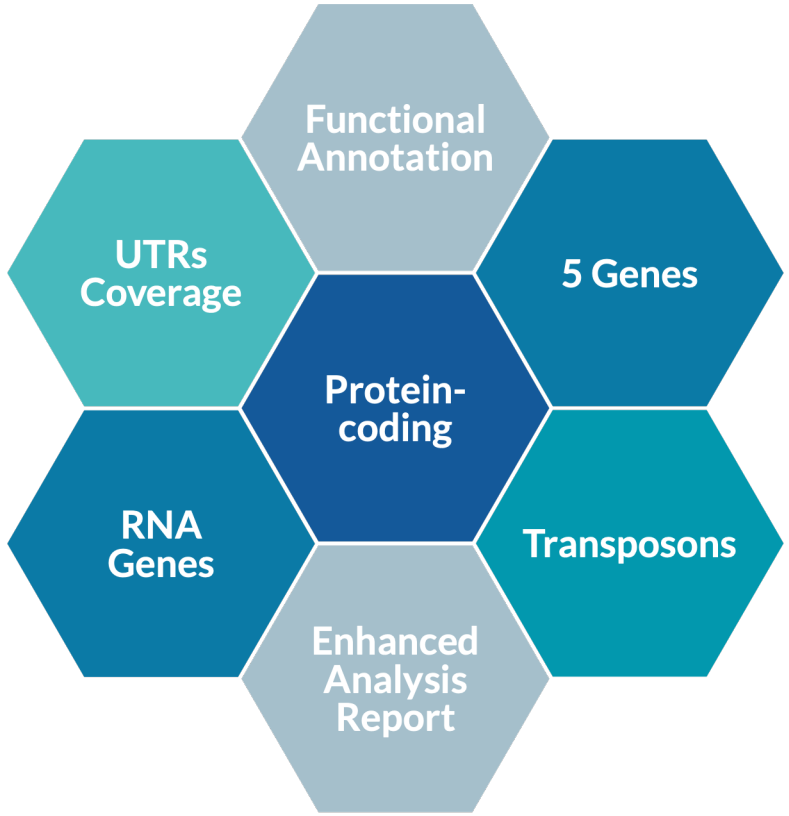
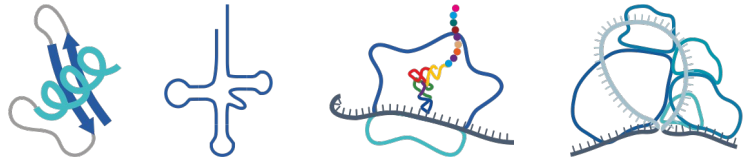
mendle® ANALYTICS



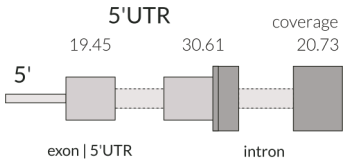
We annotate

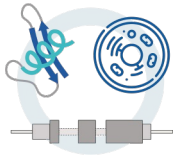
protein-coding and non-coding genes, RNA genes, transposons and pseudogenes



We assign

UTR regions, coverage and information about biological function

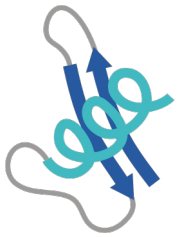




We generate

comprehensive data reports based on biological information

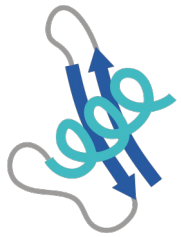
Package Protein-coding Genes



Protein-coding

File	Format	Content
protein-coding.gff	gff3	annotations of protein-coding genes, including genomic locations
protein-coding.prot.fasta	fasta	amino acid sequences of proteins encoded by the genome
protein-coding.cds.fasta	fasta	nucleotide sequences of coding DNA sequences for protein-coding genes
pseudogene.gff	gff3	annotations for pseudogenes; these pseudogenes have homology to protein-coding genes but contain frame-shifts and/or in-frame stop codons that could be due to mutations or sequencing inaccuracy
pseudogene.prot.fasta	fasta	protein sequences that are predicted from pseudogenes
non-coding.gff	gff3	annotations for long non-coding RNAs; lncRNAs are in a grey zone between protein-coding genes and erroneous transcription

Package Protein-coding Genes



Protein-coding

GFF3: attributes of an example gene in protein-coding.gff

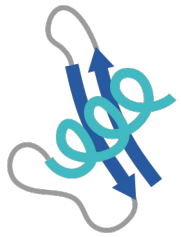
gene type: protein-coding, non-coding, pseudogene, etc.; " cov x.y"
is repeated for visualization in genome browsers

```
ID=GNX-3933;Note=protein-coding%2C cov 0.0;coverage=0.0
ID=mrna-3933;Parent=GNX-3933;coverage=0.0
Parent=mrna-3933
Parent=mrna-3933;coverage=0.0
```

coverage is filled in by the "UTRs Coverage" package

Note: Special characters in the attributes (e.g. "%2C") mask special characters for display in genome browsers.

