

На правах рукописи



Тереханова Надежда Владимировна

Неравномерность мутагенеза и отбора в геноме позвоночных

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2018

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель: **Базыкин Георгий Александрович**
кандидат биологических наук
Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

Официальные оппоненты: **Наймарк Елена Борисовна**
доктор биологических наук
Федеральное государственное бюджетное учреждение науки
Палеонтологический институт им. А. А. Борисяка Российской
академии наук

Раменский Василий Евгеньевич
кандидат физико-математических наук
Федеральное государственное автономное образовательное
учреждение высшего образования «Московский физико-
технический институт (государственный университет)»

Ведущая организация: Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова Российской
академии наук

Защита диссертации состоится ____ _____ 2019 года в ____ на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН), а также на сайте ИППИ РАН по адресу <http://iitp.ru/ru/dissertation/>

Автореферат разослан «__» _____ 2018 г.

Ученый секретарь диссертационного совета

доктор биологических наук, профессор

Рожкова Г. И.

Общая характеристика работы

Актуальность работы. Мутагенез и естественный отбор являются основными факторами эволюции. С развитием технологий секвенирования появилось большое количество геномных данных для разных видов и некоторых популяций, что открывает новые возможности для изучения мутагенеза и отбора на уровне отдельных сегментов генома.

Нуклеотидная последовательность ДНК изменяется в ряду поколений в результате мутагенеза. Мутации бывают разных типов и могут быть вызваны действием различных факторов. Поэтому важно понимать вклады различных процессов, которые участвуют в возникновении мутаций, а также учитывать неравномерность действия этих факторов в разных сегментах последовательности ДНК. В результате действия отбора мутации могут быть в разной степени вредными или полезными, или же они могут быть нейтральными. Данное свойство влияет на динамику частоты мутации внутри популяции, и видимые изменения ДНК являются результатом совместного действия мутагенеза и отбора.

Естественный отбор является процессом, который отсеивает изменения ДНК, ухудшающие приспособленность организма, и благоприятствует закреплению в популяции полезных вариантов, улучшающих приспособленность. При изменении условий окружающей среды меняется действие отбора, что может привести к тому, что определённые мутации могут стать более полезными в новой среде. Такие мутации могут либо уже присутствовать в популяции, либо возникать *de novo*. Понятно также, что мутации, обеспечивающие адаптацию, будут оказывать влияние на гены, изменение функций которых является важным при переходе в новую среду. На сегодняшний день известно, что в результате адаптации популяции происходит накопление различий между геномами популяций. Было показано, что в результате действия дизруптивного отбора в геномах образуются участки с повышенным количеством различий между популяциями ещё до момента полной репродуктивной изоляции видов.

Биоинформатический анализ нуклеотидных последовательностей молекул, полученных в результате секвенирования, позволяет делать выводы о закономерностях их структуры и эволюции. На сегодняшний день известно, что геном имеет организованную многоуровневую структуру, обеспечивающую компактное хранение большого количества ДНК в ядре. Накопление мутаций связано с повреждением ДНК, а также с молекулярными процессами репликации, репарации, рекомбинации и транскрипции, при протекании которых возникают ошибки. Факторы, вовлечённые в эти процессы, имеют свои особенности, что приводит к неравномерному накоплению мутаций в разных сегментах генома. В результате большого количества исследований, в том числе таких больших проектов, как ENCODE и RoadMap,

построено множество карт, описывающих расположение генов, структурных элементов хроматина и другие свойства сегментов ДНК.

На теоретическом уровне понимание различий скоростей мутирования и отбора между сегментами генома необходимо для построения более точных эволюционных моделей и лучшего понимания процесса видообразования, в том числе для нашего собственного вида. Помимо теоретической значимости, изучение неравномерности мутагенеза между сегментами генома имеет прикладное значение, например, в медицинской генетике. Поиск генов, ответственных за генетические заболевания, и генов-драйверов при мутагенезе требует построения правильной нуль-модели, учитывающей неравномерность скорости накопления соответственно мутаций в зародышевой линии и соматических мутаций.

Цели и задачи исследования. Целью исследования было изучение закономерностей различий скоростей мутирования и действия отбора в разных сегментах генома, а также описание факторов, которые оказывают влияние на эти процессы. Были поставлены следующие задачи:

1. Изучить закономерности изменения локальной скорости мутирования в ходе эволюции нескольких видов приматов и грызунов, а также определить причины этих изменений.

2. Исследовать структуру и свойства участков повышенной межпопуляционной изменчивости («островов дивергенции»), которые возникают вследствие действия отбора, на примере нескольких независимых популяций трёхиглой колючки, и оценить силу отбора, приведшего к их возникновению.

Научная новизна и практическая значимость. В предыдущих работах было показано, что локальная скорость мутирования (ЛСМ) является схожей в гомологичных сегментах геномов близкородственных видов, но сильно отличается у давно разошедшихся видов. Эти результаты были получены на малом числе видов, вследствие чего было трудно разобраться, как быстро распадается корреляция ЛСМ между близкородственными видами с ростом филогенетического расстояния. В настоящей работе были использованы данные из выравнивания 100 видов позвоночных, которое появилось в открытом доступе в 2014 году. Был проведён анализ ЛСМ при использовании филогении 9 видов приматов, а также были детально проанализированы вклады различных факторов мутагенеза в изменение ЛСМ на разных эволюционных расстояниях. Были качественно и количественно оценены изменения ЛСМ между близкородственными видами приматов, найдены закономерности изменений и оценены вклады в изменение ЛСМ различных факторов мутагенеза. Полученные результаты в дальнейшем могут быть использованы при разработке эволюционных моделей.

Изучение неравномерности отбора также является важной задачей для понимания процессов, связанных с изменениями в геноме. В 2005 году было показано существование геномных участков с повышенным количеством различий между адаптирующимися к разным условиям популяциями (островов дивергенции, ОД). С введением в широкую практику методов высокопроизводительного секвенирования было показано наличие ОД во множестве других пар популяций, адаптирующихся к разным условиям среды. Тем не менее, на сегодняшний день факторы, влияющие на возникновение и структуру ОД, а также их роль в процессе видообразования остаются малопонятными. Существует круг модельных организмов для данных типов исследований, и к ним относится трёхиглая колюшка *Gasterosteus aculeatus*. В данной работе впервые разработан метод поиска ОД при использовании данных полногеномного секвенирования объединенных выборок (pool-seq) популяций колюшки. Оценены изменения частоты пресноводного аллеля в ОД в молодых пресноводных популяциях, в том числе созданных экспериментально. Количественно оценена сила отбора, действующего на ОД.

Степень достоверности и апробация результатов. По материалам диссертации опубликовано 2 статьи в рецензируемых научных журналах. Результаты работы были представлены на международных конференциях SMBE'12, MCCMB'13, SMBE'14, MCCMB'15 и российских конференциях ИТИС'12, ИТИС'13, ИТИС'15.

Структура и объем диссертации. Диссертация состоит из введения, обзора литературы, 2 глав и библиографии. Общий объем диссертации 94 страницы, из них 85 страниц текста, включая 45 рисунков и 8 таблиц. Библиография включает 131 наименование на 9 страницах.

Содержание работы

Внутривидовая и межвидовая изменчивость генома представляют собой результат длительной эволюции под действием факторов, включающих мутагенез и отбор. Характер и интенсивность этих процессов различаются между сегментами генома, что влияет на плотность наблюдаемых мутаций и их частоты. Определение характеристик этих процессов представляется важной и сложной задачей.

В главе 1 исследуется изменение с ходом эволюции локальной скорости мутирования (ЛСМ), а также оценивается влияние геномных факторов на ЛСМ и её эволюцию. Один из примеров прикладного применения знаний о неравномерности ЛСМ – поиск драйверов раковых заболеваний. В ранних исследованиях в качестве драйверных генов ошибочно выделялись те, что находились в быстро мутирующих участках, такие как гены рецепторов обоняния. Учёт изменчивости ЛСМ позволил значительно улучшить качество определения драйверов.

В течение всего жизненного цикла генетического варианта, например – однонуклеотидного полиморфизма (ОНП), от момента его возникновения в результате мутации до исчезновения или закрепления на него может действовать естественный отбор. Различия в интенсивности действия отбора еще сильнее варьируют между сегментами генома, чем различия в интенсивности мутагенеза, поскольку сила и направление отбора полностью определяются функцией соответствующего участка. Одной из наиболее интересных форм отбора является тот, который можно наблюдать при адаптации к новым условиям среды. В главе 2 исследуются участки с повышенным количеством различий (острова дивергенции, ОД) между морскими и пресноводными популяциями трёхиглой колюшки, в которых предположительно находятся мишени для действия отбора.

Эволюция локальной скорости мутирования (ЛСМ) и определяющих её факторов.

