ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Межвузовская студенческая научная школа-конференция

# Информационные технологии и системы. Биоинформатика.






7-10 марта 2019 г. учебный центр «Вороново» НИУ ВШЭ

# *Программа конференции «Информационные технологии и системы. Биоинформатика» 7-10 марта 2019 года*

## 7 марта

**11:30 13:00** **приезд и заселение**

**13:00 14:00** **обед**

**14:00 15:00** **отдых**

**15:00 15:30** **открытие**

Гельфанд Михаил Предуведомления

Червонцева Зоя О проекте "Воспроизводимость на Python"

Виноградов Дмитрий Зачем биоинформатику облака?

**15:30 17:00** **Эпигеномика и транскриптомика**

- Жегалова Ирина Комплексные состояния хроматина различных линий D. melanogaster
- Маргасюк Сергей Кластеризация хроматина по данным о модификации гистонов с помощью HMM
- Бердникович Екатерина Сравнительная эпигенетическая аннотация регуляторных ДНК
- Ильницкий Иван Эволюция репертуара эпигенетических белков
- Жарикова Анастасия Полногеномный анализ взаимодействий РНК-хроматин
- Сигорских Андрей Полногеномный анализ взаимодействий РНК с хроматином
- Рябых Григорий Разработка базы данных с веб-интерфейсом для хранения и анализа данных РНК-ДНК взаимодействий
- Шишкова Светлана Распознавание эпигенетических маркеров методами машинного обучения
- Коновалов Дмитрий Роль гуаниновых квадруплексов в метилировании генома

- Ностаева Арина Поиск паттернов ассоциации между квадруплексами и эпигенетическими маркерами

- Бекназаров НазарПоиск паттернов ассоциации между Z-DNA и эпигенетическими маркерами
- Курилович Анна Предсказание структурных мутаций в раковых геномах методами машинного обучения

**17:00 17:30** **чай**

**17:30 19:00** **Эпигеномика и транскриптомика**

- Барановский Артемий Маркеры стволовых клеток ассоциированные с опухолями
- Самосюк Алексей Switch-like genes may ease integration of single-cell pseudotime and clustering
- Понамарева Ирина Dimensionality reduction methods for trajectory inference for single-cell sequencing data

- Фаворов Александр Инди-лес (случайный лес индикаторов попарных сравнений) как классификатор экспрессионных данных

**Ботанический сад**

- Пенин Алексей Ищем сумасшедших сотрудников на изучение полиплоидных растений: обзор задач
- Щелкунов Михаил Изучение ядерного транскриптома Rhopalocnemis phalloides — нефотосинтезирующего, паразитического растения
- Федотов Алексей Генетический и морфологический анализ изменения дорзовентральной полярности листьев среди представителей рода Curi
- Григорьян Максим Распределение гаплотипов LFY в дальневосточных и европейских популяциях Isoëtes
- Шнайдер Элина Анализ происхождения аллотетраплоида Capsella bursa-pastoris с использованием данных полногеномного секвенирования

**19:00 20:00** **ужин**

# 8 марта

**9:00 10:00** **завтрак**

**10:00 11:30** **Quodlibet**

- Набиева Елена Accurate Fetal Variant Calling in the Presence of Maternal Cell Contamination

**Факторы транскрипции**

- Суворова Инна Сравнительно-геномный анализ факторов транскрипции семейства IclR и их сайтов связывания: особенности структуры и ко-эволюция
- Драненко Наталья Реконструкция предковой последовательности белка OxyR
- Кадыкова Татьяна Ко-эволюция факторов транскрипции и их сайтов связывания
- Белоусова Евгения Консервативность неконсенсусных позиций в сайтах связывания факторов транскрипции
- Быкова Дарья Эволюция операторных участков в лямбдоидных фагах
- Серебренникова Мария Классификация ДНК-метилтрансфераз прокариот по сходству последовательностей
- Тутукина Мария Метаболизм гексуронатов у гаммапротеобактерий: как UxuR и ExuR делят свои регулоны между собой и с малыми РНК?
- Бессонова Татьяна Альтернативное кодирование и синтез нескольких белков с одного гена в гаммапротеобактериях на примере LeuO
- СтоляренкоАртемий Особенности формы ДНК, возникающие при связывании с транскрипционными факторами
- Шаймарданов Абусаид Идентификация видоспецифичных ДНК-регуляторных элементов в мозговой ткани

**11:30 12:00** **чай**

**12:00 13:30** **Факторы транскрипции**

- Пензар Дмитрий Предсказание влияния однонуклеотидных вариантов на экспрессию генов по данным параллельного репортерного анализа
- Кравченко Павел Улучшение предсказания сайтов связывания транскрипционных факторов с помощью машинного обучения

**Трехмерная структура хроматина**

- Храмеева Екатерина Вычислительные задачи в изучении структуры хроматина
- Храмеева Екатерина Уровень ацетилирования гистонов регулирует формирование топологически ассоциированных доменов в хроматине дрозофилы
- Галицына АлександраСтруктура хроматина: видовое разнообразие
- Галицына АлександраСтруктура хроматина единичных клеток Drosophila melanogaster
- Кононкова Анна Анализ изменения трехмерной структуры хроматина на разных стадиях сперматогенеза Drosophila melanogaster

**13:30 14:30** **обед**

**14:30 15:30** **отдых**

**15:30 17:00** **Трехмерная структура хроматина**

- Жигулев Артемий Изучение динамики ТАДов на примере сперматогенеза Drosophila melanogaster
- Быков Николай Изменения трехмерной организации хроматина в развитии эмбриона Drosophila melanogaster
- Савостьянов Антон Автоматическая разметка петель на контактных картах амебы Dictyostelium discoideum
- Плискин АлександраПредсказание петель в хроматине Dictyostelium discoideum
- Половников Кирилл Many-body contacts in fractal polymer chains
- Белан Сергей Contact Probability in Loop Extrusion Model of Interphase Chromosome
- Теванян Элен Поиск неканонических структур ДНК как нуклеосомных барьеров методами машинного обучения

- Афентьева Дарья Модели машинного обучения для распознавания положений нуклеосом на основе физико-химических и структурных характеристик ДНК
- Верезубова Виктория Построение обобщающего вероятностного профиля расположения нуклеосом методами машинного обучения

**17:00 17:30** **чай**

**17:30 19:00** **Процессинг пре-мРНК**

- Денисов Степан Тандемные альтернативные сайты сплайсинга: эволюционные свидетельства функциональности
- Миронов Алексей Тандемные альтернативные сайты сплайсинга в раке и норме
- Калмыкова Светлана Влияние соматических мутаций, разрушающих структуру РНК, на альтернативный сплайсинг
- Васюткина Ольга Роль вторичной структуры РНК в экспрессии химерных неколлинеарных транскриптов
- Микова Валерия Intronic polyadenylation and splicing in cancer
- Мазаев Лев Possible role of upstream open reading frames in NMD-dependent degradation of human mRNAs
- Иванов Тимофей Эволюция взаимно исключающих экзонов

**Вирусы**

- Корешова Алевтина Филогеография вируса гепатита А
- Сафина Ксения Молекулярная эпидемиология ВИЧ в одном Российском регионе
- Валяева Анна Влияние белка Tat вируса иммунодефицита человека (HIV-1) на клеточные процессы в B-лимфоцитах

**Quodlibet**

- Погорельская Александра Оценка скорости изменения эпигенетических свойств в одиночных клетках

**19:00 20:00** **ужин**

# 9 марта

**9:00 10:00 завтрак**

**10:00 11:30 Геномика бактерий**

- Джамалова Дильфуза Слияние бактериальных видов с помощью пангеномного анализа
- Николаева Дарья Связь структуры пангенома и разнообразия местообитаний бактерий микробиома Земли
- Перевощикова Кристина Реконструкция геномных перестроек в бактериях рода Vibrio
- Жиенбаева Молдир Comparative genomics of Bacillus spp
- Сефербекова Заира Геномные перестройки Shigella sp.
- Афасижев Роберт Гомологичная рекомбинация в геноме Escherichia coli
- Валиев Иван Поиск паттернов в объединённой сети транскрипционной регуляции и метаболизма E. coli
- Ходжаева Евгения Перестройки оперонов метаболических путей бактерий
- Рыбина Анна Филогенетический анализ кассеты генов Escherichia coli, участвующей в деградации сульфоквиновозы и лактозы
- Шевкопляс Алексей Предсказание и сравнительный анализ лидерных пептидов триптофанового оперона Rhizobiales
- Транкова Наталья Исследование комплементарных взаимодействий продуктов транскрипции 6S-РНК с мРНК

**11:30 12:00 чай**

**12:00 13:30 Эволюция последовательностей**

- Гарушянц Софья Мутационные паттерны в эволюции Escherichia coli
- Волобуева Мария Анализ частот встречаемости мутации в зависимости от контекста на примере 55 геномов 9 видов бруцелл
- Семенченко Егор Evolution of germline mutational spectra among great apes
- Потапова Надежда Микроинверсии в мире человека и других приматов
- Потапова Надежда Эволюция последовательностей, окружённых микросателлитами

- Селифанова Мария Изучение консервативных геномных участков в популяции человека
- Колотова Арина Кратность межгенных областей и альтернативные стоп-кодоны в Enterobacteriales
- Соколов Александр Оценка действия отбора на межгенные участки Saccharomyces cerevisiae
- Червова Альмира Susceptibility of single-stranded DNA to APOBEC mutagenesis during transcription

**Белки**

- Биба Дмитрий Поиск скомпенсированных сдвигов рамки считывания и определение их роли в образовании новых аминокислотных последовательностей
- Раменский Василий Анализ и предсказание эффекта коротких инделов в белках
- Столярова Анастасия Что стоит за предпочтениями аминокислот в белковых сайтах
- Сафронов Вячеслав Prediction of structural susceptibility of proteins to regulatory proteolysis

**13:30 14:30** **обед**

**14:30 15:30** **отдых**

**15:30 17:00** **Белки**

- Молдован Михаил Эволюция фосфорилируемых аминокислот
- Ириоглов Роман Были ли многодоменные белки у LUCA

**Филогенетика и молекулярная эволюция**

- Безменова Александра Накопление мутаций в экспериментальной эволюции базидиомицета Schizophyllum commune
- Безменова Александра Зависимость скорости гомологичной рекомбинации от уровня гетерозиготности хромосомы в базидиомицете Schizophyllum commune
- Столярова Анастасия Неравновесие по сцеплению с высоким разрешением в Schizophyllum commune
- Худякова Ксения Адаптивная динамика на ландшафте, заданном матрицей эпистатических взаимодействий

- Столярова Анастасия Изменение приспособленности текущего аллеля в ходе эволюции по экспериментальным данным
- Клинк Галина Распределение аминокислот на филогенетическом дереве как отражение однопозиционного адаптивного ландшафта
- Безсуднова Ольга Роль горизонтального переноса в эволюции систем рестрикции-модификации
- Гусева Екатерина Горизонтальный перенос и вертикальное наследование систем рестрикции-модификации двух гомологичных классов, включающих эндонуклеазу с доменом RE_TdeIII
- Русинов Иван Паттерн избегания сайтов рестрикции как способ предсказания хозяина бактериофага

**17:00 17:30** **чай**

**17:30 19:00** **Филогенетика и молекулярная эволюция**

- Никитин Иннокентий Предсказание качества филогенетической реконструкции методом машинного обучения
- Котюргин Александр Классификация и реконструкция филогении белковых последовательностей низкой сложности методами, не использующими выравнивания
- Панова Вера Разработка алгоритма и компьютерной программы, отличающей последовательности реальных белков от ошибочно предсказанных

**Генетика человека**

- Безуглов Виталий Анализ генома Эци и поиск отличий его с референсным геномом современного человека
- Кузнецов Иван Поиск признаков ассортативного скрещивания в геноме человека
- Славский Сергей Анализ роста человека как сложного генетического признака в UK Biobank
- Шашкова Татьяна GWAS-MAP: Анализ результатов полногеномных исследований ассоциаций с целью получение нового биологического знания
- Алексеева Евгения Филогенетический анализ B-клеточных линий человека

Жданова Анна Сравнение различных методов обогащения экзомов путём анализа с использованием GATK pipeline

**Quodlibet**

- Мыларщиков Дмитрий Архитектурные РНК млекопитающих
- Медведева Ксения Регуляция генов теплового шока у бактерий
- Быкова Ульяна Модели машинного обучения для распознавания структур стебель-петля на 3’- конце транспозонов SINE и LINE в геноме Danio rerio

**19:00 20:00 ужин**

# 10 марта

**9:00 10:00 завтрак**

**10:00 11:30 Метагеномика**

- Хачатурян Марина Исследование бактериальных защитных систем в Мировом океане
- Шатов Владислав Оценка топологических свойств пространства метагеномнов микроорганизмов кишечника человека на их кластеризацию в энтеротипы
- Шелякин Павел Сравнение бактериального метагенома больных и здоровых кораллов рода Porites
- Поздышев Арсений Микробиомный подход для определения профиля притока в горизонтальной нефтяной скважине
- Сарана Юлия Сравнительный анализ микробиомов тлей

**Зоопарк**

- Попов Алексей Изучение диапаузы у яиц Daphnia magna
- Шайхутдинов Нурислам Популяционная геномика “спящей” хирономиды (Polypedilum vanderplanki)
- Вахрушева Ольга Популяционная геномика бделлоидных коловраток вида Adineta vaga

- Бурская Валентина The search for bird mitochondrial genome adaptations to high altitude, migration, diving, wintering and flight
- Бочкарева Ольга Эволюция глобиновых локусов у рыб
- Молдован Михаил Редактирование мРНК головоногих моллюсков какпример преадаптации
- Ногина Дарья Влияние редактирования мРНК на процесс трансляции у мягкотелых головоногих моллюсков

**11:30 12:00 чай**

**12:00 13:30 Зоопарк**

- Гайдукова Софья Эволюция сдвигов рамки считывания в транскриптомах инфузорий
- Юдина Софья Морфологическая и геномная эволюция микогетеротрофных групп растений на примере рода Thismia (Thismiaceae, Dioscoreales)

**Нейронные сети**

- Алишев Наиль Использование глубинного машинного обучения для восстановления пропусков в Hi-C картах
- Горохов Никита Улучшение качества результатов экспериментов Hi-C единичных клеток с помощью нейронных сетей
- Червонцева Зоя Предсказание сигнала eCLIP с помощью нейронных сетей
- Литвин Анна Использование сверточных нейросетей для предсказания нуклеотидных последовательностей
- Преображенская Юлия Предсказание консервативности нуклеотидов в геноме кишечной палочки по контексту
- Григорашвили Елизавета Предсказание вторичной структуры тРНК с помощью глубокого обучения
- Бочкарева Мария Распознавание участков H-DNA сверточными нейронными сетями (CNN)
- Латышев Павел Распознавание квадруплексов сверточными нейронными сетями (CNN)
- **Транспозоны**
- Исаев Сергей Поиск активной генетической рекомбинации, основанной на транспозонах семейства RAG, в геномах моллюсков

- Заикин Антон Модели машинного обучения для распознавания структур стебель-петля на 3’-концах транспозонов L1 и Alu в геноме человека
- Воронкова Анастасия Модели машинного обучения для распознавания 3’-конца псевдогенов и транспозонов человека
- Черницов Александр Распознавание классов SINE по всему дереву жизни с помощью моделей машинного обучения

**13:30 14:30 обед**

**14:30 15:30 отдых**

**15:30 16:00 backup**

**16:00 отъезд**

# Эпигеномика и транскриптомика

## Комплексные состояния хроматина различных линий *D. melanogaster*

***Ирина Жегалова*, *Екатерина Храмеева***

Хроматин состоит из ДНК и множества модифицированных гистонов и негистоновых белков, которые влияют на дифференцировку клеток, регуляцию генов и другие ключевые клеточные процессы. Дрозофила использовалась в качестве модельной системы на протяжении долгого времени для изучения структуры и функции хромосом, регуляции генов, развития и эволюции.

Профилирование компонентов хроматина по всему геному создало богатую аннотацию потенциальных функций основных последовательностей ДНК. Комплексные «состояния» хроматина определяются как процессы на разных уровнях организации, от отдельных регуляторных единиц до хромосом, и связывают их отдельные состояния с функциями генома.

Мы определяем комплексные «состояния» хроматина для Drosophila melanogaster, основанные на восемнадцати гистонных модификациях и двенадцати структурных белках, представленных восемью преобладающими комплексными состояниями. Анализ проводится для клеточных линий S2-DRSC (поздние эмбриональных ткани самца), ML-DmBG3-c2 (центральная нервная система личинок самцов) и Kc167 (самки в возрасте 10-12 часов), а также для эмбрионов в возрасте 14-16 часов.

Согласно литературным данным, анализ может выявить группы взаимосвязанных признаков, в том числе связанных с регионами концентрации гетерохроматина, с репрессией, опосредованной Поликомбом, и с активной транскрипцией, аналогичные тем, которые наблюдаются у других организмов. Получение подобных данных подтвердит то, что специфические модификации гистонов работают вместе и образуют различные «состояния» хроматина.

Планируется сравнить полученные данные с имеющейся разметкой для S2-DRSC и ML-DmBG3-c2, представленной Kharchenko et al. 2010, а также провести подобный анализ для клеточных линий человека, для получения сопоставимых разметок.

## Кластеризация хроматина по данным о модификации гистонов с помощью HMM

***Сергей Маргасюк*, *Андрей Александрович Миронов***

Задача – раскраска хроматина – по данным об интенсивности модификаций гистонов вдоль генома (получены из экспериментов ChIP) разделить геномные позиции на кластеры. Известен комбинаторный характер совместного действия различных модификаций: модификации не «считываются» регуляторными белками независимо, а могут образовывать паттерны, скоррелированные с некоторыми функциональными особенностями хроматина. Таким образом, полученные кластеры могут быть использованы для поиска паттернов взаимодействия модификаций, важных для описания последовательности в целом. Кроме того,

поскольку модификации коррелируют между собой, кластеризация может быть использована для понижения размерности этих данных в других исследованиях (вместо полного набора данных использовать только идентификатор кластера). Для кластеризации используется обучение скрытой марковской модели (HMM) без учителя (наблюдаемая последовательность – интенсивность модификации, скрытое состояние – требуемый класс).

Для кластеризации данных с помощью HMM необходимо решить техническую задачу – представление входных данных в виде выборки из смеси (mixture) нескольких распределений из фиксированного параметрического семейства. При этом сами данные – пара [p-value (отличия от пуассоновской модели на основе контроля), fold-change (по сравнению с контролем)] для каждого типа модификации для каждой позиции (в результате обработки MACS) – имеют неопределенный вид: значимость, вероятно, правильно рассматривать как дискретную величину, а fold-change принимает очень большое число различных значений (то есть имеет смысл приближение непрерывным распределением), но во многих точках равен 0. Простейший вариант, описанный в работе (Ernst, 2012) – дискретизация по p-value (считаем, что сигнал равен 1, если p-value ниже порога, 0 иначе); первым шагом в нашей работе должно стать приблизительное воспроизведение результатов из этой статьи. В дальнейшем хотелось бы построить кластеризацию, основанную на непрерывном сигнале – возможно, она позволит различать больше особенностей модификации (например, модификацию в данной позиции только в одной из хромосом).

## Сравнительная эпигенетическая аннотация регуляторных ДНК

***Екатерина Андреевна Бердникович, Анастасия Александровна Жарикова, Андрей Александрович Миронов***

*Цель:*

В эукариотических клетках геномная ДНК и гистоны изменены множеством химических модификаций, которые называют эпигенетическими модификациями. Эти модификации добавляют новый уровень информации к геномным последовательностям, позволяя закодировать больше программ генной регуляции. Разнообразные эпигенетические модификации влияют на взаимодействие факторов транскрипции с ДНК. Многое остаётся неизвестным. К тому же геном далёк от того, чтобы быть полностью аннотированным на функциональном уровне, тем самым делая необходимым сначала найти регуляторные последовательности перед тем как осознать их комплексную регуляторную роль.

Сравнение последовательностей генома позволило функционально его аннотировать. Таким образом перед нами встает вопрос: можем ли мы применить эволюционное исследование, чтобы исследовать функции эпигенома?

Если так, базовые эволюционные свойства эпигенома должны быть установлены первыми предпочтительнее в контексте обеих геномной и транскриптомной эволюции [1]. Таким образом, можно поставить конечную цель:

Разработка метода для систематической оценки функции эпи-модификаций. И,что более важно, функций комбинаций эпи-модификаций.

*Задача 1*

Оценка эпигенетической консервативности.

Для достижения поставленной цели мы провели сравнительное эпигеномное исследование с акцентом на эволюцию. Мы собрали данные по эпигенетическим модификациям, включающим Cm, H2A.Z, H3K4me1/2/3, H3K9me3, H3K27me3, H3K27ac, H3K36me3 в плюрипотентных стволовых клетках человека, мыши и свиньи (Sus scrofa) [2-4].

*Что сделано:*

1) найдены все данные, используемые в статье.
2) Проведен сравнительный анализ модификаций гистонов у мыши и человека. Иерархическая кластеризация показала, что линии клеток разных видов кластеризуются отдельно, что доказывает строгую межвидовую разницу и внутривидовую схожесть в эпигенетических модификациях. Также была проведена кластеризация уровня генной экспрессии, которая показала такую же закономерность.

*Планы:*

*Подзадача 1:*

Отследить следы эволюционной сохранности эпигенетических модификаций у разных видов.

*Подзадача2:*

Отследить постоянное совместное появление эпи-модификаций во всех видах.

*Подзадача3:*

Протестировать перекрывание совместного появления эпи-модификаций с консервативными регионами.

*Подзадача4:*

Проверка генов-ортологов на схожесть эпи-модификации.

*Список литературы:*

[1] *Shu Xiao 6, Dan Xie 6, Xiaoyi Cao 6, Pengfei Yu 6, Xiaoyun Xing, Chieh-Chun Chen, Meagan Musselman, Mingchao Xie, Franklin D. West, Harris A. Lewin, Ting Wang, Sheng Zhong*
Comparative Epigenomic Annotation of Regulatory DNA
*Cell:June 7, 2012*
https://doi.org/10.1016/j.cell.2012.04.029

*[2] Xi Chen 6, Han Xu 6, Ping Yuan, Neil D. Clarke, Chia-Lin Wei, Huck-Hui Ng*
Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells
*Cell:June 12, 2008*
https://doi.org/10.1016/j.cell.2008.04.043 Chen et al., 2008

*[3] Alon Goren, Fatih Ozsolak, Noam Shoresh, Manching Ku, Mazhar Adli, Chris Hart, Melissa Gymrek, Or Zuk, Aviv Regev, Patrice M Milos & Bradley E Bernstein*

Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA *Nature:29 November, 2009*
http://www.nature.com/articles/nmeth.1404 Goren et al., 2010

*[4] Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren & Joseph R. Ecker*
Human DNA methylomes at base resolution show widespread epigenomic differences
*Nature:14 October, 2009*
http://www.nature.com/articles/nature08514

## Эволюция репертуара эпигенетических белков

***Иван Ильницкий, Анастасия Александровна Жарикова, Андрей Александрович Миронов***

Эпигенетические процессы важны для функционирования живых систем и поддержания стабильности генома, влияют на экспрессию белков и РНК. В 2006 году был открыт новый класс малых некодирующих РНК — РНК, взаимодействующие с белками класса PIWI (PIWI-interacting RNA, piРНК), отмеченных в процессах модификации хроматина. piРНК транскрибируются со специальных кластеров, после чего их созревание сопровождается взаимодействием с группой белков. Связываясь с белком PIWI, они способны лимитировать экспрессию ретротранспозонов на посттранскрипционном уровне с помощью эпигенетического сайленсинга (рекрутирование ДНК-метилтрансфераз) в развивающихся эмбриональных клетках. Также исследования на дрозофилах показывают, что белок PIWI активирует фактор HP1, стимулирующий формирование гетерохроматина. Проводимые исследования показывают, что piРНК—PIWI система имеет много аспектов в эпигенетике, однако ее взаимодействие с основными факторами ремоделирования хроматина на сегодняшний день недостаточно изучено. С целью рассмотреть коэволюционные паттерны между белками системы piРНК—PIWI и факторами эпигенетики нами была собрана коллекция белков, представляющих каждую из данных групп, а также были написаны скрипты на языках программирования Python и SPARQL для автоматизации последующего анализа. Дальнейшая работа направлена на разработку веб-сервиса и расширение базы данных белков.
Из полученных данных о наличии ортологов белков в выборке из более 400 видов многоклеточных животных, заметны следующие паттерны их представленности:
а) У трематоды Schistosoma japonicum отсутствует аппарат биогенеза piРНК;
б) Выделена группа специфичных белков для дрозофилы и близкородственных ей видов: *Krimper, Vreteno, Qin, Panoramix, Rhino, Tejas, Deadlock, Yb*;
в) У птиц имеются заметные отличия в PIWI аппарате в сравнении с другими позвоночными, у них нет белков *UAP56* и *Cutoff*, мало представлен и белок *MitoPLD*;
г) Проявляется корреляция между белками первичного биогенеза piРНК и ДНК-метилтрансферазами.
Работа имеет поисковый характер, создание набора инструментов для автоматизированного

анализа коэволюционных паттернов и расширение группы исследуемых белков поможет сделать биологически обоснованные выводы.

Литература
1. Aravin A.A., Sachidanandam R., Bourc'his D., Schaefer C., Pezic D., Toth K.F., Bestor T,Hannon GJ. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008 Sep 26;31(6):785-99. doi: 10.1016/j.molcel.2008.09.003. PubMed PMID: 18922463; PubMed Central PMCID: PMC2730041.
2. Brower-Toland B, Findley SD, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin SC, Lin H. Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes Dev. 2007 Sep 15;21(18):2300-11. PubMed PMID: 17875665; PubMed Central PMCID: PMC1973144.
3. Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T., & Kandel, E. R. (2012). A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. Cell, 149(3), 693–707. doi: 10.1016/j.cell.2012.02.057

## Полногеномный анализ взаимодействий РНК-хроматин

***Анастасия Александровна Жарикова, А.А. Галицына, А.А. Гаврилов, М.Д. Логачева, С.В. Разин и А.А. Миронов***

Известно, что значительная часть генома эукариот транскрибируется с образованием большого количества разнообразных РНК, включая мРНК и различные длинные и короткие некодирующие РНК [1]. Молекулы РНК могут выполнять свои функции не только в цитоплазме, но и оставаясь в ядре клетки, где они активно участвуют в процессах регуляции транскрипции, а также ремоделирования и поддержания пространственной структуры хроматина [2]. Классическими примерами таких РНК могут служить XIST, HOTAIR, MALAT1, TERC и другие [3].

На сегодняшний день существует целый спектр разработанных ранее методик, позволяющих выявить локусы ДНК, с которыми взаимодействует одна или несколько заранее известных РНК [4, 5]. В последнее время появляются методы, позволяющие в рамках одного эксперимента полногеномно определить РНК-ДНК интерактом: MARGI [6], ChAR-Seq [7] и GRID-Seq [8], а также метод, предложенный нами ранее [9].

Данная работа посвящена анализу результатов эксперимента по определению полногеномного спектра РНК-ДНК контактов. Ранее нами был разработан алгоритм анализа такого типа данных, охватывающий все этапы от фильтрации сырых чтений до аннотации отобранного пула РНК-ДНК контактов [10], который был доработан внедрением нескольких подходов по удалению неспецифических РНК-ДНК взаимодействий. С помощью предложенного алгоритма был проведен анализ собственных данных для клеток K562 и фибробластов человека.

Нами были обнаружены ранее неаннотированные хроматин-ассоциированные РНК, отсутствующие в RNA-seq эксперименте для того же типа клеток. Было исследовано предпочтение групп РНК к образованию цис- и/или транс-контактов с хроматином. Показано, что белок-кодирующие РНК контактируют в основном рядом с кодирующим их геном, а, например, snoRNA предпочитают образовывать транс-контакты. Также было исследовано

предпочтение РНК образовывать контакты с рядом специфических типов хроматина: промоторы разной силы, энхансеры, локусы, репрессированные с помощью белков группы Polycomb, гетерохроматин и др, определенные в статье Ernst et al. [11]. Например, показано, что сплайсосомна РНК U2 предпочитает контактировать с активным хроматином, а XIST часто взаимодействует с Polycomb-репрессированными участками. На основании распределения контактов индивидуальных РНК вдоль генома был разработан способ обнаружения групп РНК, контактирующих с геномом сходным образом. Сосредоточившись на белок-кодирующих генах было показано, что РНК чаще контактируют с областями, следующими за стартом транскрипции (по ходу транскрипции), чем с предшествующими ему, что можно объяснить тем, что РНК тянется за РНК-полимеразой в процессе транскрипции гена.

Данная работа была поддержана грантом РФФИ 17-00-00180, грантом РНФ 18-14-00011, а также грантом “Systems biology Fellowship Program”.

Ссылки:

1. Djebali S., Davis C.A., Merkel A., Dobin A., Lassmann T., Mortazavi A., Tanzer A., Lagarde J., Lin W., Schlesinger F., Xue C., Marinov G.K., Khatun J., Williams B.A., Zaleski C., Rozowsky J., Röder M., Kokocinski F., Abdelhamid R.F., Alioto T., Antoshechkin I., Baer M.T., Bar N.S., Batut P., Bell K., Bell I., Chakrabortty S., Chen X., Chrast J., Curado J., et al: **Landscape of transcription in human cells**. Nature. 2012, 489: 101-108. 10.1038/nature11233.

2. Engreitz, J.M.: **Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression**. Nature Reviews Molecular Cell Biology, 17, 756-770 (2016)

3. Quinn J.J., Chang H.Y.: **Unique features of long non-coding RNA biogenesis and function. Nature Reviews Genetics**. 2016 Jan;17(1):47-62. doi: 10.1038/nrg.2015.10

4. Engreitz, J.M.: **RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites**. Cell 159(1), 188-199 (2014).

5. Chu, C.: **Technologies to probe functions and mechanisms of long noncoding RNAs.** Nature Structural & Molecular Biology 22, 29-35 (2015)

6. Sridhar D.: **Systematic Mapping of RNA-Chromatin Interactions In Vivo**. Current Biology

27, 1-8 (2017)

7. Bell J.C.: **Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNAto-DNA contacts**. bioRxiv (2017)

8. Li, X.: **GRID-seq reveals the global RNA–chromatin interactome**. Nature Biotechnology, 35, 940–950 (2017)

9. Жарикова, А. А.: **Взаимодействие ДНК и РНК в хроматине**. Тезисы конференции ИТИС (2017)

10. Жарикова А.А.: **Исследование свойств РНК-ДНК контактов в хроматине**. Тезисы конференции ИТИС (2018)

11. Ernst J., Kheradpour P., Mikkelsen T.S., Shoresh N., Ward L.D., Epstein C.B., Zhang X., Wang L., Issner R., Coyne M., Ku M., Durham .T, Kellis M., Bernstein B.E.: **Mapping and analysis of chromatin state dynamics in nine human cell types.** Nature. 2011 May 5;473(7345):43-9. doi: 10.1038/nature09906. Epub 2011 Mar 23.

## Полногеномный анализ взаимодействий РНК с хроматином

***Андрей Сигорских*, *Анастасия Александровна Жарикова***

Взаимодействие РНК с хроматином и другими ядерными структурами хорошо изучено только на отдельных примерах, и работы, охватывающие одновременный анализ сравнительно большого числа разных РНК, представляют отдельный интерес. Целью данной работы является поиск предположительно функциональных РНК-хроматин взаимодействий на основании данных различных экспериментов по поиску РНК-белковых (в частности РНК-хроматин) контактов.

В данный момент в работе можно выделить следующие подзадачи:

- Разработать процедуру фильтрации данных полногеномного поиска РНК-хроматин взаимодействий, исследовать поведение предполагаемых неизвестных ранее длинных некодирующих РНК.
- Исследовать предположение о том, что мРНК перемещиются к ядерной мембране по областям контакта разных хромосомных территорий, а не через хромосомную территорию "своей" хромосомы.
- Исследовать динамику "движения" отобранных при выполнениии подзадачи 1 длинных некодирующих РНК по хроматину от места её транскрипции до предполагаемого места функционального взаимодействия с хроматином.

В ходе выполнения подзадачи 1 был написан тул для фильтрации данных РНК-хроматин контактов, а также эмпирически предположены пороги для отбора для дальнейшего исследования интересующих фрагментов - предположительно ранее неизвестных функциональных длинных некодирующих РНК. Задачи 2 и 3 находятся в стадии разработки.

## Разработка базы данных с веб-интерфейсом для хранения и анализа данных РНК-ДНК взаимодействий.

***Григорий Рябых, студент 6 курса ФББ МГУ***
***А.А. Жарикова, преподаватель ФББ МГУ***
***А.А. Миронов, проф. ФББ МГУ, д.б.н, д.ф-м.н***

Известно большое количество некодирующих РНК: микроРНК, пиРНК, малые ядерные РНК, длинные некодирующие РНК, энхансерные РНК и другие. Они имеют различную клеточную локализацию и играют важную роль почти во всех процессах жизнедеятельности клетки. Например, одна из главных функций длинных некодирующих РНК – это регуляция экспрессии генов на разных уровнях, включая привлечение аппарата транскрипции, посттранскрипционные модификации и эпигенетику. Одни РНК могут взаимодействовать с хроматином «цис» (например, РНК XIST инактивирует Х-хромосому, на которой сама и закодирована [1]), другие (например, MALAT1 и NEAT1 [2]) способны образовывать контакты с соседними хромосомами – «транс».

