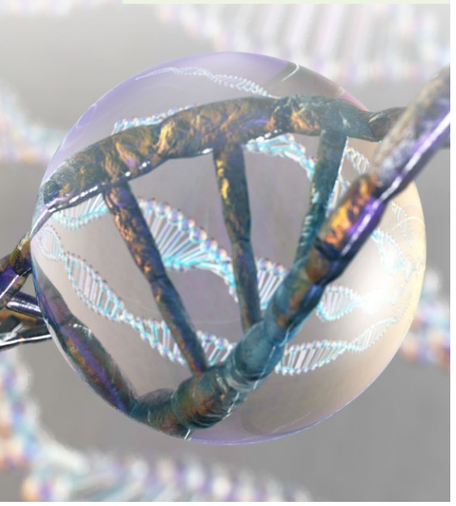
How new RNA genes are born

The study of gene birth and evolution focuses on the identification of ancestral genetic sequences, highly conserved during evolution, that can serve as a foundation for gene development. Nicholas Delihas, Professor Emeritus at the Renaissance School of Medicine at Stony Brook University, New York, has identified one such ancestral element and presented data and a model to show how new RNA genes were created with the ancestral sequence serving as a foundation. Evolutionarily, this small sequence also forms an important part of two ancient genes, the gamma-glutamyltransferase (GGT5) and the ubiquitin specific peptidase 18 (USP18), but how these protein genes were created has yet to be determined.



ustas77777777/Shutterstock.com

Gene duplication, the formation of new genes from an exact copy of existing ones, has long been considered the major process behind gene formation. However, it has been shown that genes may be born from non-coding DNA, regions of DNA that have open reading frames and display translational activity but do not encode proteins. This process is known as de novo protein gene formation based on a protogene sequence or a sequence that does not represent an existing gene. The concept has been formulated by Dr Anne-Ruxandra Carvunis and co-workers at the University of Pittsburgh and by investigators at other institutions. It constitutes a model to explain how during evolution new and different protein genes arose by a de novo mechanism instead of by existing gene sequence duplication.

Dr Nicholas Delihas, at the Renaissance School of Medicine at Stony Brook University, New York, initially was trying to understand a repeat DNA sequence present in an RNA gene family, the FAM230 long intergenic non-coding RNA (lincRNA) family that is found in human chromosome 22. This repeat sequence and the RNA gene family may be related to the onset of human genetic disorders involving aberrant chromosomal recombination and subsequent chromosomal deletions. However, these studies led to an unexpected discovery of an ancestral DNA repeat sequence, termed FAM247, that can serve as a gene-forming element: a nucleation site for new RNA gene development that parallels the concept envisioned for de novo protein birth by Dr Carvunis, co-workers and of scientists at other institutions.

Evolutionary scientists routinely survey the genetic architecture of human and primate populations, along those of other species, to find out the specific roles that different genes play in the way different populations adapt to the environment and to map any changes to the genome that occurred during evolution. Lately, there has been a major focus on the genomic determination and sequencing of long non-coding RNA (lncRNAs) genes, which are now considered to be key players in numerous biochemical pathways. This leads to the question of how many lncRNA genes are present in the genome, their age in evolutionary time and how they were born.

With an increased interest in the study of de novo genes, that is genes made from scratch and not from a template, this has also led to questions on what in fact constitutes a gene. It is generally agreed that a genetic sequence is one that leads to the formation of a functional product, which might be RNA or protein. There are different methods that are employed to confirm that a gene is a functional entity. One approach is to confirm gene expression at the RNA and protein level through biochemical techniques. Another method would be to disrupt a specific genetic sequence and to observe if any changes occur in the phenotype. This could, however, be problematic when analysing entire genomes. Evolutionary approaches look at the presence of specific genetic 'signatures' that provide evidence of selection in an attempt to confirm the presence of a gene. On the other hand, despite the challenges associated with the identification of genes, there is now plenty of evidence that the phenomenon of de novo gene formation has occurred throughout evolutionary history.