Было исследовано множественное выравнивание последовательностей ДНК восьми видов приматов: шимпанзе (*P. troglodytes*), гориллы (*G. gorilla gorilla*), орангутанга (*P. pygmaeus abelii*), гиббона (*N. leucogenys*), макаки (*M. mulatta*), зелёной мартышки (*C. sabaeus*), боливийского саймири (*S. boliviensis*) и игрунки (*C. jacchus*), с геномом человека (*H. sapiens*). Выравнивание было разделено на неперекрывающиеся окна размером 1 Мб или 100 Кб. Экзоны, нетранслируемые области, повторы, непрочитанные позиции (N), ЦпГ-динуклеотиды и гэпы были маскированы. После этого окна, содержащие <20% немаскированных позиций, были удалены из анализа. Дополнительно также было получено множественное выравнивание тех же девяти геномов приматов с добавлением геномов мыши (*M. musculus*), крысы (*R. norvegicus*) и

китайского хомячка (*C. griseus*). После маскирования были удалены окна, содержащие <10% немаскированных позиций.

Для каждого 1 Мб геномного окна были рассчитаны значения геномных разметок человека. Были использованы разметки, полученные в результате проекта ENCODE: время репликации, участки гиперчувствительные к ДНКазе, метилирование, CTCF, PolII и модификации гистонов. Скорость рекомбинации была взята из проекта НаpMap. Также использовались данные некоторых разметок из других работ: ГЦ-смещённой генной конверсии (Capra et al, 2013), распределения нуклеосом (Yazdi et al, 2015), консервативности (Lindblad-Toh et al, 2011), Hi-C (Dixon et al, 2012). Карта простых повторов была взята из UCSC. Частота минорного аллеля рассчитывалась по данным 1000 геномов человека. Для 1 Мб окон были получены также значения скоростей рекомбинации шимпанзе (Auton et al, 2012) и мыши (Brunschwig et al, 2012). Дополнительно, для мыши были использованы разметки проекта ENCODE: участки гиперчувствительные к ДНКазе, время репликации и модификации гистонов H3K9me3, H3K27ac, H3K27me3.

Для вычисления ЛСМ были использованы данные по дивергенции (дЛСМ) и по полиморфизму (пЛСМ). Расчёт ЛСМ производился с подсчётом количества однонуклеотидных замен, произошедших в данном окне на данной филогенетической линии при помощи метода «максимальной экономии», как плотность замен среди всех немаскированных нуклеотидов окна. Для исключения эффектов, связанных с ГЦ-смещённой генной конверсией, из анализа были исключены все замены W↔S (замены между парами нуклеотидов Г/Ц и А/Т).

ЛСМ сильно скоррелированы между человеком и другими приматами

ЛСМ сильно скоррелированы между человеком и шимпанзе ($R^2=0,82$ для дЛСМ, $P<2,2\times 10^{-16}$; $R^2=0,46$ для пЛСМ, $P<2,2\times 10^{-16}$; рисунок 1, А и Б). Для менее близкородственных видов данная корреляция падает с ростом филогенетического расстояния, достигая минимального значения среди приматов у игрунки ($R^2=0,27$ для дЛСМ, $P<2,2\times 10^{-16}$; $R^2=0,04$ для пЛСМ, $P<2,2\times 10^{-16}$; рисунок 1, А и Б), и это значение ещё меньше для мыши ($R^2=0,11$ для дЛСМ, $P<2,2\times 10^{-16}$; $R^2=0,02$ для пЛСМ, $P<3,13\times 10^{-7}$). Количество дисперсии ЛСМ, объясняемой ЛСМ человека, уменьшается на половину на филогенетическом расстоянии ~0,04 замены на сайт или ~16 миллионов лет (dos Reis et al, 2012), что примерно соответствует последнему общему предку человека и орангутанга.

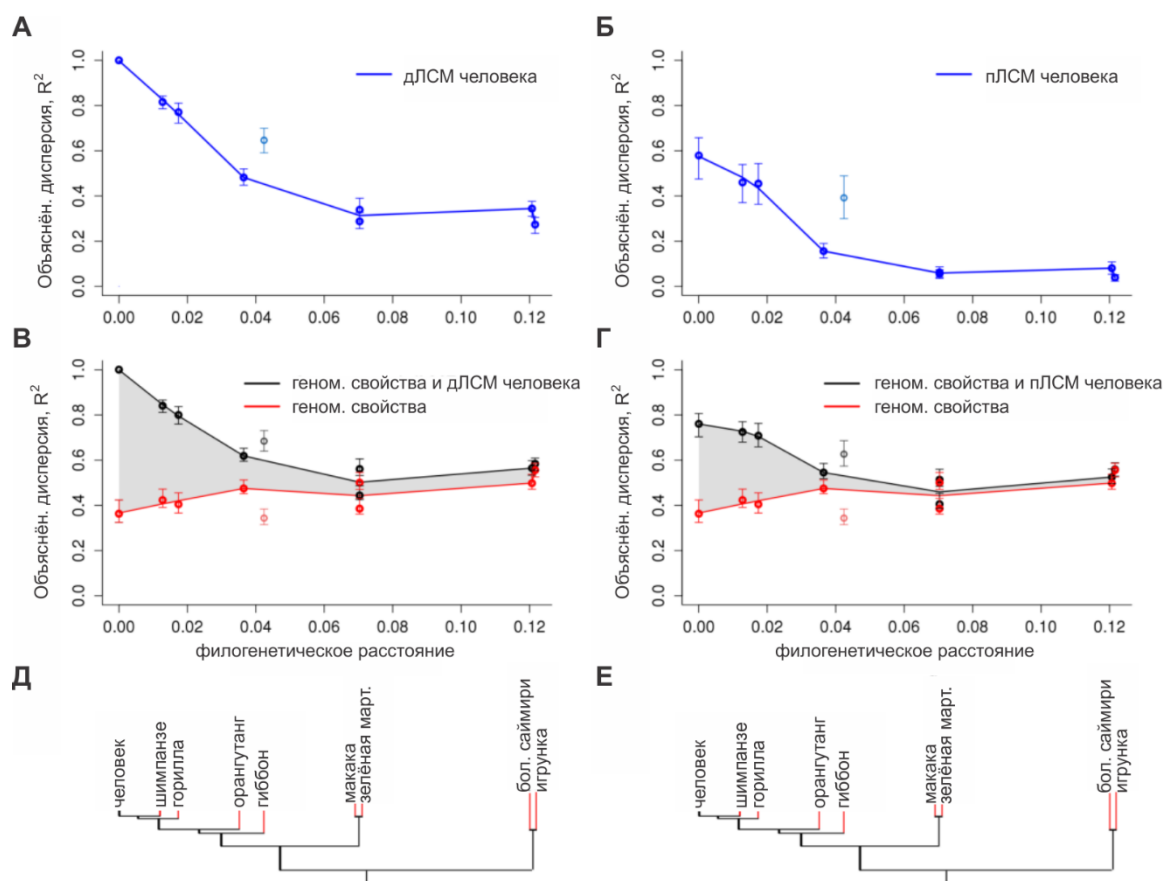


Рисунок 1. Объяснение дисперсии ЛСМ в видах приматов при помощи карт геномных свойств и ЛСМ в линии человека. (А–Г) Для каждого вида приматов доля объясняемой дисперсии в ЛСМ (R^2 , вертикальная ось) изображена в зависимости от филогенетического расстояния этого вида от человека (горизонтальная ось). Значения R^2 для гиббона представлены на графике, но не были включены в регрессию, так как у него наблюдается повышенное по сравнению с другими видами изучаемых приматов количество геномных перестроек. (А–Б) Дисперсия в дЛСМ, объясняемая при помощи дЛСМ (А) или пЛСМ (Б) человека. (В–Г) Дисперсия в дЛСМ, объясняемая при помощи только геномных свойств человека (красный цвет) или в сочетании с дЛСМ (В) или пЛСМ (Г; чёрный цвет) человека. Следующие геномные характеристики были включены в модель: ГЦ-состав, скорость рекомбинации, количество экзонных нуклеотидов, время репликации, количество сайтов, чувствительных к ДНКазе, частота минорного аллеля человеческой популяции, плотность меток модификаций гистонов H3K27ac, H3K27me3 и H3K9me3. Окрашенная серая область обозначает часть дисперсии приматов, за исключением человека, которая может быть дополнительно объяснена при добавлении в модель ЛСМ человека уже после учёта дисперсии, объясняемой геномными свойствами. Планки погрешностей соответствуют 95% доверительным интервалам, полученным за счёт процедуры бутстрэпа. (Д–Е) Филогенетическое дерево для видов, участвующих в анализе. Красным цветом обозначены ветви, для которых производился расчёт ЛСМ.

ЛСМ зависит от свойств ДНК. Линейная модель, предсказывающая дЛСМ человека при помощи разметок геномных свойств ткани эмбриональных стволовых клеток, объясняет только

33% дисперсии дЛСМ. Добавление ЛСМ в модель позволяет значительно повысить количество объясняемой дисперсии (рисунок 1В–Г): от $R^2=0,42$ до 0,72 (для пЛСМ) или до 0,84 (для дЛСМ) для шимпанзе, и от 0,40 до 0,71 (для пЛСМ) или до 0,80 (для дЛСМ) для гориллы. Таким образом, для близких видов семейства *Hominidae* линейная модель, которая включает данные геномных свойств и данные, полученные при использовании человеческого полиморфизма или дивергенции, позволяет объяснить до ~2 раз больше дисперсии ЛСМ по сравнению с моделью, включающую только геномные свойства. Остальная часть дисперсии (рисунок 1 В–Г, серая область) может быть случайной или же быть связана с геномными свойствами, не использованными в данном анализе. Часть дисперсии, которая не объясняется использованными геномными свойствами, вероятно, содержит компоненту, консервативную между близкими видами.

Анализ корреляции дЛСМ между видами также был проведён на геномных окнах 100 Кб. Корреляция между дЛСМ человека и шимпанзе в 100 Кб окнах ($R^2=0,36$, $P<2,2\times 10^{-16}$; рисунок 2) была ниже наблюдаемой для 1 Мб окон, что частично может объясняться меньшей зависимостью от локальных эффектов мелкого масштаба более длинных окон. Снижение значений коэффициентов корреляций с ростом филогенетического расстояния было схожим с наблюдаемым для 1 Мб окон: наименьшее значение объяснённой дисперсии среди приматов было для игрунки ($R^2=0,11$, $P<2,2\times 10^{-16}$, рисунок 2А), и это значение было ещё меньше для мыши ($R^2=0,06$, $P<2,2\times 10^{-16}$ для 100 Кб окон; рисунок 2Б).

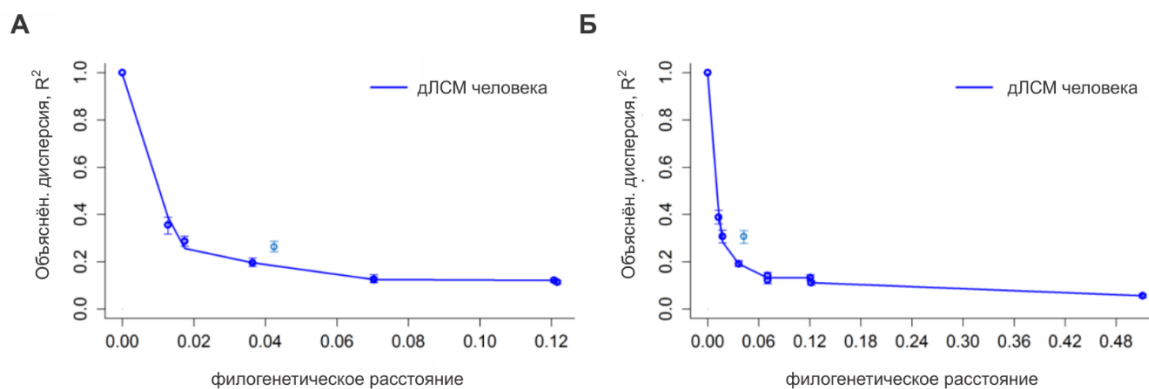


Рисунок 2. Объяснение дисперсии ЛСМ у видов приматов (А) и у видов приматов с добавлением мыши (Б) при помощи ЛСМ человека в 100 Кб геномных окнах. Обозначения те же, что и на рисунке 1.

Объяснение дисперсии в ЛСМ при помощи геномных свойств

ЛСМ в геномных участках изменяется с течением времени (рисунок 1). Количество дисперсии ЛСМ, объясняемой при помощи человеческих геномных свойств, является схожим как для близкородственных видов, так и для более далёких на филогении приматов. Тем не менее есть вероятность, что отдельные геномные свойства, а также их способность предсказывать ЛСМ, могут меняться на других временных промежутках и эти эффекты будут заметны при сравнении более давно разошедшихся по филогении видов. Мы исследовали, насколько изменения в ЛСМ зависят от изменений геномных свойств. Данные для геномных свойств приматов имеются лишь в ограниченном количестве, и поэтому для исследования использовался особый подход.

Вначале оценивалось количество дисперсии ЛСМ, объясняемой отдельными геномными свойствами. Известно, что геномные свойства скоррелированы друг с другом. Поэтому для выяснения вкладов геномных свойств в объяснение дисперсии ЛСМ, независимых от других геномных свойств, был проведён анализ ANOVA III типа (рисунок 3). Полученная оценка для относительных вкладов геномных свойств в предсказание дЛСМ человека согласуется с результатами предыдущих исследований. Вклады геномных свойств в предсказание дЛСМ остальных приматов являются схожими и для большинства свойств не зависят от филогенетического расстояния от человека до рассматриваемого вида (рисунок 3). Единственной из рассматриваемых геномных характеристик, для которой вклад в объяснение дисперсии снижается с ростом филогенетического расстояния, является скорость рекомбинации ($P=0.009$ для корреляции между R^2 и филогенетическим расстоянием). Количество дисперсии, объясняемой данной характеристикой, является более высоким для более близких видов, что соотносится с наблюдением о её большом влиянии на ЛСМ. Тем не менее, доля объясняемой рекомбинацией дисперсии очень быстро снижается с ростом филогенетического расстояния (от 6,05% для человека до 0,01% для игрушки, рисунок 3). Данное снижение, вероятно, связано непосредственно с рекомбинацией, а не со связанным с ней процессом генной конверсии, так как в нашем анализе были исключены все замены, которые могут быть результатами смещённой генной конверсии. Полученные результаты согласуются с наблюдениями влияния рекомбинации на ЛСМ, независимой от смещённой генной конверсии. Наблюдаемое снижение объясняемой рекомбинацией доли дисперсии ЛСМ даёт основания предположить, что изменения скорости рекомбинации происходят быстро, что соответствует её известной способности к быстрой эволюции. Такие изменения способны влиять на изменения ЛСМ.

Геномное свойство	R Пирсона	P-значение
ЛСМ	0,34 (0,29; 0,38)	$<2,2 \times 10^{-16}$
Скорость рекомбинации	0,01 (-0,04; 0,06)	0,67
ДНКаз	0,83 (0,82; 0,85)	$<2,2 \times 10^{-16}$
Время репликации	0,71 (0,69; 0,74)	$<2,2 \times 10^{-16}$
ГЦ-состав	0,94 (0,93; 0,94)	$<2,2 \times 10^{-16}$
Плотность экзон. нукл.	0,95 (0,94; 0,95)	$<2,2 \times 10^{-16}$
H3K9me3	0,32 (0,28; 0,37)	$<2,2 \times 10^{-16}$
H3K27ac	0,73 (0,70; 0,75)	$<2,2 \times 10^{-16}$
H3K27me3	0,66 (0,63; 0,69)	$<2,2 \times 10^{-16}$

Таблица 1. Коэффициенты корреляций Пирсона между геномными свойствами человека и мыши по 1 Мб геномным окнам. 95% доверительные интервалы показаны в скобках. ДНКаз – плотность участков, чувствительных к ДНКазе.

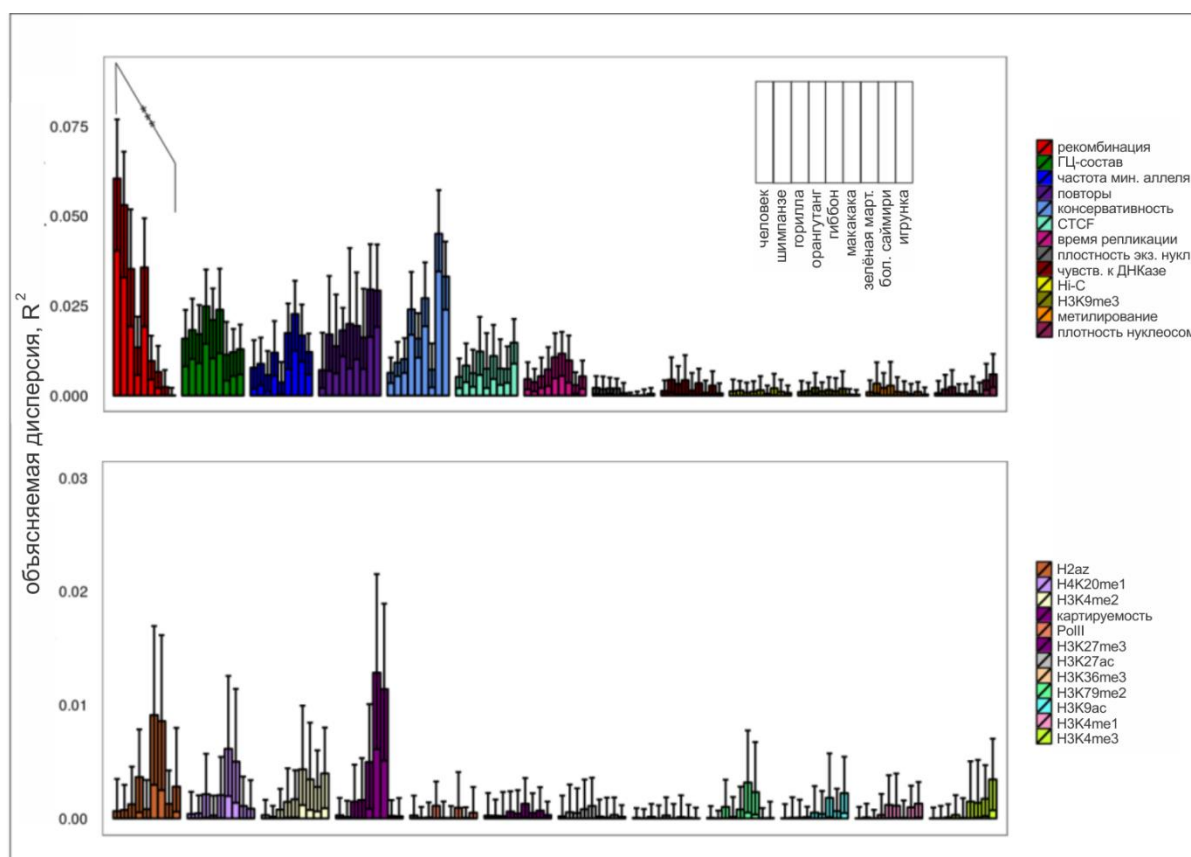


Рисунок 3. Дисперсия ЛСМ, объясняемая 25 геномными свойствами. Для каждого геномного свойства доля объясняемой дисперсии ЛСМ, определённой при помощи анализа ANOVA III, показано для каждого вида приматов (виды расположены для каждого геномного свойства слева направо – с увеличением филогенетического расстояния от человека, как на дереве рисунка 1). Порядок расположения геномных характеристик соответствует уменьшению объясняемой дисперсии в линии человека.

Звёздочками обозначена значимость отрицательной корреляции между R^2 и филогенетическим расстоянием.

Далее была произведена оценка влияния изменений скоростей рекомбинации на эволюцию ЛСМ между человеком и шимпанзе или человеком и мышью. Было найдено, что рекомбинация сильно различается между двумя видами: скорость рекомбинации человека только слабо скоррелирована со скоростью рекомбинации шимпанзе ($R^2=0,24$, $P<2,2\times 10^{-16}$) и совсем не скоррелирована с соответствующей у мыши (таблица 1).

Было проведено сравнение разницы в дЛСМ с разницей в скорости рекомбинации для каждого 1 Мб геномного окна. В обоих сравнениях (человек-шимпанзе, человек-мышь) данные факторы имели слабую положительную корреляцию ($R^2=0,01$, $P<7,0\times 10^{-6}$ для шимпанзе, рисунок 4А; и $R^2=0,1$, $P<2,2\times 10^{-16}$ для мыши, рисунок 4Б). Данный результат показывает, что повышение скорости рекомбинации геномного окна связано с повышением его ЛСМ и наоборот.

Для сравнения человек-мышь были проанализированы также и другие геномные характеристики на предмет ассоциации их изменений в распределении между видами с изменениями в ЛСМ. В целом $\sim 16,4\%$ всей дисперсии изменения ЛСМ между видами может быть объяснено изменениями в геномных свойствах (рисунок 4). Когда рассматривались отдельные вклады геномных характеристик в объясняемую дисперсию, было выявлено, что изменения скорости рекомбинации больше других свойств объясняют дисперсию изменения ЛСМ ($\sim 10\%$, рисунок 4, А–Б и К). Для выделения независимых вкладов изменений в геномных свойствах между человеком и мышью в объяснение дисперсии изменений ЛСМ между этими двумя видами был использован анализ ANOVA III (рисунок 4Л). Изменения только нескольких геномных свойств вносили существенный вклад в изменения ЛСМ. Наибольшее количество дисперсии было объяснено изменением в скорости рекомбинации; вклад этого свойства был гораздо больше, чем следующего по значению фактора (ГЦ-состав). Полученный результат показывает, что, несмотря на быструю эволюцию распределения горячих точек рекомбинации и то, что карта рекомбинации человека нисколько не предсказывает карту рекомбинации мыши (таблица 1), изменения рекомбинации, которые происходят между двумя этими видами, объясняют изменения в ЛСМ сильно лучше, чем изменения других геномных свойств.

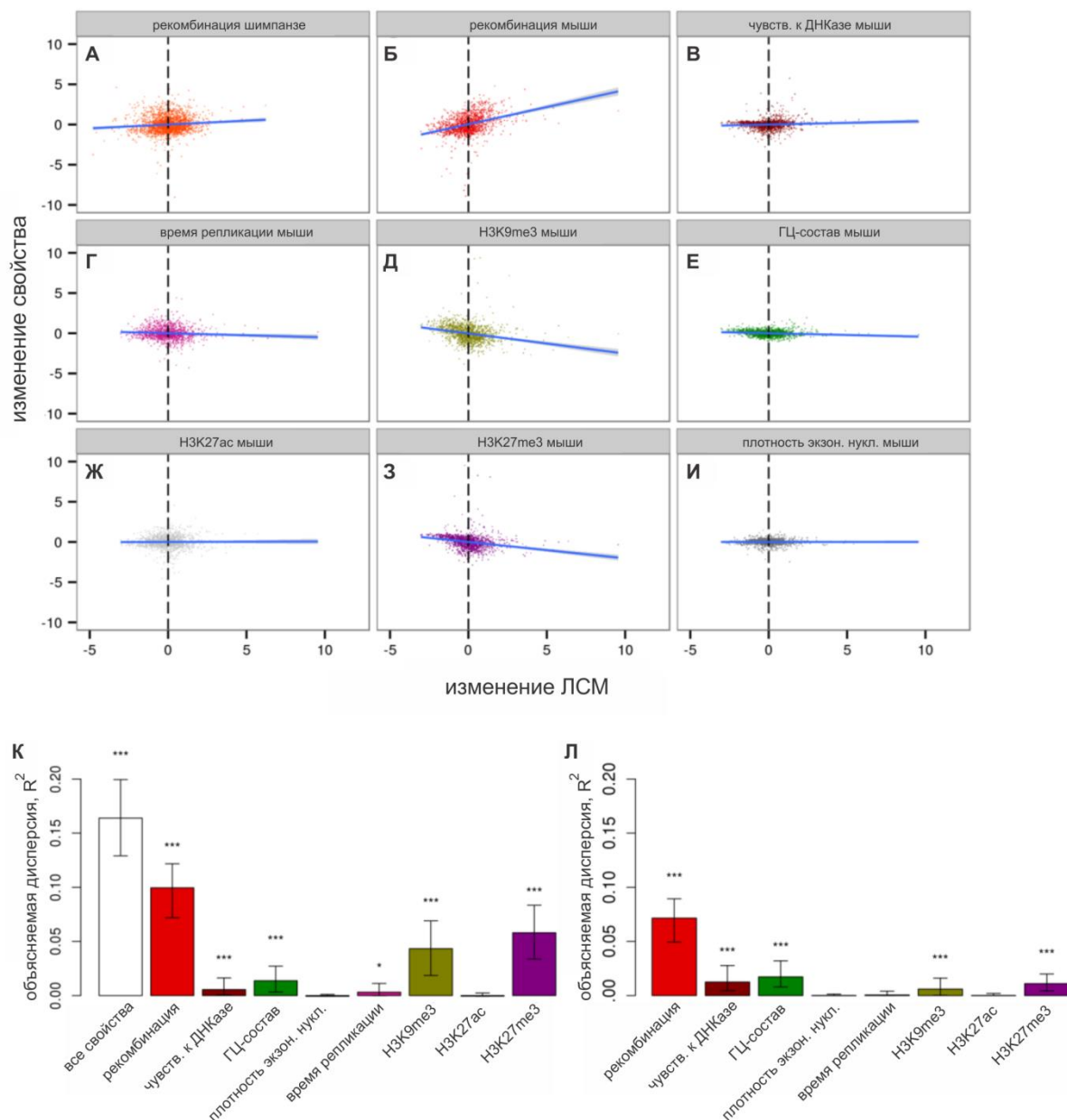


Рисунок 4. (А-И) Графики корреляций между изменениями ЛСМ и изменениями некоторых геномных свойств между видами человек и мышь (или шимпанзе в случае рекомбинации). Каждая точка соответствует одному 1 Мб геномному окну. (К-Л) Изменения в ЛСМ, объясняемые изменениями в некоторых геномных свойствах между человеком и мышью. Вертикальная ось соответствует количеству дисперсии в изменении ЛСМ, объясняемой изменениями геномных свойств между человеком и мышью. Каждый столбец соответствует количеству дисперсии, объясняемой конкретной геномной характеристикой, посчитанной через R^2 (К) или по анализу ANOVA тип III (Л). Планки погрешностей соответствуют 95% доверительным интервалам, полученным через процедуру бутстрэпа. Звёздочки показывают уровень значимости отклонения от 0 полученных количеств объяснённой дисперсии (*: $P < 0,05$, **: $P < 0,01$, ***: $P < 0,001$).

