Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта»

На правах рукописи

Шаманский Виктор Анатольевич

Влияние нуклеотидных мотивов и структуры митохондриального генома на образование делеций

Специальность 1.5.8. —
«Математическая биология, биоинформатика»

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель
кандидат биологических наук
Гунбин К.В.

Консультант
кандидат биологических наук
Попадьин К.Ю.

Калининград – 2025

Оглавление

Список картинок

Список таблиц

Введение

Старение связано с накоплением повреждений ДНК, накоплением соматических мутаций. Особенно выражен этот процесс в митохондриальных геномах (мтДНК) постмитотических клеток, где накопление масштабных соматических митохондриальных делеций связано как с процессом старения, так и с возрастными митохондриальными энцефаломиопатиями. Митохондриальный геном (мтДНК), существующий внутри клетки в большом количестве копий, сильно предрасположен к накоплению таких возрастных повреждений из-за постоянного обновления [1] и высокой частоты мутаций [2]. Сосуществование разных вариантов мтДНК внутри одной клетки (гетероплазмия) [3] обеспечивает внутриклеточную конкуренцию мтДНК, которая особенно влиятельна в медленно делящихся тканях, где «эгоистичные» мутанты мтДНК с преимуществом репликации, но функциональными недостатками имеют время для клональной экспансии [4]. Одним из наиболее изученных примеров эгоистичных мутаций мтДНК являются делеции — удаление части митохондриального генома. Например, в нейронах черной субстанции первые делеции мтДНК были обнаружены примерно в 50-летнем возрасте [4, 5]. Ежегодно эта фракция гетероплазмии увеличивалась на 1–2%, пока через несколько десятилетий не был достигнут фенотипически существенный порог в 50–80% [4], приводящий к нейродегенерации. Скелетные мышцы — еще одна ткань, предрасположенная к накоплению соматических делеций мтДНК: увеличение соматических делеций мтДНК внутри миофибрилл связано с саркопенией — потерей мышечной массы и силы с возрастом [6, 7]. Другие ткани с медленно делящимися клетками, на которые также влияют делеции мтДНК, включают экстраокулярные мышцы [8] и ооциты [9, 10, 11]. В случае ооцитов распространение делеций мтДНК потенциально может проявляться во всех тканях, включая пролиферативные, приводя к мультисистемным нарушениям [12, 13]. Замедление скорости возникновения соматических делеций может способствовать увеличению продолжительности жизни человека и здоровому старению.

Существует несколько доказательств, подтверждающих гипотезу о том, что соматические делеции мтДНК вызывают дегенерацию клеток-хозяев и несколько соответствующих возрастных фенотипов. (i) Версия мтДНК-полимеразы с дефицитом коррекции приводит к накоплению соматических точковых мутаций и делеций у мышей, за которыми следует сокращение продолжительности жизни и преждевременное появление специфичных для старения фенотипов [14, 15]. Однако уровень точечных соматических мутаций у нормальных мышей довольно низок, что ставит под сомнение роль мутаций мтДНК в нормальном старении [16, 17]. (ii) Наблюдение локализации делеций мтДНК в областях разрыва мышечных волокон [18] и нейронах с дефицитом дыхательной цепи [4] подтверждает гипотезу о причинном влиянии делеций мтДНК на старение. (iii) Сообщаемый дефицит нейронов, несущих чрезвычайно высокую (> 80%) нагрузку делеций, позволяет предположить, что такие клетки деградируют и больше не присутствуют в анализируемой ткани [4]. В целом, высокая доля делеций мтДНК не является нейтральным признаком стареющих клеток, а, скорее, является причиной их появления (causative agent). Таким образом, понимание молекулярных механизмов, лежащих в основе происхождения соматических делеций мтДНК, а также скорости их распространения, имеет первостепенное значение [19, 20].

Было показано, что большинство соматических делеций мтДНК фланкированы прямыми нуклеотидными повторами [21] или длинными несовершенными дуплексами, состоящими из коротких участков прямых повторов [22] и было выдвинуто несколько гипотез, объясняющих образование делеции с участием прямого повтора. Поскольку прямые повторы предрасполагают мтДНК к соматическим делециям, являются факторами определяющими точки разрыва, они считаются примером аллелей «вредных в позднем возрасте» (DILL): нейтральных или слегка вредных в репродуктивном возрасте, но вредных в позднем возрасте [23]. Соответственно, уменьшение количества этих аллелей DILL может привести к снижению

образования соматических делеций и увеличению продолжительности жизни. Отрицательная корреляция между количеством прямых повторов в мтДНК и видоспецифичной продолжительностью жизни млекопитающих [24, 25] интерпретируется как дополнительное свидетельство вредного воздействия повторов в мтДНК долгоживущих млекопитающих. Учитывая дефицит прямых повторов в мтДНК долгоживущих млекопитающих, мои коллеги ранее предположили, что уменьшение количества прямых повторов в митохондриальном геноме некоторых гаплогрупп человека может быть связано с меньшей распространенностью соматических делеций мтДНК, тем самым обеспечивая более здоровое старение и отсрочку процесса старения [26].

У большинства людей также обнаружен общий прямой повтор, который, по-видимому, вызывает большинство делеций в мтДНК человека [24]. Но существуют гаплогруппы японских долгожителей [27, 28] у которых этот повтор нарушен. В ходе одного из исследований мои коллеги предположили, что нарушение прямого повтора не приводит к возникновению большинства делеций, что в конечном итоге может приводить к увеличению продолжительности жизни.

Актуальность темы исследования

Митохондрии играют важную роль в метаболизме клетки, а также участвуют в регуляции клеточного цикла и запрограммированной клеточной гибели. Митохондриальная ДНК (мтДНК) является ключевым объектом для изучения эволюции, популяционной генетики и молекулярных механизмов наследственных заболеваний человека. Митохондриальная ДНК (мтДНК) во многом является уникальной, а также она содержит в своей структуре множество повторяющихся элементов и мотивов, которые могут приводить к мутациям и делециям.

Одна из ключевых тем моей научной работы это повторы в митохондриальном геноме, так как изучение влияния нуклеотидных мотивов и структуры митохондриального генома на образование делеций имеет важное значение для понимания механизмов возникновения многих заболеваний, связанных с нарушениями в функционировании митохондрий, а несовершенные повторы, играющие потенциально ключевую роль в мутагенезе и стабильности генома, до сих пор остаются одним из наименее изученных аспектов организации и эволюции мтДНК. Накопление больших объемов информации о связи повторяющихся последовательностей (повторов) митохондриальной ДНК как с продолжительностью жизни различных видов позвоночных животных, так и со старением человека ставит вопрос комплексного сравнительного исследования этих последовательностей у человека и животных. Универсальный и, в тоже время биологически обоснованный, алгоритмический метод к идентификации повторов в мтДНК в сочетании со сравнительным межвидовым анализом их природы, численности, а главное взаимодействий является основным подходом к анализу системной организации и эволюции повторов в мтДНК.

Актуальность данного исследования обусловлена наличием фундаментальных пробелов в данной области: (1) методологического - отсутствие специализированного биоинформатического инструментария, адаптированного для комплексного анализа всех типов несовершенных повторов (прямых, инвертированных, зеркальных, комплементарных) в кольцевой мтДНК, ограничивает масштабные сравнительные исследования; (2) эмпирического - не было создано единого репрезентативного ресурса, содержащего аннотированные повторы для широкого спектра видов, что затрудняет выявление универсальных закономерностей их организации; (3) теоретического - мутагенные механизмы, опосредованные повторами, в частности образование делеций, ассоциированных с возраст-зависимыми патологиями и нейродегенерацией, остаются плохо изученными, не было единой модели, объясняющей роль пространственной структуры мтДНК в этом процессе; (4) эволюционного - вклад мутационного давления и отбора в формирование мутационного спектра мтДНК, который напрямую формирует повторенные участки мтДНК, и его связь с такими фенотипическими признаками, как продолжительность жизни, являются предметом дискуссий и требуют количественной оценки.

Таким образом, разработка нового методологического подхода, создание масштабной базы данных и комплексный анализ повторов мтДНК с привлечением данных по тысячам видов являются крайне актуальными задачами, решение которых необходимо для прогресса в понимании эволюции, мутагенеза и функциональной роли митохондриального генома. Моя диссертация представляет собой важную работу в области исследования митохондриальной генетики потому что имеет широкий спектр теоретических и экспериментальных исследований, направленных на понимание роли нуклеотидных мотивов и структуры митохондриального генома в образовании делеций и может положительно влиять на дальнейшее развитие биологии и медицины. Также мое исследование может дать возможность разработки высокоэффективных и индивидуальных методов диагностики и лечения заболеваний, связанных с мутациями мтДНК. Таким образом можно заключить что тема моего исследования имеет высокую актуальность так как с пониманием механизма образования делеций начнут выявляться возможности для противодействия этому процессу.

Степень разработанности

Существует достаточное количество исследований, посвященных влиянию нуклеотидных мотивов и структуры митохондриального генома на образование делеций и продолжительность жизни. Однако существует множество вопросов которые до сих пор остаются неотвеченными: Играют ли повторы главную роль в образовании делеций? Существует ли отбор против повторов? Как влияет нуклеотидный состав повторов на их силу в образовании делеций? Каким образом повторы могут взаимодействовать друг с другом? Какие структуры имеет мтДНК и какие конформации принимает во время репликации в контексте образования делеций? Поэтому, изучение данной темы еще далеко от полного понимания механизма образования делеций, поэтому дальнейшие исследования в данной области могут иметь большую значимость для биологии и медицины.

Расширения моей работы были размещены в системе предварительного ознакомления bioRxiv:

- Shamanskiy et al., Mitochondrial direct repeat reduction as a strategy for enhancing human longevity: the case of the common repeat, bioRxiv, doi: https://doi.org/10.1101/2024.09.02.610808
- Mikhailova A.G. et al., Mammalian mitochondrial mutational spectrum as a hallmark of cellular and organismal aging, doi: https://doi.org/10.1101/589168
- Mikhailova A.G. et al., A mitochondrial mutational signature of temperature in ectothermic and endothermic vertebrates, doi: https://doi.org/10.1101/2020.07.25.221184;
- Mikhailova A.A. et al., Deleterious in late life mitochondrial alleles and aging: secrets of Japanese centenarians, doi: https://doi.org/10.1101/603282

Цели и задачи

Объект исследования: митохондриальный геном в целом и митохондриальный геном человека, млекопитающих и других позвоночных в частности.

Предмет исследования: влияние нуклеотидных мотивов и структуры митохондриального генома на образование делеций.

Цель исследования: разработка и применение инновационного биоинформатического инструментария для комплексного анализа нуклеотидных мотивов (таких как несовершенные повторы) и структуры митохондриального генома с целью выявления универсальных закономерностей их организации, установления их роли в мутагенезе (образовании делеций) и определения связи их характеристик с эволюционными и фенотипическими признаками (такими как продолжительность жизни).

Задачи исследования:

1. Разработать и верифицировать оригинальный алгоритм для детекции и классификации всех типов несовершенных повторов (прямых, инвертированных, зеркальных, комплементарных) в кольцевых геномах, адаптированный к специфике мтДНК.

2. Создать и развернуть общедоступную базу данных повторов мтДНК для репрезентативной выборки позвоночных (>3500 видов), оснащенную инструментами для визуализации, сравнительного анализа и экспорта данных.
3. Провести масштабный сравнительный анализ и выявить универсальные закономерности в организации повторяющихся элементов мтДНК (длина, нуклеотидный состав, распределение, ассоциация с функциональными областями).
4. Экспериментально (in silico) подтвердить ключевую роль пространственной структуры одноцепочечной мтДНК в обеспечении пространственной близости участков разрыва и образовании делеций; предложить структурную модель, объясняющую локализацию "горячих точек" делеций.
5. Исследовать мутагенный потенциал различных комбинаций повторов (в частности, паттернов DI...ID) и других нуклеотидных мотивов и оценить их вклад в возникновение соматических делеций, ассоциированных с возраст-зависимыми заболеваниями и выявить места наибольшей склонности к образованию делеций.
6. Установить корреляцию между характеристиками повторов и продолжительностью жизни на межвидовом уровне (млекопитающие) и на уровне человеческих гаплогрупп, проверив гипотезу об очищающем отборе.
7. Определить относительный вклад мутационного давления и отбора в формирование мутационного спектра мтДНК (на примере замен A>G), оценив влияние времени нахождения в одноцепочечном состоянии и окислительного повреждения.

Основной вопрос, который я ставлю перед собой в своем исследовании следующий: как влияет структура митохондриального генома на процесс образования делеций? Для ответа на данный вопрос необходимо не только понять какие структурные особенности могут влиять, но и как они взаимодействуют друг с другом. Для этого я выделяю для рассмотрения следующие подвопросы:

1) какие потенциальные факторы образования делеций имеют место и как эти факторы взаимодействуют друг с другом?
2) есть ли области мтДНК где образование делеций предопределено ее структурой?
3) действительно ли прямые повторы являются основным фактором образования делеций?
4) как образуются и исчезают повторы и имеет ли место отбор?
5) возможно ли найти мтДНК которая имеет наименьшую вероятность образования делеций и может ли она быть использована в медицинских целях?

Научная новизна

Результаты исследования показали, что нуклеотидные мотивы и структура митохондриального генома действительно оказывают влияние на образование делеций в митохондриальном геноме. Была выявлена связь между структурами митохондриальной ДНК и уровнем образования делеций.

С фундаментальной точки зрения, реализация данного проекта поможет приблизиться к пониманию механизма образования делеций. С прикладной точки зрения, результаты моей работы помогут более точно оценивать риски сложных заболеваний и выделять когорты людей с повышенным риском. Предложенные методы будут иметь большую значимость для медицины, т.к. они могут быть применены для создания системы предсказания заболеваний, вызываемых митохондриальными делециями.

Таким образом, исследование имеет высокую научную новизну и достоверность, и его результаты могут быть использованы в дальнейших исследованиях в области генетики и медицины.

Теоретическая и практическая значимость работы

Теоретическое значение:

Нуклеотидные мотивы и структура митохондриального генома могут влиять на частоту и местоположение делеций. Моя работа является хорошим шагом на пути к тому чтобы выяснить

полный механизм образования возрастных делеций в мтДНК человека и других живых существ, понять их влияние на функционирование стареющего организма, а также получить понимание того как меняется набор повторов и других структурных особенностей мтДНК в ходе эволюции.

Прикладное значение:

Результаты моей диссертации имеют большое практическое значение в медицине. Например, обнаружение конкретных нуклеотидных мотивов, которые предрасполагают к делециям в митохондриальном геноме, может привести к созданию новых диагностических тестов, которые позволят определять риск развития митохондриальных заболеваний. Кроме того, понимание механизмов образования делеций может помочь предсказать риски определенных заболеваний, связанных с мутациями мтДНК, что может привести к улучшению методов лечения и профилактики, например, путем корректировки нуклеотидных мотивов, что может привести к уменьшению частоты делеций в митохондриальном геноме.

Степень достоверности и апробация результатов

Исследования которые были проведены в рамках моей диссертации "Влияние нуклеотидных мотивов и структуры митохондриального генома на образование делеций" прошли следующие этапы апробации:

1. Публикация статей в научных журналах. При написании нескольких из них в анализах был использован разработанный мной алгоритм поиска повторов.

- Shamanskiy V, Mikhailova AA, Tretiakov EO, Ushakova K, Mikhailova AG, Oreshkov S, Knorre DA, Ree N, Overdevest JB, Lukowski SW, Gostimskaya I, Yurov V, Liou CW, Lin TK, Kunz WS, Reymond A, Mazunin I, Bazykin GA, Fellay J, Tanaka M, Khrapko K, Gunbin K, Popadin K. Secondary structure of the human mitochondrial genome affects formation of deletions. BMC Biol. 2023 May 8;21(1):103. doi: 10.1186/s12915-023-01606-1. PMID: 37158879; PMCID: PMC10166460;
- Mikhailova AG, Mikhailova AA, Ushakova K, Tretiakov EO, Iliushchenko D, Shamansky V, Lobanova V, Kozenkov I, Efimenko B, Yurchenko AA, Kozenkova E, Zdobnov EM, Makeev V, Yurov V, Tanaka M, Gostimskaya I, Fleischmann Z, Annis S, Franco M, Wasko K, Denisov S, Kunz WS, Knorre D, Mazunin I, Nikolaev S, Fellay J, Reymond A, Khrapko K, Gunbin K, Popadin K. A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand. Nucleic Acids Res. 2022 Oct 14;50(18):10264-10277. doi: 10.1093/nar/gkac779. PMID: 36130228; PMCID: PMC9561281;
- Shamanskiy VA, Timonina VN, Popadin KY, Gunbin KV. ImtRDB: a database and software for mitochondrial imperfect interspersed repeats annotation. BMC Genomics. 2019 May 8;20(Suppl 3):295. doi: 10.1186/s12864-019-5536-1. Erratum in: BMC Genomics. 2019 Jul 8;20(1):556. doi: 10.1186/s12864-019-5950-4.. Shamanskiy VN [corrected to Shamanskiy VA]. PMID: 31284879; PMCID: PMC6614062.

2. На разработанный алгоритм и базу данных получены авторские свидетельства

- "Программа идентификации контекстных характеристик вырожденных повторов в митохондриальных геномах Позвоночных", свидетельство о государственной регистрации программы для ЭВМ №198619009, выдано Федеральной службой интеллектуальной собственности;
- «База данных вырожденных повторов в митохондриальных геномах Позвоночных», свидетельство о государственной регистрации базы данных №2016021803, выдано Федеральной службой интеллектуальной собственности.

3. По теме исследования сделаны устные и постерные доклады на научных конференциях:

- "LIFE SCIENCES TODAY 2025", Астана (Казахстан), постерный доклад: "Towards a quantitative assessment of the mitochondrial component of aging: the fragility score as an indicator of the risk of mtDNA deletion formation";

- Moscow Conference on Computational Molecular Biology (MCCMB-2025), Москва, постерный доклад: "Сравнение вероятности делеций мтДНК в регионах, фланкированных различными комбинациями прямых и инвертированных повторов";
- “X International Conference of Young Scientists: Biophysicists, Biotechnologists, Molecular Biologists and Virologists (OpenBio-2023), Новосибирск, онлайн, устный доклад: «На пути к количественной оценке митохондриальной компоненты старения: оценка хрупкости как фактора риска образования делеции мтДНК»;
- Moscow Conference on Computational Molecular Biology (MCCMB-2023), Москва, устный доклад: "Влияние нарушенного общего повтора на здоровое старение на примере гаплогрупп японских долгожителей";
- The Society for Molecular Biology & Evolution conference (SMBE-2023), онлайн, постерный доклад: "Towards quantification of the mitochondrial component of aging: fragility score as a risk of mtDNA deletion formation";
- Форум молодых исследователей ХимБиоSeasons 2023, Калининград, устный доклад: "Влияние нуклеотидных мотивов и структуры митохондриального генома на образование делеций";
- The European Human Genetics Conference (ESGH) 2021, онлайн, постерный доклад: "Risk of mitochondrial deletions is affected by the global secondary structure of the mitochondrial genome";
- Moscow Conference on Computational Molecular Biology (MCCMB) 2021, Москва, постерный доклад: "Somatic deletions in the human mitochondrial genome: the global secondary structure, G-quadruplexes and direct nucleotide repeats explain majority of breakpoints";
- School of Bioinformatics - 2021 (Institute of Bioinformatics), Санкт-Петербург, постерный доклад: «Risk of mitochondrial deletions is affected by the global secondary structure of the mitochondrial genome»;
- The Society for Molecular Biology & Evolution conference - 2021 (SMBE-2021), онлайн, постерный доклад: "Somatic deletions in the human mitochondrial genome: the global secondary structure, G-quadruplexes and direct nucleotide repeats explain majority of breakpoints";
- XXVII International Conference of Students, Postgraduates and Young Scientists "Lomonosov" (Lomonosov Moscow State University, Moscow), устный доклад: «Risk of somatic mitochondrial deletions is affected by the secondary structure of the mitochondrial genome»;
- “VII International Conference of Young Scientists: Biophysicists, Biotechnologists, Molecular Biologists and Virologists (OpenBio-2020), Новосибирск, постерный доклад: «Risk of somatic mitochondrial deletions is affected by the secondary structure of the mitochondrial genome»;
- The Society for Molecular Biology & Evolution conference - 2019 (SMBE-2019), Манчестер (Великобритания), постерный доклад: "Seeking for mtDNA structural determinants of organisms longevity";
- Systems Biology and Bioinformatics – 2018 (SBB-2018), Новосибирск, устный доклад: "ImtRDB: a database and software for mitochondrial imperfect interspersed repeats annotation";
- Bioinformatics of Genome Regulation and Structure\Systems Biology – 2018 (BGRS\SB-2018), Новосибирск, устный доклад: "ImtRDB: a database and software for mitochondrial imperfect interspersed repeats annotation".

Глава 1. Обзор литературы

1.1 Биология старения связана с делециями

Старение характеризуется накоплением повреждений ДНК, причем митохондриальный геном (мтДНК) особенно уязвим из-за его постоянного обновления (turnover). Это обновление, в ходе которого возникают ошибки репликации и окислительные повреждения, напрямую связывает скорость оборота мтДНК с частотой появления новых мутаций, особенно в постмитотических клетках, таких как клетки сердца. В отличие от экспериментальных данных, показывающих сильно различающиеся скорости обновления различных митохондриальных компонентов (от нескольких дней до года), авторы [1] с помощью стохастической модели на основе химического мастер-уравнения (CME) демонстрируют, что реальная скорость обновления мтДНК значительно ниже — с периодом полураспада в несколько месяцев. Моделирование подтверждает, что именно такая скорость соответствует наблюдаемым уровням мутаций, в то время как более высокая скорость обновления приводит к увеличению мутационной нагрузки, потенциально вызывая структурные изменения мтДНК. Таким образом, исследование подчеркивает, что скорость обновления мтДНК является критическим фактором, определяющим накопление мутаций в стареющих тканях.

Митохондриальная гетероплазмия — наличие в клетке нескольких вариантов мтДНК — широко распространена у человека и определяет проявление связанных с мтДНК заболеваний. Однако количественно прогнозировать её наследование и динамику в организме до сих пор было сложно из-за отсутствия адекватной модели. В исследовании [3] для решения этой проблемы предложена популяционно-генетическая модель, рассматривающая гетероплазмию в контексте онтогенетической филогении, где генетический дрейф и мутации изменяют частоты аллелей на разных стадиях развития. С помощью байесовского вывода на основе экспериментальных данных авторы выявили два ключевых процесса: сильное «бутылочное горлышко» (резкое сокращение эффективного размера популяции) при материнской передаче мтДНК и то, что основной генетический дрейф, формирующий вариативность тканей, происходит на ранних, а не на поздних стадиях развития.

Авторы другой статьи тоже показали что ключевым фактором наследования мтДНК является сильное сокращение эффективной численности ее копий при передаче от матери к ребёнку, известное как «бутылочное горлышко» зародышевой линии, что затрудняет прогнозирование. Исследование [2] 39 пар мать-ребёнок показало, что это «бутылочное горлышко» оценивается всего в 30-35 митохондрий, что объясняет частые резкие сдвиги в частоте гетероплазмии между поколениями. Эти сдвиги могут превратить низкую, доброкачественную частоту варианта у матери в высокую, вызывающую заболевание у ребёнка. Кроме того, частота мутаций мтДНК была оценена как высокая (1.3×10^{-8} на сайт в год), превосходящая ядерную ДНК, и обнаружена положительная связь между количеством гетероплазмий у ребёнка и возрастом матери, что связывают со старением ооцитов.

Несмотря на первоначальный энтузиазм, вызванный фенотипом преждевременного старения у мышей-мутаторов мтДНК (несущих генетические дефекты корректирующей экзонуклеазной активности митохондриальной ДНК-полимеразы), дальнейшие исследования [16, 17] поставили под сомнение роль точечных мутаций мтДНК в нормальном старении. Было обнаружено [16], что уровни этих мутаций у старых нормальных мышей остаются крайне низкими — на порядок ниже, чем у мутаторов, и даже ниже, чем у пожилых людей. Это поразительное несоответствие, выявленное с помощью более точных методов, таких как случайный захват мутаций (RMC), указывает на отсутствие прямой причинно-следственной связи между точечными соматическими мутациями мтДНК и нормальным процессом старения.

Несмотря на продолжающиеся дебаты о причинной роли митохондриальных мутаций в

старении, применение новых высокочувствительных методов (таких как RMC) в другом исследовании [17] показало, что реальная частота точечных мутаций мтДНК у мышей дикого типа более чем в 10 раз ниже, чем считалось ранее, а гетерозиготные мыши не демонстрируют статистически значимого удлинения срока жизни при том имеют в сотни раз большую частоту мутаций (примерно в 220 раз выше в сердце и в 500 раз в мозге) по сравнению с контрольными мышами. Хотя с возрастом наблюдается 11-кратное увеличение количества мутаций, ключевым аргументом против их ведущей роли является тот факт, что мыши-мутаторы способны переносить экстремально высокую — в 500 раз выше нормы — мутационную нагрузку без явных признаков ускоренного старения или сокращения продолжительности жизни. Эти результаты убедительно свидетельствуют, что накопление митохондриальных точечных мутаций с возрастом не является лимитирующим фактором продолжительности жизни мышей дикого типа в условиях нормального старения.

Одним из наиболее изученных примеров эгоистичных мутаций мтДНК являются делеции – удаление части митохондриального генома. Например, в нейронах черной субстанции первые делеции в мтДНК были обнаружены примерно в 50-летнем возрасте [4, 5]. Авторы одной из статей [4] выяснили что в процессе старения в медленно обновляющихся тканях, таких как нейроны черной субстанции, происходит клональная экспансия «эгоистичных» делеций мтДНК, которые имеют преимущество в репликации, но нарушают функцию дыхательной цепи. Исследования показывают, что доля гетероплазмии с делециями в нейронах прогрессивно увеличивается с возрастом на 1-2% в год, и по достижении порога в 50-80% это приводит к дефициту цитохром-с-оксидазы (COX) и ускоряет нейродегенерацию. Высокие уровни делеций, вплоть до >80%, обнаруживаются в пигментированных нейронах пожилых людей, что связывают с функциональными нарушениями, гибелью клеток и появлением признаков, сходных с болезнью Паркинсона. Молекулы мтДНК с делециями внутри каждого нейрона в основном клональны, то есть они происходят от единственной молекулы мтДНК с делецией, которая подверглась клональной экспансии.

Авторы другого исследования [5] показали, что в нейронах черной субстанции у пожилых людей и пациентов с болезнью Паркинсона наблюдается высокий уровень делеций митохондриальной ДНК (52.3%±9.3% против 43.3%±9.3% в контрольной группе). Эти повреждения являются соматическими - то есть возникают в течение жизни - и представлены различными клонально расширенными делециями в отдельных нейронах. Это означает, что в каждой клетке происходит размножение (клональная экспансия) одной уникальной делетированной молекулы мтДНК, что в конечном итоге приводит к дефициту COX и нарушению работы дыхательной цепи. Нейроны черной субстанции особенно уязвимы для накопления таких мутаций из-за их высокой метаболической активности, сопряженной с генерацией активных форм кислорода, которые повреждают мтДНК. Таким образом, накопление соматических делеций мтДНК играет важную роль в избирательной гибели нейронов, наблюдаемой как при нормальном старении мозга, так и при болезни Паркинсона.

Скелетные мышцы, наряду с нейронами, являются тканью, особенно предрасположенной к накоплению соматических делеций мтДНК с возрастом. Одно из исследований [6] показывает, что в отдельных мышечных волокнах пожилых особей происходят сегментарные, клональные внутриклеточные экспансии уникальных делеций. Когда доля таких мутантных молекул мтДНК превышает пороговый уровень (~90%), это приводит к катастрофическим последствиям для клетки: потере активности COX, нарушению ферментативного баланса и появлению аномальной морфологии — атрофии, расщеплению и разрыву волокон. Таким образом, накопление делеций мтДНК может являться не просто маркером старения, а прямой причиной дисфункции и разрушения мышечных волокон, что вносит ключевой вклад в развитие саркопении — возрастной потери мышечной массы и силы.

Потеря мышечной массы (миопатия скелетных мышц) и силы с возрастом также была показана в другой статье [7]. Авторы демонстрируют, что накопление делеционных мутаций мтДНК является значимым фактором возрастной потери мышечной массы и силы. В стареющих скелетных мышцах делеции мтДНК клонально накапливаются в отдельных

волокнах, и при достижении высокой концентрации они нарушают клеточное дыхание, что связано с активацией апоптоза, атрофией, разрывом и некрозом волокон. Для проверки этой причинно-следственной связи авторы фармакологически индуцировали накопление делеций у крыс, что привело к резкому (на 1200%) увеличению числа волокон с дефицитом дыхательной цепи, сокращению количества волокон на 18% и потере мышечной массы на 22%. Эти данные подтверждают, что делеционные мутации мтДНК играют ключевую причинную роль в этиологии необратимой потери мышечных волокон при старении.

Создание сверхчувствительного метода LostArc [30] позволило обнаружить в скелетных мышцах человека около 35 миллионов делеций мтДНК (~470 000 уникальных участков), что количественно достаточно для объяснения возрастных изменений и заболеваний, связанных с мутациями Polg. Биоинформатический анализ выявил характерные паттерны делеций, коррелирующие с возрастом и патологией, которые указывают на репликацию с участием ДНК-полимеразы γ как основной драйвер образования делеций. Было установлено, что возраст-ассоциированные делеции характеризуются четырьмя ключевыми особенностями: они расположены в пределах большой дуги мтДНК, демонстрируют высокую степень микрогомологии в точках разрыва, имеют специфическое распределение относительно начала репликации и преимущественно встречаются у пожилых людей. Наблюдаемая картина лучше всего объясняется моделью остановки репликативной вилки со смещением цепи во время синтеза мтДНК, причем экзонуклеазная функция ДНК-полимеразы γ играет критическую роль в предотвращении их образования, а делеции практически не элиминируются митофагией в постмитотических мышечных волокнах. Эти данные подтверждают, что именно делеции мтДНК, а не точечные мутации, играют ключевую роль в старении и развитии митохондриальных заболеваний, непосредственно связывая эти процессы с ошибками репликации митохондриальной ДНК.

Экстраокулярные мышцы это также ткань с медленно делящимися клетками, на которые также влияют делеции соматической мтДНК накапливаясь до высоких уровней при старении и вызывая слабость экстраокулярных мышц [8]. Исследование экстраокулярных мышц (EOM) выявило ускоренный по сравнению со скелетной мускулатурой процесс старения, связанный с накоплением делеций мтДНК. COX -негативные волокна, указывающие на дефицит COX, появляются в EOM уже в третьем десятилетии жизни, и их доля значительно возрастает с возрастом, достигая 3.34% после 60 лет. Большинство этих волокон (около 72%) содержат высокие уровни делеций мтДНК (>70%), в то время как вклад точечных мутаций незначителен. Это демонстрирует, что именно делеции соматической мтДНК являются основной причиной дыхательной недостаточности и возрастной слабости экстраокулярных мышц. Длина делеций варьировалась от 3 до 7 кб, некоторые из них содержали прямые повторы (класс I), а другие не содержали фланкирующих повторов (класс III).

Ооциты так же являются медленно делящимися и подверженны делециям [9, 10, 11]. Исследование генетически идентичных братьев-близнецов [9], у одного из которых развилась хроническая прогрессирующая наружная офтальмоплегия (CPEO), а другой оставался здоровым, показало, что у обоих присутствовала идентичная крупная делеция мтДНК (около 4.1 килобаз, от позиции 11262 до 15375). Ключевое различие заключалось в уровне гетероплазмии и тканевой специфичности: у больного близнеца делеция с высоким уровнем гетероплазмии была обнаружена в мышцах, в то время как у бессимптомного брата она присутствовала в минимальных количествах. Этот случай демонстрирует, что делеция, вероятно, возникла на очень ранней стадии развития (возможно, в ооците), а ее последствия варьируются даже у генетически идентичных индивидов в зависимости от тканеспецифичного распределения и уровня гетероплазмии.

Исследования [10] показывают, что перестройки мтДНК (делеции, инсерции и дупликации) широко распространены в человеческих ооцитах и эмбрионах [10]. Они были выявлены с помощью двухэтапной ПЦР в 50.5% ооцитов (n = 295) и реже — в 32.5% эмбрионов (n = 197), что указывает на значительное снижение частоты перестроек после оплодотворения. Среди них наиболее распространена делеция мтДНК (4977 п.н.), обнаруженная в 47% ооцитов и

20% эмбрионов. При этом не наблюдалось возраст-зависимого увеличения вероятности появления перестроек, но было отмечено их значительное снижение по мере созревания ооцита. Значительное снижение количества ооцитов, содержащих перестройки мтДНК, происходило по мере развития ооцита от зародышевого пузырька до зрелого ооцита метафазы II. Также анализ показывает, что митохондриальные перестройки снижаются в эмбрионах на третий день после оплодотворения. Эти данные свидетельствуют о динамичном процессе элиминации митохондриальных перестроек на ранних этапах эмбриогенеза.

Другое исследование [11] неоплодотворенных ооцитов (метафазы II) человека выявило четкую связь между возрастом донора и состоянием митохондриальной ДНК. У женщин в возрасте 35 лет и старше наблюдалась значительно более высокая частота делеции мтДНК (4977 п.н.), которая была обнаружена в 34.6% ооцитов, а также меньшее общее количество копий мтДНК по сравнению с более молодыми женщинами, что указывает на возрастную зависимость мутаций мтДНК в ооцитах. При этом количество копий мтДНК имело отрицательную корреляцию с возрастом. Эти два параметра — возрастающая частота специфической делеции и снижение содержания мтДНК — позволяют предположить, что они являются маркерами старения яичников, которое обратно пропорционально потенциалу репликации и напрямую связано с метаболическим состоянием клетки.

Существуют убедительные доказательства гипотезы о том, что соматические делеции мтДНК являются прямой причиной возрастной дегенерации клеток [14, 15]. Одно из подтверждений получено в результате исследования мышей-мутаторов [14] ($PolgA^{mut/mut}$, мутантных мышей со значительно повышенным уровнем мутагенеза мтДНК) с дефектной версией мтДНК-полимеразы (PolgA), лишенной функции коррекции ошибок (из-за полимутаций в этом гене). У таких мышей наблюдается ускоренное накопление делеций мтДНК приводящих к преждевременному старению и сокращению продолжительности жизни, что подтверждает важную роль PolgA в процессе ремонта двойных разрывов цепей ДНК (DSBs). Большинство делеций у мышей дикого типа (WT) и гетерозиготных мышей ($PolgA^{+/mut}$) происходит между гомологичными последовательностями, особенно между двумя прямыми повторами длиной 15 п.н., которые являются точками активного образования делеций, напротив у мышей-мутаторов делеции преимущественно возникают между негомологичными последовательностями. Исследование выявило, что эти делеции часто образуются посредством механизма гомологически направленной репарации (HDR), а нормальная функция PolgA подавляет образование делеций между негомологичными последовательностями. При этом скорость, с которой мутации мтДНК проявляются фенотипически, существенно варьируется в разных тканях, что объясняет их различную восприимчивость к возрастной дегенерации.

Прямая причинно-следственная связь между накоплением соматических мутаций мтДНК и старением была экспериментально доказана в другом исследовании [15] на модели мышей-мутаторов (гомозиготных нокаутированных мышей) с дефектной версией мтДНК-полимеразы (PolgA), дефицитом коррекционной активности (proof-reading-deficient). У таких мышей наблюдалось трех-пятикратное увеличение уровня точечных мутаций и делеций мтДНК, что привело к сокращению продолжительности жизни и преждевременному появлению комплекса возрастных фенотипов, включая потерю веса, алопецию, кифоз, остеопороз, снижение фертильности и кардиомегалию. Эти результаты убедительно демонстрируют, что накопление мутаций мтДНК является не просто корреляцией, а именно причинным фактором в развитии характерных признаков старения у млекопитающих. Делеции мтДНК локализованы в определенном регионе, что указывает на роль нуклеотидных повторов в формировании делеций.

Наблюдение локализации делеций мтДНК в областях разрыва мышечных волокон [18] и нейронах с дефицитом дыхательной цепи [4] подтверждает гипотезу о причинном влиянии делеций мтДНК на старение. Профилирование экспрессии генов с последующим параметрическим анализом обогащения набора генного набора (PAGE) в исследовании [18] на мышах с мутантной митохондриальной ДНК-полимеразой (D257A), ускоряющей накопление делеций мтДНК, подтверждают причинную роль этих мутаций в развитии возрастной

саркопении. Наблюдаемая локализация делеций в зонах разрыва мышечных волокон и нейронах с дыхательной недостаточностью соответствует глубокому подавлению митохондриальных генов, снижению содержания комплексов электрон-транспортной цепи I, III и IV на 35-50%, нарушению синтеза АТФ и падению мембранного потенциала. Примечательно, что эта дисфункция развивается без увеличения продукции активных форм кислорода, что ставит под сомнение теорию "порочного круга" окислительного стресса. Таким образом, делеции мтДНК вызывают саркопению через нарушение сборки дыхательных комплексов и биоэнергетический дефицит, таким образом запуская митохондриальный-связанный путь апоптоза мышечных волокон, а следовательно потере мышечной массы и старению.

В целом, высокая доля делеций мтДНК не является нейтральным признаком стареющих клеток, а, скорее, причиной их появления. Таким образом, понимание молекулярных механизмов, лежащих в основе происхождения соматических делеций мтДНК, а также скорости их распространения, имеет первостепенное значение [29, 20]. По результатам одного из исследований [20], клональная экспансия делеций мтДНК, приводящая к их концентрациям нарушающим функцию митохондрий, может происходить по двум сценариям: при медленном процессе мутации-основатели возникают в раннем возрасте, а при быстром — преимущественно в позднем. Хотя первоначальные данные высокого разрешения интерпретировались как подтверждение раннего происхождения мутаций, повторный анализ с помощью метода цифрового обнаружения делеций (Digital Deletion Detection) указывает на преимущественно позднее возникновение мутаций-основателей. Это фундаментальное противоречие в определении критического периода для защиты от мутаций (раннее развитие либо поздний возраст) требует разрешения в дальнейших исследованиях. Скорость клональной экспансия может зависеть от типа клетки, индивидуальной активности, уровня активных форм кислорода (ROS) и длины делеций мтДНК. Авторы обнаружили, что количество различных типов делеций не увеличивается с возрастом, в то время как их распространение увеличивается. Данные свидетельствуют о том, что клональная экспансия делеций мтДНК больше соответствует сценарию «быстрого» распространения, при котором делеции быстро распространяются в начале жизни, а затем, уже существующие делеции, более медленно распространяются в с возрастом.

Высокая доля делеций мтДНК является не нейтральным маркером, а причиной старения клеток, что делает понимание механизмов их возникновения и распространения первостепенно важным. Одно из исследований [29] показывает, что эти соматические мутации часто возникают на очень ранних этапах жизни — во время внутриутробного развития или даже оогенеза. Подтверждением этой гипотезы служат данные о том, что мутагены, такие как нуклеозидные аналоги принимаемые ВИЧ-инфицированными пациентами (класс противовирусных препаратов использующийся для лечения ВИЧ и гепатита В), не столько создают новые мутации (которые затем расширяются независимо от воздействия NRTIs), сколько ускоряют клональную экспансию уже существующих делеций. Таким образом, патогенные делеции мтДНК изначально присутствуют в клетках, а процесс старения представляет собой их длительное и медленное клональное распространение, приводящее к митохондриальной недостаточности.

Исследование [31] пациентов с успешно вылеченной ВИЧ-инфекцией показывает, что применяемые антиретровирусные препараты (нуклеозидные аналоги) вызывают ускоренное старение митохондрий, а именно - прогрессивное накопление мутаций соматической мтДНК, представляющих собой те же самые мутации что и наблюдаемые гораздо позже в жизни, то есть вызванные нормальным старением. Это происходит не за счет усиления мутагенеза, а благодаря ускоренному обороту мтДНК, который приводит к клональной экспансии ранее существовавших соматических мутаций. Такой механизм вызывает биохимический дефект, поражающий до 10% клеток, и воспроизводит картину нормального старения, но в ускоренном темпе. Это приводит к ускоренному увеличению частоты COX-дефицитных мышечных волокон и повышению частоты связанных с митохондриальной функцией клинических осложнений у пожилых пациентов. Эти наблюдения подтверждают роль соматических мутаций

мтДНК в процессе старения и указывают на риск прогрессирующих митохондриальных заболеваний у данной категории пациентов. Помимо ускоренного клонального расширения предсуществующих мутаций, возможно, что существуют дополнительные механизмы, такие как преимущество репликации в пользу молекул с делециями.

Если основной причиной старения является возникновение делеций, то можно предположить что при отсутствии факторов увеличивающих вероятность образования делеций - старение будет отсрочено. Существуют японские гаплогруппы долгожителей и сверхдолгожителей которые предположительно имеют меньше факторов возникновения делеций [27, 28]. Исследование [27] митохондриальных гаплогрупп японских долгожителей показало, что хотя гаплогруппа D4a является маркером чрезвычайного долголетия, эта корреляция обусловлена не функциональными полиморфизмами в мтДНК, а популяционной структурой. Статистический анализ полных последовательностей мтДНК 112 сверхдолгожителей (старше 105 лет) не выявил значимых функциональных SNP, связанных с долголетием. Для фенотипа сверхдолгожителей наблюдается единственный статистически значимый сигнал: постепенное обогащение определенных «полезных» паттернов у долгожителей и сверхдолгожителей в гаплогруппе D4a, которая является маркером чрезвычайного долголетия в Японии. Наблюдаемое обогащение гаплогруппой D4a может объясняться "эффектом автостопа" - связью с неидентифицированным аутосомным генетическим событием в истории популяции, датируемым после появления мегагруппы D. Таким образом, несмотря на статистическую ассоциацию определенных митохондриальных гаплогрупп с долголетием, причиной этого, вероятно, являются ядерные генетические факторы, а не специфические защитные варианты мтДНК.

Анализ 672 полных митохондриальных геномов неродственных японцев (стратифицированных на семь равноразмерных групп по фенотипам: пациенты с диабетом, пациенты с диабетом с тяжелой ангиопатией, здоровые молодые мужчины, не страдающие ожирением, молодые мужчины с ожирением, пациенты с болезнью Альцгеймера, пациенты с болезнью Паркинсона и долгожители) от тех же авторов [28] в котором они составили полный список «паттернов мутаций» для долгожителей - выявил значимое обогащение гаплогруппами D4a, D5 и особенно подкластером D4b2b (отмеченного синонимической мутацией 9296C>T) среди долгожителей. Предполагается, что связь этих митохондриальных гаплогрупп с долголетием, вероятно, является результатом эффекта «автостопа» — ко-селекции с некой адаптивной ядерной мутацией, координацией между мтДНК и ядерными мутациями, а не следствием функциональных особенностей самой мтДНК. По оценкам авторов, это гипотетическое аутосомное событие отбора, предрасполагающее к долголетию, произошло примерно 24.4 ± 0.9 тысяч лет назад, после возникновения мегагруппы D, но до дивергенции её субклад D4a, D5 и D4b2b.

1.2 Роль повторов в образовании делеций

Хотя уже давно известно, что прямые нуклеотидные повторы (или длинные несовершенные дуплексы) способствуют образованию делеций мтДНК, они до сих пор объясняют лишь небольшую часть наблюдаемого распределения делеций. Это поднимает вопрос о том, почему некоторые повторы приводят к делециям, а другие — нет, и какие еще факторы могут участвовать в формировании делеций. Необходимо понять основные факторы влияющие на образование делеций мтДНК. В этом может помочь анализ частоты «синдромов делеции мтДНК» с ранним началом, классически состоящих из синдрома Кернса-Сейра (KSS), синдрома Пирсона и прогрессирующей наружной офтальмоплегии (PEO) [12, 13]. В случае ооцитов распространение делеций мтДНК потенциально может проявляться во всех тканях, в том числе пролиферативных, приводя к мультисистемным нарушениям.

Несмотря на известную роль прямых нуклеотидных повторов в образовании делеций

мтДНК, они объясняют лишь небольшую часть случаев. Исследование синдромов Кернса-Сейра (KSS) и прогрессирующей наружной офтальмоплегии (PEO) [12] выявило, что распространенная делеция длиной 5 тыс. пар оснований, встречающаяся у более чем трети пациентов, фланкирована идеальным прямым 13-парным повтором. Это указывает, что механизм подобный гомологичной рекомбинации, ранее наблюдавшейся только у низших эукариот, действует и в митохондриях млекопитающих. Все делеции локализовались в областях мтДНК, содержащих компоненты дыхательной цепи, и не были обнаружены в областях рибосомальной РНК или в областях начала репликации или транскрипции. Предполагается, что механизмом образования делеций может быть проскальзывание репликации (slipped mispairing или replication slippage). Предполагается, что делеции в мтДНК могут происходить в два этапа: спонтанное ошибочное спаривание перед репликацией и разрешение ошибочно спаренных промежуточных фрагментов во время репликации. Области митохондриальной ДНК, содержащие длинные участки гомопуриновых/гомопиримидиновых участков, могут быть особенно подвержены проскальзыванию репликации. Последовательности ДНК в полипиримидиновом блоке в области консервативной последовательности D-петли, которые демонстрируют гетерогенность (5'-CCCCCCCTCCCCCGCTTCT-3' на L-цепи), удаленный 9-пн повтор в гене COII (5'-CCCCCTCTA-3') и ядро самого 13-пн повтора KSS/PEO (5'-ACCTCCCTCACCA--3') - все они обладают потенциалом образовать изогнутую ДНК. Предполагается, что эти структуры подвержены изгибу и плавлению ДНК, что потенциально приводит к образованию одноцепочечной ДНК при суперспирализации. Можно предположить что 13-пн совершенный прямой повтор, обнаруженный в мтДНК в позициях 8470–8483 и 13447–13460, и фланкирующий общую делецию - является горячей точкой делеции, а её формированию дополнительно способствуют специфические структурные особенности ДНК: АТ-обогащенные участки, способные к изгибу и плавлению, а также последовательности, образующие тройные спирали (Н-ДНК), что создаёт условия для рекомбинации через ошибочное спаривание.

Исследователями были собраны клинические характеристики синдромов делеции мтДНК [13], включая прогрессивную PEO, KSS, и другие наследственные заболевания, связанные с мтДНК. Синдромы единичной крупномасштабной делеции мтДНК (SLSMDS), включающие синдромы Кернса-Сейра, Пирсона и хроническую прогрессирующую наружную офтальмоплегию, представляют собой спектр перекрывающихся заболеваний, возникающих из-за de novo делеций в зародышевой линии или эмбриогенезе. Эффект бутылочного горлышка во время развития ооцита и эмбриона позволяет лишь небольшому количеству молекул мтДНК проникнуть в плод, с возможностью «проскока» поврежденной мтДНК. Эти делеции, выявляемые в ДНК лейкоцитов крови и мочи, делятся на два класса: делеции I класса, окруженные прямыми повторами и возникающие в результате гомологичной рекомбинации, и делеции II класса с неизвестным механизмом образования. Наиболее распространенная делеция (m.8470_13446 del4977), как и другие, часто затрагивает гены тРНК и мРНК, кодирующие субъединицы дыхательной цепи, что приводит к нарушению синтеза белка, дефициту энергии и широкому спектру клинических проявлений.

Общим повтором [21, 22, 26] называется самый длинный совершенный прямой повтор в мтДНК человека (длиной 13-пн, плечи начинаются с позиций 8470 и 13447). Было показано, что большинство соматических делеций мтДНК фланкированы прямыми нуклеотидными повторами [21] или длинными несовершенными дуплексами, состоящими из коротких участков прямых повторов [22]. На основании этого наблюдения была выдвинута гипотеза, что общий повтор (совершенный прямой повтор длиной 13 пар оснований) по-видимому, является основным фактором, ответственным за образование большинства делеций. Анализ 263 различных делеций [21] показал, что их точки разрыва поразительно соответствуют положению плеч общего повтора, создавая пики в распределении повреждений внутри главной дуги мтДНК, то есть подавляющее большинство различных делеций мтДНК, по-видимому, связано с этим повтором, что указывает на общий механизм, связанный с репликацией, при котором экспонирование одноцепочечной ДНК этого повтора может приводить к образованию тройной

спирали (по модели Хугстина [207] и последующему делеционному событию. Это может происходить во время асимметричной репликации или в репликационном «пузыре», образуемом симметричным процессом репликации. Область, прилегающая к предполагаемому стоп-сигналу, участвующему в синтезе H-цепи в районе D-петли, высококонсервативна и потенциально способна формировать шпильковую структуру. Эта область связана с барьером репликативной вилки (RFB, Replication Fork Barrier), что указывает на связь между формированием делеции и репликацией мтДНК. Таким образом, общий повтор считается наиболее мутагенным прямым повтором в митохондриальном геноме, и его наличие критически предопределяет индивидуальную склонность к накоплению делеций.

