

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
“САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ”**

На правах рукописи

Добрынин Павел Владимирович

**СБОРКА И АННОТАЦИЯ ГЕНОМА АФРИКАНСКОГО ГЕПАРДА,
ACINONYX JUBATUS.**

1.5.7 – Генетика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:

Doctor of Philosophy (PhD), Cornell University, USA

кандидат биологических наук, Корнеллский Университет, США

профессор,

Стивен Джеймс О’Брайен

Санкт-Петербург – 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	5
Степень разработанности темы исследования.....	7
Цель и задачи диссертационной работы.....	9
Научная новизна.....	10
Теоретическая и практическая значимость работы.....	10
Методология и методы исследования.....	11
Положения, выносимые на защиту.....	12
Степень достоверности и апробация результатов.....	12
Личный вклад автора в проведенные исследования.....	14
Благодарности.....	15
ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ.....	16
1.1. Филогенетика и естественная история гепардов.....	16
1.2. Снижение численности популяций и генетическое разнообразие африканских гепардов в 19-20 веках.....	18
1.3. Обзор методологии работы с геномными данными.....	20
1.4. Обзор методов работы с геномными данными.....	21
1.4.1. Обзор методов получения сборки хромосомного уровня на основе референса.....	21
1.4.2. Обзор методов для поиска повторов.....	23
1.4.3. Обзор методов аннотации генов.....	24
1.4.4. Методы реконструкции демографической истории.....	25
ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ.....	27
2.1. Образцы тканей.....	27
2.2. Выделение ДНК, создание библиотек и секвенирование.....	27
2.3. Список биоинформатических инструментов.....	28
2.4. Оценка качества прочтений.....	29
2.4. Поиск и аннотация SNV.....	30
2.5. Способы оценки генетического разнообразия.....	31
2.6. Дополнительно использованные геномные данные.....	31

2.7. <i>De novo</i> сборка генома африканского гепарда.....	32
2.8. Сборка на основе референса.....	33
2.9. Оценка качества сборки генома.....	35
2.10. Поиск блоков синтении.....	36
2.11. Идентификация повторов.....	36
2.12. Аннотация генов.....	37
2.13. Поиск сигналов позитивного отбора.....	38
2.14. Анализ генных семейств.....	38
2.15. Анализ генов связанных с размножением.....	39
2.16. Анализ состава генов и генетического разнообразия для района МНС.....	40
2.17. Реконструкция демографической истории.....	40
ГЛАВА 3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ.....	43
3.1. Оценка данных секвенирования и сборка генома африканского гепарда, <i>Acinonyx jubatus</i>	43
3.1.1. Анализ данных секвенирования африканского гепарда для референса образца и шести дополнительных особей.....	43
3.1.2. <i>De novo</i> сборка генома африканского гепарда.....	45
3.1.3. Сборка генома африканского гепарда на основе референса домашней кошки.....	47
3.1.4. Оценка качества геномной сборки.....	49
3.2. Анализ генов и повторов в геноме африканского гепарда, <i>Acinonyx jubatus</i>	56
3.2.1. Идентификация повторов и генов в геноме африканского гепарда.....	56
3.2.2. Поиск сигналов положительного отбора.....	58
3.2.3. Анализ экспансии и сокращения генных семейств.....	60
3.2.4. Анализ генов связанных с размножением.....	63
3.3. Изучение генетического разнообразия в свободноживущих популяциях африканского гепарда, <i>Acinonyx jubatus</i>	71
3.3.1. Изучение генетического разнообразия.....	71

3.3.2. Сборка района МНС при помощи референсной последовательности...	78
3.4. Изучение популяционной истории африканского гепарда на основе данных полногеномного секвенирования.....	84
ЗАКЛЮЧЕНИЕ.....	90
ВЫВОДЫ.....	93
Список сокращений.....	95
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ.....	96

ВВЕДЕНИЕ

Чарльз Дарвин описывал фенотипическую изменчивость как отражение процесса естественного отбора, в котором биологические признаки становятся более или менее распространенными для всей популяции. В ходе естественного отбора виды должны адаптироваться к условиям окружающей среды и вступать в борьбу за "выживание наиболее приспособленного", где вариация признаков неизбежно вызывает различия в способности индивидуумов к выживанию и размножению, в чем и выражается приспособленность.

Африканские гепарды (*Acinonyx jubatus*) являются уникальными обитателями Африки к югу от Сахары и Юго-Западной Азии. Данный вид относится к семейству Кошачьих (*Felidae*), одному из 16 семейств, входящих в отряд Хищных млекопитающих (*Carnivora*). По данным IUCN Red List за 2021 год, 39 представителей семейства *Felidae* находятся под угрозой вымирания в той или иной степени (10 – CR, 14 – EN, 15 – VU). Представители данного семейства встречаются на всех континентах, кроме Австралии и Антарктиды.

Гепард эволюционировал как превосходный, стремительный хищник, достаточно проворный, чтобы преследовать добычу и избегать конкуренции с другими хищниками. Однако, сегодня он все еще борется за выживание в своей среде обитания. Одним из приспособлений гепарда к выживанию являются морфо-функциональные особенности строения ногтей, рассчитанные на быстрое передвижение. Когти гепарда, в отличие от когтей остальных кошек, втягиваются только частично и приспособлены как для охоты, так и для лазания по деревьям, чтобы поймать добычу или избежать нападения. Их когти обеспечивают надежное сцепление с поверхностью и делают возможным ускорение до 100 километров в час за три секунды [Wilson et al., 2013]. Крепкое сцепление с поверхностью повышает маневренность и

дает возможность совершать резкие повороты на высокой скорости в случае, когда добыча меняет направление движения, чтобы избавиться от преследования. Скорость, которой может достигнуть гепард, поразительна и используется им для ловли добычи, которая легче и быстрее, чем Африканские буйволы (*Cape buffalo*), зебры (*Equus quagga*) и антилопы гну (*Connochaetes taurinus*). Типичная для гепарда добыча - газели (*Eudorcas thomsonii*) и небольшие антилопы (*Raphicerus campestris*, *Cephalophus natalensis*, *Antidorcas marsupialis*). Данный навык также помогает избежать сильных хищников, таких как львы (*Panthera leo*) и пятнистые гиены (*Crocuta crocuta*). Другими адаптациями гепардов, появившимися в ходе сотен тысяч лет эволюции, являются: небольшие зубы, так как их добыча достаточно легкая и небольшая; увеличенные полости носа, позволяющие вдыхать большие объемы воздуха в процессе погони; пятнистый окрас шерсти, позволяющий как скрываться во время охоты, так и уклоняться от хищников; а также «отпечатки слез» или черные вертикальные линии на морде, защищающие от солнечных бликов. Помимо вышеперечисленных, ряд физиологических особенностей демонстрирует глубокую специализацию, связанную с уникальным способом охоты этих животных [Hudson et al., 2011; Goto et al., 2012; Hyatt et al., 2010]. Каждый раз процесс охоты оказывает серьезное влияние на физиологическое состояние гепарда, особенно на частоту дыхания, артериальное давление и регуляцию температуры тела. В результате гепарды находятся в состоянии крайнего истощения после погони за своей добычей. В естественной среде конкуренты гепардов, такие как пятнистые гиены, африканские львы и гиеновидные собаки (*Lycaon pictus*), используют эту особенность гепарда как возможность забрать у него добычу [Scantlebury et al., 2014].

Популяционная история вида и разнообразие специализированных адаптационных механизмов делают гепардов крайне интересным объектом изучения на уровне генома, в частности, в связи с возможностью обнаружения адаптаций, оставивших след в геноме гепардов. Данное исследование является важным, так как касается вопросов эволюции генома и молекулярных основ

механизмов физиологической адаптации гепардов, давая возможность совершенно по-новому взглянуть на биологию этого интересного вида.

Степень разработанности темы исследования

Риски, связанные с небольшой численностью популяции и преобладанием в ней скрещиваний среди близкородственных особей (инбридинг), были известны человечеству на протяжении тысячелетий. В начале XIX века Дарвин описал возможные последствия близкородственных скрещиваний у домашних животных в книге «Происхождение видов» [Darwin, 1859]. Экспериментальные доказательства морфологических и/или физиологических отклонений, связанных с инбридингом, были получены спустя четверть века. Затем, во второй половине XX века, коллектив ученых во главе с Кэти Роллс изучил корреляцию уровня смертности потомства и генетическую близость родителей у разных видов животных в зоопарках [Ralls et al., 1988; Ralls and Ballou, 1982]. Они не только показали, что уровень смертности потомства от близкородственных пар был выше, но и определили вклад инбридинга в уровень смертности новорожденных в зоопарке. Как выяснилось, африканские гепарды имели не только один из наиболее высоких уровней смертности детенышей (30-40%), но один из наиболее высоких уровней инбридинга [O'Brien et al, 1983].

Генетические исследования гепардов, начатые в 80-е годы XX века, показали, что у данного вида практически полностью отсутствует генетическое разнообразие, но применяемые тогда методы не имели достаточного разрешения, чтобы его обнаружить. С тех пор, в течение 30 лет, были проведены исследования с использованием различных генетических маркеров, полностью подтвердившие первоначальное заявление о том, что гепарды имеют низкое генетическое разнообразие, а по многим локусам не имеют его вовсе. Используемые маркеры и методы включали: анализ аллозимов в фибробластах с использованием двумерного электрофореза [O'Brien et al., 1983]; оценку полиморфизма длин рестрикционных фрагментов

(ПДРФ) в локусах МНС [Yuhki and O'Brien, 1990]; изучение асимметрии костей черепа [Wayne et al., 1986], секвенирование генов *МНС-I* и *МНС-II* [Castro-Prieto et al., 2011]; секвенирование митохондриальной ДНК [Freeman et al., 2001]; изучение полиморфизма генотипов, полученных при амплификации микросателлитных локусов [Driscoll et al., 2002].

Результаты всех вышеперечисленных тестов показали снижение уровня генетического разнообразия на 90-99% по сравнению с человеком или другими видами диких животных. Один из наиболее наглядных экспериментов связан с трансплантацией кожи между неродственными особями. Этот эксперимент показал влияние низкого уровня генетического разнообразия у гепардов. В ходе эксперимента исследователи обнаружили, что небольшие кусочки трансплантируемой кожи не были отторгнуты реципиентами. Это свидетельствует о том, что аллели генов иммунного распознавания были настолько идентичны у неродственных особей, что не воспринимались как чужеродные [O'Brien et al., 1985]. У гепардов очень мало аллелей генов *МНС-I* и *МНС-II*, что нетипично для диких животных. Кроме того, низкий уровень генетического разнообразия гепардов коррелирует с высоким уровнем смертности среди новорожденных, трудностями с размножением в неволе, серьезными проблемами в образовании сперматозоидов и повышенной уязвимостью к вспышкам инфекционных заболеваний [Heeney et al., 1990; Pearks Wilkerson et al., 2004]. В настоящее время гепарды являются ярким примером влияния низкого генетического разнообразия на физиологию. Потеря генетического разнообразия у современных гепардов была широко изучена и проверена [May, 1995]. Тем не менее, результаты большинства проведенных анализов согласуются с гипотезой о том, что снижение генетического разнообразия гепардов является результатом прохождения популяции через одно или несколько бутылочных горлышек начиная с Плейстоцена [Castro-Prieto et al., 2011; Caro and Laurenson, 1994; O'Brien et al., 2005].

Изучение биологии гепарда и оценка вероятности вымирания вида проводилось учёными и экологами на протяжении нескольких десятилетий. Идентификация низкого уровня генетического разнообразия гепардов, нарушенной морфологии сперматозоидов и, в конечном счете, их восприимчивости к летальным вирусным инфекциям, обеспечивают уникальные условия для детального анализа рисков и последствий низкого генетического разнообразия методами секвенирования следующего поколения на уровне полного генома.

Цель и задачи диссертационной работы

Основной целью данного исследования было полногеномное изучение генетического разнообразия, адаптационных механизмов и демографической истории африканского гепарда, находящегося под угрозой исчезновения.

Для достижения поставленной цели были выдвинуты следующие задачи:

- 1) Секвенирование и сборка высококачественного референсного генома на основе одного образца африканского гепарда;
- 2) Ресеквенирование шести дополнительных образцов, представляющих два подвида, и картирование ридов на референсный геном;
- 3) Аннотация генов, повторов и других элементов генома методами, базирующимися на поиске гомологов, и алгоритмами для *ab initio* предсказания генов;
- 4) Идентификация генетических основ адаптационных механизмов в геноме гепарда с применением методов сравнительного геномного анализа;
- 5) Изучение генетического разнообразия среди семи геномов гепардов методом детекции однонуклеотидных вариантов (SNV);
- 6) Сборка районов *MHC* гепарда и сравнение их с гомологичными районами других представителей млекопитающих;
- 7) Анализ демографической истории двух субпопуляций африканских гепардов.

Научная новизна

В рамках работы впервые представляется первичная сборка генома *de novo* и подробная аннотация генов и генетических элементов африканского гепарда. Собранный геном гепарда представляет собой первый публично доступный геном вида, относящегося к линии *Puma*. В данном исследовании был разработан новый метод сборки с применением референсного генома, который оказался более точным по сравнению с другими методами. В геноме гепарда были описаны ранее неизвестные способы адаптации к высокоскоростному бегу и плотоядному типу питания. Нами не только была точно описана потеря генетического разнообразия по всему геному, но и реконструирована цепочка исторических событий, приведших африканского гепарда к нынешнему состоянию.

Теоретическая и практическая значимость работы

Теоретическая значимость работы определяется тем, что она посвящена одной из основных фундаментальных проблем сохранения существующих в настоящее время биологических видов – пониманию эффектов снижения генетического разнообразия на геном и фенотип у исчезающих видов в результате резкого сокращения численности популяции – феномена бутылочного горлышка и эффекта основателя.

Нами были разработаны методический и биоинформатический вычислительные подходы для аннотирования генома млекопитающих и анализа генетического разнообразия. Следует отметить, что, несмотря на увеличивающиеся объемы геномных данных, проблемы при работе с таким типом данных на немодельных организмах все еще остаются.

Данные полногеномного секвенирования и первичной сборки генома африканского гепарда были размещены в базе данных NCBI GenBank, и в настоящее время находятся в открытом доступе. Вся информация об описанных вариантах, повторяющихся элементах, генах и структурных

вариантах доступна в геномном браузере GARFIELD (<http://garfield.dobzhanskycenter.org/>).

Описаны вредные мутации в гене *AKAP4*, которые присутствуют у большинства гепардов. Подобраны праймеры для амплификации локуса с данными мутациями. Эффективность праймеров была подтверждена на двадцати дополнительных образцах диких африканских гепардов.

Результаты, полученные для генома гепарда, могут быть использованы в последующих работах, связанных с биологией и сохранением исчезающих видов.

Методология и методы исследования

В работе использовались современные и актуальные методы, большая часть которых была специально разработана или модернизирована под цели и задачи данного проекта.

К ним относятся:

- Сборка генома *de novo* и с применением референсного генома;
- Выравнивание последовательностей ДНК и белков;
- Множественное полногеномное выравнивание;
- Идентификация генов, повторов и других геномных элементов, включая анализ семейств генов и признаков положительного отбора;
- Оценка генетического разнообразия на уровне генома;
- Реконструкция исторической демографии с использованием аллель частотных спектров.

Положения выносимые на защиту

1. Геном африканского гепарда отличается крайне низким генетическим разнообразием по сравнению с другими близкородственными видами и иными млекопитающими.
2. Геномный район *MHC* у гепарда не только содержит сравнительно небольшое число аллелей *MHC* классов I и II, но и демонстрирует потерю до 4 генов *MHC* класса I, по сравнению с геномом домашней кошки.
3. В результате демографического анализа генома африканского гепарда были найдены доказательства прохождения популяции через два бутылочных горлышка.
4. Семейства генов, связанных с размножением, у африканских гепардов аккумулируют избыточное количество несинонимичных замен, часть из которых оказалась предположительно вредной по своей природе. Кроме того, ряд серьезных мутаций в гене *AKAP4* скорее всего являются причиной формирования дефектных сперматозоидов у гепардов.
5. Признаки положительного отбора и амплификации генных семейств показали крайне специфическую адаптацию, ассоциированную с быстрым бегом и физической выносливостью гепарда.

Степень достоверности и апробация результатов

Основные результаты данной работы были опубликованы в семи рецензируемых публикациях в международных журналах:

1. Tamazian G. Annotated features of domestic cat – *Felis catus* genome / G. Tamazian, S. Simonov, P. Dobrynin et al. // GigaScience. – 2014. – Т.3. – №.1. – С.13.
2. Dobrynin, P. Genomic legacy of the African cheetah, *Acinonyx jubatus* / P. Dobrynin, S. Liu, G. Tamazian et al. // Genome biology. – 2015. – Т.16. - №.1. – С.277.

3. O'Brien, S.J. Response to Comment by Faurby, Werdelin and Svenning / S.J. O'Brien, K.P. Koepfli, E. Eizirik et al. // *Genome biology*. – 2016. – T.17. - №.1. – C.90.
4. Tamazian G. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences // G. Tamazian, P. Dobrynin, K. Krashennnikova et al. // *GigaScience*. – 2016. – T.5. - №.1. – C.38.
5. Noskova E. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data / E. Noskova, V. Ulyantsev, K.P. Koepfli et al. // *GigaScience*. – 2020. – T.9. - №.3. – C.giaa005.
6. Tamazian G. Draft *de novo* Genome Assembly of the Elusive Jaguarundi, *Puma yagouaroundi* / G. Tamazian, P. Dobrynin, A. Zhuk et al. // *Journal of Heredity*. – 2021. – T.112. - №.6. – C.540-548.
7. Fyodorova A. Genome-wide detection of positive selection in new African cheetah assembly / A. Fyodorova, A. Zhuk, P. Dobrynin // *BMC Bioinformatics*. – 2021. – T.22. - №.S16.

Данные результаты также были представлены на следующих конференциях: Society for Molecular Biology and Evolution meeting (Австрия, 2015), Recent Advances in Conservation Genetics Course (Венгрия, 2016) и BiATA (Санкт-Петербург, 2021).

Обсуждение диссертации проводилось на семинарах в Центре геномной биоинформатики им. Ф.Г.Добржанского, Санкт-Петербургский Государственный Университет и научно образовательном центре - Геномное разнообразие в университете ИТМО.

Личный вклад автора в проведенные исследования.

Выбор темы диссертационной работы, определение цели и постановка задач исследования автор провел совместно с научным руководителем PhD, профессором, Стивеном Джеймсом О'Брайеном. Автор работы изучил литературу по теме, непосредственно участвовал в написании и подготовке материалов к публикациям, написал рукопись работы. Основная часть экспериментальной работы: сборка генома и его фрагментов, выравнивание последовательностей, поиск генов и определение генных семейств, анализ сигналов позитивного отбора, анализ демографической истории выполнены автором лично. Общий вклад автора в работу - более 80%.

Благодарности

Эта диссертационная работа не была бы возможна без поддержки и неоценимой помощи многих людей, которых я хочу искренне поблагодарить.

Прежде всего, я глубоко признателен моему научному руководителю, доктору Стивену О'Брайену за его терпение и отзывчивость на протяжении всего периода подготовки диссертационной работы, а также за бесценные советы и опыт, которыми он всегда охотно делился.

Также я хочу выразить искреннюю благодарность доктору Клаус-Питеру Копфли, исследователю из Смитсоновского института и старшему научному сотруднику НОЦ Геномное разнообразие в университете ИТМО, под руководством которого был освоен и проведен филогенетический анализ, за его терпение, поддержку и всестороннюю помощь.

Огромное спасибо Шипинг Лиу и его команде Геномного анализа из института BGI (Китай), благодаря помощи которых были проведены первичная сборка и анализ исходных данных. Я восхищен дружественной рабочей атмосферой лаборатории Института, которая позволила мне выполнить огромную часть работы во время моего визита.

Я благодарен коллективу Центра геномной биоинформатики им. Ф.Г.Добржанского (Россия), Фонду сохранения гепардов (Намибия) и коллективу НОЦ Геномное разнообразие в ИТМО. Ценный и разносторонний опыт и личные усилия многих коллег сделали эту работу возможной.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Филогенетика и естественная история гепардов

В 1777 гепардам присваивают имя – *Felis jubatus*, позже в 1828 году их перемещают в отдельный род *Acinonyx*, спустя практически сто лет в 1917 году, на основании морфологических данных, гепардов выделяют в отдельное подсемейство – *Acinonychinae*. В последующие годы с появлением новых методов и молекулярных маркеров систематика гепардов была снова пересмотрена и их классифицировали в подсемейство – *Felinae*. Сейчас гепард является единственным представителем своего рода *Acinonyx*, которого вместе с двумя представителями рода *Puma* (*P. concolor* и *P. yagouaroundi*) выделяют в отдельную линию Пумы (*Puma lineage*). Время разделения гепардов, пум и ягуарунди на отдельные виды оценивается в пределах 6,7 млн. л. н.

Возраст самых старых окаменелостей гепарда, обнаруженных в восточной и южной частях Африки, составляет 3,5–3 млн. л. н.; самый ранний известный образец из Южной Африки происходит из самых нижних отложений грота Зильберберг [Krausman et al., 2005]. Несмотря на ограниченность и фрагментарность этих данных, можно сделать вывод о том, что предковая форма гепарда была крупнее и менее адаптирована к бегу [Valkenburgh et al., 2018]. Ископаемые останки из Европы ограничены несколькими образцами среднего плейстоцена из Хундсхайма (Австрия) и песков Мосбах (Германия) [Hemmer et al., 2008]. Известны кошки, похожие на гепардов, обитавшие 10.000 лет назад в Старом Свете. Гигантский гепард (*A. pardinensis*), значительно более крупный и вероятно более медленный по сравнению с современным гепардом, обитал в Евразии, а также на востоке и юге Африки примерно 3,8–1,9 млн. л. н. [Cherin et al., 2014] В среднем плейстоцене менее крупный гепард, *A. intermedius*, обитал на территориях от Европы до Китая [Krausman et al., 2005]. Современный гепард появился в Африке около 1,9 млн лет назад; его летопись окаменелостей ограничена Африкой [Valkenburgh et al., 2018].

Вымершие североамериканские гепардоподобные кошки исторически классифицировались как *Felis*, *Puma* или *Acinonyx*; два таких вида, *F. studei* и *F. trumani*, считались более близкими к пуме, чем к гепарду, несмотря на их близкое сходство с последним. Дэниел Адамс предложил *Miracinonyx*, как новый подрод *Acinonyx*, позже он был возведен в ранг рода [Adams, 1979]. Адамс указал, что гепардоподобные кошки Северной Америки и Старого Света могли иметь общего предка, а гепард мог появиться в Северной Америке, а не в Евразии. Однако, до сегодняшнего времени нет однозначных данных, позволяющих достоверно доказать, что *Miracinonyx* ближе к пумам или гепардам [ref]. В публикации [Dobrynin et al., 2015], лежащей в основе текущего исследования, мы склоняемся к тому, что гепард появился в Северной Америке, но этот сценарий не является единственно возможным, который способен описать наблюдаемые данные.

В настоящее время гепарды встречаются только в Африке и Иране, мы считаем, что имеется достаточно доказательств того, что представители рода возникли от одного общего предка, который жил на территории Северной Америки в Миоцене [Johnson et al., 2006]. Такая оценка основана на сравнительном анализе молекулярных маркеров митохондриальной и ядерной ДНК, ископаемых данных и сведений о современном распространении видов. Палеонтологические следы существования гепарда и гепардоподобных кошек можно встретить в Северной и Южной Америке, Европе и Азии вплоть до позднего Плейстоцена (10.000 – 12.000 лет назад). Это время примерно соответствует периоду вымирания мегафауны, которое затронуло более чем 40 видов крупных млекопитающих, включая гепардо-подобных кошек (*Miracinonyx*) и пум в Северной Америке [Culver, 2000; Werdelin et al., 2010; Hewitt, 2000].

1.2. Снижение численности популяций и генетическое разнообразие африканских гепардов в 19-20 веках

История тесного взаимодействия человека и гепардов насчитывает около 5 тысяч лет. На протяжении сотен лет этих хищников содержали при королевских дворах и готовили к охоте, в то же время на них велась охота, которая продолжается и сегодня [Marker et al., 2003]. С начала 20 века ареал обитания гепардов стремительно сокращается и в настоящее время представляет собой отдельные фрагментированные области. Общая численность гепардов в начале 20 века в Африке и Азии достигала 100.000 особей, однако в течение 100 лет численность уменьшилась почти на 90%, в то время как население земли удвоилось всего за четверть века. В настоящее время численность гепардов варьирует от 9.000 до 12.000 особей [Marker et al., 2003; Marker, 2005; Durant, 2000; Durant et al., 2017]. Гепард занесен в приложение I Конвенции о международной торговле видами дикой фауны и флоры, находящимися под угрозой исчезновения (СИТЕС) и классифицирован Международным союзом охраны природы и природных ресурсов (МСОП) как вымирающий вид [Marker, 2005].

Гепарды сталкиваются с различными трудностями в дикой природе [Durant et al., 2017]. Два основных фактора ответственны за то, что сейчас гепарды оказались на грани вымирания: 1) Потеря среды обитания, незаконная международная торговля и конфликты с владельцами скота 2) Растущие темпы охоты на антилоп, газелей и импал приводят к сокращения кормовой базы. Таким образом африканские гепарды сталкиваются с комплексной проблемой - их убивают напрямую, а затем в этих районах саванны убивают и их добычу. Кроме этого, существует ряд особенностей, связанных с их поведением и физиологией, что делает гепардов особенно чувствительными к сокращению численности. Низкая репродуктивная способность самок гепардов ограничивает рост популяции, так же как и конкурентные отношения с другими хищниками [Laurenson, 1994; Caro et al., 1994]. Дополнительным фактором является предположительно резкое и существенное сокращение

численности и, как следствие, потеря генетического разнообразия около 15 т.л.н. Небольшая эффективная численность популяции в сочетании с уменьшением количества гетерозиготных аллелей может также влиять на приспособленность и демографическую ситуацию. Сочетание представленных факторов оказывает влияние на выживаемость гепардов в современных условиях в связи с дополнительным снижением количества аллелей в генофонде, что приводит к уменьшению генетического разнообразия данной популяции или вида в долгосрочной перспективе и, следовательно, снижает эволюционный потенциал. Кроме того, конкуренция за пищевые ресурсы также имеет серьёзное влияние. В настоящее время гепарды образуют фрагментированные популяции и вынуждены вступать в конкурентную борьбу с другими плотоядными животными за пищевые ресурсы, которых может быть недостаточно в некоторых частях Африки. Более того, во время охоты гепарды подвержены нападениям со стороны других хищников. Ощутимое влияние на численность гепардов оказывают львы и гиены, либо из-за эксплуатационной конкуренции, либо в результате прямой атаки [Durant, 2000]. Пятнистые гиены, львы и леопарды совершают нападения на гепардов, превосходя их в количестве, а иногда и в силе, убивая молодых особей и выступая в качестве еще одного конкурента за пищевые ресурсы. В последнее время самым сильным из факторов, влияющих на существование гепардов, является антропогенный. Люди вытесняют гепардов из их мест обитания, охотятся на них ради трофеев и отлавливают их детенышей для продажи в качестве экзотического домашнего животного.

Проведенные в начале 80-х годов 20 века первые исследования низкого генетического разнообразия в популяциях гепардов, находящихся под угрозой исчезновения, имели большое значение. Гепард стал символом работ, направленных на сохранения видового разнообразия, а также способствовал расширению наблюдений за генетическим разнообразием всех видов, находящихся под угрозой исчезновения. В настоящее время каждая схема контроля за исчезающим видом основной целью в долгосрочной перспективе

ставит максимизацию оптимальных скрещиваний для поддержания генетического разнообразия. Предыдущие работы были сделаны на основании ограниченного набора локусов, и таким образом, среди основных задач данного исследования было определение и трактовка численных параметров полного геномного разнообразия среди гепардов на уровне полногеномных данных.

1.3. Обзор методологии работы с геномными данными

Появление высокопроизводительных технологий секвенирования позволило без больших финансовых затрат расшифровать и собрать геномы немодельных видов. Кроме того, высокопроизводительное секвенирование было использовано для повторного секвенирования геномов человека и других видов для сравнения с существующим референсным геномом для анализа генетического полиморфизма: выявления мутаций и генетических вариаций внутри и между различными популяциями. Число прочтений и их короткий размер накладывают несколько вычислительных ограничений для обнаружения и анализа некоторых геномных изменений, таких как инсерции, делеции и структурные перестройки. В этом исследовании мы использовали хорошо зарекомендовавшие себя алгоритмы и программное обеспечение для поиска одиночных нуклеотидных вариантов (SNV) в семи геномах гепарда (одна референсная сборка и шесть повторно секвенированных образцов). Идентификация небольших вставок/делеций и небольших перестроек выходит за рамки нашей работы для этого проекта.

