Comment

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

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

A new analysis traces the story of the draft genome's impact on genomics since 2001, linking its effects on publications, drug approvals and understanding of disease.

The 20th anniversary of the publication of the first draft of the human genome[1,2] offers an opportunity to track how the project has empowered research into the genetic roots of human disease, changed drug discovery and helped to revise the idea of the gene itself.

Here we distil these impacts and trends. We combined several data sets to quantify the different types of genetic element that have been discovered and that generated publications, and how the pattern of discovery and publishing has changed over the years. Our analysis linked together data including 38,546 RNA transcripts; around 1 million single nucleotide polymorphisms (SNPs); 1,660 human diseases with documented genetic roots; 7,712 approved and experimental pharmaceuticals; and 704,515 scientific publications between 1900 and 2017 (see Supplementary information; SI).

The results highlight how the Human Genome Project (HGP), with its comprehensive list of protein-coding genes, spurred a new era of elucidating the function of the non-coding portion of the genome and paved the way for therapeutic developments. Crucially, the results track the emergence of a systems-level view of biology alongside the conventional single-gene perspective, as researchers mapped the interactions between cellular building blocks (see 'No jump for big teams').

There are limitations to our analysis. For example, there is no consensus on where a gene starts and ends or, surprisingly, even what sequence exactly encodes some genes[3]. Multiple naming conventions are in use for some genomic elements, so sometimes our methodology did not connect them. And other links between publications and elements might not have been added to databases by authors. Finally, our graphs end in 2017, because there can be a time lag between an article's publication and entry into the databases we used.

However, we do not expect these issues to affect the trends we note in how genome research has changed over time. The trends remain when we control for the growth in biology publications over the same period (see SI, Fig. S6). We did not control for time since the discovery of genes, but estimate that doing so would not have altered our conclusions.

These connections offer a snapshot of the evolution of the research landscape before and after the HGP. It shows an intense focus on a small number of 'superstar' protein-coding genes, potentially to the detriment of interesting work that could be done on others. There has been a pivot towards non-protein-coding sections of the genome, and to understanding interactions between genetic material and proteins. And drug discovery has been grounded in just a few protein targets.

> **"By 2017, 22% of gene-related publications referenced just 1% of genes."**

Some of these trends are familiar to biologists, but to quantify and visualize them is to consider them anew.

There is no world without an HGP for comparison. So it is impossible to say whether these trends would have arisen anyway. Other factors, from increased computing power to sophisticated sequencing methods, also had a role in many of these developments. It is nonetheless clear that the HGP's catalogue catalysed the continuing genetic revolution.

Superstar genes

The popular perception is that the HGP marked the start of the intensive search for protein-coding genes. In fact, the 2001 draft HGP paper signalled the end of a decades-long hunt[1,2]. Indeed, evidence for the first protein-coding gene emerged in 1902, with the discovery of the hormone secretin[4] (*SCT* gene), 50 years before the structure of DNA was uncovered, and 75 years before genome sequencing became commonplace. Our analysis shows that, between the start of the HGP in 1990 and its completion in 2003 (after the draft was published in 2001), the number of discovered (or 'annotated') human genes grew drastically. It levelled out suddenly in the mid-2000s at about 20,000 protein-coding genes (see 'Twenty years of junk, stars and drugs: Non-coding elements'), far short of the 100,000-strong estimate previously adopted by many in the scientific community[2].

