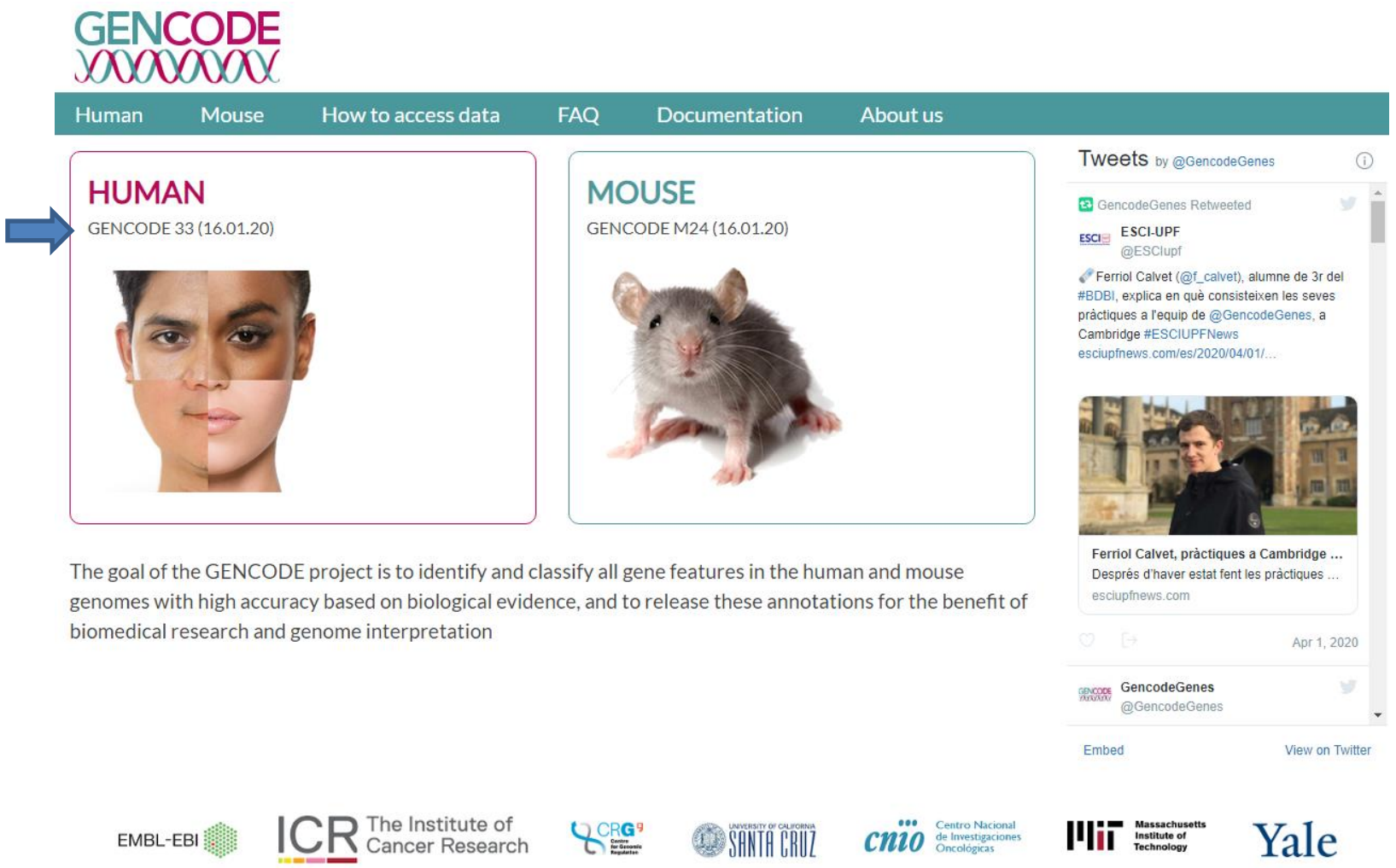
Ch 4 – L1.3

What is a gene ?

