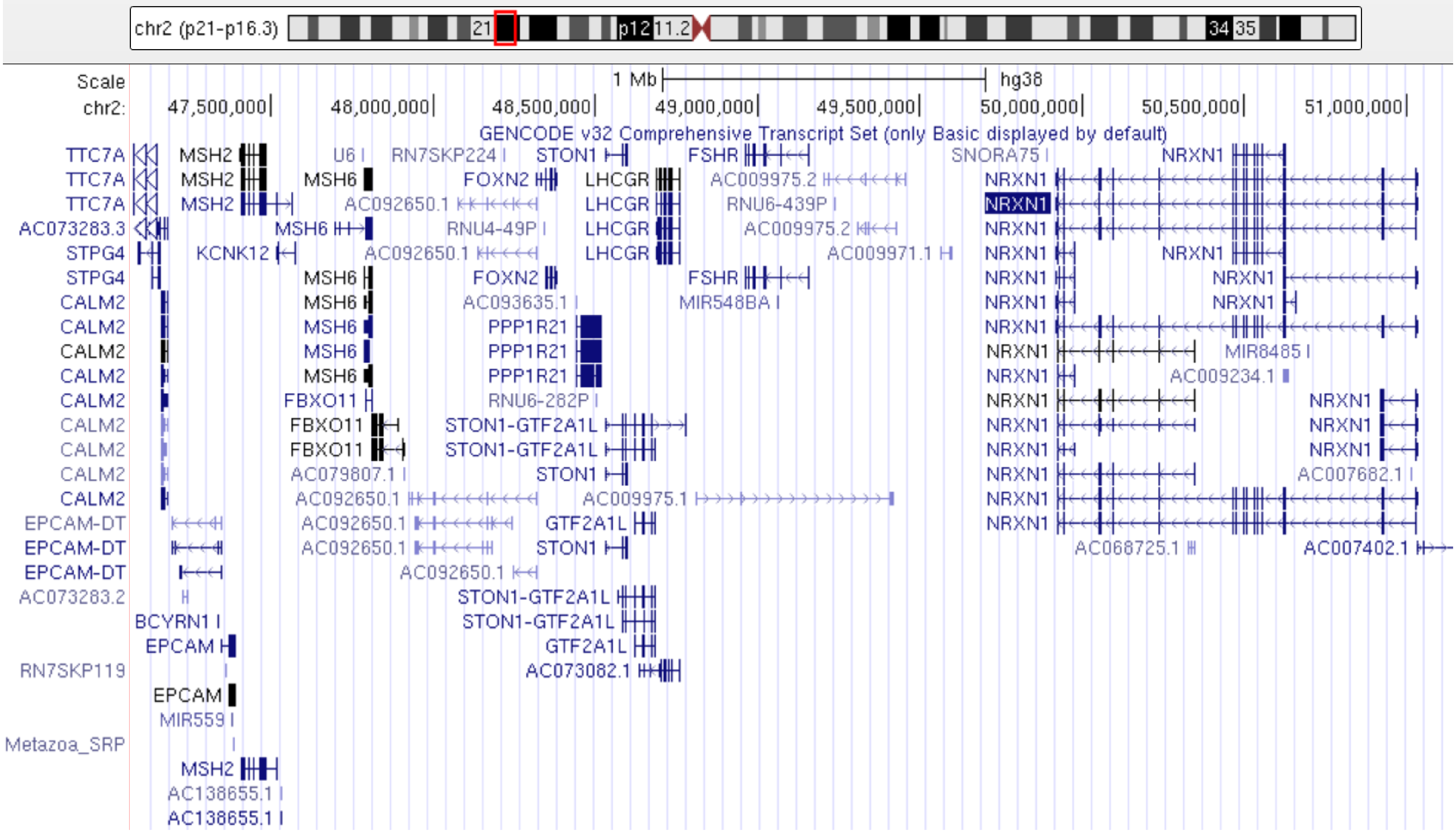
Mutations in SPACE:

Genes and consequences

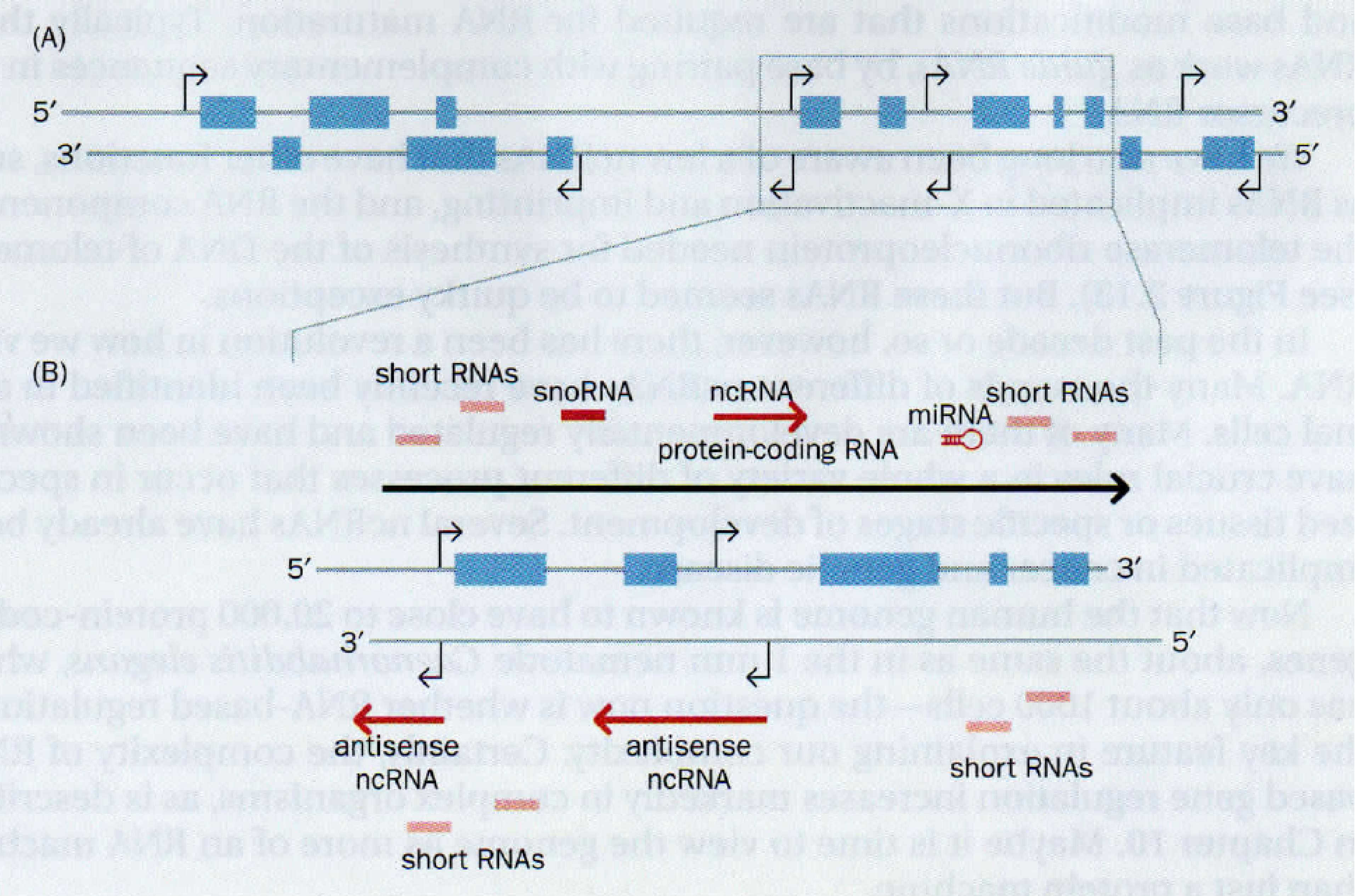
Lecture plan

- Overview of human genes structure and processing
- Alternative splicing
- Epigenetics. Chromosomal imprinting.
- Variant annotation. ENSEMBL Variant Effect Predictor: impact and consequences
- Protein-truncating and loss-of-function variants
- Missense variants, inframe indels
- Synonymous and regulatory variants
- Variant effect, dominant and recessive variants, gain- and loss-of-function

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

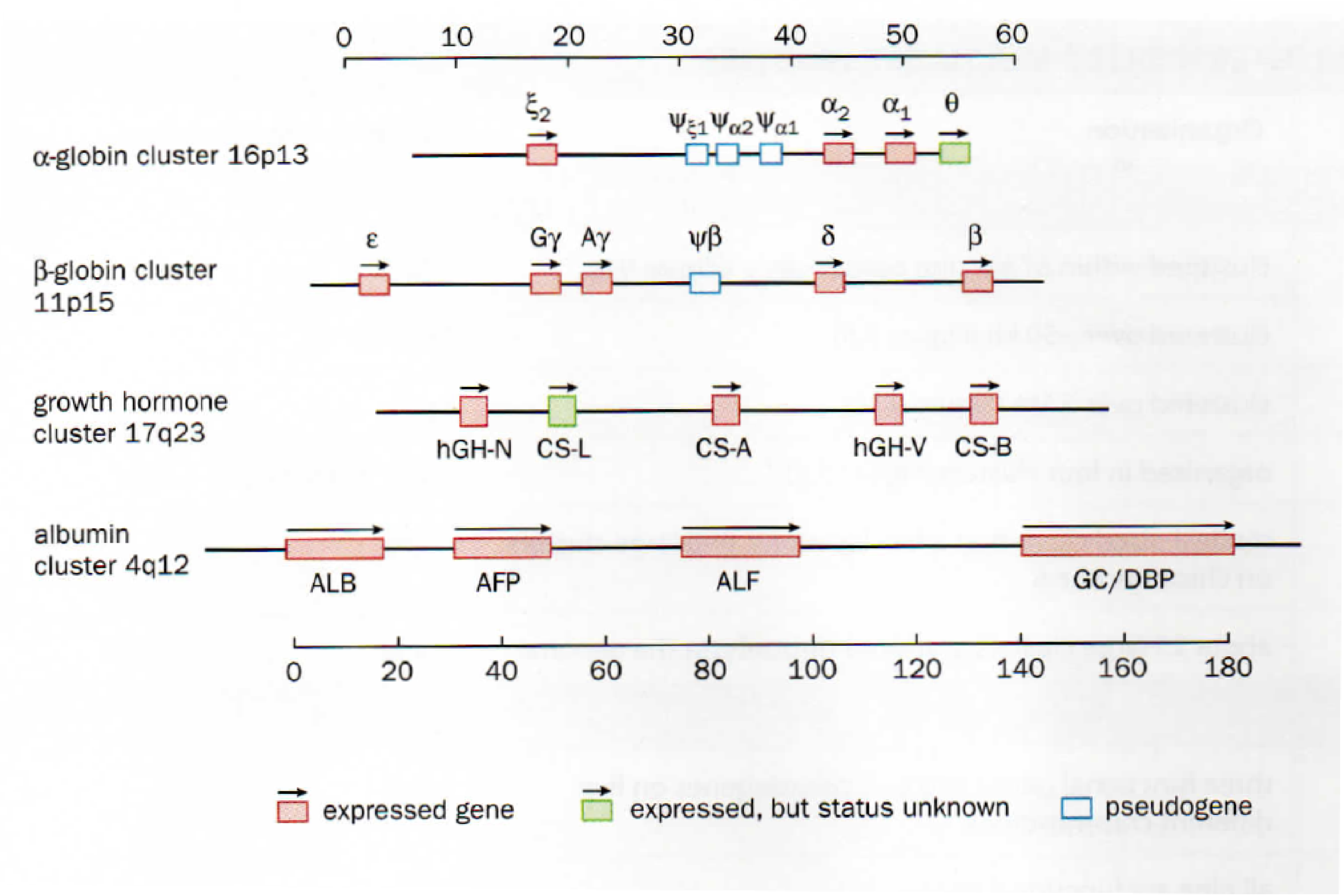


Blurring of gene boundaries



Strachan, Read – *Human Molecular Genetics*

Multigene families



Strachan, Read – *Human Molecular Genetics*

Multigene families

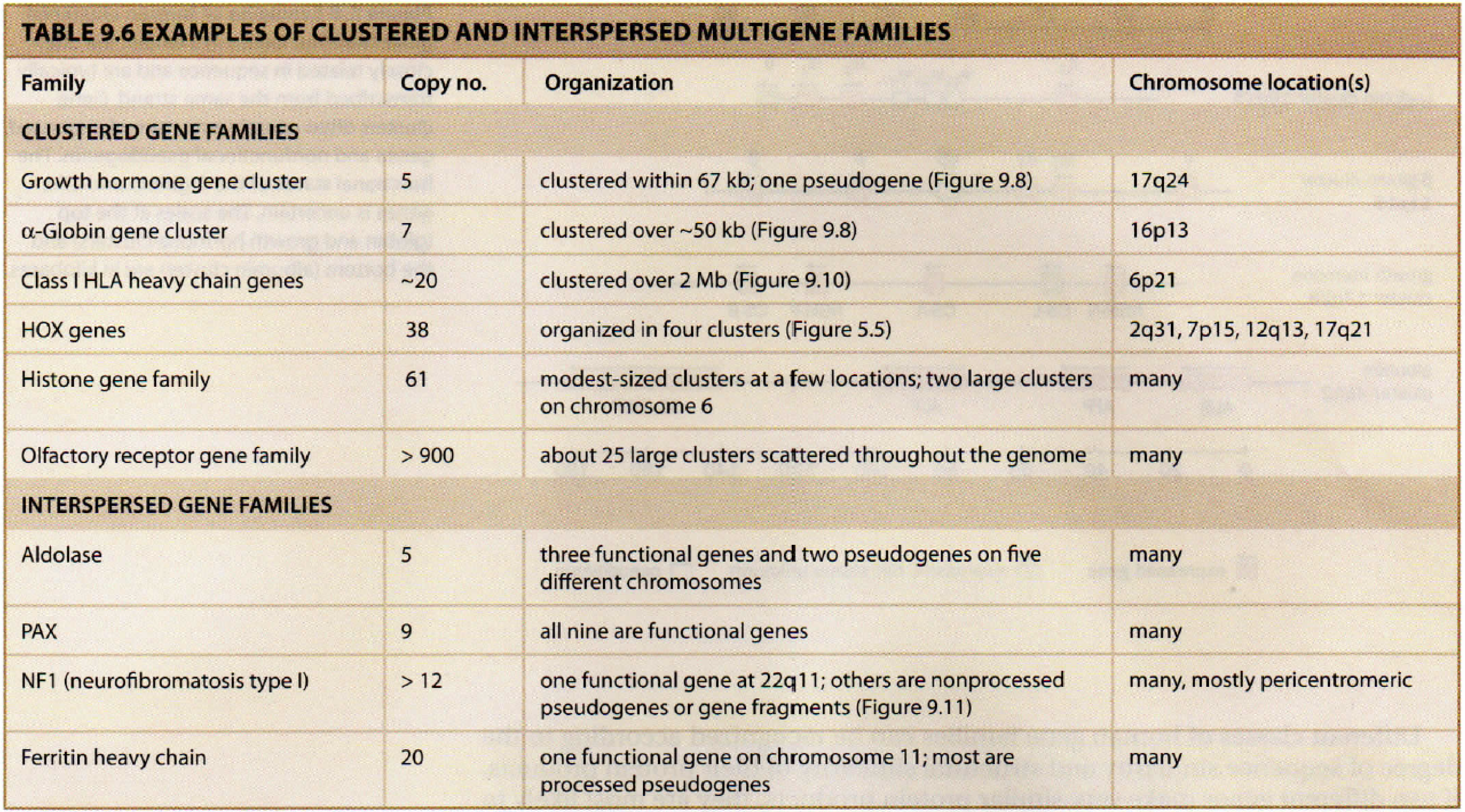


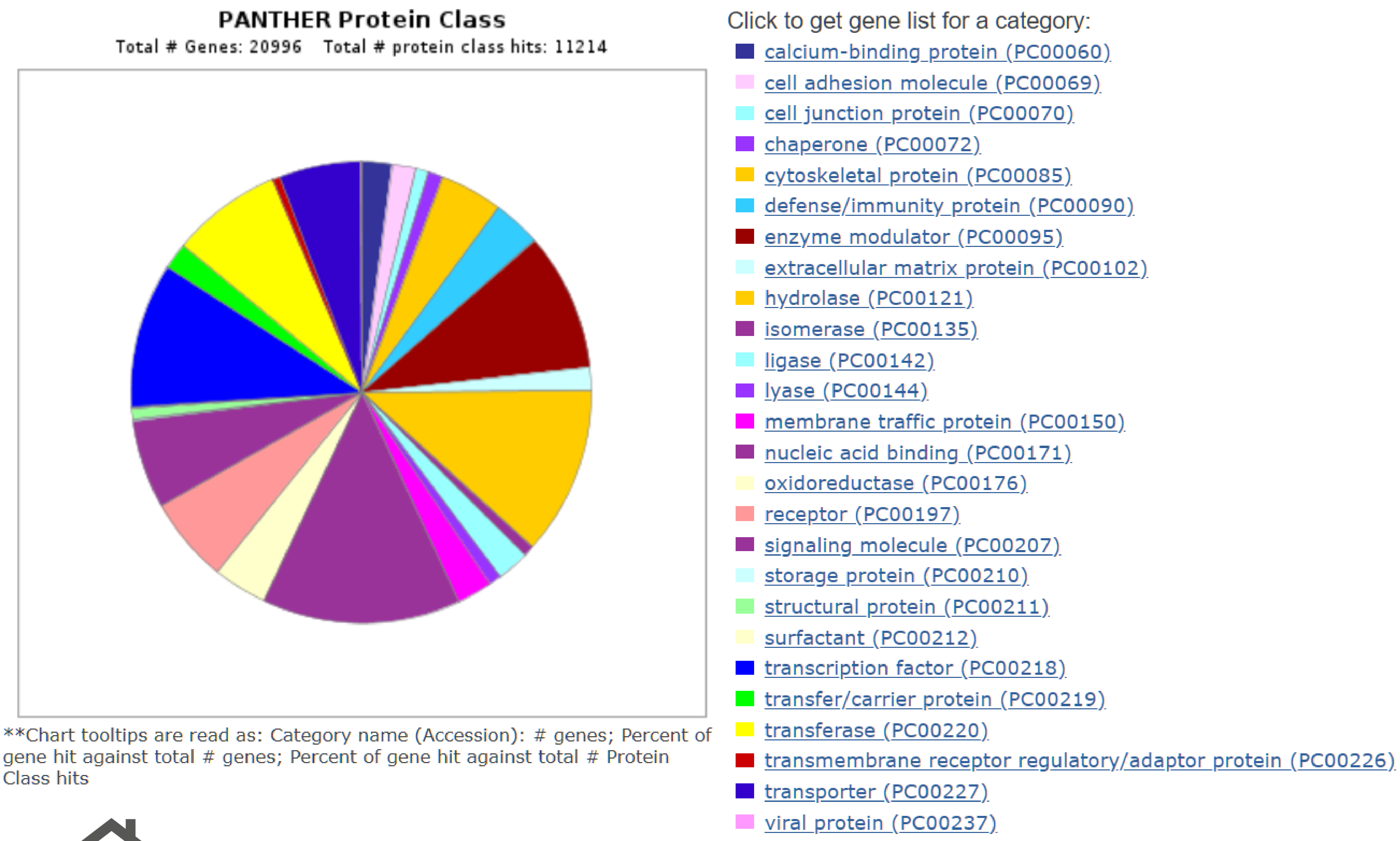
TABLE 9.6 EXAMPLES OF CLUSTERED AND INTERSPERSED MULTIGENE FAMILIES

Family	Copy no.	Organization	Chromosome location(s)
CLUSTERED GENE FAMILIES			
Growth hormone gene cluster	5	clustered within 67 kb; one pseudogene (Figure 9.8)	17q24
α-Globin gene cluster	7	clustered over ~50 kb (Figure 9.8)	16p13
Class I HLA heavy chain genes	~20	clustered over 2 Mb (Figure 9.10)	6p21
HOX genes	38	organized in four clusters (Figure 5.5)	2q31, 7p15, 12q13, 17q21
Histone gene family	61	modest-sized clusters at a few locations; two large clusters on chromosome 6	many
Olfactory receptor gene family	> 900	about 25 large clusters scattered throughout the genome	many
INTERSPERSED GENE FAMILIES			
Aldolase	5	three functional genes and two pseudogenes on five different chromosomes	many
PAX	9	all nine are functional genes	many
NF1 (neurofibromatosis type I)	> 12	one functional gene at 22q11; others are nonprocessed pseudogenes or gene fragments (Figure 9.11)	many, mostly pericentromeric
Ferritin heavy chain	20	one functional gene on chromosome 11; most are processed pseudogenes	many

Strachan, Read – *Human Molecular Genetics*

PANTHER Classification System

Human protein classes



PANTHER Protein Class

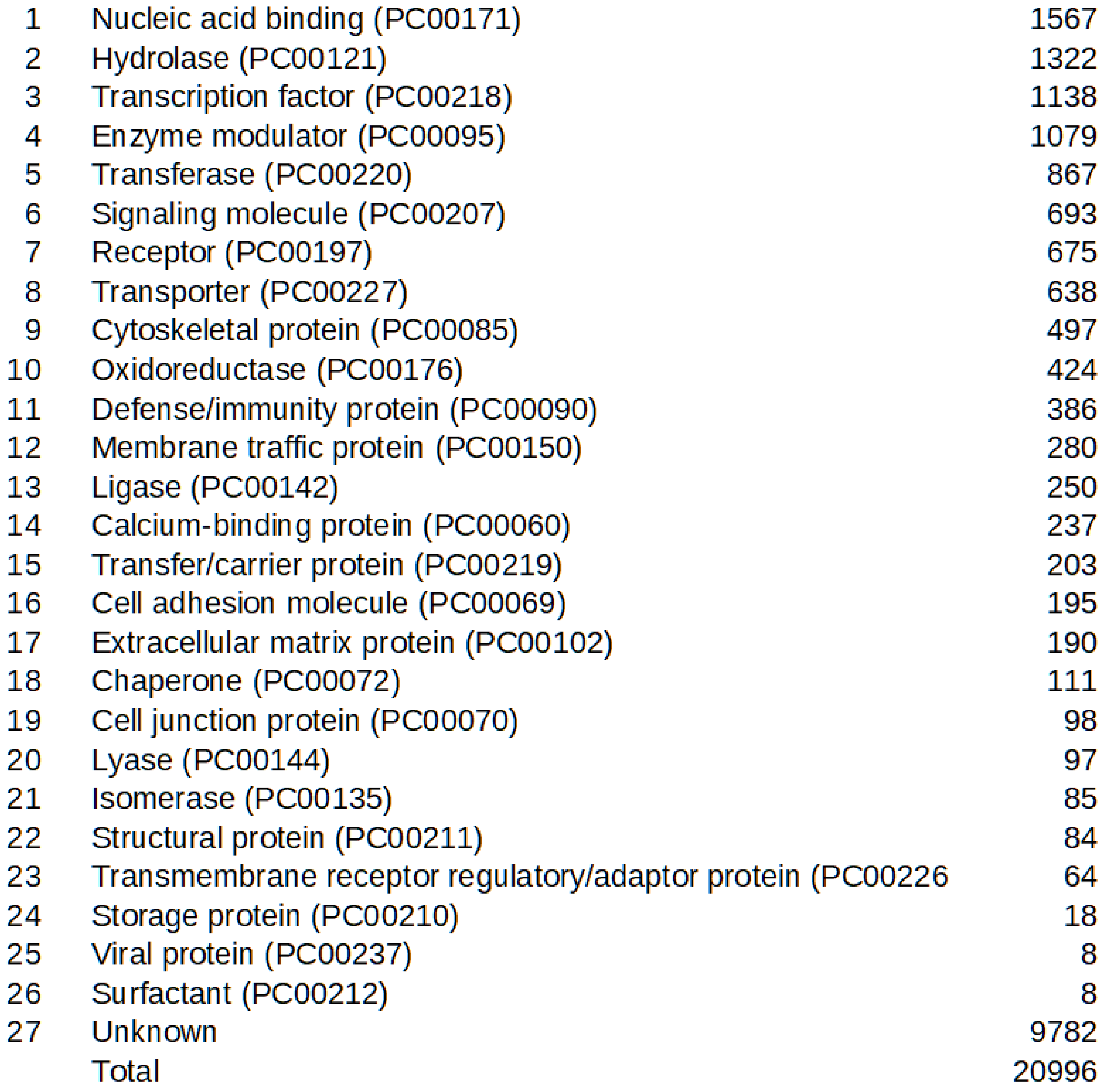
Total # Genes: 20996 Total # protein class hits: 11214

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Protein Class hits

Click to get gene list for a category:

- ■ calcium-binding protein (PC00060)
- ■ cell adhesion molecule (PC00069)
- ■ cell junction protein (PC00070)
- ■ chaperone (PC00072)
- ■ cytoskeletal protein (PC00085)
- ■ defense/immunity protein (PC00090)
- ■ enzyme modulator (PC00095)
- ■ extracellular matrix protein (PC00102)
- ■ hydrolase (PC00121)
- ■ isomerase (PC00135)
- ■ ligase (PC00142)
- ■ lyase (PC00144)
- ■ membrane traffic protein (PC00150)
- ■ nucleic acid binding (PC00171)
- ■ oxidoreductase (PC00176)
- ■ receptor (PC00197)
- ■ signaling molecule (PC00207)
- ■ storage protein (PC00210)
- ■ structural protein (PC00211)
- ■ surfactant (PC00212)
- ■ transcription factor (PC00218)
- ■ transfer/carrier protein (PC00219)
- ■ transferase (PC00220)
- ■ transmembrane receptor regulatory/adaptor protein (PC00226)
- ■ transporter (PC00227)
- ■ viral protein (PC00237)

Human protein classes



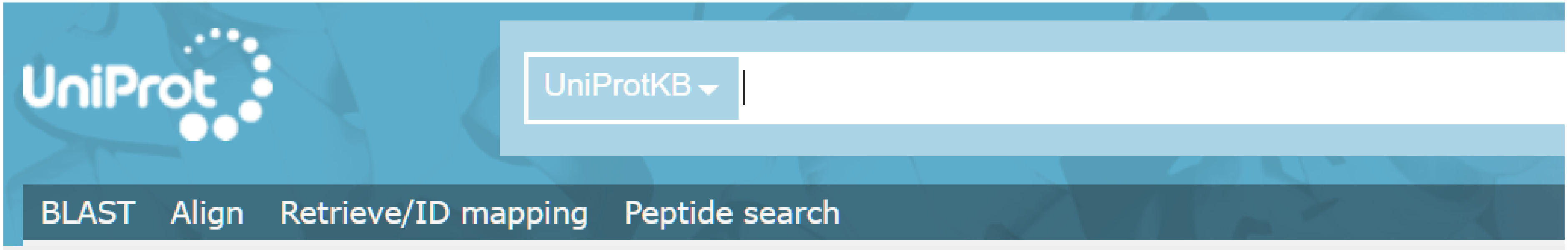
1	Nucleic acid binding (PC00171)	1567
2	Hydrolase (PC00121)	1322
3	Transcription factor (PC00218)	1138
4	Enzyme modulator (PC00095)	1079
5	Transferase (PC00220)	867
6	Signaling molecule (PC00207)	693
7	Receptor (PC00197)	675
8	Transporter (PC00227)	638
9	Cytoskeletal protein (PC00085)	497
10	Oxidoreductase (PC00176)	424
11	Defense/immunity protein (PC00090)	386
12	Membrane traffic protein (PC00150)	280
13	Ligase (PC00142)	250
14	Calcium-binding protein (PC00060)	237
15	Transfer/carrier protein (PC00219)	203
16	Cell adhesion molecule (PC00069)	195
17	Extracellular matrix protein (PC00102)	190
18	Chaperone (PC00072)	111
19	Cell junction protein (PC00070)	98
20	Lyase (PC00144)	97
21	Isomerase (PC00135)	85
22	Structural protein (PC00211)	84
23	Transmembrane receptor regulatory/adaptor protein (PC00226	64
24	Storage protein (PC00210)	18
25	Viral protein (PC00237)	8
26	Surfactant (PC00212)	8
27	Unknown	9782
	Total	20996

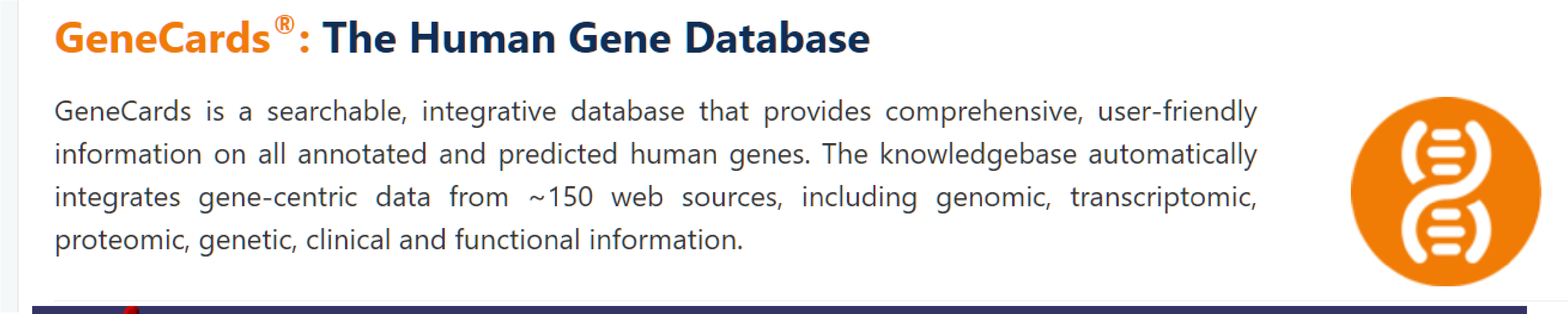
Exercise: think of appropriate questions

