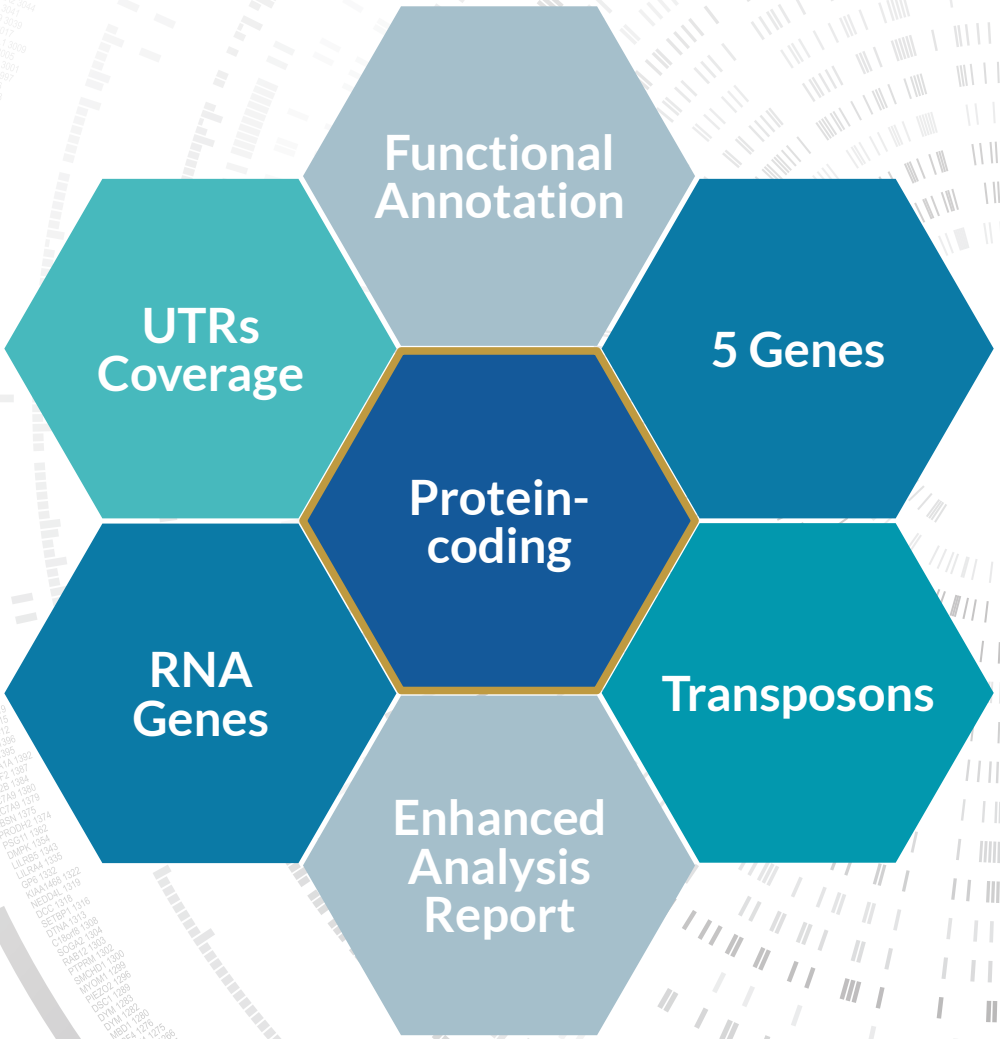
GOENOMICS

Structural annotation

identification of protein-coding genes, non-coding genes and pseudogenes



Structural annotation

identification of protein-coding genes, non-coding genes and pseudogenes

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

Protein gene prediction methods aim to identify the coding regions of genes. Technically, these methods can be categorized as ab initio, homology-based and evidence-based approaches. Ab initio approaches use statistical models and algorithms to detect sequence patterns such as codon usage, open reading frames and splice sites that are typical of protein-coding genes. Homology-based approaches use sequence similarity with known genes of other species. Evidence-based methods use experimental data such as RNA-Seq data, ESTs or full-length cDNAs, and reconstruct gene structures by mapping the experimental sequences to the genome to determine exon-intron boundaries, transcription start and end sites, and splice variants. In practice, these methods are interdependent, and state-of-the-art genome annotation relies on software that analyzes these data synchronously, or on pipelines that combine the approaches in different orders and use different software for each of the steps.

