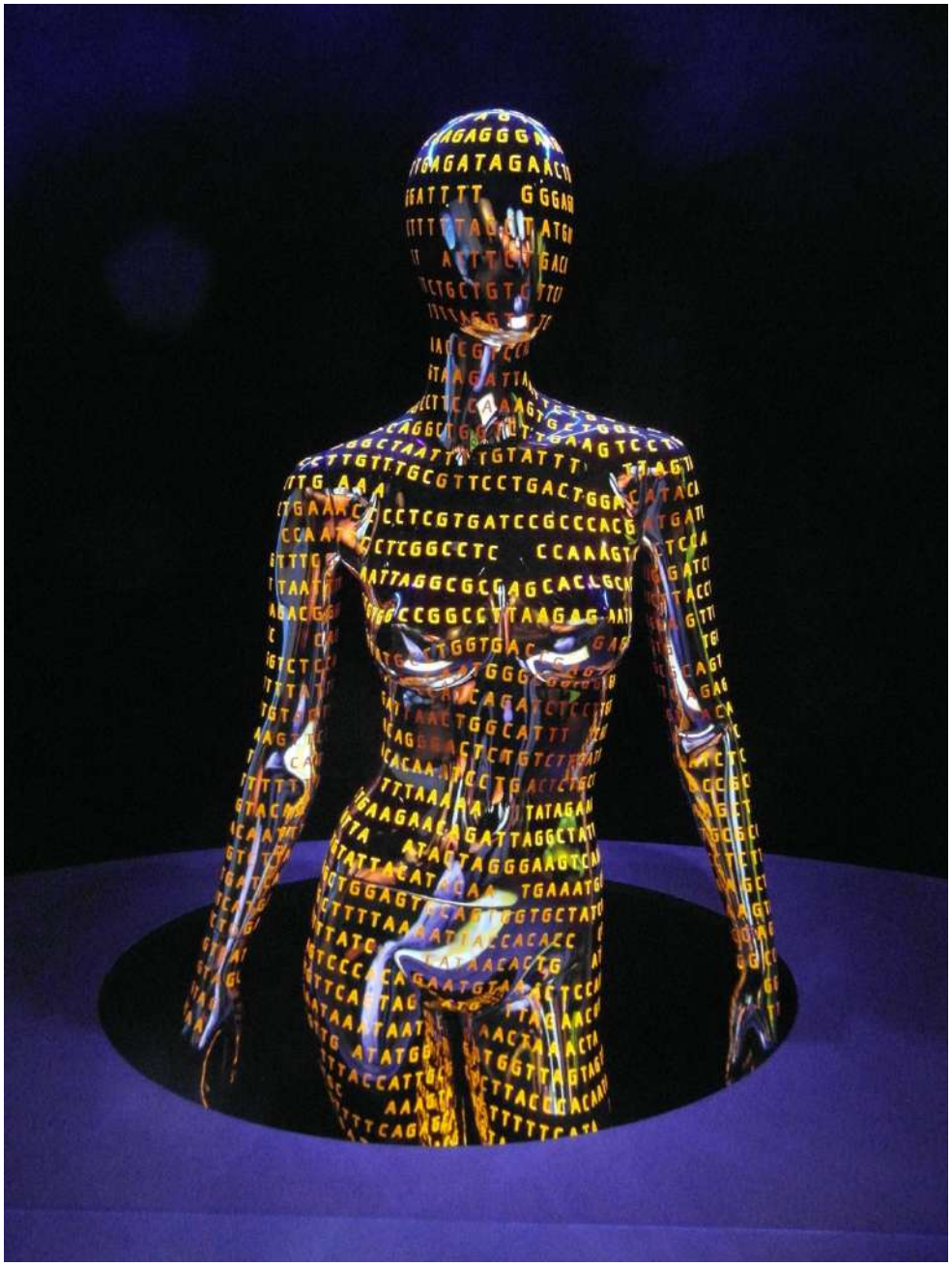
УНИВЕРСИТЕТ ИТМО

В.Б. Брюхин, Е.В. Андрусенко

ФУНКЦИОНАЛЬНАЯ ГЕНЕТИКА И ГЕНОМИКА



Санкт-Петербург
2021

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

В.Б. Брюхин, Е.В. Андрусенко

ФУНКЦИОНАЛЬНАЯ ГЕНЕТИКА И ГЕНОМИКА

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлениям подготовки 19.04.01 – «Биотехнология»; 09.04.03 – «Прикладная информатика»; 12.04.04 – «Биотехнические системы и технологии» в качестве учебно-методического пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург
2021

Брюхин В.Б., Андрусенко Е.В. Функциональная генетика и геномика. – СПБ: Университет ИТМО, 2021. – 112 с.

Рецензент: Духинова М.С., к.б.н., научный сотрудник химико-биологического кластера ФГАОУ ВО «Национальный исследовательский университет ИТМО»

Предназначено для студентов-магистрантов химико-биологического кластера Университета ИТМО, а также может быть рекомендовано студентам естественно-научного профиля при выполнении ими ряда работ в специализированных практикумах. Настоящее учебное издание также предназначено для обеспечения методического сопровождения образовательной программы: прикладная геномика и дисциплин в рамках этой программы таких, как функциональная геномика / Functional Genomics, геномика растений / Plant Genomics, эволюционная геномика растений / Evolutionary Plant Genomics, геномика развития / Development Genomics и других родственных образовательных программ и дисциплин. В настоящем издании рассмотрены предмет, цели и возникновение функциональной генетики и геномики, их связь с геномикой, протеомикой и биоинформатикой, методологию и высокотехнологичное секвенирование, которые они используют, значение функциональной генетики и геномики как науки.

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Содержание

Введение

Функциональная генетика и геномика — это область молекулярной биологии, изучающая совокупность генов организма и их функцию, то есть их экспрессию и регуляцию во времени и пространстве, следовательно на разных стадиях развития организма от зиготы (оплодотворенной яйцеклетки) до самой смерти, а также в разных органах, тканях и клетках данного организма. Сравнительная функциональная генетика и геномика, соответственно, изучает сходства и различия в работе и регуляции геномов на клеточном, организменном, популяционном и видовом уровнях. Таким образом, функциональная генетика фокусируется на динамических аспектах работы генома (совокупности всей ДНК организма), таких как транскрипция генов, трансляция и меж белковые взаимодействия, что отличает ее от генетики, которая сосредоточена на статических аспектах, таких как последовательность или структуры ДНК. В отличие от функциональной генетики, которая изучает работу отдельных групп генов и единичных генов, функциональная геномика характеризует работу всех генов в клетке, ткани или организме и использует большие количества данных. Она начала свое стремительное развитие с возникновением технологий NGS (next generation sequencing – англ. секвенирование нового поколения), микрочипов, иммунопреципитации хроматина и дальнейшего секвенирования связанных с хроматином ДНК (ChIP-seq), технологий исследования ДНК-белок взаимодействия и других, которые позволили получать полногеномные и полнотранскриптомные данные, а также информацию о генетической и эпигенетической регуляции генов и взаимодействию транскрипционных факторов и регуляторных белков с ДНК.

Развитие функциональной геномики помогает решить многие биологические проблемы, а также способствует успехам и внедрению новых технологий в медицине, сельском хозяйстве, криминалистике, фармакологии, промышленности и во многих других жизненно важных отраслях.

В настоящем издании мы рассмотрим предмет, цели и возникновение функциональной генетики и геномики, их связь с геномикой, протеомикой и биоинформатикой, методологию и высокотехнологичное секвенирование, которые они используют, значение функциональной генетики и геномики как науки.

1. Что такое геном и геномика

Геномом называется совокупность всей ДНК клетки организма, которая кодирует наследственную информацию. Соответственно, под *геномикой* понимают раздел молекулярной биологии, который изучает структуру, организацию и функцию генома, а также эволюцию геномов различных организмов и сравнение их между собой. На данных геномики базируется и молекулярная филогения. Термин *геномика* впервые использовал американский генетик Томас Родерик в 1986 году для обозначения совокупных исследований по картированию, секвенированию и характеристики геномов. *Функциональная геномика* является частью геномики, изучающей работу и функционирование геномов в различных клетках, тканях и организмах в динамике, то есть реализацию генетической информации, записанной в геноме. Как таковая геномика, в современном виде, сложилась в 80-х-90-х годах прошлого века с возникновением методов молекулярной биологии и реализацией первых геномных проектов и получила быстрое развитие в начале 21 века с приходом высокоскоростных методов секвенирования нового поколения NGS и биоинформатики, позволившей работать с большими массивами данных. Однако возникновению геномики предшествовал длительный период истории биологии и естественных наук.

1.1. Краткая история возникновения генетики и геномики

Первые дошедшие до нас представления о наследственности, то есть сходстве родителей и потомства растений, животных и человека, были сформулированы еще античными философами и учеными Древней Греции. Так, Гиппократ считал, что в зародышевых органах мужчин и женщин откладываются особые вещества, которые при слиянии определяют развитие потомства. Намного позже, во второй половине19 века, в своей теории пангенезиса английский естествоиспытатель Чарльз Дарвин (1809–1882) (рис. 1.1.1), основываясь на своих исследованиях и размышлениях, использовал термин «геммула» для обозначения микроскопической единицы наследования. Геммулы, согласно учению Дарвина, передают в половые клетки наследственную информацию об изменениях тела [1]. Теория пангенезиса впервые была представлена Дарвиным в 1868 году в книге "Изменение животных и растений при одомашнивании". Основная проблема того времени, которую он так и не смог решить из-за того, что тогда отсутствовали научные данные, - почему наследственные признаки, полученные от родителей, не «сливаются» у потомства при смешивании. Впоследствии теория пангенеза была отвергнута, хотя она просуществовала некоторое время, поскольку открытые в 1865 году монахом - августинцем Грегором Иоганном Менделем (1822-1884) (рис.

1.1.2) законы генетического наследования стали широко известными намного позже [2]. Его открытия, сделанные на основе опытов, произведенных на горохе, заложили начало новой науке - генетике. Изложив три основных закона о передаче наследственных признаков от родителей к потомству: 1) о единообразии первого поколения гибридов, 2) о расщеплении признаков во втором поколении и 3) о независимом наследовании признаков, Мендель предложил идею бинарного кодирования, а передающие признаки материальные факторы он назвал «Anlagen» (нем. зачатки), что было прообразом открытых позже генов. Менделю удалось решить проблему смешивания, однако открытие физической структуры и химических свойств носителей наследственности было еще впереди. В 1889 году голландский ученый Хуго Де Фриз, пересматривая теорию Дарвина о пангенезисе, в своей книге «Внутриклеточный пангенезис» вводит термин «панген», под которым он понимал частицы, с помощью которых происходит наследование признаков организмами. Позже «панген» был сокращен до слова «ген». Кроме того, Хуго Де Фриз заново открыл законы Менделя, которые более тридцати лет оставались неизвестными для научного сообщества, и опубликовал их. В 1910 году американский ученый Томас Хант Морган с коллегами выдвинул свою хромосомную теорию, основанную на исследованиях на плодовых мухах дрозофилах. В этой теории Морган предположил, что расположенные в клеточном ядре хромосомы являются основой клеточной наследственности, а гены, являющиеся носителями определенных свойств, располагаются в хромосомах в определенных местах, каждому гену соответствует своя локация. Морган представил первую генетическую карту, показывающую расположение генов на хромосомах дрозофилы, основанную на скрещиваниях и частоте встречаемости каждого гена. Под геном тогда Морган понимал

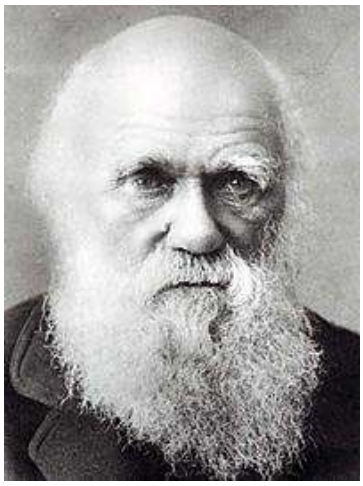


Рис. 1.1.1. Чарльз Дарвин

Рис. 1.1.2. Грегор Мендель

определенный признак. Было выявлено, что хромосомы состоят из нуклеиновых кислот и белков. Таким образом, было показано, что нуклеиновые кислоты являются носителями наследственности.

Термин *геном* был введен немецким ботаником Хансом Винкером в 1920 году, под этим термином он понимал совокупность генов в гаплоидном наборе хромосом. Позже, в 1945 году, Бидлом и Тэйтумом была предложена теория «один ген - один фермент», согласно которой определенные гены кодируют определенные белки. В 1953 году американец Джеймс Уотсон и англичанин Френсис Крик открыли двухцепочечную спиральную структуру ДНК (дезоксирибонуклеиновой кислоты), далее ими была предложена модель, согласно которой информация идет от ДНК к РНК, а от последней к белку. Открытие триплетного генетического кода, согласно которому каждую аминокислоту в ДНК и РНК кодируют три определенных нуклеотида (триплета), а также кодирующей (гены) и некодирующей («мусорная») частей ДНК, подразделение генов на экзоны и интроны, которые вырезаются при трансляции благодаря сплайсингу, подготовило почву к возникновению новых наук, молекулярной биологии и молекулярной генетике, которые, в свою очередь, дали начало геномике. Первым полностью отсеквенированным (расшифрованным) геномом в 1977 году был геном бактериофага Ф-Х174, размер которого чуть больше 5 тысяч нуклеотидов. В 1990 году стартовали международные научные проекты по секвенированию геномов эукариот, прежде всего, генома модельного растения арабидопсиса и генома человека. Оба проекта продолжались более 10 лет, в их реализацию были вовлечены тысячи ученых из разных стран мира, и стоили эти проекты миллионы долларов. В результате этих проектов было выяснено, что размер генома арабидопсиса - порядка 125 миллионов пар нуклеотидов, тогда как геном человека в гаплоидном состоянии содержал 3,2 миллиарда пар нуклеотидов, при этом количество генов в геноме арабидопсиса и человека оказалось примерно одинаковым - 22-25 тысяч. Геномные проекты позволили обнаружить то, что гены часто перекрываются, при этом отдельные гены имеют несколько продуктов. Также было выявлено, что значительную часть геномов составляют некодирующие белок повторы нуклеотидов, как короткие, так и длинные. Например, кодирующая белки в геноме человека часть составляет менее 2%. По результатам первого отсеквенированного генома человека был составлен так называемый синтетический эталонный или «референсный» геном несуществующего человека, поскольку куски этого генома состояли из ДНК разных людей. Впоследствии первая версия многократно улучшалась и корректировалась, таким образом, что сейчас мы уже имеем 38 версию референсного генома. С приходом технологий нового поколения секвенирования цена секвенирвания генома начала стремительно падать год от года, а скорость его значительно возросла. Если секвенирование первого генома человека заняло порядка 12 лет и стоило несколько

миллионов долларов, то с использованием новых технологий высокоскоростного секвенирования вся процедура занимает около 15 часов и стоит чуть выше 500 долларов. Технологические достижения, а также успехи в компьютерной биологии и биоинформатике, которые обеспечили ученых новыми программами и алгоритмами для сборки геномов и работы с большими массивами данных, значительно ускорили совершенствование геномики как прикладной и фундаментальной науки и открыли новые горизонты для ее поступательного развития.

Целью геномики является изучение геномов живых организмов в целом, а также отдельных его частей, как, например, больших хромосомных сегментов (фрагментов хромосом), содержащих семейства генов и другие структуры. Исходная задача геномики - определение полногеномной последовательности нуклеотидов, а также картирование и упорядочение генетических структур на всей последовательности ДНК. Функциональная геномика направлена на получение информации о функциях последовательностей ДНК. Для начала рассмотрим строение генома эукариот и его отличие от прокариотического генома.

Вопросы:

1) Сформулируйте три закона Г. Менделя.
2) Кто разработал хромосомную теорию?
3) Кто и когда предложил термин «геном»?

1.2. Структура генома

Как было сказано выше, геном состоит из совокупности всей ДНК клетки, то есть ДНК, заключенной в хромосомах, включая гены, межгенные последовательности и повторы. Кроме того, у *эукариот* (организмов, чьи клетки содержат ядро, окруженное мембраной) бывает два или три генома, ядерный и митохондриальный геномы (у животных и растений) плюс пластидный геном (у растений). Хотя, как правило, если не указано дополнительно, термин *«геном»* обычно относится к ядерному геному.

ДНК всех организмов состоит из четырех нуклеотидов, ароматических гетероциклических соединений, аденина (A), гуанина (G), цитозина (C) и тимина (T), представляющих собой моноэфиры ортофосфорной кислоты, соединенные фосфодиэфирной связью, также в состав молекулы нуклеотида входит моносахарид *дезоксирибоза* (тиминоза) $C_5H_{10}O_4$ (рис. 1.2.1). При этом аденозин и гуанозин являются β-N-гликозидами пуринов, а цитидин и тимидин β-N-гликозидами пиримидинов.

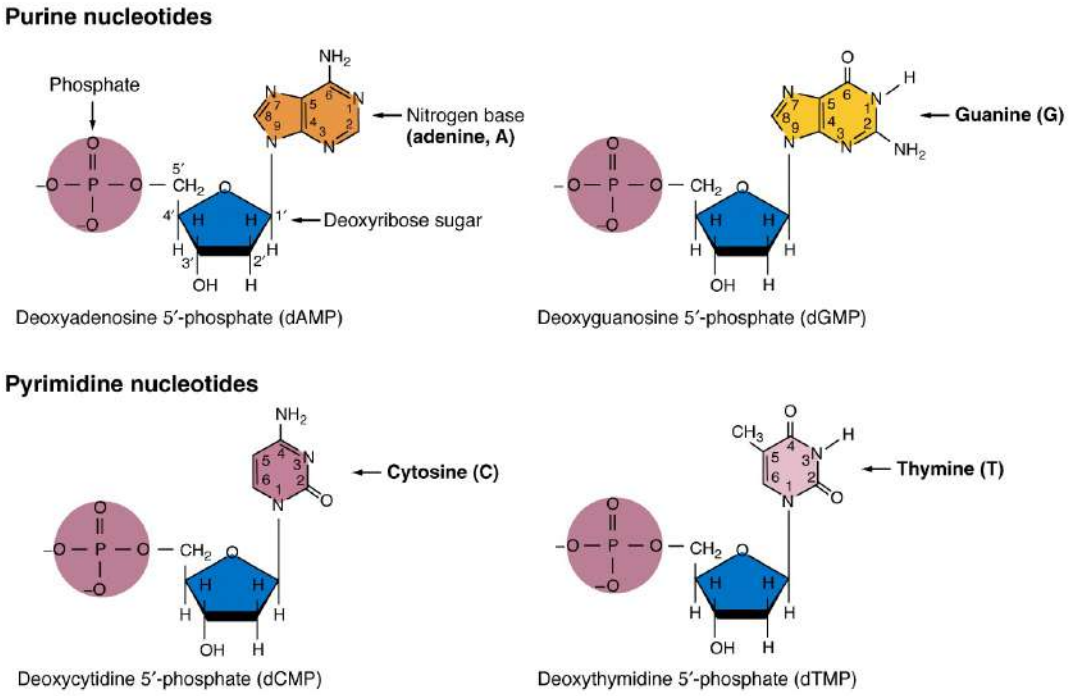


Рисунок 1.2.1. Структурные элементы четырёх нуклеотидов, из которых состоит ДНК: аденозин (A), гуанозин (G), цитидин (C) и тимидин (T).

ДНК (дезоксирибонуклеиновая кислота) представляет собой линейный полимер, имеющий форму двойной спирали (рис. 1.2.2.b). Комплементарные цепи ДНК ориентированы в противоположные стороны относительно друг друга (направление от 5'-конца к 3'-концу одной цепи соответствует направлению от 3'-конца к 5'-концу другой цепи, рис. 1.2.2.c), при этом между пуриновыми азотистыми основаниями одной цепи и пиримидиновыми основаниями другой образуются водородные связи по принципу комплементарности (рис. 1.2.2.a), то есть «дополнительности». Нуклеотид A образует водородные связи с нуклеотидом T, а G с C. Ширина двойной спирали ДНК чуть меньше 2,5 нм (нанометров), а длина может сильно варьироваться в зависимости от длины хромосомы или плазмиды, которые она составляет. У эукариот длина хромосом обычно составляет несколько микрон.

Ядерная ДНК прокариот локализована в *хромосомах*, нуклеопротеидных структурах, в которых хранится и реализуется генетическая информация [3]. Различные эукариотические организмы имеют разное, но стабильное число хромосом в составе своего ядра. Определенное количество хромосом в клетке является видоспецифичным признаком, то есть постоянным для клеток данного вида, а набор всех хромосом клетки называется *кариотипом*. На рис. 1.2.3. изображен растения из рода *Boechera* семейства капустных (Brassicaceae), который в

диплоидном состоянии насчитывает 14 хромосом и человека 2n=46. Хромосомы в соматических клетках образуют *гомологичные пары*, соответствующие *диплоидному* (двойному) набору (2n), в которых одна хромосома наследуется от материнского, а другая от отцовского организма. В половых клетках (гаметах) содержится *гаплоидное* (одинарное) число хромосом (n). При оплодотворении, то есть слиянии двух гамет (яйцеклетки и сперматозоида), происходит восстановление двойного набора хромосом и образуется диплоидная *зигота*.

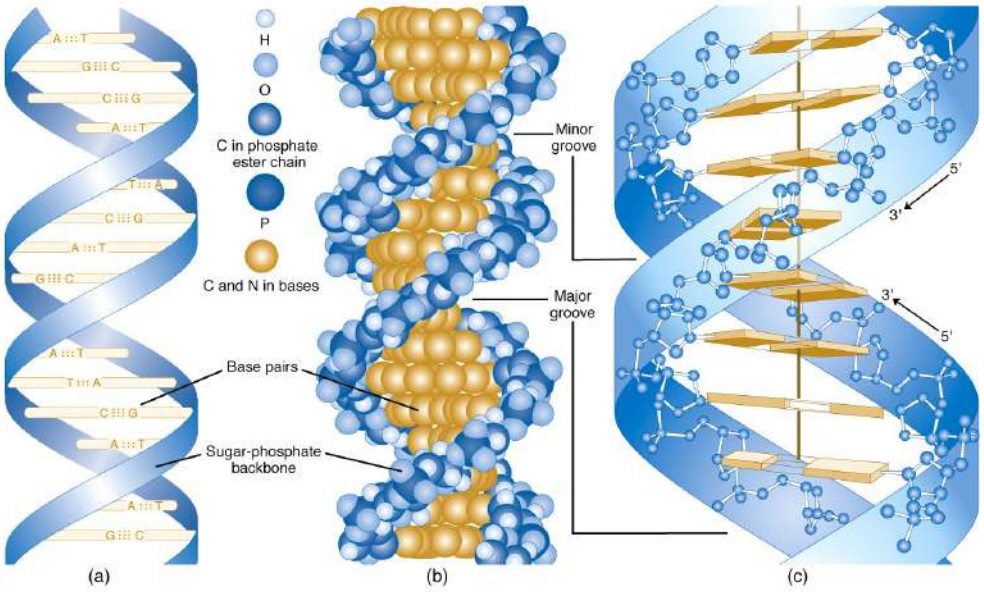
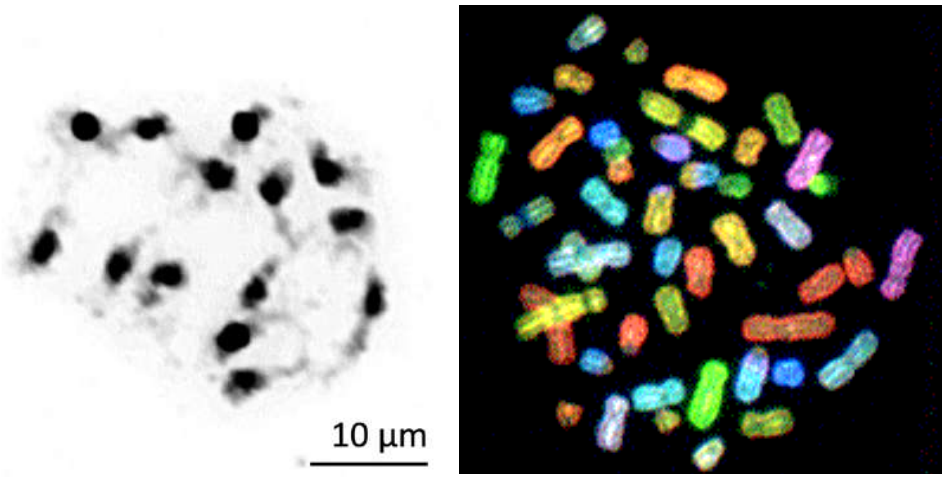


Рисунок 1.2.2. Структура двухцепочечной молекулы ДНК: (a) водородные связи между пуриновыми азотистыми основаниями одной цепи и пиримидиновыми основаниями другой; (b) ДНК - линейный полимер, имеющий форму двойной спирали; (c) Комплементарные цепи ДНК ориентированы в противоположные стороны относительно друг друга: направление от 5'-конца к 3'-концу одной цепи соответствует направлению от 3'-конца к 5'-концу другой цепи.

Морфологически хромосомы разделяются на разные типы по положению *центромеры* (участка хромосомы, связывающего две сестринские хроматиды и играющего большую роль при делении клеток): *метацентрические* - центромера расположена в середине хромосомы; *телоцентрические* - центромера на конце хромосомы; *акроцентрические* - центромера расположена близко к концу хромосомы и др., при этом буквой «p» обозначается короткое плечо хромосомы, а буквой «q» - длинное плечо.



(a) (b)

Рисунок 1.2.3. Диплоидные наборы хромосом: (a) растения из рода *Boechera* семейства капустных (Brassicaceae) 2n=14; (b) человека 2n=46.

Вопросы:

1) Какие типы геномов вам известны?
2) Что такое ДНК и как она устроена?

1.2.1. Структура генов: экзоны, интроны, промоторы

В биохимическом понимании *ген* – это участок ДНК, ответственный за синтез определенного белка. Бывают кодирующие и не кодирующие белок гены, последние называются *псевдогенами*. Структура эукариотических генов, кодирующих белки, обозначается слева направо и состоит: из 5’ верхнего региона, участвующего во взаимодействии с регуляторными сигналами, области промотора с сайтом инициации транскрипции (например, последовательность TATA), который распознается *РНК-полимеразой*, 5’ нетранслируемой области (5’ UTR), сайта инициации трансляции (включает стартовый кодон ATG), переменной последовательности экзонов (кодирующих белок) и интронов (некодирующие белок участки, которые вырезаются при трансляции), сайта остановки трансляции (стоп-кодона TAA, TAG или TGA), нетранслируемой области 3 ’UTR, сигнала полиаденилирования (поли-A хвост), сайта остановки трансляции (рис. 1.2.1.1).

Процесс синтеза РНК (рибонуклеиновой кислоты) на матрице ДНК называется *транскрипцией* (рис. 1.2.1.1.) [4]. Начинается он с того, что к промотору прикрепляется РНК-полимераза, необходимая для синтеза РНК, и транскрипционные факторы, которые активируют работу промоторов. В

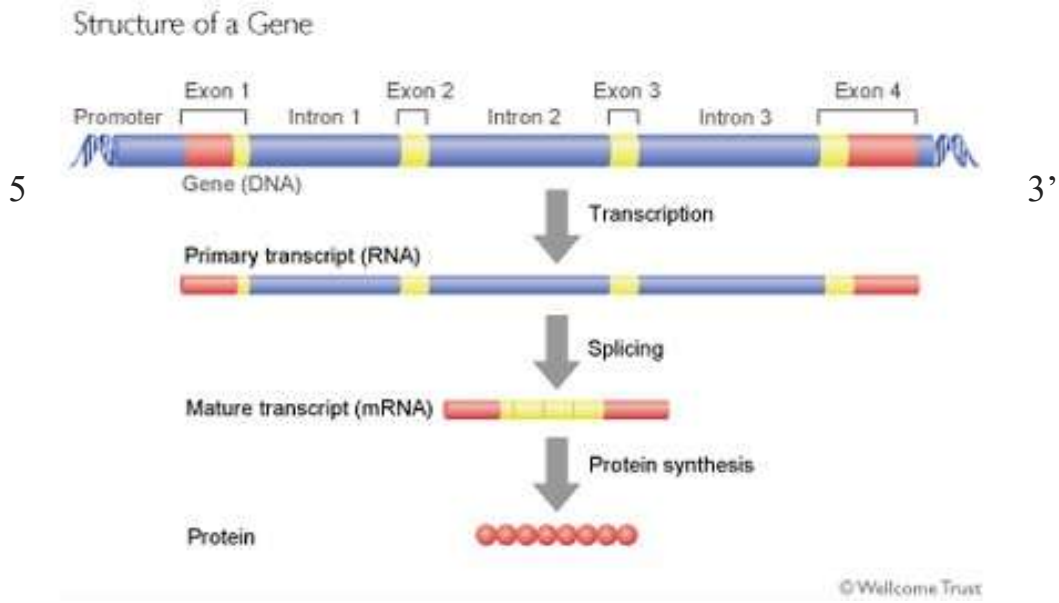
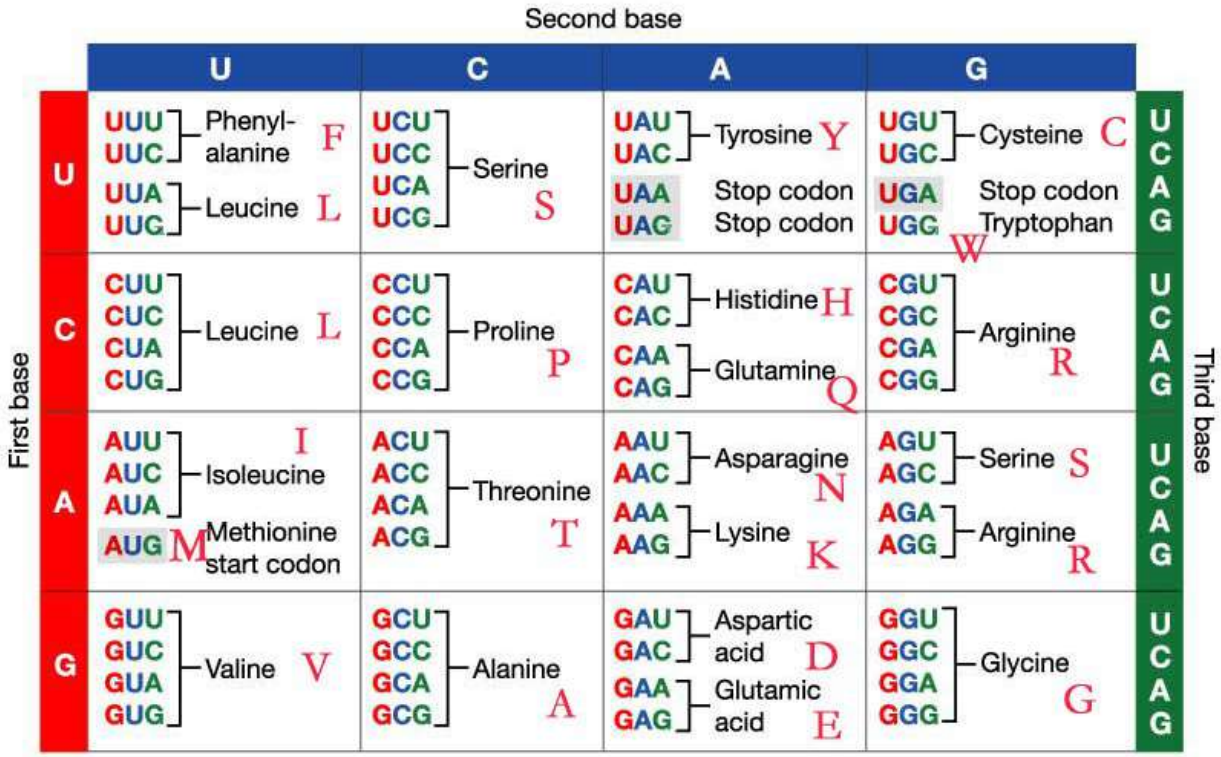


Рисунок 1.2.1.1. Схематическое изображение структуры эукариотического гена, транскрипция и трансляция.

результате транскрипции синтезируется одноценочечная мРНК (матричная = информационная РНК), комплементарная одной из цепей ДНК, на которой она синтезируется, при этом в РНК азотистые основания, содержащие тимин (Т) в ДНК, заменены на основание урацил (U). В отличие от ДНК, молекула РНК является одиночной, то есть не содержащей второй комплементарной цепи. Далее идет процесс созревания мРНК или *процессинг*, при котором к 5' концу транскрипта прикрепляется измененный нуклеотид кэп (англ. cap - колпачок), а к 3' концу - несколько нуклеотидов, содержащих основание аденин, что приводит к формированию поли-(А) хвоста, защищающего мРНК от расщепления в цитоплазме, затем в результате *сплайсинга* удаляются интроны, фрагменты гена, не несущую информацию о белке, в результате чего экзоны, кодирующие белок, сшиваются в один фрагмент (рис. 1.2.1.1.). На этом процессинг завершается, и созревшая РНК готова к *трансляции* (синтезу белка).

Если транскрипция происходит в клеточном ядре, то трансляция протекает в цитоплазме. Синтез полипептидной цепи из аминокислот происходит с помощью рибосом на мРНК и с участием тРНК (транспортной РНК). *Рибосомы*, состоящие у эукариотических организмов из двух субъединиц, большой и малой (60S и 40S), движутся по цепочке мРНК (рис. 1.2.1.1), а молекулы-адаптеры тРНК (транспортная РНК) доставляют аминокислоты к рибосомам для синтеза полипептидной цепи. Причем природа кодирования заключается в том, что каждому *кодону*, то есть трем последовательным нуклеотидам в РНК (триплету), соответствует одна аминокислота (таблица 1.2.1.1.).

Таблица 1.2.1.1. Кодоны мРНК и соответствующие им аминокислоты [5]



Из таблицы 1.2.1.1 видно, что многие аминокислоты кодируются несколькими типами триплетов, представленных разными сочетаниями нуклеотидов. Нуклеотидные замены, которые не приводят к изменению аминокислоты в полипептиде, называются *синонимичными*, они часто являются молчащими мутациями. Если же замена нуклеотида является причиной замены аминокислоты, то она называется *несинонимичной*.

Гены сильно различаются по своему размеру и организации экзонов и интронов. Некоторые гены совсем не имеют интронов, например, гены, кодирующие *гистоны*. Гистоны являются белками, принимающими участие в организации ядерной ДНК и эпигенетической регуляции экспрессии генов. Некоторые гены имеют довольно большой размер, например, ген, кодирующий у человека белок *дистрофин*, нарушение структуры которого является основной причиной мышечной дистрофии Дюшенна. Его размер составляет 2,3 миллиона пар нуклеотидов (Mb), а транскрипция этого гена занимает 16 часов, в результате чего образуется белок длиной почти в 3,7 тысячи аминокислот. Всего этот ген содержит 79 экзонов, при этом более 99% длины дистрофина занимают интроны. Известно, что гены с высокой экспрессией обычно имеют короткие интроны. Большинство экзонов - небольшого размера, в среднем 200 п.н.

Размеры интронов варьируют от десятков до миллионов пар оснований.

Рисунок 1.2.1.2. Скульптура «Танец рибосом», расположенная на территории лаборатории Колд Спринг Харбор, в штате Нью-Йорк. На скульптуре изображены рибосомы, состоящие из двух субъединиц, движущиеся по матричной РНК. Фото В. Брюхина

Похожие гены со схожими биохимическими функциями часто объединяются в *семейства генов*, образованные за счет дупликации исходного одиночного гена. Примером может быть семейство генов, кодирующих гемоглобин человека, состоящее из десяти генов, локализованных на разных хромосомах, образующих два локуса, α-глобина и β-глобина. Эти гены находятся под общей регуляцией. Обычно гены одного семейства состоят из похожих нуклеотидных последовательностей, но при этом независимы друг от друга, кроме того, они могут содержать общие консервативные локусы в виде *мотива* или *домена*. При этом мотив не является стабильной единицей, тогда как домен - стабильная и компактная структура, отвечающая за функцию кодируемого белка. Мотив может быть частью домена, тогда как домен никогда не бывает частью мотива.

Псевдогены — это, как правило, дефектные копии белковых генов. Они содержат большую часть последовательности гена, но в их кодирующей части встречаются стоп-кодоны, либо наблюдаются сдвиги рамки считывания, либо у них отсутствуют промоторы, либо они усечены,

или являются просто фрагментами генов. Дуплицированные псевдогены могут быть результатом тандемной дупликации оригинальных одиночных генов или следствием перемещения мобильных элементов. Глобальная дупликация геномов покрытосеменных растений, которая случилась 200-170 миллионов лет назад, привела к тому, что у многих растений гены находятся в удвоенном, а иногда утроенном, учетверенном и т. д. состоянии. Когда функциональный ген дублируется, одна его копия становится функционально не нужной. Иногда копия может приобретать новую функцию (например, гены, кодирующие β-глобин). В других случаях дополнительная копия будет инактивирована случайной мутацией и станет псевдогеном. Псевдогены не обладают очень большой продолжительностью жизни: как только область ДНК перестает функционировать, она быстро улавливает новые мутации и в конечном итоге становится неузнаваемой.

Вопросы:

1) Как устроены белок кодирующие гены?
2) В чем заключается генетическая транскрипция?
3) Что такое кодон?
4) Чем отличаются псевдогены от генов?

1.2.2. Гены, кодирующие РНК

Гены, кодирующие белки, транскрибируются РНК-полимеразой II, тогда как гены РНК транскрибируются с помощью РНК-полимеразы I и III. Наиболее известными из них являются гены рибосомной и транспортной РНК (рРНК и тРНК соответственно). Субъединицы рРНК транскрибируются с единиц генома, включающих до нескольких десятков тысяч копий генов, которые организуются в массивы тандемных повторов. Эти массивы характеризуются большим скоплением рибосом благодаря высокой активности рибосомальных генов и сборке рибосом на месте. Скопления рибосом в этом регионе образуют *ядрышки,* которые иногда называют областями ядрышкового организатора и которые при окрашивании цитологическими красителями хорошо видны под микроскопом. Гены транспортной РНК рассредоточены по геному, обычно они объединяются в небольшие кластеры. Всего существует 49 семейств генов тРНК.

Каталитические молекулы РНК (рибозимы) являются комплексом малой ядерной РНК (мяРНК) и белков, они участвуют в сплайсинге РНК (удалении интронов из мРНК при процессинге) и модификации оснований РНК. Также мяРНК принимают участие в регуляции некоторых транскрипционных факторов, поддержании целостности теломер

(концевых участков хромосом), инактивации Х-хромосомы, импринтинге и других важных процессах.

МикроРНК (миРНК) и малые интерферирующие РНК, будучи антисмысловыми РНК, комплементарными «смысловой» цепи мРНК, регулируют трансляцию специфических мРНК путем связывания с ними. миРНК играет важную роль в эпигенетической регуляции экспрессии генов. Использование малых интерферирующих РНК лежит в основе широко распространенного метода, называемого интерференцией РНК, который позволяет инактивировать определенные гены.

Вопросы:

1) В чем особенность генов, кодирующих РНК?
2) Какова роль микроРНК?

1.2.3. Мобильные генетические элементы и повторяющиеся последовательности

Мобильные элементы (МЭ), или *транспозоны (ТЕ)*, можно назвать «эгоистичными генами», поскольку при активации они встраивают свои копии в геном эукариот (рис. 1.2.3.1). Особенно распространены траспозоны в геномах растений [6]. Если встраивание происходит в геномы половых клеток, тогда образующиеся вставки передаются следующему поколению. ТЕ бывают двух типов: 1-типа -ретротранспозоны, передающиеся с помощью РНК-интермедиатов, которые встраиваются в геном за счет обратной транскрипции; 2-типа - ДНК транспозоны. ТЕ 1-го типа различаются по своей организации и классифицируются на три основных семейства, согласно механизму, с помощью которого они копируют себя: *LINEs* (Long Interspersed Nuclear Elements, длинные чередующиеся элементы); *SINEs* (Short Interspersed Nuclear Elements, короткие чередующиеся элементы); *LTR* элементы (Long terminal repeats, длинные концевые повторы), производные ретровирусов. LINEs являются одними из самых распространенных мобильных элементов размером от 6 до 8 тысяч нуклеотидов, в среднем занимая свыше 20% генома (рис. 1.2.3.1). LINEs могут реплицироваться автономно, в то время как SINEs не являются автономными транспозонами, их воспроизведение зависит от белков LINEs. При этом SINEs занимают второе место по представленности и в среднем составляют 13% генома. LTR-ретротранспозоны могут воспроизводиться как автономно, так и неавтономно: воспроизводящиеся автономно LTR являются самыми большими мобильными элементами размером от 6 до 11 тысяч нуклеотидов, составляя в среднем 8% генома.

В геномах ретротранспозонов содержатся три основных гена, унаследованных от ретровирусов: *Gag, Pol* и *Env* (рис. 1.2.3.1). *Gag* (*Group*

Antigens (ag)) представляет собой полипротеин, который поступает в цитоплазматический матркс и является частью основных белков ретровируса. *Pol* - кодирует обратную транскриптазу, а также интегразу и РНКазу Н, которая функционирует во время обратной транскрипции генома. Ген *Env* в ретровирусах кодирует белок оболочки, который находится в липидном слое и определяет вирусный тропизм, в LTR транспозонах *Env* является нефункциональным либо отсутствует [7].

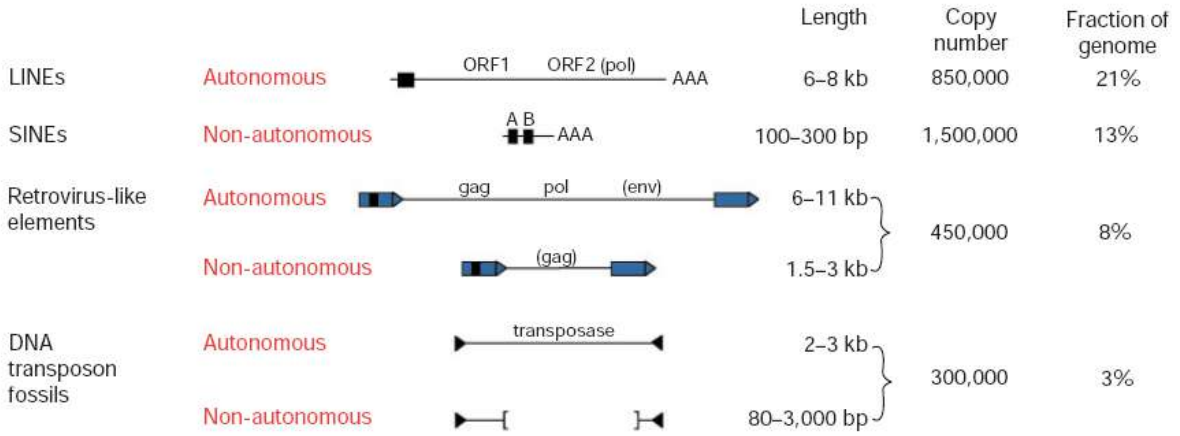


Рисунок 1.2.3.1. Схематическое изображение четырех типов транспозонов (TE): 1-типа LINEs, SINEs и LTRs и 2-типа ДНК траснпозоны.

ДНК-транспозоны перемещаются по геному посредством механизма вырезания-инсерции либо за счет репликации. При этом они делятся на автономные, транспозиция которых происходит независимо, и неавтономные, которым для перемещения нужен фермент *транспозаза*. Размер автономных транспозонов - 2-3 тысячи пар нуклеотидов, тогда как длина неавтономных транспозонов бывает от 80 до 3 тысяч п.н. ДНК транспозоны в среднем составляют 3% эукариотического генома (рис. 1.2.3.1). Особенностью транспозонов, которые перемещаются за счет вырезания-встраивания, является их фланкирование инвертированными повторами (Terminal Inverted Repeats, TIR). В отличие от ретротранспозонов, ДНК-транспозоны обычно сами вырезаются из генома и сами же повторно встраиваются в новое место. Источником транспозазы для неавтономных ДНК-транспозонов могут быть другие ДНК транспозоны, например *Ac* (activator) элементы кукурузы, которые способствуют мобилизации транспозонов *Ds* (dissociator) с последующим встраиванием последних в новую локацию. Было показано, что элементы *Ac-Ds* системы транспозонов, изначально обнаруженные в кукурузе, при биоинженерном переносе работают и на других растениях, включая табак,

арабидопсис и рис. Такие системы используются в научных целях, например, для генерации транспозированных линий растений арабидопсиса с целью точечного накаута генов и изучения их функции [8].

Вопросы:

1) Что такое транспозоны?
2) Какие типы транспозонов бывают?
3) Как устроены транспозоны?

1.2.4. Митохондриальная и пластидная ДНК

Помимо ядерной ДНК, содержащейся в хромосомах, в эукариотических клетках ДНК также содержится в *митохондриях (мтДНК),* клеточных органеллах, расположенных в цитоплазме, которые преобразуют химическую энергию питательных веществ в доступную для клеток форму *аденозинтрифосфата (АТФ)*. Совокупность этой ДНК также называют *митохондриальным геномом*. Последний намного меньше ядерного генома, так, например, митохондриальный геном человека

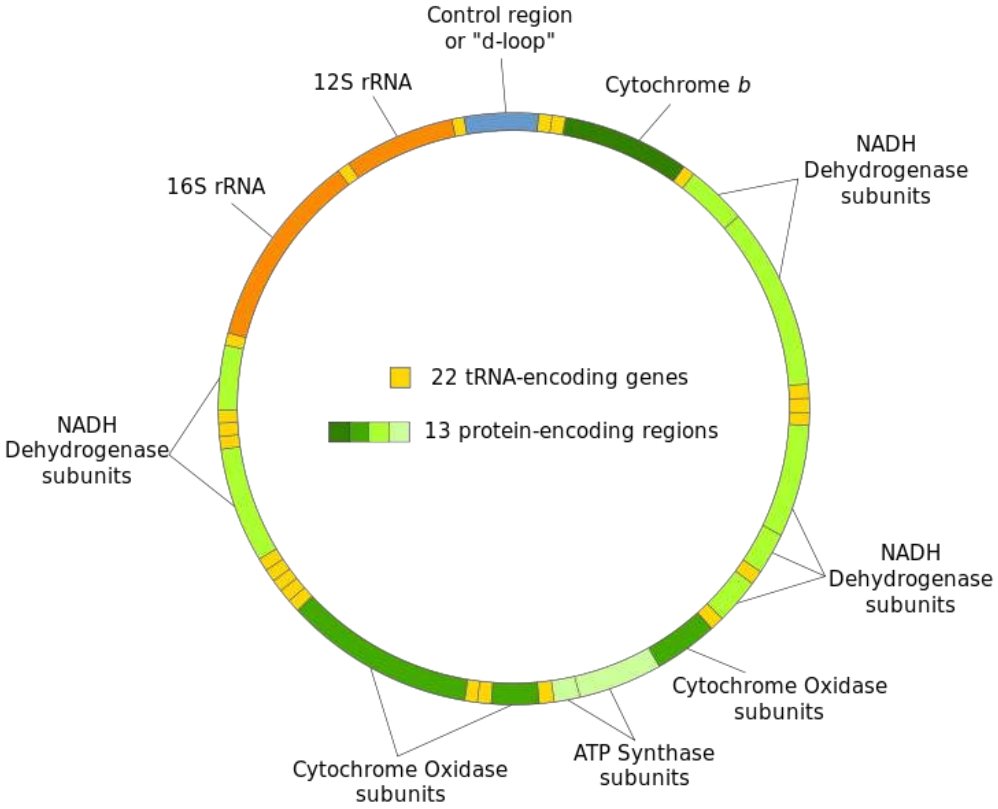


Рисунок 1.2.4.1. Схематическое изображение митохондриального генома человека. (Из Википедии)

состоит из 16 569 п.н., по сравнению с 3,2 миллиардами п. н. ядерного генома. Митохондриальный геном человека содержит всего 37 генов, тогда как в ядерном геноме содержится более 21 тысячи генов (рис. 1.2.4.1). 23 гена митохондриального генома человека кодируют тРНК, а остальные 14 генов кодируют ферменты, необходимые для дыхания клетки. мтДНК происходит от кольцевых геномов бактерий, поэтому у всех эукариотических организмов она имеет кольцевую форму. Поскольку мтДНК располагается в цитоплазме клеток, она наследуется исключительно по материнской линии. Поэтому ее можно использовать для идентификации происхождения по женской линии. В связи с тем, что мтДНК животных эволюционирует быстрее ядерного генома, она представляет собой основу для исследований по филогенетике, эволюционной биологии, а также антропологии и биогеографии, позволяя идентифицировать степень родства популяций.

Пластиды являются внутриклеточными органеллами, свойственными только клеткам растений. Они подразделяются на несколько типов, в зависимости от выполняемой функции. Наиболее известными пластидами являются зеленые *хлоропласты*, принимающие участие в фотосинтезе; *хромопласты*, накапливающие красные, желтые и оранжевые ферменты, благодаря которым осенью окрашиваются листья; *амилопласты*, накапливающие гранулы крахмала и встречающиеся в запасающих органах растений. Пластиды, как и митохондрии, передаются по материнской линии через цитоплазму яйцеклетки. Хотя у некоторых голосеменных, таких как саговники и гинкго, передача пластид происходит по мужской линии. Поскольку пластиды в процессе эволюции, подобно митохондриям, вероятнее всего, образовались в результате поглощения древней цианобактерии клеткой-«хозяйкой», пластидный геном по структуре похож на геном цианобактерии и состоит из кольцевой двуцепочечной ДНК. Размер пластидной ДНК колеблется от 75 до 290 тыс. п. н. [9]. На рисунке 1.2.4.2 изображена схема генома хлролпластов растения из рода *Boechera,* собранного сотрудниками группы геномики растений ИТМО. Как и у большинства покрытосеменных, в геноме *Boechera* присутствуют два инвертированных повтора IR_L и IR_R, которые делят геном на две неравные части. Во фрагментах инвертированных повторов содержатся гены, кодирующие рРНК. Некоторые гены организованы в опероны, то есть в группы генов, экспрессирующиеся под общим промотором. Хлоропластная ДНК, изображенная на рис. 1.2.4.2, имеет размер 149331 п.н. и содержит 124 гена, кодирующих белок, а также 39 генов, кодирующих тРНК (группы этих генов выделены на рисунке различными цветами, а их функция указана во вкладке к рисунку).

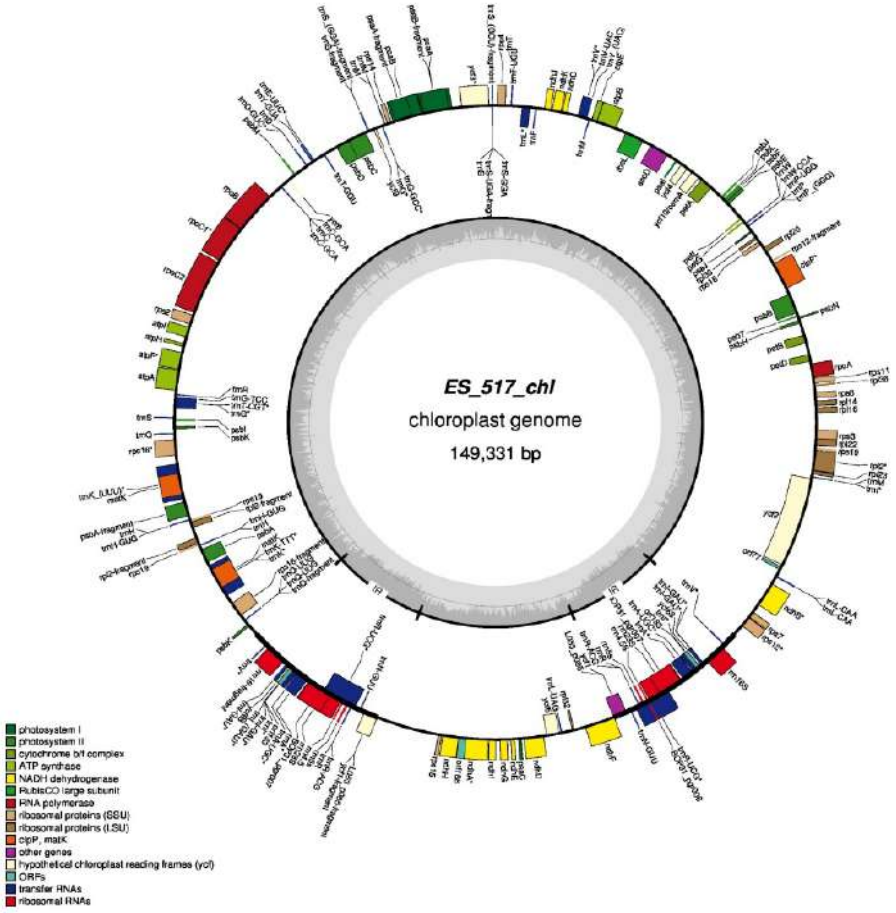


Рисунок 1.2.4.2. Схема генома хлоропластов растения из рода *Boechera*. Разным цветом выделены фрагменты хлоропластной ДНК, содержащие гены, отвечающие за разные функции.

Фрагмент хлоропластной ДНК, называемй *trnL /F* локус (размером 1 тыс. п.н.) носит консервативный характер, таким образом, этот регион стал одним из наиболее широко используемых маркеров хлоропластов для филогенетического анализа растений [10]. Этот регион включает ген транспортной РНК *trn-L*, *trn-L / F* межгенный спейсер и первые 18 п.н. гена транспортной РНК *trn-F*. Пример филогенетического дерева для видов из рода *Boechera*, построенного на основе вариабельности *trnL / F* локуса, представлен на рис. 1.2.4.3.

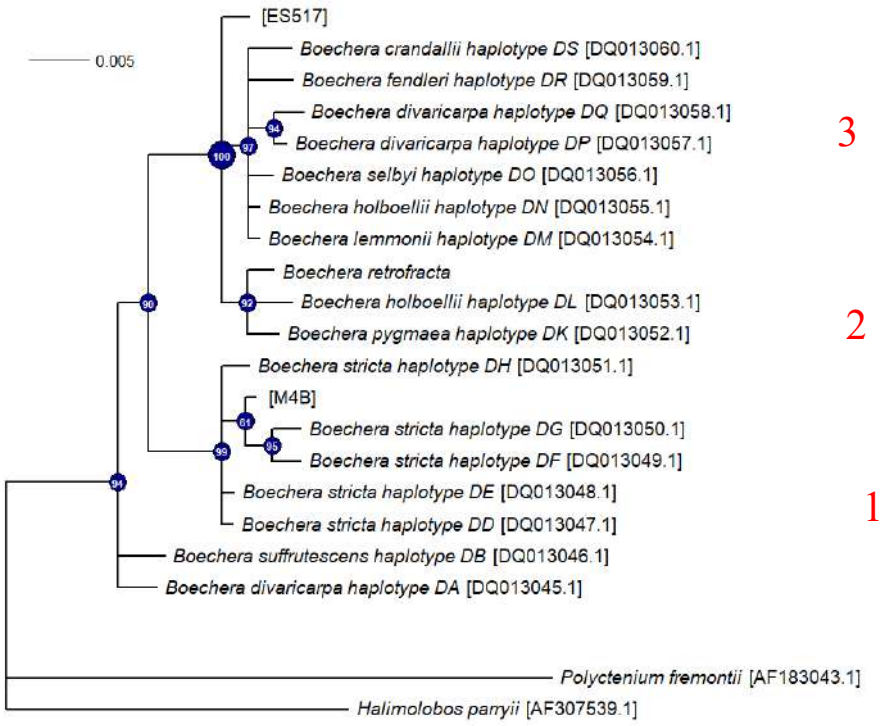


Рисунок 1.2.4.3. Филогенетическая характеристика хлоропластной ДНК видов *Boechera* на основе *trnL /F* локуса. Выделяются три клады согласно генетической вариации этого локуса в пластидной ДНК.

Вопросы:

1) Как возникли митохондриальный и пластидный геномы, в чем их отличие от ядерного генома?
2) Что такое гаплогруппы и как их используют ученые?
3) Как используется пластидный геном для филогении?

1.2.5. Геном человека

Структура генома человека была определена благодаря его секвенированию в рамках международного проекта «Геном человека», который продолжался с 1990 по 2000 год, результаты проекта были опубликованы в 2001-2003 годах. Дальнейшие технологические достижения в области секвенирования следующего поколения (NGS) в последнее десятилетие способствовали быстрому развитию геномики человека, которое началось с расшифровки индивидуальных геномов известных ученых Крейга Вентера и Джеймса Ватсона и далее способствовало выполнению глобальных исследований отдельных геномов в сравнительном аспекте, лучше всего представленных в международном

проекте «1000 геномов» [11,12], первом проекте, который исследовал разнообразие геномов человека на основе изучения 26 популяций человека, проживающих в разных частях Земного шара. Основными целями проекта было обнаружение генетических вариантов, являющихся уникальными для определённых этнических и региональных групп населения планеты; изучение генетических вариантов, влияющих на частоту отдельных заболеваний в популяциях людей; интерпретация обнаруженной вариабельности геномов для расшифровки исторических путей миграции и оседлостей отдельных популяций человека. Позже внимание ученых, изучающих геном человека, переключилось на национальные геномные проекты. Такие проекты были выполнены и продолжают реализовываться в десятках стран мира в Европе, Азии, Австралии и Америке. Все они обеспечивают создание глобального справочного ресурса, содержащего информацию о генетических вариациях человека, и предоставляют дорожную карту и возможности для открытия новых вариантов генома, ассоциированных с различными заболеваниями [13]. На рис. 1.2.5.1 представлен один из результатов проекта «Российские геномы» по исследованию генетического родства некоторых российских популяций и сравнению их геномной структуры с уже исследованными мировыми популяциями людей [14]. Проект реализовывался нынешними сотрудниками ИТМО, ранее работавшими в Центре геномной биоинформатики им. Ф.Г. Добржанского под руководством профессора Стива О'Брайена.

Определенные заболевания и наследственные признаки в разных популяциях встречаются с различной частотой благодаря генетическому дрейфу, адаптации и миграции. Частота вариантов, характерная для конкретной популяции, может привести к различиям в сценарии или распространенности заболевания в популяциях, что может повлиять на индивидуальное клиническое лечение, специфичное для определенных групп населения. На сегодняшний день исследования геномов различных групп россиян на наличие и частоту клинически значимых вариантов не проводились. Тем не менее, проект «Российские геномы», с участием одного из авторов настоящего пособия, сделал первый шаг для создания индивидуального подхода в геномной медицине для российских популяций [14]. Чтобы проиллюстрировать, как различия в истории популяции могут влиять на частоту важных физиологических признаков, были подробно изучены четыре генетических локуса: *MCM6, VCORC1, SLC45A2* и *DHDDS* (рис. 1.2.5.2). Ген *LCT- MCM6*, регулирующий толерантность взрослых к лактозе и молочным продуктам, является хорошо известным примером дифференциации на основе отбора. Мутация, приводящая к устойчивости людей к молочному сахару лактозе, возникла в гене *LCT* 7-10 тыс. лет тому

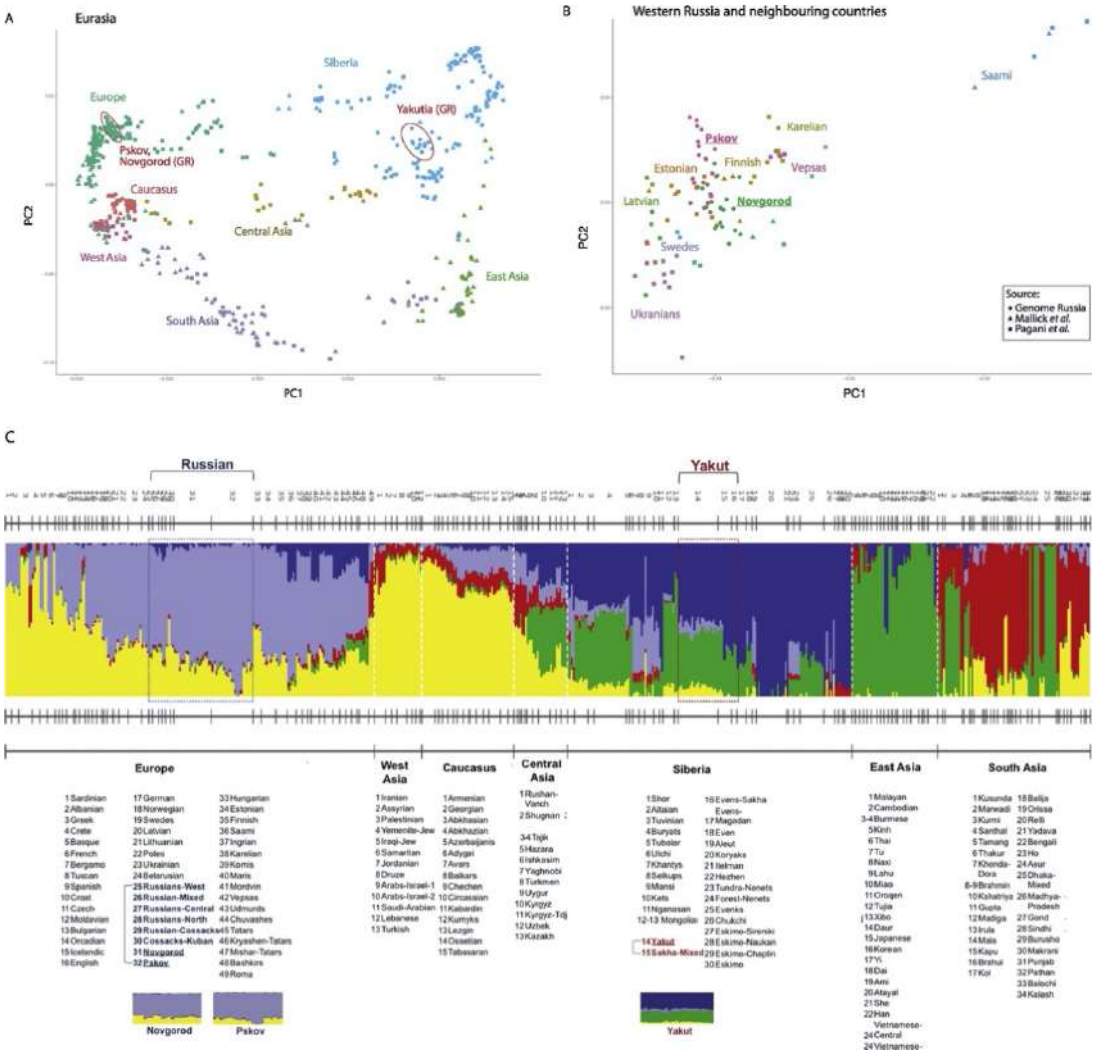


Рисунок 1.2.5.1. Родство популяций, основанное на данных генотипа. (A, B) Графики главных компонент 574 российских геномов. Цвета соответствуют географическим регионам сбора образцов. Красные овалы показывают расположение образцов, собранных по проекту «Российские геномы». (A) Евразия; (B) Западная Россия и соседние страны. (C) Структура населения по выборкам из 178 популяций из пяти основных географических регионов (k = 5). Выборки объединены в три различных исследования, охвативших территорию Российской Федерации (Mallick et al., 2016, Pagani et al., 2016, проект «Российские геномы»). Оптимальное значение k было выбрано по значению ошибки перекрестной проверки. Российские образцы из всех исследований (выделены темно-синим цветом) показывают небольшой градиент, направленный от структур Восточной Европы (украинская, белорусская, польская) к североевропейской (эстонская карельская, финская), что отражает историю продвижения населения на север. Якутские образцы (выделены красным) также показывают небольшой градиент от монголов к сибирскому народу, эвенам, как и ожидалось, благодаря их первоначальному смешиванию с этими популяциями и постепенному продвижению на север. Образцы, взятые из проекта «Российские геномы», выделены и помещены в отдельные прямоугольники внизу рисунка [14].

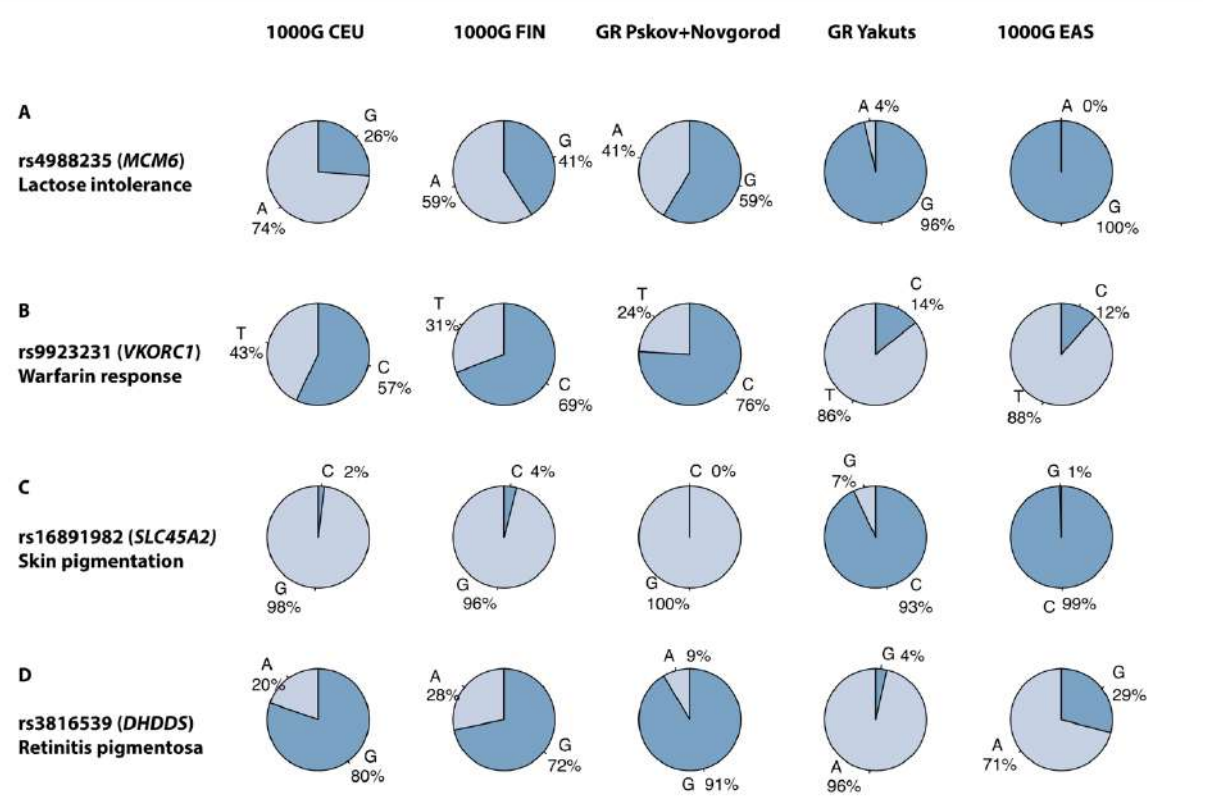


Рисунок 1.2.5.2. Различия в частотах аллелей, важных для физиологии и здоровья человека генов (различающихся по однонуклеотидному полиморфизму (SNPs)), и их дифференциация у евразийских этнических групп. Частоты аллелей для европейцев (CEU), финнов (FIN), объединённой популяции жителей Псковской и Новгородской областей, Якутии и жителей Юго-Восточной Азии (EAS) для четырех SNPs этих генов: (A) rs4988235, расположенный в гене *MCM6*. Этот SNP связан с непереносимостью лактозы у взрослых. Аллель G обозначает гаплотип непереносимости лактозы; (B) rs9923231, расположенный в гене *VKORC1*. Этот SNP связан с реакцией на антикоагулянт непрямого действия *варфарин*. Носители аллели T нуждаются в сниженной дозе варфарина; (C) rs16891982, расположенный в гене *SLC45A2*. Аллель G связана с более светлой пигментацией кожи; (D) rs3816539, расположенный в гене *DHDDS*. Аллель A связана с заболеванием пигментным ретинитом [14].

назад на Севере Европы в связи с развитием молочного животноводства, в то время как у кочевых азиатских народов такая мутация отсутствовала. Как видно на рис. 1.2.5.2, северные и европейские популяции намного более толерантны к лактозе, чем якутские и азиатские популяции [14]. Другой пример популяционных различий в частоте встречаемости аллелей, важных для прогнозирования дозировки медицинских препаратов, — это реакция на варфарин. Варфарин - популярный антикоагулянт, который имеет серьезные побочные эффекты, такие как кровотечение, при использовании

в неправильной дозировке. Ответ на варфарин зависит от нескольких факторов, включая генетические варианты генов *CYP2C9* и *VKORC1*, которые обычно используются для прогнозирования подходящей дозы. Носители аллели *VKORC1-T*, которая преобладает у представителей азиатских популяций, нуждаются в значительно более низких дозах варфарина, чем европейцы, где преобладает аллель *VKORC1-C*. Как и ожидалось, в популяциях Юго-Восточной Азии и якутов была выявлена более высокая частота встречаемости аллели *VKORC1-T* (86% и 88% соответственно) по сравнению с европейцами (43%) и финнами (31%). В двух русских популяциях с Запада России (Псковской и Новгородской областей) частота встречаемости аллели T (24%) чуть ниже таковой в финской популяции (рис. 1.2.5.2). Полученные данные означают, что, вероятно, дозировка варфарина для жителей Псковской и Новгородской областей должна быть аналогична дозировке для финнов, в то время как для якутов и для населения Восточной Азии эффективной будет более низкая дозировка [13]. Резкое расслоение популяций также очевидно и для частот аллелей гена *SLC45A2*, связанного с более светлой пигментацией кожи, и гена, кодирующего цис-пренилтрансферазу (*DHDDS*), мутации в котором являются причиной возникновения пигментной дистрофии сетчатки, поражающей глаза (рис. 1.2.5.2) [14]. Однако неясно, являются ли различия в частоте наследования аллелей *DHDDS* результатом генетического дрейфа или естественного отбора.

Итак, ядерный геном человека в гаплоидном состоянии содержит 3200 миллионов п.н. ДНК, помимо митохондриального генома, расположена в 46 хромосомах (23 парах гомологичных хромосом), при этом 22 пары называются *аутосомами*, а последнюю пару образуют *половые хромосомы X и Y*, в диплоидном состоянии сочетание XX обнаруживается в геномах у женщин, а XY у мужчин. Белок кодирующих генов в геноме человека насчитывается порядка 21 тысячи, при этом они занимают менее 2% от всего генома. Хотя белков эти гены производят намного больше, благодаря альтернативному сплайсингу. Остальные более 98% генома составляют повторы (повторяющиеся нуклеотидные последовательности); тандемные повторы, подразделяющиеся на сателитную ДНК, мини- и микросателиты; описанные выше диспергированные повторы (LINEs и SINEs); LTR-содержащие ретротранспозоны, которые очень похожи на ретровирусы. Разница заключается в том, что у ретротранспозонов человека отсутствует ген *env*, который отвечает за синтез белковой оболочки и позволяет вирусу перемещаться за пределы клетки. LTR ретротранспозоны человека также называют *эндогенными ретровирусными последовательностями (ERV)*. Большинство копий *ERV* встроились в геном предков человека много миллионов лет назад и в современном геноме являются дефектными, то есть с мутированными или отсутствующими генами *gag* и *pol*. Однако некоторые транспозоны при определенных

условиях способны перемещаться по геному. Активация и перемещение транспозонов в некоторых случаях является причиной возникновения заболеваний, включая онкологические. *ERV* занимают около 1% человеческого генома, а все повторяющиеся последовательности в сумме составляют почти две трети генома.

Полиморфизм митохондриального генома человека, состоящего всего из менее чем 16,6 тысяч п.н., используется для выделения *митохондриальных гаплогрупп*, то есть схожих гаплотипов, которые произошли от общего предка по материнской линии, аналогично тому, как полиморфизм мужской Y хромосомы находит применение в выделении *Y-гаплогрупп*, происходящих от общего предка по мужской линии. Анализ гаплогрупп широко применяется в популяционной генетике и генетической генеалогии, изучающей историю происхождения человека и родство его различных популяций.

Одной из основных задач современной молекулярной биологии, генетики и компьютерной биологии (биоинформатики) является определение структуры генома на основе сборки и аннотации полногеномной последовательности нуклеотидов, открытие новых генов и определение их функции, выяснение сетей регуляции генов, а также эволюции различных элементов генома.

Вопросы:

1) Каков размер генома человека?
2) Для чего нужно полногеномное секвенирование людей?
3) Приведите пример различий частоты встречаемости вариантов у человека.
4) Что такое аутосомы?
5) Расскажите о мужских глаплогруппах, и как они используются в науке?

1.2.6. Особенности геномов растений

Зеленые растения, иначе по-латыни называемые Viridiplantae, насчитывают на Земле почти полмиллиона видов. Их эволюционная история продолжается уже почти миллиард лет. Размер растений варьирует от мельчайших одноклеточных фотосинтезирующих организмов, таких как хламидомонада, до огромных как в высоту, так и по массе деревьев - как, например, гигантская секвойя (лат. *Sequoidendron giganteum*). Геномы растений являются ключом к пониманию эволюционной истории растений, которой занимается наука *молекулярная филогения,* основным объектом изучения ее являются геномы, транскриптомы и протеомы различных видов в сравнении между собой. Данные по анализу геномов одновременно

раскрывают загадки о прошлой жизни растений и являются мощным инструментом прогнозирования как для фундаментальной, так и для прикладной науки о растениях. Результаты исследований по геномике и молекулярной филогении направляют наши усилия на улучшение сельскохозяйственных культур, открытие новых лекарств, поиск перспективных источников для создания биотоплива, растительных волокон, продуктов питания и разработку эффективных стратегий сохранения видов. Геномы растений демонстрируют удивительное разнообразие своих размеров, состава и сложности. Вероятно, это связано с разнообразием форм и функций растений, учитывая, что, в отличие от животных, большинство растений ведут прикрепленный образ жизни, не способны поддерживать самостоятельно свою температуру и произрастают в самых разных экологических нишах от полярных тундр до тропиков, от высокогорий до морских глубин. Однако связь межу особенностью геномов и разнообразием форм и функций растений все еще остается плохо изученной. Геномы, как древние книги, несут в себе информацию об эволюционной истории, полногеномных дупликациях, процессах былых адаптаций, популяционных процессов и других событий в *филогенезе* (историческом развитии) растений. Поскольку постгеномная эра только начинается, мы пока что лишь только учимся читать эту историческую информацию, используя соответствующие генетические маркеры. Насколько впечатляюще фенотипическое разнообразие растений, настолько широким является разнообразие их геномов, то есть их размеров, содержания и структуры. Самый маленький геном эукариотического растения был обнаружен у одноклеточной зеленой водоросли, являющейся частью океанического планктона, *Ostreococcus tauri*, и составлял всего ~ 12 миллионов п.н. (Mb). Размер этой водоросли меньше 1 микрона (мкм), она содержит всего одну митохондрию и один хлоропласт.

У покрытосеменных растений размеры генома различаются между собой почти в 2000 раз: от самого маленького генома размером в ~ 82 миллионов п.н. (Mb) (2n = 28) у пузырчатки *Utricularia gibba*, рода плотоядных растений (семейство Lentibulariaceae) (рис. 1.2.6.1 A), до 149 миллиарда п.н. (Gb) (2n=40) хромосом у однодольного растения, японского вороньего глаза *Paris japonica* (семейство Melanthiaceae) (рис. 1.2.6.1 B) Также крупными геномами характеризуются папоротникии и голосеменные растения. Средний размер генома папоротника составляет ~ 12–14 Gb (диапозон от 260 Mb до146.5 Gb). Средний размер генома голосеменных составляет ~ 18 Gb, особенно большие геномы наблюдались у хвойных растений. Один из наиболее сложных геномов наблюдается у сосны ладанной (*Pinus taeda*), размер генома 22 Gbp.

Рисунок 1.2.6.1. Растения с экстремальным размером генома среди покрытосеменных = цветковых растений (Angiospermae). A) Пузырчатка, *Utricularia gibba*, рода плотоядных растений (Lentibulariaceae), имеющее один из самых маленьких геномов ~ 82 Mb; B) Японский вороний глаз, *Paris japonica* (Melanthiaceae), имеющее самый большой геном ~149 Gb.

Многие растения имеют очень большой и сложнй геном, а также обычно гораздо более высокую плоидность, более высокие показатели гетерозиготности и повторяющихся элементов, чем представители других царств. Сборка высоко гетерозиготных геномов de novo, то есть из коротких фрагментов (~ 100 п.н.), полученных путем полногеномного секвенирования методами NGS, без использования эталонного (референсного) генома, на который эти фрагменты (риды) могут быть выровнены, представляет собой нетривиальную и сложную задачу, требующую использования специальных алгоритмов и биоинформатических программ. В нашей группе Геномики растений ХимБио кластера ИТМО мы собираем до хромосомного уровня de novo геном высоко гетерозиготных растений из рода *Boechera*, которые размножаются с помощью *апомиксиса*, то есть бесполым путем посредством семян, проблемы такой сборки обсуждаются в обзоре Brukhin et al. 2019 [15]. Мы рассмотрим некоторые из методов сборки геномов de novo позже в разделе Биоиформатические методы исследования генома. Проблемы со сборкой геномов растений связаны в том числе и с их большой фрагментацией, которая возникает благодаря огромному количеству повторов. Особую трудность при сборке геномов представляют высокоповторяющиеся последовательности, часто размером > 10 тыс. п.н. Также сложность растительных геномов связана и с активностью мобильных генетических элементов, особенно преобладают у растений

ретротранспозоны с длинными концевыми повторами (LTR), которые составляют от 15% до 90% генома. У кукурузы 49-78 % генома состоит из ретротранспозонов, а у пшеницы около 90 % генома представлены повторяющими последовательностями, из них 68 % — подвижными генетическими элементами. У млекопитающих практически половина генома (45-48 %) состоит из транспозонов или остатков транспозонов. Примерно 42 % генома человека состоит из ретротранспозонов, и около 2-3 % — из ДНК-транспозонов. Полиплоидия является еще одной проблемой при сборке геномов растений, ~ 80% исследованных растений являются полиплоидами.

Как уже упоминалось выше, геном растений, помимо ядерного и митохондриального, включает в себя также и пластидную ДНК, что является важным отличием царства растений от других царств живых организмов. ДНК-маркеры пластидного генома успешно используются для изучения филогении растений уже свыше 35 лет. Обилие пластидной ДНК в клетках растений, ее консервативная структура и типичное наследование пластидного генома по материнской линии обеспечивают аналитические преимущества для его использования в молекулярно-филогенетических исследованиях. Филогенетическое древо, основанное на пластидных генах, обычно представляет *матрилинейные отношения*, в отличие от генеалогического дерева, построенного на основе вклада обоих родителей, то есть ядерного генома. Помимо описанного ранее *trnL /F* локуса хлоропластной ДНК, для *баркодирования ДНК*, то есть молекулярной идентификации растений, используют еще один ген *rbcL*, который кодирует большую субъединицу самого распространенного фермента на Земле, *рибулозо-1,5-бисфосфаткарбоксилазы/ оксигеназы*, также называемого *RuBisCO* (рис. 1.2.6.2).

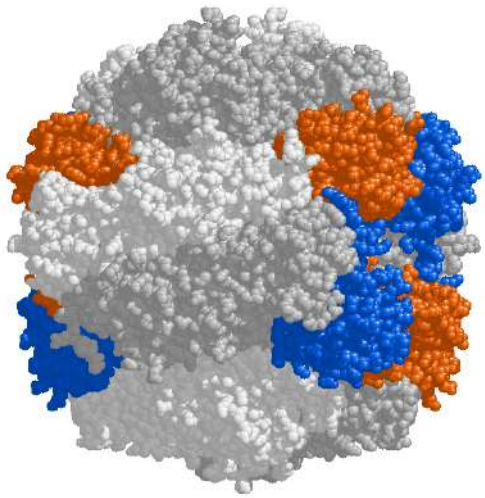


Рисунок 1.2.6.2. Модель молекулы рибулозо-1,5-бисфосфаткарбоксилазы / оксигеназы (RuBisCO), самого распространенного белка на Земле, состоящего из двух больших (выделено серым и белым) и двух малых (выделено синим и красным) субъединиц. (из Wikipedia)

Современные выводы о филогении растений, во многом основаны на анализе генома пластид. Полученные таким образом знания служат ориентиром для широкомасштабного понимания диверсификации растений и для пересмотра классификации покрытосеменных растений. Почти 35-летний период исследований в области молекулярной филогенетики пластид позволил прийти к заключению, что многое из того, что ранее было известно о филогении растений, оказалось неверным [16].

До начала 90-х годов 20 века ученые в основном работали с отдельными генами и группами генов, полезную информацию также давала кариология (наука, изучающая морфологию и поведение хромосом во время клеточных делений и интерфазы), но, тем не менее, целостного понимания устройства, функционирования и эволюции геномов не было. Для этого необходимо было получить полногенеомную последовательность, WGS (Whole Genome Sequencing) для максимального количества растений. Но расшифровка всего генома 30 лет назад была безумно дорогим мероприятием, требовавшим, кроме финансов, большого количества высококвалифицированных специалистов из разных областей и сложной современной техники. Первый многонациональный координированный проект по расшифровке генома популярного модельного растения *Arabidopsis thaliana* (Brassicaceae) (рис. 1.2.6.3) начался в 1990 году и продлился более 10 лет. Первые результаты этого проекта были опубликованы в журнале Nature в 2000 году [17]. Основными из которых явилась сборка и предварительная аннотация ДНК всех 5 хромосом (2n=10) арабидопсиса, выяснлось, что размер его генома составляет 125 млн. п.н. (Mb), секвенированные области охватывали 115,4 млн. п.н. и простирались до центромерных областей. Сообщалось также, что эволюция арабидопсиса включала дупликацию всего генома с последующей потерей ряда генов и обширными локальными генетическими дупликациями, что привело к возникновению динамического и сложного генома. Результаты проекта выявили, что геном арабидопсиса содержит 25 498 генов, кодирующих белки из 11 000 семейств, что было аналогичным функциональному разнообразию геномов других отсеквенированных к тому времени многоклеточных эукариот, животных модельных объектов, мухи дрозофилы (*Drosophila)* и круглого червя *Caenorhabditis elegans*. Было выяснено, что *Arabidopsis* имеет множество семейств новых белков по сравнению с дрозофилой и *C. elegans,* но при этом в его геноме отсутствовали гены, кодирующие ряд семейств белков, находящикся в их геномах, что указывает на то, что некоторые семейства общих белков в геномах трех многоклеточных эукариот претерпели дифференциальное расширение и сокращение [17]. Расшифровка полного генома арабидопсиса обеспечила основу для более полного сравнения консервативных процессов у всех эукариот,

определения широкого спектра функций генов, специфичных для растений, и для разработки быстрых систематических способов идентификации генов для улучшения сельскохозяйственных культур.

Рисунок 1.2.6.3. Модельное растение *Arabidopsis thaliana* (Brassicaceae) и его соцветие

Вслед за расшифровкой первого генома модельного растения арабидопсиса последовали другие проекты, задачей которых было уже расшифровать и проанализировать геномы большого количества растений. Вот несколько недавно инициированных международных проектов в области генома растений: проект «1001 геном *Arabidopsis thaliana»* (исследование различий и сходств множества экотипов арабидопсиса); 1K Plant Transcriptomes Initiative (задача - исследование транскриптома, то есть экспрессии всех генов тысячи растений; проект инициирован университетом Флориды, США); Инициатива открытого зеленого генома (задача - расшифровка и сборка геномов растений, пригодных в качестве сырья для экологических видов биотоплива; проект, выполняемый под эгидой Объединенного института исследования генома Министерства энергетики США); 10K геномов растений (проект, инициированный китайским геномным институтом BGI), цель проекта состоит в секвенировании 10 тысяч геномов различных растений с целью объяснения геномного разнообразия и сравнительной геномики большого множества видов растений из различных частей мира.

Полногеномное секвенирование и филогенетический анализ на основе ядерного генома растений, выполненый, в том числе, и в процессе реализации упомянутых выше международных проектов, показал, что многие геномы возникли благодаря гибридизации и / или *интрогрессии* генов или фрагмента генома одного вида растений в другой. Расширения и

сокращения семейств генов могут способствовать изменению размера генома, количества генов и их функции. Такое расширение-сокращение в процессе эволюции стало драматическим для зеленых растений, поскольку сильно повлияло на адаптацию, вымирание видов и на биоразнообразие растений на планете. Кроме того, было показано, что мобильные генетические элементы могут запускать изменения в экспрессии генов, а также объясняют различия в размере генома, особенно больших геномов хвойных пород. Дупликация геномов растений (WGD – whole genome duplication) может оказаться полезной, поскольку обеспечивает организм дополнительной копией каждого гена, хотя впоследствии дупликаты могут сильно отличаться, часто возникшая копия становится псевдогеном, то есть спящим, нетранскрибируемым геном. Кроме того, дупликации способствовали возникновению *генетической избыточности* (редундантности, от англ. redundancy), то есть наличию в геноме множества генов, выполняющих одну и ту же функцию. Полиплоидия также приводит к генетической избыточности, она возникает в результате ошибок во время мейоза или митоза, либо мутаций, которые оказывают влияние на деление клеток. Хотя большинство событий полиплоидизации в природе быстро отсеивается селекцией. Природное явление апомиксиса (бесполого размножения семенами, при котором отсутствует мейоз, а задорыш генетически идентичен материнскому растению) помогает избегать триплоидного блока при мейозе и обеспечивает сохранение и воспроизведение полиплоидов в популяциях [15,18]. В настоящее время удвоение полного хромосомного набора организма обнаружено у 15% покрытосеменных и 31% папоротников, но цифры могут быть намного выше, когда появятся данные по многим другим, еще неизученным и неотсеквенированным видам. События удвоения генома и полиплоидизации не возникают случайным образом во времени: многие из них связаны с драматическими периодами геологических и климатических изменений, такими как массовое вымирание, которое произошло на границе мелового и палеогенового периодов, порядка 70-60 млн лет назад, а также в результате ледниковых эпох, возникавших несколько раз в геологической истории Земли. При этом в период глобальных изменений климата полиплоидия играла важную роль в реакции на стресс, быстрой адаптации и выживании, в то же время полиплоидия могла способствовать усилению реакции на стресс и приводить к вымиранию таксонов.

Исходя из всего сказанного в этом разделе, геномы можно рассматривать как драйверы эволюционных изменений. Однако важная информация для понимания эволюционной истории растений, движущих сил диверсификации растений и связи между генотипом и фенотипом до сих пор отсутствует. Кроме того, планка того, что следует считать «геномом», продолжает расти с появлением новых знаний и новых технологий исследования геномов. Тем не менее, геномы растений

остаются ключом к решению многих вопросов, имеющих важное значение для общества: понимание и прогнозирование реакции видов растений на изменение климата, обеспечение продовольственной безопасности для быстро растущего населения планеты, открытие новых источников лекарств и содействие эффективным стратегиям сохранения видов.

Вопросы:

1) Что отличает геномы растений от геномов животных?
2) У какого растения самый большой геном?
3) Назовите самый распространенный белок на планете. Где располагаются гены, которые его кодируют?
4) Самое популярное модельное растение. Почему оно широко используется в исследованиях?

1.2.7. Отличия генома эукариот и прокариот

Сложные геномы *эукариот* (организмы, в клетках которых есть ядро, окруженное ядерной мембраной) содержат в десятки раз больше ДНК, чем требуется для кодирования всех необходимых организму РНК и белков. При этом некодирующая ДНК включает в себя содержащиеся в генах интроны; регуляторные элементы генов; множественные копии генов, включая псевдогены; межгенные последовательности (*спейсеры*), рассеянные повторы и другие структурные элементы. Геномы эукариот состоят из двойного (соматические клетки) или одинарного (гаметы) набора линейных хромосом, ограниченных ядром, тогда как геномы *прокариот* в основном представляют собой одиночные кольцевые хромосомы, состоящей из двойной спирали ДНК, занимающей *нуклеоидную область* клетки и прикрепленной к плазматической мембране. Размеры прокариотических геномов обычно намного меньше геномов эукариот. В эукариотических клетках большинство генов включает в себя как экзоны, так и интроны, тогда как прокариотические гены интронов не имеют (рис. 1.2.7.1).

Хотя эволюционное происхождение интронов остается малопонятным, тем не менее, их наличие в генах эукариот обеспечивает альтернативный сплайсинг, при помощи которого один ген может образовывать несколько белков. Существует две основные теориии возникновения интронов: 1) теория *раннего появления интронов*, согласно

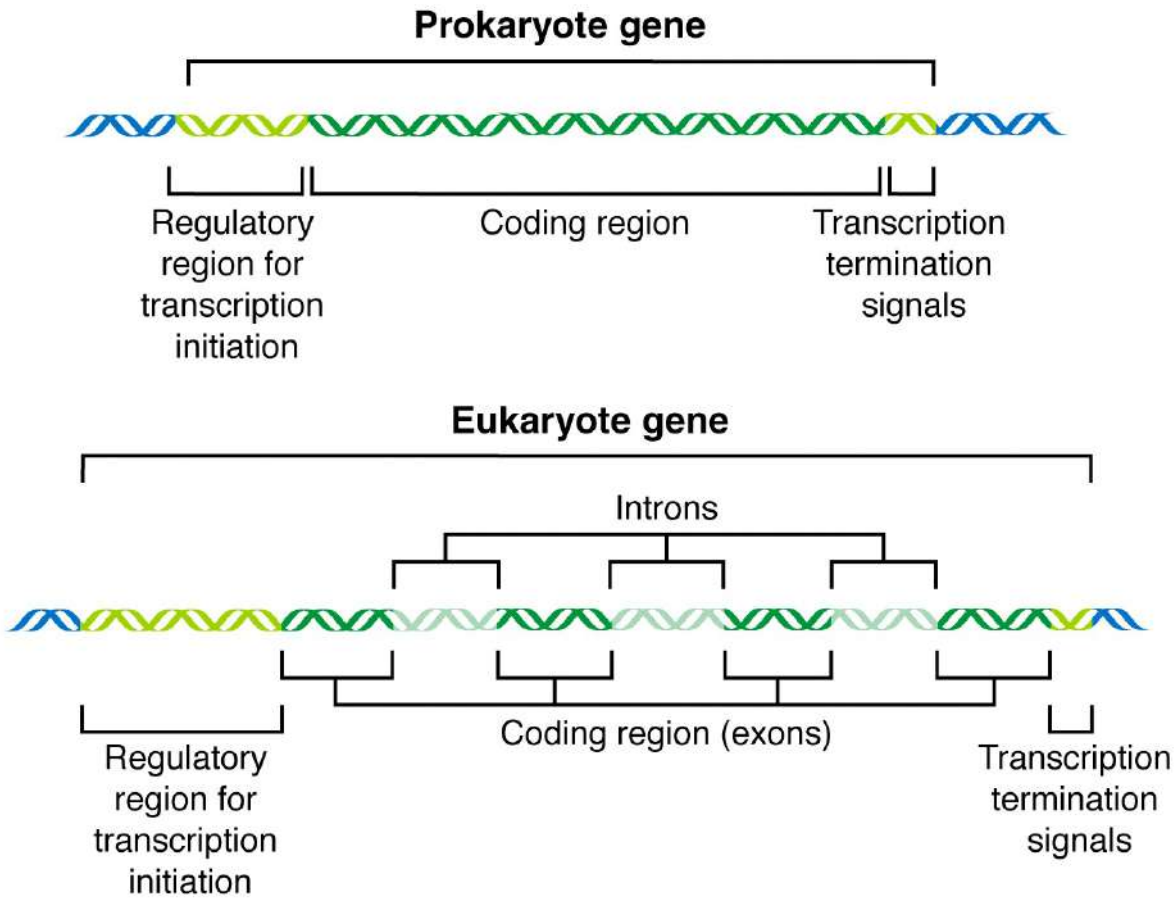


Рисунок 1.2.7.1. Структурное изображение генов прокариот (вверху) и эукариот (внизу).

которой интроны в современных организмах унаследованы от общего древнего предка и 2) теория *позднего появления интронов,* которая утверждает, что интроны появились путем вставок после возникновения прокариотов. До сих пор нет единого мнения, какая из этих гипотез наиболее верна. В настоящее время более широко распространено мнение, что интроны возникли в генах эукариот как «эгоистичные элементы», то есть генетические элементы, которые могут усиливать свою передачу в геномах, несмотря на то, имеет ли это положительный или отрицательный эффект для приспособленности организма. Одной из основных сил в эволюции прокариот является горизонтальный перенос генов. Отсутствие же появления интронов в прокариотах можно объяснить и тем, что передача сплайсосом, сложных частиц, состоящих из нескольких белков и РНК и необходимых для сплайсинга (удаления) интронов, не представляется возможной посредством горизонтального переноса. Еще одним свойством организации геномов прокариот является наличие *оперонов*. Последние представляют собой функциональные единицы ДНК, в которых несколько

генов (*цистроны*) работают под одним промотором и транскрибируются одновременно либо последовательно.

У прокариот гены часто расположены тандемно с небольшими по размеру разделительными последовательностями (спейсерами) между ними или без спейсеров, в то время как у эукариот между генами имеется значительная по размеру спейсерная ДНК. Некоторые из них — это повторяющаяся ДНК, идентичные или почти идентичные, диспергированные и тандемные повторы. Многие из спейсеров произошли от мобильных генетических элементов.

Вопросы:

1) В чем отличие геномов прокариот и эукариот?
2) Что такое интроны и как они произошли?
3) Что такое цистроны?
4) Зачем нужен сплайсинг?

1.2.8. Вирусные геномы

Вирусы представляют собой внеклеточную форму жизни. Они являются не просто разрушающими агентами, но важными компонентами глобальных экосистем. Вирусы — это облигатные внутриклеточные паразиты, не имеющие молекулярного механизма для своей репликации. Они состоят из белковой оболочки, *капсида*, и генома, представленного ДНК либо РНК. Средний диаметр сферического вируса ~ 30 нм. Палочка вируса табачной мозаики имеет размер 300 нм и диаметр 18 нм, содержит РНК-геном, состоящий из примерно 6400 нуклеотидов, окруженный оболочкой, состоящей из 2130 копий белка (рис. 1.2.8.1). Некоторые из нитчатых вирусов достигают длины почти 2000 нм = 2 мкм. Различное происхождение вирусных групп позволяет произвести их классификацию в пределах определенных типов, таких как вирусы, которые имеют геномы с положительной или отрицательной РНК- цепью, двуцепочечной РНК, а также одноцепочечные и двуцепочечные ДНК геномы.

Для примера распространенности вирусов с различными геномами 75% вирусов растений имеют геномы, состоящие из одной цепочки РНК (ssRNA), 65% из них имеют положительную цепь РНК, то есть с той же ориентацией, что и мессенджерная РНК, а 10% имеют отрицательную цепь РНК, то есть она должны быть преобразована в положительную цепь РНК,

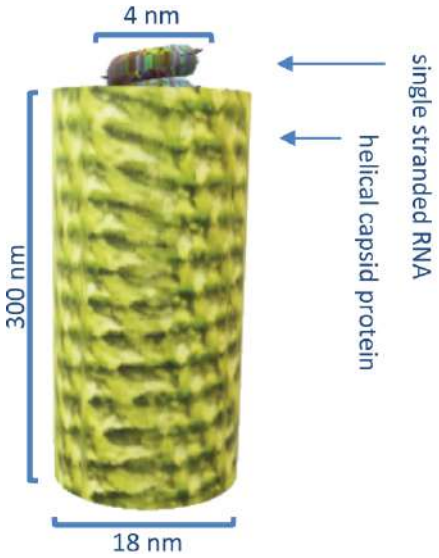


Рисунок 1.2.8.1. РНК-содержащий вирус табачной мозаики [19]

прежде чем ее можно будет транслировать. 5% вирусов содержат двуцепочечную РНК (dsRNA), она может быть напрямую транслирована, как и у вирусов с положительной цепью РНК. 17% растительных вирусов содержат геном из одноцепочечной ДНК (ssDNA), и очень немногие - из двуцепочечной ДНК (dsDNA). У животных вирусов четверть, а у бактериофагов (вирусов, поражающих бактерии) три четверти вирусов содержат геном, состоящий из двойной цепи ДНК. Вирусы используют рибосомы клетки-хозяина для производства обычно от 4 до 10 белков, кодируемых их геномом. Поскольку вирусные геномы, как и бактериальные, не содержат интронов, а гены транслируются подряд, многие из этих белков кодируются в одной цепи (*полицистронные* гены), то есть рибосома продуцирует только один белок, а трансляция прекращается на первом стоп-кодоне, расположенном в конце последнего белка. Потом происходит процессинг (нарезка) полипротеина.

Для трансляции эукариотической мРНК требуется наличие *5′ Cap-структуры* (7MeGpppN, где Me – метильная группа, N -аденин или гуанин). Вирусы кодируют белок, обычно репликазу с активностью метилтрансферазы, для осуществления трансляции. Некоторые вирусы являются похитителями 5' Cap-структуры. При этом 7MeG-capped мРНК хозяина используется вирусным транскриптазным комплексом и отщепляется эндонуклеазой, кодируемой вирусом. Полученная лидерная РНК служит для первичной транскрипции вирусного генома. Однако некоторые вирусы не используют Cap, а эффективно транслируются благодаря независимым от Cap усилителям трансляции, присутствующим в 5 'и 3' нетранслируемых областях вирусной мРНК.

В заключении можно отметить, что геномы вирусов бывают как линейными, так и кольцевыми. Кроме того, они крайне компактны с очень небольшими спейсерами между генами.

Вопросы:

1) Расскажите об особенностях строения вирусов?
2) Какие типы вирусов вам известны?
3) Что такое полицистронные гены?

1.3. Размеры геномов и хромосомы

Для каждого вида эукариот характерен определенный размер генома и стандартный хромосомный набор (таблица 1.3.1). Например, как уже было сказано, гаплоидный геном человека состоит из 3,2 миллиардов пар оснований (Gb), а ядра диплоидных клеток человека содержат 46 хромосом.

Таблица 1.3.1. Хромосомный набор клеток различных организмов

Организм	2n хромосом	Организм	2n хромосом
Цингерия (злак)	4	Рыбка Данио	50
Дрозофила	8	Лошадь	64
Абрикос, лук	16	Собака	78
Лисица, подсолнечник, кошка	38	Голубь	80
Ясень, человек	46	Карп	100
Шимпанзе	48	Краб камчатский	208

Еще с тех пор, когда данные по размерам генома и количеству хромосом у различных групп организмов не были доступными, ученых беспокоил вопрос: как соотносится размер генома и содержание в нем кодирующих белок генов с количеством хромосом в клетках, а также с эволюционной продвинутостью организмов? Казалось бы, человек, стоящий на вершине эволюционной лестницы, должен иметь наиболее сложный геном, большее количество генов и хромосом по сравнению с другими видами животных и растений. Однако, как видно из таблицы 1.3.1 и рисунка 1.3.1, это совсем не так. Человек занимает самое среднее положение среди живых организмов, принадлежащих к различным систематическим группам, как по размеру генома, так и по количеству

хромосом. Как видно из рис. 1.3.1, наибольший размах генома наблюдается у цветковых растений (зеленая полоса вверху), от десятков и сотен миллионов пар оснований (Mb) до сотен миллиардов пар оснований (Gb). Размер геномов большинства амфибий (полоса салатного цвета) и ряда рыб (голубая полоса) значительно превышает размер генома человека и других млекопитающих.

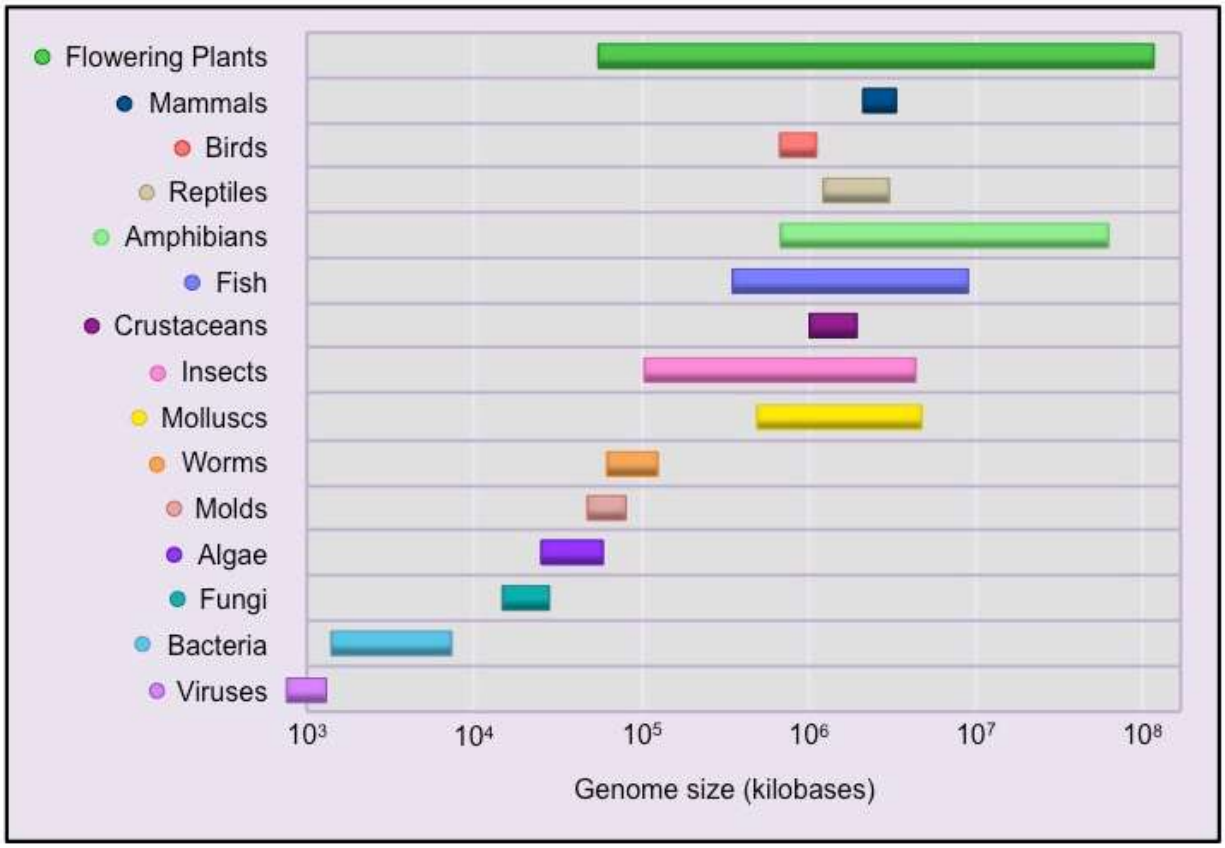


Рисунок 1.3.1. Вариабельность генома в различных группах организмов (в тысячах пар нуклеотидов) [20]

Как и ожидалось, наименьшим геномом обладают вирусы и бактерии. Размер вирусных геномов составляет всего несколько тысяч пар оснований, а бактериальных - от нескольких десятков тысяч до нескольких миллионов пар оснований.

Отсутствие связи между фенотипической сложностью эукариот и физическим размером их ядерного генома назвали *С-парадоксом*. С – это содержание ДНК в гаплоидном геноме, которое является постоянным для каждого вида. Геномы у наземных растений варьируют почти в 2000 раз, а у животных более, чем в 3000 раз. Кроме того, было обнаружено отсутствие корреляции между количеством кодирующих белок генов, размером генома и сложностью фенотипа организмов. Например, общий размер генома человека близок к геному ящерицы, а количество кодирующих

белок генов в геноме человека примерно равняется количеству генов в геномах модельных объектов, растения арабидопсиса и круглого червя *Caenorhabditis elegans* и соответствует 20-25 тысячам. Феномен отсутствия связи между количеством генов в геноме и сложностью фенотипа был назван *G-парадоксом*. Различается у видов и соотношение кодирующей белки ДНК к размеру всего генома.

Вопросы:

1) Сколько хромосом у человека?
2) У организмов какого царства самый большой разброс по размеру генома?
3) У какого растения геном содержит 2 хромосомы?
4) Что такое С-парадокс?

1.4. Модификация хроматина и метилирование ДНК

ДНК в хромосомах эукариотических организмов высоко упорядочена. В процессе ее компактизации имеет место разная степень упаковки. Первоначально нить ДНК «наматывается» на высоко консервативные белки *гистоны*, представляющие собой октамеры, образующие *нуклеосомы*, «бусинки» диаметром 11 нм. Эти «бусинки» разделены спейсерной последовательностью длиной 50 пар нуклеотидов. Существует четыре основных типа гистонов H2A, H2B, H3, H4. Вокруг каждого гистонового октамера оборачиваются около 165 нуклеотидов. Далее гистоновые октамеры с обернутой вокруг них ДНК сворачиваются во вторичную структуру, называемую *соленоид*, волокна диаметром 30 нм. В дальнейшем соленоиды упаковываются в волокна еще более высокого порядка, в итоге образуются хромосомы с участками разной степени компактизации ДНК. Набор хромосом одной клетки человека составляет примерно 1,8 метра ДНК, которая в компактном состоянии занимает всего 8-20 микрон. Каждый гистон имеет N-концевой хвост и С-конец. Оба этих важных компонента контактируют с ДНК посредством слабых взаимодействий, таких как водородные связи и солевые мостики. Взаимодействия ДНК и гистонов могут меняться, влияя на компактизацию ДНК и доступность транскрипционных факторов и РНК для белок кодирующих генов, регулируя таким образом, их транскрипцию.

Совокупность связанных с ДНК белков, в основном гистоновых, называют *хроматином*. Плотноупакованный хроматин, который хорошо окрашивается ядерными красителями традиционно называют *гетерохроматином*. ДНК в этих участках состоит из повторов или нетранскрибирующихся генов. Слабо конденсированные участки ДНК, плохо прокрашиваемые ядерными красителями, называются

эухроматином, в котором располагаются транскрипционно активные гены. Влияющая на транскрипционную активность генов плотность упаковки ДНК во многом определяется метилированием, фосфорилированием и ацетилированием N-концевых хвостов гистонов, что является одной из разновидностей *эпигенетической регуляции* экспрессии генов. Существует так называемый открытый (рыхлый) и закрытый (плотный) хроматин. Первый соответствует эухроматиновым, а второй гетерохроматиновым участкам. В таблице 1.4.1 указаны некоторые гистоновые маркеры с известной функцией для определения эталонного эпигенома. Активирующие и репрессивные гистоновые метки обычно расположены в начальной части генов. Наличие таких меток в геноме отражает эпигенетическую гетерогенность клеток. Баланс между ними определяет состояние экспрессии генов.

Таблица 1.4.1. Модификация гистонов, определяющая функцию эпигенома. H3, H4 обозначают гистон, который ацетилирован (ac) или метилирован (me) в положениях N (в данном случае N=9, 16, 4, 27) лизина (K), который расположен в N-хвостовой части гистона.

Модификация гистона	Функция
H3K9ac, H4K16ac, H3K4me	Состояние транскрипции "Включено"
H3K9me, H3K27me	Состояние транскрипции "Выключено"

Другим фактором, определяющим эпигенетическую регуляцию активности генов, может быть метелирование ДНК, а именно присоединение метильной группы к азотистому основанию цитозину в положении 5 гетероцикла (рис. 1.4.1). Метилирование в области генетического промотора или первого экзона часто меняет активность гена или группы генов, обычно приводя к подавлению их транскрипции.

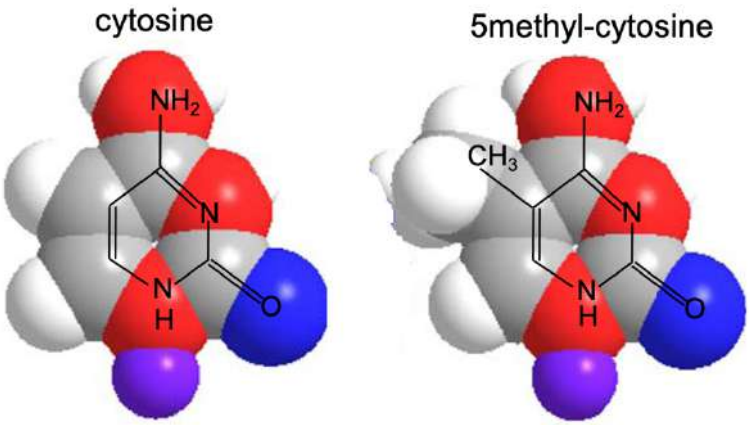


Рисунок 1.4.1. Нуклеотид цитозин и метилированный цитозин

Изучением эпигенетической регуляции экспрессии генов и фенотипа, вызванных механизмами, отличными от изменений в последовательности ДНК, занимается наука *эпигенетика*. Кроме того, она изучает каким образом паттерны экспрессии генов передаются от клетки к ее потомкам, как изменяется экспрессия генов при дифференцировке одного типа клеток в другой, каким образом факторы окружающей среды могут изменить способ экспрессии генов. Благодаря эпигенетическим модификациям ДНК и гистонов один геном может генерировать множество «эпигеномов». При развитии живых организмов эпигенетические механизмы постепенно ограничивают *тотипотентность* клеток (способность образовывать любые типы клеток), обеспечивая их специализацию и, в конечном итоге, формирование тканей и органов организма.

Яркий пример роли эпигенетической регуляции на переключение активности поведения может быть представлен функциональной дифференциацией пчел в улье. Как известно, в улье имеется одна матка, которая может производить потомство, а также тысячи рабочих пчел, нянек, сборщиц нектара, кормилиц и других неспособных к репродукции пчел. Когда в улье умирает старая матка, пчелы «выбирают» новую из новорожденных пчел и кормят ее так называемым «маточным молочком», производимы пчёлами-кормилицами в своей верхнечелюстной железе. Маточное молочко содержит, в том числе, и тиамин (B1), рибофлавин (B2), никотин, фолиевую кислоту и многое другое. Такое питание организма перепрограммирует его геном, в частности, и за счет сокращения активности фермента ДНК-метилтрансферазы (DNTM3), который катализирует перенос метильной группы на ДНК. В итоге несколько сотен генов матки отличаются по паттернам метилирования от этих же генов у рабочих пчел. В результате, помимо способности к размножению, матка может жить до 6 лет, тогда как рабочие пчелы - не более 6 недель. Между различными типами рабочих пчел также наблюдается различия в паттернах метилирования в нескольких десятках генов.

Другой интересный пример – это эпигенетическая регуляция пола у тропических рыб *синеголовых талассом*. В стае большинство рыб – это самки. Если из стаи пропадает самец, то одна из самок меняет окраску с желтой на голубую, а также пол: яичники заменяются семенниками. Исследования показали, что замена женских гонад на мужские происходит благодаря изменению в них транскрипции генов, вызванной метилированием ДНК. В результате смены гонад имеет место активация *генов Polycomb*, эпигенетических регуляторов дедифференцировки, превращающих клетки в плюрипотентные, то есть способные к различным типам морфогенеза, напоминающие стволовые клетки.

Помимо регуляции развития организмов и специализации клеток, эпигенетические модификации ДНК являются причиной *импринтинга*, то есть моноаллельной экспрессии генов, при которой происходит

транскрипция только отцовской или материнской аллели гена. Бывает полный и частичный импринтинг, при последнем импринтированная аллель подавлена не полностью, а лишь уменьшает уровень своей экспрессии. Импринтинг возникает в половых клетках благодаря метилированию гистонов либо ДНК и сохраняется какое-то время в нескольких поколениях соматических клеток, образовавшихся после слияния яйцеклетки и сперматозоида. Генетический импринтинг обнаружен в клетках человека, животных и цветковых растений. В настоящее время у человека известно около трехсот импринтированных генов. Интересный импринтированный ген *Medea (MEA)* был обнаружен у модельного растения арабидопсиса [21]. Гаметофитный мутант по материнской аллели этого гена демонстрирует аберрантную регуляцию роста во время эмбриогенеза у *Arabidopsis thaliana*. Зародыши, образовавшиеся из яйцеклеток мутантов *mea*, растут с черезвычайно большой скоростью и погибают во время высыхания семян. Летальность зародыша не зависит от отцовского вклада и дозировки этого гена. Фенотип *mea* согласуется с теорией «родительского конфликта» в эволюции влияния, зависящего от происхождения по материнской или отцовской линии. *MEA* кодирует белок домена SET, члена группы белков Polycomb [21]. У животных белки группы Polycomb обеспечивают стабильное наследование паттернов экспрессии в делящихся клетках и регулируют контроль пролиферации клеток.

Нарушение нормального метилирования ДНК или гистонов может приводить к возникновению ряда заболеваний у человека. Так, например, сидром Ретта, причиной которого является эпигенетическая мутация в гене *MeCP2*, расположенном в локусе Xq28 на половой хромосоме X, возникает у детей женского пола с частотой 1-1,5x10^{-3}. Симптомы этого заболевания - аутизм, регрессия развития, стереотипные движения рук. Другой синдром человека, называемый ICF (Immunodeficiency, Centromere instability and Facial anomalies), вызывается мутацией в гене ДНК-метилтрансферазы-3b (*DNMT3b*), который располагается на хромосоме 20q11.2. Заболевание передается по аутосомно-рецессивному типу. Симптомы этого редкого генетического заболевания включают легкий лицевой дисморфизм, задержку роста и развития и психомоторную отсталость. Ну и, наконец, нарушение метилирования промотора импринтированного на материнской хромосоме гена *IGF2* (инсулиноподобного фактора роста II) при внутриутробном развитии плода является одной из причин возникновения сахарного диабета, ожирения и появления сердечно-сосудистых заболеваний. Подобные заболевания наблюдались у детей, рожденных голодной зимой 1944 года в Нидерландах (нидерл. hongerwinter), матери которых испытывали сильный голод. Метилирование гена *IGF2* обеспечивает рост и дифференциацию клеток плода во время внутриутробного развития и в норме остается стабильным в течение всей

жизни, импринтинг поддерживается благодаря дифференциально метилированному району (DMR).

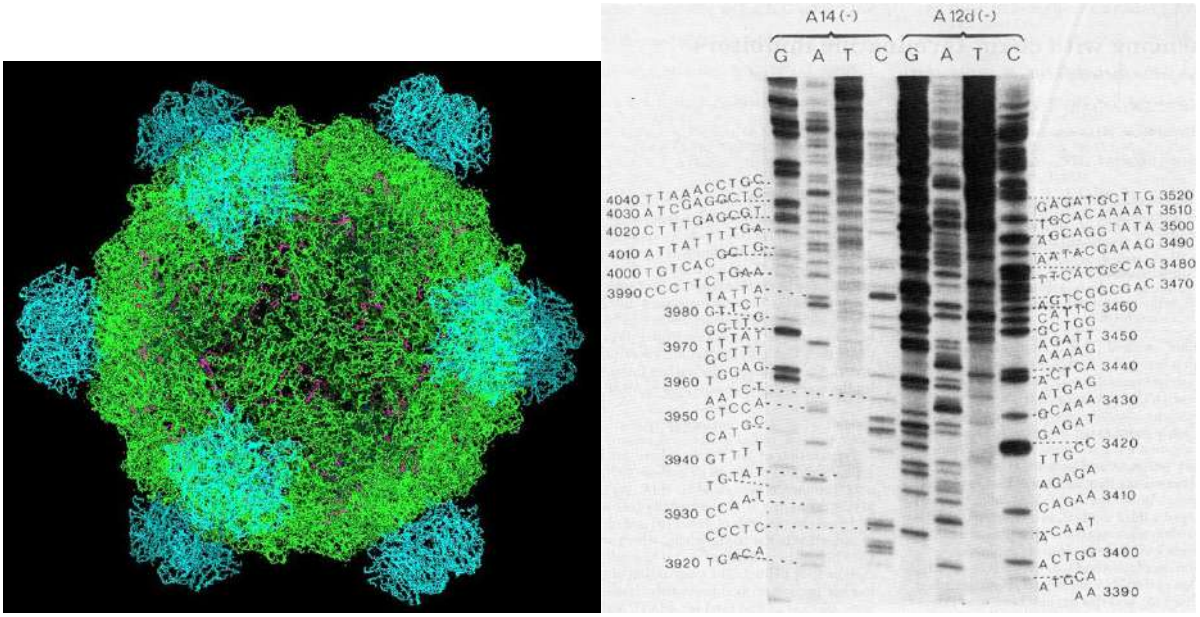
Эпимутации часто являются предшественниками собственно мутаций, в их возникновение вовлечены такие эпигенетические события, как супрессия и ко-супрессия генов, инактивация транспозонов, генетический импринтинг, подавление трансгенов и др., вызванные метелированием ДНК, ремоделированием хроматина и сайленсингом с помощью РНК-интерференции. Благодаря быстрому прогрессу в области технологий секвенирования следующего поколения, вычислительной биологии и методов машинного обучения изучение эпигеномов стало возможным не только на уровне целого организма, но и на уровне целых сообществ и популяций. Все эти достижения оставляют надежду на новые крупные открытия в области эпигеномики, которые будут происходить намного быстрее, чем это было в предыдущие десятилетия.

Вопросы:

1) Что такое эухроматин и гетрохроматин?
2) Что такое метелирование ДНК и его функция в клетках?
3) Что такое гистоны?
4) Какие виды модфикации гистонов бывают?
5) Приведите примеры эпимутаций.

1.5. Полногеномное секвенирование

Появление возможности секвенирования ДНК произвело революцию в биологии и способствовало множеству дальнейших открытий в молекулярной генетике, прежде всего благодаря точной расшифровке нуклеотидных последовательностей и определению структуры генов. Секвенирование считается «золотым стандартом» для идентификации как известных, так и новых вариантов геномной ДНК. Первым широко используемым методом секвенирования ДНК было секвенирование по Сэнгеру, известный как метод «терминирования» цепи ДНК [22], разработанный в середине 70-х годах, этим методом был секвенирован первый геном *бактериофага ФХ174* (бактериофаг, представляющий собой одноцепочечную ДНК (оцДНК вирус) (Рис. 1.5.1.)). Несмотря на высокую точность, недостатком этого метода было возможность расшифровки лишь коротких фрагментов ДНК, размером максимум несколько сотен нуклеотидов.



(a) (b)

Рисунок 1.5.1. (a) Модель капсида фага ФХ174; (b) Определение нуклеотидной последовательности ДНК в исходной матрице методом авторадиографии [22].

Развитие методов секвенирования ДНК привело к ускорению техники расшифровки ДНК в 1990-х годах, что позволило в дальнейшем осуществлять секвенирование всей геномной последовательности организма (whole genome sequencing - WGS) [23]. С 1984 по 1986 год в ходе дискуссий на научных встречах, организованных Министерством энергетики США и другими организациями, впервые была предложена идея секвенирования всего генома. Комитет, назначенный Национальным исследовательским советом США, одобрил эту концепцию, но рекомендовал расширить программу, чтобы она включала создание генетических и физических карт генома человека; а также проведение параллельных экспериментов с ключевыми модельными организмами, такими как бактерии, дрожжи, растение арабидопсис, мухи дрозофилы, круглые черви и мыши (рисунок 1.5.2).

Однако ранние методы полногеномного секвенирования (WGS) были медленными, трудоемкими и дорогими, особенно в эпоху экстенсивного роста геномных данных [25]. Появление и широкое использование методов высокопроизводительного секвенирования или ***следующего поколения*** (next generation sequencing – NGS) в значительной степени облегчило, ускорило и увеличило возможности для выполнения эффективных и недорогих процедур WGS [26]. На рисунке 1.5.3. представлена схема полногеномного секвенирования.

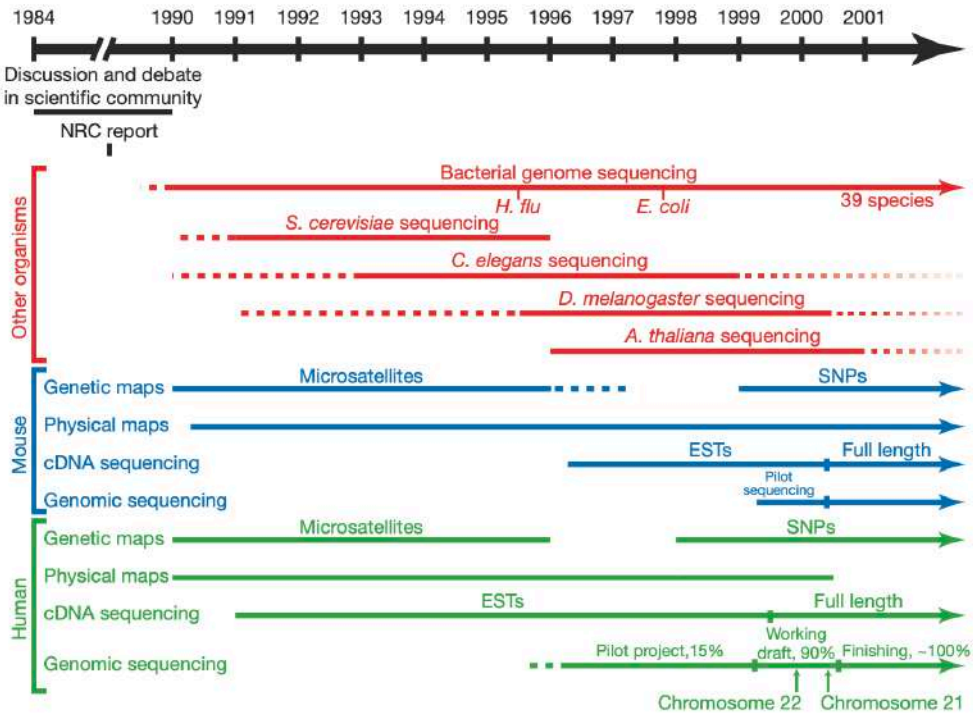


Рисунок 1.5.2. График крупномасштабных геномных анализов. Показаны работы с несколькими модельными организмами бактериями, дрожжами, растениями и беспозвоночными животными (красный), мышами (синий) и человеком (зеленый) с 1990 года [24]

Концепция методов секвенрования по Сэнглеру и NGS схожа. Во время полимеразной цепной реакции, которая состоит из нескольких циклов последовательной репликации ДНК, ДНК-полимераза катализирует дополнительное включение флуоресцентно меченых дезоксирибонуклеозид-5’-трифосфатов (dNTP) в матрицу ДНК. Для каждого цикла цвет помеченного фрагмента ДНК регистрируется детектором, таким образом определяя нуклеотид в последовательности. Основное различие между традиционной технологией Сэнгера и NGS заключается в том, что последняя не ограничивается одним фрагментом ДНК, а анализирует миллионы фрагментов с помощью технологии массового параллельного секвенирования [28,29]. Считается, что в небольшом проекте более целесообразно использовать систему секвенирования Сэнгера из-за ее точности. С другой стороны, в крупномасштабных проектах этот метод исследования будет дорогостоящим и требует много времени, поэтому необходимо применять NGS [30,31].

Десять лет назад наиболее широко используемыми платформами были Roche 454 Life Science[32], Applied Biosystems SOLiD (Sequencing by Oligonucleotide Ligation and Detection)[33] и Illumina Genome Analyzer[34], Ion Torrent [35].

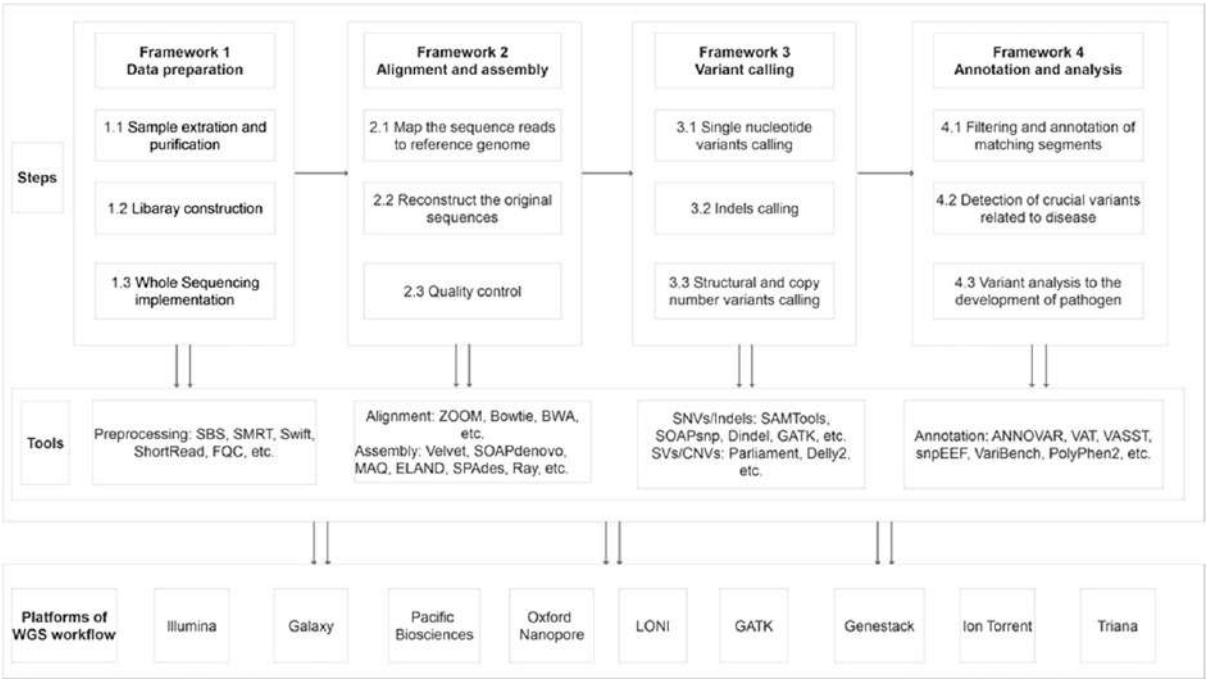
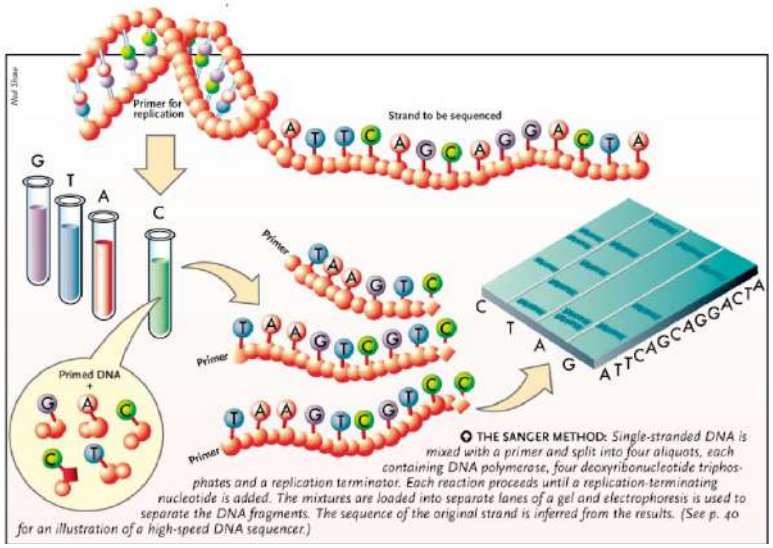
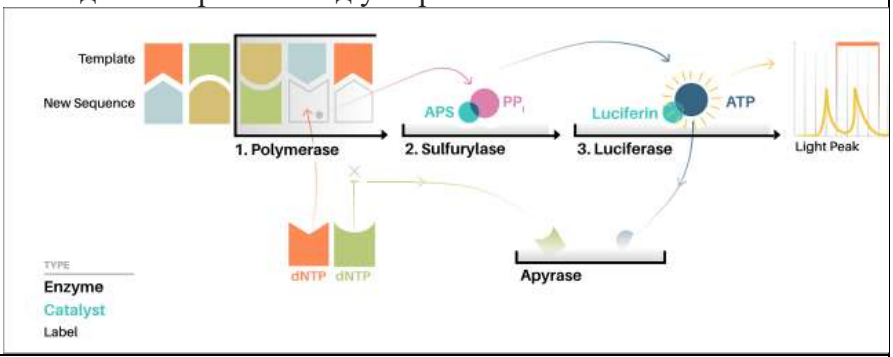


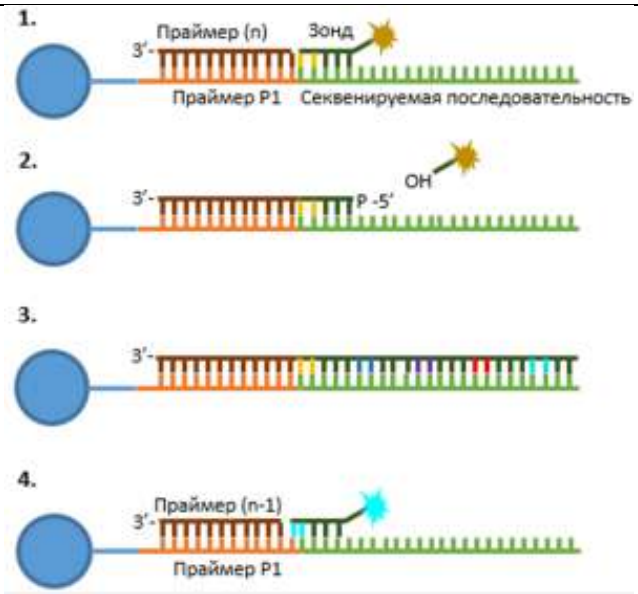
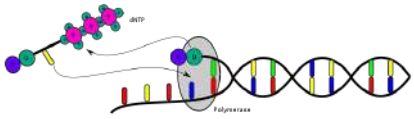
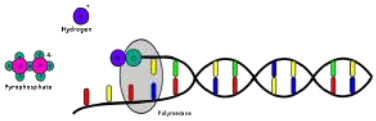
Рисунок 1.5.3. Схема последовательности этапов полногеномного секвенирования для различных платформ [27]

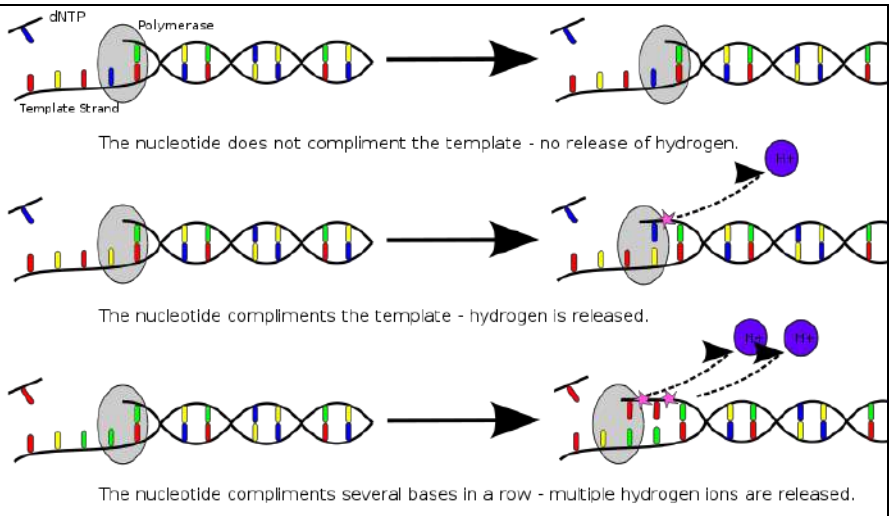
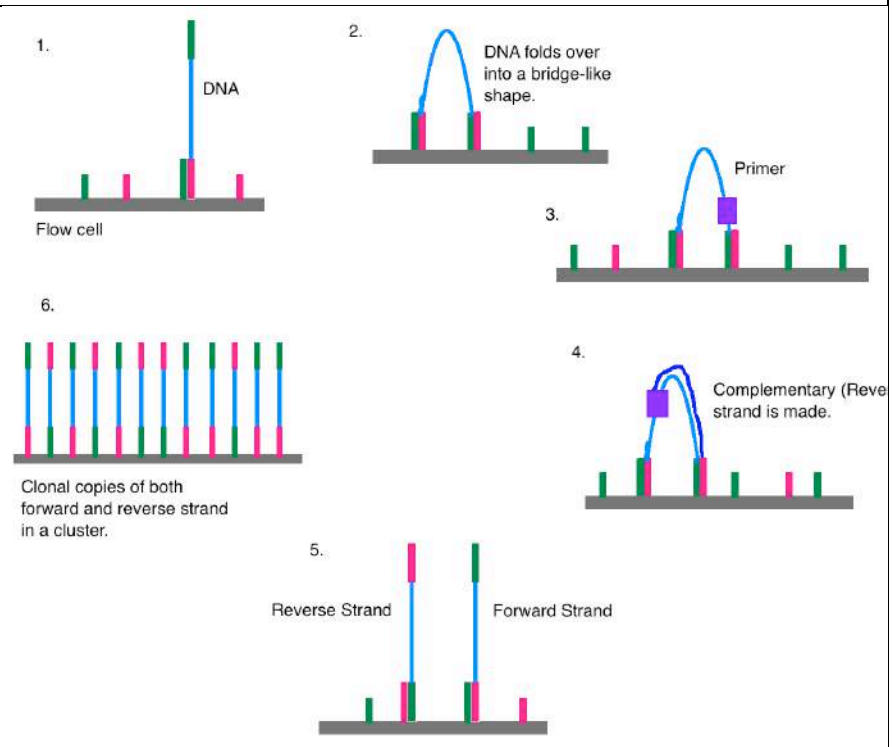
Следующим шагом в развитии секвенирования стала технология NGS (секвенирование длинными ридами) ***третьего поколения*** (NNGS), которая дала возможность обойти недостатки методов второго поколения: необходимость усиления сигнала от каждого из анализируемых фрагментов ДНК путем их амплификации, трудоемкую пробоподготовку, небольшую длину единичных ридов (= прочтений, то есть фрагментов геномной ДНК, которые получаются в результате секвенирования) и т.д. [36]. В 2011 году компания Pacific Biosystems представила концепцию секвенирования одиночных молекул в реальном времени (SMRT) на секвенаторе PacBio RSII [37]. Кроме того, эта технология позволяет получать длинные риды (со средним размером до 30 тыс. п.н. (Kb)), обеспечивает сборку de novo и прямое обнаружение гаплотипов; высокую согласованную точность и возможность эпигенетической характеристики (прямое обнаружение модификаций оснований ДНК с разрешением в одно основание) [38,39].

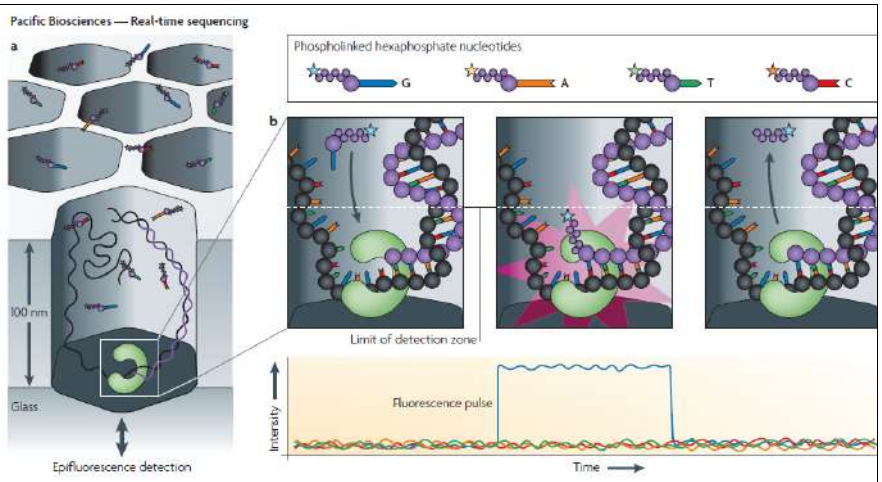
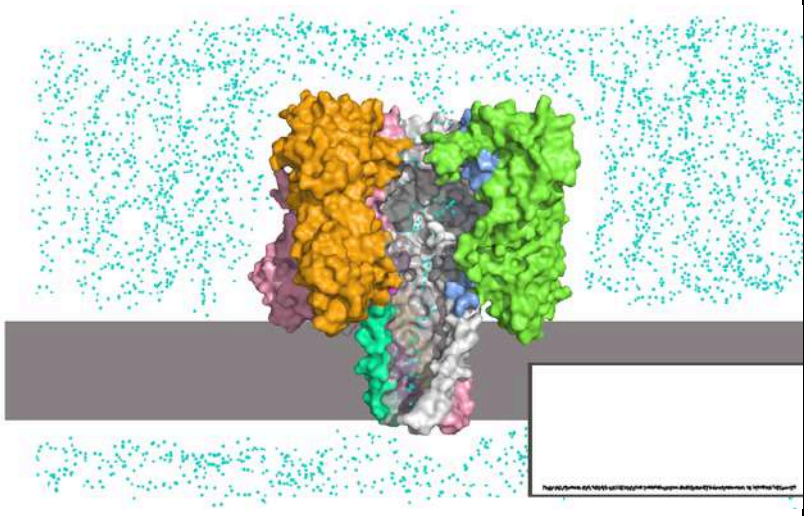
Еще один подход к секвенированию одиночных молекул был предложен Oxford Nanopore Technologies (некоторые авторы называют его ***четвертым поколением***) под названием MinION [40]. Принцип действия основан на изменении тока при прохождении нуклеотидов (A, C, T и G) через нанопору [41]. В мае 2017 года Oxford Nanopore Technologies выпустила GridION Mk1, портативное настольное устройство для секвенирования и анализа, которое предлагает в реальном времени точное секвенирование ДНК и РНК. Прибор позволяет проводить до пяти экспериментов одновременно или по отдельности, с простой подготовкой

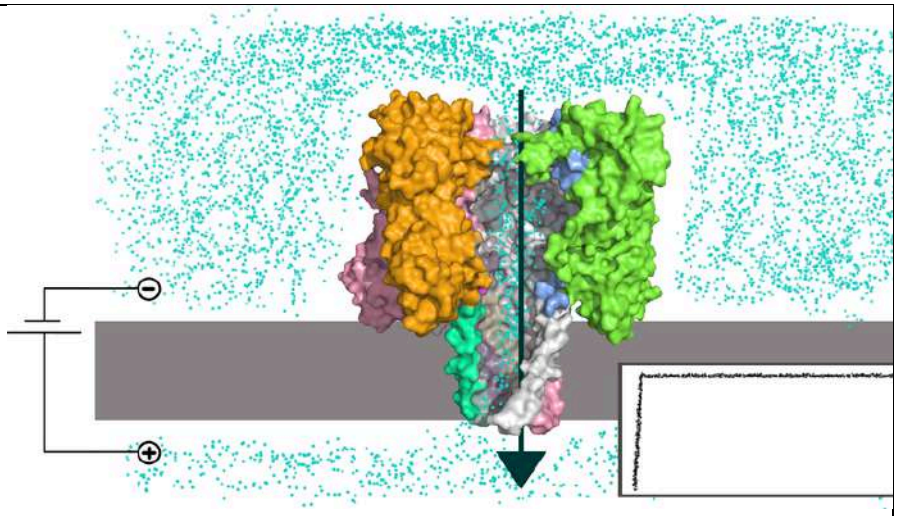
Таблица 1.5.1. Популярные технологии секвенирования

Традиционное секвенирование	
Секвенирование по Сэнгеру	 • По-прежнему широко используется для секвенирования коротких фрагментов ДНК (до 1000 п.н.) из-за простоты и точности секвенирования. • Полногеномное секвенирование требует больших затрат труда, времени и средств.
Секвенирование «Shotgun»	Фрагментация длинных цепей ДНК на многочисленные более мелкие сегменты для секвенирования по Сэнгеру
Секвенирование *второго поколения*	
Pyrosequencing (Roche 454)	• Детектирует высвобождение пирофосфата при добавлении комплементарного нуклеотида к секвенируемой последовательности • Более низкая пропускная способность и, как следствие, более высокая стоимость секвенирования на одно основание • В данное время метод устарел 
SOLiD sequencing (Life Technologies)	• Секвенирование SOLiD использует подход, основанный на лигировании • Менее популярна, чем другая платформа Life Technologies, Ion Torrent.

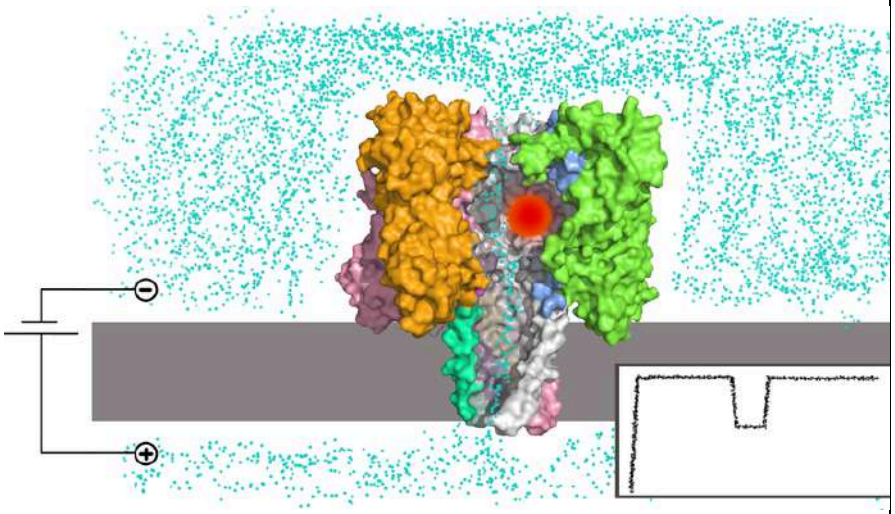
	
Ion semiconductor sequencing (Life Technologies Ion Torrent)	• Метод секвенирования основан на детектировании изменения pH из-за высвобождения ионов водорода во время полимеризации. • Требуется отдельная амплификация библиотеки эмульсионной ПЦР перед секвенированием (медленная и сложная). • Более высокий уровень ошибок для гомополимеров, короткая длина читаемого фрагмента.  Polymerase integrates a nucleotide.  Hydrogen and pyrophosphate are released.

	
Illumina sequencing	
Секвенирование *третьего поколения*	
Single molecule real-time sequencing (Pacific Biosciences)	

	
Секвенирование *четвертого поколения*	
Nanopore sequencing (Oxford Nanopore)	Секвенирование нанопор — это технология, которая позволяет проводить прямой анализ в реальном времени длинных фрагментов ДНК или РНК. Оно работает, отслеживая изменения электрического тока при прохождении нуклеиновых кислот через нанопоры белка. Результирующий сигнал декодируется для получения конкретной последовательности ДНК или РНК.  Нанопора альфа-гемолизина в мембране, напряжение отсутствует



Под действием приложенного напряжения возникает ток ионов



При прохождении большой молекулы через пору (показана красным) ток ионов уменьшается

библиотеки, что позволяет генерировать до 150 Гб данных секвенирования. В 2019 году была запущена система PromethION 48, которая представляет 48 проточных ячеек, каждая проточная ячейка позволяет одновременно обрабатывать до 3000 нанопор, что может обеспечить выход до 7,6 ТБ данных с результатами секвенирования за 72 часа. Хотя точность секвенирования нанопор еще не сопоставима с точностью секвенирования с коротким ридами (например, на платформы Illumina заявляют о точности секвенирования в 99,9%), хотя регулярно происходят обновления, а постоянные разработки направлены на расширение диапазона геномов и дальнейшее повышение точности технологии [42]. Наиболее популярные технологии секвенирования перечислены в таблице 1.5.1.

В 2016 году компания 10X Genomics выпустила прибор Chromium, в котором была реализована технология создания библиотек секвенирования с индивидуальными штрих-кодами, которые помещали в сотни тысяч капель масла в виде гелевых шариков (GEM) нанолитрового объема. Такая технология меченных библиотек позволяет получить длинные риды высокого качества с последующей возможностью сборки генома с минимумом ошибок [43]. Технология 10X Genomics открывает новые области применения, она особенно полезна для разработки новых методов в эпигенетике [44], сборки генома de novo [45] и получения длинных ридов [46].

Вопросы:

1) Опишите принцип секвенирования по Сэнглеру.
2) Перечислите этапы развития методов секвенирования.
3) В чем заключаются принципы методов секвенирования второго, третьего и четвертого поколений?

2. Биоинформатические методы исследования генома

Полногеномное секвенирование с помощью NGS на выходе выдает миллионы коротких (100-150 п.н.) или длинных (до 25 тыс. п.н. и выше) ридов в зависимости от платформы, на которой производилось секвенирование.

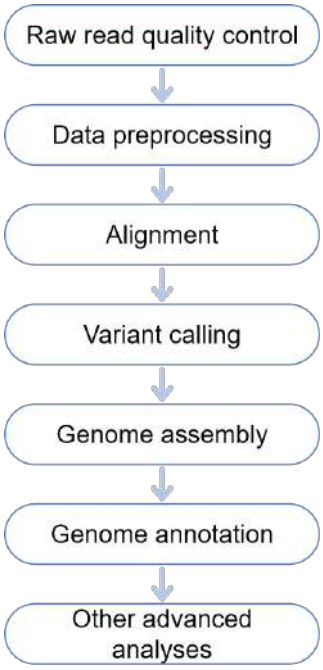


Рисунок 2.1. Рабочий биоинформатический процесс для сборки и аннотации генома на основе данных полногеномного секвенирования.

Эти риды можно сравнить с мелкими кусочками порванных книг из библиотеки, содержащей тысячи томов. Задачу восстановления этих книг из обрывков до читаемого состояния и размещение их по первоначальным полкам, согласно тематике, можно сравнить с задачей сборки генома из множества ридов до уровня хромосом и аннотацию структуры этого генома. Поэтому существует потребность в глубоком и точном анализе ридов с использованием сложных алгоритмов, чтобы каждый кусочек (рид) встал на свое место и не был перепутан с другими ридами. Осуществлением данной задачи занимается наука биоинформатика. Если есть эталонный геном (референс) близкородственного организма, на который можно выровнять полученные риды, тогда это несколько облегчает задачу, если же референса нет, то приходится делать сборку de novo, используя другие программы и алгоритмы, отличные от выравнивания. После того как сборка генома завершена, обычно делают ее аннотацию, то есть выделяют структурные компоненты генома либо функциональные компоненты. Для

этого последовательности можно сканировать, чтобы найти значимые совпадения между компонентами и сигнатюры, которые ранее были описаны и являются маркерами геномных структур [47,48]. Для сравнения данных необходим поиск информации в различных биомедицинских базах данных. Одним из крупнейших источников биомедицинской и геномной информации является NCBI (National Center for Biotechnology Information), который обеспечивает доступ к другим базам данных, таким как PubMed, Entrez Gene, OMIM, Variation Viewer, dbSNP и другим [49].

Биоинформатический процесс для полногеномного секвенирования состоит из следующих этапов (Рис 2.1): (1) контроль качества необработанного данных (ридов); (2) отфильтровывание некачественных ридов; (3) выравнивание; (4) поиск вариантов; (5) сборка генома; (6) аннотация генома; (7) другой расширенный анализ, например филогенетический анализ.

2.1. Контроль качества ридов и предварительная обработка

Как уже было сказано, высокопроизводительное секвенирование на разных платформах не расшифровывает всю ДНК организма как один длинный отрезок, а вместо этого создает большой набор коротких фрагментов, каждый из которых содержит лишь малую часть генетической информации. Эти фрагменты называются «последовательным чтением» и хранятся в так называемых файлах «FASTQ».

Риды, полученные в результате секвенирования, могут быть собраны de novo в полный геном или сопоставлены с уже собранным эталонным геномом родственного организма. Однако ни одна технология секвенирования не является идеальной, и сырые риды неизбежно содержат ошибки секвенирования. Вероятность ошибки для каждого нуклеотида при каждом чтении всегда записывается в файл FASTQ. Поэтому самым первым шагом фрагментарного анализа является контроль качества и фильтрация. Этот шаг направлен на удаление некачественных ридов.

Файлы FASTQ всегда содержат 4 строки на последовательность (Рис. 2.1.1). В первой строке отображается идентификатор последовательности и описание. Вторая строка содержит последовательность нуклеотидов. Третья строка обычно содержит только символ «+» и иногда тот же идентификатор и описание последовательности, что и первая строка. В четвертой строке отображается оценка качества каждого нуклеотида, показанного во второй строке.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Рисунок 2.1.1. Структура файла FASTQ

Показатели качества в четвертой строке представляют вероятность ошибки секвенирования в каждом положении нуклеотида. Если нам известна вероятность такой ошибки, то легко получить оценку качества, используя следующее уравнение:

$$Q_{\text{sanger}} = -10 \log_{10} p \quad .$$

Например, если вероятность ошибки (p) равна 0,01, то соответствующий показатель качества будет равен 20; если p = 0,001, то Q=30. Но, как мы видим в четвертой строке примера, файл FASTQ не содержит двузначных чисел, а вместо этого содержит только символы. Это специальные символы ASCII, которые используются для кодирования значений качества с помощью символов, а не двух- или трехзначной цифр. Такая кодировка одного необходима для того, чтобы иметь однозначное соответствие между каждым нуклеотидом в строке 2 и оценкой в строке 4.

Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0	0	000	NULL	32	20	040	&#032;	Space	64	40	100	&#064;	@	96	60	140	&#096;	`
1	1	001	Start of Header	33	21	041	&#033;	!	65	41	101	&#065;	A	97	61	141	&#097;	a
2	2	002	Start of Text	34	22	042	&#034;	"	66	42	102	&#066;	B	98	62	142	&#098;	b
3	3	003	End of Text	35	23	043	&#035;	#	67	43	103	&#067;	C	99	63	143	&#099;	c
4	4	004	End of Transmission	36	24	044	&#036;	$	68	44	104	&#068;	D	100	64	144	&#100;	d
5	5	005	Enquiry	37	25	045	&#037;	%	69	45	105	&#069;	E	101	65	145	&#101;	e
6	6	006	Acknowledgment	38	26	046	&#038;	&	70	46	106	&#070;	F	102	66	146	&#102;	f
7	7	007	Bell	39	27	047	&#039;	'	71	47	107	&#071;	G	103	67	147	&#103;	g
8	8	010	Backspace	40	28	050	&#040;	(	72	48	110	&#072;	H	104	68	150	&#104;	h
9	9	011	Horizontal Tab	41	29	051	&#041;	)	73	49	111	&#073;	I	105	69	151	&#105;	i
10	A	012	Line feed	42	2A	052	&#042;	*	74	4A	112	&#074;	J	106	6A	152	&#106;	j
11	B	013	Vertical Tab	43	2B	053	&#043;	+	75	4B	113	&#075;	K	107	6B	153	&#107;	k
12	C	014	Form feed	44	2C	054	&#044;	,	76	4C	114	&#076;	L	108	6C	154	&#108;	l
13	D	015	Carriage return	45	2D	055	&#045;	-	77	4D	115	&#077;	M	109	6D	155	&#109;	m
14	E	016	Shift Out	46	2E	056	&#046;	.	78	4E	116	&#078;	N	110	6E	156	&#110;	n
15	F	017	Shift In	47	2F	057	&#047;	/	79	4F	117	&#079;	O	111	6F	157	&#111;	o
16	10	020	Data Link Escape	48	30	060	&#048;	0	80	50	120	&#080;	P	112	70	160	&#112;	p
17	11	021	Device Control 1	49	31	061	&#049;	1	81	51	121	&#081;	Q	113	71	161	&#113;	q
18	12	022	Device Control 2	50	32	062	&#050;	2	82	52	122	&#082;	R	114	72	162	&#114;	r
19	13	023	Device Control 3	51	33	063	&#051;	3	83	53	123	&#083;	S	115	73	163	&#115;	s
20	14	024	Device Control 4	52	34	064	&#052;	4	84	54	124	&#084;	T	116	74	164	&#116;	t
21	15	025	Negative Ack.	53	35	065	&#053;	5	85	55	125	&#085;	U	117	75	165	&#117;	u
22	16	026	Synchronous idle	54	36	066	&#054;	6	86	56	126	&#086;	V	118	76	166	&#118;	v
23	17	027	End of Trans. Block	55	37	067	&#055;	7	87	57	127	&#087;	W	119	77	167	&#119;	w
24	18	030	Cancel	56	38	070	&#056;	8	88	58	130	&#088;	X	120	78	170	&#120;	x
25	19	031	End of Medium	57	39	071	&#057;	9	89	59	131	&#089;	Y	121	79	171	&#121;	y
26	1A	032	Substitute	58	3A	072	&#058;	:	90	5A	132	&#090;	Z	122	7A	172	&#122;	z
27	1B	033	Escape	59	3B	073	&#059;	;	91	5B	133	&#091;	[	123	7B	173	&#123;	{
28	1C	034	File Separator	60	3C	074	&#060;	<	92	5C	134	&#092;	\	124	7C	174	&#124;	|
29	1D	035	Group Separator	61	3D	075	&#061;	=	93	5D	135	&#093;	]	125	7D	175	&#125;	}
30	1E	036	Record Separator	62	3E	076	&#062;	>	94	5E	136	&#094;	^	126	7E	176	&#126;	~
31	1F	037	Unit Separator	63	3F	077	&#063;	?	95	5F	137	&#095;	_	127	7F	177	&#127;	Del

asciicharstable.com

Рисунок 2.1.2. Структура файла FASTQ

Первые 32 элемента в таблице (рис. 2.1.2) ASCII - непечатаемые символы. Следовательно, чтобы присвоить символу печати значение качества, мы должны начать с 33-го пункта. Это означает, что значение качества 1 будет соответствовать символу «!», значение качества 10 - «*» и так далее. Этот тип кодирования качества называется Phred33 и является наиболее распространенным способом хранения информации о качестве последовательности. Например, Phred33 используется в выходных файлах платформ секвенирования Illumina.

Для контроля качества сырых ридов чаще всего используют инструмент FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastq), который генерирует результаты статистических данных, включая базовую статистику, качество последовательности, показатели качества, содержимое последовательности, распределение их длин, избыточно представленные последовательности, графики уровня их дублирования, содержимое адаптера и содержимое k-мер. Для обрезки чтения можно использовать такие инструменты, как Fastx_trimmer и cutadapt. Более подробно с этим этапом можно ознакомиться здесь [50,51].

2.2. Выравнивание ридов

Выравнивание – это способ упорядочения последовательностей ДНК, РНК или белка для выявления областей сходства, которые могут быть следствием функциональных, структурных или эволюционных отношений между последовательностями [52]. Для процедуры выравнивания используются эталонные геномы, таким образом, алгоритм сопоставления будет пытаться найти место в эталонной последовательности, которое соответствует риду, допуская определенное количество несовпадений, чтобы обеспечить обнаружение изменений подпоследовательности. Было разработано более 60 инструментов для картирования генома, и по мере обновления платформ NGS будет появляться все больше и больше инструментов, которые постоянно развиваются (более подробно см. [53,54]). Среди часто используемых методов для выполнения выравнивания коротких чтений выделяют Burrows-Wheeler Alignment (BWA) [55] и Bowtie2 [56].

Результатом работы BWA и Bowtie2 является стандартный формат файла сопоставления последовательностей, известный как BAM файл, который упрощает выполнение следующих шагов. В качестве альтернативы для локального выравнивания широко используется алгоритм BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Таблица 2.2.1. Распространенные вычислительные программы для выравнивания ридов.

Программа	Source type	Website
Bowtie2	Open source	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SEAL	Open source	http://compbio.case.edu/seal/
SOAP3	Open source	http://www.cs.hku.hk/2bwt-tools/soap3/; http://soap.genomics.org.cn/soap3.html
BWA, BWA-SW	Open source	http://bio-bwa.sourceforge.net/
Novoalign	Commercially available	http://www.novocra.com/
SHRiMP/SHRiMP2	Open source	http://compbio.cs.toronto.edu/shrimp/
MAQ	Open source	http://maq.sourceforget.net/
Stampy	Open source	http://www.well.ox.ac.uk/project-stampy/
ELAND	Commercially available	http://www.illumina.com/
SARUMAN	Open source	http://www.cebitec.uni-bielefeld.de/brf/saruman/saruman.html

2.3. Определение вариантов

После прохождения «технического контроля» могут быть получены новые данные путем сравнения полученного генома с эталонным геномом. Такие данные могут быть связаны с заболеванием или просто быть нефункциональным геномным шумом. На этом этапе формируются файлы формата VCF (variant calling file) - это в основном текстовый файл, содержащий строки метаинформации, строку заголовка, за которой следуют строки данных, каждая из которых содержит информацию о положении на хромосоме, справочную базу, идентифицированную альтернативную базу или базы (рисунок 2.3.1). Формат также содержит информацию о генотипе образцов для каждой позиции [57].