В течение последних 30 лет было проведено несколько исследований, посвященных изучению низкого генетического разнообразия гепарда, которое является крайне необычным для диких животных. Также их низкое генетическое разнообразие коррелирует с увеличением смертности среди новорожденных детёнышей, экстремальными аномалиями в развитии сперматозоидов и повышенной уязвимостью к вспышкам инфекционных

заболеваний [Wayne et al., 1986; O'Brien et al., 1985; Pearks Wilkerson et al., 2004; O'Brien et al., 1983].

1.4. Обзор методов работы с геномными данными

1.4.1. Обзор методов получения сборки хромосомного уровня на основе референса

Хромосомная сборка является заключительной и наиболее значимой стадией практически каждого геномного анализа эукариотического организма. С использованием коротких прочтений и имеющихся алгоритмов сборки генома *de novo* в лучшем случае получают длинные скаффолды, в то время как для реконструкции скаффолда хромосомного уровня необходимы дополнительные источники информации, такие как физические карты или карты сцепления [Lewin et al., 2009], или дополнительные очень длинные прочтения. Для млекопитающих за последние пять лет были сделаны несколько сборок генома *de novo*, часть из которых связана с физическими картами хромосом [Pontius et al., 2007; Mavrich et al., 2008; McPherson et al., 2001]. Наличие подобных высококачественных «референсных геномов» предоставляет возможность для альтернативного подхода - сборки скаффолдов хромосомного уровня на основе референса геномов других видов. Сборка с применением референсного генома, как правило, включает несколько этапов. Прежде всего, в геномах близкородственных видов необходимо найти последовательности, гомологичные имеющимся собранным *de novo* фрагментам. Затем использовать информацию о положении и ориентации гомологичных фрагментов для сборки более продолжительных последовательностей изучаемого (целевого) генома.

Использованный нами подход обладает многими преимуществами для работы с собранными последовательностями хромосом, даже без применения дополнительного секвенирования или построения физической карты. К таким преимуществам относятся определение групп сцепления и оценка расстояний между маркерами, что необходимо для оценки степени вероятности сцепленного наследования признаков и синтении. Сборка генома с

использованием референса также упорядочивает, ориентирует и связывает многие контиги и/или скаффолды на основе результатов сравнительного анализа. Для многих проектов классический алгоритм сборки генома будет производить до нескольких сотен тысяч контигов и скаффолдов. Такая сильно фрагментированная сборка генома усложняет процесс анализа данных после предварительной обработки [Murchison et al., 2012]. Однако, упорядочивание контигов и скаффолдов в предполагаемые хромосомы с использованием информации о референсном геноме близкородственного вида уменьшит общее количество фрагментов с тысяч до сотен, или даже десятков, а также упростит аннотацию и анализ геномных элементов, таких как тандемные и множественные повторы, гены, однонуклеотидные замены, CNV и сегментные дупликации.

К сожалению, в описанном выше подходе существует серьезный недостаток – вероятность появления дополнительных ошибок в геномной сборке из-за наличия хромосомных перестроек между референсным и изучаемым видом. Даже близкородственные виды могут в некоторой степени отличаться организацией хромосом от изучаемого генома. Мы ожидаем, что количество привнесенных в геномную сборку ошибок в целом будет коррелировать с эволюционным расстоянием между целевым и референсным геномами [Luo et al., 2012]. Тем не менее, артефакты такого рода нетрудно скорректировать в дальнейшем, при наличии физической карты или карты сцепления. В настоящее время существуют методы третьего поколения, которые могут помочь с созданием карт такого типа. К ним относятся: оптическое картирование (BioNano Genomics) [Li et al., 2016], секвенирование сцепленных ридов (10X Genomics) [Coombe et al., 2016] и Hi-C, метод фиксации конформации хромосом (например, Phase Genomics) [Korbel and Lee, 2013].

1.4.2. Обзор методов для поиска повторов

После получения высококачественной референсной сборки для гепарда следующим этапом в исследовании генома является аннотация различных элементов, которые содержит в себе геном, включая: повторы ДНК, кодирующие белки гены, консервативные последовательности, мобильные генетические элементы, однонуклеотидные полиморфизмы (SNV), и вариации числа копий генов (CNV). Мы начали наш вычислительный анализ с идентификации повторов ДНК в собранном геноме гепарда.

Термин «повтор» в биоинформатике используется для определения двух различных типов последовательностей: 1) тандемных повторов, также известных как последовательности с низкой сложностью, где многократная повторяемость одного или тысяч нуклеотидов повторяется множество раз, и повторения примыкают друг к другу (например, микросателлиты, минисателлиты) [Benson, 1999; Parisi et al., 2003]; 2) рассеянные повторы (TE), также известные как «прыгающие гены» или транспозоны, представляющие собой последовательности ДНК, которые перемещаются (или прыгают) из одного места в геноме к другому посредством различных механизмов вырезания и способов амплификации ДНК [Finnegan, 1992]. Эта группа состоит из нескольких различных типов повторов, таких как эндогенные ретровирусы (ERV), ретротранспозоны с длинным концевым повторением (LTR), длинные диспергированные ядерные элементы (LINE) и короткие диспергированные ядерные элементы (SINE). [расширить описание]

Известно, что геномы млекопитающих различаются по своему содержанию повторов. Например, почти 50% секвенированных геномов приматов состоит из повторов [Rogers and Gibbs, 2014], тогда как для домашней кошки и домашней собаки содержание повторов составляет 39,5% и 40% соответственно [Tamazian et al., 2014; Cho et al., 2013]. Вероятно, большинство современных методов, используемых для идентификации повторов, недооценивают количество повторов в геноме из-за того, что эти методы используют только собранные скаффолды и контиги для анализа, а

чаще всего именно сложные повторы или богатые повторами области генома (например, прицентромерные и прителомерные) оказываются не собранными и, следовательно, исключаются из дальнейшего анализа. В результате эти оценки содержания следует рассматривать как нижний предел доли повторов, содержащихся в геноме. Также существует проблема, что границы повторов, как правило, недостаточно четко определены, или просто отсутствуют в финальной сборке. Чаще всего повтор обнаруживается только в виде фрагментов. В общем случае повторы, которые не замаскированы должным образом, затрудняют дальнейший анализ генома, поэтому они должны быть правильно аннотированы [Yandell and Ence, 2012].

1.4.3. Обзор методов аннотации генов

Методы *ab initio* предсказания генов в последовательностях ДНК появились в 1990-х [Guigo et al., 1992] и открыли совершенно новую область анализа генома. Инструменты предсказания гена *ab initio* называются так, поскольку они не требуют сторонних доказательств для обнаружения кодирующих последовательностей (CDS), таких как последовательности мРНК или белков. Для определения экзон-интронной структуры гена в этих методах используются математические модели, созданные на основе ранее известных генов. Несмотря на то, что они обеспечивают легкий доступ к последовательностям генов и их структуре во вновь собранных геномах, эти инструменты имеют некоторые существенные практические ограничения: 1) обычно идентифицируется только один CDS на ген, что затрудняет предсказание альтернативных вариантов сплайсинга; 2) последовательности из нетранслируемых областей (UTR) не включены в предсказание; 3) и, самое главное, все инструменты *ab initio* чрезвычайно чувствительны к качеству математической модели, используемой для определения генной структуры. Существует ряд предварительно вычисленных моделей генов для некоторых классических видов и модельных организмов. Для предсказания гена, примененного к новой сборке генома, нужно будет обучить программу,

используя набор генов, специфичных для собранного генома. Можно также использовать гены из родственных видов, но это приведет к менее точным предсказаниям, поскольку длина экзона / интрона, частоты используемых кодонов и GC состав могут отличаться даже у близких видов [Korf, 2004]. Внешние источники информации, такие как мРНК, EST и множественные геномные выравнивания, могут быть использованы в качестве дополнительного доказательства в процессе предсказания гена для существенного улучшения результатов [Stanke et al., 2006].

Качество предсказанных последовательностей генов зависит от двух факторов: размера набора обучающих данных и качества *de novo* сборки генома. Методами предсказания генов можно достигать 100% точности даже в эукариотических геномах [Guigo and Reese, 2005].

1.4.4. Методы реконструкции демографической истории

Чтобы объяснить эволюционные изменения видов и их популяций, очень важно знать, какие события произошли в их прошлом. Генетическая структура и разнообразие видов формировались в результате таких событий, как миграция, дивергенция популяций и изменения (эффективного) размера популяции в течение тысяч или миллионов лет. Следы демографической истории запечатлены в геномах представителей каждого вида. С ростом технологий NGS и обширными данными о строении генома стало возможным исследование сложных демографических моделей, которые включают в себя оценку скорости мутаций, изменений в численности популяции, не случайное спаривание, миграцию и отбор. Существует две группы методов, позволяющих определить демографическую историю: один основан на данных гаплотипов, другой основан на спектрах частот аллелей (SFS).

Один из самых популярных подходов из первой группы был представлен в статье Li и Durbin [Li and Durbin, 2011]. Он основан на попарно последовательной марковской коалесцентной модели (Pairwise Sequentially Markovian Coalescent) или PSMC. Метод PSMC интерполирует

демографическую историю, используя полногеномные данные от одного диплоидного образца. Этот метод способен учитывать пропуски в данных, вызванные повторяющимися последовательностями и структурными вариациями в геноме. PSMC применяет скрытую марковскую модель (HMM), которая последовательно анализирует геном, оценивая время наиболее раннего общего предка (или время коалесценции) между двумя гаплотипами, составляющими диплоидную последовательность. Следовательно, PSMC может быть использован для определения истории изменения эффективного размера популяции для конкретного образца и, соответственно, популяции и вида, из которых этот образец получен [Nadachowska-Brzyska et al., 2016].

Спектр частот аллелей (SFS) является очень полезным способом обобщения демографических геномных данных. SFS представляет собой матрицу размерности n , содержащую информацию о том, сколько из обнаруженных аллелей в выборке размера n найдены у $1/n$, $2/n$, $(n-1)/n$ индивидуумов. Многие часто используемые метрики в популяционной генетике, включая число сегрегирующих сайтов и среднее разнообразие нуклеотидов, сами являются обобщением SFS [Gattepaille et al., 2013].

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Образцы тканей

Для данной работы были отобраны пробы крови семи неродственных особей африканских гепардов. Чтобы понять, насколько схожи между собой генотипы гепардов, были взяты образцы из двух регионов - Намибии ($n = 4$) и Танзании ($n = 3$), на территории которых проживают популяции, принадлежащие к двум подвидам: *Acinonyx jubatus jubatus* и *A. j. raineyii*, соответственно (Табл. 1).

Таблица 1. Список образцов, использованных в работе. Образец AJU981 был использован в качестве референса, остальные были секвенированы с относительно низким покрытием.

Код образца	Место сбора	Субпопуляция	Пол
AJU981	Намибия	Южная	М
AJU884	Намибия	Южная	М
AJU878	Намибия	Южная	Ж
AJU955	Намибия	Южная	М
AJU171	Танзания	Восточная	М
AJU172	Танзания	Восточная	М
AJU173	Танзания	Восточная	Ж

2.2. Выделение ДНК, создание библиотек и секвенирование

Для выделения высокомолекулярной геномной ДНК из образцов крови гепарда использовали фенол-хлороформ с последующим осаждением ацетатом натрия. Секвенировали на платформе Illumina HiSeq2000, так как она обеспечивает хорошую глубину покрытия данными и высокое качество секвенирования.

Используя выделенную геномную ДНК Chewbaаса самца гепарда, проживающего в Намибии, были сконструированы как обычная (pair-end), так

и инвертированная (mate-pair) библиотеки для *de novo* сборки генома вида с высоким уровнем покрытия. Использовались два варианта обычных библиотек с длиной вставки 170 и 500 п.н. Для улучшения распознавания повторов на этапах сборки генома и скаффолдинга сконструировали 5 инвертированных библиотек с длиной вставки от 800 п.н. до 20 т.п.н (Табл. 2).

Таблица 2. Типы библиотек, созданных для генома гепарда, и количество дорожек, которые были просеквенированы. PE – парно-концевые, MP – инвертированные.

Тип библиотеки	Размер вставки (п.н.)	Кол-во дорожек
PE	170	7
PE	500	3
MP	800	4
MP	2000	5
MP	5000	7
MP	10000	4
MP	20000	3

Для анализа генетического разнообразия, изучения демографической истории и вариаций числа копий генов (CNV) мы дополнительно взяли шесть образцов, представляющих два подвида. Для получения геномов с низким уровнем покрытия были сконструированы и секвенированы с 5-6-кратной глубиной покрытия обычные библиотеки с длиной вставки 500 п.н.

2.3. Список биоинформатических инструментов

Основные результаты работы были получены с использованием баз данных GenBank и RefSeq (NCBI) [Wheeler et al., 2007], Ensembl [Yates et al., 2020], GeneCards [Safran et al., 2010], MalaCards [Rappaport et al., 2013], OMIM [Hamosh et al., 2005], KEGG [Kanehisa et al., 2007], Repbase [Jurka et al., 2005], miRbase [Kozomara et al., 2019], Rfam [Griffiths-Jones et al., 2003] и TreeFam [Schreiber et al., 2014].

Сборка генома – SOAPdeovo [Luo et al., 2012] и Chromosomer [Tamazian et al., 2016].

Выравнивания и поиск по базам данных - BLAST+ [Camacho et al., 2009] и Blat [Kent, 2002].

Получение базовых статистик о сборке генома, оценка качества – Quast [Gurevich et al., 2013], CEGMA [Parra et al., 2007], BUSCO [Simão et al., 2015].

Выравнивание прочтений, поиск SNV и аннотация эффектов SNV - bwa [Misale, 2014], bowtie2 [Misale, 2014], samtools [Li et al., 2009], SOAPsnp [Langmead et al., 2009], snpEff [Cingolani et al., 2012], PolyPhen2 [Adzhubei et al., 2013], ANGSD [Korneliussen et al., 2014], Beagle [Altmann et al., 2012].

Анализ к-меров в прочтениях – Jellyfish [Marcais et al., 2012]

Полногеномное выравнивание и синтения - LASTZ [Schwartz et al., 2003], ProgressiveCactus [Armstrong et al., 2020], GRIMM [Tesler, 2002]

Поиск и аннотация генов и повторов - RepeatMasker [Chen, 2004], TRF [Benson, 1999], tRNAscan-SE [Lowe et al., 1997], Infernal [Nawrocki et al., 2009], InterProScan [Mitchell et al., 2015], Augustus [Stanke et al., 2006].

Поиск сигналов отбора и анализ семейств генов - PAML [Yang, 1997], , HMMER3 [Finn et al., 2011], Proteinortho [Lechner et al., 2011], Gblocks, ClustalW [Thompson et al., 2003], PRANK [Löytynoja, 2014]

Изучение демографической истории - PSMC [Li and Durbin, 2011], DaDi [McCoy et al., 2014] и GADMA [Noskova et al., 2020]

Реализацию авторских скриптов проводили в программной средах Python, Perl и R.

2.4. Оценка качества прочтений

Прочтения считались низкокачественными в случае, если показатель качества был ниже 30 по шкале Phred [Ewing and Green, 1998], что говорит о точности прочтения данного основания ниже 99.9%. Избыток значений «N» (неопределенных пар нуклеотидов) в сырых прочтениях также был сигналом низкого качества.

Прочтения подвергали процедуре контроля качества с использованием собственных скриптов, написанных на Perl. Мы удалили все полные или

частичные низкокачественные прочтения, которые соответствовали одному или нескольким из следующих критериев: 1) содержание непрочитанных букв (N) более 10%; 2) более 40% длины прочтения было с качеством ниже Q7; 3) прочтения перекрывались более чем на 10 п.н. с последовательностью адаптера при максимальном редакционном расстоянии не более 2 п.н.; 4) парные прочтения, которые перекрывается более чем на 10 п.н. между двумя концами; 5) дублированные прочтения.

Мы заметили, что оба хвоста прочтения с общей длиной 90 п.н. всегда имели низкие оценки качества, особенно на 3'-конце. Поэтому мы удаляли максимум 10 п.н. от 5'-конца прочтения, если показатель качества был меньше Q20. Аналогично, мы отрезали максимум 40 п.н. от 3' конца прочтения, если показатель качества был меньше Q20. Таким образом, мы сохранили достаточно последовательностей нуклеотидов с высоким качеством для последующего картирования на референсный геном.

2.4. Поиск и аннотация SNV

Мы использовали программу BWA [Misale, 2014] для сопоставления исходных прочтений с геномом собранного референсного гепарда. Выравнивания были дополнительно отфильтрованы для последующего поиска однонуклеотидных замен в соответствии со следующими критериями: 1) были удалены выравнивания с оценкой качества менее 20 ($<MQ20$); 2) были удалены неуникальные выравнивания, то есть любые выравнивания, которые картируются в несколько мест в последовательности генома; 3) дублированные выравнивания, то есть два или более прочтения, координаты начала и конца которых относительно референса совпадают, также были удалены.

Поиск SNV с использованием образцов с низким покрытием ($<10X$) является трудной задачей при применении большинства современных методов. Подход, описанный в Nielson et al. [Nielson et al., 2012], является надежным и высокоточным методом поиска SNV при низком покрытии с

использованием спектра частот аллелей (SFS). Для оценки максимальной вероятности генотипа в каждом локусе в этой модели используется критерий максимального правдоподобия. Мы использовали программу SOAPsnp для создания первичного файла формат GLFv2 для каждого сайта, а затем использовали ANGSD [Korneliussen et al., 2014] и Beagle [Altmann et al., 2012] для определения генотипа (SNV).

Все SNV были аннотированы для каждого образца, используя snpEff [Cingolani et al., 2012] с базой данных, которая была построена из аннотированных генов гепарда.

2.5. Способы оценки генетического разнообразия

Для всесторонней оценки степени генетического разнообразия мы использовали шесть различных способов: 1) Встречаемость однонуклеотидных замен на всю протяженность генома 2) Плотность однонуклеотидных замен в окнах генома протяженностью в 50 т.п.н. 3) Количество вариативных позиций в кодирующих областях 4) Протяженность участков гомозиготности 5) Количество достоверных гетерозигот на геном 6) Генетическое разнообразие с использованием митохондриальных последовательностей.

2.6. Дополнительно использованные геномные данные

В работе использовалось множество генетических данных из различных источников. Наиболее активно применяемым ресурсом была сборка и аннотация домашней кошки [Tamazian et al., 2014], работа над которой была подготовительным этапом текущего проекта.

Сборка генома домашней кошки (Fca-6.2) была многократно улучшена в процессе ряда правок, так что существующая в настоящее время актуальная версия генома базируется на результатах нескольких методов секвенирования следующего поколения. Кроме того, данная версия генома привязана к хромосомным координатам с помощью двух физических карт, карты

радиационных гибридов [Davis et al., 2009] и карты сцепления коротких tandemных повторов (STR) [Menotti-Raymond et al., 2009]. Дополнительно, на основе микрочипа для генотипирования домашних кошек от компании Illumina, содержащего около 60 тысяч однонуклеотидных замен (SNP), было идентифицировано 1.943 новых участка сцепления [Pontius et al., 2007].

Кроме генома домашней кошки мы так же использовали несколько других геномных сборок - Человек (GRCh37), Собака (CanFam3.1) и Амурский тигр (ATCQ01000000).

Мы так же использовали данные секвенирования для следующих видов: 1) Дикий кот – PRJNA253950 2) Амурский тигр, Бенгальский тигр, Африканский лев, Снежный леопард - PRJNA182708 3) Горилла - PRJEB12821.

Данные секвенирования генома азиатского льва, а так же сборка генома и аннотация были предоставлены BGI и доступны по запросу.

2.7. *De novo* сборка генома африканского гепарда

Для *de novo* сборки генома гепарда из коротких ридов мы использовали программу для сборки последовательностей SOAPdenovo [Luo et al., 2012], в основе которой лежит структура данных граф де Брюйна. Все отфильтрованные прочтения были дополнительно откорректированы с использованием частотного анализа k-меров. K-меры длиной 17 п.н. и частотой встречаемости менее 4 раз считались ошибочными. После коррекции прочтения были объединены в длинные непрерывные последовательности - ДНК-контиги. Затем мы использовали информацию из пяти инвертированных библиотек для поэтапного ориентирования и объединения контигов в скаффолды. Скаффолды состоят из контигов и одного или более разрывов известной длины между ними. Мы убрали часть разрывов посредством разрешения tandemных и диспергированных повторов за счет информации из библиотеки парно-концевых прочтений. Далее мы картировали прочтения,

использованные для сборки генома, на сам геном и заполняли ту часть разрывов, для которой было достаточно доказательств разрешения повторов.

2.8. Сборка на основе референса

Среди представителей Felidae наблюдается высокий уровень консерватизма в организации генома [Dobrynin et al., 2015, Cho et al., 2013, Hsu et al., 1963]. Для картирования скаффолдов генома африканского гепарда на референсный геном домашней кошки в первую очередь мы использовали результаты парного полногеномного выравнивания двух видов, полученного в программе BLAST+ [Camacho et al., 2009]. Затем мы присвоили величину сходства для каждого картированного фрагмента. Величина сходства основывается на длине и идентичности выровненных регионов. После этапа оценивания положение каждого скаффолда на хромосомах домашней кошки определяется в соответствии с этапами, представленными на Рис. 1. Во-первых, мы предположили, что каждый скаффолд может иметь уникальное расположение в специфическом регионе на референсном геноме домашней кошки в случае, если отношение между лучшим выравниванием и вторым по степени сходства превышает заданный порог [1.2]. В случае выполнения данного условия скаффолд закрепляется на позиции, соответствующей его выравниванию с наибольшей оценкой.

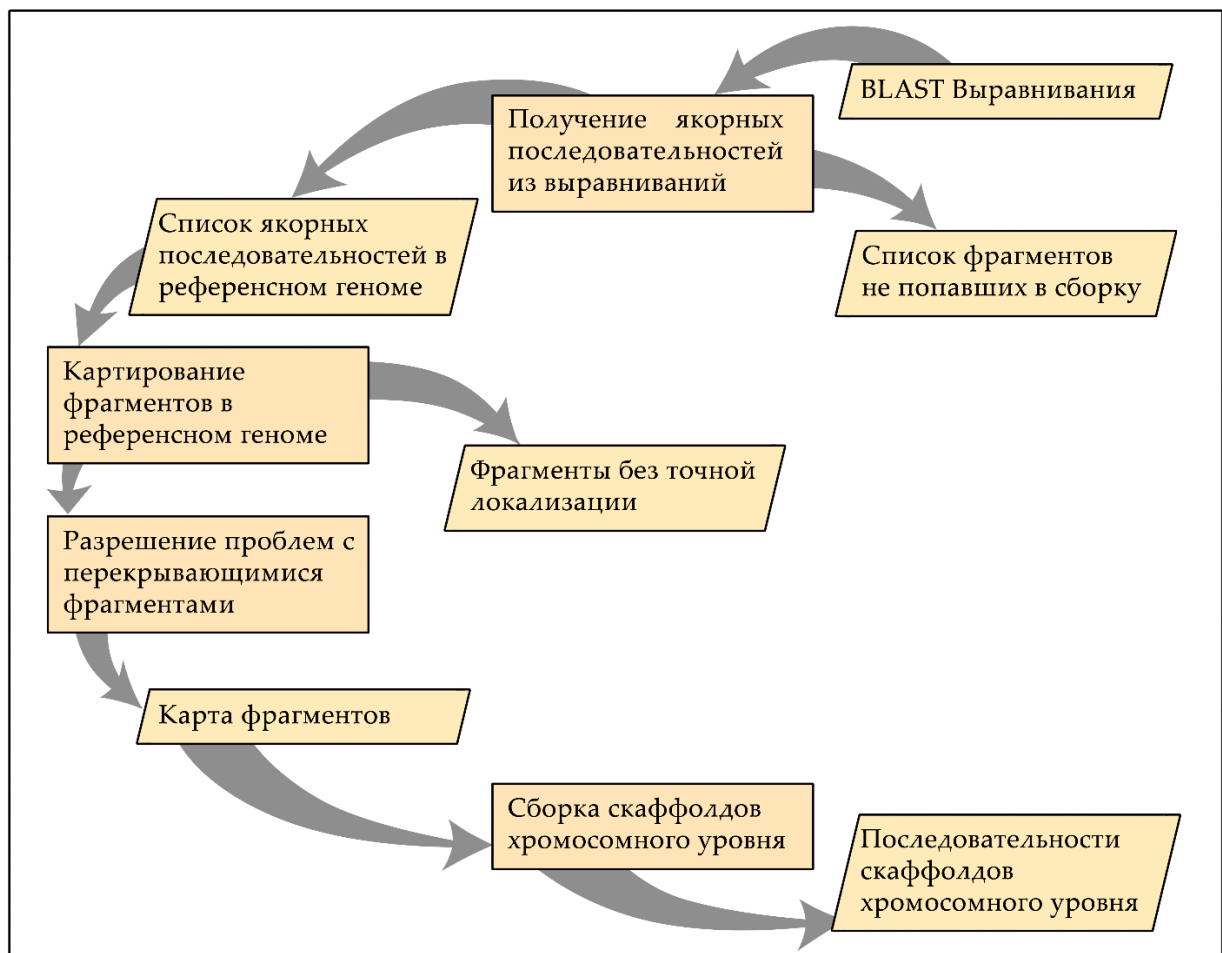


Рисунок 1. Схема работы программы Chromosomer в процессе сборки генома на основе референса. Прямоугольники соотносятся с процедурами, которые применяются к данным, соотносящимся с параллелограммами.

Скаффолды, не имеющие локализации и конкретной позиции, были исключены из дальнейшего анализа. С использованием позиционной карты мы расположили каждый скаффолд, как показано на Рис. 2 (А – Б). В случае наложения скаффолдов мы добавляли пропуски между перекрывающимися скаффолдами Рис. 2. (В – Г)

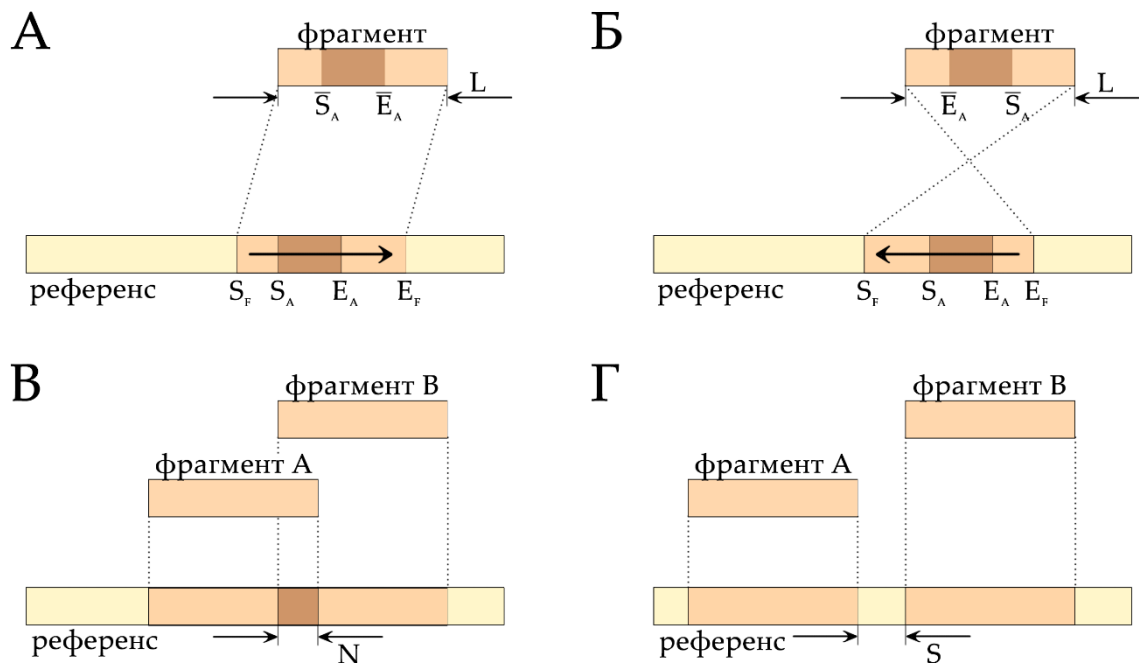


Рисунок 2. Картирование фрагмента на основе его выравнивания и разрешение перекрытий между фрагментами. Выравнивание, которое используется для определения положения, отмечено темно-серым цветом. Схема А и Б демонстрирует случаи прямой и обратной ориентации фрагмента относительно хромосомы. Схема В и Г, если два фрагмента образуют перекрытие на N п.н., то фрагмент сдвигается на $N+S$ п.н. для того, чтобы разрешить перекрытие. S описывает размер пропуска, который будет добавлен, этот параметр задается при запуске программы.