На сегодняшний день существует несколько методов, с помощью которых можно определить полногеномную локализацию одной РНК на хроматине: ChIRP [3], CHART [4],

RAP-DNA [5]. Однако эти методы позволяют анализировать только одну известную РНК за один эксперимент, и, следовательно, они не дают возможности полногеномно посмотреть на взаимодействия всех РНК с ДНК. Однако в 2017 году появились первые работы, которые предлагают методы, позволяющие получить данные обо всех потенциальных РНК-ДНК контактах в клетке: MARGI [6] и GRID-seq [7].

Данная работа посвящена разработке базы данных, предназначенной для накопления данных РНК-ДНК контактов, их быстрого и удобного анализа. За основу мы взяли колоночно-ориентированную систему управления базами данных (СУБД) ClickHouse, позволяющую выполнять аналитические запросы в режиме реального времени на больших данных.

На данный момент мы можем работать с полногеномными данными РНК-ДНК контактов (данные GRID[7], клеточная линия MDA-MB-231; данные, полученные от наших коллабораторов из лаборатории Сергея Владимировича Разина, клеточная линия K562), а также с результатами экспериментов CHART для длинной некодирующей РНК NEAT1, клеточная линия MCF-7[8].

Разработанная нами база данных предоставляет пользователю:

- выбирать разные модели учета неспецифических контактов
- определять процентное соотношение контактов РНК с хроматином, участвующих в «локальных», «цис» и «транс» взаимодействиях
- строить профили контактов выбранной РНК вдоль всего генома или его участка
- строить распределения контактов единичной РНК, или группы, по геному или любому выбранному участку

Кроме этого мы разрабатываем связанный с базой данных веб-интерфейс, который позволит пользователю формировать треки контактов из разных экспериментов по одной\группе РНК, которые можно будет визуализировать в Genome Browser или построить интерактивные графики, описанные выше.

В планах у нас:

- закончить сбор необходимых для анализа данных
- доделать и сделать общедоступным веб-интерфейс для базы данных
- добавить возможность сравнивать эксперименты между собой
- сделать возможным выкачивание наших данных, если пользователь захочет провести свою нормировку, фильтрацию или любой другой анализ

Эта работа актуальна, так как аналитической базы данных, содержащей все имеющиеся данные РНК-ДНК взаимодействий, препроцессированные единым образом, с веб-интерфейсом еще не существует, и она позволит глобально взглянуть на данные РНК-ДНК контактов, а также провести масштабный и быстрый анализ некодирующих РНК.

**Источники и литература**

1) M. D. Simon, S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V. K. Maier, R. E. Kingston, and J. T. Lee, “Highresolution Xist binding maps reveal

two-step spreading during X-chromosome inactivation," Nature, vol. 504, no. 7480, pp. 465–469, 2013.

2)J. A. West, C. P. Davis, H. Sunwoo, M. D. Simon, R. I. Sadreyev, P. I. Wang, M. Y. Tolstorukov, and R. E. Kingston, "The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites," Mol. Cell, vol. 55, no. 5, pp. 791–802, 2014.

3) Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. Mol. Cell 44, 667–678 (2011).

4) Simon, M.D. et al. The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA 108, 20497–20502 (2011).

5) Engreitz, J.M. et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973 (2013).

6) Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong S. Systematic mapping of RNA-chromatin interactions in vivo. Curr Biol. 2017;27:602–609.

7) Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., and Fu, X.-D. (2017). GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol.

8) Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).

## Распознавание эпигенетических маркеров методами машинного обучения

***Светлана Шишкова***

магистр 1 года кафедры биоифизики, физический ф-т, МГУ

Модификации гистонов играют важную роль в регуляции транскрипционной активности генов. Расшифровка гистонового кода является актуальной задачей, которая в настоящее время не решена. Целью настоящей работы является построение моделей машинного обучения для распознавания областей ДНК, содержащих модификации гистонов определенного типа в разных типах тканей. В качестве обучающей выборки будут использоваться данные консорциумного проекта The Roadmap Epigenomics. Будут исследованы разные типы моделей машинного обучения, включая SVM, Random Forest и сверточные нейронные сети.

## Роль гуаниновых квадруплексов в метилировании генома

***Коновалов Дмитрий***

студент магистратуры кафедры биофизики физического факультета МГУ им. М.В. Ломоносова.

Гуаниновые квадруплексы это неканонические вторичные структуры ДНК образующиеся из богатых гуанином последовательностей ДНК. Квадруплексы широко распространены в геноме человека и ассоциированы с рядом генетических заболеваний [1]. Роль многих таких структур в геноме, на данный момент является неизвестной. При этом существует множество свидетельств связи квадруплексов с эпигенетической регуляцией, но на данный момент не выявлено общего правила. Такое правило, по всей видимости, является сложным (как

минимум не линейным, так сообщалось об увеличении числа квадруплексов при гипометилировании[2], и о том, что стабильность квадруплексов увеличивалась при гиперметилировании [3], [4]).

В связи с этим существует необходимость уточнения устройства механизма эпигенетической регуляции активности квадруплексов, что и является целью данной работы. Для этого был проведен статистический анализ эпигенетических данных проекта «Roadmap Epigenomics»[5], в том числе результатов бисульфитного секвенирования 145-ти образцов 43-х различных человеческих тканей.


1. Maizels, Nancy. "G4-associated human diseases." *EMBO reports* 16.8 (2015): 910-922.
2. François, Maxime, et al. "Folate deficiency and DNA-methyltransferase inhibition modulate G-quadruplex frequency."*Mutagenesis* 31.4 (2016): 409-416.
3. Li, Pei-Tzu, et al. "Expression of the human telomerase reverse transcriptase gene is modulated by quadruplex formation in its first exon due to DNA methylation." *Journal of Biological Chemistry*(2017): jbc-M117.
4. Lin, Jing, et al. "Stabilization of G-quadruplex DNA by C-5-methyl-cytosine in bcl-2 promoter: implications for epigenetic regulation."*Biochemical and biophysical research communications* 433.4 (2013): 368-373.
5. http://www.roadmapepigenomics.org


## Поиск паттернов ассоциации между квадруплексами и эпигенетическими маркерами


***Арина Ностаева, 1 курс магистерской программы АДБМ НИУ ВШЭ***


Квадруплексы играют важную роль в регуляции клеточных процессов. Они обогащены в теломерах, промоторных областях, на границе интронов и экзонов, в окрестностях начала репликации. Связь квадруплексов с эпигенетическими маркерами остается неизученной. В данной работе мы планируем осуществить полногеномный поиск паттернов ассоциации квадруплексов и эпигенетических маркеров (31 на сайте The Roadmap Epigenomics) в разных типах тканей. Мы планируем выделить общие и тканеспецифичные паттерны, а также определить, с какими метаболическими путями ассоциируются обнаруженные общие и тканеспецифичные паттерны. Также мы планируем построить модели машинного обучения, распознающие данные паттерны. Большой интерес в этом плане для нас представляют сверточные нейронные сети (CNN), а также исследование влияния 2D репрезентации последовательности ДНК в качестве входных данных на качество обучения.

## Поиск паттернов ассоциации между Z-DNA и эпигенетическими маркерами


***Назар Бекназаров, 4 курс бакалавриата ФКН ВШЭ***


Участки левозакрученной формы Z-DNA были обнаружены в геномах разных видов. Существуют экспериментальные доказательства, что Z-DNA играет роль в транскрипции,

ремоделировании хроматина и рекомбинации. Связь эпигенетических факторов с участками Z-DNA остается малоизученной. Целью данной работы является определение участков Z-DNA в геноме человека, ассоциированными с эпигенетическими маркерами, и построение моделей машинного обучения для распознавания данных паттернов. В частности, планируется сравнить эффективность применения сверточных нейронных сетей (CNN) по сравнению с моделями машинного обучения, осуществляющую классификацию на основе заранее заданных признаков.

## Предсказание структурных мутаций в раковых геномах методами машинного обучения

***Курилович Анна, физтех, m2, ankurae@mail.ru***

Известно, что геном онкологических больных нестабилен, при этом в большинстве случаев нестабильность проявляется не только на нуклеотидном уровне в виде соматических мутаций, но и на структурном уровне в виде хромосомных перестроек. В последние годы повышенное внимание исследователей привлекают вторичные структуры ДНК в связи с их разнообразными регуляторными функциями. Работа посвящена предсказанию структурных мутаций в раковых клетках печени на основе данных о расположении вторичных структур ДНК и эпигенетических маркёрах при помощи методов машинного обучения.

## Маркеры стволовых клеток ассоциированные с опухолями.

***Артем Барановский, Дмитрий Папаценко, Дмитрий Первушин.***

Злокачественные образования часто характеризуются возобновлениемэкспрессии генов ассоциированных со стволовостью. Параллельно,возрастает количество наблюдений связывающих реактивацию геновстволовости с развитием раковой опухоли в направлении болееагрессивного фенотипа. На первый взгляд, подобная реактивация неследует какой-либо системе и больше напоминает случайную. Но, хотьэлемент случая, бесспорно, присутствует в этом процессе, егонаправленность поддается определению. Одна из теорией предполагаетконечной целью развития раковой опухоли эмбриональную стволовуюклетку. Другими словами, раковые клетки повторяют нормальное развитиеткани, из которой они образованы, но в обратном направлении. В нашейработе мы обнаружили непрямое подтверждение этой теории. Используяданные экспрессии раковых опухолей и нормальных тканей из The CancerGenome Atlas (TCGA) и стволовых клеток из GEO было обнаружено, что впределах экспрессии ~300 генов стволовости раковые образцы находятсямежду стволовыми клетками и соответствующими нормальными тканями. Внастоящий момент мы проводим анализ альтернативного сплайсинга внутрисобранного множества генов стволовости с предположением, что помимообщей экспрессии, раковые образцы могут быть более схожи со стволовымиклетками и по включению отдельных экзонов.

## Switch-like genes may ease integration of single-cell pseudotime and clustering

***Alexey Samosyuk and Dmitri Pervouchine***

In spite of recent advances in single cell RNA-seq data analysis, it is believed that we are only scratching the surface of what we can ask single cell data. This work considers a problem of pseudotime-clustering integration. Available methods classify cells into distinct subpopulations, independently infer interaction trajectories and then integrate these types of information. We propose an approach that allows balancing between pseudotime and clustering representation of the data within the same workflow. This method assumes that switch-like genes with a similar dropout percentage came from the same cell subpopulations. Since the total number of cells is known, it is possible to find gene sets resembling alternative stable clustering. For the developing embryo dataset and mouse fibroblasts de-differentiation this approach allows phenotype-correct clustering on a gene set of just 50-100 genes and constructing multiple pseudotime trajectories, resembling different biological processes.

## Dimensionality reduction methods for trajectory inference for single-cell sequencing data

***Irina Ponamareva, Helena Todorov, Ivan Saeys***

Благодаря данным single-cell sequencing мы можем наблюдать не только за целой популяцией клеток как за одним целым, но и получать результаты для отдельных клеток. Из этих данных мы можем наблюдать так называемые траектории: выравнивая клетки вдоль траекторий в пространствах низких размерностей, мы можем видеть, как они дифференцировались или эволюционировали. Это может быть важно, например, для задач, связанных с исследованием рака, так как раковые клетки быстро эволюционируют.

Многие из методов, разработанных для извлечения информации о траекториях в single-cell data, уже были сравнены друг с другом. Однако неисследованным остается большое количество методов понижения размерности, которые не нацелены на извлечение траекторий из данных. Большинство из этих методов, однако, могут справляться с этой задачей, и существует необходимость сравнения их результатов.

Задачи проекта:

1. Придумать способы (метрики) для сравнения результатов методов понижения размерности для различных синтетических данных (где структура траектория заранее известна)
2. Сравнить методы понижения размерности по этим метрикам, следующие методы предполагается исследовать:

a. MDS
b. tSNE, approximated tSNE
c. Diffusion maps
d. ZINB-WaVE
e. SSLE (Supervised Locally Linear Embedding)
f. UMAP (Uniform Manifold Approximation and Projection)

3. Также, после сравнения существующих методов планируется исследовать возможность использования нейронных сетей (автоэнкодеров) для решения задачи понижения размерности и извлечения траекторий из данных.

   На данный момент

(a) изучены основы этих методов, разработаны критерии для их сравнения: способность сохранять расстояния между объектами, способность восстанавливать структуру траекторий в (1, 2), (2, 3) или (1, 3) измерениях, скорость работы метода. Для того, чтобы определить, восстанавливается ли структура, мы работаем с данными после уменьшения размерности как с графическими данными, пытаясь определить граф в «редуцированном» изображении.

(b) Были извлечены траектории при помощи Diffusion Maps и MDS. Оба эти метода неидеально справляются с задачей в двух измерениях.

   Дальнейшие планы: сравнить больше методов, взять больше датасетов синтетических данных, попробовать использовать нейронные сети.

**Инди-лес (случайный лес индикаторов попарных сравнений) как классификатор экспрессионных данных.**

***Александр Фаворов*, *Bahman Afsari, Leslie Cope***

Случайный лес (RF) как метод построения классификаторов из обучающей выборки (Breiman, 2001) хорошо зарекомендовал себя в очень широком спектре приложений, в том числе, и в биоинформатических задачах. Лес — это голосование большого числа (тысяч) простых классификационных деревьев, каждое из которых строится на случайной выборке образцов и случайной же выборке переменных. Такой подход во многих случаях позволяет избежать переобучения. Где лес, там и леший. Boruta (леший, pl) (Kursa and Rudnicki, 2010) выбирает важные для классификации переменные, строя несколько случайных лесов.

Элементарные решения, из которых состоит каждое дерево, сравнивают значения переменных с порогами, при этом пороги подбираются при обучении и становятся частью классификатора. Из-за этого классифицируемые экспрессионные данные надо нормализовать вместе с обучающей выборкой, что ограничивает применимость RF для экспрессионных задач. Для этих задач, с другой стороны, существует семейство непараметрических методов (Afsari et al., 2015, 2014; Leek, 2009), основанных на попарных сравнениях значений экспрессий разных генов внутри одного сэмпла, и не зависящих от нормализации, пока она монотонна.

Идея этого проекта — соединить эти два подхода, строить случайный лес, используя попарные сравнения как элементарные предикторы. Для пилотной версии использовать стандартные библиотеки для RF, затем, если получившийся инди-лес (IndiForest, это название проекта) будет эффективен, оптимизировать библиотеку для случая бинарных предикторов или найти существующую. Boruta сможет переводить обученные инди-леса на язык наборов генов, привычный для экспрессионного анализа.

Afsari, B., Fertig, E.J., Geman, D., Marchionni, L., 2015. switchBox: an R package for k–Top Scoring Pairs classifier development. Bioinformatics 31, 273–274. https://doi.org/10.1093/bioinformatics/btu622

Afsari, B., Geman, D., Fertig, E.J., 2014. Learning Dysregulated Pathways in Cancers from Differential Variability Analysis. Cancer Inform. 13, 61–67. https://doi.org/10.4137/CIN.S14066

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324
Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. J. Stat. Softw. 36, 1–13. https://doi.org/10.18637/jss.v036.i11
Leek, J.T., 2009. The tspair package for finding top scoring pair classifiers in R. Bioinforma. Oxf. Engl. 25, 1203–1204. https://doi.org/10.1093/bioinformatics/btp126

# Ботанический сад

## Изучение ядерного транскриптома *Rhopalocnemis phalloides* — нефотосинтезирующего, паразитического растения

***Михаил Щелкунов*[1, 2], *Максим Нуралиев*[3, 4], *Мария Логачёва*[1, 2, 3]**
1 — Сколтех, 2 — ИППИ, 3 — МГУ, 4 — Российско-вьетнамский тропический центр

В предыдущей работе мы обнаружили, что нефотосинтезирующее, паразитическое растение *Rhopalocnemis phalloides* обладает пластидным геномом с необычными свойствами. АТ-состав этого генома один из высочайших среди всех известных геномов - 86.8%, а скорость нуклеотидных замен в нём в несколько сотен раз выше, чем у близкородственных фотосинтезирующих растений. Причины этих явлений неизвестны. В новой работе, представленной на данной конференции, мы изучили свойства ядерного генома *Rhopalocnemis phalloides* путём изучения его транскриптома.

Транскриптомные риды были получены путём секвенирования кДНК на NextSeq 500. Было проведено сравнение транскриптома *Rhopalocnemis phalloides* с транскриптомом *Balanophora fungosa* (также нефотосинтезирующее растение из того же семейства, что *Rhopalocnemis phalloides*) и трёх фотосинтезирующих растений из близкородственных семейств.

Анализ наличия транскриптов генов, связанных с репарацией и рекомбинацией пластидного генома, показал, что в *Rhopalocnemis phalloides* примерно в 3 раза меньше таких генов. Это может объяснять, почему у *Rhopalocnemis phalloides* столь высокая скорость накопления замен. Также, поскольку в пластидных геномах генная конверсия ГЦ-смещённая, потеря этих генов может объяснять повышенный АТ-состав. Причина, по которой могла произойти потеря большого количества генов, связанных с пластидной репарацией и рекомбинацией, сейчас обсуждается.

Мы показали потерю у *Rhopalocnemis phalloides* и *Balanophora fungosa* абсолютно всех известных генов, связанных с фотосинтезом. Также, произошло исчезновение части генов, связанных с циркадными ритмами. Исчезновение генов, связанных с циркадными ритмами, может быть связано с тем, что *Rhopalocnemis phalloides* и *Balanophora fungosa*, как нефотосинтезирующие растения, бОльшую часть жизни проводят полностью под землёй. Они образуют короткоживущую надземную часть только на период цветения и плодоношения. В связи с этим, у них меньше потребность в регуляции жизни в соответствии с освещением.

Скорость нуклеотидных замен в ядерных генах *Rhopalocnemis phalloides* и *Balanophora fungosa* примерно в 3 раза выше, чем фотосинтезирующих родственников. Возможно, она

увеличена по тем же причинам, что и скорость замен в пластидном геноме. Мы предполагаем, что увеличение скорости замен может быть адаптивно, поскольку потеря фотосинтеза сопровождается разрушением генов, связанных с фотосинтезом. Ген, находящийся в процессе деградации, может давать вредный белок. Увеличение же скорости замен (например, путём снижения качества репарации генома) поможет быстрее довести деградирующие гены до состояния, при котором уже невозможна их транскрипция или трансляция, таким образом снижая количество вредных белков.

Анализ транскриптома *Rhopalocnemis phalloides* сейчас в процессе. Планируется, к примеру, проверить, произошло ли в ядерных генах увеличение АТ-состава, подобное тому, что случилось в пластидных генах. Также будет проведён сравнительный анализ количества GO терминов (GO terms enrichment analysis), который позволит дать общее описания изменениям в транскриптомах *Rhopalocnemis phalloides* и *Balanophora fungosa* в сравнении с фотосинтезирующими родственными видами.

[Работа проведена на средства гранта РФФИ № 16-34-01003 "Исследование изменений генома *Rhopalocnemis phalloides*, сопровождающих потерю способности к фотосинтезу"]

## Генетический и морфологический анализ изменения дорзовентральной полярности листьев среди представителей рода *Curio*

*Федотов А. П., Клепикова А. В., Пенин А.А., Тимонин А. К.*

Целью настоящей работы является проведение генетического и детального анатомо-морфологического исследования развития первично бифацаильных, субунифациальных, унифациальных и вторично бифациальных листьев у южноафриканских суккулентов из рода *Curio* и их сравнение для выявления пути перехода от одного типа листа к другому.

Листья большинства покрытосеменных растений закладываются на апексе побега в виде листового примордия морфологически недифференцированного на адаксиальную и абаксиальную стороны. Затем в процессе развития лист морфологически приобретает бифациальное строение, дифференцируется на Unterblatt и Oberblatt. Первый дает начало листовому основанию и прилистникам, а последний формирует черешок и листовую пластинку, которые в основном проявляют морфологические черты дорзо-вентрального строения. Однако у некоторых представителей рода Curio наблюдается отклонение от этой типичной схемы развития. У этих видов Unterblatt и образующееся из него очень короткое листовое основание имеют выраженное бифациальное строение. В то время как Oberblatt разрастается в вертикальной плоскости, приобретает более-менее цилиндрическую форму или становится шаровидным. При этом сформировавшийся первичный край уплощенного листового примордия исчезает. Этот процесс происходит за счет разрастания абаксиальной стороны при сохранении исходной ширины адаксиальной стороны, которая оказывается обхваченной первой на всем протяжении. Такие листья называют субунифациальными. Небольшая верхняя сторона у таких растений представлена так называемым «световым окном», представляющим собой паренхимную ткань с небольшим числом слабо развитых хлоропластов, в которой не развиваются проводящие пучки. У некоторых видов такое «световое окно» не развивается и выделить верхнюю сторону на морфологическом и анатомическом уровне становится невозможно. Такие листья называют унифациальными.

Также у группы видов в ходе разрастания происходит формирование вторичного края, что можно рассматривать с точки зрения морфологии как возвращение к бифациальности. Такой тип листьев называют вторично бифациальными. Стоит отметить, что у многих видов, в том числе первично бифациальных, самый кончик листа имеет унифациальное строение и предполагают, что он состоит только из нижней стороны листа.

В ходе работы предполагается провести исследование развития отдельных сторон листа методами сравнительной транскриптомики: de novo сборку транскриптома, анализ дифференциальной экспрессии и анализ регуляторных путей. Также планируется анатомо-морфологическое исследование ранних стадий развития листьев стандартными методами световой и электронной микроскопии.

## Распределение гаплотипов LFY в дальневосточных и европейских популяциях *Isoëtes*

***Григорьян М. Ю.***

Принято выделять 4 вида *Isoëtes*, которые встречаются на территории России. Все они занесены в Красную книгу России и нуждаются в охране. Из-за трудностей изучения, связанных с аллоплоидией, гибридизацией и морфологической схожестью, таксономический статус некоторых видов спорен.

Наибольшие проблемы вызывает таксономический статус дальневосточного *I. asiatica*. Морфологически *I. asiatica* не отличим от циркумбореального *I. echinospora* (Мочалова, 2006). Однако анализ AFLP и хлоропластных маркеров этих видов показал генетические различия (Kim et al, 2009). Впрочем, результаты AFLP не вызывают доверия.

Также необходима в уточнении видовая принадлежность некоторых популяций *Isoëtes* Командорских островов. Ранее они выделялись в отдельный вид *I. beringensis*, но из-за отсутствия морфологических различий сейчас их принято относить к *I. maritima* с западного побережья Северной Америки и Алеутских островов.

Помимо таксономических сложностей, вызывает интерес происхождение многих полиплоидных *Isoëtes*. В частности, *I. lacustris*, распространённый в Северной Атлантике, является декаплоидом. Видимо, он сформировался в результате многократной аллоплоидических скрещиваний. Предковые виды *I. lacustris* ранее не были установлены.

В прошлом году мы поставили задачу найти генетические различия *I. asiatica* и *I. echinospora,* а также определить вид Командорских популяций *Isoëtes* и выявить происхождеие *I. lacustris* при помощи клонирования половины второго интрона однокопийного гена LFY. Этот участок ранее использовался в работах по видообразованию *Isoëtes*.

В результате, мы получил 62 клона из 10 образцов. В составе генома *I. lacustris* присутствуют гаплотипы *I. prototypus, I. echinospora* и тетраплоида *I. tuckermanii*. Из-за недостаточного числа поднятых копий, необходимо продолжать изучение гаплотипов *I. lacustris,* возможно, более эффективными способами. Попутным результатом работы стало нахождение нового вида для республики Марий-Эл. По результатам клонирования, галотипы образца из Марий-Эл совпадали с гаплотипами *I. lacustris*. *I. lacustris* ранее не был найден на территории республики. Определение вида было подкреплено морфологическими данными. Сведения о находке уже опубликованы (Григорьян et al, 2018). Во всех дальневосточных

образцах была найдена последовательность, которая по некоторым оценкам (Larsen et al, 2016) отделилась от других последовательностей *Isoëtes* более 22 млн лет назад. В Командорских популяциях, эта копия присутствовала вместе с типичными для *I. maritima* последовательностями, а в *I. asiatica* – вместе с последовательностями *I. echinospora*. Возможны две интерпретации такого результата. Так как мы использовали немного отличные от предыдущих работ праймеры, мы могли поднять ранее не найденную копию LFY. Также возможно, что дальневосточные популяции в прошлом скрестились с древним ныне вымершим видом *Isoëtes*, и эта древняя копия является частью его генома.

Для определения природы этой копии мы постарались в этом году применить комплексный подход к проблеме – для 24 растений с Дальнего Востока были получены снимки скульптуры микроспор и макроспор при помощи SEM. Кроме того, планируется секвенирование ранее изучаемого участка на Illumina MiSeq, а также определение плоидности при помощи цитометрии и подсчёта хромосом.

## Анализ происхождения аллотетраплоида *Capsella bursa-pastoris* с использованием данных полногеномного секвенирования

***Шнайдер Э.Д., Клепикова А.В., Касьянов А.С., Логачева М.Д., Пенин А.А.***

Полиплоидизация является ключевым событием в эволюции высших растений. Тем не менее, вопрос о том, что происходит с геномом после полногеномной дупликации у представителей различных популяций, по-прежнему остается открытым. *C. bursa-pastoris* – недавний аллотетраплоид, произошедший 100.000–300.000 лет назад в результате межвидовой гибридизации двух диплоидных видов C. orientalis и предка C. rubella/grandiflora. В отличие от своих родителей, имеющих небольшой ареал обитания, *C. bursa-pastoris* является одним из самых распространенных растений на планете, что свидетельствует о колоссальной экологической пластичности представителей этого вида. Однако, действительно ли представители разных популяций - это один вид? Морфологические и геномные данные показывают, что под названием «*C. bursa-pastoris»* могут объединяться как минимум 3 вида, дивергировавших не так давно, а, возможно, даже имеющих разное происхождение. Задачей работы является анализ происхождения аллотетраплоида Capsella bursa-pastoris с использованием последовательностей пластидного, митохондриального и ядерного геномов. К настоящему времени проведено полногеномное секвенирование 29 линий из разных частей света и начат анализ полученных данных. В том числе проведено изучение полиморфизма пластидного генома. Результат анализа мононуклеотидных сайтов показывает, что линии из ближневосточных (Израиль, Марокко), а так же английских популяций резко отличаются от европейских и дальневосточных линий. В дальнейшем планируется расширить анализ на ядерные и митохондриальные геномы.

## Accurate Fetal Variant Calling in the Presence of Maternal Cell Contamination

***Elena Nabieva*[1,2]*, *Satyarth Mishra Sharma*[1*], *Yermek Kapushev*[1], *Sofya K. Garushyants*[1,2], *Georgii A. Bazykin*[1,2], *and Dmitry Yarotsky*[1,2]**

[1] Skolkovo Institute of Science and Technology, Skolkovo, Russia, and

[2]Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow, Russia

*

These authors contributed equally to this work.

High-throughput sequencing of fetal DNA is a promising and increasingly common method for the discovery of all (or all coding) genetic variants in the fetus, either as part of prenatal screening or diagnosis, or for

genetic diagnosis of spontaneous abortions. In many cases and depending on the tissue and the laboratory technique used, the fetal DNA (from chorionic villi, amniotic fluid, or abortive tissue) can be contaminated

with maternal cells, resulting in the mixture of fetal and maternal DNA. This maternal cell contamination (MCC) undermines the assumption, made by traditional variant callers, that each allele in a heterozygous

site is covered, on average, by 50% of the reads, and therefore can lead to erroneous genotype calls.

:

We present a panel of methods for reducing the genotyping error in the presence of MCC. All methods start with the output of GATK HaplotypeCaller on the (contaminated) fetal sample and sequencing data

for both parents, and additionally rely on information about the MCC fraction (which is readily estimated from the HTS data). The first of these methods uses an explicit formula based on simple probabilistic

assumptions to "recalibrate" the fetal genotype calls produced by MCC-unaware HaplotypeCaller. The other two methods "learn" the recalibration model from examples. We use simulated contaminated fetal

data to train and to test the models. Using the test sets, we show that all three methods lead to substantially improved accuracy when compared with the original MCC-unaware HaplotypeCaller calls, as judged by the

concordance with genotype calls made on uncontaminated child data. We then apply the best-performing method to three chorinonic villus samples from spontaneously terminated pregnancies.

# Факторы транскрипции

## Сравнительно-геномный анализ факторов транскрипции семейства IclR и их сайтов связывания: особенности структуры и ко-эволюция

***Суворова И.А.***

Семейство IclR – крупная группа бактериальных факторов транскрипции. Регуляторы этой группы действуют как репрессоры и/или активаторы, имеют в составе N-терминальный HTH домен и связываются с ДНК в виде димеров или тетрамеров. Известно, что факторы транскрипции семейства IclR регулируют самые разнообразные клеточные процессы, включая, например, метаболизм ароматических соединений, кворум-сенсинг, механизмы лекарственной устойчивости, вирулентность и т.д. Наиболее полно описанным является репрессор глиоксилатного шунта IclR у *E. coli*. Несмотря на то, что ряд факторов транскрипции семейства IclR достаточно хорошо изучен, для многих других неизвестна функция и/или мотив связывания. Предполагаемым общим консенсусным мотивом связывания для регуляторов группы IclR часто считается последовательность TGRAAAWNNTTTYY, несмотря на то, что ей не соответствует ряд известных мотивов, например, MhpR (GGTGCACCTGGTGCACA), PcaR (GTTCGATAATCGCAC), PobR (TGTCCGATGATCGGACA) и т.д.

Целью данной работы является предсказание мотивов связывания и реконструкция регулонов для факторов транскрипции семейства IclR, а также выявление общих закономерностей ДНК-белковых взаимодействий регуляторов этого семейства и сопоставление их с таковыми у регуляторов семейства GntR, также имеющих в составе ДНК-связывающий HTH домен.

Было исследовано 1443 регулятора семейства IclR в 327 полных геномах бактерий, для которых удалось предсказать потенциальные мотивы связывания методом филогенетического футпринтинга. Из 4437 предсказанных сайтов связывания: 971 соответствует консенсусу NNWWY**GT**N**CG**WWWWW**CG**N**AC**NNWNN; 414 – WWT**T**C**CG**CWWWG**CG**G**A**AWW; 464 – GTWN**GT**YYNNNWWNNNRR**AC**NWAC; 1492 – TKNN**RT**TY**CR**YWWWR**YG**RA**AY**NNMA; и 1096 прочих, включая такие варианты мотивов, как ATGGAAWWWWTTCCAT, TGAAAAANTTTTTCA, AATGAAAGTNACTTTCATT, WWWGGAAYNNNRTTCCWWW, GTTGGAAAAATTTTCCAAC, имеющие сходство с ранее предполагаемым консенсусом TGRAAAWNNTTTYY.

Планы дальнейшего исследования:

- Анализ корреляций последовательностей HTH доменов регуляторов и соответствующих сайтов связывания для каждого из основных вариантов консенсусных мотивов (с помощью ProtDNAKorr)
- Сопоставление полученных результатов для каждого из вариантов, а также сравнение с данными для семейства GntR
- Более детальная метаболическая реконструкция и описание функций регуляторов для наиболее крупных ортологических групп семейства IclR

## Реконструкция предковой последовательности белка OxyR

***Н.О. Драненко, О.О. Бочкарёва***

Постановка задачи:

Реконструировать предковые последовательности белка OxyR для различных бактериальных таксонов для последующей экспериментальной проверки кинетики реакции связывания полученных белков с лигандом. Белок является биосенсором для перекиси водорода, результат исследования может быть полезен в различных медицинских анализах для измерения уровня окислительного стресса в клетке.

Результаты:

Отобрано 499 последовательностей белка OxyR из базы данных OMA, удовлетворяющих установленному порогу сходства с хорошо изученным белком OxyR из *E. coli* K-12 и прошедших проверку на наличие сайта связывания с перекисью водорода.

Для этих последовательностей построено филогенетическое дерево методом NJ.

Качество множественного выравнивания всех 499 последовательностей не позволяет достоверно произвести реконструкцию общего предка, поэтому на полученном дереве было выделено несколько крупных узлов, в каждом из которых была восстановлена предковая последовательность.

Для уточнения выравнивания были использованы данные о пространственной и вторичной структурах 6 белков OxyR из базы данных PDB. Кроме того, при анализе структур белка OxyR был найден белок с высоким сходством пространственной структуры и последовательности со структурой и последовательностью OxyR из *E.coli*, белок CynR. Для белка CynR была проведена аналогичная процедура реконструкции предковых последовательностей в узлах дерева.

Дальнейшие планы:

Получить общего предка белков OxyR и CynR.

Для белка OxyR получить кодирующие последовательности ДНК, оптимизированные под *E.coli*, синтезировать белки, соответствующие полученным предковым последовательностям, и померить кинетику реакции их связывания с перекисью

## Ко-эволюция факторов транскрипции и их сайтов связывания

***Инна Суворова, Кадыкова Татьяна***

Исследования ДНК-белковых взаимодействий популярны и актуальны в молекулярной биологии. Многие особенности этих взаимодействий не идентифицированы до сих пор.

Цель проекта – изучение ко-эволюции транскрипционных факторов и их сайтов связывания в ДНК и исследование особенностей их взаимодействий. В качестве объекта исследования было выбрано семейство транскрипционных факторов GntR.

Белки семейтсва GntR имеют очень схожий N-концевой HTH домен связывания, но различаются в С-концевом домене связывания с эффектором. Оно делится на многие подсемейства. В этом проекте исследовались подсемейства FadR, HutC, YtrA.

В подсемействе HutC ко-эволюция не выражена, так как деревья не очень похожи по своей структуре. Однако некоторые ортологические группы находятся рядом как на дереве белков, так и на дереве мотивов. Например, NagR, AgaR и AgaR3, а также PhnF2 и DasR, находятся близко.

