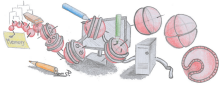
Gene Regulation

Sonja J. Prohaska

Computational EvoDevo, University of Leipzig
Santa Fe Institute, Santa Fe, NM, U.S.



WS 2026/27

Genes and Gene Expression

Today you will learn

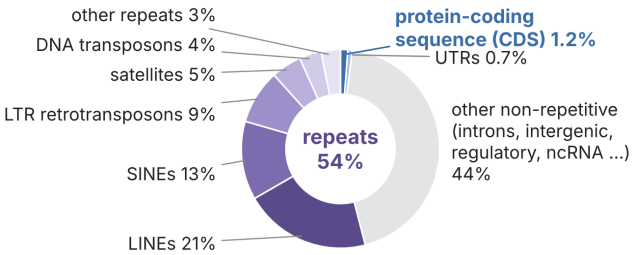
- How a gene is/is not defined
- How gene expression turns a gene into a functional product
- That borders of genic features are defined by gene expression
- How bioinformaticians use evolution (selection) to infer function
- How bioinformatics is used to annotate and predict genes

Slides developed with the help of Claude (Anthropic), an AI assistant; content checked by S. J. Prohaska (2026).

Most Prominent Elements in the Genome – Genes?

Example: the human genome
(complete T2T-CHM13 assembly, 3.05 Gbp)

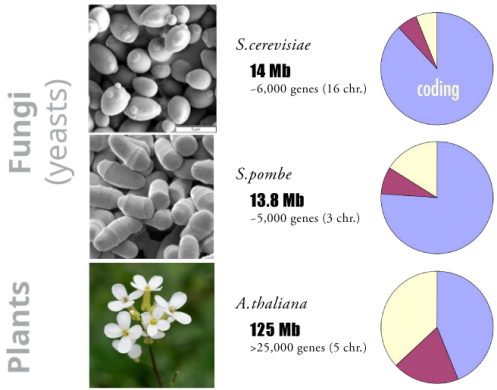
- about **1.2% protein-coding sequence**
  another 0.7% are UTR
- about **54% repeats**
  transposable elements (∼47%)
  satellites (simple, tandem repeats, ∼5%)
- about 75% of the genome is transcribed into RNA (ENCODE 2012)!



ENCODE claimed "biochemical function" for the 75% of the transcribed genome in 2012, BUT does transcription imply function?

Sources: repeats: Nurk et al. (2022) Science 376:44; CDS, UTRs: IHGSC (2004) Nature 431:931; transcription: ENCODE Project Consortium (2012) Nature 489:57.

Genome Composition Varies Between Species



What is a "gene"?

- *What is a gene?*

  (the physical basis of) the heritable unit of function

  sloppy: a DNA sequence encoding a protein (or ncRNA)*
- *What is it made of?*

  nucleic acid, specifically deoxyribonucleic acid, DNA (or RNA, in viruses)
- *What does it do?*

  nothing, it is very passive

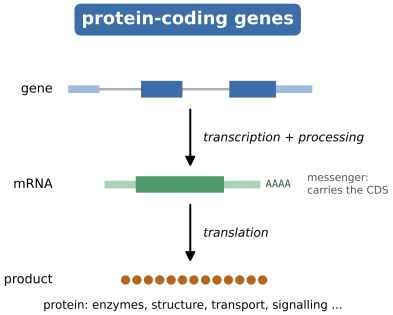
  it holds, provides and stores information about a protein (or ncRNA)*

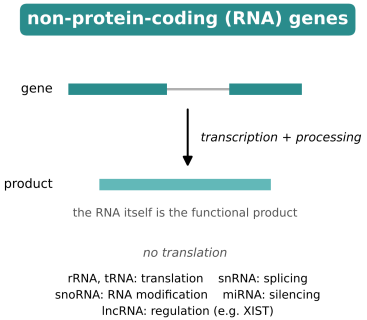
Defining a Gene Today

Modern gene definitions tie a gene to its products.

The **gene** is a construct with a **structural** and a **functional** component, i.e. the DNA sequence(s) and the function of the gene product, respectively.
The components are set into relation by the process of gene expression.