Для того чтобы лучше разобраться в связи изменений в скорости рекомбинации с изменениями в ЛСМ, был проведён анализ геномных окон, для которых значения ЛСМ были значительно повышены или снижены в человеке или шимпанзе с момента их дивергенции от общего предка человека и шимпанзе. При проведении поиска сначала рассчитывалась ожидаемая ЛСМ, $u_{b,h}(exp)$ в виде b для каждого окна h , с нормализацией на среднюю ЛСМ для данного окна по всем видам, $u_h = \frac{\sum_i u_{i,h}}{i}$, а также с нормализацией на длину филогенетической ветви данного вида, $v_b = \frac{\sum_j u_{b,j}}{j}$, и среднюю ЛСМ по всем окнам во всех видах, $u_{b,h}(exp) = \frac{\sum_i u_{i,h} \sum_j u_{b,j}}{\sum_i \sum_j u_{i,j}}$.

Затем проводилось сравнение полученных ожидаемых значений ЛСМ, $u_{b,h}(exp)$ человека или шимпанзе, с соответствующими наблюдаемыми значениями ЛСМ, $u_{b,h}(obs)$. Для выбора окон, в которых произошли наибольшие изменения ЛСМ в интересующем виде, отбирались те, для которых размер изменения в данном виде sp_1 (человек или шимпанзе) был большим по сравнению с соответствующим изменением в наиболее близком виде sp_2 (соответственно, шимпанзе или человек):

$$\left| \lg \frac{u_{sp_1,h}(obs)}{u_{sp_1,h}(exp)} \right| > \left| \lg \frac{u_{sp_2,h}(obs)}{u_{sp_2,h}(exp)} \right|$$

Окна ранжировались при помощи статистики, которая показывает, во сколько раз отношение наблюдаемой к ожидаемой ЛСМ в виде sp_1 больше соответствующего отношения для вида sp_2 :

$$\Delta_h = \frac{u_{sp_1,h}(obs)}{u_{sp_1,h}(exp)} / \frac{u_{sp_2,h}(obs)}{u_{sp_2,h}(exp)},$$

В качестве окон с наиболее ускорившейся или замедлившейся ЛСМ были отобраны 100 окон с самыми высокими или с самыми низкими значениями Δ_h .

В окнах с ускорившейся ЛСМ в линии человека скорость рекомбинации человека была значительно выше соответствующего среднего значения по геному (односторонний тест Манна-Уитни $P=2,9 \times 10^{-9}$, рисунок 5А), и это говорит о том, что такие окна чаще среднего содержат в себе горячие точки рекомбинации. Напротив, в окнах, для которых значение ЛСМ повышено у шимпанзе, скорость рекомбинации человека не так сильно повышена по сравнению со средним значением по геному ($P=8,0 \times 10^{-3}$, рисунок 5Б). Полученные результаты дают основания предположить, что окна с повышенной ЛСМ в человеке часто связаны с присутствием горячих точек рекомбинации, которые являются короткоживущими и поэтому повышают ЛСМ в человеке гораздо сильнее, чем в шимпанзе. Аналогично, окна с повышенной ЛСМ в шимпанзе

также имеют повышенную скорость рекомбинации ($P=3,7 \times 10^{-7}$, рисунок 5Г), немного превышающую соответствующее значение в окнах с ускорившейся ЛСМ в человеке ($P=7,4 \times 10^{-6}$, рисунок 5В).

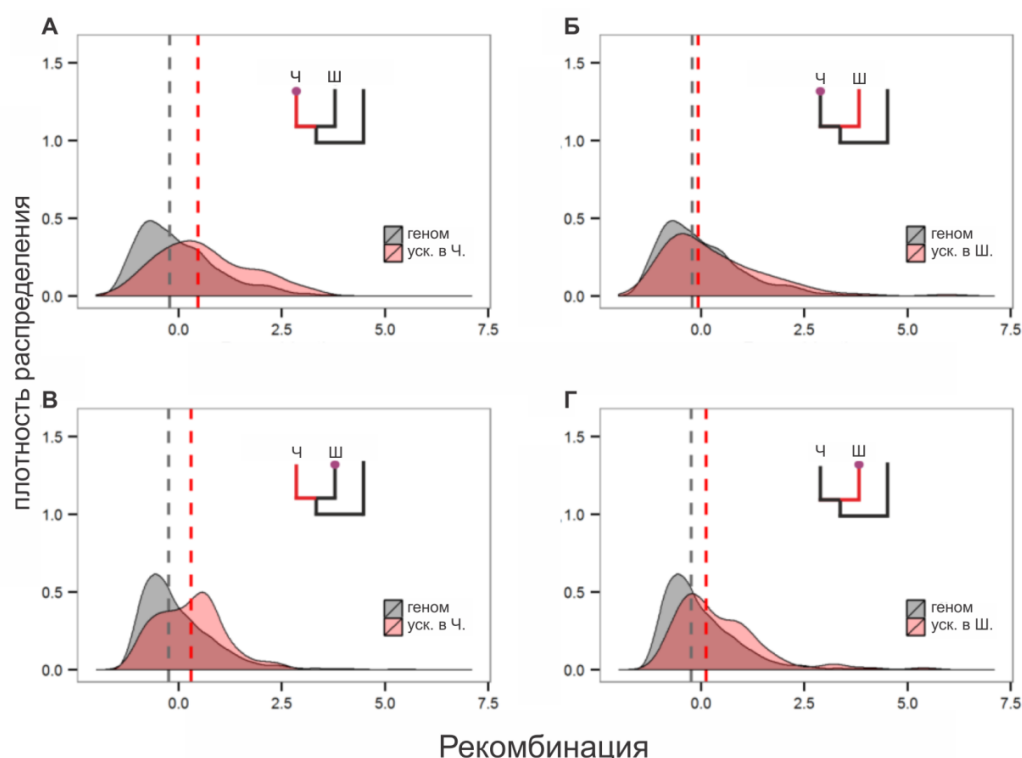


Рисунок 5. Распределения значений скоростей рекомбинации человека (А–Б) в окнах с ускоренной ЛСМ в человеке (уск. в Ч.; А) и шимпанзе (уск. в Ш.; Б), и распределение значений скоростей рекомбинации шимпанзе (В–Г) в окнах с ускоренной ЛСМ в человеке (В) и шимпанзе (Г). Схематические филогении показывают линию вида, в котором ЛСМ была повышена, красным цветом (Ч, человек или Ш, шимпанзе); виды, в которых измерялась рекомбинация, обозначены кругами. Штрихованная линия отмечает медианное значение скорости рекомбинации в ускоренных участках и по всему геному.

Геномная архитектура адаптации у трёхиглой колюшки

Сбор образцов популяций трёхиглой колюшки производился в июне-августе 2011 г. (таблица 2). Перед приготовлением библиотек образцы ДНК для каждого объединенного образца, содержащего от 8 до 20 индивидуумов, смешивались в равных соотношениях. Геномные чтения были картированы на референсный геном вида *G. aculeatus*, полученный из геномного браузера UCSC. Частоты аллелей рассчитывались для каждого образца в каждой позиции. Маркерный ОНП определялся как полиморфная позиция, в которой обе морские популяции содержали один и тот же аллель на частоте $\geq 80\%$, в то время как обе пресноводные популяции

содержали другой аллель с частотой $\geq 80\%$ («сильный критерий») или же на частоте $\geq 50\%$, что соответствовало «слабому критерию».

Образец	Описание	Количество особей	Среднее покрытие	Нуклеотидное разнообразие (π)
Нильма ^a	Морские	16	44	0.00210
Ершовское (проходные) ^a	Проходные, собранные в Ершовском	10	17	0.00188
Ершовское (резидентные) ^b	Пресноводные, возраст ~ 34 г.	12	52	0.00219
Марцы ^b	Пресноводные, возраст ~ 250 лет	10	24	0.00181
Голубой ^c	Карьер, популяция заселена в 1978 г. 20-ю морскими и 20-ю пресноводными рыбами	20	48	0.00204
Малыш ^c	Карьер, популяция заселена в 1978 г. 1-й морской и 1-й пресноводной рыбами	20	63	0.00186
Лобанешское ^d	Пресноводные, ~ 600 лет	8	23	0.00167
Машинное ^d	Пресноводные, ~ 700 лет	10	19	0.00158

Таблица 2. Места сбора и характеристики образцов. ^aморская популяция; ^bмолодые природные пресноводные популяции; ^cискусственные пресноводные популяции; ^dболее старые природные пресноводные популяции.

