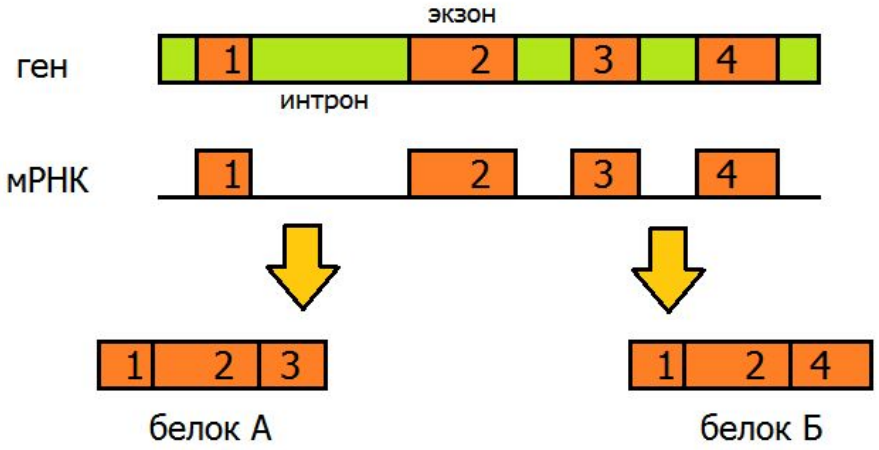
ГЛОССАРИЙ ИНФОРМАТИКИ

BIOINFORMATICS
INSTITUTE

А

Аллель — одна из двух (или более) альтернативных форм гена, каждая из которых характеризуется уникальной последовательностью нуклеотидов. Одна аллель для каждого гена наследуется от каждого из родителей.

Альтернативный сплайсинг — процесс, позволяющий одному гену производить несколько мРНК и, соответственно, белков благодаря разным наборам экзонов.



Аминокислоты — органические кислоты, содержащие минимум одну аминогруппу ($-NH_2$), то есть основную, и одну карбоксильную группу (-COOH), то есть кислотную. Среди многообразия аминокислот только 20 участвуют во внутреклеточном синтезе белков.

Амплификация — увеличение числа копий конкретного фрагмента ДНК; может быть в естественных условиях (in vivo) или в пробирке (in vitro).

Аннотация генома — процесс “привязывания” биологической информации к известным нуклеотидным последовательностям ДНК, предполагает поиск (или предсказание) генов, а так же регуляторных элементов и повторов. Выделяют структурную (идентификация геномных элементов: рамок считывания, кодирующих участков, повторов, мотивов и др.) и функциональную (определение биохимических или биологических функций белков, уровней экспрессии генов, а также механизмов ее регуляции и др.) аннотацию.

Антиген — вещество (обычно белки, реже полисахариды), вызывающее у животных иммунный ответ (образование антител).

Антитело — белок (иммуноглобулин), образуемый иммунной системой организма животных в ответ на введение антигена и способный вступать с ним в специфическое взаимодействие.

Апоптоз — запрограммированная смерть клетки. Происходит с минимальным вредом для окружающих клеток (в отличие от некроза) и часто является закономерным этапом жизненного цикла клеток (например, происходит при исчезновении перепонок между пальцами во внутриутробном развитии человека).

Аутосома — любая неполовая хромосома. У человека имеется 22 пары аутосом и одна пара половых хромосом.

Б

База — пара двух оснований нуклеотидов на комплементарных цепочках нуклеиновых кислот (противоположных ДНК или одинаковых РНК), соединённых с помощью водородных связей. Единица измерения длины последовательностей ДНК или РНК (например, длина генома бактерий варьируется от 139 до 13000 килобаз). В английском обозначении: **bp** (base pair - пара оснований).

Библиотека (геномная библиотека) — коллекция клонов, сделанных из набора случайно сгенерированных перекрывающихся фрагментов ДНК, которые представляют весь геном организма.

Бисульфитное секвенирование — общее название группы методов, направленных на изучение паттерна метилирования ДНК. Предполагает обработку ДНК бисульфитом с последующим секвенированием. Обработка бисульфитом приводит к конвертированию всех цитозинов (C), не защищенных метильной группой, в урацил. Сравнивая полученную после обработки последовательность с исходной, можно выявить метилированные участки ДНК.

Болезни, сцепленные с полом — болезни, обусловленые дефектом генов, локализованных в X- или Y-хромосомах.

В

Вектор — молекула нуклеиновой кислоты (ДНК или РНК), служащая инструментом для введения генетической информации в клетку.

Вирусы — инфекционные агенты неклеточной природы, способные в процессе реализации генетической информации, закодированной в их геноме, перестроить метаболизм клетки хозяина, направив его в сторону синтеза вирусных частиц. Вирусы состоят из белковой оболочки (капсида), внутри которой находится ДНК или РНК.

Выравнивание (alignment) — способ сопоставления последовательностей ДНК, РНК или белков для выявления областей сходства, которые могут быть следствием функциональных, структурных, или эволюционных отношений между последовательностями. Выровненные последовательности нуклеотидов или аминокислот обычно представлены в виде строк в матрице.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                     ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Выравнивание глобальное — выравнивание по всей длине всех исходных последовательностей. Наиболее распространены следующие программы: ClustalW, MUSCLE, Kalign, MAFFT и др.

Выравнивание локальное — поиск сравнительно коротких сходных областей в пределах длинных последовательностей, в целом далеких друг от друга. Наиболее распространены следующие программы: BLAST, DIALIGN, MAFFT и др. Отдельная сложная задача — полногеномное выравнивание, обычно при этом используются алгоритмы локального выравнивания (например, BlastZ, LastZ, MUMMER, MAUVE).

Г

Гамета — зрелая половая клетка с гаплоидным набором хромосом. У человека их 23.

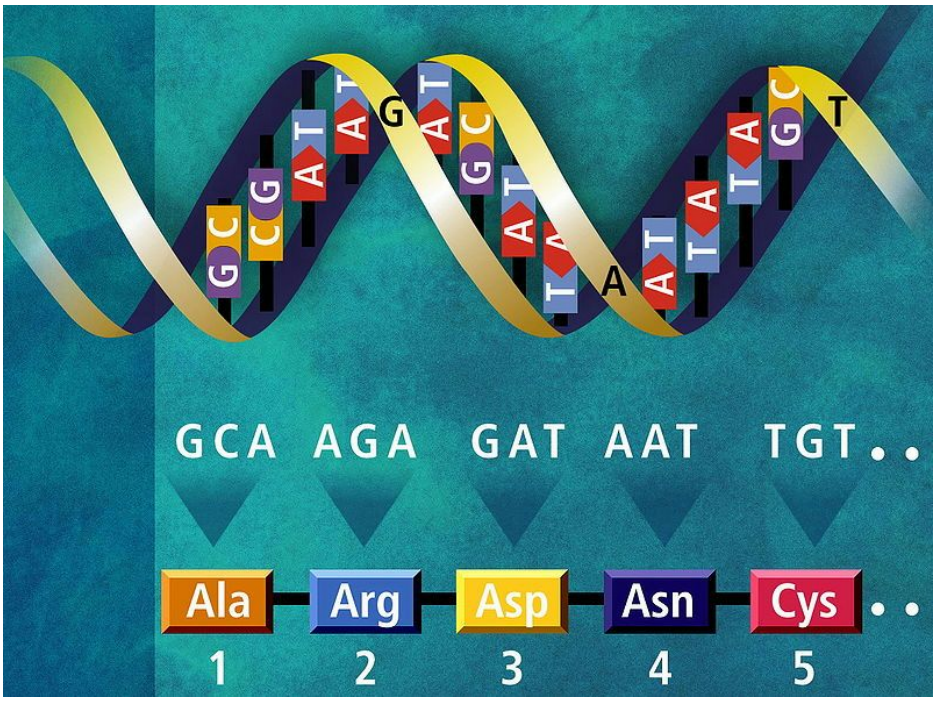
Гаплоид — клетка, ткань или организм с одинарным набором генов или хромосом (например, сперматозоиды и яйцеклетки).

Гаплотип (сокр. от «гаплоидный генотип») — совокупность аллелей на локусах одной хромосомы, обычно наследуемых вместе. Если же при кроссинговере комбинация аллелей меняется (что происходит очень редко), говорят о возникновении нового гаплотипа. Гаплотип может быть как у одного локуса, так и у целого генома. Генотип определенных генов диплоидной особи состоит из двух гаплотипов, расположенных на двух хромосомах, полученных от матери и отца соответственно.