Coding and Non-(Protein-)Coding Genes



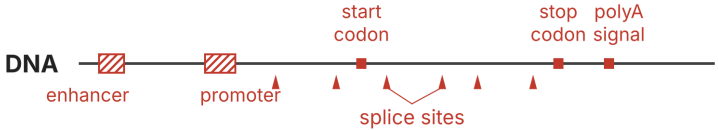


Human genes (GENCODE release 50, total 78,733)



non-coding RNA genes outnumber protein-coding genes

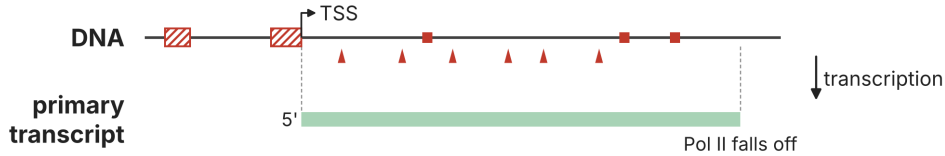
The gene: information encoded in the DNA



Signals (red) are present in the sequence, but no gene feature is defined yet.

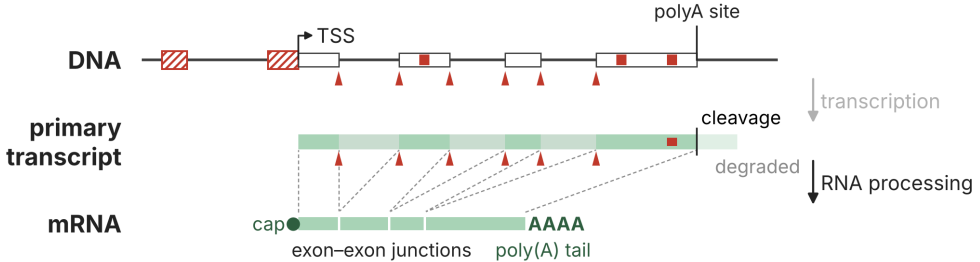
Each step of expression is carried out by its own machinery (trans), reads its own signals (cis) and defines its own features.

Step 1: Transcription



reads (cis)	machinery (trans)	defines	ignores
promoter, enhancers	RNA polymerase II, transcription factors	TSS, primary transcript	start/stop codons, splice sites, polyA signal

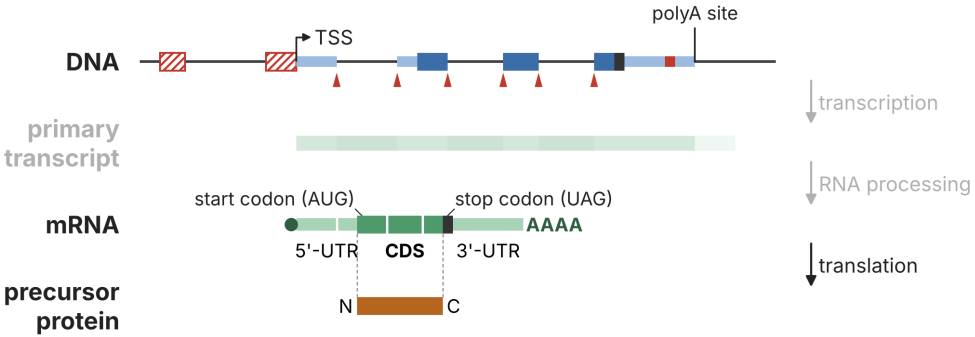
Step 2: RNA processing – from primary transcript to mRNA



	reads (cis)	machinery (trans)	defines
capping	5' end (no signal)	capping enzymes	cap = 5' end of the mRNA
splicing	5'/3' splice sites (GU/AG), branch point	spliceosome	exons (kept), introns (removed)
cleavage & polyadenylation	polyA signal (AAUAAA)	CPSF/CstF, poly(A) polymerase	polyA site = 3' end of the mRNA