В то же время известны гаплогруппы у которых общий повтор утратил совершенность которая наблюдается у подавляющего большинства человеческой популяции. Например повтор нарушен у гаплогруппы N1b1 [22], а также у двух гаплогруппы японских долгожителей (D4a, D5a) [27, 28]. Исследование [22] опровергает устоявшееся представление о ключевой роли общего повтора в образовании большинства делеций мтДНК. Было показано, что распределение точек разрыва делеций у индивидуумов без данного повтора практически идентично таковому у людей с распространённым 13-пн повтором. Это указывает на то, что основной механизм делециогенеза связан не только с совершенными повторами, но и с образованием стабильных, хотя и несовершенных, дуплексов между отдаленными участками мтДНК, на что указывает сильная корреляция между стабильностью таких межсегментных дуплексов и распределением делеций. В исследовании были изучены два варианта митохондриального генома: один содержит 13-пн повтор, а другой принадлежит к гаплогруппе N1b, в которой 5'-конец повтора изменен на 8472C>T полиморфизм, который находится рядом с D4a полиморфизмом (8473T>C). В ходе исследования были проанализированы сегменты мтДНК длиной 100 п.н. из точек разрывов делеций (5700–10737 и 11400–16100) и обнаружено, что наиболее подходящие дуплексы обычно включают короткие двухцепочечные области, разделенные петлями, что позволяет предположить, что на делеции может влиять стабильность этих структур. Количество прямых повторов в мтДНК обратно коррелирует с продолжительностью жизни у различных видов млекопитающих. Общий повтор часто ассоциируется с наиболее частыми делециями, но его отсутствие у носителей митохондриальной гаплогруппы D4a среди долгожителей Японии позволяет предположить, что делеции могут быть связаны со старением, но их влияние на продолжительность жизни незначительно. Таким образом, формирование делеций определяется в большей степени общей структурной стабильностью дуплексов, что объясняет, почему их накопление, вероятно, оказывает минимальное влияние на продолжительность жизни.

Была выдвинута нейтральная теория которая предсказывает мультигенное старение и увеличение концентрации вредных мутаций в митохондриях и Y-хромосомах [23]. В статье исследуется эволюционная история человека, начиная от дивергенции от приматов 5 миллионов лет назад до начала сельского хозяйства 10 000 лет назад. Авторы утверждают, что в течение большей части этой эволюции человек не достигал значительных численности в старшем возрасте, что приводило к тому, что гены, отвечающие за старческие признаки, не могли быть отфильтрованы естественным отбором. В результате, эти гены были фиксированы на нейтральном уровне в течение большей части эволюции человека. Основные факторы влияющие на эволюцию человека включают низкую численность популяции, периодический голод и интенсивное перемещение, характерные для образа жизни охотников-собирателей. В малых популяциях, таких как охотники-собиратели, фиксация нейтральных и слегка вредных генов происходит быстрее, что обусловлено генетическим дрейфом. Предполагается, что эти гены фиксируются в результате генетического дрейфа, а не под влиянием положительного отбора. С эволюционной точки зрения, прямые повторы в мтДНК, такие как общий повтор, представляют собой пример аллелей, вредных в позднем возрасте (DILL). Эти аллели не отсеивались естественным отбором на протяжении большей части эволюции человека, так как их вредное воздействие проявлялось после репродуктивного возраста. Вследствие низкого эффективного размера популяции (Ne) и отсутствия рекомбинации, мтДНК и Y-хромосома

подвержены ускоренной фиксации таких умеренно вредных мутаций. Это объясняет как высокую скорость эволюции мтДНК и частоту митохондриальных заболеваний, так и меньшую продолжительность жизни мужчин. Высокая вероятность фиксации нейтральных и умеренно вредных мутаций в митохондриальном геноме частично объясняет быструю скорость его эволюции, высокую наблюдаемую частоту митохондриальных заболеваний в связи с небольшим размером этого генома и может быть основной причиной переноса митохондриальных генов в течение эволюционного времени в ядро. Таким образом, высокая концентрация вредных мутаций в мтДНК, включая структурно нестабильные повторы, является результатом ослабленного очищающего отбора и служит одной из основных причин митохондриальной дисфункции при старении.

Была выявлена отрицательная корреляция между количеством прямых повторов в мтДНК и видоспецифической продолжительностью жизни млекопитающих [24, 25], что интерпретируется как дополнительное свидетельство вредного воздействия повторов в мтДНК долгоживущих млекопитающих. Анализ митохондриальных геномов 61 вида млекопитающих [24] выявил отрицательную корреляцию между количеством длинных прямых повторов в мтДНК и видоспецифической продолжительностью жизни. Это интерпретируется как прямое свидетельство того, что данные повторы, способствуя образованию делеций (таких как общая делеция 4977 п.н. у человека), являются одним из ключевых факторов, ограничивающих максимальную продолжительность жизни млекопитающих значениями около 80–100 лет. Накопление этих делеций, ведущее к гетероплазмии и нарушению митохондриальной функции, создает общевидовой паттерн, в котором виды с более короткой продолжительностью жизни как правило, имеют больше прямых повторов в последовательностях мтДНК и структурно более нестабильную мтДНК, предрасположенную к повреждениям по сравнению с видами с более длинной продолжительностью жизни. Они также показывают, что количество повторов, происходящих в случайно перетасованных последовательностях, является грубым нижним пределом количества повторов в реальных последовательностях. Исследование предполагает, что образование делеций через прямые повторы может быть более значимым у видов с более короткой продолжительностью жизни, потенциально влияя на их клональную экспансию и генетический дрейф.

Исследование связи между прямыми повторами в мтДНК и продолжительностью жизни 65 видов млекопитающих [25], в котором проанализировали последовательности мтДНК млекопитающих на наличие прямых повторов, учитывая их относительную мутагенность с поправкой на длину и общие показатели мутагенности (TMS) для сравнения видов, выявило значительную отрицательную корреляцию, особенно выраженную у близкородственных видов. Это указывает на то, что прямые повторы, являющиеся мощным источником независимого от АФК мутагенеза и способствующие образованию делеций, служат одной из ключевых детерминант видовой продолжительности жизни. Подтверждением этой гипотезы служит пример долгоживущих летучих мышей, у которых низкое количество прямых повторов и, как следствие, более низкий общий мутагенный потенциал (TMS) прогнозирует медленное возрастное ухудшение функции митохондрий с возрастом.

Анализ митохондриальных геномов 762 неродственных японцев [26] выявил внутривидовую отрицательную корреляцию: меньшее количество прямых совершенных повторов в мтДНК статистически связано с большей продолжительностью жизни. Ключевым доказательством этого механизма служит гаплогруппа D4a, характеризующаяся долголетием, в которой точечная мутация 8473C нарушает структуру общего повтора (главной горячей точки общей делеций). Это позволяет предположить, что уменьшение количества таких повторов в определенных гаплогруппах, подобно пониженному количеству прямых повторов в мтДНК долгоживущих млекопитающих, снижает частоту соматических делеций мтДНК, тем самым способствуя более здоровому старению и отсрочке возрастных патологий. Авторы предполагают, что негативное воздействие этих повторов, вероятно, проявляется с опозданием до позднего периода жизни из-за внутриклеточного отбора, который усиливает негативное воздействие повторов на организм.

Исследование с использованием метода mito-SMARD для изучения репликации мтДНК в клетках человека и мыши [32] продемонстрировало, что образование общей делеции мтДНК (4977 п.н.) непосредственно связано с остановкой репликативной вилки в области общего повтора. Нарушение репликации в этом локусе создает условия для репликационно-зависимого пути репарации, который опосредуется митохондриальной реплисомой и протекает независимо от канонического репаративного пути двунитевых разрывов (DSB). Этот процесс, вероятно, включает механизм микрогомологически опосредованного соединения концов (MMEJ), где остановка вилки приводит к разрывам ДНК, последующее восстановление которых и формирует делецию. Таким образом, общая делеция возникает в результате уникального пути репарации, инициированного специфическим нарушением репликации мтДНК.

1.3 Связь механизма образования делеций со структурой митохондриального генома

Детальное изучение распределения делеций мтДНК ставит под сомнение абсолютную роль повтора длиной 13 п.н. и других идеальных повторов в возникновении делеций мтДНК. Делеции, по-видимому, зависят от длинных и стабильных, хотя и несовершенных, дуплексов между отдаленными сегментами мтДНК. Более того, значительные различия в распределении точек останова позволяют предположить, что в создании делеций мтДНК участвуют несколько механизмов.

Исследование 202 видов млекопитающих [51] выявило, что инвертированные повторы (IR) в мтДНК имеют значительно более сильную отрицательную корреляцию с максимальной продолжительностью жизни (MLS), чем прямые повторы (DR). Показано, что IR вызывают репликационно-зависимые инверсии, которые с возрастом накапливаются в постмитотических тканях, таких как мозг и сердце. Эти инверсии признаны более мутагенными, чем DR-опосредованные делеции, поскольку могут генерировать два аберрантных продукта и вызывать сложные перестройки, нарушающие гены и митохондриальную функцию. Таким образом, нестабильность генома, вызванная IR во время репликации мтДНК, идентифицируется как ключевой фактор, налагающий более строгое ограничение на продолжительность жизни. Исследование идентифицирует индуцированную IR нестабильность митохондриального генома во время репликации мтДНК как потенциальную причину митохондриальных дефектов.

Проводя параллели между делециями у бактерий [33], общим повтором мтДНК [34] и ядерной ДНК [35], можно предположить, что прямые повторы могут с большей вероятностью вызывать делеции, когда они находятся в непосредственной близости друг от друга. Таким образом, повышенная вероятность появления делеций вблизи общего повтора поддерживается не самим общим повтором, а независимым топологическим фактором. Делеции могут быть вызваны специфическими, например, С-богатыми мотивами [34] внутри прямых повторов. Возможно что на образование делеций влияет не только сходство ДНК между областями точек разрыва, но и пространственная структура. Это бы подтвердило механизм проскальзывания репликации, при котором вложенный паттерн прямых и инвертированных повторов (далее DIID: Direct Inverted Inverted Direct) может привести к образованию делеций [33, 34].

Исследование спонтанных делеций (длиной 700–1000 п.н.) в Escherichia coli [33] показало, что короткие гомологичные последовательности (прямые повторы) играют ключевую роль в образовании крупных. Наибольшая частота образования делеций наблюдается в RecA+ штаммах, где рекомбинация усиливает этот процесс в 25 раз, причем самая активная горячая точка образуется между участками с гомологией 14 из 17 пар оснований. Замена всего одного нуклеотида в этой области снижает частоту делеций на порядок, что подтверждает критическую зависимость процесса от степени гомологии между повторами. Эти данные раскрывают механизм, при котором вторичная структура ДНК, поддерживаемая

инвертированными повторами, создает уязвимые сайты для образования делеций. Типы делеций, которые могут восстановить ферментативную активность, включают тип A (делеции, которые восстанавливают нормальную рамку считывания, то есть отсутствие сдвига) и тип B (делеции, которые устраняют только одну из начальных мутаций, но сами вызывают изменение фазы, противоположное знаку оставшейся мутации сдвига рамки считывания). Минимальный размер типов A делеций составляет 700-1000 bp, а тип B делеций может быть как маленьким (450 bp), так и большим, но в этом случае он должен быть ограничен чтением рамки +1.

Эксперимент на мтДНК пациентов с генетическими мутациями в гене POL γ (in vivo и in vitro) [34] демонстрируют, что образование делеций мтДНК происходит посредством рекомбинации с избирательным копированием (copy-choice recombination, рекомбинация выбором копии) во время репликации легкой цепи (синтеза L-цепи). Этот процесс, усиливающийся при мутациях в POL γ, характеризуется специфической направленностью и преимущественно происходит между прямыми повторами (особенно 13-пн повторами "общей делеции"), причем вторичные структуры ДНК, такие как G-квадруплексы (G4), дополнительно стимулируют делециогенез. Обнаружены повторы, включая 13-пн повторы и короткие повторяющиеся последовательности, которые могут быть связаны с процессом выбора копии. Обнаружены последовательности, богатые цитозинами (например, CCC, ACCC, ACCCC), которые также часто встречаются в областях делеций, образованных в живых клетках. При использовании ПЦР были обнаружены продукты длиной 1500 бп и 750 бп, что указывает на возможное участие процесса выбора копии в формировании делеций. Делеции часто наблюдаются между прямыми повторами, особенно в областях с non-B-структурами, такими как шпильковые структуры и G4. Эти делеции подразделяются на три типа: класс I (между гомологичными прямыми повторами), класс II (между несовершенными повторами) и класс III (между неповторяющимися последовательностями). 3'-конец D-петли определяется как регуляторный центр для повторов, расположенных близко друг к другу. Во время репликации, если репликативный аппарат диссоциирует, 3'-конец вновь синтезированной ДНК может неправильно спариваться со вторым повтором, что приводит к образованию гетеродуплексной молекулы и удаления плеча повтора, ближайшего к OriH. Исследование подтверждает модель рекомбинации с избирательным копированием, показывая, что делеции зависят от активной репликации мтДНК и что делеции образуются при разрезе на 3'-конце 13-пн повтора рядом с OriL. Полученные результаты критикуют модель "слипшейся цепи" (slipped-strand model) и подтверждают, что делеции являются следствием ошибок репликационного аппарата, когда диссоциация реплисомы приводит к неправильному спариванию 3'-конца вновь синтезированной ДНК со вторым плечом повтора, что объясняет накопление делеций при митохондриальных заболеваниях и в процессе старения.

Исследование [35] показывает, что структурная организация ядерного генома сама по себе создает хрупкие участки: эволюционно консервативные якоря хромосомных петель, удерживаемые белками CTCF (фактор связывания CCCTC) и когезином, являются предопределенными сайтами двухцепочечных разрывов ДНК (DSB). Эти разрывы опосредованы топоизомеразой 2B (TOP2B), которая разрешает топологическое напряжение, возникающее при сворачивании хромосом, и этот процесс не зависит от транскрипции, репликации и типа клеток. Полиморфизмы, изменяющие связывание CTCF, могут перераспределять эти хрупкие сайты. Исследование показывает, что ETO (этопозид, ингибитор топоизомеразы II, противораковый препарат) вызывает значительное увеличение количества DSB, особенно в областях, где присутствует CTCF, что указывает на важную роль этого белка в организации хроматина и формировании топологических структур. Более того, исследование выявляет, что DSB наиболее часто локализуются в областях, где присутствует белок CTCF, а также часто локализуются вблизи сайтов начала транскрипции (TSS, transcriptional start sites) и активных промоторов, что может указывать на роль TOP2B в регуляции транскрипции и релаксации положительной суперспирализации ДНК и формировании топологических структур, таких как петли и кольца хроматина. Полиморфизмы в мотивах связывания CTCF влияют на образование разрывов ДНК, так как они могут изменять аффинность белка к ДНК и,

следовательно, регулировать топологические стрессоры. Таким образом, якоря петель служат универсальным механизмом генерации DSB и хромосомных перестроек, которые часто транслоцируются при раке. Аналогичный механизм разрыва ДНК может быть на 5’ концах делеций мтДНК.

1.4 Инструменты предсказания вторичной структуры митохондриальной ДНК

Так как во время репликации митохондриальная ДНК частично находится в одноцепочечном состоянии - имеет смысл для понимания процессов способствующих мутагенезу мтДНК использовать инструменты предсказывающие структуру РНК.

Как правило существующие инструменты предсказания вторичной структуры цепей нуклеиновых кислот опираются на минимизацию свободной энергии, либо структуру уже известных паттернов, а принципиально новых методик, кроме использования машинного обучения, не появлялось уже давно. Все рассмотренные мной инструменты можно разделить на следующие категории: (1) Базы данных для размещения и аннотирования микросателлитов; (2) Программные средства, предназначенные для обнаружения и аннотирования совершенных микросателлитов; (3) Программные средства, предназначенные для обнаружения и аннотирования совершенных микросателлитов; (4) Базы данных перемежающихся повторов в геномах органелл и ядерном геноме; (5) Алгоритмы поиска несовершенных перемежающихся повторов всех четырех классов; (6) Инструменты обнаружения повторов, основанные на локальном попарном выравнивании и / или самовыравнивании; (7) Дот-матрица визуализации и анализа повторов; (8) Идентификация повторов с использованием скорректированной частоты k-мер; (9) Алгоритмы на основе преобразований Фурье; (10) Поиск повторов в цикличной ДНК.

Веб-сервер Mfold [44] предоставляет набор тесно связанных инструментов для прогнозирования вторичной структуры одноцепочечных нуклеиновых кислот (РНК и ДНК) на основе алгоритмов минимизации свободной энергии, рассчитывающих наиболее стабильные конформации, включая структуры с минимальной свободной энергией (ΔG). Сервер визуализирует результаты в различных форматах — структурные графики, диаграммы энергии и точечные диаграммы, а также поддерживает пакетную обработку последовательностей через опцию быстрого сканирования (Quikfold) для предсказания множества последовательностей под одним и тем же условием, что делает его удобным инструментом для анализа вторичных структур, включая исследования стабильности и фолдинга митохондриальной ДНК.

Пакет ViennaRNA 2.0 [45] представляет собой комплекс вычислительных инструментов для точного прогнозирования и анализа вторичных структур РНК на основе термодинамических параметров и алгоритмов динамического программирования. Обновленная версия сохранила вычислительную эффективность, добавив поддержку параллельных вычислений через OpenMP, расширенные методы оценки взаимодействий РНК-РНК, анализ ограниченных структурных ансамблей, а также новые форматы вывода, включая структуры центроида и максимальной ожидаемой точности. Эти возможности позволяют использовать пакет для оценки энергии Гиббса гибридизации участков мтДНК, предоставляя количественную метрику для оценки стабильности взаимодействий между одноцепочечными областями ДНК.

В одной из статей [36] произведен обзор важности предсказания структуры РНК, рассмотрены группы существующих биоинформатических методов используемых наравне с экспериментальными методами определения структуры РНК (биофизическими и биохимическими), которые в совокупности позволяют делать предсказания всё более точно, а также озвучены проблемы с которыми сталкиваются подобные инструменты. Для вычислительного моделирования трехмерных структур РНК на основе данных перечислены

ограничения, существующие подходы, недавние успехи, проблемы и перспективы, а также бенчмарк для оценки методов прогнозирования трехмерной структуры РНК - RNA-Puzzles (объединился с CASP).

Авторы выделяют следующие инструменты и подходы вычислительного моделирования трехмерных структур РНК сгруппированные по используемым типам структурной информации дополняющим друг друга:

1) Трехмерная форма молекулы варьируется от рентгеновской кристаллографии высокого разрешения до SAS, которая описывает только глобальную топологию структуры
   - карты плотности на основе атомных координат полученных в ходе рентгеновской кристаллографии и крио-ЭМ высокого разрешения
2) Характеристики одного нуклеотида, такие как спаренное/неспаренное состояние и скрытое/выставленное состояние, могут быть выведены из данных химического зондирования
   - метод RNAsc в соответствии с данными SHAPE, который включает псевдоэнергетические термины (pseudo-energy terms) и положения укладки оснований (base stacking positions), что позволяет точнее прогнозировать вторичную структуру
   - использование псевдоэнергетической информации, полученной из химического зондирования, и интеграция с термодинамическими алгоритмами сворачивания для достижения хорошего прогноза вторичной структуры
   - подход для фильтрации данных SHAPE от шума
   - метод RING-MaP выявляет разнообразные взаимодействия как на вторичном, так и на третичном уровнях
   - FoldAtlas разработанный для высокопроизводительной обработки данных химического зондирования
   - алгоритм SNPfold разработаный для распознавания конформационных изменений, вызванных SNP, при полногеномном анализе
   - алгоритм DREEM который может быть использован для характеристики различных конформаций из данных DMS-MaPseq
   - использование машинного обучения для прогнозирования вторичной структуры посредством интеграции анализа прямого связывания и данных SHAPE (Selective 2′-Hydroxyl Acylation analyzed by Primer Extension)
3) Взаимодействие пар оснований не может быть выведено только из ковариаций последовательностей, но также может быть определено из биохимических экспериментов или предсказания вторичной структуры
   - предсказание вторичной структуры MFE (минимум свободной энергии) с помощью комбинации энергетической модели на основе петель и алгоритма динамического программирования
   - использование комбинации экспериментов по химическому зондированию и ковариации последовательностей
   - MOHCA-seq продемонстрировал свою способность определять как Уотсон-Криковские, так и не-Уотсон-Криковские пары оснований дальнего радиуса действия
   - информация о парах оснований часто используется вместе с особенностями одного нуклеотида. Например, M2-seq сочетается с DMS для зондирования структуры РНК.

Типы подходов к прогнозированию структуры:

1) Сравнительное моделирование основанное на предположении, что гомологичные молекулы РНК представляют схожие структуры в относительно консервативных регионах. РНК можно предсказать, используя известную гомологичную структуру в

базе данных в качестве шаблона
- использовать попарное выравнивание последовательностей для моделирования на основе структуры шаблона (ModeRNA)
- преобразовать структуру шаблона в структурные ограничения в предсказании (RNABuilder)

2) Сборка фрагментов - разлагает известные структуры РНК на фрагменты и создает библиотеку фрагментов, а также предсказывает структуру РНК путем поиска фрагментов, похожих на целевую последовательность, и сборки их вместе
   - использование вторичной структуры в качестве предопределенного ограничения (например, RNAComposer, 3dRNA и VfoldLA)
3) Моделирование *de novo* — это собирательное название для прогнозов без шаблонов
   - моделирование силового поля и поиск идеальной структуры путем выборки конформационного пространства (например, NAST, iFoldRNA и SimRNA)
4) Интегративное моделирование - если гомологичный шаблон недоступен, сборка фрагментов и моделирование *de novo* могут интегрировать различные типы информации для помощи в прогнозировании
   - сначала предсказать вторичную структуру и предсказать трехмерную структуру в соответствии с вторичной структурой. Одним из простых случаев является RNAComposer: можно использовать RNAfold из ViennaRNA для прогнозирования вторичной структуры и использовать RNAComposer для сборки трехмерной модели в соответствии с предсказанной вторичной структурой с использованием библиотеки фрагментов RNA FRABASE
   - Более сложная версия может использовать DMS-seq и M2-seq для ограничения вторичной структуры, в то время как использовать MOHCA-seq для исследования третичных взаимодействий. Затем использовать сборку фрагментов для построения начальной трехмерной модели и оптимизировать ее с помощью конформационного поиска с использованием Rosetta или других методов, основанных на силовом поле
   - использовать ковариации из множественного выравнивания последовательностей для подтверждения пар оснований/взаимодействий и использовать методы модульного прогнозирования для назначения неканонических пар оснований. В таком подходе интегрируется несколько типов информации (особенности нуклеотидов, взаимодействие пар оснований и неканонические взаимодействия) в нескольких методах прогнозирования (сборка фрагментов и моделирование *de novo*)
   - вычислительный рабочий процесс EvoClustRNA эффективно интегрирует *de novo* моделирование и сборку фрагментов и статистический потенциал для достижения точных прогнозов. EvoClustRNA сначала выбирает набор последовательностей, гомологичных целевой последовательности для моделирования, и использует как FARFAR (сборка фрагментов), так и SimRNA (моделирование *de novo*) для моделирования всех последовательностей
   - Интегративная платформа моделирования представляет собой вычислительную платформу для интеграции данных SAS, EM, рентгеновской кристаллографии или ЯМР путем сравнительного моделирования
   - PLUMED-ISDB, который основан на библиотеке молекулярной динамики PLUMED . Он может интегрировать экспериментальные данные из ЯМР, FRET, SAXS или крио-ЭМ для моделирования структуры и динамики РНК, а также других биомакромолекул.
5) молекулярная динамика (МД) и квантовая механика (КМ).

Конформации неканонических форм ДНК (non-B DNA) — это молекулярные структуры, которые не следуют канонической двойной спирали ДНК. Мутагенная нестабильность в

геномах ядерной и митохондриальной ДНК (мтДНК) связана с простыми конформациями non-B ДНК, такими как шпильки, или более сложными структурами, такими как G-квадруплексы. Чтобы лучше охарактеризовать Структуру A (non-B-конформация, похожая на клеверный лист, предсказанная для домена контрольной области мтДНК), авторы одной из статей [37] предсказали ее 3D-конформацию с использованием моделирования молекулярной динамики. Также были рассчитаны оценки сохранения выравниваний области Структуры A у людей, приматов и млекопитающих. В данном исследовании был использован пакет Unified Nucleic Acid Folding and hybridisation (UNAFold) — хорошо зарекомендовавшее себя программное обеспечение, которое использует экспериментально определенные термодинамические параметры для прогнозирования неканонических вторичных структур РНК и ДНК. Данный инструмент использует минимизацию свободной энергии (MFE) с использованием термодинамических правил ближайшего соседа. Точность пакета UNAFold оценивалась с использованием базы данных из 22 экспериментально определенных структур одноцепочечной ДНК, полученных с помощью спектроскопии ядерного магнитного резонанса (ЯМР) со шпильками и стеблями. Эти трехмерные структуры сравнивались с предсказанием вторичной структуры UNAFold. {Шаг 1} Вторичная структура, предсказанная программным обеспечением UNAFold, была использована для создания 3D-моделей конформаций non-B ДНК с использованием веб-сервера RNAComposer. {Шаг 2} Преобразование созданных 3D-моделей одноцепочечной РНК в одноцепочечную ДНК с использованием LEaP из программного обеспечения AmberTools. {Шаг 3} Молекулярно-динамическое моделирование трехмерных non-B-конформаций ДНК с использованием TIP3P, LeaP, AmberTools18 и SHAKE. Авторы считают что использованные вычислительные методы способны измерять стабильность non-B-конформаций, используя уровень спаривания оснований во время молекулярной динамики. Структура A показала высокую стабильность и низкую гибкость, коррелирующую с высокими показателями консервативности у млекопитающих, а точнее, у приматов. Авторы показали, что 3D non-B-конформации могут быть предсказаны и охарактеризованы использованной методологией. Это позволило провести глубокий анализ структуры A, и основные результаты показали, что структура остается стабильной во время симуляционного моделирования. Моделирование показывает, что структура A обладает высокой стабильностью со средним среднеквадратичным отклонением (RMSD) 10 Å по сравнению с исходной предсказанной структурой. Анализ результатов молекулярной динамики показывает, что стабильность некоторых участков структуры мтДНК может значительно варьироваться, со значительными вариациями RMSD в определенных сегментах. Оценки консервативности и вариации RMSF (мера среднего смещения атомов при моделировании молекулярной динамики) в структуре A указывают на то, что она играет решающую роль в модуляции репликации мтДНК и высококонсервативна у разных видов. Стабильность определенных участков в структуре мтДНК, как предсказано UNAFold, коррелирует с образованием non-B-конформаций и возникновением делеций. Состав нуклеотидов влияет на образование non-B-конформаций, при этом определенные последовательности приводят к более высоким вариациям RMSD.

Анализ 9655 делеций мтДНК млекопитающих и 1307 делеций у C. elegans с использованием термодинамики гибридизации ДНК [38] показал, что двухцепочечные разрывы (DSB) являются ключевыми инициаторами делеционных мутаций. В исследовании рассматривали две основные модели: модель слипания нитей (slip-strand model) и модель ошибок репарации DSB. Модель слипания нитей предполагает, что делеции возникают во время репликации мтДНК из-за ошибочного спаривания одноцепочечных участков, тогда как модель ошибок репарации DSB объясняет делеции ошибочной репарацией DSB. Было обнаружено значительное преобладание микрогомологий длиной 0-25 п.н. в точках разрыва, что подтверждает ведущую роль ошибочного негомологичного соединения концов и микрогомологически-зависимого репаративного пути в формировании делеций. Эти данные, подкрепленные экспериментами с трансгенными мышами, экспрессирующими митохондриальные рестриктазы, свидетельствуют о механизме, при котором делеции возникают вследствие неправильного соединения концов DSB или инвазии этих концов в

открытые участки мтДНК, что опровергает модель слипания нитей и устанавливает ошибки репарации DSB как основной источник делеционного мутагенеза. Проведенный авторами анализ (с использованием UNAFold) показывает, что большую часть положений точек разрыва мтДНК можно объяснить термодинамикой коротких ≤ 5-нуклеотидных смещений. Значимость коротких несовпадений ДНК подтверждает важную роль ошибочной негомологичной и микрогомологически-зависимой репарации DSB в образовании делеций мтДНК.

Исследование того как митохондриальное происхождение и тип ткани в процессе коэволюции ядерного и митохондриальных геномов влияют на соматическую эволюцию митохондриального генома в различных тканях в процессе старения (с использованием сверхчувствительного дуплексного секвенирования - DupSeq) [39] показало, что митохондриальные геномы представляют собой динамичную популяцию, эволюционирующую на протяжении жизни организма под влиянием соматических мутаций и отбора. Анализ ~1.2 миллиона мутаций в пяти гаплотипах и трёх тканях мышей выявил специфичные для гаплотипа мутационные паттерны, где главной горячей точкой неизменно выступает область начала репликации легкой цепи (OriL), сохраняющая консервативную структуру «стебель-петля» несмотря на видовые различия в последовательности. Показано, что грызуны демонстрируют уникальный спектр соматических мутаций с преобладанием G>T/C>A замен, связанных с окислительным стрессом, в то время как мутации в кодирующих белки генах несут сигнатуры отрицательного отбора. Важным открытием стало выявление обширной соматической реверсии за счет мутаций, которые в течение жизни изменяют мтДНК гаплогруппу (митохондриальное происхождение), отражая непрерывный процесс мито-ядерной коэволюции на соматическом уровне. Кроме того, выявлены области с повышенной частотой мутаций в гене MT-ND2 и tRNAArg, что может быть связано с консервативной структурой этих областей. Минимальная частота мутаций обнаружена в функциональных областях генома, таких как генетические кодирующие области, рибосомальные РНК (рРНК) и транспортные РНК (тРНК). В статьеопределены точки мутаций в регионе OriL, консервативной структуре мтДНК. Эти точки связаны с возрастом и гаплотипом и одинаковы для разных тканей и линий мышей. Предполагается, что мутации в регионе OriL вызваны праймированием репликации РНК, подверженной проскальзыванию, что указывает на специфическую мутационную горячую точку из-за консервативной структуры стебля-петли. В работеподчёркивается, что по мере старения клеток в них накапливаются мутации в мтДНК, которая становится более подверженной повреждениям из-за отсутствия защитных гистонов и более высокой скорости репликации. Это согласуется с наблюдаемым увеличением частоты соматических мутаций мтДНК с возрастом. Предполагается, что мито-ядерные геномные взаимодействия могут формировать мутационный ландшафт мтДНК, при этом несоответствие между ядерным и митохондриальным происхождением приводит к снижению числа копий мтДНК и снижению её функции.

Определение трехмерных структур ДНК становится все более необходимым для понимания их функций, проектирования новых молекул ДНК или выбора ДНК-аптамеров. Но для ДНК такие методы недоступны, прогресс в этом направлении не такой значительный как в методах прогнозирования трехмерных структур РНК (iFoldRNA, Assemble, RNA Frabase, 3dRNA, SimRNA, HiRE-RNA, ModeRNA, RNAComposer, Vfold, RNA2D3D, Rosetta, MC-Fold, MC-Sym, 3dRNA). Поскольку количество экспериментальных структур ДНК в настоящее время ограничено, авторы одной из статей [40] предложили вычислительный метод 3dDNA, который решает проблему предсказания трёхмерных структур ДНК — области, значительно отстающей от прогнозирования структур РНК. Предположительно трехмерные структуры РНК находятся в состоянии минимальной свободной энергии, поэтому их прогнозирование обычно включает два этапа: выборку конформационного пространства и выбор модели минимальной свободной энергии. Современные методы прогнозирования трехмерных структур РНК можно условно разделить на два класса: подход ab initio и подход на основе шаблонов. Первый ищет трехмерную структуру РНК, используя молекулярную динамику для моделирования процесса ее сворачивания. Второй ищет трехмерную структуру РНК, находя и собирая трехмерные

шаблоны из экспериментальных структур РНК, которые имеют схожую последовательность или фрагменты последовательности с целевой РНК. В отличие от существующих косвенных подходов, которые сначала предсказывают структуру РНК с последующей заменой урацила на тимин и энергетической минимизацией, 3dDNA использует прямой шаблонный подход. Метод автоматически собирает 3D-структуру ДНК из библиотеки шаблонов, объединяющей фрагменты экспериментально определённых структур ДНК и РНК (с заменой U на T). Рабочий процесс включает поиск шаблонов для элементов вторичной структуры, их сборку и оптимизацию, что позволяет достичь средней точности предсказания около 2.36 Å RMSD для одно- и двуцепочечных структур и менее 4 Å для многоцепочечных комплексов.

Метод EvoClustRNA [41] использует эволюционную информацию, получаемую из гомологичных последовательностей одного семейства РНК, которые, как известно, сворачиваются в консервативную трёхмерную структуру, для улучшения точности предсказания структур РНК ab initio. В этом многоэтапном процессе гомологичные последовательности для целевой РНК сначала отбираются с помощью базы данных Rfam, после чего для каждого гомолога независимо проводятся симуляции сворачивания с использованием методов Rosetta FARFAR и SimRNA. Заключительная модель для целевой последовательности выбирается не по критерию минимальной энергии, а на основе наиболее распространённого структурного расположения общих спиральных фрагментов, выявляемого путём кластеризации всех полученных моделей. Эффективность подхода была подтверждена в слепых тестах RNA-Puzzles, где предсказания EvoClustRNA заняли первое место для рибопереключателя L-глутамина и второе — для ZMP-рибопереключателя. Более того, авторы обнаружили, что для некоторых целевых последовательностей моделирование их конкретных гомологов может дать более точные структурные модели, чем моделирование исходной последовательности ab initio.

Как уже было сказано, текущие методы предсказания 3D-структуры РНК можно разделить на две группы: методы на основе шаблонов и методы de novo. Методы на основе шаблонов предсказывают целевую структуру с использованием гомологичных шаблонов в PDB. Например, репрезентативные методы, такие как ModeRNA и MMB, работают за счет сокращения пространства выборки с помощью гомологичных структур. В целом, модели предсказанных структур с помощью методов на основе шаблонов точны, когда гомологичные шаблоны существуют в PDB. Напротив, методы de novo создают 3D-конформации, моделируя процесс сворачивания с нуля. При молекулярно-динамическом моделировании и/или сборке фрагментов такие методы, как FARNA, FARFAR, FARFAR2, SimRNA, iFoldRNA, RNAComposer и 3dRNA, хорошо работают для определенных малых РНК (<100 нуклеотидов). Тем не менее, сложно сгенерировать точные 3D-структуры для больших РНК со сложной топологией из-за неточных параметров силового поля и огромного пространства выборки. Чтобы частично решить эту проблему, для управления моделированием структуры использовались межнуклеотидные контакты, предсказанные с помощью анализа прямого сцепления (DCA). Кроме того, учитывая иерархическую природу сворачивания структуры РНК, несколько методов выводят 3D-структуры из вторичных структур, таких как Vfold и MC-Fold. Они очень быстры, но точность моделирования во многом зависит от качества входных вторичных структур. Глубокое обучение недавно было использовано для улучшения предсказания 3D-структуры de novo РНК. Предсказанные межнуклеотидные контакты остаточной сверточной сетью (ResNet) примерно в два раза точнее, чем DCA, что в некоторой степени улучшает предсказание 3D-структуры. Было показано, что при выборе модели из геометрической системы оценки на основе глубокого обучения (ARES) протокол FARFAR2 предсказал наиболее точные модели для четырех целей в слепом тесте экспериментов RNA-Puzzles. Недавно, вдохновленные успехом AlphaFold2, были разработаны несколько новых методов на основе глубокого обучения, таких как DeepFoldRNA, RoseTTAFoldNA и RhoFold. Авторы одной из статей [42] разработали trRosettaRNA - автоматизированный подход для предсказания трехмерной структуры РНК, основанный на глубоком обучении. Он частично вдохновлен успешным применением глубокого обучения в предсказании структуры белка, особенно в AlphaFold2 и предыдущем методе авторов - trRosetta. Конвейер trRosettaRNA

состоит из трех основных этапов: (1) подготовка нескольких множественных выравниваний против баз данных NCBI, Rfam и RNAcentral и отбор лучшего; (2) прогнозирование 1D- и 2D-геометрии с помощью нейронной сети состоящей из трансформеров; (3) генерация моделей полноатомной структуры - складывание трехмерной структуры за счет минимизации энергии. Вторичная структура предсказывается SPOT-RNA на основе запросов, визуализация вторичной структуры РНК выполняется с помощью forna, а поиск шаблона и моделирование 2D-структуры с помощью программы R2DT.

По результатам авторитетных международных конкурсов алгоритмов предсказания структур, RNA-Puzzles и этапов CASP15/CASP16 [208,209], себя наиболее эффективно показали такие инструменты как: AIchemy_RNA2, использующий базовый потенциал знаний на основе ротамерной и квантовой механической энергии (BRiQ) для уточнения структуры РНК; Vfold, который применяет методологии, основанные на законах физики и термодинамики для моделирования петель и сложных топологий; пакет GuangzhouRNA-human объединяющий вычислительные инструменты с экспертными знаниями; а также проверенные временем платформы GeneSilico и RNAPolis, известные своими гибридными стратегиями, комбинирующими биоинформатический анализ и сравнительное моделирование.

В одной из статей [43] был проанализирован полный митохондриальный геном находящейся под угрозой исчезновения европейской норки Mustela lutreola. Акцент исследования был сделал на контрольный регион (CR) и мотивы, которые могут образовывать стабильные вторичные шпилечные структуры. В ходе анализа мтДНК были идентифицированы прямые и инвертированные повторы. В пределах региона между позициями 16 035 и 16 180 п.н. была обнаружена область из 146 п.н., богатая инвертированными повторами. Для данной области была предсказана структура образующая одноцепочечные шпилечные структуры. Вторичные структуры тРНК и рРНК были исследованы с помощью MITOS WebServer, а программные пакеты RNAstructure и RNAfold были использованы для прогнозирования потенциальных вторичных структур контрольной области и начала репликации легкой цепи (OL), при том что когда было возможно более одной вторичной структуры, использовалась та, которая имела наименьшую оценку свободной энергии. CR разделен на три домена: ETAS (расширенная терминирующая последовательность), CD (центральный консервативный домен) и CSB (три консервативных участка). Домен CD содержит пять консервативных боксов (F-бокс, E-бокс, D-бокс, C-бокс, B-бокс) и три консервативных блока последовательностей (CSB1, CSB2, CSB3). Домен CSB включает два предполагаемых промотора инициации транскрипции (HSP и LSP). Мотив в домене CSB может образовывать стабильные вторичные шпилечные структуры со стеблем, содержащим 12 пар оснований, и петлей, содержащей 3 нуклеотида. CR содержит VNTR (участки ДНК, состоящие из повторяющихся коротких последовательностей нуклеотидов, длина которых варьируется). Результаты исследования авторы связывают с вторичной структурой мтДНК, подчеркивая, как вторичная структура CR и вариабельные участки способствуют образованию делеций, которые могут влиять на общую длину и изменчивость последовательности мтДНК.

Помимо инструментов предсказывающих структуру цепей нуклеиновых кислот на основе последовательности могут быть полезны и другие инструменты и данные. Перестройки митохондриальной ДНК (мтДНК) являются ключевыми событиями в развитии многих заболеваний. Исследования областей мтДНК, затронутых перестройками (т.е. точек разрыва), могут привести к важным открытиям о механизмах перестроек и дать важные подсказки о причинах митохондриальных заболеваний. Авторы [46] представили базу данных точек разрыва митохондриальной ДНК (MitoBreak; http://mitobreak.portugene.com), бесплатный, доступный онлайн, полный список точек останова трех классов соматических перестроек мтДНК: кольцевых молекул мтДНК с делецией, кольцевых молекул мтДНК с частичными дупликациями и линейных молекул мтДНК. Точки разрыва часто располагаются внутри прямых повторов, и в соединяющих их последовательностях могут сохраняться одна или две копии этих повторов. В настоящее время MitoBreak содержит >1400 перестроек мтДНК семи видов (Homo sapiens, Mus musculus, Rattus norvegicus, Macaca mulatta, Drosophila melanogaster,

Caenorhabditis elegans и Podospora anserina) и связанную с ними фенотипическую информацию, собранную из почти 400 публикаций. Для каждого зарегистрированного случая MitoBreak также документирует точные положения точек разрыва, последовательность соединений разрывов, заболевание или связанные с ним симптомы и дает ссылки на соответствующие публикации, предоставляя полезный ресурс для изучения причин и последствий структурных изменений мтДНК. База данных MitoBreak, содержащая коллекцию делеций мтДНК человека может быть использована для сопоставления предсказанных структур с подтвержденными местами образования делеций.

Карты Hi-C (High-Resolution Interactions Capture) представляют контакты ДНК-ДНК в виде матрицы контактов, где элементы указывают количество контактов между локусами. Применение метода Hi-C для для картирования структуры бактериальных хромосом высокого разрешения [47] выявило организацию хромосомы Caulobacter crescentus в стабильные, независимые домены, сохраняющиеся на протяжении клеточного цикла и восстанавливающиеся параллельно с репликацией ДНК. Эти домены, вероятно, состоят из суперспирализованных плектонем, образующих структуру, напоминающую ёршик (bottle brush–like fiber). Эти домены стабильны на протяжении всего клеточного цикла и восстанавливаются одновременно с репликацией ДНК. Установлено, что границы доменов формируются высокоэкспрессируемыми генами и участками, свободными от плектонем, в то время как гистоноподобный белок HU и SMC-комплекс обеспечивают краткодистантную компактизацию и связывание плеч хромосомы соответственно. Эти результаты раскрывают фундаментальные принципы компактизации и пространственной организации минихромосом (в т.ч. митохондриальных) in vivo, подчеркивая необходимость исследований с высоким разрешением для понимания того, как хромосомы компактизируются и организуются внутри клеток.

В исследовании, создавшем трёхмерную карту человеческого генома с разрешением в 1 килобазу [48], авторы применили усовершенствованную методику Hi-C in situ к линии лимфобластоидных клеток человека, получив беспрецедентно плотную карту из 4,9 миллиардов контактов. Анализ выявил, что геном организован в локальные домены, которые соответствуют различным образцам гистоновых меток и группируются в шесть субкомпартментов. Ключевым открытием стало идентификация около 10 000 петель хроматина. Эти петли часто соединяют промоторы и энхансеры, что коррелирует с активацией генов, и демонстрируют высокую степень консервативности у разных типов клеток и видов. Якоря этих петель расположены на границах доменов и ассоциированы с сайтами связывания белка CTCF, при этом более 90% CTCF-мотивов в якорях находятся в конвергентной ориентации, будучи «обращёнными» друг к другу. Полученная общедоступная карта взаимодействий высокой плотности предоставляет уникальный ресурс для проверки пространственной организации генома.

Эксперименты по картированию контактов, такие как Hi-C, исследуют, как геномы сворачиваются в 3D. Авторы [49] разработали Juicebox.js, облачную систему для визуализации и изучения полученных наборов данных Hi-C, которая позволяет пользователям увеличивать и уменьшать масштаб таких наборов данных, используя интерфейс, аналогичный Google Earth, а также включает специальные функции для обеспечения воспроизводимости и обмена данными.. Ключевой особенностью является кодирование точного состояния браузера в публичном URL-адресе, что позволяет создавать общедоступные браузеры для новых наборов данных Hi-C. В сочетании с инструментом Juicer этот подход обеспечивает полную прозрачность анализа 3D организации ДНК — от необработанных данных до публикации результатов, — что делает Juicebox.js ценным ресурсом для визуализации пространственной организации, в том числе и для анализа контактных матриц мтДНК.

1.5 Заключение к обзору литературы

Мутации митохондриальной ДНК, особенно делеции, являются основной причиной митохондриальных заболеваний и старения. Традиционно главным механизмом их образования считался неравный кроссинговер между прямыми нуклеотидными повторами. Хотя изначально основное внимание уделялось совершенным прямым повторам, последующие исследования выявили значимость длинных несовершенных повторов и других типов взаимодействий, включая инвертированные повторы, способные вызывать инверсии. Более того, актуальные работы указывают на вклад неканонических структур (таких как G-квадруплексы) и удаленных взаимодействий между сегментами мтДНК, что расширяет представление о возможных детерминантах долголетия и нестабильности мтДНК. Накопленные данные свидетельствуют о том, что одних лишь повторов недостаточно для полного объяснения паттернов соматического мутагенеза, и что ключевую роль, вероятно, играет вторичная структура одноцепочечной мтДНК. Проведенные авторами исследования не только не объясняют механизма образования делеций, но даже не показывают однозначной связи с причинами их возникновения, поэтому требуются новые, более сложные модели образования делеций и их проверка.

Современные вычислительные методы предоставляют мощный инструментарий для проверки новой парадигмы, согласно которой вторичная и третичная структура мтДНК является критическим фактором мутагенеза. Для этого могут быть применены инструменты предсказания структуры РНК/ДНК, основанные на минимизации свободной энергии (Mfold, ViennaRNA), а также передовые подходы, включая методы de novo (SimRNA, FARFAR), шаблонное моделирование (3dDNA) и интегративные платформы с использованием машинного обучения (EvoClustRNA, trRosettaRNA). Комплексное использование этих инструментов позволит перейти от анализа первичной последовательности к моделированию конформационных особенностей мтДНК, что откроет новые возможности для понимания механизмов образования делеций, а следовательно процесса старения и митохондриальных патологий.

Глава 2. ImtRDB: база данных и программное обеспечение для аннотации митохондриальных несовершенных вкрапленных повторов

2.1 Проблематика

В целом, до сих пор нет единой и устоявшейся модели, объясняющей влияние повторов в мтДНК на продолжительность жизни животных. Чтобы ответить на этот вопрос, возникла необходимость иметь набор данных со всеми типами нуклеотидных повторов в полностью отсеквенированных митохондриальных геномах. Выделяют четыре основных типа вкрапленных повторов, различающихся с точки зрения расположения их плеч (одна и та же цепь: прямой и зеркальный или разные цепи: инвертированный и комплементарный) и направления (одного направления: прямой и комплементарный; противоположного направления: зеркальный и инвертированный); кроме того, каждый тип повтора может характеризоваться уровнем его деградации (совершенным и несовершенным). Описанные типы повторов изображены на **Рисунке 1**.

Рисунок 1. Четыре типа перемежающихся повторов. Цвета обозначают повторяющийся нуклеотидный рисунок, стрелки указывают направление рисунка

2.2 Аналитическое сравнение существующих инструментов для поиска повторов и баз данных содержащих информацию о повторах в мтДНК

В настоящее время, насколько мне известно, не существует баз данных, хранящих и аннотирующих как совершенные, так и несовершенные вкрапленные повторы мтДНК любой длины и природы. Например, две наиболее популярные и полные базы данных, содержащие вкрапления повторов в органелларном и ядерном геномах, ориентированы в основном на повторы транспозонной и тРНК природы (RepBase [63], Dfam [64]). Также в существующих в настоящее время базах данных учитывается только ограниченное число типов повторов (например, только прямые) и/или уровень их вырожденности (например, только совершенные повторы). Объединение таких специализированных баз данных нецелесообразно из-за разных алгоритмов, используемых для идентификации различных типов повторов, а также из-за разных таксономических групп анализируемых видов.

Мной было проанализировано около 70 статей по алгоритмам идентификации вырожденных повторов. Основная проблема в том, что существующие алгоритмы поиска вырожденных повторов либо сильно ограничены степенью вырожденности (например

максимум 5% нуклеотидов могут быть некомплементарными) либо, как RepeatMasker, основываются на уже существующей базе последовательностей повторов-образцов. Для митохондрий не существует баз данных с консенсусными последовательностями вырожденных повторов-образцов. Для поиска повторов существует множество инструментов, но их систематизацию производит только RepeatMasker.

