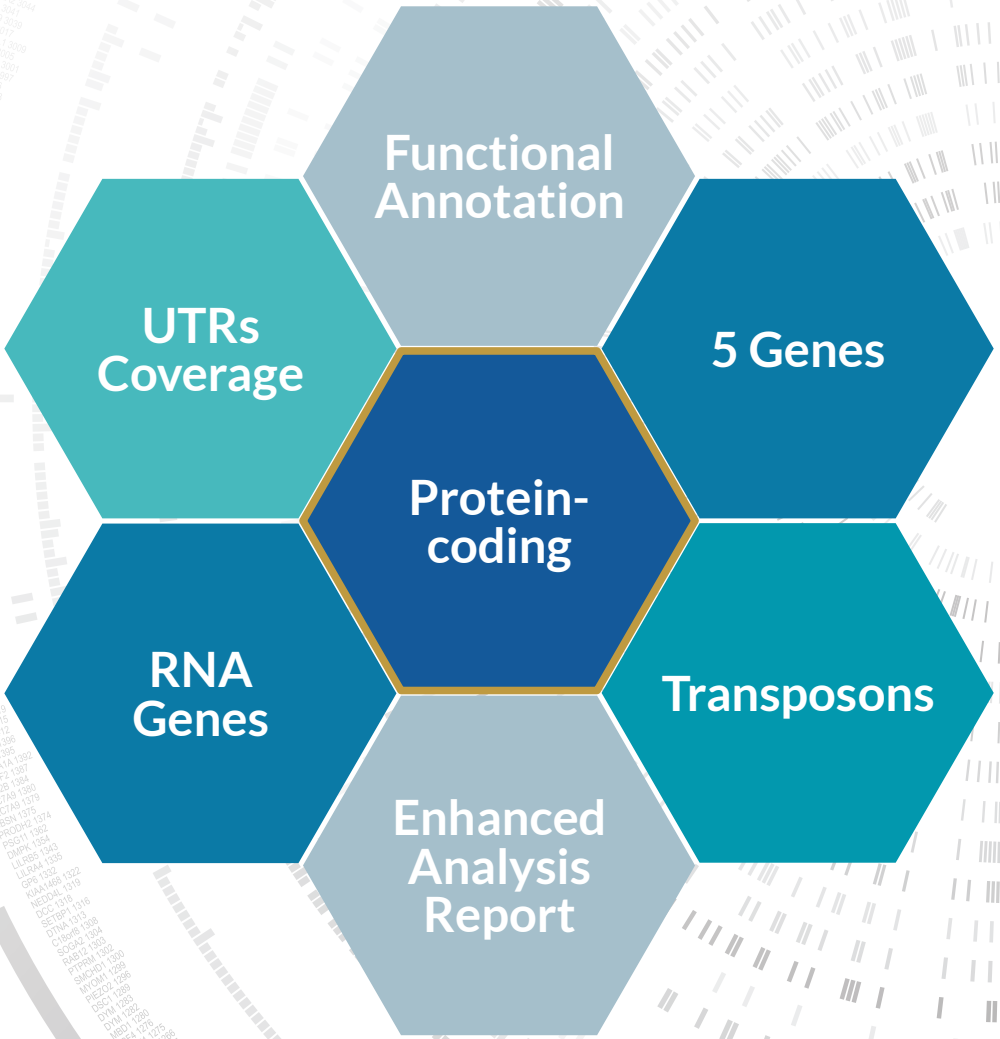
GOENOMICS

Gene categories explained

protein-coding genes, non-coding genes and pseudogenes



Gene categories explained

protein-coding genes, non-coding genes and pseudogenes

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

Protein-coding genes are traditionally classified by features like coding sequence length and exon count, with a focus on canonical genes with more than 100 codons. Short open reading frames (sORFs), which are less than 100 codons long, are often overlooked despite their potential biological significance. sORFs can be found in various genomic regions, including untranslated regions, introns, and non-coding RNAs. Detecting them requires specialized techniques like ribosome profiling and mass spectrometry. Non-coding genes, which are classified by function rather than structure, include various RNA types such as miRNAs, snRNAs, snoRNAs, tRNAs, rRNAs, piRNAs, circRNAs, eRNAs, lncRNAs, and antisense RNAs. Pseudogenes, categorized as processed, non-processed, or unitary, are mutated, non-functional copies of genes but may have regulatory roles. Errors in gene annotation are common, affecting around 20–30% of public database entries.

