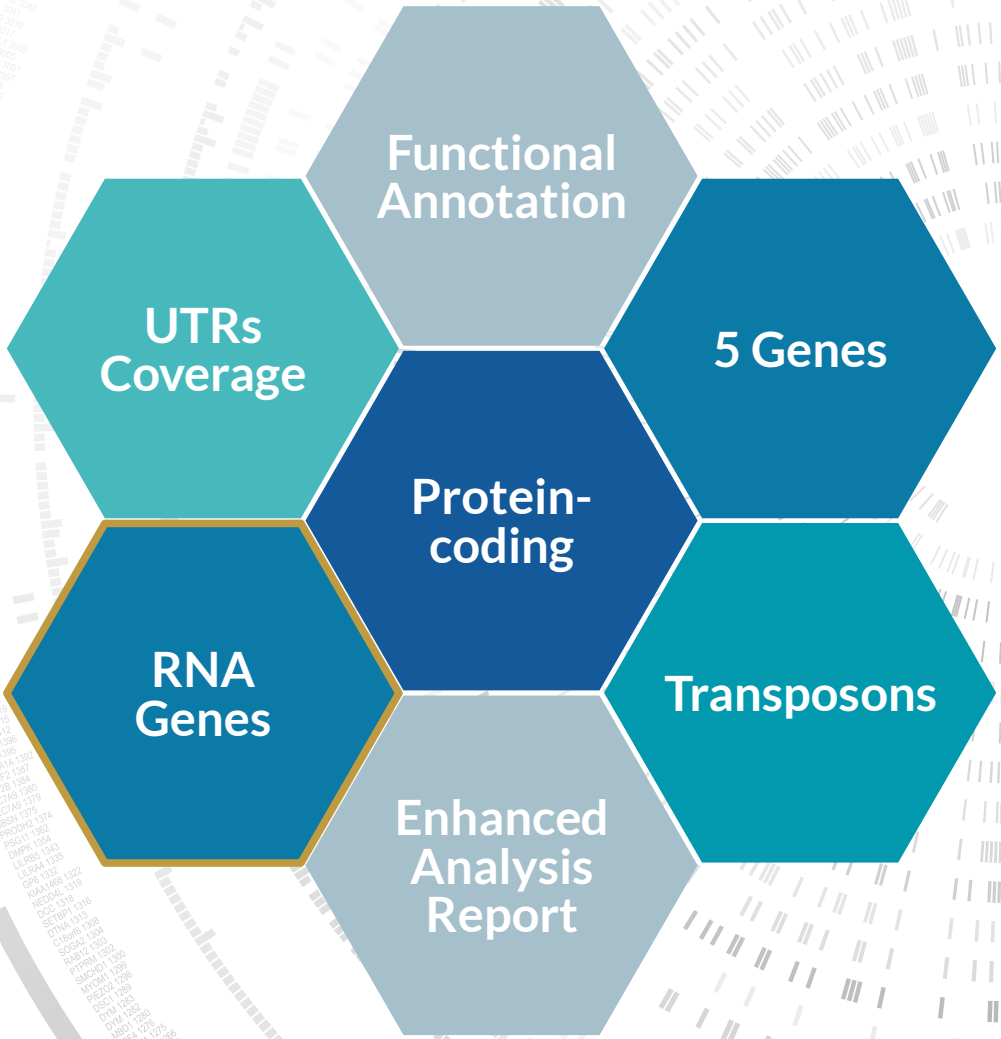
GOENOMICS

tRNA genes

identification, classification, and decoding mRNA



tRNA genes

identification, classification, and decoding mRNA

mendle TECHNOTE

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

tRNA (transfer RNA) genes encode RNA molecules that play a key role in protein synthesis. Each tRNA molecule carries a specific amino acid and recognizes codons on the mRNA through its anticodon sequence. During translation, tRNAs deliver the correct amino acids to the ribosome, where the amino acids are linked together to form a protein. The anticodon-codon pairing ensures that the amino acids are added in the proper sequence according to the mRNA template. Thus, tRNA genes are essential for translating genetic information into functional proteins.