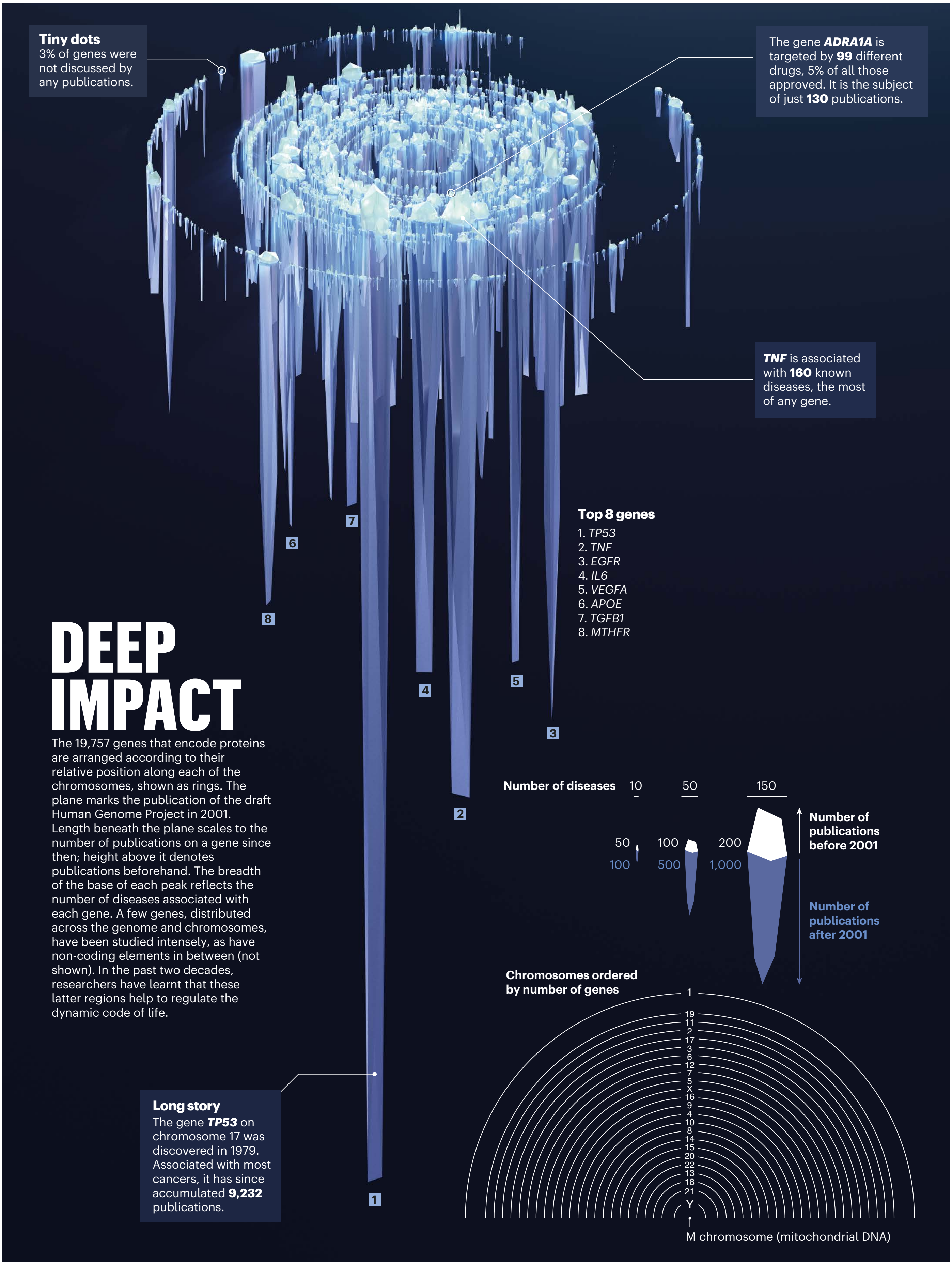
Although discoveries of protein-coding genes reached a plateau, interest in individual genes grew rapidly following the HGP. Each year since 2001, between 10,000 and 20,000 papers mentioning protein-coding genes have been published (see SI; Fig. S3).

However, that interest has focused largely on just a few genes. Before 1990, *HBA1* was the most studied because it encodes one of the proteins in adult haemoglobin. From 1990, attention then shifted to *CD4* (based on the cumulative number of publications) given the protein's involvement in T-cell immunity and as the cell receptor for HIV. Yet the interest in these two genes pales next to the explosion of attention on individual genes following the draft 2001 HGP sequence. Some superstar genes, including *TP53*, *TNF* and *EGFR*, became the subject of hundreds of publications a year, with most other genes receiving scant attention (see 'Deep impact' and 'Twenty years of junk, stars and drugs: Star genes'). We find that, by 2017, 22% of gene-related publications referenced just 1% of genes.