Evidence-based approaches of protein-coding gene prediction

It is assumed that evidence-based approaches provide the most reliable gene structures. However, transcript data are notoriously noisy and contain biological noise (incorrectly spliced and partially processed transcripts, incomplete transcripts, contamination with all possible remnants of transcription and mRNA decay) and computational noise (sequencing errors, repeats and all possible other sequencing artifacts). In addition, mapping this data to the genome assemblies generates further noise due to inaccurate identification of splice sites. Most importantly, evidence-based data is never

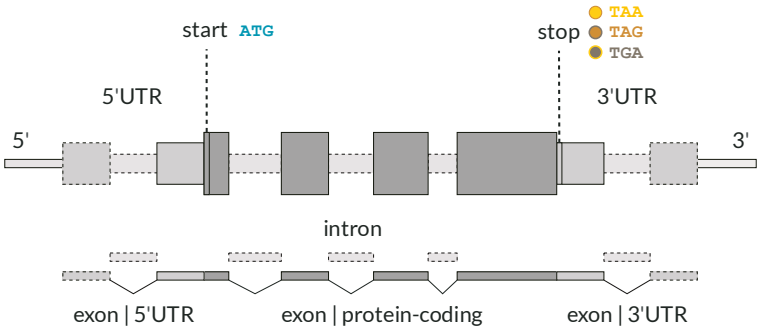


Figure 1: Gene structure with exons and introns from 5' to 3'. Small punctuated boxes represent introns removed during splicing. CDS features and non-coding exons are represented by large, dark-grey boxes and medium-sized boxes, respectively. UTR regions can consist of spliced and non-spliced exons.

Version 2024.11.11

complete due to differential expression of genes and the limited ability to obtain transcript data from all organs and tissues. An increase in sequenced transcript data will increase the completeness of transcripts, but will also increase the level of detected biological noise. The creation of annotations from short read data takes place in two steps: First, the transcript reads are mapped to the genome using software such as TopHat, GSNAP, STAR or HISAT and then the mapped reads are combined into potential genes/transcripts using tools such as Scallop, Cufflinks or StringTie. The mapping of full-length transcripts is usually done with BBmap, Spaln2, exonerate or GMAP.

Homology-based protein-coding gene prediction

While there are over a hundred tools for short-read alignment, there are only a few tools for protein-to-genome alignment, which forms the basis for homology-based gene reconstruction. Protein-to-genome alignment is computationally intensive because differences in query protein and target genome lengths must be accounted for (e.g., the homologs might have shorter or longer protein surface loop regions, so the length of the translated target does not match the length of the protein homolog), the target genome must be translated in all reading frames (introns have random lengths, so only a third of them are divisible by three), and splice sites must be modeled. Tools such as Scipio and GeneWise use BLAT and BLAST, respectively, to identify gene regions and then combine the hits into gene structures and refine exon boundaries and transcription start and end. Tools such as Exonerate, ProSplign, Spaln2 and miniprot align proteins to the genome sequence and refine exon-intron boundaries, including frameshift mutations and sequencing errors. GeMoMa also requires gene structure files of the query proteins as input.

Ab initio protein-coding gene prediction

Genes that are not present in the available RNA-Seq datasets and have no significant homology to known proteins can only be predicted by ab initio methods. Tools such as AUGUSTUS, Genscan, GeneID, GlimmerHMM, GeneMark and SNAP use hidden Markov models for intrinsic features of protein-coding genes, such as codon usage, GC content and sequence motifs such