2.9. Оценка качества сборки генома

Оценка качества геномной сборки является серьёзной проблемой. Наиболее распространенный показатели, используемые для сравнения геномных сборок между собой, являются показатели N50 и L50. В полученном наборе контигов, расположенном в порядке убывания длины, значение N50 определяется как длина последовательностей, составляющих 50% генома. Показатель L50 определяется как наименьшее число контигов, сумма длины которых составляет N50 [Gurevich et al., 2013]. Однако значения N50 и L50 и размер собранного генома не всегда коррелируют с качеством последовательностей, количеством собранных генов или ошибочных скаффолдов.

Для дополнительной оценки качества геномной сборки африканского гепарда мы использовали ряд методов, используемых для подсчета вышеперечисленных показателей:

1. Локальную точность и согласованность измеряли посредством картирования коротких прочтений шести дополнительных образцов гепардов на референсный геном гепарда.

2. Существует несколько методов оценки качества отдельной геномной сборки и ассоциированного генного состава [Parra et al., 2007; Simão et al., 2015]. Мы использовали протокол CEGMA [Parra et al., 2007], который представляет собой конвейер для точной аннотации набора хорошо изученных консервативных генов, обнаруженных в геномах эукариот.

2.10. Поиск блоков синтении

Для поиска возможных перестроек, возникших между гепардом и домашней кошкой, амурским тигром, африканским львом и собакой, мы применили алгоритм GRIMM [Tesler, 2002] к результатам выравнивания, полученным при помощи ProgressiveCactus [Armstrong et al., 2020]. Так как только геномы домашней кошки и собаки были собраны до уровня хромосом, в то время как три оставшихся вида были собраны только до уровня скаффолдов, мы должны были рассматривать перестройки только как события, происходящие внутри смежных контигов.

2.11. Идентификация повторов

Комбинация программного обеспечения RepeatMasker [Chen, 2004] и библиотеки Repbase Update [Bao et al., 2015] позволила нам идентифицировать все ранее известные повторы, специфичные для хищников и млекопитающих, в сборке генома африканского гепарда. Мы изучили повторы в следующих геномах: африканский гепард (*Acynonyx jubatus*), африканский лев (*Panthera*

leo), амурский тигр (*Panthera tigris*), домашняя кошка (*Felis catus*) и домашняя собака (*Canis lupus familiaris*).

Анализ tandemных повторов в геномах домашней кошки, гепарда, собаки, льва и тигра основывался на программном обеспечении TRF [Benson, 1999] и на алгоритме для обработки результатов TRF, который был опубликован нами ранее [Tamazian et al., 2014].

Все обнаруженные tandemные повторы были классифицированы как члены одной из трех следующих групп:

1. Микросателлиты с длиной мономера менее 5 п.н., включая совершенные микросателлиты с длиной мономера менее 5 п.н.
2. Сложные tandemные повторы, которые включали крупные tandemные повторы и сателлитную ДНК, характеризующиеся GC% от 20 до 80, длиной массива более 100 п.н., вариациями числа копий в длину более 4 п.н., энтропией массива более 1,76, длиной мономера более 4 п.н. и несовершенной организацией массива повтора.
3. Макросателлиты, характеризующиеся большими, успешно собранными tandemно повторяющимися массивами, которые были разделены на три подгруппы по длинам массива 1, 3 и 10 т.п.н.

Сложные tandemные повторы классифицировали по семействам на основе подобия последовательностей, которое вычислялось с использованием NCBI BLAST+ согласно протоколу [Tamazian et al., 2014].

2.12. Аннотация генов

Для аннотации белок-кодирующих генов в геноме гепарда, мы использовали комбинацию инструментов предсказания *ab initio* с внешними доказательствами из геномов человека, домашней собаки и кошки, загруженных из базы данных Ensembl (<http://www.ensembl.org>). Потенциальные miRNA гены были отфильтрованы против miRBase (<http://www.mirbase.org>) для удаления ложноположительных результатов.

Для аннотации генов некодирующих РНК-элементов мы использовали: tRNAscan-SE (версия 1.23) для поиска последовательностей, кодирующих тРНК. рРНК позвоночных из Европейской базы данных рибосомной РНК использовалась для предсказания последовательностей DBM (Domain Binding Motif). Rfam (версия 9.1) в сочетании с Infernal были использованы для аннотации малых ядерных РНК (snRNAs) и микроРНК (miRNAs).

2.13. Поиск сигналов позитивного отбора

Мы искали гены, которые предположительно эволюционировали при положительном отборе, используя метод максимального правдоподобия для анализа молекулярной эволюции [Yang, 1997; Yang, 2007]. В частности, мы использовали программный пакет PAML для тестирования сигналов позитивного отбора вдоль филогенетической линии гепарда. Мы сравнивали модель A1, в которой сайты могут эволюционировать нейтрально и под очищающим отбором, с моделью A, что позволяет сайтам также быть под положительным отбором. Критерии значимости (p) были рассчитаны с использованием статистики χ^2 , и скорректированы на множественное тестирование поправкой FDR.

2.14. Анализ генных семейств

Анализ экспансии/сокращения генных семейств требует, чтобы гены были сгруппированы в семейства. Для ряда секвенированных геномов семейства генов могут быть найдены в публичных базах данных, таких как TreeFam [Li et al., 2006] или Ensembl Compara [Cunningham et al., 2015]. Для *de novo* сборки генома гепарда мы должны были сами выполнять кластеризацию найденных генов. В настоящее время существует два широко используемых способа формирования семейств генов. Оба варианта используют только по одному, самому длинному белку, из каждого известного гена. Первый подход предполагает поиск генных семейств в публичных базах данных. А второй подход, возникший позже, основан на выравнивании белков с набором HMM

профилей уже аннотированных генных семейств для того, чтобы найти к какому из известных семейств может принадлежать конкретный ген. Выравнивания выполняются независимо для каждого вида при помощи программы HMMER3 [Finn et al., 2011]. Этот второй метод намного быстрее первого и позволяет использовать информацию, содержащуюся в базах данных.

2.15. Анализ генов связанных с размножением

Чтобы проанализировать белок-кодирующие гены в геноме гепарда, участвующие в размножении, мы использовали все человеческие гены, принадлежащие к термину генной онтологии GO: 0000003 (Reproduction) из базы данных Ensembl Genes [Cunningham et al., 2015]. Мы получили в общей сложности 1730 транскриптов из 964 белок-кодирующих генов. Этот набор был использован для поиска однокопийных ортологичных генов у гепарда, кошки, тигра, собаки и человека. Чтобы выявить ортологичные группы генов, был использован метод, реализованный в программе Proteinortho [Lechner et al., 2011], который основан как на результатах выравнивания BLAST, так и на анализе синтенных блоков. Из 1730 транскриптов поиск привел к 656 однокопийным ортологам, присутствующих у всех пяти видов.

Мы использовали PAML, чтобы найти гены с ускоренным накоплением несинонимичных замен (kA/kS) в филогенетической линии гепарда относительно средней скорости у остальных четырех видов. Ускоренное накопление несинонимичных замен может указывать на увеличение числа умеренно вредных и вредных мутаций, которые могут оказать влияние на репродуктивную физиологию у гепардов. Для оценки скорости накопления несинонимичных замен использовалась модель свободного коэффициента, реализованная в PAML. Модель допускает различную скорость накопления замен для каждой ветви дерева. Все гены были объединены в один «суперген», а kA/kS оценивали для каждого вида.

2.16. Анализ состава генов и генетического разнообразия для района МНС

Область МНС также уникальна по составу генов. Относительно большое количество как одиночных, так и многокопийных генов создает условия, при которых инструменты предсказания генов ведут себя очень неточно. Прежде чем аннотировать гены, обнаруженные в области МНС, сначала было проведено маскирование последовательностей, с помощью программы RepeatMasker и базы данных RepBase3.0 [Bao et al., 2015]. Первичная аннотации генов в области МНС была сделана с использованием программ Augustus [Stanke et al., 2006], Blat [Kent, 2002], Blast [Camacho et al., 2009] и пользовательских скриптов на Python. Для поиска гомологичных последовательностей мы использовали всю базу данных RefSeq [Pruitt et al., 2007]. Гены, относящиеся к МНС классов I и II, дополнительно проверялись и, при необходимости, корректировались вручную. Кроме того, мы проанализировали консервативные домены в генах МНС классов I и II с использованием базы данных NCBI [Pruitt et al., 2007] и CLUSTALW [Thompson et al., 2003].

Прочтения от образцов с низким покрытием были картированы на оба варианта сборки, как на *de novo* сборку до уровня скаффолдов, так и на консенсус, полученный путем картирования прочтений на референсную сборку домашней кошки (FCA - GCA_000181335.4) [Tamazian et al., 2014]. Обнаруженные SNV фильтровали по глубине покрытия (≥ 2), качеству (≥ 20) и местоположению.

2.17. Реконструкция демографической истории

Мы применили комбинацию из трех методов для изучения демографической истории.

Мы использовали метод PSMC, чтобы оценить историю изменения эффективной численности популяции, используя геном гепарда с высоким покрытием (AJU 878, Chewbaaca). Мы использовали программы BWA и

samtools для картирования и генотипирования коротких прочтений. Использовались следующие значения параметров: интервал между поколениями равный трем годам, скорость мутаций в районе $0,3 \times 10^{-8}$. Оценка скорости мутаций основана на полногеномном выравнивании между гепардом и домашней кошкой, которое было сделано с помощью LASTZ [Schwartz et al., 2003], затем количество различий между двумя видами усредняли на время с момента их дивергенции (7 млн лет).

Два метода DaDi и GADMA работают со спектрами частот аллелей. Для двух проанализированных популяций африканского гепарда (Намибия или Южная Африка и Танзания или Восточная Африка) SFS представляет собой многомерную матрицу X , где каждая запись x_{ij} содержит число SNV с наблюдаемым числом аллелей i в популяции 1 (Намибия), и j в популяции 2 (Танзания). Для каждой эволюционной модели рассчитывается функция максимального правдоподобия с учетом ожидаемого и наблюдаемого сайт частотного спектра (SFS). Каждая ячейка в матрице, ожидаемой SFS, отражает вероятность наблюдения SNV в этой ячейке. Метод предполагает, что все SNV нейтральны, и между ними возможна свободная рекомбинация. Тогда ожидаемые вероятности могут быть вычислены из распределения частот аллелей каждой популяции, которые, в свою очередь, могут быть найдены с помощью диффузной аппроксимации эволюционных процессов, таких как сроки и размер демографических изменений.

Чтобы сделать вывод о демографической истории двух популяций гепарда (восточной и южной), мы использовали инструменты GADMA и DaDi [Noskova et al., 2020; McCoy et al., 2014]. Итак, DaDi может генерировать SFS в соответствии с одним или несколькими демографическими сценариями. Задача состоит в том, чтобы максимизировать сходство между предполагаемой моделью и наблюдаемыми частотами аллелей в SFS. Степень сходства оценивали путем вычисления функции правдоподобия среди разных демографических сценариев.

Чтобы определить, какая из моделей была наиболее близка к наблюдаемой SFS гепарда, мы тестировали модели в пяти разных демографических сценариях, используя SFS двух популяций гепарда, которые включали в себя информацию о предковых аллелях. Чтобы оценить время и взаиморасположение между событиями разделения предковой популяции на субпопуляции и эффектом бутылочного горлышка, мы протестировали четыре двумерные (2D) модели:

1) Предковая популяция распадается на две субпопуляции, далее следует ограниченная миграция из одной субпопуляции в другую (модель 2D IM).

2) Предковая популяция сначала переживает эффект бутылочного горлышка, а затем разделяется на две субпопуляции (модель 2D BIM).

3) Предковая популяция сначала распадается на две субпопуляции, далее следует эффект бутылочного горлышка, а затем происходит восстановление численности в каждой субпопуляции (модель 2D SBR).

4) Предковая популяция, начиная с ~ 100.000 лет до разделения на две субпопуляции растет в размере, после разделения следует эффект бутылочного горлышка, который две субпопуляции переживают независимо (модель 2D ISB).

ГЛАВА 3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ

3.1. Оценка данных секвенирования и сборка генома африканского гепарда, *Acinonyx jubatus*

3.1.1. Анализ данных секвенирования африканского гепарда для референса образца и шести дополнительных особей.

Для получения высокого покрытия генома каждая библиотека была секвенирована на нескольких дорожках (Табл. 2), что обеспечило более 300 секвенированных млрд. п.н. (Табл. 3) первичных данных до какой-либо фильтрации и анализа. Для получения данных высокого качества из объема первичных прочтений мы применяли ряд фильтров для отсева дубликатов ПЦР, химерных последовательностей и прочтений с высоким уровнем ошибок.

Таблица 3. Прочтения генома гепарда, использованные для *de novo* сборки. Глубина секвенирования – суммарный объём всех прочтений высокого качества, поделенный на размер генома.

Длина вставки (п.н.)	Длина прочтения (п.н.)	Первичные данные (млрд. п.н.)	Данные высокого качества (млрд п.н.)	Глубина секвенирования (×)
170	100	102,20	67,97	28,56
500	100	50,48	38,70	16,26
800	100	38,49	26,40	11,09
2000	49	33,89	18,53	7,79
5000	49	45,71	10,44	4,39
10000	49	27,27	6,51	2,73
20000	49	24,26	9,76	4,10

Библиотеки для шести дополнительных образцов были секвенированы совместно на одном чипе секвенатора Illumina HiSeq2000, в результате чего

получены первичные данные с умеренным выходом и 5-6-кратным уровнем покрытия генома (Табл. 4).

Таблица 4. Прочтения после ресеквенирования шести дополнительных образцов для популяционного анализа. Глубина секвенирования – суммарный объем всех прочтений высокого качества, поделенный на размер генома.

Размер вставк и (п.н.)	Длина прочтения (п.н.)	Первичные данные (млн. п.н.)	Данные с высоким качеством (млн. п.н.)	Глубина секвенирования (×)	Доля выровненных прочтений (%)
500	100	13968,82	13060,75	5,49	97,76
500	100	17787,33	16435,72	6,91	97,57
500	100	19420,52	17799,68	7,48	97,57
500	100	16654,33	15391,33	6,47	96,98
500	100	15598,71	14520,59	6,10	97,32
500	100	18892,26	17252,65	7,25	97,14

Точность расшифровки коротких прочтений у Illumina считается достаточно высокой для технологии секвенирования второго поколения и уступает только секвенированию методом Сэнгера. Но часть сырых прочтений содержала артефакты, связанные с загрязнением адаптерами, или имела низкие оценки качества прочтений. Загрязнение сырых прочтений адаптерами является общей проблемой и существует в некоторой степени почти во всех наборах данных, полученных технологиями высокопроизводительного секвенирования. Загрязнение данных адаптерами происходит в процессе подготовки библиотеки, где на этапах клонирования и амплификации необходимо добавить искусственно созданные последовательности нуклеотидов, такие как адаптеры или линкеры. Такое загрязнение, оставленное без внимания, может привести к появлению артефактов при сборке генома. Лучшим способом идентификации такого рода загрязнений

является сравнение всех полученных прочтений с базой данных последовательностей адаптеров и последующая обрезка или удаление всех совпавших последовательностей. Учитывая степень влияния данных факторов, мы отфильтровали и обрезаем сырые прочтения для получения чистого набора данных, используемого в дальнейшей работе.

3.1.2. *De novo* сборка генома африканского гепарда

Для образца (Chewbacca) с высоким покрытием при сборке генома *de novo* была использована только часть полученных сырых прочтений. Поскольку нашей главной задачей было получение высококачественной геномной сборки гепарда *de novo*, мы применили более строгие стандарты качества в нашем исследовании по сравнению с теми, которые обычно используют в экспериментах. Применяя строгие протоколы фильтрации прочтений, мы значительно снизили вероятность обнаружения артефактов в нашей сборке и обеспечили получение высококачественных контигов и скаффолдов.

Откорректированные прочтения были собраны в 97981 контигов (N50 = 28,2 т.п.н.) и 2332 скаффолда (N50 = 3,1 млн. п.н.), причем длина самого протяженного собранного скаффолда составила около 13 млн.п.н. Общий размер собранного генома составил 2,376 млрд.п.н. (Табл. 5) с менее чем 2% неизвестных пар оснований. Среднее покрытие генома, рассчитанное по картированным прочтениям, составило 73× (Рис. 3) для референса и 5-6 для остальных образцов (Рис. 4).

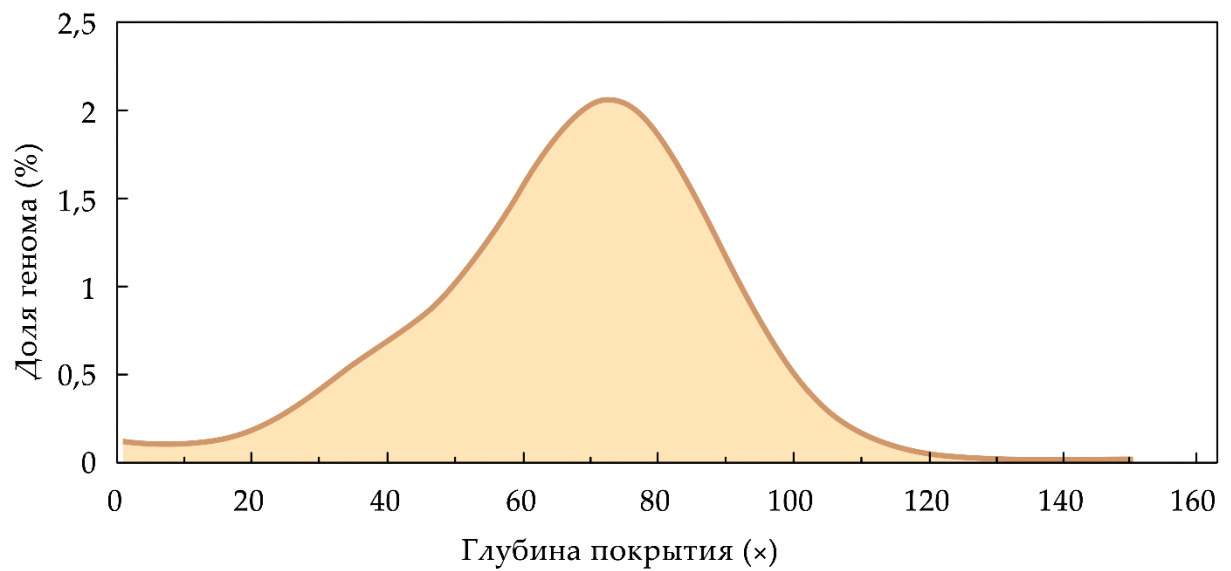


Рисунок 3. Распределение глубины секвенирования в геноме гепарда. Средняя глубина секвенирования 73×.

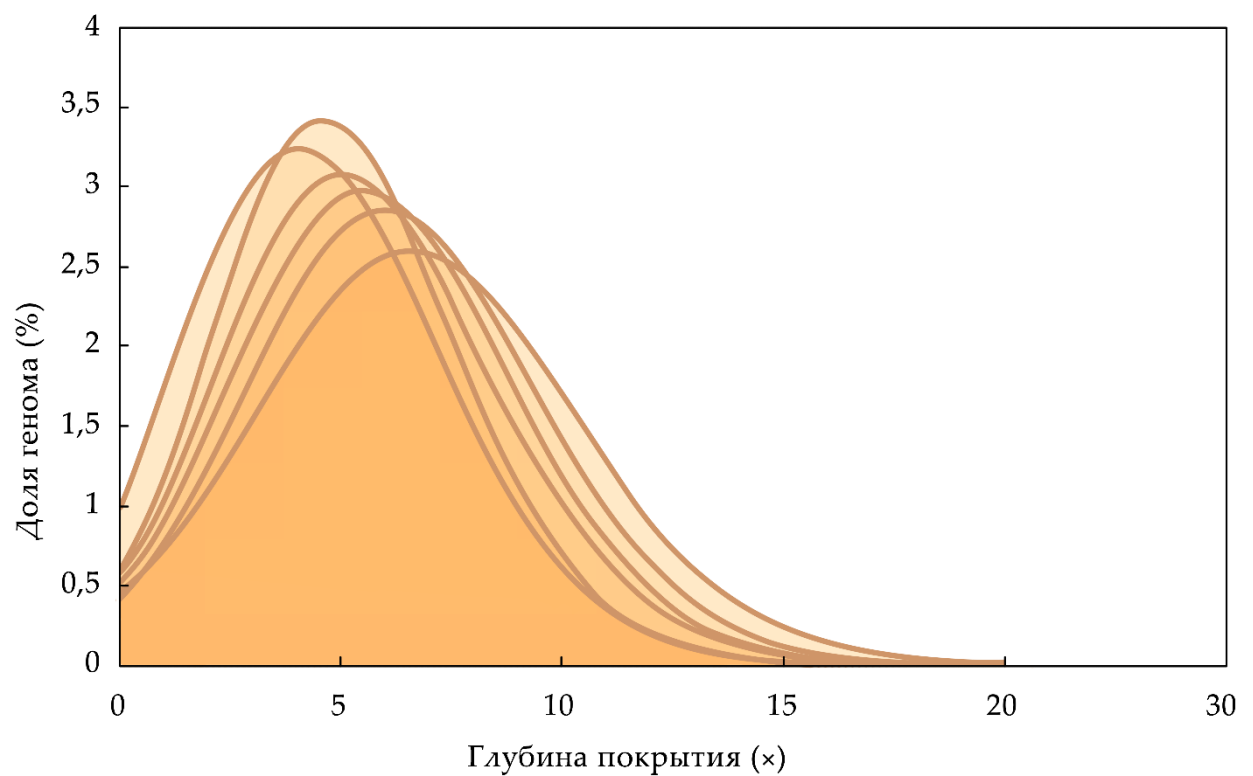


Рисунок 4. Распределение глубины покрытия секвенирования для дополнительных шести образцов. График основан на результатах картирования коротких прочтений на *de novo* собранные скаффолды генома гепарда. Средняя глубина секвенирования для образцов с низким покрытием составляет 5-6×.

Таблица 5. Информация о собранном геноме. Общий размер генома, исходя из результатов сборки, оценивается в 2,38 млрд п.н.

	Контиги		Скаффолды	
	Размер (п.н.)	Количество	Размер (п.н.)	Количество
N90	6675	87953	817533	784
N80	11732	61999	1417882	568
N70	16749	45401	1986090	427
N60	22136	33294	2460097	320
N50	28224	23952	3121442	233
Самый протяженный фрагмент	304265		13046067	
Итоговый размер (п.н.)	2333696654		2375786546	

Картирование прочтений на референсный геном гепарда показало, что около 97% всех прочтений имели уникальное положение. Значительная пропорция однозначно картированных прочтений является косвенным признаком хорошей геномной сборки (Табл. 4).

3.1.3. Сборка генома африканского гепарда на основе референса домашней кошки

Учитывая очень консервативную организацию генома у представителей *Felidae* [O'Brien and Johnson, 2005], мы решили использовать геном домашней кошки в качестве референса для сборки хромосом гепарда. Используя такой полный и хорошо описанный источник информации в проекте по сборке генома африканского гепарда, мы создали открытое кроссплатформенное программное обеспечение – Chromosomer [Tamazian et al., 2016]. Данная программа создает первичную сборку скаффолдов хромосомного уровня на

основе карты выравненных фрагментов (контигов или скаффолдов), которые должны быть как в целевом, так и в референсном геноме, тем самым позволяя провести более детальный анализ и более подробную аннотацию сборки изучаемого вида.

Применяя разработанный нами подход с использованием референсного генома, мы получили сборку скаффолдов хромосомного уровня для африканского гепарда (Табл. 6). Мы успешно идентифицировали и картировали 1519 фрагментов, входящих в состав 18 аутосом и X хромосомы. В процессе сборки с использованием референсного генома мы внесли пропуски общей длины 139 млн. п.н. Размер первичной хромосомной сборки составил 2,48 млрд. п.н. Общая длина всех нелокализованных скаффолдов составила 34 млн. п.н., что составляет менее 1,5% длины всех собранных скаффолдов.

Таблица 6. Статистика по собранным скаффолдам хромосомного уровня африканского гепарда. Интрапропуски – пропуски в скаффолдах. Интерпропуски – пропуски, добавленные для разрешения наложений скаффолдов.

Название хромосомы	Интрапропуски (млн.п.н.)	Интерпропуски (млн.п.н.)	Кол-во скаффолдов	Общий размер (млн.п.н.)
A1	2,6	21,8	95	244,6
A2	3,1	3,4	94	169,5
A3	2,2	6,5	72	146,6
B1	2,4	15,8	111	208,9
B2	2,1	5,8	94	158,1
B3	2.4	5,0	82	153,6
B4	2.1	8,8	58	145,7
C1	3.1	12,6	114	225,9
C2	1.8	4,9	79	161,0
D1	2.2	7,4	80	119,0

D2	1.2	4,6	42	90,7
D3	1.4	3,5	68	98,2
D4	1.8	4,1	61	98,5
E1	1.7	8,0	65	64,1
E2	1.5	7,0	60	65,9
E3	1.3	3,9	21	43,6
F1	1.5	2,9	44	72,4
F2	1.2	2,3	46	84,1
X	4.6	10,7	233	131,4
Всего:	40.3	138,9	1519	2,481
Общий размер без интерпропусков:				2,342
Общий размер всех фрагментов:				2,475
Общий размер фрагментов, не вошедших в сборку:				32,9

Геном африканского гепарда представляет собой первый референсный геном представителя из филогенетической линии пумы, одной из восьми основных клад, представляющих дошедших до нас кошачьих [Johnson et al., 2006]. Референсный геном самца гепарда основывается на коротких прочтениях из семи типов геномных библиотек, в результате чего было получено 73-кратное покрытие генома (Рис. 3).

3.1.4. Оценка качества геномной сборки

В случае отсутствия физической карты сцепления или референсного генома достаточно сложно определить точность сборки скаффолдов. Но нам важно проанализировать фактическую долю генома, которую мы смогли собрать. Для предсказания размера генома по данным первичных прочтений мы анализировали распределение 17-меров. Наши результаты показали, что размер генома африканского гепарда составляет около 2,395 млрд.п.н. (Рис. 5),

что близко к размерам геномов ранее секвенированных представителей семейства кошачьих – домашней кошки (2,641 млрд.п.н.), амурского тигра (2,400 млрд.п.н.) и амурского леопарда (2,450 млрд.п.н.) [Pontius et al., 2007; Tamazian et al., 2014; Cho et al., 2013]. Предсказанный нами размер генома африканского гепарда может быть немного больше, чем фактический геном, из-за ошибок секвенирования, в результате которых могут появляться более редкие k-меры. Общая длина всех собранных скаффолдов составила 2,375 млрд п.н., что на 20 млн. п.н. меньше по сравнению с предсказанием, составленным по первичным прочтениям. Данное несоответствие может быть вызвано не только за счет переоценки размера предсказанного генома, но и ввиду недостаточной представленности некоторых центромерных районов. В связи с размером и сложностью нуклеотидной последовательности оказывается почти невозможным собрать перицентромерные районы генома, используя технологии, основанные на коротких прочтениях, например, Illumina [Illumina, 2014]. Кроме того, размер собранных нами хромосом гепарда (Табл. 6) оказался гораздо больше, чем было предсказано, так как мы добавили несколько пропусков в полученную геномную сборку для устранения перекрытий скаффолдов. Для определения расположения скаффолдов половых хромосом мы сравнили степень покрытия и GC состав всех собранных фрагментов (Рис. 6). Нами было замечено, что существует два распределения, одно с 20-30-кратным и второе с 40-60-кратным покрытием, со снижением степени покрытия в регионах с высоким GC составом. Снижение глубины секвенирования в GC насыщенных регионах является общей проблемой технологии секвенирования компании Illumina. Мы предположили, что распределение с 20-30-кратной глубиной охвата соответствует X и Y хромосомам, не считая псевдоаутосомной области.

В связи с высоким содержанием повторов и небольшими размерами Y хромосомы, мы не смогли собрать её с помощью программного обеспечения Chromosomer. Для поиска скаффолдов, которые могут быть ассоциированы с Y хромосомой гепарда, мы сравнили гены, расположенные в Y хромосоме

человека, с некартированными скаффолдами гепарда. Из 54 экспрессирующих белковые продукты генов Y хромосомы человека была предсказана локализация 21 гена в скаффолдах гепарда. Таким образом, нами были найдены пять скаффолдов Y хромосомы гепарда общей длиной 1.5 млн. п.н.

