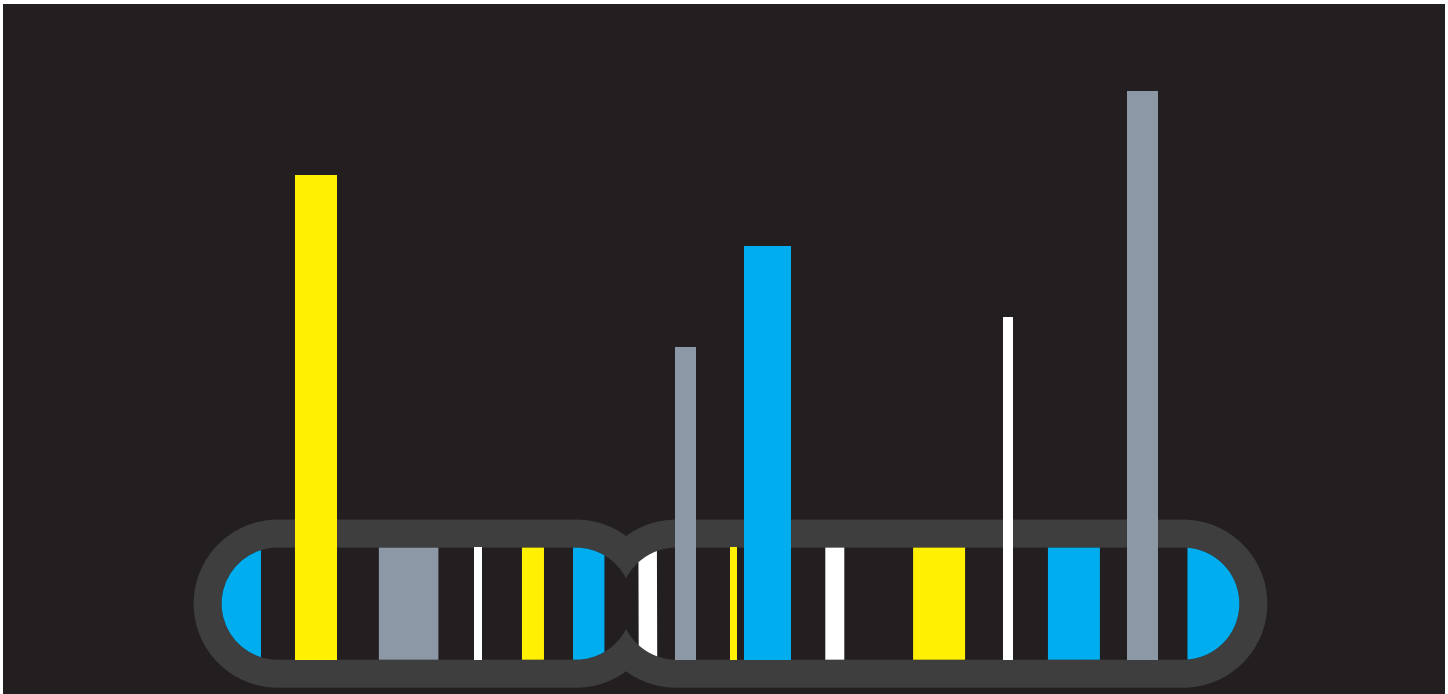


THE GREATEST HITS OF THE HUMAN GENOME

A tour through the most studied genes in biology reveals some surprises.

BY ELIE DOLGIN

Peter Kerpedjiev needed a crash course in genetics. A software engineer with some training in bioinformatics, he was pursuing a PhD and thought it would really help to know some fundamentals of biology. "If I wanted to have an intelligent conversation with someone, what genes do I need to know about?" he wondered.

Kerpedjiev went straight to the data. For years, the US National Library of Medicine (NLM) has been systematically tagging almost every paper in its popular PubMed database that contains some information about what a gene does. Kerpedjiev extracted all the papers marked as describing the structure, function or location of a gene or the protein it encodes.