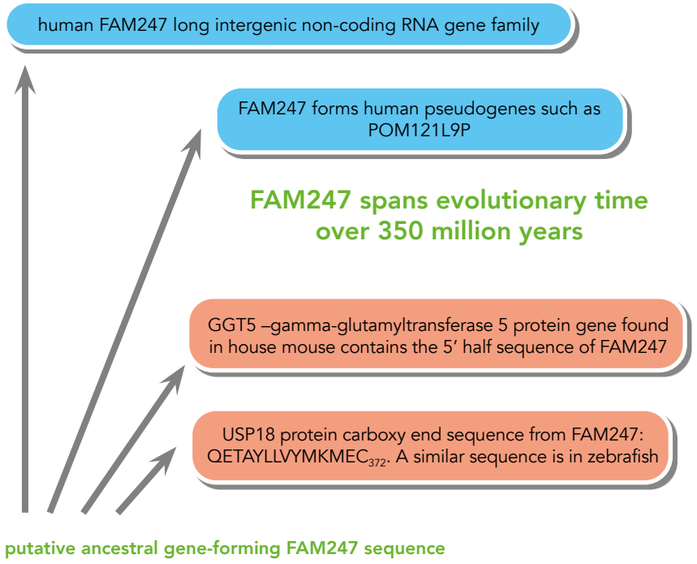


Fig 1. The FAM247 sequence forms various genes over evolutionary time. The pseudogenes, GGT5 and USP18, only contain segments of the FAM247 sequence whereas the FAM247 lincRNA genes contain the full sequence. The USP18 protein carboxyl end sequence is crucial for the regulation of the immune system.

An evolutionarily conserved sequence forms part of diverse genes and appears to serve as a nucleation site for the development of new genes during evolution.

DISCOVERY OF AN EVOLUTIONARY CONSERVED NUCLEATION SITE

There are five human FAM247 lincRNA family genes that were recently formed in humans and this gene family appears to have originated by gene sequence duplication of the FAM247 sequence, a sequence that is 11,231 bp in length. This conforms with the established model for gene family origins via gene duplications. However, other genes, pseudogenes and protein genes were found to contain segments of the FAM247 lincRNA family sequence *(see figure 1)*. The pseudogenes are of particular interest as they are unique and contain extraneous chromosomal sequences unrelated to FAM247 or the parent gene of origin; thus, they significantly differ from true pseudogenes.

Pseudogenes are defined as genes that have copies of sequences of a protein gene but cannot form a protein product because of key mutations within the sequence.

A model of gene formation involving these so-called pseudogenes has been presented by Dr Delihas whereby the FAM247 fragment serves as a nucleation element or a foundation site. Other sequence blocks from other parts of the genome are added to the FAM247 sequence to form the mature gene *(see figure 2)*. These FAM247 fragments are in genomic regions that display evolutionarily conserved sequence signatures. The FAM247 sequence appears to carry the information for the attachment of these extraneous sequences and it represents a focal point for the de novo genesis of this and other pseudogenes. These pseudogenes display RNA transcript expression, and several in a very broad and robust manner in various tissues. Functions of these genes are not known. The determination of gene function is important, however, to assess the degree of relevance of these pseudogenes to cellular molecular processes.

Surprisingly, two ancient protein genes, gamma-glutamyltransferase (GGT5) and ubiquitin specific peptidase 18 USP18 also found to contain segments of the FAM247 sequence *(see figure 1)*. Both genes date back from one hundred to several hundred million years in evolutionary time. Thus, the FAM247 sequence has formed a part of diverse genes through much of vertebrate evolution. Unlike the modified pseudogenes that are formed

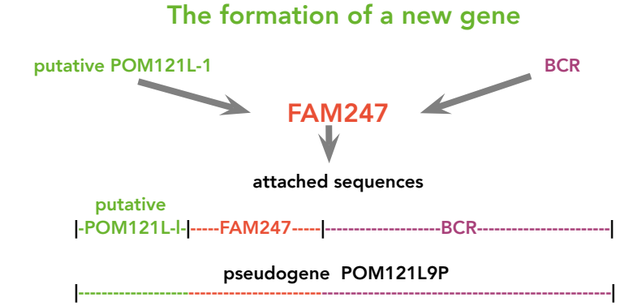


Fig 2. The birth of gene POM121L9P can be visualised by the addition of sequences from parts of other genes termed putative POM121L-1 and BCR to the FAM247 sequence. The FAM247 sequence is situated at a human chromosomal site that displays sequence synteny with the comparable chromosomal site found in the chimpanzee and represents a nucleation site for gene formation. POM121L1 is a POM121-like protein 1 gene. POM121 is a membrane component of the vertebrate nuclear pore complex. The POM121L1 sequence is attached on the left (5′) side of the FAM247 sequence. A section of the BCR gene, termed BCR activator of RhoGEF and GTPase is on the right (3′) side of FAM247.

www.researchoutreach.org

Kazakov Maksim/Shutterstock.com

The FAM247 sequence, or part of it, was present in the zebrafish USP18 gene. This may imply that certain functions of the ubiquitin specific peptidase USP18 originated in vertebrates several hundred million years ago.

Photo Credit, left image: Wiki Commons
Photo Credit, right image: Lefteris Papaulakis/Shutterstock.com

Marble statues of Phrasikleia Kore (left, c. 550–540 BC) and Anavysos (Kroisos) Kouros (right, c. 530 BC). Long non-coding RNA genes formed from FAM247 nucleation sequences in humans.

de novo and the five human FAM247 lincRNA family genes formed by gene duplication, the mechanism that led to the formation of the protein genes in vertebrate ancestors remains unclear.