tRNA identification

Currently, the most accurate tool for identifying tRNAs is tRNAscan-SE. tRNAscan-SE outputs a table with a tRNA classification based on the predicted anticodon and a classification based on the isotype covariance model with the highest score. In addition, several scores are provided

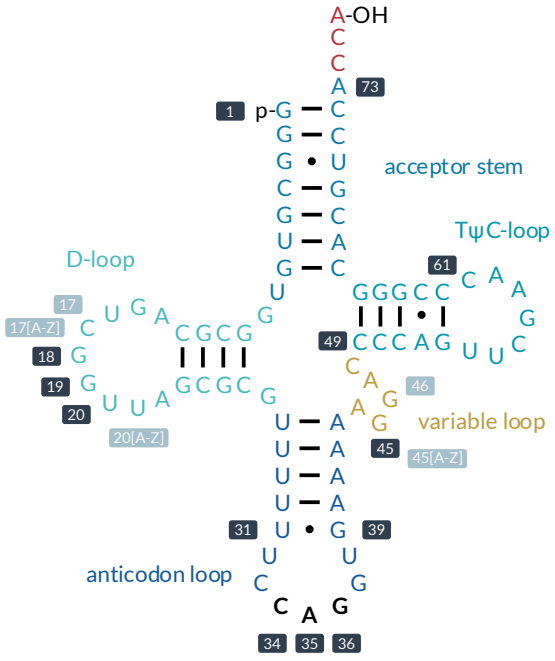


Figure 1: Cloverleaf scheme of a tRNA molecule demonstrating the official nucleotide numbering (dark-grey boxes). Light-grey boxes represent optional nucleotides.

that are calculated as part of the tRNA identification process: the infernal score, the HMM score, the secondary structure score and the isotype score. tRNAscan-SE classifies many tRNA regions derived from SINEs (short interspersed repeated elements) as pseudo-genes. Although there are updated covariance models in the latest version of tRNAscan-SE, for eukaryotic genomes the number of mismatches between anticodon and isotype model increases with the number of recognized tRNA genes. The reason for this behavior is the process of generating the covariance models, in which the 50 tRNAs with the highest score per isotype and genome were selected for the training set. Many eukaryotes have dozens to hundreds of specific tRNA isotypes, from which the same ones were always selected, making the covariance models too stringent.

tRNA classification

A manual review of all questionable classifications of tRNAs is not possible. Building better covariance models would require a statistically balanced inclusion of dozens, if not hundreds, of genomes per taxon. Taxa are terms used to classify groups of organisms based on shared characteristics and represent units in the hierarchy of species classification. Taxa do not contain quantitative information, such as the number of species

Version 2024.11.11

and sub-taxa included or the time at which the respective taxon originated on earth. Therefore, it is very difficult to obtain a representative representation of taxa. In addition, genome assemblies may contain mixtures of nuclear and organellar genomes and may have surprises in terms of tRNA and other gene content. To maximize the predictive power of tRNAscan-SE, we recommend identifying tRNAs with most of the offered search modes and classifying tRNA regions based on parsimony. For example, a tRNA region that shows a mismatch between anticodon and isotype model in one mode (e.g. eukaryotic) may have a clear classification in another mode (e.g. bacterial).

mendle-analytics tRNA annotation

For mendle-analytics, we follow the described process of classification based on parsimony. Remaining tRNA regions with mismatch of anticodon and isotype model are annotated as non-cognate tRNAs. tRNA regions with the tRNAscan-SE annotation "pseudo" are annotated as pseudogenes.

tRNA genes and the translation of mRNA information

A separate tRNA for each of the 61 amino acids would come at a significant cost for maintaining each tRNA gene and minimizing translation errors. Instead, the recognition accuracy of the third codon position is freed up to a certain extent so that a single tRNA can couple to two or more codons by wobble base pairing.

A-to-I editing (deamination of A34 to inosine) occurs in all eukaryotes. The tRNAs with INN anticodons generated in this way can form a base-pair with C, U and A at the third codon position. All boxes of the three and four-codon families are usually decoded by an A-to-I edited $tRNA_{INN}$ and by both $tRNA_{UNN}$ and $tRNA_{CNN}$. The exception is the glycine four-codon box, where the glycine codons are never decoded by $tRNA_{ACC}$, but by combinations of the other three tRNAs in this family box. $tRNA_{GNN}$, $tRNA_{UNN}$ and $tRNA_{CNN}$ are required to decode all two-codon sets. A $tRNA_{ANN}$ in a two-codon set would be susceptible to A-to-I editing and subsequent mistranslation of codons from the neighboring two-codon set.