Sorting through the records, he compiled a list of the most studied genes of all time — a sort of 'top hits' of the human genome, and several other genomes besides.

Heading the list, he found, is a gene called *TP53*. Three years ago, when Kerpedjiev first did his analysis, researchers had scrutinized the gene or the protein it produces, p53, in some 6,600 papers. Today, that number is at about 8,500 and counting. On average, around two papers are published each day describing new details of the basic biology of *TP53*.

Its popularity shouldn't come as news to most biologists. The gene is a tumour suppressor, and widely known as the 'guardian of the genome'. It is mutated in roughly half of all human cancers. "That explains its staying power," says Bert Vogelstein, a cancer geneticist at the Johns Hopkins University School of Medicine in Baltimore, Maryland. In cancer, he says, "there's no gene more important".

But some chart-topping genes are less well known — including some that rose to prominence in bygone eras of genetic research, only to fall out of fashion as technology progressed. "The list was surprising," says Kerpedjiev, now a postdoc studying genomic-data visualization at Harvard Medical School in Boston, Massachusetts. "Some genes were predictable; others were completely unexpected."

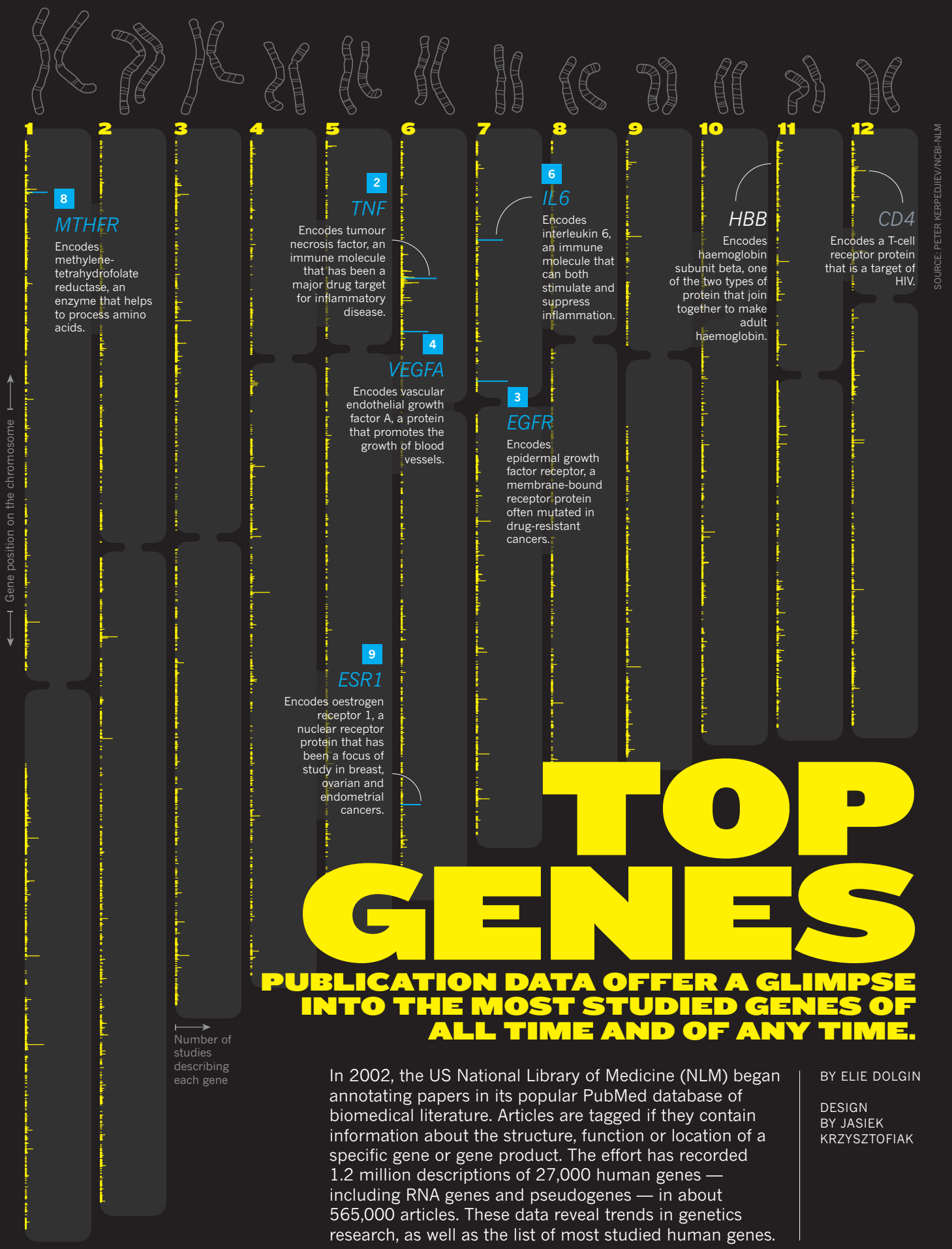
To find out more, *Nature* worked with Kerpedjiev to analyse the most studied genes of all time (see 'Top genes'). The exercise offers more than a conversation starter: it sheds light on important trends in biomedical research, revealing how concerns over specific diseases or public-health issues have shifted research priorities towards underlying genes. It also shows how just a few genes, many of which span disciplines and disease areas, have dominated research.

