Медицинская геномика

Василий Евгеньевич Раменский
Анастасия Александровна Жарикова и Мария Ильинична Зайченока

ramensky@gmail.com, azharikova89@gmail.com

НМИЦ Терапии и профилактической медицины
Факультет биоинженерии и биоинформатики МГУ
Институт искусственного интеллекта МГУ

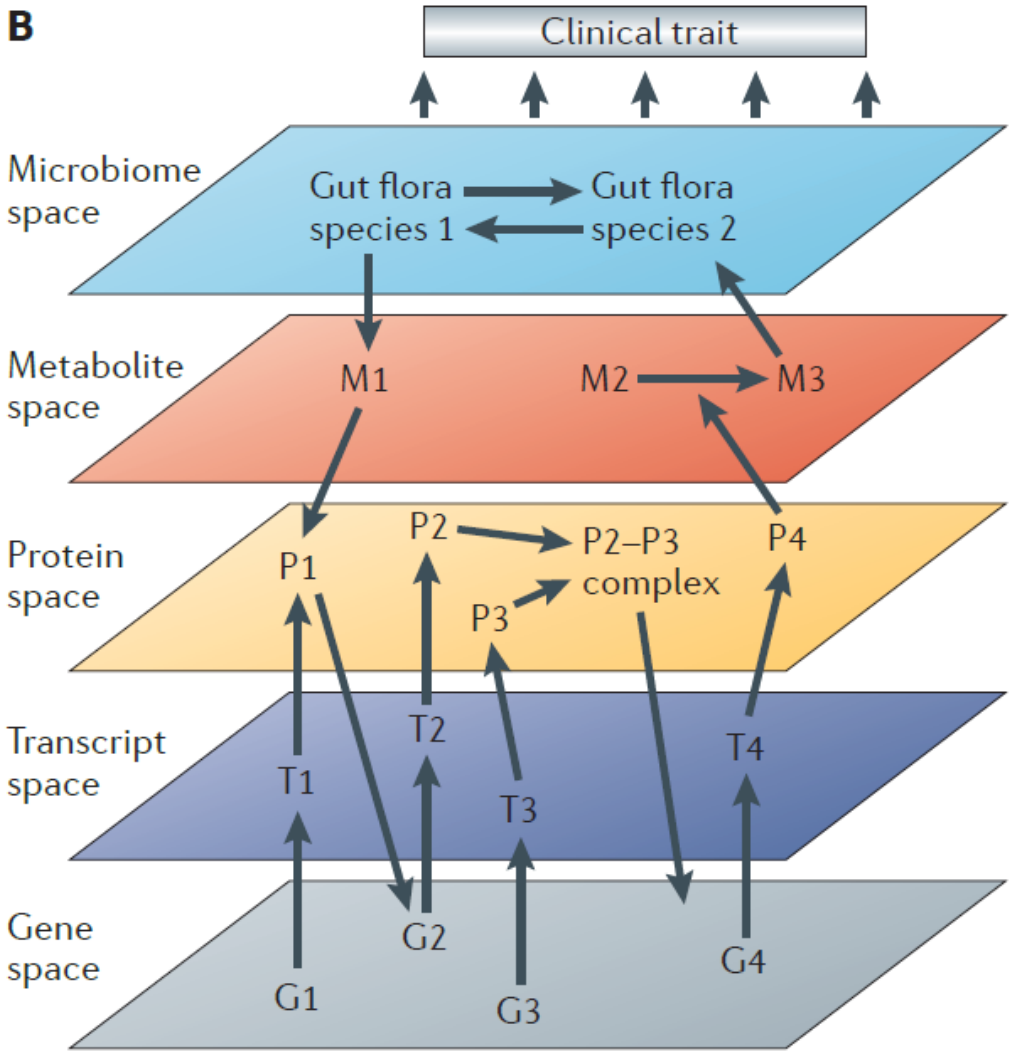
2024

Варианты в пространстве генома и их последствия

План лекции

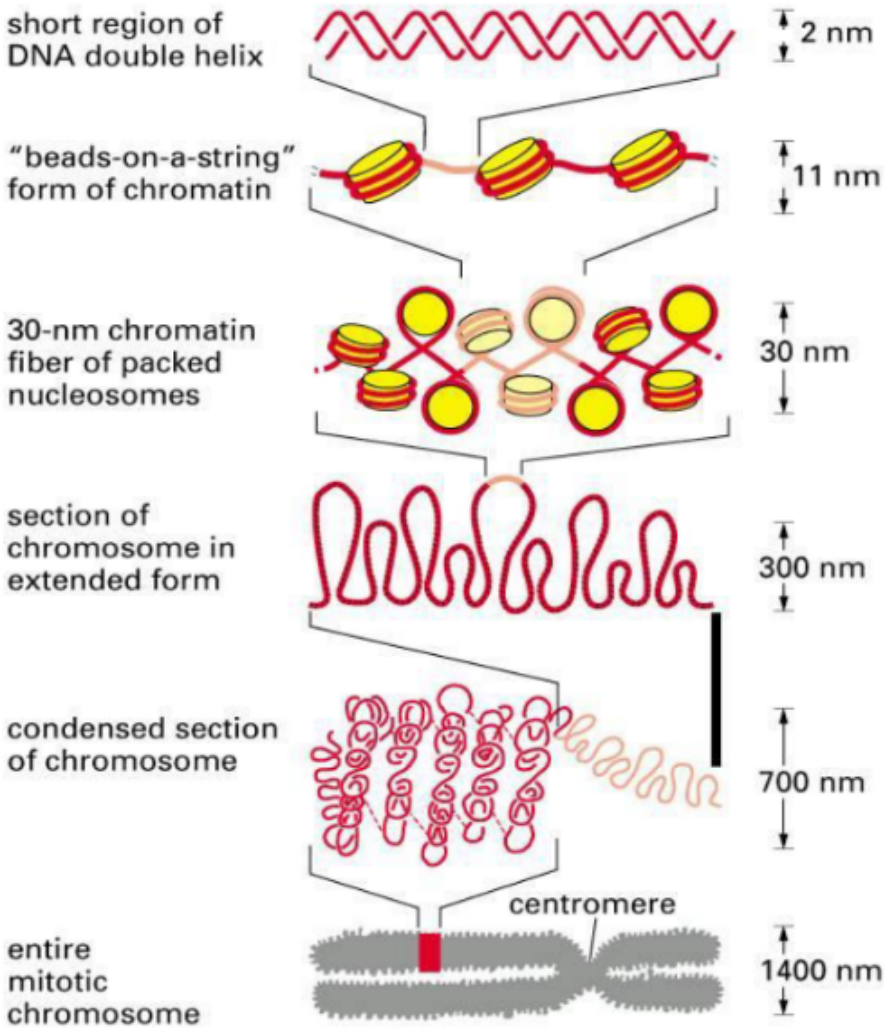
1. Гены в геноме человека
2. Альтернативный сплайсинг
3. Эпигенетика
4. Аннотация вариантов
5. Укорачивающие белок варианты
6. Несинонимичные варианты и заболевания
7. Предсказание эффекта несинонимичных вариантов
8. Варианты других типов

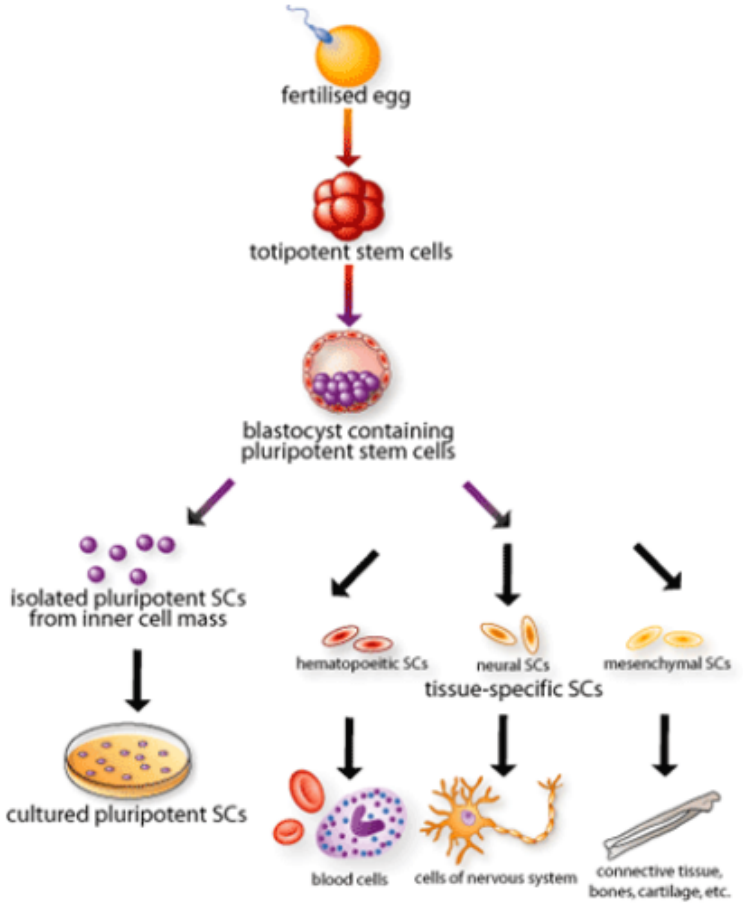
Геном человека в действии



Civelek (2014) *Nat Rev Genet*

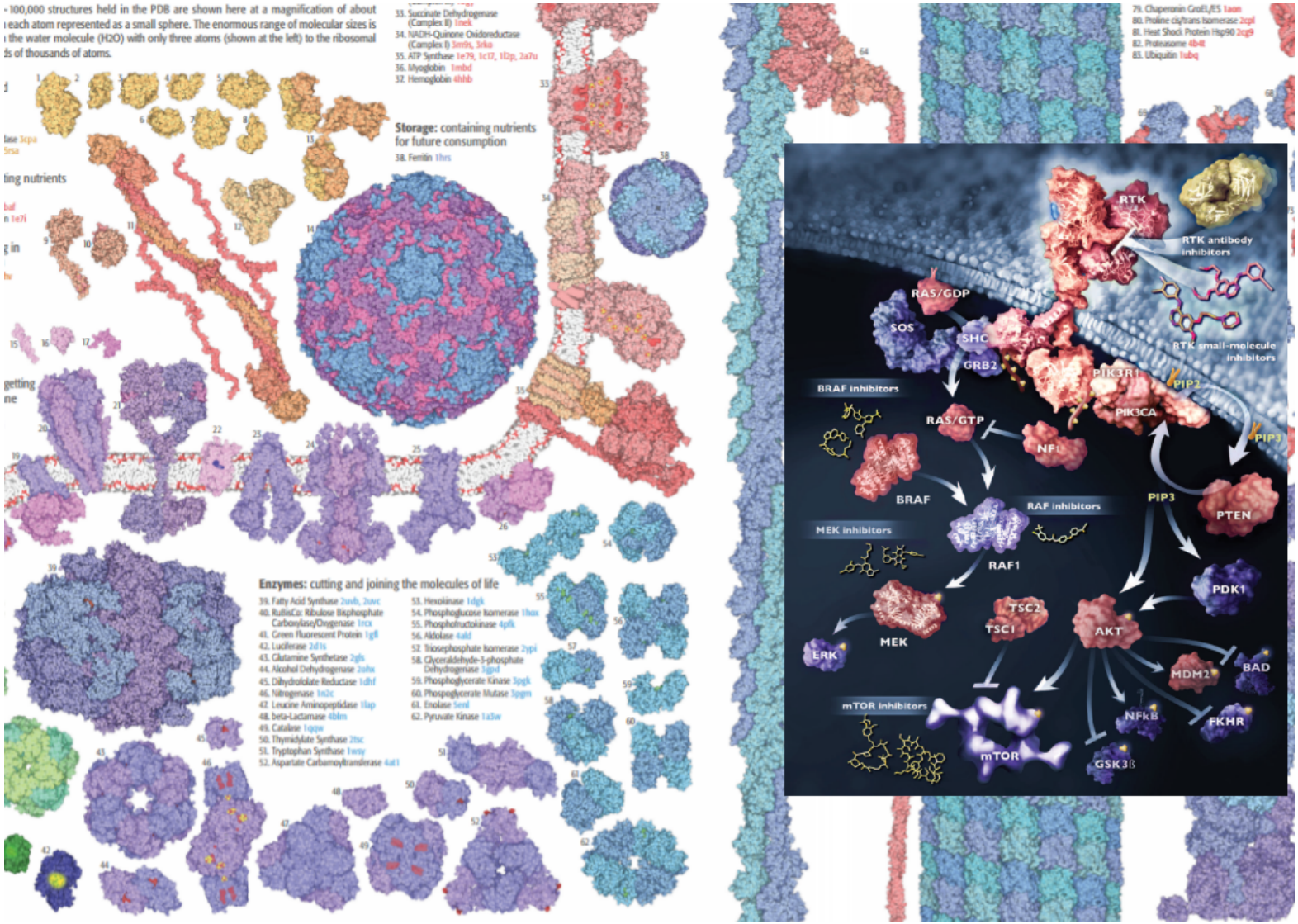
Геном человека в действии: более реалистичная картина