Существует ряд разработанных вычислительных алгоритмов для поиска несовершенных повторов всех четырех классов. По стратегии обнаружения повторов их можно условно разделить на четыре группы: (1) алгоритмы, основанные на локальном парном выравнивании (с использованием, например, эвристического поиска по BLAST [65] или точного поиска по PALS [66] или построения суффикса деревья, опосредованные поиском максимального уникального совпадения (MUM), реализованным в MUMmer [67]); (2) инструменты, использующие матричный анализ; (3) алгоритмы, основанные на анализе избыточной представленности k-меров, и (4) алгоритмы поиска периодичностей с использованием преобразований Фурье. Наиболее известным программным инструментом для идентификации повторов на основе предварительно скомпилированной базы данных повторов является RepeatMasker [68]. Инструменты обнаружения повторов, основанные на локальном парном выравнивании и/или самовыравнивании по сравнению инструменты анализа чрезмерного избыточной представленности (обогащения) k-мерами (или l-мерами, N-мерами, т.е. короткими подстроками нуклеотидов), потенциально могут позволить более точно идентифицировать короткие копии повторяющихся последовательностей и более точное распознавание фланговых регионов. Предварительный этап локального попарного выравнивания необходим, например, для расчетов RECON [69], REPET [70], PRAP [71] и PILER [72]; тогда как для RepEx [73] требуется предварительная идентификация MUM. Матричная визуализация и анализ реализованы, например, в DOTTER [74], Adplot [75], Gepard [76], JDotter [77], PLOTREP [78], r2cat [79], D-GENIES [80]. Матричный анализ обычно требовал визуального контроля полученных графиков, что связано с различными недостатками идентификации повторяющихся фланкирующих областей из-за большого размера окна. Идентификация повторов может быть выполнена с использованием скорректированной частоты k-меров в качестве начальных чисел и жадного расширения каждой начальной затравки (с помощью наивного алгоритма или построения суффиксных массивов/деревьев или путем идентификации элементарных повторов) до более длинной консенсусной последовательности. Этот подход был реализован, например, в RepeatScout [81], REPuter [82], SPADE [83], WindowMasker [84], Vmatch [85], phRAIDER [86]. Алгоритмы, основанные на преобразованиях Фурье, реализованы, например, в программном обеспечении Spectral Repeat Finder [87] и SBARS [88] на основе нуклеотидов и DNADU на основе динуклеотидов [89]. Следует отметить, что спектр мощности Фурье не может точно характеризовать повторы из-за невозможности идентифицировать повторяющийся образец, количество копий и уровень вырождения.

Важно отметить что все вышеописанные алгоритмы работают с линейной ДНК. Однако митохондриальная ДНК имеет кольцевую форму, и это свойство важно учитывать при поиске повторов в мтДНК. Единственная программа, работающая с кольцевыми молекулами, — это RepeatAround [90], однако она не позволяет находить несовершенные повторы.

Еще одной особенностью повторов мтДНК является их довольно короткая длина, большинство из них короче 20 оснований [91, 51, 24, 53, 25, 22, 92, 93]. Кроме того, из множества экспериментальных данных DNAseq и / или RNAseq, а также из особенностей взаимодействия miRNA / mRNA [94, 95, 96, 97, 98, 99, 100, 101, 102] известно, что существует ограничение на минимальную длину участков ДНК или РНК, которые полезны для спаривания оснований (например, для правильного взаимодействия miRNA / mRNA требуется идеальное спаривание оснований из 7-8 нуклеотидов коровой (seed) области miRNA [94]). Таким образом, важно было разработать алгоритм, ориентированный на короткие повторы – алгоритмы способные работать с несовершенными повторами до 10 нуклеотидов практически отсутствуют.

Среди огромного спектра программных продуктов поиска и анализа повторов есть интересные для поставленной цели. Наиболее подходящими являются Vmatch [85], RepEx [73], Gepard [76] и RepeatAround [90]. Однако все они нацелены на поиск специфических групп достаточно протяженных повторов, именно поэтому с моей точки зрения они имеют существенные недостатки, следовательно необходимо разработать новый алгоритм специализированный именно на митохондриальных геномах с их короткими повторами. Для меня важно разработать эффективный и чувствительный алгоритм для обнаружения четырех типов диспергированных вырожденных повторов, чтобы они в итоге могли потенциально получить дальнейшее изучение.

2.3 Создание алгоритма и базы

В результате сравнительного анализа существующих подходов поиска повторов и подстрок в тексте, в качестве основы для создания алгоритма идентификации повторов в мтДНК были выбраны наиболее быстрые подходы построения матриц точечной гомологии, основанные на адаптивном (по детализации) поиске областей самоподобия. Анализ протяженности ранее идентифицированных повторов в мтДНК, а также особенностей минимально необходимых областей для комплементарного спаривания нитей ДНК, позволил биохимически оптимизировать этот алгоритм в контексте уникальной природы мтДНК.

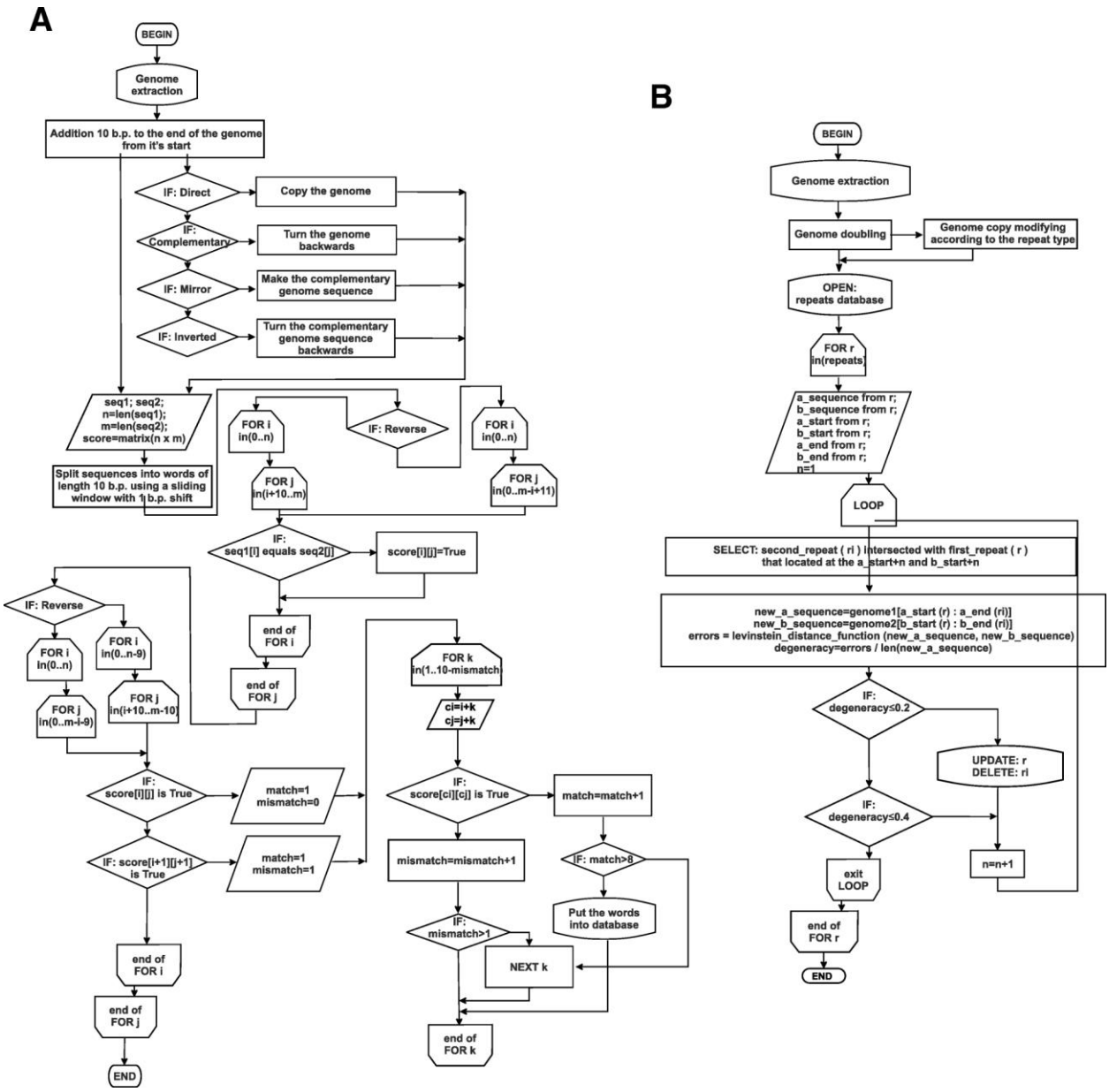

Рисунок 2. Блок-схема алгоритма поиска повторов. А - распознавание похожих коротких нуклеотидных паттернов; Б - слияние коротких паттернов

Я реализовал на языке Python наивный алгоритм распознавания образов по аналогии со стандартными процедурами построения точечных диаграмм. Алгоритм состоит из двух этапов: (I) распознавание сходных коротких нуклеотидных паттернов и (II) слияние коротких паттернов (**Рисунок 2**). Распознавание подобных коротких нуклеотидных шаблонов основано на заранее заданном скользящем окне длиной 10 п.о. и максимальной вырожденности этой длины 10%. Этот порог длины и вырожденности был выбран, потому что минимальное биологически значимое спаривание составляет около 8-10 оснований (например, канонические коровые сайты взаимодействия miRNA- mRNA имеют длину 7–8 оснований). [94], однако спаривание оснований за пределами корового региона необходимо для функции miRNA [95], оптимальная длина праймеров для техники случайной амплификации полиморфной ДНК - 10 оснований [96, 97, 98], минимальная длина совпадения для значимой интенсивности гибридизации зонда составляет около 10 оснований [99]). Для более длинных повторов учитывался максимум 20% вырожденности, которая позволяет образовывать высокостабильные структуры, несмотря на наличие неспаренных оснований. Чтобы рассмотреть кольцевую мтДНК на {*шаге 1*} этапа I копируется 10 нуклеотидов из начальной позиции митохондриального генома в конец генома, удлиняя таким образом геном на 10 п.н.. На основе

удлиненной последовательности мтДНК на {*шаге 2*} генерируются четыре дополнительные последовательности: идентичная (копия), комплементарная, зеркальная и зеркально-комплементарная. Основная процедура поиска похожих коротких нуклеотидных последовательностей {*шаг 3*} проводится как анализ скользящего окна на половине *l*l* квадратной симметричной «точечной» матрицы, где *l* длина мтДНК +10 п.н. Дополнительные последовательности используются в матрице «точечных графиков» для поиска различных типов повторов: используя копию основной последовательности для прямого поиска повторов, комплементарная последовательность использовалась для комплементарных повторов, зеркальная последовательность для обнаружения зеркальных повторов, и последовательность зеркально-комплементарная для обнаружения инвертированных повторов. Если был обнаружен повторяющийся шаблон {*шаг 4*} - фиксируются координаты обеих последовательностей (запросной и целевой) в геноме как запись базы данных SQLite, содержащая местоположения двух сегментов генома. Общая вычислительная сложность этапа распознавания подобных коротких нуклеотидных последовательностей составляет $O((m-10)^4)$, где m — размер анализируемого генома.

Чтобы найти повторяющиеся паттерны длиной более 10 п.н., итеративно проверяются все короткие повторы (полученные на этапе I) на предмет их пересечения. Пересечение двух коротких повторов подтверждаются, если повторяющиеся последовательности обоих повторов в двух и более местах генома имеют равное и коллинеарное (по типу повторов) смещение {*шаг 1*} относительно друг друга. В результате каждого раунда объединения коротких повторов удлиняются (меняются координаты генома) один повтор в базе данных SQLite и удаляется другой, пересекающийся с этим. Делается это тогда и только тогда, когда {*шаг 2*} полученный объединенные повтор имеет вырожденность не более 20%. Повторяется *шаг 1* и *шаг 2* стадии слияния коротких шаблонов до тех пор, пока не перестанут генерироваться новые объединенные повторы. Примечательно, что для рассмотрения цикличности мтДНК этап слияния коротких паттернов был основан на дублированном геноме, генерируемом путем конкатенации двух геномных последовательностей. В целом сложность вычислений на этапе слияния коротких шаблонов составляет $O(2m^2)$, где m — количество простых повторов, найденных на первом этапе.

Допускались только уникальные единичные (нетандемные) несовпадения повторов из-за того что: (1) средняя длина ранее известных несовершенных митохондриальных повторов не более 20 оснований [91, 51, 24, 53, 25, 22, 92, 93]; (2) термодинамика дуплексообразования в случае перемежающихся несовпадений является аддитивной, линейной и хорошо изученной [99, 100]; (3) имеется зависимость стабильности дуплекса тандемного (множественного) несовпадения нуклеотидов от идентичности, длины и специфичности контекста фланкирующих пар оснований [101]. Кроме того, не анализируются повторы отличающиеся делециями по тем же причинам (контекстно-специфическая зависимость устойчивости спаривания таких мотивов [99, 102]). Эти упрощения биологически значимы из-за короткой длины несовершенных митохондриальных повторов, обнаруженных в ходе данного исследования (в среднем 12 оснований).

Я сравнил свой алгоритм с другими опубликованными алгоритмами. Для этой цели я выбрал два хорошо описанных митохондриальных генома – геном Homo sapiens (NC_012920) и Mus musculus (AY172335). Для сравнения я выбрал три алгоритма: современный широко востребованный алгоритм Vmatch [85], RepeatAround [90], предназначенный для анализа кольцевой ДНК, универсальный алгоритм RepEx [73], основанный на максимальных уникальных совпадениях. Результаты сравнения отображены в **Приложения А и Б**. Все идеальные повторы найденые Vmatch и RepEx были найдены также моим алгоритмом, а также все случайно выбранные для визуального контроля повторы, найденные RepeatAround, также были найдены моим алгоритмом. Vmatch обнаруживает значительно большее количество несовершенных повторов с тандемными заменами, в то время как мой алгоритм отбрасывает такие случаи, в то же время Vmatch не обнаружил подавляющего большинства повторов

размером 10 п.н. с одним несоответствием и более длинных повторов с рассредоточенными несоответствиями, в то время как мой алгоритм эффективно нашел такие повторы.

Созданным мной алгоритмом было проанализировано 3715 митохондриальных геномов позвоночных полученных с использованием электронные утилиты NCBI [103]. Анализируемые виды имеют следующий состав: 7 классов, 139 отрядов, 659 семейств.

Чтобы структурировать данные в базе данных, сделать их интерактивными и общедоступными, создан интернет-ресурс (http://bioinfodbs.kantiana.ru/ImtRDB/). Для этого использован веб-сервер Apache, MySQL 5, Perl 5.24 (модуль CGI), HTML5 и JavaScript для динамической генерации веб-страниц. Использован jBrowse [104, 105] для интерактивной визуализации митохондриальных геномов и нескольких повторяющихся треков. База данных MySQL очень проста, она состоит из 3 реляционных таблиц: таблица, содержащая таксономию видов, извлеченную из Таксономии NCBI, которая используется для поиска видов по таксономии; таблица, связывающая идентификатор NCBI мтДНК с названиями видов; и таблицу, содержащую результаты корреляционного анализа (*p* - коэффициент корреляции Пирсона и *rho* - коэффициент ранговой корреляции Спирмена) между плотностью повторов и физико-химическими характеристиками повторов, которые используются для создания страниц видоспецифичной корреляции (подробности см. ниже).

Чтобы охарактеризовать распределение повторов вдоль каждого митохондриального генома, скоррелирована плотность повторов в конкретном регионе с несколькими физико-химическими характеристиками этого региона. Сначала вычисляется положение средней точки для каждого плеча (мономера) каждого повтора как целое число (start_position + (end_position-start_position) / 2). В каждом повторе имеется по крайней мере две копии (мономера), поэтому каждый повтор характеризуется по крайней мере двумя срединными точками, распределенными в митохондриальном геноме. Во-вторых, для каждой последовательности мономеров (копии) каждого повтора вычислены различные физико-химические особенности (Emboss package v. 6.6 [106]) и присвоены эти значения положению средней точки плеча. Использованы следующие программы Emboss и соответствующие функции: (1) btwisted (для расчета общей энергии укладки; средней энергии укладки на динуклеотид; общего числа витков; средних оснований за виток и полного витка в градусах); (2) dan по параметру «-thermo» (для расчета содержания GC,%; температуры плавления спаренных оснований повторяющихся последовательностей; изменение свободной энергии Гиббса, энтальпии и энтропии в основаниях повторяющихся последовательностей ДНК); (3) compseq (для расчета 16 фракций динуклеотидов). В-третьих, сопоставлено число повторов в середине отрезков (плотность мономеров повторов, имеющих одинаковую среднюю точку) со средними физико-химическими характеристиками, назначенными каждой средней точке на втором этапе. Все статистические анализы были выполнены в R v. 3.4.1.

Самый простой способ представить распределение повторов в митохондриальном геноме – нарисовать арки между каждой парой мономеров. Однако этот способ визуализации трудно воспринимать из-за большого количества повторов в каждом геноме и большого количества плеч в каждом повторении (два плеча – одна дуга, три плеча - три дуги, четыре плеча – шесть дуг и т. д.). Чтобы свести к минимуму количество дуг, которые нужно нарисовать, фокус сделан на трех показателях (минимальное и максимальное расстояние между плечами данного повтора и расстояние между плечами с максимальным сходством) и реализовалась вероятностная процедура соединения плечей повторов так, чтобы ближайшие плечи повторов были наиболее вероятно связаны дугой. Вероятность связывания мономеров с помощью дуги определяется экспоненциальным распределением вероятности со средним значением распределения, равным 1/16 длины митохондриального генома. Сгенерированы эти вероятности связывания на участке от 0 до 1/2 длины митохондриального генома (из-за цикличности генома две точки генома, удаленные на половину длины генома, являются наиболее удаленными точками). Таким образом, если расстояние между комплементарными повторяющимися областями значительно превышает 1/16 митохондриального генома, то вероятность того, что комплементарные области связываются с помощью арки, имеет тенденцию стремится к нулю.

Пользовательский интерфейс видоспецифических данных основан на jBrowse [104, 105]. Есть возможность сравнивать числа повторов на нуклеотид между выбранными видами путем множественного выравнивания мтДНК. Кроме того, для каждого аннотированного вида можно просмотреть корреляцию между плотностью повторов и физико-химическими характеристиками.

С помощью созданной базы данных можно попытаться ответить на такие вопросы как: (I) Как менялись плотности повторов в ходе эволюции мтДНК гоминид? (II) Как плотность повторов у Pongo abelii коррелирует с их физико-химическими свойствами?

Каждый аннотированный митохондриальный геном имеет 33 характеристики визуализированных с помощью jBrowse: (1) местоположения и описания генов, как в файле genbank; (2) плотность повторов на нуклеотид; (3) плотность средних точек повторов на нуклеотид; (4) средняя доля 16 динуклеотидов, картированных в положениях средней точки повторов в мтДНК; (5) средний процент GC повторов, нанесенных на карту для средних точек повторов; (6) средняя температура плавления повторяющихся областей спаривания оснований, сопоставленных с средними точками повторов; (7) среднее изменение свободной энергии Гиббса, энтальпии и энтропии спаренных оснований в областях повторов, умноженное на -1 и сопоставленное со средними точками повторов; (8) средняя полная энергия укладки и энергия укладки на динуклеотид спаренных оснований в областях повторов, умноженная на -1 и нанесенная на карту для повторяющихся средних точек; (9) среднее общее число витков и оснований на виток спаренных оснований в областях повторов, отображенной на средних точках повторов; (10) среднее общее число витков спаренных оснований в областях повторов, в градусах, сопоставленных с средними точками повторов; (11) минимальное и максимальное расстояние между плечами повторов (или мономерами); (12) расстояние между плечами повторов с максимальной комплементарностью.

2.4 Результаты

Был проведен статистический анализ повторов по физико-химическим характеристикам и содержанию. Использован метод изучения распределения повторяющихся последовательностей ДНК в митохондриальных геномах. Рассчитано положение средней точки для каждой повторяющейся последовательности и присвоены каждой средней точке различные физико-химические характеристики. Затем сопоставлено количество повторов в средних точках со средними физико-химическими характеристиками, присвоенными каждой средней точке, чтобы проанализировать распределение повторов в митохондриальных геномах.

Большинство таксонов имеют в среднем 18-28 коротких несовершенных повторов на нуклеотид митохондриальной ДНК. Эти повторы имеют длину около 10-12 нуклеотидов, что эквивалентно одному неполному витку спирали ДНК. Это говорит о том, что митохондриальные повторы обогащены развернутыми структурами ДНК.

Исследование дало понимание что существует сильная корреляция между обилием прямых и зеркальных повторов и обилием инвертированных и комплиментарных повторов. Эту корреляцию можно объяснить сходством нуклеотидного состава этих пар.

Также выяснилось что большинство повторов, обнаруженных в выбранных мтДНК, являются несовершенными. Это говорит о том, что идеальные повторы могут быть результатом отрицательного отбора, что согласуется с предыдущими исследованиями. Другими словами, наличие совершенных повторов может быть невыгодным и, следовательно, менее распространенным в мтДНК.

Описанным алгоритмом было проанализировано 3715 геномов позвоночных. Анализируемые виды имеют следующий состав:7 классов, 139 отрядов, 659 семейств. В **Таблице 1** отображено среднее количество повторов в некоторых таксонах.

Обнаружено статистически значимая отрицательная корреляция между обилием повторов и содержанием GC, и эта корреляция была сильнее для инвертированных и комплементарных повторов. Это может быть связано с более сильным негативным отбором против богатых GC инвертированных и комплементарных повторов, но для подтверждения этого наблюдения необходим дальнейший анализ.

Таблица 1 – Среднее количество повторов в некоторых таксонах

	Прямые	**Комплементарные**	**Зеркальные**	**Инвертированные**
Chondrichthyes	6980	3882	2916	2061
Aves	7164	1976	2657	1121
Mammalia	6465	3157	2742	1779
(Hominidae)	6652	1783	2511	1125
ALL	6094	2880	2456	1586

Млекопитающие и хрящевые рыбы имеют намного большее количество всех видов повторов чем среднее число повторов по всей выборочной совокупности. При этом гоминиды имеют значительно большее количество прямых и симметричных повторов, но значительно меньшее количество комплементарных и инвертированных повторов относительно средних по всей выборочной совокупности. На **Рисунке 3** также отображены количества повторов по классам. Данные по минимальным и максимальным количествам повторов представлены в **Таблице 2 и Таблице 3**. Как видно, колоссальные количества повторов имеют Ascidiacea, а относительно малым количеством повторов обладают различные мелкие рыбы. В **Таблице 4** отображены данные по количеству повторов у гоминид. На **Рисунке 4** представлен двойной логарифмический график по количествам повторов. Как видно на графике – жесткая точка перегиба – это длина 21 нуклеотид. Это очень биологически осмысленно, так как все что более двух витков ДНК должно жестко отбираться, потому что все такие повторы образуют очень стабильные структуры. Иными словами полученные данные согласуется с биологией.

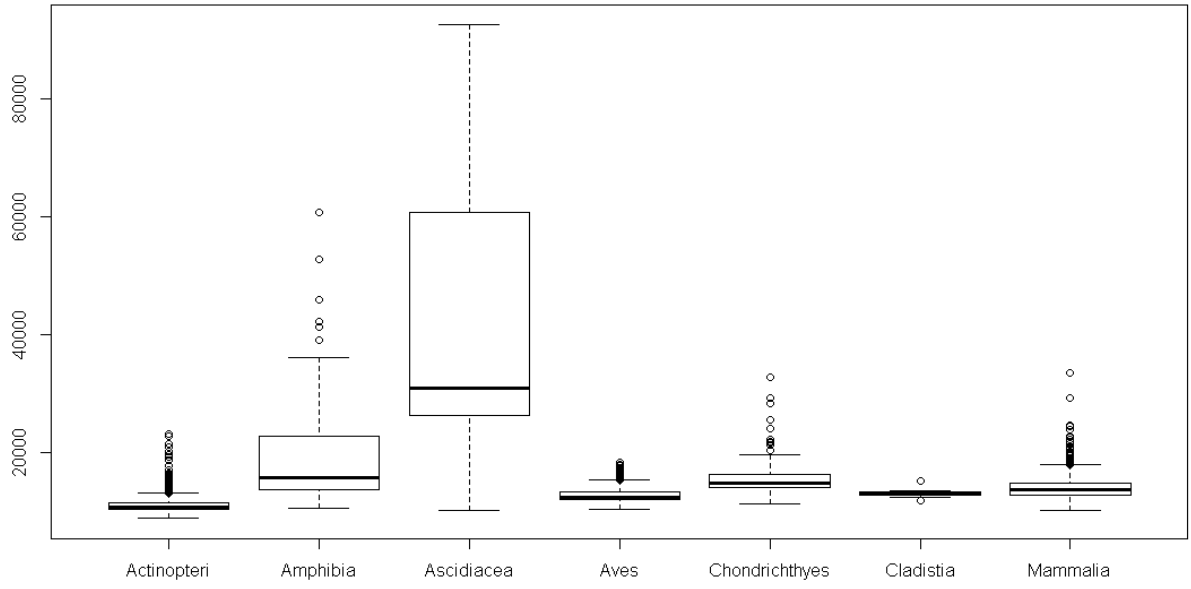

Рисунок 3 – Количества повторов по классам

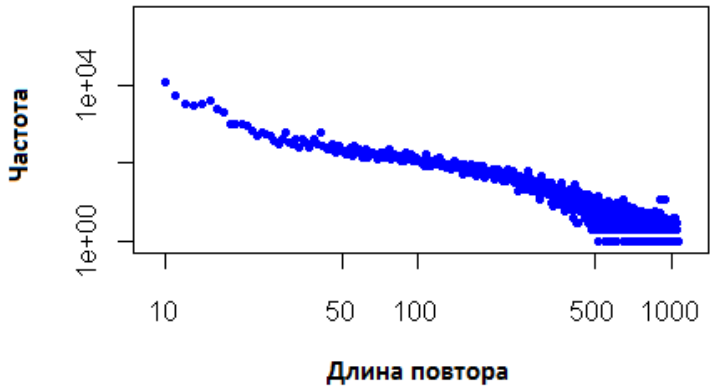

Рисунок 4 – Двойной логарифмический график по количествам повторов

Таблица 2 – Минимальные количества повторов

	Прямые	Комплементарные	Зеркальные	Инвертированные
Asymmetron_sp_A_TK 2007	3404	2700	1542	1404
Cyprinella_spiloptera	3477	2454	1453	1455
Meda_fulgida	3513	2576	1438	1443
Notropis_stramineus	3544	2527	1477	1433
Avocettina_infans	3545	2653	1489	1443

Таблица 3 – Максимальные количества повторов

	Прямые	Комплементарные	Зеркальные	Инвертированные
Didemnum_vexillum	64505	44181	35539	21982
Clavelina_lepadiformis	51755	11502	24946	6449
Clavelina_phlegraea	44751	19746	18189	9817
Aplidium_conicum	41625	15287	19587	6867
Ciona_intestinalis	32611	21544	16755	11672

Таблица 4. Количества повторов у гоминид

	Прямые	Комплементарные	Зеркальные	Инвертированные
Gorilla_gorilla	6256	1889	4825	2465
Gorilla_gorilla_gorilla	6335	1900	4867	2441
Pongo_pygmaeus	6990	1634	5113	1832
Pongo_abelii	7438	1616	5329	1947
Pan_troglodytes	6624	1842	5025	2228
Pan_paniscus	6831	1900	5084	2361
Macaca_cyclopis	6945	1908	5340	2154
Macaca_leonina	7613	1936	5857	2390
Macaca_arctoides	6720	1918	5200	2353
Macaca_fascicularis	6773	1930	5279	2243
Macaca_fuscata	6829	1871	5265	2243
Macaca_mulatta	6741	1821	5138	2215
Macaca_nemestrina	7344	1778	5552	2234
Macaca_sylvanus	7122	1893	5329	2251
Macaca_assamensis	6271	1886	4912	2243
Macaca_tonkeana	6856	1880	5343	2177
Macaca_nigra	6689	1897	5113	2270
Macaca_silenus	6764	1856	5294	2269
Macaca_thibetana	6668	1851	5074	2272
Homo_sapiens	6722	1804	5171	2299

База данных ImtRDB на данный момент содержит 4694 записей (митохондриальные геномы позвоночных), 3716 из них обработаны (аннотированы). Чтобы сравнить количество повторов между видами с разным размером генома, нормализовано количество повторов по

длине генома и получено количество повторов на нуклеотид. Это количество повторов на нуклеотид в семи таксонах позвоночных показано на **Рисунке 5**.

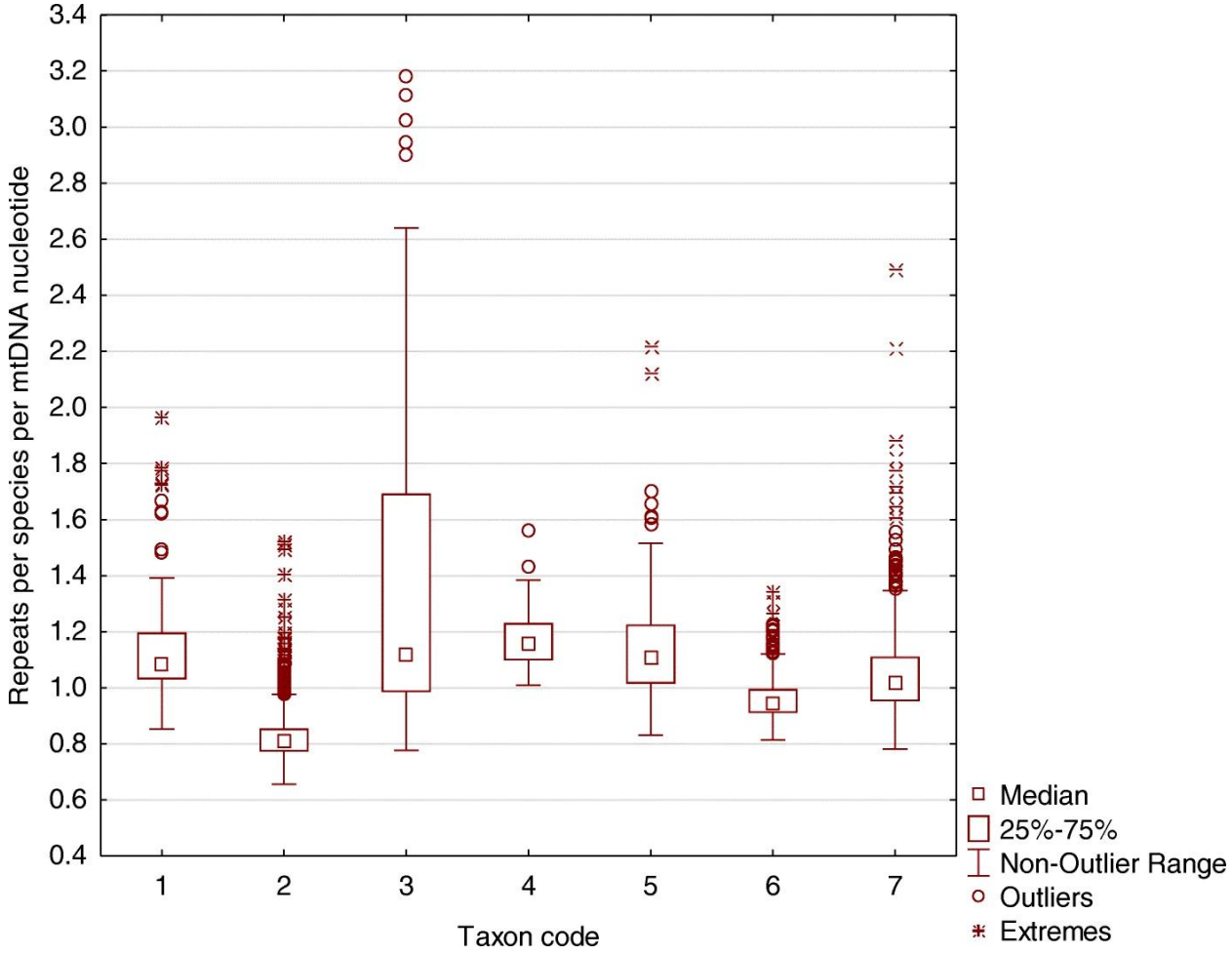

Рисунок 5. Число всех четырех типов повторов, нормированных по длине мтДНК каждого вида. Коды таксонов:

1 - Chondrichthyes; 2 - Actinopterygii; 3 - Amphibia; 4 - Testudines; 5 - Squamata; 6 - Aves; 7 - Mammalia

2.4.1 Обилие повторов в мтДНК видов позвоночных

Чтобы сравнить количество повторов между видами, нужно получить метрику, которая учитывает различия в размере генома, а также потенциальные различия в длине повторов. Чтобы сделать это, выведена средняя плотность повторов для каждого вида следующим образом: для каждого нуклеотида данного генома оценено количество перекрывающихся повторов и усреднено это число по всем нуклеотидам генома. Наконец, для каждого вида имеется метрика, представляющая количество повторов, перекрывающих средний нуклеотид (**Рисунок 6**).

2.4.2 Митохондриальные повторы обогащены развернутыми структурами ДНК

На **Рисунке 6** показано, что все таксоны, кроме Amphibia (более высокое число) и Actinopterygii (более низкое число), имеют в среднем 18-28 коротких несовершенных повторов на один нуклеотид мтДНК, длина таких повторов составляет в среднем 10-12 нуклеотидов, что эквивалентно одному неполному витку спирали ДНК в соответствии с данными, показанными на **Рисунке 6**. Интересно, что в среднем один повтор соответствует ~ 0,9 оборота ДНК. Этот факт указывает на то, что повторы могут концентрироваться в развернутых областях ДНК. Чтобы детально проанализировать этот вопрос по каждому анализируемому виду, соотносится число поворотов ДНК для средних точек повторов с числом повторов, перекрывающих эти средние точки повторов. Если существует заметная положительная корреляция между числом поворотов и плотностью повторов, то повторы предпочтительно располагаются в скрученных

областях или, альтернативно, повторы предпочтительно располагаются в развернутых областях. Полные результаты приведены в (Supplementary Tables 1 на сайте ImtRDB). Результаты показывают, что участки насыщенные повторами имеют тенденцию находиться в скрученных областях мтДНК, однако эти отношения не подтверждаются показательным значением Rho Спирмена (среднее значение Rho ~ 0,09). Следовательно, наиболее вероятно, что основная часть повторов мтДНК расположены не плотно друг другу (гетерогенно), и поэтому они находятся в развернутых структурах или, другими словами - большинство областей мтДНК, содержащих повторы, редко имеют скрученную (суперсплетенную) форму ДНК.

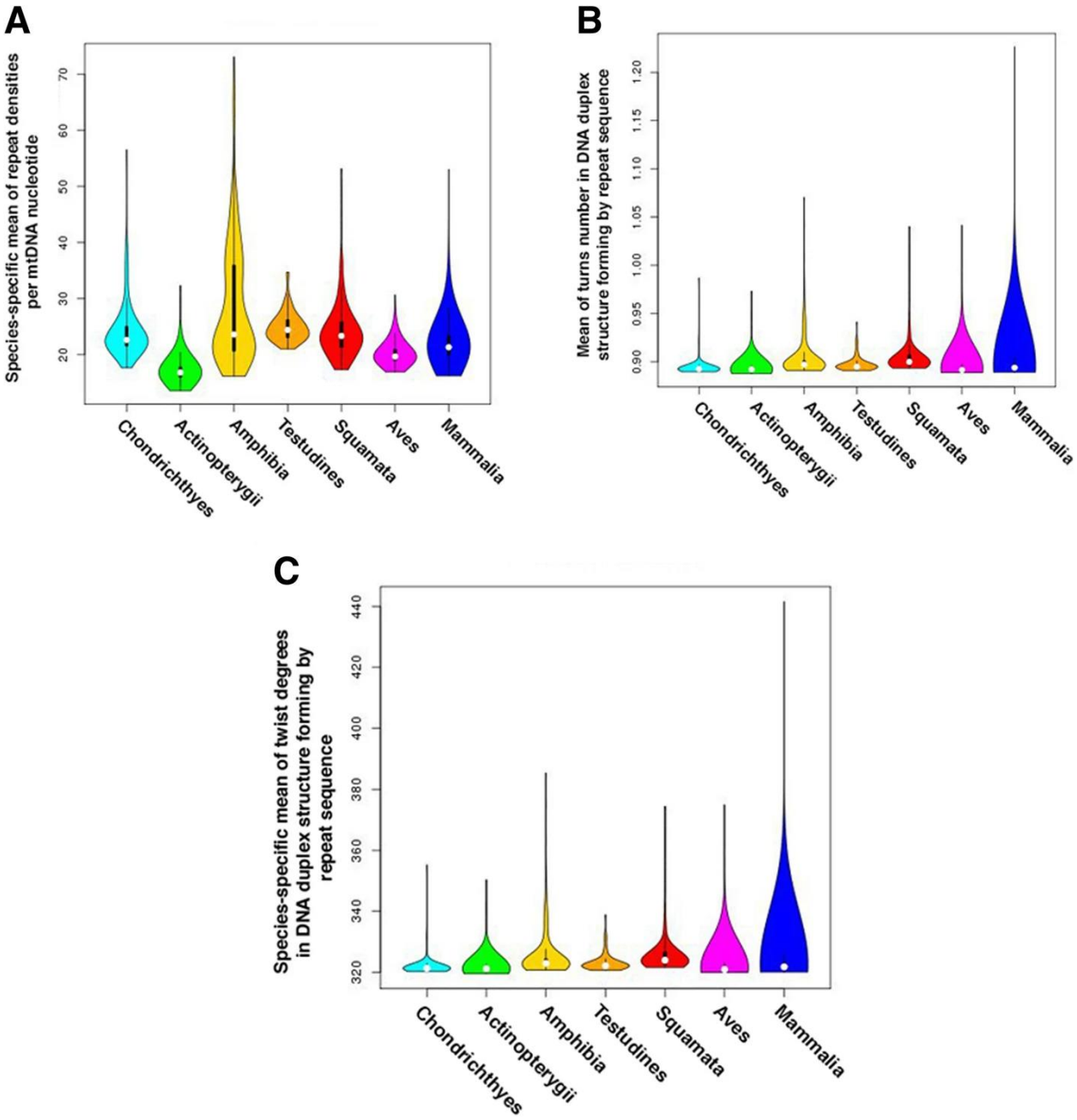

Рисунок 6. Плотность повторов, характерная для таксонов, и характеристики длины повторов.
A – видоспецифическое среднее значение плотностей повторов на нуклеотид мтДНК;
B – видоспецифическое среднее число витков в структуре дуплекса b-днк, образующейся повторяющейся последовательностью; C – видоспецифическое среднее значение степеней изгиба в структуре дуплексной днк, формируемой повторяющейся последовательностью.

2.4.3 Все типы повторов положительно коррелируют друг с другом, но эквивалентные повторы коррелируют сильнее

Проверено, коррелирует ли обилие различных типов повторов друг с другом. Продемонстрированы статистически значимые положительные попарные корреляции между всеми типами повторов (**Таблица 5**). Наиболее сильные корреляции, которые наблюдаются - между прямыми и зеркальными повторами, а также между перевернутыми и комплиментарными повторами. Эти пары повторов имеют общие черты: (1) общий нуклеотидный контекст; (2) общее местоположение (одна и та же нить ДНК: прямые и зеркальные, противоположные нити ДНК: инвертированные и комплиментарные - см. **Рисунок 1**); (3) их короткая длина (**Рисунок 6**). Благодаря общей природе этих пар повторов можно использовать их сходство в качестве важной нулевой гипотезы, утверждая, что при прочих равных условиях (одинаковый показатель происхождения и одинаковый положительный или отрицательный отбор) ожидается одинаковое количество эквивалентных повторов на геном. Любые отклонения от этого равновесия должны быть биологически информативными и указывать на различную силу мутагенеза или отбора.

Таблица 5. Парная корреляция содержания несовершенных повторов и корреляции с содержанием GC, все 3716 проанализированных видов, Rho Спирмена выше диагонали, значения p ниже диагонали.

	Содержание GC, п.н.	Прямые повторы, п.н.	Комплиментарные повторы, п.н.	Зеркальные повторы, п.н.	Инвертированные повторы, п.н.
GC-контент		-0.2864	**-0.6965**	-0.3415	**-0.6685**
Прямые повторы	< 2.2e-16		0.1247	**0.9838**	0.0919
Дополнительные повторы	< 2.2e-16	2.376e-14		0.1869	**0.9785**
Зеркало повторяется	< 2.2e-16	< 2.2e-16	< 2.2e-16		0.1426
Инвертированные повторы	< 2.2e-16	1.971e-08	< 2.2e-16	< 2.2e-16	

2.4.4 Все типы повторов отрицательно коррелируют с GC-составом, но инвертированные и комплементарные повторы коррелируют сильнее

Проверено, соотносится ли количество различных типов повторов с содержанием GC. Наблюдались статистически значимые отрицательные корреляции между количеством повторов и их GC-составом (**Таблица 6**). Интересно, что отрицательная корреляция была значительно сильнее для инвертированных и комплементарных повторов (**Таблица 6**). Это может быть объяснено более сильным отрицательным отбором против инвертированных и комплементарных повторов, богатых GC, однако необходимы дополнительные анализы, чтобы пролить свет на это наблюдение.

Далее, принимая во внимание потенциально важную роль содержания нуклеотидов в эволюции генома мтДНК [96], проверено видоспецифическое содержание GC в идентифицированных повторах и их относительные физико-химические особенности (**Рисунок 7**). Графики особенностей повторов по таксонам (**Рисунок 7**) демонстрируют, что все таксоны имеют свое конкретное оптимальное содержание GC в повторах, например, Actinopterygii и Aves имеют максимальное содержание GC, в то время как млекопитающие имеют минимальное содержание. Оптимальное содержание GC, специфичное для таксонов, напрямую влияет на изменение свободной энергии Гиббса (dG) и температуры плавления (Tm) спаренных оснований в областях повторов (**Рисунок 7**). Интересно, что энергия укладки спаренных оснований повторов и энтропии/энтальпии парных оснований (dC/dG) значительно варьирует в кладе млекопитающих и Actinopterygii/Amphibia соответственно.

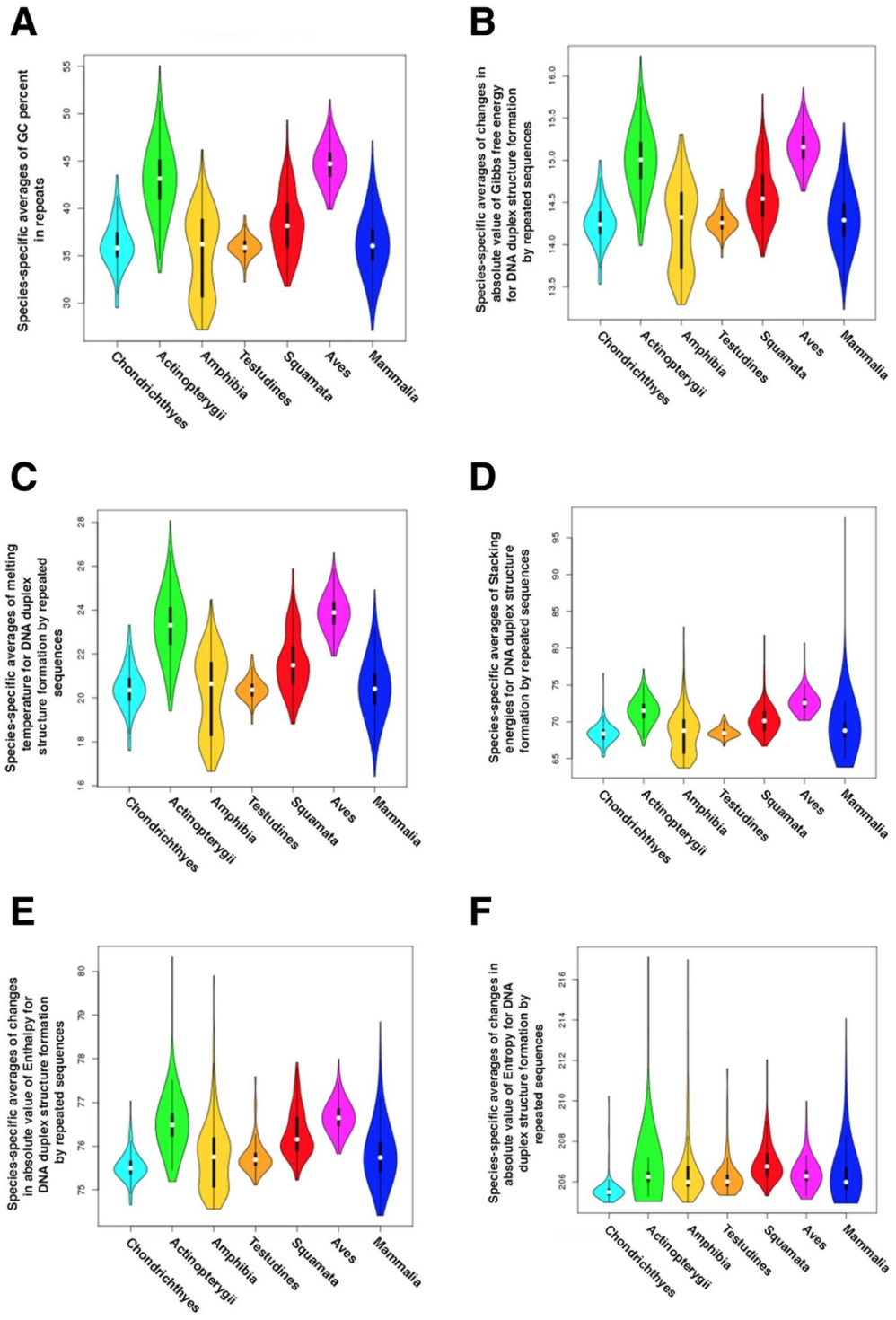

Рисунок 7. Особенности повторов, характерные для таксонов, связанные с содержанием GC. A – видоспецифичные средние значения процента gc в повторах; B – видоспецифичные средние изменения абсолютной величины свободной энергии гиббса для формирования дуплексной структуры днк с помощью повторяющихся последовательностей; C – видоспецифичные средние температуры плавления для формирования дуплексной структуры днк из повторяющихся последовательностей; D – видоспецифичные средние энергии суммирования для формирования дуплексной структуры днк с помощью повторяющихся последовательностей; E – видоспецифичные средние изменения абсолютного значения энтальпии для формирования дуплексной структуры днк с помощью повторяющихся последовательностей; F – видоспецифичные средние изменения абсолютного значения энтропии для формирования дуплексной структуры днк с помощью повторяющихся последовательностей.

Затем выяснено, равны ли частоты комплементарных типов динуклеотидов в несовершенных повторах, расположенных в мтДНК позвоночных (**Таблица 6**). Этот вопрос очень важен для идентификации возможных сил отбора, действующих на повторяющиеся последовательности. Из-за компактного расположения геномов с ненулевыми частотами динуклеотидов, содержащимися в повторах (дополнительные таблицы 2 на сайте ImtRDB), решить вопрос удалось, используя динуклеотидные частоты всего генома, однако пользователь может выполнять анализ частот динуклеотидов по отдельным регионам Сравнены пары комплементарных динуклеотидов, наблюдаемых в повторах в каждом анализируемом геноме, с помощью U-критерия для оценки статистической значимости и d-значения Коэна для оценки величины эффекта. В **Таблице 6** приведены средние соотношения, величины эффекта и уровень статистической значимости для пар комплементарных динуклеотидов для всех проанализированных таксонов (получены путем усреднения видоспецифических данных). Сравнение комплементарных пар динуклеотидов выявило, что наибольший размер эффекта наблюдался для пар CA/GT, CC/GG и AC/TG (**Таблица 6**). Все эти соотношения могут быть объяснены положительным коэффициентом асимметрии AT (превышение A по сравнению с T) и отрицательным коэффициентом асимметрии GC (дефицит G по сравнению с C) на легкой цепи (цепь, которая обычно хранится в генбанке) мтДНК, что согласуется с механизмом репликации мтДНК [107, 108]. Действительно, можно увидеть, что во всех этих соотношениях (которые построены таким образом, чтобы это соотношение было больше 1) G и T присутствовали только в знаменателях, а C и A - только в числителях. Симметричные пары (с идентичным набором нуклеотидов), такие как TA/AT или GC/CG, не подвержены перекосу нуклеотидов и, следовательно, их отношения должны быть близки к единице. Действительно, соотношение TA/AT не отличается от единицы, но, что интересно, соотношение GC/CG значительно выше, чем единица во всех таксонах. Причину превышения GC над CG-динуклеотидами в легкой цепи повторов мтДНК всех анализируемых видов стоит выяснить в будущих анализах. Частота GC в два раза выше, чем частота CG (**Таблица 6**). Этот факт является неожиданным. Причиной этой проблемы может быть в специфической регуляции экспрессии гена мтДНК в направлении ДНК / направлении цепи (в случае, если участки мтДНК с высокой плотностью повторов могут быть вовлечены в регуляцию экспрессии гена мтДНК).

Также судя по всему, похоже что повторы необходимы для функционирования митохондриальной ДНК, так как чем больше повторов на сайт в геноме тем: (1) больше температура плавления; (2) число нуклеотидов в витке повтора; (3) выше энергия взаимодействия в спаренном участке.

Таблица 6. Попарное сравнение комплементарных пар динуклеотидов в повторах мтДНК.