Out of the 20,000 or so protein-coding genes in the human genome, just 100 account for more than one-quarter of the papers tagged by the NLM. Thousands go unstudied in any given year. "It's revealing how much we don't know about because we just don't bother to research it," says Helen Anne Curry, a science historian at the University of Cambridge, UK.

IN AND OUT OF FASHION

In 2002, just after the first drafts of the human genome were published, the NLM started systematically adding 'gene reference into function', or GeneRIF, tags to papers[1]. It has extended that annotation back to the 1960s, sometimes using other databases to help fill in the details. It is not a perfectly curated record. "In general, the data set is somewhat noisy," says Terence Murphy, a staff scientist at the NLM in Bethesda, Maryland. There's probably some sampling bias for papers published before 2002, he warns. That means that some genes are over-represented and a few may be erroneously missing. "But it's not awful," Murphy says. "As you aggregate over multiple genes, that potentially reduces some of these biases."

With that caveat noted, the PubMed records reveal a few distinct historical periods in which gene-related papers tended to focus on particular hot topics (see 'Fashionable genes through the years'). Before the mid-1980s, for



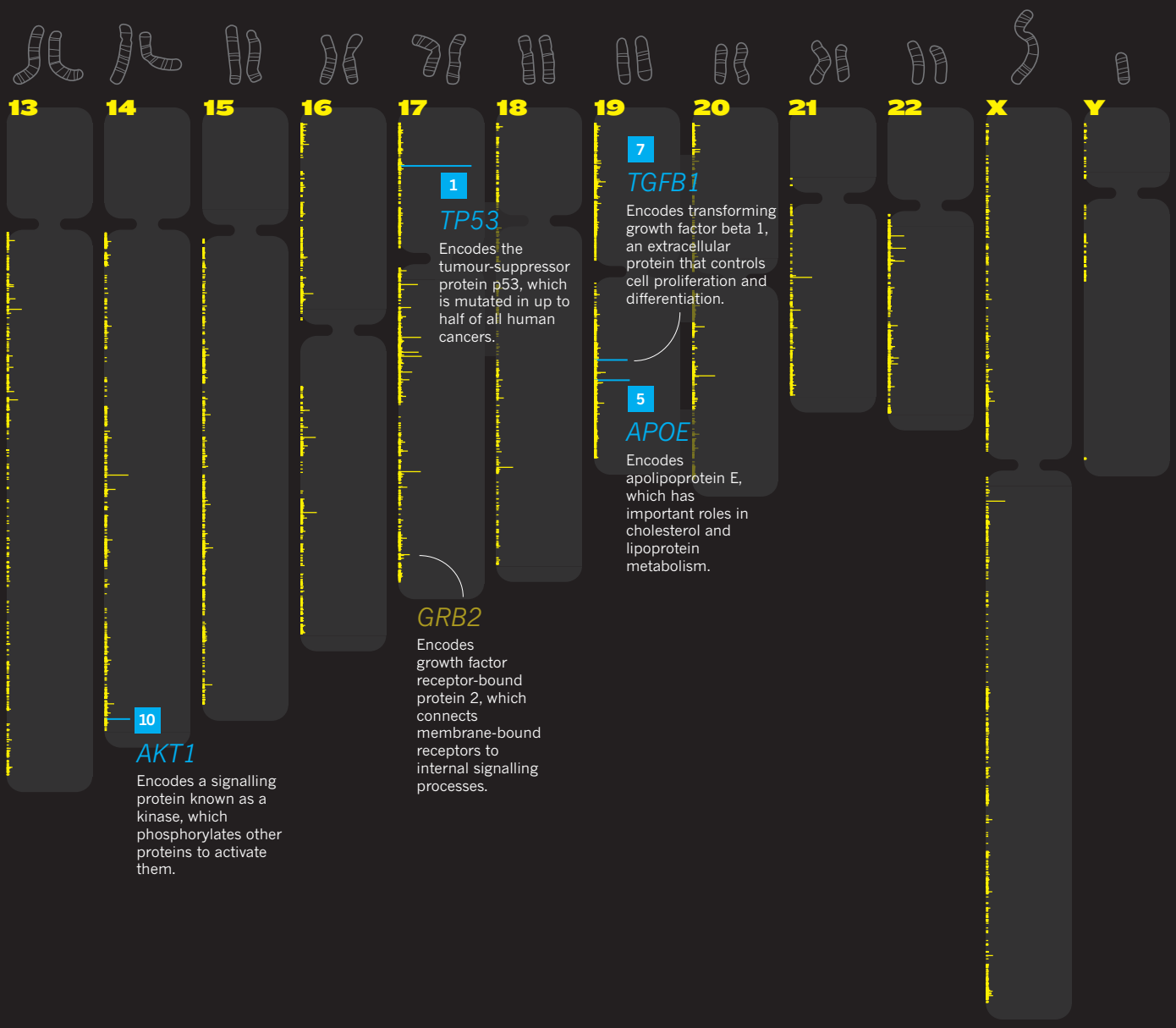
TOP GENES

PUBLICATION DATA OFFER A GLIMPSE INTO THE MOST STUDIED GENES OF ALL TIME AND OF ANY TIME.

In 2002, the US National Library of Medicine (NLM) began annotating papers in its popular PubMed database of biomedical literature. Articles are tagged if they contain information about the structure, function or location of a specific gene or gene product. The effort has recorded 1.2 million descriptions of 27,000 human genes — including RNA genes and pseudogenes — in about 565,000 articles. These data reveal trends in genetics research, as well as the list of most studied human genes.

