

Glossary of terms

Topic	Definition
3'	Refers to the third carbon of the nucleic acid sugar moiety to which additional nucleotides may be added by polymerase, often used to refer to that end of a single-stranded DNA or RNA molecule where the 3' carbon is unattached to an adjacent nucleotide; cf. 5'.
5'	Refers to the fifth carbon of the nucleic acid sugar moiety, to which the triphosphate is attached in a nucleotide triphosphate, often used to refer to that end of a single-stranded DNA or RNA molecule where the 5' carbon's phosphate group is unattached to an adjacent nucleotide; cf. 3'.
alternative splicing	The inclusion or exclusion of certain exons in the splicing reactions that determine the sequences included in the final mRNA product. This mechanism is utilized to generate a series of closely related protein isoforms, which differ by the inclusion or exclusion of the particular protein domains encoded by those exons. Alternative splicing is directed by RNA-binding proteins that block, or stimulate, utilization of a particular splice site.
amino acid	The basic building block of proteins, a small molecule with a -C-C- core, an amino group at one end and a carboxylic acid group at the other end. The basic structure can be represented as NH ₂ -CHR-COOH, where R can be any of 20 different moieties, including acidic, basic, or hydrophobic groups.
annotation	Gene annotation is the process of indicating the location, structure, and identity of genes in a genome. As this may be based on incomplete information, gene annotations are constantly changing with improved knowledge. Gene annotation databases change regularly, and different databases may refer to the same gene/protein by different names, reflecting a changing understanding of protein function.
antisense strand	Also called the negative, template, or non-coding strand. This strand of the DNA sequence of a single gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.
base	Although formally incorrect (the nitrogenous base which gives each nucleotide its name is only part of the nucleotide), this is often used as a synonym for "nucleotide."
base pair (base pairing)	The hydrogen bonding of one of the bases (A, C, G, T, U) with another, as dictated by the optimization of hydrogen bond formation in DNA (A-T and C-G) or in RNA (A-U and C-G). Two polynucleotide strands, or regions thereof, in which all the nucleotides form such base pairs are said to be complementary. In achieving complementarity, each strand of DNA can serve as a template for synthesis of its partner strand- the secret of DNA replication's extremely high accuracy and thereby of inheritance.
cDNA	"complementary DNA," a double-stranded DNA molecule prepared in vitro by copying an RNA molecule back into DNA using reverse transcriptase. The RNA component of the resulting RNA-DNA

	hybrid is then destroyed by alkali, and the complementary strand to the remaining DNA strand synthesized by DNA polymerase. The resulting double-stranded DNA can be used for cloning and analysis.
CDS	"Coding sequence", that part of the DNA sequence of a gene which is translated into protein.
coding exon	In a gene, any exon which contains some part of the CDS; in contrast, an exon which has no part translated into protein is called a "non-coding exon."
coding strand	In a gene, the DNA strand that has the sequence found in the RNA molecule. Also called the sense, positive, or non-template strand.
codon	The sequence of three nucleotides in DNA or RNA that specifies a particular amino acid.
coordinates	Numerical position within a biological sequence, e.g. the first base in a DNA sequence would have the coordinate 1.
exon	An exon is a contiguous segment of eukaryotic DNA that corresponds to a portion of the mature (processed) RNA product of that gene. Exons are found only in eukaryotic genomes, and are separated by introns. Although the introns are transcribed with the exons, the latter are spliced out and discarded during RNA processing.
frame	A frame is a single series of adjacent nucleotide triplets in DNA or RNA: one frame would have bases at positions 1, 4, 7, etc. as the first base of sequential codons. There are 3 possible reading frames in an mRNA strand and six in a double stranded DNA molecule due to the two strands from which transcription is possible. Different computer programs number these frames differently, particularly for frames of the negative strand, so care should be taken when comparing designated frames from different programs.
initiation codon (start codon)	The first codon of a coding sequence. In eukaryotes this is almost always ATG, which codes for Methionine.
intron	Non-coding sections of a eukaryotic nucleic acid sequence found between exons. Introns are removed ("spliced out") of mRNA after transcription and before the molecule is exported to the cytoplasm for translation; cf. exon.
isoform	Alternate forms of a gene that are produced by alternative splicing of a particular mRNA, or different transcription start sites. Isoforms of a gene always have different mRNA sequences, but they may have the same protein sequence.
mature mRNA	Messenger RNA that has been completely processed; it has a 7-methylguanosinecap at its 5' end, a poly (A) tail at its 3' end, and has all its introns spliced out from it.
non-coding strand	Also called the negative, template, or anti-sense strand. This strand of the DNA sequence of a single gene is the complement of the 5' to 3' DNA strand known as the sense, positive, non-template, or coding strand. The term loses meaning for longer DNA sequences with genes on both strands.
ORF	"Open Reading Frame", a long stretch of codons in the same reading frame uninterrupted by stop codons; an ORF may reflect

	the presence of a gene.
phase	The phase describes the relationship between the translation frame of an exon and the position of a splice junction. In the GEP we define the term to describe the number of bases between the end of the exon (defined by the splice site) and the full codon nearest that splice site. The number of bases between the adjacent full codon at an exon/site junction can be either 0, 1 or 2. The phase of an exon/splice donor junction will determine which frame is translated in the downstream exon as it will indicate how many bases are used after the acceptor splice site to create a full codon of 3 bases.
poly(A) tail	The segment of adenylate residues that is posttranscriptionally added to the 3' end of eukaryotic mRNA. About 250 nucleotides of (A) are added by poly (A) polymerase following cleavage of the newly synthesized RNA about 20 nucleotides downstream of an AAUAAA signal sequence.
pre-mRNA	The initial transcript from a protein-coding gene is often called a pre-mRNA and contains both introns and exons. Pre-mRNA requires processing (addition of 5' cap and 3' poly (A) tail, removal of introns} to produce the final mRNA molecule containing only exons.
promoter	A segment of DNA to which RNA polymerase binds to initiate transcription of the downstream gene(s).
read	A raw DNA sequence
splicing	The process by which introns are removed and exons are joined to produce a mature, functional RNA from a primary transcript. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.
splice acceptor site	The boundary between an intron and the exon immediately downstream (i.e. on the 3' side of the intron).
splice donor site	The boundary between an intron and the exon immediately upstream (i.e. on the 5' side of the intron).
splice junction	Either a splice acceptor site or a splice donor site.
stop codon (termination codon)	A codon that specifies the termination of peptide synthesis; sometimes called "nonsense codons," since they do not specify any amino acid.
transcription	The process of copying one strand of a DNA double helix by RNA polymerase, creating a complimentary strand of RNA called the transcript.
translation	The process by which codons in an mRNA are used by the ribosome to direct protein synthesis.
UTR	"Untranslated region", a segment of DNA (or RNA) which is transcribed and present in the mature mRNA, but not translated into protein. UTRs may occur at either or both the 5' and 3' ends of a gene or transcript.