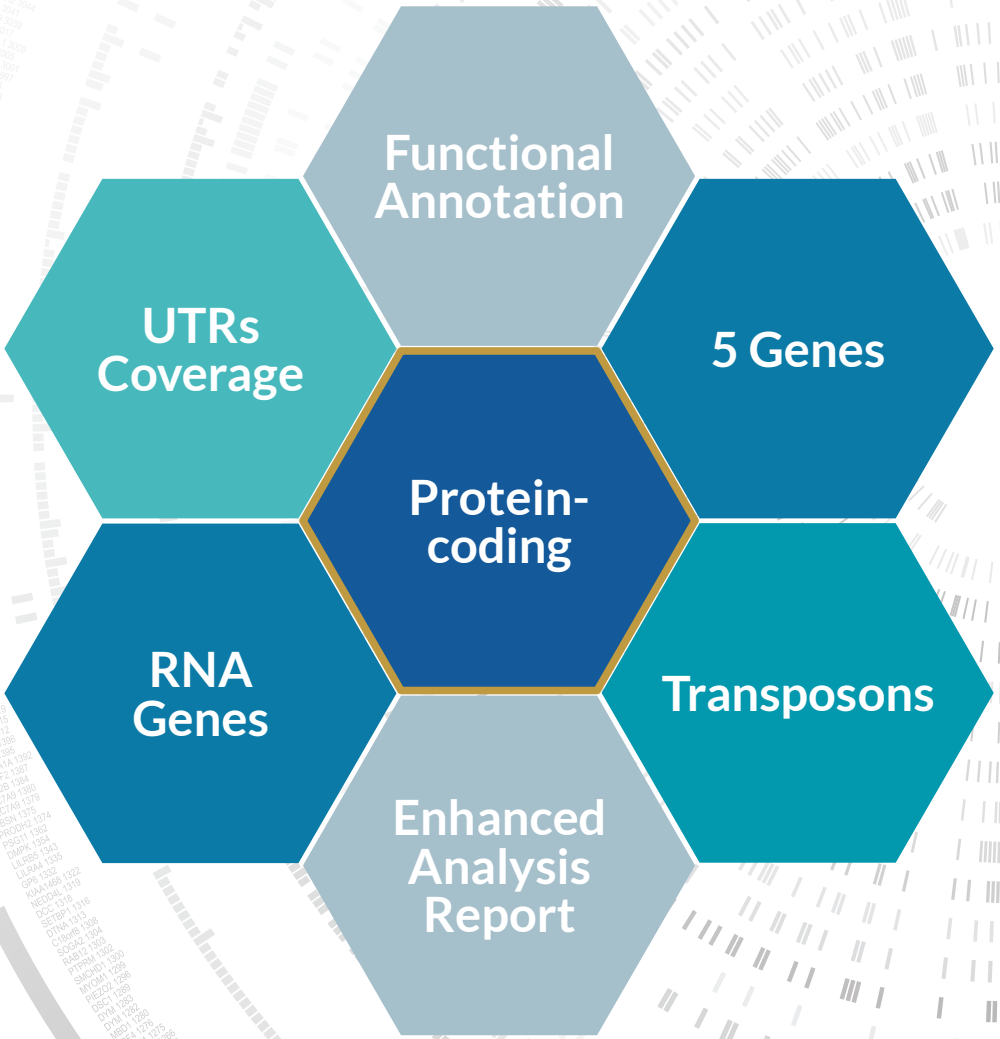
GOENOMICS

Annotation myths debunked

annotated proteins without start and/or stop, highly repetitive protein sequences



Annotation myths debunked

annotated proteins without start and/or stop, highly repetitive protein sequences

Martin Kollmar, Dominic Simm
GOENOMICS GmbH

Incomplete and fragmented genome assemblies are caused by technical and biological challenges. Technical issues include degraded DNA, short-read sequencing limitations, errors in base-calling, low or uneven coverage, limitations in assembly software, and contamination. Biological challenges involve repetitive sequences, structural variations, and high heterozygosity, complicating gene annotation. Partial genes often lack 5' or 3' ends, especially at scaffold boundaries, and complete telomere-to-telomere assemblies are rare. Translation can start at non-AUG codons, challenging the assumption that all proteins begin with methionine. Approximately 12.1% of annotated human proteins do not start with methionine. Repetitive protein sequences, including homorepeats and single alpha-helices (SAH-domains), influence protein interactions and structural dynamics. SAH-domains, composed mostly of charged amino acids, are stable, rigid connectors used in various species, including humans.