HGNC

HUGO Gene Nomenclature Committee

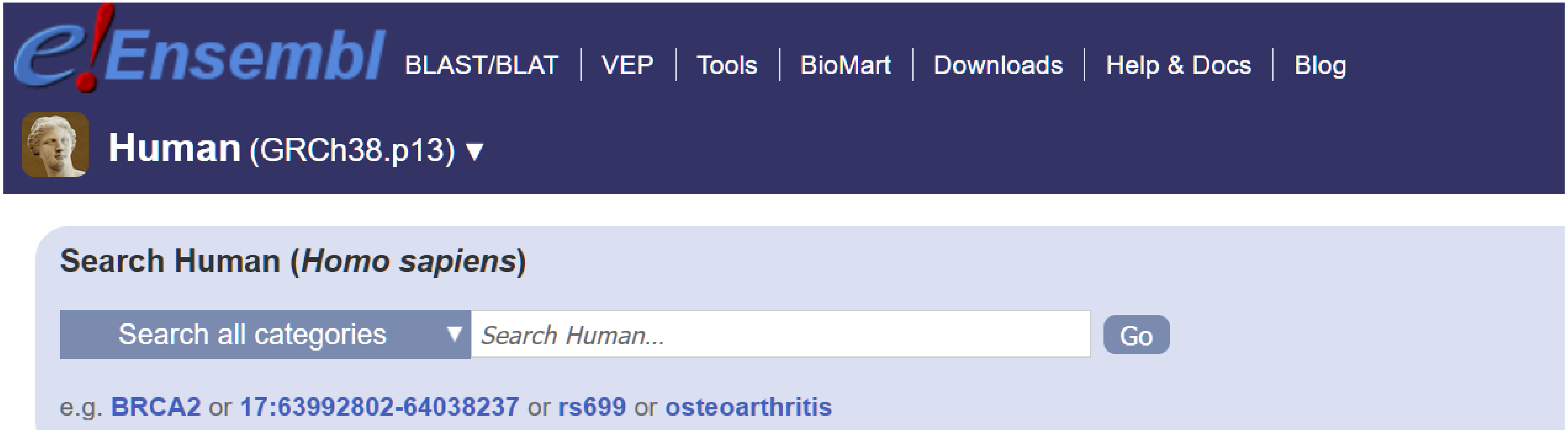
The resource for approved human gene nomenclature





GeneCards®: The Human Gene Database

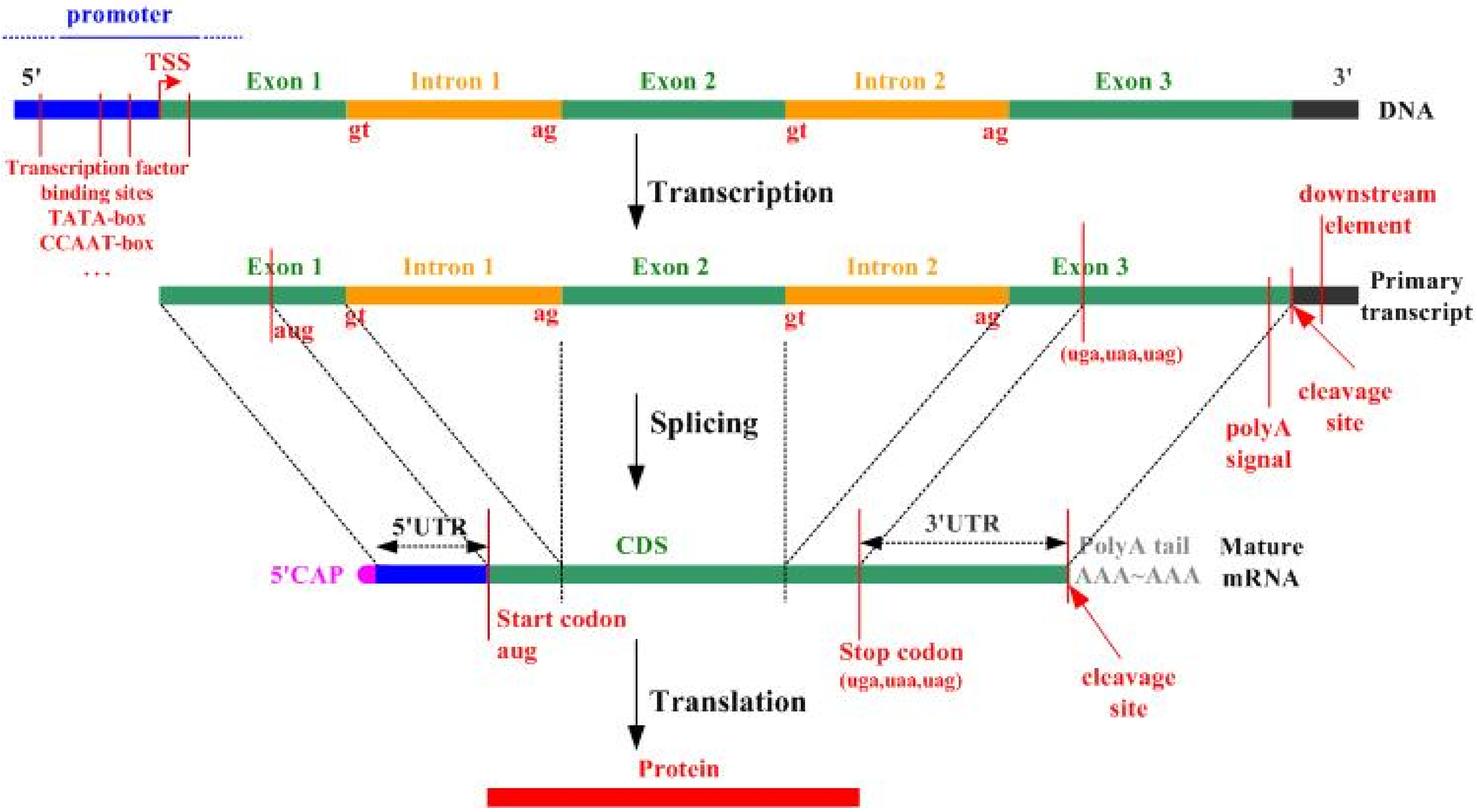
GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledgebase automatically integrates gene-centric data from ~150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.



Search Human (*Homo sapiens*)

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Human gene structure and processing



Note: CDS (coding sequence) vs. mRNA, splicing sites, stop and start codons

Exercise: draw a typical human gene

Human gene structure and processing

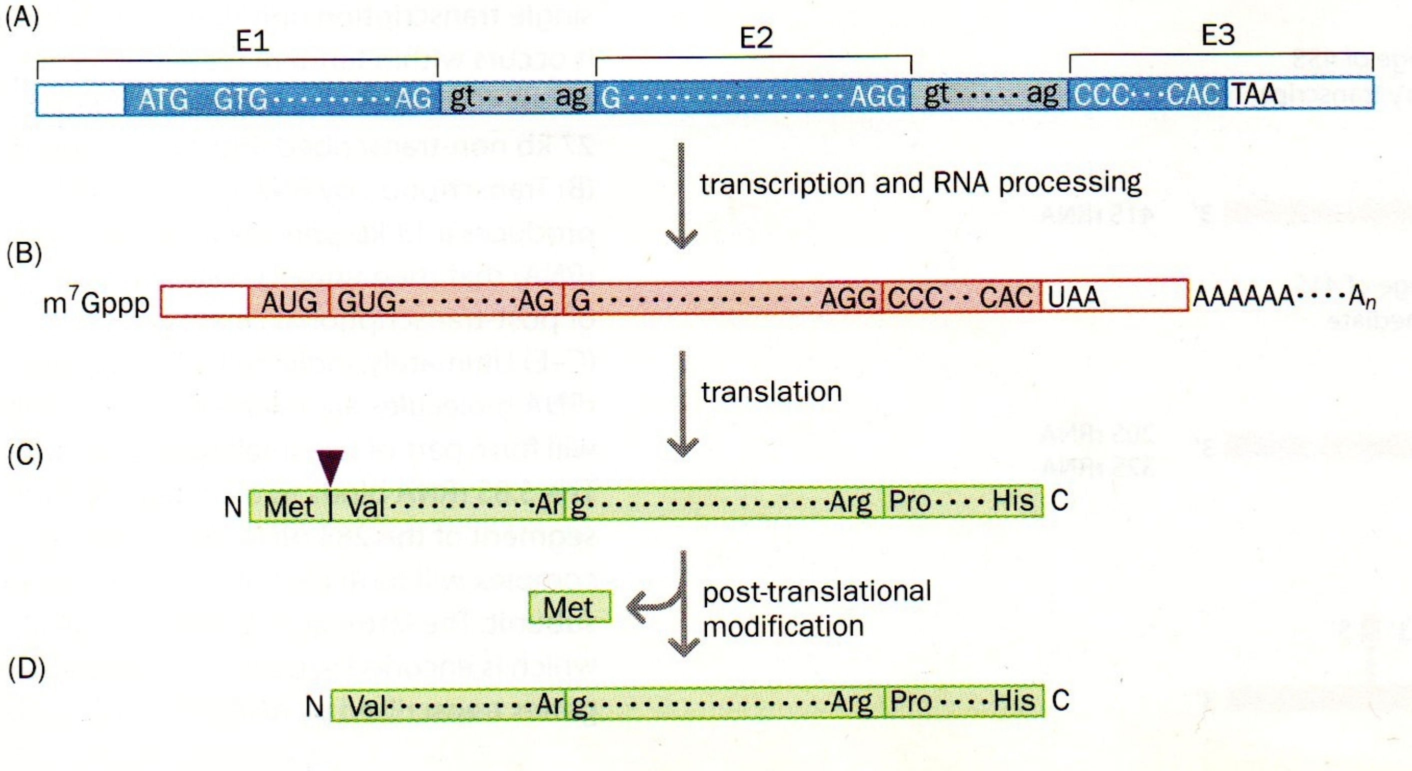

(A)
E1
E2
E3
ATG GTG··········AG gt·····ag G················AGG gt·····ag CCC···CAC TAA
transcription and RNA processing
(B)
m7Gppp AUG GUG·········AG G················AGG CCC··CAC UAA AAAAAA····An
translation
(C)
N Met Val···········Ar g················Arg Pro····His C
Met
post-translational modification
(D)
N Val···········Ar g················Arg Pro····His C

Figure 1.23 Transcription and translation of the human β-globin . (A) The β-globin gene comprises three exons (El-E3) and two introns. The 5'-end sequence of El and the 3' end sequence of E3 are noncoding sequences (unshaded sections). (B) These sequences are transcribed and so occur at the 5' and 3' ends (unshaded sections) of the β-globin mRNA that emerges from RNA processing. (C) Some codons can be specified by bases that are separated by an intron. The Arg104 is encoded by the last three nucleotides (AGG) of exon 2 but the Arg30 is encoded by an AGG codon whose first two bases are encoded by the last two nucleotides of exon 1 and whose third base is encoded by the first nucleotide of exon 2. (D) During post-translational modification the 147·amino acid precursor polypeptide undergoes cleavage to remove ils *N*-terminal methionine residue, to generate the mature 146-residue β-globin protein. The flanking *N* and *C* symbols to the left and right, respectively, in (C) and (D) depict the *N*-terminus and *C*-terminus.

Strachan, Read – *Human Molecular Genetics*

Human gene structure and processing

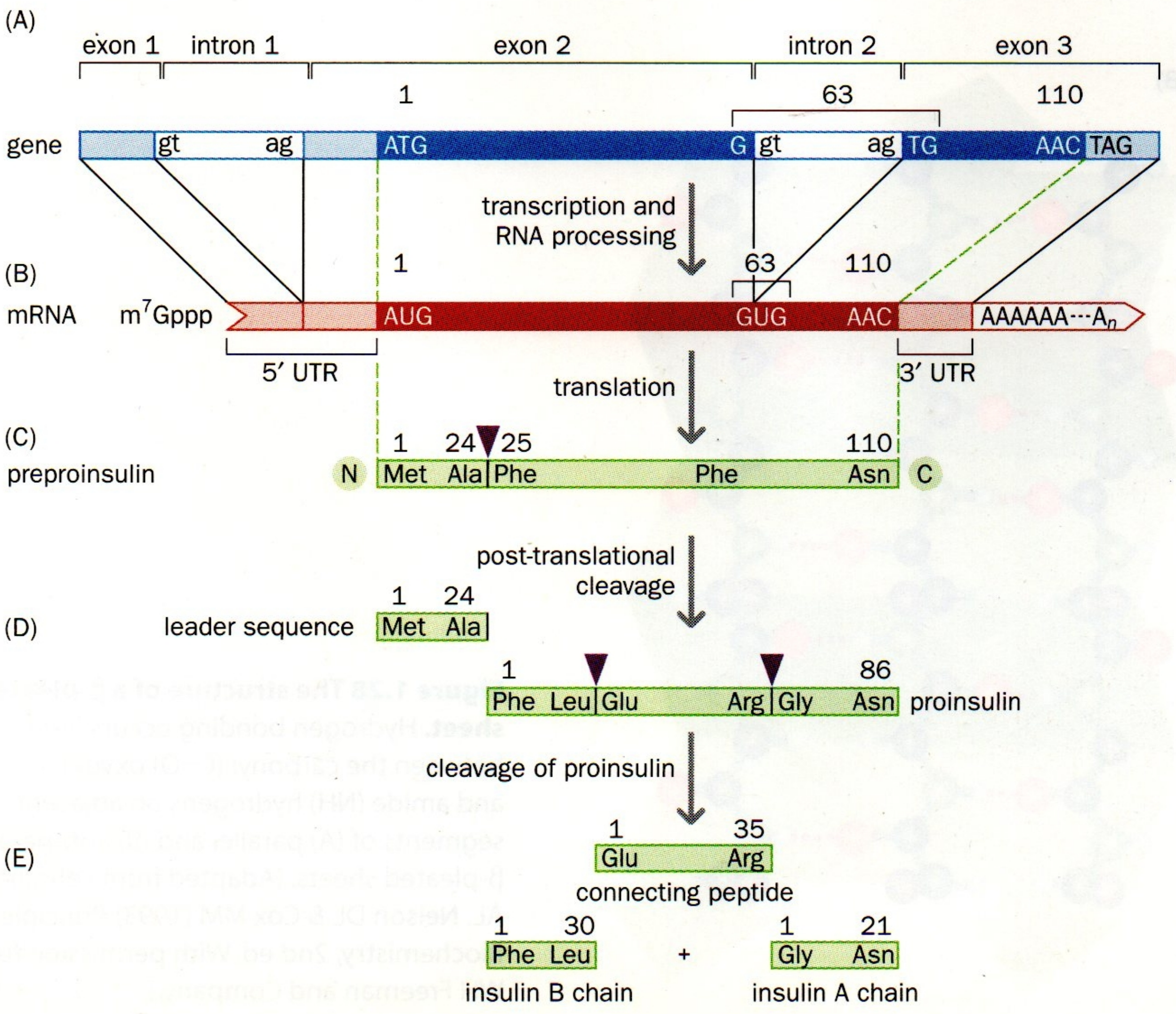


Figure 1.26 Insulin synthesis involves multiple post-translational cleavages of polypeptide precursors. (A) The human insulin gene comprises three exons and two introns. The coding sequence (the part that will be used to make polypeptide) is shown in deep blue. It is confined to the 3′ sequence of exon 2 and the 5′ sequence of exon 3. (B) Exon 1 and the 5′ part of exon 2 specify the 5′ untranslated region (5′ UTR), and the 3′ end of exon 3 specifies the 3′ UTR. The UTRs are transcribed and so are present at the ends of the mRNA. (C) A primary translation product, preproinsulin, has 110 residues and is cleaved to give (D) a 24-residue N-terminal *leader sequence* (that is required for the protein to cross the cell membrane but is thereafter discarded) plus an 86-residue proinsulin precursor. (E) Proinsulin is cleaved to give a central segment (the connecting peptide) that may maintain the conformation of the A and B chains of insulin before the formation of their interconnecting covalent disulfide bridges (see Figure 1.29).

Examples of post-translational processing

Strachan, Read – *Human Molecular Genetics*

Human gene structure and processing

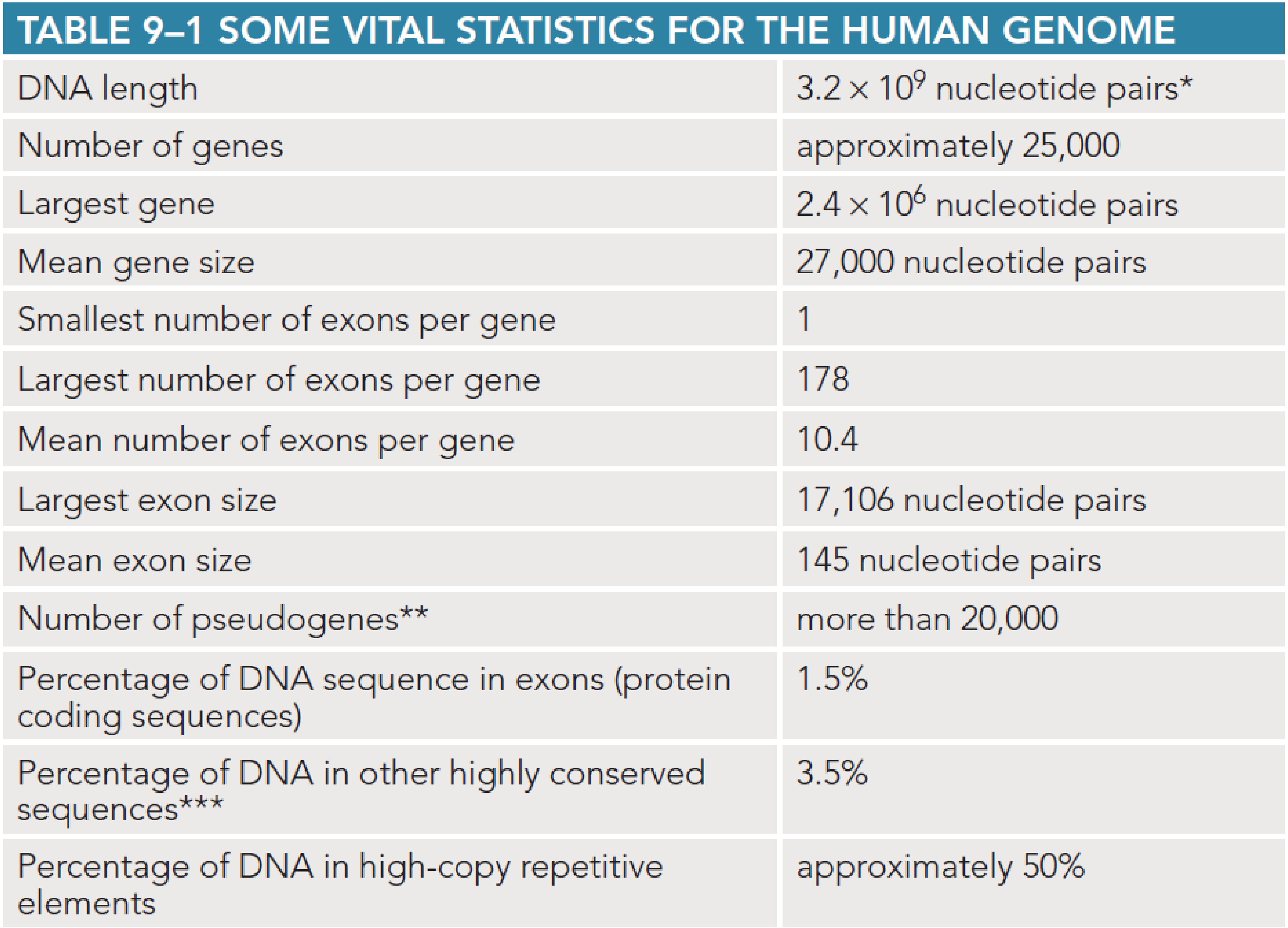


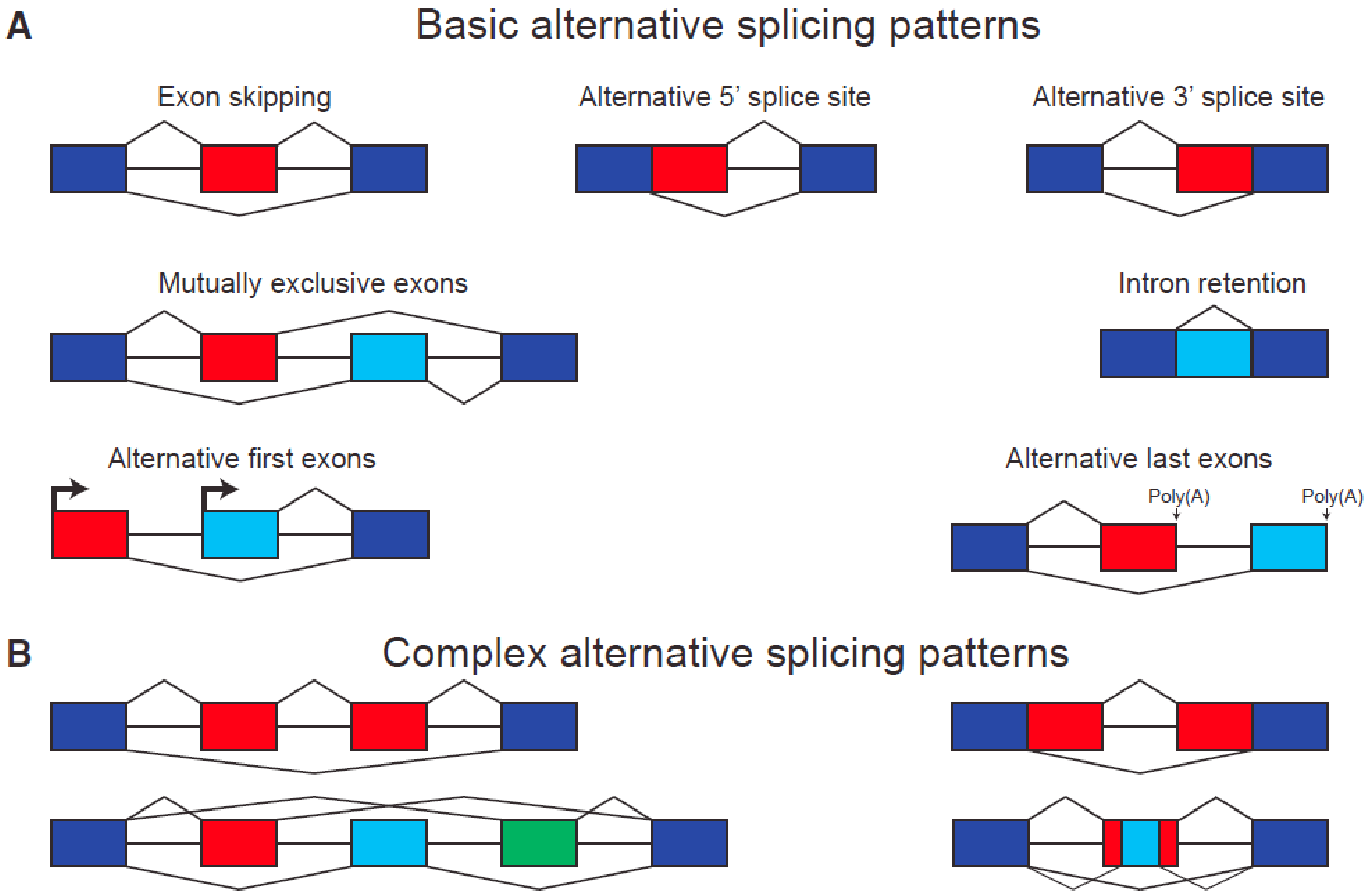
TABLE 9–1 SOME VITAL STATISTICS FOR THE HUMAN GENOME

DNA length	3.2×10^9 nucleotide pairs*
Number of genes	approximately 25,000
Largest gene	2.4×10^6 nucleotide pairs
Mean gene size	27,000 nucleotide pairs
Smallest number of exons per gene	1
Largest number of exons per gene	178
Mean number of exons per gene	10.4
Largest exon size	17,106 nucleotide pairs
Mean exon size	145 nucleotide pairs
Number of pseudogenes**	more than 20,000
Percentage of DNA sequence in exons (protein coding sequences)	1.5%
Percentage of DNA in other highly conserved sequences***	3.5%
Percentage of DNA in high-copy repetitive elements	approximately 50%

Q: what gene (exon) is the largest?

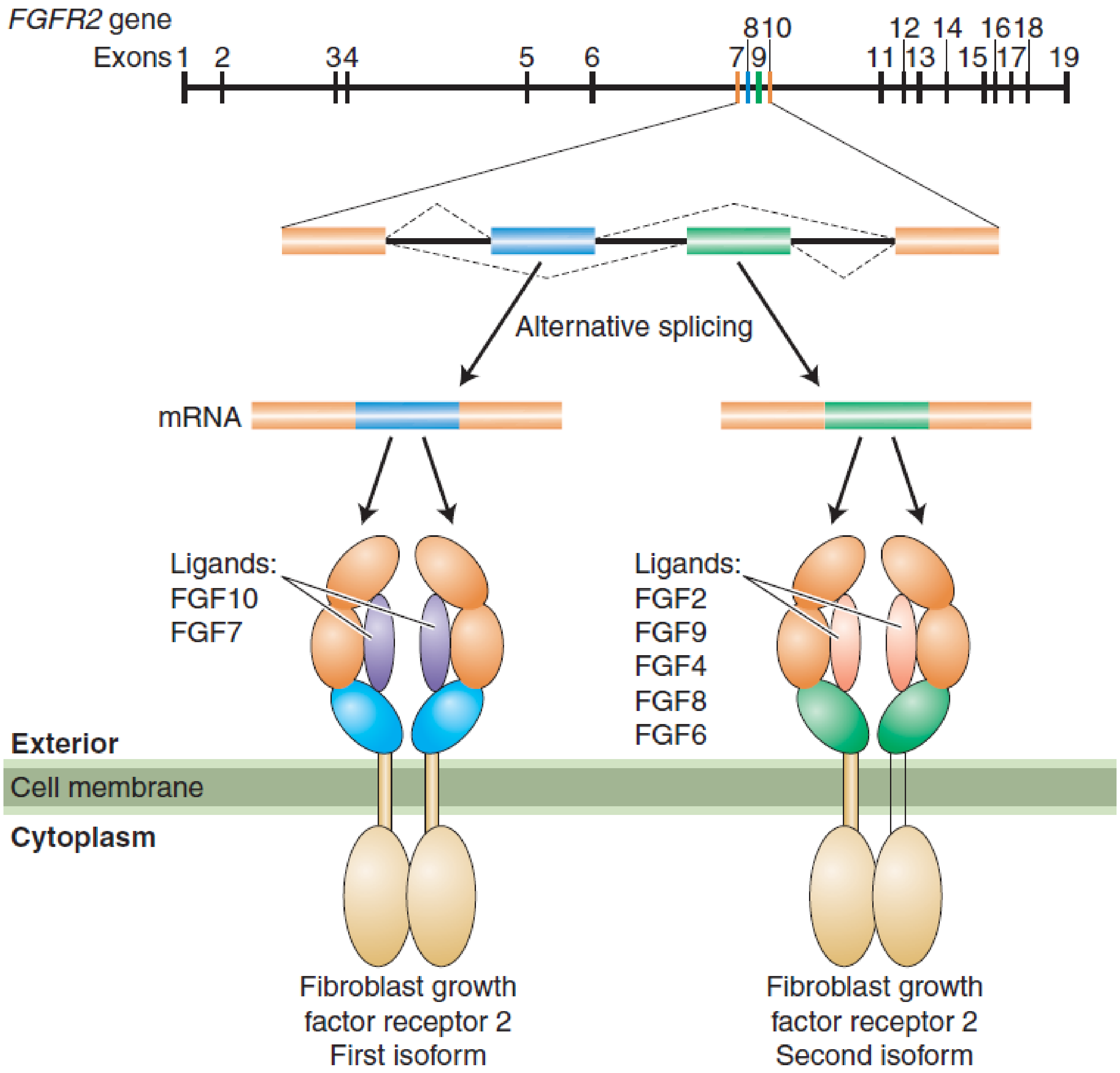
Alberts – *Essential Cell Biology*

Alternative splicing of human genes


A
Basic alternative splicing patterns
Exon skipping
Alternative 5' splice site
Alternative 3' splice site
Mutually exclusive exons
Intron retention
Alternative first exons
Alternative last exons
Poly(A)
Poly(A)
B
Complex alternative splicing patterns