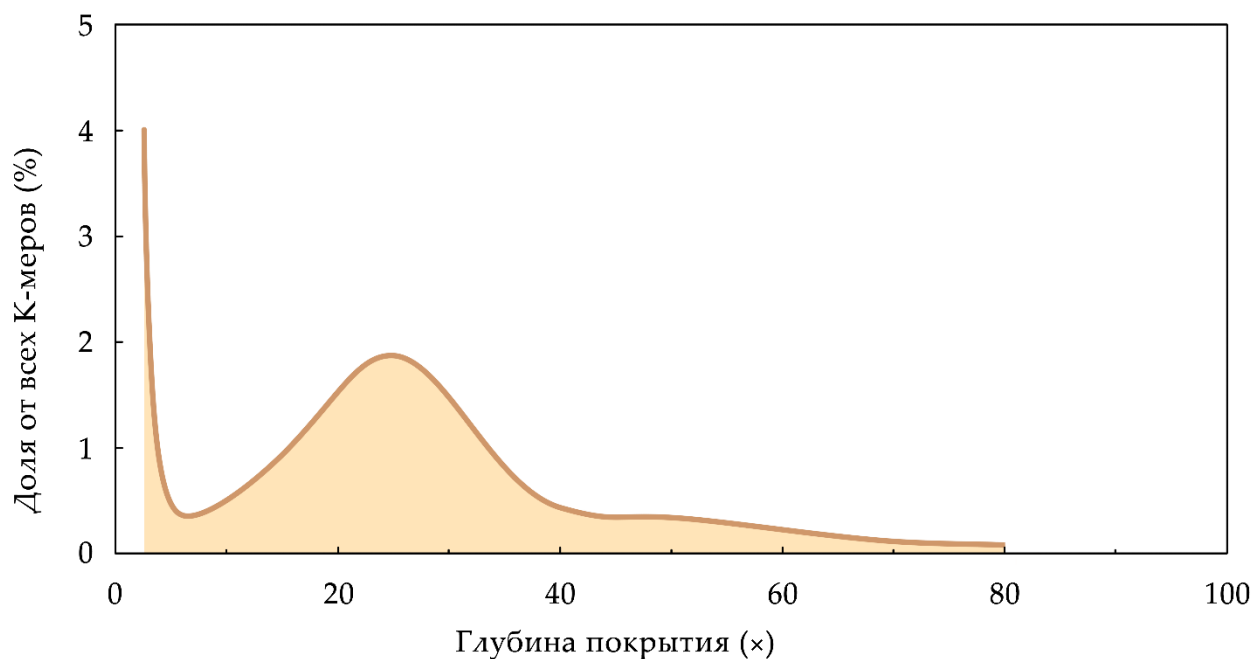


Рисунок 5. Оценка размера генома на основе распределения 17-меров. По оси X откладывается глубина секвенирования, а по оси Y – пропорция, которая является отношением частоты покрытия данного к-мера к общей частоте покрытия всех к-меров.

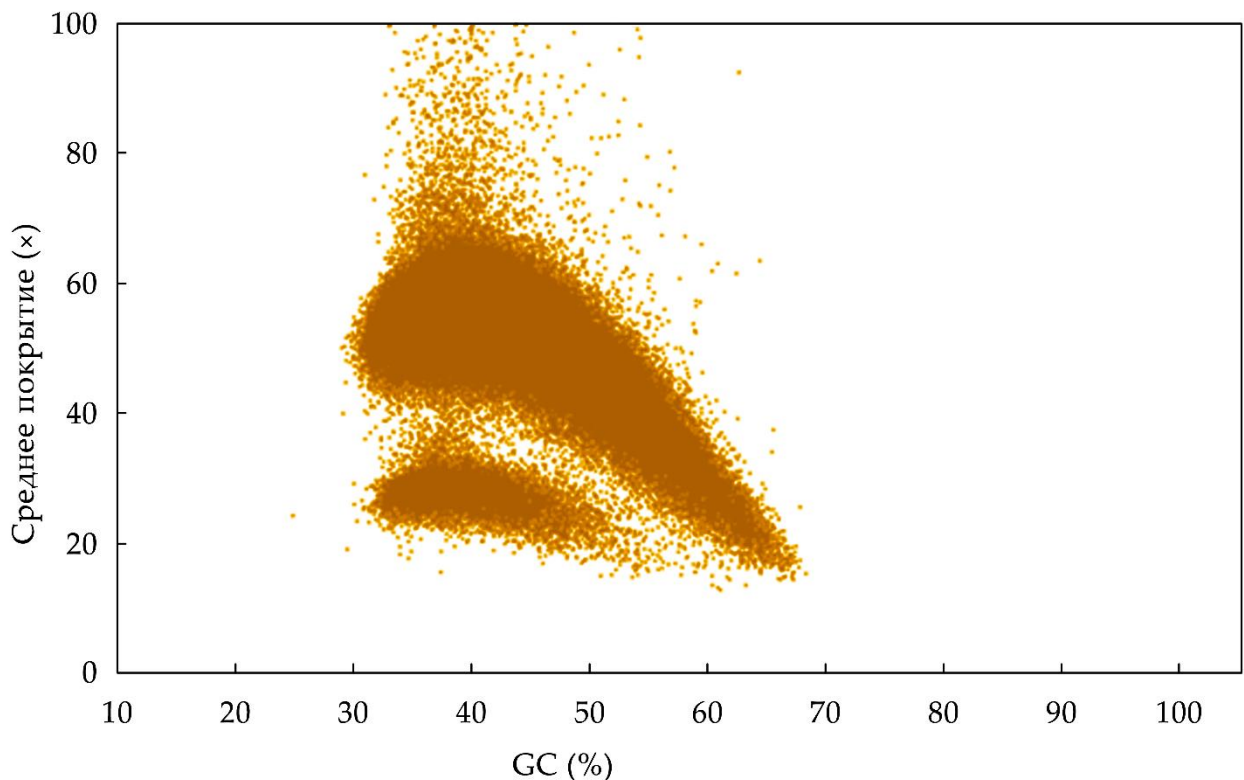


Рисунок 6. GC состав и глубина секвенирования. Основное распределение GC состава расположено в области 35%~40% (облако сверху). Мы также можем наблюдать меньшее распределение (облако снизу) в половину покрытия основного, которое вероятно представляет половые хромосомы (X and Y).

Более 91% генов, консервативных для эукариот, присутствовали в нашей сборке генома, что говорит о высоком уровне сходства качества нашей сборки с геномами других представителей отряда хищных: собаки - 93%, домашней кошки - 95%, тигра - 93%, льва - 89%, процент обнаруженных генов предполагает аналогичную точность сборки геномов данных видов.

Мы произвели выравнивание геномов африканского гепарда, домашней кошки, тигра, льва и домашней собаки для подсчета количества и размера блоков синтении. Мы ожидали, что количество и размер блоков синтении будет больше между близкими видами, чем между далекими. Кроме того, мы искали крупные геномные перестройки в геномах домашней кошки и африканского гепарда, поскольку наличие большого количества таких перестроек будет говорить о возможных ошибках, полученных в процессе сборки скаффолдов. Как было показано ранее, алгоритм Progressive Cactus

[Paten et al., 2011] обеспечивает высокоточные множественные геномные выравнивания, особенно для позвоночных животных. Процент совпадения генома африканского гепарда по сравнению с геномами других животных составил: 93.6 % для тигра (*Panthera tigris*), 91.6 % для льва (*Panthera lion*), 91.1 % для домашней кошки (*Felis catus*) и 74.1 % для собаки (*Canis familiaris*). Как и предполагалось, наибольшая разница была найдена между наиболее отдаленными видами (гепардом и собакой). Процент совпадения отражает долю генома гепарда которая вошла в синтенные блоки в процессе анализа.

Блоки синтении между геномами домашней кошки и гепарда покрывают большую часть генома гепарда (98,6%), вероятно из-за того, что сборка генома домашней кошки гораздо полнее, чем сборки льва и тигра. Данные результаты также согласуются со сравнительно коротким эволюционным расстоянием между кошкой и гепардом, которое оценивается в 6.7 млн лет [Johnson et al., 2006]. Для сравнения, блоки синтении в выравнивании человека и мыши покрывают только 82% человеческого генома [Alkan et al., 2009], что соответствует времени расхождения человека и мыши в 96 млн. лет.

Мы также проанализировали распределение длин блоков синтении для блоков с длиной более 10 т. п.н. (Рис. 7). Пики на графиках соответствуют количеству блоков синтении соответствующей длины. Данные графики показывают, что короткие блоки синтении встречаются чаще, чем длинные. Мы обнаружили, что геном льва наиболее фрагментирован, что объясняет, почему большинство блоков синтении при сравнении гепарда и льва имеют длину менее 1.5 млн п.н. Графики для кошки и собаки схожи, а их блоки синтении длиннее по сравнению с львом и тигром в связи с высоким качеством сборки первых двух видов. Также мы нашли десять крупных геномных перестроек при сравнении геномов домашней кошки и гепарда (Рис. 8), что подтверждает идею о консервативности организации генома кошачьих.

Принимая во внимание полученные результаты, мы уверены в высоком качестве *de novo* геномной сборки африканского гепарда.

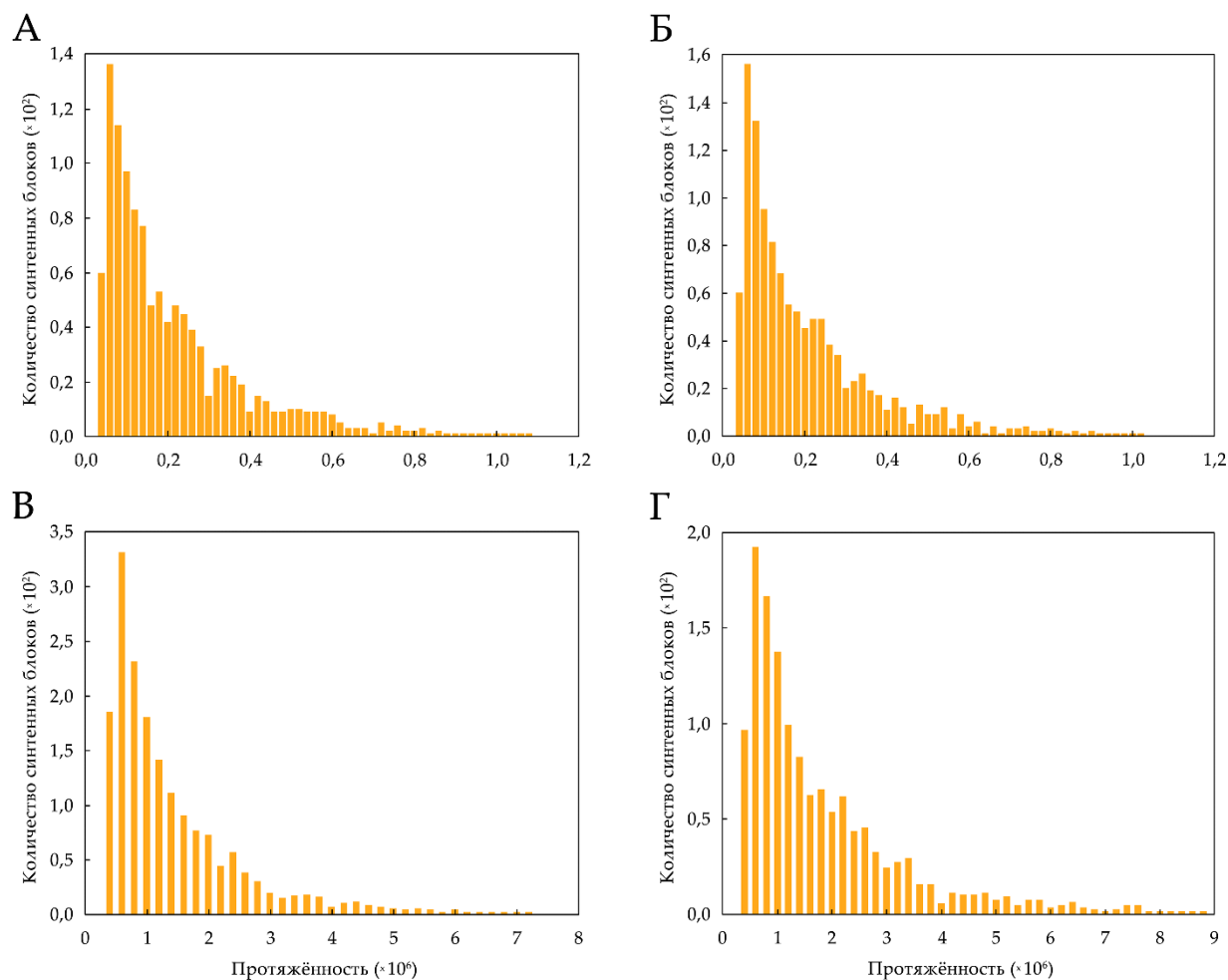


Рисунок 7. Распределение количества синтаксических блоков в неперекрывающихся окнах с размером в 10 т.п.н. в геноме гепарда. Были рассмотрены синтаксические блоки относительно четырех видов хищников: домашней кошки, гепарда, домашней собаки и льва. На оси X отложено количество окон в геноме, размер которых указан на оси Y. (А) Африканский гепард – домашняя кошка (Б) Африканский гепард – домашняя собака (В) Африканский гепард – африканский лев (Г) Африканский гепард – амурский тигр.

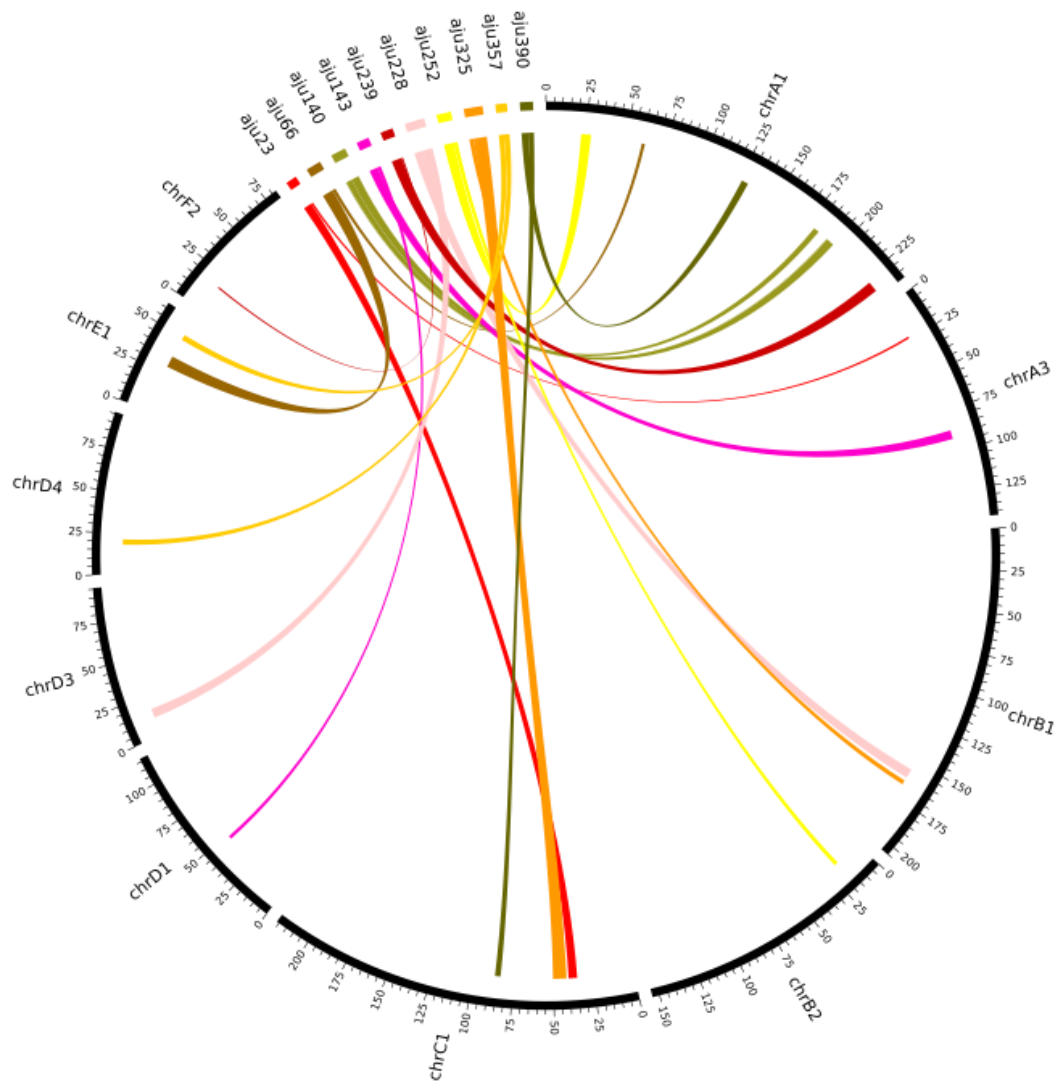


Рисунок 8. Десять самых крупных геномных перестроек между домашней кошкой и африканским гепардом. Перестройки показаны между хромосомами домашней кошки (черный) и скаффолдами африканского гепарда. Данные перестройки могут быть результатами ошибок в процессе сборки генома и должны быть дополнительно проверены методами ПЦР и/или FISH.

3.2. Анализ генов и повторов в геноме африканского гепарда, *Acinonyx jubatus*

3.2.1. Идентификация повторов и генов в геноме африканского гепарда

Результаты анализа повторов с использованием RepeatMasker и TRF приведены в Табл. 7 и 8 соответственно.

Таблица 7. Содержание повторов в геномах, исследованных пяти хищников. Все пять геномов имеют сходные значения как доли замаскированного генома, так и доли отдельных семейств повторов.

	Гепард	Лев	Кошка	Тигр	Собака
Доля повторов	39,48%	38,94%	39,17%	39,15%	39,97%
SINE	10,80%	10,82%	10,98%	10,84%	10,37%
LINE	20,28%	19,85%	20,02%	19,95%	20,18%
LTR транспозоны	5,27%	5,19%	5,08%	5,23%	4,94%
ДНК транспозоны	2,94%	2,89%	2,82%	2,93%	2,79%
Не классифицированные	0,05%	0,05%	0,05%	0,05%	0,05%
pPНК	8,03%	8,1%	8,32%	8,09%	7,81%

Таблица 8. Количество локусов с тандемными повторами (ТП) в геномах пяти хищников. Геномы хищников обладают похожим содержанием тандемных повторов, за исключением генома домашней собаки.

	Гепард	Лев	Тигр	Кошка	Собака
Все ТП	851094	955312	891574	882953	1085732
Микросателлиты	329761	366015	339997	361992	309032
Совершенные микросателлиты	132548	150769	135625	152612	135903
Сложные ТП	3457	3555	3270	4038	7492
Большие ТП >1 т.п.н.	1284	2113	1727	2077	4733
Большие ТП >3 т.п.н.	165	263	196	245	2931
Большие ТП >10 т.п.н.	11	8	6	7	715

Сложные tandemные повторы классифицировали по семействам согласно протоколу [Tamazian et al., 2014]. Каждое семейство было названо согласно номенклатуре, основанной на самой частой длине мономера. Семейство Ajub483A является наиболее сходным с повтором домашней кошки (FA-SAT), и оно предположительно расположено в перицентромерных и прителомерных областях. Семейства Ajub33A и Ajub113A предположительно расположены в перицентромерных областях. Семейство Ajub84A основано на tandemно повторяющемся мотиве, гомологичном последовательностям цинкового пальца (Табл. 9).

Геном собаки содержит примерно на 20% больше подтвержденных tandemных повторов и значительно больше собранных крупных повторов по сравнению с четырьмя опубликованными геномами кошачьих.

Таблица 9. Семейства сложных tandemных повторов. Основные семейства сложных tandemных повторов, идентифицированных в геноме африканского гепарда.

Семейство	Кол-во массивов в геноме	Размер мономера	GC%	Предсказанные черты и локализация
Ajub33A	177	33 п.н.	51%	перицентромерная
Ajub483A	65	483 п.н.	50,5%	прителомерная
Ajub25A	50	25 п.н.	65%	-
Ajub113A	47	113 п.н.	39%	перицентромерная
Ajub84A	277	84 п.н.	40%	Домены типа цинковых пальцев

Всего в геноме гепарда было идентифицировано 20343 белок-кодирующих генов и 110431 генов некодирующих РНК-элементов (Табл. 10 и 11).

Таблица 10. Статистика по аннотации белок-кодирующих генов. Всего было обнаружено более 20.000 белок-кодирующих генов.

	Количество	Размер (млн. п.н.)	Доля в геноме (%)
Гены	20343	601,20	25,30
CDS	174408	29,00	1,22

Таблица 11. Статистика по аннотации некодирующих РНК-элементов.

Тип	Число копий	Средний размер (п.н.)	Суммарный размер (т.п.н.)	Доля в геноме (%)
miRNA	43878	102	4,46	0,19
tRNA	62321	84	5,23	0,22
rRNA	530	81	0,0043	0,000017
snRNA	3702	134	0,042	0,00017

3.2.2. Поиск сигналов положительного отбора

Определенные гены, которые связаны с энергетическим метаболизмом, показали сигналы избирательного роста отношения kA/kS и являются кандидатами на объяснение адаптации гепарда к высокоскоростному преследованию. Мы идентифицировали 947 генов с сигналами положительного отбора ($p < 0,05$ с учетом множественного тестирования). Более подробное описание методологии и используемых инструментов описано в нашей статье [Dobrynin et al., 2015].

Анализ обогащения специфичными GO (максимальное значение p было 10^{-3}) метками показал, что среди 947 генов с сигналами положительного отбора было семь генов, связанных с регуляцией сокращения мышц (*ADORA1*, *RGS2*, *SCN5A*, *ADRA1B*, *CACNA1C*, *TAOK2* и *SCAI*), которые, вероятно, играют важную роль в передвижении гепарда и сокращении сердечной мышцы, два

гена (*TAOK2* и *ADORA1*) связаны с активностью *MAPKK* (Mitogen-activated protein kinase kinase), которая важна в ответе на стрессовые воздействия, включая тепловой стресс.

Мы искали следы недавнего естественного отбора по всем генам гепарда, оценивая соотношения kA/kS в соответствии с ортологами из геномов льва, тигра, кошки, человека и мыши. В частности, мы использовали PAML для тестирования сигналов позитивного отбора вдоль филогенетической линии гепарда и обнаружили 946 генов со значимыми сигналами ($p < 0,05$, скорректированных для множественного тестирования), десять из которых показали обогащение в конкретных GO-терминах. Пять генов со следами отбора были связаны с регуляцией сокращения сердечной и поперечнополосатой мускулатуры (*ADORA1*, *RGS2*, *SCN5A*, *ADRA1* и *CACNA1C*); два гена (*TAOK2* и *ADORA1*) были связаны с активностью *MAPKK*, важной для ответа на стрессовые воздействия, включая тепловой стресс, и четыре гена (*APOC3*, *DDIT4*, *SUFU* и *PPARA*) были связаны с негативной регуляцией катаболических процессов. Скрининг вариаций числа копий генов (CNV) показал, что 12,4 МБ генома гепарда было включено в сегментные дупликации (SD) (общие среди семи гепардов), указывающие на генные регионы и отдельные гены, которые могут влиять на процессы метаболизма, пищеварения и восприятия. Все эти гены, находящиеся под отбором, в расширенных генных семействах или в сегментных дупликациях дают возможность объяснить такие адаптации гепарда, как способность развивать и переносить огромное ускорение и поддерживать краткосрочную выносливость.

3.2.3. Анализ экспансии и сокращения генных семейств

Семейства генов представляют собой группы генов-паралогов, которые имеют очень схожие функции. Анализ динамики экспансии и сокращения генных семейств может дать ключ к пониманию эволюционных событий, которые сформировали геномы млекопитающих. Предполагается, что дупликация гена является мощным двигателем эволюционных изменений и инноваций [Ohno, 1970; Hurles, 2004; Magadum et al., 2013], в то время как потеря гена считается распространенной реакцией на изменения в селективном давлении в процессе эволюции, например, в происхождении человека [Hahn et al., 2007].

Для того чтобы сделать более подробную аннотацию генома гепарда, мы провели анализ генных кластеров, используя семь других геномов млекопитающих (гепард, тигр, лев, кошка, человек, собака, мышь и опоссум). Геном гепарда содержит 17863 семейства ортологичных генов. Среди них 10983 семейства ортологичных генов имелись у всех восьми геномов и 12114 генных семейства у кошачьих, в то время как 112 были уникальными исключительно для гепарда и домашней кошки (Рис. 9). Было обнаружено 1335 генов, уникальных для генома гепарда; 812 из них содержали 2293 белковых домена, идентифицированных с помощью InterProScan [Mitchell et al., 2015]. Геном гепарда имеет 814 расширенных и 2169 сокращенных генных семейств по сравнению с общим предком кошачьих (Рис. 10).

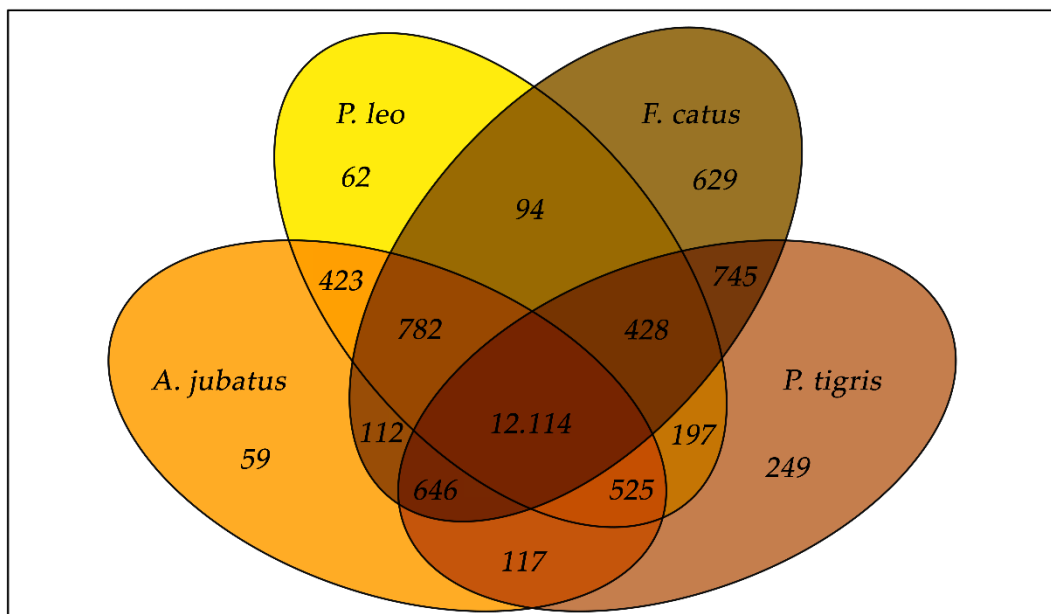


Рисунок 9. Анализ генных семейств ортологических генов. Уникальные и общие генные семейства в геноме гепарда относительно трех других представителей кошачьих.

