Glossary of Terms

Note: Words in **bold** in a definition indicate terms also defined in this Glossary

Term	Definition
3'	"3 prime"; Refers to carbon 3 of the nucleic acid sugar component (either ribose in RNA or deoxyribose in DNA) 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 retains its hydroxyl group (-OH) and no further nucleotides are bonded.
5'	"5 prime"; Refers to carbon 5 of the nucleic acid sugar component (either ribose in RNA or deoxyribose in DNA), 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(s) is/are unattached to a preceding nucleotide.
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 regions encoded by those exons. Alternative splicing is directed by RNA-binding proteins that may 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 amine group ($-NH_2$) at one end and a carboxylic acid group (-COOH) at the other end. The general structure can be represented as NH_2-CHR-COOH, where R can be any of 20 different functional groups of acidic, basic, or nonpolar character.
annotation	Gene annotation is the process of notating the location, structure, and identity of genes in a genome. As initial attempts may be based on incomplete information, gene annotations are constantly changing as further data becomes available. Gene annotation databases are updated regularly, and different databases may refer to the same gene/protein by different names, reflecting new knowledge and improved understanding of protein function.
base	Although formally incorrect (the nitrogenous base that defines A, C, G, T and U is only part of the whole **nucleotide**), this is often used as a synonym for "nucleotide" in referring to the A, C, G, T, and U components of DNA and RNA.
base pair/base pairing	The hydrogen bonding of one of the **bases** (A, C, G, T, U) with another, as dictated by optimal 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.

canonical	In agreement with existing principles and standards generated from data and evidence. For example, the canonical **splice donor sequence** is GT. In rare instances, however, the sequence GC is used instead; since GC is not the "standard", it would be referred to as **non-canonical**.
cDNA	"complementary DNA"; a double-stranded DNA molecule prepared *in vitro* ("outside of the body"; i.e., in a test tube) by employing an RNA molecule as a template to synthesize DNA using reverse transcriptase. The RNA component of the resulting RNA-DNA hybrid is enzymatically degraded, and the complementary strand then 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 that is translated into protein.
coding exon	In a gene, any **exon** that contains some part of the **CDS**; in contrast, an exon that has no part translated into protein is called a "**non-coding exon**."
coding strand/ positive 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**.
coordinate	Numerical position within a biological sequence; for example, the first base in a DNA sequence would have the coordinate "1".
downstream	Refers to the genomic region that comes after the **feature** being examined.
exon	In eukaryotes, a contiguous segment of DNA that corresponds to a portion of the mature (processed) RNA product of that gene. Exons in eukaryotic genomes are often, but not always, separated by **introns**. Although exons are transcribed with the introns, the latter are spliced out during RNA processing and degraded.
feature	Any region of defined structure/sequence in a genomic fragment of DNA. Inherent features would include genes, pseudogenes, and repetitive elements. A feature may also be predicted by computational algorithms, such as those aimed at identifying protein-coding genes.
intron	Non-coding section of a eukaryotic nucleic acid sequence found between exons. Introns are removed ("spliced out") from the primary transcript/pre-mRNA after transcription and before the molecule is exported to the cytoplasm for translation.
isoforms	Potentially different versions of a protein encoded by a single gene. Isoforms result from alternative splicing of a particular **pre-mRNA**, and/or the use of a different transcription start site.
mRNA	Mature messenger RNA that has been completely processed and is ready for translation; it has a 7-methylguanosine cap at its **5'** end, a **poly(A) tail** at its **3'** end, and has all its i**ntrons** spliced out.

non-coding strand/ negative strand	Also called the anti-sense, 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 positive, sense, non-template, or **coding strand**. The term loses meaning for longer DNA sequences with genes on both strands.
nucleotide	The basic building block of DNA (A, C, G, T) and RNA (A, C, G, U). Nucleotides consist of a nitrogenous base, a 5-carbon sugar (either ribose in RNA or deoxyribose in DNA), and phosphate group(s).
ORF	"Open reading frame"; a long stretch of **codons** in the same reading **frame** uninterrupted by **termination codons**; an ORF may reflect the presence of a gene.
phase	The phase describes 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 and an exon/splice site can be 0, 1 or 2. The phase of an **upstream** exon will determine which **frame** is translated in the **downstream** exon by indicating how many bases after the splice acceptor site are needed to create a full codon of 3 bases.
poly(A) tail	About 250 adenine **nucleotides** that are post-transcriptionally added by poly(A) polymerase to the **3'** end of eukaryotic transcripts, following cleavage of the newly synthesized RNA ~20 nucleotides **downstream** of an AAUAAA polyadenylation signal sequence.
pre-mRNA (primary transcript)	The initial transcript from a protein-coding gene that contains both **introns** and **exons**. Pre-mRNA requires the addition of a **5'** cap and **3' poly (A) tail** and the removal of introns to produce the final **mRNA** molecule containing joined exons.
promoter	A segment of DNA to which RNA polymerase binds to initiate **transcription** of the **downstream** gene(s).
putative	Something that may be predicted or inferred but that requires more evidence to confirm or refute.
read	A raw DNA sequence.
reading frame/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 three 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, so care should be taken when comparing designated frames from different programs. One common way is to refer to the three possible left-to-right reading frames as +1, +2, and +3 and the three possible right-to-left reading frames as -1, -2, and -3.
splicing	The process by which **introns** are removed and **exons** are joined to produce a mature, functional RNA (**mRNA**) from a **primary transcript**. Some RNAs are self-splicing, but most require a specific ribonucleoprotein complex to catalyze the reaction.

splice acceptor site	The **splicing** site at the **3'** end of an **intron**, at the boundary between an intron and the **exon** immediately **downstream**. The **canonical** splice acceptor site dinucleotide sequence is AG.
splice donor site	The splicing site at the 5' end of an **intron**, at the boundary between an intron and the **exon** immediately **upstream**. The **canonical** splice donor site dinucleotide sequence is GT; in rare cases, the **non-canonical** sequence GC is used instead.
splice junction	Either a **splice acceptor site** or a **splice donor site**.
start codon (initiation codon)	The first codon of a **CDS**. In eukaryotes this is almost always ATG, which codes for methionine (one of the 20 **amino acids**).
stop codon (termination codon)	A codon that specifies the termination of protein synthesis; sometimes called a "nonsense codon" since it does not specify an **amino acid**.
transcription	The process of copying one strand of a DNA double helix by RNA polymerase, creating a complementary strand of RNA called the transcript.
translation	The process by which codons in an mRNA are "read" by the ribosome and tRNAs to direct protein synthesis.
TSS (transcription start site)	The location in DNA, generally **upstream** of a gene's coding sequence, where RNA polymerase begins **transcription**.
UTR	"Untranslated region"; a segment of DNA (or RNA) that is transcribed and present in the mature **mRNA** but is not translated into protein. UTRs may be found at either or both of the **5'** and **3'** ends of a gene or transcript.
upstream	Refers to the genomic region prior to the **feature** being examined.