Package Protein-coding Genes



Protein-coding

GFF3: attributes of an example gene in pseudogene.gff

protein-coding genes are marked as “potential pseudogenes“ if in-frame stop codons and/or frameshifts interrupt the CDS

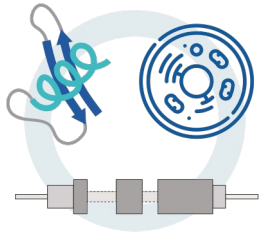
list of issues for “pseudogene“ classification with location

```
ID=GNX-17528;Note=protein-coding%2C potential pseudogene%2C cov 0.0;coverage=0.0;issues=potential
pseudogene%2C in-frame stop codons at pos 67%2C frameshifts at aa 41
ID=mrna-17528;Parent=GNX-17528;coverage=0.0
Parent=mrna-17528
Parent=mrna-17528;coverage=0.0
```

coverage is filled in by the “UTRs Coverage“ package

Note: Special characters in the attributes (e.g. “%2C“) mask special characters for display in genome browsers.

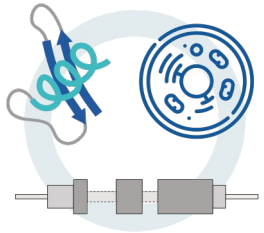
Package Functional Annotation



Functional Annotation

File	Format	Content
protein-coding.w_func.gff	gff3	annotations of protein-coding genes, including genomic locations and functional annotations
protein-coding.w_func.prot.fasta	fasta	amino acid sequences of proteins encoded by the genome, with functional annotations
protein-coding.w_func.cds.fasta	fasta	nucleotide sequences of coding DNA sequences for protein-coding genes, along with functional annotations
protein-coding.functions.xlsx	xlsx	functions related to protein-coding genes, including the closest known homolog with SwissProt-accession, species name, and taxonomy, a standardized majority consensus protein name, EC numbers, GO terms, and protein domains

Package Functional Annotation



Functional Annotation

GFF3: attributes of an example gene in .w_func GFF files

gene type: protein-coding, non-coding, pseudogene, etc.; " cov x.y" is repeated for visualization in genome browsers

protein description of first hit in result list after BLASTing against latest SwissProt DB

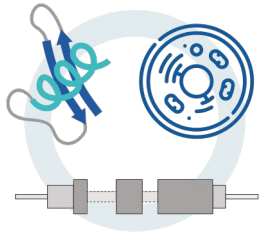
accession and species name of first hit in result list

```
ID=GNX-3933;Name=Neuronal vesicle trafficking-associated protein 2 [Q9Y328.1:Homo sapiens (Human)]; \
Note=protein-coding%2C cov 0.0;coverage=0.0
ID=mrna-3933;Parent=GNX-3933;coverage=0.0
Parent=mrna-3933
Parent=mrna-3933;coverage=0.0
```

coverage is filled in by the "UTRs Coverage" package

Note: Special characters in the attributes (e.g. "%2C") mask special characters for display in genome browsers.

Package Functional Annotation



Functional Annotation

FASTA: header of example entry

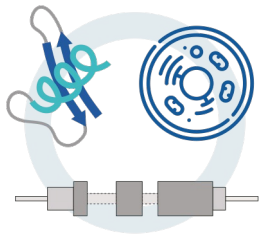
protein description of first hit in result list
after BLASTing against latest SwissProt DB

accession and species name of first hit in result list

```
>GNX-3933 Neuronal vesicle trafficking-associated protein 2 [Q9Y328.1:Homo sapiens (Human)] [organism name]
```

name of the organism to which
the cds/protein fasta belongs

Package Functional Annotation



Functional Annotation

EXCEL: annotation in tabular format

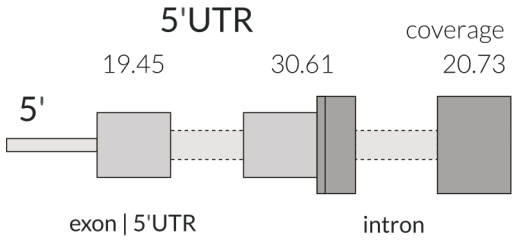
protein description of first hit in result list after BLASTing against latest SwissProt DB

organism and taxonomy of the first hit

	best_hit_acc	best_hit_desc	major_consensus	length	organism	taxonomy	ec	go	domains
GNX-2	Q9ZQH0.1	Dephospho-CoA kinase	Dephospho-CoA kinase	2257	Arabidopsis thaliana (Mouse-ear cress)	Eukaryota > Viridiplantae > Streptophyta > Embryophyta > Tracheophyta > Spermatophyta > Magnoliopsida > eudicotyledons > Gunneridae > Pentapetalae > rosids > malvids > Brassicales > Brassicaceae > Camelineae > Arabidopsis	2.7.1.24	GO:0004140, GO:0005524, GO:0005737, GO:0005739, GO:0005741, GO:0005773, GO:0005777, GO:0009507, GO:0015937, GO:0016020, GO:0016310	[{'analysis': 'Pfam', 'signature_acc': 'PF01121', 'signature_desc': 'Dephospho-CoA kinase', 'start': 3, 'stop': 182, 'interpro_acc': 'IPR001977', 'interpro_desc': 'Dephospho-CoA kinase', 'go': 'GO:0004140'}]