Reasons for annotated proteins without start and/or stop

Assembly problems resulting in fragmentation

Genome assemblies are often incomplete and fragmented due to several technical and biological challenges (Figure 1). Some key reasons for technical difficulties are

A) DNA quality and extraction bias: Degraded DNA or biased extraction methods can affect the quality of sequencing reads, resulting in incomplete assemblies.

B) Read length limitations: Short-read sequencing technologies (e.g., Illumina) produce reads that are too short to span repetitive regions or structural variants, leading to fragmented assemblies.

C) Sequencing errors: Errors in base-calling, especially in long-read sequencing (e.g., Pacific Biosciences, Oxford Nanopore), can lead to assembly mistakes. High-error rates during the sequencing process can complicate the assembly process and reduce accuracy.

D) Low coverage or uneven coverage: Insufficient sequencing depth or uneven coverage across the genome can leave gaps in the assembly or create regions with low confidence.

E) Assembly algorithms: Limitations in assembly software (e.g., incorrect scaffolding, misassembly of repeat regions) can result in incomplete or erroneous assemblies.

F) Contamination: Presence of contaminant DNA (e.g., from other organisms, environmental sources) can interfere with assembly and produce misleading results.

Reasons for biological challenges leading to assembly errors are

A) Repetitive sequences: Highly repetitive regions (e.g., transposable elements, tandem

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms at https://goenomics.com/terms-of-use. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

Version 2024.11.11

repeats, centromeres, telomeres) are difficult to assemble accurately because short reads cannot distinguish between identical sequences.

B) Structural variations: Large insertions, deletions, inversions, or duplications are hard to detect and assemble correctly, particularly with short-read technologies.

C) Heterozygosity: High levels of heterozygosity in diploid or polyploid organisms can make assembly difficult, leading to fragmented or mixed contigs.

These technical and biological challenges are the reason why genes can often only be partially annotated at the scaffold boundaries (Figure). In these partial genes, either the 5‘ or the 3’ end of the gene or both ends are missing. The incompleteness often also affects the edges of the so-called chromosome-scale scaffolds. The problem can only be solved by creating telomere-to-telomere assemblies, which are currently only available for a limited number of genomes.

Assembly problems within scaffolds

Scaffold assembly problems, especially tandem duplication of regions and missing sequences, are usually the result of limited long-read sequence data and limitations in the assembly software. Most commonly, these problems affect DNA sequence repeat regions, but sometimes tandem gene regions are also affected, both protein coding and non-coding (e.g. RNA) regions (Figure 2).

