Comment

Supplementary information to:

A wealth of discovery built on the Human Genome Project — by the numbers

A Comment published in *Nature* **590**, 212–215 (2021)
https://doi.org/10.1038/d41586-021-00314-6

Alexander J. Gates, Deisy Morselli Gysi, Manolis Kellis & Albert-László Barabási

SUPPLEMENTARY INFORMATION

A WEALTH OF DISCOVERY BUILT ON THE HUMAN GENOME PROJECT — BY THE NUMBERS

Alexander J. Gates[1,*], Deisy Morselli Gysi[1,3,*], Manolis Kellis[4,5], and Albert-László Barabási[1,2,3,5,†]

[1]Network Science Institute, Northeastern University, Boston, Massachusetts 02115, USA
[2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA USA
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
[5]Department of Network and Data Science, Central European University, Budapest 1051, Hungary
[†] To whom correspondence should be addressed: a.barabasi@northeastern.edu
[*] Authors contributed equally

ACKNOWLEDGEMENTS

A.J.G. and A.-L.B. were supported in part by the Templeton Foundation, contract #61066, and the Air Force Office of Scientific Research under award number FA9550-19-1-0354. A.-L.B. is partly supported by European Union's Horizon 2020 research and innovation program under grant agreement No. 810115 DYNASNET and by NIH grant 1P01HL132825. We also express our gratitude to Alice Grishchenko and Csaba Both for help in designing the figures.

TABLE OF CONTENTS

S1 Data and Data Sources

Our genomic reference is comprised by the 38,546 transcripts annotated by GENCODE v35 (GRCh38.p13)[1], 1 million SNPs present in the GWAS catalogue[2], and the ENCODE project list of promoters and enhancers[3], 7712 drugs retrieved from DrugBank[4], 1660 diseases collected from CTD[5], OMIM[6], DisGeNet[7,8], Orphanet, ClinGen[9], ClinVar[10], GWAS catalogue[2], PheGenI[11], lncRNADisease[12,13] and HMDD[14]. We scraped the PubMed ids for all references linking a transcript to a disease (GWAS catalogue & OMIM), or on the National Institute of Bioinformatics (NCBI), uncovering a total of 2,386,046 genome-publication relationships (Table S 1).

Type	Total Number of Elements	Unreferenced Elements	Publication References	Unique Publications
Enhancer	809,429	0	1	1
SNP	125,128	0	157,975	4,147
Promoter	58,565	0	1	1
protein coding	19,757	608	1,534,274	663,835
pseudogene	9,127	8,491	3,465	2,038
regulatory RNAs	6264	1,998	49,210	29,949
post-transcriptional modification RNAs	2263	1,846	1675	298
misc_RNA	1033	1,007	87	47
Other	62	14	218	174
rRNA	40	21	132	52

Table S 1 Genomic Elements. The number of genomic elements by type and the number of publications referencing those elements.

The publication meta-data was gathered for the resulting 704,515 PubMed publications, giving the year of publication. Additionally, the publications were linked to the Microsoft Academic Graph (MAG) to gather author information and field of study (level 0 and 1 fields). We see in Table S 2 that most protein-coding genes had their first publication before the first draft of the HGP in 2001, while the vast majority of the other genomic elements started drawing attention from the scientific community after that.

Type	Before 2001	After 2001
protein coding	15,068	4,156
regulatory RNAs	445	3,821
post-transcriptional modification RNAs	187	230
pseudogene	184	452
Other	15	33
misc_RNA	9	17
rRNA	5	14
Enhancer	0	809,429*
Promoter	0	58,565
SNP	0	125,128

Table S 2The number of genomic elements discovered before and after the HGP.

*NOTE: enhancers all come from a 2020 publication and therefore are not included in the trend graphs.

We further linked the identified gene transcripts to 1,600 diseases with documented genetic roots, collected from CTD[5], OMIM[6], DisGeNet[7,8], Orphanet, ClinGen[9], ClinVar[10], GWAS catalogue[2], PheGenI[11], lncRNADisease[12,13] and HMDD[14]. Using data from the Genome Wide Association Studies (GWAS) catalogue, we mapped each SNP to the GENCODE human assembly and identified where each of the SNPs are located in the genome. We find that more than 90% of the SNPs associated with traits are located in protein-coding regions, enhancers or lncRNAs (Table S 3).