Protein-coding gene categories

In the annotation of genomes, the group of protein-coding genes has always been defined by measurable features such as the length of the coding sequence, the number of segments containing coding sequences and the number of alternatively spliced transcripts. For historical reasons, the length of the coding sequence proved to be the most important classification feature. In the early days of the development of gene prediction methods, the prediction of genes in pieces was extremely difficult, and most methods relied on the prediction and evaluation of the longest open reading frames across all six frame translations. To reduce the number of false positives, a limit of one hundred codons was set, although it was clear that hundreds to thousands of shorter genes were likely to be missed

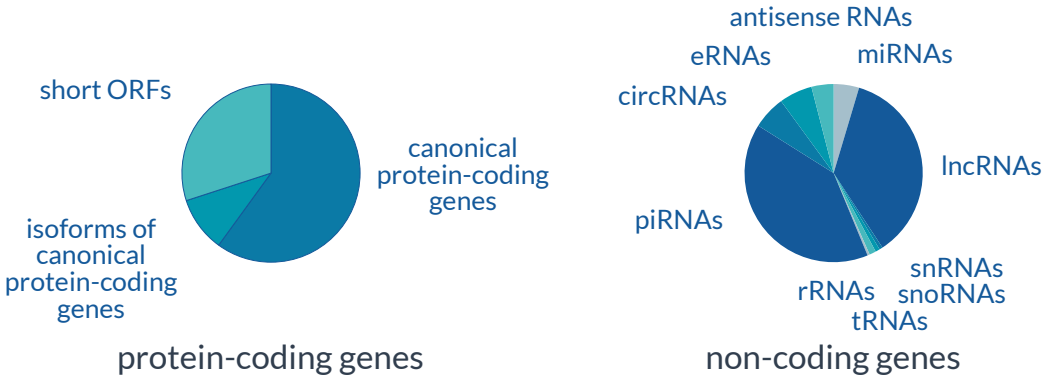


Figure 1: Common classification of protein-coding genes (left) and non-coding genes (right). The segments do not represent real numbers, but are intended to highlight the estimated proportions of gene types for the human genome.

Version 2024.11.11

(Basrai *et al*, 1997). It was simply not possible to distinguish the real positives from the false positives.

Accordingly, there is the large group of canonical protein-coding genes with a length of usually more than one hundred codons and the rather unexplored group of short open reading frames (sORFs), also known as small ORFs (Figure 1). Some of the canonical protein-coding genes encode alternatively spliced isoforms of these genes, which lead to different transcripts and, as a result, sometimes to different translations.

sORFs generally encode peptides of fewer than hundred amino acids. They can be found in various genomic regions, including the 5' and 3' untranslated regions of canonical protein-coding genes, they can overlap and be out-of-frame with the canonical protein-coding genes, they can be in introns or intergenic regions, and they can overlap or be translations of non-coding RNAs including long non-coding RNAs (lncRNAs). The translation of genes previously classified as pseudogenes has also been demonstrated. Many sORFs are highly conserved across species, suggesting important biological functions. Since their identification using standard gene prediction algorithms is difficult, ribosome profiles and mass spectrometry are often used for their detection (Hanada *et al*, 2013; Chen *et al*, 2020; Pennisi, 2024; Schlesinger *et al*, 2025). Despite their size, sORFs can have a variety of biological functions, including cell signalling and regulation of gene expression, modulation of protein-protein interactions, response to stress conditions, regulation of metabolic pathways and function as structural components of protein complexes.