BY ELIE DOLGIN

DESIGN BY JASIEK KRZYSZTOFIAK



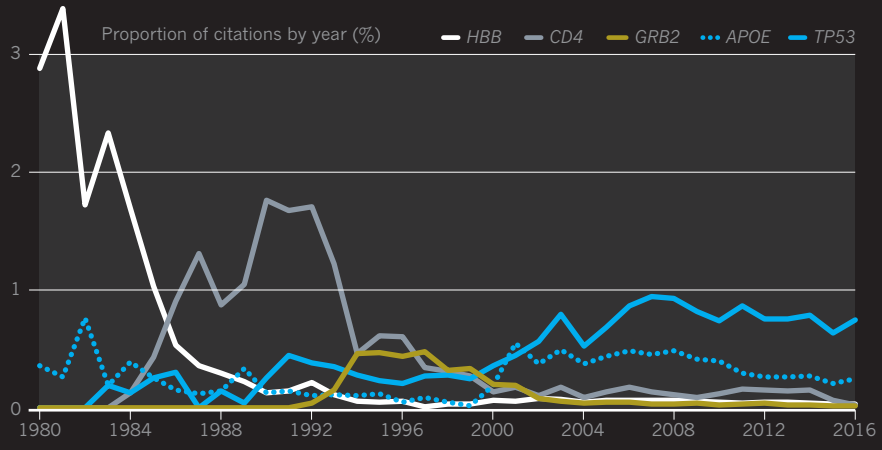
THE TOP 10

The ten most studied genes of all time are described in more than 40,000 papers.

1	*TP53*	8,479 citations
2	*TNF*	5,314
3	*EGFR*	4,583
4	*VEGFA*	4,059
5	*APOE*	3,977
6	*IL6*	3,930
7	*TGFB1*	3,715
8	*MTHFR*	3,256
9	*ESR1*	2,864
10	*AKT1*	2,791

FASHIONABLE GENES THROUGH THE YEARS

Different genes have dominated the research literature in different eras. The trends reflect new understanding about the development of genetic diseases such as sickle-cell anaemia (*HBB*), concerns about new infectious diseases (*CD4*), breakthroughs in cellular signalling (*GRB2*) and more.



BEYOND HUMAN

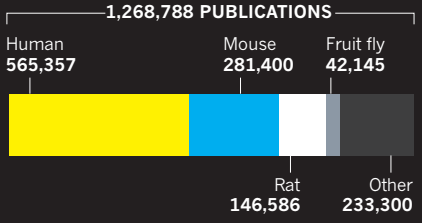
The US National Library of Medicine has tracked references to genes from dozens of species, including mice, flies and other important model organisms, as well as viruses. Looking at genes from all species, more than two-thirds of the 100 most studied genes over the past 50 years have been human (see 'The gene menagerie'). But non-human genes do appear quite high on the list. Often, these have a clear link to human health, as with mouse versions of *TP53*, or *env*, a viral gene that encodes envelope proteins involved in gaining entry to a cell.

Others became foundational to broader genetic studies. A gene from the fruit fly *Drosophila melanogaster* known simply as *white* has been the focus of about 3,600 papers — dating back to when biologist Thomas Hunt Morgan, working at Columbia University in New York City, peered through a hand lens one day in 1910 and saw a single male fly with white eyes instead of red[11]. Because its product causes an easily observable change in the fly, the *white* gene serves as a marker for scientists looking to map and manipulate the fly genome. It has been involved in many fundamental discoveries[12], such as the demonstration that large stretches of DNA can be duplicated because of unequal exchange between matching chromosomes.

The most popular non-human gene of all time is actually a spot in the mouse genome whose normal function remains poorly understood. *Rosa26* comes from an experiment published[13] in 1991, in which cell biologists Philippe Soriano and Glenn Friedrich used a virus to insert an engineered gene randomly into mouse embryonic stem cells. In one cell line, dubbed ROSA26, the engineered gene seemed to be active at all times and in nearly every cell type. The discovery served as a building block for the creation of tools to make and manipulate transgenic mice. "People starting using it like crazy," recalls Soriano, who is now at the Icahn School of Medicine at Mount Sinai in New York City. So far, the genetic locus known as *Rosa26* has been involved in some 6,500 functional studies. It is second only to *TP53*. **E.D.**

THE GENE MENAGERIE

Of about 1.3 million publications pertaining to genes in any species, nearly half are on human genes.



SOURCE: PETER KERPEDJIEV/NCBI-NLM

example, much genetic research centred on haemoglobin, the oxygen-carrying molecule found in red blood cells. More than 10% of all studies on human genetics before 1985 were about haemoglobin in some way.