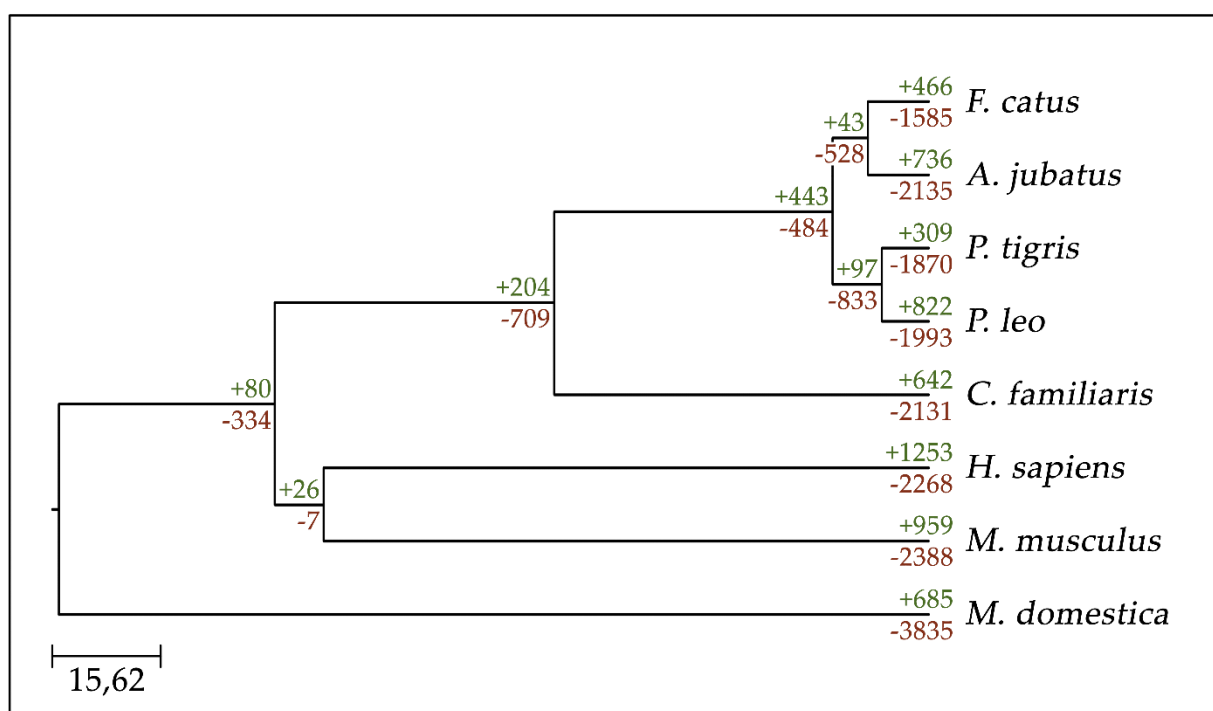


Рисунок 10. Экспансия и сокращение генных семейств в геноме гепарда. Кладограмма, отражающая филогенетические отношения между изучаемыми видами млекопитающих. Нанесенные на нее цифры, отмеченные зеленым и красным цветом, отражают количество генных семейств, подвергшихся экспансии и сокращению соответственно.

Расширенные генные семейства включали множество ГО-терминов, таких как обонятельные и G-связанные белковые рецепторы (также

подвергшиеся экспансии у других представителей *Felidae* [Cho et al., 2013; Kim et al., 2016]), которые могут быть связаны с физиологией гепарда. Семейство генов *LDH-A* и *LDH-B*, кодирующих ферменты лактатдегидрогеназы, являются важным элементом анаэробного метаболического пути. Показало двукратное увеличение числа генов у некоторых видов *Felidae* (кошка, гепард и лев) по сравнению с другими млекопитающими, что потенциально является характерной адаптацией для хищнического образа жизни, типичного для большинства видов кошачьих (Рис. 11). Интересно, что количество генов у тигра не демонстрирует существенного увеличения, что может быть связано как с качеством сборки генома, так и с другими особенностями физиологии, на текущий момент ответить на этот вопрос однозначно не представляется возможным.

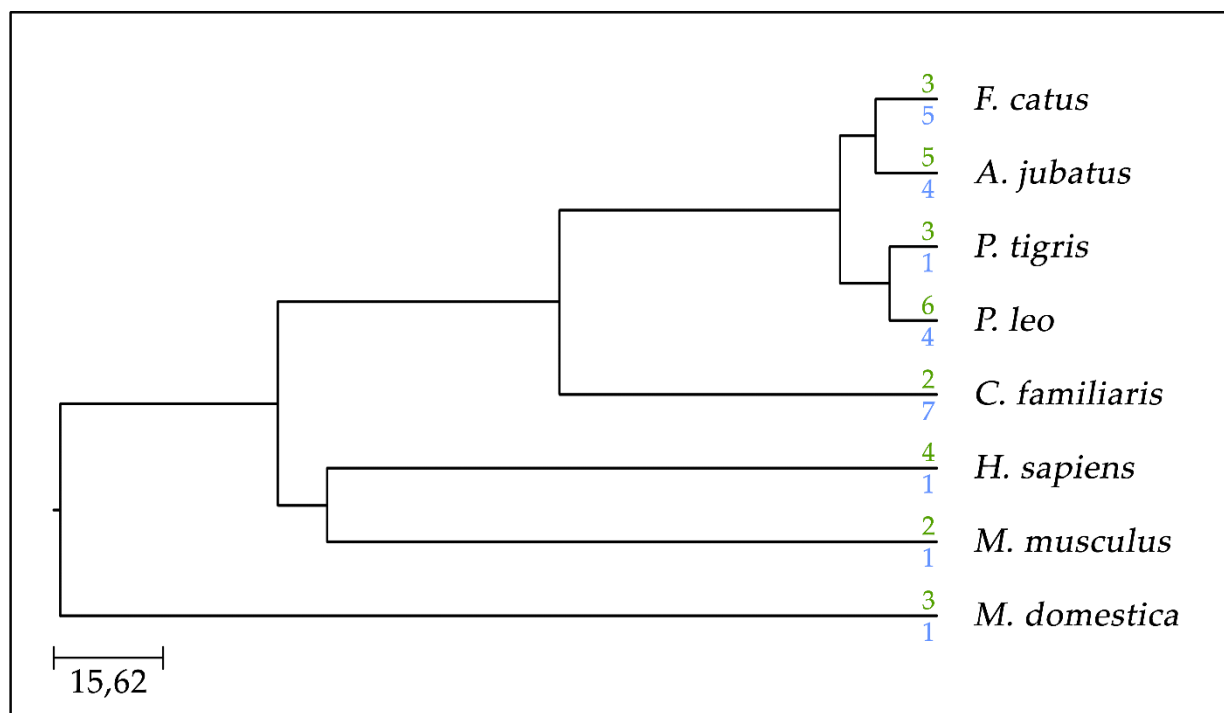


Рисунок 11. Кладограмма для восьми видов млекопитающих с количеством генов LDH, представленных в их геномах. Зеленый – *LDHA/LDHC*, голубой – *LDHB*.

3.2.4. Анализ генов связанных с размножением

Удивительно, что гепард имел самые высокие значения kA/kS среди пяти видов, используемых в анализе. Чтобы проверить надежность этих результатов, было проведено 500 бутстреп репликаций (Рис. 12).

Для проверки гипотезы о том, что в филогенетической линии гепарда увеличено значение kA/kS , был создан набор из 6348 однокопийных ортологов из следующих видов: домашняя кошка, амурский тигр, африканский гепард и домашняя собака. После фильтрации неоднозначно картированных областей было получено выравнивание кодирующих последовательностей. На основе этого выравнивания было выполнено 200 репликаций бутстрепа, и полученный набор данных был использован для анализа модели свободного коэффициента в PAML. Для полногеномного набора данных получены аналогичные результаты: гепард показал более высокие значения отношения kA/kS относительно других трех видов (Рис. 13).

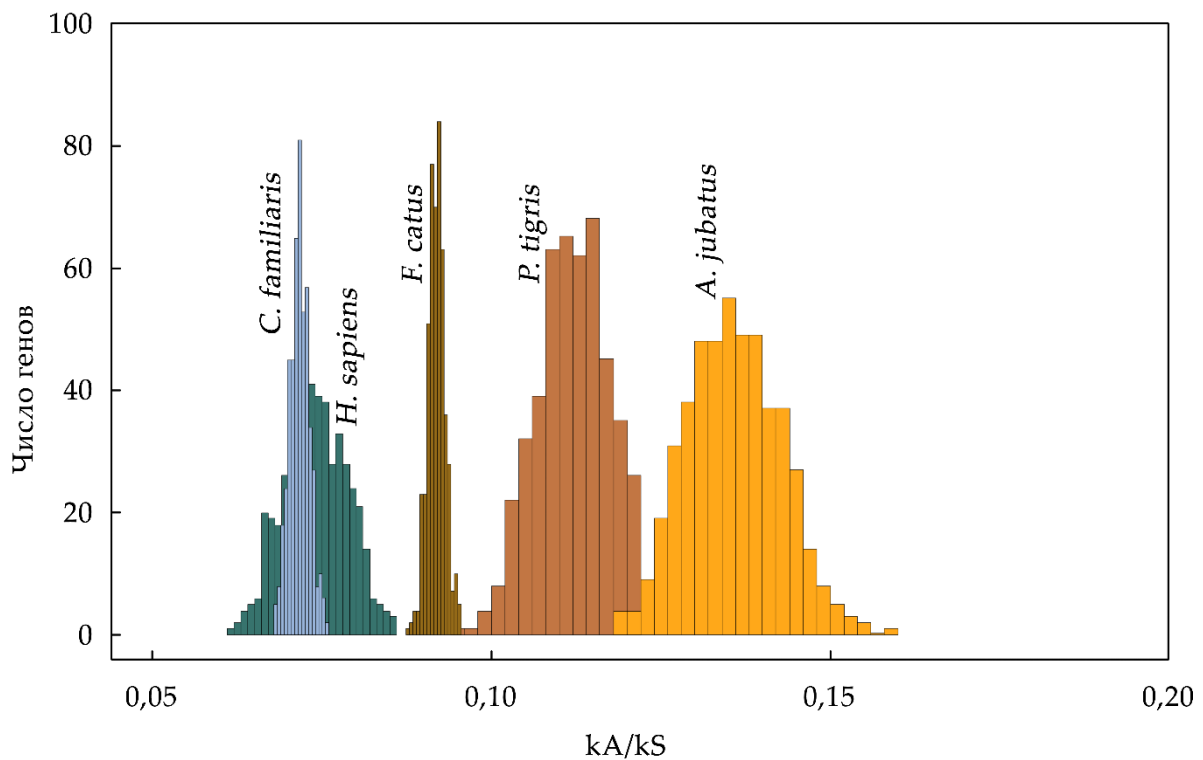


Рисунок 12. Распределение kA/kS значений в генах, связанных с размножением. kA/kS отношения были подсчитаны для пяти видов (домашняя собака, человек, домашний кот,

амурский тигр и африканский гепард) и основаны на 500 бутстреп репликациях модели свободного коэффициента в PAML.

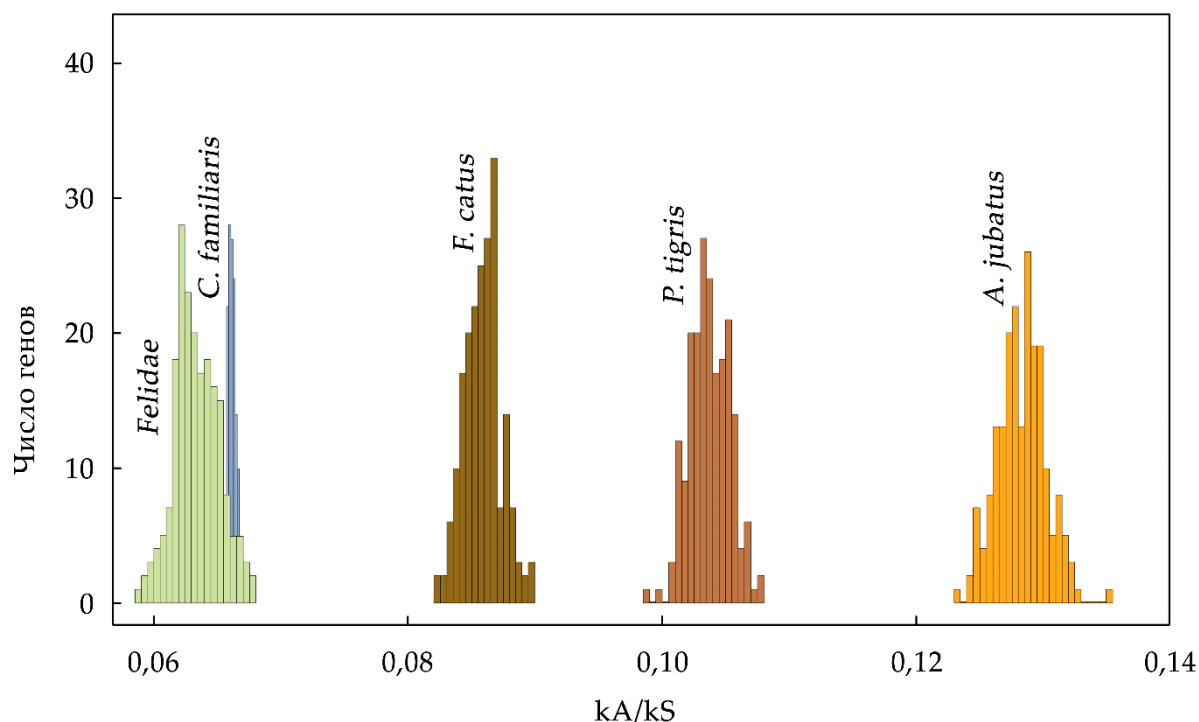


Рисунок 13. Распределение kA/kS значений для пяти видов (домашняя собака, человек, домашний кот, амурский тигр и африканский гепард) и для реконструированной линии *Felidae*. kA/kS значения основаны на 200 бутстреп репликациях для белок кодирующей последовательности “супергена” размером в 10 млн п.н.

Мы обнаружили в геноме гепарда 92 гена, связанных с размножением (например, оогенез и сперматогенез), с повышенными значениями kA/kS . В дальнейшем эти гены были вручную проверены с использованием общедоступных баз данных (GeneCards и Ensembl), чтобы найти гены, непосредственно связанные со сперматогенезом, азооспермией, олигоспермией, оогенезом и дисфункцией гонад. Был сформирован финальный список из 18 генов (Табл. 12) для поиска в базах данных генетических заболеваний, включая OMIM [Hamosh et al., 2000], KEGG [Ogata et al., 1999] и MalaCards [Rappaport et al., 2013], а также для скрининга всех несинонимичных замен и делеций. Патогенность мутаций оценивали с использованием PolyPhen2 [Adzhubei et al., 2013] с белками человека в качестве модели.

Таблица 12. Список генов, связанных с размножением, в которых были обнаружены мутации. Тяжесть мутации оценивалась при помощи базы данных PolyPhen2.

Идентификатор	Имя гена	P-значение	Описание эффекта
ENST00000295228	<i>INHBB</i>	3.58E-07	2 вероятно вредных замены
ENST00000377847	<i>PAX5</i>	1.35E-06	1 вероятно вредная замена
ENST00000399635	<i>TSSK2</i>	2.71E-03	Этот ген играет важную роль в сперматогенезе и ассоциирован с мужской фертильностью. Были найдены 3 консервативные мутации, специфичные для гепардов.
ENST00000396368	<i>KIA0430</i>	2.84E-03	Ген играет важную роль в оогенезе. Было обнаружено несколько вероятно вредных мутаций.
ENST00000394337	<i>CCDC135</i>	3.46E-03	Важный неконсервативный ген, содержит несколько мутаций, специфичных для гепардов.
ENST00000225538	<i>P2RX1</i>	6.08E-03	Три мутации, специфичные для гепардов. Одна из них в консервативной области.
ENST00000358526	<i>AKAP4</i>	1.58E-02	Ген, играющий важную роль в подвижности сперматозоидов. У гепардов обнаружено 5

			различных миссенс мутаций.
ENST00000252677	<i>BMP15</i>	1.62E-02	Ген играет важную роль в развитии яйцеклетки. Гепард имеет одну вероятно вредную мутацию.
ENST00000266991	<i>DHH</i>	2.13E-02	Ген ассоциирован с дисгенезией гонад. У гепарда одна вероятно вредная мутация.
ENST00000274008	<i>SPATA5</i>	2.31E-02	Одна вероятно вредная мутация
ENST00000341206	<i>LCN6</i>	2.77E-02	Одна вероятно вредная мутация
ENST00000395858	<i>MOV10L1</i>	4.88E-02	Одна вероятно вредная мутация

Среди 18 генов мы обнаружили один ген, который показал избыток возможных повреждающих миссенс-мутаций, и был связан с важными функциями сперматогенеза: *AKAP4* (Рис 14). Мы использовали секвенирование по Sanger для подтверждения мутаций *AKAP4* у 10 дикоживущих гепардов из Намибии. Четыре из пяти несинонимичных замен были подтверждены в 9 образцах и оказались гомозиготными. Пятая мутация пока не была изучена, поскольку она была расположена в одной из праймерных последовательностей, которые мы сконструировали для амплификации *AKAP4*.

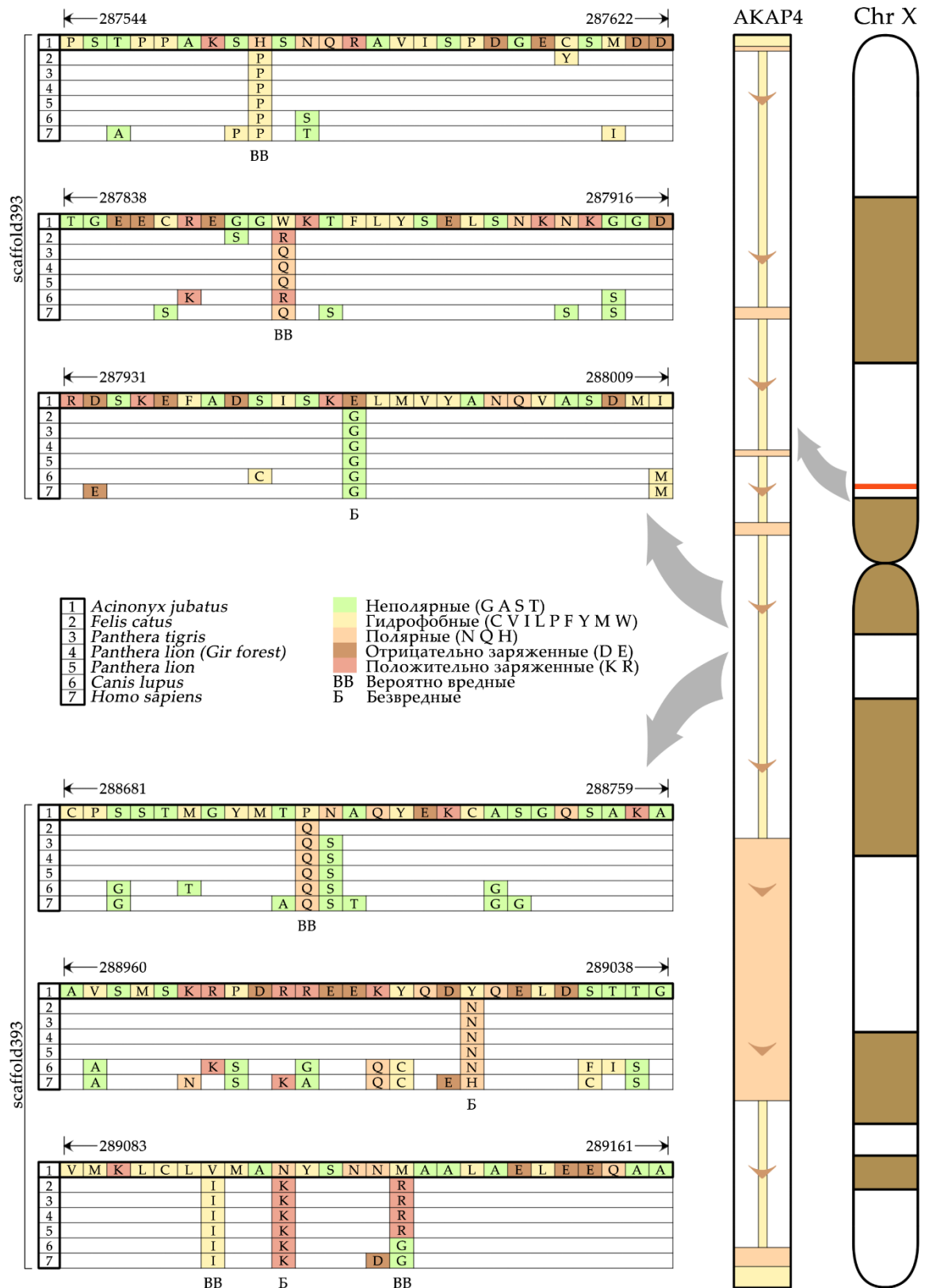


Рисунок 14. Белковое выравнивание гена *AKAP4* для кошки, тигра, собаки, гепарда, человека, африканского и азиатского льва, демонстрирующее избыток вероятно вредных мутаций в геноме гепарда. Используются следующие сокращения. Б — Безвредная (benign), BB — Вероятно вредная (possibly damaging).

Современные представители вида *Acinonyx jubatus* демонстрируют множественные физиологические корреляты инбридинга как в неволе, так и в диких популяциях. По сравнению с другими видами *Felidae*, гепарды проявляют конститутивные нарушения в размножении, включая низкую плодовитость в неволе, в среднем 80% сперматозоидов с аномальной морфологией не эякулят и повышенную частоту акросомных дефектов, как это наблюдалось у других инбредных природных популяций [O'Brien et al., 1983; Crosier et al., 2007; Fitzpatrick and Evans, 2009]. Чтобы исследовать гены, связанные с репродуктивными проблемами гепарда, мы идентифицировали 964 гена человека с терминами генной онтологии (GO), относящимися к размножению, и кодирующие 1.730 РНК-транскриптов. Список был сужен до 656 генов, у которых был однокопийный ортолог среди геномов гепарда, кошки, тигра, собаки и человека. В целом гепарды показали гораздо более ускоренное накопление несинонимичных замен в этих генах по сравнению с другими видами. Мы идентифицировали 92 гена гепарда со статистически значимым повышенным kA/kS отношением, для них мы определили тип и частоту повреждающих мутаций. Восемнадцать генов, связанных со сперматогенезом и оогенезом, содержали распространенные и/или зафиксированные в популяции мутации, ранее ассоциированные с азооспермией, олигоспермией и дисфункцией гонад (Табл. 12).

Среди них один ген (*AKAP4*) показал особенно ускоренное накопление потенциально повреждающих мутаций. Выравнивание аминокислотной последовательности этого гена вместе с ортологами из генома человека, амурского тигра, азиатского льва, домашней и дикой кошки, представлено на Рис. 14. Мутации в гене *AKAP4*, обнаруженные в геноме гепарда, не наблюдались у тигра, домашней и дикой кошки (*Felis silvestris*), а также у азиатского льва, чья единственная популяция, обитает в национальном парке Гир-Форест, Индия, и демонстрирует крайнее генетическое истощение и обширные репродуктивные дефекты [O'Brien, 1987]. Секвенирование по Сангеру подтвердило четыре из пяти гомозиготные замены в *AKAP4* у 9

намибийских гепардов. Пятая мутация не была явно проверена. Обнаруженные нами нарушения репродуктивных генов являются важными фактами для описания особенностей репродуктивной физиологии, которая наблюдается у всех гепардов.

Мы использовали еще один подход, основанный на аннотации генных эффектов, в семи секвенированных геномах для выявления вредных мутаций, которые сегрегированы в популяциях гепарда. Единичные нуклеотидные варианты (SNV), показывающие возможные вредные эффекты, были идентифицированы с использованием snpEff [Cingolani et al., 2012], и отфильтрованы с именами 656 ранее идентифицированных однокопийных ортологов для пяти видов, которые связаны с репродукцией и имеют потенциально вредные эффекты (например, преждевременный стоп-кодон и/или поврежденные сайты сплайсинга). Всего был найден 61 ген, среди них 20 показали первичную связь с репродуктивными аномалиями, обнаруженными у гепардов. Эти мутации обеспечивают ценную основу для будущих ассоциативных исследований репродуктивных нарушений в популяциях гепардов.

Геномный анализ выявил убедительные данные для групп генов, участвующих в размножении, которые накапливали избыток аминокислотных изменений у гепардов по сравнению с другими кошачьими. Мы идентифицировали восемь фиксированных мутаций в локусе гена *AKAP4*, который экспрессируется исключительно в семенниках. Гомологи *AKAP4* играют критическую роль в развитии и подвижности сперматозоидов и возникновении аномалий у сперматозоидов для некоторых видов млекопитающих [Miki et al., 2002]. Пять гомозиготных мутаций в гене *AKAP4*, весьма вероятно, объясняют высокую долю дефектных сперматозоидов (в среднем 81,6% поврежденных сперматозоидов), обнаруженную у гепарда.

3.3. Изучение генетического разнообразия в свободноживущих популяциях африканского гепарда, *Acinonyx jubatus*

3.3.1. Изучение генетического разнообразия

Первоначально мы получили список SNV для сайтов с высоким покрытием по всему геному, в котором минимальная и максимальная глубина секвенирования для каждого образца были установлены на 5X и 30X соответственно. После применения наших критериев и фильтров было установлено в общей сложности 3,44 млн. сайтов SNV (Табл. 13).

Таблица 13. Статистика по обнаруженным SNV. Общее количество SNV включает все SNV в кодирующих областях и повторяющихся элементах. Синглтоны - это SNV, встречающиеся только в одном образце. Частота SNV рассчитывается как общее количество SNV, деленное на размер генома.

Всего SNV	Синглтоны	Размер генома	Частота SNV
3438824	1432827	2048690294	0,0016

Кроме того, были отфильтрованы позиции, расположенные в регионах, содержащих повторы, в результате чего был сформирован окончательный набор из 1820419 SNV, что составляет 53% от первоначального числа SNV (3438824). Мы оценили распределение SNV по геному у всех особей (Табл. 14).

Таблица 14. Распределение гетерозиготных SNV в геноме гепарда. Для каждого образца мы рассматривали SNV, генотип которых был гетерозиготным. Для оценки распределения SNV в геноме мы разделили геном на неперекрывающиеся окна длиной 50 тысяч п.н. и вычислили минимальное, медиану, среднее и максимальное количество SNV в каждом окне.

Образец	Количество гетерозиготных SNV						Частота SNV
	Мин.	Медиана	Среднее	Макс.	В окнах	Всего	
AJU878	0	9	10,35	248	484403	494308	0,00020

AJU171	0	8	9,82	330	459588	469343	0,00019
AJU884	0	9	10,51	337	491758	501858	0,00021
AJU172	0	9	10,49	253	490857	500450	0,00021
AJU173	0	8	9,75	343	456263	466218	0,00019
AJU955	0	9	10,44	348	488607	498729	0,00020
AJU981	0	9	10,37	335	484951	494299	0,00020

Среди всех обнаруженных SNV 73,7% были расположены вне белок кодирующих генов, только 1,3% были внутри экзонов, а основная их часть, 24,92%, была обнаружена внутри интронов (Табл. 15).

Таблица 15. Распределение SNV по разным функциональным участкам генов в геноме гепарда. Эффекты SNV, сгруппированные по положению в геноме. Были проанализированы как некодирующие районы внутри и снаружи генов, так и белок кодирующие участки.

Образец	5' прилежащие районы	Эзоны	Внутригенные некодирующие области	Интроны	3' прилежащие районы
AJU878	20906	6761	364130	122791	18370
AJU171	19804	6243	347735	114904	17347
AJU884	21203	6735	370760	123953	18624
AJU172	21253	6879	370573	122575	19204
AJU173	19661	6350	343955	115369	17635
AJU955	20971	6798	369291	122072	18721
AJU981	21128	7625	363638	122511	18912

Три дополнительных образца гепарда из Танзании и три из Намибии были секвенированы с низким покрытием, в результате было выявлено 1.820.419 различных однонуклеотидных варианта (SNV) и проведено

сравнение с SNV у других представителей *Felidae* и прочих млекопитающих. Мы оценили степень и структуру геномного разнообразия, используя семь различных мер, описанных ниже, каждая из которых подтвердила значительное снижение уровня генетической и геномной изменчивости у гепарда.

1) Гепарды демонстрируют самую низкую встречаемость SNV в геноме среди 11 видов, включая человека, домашнюю кошку, гориллу (*Gorilla gorilla*), льва (*Panthera leo*) и тасманского дьявола (*Sarcophilus harrisii*), частота встречаемости SNV на 90% меньше, чем, например, у изученного ранее домашнего кота (Boris) из Санкт-Петербурга [Tamazian et al., 2014], Рис. 15.