Ген — последовательность нуклеотидов в ДНК, кодирующая определенную РНК, которая обусловливает какую-либо функцию в организме или регулирует транскрипцию другого гена.

Генетическая карта — схема расположения структурных генов и регуляторных элементов в хромосоме.

Генетический код — способ кодирования аминокислотной последовательности белков при помощи последовательности нуклеотидов. Реализация генетической информации в живых клетках (то есть синтез белка, кодируемого геном) осуществляется при помощи двух матричных процессов: транскрипции (то есть синтеза мРНК на матрице ДНК) и трансляции генетического кода в аминокислотную последовательность (синтез полипептидной цепи на мРНК). Для кодирования 20 аминокислот, а также сигнала «стоп», означающего конец белковой последовательности, достаточно трёх последовательных нуклеотидов. Набор из трёх нуклеотидов называется триплетом. Каждую аминокислоту кодирует определенный триплет (один или несклько).



Генная терапия — процедура, направленная на замену, манипулирование или дополнение нефункциональных или неправильно функционирующих генов здоровыми генами.

Генная инженерия (генетическая инженерия) - совокупность приёмов, методов и технологий получения рекомбинантных РНК и ДНК (искусственно созданных человеком) , выделения генов из организма (клеток), осуществления манипуляций с генами и введения их в другие организмы.

Геном — общая генетическая информация, содержащаяся в генах организма, или генетический состав клетки. Термин «геном» иногда употребляется для обозначения гаплоидного набора хромосом.

Генотип — (1) вся генетическая информация организма; (2) генетическая характеристика организма по одному или нескольким изучаемым локусам.

Гетерозигота — клетка (или организм), содержащая две различных аллели в конкретном локусе гомологичных хромосом.

Гибридизация in situ — использование ДНК или РНК зондов для обнаружения присутствия комплементарной последовательности ДНК в клетках.

Гибридизация ДНК — процесс образования двуцепочечной ДНК или дуплексов ДНК-РНК в результате взаимодействия комплементарных нуклеотидов по правилу G-C, A-T(U).

Гистон - обширный класс ядерных белков, выполняющих две основные функции: участие в упаковке нитей ДНК в ядре и в эпигенетической регуляции таких ядерных процессов, как транскрипция, репликация и репарация (см. ниже).

Гомозигота — клетка (или организм), содержащая две одинаковые аллели в конкретном локусе гомологичных хромосом.

Гомологичные хромосомы — хромосомы, одинаковые по набору составляющих их генов.

Группа сцепления — все гены, локализованные в одной хромосоме. У человека 25 групп сцепления: 22 аутосомы, две разных половых хромосомы и митохондриальная хромосома.

Д

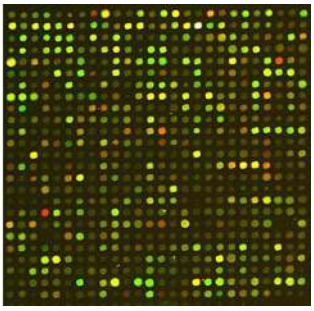
Делеция — тип хромосомной мутации, при которой утрачивается участок хромосомы.

Денатурация — нарушение пространственной структуры молекулы в результате разрыва внутри- или межмолекулярных нековалентных связей.

Диплоидный набор — полный набор генетического материала, состоящего из парных хромосом, по одному от каждого родительского набора. Большинство клеток животных, кроме половых клеток, имеют диплоидный набор хромосом. Диплоидный набор хромосом человека содержит 46 хромосом.

ДНК (дезоксирибонуклеиновая кислота) — молекула, кодирующая генетическую информацию. ДНК является двухцепочечной молекулой, которая

удерживается слабыми связями между парами оснований нуклеотидов. Четыре нуклеотида ДНК содержат основания: **аденин (A), гуанин (G), цитозин (C), и тимин (T)**. В природе пары оснований образуются только между A и T или между G и C.



ДНК-микрочип — технология, используемая в молекулярной биологии и медицине. Он состоит из тысяч дезоксиолигонуклеотидов (зондов или проб), сгруппированных в виде микроскопических точек и закреплённых на твёрдой подложке. Каждая точка содержит определённую нуклеотидную последова- тельность, которая может быть коротким участком гена или других функциональных элементов ДНК и используется для гибридизации с кДНК или мРНК. Гибридизация зонда и мишени регистрируется и количественно характеризуется при помощи флюоресценции или хемилюминесценции, что позволяет определять относительное количество нуклеиновой кислоты с заданной последовательностью в образце.

ДНК-полимераза — фермент, осуществляющий матричный синтез ДНК.

Доминирование — тип наследования, при котором признак, кодируемый одним аллелем, проявляется, а кодируемый другим, парным — нет.

Дрейф генов — изменение частот генов в ряду поколений, обусловленное случайными событиями митоза, оплодотворения и размножения.

Дупликация — тип мутации, при которой удвоен какой-либо участок ДНК.

З

Зонд генетический — короткий отрезок ДНК или РНК известной структуры или функции, меченный каким-либо радиоактивным или флуоресцентным соединением.

И

Импринтинг — эпигенетический процесс, при котором экспрессия определённых генов осуществляется в зависимости от того, от какого родителя получены аллели. Наследование признаков, определяемых импринтируемыми генами, происходит не по Менделю.

Интерфероны — белки, синтезируемые клетками позвоночных в ответ на вирусную инфекцию, и подавляющие их развитие.

Интрон — некодирующий участок гена, который транскрибируется, а затем удаляется из предшественника мРНК при сплайсинге.

К

Картирование генов — определение относительного положения генов в молекуле ДНК (хромосоме или плазмиде) и расстояния в единицах сцепления или физических единицах между ними.

кДНК (комплементарная ДНК) — ДНК, синтезируемая по матрице РНК с помощью обратной транскриптазы.

Кишечная палочка (лат. ***Escherichia coli, E. coli***) — распространенная бактерия, которая интенсивно используется в исследованиях, в частности, из-за небольшого размера генома и простоты выращивания в лаборатории.

Клетка — элементарная единица строения и жизнедеятельности всех организмов (кроме вирусов), обладающая собственным обменом веществ, способная к самостоятельному существованию, самовоспроизведению и развитию.

Клон — группа генетически идентичных клеток, возникших неполовым путем от общего предка.

Кодон — тройка расположенных подряд нуклеотидных остатков в ДНК или РНК, кодирующая определенную аминокислоту или являющаяся сигналом окончания трансляции (стоп-кодон).

Комплементарность — свойство азотистых оснований образовывать с помощью водородных связей парные комплексы аденин — тимин (или урацил) и гуанин — цитозин при взаимодействии цепей нуклеиновых кислот.

Контиг — набор перекрывающихся последовательностей фрагментов ДНК (ридов), полученных из одного биологического источника (организма, ткани, клетки) в результате секвенирования.

Кроссинговер (от англ. *crossing over* — пересечение) — процесс обмена участками гомологичных хромосом во время мейоза.

Л

Линия клеток — генетически однородные клетки животных или растений, которые можно выращивать in vitro в течение неограниченно долгого времени.

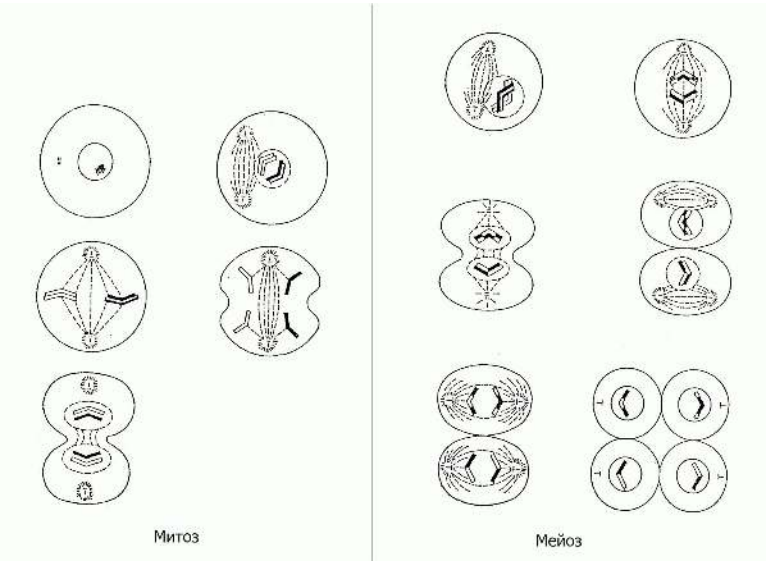
Локус — местоположение определённого гена или другой последовательности ДНК на генетической или цитологической карте хромосомы.

