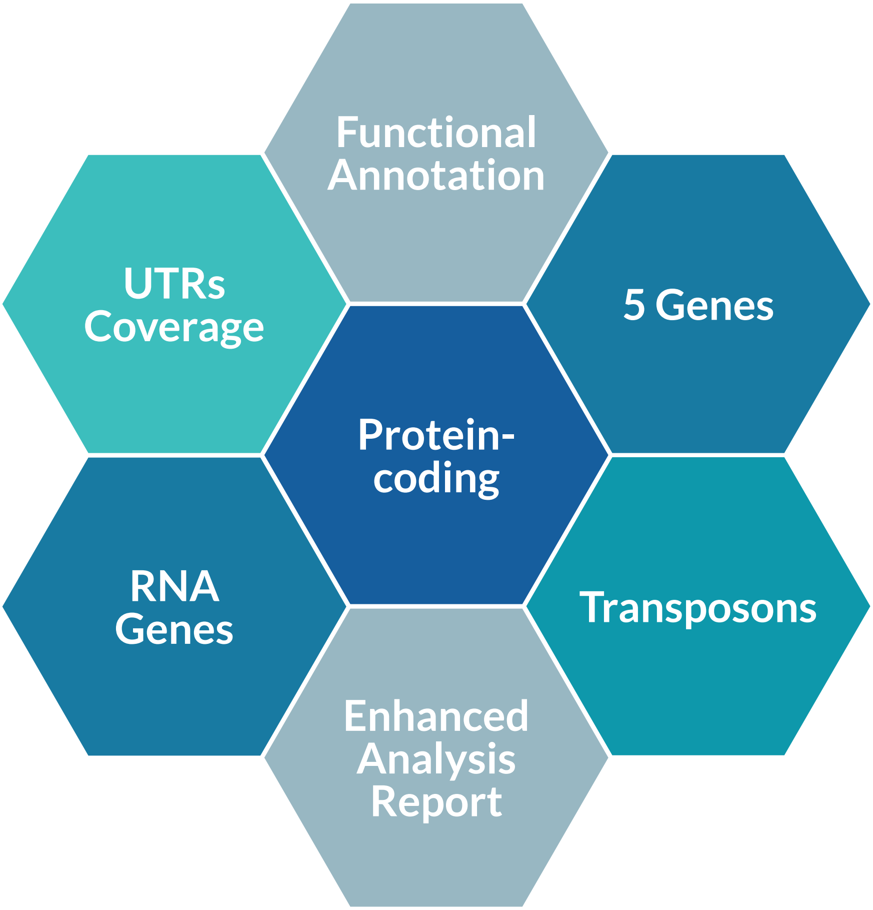
mendle®

ANALYTICS

Genome annotation tailored to your needs



Functional Annotation

UTRs Coverage

5 Genes

Protein-coding

RNA Genes

Transposons

Enhanced Analysis Report

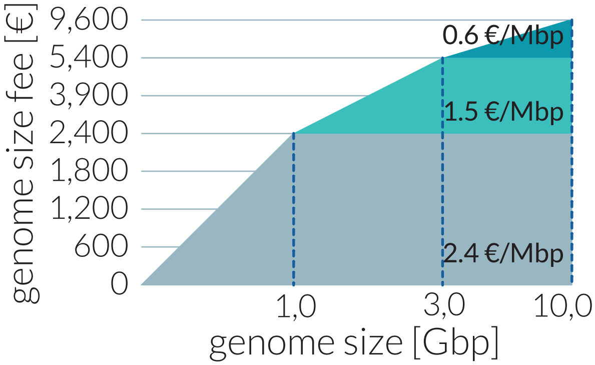
GENOME ANNOTATION PLAN

STARTING DATA evaluation / analysis

A basic evaluation of the provided genome assembly or the assembly selected from mendle® database.