В подсемействе YtrA такая же ситуация, как в HutC. Ортологические группы EF1676, XAC1548, TM0766, TTE0438 остаются рядом, но в дереве мотивов к этой группе присоединяется SO0072.

Однако в подсемействе FadR дерево мотивов поделилось на две половины: верхнюю и нижнюю. Дерево мотивов также делится на две половины, коррелируя с разделением дерева транскрипционных факторов. Предполагается, что это разделение связано с важными различиями в аминокислотах и нуклеотидах, отвечающих за связь между транскрипционным фактором и сайтом связывания.

Были построены лого двух половин сайтов. Наиболее весомые нуклеотиды немного отличаются. В верхней половине выражены TGGT и ACCA, в то время как в нижней – TTGT и ACAA. Эти нуклеотиды имеют наиболее сильно взаимодействуют с транскрипционным фактором.

Далее были проанализированы деревья транскрипционных факторов и выявлены позиции HTH домена, на которых группы белков имеют наиболее важное различие. Получилось, что они различаются в 26, 28, 39, 40, 58, 59 и 62 позициях. Оказалось, что все эти позиции, за исключением 26, отвечают за связывания с сайтами. Это было выявлено в диссертации на примере *E. coli.*

Подводя итог, подсемейство транскрипционных факторов FadR делится на две половины, как и их сайты, в зависимости от аминокислот и нуклеотидов, которые связываются друг с другом. Однако в подсемействах HutC и YtrA ко-эволюция сильно не была выражена, только на уровне некоторых совпадений ортологичеких групп.

## Консервативность неконсенсусных позиций в сайтах связывания факторов транскрипции.

***Белоусова Е. А.***[1]

[1]Факультет Биоинженерии и Биоинформатики, МГУ им. М. В. Ломоносова, Москва, Россия bilyius@mail.ru

Ранее [1] был замечен феномен консервативности неконсенсусных позиций в сайтах связывания факторов транскрипции: некоторые неконсенсусные основания сайта медленнее эволюционируют в конкретных группах бактерий. Это явление хорошо заметно только для глобальных регуляторов, для которых существуют большие регулоги – группы регулонов из близких видов.

Чтобы оценить, насколько значима консервативность неконсенсусных позиций, эти позиции можно сравнить с третьими позициями четырехвырожденных кодонов генов, транскрипция которых регулируется данным фактором, что и было сделано в 2005 году [1]. Оказалось, что изучаемые неконсенсусные основания, действительно, более консервативны, чем третьи позиции. Этому может существовать несколько объяснений. Во-первых, возможно

перекрывание сайтов связывания регуляторов транскрипции, и тогда консервативные «неконсенсусы» относятся к еще неизвестным регуляторным элементам. Во-вторых, замена неконсенсусного основания на консенсусное меняет сродство регулятора к сайту, а значит, и уровень транскрипции гена. То есть, если неконсенсусное основание однажды начало обеспечивать нужный уровень регуляции, то оно уже не может «переместиться» на другую позицию.

Благодаря развитию полногеномных технологий, число известных сайтов регуляторов транскрипции в последние годы существенно возросло. В данной работе стоит задача найти хорошо выраженные консервативные неконсенсусы в ортологичных сайтах, сравнить их консервативность с нейтрально (или почти нейтрально) эволюционирующими элементами и объяснить это явление. Мы начали с регулятора CcpA. Этот фактор обеспечивает транскрипционный ответ на появление быстроусвояемых углеводородов в среде [2]. Логотип сайта связывания CcpA, полученный с помощью WebLogo [3], представлен на рисунке 1. В данном сайте были обнаружены позиции с консервативным неконсенсусным основанием в нескольких группах ортологичных сайтов. По предварительной грубой оценке они, действительно, более консервативны, чем третьи основания четырехвырожденных кодонов (пример на рисунке 2). Однако, это наблюдение требует дальнейшей проверки.

Данная работа производится совместно с М. С. Гельфандом. Исследование поддержано грантами РНФ 18-14-00358 и РФФИ 18-34-01006.

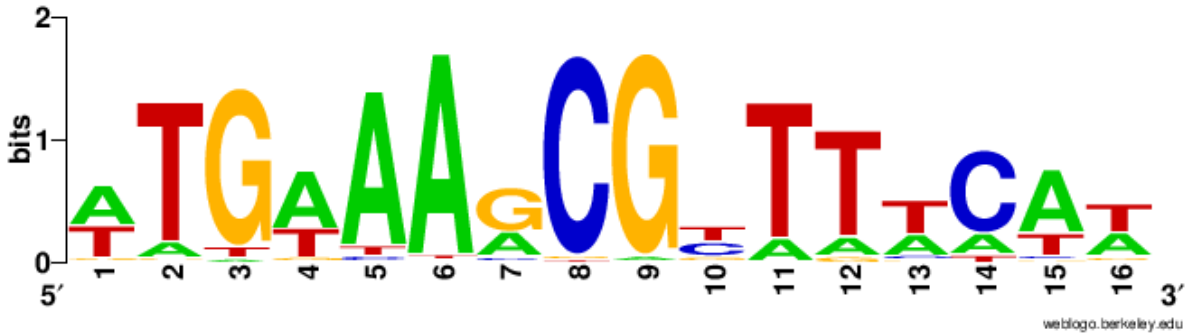



**Рис. 1.** Логотип сайта связывания CcpA, полученный с помощью WebLogo.

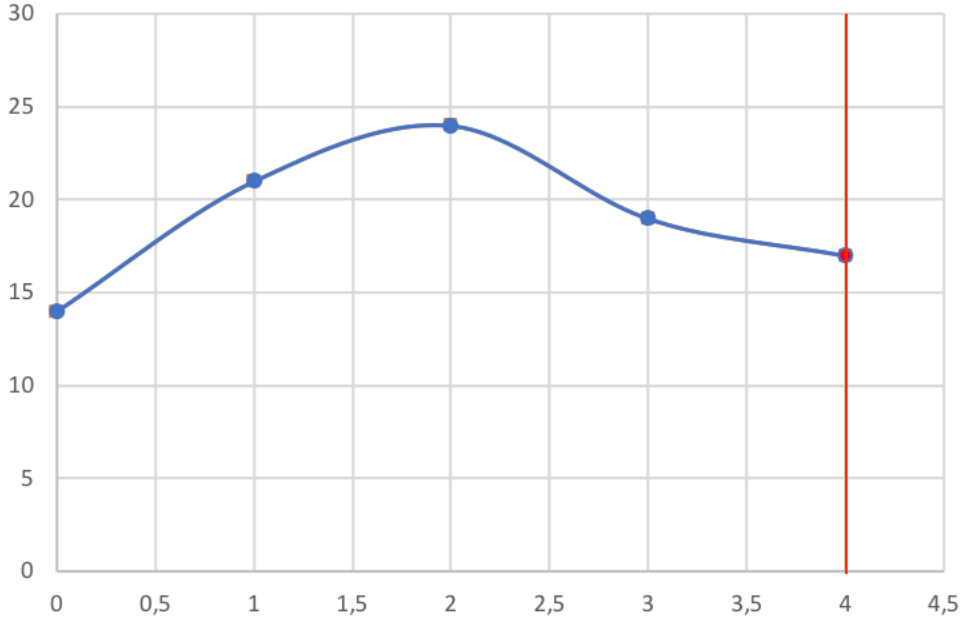


**Рис. 2.** Распределение числа «незамен» аденина в третих позициях четырехвырожденных кодонов в шести ортологах гена *citZ*. А в девятой позиции четырех из шести ортологичных сайтов связывания CcpA находится неконсенсусный аденин. Четыре «незамены» находятся на самом краю распределения.

**Ссылки**


1. Kotelnikova, E., Makeev, V., Gelfand, M.: Evolution of transcription factor DNA binding sites. Gene 347(2):255-63 (2005)
2. RegPrecise, http://regprecise.lbl.gov/RegPrecise/index.jsp
3. WebLogo, http://weblogo.berkeley.edu/logo.cgi


## Классификация ДНК-метилтрансфераз прокариот по сходству последовательностей.


***М. Ю. Серебренникова, И. С. Русинов, О.И. Безсуднова, А.С.Попенко, А.В. Алексеевский***


ДНК прокариотических и эукариотических клеток и их вирусов часто модифицируется путем ферментативного метилирования, осуществляемого S-аденозил-L-метионин (AdoMet) - зависимыми ДНК-метилтрансферазами (MTases). У прокариот метилирование ДНК выполняет роль меток сайтов рестрикции систем рестрикции-модификации, для предотвращения их расщепления родственными рестрикционными эндонуклеазами, а также может участвовать в репарации несоответствия ДНК, регуляции экспрессии генов и контроле времени репликации ДНК [1].

Существует несколько классификаций МТаз, основанных на различных свойствах: по метилированной группе атомов ДНК; по последовательности консервативных мотивов; по типам рестрикционно-модификационных систем, по особенностям процесса метилирования. Однако общепринятой классификации на основе сходства последовательностей не существует.

Для разработки такой классификации мы загрузили все 132835 последовательностей МТаз (версия от 03.10.2018) из REBASE [2]. В Pfam, базе данных семейств доменных доменов, нами были обнаружены девять семейств, аннотированных как каталитические домены ДНК-метилтрансфераз: PF05869, PF00145, PF07669, PF13651, PF02086, PF05063, PF02384, PF01555, PF12564.

Используя профили HMM от Pfam, мы определили архитектуру домена, то есть последовательность всех доменов Pfam в каждой из 132835 последовательностей МТаз. Мы разделили все МТазы на три группы: группа (1) включает 25318 МТаз без доменов Pfam в последовательности; группа (2) включает 6116 МТаз, не имеющих каталитического ДНК-метилтрансферазного домена из нашего списка; группа (3) включает 101401 МТаз, имеющих каталитический домен ДНК-метилтрансферазы из нашего списка, некоторые из них включают более одного каталитического домена.

МТазы групп (2) и (3) были разделены на группы с одинаковыми доменными архитектурами. Для архитектуры из 100 доменов с более чем 10 МТазами (всего 106752 МТаз) мы построили множественное выравнивание последовательностей.

Наши дальнейшие планы - решить следующие задачи:

Почему последовательности группы (1) вообще не имеют доменов Pfam? Они были пропущены или не найдены?

Правда ли, что последовательности группы (2) действительно имеют каталитические домены, не учтенные в списке из 9 каталитических доменов ДНК-метилтрансфераз? Если это так, то мы должны их обнаружить.

Каков источник МТаз из группы (3), имеющей более одного каталитического домена? Одним из предполагаемых источников является довольно частая ошибка аннотации в Pfam: фактический домен в последовательности белка может быть представлен в Pfam как два последовательных одинаковых домена, соответствующих N-концевой и C-концевой частям домена, из-за низкого сохранения последовательности между их.

Являются ли МТазы с одинаковой доменной архитектурой гомологичными? В белках каталитические домены являются наиболее стабильными. Поэтому мы планируем проверить, что каталитические домены из последовательностей с одинаковой архитектурой доменов образуют клады в филогенетическом дереве этого каталитического домена.

Работа поддержана РНФ грантом 16-14-10319.

1. Bujnicki JM. Sequence permutations in the molecular evolution of DNA methyltransferases. BMC evolutionary biology. 2002 Dec;2(1):3.

**2.** Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2015 (Database issue):D298-9.

## Метаболизм гексуронатов у гаммапротеобактерий: как UxuR и ExuR делят свои регулоны между собой и с малыми РНК?

***Мария Тутукина*, Инна Суворова, Анна Казнадзей, Ульяна Швырева, О.Н. Озолинь, М.С. Гельфанд**

Метаболизм гексуроновых кислот по пути Эшвелла и Энтнера-Дудорова обеспечивает энергетически более выгодную альтернативу гликолизу и играет важную роль в подвижности бактерий и колонизации ими организмов хозяев. Ранее было предсказано, что гомологичные белки UxuR и ExuR подавляют синтез ферментов, необходимых для метаболизма гексуронатов, но как именно они взаимодействуют для обеспечения сбалансированной регуляции, до сих пор неясно.

Нашей задачей было выявить регулоны UxuR и ExuR , используя данные ChIP-seq и RNA-seq, а также проанализировать возможность формирования ими гетеродимеров с помощью электрофореза с задержкой в геле и сравнения данных ChIP-seq и SELEX.

Анализ ChIP-seq и экспрессионных данных выявил около 360 мишеней UxuR в геноме *E. coli* K-12 MG1655. Среди них есть не только гены, кодирующие ферменты сахарного метаболизма, но также гены, отвечающие за метаболизм и транспорт железа, образование жгутиков и флагелл, и несколько регуляторов. При этом около 70% всего белка в клетке связывается с 11 основными мишенями, все из которых отвечают за метаболизм сахаров, и это связывание было чувствительным к D-глюкуронату, интермедиату пути Эшвелла. Наиболее сильный эффект был обнаружен для оперона *uxuAB*, кодирующего D-маннонатдегидратазу и оксидоредуктазу. По данным экспериментов SELEX *in vitro, uxuAB* является единственной мишенью UxuR. По нашим данным EMSA и плазмонного резонанса получилось, что ExuR не

способен взаимодействовать с его промоторной областью ни отдельно, ни в комплексе с UxuR. При этом со всеми остальными основными мишенями UxuR ExuR связывается, но с гораздо большей константой. Для самого ExuR было обнаружено более 300 мишеней как по результатам ChIP-seq, так и SELEX, эффективность связывания с которыми почти не зависела от добавления D-глюкуроната или D-галактуроната. Это указывает на то, что ExuR является менее специфическим глобальным регулятором. При росте на глюкозе мишени UxuR и ExuR пересекаются всего на 12%. Однако во время роста с D-глюкуронатом более 60% их мишеней становятся общими, что, в совокупности с данными SELEX для UxuR, может говорить о лиганд-зависимом образовании гетеродимера UxuR-ExuR. Их способность образовывать гетеродимеры была подтверждена EMSA.

Несмотря на исходно предполагаемую узкую специфичность UxuR, его регулон включает гены, кодирующие FNR, Fis, системы усвоения железа и малые РНК, например, DsrA. Это может быть неслучайно, так как в конце самого UxuR инициируется синтез нескольких потенциальных малых РНК, как в прямом, так и в антисмысловом направлении. Одна из них - UxuT – начинается в терминаторной шпильке гена, имеет длину 79 нт и способна регулировать транскрипцию *uxuAB* за счет связывания с промоторной областью, а также стабилизировать *uxuR*- и *hns*-мРНК за счет взаимодействия с их 3'-концами. Недавние результаты секвенирования тотальной фракции всех РНК кишечной палочки длиной менее 40 нт показали, что UxuT перекрывается с другими короткими транскриптами, которые синтезируются из гена *uxuR* в антисмысловом направлении. Удивительно, но эти короткие РНК были значительно представлены во фракции секретируемых внеклеточных РНК во время роста бактерий на среде M9/глюкоза, при этом они отсутствовали при росте кишечной палочке на богатой среде LB. При сравнении данных РНК-сек в штамме с полностью удаленным геном *uxuR* (удалены и все РНК) и штамме с его выключенной трансляцией оказалось, что изменение экспрессии большого количества генов опосредовано именно РНК – вероятно, за счет их связывания со шпилечными структурами в их регуляторных областях. Таким образом, по-видимому, сеть регуляторов UxuR и ExuR намного сложнее, чем ожидалось, и может включать новый класс регуляторных РНК с функцией, которую еще предстоит понять.

## Альтернативное кодирование и синтез нескольких белков с одного гена в гаммапротеобактериях на примере LeuO

***Татьяна Бессонова, Евгения Белоусова, Софья Гарушянц, Мария Тутукина***

Белок LeuO - это глобальный регулятор транскрипции двойного действия, принадлежащий к семейству LysR и предположительно регулирующий гены вирулентности, кворум-сенсинга и общего метаболизма. Изначально наш интерес к *leuO* возник из-за сильно возросшего уровня его внутригенной транскрипции как в прямом, так и антисмысловом направлении, в штамме с удаленным геном нуклеоидного белка Dps. Это интересно, так как LeuO может взаимодействовать с белками нуклеоида, тем самым тоже осуществляя регуляцию на уровне изменения конформации ДНК. Кроме того, при суперпродукции этого белка с плазмиды было обнаружено несколько его изоформ, что очень похоже на YjjM (в *E. coli*) и VirF (в *Shigella*), которые также синтезируются в виде нескольких форм белка с укороченных мРНК транскриптов в различных условиях и отвечают за вирулентность в протеобатериях.

Нашей задачей стала проверка возможности альтернативного кодирования в локусе гена, кодирующего LeuO.

На первом этапе мы составили полную карту промоторов для локуса *leuO*, в том числе, стартов антисмысловой транскрипции. Был обнаружен ранее неизвестный промотор для синтеза в прямом направлении на расстоянии 2 bp даунстрим от основного ATG кодона, с которого может инициироваться синтез укороченной мРНК гена. Для анализа изоформ LeuO, синтезирующихся в кишечной палочке в разных условиях, были получены штаммы *Escherichia coli* K-12 MG1655 с His-tag на С-конце белка LeuO. С помощью Вестерн-блот был детектирован синтез трех форм белка разной длины *in vivo*, что соотносится с данными множественного выравнивания по консервативным стартам трансляции среди гаммапротеобактерий. Играет ли регуляторную роль неосновные формы белка (массой 31.2 кДа и 29 кДа), ещё предстоит выяснить. При сравнении данных SELEX (LeuO) и Chip-seq и SELEX (YjjM) было найдено 6 общих генов мишеней для LeuO и YjjM, связанных с вирулентностью бактерий. Влияние LeuO на эти мишени было подтверждено qRT-PCR. Для *yjjQ* и *acrE* он выступает в качестве активатора, а для *tsr* и *sdiA* репрессором, данные для *fes* и *fepA* неоднозначны и требуют дополнительной статистической проверки.

В ближайшем будущем необходимо построить более масштабные филогенетические деревья и выравнивания для оценки консервативности потенциальных изоформ белка. Мы планируем очистить белок LeuO (если возможно с разделением его форм), получить белок-специфические антитела и провести Chip-seq анализ для выявления всех возможных мишеней LeuO (или опять же его форм) уже *in vivo*. Для оценки взаимовлияния LeuO, YjjM и белков нуклеоида планируется сравнить данные Chip-seq анализа для LeuO, YjjM, Dps и H-NS. Кроме того, видимо, необходим ДНК-сэмплинг анализ для выявления всех белков связанных с регуляторной областью *leuO*.

Исследования поддержаны грантом РФФИ 18-34-01006

## Особенности формы ДНК, возникающие при связывании с транскрипционными факторами

***Столяренко Артемий, 1 курс магистерской программы АДБМ НИУ ВШЭ***
Научный руководитель – Спирин Сергей Александрович

Планируется исследовать параметры формы двойной спирали ДНК в комплексах с транскрипционными факторами. Будут рассмотрены структуры комплексов, включающих транскрипционные факторы из различных семейств. Семейства предполагается определять согласно базе белковых семейств Pfam (http://pfam.xfam.org/). Целью является поиск семейств, белки из которых вызывают при связывании с ДНК достоверные специфические изменения формы двойной спирали, и описание этих изменений. Для анализа пространственных структур ДНК будут использоваться пакеты x3dna и Curves+. Полученные результаты будут отображены в базе данных структур комплексов белков с нуклеиновыми кислотами NPIDB http://npidb.belozersky.msu.ru/.

## Предсказание влияния однонуклеотидных вариантов на экспрессию генов по данным параллельного репортерного анализа

***Пензар Д.Д, Зинкевич А.О, Воронцов И. Е, Василий В. Ситник В. В, Александр В. Ф., Макеев В. Ю., Иван В. Кулаковский И. В.***

Предсказание влияния однонуклеотидных замен в геноме на экспрессию генов является одной из важных задач регуляторной геномики, решение которой имеет практическое значение для персонифицированной медицины.

Мы использовали современные методы машинного обучения и данные параллельного репортерного анализа [1, 2] для предсказания влияния однонуклеотидных замен в регуляторных областях генома на экспрессию генов, находящихся под их контролем.

В качестве признаков для машинного обучения использовались различные источники данных: эпигенетические разметки (DNase-Seq, ATAC-Seq, ChIP-Seq), анализ мотивов в последовательностях (с помощью Perfectos-APE [3] и HOCOMOCO [4]) и признаки-результаты работы существующих алгоритмов машинного обучения. В последнюю категорию входили признаки, полученные с последнего слоя нейронной сети DeepSEA[5] и предсказания моделей, построенных с помощью метода опорных векторов deltaSVM [6].

Учебная выборка соревнования CAGI включала в себя информацию из различных районов 14 репортерных конструкций, другие районы которых составляли валидационную выборку.

В ходе сравнения различных моделей, нами было показано, что признаки из DeepSEA позволяют наилучшего качества предсказания, по сравнению с опубликованными решениями соревнования CAGI2018 – Regulation Saturation.

Как нам удалось показать, высокое качество предсказаний по большей части достигается за счет утечки информации из соседних регионов репортеров, доступных в обучающей выборке. Исключение регионов предсказываемого репортера из обучения существенно снижало качество предсказания (для некоторых конструкций - вплоть до уровня случайного предсказания).

Валидация на данных двух независимых репортеров [2] подтвердила, что утечка информации вносит существенный вклад с достижимое качество.

В ходе дальнейшей работы перспективным представляется минимизировать утечку информации на этапе обучения (с помощью комбинации различных предсказателей и использовании метапредсказателей (использующих информацию от других предсказателей)) и поиск альтернативных источников данных для более точной валидации полученных моделей.

**Ссылки**

1. **The Critical Assessment of Genome Interpretation**. https://genomeinterpretation.org

2. **Massively parallel functional dissection of mammalian enhancers in vivo.** Patwardhan P. P. et al., 2012

3. **PERFECTOS-APE - Predicting Regulatory Functional Effect of SNPs by Approximate P-value Estimation.** Vorontsov et al., 2015

4. **HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis.** Kulakovskiy et al., 2017

5. **Convolutional neural network architectures for predicting DNA–protein binding.** Zeng et al., 2016

6. **A method to predict the impact of regulatory variants from DNA sequence.** Lee et al., 2015

# Улучшение предсказания сайтов связывания транскрипционных факторов с помощью машинного обучения

***Кравченко П.А., Пензар Д.Д., Воронцов И.Е, Кулаковский И.В.***
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики; Институт общей генетики им. Н.И.Вавилова РАН. Москва, Россия
E-mail: pavel-kravchenk0@yandex.ru

ДНК-паттерны, распознаваемые белками-регуляторами транскрипции (транскрипционными факторами, ТФ), традиционно представляются в виде позиционно-весовых матриц (ПВМ), которые предполагают независимость соседних нуклеотидов в сайтах связывания. В настоящее время предложено множество альтернативных моделей, учитывающих корелляции между соседними позициями, однако ПВМ продолжают широко использоваться на практике.

В частности, одним из способов построения уточненных моделей является объединение нескольких ПВМ в решающее дерево [1], при этом исходный подход не показал значительного прироста точности распознавания сайтов связывания ТФ, по сравнению с простой ПВМ. В то же время, наличие результатов десятков независимых экспериментов для одного фактора транскрипции позволяет построить множество ПВМ и затем применить современные методы машинного обучения, такие как градиентный бустинг, для построения объединенного классификатора [2].

В нашей работе мы использовали ПВМ, построенные по данным ChIP-Seq экспериментов, представленных в базе

GTRD (Gene Transcription Regulation Database)[3] и ПВМ, полученные на их основе в ходе построения коллекции мотивов связывания ТФ мыши и человека HOCOMOCO [4]. Предсказания индивидуальных ПВМ использовались как признаки для обучения итоговой модели. В качестве негативной выборки использовались последовательности схожих длин, являющиеся сайтами связывания факторов транскрипции других структурных семейств..

Нам удалось продемонстрировать, что модель, построенная с использованием множества ПВМ,, позволяет значительно улучшить точность предсказания сайтов для различных ТФ, причем эффект сохраняется при предсказании сайтов связывания ТФ мыши при обучении на ChIP-Seq для ТФ человека и обратно.

Реализация проекта доступна в репозитории GitHub по адресу:
github.com/Pavel-Kravchenko/TF-ML

**Список литературы**

[1] "Tree-Based Position Weight Matrix Approach to Model Transcription Factor Binding Site Profiles" Yingtao Bi, et al. (2011)

[2] XGBoost: A Scalable Tree Boosting System. Tianqi Chen and Carlos Guestrin (2016)

[3] GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. I.S. Yevshin, R.N. Sharipov, T.F. Valeev, A.E. Kel, F.A. Kolpakov. (2016)

[4] "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis" Ivan V Kulakovskiy, et al. (2018)

# Трехмерная структура хроматина

## *Уровень ацетилирования гистонов регулирует формирование топологически ассоциированных доменов в хроматине дрозофилы*

*Екатерина Храмеева, Александра Галицына*

Recent progress of next-generation sequencing methods for 3D chromatin organization analysis [2] has unraveled many details of its fine structure. In particular, chromosomes of higher eukaryotes have been shown to be organized into spatially compact Topologically Associating Domains (TADs) [1]. *Drosophila melanogaster* is a popular model organism for chromatin studies, however, the mechanism of TAD formation is not yet well-known there. Here, we test the hypothesis that the mechanism of TAD self-assembly is based on the ability of nucleosomes from inactive chromatin to aggregate, and on the lack of this ability in acetylated nucleosomal arrays [3]. We analyzed data of Hi-C and Chip-Seq (with antibodies against pan acetylated H3 histone) experiments in control *D. melanogaster* late embryonic (Schneider-2) cells, as well as in HDAC1-depleted cells and in cells treated with histone acetyltransferase inhibitor curcumin or histone deacetylase inhibitor trichostatin A. Acetylation level changes were studied, in association with TAD position and density differences. Inhibition of HDAC1 was found to lead to an increase of acetylation level in interTAD regions, and coordinated changes of TAD structure. Thus, histone acetylation plays a key role in the mechanism of TAD formation.

1. J.R. Dixon et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature, 485:376-380.
2. E. Lieberman-Aiden et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science, 326:289-293.
3. S.V. Ulianov et al. (2016) Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains, Genome Research, 26(1):70-84.

## Структура хроматина: видовое разнообразие

***Александра Галицына, Сколтех, ИППИ, ИБГ***

Благодаря развитию методов биоинформатики для анализа данных секвенирования и фиксации конформации хромосом, становятся доступными геномные и интерактомные данные для ДНК организмов разнообразных групп: насекомых, хордовых, круглых червей, бактерий; а также удобные и эффективные способы их анализа.

Данная работа представляет собой набор проектов, общим лейтмотивом которых является мотивация исследования структуры хроматина, а также молекулярных факторов, влияющих на эту структуру. Кроме того, проекты объединены общей технической

поддержкой, для всех них используется или планируется использование современнейших программ вычислительной биологии: картирование данных секвенирования с помощью bwa mem [1], анализ контактов ДНК с помощью pairsamtools [2], cooltools [3], анализ мотивов связывания белков с помощью программ пакета autosome [4], визуализация в лабораторном браузере HiGlass [5].

Далее приводится короткое описание и статус каждого из проектов: (1) эволюция CTCF круглых червей, (2) структура хроматина в эмбриогенезе *Danio rerio*, (3) факторы хроматина *Escherichia coli*.

(1) CTCF – архитектурый белок хроматина, играющий главную роль в инсуляции регуляторных единиц генома. Его практически повсеместное присутствие в геномах билатерий позволяет предположить его главную роль в формировании укладки трехмерной структуры хроматина (в частности, по механизму экструзии петель). Единственная группа билетарий без найденного CTCF в геноме – круглые черви (в частности, *Caenorhabditis elegans*). Анализ структуры хроматина *C. elegans* с помощью Hi-C позволяет предположить отсутствие ТАДов у этого организма. Важным наблюдением является также отсутствие обогащения генома мотивами связывания CTCF практически у всех круглых червей. [6]

Однако у одного представителя круглых червей, *Trihinella spiralis,* CTCF найден, однако нет обогащения мотивами CTCF. Данный феномен позволяет предположить промежуточную стадию эволюции круглых червей на пути к потере CTCF. Проект посвящен исследованию этого явления и верификации находок работ предшественников.
На данный момент подготовлен сравнительный анализ геномов *C. elegans* и *T. spiralis,* проведен поиск мотивов CTCF в геномах этих организмов, сделано полногеномное выравнивание для поисков участков синтении.

Работа выполняется в коллаборации с Сергеем Ульяновым (ИБГ).

(2) Эмбриогенез – важнейший процесс развития хордовых организмов, в который вовлечены хорошо изученные регуляторные факторы (Nanog, MZT и др.). Один из ключевых этапов этого процесс – зиготическая активация генома эмбриона, когда оплодотворенное яйцо начинает работу собственных ядерных генов. В работах ранее было показано радикальное изменение структуры хроматина в ходе этого процесса у *D. melanogaster* и *D. rerio* [7, 8, 9], однако недостаточно внимания уделено факторам и механизмам формирования структуры хроматина в ходе этого процесса.

В данном проекте предложено разобраться с деталями этого процесса в свете новых полученных данных на разных стадиях эмбриогенеза этих организмов. На данный момент в работе проведен первичный анализ опубликованных и новых данных, проведен контроль качества и ассоциация особенностей данных (ТАДы, компартменты) с факторами хроматина (мотивы связывания регуляторов эмбриогенеза).

Работа выполняется в коллаборации с Сергеем Ульяновым (ИБГ), Daria Onichtchouk (University of Freiburg Faculty of Biology) и Leonid Mirny (MIT).

(3) ДНК бактериальных организмов демонстрирует большое разнообразие принципов укладки по данным фиксации конформации хромосом. В данной работе предложено разобраться с пространственной укладкой ДНК *E. coli* [10]. На данный момент проведен первичный анализ опубликованных и новых данных фиксации конформации хромосом этого организма, впервые обнаружены компартмент-подобные стурктуры и обнаружена ассоциация этого феномена с активными генами.

Работа выполняется в коллаборации с Дмитрием Суторминым (Сколтех).

Работа проводится при поддержке Skoltech Fellowship in System Biology. Автор выражает благодарность Юрию Коростелеву за поддержку лабораторного браузера HiGlass и техническое обеспечение кластера meatgrinder.

Литература:

1. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168]
2. https://github.com/mirnylab/pairsamtools
3. https://github.com/mirnylab/cooltools
4. http://autosome.ru/
5. http://mg.uncb.iitp.ru:8889/
6. Heger et al. “The chromatin insulator CTCF and the emergence of metazoan diversity”, 2012, PNAS
7. Schulz & Harrison “Mechanisms regulating zygotic genome activation”, 2018, Nature reviews
8. Kaaij et al. “Systemic Loss and Gain of Chromatin Architecture throughout Zebrafish Development”, 2018, Cell reports
9. Hug et al. “Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription”, 2017, Cell
10. Lioy et al. “Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins”, 2018, Cell

## Структура хроматина единичных клеток *Drosophila melanogaster*

***Александра Галицына, Сколтех***

Анализ данных Hi-C единичных клеток *Drosophila melanogaster* представляет собой интересную и важную проблему в фундаментальном исследовании структуры хроматина, описанную нами неоднократно, напр. [1].

В текущей работе обобщены накопленные знания и факты о методе Hi-C единичных клеток по протоколу single-cell Hi-C без обогащения контактами [2] и структуре хроматина единичных клеток *D. melanogaster*. Проведена связь с эпигенетическими подписями ДНК, приведены доводы, подтверждающие непротиворечивость наблюдений с существующими гипотезами о формировании трехмерной структуры ДНК.

Литература:

1. Zakharova V.S. et al., «Single-cell Hi-C demonstrates that TADs are stable units of Drosophila genome folding that persist in individual cells», 2018, 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
2. Flyamer I. M. et al., “Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition.”, 2017, Nature

## Анализ изменения трехмерной структуры хроматина на разных стадиях сперматогенеза Drosophila melanogaster.

***Кононкова Анна Дмитриевна***

На фоне большого количества исследований, посвященных сравнительному анализу трехмерной структуры хроматина в разных видах или тканях одного и того же организма, изменение конформации хроматина на разных стадиях развития изучена в меньшей степени [1,2,8].

Цель данной работы – изучение трехмерной структуры хроматина для двух стадий развития семенников Drosophila melanogaster (vas - сперматоциты, bam - сперматогонии). Исследование проводится для двух разрешений – 5 и 10 кб. Анализ конформации хроматина для разных стадий у данного организма на таком разрешении проводится впервые. Экспрессия генов значительно меняется в ходе сперматогенеза [6]. Известно, что между экспрессией генов и пространственной структурой хроматина имеется взаимосвязь [7]. Анализ последней является наиболее важной составляющей данной работы.

Исследование проводилось в следующих группах генов: дифференциально экспрессирующиеся на стадии vas (DV), на стадии bam (DB), не различающиеся между стадиями (ND) и сперматоцит-специфичные гены (SSG)

Результаты можно условно разделить на 3 группы:

1) Феноменологические

2) Отрицательные

- не было обнаружено корреляции частот контактов, IS, плотности ТАДов с экспрессией и сигналом DamID
- профиль инсуляции вокруг SSG не имеет существенных различий по форме в vas и bam и не отличается от других групп генов
- кластер SSG может располагаться как в одном ТАДе, так и в разных. Подобное отсутствие закономерности ранее показано для другого организма [5].

-прочее

3) Положительные

-частоты контактов выше в bam по сравнению с vas в активном компартменте bam и ниже в неактивных областях bam

- для участков, перешедших из неактивного компартмента в bam в активный в vas, доля ССГ генов в 3 раза больше, а DB-генов в 2 раза меньше по сравнению с фоном (фон – пропорция генов в целом, без учета каких-либо компартментов). Для участков, перешедших из активного компартмента в bam в неактивный в vas, нет перепредставленности определенной группы генов по сравнению с фоном.
- наблюдается небольшая положительная корреляция между изменением значения первой главной компоненты и изменением экспрессии (коэф. кор. Спирмена 0.23). Общая тенденция для медиан соответствует наблюдению в предыдущем пункте (рисунок в приложении).