It is worth spending few minutes on the Statistics to consider how many different types of long- and short-noncoding RNA have been catalogued



http://www.gencodegenes.org/

Human

Statistics about the current GENCODE Release (version 33)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

Total No of Genes	60662	Total No of Transcripts	227912
Protein-coding genes	19957	Protein-coding transcripts	84107
Long non-coding RNA genes	17952	- full length protein-coding	58048
Small non-coding RNA genes	7576	- partial length protein-coding	26059
Pseudogenes	14768	Nonsense mediated decay transcripts	15937
- processed pseudogenes	10672	Long non-coding RNA loci transcripts	48438
- unprocessed pseudogenes	3554		
- unitary pseudogenes	232		
- polymorphic pseudogenes	55		
- pseudogenes	18	Total No of distinct translations	62357
		Genes that have more than one distinct translations	13739
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments	408		
- pseudogenes	237		

The Gencode is mirrored and searcheable (Browser) at different locations, including the UCSC Gencode Browser: https://genome.ucsc.edu/

RNA Biotypes

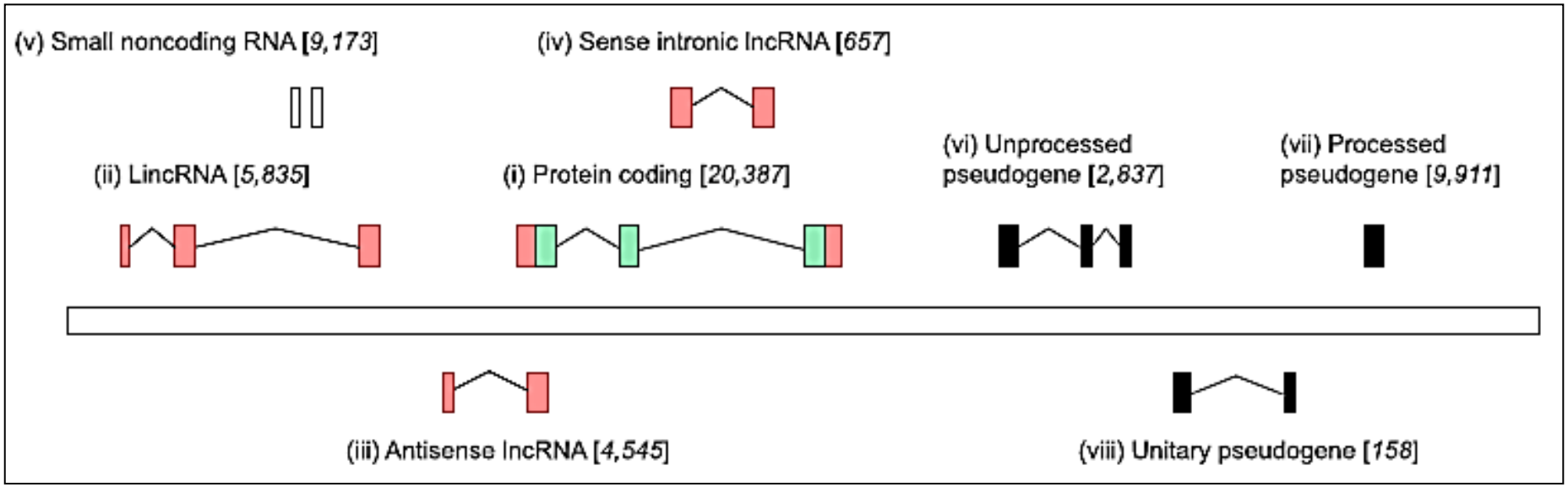


Figure 2. A summary of locus biotypes in GENCODE.

From Mudge et al., 2013

What is a gene ?

please go to: https://genome.ucsc.edu/FAQ/FAQgenes.html

Perspective

Functional transcriptomics in the post-ENCODE era

Jonathan M. Mudge,[1] Adam Frankish, and Jennifer Harrow

Department of Informatics, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom

The last decade has seen tremendous effort committed to the annotation of the human genome sequence, most notably perhaps in the form of the ENCODE project. One of the major findings of ENCODE, and other genome analysis projects, is that the human transcriptome is far larger and more complex than previously thought. This complexity manifests, for example, as alternative splicing within protein-coding genes, as well as in the discovery of thousands of long noncoding RNAs. It is also possible that significant numbers of human transcripts have not yet been described by annotation projects, while existing transcript models are frequently incomplete. The question as to what proportion of this complexity is truly functional remains open, however, and this ambiguity presents a serious challenge to genome scientists. In this article, we will discuss the current state of human transcriptome annotation, drawing on our experience gained in generating the GENCODE gene annotation set. We highlight the gaps in our knowledge of transcript functionality that remain, and consider the potential computational and experimental strategies that can be used to help close them. We propose that an understanding of the true overlap between transcriptional complexity and functionality will not be gained in the short term. However, significant steps toward obtaining this knowledge can now be taken by using an integrated strategy, combining all of the experimental resources at our disposal.