М

Метаболизм — совокупность химических реакций, обеспечивающих существование и воспроизведение клетки.

Мейоз — деление эукариотической клетки, в результате которого из диплоидных клеток образуются гаплоидные, которые, как правило, являются гаметами.

Митоз — непрямое деление клетки, наиболее распространенный способ репродукции эукариотических клеток. Биологическое значение митоза состоит в строго одинаковом распределении хромосом между дочерними ядрами, что обеспечивает образование генетически идентичных дочерних клеток и сохраняет преемственность в ряду клеточных поколений.



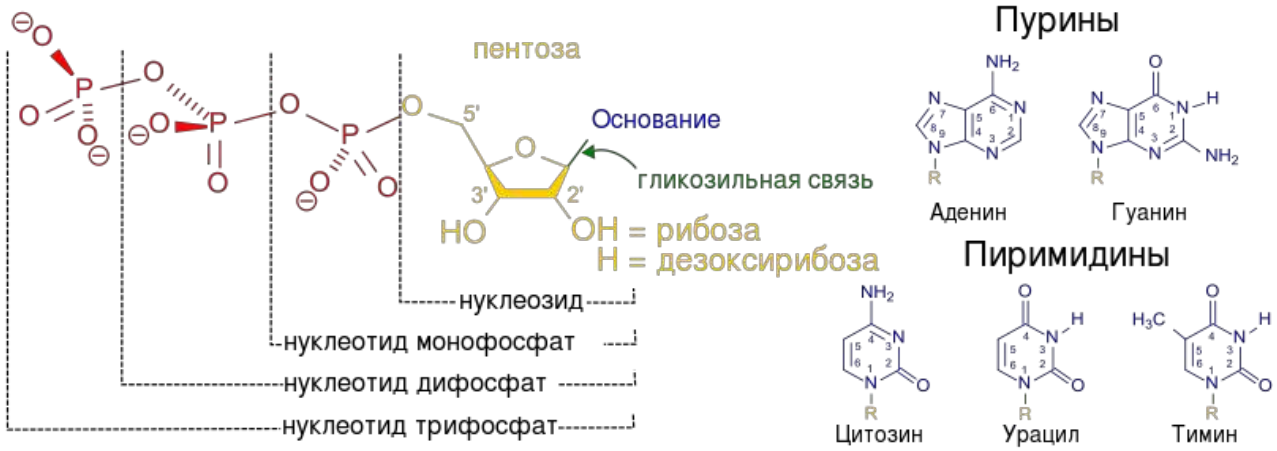
Мутагены — физические, химические или биологические агенты, увеличивающие частоту возникновения мугаций.

Мутация — стойкое (т.е. которое может быть унаследовано потомками данной клетки или организма) преобразование генотипа, происходящее под влиянием внешней или внутренней среды.

Н

Нуклеотид (нуклеотидное основание) - сложные эфиры нуклеозидов и фосфорных кислот. Нуклеозиды, в свою очередь, являются N-гликозидами, содержащими гетероциклический фрагмент, связанный через атом азота с С-1 атомом остатка сахара. Названия нуклеотидов представляют собой аббревиатуры в виде стандартных трёх- или четырёхбуквенных кодов. Если аббревиатура начинается со строчной буквы «д» (англ. *d*), значит подразумевается дезоксирибонуклеотид; отсутствие буквы «д» означает рибонуклеотид. Если аббревиатура начинается со строчной буквы «ц» (англ. *c*), значит речь идёт о циклической форме нуклеотида (например, цАМФ). Первая прописная буква аббревиатуры указывает на конкретное азотистое основание или группу возможных нуклеиновых оснований, вторая буква — на количество остатков фосфорной кислоты в структуре (М — моно-, Д — ди-, Т — три-), а третья прописная буква — всегда буква Ф («-фосфат»; англ. *P*). Латинские и русские коды для нуклеиновых оснований:

- A — А: Аденин;
- G — Г: Гуанин;
- C — Ц: Цитозин;
- T — Т: Тимин (5-метилурацил), встречается в РНК, занимает место урацила в ДНК;
- U — У: Урацил, встречается в ДНК у бактериофагов, занимает место тимина в РНК.



О

Обратная транскриптаза — фермент, катализирующий реакцию синтеза ДНК по РНКовой матрице.

Олигонуклеотид — цепь, состоящая из нескольких (обычно от 2 до 20) нуклеотидных остатков.

Омы (омики) -- совокупность данных целого организма. Самыми распростарненными являются геном (совокупность всей генетической информации), транскриптом (совокупность всех транскриптов) и протеом(совокупность всех белков организма).

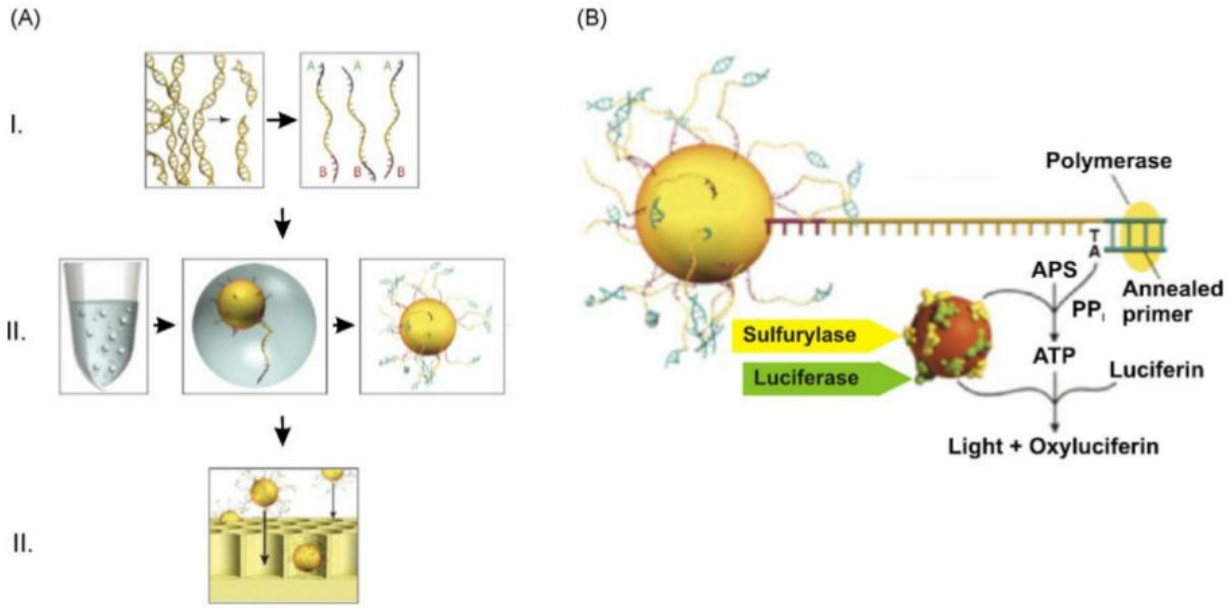
Омиксные данные -- данные, полученные при анализе омов (омиков)

Онкогены — гены, чьи продукты обладают способностью трансформировать эукариотические клетки так, что они приобретают свойства опухолевых клеток.

Оперон — совокупность совместно транскрибируемых генов, обычно контролирующих родственные биохимические функции.

П

Пиросеквенирование — метод секвенирования путем синтеза, основанный на детекции при помощи хемолюминесценции пирофосфатов, освобождающихся при присоединении к растущей цепи ДНК нуклеотидов: **(A) I** — приготовление ДНК-библиотек (фрагментация геномной ДНК и пришивание адептеров); **II** — эмульсионная ПЦР, в результате которой получают сферы с прикрепленными к ним клонами одной молекулы ДНК; **III** — помещение сфер на специальную подложку с ячейками, так что в каждой ячейке оказывается одна сфера. **(B)** Процесс секвенирования: в каждом цикле в ячейку добавляется только один нуклеотид, если он прикрепляется ДНК-полимеразой к растущей цепи ДНК, происходит освобождение пирофосфата, вступающего в реакцию с ферментной системой, состоящей из АТФ-сульфурилазы и люциферазы, в результате которой высвобождается детектируемый видимый свет. Таким образом, свет образуется в тот момент, когда добавленный нуклеотид соответствует первому неспаренному основанию матричной молекулы ДНК. Основная сложность — прочтение гомополимерных последовательностей, т.к. с увеличением числа присоединяемых одновременно нуклеотидов интенсивность свечения возрастает неравномерно.



Полупроводниковое секвенирование — метод секвенирования путем синтеза, с работающий, как и пиросеквенирование, с помощью детекции присоединения каждого следующего нуклеотида. Основан на идее, что при присоединении нуклеотида ДНК-полимеразой к растущей цепочке ДНК высвобождается ион водорода, изменяющий pH раствора.

Покрытие нуклеотида — сколько раз нуклеотид считывается в процессе секвенирования. Отсюда можно посчитать, например, среднее покрытие генома.

Плазмида — кольцевая или линейная молекула ДНК, реплицирующаяся автономно от клеточной хромосомы.

Полимеразы — ферменты, ведущие матричный синтез нуклеиновых кислот.

Полимеразная цепная реакция (ПЦР) — экспериментальный метод молекулярной биологии, способ значительного увеличения малых концентраций определённых фрагментов ДНК в биологическом материале (пробе).

Полиморфизм — разница в последовательности ДНК среди отдельных лиц, групп или групп населения (например, гены голубого цвета глаз против карего). Иными словами - одновременное существование в популяции разных аллелей одного гена.

Полипептид — полимер, состоящий из аминокислотных остатков, связанных пептидными связями.

Праймер — короткий фрагмент нуклеиновой кислоты, комплементарный ДНК- или РНК-мишени. Служит затравкой для синтеза комплементарной цепи с помощью ДНК-полимеразы (при репликации ДНК).

Прокариоты — организмы, у которых нет клеточного ядра.

Р

Рак — заболевания, при которых клетки начинают бесконтрольно делиться и расти.

Референс (референсная сборка, референсный геном; от англ. ***reference*** — начало отсчёта, эталон, образец) — общий репрезентативный пример генетического кода того или иного организма, хранящийся в цифровом виде. Например, для генома человека на данный момент последняя версия референса — GRCh38/hg38 (Genome Reference Consortium human genome 38), вышедшая в декабре 2013 года. С каждой следующей версией уточняется последовательность генома (главным образом, в таких сложных для сборки участках, как центромеры и теломеры). Тем не менее, есть некоторые трудности, связанные с переходом между версиями, т.к. геномные координаты для них могут не совпадать.

Рекомбинантный белок — белок, полученный в результате экспрессии с рекомбинантной молекулы ДНК, часто получаемый в кишечной палочке.

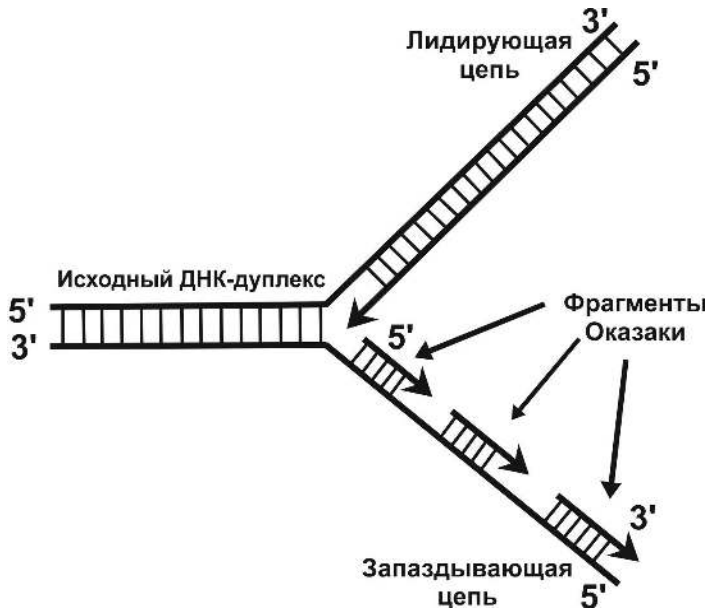
Рекомбинантная молекула ДНК — молекула ДНК, в которой содержатся различные гены, соединенные методами генетической инженерии.

Рекомбинация гомологическая — обмен генетическим материалом между двумя гомологичными молекулами ДНК.

Рекомбинация сайт-специфическая — объединение путем разрыва и слияния двух молекул ДНК или участков одной молекулы, происходящее по определенным сайтам.

Репарация ДНК — исправление повреждений молекулы ДНК, восстанавливающее её первоначальную структуру.

Репликация — процесс удвоения молекул ДНК или геномных вирусных РНК.



Рестрикция - процесс разрезания цепочки ДНК, осуществляемый рестриктазой.

Рестриктаза (эндонуклеаза рестрикции) - группа ферментов, которые расщепляют нуклеиновые кислоты. При этом каждая рестриктаза узнаёт определённый участок ДНК длиной от четырёх пар нуклеотидов и расщепляет нуклеотидную цепь внутри участка узнавания или вне его. Активно используется в биотехнологии для модификации бактериальных плазмид.

Ретровирусы — семейство РНК-содержащих вирусов, заражающих преимущественно позвоночных. Наиболее известный и активно изучаемый представитель — вирус иммунодефицита человека (ВИЧ).

Рецессивное наследование — тип наследования признака или болезни, при котором мутантный аллель должен быть унаследован от обоих родителей для проявления признака.

Риды (от англ. **read**) — результат работы секвенатора: строки, содержащие прочитанные последовательности нуклеотидов.

РНК (рибонуклеиновая кислота) - одна из трёх основных макромолекул (две другие — ДНК и белки), которые содержатся в клетках всех живых организмов, которая состоит из нуклеотидов. Есть несколько типов РНК: матричные РНК (мРНК) образуются в ходе транскрипции, осуществляемого РНК-полимеразами. Затем они принимают участие в трансляции. Транспортные РНК (тРНК) служат для узнавания кодонов и доставки

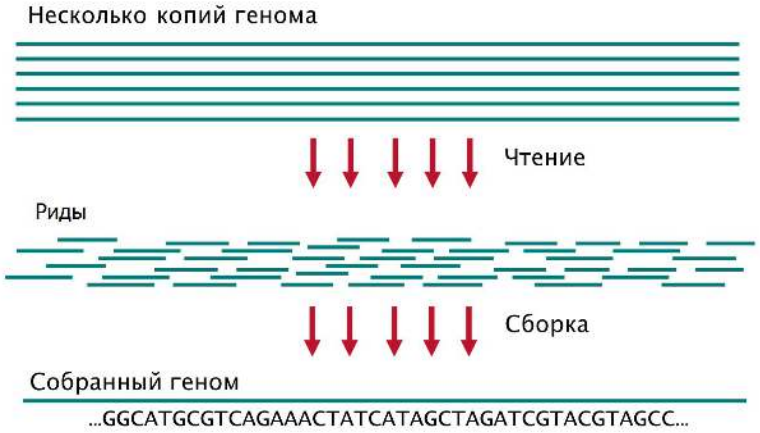
соответствующих аминокислот к месту синтеза белка, а рибосомные РНК (рРНК) служат структурной и каталитической основой рибосом. Малые ядерные РНК (мяРНК) принимают участие в сплайсинге эукариотических матричных РНК.

Рибосома - немембранный органоид живой клетки, служащий для биосинтеза белка из аминокислот по заданной матрице на основе генетической информации, предоставляемой матричной РНК (мРНК). Состоит из большой и малой субъединицы, которые различаются по размеру.

С

Сайт — участок молекулы ДНК, белка и т. п.

Сборка (assembly) — процесс объединения большого количества коротких фрагментов ДНК или РНК (ридов), полученных в результате секвенирования, в одну или несколько длинных последовательностей (контигов и скаффолдов) с целью восстановления исходных последовательностей.

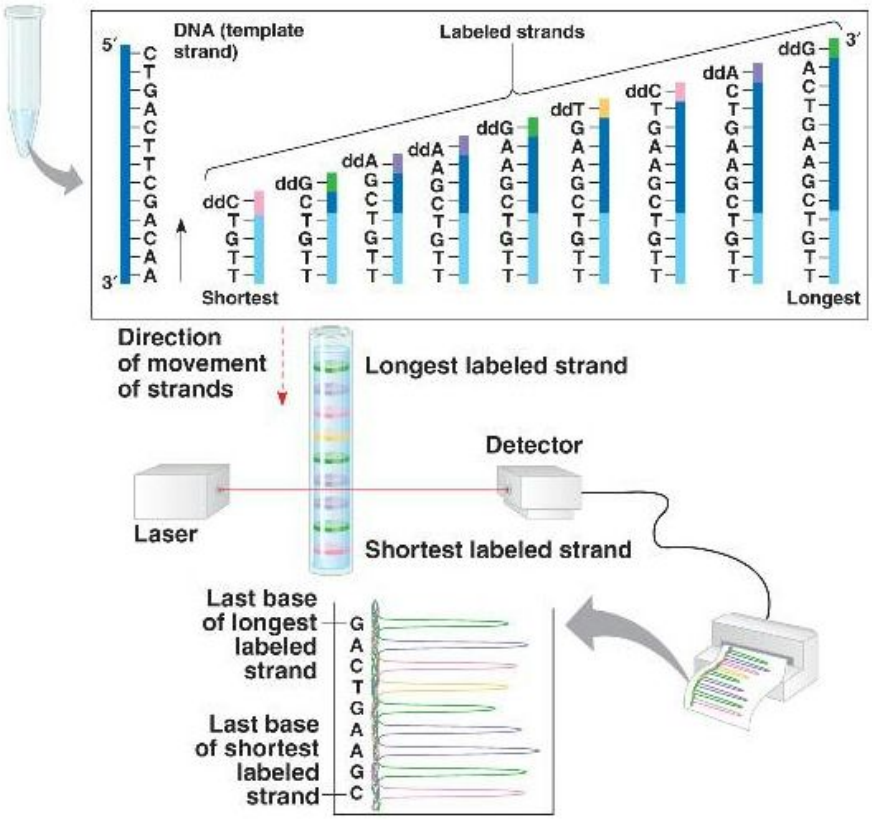


Секвенирование — определение аминокислотной (белки) или нуклеотидной (ДНК / РНК) последовательности.

Секвенирование РНК (RNA-seq) — определение первичной структуры молекулы РНК. Под этим может подразумеваться как секвенирование мРНК, так и определение последовательности некодирующих РНК. Современное полногеномное секвенирование РНК основано на прямом секвенировании фрагментов кДНК

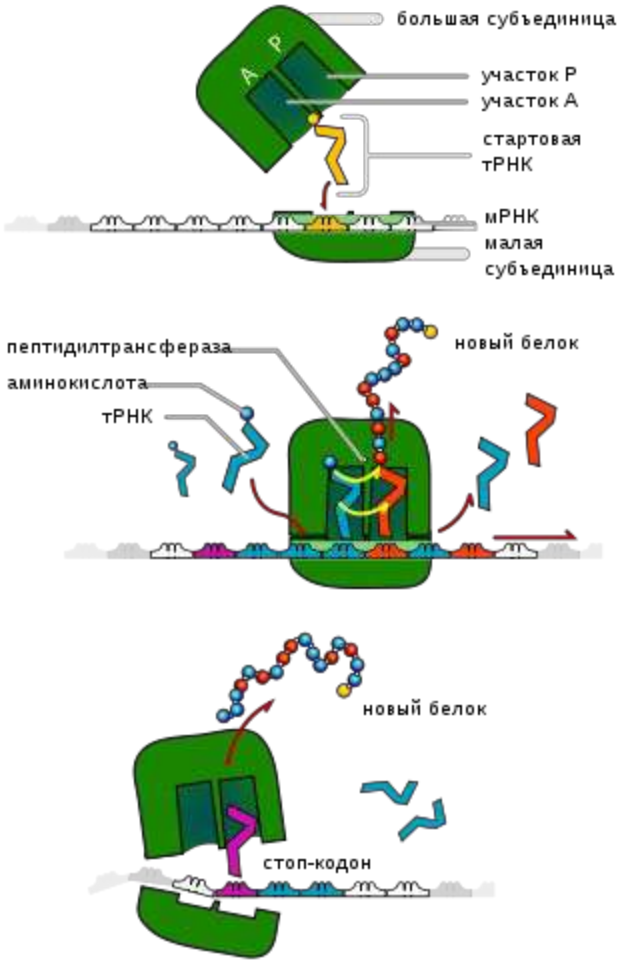
Секвенирование по Сэнгеру — метод секвенирования первого поколения, известен как “метод терминаторов”, “дидезокси-метод” или “метод обрыва цепи”, был предложен Ф. Сэнгером в 1977 году (Нобелевская премия по химии 1980 г.). Основан на использовании

дидезоксинуклеотидов для терминации цепи в ходе ПЦР. При этом в результате реакции получается набор копий одного фрагмента, отличающихся длиной на один нуклеотид, причем крайний нуклеотид на каждом из фрагментов оказывается флуоресцентно (первоначально — радиокативно) меченым. В современном варианте каждый сорт нуклеотидов метят своей флуоресцентной меткой, детекция флуоресеценции осуществляется в ходе капиллярного электрофореза.



Секвенирование следующего поколения (NGS, Next Generation Sequensing) — технология секвенирования (может быть основана на разных методах), позволяющая одновременно “прочитать” сразу большое число фрагментов ДНК. При этом за один рабочий цикл прибора происходит получение от сотен мегабаз до гигабаз нуклеотидных последовательностей.

Синтез белка - сложный многостадийный процесс синтеза полипептидной цепи из аминокислот, происходящий на рибосомах с участием молекул мРНК и тРНК.



Системная биология — междисциплинарное научное направление, образовавшееся на стыке биологии и программирования, ориентированное на изучение сложных взаимодействий в живых системах.

Синтения - это структурное сходство групп сцепления генов у организмов разных биологических видов. В частности, в геномах человека и мыши известно несколько десятков синтеничных групп генов. Наличие феномена синтении позволяет суживать круг поиска места локализации исследуемого гена на хромосомах, ограничивая его областью известных генов, принадлежащих к конкретной синтеничной группе.(см. также синтенный блок).

Скафолд (англ. scaffold) — серия контигов, которые находятся в правильном порядке по отношению к геному, но не обязательно связаны в одну непрерывную последовательность.

«Снип» (англ. Single nucleotide polymorphism, **SNP**) — отличия последовательности ДНК размером в один нуклеотид (A, T, G или C) в геноме (или в другой сравниваемой последовательности) представителей одного вида или между гомологичными участками гомологичных хромосом.

Соматические клетки — клетки тканей многоклеточных организмов, не относящиеся к половым.

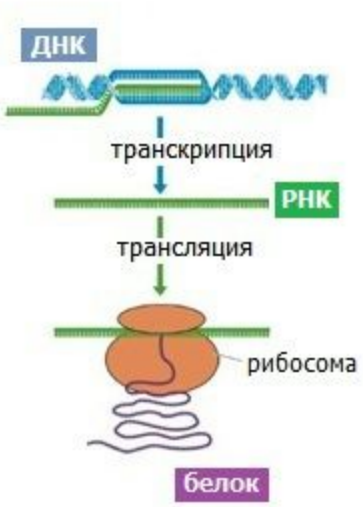
Сплайсинг — процесс формирования зрелой мРНК или функционального белка путем удаления внутренних частей молекул — интронов РНК.

Стволовые клетки — недифференцированные (незрелые) клетки, имеющиеся во всех многоклеточных организмах. Стволовые клетки способны самообновляться, образуя новые стволовые клетки, делиться посредством митоза и дифференцироваться в специализированные клетки, т.е. превращаться в клетки различных органов и тканей.

Т

Транскрипция — синтез РНК на ДНК-матрице; осуществляется РНК-полимеразой.

Транскрипт — продукт транскрипции, т. е. РНК, синтезированная на данном участке ДНК как на матрице и комплементарная одной из его нитей.



Транскриптаза обратная — фермент, синтезирующий по РНК как по матрице комплементарную ей однонитевую ДНК.

Транскриптом — совокупность всех транскриптов, синтезируемых одной клеткой или группой клеток, включая мРНК и некодирующие РНК. В отличие от генома, который, как правило, одинаков для всех клеток одной линии, транскриптом может сильно меняться в зависимости от ткани, этапа развития и условий окружающей среды.

Трансляция — процесс синтеза полипептида, определяемый матричной РНК.

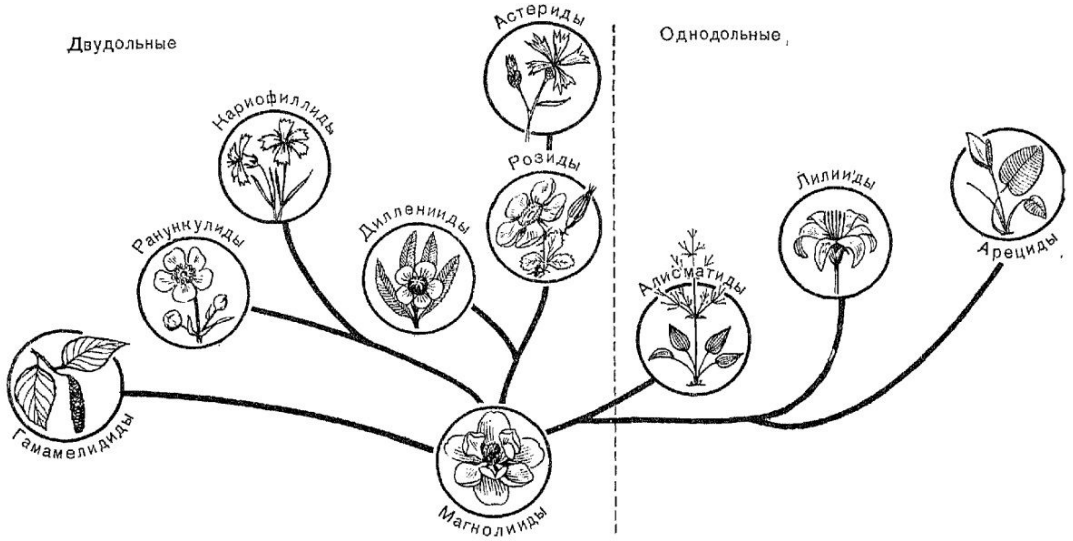
Транспозон — генетический элемент, реплицируемый в составе репликона и способный к самостоятельным перемещениям (транспозиции) и интеграции в разные участки хромосомной или внехромосомной ДНК.

Ф

Фенотип — внешнее проявление свойств организма, зависящих от его генотипа и факторов окружающей среды.

Фермент - обычно белковые молекулы или молекулы РНК (рибозимы) или их комплексы, ускоряющие (катализирующие) химические реакции в живых системах.

Филогенетическое дерево (эволюционное дерево) — дерево, отражающее эволюционные взаимосвязи между различными видами или другими сущностями, имеющими общего предка.



Финишированный геном — полная последовательность ДНК всех хромосом без разрывов, с высоким качеством и низким процентом ошибок.

Форматы файлов, используемые в биоинформатике:

bcf — Binary VCF, представляет собой сжатый vcf файл.

bam — бинарная (сжатая) форма sam-формата, позволяющая компактно хранить информацию о выравнивании последовательностей.

bed — текстовый файл, предназначенный для хранения информации об аннотации генома, включает три обязательные поля (хромосома, координата начала фрагмента и координата конца) и девять дополнительных, которые могут содержать различную информацию о заданном участке ДНК.

csv — Comma-Separated Values, представляет собой текстовый формат для представления табличных данных, где столбцы таблицы разделены запятыми.

fasta — текстовый формат хранения данных о последовательностях, как нуклеотидных, так и аминокислотных, представленных в виде однобуквенного кода. Каждая последовательность в файле начинается с описания в одну строку (выделенного знаком “>”), затем идет произвольное число строк последовательности.

fastq — текстовый формат, позволяющий хранить не только нуклеотидную последовательность, но и данные о phred quality score для каждого из нуклеотидов. Состоит из повторяющихся четверок строк, первая из них начинается с символа '@', за которым следует идентификатор последовательности и комментарий, затем - строка с собственно нуклеотидной последовательностью, затем - строка, начинающаяся с символа '+', за которым может следовать комментарий, и, наконец, строка, содержащая значения phred quality. Длина этой строки должна соответствовать длине строки с нуклеотидной последовательностью.

gff/gtf — General Feature Format/General Transfer Format, текстовый формат, используемый для описания генов, повторов и других характеристик ДНК, РНК и белков, содержит девять обязательных полей. В целом, похож на bed формат, но отличается порядком полей и более жесткой структурой.

newick — текстовый формат, позволяющий хранить филогенетические (и не только) деревья с длинами ветвей, используя комбинации скобок и запятых.

sam — Sequence Alignment/Map формат, предназначенный для хранения больших выравниваний последовательностей.

vcf — Variant Call Format, представляет собой текстовый файл, содержащий строки с метаинформацией (начинающиеся с ‘##’), строку заголовка (начинается с ‘#’) и строки с данными, в каждой из которых хранится информация о позиции в геноме (обязательно — номер хромосомы и координата на ней, референсная последовательность, последовательность варианта). Кроме того, в каждой строке возможно хранение дополнительной информации (например, о качестве варианта, об образце или генотипе). Как правило, в данном формате хранятся результаты SNP-calling.

Х

Хромосомы — структуры клетки, в которых сосредоточена бо́льшая часть наследственной информации и которые предназначены для её хранения, реализации и передачи. Иными словами, форма упаковки наследственной информации. Основной составляющей каждой хромосомы является ДНК. У прокариот хромосомная ДНК является кольцевой, и весь геном находится обычно на одной хромосоме. Геномы эукариот состоят из нескольких хромосом, каждая из которых ассоциируется с различными видами белков.

Ш

Штамм — клональная по происхождению культура микроорганизмов, изолированнная в определенное время в определенном месте и сходная по морфологическим, биохимическим, генетическим и другим признакам.

Э

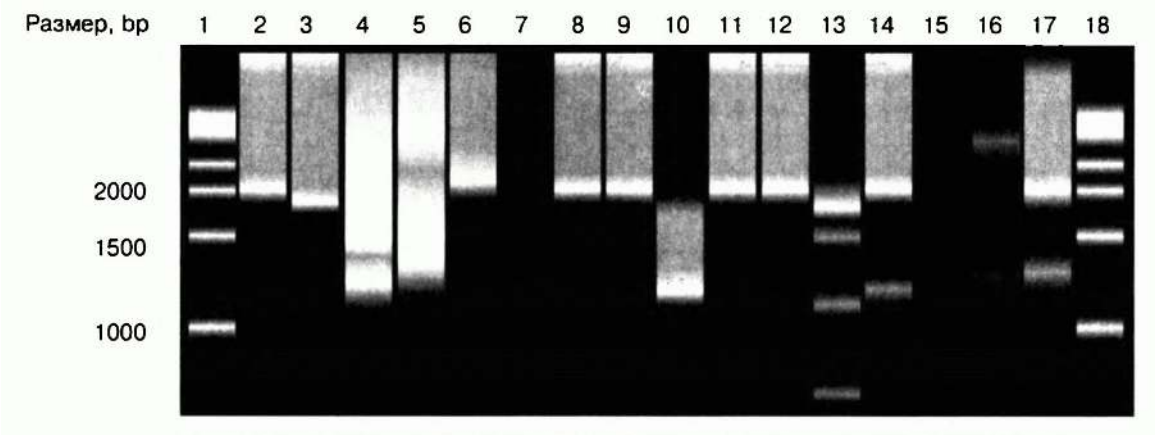
Экзон — сохраняющаяся при сплайсинге часть интронированного гена и несущая генетическую информацию.

Экзомное секвенирование -- стратегия секвенирования всех белок-кодирующих генов в геноме (т. е. экзома), предполагающая выбор только тех участков ДНК, которые кодируют белки (экзонов) и их последующее секвенирование

Эпигенетика (греч. ἐπί — над, выше) — изучение закономерностей эпигенетического наследования — изменения экспрессии генов или фенотипа клетки, вызванных механизмами, не затрагивающими последовательности ДНК.

Экспрессия гена — процесс реализации информации, закодированной в гене. Состоит из двух основных стадий .— транскрипции и трансляции.

Электрофорез — метод для разделения больших молекул (например, белки или фрагменты ДНК) из смеси подобных молекул. Электрический ток пропускают через среду, содержащую смесь, и каждый вид молекул проходит через среду с различной скоростью в зависимости от его электрического заряда и размера.



Эукариоты — организмы, клетки которых содержат ядра. Эукариоты включают в себя все организмы, кроме вирусов, бактерий и архей.

B

BLAST (Basic Local Alignment Search Tool) — алгоритм (и одноименный инструмент), используемый для поиска последовательностей в базах данных, оптимальных для локального выравнивания с запрашиваемой строкой. Часто применяется для поиска гомологичных (похожих) генов.

C

ChIP-seq -- метод, используемый для анализа ДНК-белковых взаимодействий и поиска потенциальных сайтов связывания белка

CNV (copy-number variation) — вариация числа копий — вид генетического полиморфизма, к которому относят различия индивидуальных геномов по числу копий хромосомных сегментов размером от 1 тыс. до нескольких млн. пар оснований. CNV возникают в результате несбалансированных хромосомных перестроек, таких как делеции и дупликации.

CRISPR (*Clustered Regularly Interspaced Short Palindromic Repeats*) **--** прямые повторы и разделяющие их уникальные последовательности в ДНК бактерий и архей, которые совместно с ассоциированными генами обеспечивают защиту клетки от чужеродных генетических элементов (бактериофагов, плазмид). В настоящее время разработаны

способы высокоизбирательного активирования и ингибирования генов, базирующиеся на этой системе.

G

Gap — (1) разрыв или пропуск в последовательности ДНК, (2) участок, либо отсутствующий в исследуемой последовательности, либо неизвестный.

GC-состав — процентный состав суммы всех гуанинов (G) и цитозинов (C) по отношению к длине исследуемого участка нуклеиновых кислот.

Genome-wide association study (GWAS) — исследование связи генотипа с различными фенотипическими признаками (в первую очередь, с наследственными заболеваниями) в масштабе всего генома. При этом сравнивают геномы людей, подверженных болезни (cases), с геномами здоровых людей (controls). В результате выявляют отличия, статистически значимо связанные с развитием заболевания.

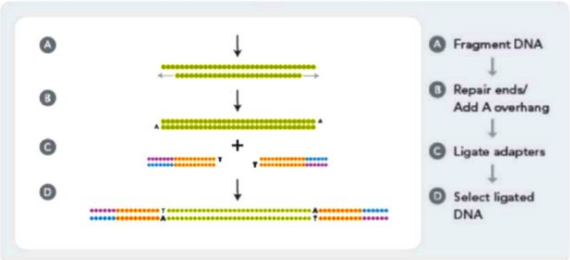
H

Hi-C - технология для оценки пространственной близости локусов в геноме.

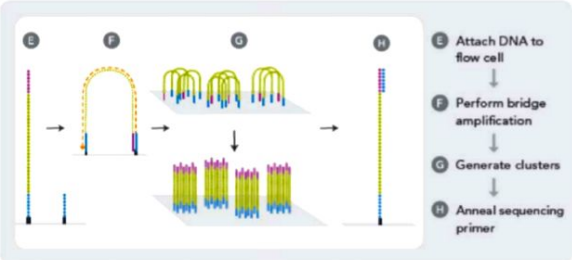
I

Illumina (ранее Solexa) — платформа секвенирования, коммерциализованная в 2006 году. Принцип работы основан на “секвенировании путем синтеза”: **I** — приготовление библиотек из фрагментов ДНК; **II** — прикрепление фрагментов ДНК к твердой подложке проточной ячейки; **III** — амплификация фрагментов с использованием 3’-блокированных флуоресцентно-меченых нуклеотидов, так, что на каждом этапе амплификации к синтезируемой цепочке ДНК может быть присоединен только один нуклеотид. После каждого шага амплификации происходит измерение флуоресценции. В конце каждого цикла блокировка с конца растущей цепи снимается, так что далее может быть присоединен еще один нуклеотид и т. д. Первоначально данный метод позволял последовательно “прочитать” около 35 нуклеотидов, на данный момент — более 200; при этом параллельно происходит чтение десятков миллионов таких фрагментов.

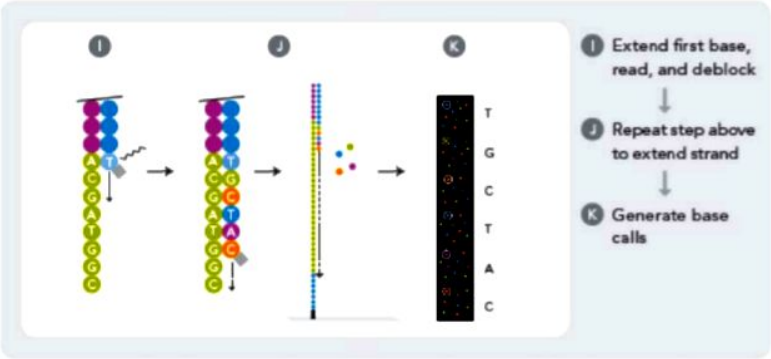
I.



II.



III.



Indel — событие, заключающееся во вставке или удалении одного или нескольких подряд идущих элементов последовательности. Данное понятие объединяет insertion (вставка одного или нескольких нуклеотидов, вплоть до крупных фрагментов хромосом, содержащих миллионы нуклеотидов) и deletion (потеря одного или нескольких нуклеотидов, вплоть до крупных фрагментов хромосом). В общем случае при сравнении нескольких последовательностей неизвестно, какое событие произошло — потеря фрагмента или его вставка.

In vitro (лат. «в стекле») — технология выполнения экспериментов, при которой опыты проводятся «в пробирке» — вне живого организма.

In vivo (лат. «в (на) живом») — проведение экспериментов на (или внутри) живой ткани при живом организме.

In silico (искаж. лат. «в кремнии») — компьютерное моделирование экспериментов.

Ion Torrent — платформа секвенирования следующего поколения, выпущенная в 2010 году, основана на технологии полупроводникового секвенирования. Максимальная длина прочтения составляет от 200 до 400 bp.

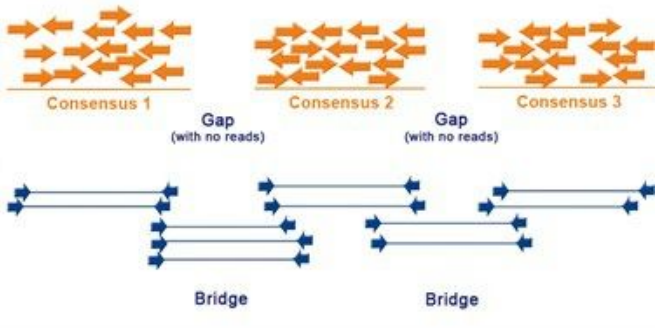
K

k-mer — подстрока рида длины k. Например 4-мерами последовательности ДНК ACCAGTA являются ACCA, CCAG, CAGT, AGTA

M

Mate pair sequencing — метод, позволяющий получать ДНК-библиотеки для секвенирования, аналогичные paired-end, но с очень большим размером вставки (до десятков килобаз). При этом секвенируют только два коротких фрагмента по краям

вставки. Также известно, какие из ридов образуют пары, и среднее расстояние между ними.

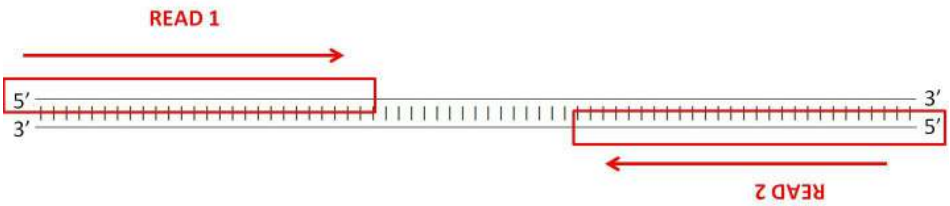


Multiple sequence alignment (MSA) — множественное выравнивание. Обычно используется для поиска консервативных регионов в группах последовательностей, гипотетически связанных эволюционно. Множественные выравнивания вычислительно сложны и в большинстве формулировок сводятся к NP-полным задачам. На практике решаются с использованием эвристических методов. Программы: ClustalW, Kalign, MUSCLE, HMMER, MAFFT и др.

P

PacBio — платформа секвенирования, представленная в 2010 году, основана на технологии “single molecule real time sequencing (SMRT)", использующей “zero-mode waveguides” (ZMW) — элементы, проводящие электромагнитные волны оптического диапазона с малой энергией. Они позволяют “наблюдать” за ячейкой объемом порядка 20 зептолитров (10^{-21} литров), детектируя присоединение ДНК-полимеразой единичного флуоресцентно-меченого нуклеотида к растущей цепочке ДНК. При присоединении нуклеотида флуоресцентная метка “отрезается” полимеразой и покидает ячейку, пропадая из поля зрения детектора. Таким образом, детектируется момент присоединения каждого нуклеотида, а его тип определяется типом флуоресцентной метки. Технология позволяет “читать” длинные фрагменты ДНК (до 7 килобаз). Основной ее проблемой является большое число ошибок чтения.

Paired-end sequencing — метод, позволяющий секвенировать оба конца молекулы ДНК (рисунок 3). При этом он сохраняет информацию о принадлежности ридов к одной паре. Участок ДНК между парными ридами, который не секвенируется, называется “вставкой”. Ее размер может варьироваться от нескольких десятков до нескольких сотен пар оснований и зависит от метода приготовления библиотеки, т.е., как правило, он известен.



Pairwise alignment — выравнивание двух последовательностей, может быть как глобальным, так и локальным. В отличие от множественного выравнивания парное выравнивание эффективно решается за полиномиальное время. В случае глобального выравнивания используется алгоритм Нидлмана-Вунша; в случае локального — алгоритм Смита-Ватермана.

Phred quality score — оценка качества прочитанного нуклеотида ДНК, изначально введенная для автоматизации процесса секвенирования ДНК в проекте “Геном человека”. Phred quality score *Q* определяется как свойство, логарифмически связанное с вероятностью ошибки *P* при определении данного нуклеотида. Таким образом, значение Q оказывается равным 20 при 99% точности определения данного нуклеотида и 30 — при 99,9% точности. Для хранения quality score для каждого нуклеотида в риде используется специальный формат fastq.

$$Q = -10 \log_{10} P$$

R

Roche 454 — первая эффективно используемая на коммерческой основе NGS-платформа. Принцип ее работы основан на пиросеквенировании. Долгое время была единственной NGS-платформой, позволяющей получать риды длиной до 500 пн, основной ее недостаток — большое число ошибок при секвенировании гомополимеров и повторов.

S

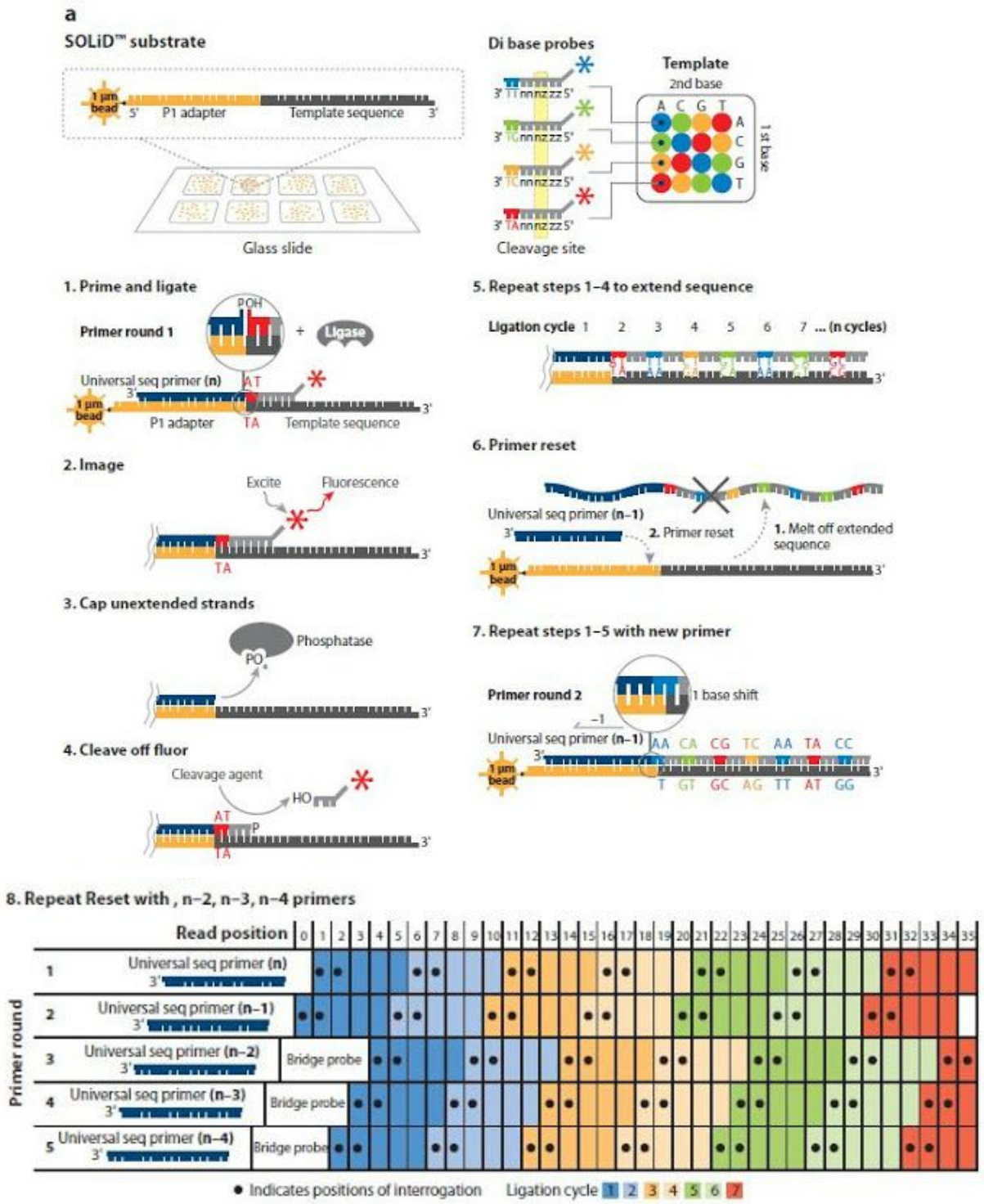
SNP (произносится “снип”, single nucleotide polymorphism) — однонуклеотидный полиморфизм — отличия размером в один нуклеотид (A, T, G или C) в геноме (или в другой сравниваемой последовательности) представителей одного вида или между гомологичными участками гомологичных хромосом. Предполагается, что встречается в исследуемой популяции с частотой более 1%.

SNP-calling — процесс поиска SNP между двумя известными последовательностями (референсом и образцом). На практике включает в себя такие этапы, как оценка качества входных данных, тримминг (удаление последовательностей низкого качества, адаптеров, контаминации и др.), выравнивание ридов образца на референс, SNP-calling (сравнение

референса и образца и выявление различающихся участков ДНК), и затем — фильтрацию полученных результатов по качеству, функциональной значимости и т. п. Включает использование таких программ, как fastqc, trimmomatic или TrimGalore, bowtie2 или BWA, samtools. В результате получают файл в формате vcf, работа с которым далее зависит от целей исследования.

SNV (single nucleotide variant) — также как и SNP, это однонуклеотидный полиморфизм, но встречающийся в исследуемой популяции с частотой менее 1%. Как правило, такие варианты недостаточно хорошо охарактеризованы, например, обнаружены только у одного индивида. Зачастую возникают сложности разделения SNV и ошибок секвенирования.

SOLiD (Supported Oligonucleotide Ligation and Detection System 2.0) — технология высокопроизводительного секвенирования путем легирования, предложена в 2005 году. Основная ее особенность заключается в том, что присоединяется одновременно по два нуклеотида (т.е. существует 16 возможных сочетаний). Они кодируются в виде матрицы преобразований из 4-х цветов. Одним цветом кодируются: пара нуклеотидов и она же в обратном порядке (например, CA и AC), пара нуклеотидов и комплементарная ей пара (например, CA и GT), пара нуклеотидов и обратно комплементарная ей пара (например, CA и TG). Для преобразования последовательности цветов в последовательность нуклеотидов нужно знать один нуклеотид из каждой пары. Преимуществом метода является то, что каждый нуклеотид читается дважды. Это увеличивает точность прочтения. Недостаток — введение цветовой кодировки требует использования специфического ПО, что заметно усложняет жизнь биоинформатикам.



Synteny block (синтенный блок) — неформально говоря, крупномасштабные гомологичные участки между разными геномами. Формально блок синтении строится

следующим образом: находятся гомологичные участки между парой (или большим количеством) геномов (это могут быть участки с хорошим выравниванием или просто гены). Эти участки называются якорями. Дальше эти якоря группируются в синтени блоки по различным правилам и каждому блоку присваивается свой номер для комбинаторной интерпретации задачи поиска геномных перестроек. Один из возможных наборов правил: 1) якоря должны находиться ближе от уже лежащего в блоке якоря, чем на N bp, 2) в блоках разных организмов с одним и тем же номером лежат одни и те же якоря.