Standardized protein description of the majority of the first 20 hits in the result list after BLASTing against the latest SwissProt DB. Many proteins are part of large protein families (e.g. actin, tubulin, myosin) and correct classification is only possible by a thorough phylogenetic analysis. Specific names are the result of historical annotations and are very often misleading.

Package UTRs Coverage



UTR exons and coverage are added to the protein-coding and pseudogenes gff files.

GFF3: attributes of an example gene in .w_func GFF files

```
ID=GNX-3933;Name=Neuronal vesicle trafficking-associated protein 2 [Q9Y328.1:Homo sapiens (Human)]; \
Note=protein-coding%2C cov 0.0;coverage=11.7
ID=mrna-3933;Parent=GNX-3933;coverage=9.8
Parent=mrna-3933
Parent=mrna-3933;coverage=12.3
```

"gene coverage" is the mean value of all exon coverages

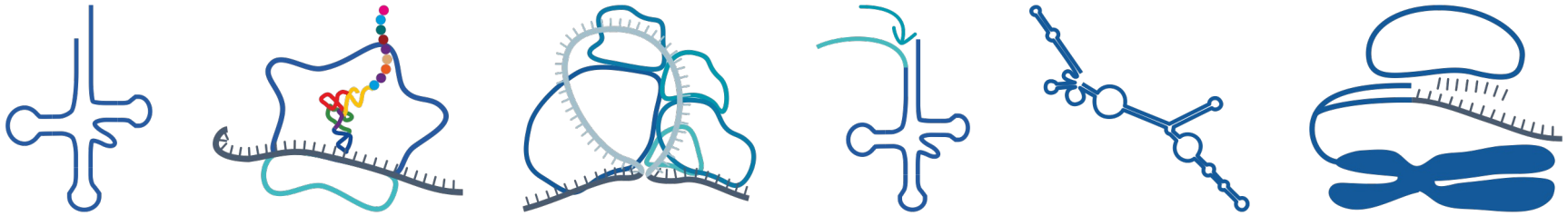
"exon coverage" is only assigned if this exon with exactly these borders is found in RNA-Seq data

coverage of gene/mRNA/exon by RNA-Seq data
The coverage found in one of the provided or used RNA-Seq data is taken without further modification. The coverage is intended as validation help and not for e.g. differential expression studies.

Note: Special characters in the attributes (e.g. "%2C") mask special characters for display in genome browsers.

Package RNA Genes

File	Format	Content
`rna_genes.gff`	gff3	annotations for tRNA, ribosomal RNA, spliceosomal RNA, telomerase, RNase P, RNase MRP and U3 snoRNA genes
`cognate_trna.gff`	gff3	annotations for cognate (=anticodon matches isotype) tRNA genes



Package RNA Genes

RNA Genes

GFF3: attributes of example genes

anticodon from 5‘ to 3‘
isotype according to tRNA HMM

note if anticodon does not match
the isotype; resulting gene could be a
pseudogene or a misdecoder

```
ID=GNX-13;Note=isotype:Gly_tRNA %28anticodon AGG%29 anticodon does not match isotype;gene_biotype=Pseudo_tRNA
ID=GNX-14;Note=isotype:Gln_tRNA %28anticodon CTG%29;gene_biotype=tRNA
ID=rna-14;Parent=GNX-14
ID=exon-14;Parent=rna-14
```

gene predictions with low score
are listed as pseudogenes

Note: Special characters in the attributes (e.g. “%28“) mask special characters for display in genome browsers.

Package Transposons

File	Format	Content
`transposon.gff`	gff3	annotations of transposon regions, including genomic locations and functional annotations (e.g. potentially non-functional)
`transposon.prot.fasta`	fasta	amino acid sequences of transposons encoded by the genome
`transposon.cds.fasta`	fasta	nucleotide sequences of coding DNA sequences for transposons

Package Data Analysis

Enhanced Analysis Report

File	Format	Content
`report.pdf`	pdf	Extensive analysis of the annotation. Highlights include analysis of intron patterns, annotation completeness based on gene homology, biological functions and domain architectures, codon usage, RNA-Seq mapping, analysis of the tRNA gene decoding potential (e.g. presence of tRNAs for all codons, absence of tRNAs leading to potential mistranslation) and evaluation for completeness of major RNA gene containing complexes (e.g. presence/absence of all components required for the major and minor spliceosome as well as the `AU--AC` subtype).