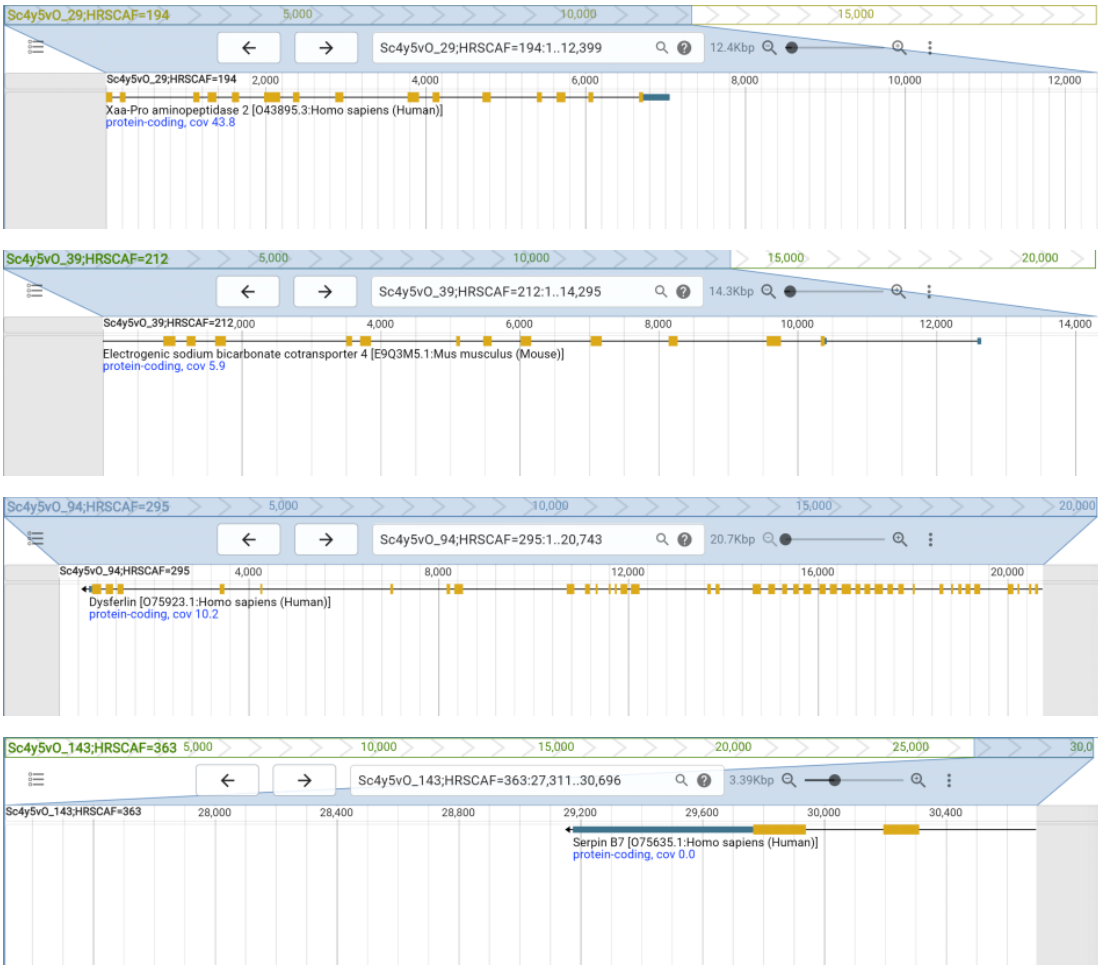


Figure 1: Annotation of partial genes at scaffold edges.

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms at https://goenomics.com/terms-of-use. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

Translation not starting with methionine (ATG)

Although it has been assumed for many decades that translation starts with the methionine codon ATG, ribosome profiling data have shown that up to 50 % of translation starts at non-AUG codons (Cao & Slavoff, 2020). This figure should not be misunderstood to mean that most cellular proteins originate from non-AUG translation events or that initiation at non-AUG start codons is more efficient than at standard AUG codons. Rather, the data merely indicate that when all initiation sites in the entire transcriptome are counted - without considering the efficiency of individual events - translation initiation occurs more frequently at non-AUG start codons than at AUG codons. Accordingly, when annotating protein-coding genes and transcripts, the start codons for translation and the first coding sequence features (CDS region) are often misinterpreted and therefore incorrectly assigned. The Assignment of all start codons to ATG is a purely computer-induced bias. See below the section "Numbers for the GENCODE annotation of the human genome (v43)" to get an overview on numbers of translated transcripts not starting with ATG.

Short open reading frames (sORFs)

Microproteins are polypeptides derived from short open reading frames (sORFs) consisting of fewer than a hundred codons (Basrai *et al*, 1997). They have long been overlooked due to the difficulty of distinguishing coding sORFs from non-coding ones. However, advancements in ribosome profiling, along with improved bioinformatics and proteomics methods, have significantly narrowed the number of potentially translated sORFs from hundreds of thousands or millions to several thousand (Hanada *et al*, 2013; Chen *et al*, 2020; Pennisi, 2024; Schlesinger *et al*, 2025). As a result, efforts are underway to include sORFs with strong evidence of translation into databases like GENCODE. Although the field has been steadily growing, the exact number of functional coding sORFs in the human genome remains unknown, and only a small number of microproteins have been characterized so far. Microproteins can originate from various sources, including: (1) canonical protein-coding transcripts where they are translated as upstream open reading frames (uORFs) from the 5' "untranslated" region; (2) sORFs that overlap with canonical ORFs but are translated out-of-