Methionine tRNAs are a specialized type of tRNA that carry the amino acid methionine to the ribosome during protein synthesis. There are typically two types: initiator tRNA ($tRNA^{iMet}$ in eukaryotes and $tRNA^{fMet}$ in prokaryotes), which is responsible for recognizing the start codon (AUG) in mRNA and initiating translation, and elongator tRNA ($tRNA^{Met}$), which incorporates methionine into the growing polypeptide chain during elongation. The initiator tRNA is unique in its ability to bind directly to the ribosome's P-site, setting the stage for translation to begin. Both types of tRNAs have no common evolutionary origin, but are loaded by the same methionyl-tRNA synthetase.

Annotation of methionine tRNA genes

The annotation of methionine tRNAs in eukaryotes is not yet fully understood. Methionine tRNAs, which are annotated as $tRNA^{fMet}$ when using bacterial covariance models, are consistently annotated as elongator $tRNA^{Met}$ when using eukaryotic covariance models. The current version of tRNAscan-SE does not include a covariance model for a eukaryotic $tRNA^{Ile2}$. tRNA genes that are annotated as $tRNA^{Ile2}$ when using bacterial covariance models are annotated as all types of non-cognate $tRNA^{Met}$ when using eukaryotic covariance models. tRNAs of type $tRNA^{Ile2}$ contain CAU anticodons, but their cytidine nucleotides at position 34 are modified and become lysidine in bacteria, agmatidine in archaea,

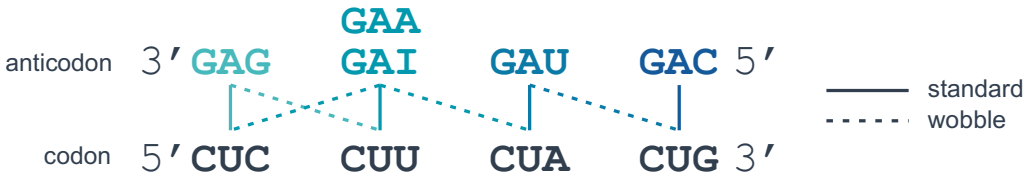


Figure 2: Base pairing between mRNA codons and tRNA anticodons. Watson-Crick base pairing is preferred, but other pairings are also possible, of which the most prominent are the wobble base pairings.

and 2-aminovaleramidine in chloroplasts. The modified $tRNA^{Ile2}$ recognizes the UAU codon. Whether eukaryotes or some of their lineages contain $tRNA^{Ile2}$, and if so, how the cytidine is modified to match adenine, remains an open question.

	T		C		A		G	
T	Phe	29	Ser	15	Tyr	26	Cys	20	T
				8					C
	Leu	7		16	stop		stop		A
		8		4			Trp	12	G
C	Leu	13	Pro	11	His	24	Arg	11	T
								5	C
		8		11	Gln	9			A
		6		6		9		4	G
A	Ile	25	Thr	11	Asn	31	Ser	17	T
		5		7					C
				15	Lys	16	Arg	10	A
	Met	38		8		18		12	G
G	Val	16	Ala	23	Asp	1	Gly		T
						111		21	C
		7		14	Glu	13		11	A
		8		8		17		8	G

38 tRNAs for ATG contain Met- and iMet-tRNAs

2 tRNAs for TGA are usually tRNAs for selenocysteine

7 tRNAs not required for decoding; absent in most genomes

1 tRNAs should be absent to avoid mistranslation; if present, tRNAs might be non-functional or immune to A-to-I editing.

Figure 3: Genetic code matrix with the numbers of cognate-tRNA genes per respective codon (nucleotide triplett). Because of efficiency reasons, some tRNAs are rarely found in four-codon-boxes (dark turquoise contour). Because of potential mistranslation through wobble base pairing, $tRNA_{ANN}$ of two-codon-boxes (decoding NNU codons) are usually absent in genomes.