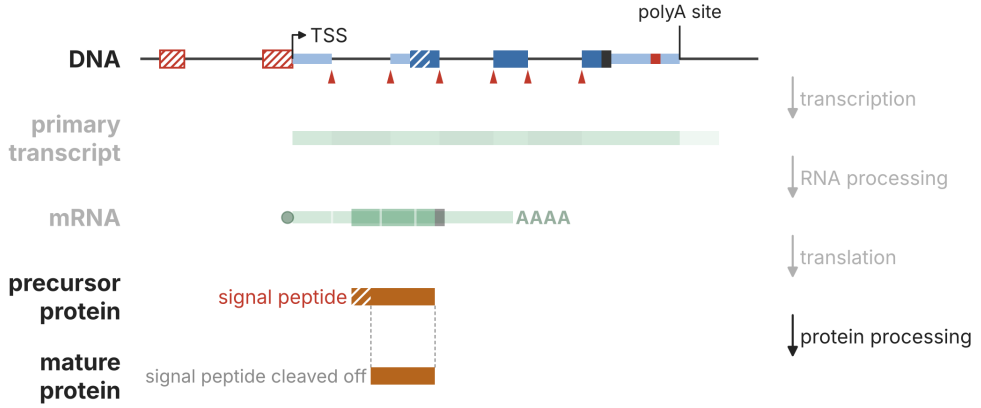
ignores: codons (exon $\neq$ coding). All three steps act co-transcriptionally.

Step 3: Translation



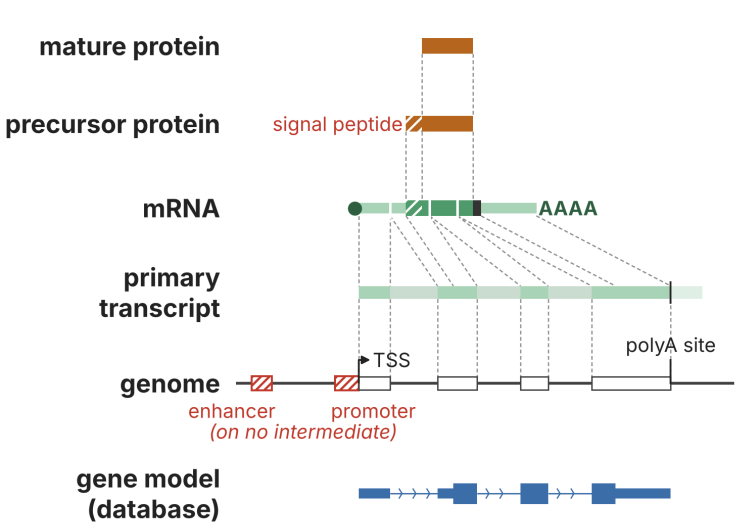
reads (cis)	machinery (trans)	defines	ignores
start codon (Kozak context), reading frame, stop codon	ribosome, tRNAs	TLS, CDS, 5'/3'-UTR	introns: it never sees them

Step 4: Protein processing



reads (cis)	machinery (trans)	defines	ignores
signal peptide, cleavage site	signal recognition particle, signal peptidase	signal peptide, mature protein	UTRs: they are not part of the protein

Step 5: Gene expression defines the borders of the gene features



mapping defines

→ signal peptide

→ CDS (TLS, stop)

→ exons, introns, UTRs

→ gene span: TSS, polyA site

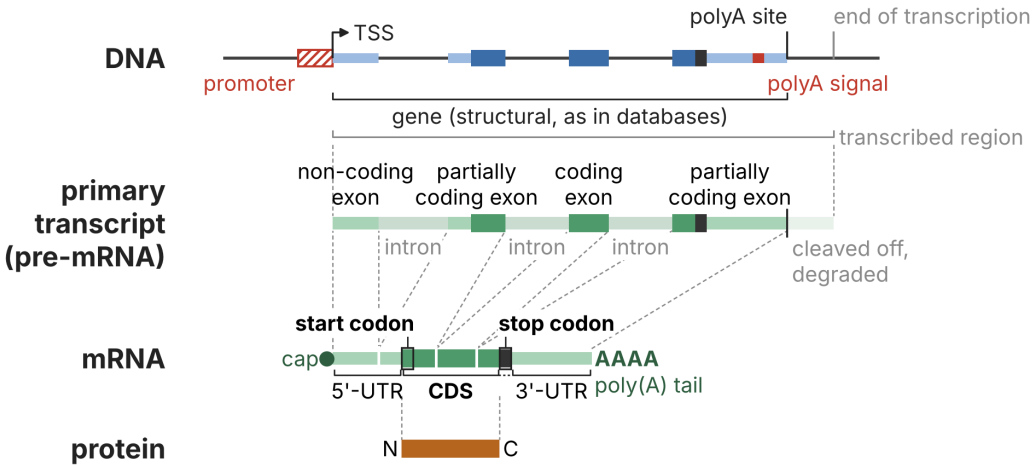
coding (CDS)
signal peptide
UTR
stop codon
regulatory DNA