Figure 2: Region around the ribosomal DNA operon in a eukaryote. Eukaryotic RNA genes are organised in a cluster of the 18S RNA (SSU, approx. 2,000 nt long), the 5.8S RNA (150 nt) and the 28S RNA (LSU, 3,000-5,000 nt). The SSU is present here, but only nucleotides ~2600 to 3400 of the expected 3400 bp are present from the LSU. This is the only region of the ribosomal DNA operon in this 850 Mbp genome, indicating a problem with internal genome assembly.

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms at https://goenomics.com/terms-of-use. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

frame; and (3) transcripts previously classified as non-coding, such as long non-coding RNAs (lncRNAs) or microRNAs. Additionally, sORF translation from pseudogenes has also been observed.

Proteins in the human genome annotation without starting methionine

Researchers normally expect translated transcripts (CDS sequences) to start with a methionine. This is not even true for the official human reference annotation (Table 1). 12.1% of annotated human protein sequences do not start with a methionine, and more than two-thirds of these are longer than one hundred amino acids. Although only 589 of the longest transcripts per gene (2.9 %) are shorter than 100 amino acids, about 20 % of the annotated protein sequences (including alternative isoforms) are shorter than this number. This indicates that a very large proportion of the annotated alternative transcripts are only short pieces compared to their respective longest isoform. It is doubtful that these short isoforms represent functional proteins and not function-restricting isoforms that are never translated.

Highly repetitive protein sequences

Sequences that frequently repeat the same amino acid are known as homorepeats or homopolymeric tracts (Lee *et al*, 2022). Simple repeats often consist of a single amino acid repeated multiple times (e.g., polyQ, polyA, polyG). As low complexity regions they are generally considered part of intrinsically disordered regions (IDRs) that lack a fixed structure. Homorepeats can mediate specific binding events or promote the formation of multiprotein complexes. For example, PolyQ tracts can facilitate interactions in transcription factors and coactivators. Repetitive sequences can also influence gene expression, mRNA stability, and translation efficiency. An example for this function is PolyA tracts that can act as transcriptional activators or repressors. Homorepeats often contribute to protein flexibility or elasticity, enabling proteins to adopt various conformations. They may also form fibrils or aggregates under specific conditions. Homorepeats are thought to evolve rapidly and may provide a mechanism for adaptive flexibility in proteins. They are often observed in proteins involved in transcription, signaling, and development, where functional diversity may be advantageous.

Single alpha-helices: repeat sequences with high percentage of E, K and R

A specific type of repeat sequences form so-called single alpha-helices. Stable single-alpha helices (SAH-domains) function as rigid

Table 1: Numbers for the GENCODE annotation of the human genome (v43).

Counting only the longest translation for each of the 20,366 genes	
shortest translation	2 aa (gene TRDD1)
#nanoproteins (2-10 aa)	26 (none starting with Met)
#microproteins (11-50 aa)	191 (76 starting with Met)
#short proteins (51-100 aa)	589 (2.9% of all genes)
Counting from all genes/transcripts	
#genes	20,366
#translated transcripts	111,276 (Ø 5.46 transcripts/gene)
#transcripts shorter than 50 aa	6,831 (6.1% of all transcripts)
#transcripts shorter than 100 aa	21,586 (19.4% of all transcripts)
#transcripts not starting with Met	13,469 (12.1% of all transcripts)
#transcripts no start-Met and <100 aa	4,464 (4.0% of all transcripts)
Annotated isoforms for human in SwissProt	~1.1 transcripts/gene

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms at https://goenomics.com/terms-of-use. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