AN EVOLUTIONARY HIGHLY CONSERVED SEQUENCE PROVIDED BY FAM247 IS ESSENTIAL FOR THE REGULATION OF THE IMMUNE SYSTEM

A major highlight of these studies is the important role of FAM247 in providing the amino acid sequence essential for the USP18 gene to function in the regulation of the immune system. USP18 is the ubiquitin specific peptidase gene, a member of the deubiquitinating protease family. Many proteins are modified by the addition of ubiquitin. When attached to proteins, it acts as a signalling molecule that controls cell differentiation and has also a role in the regulation of the immune system. Deubiquinating proteases like USP18 remove ubiquitin from proteins regulating their functions by switching off the signals acquired upon ubiquitination.

This small genomic DNA sequence appears to carry information for the attachment of other genomic segments, which results in the formation of new genes. It also carries information in its open reading frames that form protein exons.

Near D Krasaesorn/Shutterstock.com

Scientists routinely survey the genetic architecture of human and primate populations to find out the specific roles that different genes play in evolution.

There is a very high identity between the FAM247 nucleotide sequence and the human/primate carboxy terminal end exon nucleotide sequence of USP18, in most cases 100% identity. In zebrafish, a comparison of amino acid sequence shows a small number of amino acid changes in the terminal exon compared to the human USP18 terminal exon and the FAM247 translated sequence, but most amino acids are unaltered. The high number of phylogenetically conserved amino acid positions in the terminal exon, for one, suggests that the FAM247 sequence, or part of it, was present in the zebrafish USP18 gene. However important to protein function, that much of the amino acid sequence of the last exon is evolutionarily conserved may imply that certain functions of the ubiquitin specific peptidase USP18 originated in vertebrates several hundred million years ago. In a series of papers, Dr Dong-Er Zhang at the University of California San Diego and colleagues at other institutions reported the elucidation of functions of the carboxy end of the USP18 protein. These functions are of major significance because of the influence they have in regulating interferon. Interferon is released in response to viral infections and is responsible for an inflammatory cascade that heightens anti-viral defenses. USP18 is a negative regulator of interferon-signalling, acting as an anti-inflammatory protein. In addition, USP18 has an important role in pathogen control and in the development of autoimmune diseases. This protein could therefore act as an important pharmacological target in a variety of diseases linked to inflammation.

Behind the Research

Dr Nicholas Delihas

E: nicholas.delihas@stonybrook.edu **T:** +1 631 286 9427
W: https://orcid.org/0000-0002-1704-2587

Research Objectives

Prof Delihas studies long non-coding RNA genes and their evolutionary development.

Detail

Nicholas Delihas
Department of Micobiology and Immunology
Life Sciences Building
Stony Brook University
Stony Brook, NY USA 11794

Bio

Nicholas Delihas received his PhD in Molecular Biophysics and Biochemistry from Yale University in 1961, then held a post-doctoral fellowship position at the Sloan Kettering Institute in New York for 2 years. In 1964/5 he was awarded a National Institutes of Health Special Post-Doctoral Fellowship to study RNA chemistry at the CIBA Research Labs (RNA Lab of Mattys Staehelin) in Basel, Switzerland. He continued his RNA research at the Brookhaven National Laboratory in New York. From 1971-present he has been a faculty member of the Department Immunology and Microbiology, Renaissance School of Medicine Stony Brook, and currently is Professor Emeritus. Dr Delihas has previously served as Associate Dean for Basic Sciences at the Renaissance School of Medicine for 9 years. He currently is on the editorial boards of the *International Journal of Molecular Sciences* and the journal *Non-Coding RNA*.

Renaissance School *of* Medicine
Stony Brook University

References

Delihas, N. (2020). Genesis of Non-Coding RNA Genes in Human Chromosome 22 – A Sequence Connection with Protein Genes Separated by Evolutionary Time. *Non-Coding RNA*, 6(3), 36. Available at: https://doi.org/10.3390/ncrna6030036

Delihas, N. (2020). Formation of human long intergenic non-coding RNA genes, pseudogenes, and protein genes: Ancestral sequences are key players. *PLoS ONE*, 15(3), e0230236. Available at: https://doi.org/10.1371/journal.pone.0230236

Carvunis, A.R. *et al.* (2012). Proto-genes and de novo gene birth. *Nature*, 487(7407), 370–4. Available at: https://www.doi.org/10.1038/nature11184

Honke, N. *et al.* (2016). Multiple functions of USP18. *Cell Death & Disease*, 7(11), e2444. Available at: https://www.doi.org/10.1038/cddis.2016.326

Personal Response

Are you currently focused on repeat sequences in chromosome 22 or is there any other project on the horizon?

"I am currently focused on finding additional repeat sequences in chromosome 22 that may serve as protogene elements."

hywards/Shutterstock.com