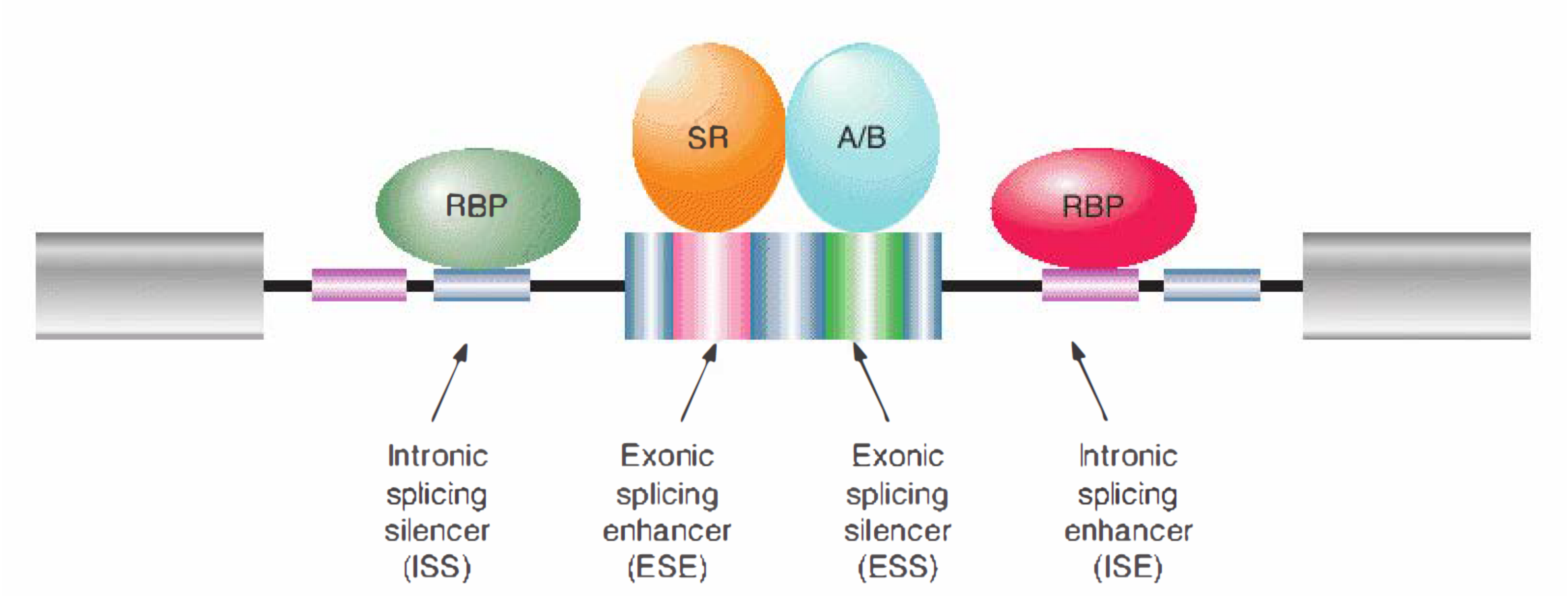
Park (2018) *Am J Hum Genet*

Alternative splicing of human genes

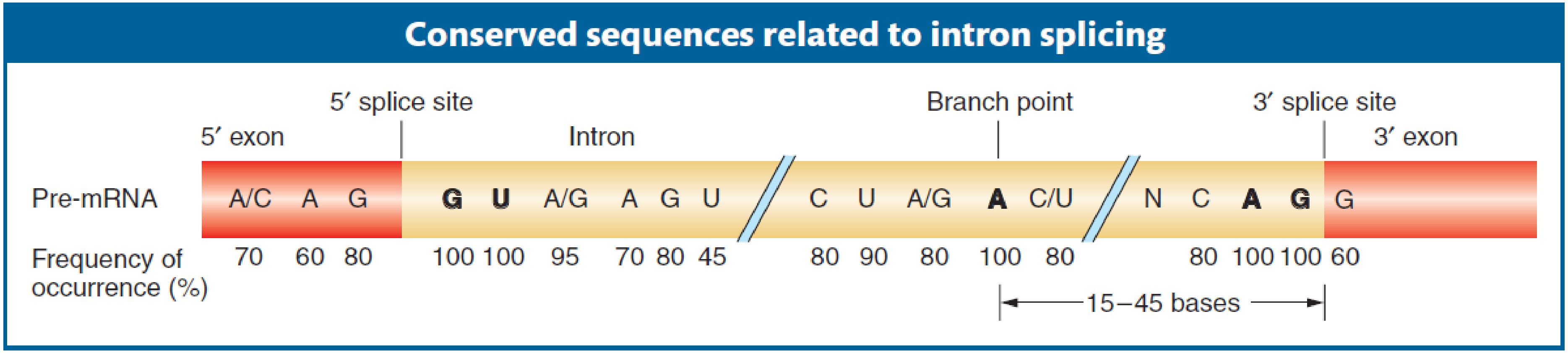


Griffiths -- *Introduction to Genetic Analysis*

Alternative splicing of human genes


SR
A/B
RBP
RBP
Intronic splicing silencer (ISS)
Exonic splicing enhancer (ESE)
Exonic splicing silencer (ESS)
Intronic splicing enhancer (ISE)

Lewin – *Genes XI*



Griffiths -- *Introduction to Genetic Analysis*

Alternative splicing of human genes

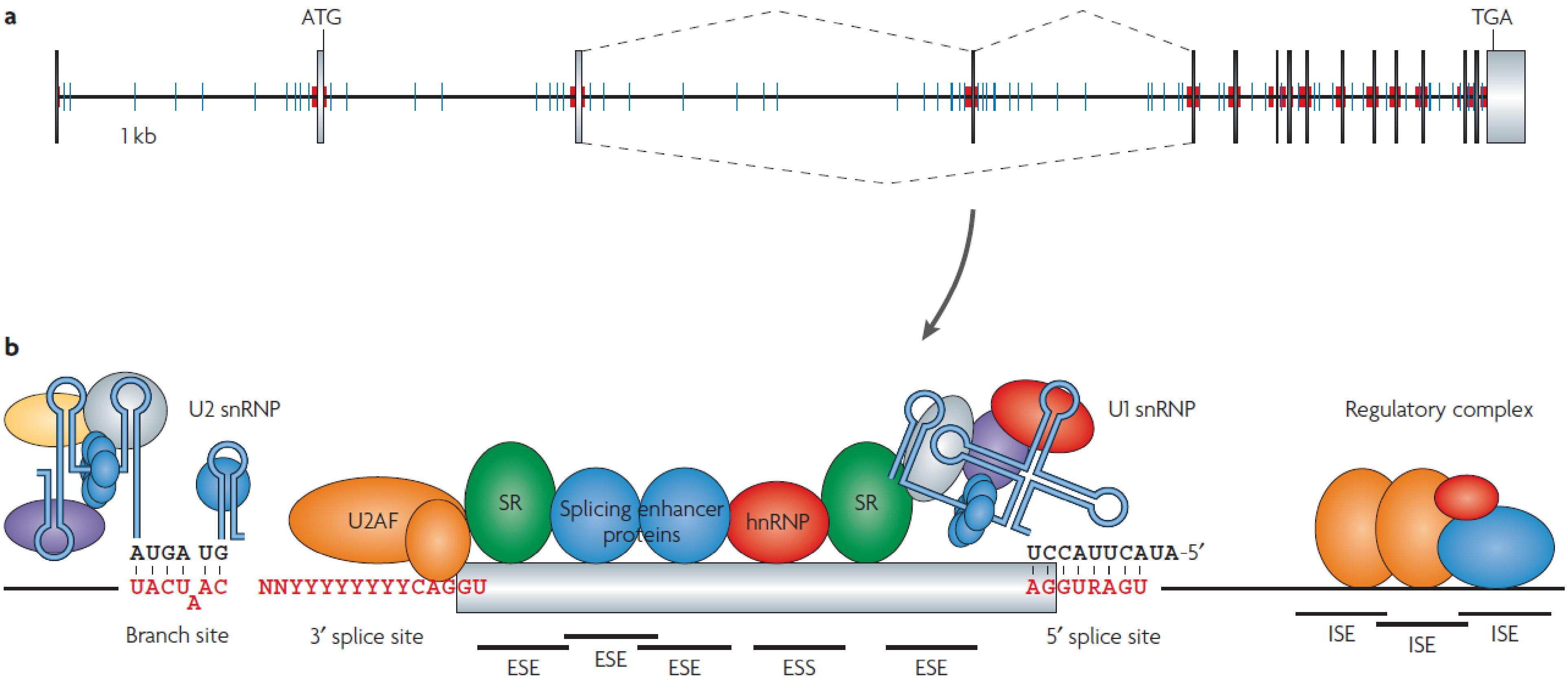


Figure 1 | **The splicing code. a** | A pre-mRNA as it might appear to the spliceosome. Red indicates consensus splice site sequences at the intron–exon boundaries. Blue indicates additional intronic cis-acting elements that make up the splicing code. **b** | cis-elements within and around an alternative exon are required for its recognition and regulation. The 5′ splice site and branch site serve as binding sites for the RNA components of U1 and U2 small nuclear ribonucleoprotein (snRNPs), respectively. This RNA:RNA base pairing determines the precise joining of exons at the correct nucleotides. Mutations in the pre-mRNA that disrupt this base pairing decrease the efficiency of exon recognition. Exons and introns contain diverse sets of enhancer and suppressor elements that refine bone fide exon recognition. Some exon splicing enhancers (ESEs) bind SR proteins and recruit and stabilize binding of spliceosome components such as U2AF. Exon splicing suppressors (ESSs) bind protein components of heterogeneous nuclear ribonucleoproteins (hnRNP) to repress exon usage. Some intronic splicing enhancers (ISEs) bind auxiliary splicing factors that are not normally associated with the spliceosome to regulate alternative splicing.

Wang (2007) *Nat Rev Genet*

Alternative splicing of human genes

- ENSEMBL GRCh38 v.99, protein-coding genes and transcripts:
  - 1 transcript: 22.2% (no alternative splicing)
  - 2-5 transcripts: 52.9%
  - >5 transcripts: 24.9%
  - More than 75 transcripts: *ADGRG1, ANK2, KCNMA1, MAPK10, NDRG2, PAX6, TCF4*

- Longest transcript designated as **canonical** (≠ most biologically relevant)
- AS contribution to proteome complexity and transcript functionality is still debated: transcripts ≠ protein isoforms
- AS transcripts that introduce premature stop codon are subject to NMD (**nonsense-mediated decay**)
- Microexons (3-30 nt): misregulated in autistic brain (Irimia (2014) *Cell*).

Aberrant splicing in disease

- **Cis-acting splicing mutations**: disruption of the splicing code, **15-60% of human disease mutations** (Wang 2007 *Nat Rev Genet*)

Examples: synonymous mutations in *CFTR* ⟹ cystic fibrosis;

Splice site mutations in *MITF* ⟹ Waardenburg syndrome type 2 (WS2), a dominantly inherited syndrome of hearing loss and pigmentary disturbances

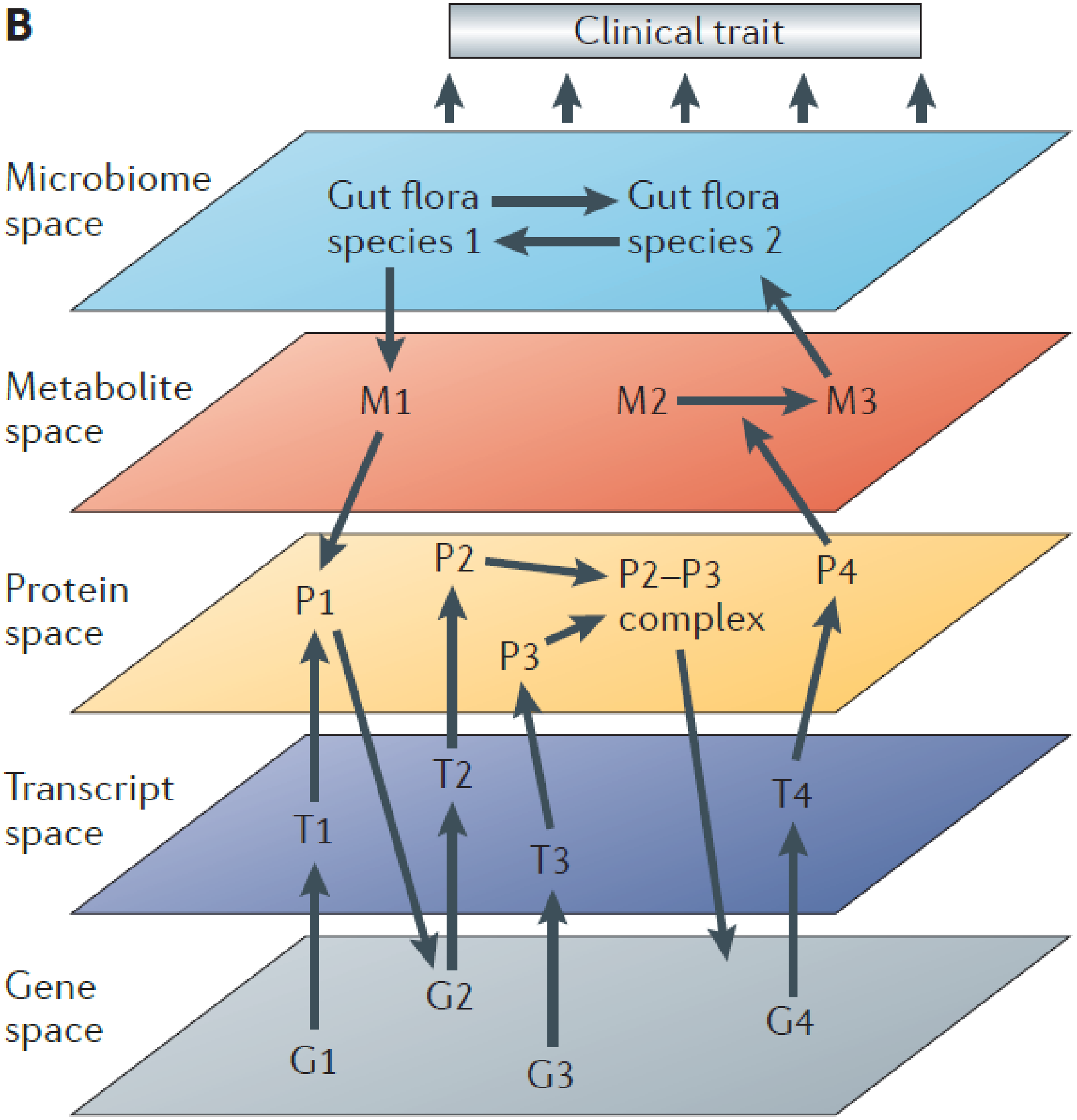
- **Trans-acting mutations**: disruption of the splicing RNA-protein machinery.

Example: mutations in *SMN1* ⟹ loss of snRNP production ⟹ spinal muscular atrophy (SMA). Nusinersen, an antisense oligonucleotide drug for correcting splicing in spinal muscular atrophy.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet*. 102, 11–26.

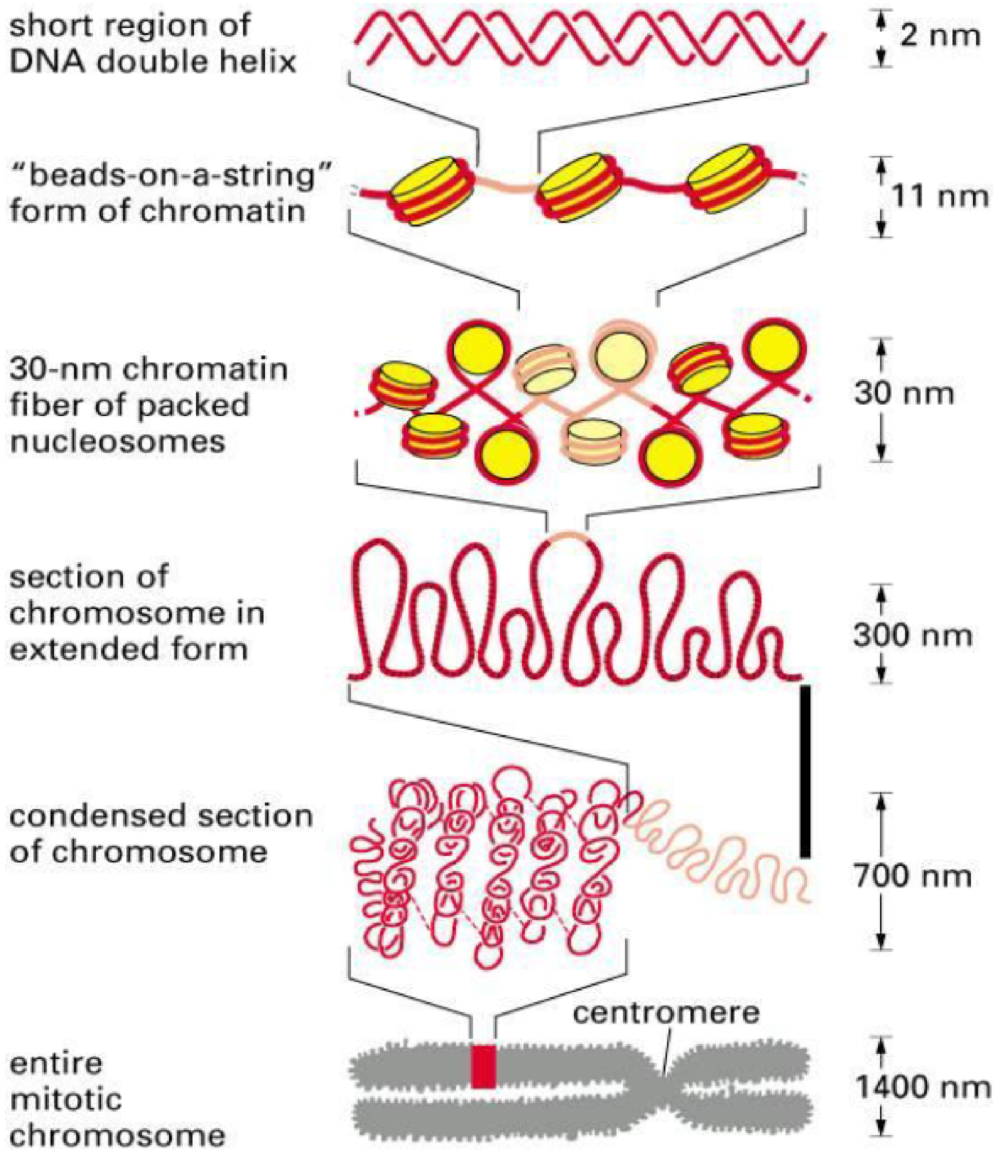
Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet*. 8, 749–761.

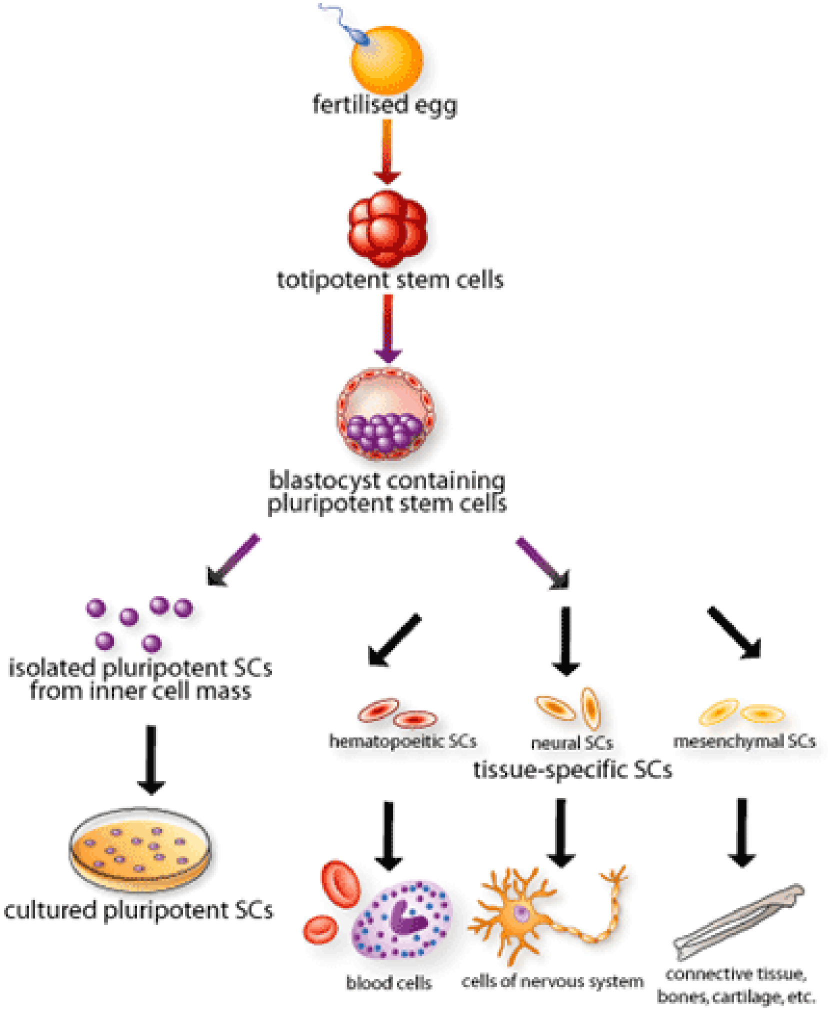
Human genome in action



Civelek (2014) *Nat Rev Genet*

More realistic picture


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm



Molecular Biology of the Cell, 4th ed.

Chaudrey (2004) *Stem Cell Bioeng*

Epigenetics

Epigenetics: heritable phenotype changes that do not involve alterations in the DNA sequence

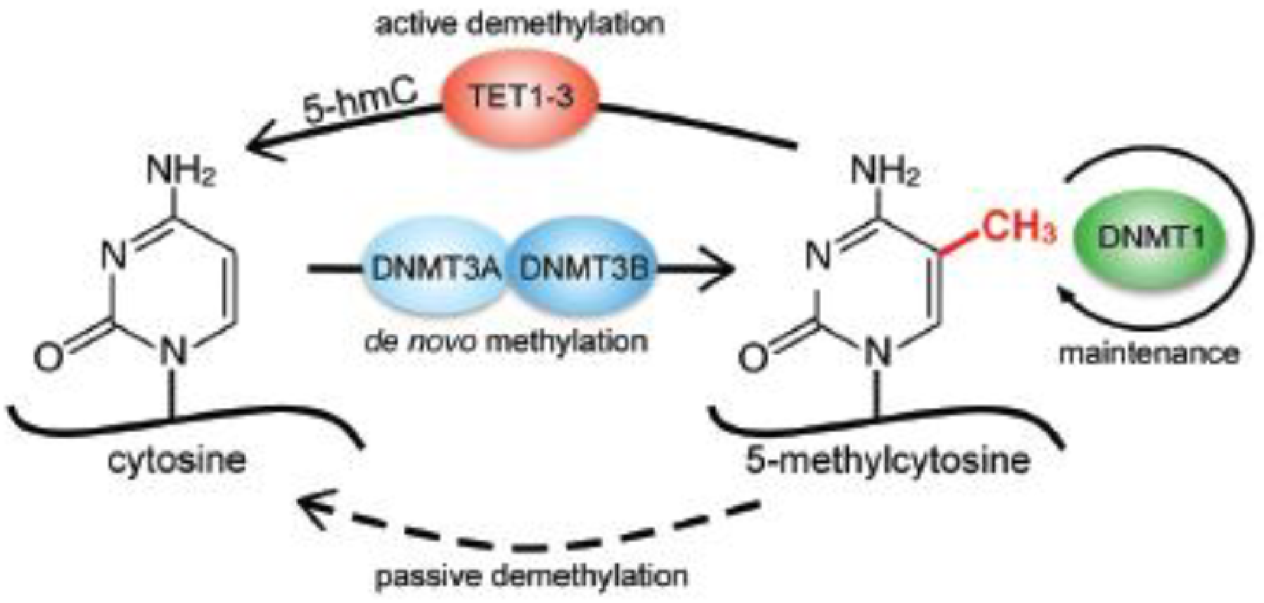
Epigenetic regulation:

1. DNA methylation at CpG dinucleotides
2. Covalent modification of histone proteins
3. Noncoding RNAs

- *Above the genetis*: instructions on using instructions, or gene expression control mechanisms
- Methylation and histone modifications are reversible
- Maintained at cell division and erased during early embriogenesis
- Affected by internal (development, aging) and environmental (chemicals, drugs, diet, lifestyle) factors

DNA methylation

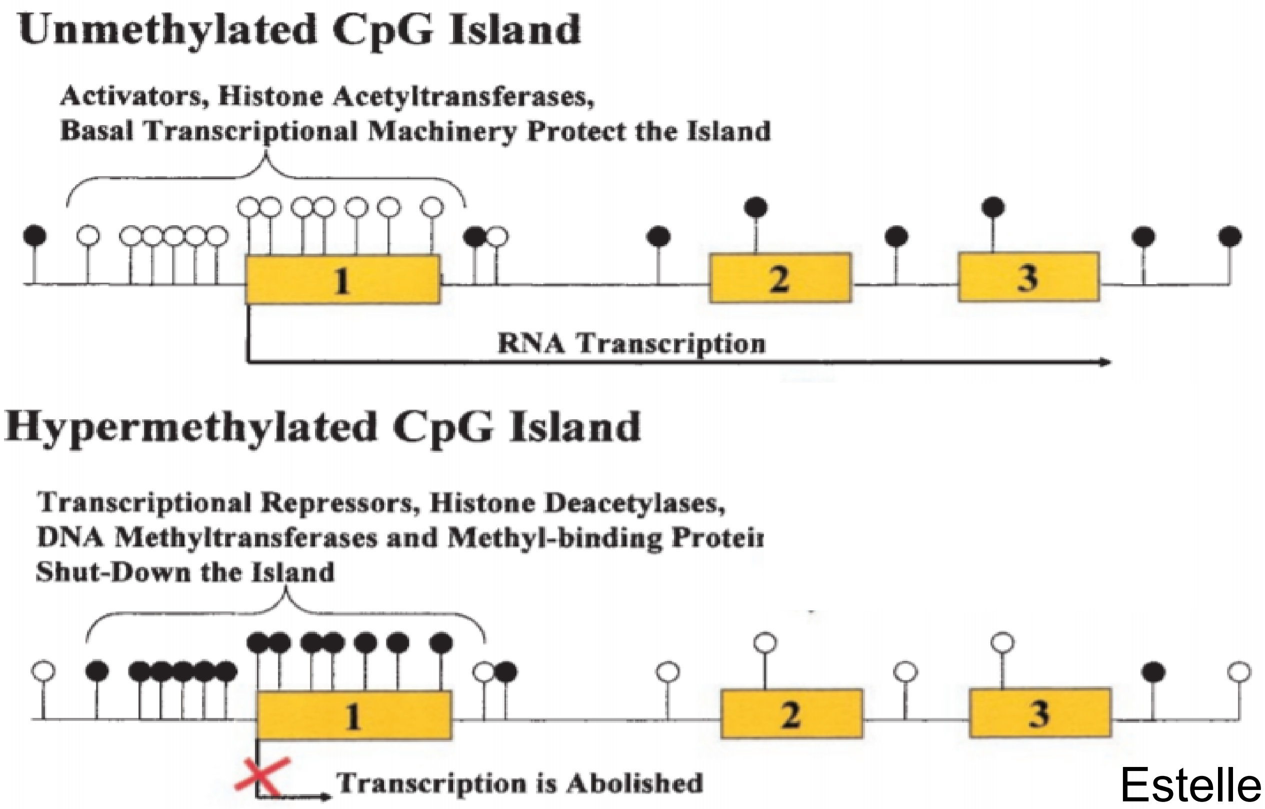
- The only known epigenetic modification of DNA in mammals is methylation of cytosine at position C_5 in CpG dinucleotides
- DNA methyltransferases (DNMTs) establish and maintain DNA methylation patterns
- Methyl-CpG binding proteins (MBDs) read them
- Patterns of CpG methylation may be person-specific, tissue-specific, or locus-specific



Ambrosi (2017) *J Mol Biol*

CpG dinucleotides and islands

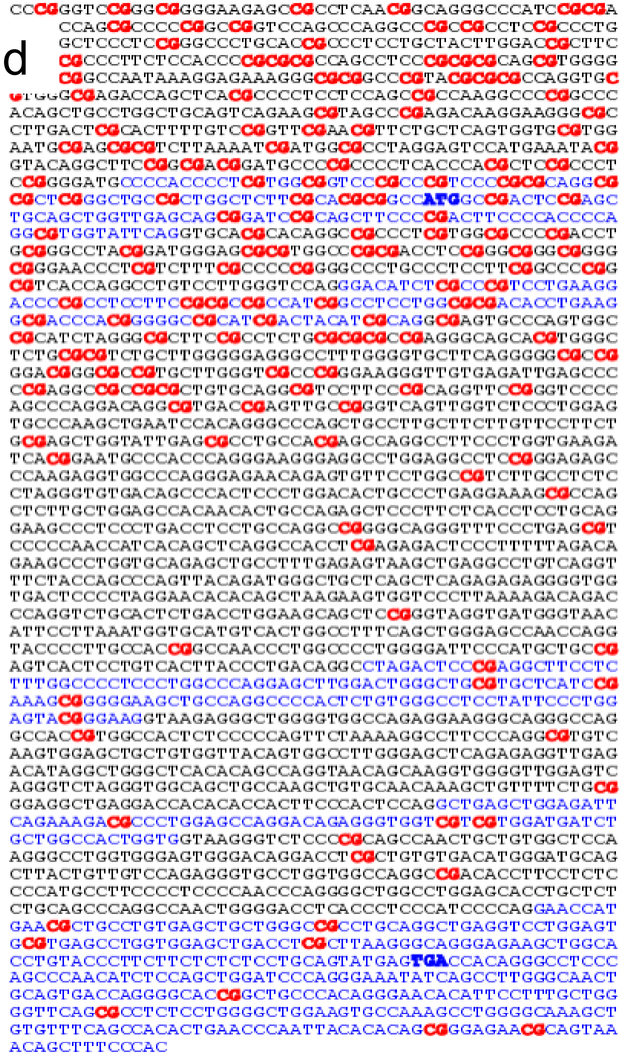
- **CpG island** *ad hoc* definition: length >200 bp, CG >50%, observed-to-expected CpG ratio >60%
- ~30,000 CpG islands in the human genome
- ~70% of human promoters have high CpG content (Saxonov 2006 *PNAS*)
- **Methylation of CpG islands silences gene expression**


Unmethylated CpG Island
Activators, Histone Acetyltransferases,
Basal Transcriptional Machinery Protect the Island
1
2
3
RNA Transcription
Hypermethylated CpG Island
Transcriptional Repressors, Histone Deacetylases,
DNA Methyltransferases and Methyl-binding Proteir
Shut-Down the Island
1
2
3
Transcription is Abolished