These results show the problems of classifying gene types solely on the basis of parameters such as length and number of exons. If translation is proven, there is no reason not to label these genes and gene regions as protein-coding genes independent of their lengths. Currently, any gene with an ORF shorter than the cut-off value is either ignored or labelled as non-coding. If the sORFs were labelled as protein-coding in a similar way to the canonical ones, the number of protein-coding genes per organism would increase considerably. In reality, there are overlaps and intersections of genes of all categories, canonical protein-coding genes, sORFs, all types of non-coding genes (see below) and pseudogenes in many genomic regions. There is consensus on the overlap of RNA genes such as tRNAs and miRNAs with canonical protein-coding genes and the overlap of transposons, mobile elements and viral relics, but there is a lack of literature and consensus on the overlap of canonical protein-coding genes with other types of genes, other protein-coding genes, non-coding genes, especially from the lncRNA class, and pseudogenes.

Protein-coding genes would be classified completely differently on the basis of function, structure or evolutionary origin. Classes would include enzymes, structural proteins, transport proteins, signalling proteins, regulatory proteins, receptor proteins, motor and contractile proteins, immune system proteins, chaperones, storage proteins, antifreeze proteins, toxins and defense proteins or adhesion proteins. There are many possibilities for different groupings.

non-coding genes

In contrast to protein-coding genes, non-coding genes have never been classified by measurable numbers such as gene lengths or number of gene structure features (Figure 1). Non-coding genes have always been classified by their functions. The main types of non-coding genes are miRNAs, snRNAs, snoRNAs, tRNAs, rRNAs, piRNAs, circRNAs, eRNAs, lncRNAs, and antisense RNAs.

Processed miRNAs usually have lengths of about 22 bp. miRNA precursors are considerably longer. Most miRNAs function in post-transcriptional gene regulation. There are about 2,300 miRNAs in the human genome.

Small nuclear RNAs (snRNAs) comprise the components of the spliceosome that does the pre-mRNA splicing. Their combining feature is their inclusion in one of the many complexes that build and re-build during the splicing process. But there is no homology between the various snRNA subtypes. The human genome contains 200-300 snRNAs.

Small nucleolar RNAs (snoRNAs) are involved in the processing and modification of the ribosomal

RNAs. The human genome contains about 400 snoRNAs.

Transfer RNAs (tRNAs) carry amino acids to the ribosome and are essential for translation. There are dedicated tRNAs for most of the 61 codons coding for amino acids. There are 600 – 650 tRNAs in the human genome including tRNA-like pseudogenes.

Ribosomal RNAs (rRNAs) are the structural and catalytic components of ribosomes. There are four classes of rRNAs, 5S, 5.8S, 18S, and 28S rRNAs, and all of them are present in multiple copies in each genome.

Piwi-interacting RNAs (piRNAs) are involved in the silencing of transposons and other genetic elements, particularly in the germline. piRNAs are usually not well characterised, and estimates for the human genome range up to tens of thousands of loci that generate piRNAs.

Circular RNAs (circRNAs) have closed-loop structures formed by back-splicing, and are involved in regulatory functions and potential sponging of miRNAs. The human genome likely contains thousands, which are highly variable and often generated from exonic or intronic sequences.

Enhancer RNAs (eRNAs) are non-coding RNAs transcribed from enhancer regions of the genome. There number is difficult to quantify, but thousands of enhancers are known to produce eRNAs in the human genome.

Long non-coding RNAs (lncRNAs) are a group of RNAs longer than 200 nucleotides that regulate gene expression at different levels (transcriptional, post-transcriptional, epigenetic). The length of the cut-off is rather arbitrary in order not to include too many false-positive cases. The human genome contains 18,000 to 20,000 lncRNAs, most of which have not yet been characterised.

Antisense RNAs are transcripts that overlap protein-coding or non-coding genes in opposite directions and thus influence gene expression. There are thousands of them in the human genome, particularly in the lncRNA category.

pseudogenes

Pseudogenes are non-functional, mutated copies of protein-coding genes or functional RNA genes that have lost their original coding ability or biological function. They are often regarded as genetic fossils, but some may still have regulatory functions. Pseudogenes are generally categorised into three main categories: processed pseudogenes, non-processed pseudogenes and unitary pseudogenes. As with all categories, there is a considerable grey area, such as the partially processed pseudogenes.