Gene expression: each step defines its own features

Step	Machinery (trans)	Reads (cis)	Defines
Transcription	RNA Pol II, transcription factors	promoter, enhancers	TSS, primary transcript
Capping	capping enzymes	the 5' end itself	5' end of the mRNA
Splicing	spliceosome	5'/3' splice sites (GU/AG), branch point	exons, introns
Cleavage & polyadenylation	CPSF/CstF, poly(A) polymerase	polyA signal (AAUAAA)	polyA site = 3' end of the mRNA
Translation	ribosome	start codon, reading frame, stop codon	TLS, CDS, 5'/3'-UTR
Protein processing (broad sense)	signal peptidase, proteases, ...	signal peptide, cleavage sites	mature protein

Every term of gene structure names a feature defined by one step of expression.

A Eukaryotic Gene: Structure and Genic Features



The start codon is translated while the stop codon is not.

The content and vocabulary on this slide are very important!

A Eukaryotic Gene: Structure and Genic Features

The structural gene as used in many gene/genome databases

A gene is transcribed into a primary transcript (pre-mRNA) running from the transcription start site (TSS), past the polyadenylation signal, and ends where the polymerase falls off the DNA. The primary transcript is cleaved at the polyadenylation site and a poly(A) tail is added; the downstream part is degraded. The gene therefore spans the region from the TSS to the polyadenylation site. The primary transcript can be partitioned into exonic and intronic sequences. Exons are found on the (m)RNA sequence after splicing while introns are removed. Exons can be (protein-)coding, non-(protein-)coding or partially (protein-)coding. Non-coding, exonic sequences are found at the 5'- and 3'-end of the mRNA sequence and are called 5'- and 3'-untranslated regions (UTRs), respectively. The coding sequence (CDS) lies in between. It starts with the start codon (e.g. AUG), which encodes the first amino acid (Met) and is part of the CDS, and ends at the stop codon (e.g. UAG), which is not translated. Together they form an open reading frame (ORF): a contiguous stretch of nucleotides that is a multiple of 3.

The content and vocabulary on this slide are very important!

Typical pitfalls describing a (eukaryotic) gene (1)

- The (protein) coding sequence (CDS) starts at the **transcription** start site (TSS).

  Wrong! It starts at the **translation** start site (TLS).

  There is always non-coding sequence between TSS and TLS,

  – **on DNA level**, e.g. non-coding exons, non-coding parts of exons, and introns

  – **on mRNA level**, i.e. the 5-prime untranslated region (5'-UTR)

- Exons are **identical to** coding sequences (CDSs).

  Wrong! An exon can be ...

  – (completely) non-coding ("non-coding exon")

  – partially coding and non-coding (often called "coding exon")

  – (completely) coding ("coding exon")

- Introns are spliced out of mRNA.

  Wrong! Introns are spliced out **before** a transcript is called *mRNA*.

  Introns are spliced out of primary transcripts/pre-mRNAs to give rise to mRNAs.

  An exception is **intron retention**. However, from an mRNA perspective the retained "intron" becomes untranslated (UTR) or translated (coding) sequence.

The gene in the figure (previous slides) has 4 exons: 1 non-coding, 2 partially coding, 1 coding.

Typical pitfalls describing a (eukaryotic) gene (2)

Remember: **exons** are the parts of the primary transcript that are **kept** in the mRNA, **introns** are the parts that are **removed**. The following rules follow from this.

