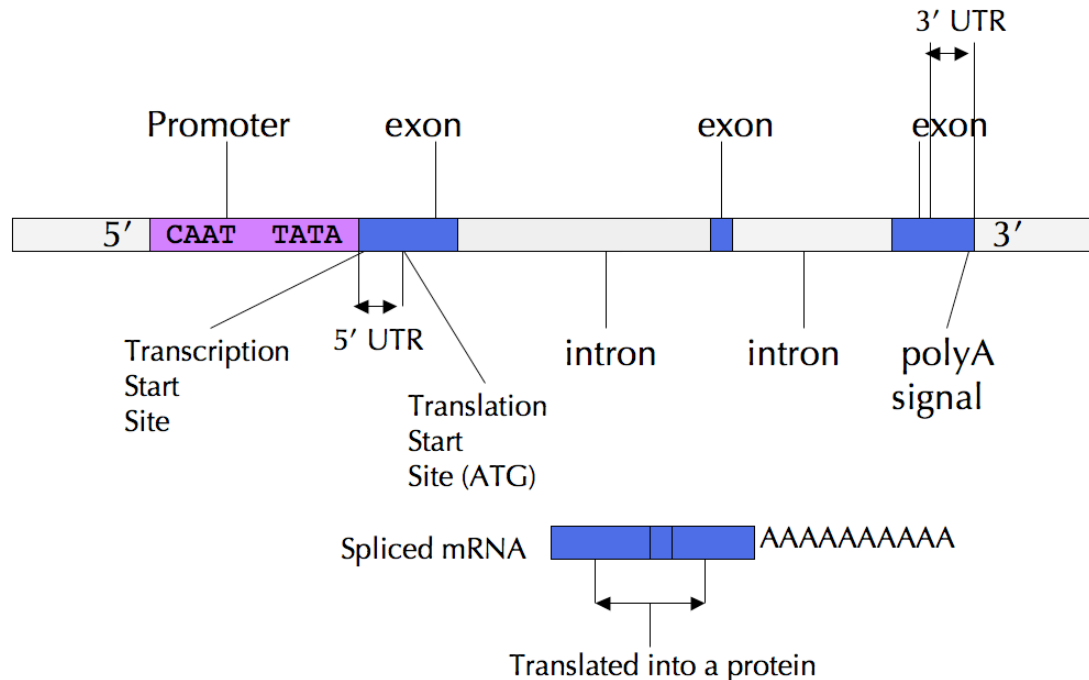


Gene structure/splicing

Genes have parts:



TSS = transcription start site = where the RNA polymerase sits down on the DNA and starts to make an RNA copy of the DNA

ATG = translation start site = where the ribosome starts reading the mRNA to follow its codons to make a protein

UTR = untranslated region = region of the mRNA that does not get translated into a protein. Often regulatory.

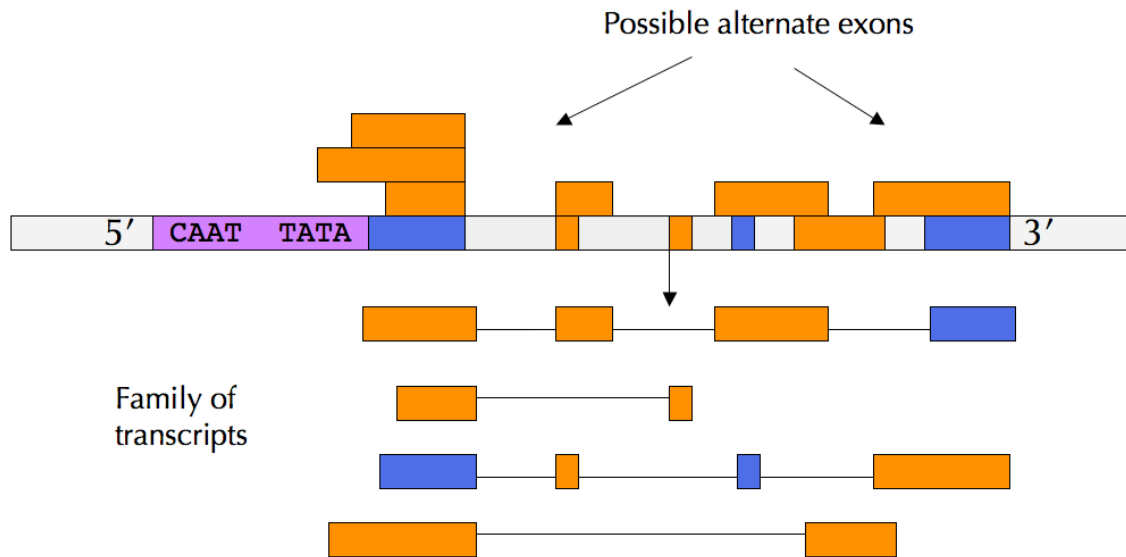
Exon = the coding part of a gene = the gene sequence that is decoded to make a protein

Intron = “intrusive” = the parts of a gene that lie between exons and get spliced out of the transcript to produce the mature mRNA. Introns are essential in higher eukaryotes; yeast only have ~30 introns in their whole genome; nobody quite knows what they’re for. Unicellular organisms generally have to replicate quickly and cannot tolerate all that extra DNA (or so it’s thought).