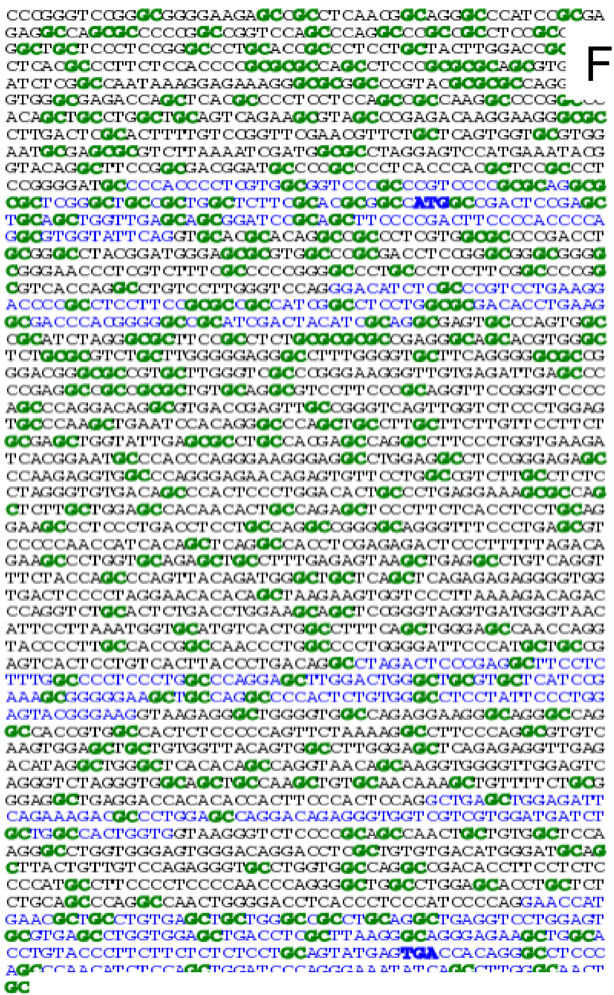
Esteller (2002) *Oncogene*

CpG dinucleotides and islands

Rev strand

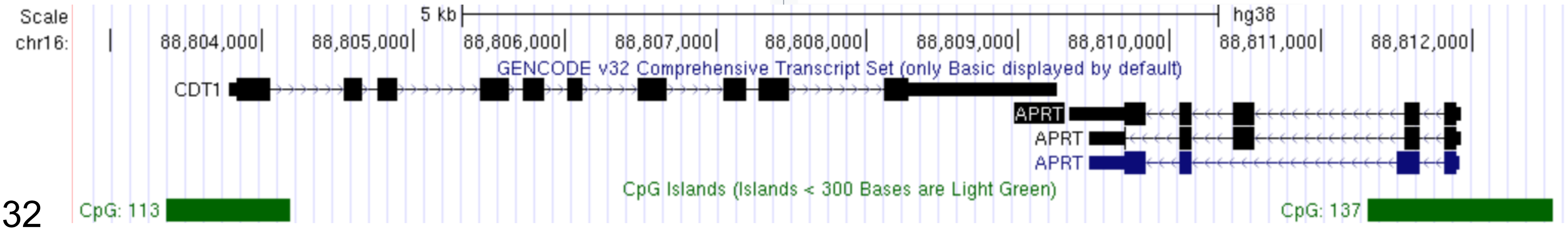


```
CCCGGGTCCGGGCGGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
    CCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
    GCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
    CGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
    CGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
GAAGCCCTCCCTGACCTCCTGCCAGGCCGGGGCAGGGTTTCCCTGAGCGT
CCCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCCTTTGAGAGTAAGCTGAGGCCTGTCAGGT
TTCTACCAGCCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCCTAGGAACACACAGCTAAGAAGTGGTCCCTTAAAAGACAGAC
CCAGGTCTGCACTCTGACCTGGAAGCAGCTCCGGGTAGGTGATGGGTAAC
ATTCCTTAAATGGTGCATGTCACTGGCCTTTCAGCTGGGAGCCAACCAGG
TACCCCTTGCCACCGGCCAACCCTGGCCCCTGGGGATTCCCATGCTGCCG
AGTCACTCCTGTCACTTACCCTGACAGGCCTAGACTCCCGAGGCTTCCTC
TTTGGCCCCTCCCTGGCCCAGGAGCTTGGACTGGGCTGCGTGCTCATCCG
AAAGCGGGGGAAGCTGCCAGGCCCCACTCTGTGGGCCTCCTATTCCCTGG
AGTACGGGAAGGTAAGAGGGCTGGGGTGGCCAGAGGAAGGGCAGGGCCAG
GCCACCGTGGCCACTCTCCCCCAGTTCTAAAAGGCCTTCCCAGGCGTGTC
AAGTGGAGCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTTCTGCG
GGAGGCTGAGGACCACACACCACTTCCCACTCCAGGCTGAGCTGGAGATT
CAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTCGTCGTGGATGATCT
GCTGGCCACTGGTGGTAAGGGTCTCCCCGCAGCCAACTGCTGTGGCTCCA
AGGGCCTGGTGGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATGCAG
CTTACTGTTGTCCAGAGGGTGCCTGGTGGCCAGGCCGACACCTTCCTCTC
CCCATGCCTTCCCCTCCCCAACCCAGGGGCTGGCCTGGAGCACCTGCTCT
CTGCAGCCCAGGCCAACTGGGGACCTCACCCTCCCATCCCCAGGAACCAT
GAACGCTGCCTGTGAGCTGCTGGGCCGCCTGCAGGCTGAGGTCCTGGAGT
GCGTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCA
CCTGTACCCTTCTTCTCTCTCCTGCAGTATGAGTGACCACAGGGCCTCCC
AGCCCAACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGGCAACT
GCAGTGACCAGGGGCACCGGCTGCCCACAGGGAACACATTCCTTTGCTGG
GGTTCAGCGCCTCTCCTGGGGCTGGAAGTGCCAAAGCCTGGGGCAAAGCT
GTGTTTCAGCCACACTGAACCCAATTACACACAGCGGGAGAACGCAGTAA
ACAGCTTTCCCAC
```

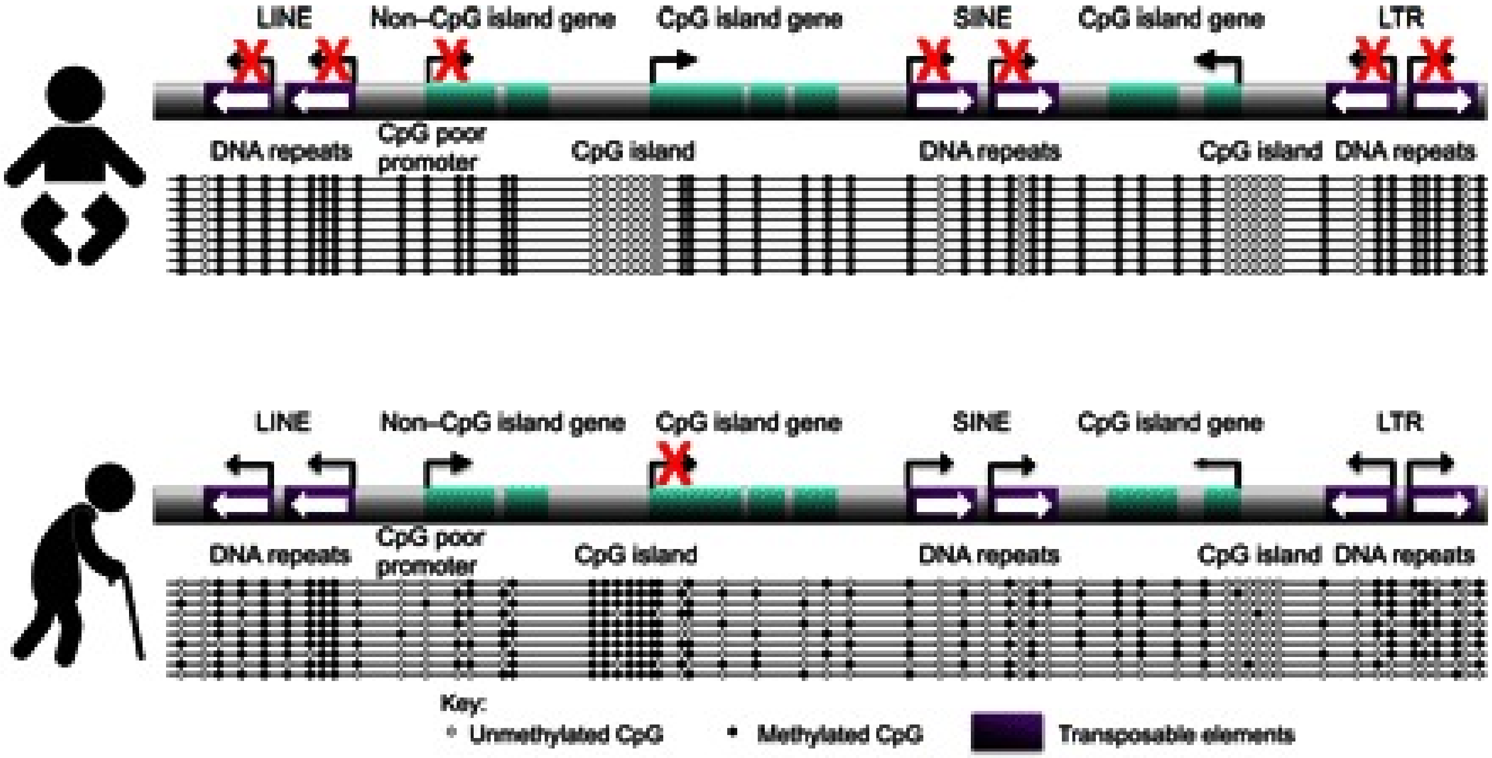


Fwd strand

APRT: Adenine Phosphoribosyltransferase



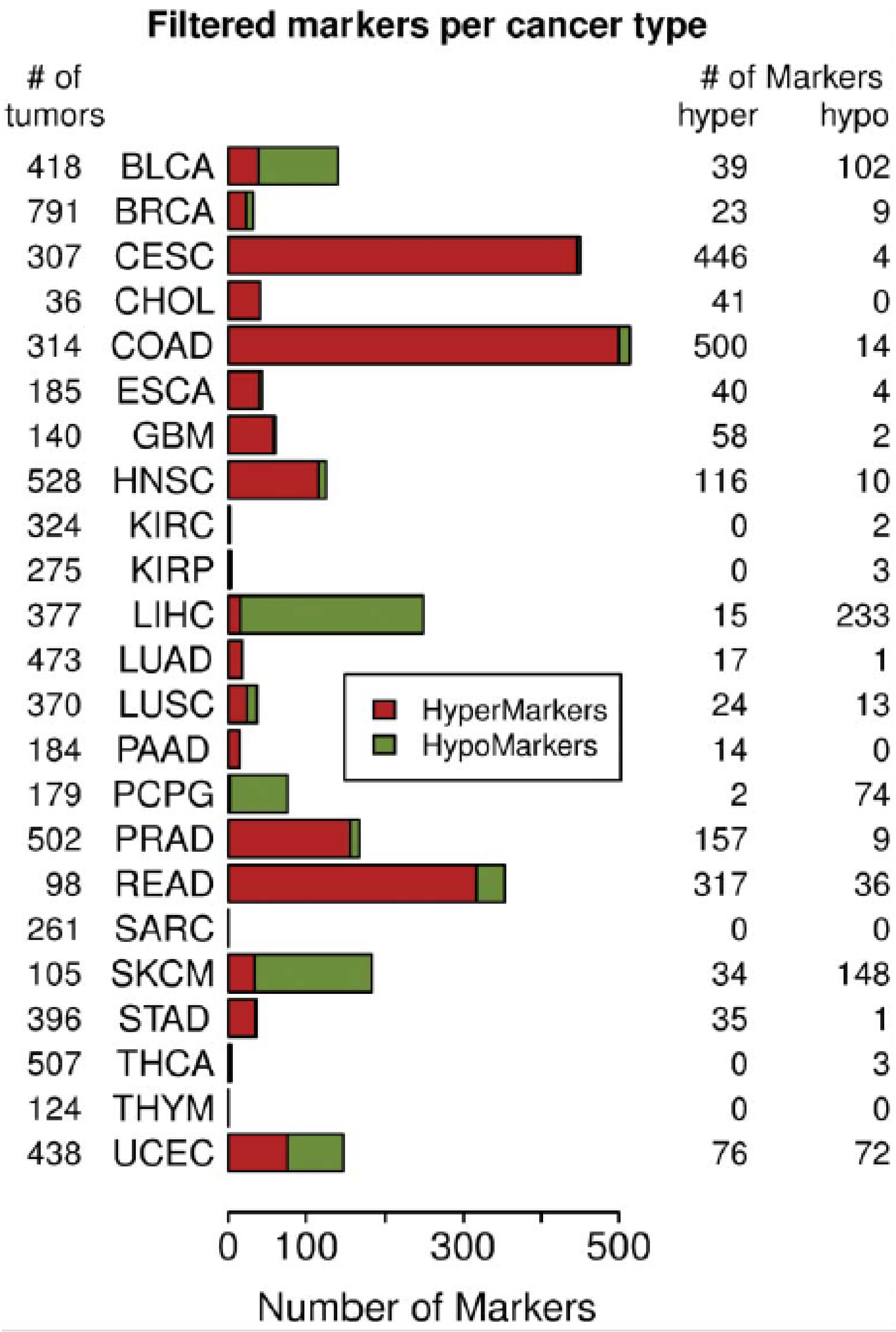
DNA methylation and aging



Young mammalian cells are characterized by DNA hypermethylation over the genome, with the exception of CpG islands within the promoters of expressed genes. In particular, DNA repeats, such as LINE, SINE, and long terminal repeat (LTR) transposable elements, are heavily DNA-methylated, helping to maintain them in a constitutive heterochromatin state. **During aging, there is general DNA hypomethylation over the genome, which mostly occurs in a stochastic manner within the cell population.** Loss of DNA methylation leads to activation of normally silenced DNA sequences like the transposable elements. However, DNA methylation also increases in a nonstochastic manner over the CpG islands of certain genes, correlating with their heterochromatinization and silencing.

Pal & Tyler (2016) *Sci Adv*

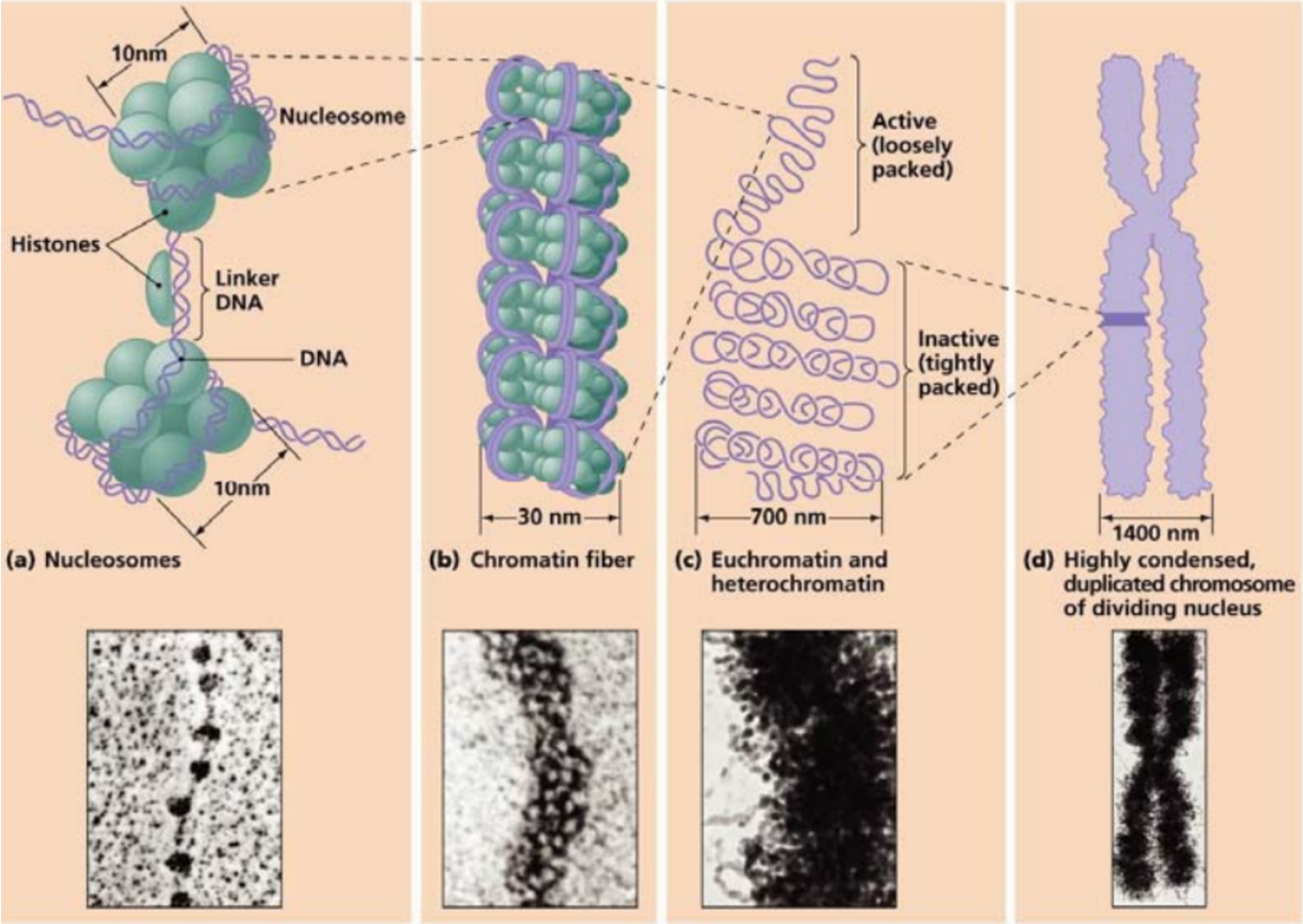
DNA methylation and cancer



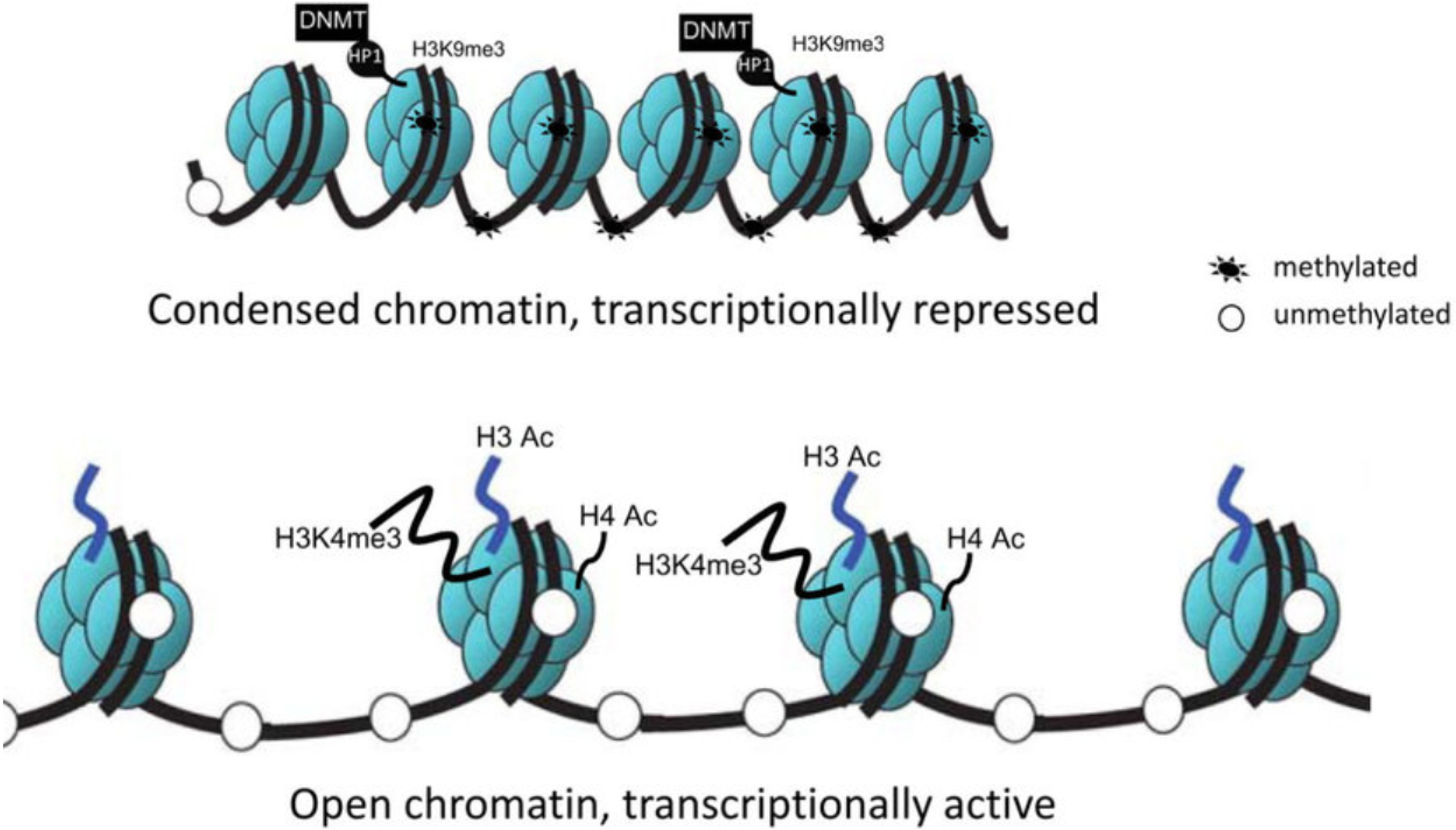
We identified **differentially methylated regions for individual cancer types** and those were further filtered against data from normal tissues to obtain marker regions with cancer-specific methylation, resulting in a total of 1,250 hypermethylated and 584 hypomethylated marker CpGs. From hypermethylated markers, optimal sets of six markers for each TCGA cancer type were chosen that could identify most tumors with high specificity and sensitivity [area under the curve (AUC): 0.969-1.000] and a universal 12 marker set that can detect tumors of all 33 TCGA cancer types (AUC >0.84).

Vrba & Futscher (2018) *Epigenetics*

Histone modifications, histone code


10nm
Nucleosome
Histones
Linker DNA
DNA
10nm
(a) Nucleosomes
30 nm
(b) Chromatin fiber
Active (loosely packed)
Inactive (tightly packed)
700 nm
(c) Euchromatin and heterochromatin
1400 nm
(d) Highly condensed, duplicated chromosome of dividing nucleus

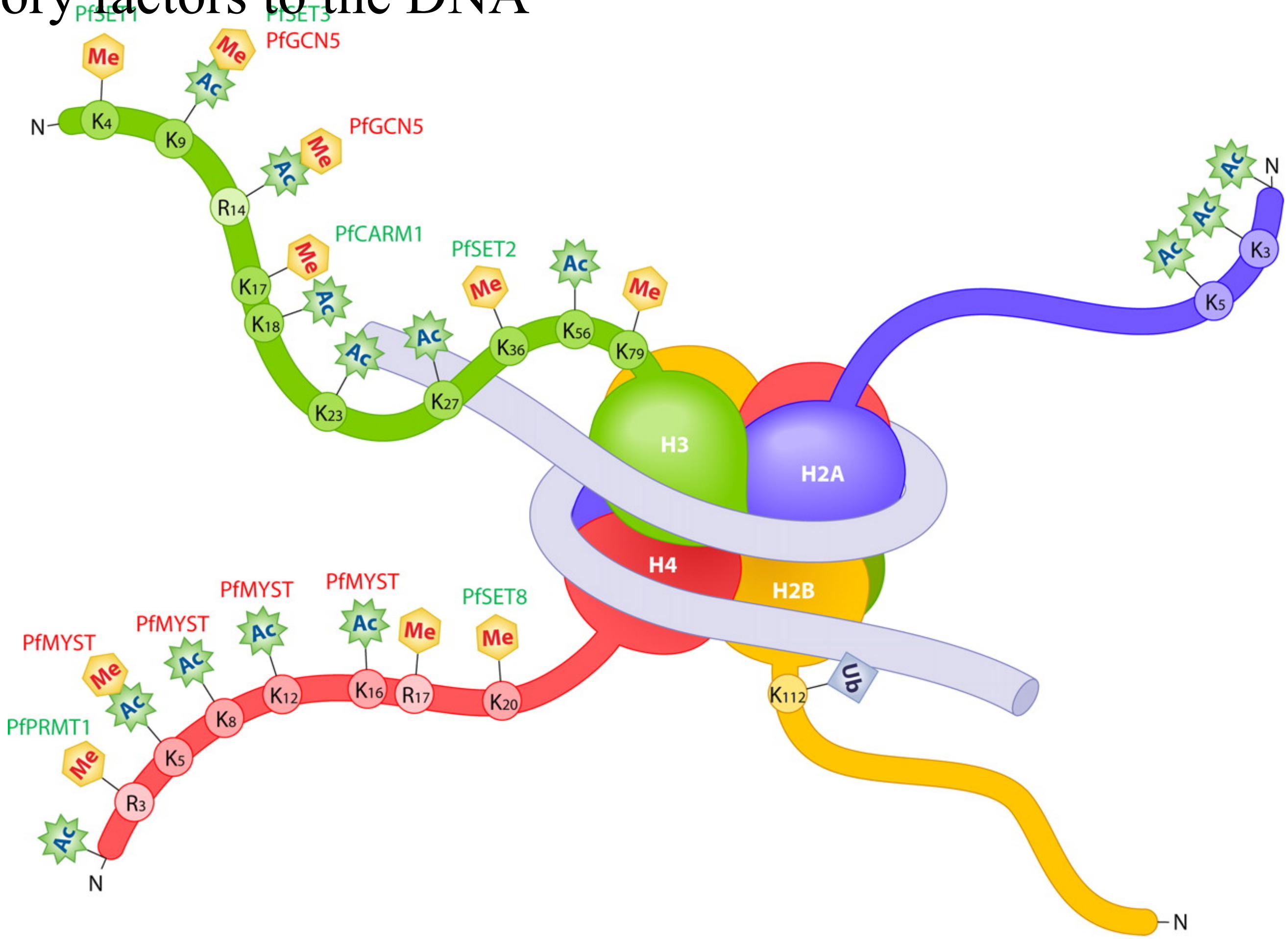
Histone modifications, histone code


DNMT
HP1
H3K9me3
DNMT
HP1
H3K9me3
Condensed chromatin, transcriptionally repressed
✹ methylated
○ unmethylated
H3 Ac
H3K4me3
H4 Ac
H3 Ac
H3K4me3
H4 Ac
Open chromatin, transcriptionally active

Bansal (2017) *Pediatric Diabetes*

Histone modifications, histone code

- **Histone code**: post-translational modifications of histone N-ends (Lys, Arg, Cys) by phosphorylation, acetylation, methylation and ubiquitylation.
- These changes regulate gene expression by modulating the access of regulatory factors to the DNA



Histone modifications, histone code

The eukaryotic genome is organized in what is known as a **nucleosome**, the first level of condensation. The nucleosome is composed of 147 base pairs of negatively-charged DNA wrapped twice around an octamer of positively-charged proteins called **histones**. It consists of two H2A and H2B dimers, and a H3 and H4 tetramer. The nucleosomes are separated by 1,016 base pairs (bp) of DNA called "linker DNA", which constitutes an arrangement referred to as "beads on a string", that is around 10nm in diameter. DNA can be further condensed at different points during the cell cycle, forming a 30nm chromatin fiber composed of packed nucleosomes using the histone H1, which binds to the linker DNA. These 30nm fibers can form scaffolds and further condense until chromosomes are formed, which are the highest form of DNA organization within a cell.