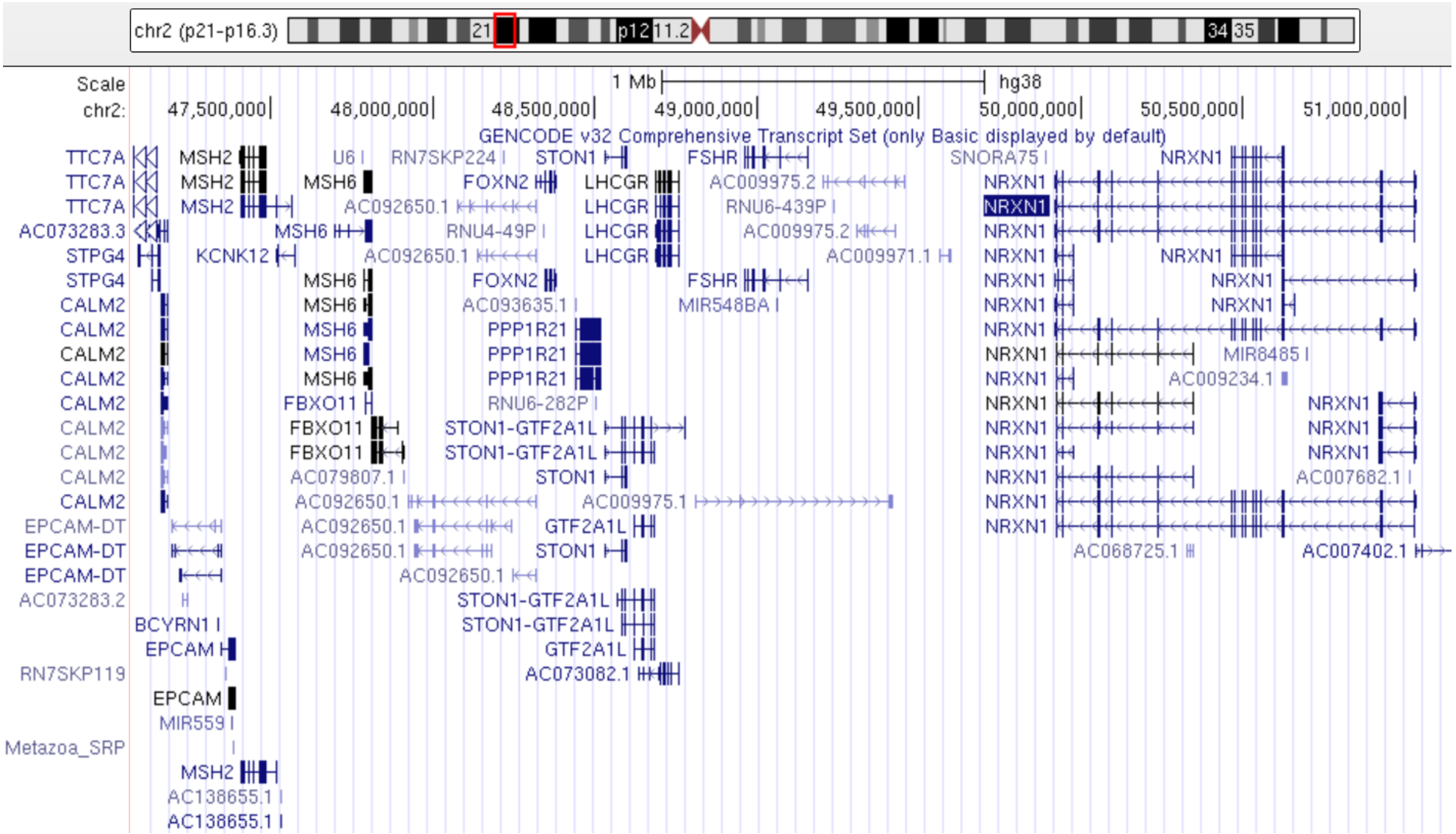
Chaudrey (2004) *Stem Cell Bioeng*

Геном человека в действии: более реалистичная картина

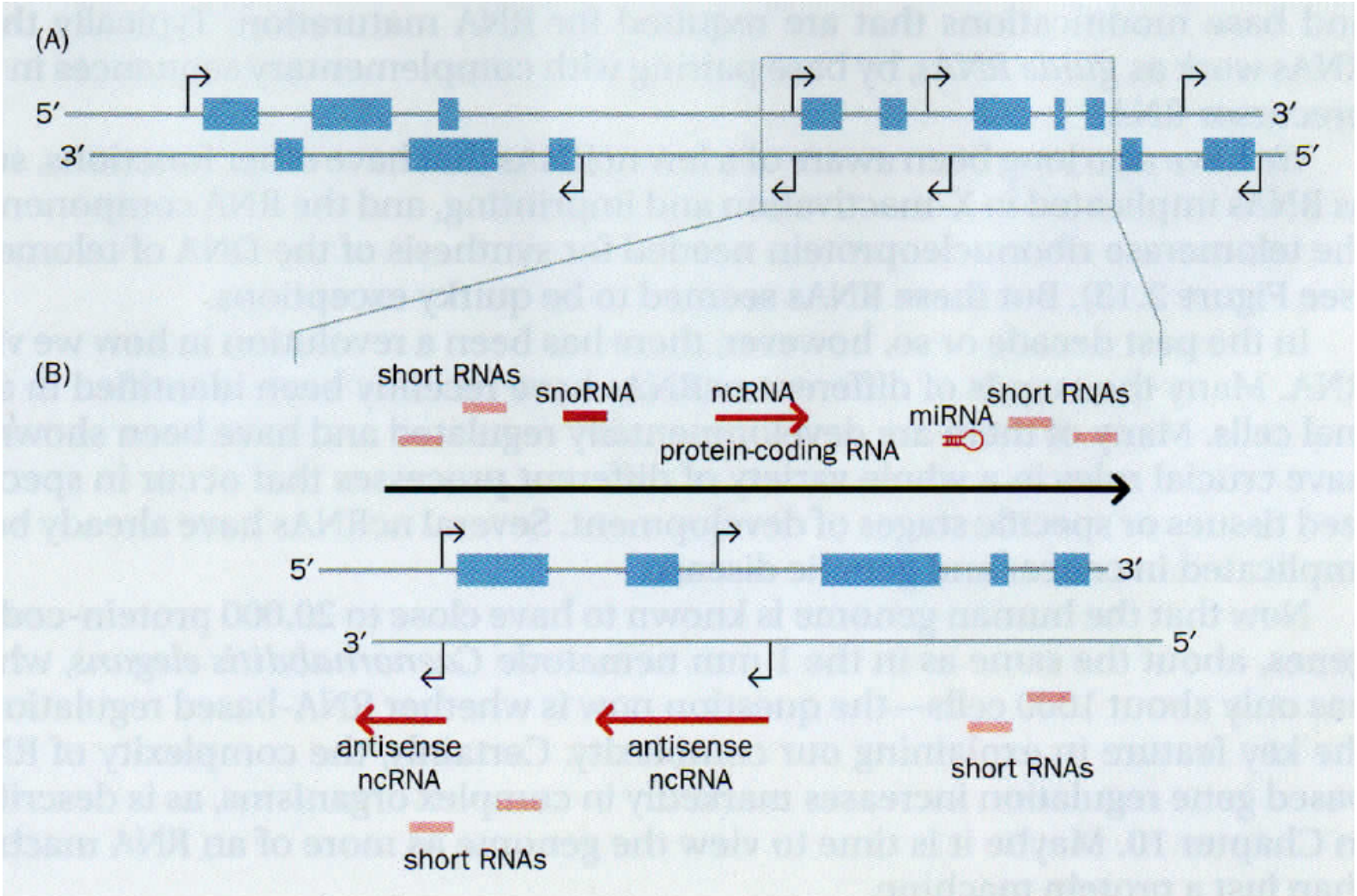


Vogelstein (2013) Science 74 Protein Data Bank rcsb.org

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

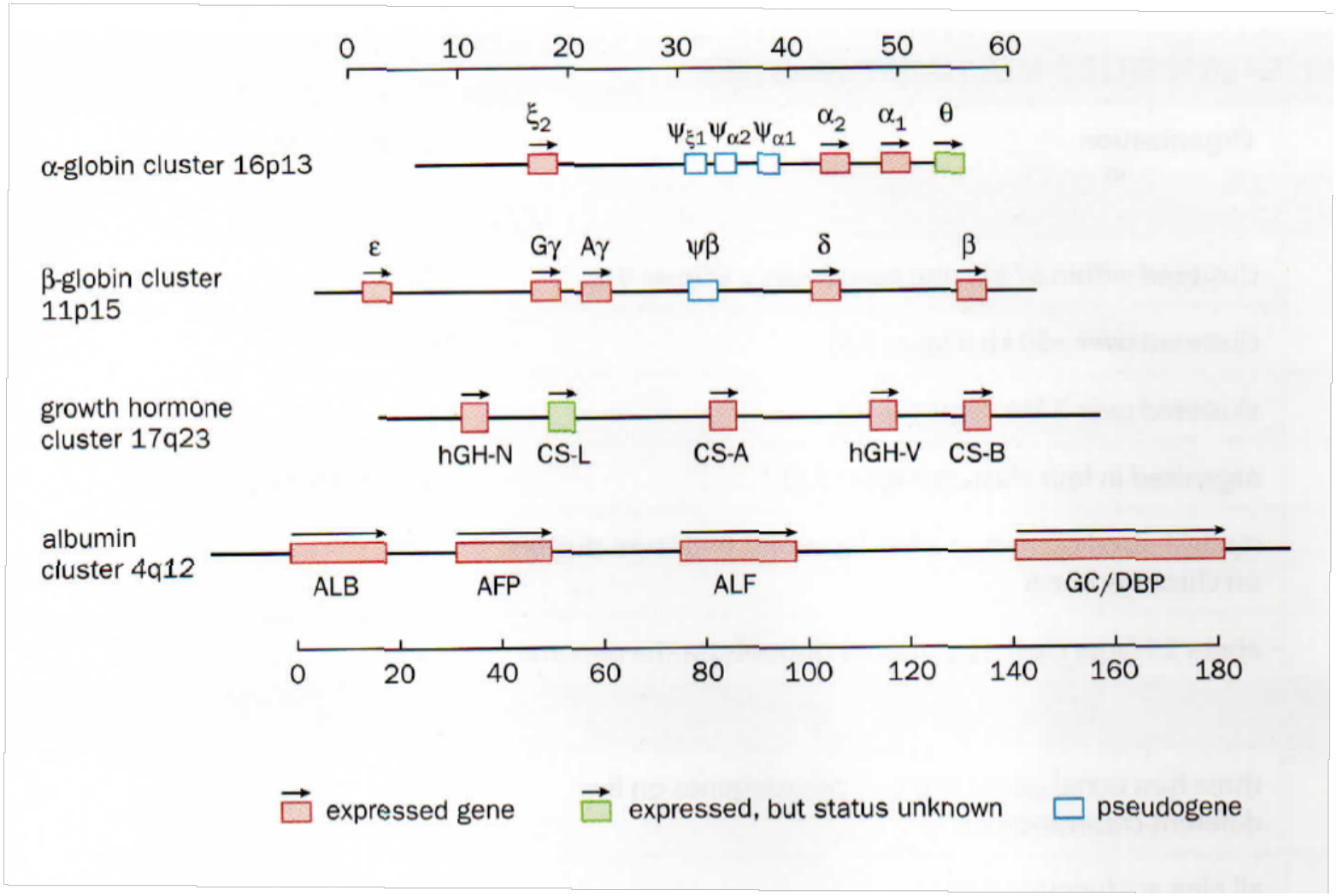


Размытие границ генов



Strachan, Read – *Human Molecular Genetics*

Мультигенные семейства



Strachan, Read – *Human Molecular Genetics*

Мультигенные семейства

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*

Классы белков человека

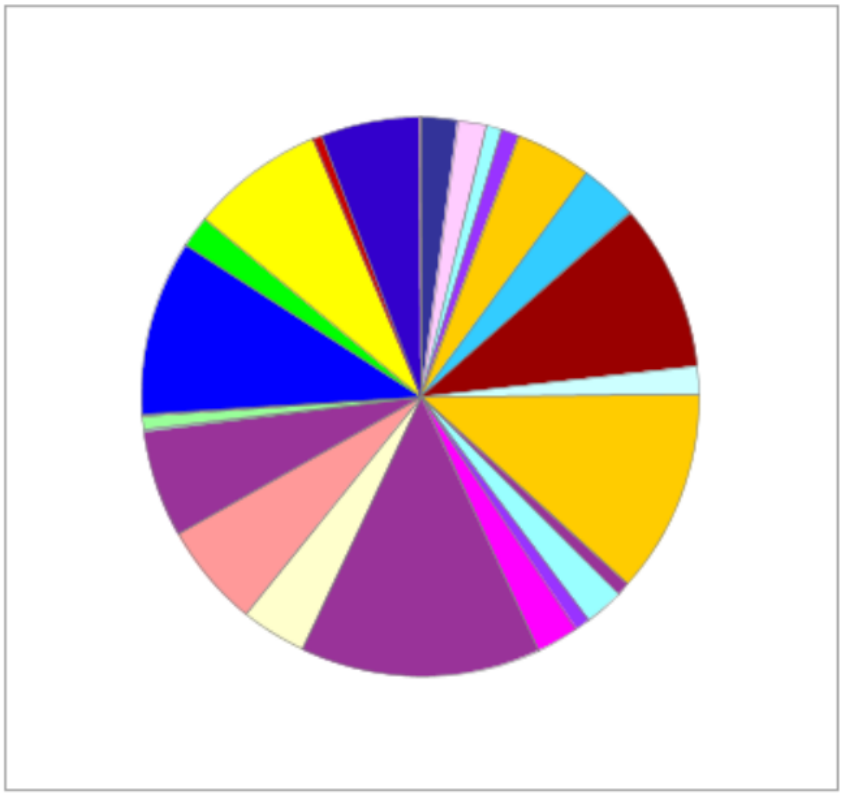


**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Protein Class hits

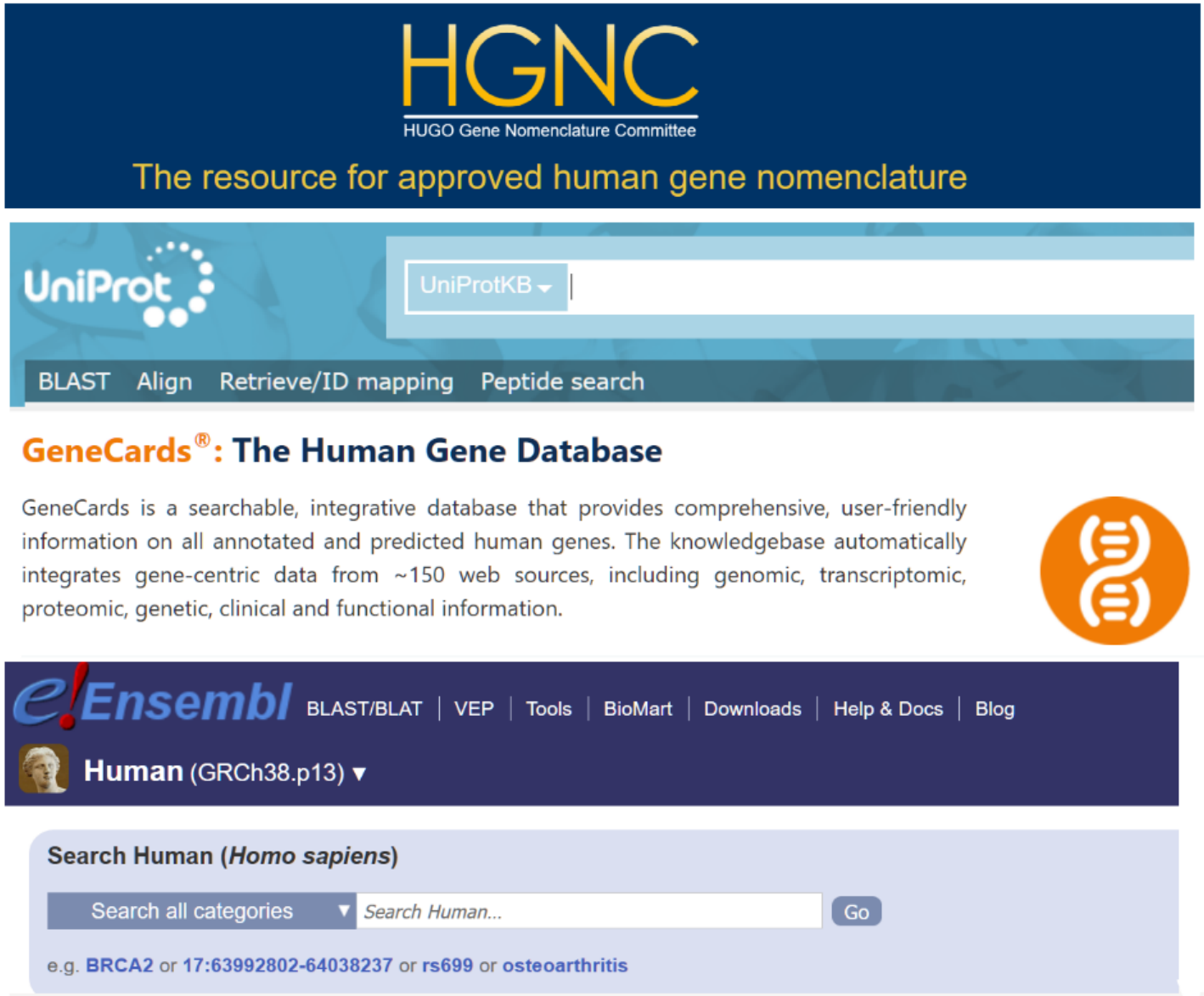
Click to get gene list for a category:

- 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)

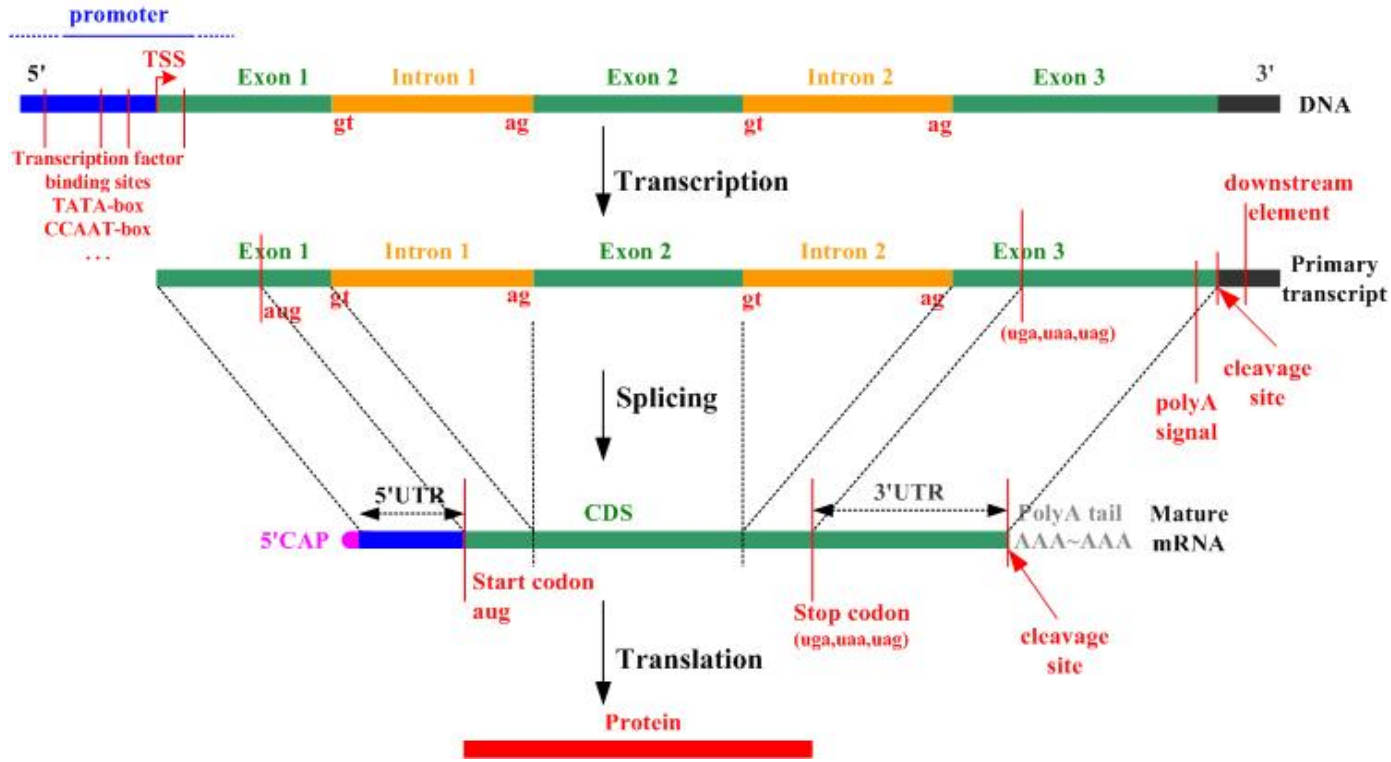
Классы белков человека

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

Номенклатура генов



Структура и процессирование генов человека



Carol Guze – *Biology 442 – Human Genetics*

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

Упражнение

Нарисуйте типичный ген человека

Структура и процессирование генов человека

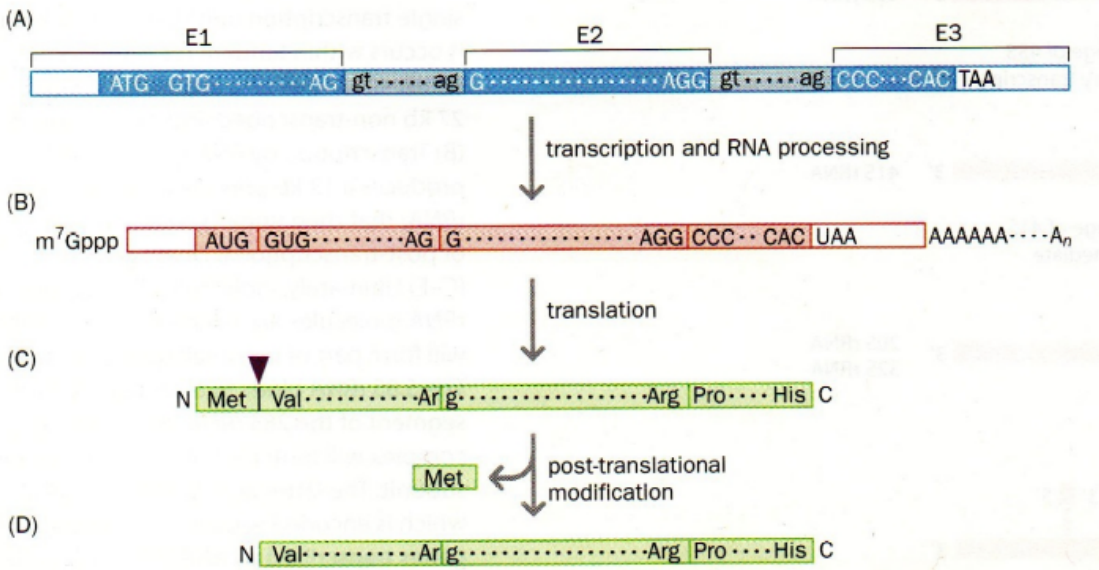


Figure 1.23 Transcription and translation of the human β-globin . (A) The β-globin gene comprises three exons (El-E3) and two introns. The 5'-end sequence of El and the 3' end sequence of E3 are noncoding sequences (unshaded sections). (B) These sequences are transcribed and so occur at the 5' and 3' 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*

Структура и процессирование генов человека

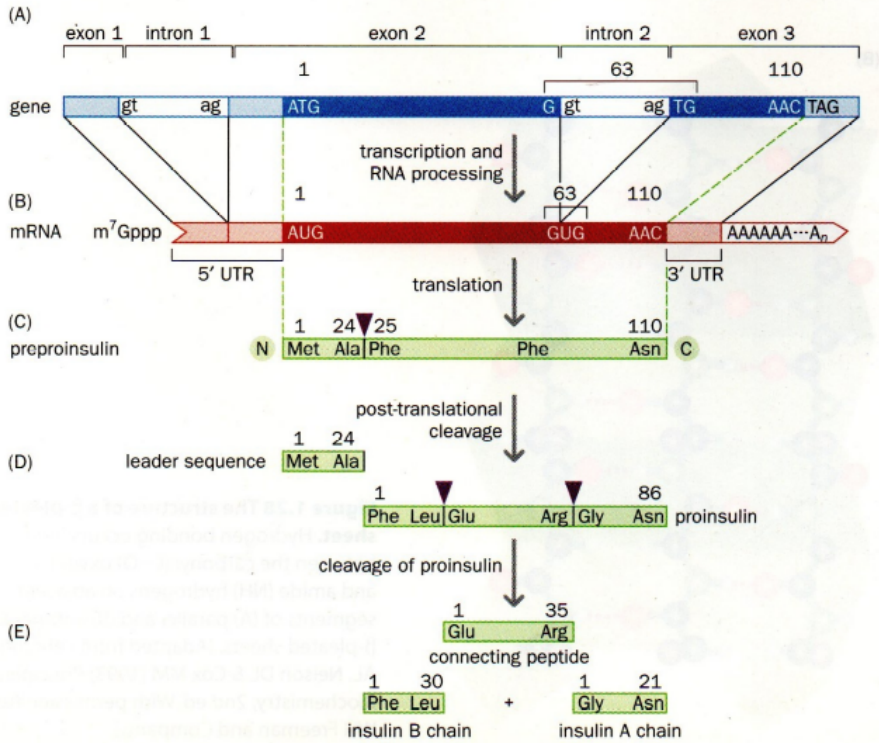


Figure 1.26 Insulin synthesis involves multiple post-translational cleavages of polypeptide precursors. (A) The human insulin gene comprises three exons and two introns. The coding sequence (the part that will be used to make polypeptide) is shown in deep blue. It is confined to the 3′ sequence of exon 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).

Strachan, Read – *Human Molecular Genetics*

Упражнение

Назовите примеры пост-трансляционных модификаций

Структура и процессирование генов человека

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%

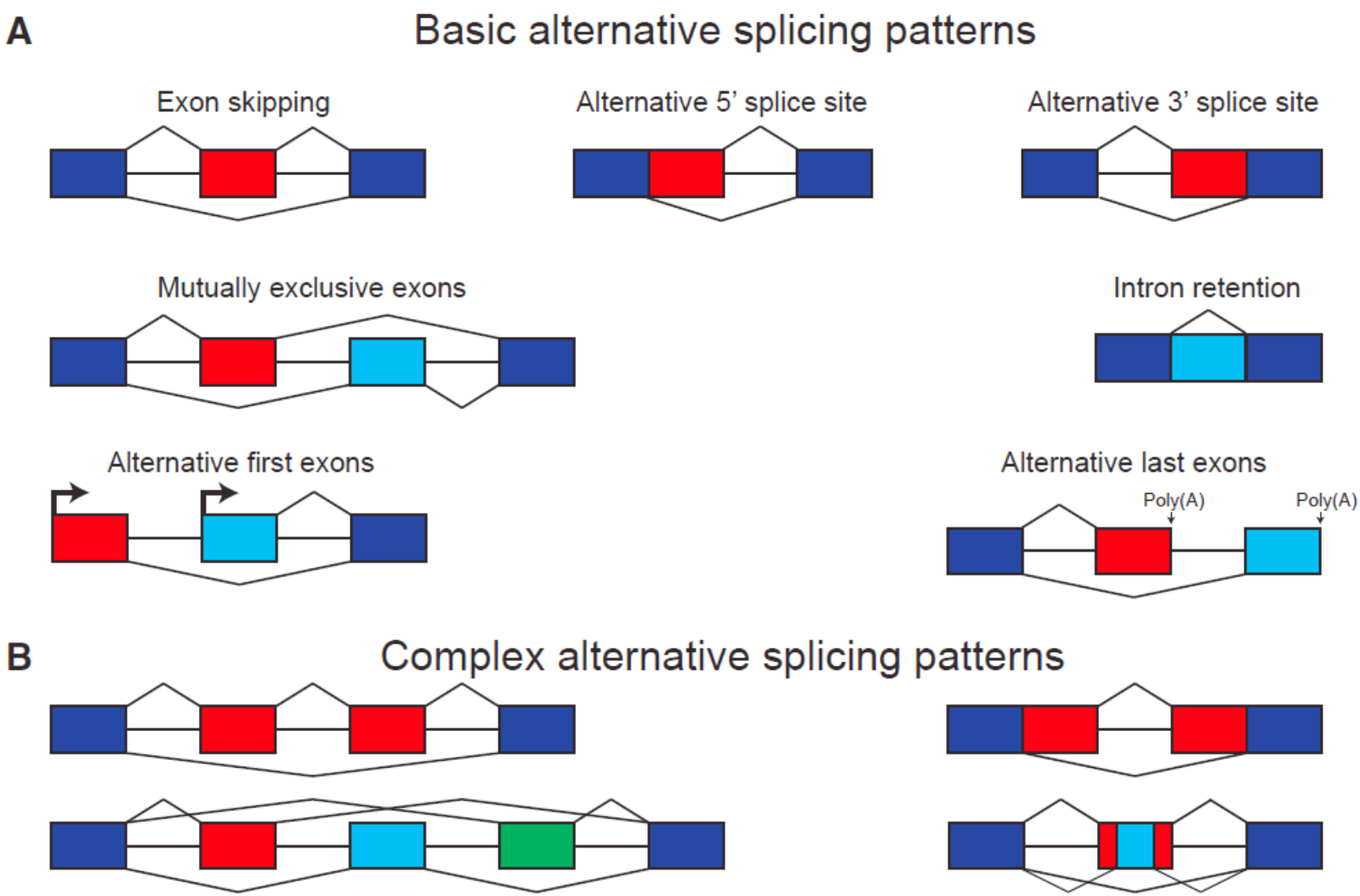
Alberts – *Essential Cell Biology*

Вопрос

Какой ген (экзон) самый большой?