Соотношение частот комплементарных динуклеотидов с размером эффекта *	Chondrichthyes	Actinopteri	Amphibia	Testudines	Squamata	Aves	Mammalia
CA/GT	3.1395	2.6701	2.8525	4.9014	4.9034	4.8671	3.6059
Coohen d	**0.1601**	**0.1419**	**0.1475**	**0.234**	**0.2351**	**0.2334**	**0.188**
CC/GG	3.7044	3.0336	2.9595	3.9861	4.0513	5.3133	3.7551
Coohen d	**0.1656**	*0.1696*	**0.1283**	**0.1687**	**0.1789**	**0.2548**	**0.1588**
AC/TG	2.3030	1.9380	2.0039	3.7421	3.6465	3.4388	2.8990
Coohen d	**0.1282**	*0.1078*	**0.1086**	**0.2137**	**0.2126**	**0.1981**	**0.1657**
CT/GA	2.2609	2.0033	2	2.1529	2.3051	2.5639	2.2432
Coohen d	**0.1414**	*0.1215*	**0.1115**	**0.1232**	**0.1302**	**0.1668**	**0.1337**
AA/TT	1.1184	1.1642	1.0196	1.7303	1.6796	1.9638	1.4414
Coohen d	0.0287	0.0326	0.0055	**0.1298**	0.118	*0.1258*	0.0857
TC/AG	1.7016	1.3023	1.4356	1.4441	1.4948	1.6962	1.5327
Coohen d	**0.0901**	0.0437	0.0551	**0.0568**	*0.0601*	**0.0944**	**0.0705**
GC/CG	1.92	1.7840	1.9459	1.9579	1.7984	1.7919	1.8598
Coohen d	**0.0494**	**0.0614**	**0.0544**	**0.049**	**0.0505**	**0.0581**	**0.0493**
TA/AT	1.0345	1.1097	1.0316	1.0741	1.0871	1.0926	1.0635
Coohen d	0.0094	0.022	0.0086	0.0201	0.0208	0.0186	0.0174

* жирный шрифт - n<1E-10, курсив - 1E-5<n<1E-10

2.5 Выводы

Так как не было обнаружено удовлетворяющих целям исследования баз данных повторов и алгоритмов, способных к поиску всех типов несовершенных повторов в кольцевой мтДНК, это привело к необходимости разработать свой собственный алгоритм и собственную единую базу данных, хранящую эти повторы мтДНК для всех видов хордовых с секвенированной полной мтДНК.

Мной был реализован алгоритм позволяющий находить как совершенные, так и несовершенные повторы четырех основных типов: прямой, инвертированный, зеркальный и комплиментарный. Разработанный мной алгоритм поиска вырожденных повторов написан на языке программирования Python и основан на распознавании паттернов по аналогии со стандартными процедурами построения дот-матрицы. Мой алгоритм адаптирован к специфическим характеристикам мтДНК, таким как кольце-образность и избыток коротких повторов. Организована интерактивная базу данных с онлайн-доступом содержащая позиции совершенных и несовершенные повторов в мтДНК более 3500 видов позвоночных. Дополнительные инструменты, такие как визуализация повторов в геноме, сравнение плотности повторов среди разных геномов и возможность скачать все результаты делают эту базу данных полезной для многих биологов. Разработанный мной алгоритм и база данных являются инструментами в которых учтены те аспекты анализа повторов в митохондриальных геномах которым в других инструментах либо не было уделено внимание, либо оно было уделено не в полной мере. Данные инструменты имеют широкую область применения, от анализа геномов в научных целях до интеграции в системы персонализированной медицины. База данных и разработанный алгоритм могут быть полезным инструментом в различных областях исследования ДНК митохондрий и хлоропластов.

Первые анализы базы данных продемонстрировали, что несовершенные повторы мтДНК (i) обычно короткие; (ii) связаны с развернутыми структурами ДНК; (iii) четыре типа повторов положительно коррелируют друг с другом, образуя две эквивалентные пары: прямые и зеркальные против инвертированных и комплементарных, с идентичным содержанием нуклеотидов и сходным распределением между видами; (iv) обилие повторов отрицательно связано с GC-составом; (v) динуклеотиды GC против CG перепредставлены на легкой цепи мтДНК покрытой повторами.

Используя базу данных, продемонстрирована сильная положительная корреляция между количеством прямых и зеркальных повторов, а также между инвертированными и комплементарными повторами. Можно отметить, что эти пары (прямая и зеркальная; инвертированная и комплементарная) имеют идентичный нуклеотидный состав плеча повтора (см. **Рисунок 1**: первое плечо прямого повтора имеет два «A», пять «T», два «G» и два «C»; одинаковое содержание на одной и той же цепи будет наблюдаться как в случае второго плеча прямого повтора, так и во втором плече зеркального повтора) и, таким образом, их можно рассматривать как эквивалентные повторы, т. е. если предположить одинаковую скорость происхождения (мутагенеза), а также одинаковую скорость распада (коэффициенты отбора), то ожидается одинаковое количество эквивалентных повторов. Можно считать, что равенство эквивалентных повторов является полезной нулевой гипотезой, которую можно проверить в будущем. Кроме того, подтверден дефицит нуклеотидов C и G в геномах, богатых повторами, и продемонстрировано ранее неизвестное превышение GC над динуклеотидами CG в легкой цепи мтДНК, покрытой повторами. Также подтверждено, что повторы мтДНК обычно короткие и связаны с развернутыми структурами ДНК. Полученная база данных, а также первичные наблюдения будут способствовать будущим открытиям функциональной роли повторов мтДНК.

Стоит заметить что поскольку исходные данные не соответствуют нормальному распределению, использование параметрических методов привело бы к некорректным

значениям. Поэтому для анализа был выбран непараметрический коэффициент корреляции Rho Спирмена (**Таблица 5**). Несмотря на его устойчивость к нарушению нормальности, этот метод не чувствителен к нелинейным или мультимодальным закономерностям. Учитывая недостаток данных о взаимодействии повторов и руководствуясь принципом парсимонии, было принято решение ограничиться этим методом для выявления базовых взаимосвязей. Для более полной картины была рассчитана величина эффекта (d Коэна, **Таблица 6**), которая показала незначительные размеры эффекта, что свидетельствует об отсутствии выраженных различий между группами.

Замечено, что несовершенные повторы митохондриальной ДНК обычно короткие, часто встречаются и обогащены релаксированными структурами ДНК.

Обнаружены сильные отрицательные корреляции между количеством повторов и содержанием в них GC. Это можно объяснить отрицательным отбором в отношении богатых GC повторов, который, вероятно, более выражен в случае инвертированных и комплементарных повторов по сравнению с прямыми и зеркальными. Это соответствует общепринятой точке зрения, согласно которой потенциально вредный эффект повторов зависит как от длины повтора, так и от содержания GC в повторе.

Также наблюдается, что распределение большинства комплементарных динуклеотидов в легкой цепи повторяющихся участков мтДНК определяется положительным перекосом AT и отрицательным GC, однако сильное и равномерное превышение GC над динуклеотидами CG не может быть объяснено перекосом и, следовательно, должен быть исследован дополнительно.

Созданная база данных позволяет более подробно и точно отвечать на вопросы, касающиеся расположения повторов, а также взаимодействия между различными типами повторов и взаимодействия между количеством повторов и их физико-химическими свойствами.

Учитывая высокую плотность повторов, имеет смысл также оценить долю нуклеотидов мтДНК, которые не пересекаются с повторяющимися последовательностями, и отследить изменение значения этой метрики при использовании разных пороговых значений на минимальную длину найденных повторов.

Глава 3. Влияние вторичной структуры митохондриального генома человека на образование делеций

3.1 Проблематика

Хотя уже давно известно, что прямые нуклеотидные повторы (или длинные несовершенные дуплексы) способствуют образованию делеций мтДНК, они до сих пор объясняют лишь небольшую часть наблюдаемого распределения делеций. Это поднимает вопросы о том, почему некоторые повторы приводят к делециям, а другие — нет, и какие еще факторы могут участвовать в формировании делеций. Чтобы понять основные факторы формирования делеции мтДНК, разумно начать с анализа наиболее мутагенного прямого повтора в геноме человека: общего повтора [21, 22, 26], который является самым длинным совершенным прямым повтором в мтДНК человека. Важной особенностью общего повтора является то, что его «плечи» (первое плечо на 8470–8482 п.н. и второе плечо на 13447–13459 п.н.) расположены точно на пиках распределения всех точек разрыва делеции по мтДНК. На основании этого наблюдения была выдвинута гипотеза, что общий повтор «кажется, является основным фактором, ответственным за образование большинства делеций» [21]. Это означает, что общий повтор может быть важен не только для формирования общей делеции, но и играть роль в возникновении всех других делеций мтДНК. Чтобы проверить эту гипотезу, проанализирован спектр делеции мтДНК в лобной коре образцов гаплогруппы N1b1, где повтор был нарушен (проксимальное плечо было acTtccctcacca против acCtccctcacca, как у подавляющего большинства человеческой популяции). Если бы существовала особая структурная роль общего повтора, ожидалось что распределение всех делеций мтДНК в образцах N1b1 отличалось бы от других гаплогрупп с идеальным общим повтором. В пределах имеющегося размера выборки (два случая и два контроля) наблюдалось почти полное отсутствие общей делеции как таковой в образцах N1b1; однако не обнаружено никаких изменений в распределении других делеций [22]. Таким образом, отвергнута гипотеза о том, что общий повтор является основным фактором образования большинства делеций [21]. Отвержение этой гипотезы оставило необъяснимым главное наблюдение, подчеркнутое Сэмюэлсом и соавторами, а именно: почему распределение делеций внутри большой дуги сильно неравномерно? Эта неравномерность распределения делеций требует нового объяснения.

Здесь, проводя параллели между делециями в бактериях [33], общим повтором мтДНК [34] и ядерной ДНК [35], можно предположить, что прямые повторы могут с большей вероятностью вызывать делеции, когда они находятся в непосредственной близости друг от друга. Таким образом, повышенная вероятность появления делеций вблизи общего повтора поддерживается не самим общим повтором, а независимым топологическим фактором. Чтобы проверить эту гипотезу, принято решение реконструировать потенциальную пространственную структуру одноцепочечной большой дуги мтДНК и сделано предположение, что она организована как крупномасштабная шпилькообразная петля (**Рисунок 8**). Предположительно именно такая форма мтДНК (в виде символа бесконечности) может влиять на мутагенный потенциал прямых повторов во время репликации и, таким образом, формирует распределение делеций.

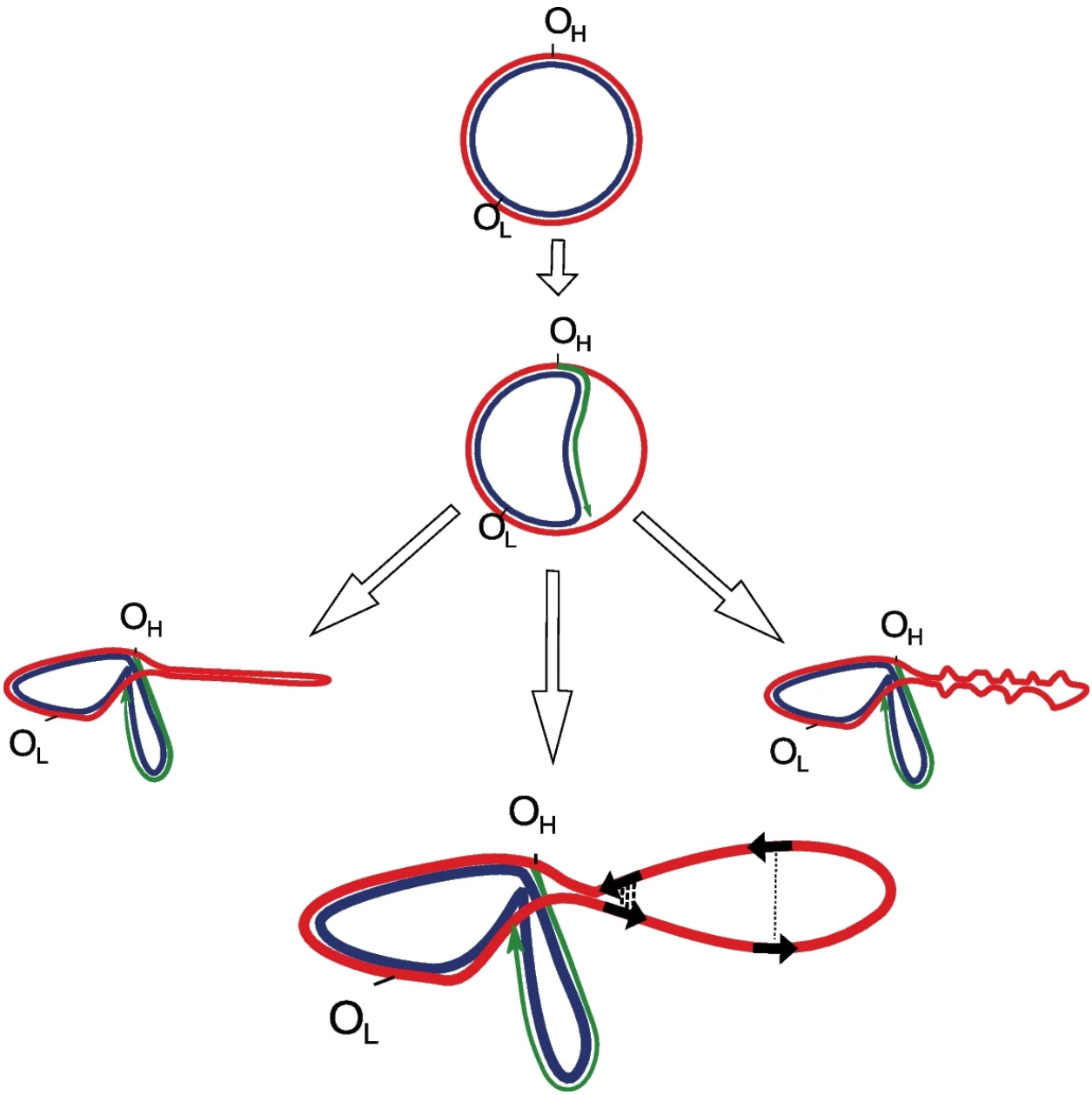

Рисунок 8. Потенциальные вторичные структуры, образованные одноцепочечной родительской тяжелой цепью во время репликации мтДНК. Нижняя панель показывает, что прямые повторы, отмеченные черными стрелками, имеют разные шансы реализоваться в делеции в зависимости от пространственной структуры. Тесная пространственная близость повторов (жирные пунктирные линии) увеличивает вероятность образования делеций, тогда как для повторов, которые пространственно разделены большим расстоянием, эта вероятность уменьшается (тонкая пунктирная линия)

3.2 Методы

Весь код и данные, сгенерированные или проанализированные в ходе этого исследования, включены в общедоступный репозиторий (https://mitoclub.github.io/GlobalStructure/). Релизная версия 1 (https://github.com/mitoclub/GlobalStructure/releases/tag/v.1) был депонирован на Figshare (DOI: 10.6084/m9.figshare.22559710).

Распределение центров: Для каждой делеции из MitoBreak в большой дуге (5781-16569) была найдена ее середина. Далее каждая из реальных делеций произвольно перемещалась в пределах большой дуги, и также были получены их средние точки. Для наблюдаемых средних наблюдаемых делеций и случайно смоделированных делеций были получены и сопоставлены соответствующие стандартные отклонения.

Матрица контактов мтДНК Hi-C: Общедоступная матрица мтДНК была визуализирована с помощью Juicebox [49]. Методика получения Hi-C данных из линии лимфобластоидных клеток человека описана в Rao et al. [48]. Кроме того, получилено шесть контактных матриц Hi-C мтДНК из обонятельных рецепторов пациентов с COVID и контрольной группы. Подробная информация о протоколе Hi-C in situ, а также биоинформатический анализ описаны в оригинальной статье [109]. Матрицы были визуализированы с помощью Juicebox [49].

Фолдинг in silico: Использована тяжелая цепь эталонной последовательности мтДНК человека (NC_012920.1), поскольку она проводит большую часть времени в одноцепочечном состоянии в соответствии с асимметричной моделью репликации мтДНК [34]. Используя Mfold [44] с параметрами, установленными для сворачивания ДНК и кольцевой последовательности, было ограничено все, кроме главной дуги, от образования пар оснований. Получена глобальная (по всему геному) вторичная структура, которую затем перевели в количество водородных связей, соединяющих интересующие области (окна 100 п.н. для анализа и визуализации).

Затем внутри одноцепочечной тяжелой цепи основной дуги определилены окна размером 100 п.н. и гибридизованы все потенциальные пары таких окон с использованием пакета ViennaRna Package 2 [45]. Полученные энергии Гиббса для каждой пары таких окон использовались как метрика силы потенциального взаимодействия между двумя одноцепочечными участками ДНК.

Плотность инвертированных/прямых повторов: Для каждой пары окон размером 100 п.н. оценено количество нуклеотидов, участвующих как минимум в одном инвертированном/прямом деградированном повторе. Соответствующий повтор должен иметь одно плечо, расположенную в первом окне, и другое плечо, расположенную во втором окне. Все деградированные (с максимальным уровнем несовершенства 80%) повторы в мтДНК человека были найдены с использованием разработанного мной алгоритма, описанного ранее [110].

Кластеризация делеций: Для кластеризации использованы все MitoBreak [46] делеции из большой дуги. Использованы координаты 5’ и 3’ в качестве входных данных для алгоритма кластеризации на основе иерархической плотности (python hdbscan v0.8.24). DBSCAN — это хорошо известный алгоритм кластеризации на основе плотности вероятности, который обнаруживает кластеры как области с более плотно расположенными выборочными данными, а также выборки с выбросами. Преимуществом этого метода является мягкая кластеризация. Варьировались параметры плотности кластеров, чтобы обеспечить стабильность кластеров, и было обнаружено, что кластерные образования остаются относительно стабильными для широкого диапазона параметров. Таким образом, DBSCAN создает надежный набор кластеров, предоставляя дополнительные доказательства для областей с повышенным уровнем делеций. Также выполнена кластеризация с аффинным распространением [111] в качестве эксперимента по исследованию данных, который также дает надежную кластеризацию.

*Совершенные прямые повторы мтДНК человека:*список идеальных прямых повторов длиной 10 и более пар оснований был использован из алгоритма, описанного в Guo et al. [22].

Реализованные и нереализованные прямые деградированные повторы: Использована база данных деградированных повторов мтДНК полученная с использованием разработанного алгоритма [110] длиной 10 п.н. и более и сходством 80% или больше. Учитывались только прямые повторы с расположением обеих плеч в большой дуге. Сгруппированы повторы с похожими мотивами в кластеры так, чтобы каждый рассматриваемый кластер содержал не менее трех ветвей повтора, а с двумя из них была связана хотя бы одна делеция. Дополнительно ограничено нужное подмножество кластеров, принимались к рассмотрению только нереализованные повторы как пары плеч, где хотя бы одно из них (первое или второе) совпадает с реализованным повтором. Визуально на **Рисунке 9B** это означает, что внутри каждого кластера сравниваются реализованные повторы (красная точка) с нереализованными (серая точка), расположенными на одной и той же горизонтальной (одна и та же координата Y) или вертикальной (одна и та же координата X) оси. Получилось 618 таких кластеров.

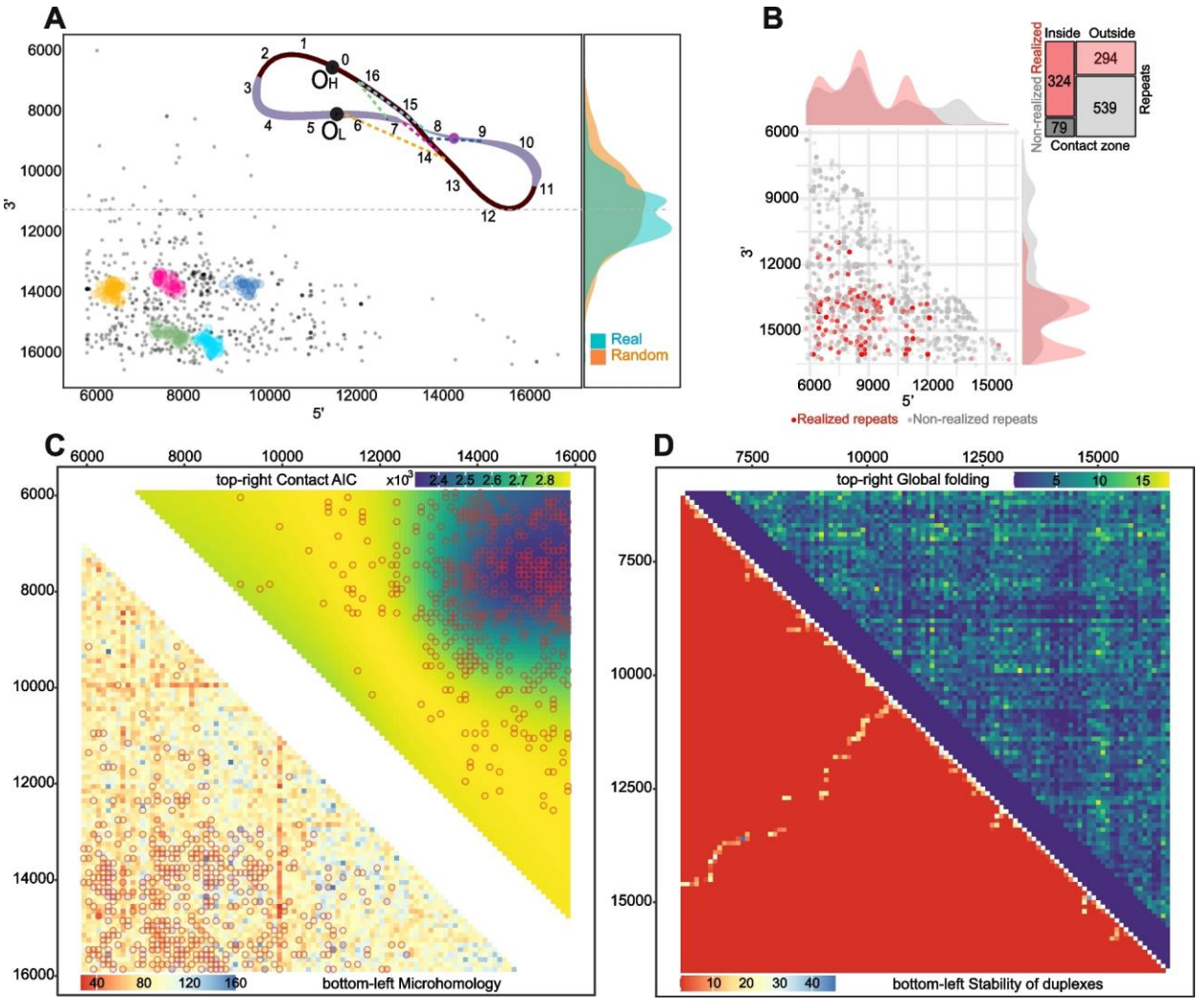

Рисунок 9. Вторичная структура мтДНК. A Кластеры делеций внутри большой дуги. Большинство кластеров расположены близко друг к другу в пределах потенциальной зоны контакта. Цвета на схеме сверху соответствуют кластерам. Вертикальные графики плотности в правой части рисунка демонстрируют распределение центров делеций: реальные (наблюдаемые) и случайные (ожидаемые). B Реализованные (красные) и нереализованные (серые) повторы, как правило, обогащаются в потенциальной зоне контакта. Мозаичный график повторов (реализованные и нереализованные) внутри и снаружи потенциальной зоны контакта. C Внизу слева: тепловая карта микрогомологий между окнами 100 п.н. в пределах большой дуги. Одна только микрогомология плохо объясняет распределение делеций (пустые кружки). Вверху справа: тепловая карта управляемой данными зоны контакта, основанная на AIC сравниваемых моделей. D Подход in silico для глобального прогнозирования складок большой дуги. Нижний левый треугольник: контактная матрица, полученная в результате оценки in silico

сворачивания всей одноцепочечной тяжелой цепи основной дуги; верхний правый треугольник: контактная матрица, полученная в результате оценки in silico сворачивания окон размером 100 п.н. одноцепочечной тяжелой цепи основной дуги

Парные выравнивания для матрицы микрогомологии: Мера степени сходства между сегментами большой дуги была получена путем выравнивания небольших окон последовательности митохондриальной большой дуги друг с другом. Была нарезана последовательность главной дуги митохондрий на 100 нуклеотидных фрагментов и выровняли их друг относительно друга с помощью EMBOSS Needle [112] с параметрами по умолчанию (совпадение +5, открытие гэпа - 10, расширение гэпа - 0.5), были проанализированы баллы выравнивания, получая таким образом данные для матрицы микрогомологии.

3.3 Результаты

Анализируя делеции мтДНК человека в большой дуге мтДНК, которая является одноцепочечной во время репликации и характеризуется большим количеством делеций, обнаружено неравномерное распределение с «горячей точкой», где одна точка делеционного разрыва находится в пределах области. 6-9кб и еще одна в пределах 13-16кб от мтДНК. Это распределение не было объяснено наличием прямых повторов, что позволяет предположить, что причиной могут быть другие факторы, такие как пространственная близость этих двух областей. Анализ in silico показал, что одноцепочечная главная дуга может быть организована в виде крупномасштабной петли, похожей на шпильку, с центром, близким к 11 т.п.н., и контактными областями между 6-9 т.п.н. и 13-16 т.п.н., что объясняет высокую делеционную активность в этой зоне контакта. Прямые повторы, расположенные в зоне контакта, такие как хорошо известный общий повтор с первым плечом 8470-8482 п.н. (пара оснований) и вторым плечом 13447-13459 п.н., в три раза чаще вызывают делеции по сравнению с прямыми повторами, расположенными вне контактной зоны. Сравнение делеций, связанных с возрастом и заболеванием, показало, что контактная зона играет решающую роль в объяснении возрастных делеций, подчеркивая ее важность в скорости здорового старения.

3.3.1 Спектр делеций неоднороден и плохо объясняется прямыми повторами.

Если образование делеций зависит от пространственной близости одноцепочечных участков ДНК, ожидается, что распределение делеций будет неравномерным и будет повторять структуру ДНК (**Рисунок 8**). Чтобы понять потенциальную структуру одноцепочечных участков ДНК, проанализировано распределение делеций внутри основной дуги мтДНК человека, где происходит большинство делеций. Для этого использованы данные базы данных MitoBreak [46], содержащей коллекцию делеций мтДНК человека [46]. Исследовано распределение центров каждой делеции внутри большой дуги. Обнаружено, что медианный центр располагался на расстоянии 11 463 п.н. (**Рисунок 9А**, правая вертикальная панель, N = 1060), а распределение было относительно узким, что указывает на низкую изменчивость положения центров и их тенденцию к кластеризации. Для подтверждения этого сравнено наблюдаемое изменение положения центров со случайно сгенерированными (см. раздел «Методы» выше). Анализ показал, что наблюдаемое изменение действительно было значительно ниже ожидаемого (значение $p < 0{,}0001$). Наблюдение, что большинство кластеров делеций расположены около 11 463 п.н., позволяет предположить, что одноцепочечная большая дуга может быть свернута в шпилькоподобную структуру с осью сгиба около 11 463 п.н. (**Рисунок 8**).

Точки делеционного разрыва: ожидается, что координаты 3' и 5' будут более распространены в регионах, пространственно близких друг к другу. Чтобы выявить возможную неравномерность в распределении точек разрыва, сгруппированы отдельные делеции в

кластеры (**Рисунок 9А**. Кластеры представлены цветными точками, см. раздел «Методы» выше). Замечено, что самые большие кластеры расположены близко друг к другу в определенной области, с точками разрыва 5' между 6–9 т.п.н. и точками разрыва 3' между 13–16 т.п.н. Это позволяет предположить, что одноцепочечная большая дуга образует стебель, в котором 6–9 и 13–16 т.п.н. пространственно близки друг к другу; дефицит точек разрыва за пределами этой области (9–13 т.п.н.) предполагает, что этот участок одноцепочечной большой дуги может сохраняться как разомкнутая петля. Важно отметить, что приблизительный центр этой петли (11 т.п.н.) соответствует предсказанной оси сворачивания (11,463 п.н.) на основе анализа центров (**Рисунок 9А**. Правая вертикальная панель; см. также схему мтДНК в правой верхней части **Рисунок 9А**).

Во время асинхронной репликации мтДНК родительская тяжелая цепь большой дуги постепенно становится одноцепочечной. Область, ближайшая к началу репликации тяжелой цепи (с примерными координатами 16.5 т.п.н.), является областью ранней репликации, которая сначала становится одноцепочечной, а по мере движения репликационной вилки вся большая дуга становится одноцепочечной. причем последний участок близок к началу репликации легкой цепи (приблизительные координаты 6 т.п.н.).

Если для образования делеции важно время нахождения в одноцепочечном состоянии, как это предполагается для мутагенеза однонуклеотидных замен [113, 114], то можно предположить, что ранне-реплицирующаяся область (~16,5 т.п.н.) может быть более мутагенной по сравнению с поздно-реплицирующейся областью (~ 6 т.п.н.). Повышенная делеционная мутагенность раннереплицирующихся участков (16.5 т.п.н., что соответствует 3'-точке разрыва) по сравнению с позднореплицирующимися участками (6 т.п.н., что соответствует 5'-точке разрыва) может быть реализована в том, что ранне-реплицирующаяся область менее избирательна: эта область может ассоциироваться с любыми другими открытыми областями основной дуги, что означает повышенную вариабельность точки разрыва 5'-делеции по сравнению с точкой разрыва 3'-делеции. Анализ диаграммы рассеяния точек делеционного разрыва и кластеров (**Рисунок 9А**) подтверждает это, показывая, что цветные кластеры лучше описываются как овал с увеличенной длиной вдоль оси X, что указывает на повышенную вариабельность точек разрыва 5'. В целом, повышенная вариабельность точек разрыва 5' по сравнению с точками разрыва 3' (**Рисунок 9А**) позволяет предположить, что на образование делеций также может влиять продолжительность времени, в течение которого цепь остается одноцепочечной: этот показатель выше для 3'-точек разрыва, что позволяет им ассоциироваться с более широким диапазоном 3'-точек разрыва.

Все приведенные выше результаты позволяют предположить, что одноцепочечная большая дуга может сворачиваться в большую петлю с центром, близким к 11,5 т.п.н., и стеблем, образованным областями 6–9 и 13–16 т.п.н., где рано реплицированная часть стебля может ассоциироваться с широким спектром поздно реплицируемых областей: например, не только 6–9 т.п.н., но также 10 и 11 т.п.н. Однако до сих пор все проведенные анализы были агностическими - без учета информации о том, что распределение делеций частично объясняется прямыми повторами в мтДНК человека [21, 22] и бактериальных геномах [33]. Чтобы проверить важность пространственной структуры ДНК как фактора, влияющего на образование делеций, необходимо принять во внимание и прямые повторы. Для этого сравнено распределение идеальных прямых повторов (см. раздел «Методы» выше) и делеций из базы данных MitoBreak. Хотя прямые повторы могут объяснить локальное распределение делеций внутри определенных регионов, таких как область 6–10 по сравнению с 12–16 т.п.н. [22], в глобальном масштабе, в масштабе всей большой дуги, они плохо коррелируют с распределением делеций. Действительно, наблюдалось приблизительно однородное глобальное распределение прямых повторов внутри главной дуги по сравнению с сильно смещенным распределением делеций (**Рисунок 10**). Это наблюдение согласуется с предыдущими выводами Samuels et al. [21] и подтверждает, что сами по себе прямые повторы не полностью объясняют распределение делеций в мтДНК, и подчеркивает необходимость учитывать другие факторы, такие как роль пространственной структуры ДНК в формировании делеций мтДНК.

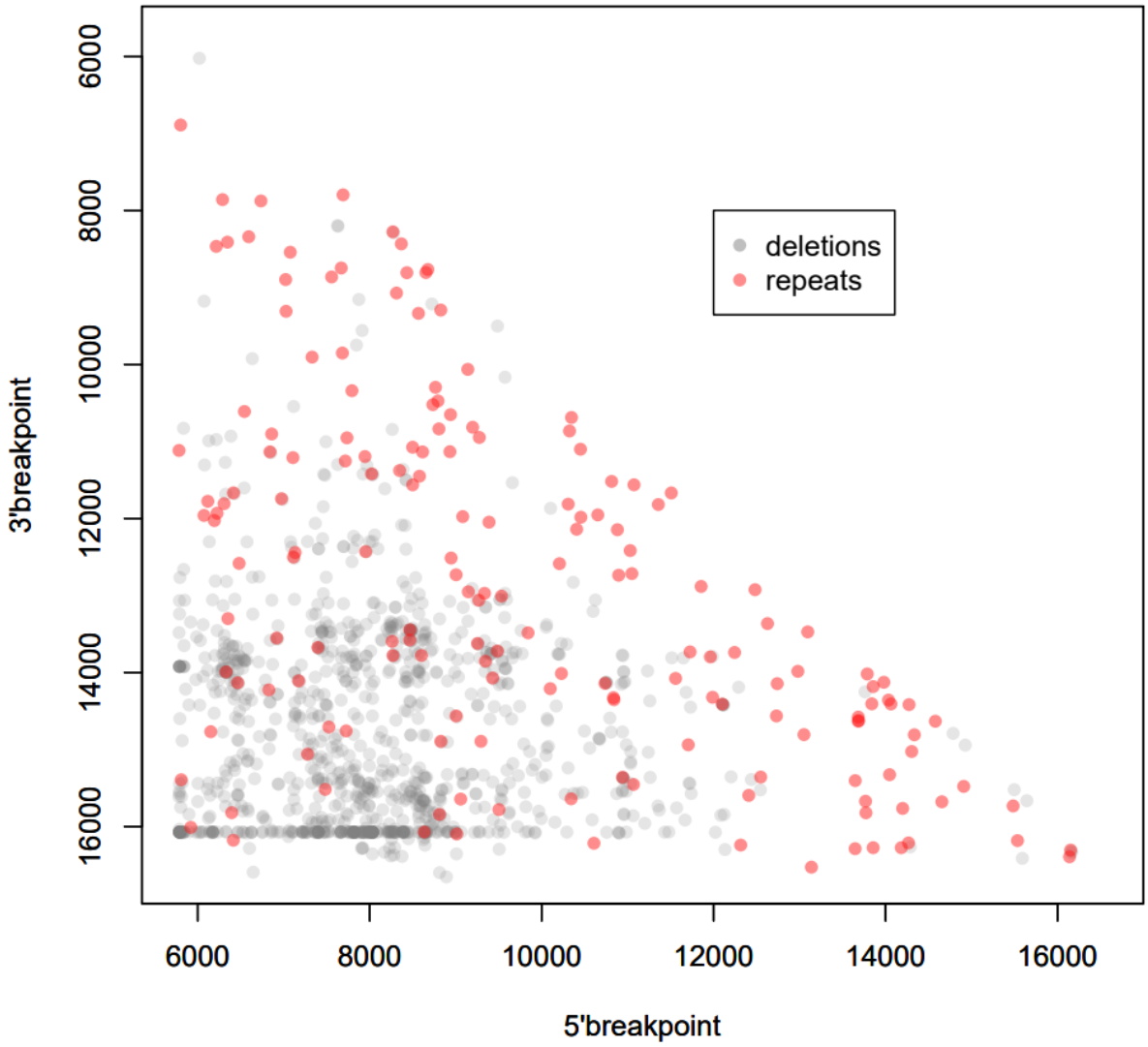

Рисунок 10. Распределение совершенных прямых повторов и делеций из MitoBreak по главной дуге.

Делеции могут быть вызваны специфическими, например, С-богатыми мотивами [34] внутри прямых повторов. Таким образом, существует вероятность, что смещенное распределение делеций может быть объяснено смещенным распределением таких мотивов - например, горячие, индуцированные делецией мотивы (deletion-induced motifs) могут быть расположены преимущественно в областях 6–9 и 13–16 т.п.н. Чтобы проверить потенциальное влияние специфических мотивов в прямых повторах на образование делеций, проанализирована база данных вырожденных повторов мтДНК человека [110], сгруппированных их по мотивам. Затем объединены все повторы с одним и тем же мотивом во все возможные пары, где один повтор был «реализован» (если есть делеция MitoBreak, фланкированная этими нуклеотидными последовательностями), а другой был «нереализован» (если в MitoBreak нет соответствующей делеции). Проведенное сравнение положений реализованных и нереализованных повторов показало, что реализованные повторы с большей вероятностью располагались вблизи области 6–9 и 13–16 т.п.н., тогда как нереализованные повторы были более равномерно распределены по всей большой дуге (**Рисунок 9А**). Парный анализ, специфичный для мотива, в котором сравнивались свойства реализованных и нереализованных повторов с одним и тем же мотивом, показал, что нереализованные повторы имеют тенденцию начинаться на 700 п.н. позже и заканчиваться на 1300 п.н. раньше, что приводит к сокращению расстояния на 2000 п.н. между плечами нереализованных повторов по

сравнению с реализованными повторами (все значения p < 2,2e-16, парный U-критерий Манна-Уитни; количество кластеров = 618).

В целом замечено, что распределение делеций крайне неоднородно (**Рисунок 9A**), и эта неравномерность не связана ни с прямым содержанием повторов (**Рисунок 10**), ни с мотивами прямых повторов (**Рисунок 9B**). Учитывая, что 80% реализованных повторов начинаются в интервале 6465–10954 п.н. и заканчиваются в интервале 13286–15863 п.н., можно предположить, что такое необъективное распределение объясняется потенциальной макромолекулярной зоной контакта одноцепочечной ДНК между 6–9 и 13–16 кб. Действительно, существует сильный избыток реализованных повторов в области 6–9 и 13–16 т.п.н. (**Рисунок 9B**, мозаичный график; отношение шансов Фишера = 7,5, p < 2,2e - 16).

3.3.2 Вероятность делеций зависит как от микрогомологии ДНК, так и от близости к точке контакта.

Показано, что распределение делеций внутри большой дуги плохо объясняется только распределением прямых повторов, в то время как потенциальная глобальная структура одноцепочечной мтДНК может быть дополнительным фактором, влияющим на образование делеций (**Рисунки 8 и 9A, 9B**). Здесь целью было построить множественную модель, включающую как повторы, так и вторичную структуру в качестве основных факторов, влияющих на образование делеций. Вместо прямых повторов получен более биологически значимый показатель «микрогомологического сходства», который представляет собой (i) интегральный показатель сходства между двумя областями ДНК и (ii) он фиксирован к окнам размером 100 п.о. для облегчения последующего анализа (см. раздел «Методы» выше). Оценены все попарные сходства микрогомологии между всеми окнами размером 100 п.н. внутри большой дуги (**Рисунок 9C**, нижний левый треугольник) и, прежде всего, как и ожидалось, получена положительная корреляция между сходством микрогомологии и плотностью прямых повторов в соответствующие окна (см. раздел «Методы» выше, rho Спирмена = 0,07, P = 1,698e - 06, N = 4950 регионов, окна 100 п.н. × 100 п.н.). Далее, чтобы проанализировать связь между делециями и сходством микрогомологии, расчитана логистическая регрессия, где наличие или отсутствие делеций в каждой ячейке размером 100 × 100 п.н. (кодируется как 1, если в клетке есть хотя бы одна делеция, N = 484; кодируемый как 0, если в клетке нет делеций, N = 4466) оценивалась как функция микрогомологического сходства (MS):

*log(p/(1-p)) = -2,25264 + 0,27442 * PC, все значения p (отрезок, коэффициент) меньше или равны 5.13e-09, N =4005. (Уравнение 1)*

Этот результат подтверждает полученные ранее результаты [22] и показывает, что высокое микрогомологическое сходство положительно коррелирует с более высокой вероятностью делеции в масштабе всей большой дуги.

На следующем этапе целью было включить в модель вторую независимую переменную, называемую зоной контакта (CZ). Переменная CZ кодировалась для каждой ячейки размером 100 × 100 п.н. как 1 внутри зоны (6–9 т.п.н. и 13–16 т.п.н.) и 0 для областей за пределами этой зоны.

*log(p/(1-p)) = -2,38296 + 0,32592 * MS + 0,90579 * CZ, все значения p меньше или равны 3,8e-10, N = 4005. (Уравнение 2)*

Полученные результаты показывают, что наличие контактной зоны оказывает значительное и положительное влияние на вероятность делеций. Используя стандартизированные переменные в уравнении, можно сравнить коэффициенты и определить, что переменная зоны контакта влияет на шансы вероятности в три раза сильнее (0,91 против 0,33), чем сходство микрогомологии.

Чтобы точно определить местоположение зоны контакта макромолекул, проведен дополнительный анализ данных. Вместо использования переменной зоны контакта введена переменная с евклидовым расстоянием от точки контакта до каждой ячейки матрицы контактов. Сделано предположение, что существует одна точка контакта, которая в сочетании с показателем микрогомологии наиболее эффективно объясняет распределение делеций (как показано на схеме на **Рисунке 9A**). Чтобы проверить эту гипотезу, выполнено 4005 логистических регрессий, каждая из которых имела разную точку контакта в качестве центра всех 4005 ячеек матрицы (все ячейки, исключая диагональную зону). Замечено, что самая сильная точка контакта (т. е. точка контакта, соответствующая модели с минимальным информационным критерием Акаике, AIC) имеет координаты 7550 п.н. как 5' и 15150 п.н. как 3'. Построенная тепловую карту с помощью AIC для каждой точки контакта, продемонстрировала, что зона контакта, основанная на данных, аналогична визуально полученной зоне контакта размером 6–9 КБ против 13–16 КБ (**Рисунок 9C**, верхний правый треугольник). В целом можно предположить, что распределение делеций мтДНК человека определяется как зоной макромолекулярного контакта одноцепочечной большой дуги, так и локальными микрогомологиями между участками ДНК (**Рисунок 9C**).

3.3.3 Одноцепочечная большая дуга мтДНК может быть свернута в крупномасштабную петлю благодаря свойствам ДНК, таким как инвертированные повторы

Одноцепочечная ДНК может сохранять свою структуру благодаря различным факторам, например, благодаря специфическим белкам, таким как SSB. Однако, когда одноцепочечная ДНК не покрыта никакими белками, она может стать более восприимчивой к структурным изменениям и делециям.

Учитывая, что делеции в мтДНК происходят нечасто (и могут быть связаны с колебаниями количества SSB - одноцепочечного связывающего белка) и, скорее всего, во время динамического процесса репликации мтДНК, когда одноцепочечная ДНК не полностью покрыта защитными белками, целью было исследовать пространственную структуру одноцепочечной большой дуги мтДНК: изучить форму, которую примет большая дуга, основываясь исключительно на свойствах ее ДНК последовательности.

Для реконструкции пространственной структуры одноцепочечной родительской тяжелой цепи большой дуги мтДНК проведено сворачивание in silico с использованием Mfold (см. раздел «Методы» выше). Используя результаты Mfold, полученные для одноцепочечной молекулы ДНК родительской тяжелой цепи, получена контактная матрица как количество водородных связей между двумя областями ДНК (**Рисунок 9D**, нижний левый треугольник). Наблюдалась интересная закономерность в матрице контактов: рисунок, представляющий собой диагональ из нижней левой части матрицы в верхнюю правую, перекрывал зону контакта между 6–9 и 13–16 кб. Этот крестообразный граф контактной матрицы напоминает бактериальные данные Hi-C [47] и предполагает, что одноцепочечная тяжелая цепь образует структуру, подобную шпильке, с центром, близким к 11 т.п.н., и крупномасштабным стеблем, образованным областями, которые выровнены. друг с другом, например, 9,5 КБ напротив 11,5 КБ, 8,5 КБ напротив 12,5 КБ, 7,5 КБ напротив 13,5 КБ, и самый сильный контакт обнаружен на 6,5 КБ напротив 14,5 КБ (нижний левый треугольник на **Рисунок 9D**) .

Считается, что глобальная вторичная структура одноцепочечной ДНК поддерживается за счет микрогомологии, включая инвертированные повторы, которые могут гибридизоваться друг с другом с образованием стеблей. Программа Mfold использует избыток и сходство различных инвертированных повторов для реконструкции формы одноцепочечной ДНК. Чтобы проверить результаты, полученные с использованием Mfold, скоррелирована плотность инвертированных повторов в окнах размером 100 п.н. с соответствующими контактными плотностями сворачивающейся матрицы in silico Mfold (нижний левый треугольник на **Рисунок 9D**). Наблюдалась положительная корреляция между двумя переменными (rho Спирмена = 0,05, p = 0,0017, N = 4005, диагональные элементы были удалены из анализа), что подтверждает, что

как предсказания Mfold, так и плотность инвертированных повторов демонстрируют аналогичную тенденцию.

Фолдинг (сворачивание) in silico очень длинной (~ 10 т.п.н.) одноцепочечной молекулы ДНК, использованное для получения результата на **Рисунок 9D** (левый нижний треугольник), может иметь вычислительные ограничения и искусственно форсировать происхождение шпильки как структуры. Чтобы избежать этой потенциальной проблемы, главная дуга была разбита на короткие (100 п.н.) окна, все парные комбинации были свернуты, оценены энергии Гиббса для каждой пары и, наконец, реконструирована мелкомасштабная матрица контактов (верхний правый треугольник на **Рисунок 9D**). На мелкомасштабном графике матрицы контактов видно несколько полос, соответствующих наиболее сильным контактам (три горизонтальные линии с ординатами 6100, 6900 и 7900 и одна вертикальная с абсциссой 15 000), причем пересечение этих линий хорошо перекрывается с зоной контакта. В целом подход фолдига in silico подтверждает существование зоны контакта между 6–9 и 13–16 т.п.н. мтДНК на основе свойств одноцепочечной ДНК.

3.3.4 Контактная зона описывает динамику делеций, возникших при здоровом старении

В недавнем исследовании с помощью сверхчувствительного метода было обнаружено около 470 000 уникальных делеций в мтДНК человека [30]. Биоинформатический анализ этого набора данных выявил три основных компонента, описывающих основные свойства делеций: (i) делеции, связанные с заболеванием и здоровьем, (ii) расположены в малых или больших дугах; (iii) молодой или пожилой возраст на момент биопсии.

Авторы [30] обнаружили, что делеции с высокими баллами по третьему основному компоненту (a) связаны с пожилым возрастом, (б) расположены преимущественно в пределах большой дуги мтДНК, (в) имеют высокое микрогомологическое сходство между точками разрыва, и (d) были расположены определенным образом в большой дуге, где точки разрыва вблизи начала репликации и в середине большой дуги в основном были представлены в меньшей степени (рис. 4C в оригинальной статье [30]). Этот специфический способ расположения делеций сильно напоминает контактную зону, полученную в моем исследовании. Это сходство предполагает, что образование возрастных делеций — наиболее распространенных делеций в человеческой популяции — обусловлено главным образом контактной зоной. Действительно, используя метаданные анализа главных компонентов, предоставленные авторами, подтвердено, что баллы третьего главного компонента главной дуги были значительно выше для ячеек, расположенных внутри зоны контакта макромолекул, по сравнению с ячейками, расположенными вне зоны контакта (p - значение <4,48 -13, U-критерий Манна-Уитни, **Рисунок 11**). Это показывает, что пространственная структура одноцепочечной мтДНК, и особенно контактная зона, играет важную роль в формировании здоровых возрастных делеций. Важно также подчеркнуть, что механизм возникновения этого класса делеций, как предполагается, происходит как проскальзывание праймера во время асинхронного смещения цепи репликации мтДНК [30], что полностью подтверждает сделанные выводы о том, что пространственная структура одноцепочечной родительской тяжелой ДНК цепь имеет большое значение для образования делеций (**Рисунки 8 и 9**, см. также интегральную схему образования делеций на **Рисунке 12**).