Processed pseudogenes (or retrotransposed pseudogenes) result from the reverse transcription of mRNA and the integration of the cDNA copy into a new genomic position. These pseudogenes lack introns because they are derived from spliced mRNA, are often flanked by direct repeats generated during integration, and have poly-A tails or remnants thereof. One example is the PTENP1 pseudogene, which can regulate the PTEN tumor suppressor gene by acting as a miRNA sponge.

Non-processed pseudogenes are created by gene duplication, followed by mutation and loss of function. In contrast to processed pseudogenes, they retain exon-intron structures similar to those of their parent genes. However, they often contain frameshift mutations, premature stop codons or deletions that impair their coding potential. Non-processed pseudogenes can remain in close proximity to the original gene or be relocated to other regions. Examples of this are the pseudogenes of the olfactory receptor genes, which are numerous in the human genome.

Unitary pseudogenes result from the accumulation of mutations within a single-copy gene, causing the gene to lose its function without duplication or retrotransposition. These pseudogenes are usually species-specific and reflect evolutionary gene loss. One example is GULOP (L-gulono-γ-lactone oxidase), a pseudogene that is responsible for the inability to synthesise vitamin C in humans.

In genome annotation, genes from families of orthologous genes are often labelled as pseudogenes if they have a less complex gene structure (e.g. fewer exons) than the most

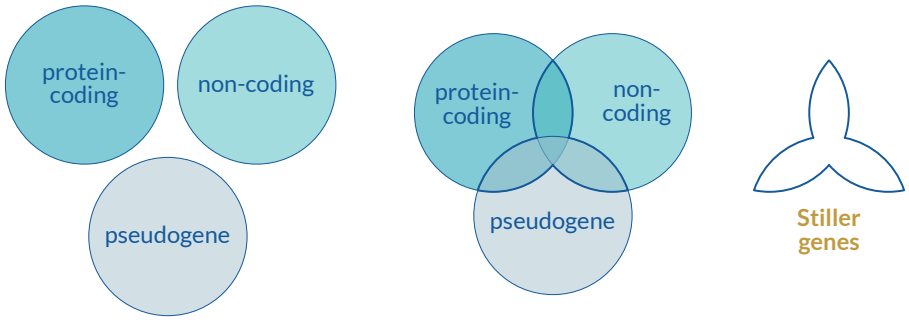


Figure 2: In real world annotations, there is a grey area of genomic regions that are either annotated as protein-coding, non-coding or pseudogenic, although the category could be completely wrong (e.g. due to applied length cut-offs) or the regions could have ambiguous/dual functions.

complex family member or if they have shorter ORFs, e.g. due to an earlier stop codon. There is usually no verification of the accuracy of the genome assembly in the respective region, which could also be the reason for function-impairing mutations. There is also a huge knowledge gap about pseudogenes of non-coding genes. For tRNAs, for example, it is known that there are tens of thousands of gene copies in some mammals and other animals, the majority of which must be pseudogenes. The same is probably true for the other non-coding gene groups, but the identification of function-impairing mutations in non-coding transcripts is complicated.

Errors in genome and gene annotations

All genome annotations are provisional. This applies to complex genomes such as animal genomes (including the human genome) and plant genomes as well as simpler genomes such as those of yeasts and bacteria. Without experimental evidence, any gene annotation remains a prediction. For most animal, plant and fungal/yeast gene loci, there is now good evidence based on RNA-Seq data and evolutionary conservation/comparative genomics. Although the gene loci are known, there is less evidence for the individual exons of each gene. Due to the many errors in the annotations in public databases, these errors are transferred to the annotations of the next genomes in evidence-based annotation approaches. Studies have shown that about 40-50% of gene annotations in public databases contain deletions, insertions and erroneous segments where part of the correct protein sequence is replaced by an alternative, erroneous sequence compared to the best annotated reference (Meyer *et al*, 2020). Aberrant splicing leads to massive misannotations of putative alternative isoforms. This led to a high number of isoforms for the human genome in the RefSeq annotation and GENCODE annotation, although only a small number of these isoforms are supported by biochemical and high-throughput proteomics data (Tress *et al*, 2017).