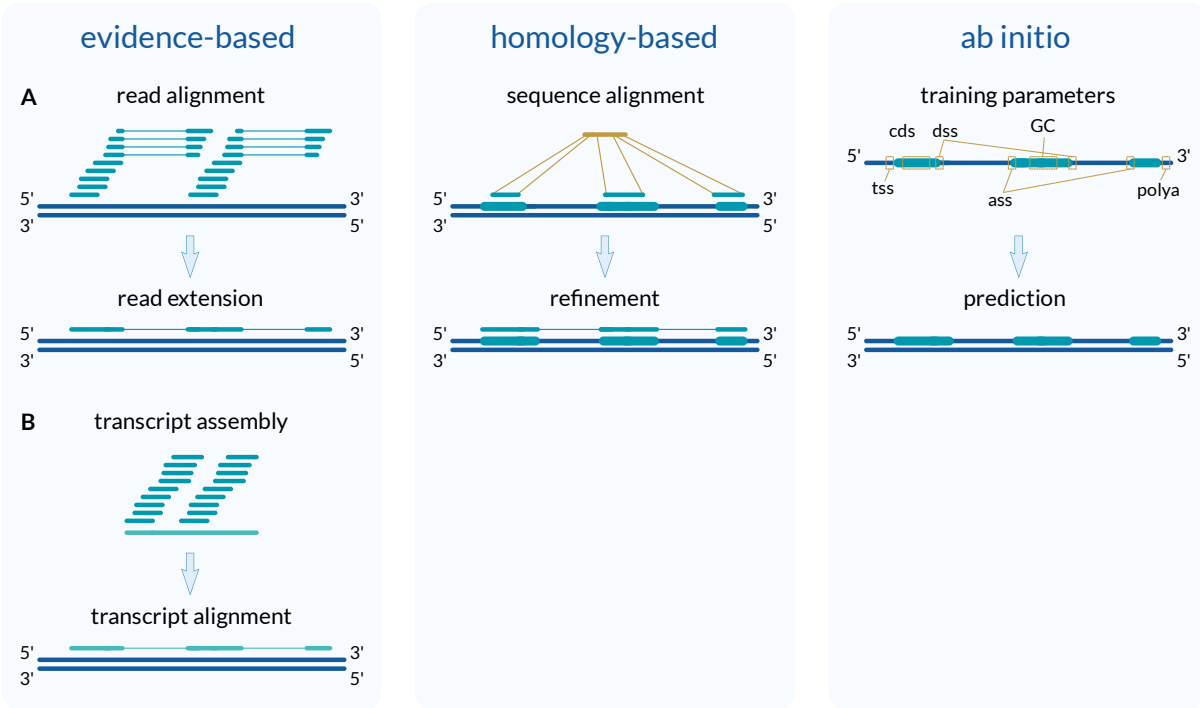


Figure 2: Gene prediction approaches.

Version 2024.11.11

as promoters, start codons and splice sites, which have been trained using different sets of known protein sequences. However, since ab initio methods rely solely on computational predictions, they can miss genes with unusual structures and usually have a high rate of false positives. Most of the current tools use additional data from RNA-Seq and protein homologs to filter and evaluate potential gene structure signals.

Prediction of protein-coding genes in combinatorial approaches

There are several pipelines that attempt to utilize the advantages of all three approaches. The MAKER pipeline generates gene predictions from all approaches in parallel and finally selects the best fitting gene model for each gene region using EVidenceModeler. Ab initio software such as AUGUSTUS is not trained for the specific genome in the MAKER pipeline, but available RNA-Seq data is integrated into the gene prediction process of AUGUSTUS. It is recommended to run the pipeline iteratively several times. The BRAKER pipeline streamlines the entire AUGUSTUS training and prediction process. In a first step, RNA-Seq data and protein homologs are used to identify gene regions and guide ab initio gene prediction with GeneMark. The results of GeneMark are then used to train AUGUSTUS. Finally, all RNA-Seq data, protein homologs and GeneMark predictions are integrated into a final AUGUSTUS gene prediction run. Funannotate can be seen as a combination of MAKER and BRAKER. It uses the mapped transcript and protein evidence together with the available RNA-Seq data to train AUGUSTUS, SNAP and GlimmerHMM. All data is fitted into EVidenceModeler and the final models are filtered by length, gap-bridging and transposable elements.

mendle-analytics identification of protein-coding genes

With mendle-analytics, we combine all approaches in a new pipeline. Taxon-specific protein sequences are collected from multiple sources, including GOENOMICS' database of genome annotations of representative species. These sequences are mapped to the assemblies using newly developed software optimized for resolving non-canonical and fuzzy splice sites, transcript start and end, gene fragments on multiple contigs, all types of sequencing errors and mutations, and challenges due to multiple and overlapping signals from (often tandem) gene copies in the target genome assembly. UTR regions are added to the reconstructed genes by alignment with transcriptome assemblies. The expanded genes are used for thorough training, including training of the UTR regions, of a species-specific parameter set for AUGUSTUS (=> species profile). RNA-Seq data, Isoseq data and transcriptome assemblies are prepared using custom scripts as hints for AUGUSTUS gene prediction. An AUGUSTUS gene prediction is generated based on the new species profile using the expanded genes and transcript data as hints. Finally, the homology-based gene predictions, the Isoseq data and transcript assemblies and the AUGUSTUS gene prediction are merged and divided into sets for protein-coding and non-coding genes as well as (potential) pseudogenes.