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS      ID         REF    ALT     QUAL  FILTER   INFO                              FORMAT       NA00001         NA00002          NA00003
20     14370    rs6054257  G      A       29    PASS     NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51
1/1:43:5:.,.
20     17330    .          T      A       3     q10      NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3
0/0:41:3
20     1110696  rs6040355  A      G,T     67    PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2
2/2:35:4
20     1230237  .          T      .       47    PASS     NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51
0/0:61:2
20     1234567  microsat1  GTC    G,GTCT  50    PASS     NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4        0/2:17:2
1/1:40:3
```

Рисунок 2.3.1. Пример файла VCF

Этап определения вариантов может быть затруднен из-за большого количества ложноположительных и ложноотрицательных определений однонуклеотидного полиморфизма (SNV), что контролируется на предыдущем этапе. Пакеты программного обеспечения для данного этапа приведены в таблице 2.3.1 Основными факторами при выборе пакета являются: возможность формировать как список определившихся отличий от основной версии генома организма, так и выделять места полиморфизма, которые находятся в заранее сформированном списке; корректность оценки качества прочтения нуклеотида в точке, с учетом особенностей конкретного прибора; вычислительное быстродействие.

Таблица 2.3.1. Программные пакеты для определения вариантов

Программы	**Descriptions**	**Website**
GATK	Multiple-sequence realignment Quality score recalibration SNP genotyping Indel discovery and genotyping	http://software.broadinstitute.org/gatk/
SOAPsnp	Consensus calling and SNP detection Calculation of the likelihood of each genotype	http://soap.genomics.org.cn/

VarScan/VarScan2	Detects variants at 1% frequency Normalizes sequence depth at each position	http://genome.wustl.edu/tools/cancer-genomics
ALTAS 2	Variant calling of aligned data from diverse NGS platforms	http://www.genboree.org/

2.4. Сборка генома

Сборка генома представляет собой процесс реконструкции всей его последовательности из множества более коротких последовательностей

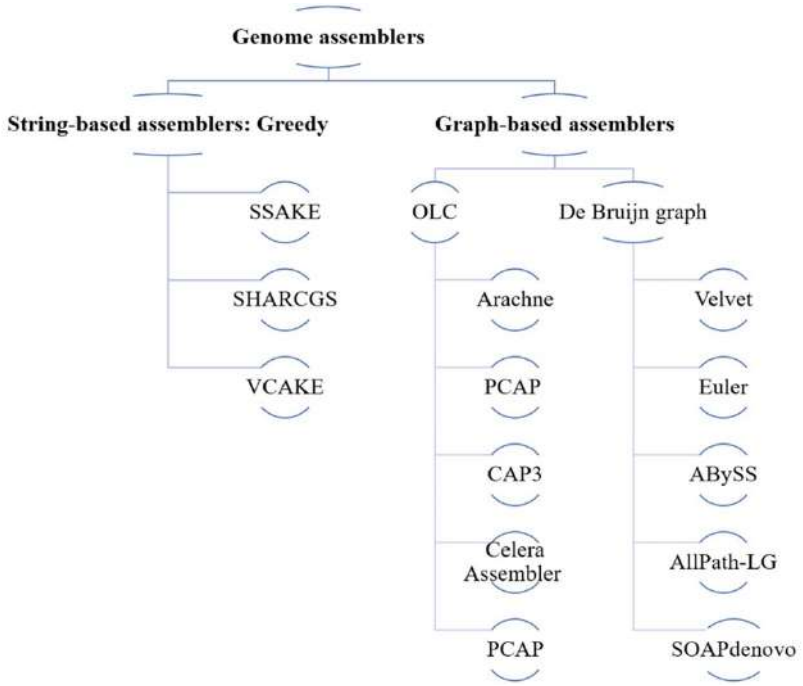


Рисунок 2.4.1. De novo асемблеры

или ридов. Большинство алгоритмов сборки de novo основаны на стратегии консенсуса перекрытия-компоновки (overlap-layout-consensus, OLC), при которой все риды попарно выравниваются, затем происходит идентификация перекрытий между всеми парами ридов, и строятся графы перекрытий. Сборка de novo предпочтительна для длинных последовательностей.

Ассемблеры de novo в большинстве случаев основаны на теории графов и могут быть разделены на три основные группы (Рисунок 2.4.1.)

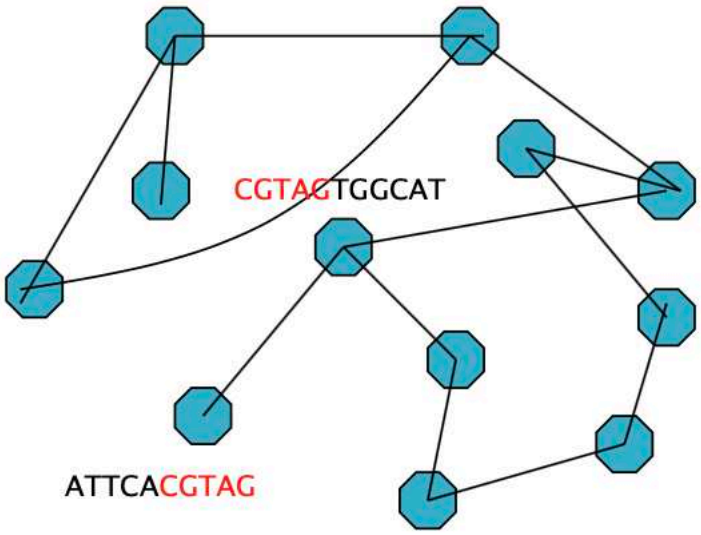
2.4.1. Алгоритмы консенсуса перекрытия-компановки (OLC)

Шаг 1: поиск пересекающихся (overlaping) фрагментов

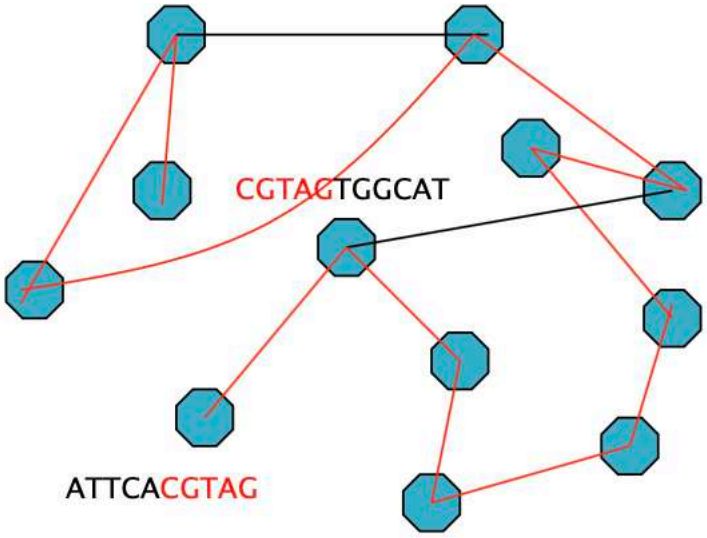
- Найти пары фрагментов, имеющих общие k-меры (строки длины k, обычно k~24).
- Увеличить область пересечения путем выравнивания строк в обе строны. Прекратить процесс если сходство строк падает ниже определенного порога (<95% сходства).

Шаг 2: построение графа, поиск контигов - последовательных участков генома, образованных пересекающимися ридами без пропусков.

Узлы - фрагменты генома (риды). Ребра - пересечения фрагментов.



- Наибольший простой путь - простой путь, который невозможно продлить.
- Гамильтонов путь - простой путь проходящий через все вершины графа.
- Поиск гамильтонова пути - NP-полная задача.



Шаг 3: определение консенсусной последовательности.

- Построение множественного выравнивания по попарным выравниваниям фрагментов.
- Корректировка ошибок.

2.4.2. Методы, использующие граф де Брюина (DBG, также известный как Eurelian)

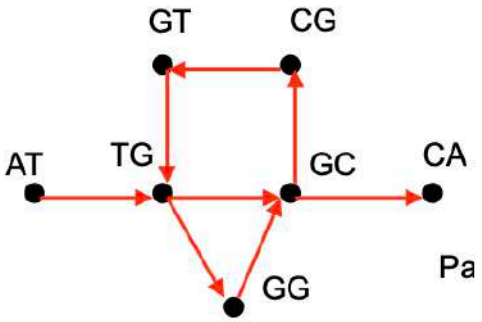
- Путь Эйлера - путь, проходящий один раз по каждому ребру графа.
- Существует линейный по времени алгоритм нахождения Эйлерова пути для графа.
- Фрагменты (риды) разбиваются на пересекающиеся k-меры.

Пример:
фрагмент 10bp: ATTCGACTCC
разбиение на k=5-меры: ATTCG
TTCGA
TCGAC
CGACT
GACTC
ACTCC

- Построение графа.

- Узлы - (k-1)- меры
- Ребра - связывающие их k-меры.

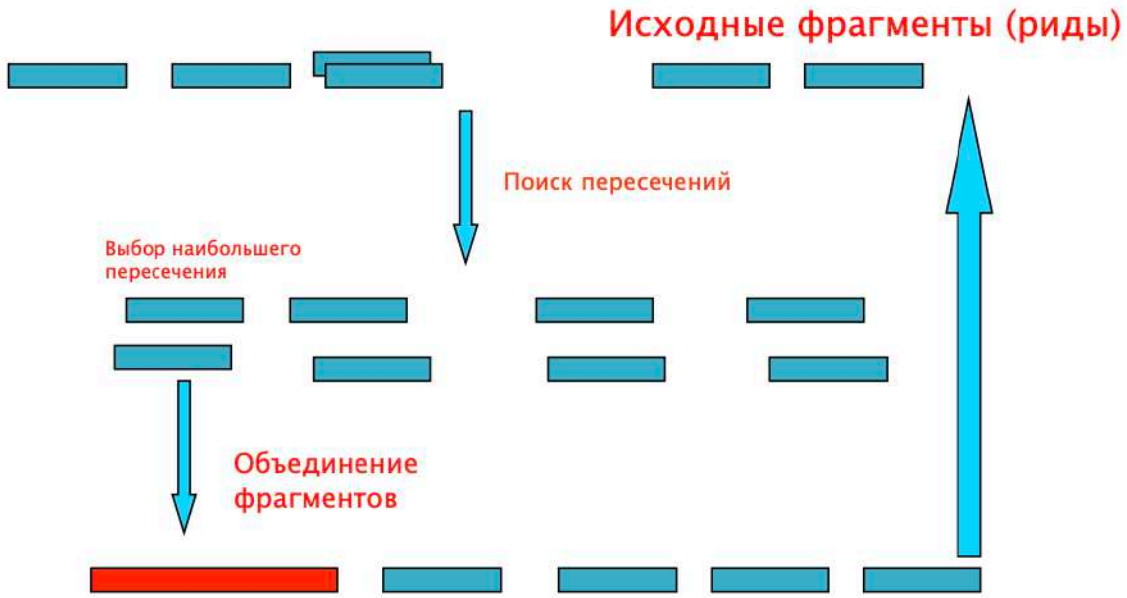
S = { ATG, TGC, GTG, GGC, GCA, GCG, CGT }



ATGGCGTGCA

2.4.3. «Жадные» алгоритмы

1. Найти пересечение фрагментов.
2. Выбрать два фрагмента с наибольшим пересечением.
3. Объединить выбранные фрагменты.
4. Повторить пункты 2 и 3 до объединения всех фрагментов.



2.4.4. Аннотация генома

После завершения этапа сборки генома и определния вариантов следующим важным шагом является аннотация генома для выявления его структуры (Рисунок 2.4.4.1). Аннотация генома включает идентификацию сегментов генома, некодирующих белки, например повторы, псевдогены и др., предсказание кодирующих белок генов и сопоставление идентифицированных элементов с обнаруженными ранее генетическими последовательностями, размещенными в существующих базах данных. С помощью аннотации генома, к примеру, можно осуществить реконструкцию метаболических путей в организме, выявить регуляторные сети и смоделировать их взаимодействие и развитие, а также предсказать динамику фенотипа организма. В свою очередь с помощью функциональной аннотации генов и белков можно установить их биологическую функцию и особенности регуляции, то есть выяснить - как и в каких условиях они работают. Заполнение каталога функциональных областей генома может облегчить его дальнейший анализ, например, оценку эффекта мутаций, вызванных однонуклеотидными вариантами, у индивидуумов или популяций.

Аннотация состоит из двух этапов.

(1) *Вычислительная фаза*. Включает в себя маскировку повторов, предсказание кодирующих гены последовательностей (CDS) и предсказание генных моделей.

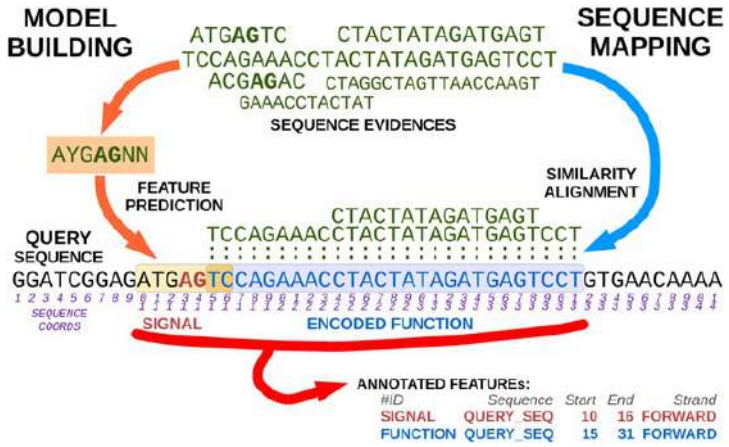


Рисунок 2.4.4.1. Для аннотирования неизвестной последовательности существует два основных подхода. Первый подход, изображенный слева, заключается в построении моделей, которые на основе известных последовательностей суммируют некоторые свойства или характеристики смещения аннотируемых функциональных элементов, а затем применяют алгоритм для оценки этих элементов по последовательности, принимая оптимальное предсказание в качестве аннотации. Второй подход использует выравнивание по запросу набора известных последовательностей, полученных либо у того же (поиск сходства), либо у других видов (поиск гомологии). Опять же, за оптимальное выравнивание принимается то, которое аннотирует функциональный элемент. Координаты относительно исходной последовательности определяют местоположение каждого из функциональных элементов, обнаруженных в анонимном геноме, также известном как набор аннотаций [58].

Маскировка повторов: поскольку повторы плохо собираются в геномах у разных видов, рекомендуется создание видоспецифичных библиотек повторов с помощью таких инструментов, как RepeatModeler, RepeatExplorer.

Для предсказания кодирующих белки генетических последовательностей (CDS) используются алгоритмы ab initio (подробнее об алгоритмах предсказания для прокариот и эукариот, а также об оценке точности предсказания и программах в [58].)

Предсказание генных моделей.

Секвенирование экспрессирующихся фрагментов EST (expressed sequence tag) или полноразмерных мРНК и белков видов, геном которых аннотируется, было основным подходом к повышению точности аннотации генов ab initio. Вместо того, чтобы полагаться на статистические свойства CDS и определяющие их сигналы, прямое выравнивание транскрибируемых или транслируемых последовательностей генома

обеспечивает экспериментальное подтверждение структуры гена. Выравнивание таких последовательностей с «голым» геномом использует алгоритм динамического программирования, который учитывает сайты сплайсинга и структуру экзон-интронной организации эукариотических генов. В частности, EST или полноразмерные транскрипты могут точно согласовываться с геномом с использованием программы EST_GENOME. Для быстроты можно использовать эвристический алгоритм BLAST для получения начальных выравниваний кодирующих экзонов в геноме, которые позже рекурсивно объединяются в цепочку для получения наилучшей структуры гена. Более новым инструментом для выравнивания транскриптов на геномную последовательность является программа EXONERATE, которая обеспечивает (более медленный) полный или (более быстрый) эвристический подход (не являющийся гарантировано точным и оптимальным, но достаточным для решения поставленной задачи) для выравнивания транскриптов с использованием различных алгоритмов динамического программирования.

(2) *Фаза аннотации*. Все данные, полученные выше (ab initio предсказание, а также выравнивание белков, EST- и РНК), затем суммируются для конечной аннотации гена. Существуют автоматизированные инструменты аннотации, такие как MAKER и PASA. Для редактирования аннотации через визуальный интерфейс можно использовать WebApollo.

После оценки конечной аннотации нового генома путем визуального осмотра авторы обычно публикуют и черновую аннотацию последовательности генома для того, чтобы у других исследователей была возможность для улучшения сборки и аннотации генома. Доступные базы данных для загрузки геномов размещены на порталах ENSEMBL и NCBI.

Вопросы:

1) Перечислите этапы биоинформатического анализа данных, полученных в ходе полногеномного секвенирования.
2) Какие существуют методы сборки генома?
3) Из каких шагов состоит процесс аннотации генома?

3. Функциональная геномика

Если посмотреть вокруг, то можно увидеть, насколько сильно люди отличаются друг от друга по внешнему виду, по росту, цвету кожи и глаз, пропорциям фигуры, весу, а также по характеру, способностям, склонности к различным заболеваниям и многим другим признакам. И это при том, что 99,9% ДНК в геноме у всех людей одинаково. Еще более интересные наблюдения можно сделать, разглядывая части собственного тела: ноги,

руки, лицо, глаза абсолютно не похожи друг на друга. Почему же так происходит, зная, что каждая клетка одного организма содержат идентичную ДНК? Все дело в том, что в разных клетках, тканях и органах организма различные гены экспрессируются по-разному. Кроме того, наличие гена в геноме вовсе не означает, что он будет транскрибирован и транслирован.

Технологическими предпосылками возникновения функциональной геномики были успехи молекулярной биологии, в частности, технология рекомбинантной ДНК, ПЦР-амплификация, методы гибридизации, а также появление методов высокоскоростного секвенирования ДНК и развитие методов компьютерной биологии для обработки и анализа больших объемов генерируемых данных. Функциональная геномика, что следует из самого названия, изучает не просто структуру генома, а именно функцию генов, сосредотачиваясь на динамических аспектах, таких как транскрипция генов, трансляция и межбелковые взаимодействия. Также исследуется гомология последовательностей, осуществляется поиск в библиотеках аналогичных последовательностей, уже описанных в других белках с известной функцией, в том числе и у других видов. Производится поиск и сравнение консервативных участков белков, например анализ доменов / мотивов, поскольку известно, что некоторые последовательности имеют определенную структуру («мотив»), например, «спираль-поворот-спираль») или функцию («домен»), например, последовательность ионного канала, область связывания ДНК и др. Целью функциональной геномики является понимание взаимосвязи между геномом организма и его фенотипом. Термин функциональная геномика часто используется в широком смысле для обозначения множества возможных подходов к пониманию свойств и функций генов и генных продуктов организма в целом. Предмет исследования функциональной геномики включает изучение естественных изменений генов, РНК и белков во времени (например, при развитии организма) или в пространстве (например, в различных частях его тела), а также исследования естественных или экспериментальных функциональных нарушений, влияющих на гены, хромосомы, РНК или белки.

3.1. Идентификация функции гена

Как было сказано выше, основной задачей функциональной геномики является определение функции генов и белков во времени и пространстве и, в конечном итоге всех компонентов генома. В зависимости от подхода функциональная генетика подразделяется на *прямую генетику* (англ. forward genetics, то есть прямой генетический скрининг) и *обратную генетику* (англ. reverse genetics, то есть обратный генетический скрининг). Прямая генетика идет от фенотипа к генотипу, то есть идентифицирует

гены, ответственные за определенный фенотип организма. В то же время обратная генетика анализирует фенотип организма, выявляя последствия нарушения работы известного гена.

3.1.1. Методы потери функции. Мутагенез

Одним из направлений функциональной генетики является метод анализа мутантов, полученных путем «нокаута» исследуемых генов. Генетические нокауты осуществляются посредством делеции гена, либо путем нарушения его функции, например, с помощью инсерционного мутагенеза, далее изучается фенотип полученных мутированных организмов, что дает ключ к разгадке функции нарушенного гена.

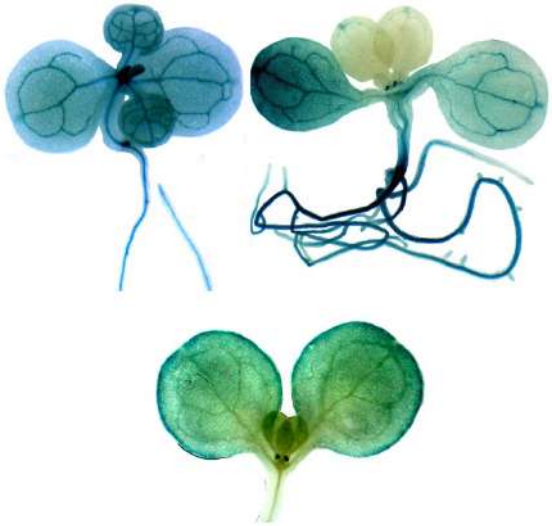
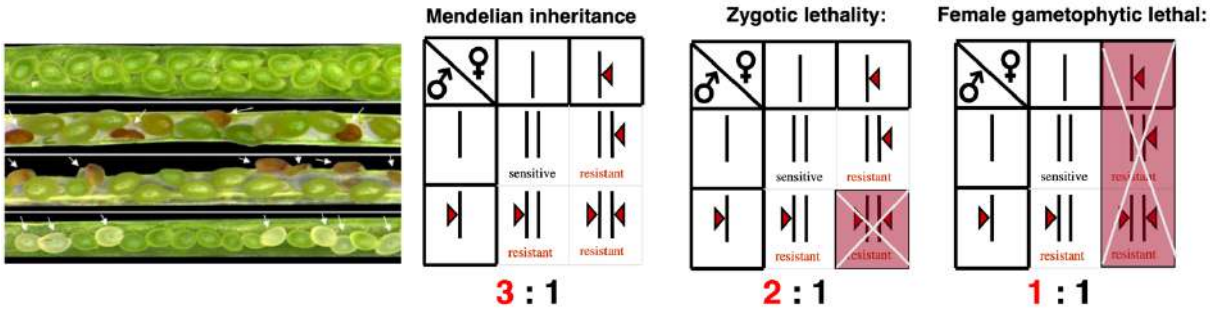


Рисунок 3.1.1.1 Паттерны экспрессии гена репортера *GUS*, встроенного в *Ds*-транспозон, соответствующие локализации экспресии прерванных транспозоном генов у растения арабидопсиса.

В этом параграфе мы рассмотрим метод анализа мутантов, полученных нами с помощью инсерционного мутагенеза распространенного модельного растения *Arabidopsis,* который очень хорошо изучен с точки зрения молекулярной генетики и функциональной аннотации генов. Коллекция мутантов была произведена с использованием системы транспозонов-ловушек (англ. gene trap), сконструированных методом биоинженерии на основе природных транспозонов кукурузы. Система состояла из двух независимых мобильных элементов, содержащих гены активации и диссоциации (*Ac/Ds*). При скрещивании растений, содержащих в своем геноме встроенный транспозон *Ds,* с растениями,

содержащими источник траспозазы *Ac, Ds* элемент приходил в движение с дальнейшим встраиванием в новую локацию. При условии исключения *Ac* элемента из генома в последующем поколении в результате рекомбинации *Ds* элемент становился иммобилизованным. Поскольку *Ds*-элемент был сконструирован таким образом, что в его состав входили селективный маркер - ген устойчивости к антибиотику канамицину *NPTII* и ген – репортер *GUS* (ген, кодирующий β- глюкуронидазу), встраивание *Ds* в рамку считывания какого-либо гена растения приводило к появлению устойчивости растения к канамицину и возможности выявления



(a) (b)

Рисунок 3.1.1.2. Стратегия двухэтапного скрининга для идентификации гаметофитных мутантов в GT-линиях модельного растения арабидопсиса. (а) На первом этапе оценивали стручки на предмет семистерильности (уменьшения завязываемости семян). В случае гетерозиготности по летальной гаметофитной мутации около половины семязачатков останавливают свое развитие и не образуют семя. У растений, гетерозиготных по летальной зиготическо-эмбриональной мутации, около 25% семян прерывают свое развитие, тогда как мутации с гаметофитным материнским типом наследования обнаруживают около 50% абортивных семян. Фенотип определялся путем вскрытия зеленых или сухих стручков мутантных растений. Стрелки указывают на абортивные семена. (б) На втором этапе коэффициент расщепления в сегрегирующейся популяции встроенного *Ds*-транспозона определялся благодаря содержащемуся в нем доминантному маркеру, гену устойчивости к антибиотику канамицину *NPTII*. Если инсеция *Ds*-транспозона не влияет на передачу гена через гаметофит, соотношение устойчивых к канамицину проростков к неустойчивым будет наблюдаться в менделевской пропорции 3: 1, если инсерция вызывает эмбриональную летальность, то расщепление будет в соотношении 2: 1, а если она вызывает женскую гаметофитную летальность, то соотношение устойчивых и неустойчивых к канаимицину проростков будет 1: 1 [59,60].

локализации экспресии этого гена за счет окраски β- глюкуронидазы, производимой геном *GUS*, специфическими красителями (рис. 3.1.1.1). Физическая локализация нокаутированного при помощи *Ds* элемента гена в геноме осуществлялась с помощью амплифиции методом TAIL- ПЦР (thermal asymmetric interlaced PCR), который помогал определить хромосомные последовательности, фланкирующие вставки *Ds* элемента. Благодаря описанной выше системе нами была создана коллекция из более чем 2,5 тыс. новых вставок *Ds*, распределенных случайным образом по всему геному *Arabidopsis* [8]. Полученные транспозиционные линии мутантов (GT-линии от англ. GT-lines) исследовали для выявления генов, влияющих на развитие и функцию женских гаметофитов (зародышевых мешков), а также анализировались фенотипы этих линий. У семенных растений жизненный цикл чередуется между доминирующей диплоидной фазой (спорофит) и сильно редуцированной гаплоидной фазой (гаметофит). У покрытосеменных гаплоидный женский гаметофит (зародышевый мешок) образуется из функциональной мегаспоры, единственной выжившей клетки, возникшей благодаря мейозу, путем трех митотических делений. Женский гаметофит обычно состоит всего из семи клеток, двух женских гамет (яйцеклетки и центральной клетки) и пяти дополнительных клеток. Он труднодоступен, поскольку развивается внутри семязачатка, внедренного глубоко в гинецей цветка. Мутации, влияющие на гаметофитную фазу жизненного цикла, были идентифицированы двухэтапным скринингом, направленным на поиск: (1) снижения фертильности (абортивные семена или неразвитые семязачатки) и (2) нарушения соотношения сегрегации доминирующего маркера (в данном случае устойчивости к антибиотику канамицину), присутствующего на транспозоне *Ds* (рис. 3.1.1.2).

Отклонение от менделевского расщепления 3: 1 соотношения устойчивых к канамицину (Kan^r) относительно чувствительных (Kan^s) проростков в сторону соотношения менее 2: 1 является характеристикой мутаций, затрагивающих один или оба гаметофита (мужской и женский). Оценка всей популяции, состоящей из 2511 GT-линий, относительно расщепления гена устойчивости к антибиотику Kan^r, присутствующего на *Ds* элементе, показало, что у 9,5% GT-линий соотношение Kan^s /Kan^r проростков было от 2: 1 до 0,14: 1. В результате двухэтапного скрининга нам удалось выделить 12 линий гаметофитных мутантов и 12 эмбриолетальных линий.

Ниже дается описание шести мутантов, фенотипы которых, преимущественно затрагивают развитие или функцию гаметофитов, а также одного мутанта, в котором гомозиготное состояние нарушенного гена приводит к летальности зародыша (таблица 3.1.1.1). Гаметофитные мутанты были названы в честь богов или богинь из различных пантеонов, которые, согласно легенде, оказывали влияние на плодородие и

размножение. Были получены генетические и цитологические характеристики этих мутантов. Согласно фенотипу и сценарию ареста зародышевых мешков мутанты были классифицированы на *митотический* (обнаруживают отклонения в одном или нескольких из трех митотических делений при формировании зародышевого мешка), *кариогамный* (затрагивают слияние полярных ядер зародышевого мешка), *гаметофитный материнский* (англ. gametophytic maternal; фенотип проявляется только после оплодотворения, абортивность семян зависит только от генотипа женского гаметофита, полученного по материнской линии, независимо от отцовского вклада) и *дегенеративный* (демонстрирует спонтанную аномальную дегенерацию ядер при развитии зародышевого мешка) классы.

Таблица 3.1.1.1. Фенотип, генетическая сегрегация и трансмиссионный анализ гаметофитных и эмбриолетальных мутантов, а также функциональная идентификация нарушенных генов.

мутант	фенотип	Kanr : Kans	Значение P	TEF	TEM	ген инсерции Ds	Функция нарушенного гена
kupalo	58% N+42% UO	1.04:1	0.81	40%	68%	At2g01070	PTM1-подобный белок
astlik	49%N+51%UO	1.66:1	0.23	61%	98%	Делеция в At3g03030 At3g03040 At3g03050 At3g03060 Ds3'At1g75990	семейство белков F-box; семейство белков F-box; целлюлозасинтаза-подобный белок; семейство белков АТФазы ААА-типа; семейство белков АТФазы ААА-типа; RPN3b субединица протеосомы 26S;
amon	59%N+36%UO+5%A	1.13:1	0.39	56%	24%	At4g02700	
apis	61%N+31%UO+8%A	0.32:1	NA	5%	8%	At5g44520	
yarilo	68%N+29%UO+3%A	0.30:1	NA	15%	9%	Ds 5' At2g34680	AIR9 индуцированный ауксином в корневой культуре
didilia	70%N+5%UO+25%A	1.05:1	0.71	34%	59%	At2g01110	TATC–подобный белок, НЕОПЛОДОТВОРЕННЫЙ ЗАРОДЫШЕВЫЙ МЕШОК-3
rpn1a	76%N+24%A	1.95:1	0.47	98%	NA	At2g20580	RPN1a субединица протеосомы 26S
wt	93%N+6%UO+1%A	3.0:1		100%	100%		

N - нормальные семена; A- абортированные семена; UO - неразвившиеся семязачатки. Wt – растения дикого типа. Значение P ≥0,05 основано на ожидаемом соотношении сегрегации Kan^r / Kan^s 1: 1 для летальной гаметофитной мутации или 2:1 для эмбриолетальной мутации. Эффективность трансмиссии рассчитывалась согласно: TE = Kan^r / Kan^s x 100%; Kan^r - проростки, устойчивые к канамицину; Kan^s, проростки, чувствительные к канамицину; TEF - эффективность женской трансмиссии, TEM - эффективность мужской трансмиссии [8,60].

Оценка фертильности плодов показала, что у шести исследуемых мутантов, отнесенных к типу гаметофитных, неразвившиеся семена составляли примерно от 30% до 50% от всех семян, что является типичным для гаметофитных мутнтов, тогда как в мутантной GT-линии *rpn1a* абортивность семян составляла 24%, что характерно для зиготических эмбриолеталей.

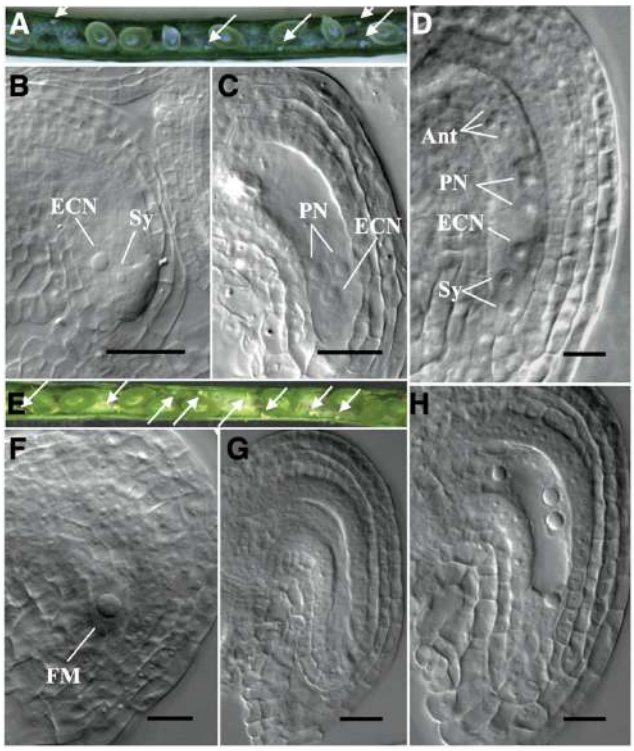


Рисунок 3.1.1.3. Фенотипы мутантов митотического класса *kupalo* и *astlik*. (A-C) *kupalo*, (D) дикий тип, (E-H) *astlik*. (A) Стручок, содержащий нормальные семена и бесплодные семязачатки (стрелки). (B) Дефектная клетка синергиды. (C) Аномальная целлюляризация зародышевого мешка с отсутствующими ядрами синергид. (D) Семиклеточный, 8-ядерный зародышевый мешок дикого типа, содержащий яйцеклетку, две синергиды, два полярных ядра и три антиподальные клетки. (E) Стручок, содержащий нормальные семена и бесплодные семязачатки (стрелки). (F) Арест функциональной мегаспоры. (G) Арестованный семязачаток без зародышевого мешка, но с продолжающимся ростом интегументов. (H) Четырехядерный зародышевый мешок, остановивший развитие. ECN, ядро яйцеклетки; FM - функциональная мегаспора; PN - полярные ядра; Ant, антиподы; Sy, синергиды. Масштабная линейка 10 мкм. (B-D и F-H) Просветленные хлоралгидратом семязачатки под дифференциальной интерференционно-контрастной оптикой (DIC) [8].

Таким образом, мутанты *kupalo* и *astlik* были отнесены к митотическому классу, поскольку при их развитии наблюдался арест

делящихся ядер зародышевого мешка (рис. 3.1.1.3). *amon* и *apis*, в которых обнаружены нарушения слияния полярных ядер центральной клетки зародышевого мешка, были выделены в кариогамный класс. Абортивность семязачатков в мутантах *yarilo* была вызвана дегенерацией центральной клетки зародышевого мешка, яйцеклетки и синергид, поэтому этот мутант был отнесен к дегенративному классу. Мутант *didilia* демонстрировал гаметофитный материнский дефект, который проявляется на стадии после оплодотворения и зависит только от генотипа женского гаметофита, что приводит к абортивности семян (таблица 3.1.1.1.) [8]. После сегрегационного, трансмиссионного и цитоэмбриологического анализа мутантов (таблица 3.1.1.1., рис. 3.1.1.3.) важно было определить генетическую основу фенотипа, то есть выяснить функцию генов, нарушение работы которых приводит к наблюдаемым фенотипам.

Для идентификации последовательностей, фланкирующих вставки *Ds*-транспозонов в прерванных им генах, хромосомные последовательности, фланкирующие *Ds*, амплифицировали с помощью TAIL-PCR (thermal asymmetric interlaced ПЦР) согласно Liu et al. [61] с модификациями, описанными Grossniklaus et al. [21], с использованием вложенных праймеров на границах *Ds*-элементов. Также применяли метод обратной ПЦР (iPCR) с разрезанием ДНК рестриктазой *BstYI* для идентификации 5 ’конца и рестриктазой *NcoI* для 3’-конца [8]. Нарушенные гены были физически картированы и сопоставлены с геномной последовательностью *Arabidopsis* с помощью алгоритма поиска BLAST (www.Arabidopsis.org). Функция обнаруженных генов и их идентификационный шифр указаны в таблице 3.1.1.1, а структура этих генов показана на рис. 3.1.1.4. [8].

Анализ полученных мутантов свидетельствует о *плейотропизме* при аресте гаметофитов. Хотя женский гаметофит имеет небольшой размер и состоит из малого количества гаплоидных клеток (восемь ядер, семь клеток), в его развитии и функционировании принимают участие многие клеточные процессы. Несмотря на преобладающую стадию ареста, фенотип гаметофитных мутантов затрагивает также и другие стадии. То есть процесс развития редко останавливается на одной конкретной стадии, поскольку в гаметофите экспрессируется большое количество генов. Фенотипическая изменчивость также может быть вызвана и различной степенью переноса мРНК, образовавшейся в результате более ранней экспрессии гена в материнской клетке мегаспоры, и / или генетической избыточностью в геноме *Arabidopsis*. Кроме того, многие мутанты обнаруживают гаметофитные материнские эффекты, которые вызывают абортивность семян на более поздних стадиях. Так, например материнский фенотип *did* может быть вызван нарушением экспрессии гена *At2g01110* (таблица 3.1.1.1.), что приводит к формированию дефектных пластид и / или митохондрий, наследуемых по материнской линии.

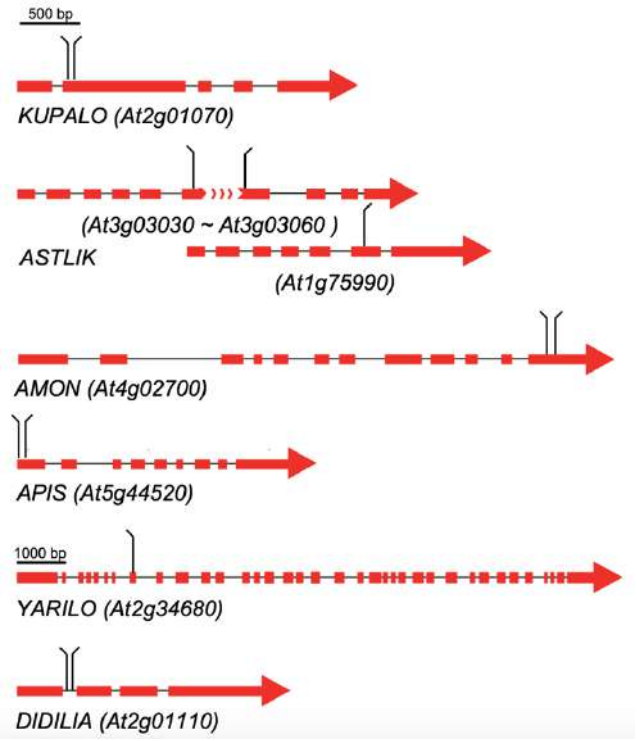


Рисунок 3.1.1.4. Схематическое изображение генов арабидопсиса и сайтов инсерции *Ds*-элементов. Показана интрон-экзонная структура нарушенных генов. 3’- и 5’-концы инсерции *Ds*-элементов обозначены вертикальными линиями, отмеченными слева и справа соответственно [8].

Поздний материнский эффект, который влияет на зрелые семена, может объяснять несоответствие между низкой эффективностью трансмиссии и мягким семистерильным фенотипом, наблюдаемым у *yar* и *aps* (таблица 3.1.1.1.). У этих мутантов семена имели нормальный внешний вид, но не прорастали, либо проростки погибали вскоре после прорастания. Многие женские гаметофитные мутации также затрагивали формирование пыльцы. Отклонения наблюдались не только во время развития пыльцы, но и в прогамной фазе, то есть во время прорастания пыльцевой трубки или распознавании ею зародышевого мешка и яйцеклетки.

Информативный метод для понимания организации, регуляции и функции генов – это метод гибридизации in situ. Он используется для выявления местоположения конкретных транскриптов в тканях либо известных фрагментов нуклеиновых кислот на хромосомах. Для локализации определенной мРНК в клетках или тканях (in situ) применяют меченую комплементарную цепь ДНК или РНК (пробу). Метод особенно эффективен для детекции генетических транскриптов в отдельных клетках или небольших тканях таких структур, как зародышевый мешок, семязачаток и семя. Гибридизация in situ со смысловой и антисмысловой пробами показала, что профили экспрессии генов *RPN1a* и *RPN1b* при

эмбриогенезе были очень похожи (рис. 3.1.1.7). И *RPN1a,* и *RPN1b* транскрипты были обнаружены в семенах, при этом наиболее сильные сигналы гибридизации наблюдались в халазальной части эндосперма и зародышах вплоть до поздней глобулярной стадии. Затем на стадиях сердечковидного, торпедовидного и зрелого зародыша уровни гибридизации значительно снижались, лишь немного превышая фоновый уровень антисмысловой пробы [60].

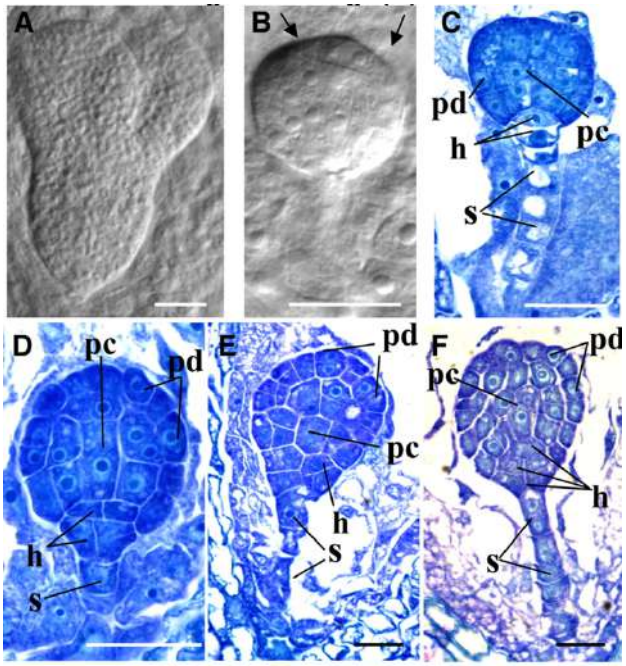


Рисунок 3.1.1.5 Развитие семян и зародышей растений мутантной линии *rpn1a-1 / RPN1a.* (A) и (B) Просветленные хлоралгидратом семена, полученные из одного и того же стручка: стадия торпедовидного зародыша в семени дикого типа. Семена наблюдали с помощью микроскопа Leica DMR (Leica Microsystems) под дифференциальной интерференционно-контрастной оптикой (DIC) (A), абортированные семена с задержкой развития зародыша на глобулярной стадии (B). Обратите внимание на нерегулярный слой протодермы (стрелки). (C) - (F) окрашенные толуидиновым синим срезы зародыша дикого типа на ранней глобулярной стадии (C) и мутантных зародышей линии *rpn1a* на глобулярной стадии с аномалиями ([D] - [F]). h, производные клетки гипофиза; pc - прокамбиальные клетки; pd, протодерма; s, суспензор. Масштабная линейка = 25 мкм в (A) и 50 мкм в (B) - (F) [60].

Еще одно исследование с использованием инсерционных мутантов арабидопсиса позволило нам обнаружить, что субъединица RPN1 протеасомы 26S арабидопсиса важна для нормального эмбриогенеза [21]. У эукариотических организмов протеасома 26S играет центральную роль в

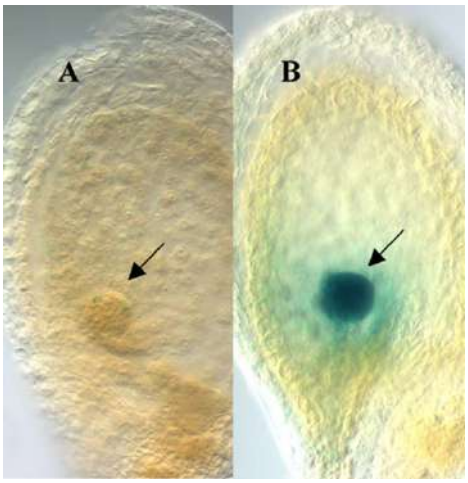


Рисунок 3.1.1.6. Экспрессия *циклина B1* в семенах мутанта *rpn1a*. Семена просветляли хлоралгидратом и наблюдали с помощью микроскопа Leica DMR (Leica Microsystems) под дифференциальной интерференционно-контрастной оптикой (DIC). (A) Зародыш, несущий по крайней мере одну аллель *RPN1a* дикого типа. Рекомбинантный белок *cyclinB1; 1-GUS* расщепляется и практически не обнаруживается в клетках зародыша (стрелка). (B) *rpn1a* гомозиготный мутантный зароды накапливает рекомбинантный белок *cyclinB1; 1-GUS* (стрелка) [60].

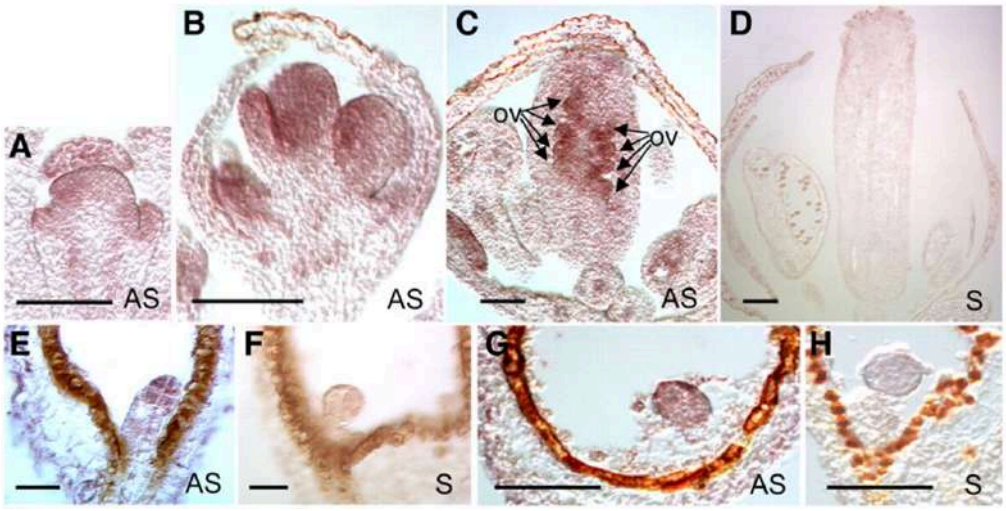


Рисунок 3.1.1.7. Детекция мРНК *RPN1a* in situ в развивающихся цветках, семенах и стручках. (A) - (D) Продольные срезы цветочной меристемы и бутонов на разных стадиях. (E) и (F) Зародыш на стадии октанта. (G) и (H) Зародыш на глобулярной стадии. AS, антисмысловая проба; S, смысловая проба; ch, халаза; ov, семзачатки. Масштабная линейка = 25 мкм на (A), (B), (E) - (H) и 50 мкм на (C), (D) [60].