23:1961–1973 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/13; www.genome.org

«Classical» versus modern view of a gene

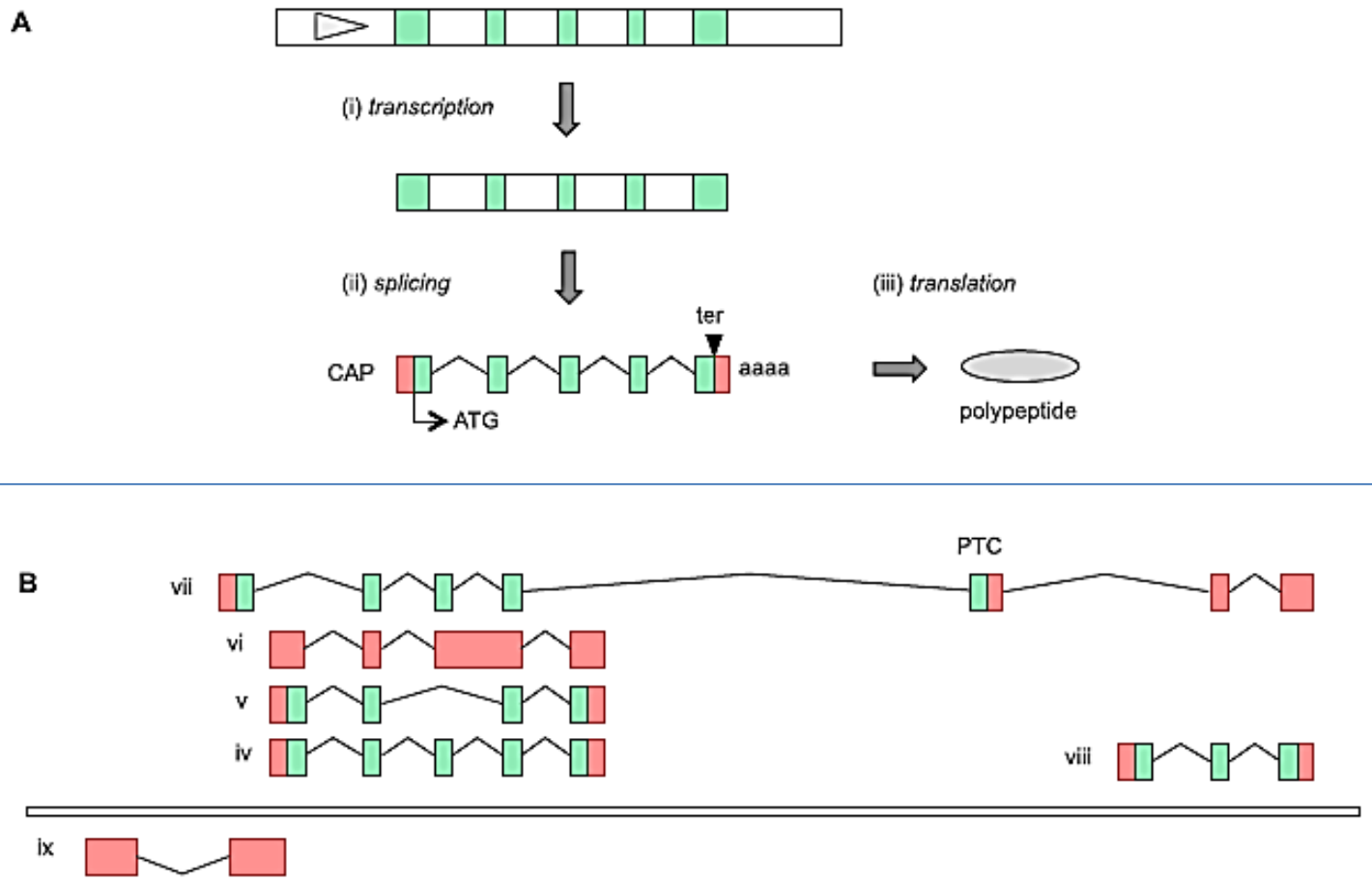
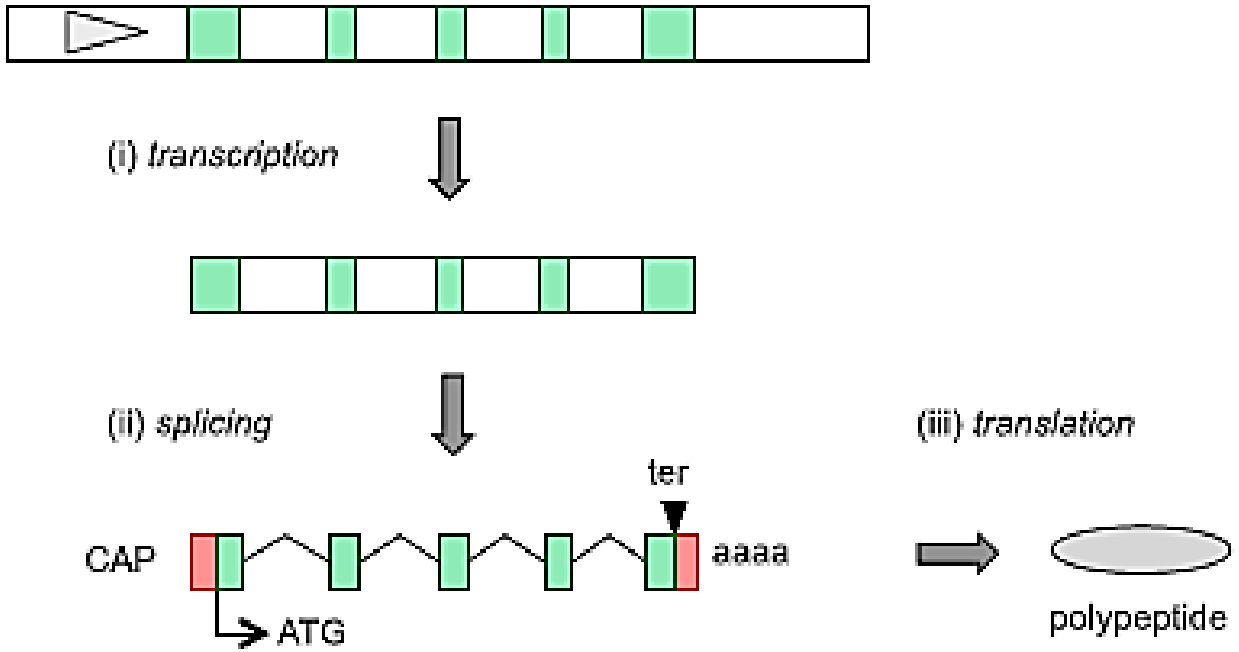