Drugs can have multiple targets, with different pharmacological functions. DrugBank classifies targets into four main categories: Polypeptides, Enzymes, Carries and Transporters (Figure S 1). Polypeptides are mostly related with disease modifications, while the other categories are related to the metabolism of the drug. Therefore, for the results discussed in the main paper we present only approved drugs and their polypeptide targets, unless stated otherwise.

Genomic Element	SNP count	%
Protein-coding	69,886	71.12
Enhancer	15,911	16.20
lncRNA	9,633	9.80
Promoter	1,497	1.53
Pseudogene	1,302	1.32
miRNA	13	0.01
snRNA	13	0.01
Other RNAs	12	0.01

Table S 3 Where do SNPs with associated traits land? The number of SNPs associated with traits, and the genomic elements by type where they are located. Protein-coding genes, enhancers and lncRNAs are the ``mutation hot-spots'' associated to traits.

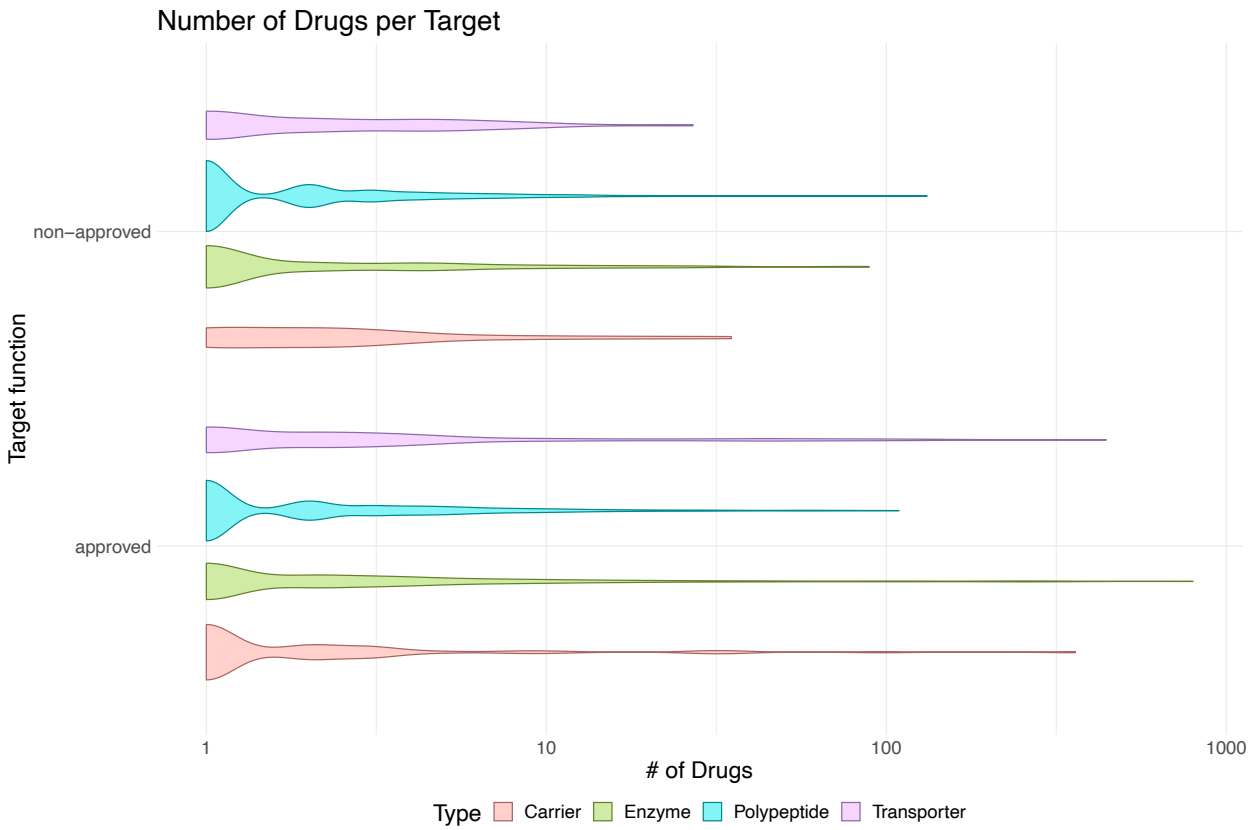

Figure S 1 Genes are targets from multiple drugs. Most of the genes that are targeted by pharmacological drugs are targeted by multiple drugs, independent of their role in the drug delivery. Genes associated to approved drugs tend to be associated to more drugs than non-approved drugs, showing that there is a bias towards known and approved targets.

