GENES FROM THE JUNKYARD

Scientists long assumed that new genes appear when evolution tinkers with old ones. It turns out that natural selection is much more creative.

BY ADAM LEVY

In the depths of winter, water temperatures in the ice-covered Arctic Ocean can sink below zero. That's cold enough to freeze many fish, but the conditions don't trouble the cod. A protein in its blood and tissues binds to tiny ice crystals and stops them from growing.

Where codfish got this talent was a puzzle that evolutionary biologist Helle Tessand Baalsrud wanted to solve. She and her team at the University of Oslo searched the genomes of the Atlantic cod (*Gadus morhua*) and several of its closest relatives, thinking they would track down the cousins of the antifreeze gene. None showed up. Baalsrud, who at the time was a new parent, worried that her lack of sleep was causing her to miss something obvious.

But then she stumbled on studies suggesting that genes do not always evolve from existing ones, as biologists long supposed. Instead, some are fashioned from desolate stretches of the genome that do not code for any functional molecules. When she looked back at the fish genomes, she saw hints this might be the case: the antifreeze protein — essential to the cod's survival — had seemingly been built from scratch[1]. By that point, another researcher had reached a similar conclusion[2].

The cod is in good company. In the past five years, researchers have found numerous signs of these newly minted '*de novo*' genes in every lineage they have surveyed. These include model organisms such as fruit flies and mice, important crop plants and humans; some of the genes are expressed in brain and testicular tissue, others in various cancers.

De novo genes are even prompting a rethink of some portions of evolutionary theory. Conventional wisdom was that new genes tended to arise when existing ones are accidentally duplicated, blended with others or broken up, but some researchers now think that *de novo* genes could be quite common: some studies

PAUL NICKLEN/NG IMAGE COLLECTION

Some cod species have a newly minted gene involved in preventing freezing.

suggest at least one-tenth of genes could be made in this way; others estimate that more genes could emerge *de novo* than from gene duplication. Their existence blurs the boundaries of what constitutes a gene, revealing that the starting material for some new genes is non-coding DNA (see 'Birth of a gene').

The ability of organisms to acquire new genes in this way is testament to evolution's "plasticity to make something seemingly impossible, possible", says Yong Zhang, a geneticist at the Chinese Academy of Sciences' Institute of Zoology in Beijing, who has studied the role of *de novo* genes in the human brain.

But researchers have yet to work out how to definitively identify a gene as being *de novo*, and questions still remain over exactly how — and how often — they are born. Scientists also wonder why evolution would bother making genes from scratch when so much gene-ready material already exists. Such basic questions are a sign of how young the field is. "You don't have to go back that many years before *de novo* gene evolution was dismissed," Baalsrud says.

NEW ARRIVALS

Back in the 1970s, geneticists saw evolution as a rather conservative process. When Susumu Ohno laid out the hypothesis that most genes evolved through duplication[3], he wrote that "In a strict sense, nothing in evolution is created *de novo*. Each new gene must have arisen from an already existing gene."

Gene duplication occurs when errors in the DNA-replication process produce multiple instances of a gene. Over generations, the versions accrue mutations and diverge, so that they eventually encode different molecules, each with their own function. Since the 1970s, researchers have found a raft of other examples of how evolution tinkers with genes — existing genes can be broken up or 'laterally transferred' between species. All these processes have something in common: their main ingredient is existing code from a well-oiled molecular machine.

But genomes contain much more than just genes: in fact, only a few per cent of the human genome, for example, actually encodes genes. Alongside are substantial stretches of DNA — often labelled 'junk DNA' — that seem to lack any function. Some of these stretches share features with protein-coding genes without actually being genes themselves: for instance, they are littered with three-letter codons that could, in theory, tell the cell to translate the code into a protein.

It wasn't until the twenty-first century that scientists began to see hints that non-coding sections of DNA could lead to new functional codes for proteins. As genetic sequencing advanced to the point that researchers could compare entire genomes of close relatives, they began to find evidence that genes could disappear rather quickly during evolution. That made them wonder whether genes could just as quickly spring into being.

