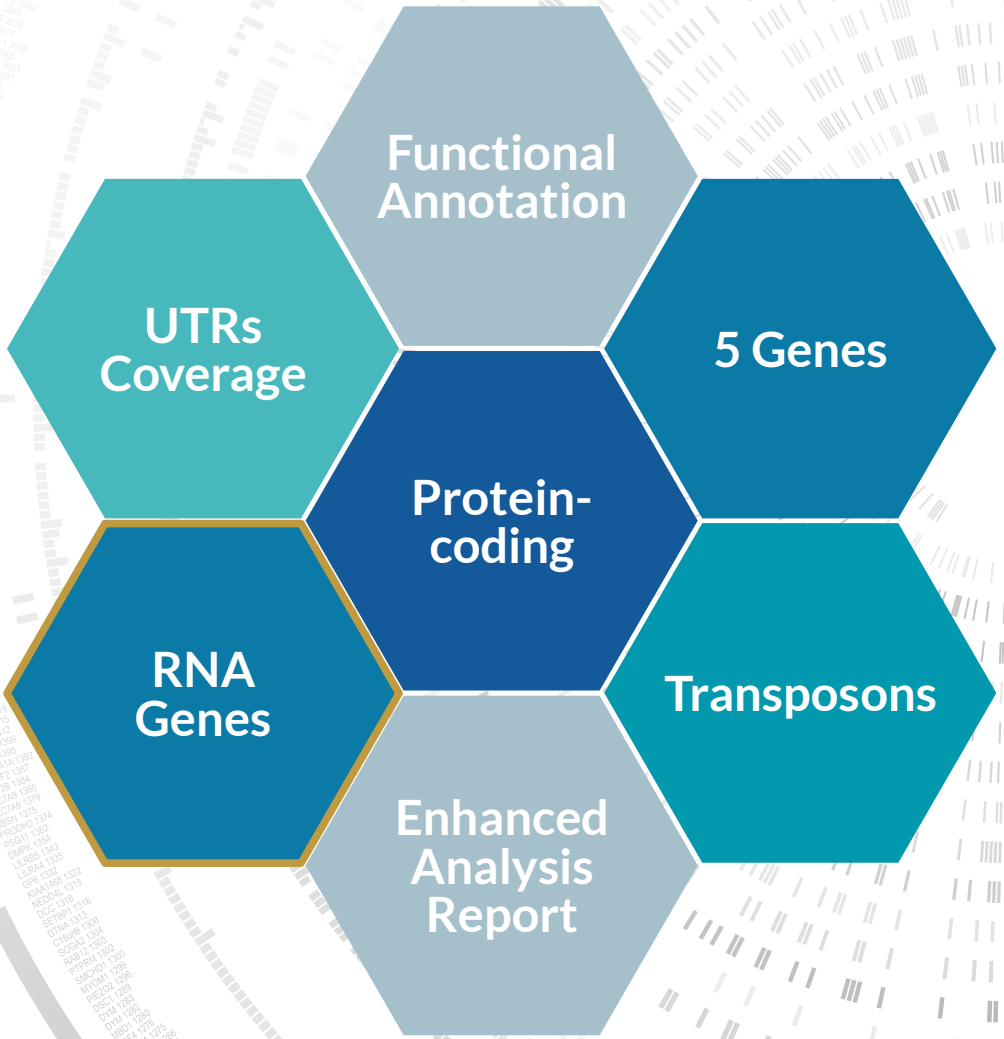
GENOMICS

RNA genes

identification and classification of ribosomal, spliceosomal, RNase-P and SRP genes, snoRNAs, miRNAs, telomerase, and more



RNA genes

identification and classification of ribosomal, spliceosomal, RNase-P and SRP genes, snoRNAs, miRNAs, telomerase, and more

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

RNA genes encode RNA molecules that perform various cellular functions. Important RNA types include tRNA (transfer RNA) and rRNA (ribosomal RNA), both of which are essential in the process of translation. Additionally, non-coding RNAs like microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) regulate gene expression by controlling mRNA stability and translation.

RNA gene identification and classification

The most accurate method for predicting RNA genes is to use the covariance models from the Rfam database. The program cmscan from the infernal package identifies and annotates alignments that match the covariance models on both strands of the genome sequence. For many RNA gene types there are covariance models for different taxonomic branches, which are grouped into so-called clans. The disadvantage of this approach is the covariance models and, as a consequence, the hundreds of CPU hours for plant and animal genomes. To reduce the number of false positives and overlapping annotations, Ensemble and other large genome databases apply numerous output filters, such as filters for non-gene biotypes, taxon-specific thresholds (e.g. for divisions), filters for overlapping hits (hits from different covariance models and overlaps with e.g. protein-coding exons) and the removal of partial hits and hits in regions with skewed nucleotide distribution (e.g. regions with high GC or AT content). The classification of the predicted RNA genes is given by the name of the covariance model.