Type	approved	non-approved
Polypeptide	2149	1905
Enzyme	368	93
Carrier	78	16
Transporter	252	54
Total Targets	**2467**	**1988**
Total Drugs	**2208**	**3894**

Table S 4 Targets for approved and non-approved drugs.

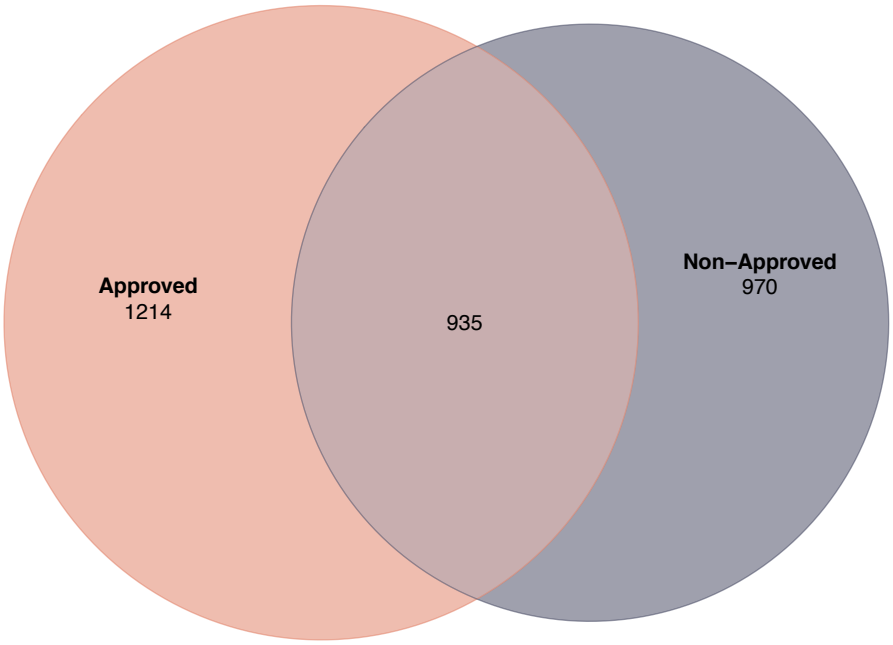

Figure S 2 Targets from approved and non-approved drugs have a small overlap. Showing that we are still exploring new target options for drugs.