- A primary transcript can start or end with an intron.

  Wrong! Its 5'- and 3'-ends become the ends of the mRNA, so the first and last segments are always exons. Splicing removes an intron by joining the two exons flanking it.
- The DNA upstream of the TSS (or downstream of the transcript) is intronic.

  Wrong! It is not transcribed at all. Introns exist only **within** the transcribed region.
- Two introns (or two exons) can be adjacent.

  Wrong! They would simply form one longer intron (exon): exons and introns always alternate.

Hence: a gene with n exons has $n-1$ introns (a single-exon gene has no intron).

The gene in the narrow sense

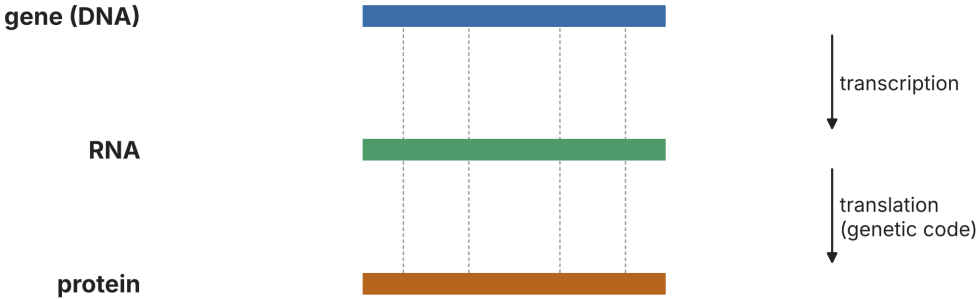
- A gene holds the information about (the sequence of) the gene product.
- Gene expression is the decoding process: DNA $\rightarrow$ RNA ($\rightarrow$ protein)
- One gene – one polypeptide; gene and product are colinear
  (Beadle & Tatum 1941; Yanofsky 1964)

$$\text{product} = f(\text{gene})$$

The price

- cannot handle split genes, alternative splicing (one gene, many products), RNA editing, protein processing
- UTRs, introns and all regulatory signals are not part of the gene

The gene in the narrow sense



product = f(gene)

The gene in the broad sense

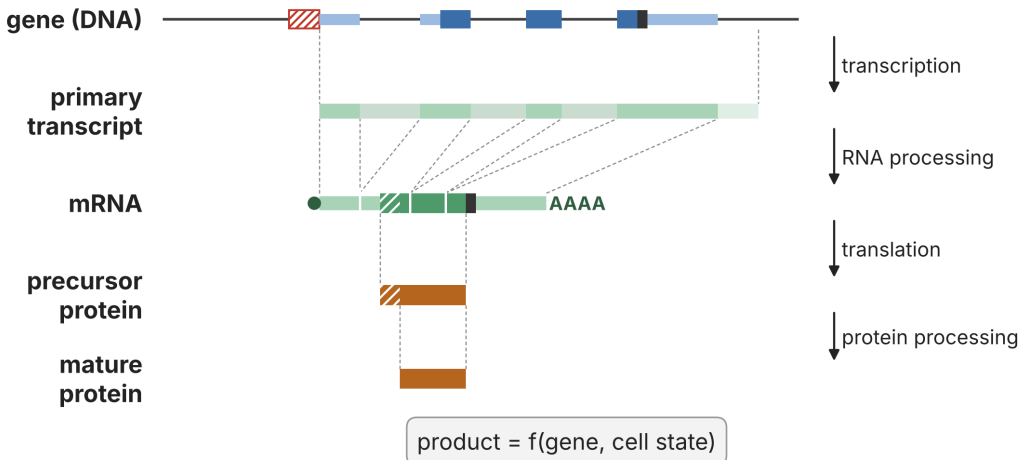
- A gene holds the information used in the synthesis of a gene product, which it specifies together with the cellular machinery.
- Gene expression also includes the processing steps needed to arrive at the functional gene product.
- The gene comprises all cis-acting information: *what* the product is, and *when, where, how much*.