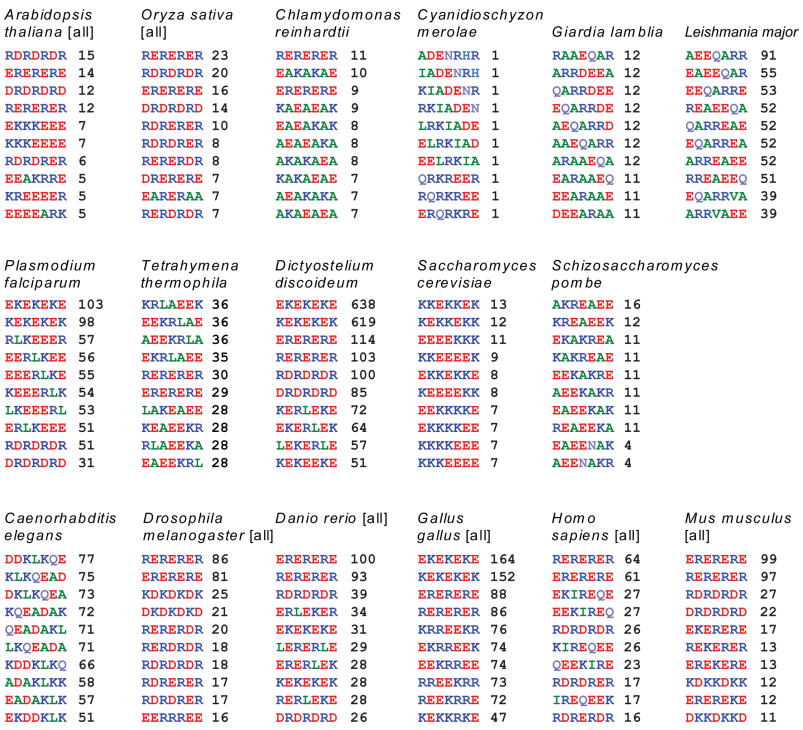


Figure 3: The ten most common heptad repeats found in SAH-domains per species.

connectors and constant force springs between structural domains, and can provide contact surfaces for protein-protein and protein-RNA interactions. SAH-domains mainly consist of charged amino acids and are monomeric and stable in polar solutions, characteristics which distinguish them from coiled-coil domains and intrinsically disordered regions. SAH-domains were predicted in 0.5 to 3.5% of the protein-coding content in 24 species across eukaryotes (Simm & Kollmar, 2018). In human, SAH-domains are mainly used as alternative building blocks not being present in all transcripts of a gene. Another characteristic of SAH-domains distinguishing them from any other domain is their amino acid distribution with up to 80% of the residues being E, K and R.

References

Basrai MA, Hieter P & Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res* 7: 768–771

Cao X & Slavoff SA (2020) Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. *Experimental Cell Research* 391: 111973

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, *et al* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science* 367: 1140–1146

Hanada K, Higuchi-Takeuchi M, Okamoto M, Yoshizumi T, Shimizu M, Nakaminami K, Nishi R, Ohashi C, Iida K, Tanaka M, *et al* (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* 110: 2395–2400

Lee B, Jaberi-Lashkari N & Calo E (2022) A unified view of low complexity regions (LCRs) across species. *eLife* 11: e77058

Pennisi E (2024) 'Dark proteome' survey reveals thousands of new human genes. *Science* 386: 951–952

Schlesinger D, Dirks C, Navarro C, Lafranchi L, Spinner A, Raja GL, Tong GM-S, Eirich J, Martinez TF & Elsässer SJ (2025) A large-scale sORF screen identifies putative microproteins involved in cancer cell fitness. *iScience* 28

Simm D & Kollmar M (2018) Waggawagga-CLI: A command-line tool for predicting stable single α-helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS ONE* 13: e0191924

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms at https://goenomics.com/terms-of-use. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

CEO
Dr. Martin Kollmar
kollmar@goenomics.com

CTO
Dr. Dominic Simm
simm@goenomics.com

GOENOMICS GmbH
Benfeyweg 9
37075 Göttingen
Germany

Research use only. Not for use in diagnostic procedures. Information in this document is subject to change without notice. GOENOMICS assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of GOENOMICS products and/or third party products. Refer to the applicable GOENOMICS terms and conditions of sale and to the applicable license terms. GOENOMICS and mendle are trademarks of GOENOMICS GmbH.

© 2024, GOENOMICS GmbH. All rights reserved.

Version 2024.11.11