Histones have very dynamic N-terminal "tails" extending from the surface of the nucleosome that are rich in basic amino acids. These tails can be modified by post-translational modifications (PTM's) catalyzed by a variety of enzymes, by adding either methyl, acetyl or phosphoryl groups. Aditionally, lysines can be mono, di or tri-methylated, while arginine can accept up to two methyl groups which adds to the complexity. Methylation of DNA at cytosine residues, as well as PTMs of histones, including phosphorylation, acetylation, methylation and ubiquitylation, contributes to the epigenetic information carried by chromatin. These changes play an important role in the regulation of gene expression by modulating the access of regulatory factors to the DNA. Many modification sites are close enough to each other and it seems that modification of histone tails by one enzyme might influence the rate and efficiency at which other enzymes use the newly modified tails as a substrate.

http://themysteryofepigenetics.blogspot.com/2016/10/the-histone-

Histone modifications, histone code

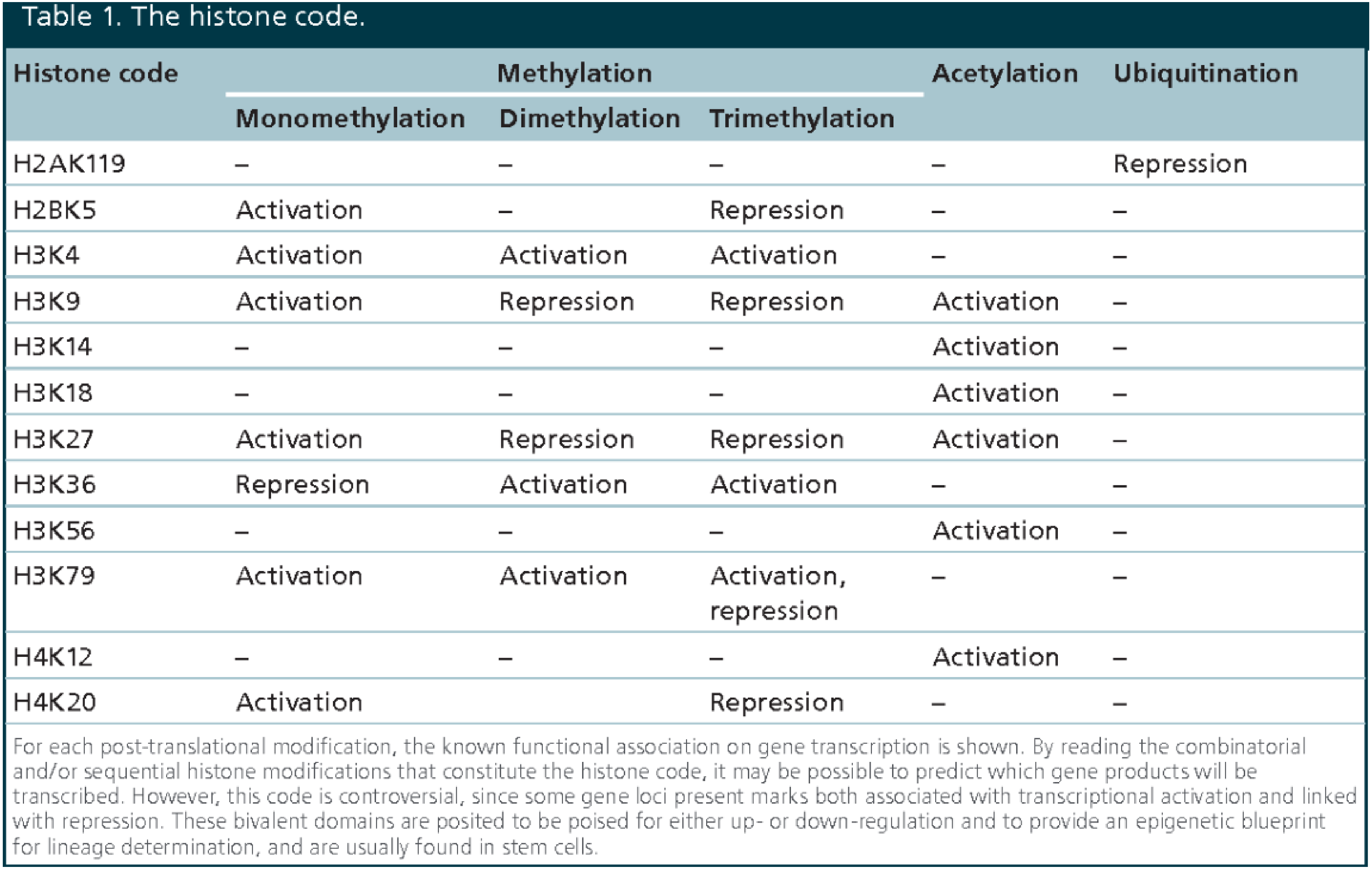


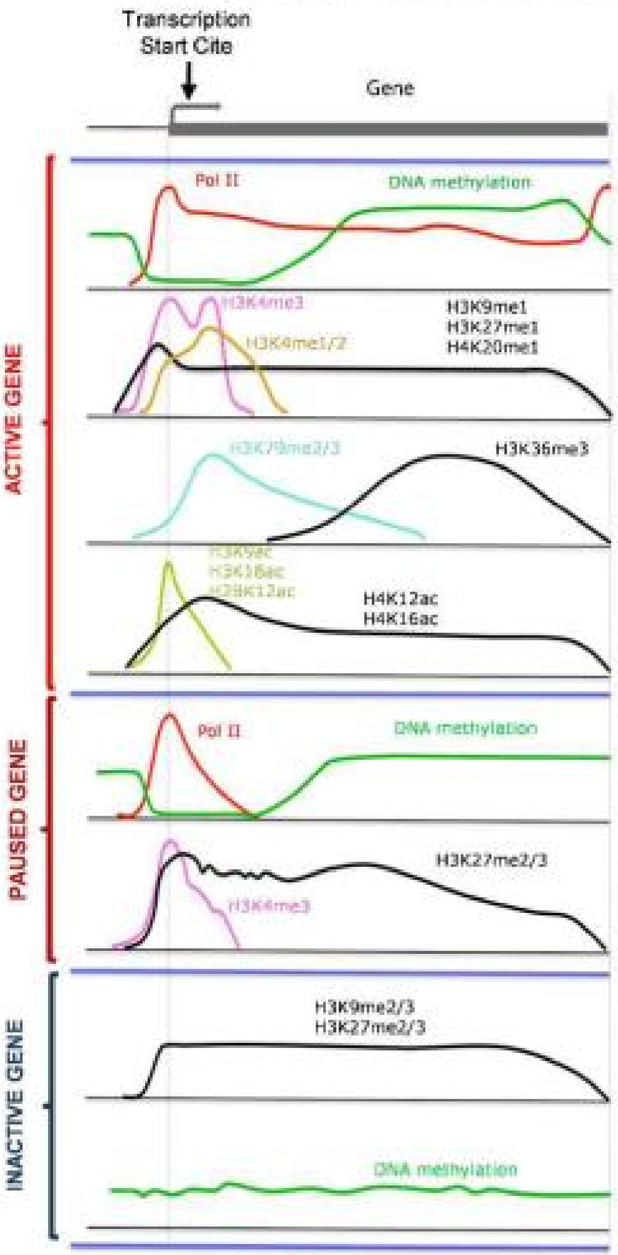
Table 1. The histone code.

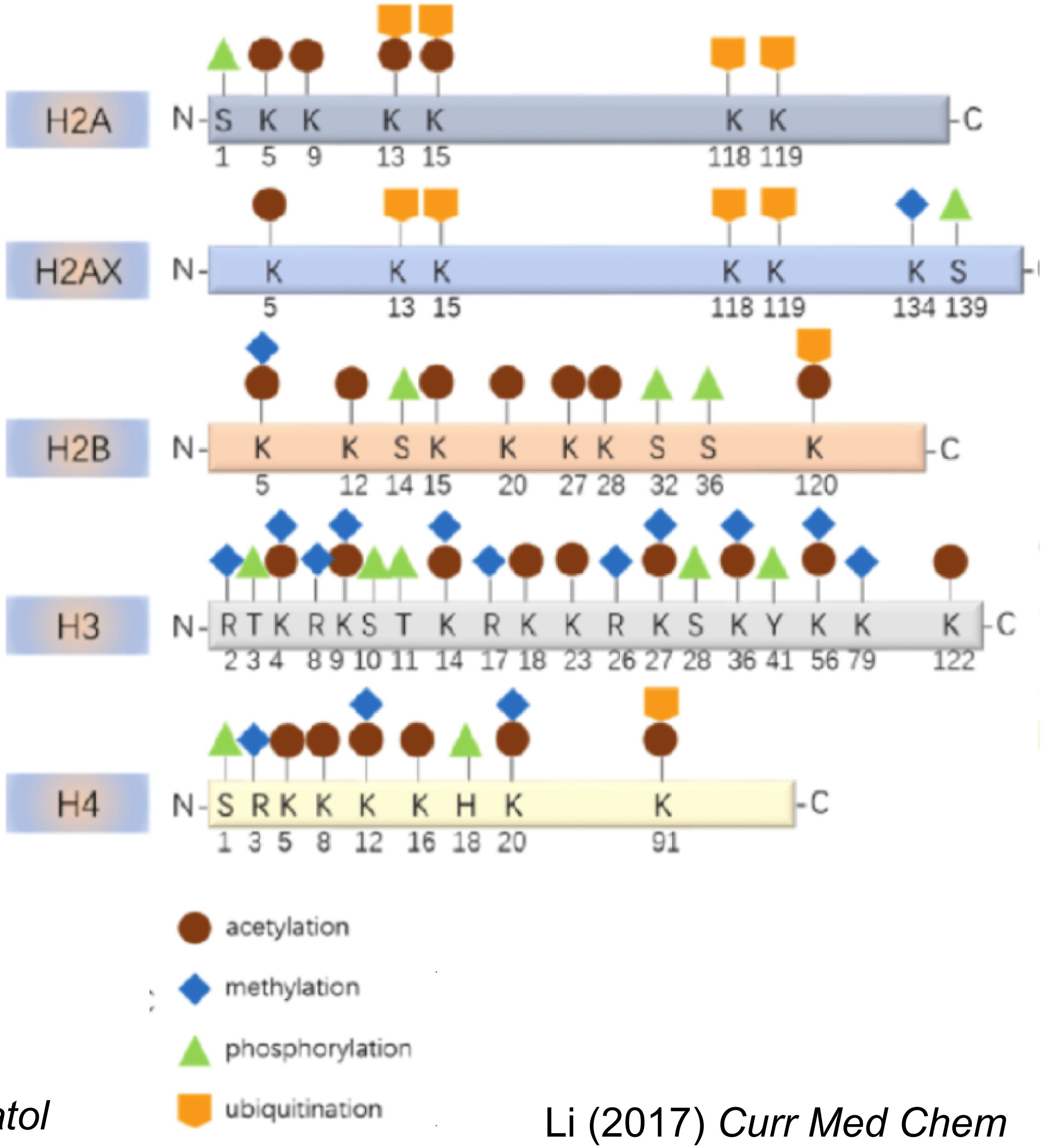
Histone code	Methylation			Acetylation	Ubiquitination
	Monomethylation	Dimethylation	Trimethylation		
H2AK119	–	–	–	–	Repression
H2BK5	Activation	–	Repression	–	–
H3K4	Activation	Activation	Activation	–	–
H3K9	Activation	Repression	Repression	Activation	–
H3K14	–	–	–	Activation	–
H3K18	–	–	–	Activation	–
H3K27	Activation	Repression	Repression	Activation	–
H3K36	Repression	Activation	Activation	–	–
H3K56	–	–	–	Activation	–
H3K79	Activation	Activation	Activation, repression	–	–
H4K12	–	–	–	Activation	–
H4K20	Activation		Repression	–	–

For each post-translational modification, the known functional association on gene transcription is shown. By reading the combinatorial and/or sequential histone modifications that constitute the histone code, it may be possible to predict which gene products will be transcribed. However, this code is controversial, since some gene loci present marks both associated with transcriptional activation and linked with repression. These bivalent domains are posited to be poised for either up- or down-regulation and to provide an epigenetic blueprint for lineage determination, and are usually found in stem cells.

Bauge (2014) *Future Med Chem*

Histone modifications, histone code



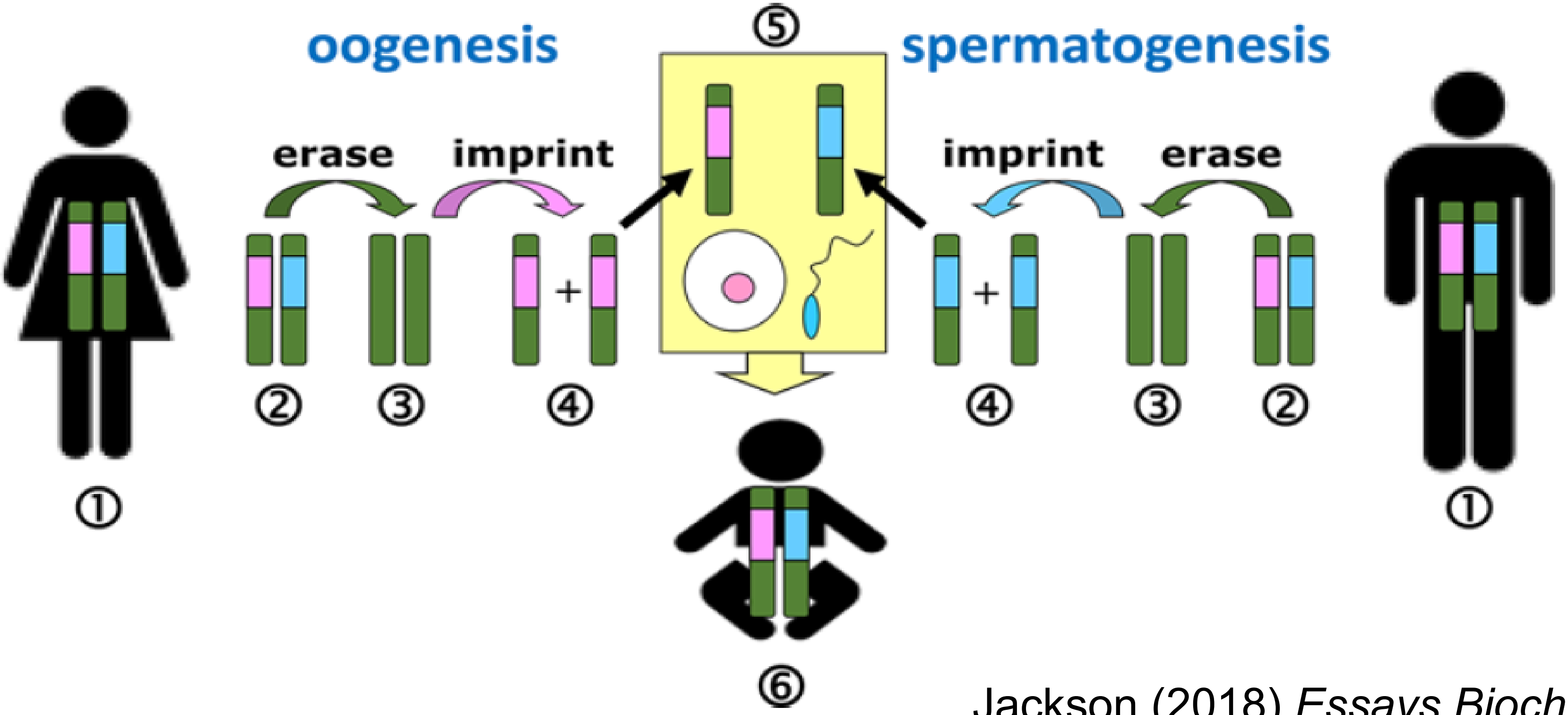


Botchkarev (2012) *J Invest Dermatol*

Li (2017) *Curr Med Chem*

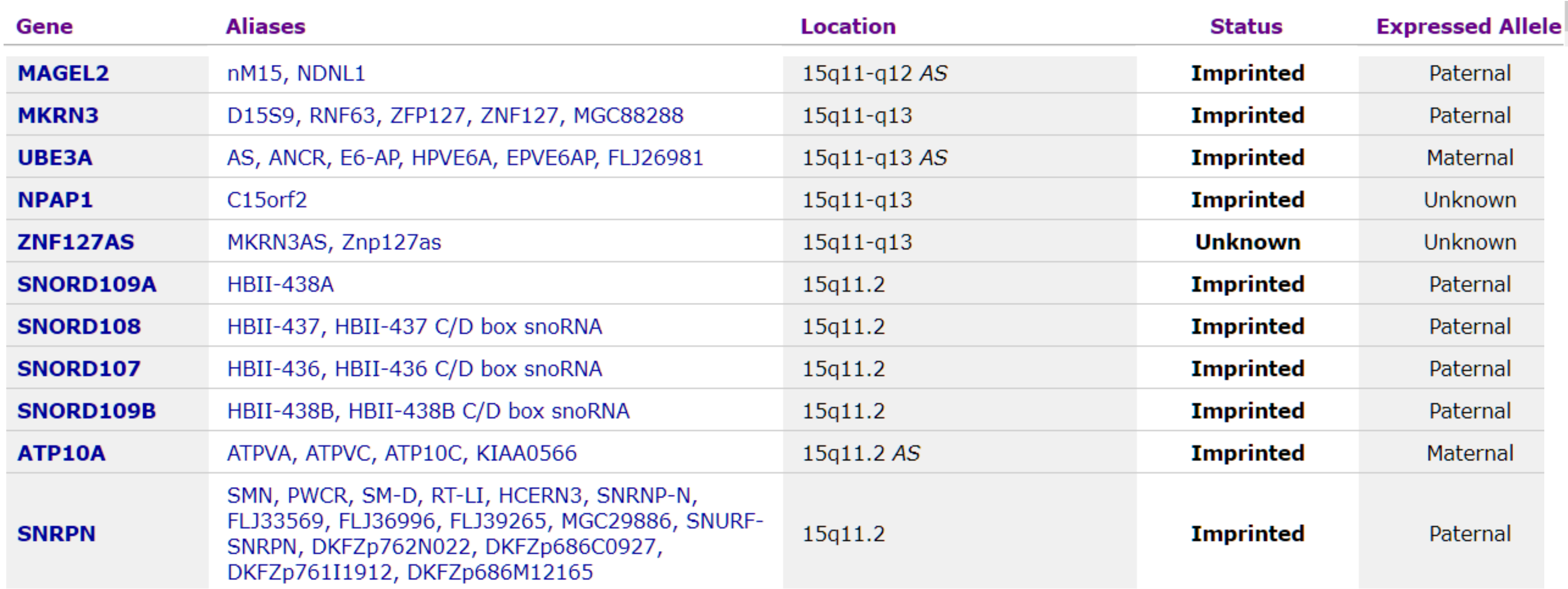
Chromosomal imprinting

- **Chromosomal imprinting, or imprints**: ~100 genes on various chromosomes, one copy is inactive by epigenetic mechanisms depending upon parent of origin
- For some genes (~70) only the paternal allele is active, while the maternal copy is epigenetically silenced throughout the life of the individual, and vice versa (~30 genes)
- Mutations in an active copy of a gene result in **imprinting disorders**



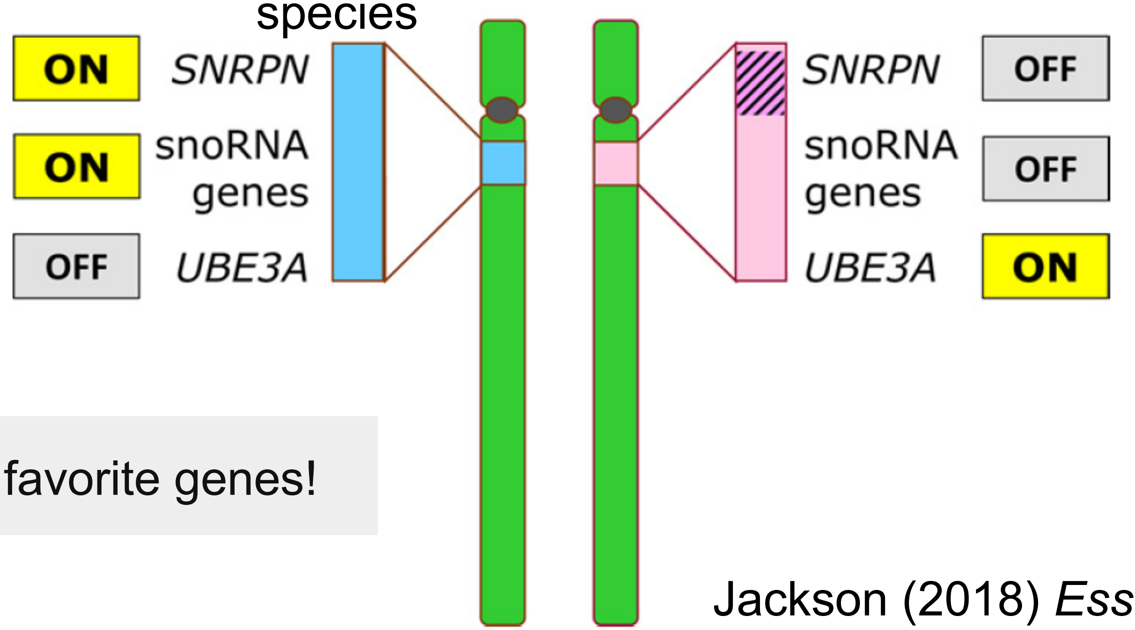
Jackson (2018) *Essays Biochem*

Chromosomal imprinting



Gene	Aliases	Location	Status	Expressed Allele
MAGEL2	nM15, NDNL1	15q11-q12 *AS*	**Imprinted**	Paternal
MKRN3	D15S9, RNF63, ZFP127, ZNF127, MGC88288	15q11-q13	**Imprinted**	Paternal
UBE3A	AS, ANCR, E6-AP, HPVE6A, EPVE6AP, FLJ26981	15q11-q13 *AS*	**Imprinted**	Maternal
NPAP1	C15orf2	15q11-q13	**Imprinted**	Unknown
ZNF127AS	MKRN3AS, Znp127as	15q11-q13	**Unknown**	Unknown
SNORD109A	HBII-438A	15q11.2	**Imprinted**	Paternal
SNORD108	HBII-437, HBII-437 C/D box snoRNA	15q11.2	**Imprinted**	Paternal
SNORD107	HBII-436, HBII-436 C/D box snoRNA	15q11.2	**Imprinted**	Paternal
SNORD109B	HBII-438B, HBII-438B C/D box snoRNA	15q11.2	**Imprinted**	Paternal
ATP10A	ATPVA, ATPVC, ATP10C, KIAA0566	15q11.2 *AS*	**Imprinted**	Maternal
SNRPN	SMN, PWCR, SM-D, RT-LI, HCERN3, SNRNP-N, FLJ33569, FLJ36996, FLJ39265, MGC29886, SNURF-SNRPN, DKFZp762N022, DKFZp686C0927, DKFZp761I1912, DKFZp686M12165	15q11.2	**Imprinted**	Paternal

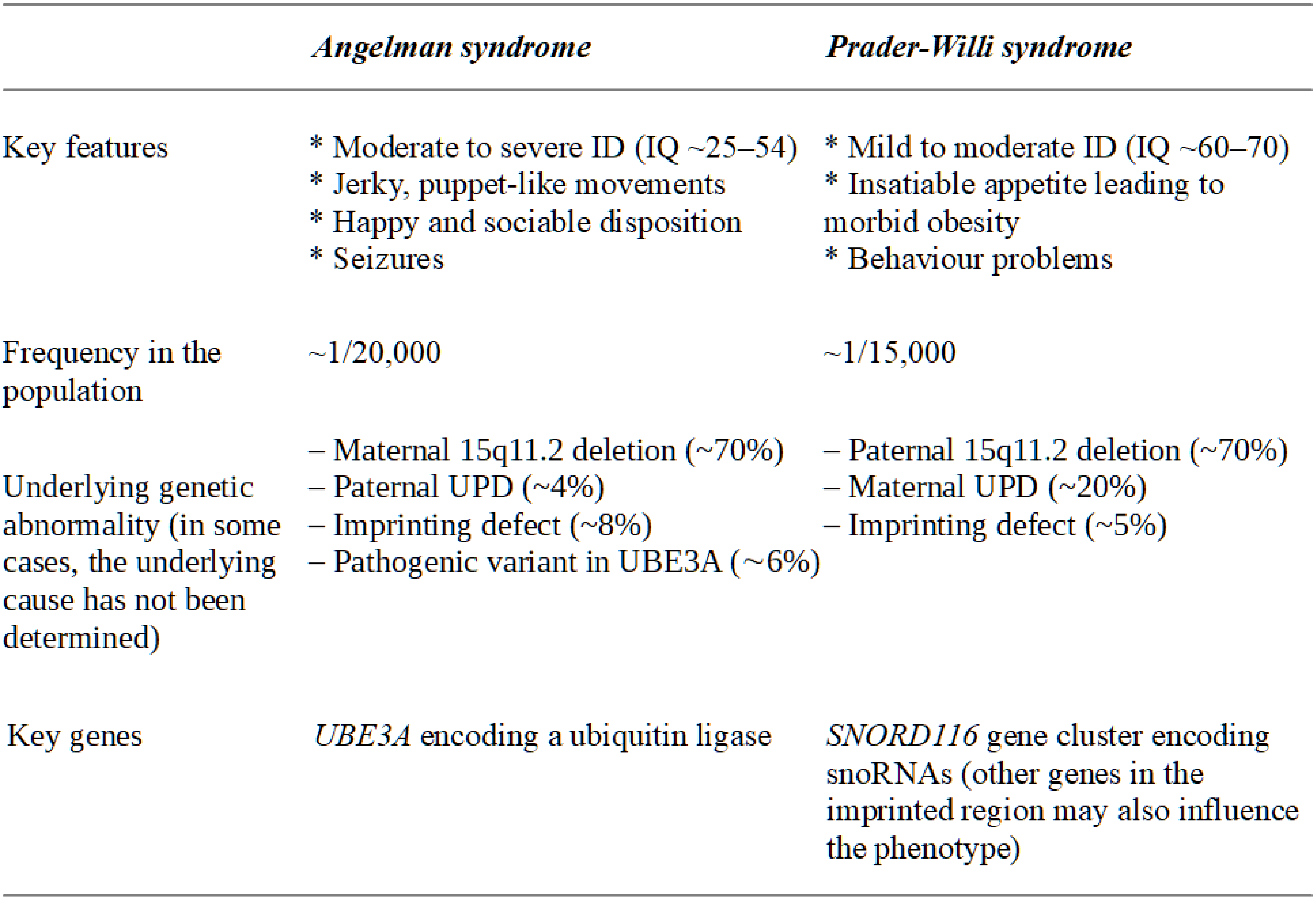
http://www.geneimprint.com/site/genes-by-species



Exercise: check your favorite genes!

Jackson (2018) *Essays Biochem*

Imprinting disorders



	Angelman syndrome	*Prader-Willi syndrome*
Key features	* Moderate to severe ID (IQ ~25–54) * Jerky, puppet-like movements * Happy and sociable disposition * Seizures	* Mild to moderate ID (IQ ~60–70) * Insatiable appetite leading to morbid obesity * Behaviour problems
Frequency in the population	~1/20,000	~1/15,000
Underlying genetic abnormality (in some cases, the underlying cause has not been determined)	– Maternal 15q11.2 deletion (~70%) – Paternal UPD (~4%) – Imprinting defect (~8%) – Pathogenic variant in UBE3A (~6%)	– Paternal 15q11.2 deletion (~70%) – Maternal UPD (~20%) – Imprinting defect (~5%)
Key genes	*UBE3A* encoding a ubiquitin ligase	*SNORD116* gene cluster encoding snoRNAs (other genes in the imprinted region may also influence the phenotype)

Jackson (2018) *Essays Biochem*

Imprinting disorders

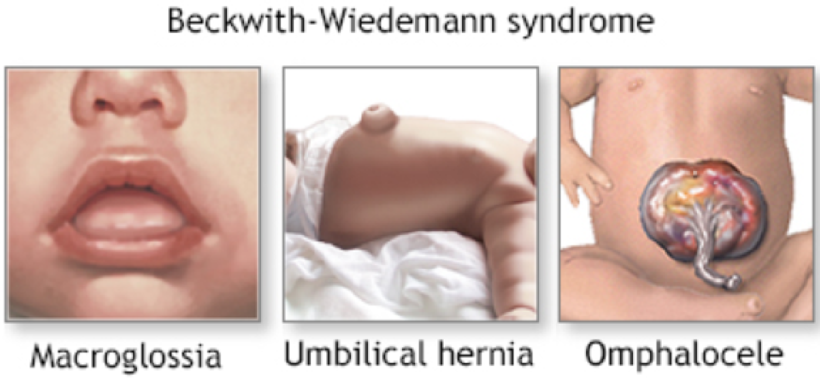
- IGF2 is a hormone that stimulates growth during embryonic and fetal development // not the IGF2 receptor gene!
- Normally maternally silenced in humans
- **Epimutation** (missing methyl tags) can result in two active copies

Activation of the maternal *IGF2* gene during egg formation or very early in development causes **Beckwith-Wiedemann Syndrome (BWS):**

– overgrowth

– an increased risk of cancer, especially during childhood

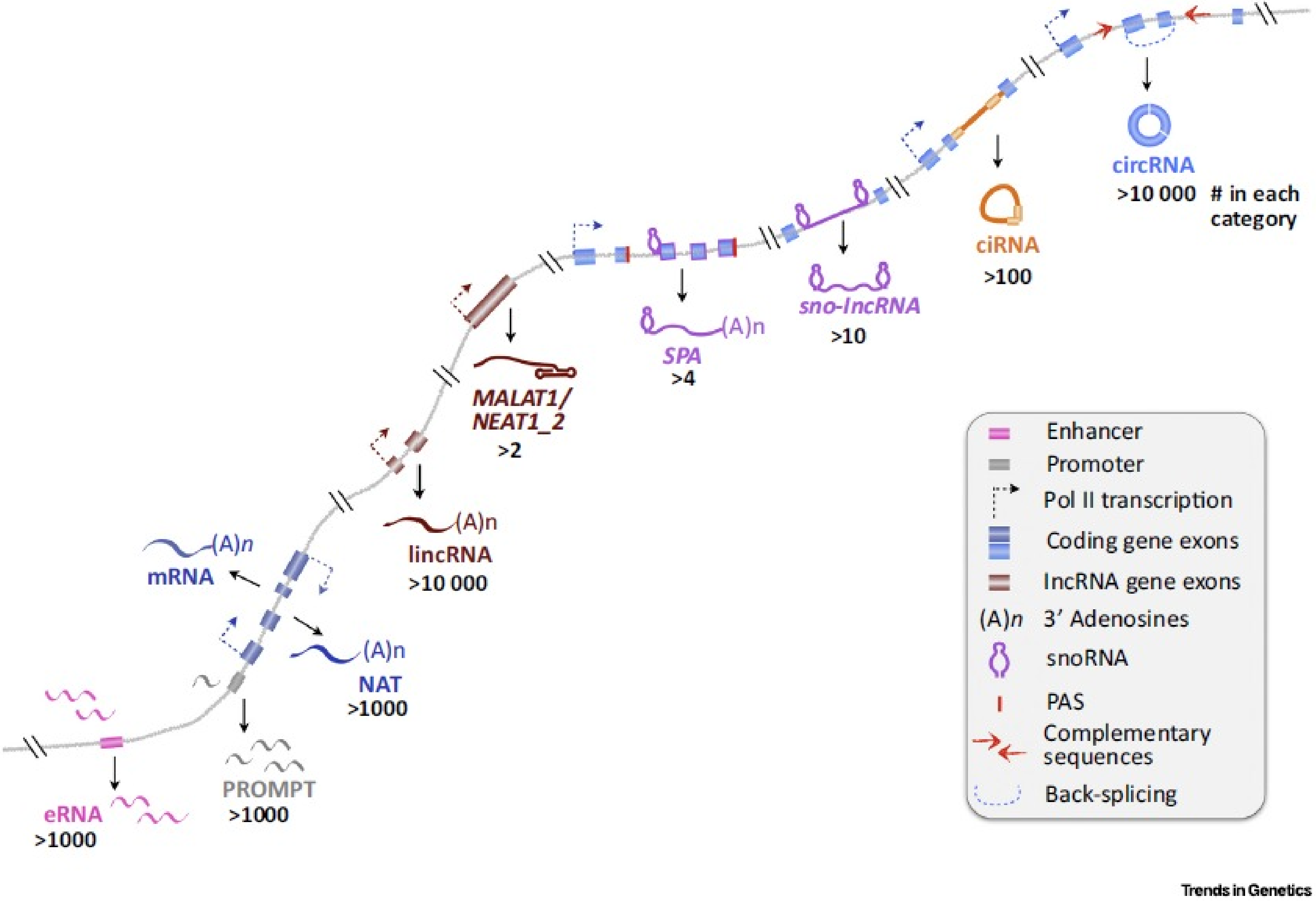
– variety of other symptoms



Frequency: ~15,000 births. However, in babies that were conceived in the laboratory with the help of artificial reproductive technology, the rate of BWS may be as high as 1/4,000.