Intense study is, of course, justified for genes that have profound biological importance. A good example is *TP53* — it is crucial to cell growth and death, and leads to cancer when inactivated or altered. Variations in this

DEEP IMPACT

The 19,757 genes that encode proteins are arranged according to their relative position along each of the chromosomes, shown as rings. The plane marks the publication of the draft Human Genome Project in 2001. Length beneath the plane scales to the number of publications on a gene since then; height above it denotes publications beforehand. The breadth of the base of each peak reflects the number of diseases associated with each gene. A few genes, distributed across the genome and chromosomes, have been studied intensely, as have non-coding elements in between (not shown). In the past two decades, researchers have learnt that these latter regions help to regulate the dynamic code of life.



Tiny dots
3% of genes were not discussed by any publications.

The gene ***ADRA1A*** is targeted by **99** different drugs, 5% of all those approved. It is the subject of just **130** publications.

TNF is associated with **160** known diseases, the most of any gene.

Top 8 genes
1. *TP53*
2. *TNF*
3. *EGFR*
4. *IL6*
5. *VEGFA*
6. *APOE*
7. *TGFB1*
8. *MTHFR*

Long story
The gene ***TP53*** on chromosome 17 was discovered in 1979. Associated with most cancers, it has since accumulated **9,232** publications.

1
2
3
4
5
6
7
8
Number of diseases
10
50
150
50
100
200
100
500
1,000
Number of publications before 2001
Number of publications after 2001
Chromosomes ordered by number of genes
1
19
11
2
17
3
6
12
7
5
X
16
9
4
10
8
14
15
20
22
13
18
21
Y
M chromosome (mitochondrial DNA)

SOURCE: BARABASI LAB

gene are found in more than 50% of tumour sequences. It is mentioned in 9,232 publications between 1976 and 2017 (see SI, Fig. S4).

One might assume that the more that is known about the same genes, the greater the incentive would be to explore the rest of the genome. Instead, the opposite happened during the past two decades: more attention was lavished on a select few. Despite this being flagged as a potential problem on the tenth anniversary[5] of the draft genome's publication, there has been no course correction.

Our previous work on other, very different systems from human social networks to the World Wide Web indicates that this vast imbalance can be explained by a 'rich-gets-richer' dynamic[6,7] rooted in social factors. It is likely that as the number of papers focusing on *TP53* increases, the easier it is to secure funding, mentorship, tools and citations for further *TP53* work because it is a safe bet (see SI; Fig. S4). In network science, this phenomenon is called preferential attachment[7]. Indeed, we find that the number of new yearly publications focusing on a given gene is linearly proportional to the size of previous literature on it (see SI, Fig. S6).

A challenge now for biology is to disentangle the motivations for what gets studied next (see page 209). Are researchers putting money, time and effort into what is most important or urgent, or into more of the same because that will reliably win grants and plaudits?

Not junk

A great debate pre-dated the start of the HGP: was it worth mapping the vast non-coding regions of genome that were called junk DNA, or the dark matter of the genome? Thanks in large part to the HGP, it is now appreciated that the majority of functional sequences in the human genome do not encode proteins. Rather, elements such as long non-coding RNAs, promoters, enhancers and countless gene-regulatory motifs work together to bring the genome to life. Variation in these regions does not alter proteins, but it can perturb the networks governing protein expression.