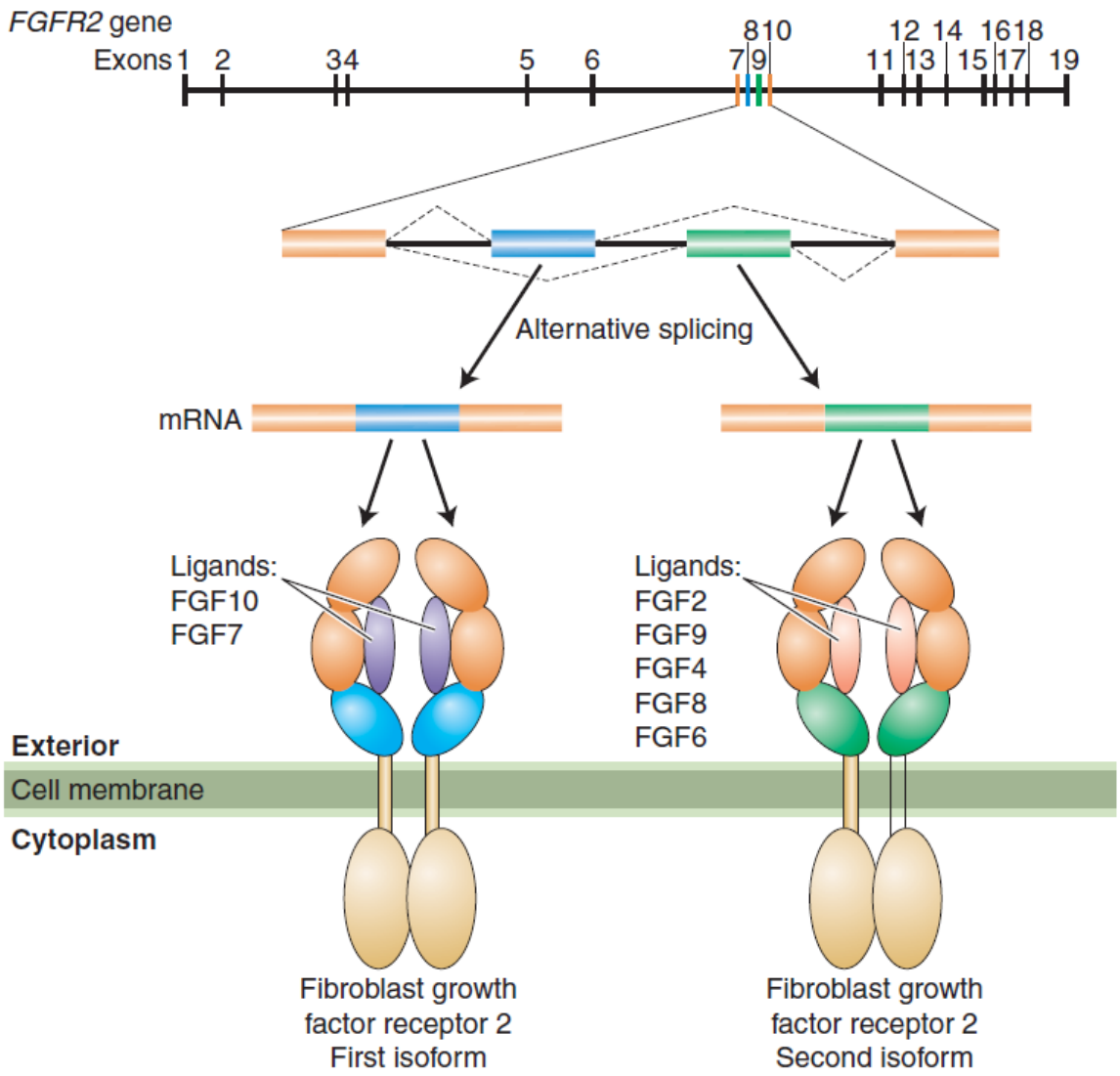
Альтернативный сплайсинг генов человека

Альтернативный сплайсинг генов человека



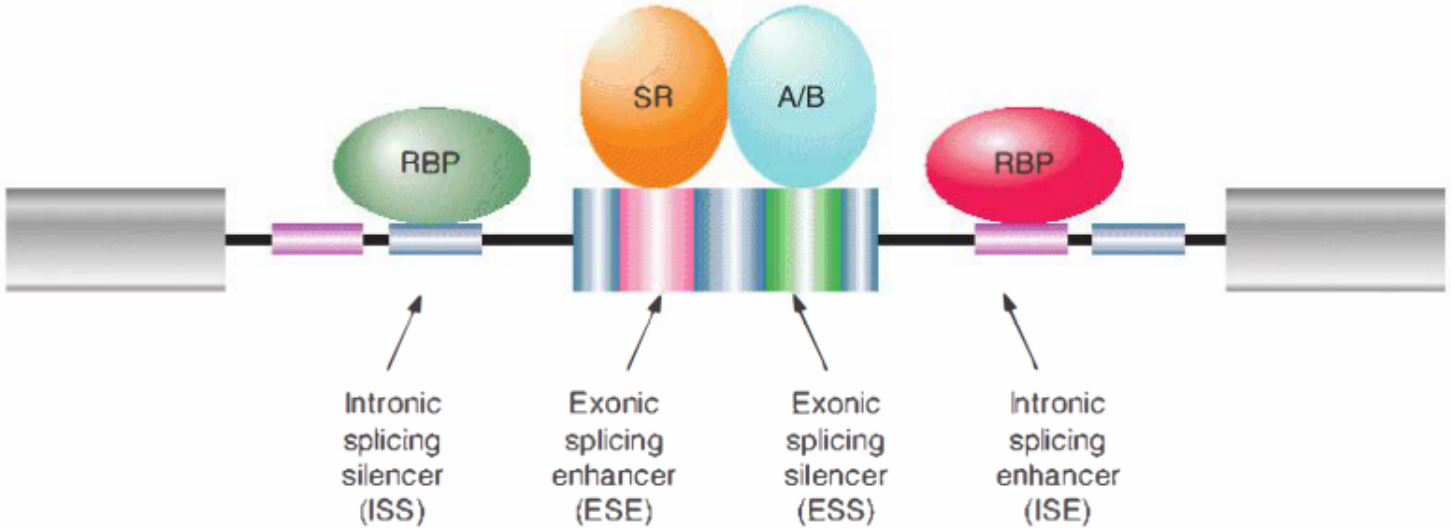
Park (2018) *Am J Hum Genet*

Альтернативный сплайсинг генов человека

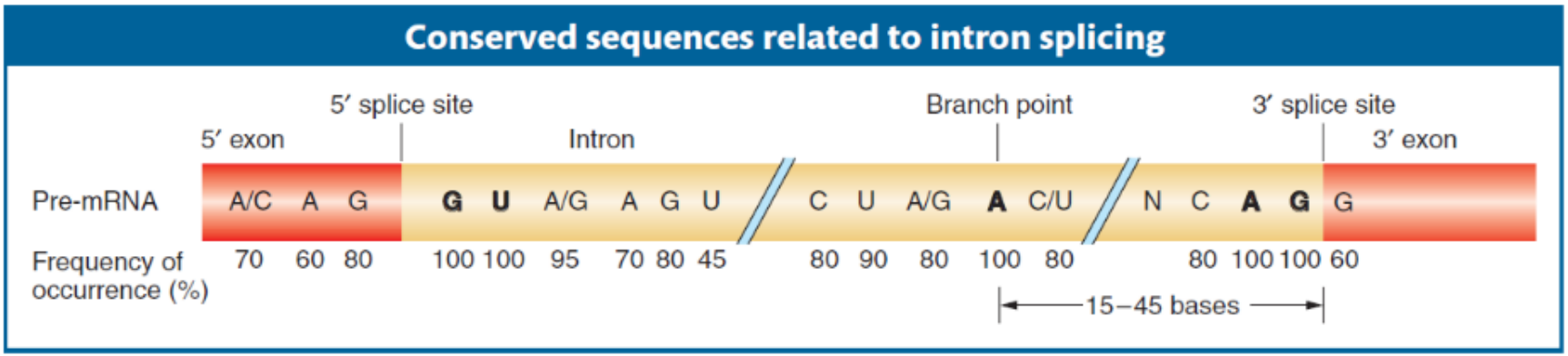


Griffiths – *Introduction to Genetic Analysis*

Альтернативный сплайсинг генов человека



Lewin – *Genes XI*



Griffiths – *Introduction to Genetic Analysis*

Альтернативный сплайсинг генов человека

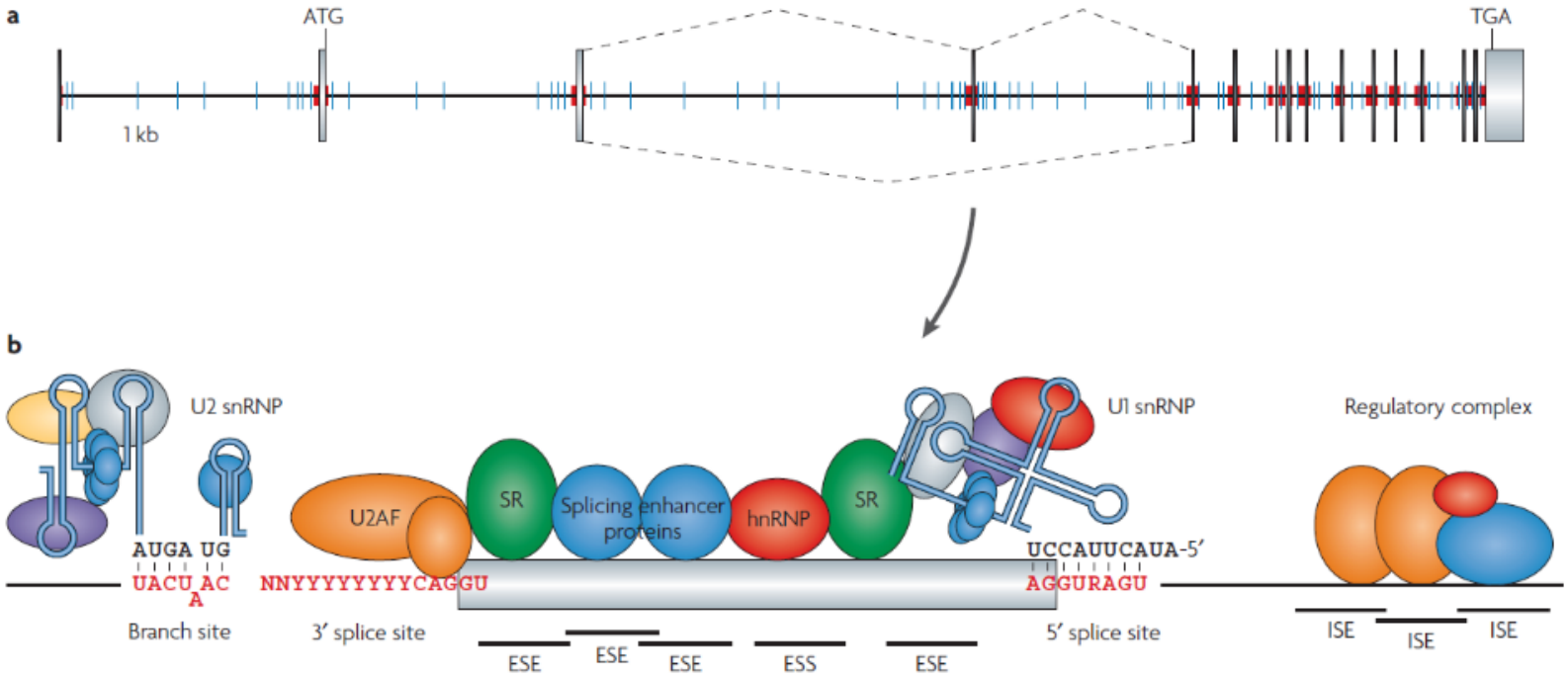


Figure 1 | **The splicing code. a** | A pre-mRNA as it might appear to the spliceosome. Red indicates consensus splice site sequences at the intron–exon boundaries. Blue indicates additional intronic cis-acting elements that make up the splicing code. **b** | cis-elements within and around an alternative 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*

Альтернативный сплайсинг генов человека

- ENSEMBL GRCh38 v.99, белок-кодирующие гены и транскрипты:
  - 1 транскрипт: 22% (нет альтернативного сплайсинга)
  - 2–5 транскриптов: 53%
  - >5%: 25%
  - Более 75 транскриптов: *ADGRG1, ANK2, KCNMA1, MAPK10, NDRG2, PAX6, TCF4*
- Самый длинный транскрипт назначается **каноническим** (≠самый биологически значимый)
- Вклад альтернативного сплайсинга в сложность протеома все еще обсуждается (количество транскриптов ≠ изоформы белков)
- Транскрипты, появившиеся благодаря альтернативному сплайсингу, содержащие преждевременные стоп-кодоны, подвержены NMD (**нонсенс-опосредованному распаду мРНК**)
- Микроэкзоны (3-30 нуклеотидов): неправильно регулируются в мозге у больных PAC (Irimia (2014) *Cell*)

Нарушения сплайсинга и заболевания

- **Цис-действующие мутации сплайсинга**: нарушение кода сплайсинга, **15-60% мутаций в заболеваниях человека** (Wang (2007) *Nat Rev Genet*). Примеры:
  - Синонимичные замены в гене *CFTR* $\Rightarrow$ муковисцидоз
  - Мутации в гене *MITF* $\Rightarrow$ синдром Ваарденбурга 2 типа, заболевание с доминантным типом наследование, проявляющееся в нарушениях пигментации и потере слуха
- **Транс-действующие мутации сплайсинга**: нарушения машинерии сплайсинга на уровне мРНК-белок. Пример:
  - Мутации *SMA* $\Rightarrow$ потеря производства snRNP (Small nuclear ribonucleoproteins) $\Rightarrow$ спинальная мышечная атрофия. Нусенерсен, лекарство, содержащее антисмысловой олигонуклеотид для корректировки сплайсинга при спинальной мышечной атрофии.

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.

Эпигенетика

Эпигенетика

Эпигенетика – наследуемые изменения фенотипа, которые не включают в себя изменения в последовательности ДНК.
Над генетикой: инструкции по использованию инструкций, или механизм контроля экспрессии генов

Эпигенетическая регуляция:

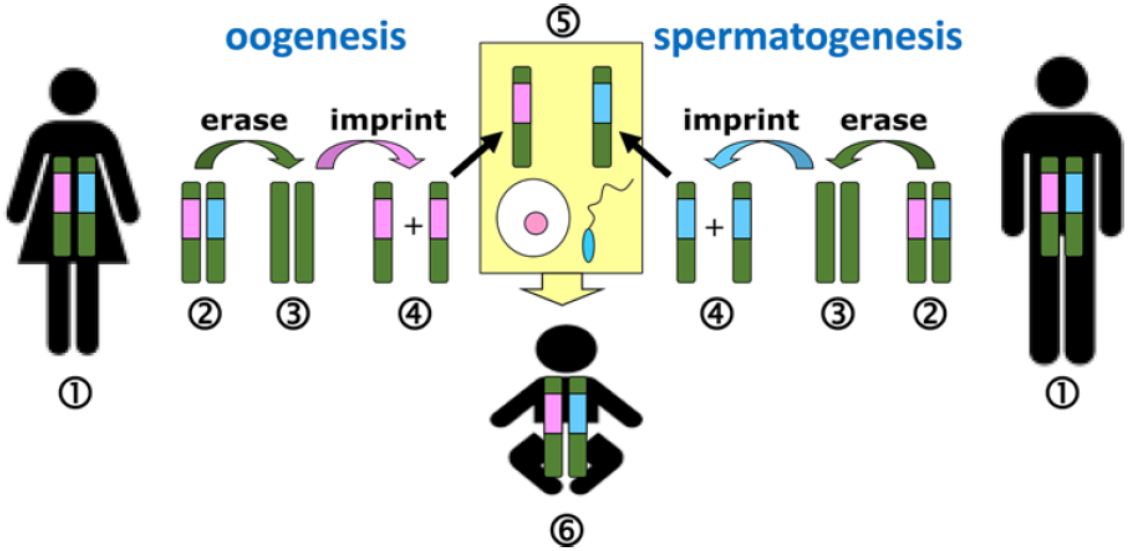
1. Метилирование ДНК в CpG динуклеотидах
2. Ковалентные модификации гистоновых белков
3. Некодирующие РНК

Замечания:

- Метилирование и гистоновые модификации обратимы
- Поддерживается при делении клеток и уничтожается при раннем эмбриогенезе
- Подвержено внутренним (развитие, старение) и внешним (химикаты, лекарства, диета, образ жизни) факторам

Хромосомный импринтинг

- **Хромосомный импринтинг**: $\sim$ 100 генов, одна копия которых остается неактивной благодаря механизмам эпигенетики, выбор — в зависимости от того, от кого из родителей эта копия пришла.
- Для некоторых ($\sim$ 70) активна только отцовская копия гена, в то время как активность материнской подавляется за счет эпигенетических механизмов, и наоборот ($\sim$ 30)
- Мутации в активных копиях генов приводят к **заболеваниям, связанным с импринтингом**

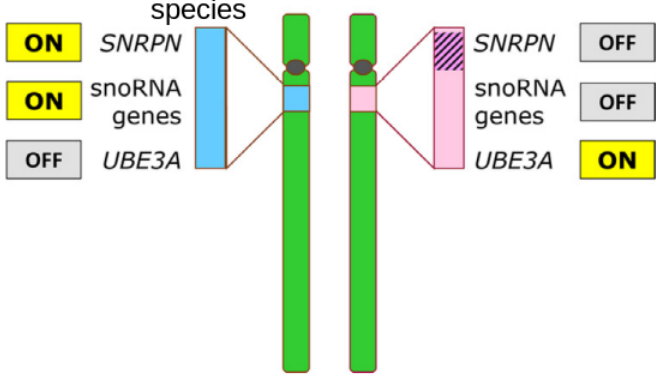


Jackson (2018) *Essays Biochem*

Хромосомный импринтинг

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



Jackson (2018) *Essays Biochem*

Упражнение

Проверьте свои любимые гены

Хромосомный импринтинг

	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*

Заболевания, связанные с импринтингом

- *IGF2* – гормон, стимулирующий рост во время развития эмбриона и плода (не путать с геном рецептора IGF2)
- В норме подавляется материнская копия
- **Эпимутации** (отсутствующие метки метилирования) могут привести к двум активным копиям

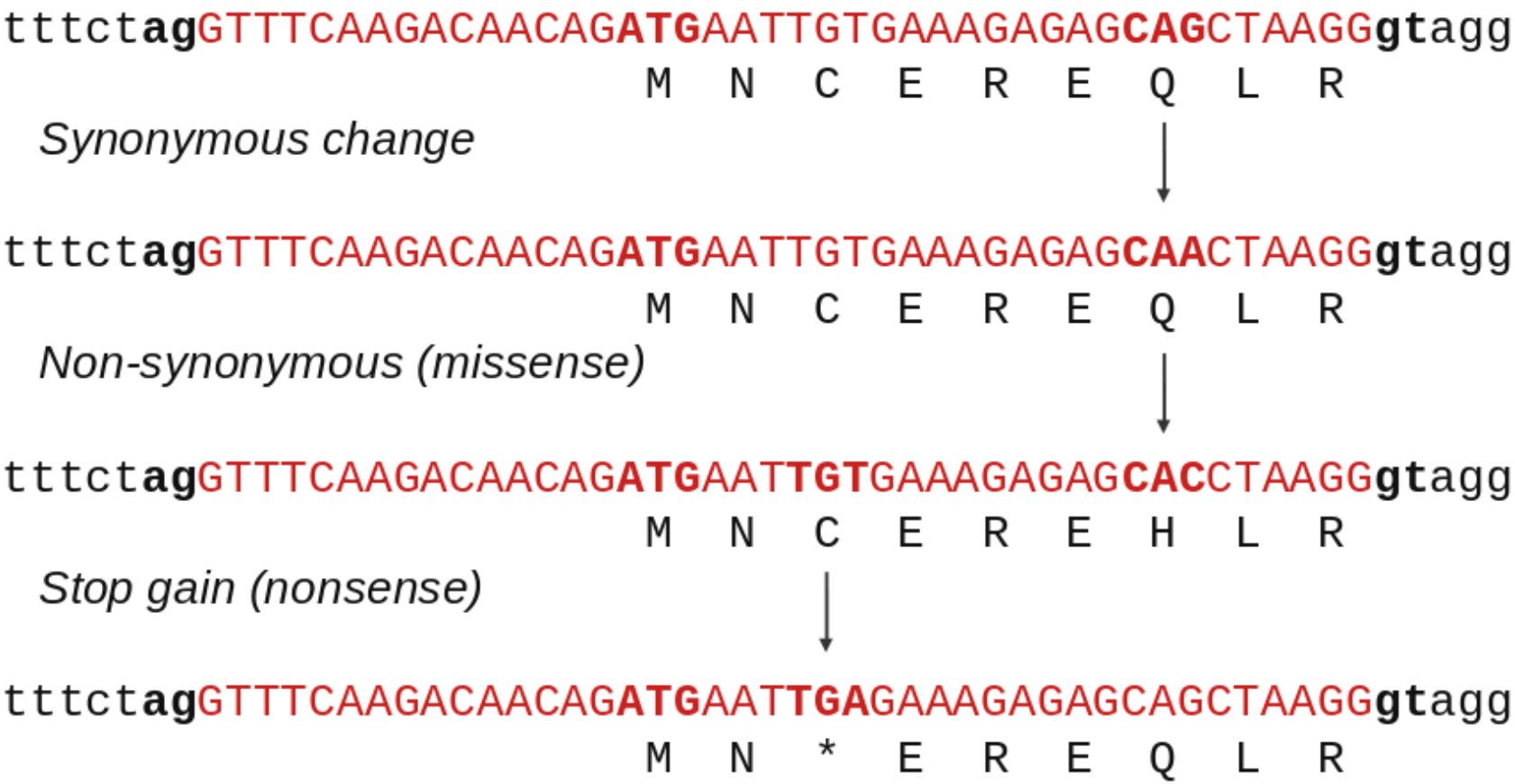
Активация материнской копии при формировании яйцеклетки или на ранних стадиях развития вызывает **синдром Беквита — Видемана**:

- Морфологические нарушения
- Повышенный риск развития рака в раннем возрасте
- Другие симптомы

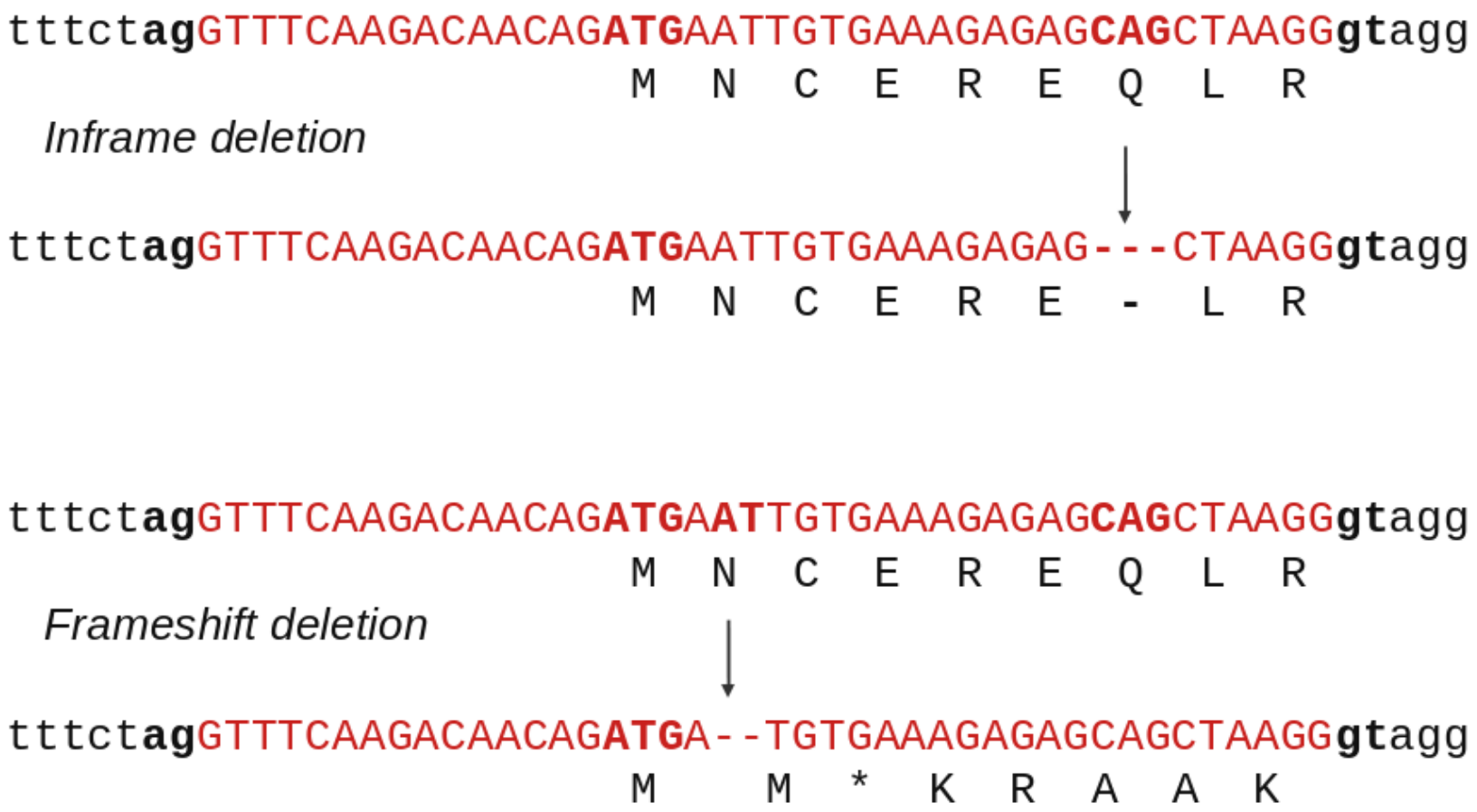
Частота встречаемости: $\sim$ 1/15,000 новорожденных. Однако для детей, зачатых методом ЭКО, частота встречаемости синдрома Беквита-Видемана может достигать 1/4,000.
`https://learn.genetics.utah.edu/content/epigenetics/imprinting`

Аннотация вариантов

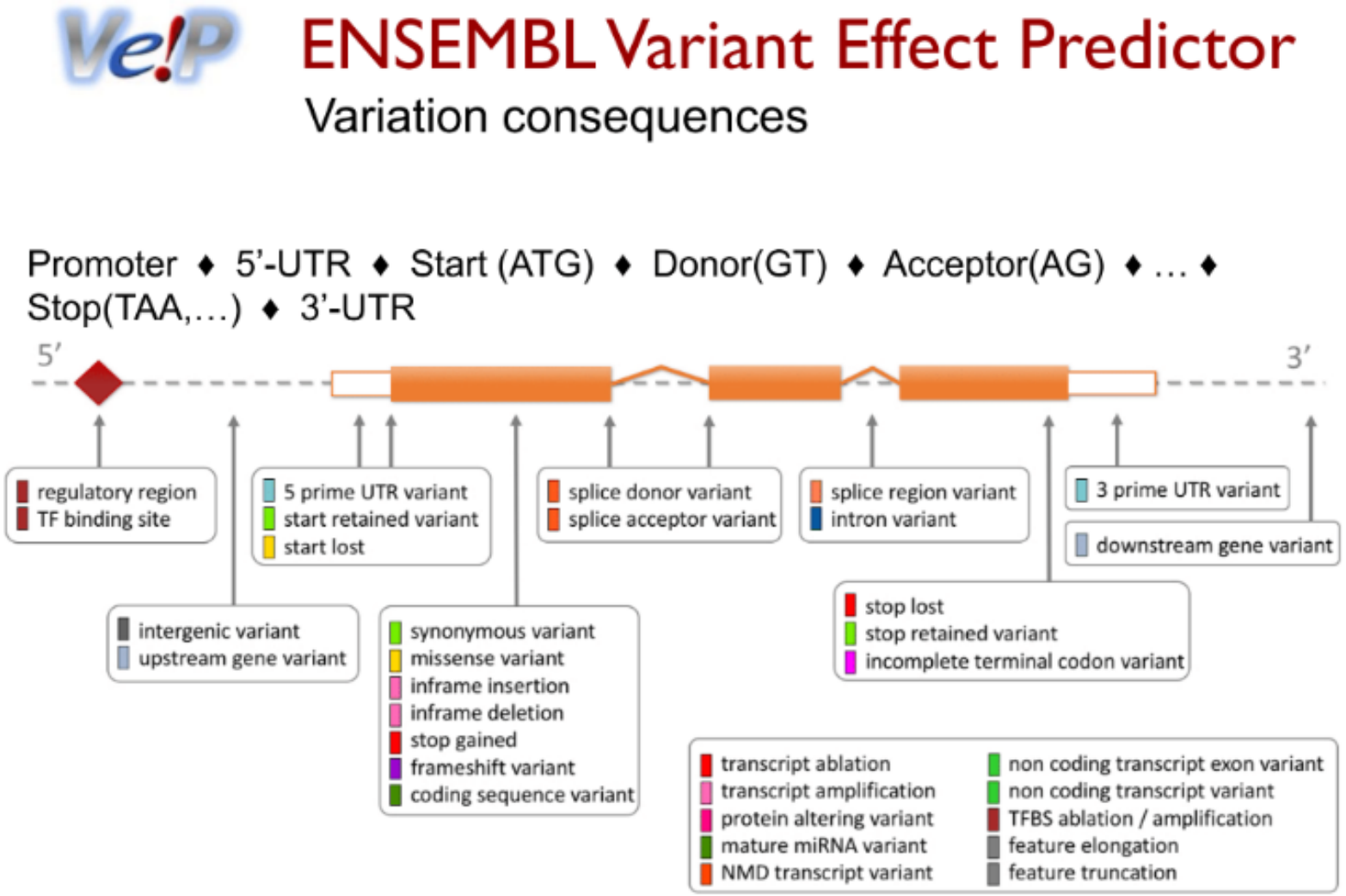
Примеры кодирующих замен в *RBFOX1*



Примеры кодирующих замен в *RBFOX1*



ENSEMBL Variant Effect Predictor



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

ENSEMBL Variant Effect Predictor

*	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

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

ENSEMBL Variant Effect Predictor

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	#Chr	Pos	Variant	Ref	Alt	Feature	Consequence	Existing_varia	IMPACT	STRAN	VARIANT_C	SYMBOL	BIOTYPE
2	chr1	237445489	chr1:237445489_G/A	G	A	ENST00000366574	missense_variant	rs794728721	MODERATE	1	SNV	RYR2	protein_
3	chr1	237783995	chr1:237783995_G/A	G	A	ENST00000366574	missense_variant	rs753850982	MODERATE	1	SNV	RYR2	protein_
4	chr3	14141665	chr3:14141665_C/T	C	T	ENST00000306077	missense_variant	rs63750743	MODERATE	1	SNV	TMEM43	protein_
5	chr6	7579989	chr6:7579989_C/T	C	T	ENST00000379802	stop_gained	rs121912997	HIGH	1	SNV	DSP	protein_
6	chr11	2445385	chr11:2445385_C/G	C	G	ENST00000155840	missense_variant	rs1337409061	MODERATE	1	SNV	KCNQ1	protein_
7	chr11	2585275	chr11:2585275_C/T	C	T	ENST00000155840	missense_variant	rs199473411	MODERATE	1	SNV	KCNQ1	protein_
8	chr11	2778015	chr11:2778015_G/T	G	T	ENST00000155840	missense_variant	rs199472814	MODERATE	1	SNV	KCNQ1	protein_
9	chr11	47351507	chr11:47351507_T/C	T	C	ENST00000545968	splice_acceptor_vari	rs376395543	HIGH	-1	SNV	MYBPC3	protein_
10	chr13	32340300	chr13:32340300_GT/G	GT	G	ENST00000380152	frameshift_variant	rs80359550	HIGH	1	deletion	BRCA2	protein_
11	chr14	23415652	chr14:23415652_G/A	G	A	ENST00000355349	missense_variant	rs121913650	MODERATE	-1	SNV	MYH7	protein_
12	chr17	43057062	chr17:43057062_T/TG	T	TG	ENST00000471181	frameshift_variant	rs80357906	HIGH	-1	insertion	BRCA1	protein_
13	chr1	26061184	chr1:26061184_ACT/A	ACT	A	ENST00000374272	frameshift_variant	rs540072010	HIGH	-1	deletion	TRIM63	protein_
14	chr1	45508806	chr1:45508806_G/C	G	C	ENST00000401061	missense_variant	rs140522266	MODERATE	1	SNV	MMACHC	protein_
15	chr1	99902746	chr1:99902746_C/T	C	T	ENST00000294724	stop_gained	rs771853367	HIGH	1	SNV	AGL	protein_
16	chr1	99916398	chr1:99916398_A/G	A	G	ENST00000294724	intron_variant	rs369973784	MODIFIER	1	SNV	AGL	protein_
17	chr1	114716123	chr1:114716123_C/A	C	A	ENST00000369535	missense_variant	rs121434596	MODERATE	-1	SNV	NRAS	protein_
18	chr2	43872094	chr2:43872094_G/A	G	A	ENST00000272286	stop_gained	rs137852987	HIGH	1	SNV	ABCG8	protein_
19	chr2	43875377	chr2:43875377_G/A	G	A	ENST00000272286	missense_variant	rs137852988	MODERATE	1	SNV	ABCG8	protein_
20	chr2	73385920	chr2:73385920_G/GAGG	G	GAGG	ENST00000613296	stop_gained,inframe	?	HIGH	1	insertion	ALMS1	protein_
21	chr2	73424857	chr2:73424857_TGGACA	TGGACA	T	ENST00000613296	frameshift_variant	rs761292021	HIGH	1	deletion	ALMS1	protein_
22	chr2	73450679	chr2:73450679_T/TA	T	TA	ENST00000613296	frameshift_variant	rs797045228	HIGH	1	insertion	ALMS1	protein_
23	chr2	73453759	chr2:73453759_T/TA	T	TA	ENST00000613296	frameshift_variant	rs1553404426	HIGH	1	insertion	ALMS1	protein_
24	chr2	73573187	chr2:73573187_TAGAG/T	TAGAG	T	ENST00000613296	frameshift_variant	rs747272625	HIGH	1	deletion	ALMS1	protein_
25	chr2	178542263	chr2:178542263_C/G	C	G	ENST00000589042	splice_donor_variant	rs727505319	HIGH	-1	SNV	TTN	protein_
26	chr2	178704946	chr2:178704946_CCTCT	CCTCT	C	ENST00000589042	frameshift_variant	rs777924443	HIGH	-1	deletion	TTN	protein_
27	chr3	33051989	chr3:33051989_A/C	A	C	ENST00000307363	missense_variant	rs376663785	MODERATE	-1	SNV	GLB1	protein_

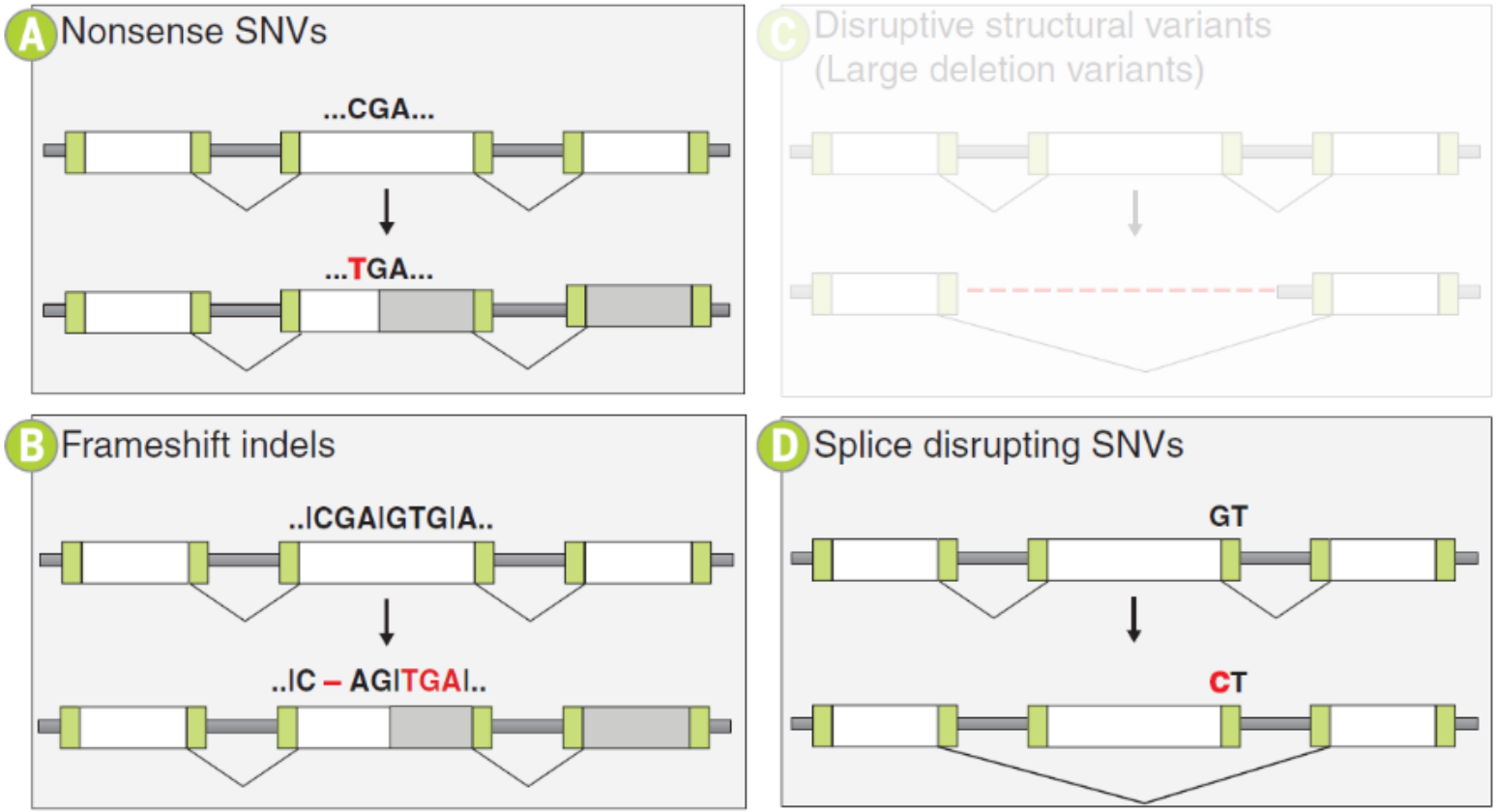
Выдача VEP

Укорачивающие белок варианты

Укорачивающие белок варианты

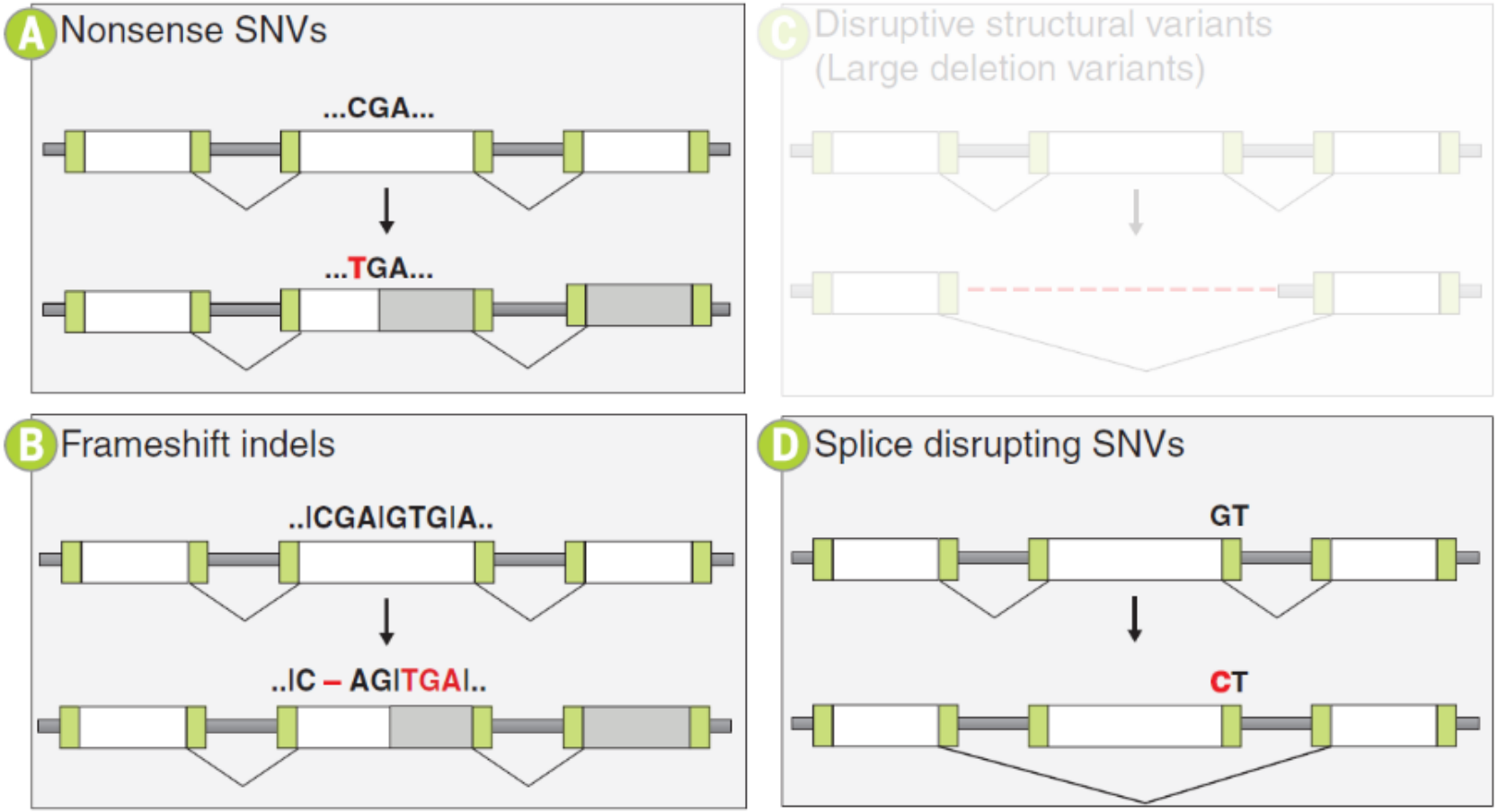
Укорачивающие варианты (protein-truncating variants, PTV) // Rivas (2015) *Science*

- Появление или уничтожение стоп-кодона
- Замены в канонических сайтах сплайсинга
- Инделы со сдвигом рамки считывания



Укорачивающие белок варианты

Укорачивающие белок варианты (protein-truncating variants, PTV)



Rivas (2015) *Science*

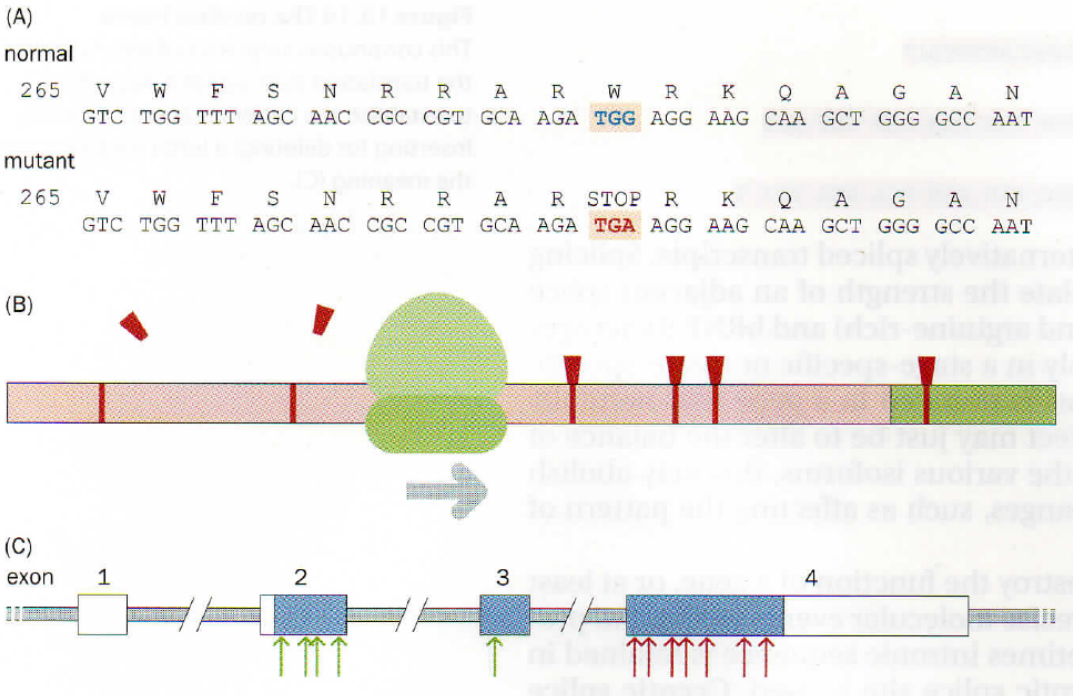
Варианты PTV и LoF

Не все PTV являются мутациями с потерей функции (loss-of-function, LoF)
`LOFTEE` (K.Karczewski et al) `https://github.com/konradjk/loftee`

PTV, не предсказываемые как pLoF (putative LoF), примеры:

- Появление стоп-кодона или сдвига рамки считывания ближе, чем 50 п.н. от конца транскрипта (NMD)
- Варианты в экзоне с неканоническим сайтом сплайсинга (не GT, AG)
- Варианты в сайтах сплайсинга, «спасаемые» лежащими рядом другими сайтами в той же рамке считывания, например, NAGNAG
- Варианты в сайтах сплайсинга в коротком интроне
- ...или в интроне с неканоническим сайтом сплайсинга

PTV и NMD



Strachan, Read – *Human Molecular Genetics*

(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.

PTV и NMD

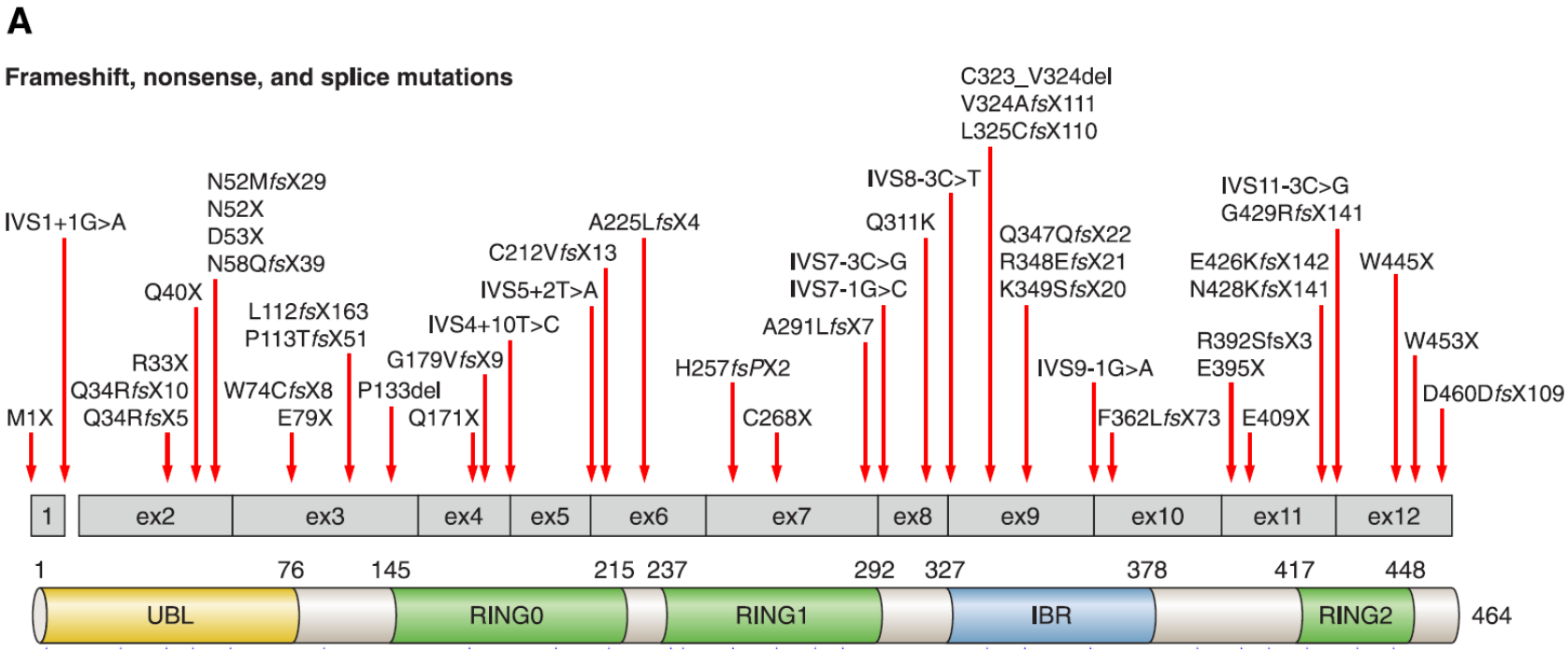
В идеале: **PTV** → **NMD** → **Концентрация транскриптов** → **Концентрация белка** → **Функционирование клетки**

Однако изменения количества мРНК не всегда коррелируют с уровнями экспрессии белков.

Battle, A., Khan, Z., *et al.* (2015). Impact of Regulatory Variation from RNA to Protein. *Science* 347, 664–667.

Narasimhan, V.M., Xue, Y., Tyler-Smith, C. (2016) Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends in Molecular Medicine*. 22, 341-351.

Примеры эффекта PTV



Corti (2011) *Physiol Rev*

Мутации в гене паркин-RBR-убиквитинлигазы-E3 (*PRKN*) — самая частая известная причина раннего (40–50 лет) развития **болезни Паркинсона**. Болезнь Паркинсона – второе по частоте встречаемости ($\sim 0,3\%$ в развитых странах) нейродегенеративное заболевание после болезни Альцгеймера.

Примеры эффекта PTV

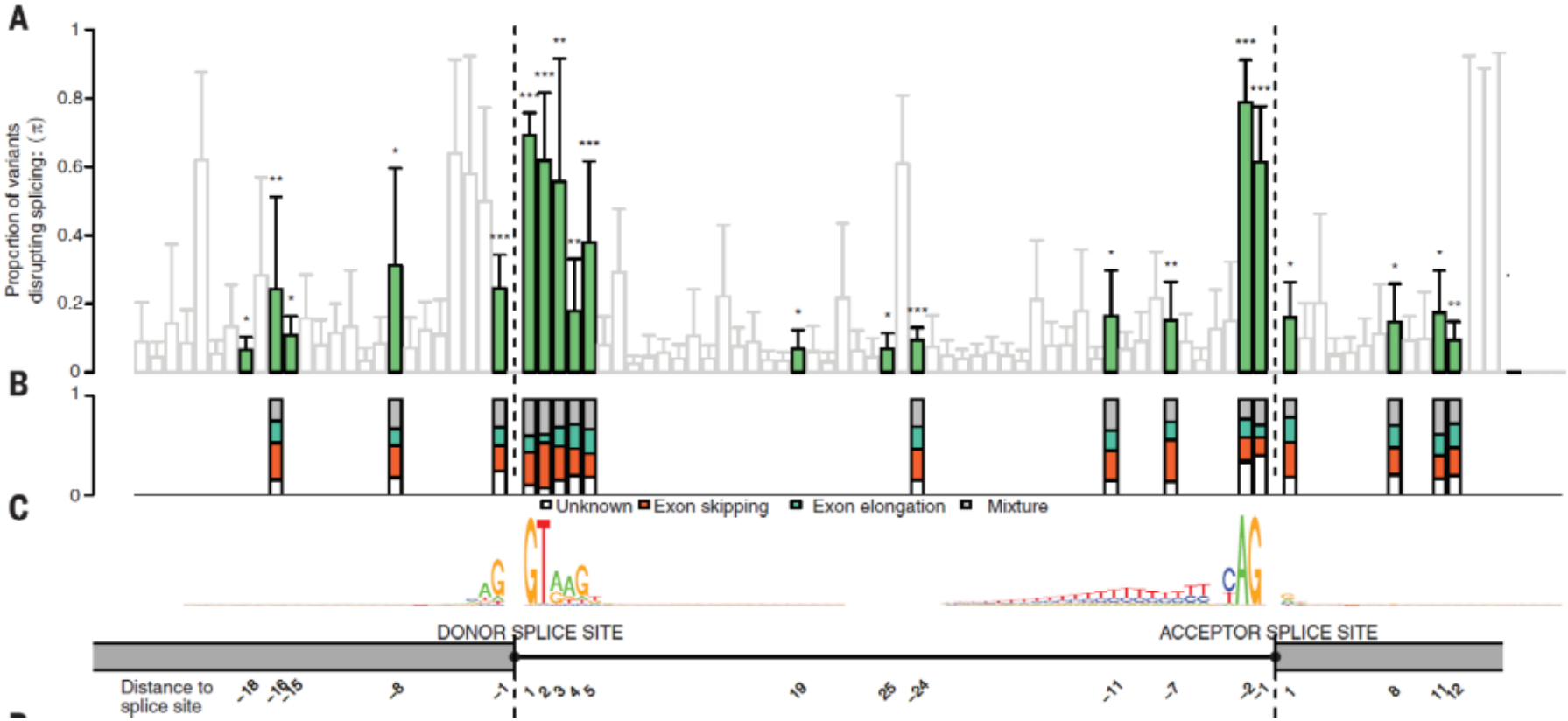


Fig. 3. Splicing disruption. (A) Proportion of variants disrupting splicing at each distance +/- 25 bp from donor and acceptor site (B) Classification of splice disruption events: exon skipping, exon elongation and mixture (C) Diagram of donor and acceptor splice junctions and sequence logo of represented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

Rivas (2015) *Science*

Примеры эффекта PTV

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

- Было показано, что нокаут гена иммунной системы *IRF7* **увеличивает восприимчивость к вирусу гриппа**, приводя к угрожающей жизни инфекции у в остальном здорового ребенка (Ciancanelli 2015 *Science*)
- Случаи, когда произошедшие **естественным образом PTV оказались благоприятными для здоровья**. Эти открытия способствовали развитию разработки лекарств:
  - Понижение уровней ЛПНП: *PCSK9*
  - Уменьшение восприимчивости к вирусу ВИЧ: *CCR5*
  - Увеличение выносливости: *ACTN3*
  - Увеличение сопротивления сепсису: *CASP12*
  - Понижение уровня триглицеридов: *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 $\times$ 135 фенотипов; найдено 27 ассоциаций между клиническими фенотипами и PTV в генах вне MHC

Примеры эффекта PTV

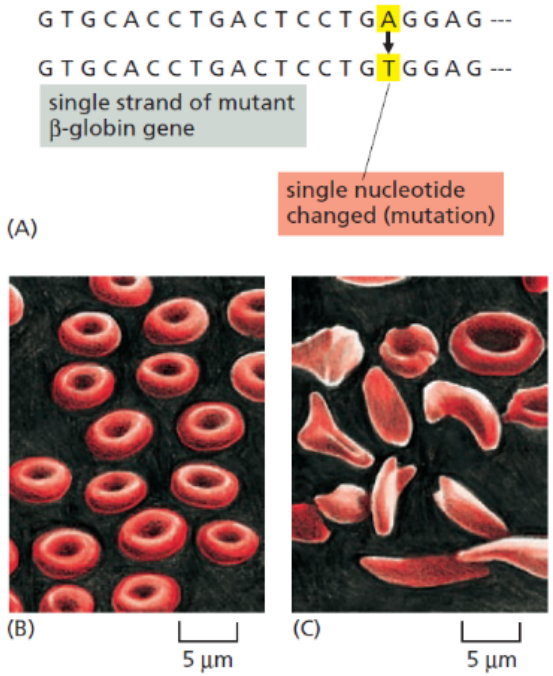
1. The stop-gain variant in *GNAS* (MIM:139320) is present in the highly variable first exon of the gene and is likely to result in nonsense-mediated RNA decay; in contrast, pathogenic *GNAS* variants that cause Albright hereditary osteodystrophy (MIM:103580) are located in later, highly constrained 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.
4. 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*

Несинонимичные варианты (миссенсы) и заболевания

Классический пример: серповидноклеточная анемия

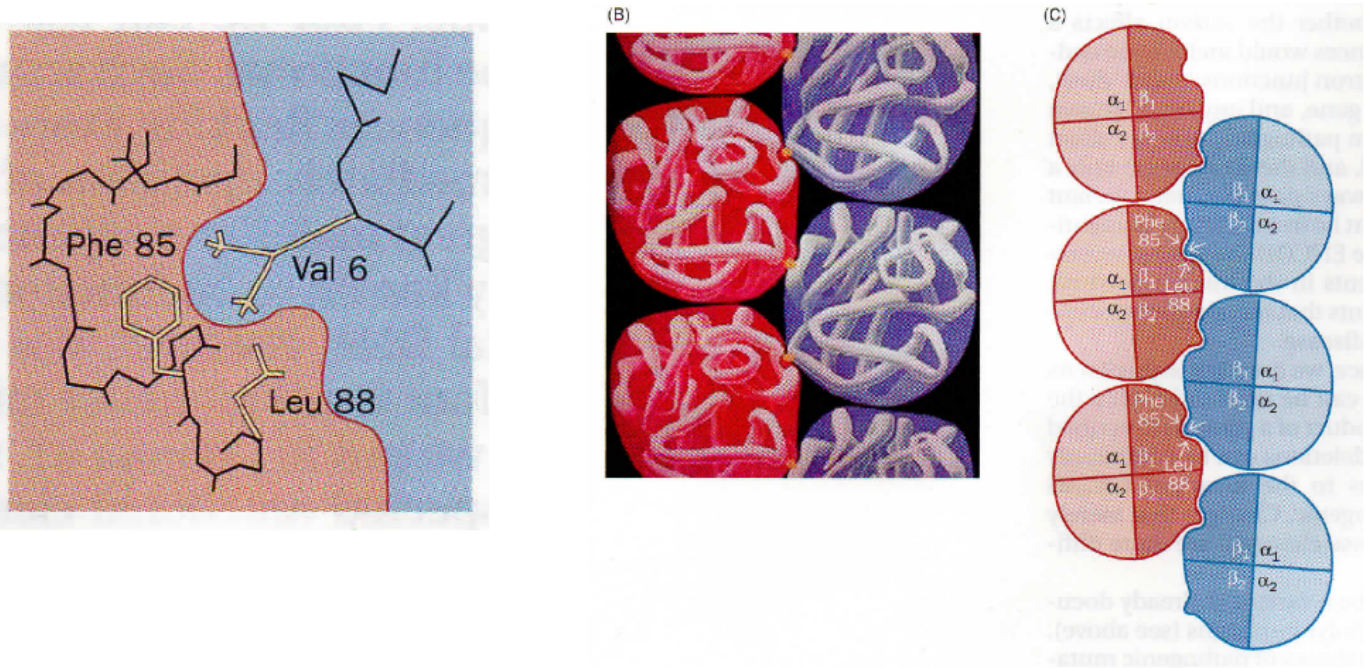
Figure 6–19 **A single nucleotide change causes the disease sickle-cell anemia.** (A) β-globin is one of the two types of subunit that form hemoglobin (see Figure 4–20). A single nucleotide change (mutation) in the β-globin gene produces a β-globin subunit that differs from normal β-globin only by a change from glutamic acid to valine at the sixth amino acid position. (Only a small portion of the gene is shown here; the β-globin subunit contains a total of 146 amino acids.) Humans carry two copies of each gene (one inherited from each parent); a sickle-cell mutation in one of the two β-globin genes generally causes no harm to the individual, as it is compensated for by the normal gene. However, an individual who inherits two copies of the mutant β-globin gene displays the symptoms of sickle-cell anemia. Normal red blood cells are shown in (B), and those from an individual suffering from sickle-cell anemia in (C). Although sickle-cell anemia can be a life-threatening disease, the mutation responsible can also be beneficial. People with the disease, or those who carry one normal gene and one sickle-cell gene, are more resistant to malaria than unaffected individuals, because the parasite that causes malaria grows poorly in red blood cells that contain the sickle-cell form of hemoglobin.



Alberts - *Essential Cell Biology*

HBB.Glu7Val Sickle cell anemia [MIM:603903]: Characterized by abnormally shaped red cells resulting in **chronic anemia and periodic episodes of pain, serious infections and damage to vital organs**. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can led to microvascular occlusion thus cutting off the blood supply to nearby tissues // `www.genecards.org`

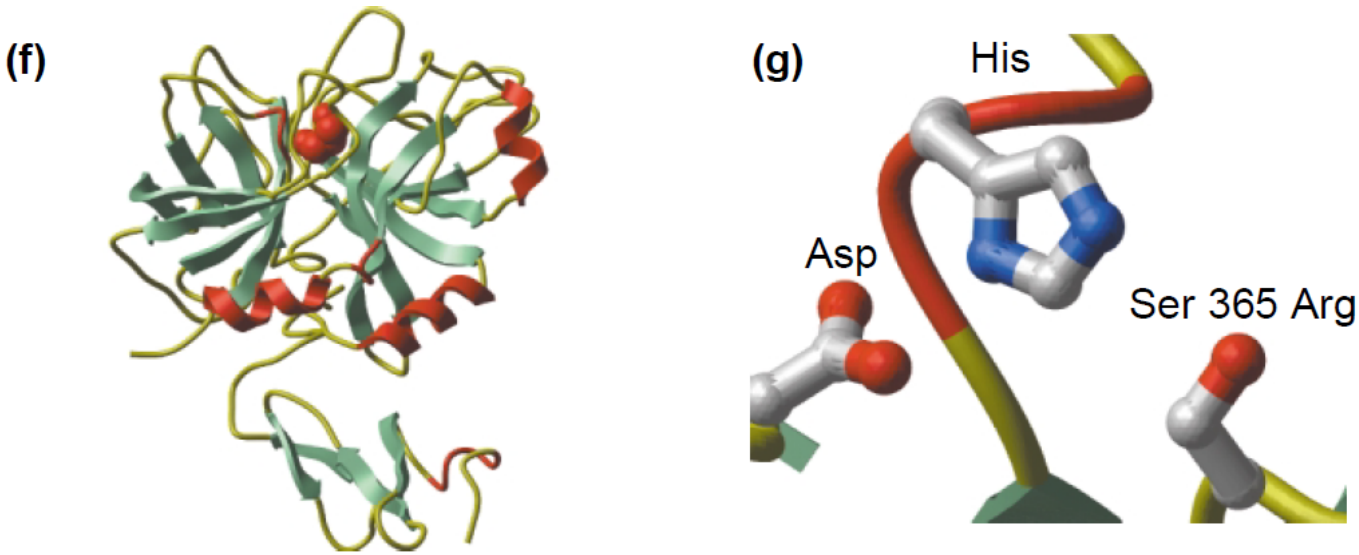
Классический пример: серповидноклеточная анемия



Strachan, Read – *Human Molecular Genetics*

Мутация A>T в гене β-глобина (*HBB*) вызывает замену аминокислоты в белке β-глобина. Глутаминовая кислота (гидрофильная заряженная) заменяется валином (гидрофобной неполярной аминокислотой). Эта замена на поверхности белка глобина вызывает адгезивные взаимодействия молекул гемоглобина.

Другие примеры



Vogelstein (2013) Science 74 Protein Data Bank rcsb.org

Factor IX F9 is a serine protease with Ser-His-Asp catalytic triade that participates in the intrinsic pathway of blood coagulation by converting factor X to its active form Xa. Disease mutations in F9 are associated with the X-linked recessive bleeding disorder haemophilia B (OMIM:306900).

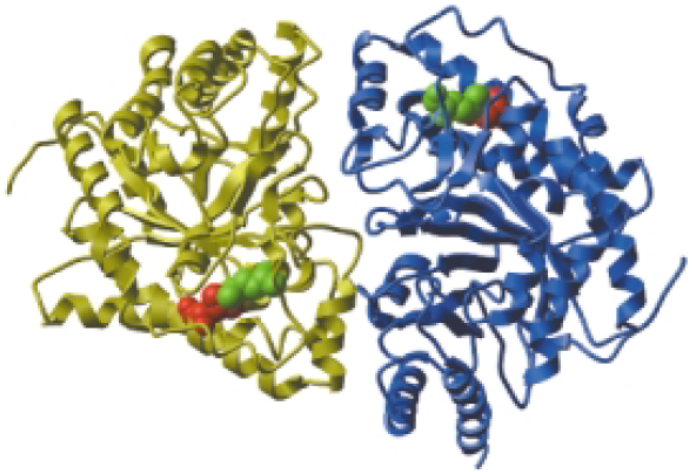
Disruption of catalytic residues. Mutations of the catalytic serine residue to an arginine results in the loss of enzyme activity and a severe haemophilia phenotype.

Steward (2003) Trends Genet

Другие примеры

Introduction of buried charged residues:
Met165Arg ⇒ arginine sidechain cannot be accommodated in a hydrophobic pocket ⇒ no soluble protein.

Size changes in the hydrophobic core:
Leu195Phe ⇒ rearrangement of surrounding side-chains ⇒ 30% of the wild-type activity.



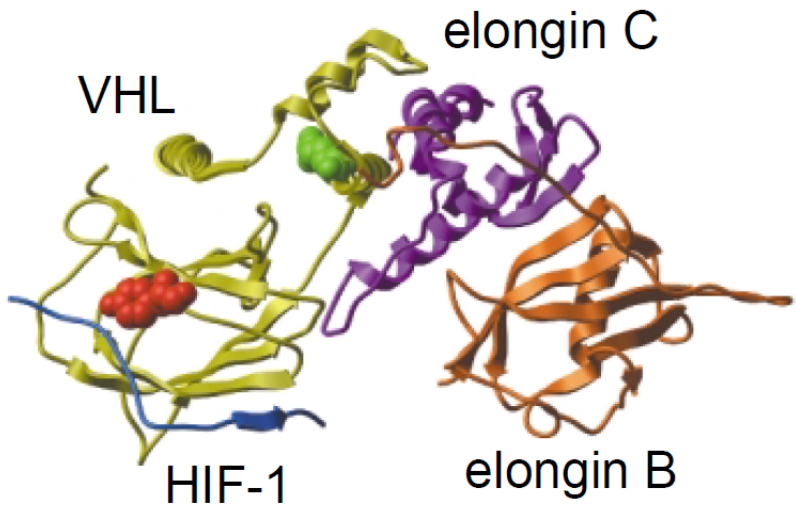
Vogelstein (2013) *Science* 74 Protein Data Bank `rcsb.org`

Mutations in the uroporphyrinogen decarboxylase UROD are associated with Porphyria cutanea tarda (OMIM:176100), accumulation of uroporphyrins in the liver and plasma, leading to skin fragility and photosensitive dermatitis.

Steward (2003) Trends Genet

Другие примеры

Disruption of protein–protein interactions: Tyr98His destroys binding between HIF and VHL ⇒ HIF not degraded ⇒ over-expression of angiogenic growth factors ⇒ local proliferation of blood vessels.



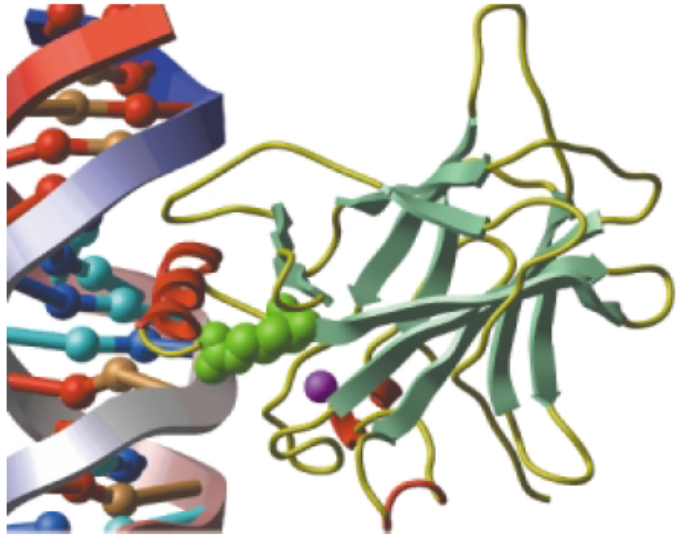
Vogelstein (2013) Science 74 Protein Data Bank rcsb.org

Von Hippel-Lindau syndrome (OMIM:193300) is an inherited pre-disposition to a variety of cancers. Von Hippel-Lindau disease tumor suppressor VHL codes for a protein with two structural domains. The β-domain of VHL binds to hypoxia-inducible transcription factor HIF, ultimately leading to HIF degradation.

Steward (2003) Trends Genet

Другие примеры

Disruption of DNA binding Arg273 contacts the DNA phosphate backbone with its charged side-chain. Arg273His is associated with low p53 DNA-binding and Li-Fraumeni syndrome.



Vogelstein (2013) Science 74 Protein Data Bank rcsb.org

Li-Fraumeni syndrome (OMIM 191170), a predisposition to a broad spectrum of cancers at an early age. Cellular tumor antigen p53 (*TP53*) is a tumor suppressor in many tumor types, induces growth arrest or apoptosis. Three functional domains: an N-terminal transcription factor domain, a DNA-binding core domain, and a C-terminal homooligomerization domain.