https://learn.genetics.utah.edu/content/epigenetics/imprinting

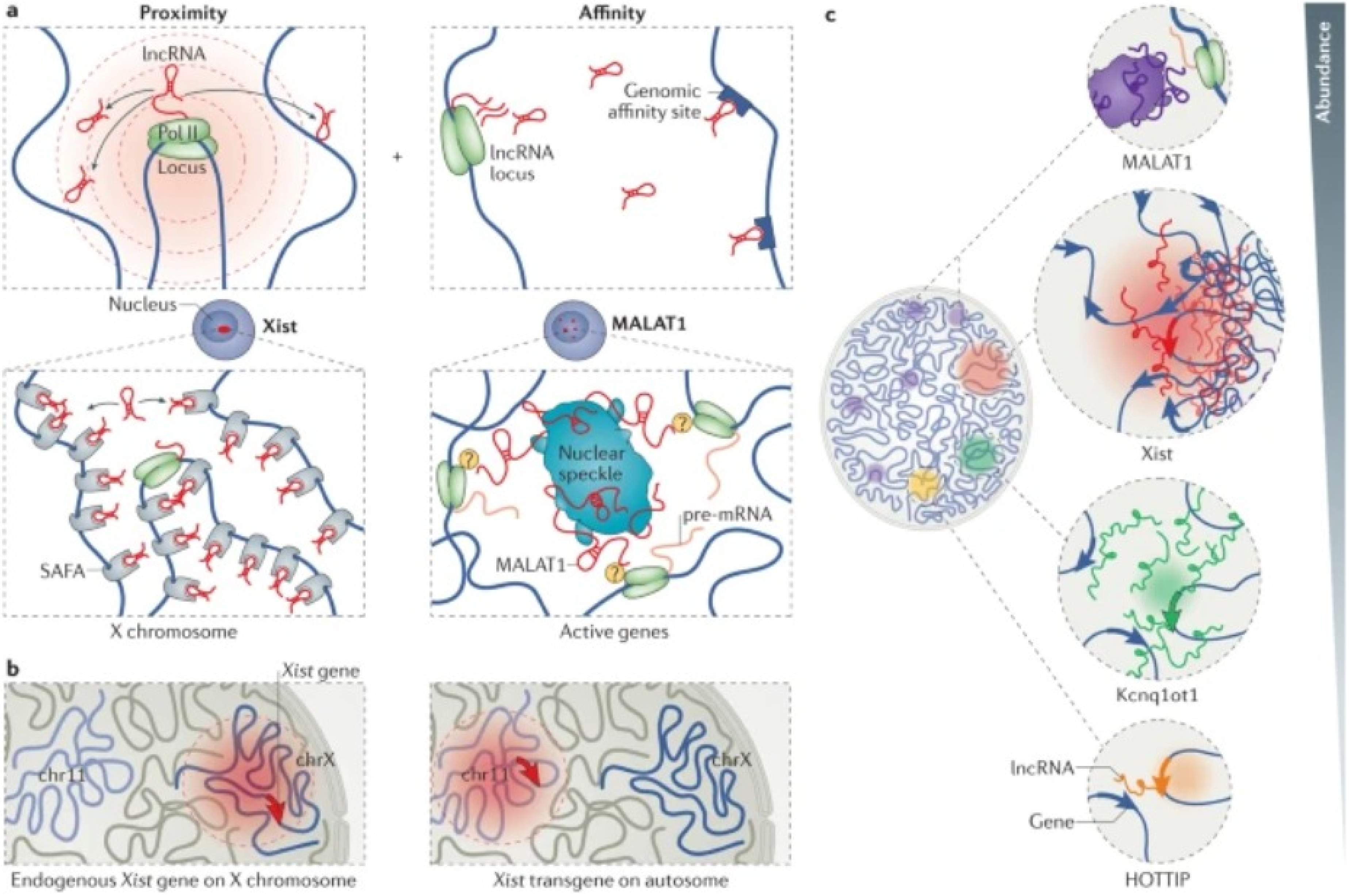
Non-coding RNAs in the genome



Huang Wu (2017) *Trends Genet*

Non-coding RNAs in the genome

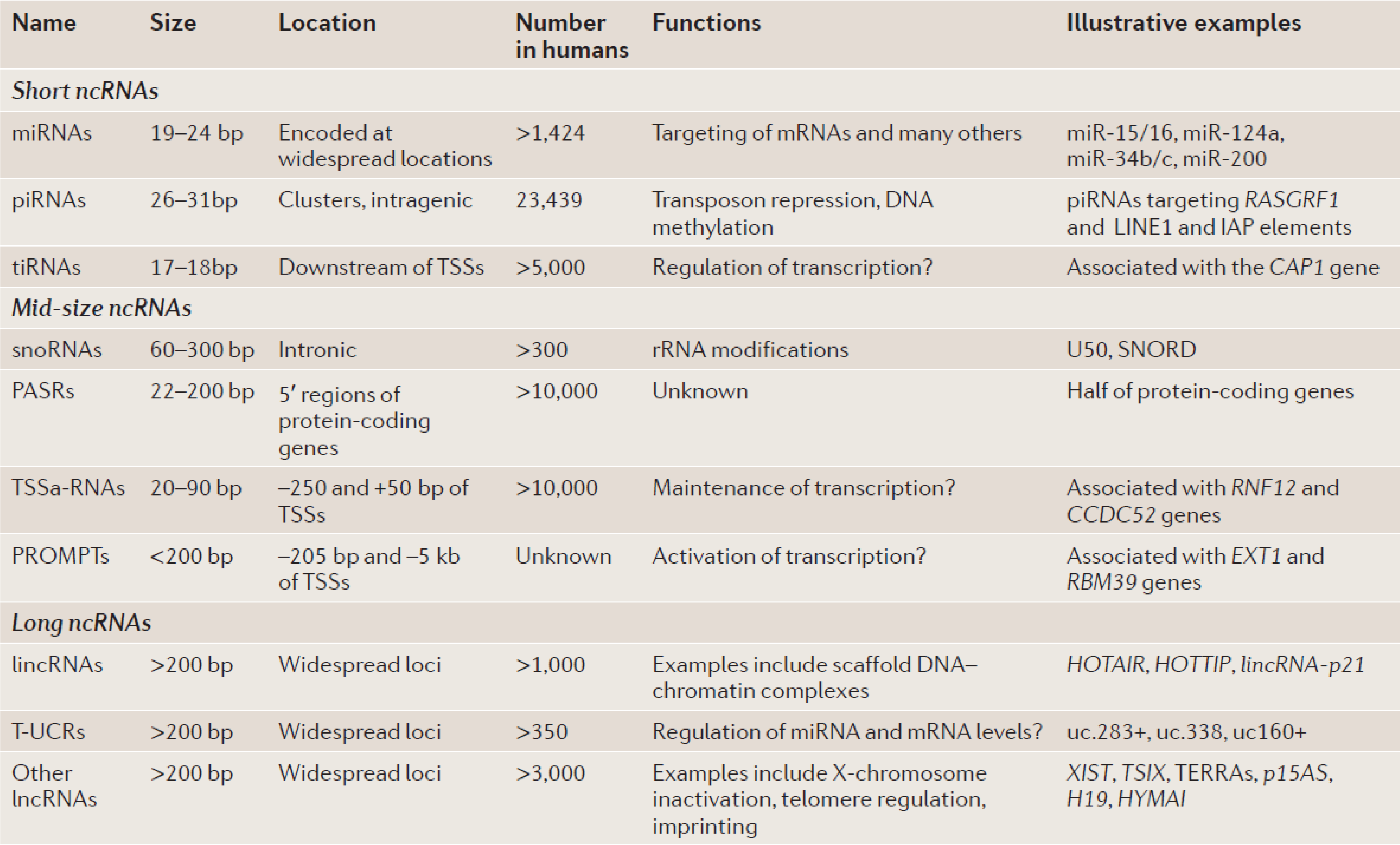
Mechanisms of long non-coding RNA localization to chromatin



Nature Reviews | Molecular Cell Biology

Engreitz (2016) *Nat Rev Mol Cell Biol*

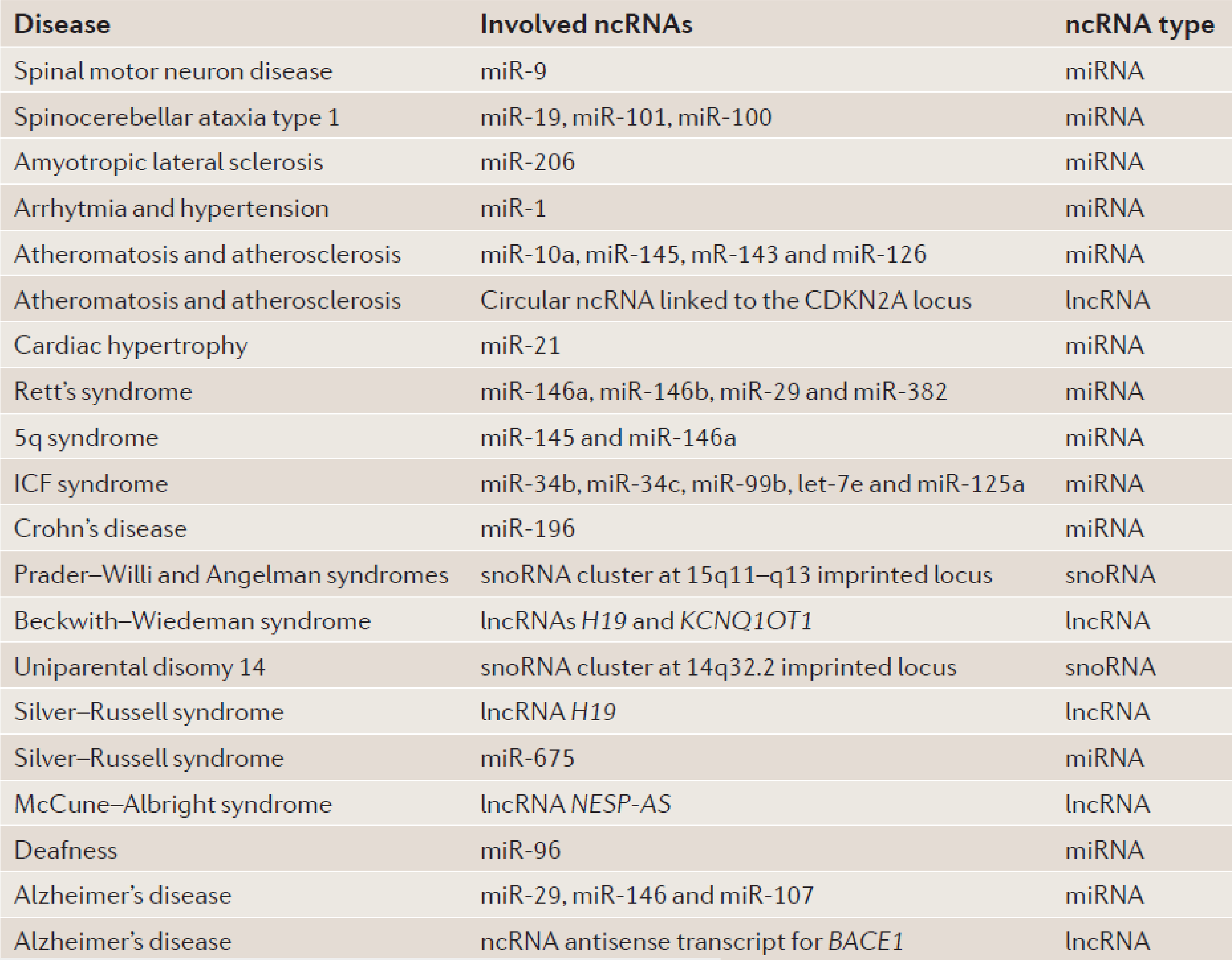
Non-coding RNAs in the genome



Name	Size	Location	Number in humans	Functions	Illustrative examples
Short ncRNAs					
miRNAs	19–24 bp	Encoded at widespread locations	>1,424	Targeting of mRNAs and many others	miR-15/16, miR-124a, miR-34b/c, miR-200
piRNAs	26–31bp	Clusters, intragenic	23,439	Transposon repression, DNA methylation	piRNAs targeting *RASGRF1* and LINE1 and IAP elements
tiRNAs	17–18bp	Downstream of TSSs	>5,000	Regulation of transcription?	Associated with the *CAP1* gene
Mid-size ncRNAs					
snoRNAs	60–300 bp	Intronic	>300	rRNA modifications	U50, SNORD
PASRs	22–200 bp	5′ regions of protein-coding genes	>10,000	Unknown	Half of protein-coding genes
TSSa-RNAs	20–90 bp	–250 and +50 bp of TSSs	>10,000	Maintenance of transcription?	Associated with *RNF12* and *CCDC52* genes
PROMPTs	<200 bp	–205 bp and –5 kb of TSSs	Unknown	Activation of transcription?	Associated with *EXT1* and *RBM39* genes
Long ncRNAs					
lincRNAs	>200 bp	Widespread loci	>1,000	Examples include scaffold DNA–chromatin complexes	*HOTAIR*, *HOTTIP*, *lincRNA-p21*
T-UCRs	>200 bp	Widespread loci	>350	Regulation of miRNA and mRNA levels?	uc.283+, uc.338, uc160+
Other lncRNAs	>200 bp	Widespread loci	>3,000	Examples include X-chromosome inactivation, telomere regulation, imprinting	*XIST*, *TSIX*, TERRAs, *p15AS*, *H19*, *HYMAI*

Esteller (2011) *Nat Rev Genet*

Non-coding RNAs in non-cancer disease

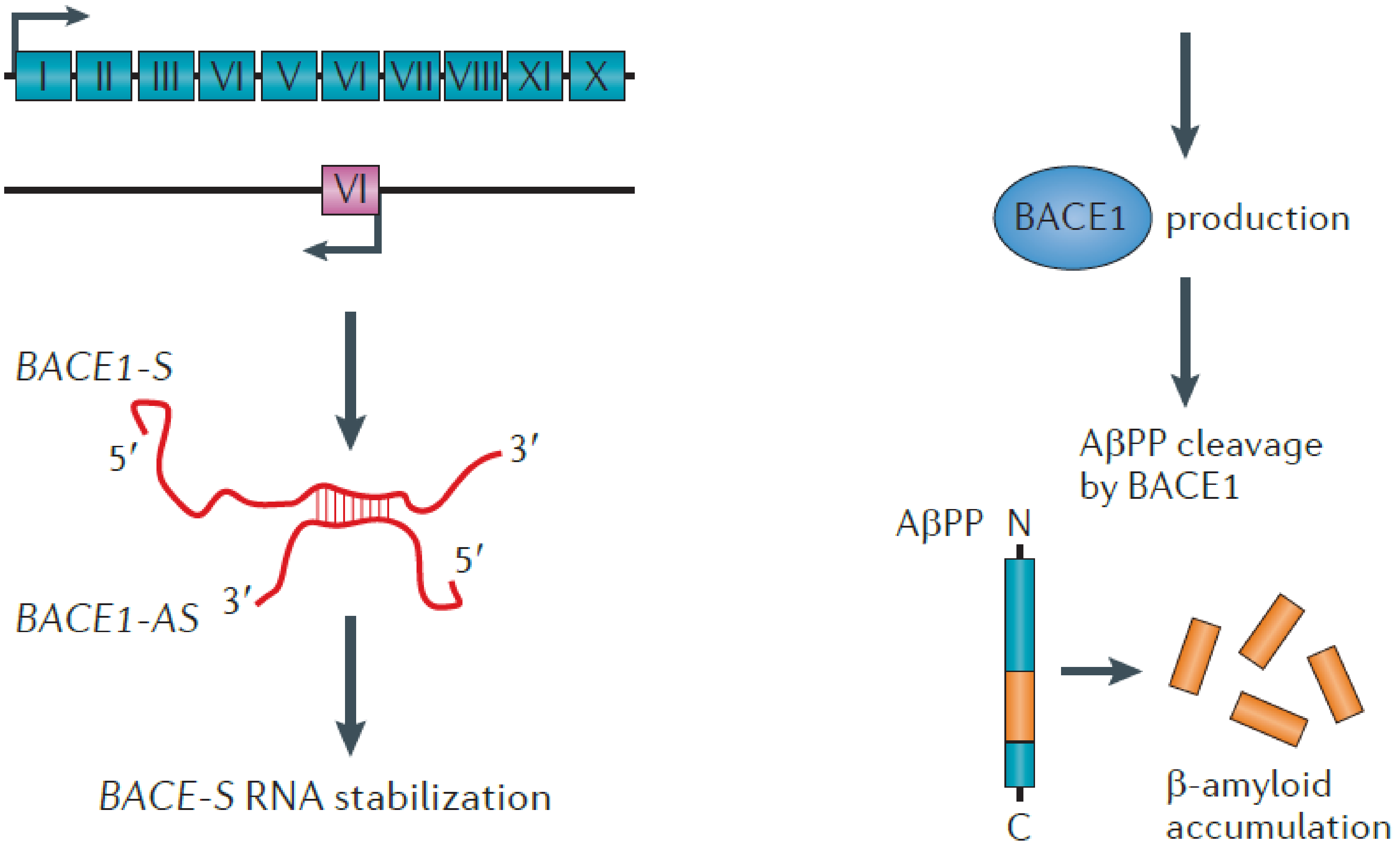


Disease	Involved ncRNAs	ncRNA type
Spinal motor neuron disease	miR-9	miRNA
Spinocerebellar ataxia type 1	miR-19, miR-101, miR-100	miRNA
Amyotropic lateral sclerosis	miR-206	miRNA
Arrhytmia and hypertension	miR-1	miRNA
Atheromatosis and atherosclerosis	miR-10a, miR-145, mR-143 and miR-126	miRNA
Atheromatosis and atherosclerosis	Circular ncRNA linked to the CDKN2A locus	lncRNA
Cardiac hypertrophy	miR-21	miRNA
Rett's syndrome	miR-146a, miR-146b, miR-29 and miR-382	miRNA
5q syndrome	miR-145 and miR-146a	miRNA
ICF syndrome	miR-34b, miR-34c, miR-99b, let-7e and miR-125a	miRNA
Crohn's disease	miR-196	miRNA
Prader–Willi and Angelman syndromes	snoRNA cluster at 15q11–q13 imprinted locus	snoRNA
Beckwith–Wiedeman syndrome	lncRNAs *H19* and *KCNQ1OT1*	lncRNA
Uniparental disomy 14	snoRNA cluster at 14q32.2 imprinted locus	snoRNA
Silver–Russell syndrome	lncRNA *H19*	lncRNA
Silver–Russell syndrome	miR-675	miRNA
McCune–Albright syndrome	lncRNA *NESP-AS*	lncRNA
Deafness	miR-96	miRNA
Alzheimer's disease	miR-29, miR-146 and miR-107	miRNA
Alzheimer's disease	ncRNA antisense transcript for *BACE1*	lncRNA

Exercise: research a ncRNA-related disease

Esteller (2011) *Nat Rev Genet*

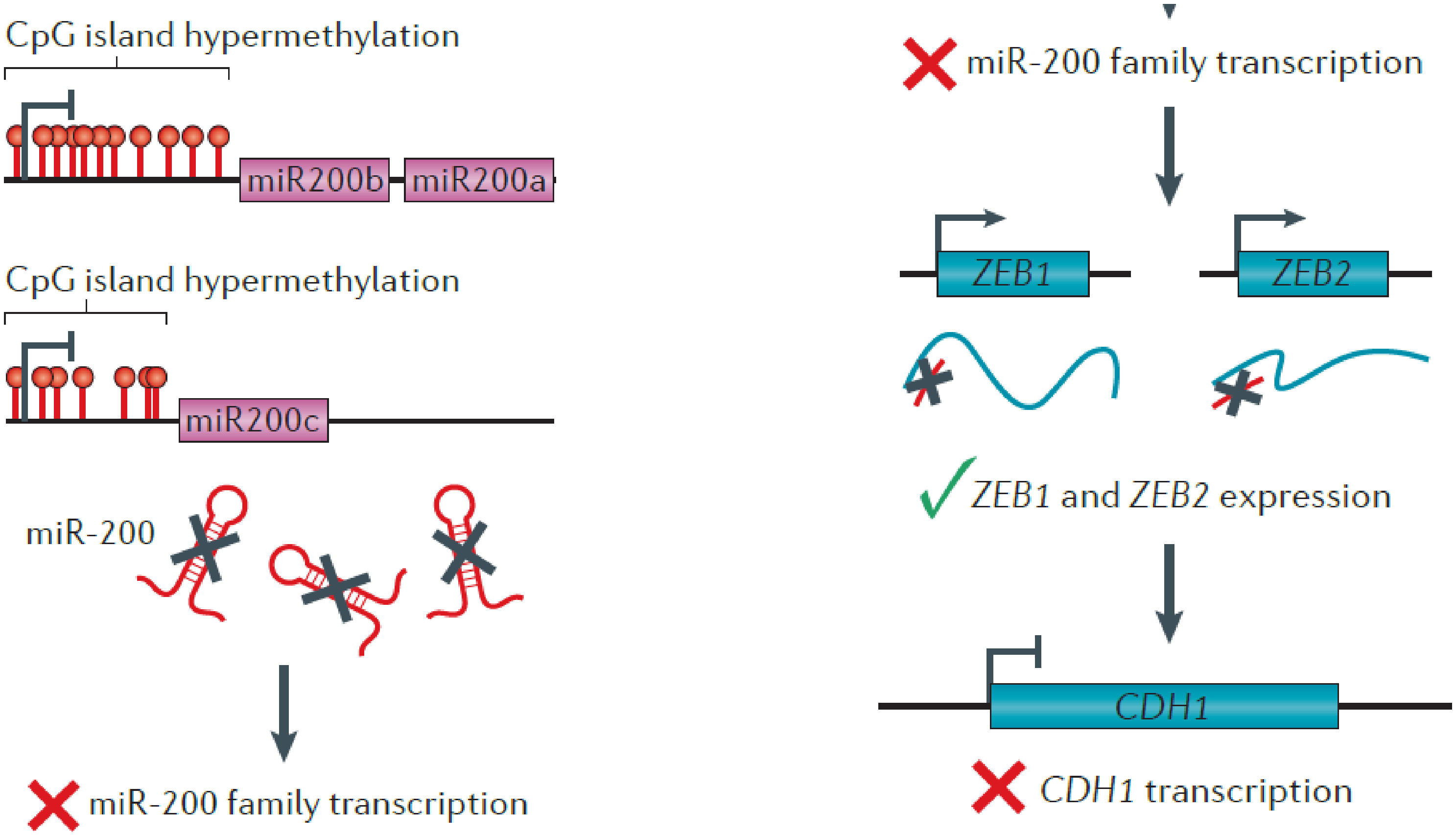
Non-coding RNAs in Alzheimer’s disease



An antisense lncRNA, *BACE1□AS*, regulates the expression of the sense *BACE1* gene (labelled *BACE1□S* in the figure) through the stabilization of its mRNA. *BACE1□AS* is elevated in Alzheimer’s disease, increasing the amount of BACE1 protein and, subsequently, the production of β□amyloid peptide.

Esteller (2011) *Nat Rev Genet*

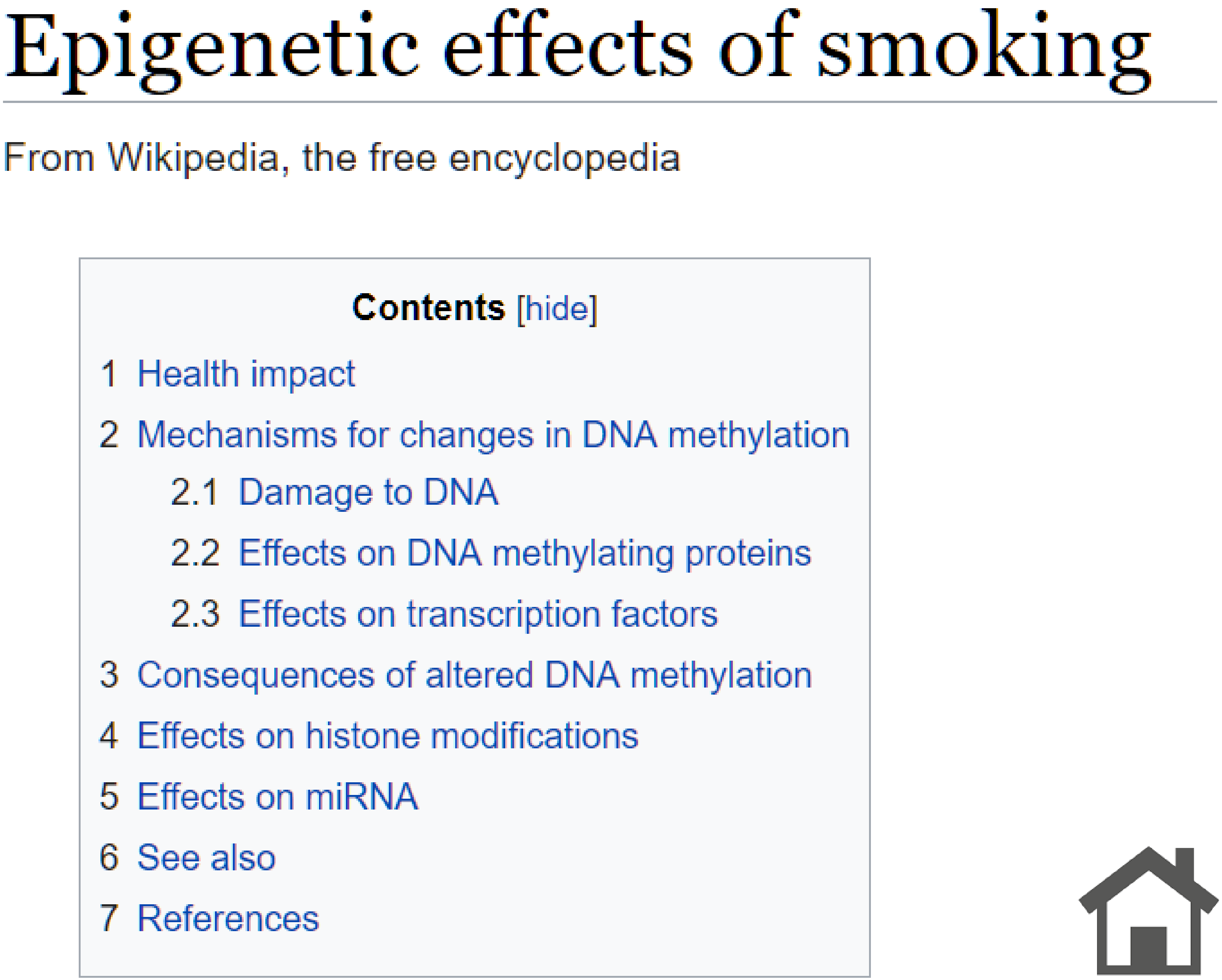
Non-coding RNAs in cancer



CpG island hypermethylation
miR200b
miR200a
CpG island hypermethylation
miR200c
miR-200
✗ miR-200 family transcription
✗ miR-200 family transcription
ZEB1
ZEB2
✓ ZEB1 and ZEB2 expression
CDH1
✗ CDH1 transcription

Alterations in the epigenetic regulation of the miR□200 family are involved in epithelial-to-mesenchymal transition in cancer. Specifically, CpG island hypermethylation-associated silencing of these miRNAs in human tumours causes an upregulation of the zinc finger E-box-binding homeobox (HOX) 1 (*ZEB1*) and *ZEB2* transcriptional repressors, which, in turn, leads to a downregulation of E-cadherin *CDH1*

Esteller (2011) *Nat Rev Genet*

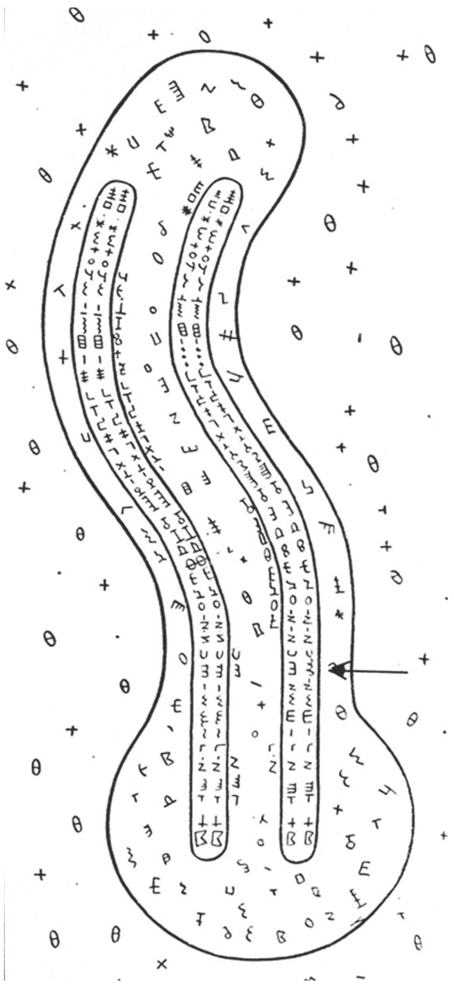


Epigenetic effects of smoking

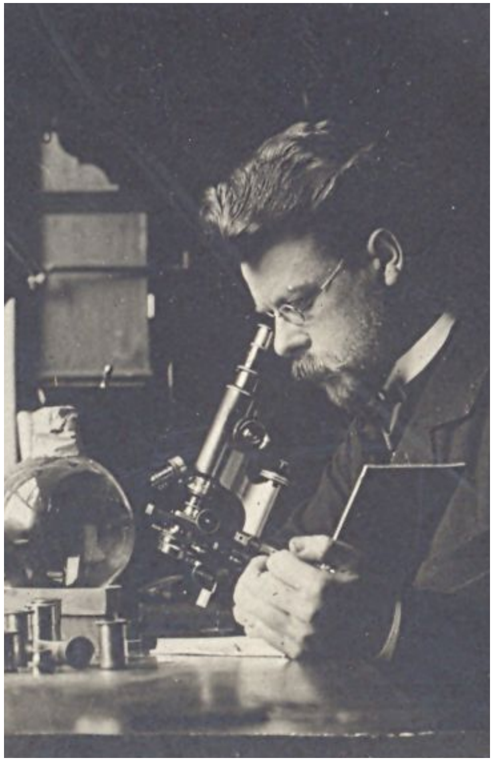
From Wikipedia, the free encyclopedia

Contents [hide]