Генетические различия между морской и пресноводными популяциями

Всего было получено более 18.000 маркерных ОНП, которые отличали морскую популяцию от пресноводных. Маркерные ОНП были распределены неравномерно вдоль генома, и образовывали ярко выраженные кластеры («острова дивергенции», ОД), для которых наблюдались повышенные плотности распределения маркерных ОНП. Всего было использовано два критерия для определения ОД. ОД, найденный по сильному (слабому) критерию, определялся как непрерывный участок, в котором в каждой 10 Кб рамке генома было как минимум 10 сильных (20 слабых) маркерных ОНП. После этого полученные участки на расстоянии < 40 Кб объединялись, так как рекомбинация маловероятна на таких коротких расстояниях. Такой алгоритм достаточно хорошо описывает наблюдаемые данные (рисунок 6), а также приводит к выделению меньшего числа более длинных островов. Среди 6.107 маркерных ОНП, полученных

при помощи сильного критерия, 5.801 (95%) было расположено внутри ОД. Всего при объединении сильного и слабого критериев было обнаружено 19 ОД, которые были расположены на 10 из 21 хромосом генома *G. aculeatus* (рисунок 6, таблица 3) и покрывали в целом 3.301.948 нуклеотидов последовательности или 0,74% от всего генома.

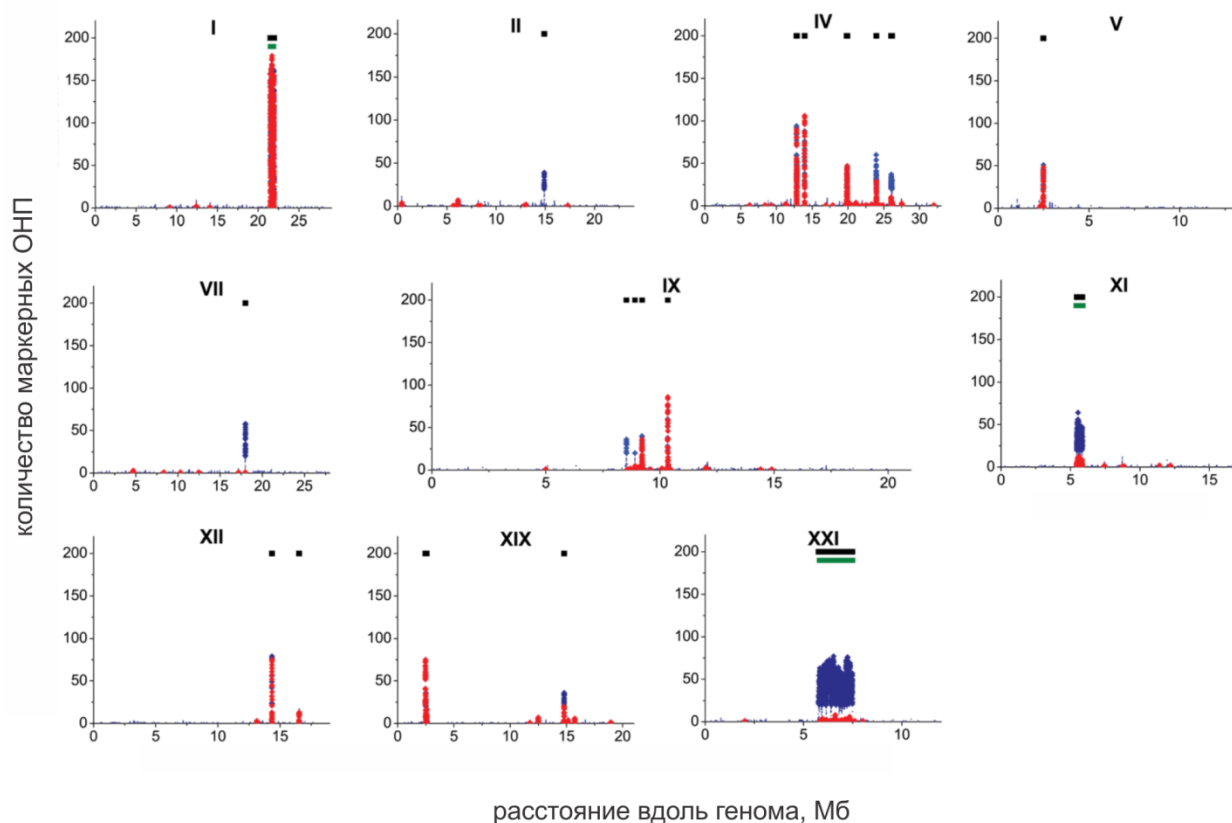


Рисунок 6. Распределение маркерных ОНП, различающих морскую и пресноводные популяции, в геноме *G. aculeatus*. На рисунке присутствуют все хромосомы, которые содержат хотя бы один ОД. Для каждой хромосомы горизонтальная ось показывает координаты расположения вдоль генома 10 Кб окон, а вертикальная ось показывает число маркеров внутри каждого из этих окон. Красный цвет используется для значений, определённых по «сильным» маркерам, а синий цвет для соответствующих значений, определённых по «слабым» маркерам. На данном рисунке более мелкими ромбами изображены значения <20 для слабых маркеров. Чёрными полосами сверху обозначены границы определённых ОД, а зелёные полосы отмечают места известных инверсий на хромосомах I, XI и XXI.

ОД	Начало	Конец	# маркер-ных ОНП	Также были найдены в		# генов, перекр. с ОД	# нс ОН П	s _E	s _G
				[1]	[2]				
I-1	21.487.998	21.960.119	4.186	+	+	27	100	0,247	0,212
II-1*	14.874.366	14.898.826	73	+	+	0	0	0,115	0,146
IV-1	12.803.780	12.881.296	285	+	+	7	19	0,188	0,094
IV-2	13.930.002	13.959.331	168		+	2	1	0,218	0,107
IV-3	19.811.922	19.914.666	209	+	+	6	0	0,212	0,163
IV-4	23.954.634	23.981.981	48	+	+	3	1	0,227	0,1
IV-5*	26.016.955	26.166.536	252	+	+	4	1	0,196	0,036
V-1	2.482.209	2.501.295	65		+	2	3	0,255	0,239
VII-1*	17.982.351	18.002.671	84	+	+	4	6	<0	0,273
IX-1*	8.521.935	8.537.559	44		+	2	0	0,062	0,026
IX-2*	8.901.816	8.910.115	20		+	2	0	0,108	0,023
IX-3	9.208.158	9.227.809	46		+	1	0	0,198	0,131
IX-4	10.334.101	10.353.801	114		+	1	0	0,141	0,101
XI-1*	5.445.757	5.855.124	1.237	+	+	21	10	0,054	0,147
XII-1	14.338.229	14.358.336	91	+	+	2	0	0,023	0,193
XII-2	16.522.028	16.538.810	24		+	2	0	0,162	
XIX-1	2.449.903	2.581.858	277	+	+	5	11	0,178	0,131
XIX-2	14.787.904	14.799.088	21	+	+	0	0	0,117	0,056
XXI-1*	5.759.879	7.486.635	6.900	+	+	79	63	0,177	<0
Всего			14.144	12	19	170	215		

Таблица 3. 19 ОД, отличающих морские и пресноводные популяции. Идентификаторы ОД представлены в следующем формате '№ хромосомы-порядковый № ОД'. Звёздочкой отмечены ОД, которые были определены при помощи слабого критерия; все остальные ОД были определены при помощи сильного критерия. '+' отмечает ОД, которые были также найдены в работе [1] (Hohenlohe et al, 2010) или в работе [2] (Jones et al., 2012); s_E – коэффициент отбора, благоприятствующий закреплению пресноводного аллеля, рассчитанный для популяции озера Ершовского; s_G – коэффициент отбора, благоприятствующий закреплению пресноводного аллеля, рассчитанный для популяции карьера Голубой.

17 из 19 найденных ОД перекрывались с границами 170 генов. Среди маркерных ОНП, находящихся внутри этих генов, 139 были несинонимическими и 146 были синонимическими (таблица 4). Отношение несинонимических к синонимическим маркерным ОНП было выше, чем

то же отношение для немаркерных ОНП в ОД, которые были полиморфными внутри морской популяции (290 несинонимических и 528 синонимических; таблица 4). Данный факт даёт основания предполагать, что большее количество позиций находится под положительным отбором среди маркерных ОНП, чем среди немаркерных ОНП. Тем не менее, 7 из 17 ОД не содержат несинонимических маркерных ОНП и 2 ОД совсем не перекрываются ни с одним из известных генов. Предположительно гены, которые перекрываются с ОД, могут быть ответственны за фенотипические различия между морской и пресноводной формами колюшки.

	Маркерные ОНП			ОНП морской популяции		
	Несинонимич.	Синонимич.	Отношение	Несинонимич.	Синонимич.	Отношение
Внутри ОД	139	146	0,95	290	528	0,55
Вне ОД	17	4	4,25	28.616	41.902	0,68

Таблица 4. В таблице представлены числа несинонимических и синонимических замен, расположенных внутри и вне ОД, для маркерных ОНП и для немаркерных ОНП морской популяции.