Steward (2003) Trends Genet

Эффект несинонимичных вариантов

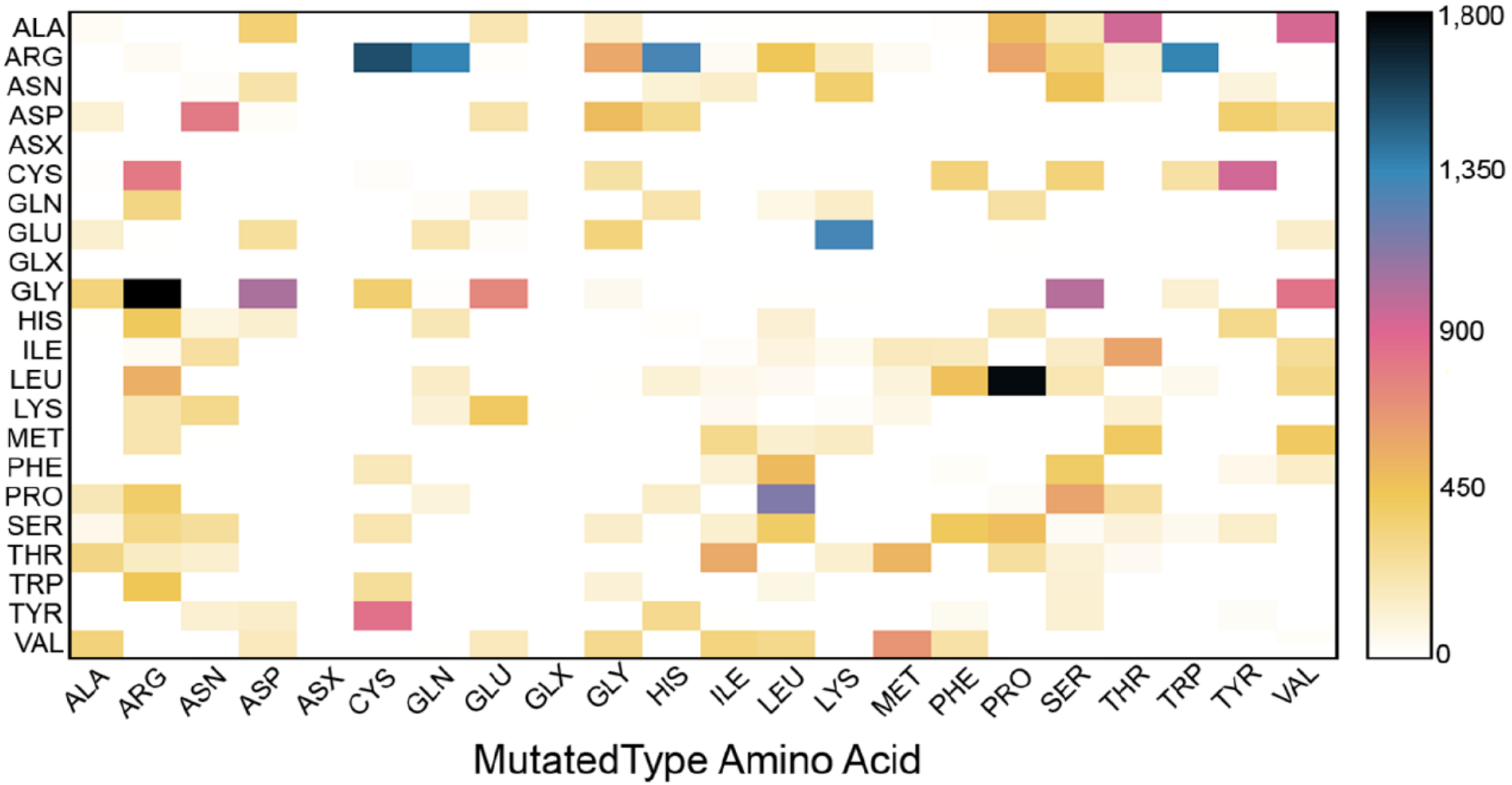
Fig. 5 Protein structural variation. Organization of the descriptive VariO terms, which facilitate very detailed annotation of observed effects

Vihinen (2015) *Human Gene*
`www.variationontology.org`

VariO - эффект вариантов на белок:

- Динамика
- Четвертичная структура
- Количество аминокислот
- Скорость фолдинга
- Взаимодействия
- Пост-трансляционные модификации
- Вторичная структура
- Фолдинг
- Эпигенетические модификации
- Количество
- Доступность
- Активность
- Заряд
- Деградация
- Растворимость
- Стабильность
- Местоположение в клетке

Болезнетворные несинонимичные // Peterson (2013) *J Mol Biol*



Упражнение

Укажите 10 самых частых вызывающих болезни миссенс-вариантов

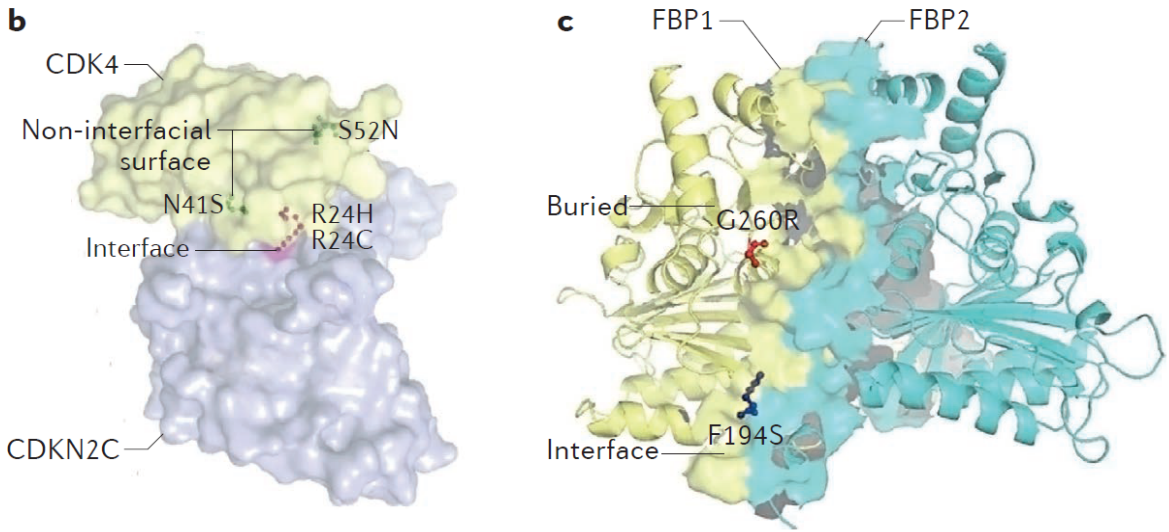
Болезнетворные несинонимичные: стабильность белка и ББВ

Table 1 | **Human diseases caused by defects in protein folding, stability and aggregation**

Disease	Protein affected	Description	References
Cystic fibrosis	Cystic fibrosis transmembrane conductance regulator (CFTR)	The ΔPhe508 mutant has wild-type activity, but impaired folding in the endoplasmic reticulum leads to degradation.	97
α1 Antitrypsin deficiency	α1 Antitrypsin (also known as SERPINA1)	80% of Glu342Lys mutants misfold and are degraded. Pathology is due to aggregation in patients with a reduced degradation rate.	97
SCAD deficiency	Short-chain acyl-CoA dehydrogenase (SCAD)	Impaired folding of Arg22Trp mutants leads to rapid degradation.	98
Alzheimer disease	Presenilin, γ-secretase	Mutations cause incorrect cleavage by the γ-secretase protease to produce the amyloid β-peptide; this aggregates into extracellular amyloid plaques.	99,100
Parkinson disease	α-Synuclein	Oxidative damage causes misfolding and aggregation. Hereditary forms are linked to deficiency in ubiquitin-mediated degradation.	101
Huntington disease	Huntingtin	CAG expansions in the Huntingtin gene lead to an abundance of polyglutamine fragments that aggregate and associate non-specifically with other cellular proteins.	101,102
Sickle cell anaemia	Haemoglobin	The Glu6Val mutation leads to aggregation in red blood cells.	103

De Pristo (2005) *Nat Rev Genet*

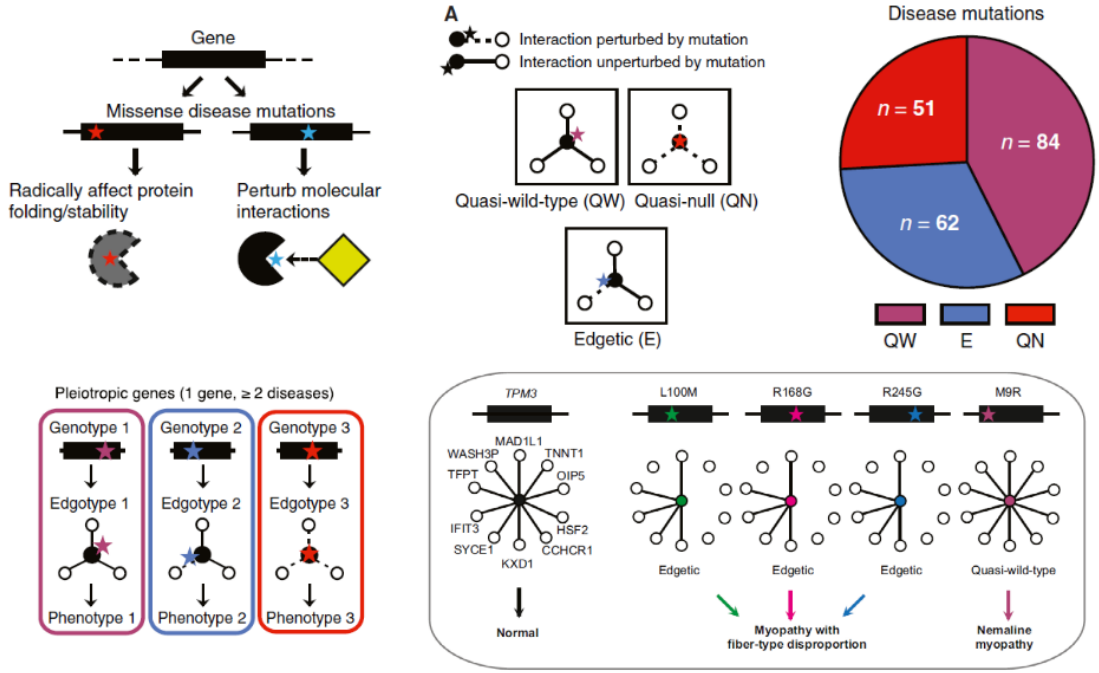
Болезнетворные несинонимичные: стабильность белка и ББВ



Yi (2017) *Nat Rev Gene*

b | Locations of residues affected by mutations are highlighted on the cyclin-dependent kinase 4 (CDK4) structure based on homology modelling (PDB: 1bi7). CDKN2C, CDK inhibitor 2C. **c** | Locations of residues affected by mutations are highlighted on the fructose bisphosphatase 1 (FBP1) structure (PDB: 1fpi).

Болезнетворные несинонимичные: стабильность белка и ББВ



Sahni (2015) *Cell*

Эффекты болезнетворных миссенс-мутаций на молекулярные взаимодействия могут разниться от отсутствия видимых изменений во взаимодействиях **(quasi-WT)**, до специфических потерь определенных взаимодействий **(edgetic)**, и до видимой полной потери взаимодействий **(quasinull)**.

Предсказание эффекта несинонимичных вариантов

Предсказание эффекта несинонимичных вариантов

Применения:

- Поиск генов, связанных с заболеваниями
- Клиническое секвенирование: $\sim$11 000 несинонимичных SNP на индивидуума, в том числе и редких
- Эволюционная, популяционная генетика
- Дизайн белков

Эффекты миссенс-вариантов могут быть крайне различными, эксперименты не представляются возможными. **Какие эксперименты?**

In vivo

- Клиническое влияние (редкие мутации, зависят от контекста и вида наследуемости)
- Модельные организмы: применимость?

In vitro

- Функциональное тестирование: применимость?

In silico: повреждающие | безвредные

- Источники данных и признаков
- Методы предсказания
- Оценка

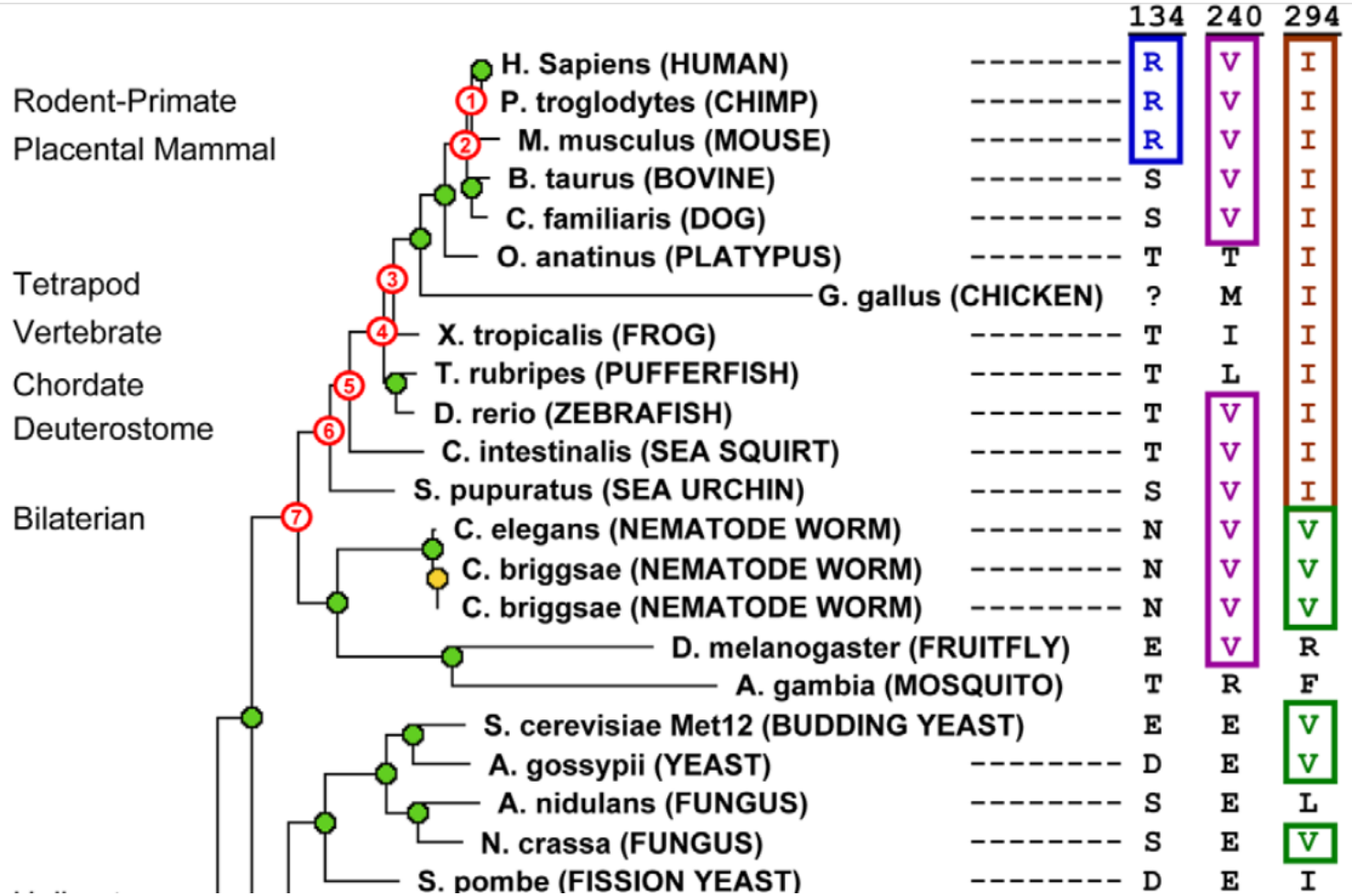
Постановка задачи

Источник данных	
Клинический эффект: ClinVar, HGMD	Болезнетворный (патогенный)
Биохимические исследования Публикации, Protein Mutant Database	Функциональный
Глубокое мутационное сканирование Публикации, MAVEdb	Функциональный
Популяционные данные dbSNP, ExAC/gnomAD, другие виды	Вредный
Филогенетические данные NCBI nr, UniPto, UCSC, MultiZ	Вредный

Признаки

1. Замены
   - Консервативные/радикальные (BLOSUM, Grantham score)
   - Изменение объема и гидрофобности
2. Сайты
   - Консервативность
   - Расположение: ядро/поверхность (относительная площадь поверхности)
   - Контакты: белки, лиганды, ДНК/РНК
   - Вторичная структура, неупорядоченность
   - B-фактор
3. Белки
   - Количество взаимодействий
   - Количество упоминаний в PubMed

Филогенетическая информация



Marini (2010) *PLOS Genet*

Филогенетическая информация в виде профильной матрицы

Protein → Multiple Sequence Alignment

N	E	L	V	T	L	T	C	L	A	R	G	F	S	-	P	K	D	V	L	V	R	W	L
R	E	S	A	T	I	T	C	L	V	T	G	F	S	-	P	A	D	V	F	V	Q	W	M
G	G	S	L	R	L	S	C	V	A	S	G	I	T	-	F	S	G	Y	D	M	Q	W	V
T	P	G	L	T	L	T	C	T	V	S	G	F	S	-	L	S	S	Y	D	M	G	W	V
G	Q	K	A	K	M	R	C	I	P	E	-	-	-	-	K	G	H	P	V	V	F	W	Y
G	Q	E	A	T	L	W	C	E	P	I	-	-	-	-	S	G	H	S	A	V	F	W	Y
G	Q	Q	V	T	L	S	C	F	P	I	-	-	-	-	S	G	H	L	S	L	Y	W	Y
R	K	D	V	S	L	T	C	L	V	V	G	F	N	-	P	G	D	I	S	V	E	W	T
G	Q	K	L	T	L	K	C	Q	Q	N	-	-	-	-	F	N	H	D	T	M	Y	W	Y
R	D	K	A	T	F	T	C	F	V	V	G	S	D	-	L	K	D	A	H	L	T	W	E
S	K	S	A	T	L	T	C	R	V	S	N	M	V	N	A	D	G	L	E	V	S	W	W
G	A	R	T	S	L	N	C	T	F	S	D	-	-	-	S	A	S	Q	Y	F	W	W	Y
G	A	S	L	Q	L	R	C	K	Y	S	Y	-	-	-	S	A	T	P	Y	L	F	W	Y
N	G	A	P	K	L	T	C	L	V	V	D	L	E	S	E	K	N	V	N	V	T	W	N
E	A	T	V	T	L	T	C	V	V	S	N	-	-	A	P	Y	G	V	N	V	S	W	T

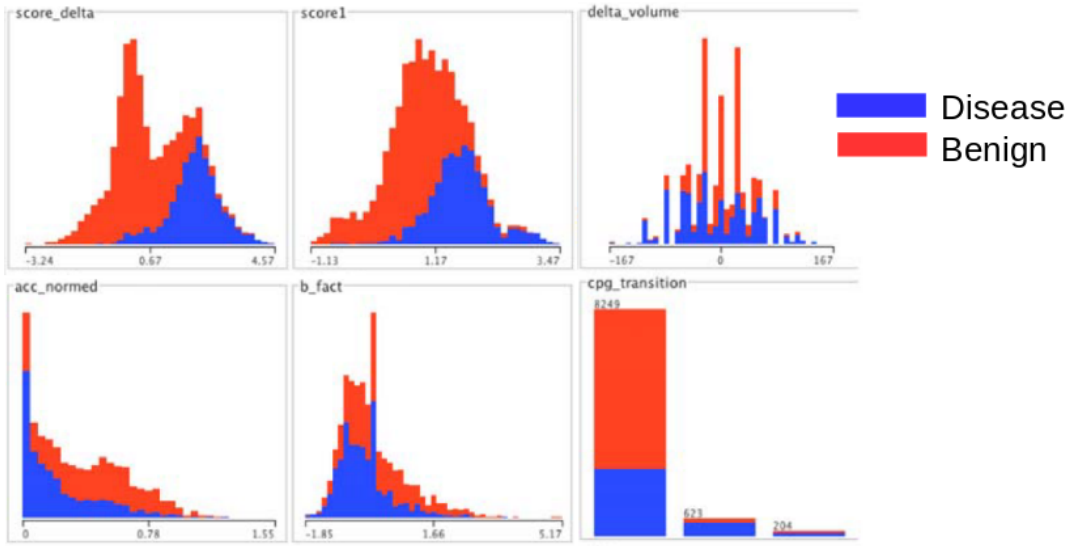
Profile

Ala	-1.2	1.1	-0.6	-0.8	0.3	...	...
Arg	0.6	-0.3	-0.3	-0.5	0.6	...	...
Asn	-1.1	-0.5	-0.5	-0.7	0.4	...	...
Asp	-0.9	-0.3	-0.3	-0.5	0.6	...	...
Cys	0.4	-0.5	0.6	0.8	-0.3	...	...
Gln	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

PSIC (Position Specific Independent Counts) profile scores matrix

Sunyaev (1999) *Protein Eng*

Задача обучения с учителем

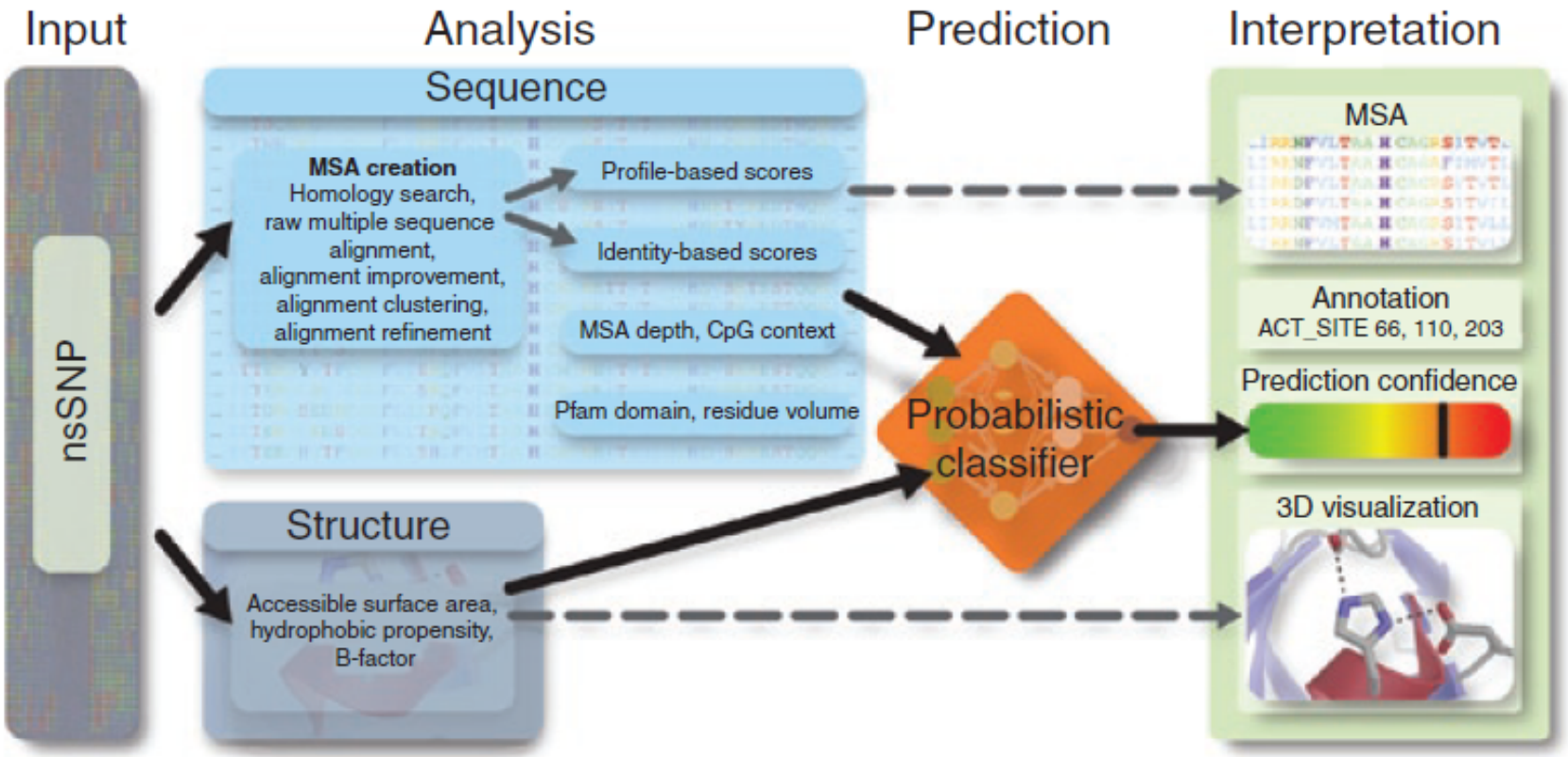


Adzhubei (2010) *Nat Methods*

Предсказание эффекта миссенс-вариантов PolyPhen-2

score_delta:	PSIC(AA1)-PSIC(AA2)
score1:	PSIC(AA1)
delta_volume:	изменение объема боковой цепи
cpg_transition:	CpG контекст (0:нет, 1: удаляет CpG, 2:создает)
acc_normed∗:	нормализованная доступная площадь поверхности (если известна 3D структура)
b_fact∗:	Средний температурный фактор

PolyPhen/PolyPhen-2

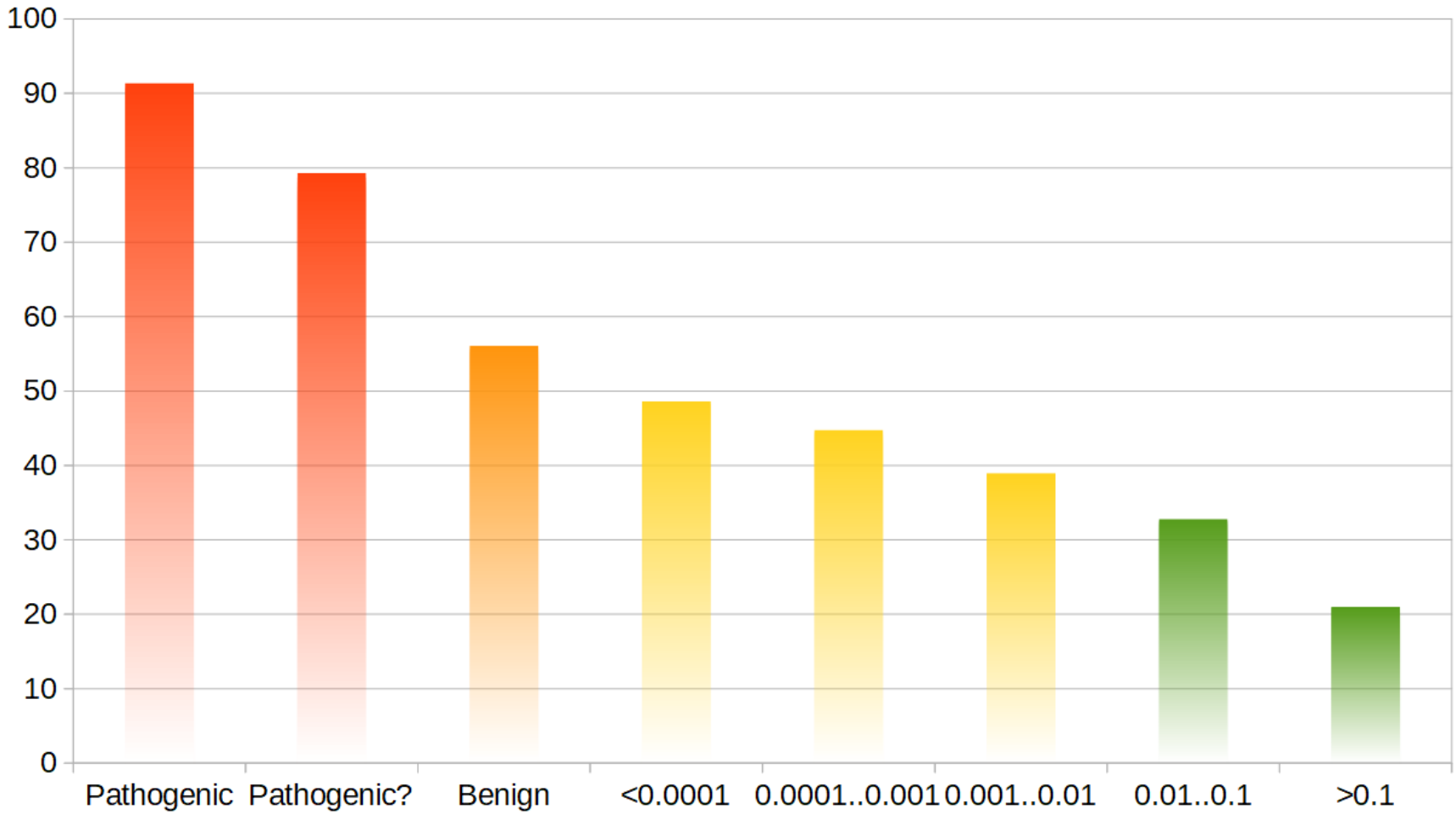


Adzhubei (2010) *Nat Methods*

Тренировочный набор данных (HumDiv): 3,155 болезнетворных мутаций; 6,321 замен между человеком-ортологом

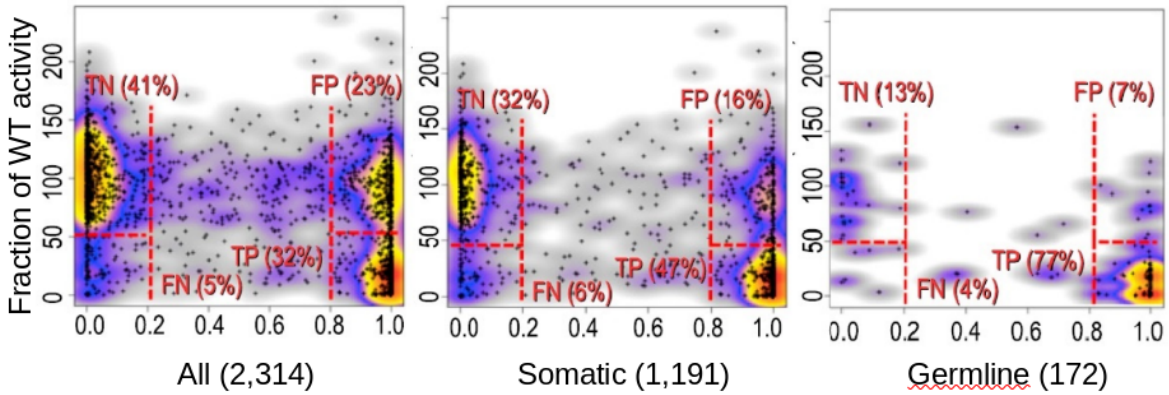
Показатели: FPR=10%, TPR=77%; FPR=20%, TPR=92%

Предсказание эффекта и популяционная частота варианта



ClinVar: болезнетворные мутации | ExAC: популяционные варианты по частоте аллеля

Что мы предсказываем?



Miosge (2015) *PNAS*

- Эксперимент: *in vitro* активность TP53 сравнивается с предсказаниями PolyPhen-2, порог: 50% активности WT
- Низкий уровень ложноотрицательных предсказаний, но
- 42% мутаций, предсказанных PolyPhen-2 как повреждающие, имели малые измеримые последствия на транскрипцию, регулируемую TP53
- Предсказания не могут как следуют различать мутации, которые непосредственно клинически важны (отменяют или значительно уменьшают способность выполнять функцию) и те, которые почти нейтральны (уменьшают способность соответствующего белка выполнять функцию на 10%)

Что мы предсказываем? Терминология

Damaging, deleterious, pathogenic, detrimental

The effect of a missense mutation on an organism is always multifaceted and can be considered from multiple perspectives – **biochemical, medical, and evolutionary**. The relationship between the effects of amino acid substitution on protein activity, human health, and an individual's evolutionary fitness is not trivial.

A mutation that damages protein structure does not necessarily lead to a detectable human-disease phenotype, and a mutation that predisposes an individual toward a disease is not necessarily evolutionarily deleterious. <...> Substitutions leading to abnormal hemoglobin function that cause sickle-cell anemia are apparently negative from both biochemical and medical points of view. Nevertheless, they cannot be considered negative from an evolutionary point of view, because balancing selection has brought them to high frequency in many parts of the world as a result of malaria resistance in heterozygotes.

To clearly distinguish different aspects of negative mutations, we use the term **damaging** to refer to a mutation that decreases protein activity, the term **detrimental** to refer to a mutation that predisposes an individual toward a disease, and the term **deleterious** to refer to a mutation that has been subject to purifying selection.

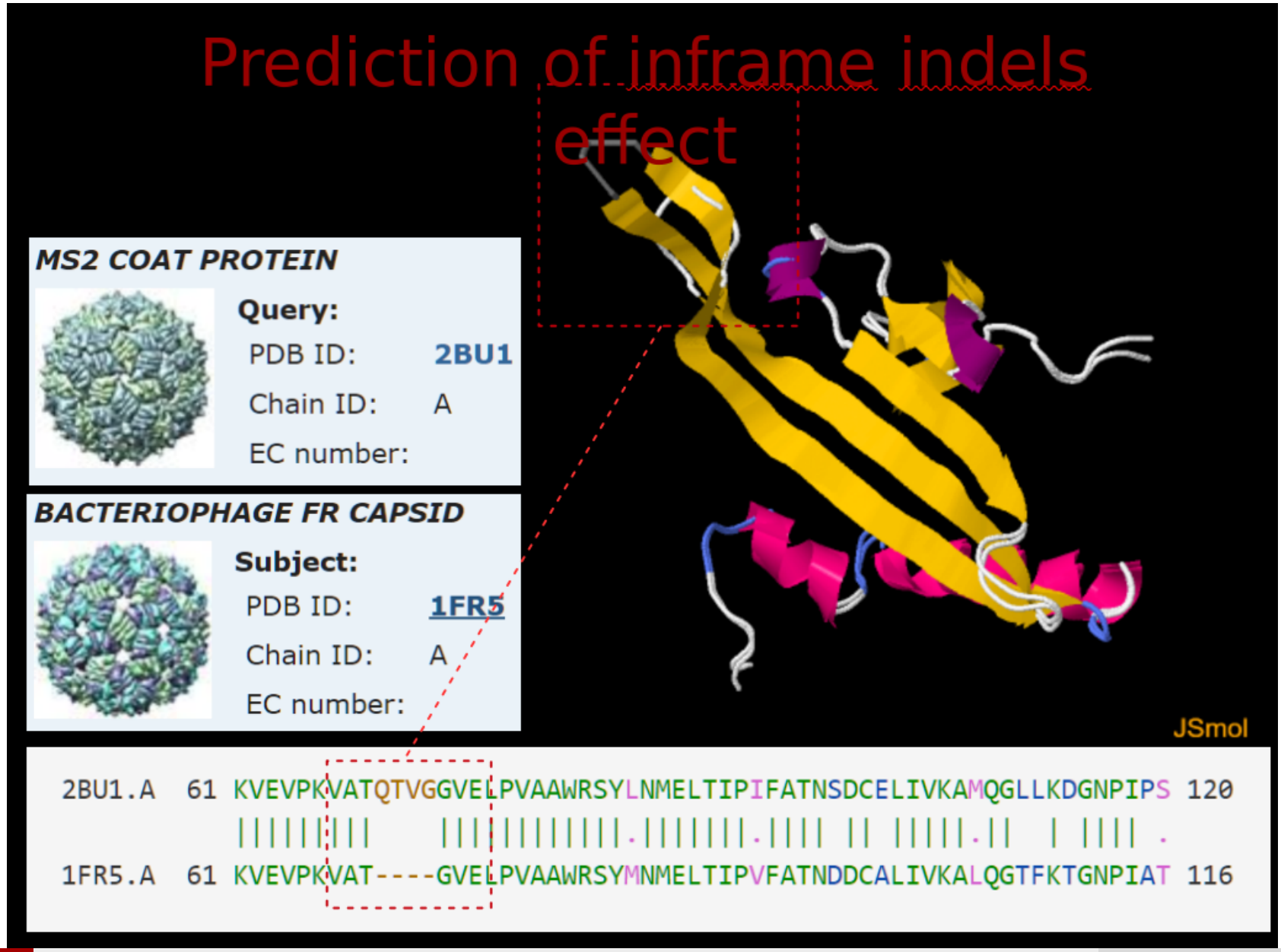
Kryukov (2007) *Am J Hum Genet*

Разное новое

- **Предсказание для полного протеома**: dbNSFP, 84 млн миссенс- и сплайс-сайт SNP
- **Ensemble (мета-) предсказатели**: MetaSVM, MetaLR, ReVel, M-CAP, и т.д.
- **Нейросети и другие методы ML**: PrimateAI, $\sim$ 380,000 частых миссенс-вариантов человека и приматов, градиентный бустинг
- **Ковариация**: EVmutation учитывает эпистаз, напрямую моделирует взаимодействия между всеми парами остатков
- **Предсказание количественного эффекта**: Envision измерил эффект 21,026 вариантов из 9 крупномасштабных экспериментальных наборов данных по мутагенезу
- **Клиническое применение**: M-CAP, 9 инструментов, 7 оценок консервативности, 298 признаков из множественного выравнивания последовательностей, градиентный бустинг

Варианты других типов

Инделы без сдвига рамки считывания



Инделы без сдвига рамки считывания

	Insertions, duplications	Deletions
ClinVar, 21 Oct 2019 (hg38)		
Pathogenic, Likely pathogenic	303	1,193
Benign, Likely benign	306	483
Other	1,291	3,566
GnomAD 2.1.1 (hg38)		
AF_POPMAX<1%	30,489	79,023
AF_POPMAX≥1%	742	1,517
Unknown	7,389	10,640
Individual exome (GiaB)	228	275

Вопрос

Какой самый известный индел без сдвига рамки считывания, вызывающий болезнь?

Инделы без сдвига рамки считывания

Gene	*ClinVar*	*gnomAD*
KCNH2 Potassium Voltage-Gated Channel Subfamily H Member 2	Pathogenic (4) Unknown (8)	Rare (11)
PHOX2B Paired Like Homeobox 2B	Benign (7) Pathogenic (4) Unknown (2)	Common (2) Rare/Unknown (14)
CACNA1A Calcium Voltage-Gated Channel Subunit Alpha1 A	Benign (5) Pathogenic (2)	Common (4) Rare/Unknown (42)
FOXC1 Forkhead Box C1	Benign (5) Pathogenic (3) Unknown (4)	Common (2) Rare/Unknown (49)

Предсказание эффекта инделов без сдвига рамки считывания

Название	Версия генома	Координаты	Реализация	Публикация	Последнее обновление
VEST-Indel	37, 38	Геномные	Web/ Local	2016	2019
CADD	37, 38	Геномные	Web/ Local	2013	2019
SIFT Indel	37, 38	Геномные	Web/ Local	2013	2016
MutPred-Indel	37?	Белковые	Web/ Local	2019	-
DDIG-in	37	Геномные	Web	2013	2017
Provean	37	Геномные	Web/ Local	2012	2017

Предсказание эффекта инделов без сдвига рамки считывания

Method	ML	Best features
VEST-Indel	Random forest	Log10 of count of publications in PubMed where gene name is mentioned, Exon Conservation, protein local regional sequence composition
CADD	SVM	cDNApos, ProtPos, PolyPhenVal, SIFTVal, Relative position in coding sequence
SIFT Indel	Decision tree	Repeat, DNA Conservation score, Protein disorder region, Fraction of all Pfam domains affected due to indel
MutPred-Indel	Neural Network	PSSM*, sequence conservation indices, number of homologs in the human and mouse genomes, relative position in protein
DDIG-in	SVM	Disorder, ASA*, DNA Conservation, Neff*, Probabylity of sheet
PROVEAN	Not ML	PROVEAN score

* PSSM – position-specific scoring matrix, ASA – solvent accessible surface area, Neff – number of effective homologous sequences aligned to residues

Предсказание эффекта инделов без сдвига рамки считывания

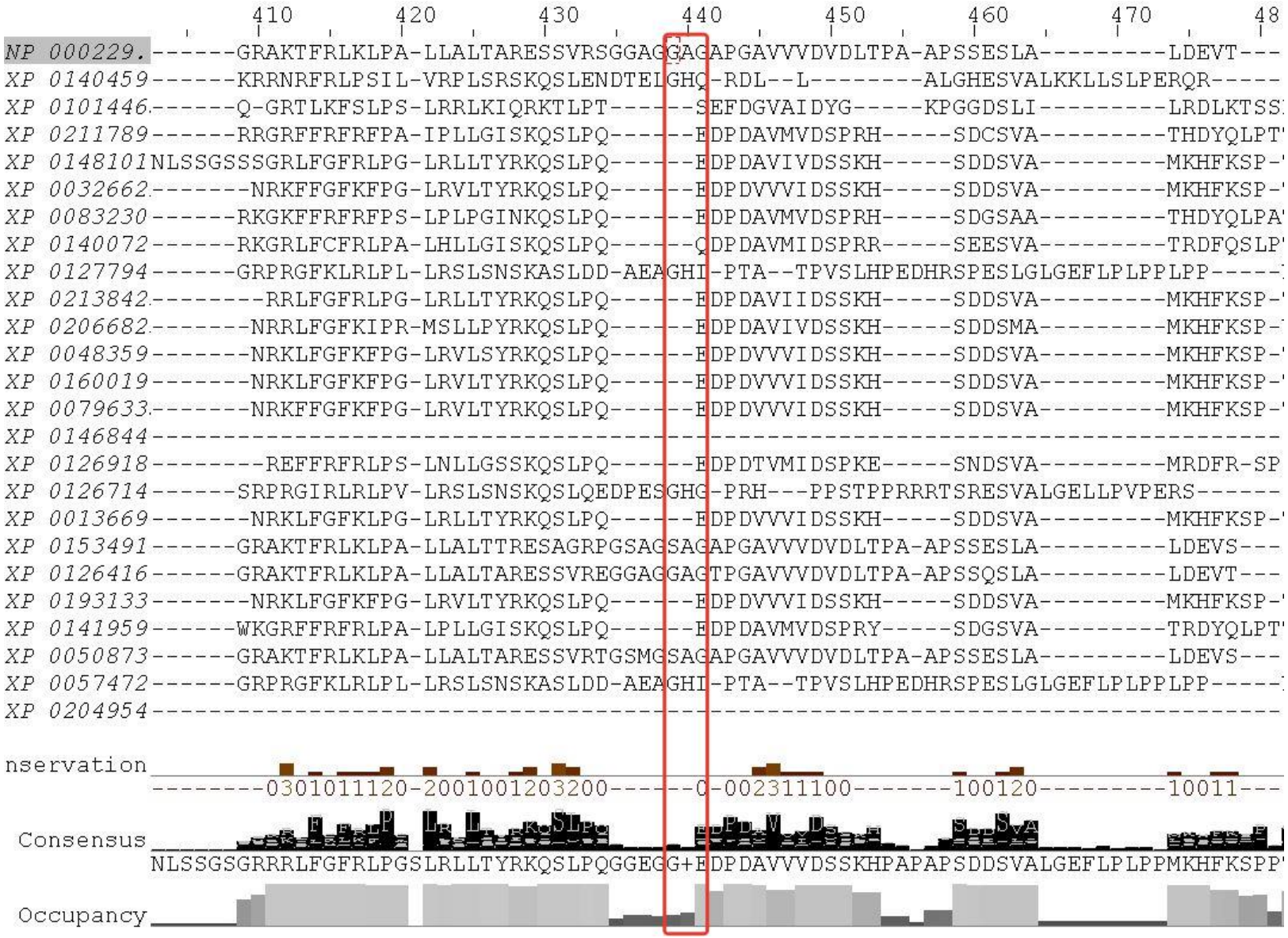
Meta-Predictors that Combine Classifications of Multiple Methods

In these Boolean expressions, each method is represented by a variable X_i, which is set to TRUE when the method classifies an example as pathogenic and FALSE when the method classifies an example as benign. For combinations of two methods, candidate meta-predictors were (X_1 and X_2) and (X_1 or X_2). For combinations of three methods, candidate meta-predictors (X_1 and X_2 and X_3), (X_1 or X_2 or X_3),(X_1 or X_2 or X_3), ((X_1 and X_2) or X_3), ((X_1 or X_2) and X_3), ((X_1 and X_3) or X_2),((X_1 or X_3) and X_2), ((X_2 and X_3) or X_1), ((X_2 or X_3) and X_1). For combinations of four methods, there are 64 possible combinations (Supp. Table S4). We used a brute-force approach and limited the number of methods in the meta-predictor to a maximum of four to avoid a combinatorial explosion. All possible four-way combinations of the five methods were explored.

Method	Sensitivity	Specificity	Balanced Accuracy
(VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)	0.930	0.974	0.952
(VEST-indel OR CADD) AND PROVEAN	0.947	0.955	0.951
(VEST-indel OR CADD) AND (PROVEAN OR DDIG-in)	0.947	0.949	0.948
VEST-indel OR (CADD AND PROVEAN AND DDIG-in	0.930	0.955	0.942
VEST-indel OR (CADD AND DDIG-in)	0.930	0.949	0.939
VEST-indel OR (DDIG-in AND CADD)	0.930	0.949	0.939
VEST-indel OR (CADD AND PROVEAN)	0.947	0.929	0.938
(VEST-indel OR DDIG-in) AND PROVEAN	0.930	0.942	0.936

Douville (2016) *Hum Mutation*

Предсказание эффекта инделов без сдвига рамки считывания



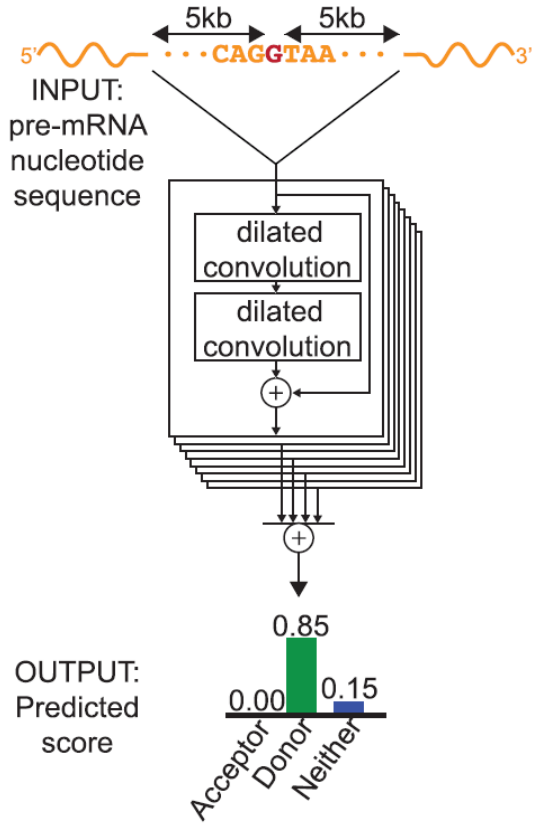
SpliceAI: предсказание сплайсинга по последовательности

Варианты в канонических сайтах сплайсинга (GT, AG) нарушают их.

Скрытые сплайс-варианты: некодирующие (интронные, синонимичные) варианты вне канонических сайтов сплайсинга, которые нарушают нормальный сплайсинг мРНК.

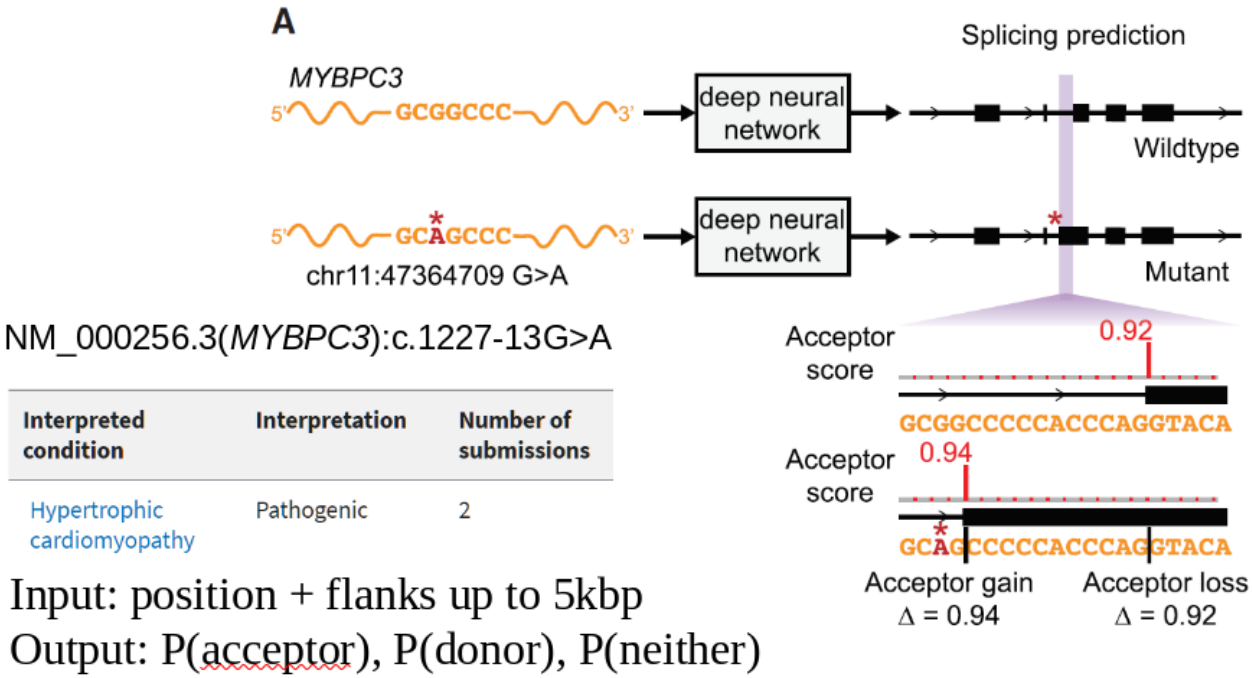
SpliceAI: 32-слойная глубокая нейронная сеть, которая предсказывает точки сплайсинга из любой последовательности транскрипта пре-мРНК.