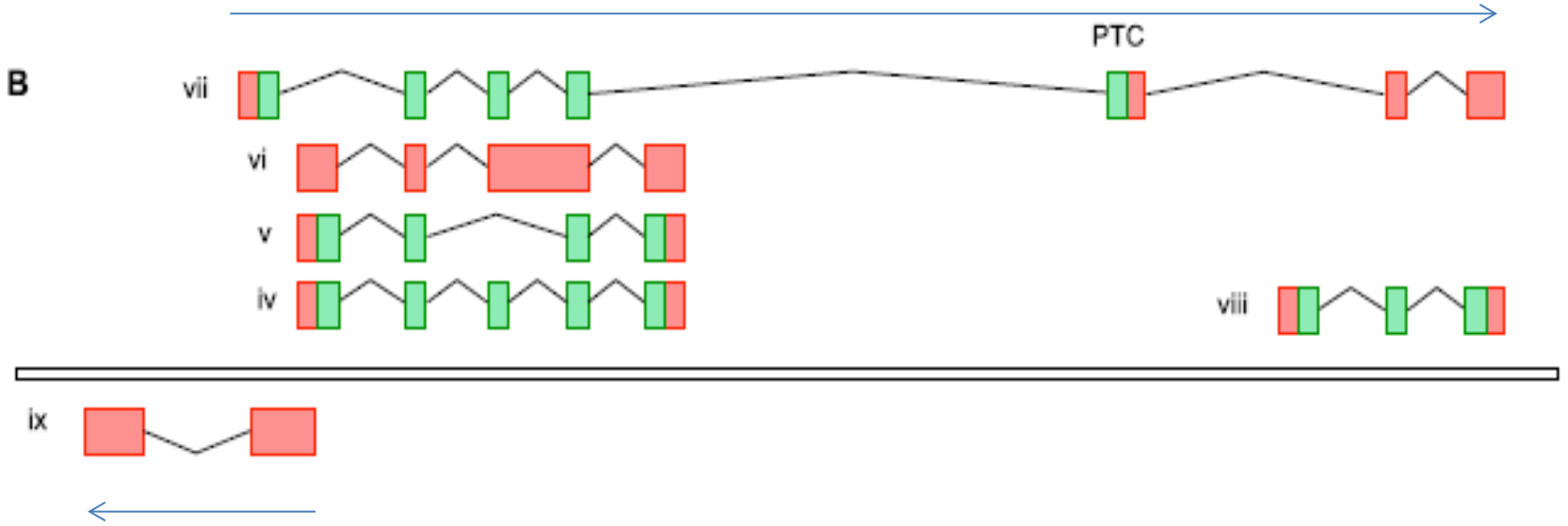
Figure 1. The evolving dogma of gene transcription.

