1.9.2.3. Сведения о фактическом выполнении плана работы на год (фактически проделанная работа, до 10 стр.)

Тема 1

Получение библиотек для секвенирования из собранных образцов при стрессовых воздействиях

Для изучения ответа разных частей и органов *A. thaliana* на холодовое воздействие было собрано три временных точек для шести органов растения по две повторности каждая (всего 36 проб, каждая состояла из объединенной выборки по 15 растений).

Из данных 36 образцов была выделена тотальная РНК, качество которой было оценено с помощью параметра RIN (RNA Integrity Number, значение целостности РНК). Параметр RIN имел величину не менее 8, что говорит о высоком качестве РНК. Библиотеки для секвенирования из РНК такого качества могли быть приготовлены с помощью протокола, основанного на выделении мРНК из тотальной на микрочастицах с олигоТ-нуклеотидами. Кроме того, использование этого протокола позволило проводить сравнения с транскриптомной картой *A. thaliana,* полученной этим же методом на предыдущих этапах исследования.

Полученные библиотеки были секвенированы на платформе Illumina HiSeq2000 с глубиной секвенирования, достаточной для получения не менее 15 миллионов уникально картированных чтений на образец. Исходные чтения были отфильтрованы по качеству с удалением адаптерных последовательностей, необходимых для секвенирования. После этого полученные высококачественные чтения были картированы на геном *A. thaliana* со средним числом чтений, уникально картирующихся на гены, около 18 млн. на образец.

Сходство биологических повторностей было оценено с помощью квадрата коэффициента корреляции Пирсона r^2, который варьировался от 0,9 до 1 (среднее 0,97), что указывает на пригодность биологических повторностей для дальнейшего анализа. Дерево, полученное кластеризацией, также продемонстрировало сходимость биологических повторностей и подтвердило корректность сбора образцов. В структуре дерева присутствует четыре кластера: все образцы цветков (12 проб); все образцы семян (6 проб); листоподобные образцы при 3 часах обработки холодом и их контроли (листья, гипокотили и семядоли, 12 проб); листоподобные образцы при 27 часах обработки холодом (6 проб).

Создание общедоступной базы данных, содержащей информацию об уровне экспрессии генов в разных элементах растения и инструменты для ее анализа

Общедоступная база данных TRAVA (http://travadb.org/) реализована в виде реляционной базы данных на основе открытой СУБД SQLite. Для обеспечения удобного доступа разработано веб приложение на основе свободного программного каркаса на языке Python – Django. Ключевым преимуществом использованного каркаса является возможность разнесения программного кода, ответственного за работу с базой данных, осуществляющего подготовку данных для показа пользователю и отвечающего за отображения информации. Архитектура разработанного веб-приложения полностью использует предоставляемые Django возможности. Часть программного обеспечения, отвечающая за работу с базой данных, была организована в виде модуля model.py; программный код, ответственный за подготовку данных для показа пользователю – в виде модуля views.py; модуль, отвечающий за отображение информации – в виде набора шаблонов веб-страниц, содержащих специальные инструкции: browse.html, browse.multiple.html, index.html, how_to_use.html, samples.html и contacts.html. Данная архитектура позволит в перспективе значительно снизить издержки на обновление и модернизацию веб-приложения.

Для сбора статистики по использованию веб-приложения использовалась система Google Analytics. Статистика собиралась за период с 1 августа по 23 ноября 2016 года. Число уникальных пользователей – 1148. Веб приложение использовали пользователи из 47 стран, среди них 202 из США, 153 из Китая, 112 из РФ, 87 из Германии, 82 из Великобритании, 59 из Франции, 54 из Японии. Пользователями было просмотрено 8773 страницы и проведено 2261 сеансов работы. В среднем пользователь просматривал около 4 страниц за сеанс.

Созданная база данных содержит подробное описание всех образцов, составляющих транскриптомную карту *A. thaliana* с микрофотографиями. Описание образцов сгруппировано по органам и доступно по ссылке http://travadb.org/samples/.

База данных позволяет искать гены как по стандартному идентификатору TAIR (в формате ATxGxxxxx), так и по тривиальным названиям, по ссылке http://travadb.org/browse/. Для выбранного гена отображаются сгруппированные по органам образцы, для каждого из которых по умолчанию приведены значения экспрессии, нормализованные по методу DESeq, усреденные по двум биологическим повторностям и деленные на максимум экспрессии для данного гена. Образцы имеют цветовое обозначение, соответствующее градиенту от минимального уровня экспрессии до максимального. База предоставляет возможность менять нормализацию на размер библиотеки с метода DESeq на TMM; отображать исходное число чтений на ген вместо нормализованного на максимум; отключать цветовое выделение.

Кроме того, существует возможность анализа дифференциальной экспрессии выбранного гена между выбранным образцом и остальными образцами карты. Для каждого образца отображается уровень изменения экспрессии относительного выбранного образца (и прочерк в случае отсутствия дифференциальной экспрессии между образцами). Для расчета дифференциальной экспрессии могут быть выбраны пакеты статистической среды R DESeq (по умолчанию), DESeq2 или baySeq. Пользователь может выбрать цветовое кодирование, основанное как на градиенте уровня экспрессии, так и на градиенте уровня изменения экспрессии.

Наконец, отображение уровней экспрессии может быть применено для нескольких генов сразу, что позволяет визуально оценивать сходство профилей их экспрессии.

Анализ генов, изменяющих уровень экспрессии в ответ на холодовое воздействие в разных частях растений

Дифференциальная экспрессия генов *A. thaliana* в ответ на холодовое воздействие была проанализирована с помощью сравнения образцов, собранных после 3 и 27 часов обработки холодом, с контрольными образцами (для 6 органов, всего 12 сравнений). Анализ проводился пакетом статистической среды R DESeq2; в качестве порогов статистически достоверной дифференциальной экспрессии были взяты значения FDR<0.05 и уровень изменения экспрессии >2. Всего было найдено 15459 гена, дифференциально экспрессирующихся хотя бы в одном из сравнений. Для более полного обнаружения дифференциальной экспрессии для тех генов, которые были распознаны как дифференциально экспрессирующиеся при таких порогах, был проведен повторный анализ экспрессии со смягченными порогами (FDR <0.25 и уровень изменения экспрессии >1.5).

Обнаруженные дифференциально экспрессирующиеся гены показывают хорошее совпадение с ранее полученными данными: до 90% совпадения списков с данными по измерению экспрессии в листьях при холодовом стрессе (Barah et al., 2013) и до 96% совпадения с данными по анализу трансгенных линий *A. thaliana* со сверхэкспрессией главных регуляторов холодового ответа генов CBF1-3 (Park et al., 2015).

Сравнение списков дифференциально экспрессирующихся генов показало крайне низкое (309 (3%) и 1083 (7%) для 3 и 27 часов обработки холодом соответственно) число генов, общих для

всех органов. Эти гены имели сильную перепредставленность категорий GO, связанных со стрессовым ответом.

4685 (45%) генов являлись орган-специфичными в ответе на стресс при 3 часах холода и 3101 (30%) при 27 часах, что показывало большое разнообразие в ответе на холод в разных частях растения. Для многих списков генов было обнаружена соответствующее органу обогащение категориями GO. Например, гены, экспрессия которых уменьшалась в ответ на стресс в листьях, были обогащены категориями, связанными с фотосинтетическими процессами, а гены, экспрессия которых увеличивалась в ответ на стресс в семенах – с биогенезом липидов.

Было показано, что многие из генов, отвечающих на холодовой стресс, не имеют соответствующей аннотации GO, что было особенно заметно для нелистовых проб (до 66% неаннотированных генов в цветках).

Из известных регуляторов раннего ответа на холод только 8 из 27 генов являлись дифференциально экспрессирующимися во всех органах после 3 часов холодового воздействия. Экспрессия остальных генов значительно отличалась между органами, причем наибольшее сходство с ранее известными данными, полученными при изучении листьев, показывали пробы листьев и листоподобных органов, в то время как цветки и семена заметно отличались.