Poly (A) signal = an untemplated run of adenosines that gets attached to the 3’ end of an mRNA transcript (except in the case of histone mRNAs which are treated differently). Thought to be critical for mRNA stability and processing.

How we know:

TSS are really hard to figure out, mostly because the 5' UTR does not make it into the final protein. 5' RACE is a common lab technique that's used to find TSSs, but this can slightly miscalculate the TSS position. Additionally, some genes use only one TSS but many do use more than one TSS, in a regulated way. In the picture, the blue exons are the most commonly accepted annotated exons and the orange exons are a bunch of alternate exons. There are many, many, many transcripts that could come from this gene, and they will have different starts/stops/proteins. This is why "gene" is a nebulous and not always very helpful concept.



ATG are easier because that's where the protein starts, though that piece of the protein usually gets cleaved quickly. Many genes have more than one place where translation could start and still give pretty much the same protein product.

Poly(A) signals are not too hard to find because we can use oligo d(T) binding methods to pull polyadenylated transcripts out of a mixture and then sequence them from the 3' end. This is one way to make ESTs (another is to use random hexamers to start sequencing transcripts from the middle). Using this method it's easy to see that many genes can use a family of different 3' UTRs.

Exons can be predicted from genomic sequence but this is not very reliable. Alternative splicing is still extremely difficult to predict and is still not well understood. The picture is made only very slightly easier by the fact that the 3 base genetic code means that for each exon, not every possible downstream exon can be used if that spliceform is to code for a protein. Not all spliceforms code for proteins, though. Only very rarely in higher eukaryotes do exons overlap on the plus and minus strands, partly because this situation severely constrains the exon sequence (a silent mutation on one

strand is only rarely not detrimental to the exon on the opposite strand). If exons are predicted (using one of many possible gene prediction programs), their boundaries are not always precise.

Introns are traditionally defined as noncoding but can contain stretches of open reading frames that can be spliced into the final mRNA; introns can contain functionally relevant antisense transcripts (coding or noncoding); introns can contain critical regulatory elements; and intron length, sequence, position, and number is not generally conserved even between closely related species (this is something I find astounding).

UTRs are determined by default once all the other signals are known. Many of a gene's regulatory elements lie in the UTRs, though promoter and enhancer architecture (as well as other cis- and trans- factors) is critical.

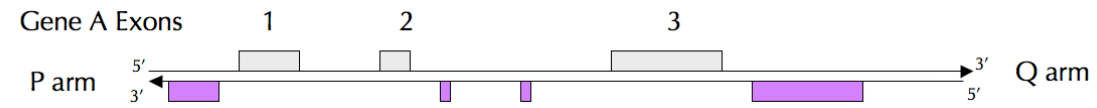
Which feature files to use depends on the question being asked. If someone is looking very closely at transcription start sites and their relationship to some feature of interest, then all annotated TSSs (except those marked "dubious") should probably be considered. When the question is more general, something like RefSeqs or CCDSs might be much more helpful because the annotation is clear and widely accepted. Those annotations contain only one transcript per gene, with nonoverlapping exons, though an exon from a minus strand gene may show up in the intron of a plus strand gene.

Genes comprise a tiny fraction (2%, for exons) of the genome, so having overlapping or interleaved exons is unexpected, to say the least, but interleaved exons are fairly common.

A note on numbering:

All chromosomes in all species (except for circular chromosomes) have agreed-upon numbering that goes from one end to the other (the direction generally determined by chromosomal structural properties), so that the plus strand coordinates increase along the 5' to 3' direction and the minus strand coordinates increase from 3' to 5' along that strand. 5' and 3' are chemical designations for the type of sugar linkage that happens along the DNA and RNA backbones; this gives DNA and RNA a directionality. This means that for a gene on the plus strand, the 5' exons have lower numbered coordinates than the 3' exons, and for a gene on the minus strand, the opposite is the case.

Minus strand genes are often written with the exon start and stop coordinates "backward," that is, stop is a lower number than start, to indicate that the exon (or any feature) is on the minus strand.



Gene B Exons

4 3 2 1

0 1 2 numbering . . . 10000 10001 . . .

1	80	-	geneB_ID
100	300	+	geneA_ID
700	775	+	geneA_ID
800	850	-	geneB_ID
1200	1250	-	geneB_ID
2000	2500	+	geneA_ID
2800	3750	-	geneB_ID

100	300	geneA_ID ex1	3750	2800	geneB_ID ex1
700	775	geneA_ID ex2	1250	1200	geneB_ID ex2
2000	2500	geneA_ID ex3	850	800	geneB_ID ex3
			80	1	geneB_ID ex4