed source of mapping the disease to a gene. When distinguishing any crossover between OMIM and GWAS, we went with whichever had the earliest marked year for the gene entry. In cases where an OMIM gene had an ID below 600000, it was marked as being from 1994, as it was from before the online version of OMIM. In cases where years matched between GWAS and OMIM, the gene was considered a GWAS gene since our findings suggest that most Mendelian disease genes were discovered early. More info for these filters can be found on the OMIM FAQ.

This resulted in 4917 GWAS genes and 2013 OMIM genes in our datasets.

Publication Rate Model

The model of publication rates comes from Pfeiffer and Hoffman (2007) [7].

\[
\Delta P_{i,t+1} = \frac{k_1P^*_t + k_2P_{i,t} + k_3}{1 + (P^*_t / P_S)^\alpha} \tag{1}
\]

The publication rate model estimates the mean number of publications assuming they are Poisson distributed \(\Delta P_{i,t+1}\), which is the expected number of publications for gene \(i\) in year \(t+1\) based on all publications for gene \(i\) up to and including to year \(t\) \((P_{i,t})\) and the average number of publications for all genes in an organism up to and including year \(t\) \((P^*_t)\). The parameters in the numerator describe the exponential growth of publications. \(k_1\), the rate of publication on gene \(i\) due to publications on all other genes up to year \(t\), \(k_2\), the rate at which publications on gene \(i\) occur due to prior publications on gene \(i\), and \(k_3\), the rate of spontaneous publications to describe the initial publication on gene \(i\). The parameters in the denominator, \(P_S\) and \(\alpha\), are saturation terms to represent a point in time publications will start to saturate and how quickly that saturation occurs. It is assumed that the actual number of publications will follow a Poisson distribution giving the function \(f(\lambda,n) = e^{-\lambda}\lambda^n/n!\). We fit the parameters to publication data from 1949 to 2014 for all genes never associated to disease through GWAS using a maximum likelihood estimator with SciPy [29] on the product of the function \(f(\lambda = \Delta P_{i,t+1}; n = \text{actual number of publications for gene }i\text{ in year }t+1)\) for each gene \(i\) and each year \(t\). Values for our parameters can be found in Table 1.

To test for gene-years where the number of observed publications was statistically significant from the models predicted number of publications was determined by calculating the probability of getting \(k\) or more publications given the expected number of publications \(\lambda\) in a Poisson distribution; where \(k\) is the number of observed publications for gene \(i\) on year \(t\) and \(\lambda\) is the predicted number of publications for gene \(i\) on year \(t\). A Bonferroni correction was used to set our cutoff at 0.05/\((N_gN_y)\), where \(N_g = 20907\) genes and
Table 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>0.036</td>
</tr>
<tr>
<td>$k_2$</td>
<td>0.22</td>
</tr>
<tr>
<td>$k_3$</td>
<td>0.002</td>
</tr>
<tr>
<td>$P_S$</td>
<td>13.126</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.619</td>
</tr>
</tbody>
</table>

$N_y = 65$ years. Here our statistical significance cutoff is $-3.67 \times 10^{-8}$ after the Bonferroni correction. This resulted in 1129 gene-years having a statistically significant number of publications among 550 genes.

**DSSP Homology Model**

The proteomes of the 16 species used for analysis of the spurious phosphorylation hypothesis were downloaded from Ensembl release 72 [30]. In cases of multiple splice forms of a gene, the longest protein was chosen. Orthologs in the DSSP database [17, 31] were retrieved using NCBI BLAST 2.2.26 [32]. Matching tyrosine sites was done by aligning orthologs with MUSCLE 3.8.31 [33]. If 40% of bases in orthologs matched when aligned then they were used in our analysis. DSSP accessibility score was assigned only to matching residues.

**References**


30. Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García

