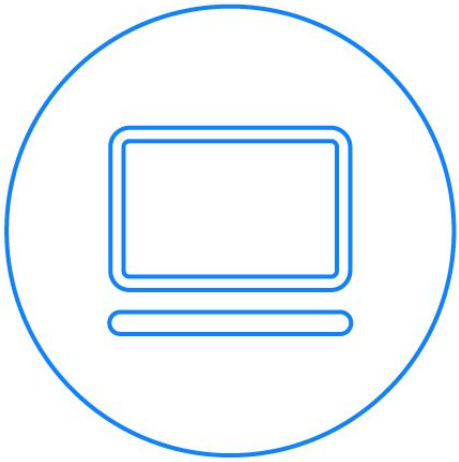


ГЛОССАРИЙ БИОЛОГИ

BIOINFORMATICS
INSTITUTE

А

Алгоритм — формализованный набор инструкций, описывающий порядок действий исполнителя для достижения результата решения задачи за конечное число шагов.

Алгоритм Нидлмана–Вунша (Needleman–Wunsch algorithm) — алгоритм глобального выравнивания двух последовательностей (*динамическое программирование*).

Алгоритм Смита–Ватермана (Smith–Waterman algorithm) — алгоритм локального выравнивания двух последовательностей (*динамическое программирование*).

Аннотация генома — процесс добавления 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.

Б

База данных — совокупность данных, организованных для удобного поиска и обработки хранимой информации с помощью ЭВМ. Наиболее крупные базы данных, где можно найти информацию о нуклеотидных последовательностях ДНК: GenBank, DDBJ, ENA SRA, Ensembl, dbSNP.

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 (U). Метилированные участки можно выявить с помощью сравнения последовательностей ДНК до и после обработки.

В

Выравнивание последовательностей — такая компоновка этих последовательностей, которая позволяет обнаруживать области сходства между ними, связанные с их структурным, функциональным или эволюционным родством (рис. 1). Исследуемые последовательности обычно представляют собой строки матрицы, в столбцах которой находятся одинаковые или сходные символы или пропуски (gaps). Существует большое число алгоритмов, используемых для выравнивания последовательностей. Их можно

разделить на две категории, в зависимости о того, локальное или глобальное выравнивание они позволяют получить (см. ниже). Существующие алгоритмы можно разделить на медленные, но точные (например, основанные на динамическом программировании), либо более эффективные, но не гарантирующие получение оптимального решения. Они основаны на эвристических или вероятностных методах. Важно помнить, что не всегда оптимальное решение биологически оправдано!

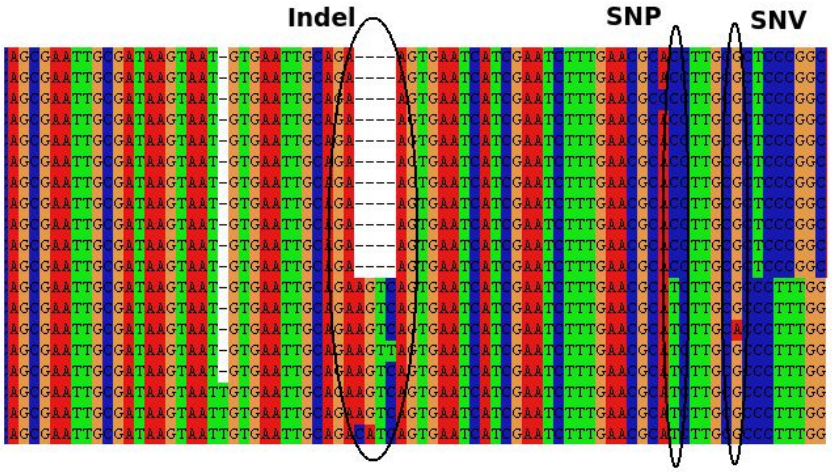


Рисунок 1 . Множественное выравнивание (MSA) нуклеотидных последовательностей.

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).

Высокоуровневый язык программирования — язык программирования, разработанный для быстроты и удобства использования. Основная черта высокоуровневых языков — абстракция, т.е. введение смысловых конструкций, кратко описывающих структуры данных и операции над ними, описания которых на машинном коде (или низкоуровневом языке программирования) слишком длинные и сложные для понимания.

Г

Гамильтонов путь в графе — путь, содержащий каждую вершину графа ровно один раз.

Граф — множество вершин (узлов), соединенных ребрами (дугами), может быть как ориентированным (рис. 2), то есть иметь направленные ребра, так и неориентированным, если направление ребра не задаётся. Неориентированный граф называется связным в том случае, если из каждой его вершины можно достичь любую другую вершину, передвигаясь по ребрам. Теория графов широко применяется при решении различных биоинформатических задач, например, при сборке генома (граф де Брёйна).

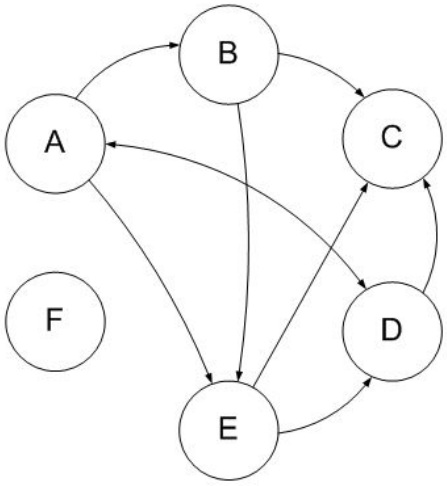


Рисунок 2. Несвязный ориентированный граф.

Граф де Брёйна — ориентированный граф, описывающий перекрытия между последовательностями символов (например, нуклеотидными последовательностями). Вершины такого графа являются *k-мерами*. Ребро соединяет две вершины в том случае, если k-меры в этих вершинах перекрываются на k-1 символ (k-1 суффикс совпадает с k-1 префиксом).

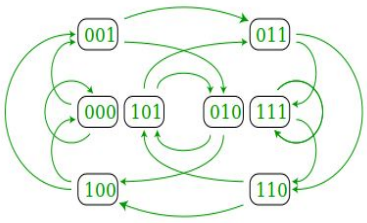


Рисунок 3. Пример графа де Брёйна для двоичных последовательностей длиной три символа.

Д

Динамическое программирование — способ решения сложных задач, предполагающий их разбиение на набор более простых перекрывающихся подзадач. Каждая подзадача

при этом решается один раз, полученный промежуточный результат сохраняется, что позволяет сократить время решения.

И

Интерпретируемый язык программирования — язык программирования, в котором исходный код программы не преобразуется предварительно полностью в машинный код, как в компилируемом языке, а исполняется последовательно с помощью специальной программы-интерпретатора, т.е., фактически, “переводится” на понятный машине язык по мере выполнения.

Исходный код — текст программы на одном из языков программирования. В случае интерпретируемых языков текст является программой сам по себе, а программу, написанную на компилируемом языке, сначала необходимо перевести в бинарный исполняемый файл (скомпилировать).

К

Класс сложности задачи — множество вычислительных задач, примерно одинаковых по сложности вычисления, т.е. для решения которых требуется примерно одинаковое количество ресурсов.

Командная строка — способ взаимодействия с компьютером без использования мыши и графического интерфейса с помощью написания специальных команд в окне терминала (командной строки). Чаще всего командная строка ассоциируется с компьютерами под управлением UNIX/Linux, хотя и в Windows, и в Mac OS она тоже есть.

Компилируемый язык программирования — язык программирования, исходный код которого преобразуется компилятором в машинный код и записывается в файл, и лишь может исполняться.

Компиляция исходного кода — трансляция программы, составленной на исходном языке программирования, в эквивалентную программу на низкоуровневом языке (машинно-ориентированном языке), близком машинному коду. Другими словами, это перевод программы из формы, понятной человеку, в форму, понятную компьютеру.

Конвейер (pipeline, workflow) — способ решения конкретных задач по обработке данных, с помощью объединения программ более общего назначения в цепочку таким образом, чтобы информация, выдаваемая одной программой, попадала на вход следующей.

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

М

Машинный код (машинный язык) — система команд (набор кодов операций) конкретной вычислительной машины, которая интерпретируется непосредственно процессором или микропрограммами этой вычислительной машины.

О

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

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

П

Покрытие нуклеотида — количество раз, которое был прочитан нуклеотид в процессе секвенирования. Отсюда можно посчитать, например, среднее покрытие генома.

Пиросеквенирование — метод секвенирования путем синтеза, основанный на обнаружении пирофосфатов, освобождающихся при присоединении к растущей цепи ДНК нуклеотидов при помощи хемолюминесценции (рисунок 4): **(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 таким образом, чтобы в каждой ячейке оказалась 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.

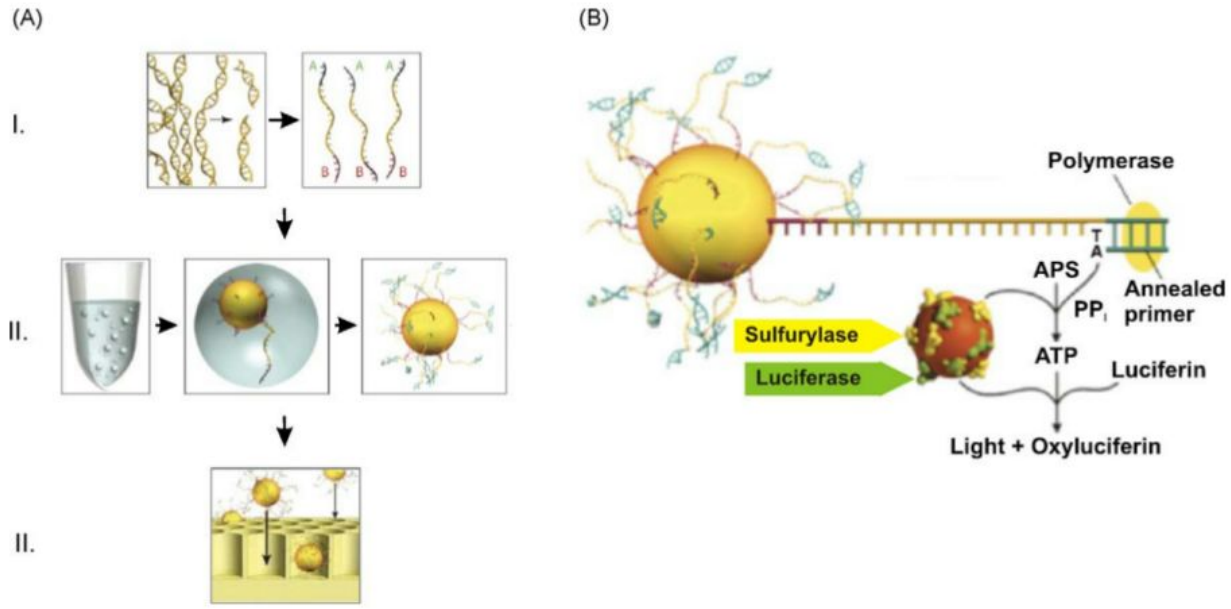


Рисунок 4. Основные принципы пиросеквенирования.

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

Р

Расстояние Левенштейна (редакционное расстояние, расстояние редактирования) между двумя строками — минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для преобразования одной строки в другую.

Рекурсия — см. рекурсия.

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.

Рид (read, прочтение) — отдельная последовательность нуклеотидов, полученная в результате секвенирования.

С

Сборка генома — объединение большого числа ридов в одну (хромосома) или несколько (контиги, скаффолды) длинных последовательностей, в результате чего должна быть восстановлена исходная последовательность генома. Выделяют **сборку *de novo*** (рисунок 5), при которой последовательность собирается только на основе имеющихся коротких прочтений, и **сборку с использованием референса**, при которой риды выравниваются на уже собранный геном того же или близкого вида. Сборка генома *de novo* является весьма сложной задачей, для ее решения используется большое число различных алгоритмов, многие из которых основаны на решении задачи поиска Эйлерова пути в графе де Брёйна. Примеры программ (ассемблеров): MaSuRCA, SPAdes, Velvet, SOAP-denovo и др. Несмотря на сложность, задача сборки является лишь первым шагом в изучении генома организма.

Секвенирование — определение первичной структуры биополимера.

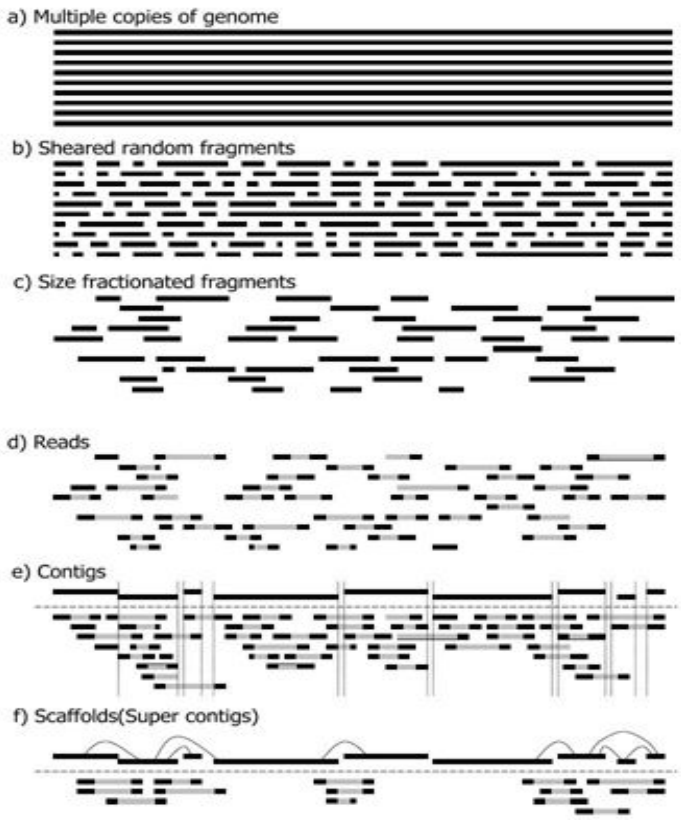


Рисунок 5. Основные этапы секвенирования и сборки *de novo* генома: фрагментация ДНК, выделение фрагментов подходящей для секвенирования длины, затем – собственно секвенирование, результатом которого является большое число коротких ридов, которые уже программно собирают сначала до контигов, а затем – до скаффолдов.

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 (рисунок 6).

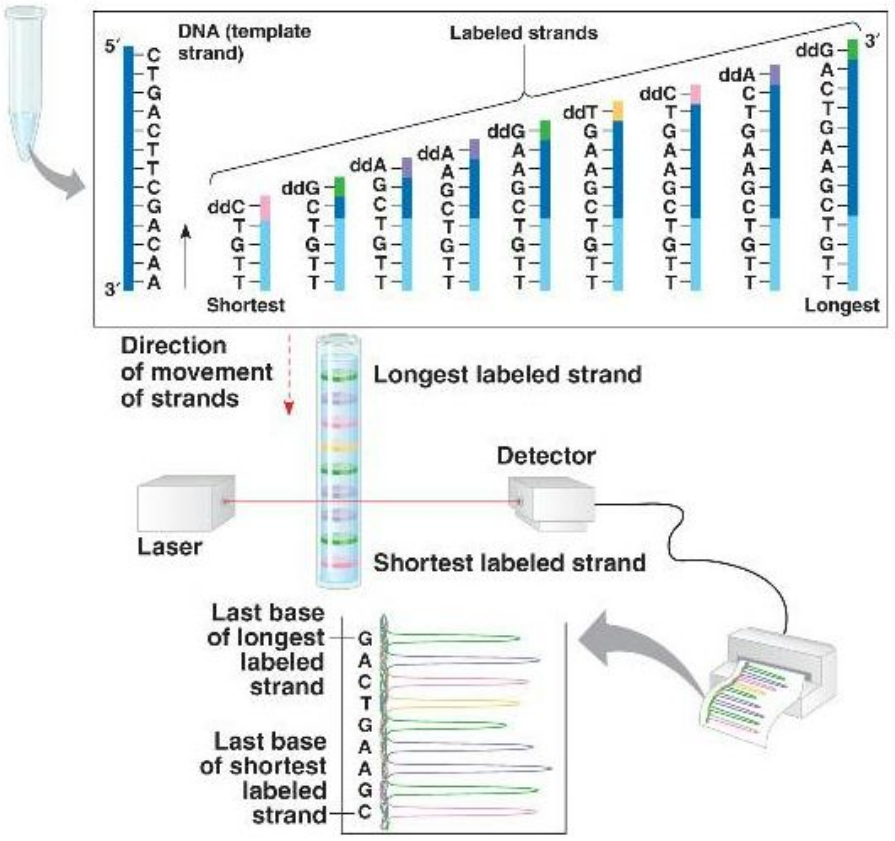


Рисунок 6. Основные принципы секвенирования по Сэнгеру.

Секвенирование следующего (нового) 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.

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 сократить круг поиска места локализации исследуемого гена на хромосомах, ограничивая его областью известных генов, принадлежащих к конкретной синтеничной группе (см. также синтенный блок).

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

Система контроля версий — компьютерная система управления версиями файлов, позволяющая в любой момент обратиться к произвольной версии этих файлов, а также

управляющая изменениями, поступающими от нескольких пользователей системы. Подобные системы используются в таких известных программах общего пользования, как dropbox, ЯндексДиск, но существуют и более специализированные системы, используемые для управления версиями текстовых файлов, например, Git или SVN.

Скаффолд (scaffold, с англ. "*строительные леса"***)** — серия контигов, расположенных в правильном порядке, но необязательно соединенных в одну непрерывную последовательность.

Скрипт — разновидность программы, написанная на интерпретируемом языке (а значит, не требующая специальной компиляции) и используемая биоинформатиками для автоматизации своих задач.

Ф

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 качестве (phred quality score) каждого из нуклеотидов. Состоит из повторяющихся блоков строк, первая часть блока

содержит идентификатор последовательности, начинающийся с символа ‘@’. После этой строки записывается последовательность нуклеотидов. Вторая часть блока также содержит идентификатор последовательности, который предваряется символом ‘+’, на следующих строках блока уже идут символы, описывающие качество (phred quality) каждого нуклеотида.

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.

Э

Эвристика (эвристический алгоритм) — алгоритм решения задачи, не имеющий строгого обоснования (его правильность во всех случаях не доказана), позволяющий достаточно быстро получать приближенное решение сложной задачи (или точное, но в частных случаях). При этом возможно, что в отдельных случаях эвристический алгоритм может давать неоптимальное или даже неверное решение.

Эйлеров путь в графе — это путь, проходящий по всем ребрам графа только один раз.

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

B

BLAST (Basic Local Alignment Search Tool) — алгоритм, опубликованный в 1990 году, был разработан для быстрого приблизительного (эвристического) поиска в базах данных белков или последовательностей ДНК (например, Nucleotide, http://www.ncbi.nlm.nih.gov/nuccore/). В настоящее время под названием BLAST объединено целое семейство программ, специализированных под различные виды данных.

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.

Git — распределенная система контроля версий. Она позволяет регистрировать изменения в файлах, добавлять к ним комментарии, и, при необходимости, возвращаться к старым версиям этих файлов. Распределенной система называется потому, что у каждого пользователя репозитория хранится полная история изменений файлов. Подобный подход позволяет легко восстанавливать основной репозиторий (как правило, расположенный на сервере) в случае его

утраты, а так же одновременно работать нескольким группам людей в рамках одного проекта.

GitHub — основанный на Git веб-сервис для размещения IT-проектов и их совместной разработки. С одной стороны, позволяет публиковать собственные разработки, с другой — читать и комментировать чужие, а также объединять работу нескольких людей в одном проекте.

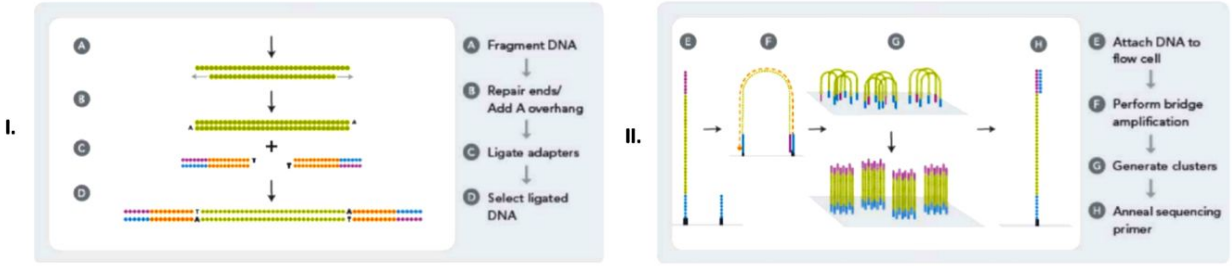
Genome-wide association study (GWAS) — исследование связи генотипа с различными фенотипическими признаками (в первую очередь, с наследственными заболеваниями) в масштабе всего генома. При этом сравнивают геномы людей, подверженных болезни (cases), с геномами здоровых людей (controls). В результате выявляют 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” (рисунок 7): **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.



III.

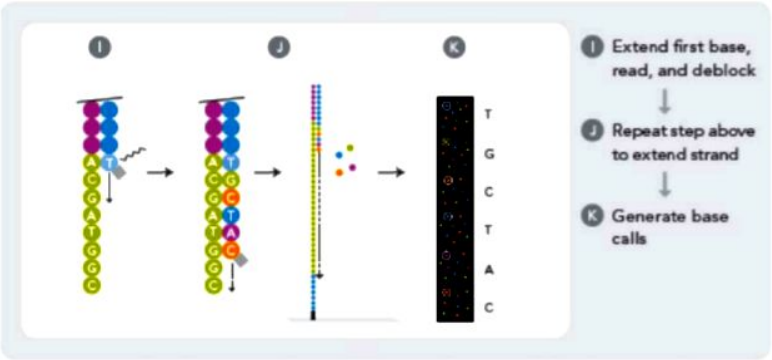


Рисунок 7. Основные принципы секвенирования технологии Illumina.

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.

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 (прочтения) 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 (рисунок 8).

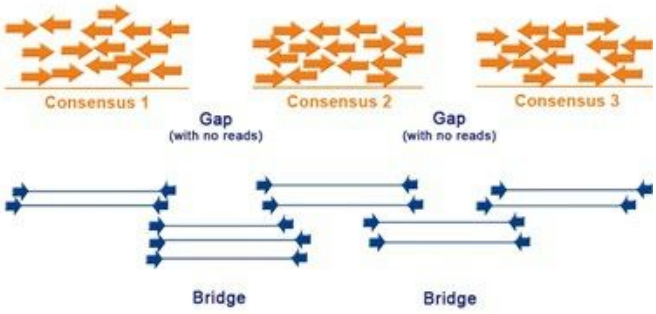


Рисунок 8. Сборка участка ДНК с использованием контигов, собранных из обычных коротких ридов (показаны оранжевым), при помощи mate pair ридов (показаны синим).

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.

N

NP-задача — неформально говоря, такая задача, которая может быть решена достаточно быстро (за полиномиальное время), если всякий раз, когда компьютер будет сталкиваться с необходимостью выбора следующего шага, некий “оракул” будет подсказывать ему, какой из путей решения правильный. Соответственно, так как в реальности такого “оракула” нет, поиск решения задачи будет гораздо более долгим. Ускорить поиск можно с помощью некоторого алгоритма, который позволит всякий раз делать правильный выбор. На практике, подобные задачи решают, используя эвристические методы.

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 9). 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.

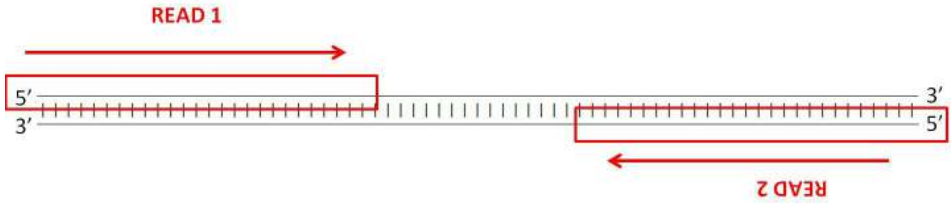


Рисунок 9. Прямой и обратный paired-end риды.

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.

Perl — высокоуровневый интерпретируемый динамический язык программирования, разработанный Ларри Уоллом, лингвистом по образованию, специально для работы с текстом. Согласно Ларри, у Perl есть два девиза: первый — “Есть больше одного способа это сделать” (“*There’s more than one way to do it*”), второй — “Простые вещи должны оставаться простыми, а сложные — стать выполнимыми” (“*Easy things should be easy and hard things should be possible*”). Этот язык широко применялся (и часто до сих пор применяется) для анализа нуклеотидных и аминокислотных последовательностей.

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.

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

Python — высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода.

Название языка произошло вовсе не от вида пресмыкающихся: автор назвал язык в честь британского комедийного телешоу 1970-х «Летающий цирк Монти Пайтона».

P-задача — неформально говоря, это задача, которая достаточно быстро (за полиномиальное время) решается современными компьютерами.

R

R — язык программирования для статистической обработки данных, широко используется для обработки биологических и биоинформатических данных.

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 (рисунок 10), 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.

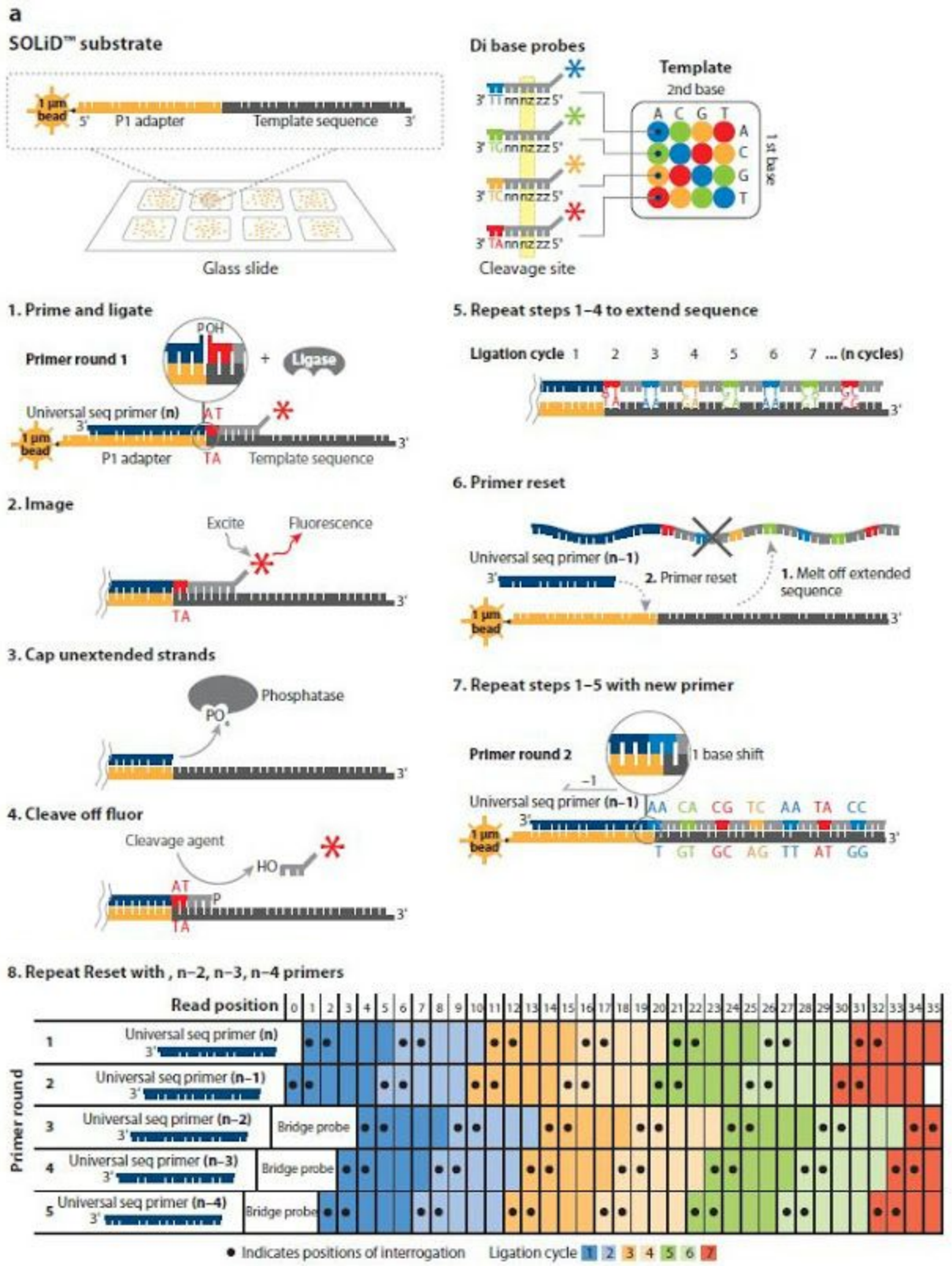


Рисунок 10. Основные принципы секвенирования технологии SOLiD.

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.

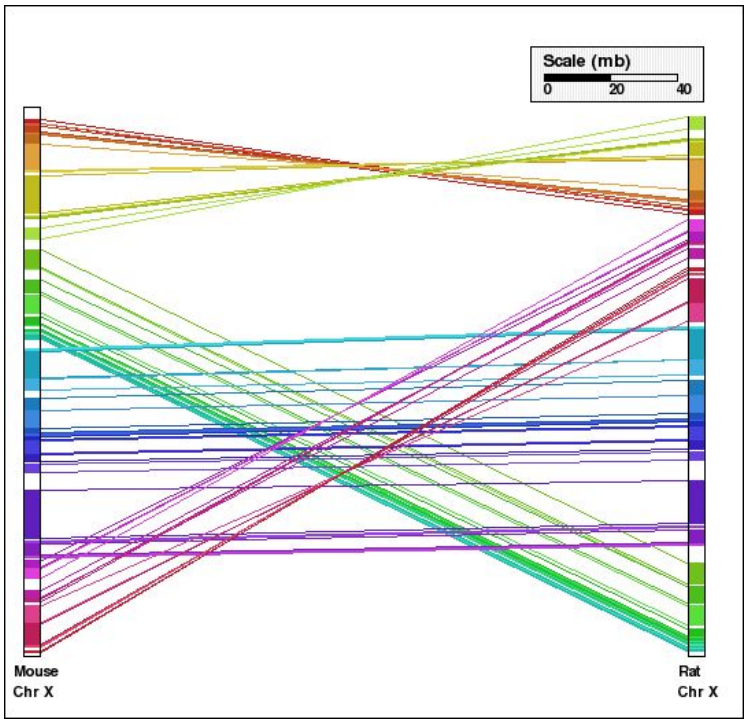


Рисунок 11. Сравнение хромосомы X мыши и крысы при помощи программы NOAGG: 49 synteny blocks.

U

UNIX/Linux — семейство изначально многопользовательских и многозадачных операционных систем (ОС). Чаще всего используется на серверах и вычислительных кластерах, однако может устанавливаться и на персональные компьютеры в качестве альтернативы коммерческим ОС (таким как Windows). Особенностью этих систем является модель разработки. Поскольку ОС имеют открытый исходный код, в их создании участвуют программисты-добровольцы со всего мира. Впрочем, число версий так велико, что есть и “закрытые” ветви — например, Mac OS, которая с некоторого времени внезапно стала “потомком” UNIX-систем.