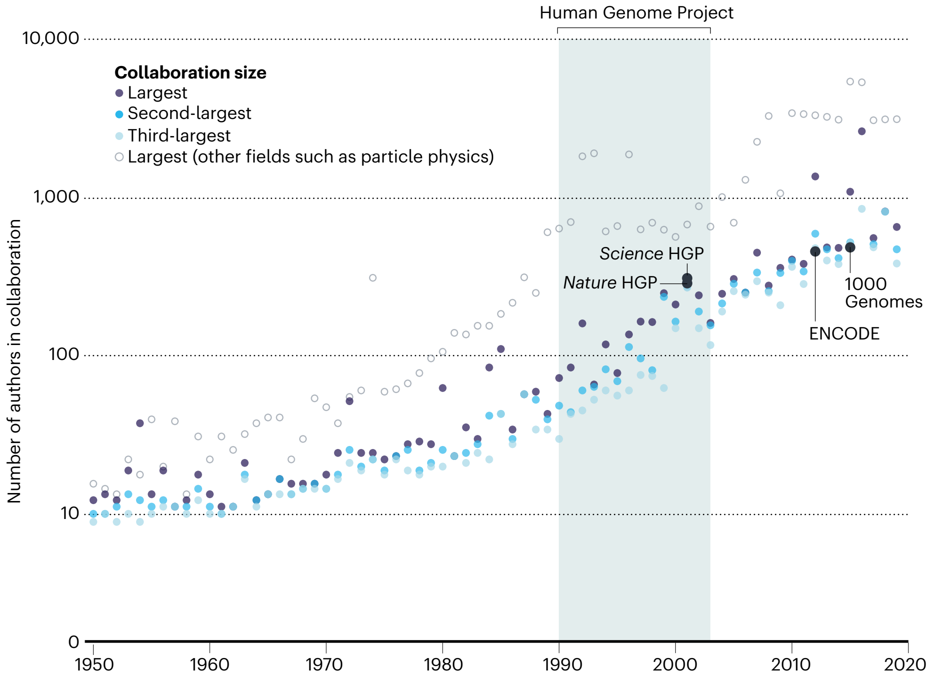
> **"The discovery of non-protein-coding elements exploded."**

With the HGP draft in hand, the discovery of non-protein-coding elements exploded. So far, that growth has outstripped the discovery of protein-coding genes by a factor of five, and shows no signs of slowing. Likewise, the number of publications about such elements also grew in the period covered by our data set (1900 to 2017; see SI, Fig. S3a). For example, there are thousands of papers on non-coding RNAs, which regulate gene expression.

The HGP also offered a way to catalogue human genetic variation, including that of SNPs. Other big efforts slashed the cost of profiling common differences across thousands of individuals; these included the International HapMap Project[8] (the third and final phase of which was completed in 2010) and the 1000 Genomes Project[9] (completed in 2015). These data sets, combined with advances in statistical analysis, ushered in genome-wide association studies (GWAS) of countless traits, including height[10], obesity[11] and susceptibility to complex diseases such as schizophrenia[12].

NO JUMP FOR BIG TEAMS

There is a common perception that the number of authors collaborating on papers about the Human Genome Project (HGP) marked a step change. In fact, team sizes in biology have grown consistently since the 1950s.



SOURCE: BARABASI LAB

There are now more than 30,000 papers per year linking SNPs and traits. A large fraction of these associations are in the once-dismissed non-coding regions (see SI, Table S3).

Cellular function relies on weak and strong links between genetic material and proteins. Mapping out this network now complements the Mendelian perspective (see page 218). Today, more than 300,000 regulatory network interactions have been charted — proteins binding with non-coding regions or with other proteins.

Drug discovery

Before about the 1980s, drugs were found largely by serendipity. Their molecular and protein targets were usually unknown. Until 2001, the probability of knowing all of a drug's protein targets was less than 50% in any given year. The HGP changed this. Now, the targets are known for almost all drugs licensed in the United States each year (see 'Twenty years of junk, stars and drugs: Drug targets').