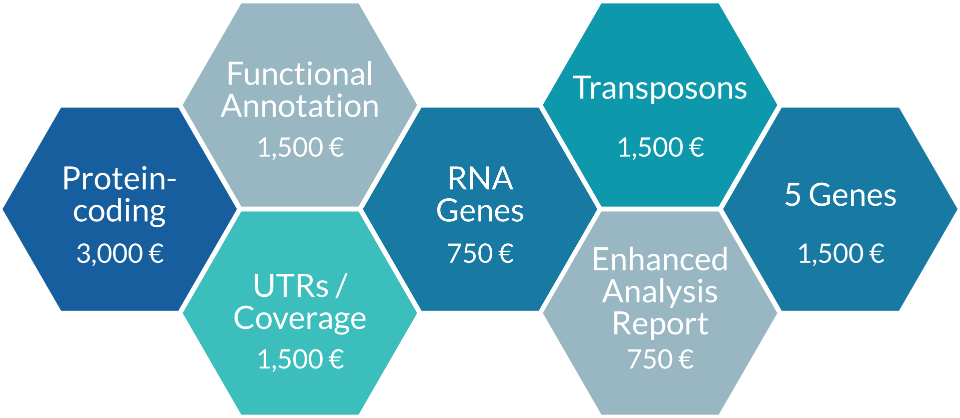
GENOME SIZE dependent fee

The genome size fee depends on the size of the genome to account for the different computing resources needed for small and large genomes.



ANNOTATION packages

Choose from the genome annotation packages.



Protein-coding
3,000 €

Functional Annotation
1,500 €

UTRs / Coverage
1,500 €

RNA Genes
750 €

Transposons
1,500 €

Enhanced Analysis Report
750 €

5 Genes
1,500 €

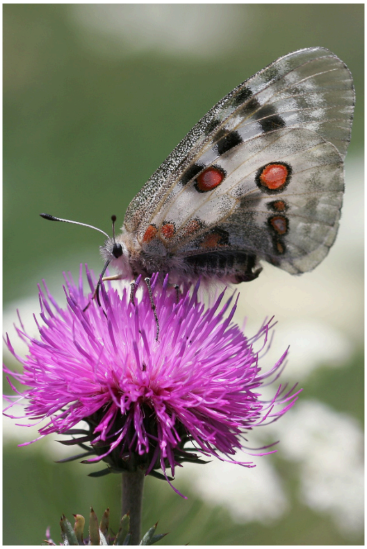
EXAMPLE

genome size: 1.39 GBp APOLLO BUTTERFLY

starting data evaluation	600 €
genome size dependent fee	
1 GBp x 2.4 €	2,400 €
0.39 GBp x 1.5 €	585 €
protein-coding genes	3,000 €
transposons	1,500 €
net	8,085 €

Eukaryota / Metazoa / Ecdysozoa / Arthropoda / Hexapoda / Insecta / Pterygota / Neoptera / Endopterygota / Lepidoptera / Glossata / Ditrysia / Papilionoidea / Papilionidae / Parnassiinae / Parnassini / Parnassius / Parnassius

AR	أبولو; فراشة أبولو
DE	Apollo; Roter Apollo
EN	Apollo; Mountain Apollo; Apollo Butterfly
ES	Apolo
FR	Apollon
JP	アポロウスバ アポロウスバシロチョウ
PT	Borboleta-apolo
RU	Аполлон