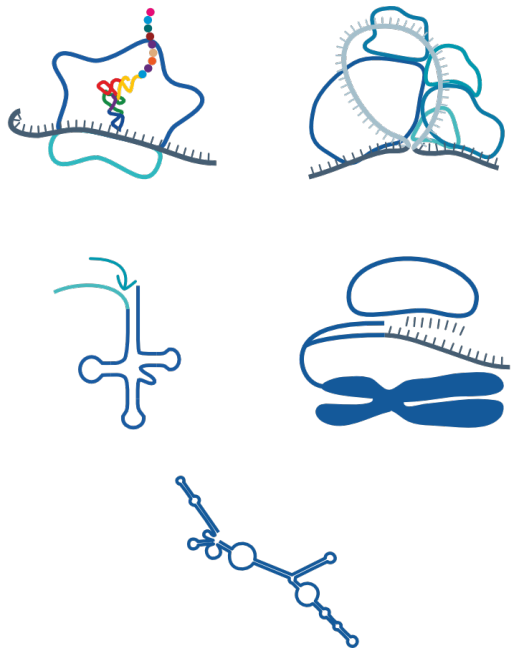


Figure 1: RNA genes conserved throughout the eukaryotic tree of life. The ribosome contains four RNA genes (top left), the spliceosome contains five (top right), RNase-P (middle left), telomerase (middle right) and snoRNA U3 are independently functioning molecules.

Ribosomal RNA genes

In eukaryotic ribosomes, ribosomal RNA (rRNA) genes play a critical role in ribosome structure and function. These genes encode the rRNAs that form the ribosome's core and catalyze protein synthesis. The ribosome comprises two

Version 2024.11.11

subunits: the large (60S) and the small (40S). The 60S subunit includes the 28S, 5.8S, and 5S rRNAs, while the 40S subunit contains the 18S rRNA.

Eukaryotic rRNA genes are transcribed by RNA polymerase I (for 28S, 18S, and 5.8S rRNAs) and RNA polymerase III (for 5S rRNA). The genes for 18S, 5.8S and 28S rRNA form a single transcription unit that is transcribed into a single large RNA (45S pre-rRNA). Between the transcription units for the 45S rRNA are spacer regions.

The rRNAs are assembled with ribosomal proteins in the nucleolus to form functional ribosomes. Variations and modifications of rRNA, such as methylation and pseudouridylation, are crucial for ribosomal accuracy and efficiency. Ribosomal RNA genes are found in multiple copies in the genome, ensuring the high output required for protein synthesis in eukaryotic cells.

Spliceosomal RNA genes

The major spliceosome complexes consist of the U1, U2, U4, U5 and U6 RNA genes. U11, U12, U4atac and U6atac are the minor spliceosome analogs of U1, U2, U4 and U6 of the major spliceosome. U5 is shared by both complexes. The minor spliceosome removes introns of the rarer type (`AT---AC`, U12-type). The first identified introns of this type had the splice sites AT (5' end) and AC (3' end), but other splice sites are now known.

Ribonuclease P genes

Ribonuclease P (RNase P) is an endoribonuclease, whose best characterised activity is the cleavage of the 5'-leader elements of precursor-tRNAs to generate mature 5'-ends of tRNAs. RNase MRP is involved in precursor rRNA processing, where it cleaves the internal transcribed spacer 1 between 18S and 5.8S rRNAs.

Telomerase genes

Telomerase is a ribonucleoprotein that adds a species-dependent telomere repeat sequence to the 3' end of telomeres.

snoRNA genes

The most prominent and wildly distributed type of snoRNAs is the U3 snoRNA gene. U3 snoRNA is predominantly found in the nucleolus and is thought to guide site-specific cleavage of ribosomal RNA (rRNA) during pre-rRNA processing.

miRNA genes

miRNA genes encode small, non-coding RNAs (~21-23 nucleotides) that regulate gene expression post-transcriptionally. miRNA genes are

SPLICEOSOMAL RNA GENES

MAJOR SPLICEOSOME

U1	U2	U4	U5	U6	
21	26	4	18	35	complete
0	0	0	0	0	incomplete

MINOR (U12-TYPE) SPLICEOSOME

U11	U12	U4ATAC	U6ATAC	
1	3	0	3	complete
0	0	0	0	incomplete

Figure 2: Annotation of the RNA genes of the spliceosome complexes of a plant genome. The covariance model of the U4ATAC gene is too restrictive for plants, so that these genes are rarely detected in plant genomes.

found in various genomic contexts: within introns of protein-coding genes (mirtrons), intergenic regions, or exons of non-coding RNAs. Some miRNAs exist as single genes, while others are clustered, allowing their co-transcription.

The processing of miRNA involves multiple steps, with distinct precursors at each stage: 1. The initial transcript is the primary miRNA (pri-miRNA), which may span hundreds to thousands of nucleotides. Pri-miRNAs contain one or more stem-loop structures where the miRNA sequences are embedded. 2. The pri-miRNA is processed in the nucleus by the Drosha-DGCR8 complex. Drosha cleaves the pri-miRNA at the stem-loop base, releasing a shorter hairpin-shaped pre-miRNA (~70 nucleotides). 3. The pre-miRNA is exported to the cytoplasm and further cleaved by Dicer, generating the mature ~22-nucleotide miRNA duplex.

Signal recognition particle

The signal recognition particle (SRP) is a ribonucleoprotein complex essential for targeting specific proteins to the endoplasmic reticulum in eukaryotes. The SRP includes both protein components and a crucial RNA molecule encoded by SRP RNA genes. The RNA molecule in the SRP is known as 7SL RNA in eukaryotes.

mendle-analytics RNA gene annotation

The annotation of tRNA genes is performed with tRNAscan-SE (see Technote tRNA genes). For the annotation of other RNA genes, covariance models of biologically well-characterized RNA gene families were selected manually regardless of their taxonomic restriction. Cmscan is run with these models instead of all Rfam models, which significantly reduces CPU runtime. In addition, this approach ensures that false-positive hits can be virtually eliminated. Overlapping hits are filtered out. This approach helps in detecting potential genome contamination when matches are found with bacterial or archaeal models, and provides some model flexibility in case genomes encode very different gene homologs.