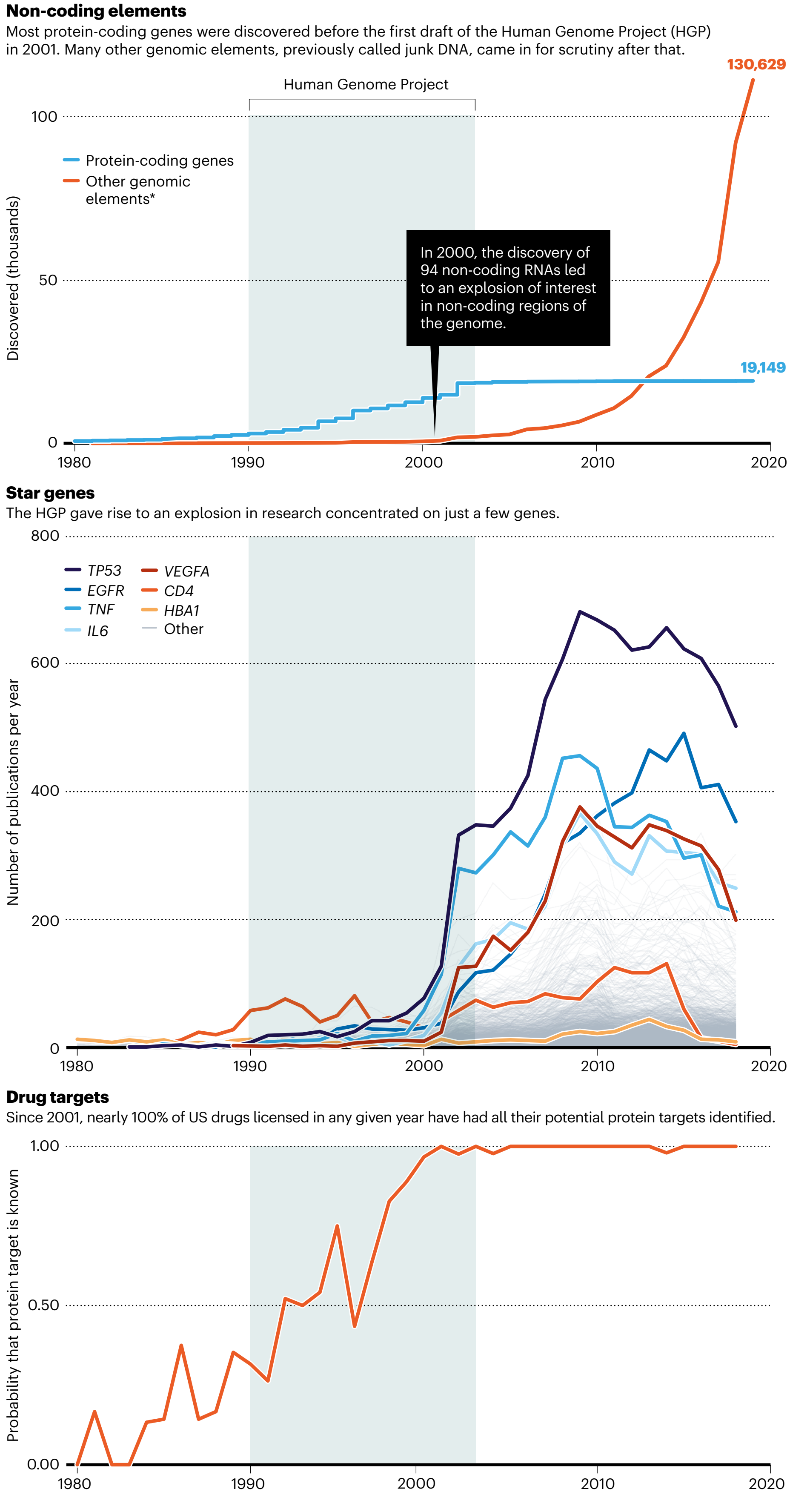
Of the roughly 20,000 proteins revealed by the HGP as potential drug targets, we show that only about 10% — 2,149 — have so far been targeted by approved drugs (see SI, Table S4 and Fig. S1). That leaves 90% of the proteome untouched by pharmacology[13]. Experimental drugs in our data set increase this number to 3,119 (SI, Fig. S2). Again, the attention given to these is highly uneven. Five per cent of all approved drugs currently approved (99 distinct molecules) target the protein ADRA1A, which is involved in cell growth and proliferation.

As previously, there could be good reasons for this skew. Some proteins might be more important to human health or more likely to act as drug targets. Some might not be druggable. Or it could be that there are many more proteins worth exploring as drug targets if only researchers, funders and publishers were less risk-averse.