In 2006 and 2007, evolutionary geneticist David Begun at the University of California, Davis, published what many regard as the first papers to make the case for particular genes arising *de novo* in fruit flies[4,5]. The studies linked these genes to male reproduction: Begun found they were expressed in the testes and the seminal fluid gland, where it seemed the powerful evolutionary force of sexual selection was driving gene birth.

Shortly before that, evolutionary genomicist Mar Albà at the Hospital del Mar Medical Research Institute in Barcelona, Spain, had shown that the younger a gene is, evolutionarily speaking, the faster it tends to evolve[6]. She speculated that this might be because the molecules encoded by younger genes are less polished and need more tuning, and that this could be a consequence of the genes having arisen *de novo* — they were not tied to a previous function as tightly as those that had evolved from older genes. Both Albà and Begun recall that it was challenging to publish their early work on the topic. "There was a lot of scepticism," says Albà. "It's amazing how things have changed."

> "IT'S LIKE A BETA LAUNCH."

Studies have also started to unpick what *de novo* genes do. One gene allows the thale cress plant (*Arabidopsis thaliana*) to produce starch, for instance, and another helps yeast cells to grow. Understanding what they are doing for their hosts should help to explain why they exist — why it is advantageous to create from scratch rather than evolve from existing material. "We're not going to understand why these genes are evolving if we don't understand what they're doing," says Begun.

GENES-IN-WAITING

Studying *de novo* genes turns out to be part genetics, part thought experiment. "Why is our field so difficult?" asks Anne-Ruxandra Carvunis at the University of Pittsburgh in Pennsylvania. "It is because of philosophical issues." At its heart is a question that Carvunis has been asking for a decade: what is a gene?

A gene is commonly defined as a DNA or RNA sequence that codes for a functional molecule. The yeast genome, however, has hundreds of thousands of sequences, known as open reading frames (ORFs), that could theoretically be translated into proteins, but that geneticists assumed were either too short or looked too different from those in closely related organisms to have a probable function.

When Carvunis studied yeast ORFs for her PhD, she began to suspect that not all of these sections were lying dormant. In a study[7] published in 2012, she looked at whether these ORFs were being transcribed into RNA and translated into proteins — and, just like genes, many of them were — although it was unclear whether the proteins were useful to the yeast, or whether they were translated at high enough levels to serve a function. "So what is a gene? I don't know," Carvunis says. What she thinks she has found, though, is "raw material — a reservoir — for evolution".

Some of these genes-in-waiting, or what Carvunis and her colleagues called proto-genes, were more gene-like than others, with longer sequences and more of the instructions necessary for turning the DNA into proteins. The proto-genes could provide a fertile testing ground for evolution to convert non-coding material into true genes. "It's like a beta launch," suggests Aoife McLysaght, who works on molecular evolution at Trinity College Dublin.

Some researchers have gone beyond observation to manipulate organisms into expressing non-coding material. Michael Knopp and his colleagues at Uppsala University, Sweden, showed that inserting and expressing randomly generated ORFs into *Escherichia coli* could enhance the bacterium's resistance to antibiotics, with one sequence producing a peptide that increased resistance 48-fold[8]. Using a similar approach, Diethard Tautz and his team at the Max Planck Institute for Evolutionary Biology in Plön, Germany, showed that half of the sequences slowed the bacterium's growth, and one-quarter seemed to speed it up[9] — although that result is debated. Such studies suggest that peptides from random sequences can be surprisingly functional.

But random sequences of DNA could also code for peptides that are "reactive and nasty and have a tendency to aggregate and do bad things", says evolutionary biologist Joanna Masel of the University of Arizona in Tucson. Expressing these sequences at low levels could help natural selection to weed out potentially dangerous portions — those that create messy or misfolded proteins — so that what remains in a species is relatively benign.

Creating genes from non-coding regions could have some benefits over other gene-making methods, says Albà. Gene duplication is a "very conservative mechanism" she says, producing well-adapted proteins cut from the same cloth as their ancestors; *de novo* genes, by contrast, are likely to produce markedly different molecules. That could make it difficult for them to fit into well-established networks of genes and proteins — but they could also be better suited to certain new tasks.