Анализ поведения генов, экспрессия которых меняется в ответ на холодовой стресс, в транскриптомной карте позволил обнаружить группы генов, имеющих иное, нежели остальные, распределение различных параметров экспрессии. Так, например, среди генов, увеличивающих экспрессию после 3 часов холода в листьях, выделялась группа генов с низкими значениями энтропии Шеннона, являющейся мерой ширины паттерна экспрессии. Малая величина энтропии свидетельствует о специфичной экспрессии гена в небольшом числе образцов. Гены с низкой энтропией, отвечающие на стресс в листьях, в норме экспрессировались в пыльниках и участвовали в контроле развития клеточной стенки. В семенах гены имели пик распределения энтропии около 1. В число генов с энтропией, близкой к 1, входили гены, специфически экспрессирующиеся на разных стадиях развития семян и связанных с защитой растения от патогенов.

Изучение динамики экспрессии генов при холодовом стрессе и восстановлении после него

Для изучения поведения генетических сетей, контролирующих ответ на воздействие холодом при снятии стрессового воздействия, был проведен анализ экспрессии генов в листьях растений возраста 21 день под воздействием и после его снятия. В этом возрасте растения уже переходят к цветению, что позволяет избежать наложения влияния этого процесса на стрессовый ответ. Для унификации материала с целью минимизации влияния морфологического разнообразия и возраста листьев на результаты эксперимента в анализ брали только пятый лист розетки; эти листы собирались в пулы с 15 разных растений.

Были проанализированы следующие серии образцов: контрольная серия при 22C – 1, 3, 6, 12, 24 часа (за точку отсчета брался момент через 2 часа после включения света); после снижения температуры до 4C – 1, 3, 6, 12, 24 часа; и после обратного повышения температуры до 22C – 1, 3, 6, 12, 24 часа.

Анализ полученных данных показал, что состояние генетической сети возвращается к исходному через 24 часа после снятия стрессового воздействия. Однако этот возврат не заключается прекращении функционирования элементов генетических сетей, отвечающих за ответ на холодовое воздействие – происходит активация других путей регуляции. Всего в процессе адаптации к вернувшимся нормальным условиям задействованы более 4 тыс генов, не участвовавших в самом ответе на холодовой стресс. В том числе это такие гены, как гены теплового шока.

Тема 2

1. Разработана и тестирована указанная в задании модель, которая позволяет описывать процесс транскрипции и сопряжённый процесс трансляции у бактерий и в полуавтономных органеллах, ведущих своё происхождение от бактерий, – митохондриях и пластидах. Исследована зависимость числа лидерных генов от расстояния до структурного гена у многих групп бактерий. Выполнен широкомасштабный анализ лидерных генов у бактерий (коротких генов, не отмеченных в аннотациях геномов, которые не кодируют стабильных белков, имеющих самостоятельное значение).

2. Каждую цепь ДНК можно отождествить с последовательностью букв в алфавите {A, C, G, T} с известной комплементарностью. Инвертированный повтор – это участок ДНК чётной длины, начало которого комплементарно концу, а вырожденный инвертированный повтор – участок произвольной длины, близкий в метрике Левенштейна к инвертированному повтору. Расстояние Левенштейна между двумя последовательностями – минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для преобразования одной последовательности в другую. Расстояние между двумя строками длины m и n вычисляется за $O(mn)$ операций с линейной памятью $O(\min(n, m))$. Участок РНК, соответствующий вырожденному инвертированному повтору на ДНК, может образовать шпильку, в которой присутствует петля. Петля не может быть короче трёх и обычно содержит не меньше четырёх нуклеотидов. Большие петли и некомплементарность нуклеотидов плеч уменьшают стабильность шпильки. Мы классифицировали шпильки по параметрам: длине плеча, длине петли и расстоянию между одним плечом и участком, комплементарным другому плечу в метрике Левенштейна.

Геномные данные были получены из базы данных GenBank. Рассмотрены полные геномы следующих сорока видов или штаммов микобактерий: *Mycobacterium tuberculosis* H37Rv (NC_000962), *Mycobacterium leprae* TN (NC_002677), *Mycobacterium tuberculosis* CDC1551 (NC_002755), *Mycobacterium avium subsp. paratuberculosis* K-10 (NC_002944), *Mycobacterium bovis* AF2122/97 (NC_002945), *Mycobacterium* sp. MCS (NC_008146), *Mycobacterium avium* 104 (NC_008595), *Mycobacterium smegmatis* str. MC2 155 (NC_008596), *Mycobacterium ulcerans* Agy99 (NC_008611), *Mycobacterium* sp. KMS (NC_008705), *Mycobacterium vanbaalenii* PYR-1 (NC_008726), *Mycobacterium bovis* BCG str. Pasteur 1173P2 (NC_008769), *Mycobacterium* sp. JLS (NC_009077), *Mycobacterium gilvum* PYR-GCK (NC_009338), *Mycobacterium tuberculosis* H37Ra (NC_009525), *Mycobacterium tuberculosis* F11 (NC_009565), *Mycobacterium abscessus* ATCC 19977 (NC_010397), *Mycobacterium marinum* M (NC_010612), *Mycobacterium leprae* Br4923 (NC_011896), *Mycobacterium bovis* BCG str. Tokyo 172 (NC_012207), *Mycobacterium tuberculosis* KZN 1435 (NC_012943), *Mycobacterium gilvum* Spyr1 (NC_014814), *Amycolicicoccus subflavus* DQS3-9A1 (NC_015564), *Mycobacterium* sp. JDM601 (NC_015576), *Mycobacterium africanum* GM041182 (NC_015758), *Mycobacterium canettii* CIPT 140010059 (NC_015848), *Mycobacterium rhodesiae* NBB3 (NC_016604), *Mycobacterium tuberculosis* KZN 4207 (NC_016768), *Mycobacterium bovis* BCG str. Mexico (NC_016804), *Mycobacterium intracellulare* ATCC 13950 (NC_016946), *Mycobacterium intracellulare* MOTT-02 (NC_016947), *Mycobacterium intracellulare* MOTT-64 (NC_016948), *Mycobacterium tuberculosis* CCDC5180 (NC_017522), *Mycobacterium tuberculosis* CCDC5079 (NC_017523), *Mycobacterium tuberculosis* CTRI-2 (NC_017524), *Mycobacterium* sp. MOTT36Y (NC_017904), *Mycobacterium chubuense* NBB4 (NC_018027), *Mycobacterium tuberculosis* KZN 605 (NC_018078), *Mycobacterium tuberculosis* H37Rv (NC_018143), *Mycobacterium smegmatis* str. MC2 155 (NC_018289).

Независимо рассматривались кодирующие и некодирующие (межгенные) области трёх типов: между сходящимися генами, между расходящимися генами, между последовательно расположенными генами. Разработана программа, реализующая оригинальный алгоритм поиска

вырожденных инвертированных повторов и привязки их к областям генома. Программа весьма быстрая: время счёта на процессоре с двумя ядрами составило примерно 10 минут на один геном.

3. Проведена кластеризация белков из следующих геномов оригинальным методом. Геномы пластид получены из базы данных GenBank. Среди пластид апикопласты из Piroplasmida: *Babesia orientalis* strain Wuhan (NC_028029.1), *Babesia microti* strain RI (LK028575.1), *Babesia bovis* T2Bo (NC_011395.1), *Theileria parva* strain Muguga (NC_007758.1); из Coccidia: *Eimeria tenella* (NC_004823.1), *Cyclospora cayetanensis* (KP866208.1), *Toxoplasma gondii* RH (NC_001799.1); из Haemosporida: *Leucocytozoon caulleryi* (NC_022667.1) и *Plasmodium chabaudi*. Также рассмотрены недавно секвенированные пластиды из родофитной ветви, в том числе *Lepidodinium chlorophorum* (класс Dinophyceae, GenBank: NC_027093), *Choreocolax polysiphoniae* (отдел Rhodophyta, GenBank: NC_026522), *Vertebrata lanosa* (отдел Rhodophyta, GenBank: NC_026523) и *Trachydiscus minutus* (отдел Eustigmatophyceae, GenBank: NC_026851).

Многие полученные кластеры удивительно плотные. Параметры кластеризации выбирались так, чтобы кластеры соответствовали традиционным семействам белков во всех известных случаях.

Поиск промоторов осуществлён оригинальным алгоритмом, основанном на консервативности известных промоторов и экспериментальных данных о влиянии нуклеотидных замен на эффективность инициации транскрипции РНК-полимеразой бактериального типа *psbA* в пластидах горчицы.

4. Впервые рассмотрена общая модель перестроек хромосомной структуры: структуры могут иметь любое число линейных и кольцевых хромосом, неравный генный состав и содержать любое число паралогов. Получен точный почти линейный по времени алгоритм наиболее экономного (относительно заданных цен операций) преобразования одной данной хромосомной структуры в другую. Получен кубический по времени алгоритм реконструкции предковых хромосомных структур на филогенетическом дереве с произвольно заданными структурами в листьях, строящий наиболее экономный эволюционный сценарий. Программы тестировались на биологических данных для реконструкции эволюции хромосомных структур митохондрий и пластид. Полученные деревья эволюции хромосомных структур, как и предковые состояния структур, представляются адекватными. Создана и тестирована соответствующая задания программная система.

Получено важное в прикладных вопросах обобщение, когда в структурах разрешается повторение имён, что соответствует наличию в них паралогов. В силу NP-трудности задачи с повторениями нельзя найти точный полиномиальный алгоритм её решения. Однако мы показываем, как математически строго свести её к задаче целочисленного линейного программирования. Как известно, для последних задач доступны компьютерные программы, выдающие, как правило, точное решение за время близкое к линейному.

5. Разработан алгоритм поиска локусов, в которых синтения сохраняется у одних видов и нарушается у других видов; эти два непересекающихся множества видов могут быть любыми и задаются заранее. Часть проблемы состоит в эффективном определении ортологичных генов. Проведён поиск у позвоночных животных генов, потерянных или перемещённых с нарушением синтении у амниот, но сохраняющимся у лягушки и рыб. Получены весьма короткие списки таких генов.

Наряду с созданием алгоритма поиска локусов сохранения и нарушения синтении, нами начата работа по созданию алгоритма поиска высоко консервативных элементов, ранее запланированная на 2017 год.

6. Разработан алгоритм поиска высоко консервативных участков (ВКЭ) у далёких видов. Рассмотрим граф, рёбрам которого приписаны положительные веса. Вершины этого графа соответствуют участкам ДНК, рёбра соединяют участки с близкими последовательностями из разных геномов. «Близость» последовательностей подразумевает, что редакционное расстояние

между ними не превышает заданной величины *r*, т.е. одну последовательность можно получить из другой последовательным применением не более чем r элементарных операций редактирования: замены, вставки или удаления одной буквы. Если элементарные операции неравноценны, то суммарная стоимость всех операций не должна быть больше *r*.

Участки на концах любого ребра ограничиваются так, что их нельзя продолжить без того, чтобы расстояние не превысило порога *r*. При этом в качестве веса ребра используется длина этих участков (бóльшая из двух, если длины разные). Кластеры – индуцированные подграфы, которые выбираются так, чтобы внутри кластера рёбер было больше (лучший кластер – клика) и они имели больший суммарный вес, а между кластерами – меньше (в лучшем случае кластеры изолированные) и с меньшим весом. Каждый кластер соответствует набору достаточно длинных похожих участков сразу в нескольких геномах и называется высоко консервативным элементом (ВКЭ).

ВКЭ часто отвечают участкам генома, которые выполняют одинаковую функцию в разных организмах, причём во многих случаях эта функция неизвестна. Это объясняет важность нахождению ВКЭ, в том числе в сравнительно далёких друг от друга видах. Для поиска ВКЭ мы используем оригинальный метод кластеризации многодольных графов (Rubanov et al. 2016).

При поиске ВКЭ у простейших из надтипа Alveolata использованы суперкомпьютеры MVS-100K и MVS-10P Межведомственного суперкомпьютерного центра Российской академии наук http://www.jscc.ru/scomputers.shtml В целом этот пример потребовал около 200 часов с использованием 512 процессоров.

7. Рассмотрены 93 вида первичноротых животных, представляющих все типы из таксономической группы Lophotrochozoa (синоним: Spiralia), включая ортонектид (*Intoshia linei*) и дициемид (два представителя *Dicyema* spp.). Методом максимального правдоподобия (ML) реконструировано устойчивое дерево этих видов, согласующееся с известными данными об эволюции. Дерево позволило предположить близость дициемид к гастротрихам (Gastrotricha) и плоским червям (Plathelminthes), вместе образующим предполагаемую кладу Rouphozoa, но не к целомическим Lophotrochozoa (Trochozoa), таким как кольчатые черви (Annelida), моллюски (Mollusca), плеченогие (Brachiopoda) и пр. Вместе с тем различные варианты ML-реконструкции никогда не включали дициемид *внутрь* других типов, что говорит против гипотезы, что дициемиды – дивергентные представители плоских или кольчатых червей, коловраток или других известных типов и подтверждает их независимое происхождение.

Однако ML-реконструкция приводит к объединению с высоким уровнем поддержки дициемид с ортонектидами, а также с аннелидой *Myzostoma*. Такое объединение трёх видов с крайне дивергентными геномами является типичным проявлением известного эффекта «притяжения длинных ветвей» (LBA), которого в случае ML-анализов пока не удалось избежать, несмотря на добавление множества новых видов (для чего нами была выполнена сборка десятков транскриптомов из разных библиотек чтений). Эффект LBA удалось преодолеть лишь с помощью гораздо более ресурсоёмкого метода байесовской реконструкции (BI) с применением модели CAT, реализованной в программе PhyloBayes, что потребовало многомесячного счёта с привлечением пятисот вычислительных ядер суперкомпьютера МВС-10П МСЦ РАН в режиме повышенного приоритета.

В результате BI-реконструкции основные предварительные выводы о положении дициемид остались в силе: дициемиды по-прежнему уверенно располагаются вне других типов и скорее близки к Rouphozoa, чем к Trochozoa. Вместе с подтверждением того, что положение дициемид в ML-дереве не является следствием LBA, удалось получить новые результаты об эволюционном происхождении ортонектид. На BI-дереве *Intoshia* с апостериорной вероятностью 1 располагается в глубине клады Annelida. Для преодоления артефакта LBA оказалось важным наличие добавленной нами мизостомы, т. к. в её отсутствие даже модель CAT объединяет интошию с дициемами. С другой стороны, положение интошии среди аннелид *не* вызвано эффектом LBA, так

как, во-первых, она не объединяется в кладу с мизостомой, а во-вторых, при одновременном исключении мизостомы и дицием её положение в аннелидах сохраняется; остальные рассмотренные аннелиды расположены на ветвях весьма умеренной длины.

Тема 3

Согласно плану исследований по данному проекту, мы провели полногеномное исследование представителей нескольких ранее не изученных групп животных. Мы закончили анализ и опубликовали статью в высокорейтинговом научном журнале *Current biology*, описывающую полный геном представителя мало изученной группы животных – ортонектид с исключительно малой нервной системой (Mikhailov et al 2016). Ортонектиды – это редкие паразиты морских беспозвоночных, которые обычно рассматриваются в учебниках как тип неопределенного филогенического положения. Трофические формы ортонектид обитают в тканях своих хозяев в форме многоядерных плазмодиев. Плазмодии порождают червеобразных, покрытых ресничками, организмов, которые выходят в окружающую среду для копуляции. Эти эфемерные самцы и самки состоят всего из нескольких сотен клеток и лишены пищеварительной, кровеносной и выделительной систем. Со времени открытия в XIX веке ортонектиды считались частью мезозой, предполагаемой переходной группы между многоклеточными животными и их одноклеточными родственниками. Совсем недавно эта точка зрения была нами оспорена, и наши данные свидетельствуют, что ортонектиды – это животные, которые упростились в связи с паразитическим образом жизни.

Мы исследовали уникальную стадию жизненного цикла ортонектид - многоядерный плазмодий. Получить РНК «чистого» изолированного плазмодия технически невозможно. Плазмодий разветвлен в теле хозяина так, что их ткани фактически перемешаны. Для наших исследований мы применили специальный приём. РНК была выделена из свободных, покинувших хозяина особей ортонектиды и из целого зараженного хозяина. В последнем случае мы получили смесь РНК хозяина и паразита. Поскольку геном паразита нами уже прочитан, можно отфильтровать только те транскрипты, что относятся к паразиту. Такие опыты уже поставлены и позволят провести анализ генов, дифференциально работающих в этих двух радикально различающихся стадиях.