деградации регуляторных белков, участвующих во множестве процессов при развитии организма. Протеосома состоит из мультибелковых комплексов: протеолитической коровой протеазы и регуляторной частицы (RP), которую образуют несколько субъединиц. Субъединица RPN1у *Arabidopsis thaliana* кодируется двумя паралогическими генами, *RPN1a* и *RPN1b*. Нокаут *RPN1a* вызывал летальность зародыша, в то время как мутанты по гену *RPN1b* не демонстрировали явного аномального фенотипа. Мутация *rpn1a* расщепляется как рецессивный моногенный признак с соотношением Kan^r:Kan^s ~ 2:1 (таблица 3.1.1.1). Эмбриональная летальность, вызванная нарушением гена *RPN1a,* была подтверждена реверсией мутантного фенотипа посредством удаления транспозона, скрещиванием *rpn1a* мутанта с растением, содержащими *Ac*-элемент и ремобилизацией *Ds*-элемента, а также анализом двух дополнительных мутантных аллелей *rpn1a,* обнаруживших аналогичный фенотип. Зародыши, гомозиготные по *rpn1a*, останавливали свое развитие на глобулярной стадии, что сопровождалось дефектами формирования зародышевого корня, протодермы и прокамбия (рис. 3.1.1.5). При этом в зрелых плодах наблюдалось 24% абортированных семян. Было обнаружено, что в клетках зародышей абортированных семян не расщеплялся белок - регулятор клеточного цикла, *циклин B1*, который как известно, экспрессируется только во время перехода эукариотических клеток от стадии G2 клеточного цикла к M, возможно, что избыток циклина B1 стал причиной неспособности клеток зародыша к дальнейшему делению и их ареста на этой стадии (рис. 3.1.1.6). Полученные методом трансформации двойные мутантные с нокаутом обоих генов *rpn1a* (*rpn1a / RPN1a)* и *rpn1b* (*rpn1b / rpn1b*) образовывали зародыши с фенотипом, неотличимым от фенотипа одиночного мутанта *rpn1a*. Таким образом, несмотря на то, что они в значительной степени перекрывают паттерны своей экспрессии в цветках и развивающихся семенах, эти две изоформы не имеют общих избыточных функций во время гаметогенеза и эмбриогенеза. Однако комплементация мутации *rpn1a* кодирующей областью *RPN1b*, экспрессируемой под контролем промотора *RPN1a*, указывает на то, что две изоформы RPN1 функционально эквивалентны. В целом полученные результаты указывают на то, что активность частицы RPN1 является необходимой во время эмбриогенеза, где она возможно участвует в деградации определенных белковых субстратов.

Вопросы:

1) Что такое прямая и обратная генетика?
2) Как работает метод инсерционного мутагенеза и для чего он используется?
3) Что такое сегрегационный анализ?

4) Как скринируются гаметофитный и зиготические мутации?
5) В чем особенности метода гибридизации in situ?

3.1.2. Генетическая гомология. Сходство генов и белков

Сравнительный анализ последовательностей нуклеотидов ДНК в генах и аминокислот в белках у организмов из разных систематических групп позволяет выявить определенные сходства, которые по аналогии со сходством в строении частей и органов тела у разных организмов, имеющих общее происхождение, назвали *гомологичными*. При этом гомологичные последовательности могут быть *ортологичными*, которые возникли за счет мутационных изменений при образовании новых видов, и *паралогичными*, которые возникли за счет увеличения копий гена в геноме одного организма и появления в них мутаций.

Примером гомологичных генов (гомологов) может быть описанный выше ген арабидопсиса *KUPALO* (*At2g01070*), кодирующий PTM1-подобный трансмембранный белок, и ген *Lung seven* трансмембранного белка риса, PTM1-подобного трансмембранного белка дрожжей и ряда генов у других организмов, поскольку эти гены имеют общего предка и выполняют сходную функцию.

Типичный пример паралогичных генов (пралогов) — это описанные выше гены *RPN1a* и *RPN1b*. Оба этих гена и их белки имели высокое сходство на уровне нуклеотидов и аминокислот соответственно. Однако эти гены не являлись избыточными, поскольку *RPN1b* не компенсировал функцию поврежденного паралога *RPN1a,* кроме того, эти гены располагались на разных хромосомах, а биоинформатический анализ показал, что они разошлись за счет дупликации генома растений, произошедшего миллионы лет назад. Тем не менее, оба этих гена кодировали белок субъединицы RPN1 протеасомы 26S и являлись функционально эквивалентными, поскольку трансляция гена *RPN1b* под более сильным промотором вызывала комплементацию нарушенного гена *rpn1a.* По-видимому, *RPN1b* является «спящим» геном и имеет очень слабый промотор, так как его экспрессия у растений дикого типа намного ниже экспрессии его паралога *RPN1a* [60].

Благодаря появлению компьютерной биологии (биоинформатики) и наличию специально разработанных алгоритмов, стало возможным осуществление выравнивания белков друг на друга с целью поиска участков гомологии путем определения консервативных аминокислотных последовательностей, то есть последовательностей, которые наиболее редко подвержены изменениям у различных видов организмов. Для определения гомологичных последовательностей в геномных и белковых базах данных производится поиск аналогичных последовательностей, уже описанных в других белках с известной функцией у других видов. Также

осуществляется анализ мотива и/или домена, поскольку, как уже было сказано выше, некоторые аминокислотные последовательности имеют общую структуру в виде «мотивов» или «доменов». На рис. 28 представлена круговая диаграмма анализа всех ~25 тысяч генов арабидопсиса, для которых обнаружена открытая рамка считывания (ORF), согласно их функциональной гомологии и пропорциональному распределению относительно общего количества генов. Из диаграммы видно, что около половины генов до сих пор не классифицировано. Среди генов, функциональная принадлежность которых определена или предсказана по гомологии, большая доля приходится на гены, кодирующие транскрипционные факторы, регуляторы синтеза ДНК и деления клеток, а также на гены, ответственные за защиту клетки от стресса, старение и гибель клеток и за сигнальную трансдукцию. Функция межклеточного транспорта и доставки белков также занимает значительную долю среди всех белок кодирующих генов. Меньше всего приходится на гены, ответственные за ионный гомеостаз, синтез структурных белков и выработку энергии (рис. 3.1.2.1).

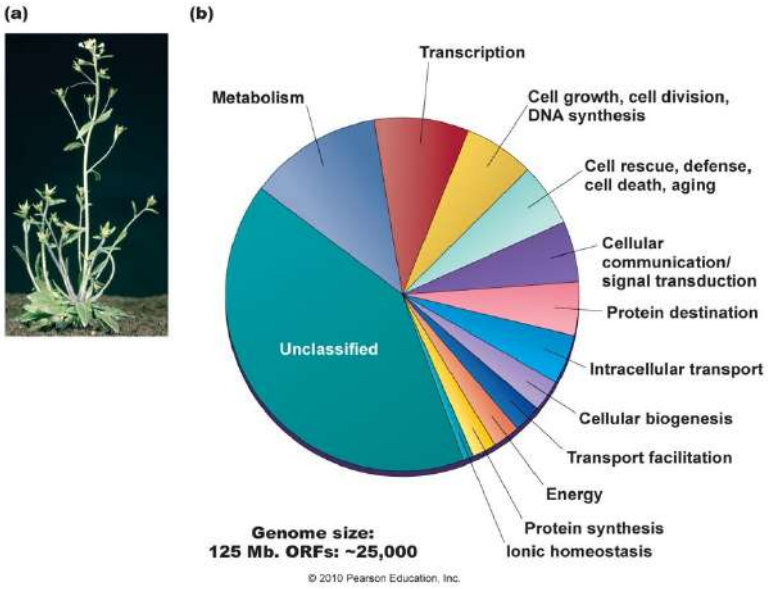


Рисунок 3.1.2.1. Модельное растение арабидопсис (a) и круговая диаграмма, представляющая пропорциональное распределение генов арабидопсиса согласно их функциональной гомологии (b).

В нашей работе по сборке de novo и аннотации генома растения *Boechera retofracta* из семейства капустных (Brassicaceae) для предсказания генов, кодирующих белок, мы использовали комбинированный подход, основанный на предсказании на основе гомологии и предсказании de novo, последний использовался только для заполнения пробелов в сборке генома и для расширения предсказаний на основе гомологии [62]. Для

доказательства наличия генов на основе гомологии использовали белки и транскрипты четырех близкородственных видов из семейства Brassicaceae: *Arabidopsis thaliana* (assembly TAIR10), *Brassica rapa* (Brapa_1.0), *Capsella rubella* (Caprub1_0), *Eutrema salsugineum* (Eutsalg1_0), которые были выровнены на сборку *B. retrofracta* с помощью программы Exonerate с использованием модели protein2genome, при этом делали максимум три выборки на белок. В итоге в геноме *B. retrofracta* было предсказано 27048 генов и 28269 транскриптов, разница наблюдалась за счет генов, кодирующих тРНК и рРНК. В процессе распределения белков по ортологическим группам выбирались самые длинные белки, соответствующие каждому предсказанному гену у *B. retrofracta* и шести других видов из семейства Brassicaceae: *B. stricta* (сборка v1.2), *Arabidopsis thaliana* (TAIR10), *Arabidopsis lyrata* (v.1.0), *Capsella rubella* (Caprub1_0), *Cardamine hirsuta* (v1.0), *Eutrema salsugineum* (Eutsalg1_0). Белки были выровнены по профилю HMM (англ. Hidden Markov Model) подмножества braNOG из базы данных eggNOG с использованием HMMER. Наиболее близкие совпадения из выравниваний были извлечены и использованы для сопоставления с соответствующими ортологическими группами белков с последующим выделением ортологов с одной копией.

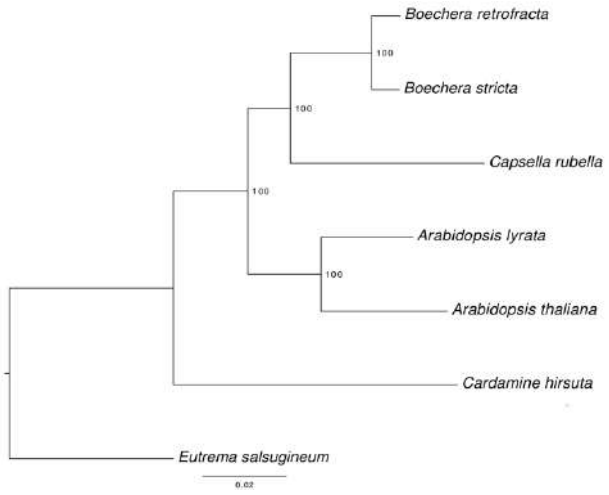


Рисунок 3.1.2.2. Филогенетическое дерево семи видов Brassicaceae. Дерево было укоренено с помощью *E. salsugineum* в качестве внешней группы. Числа в узлах соответствуют поддержке бутстрапов [62].

Однокопийные ортологичные белки семи видов, включенных в анализ, сравнивались путем множественного выравнивания при помощи быстрого преобразования Фурье (MAFFT) и программы Exonerate с использованием модели cDNA2Genome. Предсказание проводилось с помощью программного пакета AUGUSTUS. Белки были транслированы из предсказанных генов и сопоставлены с помощью HMMER 3.1 и BLAST

алгоритмов с базами данных Pfam и Swiss-Prot соответственно. Гены транспортной РНК (тРНК) и рибосомной РНК (рРНК) были предсказаны с помощью программ tRNAscan-SE v1.3.1 и Barrnap v0.6 соответственно. Всего у семи исследованных видов было идентифицировано 8959 однокопийных ортологов. На их основе с помощью программы RAxML 1000 раз было реконструировано и протестировано дерево максимального правдоподобия, которое было укоренено с помощью *Eutrema salsugineum* в качестве внешней группы (рис. 3.1.2.2) [62].

Кроме того, мы исследовали локус гистидиновой экзонуклеазы *APOLLO*, который ассоциирован с апомиксисом у *Boechera*, и предложили модель его эволюции через серию дупликаций. *APOLLO* наследуется биаллельно, через «апоаллели» и «секс-аллели», различающиеся по своему полиморфизму и кодирующие факторы транскрипции. Все апомиктические линии *Boechera spp.* гетерозиготны по аллелям *APOLLO* (то есть имеют по крайней мере одну апоаллель и одну половую аллель), в то время как все половые генотипы были гомозиготными по секс-аллелям [62]. Эволюционная история гена *APOLLO* была проверена с помощью метода максимального правдоподобия. Первоначальное выравнивание соответствующих кодирующих последовательностей (CDS) было выполнено с помощью программы prank v.140110 с учетом кодонов. Результат выравнивания в дальнейшем был использован для построения филогенетического древа на основе модели Тамура-Нея. Было выбрано дерево с наибольшим логарифмическим правдоподобием (-12 153,79). Исходное дерево (a) для эвристического поиска было получено автоматически путем применения алгоритмов Neighbor-Join и BioNJ к матрице попарных расстояний, оцененных с использованием подхода максимального сложного правдоподобия (MCL), а затем выбора топологии с превосходным значением логарифмического правдоподобия. Все позиции, содержащие пробелы и недостающие данные, были удалены. В окончательном наборе данных было всего 1158 позиций. Эволюционный анализ проводился программой MEGA7 [62].

Ветви на дереве были сгруппированы по генам, а не по видам, предполагая, что событие трипликации имело место до разделения видов Brassicaceae. Ветвь, ведущая к апо-аллелям, находится под положительным отбором (Ka / Ks = 1.4646), что типично для паралогов, которые необходимы для выполнения новой функции. Коэффициент отношения Ka/Ks используется для оценки баланса между нейтральными мутациями, очищающим отбором и полезными мутациями, которые воздействуют на гомологичные кодирующие белок гены. Коэффициент рассчитывается как отношение количества замен несинонимичных аминокислот в несинонимичном сайте (Ka) за данный период времени к количеству синонимичных замен в синонимичном сайте (Ks) за тот же период (см.

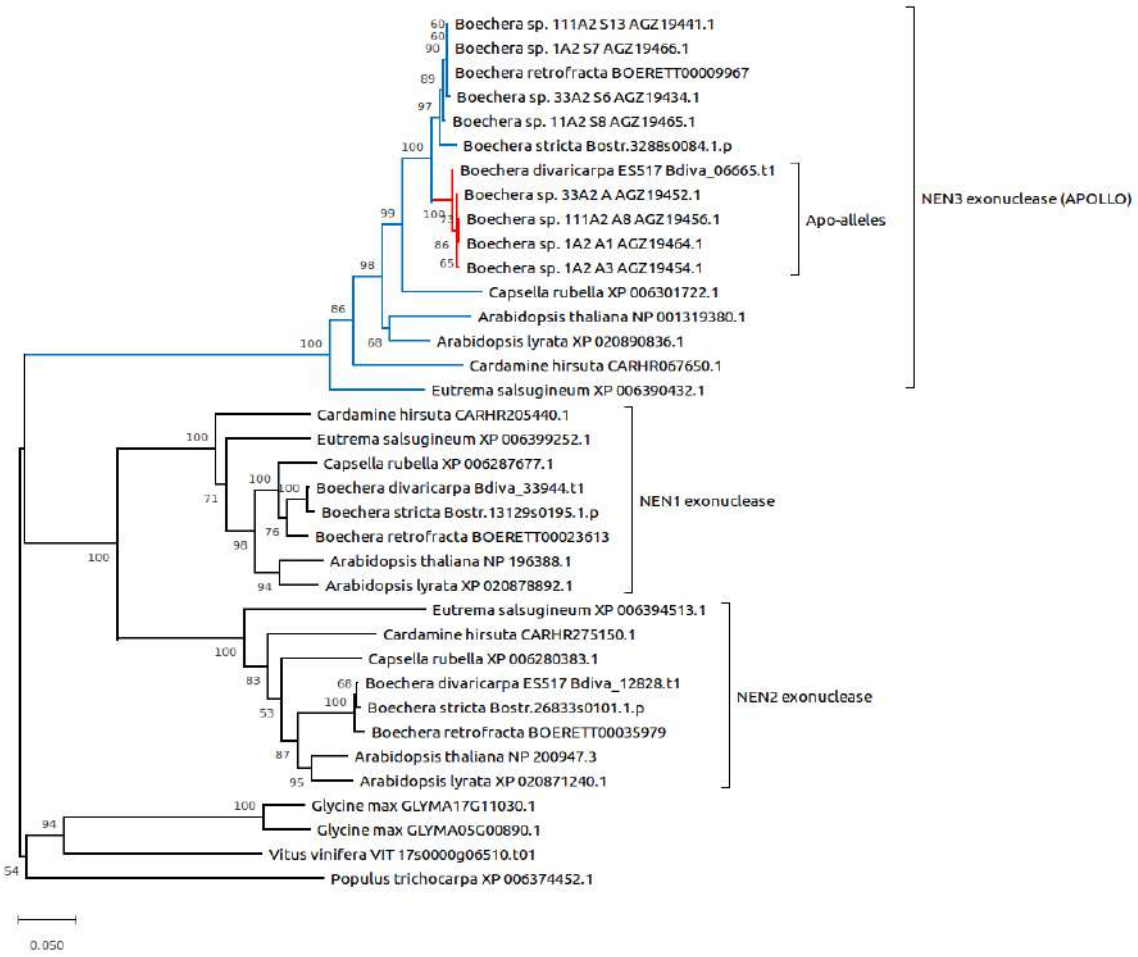


Рис. 3.1.2.3. Филогенетическое дерево эволюции изоформ локуса *APOLLO* (экзонуклеаза NEN) у семи видов и аллелей локуса *APOLLO* апомиктических видов *Boechera*. Последовательности *Populus trichocarpa, Vitus vinifera* и *Glycine max* использовали в качестве внешней группы. Клада, относящаяся к локусу *APOLLO*, выделена голубым цветом, а апоаллели - красным. Числа возле узлов представляют соответствующую поддержку бутстрапов [62].

параграф 1.2.1). Синонимичные замены считаются нейтральными, так что соотношение указывает на чистый баланс между вредными и полезными мутациями. Значения Ka/Ks, значительно превышающие 1, маловероятны и могут наблюдаться, если по крайней мере некоторые из мутаций будут полезными. Если допустить, что полезные мутации вносят небольшой вклад, то Ks оценивает степень эволюционных ограничений.

Мутационный анализ, сравнение генетических гомологов и изучение их филогении определяют незаменимую стратегию исследований в функциональной генетике. Однако для изучения функции генов на полногеномном уровне основными методами исследования являются анализ транскриптомов и протеомов.

Вопросы:

1) Что такое гомология, ортология и парология?
2) Как используются ортологи для филогении?
3) Что такое синонимичные и несинонимичные аминокислоты?
4) Как определяются гены под положительным эволюционным отбором?

4. Функциональная геномика и современные методы анализа

4.1. Функциональный анализ транскриптома

Когда геном человека был полностью секвенирован, фокус внимания сместился на идентификацию и аннотирование его функциональных элементов, включая те, которые регулируют экспрессию генов. Идентификация таких элементов, в том числе выявление геномных вариантов, связанных с устойчивостью и восприимчивостью к различным заболеваниям, является важным шагом к развитию персонализированной прецизионной медицины (Проект «Энциклопедия элементов ДНК» (ENCODE) [63].

Все процессы на уровне РНК, включая активацию или ингибирование транскрипции, процессинг мРНК и ее транспорт, регулируются различными функциональными элементами геномной ДНК. Тем не менее, наиболее важная регуляция происходит на уровне инициации транскрипции с помощью регуляторных элементов, которые называются *цис*-регуляторными элементами/ последовательностями (участки некодирующей ДНК, которые регулируют транскрипцию соседних генов: IRES, ТАТА-бокс и др.) и *транс*-регуляторными элементами/ последовательностями - представляют собой последовательности ДНК, кодирующие вышестоящие регуляторы, т. е. *транс*-действующие факторы, которые могут изменять или регулировать экспрессию отдаленных генов [64].

Факторы транскрипции (TF), такие как активаторы и репрессоры (включая коактиваторы и корепрессоры), взаимодействуют со специфическими участками ДНК, то есть с *цис*-действующими регуляторными последовательностями/элементами, которые включают основной промотор (с ТАТА-боксом и другими связывающими элементами), проксимальный промотор, энхансер, сайленсер, инсулятор и область контроля локуса (LCR). Исследование перечисленных регуляторных элементов является сложной задачей для ученых ввиду трудностей в идентификации положения региона старта транскрипции (TSS - transcription start sites) и региона связывания факторов транскрипции (TFBS - transcription factor binding site) в коровом промоторе. Однако на данный момент существует несколько экспериментальных (рисунок 4.1.1.) и биоинформатических подходов для решения этой задачи [64].

4.1.1. Функциональные анализы, которые измеряют активность регуляторного элемента транскрипции

Один из наиболее универсальных методов определения и анализа активности регуляторного элемента транскрипции основан на

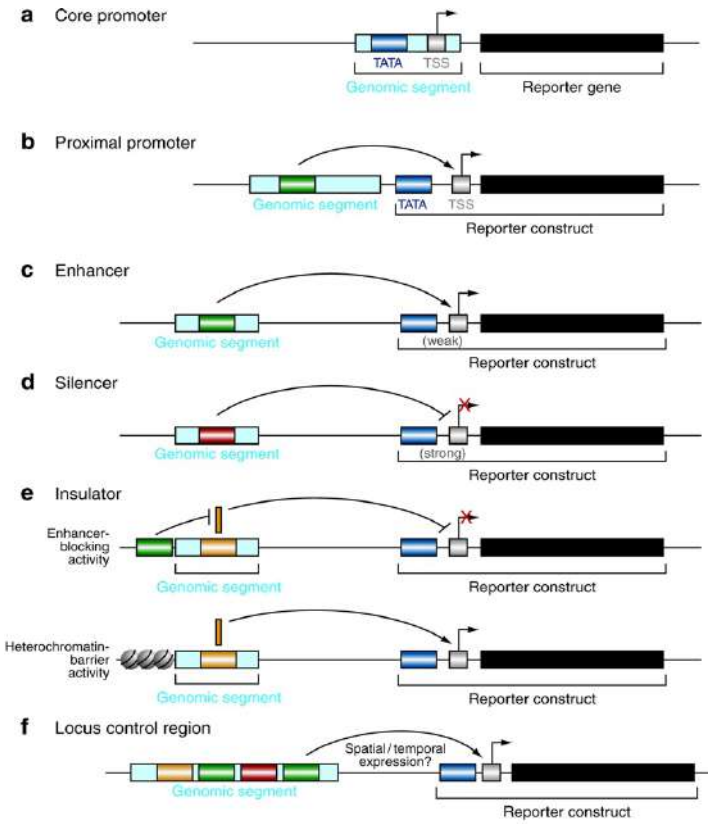


Рисунок 4.1.1. Принципы функционального анализа активности регуляторных элементов транскрипции. Традиционные методы анализа активности элемента регуляции транскрипции базируются на использовании анализов на основе плазмид (черная основная линия на схемах) или трансгенных репортерных генов. (**a**) Для анализа активности корового промотора тестируемый геномный сегмент (голубой цвет) клонируют в плазмиду непосредственно перед репортерным геном, в котором отсутствует эндогенный промотор. (**b – d**) Проксимальные промоторы, энхансеры и сайленсеры могут быть проанализированы аналогичными методами, когда геномный сегмент клонируется выше репортерного гена, управляемого соответствующим промотором. (**e**) Активность по блокированию энхансера инсулятора может быть измерена с помощью анализа на основе плазмид, который отслеживает способность клонированного инсулятора вмешиваться в коммуникацию энхансер-промотор, тогда как методы, которые измеряют активность гетерохроматинового барьера, требуют анализа трансгенного репортера для определения способности инсулятора, чтобы защитить трансген от репрессивных эффектов гетерохроматина. (**f**) Способность области контроля локуса преодолевать эффекты положения и обеспечивать пространственную и/или временную экспрессию измеряют с помощью анализа трансгенного репортера [64].

использовании анализа репортерного гена. Выше мы уже упоминали использование репортерного гена GUS. Для подобного анализа участок ДНК, который будет тестироваться на регуляторную активность, клонируют в плазмиду, расположенную выше легко анализируемого репортерного гена, такого как хлорамфениколацетилтрансфераза (CAT), β-галактозидаза, зеленого флуоресцентного белока (GFP) или гена люциферазы. Для целей крупномасштабного скрининга геномные сегменты могут генерироваться случайным, ферментативным либо физическим способами. Полученная конструкция затем трансфицируется (временно или стабильно) в культивируемые клетки, и измеряется активность переносчика, чтобы определить, содержит ли тестируемый сегмент элементы, которые изменяют экспрессию репортергена (рисунок 4.1.1). Точная конфигурация репортерной конструкции зависит от идентифицируемого регулирующего элемента.

4.1.2. Геномный анализ сайтов связывания транскрипционных факторов

Было разработано несколько методов для идентификации сайтов связывания транскрипционных факторов TFBS (transcription factor binding site) в масштабе всего генома. Например, картирование сайтов гиперчувствительности к дезоксирибонуклеазе I - метод основан на обнаружении участков геномной ДНК, в которых состояние хроматина нарушено. Гиперчувствительное картирование сайтов к ДНКазе I также использовалось для обнаружения сайленсеров, инсуляторов и LCR [65]. Также в 2004 г была разработана методика высокопроизводительного обнаружения гиперчувствительных сайтов к ДНКазе I в полногеномном масштабе [66]. Такой подход является действенным по своей способности обнаруживать любой регуляторный элемент, связанный с нарушением хромотина; однако он ограничен, поскольку наличие гиперчувствительности к ДНКазе I в каком либо сайте подразумевает, но не доказывает, наличие лежащего в основе функционального элемента регуляции транскрипции.

Следующим шагом в методах анализа связывания транскрипционных факторов был метод ChiP-chiP - технология, сочетающая иммунопреципитацию хроматина (ChIP - Chromatin immunoprecipitation) с ДНК-микрочипом («чипом») (рисунок 4.1.2.1) [67]. ChIP-on-chip используется для исследования взаимодействий между белками и ДНК in vivo. В зависимости от белка, который служит мишенью для иммунопреципитации, метод может обнаруживать энхансеры, а также ядерные промоторы.

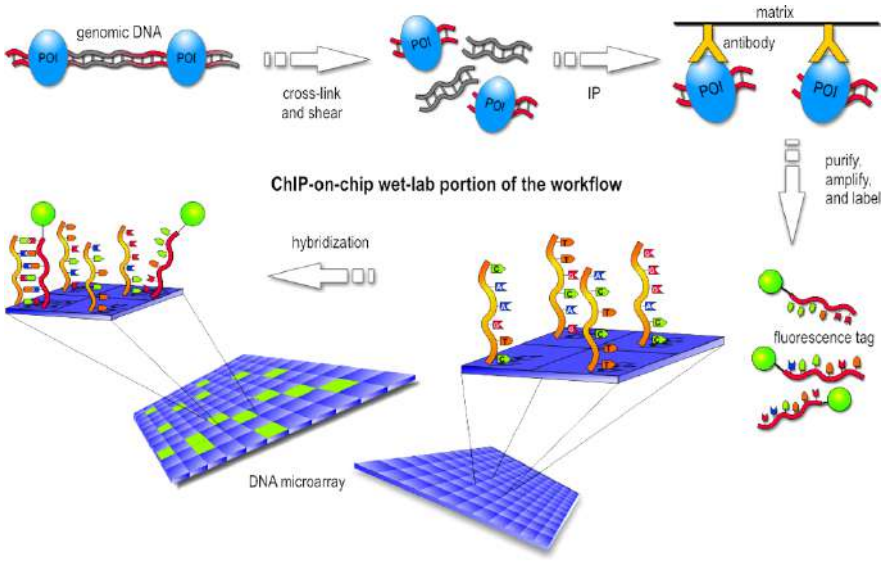


Рисунок 4.1.2.1. Рабочий процесс метода ChiP-chiP. POI - the protein of interest (интересующий белок).

ChIP-секвенирование, также известное как ChIP-seq широко применяется для точного картирования глобальных сайтов связывания для любого интересующего белка [68] рисунок 4.1.2.2. Прежде всего этот метод незаменим для идентификации эпигенетической регуляции генов, ассоциированных с закрытым или открытым хроматином, то есть метод дает возможность глобально оценить, какие из генов, регулируемых эпигенетически за счет ремоделинга хроматина, находятся в рабочем состоянии, а какие в подавленном. ChIP-seq сочетает в себе иммунопреципитацию хроматина (ChIP) с секвенированием ДНК для идентификации сайтов связывания ДНК-ассоциированных белков.

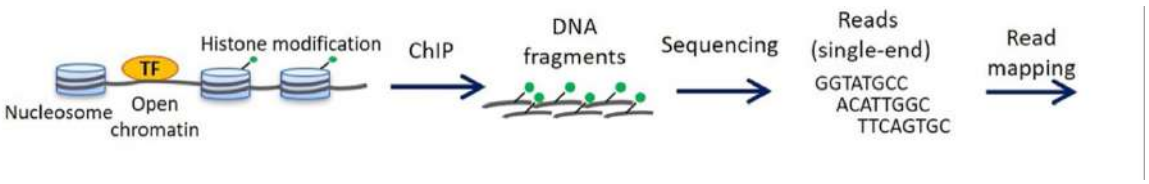


Рисунок 4.1.2.2. Рабочий процесс анализа ChIP-seq. Подготовка проб и секвенирование [69]

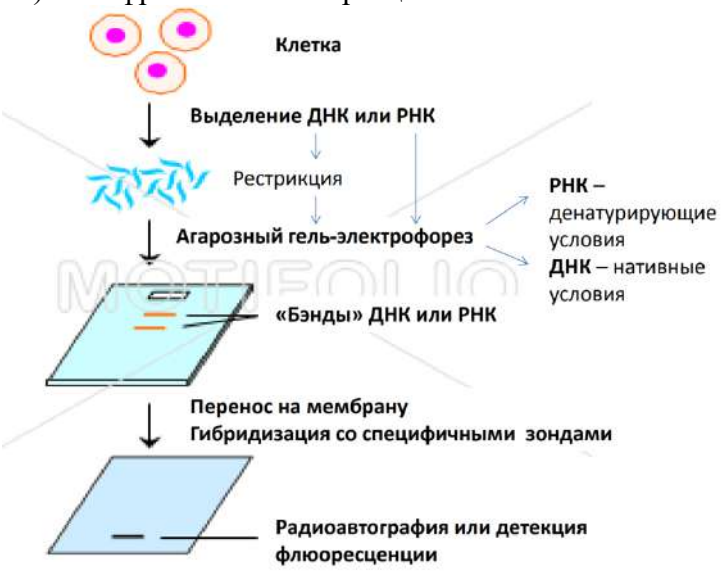
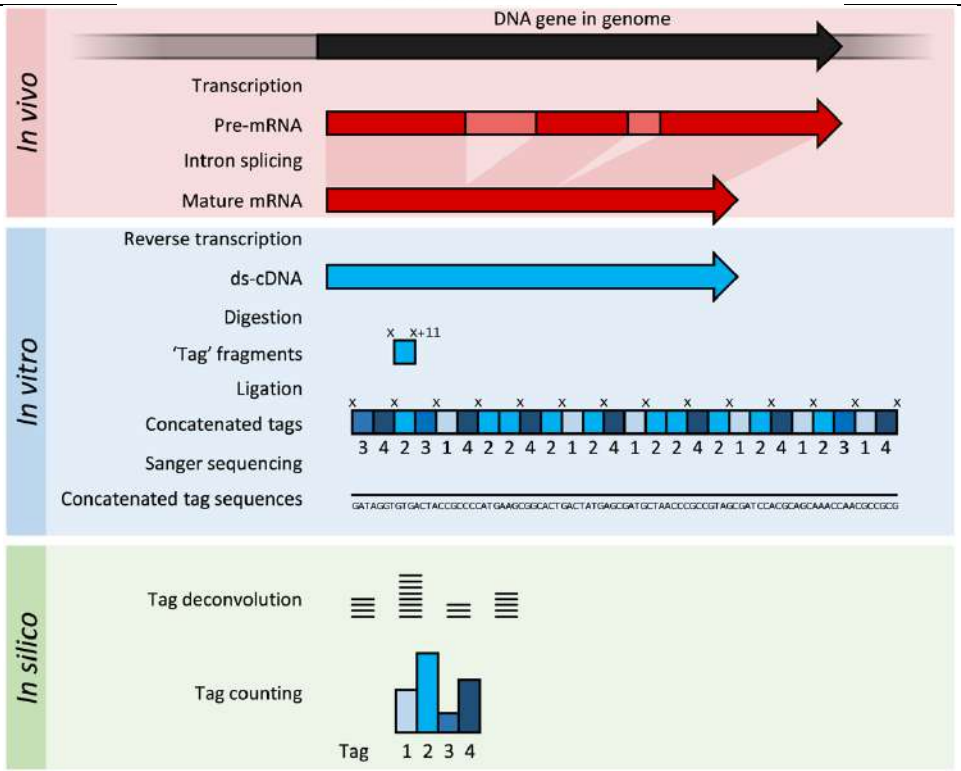
Многие типы клеток, включая иммунные клетки, могут играть вспомогательную функцию как в нормальных тканях, так и опухолях [70]. Для этих клеток scChIP-seq (Single-cell ChIP-seq) позволяет проводить оценку модификаций гистонов и других хроматин-связывающих белков в масштабе всего генома на уровне отдельных клеток, полученных из образцов с низким содержанием клеток. Недавно были разработаны методы подготовки библиотеки ChIP-seq (Таблица 4.1.2.1) с отдельными мечеными клетками, которые используют микрофлюидные системы, метку транспозазой Tn5 и стратегии без использования ChIP. Первый метод scChIP-seq, scDrop-ChIP [104], использует микрофлюидные системы для метки клеток в сочетании с каноническими методами ChIP для генерации ~ 800 недублируемых считываний на клетку. Капельный микрожидкостный метод [105] обеспечивает более высокое разрешение, производя около 10000 неповторяющихся считываний на клетку. Ограничением этих методик является то, что специализированные микрофлюидные устройства обычно недоступны в большинстве лабораторий.

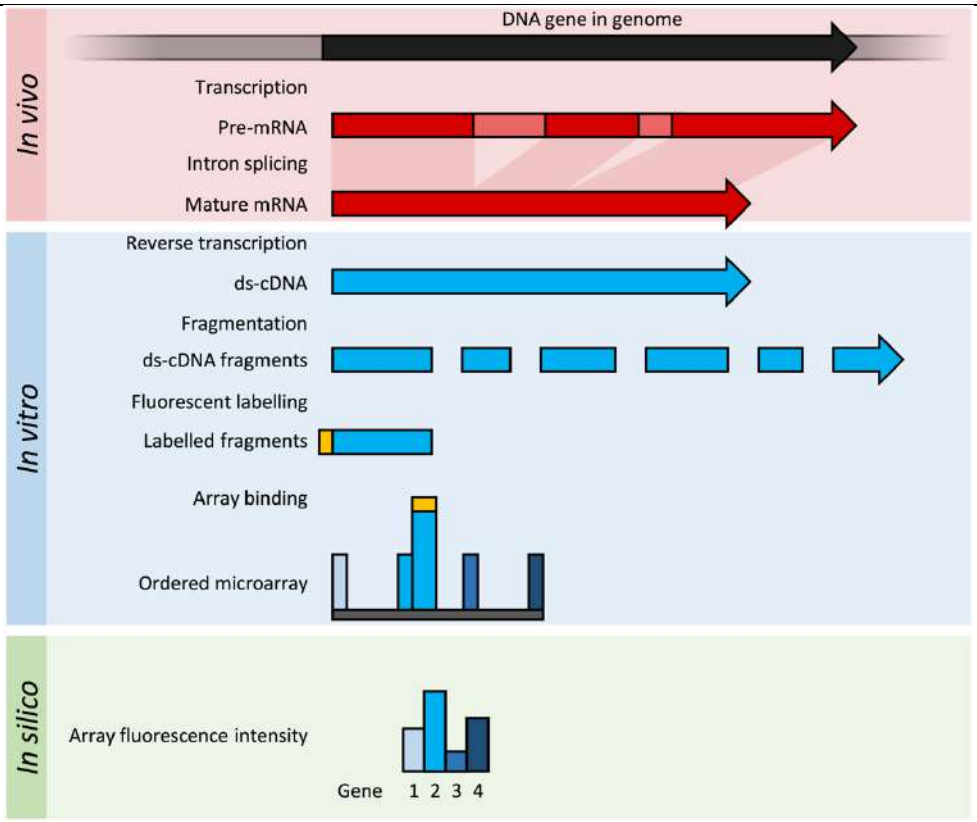
Таблица 4.1.2.1. scChIP-seq (Single-cell ChIP-seq) методы

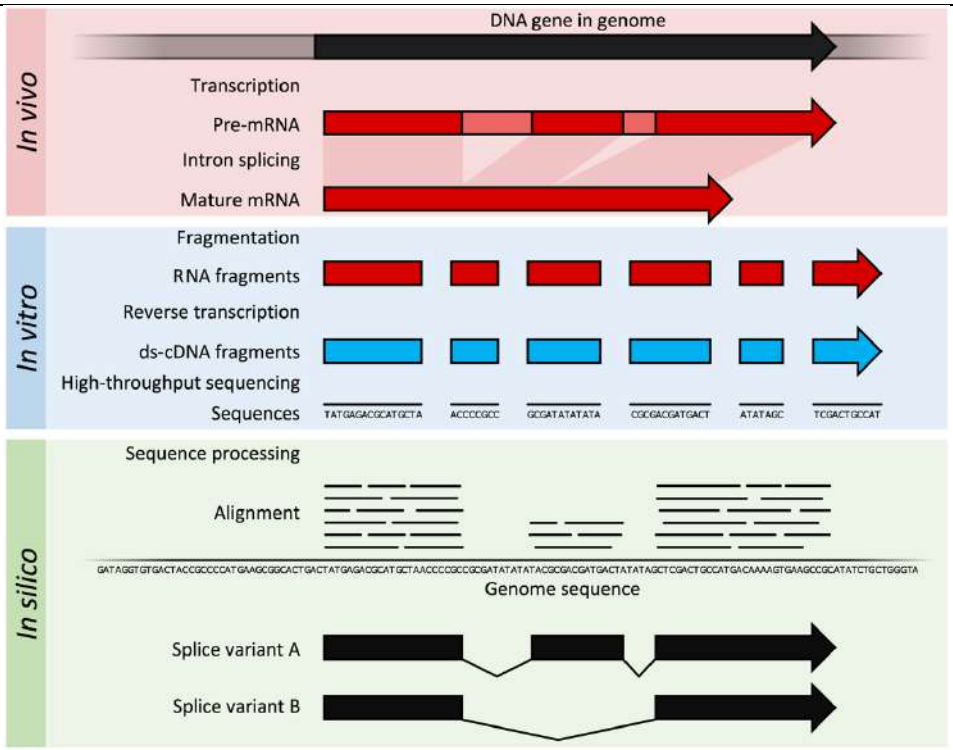
Методы	Принцип	Ссылки
scDrop-ChIP	ChIP и микрофлюидная система	[71]
sc-itChIP-seq	ChIP и тагментация	[72]
scChIC-seq	ChIP-free (расщепление с помощью монококковой нуклеазы)	[73]
CUT&Tag	ChIP-free (тагментация)	[74]
ACT-seq	ChIP-free (тагментация)	[75]
CoBATCH	ChIP-free (тагментация)	[76]

В таблице 4.1.2.2. перечислены технологии, используемые для транскриптомики.

Таблица 4.1.2.2. Технологии транскриптомики

Технология	Принцип
Northern Blot	Метод исследования экспрессии генов путём тестирования молекул РНК (мРНК) и их фрагментов в образцах 
SAGE/CAGE	 Серийный анализ экспрессии генов (SAGE) [77] – метод, основанный на секвенировании по Сэнгеру соединённых фрагментов случайных транскриптов.

	Метод кэп-анализа экспрессии генов (CAGE) представляет собой вариант SAGE, который устанавливает последовательность тегов только с 5'-конца транскрипта мРНК [78].
qPCR	Быстрый, точный, чувствительный и воспроизводимый метод количественного определения мРНК в режиме реального времени
cDNA microarray	 Микрочипы состоят из коротких нуклеотидных олигомеров, известных как «зонды», которые расположены на твердой подложке (например, стекле). Транскрипты из пробы гибридизуются с флуоресцентно меченными транскриптами микрочипа [77].

RNA-Seq	 RNA-Seq относится к комбинации методов высокопроизводительного секвенирования с вычислительными методами для идентификации и количественной оценки транскриптов, присутствующих в общей РНК, выделенной из определенных тканей или клеток. Генерируемые нуклеотидные последовательности обычно имеют длину около 100 п.н., но могут варьироваться от 30 п.н. до более чем 10 000 п.н., в зависимости от используемого метода секвенирования. RNA-Seq использует глубокую выборку транскриптома со множеством коротких фрагментов, чтобы обеспечить вычислительную реконструкцию исходного транскрипта РНК путем выравнивания ридов на эталонный геномом или друг на друга (сборка de novo) [77].

4.2. Бионформатический анализ транскриптома

Методы транскриптомики требуют больших компьютерных вычислений как для обработки данных, полученных на микрочипах (microarrays), так и для обработки результатов RNA-seq. Схемы процессов биоинформатического анализа данных, полученных на микрочипах cDNA и RNA-seq, представлены на рисунках 4.2.1 и 4.2.2 соответственно.

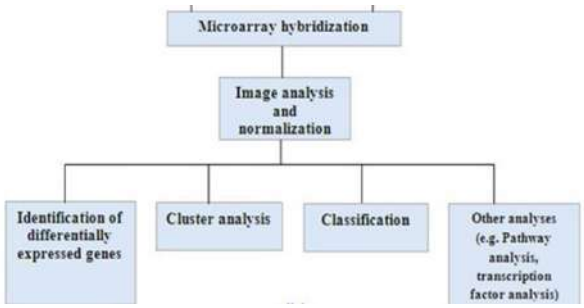


Рисунок 4.2.1. Схема анализа данных, полученных технологией cDNA microarray [79]. Со списком программ для работы с данными можно ознакомиться там же.

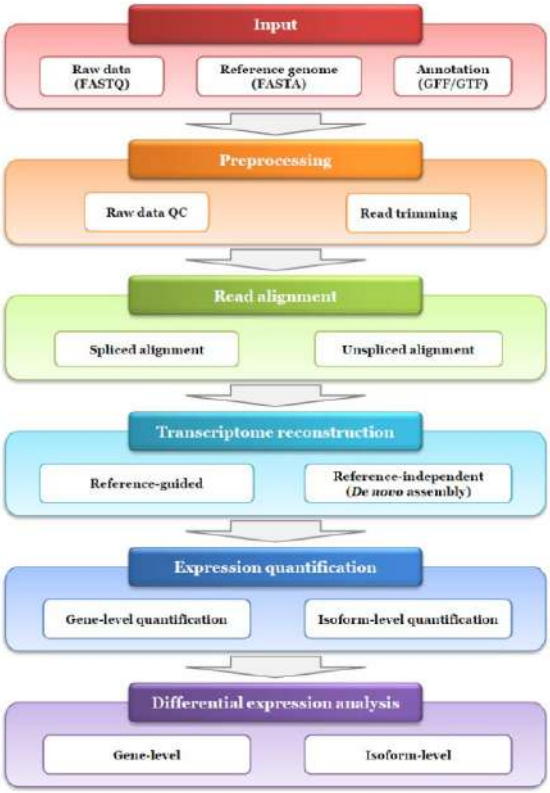


Рисунок 4.2.2. Схема анализа данных, полученных технологией RNA-Seq [80]. Со списком программ для работы с данными можно ознакомиться там же. Красным обозначены исходные данные, референсный геном, аннотация. Оранжевым (обработка полученных сырых ридов) - контроль качества ридов, уаление индексов. Зеленым (выравнивание ридов) - выравнивание после сплайсинга, выравнивание до сплайсинга. Голубым (сборка транскриптома) - выравнивание на референс, выравнивание без референса (сборка de novo). Синим (количественная оценка экспрессии) - количественная анализ на уровне генов, количественный анализа на уровне изоформ. Фиолетовым (дифференциальный анализ экспрессии) - на уровне генна, на уровне изоморфм.

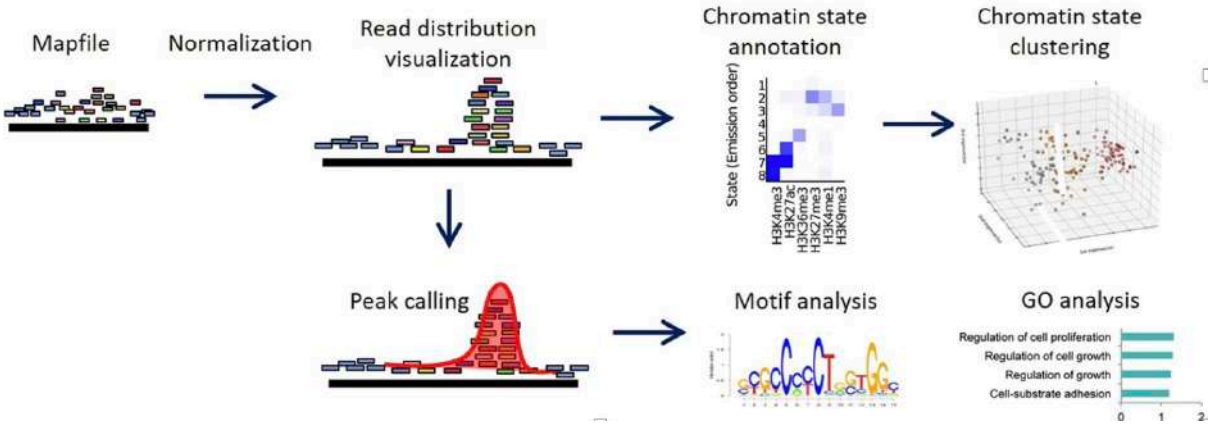


Рисунок 4.2.3. Компьютерный анализ данных ChIP-seq [69]. Программное обеспечение для биоинформатического анализа данных ChIP-seq представлено в [80].

Вопросы:

1) Перечислите методы функционального анализа транскриптома.
2) В чем заключается принцип методов, измеряющих активность регуляторного элемента транскрипции?
3) Какие методы используются для геномного анализа сайтов связывания транскрипционных факторов?
4) Перечислите этапы биоинформатического анализа данных, полученных на микрочипах cDNA и RNA-Seq.
5) Как осуществляется анализ данных ChIP-seq?

4.3. Протеомика

4.3.1. Функциональный анализ протеома

Протеомикой называют науку, изучающую все белки клетки, ткани или организма, то есть осуществляющую систематический крупномасштабный анализ белков. Под протеомом понимают полный набор белков, продуцируемых данной клеткой или организмом при определенных условиях. Белки являются продуктом функционирующих генов и непосредственно участвуют почти в каждом биологическом процессе, поэтому их всесторонний анализ в клетке дает понимание того, как эти молекулы взаимодействуют для создания и поддержания работающей биологической системы. Таким образом, функция каждого

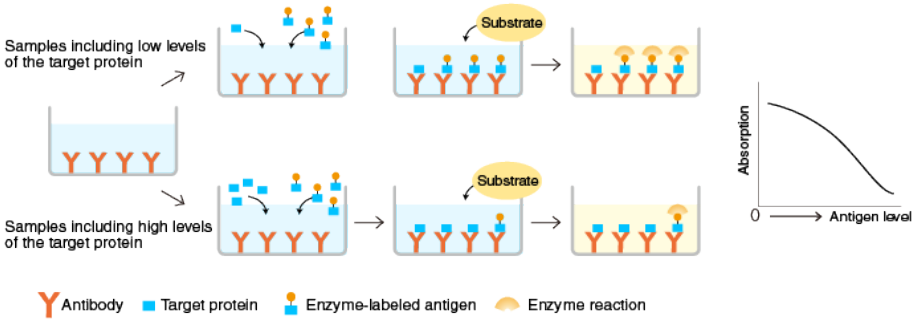
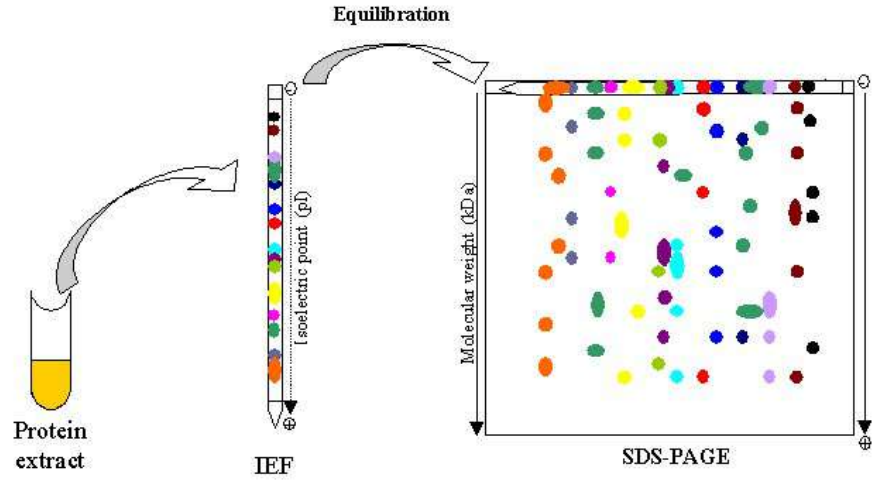
гена определяется тем, какой белок он кодирует и каким образом этот белок влияет на биохимические реакции в клетке и организме. Клетка реагирует на внутренние и внешние изменения, регулируя уровень и активность своих белков, поэтому качественные и количественные изменения в протеоме как системе обеспечивают моментальный снимок этой системы в действии. Белки выполняют широкий спектр функций внутри организмов, а аномальная экспрессия белка вызывает отклонения посттранскрипционной модификации или взаимодействия белка с другим белком или нуклеиновыми кислотами и таким образом может нарушать функцию клетки.

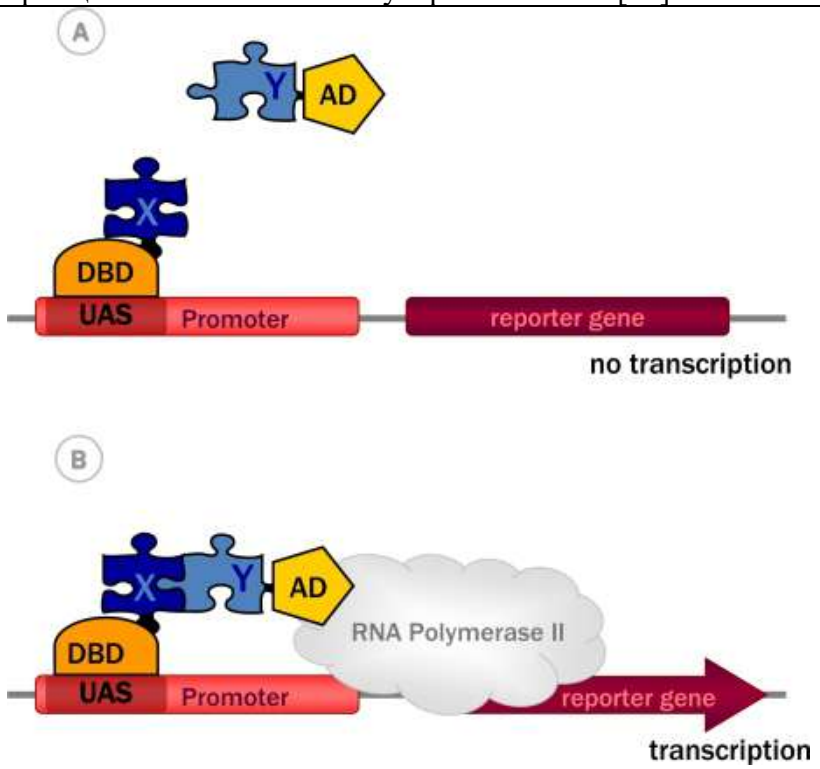
Основные методы протеомики перечислены в таблице 4.3.1.1. Немного остановимся на некоторых из них. В зависимости от цели исследования существуют две стратегии количественного определения белка: *иммуноанализ* (с антителами) и методы определения белков без антител. Иммуноанализ или *иммуноферментный анализ (ИФА)* (англ. ELISA) является широко используемым методом из-за его высокой чувствительности и специфичности. Однако иногда исследователи могут столкнуться с проблемой, если не существует антител к интересующему белку. В таких случаях решением являются методы, которые основаны на детекции белка без антител. Наиболее распространенным аналитическим инструментом для обнаружения, идентификации и количественной оценки белков является *масс-спектрометрия (МС)*, которая основана на зависимости скорости перемещения белков или их частей от массы и заряда, метод измеряет отношение масс-заряда (m/z) ионов. Развитие метода МС дало возможность добиться большей пропускной способности анализа образцов, содержащих тысячи белков и полипетидов в своем составе, с индетификацией каждого из них с высокой точностью. Кроме того, считается, что методология МС является быстрой и надежной для крупных исследований [81,82]. Поэтому МС очень часто сочетается в исследованиях с другими методами.

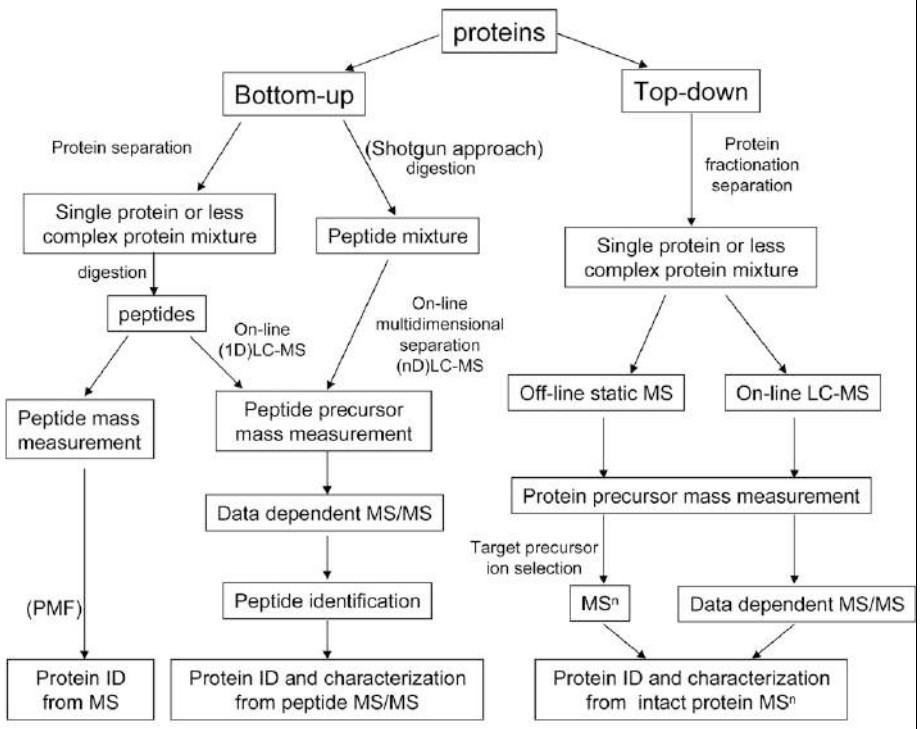
Важным шагом на пути к характеристике функции белка является идентификация путей взаимодействия белков с другими малыми и полимерными молекулами, например, с нуклеиновыми кислотами, ДНК и РНК. При функционально генетических исследованиях наиболее важным представляется выявление взаимодействий, которые происходят между ДНК и факторами транскрипции или регуляторными элементами. В случае РНК очевидна необходимость проверки степени взаимодействия между нуклеиновой кислотой и рибосомой или другими РНК-связывающими белками. Анализ взаимодействия как ДНК, так и РНК с белками основан на аналогичных методах [83,84]. Ранее упомянутая высокопроизводительная иммунопреципитация нуклеиновой кислоты и белкового комплекса все чаще становится предпочтительным методом для обнаружения транскрипционных факторов и модификаций гистонов. Последующий

NGS-анализ или анализ микрочипов позволяют идентифицировать конкретный локус, то есть область, которая специфически взаимодействует с интересующим белком.

Таблица 4.3.1.1. Технологии протеомики

Технология	Принцип
ELISA (enzyme-linked immuno-sorbent assay) = иммуноферментный анализ (ИФА)	 Антитело, специфичное к целевому белку, иммобилизуют на поверхности лунок микропланшета и инкубируют с образцами, содержащими целевой белок и известное количество меченного ферментом целевого белка в качестве контроля. После реакции измеряют активность связанного фермента, находящегося на микропланшете. Если уровень белкового антигена в исследуемом образце высокий, то и количество связанного с антителом антигена, меченного ферментом, ниже и окраска светлее. И наоборот, если количество белка низкое, то уровень связанного с антителом антигена, меченного ферментом, выше, а цвет темнее. График вверху и справа иллюстрирует корреляцию между абсорбцией и уровнями исследуемого антигена в образцах [85].
2-DE	 Двумерный гель-электрофорез (2-DE) - удобный инструмент протеомики. Он используется для разделения и фракционирования сложных белковых смесей из биологических образцов. 2-DE разделяет белки в два этапа: первый называется изоэлектрическое