Обучающий набор: транскрипты пре-мРНК; алгоритм выучивает контекст реальных сайтов сплайсинга.



Jaganathan (2019) *Cell*

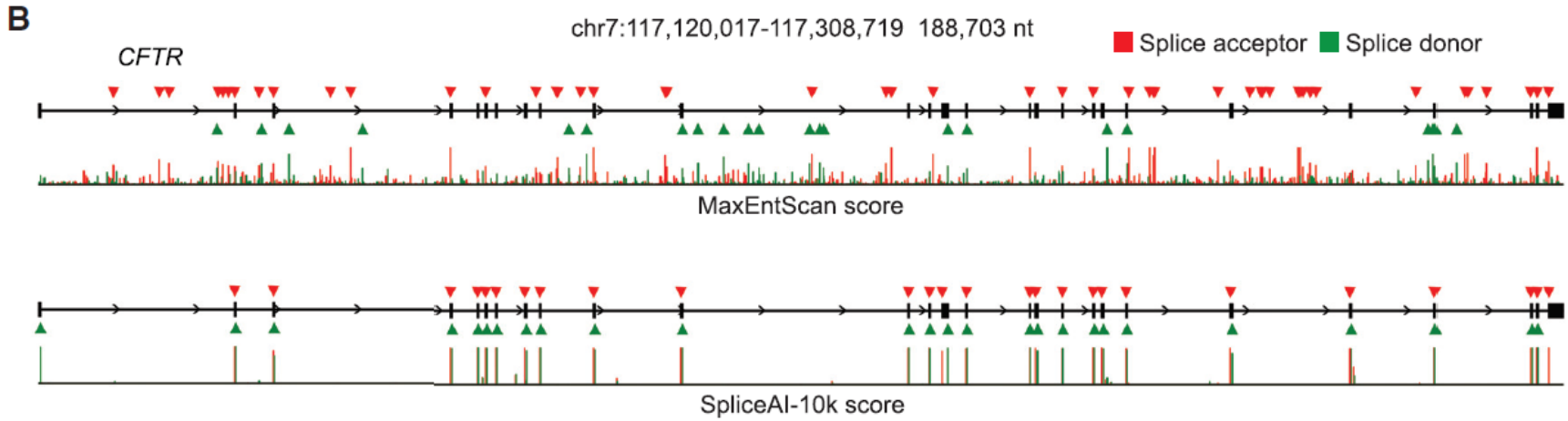
SpliceAI: предсказание сплайсинга по последовательности



Jaganathan (2019) *Cell*

SpliceAI-10k предсказывает вероятности акцепторов и доноров на каждой позиции последовательности пре-мРНК гена с мутацией и без нее, как показано для rs397515893, патогенного скрытого сплайс-варианта в интроне гена *MYBPC3*, связанного с кардиомиопатией. Значение D score для мутации равняется наибольшему изменению в оценке сплайс-предсказания в 50 п.н. от варианта.

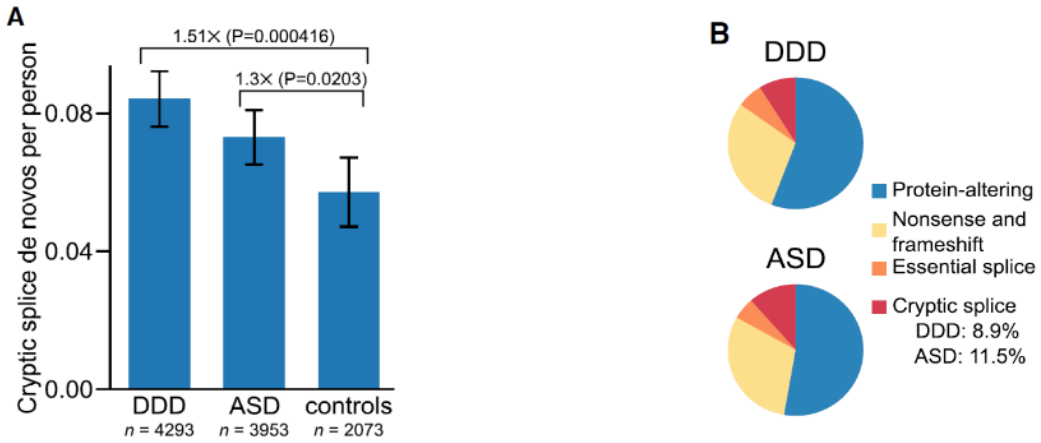
SpliceAI: предсказание сплайсинга по последовательности



Jaganathan (2019) *Cell*

Показан полный транскрипт пре-мРНК для гена *CFTR*, оцененный с использованием MaxEntScan (сверху) и SpliceAI-10k (снизу), вместе с предсказанными акцепторными (красные стрелки) и донорными (зеленые стрелки) сайтами и настоящим положением экзонов (черные прямоугольники). Для каждого метода был применен порог таким образом, чтобы количество предсказанных сайтов равнялось общему количеству реальных сайтов.

SpliceAI: предсказание сплайсинга по последовательности



Jaganathan (2019) *Cell*

(A) Предсказанные скрытые *de novo* сплайс-мутации (на 1 человека) среди пациентов Deciphering Developmental Disorders cohort (DDD), людей с расстройствами аутистического спектра (ASDs) из Simons Simplex Collection и the Autism Sequencing Consortium, а так же здоровых контролей.
(B) Оценочная доля патогенных *de novo* мутаций по функциональным категориям в когортах DDD и ASD, основываясь на сравнении с контрольной выборктой.
Скрытый сплайсинг может составлять до 10% патогенных вариантов в расстройствах нервно-психического развития.

Предсказание эффекта некодирующих вариантов

CADD: Combined Annotation–Dependent Depletion интегрирует различные аннотации и признаки генома для любого возможного SNV или индела человека.

Количество вредных вариантов (уменьшающих приспособленность организма)уменьшается за счет естественного отбора в фиксированных, но не симулированных данных.

Наблюдаемые варианты (15 млн SNV, 0,63 млн инсерций и 1,1 млн делеций):

- Различия между человеком и шимпанзе, исключая варианты с MAF>5%
- SNP с часотой производного аллеля > 95% (<5% общей)

Симулированные варианты (44 млн SNV, 2,1 млн инсерций и 3,1 млн делеций):

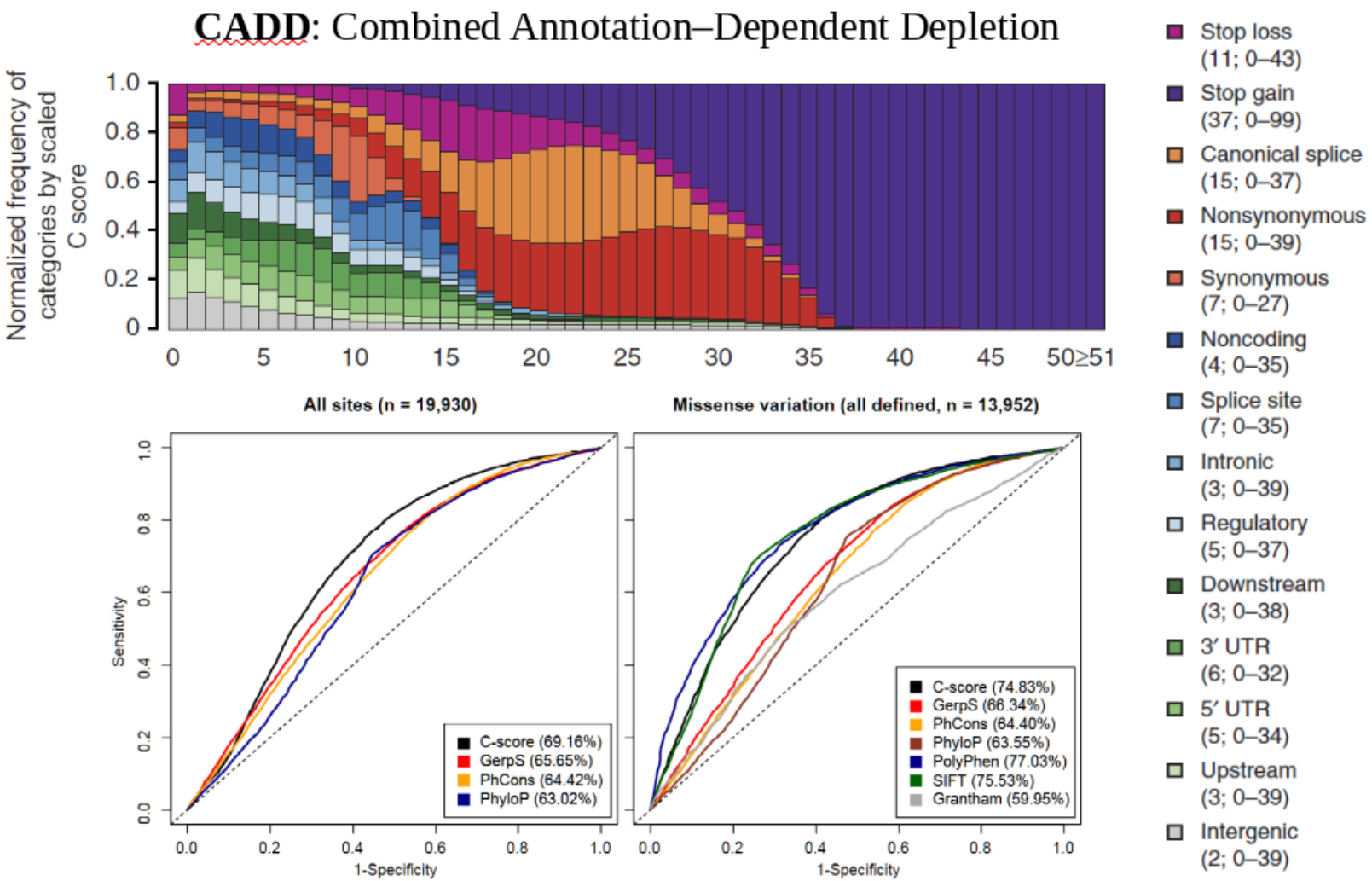
- Полностью эмпирическая модель эволюции последовательности с отдельной частотой для CpG динуклеотидов и локальной оценки скоростей мутирования

Признаки: Аннотации VEP, SIFT, PolyPhen-2, оценки консервативности, метилирование и гистоновые модификации в различных типах клеток и тканей, сайты связывания транскрипционных факторов и т.д.

Выдача: C-оценки, которые измеряют вредность 8.6×10^9 вариантов.

Kircher (2014) *Nat Genet*

Предсказание эффекта некодирующих вариантов



Kircher (2014) *Nat Genet*

Предсказание эффекта некодирующих вариантов

Score	Data sources	Approach
Eigen	• Uses data from the ENCODE and Roadmap Epigenomics projects	• Weighted linear combination of individual annotations • Unsupervised learning method
FunSeq2	• Inter- and Intra-species conservation • Loss- and gain-of-function events for transcription factor binding • Enhancer–gene linkage	• Weighted scoring system
LINSIGHT	• Conservation scores (phastCons, phylopP), predicted binding sites (TFBS, RNA), regional annotations (ChIP-seq, RNA-seq)	• Graphical model • Selection parameter fitting using generalized linear model based on 48 genomic features
CADD	• Ensembl variant effect predictor • Protein-level scores: Grantham, SIFT, PolyPhen • DNase hypersensitivity, TFBS, transcript information • GC content, CpG content, histone methylation	• Support vector machine
FATHMM	• 46-way sequence conservation • ChIP-seq, TFBS, DNase-seq • FAIRE, footprints, GC content	• Hidden Markov models
ReMM	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • 26 features: PhastCons, PhyloP, CpG, GC, regulation annotations	• Random forest classifier
Orion	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • Independent from annotation and features	• Expected and observed site-frequency spectrum of a given stretch of sequence
CDTS	• Identify constrained non-coding regions in the human genome and deleteriousness of variants • Independent from annotation and features. Uses *k*-mers	• Expected and observed site-frequency spectrum of a given heptamer

Telenti (2018) *Hum Mol Genet*

Предсказание эффекта некодирующих вариантов

Table 2 Summary of genomic features used for LINSIGHT scores

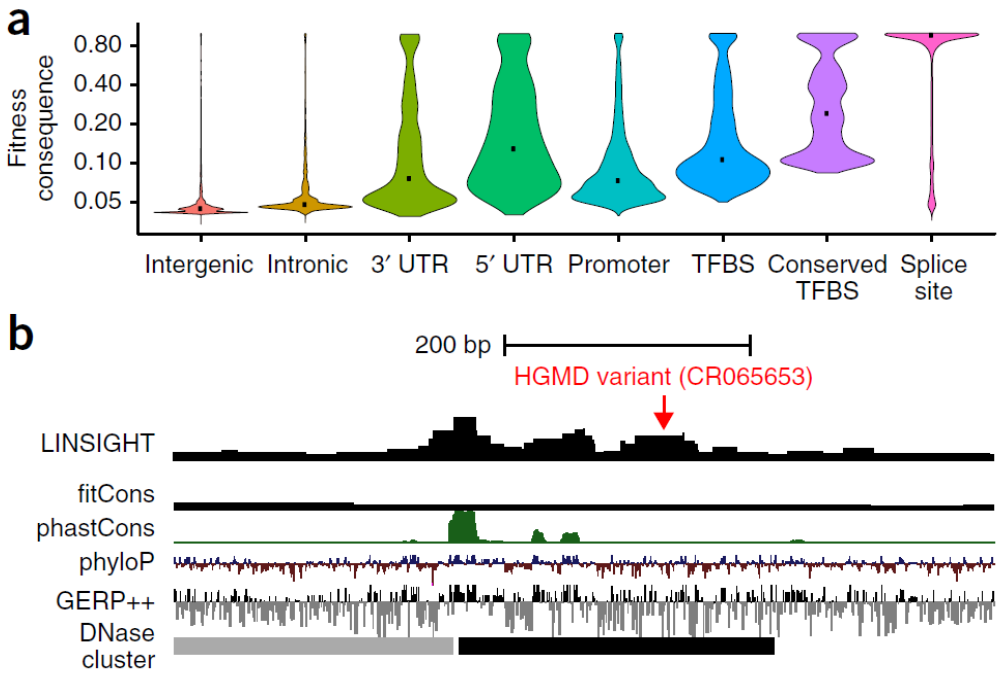
Class	Genomic feature[a]	Spatial resolution
Conservation	phyloP score	High
	phastCons element	High
	SiPhy element	High
	CEGA element	High
Binding site	Conserved TFBS	High
	rVISTA TFBS	High
	SwissRegulon TFBS	High
	Predicted TFBS within ChIP-seq peak	High
	Conserved miRNA binding site	High
	Splicing site predicted by SPIDEX	High
Regional annotation	ChIP-seq peak of transcription factor	Low
	DNase-I hypersensitive site	Low
	UCSC FAIRE peak	Low
	RNA-seq signal	Low
	Histone modification peak	Low
	FANTOM5 enhancer	Low
	Predicted distal regulatory module	Low
	Distance to nearest TSS	Low

[a]Each 'genomic feature' listed here may actually correspond to multiple features in the model. For example, four features are derived from phyloP scores: two from the mammalian phyloP scores and two from the vertebrate phyloP scores. See **Supplementary Table 3** for complete details.

Huang (2017) *Nat Genet*

LINSIGHT интегрирует функциональные геномные данные с оценками консервативности и другими признаками.

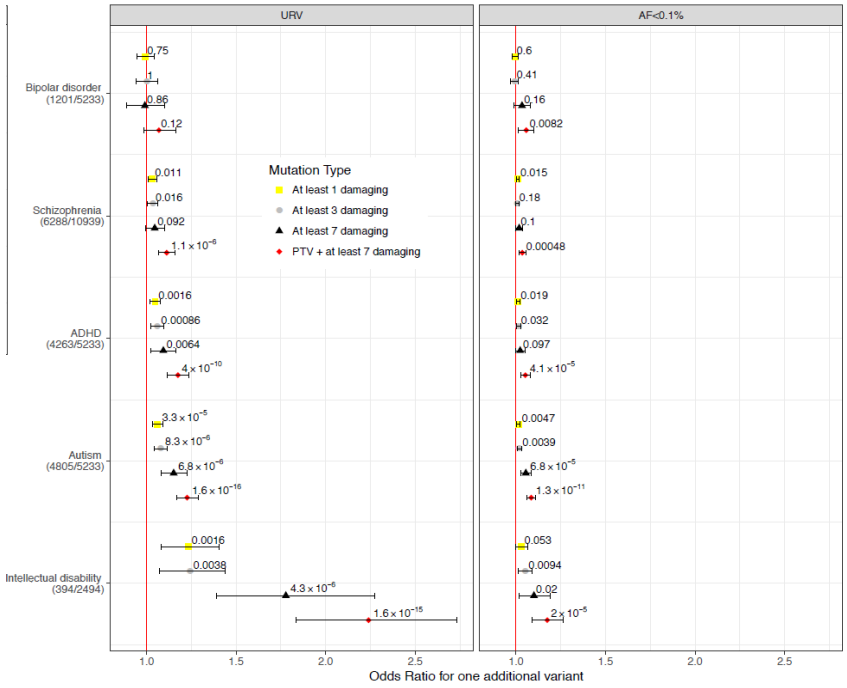
Предсказание эффекта некодирующих вариантов



Huang (2017) *Nat Genet*

(a) Распределение оценок LINSIGHT для различных участков геномов. Межгенные регионы, интронные участки, UTRs, и промоторы длиной в 1 кб: GENCODE 19; сайты связывания транскрипционных факторов (TFBSs): пики ChIP-seq (Ensembl Regulatory Build); консервативные TFBSs: UCSC Genome Browser. **(b)** LINSIGHT единственный метод, которые выделяет вариант из HGMD (CR065653), который ассоциирован с повышенной регуляцией гена *TERT*.

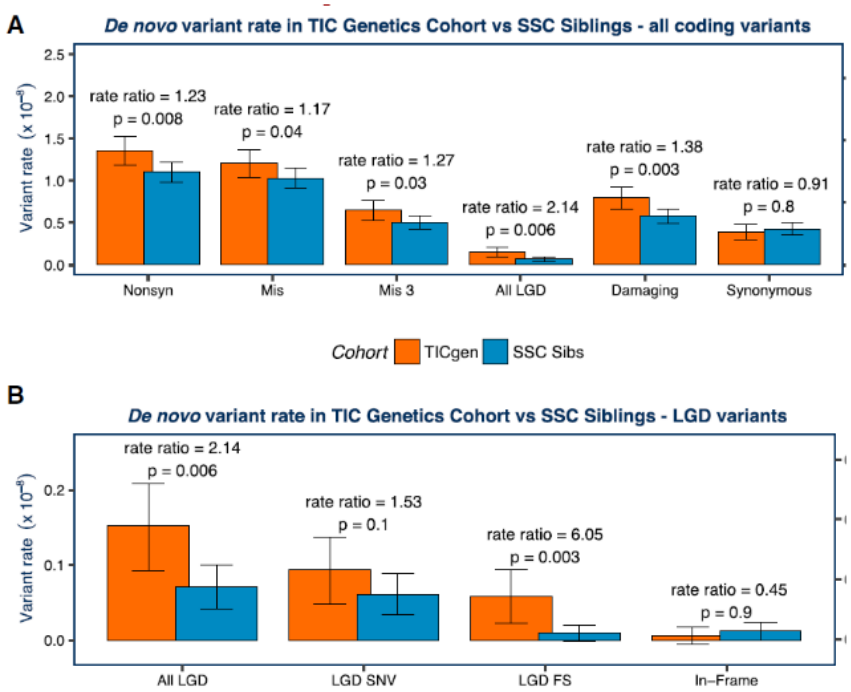
Эффект вариантов и ассоциация с фенотипами



Ganna (2018) *Am J Hum Genet*

Мета-анализ ассоциаций ультра-редких и редких вредных миссенс-вариантов в генах, выносящих PTV и 5 заболеваний. **Степень ассоциации увеличивается как функция от нескольких алгоритмов, и особенно велика среди ультра-редких вариантов.**

Эффект вариантов и ассоциация с фенотипами



Willsey (2017) *Neuron*

Все классы *de novo* несинонимичных вариантов показывают большую частоту среди пробандов с расстройством Туретта (оранжевый) против SSC братьев и систер (контроли, синие). LGD: likely gene disrupting variants: insertion of premature stop codon, frameshift, or canonical splice-site variant; **FS**: frameshift indels; **Damaging**: variants predicted by PolyPhen2; **Mis3**: LGD or damaging; **Nonsyn**: missense or nonsense.

Выводы

1. Последовательность генома человека все еще обновляется. Возможно, скоро мы переключимся с одной референсной последовательности на несколько сразу.
2. Белоккодирующие гены представляют лишь маленькую часть всех генов человека и ничтожную часть генома.
3. Примерно половина генома — повторяющиеся последовательности.
4. Структура и процессирование генов человека весьма сложны и разнообразны.
5. Сразу несколько участков генома могут участвовать в сплайсинге: интронные и экзонные энхансеры и сайленсеры сплайсинга. Значительная часть мутаций, вызывающих заболевания у человека, считаются ассоциированными со сплайсингом.
6. Эпигенетика отвечает за наследуемые изменения фенотипов, которые не связаны с изменениями в последовательности ДНК: метилирование ДНК в CpG участках, ковалентные модификации гистоновых белков. Некодирующие РНК считаются частью эпигенетической машинерии.

Список литературы

- Strachan, Read – Human Molecular Genetics, Chapter 13
- Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348, 666–669.
- Saleheen, D., Natarajan, P., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544, 235–239
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535-548.e24.
- Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Human Mutation 37, 579–597.

Список литературы

- Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y., Wang, X., and Sun, Z. (2018). Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 46, 7793–7804.
- DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet 6, 678–687.
- 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.
- Lee, P., Lee, C., Li, X., Wee, B., Dwivedi, T., and Daly, M. (2018). Principles and methods of in-silico prioritization of non-coding regulatory variants. Hum Genet 137, 15–30.
- Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics 18, 599.