Текущие выводы.

Анализ литературы ( безотносительно определенного вида ) показывает отсутствие ярко выраженной корреляции между изменением экспрессии и изменением конформации хроматина для разных стадий развития [1,2] либо при исследовании ортологичных групп генов у разных организмов [4]. В статье [2] описывается регуляции экспрессии глобинового локуса посредством изменения трехмерной структуры в двух стадиях развития эритробластов, в данном случае речь идет об одном (!) ТАДе. Также показано, что принадлежность к одному классу ко-экспрессии плохо предсказывается частотой контактов (напрямую, без усложнения модели) [3,5]. Результаты, полученные в данной работе, в большей степени говорят о

существовании взаимосвязи между изменением экспрессии и изменением конформации хроматина (возможно из-за более высокого разрешения), причем различия проявляются при исследовании компартментализации. В статье [4] высказано предположение о том, что изменение конформации в большей степени объясняет изменение в экспрессии, чем наоборот. Полученные результаты согласуются с этим предположением, но оно требует дальнейшей проверки.

Планы.

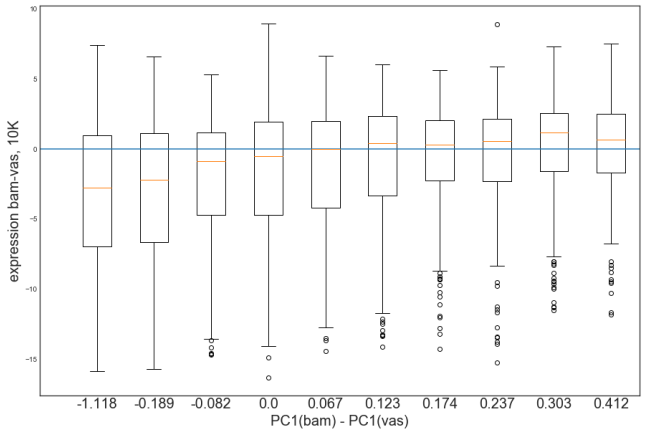
-исследовать дальние взаимодействия в vas и bam

-посчитать корреляцию профилей экспрессии для генов внутри кластеров для двух стадий

-получить 2 реплики vas и провести дифференциальный анализ.

-применить линейную модель и проверить вклад частот контактов в изменение экспрессии

Приложение.




Литература.


1. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., … Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. https://doi.org/10.1038/nature14222
2. Huang, P., Keller, C. A., Giardine, B., Grevet, J. D., Davies, J. O. J., Hughes, J. R., … Blobel, G. A. (2017). Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. *Genes and Development*, *31*(16), 1704–1713. https://doi.org/10.1101/gad.303461.117
3. Babaei S, Mahfouz A, Hulsman M, Lelieveldt BP, de Ridder J, Reinders M. Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex. *PLoS Comput Biol*. 2015;11(5):e1004221. Published 2015 May 12. doi:10.1371/journal.pcbi.1004221
4. Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E., & Gilad, Y. (2018). Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates. *BioRxiv*, 474841. https://doi.org/10.1101/474841
5. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V., & Reyes, J. C. (2017). Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Computational Biology*, *13*(9), 1–25. https://doi.org/10.1371/journal.pcbi.1005708

6. Laktionov, P. P., Maksimov, D. A., Romanov, S. E., Antoshina, P. A., Posukh, O. V, White-Cooper, H., … Belyakin, S. N. (2018). Genome-wide analysis of gene regulation mechanisms during Drosophila spermatogenesis. *Epigenetics & Chromatin*, *11*(1), 14. https://doi.org/10.1186/s13072-018-0183-3
7. Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., … Razin, S. V. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*, *26*(1), 70–84. https://doi.org/10.1101/gr.196006.115
8. Bonev, B., Cohen, N. M., & Cavalli, G. (n.d.). Multiscale 3D Genome Rewiring during Mouse Neural Development Graphical Abstract Introduction Global Reorganization of the 3D Genome during Neural Differentiation.

## Изучение динамики ТАДов на примере сперматогенеза *Drosophila melanogaster*

***Жигулев А.А., Галицына А.А., Кононкова А.Д., Храмеева Е.Е., Гельфанд М.С.***

Организация хроматина играет важную роль в регуляции экспрессии генов. Новейшие молекулярно-биологические методы, такие как Hi-C, позволили расширить наше представление об устройстве хроматина. Таким образом, на смену классическому представлению об уровнях организации хроматина пришли новые понятия, в том числе топологически ассоциированные домены (ТАДы). Параллельно с совершенствованием методов, важно найти биологическое обоснование получаемых результатов, в частности динамики образования ТАДов. Одним из наиболее сложных процессов дифференциации клеток является сперматогенез. В процессе сперматогенеза хроматин половых клеток претерпевает серьезные динамические изменения, помогающие осуществлять сложную программу регуляции экспрессии генов в развитии. Поэтому именно в этом процессе мы ожидаем увидеть яркие изменения биологических механизмов, которые могли бы лежать в основе динамики образования ТАДов. В число этих механизмов входит дифференциальная экспрессия генов, а также эпигенетические изменения.

ТАДы были аннотированы для двух стадий сперматогенеза Drosophila melanogaster: Bam (сперматоциты) и Vas (сперматоциты и сперматогонии), программой Armatus на основании экспериментальных данных Hi-C, полученных в лаборатории С.В. Разина в ИБГ РАН. В данный момент ведется работа по кластеризации предсказанных ТАДов на основе динамики их изменений на 4 основные группы: консервативные, полуконсервативные ТАДы, новые ТАДы, появившиеся на стадии Vas и исчезнувшие после стадии Bam. Также, изучается распределение размеров ТАДов в этих группах, и количество семенник-специфичных генов в ТАДах и их окружении. Отдельной задачей будет изучение появившихся и исчезнувших ТАДов, которые, вероятно, выполняют крайне важную функцию в процессе дифференциации клеток.

## Изменения трехмерной организации хроматина в развитии эмбриона *Drosophila melanogaster*

***Николай Быков***, *НИУ ВШЭ – Анализ данных в биологии и медицине*
*Александра Галицына, Сколтех*

a) В последние несколько лет создание и развитие новых методов молекулярной биологии для анализа энхансер-промоторных взаимодействий, таких как FISH и 3C (в том числе Hi-C и его всевозможные модификации), позволило ученым глубже изучать проблему укладки хроматина, его роль в эпигенетике, изменение конформации с развитием [1]. Одна из важнейших изучаемых сейчас проблем, связанных с хроматином – взаимное влияние эпигенетической регуляции биологических процессов и динамики трехмерной организации хроматина при развитии организма [2]. Наш проект будет посвящен исследованию этой проблемы для *Drosophila melanogaster*.

b) Проблема динамики трехмерной организации хроматина с развитием организма интересна научному сообществу во многом благодаря ее связи с эпигенетикой. От одной стадии развития организма к другой меняется не только пространственная структура хроматина (включая организацию ТАДов), но и ее соответствие различным эпигенетическим маркерам. Так, умение предсказывать паттерны временных изменений ТАДов по содержанию эпигенетических меток поможет исследователям, которые занимаются вопросами эмбрионального развития отдельных организмов (например стволовых клеток, для получения которых выращивают специальные колонии эмбрионов [3]), соотносить определенные паттерны в изменении конформации ТАДов с конкретной эпигенетикой этих организмов.

c) В рамках проекта планируется:

i. Осуществить поиск ТАДов в данных Hi-C по развитию эмбриона *Drosophila melanogaster* из статьи [2] с помощью программы lavaburst [4] (варьирование параметров поиска ТАДов, контроль устойчивости поиска).

ii. Реализовать кластеризацию найденных ТАДов (варьирование алгоритмов и параметров кластеризации по различным критериям, контроль качества и устойчивости кластеризации).

iii. Разработать консольную утилиту для кластеризации ТАДов по D-score в ряде экспериментов.

iv. Проверка биологической гипотезы о различии ТАДов по времени созревания.

v. В случае неуспеха в проверке биологической гипотезы – переключиться на другой датасет (мышь с индукцией плюрипотентности из статьи [3]) и выполнить заново пункты i-iii.

vi. Предсказание изменения организации ТАДов во времени по содержанию эпигенетических меток с помощью методов машинного обучения.

В данный момент ведется работа над пунктами i-ii. Проведено первичное исследование данных и их визуализация. В ближайших планах – провести все необходимые в рамках i-ii эксперименты, получив устойчивые результаты, и начать разработку консольной утилиты, описанной в пункте iii.

**Список литературы**

1. Lieberman-Aiden E et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. 2009, Science.

2. Clemens B.Hug et al., Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription, 2017, Cell.
3. Boyan Bonev et al., Multiscale 3D Genome Rewiring during Mouse Neural Development, 2017, Cell.
4. Nezar Abdennur et al., Lavaburst, 2014, https://github.com/nvictus/lavaburst.

## Автоматическая разметка петель на контактных картах амебы Dictyostelium discoideum

***Е.Храмеева, О.Цой, А.Галицына, С.Ульянов, А.Савостьянов***

Корректная аннотация трехмерной структуры укладки хроматина является одним из качественно важных вопросов клеточной биологии. В частности, известно, что топологическая близость функциональных участков ДНК-молекулы в трехмерном пространстве оказывает существенное влияние на регуляцию генов. Для количественного описания трехмерной структуры участков (бинов) ДНК используются данные эксперимента Hi-C в виде контактных карт. Одной из широко известных структур хроматина являются топологически ассоциированные домены, состоящие из сильно контактирующих участков на главной диагонали карты. В то же время, для общественной амебы Dictyostelium discoideum на картах заметны скопления сильно контактирующих пар бинов, отделимые от главной диагонали – петель. Таким образом, возникает естественная задача автоматической генерации трека петель по картам контактов.

В рамках работы был реализован метод автоматической разметки петель по контактной карте Hi-C в формате cooler. Полагая корректным эвристическое предположение об отсутствии петель в существенном отдалении от главной диагонали, метод модифицирует классический алгоритм компьютерного зрения Laplacian of Gaussian (представляющий из себя свертку контактной матрицы с двумерным гауссовским ядром с перебираемой дисперсией для поиска локальных скоплений отделимо более контактирующих пар бинов – петель) в скользящей рамке относительно главной диагонали. Для уменьшения количества найденных несодержательных особенностей карты (вызванных, например, зашумленностью входных данных) метод искусственно поднимает разрешение карты при помощи сглаживаемой интерполяции и дополнительно фильтрует недостаточно выделяющиеся на фоне всей рамки и локальной окрестности петли. Полученное множество петель, избыточное ввиду многократного вхождения одной петли в скользящую рамку, подвергается процедуре объединения и удаления петель, включающих главную диагональ, как соответствующих фрагменту топологически ассоциированного домена.

В результате работы метода был получен трек петель для контактной карты амебы Dictyostelium discoideum для всех 6 хромосом; в итоговой автоматической разметке обнаружено 78% петель из имеющейся ручной разметки, в то время как 50% автоматической разметки не соответствует ни одной из имеющихся петель. При анализе результатов было отмечено скопление новых петель в окрестностях областей, содержащих экспериментальные артефакты контактных карт, которые приводят к малой достоверности как и самих карт, так и результатов разметки в этих областях. Исключение подобных участков сокращает долю новых петель в разметке до 25-40% в зависимости от размера исключенных участков. Для оценки false discovery rate распределение найденных петель по диагоналям было решено приближать

гамма-распределением, на основе чего доля невязки в гамма-распределении использовалась для оценки FDR. По автоматической разметке коллегами был построен профиль эпигенетических маркеров, в достаточной степени совпадающий с профилем относительно ручной разметки.

Дальнейшая работа будет включать изучение влияние фазы эксперимента, для каждой из которых получены результаты алгоритма, на разметку петель (как минимум известно, что фаза заметно сказывается на количестве обнаруженных петель); также необходимо провести тест на независимость результатов от порядка обхода скользящим окном.

## **Предсказание петель в хроматине** *Dictyostelium discoideum*

***Плискин Александр, НИУ ВШЭ, Анализ данных в биологии и медицине***
***Галицына Александра, Сколтех***

*Dictyostelium discoideum* – клеточный слизевик, один из важных модельных организмов в биологии и генетике. Интересен тем, что большую часть времени проводит в виде одиночной амёбы, однако при определенных условиях объединяется с другими диктиостелиумами, образуя подвижные агрегаты для совместного существования в неблагоприятных условиях. Hi-C метод [1], который исследует трехмерную архитектуру целых геномов, соединяя лигирование на основе близости с массивно-параллельным секвенированием, позволяет получить одноименные карты, которые визуализируют участки генома оказывающиеся рядом, даже если они расположены далеко друг от друга на одной хромосоме или даже на разных хромосомах. По контактам на Hi-C картах можно определить петли в геноме человека [2].

Имеются Hi-C карты хроматина диктиостелиума, полученные в лаборатории С.В. Разина (ИБГ РАН) на которых вручную размечены координаты, размеры и типы петель (на хромосомах 1 и 6). На данный момент не существует программы, которая могла быть предсказать петли на Hi-C карте диктиостелиума. Задачей данной работы является разработка такой программы.

Планируется свести задачу нахождения петель на Hi-C карте к задаче распознавания образов на изображении. В качестве образов будут выступать участки с петлями, в качестве изображения - Hi-C карта. В качестве образцов будут выделены образы петель на карте, по которым имеются вручную размеченные координаты. По эти образцам планируется применить глубинное обучение с использованием свёрточных нейронных сетей R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN [3]. Полученную нейронную сеть планируется применить для нахождения петель на других хромосомах и других картах диктиостелиума. В случае успеха подобную программу можно будет применять на Hi-C картах и других организмов.

Список литературы:

1. Lieberman-Aiden E, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, 2009, Science.
2. Rao S. S. P, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, 2014, Cell.
3. https://github.com/tensorflow/models/tree/master/research/object_detection

# MANY-BODY CONTACTS IN FRACTAL POLYMER CHAINS

*Polovnikov K.1,2, Nechaev S. 3,4, Tamm M.V.1,5*

***1 - Physics Department, Moscow State University, 119992, Moscow***

***2 - Skolkovo Institute of Science and Technology, 143005, Skolkovo***

***3 - Interdisciplinary Scientific Center Poncelet (ISCP), 119002, Moscow***

***4 - Lebedev Physical Institute RAS, 119991, Moscow***

***5 - Department of Applied Mathematics, National Research University Higher School of Economics, 101000, Moscow***

*kipolovnikov@gmail.com*

Modern experimental techniques of genome-wide chromosome conformation capture (Hi-C) allow to describe a conformation of the chromatin by tabulating a full collection of *in vivo* pairwise contacts of the chromatin fiber with itself. The Hi-C data has proven to be hugely beneficial for our understanding of chromosome conformations under different conditions. There exist a broad range of theories used to rationalize the existing Hi-C data, however all these models seem to agree that in a wide range of length scales the resulting equilibrium chromatin packing is approximately fractal with transient fractal dimension $ddff$ lying in the interval $2 \leq ddff \leq 3$, with $ddff=3$, which corresponds to a space-filling curve in the three-dimensional space, being the most obvious candidate for the true fractal dimension in the limit of infinitely long chains.

Additional evidence concerning chromatin conformations should come from consideration of triple and many-body contacts, which, generally speaking, contain information on the properties of the fiber packing, which is not reducible to the information obtained from two-loci contacts. Recent advances in experimental techniques [1, 2] allow to expect that soon it will be possible to capture triple contacts in Hi-C experiments. If clusters of several (more than 2) chromatin loci become detectable in each individual cell, then, after averaging over ensemble of the cells, one can obtain the statistics of triple contacts. Therefore, theoretical approaches allowing to make sense of this upcoming data are urgently needed.

Here we calculate the three-body and many-body contact probabilities in the framework of a Gaussian polymer chain [3], which can be used as a basic benchmark to compare the experiment data with. In particular, we propose here to measure experimentally the two-loop correlation factor (the ratio of a three-body contact probability to the product of the probabilities of two independent loops of the same size) as an important characteristic of the chromatin conformation statistics, and provide concrete predictions for the value of this factor within our Gaussian approach.

[1] P. Olivares-Chauvet, Z. Mukamel, A. Lifshitz, O. Schwartzmann, N.O. Elkayam, et al., Capturing pairwise and multi-way chromosomal conformations using chromosomal walks, Nature, 540, 296, (2016);

[2] A.M. Oudelaar, J.O.J. Davies, L.L.P. Hanssen, J.M. Telenius, R. Schwessinger, et al., Single-cell chromatin interactions reveal regulatory hubs in dynamic compartmentalized domains, Nature Genetics, 50, 1744 (2018);

[3] K. Polovnikov, S. Nechaev, and M. V. Tamm, Effective Hamiltonian of topologically stabilized polymer states, Soft Matter, 14, 31 (2018).

## Contact Probability in Loop Extrusion Model of Interphase Chromosome

***Sergey Belan***
*(in collaboration with Mirny Lab, MIT)*

Due to the development of the chromosome conformation capture (Hi-C) method, it has become possible to get insight into the chromatin organization by measuring the frequency of physical contacts between different parts of genome. The mechanism of active loop extrusion holds great promise for explaining the key features of the contact maps obtained from the Hi-C data. The loop extrusion model assumes that ATP-dependent process allows nanometer-size molecular machines to organize chromosomes by producing dynamically expanding chromatin loops. In this talk I will give a brief introduction into the loop extrusion model and demonstrate that analytical predictions extracted from this model in its simplest version, where chromatin fiber is treated as an ideal Gaussian chain, are in agreement with experimentally measured statistics of contacts in the interphase chromosomes.

## Поиск неканонических структур ДНК как нуклеосомных барьеров методами машинного обучения

***Элен Теванян,*** *аспирантка ФКН*

Мы обучили алгоритм случайного леса для распознавания паттернов взаимного расположения нуклеосом и вторичных структур ДНК, которые могут служить барьерами для нуклеосом, в геноме мыши. Мы показали, что среди четырех типов рассмотренных структур (Z-ДНК, H-ДНК, G-квадруплексов и участков SIDD) наилучшая предсказательная сила моделей достигается для G-квадруплексов и H-ДНК. Настоящей задачей является построение модели, основанной на физико-химических и структурных свойствах последовательности ДНК, а также тестирование моделей на другом типе ткани (в частности, стволовых клетках) и определение как общих, так и тканеспецифичных паттернов.

## Модели машинного обучения для распознавания положений нуклеосом на основе физико-химических и структурных характеристик ДНК.

***Даша Афентьева***
бакалавр кафедры биоифизики, физический ф-т, МГУ

Факторы, влияющие на расположение нуклеосом в разных типах ткани до конца не понятны. Целью настоящей работы является построение моделей машинного обучения, предсказывающих расположение нуклеосом в разных типах тканей. Предполагается построение моделей на основе физико-химических и структурных характеристик ДНК, таких как структурные параметры динуклеотидов Rise, Slide, Shift, Twist, Roll, Tilt и физические параметры энтальпия, энтропия, свободная энергия Гиббса, электрический потенциал,

гидрофильность. Планируется исследование моделей отдельно как на хорошо позиционированных, так и на динамически перемещаемых нуклеосомах.

### Построение обобщающего вероятностного профиля расположения нуклеосом методами машинного обучения

***Виктория Верезубова**, ФКН ВШЭ, магистр 2 года обучения, vika.verezubova@gmail.com*

Были реализованы модели машинного обучения для предсказания нуклеосомной разметки генома на основе физико-химических и геометрических свойств динуклеотидов, а также на основе статистики k-меров в последовательностях. В качестве обучающей выборки были использованы экспериментальные данные MNase-Seq генома человека, содержащие информацию о расположении нуклеосом в однонуклеотидном разрешении. Настоящая задача состоит в построении, с помощью обученных моделей, обобщающего вероятностного профиля расположения нуклеосом, и валидация данного профиля на нуклеосомных картах разных типов тканей.

# Процессинг пре-мРНК

### Тандемные альтернативные сайты сплайсинга: эволюционные свидетельства функциональности

***Степан Денисов**, Алексей Миронов, Дмитрий Первушин*

Мы определили неаннотированные сайты сплайсинга как сайты сплайсинга (СС), подтвержденные данными высокопроизводительного секвенирования РНК (RNAseq) при этом не представленные в базе данных GENCODE. Тандемные альтернативные сайты сплайсинга (TASS) – это пары из аннотированных и неаннотированных сайтов сплайсинга, находящиеся недалеко друг от друга (<30 нт). При этом аннотированный СС используется в большинстве случаев (в разных тканях и индивидуумах). Возникает вопрос: неаннотированные TASS представляют собой функциональные элементы или же они являются следствием случайных ошибок сплайсосомы (шум сплайсинга)? Мы попытались подойти к этому вопросу с эволюционной точки зрения: если это ошибки сплайсосомы, то мы не ожидаем, что они консервативны, напротив, если они функциональны, то они должны быть консервативны.

Мы применили разработанный ранее тест на положительный и отрицательный отбор в СС и получили следующие результаты. На неаннотированные СС действует отбор схожий по порядку величины с таковым, действующим на аннотированные сайты. Однако есть интересные отличия.

Так на консенсусные нуклеотиды в неаннотированных акцепторных СС действует даже более сильный отрицательный отбор, чем в аннотированных (для донорных СС отличия статически не значимы). Из предыдущих исследований известно, что сила отбора,

действующего на нуклеотиды в той или иной позиции зависит от общей силы СС (в силу эпистаза между позициями), в частности на слабые СС действует более сильный отрицательный отбор, стремящийся сохранить консенсусные нуклеотиды. В среднем аннотированные СС сильнее неаннотированных, поэтому возникла гипотеза, что на разницу в силе отбора существенно влияют различия в распределениях сил сайтов. Мы проверили эту гипотезу, выбрав только СС с очень похожими силами из аннотированных и неаннотированных и сравнили действие отбора на них. Оказалось, что на неаннотированные донорные СС действует более слабый отрицательный отбор на консенсусные нуклеотиды, чем на аннотированные с похожей силой сайта. Это согласуется с предположением, что среди неаннотированных донорных СС больше шумовых событий сплайсинга. Однако, у акцепторных СС (уравненных по силе) отбор на консенсусные нуклеотиды практически не отличается между группами.

Интересно, что на неаннотированные донорные СС действует положительный отбор, способствующий заменам неконсенсусных нуклеотидов на консенсусные, сила которого даже больше, чем в аннотированных СС. Этот эффект сохраняется после выравнивания по силе сайта (см. выше). Кроме того, если рассмотреть только неаннотированные донорные СС и разделить их по силе сайта, то оказывается, что на слабые (обедненные консенсусными буквами) действует более сильный положительный отбор на переходы из неконсенсуса в консенсус. Предположительно это связано с тем, что неаннотированные СС могут быть обогащены недавно возникшими в эволюции СС, которые ещё не достигли оптимума приспособленности.

Также мы проверили, как такие характеристики, как частота использования неаннотированных СС по сравнению с аннотированными, сохранение рамки считывания влияют на силу отбора. Как и ожидалось, на более часто используемые (в среднем по всем тканям и индивидуумам) неаннотированные СС действует более строгий отбор. Видно, что те неаннотированные донорные СС, которые используются в большом количестве тканей и индивидуумов отрицательный отбор в большей степени сохраняет консенсусные нуклеотиды. Также отбор на неаннотированный СС зависит от того, находится ли СС в рамках кодирующей последовательности или вне её (что согласуется с предыдущими результатами на аннотированных СС).

## Тандемные альтернативные сайты сплайсинга в раке и норме

***Миронов Алексей, Степан Денисов, Дмитрий Первушин***

Хорошо известно, что более 90% генов человека альтернативно сплайсируются, но функциональное значение альтернативных транскриптов все еще остается предметом дискуссий. С одной стороны, многие альтернативные изоформы ассоциированы с болезнями, что указывает на их функциональность. С другой стороны, протеомные исследования выявили лишь небольшую долю изоформ мРНК, которые транслируются в белки.

Среди массивного репертуара неаннотированных изоформ РНК, обнаруженных в данных NGS, есть специфические события сплайсинга, которые характеризуются близким (~30 нт) тандемным расположением альтернативных сайтов сплайсинга (TASS – кластер таких альтернативных сайтов). Мы исследуем сплайсирование альтернативных сайтов в TASS,

используя данные большой панели экспериментов РНК-сек из проектов GTEx (норма) и TCGA (рак).

Исследовав нормальные ткани, мы оценили последствия использования альтернативных сайтов в TASS на уровне белка, используя структурную аннотацию человеческого протеома (рис. 1). Среди 44 тысяч TASS, содержащих хотя бы один аннотированный сайт, лишь в 12.5 тысячах TASS использование альтернативных сайтов не приводит к смещению рамки считывания. Среди последних только 10% (~2.5 тысячи) альтернативно сплайсируются, т.е. наблюдается значимое использование альтернативных сайтов. Мы проанализировали представленность таких TASS (зеленые столбцы) в категориях структурной аннотации протеома по сравнению с остальными TASS, где доминирует экспрессия одного сайта (синие столбцы). В качестве контроля мы использовали сайты сплайсинга, не входящие ни в один TASS (красные столбцы).

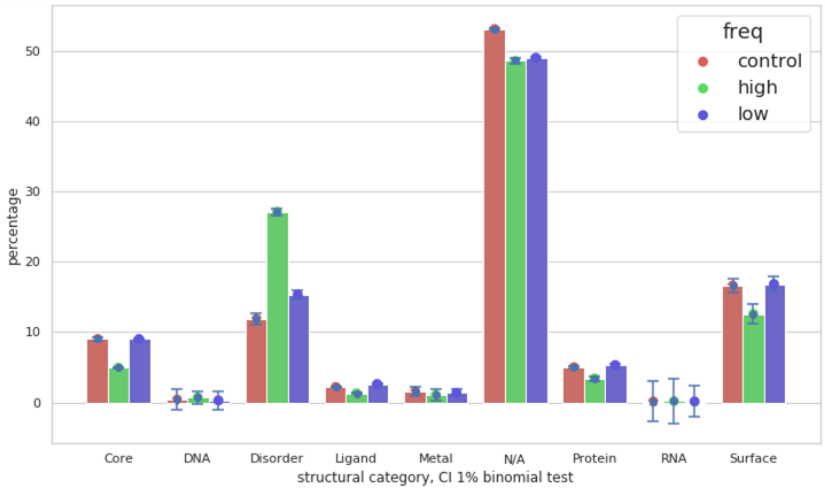



*Рисунок 1 Анализ пере- и недопредставленности альтернативно сплайсирующихся TASS в категориях белковых структур*

Альтернавтивно сплайсирующиеся TASS перепредставлены в категории disorder (часть белка, не свернутая во вторичную или третичную структуру) и недопредставлены в функциональных категориях белка (core, protein, surface). Это указывает на то, что альтернативный сплайсинг в TASS в большинстве случаев является следствием ошибок сплайсосомы, является слаботоксичным и в норме подавляется в функционально важных сегментах гена.

В раковых клетках же использование альтернативных сайтов в TASS может быть важным механизмом регуляции функции генов. По предварительному анализу данных РНК-сек гепатоцеллюлярной карциномы (TCG-LIHC) доля транскриптов, сплайсированных с использованием минорных сайтов из TASS, больше в среднем в 1.3 раза по сравнению со здоровой тканью печени больных. Мы обнаружили более 50 значительно ракоспецифически сплайсируемых TASS, среди которых, в частности, присутствуют гены семейства BCL2, участвующие в ингибировании апоптоза.

Подводя итог, проект направлен на выяснение доли и роли шума в альтернативном сплайсинге и выявление новых механизмов регуляции генов в болезнях, использующих аберрантный сплайсинг, в частности, сплайсирование альтернативных минорных сайтов в TASS.

## Влияние соматических мутаций, разрушающих структуру РНК, на альтернативный сплайсинг

***Калмыкова С.Д., Калинина М.А., Скворцов Д.А., Первушин Д.Д.***

Известно, что на сплайсинг пре-мРНК могут влиять мутации в цис-регуляторных элементах, таких как сайты сплайсинга, однако глубокие интронные мутации также оказывают влияние на сплайсинг. Один из возможных механизмов - нарушение структуры пре-мРНК, включая в частности внутримолекулярные спаривания оснований на больших расстояниях. Целью данной работы является предсказание таких внутримолекулярных спариваний и описание статистических взаимосвязей между структурой пре-мРНК и соматическими мутациями, ассоциированными с изменениями сплайсинга. Нами был разработан алгоритм динамического программирования, использующий k-меры нуклеотидов и вычисленную заранее матрицу их стекинг-энергий, который позволяет быстро находить потенциальные структуры РНК в консервативных интронных участках в масштабах всего генома. С помощью него было найдено около 930 тысяч возможных структур РНК в геноме человека. Была исследована взаимосвязь
между потенциальными структурам РНК и соматическими мутациями в гепатоцеллюлярной карциноме и раке почки. Было обнаружено всего около 12000 однонуклеотидных замен, разрушающих РНК-структуры, для рака печени и 3000 для рака почки. Было показано, что структурированные участки обеднены мутациями. Кроме того, однонуклеотидные замены, которые наблюдаются при раке, предположительно дестабилизируют структуры сильнее, чем случайные. Также был получен список из 200 дифференциально экспрессирующихся между опухолевой и здоровой тканью экзонов, из которых 9 принадлежат известным онкогенам. Мы получили косвенное подтверждение гипотезы о том, что внесение разрушающей мутации в структуру РНК, выпетливающую экзон, может вести к увеличению экспрессии этого экзона.

## Роль вторичной структуры РНК в экспрессии химерных неколлинеарных транскриптов

***Ольга Васюткина, Дмитрий Первушин***

Транс-сплайсинг – это редкий и малоизученный пост-транскрипционный процесс, при котором экзоны двух молекул пре-мРНК соединяются и образуют одну зрелую мРНК. В частности, транс- сплайсинг может происходить между транскриптами в одном локусе, что может привести к возникновению транскриптов с порядком экзонов, не совпадающим с их порядком в геноме. Такие транскрипты описаны в литературе под названием post-transcriptional exon shuffling (PTES) и rearrangements or repetition in exon order (RREO). Мы предполагаем, что события транс-сплайсинга с большей вероятностью могут происходить в областях генома, где РНК-полимераза II движется с относительно низкой скоростью. В таком случае различные копии одного и того же транскрипта сближены в пространстве, что увеличивает вероятность транс-сплайсинга
между ними. Цель нашей работы - оценить вклад событий транс-сплайсинга в образование транскриптов с PTES или RREO, и отличить такие события от образования кольцевых РНК,

что возможно при использовании парно-концевых чтений РНК-секвенирования. Далее мы планируем проверить, есть ли накопление транскриптов с PTES или RREO в областях
с сильным сигналом данных ChIP-Seq для РНК-полимеразы II или областях, где вероятно образование вторичных структур РНК. Мы обнаружили существенное количество транскриптов с PTES или RREO, которые нельзя объяснить образованием кольцевых РНК, в данных РНК-секвенирования ENCODE клеточных линий человека K562 и HepG2. Есть данные о том, что
экзоны в РНК способны формировать кольцевые структуры, если в окружающих их интронах есть взаимно комплементарные участки. Мы обнаружили, что найденные нами кольцевые РНК предпочтительно находятся внутри областей с возможностью образования вторичной структуры, а для транскриптов, возможно образованных в результате транс-сплайсинга,
такое не наблюдается.

## Intronic polyadenylation and splicing in cancer

***Valeriya Mikova, Kim Adameyko, Dmitri Pervouchine***

Recent studies have shown that intronic polyadenylation (IPA), i.e. premature termination of transcription, is frequently observed in diverse cancer types and can mimic the functional outcome of genetic alterations that lead to truncated proteins. In particular, these products may lack tumor suppressor functions, which they otherwise would have had in the case of translating full-length transcripts. The goal of this project is to use public cancer data sources (The Cancer Genome Atlas, The Pan-Cancer Analysis of Whole Genomes) to identify IPA events that are associated with cancers. Current approaches use the combination of 3-'seq and poly(A)-seq data to identify IPA. The aim of this project is to identify tumor-associated IPA events using RNA-seq data alone, namely by extracting short reads that contain poly(A)-stretches that partially align to the genome.

## Possible role of upstream open reading frames in NMD-dependent degradation of human mRNAs

***Lev Mazaev, Dmitri Pervouchine, Stepan Denisov***

Transcripts with premature stop-codons are the target of a post-transcriptional regulatory mechanism of gene expression called nonsense-mediated decay (NMD) which is present in all eukaryotes. Some of such transcripts contain uORF – an open reading frame with both its start and stop codons in 5’-UTR. uORF normally decrease probability of translation initiation of the main ORF leading to lower level of the protein product.

We hypothesize that when NMD surveillance pathway is active the coding sequence of uORF-containing mRNA is normally unable to be translated to the functional protein due to fast degradation of the transcript. But actually, it remains a possibility of translation reinitiation on the main ORF and leaky scanning of uORF, so eventually the gene may be expressed. NMD of the transcript may depend on translation of the main ORF, modulated by presence and of an uORF and strength of a reinitiation context. To determine the role of NMD-dependent mRNA degradation and the magnitude

of possible interdependency between reinitiation/leaky scanning and NMD we will compare the expression of uORF and non-uORF genes using RNA-sequencing data obtained from human cell lines with depletion of UPF1 and SMG6 (main factors involved in NMD) and control cell lines, where NMD stays active.