Overlap and intersection of gene categories

In an ideal world, every region in a genome is assigned to a certain category of genes (protein-coding, non-coding, pseudogene) or repeat/transposon type (Figure 2; repeat regions and transposons are not considered genes and will be discussed elsewhere). In real world annotations, many regions do not get annotated at all and for many other regions the annotation category remains unclear. For example, sORF genes are usually not annotated and their loci are accordingly missing in the annotations or are annotated as a different category, e.g. non-coding (lncRNA). Many tRNA genes are annotated as real tRNAs, although the corresponding genome regions could rather represent pseudogenes. This is true for the tens of thousands of identified tRNA gene regions in animals and plants and also for most, if not all, tRNAs where the anticodon triplet does not match the tRNA type as determined by the set of tRNA discriminator nucleotides (so-called non-cognate tRNAs. We propose to call these genes in the grey area of categorisation Stiller genes, in reference to the famous book by Max Frisch.

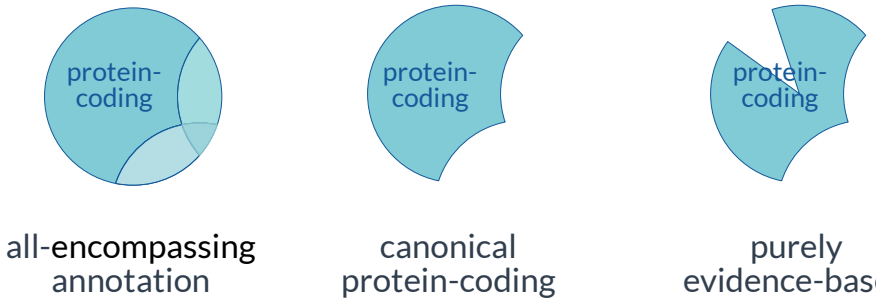


Figure 3: Different annotation approaches for protein-coding genes include/exclude part of the genes. When focusing on the canonical protein-coding genes, the short open reading frames (sORFs) are usually completely ignored. The regions encoding sORFs are usually annotated as non-coding or pseudogene, if annotated at all. Purely evidence-based approaches are not able to identify and annotate regions, for which expression data (e.g. RNA-Seq or EST) is not available and which lack high homology to known genes in public databases.

When evaluating and comparing genome annotations, it is important to consider the various limitations of current software for identifying and annotating gene regions. Different total numbers of genes do not necessarily indicate false-positive annotations, but may be the result of a different approach (Figure 3). It depends on the user's preference whether a subset of the canonical protein-coding genes (e.g. as determined by a purely evidence-based approach), the full set of canonical protein-coding genes or even an all-encompassing set is best suited for the intended analyses.

References

Basrai MA, Hieter P & Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res* 7: 768–771

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, *et al* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science* 367: 1140–1146

Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, *et al* (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* 110: 2395–2400

Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O & Thompson JD (2020) Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. *BMC Bioinformatics* 21: 513

Pennisi E (2024) 'Dark proteome' survey reveals thousands of new human genes. *Science* 386: 951–952

Schlesinger D, Dirks C, Navarro C, Lafranchi L, Spinner A, Raja GL, Tong GM-S, Eirich J, Martinez TF & Elsässer SJ (2025) A large-scale sORF screen identifies putative microproteins involved in cancer cell fitness. *iScience* 28

Tress ML, Abascal F & Valencia A (2017) Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends Biochem Sci* 42: 98–110

CEO
Dr. Martin Kollmar
kollmar@goenomics.com

CTO
Dr. Dominic Simm
simm@goenomics.com

GOENOMICS GmbH
Benfeyweg 9
37075 Göttingen
Germany

Version 2024.11.11