From Mudge et al., 2013



(A) The **historical "central dogma"** of molecular biology. By this model, (i) transcription generates the primary transcript (exons in green, introns in white), with the initial interaction between the RNA polymerase complex and the genome being mediated by a promoter region (gray triangle). (ii) The introns of the primary transcript are removed by the spliceosome, and a mature mRNA is generated by 5' end capping (CAP) and polyadenylation (aaaa) (coding region [CDS] shown in green, untranslated 5' and 3' UTRs in red). (iii) The mRNA is translated into a polypeptide by the ribosome complex, with translation proceeding from the initiation codon (ATG) and ending at the termination codon (ter).

From Mudge et al., 2013

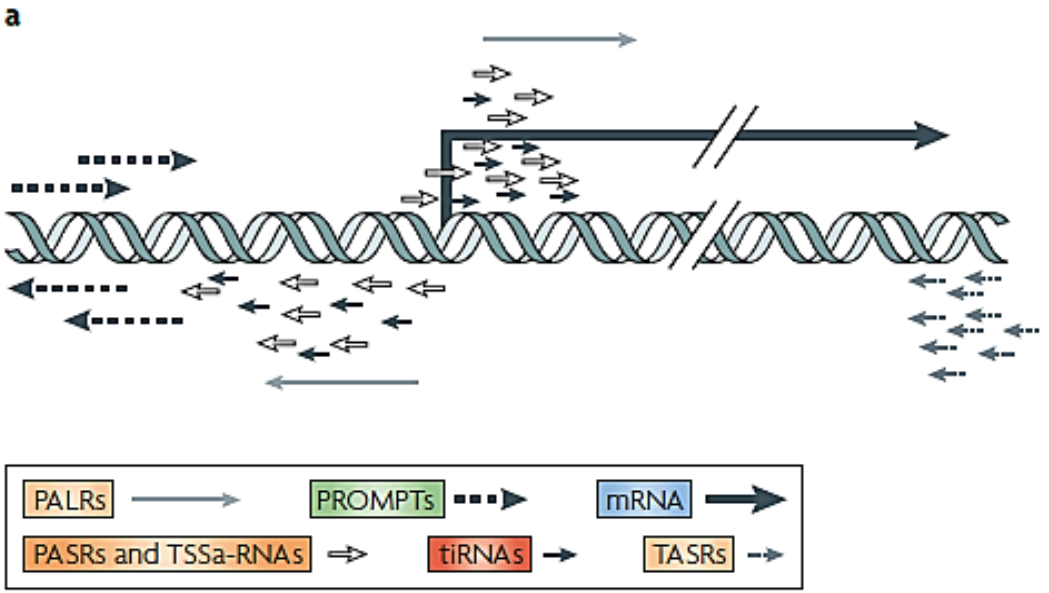


(B) An **updated model** reflecting a modern view of transcriptional complexity. Here, the same gene (iv) undergoes alternative splicing (AS), for example an exon skipping event that does not change the frame of the CDS (v); this event thus has the potential to generate an alternative protein isoform. However, products of AS cannot be assumed to be functional; this gene has generated a retained intron transcript (vi), perhaps due to the failure of the spliceosome to remove this intron. Further complexity comes from a read-through transcription event (vii), whereby a transcript is generated that also includes exons from a neighboring protein-coding locus (viii). In this example, the read-through transcript has an alternative first exon compared with the upstream gene that contains a potential alternative ATG codon, although the presence of a subsequent premature termination codon (PTC) prior to two splice junctions indicates that this transcript is likely subjected to the nonsense mediated decay (NMD) degradation pathway. Finally, model ix is a transcript that is antisense to the upstream gene; both loci are potentially generated under the control of a bidirectional promoter.