$$\text{product} = f(\text{gene}, \text{cell state})$$

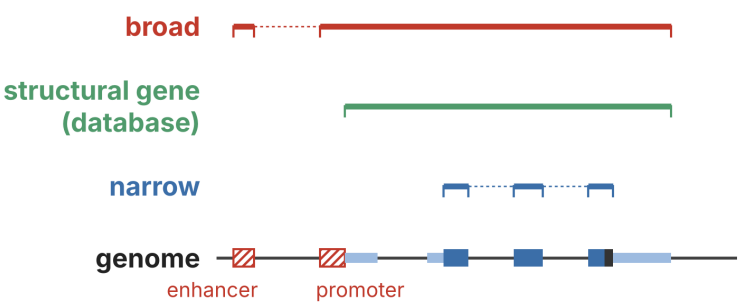
The price

- boundaries depend on the condition (cell type, stage)
- genes become discontinuous, overlap, and share elements (enhancers)
- gene counts become ambiguous; hard to use for annotation

The gene in the broad sense



The structural gene lies between the two extremes



all cis-information used:
+ promoter, enhancers

union of mapped transcripts:
introns, UTRs included

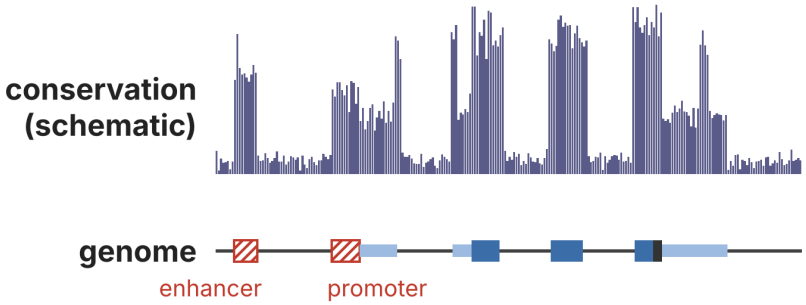
coding sequence only:
'what' the product is

regulatory DNA (an enhancer may serve several genes) is annotated separately and only linked to the gene model

Does transcription imply function? Evolution as a tool

- Biochemical activity (being transcribed, being bound by a protein) is **not** function: about 75% of the human genome is transcribed, much of it probably noise.
- **Function** in the evolutionary (selected-effect) sense: a sequence is functional if selection maintains it.
- Expression is necessary over evolutionary time: a gene that is never expressed is invisible to selection and decays (pseudogene).
- **Evolution is the experiment**: over millions of years mutations hit every position, purifying selection removed the harmful ones. **Conservation** across species is the readout.
- For bioinformaticians without a lab, comparative genomics is *the* way to infer function, e.g. `PhyloCSF` recognises coding regions from multi-species alignments (codon substitution patterns) – without any RNA data.

Selection leaves a footprint



CDS: highest

splice sites,
polyA signal
UTRs: moderate

introns: low

enhancer/promoter:
conserved, but no
transcript shows it

Conservation marks coding *and* regulatory elements – including those that no transcript reveals (enhancers, promoters).

Finding genes in a genome sequence

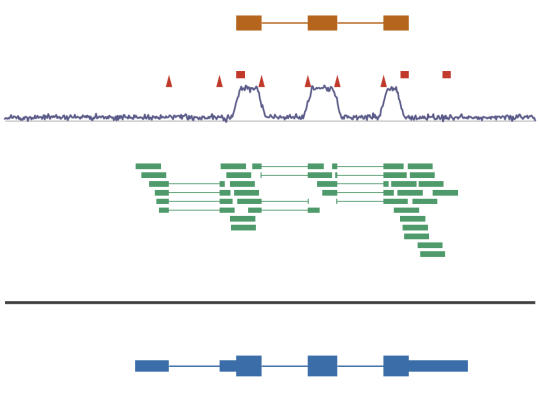
Experimentally: painstaking work on living cells and organisms. Computationally: **gene prediction** – two sources of information:

- **Homology-based prediction** (extrinsic information)
  Genes (proteins, transcripts) already annotated in related species are aligned to the target genome; what is conserved is likely a gene.
  example tool: Gene Model Mapper `GeMoMa`
- **Ab initio prediction** (intrinsic information)
  Conserved gene signatures: the signals read during gene expression (splice sites, start/stop codons, polyA signal) and statistical properties of protein-coding sequence (codon usage, reading frame).
  example tools: `GENSCAN`, `AUGUSTUS`
- **Combined pipelines** use both, e.g. `Gnomon` (NCBI), `BRAKER`, `MAKER`
- specialized tools for particular classes of ncRNA genes (rRNA, tRNA, ...)