Продолжая согласно плану исследований полногеномное исследование представителей нескольких ранее не изученных групп животных, мы провели высокопроизводительное чтение ДНК двух видов представляющих тип кининохов. В настоящий момент мы полностью собрали митохондриальные геномы этих двух видов и опубликовали статью (Popova et al 2016).

Впервые прочитаны и аннотированы полные митохондриальные геномы пяти представителей малоизученной группы животных – волосатиков с уникальными нуклеотидными палиндромными последовательностями в кодирующих генах.

Палиндромная последовательность ДНК (РНК) совпадает с последовательностью, комплиментарной самой себе, записанной в обратном направлении. Палиндром ДНК (РНК) способен формировать шпильку. Мы обнаружили необычные точные длинные палиндромы (до ~250 нт.) внутри кодирующих белки последовательностей митохондриальной ДНК червей волосатиков (Nematomorpha). Такие палиндромы были найдены в семи типах митохондриальных генов четырёх видов волосатиков, и существование этих палиндромов было подтверждено несколькими экспериментальными методами. Белки, кодируемые такими последовательностями, консервативны, что предполагает, что отбор одновременно сохраняет основу их аминокислотной последовательности и в то же время создаёт и поддерживает палиндромы в кодирующей их ДНК. Ещё более удивительно, что некоторые белок кодирующие ДНК волосатиков содержат не один, а два перекрывающихся палиндрома. Насколько нам известно, ничего подобного ранее не наблюдалось ни в митохондриальных, ни в ядерных генах. Наше открытие предполагает возможность существования нового неизвестного механизма геномной регуляции у животных. Работа была доложена на конференции и статья находится на стадии написания.

Тема 4

В рамках задачи 4.2 завершена разработка метода анализа глобального эпистаза по распределению частот вредных аллелей в геноме. Метод применен к популяционно-геномным данным *Drosophila melanogaster*. Показано, что отрицательный отбор в сочетании с синергическим эпистазом должен приводить к отрицательному неравновесию по сцеплению между вредными аллелями, и к пониженной дисперсии распределения числа вредных аллелей в геноме. В соответствии с этими ожиданиями, число редких аллелей потери функции в геномах *Drosophila melanogaster* обладает более низкой дисперсией, чем ожидается при независимом влиянии вредных мутаций на приспособленность. Этот эффект усиливается, если рассматривать только локусы, находящиеся под более сильным отбором; отбор мы оценивали, используя отношение числа несинонимических и сионимических межвидовых различий на сайт (dN/dS) и прямые данные по незаменимости (essentiality) генов. Путем моделирования популяций с учетом популяционной структуры мы показали, что структурированность популяций и техническая гетерогенность при секвенировании также могут отклонять дисперсию от ожидаемой, однако в этом случае ожидается повышение дисперсии по сравнению с независимостью, то есть противоположный эффект. Мы также рассмотрели альтернативные объяснения пониженной дисперсии: стабилизирующий отбор и интерференцию Хилла-Робертсона, и показали, что эти объяснения маловероятны. Действительно, стабилизирующий отбор предполагает, что оптимальным является ненулевое число аллелей, снижающих приспособленность, в геноме, а интерференция Хилла-Робертсона возможна только между сцепленными локусами, в то время как часть наблюдаемого сигнала приходит с разных хромосом. Таким образом, наблюдаемое занижение дисперсии совместимо только с действием синергического эпистаза. Эти результаты являются первым свидетельством эпистатических взаимодействий между мутациями на уровне полного генома, и могут объяснить способность видов выживать несмотря на высокую скорость возникновения вредных мутаций в геноме.

В рамках задачи 4.4 мы исследовали локальный эпистатический отбор, то есть отбор, действующий не на уровне полного генома, а на уровне отдельных функциональных элементов, у *D. melanogaster*. В соответствии с планом работы, мы использовали полногеномные данные по делеционным полиморфизмам и по однонуклеотидным полиморфизмам (ОНП) внутри сегментов ДНК, содержащих полиморфные делеции, чтобы оценить силу эпистатического отбора. Спектры аллельных частот свидетельствуют о том, что отбор против делеции усиливается с длиной делеции, что не удивительно: более длинная делеция чаще захватывает функциональный участок ДНК. Однако вредность делеции в среднем гораздо ниже, чем можно ожидать из данных по ОНП, и этот контраст выражен сильнее для более длинных делеций. Другими словами, удаление всего функционального элемента менее вредно, чем ожидается из произведения эффектов ОНП на приспособленность. Таким образом, ОНП, расположенные вблизи друг от друга, участвуют в эпистатических взаимодействиях, причем этот эпистаз антагонистический. Это может означать, что функциональный элемент действует как единое целое, и любая точечная мутация в нем нарушает его функцию в той же степени, что и делеция всего элемента, так что последующие дополнительные мутации не приводят к дополнительному снижению приспособленности. Межвидовая дивергенция нуклеотидов в пределах сегментов полиморфных делеций и в их ближайшей окрестности приближается к селективной нейтральности с ростом аллельной частоты делеции. Это означает, что нуклеотиды, вложенные в высокочастотных делециях, не только испытывают более слабый отрицательный отбор, но и вовлечены в более слабые эпистатические взаимодействия.

В рамках задачи 4.6 продолжен анализ нонсенс-аллелей (последовательностей, содержащих стоп-кодоны) в белок-кодирующих генах *D. melanogaster*. Вдобавок к результатам, полученным в прошлом году, мы проаннотировали полиморфные нонсенс-аллели в 196 природных линиях *D. melanogaster (*замбийские популяции). Отношение числа несинонимических и синонимических ОНП на сайт (pN/pS) среди мутаций, расщепляющихся только в нонсенс-аллелях («вложенных» мутаций), было высоким, что свидетельствует об ослаблении отбора в генотипах, несущих такие

аллели. В одноэкзонных генах, а также в многоэкзонных генах для ОНП в том же экзоне, что и нонсенс-мутация, pN/pS≈1, что свидетельствует о потери функции этих аллелей. Однако в других экзонах многоэкзонных генов оно было ниже (pN/pS≈0,5). Различие между экзонами не связано с альтернативным сплайсингом нонсенс-содержащих экзонов, так как оно наблюдается и для конститутивно сплайсируемых экзонов. Таким образом, оно свидетельствует о роли рекомбинации: в более удаленных экзонах понижение pN/pS связано с рекомбинацией с аллелями дикого типа. Анализ связи между частотой нонсенс-аллеля и pN/pS позволяет определить возраст нонсенс-мутации.

В рамках задачи 4.7 …

Тема 5

Продолжена работа по теме «Изучение изменчивости скорости и паттернов мутагенеза в геноме человека». В рамках этой темы была исследована роль белка APOBEC3A/B в накоплении наследственных мутаций. Ранее было известно, что цитидиновая деаминаза APOBEC3A/B отвечает за большинство соматических мутаций во многих образцах опухолевой ткани; однако её роль в накоплении наследственных мутаций оставалась неясной. Использовав данные по низкочастотным ОНП человека, межвидовой дивергенции и de novo мутациям, мы исследовали мутации в нуклеотидных контекстах, подверженных атаке APOBEC3A/B. Мы показали, что такие мутации чаще возникают на той цепи ДНК, которая является запаздывающей в ходе репликации. Кроме того, APOBEC3A/B-подобные мутации часто возникают кластерами, которые также более часты на запаздывающей цепи. Эти результаты означают, что около 20% мутаций C->T и C->G, расщепляющихся как ОНП в популяции человека, возникли в результате действия APOBEC3A/B.

Кроме того, в рамках этой темы исследована асимметрия, связанная с действием системы репарации ошибочно спаренных оснований (mismatch repair, MMR). MMR – это одна из основных систем, определяющих точность репликации ДНК. Мы использовали данные секвенирования образцов опухолевой ткани из пациентов с функционирующей (MSS) и нарушенной (MSI) системой MMR, чтобы исследовать свойства MMR у человека. В MSI-образцах (но не в MSS-образцах) мы выявили неравномерность скоростей мутирования между лидирующей и запаздывающей цепями. Направлении асимметрии между цепями в MSI-образцах соответствует направлению асимметрии в опухолевых образцах с мутированным экзонуклеазным доменом полимеразы δ, что означает, что полимераза δ, преимущественно реплицирующая запаздывающую цепь, вносит больше мутаций, чем полимераза ε, преимущественно реплицирующая лидирующую цепь. Это означает, что мутации, возникающие в нормальных клетках при репликации запаздывающей цепи, репарируются MMR в ~3 раза эффективнее, чем возникающие на лидирующей цепи.