Динамика адаптации *G. aculeatus* к пресноводной среде обитания

Помимо образцов из озёр Машинное и Лобанешское также были секвенированы образцы *G. aculeatus* из четырёх более молодых популяций: карьеров Голубой и Малыш, озёр Марцы и Ершовское, все из которых располагаются вблизи от Белого моря. Карьерные популяции были основаны В. Зюгановым в 1978 г. в заброшенных изолированных карьерах, в которых до заселения не было рыбы этого вида. Таким образом, возраст этих популяций составляет 34 года на момент забора образцов. Карьер Голубой (площадью ~70.000 м² и вместимостью ~1.000 рыб) был изначально заселён 20 морскими (10 самок и 10 самцов) и 20 пресноводными (10 самок и 10 самцов) особями. Карьер Малыш (площадью ~75 м² и вместимостью ~100 рыб, с очень малым количеством мест пригодных для постройки гнёзд) был заселён 1 морской самкой и 1 пресноводным самцом. Морские особи были взяты из популяции Белого моря, а пресноводные особи были из популяции озера Машинное. Озёра Марцы и Ершовское образовались вследствие изоляции морских заливов в результате постепенного поднятия земной коры над уровнем моря со скоростью ~3,8 мм за год. По оценкам, возраст пресноводной популяции озера Марцы составляет ~250 лет, вследствие наблюдения, что поверхность озера на данный момент находится на возвышении ~1 м над уровнем моря. Озеро Ершовское на данный момент находится на возвышении ~13 см над уровнем моря, и мы предполагаем возраст популяции озера Ершовское

равным 34 годам. Таким образом, вероятно, что во всех вышеперечисленных молодых пресноводных популяциях процесс адаптации все ещё продолжается.

Частоты пресноводных аллелей в маркерных ОНП внутри островов во всех четырёх молодых пресноводных популяциях значительно увеличены относительно соответствующих в морской популяции (рисунок 7). Наблюдаемая динамика роста частоты пресноводного аллеля с увеличением возраста популяции предполагает, что в каждой из них отбор благоприятствует одним и тем же аллелям. Первоначальные частоты пресноводных аллелей в популяциях озёр Ершовское и Марцы предположительно были равными наблюдаемым в морской популяции ($\sim 0,1$, таблица 3). В двух искусственных популяциях, населённых равными количествами морских и пресноводных индивидуумов, изначальная средняя частота пресноводного аллеля составляла 0,5. Измеренные средние частоты пресноводных аллелей искусственных популяций составляли 0,56 для карьера Малыш и 0,73 для карьера Голубой (рисунок 7). В карьере Малыш частоты пресноводных аллелей имели меньшее среднее и большую дисперсию по сравнению с соответствующими для карьера Голубой, что вероятно может являться результатом более низкой эффективной численности карьера Малыш и вследствие этого более сильным эффектом дрейфа в этой популяции. Частоты пресноводных аллелей в озере Марцы в среднем выше, чем в озере Ершовское, вероятно вследствие того, что последняя популяция имеет меньший возраст.

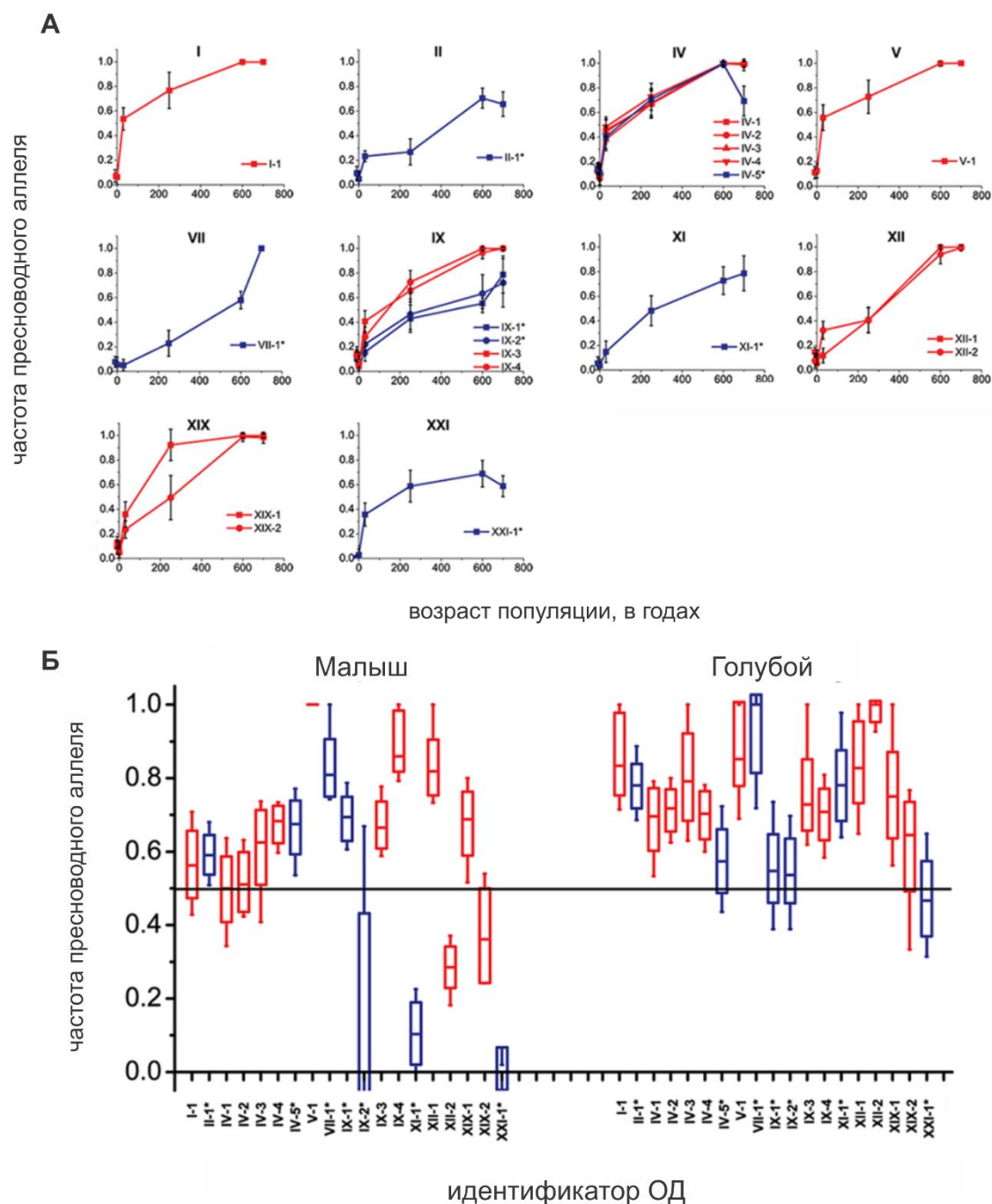


Рисунок 7. Средние частоты пресноводных аллелей в маркерных ОНП внутри ОД для популяций разных возрастов. (А) Природные популяции. Горизонтальная ось показывает примерный возраст популяций, слева направо идут: морская популяция (~0 лет), молодые пресноводные популяции (~34 и ~250 лет) и более старые пресноводные популяции (~600 и ~700 лет). Усы соответствуют стандартным отклонениям. (Б) Искусственные популяции. Для каждого ОД первоначальная частота пресноводных аллелей считается равной 50% (чёрная линия) а частоты в настоящем времени показаны: карьер Малыш (слева) и карьер Голубой (справа), популяция в каждом из которых была основана в 1978 г. Усы, прямоугольники и прямые внутри них соответствуют 5-й и 95-й перцентилям, стандартному отклонению и медиане соответственно. Красный цвет – ОД, определённые по сильному критерию; синий – ОД, определённые по слабому критерию.

Сила отбора, способствующего адаптации к пресноводной среде обитания

При знании изменений частот пресноводных аллелей можно рассчитать силу отбора, который влияет на их изменения. Вследствие того, что возраст озера Марцы известен только приблизительно, было решено не использовать эту популяцию для расчёта силы отбора. Также для расчётов не была использована популяция озера Малыш, так как она имела изначально очень низкую численность, и в таком случае ожидается сильный эффект действия генетического дрейфа. Тем самым, расчёты проводились только для молодых популяций озера Ершовское и карьера Голубой. При оценке коэффициента отбора в качестве значения длины поколения G , *aculeatus* было взято 2 года, а в качестве коэффициента доминирования был выбран $h=0,5$, что согласуется с результатами недавних исследований для множества морфологических признаков *G. aculeatus*.

Оценки коэффициентов отбора были рассчитаны для ОД, а не для отдельных маркерных ОНП, так как оценка частоты в индивидуальных ОНП будет смещена с большей вероятностью. Коэффициент отбора s рассчитывался на основе средней частоты пресноводных аллелей внутри ОД. Стоит понимать, что частоты маркерных ОНП не являются независимыми внутри ОД, и это позволяет более точно рассчитать среднюю частоту пресноводных аллелей внутри ОД, и соответственно получить более достоверную оценку коэффициента отбора.