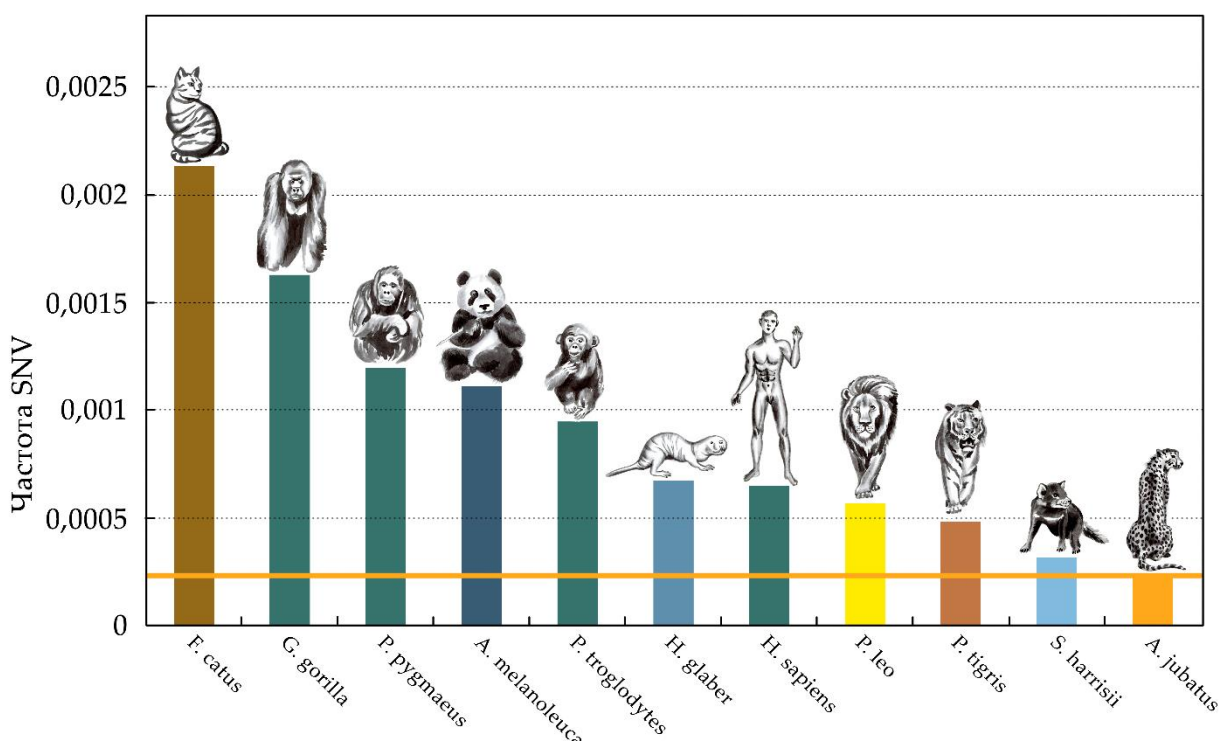


Рисунок 15. Частота SNV у одиннадцати видов млекопитающих. Частота SNV для каждого вида была оценена как число вариативных позиций, усредненное на размер генома. Повторы не фильтровались.

2) Референсный геном африканского гепарда разбивали окнами по 50 тысяч п.н., которые использовали для оценки плотности SNV. В общей сложности 46787 окон составили 2,337 Гб или 99,12% от общей длины генома.

Анализ большинства окон показал, что гепарды в 8-15 раз менее генетически разнообразны, чем человек, домашний кот (Борис) или Европейский дикий кот (*Felis silvestris*) (Рис. 16). Единственным видом или популяцией с сопоставимой, либо более низкой плотностью SNV, чем гепард, был азиатский лев из Национального парка Гир-Форест в Индии. Из популяции, которая, как известно, претерпела сильное снижение численности и экстремальную генетическую гомогенизацию в своей недавней истории [Rangarajan, 2013].

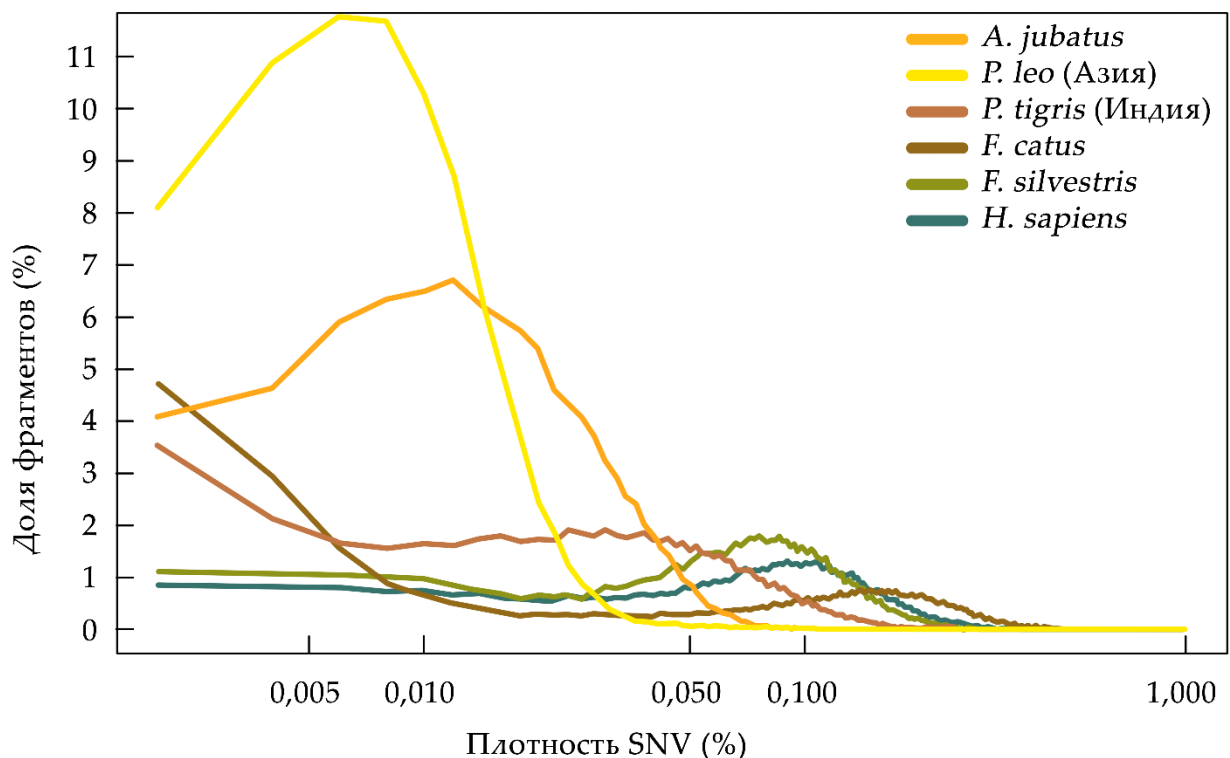


Рисунок 16. Плотность SNV в скользящих окнах по 50 тысяч п.н. у гепарда, тигра, дикой и домашней кошки (Борис), и человека. Фрагменты, имеющие длину меньше указанного размера окна, были исключены из дальнейшего анализа; Большинство из этих фрагментов представляло собой контиги с длиной менее 500 п.н. Всего было построено и проанализировано 46787 окон общей длиной 2,337 млрд п.н.

3) Кодированные гены гепарда обнаружили резкое 50-кратное (~98%) сокращение генетической вариации к уровню вариации в геноме у домашней или дикой кошки (Рис. 17). Экстремальное сокращение вариативных позиций в кодирующих генах объясняет первоначально открытую низкую генетическую изменчивость гепарда, обнаруженную три десятилетия назад в исследованиях с использованием аллозимов, электрофоретических вариантов

клеточного белка и полиморфизма длины рестрикционных фрагментов на основе генов (RFLPs) [O'Brien et al., 1983; O'Brien et al., 1985].

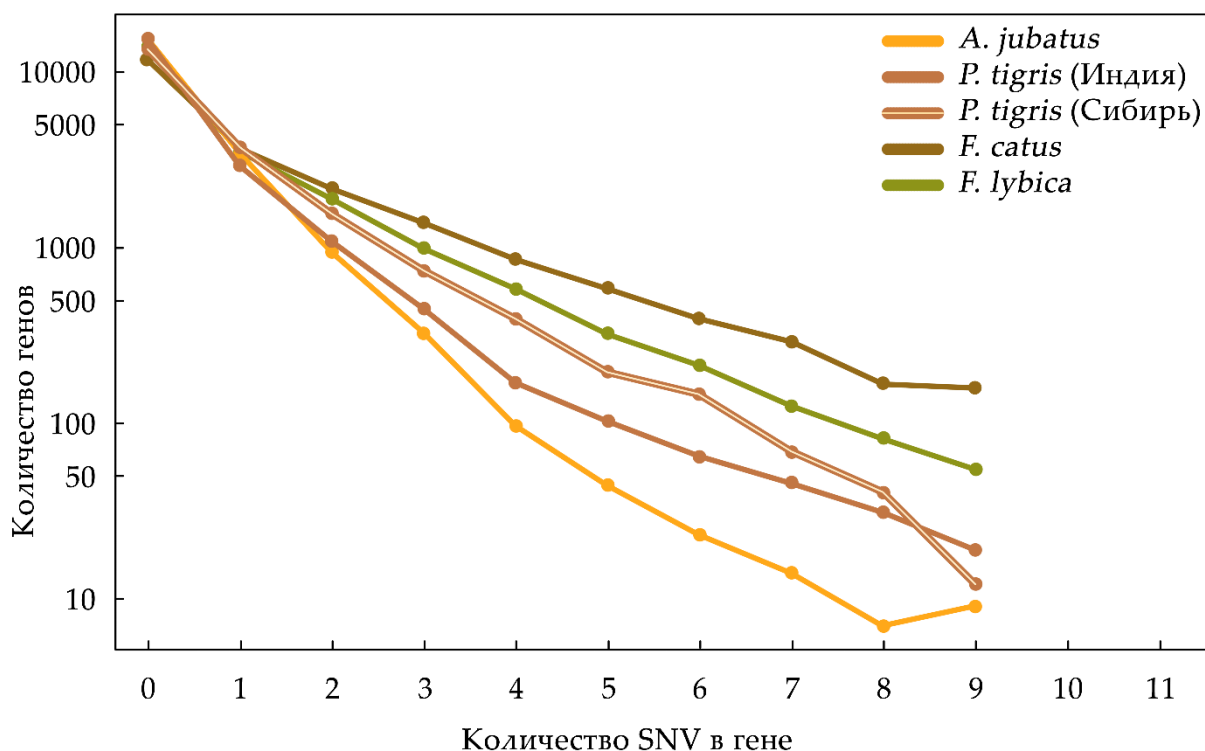


Рисунок 17. Количество несинонимичных SNV в белок-кодирующих генах в геномах кошачьих. Значения для генов с числом SNV больше 9 не изображены.

4) Гепарды демонстрируют в среднем в 10 раз более протяженные участки гомозиготности по сравнению с геномом дикой и домашней кошки. В среднем 93% генома каждого гепарда состоит из областей гомозиготности (RON - то есть областей генома, в которых не было обнаружено SNV) с медианной длиной около 2000 п.н. Геном домашнего бродячего кота (Борис), живущего в Санкт-Петербурге, сравнивали с высокоинбредной абиссинской кошкой (Cinnamon), и гепардом (Chewbaaca, AJU878). Медиана длины регионов гомозиготности (RON) у гепардов (семь особей), африканских львов (пять особей), сибирских и бенгальских тигров и домашней кошки представлены на Рис. 18. Плотность SNV для гепарда, дикой и домашней кошки была оценена приблизительно для 15000 областей по 100 млн.п.н каждая по всему геному (Рис. 19, 20 и 21).

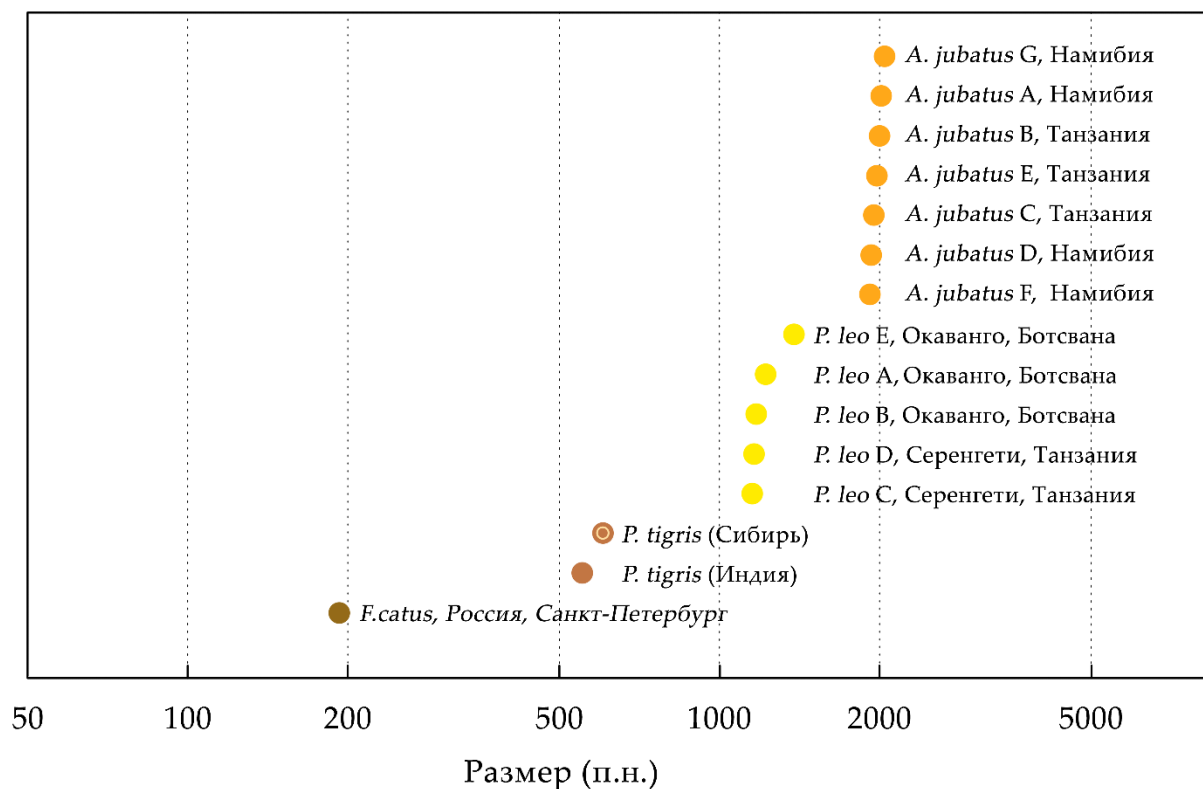


Рисунок 18. Размер участков гомозиготности в геномах. Для каждого образца отмечена медиана длин всех регионов гомозиготности.

5) В среднем в геноме гепарда обнаружено очень мало гетерозиготных SNV, 0,019-0,021%, что составляет до 50-61% гетерозигот у тигров, 30% у людей и 15% у домашних кошек [Tamazian et al., 2014; Cho et al., 2013].

6) Полные митохондриальные геномы гепарда аналогичным образом показывают снижение количества SNV по отношению к другим видам в среднем на 90% (Табл. 16).

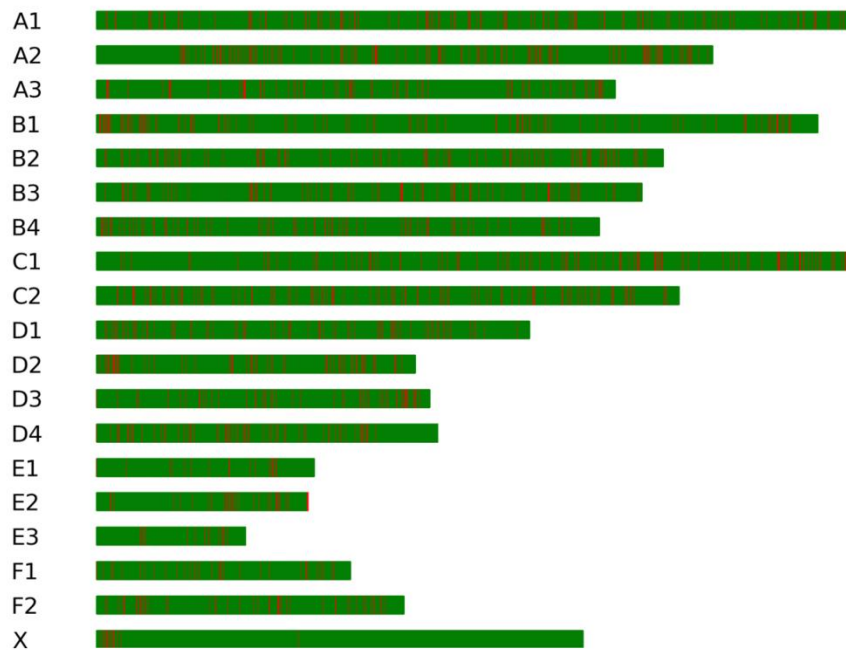


Рисунок 19. Регионы гомозиготности в геноме гепарда. Области высокой изменчивости (> 40 SNVs / 100 т.п.н.) окрашены в красный цвет; Яркие выраженные гомозиготные области (≤ 40 SNVs / 100 т.п.н.) окрашены в зеленый цвет.

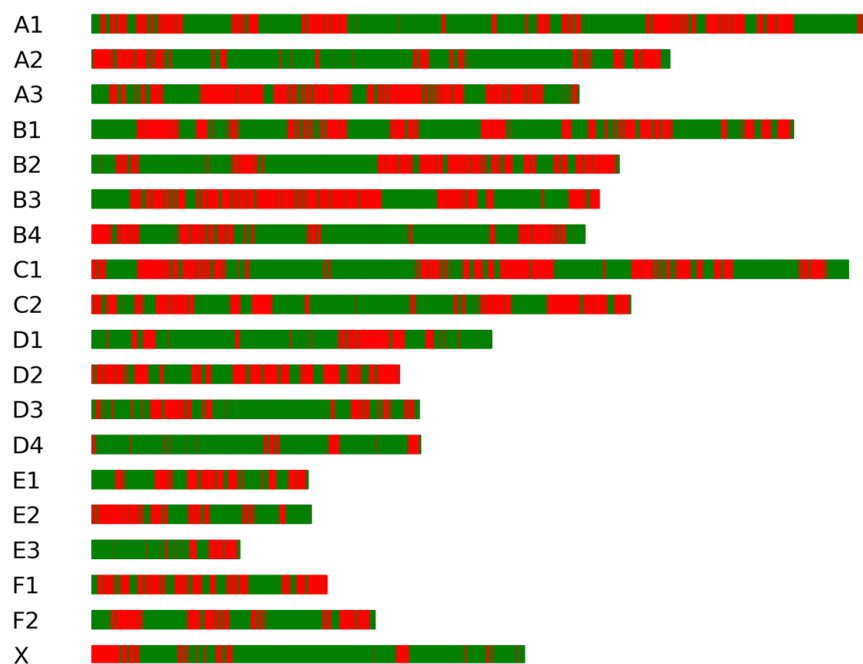


Рисунок 20. Районы гомозиготности в геноме инбредной домашней кошки (Cinnamon). Области высокой изменчивости (> 40 SNVs / 100 kbp) окрашены в красный цвет; ярко выраженные гомозиготные области (≤ 40 SNVs / 100 kbp) окрашены в зеленый цвет.

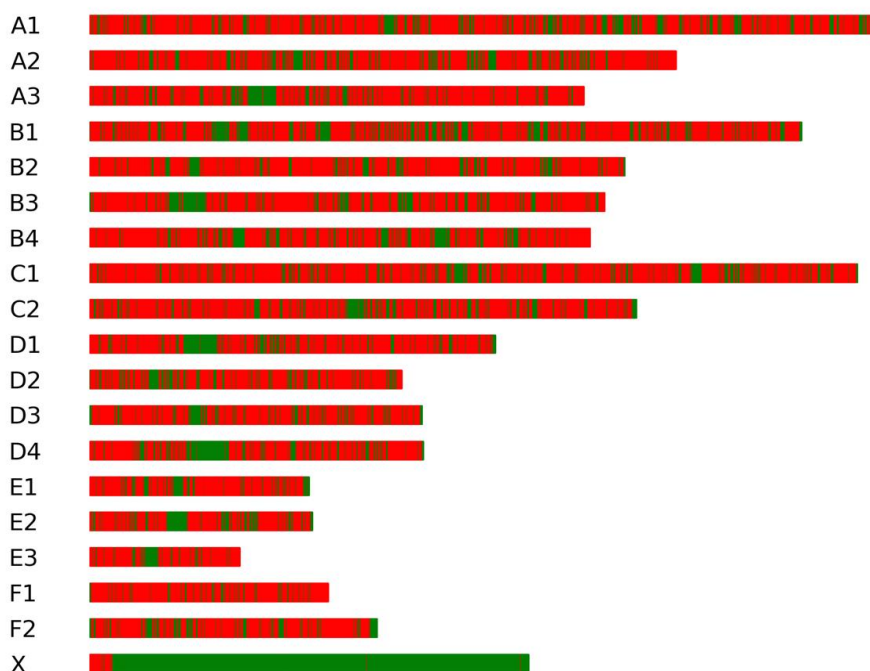


Рисунок 21. Районы гомозиготности в геноме бродячего кота из Санкт-Петербурга (Борис). Области высокой изменчивости (> 40 SNVs / 100 т.п.н.) окрашены в красный цвет; ярко выраженные гомозиготные области (≤ 40 SNVs / 100 т.п.н.) окрашены в зеленый цвет.

Таблица 16. Генетическое разнообразие (P_i) в митохондриальном геноме гепарда в сравнении с восемью другими млекопитающими.

Виды	P_i	Стандартное отклонение
<i>Acinonyx jubatus</i> (Tanzania)	8.00E-05	3.00E-05
<i>Acinonyx jubatus</i> (Namibia)	7.10E-04	3.30E-04
<i>Bison bison</i>	7.90E-04	7.00E-05
<i>Bos taurus</i>	8.60E-04	1.50E-04
<i>Orcinus orca</i>	1.20E-03	2.10E-04
<i>Homo sapiens</i>	1.65E-03	8.00E-05
<i>Canis lupus familiaris</i>	3.31E-03	1.00E-04

<i>Equus caballus</i>	3.80E-03	1.00E-04
<i>Canis lupus</i>	5.18E-03	3.70E-04
<i>Pan troglodytes</i>	5.38E-03	2.00E-07

3.3.2. Сборка района МНС при помощи референсной последовательности

Строение генома в области МНС у млекопитающих создает дополнительные проблемы при сборке и аннотации. В большинстве случаев собранные геномные последовательности с генами МНС классов I и II достаточно сильно фрагментированы, даже если для сборки скаффолдов используется несколько различных библиотек. Чтобы повысить качество дальнейшего анализа в области МНС, мы создали точную сборку на основе референса специально для МНС региона гепарда.

Мы идентифицировали скаффолды, предположительно принадлежащие к области МНС, на основе гомологии с FLA-последовательностью домашней кошки [Pontius et al., 2007]. Чтобы восстановить область МНС класса I, мы картировали прочтения гепарда на исходную последовательность FLA в качестве референса. Для выравнивания и создания консенсусной последовательности мы использовали программы bowtie2 [Misale, 2014] и samtools [Li et al., 2009] соответственно. Последовательности ВАС от домашней кошки использовали для дополнительной поддержки ориентации и порядка идентифицированных фрагментов (Рис 22).

Из сборки гепарда удалось выделить 13 скаффолдов и 7 контигов (общий размер 8,3 млн п.н.), которые содержали 278 генов, относящихся к области FLA, причем относительный порядок генов у гепарда был аналогичен таковому в геноме *Felis catus*. Расширенный класс II, класс II и класс III были хорошо представлены в нескольких крупных скаффолдах. Наиболее проблемным регионом для сборки был район класса I, где мы смогли собрать

и идентифицировать всего несколько генов и несколько фрагментов-гомологов *FLA-A*. Консенсусная последовательность для сборки на основе референса области МНС была основана на прочтениях, полученных из библиотек размером в 500 п.н. из семи гепардов (один с высоким покрытием и шесть с низким).

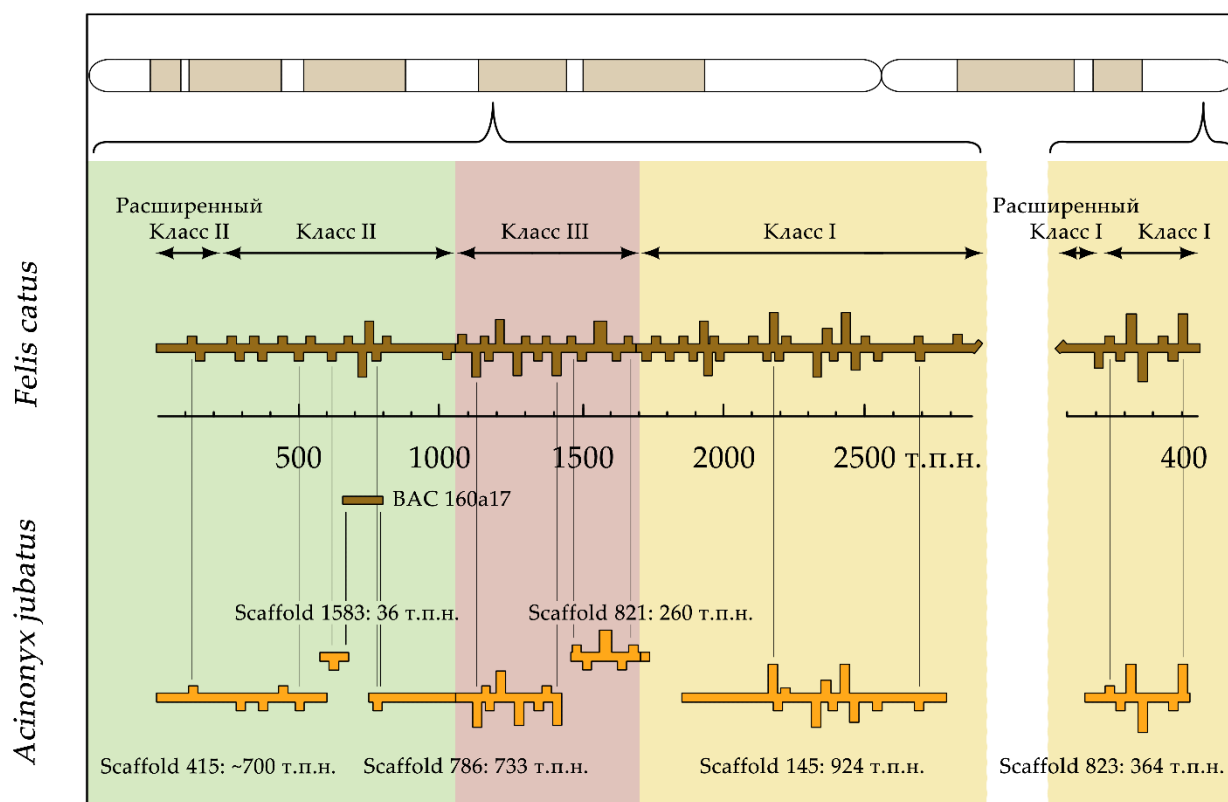


Рисунок 22. Часть региона МНС гепарда, собранного на основе референса. Сверху – референсный район МНС из домашней кошки (*Felis catus* FLA). Снизу – Скаффолды из de novo сборки генома гепарда, ориентированные по порядку генов и BAC домашней кошки.

Кроме того, мы проанализировали консервативные домены в генах МНС классов I и II с использованием базы данных NCBI [Pruitt et al., 2007] и CLUSTALW [Thompson et al., 2003] (Табл. 17)

Таблица 17. Статистика по предсказанным генам в районе МНС. Статистика по размеру (п.н.) предсказанных генов в двух вариантах сборки.

Использованная программа		#	Мин. длина п.н.	Медиана	Среднее	Макс. длина п.н.
Exonerate	Скаффолды	129	512	4046	6351	17970
Augustus		145	208	11300	22020	211100
Exonerate	Контиги	116	512	3700	5530	17900
Augustus		130	312	8822	16790	140700

Генетическое разнообразие в МНС области генома гепарда соответствовало уровню разнообразия по всему геному, и мы смогли найти лишь несколько полиморфных позиций (Табл. 18).

Таблица 18. Генетическое разнообразие в предсказанных генах МНС I и II класса. Полноту структуры гена оценивали по наличию консервативных доменов: если домены присутствовали частично, то ген обозначали как усеченный (У), в противном случае ген обозначали как полноразмерный (П). S и NS обозначают синонимичные и несинонимичные SNV соответственно. Аннотацию гена МНС для домашней кошки (*F. catus*) взяли у Yuhki et al [Yuhki et al., 1989]. Значение NA для гена кошки МНС означает, что ген не был аннотирован в оригинальной статье.