A newly minted gene could help an

organism to respond to a change in its environment, for instance. This seems to have been the case for the cod, which acquired its antifreeze protein as the Northern Hemisphere cooled some 15 million years ago.

BIRTH RATE

To trace which of an organism's genes were made *de novo*, researchers need comprehensive sequences for the organism and its close relatives. One crop plant that fits the bill is rice. The sweltering heat of Hainan, a tropical island in southern China, is the perfect environment for growing the crop — although the working conditions can be trying. "It's horrible," says evolutionary geneticist Manyuan Long of the University of Chicago, Illinois. It's so hot "you can cook your egg in the sand".

Long's team wanted to know how many genes had emerged *de novo* in the strain *Oryza sativa japonica*, and what proteins those genes might be making. So the team lined up its genome against those of its close relatives and used an algorithm to pick out regions that contained a gene in some species but lacked it in others. This allowed the researchers to identify the non-coding DNA that led to the gene in question, and track its journey to being a gene. They could also tot up the number of *de novo* genes that appeared in the strain: 175 genes over 3.4 million years of evolution[10] (over the same period, the strain gained 8 times as many genes from duplication).

The study gets at one of the field's biggest preoccupations: how to tell whether a gene is truly *de novo*. Answers vary wildly, and approaches are still evolving. For example, an early study found 15 *de novo* genes in the whole primate order[11]; a later attempt found 60 in humans alone[12]. One option for finding candidate *de novo* genes is to use an algorithm to search for similar genes in related species. If nothing shows up, then it's possible that the gene arose *de novo*. But failing to find a relative doesn't mean no relative is there: the gene could have been lost along the way, or might have shape-shifted far away from its kin. The rice study got around this by explicitly identifying the pieces of non-coding DNA that became *de novo* genes.

Over long evolutionary timescales — much longer than the few million years of rice evolution — it is hard to distinguish between a *de novo* gene and one that has simply diverged too far from its ancestors to be recognizable, so determining the absolute number of genes that have arisen *de novo* rather than from duplication "is an almost unanswerable question", says Tautz.

To demonstrate how varied the results of different methods can be, evolutionary geneticist Claudio Casola at Texas A&M University in College Station used alternative approaches to reanalyse the results of previous studies, and failed to verify 40% of the *de novo* genes they had proposed[13]. To Casola, this points to the need to standardize tests. Currently, he says, "it seems to be very inconsistent".

Counting *de novo* genes in the human genome comes with the same trail of caveats. But where *de novo* genes have been identified, researchers are beginning to explore their roles in health and disease. Zhang and his colleagues have found that one gene unique to humans is expressed at a greater level in the brains of people with Alzheimer's disease[14], and previous work[15] had linked certain variants of the gene to nicotine dependence. For Zhang, research that links *de novo* genes to the human brain is tantalizing. "We know that what makes us human is our brains," he says, "so there must be some genetic kit to push the evolution of our brain." That suggests an avenue for future studies. Zhang suggests that researchers could investigate the genetic kit through experiments with human organoids — cultured cells that serve as a model organ.

De novo genes could have implications for understanding cancer, too. One such gene — unique to humans and chimpanzees — has been linked to cancer progression in mouse models of neuroblastoma[16]. And cancer-causing versions of human papillomavirus include a gene that is not present in non-cancer-causing forms[17].

Many *de novo* genes remain uncharacterized, so the potential importance of the process to health and disease is unclear. "It will take some time before we fully understand to what extent it contributes to human health and to what extent it contributes to the origin of the human species," says Carvunis.

Although *de novo* genes remain enigmatic, their existence makes one thing clear: evolution can readily make something from nothing. "One of the beauties of working with *de novo* genes," says Casola, "is that it shows how dynamic genomes are." ■

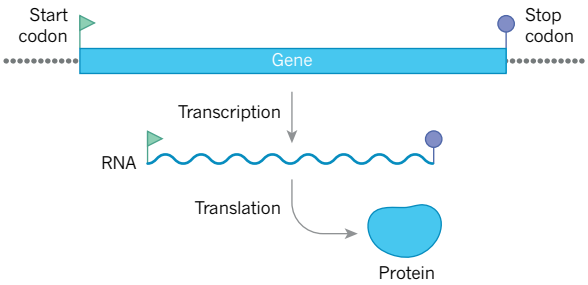
Adam Levy *is a science journalist based in London.*

BIRTH OF A GENE

Scientists long assumed that evolution made new genes from old ones — by copying them in error, or by fusing together or breaking apart existing ones. Now, more and more examples are emerging of genes being created '*de novo*', from barren non-coding portions of the genome.

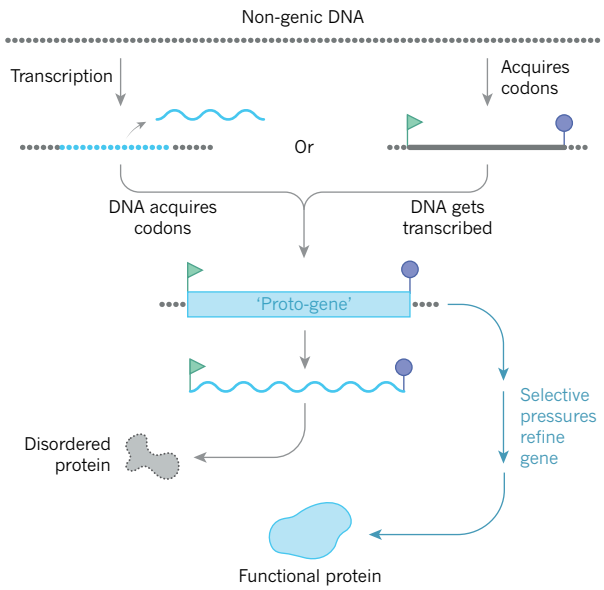
How genes work

Genes are usually considered any stretches of DNA that code for useful molecules. To make a protein, DNA is transcribed into RNA, which is then translated. Three-letter pieces of sequence called codons dictate which portions of the RNA to translate.



Making a gene *de novo*

Genes can evolve from non-coding portions of DNA by gaining transcription and codons, in either order. At first, the products of these 'proto-genes' might be dysfunctional or disordered.



NIK SPENCER/*NATURE*

1. Baalsrud, H. T. *et al. Mol. Biol. Evol.* **35**, 593–606 (2018).
2. Zhuang, X. *Creating sense from non-sense DNA:* de novo *genesis and evolutionary history of antifreeze glycoprotein gene in northern codfishes (gadidae).* PhD thesis, Univ. Illinois Urbana-Champaign (2014).
3. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
4. Begun, D. J., Lindfors, H. A., Thompson, M. E. & Holloway, A. K. *Genetics* **172**, 1675–1681 (2006).
5. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. *Genetics* **176**, 1131–1137 (2007).
6. Albà, M. M. & Castresana, J. *Mol. Biol. Evol.* **22**, 598–606 (2005).
7. Carvunis, A.-R. *et al. Nature* **487**, 370–374 (2012).
8. Knopp, M. *et al. mBio* **10**, e00837-19 (2019).
9. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. *Nature Eco. Evol.* **1**, 0127 (2017).
10. Zhang, L. *et al. Nature Ecol. Evol.* **3**, 679–690 (2019).
11. Toll-Riera, M. *et al. Mol. Biol. Evol.* **26**, 603–612 (2009).
12. Wu, D.-D., Irwin, D. M. & Zhang, Y.-P *PLoS Genet.* **7**, e1002379 (2011).
13. Casola, C. *Genome Biol. Evol.* **10**, 2906–2918 (2018).
14. Li, C.-Y. *et al. PLoS Comput. Biol.* **6**, e1000734 (2010).
15. Wang, D., Ma, J. Z. & Li, M. D. *Pharmacogenomics J.* **5**, 166–172 (2005).
16. Suenaga, Y. *et al. PLoS Genet.* **10**, e1003996 (2014).
17. Willemsen, A., Félez-Sánchez, M. & Bravo, I. G. *Genome Biol. Evol.* **11**, 1602–1617 (2019).

CLARIFICATION

After this story was published, we learnt that a scientist had suggested the *de novo* origin of antifreeze protein genes before Helle Baalsrud's team. The text and reference list have been modified accordingly.