	фокусирование (IEF), с помощью которого белки делятся в соответствии с изоэлектрическими точками (pI- такой показатель pH, при котором общий заряд белка нейтрален. Соответственно, выше такой точки, то есть при более высоком pH общий заряд белка является отрицательным, а ниже – положительным). Второй этап - электрофорез в полиакриламидном геле в присутствии додецилсульфата натрия (SDS-PAGE), который разделяет белки на основе молекулярных масс (относительная молекулярная масса, Mr). Таким образом, можно разделить тысячи белков и получить информацию об их IEF и молекулярных массах. [86]
Y2H (east two hybrids)	 Двугибридный анализ позволяет детектировать белок-белковые и ДНК-белковые взаимодействия, используя клетки дрожжей. На схеме: (A) Интересующий белок X слит с ДНК-связывающим доменом (DBD), конструкцией, называемой приманкой. Потенциально взаимодействующий белок Y слит с доменом активации (AD) и называется «добыча». Оба конструкта переносятся в клетки дрожжей методом трансфекции. (B) Приманка, то есть слитый белок DBD-X, связывает вышестоящую активаторную последовательность (UAS) промотора. В случае взаимодействия приманки с добычей, то есть слитым белком AD-Y, восстанавливает функциональный фактор транскрипции, что приводит к дальнейшему привлечению РНК-полимеразы II и последующей транскрипции репортерного гена и окраске дрожжевых клеток обычно в синий цвет. Если взаимодействия нет, то клетки остаются бесцветными [87].

MS = Масс-спектрометрия	 Белки, извлеченные из биологических образцов, можно анализировать восходящими или нисходящими методами. При восходящем подходе белки в сложных смесях разделяются и затем расщепляются ферментативным (или химическим) способом с последующим сбором пептидных масс-«отпечатков пальцев» или разделением пептидов в режиме онлайн в сочетании с тандемной масс-спектрометрией. В качестве альтернативы белковая смесь может быть непосредственно преобразована в набор пептидов (подход «дробовика»), которые затем разделяются многомерной хроматографией в режиме онлайн в сочетании с тандемным масс-спектрометрическим анализом. При нисходящем подходе белки в сложных смесях фракционируют и разделяют на чистые отдельные белки или менее сложные белковые смеси с последующей статической инфузией образца в масс-спектрометре для измерения массы неповрежденного белка или его фрагментов [87]. Детектированные по массе, заряду и скорости перемещения полипептиды идентифицируются сравнением с базами данных белков, чтобы определить последовательность, которая лучше всего соответствует шаблону.

4.3.2. Биоинформатический анализ данных протеома

Изменения различных физиологических состоянияй в биологической системе являются очень сложными по своей многофакторной природе. Поэтому их точная количественная оценка является крайне сложной. За последние несколько лет в сочетании с классическими методами

исследования, основанные на масс-спектрометрии, стали незаменимым инструментом для определения точного качественного и количественного содержания белков в сложных смесях. В настоящее время этот метод хорошо зарекомендовал себя, использует стратегии количественного определения как на основе меток, так и без них и является самым популярным методом на сегодняшний день. Методы количественного определения белков основаны на использовании стабильных меченых изотопов, включенных в пептиды, что приводит к появлению ожидаемой разнице масс. Напротив, протеомика без использования меток позволяет определять как относительное, так и абсолютное количество белка, используя интенсивность сигнала и спектральный подсчет пептидов (общее количество спектров, идентифицированных для белка).

В отличие от всего генома, структура протеома может нарушаться как со временем, так и под воздействием условий внутри клетки. Природа таких нарушений сложна и зависит от эпигенетического статуса, посттранскрипционных, посттрансляционных и физиологических процессов. Следовательно, функциональная и количественная характеристика каждого белка на основе его изоформ, посттрансляционных модификаций, субклеточной локализации, тканевой экспрессии и белков-партнеров по взаимодействию является сложной задачей.

Количественную протеомику, основанную на масс-спектрометрии, можно разделить на две широкие категории - методы, основанные на метках, и методы без меток. Методы протеомики на основе меток стали доступны благодаря введению меток ICAT группой Эберсольда [88]. По сути, количественный анализ белков основан на соотношении интенсивности легких/тяжелых пептидов. В методах, использующих метки, образцы сначала дифференциально маркируются, объединяются и подвергаются анализу и количественному определению с помощью масс-спектрометрии. Наиболее широко используемые методы маркировки включают метаболические, протеолитические и химические стратегии. В последнее время значительный интерес к вызывает безметочная количественная протеомика, что связано с появлением масс-спектрометров с высоким разрешением и точностью, а также простотой использования и воспроизводимостью этих методов. В этом случае образцы обрабатываются и анализируются независимо с помощью масс-спектрометрии. Последующая количественная оценка выполняется путем измерения площади пика и/или анализа количества спектров MC/MC (двойной квадруполь) от каждого пептида [89]. Для эффективного количественного анализа применяются различные программы способные обрабатывать большое количество данных, полученных с использованием как методов, основанных на метках, так без них (Таблица 4.3.2.1). Более детально ознакомиться со стратегиями, базами данных для биоинформатического анализа протеома можно в обзоре [89].

Таблица 4.3.2.1. Доступны общие программные пакеты для анализа данных количественной протеомики

На основе меток	SILAC	BioWorks, Census, Mascot Distiller, Elucidator, MaxQuant, MaXIC-Q, OpenMS, PeakQuant, MFPaQ, PEAKS Q, ProteinPilot, ProteoIQ, TPP-ASAPRatio, WARP-LC, MSQuant
	^{15}N, ^{18}O	^{18}O: Mascot Distiller, MSQuant, PEAKS Q, ProteoIQ, QUIL, STEM, VIPER, ZoomQuant, ProRata
		^{15}N: MSQuant, Census, PeakQuant, ProRata, ProteoIQ, Qupe, TPP-XPRESS, X-TRacker
	ICAT, ITRAQ, TMT	ICAT: BioWorks , Elucidator, MaxQuant, MaXIC-Q, MFPaQ, PEAKS Q, ProteinPilot, ProteoIQ, TPP-ASAPRatio, MSQuant, QUIL, TPP-XPRESS, VIPER, ProRata
		ITRAQ: BioWorks, Census, ITracker, OPenMS, PeakQuant, PEAKS Q, Pro Quant, ProteinPilot, Proteios, ProteoIQ, Proteome Discoverer, TPP-Libra, X-Tracker
		TMT: Proteios, ProteoIQ, Proteome Discover, PEAKS Q
«Без меток»	Основанные на интенсивности пика	SpecArray, MSight, PEPPeR, MSInspect, MSQuant, Census, Corra, Serac, SuerHIrn, MzMine, BioWorks, Elucidator, Mascot Distiller, OpenMS, ProteoIQ, SIEVE, Skyline
	Основанные на спектральном счете	SEQUEST, MASCOT, X!Tandem, ProteoIQ, Census, PepC, emPAI Calc, Elucidator, MFPaQ

Вопросы:

1) Какие две основные стратегии анализа протеома существуют?
2) Перечислите методы анализа протеома.
3) Какие методы количественного определения белка используются в настоящее время?

4.4. Модельные системы

Широко развитые технологии высокопроизводительного секвенирования позволяют создать подробный каталог индивидуальных генетических вариантов свойственных одельному индивида или популяции. Однако главный вопрос касается взаимосвязи генотипа и фенотипа. Чтобы получить ответ на этот вопрос, необходимо осуществить функциональное исследование конкретных биологических объектов. В случае людей такие исследования часто не представляются возможными из-за некоторых возникающих биоэтических проблем. Поэтому на помощь приходят экспериментальные исследования на модельных системах, таких как культура клеток in vitro или животные модели, которые позволяют осуществить исследования по функциональной интерпретации интересующих вариантов последовательности генома (Таблица 4.4.1).

«Животные» модели уже давно используются в различных исследованиях для изучения биологических и патогенетических механизмов, а также для разработки эффективных методов лечения различных заболеваний. Чаще всего для исследований по функциональной геномики используются модельные животные: мыши (Mus musculus), плодовые мушки (Drosophila melanogaster) и рыбка Данио (Danio rerio). Упомянутые модельные системы имеют ряд преимуществ. Например, мутация может быть вызвана искусственно, а мутантный фенотип легко распознан, кроме того, возможно осуществить клонирование генов с использованием стандартных процедур, и последнее. Животное производит большое количество потомства за относительно короткий период времени, что позволяет использовать достаточное количество экспериментального материала, необходимое для достоверной биостатистики [90]. Существуют две основные стратегии использования животных моделей: упомянутый в главе 3.1.1 метод генетического *«нокаута»*, то есть подавления интересующего исследователей гена, или включение той же мутации, которая наблюдается у человека. Например, был проведен ряд исследований по созданию животных моделей заболеваний человека с помощью химического мутагенеза (например, с помощью метил нитромочевины; ENU), вызывающего случайные аллельные точечные мутации у мышей. Однако основным ограничением использования животных моделей являются особенности их фенотипа, которые часто не отражает фенотип человека [91].

В настоящий время эффективной технологией для получения создания направленного мутагенеза и редактирования геномов является система CRISPRCas9 (Clustered Regularly Interspaced Short Polindromic Repeats/CRISPR associated), которая представляет собой элементы прокариотической иммунной системы против вирусов. Эта система состоит из небольшого кластера генов cas (кодирующих белки, ассоциированные с

CRISPR) и специфической последовательности ДНК, называемой локусом CRISPR, которая включает короткие повторы, разделенные уникальными спейсерами [92]. При вирусной инфекции его уникальный спейсер интегрируется в бактериальный локус CRISPR. Впоследствии этот локус транскрибируется в предшественник РНК CRISPR. После процессинга зрелая crisprРНК может распознавать и разрушать нуклеиновую кислоту-мишень, взаимодействуя с белками Cas. Таким образом, локус CRISPR содержит информацию о предыдущих вирусных инфекциях, что дает бактериям возможность распознавать и инактивировать вирус в случае повторного заражения. В настоящее время многие научные исследования показывают, что можно использовать систему CRISPR для распознавания и разрезания ДНК в желаемом локусе [93]. Благодаря этим свойствам появилась возможность применения этой системы in vitro в клеточных линиях, в том числе и для изучения генетических заболеваний человека, а также редактирования генома (то есть исправления генетических дефектов) без каких-либо негативных последствий для клеток.

Таблица 4.4.1. Модельные системы

Химический мутагенз	Мутация может быть вызвана искусственно, и мутантный фенотип можно легко распознать. Гены можно клонировать с помощью стандартных процедур. Фенотип не всегда отражает человека
CRISPR-Cas9	Возможность конструировать белковые и РНК-компоненты бактериальной системы CRISPR для распознавания и разрезания и сшивания ДНК в желаемом локусе.

Вопросы:

1) Какие существуют модельные системы для функциональной геномики?
2) Какие существуют способы редактирования генома?

Заключение

Появление «омиксных» технологий, геномики, транкскриптомики, протеомики, эпигеномики, а также высокопроизводительных методов секвенирования следующего поколения (NGS) и биоинформатики значительно ускорили развитие наук о жизни за последние два десятилетия и приблизили так называемую «постгеномную эру». Однако наступление этой эры было бы невозможным без достижений функциональной генетики и геномики. Функциональная генетика - важная отрасль молекулярной биологии, поскольку она изучает не только структуру и устройство генов, но и конечный результат их работы, то есть функцию, которая связана с влиянием конкретного гена на фенотип. Кроме того, функциональная генетика исследует метаболические и регуляторные цепи каждого гена во взаимодействии с другими генами и, тем самым, является неотъемлемой частью системной и сетевой биологии, которые в совокупности с компьютерной биологией (биоинформатикой) могут предсказывать фенотипы и их изменения в результате взаимодействия с окружающей средой за счет переключения регуляторных сигналов и их трансдукции. Например, с помощью функциональной генетики можно предсказывать формулы новых лекарств и эффект их воздействия на организм. Причем, в случае наличия отсеквенированного генома пациента и знания индивидуального состава его аллелей можно выяснить индивидуальный эффект лекарства на организм каждого отдельного пациента. Таким образом, функциональная генетика является базой для создания персонализированной медицины. Некоторые примеры этого рассмотрены нами в параграфе «Геном человека». Также функциональная генетика крайне важна для интенсификации и «экологизации» сельского хозяйства. Имея информацию о взаимодействии метаболических путей ключевых белков и регуляции кодирующих их генов, можно предсказать биохимические свойства и качество урожая, наличие или отсутствие токсичности растительной биомассы, влияние растения на почву и воздействие удобрений на растение. Однако для более точного прогнозирования взаимодействия метаболических путей организма и способах их регуляции необходима полная информация о совокупности работы всех генов генома клетки, ткани и организма во времени и пространстве, то есть на разных стадиях онтогенеза (индивидуального развития организма) и в различных тканях и органах. Работой и регуляцией всех генов клетки занимается функциональная геномика. Она помогает выяснить качественные и количественные и отличия в накоплении транксриптов и белков в различных клетках. Суммируя всю информацию о функции и регуляции каждого отдельного гена, полученную исследованиями функциональной генетики, функциональная геномика фактически позволяет с высокой точностью прогнозировать и ставить

эксперименты in silico, то есть на компьютере, а не в лаборатории, тем самым ускоряя и удешевляя развитие медицины, фармакологии, микробиологии, сельского хозяйства и других биологических дисциплин. Функциональную геномику вместе с биоинформатикой и биотехнологией можно назвать науками будущего, от которых в ближайшее время можно ожидать крупнейшие открытия, которые помогут значительно улучшить качество нашей жизни.

Библиография

1. Holterhoff K. The History and Reception of Charles Darwin's Hypothesis of Pangenesis // Journal of the History of Biology. Kluwer Academic Publishers, 2014. Т. 47, № 4. С. 661–695.
2. Mendel's Paper (English - Annotated) [Электронный ресурс]. URL: http://mendelweb.com/Mendel.html (дата обращения: 26.02.2021).
3. Коряков Д. Е. Ж.И.Ф. Хромосомы. Структура и функции. Новосибирск: Изд-во Сибирского отд-ния Российской акад. наук, 2009. 258 с.
4. Генетика с основами селекции : учебник для студентов высших... [Электронный ресурс]. URL: https://mdk-arbat.ru/book/846721 (дата обращения: 04.04.2021).
5. Codons to Amino Acids - Gen. BIO #2 [Электронный ресурс]. URL: https://sites.google.com/site/bio1040genbio2/chapter-17-from-gene-to-protein/codons-to-amino-acids (дата обращения: 03.04.2021).
6. Бадаева Е.Д., Салина Е.А. Вавиловский журнал генетики и селекции. 2013. Т. 17, № 2.
7. Lerat E., Capy P. Retrotransposons and retroviruses: Analysis of the envelope gene // Mol. Biol. Evol. Society for Molecular Biology and Evolution, 1999. Т. 16, № 9. С. 1198–1207.
8. Brukhin V.B. и др. Female gametophytic mutants of Arabidopsis thaliana identified in a gene trap insertional mutagenesis screen // Int. J. Dev. Biol. Int J Dev Biol, 2011. Т. 55, № 1. С. 73–84.
9. под ред. И. П. Ермакова. Физиология растений. «Академия». М, 2007. 640 с.
10. YULITA K.S. Secondary Structures of Chloroplast trnL Intron in Dipterocarpaceae and its Implication for the Phylogenetic Reconstruction // HAYATI J. Biosci. Institut Pertanian Bogor, 2013. Т. 20, № 1. С. 31–39.
11. Auton A. и др. A global reference for human genetic variation // Nature. Nature Publishing Group, 2015. Т. 526, № 7571. С. 68–74.
12. Green E.D., Watson J.D., Collins F.S. Human Genome Project: Twenty-five years of big biology // Nature. Nature Publishing Group, 2015. Т. 526, № 7571. С. 29–31.
13. Oleksyk T.K., Brukhin V., O'Brien S.J. The Genome Russia project: closing the largest remaining omission on the world Genome map // Gigascience. BioMed Central Ltd., 2015. Т. 4, № 1. С. 53.
14. Zhernakova D. V. и др. Genome-wide sequence analyses of ethnic populations across Russia // Genomics. Academic Press Inc., 2020. Т. 112, № 1. С. 442–458.
15. Brukhin V. и др. The Boechera genus as a resource for apomixis research // Frontiers in Plant Science. Frontiers Media S.A., 2019. Т. 10.
16. Soltis P.S., Soltis D.E. Plant genomes: Markers of evolutionary history and

drivers of evolutionary change // PLANTS, PEOPLE, PLANET. Wiley, 2021. Т. 3, № 1. С. 74–82.
17. Kaul S. и др. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana // Nature. Nature Publishing Group, 2000. Т. 408, № 6814. С. 796–815.
18. Молекулярно-генетическая регуляция апомиксиса, «Генетика» // Генетика. Akademizdatcenter Nauka, 2017. № 9. С. 1001–1024.
19. Namba K., Stubbs G. Structure of tobacco mosaic virus at 3.6 Å resolution: Implications for assembly // Science (80-.). American Association for the Advancement of Science, 1986. Т. 231, № 4744. С. 1401–1406.
20. Genome Size | BioNinja [Электронный ресурс]. URL: https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html (дата обращения: 03.04.2021).
21. Grossniklaus U. и др. Maternal control of embryogenesis by MEDEA, a Polycomb group gene in Arabidopsis // Science (80-.). Science, 1998. Т. 280, № 5362. С. 446–450.
22. Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 1977. Т. 74, № 12. С. 5463–5467.
23. Heather J.M., Chain B. The sequence of sequencers: The history of sequencing DNA // Genomics. Academic Press Inc., 2016. Т. 107, № 1. С. 1–8.
24. Lander E.S. и др. Initial sequencing and analysis of the human genome // Nature. Nature Publishing Group, 2001. Т. 409, № 6822. С. 860–921.
25. Kwong L.N. и др. Co-clinical assessment identifies patterns of BRAF inhibitor resistance in melanoma // J. Clin. Invest. American Society for Clinical Investigation, 2015. Т. 125, № 4. С. 1459–1470.
26. Schuster S.C. Next-generation sequencing transforms today’s biology // Nature Methods. Nat Methods, 2008. Т. 5, № 1. С. 16–18.
27. Yin R., Kwoh C.K., Zheng J. Whole genome sequencing analysis // Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Elsevier, 2018. Т. 1–3. С. 176–183.
28. Nakazato T., Ohta T., Bono H. Experimental Design-Based Functional Mining and Characterization of High-Throughput Sequencing Data in the Sequence Read Archive // PLoS One / под ред. Aziz R.K. Public Library of Science, 2013. Т. 8, № 10. С. e77910.
29. Ihle M.A. и др. Comparison of high resolution melting analysis, pyrosequencing, next generation sequencing and immunohistochemistry to conventional Sanger sequencing for the detection of p.V600E and non-p.V600E BRAF mutations // BMC Cancer. BMC Cancer, 2014. Т. 14, № 1.
30. Morozova O., Marra M.A. Applications of next-generation sequencing technologies in functional genomics. 2008.

31. Shendure J., Ji H. Next-generation DNA sequencing // Nature Biotechnology. Nature Publishing Group, 2008. Т. 26, № 10. С. 1135–1145.
32. Roche - Roche 454 Life Sciences and SoftGenetics Sign Co-Promotion Agreement for Next-Gen Sequencing Software Tools [Электронный ресурс]. URL: https://www.roche.com/media/releases/med-cor-2012-05-09t.htm (дата обращения: 18.03.2021).
33. Instrument Operation Templated Bead Preparation SAGE™ Tag Preparation SOLiD ™ SAGE ™ Guide Applied Biosystems SOLiD ™ 3 System SOLiD ™ SAGE ™ Guide.
34. SPECIFICATION SHEET: ILLUMINA ® SEQUENCING.
35. Ion Torrent | Thermo Fisher Scientific - RU [Электронный ресурс]. URL: https://www.thermofisher.com/ru/ru/home/brands/ion-torrent.html (дата обращения: 18.03.2021).
36. Pushkarev D., Neff N.F., Quake S.R. Single-molecule sequencing of an individual human genome // Nat. Biotechnol. Nature Publishing Group, 2009. Т. 27, № 9. С. 847–850.
37. Mccarthy A. Third generation DNA sequencing: Pacific biosciences' single molecule real time technology // Chemistry and Biology. Elsevier Ltd, 2010. Т. 17, № 7. С. 675–676.
38. Nakano K. и др. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area // Human Cell. Springer Tokyo, 2017. Т. 30, № 3. С. 149–161.
39. Schadt E.E., Turner S., Kasarskis A. A window into third-generation sequencing // Hum. Mol. Genet. Hum Mol Genet, 2010. Т. 19, № R2.
40. Company history [Электронный ресурс]. URL: https://nanoporetech.com/about-us/history (дата обращения: 18.03.2021).
41. Deamer D., Akeson M., Branton D. Three decades of nanopore sequencing // Nature Biotechnology. Nature Publishing Group, 2016. Т. 34, № 5. С. 518–524.
42. Xu L., Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer // Journal of Human Genetics. Springer Nature, 2020. Т. 65, № 1. С. 25–33.
43. The power of single cell partitioning.
44. Laurentino S. и др. High-resolution analysis of germ cells from men with sex chromosomal aneuploidies reveals normal transcriptome but impaired imprinting // Clin. Epigenetics. BioMed Central Ltd., 2019. Т. 11, № 1.
45. Wang X. и др. Genome Assembly of the A-Group Wolbachia in Nasonia oneida Using Linked-Reads Technology // Genome Biol. Evol. Oxford University Press, 2019. Т. 11, № 10. С. 3008–3013.
46. Delaneau O. и др. Accurate, scalable and integrative haplotype estimation // Nat. Commun. Nature Research, 2019. Т. 10, № 1. С. 1–10.
47. Wheeler D.L. и др. Database resources of the National Center for

Biotechnology Information // Nucleic Acids Res. Nucleic Acids Res, 2007. T. 35, № SUPPL. 1.
48. Genetics : analysis & principles (Book, 2009) [WorldCat.org] [Электронный ресурс]. URL: https://www.worldcat.org/title/genetics-analysis-principles/oclc/173243854 (дата обращения: 18.03.2021).
49. Human molecular genetics - NLM Catalog - NCBI [Электронный ресурс]. URL: https://www.ncbi.nlm.nih.gov/nlmcatalog/101523906 (дата обращения: 18.03.2021).
50. Analysing raw sequencing reads with FASTQC for quality control and filtering | by shilparaopradeep | Medium [Электронный ресурс]. URL: https://medium.com/@shilparaopradeep/analysing-raw-sequencing-reads-with-fastqc-for-quality-control-and-filtering-cacaf06b8988 (дата обращения: 18.03.2021).
51. Trivedi U.H. и др. Quality control of next-generation sequencing data without a reference // Front. Genet. Frontiers Research Foundation, 2014. T. 5, № MAY. C. 111.
52. Bioinformatics: Sequence and Genome Analysis, Second Edition [Электронный ресурс]. URL: https://www.cshlpress.com/default.tpl?cart=161607647644015726 8&fromlink=T&linkaction=full&linksortby=oop_title&--eqSKUdatarq=466 (дата обращения: 18.03.2021).
53. Miller J.R., Koren S., Sutton G. Assembly algorithms for next-generation sequencing data // Genomics. Genomics, 2010. T. 95, № 6. C. 315–327.
54. Fonseca N.A. и др. Tools for mapping high-throughput sequencing data // Bioinformatics. Bioinformatics, 2012. T. 28, № 24. C. 3169–3177.
55. Li H., Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform // Bioinformatics. Oxford Academic, 2010. T. 26, № 5. C. 589–595.
56. Langmead B. и др. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biol. BioMed Central, 2009. T. 10, № 3. C. R25.
57. The Variant Call Format (VCF) Version 4.2 Specification. 2021.
58. Abril J.F., Castellano S. Genome annotation // Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Elsevier, 2018. T. 1–3. C. 195–209.
59. Brukhin V., Curtis M.D., Grossniklaus U. The angiosperm female gametophyte: No longer the forgotten generation // SPECIAL SECTION: EMBRYOLOGY OF FLOWERING PLANTS CURRENT SCIENCE. 2005. T. 89, № 11.
60. Brukhin V. и др. The RPN1 subunit of the 26S proteasome in Arabidopsis is essential for embryogenesis // Plant Cell. American Society of Plant Biologists, 2005. T. 17, № 10. C. 2723–2737.
61. Liu Y. -G и др. Efficient isolation and mapping of Arabidopsis thaliana T-

DNA insert junctions by thermal asymmetric interlaced PCR // Plant J. Plant J, 1995. Т. 8, № 3. С. 457–463.
62. Kliver S. и др. Assembly of the boechera retrofracta genome and evolutionary analysis of apomixis-associated genes // Genes (Basel). MDPI AG, 2018. Т. 9, № 4.
63. Dunham I. и др. An integrated encyclopedia of DNA elements in the human genome // Nature. Nature Publishing Group, 2012. Т. 489, № 7414. С. 57–74.
64. Maston G.A., Evans S.K., Green M.R. Transcriptional regulatory elements in the human genome // Annual Review of Genomics and Human Genetics. 2006. Т. 7. С. 29–59.
65. Gross D.S., Garrard W.T. Nuclease Hypersensitive Sites in Chromatin // Annu. Rev. Biochem. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 1988. Т. 57, № 1. С. 159–197.
66. Crawford G.E. и др. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites // Proc. Natl. Acad. Sci. U. S. A. Proc Natl Acad Sci U S A, 2004. Т. 101, № 4. С. 992–997.
67. Aparicio O., Geisberg J. V., Struhl K. Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo // Curr. Protoc. Cell Biol. Wiley, 2004. Т. 23, № 1. С. 17.7.1-17.7.23.
68. Muhammad I.I. и др. RNA-seq and CHIP-seq as complementary approaches for comprehension of plant transcriptional regulatory mechanism // International Journal of Molecular Sciences. MDPI AG, 2020. Т. 21, № 1.
69. Nakato R., Sakata T. Methods for ChIP-seq analysis: A practical workflow and advanced applications // Methods. Academic Press Inc., 2020. Т. 187. С. 44–53.
70. Azizi E. и др. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment // Cell. Cell Press, 2018. Т. 174, № 5. С. 1293-1308.e36.
71. Rotem A. и др. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state // Nat. Biotechnol. Nature Publishing Group, 2015. Т. 33, № 11. С. 1165–1172.
72. Ai S. и др. Profiling chromatin states using single-cell itChIP-seq // Nat. Cell Biol. Nature Publishing Group, 2019. Т. 21, № 9. С. 1164–1172.
73. Ku W.L. и др. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification // Nat. Methods. Nature Publishing Group, 2019. Т. 16, № 4. С. 323–325.
74. Kaya-Okur H.S. и др. CUT&Tag for efficient epigenomic profiling of small samples and single cells // Nat. Commun. Nature Publishing Group, 2019. Т. 10, № 1. С. 1–10.
75. Carter B. и др. Mapping histone modifications in low cell number and

single cells using antibody-guided chromatin tagmentation (ACT-seq) // Nat. Commun. Nature Publishing Group, 2019. Т. 10, № 1. С. 1–5.
76. Wang Q. и др. CoBATCH for High-Throughput Single-Cell Epigenomic Profiling // Mol. Cell. Cell Press, 2019. Т. 76, № 1. С. 206-216.e7.
77. Lowe R. и др. Transcriptomics technologies // PLoS Comput. Biol. Public Library of Science, 2017. Т. 13, № 5. С. e1005457.
78. Shiraki T. и др. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2003. Т. 100, № 26. С. 15776–15781.
79. Selvaraj S., Natarajan J. Microarray Data Analysis and Mining Tools // Bioinformation. Biomedical Informatics Publishing Group, 2011. Т. 6, № 3. С. 95–99.
80. Yang I.S., Kim S. G&I Genomics & Informatics Analysis of Whole Transcriptome Sequencing Data: Workflow and Software // Genomics Inform. 2015. Т. 13, № 4. С. 119–125.
81. Stanczyk F.Z., Clarke N.J. Advantages and challenges of mass spectrometry assays for steroid hormones // Journal of Steroid Biochemistry and Molecular Biology. J Steroid Biochem Mol Biol, 2010. Т. 121, № 3–5. С. 491–495.
82. Yates J.R., Ruse C.I., Nakorchevsky A. Proteomics by mass spectrometry: Approaches, advances, and applications // Annual Review of Biomedical Engineering. Annu Rev Biomed Eng, 2009. Т. 11. С. 49–79.
83. Helwa R., Hoheisel J.D. Analysis of DNA-protein interactions: From nitrocellulose filter binding assays to microarray studies // Analytical and Bioanalytical Chemistry. Anal Bioanal Chem, 2010. Т. 398, № 6. С. 2551–2561.
84. Ascano M., Gerstberger S., Tuschl T. Multi-disciplinary methods to define RNA-protein interactions and regulatory networks // Current Opinion in Genetics and Development. Elsevier Current Trends, 2013. Т. 23, № 1. С. 20–28.
85. The principle and method of ELISA | MBL Life Science -JAPAN- [Электронный ресурс]. URL: https://ruo.mbl.co.jp/bio/e/support/method/elisa.html (дата обращения: 01.04.2021).
86. Two dimensional gel electrophoresis (2-DE) – Creative Proteomics Blog [Электронный ресурс]. URL: https://www.creative-proteomics.com/blog/index.php/two-dimensional-gel-electrophoresis-2-de/ (дата обращения: 26.03.2021).
87. Brückner A. и др. Yeast two-hybrid, a powerful tool for systems biology // International Journal of Molecular Sciences. Multidisciplinary Digital Publishing Institute (MDPI), 2009. Т. 10, № 6. С. 2763–2788.
88. Gygi S.P. и др. Quantitative analysis of complex protein mixtures using

isotope-coded affinity tags // Nat. Biotechnol. Nat Biotechnol, 1999. Т. 17, № 10. С. 994–999.
89. Bantscheff M. и др. Quantitative mass spectrometry in proteomics: A critical review // Anal. Bioanal. Chem. Anal Bioanal Chem, 2007. Т. 389, № 4. С. 1017–1031.
90. Meneely P. Genetic Analysis: Genes, Genomes, and Networks in Eukaryotes, 2nd edition // Fac. Publ. 2014.
91. Claij N., Peters D.J.M. Teaching molecular genetics: Chapter 2-Transgenesis and gene targeting: Mouse models to study gene function and expression // Pediatr. Nephrol. Springer, 2006. Т. 21, № 3. С. 318–323.
92. Rath D. и др. The CRISPR-Cas immune system: Biology, mechanisms and applications // Biochimie. Elsevier B.V., 2015. Т. 117. С. 119–128.
93. Gasiunas G. и др. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2012. Т. 109, № 39. С. E2579–E2586.

Брюхин Владимир Борисович, Андрусенко Елена Владимировна

Функциональная генетика и геномика

Учебно-методическое пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ No
Тираж: 30 экземпляров
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49