Николай Конст. Кольцов (1872-1940)



- 1915: «Следует признать гены способными... к мутациям. Ведь во всяком органическом соединении атом водорода может быть скачкообразно заменен группой CH_3»
- 1927: *Omnis molecula ex molecula:* гипотеза о матричном воспроизведении молекул наследственности



Кольцов
1927

→

Тимофеев-Ресовский, Циммер,
Дельбрюк, Шредингер
1935-1945

→

Уотсон, Крик
1953

Examples of coding changes in *RBFOX1*

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAGCAGCTAAGGgtagg
                       M  N  C  E  R  E  Q  L  R
```

Synonymous change

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAGCAACTAAGGgtagg
                       M  N  C  E  R  E  Q  L  R
```

Non-synonymous (missense)

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAGCACCTAAGGgtagg
                       M  N  C  E  R  E  H  L  R
```

Stop gain (nonsense)

```
tttctagGTTTCAAGACAACAGATGAATTGAGAAAGAGAGCAGCTAAGGgtagg
                       M  N  *  E  R  E  Q  L  R
```

Examples of coding changes in *RBFOX1*

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAGCAGCTAAGGgtagg
                       M  N  C  E  R  E  Q  L  R
```

Inframe deletion

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAG---CTAAGGgtagg
                       M  N  C  E  R  E  -  L  R
```