# Вирусы

## Филогеография вируса гепатита А

***Алевтина Корешова, Алексей Неверов, Георгий Базыкин***

Вирус гепатита А (ВГА) принадлежит семейству Picornaviridae и является единственным представителем рода Hepatovirus. Геном ВГА состоит из (+) РНК цепи длиной в 7.5 кб [1]. Вирус передается фекально-оральным путем и вызывает острый гепатит у взрослых [2]. У детей до 6 лет инфекция обычно протекает бессимптомно [3,4]. ВГА преобладает в странах с низким уровнем гигиены, но не приводит к эпидемиям, потому что большинство людей инфицируются в раннем детстве и приобретают пожизненный иммунитет. Таким образом, ВГА представляет опасность только для развитых стран, где вызывает вспышки заболевания среди взрослого населения, как, например, в России.

В отличие от других пикорнавирусов, ВГА имеет очень низкую скорость мутации (примерно 1.99 x ) [5]. Генотипирование ВГА проводится на коротких вариабельных фрагментах генома VP1/2A и 2C/3A. ВГА включает в себя 6 генотипов, из которых I и III чаще всего встречаются у людей. Оба делятся на субгенотипы A и B, и IA и IIIA наиболее распространены в России, причем IA преобладает в европейской части, а IIIA – в азиатской. Мы предполагаем, что ВГА имеет географическую кластеризацию, и IA субгенотип попал в Россию из Южной Европы и распространяется на восток, в то время как IIIA зародился в Индии и является более древним на территории Сибири.

Филогеографический анализ был проведен на фрагментах 2C/3A длиной в 650 нуклеотидов, полученных из ВГА России и СНГ в 1999-2015 гг, а также фрагментах из доступных полных геномов из Genbank. Мы использовали пакет программ BEAST, основанный на байесовском подходе. BEAST позволяет добавлять к анализу дополнительные данные, например, географические координаты, и предсказывать их значения в узлах дерева, а также датировать эти узлы [6,7].

Результаты подтверждают гипотезу относительно IIIA субгенотипа, но не IA. Вероятно, для IA субгенотипа у нас не хватает географического покрытия (нет образцов из Средиземноморья и Каспия). Следующими нашими шагами будут:

1) Определение географической кластеризации и направления распространения субгенотипов ВГА по России с помощью пакета программ GenGis [8].
2) Анализ на еще одном имеющемся фрагменте генома — VP1 (400 п.н.). Это несколько

увеличит географическое покрытие, т.к. этих образцов больше, но снизит разрешение.
3) Секвенирование полных геномов из тех же образцов, откуда были получены частичные последовательности, и добавление их к анализу.
4) Добавление частичных последовательностей из Genbank для увеличения географического покрытия. Скорее всего, придется взять только последовательности, которые покажут существенное пересечение с нашими фрагментами.

1. Knowles, et al. (2012) Family Picornaviridae. Virus taxonomy. Ninth Report of the International Committee on Taxonomy of Viruses, eds King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (Elsevier/Academic Press, Amsterdam), pp 855–880.
2. Feng, et al. (2013) A pathogenic picornavirus acquires an envelope by hijacking cellular membranes. Nature 496(7445):367–371
3. Hadler, et al. (1980) Hepatitis A in day-care centers: a community-wide assessment. N. Engl. J. Med. 302:1222–1227.
4. Lednar, et al. (1985) Frequency of illness associated with epidemic hepatitis A virus infections in adults. Am. J. Epidemiol. 122:226–233.
5. Kulkarni et al. (2009) Full length genomes of genotype IIIA Hepatitis A Virus strains (1995–2008) from India and estimates of the evolutionary rates and ages. Infection, Genetics and Evolution 9, 1287–1294
6. Suchard et al. (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4, vey016.
7. Lemey et al. (2010) Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27, 1877–1885.
8. Parks et al (2013) GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. PLoS One 29;8(7).

## Молекулярная эпидемиология ВИЧ в одном Российском регионе

***Сафина К.**, Неверов А., Базыкин Е., Киреев Д.*

В России зарегистрировано около одного миллиона лиц, живущих с ВИЧ, но динамика эпидемии ВИЧ остаётся плохо изученной, что во многом связано с ограниченным количеством доступных для анализа данных. Классические эпидемиологические данные — анамнез (предполагаемые дата, место и способ заражения) могут быть дополнены нуклеотидными последовательностями вируса, полученными из крови заражённого пациента. Современные методы филогенетического анализа позволяют использовать эти данные для изучения динамики эпидемии и оценки эффективности проводимых профилактических мероприятий, однако их применение требует высокого покрытия (не менее 30%) инфицированной популяции. Таких данных по России нет.

Мы планируем взять отдельный Российский регион, в котором количество ВИЧ-инфицированных пациентов не превышает нескольких тысяч, и отсеквенировать фрагмент генома ВИЧ не менее чем у половины этих пациентов. Мы проведём филогенетический и кластерный анализы, используя полученные последовательности и сопутствующую им

эпидемиологическую информацию, что позволит нам сделать выводы о тенденциях эпидемии ВИЧ в этом Российском регионе как в целом, так и в отдельных уязвимых группах, оценить эффективность АРВ терапии и распространение устойчивых вариантов вируса.

## Влияние белка Tat вируса иммунодефицита человека (HIV-1) на клеточные процессы в B-лимфоцитах

***Анна Валяева, Анастасия Жарикова, Алексей Пенин, Мария Логачева, Анна Клепикова, Андрей Миронов***

Вирус иммунодефицита человека (ВИЧ) инфицирует клетки иммунной системы человека, что приводит к развитию синдрома приобретенного иммунного дефицита (СПИД). Применение высокоактивной антиретровирусной терапии (ВААРТ) позволяет предотвращать развитие основных симптомов СПИД. Однако у многих пациентов на фоне ВААРТ развиваются другие ВИЧ-ассоциированные заболевания, в том числе В-клеточные лимфомы: лимфома Ходжкина, лимфома Беркитта и диффузная В-крупноклеточная лимфома. При этом, В-лимфоциты не инфицируются ВИЧ, то есть развитие этих лимфом не является следствием заражения клеток вирусом. Ключевую роль в развитии этих лимфом, по-видимому, играет вирусный белок Tat. Этот белок участвует в регуляции транскрипции генов ВИЧ и модулирует экспрессию ряда клеточных генов. Tat белок может секретироваться из ВИЧ-инфицированных клеток, накапливаться во внеклеточном матриксе и проникать в незараженные клетки.

Данная работа направлена на поиск генов В-лимфоцитов, экспрессия которых изменяется под влиянием Tat белка, и соответствующих клеточных процессов, в которых продукты данных генов участвуют. Для этого используются данные РНК-секвенирования по 4 линиям В-лимфоцитов: В-лимфоциты, экспрессирующие белок Tat-EGFP ($RPMI^{Tat}$), TatC22A-EGFP ($RPMI^{Cys}$) или EGFP($RPMI^{EGFP}$), и исходная линия В-лимфоцитов RPMI 8866 (RPMI). В ходе работы была проведена оценка качества РНК-ридов программой FastQC, картирование ридов на геном человека с помощью HISAT2. Используя программу HTSeq-count, было подсчитано количество ридов, выровненных с каждым геном.

Данные по генной экспрессии были визуализированы с помощью метода главных компонент (PCA), который указал на сходство образцов RPMI и $RPMI^{EGFP}$ между собой. Анализ дифференциальной экспрессии в данных образцах так же выявил лишь незначительные различия в профилях экспрессии генов, поэтому влиянием белка EGFP на экспрессию генов в В-лимфоцитах было решено пренебречь и сравнивать клеточные линии $RPMI^{Tat}$ и $RPMI^{Cys}$ с RPMI.

В результате анализа дифференциальной экспрессии, проведенного с помощью R пакета DESeq2, было найдено 569 дифференциально экспрессирующихся (ДЭ; padj < 0.5 и |fold change| > 2) под влиянием Tat белка генов (уровень экспрессии 206 генов повысился, а у 363 генов снизился), из них 374 - белок кодирующие гены (141 ген – экспрессия возросла, 233 – снизилась). Чтобы узнать, какие клеточные процессы оказались затронуты при изменении экспрессии найденных генов, был проведен GSEA (Gene Set Enrichment Analysis) анализ, используя аннотации GO (Gene Ontology) и KEGG. Повышенная экспрессия генов, участвующих в клеточном ответе против вирусных инфекций, процессировании РНК, пострансляционной модификации, регуляции перехода от $G_1$ к S фазе митоза, и пониженная

экспрессия генов, чьи продукты участвуют в клеточной адгезии, метаболизме, могут говорить о том, что в клетке, экспрессирующей Tat белок, запускаются как ативирусные ответы, так и провирусные реакции.

Помимо анализа белок-кодирующих генов, были проанализированы не белок кодирующие РНК – среди них нашлись длинные некодирующие РНК MALAT1 и NEAT1, уровень экспрессии которых вырос под воздействием Tat белка.

Кроме того, было проведено сравнение списков ДЭ генов, полученных в ходе нашего анализа, проводимого на В-клетках, и ДЭ генов, полученных в результате аналогичного исследования на Т-клетках. Сравнение показало лишь незначительные пересечения между множествами активированных и репрессированных под действием Tat белка генов в В- и Т-лимфоцитах. GO аннотации, полученные в ходе нашего анализа В-клеток и анализа Т-клеток, также не совпадают, что, вероятно, свидетельствует о том, Tat белок в разных типах клеток ведет себя по-разному, использует различные механизмы контроля клеточной генной экспрессии для создания благоприятствующей размножению и распространению вируса среды во всем организме.

# Геномика бактерий

## Слияние бактериальных видов с помощью пангеномного анализа

*Дильфуза Джамалова, Михаил Молдован, Михаил Гельфанд*

## Связь структуры пангенома и разнообразия местообитаний бактерий микробиома Земли

*Николаева Дарья Дмитриевна, Гарушянц Софья Константиновна*

Пангеном - это совокупность белок-кодирующих генов, присутствующих в наборе геномов одного вида или рода бактерий. Традиционно в структуре пангенома выделяют “универсальный геном” - гены, которые встречаются почти во всех рассматриваемых штаммах, и “периферию” - гены, встречающиеся у небольшого количества штаммов. Соотношение размера периферии к размеру ядра отличается у разных бактерий, какие именно факторы определяют это соотношение, до сих пор остается непонятным.

Мы предположили, что одним из определяющих факторов может являться разнообразие бактериальных сообществ, в которых встречается данный вид. Так, виды бактерий, способные существовать в разнообразных физико-химических условиях (виды-генералисты), должны, с одной стороны, иметь гены, необходимые для приспособления к конкретному естообитанию, а, с другой стороны, могут взаимодействовать с большим количеством бактерий других видов, от которых могут получать более разнообразные гены в результате горизонтального переноса. В то же время существуют виды-специалисты, которые привязаны к единственной экологической нише, хорошо приспособлены к ней и поэтому, вероятно, генетически более днородны. Получить информацию о структуре сообществ можно из метагеномных данных, когда совместно секвенируются все нуклеотидные последовательности, выделенные из

данного местообитания. Такой подход позволяет определить как качественный, так и количественный состав организмов разных экосистем.
Идея работы заключается в том, чтобы установить, существует ли связь между соотношением элементов структуры пангенома (универсального генома и периферии) и количеством местообитаний, в которых данный вид бактерий встречается. Ранее мы уже попытались ответить на этот вопрос, используя данные о бактериях-генералистах и специалистах [1] и предварительно получили положительный ответ, однако он требует подтверждения с использованием более масштабных данных. На этот раз используются данные Earth Microbiome Project [2], на основе которых будут выбраны виды бактерий-генералистов и специалистов для построения и анализа пангеномов.
[1] Sriswasdi S., Yang C., Iwasaki W. Generalist species drive microbial dispersion and evolution //Nature communications. – 2017. – Т. 8. – No. 1. – С. 1162.
[2] Thompson L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity //Nature. – 2017. – Т. 551. – №. 7681.

## Реконструкция геномных перестроек в бактериях рода *Vibrio*

***Кристина Перевощикова*, *Ольга Бочкарёва***

*Vibrio* – род бактерий, относящийся к классу *Gammaproteobacteria*. Многие вибрионы являются патогенами, вызывающими заболевания у человека и животных. Особенностью данного рода является наличие двух хромосом, первая из которых имеет больший размер, содержит больше базовых для жизнедеятельности бактерии генов и имеет первичное происхождение в эволюционной истории рода. Вторая же хромосома имеет плазмидное происхождение, содержит в основном нишеспецифичные гены наряду с небольшим количеством генов базового метаболизма и имеет значительно меньший и довольно широко варьирующийся среди разных видов размер. Некоторые виды *Vibrio* (3 из 114 исследуемых штаммов) имеют одну хромосому, сформированную в результате слияния двух. В связи с большим размером первая хромосома начинает реплицироваться раньше второй, что способствует локализации на ней наиболее активно транскрибируемых генов. Меньший уровень транскрипции генов второй хромосомы способствует их более быстрой эволюции. Такая гетерогенная организация генома позволяет поставить задачу сравнения пангеномов первых и вторых хромосом, реконструкции межхромосомных транслокаций и проверки гипотезы о действии отбора, сохраняющего положение транслоцированного участка относительно лидирующей/запаздывающей цепи.
С целью сравнить пангеномы обеих хромосом были построены U-кривые, пики которых были проассоциированы с филогенетическими группами (филогенетические взаимоотношения были установлены исходя из выравнивания панортологов). Пики на U-кривой для первых хромосом хорошо проявляются даже для больших филогенетических групп; для вторых они пропадают уже начиная с 26 видов, что говорит о большом генетическом и функциональном разнообразии вторых хромосом. Пики не соответствующие единым филогенетическим группам являются следствием фаговых вставок, событий интеграции F плазмид в геномную ДНК, и распространения общих путей метаболизма ароматических соединений, жирных кислот и углеводов. Ряды первых хромосом, соответствующие веткам *V. anguillarum*, *V. cholerae, V. mimicus* имеют перепредставленность GO категорий связанных с хемотаксисом и

двухкомпонентными сигнальными системами; *V. parahaemolyticus, V. campbellii, V. harveyi, V. alginolyticus* GO категорий связанных с системами секреции II и III типа; V. cholerae GO категорий TCP белков. Ряды вторых хромосом соответствующие *V. natriegens* имеют перепредставленность GO категорий, связанных с азотфиксацией*; V. anguillarum* GO категорий связанных с PTS системами для транспорта сорбитола. В целом же ряды первых хромосом имеют перепредставленность GO категорий рибосомных белков и белков связывания РНК и недопредставленность генов транскрипционных факторов, а ряды вторых хромосом имеют перепредставленность GO категорий транскрипционных факторов и клеточной коммуникации (TCRS). Это говорит о том, что вторые хромосомы обладают большим генетическим разнообразием даже внутри филогенетически единых групп, в то время как первые хромосомы несут общие гены вирулентности и клеточной коммуникации.

Дальнейшая работа предполагает реконструкцию межхромосомных транслокаций, проверку гипотезы о положительном отборе на транслокации с сохранением цепи и симметричные инверсии вокруг ori/ter, более глубокий анализ процесса слияния двух хромосом в одну у однохромосомных штаммов.

## Comparative genomics of Bacillus spp

*Moldir Zhiyenbayeva, Olga Bochkareva*

**Introduction**

Studies for individual species of Bacillus genera have been presented earlier, but a fullpangenomic analysis of such a large number of species is still lacking. We analyzed 103*Bacillus* species strains to find patterns that characterize particular groups of strains. Based onthese combined data phylogenetic and pan-genome analysis of *Bacillus* species werepresented in order to detect fractions of genes distinguishing one species of the genus*Bacillus* from another.

**Results**

1. Phylogenetic tree of Bacillus species were constructed. The tree turned out to be not homogeneous and strains from different species correspond to the same clade. However two well-separated clades were identified: clade of *Bacillus subtilis, Bacillus licheniformis, Bacillus amyloliquefaciens, Bacillus velezensis, Bacillus atrophaeus* species and clade of *Bacillus cereus, Bacillus thuringiensis, Bacillus anthracis, Bacillus cytoxicus, Bacillus mycoides, Bacillus megaterium* species. Remaining strains of *Bacillus xiamenensis, Bacillus pumilus* species were considered as outgroup. It was supposed that the main separation pattern of these species was pathogenicity/non-pathogenicity pattern.

2. Additionally, in order to represent a pangenome the gene frequency spectrum function was considered. Two main peaks in the spectrum function were found and they correspond to the large clade of *Bacillus cereus, Bacillus anthracis, Bacillus mycoides и Bacillus thuringiensis species.* Another visible peak is responsible for the second clade including strains of *Bacillus subtilis, Bacillus licheniformis, Bacillus amyloliquefaciens, Bacillus velezensis, Bacillus atrophaeus* species in combination with *Bacillus xiamenensis, Bacillus pumilus outgroup species.*

3. Orthologous groups of genes and annotations of each group were found. And now we are trying to find common functions of genes that form peaks of the U-curve.

**Future plans**

Rearrangements in the genomes of *Bacillus subtilis* species were identified, and they showed the considerable level of collinearity between them. According this, intergenic spaces of orthologous genes can be analyzed in the future work. Strain-specific functions of *Bacillus* species are also of particular interest for research because they allow to understand the mechanism of development of these species throughout the entire process of their genus evolution. To sum up, we can emphasize two points for future work:

1. To study rearrangements in clades of the phylogenetic tree of species;
2. To analyze intergenic spaces of orthologous genes.

## Геномные перестройки *Shigella* sp.

***Заира Сефербекова, Ольга Бочкарева***

Бактерии *Shigella* sp. вызывают бактериальную дизентерию и подразделяются на четыре группы — *S. dysenteriae, S. flexneri, S. boydii* and *S. Sonnei*. Несмотря на то что эти бактерии являются парафилетической группой внутри рода *Escherichia,* исторически они были выделены в отдельный род из-за медицинской значимости. Геномы бактерий *Shigella* содержат большое количество IS и характеризуются повышенной динамичностью, обеспечивающей возможность адаптироваться к изменяющимся условиям внутри клетки-хозяина.

Мы проанализировали 553 генома, в том числе 33 генома *Shigella*. Все геномы были поделены на 6 случайных групп, внутри которых был выполнен поиск ортологов. Из найденных ортологичных рядов было выбрано 239 универсальных. Внутри каждого универсального ряда соответствующие нуклеотидные последовательности были выравнены алгоритмом mafft. По конкатенату полученных выравниваний было построено дерево с помощью RAxML. С использованием пакета R GGRaSP на дереве было выделено 19 кластеров. *E.coli* разных филогрупп попали в разные кластеры, расположение *Shigella* на дереве также согласуется с литературными данными. Внутри каждой из групп мы построили синтенные блоки, реконструировали историю перестроек и сравнили их частоты на разных ветвях. В дальнейшем также планируется произвести поиск событий гомологичной рекомбинации и оценить ее частоты между штаммами *E.coli* и *Shigella*.

## Гомологичная рекомбинация в геноме *Escherichia coli*

***Гельфанд М.С., Афасижев Р.Н.***

**О чем:**

Мы определяем соотношения вклада рекомбинантных переносов и клонального наследования в распределение полиморфизмов по базовому геному *E. coli.*

Это **интересно**, потому что мы имеем возможность параметрически оценить частоту гомологичной рекомбинации у разных видов и вклад полиморфизмов, возникших по разным причинам.

**Что увидели**:

Параметры модели проявляют относительную устойчивость вне зависимости от выборки внутри близких филогрупп E.coli. Вклад полиморфизмов из клональных блоков в конечное распределение меньше, чем от рекомбинантных (примерно 0.2 против 0.8, соответственно).

**Планы:**

Расширить выборку на другие виды.

## Поиск паттернов в объединённой сети транскрипционной регуляции и метаболизма *E. coli*

***Валиев Иван***

Распространённым способом анализа биологических процессов (i.e. транскрипционная регуляция, метаболизм, белок-белковые взаимодействия) является их представление в виде графа. При их анализе наибольший интерес вызывают либо общие статистики, либо так называемый “профиль мотивов” - набор подграфов, перепредставленных в указанном графе. Ранее было показано, что для разных по своей природе объектов (e.g. пищевая сеть и сеть транскрипционной регуляции) профиль мотивов отличается.

Определённый интерес должны представлять графовые представления для нескольких биологических процессов одновременно, в особенности если они непосредственно связаны друг с другом. К сожалению, попытки построить подобные графовые представления единичны.

Основной идеей данного исследования было построить модель-граф, отражащую одновременно ранскрипционную регуляцию и метаболизм *E. coli*. А затем исследовать свойства этой модели.

Полученная в результате модель объединила в себе 207 транскрипционных факторов, 1516 генов, 1488 химических реакций и 1139 метаболитов. Для адекватности модели каждый из типов этих объектов был представлен в виде вершин отдельного цвета. Рёбра отражали соответствующие взаимодействия. Далее свойства модели были рассмотрены на примере циклических структур типа “метаболит-транскрипционный фактор-ген-химическая реакция-метаболит” в тех случаях, когда начальный и конечный метаболит был одним и тем же. Для этих струкур (петель обратной связи) были рассмотрены общее количество в графе, средняя длина и стандартное отклонение длины петель. Статистики петель обратной связи были сравнены с ожидаемыми с помощью пермутирования графа и перестановочного теста.

В результате указанных манипуляций было выявлено, что средняя длина петли обратной связи в реальном графе значительно ниже, чем в среднем в случайном графе, а стандартное отклонение длин - значительно выше.

В текущих планах проекта - выделить компоненты сильной связности в метаболической части графа, сколлапсировать их и посмотреть. как меняется его поведение.

## Перестройки оперонов метаболических путей бактерий

***Евгения Ходжаева*, *Зоя Червонцева***

Бактериальные опероны синхронизируют экспрессию закодированных генов. Однако за счет разной скорости деградации эффективное количество мРНК генов одного оперона может отличаться в сотни раз. Показано, что за регуляцию деградации часто отвечают специальные структурные элементы, которые часто консервативны на уровне семейства [1]. Вместе с тем состав оперонов может различаться даже у близких видов. В этой работе мы хотим проанализировать связь оперонных перестроек универсальных метаболических путей бактерий с возникновением/исчезновением подобных структур.

[1] Dar D, Sorek R. Extensive reshaping of bacterial operons by programmed mRNA decay. PLoS Genet. 2018;14(4):e1007354.

## Филогенетический анализ кассеты генов *Escherichia coli*, участвующей в деградации сульфоквиновозы и лактозы

***Анна Казнадзей*, *Анна Рыбина***

Недавно было показано, что *Escherichia coli* способна деградировать сульфо-содержащие соединения углеводов (Denger *et al*, 20014). Соответствующие гены (*yih*-гены) формируют в геноме кишечной палочки в 10-генную кассету. На основании сравнительного анализа кассет генов углеводного метаболизма бактерий выяснилось, что сочетание функций данной кассеты консервативно не только для бактерий семейства *Enterbacteriaceae*, но что также семь генов *yih*-кассеты кодируют белки, которые выполняют сходные функции с белками, закодированных в кассете генов бактерий класса *Bacilli* (Kaznadzey *et al*, 2018). Белки кассеты *Bacilli* участвуют в хорошо изученном пути катаболизма лактозы. Было выдвинуто предположение о том, что *yih*-кассета также участвует в деградации лактозы. Оно было подтверждено экспериментально - экспрессия четырех *yih*-генов значительно повышалась при росте клеточной культуры на лактозе. Также было показано, что эта кассета состоит из нескольких оперонов, и что глобальный транскрипционный фактор CRP и локальный фактор YihW из семейства регуляторных белков DeoR определяют характер их экспрессии в зависимости от присутствия в среде разных углеводов. Таким образом, *yih*-кассета оказалась вовлечена в разные метаболические процессы, контролируемые сложным регуляторным механизмом. Белки же, закодированные в кассете, по всей видимости, способны проявлять мультифункциональные свойства (известно, что некоторые бактериальные белки обладают такими характеристиками, однако масштабы этого явления еще предстоить выяснить).

До сих пор *yih*-гены не подвергали детальному филогенетическому анализу. В частности, неизвестно, как происходила эволюция кассеты, т.е. "сборка" ортологов *yih*-генов в кассету в разных геномах. Одна из целей данной работы - установить эволюционные отношения между генами yih-кассеты и оценить, насколько значимы события их ко-локализации. Вторая цель - попытаться выявить свойства, связанные с мультифункциональностью соответствующих белков. Наконец, не до конца изучены механизмы регуляции работы кассеты: в межгенных

областях кассеты было обнаружено несколько консервативных мотивов связывания для CRP, но для транскрипционного фактора YihW мотивы еще предстоит установить.

На данном этапе работы найдены ортологи генов *yih*-кассеты в выборке всех известных бактериальных геномов и начат анализ их взаимного расположения. Далее планируется построить филогенетические деревья и проследить на них события сборки и/или разборки кассеты. На следующем этапе будет осуществлено выравнивание межгенных последовательностей *yih*-кассеты разных штаммов *Eschericihia coli* и близкородственных видов и проведение филогенетического футпринтинга для поиска мотивов связывания локального транскрипционного фактора YihW. Наконец, мы собираемся установить структуры белков yih-кассеты в разных бактериях и попытаться найти связь структурных особенностей этих белков с их возможными мультифункциональными характеристиками

## Предсказание и сравненительный анализ лидерных пептидов триптофанового оперона Rhizobiales

***Алексей Шевкопляс*, *Зоя Червонцева***

Триптофановый оперон у некоторых бактерий регулируется транскрипционной аттенюацией, в ходе которой образуется так называемый триптофановый лидерный пептид. Раньше полагалось, что он уничтожается сразу же после трансляции, однако наши коллеги экспериментально обнаружили, что у *Sinorhizobium meliloti* и других представителей порядка Rhizobiales альфапротеобактерий он активирует гены устойчивости к тетрациклину и схожим антибиотикам. Интересно исследовать лидерный пептид не только нескольких модельных бактерий, но, разные лидерные пептиды, проследить их эволюцию. Это может, в частности, помочь разобраться в их физиологической функции. Также интересно научиться предсказывать не только лидерные пептиды, но и аттенюаторы.

Проанализировав геномы бактерий порядка *Rhizobiales,* мы выявили в них предположительные лидерные пептиды триптофанового оперона и сравнили их между собой, чтобы понять, могут ли они иметь сходные дополнительные функции. Лидерным пептидом считалась предшествовавшая триптофановый оперон (либо конкретно ген trpE) на той же нити ДНК возможно транслирующаяся аминокислотная последовательность, содержащая в себе по крайней мере два триптофана подряд.

Анализ 281 предположительных лидерных пептидов выявил несколько относительно консервативных групп пептидов, как правило, сопряжённых с таксономическим положением бактерий. Пептиды можно разделить на две основные группы: одна с двумя триптофанами подряд и консенсусной последовательностью MSTvvrPsRLWWRTss относится к бактериям *Bradyrhizobium sp.* и близких родов, а другая, более крупная группа, с четырьмя вариантами со слегка различной длиной и заметно различающимися последовательностями, слабо консерватным консенсусом которой является Mni(a/s)(n/i/v)WWWAR. Помимо триптофана, паттерны консервативности сильно различаются между группами.

Лидерный пептид триптофанового оперона нельзя назвать консервативным в Rhizobiales. Это контринтуитивно результатам, полученным экспериментаторами, потому что исходя из них, пептид обладает схожей функцией у различных представителей Rhizobiales.

Тем не менее, обращает на себя внимание хорошая сохранность аргинина, порой заменяющегося на лизин. Возможно, положительный заряд аргинина играет какую-то роль в образовании пептидом комплекса с мРНК, который был предсказан в экспериментальной работе.

В ходе работы были исследованы аттенюаторы Rhizobiales. Был написан скрипт, предсказывающих аттенюаторы у значительной части бактерий. У остальных бактерий аттенюаторы выявлялись полуавтоматически. Аттенюаторы, выявленные такими способами, похожи на аттенюаторы, предсказанные в более ранних работах. Предсказанные аттенюаторы служили дополнительным доказательством правомерности называть выявленные аминокислотные последовательности лидерными пептидами триптофанового оперона.

Итак, были предсказаны лидерные пептиды триптофанового оперона 281 альфапротеобактерий Rhizobiales. Несмотря на существенную физиологическую функцию пептида, последовательность пептида у представителей разных родов отличается достаточно сильно. Впрочем, обращает на себя внимание устойчивое сохранение положительно заряженных аминокислот. В дальнейшем было бы интересно сконцетрироваться на сравнении не только пептидов, но и рибонуклеотидных последовательностях аттенюаторов, а также на улучшении метода предсказания этих самых аттенюаторов.

## Исследование комплементарных взаимодействий продуктов транскрипции 6S-РНК с мРНК бактерий

***Наталья Транкова, Елена Кубарева, Дмитрий Первушин***

Одной из наиболее экспрессируемых малых некодирующих РНК у бактерий является 6s РНК, которая обладает способностью связывать холоферменты РНКП благодаря консервативной вторичной структуре в форме шпильки с большой расплетенной центральной петлей. Связывание РНКП с 6s РНК приводит к ингибированию транскрипции множества генов. Одной из особенностей 6s РНК является ее способность при определенных условиях служить для РНКП матрицей для синтеза коротких пРНК (2-25 н.о.), комплементарных центральной части молекулы. Синтез пРНК в конечном счете приводит к высвобождению РНКП, что обеспечивает обратимость ингибирования. Цель проекта заключается в том, чтобы, во-первых, методами филогенетического анализа определить, существуют ли комплементарные взаимодействия пРНК с мРНК клеток, для чего, например, сравнить компенсаторные мутации в 6s РНК и соответствующих сайтах мРНК. Вторая задача заключается в том, чтобы установить, существует ли на глобальном уровне избегание таких пРНК-мРНК взаимодействий на
уровне синонимичных замен, то есть таких замен, которые бы не влияли на последовательность белка, закодированную в мРНК, но при этом препятствовали бы ее связыванию с пРНК.

# Эволюция последовательностей

## Мутационные паттерны в эволюции *Escherichia coli*

***Софья Гарушянц**, Мария Селифанова, Арина Колотова, Анна Казнадзей, Михаил Гельфанд*

Бактерии – это идеальные объекты для изучения эволюции, потому что они достаточно легко культивируются, быстро размножаются и имеют небольшие и достаточно простые, по сравнению с эукариотами, геномы. Более того, к настоящему моменту накоплено достаточно много экспериментальных данных о мутациях в таких модельных объектах, как *Escherichia coli* и *Bacillus subtilis*. Часто изучение мутаций в бактериях в ходе эволюции, например как в долгосрочном эволюционном эксперименте (LTEE), предполагает исследование конкретных оснований, в которых произошли замены, однако контекст, в котором они произошли не учитывается. Тем не менее, подход использующий информацию о контекстах мутаций широко применяется в эукариотической геномике, особенно в исследовании раков. Обычно в таких исследованиях изучают распределение мутаций по всем возможных трехнуклеотидным (один нуклеотид слева и справа от мутации) контекстам, которые называют мутационными подписями. Использование информации о контексте мутации позволяет более точно разделять происходящие в клетке мутационные процессы. К настоящему моменту, нет работ, которые бы применили этот подход к бактериальным геномам и к дивергенции бактерий.

Целью данной работы является определение подписей мутационных процессов, происходящих в кишечной палочке и применение полученных подписей для определения основных процессов, определяющих дивергенцию штаммов.

Для определения подписей мутационных процессов использовались данные о бактериях-мутаторах LTEE и данные о мутациях, накапливаемых в линиях кишечных палочек с поломками в определенных генах. В ходе работы удалось выделить мутационные подписи, характерные для поломки системы репарации неспаренных оснований, генов *mutY*, *mutT* и *dnaQ*.

На предварительной стадии исследования вклад отдельных подписей был оценен для небольшого количества штаммов, и было показано, что в дивергенцию значительный вклад вносят поломки системы неспаренных оснований и *dnaQ*.

Для того, чтобы провести более аккуратный анализ было решено использовать все (522) полные, хорошо проаннотированные геномы кишечных палочек. В настоящий момент построено выравнивание и филогенетическое древо для всех таких штаммов и на них планируется провести более аккуратный анализ.

## Анализ частот встречаемости мутации в зависимости от контекста на примере 55 геномов 9 видов бруцелл

***Научный руководитель - Алексеевский Андрей Владимирович.***
***Волобуева Мария Евгеньевна.***

Целью моей работы является проанализировать встречаемость однонуклеотидных мутаций в зависимости от контекста (нуклеотидов, находящихся в непосредственной близости от полиморфизма), изучаемых бактерий. Для работы было взято только ядро пангенома – совокупности выравниваний блоков высокосходных нуклеотидных последовательностей, в каждый из которых входит ровно по одному фрагменту из каждого генома. В каждом блоке из ядра отобраны позиции такие что, (1) в позиции нет символов гэпов; (2) встречается ровно два разных нуклеотида; (3) разбиение геномов, определяемое тем, какой нуклеотид стоит в данной позиции, совпадает ли с разбиением, определяемым одной из ветвей филогенетического дерева геномов (Такое дерево, построенное по нуклеотидному ядру, выдается NPG-explorer'ом, оно укорененное); (4) две соседних позиции с каждой из сторон являются абсолютно консервативными. Такая позиция интерпретируется как изолированная замена нуклеотида у общего предка геномов из клады, соотвествующей ветви в контексте, определяемом соседними позициями.

Полная информация по таким мутациям собрана с помощью скриптов из нуклеотидного пангенома, и мутации сгруппированы в соответствии с кладами филогенетического дерева.

В дальнейшем планируется сравнить подписи мутаций для разных видов (виды бруцелл определяются хозяином) и, если получится, связать типы мутаций с известными мутагенами.