© Robert Kindermann, Wikipedia

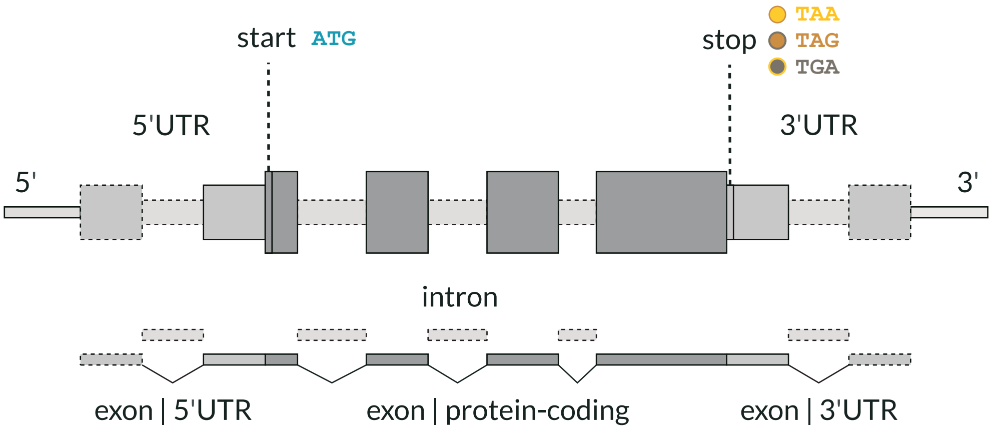
PROTEIN-CODING GENES

STRUCTURAL annotation

Structural annotation is the identification and annotation of the locations of protein-coding genes within a DNA sequence. In the vast landscape of genomic information, the accurate prediction of structural genes is crucial for unraveling the functional elements that govern an organism's biology.

Structural gene prediction takes into account various features, such as open reading frames (ORFs), splice sites, start and stop codons, and consensus sequences, to distinguish coding regions from non-coding ones.

We use a newly developed, proprietary algorithm for homology-based gene reconstruction. To adapt the homology model, we created thousands of genome annotations internally and improved them iteratively.



The gtf and gff file formats impose some restrictions on the correct description of genes. These must be taken into account when working with them or extracting data from them. For example, frameshifts in exons can be the result of sequencing inaccuracies or recent indel events. gtf/gff restrictions require CDS features divisible by three, taking into account phase information. In order to provide a gtf/gff file that is compatible with subsequent analysis tools, frameshifts are represented by artificial "introns".

What is a gene?

Definitions

“A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions.”
Sequence Ontology consortium

“A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.” **ENCODE**

However, given the transcription of almost the entire genome, genes hardly have defined boundaries. Exons from different genes can be part of the same transcript, some microbial genomes contain thousands of scrambled genes that need to be decrypted during development, the functional status of a gene can be passed down to a daughter cell, and at least mammals and plants can rewrite their DNA based on RNA inherited from past generations.

WE DELIVER

annotation in **gff3** format

Gene types are denoted as **Note** in the gff3 attributes of the gene entries. Genes without any issue are termed "protein-coding" and "non-coding" depending on the presence of CDS regions. Genes containing in-frame stop codons or frameshifts are termed "potential pseudogene". Frameshifts are characterised by an indel of 1 or 2 nucleotides. The positions of the in-frame stop codons and frameshifts with respect to their amino acid position are listed in an "issues=" attribute.

amino acid sequences in **fasta** format

The translations of the protein-coding genes are given without the terminal stop. The protein sequences of potential pseudogenes are provided in a separate file. The latter may contain internal stop codons and the letter X at possible frameshift positions. Please note that this may influence the further analysis depending on possible software restrictions.

CDS sequences in **fasta** format

The CDS sequences of protein-coding genes and potential pseudogenes are given without further manipulation to keep the divisible-by-3 character.

FUNCTIONS OF PROTEIN GENES

FUNCTIONAL annotation

Functional gene annotation involves assigning biological functions to the discovered genes within a genome. We assign among others

domain architecture

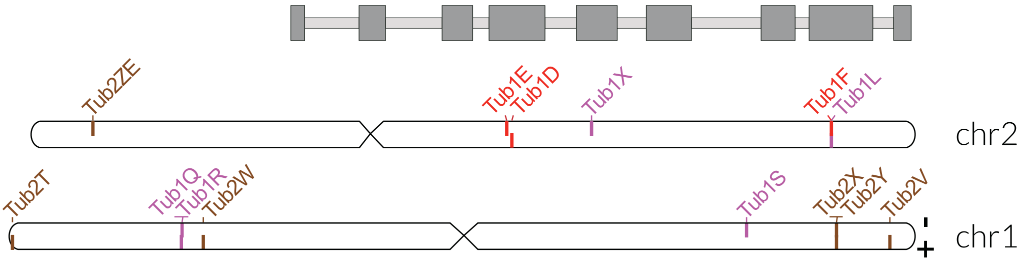
Domain profiles from Pfam, CDD, PRINTS and others.

protein/gene names

Protein naming strongly depends on the reference database used and the quality and completeness of the query sequence. In addition, naming subfamily members within protein families is error prone. In addition to subfamily designations we also provide protein family names.