1.9.2.4. Сведения о достигнутых конкретных научных результатах в отчетном году (до 5 стр.)

Тема 1

Разработана общедоступная база данных для анализа уровней экспрессии генов у модельного объекта генетики растений *Arabidopsis thaliana* (http://travadb.org). База включает ряд инструментов анализа, в т.ч. анализ дифференциальной экспрессии между образцами разными методами, сопоставление уровня изменения экспрессии и др. За 4 месяца созданной базой воспользовалось более 1000 уникальных пользователей, которыми сделано более 8 тысяч поисковых запросов.

Анализ процессов ответа на воздействие холодом в разных типах органов показал, что существенная часть происходящих процессов органоспецифична, что показывает разные стратегии адаптации к холоду.

Показано, что процесс восстановления нормального функционирования растений после холодового стресса является отдельным процессом, схожим по ряду параметров с происходящим при тепловом шоке.

Тема 2

1. Лидерные гены обычно не кодируют стабильных белков, хотя их важная роль в регуляции экспрессии бактериальных геномов широко признана; часто такие гены вовлечены в аттенюаторную регуляцию. Обнаруженное нами обилие лидерных генов позволило нам предположить, что у бактерий их роль не ограничивается регуляцией. Действительно, в случае малого расстояния между стоп-кодоном лидерного гена и инициирующим кодоном структурного гена, когда трудно предположить их регуляторную роль, мы предположили, что лидерный ген увеличивает уровень экспрессии структурного гена в результате реинициализации рибосом с лидерного пептида. Например, у актинобактерий частота расположения лидерного гена на расстоянии 10–11 п.н. примерно на 70% выше, чем средняя частота на расстояниях от 1 до 65 п.н., и плавно падает после 65 п.н. Выраженный пик такой зависимости частоты от расстояния наблюдается у протеобактерий, Bacteroidetes, Spirochaetales, Acidobacteria, группы Deinococcus–Thermus и Planctomycetes. Напротив, у Firmicutes такой пик приходится на расстояние 15–16 п.н. и слабо выражен, а у цианобактерий и тенерикутов пик отсутствует. Однако в целом такой пик характерен для многих бактерий.

Структурные гены, перед которыми обнаружены лидерные гены на расстоянии 10–11 п.н., содержат разнообразные домены, определяемые по базе данных Pfam. Нет оснований выделить тип белков, кодируемых структурными генами с такими лидерными генами, или систематически связать такую пару с регуляцией, зависящей от концентрации какой-то аминокислоты. Хотя в некоторых случаях лидерный ген, расположенный на малом расстоянии, может участвовать в регуляции. А именно, задержка рибосомы на регуляторных кодонах приводит к снижению частоты реинициализации рибосом. Мы предполагаем, что лидерные гены, расположенные так близко к структурным генам, входят в единый с ними оперон и служат для увеличения частоты инициации трансляции белка, кодируемого структурным геном. Здесь рибосома может инициировать трансляцию непосредственно или в результате реинициализации после завершения трансляции лидерного пептида. Таким образом, лидерная область играет роль “антенны” на полицистронной мРНК, увеличивающей частоту инициации трансляции белка, кодируемого структурным геном.

В целом большое число лидерных генов без явно выраженной концентрации регуляторных кодонов редких аминокислот служит основанием предполагать, что большинство из них не участвует в регуляции, зависимой от концентрации аминокислот или аминоацил-тРНК синтетаз. Обычно в случаях регуляции экспрессии структурных генов в зависимости от концентрации какой-то аминокислоты расстояние между лидерным и структурным генами колеблется в широком диапазоне и при достаточном расстоянии существенную роль играет вторичная структура РНК, как показано при нашем моделировании аттенюаторной регуляции и в экспериментах по разрушению мутацией спирали РНК. Расстояние между лидерным и структурным генами должно быть достаточно большим для формирования на мРНК терминатора транскрипции за пределами области, перекрываемой рибосомой. Если учесть, что рибосома закрывает до 12 нуклеотидов ниже стоп-кодона, а терминатор не перекрывает сайт связывания рибосомы перед структурным геном, то расстояние между лидерным и структурным генами должно составлять не менее 40 п.н.

Малое количество лидерных генов на расстояниях 6–9 п.н. объясняется тем, что этот участок перекрывается областью Шайна–Дальгарно, богатой пуринами, что противоречит присутствию здесь любого из стоп-кодонов, содержащего пиримидин. Этот эффект проявляется для всех стоп-кодонов.

В экспериментах с митохондриями гороха показана важная роль в инициации транскрипции участка из девяти нуклеотидов CRTAAGAGA, где R обозначает из двух нуклеотидов A или G. Анализ митохондриальной ДНК (NC_016740.1) из финика пальчатого (*Phoenix dactylifera*) выявил потенциальный промотор CATAAGAtA (начиная с позиции 83965), где строчные буквы отмечают несоответствие экспериментально изученному промотору в митохондриях гороха. Этот промотор расположен после гена РНК-полимеразы на комплементарной цепи и может инициировать транскрипцию в антисмысловом направлении. Координаты этого гена complement(94571..97198). Промотор может регулировать экспрессию гена РНК-полимеразы. Действительно, если РНК-полимераз достаточно много, их синтез должен сокращаться. Возможны два механизма: прерывание транскрипции из-за столкновений РНК-полимераз до завершения транскрипции своего гена, кодирующего РНК-полимеразу; или образование дуплекса из двух комплементарных РНК, препятствующих трансляции. Столкновение РНК-полимераз фагового типа не обязательно приводит к терминации транскрипции (Ma, McAllister 2009), но в любом случае экспрессия гена снижается.

2. Рассмотрено 187759 межгенных областей, из них 123737 последовательных, 31994 сходящихся и 32028 расходящихся у сорока геномов микобактерий. Большое число вырожденных инвертированных повторов, соответствующих шпилькам с длиной петли в четыре нуклеотида, позволяет предположить, что значительная доля повторов соответствуют шпилькам на РНК. Две зависимости расстояния между шпилькой и ближайшим геном от параметров шпильки существенно различаются между собой для двух типов межгенных областей. Это говорит о различной роли шпилек в зависимости от типа области. Они играют регуляторную роль в экспрессии генов или служат для стабилизации транскриптов, располагаясь на 3'-конце РНК. Среднее расстояние от шпильки до ближайшего гена также зависит от типа области и составляет около 100 п.н. для областей между сходящимися фланкирующими генами и около 70 п.н. для областей между расходящимися фланкирующими генами.

Для последовательных генов заметен рост расстояния между шпильками всех размеров и началом гена при увеличении петли. С увеличением петли растет расстояние между короткими шпильками и концом гена. Для петель в 6-8 нуклеотидов среднее расстояние между короткими шпильками и началом гена становится больше, чем между началом гена и длинными или средними шпильками. В областях между расходящимися генами шпильки находятся ближе к генам, чем в областях между сходящимися генами. В областях между последовательными генами короткие шпильки в среднем находятся ближе к началам генов, а средние и длинные – ближе к концам. Средние длины некодирующих областей между последовательно расположенными и сходящимися генами приблизительно равны для разных таксономических групп, а некодирующие области между расходящимися генами в среднем значительно длиннее у всех групп. В то же время среднее расстояние до ближайшего гена более короткое именно для расходящихся генов; этот эффект нельзя объяснить простым увеличением длин некодирующих областей.

Определены интервалы типичных значений параметров шпилек и расстояний от них до ближайших генов у микобактерий. Полученные результаты могут служить основой для дальнейшего предсказания регуляции экспрессии генов. Также полученные результаты могут быть использованы для предсказания частоты хромосомных перестроек, в результате которых возникают инвертированные повторы. Это позволяет уточнить ранее рассмотренную модель эволюции генома.

3. Выполнена кластеризация белков, кодируемых в пластидах родофитной ветви. Результаты собраны в базу данных, позволяющую проводить поиск по филогенетическому профилю и по аминокислотной последовательности. На основе полученных результатов построено дерево апикопластов (пластид простейших из типа Apicomplexa).

В целом распределение по кластерам белков, кодируемых в пластидах багрянок, показывает значительное отличие *Porphyridium purpureum* от остальных видов, что сопровождается многочисленными перестройками ДНК пластид этих водорослей, а также значительное

обособление клады из трёх видов *Galdieria sulphuraria, Cyanidium caldarium* и *Cyanidioschyzon merolae*.

По сравнению с нашими предыдущими результатами кластеры белков MoeB и Ycf28 одновременно пополнились белками, кодируемыми в пластидах *Vertebrata lanosa*; у *Choreocolax polysiphoniae* и всех рассмотренных видов вне отдела Rhodophyta ни один из этих белков не кодируется в пластидах. Так подтверждено ранее отмеченное нами совпадение филогенетических профилей этих белков.

Белок Ycf28 имеет значительное сходство с транскрипционным фактором NtcA цианобактерий. Поэтому мы предполагаем, что именно Ycf28 регулирует в пластидах транскрипцию гена *moeB,* связывая ДНК вблизи промотора, где нами предсказан консервативный мотив. Нет оснований считать, что Ycf28 связан с метаболизмом азота, то есть по сравнению с цианобактериями произошла смена специфичности транскрипционного фактора к субстрату. Отсутствие типичного -35 бокса промотора перед геном *moeB*, говорит о том, что Ycf28 является активатором транскрипции.

Транскрипционный регулятор Ycf29 кодируется в пластидах криптофитовых водорослей и багрянок кроме *Porphyridium purpureum*. Этот белок кодируется у *Calliarthron tuberculosum* (YP_007878178.1), *Chondrus crispus* (YP_007627336.1), *Choreocolax polysiphoniae* (YP_009122074.1), *Cryptomonas paramecium* (YP_003359295.1), *Cyanidioschyzon merolae* (NP_849011.1), *Cyanidium caldarium* (NP_045122.1), *Galdieria sulphuraria* (YP_009051025.1), *Gracilaria salicornia* (YP_009019567.1), *Gracilaria tenuistipitata* (YP_063559.1), *Grateloupia taiwanensis* (YP_008144796.1), *Guillardia theta* (NP_050668.1), *Porphyra purpurea* (NP_053953.1), *Pyropia haitanensis* (YP_007947873.1), *Pyropia perforate* (YP_009027627.1), *Pyropia yezoensis* (YP_537024.1), *Rhodomonas salina* (YP_001293481.1), *Teleaulax amphioxeia* (YP_009159161.1), *Vertebrata lanosa* (YP_009122313.1).

Другого белка с таким филогенетическим профилем не найдено. Близкий профиль имеет белок CemA, который есть у *Porphyridium purpureum*, но отсутствует у *Choreocolax polysiphoniae*. Белок CemA содержит домен PF03040 и найден на внутренней стороне наружной мембраны хлоропластов, но не в тилакоидной мембране. Ортологичные CemA белки у цианобактерий вовлечены в транспорт углекислого газа, но сами не являются транспортёрами. Другой белок с близким филогенетическим профилем – это мембранный белок Ycf19. Перед геном *ycf19* предсказан консервативный промотор бактериального типа, близкий к консенсусу. С другой стороны, присутствие Ycf29 у нефотосинтезирующих видов *Cryptomonas paramecium* и *Choreocolax polysiphoniae* свидетельствует о том, что он регулирует процессы, не связанные с фотосинтезом. Поскольку Ycf29 входит в двухкомпонентную систему передачи сигнала, его регулон связан с реакцией на изменения внешних условий, а не внутри пластиды.

Известно, что в пластидах многих видов водорослей кодируется транскрипционный фактор Ycf30, регулирующий экспрессию генов *rbcLS*, кодирующих субъединицы рибулозобисфосфаткарбоксилазы (КФ 4.1.1.39), и гена *cbbX*. Экспериментально показана индуцируемая светом активация транскрипции этих генов в изолированных *Cyanidioschyzon merolae* и предсказан сайт связывания фактора Ycf30 в пластидах багрянок. Построенные нами филогенетические профили этих белков согласуются с этими предсказаниями. Однако низкая консервативность сайта связывания Ycf30 не позволила точно предсказать его положение на ДНК.

Консервативный сайт найден в 5’-нетранслируемых областях генов *ycf24* (*sufB*) у *Eimeria tenella, Cyclospora cayetanensis, Toxoplasma gondii RH, Leucocytozoon caulleryi, Plasmodium chabaudi* и *Porphyra purpurea*. Вероятно, этот сайт вовлечён в регуляцию экспрессии гена на уровне трансляции.

Анализ вторичных структур некодирующих областей мРНК из пластид родофитной ветви говорит об отсутствии консервативных структур. Можно предположить, что регуляция экспрессии генов основана взаимодействии белков с другими белками и нуклеиновыми кислотами.

4. Получен точный почти линейный по времени алгоритм наиболее экономного (относительно заданных цен операций) преобразования одной данной хромосомной структуры в другую. Получен кубический по времени алгоритм реконструкции предковых хромосомных структур на филогенетическом дереве с произвольно заданными структурами в листьях, строящий наиболее экономный эволюционный сценарий. Программы тестировались на биологических данных для реконструкции эволюции хромосомных структур митоходрий и пластид. Полученные деревья эволюции хромосомных структур, как и предковые состояния структур, представляются адекватными.

5. У лягушки *Xenopus tropicalis* существуют два гена ras-dva и ras-dva-2 (http://www.ncbi.nlm.nih.gov/gene/497008 и http://www.ncbi.nlm.nih.gov/gene/733456). У человека соответствующих генов нет. У ящерицы *Anolis carolinensis* есть гомолог rap2b (малая ГТФаза Ras-dva). У курицы *Gallus gallus* есть гомолог RASL10A (Ensembl:ENSGALG00000021552 или http://www.ncbi.nlm.nih.gov/gene/417349) У *Danio rerio* рядом с геном zgc:152698 (Ensembl:ENSDARG00000058464) расположены гены ndufab1a, palb2, plk1, ern2. У *Xenopus tropicalis* рядом с геном ras-dva расположены гены palb2, dctn5. У *Xenopus tropicalis* рядом с геном ras-dva-2 расположены гены tnrc6c, sec14L5. У курицы рядом с геном RASL10A расположены гены TNRC6C, SEC14L1, MIR6516. Можно сделать вывод о том, что при определении синтении полезно смотреть микроРНК. На основании синтении можно предположить, что ген ras-dva-2 лягушки сохранился у курицы и соответствует гену RASL10A. Напротив, ген ras-dva лягушки наследован от рыб и не имеет ортологов у амниот.

6. На основе результатов поиска ВКЭ у простейших из надтипа Alveolata программой RAxML построено филогенетическое древо.

7. Нами получено: дициемиды уверенно располагаются вне других типов и скорее близки к Rouphozoa, чем к Trochozoa. Положение дициемид в ML-дереве не является следствием LBA, а также – на BI-дереве *Intoshia* с апостериорной вероятностью 1 располагается в глубине клады Annelida. Это положение сохраняется и при исключении из анализа дивергентной аннелиды *Myzoatoma*, что позволяет утверждать, что этот вывод не является следствием LBA; остальные рассмотренные аннелиды расположены на ветвях весьма умеренной длины.

Тема 3

Впервые прочитан и опубликован полный геном представителя мало изученной группы животных – ортонектид с исключительно малой нервной системой (Mikhailov et al 2016). Этот геном полностью аннотирован и доступен в базе данных NCBI (https://www.ncbi.nlm.nih.gov/genome/?term=txid33209[Organism:exp], https://www.ncbi.nlm.nih.gov/bioproject/?term=txid33209[Organism:exp])

Для изучения особенностей уникальной паразитической стадии ортонектид – синцитиального плазмодия – мы исследовали транскриптом этой стадии и сравнили с транскриптомом свободноживущих особей. В результате будут найдены и проанализированы гены ортонектид, специализированные для функционирования плазмодия.