## Evolution of germline mutational spectra among great apes.

**Semenchenko Egor**

Evolutionary and medical genomic studies depend on an understanding of mutagenesis. The picture of how mutations are distributed along the genome is essential in the evaluation of the functional role of genes and inferring the population's history. There are numerous studies showing that mutational rate in germline and somatic cell lines varies within the genome and between species. Although this field is intensively studied, there is limited information on the causal factors that may drive changes in mutational spectra.

We attempt to infer mathematically stable mutational signatures in humans and the nearest species in the germline. The ultimate goal of the project is to find associations of detected signatures with underlying biochemical processes affecting the shape of mutational spectra.

The common way to study mutagenesis is investigating single nucleotide mutations (SNV's). They are the most frequent type of mutational events. Since the data on de-novo mutations in apes are limited, as the primary data sources we use species divergence and polymorphism datasets.

At the initial steps of the project, we have built the computational framework for obtaining SNV mutational spectra within the context of 5' and 3 flanking nucleotides. We have accounted for the selection pressure and determined how the known structure of the mutational spectrum of species polymorphism reproduces the spectrum of their divergence.

Recent objectives are focused on the mutational spectra distribution within 100kb -1Mb windows along species genomes. We are testing the means of blind source decomposition: Independent Component Analysis (ICA) and Non-Negative Matrix Factorization (NMF) to achieve statistically reliable observations of mutational signatures.

## Микроинверсии в мире человека и других приматов

***Н.А. Потапова*, *А.С. Кондрашов***

Инверсии – это хромосомные перестройки, в результате которых фрагмент хромосомы поворачивается на 180º. Про большие инверсии, порой затрагивающие целые участки плеч хромосом, известно многое. А про микроинверсии длиной в несколько десятков или сотен нуклеотидов исследований мало и информация в них очень противоречивая. Мы решили узнать, сколько же микроинверсий на самом деле между человеком и шимпанзе, и какие из них появились на ветви человека, а какие – на ветви шимпанзе.

Оказалось, что микроинверсий намного меньше, чем заявлялось ранее. В противовес информации о нескольких тысячах таких событий, мы обнаружили только около 200. Видимо, это связано с фильтрацией и тем, что мы не берём во внимание потенциальные микроинверсии длиной 5-15 нп, потому как это это могут быть кластерные мутации или артефакты секвенирования или сборки. Микроинверсий, находящихся в генах, и регионов с повышенным количеством микроинверсий мы не нашли.

В дальнейшем мы посмотрим, сколько микроинверсий происходит между двумя людьми и, также, между шимпанзе и бонобо. Ожидается, что их будет меньше, чем между человеком и шимпанзе и мы сможем опустить планку для поиска микроинверсий до 8-10 нуклеотидов. И узнаем бывают ли такие небольшие микроинверсии, или нет.

## Эволюция последовательностей, окружённых микросателлитами

***Н.А. Потапова*, *А.С. Кондрашов***

Микросателлиты – это повторяющиеся последовательности с периодом 6-10, которые суммарно могут достигать длины до 1000 нуклеотидов. Они быстро эволюционируют и в исследованиях часто изучают то, как микросателлит влияет на окружающие его обычные последовательности. Но не было известно о том, что происходит с обычной небольшой по длине последовательностью, которая фланкирована двумя микросателлитами.

На выравнивании геномов человека и шимпанзе мы рассмотрели разные ситуации, в которых длина обычной последовательности, окружённой микросателлитами, была до 300, до 500, до 1000 нуклеотидов. И увидели, что присутствие двух окружающих микросателлитов увеличивает количество замен в последовательности, по сравнению со случаями, когда микросателлитов вокруг нет. Для повторов с небольшим периодом этот эффект выражался чётче, чем для повторов с периодом около 10.

Мы планируем ещё посмотреть на влияние длины микросателлитов на изменения во фланкируемой последовательности, и на то, бывают ли случаи, когда эта последовательность вообще пропадает, поглощённая микросателлитами.

## Изучение консервативных геномных участков в популяции человека

***Мария Селифанова, Дмитрий Алексеевич Коркин, Михаил Сергеевич Гельфанд***

Объект нашего исследования — Long Identical Multispecies Elements (LIMES). ЛАЙМы были найдены с помощью нового подхода, позволяющего идентифицировать ультраконсервативные элементы как в синтеничных, так и в несинтеничных областях достаточно отдаленных геномов. Таким образом были выявлены сотни идентичных многокопийных элементов, расположенных в разных областях геномов почти всех позвоночных, некоторых беспозвоночных и даже грибов. Природа, а также эволюционный механизм, лежащий в основе этого явления, до сих пор не выяснены. Кроме того, неизвестно как ЛАЙМы ведут себя в популяциях. В частности, в популяции человека.

Проект делится на две части. Цель первой состоит в том, чтобы исследовать ЛАЙМы в популяции здоровых людей и в образцах рака, где частота мутаций значительно выше, чем в нормальных клеточных линиях. Вторая часть изначально предполагала задачу классификации копий человеческих ЛАЙМов по степени их синтеничности, с целью использовать потом эти данные в первой части проекта. Сейчас эволюционная часть расширилась: мы хотим построить карту «рождения и смерти» копий ЛАЙМов в геномах позвоночных. В частном случае на основании этих данных можно сделать и классификацию человеческих ЛАЙМов.

На данный момент я работаю над эволюционной частью проекта: написан пайплайн для нахождения всех копий всех ЛАЙМов в нужных геномах, а также их окружений. Разработан метод определения синтеничных кластеров, сейчас я проверяю насколько этот метод работает. Популяционная часть проекта была временно заморожена.

На сегодняшний день есть промежуточные результаты в виде написанных алгоритмов и их выдачи с небольших сэмплов. К марту планируется в каком-то виде получить результаты эволюционной части проекта, а также разработать подробный план по популяционной: так, в доклад планируется включить то, какие базы данных будут использованы, в каком виде будут отображены результаты, и так далее.

Планы на будущее просты: выполнить цели, заявленные в моей курсовой. Параллельно хорошей идеей кажется проверить ЛАЙМы на их связь с SNP, найденных с помощью GWAS, а также посмотреть соответствующие ЛАЙМам эпигенетические метки.

## Кратность межгенных областей и альтернативные стоп-кодоны в *Enterobacteriales*

***Колотова Арина, Гарушянц Софья, Казнадзей Анна***

Ряд однонуклеотидных замен, происходящих в оперонных последовательностях, приводит к инактивации стоп-кодона. В таких случаях во время трансляции не происходит терминация на стоп-кодоне, и синтез полипептидной цепи продолжается дальше, что может приводить к слиянию соседних генов и образованию более длинного белкового продукта. Этот процесс наблюдали для некоторых бактериальных белков. Пока неясно, как устроен отбор, направленный против таких слияний. Факторы, которые могут препятствовать слиянию соседних генов – это наличие достаточно консервативных стоп-кодонов после основного стоп-кодона и кратность числа нуклеотидов в межгенных областях.

Целью работы является исследование этих факторов на примере порядка *Enterobacteriales*, т.е. изучение кратности межгенных областей внутри и между операонами и консервативности альтернативных стоп-кодонов. Помимо этого, предполагается выяснить, в каких случаях более вероятно эволюционное закрепление мутаций, приводящих к слиянию генов. В будущем интересно проследить, существует ли между склонными к слиянию генами устойчивая функциональная или структурная конгруэнтность. Данный проект может открыть неизвестные до этого закономерности в эволюции оперонных генов.

К настоящему моменту, мы проверили кратность (делимость на три) межгенных областей внутри оперонов и вне оперонов. Внутри оперонов при делении самых мелких межгенных расстояниях самый частый остаток был равен нулю, а с увеличением размеров нетранслируемой области самым частым результатом деления межгенных расстояний по модулю три становится 2. Вне оперонов никаких закономерностей ожидаемо выявлено не было. Из этого, вероятно, следует, что если расстояние между генами внутри оперона мало, эти гены более склонны к слиянию с сохранением той же рамки считывания.

Полученный результат может свидетельствовать о наличии механизма защиты от слияния генов внутри оперонов, разделенных длинными межгенными участками, в результате мутации в стоп-кодоне. В дальнейшем планируется исследование консервативности альтернативных стоп-кодонов. Наличие устойчивых альтернативных стоп-кодонов в межгенных областях внутри оперонов в различных штаммах анализируемых видов *Enterobacteriales* может быть способом защиты от слияния белковых продуктов.

## Оценка действия отбора на межгенные участки Saccharomyces cerevisiae.

***Александр Соколов, Павел Шелякин, Михаил Сергеевич Гельфанд***

Задача исследования состоит в том, чтобы посмотреть, как устроен отбор в межгенных областях дрожжей. Поскольку есть несколько сотен практически полных геномов *Saccharomyces cerevisiae* и несколько геномов других представителей рода *Saccharomyces*, то это можно делать на разных уровнях.

На первой стадии построим выравнивания ортологичных межгенных областей разных штаммов *S. cerevisiae* и оценим количество нуклеотидных замен в них. Это позволит в некоторой степени оценить силу отрицательного отбора, действующего на межгенные области.

Далее, посмотрим имеются ли в межгенных областях островки консервативных позиций, которые могут представлять собой регуляторные последовательности, например, сайты связывания факторов транскрипции или последовательности, после транскрипции формирующие вторичные структуры РНК. В таких островках мы посмотрим наличие сайтов связывания известных факторов.

Далее, в ходе исследования планируется:

1. Посмотреть отбор в генах и в межгенных участках. Построить профиль консервативности
2. Посмотреть данные о доступности хроматина и о расположении известных сайтов связывания регуляторных факторов
3. Спроектировать нейронную сеть, предсказывающую профиль консервативности и доступность хроматина по последовательности штамма.

# Susceptibility of single-stranded DNA to APOBEC mutagenesis during transcription

***Almira Chervova, Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology***

**Background**

The apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) is the family of enzymes of the human innate immune system that recently has been implicated in cancer mutageneses. APOBEC-associated mutations have been detected in many types of human cancer, including breast, lung, bladder, head/neck and cervical cancers. As the APOBEC enzymes have a strong specificity toward single-stranded DNA (ssDNA), it was suggested that enzyme can mutate host DNA in one or several cellular processes associated with the unwinding of double-stranded human DNA, such as DNA repair, replication or transcription. Indeed, evidences that APOBEC mutagenesis is associated with the replication were recently obtained in findings demonstrating that the density of APOBEC-induced mutations has a strong bias toward the lagging replication strand (Seplyarskiy et al., 2016) and is elevated in early-replicated regions (Kazanov et al., 2015). The first evidences for the connection of APOBEC mutagenesis and transcription was obtained in whole-genome, exome and transcriptome study of bladder cancer and in recent study in yeasts showed susceptibility of the non-template strand of tRNA genes to APOBEC mutagenesis. However, the study analyzing the distribution of APOBEC-induced mutations across the genomes of 119 breast and 24 lung cancer samples did not find statistically significant difference of the density of APOBEC-induced mutations between transcribed and non-transcribed genomic regions, leaving the relevance of transcription to APOBEC mutagenesis in question.

**Results**

Here, we analyzed the whole genome and transcriptome sequencing data from the study of 505 tumor genomes across 14 cancer types (Fredriksson et al., 2014), in attempt to investigate the connection of APOBEC mutagenesis and transcription. This dataset has approximately five times more cancer samples and, consequently, much more mutation data than we used in our previous study (Kazanov et al., 2015). First, we sought to validate the effect of elevated density of APOBEC-induced mutations in early replicated regions, which was observed in our previous study. We found that this effect is not universal for the samples from considered dataset, which are enriched in APOBEC-induced mutations. For the part of the APOBEC-enriched cancer samples this effect was not observed even in a small degree. Then, we performed the analysis of density of APOBEC-induced mutations in genes depending on expression level and found the elevated density of mutations in highly expressed genes only in samples, having aforementioned effect for replication. Thus, we observed that cancer samples enriched in APOBEC-induced mutations are divided into two classes – samples with the elevating density of APOBEC-induced mutations in early replicated regions and highly expressed genes, and samples, which have elevated density in late replication regions and low expression genes.

Seplyarskiy, V.B., Soldatov, R.A., Popadin, K.Y., Antonarakis, S.E., Bazykin, G.A., and Nikolaev, S.I. (2016). APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. Genome Res. 26, 1–9.

Kazanov, M.D., Roberts, S.A., Polak, P., Stamatoyannopoulos, J., Klimczak, L.J., Gordenin, D.A., and Sunyaev, S.R. (2015). APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. Cell Rep. 13, 1103–1109.

# Белки

## Поиск скомпенсированных сдвигов рамки считывания и определение их роли в образовании новых аминокислотных последовательностей.

***Дмитрий Биба*[1], *Галина Клинк*[2], *Георгий Базыкин*[2]**
***1-биофак МГУ, 2-Сколтех***

Среди мутаций встречаются сдвигающие рамку считывания. Это инсерции или делеции нуклеотидов, число которых не кратно трём. Обычно такие мутации очень вредны, потому что изменяют все аминокислоты в белке, находящиеся ближе к 3'-концу. Есть случаи, в которых они должны быть не очень вредны или даже не вредны – если аминокислот, находящихся ближе к 3'-концу, мало (мутация происходит близко к концу гена) или если случается другая инсерция или делеция неподалёку от первой, и суммарное число нуклеотидов в этих двух инсерциях (делециях) становится кратно трём. Тогда будут изменены только аминокислоты, находящиеся между этими мутациями, и это может быть не очень вредно. Я ищу такие случаи (скомпенсированные сдвиги рамки считывания) в отсеквенированных геномах позвоночных.

Это интересно, потому что если такие случаи встречаются, и нередко, то они могут быть материалом для образования новых аминокислотных последовательностей и потенциально – для образования совершенно новых белков, мало похожих на свои предковые формы.

Планируется поискать скомпенсированные инсерции/делеции в геномах позвоночных (база данных UCSC). Если они найдутся, можно будет рассуждать об их роли в эволюции позвоночных.

## Анализ и предсказание эффекта коротких инделов в белках

***Мария Молчанова (Московский физико-технический институт), Анастасия Жарикова (Факультет биоинженерии и биоинформатики МГУ), Ольга Калинина (Helmholtz Institute for Pharmaceutical Research Saarland), Василий Раменский (Национальный медицинский исследовательский центр профилактической медицины)***

Короткие вставки и удаления участков генома (инделы) являются наиболее распространенным после однонуклеотидных вариантов типом полиморфизма генома. Инделы, расположенные в кодирующих участках генов и не сдвигающие рамку считывания гена, представляют первоочередный интерес ввиду их возможности влиять на структуру и функции генов, а следовательно, на фенотипические проявления на всех уровнях организации живой системы. Поскольку масштабная экспериментальная проверка функциональности большого количества инделов невозможна, возникает задача вычислительного предсказания их эффекта

и отделения потенциально функциональных случаев от нейтральных. Существующие методы предсказания имеют ряд недостатков. Цель данного проекта заключается в поиске и анализе информативных свойств белковых инделов и разработке метода вычислительного предсказания их эффекта.

На данный момент нами были выполнены следующие задачи: создание базы данных болезнетворных и нейтральных инделов в человеческих белках на основе ресурсов ClinVar и gnomAD, тестирование разработанных ранее предсказательных методов с помощью созданной базы для независимой оценки их эффективности; разработка вычислительной процедуры для построения репрезентативных качественных выравниваний исходного белка с гомологами; поиск информативных свойств инделов, которые будут использованы для предсказания их функциональности.

На основе полученных информативных свойств инделов с помощью методов машинного обучения будет разработан метод предсказания функциональности инделов. Для этого будет выбран метод машинного обучения (метод поддерживающих векторов, метод случайного леса, байесовские методы), обеспечивающий наилучшие результаты. По результатам процедуры кросс-валидации на собранной базе данных болезнетворных и нейтральных инделов будут описаны параметры, характеризующие точность предсказания, в частности, чувствительность, специфичность и аккуратность. С помощью разработанного метода будет анализирована полная созданная выборка человееческих инделов с известными популяционными частотами и составлена таблица их параметров и предсказаний функционального эффекта. Будет проверена гипотеза о том, что по мере уменьшения популяционной частоты инделов увеличивается доля предсказанных как потенциально болезнетворные, а также описана доля таковых среди предположительно безвредных (benign) вариантов базы ClinVar.

Работа поддерживается грантом РФФИ № 18-04-00789

## Что стоит за предпочтениями аминокислот в белковых сайтах

***А. Столярова, Г. Базыкин, М. Гельфанд, А. Миронов***

Одной из основных концепций эволюционной биологии является адаптивный ландшафт - функция приспособленности на пространстве генотипов. Для одиночного аминокислотного сайта можно определить однопозиционный адаптивный ландшафт (ОПАЛ) как вектор приспособленностей всех 20 возможных вариантов.

Современные методы, такие как глубокое мутационное сканирование, позволяют получать прямые измерения ОПАЛ. На опубликованных результатах мутационного сканирования для разных белков дрожжей, бактерий и вирусов мы показали, что приспособленности схожих по физико-химическим свойствам аминокислот скорректированы положительно, в то время как приспособленности непохожих аминокислот (например, гидрофобных и заряженных) скорректированы отрицательно. Эти результаты соответствуют данным, полученным по выравниваниям (матрицы BLOSUM).

Мы хотим исследовать структуру пространства ОПАЛ: определить их консервативность среди разных наборов данных; определить, образуют ли ОПАЛ

для различных сайтов кластеры и, если да, есть ли случаи, когда один и тот же сайт белка в разных штаммах вируса принадлежит к разным кластерам.

## Prediction of structural susceptibility of proteins to regulatory proteolysis

***Vyacheslav Safronov, Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University***

**Background**

Regulatory proteolysis, a specific cleavage one or a few protein's peptide bonds, is a irreversible post-translational modification playing an important role in many biological pathways like blood coagulation, cell proliferation, and apoptosis. More than five hundred proteases are known in human. Disruption of the proteolytic pathways may lead to many pathologies, including inflammation and cancer. Thus, the proteases and their substrates are important diagnostic and therapeutic targets, and their identification is of great importance.

The main goals of the proposed project are (i) improvement of the developed early method for prediction of structural susceptibility of proteins to proteolytic processing; (ii) development of the new prediction method combining prediction of the structural susceptibility to proteolysis and prediction of proteolytic sites based on protease sequence specificity; (iii) construction of the predicted proteolytic networks by application the developed method to human proteome.

**Methods/Results**

The dataset representing registered regulatory proteolytic events in human proteins with known 3D structures was generated as follows. Information about known proteolytic events was extracted from CutDB and defined as a unique combination of three elements: (i) protein substrate (NCBI Accession), (ii) protease (MEROPS ID), and (iii) position of a cleavage site (from original publications available in PubMed). Structural information for a subset of protein substrates with experimentally solved 3D structure was obtained by scanning the PDB using BLAST. Then, BLAST output was filtered using following thresholds: sequence identity > 90%, the length of alignment cover more than 90% both query and fetched sequence. Our final data sets included 359 nonredundant protein substrates with experimentally solved structures.

Then we comprised a training set, which will be used as an input to machine learning methods, as a set of structural descriptors derived from the experimentally solved structures. This dataset included solvent accessibility, hydrogen bonding, torsion angles, and secondary structures calculated using DSSP software. Other structural descriptors, including packing, protrusion, depth index, B-factor and disordered regions were calculated.

## Эволюция фосфорилируемых аминокислот

***Михаил Молдован, Михаил Гельфанд***

Посттрансляционные модификации играют важную роль практически во всех клеточных процессах. Особое место среди них занимает фосфорилирование – самая распространенная

модификация. На сегодняшний день известно о сотнях тысяч сайтов фосфорилирования в самых разных организмах, что позволяет исследователям проводить крупномасштабные проекты, посвященные эволюции этого феномена. В ходе нескольких исследований было установлено, что частоты замен фосфорилируемых серинов на другие аминокислоты отличаются от таковых для нефосфорилируемых серинов.

В нашей работе мы исследуем различные аспекты эволюции фосфорилируемых аминокислот. Сайты фосфорилирования можно определять разными способами: как все сайты, фосфорилирующиеся при каких-либо условиях, как консервативные сайты фосфорилирования, как сайты, фософорилирующиеся в малом или в большом числе тканей или как-нибудь еще. Частоты замен фосфорилируемых серинов на отрицательно заряженные аминокислоты, как оказалось, сильно зависят от способа определения сайта фосфорилирования. Так, консервативные сайты фосфорилирования, а также сайты с высоким “показателем качества” и сайты, открытые в более ранних исследованиях демонстрируют более высокую частоту замен на аспартат или глутамат, в то время, как недавно открытые фосфорилируемые серины и серины
с низким показателем достоверности фосфорилирования меняются на другие
аминокислоты примерно с теми же частотами, что и их нефосфорилируемые
аналоги.

Термин “островки фосфорилирования” часто используется в цитологических и молекулярно-биологических работах. При этом сами островки фосфорилирования систематически не изучались. Используя различные определения сайтов фосфорилирования, мы установили, что до половины фосфорилируемых серинов находится в островках и исследовали
частоты их замен на другие аминокислоты.

Проект пока далек от завершения. Планируется проверить, перепредставлен ли положительный отбор на замены серина в какие-либо аминокислоты в случае фосфорилируемых серинов. Также интересно было бы посмотреть, как меняются паттерны фосфорилирования при образовании
паралогов и провести популяционно-генетический анализ сайтов фосфорилирования.__

## Были ли многодоменные белки у LUCA

***Научный руководитель – Алексеевский Андрей Владимирович***
***Ириоглов Роман Андреевич***

***Задача:*** подтвердить или опровергнуть гипотезу о наличии у последнего универсального общего предка (LUCA) многодоменных белков.

***Первые результаты:*** среди студенческих работ 4го семестра найдены примеры доменных архитектур, предположительно, существовавшие у общего предка всех организмов - LUCA.

Доменные архитектуры на страницах студентов можно увидеть по ссылкам:
http://kodomo.fbb.msu.ru/~Shadrina92/term4/practice11.html
http://kodomo.fbb.msu.ru/~impbios93/Term4/Practice10/report10.html

***Планы на будущее:***

1) Для всех доменов составить таблицу доменных архитектур с указанием числа представителей по таксонам археи, бактерии, эукариоты.

2) Выбрать перспективные примеры.
3) Реализовать программно методику.

Метод: пусть выбран домен А и две доменные архитектуры, включающие домен А и выбраны таксоны высокого уровня. В идеале, эукариоты, бактерии и археи. Построим филогенетическое дерево доменов А из всех белков, его содержащих. Листья пометим доменной архитектурой белка и таксоном. Если это дерево окажется состоящим из двух клад, соответствующих доменным архитектурам, то можно сделать вывод, что белки с обеими архитектурами существовали у общих предков рассматриваемых таксонов.

# Филогенетика и молекулярная эволюция

## Накопление мутаций в экспериментальной эволюции базидиомицета *Schizophyllum commune*

***Безменова А., Звягина Е., Федотова А., Неретина Т., Пенин А., Кондрашов А.***

Базидиомицет *Schizophyllum commune* – уникальный организм, обладающий наибольшим полиморфизмом среди изученных видов, который может достигать 20% в популяции, и высокой скоростью мутирования, составляющей 2.0*10-8 замен/нуклеотид/поколение. В жизненном цикле *S. commune* присутствует гаплоидная одноядерная стадия, которую можно легко культивировать на твердых средах. Мы разработали экспериментальную систему, которая позволяет нам изучать процесс накопления соматических мутаций в процессе вегетативного роста гаплоидного мицелия *S. commune* в длинных трубках фиксированного диаметра в течении длительного времени. Мы использовали трубки двух разных диаметров. Тонкие трубки с диаметром 0.8 – 0.9 мм были призваны минимизировать эффект естественного отбора и позволить оценить скорость спонтанного мутагенеза на клеточное деление. Более толстые трубки с диаметром 4 мм, в которых эффективный размер популяции растущих гиф в мицелии достаточно велик, служили системой, в которой в мицелии может действовать естественный отбор. Мы исследовали 24 производных линии 4-х исходных культур *S. commune* - половина линий культивировались в тонких трубках, половина - в толстых. Мы вырастили 75 – 100 см мицелия для каждой линии в тонких трубках и 2 – 2.5 м мицелия для линий в толстых трубках, и секвенировали фрагменты мицелия каждые ~30 см. Таким образом, к настоящему моменту наши популяции были отсеквенированы от 4 до 7 раз. Всего было проведено 112 секвенирований растущих популяций, а также 4 секвенирования геномов мицелиев- основателей. На данный момент мы обнаружили 289 новых мутаций в экспериментальных популяциях. Большинство этих мутаций (213) достигали высоких частот и фиксировались в мицелии, однако мы также наблюдаем небольшое число замен, достигших высокой частоты, однако в дальнейшем не зафиксировавшихся и исчезнувших из популяции растущих гиф, а также

замен, высоких частот не достигших. Часть обнаруженных мутаций оказались кодирующими, однако число таких мутаций оказалось недостаточным, чтобы можно было делать выводы о достоверных различиях в доле кодирующих или несинонимичных мутаций в трубках разных диаметров. Мы обнаружили, что скорость накопления мутаций трубках диаметром 0.8 мм (3.36*10-11 замен/нуклеотид/клеточное деление, 95% CI: 2.41*10-11 – 4.30*10-11) оказалась примерно в два раза больше, чем в трубках диаметром 4 мм (1.43*10-11 замен/нуклеотид/клеточное деление, 95% CI: 0.77*10-11 – 2.10*10-11).

## Зависимость скорости гомологичной рекомбинации от уровня гетерозиготности хромосомы в базидиомицете *Schizophyllum commune*

***Безменова А., Звягина Е., Федотова А., Сеплярский В., Кондрашов А.***

Базидиомицет *Schizophyllum commune* – уникальный организм, обладающий наибольшим полиморфизмом среди изученных видов, который может достигать 20% в популяции. Таким образом, *S. commune* является очень удобным объектом для исследования многих эволюционных процессов с недостижимыми доселе разрешением и точностью. В частности, мы можем с большой точностью исследовать гомологичную рекомбинацию. Стоит отметить, что в свете очень высокого полиморфизма не до конца ясно, как может работать гомологичная рекомбинация в таком виде. Было показано, что у *S. Commune* горячие точки рекомбинации чаще встречаются в генах – то есть в более консервативных областях, что обычно избегается в организмах с не таким высоким полиморфизмом, так как рекомбинация сопряжена с повышенным уровнем спонтанного мутагенеза. Мы разработали экспериментальную систему, которая позволяет нам напрямую изучать зависимость скорости гомологичной рекомбинации от уровня гетерозиготности участка хромосомы. Для этого мы скрещиваем две особи *S. commune*, и среди потомков F1 выбираем тех, у кого в некоторых хромосомах произошло два кроссинговера – то есть плечи хромосомы имеют генотип одного из родителей, а центральная часть – другого. Таких потомков мы скрещиваем с обоими родителями. При этом генотипирование плеч хромосом позволяет нам эффективно оценивать число событий рекомбинации внутри центрального фрагмента – в том числе, когда этот фрагмент полностью гомозиготен – и сравнить скорость рекомбинации в гомозиготном (при скрещивании с одним из родителей) и гетерозиготном (при скрещивании с другим родителем) фрагменте. На данный момент получены и отсеквенированы 25 потомков F1 и ведется анализ структуры хромосом.

## Неравновесие по сцеплению с высоким разрешением в *Schizophyllum commune*

***А. Столярова, М. Логачева, А. Кондрашов, Г. Базыкин***

Базидиомицет *S. commune* является самым высокополиморфным известным

эукариотическим организмом (геномы двух образцов S. commune из американской популяции отличаются в среднем на 20% по синонимическим сайтам). Этот факт, а также маленький размер генома, легкость культивирования и присутствие гаплоидной стадии в жизненном цикле гриба делает его перспективной моделью для эволюционной и популяционной геномики.

В нашей лаборатории секвенировали, собрали de novo и выровняли 54 генома S.commune. В полученном наборе данных было определено около 7 миллионов однонуклеотидных полиморфизмов, из которых около 80% - биаллельные. Высокая плотность полиморфизмов позволяет использовать полученные данные для изучения закономерностей неравновесия по сцеплению (LD) с большим разрешением.

Мы показали, что неравновесие по сцеплению между несинонимическими сайтами сильнее, чем между синонимическими на том же расстоянии, причем этот эффект наблюдается независимо от частот минорного аллеля в этих сайтах. Кроме того, LD между сайтами на одном и том же расстоянии оказалось сильнее, если эти сайты принадлежат разным экзонам, чем если они принадлежат одному экзону. Этот эффект частично, но не полностью может быть объяснен тем фактом, что LD в целом оказалось сильнее в коротких экзонах, чем в длинных. Мы также показали, что неравновесие по сцеплению между концами экзона сильнее, чем между, например, началом экзона и его серединой, что не соответствует ожидаемому монотонному снижению LD с расстоянием.

Мы планируем использовать симуляции, чтобы понять, может ли разница в силе отбора, действующего на сайты разных классов, объяснить полученные
результаты. Кроме того, мы хотим учесть влияние рекуррентных мутаций, которые вероятны в настолько полиморфной популяции, и других возможных факторов.

## Адаптивная динамика на ландшафте, заданном матрицей эпистатических взаимодействий

***Ксения Худякова*, *Алексей Кондрашов***

Адаптивный ландшафт - это отображение пространства всех возможных генотипов на приспособленность. В предположении, что в каждой позиции генома существует только два аллеля, 0 и 1, адаптивный ландшафт представляет собой (N+1)-мерный бинарный гиперкуб, где N - длина генома. Изучение свойств гиперкуба интересно потому, что адаптивную динамику популяции, то есть изменение частот аллелей в сторону большей средней приспособленности, можно представить как путь по его вершинам с возрастающей на каждом шаге приспособленностью. Такое представление называется адаптивной прогулкой. Свойства адаптивных прогулок на ландшафтах, заданных различными способами, широко изучались. Типичные вопросы, которые ставились, это: средняя длина прогулки, средняя высота прогулки (то есть средняя приспособленность, на которой прогулка заканчивается), доступность глобального пика, число пиков на ландшафте. Адаптивные ландшафты можно задавать разными способами. В некоррелированном ландшафте приспособленность последовательности - это случайная величина. В NK-ландшафте вклад каждой позиции последовательности зависит от того, что стоит в K других случайно выбранных сайтах. В ландшафте под названием “холмистая гора Фудзи” есть тренд увеличения приспособленности

последовательности с ростом числа единиц в ней, что обеспечивает существование глобального пика, но в то же время у каждой приспособленности есть случайное отклонение от этого тренда, что обеспечивает появление локальных пиков.

Мы придумали новый способ задать адаптивный ландшафт, который позволяет регулировать тип эпистаза, влияющий на форму ландшафта. Это матрица попарных эпистатических взаимодействий между локусами, в которой тип взаимодействия между каждой парой локусов выбирается случайно. Главные вопросы, на которые мы хотим ответить, -- это как часто последовательность находится в области притяжения более чем одного пика и сколько из каждой точки в среднем существует одно мутационных шагов, ведущих к повышению приспособленности. Ответ на первый вопрос определяет предсказуемость исхода эволюции, а на второй -- вероятность параллельных траекторий.

## Изменение приспособленности текущего аллеля в ходе эволюции по экспериментальным данным

***А. Столярова, Ф. Кондрашов, Г. Базыкин***

Приспособленность аминокислоты, занимающей некоторую позицию в белке, может меняться в ходе эволюции. Если данный белковый сайт задействован в эпистатических взаимодействиях с другими позициями, то замены в них будут в среднем приводить к росту приспособленности текущего аллеля со временем (“акклиматизация” или “окапывание” аллеля). С другой стороны, изменения условий окружающей среды будут в среднем снижать приспособленность текущего варианта (“старение” аллеля). Сейчас изменения адаптивного ландшафта изучают по непрямым данным, например, по распределениям частот замен (в том числе реверсий и параллельных замен) вдоль филогенетического дерева. В 2018 году группа Ф. Кондрашова опубликовала результаты экспериментального измерения ландшафта приспособленности белка His3 из *S. cerevisiae*, причем в этом эксперименте были измерены прежде всего приспособленности вариантов из близких к *S. cerevisiae* видов дрожжей и их сочетаний. Около 85% исследованных вариантов были подвержены эпистазу - их приспособленность зависела от генетического окружения. Сочетания аллелей, приспособленности которых были измерены в этом эксперименте, включают в себя в том числе сочетания, которые соответствуют реконструированным предковым последовательностям (внутренним узлам филогении) и по которым можно напрямую отследить изменения приспособленностей аллелей в ходе эволюции. Мы обнаружили, что существующие данные по все же не идеально подходят для этой задачи: во-первых, исследование не было сфокусировано на реконструкции древних последовательностей, поэтому много соответствующих сочетаний аллелей остались неизмеренными; во-вторых, последовательность была поделена на сегменты, из-за чего, возможно, значительная часть эффекта была потеряна. Тем не менее, даже по этим данным мы обнаружили свидетельства “акклиматизации” текущего аллеля. Мы планируем синтезировать последовательности *His3*, соответствующие предковым вариантам, и измерить приспособленности комбинаций современных и предковых аллелей. В результате мы хотим изучить, как менялись приспособленности вариантов вдоль филогении и исследовать эволюционный путь этого гена по адаптивному ландшафту.

## Распределение аминокислот на филогеенетическом дереве как отражение однопозиционного адаптивного ландшафта.

***Галя Клинк, Георгий Базыкин***

Что мы можем сказать про однопозиционный адаптивный ландшафт аминокислотного сайта, глядя на расположение аминокислот в этом сайте на филогенетическом дереве? Мы разработали двумерную статистику (ДС), которая отражает филогенетическое распределение аминокислоты в сайте белка. Посчитав ДС для результатов симуляции эволюции при разных известных значениях относительной приспособленности аминокислот, можно определять приспособленности аминокислот в настоящих данных по значениям ДС. Также, научившись с помощью симуляций отличать распределения ДС аминокислот при постоянных и переменных однопозиционных ландшафтах приспособленности, можно искать в настоящих данных сайты, в которых аминокислоты меняют приспособленность на филогении.

