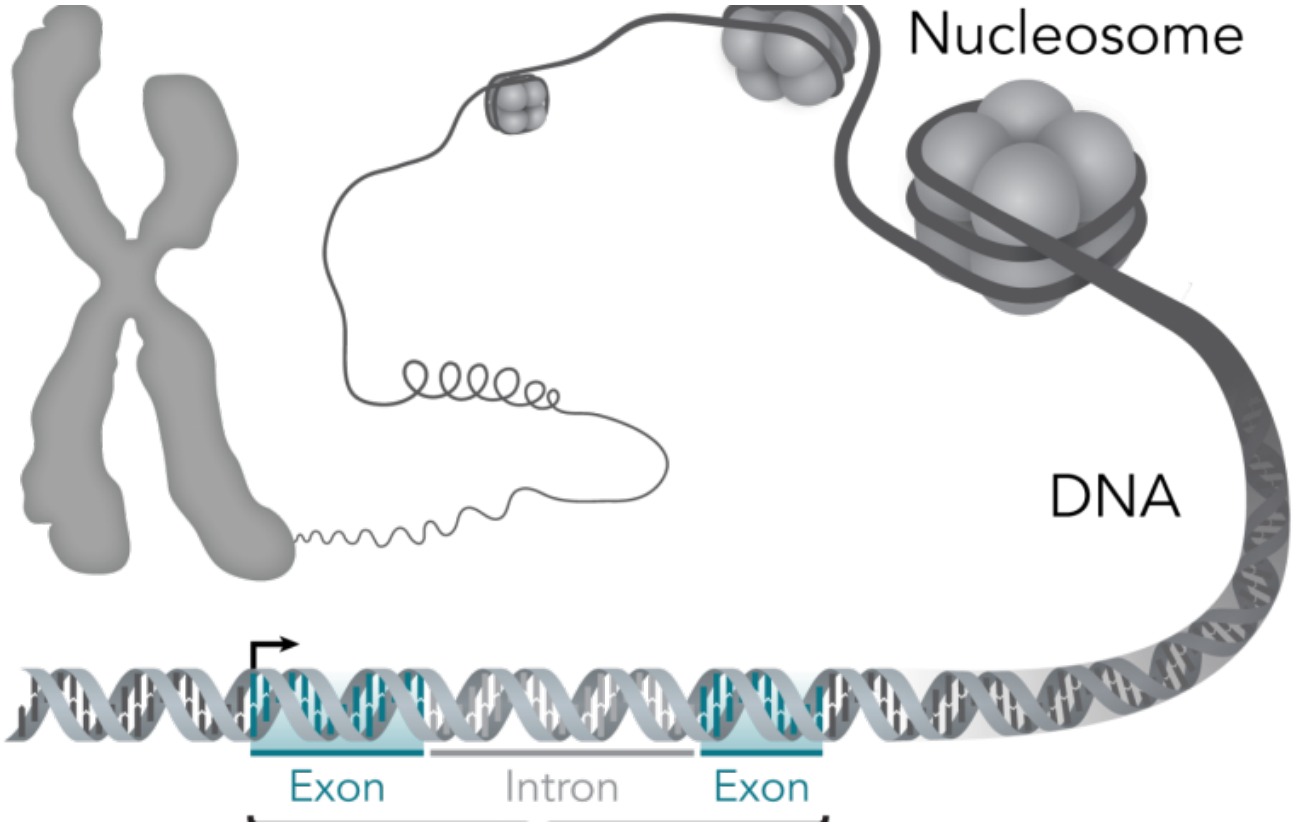
Team reveals that human genome could contain up to 20 percent fewer genes



This stylistic diagram shows a gene in relation to the double helix structure of DNA and to a chromosome (right). The chromosome is X-shaped because it is dividing. Introns are regions often found in eukaryote genes that are removed in the splicing process (after the DNA is transcribed into RNA): Only the exons encode the protein. The diagram labels a region of only 55 or so bases as a gene. In reality, most genes are hundreds of times longer. Credit: Thomas Splettstoesser/Wikipedia/CC BY-SA 4.0

A new study led by the Spanish National Cancer Research Centre (CNIO) reveals that up to 20 percent of genes classified as coding (those that produce the proteins that are the building blocks of all living things) may not be coding after all because they have characteristics that are typical of non-coding or pseudogenes (obsolete coding genes). The consequent reduction in the size of the human genome could have important effects in biomedicine, since the number of genes that produce proteins and their identification is of vital importance for the investigation of multiple diseases, including cancer and cardiovascular diseases.

The work, published in the journal *Nucleic Acids Research*, is the result of an international collaboration led by Michael Tress of the CNIO Bioinformatics Unit along with researchers from the Wellcome Trust Sanger Institute in the United Kingdom, the Massachusetts Institute of Technology in the United States, the Pompeu Fabra University and the National Center for

Supercomputing (BSC-CNS) in Barcelona, and the National Center for Cardiovascular Research (CNIC) in Madrid.

Since the completion of the sequencing of the human genome in 2003, experts from around the world have been working to compile the final human proteome (the total number of proteins generated from genes) and the genes that produce them. This task is immense, given the complexity of the human genome and the fact that humans have about 20,000 separate coding genes.

The researchers analyzed the genes cataloged as protein coding in the main reference human proteomes. The detailed comparison of the reference proteomes from GENCODE/Ensembl, RefSeq and UniProtKB found 22,210 coding genes, but only 19,446 of these genes were present in all 3 annotations.

When they analyzed the 2,764 genes that were present in only one or two of these reference annotations, they were surprised to discover that experimental evidence and manual annotations suggested that almost all of these genes were more likely to be non-coding genes or pseudogenes. In fact, these genes, together with another 1,470 coding genes that are present in the three reference catalogs, were not evolving like typical protein coding genes. The conclusion of the study is that most of these 4,234 genes probably do not code for proteins.

The study is already paying off, according to the scientists. “We have been able to analyze many of these genes in detail,” Tress explains, “and more than 300 genes have already been reclassified as non-coding.” The results are already being included in the new annotations of the human genome by the GENCODE international consortium, of which the CNIO researchers are part.

Conflicting gene numbers in recent years

The work once again highlights doubts about the number of real genes present in human cells 15 years after the sequencing the human genome. Although the most recent data indicates that the number of genes encoding human proteins could exceed 20,000, Federico Abascal, of the Wellcome Trust Sanger Institute in the United Kingdom and first author of the work, says, “Our evidence suggests that humans may only have 19,000 coding genes, but we still do not know which 19,000 genes are.”

For his part, David Juan, of the Pompeu Fabra University and participant in the study, reiterates the importance of these results: “Surprisingly, some of these unusual genes have been well studied and have more than 100 scientific publications based on the assumption that the gene produces a protein. “

This study suggests that there is still a large amount of uncertainty, since the final number of coding genes could 2,000 more or 2,000 fewer than it is now. The human proteome still requires much work, especially given its importance to the medical community.

Explore further:
A new method accelerates the mapping of genes in the ‘Dark Matter’ of our DNA

More information:
Federico Abascal et al. Loose ends: almost one in five human genes still have unresolved coding status, *Nucleic Acids Research* (2018). DOI: 10.1093/nar/gky587

Journal reference:
Nucleic Acids Research

Provided by:
The National Centre for Cancer Research