At the time, researchers were still building on the early work of Linus Pauling and Vernon Ingram, trailblazing biochemists who pioneered the study of disease at a molecular level with discoveries in the 1940s and 1950s of how abnormal haemoglobin caused sickle-cell disease. Molecular biologist Max Perutz, who won a share in the 1962 Nobel Prize in Chemistry for his 3D map of haemoglobin's structure, continued to explore how the protein's shape related to its function for decades afterwards.

According to Alan Schechter, a physician-scientist and senior historical consultant at the US National Institutes of Health in Bethesda, the haemoglobin genes — more than any others at the time — offered "an entryway to understanding and perhaps treating a molecular disease".

A sickle-cell researcher himself, Schechter says that such genes were a focus of conversation both at major genetics meetings and at blood-disease meetings in the 1970s and early 1980s. But as researchers gained access to new technologies for sequencing and manipulating DNA, they started to move on to other genes and diseases, including a then-mysterious infection that was predominantly striking down gay men.

Even before the 1983 discovery that HIV was the cause of AIDS, clinical immunologists such as David Klatzmann had noticed a peculiar pattern among people with the illness. "I was just struck by the fact that these people had no T4 cells," recalls Klatzmann, who is now at Pierre and Marie Curie University in Paris. He showed[2] in cell-culture experiments that HIV seemed to selectively infect and destroy these cells, a subset of the immune system's T cells. The question was: how was the virus getting into the cell?

Klatzmann reasoned that the surface protein (later called CD4) that immunologists used to define this set of cells might also serve as the receptor through which HIV entered the cell. He was right, as he reported[3] in a study published in December 1984, alongside a similar paper[4] from molecular virologist Robin Weiss, then at the Institute of Cancer Research in London, and his colleagues.

Within three years, *CD4* was the top gene in the biomedical literature. It remained so from 1987 to 1996, a period in which it accounted for 1–2% of all the tags tallied by the NLM.

That attention stemmed in part from efforts to tackle the emerging AIDS crisis. In the late 1980s, for example, several companies dabbled with the idea of engineering therapeutic forms of the CD4 protein that could mop up HIV particles before they infected healthy cells. But results from small human trials proved "underwhelming", says Jeffrey Lifson, director of the AIDS and Cancer Virus Program at the US National Cancer Institute in Frederick, Maryland.

An even bigger part of *CD4*'s popularity had to do with basic immunology. In 1986, researchers realized that CD4-expressing T cells could be subdivided into two distinct populations — one that eliminates cell-infecting bacteria and viruses, and one that guards against parasites such as worms, which cause illness without invading cells. "It was a fairly exciting time, because we really understood very little," says Dan Littman, an immunologist at the New York University School of Medicine. Just the year before, he had helped to clone the DNA that encodes CD4 and insert it into bacteria[5], so that vast quantities of the protein could be made for research.

A decade later, Littman also co-led one of three teams to show[6] that to enter cells, HIV uses another receptor alongside CD4: a protein identified as CCR5. These, and a second co-receptor called CXCR4, have remained the focus of intensive, global HIV research ever since, with the goal — as-yet unfulfilled — of blocking the virus's entry into cells.

FIFTEEN MINUTES OF FAME

By the early 1990s, *TP53* was already ascendant. But before it climbed to the top of the human gene ladder, there were a few years in which a lesser-known gene called *GRB2* was in the spotlight.

At the time, researchers were starting to identify the specific protein interactions involved in cell communication. Thanks to pioneering work by cell biologist Tony Pawson, scientists knew that some small intracellular proteins contained a module called SH2, which could bind to activated proteins at the surface of cells and relay a signal to the nucleus.

In 1992, Joseph Schlessinger, a biochemist at the Yale University School of Medicine in New Haven, Connecticut, showed[7] that the protein encoded by *GRB2* — growth factor receptor-bound protein 2 — was that relay point. It contains an SH2 module as well as two domains that activate proteins involved in cell growth and survival. "It's a molecular matchmaker," Schlessinger says.

