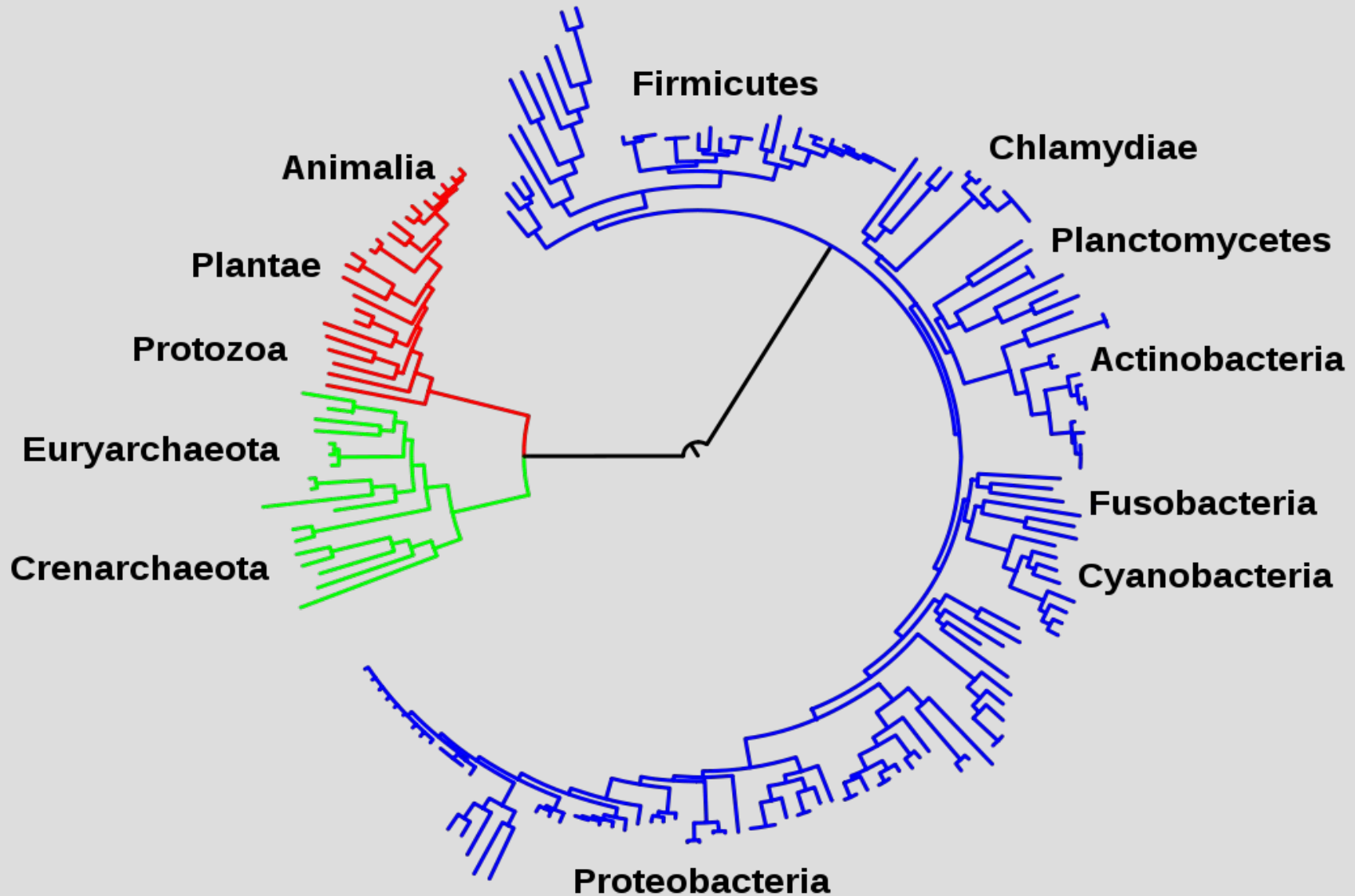


Занятие №4. Структура и полиморфизм геномов

Филогенетическое древо живых организмов



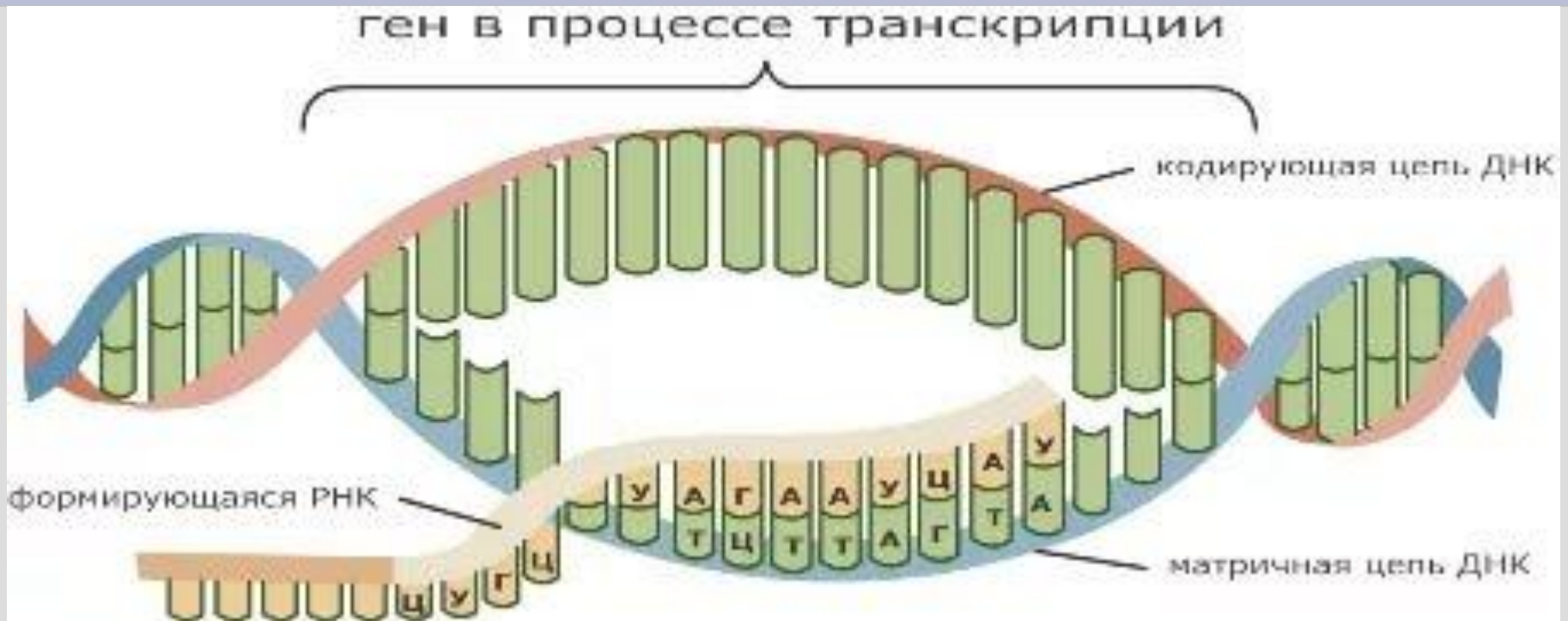
Размеры геномов

Домен/группа	Длина генома, п.н.	Количество генов
		(обычно больше из-за перекрывания генов)
		(обычно меньше из-за преобладания “мусорной” ДНК)

Открытая рамка считывания (Open Reading Frame, ORF)

- Потенциально кодирующая последовательность в геноме:
- Начинается с иницирующего кодона (старт-кодон) (ATG)
- На большом расстоянии отсутствуют терминирующие кодоны (стоп-кодоны) (TAA, TGA, TAG)
- Один иницирующий кодон даёт 6 вариантов потенциальных генов (чтение из положений +1, +2, +3 с двух цепей ДНК).

Транскрипция



Кодирующая (смысловая; +) цепь: 5'-СТГСТАГААТСАТ-3'

Матричная (-) цепь: 3'-GACGATCTTAGTA-5'

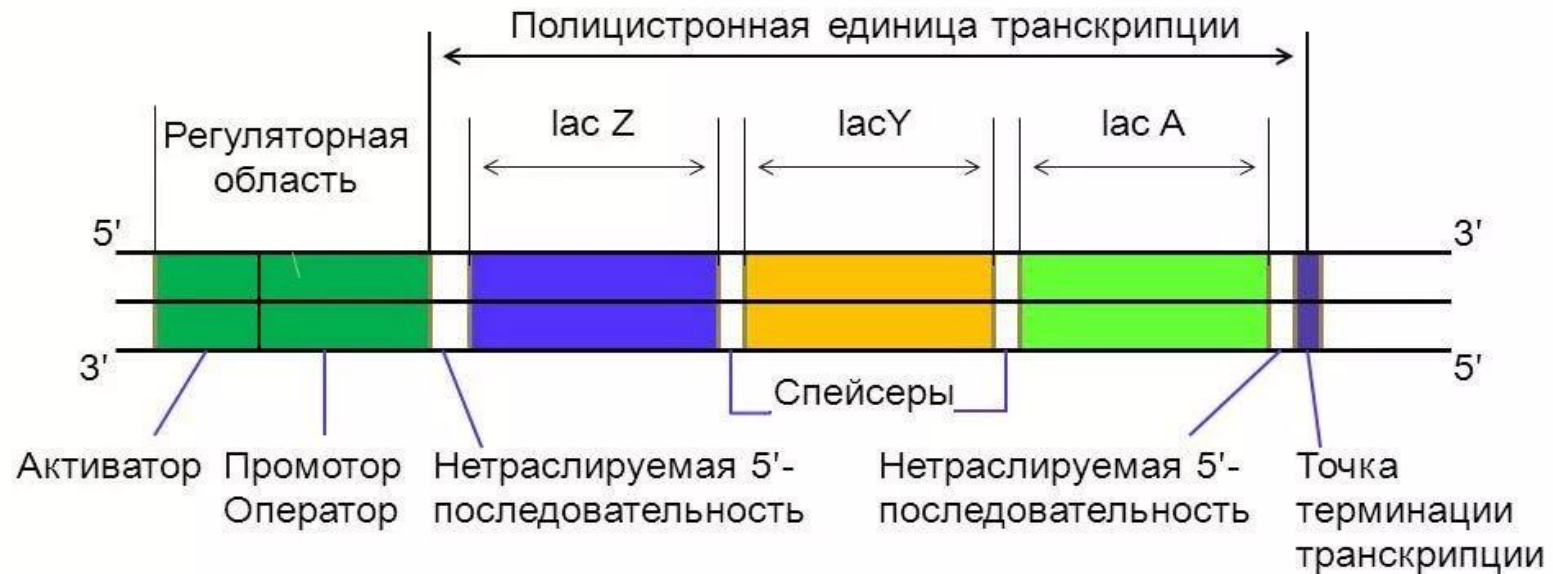
мРНК-транскрипт: 5'-CUGCUAGAAUCAU-3'

Особенности геномов прокариот (бактерий и архей)

- Сравнительно малый размер ($\sim 10^5$ - 10^7 п.н.)
- Гаплоидность
- Унипартитность/Бипартитность (одна-две разные хромосомы)
- Распространённость кольцевых геномов
- Наличие плазмидного эписомного генома ($\sim 10^3$ - 10^6 п.н.)
- Большая плотность генов (90% генома – кодирующие последовательности)
- Гены часто организованы в общую группу транскрипции (оперон)
- Гены практически не содержат интронов
- Средний размер гена ~ 1000 п.н.

Структура бактериального оперона

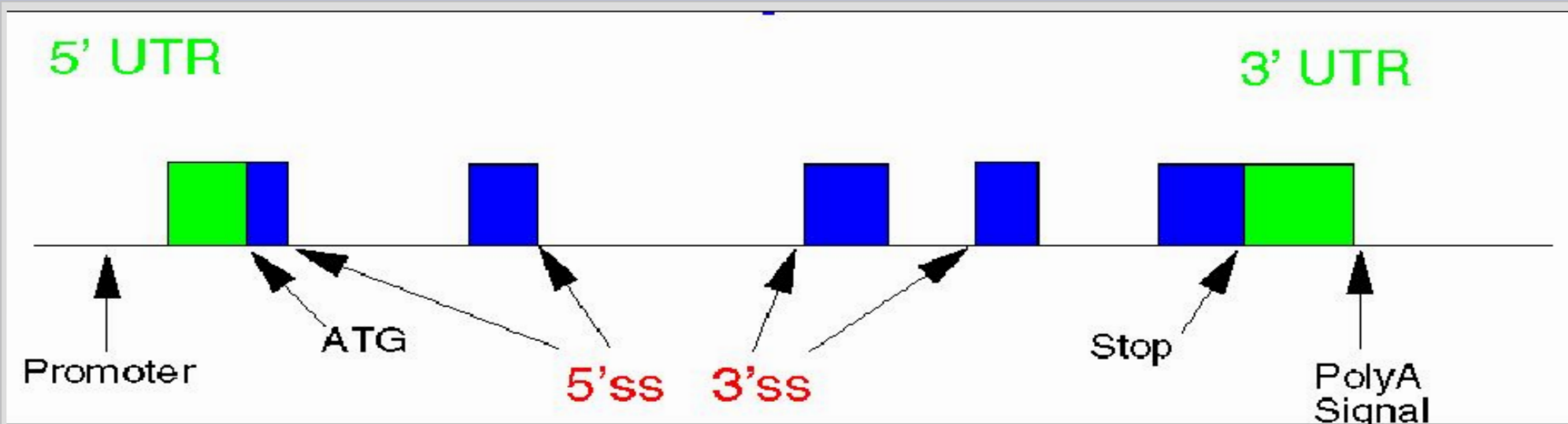
Схема строения *lac*-оперона



Особенности геномов эукариот

- Сравнительно большой размер ($\sim 10^8$ - 10^9 п.н.)
- Мультипартитность и мультиплоидность
- Организация в виде хроматина, имеющего тонкую структуру
- Малая плотность генов (90% генома – некодирующие последовательности)
- Гены крайне редко организованы в общую группу транскрипции (оперон), транскрибируются по отдельности
- Гены мозаичны – содержат кодирующие участки экзоны (~ 5 - 6) и некодирующие интроны (~ 4 - 5)
- Длина интрона (~ 2000 п.н.) в 10 раз превышает длину экзонов (~ 200 п.н.)

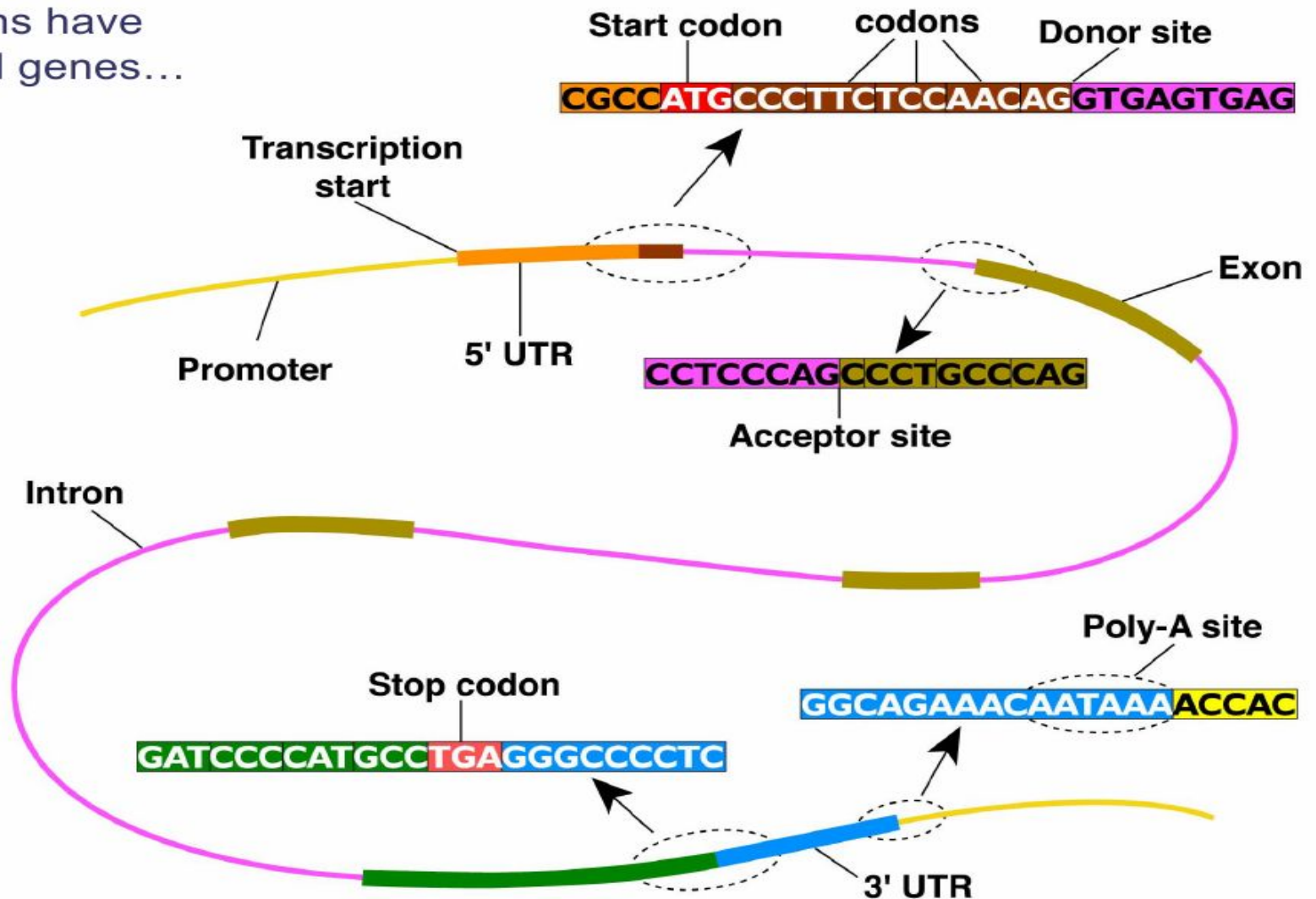
Открытая рамка считывания у эукариот



- Иницирующий кодон ATG
- Стоп-кодона TAA/TAG/TGA
- Интроны ограничены сайтами сплайсинга (SS)
 - – 5' SS: **GT**
 - – 3' SS: **AG**

Мозаичное строение эукариотного гена

Humans have spliced genes...



Инфраструктура генома

- Кодирующие последовательности (гены)
 - – белковые (структурные и регуляторные)
 - – РНК-гены
 - – гены мобильных генетических элементов и интронов (прокариоты)
- Некодирующие последовательности
 - – контролирующие генетические элементы (промоторы, терминаторы, операторы)
 - – повторяющиеся последовательности (повторы)
 - – внутренние спейсеры (прокариоты) и интроны (эукариоты)
 - – другие регуляторные последовательности

Гены белков

- Однокопийные/многокопийные
- Паралогичные
- Гены-сироты (orphan genes)
- Криптические (молчащие)
- Псевдогены
- Взаимноперекрывающиеся

РНК-гены

РНК	Длина (н.)	Функция	Источник

snRNA – малая ядерная РНК

lncRNA – длинная некодирующая РНК

miRNA – микроРНК

crRNA – CRISPR-РНК

Повторяющиеся некодирующие последовательности

- короткие (<100 п.н.) или длинные (>100 п.н.)
- прямые или инвертированные
- тандемные или диспергированные

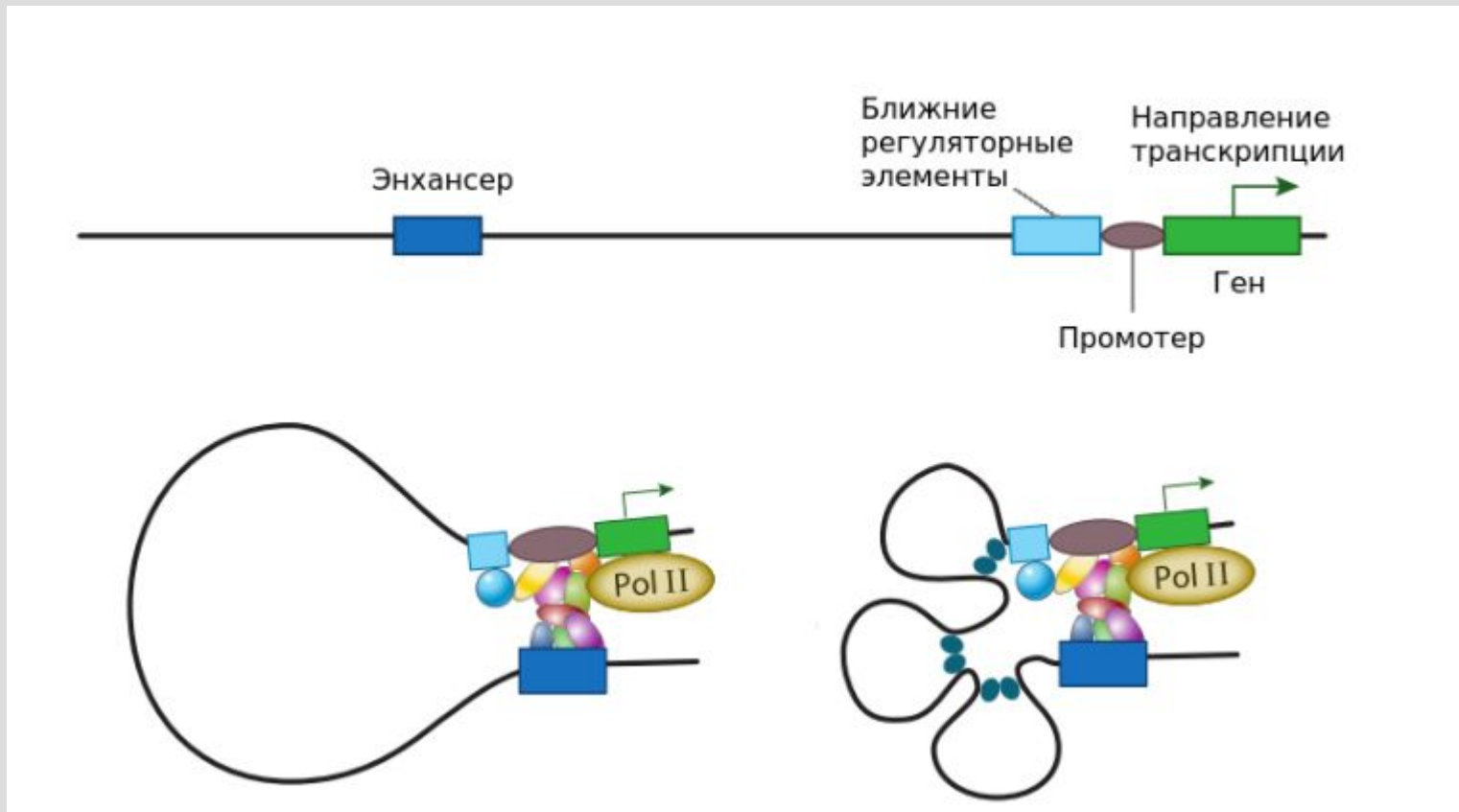
Тандемные повторы – результат ошибки в работе ДНК-полимеразы

Название	Длина повтора, п.н.	Количество копий в кластере	Расположение в геноме
Сателлиты	$\sim 1-10^3$	1.000 – 10.000	Теломеры и центромеры эукариот
Минисателлиты	10-100	5-50	Вблизи теломер, в эухроматине эукариот и любой части генома прокариот
Микросателлиты, или короткие тандемные повторы (STR)	<10	5-50	В эухроматине эукариот и любой части генома прокариот

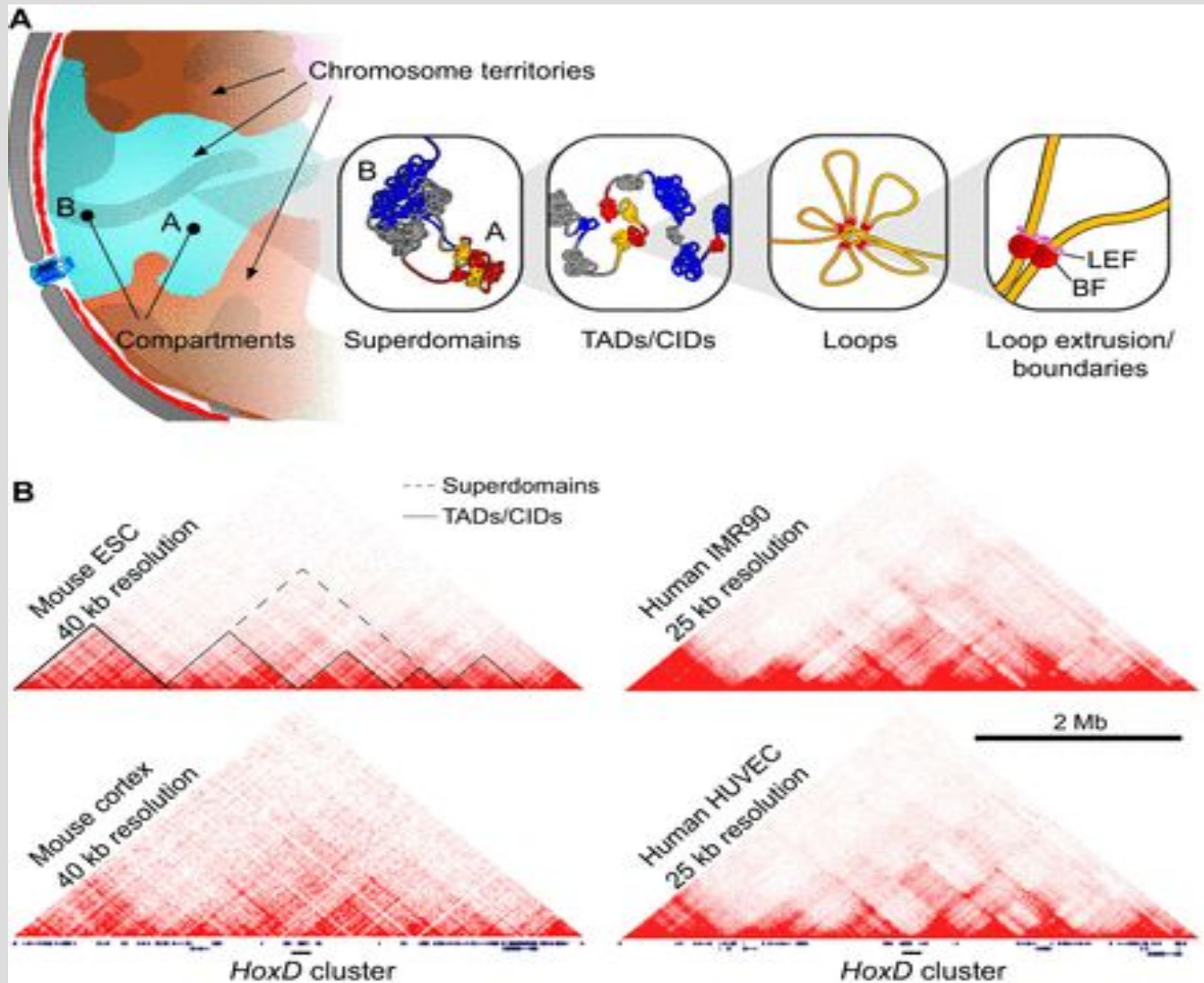
Диспергированные повторы – ретротранспозоны

Название	Длина повтора, п.н.	Количество копий	Встречаемость

Пространственное взаимодействие в геноме эукариот



Топологически ассоциированные домены (TAD) – области наиболее тесного взаимодействия в геноме



Генетический полиморфизм

- 1. Явление существования вариантов в нуклеотидной последовательности разных представителей одного вида.
- 2. Собственно вариант в нуклеотидной последовательности.

- Все представители вида *Homo sapiens* идентичны по нуклеотидной последовательности геномов на 99,9%

Виды генетических полиморфизмов

- Однонуклеотидный полиморфизм (ОНП; single nucleotide polymorphism, SNP, 'snip')
- Инсерция/делеция (индел)
 - – простая
 - – микросателлиты (полиморфизм коротких tandemных повторов; short tandem repeat polymorphisms, STRP)
 - – мини-сателлиты (переменное число tandemных повторов; variable number tandem repeats, VNTR)
 - – диспергированные повторы (напр. SINE)

SNPs

- Самый частый вид полиморфизма
- Частота: 1 нуклеотид на каждую 1000 п.н. ($\sim 3 \cdot 10^6$ различий между любыми двумя геномами человека)
- Отличие от точковой мутации: встречается в популяциях с вероятностью $\geq 1\%$ (“распространённые SNP” $> 20\%$)
- Виды SNP в экзоне:
 - – миссенс-замена (замена одной АК на другую)
 - – нонсенс-замена (замена кодона на стоп-кодон)
 - – синонимичная замена (новый кодон кодирует ту же АК)

Полиморфизм тандемных повторов

- 1. Полиморфизм числа тандемных повторов
 - – VNTR (полиморфные минисателлиты, отличающиеся числом повторов в разных локусах у разных организмов)

- 2. Полиморфизм числа копий диспергированных повторов
 - – Alu-повторы (вариант SINE)

Виды SNP в зависимости от природы меняющихся азотистых оснований

- Транзиции (~2/3 всех SNP) – пуриновое основание заменяется пуриновым или пиримидиновое – пиримидиновым:
 - – **A \rightleftharpoons G**
 - – **C \rightleftharpoons T**
- Трансверсии (~1/3 всех SNP) – пуриновое основание заменяется пиримидиновым или наоборот:
 - – **A \rightleftharpoons T**
 - – **A \rightleftharpoons C**
 - – **G \rightleftharpoons C**
 - – **G \rightleftharpoons T**

Обозначение SNP в формате FASTA

Код	Значение	Объяснение

Номенклатура полиморфизмов

- 1. Указание источника последовательности
- геномная ДНК (g.)
- комплементарная ДНК (с.)
- последовательность белка (p.)
- мтДНК (m.)
- 2. Положение полиморфизма от начала отсчёта
- для кДНК начало – аденин иницирующего кодона ATG (А имеет номер +1, вправо положительные значения, влево отрицательные: -1, -2 и т.д.)
- для гДНК начало – начало последовательности
- для белка начало – концевой метионин (соответствует ATG)
- 3. Указание на вид полиморфизма (>, del, ins)

Обозначение положения: длины экзонов неизвестны (интроны обозначаются IVS – intervening sequence)

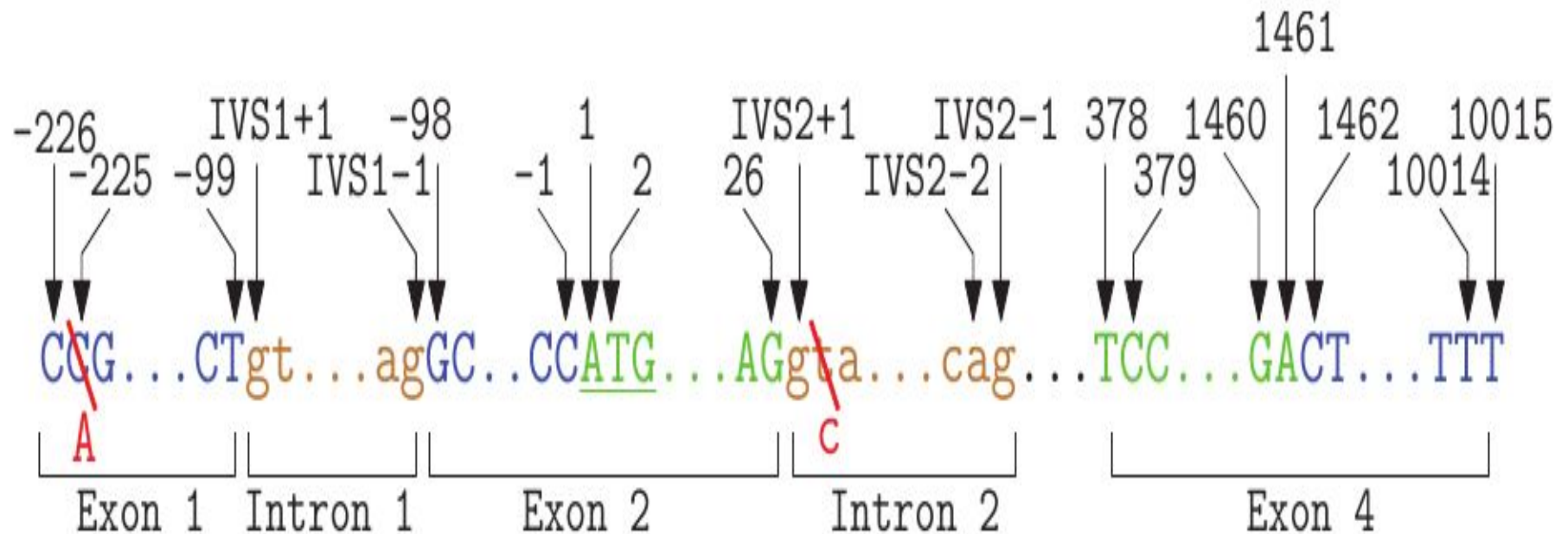


Figure 5 Illustration of the “intervening sequence” concept in human mutation nomenclature. The example annotations for parts of the gene MECP2 (NG_007107.1) are following the IVS concept. In this nomenclature variant the adenine of the start codon is used as start position. Variations located in intronic regions start with the abbreviation “IVS” followed by the number of the intron where the variation is located. The consecutive number determines the distance to the next intron/exon boundary. Using these recommendations the two SNPs are described as -225C→A and IVS2+2T→C.

Обозначение положения: длины экзонов известны

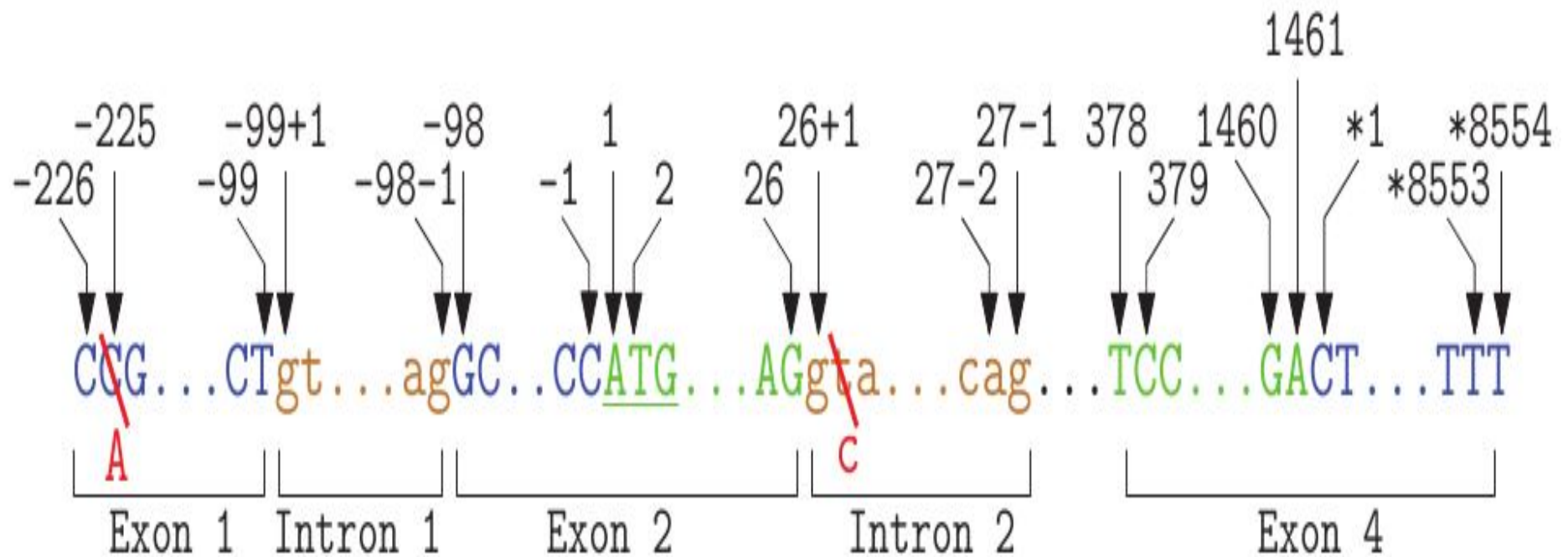


Figure 6 Illustration of the latest recommendations for human mutation nomenclature. This most recent nomenclature discards the IVS concept for intronic variations. Instead, the concept introduced earlier using two numbers is again recommended. Variations occurring in the 3'UTR are labeled with a preceding asterisk and numbering starts at the beginning of the UTR. Using these recommendations the two SNPs are described as c.-225C>A and c.26+2T>C.

Примеры записи SNPs

- **c.1777A>C**
- замена аденина на цитозин в положении 1777 кДНК
- **IVS33-2A>T**
- замена аденина на тимин во -2-м положении (слева) от интрона 33
- **c.26+2T>C**
- замена тимина на цитозин во втором положении (справа) от экзона, заканчивающегося 26 позицией
- **S549R**
- замена серина на аргинин в положении 549 белковой молекулы
- **W1282X**
- стоп-кодон вместо триптофана в положении 1282 белковой молекулы

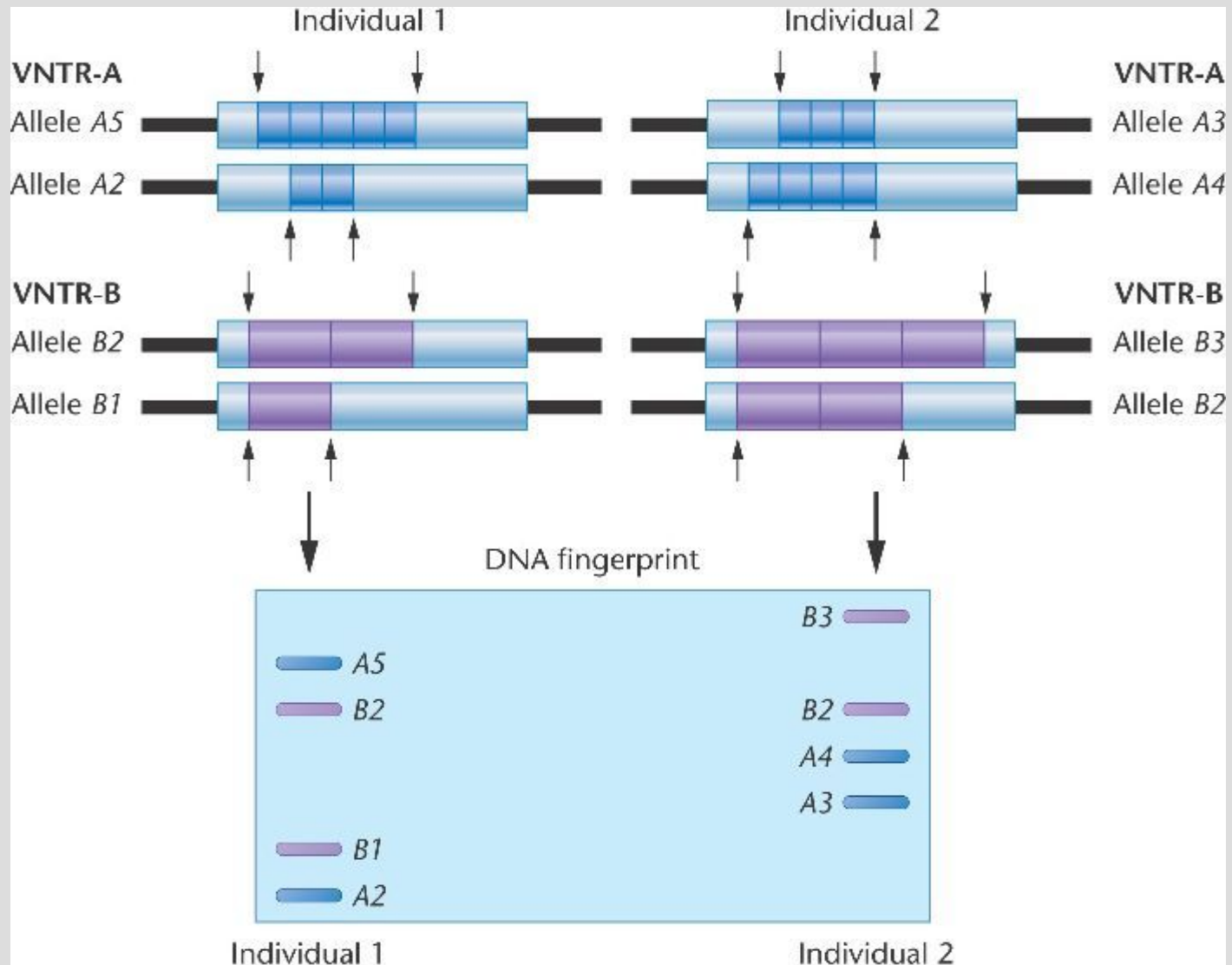
Использование уникального идентификатора для обозначения SNP

- База данных dbSNP (>55 организмов, >63 млн. SNP):
- каждый SNP имеет идентификатор rs (reserve SNP)
- **Напр.:** замена c.1805T>G имеет идентификатор rs5743618
- dbSNP связана с другими базами:
 - EntrezGene
 - GenBank
 - UniProt
 - MapMap
 - Ensembl
 - SNPedia

Простая инсерция/делеция

- Однократная вставка/удаление короткого участка ДНК. Может приводить к сдвигу рамки считывания.
- Номенклатура:
- **c.441delA**
 - делеция 441 нуклеотида в кДНК
- **c.1524-1527delCGTA**
 - делеция CGTA, занимающих положение с 1524 по 1527 в кДНК
- **c.1277-1278insTATC**
 - инсерция TATC между нуклеотидами 1277 и 1278 в кДНК
- **c.852del22**
 - делеция 22 нуклеотидов, начиная с 852 нуклеотида в кДНК
- **c.3320ins7**
 - инсерция 7 нуклеотидов после нуклеотида в положении 3320 в кДНК

Полиморфизм числа повторов



Методы определения полиморфизма генов

- MOLBIOL
- molbiol.edu.ru/review/index.html