GO terms

EC numbers



Tubulin alpha-1A chain, TUBA1A
EC 3.6.5.6 - tubulin GTPase
Tubulin/FtsZ family, GTPase domain; Tubulin C-terminal domain
GO:0000166; GO:0005198; GO:0005200; GO:0005515; ...

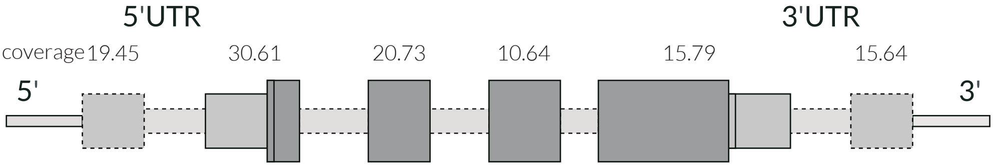
WE DELIVER

Functional annotation is provided in **csv** format. Upon request, the annotation can also be added to the attributes in the gff3 file and/or the fasta headers.

UTRs / COVERAGE

5' and 3' untranslated REGIONS

The exon regions upstream (5') and downstream (3') of the coding exon regions are called untranslated regions (UTRs), 5'UTRs (also known as leader sequences) and 3'UTRs (also known as trailer sequences) respectively. UTRs can be interrupted by introns or alternatively spliced like all other exon regions.



UTRs are determined by mapping RNASeq data or by prediction. Predictions are made from profiles trained with UTRs based on RNASeq data or from profiles of closely related species.

COVERAGE

Coverage by RNASeq data can be defined in two ways: i) Full-length coverage, i.e. the support of the gene model from the 5’ to the 3' end. In this case, the total length of the exons is defined as 100% and the coverage is the percentage of mRNA covered by the RNASeq data. ii) Base-level coverage, i.e. the support of each nucleotide by RNASeq reads.

Here we use the base-level coverage to indicate whether a feature is supported by experimental data. The value indicates whether there is strong support (high values) or low support (low values). Note: The values must not be used for differential expression analysis!

WE DELIVER

UTRs are added to the **gff3** file as additional features to each gene if they can be determined or predicted. Coverage is added as an attribute to mRNA and exon features.

RNA GENES

TRNA genes

tRNA genes are predicted using **tRNAscan-SE**. Annotation types include

cognate
anticodon matches the tRNA isotype; no pseudo genes

non-cognate isotype
anticodon does not match the tRNA isotype

pseudo
score of the tRNA-sequence match too low

undetermined
anticodon contains 'N'

truncated
tRNA not complete, sequence at either or both ends missing

suppressor
anticodon matches one of the stop codons

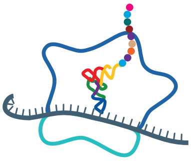
Plotting the identified cognate tRNAs onto the genetic code matrix shows whether all the tRNAs required for decoding are present in the assembly.

	T		C		A		G		
T	Phe		Ser	15	Tyr		Cys		T
		29		8		26		20	C
	Leu	7		16	stop		stop		A
		8		4			Trp	12	G
C	Leu	13	Pro	11	His		Arg	11	T
						24			C
		8		11	Gln	9		5	A
		6		6		9		4	G
A	Ile	25	Thr	11	Asn		Ser		T
				7		31		17	C
		5		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.

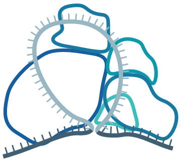
RNA genes, except tRNAs, are identified with **Infernal** using the **Rfam** profiles for ribosomal, spliceosomal, RNase P, RNase MRP, telomerase, snoRNA U3 and other genes.

ribosomal RNA



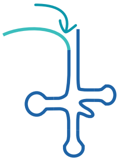
The large ribosomal subunit contains the 5S, 5.88 and 28S rRNAs, the small subunit the 18S rRNA. In eukaryotes, the genes for 18S, 5.8S and 28S rRNA form a single transcription unit that is transcribed into a single large RNA (45S pre-rRNA).

spliceosomal RNA



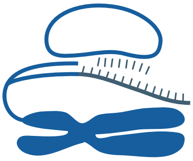
U11, U12, U4atac and U6atac are the analogs of U1, U2, U4 and U6 of the major spliceosome. U5 is shared by both complexes. The minor spliceosome removes introns of the rarer type (ATAC, U12-type).