Сейчас мы проводим симуляции при разных значениях относительной приспособленности аминокислот в сайтах с постоянным и переменным адаптивным ландшафтом для того, чтобы определить, в каких диапазонах этого параметра можно отличить сайты с постоянными и переменными однопозиционными ландшафтами приспособленности с помощью ДС. В последущем мы планируем использовать ДС в сочетании с нейронной сетью.

## Роль горизонтального переноса в эволюции систем рестрикции-модификации.

***Безсуднова О.И, Русинов И.С., Ершова А.С., Корягина А.С., Спирин С.А., Алексеевский А.В.***

Системы рестрикции-модификации (Р-М) защищают бактерии и археи от чужеродной ДНК. Классическая система состоит из двух белков: эндонуклеазы рестрикции (ЭР), расщепляющей неметилированную ДНК, и метилтрансферазы (МТ), метилирующей определенный нуклеотид ДНК. Известно, что системы Р-М могут эволюционировать независимо от хозяина благодаря горизонтальному переносу систем или отдельных генов. Целью данной работы служит описание эволюции систем Р-М на всем доступном материале полных геномов бактерии, а именно оценить роль горизонтального переноса в эволюции систем Р-М.

Данные о системах Р-М и их генах были получены из базы данных REBASE, содержащей наиболее полную информацию о системах Р-М. Из 4594 геномов прокариот были получены данные о 31508 системах Р-М. Системы Р-М были классифицированы по составу генов и объединены в классы по следующему правилу: две системы принадлежат одному классу, если каталитические домены этих двух систем совпадают. Было получено 230 классов систем Р-М. Каждый из полученных классов был распределен по отделам архей и бактерий, в соответствии с тем, к какой таксономической группе принадлежит геном с данной системой Р-М.

Самый распространенный класс систем Р-М - Класс N6_Mtase/ResIII, содержащий 3568 систем Р-М типа I. Данный класс распределен по 3 отделам архей и 26 отделам бактерий. Поскольку данный класс представлен в большинстве таксонов (29 из 35), было предположено,

что такая система возможно была у их общего предка. Для проверки данной гипотезы были построены филогенетические деревья МТ и ЭР каждого из отделов архей и бактерий (Деревья МТ и ЭР были построены методом Maximum Likelihood с бутстрэп 100). Далее деревья были объединены в танглеграммы, на основе которых было показано, что системы Р-М передаются с помощью горизонтального переноса, причем гены МТ и ER преимущественно передаются вместе. Также горизонтальный перенос систем Р-М возможен между разными таксономическими группами.

Следующим шагом данной работы является количественная оценка горизонтального переноса систем Р-М класса N6_Mtase/ResIII и исследование второго по распространённости класса систем Р-М - класс N6_N4_Mtase/ResIII типа III.

## Горизонтальный перенос и вертикальное наследование систем рестрикции-модификации двух гомологичных классов, включающих эндонуклеазу с доменом RE_TdeIII

***Е. А. Гусева, О. И. Безсуднова, А. В. Алексеевский***

Системы рестрикции-модификации (Р-М) архей и бактерий защищают клетку хозяина от проникновения чужеродной ДНК. Они ведут себя как мобильные элементы, хотя у них нет собственных механизмов переноса [1].

Две Р-М системы считаются гомологичными если каталитические домены их ключевых белков, эндонуклеазы рестрикции (ЭР) и ДНК метилтрансферазы (МТазы) относятся к одному семейству доменов согласно базе данных Pfam. Мы классифицировали Р-М системы на гомологичные классы используя 12426 Р-М систем из 4594 полных прокариотических геномов.

В этой работе мы представляем детальное описание эволюции двух классов гомологичных систем Р-М: (i) RE_TdeIII /N6_N4_MTase and (ii) RE_TdeIII /DNA_Methylase (названия классов образованы от названий каталитических доменов согласно Pfam). Эти два небольших класса, относящиеся к типу II систем Р-М содержат ЭР с одинаковыми Pfam доменами. Удивительно, но белки семейства RE_TdeIII были найдены только в 44 геномах (из 4594 изученных) из 15 филюмов.

Мы определили все случаи вертикального наследования и горизонтального переноса систем Р-М с помощью анализа филогенетических деревьев ЭР и МТаз. В изучаемых классах горизонтальный перенос систем Р-М преобладал над вертикальным наследованием. Также мы обнаружили случаи обмена одиночными ферментами систем Р-М МТазами и ЭР между геномами.

Работа выполнена при поддержке РСФ гранта 16-14-10319.

1. Naito, T., Kusano, K., and Kobayashi, I. Selfish behavior of restriction–modification systems // Science. 1995. No. 267. P. 897–899

## Паттерн избегания сайтов рестрикции как способ предсказания хозяина бактериофага

***A. Ershova, I. Rusinov, M. Khachaturyan, A. Karyagina, S. Spirin, A. Alexeevski***

Системы рестрикции-модификации (Р-М) — это прокариотические иммунные системы, обладающие высокой специфичностью к определенным коротким последовательностям ДНК (сайтам узнавания). Избегание сайтов узнавания систем Р-М в геномной последовательности — известная стратегия анти-рестрикции прокариотических вирусов. Набор систем Р-М с различными сайтами узнавания является видо- или даже штаммо-специфичным для бактерий и архей. Отсюда следует потенциальная возможность предсказания хозяина вируса по паттерну потенциальных сайтов узнавания систем Р-М, которые избегаются в его геноме. Определение хозяев прокариотических вирусов является важным, иногда даже ключевым, этапом во многих метагеномных исследованиях. Эта задача плохо решается компьютерными методами на данный момент.

Мы проанализировали избегание палиндромных сайтов узнавания систем Р-М типа II (только для таких сайтов было показано систематическое избегание) в 3870 полных геномов прокариотических вирусов с известным хозяином. Для анализа были отобраны только вирусы, в геномах которых избегается не менее 5 потенциальных сайтов узнавания. Было показано, что избегаются далеко не только сайты, специфичные для вида хозяина. Примерно для половины избегаемых сайтов подходящие по специфичности системы находятся только в организмах из другого таксономического отдела, или даже в более далеких прокариотах.

Однако, не все сайты узнавания систем Р-М обладают одинаковой специфичностью избегания. Так, например, сайты CCCGGG, GGGCCC, AGTACT, CTAG избегаются только в фагах далеких родственников организма-хозяина, а сайты CTGCAG, CGCG, GGCC, CCGG, наоборот, преимущественно избегаются в фагах того же вида или рода прокариот.

В дальнейшем мы планируем выделить наборы потенциальных сайтов узнавания систем Р-М, которые обладают высокой специфичностью избегания, и проверить возможность предсказания хозяев прокариотических вирусов по избеганию таких сайтов. Кроме того, мы планируем провести аналогичную работу на данных одного или нескольких метагеномов, чтобы убедиться, что полученный результат не является артефактом неправильной аннотации хозяев вирусов или сайтов узнавания систем Р-М в нашем наборе.

## Предсказание качества филогенетической реконструкции методом машинного обучения

***Иннокентий Никитин, Алексей Ефремов, Анна Ершова, Сергей Спирин***

Филогенетическое дерево, построенное по заданному множественному выравниванию, далеко не всегда точно описывает реальную филогению соответствующих белков. Можно ли по выравниванию угадать, насколько точной будет филогенетическая реконструкция? Мы

пытаемся решить эту задачу методом градиентного бустинга, используя различные признаки, полученные из выравниваний. Для обучения и тестирования используется набор из нескольких тысяч семейств ортологичных белков, взятых из организмов с известной филогенией.

На данный момент написана программа, предсказывающая расстояние до правильного дерева от реконструкции, полученной из выравниваний с фиксированным числом последовательностей (15, 20, 25, 30). Подобраны параметры для бустинга (LGBM, XGBoost).

В наших планах расширить набор признаков, используемых для обучения и изучить вопрос о том, насколько можно предсказывать качество реконструкции для выравниваний с различным числом последовательностей.

## Классификация и реконструкция филогении белковых последовательностей низкой сложности методами, не использующими выравнивания

***Научный руководитель - Алексеевский Андрей Владимирович***
***Котюргин Александр Павлович***

Последовательности низкой сложности широко распространены у различных организмов (из прикладных примеров можно вспомнить поверхностные антигены ВИЧ), при этом исследование их эволюции затруднено из-за практической невозможности построить оптимальное множественное выравнивание.

В такой ситуации логично обратиться к методам сравнения последовательностей, которые не требуют построения выравнивания. Результатом применения данных методов может быть, во-первых, классификация последовательностей в рамках заданной таксономии (по сути, молекулярный определитель), во-вторых, кластеризация последовательностей (без обязательного сближения филогенетически близких групп), в-третьих, реконструкция филогении.

На нынешний момент была рассмотрена возможность классификации последовательностей в рамках заданной таксономии различными методами машинного обучения (мультиномиальная регрессия, решающий лес) по набору частот n-меров в составе последовательности (n от 1 до 4), оказалось, что при размере обучающей выборки более 40 последовательностей на таксономическую группу средняя ошибка классификатора составляет около 3%. В то же время ни один из рассмотренных методов кластеризации (PCA, k-means, DBSCAN) не показали способности к классификации.

В будущем планируется, во-первых, рассмотреть способы сравнения последовательностей, базирующиеся на выделении общих подстрок, в качестве метода выделения низкоранговых таксонов. Во-вторых, предпринять попытку определить функцию расстояния между последовательностями на основе частот n-меров, позволяющую построение филогенетических деревьев. При успешном применении данных методов — протестировать на других последовательностях белков низкой сложности.

# Разработка алгоритма и компьютерной программы, отличающей последовательности реальных белков от ошибочно предсказанных

***Панова В.В, Спирин С.А.***

Аннотация кодирующих последовательностей в геномах не всегда правильна, что приводит к появлению в банках данных аминокислотных последовательностей, не отвечающих реальным белкам. В данной работе планируется создать компьютерную программу, способную быстро обрабатывать большое число аминокислотных последовательностей и выделять предположительно ошибочно предсказанные белки.

Основная идея работы основана на наблюдении, что в выравниваниях последовательностей реальных белков консервативные колонки имеют тенденцию группироваться вместе. Предполагается анализировать распределение консервативных колонок, сравнивая это распределение с равномерным по критерию Колмогорова. Предположительно для выравниваний реальных белков будет характерно низкое p-значение, а для выравниваний, например, формальных трансляций случайно образовавшихся длинных открытых рамок считывания — высокое.

На первом этапе технология отрабатывается на распределениях интервалов между аминокислотными остатками какого-то конкретного типа (например, пролинами) в отдельных последовательностях. Такие распределения предполагается сравнивать с экспоненциальным распределением, которое должно возникать при случайном и независимом расположении остатков данного типа. Предполагается проверить гипотезу о том, что p-значения (по критерию Колмогорова) будут систематически ниже у реальных последовательностей белков по сравнению со случайно перемешанными последовательностями.

К настоящему моменту написаны программные модули на языках C и Python, позволяющие сравнить с экспоненциальным распределением набор длин интервалов между аминокислотными остатками заданного типа во входной аминокислотной последовательности. Программы отлажены на нескольких последовательностях.

В ближайшее время планируется сравнить распределение получаемого p-значения на достоверных белках (последовательности из банка Swiss-Prot с флагом "Evidence at prtoein level"), с одной стороны, и на случайно перемешанных последовательностях тех же белков, с другой. Будут протестированы распределения интервалов между остатками всех достаточно часто встречающихся типов, прежде всего между пролинами и между глицинами — предполагается, что эти структурно важные остатки могут показать наиболее заметную разницу.

В дальнейшем предполагается написать модуль, анализирующий аналогичным образом консервативные колонки множественных выравниваний. С его помощью планируется проанализировать как семейства достоверных белков, так и наборы формальных трансляций случайных длинных ORF из родственных бактерий

# Генетика человека

## Анализ генома Эци и поиск отличий его с референсным геномом современного человека

***Научные руководители - Алексеевский А. В., Жарикова А.А.***
***Безуглов Виталий***

Целью моей работы является собрать заново геном Эци из открытых ридов, находящихся в базе данных ENA, найти и проанализировать полиморфизмы с новейшим референсом GRCh38, сравнить с ранее изданной публикацией. Авторы публикации Keller et al., 2012 собрали геном Эци, однако его нет в свободном доступе в интернет.

Эци – ледяная мумия человека, найденная в Тироле. Её возраст более 5000 лет. В 2011 году его геном был секвенирован 3 раза секвенатором AB SOLiD 4 и впоследствии были получены парные риды. Я решил изучить 3 полученные сборки ридов, картировать их на референсный геном GRCh38, найти и проанализировать полиморфизмы с данным референсом генома человека, а также выяснить, какие полиморфизмы встретились во всех трёх сборках ридов при отсеквенировании генома Эци. Это может быть очень важно, ведь это даёт нам возможность узнать о возможных генетических предрасположенностях первобытного человека, жившего 5000 лет назад, увидеть возможные мутации, иными словами, проследить историю эволюции генома человека за последние тысячелетия.

Я скачал fastq-файлы с ридами из базы данных ENA, обработал их и картировал на референсный геном. Далее с помощью специальных скриптов и программ я изучал сколько ридов было картировано на референсный геном и сколько раз. Из полученного файла с картированными ридами я выделил риды, которые легли на геном только один раз. Это было проделано, так как риды, которые легли на геном человека более одного раза, скорее всего, соответствуют тем или иным повторяющимся участкам и поэтому не являются функционально важными. Полученные риды были отсортированы, и был создан специальный файл с полиморфизмами с новейшим референсом. Эта процедура была выполнена со всеми тремя сборками ридов генома Эци. После этого с помощью специального скрипта мною были выделены полиморфизмы, которые встречаются во всех трёх экспериментах. Возможно, эти полиморфизмы являются важным отличием генома Эци от современного человека.

В дальнейшем я планирую аннотировать данные полиморфизмы по новейшим базам данных и узнать возможное функциональное значение данных полиморфизмов. Также я планирую создать описание данных полиморфизмов, гена, в котором они находятся, а также функции этого гена, предположения о роли данных полиморфизмов и причинах их отсутствия в референсном геноме современного человека. Также я собираюсь сравнить свои результаты с ранее полученными и опубликованными результатами.

## Поиск признаков ассортативного скрещивания в геноме человека

***Кузнецов И.А.*[1], *Кондрашов А.С.*[1,2]**

***[1] Laboratory of Evolutionary Genomics, A.N. Belozersky Institute of Physico-Chemical Biology of Lomonosov Moscow State Universit, Moscow, 119992, Russia.***

***[2] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109, USA.***

Генетическая структура популяции может быть следствием пространственной структуры, инбридинга, ассортативного скрещивания, имеющих место в данной популяции. В человеческих популяциях показано существование ассортативного скрещивания по таким признакам как рост, индекс массы тела и др. Кроме того, признаки данного явления наблюдаются также на уровне генома.

Предполагается, что у человека может существовать положительная ассортация при скрещивании по общему уровню здоровья, который может быть связан с количеством вредных вариантов в геноме. Среди задач было найти способ корректно оценить количество вредных аллелей в геноме, определить наличие или отсутствие корреляций между супругами по данному параметру. При этом важным аспектом остаётся учёт возможной популяционной структуры, возникающей по иным причинам.

Данная задача важна, так как её решение способно внести вклад в понимание процессов, происходящих в человеческих популяциях, и имеющих значение для демографии и здравоохранения.

На данный момент посчитаны корреляции между супругами по большому числу количественных признаков генома. Полученные значения говорят в пользу достаточно сильной генетической структурированности человеческих популяций. При оценке такой структурированности необходимо учитывать частотные характеристики рассматриваемых вариантов. Результаты метаанализа по популяциям свидетельствуют о повышенных значениях корреляции между супругами по количеству предсказанных вредных аллелей. Также наблюдается повышенная дисперсия по данному признаку относительно контрольных наборов генетических вариантов. Однако при рассмотрении отдельных популяций данная закономерность воспроизводится не всегда.

На данный момент остаётся не до конца решённым вопрос о корректности оценки количества вредных аллелей в геноме для разных популяций. Параллельно идёт работа с популяционно-генетическими симуляциями с целью воспроизвести наблюдаемую в реальных данных картину.

## Анализ роста человека как сложного генетического признака в UK Biobank

***Сергей Славский, Татьяна Шашкова, Георгий Базыкин, Юрий Аульченко***

Согласно классическим работам в области генетики и эпидемиологии, рост взрослого человека является нормально распределенной величиной, получающейся в результате суммирования воздействий многих факторов. Это вывод был многократно подтвержден в течение XX века на выборках размером в несколько тысяч индивидуумов. Однако, из социоэкономических и антропологических исследований известно, что хотя величина среднего роста разнится в разных популяциях, коэффициент вариации (отношение среднеквадратичного

отклонения к среднему) остается практически неизменным. Постоянство коэффициента вариации, то есть шкалирование дисперсии вместе со средним, может свидетельствовать о том, что рост описывается мультипликативной, а не аддитивной моделью.

Чтобы проверить предположение о мультипликативном характере роста человека, были использованы данные о 370 тысячах людей из когорты UK Biobank. В ходе анализа было обнаружено, что величина среднеквадратичного отклонения растет вместе со средним значением роста в группе (рис. 1А). При этом, увеличение среднеквадратичного отклонения не наблюдается, если в качестве объекта исследования используется логарифм роста (рис. 1Б). Аналогичная картина наблюдалась при анализе оцененных из линейной модели размеров эффекта факторов (пол и полигенный индекс), влияющих на рост: при анализе нетрансформированного роста наблюдалось значимое (adjusted R2=0.52) линейное увеличение размера эффекта вместе со средним значением роста в группе, линейный тренд исчезал (R2<0.001) при анализе логарифма роста.

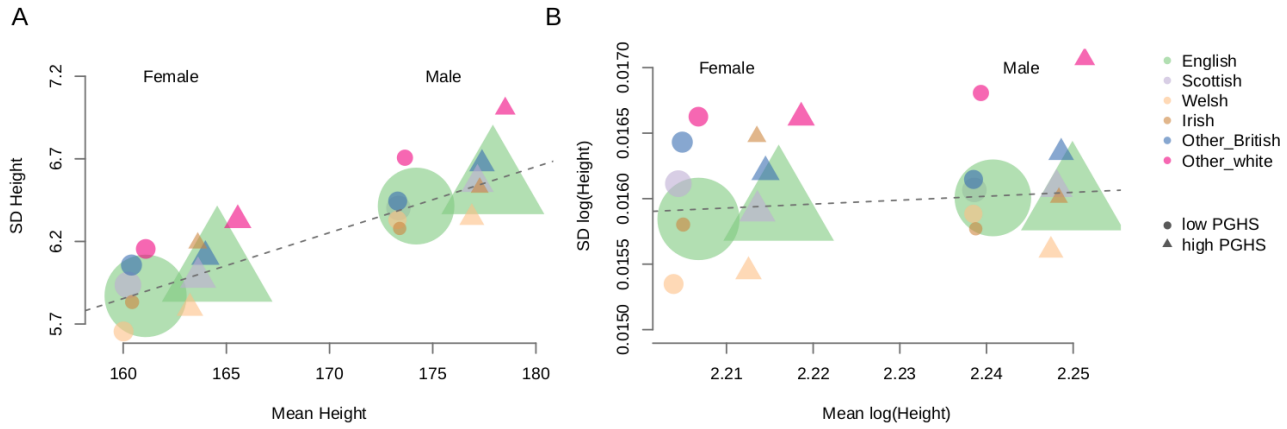



Рис 1. Зависимость среднеквадратичного отклонения от среднего роста (А) и логарифма роста (В) в шести группах индивидуумов из UK biobank, разделенных по полу и медиане полигенного индекса (всего 24 группы). Площадь фигуры пропорциональна регрессионному весу, определенному как размер группы. Взвешенная линейная регрессия была использована для оценки зависимости (A: adjusted R2=0.92; B: adjusted R2<0.02).

Помимо шкалирования размера эффектов пола и полигенного индекса, было обнаружено значимое (p-value<$10^{-13}$) взаимодействие между полом и полигенным индексом в линейной модели “Рост ~ пол + индекс + пол * индекс”. Это взаимодействие исчезало (p-value=0,9) при замене зависимой переменной в модели на логарифм роста.

Полученные результаты указывают на то, что человеческий рост является мультипликативным признаком. Это открытие имеет важное практического следствие для работ, связанных с анализом роста человека. Для уменьшения вероятности детектирования ложноположительных эффектов взаимодействий между факторами, влияющими на рост, следует анализировать рост на логарифмической шкале.

## GWAS-MAP: Анализ результатов полногеномных исследований ассоциаций с целью получение нового биологического знания

***Шашкова Т.И.*[1,2,3], *Горев Д.Г.*[1,2], *Цепилов Я.А.*[1], *Торгашева А.А.*[4], *Пахомов Е.Д.*[1], *Джоши П.*[5], *Аульченко Ю.С.*[1,4]**

*1 - Новосибирский Государственный Университет, Россия*

*2 - Московский Физико-технический Институт (ГУ), Россия*

*3 - Институт Проблем Передачи Информации им. А.А.Харкевича, РАН, Россия*

*4 - Институт Цитологии и Генетики СО РАН, Россия*

*5 - Университет Эдинбурга, Великобритания*

ПГИА (GWAS) - методология исследований, которая приобрела большую популярность за последнее десятилетие. Результаты ПГИА (РПГИА) публикуются в виде суммарных статистик -- для каждого однонуклеотидного полиморфизма (ОНП) представляется информация о частотах аллелей и характеристики эффекта аллеля на фенотип. Целью данной работы является демонстрация того, что анализ множества результатов GWAS позволяет идентифицировать биомаркеры и мишени терапевтического воздействия. Мы сосредоточились на заболевании варикозного расширения вен (ВВ), одном из клинических проявлений хронического венозного заболевания, представляющего как косметическую, так и медицинскую проблему.

Мы разработали платформу GWAS-MAP, которая позволяет интегрировать, хранить и обрабатывать результаты полногеномных исследований ассоциаций. GWAS-MAP состоит из базы данных (3,831 исследование) и трех модулей обработки данных: интеграции, контроля качества и анализа РПГИА. В платформу имплементирован широко используемые анализы: генетические корреляции, методы менделевской рандомизации, Summary-level mendelian randomization и heterogeneity in dependent instruments, Dimitrieva-Georges Theta. При помощи представленный системы нами была проведен анализ ВВ. В ходе анализа генетических корреляций были подтверждены известные положительные этиологические корреляции ВВ с ростом и весом, а также такими признаками, как тромбоз глубоких вен, курение, боль и количество перенесенных операций. Показаны отрицательные генетические корреляции ВВ с показателями интеллекта и уровнем образования. Более того, было показано, что тромбоз глубоких вен и ВВ имеют общие функциональные геномные варианты, не ассоциированы с другими вышеупомянутыми факторами. Была установлена причинно-следственная связь между ВВ и антропометрическими признаками, такими как рост, размер талии, окружность бедер и вес (как для жира, так и для мышечной массы), а также уровнями белков MICB и CD209 в плазме крови человека. Повышение значений по данным показателям ведет к увеличению риска ВВ. На основании полученных результатов, нами выдвинута гипотеза, что повышение уровней белков MICB и CD209 является фактором риска для ВВ. При этом, нами было также показано, что белок CD209 может опосредовать связь между полиморфизмом гена ABO и ВВ, а также между ВВ и группой крови А. Оба белка участвуют во врожденном и адаптивном иммунитете, что консистентно с известной ролью воспалительных процессов в развитии ВВ.

Анализ ВВ продемонстрировал потенциал системы для формулирования новых биологических гипотез, как, например, постулированная нами причинная связь между уровнями белков MICB и CD209 в плазме крови человека и ВВ.

## Филогенетический анализ B-клеточных линий человека

***Евгения Алексеева**, аспирантка первого года, Сколтех*
*Научный руководитель: Егор Базыкин*

После первичного распознавания антигена В лимфоциты получают сигнал к делению и образованию клона. В процессе нарастания численности клона, вариабельные участки иммуноглобулинов проходят через запрограммированный процесс соматических гипермутаций, в основном состоящий из нуклеотидных замен и реже из вставок и делеций.
После каждого раунда мутаций В-клетка проверяет изменившуюся аффинность к антигену и при ее увеличении получает новый сигнал к делению. Таким образом, мутации, повышающие аффинность, распространяются. В результате каждый клон В лимфоцитов неоднороден и состоит из подклонов с разными наборами соматических мутаций. С
развитием методов высокопроизводительного секвенирования появилась возможность выявлять отдельные В клеточные линии и исследовать их филогенетику. На данный момент наиболее распространенной оценкой аффинности подклона является степень его представленности в В клеточном репертуаре, однако в недавних работах было показано, что разные подклоны внутри одной линии подвержены разной силе отбора. В наши планы входит разработка метода оценки аффинности антител по силе отбора, действующему на них внутри В клеточной линии. Это позволит сравнивать эффективность линий, выработанных к разным антигенам, и может быть использовано для разработки высокоэффективных вакцин.

## Сравнение различных методов обогащения экзомов путём анализа с использованием GATK pipeline

***Жданова А.А.**, Логачёва М.Д., Набиева Е.Р., Гарушянц С.К, Базыкин Г.А.*

Экзомное секвенирование (WES), позволяющее направить анализ в сторону обогащенных таргетных последовательностей, часто применяется в медицинских исследованиях при определении взаимосвязи между наличием вариантов, отличных от генома здорового человека, и клиническими случаями.

Присутствие мутаций играет важнейшую роль в потере беременности у человека. Летальный фенотип плода может быть приобретён вследствие как унаследованных мутаций, так и de novo вариантов у родителей. При изучении случаев возникновения мутаций используется подход трио-анализа, при котором производится анализ генома либо его белок-кодирующей части обоих родителей и плода. На настоящий момент не существует оптимального протокола секвенирования экзома, применимого для определённого типа пробоподготовки, хотя существуют сравнения различных платформ полноэкзомного секвенирования (Clark et al. Nat Biotechnol. 2011 Oct; 29(10): 908–914).

В работе будет произведено сравнение трёх способов обогащения экзома трио родителей и цитогенетически нормального плода путём анализа данных полноэкзомного секвенирования. Пробоподготовка произведена путём приготовления трёх видов геномных библиотек: Illumina TruSeq, NEB Next и Agilent SureSelect. Различие библиотек состоит в

особенностях используемых адаптеров (8-нуклеотидные линейные, U-образные для первых двух), зондов (ДНК-зонды NimbleGen, Roche, Швейцария, и РНК-зонды SureSelect, Agilent, США), выборе таргетных участков (например, UTR), а также в методах их захвата.

В рамках данной работы проведена полная пробоподготовка и секвенирование на платформе Illlumina HiSeq4000 для трёх описанных протоколов. Качество обогащения библиотек определено с помощью RealTime PCR. В ходе исследования планируется определить базовые параметры секвенирования, необходимые для оценки качества анализа, после чего будут определены варианты (SNP, CNV). Статистика секвенирования свидетельствует о покрытии дедуплицированными ридами таргетных регионов, достаточном для определения вариантов в ДНК человека (>40x). Найденные мутации будут проверены методом секвенирования по Сэнгеру.

Грант: Skoltech Genomics Core Facilities, “Comparison of whole exome sequencing methods applied to trios of parents and cytogenetically normal abortuses”

## Архитектурные РНК млекопитающих

***Мыларщиков Дмитрий Евгеньевич*[1], Миронов Андрей Александрович[1,3], Шеваль Евгений Валерьевич[1,2]**

*1 – ФББ МГУ, 2 – ФХБ МГУ, 3 – ИППИ РАН*

Архитектурные РНК являются структурным и функциональным компонентом ядерных телец. Известно множество ядерных телец в различных тканях и клеточных линиях человека с предполагаемыми архитектурными РНК, но хорошо охарактеризованы лишь небольшое их число[1]. При этом предполагается, что ядерные тельца представлены у многих живых организмов, так как они играют важную роль в жизнедеятельности клетки. Таким образом, компоненты ядерных телец – белки и РНК – должны быть консервативными структурами. Однако архитектурные РНК как длинные некодирующие обладают слабой консервативностью[2]. В данной работе мы исследуем эволюцию архитектурных РНК внутри млекопитающих.

С помощью дифференциальной экспрессии по данным RNAseq линии HeLa между обычным выделением РНК и усиленным (нагревание)[3] были обнаружены несколько тысяч РНК, возможно вовлечённых в тесные взаимодействия с белками, в том числе NEAT1, некоторые miRNA. При множественном картировании данных RNAseq было показано, что РНК с короткими тандемными повторами (strRNA) PNCTR, которая участвует в образовании тельца PNC[4], проявляет дифференциальную экспрессию, как и некоторые другие strRNA. При этом были обнаружены многие ранее не аннотированные РНК, которые, возможно, взаимодействуют с белками.

В данный момент ведётся поиск гомологов предполагаемых архитектурных РНК человека у других млекопитающих по синтении окружающих участков генома с использованием базы данных OrthoDB. Планируется провести сравнительный анализ первичной и вторичной структур обнаруженных РНК, а также поиск функциональных аналогов архитектурных РНК по содержанию k-меров[5]. Предсказание РНК-белковых взаимодействий предполагается вести с помощью модели HLPI-Ensemble[6]. Достоверные

находки, обнаруженные биоинформатическими методами, планируется проверить на живых клетках методами RNAseq и FISH.

**Литература**


1. Chujo, T., & Hirose, and T. (n.d.). Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function. Moleucles and Cells, 40(12), 889–896. https://doi.org/10.14348/MOLCELLS.2017.0263;
2. Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Reports, 11(7), 1110–22. https://doi.org/10.1016/j.celrep.2015.04.023;
3. Chujo, T., Yamazaki, T., Kawaguchi, T., Kurosaka, S., Takumi, T., Nakagawa, S., & Hirose, T. (2017). Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding RNAs. The EMBO Journal, 36(10), 1447–1462. https://doi.org/10.15252/embj.201695848;
4. Yap, K., Mukhina, S., Zhang, G., Tan, J. S. C., Ong, H. S., & Makeyev, E. V. (n.d.). A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. Molecular Cell, 0(0). https://doi.org/10.1016/J.MOLCEL.2018.08.041;
5. Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., … Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. Nature Genetics, 50(10), 1474–1482. https://doi.org/10.1038/s41588-018-0207-8;
6. Huan Hu, Li Zhang, Haixin Ai, Hui Zhang, Yetian Fan, Qi Zhao & Hongsheng Liu (2018) HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy, RNA Biology, 15:6, 797-806, DOI: 10.1080/15476286.2018.1457935.


## Регуляция генов теплового шока у бактерий


***Ксения Медведева, ФКН НИУ ВШЭ***
***руководитель Илья Жаров, ИППИ РАН***


Первая часть работы посвящена реконструкции регулонов HspR в геномах бактерий. HspR является транскрипционным фактором семейства MerR. в отличие от большинства членов семейства, он подавляет, а не активирует транскрипцию. При нормальных условиях он подавляет транскрипцию генов теплового шока, а при повышенной температуре репрессия транскрипции снимается. К регулируемым генам относятся *dnaK, dnaJ, grpE, clpB, cbpA, lonA,* а также собственный ген.

В ходе работы найдено 212 потенциально регулируемых оперонов в 85 геномах бактерий. Они принадлежат прежде всего типам Actinobacteria и Epsilon-proteobacteria, а также Chloroflexi и Deinococcus-Thrermus. Геном содержит от одного до четырех регулируемых оперонов. Для Actinobacteria характерны опероны *dnaK-grpE-dnaJ-hspR* и *clpB*, для Epsilon-proteobacteria – *hrcA-grpE-dnaK, groESL, clpB* и *cbpA-hspR*, для Chloroflexi – *cbpA-hspR-(clpB),* для Deinococcus-Thrermus – *dnaK-grpE-cbpA* и *clpB*. Все указанные гены кодируют шапероны за исключением *hspR* и *clpB*, кодирующего протеазу.

В ряде случаев перед регулируемым опероном находятся два потенциальных сайта связывания HspR. Прежде всего это характерно для оперонов, содержащих гены *dnaK* и *grpE*.

Были изучены относительные координаты одиночного, проксимального и дистального сайтов. Расстояние между проксимальным и дистальным сайтом имеет вчетверо меньшую дисперсию, чем координата любого типа сайта. Среднее значение этого расстояния составляет 53,35 п.н., что соответствует 5,03 виткам спирали ДНК. Таким образом, парные сайты располагаются одинаково относительно оси двойной спирали ДНК. Это может указывать на следующий механизм репрессии транскрипции. Два димера HspR, располагающиеся на парных сайтах, контактируют между собой и изгибают молекулу ДНК. Можно предположить, что димеры HspR контактируют непосредственно между собой или посредством шаперона DnaK.

Вторая часть работы посвящена анализу встречаемости основных генов теплового шока и транскрипционных факторов, которые регулируют их экспрессию. В экспериментальных работах показано физическое взаимодействие транскрипционного фактора HspR с шаперонами DnaK (у *Mycobacterium tuberculosis* и *Streptomyces coelicolor*) и CbpA (у *Helicobacter pylori*) и транскрипционного фактора HrcA с шапероном GroE у *Bacillus subtilis*. Эти взаимодействия необходимы для нормальной работы этих регуляторных белков. Поэтому было решено исследовать, обязательно ли гены взаимодействующих белков совместно встречаются в геномах бактерий.