Впервые прочитаны и опубликованы полные митохондриальные геномы двух представителей мало изученной группы животных – киноринх (Popova et al 2016). Эти геномы полностью аннотированы и доступны в базе данных NCBI (https://www.ncbi.nlm.nih.gov/genome/50390 , https://www.ncbi.nlm.nih.gov/genome/50343).

В настоящее время проведен предварительный анализ ядерных геномов этих видов.

Впервые прочитаны и аннотированы полные митохондриальные геномы пяти представителей мало изученной группы животных – волосатиков с уникальными нуклеотидными палиндромными последовательностями в кодирующих генах. Работа находится на стадии написания статьи.

Мы осуществили широкомасштабное секвенирование геномной ДНК четырёх представителей разных видов волосатиков и транскриптома одного из этих видов. Получены сборки геномов этих видов и предварительная аннотация их генов. Эти данные позволяют находить гены, специфичные для волосатиков и неизвестные у других групп организмов.

Тема 4

В рамках задачи 4.1 (в основном завершенной в 2015 году; результаты находятся на повторном рецензировании; препринт: http://biorxiv.org/content/early/2016/04/07/047274) полученные ранее результаты по эпистатическим взаимодействиям в белках митохондрий применены к предсказанию патогенных вариантов человека. Анализ филогений, включающих несколько тысяч видов многоклеточных животных, показал, что мутации митохондриальных генов, патогенные для человека, часто являются допустимыми для других видов, по-видимому – из-за эпистатических взаимодействий с другими позициями тех же генов или же с другими генами. Вторая статья по задаче 4.1, описывающие эти результаты, готовится к публикации.

В рамках задачи 4.2 впервые выявлено действие отрицательного (синергического) эпистаза на масштабах полного генома. Показано, что результаты не объясняются структурированностью популяции или корреляциями в ошибках секвенирования. Рукопись статьи, описывающая эти результаты наряду с аналогичными результатами по популяциям человека, находится на повторном рецензировании в журнале *Science* (препринт: http://biorxiv.org/content/early/2016/07/29/066407).

В рамках задачи 4.4 выявлено отклонение от независимости между ОНП и инделами в одних и тех же сегментах генома. Полученные результаты объяснимы действием антагонистического эпистаза на уровне функциональных элементов. Результаты работы готовятся к публикации.

В рамках задачи 4.6 показано, что нонсенс-мутация в геноме «выключает» отбор на нуклеотидную последовательность того экзона, который содержит эту мутацию, так что дальнейшее накопление несинонимических и синонимических мутаций в этом экзоне происходит с равными скоростями. Разработан метод анализа возраста нонсенс-аллелей по частотам сцепленных с ними миссенс-аллелей. Результаты работы готовятся к публикации.

В рамках задачи 4.7 …

Тема 5

В рамках задачи 5 выявлена значительная роль цитидиновой деаминазы APOBEC3A/B в наследуемом мутагенезе у человека. Результаты работы находятся на повторном рецензировании (препринт: http://biorxiv.org/content/early/2016/05/19/054197).

Кроме того, в рамках задачи 5 выявлена более высокая эффективность системы MMR в репарации ошибок, возникших при репликации запаздывающей цепи. Результаты этой работы также находятся на повторном рецензировании (препринт: http://biorxiv.org/content/early/2016/03/23/045278).

3.2.1. План работы на год *(в том числе указываются запланированные командировки по проекту), до 5 стр.*

Тема 1

1. Поиск новых сайтов сплайсинга, с использованием ранее полученных данных. Будет проведена разработка критериев, направленных на их идентификацию, и оценка возможного числа

нераспознанных сайтов. Будет проведена оценка консервативности числа сайтов сплайсинга в генах разных классов.

2. Будет проведен анализ изменения экспрессии генов при недавней тетраплоидизации на примере близкого родственника *Arabidopsis* – *Capsella bursa-pastoris* (пастушьей сумки). Будет проведен анализ связи изменения в уровне экспрессии паралогичных генов с изменениями в промоторных участках.

Тема 2

1. Определение величины отклонения сайтов связывания рибосомы от стандартного консенсуса GGAGGA. Определение спиралей РНК, расположенных в 5'-лидерных областях генов и перекрывающих сайт связывания рибосомы, выявление наиболее консервативных спиралей. Определение альтернативных к консервативным спиралям вторичных структур РНК: T-боксов, LEU- и LEU1-элементов, других РНК-переключателей, G-квадруплексов. Перечисленные работы будут выполнены для актинобактерий.

2. Дальнейшее развитие модели взаимодействия РНК-полимераз для описания трансляции белков, пост-транскрипционных (деградация мРНК) и пост-трансляционных (взаимодействие с другими белками - компонентами систем передачи сигнала) преобразований с учётом вторичных структур. Разработка эффективной параллельной программной реализации модели для суперкомпьютерных систем. Тестирование модели.

3. Вычисление скорости элонгации РНК-полимераз фагового типа в пластидах растений на основании сопоставления размеров экзонов и интронов в генах пластид.

4. Построение полного эволюционного сценария для известных видов Rhizobiales с реконструкцией предковых событий дупликаций, потерь и возникновений генов и семейств генов.

Тема 3

Продолжение полногеномного исследования представителей нескольких ранее не изученных групп животных. Мы завершим описание и анализ открытых нами уникальных палиндромов в кодирующих областях митохондриальных генов волосатиков. На основе проведенной нами сборки митохондриальных геномов четырех различных видов волосатиков мы проведем анализ роли и механизмов этого явления и подготовим и опубликуем статью по этой теме.

Для выяснения роли найденных нами в митохондриальных геномах волосатиков уникальных палиндромов мы планируем проведение поведенческих, биохимических и молекулярно-биологических экспериментов на этих животных.

В опубликованной нами в 2016 году работе мы детально проанализировали гены ортонектид, связанные с функционированием и развитием нервной системы. Наши исследования указывают на то, что этот организм обладает необычайно простым мозгом и набором генов, вовлеченных в работу нервной системы. Наши новые морфологические и поведенческие эксперименты будут направлены на изучение того, насколько на самом деле упрощена нервная система ортонектид и насколько сложное поведение может демонстрировать животное при таком малом числе нервных клеток.

Мы продолжим сравнение последовательностей кДНК из разных стадий исследуемых нами организмов. Ортонектиды имеют две очень разные жизненные стадии. Одна из них нами детально изучена, и работа по этой теме подготовлена к печати. Другая уникальная стадия жизненного цикла ортонектид – это многоядерный плазмодий. Для изучения его особенностей мы исследуем

полученные нами транскриптомы этой стадии и сравним с транскриптомомами свободноживущих особей.

Разработка оптимизированного и более эффективного алгоритма для поиска высоко консервативных элементов (ВКЭ), эффективно применимого к группам с большим филогенетическим расстоянием между видами и наборам из тысяч геномов; его реализация для суперкомпьютера. Программный комплекс будет работать на суперкомпьютерах с распределённой памятью и большим числом процессоров (вычислительных кластерах). Программный комплекс дополнительно будет классифицировать ВКЭ по их расположению относительно кодирующих областей ДНК. Применение полученного комплекса для поиска ВКЭ в пластидах багрянок.

Данные будут получены из базы данных GenBank. В том числе следующие геномы пластид багрянок. Из Bangiophyceae: *Bangia atropurpurea* (NC_030221), *Cyanidioschyzon merolae* (NC_004799, *Cyanidium caldarium* (NC_001840), *Galdieria sulphuraria* (NC_024665), *Porphyra pulchra* (NC_029861), *Porphyra purpurea* (NC_000925), *Porphyra yezoensis* (NC_007932), *Porphyridium purpureum* (NC_023133), *Porphyridium sordidum* (NC_031175), *Pyropia haitanensis* (NC_021189), *Pyropia perforata* (NC_024050), *Wildemania schizophylla* (NC_029576); из Compsopogonophyceae: *Erythrotrichia carnea* (NC_031176), *Rhodochaete parvula* (NC_031180); из Florideophyceae: *Ahnfeltia plicata* (NC_031145), *Apophlaea sinclairii* (NC_031172), *Asparagopsis taxiformis* (NC_031148), *Calliarthron tuberculosum* (NC_021075), *Ceramium cimbricum* (NC_031211), *Ceramium japonicum* (NC_031174), *Chondrus crispus* (NC_020795), *Choreocolax polysiphoniae* (NC_026522), *Coeloseira compressa* (NC_030338), *Dasya binghamiae* (NC_031161), *Gelidium elegans* (NC_029858), *Gelidium vagum* (NC_029859), *Gracilaria chilensis* (NC_029860), *Gracilaria chorda* (NC_031149), *Gracilaria salicornia* (NC_023785), *Gracilaria tenuistipitata* (NC_006137), *Gracilariopsis lemaneiformis* (NC_029644), *Grateloupia taiwanensis* (NC_021618), *Hildenbrandia rivularis* (NC_031177), *Hildenbrandia rubra* (NC_031146), *Kumanoa americana* (NC_031178), *Mastocarpus papillatus* (NC_031167), *Palmaria palmata* (NC_031147), *Plocamium cartilagineum* (NC_031179), *Rhodymenia pseudopalmata* (NC_031144), *Schimmelmannia schousboei* (NC_031168), *Schizymenia dubyi* (NC_031169), *Sebdenia flabellata* (NC_031170), *Sporolithon durum* (NC_029857), *Thorea hispida* (NC_031171), *Vertebrata lanosa* (NC_026523), из Stylonematophyceae: *Bangiopsis subsimplex* (NC_031173).

Определение оптимальных параметров программ для широкомасштабного поиска ВКЭ у вышеуказанных видов. Предсказание потенциальных регуляторных участков ДНК, включая промоторы и сайты связывания транскрипционных факторов, в пластидах багрянок. Создание общедоступной базы данных ВКЭ пластидах багрянок по адресу http://lab6.iitp.ru/.По найденным ВКЭ будет построено дерево, описывающее их эволюцию с использованием программы RAxML [Stamatakis A. // Bioinformatics. 2014. V. 30, no. 9. P. 1312–1313]. Для этого используется матрица, элементы которой равны 1 или 0 и указывают на присутствие или отсутствие у данного вида представителя данного ВКЭ; число строк равно числу видов, число столбцов равно числу найденных ВКЭ. При построении дерева используется модель двоичных подстановок (binary substitution model), число бутстрэп-реплик равно 300.

Создание алгоритма поиска ультраконсервативных участков в структуре геномов и сцепленных с ними уникальных локусов. Программная реализация алгоритма для многопроцессорных систем. Тестирование алгоритма на искусственных и биологических данных. Поиск ультраконсервативных участков в геномах многоклеточных. Определение последовательностей уникальных локусов, сцепленных с ультраконсервативными участками. На этой основе будет получен набор данных для реконструкции филогении многоклеточных.

Тема 4

В соответствии с исходным планом работ, в 2017 году будут выполняться работы по следующим задачам, описанным в разделах 4.7 и 4.10 заявки.

4.3. Анализ локального эпистаза по распределению числа вредных аллелей на функциональный элемент.

4.7. Анализ глобального отрицательного эпистаза по недопредставленности аллельных пар в высокополиморфном грибе *Schizophyllum commune.*

Кроме того, предполагается работа по дальнейшему развитию проектов, в основном завершенных в 2015–2016 годах. В рамках задачи 4.1 «Анализ эпистаза по филогенетической кластеризации гомоплазий» планируется с учетом ранее полученных результатов использовать филогенетическую неравномерность гомоплазий для анализа изменения сайт-специфических адаптивных ландшафтов. В рамках задачи 4.2 «Анализ глобального эпистаза по распределению числа вредных аллелей на геном» мы проанализируем отбор против различных видов стоп-кодонов у *D. melanogaster*.

Также будет продолжена работа в рамках темы 5 «Изучение изменчивости скорости и паттернов мутагенеза в геноме человека».

Предполагается ряд командировок на конференции для представления полученных результатов.

3.2.2. Ожидаемые в конце года конкретные научные результаты *(форма изложения должна дать возможность провести экспертизу результатов и оценить степень выполнения заявленного плана работы), до 5 стр.*

Тема 1

1. Будет идентифицированы новые сайты сплайсинга и проведен их анализ. Будет проанализирована представленность изоформ генов в разных частях *Arabidopsis thaliana*, доля генов, в регуляции которых участвует процесс сплайсинга. Будет проведена оценка консервативности сплайсинга.

2. Будет показано отсутствие или наличие зависимости разницы в экспрессии паралогов у *Capsella bursa-pastoris* от изменений в промоторных участках.

Тема 2

Будут определены величины отклонения сайтов связывания рибосомы от стандартного консенсуса GGAGGA и найдены консервативные спирали РНК, расположенные в 5'-лидерных областях генов и перекрывающих сайт связывания рибосомы у актинобактерий.

Будет выполнено дальнейшее развитие модели взаимодействия РНК-полимераз; получены посредством моделирования новые результаты, связанные с подтверждением роли консервативных структур РНК.

Будут получены оценки скорости элонгации РНК-полимераз фагового типа в пластидах растений на основании сопоставления размеров экзонов и интронов в генах пластид.

Будет построен эволюционный сценарий для некоторых видов Rhizobiales с реконструкцией предковых событий дупликаций, потерь и возникновений генов и семейств генов.

Тема 3

Мы завершим описание и анализ открытых нами уникальных палиндромов в кодирующих областях митохондриальных генов волосатиков, подготовим и опубликуем статью по этой теме.

На животных типа волосатики мы проведем исследования, посвященные биологической роли и механизмам работы клеток и митохондрий, содержащих в геноме уникальные нуклеотидные палиндромы в кодирующих генах.

Мы завершим работу по детальному описанию чрезвычайно простой нервной системы ортонектид и покажем на этом уникальном модельном организме, какой уровень сложности поведения может быть достигнут с такими крохотными мозгами. Будет проведен анализ поведения этих организмов в различных условиях.

Мы завершим работу по исследованию уникальную стадию жизненного цикла ортонектид – многоядерного плазмодия. Для изучения его особенностей мы исследуем транскриптом этой стадии и сравним с транскриптомом свободноживущих особей. В результате будут найдены и проанализированы гены ортонектид, специализированные для функционирования плазмодия.

Будут разработаны алгоритм поиска ВКЭ и его компьютерная реализация, которые допускают применение к группам с большим филогенетическим расстоянием между видами; группы могут состоять из тысяч геномов. Программный комплекс будет работать на суперкомпьютерах с распределённой памятью (вычислительных кластерах). Он дополнительно классифицирует ВКЭ по их расположению относительно кодирующих областей ДНК. Набор оптимальных параметров для широкомасштабного поиска ВКЭ предложенным методом. Список ВКЭ, найденных с помощью этой программы, в пластидах багрянок. Потенциальные регуляторные участки ДНК, включая промоторы и сайты связывания транскрипционных факторов в пластидах этих видов. База данных ВКЭ в пластидах, доступная по адресу http://lab6.iitp.ru. Кластеризация белков, кодируемых в пластидах. Пополнение базы данных кластеров белков, доступной по адресу http://lab6.iitp.ru/ppc/ . Деревья белков, кодируемых в пластидах рассмотренных видов. Выводы о филогенетическом положении видов, содержащих пластиды.

Будет выполнен поиск ультраконсервативных участков в геномах многоклеточных. Будут найдены последовательности уникальных локусов, сцепленных с ультраконсервативными участками. На этой основе будет получен набор данных для реконструкции филогении многоклеточных.

Тема 4

В соответствии с исходным планом работ, в рамках задачи 4.3 будет разработана методика анализа эпистаза, в т.ч. неэпистатическая нуль-модель. Будет проведена оценка локального эпистаза в ультраконсервативных некодирующих элементах генома *D. melanogaster*.

В рамках задачи 4.7 будет составлен каталог генетической изменчивости *S. commune*, исследованы распределения аллельных частот в полиморфных сайтах.

Кроме того, ожидаются новые результаты, связанные с развитием проектов, в основном завершенных в 2015–2016 годах. В рамках задачи 4.1 будет разработан метод анализа изменения сайт-специфических адаптивных ландшафтов; метод позволит отличать изменения, связанные с эпистазом, от изменений, связанных с динамикой условий внешней среды. В рамках задачи 4.2 будет проанализирован отбор, связанный с различными видами стоп-кодонов у *D. melanogaster*.