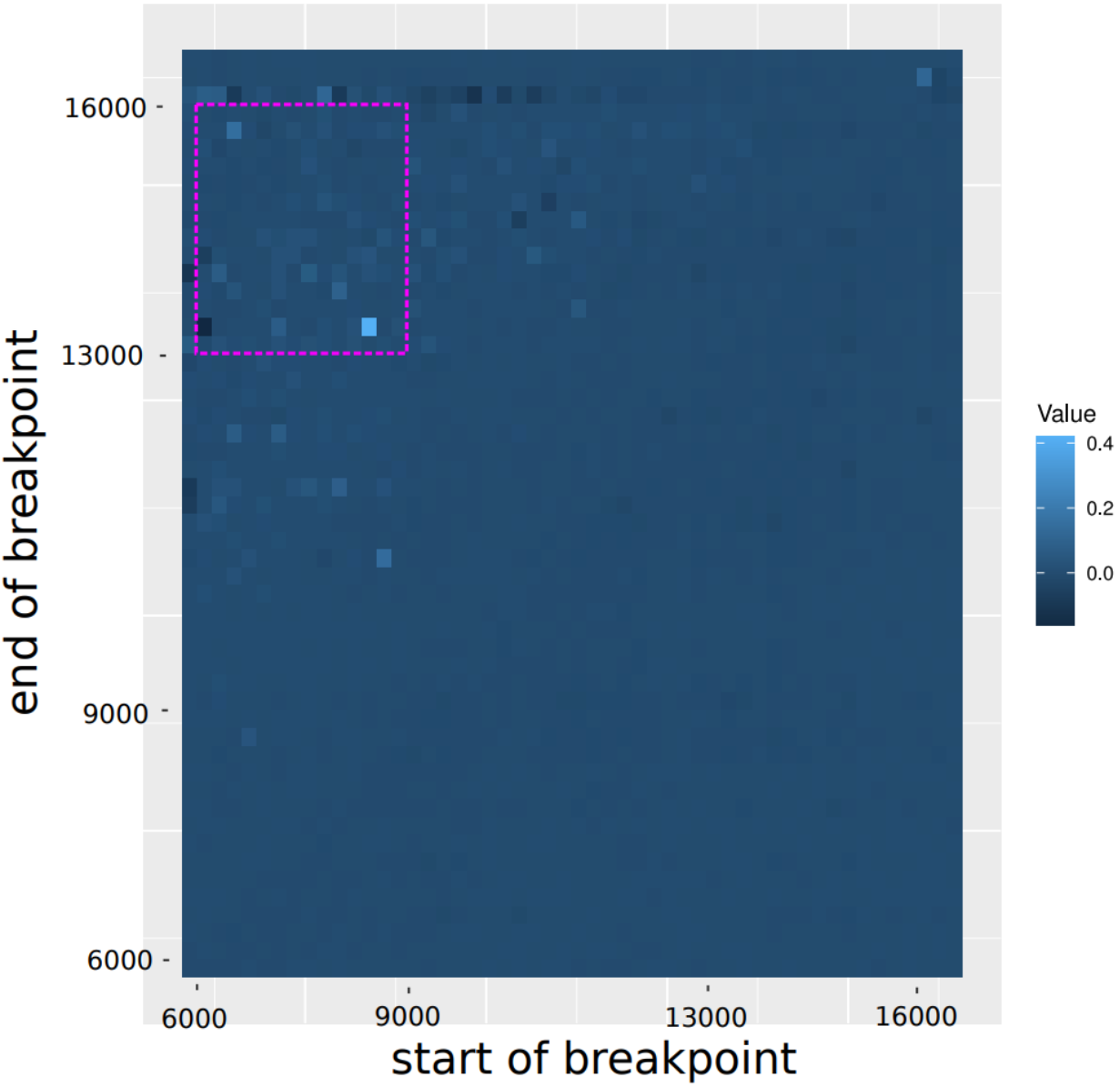

Рисунок 11. Третий главный компонент оценок, связанный с удалением здоровых образцов, связанным со старением, из статьи [48]. Контакт, отмеченный розовым квадратом, характеризуется повышенными оценками.

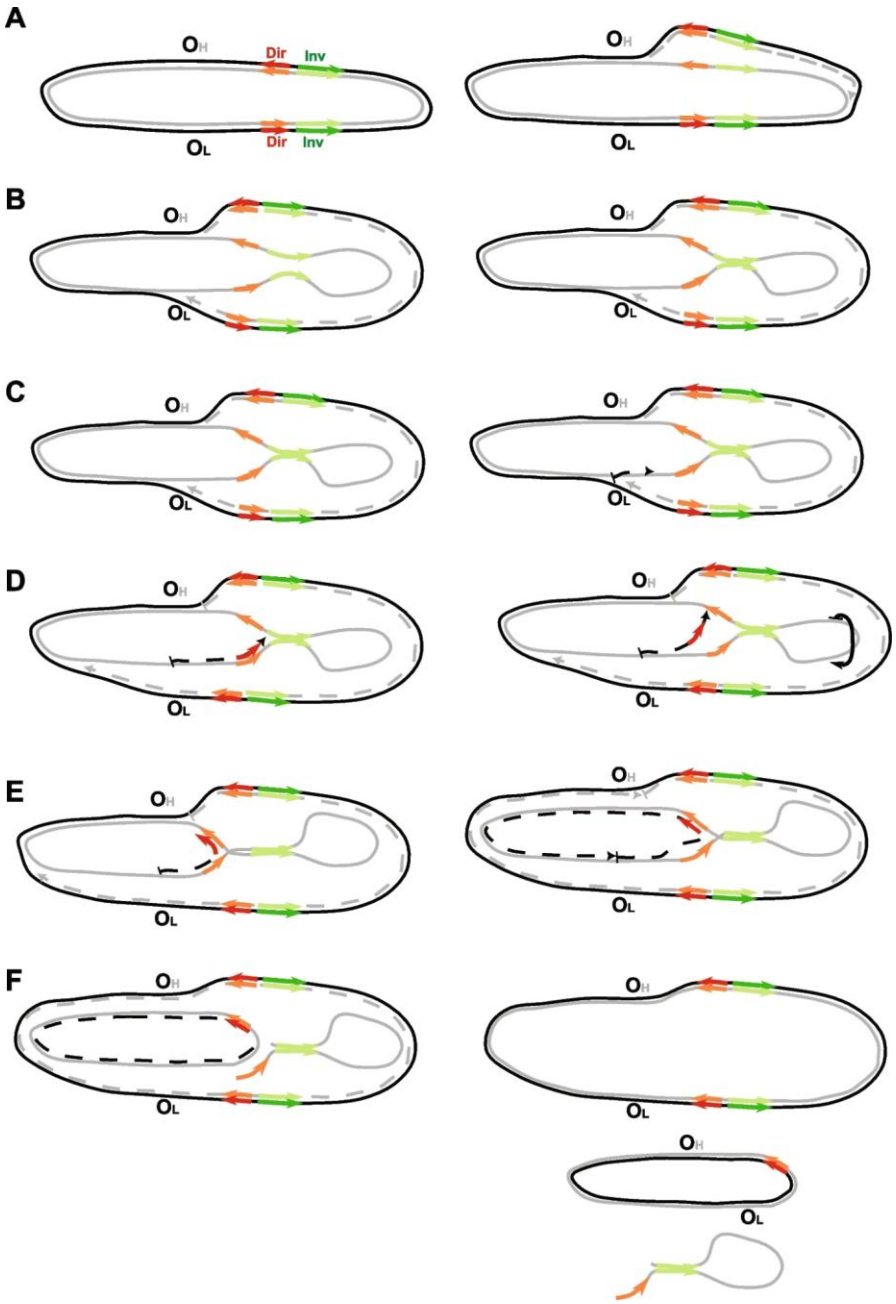

Рисунок 12. Интегральная схема происхождения делеций мтДНК. Родительская тяжелая цепь обозначена серой линией, родительская легкая цепь обозначена черной линией. Дочерние цепи обозначены пунктирными линиями. Прямые повторы нуклеотидов обозначены красными стрелками, а инвертированные повторы нуклеотидов — зелеными стрелками. Отмечены начало репликации тяжелой цепи (ОН) и начало репликации легкой цепи (ОЛ). A В начале репликации дочерняя тяжелая цепь (пунктирная серая линия) реплицируется на матрице родительской легкой цепи (черная линия), и репликационная вилка начинает двигаться от ОН к ОЛ. B В течение этого времени родительская тяжелая цепь остается одноцепочечной (ssDNA) в течение значительного периода времени, и различные типы микрогомологии, включая инвертированные повторы (зеленые стрелки), могут сворачивать пространственную структуру. C Как только первая репликативная вилка достигает OL, вторая начинает реплицировать легкую цепь дочерней цепи (пунктирная линия сзади) на матрице родительской тяжелой цепи. D Вторая репликативная вилка останавливается около стебля, инициированного инвертированными повторами. Если время остановки достаточно велико и пространственная структура одноцепочечной ДНК не может быть разрешена, вновь синтезированная дочерняя цепь частично диссоциирует и повторно выравнивается со вторым плечом прямого повтора. Если одноцепочечная ДНК образует стебель, она может вращаться (как указано черной стрелкой), потенциально еще больше сближая прямые повторы. E Это позволяет репликативной вилке продолжить репликацию. F Репликация легкой цепи дочерней цепи завершается, и область петли одноцепочечной ДНК исчезает либо во втором раунде репликации мтДНК (см. рис. 5f в [34]), либо когда мтДНК с делециями восстанавливается, а одноцепочечная ДНК деградирует

3.3.5 Двухцепочечная большая дуга мтДНК также может быть свернута в крупномасштабную петлю

Согласно недавнему сообщению [34] и полученным результатам (**Рисунки 8, 9 и 12**), можно считать, что делеции мтДНК возникают во время репликации мтДНК, когда длинный участок родительской тяжелой цепи остается одноцепочечным. Однако двухцепочечная мтДНК может также принимать форму, аналогичную форме одноцепочечной мтДНК, где оба начала репликации — тяжелый и легкий — расположены проксимально в зоне контакта. Эта близость может облегчить регуляцию репликации посредством прямых перекрестных помех между двумя источниками репликации. Чтобы проверить пространственную структуру двухцепочечной мтДНК, использована общедоступная контактная матрица высокой плотности из экспериментов Hi-C на лимфобластоидных клетках человека с доступным разрешением не менее 1 т.п.н. [48]. Наблюдались контакты высокой плотности между 0–1 т.п.н. и 15–16,5 т.п.н., что, вероятно, отражает кольцевую природу мтДНК (нуклеотиды с положениями 1 и 16569 являются соседними нуклеотидами). Это подтверждает, что пространственная реконструкция мтДНК по данным Hi-C является биологически адекватной (**Рисунок 13**, овалами отмечены контакты, отражающие кольцевую природу мтДНК). Далее наблюдался второй контакт, который находился в пределах большой дуги и сильно напоминал зону контакта: 6–9 кб против 13–16 кб (**Рисунок 13**, пунктирные квадраты обозначают потенциальную зону контакта). Наличие этой зоны контакта мтДНК позволяет предположить, что двухцепочечная большая дуга также может принимать петлеобразную форму, а вся двухцепочечная мтДНК может напоминать «символ бесконечности» с зонами контакта между положениями 6–9 т.п.н. и 13–16 т.п.н.

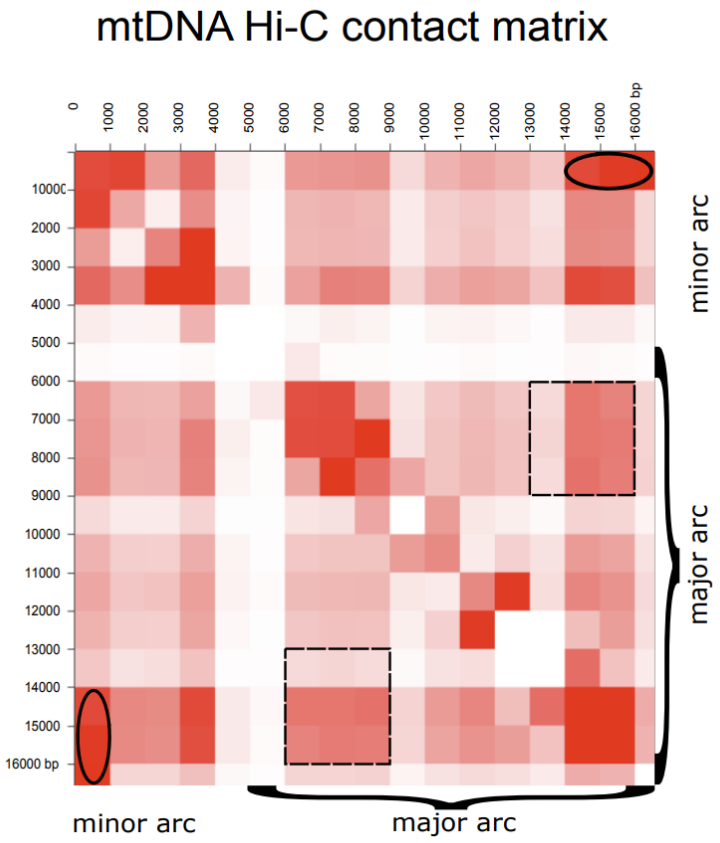

Рисунок 13. Контактная матрица Hi-C мтДНК, полученная из человеческих лимфобластоидных клеток. Пунктирные квадраты обозначают потенциальные контактные зоны. Овалы обозначают контакты, подчеркивая округлость мтДНК.

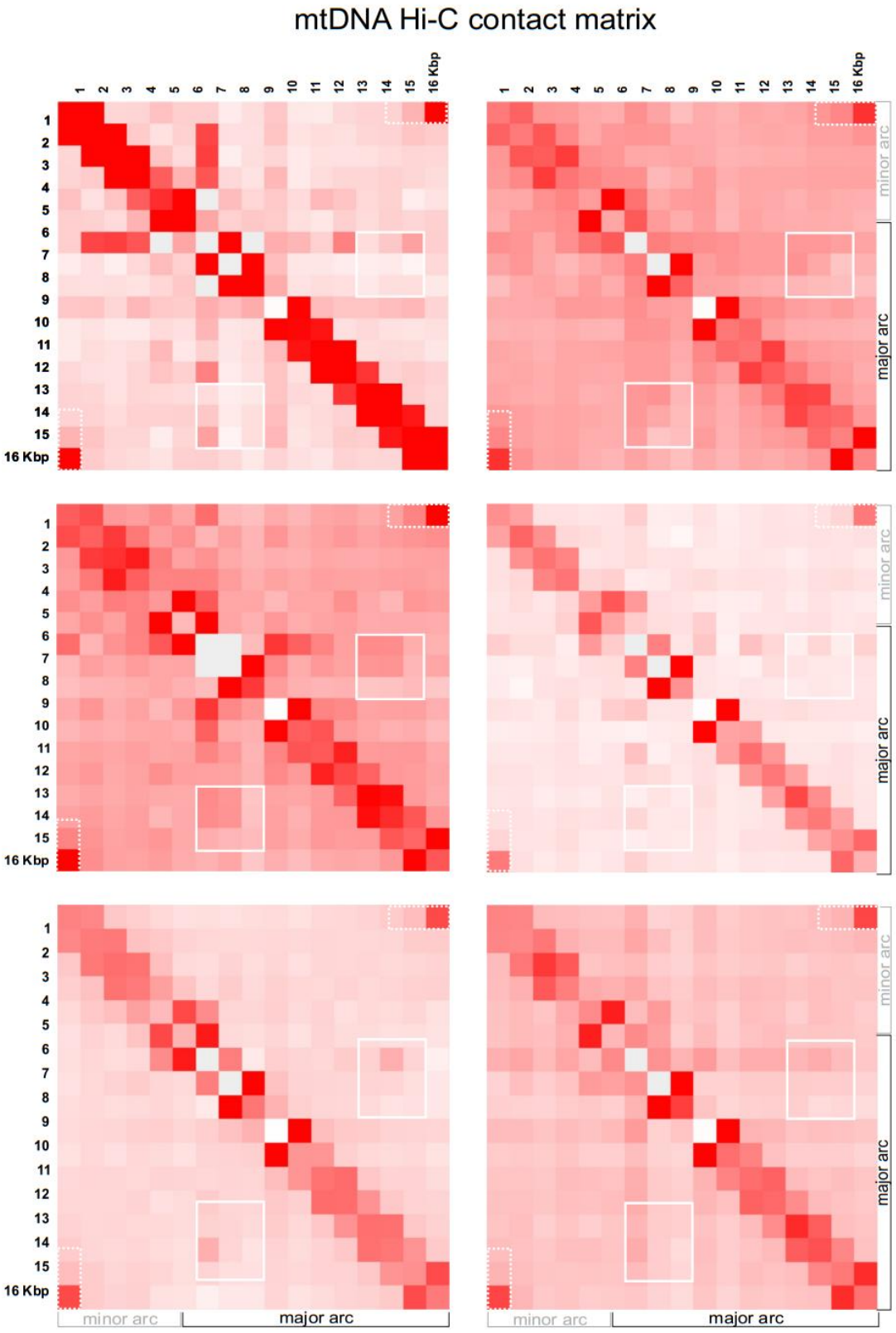

Рисунок 14. Контактная матрица Hi-C мтДНК, полученная из аутопсий обонятельного эпителия человека. Верхний ряд представляет собой две контактные матрицы от пациентов с COVID, средний и нижний ряды представляют контактные матрицы от контролей. Сплошные белые квадраты обозначают потенциальную контактную зону. Пунктирные белые прямоугольники обозначают контакты, подчеркивая округлость AQmtDNA.

Чтобы проверить надежность общедоступной матрицы Hi-C мтДНК [48], дополнительно взяты шесть контактных матриц Hi-C мтДНК, полученных из обонятельных рецепторов контрольной группы и пациентов с COVID [109]. Анализ этих контактных матриц, несмотря на низкую степень покрытия и технический шум, подтвердил существование контактов между позициями 0–1 т.п.н. и 15–16,5 т.п.н., отражающих кольцевую природу мтДНК, а также контактов между позициями 6–9 т.п.н. и 13–16 т.п.н., что поддерживает предложенную модель «символ бесконечности» (как показано в **Рисунке 14**). Никаких существенных различий между пациентами с COVID-19 и контрольной группой не наблюдалось (**Рисунок 14**). Однако для дальнейшего выяснения формы двухцепочечной мтДНК необходимы дальнейшие исследования Hi-C с высоким разрешением, сосредоточенные на мтДНК.

3.4 Вывод

Проведенные исследования позволяют пересмотреть традиционные взгляды на механизм образования делеций в митохондриальной ДНК. Если ранее основной акцент делался на локальных последовательностях, таких как прямые повторы, то полученные данные свидетельствуют о определяющей роли макромолекулярной пространственной структуры. В частности, было продемонстрировано, что на образование делеций влияет не только локальное сходство участков, но и их пространственная близость, обусловленная структурой одноцепочечной тяжелой цепи большой дуги мтДНК во время репликации. Мы предположили и подтвердили, что эта цепь формирует шпилькообразную структуру с четко выраженной контактной зоной между регионами ~6–9 т.п.н. и ~13–16 т.п.н., что и определяет неравномерный спектр делеций (**Рисунки 8, 9, 12**).

Ключевым результатом стало открытие и характеристика этой контактной зоны. Было показано, что пространственная близость в ней преобладает над локальной гомологией: наличие прямого повтора в зоне контакта увеличивает вероятность делеции в 3 раза по сравнению с аналогичным повтором вне ее. Моделирование in silico и данные Hi-C независимо подтвердили способность мтДНК сворачиваться в крупномасштабную петлю («символ бесконечности»). Наиболее точное описание формирования делеций достигается двухфакторной моделью, учитывающей как микрогомологию, так и принадлежность к контактной зоне. Важно, что эта структура объясняет и физиологически значимые события, внося основной вклад в образование делеций, ассоциированных со здоровым старением.

Для дальнейшего укрепления этих выводов при анализе данных Hi-C целесообразно учитывать зависимость частоты контактов от линейного расстояния. Перспективным подходом является расчет для каждой пары участков отношения наблюдаемой частоты контактов к ожидаемой (O/E), где ожидаемая частота оценивается на основе средних значений для данного расстояния в последовательности. Сравнение распределений O/E-отношений для участков в зонах частых делеций и контрольных участков позволило бы более строго выявить статистически значимые контакты.

Таким образом, опровергнута гипотеза, ставившая во главу угла общий повтор. Вместо этого предложена новая парадигма, в которой глобальная вторичная структура мтДНК выступает организатором мутагенеза, определяя, какие из многочисленных прямых повторов реализуются в делеции. Это объясняет давнюю загадку неравномерного распределения разрывов и открывает новые пути для исследований, связывающих архитектуру мтДНК, ее репликацию и геномную нестабильность.

Глава 4. Влияние нарушенного общего повтора на здоровое старение на примере гаплогрупп японских долгожителей

4.1 Проблематика

Чтобы раскрыть роль прямых повторов в образовании делеций мтДНК, фокус сделан на ключевых свойствах наиболее мутагенного прямого повтора мтДНК человека: «общего повтора» (common repeat) - прямого повтора длиной 13 п.о., начинающегося с нуклеотидов 8470 и 13447 (8470-8482: ACCTCCCTCACCA; 13447-13459: ACCTCCCTCACCA), обнаруженного у большинства людей. Подавляющее большинство различных делеций мтДНК, по-видимому, связано с этим повтором, что указывает на общий механизм, связанный с репликацией мтДНК [24]. Примечательно что общий повтор расположен именно в потенциальной зоне контакта (описанной ранее контактной зоне), что позволяет предположить близость обоих плеч друг к другу во время репликации. Но есть гаплогруппы японских долгожителей, у которых этот повтор нарушен [27, 28]. Можно предположить, что нарушение прямого повтора приводит к отсутствию большинства распространенных делеций, что в конечном итоге может приводить к увеличению продолжительности жизни. Используя коллекцию 43437 мтДНК человека из разных гаплогрупп [115], разделенную на кейсы (с нарушенным общим повтором) и контроли (повтор не нарушен), делается пытатка выяснить, связаны ли определенные особенности общего повтора с продолжительностью жизни человека. Это означает, что на результаты может влиять то, какие нуклеотиды уже присутствуют в повторе, а не только само наличие повтора. Здесь исследуются факторы, влияющие на скорость возникновения соматических делеций мтДНК, с конечной целью расчета риска и спектра соматических делеций для различных последовательностей мтДНК у людей. Анализируется связь между прямыми повторами мтДНК и долголетием у человека, обсуждаем связь между нарушением общих повторов в мтДНК и увеличением продолжительности жизни японской гаплогруппы D4a и предполагаем существование функциональной связи между отсутствием повтора, дефицитом соматических делеций и увеличением продолжительности жизни. Рассматривается общий прямой повтор в контексте эволюции, чтобы понять как он возник и способен ли он элиминироваться.

4.2 Методы

4.2.1 Подготовка к выравниванию

В своем исследовании я ставлю перед собой задачу сравнить скорость эволюции геномов мтДНК, инкапсулирующих основной вариант нарушенного общего повтора, с теми, которые содержат неповрежденные оба плеча общего повтора. Анализ основывался на обширной выборке последовательностей мтДНК человека, полученной из широкого спектра репозиториев. Чтобы смягчить систематическую ошибку, вызванную влиянием отбора, было обеспечено ослабление давления отбора. Это повлекло за собой исключение последовательностей с сайтами, содержащими вредные мутации или подвергшимися значительному положительному отбору, которые обычно быстро отбираются в ходе недавней митохондриальной эволюции, или последовательностей с ограниченной изменчивостью.

Использована классификация последовательностей генома мтДНК из базы данных HmtDB (получена версия от октября 2018 г., http://www.hmtdb.uniba.it) [115], который разделяет последовательности мтДНК на две группы: те, которые содержат вредные мутации, и те, которые лишены таких мутаций. Множественное выравнивание из HmtDB, содержащее 43 437

последовательностей генома мтДНК без вредных мутаций, послужило основой для данного анализа.

Чтобы убедиться, что проанализированы нейтральные или почти нейтральные сайты мтДНК для сравнения скоростей эволюции, извлечены только вариабельные сайты из этого множественного выравнивания. Это повлекло за собой исключение любых неоднозначных символов, кроме A, T, G или C. Принято три порога вариации сайта: >0,5% (791 сайт), >0,1% (1941 сайт) и >0,05% (2778 сайтов) вариации. Эти пороги приблизительно отражают степень отклонений от ожиданий почти нейтрального сайта.

4.2.2 Филогенетическая реконструкция

Одновременно для каждой мтДНК была спрогнозирована гаплогруппа с помощью программы HaploGrep v2.1.20 [116]. Кроме того, используя все множественное выравнивание полных геномов мтДНК, реконструировано *Филогенетическое дерево мтДНК человека* с помощью программного обеспечения IQ-TREE v. 1.6.1 [117]. При этом использовалась общая обратимая во времени (GTR) модель замен оснований, а также модель вариаций сайтов FreeRate [118] с учетом доли неизменяемых сайтов (вариант: -m GTR+F+I+R6).

Реконструированное общее филогенетическое дерево с выделенными гаплогруппами для каждой последовательности позволило четко разграничить пять подмножеств для сравнительного изучения скорости эволюции геномов мтДНК, содержащих нарушенные и неповрежденные общие повторяющиеся последовательности. Эти подмножества включали последовательности, имеющие нормальные общие плечи повторов (контроль) и последовательности демонстрирующие нарушение общего повтора из-за мутации m.8473T>C (случай). Используя эти подвыравнивания, созданы неукорененные филогенетические деревья с помощью программного обеспечения IQ-TREE. После того, как эти неукорененные деревья были построены, использовалось программное обеспечение Archeopteryx (версия 0.9928 бета) [119] для извлечения топологии деревьев, что дало возможность понять отношения между пятью подгруппами и определить корневые последовательности для каждого из подмножеств. Эти корневые последовательности представляют общего предка последовательностей в подмножестве. Последовательности как контролей, так и кейсов имели общего недавнего предка и были почти равны по количеству. Неукоренные деревья и идентифицированные корневые последовательности были использованы для углубленного анализа длин ветвей.

4.2.3 Анализ длины ветвей скорости эволюции

Стремясь выявить потенциальные различия в темпах эволюции между кейсами и контрольной группой, проанализированы длины ветвей филогенетических деревьев, созданных на предыдущем этапе. Длина ветвей филогенетического дерева является мерой генетических изменений, и сравнение этих длин может дать представление о темпах эволюции различных групп. Однако проведение справедливого сравнения требует определенных корректировок для учета присущих данным различий.

Модели, которые выбраны для анализа, а именно JC, F81, K80 и HKY, являются хорошо зарекомендовавшими себя моделями эволюции последовательностей ДНК. Эти модели описывают вероятности различных типов мутаций (таких как переходы и трансверсии), происходящих с течением времени. Модель JC предполагает равные базовые скорости замещения и равные базовые частоты, модель F81 предполагает равные скорости, но неравные базовые частоты, модель K80 предполагает неравные скорости перехода/трансверсии и равные базовые частоты, а модель HKY допускает неравные скорости перехода/трансверсии и неравные базовые частоты.

Чтобы укоренить эти деревья и количественно определить длину ветвей от корня, использованы Newick Utilities v1.6 [120]. Поэтому, чтобы обеспечить объективное сравнение, вычтено значение X/L из длин ветвей, представляющих последовательности с нарушенным общим повтором. Здесь X — это либо единица (в случае моделей JC и F81), либо скорость перехода/трансверсии (в случае моделей K80 и HKY), а L — количество сайтов в

выравнивании. Эта корректировка учитывает различия в частоте мутаций и оснований, которые по своей сути учитываются этими моделями.

Статистическая устойчивость проведенного анализа была затем оценена с использованием процедуры random-half-jackknife (собственный сценарий Perl реализующий метод складного ножа — метод повторной выборки используемый для оценки изменчивости выборочной статистики). В данном контексте процедура random-half-jackknife включала случайный выбор половины ветвей как из фоновой (контрольной), так и из передней (контрольной) группы и сравнение их длин. Этот процесс был повторен 100 раз, чтобы гарантировать надежность результатов.

Чтобы установить статистическую значимость любых наблюдаемых различий в темпах эволюции между основной и контрольной группами, использован знаково-ранговый критерий Уилкоксона (версия R 3.4.1; https://www.R-project.org/). Это непараметрический статистический тест, который сравнивает две связанные выборки или повторные измерения в одной выборке, чтобы оценить, различаются ли средние ранги их совокупности. Представлются результаты как направление качественного сдвига клад U6a, U2e, H1c, R/P и D4 со средним значением джекнайф-оценки и стандартным отклонением для значения p, количественно определенного с помощью теста Уилкоксона для каждой из эволюционных моделей F81, HKY, JC и K80 в трех вариантах выбора строгости вариантов сайта.

Наконец, чтобы визуализировать сравнение длин ветвей фона и переднего плана, используется пакет vioplot R версии 0.3.0 [121], что особенно полезно для сравнения распределений по разным группам, что делает его подходящим для данного анализа. Благодаря этому тщательному и комплексному подходу обеспечен надежный и глубокий анализ скорости эволюции геномов мтДНК, содержащих нарушенные и неповрежденные общие повторяющиеся последовательности.

4.3 Результаты

4.3.1 Общий повтор мтДНК может повлиять на продолжительность здоровья человека

Анализ связи между долголетием человека и свойства мтДНК показал, что самый длинный (13 пар оснований) и самый тяжелый прямой повтор (так называемые "общий повтор» наблюдается у 98% человеческой популяции) в митохондриальном геноме человека был нарушен точечной мутацией в нескольких гаплогруппах. Среди носителей нарушенного общего повтора наиболее частым вариантом является m.8473T>C, который возникал независимо несколько раз, что привело к образованию 5 больших кластеров, содержащих более 20 геномов каждый, и множества малых кластеров (см. **Таблицу 7**).

Таблица 7. Список вариантов, нарушающих проксимальное плечо общего повтора и соответствующих гаплогрупп, в базе данных HmtDB более 20 случаев (43437 мтДНК).

Последовательность проксимального плеча прямого повтора (8470–8482 п.н.)	Гаплогруппы	Количество особей
совершенный прямой повтор: ACCTCCCTCACCA	**большинство гаплогрупп, предковое состояние и RefSeq**	**42641**
м.8473T>C; ACCcCCCTCACCA; СИН	D4a (89), R2 (68), U2e (65), H1c (56), U6a (42) и спорадические случаи в других гаплогруппах (79)	399
м.8472C>T; ACtTCCCTCACCA; Про>Лей	N1b (106) и спорадические случаи в других гаплогруппах (38)	144
м.8479A>G; ACCTCCCTCgCCA; СИН	D5a (57) и спорадические случаи вдругие гаплогруппы (3)	60
м.8470A>G; gCCTCCCTCACCA; СИН	Рассредоточенные спорадические случаи в различных гаплогруппах (32)	32
м.8478C>T; ACCTCCCTtACCA; Сер> Лей	L4b (18) и спорадические случаи в других гаплогруппах (12)	30
м.8477T>C; ACCTCCCcCACCA; Сер>Про	Рассредоточенные спорадические случаи в различных гаплогруппах (26)	26

Наиболее изученная и наиболее секвенированная гаплогруппа этого варианта — D4a. Японская гаплогруппа который отличается повышенной долговечностью [27]. Это послужило основой для первоначально выдвинутой гипотезы [26]. Эта гаплогруппа хорошо известна благодаря значительному относительному избытку долгожителей (лиц, живущих 100 и более лет) и сверхдолгожителей (лиц, живущих 110 и более лет) [27, 28]. Сделано предположение, что нарушение общего повтора благотворно влияет на гаплогруппу D4a [26] поскольку это снижает вероятность возникновения соответствующей соматической делеции. Эту делецию также называют «обычной» (common) делецией, поскольку она часто наблюдается в старых постмитотических тканях (**Рисунок 15A**).

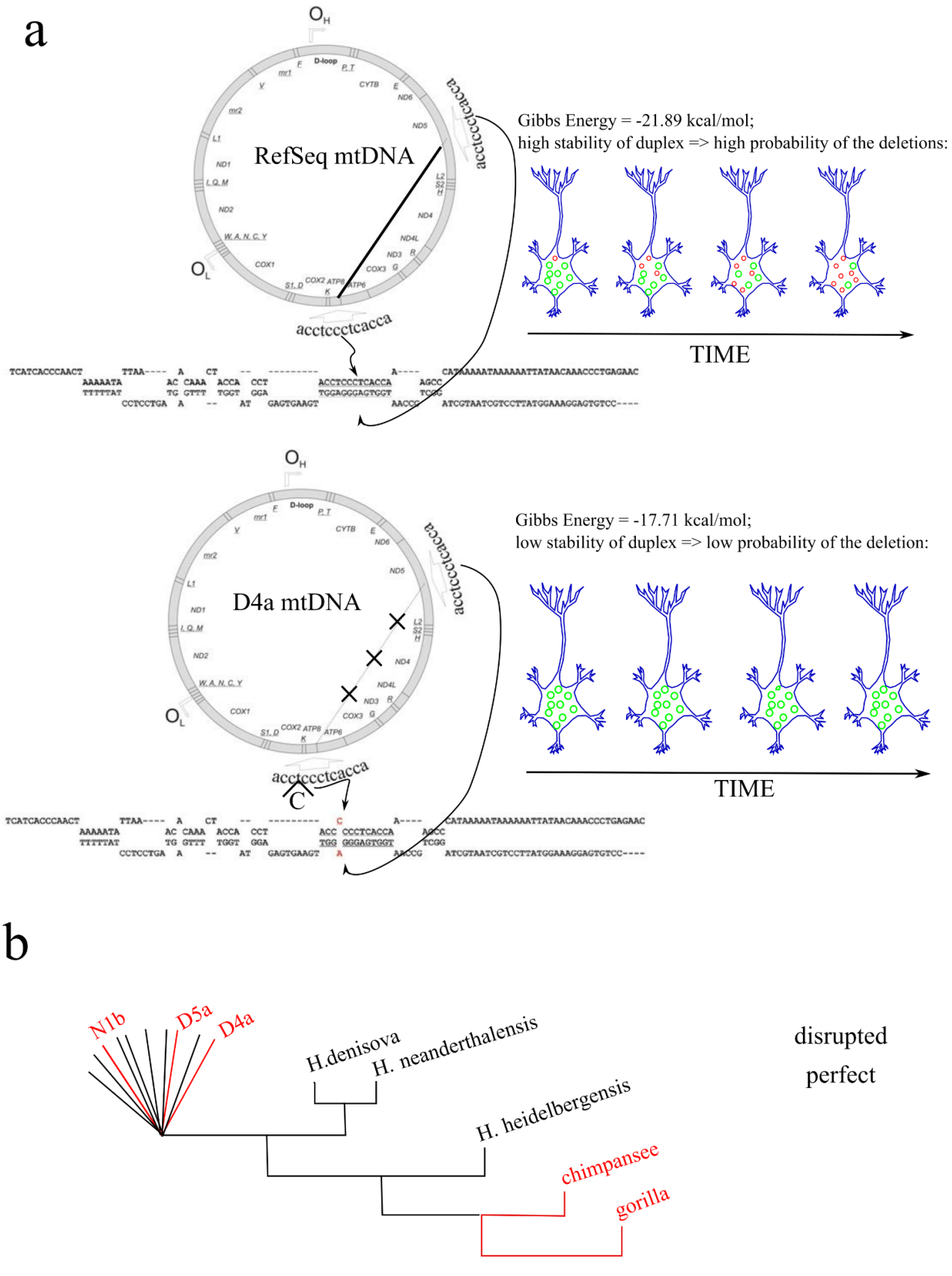

Рисунок 15. Общий повтор. a. Рабочая гипотеза: нарушенный общий повтор препятствует возникновению соматических делеций мтДНК, поддерживая постмитотические клетки (нейроны и скелетно-мышечные клетки) в более здоровых условиях (зеленые круги — дикий тип мтДНК; красные маленькие круги — мтДНК с общей делецией), тем самым откладывая возрастные фенотипы. b. Филогенетическая схема общего повтора. Совершенный общий повтор специфичен для человека

Вторым наиболее распространенным вариантом нарушенного общего повтора является m.8472C>T, который принадлежит Ашкенази-специфичной гаплогруппе N1b, где было продемонстрировано значительное снижение общей делеции в старых нейронах [22]. В качестве эксперимента для подтверждения принципа проанализированы образцы лобной коры двух пожилых людей из гаплогруппы N1b, содержащих аналогичный, но не идентичный варианту зародышевой линии D4a, m.8472C>T, нарушающий общий повтор [22]. В соответствии с выдвинутой гипотезой вообще не наблюдалось общих делеций в их митохондриальных геномах, а это означает, что нарушение повтора длиной 13 п.н. даже однонуклеотидным вариантом зародышевой линии полностью блокирует образование общей соматической делеции. Недавно была предложена новая модель образования делеций [34], который постулирует, что образование делеции мтДНК является результатом проскальзывания репликации во время активного синтеза L-цепи мтДНК. Весьма интересно, что авторы своим *in vitro* экспериментом продемонстрировали, что удаленный продукт терялся при мутации 8470 или 13447 плеч общего повтора [34]. Стоит подчеркнуть, что результаты этого *in vitro* эксперимента полностью соответствуют исходной выдвинутой гипотезе [26], а также с отсутствием общих делеций в гаплогруппе N1b [22].

Третий наиболее распространенный вариант — m.8479A>G, принадлежащий к гаплогруппе D5a, который также характеризуется повышенным долголетием независимо от D4a [28].

Недавний анализ базы данных митохондриального генома человека [122] содержащий 196554 полные записи подтверждают редкость нарушенного общего повтора. Согласно этим данным, наиболее распространенным нарушенным вариантом был m.8473T>C (1%: 2118 из 196554), а вторым наиболее распространенным вариантом был m.8472C>T (0,4%: 833 из 196554). Интересно отметить, что оба частых синонимичных варианта m.8473T>C и m.8479A>G связаны с увеличением продолжительности жизни [27**,** 28]. Синонимические варианты могут быть связаны с долголетием, поскольку их единственным ожидаемым эффектом является нарушение общего повтора без каких-либо потенциально вредных аминокислотных замен. Напротив, несинонимичный N1b-специфичный вариант (m.8472C>T) не был связан с увеличением продолжительности жизни, что может быть связано с потенциально полезным эффектом разрушения повтора, т.е. снижением делеционной нагрузки [22], было уравновешено вредным эффектом редкой аминокислотной замены.

Бактериальные делеции, которые могут иметь тот же механизм происхождения, что и делеции мтДНК, действительно, делеции у бактерий становятся на порядок реже, если в прямой повтор вводится несовпадение в один нуклеотид [33].

Еще одной особенностью общего повтора является то, что нуклеотидная последовательность повтора (aCCTCCCTCACCAx) сильно напоминает вырожденный мотив из 13 п.о. (CCNCCNTNNCCNC), который чрезмерно представлен в предсказанных горячих точках рекомбинации человека в ядерном геноме [123] и соответствует предсказанному домену связывания для белка PRDM9 [123, 124, 125]. Хотя PRDM9 не входит в митопротеом [126] и, таким образом, его участие в делециях мтДНК крайне маловероятно, было решено подчеркнуть это поразительное сходство мотивов, чтобы облегчить потенциальное будущее открытие важности мотива в митохондриях. Кроме того, мотив общего повтора очень богат C (8 из 13), что является отличительной чертой высокомутагенных повторов [34].

Сопоставив несколько ортогональных линий доказательств: (i) связь соматической делеционной нагрузки мтДНК с нейродегенерацией [4, 5] и саркопенией [6, 127]; (ii) связь между нарушенным повтором и увеличением продолжительности жизни гаплогруппы D4a [27, 28]; (iii) отсутствие общих делеций в старых образцах лобной коры в гаплогруппе N1b с нарушенным общим повтором [22]; и (iv) блокирование образования делеции мутированными плечами общего повтора в *in vitro* эксперимент [34], можно заключить что нарушенный общий повтор действительно может обеспечить более здоровое старение носителям за счет устранения соматической делеционной нагрузки.

4.3.2 Нет доказательств того, что нарушение общего повтора имеет эволюционное преимущество.

Анализ эталонных последовательностей мтДНК приматов и нескольких доступных древних мтДНК человека показывает, что общий повтор специфичен для человека. Он присутствует у Homo heidelbergensis, Homo neanderthalensis и Homo denisova, но отсутствует (т.е. нарушен) у шимпанзе (у панискусов и троглодитов) (**Рисунок 15B**). Чтобы изучить распространенность общего повтора в современной человеческой популяции, проанализированы полные митохондриальные геномы из базы данных HmtDB [115] и реконструировалась филогения мтДНК человека (см. «Методы» выше). Замечено, что более 98% (42641 из 43437) секвенированных людей имеют идеальный общий повтор, в то время как менее 2% людей имеют нарушение повтора в результате однонуклеотидных замен в проксимальном плече (8470-8482 п.н., смотреть **Таблицу 7**).

Если нарушение общего повтора увеличивает продолжительность жизни, но не влияет на приспособленность, то он не подлежит отбору и считается эволюционно нейтральным. Однако нарушение общего повтора может также повысить приспособленность (i) напрямую – если носители более здоровы в репродуктивном возрасте и/или имеют более высокую фертильность, или (ii) косвенно, если увеличение продолжительности жизни родителей, бабушек и дедушек выгодно для потомства (эффект бабушки). Потенциальная важность эффекта бабушки для человеческой популяции обсуждалась в нескольких статьях [128, 129]. Интересуют же только прямые эффекты, которые могут принести эволюционную выгоду носителям нарушенного общего повтора. Прямые механизмы предполагают, что вариация делеционной нагрузки важна даже при низком уровне гетероплазмии, характерном для репродуктивного возраста (20-50 лет). Если это так, то, например, 5% против 10% гетероплазмии в постмитотических тканях носителей по сравнению с контрольной группой могут способствовать поддержанию этих тканей в более здоровых условиях (снижение саркопении, уменьшение нейродегенерации в репродуктивном возрасте) или, например, 1% против 3% гетероплазмии в ооцитах-носительницах по сравнению с контрольными может увеличить скорость оплодотворения носителей. Выявление этих слабых различий между случаями и контролем требует глубокой и крупномасштабной фенотипической характеристики обеих когорт, чего еще не было сделано. Дополнительная идея заключается в том, что средняя частота мутаций зародышевой линии у носителей может быть снижена из-за более низкой частоты делеций, более низкого окислительного стресса в ооцитах D4a и более низкой продукции АФК (активных форм кислорода). Если это так, то можно проверить этот эффект, проанализировав дерево мтДНК человека и сравнив частоту мутаций носителей с контрольной группой.

Чтобы проверить эту возможность, использованы все митохондриальные геномы человека, доступные в базе данных HmtDB [115], реконструировалась их филогения, определена гаплогруппа для каждого генома и проанализированы оба плеча общего повтора длиной 13 п.н. (см. дополнительные материалы). Как и ожидалось, наблюдается, что более 98% (42641 из 43437) геномов человека имеют совершенные общие прямые повторы, унаследованные от общего предка с шимпанзе. Среди носителей разорванного повтора разрушено было только проксимальное плечо (8469–8482 п.н.), тогда как дистальное плечо (13447–13459 п.н.) полностью сохранилось. Наиболее частые варианты разрыва проксимального отдела плеча представлены на **Таблице 7**. Среди них 399 случаев имеют замену m.8473T>C, которая происходила много раз независимо, что привело к образованию 5 больших кластеров с более чем 20 геномами каждый и множества малых кластеров (**Рисунок 16**). Наблюдение о том, что эта редкая замена m.8473T>C не является уникальной для D4a и отмечает по крайней мере четыре больших дополнительных поддерева в митохондриальном дереве человека, должно вызвать интерес как к долголетию, так и к потенциальному снижению предрасположенности к митохондриальным энцефаломиопатиям следующих гаплогрупп.: R2, U2e, H1c и U6a. В настоящее время в литературе не найдено доказательств увеличения продолжительности жизни этих гаплогрупп. Важно отметить, что информативны и другие

варианты нарушенного общего повтора; например, гаплогруппа D5a с нарушенным общим повтором m.8479A>G (см. **Таблицу 7**) связано с увеличением продолжительности жизни [28].

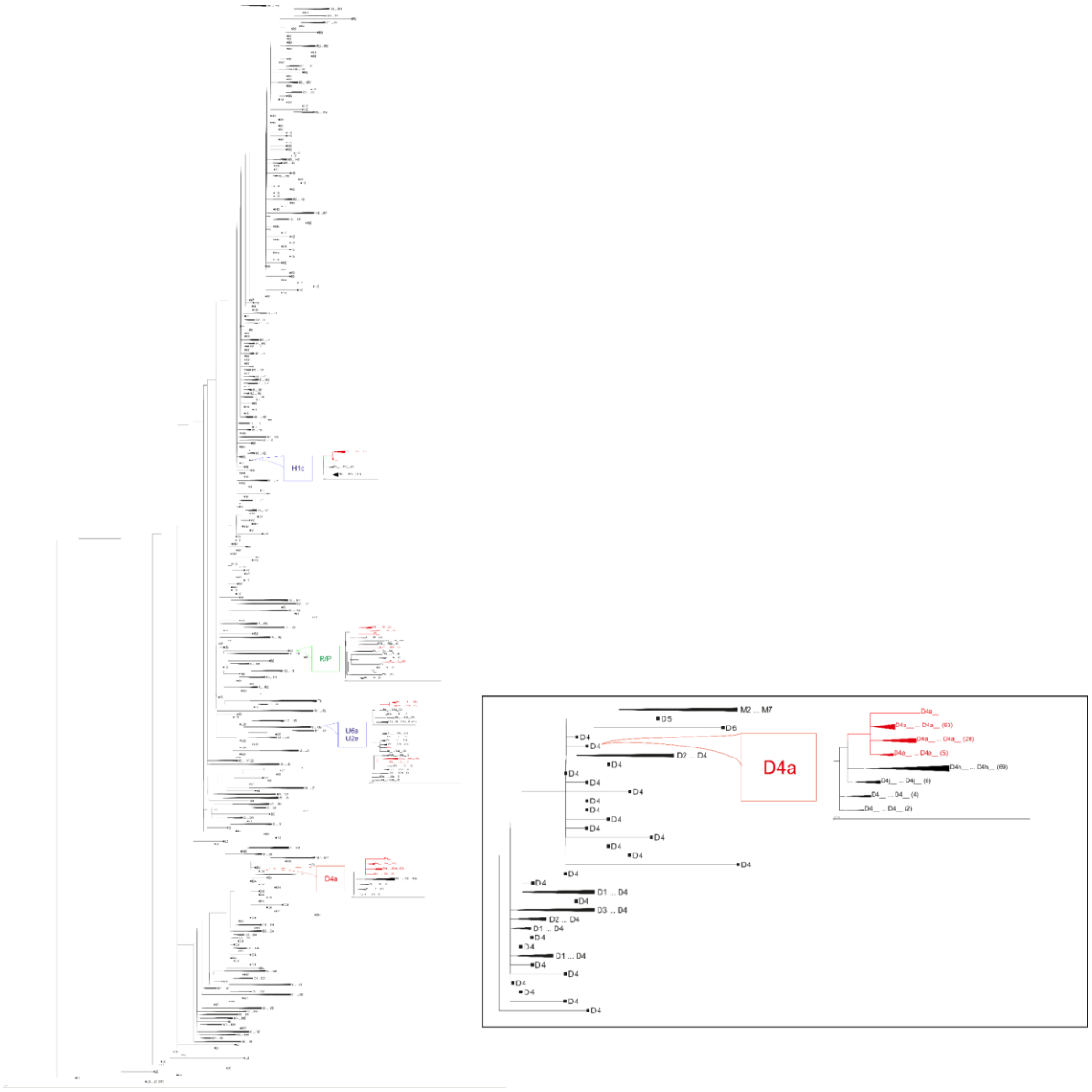

Рисунок 16 Упрощенное (со свернутыми кладами) филогенетическое дерево 43437 геномов мтДНК человека с отмеченными кладами, содержащими последовательности с нарушенным общим повтором (случаи) и последовательности, в которых оба плеча общего повтора нетронуты (контроли).

Чтобы оценить потенциальное влияние разрушенного повтора на частоту нуклеотидных замен зародышевой линии, внимание сосредоточено на пяти гаплогруппах с заменой m.8473T>C (далее — случаи) и назначено каждой из них ближайшее сестринское поддерево в качестве контроля. Затем используется модифицированный тест отношения правдоподобия. Аппроксимируется частота мутаций длиной ветвей от общего предка (предка как кейса, так и контроля) до терминальных концов случаев и контроля, используя три подмножества позиций со слабыми ограничениями (с высокими частотами вариантов аллелей в человеческой популяции) и применяя четыре матрицы замены (дополнительные материалы). Не наблюдалось универсальной тенденции: гаплогруппа D4a продемонстрировала пониженную скорость замещения, U6a и U2e продемонстрировали повышенную скорость замещения, а R2 и H1c вообще не показали никакого эффекта (**Таблица 8**). Снижение частоты мутаций зародышевой линии в D4a очень интересно и может отражать не только разрушение повтора мтДНК, обсуждаемое в этой статье, но и ядерные (POLG, TWINKLE, связанные с гаплогруппой), а

также факторы на снижение частоты делеций и увеличение продолжительности жизни. Таким образом, стоит продолжить анализ снижения частоты мутаций зародышевой и соматической мтДНК в D4a и других гаплогруппах мтДНК, связанных с увеличением продолжительности жизни. Однако в целом в этом пилотном исследовании сделан вывод, что нет никаких доказательств, подтверждающих снижение частоты мутаций во всех поддеревьях с нарушенным общим повтором, и, следовательно, нет никаких доказательств того, что нарушение общего повтора является эволюционно полезным *per se*. Однако хотелось бы подчеркнуть, что более глубокое фенотипическое описание (возникновение митохондриальных заболеваний, таких как саркопения, нейрогенерация и т. д.) всех гаплогрупп с нарушенным общим повтором (**Таблица 8**) может пролить свет на потенциальные преимущества этих замен [130].

Таблица 8. Сравнение случай-контроль эволюционных скоростей геномов мтДНК, содержащих нарушенный общий повтор (случай), с геномами, имеющими оба плеча общего повтора неповрежденными (контроль).

Модель замещения	Вариации сайтов, %	Статистическое свойство эволюционных скоростей в случае, тест знаковых рангов Вилкоксона	Состав клады				
			U6a	U2e	H1c	R/P	**D4**
F81	0.05	shift	greater	greater	greater	less	**less**
		average p	2.17E-05	9.73E-10	0.00176	0.062878	**7.75E-07**
		std.dev. of p	5.16E-05	1.48E-09	0.004146	0.091994	**3.22E-06**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.00135	3.41E-12	0.04709	0.478095	**1.23E-06**
		std.dev. of p	0.001961	7.19E-12	0.06996	0.207913	**4.38E-06**
	0.5	shift	less	greater	less	greater	**less**
		average p	0.198667	0.058184	0.034564	0.281155	**2.88E-06**
		std.dev. of p	0.171724	0.060021	0.033213	0.142158	**6.9E-06**
HKY	0.05	shift	greater	greater	greater	less	**less**
		average p	4.27E-05	8.09E-12	0.04157	0.06907	**2.07E-06**
		std.dev. of p	9.7E-05	1.19E-11	0.057705	0.091293	**5.74E-06**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.005236	5.22E-12	0.049723	0.293431	**7.91E-07**
		std.dev. of p	0.013786	6.58E-12	0.05822	0.172256	**2.48E-06**
	0.5	shift	less	less	less	greater	**less**
		average p	0.273149	0.284463	0.001448	0.418316	**5.53E-07**
		std.dev. of p	0.209097	0.141528	0.003522	0.236519	**1.02E-06**
JC	0.05	shift	greater	greater	greater	less	**less**
		average p	4E-05	6.89E-12	0.023705	0.0664	**4.62E-07**
		std.dev. of p	8.81E-05	9.34E-12	0.05136	0.071869	**1.22E-06**

	0.1	shift	greater	greater	greater	less	**less**
		average p	0.000618	2.35E-11	0.044148	0.5305	**2.71E-07**
		std.dev. of p	0.001676	4.55E-11	0.062387	0.214982	**6.51E-07**
	0.5	shift	less	greater	less	less	**less**
		average p	0.039706	0.237216	0.009861	0.461538	**8.35E-06**
		std.dev. of p	0.060181	0.186655	0.014167	0.266469	**1.69E-05**
K80	0.05	shift	greater	greater	greater	less	**less**
		average p	9.73E-06	1.66E-11	0.031127	0.087671	**1.14E-07**
		std.dev. of p	7.1E-06	5.27E-11	0.025872	0.146845	**2.04E-07**
	0.1	shift	greater	greater	greater	greater	**less**
		average p	0.000228	6.21E-10	0.060003	0.369783	**3.65E-07**
		std.dev. of p	0.000542	1.61E-09	0.106046	0.204312	**1.19E-06**
	0.5	shift	less	greater	less	greater	**less**
		average p	0.017182	0.189886	0.02839	0.554828	**3.98E-06**
		std.dev. of p	0.019592	0.11964	0.043765	0.255987	**9.78E-06**

4.3.3 Факторы не позволяющие напрямую связать скорость эволюции мтДНК гаплогруппы D4a с продолжительностью жизни

Сравнение скоростей митохондриальных замен между поддеревьями с нарушенным и ненарушенным общим повтором не позволяет надежно разделить эффекты отбора и скорости мутирования из-за влияния других факторов. Наблюдаемые различия между гаплогруппами (например, повышенное долголетие гаплогруппы D4a) могут быть обусловлены не столько наличием повтора, сколько совокупностью сторонних факторов. К ним относятся особенности медицины, рациона питания, климата, культурных практик и генетического фона, не связанного с митохондриями. Существующие выборки долгожителей слишком малы и подвержены влиянию специфической социальной среды в которой они обеспечены особым отношением, что не позволяет сформировать адекватные контрольные группы. Для проверки гипотезы требуется масштабное когортное исследование носителей D4a за пределами их традиционного ареала. Также, в случае верности гипотезы, постмутационный эффект снижения скорости замен следует ожидать и в соматических тканях, а не только в зародышевой линии.

4.3.4 Нет доказательств отрицательного отбора против прямых повторов у млекопитающих-долгожителей

Также изучено как количество повторов в мтДНК разных млекопитающих связано с продолжительностью их жизни и обнаружено, что у долгоживущих животных может не быть негативного отбора против повторов, и скорее всего это связано с различиями в нуклеотидном составе. При сравнении реальных и смоделированных геномов млекопитающих не найдено доказательств прямого отбора против повторов — это означает, что количество повторов не влияет на выживание животного напрямую.

В предыдущей главе не удалось найти доказательств отбора, способствующего нарушению общего повтора в человеческой популяции. Здесь хотелось бы экстраполировать эту логику на все виды млекопитающих и подвергнуть сомнению несколько исследований, утверждающих, что отрицательная корреляция между количеством прямых повторов и

продолжительностью жизни предполагает отрицательный отбор против прямых повторов у долгоживущих млекопитающих. Можно предположить, что отрицательная корреляция между повторами и долголетием может возникнуть в результате увеличения количества прямых повторов у короткоживущих млекопитающих из-за их более асимметричного нуклеотидного состава. Стоит сфокусироваться на содержании нуклеотидов мтДНК как на сильном потенциальном искажающем эффекте, который может объяснить большинство предыдущих результатов. Ожидается, что в случайной нуклеотидной последовательности с одинаковым содержанием нуклеотидов (A, T, G и C по 25%) обилие повторов будет минимальным, и как только частота нуклеотидов отклоняется от 25%, вероятность возникновения повторов возрастает выше из-за чисто комбинаторной природы этих повторов (в крайнем случае, если вся последовательность состоит из одного и того же нуклеотида, весь геном будет покрыт повторами). Чтобы визуализировать этот эффект и проверить потенциальную силу этого искажающего фактора, выполнен простой *in silico* эксперимент, в котором смоделированы случайные нуклеотидные последовательности длиной 16000 пар оснований и различным содержанием нуклеотидов (изменяя частоту одного нуклеотида от 10 до 50% и сохраняя все три других нуклеотида с теми же частотами) и оценивалось для них обилие прямых повторов, как и ранее [22] (**Рисунок 17A**). Наблюдается, что действительно минимальное содержание повторов соответствует 25%, а любые отклонения от этой частоты приводят к увеличению содержания прямых повторов. Влияние содержания нуклеотидов на количество повторов очень сильное, и, таким образом, этот фактор может иметь большое значение в мтДНК млекопитающих с сильно смещенным содержанием нуклеотидов. Например, было показано, что продолжительность жизни млекопитающих положительно связана с содержанием GC в мтДНК, что можно объяснить силами отбора [50] или мутационной предвзятостью [114]. Независимо от объяснения систематической ошибки, можно предположить, что чем выше отклонение от 25%, тем выше количество случайно ожидаемых прямых повторов. Ниже проверяется важность нуклеотидного состава, используя два подхода: (i) случайную перетасовку и (ii) множественные линейные модели.

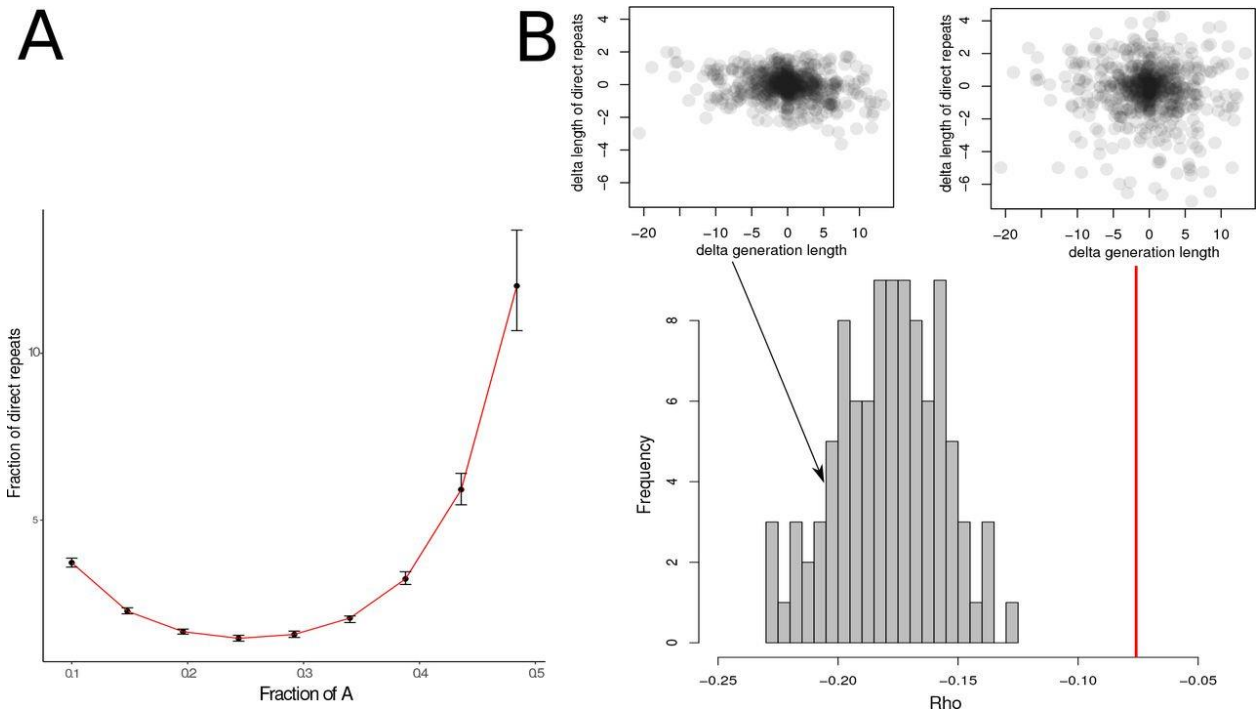

Рисунок 17. Содержание нуклеотидов может быть сильным фактором, мешающим анализу корреляции между повторами нуклеотидов и продолжительностью жизни. a. Моделируемые геномы и их нагрузка прямыми повторами. Ось X отражает частоту нуклеотида A, в то время как T, G и C всегда имели одинаковую частоту = (1-fr(A))/3. Ось Y отражает долю генома, покрытую прямыми повторами. b. Отрицательная корреляция между длиной

поколения и распространенностью повторов может быть сформирована исключительно смещением содержания нуклеотидов. наблюдаемое против ожидаемого

Прежде всего, используя 705 видов млекопитающих с секвенированным полным митохондриальным геномом и известной длиной поколения [131] проведен корреляционный анализ между длиной поколения и долей генома, покрытой прямыми повторами [110]. Как длина поколения, так и количество прямых повторов нормализовалось с использованием филогенетически независимых контрастов [132]. Наблюдалась слабая отрицательная корреляция между длиной поколения и прямыми повторами (rho Спирмена = -0,076, значение p = 0,04, **Рисунок 17B** красная вертикальная линия). Далее меня интересовал основной фактор этой корреляции: либо содержание нуклеотидов играет основную роль в этой корреляции (короткоживущие млекопитающие более богаты A, и это увеличит количество случайно ожидаемых повторов у короткоживущих млекопитающих), либо отрицательный отбор против прямых повторы у долгоживущих млекопитающих также задействованных (это уменьшит количество повторов у долгоживущих млекопитающих)? Чтобы подойти к этому вопросу, 100 раз случайным образом перетасованы все геномы млекопитающих, сохранив исходное содержание нуклеотидов и исходную длину поколения, и проверили, существует ли корреляция между количеством прямых повторов и длиной поколения. Интересно, что все 100 корреляций, основанных на перетасованных последовательностях, были намного сильнее по сравнению с реальной (**Рисунок 17B**). Это означает, что видоспецифическое содержание нуклеотидов, связанное с длиной поколения, имеет достаточно сильный эффект, чтобы искусственно создать отрицательную корреляцию между повторами и длиной поколения. Используя этот анализ, нельзя утверждать об отсутствии отбора против прямых повторов у долгоживущих млекопитающих, но можно продемонстрировать, что содержание нуклеотидов является чрезвычайно важным фактором, который приводит к сильной отрицательной корреляции без какого-либо отбора, связанного с долголетием. Хотелось бы подчеркнуть, что в использованном подходе к перетасовке не сравнивались напрямую реальные последовательности с перетасованными, перетасованные последовательности использовались только для того, чтобы продемонстрировать, насколько важным может быть нуклеотидный состав в формировании количества прямых повторов.

Во-вторых, выполнено несколько линейных моделей, в которых количество прямых повторов было объяснено как функция длины поколения и содержания нуклеотидов (**Таблица 9**). В большинстве моделей влияние длины поколения было незначительным (маргинально значимым), тогда как содержание нуклеотидов было почти всегда значимым и, что важно отметить, для нуклеотидов A и T, средняя частота которых превышает 25%, коэффициенты являются положительными, тогда как для нуклеотидов G и C, частота которых составляет менее 25%, коэффициенты были отрицательными. Другими словами, увеличение доли редких нуклеотидов связано с уменьшением количества прямых повторов, а увеличение доли частых нуклеотидов связано с увеличением количества прямых повторов (**Рисунок 17A**). Этот результат полностью соответствует представлению о том, что количество прямых повторов в основном определяется содержанием нуклеотидов и, следовательно, кажется нейтральным, а не подвергающимся сильному отрицательному отбору.

Таблица 9. Результаты множественной линейной модели: доля генома, покрытая прямыми повторами, как функция длины поколения и нуклеотидного состава. Все значения нормализованы по PIC (кроме модели 0). Все значения (доля генома, покрытая прямыми повторами; длина поколения; нуклеотидный состав) преобразованы по логарифму 2. Количество проанализированных видов млекопитающих — 705.

	Интерсепт (коэффициент, значение p) ноль означает, что регрессия была вынуждена пройти через начало координат	Коэффициент длины поколения (p-value)	нуклеотид: коэффициент (p-value)
model 0 (without PIC norm.)	-1.18028 (< 2e-16)	**-0.03930** (0.000784)	NA
model 1.A	-0.20643 (0.0317)	**-0.01349** (0.2970)	NA
model 1.B	zero	**-0.01299** (0.317)	NA
model 2.A	-3.487e-02 (0.07895)	**-0.005679** (0.6641)	A: 0.947124 (0.0013)
model 2.B	zero	**-0.005294** (0.68642)	A: 0.930128 (0.00162)
model 3.A	-0.189388 (0.0490)	**0.001839** (0.9011)	T: 0.272742 (0.0347)
model 3.B	zero	**0.003499** (0.8132)	T: 0.294066 (0.0226)
model 4.A	-0.20438 (0.0337)	**-0.01170** (0.3868)	G: -0.08649 (0.6428)
model 4.B	zero	**-0.01082** (0.424)	G: -0.10474 (0.575)
model 5.A	-0.18158 (0.0556)	**0.02153** (0.1469)	C: -0.54527 (4.39e-06)
model 5.B	zero	**0.02280** (0.125)	C: -0.55808 (2.64e-06)

4.4 Вывод

Обнаружено, что существует три различных типа изменений общего повтора, которые часто происходят у людей, и все они связаны с увеличением продолжительности жизни. Интересно, что самый длинный (и, соответственно, самый тяжелый) прямой повтор в митохондриальном геноме человека (общий нуклеотидный повтор из 13 пар оснований, наблюдаемый у 98% человеческой популяции) был разрушен точечной синонимичной мутацией m.8473T>C в японской гаплогруппе D4a, известной избытком долгожителей (людей, живущих 100 и более лет) и сверхдолгожителей (людей, живущих 110 и более лет) [27, 28]. Таким образом, можно предположить, что нарушение общего повтора благотворно влияет на гаплогруппу D4a, поскольку снижает вероятность того, что соответствующая соматическая делеция (которую еще называют «общей» делецией, поскольку она очень часто появляется в старых постмитотических тканях) возникает в течение жизни и таким образом, это может отсрочить нейродегенерацию и саркопению, объясняя, по крайней мере частично, чрезвычайную долговечность гаплогруппы D4a [26].

В соответствии с выдвинутой гипотезой [26], не обнаружено общих делеций в образцах N1b [22]. Следовательно, разрушение повтора длиной 13 п.н. даже однонуклеотидным вариантом зародышевой линии может значительно уменьшить образование общей соматической делеции. Таким образом, можно предположить, что нарушение общего повтора аннулирует общую делецию (которая является наиболее частой среди всех соматических делеций) и, по крайней мере частично, может способствовать чрезвычайному долголетию японской гаплогрупп D4a и D5a. Понимание механизмов образования делеций позволяет прогнозировать генетические риски делеций мтДНК для различных гаплогрупп человека.

Стоит учесть множество других факторов которые могут влиять на продолжительность жизни японских гаплогрупп, а также выбрать подходящие контроли для сравнительного анализа.

Возможно потенциальное применение гаплогруппы с нарушенным общим повтором в современной технологии донорства митохондрий, а также имеется возможность нарушения общего повтора в будущем. Остается вопрос, можно ли считать потерю повтора полезной мутацией с эволюционной точки зрения, и не наблюдается никаких текущих доказательств, подтверждающих это. Сделанный вывод можно экстраполировать на все виды млекопитающих, ставя под сомнение существование очищающего отбора против прямых повторов у долгоживущих млекопитающих.

Глава 5. Взаимодействие прямых и инвертированных повторов при образовании делеции

5.1 Проблематика

Можно предположить, что помимо прямых повторов инвертированные повторы также могут играть важную роль в структурировании мтДНК. Используя коллекцию делеций мтДНК человека, а также глобальные и локальные свойства мтДНК (распределение прямых и инвертированных повторов), пересматриваются риски соматических делеций мтДНК и обнаруживаем, что, помимо прямых повторов, которые, как известно, влияют на делеции, существует сильное влияние вторичной структуры одноцепочечной мтДНК во время репликации. Вторичная структура формируется инвертированными повторами, которые могут образовывать стебли и сокращать эффективное расстояние между двумя прямыми повторами, тем самым увеличивая вероятность делеций.

Было показано, что процесс проскальзывания репликации [34, 33, 133], который также называют рекомбинацией выбора копии, влияет на образование делеций у бактерий [33**,** 133] и в экспериментальных шаблонах, имитирующих мтДНК *in vitro* [34]. Этот процесс может зависеть от пространственной организации ДНК. Процесс проскальзывания репликации основан на (i) остановке репликации из-за вторичной структуры, образованной инвертированными повторами одноцепочечной ДНК, (ii) диссоциации вновь синтезированного (зарождающегося) проксимального плеча прямого повтора и (iii) перевыравнивании зарождающегося плеча к дистальному родительскому (**Рисунки 12 и 18**). Предполагаемая роль инвертированных повторов в этом процессе заключается в приостановке репликационной вилки и уменьшении эффективного расстояния между двумя плечами прямого повтора, что увеличивает шансы на успешное выравнивание и формирование делеции.

5.2 Методы

Плотность прямых и инвертированных повторов: Для каждой пары окон размером 100 п.н. оцененено количество нуклеотидов, участвующих хотя бы в одном прямом или инвертированном деградированном повторе. Соответствующий повтор должен иметь одно плечо, расположенное в первом окне, и другое плечо, расположенное во втором окне. Все деградированные (с максимальным уровнем несовершенства 20%) повторы в мтДНК человека были найдены с помощью разработанного мной алгоритма [110].

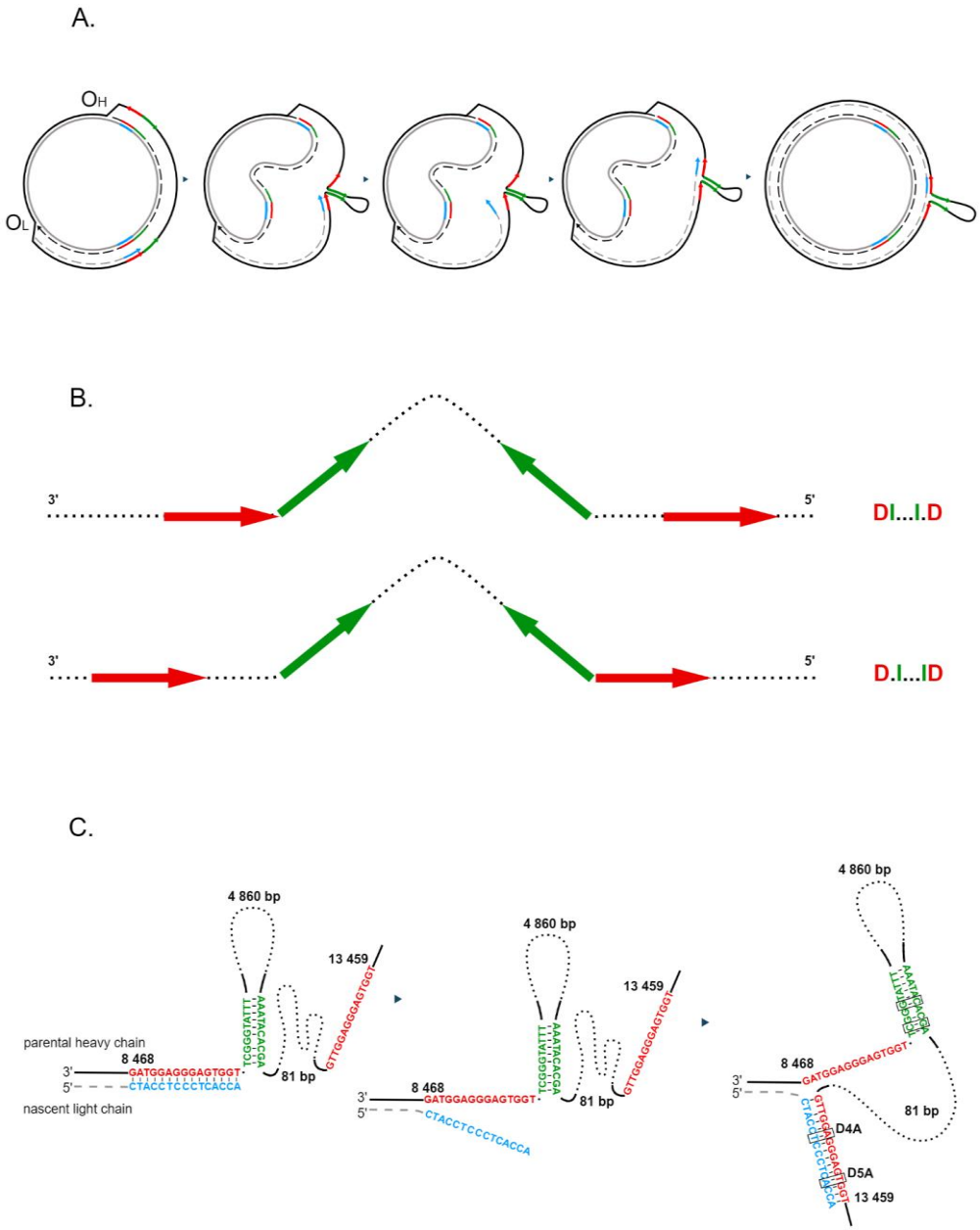

Рисунок 18. Проскальзывание репликации может объяснить образование общей делеции мтДНК, где вложенные инвертированные (зеленые) и прямые (красные) повторы играют решающую роль во время репликации отстающей (родительской тяжелой цепи) цепи. А. Основные этапы образования делеции. Первое: согласно асимметричной репликации мтДНК отстающая цепь (сплошная черная линия) провела значительное количество времени в одноцепочечном состоянии. Второе: одноцепочечная родительская тяжелая цепь образует вторичную структуру, сформированную инвертированными повторами. Этот стебель может останавливать репликационную вилку. Третье: остановка репликации приводит к частичной диссоциации недавно синтезированного проксимального плеча прямого повтора (синяя стрелка). Четвертое: диссоциированная цепь перестраивается на ближайшее дистальное плечо прямого повтора. Пятый: репликация похожа, но вновь синтезированная легкая цепь (пунктирная серая линия) имеет делецию - как инвертированные повторы, так и ДНК между ними, а также проксимальное плечо прямого повтора удалены; В. Общий повтор (красный) имеет вложенный инвертированный повтор (зеленый), что может объяснить высокую изменчивость общего повтора. Цветовые коды такие же, как на рис. 18А. Первый: синтез зарождающейся легкой цепи (синий) и остановка репликации около стебля, образованного инвертированным повтором. Второй: частичная диссоциация зарождающегося проксимального плеча общего повтора. Третий: перестройка вновь синтезированного проксимального плеча общего повтора с дистальным плечом общего повтора и продолжение репликации. Наиболее частые полиморфизмы (со 100 или более мтДНК из HmtDB), нарушающие эти повторы, отмечены прямоугольниками. Эти замены могут существенно замедлить процесс образования делеции. Распространенный повтор здесь расширен от классического идеального повтора длиной 13 п.н. (ACCTCCCTCACCA) до деградированного повтора длиной 15 п.н. с одним несовпадением (CtACCTCCCTCACCA).

Колокализация деградированных прямых и инвертированных повторов: Для всех анализов использовалась созданная база данных со всеми деградированными повторами мтДНК человека (с минимальным уровнем сходства 80%) длиной 10 п.н. и более [110]. В этом наборе данных имеется 2957 прямых и 764 инвертированных повтора в большой дуге. 207 прямых повторов, оба плеча которых расположены близко друг к другу, были исключены из анализа, поскольку у них нет невложенных инвертированных повторов. Остальные 2750 прямых повторов (DD), содержащие по крайней мере одну пару вложенных инвертированных повторов, были использованы в последующих анализах. Для каждой такой пары DD выполняется поиск всех пар инвертированных повторов, расположенных между прямыми, и находится проксимальный зазор как минимальное расстояние между первым плечом прямого повтора и первым плечом инвертированного повтора. Медиана проксимальных промежутков равна 21. Аналогично для каждой такой пары DD найден дистальный промежуток как минимальное расстояние между вторым плечом инвертированного повтора и вторым плечом прямого повтора. Медиана дистальных промежутков равна 25.

5.3 Результаты

5.3.1 Делеции происходят чаще, если прямые повторы вложены в инвертированные: комбинации DI...ID и их свойства

Учитывая, что внутренняя гомология мтДНК поддерживается в основном за счет прямых повторов, а пространственная структура поддерживается за счет инвертированных повторов, введено понятие DIID (Прямой-Инвертированный-Инвертированный-Прямой), которое относится к вложенному паттерну прямых и инвертированных повторов и является самая хрупкая структура мтДНК. Чтобы проверить участие проскальзывания репликации в формировании делеций мтДНК *in vivo,* выбраны все комбинации повторов DIID, расположенные внутри большой дуги. Используя созданную базу данных [110] всех вырожденных повторов мтДНК человека (2957 прямых и 764 обратных повтора в пределах большой дуги, см. Методы выше), найдены вложенные комбинации повторов. После экспериментальных исследований, которые подтверждают механизм проскальзывания репликации [34, 33], можно предположить, что проксимальные прямые и инвертированные повторы должны располагаться близко друг к другу (максимальное расстояние 10 п.н.), тогда как дистальные инвертированный и прямой повторы могут располагаться на расстоянии до 1000 п.н. друг от друга. Чтобы подчеркнуть это условие, в дальнейшем будет использоваться обозначение DI...ID. В пределах большой дуги наблюдалось 260 комбинаций DI...ID и 2697 обычных прямых повторов (DD) без вложенных инвертированных повторов. Далее был поставлен вопрос, действительно ли мутагенный потенциал комбинаций DI...ID выше, чем у обычного DD. Сравнивая долю комбинаций DI...ID, связанных с делециями (7 были ассоциированы и 253 не были связаны) и доли DD комбинации, связанные с делециями (20 были связаны и 2677 не были связаны), было сделано наблюдение что мутагенный потенциал комбинации DI...ID действительно выше (отношение шансов Фишера = 3,7, p-value = 0,0072). Повышенный мутагенный потенциал DI...I.D еще более выражен в потенциальной зоне контакта (DI...ID: 6/34, DD: 12/438; отношение шансов Фишера = 6,4; p-value = 0,0018).

Увеличение мутагенности из комбинаций DI...I.D внутри потенциальной зоны контакта позволяют предположить, что оба уровня вторичной структуры мтДНК могут быть важны: глобальная складка, интегрированная для всей основной дуги (**Рисунок 9**), и локальная специфичная для комбинации повторов DI...I.D (**Рисунки 12 и 18**). Далее было проверено, можно ли объяснить потенциальную зону контакта повышенной плотностью комбинаций DI...I.D. Принимая во внимание, что площадь потенциальной зоны контакта (3кб*3кб) составляет около 18% от основной дуги (10кб*10КБ/2 = 50), наблюдалась пропорциональная

доля DI...ID - 15% (40 комбинаций DI...ID в потенциальной контактной зоне из 260 в большой дуге). Это предполагает, что существование потенциальной зоны контакта не может быть объяснено за счет увеличения плотности комбинаций DI...I.D. Напротив, повышенная мутагенность DI...I.D в потенциальной зоне контакта предполагает наличие дополнительного, вероятно, структурного фактора, поддерживающего повышенную мутагенность зоны потенциального контакта одноцепочечной ДНК (**Рисунок 9**).

Чтобы дополнительно подтвердить гипотезу о том что большинство делеций происходит во время репликации отстающей цепи [34] сравнен мутагенный потенциал комбинаций DI...I.D (короткое расстояние, максимум 10 п.н., между проксимальными прямыми и инвертированными повторами и потенциально высокий, до 1000 п.н., между дистальными прямыми и инвертированными повторами) и комбинаций D.I...ID (потенциально длинное расстояние - до 1000 п.н. между проксимальными прямыми и инвертированными повторами и короткое расстояние - максимум 10 п.н. между дистальными инвертированными и прямыми повторами) (**Рисунок 18B**). Логика этого анализа заключается в том, что паттерн DI...ID на отстающей цепи эквивалентен паттерну DI...ID на ведущей цепи, и если обе цепи одинаково мутагенны, ожидается, что обе комбинации будут одинаково связанны с делециями (**Рисунок 18B**). Анализируя связь D.I...ID с делециями, было сделано наблюдение что отношение шансов Фишера 2,61 (DI...ID: 5/234, DD: 22/2696; отношение Фишера = 2,61; p=0,0615), что было менее 3,7 наблюдаемого для DI...ID (см. выше). Важно подчеркнуть что указанные здесь и выше отношения шансов (OR) верны для референсной мтДНК, а также гаплогрупп близких к ней, но могут измениться в других гаплогруппах. D.I...ID в потенциальной контактной зоне вообще не связан с делециями (D.I...ID: 3/44, DD: 15/428; коэффициент Фишера-Одда = 1,9421, p-value = 0,4), тогда как DI...I.D продемонстрировал сильный сигнал (см. выше).

Полученные результаты соответствуют первоначальному выводу о том, что большинство делеций происходит во время репликации на отстающей цепи [34].

Если комбинации DI...I.D являются сильно мутагенными, ожидается, что очищающий отбор устранит их, сохраняя количество наблюдаемых комбинаций DI...I.D меньше, чем случайно ожидалось. Совершая перестановку 2957 прямых и 764 обратных повторов внутри большой дуги, получено медианное значение 365 комбинаций DI...I.D, что на 29% выше, чем 260 наблюдаемых (p <0,0001, 10000 перестановок). В пределах потенциальной зоны контакта результаты были схожими: перестановки показали медианное значение 54 комбинаций DI...I.D, что на 25% выше, чем 40 наблюдаемых (p=0,0243, 10000 перестановок). Уменьшение количества наблюдаемых по сравнению с ожидаемым комбинаций DI...I.D можно использовать как признак очищающего отбора, устраняющего эти мутагенные комбинации. Важно подчеркнуть что описанное в данном разделе имитационное моделирование является абстрактным численным экспериментом который основан на перестановках приводящих к потере функциональности мтДНК.

На эффект прямых повторов может влиять не один, а множество инвертированных повторов: наиболее проксимальная пара инвертированных повторов будет влиять на приостановку репликации, а дистальная пара инвертированных повторов в свою очередь будет уменьшать эффективное расстояние между прямыми повторами. Чтобы протестировать этот расширенный сценарий, для каждой пары прямых повторов определены проксимальные и дистальные промежутки как соответствующие расстояния между прямыми и инвертированными повторами (см. «Методы» выше). Далее, регрессируя мутагенность прямых повторов (оцениваемую как количество связанных делеций, см. Методы) в зависимости от разрывов, получен следующий результат: мутагенность = -2,5 - 1*(проксимальный разрыв) - 0,5*(дистальный разрыв); N = 2750, все коэффициенты масштабированы, все значения p < 2.e-06, регрессия Пуассона. Результаты показывают, что риск делеции увеличивается, когда инвертированные повторы расположены близко к прямым и проксимальный разрыв в два раза важнее дистального.

Продемонстрировано, что колокализация прямых и вложенных инвертированных повторов может формировать распределение делеций. Следующий вопрос – объясняется ли

повышенная мутагенность прямых повторов внутри потенциальной зоны контакта более плотным паттерном этой колокализации? Замечено, что как проксимальные, так и дистальные промежутки короче в пределах потенциальной зоны контакта. Это демонстрирует локальный эффект инвертированных повторов – они модулируют мутагенность прямых повторов двумя способами: приостанавливают репликацию вблизи первого плеча повтора и уменьшают эффективное расстояние. Приостановка репликации – это специфический процесс, уникальный для каждого прямого повтора, а уменьшение эффективного расстояния – общее для всех прямых повторов.

5.3.2 Общий прямой повтор имеет вложенный инвертированный повтор

В комбинациях DI...ID в большой дуге участвует 8,8% прямых повторов (260/2957). Общий повтор наблюдался среди 2,3% комбинаций DI...ID, дополнительно связанных с делециями (7/260) (**Рисунок 18C**). Можно ппредположить что оба плеча общего прямого повтора могут находиться в пространственной близости за счет как глобальных, так и локальных структур. Деградированный инвертированный повтор длиной 10 п.н., связанный с общим повтором (8484-8493: AGCcCATAAA; 13354-13363: TTTATGtGCT), обладает потенциалом для формирования структуры стебля, которая может остановить репликационную вилку и уменьшить эффективное расстояние между плечами прямого повтора от примерно от 5 КБ (13447-8482=4965) до менее 100 пар оснований (**Рисунок 18C**). Обе эти роли вложенного инвертированного повтора могут быть ключевыми в объяснении высокого мутагенного потенциала общего повтора (**Рисунок 18B**). Разные гаплогруппы имеют несовпадения по общему повтору или по инвертированному. Несмотря на то, что эти замены достаточно редки, иногда оба типа несовпадений существуют в одном и том же митохондриальном геноме. Например, некоторые представители гаплогруппы R22 (идентификаторы GenBank: KP346026.1, KP346018.1), распространенные преимущественно в Никобарский архипелаг [134], а также встречается в южно-центральной Индонезии и Вьетнаме [135], содержат одну синонимичную замену, нарушающую общий повтор (m.8473T>C) и еще одна синонимичная замена, разрушающая инвертированный повтор (m.13359G>A). Ожидается эпистатическое взаимодействие между двумя этими несоответствиями: дестабилизированный стебель вряд ли остановит репликационную вилку, тогда как деградированный общий повтор вряд ли, с вероятностью, уменьшенной на один порядок [33], позволит перестроить вновь синтезированное проксимальное плечо к дистальному плечу родительской цепи (**Рисунок 18C**). В целом оба эти несоответствия как ожидается, приведут к исчезновению общей делеции. Исследование делеционной нагрузки в тканях соответствующего возраста этих и других интересных гаплогрупп поможет установить прогноз генетических рисков соматических делеций мтДНК.

5.4 Вывод

Независимо от шаблона поведения глобальной петли во время репликации, описанного выше, был отмечен также локальный шаблон. Общий повтор участвует в комбинации повторов DIID, которая представляет собой хрупкую структуру, предрасположенную к образованию делеций. Предполагаемая роль инвертированных повторов в этом процессе состоит в том, чтобы приостановить репликационную вилку и уменьшить эффективное расстояние между двумя плечами прямого повтора, что увеличивает шансы на успешное перестроение. Деградированный инвертированный повтор связанный с общим повтором потенциально может образовывать структуру стебля, которая может остановить репликационную вилку и уменьшить эффективное расстояние между плечами прямых повторов. Обе эти роли вложенных инвертированных повторов имеют решающее значение для объяснения высокого мутагенного потенциала общего повтора.

Полученные результаты совместимы с механизмом проскальзывания репликации **[34, 33]** когда вложенный паттерн прямых и инвертированных повторов (DI...ID) приводит к образованию делеций (**Рисунки 11 и 17**). Показано, что в мтДНК человека комбинации DI...I.D действительно более мутагенны по сравнению с обычными паттернами DD, и комбинации DI...I.D потенциально развиваются под давлением очищающего отбора. Ожидается, что аналогичные эффекты отбора против DI...ID будут очевидны в эволюционном и межвидовом масштабе. Первоначально было показано, что продолжительность жизни млекопитающих отрицательно коррелирует с обилием прямых повторов в мтДНК [24, 25], хотя позже было показано, что инвертированные повторы демонстрируют еще более сильную отрицательную корреляцию с продолжительностью жизни млекопитающих [51]. Недавно было показано, что оба типа повторов: прямые и инвертированные достаточно хорошо коррелируют друг с другом [110].

Если соображения, приведенные выше, верны и нарушение общего повтора и других повторов из комбинаций DI...I.D полезно для пострепродуктивного состояния здоровья человека и старения, оно все равно может быть эволюционно нейтральным, поскольку увеличение продолжительности жизни не обязательно связано с длительным репродуктивным периодом и/или повышенным количеством потомства. Таким образом, согласно нулевой гипотезе: синонимичные варианты, затрагивающие повторы из комбинаций DI...ID, нейтральны. Однако возможность того, что нарушение DI…I.D связано с некоторыми преимуществами приспособленности, весьма интригует и заслуживает дальнейшего изучения. Например, нарушение DI...I.D может повысить приспособленность (i) непосредственно, если носители более здоровы в репродуктивном возрасте и, таким образом, имеют более высокую рождаемость, или (ii) косвенно, если увеличение продолжительности жизни родителей, бабушек и дедушек выгодно для потомства из-за эффекта бабушки [128, 129]. В настоящее время потенциальная выгодная потеря комбинаций DI...I.D подтверждается (i) дефицитом наблюдаемых комбинаций DI...I.D в мтДНК человека по сравнению с ожидаемым и (ii) отрицательными корреляциями между обилием повторов и продолжительностью жизни млекопитающих [24, 25, 51]. Необходимы дальнейшие тесты, чтобы проверить эволюционное преимущество потери комбинации DI...ID у людей и других долгоживущих млекопитающих.

Выдвинутая гипотеза подтверждается несколькими экспериментами. Первый заметный эксперимент был проведен на мтДНК Nematomorpha, которая имеет высокий уровень совершенных инвертированных повторов значительной длины [136]. Исследование показало, что инвертированные повторы могут образовывать шпильки и влиять на репликацию ДНК в ПЦР (полимеразной цепной реакции). Также было показано, что паттерн DIID исчез во время ПЦР. Предположительно, что более короткие продукты (мтДНК с делецией), вероятно, являются результатом скачка ПЦР, чему способствует наличие прямых повторов, фланкирующих шпильку. Это демонстрирует, что паттерн DIID действительно является сильно мутагенным и может привести к образованию делеции (см. Дополнительный рисунок 2 в статье [136]). Второй эксперимент на мтДНК человека показал, что репликативные полимеразы могут вызывать делеции посредством рекомбинации с выбором копии между прямыми повторами и что этот эффект усиливается вторичными структурами [34], которые поддерживаются инвертированными повторами. В-третьих, предыдущее исследование показало, что гомологии коротких последовательностей (то есть прямые повторы) играют роль в образовании делеций у бактерий [33]. Кроме того, было обнаружено, что горячая точка делеций характеризуется вторичной структурой, сохраняющейся инвертированными повторами [33], который очень напоминает хрупкий паттерн DIID, предложенный в моем исследовании. В-четвертых, остановка ДНК-полимеразы вблизи общего повтора мтДНК человека была продемонстрирована как предпосылка образования общей делеции [32]. Согласно предложенному в моем иследовнии механизму, это торможение может быть инициировано контактной зоной.

Проведенные анализы представляют собой первоначальный этап исследования, основанный на принципе парсимонии, и включают применение простейших методов с целью первичной валидации гипотезы перед переходом к более сложным и комплексным моделям.

Стоит заметить что как прямые, так и инвертированные повторы характеризуются значительными ограничениями по длине. Моделирование методом молекулярной динамики показывает, что образование шпилечных структур короткими повторами энергетически невыгодно из-за конкуренции со стохастическими термическими флуктуациями, приводящими к диссоциации дуплекса с обоих его концов. Существенную роль в этом процессе играют белково-нуклеотидные взаимодействия, в частности, аффинность белков к ДНК, обусловленная электростатическими взаимодействиями между отрицательно заряженной молекулой ДНК и положительно заряженными доменами белков. Кроме того, важно учитывать что формирование глобальной архитектуры мтДНК может модулироваться присутствием других структурных элементов, в изобилии представленных в митохондриальном геноме: G-квадруплексов, H-участков (склонных к спонтанному расплетанию), и Z-ДНК (левозакрученной конформации, обладающей повышенной склонностью к суперспирализации и релаксации).

Глава 6. Митохондриально-специфический мутационный признак старения: повышенная частота замен A > G в тяжелой цепи

6.1 Проблематика

Очевидно что к таким глобальным изменениям мтДНК как делеции могут приводить такие небольшие изменения как нуклеотидные замены, которые к тому же сами по себе могут вызывать патологии в работе мтДНК. Как можно было убедиться ранее, изменения в общем повторе или другом совершенном повторе могут уменьшить вероятность образования делеций, а мутации в вырожденных повторах могут как уменьшить микрогомологию, так и увеличить.

На молекулярную эволюцию влияют как мутагенез, так и отбор. Одна из гипотез предполагает, что избыток G-нуклеотидов в митохондриальном геноме долгоживущих млекопитающих обусловлен отбором, благоприятствующим более стабильным геномам этих видов. Однако, прежде чем делать выводы, важно сравнить мутационные процессы между короткоживущими и долгоживущими млекопитающими, поскольку значительные различия в мутационных спектрах мтДНК наблюдались у разных видов без четкого объяснения вариаций.

Имеется пробел в знаниях относительно мутационных спектров митохондриальной ДНК (мтДНК) в разных тканях и видах. Это подчеркивает уникальную мутационную подпись мтДНК и отсутствие понимания основных причин. Можно предположить, что мутаген, ответственный за мутации мтДНК, и факторы, влияющие на изменение мутационных спектров, до сих пор неизвестны, что подчеркивает необходимость дальнейших исследований в области соматического, популяционного и эволюционного анализа мтДНК.

Широко распространенно мнение о том, что активные формы кислорода (АФК), вырабатываемые митохондриями, могут повредить митохондриальную ДНК (мтДНК). Хорошо документированным мутационным признаком, индуцированным АФК, является модификация основания ДНК гуанина (G) в 7,8-дигидро-8-оксо-20-дезоксигуанозин (8-oxodG), что после неправильного спаривания с аденином приводит к G>T-трансверсионным мутациям. Хотя замены G>T считаются признаком окислительного повреждения ядерной ДНК (сигнатура COSMIC 18) [137, 138, 139], они довольно редки в мтДНК [140] и незначительно увеличиваются с возрастом в мтДНК [140, 141, 142]. Таким образом, до сих пор не существует четко установленных мутационных признаков окислительного повреждения мтДНК [143].

Принимая во внимание недавний прогресс в расшифровке вариаций мутационных спектров ядерного генома в зависимости от различных типов рака [144], факторов окружающей среды [139], нокаутов генов [145], человеческих популяций [146] и видов приматов [147] мы сосредоточились на мтДНК и провели крупномасштабную реконструкцию ее мутационных спектров у сотен видов млекопитающих. Учитывая тесную связь уровня метаболизма мтДНК (и, следовательно, потенциальных мутагенов мтДНК) с видоспецифичными особенностями жизненного цикла, мы стремились выявить корреляцию между мутационным спектром мтДНК и особенностями жизненного цикла. Используя коллекции (i) соматических мутаций мтДНК у мышей и людей, (ii) полиморфных синонимичных замен у сотен видов млекопитающих и (iii) содержания нуклеотидов во всех митохондриальных геномах млекопитающих, мы наблюдали одну универсальную тенденцию: замены $A_H>G_H$ (H - обозначение тяжелой цепи) положительно коррелирует с длиной поколения. Учитывая дополнительную связь замен $A_H>G_H$ со временем пребывания в одноцепочечном состоянии (TSSS) при асинхронной репликации мтДНК и многочисленные литературные данные о заменах A>G, мы предполагаем, что повышение $A_H>G_H$ в мтДНК долгоживущих млекопитающих является мутационным признаком возрастного повреждения, специфичного для одноцепочечной ДНК. Следовательно, описанные вариации в

мутационном спектре мтДНК следует более широко учитывать в соматическом, популяционном и эволюционном анализе мтДНК.

6.2 Методы

Чтобы упростить биологическую интерпретацию мутационного спектра мтДНК, мы используем 12-компонентный спектр, основанный на обозначениях тяжелых цепей.

Все данные о соматических мутациях мтДНК, полученные с помощью метода дуплексного секвенирования, были получены от [113]. Были использованы данные дуплексного секвенирования контрольной человеческой мтДНК, полученные из двух источников [148, 149] с указанным возрастным интервалом 10-30 и 80-90 лет соответственно.

Используя все доступные внутривидовые последовательности (апрель 2016 г.) генов, кодирующих митохондриальные белки, мы получили мутационный спектр для каждого вида. Мы собрали все доступные последовательности мтДНК любых белок-кодирующих генов для любого вида хордовых, реконструировали внутривидовую филогению, используя последовательность внешней группы (ближайший вид к анализируемому), реконструировали спектры предковых состояний во всех позициях во всех узлах внутреннего дерева и, наконец, получили список однонуклеотидных замен для каждого гена каждого вида.

Используя виды, имеющие не менее 15 однонуклеотидных синонимичных замен в четырехкратно вырожденных сайтах, мы оценили мутационный спектр как вероятность мутации каждого нуклеотида в любой другой нуклеотид (вектор из 12 типов замен с суммой, равной единице) в течение более чем тысяч видов хордовых. Четырехкратно вырожденные сайты практически не оказывают селективного давления и поэтому могут считаться наиболее нейтральными и отражают мутационную предвзятость [150, 151].

Основываясь на асинхронном режиме репликации мтДНК и предполагая постоянную скорость репликации ДНК-полимеразой в пределах большой и малой дуг, мы рассчитали относительный TSSS для белок-кодирующих генов, кодируемых на тяжелой цепи мтДНК человека (все, кроме ND6).

Чтобы сосредоточиться на видоспецифической изменчивости мутационных спектров, мы подробно проанализировали ген CYTB. В качестве простейшей метрики мутационного спектра для каждого вида мы сначала рассчитали соотношение переход/трансверсия (Ts/Tv) как сумму частот всех переходов, деленную на сумму частот всех трансверсий. Чтобы понять, какой тип замены преимущественно формировал наблюдаемую корреляцию между Ts/Tv и длиной поколения, мы провели двенадцать анализов попарной ранговой корреляции между каждым типом замены и длиной поколения.

6.3 Результаты

6.3.1 Частота de novo мутаций $A_H > G_H$ увеличивается с возрастом в соме и зародышевой линии

Митохондриальный геном характеризуется сильной асимметрией цепей по содержанию нуклеотидов: тяжелая цепь (H-цепь) богата гуанином (G_H) и бедна цитозином (C_H), а легкая цепь (L-цепь) — наоборот: богата цитозином (C_L) и бедна гуанином (G_L). Мутагенное объяснение этой асимметрии основано на предположении, что тяжелая цепь мтДНК, будучи одноцепочечной во время асинхронной репликации, более восприимчива к двум наиболее распространенным мутациям в мтДНК: C_H>T_H и A_H>G_H, приводящим к дефицит C_H и избыток G_H. Анализ полных митохондриальных геномов млекопитающих дополнительно показал, что

эта нуклеотидная асимметрия образует градиент вдоль мтДНК [150, 152, 153]: глобальный дефицит C_H по сравнению с T_H и A_H над и G_H в третьих положениях кодона становится все более выраженным.

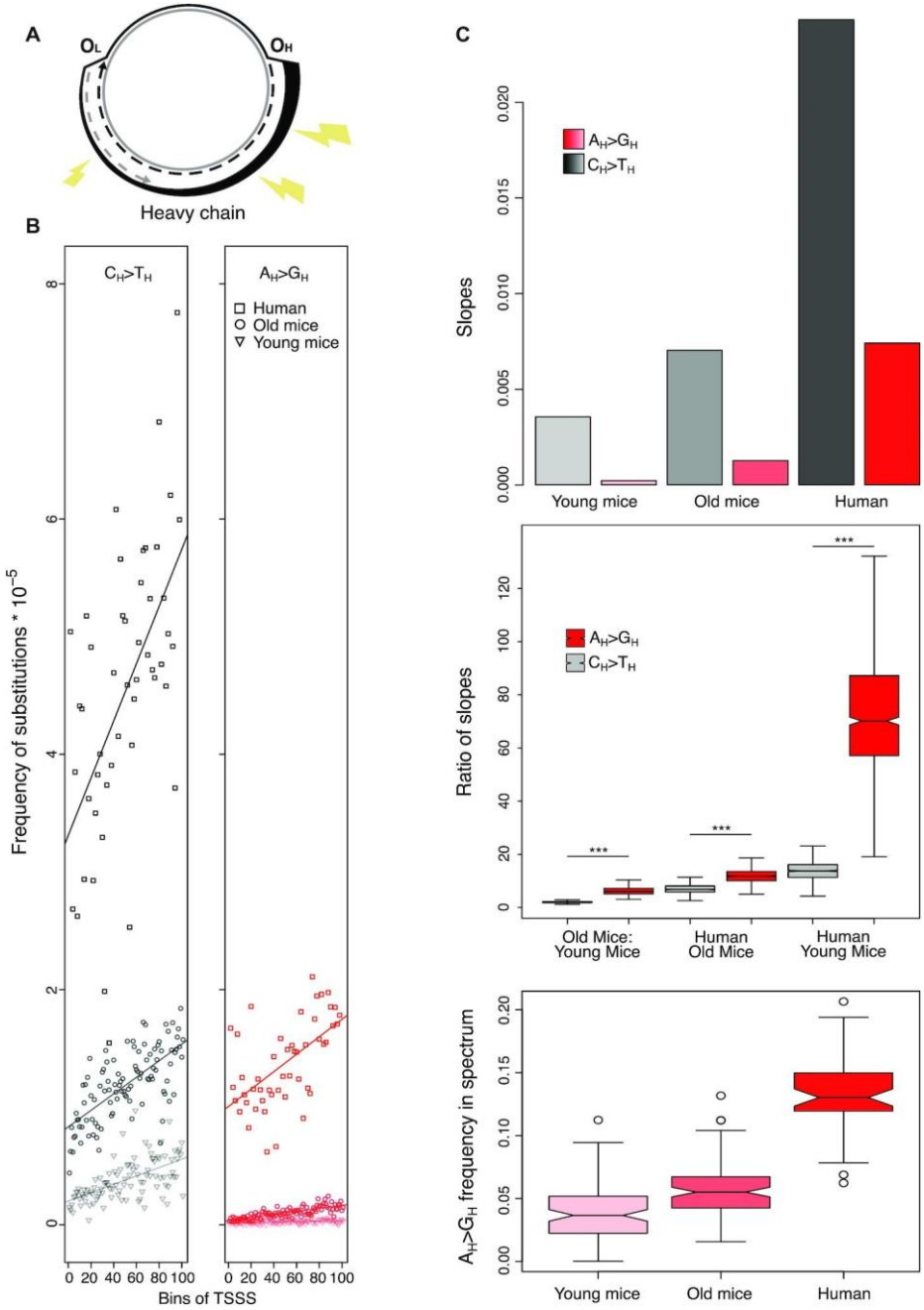

Рисунок 19. Градиент мутаций мтДНК AH>GH увеличивается с возрастом образца. (A) Асинхронная репликация мтДНК связана с длительным временем, проведенным родительской тяжелой цепью в одноцепочечной форме (TSSS). TSSS, в свою очередь, связан с высокой частотой двух наиболее распространенных переходов мтДНК: CH>TH и AH>GH (дочерняя тяжелая цепь: пунктирная черная линия; родительская тяжелая цепь: жирная утолщенная черная линия, отражающая TSSS; дочерняя легкая цепь: пунктирная серая линия; родительская легкая цепь: сплошная серая линия; OH: начало репликации дочерней тяжелой цепи, OL: начало репликации дочерней легкой цепи). (B) Градиенты мутаций CH>TH и AH>GH вдоль главной дуги мтДНК более выражены у людей по сравнению со старыми мышами и у старых мышей по сравнению с молодыми мышами. Как пересечения, так и наклоны увеличиваются с возрастом образца. (C) Скорость замены AH>GH увеличивается быстрее в старых образцах. Верхняя панель: столбчатые диаграммы визуализируют наклоны линейных регрессий между частотой мутаций и TSSS. Средняя панель: наклоны AH>GH увеличиваются быстрее с возрастом по сравнению с наклонами CH>TH. Ящичные диаграммы основаны на отношении наклонов, полученных из 1000 бутстрепированных образцов. Нижняя панель: частота AH>GH в общем мутационном спектре увеличивается с возрастом (значения P из всех трех попарных сравнений меньше 1,583e−08, U-критерий Манна–Уитни). ‘***’ обозначает значения P < 0,001.

Недавно большая коллекция соматических мутаций, полученная с помощью высокочувствительного подхода дуплексного секвенирования, позволила точно реконструировать градиенты C_H>T_H и A_H>G_H и однозначно подтвердила мутагенный эффект TSSS (времени пребывания в одноцепочечном состоянии) во время асинхронной репликации (**Рисунок 19А**) [113]. Подтверждение такой мутационной природы градиентов мтДНК [113] обеспечивает прочную основу для дальнейших исследований мутационных спектров мтДНК. Было отмечено, например, что положительный градиент G_H/A_H значительно различается между видами приматов и выше у видов с более длительным периодом беременности, в то время как градиент T_H/C_H не показывает сильных видоспецифичных вариаций [154]. Это говорит о том, что мутации A_H>G_H, формирующие градиент G_H/A_H, могут быть чувствительны к некоторым мутагенам, связанным со временем беременности или другими особенностями жизненного цикла. Благодаря существованию положительных корреляций между временем беременности, размером тела и продолжительностью жизни, которые, в свою очередь, связаны с длиной поколения и уровнем митохондриального метаболизма [155, 156, 157], мы можем ожидать различий в мутагенезе мтДНК между видами с разным жизненным циклом. Чтобы проверить эту гипотезу, мы сравнили наборы данных соматических мутаций мтДНК мышей и людей.

В целом, используя наборы данных соматических мутаций мтДНК, полученные с помощью высокочувствительного подхода дуплексного секвенирования, мы обнаружили, что A_H>G_H более чувствителен к возрасту по сравнению с C_H>T_H.

Предполагая сходство мутагенеза мтДНК в соматических и зародышевых тканях, мы ожидаем также наблюдать избыток A_H>G_H в старых тканях зародышевого типа. Действительно, недавнее глубокое секвенирование мутаций мтДНК de novo в ооцитах старых и молодых мышей подтвердило, что самым сильным признаком старения ооцитов является увеличение доли замен A_H>G_H [158].

6.3.2 A_H>G_H более распространены у млекопитающих с большой длиной поколения: данные нейтральных мутационных спектров, полученных на основе полиморфизма

Вариации в мутационных спектрах мтДНК между разными видами [159, 160] ранее не имели общего объяснения. Полученные результаты (**Рисунок 20**) и литературные данные [158] позволяют предположить, что эти изменения, и особенно часть переходов мтДНК A_H>G_H, могут быть связаны со старением. Таким образом, мы предполагаем, что видоспецифичные мутационные спектры мтДНК зависят от длины поколения, которая, в свою очередь, является хорошим показателем возраста ооцитов у млекопитающих. Поскольку ооциты млекопитающих задерживаются от рождения до полового созревания, что занимает недели (для мышей) или десятилетия (для людей) [161], мы можем использовать видоспецифичную продолжительность поколения в качестве естественного показателя долговечности ооцитов у различных видов млекопитающих. Кроме того, поскольку ооциты являются единственной линией, через которую мтДНК передается (за редкими исключениями отцовского наследования) из поколения в поколение у млекопитающих [162], мы ожидаем наблюдать корреляцию между видоспецифичными свойствами мтДНК и длиной поколения (прокси для долговечности ооцитов) у разных видов млекопитающих. Продолжительность поколения определяется как «средний возраст родителей текущей когорты», который доступен для подавляющего большинства видов млекопитающих [131, 163], а также связан с многочисленными экологическими (масса тела, размер помета, эффективный размер популяции) и физиологическими (основной уровень метаболизма) параметрами видов млекопитающих [164, 165].

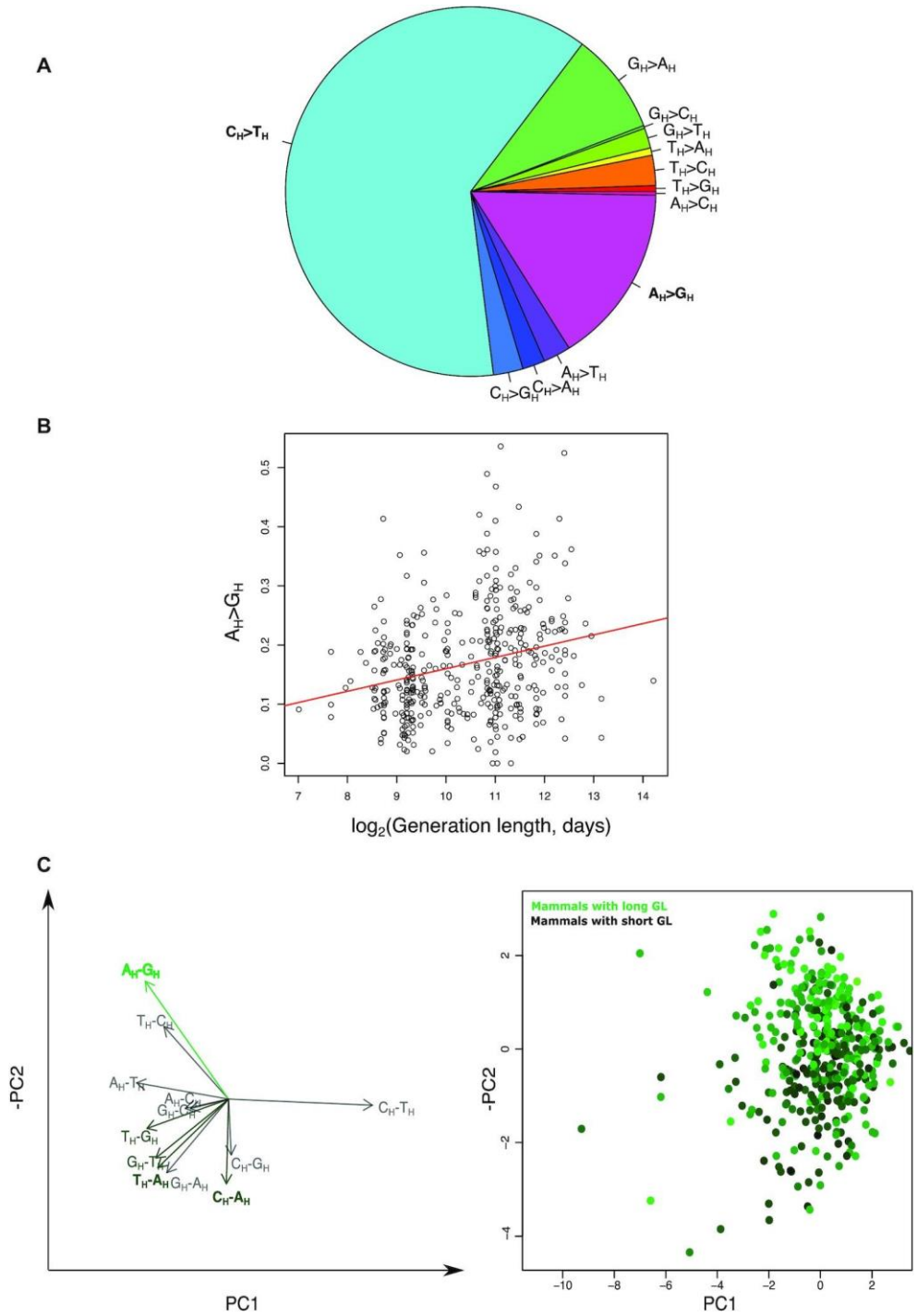

Рисунок 20. Изменчивость нейтрального спектра мутаций мтДНК млекопитающих обусловлена длиной поколения. (A) Средний спектр мутаций мтДНК видов млекопитающих (N = 611). Мутационный спектр представляет собой вероятность мутации каждого нуклеотида друг в друга на основе наблюдаемых и нормализованных частот двенадцати типов нуклеотидных замен в четырехкратно вырожденных синонимичных сайтах всех доступных внутривидовых полиморфизмов генов, кодирующих белок мтДНК. (B) Мутационные спектры варьируются в зависимости от длины поколения, специфичной для вида (N = 424). AH> GH — это тип замен, частота которых сильнее коррелирует с длиной поколения. Он показывает приблизительно двукратное различие между млекопитающими с очень короткой и очень длинной длиной поколения. (C) Анализ главных компонентов (PCA) спектров мутаций мтДНК видов млекопитающих (N = 424). Левая панель: двойной график анализа главных компонент (первый и второй компоненты объясняют 16% и 12% вариации соответственно). CH> TH имеет самую высокую нагрузку на первый главный компонент, в то время как AH> GH имеет самую высокую нагрузку на второй главный компонент. Обратите внимание, что мы построили отрицательный PC2, чтобы сделать его положительно коррелирующим с длиной поколения. Правая панель: второй главный компонент коррелирует с длиной поколения у млекопитающих. Длина поколения обозначена цветом от темно-зеленого (самая короткая длина поколения) до светло-зеленого (самая длинная длина поколения).

Многочисленные митохондриальные последовательности, полученные в результате экологических, эволюционных и популяционных генетических исследований различных видов [166], являются ценным источником полиморфизмов мтДНК, используемых в наших анализах. На основе нашей собственной разработки мы реконструировали мутационный спектр видов млекопитающих. Вкратце, мы (i) скачали все доступные нуклеотидные последовательности генов млекопитающих, кодирующих митохондриальные белки, (ii) получили множественное выравнивание кодонов для каждого гена каждого вида, (iii) укоренили внутривидовое митохондриальное дерево по последовательности ближайшего соседа из другой вид (iv) реконструировал предковые последовательности в каждом внутреннем узле, (v) получил список поляризованных однонуклеотидных замен и (vi) нормализовал их по частоте предковых нуклеотидов. Ориентируясь на наиболее нейтральные 70 053 замены, расположенные в пределах 4-кратно вырожденных синонимичных сайтов, мы реконструировали нейтральный мутационный спектр для 611 видов млекопитающих. Средний мутационный спектр всех видов млекопитающих (**Рисунок 20A**) демонстрирует сильный избыток замен $C_H>T_H$ и $A_H>G_H$, которые были показаны в предыдущих исследованиях [140, 167].

Мы заметили, что только частота $A_H>G_H$ положительно коррелировала с длиной поколения (rho Спирмена = 0,252, номинальное значение P = 1,188e-07) (**Рисунок 20B**), в то время как несколько редких трансверсий показали слабую и отрицательную корреляцию ($T_H>A_H$, $T_H>G_H$, $C_H>A_H$ и $G_H>T_H$: все значения rho Спирмена < -0,17, все номинальные значения P < 0,0003).

6.3.3 МтДНК млекопитающих с высокой длиной поколения более бедна A_H и богата G_H из-за интенсивного мутагенеза $A_H>G_H$.

Ожидается, что мутационная предвзятость, если она сильнее отбора, в долгосрочной перспективе изменит содержание нуклеотидов во всем геноме. Ниже мы проверим это предположение.

Во-первых, чтобы смоделировать возможное влияние ошибки мутации на нуклеотидный состав, мы использовали компьютерное моделирование, которое вывело ожидаемый нейтральный нуклеотидный состав на основе входного 12-компонентного мутационного спектра. Результаты этого моделирования показали, что ожидаемый нуклеотидный состав млекопитающих с высокой длиной поколения характеризуется пониженной частотой A_H. Результаты этого моделирования были подтверждены нашим аналитическим решением. Оба подхода (моделирование и аналитическое решение) также подтвердили, что ожидаемый состав нейтральных нуклеотидов в равновесии зависит исключительно от мутационного спектра и не зависит от начальных условий. Чтобы оценить, насколько млекопитающие близки к своему композиционному нуклеотидному равновесию, мы сравнили ожидаемый нуклеотидный состав с наблюдаемыми, которые были получены с использованием синонимичного четырехкратно вырожденного содержания нуклеотидов двенадцати (всех, кроме ND6) белок-кодирующих генов одного и того же вида с очень короткой и очень длинной длиной поколения (**Рисунок 21A**). Мы обнаружили, что наблюдаемый нуклеотидный состав довольно похож на ожидаемый, а это означает, что анализируемые виды достаточно близки к композиционному равновесию и продолжают стремиться к равновесию. Более того, мы заметили, что виды с короткой продолжительностью поколения имеют тенденцию быть ближе к ожидаемому равновесию (см. горизонтальные пунктирные линии на **Рисунке 21A**) по сравнению с видами с большой продолжительностью поколения, вероятно, потому, что виды с короткой продолжительностью поколения имеют повышенную скорость мутаций (увеличение количества репликаций мтДНК в единицу времени) и, таким образом, быстрее приближаются к равновесию. В целом мы заметили, что синонимичный четырехкратно вырожденный нуклеотидный состав млекопитающих близок к их нейтральному равновесию, и поэтому мы ожидаем наблюдать влияние ошибки мутации на содержание нуклеотидов в мтДНК млекопитающих.

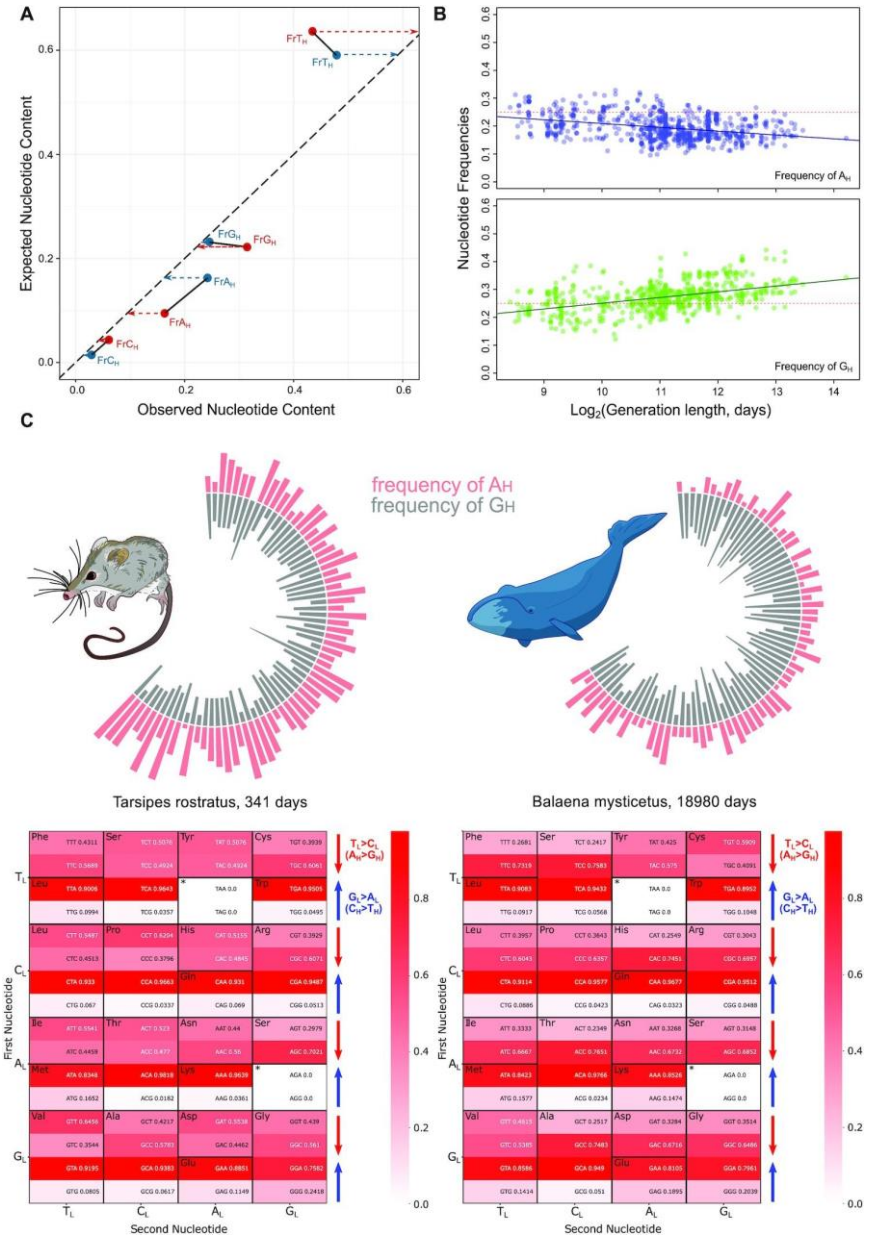

Рисунок 21. Долгосрочный эффект мутационного смещения: содержание нейтральных нуклеотидов у видов млекопитающих. (A) Корреляция между ожидаемым (полученным при моделировании) и наблюдаемым содержанием нейтральных нуклеотидов у млекопитающих с очень короткой и очень длинной длиной поколения. Из-за избытка замен A_H> G_H у долгоживущих млекопитающих (отмеченных красными кружками) они более бедны AH и богаты GH как для ожидаемых, так и для наблюдаемых значений. Расположение всех точек данных (красные и синие кружки) около диагонали показывает, что мтДНК млекопитающих достаточно близка к нейтральному равновесию. Однако короткоживущие виды (отмеченные синими кружками) находятся еще ближе к диагонали (горизонтальная пунктирная синяя линия по направлению к диагонали короче красных пунктирных линий), что позволяет предположить, что они быстрее эволюционируют в направлении нейтрального равновесия. (B) Частоты нуклеотидов в нейтральных участках всех 13 генов, кодирующих белки, как функция длины поколения — доля A_H уменьшается, а доля G_H увеличивается (N = 650). (C) Структура мтДНК двух видов млекопитающих с экстремальными длинами поколений: медоносный опоссум и кит. Верхняя панель: частоты нуклеотидов A_H (красный) и G_H (серый) вдоль главной дуги мтДНК самого короткоживущего (медоносный опоссум) и самого долгоживущего (кит) видов млекопитающих из нашего набора данных. Каждый столбец представляет собой частоту нуклеотидов в окне из 20 нуклеотидов. У обоих млекопитающих A_H уменьшается, а G_H увеличивается вдоль главной дуги мтДНК: от нижнего левого угла (начало репликации легкой цепи) до верхнего правого угла (начало репликации тяжелой цепи). Однако, в дополнение к градиенту, мтДНК кита имеет интегральный, общегеномный, дефицит A_H и избыток G_H — признак увеличенной длины поколения. Нижняя панель: тепловые карты визуализируют асимметрию использования кодонов 12 генов, кодирующих белки (все, кроме ND6). Кит

более контрастен, чем медовый опоссум, с точки зрения асимметрии, обусловленной возрастными заменами T_L>C_L (A_H>G_H). Тепловые карты обоих видов в равной степени контрастны с точки зрения асимметрии, обусловленной заменами G_L>A_L (C_H>T_H), которые имеют высокую и схожую (не связанную с возрастом) скорость замены у обоих видов.

Во-вторых, мы проверили, будет ли увеличение A_H>G_H (**Рисунок 20**) у видов с большой длиной поколения уменьшать частоты A_H и увеличивать частоты G_H в соответствующих эталонных последовательностях. Поскольку длина поколения коррелирует с силой A_H>G_H (**Рисунок 20**), мы ожидаем, что длина поколения должна демонстрировать положительную корреляцию с G_H и отрицательную с A_H. Проверяя все четыре парные корреляции между видоспецифичной длиной поколения и содержанием нуклеотидов (A_H, T_H, G_H, C_H), мы наблюдали две самые сильные корреляции: отрицательную с AH и положительную с G_H (**Рисунок 21B**). Включение всех четырех типов частот нуклеотидов в множественную линейную модель подтвердило важность только A_H и G_H, эффект которых также был устойчив к филогенетической инерции. Таким образом, мы пришли к выводу, что мтДНК млекопитающих с длинными поколениями по сравнению с короткими поколениями в большей степени бедна A_H и богата G_H (**Рисунок 21B**), что соответствует более интенсивному мутагенезу A_H>G_H у первых (**Рисунок 20**).

В-третьих, мы проверили, определяет ли избыток G_H и дефицит A_H у долгоживущих видов положительную асимметрию нуклеотидов G_HA_H. Перекос нуклеотидов G_HA_H аппроксимирует уровень асимметрии в распределении этих двух нуклеотидов и рассчитывается как (G_H–A_H)/(G_H+A_H). На основе четырехкратно вырожденных синонимических позиций 12 генов (всех, кроме ND6), мы оценили асимметрию G_HA_H для каждого вида млекопитающих и коррелировали ее с длиной поколения. Как и ожидалось, мы получили положительную корреляцию (филогенетический обобщенный метод наименьших квадратов: коэффициент = 0,13, P-value = 2,9 × 10–4; см. также **Рисунок 21C**). Чтобы визуализировать контраст в асимметрии G_HA_H между самыми короткоживущими и самыми долгоживущими видами в нашем наборе данных, мы нанесли на график фракции A_H и G_H вдоль главной дуги мтДНК медового опоссума (продолжительность поколения 341 день) и кита (продолжительность 18980 дней) (**Рисунок 21C**). Очевидно, что в среднем мтДНК медового опоссума имеет избыток A_H (красный цвет на **Рисунке 21C**), тогда как кит имеет избыток G_H (серый цвет на **Рисунке 21C**).

В целом мы продемонстрировали, что мутагенез A_H>G_H, который более выражен у долгоживущих видов, сильно влияет на формирование его эталонных последовательностей: содержание нуклеотидов (низкая частота A_H и высокая частота G_H), перекос нуклеотидов (сильный положительный перекос G_HA_H) и использование кодонов (положительная асимметрия XXC_L).

6.3.4 Перекос нуклеотидов G_HA_H зависит как от времени, проведенного в одноцепочечном состоянии (TSSS), так и от длины поколения

Было показано, что частота замен A_H>G_H зависит от того, сколько времени родительская тяжелая цепь находилась в одноцепочечном состоянии (TSSS) во время асинхронной репликации мтДНК. Гены, расположенные близко к месту начала репликации легкой цепи (O_L), такие как COX1, проводят минимальное время в одноцепочечном состоянии и демонстрируют низкую частоту A_H>G_H, в то время как гены, расположенные далеко от O_L, проводят больше времени являются одноцепочечными и демонстрируют соответственно более высокие частоты A_H>G_H (**Рисунок 19**). Таким образом, мы ожидаем, что эффективно нейтральный нуклеотидный состав мтДНК является функцией как геноспецифичного TSSS, так и видоспецифичной длины поколения. Чтобы проверить это, мы вывели для каждого гена каждого вида асимметрию G_HA_H и разделили все виды млекопитающих на виды с короткой и длинной продолжительностью поколения в соответствии с медианой (медиана = 2190 дней, N коротких = 325, N длинных = 319). Затем мы построили диаграмму распределения G_HA_H млекопитающих с короткой и длинной продолжительностью поколений для каждого гена, ранжируя их по основной дуге от COX1 (ранг равен 1) до CYTB (ранг равен 10), что соответствует увеличению TSSS. Как и

ожидалось, мы заметили, что асимметрия G_HA_H увеличивается как с геноспецифичным TSSS, так и с видоспецифичной длиной поколения (**Рисунок 22A**). Выполняя несколько линейных моделей, где асимметрия G_HA_H является функцией как TSSS, так и длины генерации, мы подтвердили, что оба фактора влияют на асимметрию, причем в очень похожей степени.

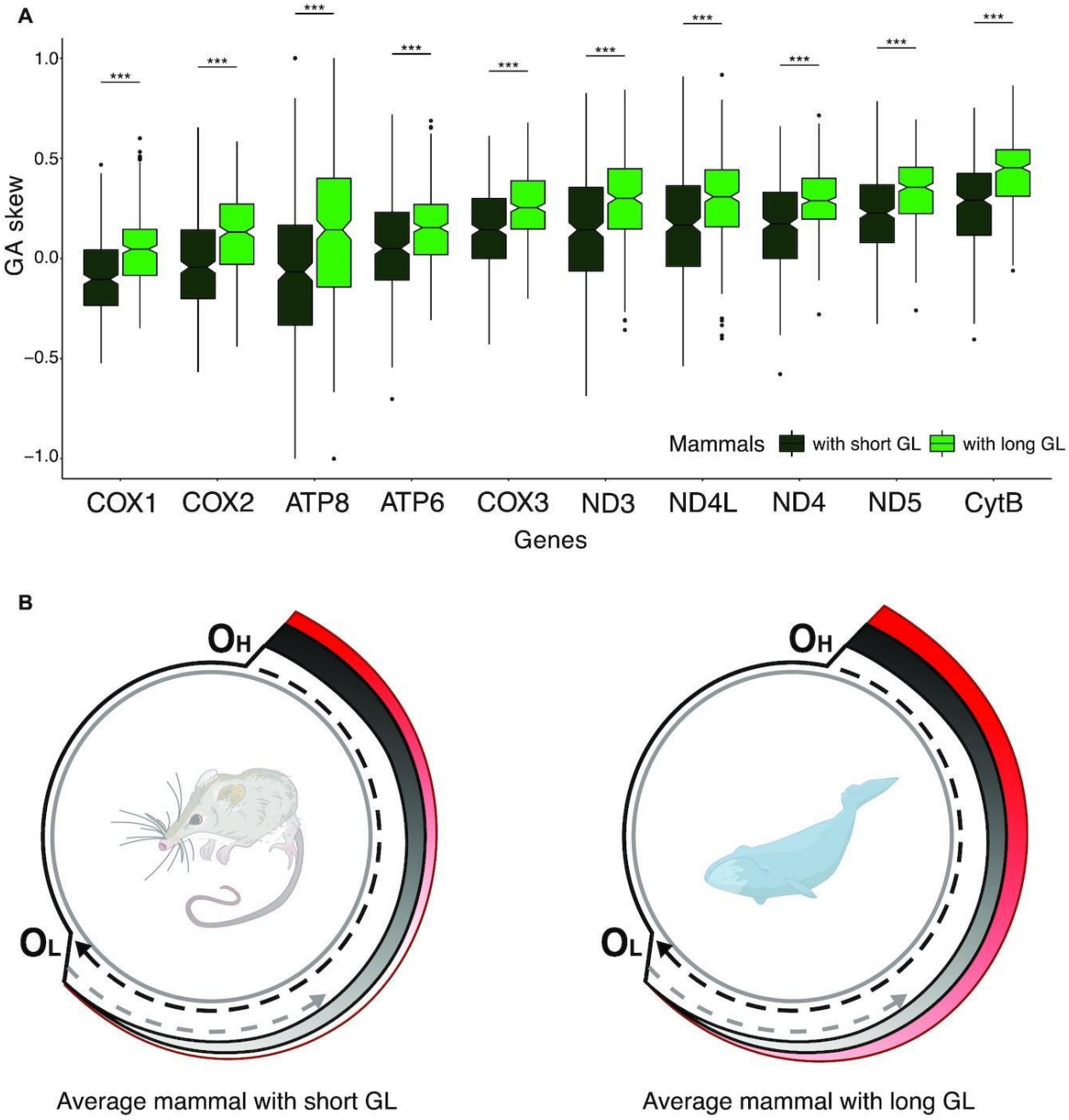

Рисунок 22. (A) Изменения в содержании нуклеотидов вдоль мтДНК коротко- и долгоживущих млекопитающих (N = 650). Все гены (за исключением ND6), расположенные в главной дуге, ранжированы в соответствии со временем, проведенным в одноцепочечной форме: от COX1 до CYTB. Пары диаграмм ящиков для каждого гена представляют перекос GHAH для коротко- и долгоживущих млекопитающих, разделенный на медианную длину поколения. GHAH увеличивается как с геноспецифическим TSSS, так и с видоспецифической длиной поколения. (B) Визуальное резюме основного вывода: скорость замены AH> GH (отмечена красным градиентом) увеличивается как с геноспецифическим TSSS, так и с видоспецифической длиной поколения. Размер эффекта GL сопоставим с размером эффекта TSSS. Скорость замены CH> TH (отмечена серым градиентом) чувствительна только к TSSS.

В линейных моделях мы не наблюдали значимого взаимодействия между TSSS и длиной поколения, что позволяет предположить, что либо эти факторы влияют на нуклеотидный состав независимо друг от друга или сигнал взаимодействия слишком слаб, чтобы быть значимым при нашем размере выборки. Наш конкретный анализ, направленный на выявление потенциального взаимодействия между TSSS и длиной поколения, действительно показал положительную тенденцию, предполагая, что млекопитающие с высокой продолжительностью поколения демонстрируют более быстрое снижение A_H и увеличение G_H вдоль генома, тот же эффект виден на **Рисунке 21B**). Более быстрые изменения (более сильные градиенты) A_H и G_H вдоль большой дуги мтДНК у долгоживущих млекопитающих можно интерпретировать как взаимодействие между TSSS и длиной поколения, как если бы скорость замещения A_H>G_H увеличивалась быстрее в зависимости от TSSS в случае большой длины генерации. В целом наши результаты показывают, что содержание нуклеотидов, сформированное мутационным смещением от A_H до G_H, положительно и сильно зависит как от TSSS, так и от длины поколения.

6.4 Вывод

В данной работе мы сконцентрировались на изучении того, как частота мутаций в митохондриальной ДНК (мтДНК) меняется с возрастом и длиной поколения у разных видов, а также как на мутационный процесс влияют различные факторы.

Мы считаем, что за наши результаты в первую очередь ответственен процесс мутагенеза мтДНК, а не отбора. Во-первых, мы не ожидаем эффекта отбора в случае чрезвычайно редких, с частотой вариантов аллелей менее 1%, мутаций мтДНК, называемых в подходе дуплексного секвенирования [113] (**Рисунок 19**). Во-вторых, из-за небольшого количества или отсутствия доказательств отбора на синонимичных четырехкратно вырожденных сайтах в мтДНК млекопитающих [150, 151] мы рассматриваем полиморфные варианты (**Рисунок 20**), а также варианты, фиксированные между видами млекопитающих (**Рисунки 21 и 22**) столь же нейтрально.

На митохондриальный мутационный спектр, и особенно замены A_H>G_H, могут влиять как ошибки гамма-ДНК-полимеразы, так и повреждения, связанные с митохондриальным микроокружением. Недавний элегантный эксперимент с мышами, гомозиготными по дефицитной по экзонуклеазе гамма-ДНК-полимеразе, показал, что градиенты A_H>G_H и CH>TH в основном формируются экзогенным мутагеном, связанным с асинхронной репликацией ДНК, а не ошибками ДНК-полимеразы [113]. Учитывая, что активные формы кислорода являются основным вредным побочным продуктом аэробного метаболизма, мы предполагаем, что наши ключевые результаты могут быть связаны с эффектами окислительного повреждения.

Наша гипотеза основана на высокой чувствительности A>G к TSSS: одноцепочечная ДНК более уязвима к агентам, повреждающим ДНК [168] ДНК существует в одноцепочечной форме в процессе репликации, транскрипции и репарации ДНК [168]. Мутационный градиент вдоль большой и малой дуг мтДНК [113] предполагает, что TSSS во время репликации более мутагенен в случае мтДНК, чем TSSS во время транскрипции и репарации. Однако ненулевые точки пересечения, наблюдаемые для обоих распространенных переходов, когда TSSS, управляемый репликацией, равен нулю [113] (см. также **Рисунок 19**), предполагают, что фоновый мутагенез, вероятно связанный с TSSS, управляемым транскрипцией или репарацией, может играть некоторую роль.

Избыток G_H в мтДНК долгоживущих млекопитающих ранее был показан Леманном и др. [169]. Они предложили объяснение этого наблюдения, основанное на отборе, предполагая повышенную стабильность геномов, богатых G_H, что может дать преимущество долгоживущим млекопитающим. Наши результаты показывают, что избыток G_H у долгоживущих млекопитающих может быть нейтральным последствием мутагенеза A_H>G_H, а не результатом

механизма, управляемого отбором (**Рисунки 21 и 22**). Кроме того, низкий эффективный размер популяции долгоживущих млекопитающих увеличивает силу случайного генетического дрейфа и скорость фиксации слегка вредных вариантов в их мтДНК [170, 171, 172], что делает объяснение, основанное на отборе, еще менее вероятным.

В целом мы продемонстрировали, что замены A>G зависят как от TSSS, так и от длины поколения, и эта связь может быть опосредована чувствительностью этого типа замены к окислительному повреждению одноцепочечной ДНК (**Рисунок 22B**).

Выводы к кандидатской диссертации

Разработан и апробирован инновационный методологический инструментарий для анализа повторов мтДНК: Создан и реализован оригинальный алгоритм на Python для детекции всех типов несовершенных повторов (прямых, инвертированных, зеркальных, комплементарных) в кольцевой мтДНК, адаптированный к её специфике. Развернута общедоступная база данных повторов для более 4000 референсных геномов позвоночных, включающая инструменты визуализации, сравнения и экспорта. Этот ресурс устраняет пробелы существующих решений и открывает новые возможности для изучения эволюции, мутагенеза и функциональной роли повторов.

Установлены универсальные закономерности организации повторов в мтДНК: Показано, что несовершенные повторы мтДНК характеризуются малой длиной, ассоциацией с релаксированными структурами ДНК и отрицательной корреляцией с GC-составом генома. Обнаружена эквивалентность пар типов повторов (прямые/зеркальные и инвертированные/комплементарные) по нуклеотидному составу и распределению, что задает новую нулевую гипотезу для эволюционных исследований. Выявлено аномальное обогащение GC-динуклеотидами (превышение GC над CG) в легкой цепи повторяющихся участков.

Раскрыта ключевая роль пространственной структуры мтДНК в образовании делеций: Экспериментально (in silico) установлено, что образование делеций в мтДНК определяется в первую очередь не микрогомологией последовательностей, а пространственной близостью участков разрыва, обеспечиваемой вторичной структурой одноцепочечной тяжелой цепи (H-цепи). Предложена модель шпильки в области 6-9 и 13-16 т.п.н. мтДНК человека как ключевого детерминанта "горячих точек" делеций. Количественно доказано преобладающее влияние фактора "зоны контакта" вторичной структуры над фактором микрогомологии (отношение шансов 0.91 против 0.33), что указывает на универсальный механизм пространственного сближения удаленных участков ДНК через стабилизированные инвертированными повторами макроструктуры как основу делеционного процесса.

Подтверждена связь между повторами мтДНК и продолжительностью жизни на определенных гаплогруппах: показано, что нарушение консервативного прямого повтора (13 п.н.) герминативной мутацией m.8473T>C в гаплогруппе D4a коррелирует с экстремальным долголетием, а разрушение паттернов DI...ID снижает риск соматических делеций, потенциально замедляя возрастные патологии (нейродегенерацию и саркопению). Выдвинута и подтверждена гипотеза, согласно которой дефицит прямых и инвертированных повторов в мтДНК долгоживущих видов возник как механизм отрицательного отбора, снижающего частоту соматических делеций и потенциально увеличивающего продолжительность здоровой жизни.

Предложены мутагенные механизмы, опосредованные повторами: Установлено, что комбинации прямых и инвертированных повторов (DI...ID) являются ключевыми драйверами делеций в мтДНК через остановку репликационной вилки и облегчение перестроек. Экспериментально подтверждена высокая мутагенность таких структур, проявляющаяся в статистически значимом увеличении частоты делеций, когда прямые повторы вложены в пары инвертированных повторов. Эволюционный дефицит этих паттернов в мтДНК человека и их отрицательная корреляция с продолжительностью жизни млекопитающих указывают на действие очищающего отбора против этих нестабильных элементов.

Установлены детерминанты мутационного спектра мтДНК: Показано, что замены A>G (A_H>G_H) в мтДНК формируются преимущественно мутагенезом (а не отбором) и зависят от двух факторов: (1) времени нахождения в одноцепочечном состоянии (TSSS) при репликации, (2) окислительного повреждения, связанного с длиной поколения. Избыток G_H-сайтов у долгоживущих видов объясняется нейтральным накоплением мутаций A_H>G_H, усиленным малым эффективным размером популяции, а не адаптивным отбором.

Заключение

Мной были созданы алгоритм поиска вырожденных повторов и база данных которые помогут исследователям в изучении структуры мтДНК различных видов и процессов мутагенеза мтДНК различных видов.

Результаты моего исследования вторичной структуры мтДНК дают подтверждение существования контактной зоны которая, по-видимому, является важным фактором в возникновении делеций, особенно у пожилых людей. Было продемонстрировано, что спектр делеций мтДНК связан не только с микрогомологией мтДНК, но и с вторичной структурой одноцепочечной тяжелой цепи мтДНК (**Рисунок 9**), а именно что нуклеотидные мотивы и структура митохондриального генома могут влиять на образование делеций митохондриальной ДНК. В частности прямые повторы, формирующие микрогомологию мтДНК, взаимодействуют с инвертированными повторами (**Рисунки 12 и 18**), которые, в свою очередь, формируют вторичную структуру одноцепочечной тяжелой цепи большой дуги (**Рисунок 8**).

Также судя по всему, похоже что повторы, несмотря на их связь с образованием делеций, необходимы для функционирования митохондриальной ДНК, так как чем больше повторов на сайт в геноме тем: (1) больше температура плавления; (2) число нуклеотидов в витке повтора; (3) выше энергия взаимодействия в спаренном участке. К тому же благодаря анализу мутационных спектров мы приближаемся к пониманию того, появления или каких повторов стоит ожидать.

Открытие об общем нарушенном общем повторе может быть связано с неким эволюционным изменением, которое произошло давным-давно. Можно предположить, что, возможно, хоть это изменение и дало людям с нарушенным общим повтором жить дольше, но не дало им преимуществ приводящих к большему количеству детей, то есть не было положительного отбора. Объединив несколько доказательств, можно предположить, что нарушение общего повтора в мтДНК человека снижает соматическую делеционную нагрузку, откладывая возрастную деградацию постмитотических клеток и старение (**Рисунок 15A**). Таким образом, соответствующие гаплогруппы мтДНК с нарушенным повтором могут быть использованы в технологиях донорства митохондрий (**Рисунок 23**). Интересно, что, несмотря на благотворный характер этих нарушений с точки зрения здоровья человека, не найдено доказательств, подтверждающих, что эти нарушения находятся под положительным отбором ни у одного человека (**Рисунок 16**) или других видов млекопитающих (**Рисунок 17**). Эта категория вариантов, полезная для старения человека, но избирательно нейтральная, важна как для лечения возрастных заболеваний, так и для понимания более глубоких эволюционных механизмов старения [173]. Понимание того, как происходят такие изменения мтДНК, может помочь найти способ устранять факторы связанные со старением. Также необходимы дополнительные исследования, чтобы понять, почему некоторые изменения в мтДНК могут быть полезны с течением времени. На данный момент ведется анализ эволюции общего повтора.

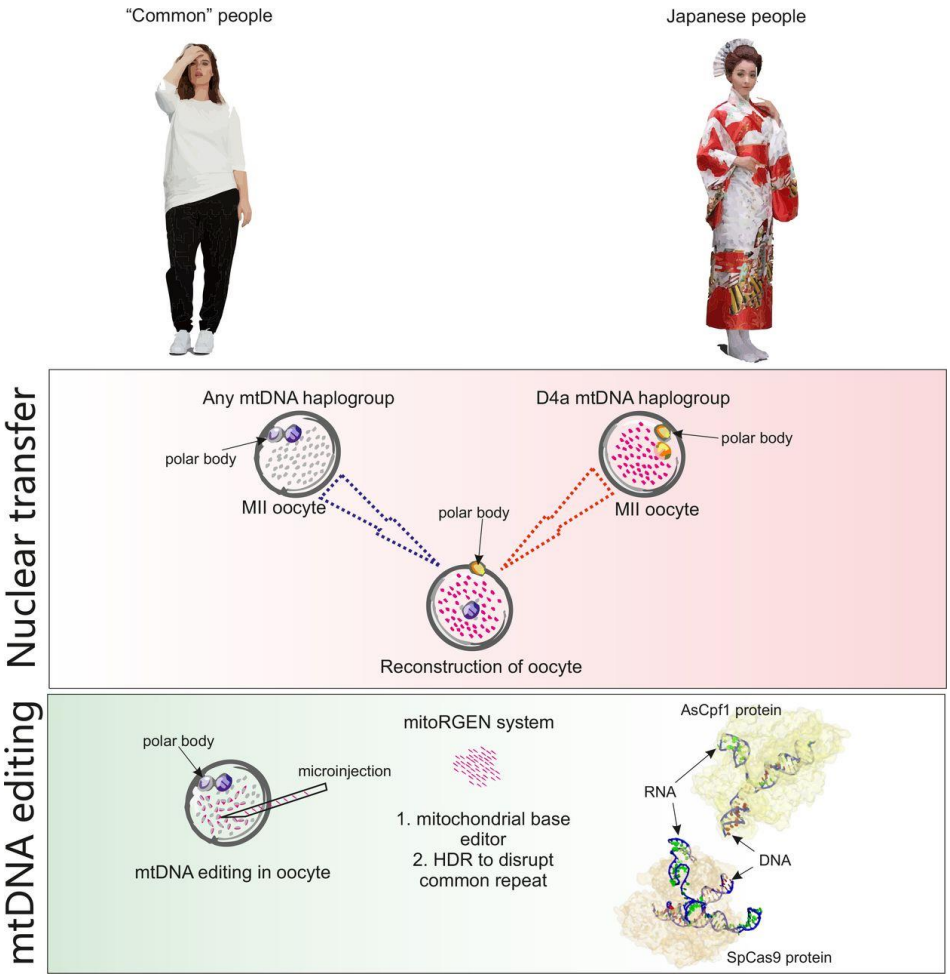

Рисунок 23. Выбор гаплогруппы мтДНК как часть вспомогательных репродуктивных технологий. «Обычные люди» означают людей с идеальным прямым повтором (ACCTCCCTCACCA), тогда как японцы — это люди с нарушенным (m.8473T>C) прямым повтором (или вообще любыми другими нарушенными прямыми повторами, Таблица 1). Перенос ядра включает удаление ядерного генома из ооцита (или зиготы), который содержит мутантную (в нашем случае «обычную») мтДНК, и перенос его в донорский ооцит (или зиготу) с мтДНК дикого типа (в нашем случае гаплогруппа D4a), у которого удален собственный ядерный геном. Методы редактирования митохондриальной ДНК включают использование митохондриально-таргетированной эндонуклеазы (в нашем случае систем mitoRGEN, митохондриально-таргетированных РНК-направляемых эндонуклеаз) для редактирования определенного участка мтДНК или расщепления и восстановления определенного участка мтДНК с помощью HDR (гомологически-направленная репарация). Ведутся работы над модификациями систем SpCas9 и AsCpf1 [174].

Анализ паттернов DIID в мтДНК у разных людей позволит получить показатель хрупкости мтДНК, который может быть количественной мерой нестабильности мтДНК. В сочетании с ядерными локусами этот показатель может быть важным дополнительным фактором при определении показателей полигенного риска различных сложных возрастных заболеваний. На данный момент ведется анализ роли G-квадруплексов в усилении эффекта инвертированного повтора вложенного в прямой повтор. Было бы полезно использовать сравнительные данные о видах, чтобы расширить выдвинутую гипотезу до эволюционного масштаба и продемонстрировать, что паттерны DIID увеличивают количество делеций в мтДНК всех видов. Первоначально сообщалось, что продолжительность жизни млекопитающих отрицательно коррелирует с обилием прямых повторов в мтДНК. [24, 25], предполагая, что прямые повторы приводят к образованию делеций мтДНК, ограничивающих продолжительность жизни. Позже было обнаружено, что инвертированные повторы имеют еще

более сильную отрицательную корреляцию с продолжительностью жизни млекопитающих [51]. В последнее время наблюдается сильная положительная корреляция между обилием прямых и инвертированных повторов [110]. В целом, мое исследование предполагает, что обилие как прямых, так и инвертированных повторов влияет на количество ломких паттернов DIID, которые, как ожидается, будут лучшими предикторами бремени соматических делеций и продолжительности жизни млекопитающих. Аннотация паттернов DIID в мтДНК всех млекопитающих или всех позвоночных откроет захватывающее потенциальное направление для будущих исследований в этой области.

Интересным шагом в контексте анализа мутационных спектров может быть анализ повторов которые в ходе мутагенеза могут стать совершенными, либо наоборот исчезнуть, а также областей близких к образованию повтора. Перспективы развития моей работы описаны в Приложении В.