From Mudge et al., 2013

other misteries...

Unstable small RNA accompanying gene transcription



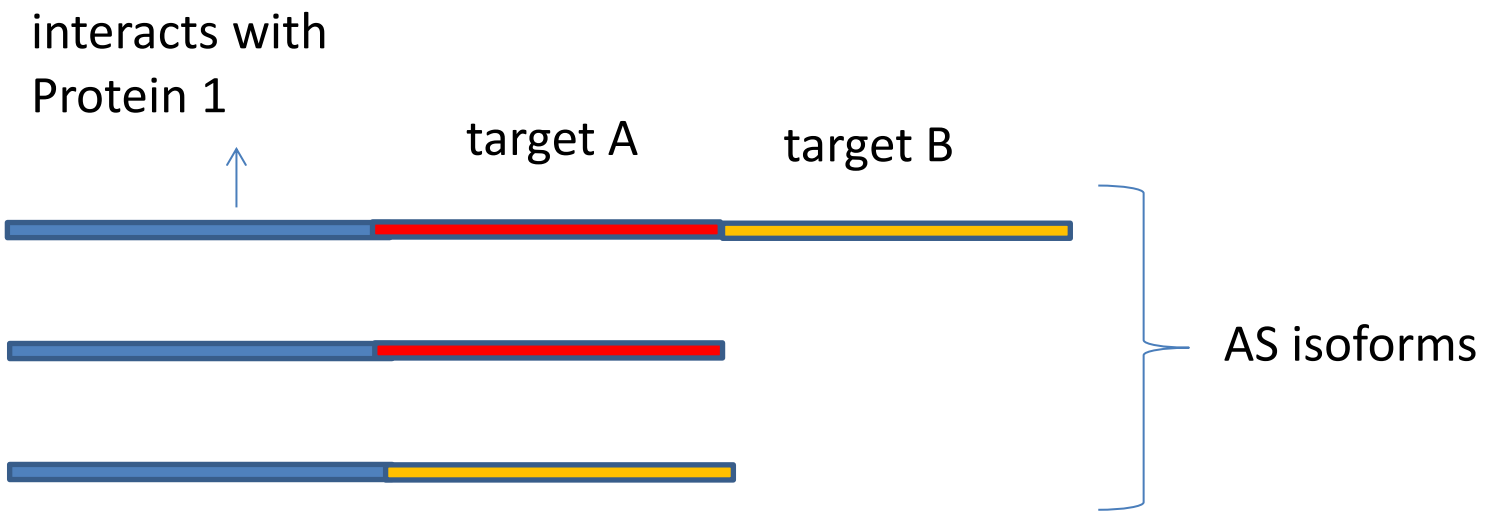
Short name of RNA classes	Full name of RNA classes	
PALRs	Promoter-associated long RNAs	Hundreds nt long RNAs spanning regions on proximal promoters to the first exon
PASRs	Promoter-associated short RNAs	20–70 nt long RNAs spanning regions around core promoters
TASRs	Termini-associated short RNAs	20–70 nt long RNAs spanning regions around transcription termination sites
PROMPTs	Promoter upstream transcripts	Unstable transcripts mapping 0.5–2 kb upstream the transcription starting sites
TSSa-RNAs	Transcription start sites antisense RNAs	RNAs, generally short and non-coding, generated from bidirectional activity of mammalian RNA Polymerase II
NRO-RNAs	Nuclear run-on assay derived RNAs	Short RNA detected by nuclear run-on assays, mapping 20 to 50 downstream to transcriptions starting sites of mRNAs
RE RNAs	Retrotransposon-derived RNAs	A heterogeneous class of RNAs which starting sites overlap retrotransposon elements
tiRNAs	Tiny transcription initiation RNAs	RNAs about 18 nt long, positioned about 20 bp after the transcription starting sites of highly expressed mRNAs

LncRNA undergo Alternative Splicing

They are capped and polyadenylated

What is the sense of making AS ?

Their function can be modulated by including/excluding certain parts.



Alternative Splicing of lncRNAs is guided by the same elements as protein-coding RNAs

However, while in protein-coding RNA alternative exons are few (average one-two on an average of 9 exons), lncRNA tend to have more alternatives.

Note that lncRNAs do not have the constraint of the coding sequence.

DSCAM-AS1

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000021.9 Chromosome 21 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank