GENETICS

Expanded human gene tally reignites debate

After 15 years, researchers still can't agree on how many genes are in the human genome.

BY CASSANDRA WILLYARD

One of the earliest attempts to estimate the number of genes in the human genome involved tipsy geneticists, a bar in Cold Spring Harbor, New York, and pure guesswork.

That was in 2000, when a draft human genome sequence was still in the works; geneticists were running a sweepstake on how many genes humans have, and wagers ranged from tens of thousands to hundreds of thousands. Almost two decades later, scientists armed with real data still can't agree on the number — a knowledge gap that they say hampers efforts to spot disease-related mutations.

The latest attempt to plug that gap uses data from hundreds of human tissue samples and was posted on the bioRxiv preprint server on 29 May (M. Pertea *et al.* Preprint at bioRxiv http://doi.org/cq5s; 2018). It includes almost 5,000 genes that haven't previously been spotted — among them nearly 1,200 that carry instructions for making proteins. And the overall tally of more than 21,000 protein-coding genes is a substantial jump from previous estimates, which put the figure at around 20,000.

But many geneticists aren't yet convinced that all the newly proposed genes will stand up to close scrutiny. Their criticisms underscore just how difficult it is to identify new genes, or even to define what a gene is.

"People have been working hard at this for 20 years, and we still don't have the answer."

"People have been working hard at this for 20 years, and we still don't have the answer," says Steven Salzberg, a computational biologist at Johns Hopkins University in Baltimore, Maryland, whose team produced the latest count.

HARD TO PIN DOWN

In 2000, with the genomics community abuzz over the question of how many human genes would be found, researcher Ewan Birney launched the GeneSweep contest. Birney, now co-director of the European Bioinformatics Institute (EBI) in Hinxton, UK, took the first bets at a bar during an annual genetics meeting, and the contest eventually attracted more than 1,000 entries and a US$3,000 jackpot. Bets on the number of genes ranged from more than 312,000 to just under 26,000, with an average of around 40,000. These days, the span of estimates has shrunk — with most now between 19,000 and 22,000 — but there is still disagreement (see 'Gene tally').

Salzberg's team used data from the Genotype-Tissue Expression (GTEx) project, which sequenced RNA from more than 30 different tissues taken from several hundred cadavers. RNA is the intermediary between DNA and proteins. The researchers wanted to identify genes that encode a protein and those that don't, but that still have an important role in cells. So they assembled GTEx's 900 billion tiny RNA snippets and aligned them with the human genome.

Just because a stretch of DNA is expressed as RNA, however, does not necessarily mean it's a gene. So the team attempted to filter out noise using a variety of criteria. For example, the researchers compared their results with genomes from other species, reasoning that sequences shared by distantly related creatures have probably been preserved by evolution because they serve a useful purpose, and so are likely to be genes.

The team was left with 21,306 protein-coding genes and 21,856 non-coding genes — many more than are included in the two most widely used human-gene databases. The GENCODE gene set, maintained by the EBI, includes 19,901 protein-coding genes and 15,779 non-coding genes. RefSeq, a database run by the US National Center for Biotechnology Information (NCBI), lists 20,203 protein-coding genes and 17,871 non-coding genes.

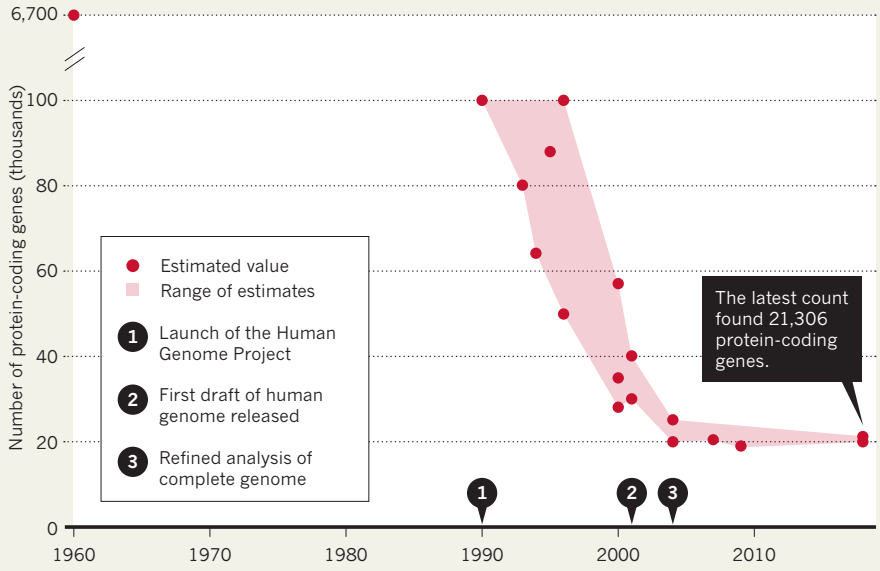
Kim Pruitt, a genome researcher at the NCBI in Bethesda, Maryland, and a former head of RefSeq, says the difference is probably due in part to the volume of data that Salzberg's team analysed. RefSeq relies on an older data set that contains 21 billion short sequences. GENCODE uses different data again: a type that makes recognizing transcripts easier, but which can miss genes. And there's another major difference. Both GENCODE and RefSeq use manual curation — a person reviews the evidence for the gene and makes a final determination. Salzberg's group relied solely on computer programs to sift the data.

"If people like our gene list, then maybe a couple years from now we'll be the arbiter of human genes," says Salzberg.

But many scientists say they need more evidence to be convinced that the latest list is

GENE TALLY

Scientists still don't agree on how many protein-coding genes the human genome holds, but the range of their estimates has narrowed in recent years.



SOURCE: M. PERTEA & S. L. SALZBERG

accurate. Adam Frankish, a computational biologist at the EBI who coordinates the manual annotation of GENCODE, says that he and his group have scanned about 100 of the protein-coding genes identified by Salzberg's team. By their assessment, only one of those seems to be a true protein-coding gene. And Pruitt's team looked at about a dozen of the Salzberg group's new protein-coding genes, but didn't find any that would meet RefSeq's criteria.

Salzberg acknowledges that the new genes on his team's list will require validation by his group and others.

Further confounding counting efforts is the imprecise and changing definition of a gene. Biologists used to see genes as sequences that code for proteins, but then it became clear that some non-coding RNA molecules have important roles in cells. Judging which are important — and should be deemed genes — is controversial, and could explain some of the discrepancies between Salzberg's count and others.

Having an accurate tally of all human genes is key for efforts to uncover links between genes and disease. Uncounted genes are often ignored, even if they contain a disease-causing mutation, Salzberg says. But hastily adding genes to the master list can pose risks, too, says Frankish. A gene that turns out to be incorrect can divert geneticists' attention away from the real problem.

Still, the inconsistencies in the number of genes from database to database are problematic for researchers, Pruitt says. "People want one answer," she adds, "but biology is complex." ■

MEDICAL RESEARCH

Silent cancer cells targeted

Researchers hunt dormant cells that break off tumours, and aim to keep them asleep.

BY HEIDI LEDFORD

After decades of designing drugs to kill rapidly dividing tumour cells, many cancer researchers are switching gears: targeting malignant cells that lie silent and scattered around the body, before they give rise to new tumours.

These cells seed the metastases responsible for about 90% of cancer deaths. They are the source of the heartbreaking cancer resurgence seen in many people whose seemingly successful initial treatment had fostered hopes that they were cured. Treatments that target proliferating tumour cells often miss these silent cells because they're not actively dividing.

Dormant cancer cells are rare, and they are difficult to sift from the trillions of normal cells in the body. For years, scientists lacked the tools to study them, says cancer researcher Julio Aguirre-Ghiso of the Icahn School of Medicine at Mount Sinai in New York City. But that is beginning to change.

From 19 to 22 June, researchers will gather in Montreal, Canada, for what Aguirre-Ghiso says is the first meeting dedicated to these sleeper cancer cells. "The mass of investigators has reached a critical number," he says. "And there is the realization that it's an important clinical need."

That demand is particularly acute in cancers — such as those in the breast, prostate and pancreas — that recur at a high rate, sometimes many years after treatment. "You remove the tumour, you irradiate, you do this, you do that," says cancer researcher Mina Bissell, of the Lawrence Berkeley National Laboratory in California. "But sooner or later the cancer metastasizes, and you say to yourself, 'Where did these things come from?'"

CELL SPOTTING

Mounting evidence suggests that dormant cells break away from a parent tumour early in its development and travel through blood vessels to new sites in the body (see *Nature Methods* **15**, 249–252; 2018). But then, after settling into other tissues or organs, such cells will effectively go to sleep, lying dormant until a trigger — as yet unknown — rouses them. Only then do they begin dividing and form a new tumour.

When cancer researchers tried to study this dormancy, they quickly ran into a problem: mouse models of cancer had been designed to generate quick-growing and highly lethal parent, or primary, tumours. Researchers studying dormancy, however, need slow-growing tumours — which have time to shed rogue cancer cells — and the ability to track those cells long after the primary tumour has been removed.

"Those sorts of animals have been very difficult to develop," says Kathy Miller, a breast-cancer specialist at Indiana University in Indianapolis. But several labs have made progress, developing models to track dormant cells in mice for more than a year.

Techniques for identifying those cells are also improving. Joshua Snyder, a cell biologist at Duke University School of Medicine in Durham, North Carolina, uses a mix of fluorescent markers to identify and trace rogue cells expressing cancer-linked genes.

"As long as those cells remain dormant, they're not killing my patient."

And at the meeting in Montreal, geneticist Jason Bielas of the Fred Hutchinson Cancer Research Center in Seattle, Washington, will present preliminary results from his efforts to barcode such cells using specific DNA sequences. The cells can then be identified using cheap DNA-detection methods at a resolution of about one in one billion cells.

IDENTIFYING INHIBITIONS

Once the silent cells are identified, new methods for determining which genes they express could help researchers to pin down the factors that induce dormancy and the triggers that can rouse sleeping cells. With that information, it might be possible to prevent the cells from waking, says Miller. "As long as those cells remain ▶

MORE ONLINE

NATURE BRIEFING

Save time — get the Nature Briefing direct to your inbox every day go.nature.com/savetime

MORE NEWS

- Controversial US alcohol study cancelled go.nature.com/2mca0gg
- Tech entrepreneur doubles down on critique of NASA mission go.nature.com/2yn2vgh
- Mammals turn to night life to avoid people go.nature.com/2lj7e35

NATURE PODCAST

Pancreatic cancer weight loss; tiny silicon cages; and bias in AI algorithms nature.com/nature/podcast