RNase P | RNase MRP



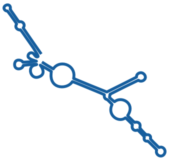
Ribonuclease P (RNase P) is an endoribonuclease, which cleaves the 5'-leader elements of precursor-tRNAs to generate mature 5'-ends. RNase MRP is involved in precursor rRNA processing, where it cleaves the internal transcribed spacer 1 between 18S and 5.8S rRNAs.

telomerase



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

snoRNA U3



U3 snoRNA is predominantly found in the nucleolus and thought to guide site-specific cleavage of ribosomal RNA (rRNA) during pre-rRNA processing.

WE DELIVER

tRNAs and the other types of RNA genes are added to the gff3 file. The type of the RNA is added as **gene_biotype** attribute. The tRNA isotype, anticodon, pseudogene character and other information are added to the **Note** attribute.

TRANSPOSONS

LTR retrotransposons

LTR retrotransposons are an important class of genetic elements that play a significant role in the dynamic evolution of genomes. These retrotransposons are a type of transposable element, a DNA sequence capable of changing its position within a genome, and they constitute a substantial portion of many eukaryotic genomes, including those of plants, animals, and fungi.

These elements have played a crucial role in shaping genome architecture and diversity over evolutionary time. The study of LTR retrotransposons provides valuable insights into the mechanisms of genetic change, adaptation, and the complex interplay between mobile genetic elements and their host organisms. Understanding the biology of LTR retrotransposons is not only essential for unraveling the intricacies of genome evolution but also holds potential implications for genetic diversity, disease, and evolutionary processes across diverse organisms.

Due to their similarity to genes and gene elements, retrotransposons are incorrectly predicted to a considerable extent by common gene prediction programmes. Depending on the genome, 5 to 10 % of the genome annotations in public databases consist of transposons and/or contain parts of transposons in their gene structure.

NON-LTR retrotransposons

Unlike their LTR (long terminal repeat) counterparts, non-LTR retrotransposons lack the characteristic terminal repeats at their ends. Instead, they employ a distinct mechanism for their mobilization within a host genome. These mobile genetic elements utilize a "copy-and-paste" mechanism, where an RNA intermediate is reverse transcribed into DNA by the enzyme reverse transcriptase. The newly synthesized DNA is then inserted back into the genome, resulting in an increase in the copy number of the retrotransposon.

WE DELIVER

Transposons are provided as separate **gff3** file.

ENHANCED ANALYSIS REPORT

analyses, statistics and PLOTS

assembly evaluation
N50 plot, A50 plot, GC content per contig plot

genome annotation metrics
gene stats, exon/intron lengths statistics
distribution of genes on strands
genome covered by genes/features
codon usage frequency
exon phase distribution

annotation quality indicators
intron type distribution (3n/3n+1/3n+2)
analysis of introns with respect to in-frame stop codons
splice site pattern distribution

only in combination with UTR package
distribution of genes with UTRs
genes covered by RNA-Seq

WE DELIVER

The data analyses are provided as a comprehensive **pdf** report. We are happy to provide you with the raw data from the analyses. Please contact us if you require further analyses.

5 GENES

You will select 5 genes for final annotation, which we will analyse manually. This includes the confirmation of exons by RNA-Seq data or comparative genomics and the annotation of the most important alternative splice variants.

WE DELIVER

Results will be delivered as **pdf**. Please contact us if you require further information/data.

CEO

Dr. Martin Kollmar

kollmar@goenomics.com

CTO

Dr. Dominic Simm

simm@goenomics.com

GOENOMICS GmbH

Benfeyweg 9
37075 Göttingen
Germany