Класс	Имя гена	Полнота	#SNV в <i>A. jubatus</i>		#SNVs в <i>F. catus</i>	
			S	NS	S	NS
II	<i>DRBp</i>	П	1	4	NA	NA
II	<i>DRAp</i>	У	0	3	NA	NA
II	<i>DMA</i>	П	0	2	NA	NA
II	<i>DMB</i>	П	0	0	NA	NA
II	<i>DOB</i>	П	1	0	NA	NA
II	<i>DRB4</i>	П	1	1	8	25
II	<i>DRB1</i>	У	0	0	1	2
II	<i>DRA1</i>	П	0	0	2	0
II	<i>DRB3</i>	П	0	0	5	13
II	<i>DRA2</i>	П	0	0	0	0
II	<i>DRA1</i>	П	0	0	0	0

I	<i>FLA-A</i>	II	1	2	1	0
I	<i>FLA-C</i>	Y	0	2	2	1
I	<i>FLA-E</i>	Y	0	0	0	0
I	<i>FLA-F</i>	Y	0	1	4	40
I	<i>FLA-H</i>	Y	0	0	15	50
I	<i>FLA-J</i>	Y	0	0	6	7
I	<i>FLA-K</i>	II	0	0	0	1
I	<i>FLA-L</i>	Y	0	3	1	1
I	<i>FLA-O</i>	II	0	0	3	4
I	<i>FLA-Q</i>	Y	0	2	NA	NA

Наконец, мы исследовали у гепарда структуру и вариации главного комплекса гистосовместимости (МНС), который состоит из кластера ~ 280 генов, связанных с иммунитетом. Предыдущее исследование продемонстрировало, что гепарды принимали реципрокные аллотрансплантаты кожи от неродственных особей, как если бы они были иммунологически идентичны, предполагая, что генетическое разнообразие в регионе МНС гепарда сильно снижено [O'Brien et al., 1985]. Сборка на основе референса и ВАС из 20 скаффолдов и контигов (общий размер 8,3 млн п.н.) показала наличие 278 генов из областей расширенного класса II, класса II, класса I и расширенного класса I. Хотя для большинства регионов было получено хорошее покрытие, полные гомологи определенных генов МНС класса (*FLA-I*, *-F*, *-H* и *-M*) не были обнаружены (Табл. 18).

Когда мы сравнивали структурную организацию и генный порядок МНС гепарда с другими видами, гепард и домашняя кошка были очень похожи, но отличались от собаки и человека. МНС гепарда и кошки включают три функциональных гена вомероназального рецептора (важные для распознавания феромонов [Keverne, 1999]) в расширенном районе МНС I класса (эти гены отсутствуют у человека, других приматов и собаки). Кошка и гепард также показали экспансию определенных генов обонятельных рецепторов (30 генов, занимающих 0.9 млн п.н.) в расширенном классе МНС-

I [Montague et al., 2014]. Мы сравнили количество обнаруженных SNV (синонимичных и несинонимичных) в генах МНС гепарда (из Намибии и Танзании), домашней и дикой кошки, человека и собаки [Tamazian et al., 2014; Montague et al., 2014]. Мы обнаружили 95-98% -ное снижение SNV разнообразия в обеих популяциях гепардов, а также у Cinnamon (высокоинбредного абиссинского кота [Tamazian et al., 2014]) по сравнению с обилием SNV у бродячего домашнего кота (Борис), человека и собаки (Рис. 24).

Сокращение количества SNV в областях МНС инбредной кошки и гепарда включает как синонимичные, так и несинонимичные аминокислотные замены. Подобные функциональные варианты обычно являются целью балансирующего отбора и отражают историю взаимодействия вида с различными патогенами [Cagliani and Sironi, 2013] (Табл. 18).

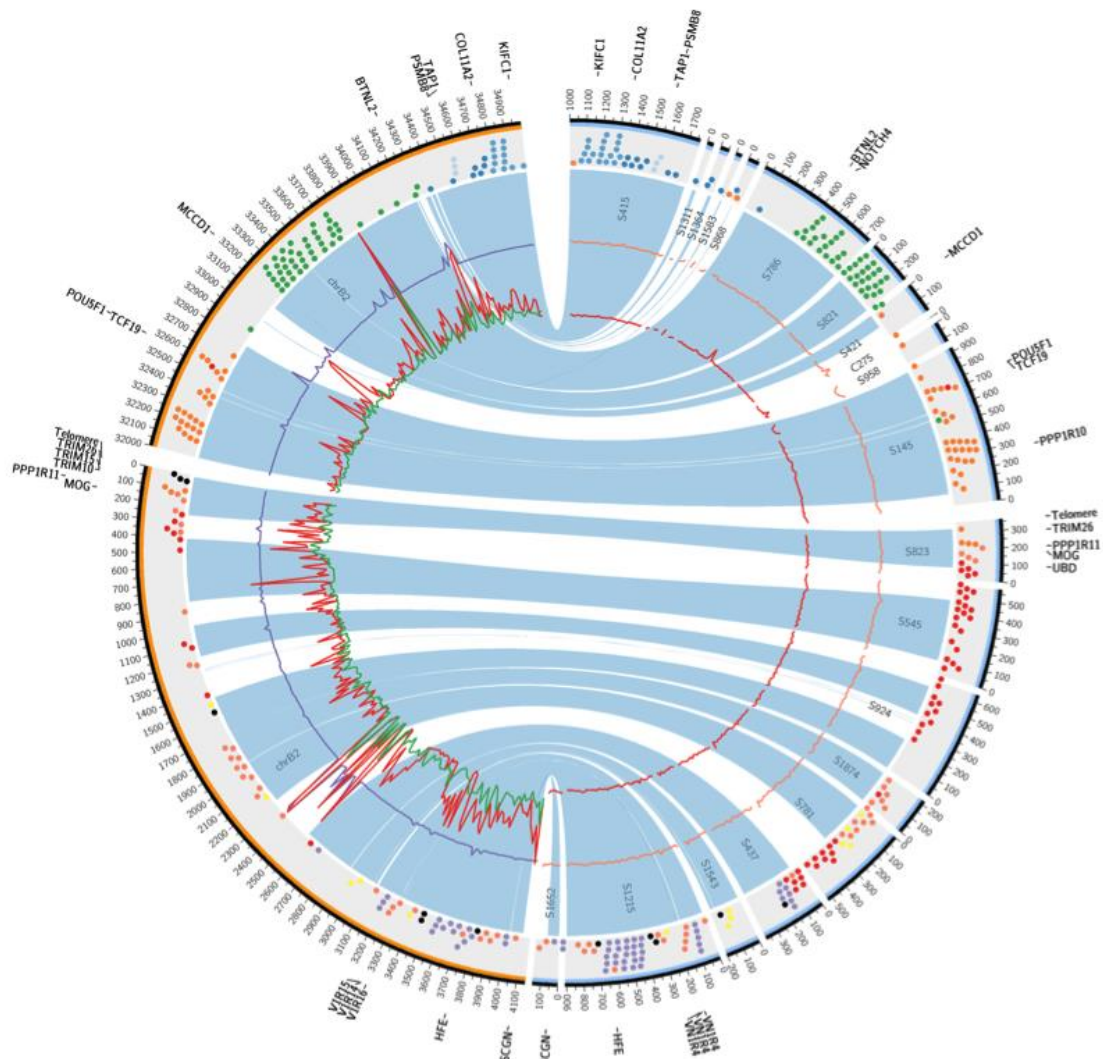


Рисунок 23. Сравнение организации МНС и генетического разнообразия между гепардом и домашней кошкой. *Левая сторона:* Два сегмента хромосомы В2, содержащие МНС гены в геноме домашней кошки. Основаны на ВАС библиотеках [Pontious et al., 2007]. *Правая сторона:* Скаффолды гепарда, относящиеся к региону МНС. Порядок и ориентация скаффолдов были определены по анализу синтении (Светло-голубые полосы). Отдельные гены отмечены точками и окрашены в зависимости от своего класса: Светло-голубой – расширенный класс II, Голубой - класс II, Зеленый - класс III, Оранжевый - класс I, Красный – обонятельные рецепторы, и Фиолетовый - гистоны. Генетическое разнообразие в регионе МНС оценивалось в неперекрывающихся окнах по 50 тысяч п.н. Результаты изображены в виде цветных линий на рисунке; для кошек: Зеленый – дикая кошка, Красный - Boris и Фиолетовый - Cinnamon; для гепардов: Красный - Танзания и Оранжевый – Намибия.

3.4. Изучение популяционной истории африканского гепарда на основе данных полногеномного секвенирования

Результаты PSMC показали постепенное сокращение эффективной численности популяции в течение времени без каких-либо свидетельств о наличии эффекта бутылочного горлышка в прошлом (Рис. 24). Подобный результат может быть обусловлен спецификой анализа PSMC, имеющего более низкую чувствительность к событиям в более недавнем прошлом (поскольку меньше коалесцентных событий оставили след в геноме) и/или, в результате того, что событие, приведшее к сокращению численности, было настолько быстрым и тяжелым, что привело к потере всех более ранних следов.

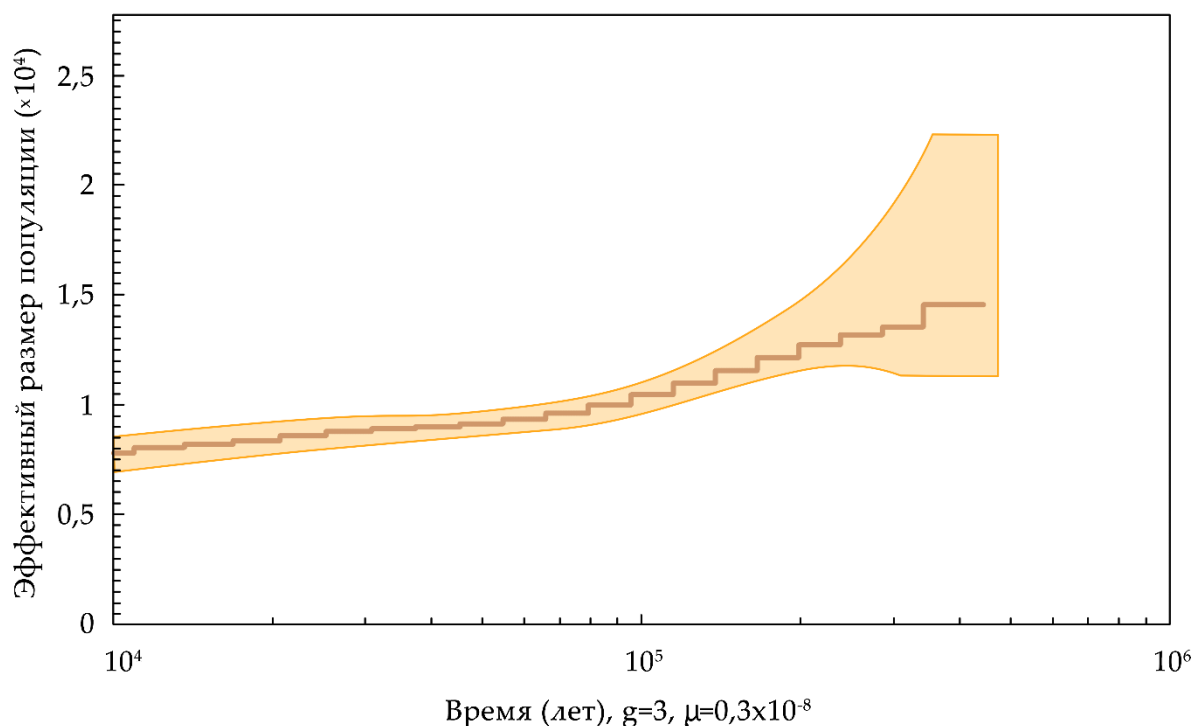


Рисунок 24. Реконструированная на основе метода PSMC эффективная численность популяции гепарда на протяжении от 1 млн. до 10 т. лет. На оси X отмечено время, оцененное по времени дивергенции, а по оси Y - эффективный размер популяции (на 10^4).

Четыре сценария были смоделированы при помощи DaDi и GADMA, чтобы сравнить соответствие каждой модели с наблюдаемыми данными SFS

гепарда. Было рассчитано значение функции правдоподобия для каждой модели. Чтобы определить, какая именно модель лучше всего подходит для объяснения картины, наблюдаемой в реальных данных для популяций из Намибии и Танзании, мы использовали два критерия:

1. Оптимизированное значение функции максимального правдоподобия (Табл. 19)
2. Дисперсия результатов, оцененная на 100 бутстреп репликациях (Рис. 25)

Таблица 19. Четыре демографические модели, которые были проверены для оценки демографической истории гепарда, и значения функции правдоподобия для них. Модель 4 имеет наибольшее значение функции правдоподобия. Мы обнаружили, что дисперсия правдоподобия для 4 модели намного меньше по сравнению с другими моделями.

Модель	Описание сценария	Значение логарифма правдоподобия
1	2D IM модель (Предковая популяция распадается на две субпопуляции, далее следует ограниченная миграция из одной субпопуляции в другую)	-51979,09
2	2D BIM модель (Предковая популяция сначала переживает эффект бутылочного горлышка, а затем разделяется на две субпопуляции)	-77657,60
3	2D SBR модель (Предковая популяция сначала распадается на две субпопуляции, за чем следует эффект бутылочного горлышка, далее происходит восстановление численности в каждой субпопуляции)	-50753,12
4	2D ISB модель (Предковая популяция, начиная с ~ 100 000 лет до разделения на две субпопуляции растет в размере, после разделения следует эффект бутылочного горлышка, который две	-43.587,07

	субпопуляции переживают независимо)	
--	-------------------------------------	--

Основываясь на дисперсии и значении правдоподобия, мы выбрали 2D-
ISB (модель 4) как наиболее вероятную. Эффективный размер популяции,
уровень миграции, сроки прохождения через бутылочное горлышко и время
разделения на субпопуляции реконструированы исходя из модели 2D-ISB
(Табл. 20).

Таблица 20. Значения параметров для демографических событий, исходя из 4 модели (2D ISB).

	Границы значений	
	Верхняя	Нижняя
Уровень миграция из восточной субпопуляции в южную субпопуляцию	0,000994	0,001095
Уровень миграции из южной субпопуляции в восточную субпопуляцию	0,001049	0,001166
Размер популяции до разделения	134873	150096
Общий размер восточной субпопуляции после разделения	35428	39959
Общий размер южной субпопуляции после разделения	119416	133272
Размер изолированной восточной субпопуляции после бутылочного горлышка	111	124
Размер изолированной южной субпопуляции после бутылочного горлышка	111	126
Время роста численности до разделения на две субпопуляции	100089	113518
Время разделения предковой популяции на субпопуляции	11084	12589
Время прохождения субпопуляций через бутылочное горлышко	11208	12728

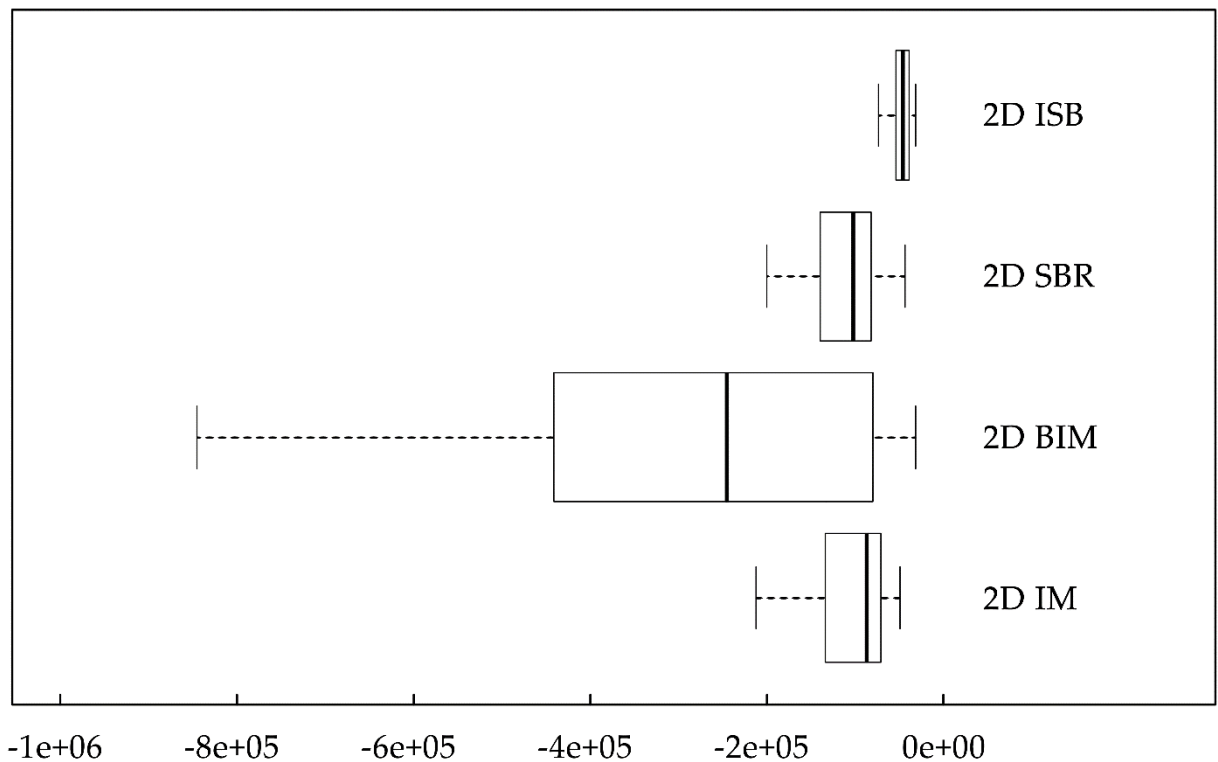


Рисунок 25. График с результатами бутстрепа для четырех демографических моделей. Мы выполнили в общей сложности 100 бутстреп репликаций, чтобы оценить дисперсию значений функции правдоподобия.

Мы использовали паттерны изменений геномных вариаций для моделирования и определения истории популяции гепардов из восточной и южной части Африки (из Танзании и Намибии, соответственно) при помощи модели диффузной аппроксимации и спектров частот аллелей (SFS). Подход, реализованный в программном обеспечении DaDi, сравнивает ожидаемую и наблюдаемую частоту аллеля в SFS, вычисляя совокупный показатель правдоподобия для наилучшего подбора нескольких различных, но правдоподобных эволюционных сценариев. Сценарии были смоделированы с использованием данных SFS, результаты были использованы для определения лучшей модели, которая наиболее правдоподобна с точки зрения наблюдаемых данных.

Модель 4 (также обозначаемая 2D ISB), двумерная (2D) модель, предполагающая расширяющуюся предковую популяцию, которая разделяется на две субпопуляции, показала наибольшую вероятность

соответствия, основываясь на низкой дисперсии и высоком значении максимального правдоподобия ($L = -43587$) (Рис. 25 и Табл. 19). Результаты моделирования DaDi предполагают наличие эффекта основателя у гепардов около 100 000 лет назад, возможно, это следствие их долгой миграционной истории, когда в позднем плейстоцене они из перешли из Северной Америки через Берингов пролив в Азию, а затем на юг в Африку, и в связи с регулярным постоянным сокращением численности, а также с ограничением потока генов (рис. 26 и табл. 20) [Johnson et al., 2006; Faurby et al., 2016; O'Brien et al., 2016].

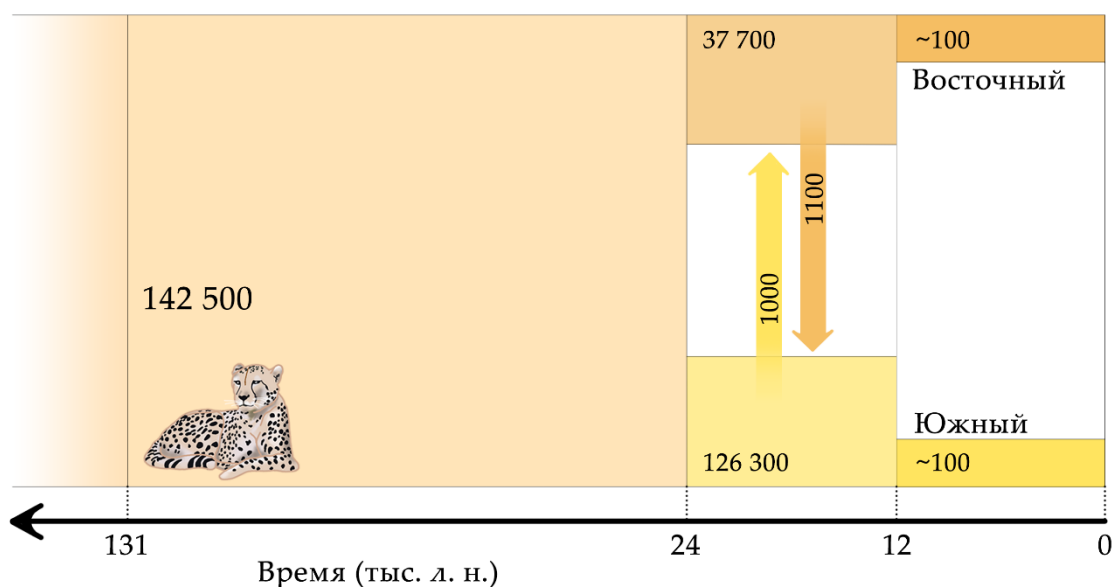


Рисунок 26. Анализ демографической истории африканского гепарда. Демографическая история двух популяций гепарда (юга Намибии и восточной части Танзании) на основе результатов DaDi. Четыре различные, но правдоподобные демографические модели были протестированы с помощью DaDi и SFS. Модель 4 (показана на этом рисунке) показала наилучшее соответствие данным, она включает в себя расчетные сроки увеличения численности популяции, эффекты бутылочного горлышка, миграцию после разделения на субпопуляции и эффективные размеры популяции с течением времени.

В качестве альтернативы Barnett et al. [Barnett et al., 2005] предложили, основываясь на результатах изучения древней ДНК *Miracinonyx trumani* (американских гепардов), что современные африканские гепарды происходят из Азии, что указывает на то, что эффект основателя 10000 лет назад совпадает с миграцией гепардов из Азии в Африку.

Прохождения южной и восточной популяций через бутылочное горлышко в позднем плейстоцене дополнительно истощает генетическое разнообразие в геноме африканских гепардов в обеих популяциях [Macdonald and Loveridge, 2010; O'Brien et al., 1985, O'Brien and Johnson, 2005]. Моделирование SFS показало заметный избыток определенных аллелей у восточной популяции по сравнению с южной, что подразумевает наличие у предковых популяций потока генов от Намибии до Танзании в интервале, оцененном в 11084 – 12589 лет назад (Табл. 20). Параллельный анализ попарно последовательной марковской коалесцентной модели (PSMC) для оценки демографической истории подтвердил вывод о снижении численности популяции гепарда за последние 100.000 лет (Рис. 25).

ЗАКЛЮЧЕНИЕ

Геномы африканских гепардов демонстрируют значительное уменьшение генетической изменчивости и следы необычной демографической истории. Семь различных показателей для оценки генетического разнообразия генома или отдельных его локусов демонстрируют наличие всего 1-10% уровня разнообразия, наблюдаемого у других диких млекопитающих, что значительно ниже, чем наблюдалось в исследованиях геномов инбредных собак и инбредных кошек, в генетически однородных Тасманских дьяволах, или геномах горилл с Вирунги [Xue et al., 2015]. Единственным исключением стало то, что у реликтовой популяции азиатских львов, которая встречается в национальном парке Гир-Форест в Гуджарате, Индия, плотность SNV была еще ниже. Эта популяция львов представляет собой отдельный подвид льва (*Panthera leo persica*) и настолько инбредна, что STR маркеры всех азиатских львов идентичны [Rangarajan, 2013].

Гепарды демонстрируют высокую аллотолерантность и способны не отторгать хирургически пересаженные трансплантаты кожи от других особей своего вида [O'Brien et al., 1983]. Это наблюдение положило начало первому глубокому исследованию почти всей области MHC гепарда. Сравнение сборки высокого разрешения для MHC домашней кошки со сборкой гепарда, выявило потерю 2-4 генов MHC класса I (*FLA-F*, *-H*, *-I* и *-M*), а также отсутствие вариабельности аминокислот в белках, кодирующихся генами I класса, среди всех семи геномов гепарда по сравнению со значительным разнообразием в регионе MHC у домашней кошки.

Анализ демографической истории, проведенный с использованием DaDi, модели PSMC, и полногеномной вариации SNV у двух популяций африканских гепардов, показал наличие двух событий, когда популяция проходила через “бутылочное горлышко”: ~ 100 и ~ 12 тыс.л.н. Предыдущие анализы с использованием RFLP и последовательностей мтДНК и микросателлитных локусов также предполагали недавнее бутылочное

горлышко 10-12 тыс.л.н., совпадающее с поздне-плейстоценовым вымиранием преимущественно крупных животных в Северной Америке и некоторых частях Евразии: мамонтов, мастодонтов, ужасных волков, короткомордых медведей, американских львов, саблезубых тигров и четырех видов плотоядных птиц [Werdelin et al., 2010; Hewitt, 2010; Driscoll et al., 2002]. Пума и гепарды также исчезли из Северной Америки в это время [Culver, 2000; O'Brien and Johnson, 2005]. Мы предполагаем, что два поздне-плейстоценовых сокращения численности популяции сильно уменьшили генетическое разнообразие предков гепарда, и оставили следы демографического сокращения в их геномах. Во-первых, ~ 100.000 лет назад миграция гепардов по всей Азии и их распространение в Африку, могло сопровождаться увеличением близкородственного скрещивания в результате заселения территорий ("Эффект основателя"). Более позднее "бутылочное горлышко" африканских популяций гепардов ~12 т.л.н. еще больше сократило численность и привело к дополнительной потере изменчивости генома, наблюдаемой у современных гепардов [Faurby et al., 2016; O'Brien et al., 2016].

Геномный анализ белок-кодирующих генов и семейств генов, критических для размножения, продемонстрировал убедительные статистические данные, подтверждающие накопление избытка несинонимичных замен у гепардов по сравнению с другими кошачьими. Мы идентифицировали десять фиксированных аминокислотных вариантов в локусе гена *AKAP4*, который экспрессируется исключительно в семенниках и формирует важный компонент жгутика сперматозоида. Гомологи *AKAP4* играют критическую роль в развитии сперматозоидов и возникновении аномалий сперматозоидов у нескольких видов млекопитающих. Пять гомозиготных вредных мутаций в кодирующей части *AKAP4*, вероятно, объясняют повышенную тератозооспермию (в среднем 81,6% поврежденных сперматозоидов), обнаруженную у гепардов-самцов.

Некоторые гены, которые опосредуют мышечную активность и энергетический обмен, обнаружили сигналы позитивного отбора и являются

кандидатами для объяснения адаптации гепарда к высокоскоростному преследованию, они заслуживают дополнительного изучения в будущем.

В целом изучение генома гепарда предлагает беспрецедентные данные и позволяет выявить генетические особенности, необходимые для понимания истории, адаптации и выживания видов, находящихся под угрозой исчезновения. Несколько десятилетий назад, в поиске объекта исследования из числа видов, находящихся под угрозой вымирания, выбор ученых пал на гепардов, содержащихся в неволе, что привело к тому, что сейчас измерение уровня генетического разнообразия и его коррелятов включено во все программы по контролю и управлению популяциями, находящимися под угрозой исчезновения. Знание уникальной истории гепардов и их генетических особенностей должно стать полезным для поддержания и увеличения численности гепардов в их нынешних и бывших местах обитания.

ВЫВОДЫ

1) За последние сто тысяч лет в результате по крайней мере двух резких сокращений эффективной численности популяции *Acinonyx jubatus* была потеряна существенная часть генетического разнообразия, которое является характерной чертой для дико живущих представителей семейства *Felidae* и других млекопитающих.

2) Высокое генетическое сходство и низкое разнообразие аллелей генов МНС классов I и II, а также потеря нескольких генов МНС класса I дают наиболее вероятное объяснение двум ранее наблюдаемым фенотипическим свойствам у *Acinonyx jubatus*: а) Приживаемость кожных трансплантатов между неродственными особями и б) Высокая смертность от коронавирусной инфекции.

3) Данные анализа спектров частот аллелей и гаплотипов между двумя популяциями Африканских гепардов позволили с высокой точностью реконструировать демографическую историю вида вплоть до событий в районе 100 тысяч лет назад. Были получены данные, подтверждающие и уточняющие ранее высказанные гипотезы о прохождении предковых популяций через бутылочное горлышко в районе 12 тысяч лет назад.

4) Анализ мутационного груза в группе генов, связанных с размножением, обнаружил избыточное количество потенциально вредных мутаций относительно генома домашней кошки и других млекопитающих, а также выявил вероятную причину высокой доли аномальных сперматозоидов у Африканских гепардов в виде гена *AKAP4*, накопившем ряд серьезных мутаций.