The output is a **gene annotation**: the genome coordinates of predicted gene models (exons, introns, CDS, UTRs).

general info: `https://en.wikipedia.org/wiki/Gene_prediction`
Gnomon used by NCBI: `https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon`

Evidence for a gene model



protein of a related species, aligned to the genome

signals (splice sites, start/stop, polyA) + coding statistics

transcripts sequenced and mapped back: exons, introns (split reads), TSS, polyA site

gene annotation (exons, CDS, UTRs)

Predicted, transcribed, functional – not the same

- A **predicted** gene is a hypothesis about gene structure.
- **Transcriptome sequencing** (RNA-seq, long reads) shows which regions are transcribed. Mapped reads define exons, introns, TSS and polyA sites (cf. step 5). They are used
  - as additional information for prediction tools, which improves the predictions,
  - as evidence that a predicted gene is **transcribed**.
- **Conservation** of a gene hints at conserved **function** (selection, see above), but:

> *"Predicting the function of a gene and confirming that the gene prediction is accurate still demands in vivo experimentation through gene knockout and other assays."*
>
> Wikipedia, *Gene prediction*

Gene Annotation vs. Genome Annotation

Gene annotation

- location, i.e. genome coordinates, of known and/or predicted genes
- that should be
  - linked to the gene products (and their functions)
  - and the intermediate steps of the gene expression cascade

    e.g. the corresponding transcript

Genome annotation

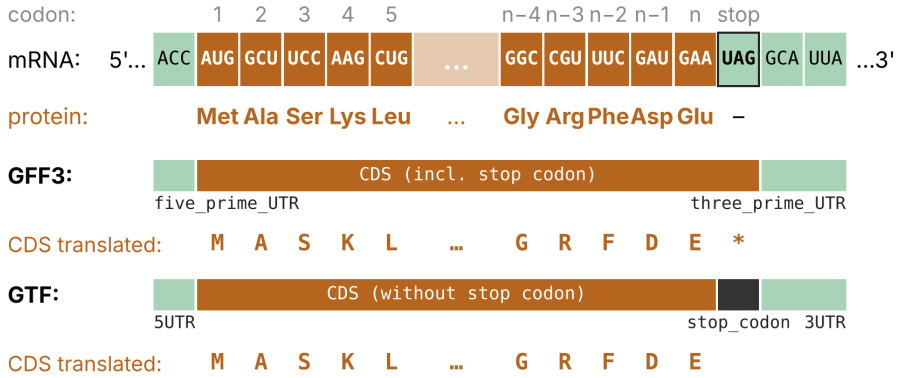
- contains any functional element or sequence features
  - regulatory elements (promoters, enhancers), linked to their target genes
  - transposable elements
  - simple repeats
  - i.e. stuff that is easy to annotate automatically
- may or may not contain the location of genes

Outlook: What is ‘gene regulation’?

It refers to the **regulation of gene expression**
and includes a wide range of mechanisms that are used by cells to increase or decrease the production of specific gene products (protein or ncRNA).

Every figure, number and definition was checked by hand: do the same with AI answers.

Backup: start codon, stop codon and the CDS



GFF3 (Sequence Ontology): CDS includes the stop codon ⇒ its translation ends with "*". **GTF** (GTF2.2): CDS excludes it; the stop codon is a separate `stop_codon` feature (in some files it is counted to the 3'-UTR instead) ⇒ no "*".

CDS lengths differ by 3 nt between the formats – check the convention before computing.

In GTF/GFF files the coding portion of each exon is a separate "CDS" line. Not every ORF is translated, and alternative transcripts have their own CDS.