That said, the majority of successful drugs do not directly target individual disease genes[14]. Instead, they target proteins one or two interactions away, modulating the consequences of faulty components. For example, large-scale screens of existing drugs that could be repurposed for use against COVID-19 found that only 1% of promising candidates targeted a viral protein — the majority were drugs that modulated human proteins not directly involved in SARS-CoV-2 activity[15]. Such network drugs hold huge potential.

TWENTY YEARS OF JUNK, STARS AND DRUGS

What genomics researchers have studied, when and why — as traced by bibliometric analysis.



Non-coding elements

Most protein-coding genes were discovered before the first draft of the Human Genome Project (HGP) in 2001. Many other genomic elements, previously called junk DNA, came in for scrutiny after that.

In 2000, the discovery of 94 non-coding RNAs led to an explosion of interest in non-coding regions of the genome.

Star genes

The HGP gave rise to an explosion in research concentrated on just a few genes.

Drug targets

Since 2001, nearly 100% of US drugs licensed in any given year have had all their potential protein targets identified.

*Including single nucleotide polymorphisms, pseudogenes, non-coding RNAs, promoters and so on.

SOURCE: BARABASI LAB

Network glimpsed

In summary, we think that the HGP is more notable for the new era of genomics it ushered in, than for the protein catalogue itself. As the theory of complex systems shows, an accurate survey of components is necessary – but not sufficient – to understand any system. Complexity arises from the diversity of the interactions between components. After 20 years of research building on the HGP, biologists now have a glimpse of the network structure and dynamics that define life.

The authors

Alexander J. Gates is an associate research scientist at the Network Science Institute, Northeastern University, Boston, Massachusetts, USA. **Deisy Morselli Gysi** is a postdoctoral research associate at the Network Science Institute, Northeastern University, Boston, Massachusetts, USA, and research trainee at the Department of Medicine, Brigham and Women's Hospital, Boston. **Manolis Kellis** is professor of computer science and principal investigator of the Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, and member of the Broad Institute of MIT and Harvard, Cambridge. **Albert-László Barabási** is professor of network science and a distinguished university professor at the Network Science Institute, Northeastern University, Boston, Massachusetts, USA; lecturer at the Department of Medicine, Brigham and Women's Hospital, Boston, and visiting professor at the Department of Network and Data Science, Central European University, Budapest, Hungary.

A.J.G. and D.M.G. contributed equally to this article. Supplementary information is available at go.nature.com/39qgndf.

e-mail: a.barabasi@northeastern.edu

1. Venter, J. C. *et al. Science* **291**, 1304–1351 (2001).
2. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
3. Portin, P. & Wilkins, A. *Genetics* **205**, 1353–1364 (2017).
4. Bayliss, W. M. & Starling, E. H. *J. Physiol.* **28**, 325–353 (1902).
5. Edwards, A. M. *et al. Nature* **470**, 163–165 (2011).
6. Bianconi, G. & Barabási, A.-L. *Europhys. Lett.* **54**, 436 (2001).
7. Barabási, A.-L. & Albert, R. *Science* **286**, 509–512 (1999).
8. The International HapMap Consortium. *Nature* **426**, 789–796 (2003).
9. The 1000 Genomes Project Consortium. *Nature* **526**, 68–74 (2015).
10. Lango Allen, H. *et al. Nature* **467**, 832–838 (2010).
11. Speliotes, E. K. *et al. Nature Genet.* **42**, 937–948 (2010).
12. Lencz, T. *et al. Mol. Psychiatry* **12**, 572–580 (2007).
13. Wishart, D. S. *et al. Nucleic Acids Res.* **46**, D1074–D1082 (2018).
14. Yildirim, M. A., Goh, K. Il, Cusick, M. E., Barabási, A. L. & Vidal, M. *Nature Biotechnol.* **25**, 1119–1126 (2007).
15. Gysi, D. M. *et al.* Preprint at https://arxiv.org/abs/2004.07229 (2020).