Список сокращений и условных обозначений

DILL – аллели «вредные в позднем возрасте»

non-B DNA – конформации ДНК, которые отличаются от канонической конформации B-ДНК

SNP – однонуклеотидный полиморфизм

SQL – формальный непроцедурный язык программирования, применяемый для создания, модификации и управления данными в произвольной реляционной базе данных, управляемой соответствующей системой управления базами данных

Tm – температура плавления

мтДНК – митохондриальная дезоксирибонуклеиновая кислота

п.н. – пара нуклеотидов

п.о. – пара оснований

ПЦР – полимеразная цепная реакция

Словарь терминов

Большая дуга мтДНК – выделяется между двумя ориджинами репликации, OriH (тяжелой цепи) и OriL (легкой цепи), расположена на 5781 - 16569 п.н. человеческой мтДНК, очень подвержена делециям

Вложеные повторы – ситуация при которой между плечами одного повтора находятся плечи другого повтора

Вырожденность повтора – количество несовпадений между плечами повтора

Гетероплазмия – наличие нескольких отличающихся друг от друга копий последовательности ДНК каких-либо органоидов (митохондрии либо пластиды) в одном и том же организме, зачастую даже в одной клетке

Гомология – похожесть

Зеркальный повтор – повтор в котором второе плечо повторятся в обратном порядке на той же цепи (см. **Рисунок 1**)

Инвертированные повторы – повтор в котором второе плечо повторятся в обратном порядке на комплиментарной первому плечу цепи (см. **Рисунок 1**)

Квадруплекс – последовательности нуклеиновых кислот, обогащенные гуанином и способные образовывать структуры из четырёх цепей

Легкая цепь мтДНК – цепь имеющая меньшую массу при денатурации двухцепочечной цепи, как правило именно она хранится в генбанке

Минимизация свободной энергии (MFE) – метод предложенный Джозайей Уиллардом Гиббсом для вычисления стабильного состояния системы

Нарушенный общий повтор – несовпадение между плечами так называемого общего повтора

Несовершенные повторы – повторы между плечами которых есть различия

Нестабильность/хрупкость ДНК – свойство ДНК которое описывает вероятность образования структур приводящих к образованию делеций

Общий повтор (common repeat) – повтор имеющийся у большинства человеческой популяции и вызывающий большинство делеций

Плечи повторов – первая и вторая последовательность которые похожи, левая и правая соответственно по отношению к анализируюемой последовательности генома

Повтор – две похожие последовательности в одном геноме, в частности в мтДНК

Прямые нуклеотидные повторы – повтор в котором второе плечо повторятся в том же порядке на той же цепи (см. **Рисунок 1**)

Реализованный повтор – повтор который привел к делеции

Тяжелая цепь мтДНК – цепь имеющая большую чем вторая (легкая) цепь массу при денатурации двухцепочечной цепи

Список литературы

1. Poovathingal, S.K., Gruber, J., Lakshmanan, L., Halliwell, B., and Gunawan, R. (2012). Is mitochondrial DNA turnover slower than commonly assumed? Biogerontology *13*, 557–564.
2. Rebolledo-Jaramillo, B., Su, M.S.-W., Stoler, N., McElhoe, J.A., Dickins, B., Blankenberg, D., Korneliussen, T.S., Chiaromonte, F., Nielsen, R., Holland, M.M., *et al.* (2014). Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proc Natl Acad Sci USA *111*, 15474–15479.
3. Wilton, P.R., Zaidi, A., Makova, K., and Nielsen, R. (2018). A population phylogenetic view of mitochondrial heteroplasmy. Genetics *208*, 1261–1274.
4. Kraytsberg, Y., Kudryavtseva, E., McKee, A.C., Geula, C., Kowall, N.W., and Khrapko, K. (2006). Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. Nat. Genet. *38*, 518–520.
5. Bender, A., Krishnan, K.J., Morris, C.M., Taylor, G.A., Reeve, A.K., Perry, R.H., Jaros, E., Hersheson, J.S., Betts, J., Klopstock, T., *et al.* (2006). High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. Nat. Genet. *38*, 515–517.
6. Herbst, A., Pak, J.W., McKenzie, D., Bua, E., Bassiouni, M., and Aiken, J.M. (2007). Accumulation of mitochondrial DNA deletion mutations in aged muscle fibers: evidence for a causal role in muscle fiber loss. J. Gerontol. A Biol. Sci. Med. Sci. *62*, 235–245.
7. Herbst, A., Wanagat, J., Cheema, N., Widjaja, K., McKenzie, D., and Aiken, J.M. (2016). Latent mitochondrial DNA deletion mutations drive muscle fiber loss at old age. Aging Cell *15*, 1132–1139.
8. Yu-Wai-Man, P., Lai-Cheong, J., Borthwick, G.M., He, L., Taylor, G.A., Greaves, L.C., Taylor, R.W., Griffiths, P.G., and Turnbull, D.M. (2010). Somatic mitochondrial DNA deletions accumulate to high levels in aging human extraocular muscles. Invest. Ophthalmol. Vis. Sci. *51*, 3347–3353.
9. Blakely, E.L., He, L., Taylor, R.W., Chinnery, P.F., Lightowlers, R.N., Schaefer, A.M., and Turnbull, D.M. (2004). Mitochondrial DNA deletion in “identical” twin brothers. J. Med. Genet. *41*, e19.
10. Barritt, J.A., Brenner, C.A., Cohen, J., and Matt, D.W. (1999). Mitochondrial DNA rearrangements in human oocytes and embryos. Mol. Hum. Reprod. *5*, 927–933.
11. Chan, C.C.W., Liu, V.W.S., Lau, E.Y.L., Yeung, W.S.B., Ng, E.H.Y., and Ho, P.C. (2005). Mitochondrial DNA content and 4977 bp deletion in unfertilized oocytes. Mol. Hum. Reprod. *11*, 843–846.
12. Schon, E.A., Rizzuto, R., Moraes, C.T., Nakase, H., Zeviani, M., and DiMauro, S. (1989). A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA. Science *244*, 346–349.
13. Goldstein, A., and Falk, M.J. (2019). Mitochondrial DNA deletion syndromes. In GeneReviews(®), R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, A. Amemiya, L. J. Bean, T. D. Bird, N. Ledbetter, H. C. Mefford, R. J. Smith, et al., eds. (Seattle (WA): University of Washington, Seattle).
14. Vermulst, M., Wanagat, J., Kujoth, G.C., Bielas, J.H., Rabinovitch, P.S., Prolla, T.A., and Loeb, L.A. (2008). DNA deletions and clonal mutations drive premature aging in mitochondrial mutator mice. Nat. Genet. *40*, 392–394.

15. Trifunovic, A., Wredenberg, A., Falkenberg, M., Spelbrink, J.N., Rovio, A.T., Bruder, C.E., Bohlooly-Y, M., Gidlöf, S., Oldfors, A., Wibom, R., *et al.* (2004). Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature *429*, 417–423.

16. Khrapko, K., and Vijg, J. (2007). Mitochondrial DNA mutations and aging: a case closed? Nat. Genet. *39*, 445–446.

17. Vermulst, M., Bielas, J.H., Kujoth, G.C., Ladiges, W.C., Rabinovitch, P.S., Prolla, T.A., and Loeb, L.A. (2007). Mitochondrial point mutations do not limit the natural lifespan of mice. Nat. Genet. *39*, 540–543.

18. Hiona, A., Sanz, A., Kujoth, G.C., Pamplona, R., Seo, A.Y., Hofer, T., Someya, S., Miyakawa, T., Nakayama, C., Samhan-Arias, A.K., *et al.* (2010). Mitochondrial DNA mutations induce mitochondrial dysfunction, apoptosis and sarcopenia in skeletal muscle of mitochondrial DNA mutator mice. PLoS ONE *5*, e11468.

19. Nicholas, A., Kraytsberg, Y., Guo, X., and Khrapko, K. (2009). On the timing and the extent of clonal expansion of mtDNA deletions: evidence from single-molecule PCR. Exp. Neurol. *218*, 316–319.

20. Popadin, K., Safdar, A., Kraytsberg, Y., and Khrapko, K. (2014). When man got his mtDNA deletions? Aging Cell *13*, 579–582.

21. Samuels, D.C., Schon, E.A., and Chinnery, P.F. (2004). Two direct repeats cause most human mtDNA deletions. Trends Genet. *20*, 393–398.

22. Guo, X., Popadin, K.Y., Markuzon, N., Orlov, Y.L., Kraytsberg, Y., Krishnan, K.J., Zsurka, G., Turnbull, D.M., Kunz, W.S., and Khrapko, K. (2010). Repeats, longevity and the sources of mtDNA deletions: evidence from “deletional spectra”. Trends Genet. *26*, 340–343.

23. Cortopassi, G.A. (2002). A neutral theory predicts multigenic aging and increased concentrations of deleterious mutations on the mitochondrial and Y chromosomes. Free Radic. Biol. Med. *33*, 605–610.

24. Samuels, D.C. (2004). Mitochondrial DNA repeats constrain the life span of mammals. Trends Genet. *20*, 226–229.

25. Khaidakov, M., Siegel, E.R., and Shmookler Reis, R.J. (2006). Direct repeats in mitochondrial DNA and mammalian lifespan. Mech. Ageing Dev. *127*, 808–812.

26. Popadin, K., and Bazykin, G. (2008). Nucleotide repeats in mitochondrial genome determine human lifespan. Nature Precedings.

27. Bilal, E., Rabadan, R., Alexe, G., Fuku, N., Ueno, H., Nishigaki, Y., Fujita, Y., Ito, M., Arai, Y., Hirose, N., *et al.* (2008). Mitochondrial DNA haplogroup D4a is a marker for extreme longevity in Japan. PLoS ONE *3*, e2421.

28. Alexe, G., Fuku, N., Bilal, E., Ueno, H., Nishigaki, Y., Fujita, Y., Ito, M., Arai, Y., Hirose, N., Bhanot, G., *et al.* (2007). Enrichment of longevity phenotype in mtDNA haplogroups D4b2b, D4a, and D5 in the Japanese population. Hum. Genet. *121*, 347–356.

29. Khrapko, K. (2011). The timing of mitochondrial DNA mutations in aging. Nat. Genet. *43*, 726–727.

30. Lujan, S.A., Longley, M.J., Humble, M.H., Lavender, C.A., Burkholder, A., Blakely, E.L., Alston, C.L., Gorman, G.S., Turnbull, D.M., McFarland, R., *et al.* (2020). Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. Genome Biol. *21*, 248.

31. Payne, B.A.I., Wilson, I.J., Hateley, C.A., Horvath, R., Santibanez-Koref, M., Samuels, D.C., Price, D.A., and Chinnery, P.F. (2011). Mitochondrial aging is accelerated by anti-retroviral

therapy through the clonal expansion of mtDNA mutations. Nat. Genet. *43*, 806–810.

32. Phillips, A.F., Millet, A.R., Tigano, M., Dubois, S.M., Crimmins, H., Babin, L., Charpentier, M., Piganeau, M., Brunet, E., and Sfeir, A. (2017). Single-Molecule Analysis of mtDNA Replication Uncovers the Basis of the Common Deletion. Mol. Cell *65*, 527-538.e6.

33. Albertini, A.M., Hofer, M., Calos, M.P., and Miller, J.H. (1982). On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell *29*, 319–328.

34. Persson, Ö., Muthukumar, Y., Basu, S., Jenninger, L., Uhler, J.P., Berglund, A.-K., McFarland, R., Taylor, R.W., Gustafsson, C.M., Larsson, E., *et al.* (2019). Copy-choice recombination during mitochondrial L-strand synthesis causes DNA deletions. Nat. Commun. *10*, 759.

35. Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., *et al.* (2017). Genome organization drives chromosome fragility. Cell *170*, 507-521.e18.

36. Li, B., Cao, Y., Westhof, E., and Miao, Z. (2020). Advances in RNA 3D structure modeling using experimental data. Front. Genet. *11*, 574485.

37. Pina, A.F., Sousa, S.F., Azevedo, L., and Carneiro, J. (2022). Non-B DNA conformations analysis through molecular dynamics simulations. Biochim. Biophys. Acta Gen. Subj. *1866*, 130252.

38. Lakshmanan, L.N., Yee, Z., Gruber, J., Halliwell, B., and Gunawan, R. (2018). Thermodynamic analysis of mitochondrial DNA breakpoints reveals mechanistic details of deletion mutagenesis. BioRxiv.

39. Serrano, I.M., Hirose, M., Valentine, C.C., Roesner, S., Schmidt, E., Pratt, G., Williams, L., Salk, J., Ibrahim, S., and Sudmant, P.H. (2024). Mitochondrial haplotype and mito-nuclear matching drive somatic mutation and selection throughout ageing. Nat. Ecol. Evol. *8*, 1021–1034.

40. Zhang, Y., Xiong, Y., and Xiao, Y. (2022). 3ddna: A computational method of building DNA 3D structures. Molecules *27*.

41. Magnus, M., Kappel, K., Das, R., and Bujnicki, J.M. (2019). RNA 3D structure prediction guided by independent folding of homologous sequences. BMC Bioinformatics *20*, 512.

42. Wang, W., Feng, C., Han, R., Wang, Z., Ye, L., Du, Z., Wei, H., Zhang, F., Peng, Z., and Yang, J. (2023). trRosettaRNA: automated prediction of RNA 3D structure with transformer network. Nat. Commun. *14*, 7266.

43. Skorupski, J. (2022). Characterisation of the Complete Mitochondrial Genome of Critically Endangered Mustela lutreola (Carnivora: Mustelidae) and Its Phylogenetic and Conservation Implications. Genes *13*.

44. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

45. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

46. Damas, J., Carneiro, J., Amorim, A., and Pereira, F. (2014). MitoBreak: the mitochondrial DNA breakpoints database. Nucleic Acids Res. *42*, D1261-8.

47. Le, T.B.K., Imakaev, M.V., Mirny, L.A., and Laub, M.T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. Science *342*, 731–734.

48. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., *et al.* (2014). A 3D map of the human

genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

49. Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P., and Aiden, E.L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. *6*, 256-258.e1.

50. Lehmann, G., Segal, E., Muradian, K.K., and Fraifeld, V.E. (2008). Do mitochondrial DNA and metabolic rate complement each other in determination of the mammalian maximum longevity? Rejuvenation Res. *11*, 409–417.

51. Yang, J.-N., Seluanov, A., and Gorbunova, V. (2013). Mitochondrial inverted repeats strongly correlate with lifespan: mtDNA inversions and aging. PLoS ONE *8*, e73318.

52. Galtier, N., Jobson, R.W., Nabholz, B., Glémin, S., and Blier, P.U. (2009). Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. Biol. Lett. *5*, 413–416.

53. Cechová, J., Lýsek, J., Bartas, M., and Brázda, V. (2018). Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. Bioinformatics *34*, 1081–1085.

54. Madsen, C.S., Ghivizzani, S.C., and Hauswirth, W.W. (1993). In vivo and in vitro evidence for slipped mispairing in mammalian mitochondria. Proc Natl Acad Sci USA *90*, 7671–7675.

55. Mita, S., Rizzuto, R., Moraes, C.T., Shanske, S., Arnaudo, E., Fabrizi, G.M., Koga, Y., DiMauro, S., and Schon, E.A. (1990). Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. Nucleic Acids Res. *18*, 561–567.

56. Lakshmanan, L.N., Gruber, J., Halliwell, B., and Gunawan, R. (2015). Are mutagenic non D-loop direct repeat motifs in mitochondrial DNA under a negative selection pressure? Nucleic Acids Res. *43*, 4098–4108.

57. Damas, J., Carneiro, J., Gonçalves, J., Stewart, J.B., Samuels, D.C., Amorim, A., and Pereira, F. (2012). Mitochondrial DNA deletions are associated with non-B DNA conformations. Nucleic Acids Res. *40*, 7606–7621.

58. Dong, D.W., Pereira, F., Barrett, S.P., Kolesar, J.E., Cao, K., Damas, J., Yatsunyk, L.A., Johnson, F.B., and Kaufman, B.A. (2014). Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. BMC Genomics *15*, 677.

59. Wang, Y., Liu, V.W.S., Ngan, H.Y.S., and Nagley, P. (2005). Frequent occurrence of mitochondrial microsatellite instability in the D-loop region of human cancers. Ann. N. Y. Acad. Sci. *1042*, 123–129.

60. Lee, J.-H., Hwang, I., Kang, Y.-N., Choi, I.-J., and Kim, D.-K. (2015). Genetic characteristics of mitochondrial DNA was associated with colorectal carcinogenesis and its prognosis. PLoS ONE *10*, e0118612.

61. Czarnecka, A.M., Klemba, A., Semczuk, A., Plak, K., Marzec, B., Krawczyk, T., Kofler, B., Golik, P., and Bartnik, E. (2009). Common mitochondrial polymorphisms as risk factor for endometrial cancer. Int. Arch. Med. *2*, 33.

62. Tipirisetti, N.R., Govatati, S., Pullari, P., Malempati, S., Thupurani, M.K., Perugu, S., Guruvaiah, P., Rao K, L., Digumarti, R.R., Nallanchakravarthula, V., *et al.* (2014). Mitochondrial control region alterations and breast cancer risk: a study in South Indian population. PLoS ONE *9*, e85363.

63. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.

64. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*,

D81-9.

65. Mount, D.W. (2007). Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc. *2007*, pdb.top17.

66. Rasmussen, K.R., Stoye, J., and Myers, E.W. (2006). Efficient q-gram filters for finding all epsilon-matches over a given length. J. Comput. Biol. *13*, 296–308.

67. Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999). Alignment of whole genomes. Nucleic Acids Res. *27*, 2369–2376.

68. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker. Available at: http://repeatmasker.org [Accessed September 2, 2024].

69. Bao, Z., and Eddy, S.R. (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. *12*, 1269–1276.

70. Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS ONE *6*, e16526.

71. Chen, G.-L., Chang, Y.-J., and Hsueh, C.-H. (2013). PRAP: an ab initio software package for automated genome-wide analysis of DNA repeats for prokaryotes. Bioinformatics *29*, 2683–2689.

72. Edgar, R.C., and Myers, E.W. (2005). PILER: identification and classification of genomic repeats. Bioinformatics *21 Suppl 1*, i152-8.

73. Gurusaran, M., Ravella, D., and Sekar, K. (2013). RepEx: repeat extractor for biological sequences. Genomics *102*, 403–408.

74. Sonnhammer, E.L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene *167*, GC1-10.

75. Taneda, A. (2004). Adplot: detection and visualization of repetitive patterns in complete genomes. Bioinformatics *20*, 701–708.

76. Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics *23*, 1026–1028.

77. Brodie, R., Roper, R.L., and Upton, C. (2004). JDotter: a Java interface to multiple dotplots generated by dotter. Bioinformatics *20*, 279–281.

78. Tóth, G., Deák, G., Barta, E., and Kiss, G.B. (2006). PLOTREP: a web tool for defragmentation and visual analysis of dispersed genomic repeats. Nucleic Acids Res. *34*, W708-13.

79. Husemann, P., and Stoye, J. (2010). r2cat: synteny plots and comparative assembly. Bioinformatics *26*, 570–571.

80. Cabanettes, F., and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ *6*, e4958.

81. Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-8.

82. Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. *29*, 4633–4642.

83. Mori, H., Evans-Yamamoto, D., Ishiguro, S., Tomita, M., and Yachie, N. (2019). Fast and global detection of periodic sequence repeats in large genomic resources. Nucleic Acids Res. *47*, e8.

84. Morgulis, A., Gertz, E.M., Schäffer, A.A., and Agarwala, R. (2006). WindowMasker: window-

based masker for sequenced genomes. Bioinformatics *22*, 134–141.

85. Abouelhoda, M.I., Kurtz, S., and Ohlebusch, E. (2004). Replacing suffix trees with enhanced suffix arrays. Journal of Discrete Algorithms *2*, 53–86.

86. Schaeffer, C.E., Figueroa, N.D., Liu, X., and Karro, J.E. (2016). phRAIDER: Pattern-Hunter based Rapid Ab Initio Detection of Elementary Repeats. Bioinformatics *32*, i209–i215.

87. Sharma, D., Issac, B., Raghava, G.P.S., and Ramaswamy, R. (2004). Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics *20*, 1405–1412.

88. Pyatkov, M.I., and Pankratov, A.N. (2014). SBARS: fast creation of dotplots for DNA sequences on different scales using GA-,GC-content. Bioinformatics *30*, 1765–1766.

89. Yin, C. (2017). Identification of repeats in DNA sequences using nucleotide distribution uniformity. J. Theor. Biol. *412*, 138–145.

90. Goios, A., Meirinhos, J., Rocha, R., Lopes, R., Amorim, A., and Pereira, L. (2006). RepeatAround: a software tool for finding and visualizing repeats in circular genomes and its application to a human mtDNA database. Mitochondrion *6*, 218–224.

91. Cortopassi, G.A., and Arnheim, N. (1990). Detection of a specific mitochondrial DNA deletion in tissues of older humans. Nucleic Acids Res. *18*, 6927–6933.

92. Sablok, G., Padma Raju, G.V., Mudunuri, S.B., Prabha, R., Singh, D.P., Baev, V., Yahubyan, G., Ralph, P.J., and La Porta, N. (2015). ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection. Database (Oxford) *2015*.

93. Kumar, M., Kapil, A., and Shanker, A. (2014). MitoSatPlant: mitochondrial microsatellites database of viridiplantae. Mitochondrion *19 Pt B*, 334–337.

94. Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

95. Broughton, J.P., Lovci, M.T., Huang, J.L., Yeo, G.W., and Pasquinelli, A.E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. Mol. Cell *64*, 320–333.

96. Kumari (2014). RANDOMLY AMPLIFIED POLYMORPHIC DNA-A BRIEF REVIEW. Am. J. Anim. Vet. Sci. *9*, 6–13.

97. Power, E.G. (1996). RAPD typing in microbiology--a technical review. J. Hosp. Infect. *34*, 247–265.

98. Atienzar, F.A., and Jha, A.N. (2006). The random amplified polymorphic DNA (RAPD) assay and related techniques applied to genotoxicity and carcinogenesis studies: a critical review. Mutat. Res. *613*, 76–102.

99. Harrison, A., Binder, H., Buhot, A., Burden, C.J., Carlon, E., Gibas, C., Gamble, L.J., Halperin, A., Hooyberghs, J., Kreil, D.P., *et al.* (2013). Physico-chemical foundations underpinning microarray and next-generation sequencing experiments. Nucleic Acids Res. *41*, 2779–2796.

100. Hooyberghs, J., Van Hummelen, P., and Carlon, E. (2009). The effects of mismatches on hybridization in DNA microarrays: determination of nearest neighbor parameters. Nucleic Acids Res. *37*, e53.

101. Fish, D.J., Horne, M.T., Brewood, G.P., Goodarzi, J.P., Alemayehu, S., Bhandiwad, A., Searles, R.P., and Benight, A.S. (2007). DNA multiplex hybridization on microarrays and thermodynamic stability in solution: a direct comparison. Nucleic Acids Res. *35*, 7197–7208.

102. Minetti, C.A.S.A., Remeta, D.P., Dickstein, R., and Breslauer, K.J. (2010). Energetic signatures of single base bulges: thermodynamic consequences and biological implications. Nucleic Acids

Res. *38*, 97–116.

103. Sayers, E. A General Introduction to the E-utilities - Entrez Programming Utilities Help. Available at: https://www.ncbi.nlm.nih.gov/books/NBK25497/ [Accessed September 2, 2024].

104. Skinner, M.E., and Holmes, I.H. (2010). Setting up the JBrowse genome browser. Curr. Protoc. Bioinformatics *Chapter 9*, Unit 9.13.

105. Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J., and Holmes, I.H. (2009). JBrowse: a next-generation genome browser. Genome Res. *19*, 1630–1638.

106. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. Trends Genet. *16*, 276–277.

107. Miralles Fusté, J., Shi, Y., Wanrooij, S., Zhu, X., Jemt, E., Persson, Ö., Sabouri, N., Gustafsson, C.M., and Falkenberg, M. (2014). In vivo occupancy of mitochondrial single-stranded DNA binding protein supports the strand displacement mode of DNA replication. PLoS Genet. *10*, e1004832.

108. Morin, J.A., Cerrón, F., Jarillo, J., Beltran-Heredia, E., Ciesielski, G.L., Arias-Gonzalez, J.R., Kaguni, L.S., Cao, F.J., and Ibarra, B. (2017). DNA synthesis determines the binding mode of the human mitochondrial single-stranded DNA-binding protein. Nucleic Acids Res. *45*, 7237–7248.

109. Zazhytska, M., Kodra, A., Hoagland, D.A., Frere, J., Fullard, J.F., Shayya, H., McArthur, N.G., Moeller, R., Uhl, S., Omer, A.D., *et al.* (2022). Non-cell-autonomous disruption of nuclear architecture as a potential cause of COVID-19-induced anosmia. Cell *185*, 1052-1064.e12.

110. Shamanskiy, V.A., Timonina, V.N., Popadin, K.Y., and Gunbin, K.V. (2019). ImtRDB: a database and software for mitochondrial imperfect interspersed repeats annotation. BMC Genomics *20*, 295.

111. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.

112. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

113. Sanchez-Contreras, M., Sweetwyne, M.T., Kohrn, B.F., Tsantilas, K.A., Hipp, M.J., Schmidt, E.K., Fredrickson, J., Whitson, J.A., Campbell, M.D., Rabinovitch, P.S., *et al.* (2021). A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. Nucleic Acids Res. *49*, 11103–11118.

114. Mikhailova, A.G., Mikhailova, A.A., Ushakova, K., Tretiakov, E.O., Iliushchenko, D., Shamansky, V., Lobanova, V., Kozenkov, I., Efimenko, B., Yurchenko, A.A., *et al.* (2022). A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand. Nucleic Acids Res. *50*, 10264–10277.

115. Clima, R., Preste, R., Calabrese, C., Diroma, M.A., Santorsola, M., Scioscia, G., Simone, D., Shen, L., Gasparre, G., and Attimonelli, M. (2017). HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. Nucleic Acids Res. *45*, D698–D706.

116. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58-63.

117. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.

32, 268–274.

118. Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., and Cooper, A. (2012). The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. *29*, 3345–3358.

119. Han, M.V., and Zmasek, C.M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics *10*, 356.

120. Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: High-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics *26*, 1669–1670.

121. Adler, D., Kelly, S.T., Elliott, T.M., and Adamson, J. (2024). vioplot: Violin Plot. The R Foundation.

122. Bolze, A., Mendez, F., White, S., Tanudjaja, F., Isaksson, M., Jiang, R., Rossi, A.D., Cirulli, E.T., Rashkin, M., Metcalf, W.J., *et al.* (2020). A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans. BioRxiv.

123. Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. *40*, 1124–1129.

124. Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840.

125. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature *485*, 642–645.

126. Calvo, S.E., Clauser, K.R., and Mootha, V.K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. *44*, D1251-7.

127. Vincent, A.E., Rosa, H.S., Pabis, K., Lawless, C., Chen, C., Grünewald, A., Rygiel, K.A., Rocha, M.C., Reeve, A.K., Falkous, G., *et al.* (2018). Subcellular origin of mitochondrial DNA deletions in human skeletal muscle. Ann. Neurol. *84*, 289–301.

128. Chapman, S.N., Pettay, J.E., Lummaa, V., and Lahdenperä, M. (2019). Limits to Fitness Benefits of Prolonged Post-reproductive Lifespan in Women. Curr. Biol. *29*, 645-650.e3.

129. Engelhardt, S.C., Bergeron, P., Gagnon, A., Dillon, L., and Pelletier, F. (2019). Using geographic distance as a potential proxy for help in the assessment of the grandmother hypothesis. Curr. Biol. *29*, 651-656.e3.

130. Raule, N., Sevini, F., Li, S., Barbieri, A., Tallaro, F., Lomartire, L., Vianello, D., Montesanto, A., Moilanen, J.S., Bezrukov, V., *et al.* (2014). The co-occurrence of mtDNA mutations on different oxidative phosphorylation subunits, not detected by haplogroup analysis, affects human longevity and is population specific. Aging Cell *13*, 401–407.

131. Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottolo Marasini, G., Visconti, P., and Rondinini, C. (2013). Generation length for mammals. NC *5*, 89–94.

132. Felsenstein, J. (1985). Phylogeniesand the Comparative Method. The American Naturalist *Vol. 125, No. 1*.

133. Viguera, E., Canceill, D., and Ehrlich, S.D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. *20*, 2587–2595.

134. Trivedi, R., Sitalaximi, T., Banerjee, J., Singh, A., Sircar, P.K., and Kashyap, V.K. (2006). Molecular insights into the origins of the Shompen, a declining population of the Nicobar archipelago. J. Hum. Genet. *51*, 217–226.

135. Peng, M.-S., Quang, H.H., Dang, K.P., Trieu, A.V., Wang, H.-W., Yao, Y.-G., Kong, Q.-P., and Zhang, Y.-P. (2010). Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. Mol. Biol. Evol. *27*, 2417–2430.

136. Mikhailov, K.V., Efeykin, B.D., Panchin, A.Y., Knorre, D.A., Logacheva, M.D., Penin, A.A., Muntyan, M.S., Nikitin, M.A., Popova, O.V., Zanegina, O.N., *et al.* (2019). Coding palindromes in mitochondrial genes of Nematomorpha. Nucleic Acids Res. *47*, 6858–6870.

137. Fraga, C.G., Shigenaga, M.K., Park, J.W., Degan, P., and Ames, B.N. (1990). Oxidative damage to DNA during aging: 8-hydroxy-2'-deoxyguanosine in rat organ DNA and urine. Proc Natl Acad Sci USA *87*, 4533–4537.

138. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., *et al.* (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

139. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P., *et al.* (2019). A compendium of mutational signatures of environmental agents. Cell *177*, 821-836.e16.

140. Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J., Weinstein, J.N., *et al.* (2020). Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat. Genet. *52*, 342–352.

141. Kennedy, S.R., Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. *9*, e1003794.

142. Gouliaeva, N.A., Kuznetsova, E.A., and Gaziev, A.I. (2006). Proteins associated with mitochondrial DNA protect it against X-rays and hydrogen peroxide. Biophysics (Oxf) *51*, 620–623.

143. Zsurka, G., Peeva, V., Kotlyar, A., and Kunz, W.S. (2018). Is There Still Any Role for Oxidative Stress in Mitochondrial DNA-Dependent Aging? Genes *9*.

144. Koh, G., Degasperi, A., Zou, X., Momen, S., and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. Nat. Rev. Cancer *21*, 619–637.

145. Zou, X., Koh, G.C.C., Nanda, A.S., Degasperi, A., Urgo, K., Roumeliotis, T.I., Agu, C.A., Badja, C., Momen, S., Young, J., *et al.* (2021). A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat. Cancer *2*, 643–657.

146. Harris, K., and Pritchard, J.K. (2017). Rapid evolution of the human mutation spectrum. eLife *6*.

147. Moorjani, P., Amorim, C.E.G., Arndt, P.F., and Przeworski, M. (2016). Variation in the molecular clock of primates. Proc Natl Acad Sci USA *113*, 10607–10612.

148. Baker, K.T., Nachmanson, D., Kumar, S., Emond, M.J., Ussakli, C., Brentnall, T.A., Kennedy, S.R., and Risques, R.A. (2019). Mitochondrial DNA Mutations are Associated with Ulcerative Colitis Preneoplasia but Tend to be Negatively Selected in Cancer. Mol. Cancer Res. *17*, 488–498.

149. Hoekstra, J.G., Hipp, M.J., Montine, T.J., and Kennedy, S.R. (2016). Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. Ann. Neurol. *80*, 301–306.

150. Faith, J.J., and Pollock, D.D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics *165*, 735–745.

151. Uddin, A., and Chakraborty, S. (2017). Synonymous codon usage pattern in mitochondrial CYB

gene in pisces, aves, and mammals. Mitochondrial DNA A DNA Mapp. Seq. Anal. *28*, 187–196.

152. Reyes, A., Gissi, C., Pesole, G., and Saccone, C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. *15*, 957–966.

153. Tanaka, M., and Ozawa, T. (1994). Strand asymmetry in human mitochondrial DNA mutations. Genomics *22*, 327–335.

154. Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.-B., and Pollock, D.D. (2005). Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res. *15*, 665–673.

155. Polishchuk, L.V., and Tseitlin, V.B. (1999). Scaling of Population Density on Body Mass and a Number-Size Trade-Off. Oikos *86*, 544.

156. Damuth, J. (1981). Population density and body size in mammals. Nature *290*, 699–700.

157. White, C.R., and Seymour, R.S. (2005). Allometric scaling of mammalian metabolism. J. Exp. Biol. *208*, 1611–1619.

158. Arbeithuber, B., Hester, J., Cremona, M.A., Stoler, N., Zaidi, A., Higgins, B., Anthony, K., Chiaromonte, F., Diaz, F.J., and Makova, K.D. (2020). Age-related accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues. PLoS Biol. *18*, e3000745.

159. Belle, E.M.S., Piganeau, G., Gardner, M., and Eyre-Walker, A. (2005). An investigation of the variation in the transition bias among various animal mitochondrial DNA. Gene *355*, 58–66.

160. Montooth, K.L., and Rand, D.M. (2008). The spectrum of mitochondrial mutation differs across species. PLoS Biol. *6*, e213.

161. Von Stetina, J.R., and Orr-Weaver, T.L. (2011). Developmental control of oocyte maturation and egg activation in metazoan models. Cold Spring Harb. Perspect. Biol. *3*, a005553.

162. Sato, K., and Sato, M. (2017). Multiple ways to prevent transmission of paternal mitochondrial DNA for maternal inheritance in animals. J. Biochem. *162*, 247–253.

163. Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., and de Magalhães, J.P. (2013). Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res. *41*, D1027-33.

164. Ollason, J.G. (1987). R. H. Peters 1986. The ecological implications of body size. Cambridge University Press, Cambridge. 329 pages. ISBN 0-521-2886-x. Price: £12.50, US$16.95 (paperback). J. Trop. Ecol. *3*, 286–287.

165. Damuth, J. (1987). Interspecific allometry of population density in mammals and other animals: the independence of body mass and population energy-use. Biological Journal of the Linnean Society *31*, 193–246.

166. Hebert, P.D., Cywinska, A., Ball, S.L., and deWaard, J.R. (2003). Biological identifications through DNA barcodes. Proc. Biol. Sci. *270*, 313–321.

167. Ju, Y.S., Alexandrov, L.B., Gerstung, M., Martincorena, I., Nik-Zainal, S., Ramakrishna, M., Davies, H.R., Papaemmanuil, E., Gundem, G., Shlien, A., *et al.* (2014). Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. eLife *3*.

168. Saini, N., and Gordenin, D.A. (2020). Hypermutation in single-stranded DNA. DNA Repair (Amst) *91–92*, 102868.

169. Lehmann, G., Budovsky, A., Muradian, K.K., and Fraifeld, V.E. (2006). Mitochondrial genome anatomy and species-specific lifespan. Rejuvenation Res. *9*, 223–226.

170. Popadin, K., Polishchuk, L.V., Mamirova, L., Knorre, D., and Gunbin, K. (2007). Accumulation

of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. Proc Natl Acad Sci USA *104*, 13390–13395.

171. Nikolaev, S.I., Montoya-Burgos, J.I., Popadin, K., Parand, L., Margulies, E.H., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, and Antonarakis, S.E. (2007). Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. Proc Natl Acad Sci USA *104*, 20443–20448.

172. Popadin, K.Y., Nikolaev, S.I., Junier, T., Baranova, M., and Antonarakis, S.E. (2013). Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. Mol. Biol. Evol. *30*, 347–355.

173. Hughes, K.A., Alipaz, J.A., Drnevich, J.M., and Reynolds, R.M. (2002). A test of evolutionary theories of aging. Proc Natl Acad Sci USA *99*, 14286–14291.

174. Verechshagina, N., Nikitchina, N., Yamada, Y., Harashima, H., Tanaka, M., Orishchenko, K., and Mazunin, I. (2019). Future of human mitochondrial DNA editing technologies. Mitochondrial DNA A DNA Mapp. Seq. Anal. *30*, 214–221.

175. Gomes, A.P., Price, N.L., Ling, A.J.Y., Moslehi, J.J., Montgomery, M.K., Rajman, L., White, J.P., Teodoro, J.S., Wrann, C.D., Hubbard, B.P., *et al.* (2013). Declining NAD(+) induces a pseudohypoxic state disrupting nuclear-mitochondrial communication during aging. Cell *155*, 1624–1638.

176. Bellanti, F., Romano, A.D., Giudetti, A.M., Rollo, T., Blonda, M., Tamborra, R., Vendemiale, G., and Serviddio, G. (2013). Many faces of mitochondrial uncoupling during age: damage or defense? J. Gerontol. A Biol. Sci. Med. Sci. *68*, 892–902.

177. Stadtman, E.R. (2006). Protein oxidation and aging. Free Radic. Res. *40*, 1250–1258.

178. Ademowo, O.S., Dias, H.K.I., Burton, D.G.A., and Griffiths, H.R. (2017). Lipid (per) oxidation in mitochondria: an emerging target in the ageing process? Biogerontology *18*, 859–879.

179. Schriner, S.E., Linford, N.J., Martin, G.M., Treuting, P., Ogburn, C.E., Emond, M., Coskun, P.E., Ladiges, W., Wolf, N., Van Remmen, H., *et al.* (2005). Extension of murine life span by overexpression of catalase targeted to mitochondria. Science *308*, 1909–1911.

180. Hjelm, B.E., Ramiro, C., Rollins, B.L., Omidsalar, A.A., Gerke, D.S., Das, S.C., Sequeira, A., Morgan, L., Schatzberg, A.F., Barchas, J.D., *et al.* (2023). Large Common Mitochondrial DNA Deletions Are Associated with a Mitochondrial SNP T14798C Near the 3' Breakpoints. Complex Psychiatry *8*, 90–98.

181. Wolf, D.P., Hayama, T., and Mitalipov, S. (2017). Mitochondrial genome inheritance and replacement in the human germline. EMBO J. *36*, 2659.

182. Tan, B.G., Wellesley, F.C., Savery, N.J., and Szczelkun, M.D. (2016). Length heterogeneity at conserved sequence block 2 in human mitochondrial DNA acts as a rheostat for RNA polymerase POLRMT activity. Nucleic Acids Res. *44*, 7817–7829.

183. Gupta, R., Kanai, M., Durham, T.J., Tsuo, K., McCoy, J.G., Chinnery, P.F., Karczewski, K.J., Calvo, S.E., Neale, B.M., and Mootha, V.K. (2023). Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. medRxiv.

184. Torres-Gonzalez, E., and Makova, K.D. (2022). Exploring the effects of mitonuclear interactions on mitochondrial DNA gene expression in humans. Front. Genet. *13*, 797129.

185. Lechuga-Vieco, A.V., Latorre-Pellicer, A., Calvo, E., Torroja, C., Pellico, J., Acín-Pérez, R., García-Gil, M.L., Santos, A., Bagwan, N., Bonzon-Kulichenko, E., *et al.* (2022). Heteroplasmy of Wild-Type Mitochondrial DNA Variants in Mice Causes Metabolic Heart Disease With Pulmonary Hypertension and Frailty. Circulation *145*, 1084–1101.

186. Mertens, J., Regin, M., De Munck, N., Couvreu de Deckersberg, E., Belva, F., Sermon, K., Tournaye, H., Blockeel, C., Van de Velde, H., and Spits, C. (2022). Mitochondrial DNA variants segregate during human preimplantation development into genetically different cell lineages that are maintained postnatally. Hum. Mol. Genet. *31*, 3629–3642.

187. Spath, K., Babariya, D., Konstantinidis, M., Lowndes, J., Child, T., Grifo, J.A., Poulton, J., and Wells, D. (2021). Clinical application of sequencing-based methods for parallel preimplantation genetic testing for mitochondrial DNA disease and aneuploidy. Fertil. Steril. *115*, 1521–1532.

188. Fan, X.-Y., Guo, L., Chen, L.-N., Yin, S., Wen, J., Li, S., Ma, J.-Y., Jing, T., Jiang, M.-X., Sun, X.-H., *et al.* (2022). Reduction of mtDNA heteroplasmy in mitochondrial replacement therapy by inducing forced mitophagy. Nat. Biomed. Eng. *6*, 339–350.

189. Shamanskiy, V., Mikhailova, A.A., Tretiakov, E.O., Ushakova, K., Mikhailova, A.G., Oreshkov, S., Knorre, D.A., Ree, N., Overdevest, J.B., Lukowski, S.W., *et al.* (2023). Secondary structure of the human mitochondrial genome affects formation of deletions. BMC Biol. *21*, 103.

190. Tanaka, M., Gong, J.S., Zhang, J., Yoneda, M., and Yagi, K. (1998). Mitochondrial genotype associated with longevity. Lancet *351*, 185–186.

191. Kang, E., Wu, J., Gutierrez, N.M., Koski, A., Tippner-Hedges, R., Agaronyan, K., Platero-Luengo, A., Martinez-Redondo, P., Ma, H., Lee, Y., *et al.* (2016). Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. Nature *540*, 270–275.

192. Hyslop, L.A., Blakeley, P., Craven, L., Richardson, J., Fogarty, N.M.E., Fragouli, E., Lamb, M., Wamaitha, S.E., Prathalingam, N., Zhang, Q., *et al.* (2016). Towards clinical application of pronuclear transfer to prevent mitochondrial DNA disease. Nature *534*, 383–386.

193. Gorman, G.S., McFarland, R., Stewart, J., Feeney, C., and Turnbull, D.M. (2018). Mitochondrial donation: from test tube to clinic. Lancet *392*, 1191–1192.

194. Gollihue, J.L., Patel, S.P., and Rabchevsky, A.G. (2018). Mitochondrial transplantation strategies as potential therapeutics for central nervous system trauma. Neural Regen. Res. *13*, 194–197.

195. Ali Pour, P., Kenney, M.C., and Kheradvar, A. (2020). Bioenergetics consequences of mitochondrial transplantation in cardiomyocytes. J. Am. Heart Assoc. *9*, e014501.

196. Bacman, S.R., Williams, S.L., Pinto, M., Peralta, S., and Moraes, C.T. (2013). Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. Nat. Med. *19*, 1111–1113.

197. Gammage, P.A., Rorbach, J., Vincent, A.I., Rebar, E.J., and Minczuk, M. (2014). Mitochondrially targeted ZFNs for selective degradation of pathogenic mitochondrial genomes bearing large-scale deletions or point mutations. EMBO Mol. Med. *6*, 458–466.

198. Peeva, V., Blei, D., Trombly, G., Corsi, S., Szukszto, M.J., Rebelo-Guiomar, P., Gammage, P.A., Kudin, A.P., Becker, C., Altmüller, J., *et al.* (2018). Linear mitochondrial DNA is rapidly degraded by components of the replication machinery. Nat. Commun. *9*, 1727.

199. Jo, A., Ham, S., Lee, G.H., Lee, Y.-I., Kim, S., Lee, Y.-S., Shin, J.-H., and Lee, Y. (2015). Efficient mitochondrial genome editing by crispr/cas9. Biomed Res. Int. *2015*, 305716.

200. Loutre, R., Heckel, A.-M., Smirnova, A., Entelis, N., and Tarassov, I. (2018). Can mitochondrial DNA be crisprized: pro and contra. IUBMB Life *70*, 1233–1239.

201. Bian, W.-P., Chen, Y.-L., Luo, J.-J., Wang, C., Xie, S.-L., and Pei, D.-S. (2019). Knock-In Strategy for Editing Human and Zebrafish Mitochondrial DNA Using Mito-CRISPR/Cas9 System. ACS Synth. Biol. *8*, 621–632.

202. Hussain, S.-R.A., Yalvac, M.E., Khoo, B., Eckardt, S., and McLaughlin, K.J. (2020). Adapting crispr/cas9 system for targeting mitochondrial genome. BioRxiv.

203. Loutre, R., Heckel, A.-M., Jeandard, D., Tarassov, I., and Entelis, N. (2018). Anti-replicative recombinant 5S rRNA molecules can modulate the mtDNA heteroplasmy in a glucose-dependent manner. PLoS ONE *13*, e0199258.

204. Costa-Borges, N., Nikitos, E., Späth, K., Miguel-Escalada, I., Ma, H., Rink, K., Coudereau, C., Darby, H., Koski, A., Van Dyken, C., *et al.* (2023). First pilot study of maternal spindle transfer for the treatment of repeated in vitro fertilization failures in couples with idiopathic infertility. Fertil. Steril. *119*, 964–973.

205. Kirillova, A., and Mazunin, I. (2022). Operation “mitochondrial wipeout” - clearing recipient mitochondria DNA during the cytoplasmic replacement therapy. J. Assist. Reprod. Genet. *39*, 2205–2207.

206. Wallace, D.C. (2018). Mitochondrial genetic medicine. Nat. Genet. 50, 1642–1649.

207. Nikolova, E.N., Zhou, H., Gottardo, F.L., Alvey, H.S., Kimsey, I.J., and Al-Hashimi, H.M. (2013). A historical account of Hoogsteen base-pairs in duplex DNA. Biopolymers 99, 955–968.

208. Das, R., Kretsch, R.C., Simpkin, A.J., Mulvaney, T., Pham, P., Rangan, R., Bu, F., Keegan, R.M., Topf, M., Rigden, D.J., et al. (2023). Assessment of three-dimensional RNA structure prediction in CASP15. Proteins 91, 1747–1770.

209. Kretsch, R.C., Hummer, A.M., He, S., Yuan, R., Zhang, J., Karagianes, T., Cong, Q., Kryshtafovych, A., and Das, R. (2025). Assessment of nucleic acid structure prediction in CASP16. Proteins.

Приложения

Приложение А. Сравнение нашего алгоритма поиска повторов с ранее опубликованными

Геном	**Тип повторов**	**Разработанный мной алгоритм[1]**	**Vmatch несовершенные**[2]	**Vmatch совершенные**[3]	**RepEx[4]**	**Repeat-Around[5]**
Homo sapiens	прямые	6304 (6135 несов. 169 сов.)	2507 (2507 несов.) 1358 общих	320 сов., общих	-	333 сов., t.l.
	комплементарные	1694 (1654 несов. 40 сов.)	-	-	70 сов., общих	7 сов., t.l.
	зеркальные	5416 (5295 несов. 121 сов.)	-	-	252 сов., общих	83 сов., t.l.
	инвертированные	1939 (1868 несов. 71 сов.)	1984 (1974 несов., 10 сов.) 1937 общих	127 сов., общих	110 сов., общих	35 сов., t.l.

Продолжение приложения А

Геном	**Тип повторов**	**Разработанный мной алгоритм[1]**	**Vmatch несовершенные[2]**	**Vmatch совершенные[3]**	**RepEx[4]**	**Repeat-Around[5]**
Mus musculus	прямые	6765 (6594 несов. 171 сов.)	2543 (2543 несов.) 1325 общих	308 сов., общих	-	323 сов., t.l.
	комплементарные	3580 (3511 несов. 69 сов.)	-	-	143 сов., общих	50 сов., t.l.
	зеркальные	6029 (5871 несов. 158 сов.)	-	-	286 сов., общих	97 сов., t.l.
	инвертированные	3873 (3772 несов. 101 сов.)	3947 (3929 несов. 18 сов.) 3853 общих	195 сов., общих	179 сов., общих	63 сов., t.l.

[1] несов. и сов. обозначает несовершенные и совершенные повторы соответственно.

[2] Варианты прогона Vmatch для поиска несовершенных повторов: 1) для длины прямого повтора 10 допустимое расстояние Хемминга 1 (идентичность 90%), для длины прямого повтора от 11 до 100 допустимое расстояние Хемминга является целым числом L / 5, где L - повторы length (80% идентичности). Опция '-supermax' использовалась для всех повторяющихся длин, 2) для инвертированных длин повторений с длинами от 10 до 100 допустимое расстояние Хемминга варьируется от 1 до 10 для каждой длины (минимальная идентичность составляет от 90% до 80% с повторным ростом длины). После поиска повторов все двойники были проигнорированы, а также внутренние повторы (или подповторения) с меньшей длиной, чем искомый; все пересекающиеся повторы были объединены в более длинные.

[3] Варианты прогона Vmatch для поиска идеальных повторов: опция «-dentity 100», длина повторов от 10 до 100 для прямых и инвертированных повторов.

«Общий» обозначает общие шаблоны повторения между нашим алгоритмом и предыдущими тремя алгоритмами.

[4] Параметры запуска RepEx: минимальная длина 10; интервалы распорки больше 0; разрешено вырождение последовательности

[5] Параметры запуска RepeatAround: длина повторов от 10 до 256

RepeatAround «t.l.» обозначает типичные местоположения или, другими словами, местоположения, графически сопоставленные вручную с повторяющимися положениями, найденными нашим алгоритмом (подробности см. В тексте).

Приложение Б. Сравнение нашего алгоритма поиска повторов с Vmatch

Количество нуклеотидов между соседними несовпадениями в несовершенных повторах, найденных по нашему алгоритму и Vmatch в (A) Homo Sapiens мтДНК и (B) Mus musculus мтДНК