5) Следы положительного отбора, экспансия некоторых генных семейств, а также структурные вариации в генах, отвечающих за регуляцию сокращения сердечной и поперечнополосатой мускулатуры, тепловой стресс и регуляцию катаболических процессов, демонстрируют генетические основы уникальных адаптаций *Acinonyx jubatus* к хищническому образу жизни и стилю охоты, который выделяет их на фоне не только группы Felidae, но и *Carnivora* в целом.

Список сокращений

тыс.л.н. - тысяч лет назад

млн.л.н. - миллион лет назад

п.н. - пар нуклеотидов

т.п.н. - тысяч пар нуклеотидов

млн.т.п. - миллионов пар оснований

млрд.п.н. - миллиардов пар нуклеотидов

ДНК - дезоксирибонуклеиновая кислота

BAC - bacterial artificial chromosome

BGI - Beijing Genome Institute

BLAST - basic local alignment search tool

CDS - coding sequences

CR - critically endangered

CNV - copy number variation

EN - Endangered

EST - expressed sequence tag

FISH - fluorescent in situ hybridization

MHC - Major Histocompatibility complex

ROH - Runs of Homozygosity

SFS - site frequency spectrum

SNV - Single nucleotide polymorphism

STR - short tandem repeat

SD - Segmental duplication

TE - transposable element

VU - Vulnerable

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Adams D.B. The cheetah: native American / D.B. Adams // *Science*. – 1979. – Т.205. – №.4411. – С.1155-1158.
2. Adzhubei I. Predicting functional effect of human missense mutations using PolyPhen-2 / I. Adzhubei, D.M. Jordan, S.R. Sunyaev // *Current protocols in human genetics*. – 2013. – Т.76. – №.1. – С.7-20.
3. Alkan C. Personalized copy number and segmental duplication maps using next-generation sequencing / C. Alkan, J.M. Kidd, T. Marques-Bonet et al. // *Nature genetics*. – 2009. – Т.41. – №.10. – С.1061
4. Altmann A. A beginner's guide to SNP calling from high-throughput DNA-sequencing data / A. Altmann, P. Weber, D. Bader et al. // *Human genetics*. – 2012. – Т.131. – №.10. – С.1541-1554.
5. Armstrong J. Progressive Cactus is a multiple-genome aligner for the thousand-genome era // J. Armstrong, G. Hickey, M. Diekhans et al. // *Nature*. – 2020. – Т.587. – №.7833. – С.246-251.
6. Barnett R. Evolution of the extinct Sabretooths and the American cheetah-like cat / R. Barnett, I. Barnes, M.J. Philips et al. // *Current Biology*. – 2005. – Т.15. – №.15. – С.89-90.
7. Bao W. Repbase Update, a database of repetitive elements in eukaryotic genomes / W. Bao, K.K. Kojima, O. Kohany // *Mobile Dna*. – 2015. – Т.6. – №.1. – С.1-6.
8. Benson G. Tandem repeats finder: a program to analyze DNA sequences / G. Benson // *Nucleic acids research*. – 1999. – Т.27. – №.2. – С.573-580.
9. Cagliani R. Pathogen-driven selection in the human genome / R. Cagliani, M. Sironi // *International journal of evolutionary biology*. – 2013.

10. Camacho C. BLAST+: architecture and applications / C. Camacho, G. Coulouris, V. Avagyan et al. // BMC bioinformatics. – 2009. – T.10. – №.1. – C.1-9.
11. Caro T.M. Cheetahs of the Serengeti Plains: group living in an asocial species / T.M. Caro // University of Chicago Press. – 1994.
12. Caro T.M. Ecological and genetic factors in conservation: a cautionary tale / T.M. Caro, M.K. Laurenson // Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information. – 1994. – T.263. – №.5146. – C.485-486.
13. Castro-Prieto A. Cheetah paradigm revisited: MHC diversity in the world's largest free-ranging population / A. Castro-Prieto, B. Wachter, S. Sommer // Molecular biology and evolution. – 2011. – T.28. – №.4. – C.1455-1468.
14. Chen N. Using Repeat Masker to identify repetitive elements in genomic sequences / N. Chen // Current protocols in bioinformatics. – 2004. – T.5. – №.1. – C.4-10.
15. Cherin M. *Acinonyx pardinensis* (Carnivora, Felidae) from the Early Pleistocene of Pantalla (Italy): predatory behavior and ecological role of the giant Plio–Pleistocene cheetah / M. Cherin, D.A. Iurino, R. Sardella, L. Rook // Quaternary Science Reviews. – 2014. – T.87. – C.82-97.
16. Cho Y.S. The tiger genome and comparative analysis with lion and snow leopard genomes / Y.S. Cho, L. Hu, H. Hou et al. // Nature communications. – 2013. – T.4. – №.1. – C.1-7.
17. Cingolani P. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3 / P. Cingolani, A. Platts, L.L. Wang // Fly. – 2012. – T.6. – T.2. – C.80-92.
18. Coombe L. Assembly of the complete Sitka spruce chloroplast genome using 10X Genomics' GemCode sequencing data / L. Coombe, R.L.

- Warren, S.D. Jackman et al. // PloS one. – 2016. – T.11. – №.9. – C. e0163059.
19. Crosier A.E. Ejaculate traits in the Namibian cheetah (*Acinonyx jubatus*): influence of age, season and captivity / A.E. Crosier, L. Marker, J. Howard J // *Reproduction, Fertility and Development*. – 2007. – T.19. – №.2. – C.370-382.
 20. Culver M. Genomic ancestry of the American puma (*Puma concolor*) / M. Culver, W.E. Johnson, J. Pecon-Slaterry, S.J. O'Brien // *Journal of Heredity*. – 2000. – T.91. – №.3. – C.186-197.
 21. Cunningham F. Ensembl 2015 / F. Cunningham, M.R. Amode, D. Barrell et al. // *Nucleic acids research*. – 2015. – T.43. – №.1. – C.662-669.
 22. Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life* / C. Darwin // 1859.
 23. Davis B.W. A high-resolution cat radiation hybrid and integrated FISH mapping resource for phylogenomic studies across Felidae / B.W. Davis, T. Raudsepp, A.J. Wilkerson et al. // *Genomics*. – 2009. – T.93. – №.4. – C.299-304.
 24. Dobrynin P. Genomic legacy of the African cheetah, *Acinonyx jubatus* / P. Dobrynin, S. Liu, G. Tamazian et al. // *Genome biology*. – 2015. – T.16. – №.1. – C.1-20.
 25. Driscoll C.A. Genomic microsatellites as evolutionary chronometers: a test in wild cats / C.A. Driscoll, M. Menotti-Raymond, G. Nelson et al. // *Genome research*. – 2002. – T.12. – №.3. – C.414-423.
 26. Durant S.M. Living with the enemy: avoidance of hyenas and lions by cheetahs in the Serengeti / S.M. Durant // *Behavioral ecology*. – 2000. – T.11. – №.6. – C.624-632.
 27. Durant S.M. The global decline of cheetah *Acinonyx jubatus* and what it means for conservation / S.M. Durant, N. Mitchell, R. Groom et al. // *Proceedings of the National Academy of Sciences*. – 2017. – T.114. – №.3. – C.528-533.

28. Ewing B. Base-calling of automated sequencer traces using phred. II. Error probabilities / B. Ewing B, P. Green // *Genome research*. – 1998. – T.8. – №.3. – C.186-194.
29. Faurby S. The difference between trivial and scientific names: There were never any true cheetahs in North America / S. Faurby, L. Werdelin, J.C. Svenning // *Genome Biology*. – 2016. – T.17. – №.1. – C.89.
30. Finnegan D.J. Transposable elements / D.J. Finnegan // *Current opinion in genetics & development*. – 1992. – T.2. – №.6. – C.861-867.
31. Finn R.D. HMMER web server: Interactive sequence similarity searching / R.D. Finn, J. Clements, S.R. Eddy // *Nucleic acids research*. – 2011. – T.39. – №.2. – C.29-37.
32. Fitzpatrick J.L. Reduced heterozygosity impairs sperm quality in endangered mammals / J.L. Fitzpatrick, J.P. Evans // *Biology letters*. – 2009. – T.5. – №.3. – C.320-323.
33. Freeman A.R. Sequence variation in the mitochondrial DNA control region of wild African cheetahs (*Acinonyx jubatus*) / A.R. Freeman, D.E. Machugh, S. Mckeown et al. // *Heredity*. – 2001. – T.86. – №.3. – C.355-362.
34. Gattepaille L.M. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks / L.M. Gattepaille, M. Jakobsson, M.G. Blum // *Heredity*. – 2013. – T.110. – №.5. – C.409-419.
35. Goto M. Distribution of muscle fibers in skeletal muscles of the cheetah (*Acinonyx jubatus*) / M. Goto, M. Kawai, M. Nakata et al. // *Mammalian biology*. – 2013. – T.78. – №.2. – C.127-133.
36. Griffiths-Jones S. Rfam: an RNA family database / S. Griffiths-Jones, A. Bateman, M. Marshall et al. // *Nucleic acids research*. – 2003. – T.31. – №.1. – C.439-441.
37. Gurevich A. QUAST: quality assessment tool for genome assemblies / A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler // *Bioinformatics*. – 2013. – T.29. – №.8. – C.1072-1075.

38. Hahn M. Accelerated rate of gene gain and loss in primates / M.W. Hahn, J.P. Demuth, S.G. Han // *Genetics*. – 2007. – T.177. – №.3. – C.1941-1949.
39. Hamosh A. Online Mendelian inheritance in man (OMIM) / A. Hamosh, A.F. Scott, J. Amberger et al. // *Human mutation*. – 2000. – T.15. – №.1. – C.57-61.
40. Heeney J.L. Prevalence and implications of feline coronavirus infections of captive and free-ranging cheetahs (*Acinonyx jubatus*) / J.L. Heeney, J.F. Evermann, A.J. McKeirnan et al. // *Journal of Virology*. – 1990. – T.64. – №.5. – C.1964-1972.
41. Hemmer H. Cheetahs in the Middle Pleistocene of Europe: *Acinonyx pardinensis* (sensu lato) *intermedius* (Thenius, 1954) from the Mosbach Sands (Wiesbaden, Hesse, Germany) / Hemmer H., Kahlke R.D. // *Neues Jahrbuch für Geologie und Paläontologie – Abhandlungen*. – 2008. – T.249. – №.3. – C.345-356.
42. Hewitt G. The genetic legacy of the Quaternary ice ages / G. Hewitt // *Nature*. – 2000. – T.405. – №.6789. – C.907-913.
43. Hsu T.C. Karyological studies of nine species of Felidae / T.C. Hsu, H. Rearden, G. Luquette // *The American Naturalist*, - 1963. – T.97. – №.895. – C. 225-234.
44. Hudson P.E. Functional anatomy of the cheetah (*Acinonyx jubatus*) hindlimb / P.E. Hudson, S.A. Corr, R.C. Payne-Davis et al. // *Journal of anatomy*. – 2011. – T.218. – №.4. – C.363-374.
45. Hurles M. Gene duplication: the genomic trade in spare parts / M. Hurles // *PLoS Biology*. – 2004. – T.2. – №.7. – C.e206.
46. Hyatt J.P. Myosin heavy chain composition of tiger (*Panthera tigris*) and cheetah (*Acinonyx jubatus*) hindlimb muscles / J.P. Hyatt, R.R. Roy, S. Rugg, R.J. Talmadge // *Ecological Genetics and Physiology*. – 2010. – T.313. – №.1. – C.45-57.

47. Johnson W.E. The late Miocene radiation of modern Felidae: a genetic assessment / W.E. Johnson, E. Eizirik, J. Pecon-Slattery et al. // Science. – 2006. – T.311 – T.5757. – C.73-77.
48. Jurka J. Repbase Update, a database of eukaryotic repetitive elements / J. Jurka, V.V. Kapitonov, A. Pavlicek et al. // Cytogenetic and genome research. – 2005. – T.110. – №.1-4. – C.462-467.
49. Sequencing methods review of publications featuring illumina® technology. 2014 // Illumina [сайт] URL: http://www.illumina.com/content/dam/illumine-marketing/documents/products/research_reviews/sequencing-methods-review (дата обращения 25.07.2014)
50. Kanehisa M. KEGG for linking genomes to life and the environment / M. Kanehisa, M. Araki, S. Goto et al. // Nucleic acids research. – 2007. – T.36. – C.480-484.
51. Kent W.J. BLAT—the BLAST-like alignment tool / W.J. Kent // Genome research. – 2002. – T.12. – №.4. – C.656-664.
52. Keverne E.B. The vomeronasal organ / E.B. Keverne // Science. – 1999. – T.286. – №.5440. – C.716-720.
53. Kim S. Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly / S. Kim, Y.S. Cho, H.M. Kim et al. // Genome biology. – 2016. – T.17. – №.1. – C.1-2.
54. Korbel J.O. Genome assembly and haplotyping with Hi-C / J.O. Korbel, C. Lee // Nature biotechnology. – 2013. – T.31. – №.12. – C.1099-1101.
55. Korf I. Gene finding in novel genomes / I. Korf // BMC bioinformatics. – 2004. – T.5. – №.1. – C.1-9.
56. Korneliussen T.S. ANGSD: Analysis of Next Generation Sequencing Data / T.S. Korneliussen, A. Albrechtsen, R. Nielsen // BMC bioinformatics. – 2014. – T.15. – №.1. – C.1-3.

57. Kozomara A. miRBase: from microRNA sequences to function / A. Kozomara, M. Birgaoanu, S. Griffiths-Jones // *Nucleic acids research*. – 2019. – T.47. – №.1. – C.155-162.
58. Krausman P.R. *Acinonyx jubatus* / P.R. Krausman, S.M. Morales // *Mammalian Species*. – 2005. – T.771. – C.1-6.
59. Langmead B. Searching for SNPs with cloud computing / B. Langmead, M.C. Schatz, J. Lin et al. // *Genome biology*. – 2009. – T.10. – №.11. – C.1-10.
60. Laurenson M.K. High juvenile mortality in cheetahs (*Acinonyx jubatus*) and its consequences for maternal care / M.K. Laurenson // *Journal of Zoology*. – 1994. – T.234. – №.3. – C.387-408.
61. Lechner M. Proteinortho: detection of (co-) orthologs in large-scale analysis / M. Lechner, S. Findeiß, L. Steiner et al. // *BMC bioinformatics*. – 2011. – T.12. – №.1. – C.1-9.
62. Lewin H.A. Every genome sequence needs a good map / H.A. Lewin, D.M. Larkin, J. Pontius, S.J. O'Brien // *Genome research*. – 2009. – T.19. – №.11. – C.1925-1928.
63. Li H. TreeFam: a curated database of phylogenetic trees of animal gene families / H. Li, A. Coghlan, J. Ruan et al. // *Nucleic acids research*. – 2006. – T.34. – T.1. – C.572-580.
64. Li H. The sequence alignment/map format and SAMtools / H. Li, B. Handsaker, A. Wysoker // *Bioinformatics*. – 2009. – T.25. – №.16. – C.2078-2079.
65. Li H. Inference of human population history from individual whole-genome sequences / H. Li, R. Durbin // *Nature*. – 2011. – T.475. – №.7357. – C.493-496.
66. Li M. Towards a more accurate error model for BioNano optical maps / M. Li, A.C. Mak, E.T. Lam et al. // *International Symposium on Bioinformatics Research and Applications*. – 2016. – C.67-79.

67. Li R. The sequence and de novo assembly of the giant panda genome / R. Li, W. Fan, G. Tian et al. // *Nature*. – 2010. – T.463. – T.7279. – C.311-317.
68. Lowe T.M. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence / T.M. Lowe, S.R. Eddy // *Nucleic acids research*. – 1997. – T.25. – №.5. – C.955-964.
69. Löytynoja A. Phylogeny-aware alignment with PRANK. In *Multiple sequence alignment methods* / A. Löytynoja // Humana Press. – 2014. – C.155-170.
70. Luo H. Phylogenetic analysis of genome rearrangements among five mammalian orders / H. Luo H, W. Arndt, Y. Zhang et al. // *Molecular phylogenetics and evolution*. – 2012. – T.65. – №.3. – C.871-882.
71. Luo R. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler / R. Luo, B. Liu, Y. Xie et al. // *Gigascience*. – 2012. – T.1. – T.1. – C.2047-217X.
72. Macdonald D. The biology and conservation of wild felids / D. Macdonald, A. Loveridge // Oxford University press. – 2010.
73. Magadum S. Gene duplication as a major force in evolution / S. Magadum, U. Banerjee, P. Murugan, D. Gangapur, R. Ravikesavan // *Journal of genetics*. – 2013. – T.92. – №.1. – C.155-161.
74. Marcais G. Jellyfish: A fast k-mer counter / G. Marcais, C. Kingsford // *Tutorialis e Manuais*. – 2012. – T.29. – №.1. – C.1-8.
75. Marker L.L. Demography of the Namibian cheetah, *Acinonyx jubatus jubatus* / L.L. Marker, A.J. Dickman, R.M. et al. // *Biological Conservation*. – 2003. – T.114. – №.3. – C.413-425.
76. Marker L. Overview of the global cheetah population / L. Marker // *In Animal Keeper's Forum*. – 2005. – T.7. – №.8. – C.284-288.
77. Mavrich T.N. Nucleosome organization in the *Drosophila* genome / T.N. Mavrich, C. Jiang, I.P. Ioshikhes et al. // *Nature*. – 2008. – T.453. – №.7193. – C.358-362.

78. May R.M. The cheetah controversy / R.M. May // *Nature*. – 1995. – T.374. – №.6520. – C.309-310.
79. McCoy R.C. Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population / R.C. McCoy, N.R. Garud, J.L. Kelley et al. // *Molecular Ecology*. – 2014. – T.23. – №.1. – C.136-150.
80. McPherson J.D. A physical map of the human genome / J.D. McPherson, M. Marra, L.D. Hillier et al. // *Nature*. – 2001. – T.409. – №.6822. – C.934-941.
81. Menotti-Raymond M. An autosomal genetic linkage map of the domestic cat, *Felis silvestris catus* / M. Menotti-Raymond, V.A. David, A.A. Schäffer et al. // *Genomics*. – 2009. – T.93. – №.4. – C.305-313.
82. Miki K. Targeted disruption of the *Akap4* gene causes defects in sperm flagellum and motility / K. Miki, W.D. Willis, P.R. Brown // *Developmental biology*. – 2002. – T.248. – №.2. – C.331-342.
83. Misale C. Accelerating Bowtie2 with a lock-less concurrency approach and memory affinity / C. Misale // 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing. – 2014. – C.578-585.
84. Mitchell A. The InterPro protein families database: the classification resource after 15 years / A. Mitchell, H.Y. Chang, L. Daugherty et al. // *Nucleic acids research*. – 2015. – T.43. – №.1. – C.213-221.
85. Montague M.J. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication / M.J. Montague, G. Li, B. Gandolfi, R. Khan et al. // *Proceedings of the National Academy of Sciences*. – 2014. – T.111. – №.48. – C.17230-27235.
86. Murchison E.P. Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer / E.P. Murchison, O.B. Schulz-Trieglaff, Z. Ning et al. // *Cell*. – 2012. – T.148. – №.4. – C.780-791.

87. Nadachowska-Brzyska K. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers / K. Nadachowska-Brzyska, R. Burri, L. Smeds, H. Ellegren // Molecular Ecology. – 2016. – T.25. – №.5. – C.1058-1072.
88. Nawrocki E.P. Infernal 1.0: inference of RNA alignments / E.P. Nawrocki, D.L. Kolbe, S.R. Eddy // Bioinformatics. – 2009. – T.25. – №.10 – C.1335-1337.
89. Nielsen R. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data / R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, J. Wang // PLoS one. – 2012. – T.7. – №.7. – C.e37558.
90. Noskova E. GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data / E. Noskova, V. Ulyantsev, K.P. Koepfli et al. // GigaScience. – 2020. – T.9. – №.3. – C. giaa005
91. O'Brien S.J. The cheetah is depauperate in genetic variation / S.J. O'Brien, D.E. Wildt, D. Goldman et al. // Science. – 1983. – T.221. – T.4609. – C.459-462.
92. O'Brien S.J. Genetic basis for species vulnerability in the cheetah / S.J. O'Brien, M.E. Roelke, L. Marker et al. // Science. – 1985. – T.227. – №.4693. – C.1428-1434.
93. O'Brien S.J. Biochemical genetic variation in geographic isolates of African and Asiatic lions / S.J. O'Brien // National Geographic Research. – 1987. – T.3. - №.1. – C.114.
94. O'Brien S.J. Big cat genomics / S.J. O'Brien, W.E. Johnson // Annu. Rev. Genomics Hum. Genet. – 2005. – T.6. – C.407-429.
95. O'Brien S.J. Evolution: a new cat species emerges / S.J. O'Brien, K.P. Koepfli // Current Biology. – 2013. – T.23. – №.24. – C.1103-1105.
96. O'Brien S.J. Response to Comment by Faurby / S.J. O'Brien, K.P. Koepfli, E. Eizirik et al. / Genome Biology. – 2016. – T.17. – №.1. – C.90.

97. Ogata H. KEGG: Kyoto encyclopedia of genes and genomes / H. Ogata, S. Goto, K. Sato et al. // *Nucleic acids research*. – 1999. – T.27. – №.1. – C.29-34.
98. Ohno S. *Evolution by gene duplication* / S. Ohno // Springer Science & Business Media. - 1970.
99. Parisi V. STRING: finding tandem repeats in DNA sequences / V. Parisi, V. De Fonzo, F. Aluffi-Pentini // *Bioinformatics*. – 2003. – T.19. – №.14. – C.1733-1738.
100. Parker H.G. Man's best friend becomes biology's best in show: genome analyses in the domestic dog / H.G. Parker, A.L. Shearin, E.A. Ostrander // *Annual review of genetics*. – 2010. – T.44 – C.309-336.
101. Parra G. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes / G. Parra, K. Bradnam, I Korf // *Bioinformatics*. – 2007. – T.23. – №.9. – C.1061-1067.
102. Paten B. Cactus: Algorithms for genome multiple sequence alignment / B. Paten, D. Earl, N. Nguyen et al. // *Genome research*. – 2011. – T.21. – №.9. – C.1512-1528.
103. Pontius J.U. Initial sequence and comparative analysis of the cat genome / J.U. Pontius, J.C. Mullikin, D.R. Smith et al. // *Genome research*. – 2007. – T.17. – №.11. – C.1675-1689.
104. Pruitt K.D. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins / K.D. Pruitt, T. Tatusova, D.R. Maglott // *Nucleic acids research*. – 2007. – T.35. – №.1. – C.61-65.
105. Ralls K. Effect of inbreeding on juvenile mortality in some small mammal species / K. Ralls, J. Ballou // *Laboratory Animals*. – 1982. – T.16. – №.2. – C.159-166.
106. Ralls K. Estimates of lethal equivalents and the cost of inbreeding in mammals / K. Ralls, J.D. Ballou, A. Templeton // *Conservation biology*. – 1988. – T.2. – №.2. – C.185-193.

107. Rangarajan M. Animals with rich histories: the case of the lions of Gir Forest, Gujarat, India / M. Rangarajan // History and Theory. – 2013. – T.52. – №.4. – C.109-127.
108. Rappaport N. MalaCards: an integrated compendium for diseases and their annotation / N. Rappaport, N. Nativ, G. Stelzer et al. // Database. – 2013.
109. Rogers J. Comparative primate genomics: emerging patterns of genome content and dynamics / J. Rogers, R.A. Gibbs // Nature Reviews Genetics. – 2014. – T.15. – №.5. – C.347-359.
110. Safran M. GeneCards Version 3: the human gene integrator / M. Safran, I. Dalah, J. Alexander et al. // Database. – 2010.
111. Schreiber F. TreeFam v9: a new website, more species and orthology-on-the-fly / F. Schreiber, M. Patricio, M. Muffato et al. // Nucleic acids research. – 2014. – T.42. – №.1. – C.922-925.
112. Schwartz S. Human – Mouse Alignments with BLASTZ / S. Schwartz, W.J. Kent, A. Smit // Genome Research. – 2003. – T.13. – №.1. – C.103 – 107.
113. Simão F.A. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs / F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov // Bioinformatics. – 2015. – T.31. – №.19. – C.3210-3212.
114. Sims D. Sequencing depth and coverage: key considerations in genomic analyses / D. Sims, I. Sudbery, N.E. Illott et al. // Nature Reviews Genetics. – 2014. – T.15. – №.2. – C.121-132.
115. Scantlebury D.M. Flexible energetics of cheetah hunting strategies provide resistance against kleptoparasitism / D.M. Scantlebury, M.G. Mills, R.P. Wilson et al. // Science. – 2014. – T.346. – №.6205. – C.79-81.
116. Stanke M. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome / M.

- Stanke, A. Tzvetkova, B. Morgenstern // *Genome biology*. – 2006. – T.7. – №.1. – C.1-8.
117. Tamazian G. Annotated features of domestic cat–*Felis catus* genome / G. Tamazian, S. Simonov, P. Dobrynin // *Gigascience*. – 2014. – T.3. – №.1. – C.2047-17X.
118. Tamazian G. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences // G. Tamazian, P. Dobrynin, K. Krasheninnikova et al. // *GigaScience*. – 2016. – T.5. – №.1. – C.13742-137016.
119. Tesler G. GRIMM: genome rearrangements web server / G. Tesler // *Bioinformatics*. – 2002. – T.18. – №.3. – C.492-493.
120. Thompson J.D. Multiple sequence alignment using ClustalW and ClustalX / J.D. Thompson, T.J. Gibson, D.G. Higgins // *Current protocols in bioinformatics*. – 2003. – T.1. – C.2-3.
121. Van Valkenburgh B. The plio-pleistocene cheetah-like cat *miracinonyx inexpectatus* of North America / B. Van Valkenburgh, F. Grady, B. Kurtén // *Journal of Vertebrate Paleontology*. – 1990. – T.10. – №.4. – C.434-454.
122. Van Valkenburgh B. The cheetah: evolutionary history and paleoecology / B. Van Valkenburgh, B. Pang, M. Cherin, L. Rook // 2017.
123. Wayne R.K. Morphological variability and asymmetry in the cheetah (*Acinonyx jubatus*), a genetically uniform species / R.K. Wayne, W.S. Modi, S.J. O'Brien // *Evolution*. – 1986. – T.40. – №.1. – C.78-85.
124. Werdelin L. Phylogeny and evolution of cats (Felidae) / L. Werdelin, N. Yamaguchi, W.E. Johnson, S.J. O'Brien // *Biology and conservation of wild felids*. – 2010. – C.59-82.
125. Wilkerson A.J. Coronavirus outbreak in cheetahs: lessons for SARS / A.J. Wilkerson, E.C. Teeling, J.L. Troyer et al. // *Current Biology*. – 2004. – T.14. – №.6. – C.227-228.

126. Wilson, J.W. Cheetahs, *Acinonyx jubatus*, balance turn capacity with pace when chasing prey / J.W. Wilson, M.G. Mills, R.P. Wilson et. al.// *Biology letters*. – 2013. – T.9. - №.5. – C. 20130620.
127. Wheeler D.L. Database resources of the national center for biotechnology information / D.L. Wheeler, T. Barrett, D.A. Benson et al. // *Nucleic acids research*. – 2007. – T.36. – C.13-21.
128. Xue Y. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding / Y. Xue, J. Prado-Martinez, P.H. Sudmant et al. // *Science*. – 2015. – T.348. – №.6231. – C.242-245.
129. Yandell M. A beginner's guide to eukaryotic genome annotation / M. Yandell, D. Ence // *Nature Reviews Genetics*. – 2012. – T.13. – №.5. – C.329-342.
130. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood / Z. Yang // *Bioinformatics*. – 1997. – T.13. – №.5. – C.555-556.
131. Yang Z. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages / Z. Yang, R. Nielsen // *Molecular biology and evolution*. – 2002. – T.19. – №.6. – C.908-917.
132. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood / Z. Yang // *Molecular biology and evolution*. – 2007. – T.24. – №.8. – C.1586-1591.
133. Yates A.D. Ensembl / A.D. Yates, P. Achuthan, W. Akanni et al. // *Nucleic acids research*. – 2020. – T.48. – C.682-688.
134. Yuhki N. Characterization of MHC cDNA clones in the domestic cat. Diversity and evolution of class I genes / N. Yuhki, G.F. Heidecker, S.J. O'Brien // *The Journal of Immunology*. – 1989. – T.142. – №.10. – C.3676-3682.
135. Yuhki N. DNA variation of the mammalian major histocompatibility complex reflects genomic diversity and population history / N. Yuhki, S.J.

O'Brien // Proceedings of the National Academy of Sciences. – 1990. –
T.87. – T.2. – C.836-838