В целом, полученные значения для s были достаточно высокими: из 19 ОД 15 в популяции озера Ершовское и 12 в популяции карьер Голубой имели $s>0,1$ (таблица 3, рисунок 8). Коэффициенты отбора, оцененные для каждого ОД, не имели значимой корреляции между двумя пресноводными популяциями (рисунок 8; $\rho=0,30$, $P=0,27$). Данный результат ожидаем, поскольку мы рассматриваем лишь небольшое количество ОД, и ошибка оценки коэффициента отбора для отдельных ОД достаточно высока. Тем не менее, средние значения для s популяций озера Ершовское ($s=0,16$) достаточно схожи с s для популяции из карьера Голубой ($s=0,13$). В частности, ОД IV-1, в котором расположен ген *EDA*, коэффициент отбора $s=0,19$ в популяции озера Ершовское, и $s=0,09$ в популяции карьера Голубой. Данные высокие значения s согласуются с ранее полученными данными о том, что изменения фенотипа по признаку, контролируемому геном *EDA*, наблюдаются уже на очень коротких временных промежутках при изменении условий среды обитания. Наибольшее изменение средней частоты пресноводного аллеля в популяции озера Ершовское было обнаружено для ОД DI V-1, в котором соответствующая частота возросла с 0,1 до 0,56, что соответствует коэффициенту отбора $s=0,255$. В данном ОД наблюдается несколько несинонимических замен в гене *HTR3A*, субъединице

серотонинового рецептора с широким кругом физиологических функций. С немного меньшим изменением частоты наблюдался ОД I-1 ($s_E=0,247$, $s_G=0,212$), который находится в инверсии длиной 470 Кб, содержащей 27 генов, включая ген *ATP1A1*, кодирующий белок Na^+/K^+ АТФазу. Маркерные ОНП внутри этого гена включают 6 несинонимических замен, которые могут быть частично ответственны за различия в осморегуляции между морской и пресноводной популяциями. Среди ОД популяции карьера Голубой наиболее сильное изменение частоты пресноводного аллеля наблюдалось для ОД XII-2 (соответствующая частота почти достигала 1, и при таких значениях s достаточно трудно рассчитать точно). Данный ОД перекрывается с участком, находящимся вблизи от 5'-конца гена *OVGP1*. Данный ген участвует в процессе размножения, и тем самым является вероятной мишенью для действия положительного отбора.

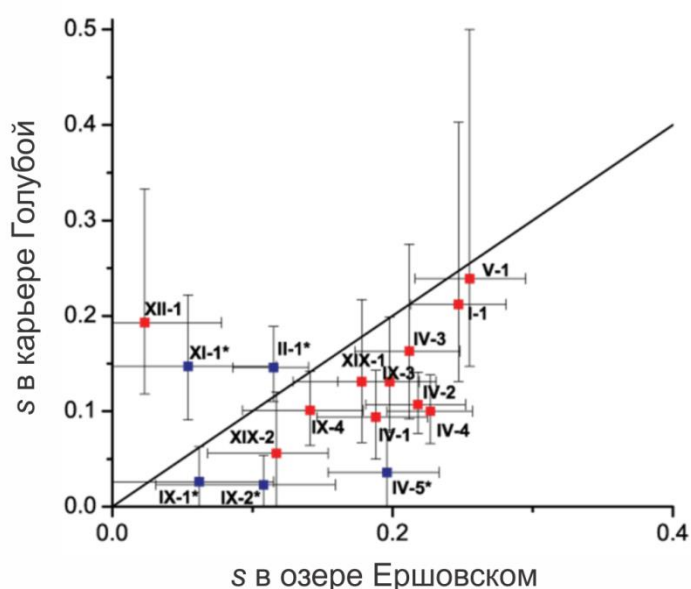


Рисунок 8. Коэффициенты отбора для ОД, определённые для популяций из озера Ершовское и карьера Голубой. Красным обозначены s в ОД, определённых по сильному критерию, синим – ОД, определённые по слабому критерию. Три ОД не обозначены на рисунке: VII-1 и XXI-1, для которых $s < 0$ в одной из популяций, и XII-2, для которой s в популяции карьера Голубой не может быть рассчитан относительно достоверно из-за частоты пресноводного аллеля, близкой к фиксации (таблица 3). Усы соответствуют значениям s , полученным для оценок средних частот пресноводных аллелей с учётом стандартных отклонений.

Выводы

1. Показано, что учёт локальной скорости мутирования (ЛСМ) близкородственных видов приматов позволяет объяснить существенно (почти в два раза) большую долю дисперсии ЛСМ человека по сравнению с использованием только карт известных геномных разметок, как это делалось в предыдущих работах.

2. Вклад скорости рекомбинации в объясняемую дисперсию ЛСМ является наибольшим среди всех исследованных геномных характеристик. Однако, если для предсказания ЛСМ в одном виде используется локальная скорость рекомбинации другого вида, то её предсказательная сила быстро падает с ростом филогенетического расстояния между видами. Изменения скорости рекомбинации между видами человек и мышь связаны с изменениями ЛСМ сильнее, чем изменения других исследованных геномных характеристик.

3. На примере человека и шимпанзе разработан метод поиска участков генома с повышенной и пониженной ЛСМ на линии вида. Показано, что участки с повышенной ЛСМ у человека или у шимпанзе в среднем имеют повышенные значения скоростей рекомбинации. Данный результат свидетельствует о том, что рекомбинация ассоциирована с изменением мутагенеза на коротких эволюционных временах.

4. Описаны участки генома, которые предположительно вовлечены в адаптацию к пресноводной среде обитания – острова дивергенции (ОД). Найдены 19 ОД, содержащих повышенное количество различий между морскими и пресноводными особями трёхиглой колюшки. Показано, что скорость эволюции повышена в ОД, и изучена динамика изменения частот пресноводных аллелей внутри ОД. Полученные результаты согласуются с гипотезой о том, что для адаптации используется предсуществующая генетическая изменчивость.

5. Показано, что на ОД при адаптации действует сильный отбор. Средние значения коэффициентов отбора составили $s=0,16$ для популяции озера Ершовское и $s=0,13$ для популяции карьера Голубой.

Список публикаций по теме диссертации

Статьи в рецензируемых научных журналах

1. Terekhanova N.V., Logacheva M.D., Penin A.A., Neretina T.V., Barmintseva A.E., Bazykin G.A., Kondrashov A.S., Mugue N.S. Fast evolution from precast bricks: genomics of young freshwater populations of threespine stickleback *Gasterosteus aculeatus* // PLoS Genet 2014. Vol. 10, no. 10, P. e1004696.
2. Terekhanova N.V., Seplyarskiy V.B., Soldatov R.A., Bazykin G.A. Evolution of local mutation rate and its determinants // Mol Biol Evol 2017. Vol. 34, no. 5, P. 1100–1109.

Тезисы конференций

1. Terekhanova N.V., Mugue N.S., Bazykin G.A., Kondrashov A.S. Population genomics of threespine stickleback after 30-year evolution experiment // “Society for Molecular Biology & Evolution” (SMBE 2012), Дублин, Ирландия, 2012.
2. Тереханова Н.В., Базыкин Г.А., Кондрашов А.С., Мюге Н.С. Genetic variation in the genomes of the threespine stickleback *Gasterosteus aculeatus* // “Информационные технологии и системы” (ИТИС 2012), Петрозаводск, Россия, 2012.
3. Terekhanova N.V., Bazykin G.A., Kondrashov A.S., Mugue N.S. Patterns of genomic divergence during adaptation to fresh water in the threespine stickleback *Gasterosteus aculeatus* // “Moscow Conference on Computational Molecular Biology” (MCCMB 2013), Москва, Россия 2013.
4. Terekhanova N.V., Bazykin G.A., Kondrashov A.S., Mugue N.S. Genome-wide patterns of divergence during adaptation to fresh water in threespine stickleback *Gasterosteus aculeatus* // “Информационные технологии и системы” (ИТИС 2013), Калининград, Россия, 2013.
5. Terekhanova N.V., Bazykin G.A., Kondrashov A.S., Mugue N.S. Fast evolution in young freshwater populations of *G. aculeatus* // “Society for Molecular Biology & Evolution” (SMBE 2014), Сан-Хуан, Пуэрто-Рико, 2014.
6. Terekhanova N.V., Bazykin G.A., Soldatov R.A., Seplyarskiy V.B. Local variation of the mutation rate across the primate phylogeny // “Moscow Conference on Computational Molecular Biology” (MCCMB 2015), Москва, Россия, 2015.

7. Terekhanova N.V., Bazykin G.A., Soldatov R.A., Seplyarskiy V.B. Evolution of local mutation rate and its determinants // “Информационные технологии и системы” (ИТИС 2015), Сочи, Россия, 2015.