```
tttctagGTTTCAAGACAACAGATGAATTGTGAAAGAGAGCAGCTAAGGgtagg
                       M  N  C  E  R  E  Q  L  R
```

Frameshift deletion

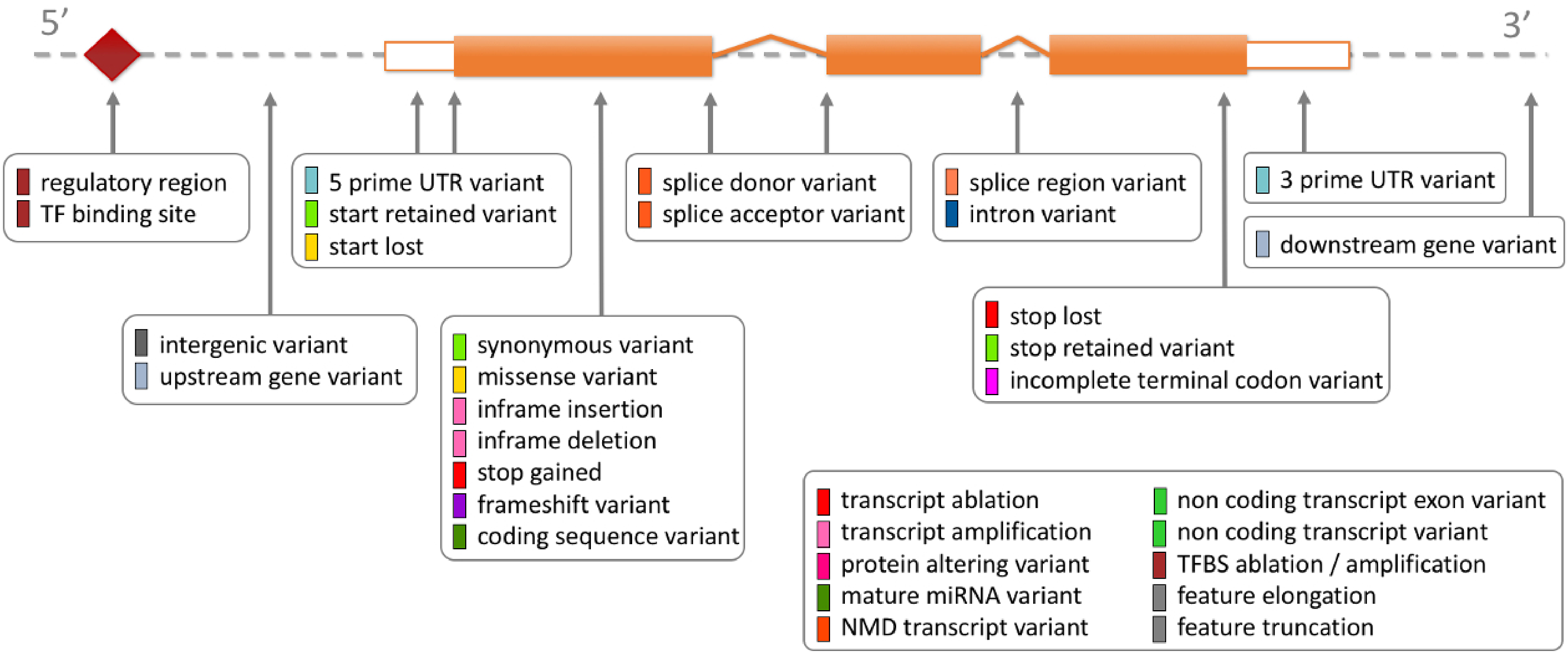
```
tttctagGTTTCAAGACAACAGATGA--TGTGAAAGAGAGCAGCTAAGGgtagg
                       M    M  *  K  R  A  A  K
```

Ve!P

ENSEMBL Variant Effect Predictor

Variation consequences

Promoter ♦ 5’-UTR ♦ Start (ATG) ♦ Donor(GT) ♦ Acceptor(AG) ♦ … ♦ Stop(TAA,…) ♦ 3’-UTR



■ regulatory region
■ TF binding site

■ intergenic variant
■ upstream gene variant

■ 5 prime UTR variant
■ start retained variant
■ start lost

■ synonymous variant
■ missense variant
■ inframe insertion
■ inframe deletion
■ stop gained
■ frameshift variant
■ coding sequence variant

■ splice donor variant
■ splice acceptor variant

■ splice region variant
■ intron variant

■ stop lost
■ stop retained variant
■ incomplete terminal codon variant

■ 3 prime UTR variant

■ downstream gene variant

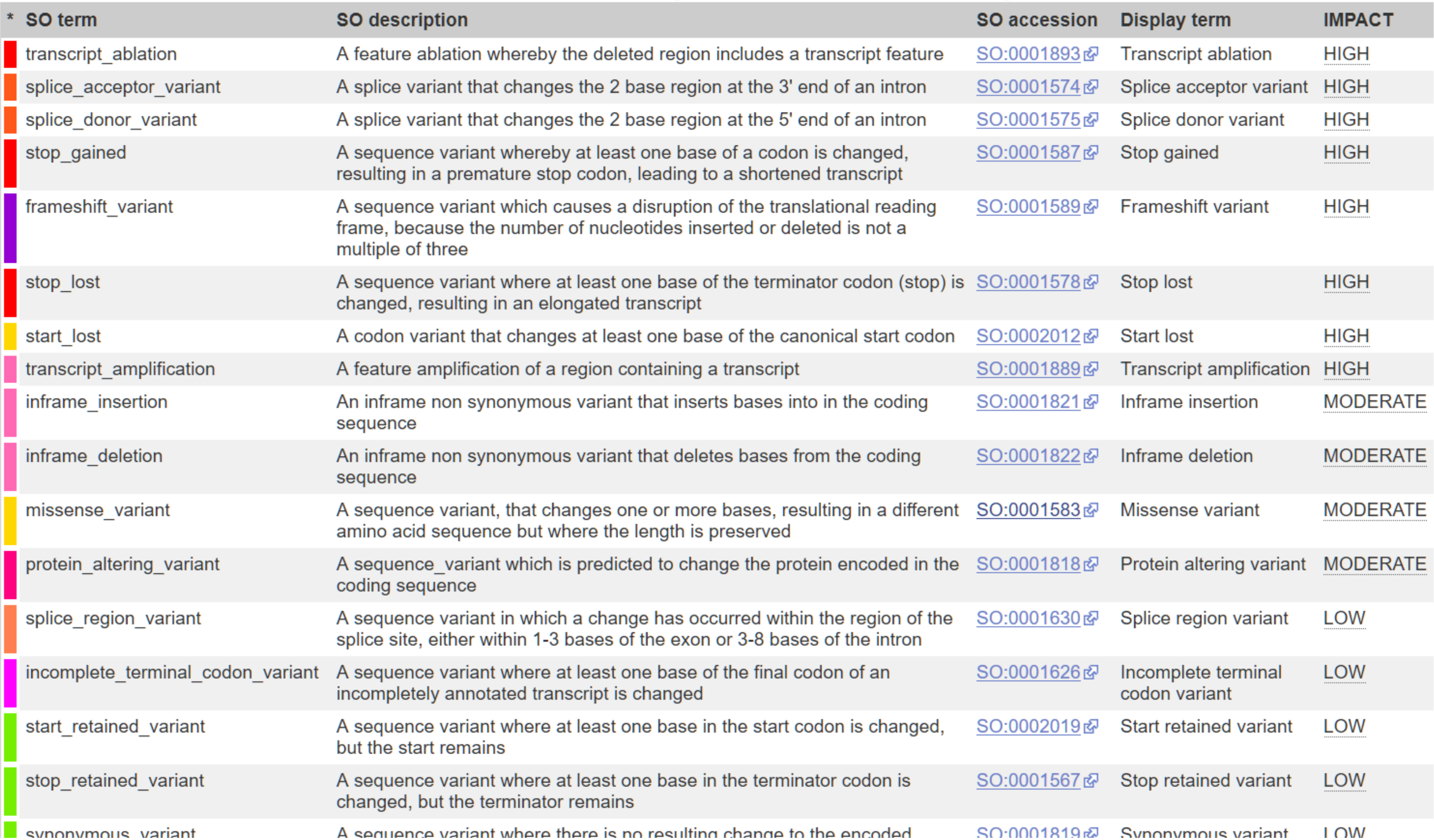
■ transcript ablation
■ transcript amplification
■ protein altering variant
■ mature miRNA variant
■ NMD transcript variant

■ non coding transcript exon variant
■ non coding transcript variant
■ TFBS ablation / amplification
■ feature elongation
■ feature truncation

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

ENSEMBL Variant Effect Predictor

Variation consequences and impact

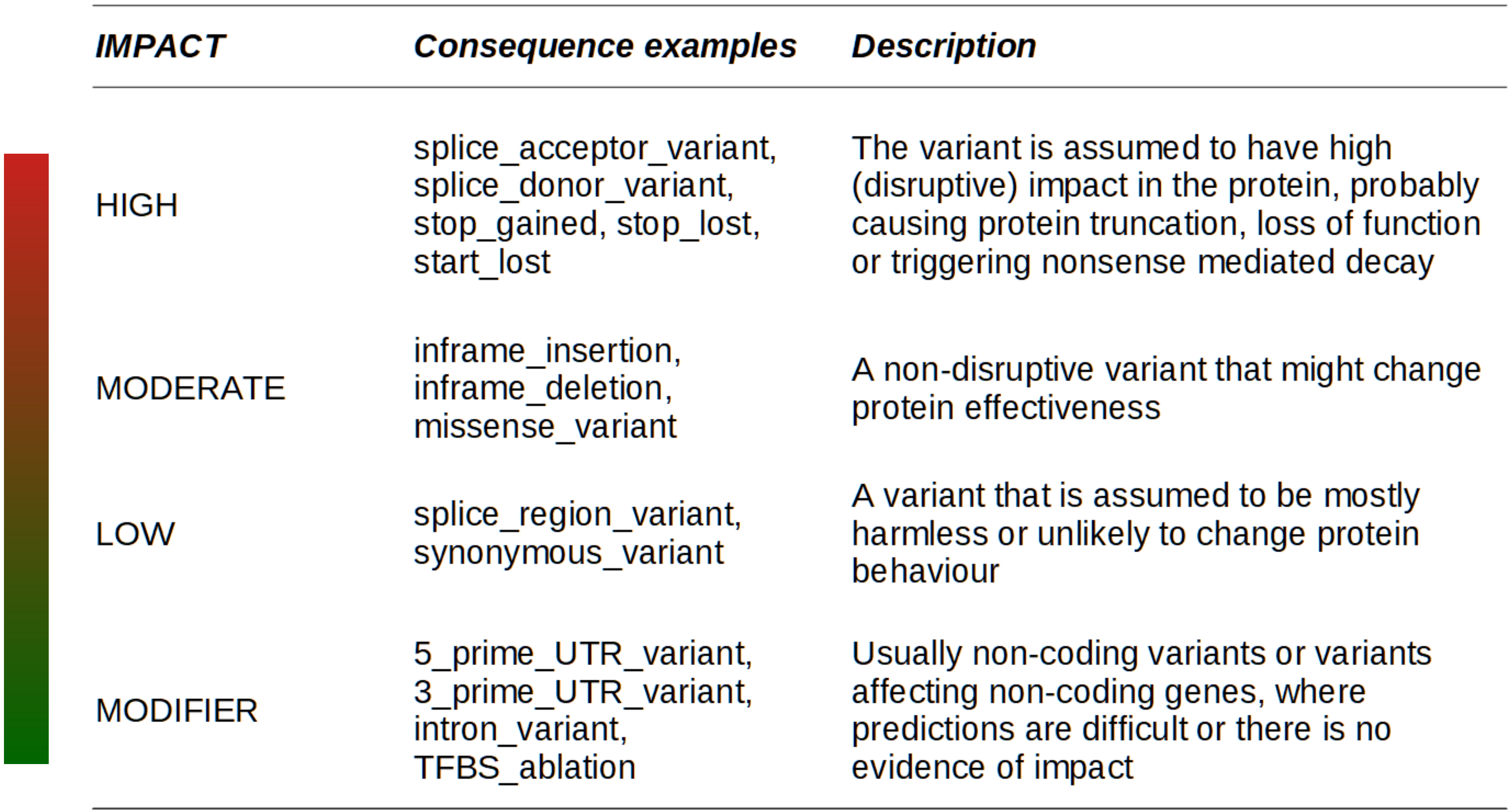


*	SO term	SO description	SO accession	Display term	IMPACT
🟥	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
🟧	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
🟧	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
🟥	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
🟪	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
🟥	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
🟨	start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
🟪	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
🟪	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
🟪	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
🟨	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
🟪	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
🟧	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
🟪	incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
🟩	start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
🟩	stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
🟩	synonymous_variant	A sequence variant where there is no resulting change to the encoded	SO:0001819	Synonymous variant	LOW

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

ENSEMBL Variant Effect Predictor

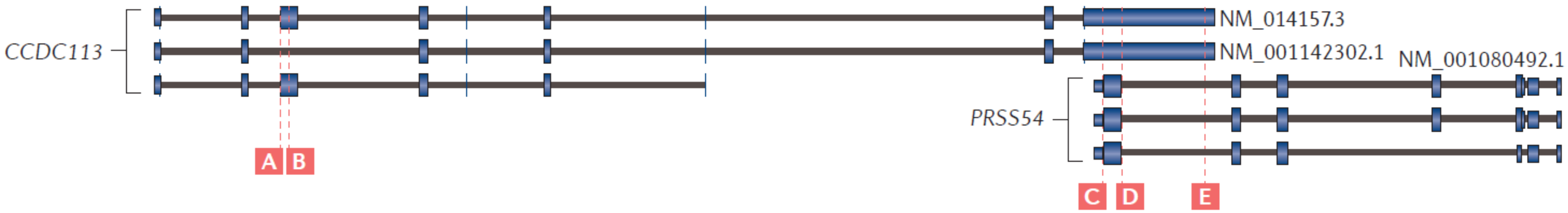
Variation consequences and impact



IMPACT	*Consequence examples*	*Description*
HIGH	splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost, start_lost	The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay
MODERATE	inframe_insertion, inframe_deletion, missense_variant	A non-disruptive variant that might change protein effectiveness
LOW	splice_region_variant, synonymous_variant	A variant that is assumed to be mostly harmless or unlikely to change protein behaviour
MODIFIER	5_prime_UTR_variant, 3_prime_UTR_variant, intron_variant, TFBS_ablation	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

Complexity of variant annotation

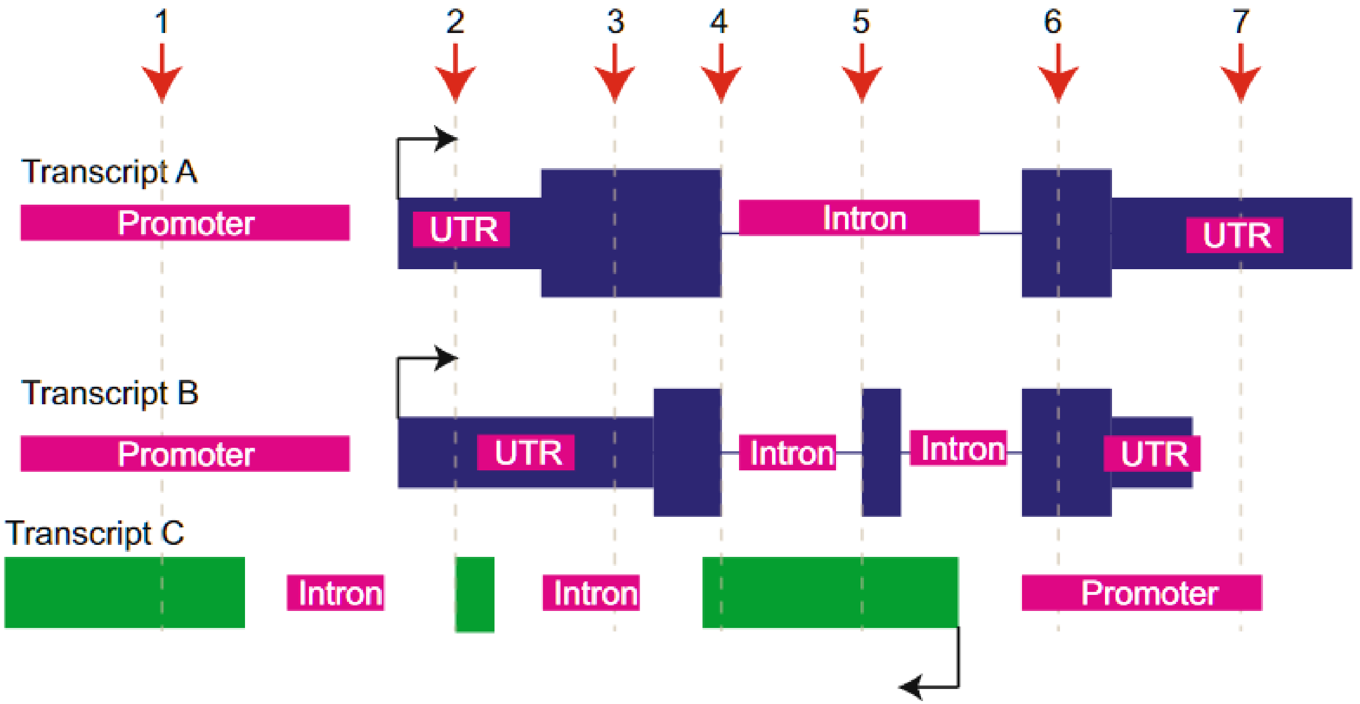


		Variant allele	Gene	Transcript change	RefSeq	Protein change	Molecular consequence
A	rs765957496	G	*CCDC113*	c.228+1143A>G	NM_001142302.1	—	Intron variant
		G	*CCDC113*	c.229•2A>G	NM_014157.3	—	Splice acceptor variant
B	rs775877153	A	*CCDC113*	c.228+1182T>A	NM_001142302.1	—	Intron variant
		A	*CCDC113*	c.266T>A	NM_014157.3	Met89Lys	Missense variant
C	rs780162055	T	*PRSS54*	c.1135G>A	NM_001080492.1	Glu379Lys	Missense variant
		T	*CCDC113*	c.*500C>T	NM_001142302.1	—	3' UTR variant
D	rs776101237	A	*PRSS54*	c.655-2A>T	NM_001080492.1	—	Splice acceptor variant
		A	*CCDC113*	c.*962T>A	NM_001142302.1	—	3' UTR variant
E	rs745863465	C	*PRSS54*	c.655-18T>G	NM_001080492.1	—	Intron variant
		C	*CCDC113*	c.*996A>C	NM_001142302.1	—	3' UTR variant

A demonstration of the multiple possible effects of a single variant across transcripts and genes. The complexity of genomic annotation adds to the complexity of variant annotation. In this example, two genes, coiled-coil domain-containing 113 (*CCDC113*) and protease serine 54 (*PRSS54*) overlap on different strands of the genome, and both have multiple observed transcripts. Variants intersecting this extent of the genome show different effects depending on the gene and the transcript inspected.

Eilbeck (2017) *Nat Rev Genet*

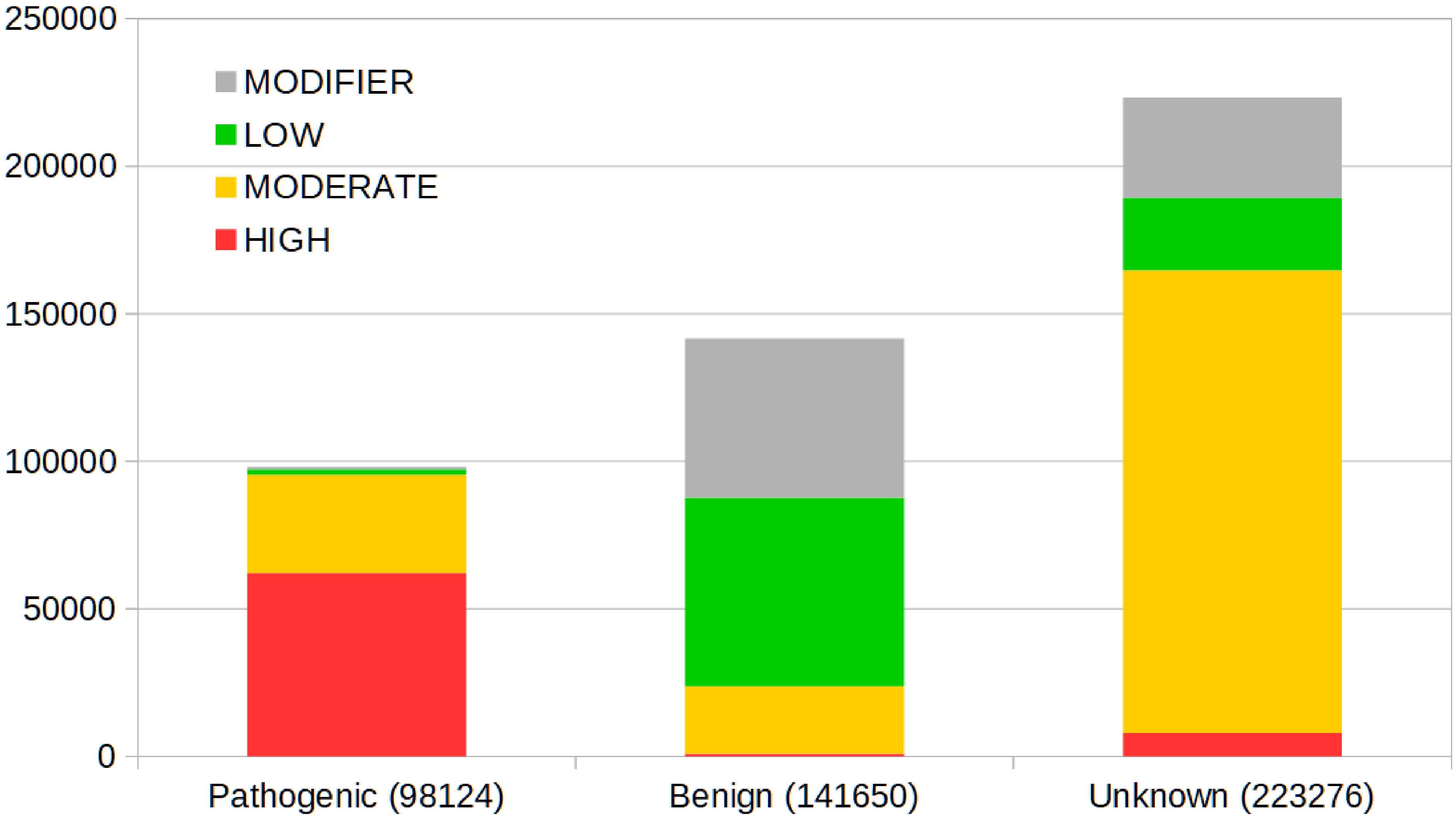
Complexity of variant annotation



Variant	Transcript A	Transcript B	Transcript C
1	Promoter	Promoter	Exon
2	Non Coding Exon	Non Coding Exon	Non Coding Splice
3	Coding Exon	Non Coding Exon	Intron
4	Coding Splice	Coding Splice	Non Coding Exon
5	Intron	Coding Splice	Non Coding Exon
6	Coding Exon	Coding Exon	Promoter
7	Non Coding Exon	Downstream	Prompter

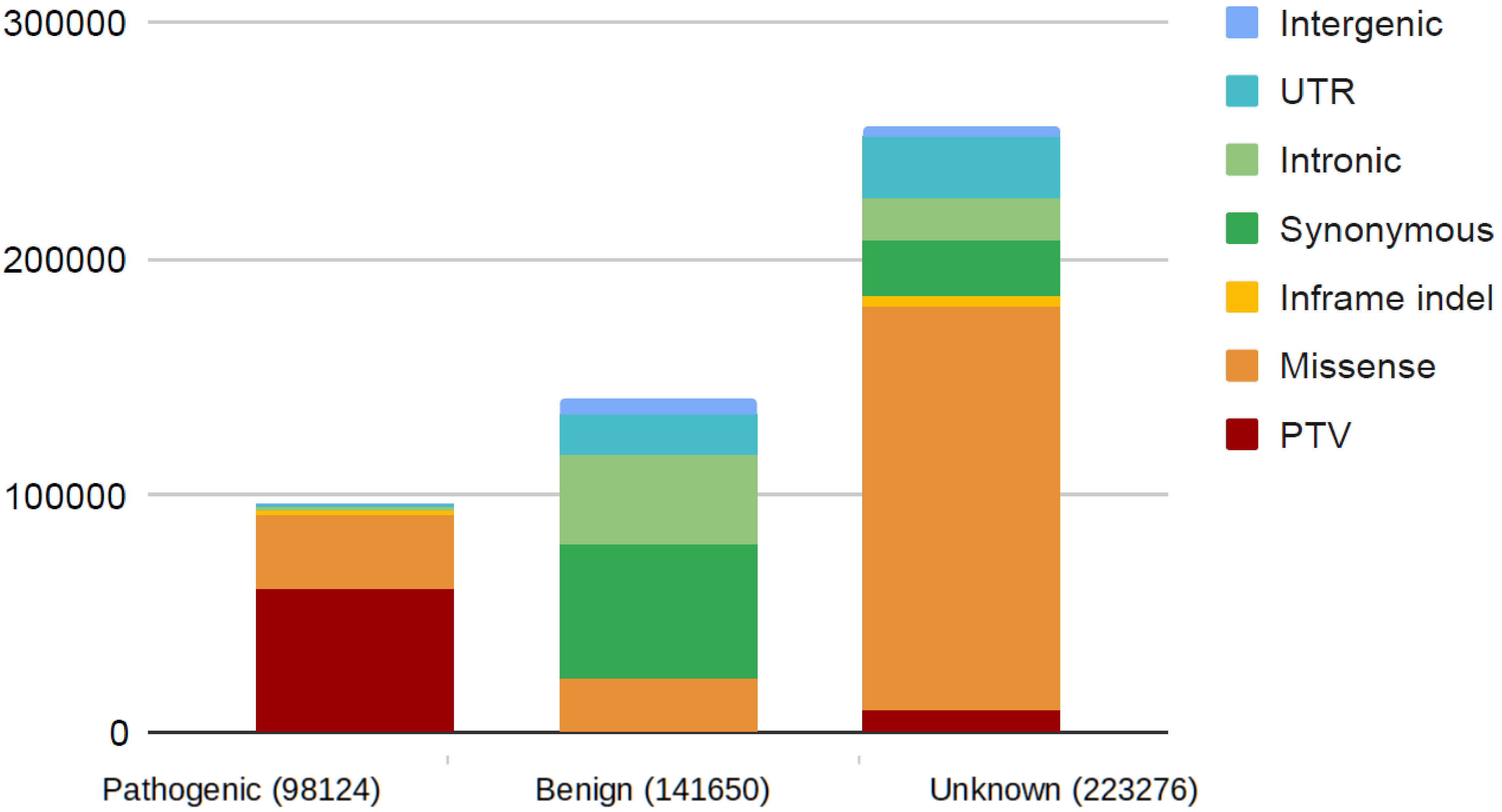
Gloss (2018) *Exp Mol Med*

Ensembl VEP annotation for ClinVar variants



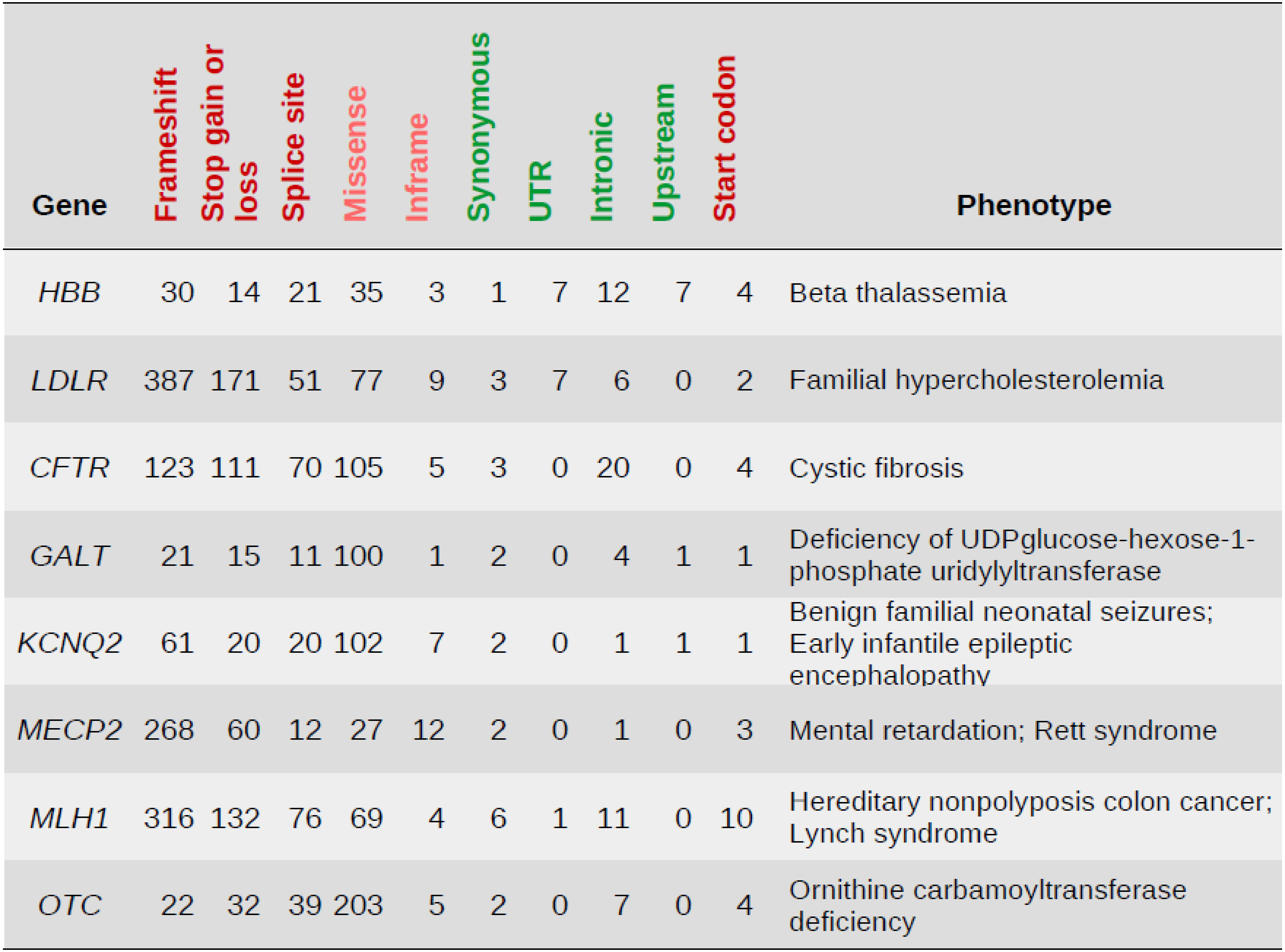
ClinVar (Oct. 2019), 498,742 variants annotated with Ensembl VEP

Ensembl VEP annotation for ClinVar variants



ClinVar (Oct. 2019), 498,742 variants annotated with Ensembl VEP

Pathogenic variants in ClinVar (Oct. 2019)



Gene	Frameshift	Stop gain or loss	Splice site	Missense	Inframe	Synonymous	UTR	Intronic	Upstream	Start codon	Phenotype
HBB	30	14	21	35	3	1	7	12	7	4	Beta thalassemia
LDLR	387	171	51	77	9	3	7	6	0	2	Familial hypercholesterolemia
CFTR	123	111	70	105	5	3	0	20	0	4	Cystic fibrosis
GALT	21	15	11	100	1	2	0	4	1	1	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase
KCNQ2	61	20	20	102	7	2	0	1	1	1	Benign familial neonatal seizures; Early infantile epileptic encephalopathy
MECP2	268	60	12	27	12	2	0	1	0	3	Mental retardation; Rett syndrome
MLH1	316	132	76	69	4	6	1	11	0	10	Hereditary nonpolyposis colon cancer; Lynch syndrome
OTC	22	32	39	203	5	2	0	7	0	4	Ornithine carbamoyltransferase deficiency

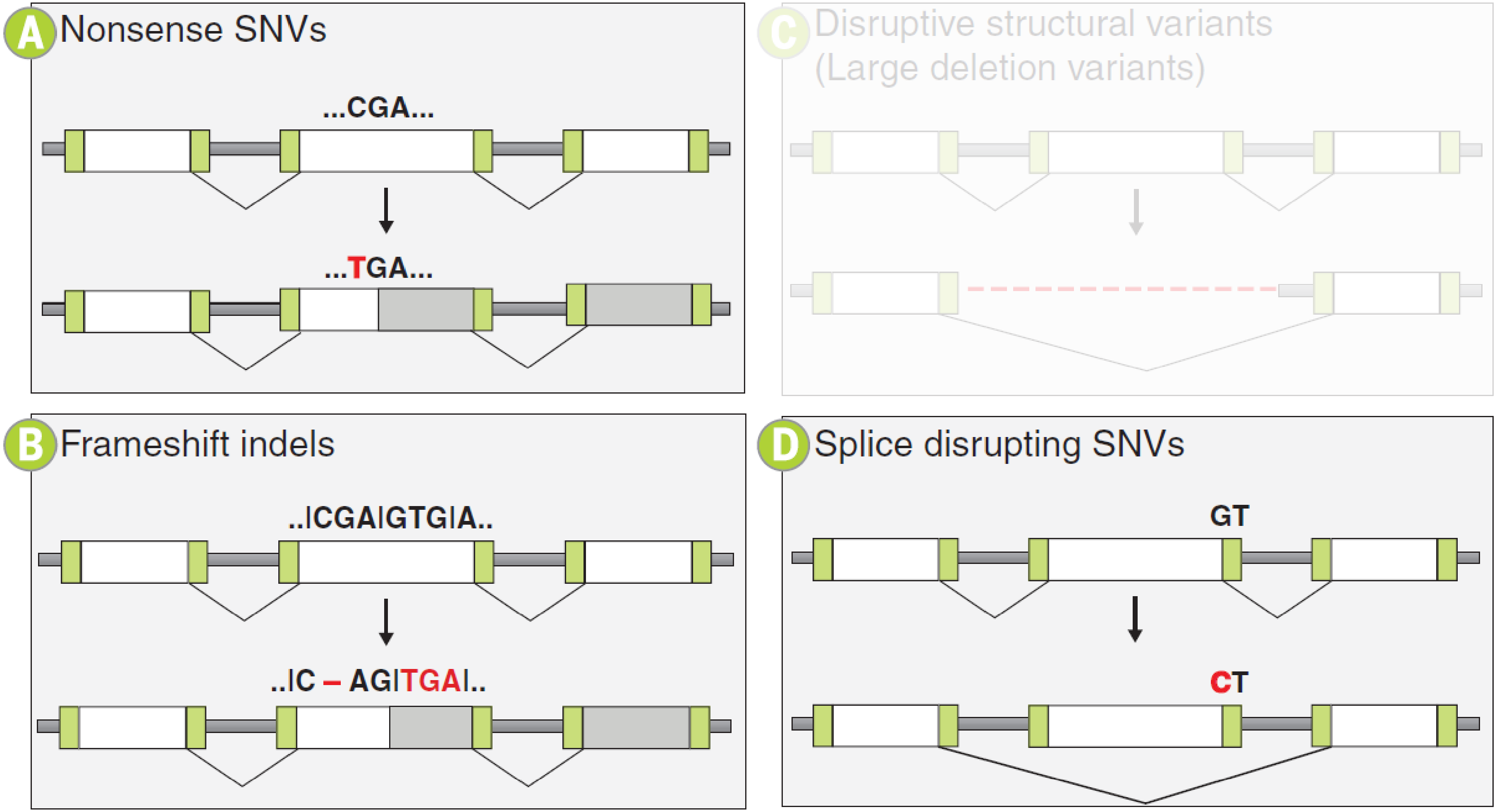
Exercise

Use ClinVar (OMIM) to find and save one example of disease-associated pathogenic mutation for *each* annotation type:

- stop-gain
- synonymous
- missense
- splice-site
- frameshift indel

PTVs and LoF variants

Protein-truncating variants: stop-gain, splice site, frameshift indels.
VEP impact: HIGH.



Rivas (2015) *Science*

PTVs and LoF variants

Protein-truncating variants: stop-gain, splice site, frameshift indels. VEP impact: HIGH. *However, not all PTVs are loss-of-function*

LOFTEE tool (K.Karczewski et al): filters and flags to predict pLoF (putative LoF) from candidate PTVs. https://github.com/konradjk/loftee

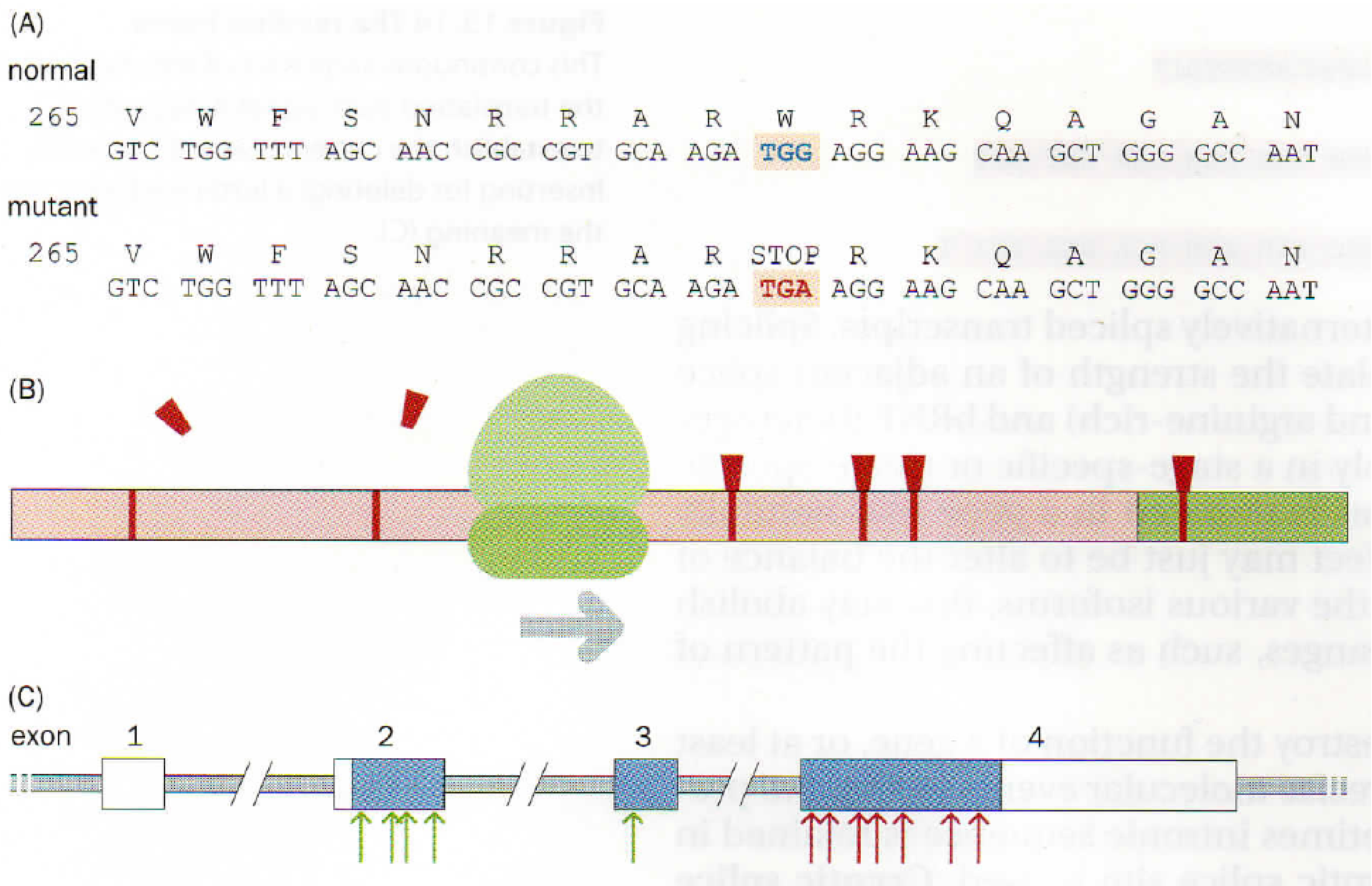
PTVs not predicted as pLoF, examples:

- Stop-gain and frameshift variants near the end of the transcript, based on the 50 bp rule
- Variants in an exon with non-canonical splice sites (GT, AG) around it
- Splice site variants rescued by nearby, in-frame splice site
- Variants in small introns

Flagged PTVs, examples:

- Variants in NAGNAG sites (acceptor sites rescued by in-frame acceptor site)
- Variants that fall in an intron with a non-canonical splice site

PTVs and nonsense-mediated decay (NMD)


(A)
normal
265 V W F S N R R A R W R K Q A G A N
GTC TGG TTT AGC AAC CGC CGT GCA AGA TGG AGG AAG CAA GCT GGG GCC AAT
mutant
265 V W F S N R R A R STOP R K Q A G A N
GTC TGG TTT AGC AAC CGC CGT GCA AGA TGA AGG AAG CAA GCT GGG GCC AAT
(B)
(C)
exon
1
2
3
4

(A) G>A change in exon 6 of the *PAX3* gene (B) Nonsense-mediated decay (NMD). Splice junctions (red bars) retain proteins of the exon junction complex (EJC, red triangles). Ribosome moves along the mRN A and displaces the EJC proteins. If it encounters a premature stop codon and detaches before displacing all EJCs, the mRNA is targeted for degradation. **Stop codons in the last exon or less than 50 nucleotides upstream of the last splice junction (the green zone) do not trigger NMD.** (C) Depending on whether or not a premature stop codon triggers NMD, the consequences of a nonsense mutation can be very different.

Strachan, Read – *Human Molecular Genetics*

PTVs and nonsense-mediated decay (NMD)

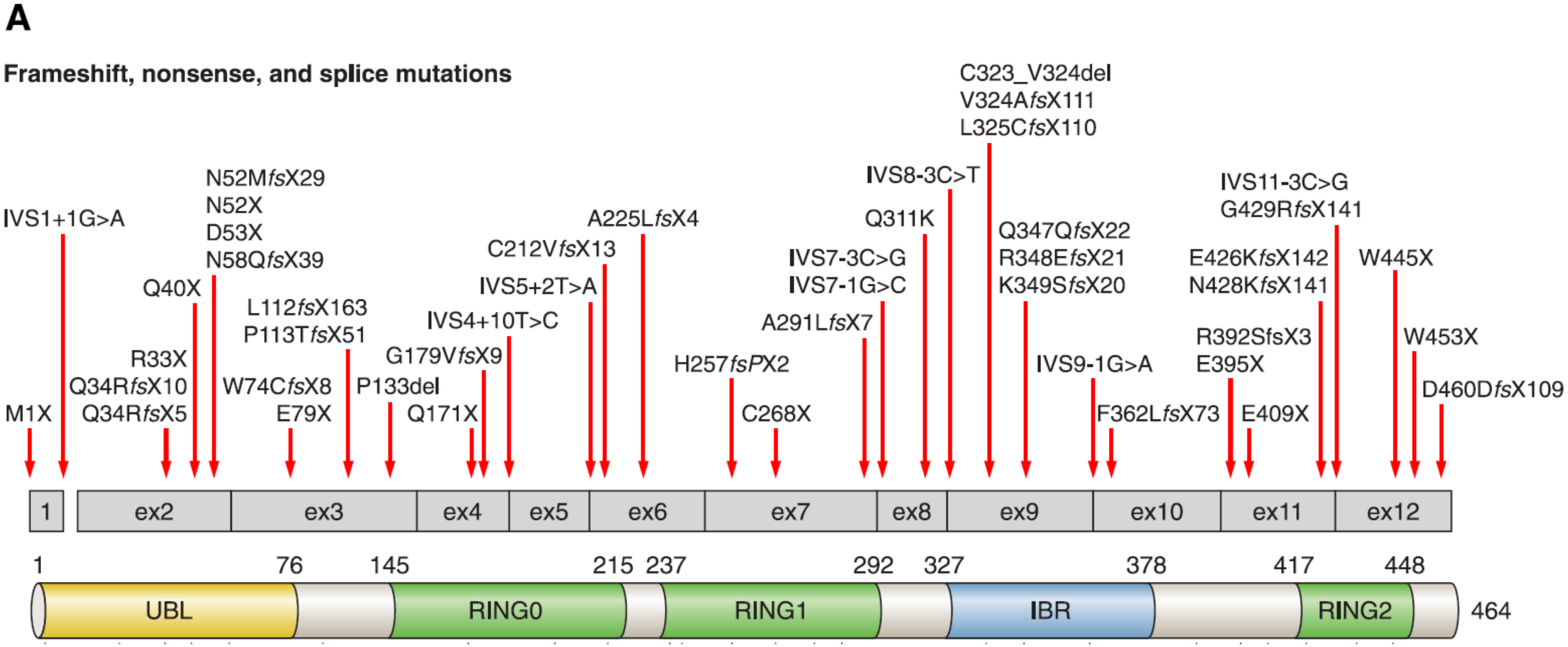
Ideally: **PTV → NMD → Transcript level → Protein level → Cellular functions**

However, variation in mRNA and protein expression levels are often uncorrelated: the reduction in RNA levels may not reduce the protein level, and vice versa

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Impact of Regulatory Variation from RNA to Protein. Science 347, 664–667.

Narasimhan VM, Xue Y, Tyler-Smith C. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? Trends in Molecular Medicine. 2016;22(4):341-351. doi:10.1016/j.molmed.2016.02.006.

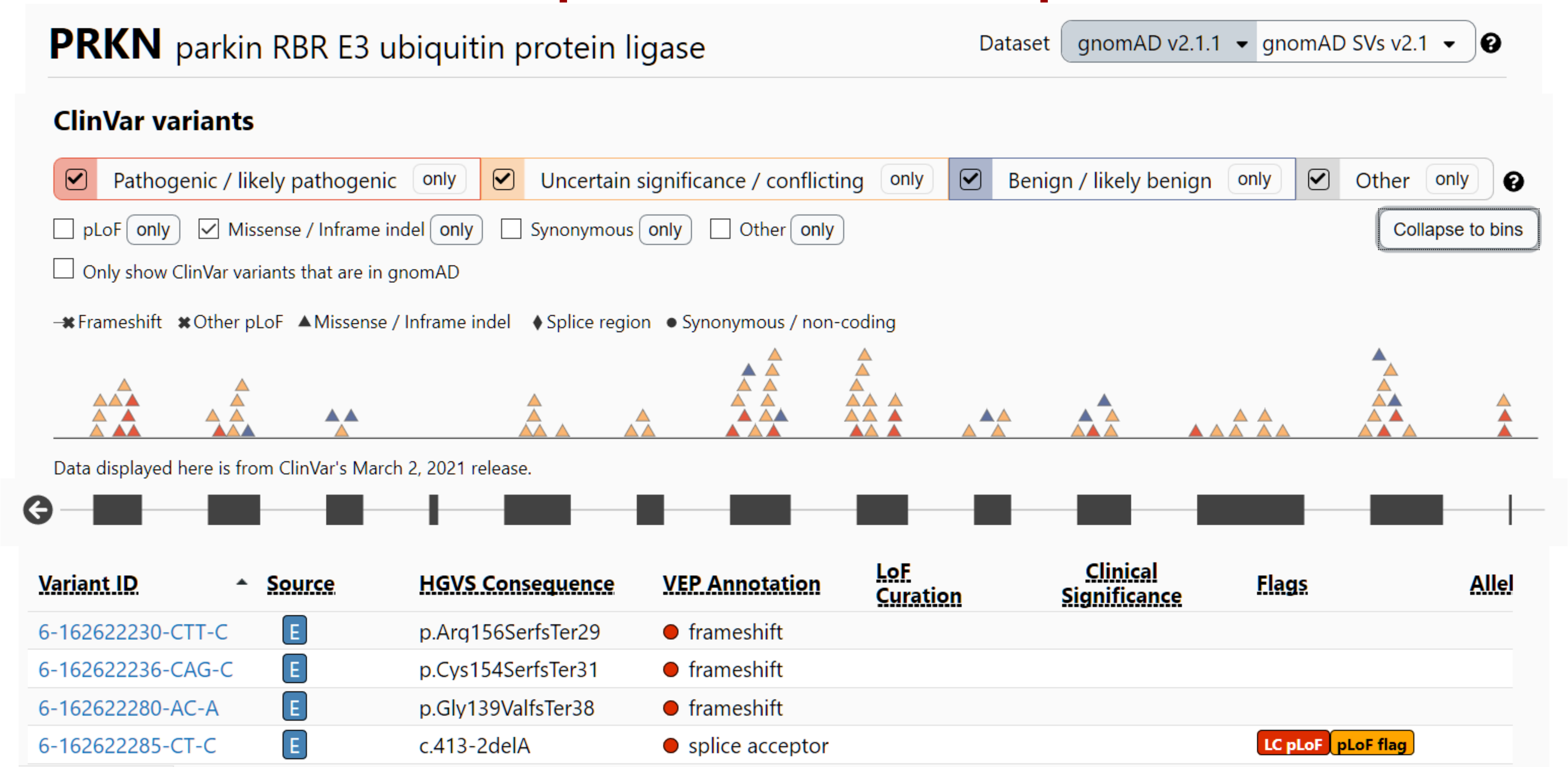
Examples of PTV impact



Mutations in the Parkin RBR E3 Ubiquitin Protein Ligase *PRKN* are the most frequent known cause of early-onset (40–50 yr) Parkinson's disease. PD is the second most common neurodegenerative disorder, after Alzheimer's disease, with prevalence in industrialized countries ~0.3%.

Corti (2011) *Physiol Rev*

Examples of PTV impact



PRKN parkin RBR E3 ubiquitin protein ligase

Dataset: gnomAD v2.1.1 ▾ | gnomAD SVs v2.1 ▾

ClinVar variants

☑ Pathogenic / likely pathogenic only | ☑ Uncertain significance / conflicting only | ☑ Benign / likely benign only | ☑ Other only

□ pLoF only | ☑ Missense / Inframe indel only | □ Synonymous only | □ Other only | Collapse to bins

□ Only show ClinVar variants that are in gnomAD

Data displayed here is from ClinVar's March 2, 2021 release.

Variant ID	Source	HGVS Consequence	VEP Annotation	LoF Curation	Clinical Significance	Flags	Allel
6-162622230-CTT-C	E	p.Arg156SerfsTer29	● frameshift				
6-162622236-CAG-C	E	p.Cys154SerfsTer31	● frameshift				
6-162622280-AC-A	E	p.Gly139ValfsTer38	● frameshift				
6-162622285-CT-C	E	c.413-2delA	● splice acceptor			LC pLoF pLoF flag	

Mutations in the Parkin RBR E3 Ubiquitin Protein Ligase *PRKN* are the most frequent known cause of early-onset (40–50 yr) Parkinson’s disease. PD is the second most common neurodegenerative disorder, after Alzheimer’s disease, with prevalence in industrialized countries ~0.3%.

Corti (2011) *Physiol Rev*

Examples of PTV impact

Protein-truncating variants: stop-gain, splice site, frameshift indels.
VEP impact: HIGH.

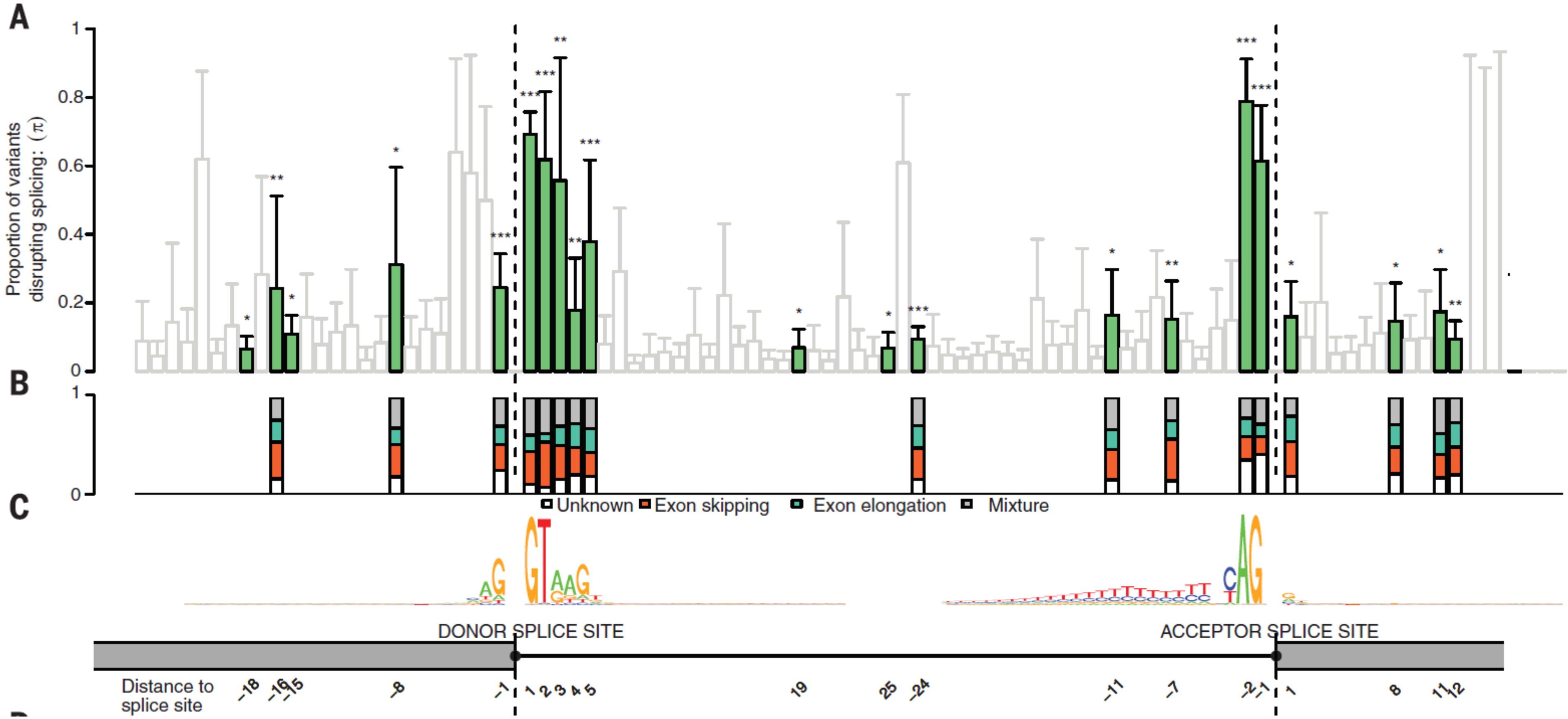


Fig. 3. Splicing disruption. (A) Proportion of variants disrupting splicing at each distance +/- 25 bp from donor and acceptor site (B) Classification of splice disruption events: exon skipping, exon elongation and mixture (C) Diagram of donor and acceptor splice junctions and sequence logo of represented sequences.

Rivas (2015) *Science*

Examples of PTV impact

1. Narasimhan VM, Xue Y, Tyler-Smith C. (2016) Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends Mol Med* 22:341-351.

- A knockout of the immune gene *IRF7* was shown to confer **susceptibility to flu viruses**, leading to life-threatening influenza in an otherwise healthy child (Ciancanelli 2015 *Science*)
- Instances where a naturally-occurring **LoF variant proves beneficial to health**. These discoveries have stimulated drug development:
  - lowering LDL levels: *PCSK9*
  - decreasing susceptibility to HIV: *CCR5*
  - increasing endurance: *ACTN3*
  - increasing sepsis resistance: *CASP12*
  - reduced triglyceride levels in humans: *APOC3*

2. DeBoever, C., Tanigawa, Y., Lindholm, M.E., et al. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat Commun* 9, 1–10.

- 18,228 PTVs × 135 phenotypes; find **27 associations between medical phenotypes and PTVs** in genes outside the MHC

Examples of PTV impact

1. The stop-gain variant in *GNAS* (MIM:139320) is present in the highly variable **first exon** of the gene and is likely to result in nonsense-mediated RNA decay; in contrast, pathogenic *GNAS* variants that cause Albright hereditary osteodystrophy (MIM:103580) are located in **later**, highly constrained exons.

2. Similarly, the stop-gain variant in *TGIF1* (MIM:602630) is located in the **first exon**, where multiple PTVs in gnomAD are also located, but *TGIF1* pathogenic variants causing holoprosencephaly are located in the **final exons**, where they affect DNA binding affinity.

3. Finally, a frameshift deletion in *HIST1H1E* (MIM:142220) is located near **the start** of the single exon of this gene; however, pathogenic *HIST1H1E* frameshift deletions that cause child overgrowth and intellectual disability are located near **the end** of the exon, where they result in a truncated histone protein with lower net charge that is less effective at binding DNA.

We believe that these three rare PTVs are benign because of their locations, despite the fact that they occur in genes that cause dominant DD via haploinsufficiency.

Wright (2019) *Am J Hum Genet*