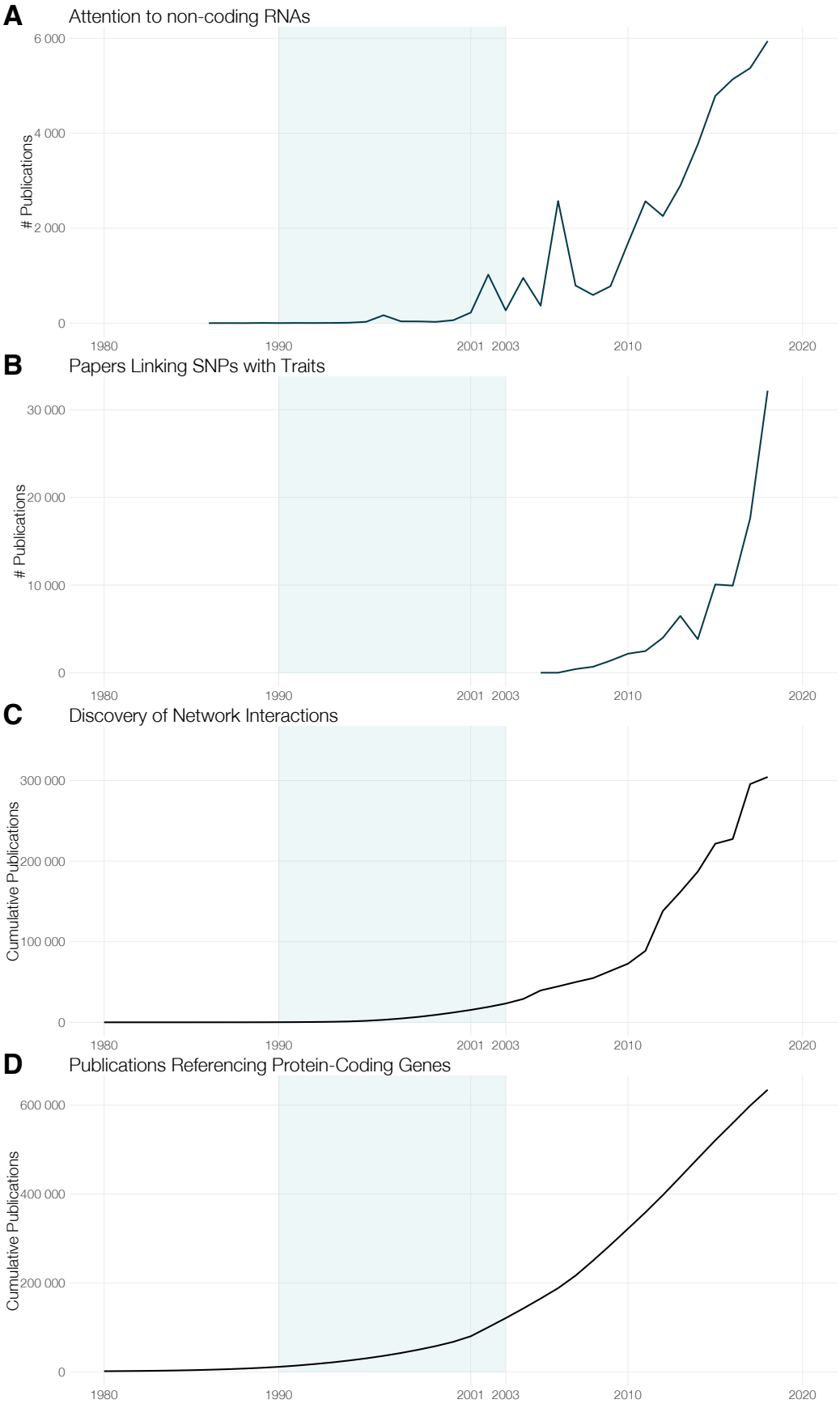
A
Attention to non-coding RNAs
Publications
6 000
4 000
2 000
0
1980
1990
2001
2003
2010
2020
B
Papers Linking SNPs with Traits
Publications
30 000
20 000
10 000
0
1980
1990
2001
2003
2010
2020
C
Discovery of Network Interactions
Cumulative Publications
300 000
200 000
100 000
0
1980
1990
2001
2003
2010
2020
D
Publications Referencing Protein-Coding Genes
Cumulative Publications
600 000
400 000
200 000
0
1980
1990
2001
2003
2010
2020

Figure S 3 (A) The number of publications each year that reference a regulatory non-coding RNA (an important subset of the elements shown in Fig Discovery of genomic elements. (B) The number of publications each year linking SNPs to traits in the GWAS catalogue. (C) The cumulative number of known protein-protein and protein-genetic interactions whose growth reflects the rise of network genomics. (D) The cumulative number of publications for all protein-coding genes.

S2 Attention Inequality

As shown in Figure S 4, the distribution of the total number of publications per gene is heavy-tailed. To measure the inequality in publications per gene represented by the distributions in Figure S 4, we calculate the Gini coefficient. The Gini coefficient is a real number between 0 and 1, whose value is 0 if all genes are mentioned in exactly the same number of publications, and 1 denotes maximum inequality among the genes, where all attention is devoted to a single gene. As we can see in Figure S 5, the Gini coefficient has steadily increased since 1960, reflecting a growing inequality in the number of publications per gene.

As shown in Figure S 6, the change in the number of publications per gene (dk) grows linearly with the accumulated number of publications per gene (k) in the log-log plot. See Barabasi 2016, Network Science for details.

The distributions of scientific interest as measured by drug targets (Figure S 7A) and disease associations (Figure S 7B) is similarly heavy-tailed. For example, the gene *ADRA1A* is targeted by 103 drugs while 1,294 genes are targeted by only one drug, and the gene *APOE* is associated with 27 diseases, while 2,471 genes are associated with only 1 disease. Similarly, the number of genes targeted by each drug (Figure S 7C) and the number of genes associated with each disease (Figure S 7D) are heavy-tailed.

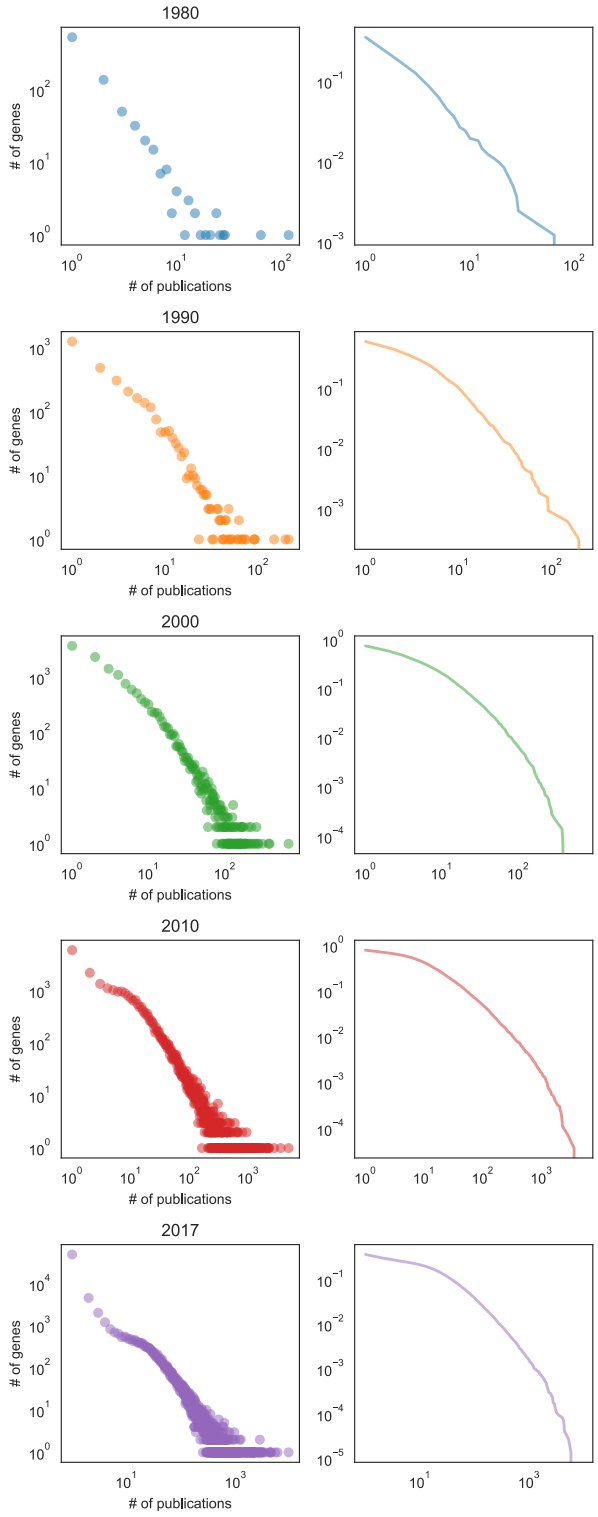

Figure S 4 Growing inequality in publications per gene. The most studied gene was the focus of 161 publications in 1990, while today the most explored gene has been referenced by almost 10,000 publications each year. (left) pdf, (right) cdf for each decade 1980-2020.

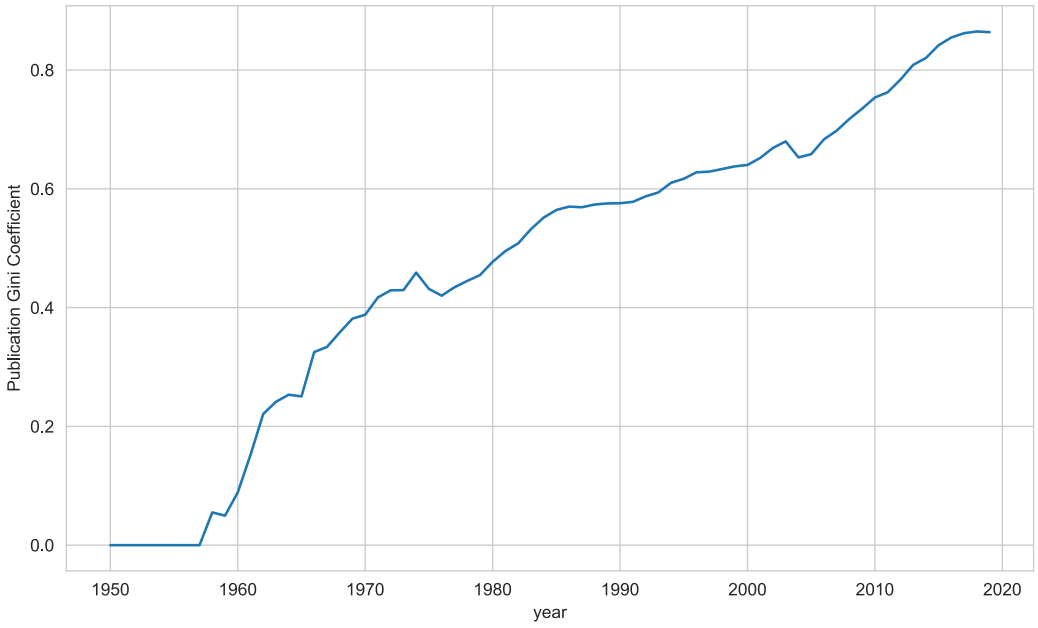

Figure S 5 ***Growing inequality in publications per gene. The Gini coefficient measuring inequality in the*** *distribution of publications per gene.*

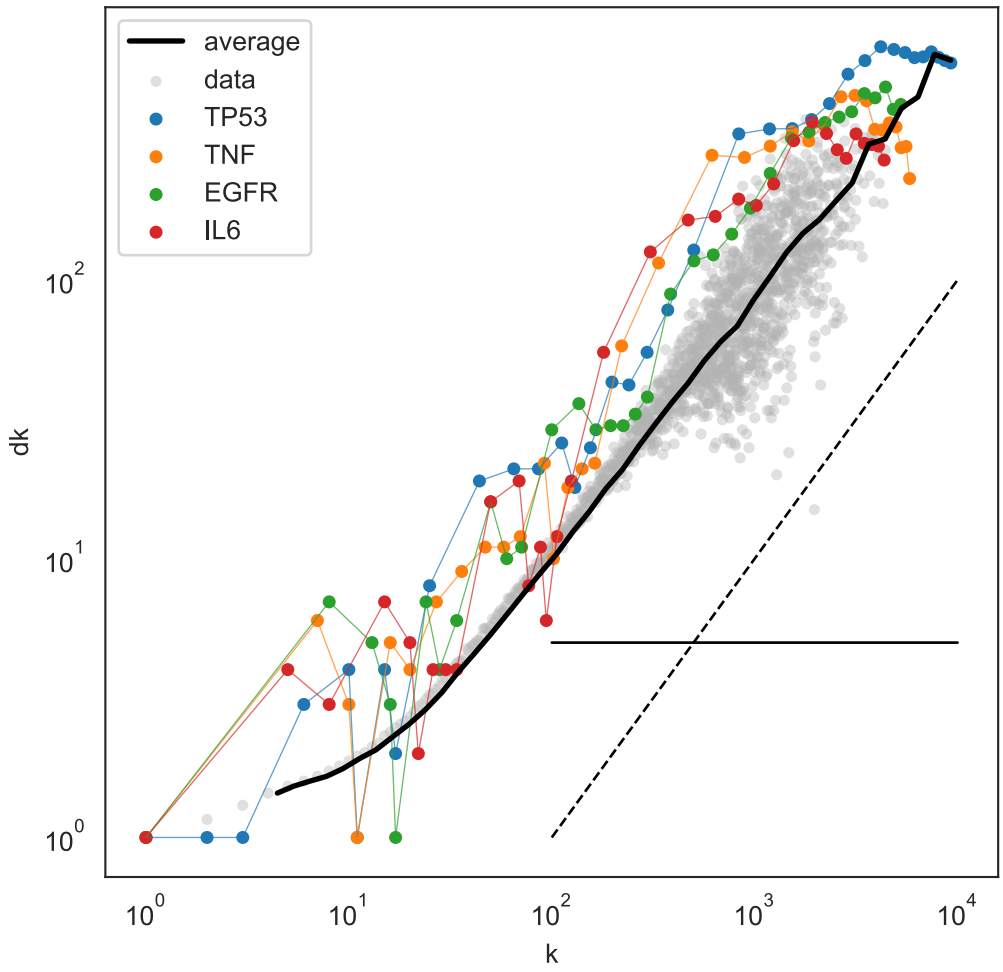

Figure S 6 Preferential attachment (PA) of attention to genes. The change in the number of publications per gene (dk) given the existing number of publications per gene (k) for all genes in the dataset (grey). Also shown is the preferential growth for the top 5 most published genes. The average increase in the number of publications per gene follows a linear trend, reflecting the presence of preferential attachment [15]. The plot shows the PA function, where the observed k dependence (dashed line, guide to the eye) is evidence of preferential attachment, while a constant k dependence (solid line, guide to the eye) would imply the lack of it.

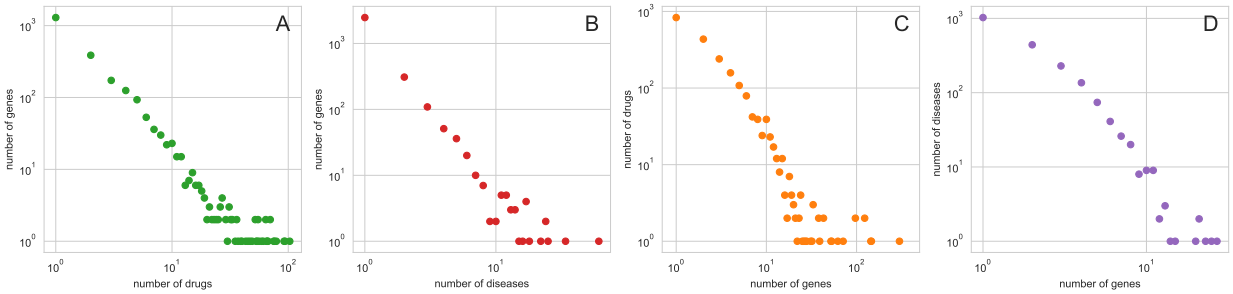

Figure S 7 Inequality in scientific interest from drugs and diseases. A) The distribution of the number of drugs targeting each gene. B) The distribution of the number of diseases associated with each gene. C) The distribution of the number of genes targeted by each drug. D) The distribution of the number of genes associated with each disease.

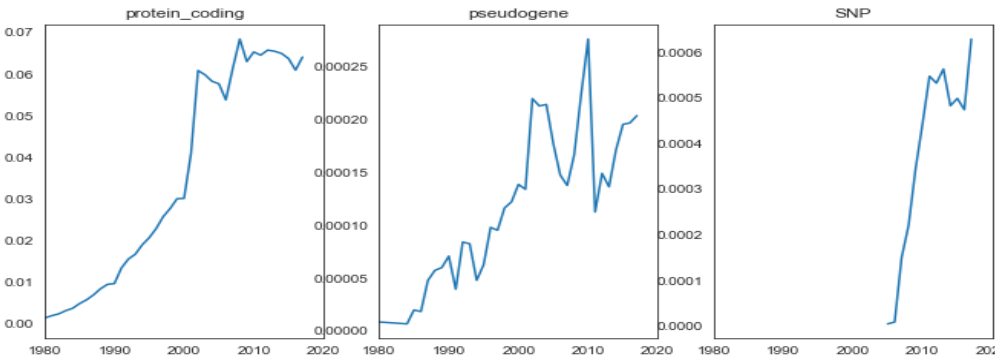

Figure S 8. The normalized yearly publications referencing (a) protein coding transcripts, (b) pseudogenes, and (c) SNPs. The number of publications in biology is used as the denominator.

S3 Largest Collaborations

The publication team size is defined as the number of authors linked to the publication on the Microsoft Academic Graph (MAG). In most cases, the MAG lists all authors in a consortium, although a few exceptions were identified in which the consortium name appeared as the sole author. To define the set of all publications, we limited the analysis to only document type "journal article" with at least 20 citations. We further defined the biology publications using the level 0 field "Biology" in the MAG.

References

1. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

2. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky1120

3. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* (2020). doi:10.1038/s41586-020-2493-4

4. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research* (2018). doi:10.1093/nar/gkx1037

5. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkaa891

6. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics* (2007). doi:10.1086/514346

7. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkz1021

8. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833–D839 (2017).

9. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *New England Journal of Medicine* (2015). doi:10.1056/nejmsr1406261

10. Landrum, M. J. *et al.* ClinVar: Improvements to accessing data. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkz972

11. Ramos, E. M. *et al.* Phenotype-genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *European Journal of Human Genetics* **22**, 144–147 (2014).

12. Chen, G. *et al.* LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Research* (2013). doi:10.1093/nar/gks1099

13. Bao, Z. *et al.* LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky905

14. Huang, Z. *et al.* HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Research* (2019). doi:10.1093/nar/gky1010

15. Barabási, A.-L. *Network Science*. (Cambridge University Press, 2016).