Other researchers soon filled in the gaps, opening a field of study in signal transduction. And although many other building blocks of cell signalling were soon unearthed — ultimately leading to treatments for cancer, autoimmune disorders, diabetes and heart disease — *GRB2* stayed at the forefront and was the top-referenced gene for three years in the late 1990s.

In part, that was because *GRB2* "was the first physical connection between two parts of the signal-transduction cascade", says Peter van

der Geer, a biochemist at San Diego State University in California. Furthermore, "it's involved in so many different aspects of cellular regulation".

GRB2 is something of an outlier in the most-studied list. It's not a direct cause of disease; nor is it a drug target, which perhaps explains why its moment in the sun was fleeting. "You have some rising stars that fall down very quickly because they have no clinical value," says Thierry Soussi, a long-time *TP53* researcher at the Karolinska Institute in Stockholm and Pierre and Marie Curie University. Genes with staying power usually show some sort of therapeutic potential that attracts funding agencies' support. "It's always like that," Soussi says. "The importance of a gene is linked to its clinical value."

It can also be linked to certain properties of the gene, such as the levels at which it is expressed, how much it varies between populations and the characteristics of its structure. That's according to an analysis by Thomas Stoeger, a systems biologist at Northwestern University in Evanston, Illinois, who reported this month at a symposium in Heidelberg, Germany, that he could predict which genes would garner the most attention, simply by plugging such attributes into an algorithm.

Stoeger thinks that the reasons for these associations largely boil down to what he calls discoverability. The popular genes happened to be in hot areas of biology and could be probed with the tools available at the time. "It's easier to study some things than others," says Stoeger — and that's a problem, because vast numbers of genes remain uncharacterized and under-explored, leaving major gaps in the understanding of human health and disease.

"THE IMPORTANCE OF A GENE IS LINKED TO ITS CLINICAL VALUE."

Curry also points to "intertwined technical, social and economic factors" shaped by politicians, drugmakers and patient advocates.

RIGHT PLACE, RIGHT TIME

Stoeger has also tracked how the general features of popular genes have changed over time. He found, for example, that in the 1980s, researchers focused largely on genes whose protein products were found outside cells. That's probably because these proteins were easiest to isolate and study. Only more recently did attention shift towards genes whose products are found inside the cell.

That shift happened alongside the publication of the human genome, says Stoeger. The advance would have opened up a larger percentage of genes to enquiry.

Many of the most explored genes, however, don't fit these larger trends. The p53 protein, for example, is active inside the nucleus. Yet *TP53* became the most studied gene around 2000. It, like many of the genes that came to dominate biological research, was not properly understood after its initial discovery — which may explain why it took several decades after the 1979 characterization of the protein for the gene to rise to the top spot in the literature.

At first, the cancer-research community mistook it for an oncogene — one that, when mutated, drives the development of cancer. It wasn't until 1989 that Suzanne Baker, a graduate student in Vogelstein's lab, showed[8] that it was actually a tumour suppressor. Only then did functional studies of the gene really begin to pick up steam. "You can see from the spike in publications that go up essentially at that point that there were a lot of people who were really very interested," says Baker, now a brain-tumour researcher at the St. Jude Children's Research Hospital in Memphis, Tennessee.

Research into human cancer also brought scientists to *TNF*, the runner-up to *TP53* as the most-referenced human gene of all time, with more than 5,300 citations in the NLM data (see 'The top 10'). It encodes a protein — tumour necrosis factor — named in 1975 because of its ability to kill cancer cells. But anticancer action proved not to be *TNF*'s main function. Therapeutic forms of the TNF protein were highly toxic when tested in people.

The gene turned out to be a mediator of inflammation; its effect on tumours was secondary. Once that became clear in the mid-1980s, attention quickly shifted to testing antibodies that block its action. Now, anti-TNF therapies are mainstays of treatment for inflammatory disorders such as rheumatoid arthritis, collectively pulling in tens of billions of dollars in annual sales worldwide.

"This is an example where the knowledge of the gene and the gene product has relatively quickly changed the health of the world," says Kevin Tracey, a neurosurgeon and immunologist at the Feinstein Institute for Medical Research in Manhasset, New York.

TP53's dominance was briefly interrupted by another gene, *APOE*. First described in the mid-1970s as a transporter involved in clearing cholesterol from the blood, the APOE protein was "seriously considered" as a lipid-lowering treatment for preventing heart disease, says Robert Mahley, a pioneer in the field at the University of California, San Francisco, who tested the approach in rabbits[9].

Ultimately, the creation of statins in the late 1980s doomed this strategy to the dustbin of pharmaceutical history. But then, neuroscientist Allen Roses and his colleagues found the APOE protein bound up in the sticky brain plaques of people with Alzheimer's disease. They showed[10] in 1993 that one particular form of the gene, *APOE4*, was associated with a greatly increased risk of the disease.

This generated much wider interest in the gene. Still, it took time to move up the most-studied chart. "The reception was very cool," recalls Ann Saunders, a neurogeneticist and chief executive of Zinfandel Pharmaceuticals in Chapel Hill, North Carolina, who collaborated with Roses, her late husband. The amyloid hypothesis, which states that build-up of a protein fragment called amyloid-β is responsible for the disease, was all the rage in the Alzheimer's-research community at the time. And few researchers seemed interested in finding out what a cholesterol-transport protein had to do with the disease. But the genetic link between *APOE4* and Alzheimer's risk proved "irrefutable", Mahley says, and in 2001, *APOE* briefly overtook *TP53*. It remains in the all-time top five, at least for humans (see 'Beyond human').

Like other popular genes, *APOE* is well studied because it's central to one of the biggest unsolved health problems of the day. But it's also important because anti-amyloid therapies have mostly flamed out in clinical testing. "I hate saying this, but what helped me were the failed trials," says Mahley, who this year raised US$63 million for his company E-Scape Bio to develop drugs that target the APOE4 protein. Those failures, he says, forced industry and funding agencies to rethink therapeutic strategies for tackling Alzheimer's.

There's the rub: it takes a certain confluence of biology, societal pressure, business opportunity and medical need for any gene to become more studied than any other. But once it has made it to the upper echelons, there's a "level of conservatism", says Gregory Radick, a science historian at the University of Leeds, UK, "with certain genes emerging as safe bets and then persisting until conditions change".

The question now is how conditions might change. What new discoveries might send a new gene up the chart — and knock today's top genes off their pedestal? ■

Elie Dolgin *is a science writer in Somerville, Massachusetts.*

1. Mitchell, J. A. *et al. AMIA Annu. Symp. Proc.* **2003,** 460–464 (2003).
2. Klatzmann, D. *et al. Science* **225,** 59–63 (1984).
3. Klatzmann, D. *et al. Nature* **312,** 767–768 (1984).
4. Dalgleish, A. G. *et al. Nature* **312,** 763–767 (1984).
5. Maddon, P. J. *et al. Cell* **42,** 93–104 (1985).
6. Deng, H. *et al. Nature* **381,** 661–666 (1996).
7. Lowenstein, E. J. *et al. Cell* **70,** 431–442 (1992).
8. Baker, S. J. *et al. Science* **244,** 217–221 (1989).
9. Mahley, R. W. *et al. J. Clin. Invest.* **83,** 2125–2130 (1989).
10. Strittmatter, W. J. *et al. Proc. Natl Acad. Sci. USA* **90,** 1977–1981 (1993).
11. Morgan, T. H. *Science* **32,** 120–122 (1910).
12. Green, M. M. *Genetics* **184,** 3–7 (2010).
13. Friedrich, G. & Soriano, P. *Genes Dev.* **5,** 1513–1523 (1991).