### Модели машинного обучения для распознавания структур стебель-петля на 3’- конце транспозонов SINE и LINE в геноме *Danio rerio*

***Ульяна Быкова, магистр 2 года АДБМ, ФКН ВШЭ***

Механизм ретротранспозиции ретротрансопозонов не до конца изучен. Известно, что 3’-конец транспозонов играет важную роль в распознавании транспозонов класса SINE и LINE, причем некоторые SINE-LINE пары обладают полностью идентичным 3’-концом, который к тому же содержит структуру-стебель петля. Мутагенный анализ показал, что нарушение структуры стебель-петля ингибирует ретротранспозицию. В настоящей работе методами машинного обучения были исследованы транспозоны рыбы *Danio Rerio*. Были построены модели машинного обучения XGBoost, распознающие 3’-концевую шпильку SINE и LINE с 85% точностью.

# Метагеномика

### Исследование бактериальных защитных систем в Мировом океане

***Марина Хачатурян, Русинов И.С., Ершова А.С., Спирин С.А., Карягина А.С., Алексеевский А.В.***

Бактерии и археи используют различные системы для защиты от бактериофагов. Наиболее известными и изученными являются CRISPR/Cas система и системы рестрикции-модификации (P-M), однако существуют и другие, для части из которых даже неизвестен механизм, но их эффективность показана экспериментально. Большинство защитных систем подвержены

сильному горизонтальному переносу, причем переносятся системы зачастую вместе в составе «защитных островов». До сих пор соотношение и устойчивые сочетания защитных систем, а также скорости их распространения, исследованы не были.

Для анализа эволюции бактериальных сообществ и распределения защитных систем был выбран проект *Tara* Ocean, содержащий наибольшее количество метагеномных данных, собранных по единым протоколам с разных глубин и по всему океану. Для данного анализа использовался бактериальный подпроект ERP001736, который содержит 29 глубоких и 106 поверхностных метагеномов.

Наши наблюдения за распределением систем P-M в океане показали, что для поверхностного слоя (до 200 метров, освещенная часть) отсутствует географическая локализация генов, что свидетельствует о быстром перемешивании поверхности океана по сравнению со скоростью горизонтального распространения генов и по сравнению с глубокими слоями, для которых, наоборот, видна идентичность отдельных метагеномов. Поэтому поверхность океана можно рассматривать как единое сообщество с точки зрения генетического анализа.

С другой стороны, хотя глубокие метагеномы собраны хуже поверхностных, в них нашлось значительно больше систем P-M и количественно (на 65%, среднее нормированное количество генов систем P-M 33 против 20) и качественно (на 40%, среднее разнообразие генов систем P-M 53 против 38), что, предположительно, является следствием гипотезы выше: постоянная «гонка вооружений» эффективна в более стабильных глубоких сообществах, в то время как на постоянно перемешивающейся поверхности «выгоднее» использовать «универсальный защитный набор». Схожая тенденция предварительно наблюдается и для систем CRISPR/Cas.

Планируется посмотреть распределение всех известных защитных систем, что даст более полную картину защитных стратегий бактериальных сообществ в различных условиях.

Работа поддержана грантом Российского Фонда Фундаментальных исследований № 18-34-00860.

## Оценка топологических свойств пространства метагеномнов микроорганизмов кишечника человека на их кластеризацию в энтеротипы

***Шатов Владислав Михайлович (Гельфанд Михаил Сергеевич)***

При анализе накопленного массива данных метагеномов микроорганизмов из кишечника человека было обнаружено, что вариации в таксономическом составе микробиоты имеют тенденцию кластеризоваться на несколько обособленных групп, названные в последствии энтеротипами. Предполагается, что энтеротипы могут иметь гендерные различия, изменяться в течение возраста, быть ассоциированными с генетическими и наследственными заболеваниями, зависеть от типа питания и образа жизни и других факторов. С клинической точки зрения энтеротипы могут представлять особый интерес, так как их анализ может быть применен при постановке диагноза, оценки факторов риска развития заболеваний и при изучении хода его развития. Кроме того, энтеротипы микробиоты могут влиять на процессы всасывания и разрушения лекарственных препаратов, изменяя из факмакокинетику или факмакодинамику.

Существование энтеротипов до сих пор вызывает сомнения, так для анализа данных метагеномов применяют метод главных компонент, редуцирующий N-мерное пространство всех образцов микробиоты в низкоразмерное пространство, доступное для визуализации и поиска кластеров. При выделении главных компонент в сложном наборе данных может возникать проблема влияния топологических свойств пространства на конечный результат анализа. Так, например, если в областях повышенной кривизны многомерного пространства собрано большое количество точек, то его редукция может автоматически привести к искусственной кластеризации. Для предотвращения такого рода ошибок необходимо понимать топологическое устройство такого рода пространств.

Цель данной работы заключается в анализе топологических свойств пространства выборок метагеномов микроорганизмов из кишечника человека и попытке разработать подходы для оценки влияния топологических неоднородностей этого пространства на конечную кластеризацию образцов.

Планы:

В ходе работы планируется собрать разнообразную базу данных анализа метагеномов микробиоты кишечника человека. Основу массива данных составляют результаты масштабных проектов Human metagenome project (HMP) и American guts project(AGP). Суммарный объем баз данных составляет порядка 10000 образцов. Кроме того, проект HMP предоставляет возможность работать с когортами образцов от пациентов с наследственными заболеваниями (болезнь Крона и др). Дополнить проект должны менее масштабные проекты, например, UKTwis project. Планируется также конвертировать результаты недавнего исследования группы Nicola Segata (2019г) по реконструкции полных геномов микроорганизмов из метагеномов более 9000 образцов. Анализ топологии пространства предполагается проводить в пространстве редуцированных координат. Планируется найти оптимальную степень уменьшения размерности пространства, достаточную для достоверной оценки кластеризации образов.

## Сравнение бактериального метагенома больных и здоровых кораллов рода Porites

***Павел Шелякин*, *Софья Гарушянц*, *Михаил Сергеевич Гельфанд***

В настоящее время по всему свету наблюдается массовая гибель коралловых рифов. Во многом это вызвано увеличением температуры океана, снижением pH воды и загрязнением вод органическими и неорганическими веществами. На фоне этих стрессовых воздействий колонии кораллов становятся более подверженными бактериальным, вирусным и прочим заболеваниям. Цель настоящей работы состоит в анализе изменения населяющего коралл бактериального сообщества при заболевании коралла и в выявлении потенциальных патогеннов.

Для этого были получены 90 парных образцов тканей (здоровая и поражённая ткань) с 45 колоний кораллов. Среди них 14 колоний с реюньонской белопятнистой болезнью, 10 с австралийской белопятнистой болезнью, по 9 с язвенной и розовополосной болезнями и 3 с чернополосной болезнью. В ДНК образцов будем секвенировать V3-V4 и V2-V3 вариабельные

участки 16S рРНК для последующего анализа метагенома. Основной анализ будем проводить с помощью пакета программ QIIME2.

В результате планируется оценить насколько схожи метагеномы у здоровых кораллов, как сильно они изменяются при разных болезнях и есть ли характерные для каждой болезни изменения.

Работа проводиться совместно с Михаилом Никитиным и Вячеславом Иваненко

## Микробиомный подход для определения профиля притока в горизонтальной нефтяной скважине

***Поздышев Арсений Станиславович***

Для определения интервала происхождения нефтенасыщенного флюида и как следствие определения профиля притока в горизонтальной скважине было решено применить новый подход, в основе которого лежат данные микробиома выбуренной породы и флюида, собранных в устье скважины при бурении и на различных этапах эксплуатации соответственно. В отличие от классических методов записи профиля притока, разрабатываемый подход позволит проводить мониторинг без погружения в скважину специальных каротажных приборов и остановки эксплуатации. В конечном итоге, разрабатываемый подход позволит: сократить время проведения работ по геолого-техническим исследованиями, снизить риски, связанные с использованием дорогостоящего оборудования и повысить эффективность работ по разработке месторождения.

Таксономический анализ образцов бурового шлама показал высокий уровень разнообразия, о чем свидетельствует среднее значение индекса Шаннона равное 5,6. Кластерный анализ представителей бактерий, населяющих выбуренную породу, показал отсутствие групп, между которыми существует статистически значимая разница. Однако качественный (unweighted) UniFrac показал, что первые два образца, отобранные при бурении до вхождения в продуктивный пласт группируются отдельно от остальных. Среди микроорганизмов, населяющих образцы выбуренной породы, наиболее часто встречаются представители следующих родов: *Halomonas, Marinobacter, Aliidiomarina, Marinobacter alkaliphilus, Pseudomonas, Belliella, Dietzia, Alcanivorax, Azoarcus.* Для некоторых из них характерна солеустойчивость и способность к деградации углеводородов.

Дальнейшие планы включают в себя: исследование микробного сообщества бурового раствора, который использовался при вскрытии пласта, сравнение микробиомов флюида, собранного на этапах эксплуатации скважины, и выбуренной породы, исключение влияния бурового раствора на полученные результаты (биоинформатически), сопоставление полученных результатов с данными проведения каротажа профиля притока, осуществленного стандартными методами.

## Сравнительный анализ микробиомов тлей

***Юлия Сарана, Павел Шелякин, Михаил Гельфанд***

Целью нашей работы является изучение и сравнение микробиомов тлей различных видов, пищевых специализаций и цветовых морф.

Микробиом тлей играет существенную роль в биологии хозяев, поставляя необходимые питательные вещества, участвуя в метаболизме токсинов, обеспечивая защиту от паразитов и неблагоприятных условий среды [1, 2]. Тли (Aphidoidea) представляют собой классическую модель для изучения взаимодействий микробиом - хозяин. Для тли единственным источником питательных веществ является флоэмный сок, богатый простыми углеводами и бедный аминокислотами. Для выживания на такой диете тли выработали устойчивые симбиотические взаимодействия с первичными симбионтами - бактериями из рода *Buchnera* [3], а также с рядом вторичных симбионтов, участвующих в развитии устойчивости к паразитам и фитотоксинам [4]. Кроме того, эти насекомые часто являются вредителями культурных растений и переносчиками инфекционных заболеваний растений и склонны вырабатывать устойчивость к инсектицидам.

На данный момент проанализированы сырые риды, полученные при полногеномном секвенировании *Aphis glycines* [5, 6]. Проанализированный датасет состоит из 21 образца тлей, относящихся к 1 или 4 биотипу (неустойчивые либо устойчивые к Rag1/Rag2 сое соответственно) из разных локалитетов. Все сырые риды после анализа качества были картированы при помощи Bowtie2 на все доступные полные референсные геномы бактерий. Были выделены наиболее представленные роды бактерий по количеству уникально выровнявшихся ридов на бактериальные геномы. После чего было проанализировано различие состава микробиомов для образцов, относящихся к разным биотипам и была произведена попытка определить таксономические группы, частота которых значимо различается между между этими группами. Достоверных различий между данными группами найти не удалось (PERMANOVA, ANCOM), микробиом больше зависит от места сбора, чем от биотипа A. glycines, что может быть связано с недостаточным покрытием при полногеномном секвенировании. Однако удалось найти определенные закономерности:

- В каждом образце было идентифицировано от 9 до 13 бактериальных родов (относительная бедность микробиома тлей согласуется с ранее опубликованными данными [7]).
- В большинстве образцов преобладающими бактериальными родами являются: *Arsenophonus, Pseudomonas, Buchnera, Stenotrophomonas* (представители этого рода могут быть ассоциированы со способностью к гидролизу неоникотиноидов [8]).
- В некоторых образцах род *Buchnera* не был самым преобладающим.
- Минорные по встречаемости роды сильно различаются между образцами.

Планируется проанализировать еще один датасет, содержащий сырые риды от полногеномного секвенирования *Acyrthosiphon pisum* (33 образца с различных кормовых растений) [9, 10]. Также планируется получить и проанализировать данные секвенирования вариабельных участков 16S рРНК (V3-V4, V4-V5) нескольких видов тлей, среди которых есть полифаги, живущие на разных кормовых растениях в природе (*Aphis pomi*) и в лаборатории (*Myzus persicae*), монофаги (*Aphis intybi, Uroleucon cichorii*), живущие на одном растении (цикорий обыкновенный) в разных географических точках и разных цветовых морф одного вида (*Aphis pomi*).

**Список литературы**

1. Oliver, K. M., Degnan, P. H., Burke, G. R., & Moran, N. A. Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. Annual review of entomology, 55, 247-266. (2010).
2. Ceja-Navarro, J. A., Vega, F. E., Karaoz, U., Hao, Z., Jenkins, S., Lim, H. C., ... & Brodie, E. L.: Gut microbiota mediate caffeine detoxification in the primary insect pest of coffee. Nature communications, 6, 7618. (2015).
3. Douglas, A. E.: Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria Buchnera. Annual review of entomology, 43(1), 17-37. (1998).
4. Guo, J., Hatt, S., He, K., Chen, J., Francis, F., & Wang, Z. Nine: facultative endosymbionts in aphids. A review. Journal of Asia-Pacific Entomology, 20(3), 794-801. (2017).
5. Jacob, A. W., Bryan J. C., Fabrice, L., J. Spencer, J., Raman, B., Ashley, D. Y., … & Andy, M.: Whole genome sequence of the soybean aphid, Aphis glycines. Insect Biochemistry and Molecular Biology. (2017).
6. https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP070881, last accessed 2018/12/22.
7. Guidolin, A.S. & Cônsoli, F.L.: Symbiont Diversity of Aphis (Toxoptera) citricidus (Hemiptera: Aphididae) as Influenced by Host Plants. Microb Ecol 73(1), 201-210 (2017).
8. Zhao, YJ., Dai, YJ., Yu, CG. et al.: Hydroxylation of thiacloprid by bacterium Stenotrophomonas maltophilia CGMCC1.1788. Biodegradation 20: 761. (2009).
9. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA255937, last accessed 2019/01/23.
10. Gouin, A., Legeai, F., Nouhaud, P., Whibley, A., Simon, J. C., & Lemaitre, C. Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads. *Heredity*, *114*(5), 494-501. (2014).

# Зоопарк

## Изучение диапаузы у яиц *Daphnia magna*

***Попов Алексей Алексеевич, Галимов Ян Рудольфович, Гусев Олег Александрович***

Диапауза - это состояние, характеризующееся торможением метаболизма и задержкой развития, которое наступает в ответ на сигналы о неблагоприятных условиях окружающей среды, такие как высыхание, замерзание и недостаток кислорода. Изучение диапаузы обуславливается необходимостью борьбы в насекомыми - вредителями сельского хозяйства. Понимание молекулярной составляющей процесса также может вести к прогрессу в области космической биологии, в частности, к новым разработкам, связанным с гибернацией живых организмов.

В последнее время ведётся охота на перспективные модельные организмы, подходящие для изучения этого процесса. Одним из таких организмов являются веслоногие ракообразные *Daphnia magna*, яйца которых способны храниться в диапаузе при заморозке несколько сотен лет.

Основной целью данного исследования является поиск и изучение молекулярных механизмов диапаузирования яиц D.magna.

Для изучения данного процесса с четырёх временных точек заморозки яиц (0, 7, 30, 90 дней) были взяты три выборки по 50 яиц *D. magna*, которые подверглись транскриптомному анализу с подсчётом дифференциальной экспрессии. На основании полученных данных можно сделать простой, но неочевидный вывод о том, что в течение заморозки в яйцах *D. magna* меняется уровень некоторых транскриптов. Возможна связь данных транскриптов с процессом выхода из диапаузы.

В дальнейшем будет произведён поиск генов, биохимических путей, в которые вовлечены их продукты, а потом и некодирующих транскриптов, которые приблизят нас к пониманию механизмов диапаузирования.

Планируется также произвести сравнение кандидатных генов в различных популяциях *D. magna* для поиска и возможности предсказания закономерностей в фенотипах их яиц и способах выхода из диапаузы.

## Популяционная геномика “спящей” хирономиды *(Polypedilum vanderplanki)*

***Базыкин Г.А., Клинк Г.В., Гарушянц С.К., Гусев О.А, Шайхутдинов Н.М.***

Эволюционный анализ геномов экстремофильных организмов является активно развивающимся направлением в эволюционной биологии, так как их уникальная способность выживать в экстремальных условиях, зачастую связана с организацией генома и определенными эволюционными приспособлениями для существования в таких условиях. В частности, рядом особенностей организации генома обладает комар-звонец (*Polypedilum vanderplanki)* из семейства *Chironomidae*, который в личиночной стадии способен переживать засуху, в состоянии полного обезвоживания. В качестве особенностей генома комара, можно назвать множественные дупликации высоко экспрессируемых (в ответ на высыхание) генов, в большинстве своем организованных в кластеры на четвертой хромосоме комариного генома; малый размер генома (120 Мб); низкий GC состав (около 28%). Изучение комара с точки зрения популяционной геномики поможет лучше взглянуть на биологию экстремофилов и их эволюцию.

Был проведен популяционный анализ данных, полученных в ходе высокопроизводительного секвенирования полных геномов объединённых особей 5 популяций в северной части Нигерии. Установлено, что популяции имеют структуру и они кластеризуются в соответствии с географической локализацией, то есть северные и южные популяции изолированы ($F_{ST}$ = 0.4-0.5). Стоит отметить, что используя общегеномное сканирование дивергенции популяций в режиме скользящих окон, была определена зона значительной генетической вариабельности в правом плече второй хромосомы, которая по данным кариологии комара подверглась большой парацентрической инверсии. Для всех популяций в данном регионе характерен практически одинаковый уровень нуклеотидного разнообразия, что может быть связано с неравновесным сцеплением генов.

Для определения модальности отбора была проанализирована группа генов, участвующая в развитии ангидробиоза и выборка генов с высоко экспрессируемыми генами в ответ на высыхание личинки комара по данным CAGE. Было обнаружено, что при наличии небольшого количества генов, которые могут находится под отбором, практически все гены из

двух выборок, находятся в режиме отрицательного отбора, так как вероятно, что возникающие нонсенс-мутации могут только снижают приспособленность комара.

В дальнейшем планируется провести сравнительный популяционный анализ другого ангидробиотического комара *(P.pembai)* из Малави, оценить отбор с помощью теста MK, а также определить наличие свипов в паралогах, задействованных в развитии ангидробиоза.

## Популяционная геномика бделлоидных коловраток вида *Adineta vaga*

***О.А. Вахрушева, Е.А. Мнацаканова, Я.Р. Галимов, Т.В. Неретина, Е.С. Герасимов, С.Г. Озерова, А.О. Залевский, И.А. Юшенова, И.Р. Архипова, А.А. Пенин, М.Д. Логачева, Г.А. Базыкин, А.С. Кондрашов***

Переходы к бесполому размножению случались многократно в эволюции эукариот, но отказ от полового размножения в большинстве случаев приводит к быстрому вымиранию. В связи с этим переход к партеногенезу обычно рассматривается как эволюционно тупиковая стратегия. Однако существование древних бесполых таксонов противоречит гипотезе о необходимости полового размножения для долгосрочного эволюционного успеха вида. Коловратки класса *Bdelloidea* считаются одним из немногочисленных примеров группы древних бесполых организмов. По всей видимости, бделлоидные коловратки отказались от классического мейоза более 10 миллионов лет назад. Бесполый статус группы бделлоидных коловраток основывается в первую очередь на том, что среди нескольких сотен тысяч проанализированных разными исследователями особей, относящихся к классу *Bdelloidea*, не было обнаружено ни одного самца. Тем не менее существует вероятность того, что какие-то формы обмена генетическим материалом у коловраток класса *Bdelloidea* все-таки существуют. Действительно, в двух работах, опубликованных в последние годы, на основании последовательностей нескольких геномных локусов было сделано заключение о возможном обмене генетическим материалом у бделлоидных коловраток. Однако результаты одной из статей были подвергнуты критике. Кроме того, полногеномные данные по полиморфизму у бделлоидных коловраток отсутствовали. Для того, чтобы убедительно ответить на вопрос о существовании обмена генетическим материалом у бделлоидных коловраток, мы отсеквенировали полные геномы 11 особей, относящихся к виду *Adineta vaga*. Анализ полученных данных показал, что структура генетической изменчивости вида *A. vaga* не совместима с облигатным партеногенезом. В частности, мы показали, что неравновесие по сцеплению между однонуклеотидными полиморфизмами в популяции *A. vaga* падает с расстоянием между сайтами. Это падение не может быть объяснено исключительно действием генной конверсии. Кроме того, среди отсеквенированных особей наблюдается значительно превышающее ожидание число триаллельных сайтов, представленных всеми тремя возможными гетерозиготными генотипами. Помимо этого на присутствие обмена генетическим материалом у *A. vaga* указывает анализ топологии филогенетических деревьев, построенных для двух гаплотипов из одного локуса. В случае облигатного бесполого размножения ожидается, что топологии, построенные для двух гаплотипов, будут совпадать. Нами было показано, что в большом количестве локусов генома *A. vaga* два гаплотипа имеют различные филогении, что противоречит гипотезе о том, что особи *A. vaga* не обмениваются генетическим материалом.

Таким образом, проведенный анализ полногеномных данных изменчивости у бделлоидных коловраток вида *A. vaga* позволяет сделать заключение о присутствии у *A. vaga* обмена генетическим материалом.

Данная работа была частично поддержана грантом РФФИ № 16-34-01303 мол_а и грантом по программе Президиума РАН “Молекулярная и клеточная биология”.

## Эволюция глобиновых локусов у рыб

***О.О. Бочкарёва, О.В. Яровая***

***Научная проблема.*** У теплокровных позвоночных α- и β- глобиновые гены расположены на разных хромосомах, организованы в домены разных типов, и механизмы их тканеспецифичной активации и репрессии различаются. α-глобиновые гены теплокровных расположены рядом с генами домашнего хозяйства и тесно с ними интегрированы –главный эритроидный энхансер α -глобиновых генов находится в соседнем с глобиновыми генами гене NPRL3. β-глобиновые гены теплокровных окружены кластеризованными генами обонятельных рецепторов и изолированы от них инсуляторами.

В геноме лучеперых рыб α - и β -глобиновые гены не сегрегированы и сосредоточены в двух общих кластерах α/β-глобиновых генов, которые расположены на разных хромосомах. Оба локуса α/β -глобиновых генов рыб синтенны локусу α-глобиновых генов теплокровных. У рыб кластер обонятельных рецепторов не содержит глобиновых генов.

Считается, что предковый кластер глобиновых генов гомологичен кластеру α/β глобиновых генов рыб и что кластер β-глобиновых генов теплокровных позвоночных произошел путем транслокации одного или нескольких глобиновых генов из предкового α/β-глобинового локуса в предсуществующий локус обонятельных рецепторов. Однако неясно, как в ходе эволюции возникли регуляторные элементы, изолирующие глобиновые гены и гены одорантных рецепторов.

***Постановка задачи.*** Описать структуру глобиновых локусов у различных рыб, чтобы найти подтверждения (или опровержения) гипотезы, согласно которой связка глобиновых генов с обонятельными рецепторами существовала намного раньше в эволюции, и один из предковых локусов глобиновых генов, схожий с минорным локусом глобиновых генов *Danio rerio*, дал начало домену β-глобиновых генов теплокровных, интегрированному в кластер обонятельных рецепторов.

***Результаты.*** В геноме *Danio rerio* в сегменте ДНК между геном *rhbdf* и первым глобиновым геном минорного локуса найден псевдоген одорантного рецептора (семейство F) на расстоянии 5,5 kb от глобинового локуса. На расстоянии 14kb от глобинового локуса находится участок, сходный с геном цитоглобина, что позволяет предположить следующую конфигурацию предкового глобинового локуса: *rhbdf* + *mpg, nprl* + α-globin-β-globin + предок гена обонятельного рецептора, имеющий трансмембранный домен + cytoglobin.

В части геномов рыб псевдогены нашлись. В частности, в геноме панцирной щуки есть только один локус, содержащий глобиновые гены. Его конфигурация соответствует главному локусу лучеперых рыб и в сегменте ДНК между генами *rhbdf и mpg* найдена последовательность, сходная с геном одорантного рецептора.

В части геномов рыб псевдогены не нашлись. Возможно, в случаях, когда гены *rhbdf, mpg* и глобинов расположены очень близко на хромосоме, такая картина может объясняться делецией соответствующих сегментов.
***Дальнейшие планы.*** Описать ситуацию во всех доступных геномах и оценить статистическую значимость наблюдений.

## Редактирование мРНК головоногих моллюсков как пример преадаптации

***Михаил Молдован, Зоя Червонцева, Михаил Гельфанд***

Под преадаптациями понимаются признаки, присутствующие в предковой популяции, изменяющие свою функцию в процессе эволюции, фактически при этом не меняясь. Впервые эта идея была высказана Чарльзом Дарвином, а развил ее в начале двадцатого века французский биолог Люсьен Куэно, который и ввел термин «преадаптация». Несмотря на то, что поначалу
преадаптация обсуждалась Куэно в терминах Ламаркизма и вызывала недоверие у пионеров эволюционной биологии, было найдено множество примеров, как анатомических признаков, так и молекулярных, попадающих под определение, данное Дарвином. Примеры преадаптаций на сегодняшний день представлены единичными генами или признаками, и нет оснований полагать, что преадаптации распространены и могут существенно влиять на эволюцию всего генома.
В своей работе мы исследуем редактирование мРНК мягкотелых головоногих моллюсков, для которых были показаны чрезвычайно большие количества сайтов редактирования “A на I” , в которых аденин заменяется на инозин. Инозин распознается всей клеточной машинерией как гуанин. Мы предположили, что редактируемый аденин может служить переходным состоянием для замены аденина на гуанин. Если сайты редактирования действительно являются преадаптациями, то, в первую очередь, частоты замен редактируемого аденина на гуанин должны быть выше, чем таковые для нередактируемых аденинов. Также известно, что у близких организмов больше общих преадаптаций, чем у далеких. Сравнивая частоты замен, мы обнаружили, что замены A на G редактируемых аденинов действительно происходят чаще, чем таковые для нередактируемых аденинов, а также этот эффект более выражен, если рассматриваются более близкие пары видов. Помимо этого, с усилением редактируемости сайта (доли транскриптов, в которых данный сайт редактируется), наблюдается дальнейшее увеличение частоты замен редактируемых аденинов на гуанины. Если рассматривать замены аденинов на пиримидиновые нуклеотиды (тимин и цитозин), то наблюдается обратная картина: редактируемые аденины меняются на пиримидины значительно реже, чем нередактируемые, и при рассмотрении более близких пар видов эффект более выражен. Более выражен он и для более редактируемых сайтов. В высокоредактируемых сайтах наблюдается и сильный положительный отбор, но только для замен на гуанин. Для замен на пиримидиновые нуклеотиды наблюдается повышенная разность с уровнем положительного отбора на замены на гуанин, с явным преобладаение последнего. Проведеннная нами экстраполяция эволюционного процесса показала, что скорость накопления гуанинов заметно замедляется, если редактируемые аденины перестают быть таковыми. Известно, что редактирование A на I производится ферментативным комплексом, распознающим специфический контекст и

локальную вторичную структуру РНК. Мы обнаружили зависимость как контекста, так и стабильности вторичной структуры от редактируемости сайта, а также усиление вторичной структуры для редактируемых аденинов, меняющихся на гуанин в других видах по сравнению с их нередактируемыми аналогами.
Таким образом, наши результаты показывают, что эволюция редактируемых аденинов заметно отличается от эволюции нередактируемых аденинов, и множество сайтов редактирования мягкотелых головоногих моллюсков действительно представляет из себя первый пример преадаптации, затрагивающей значительную часть генома.__

### Морфологическая и геномная эволюция микогетеротрофных групп растений на примере рода Thismia (Thismiaceae, Dioscoreales)

***С.В. Юдина,*** ***М.С. Нуралиев, М.Д. Логачёва***

*Thismia* – род микогетеротрофных растений из семейства Burmanniaceae, включающий около 30 видов, преимущественно описанных в тропических регионах Азии и Америки, и несколько видов из субтропических регионов Северной Америки, Японии, Новой Зеландии, Тасмании. Наибольшее видовое разнообразие приходится на Юго-Восточную Азию. Растения очень мелкие, обладают эфемерным цветком и ведут паразитический образ жизни (бесхлорофильные, облигатные паразиты грибов).
Среди нефотосинтезирующих растений наблюдается тенденция к изменениям в пластидном геноме разной степени выраженности: потеря генов, мутации, резкое уменьшение размеров. Так, по оценкам, размер генома Th. tentaculata – 16 kb. Тем не менее, геном остаётся функционален, так как сохраняется необходимость синтеза продуктов, не связанных с фотосинтезом.
Мы поставили цель провести комплексное исследование клады и составить филогенетическое дерево группы растений на основании данных морфологии и молекулярной биологии. Одна из задач – проследить редукцию пластома. На данный момент мы имеем результаты секвенирования для 9 видов, из которых предстоит собрать аннотированные геномы.

# Нейронные сети

### Использование глубинного машинного обучения для восстановления пропусков в Hi-C картах

***Алишев Наиль, НИУ ВШЭ, Анализ данных в биологии и медицине***
***Екатерина Храмеева, Сколтех***
***Александра Галицына, Сколтех***

Изучение трехмерной структуры хромосом имеет важное значение для понимания механизмов регуляции экспрессии генов, репликации и репарации ДНК, рекомбинации. Метод

Hi-C позволяет анализировать пространственную структуру хромосом с высокой точностью, однако имеет свои недостатки [1]. Один из них заключается в том, что в конечных Hi-C картах могут иметь место пропуски. Пропуск – это та область в Hi-C карте, в которой отсутствуют значения. Такие пропуски мешают эффективной работе с Hi-C картами. По этой причине было решено попробовать заполнить эти пропуски, используя знания о пропущенном сегменте, полученные из окружающих сегментов карты – иначе говоря, предсказать их значения.

В качестве предсказательной модели была выбрана модель нейронной сети, так как известно, что эта модель может «выучивать» сложные нелинейные соответствия между входными и выходными данными. В ходе работы планируется провести множество экспериментов по определению наиболее эффективного подхода предсказания пропущенных сегментов. Планируется попробовать разные наборы предикторов – тех соседних сегментов, которые будут использоваться для обучения модели. Также, планируется экспериментально выяснить оптимальную архитектуру нейронной сети (количество слоев, количество нейронов в скрытых слоях). Помимо этого, планируется определить оптимальное значение таких гиперпараметров как скорость обучения, тип градиентного спуска, тип регуляризации и так далее.

Работа над этим проектом только началась, но уже были получены первые результаты с использованием простой нейронной сети. В ближайших планах стоят задачи нормализации и «очистки» исходных данных, формирования оптимального тренировочного и тестового множеств и реализации оценки качества предсказаний.

Список литературы:

1. Lieberman-Aiden E, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, 2009, Science

## Улучшение качества результатов экспериментов Hi-C единичных клеток с помощью нейронных сетей

***Горохов Н.С.[1], Галицына А.А[2]***

*[1]) НИУ ВШЭ, Анализ данных в биологии и медицине cycloner777@yandex.ru*

*[2]) Сколтех agalitzina@gmail.com*

### Введение

В 2017 году были опубликованы результаты метода single-cell Hi-C [1], ко- торый позволил наблюдать хроматин в индивидуальных клетках. Данные отличаются от обычного Hi-C тем, что они сильно прорежены, обладают сильной вариабельностью и некоторым количеством шума. Однако некото- рые вещи от клетки к клетке остаются универсальными, например, распре- деление ТАДов, присутствие компартментов и т.п. Еще одно примечатель- ное свойство single-cell Hi-C в том, что суммирование многих клеток дает карту популяционного Hi-C.

### Постановка задачи

Целью исследования является удаление шума из карт single-cell Hi-C с помо- щью

автоэнкодера. И дальнейшего сопоставления полученных результатов с популяционными данными.

### Что сделано

Была сделана обзорная презентация на тему структуры и применимости ав- тоэнкодеров. Также были воспроизведены результаты sn Hi-C, полученные в [1]

### Список литературы


[1] Ilya Flyamer et al. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 2017.

[2] Erez Lieberman-Aiden et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009.

[3] Ian Goodfellow et al. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

[4] Carl Doersch. Tutorial on Variational Autoencoders. *arXiv e-prints*, 2016.

[5] Ian Goodfellow et al. Generative Adversarial Networks. *arXiv e-prints*, June 2014.


## Использование сверточных нейросетей для предсказания нуклеотидных последовательностей


***Анна Литвин*, Зоя Червонцева**


Представим, что мы хотим оценить, сколько информации о нуклеотиде в ДНК содержится в его нуклеотидном окружении. Один из способов решения этой задачи заключается в следующем: давайте научимся предсказывать нуклеотид по его контексту и посмотрим, насколько хорошо это получается. Кажется вероятным, что степень “информированности” соседних нуклеотидов будет разной в зависимости от того, на какой именно участок генома мы смотрим, — закодирован ли там белок-кодирующий ген, РНКовый ген, регуляторная последовательность, или это участок, не имеющий очевидной функции. Поэтому, если нам удастся получить адекватный предсказатель, дающий разные точности предсказания нуклеотидов для разных биологических объектов, мы сможем использовать его для аннотации новых функциональных участков.

Задача предсказания или восстановления значения по его окружению часто встречается в работе с изображениями. Типичными примерами таких классов задач являются super-resolution (увеличение разрешения) или задача восстановления поврежденных частей изображения. Глубокие сверточные нейросети хорошо зарекомендовали себя в этой области. Недавно предложенный метод Deep Image Prior [1], основанный на сверточных нейронный сетях типа “hourglass” с обходными соединениями, позволяет восстанавливать и улучшать качество изображений, используя только *одно* начальное изображение. Несмотря на то, что алгоритм использует минимум начальной информации, такой подход дает впечатляющие результаты. В этой работе мы хотим попробовать применить его к последовательностям ДНК.

На данный момент мы протестировали два подхода. Суть первого подхода заключается в представлении последовательности в виде картинки (кривая Гильберта) и анализе полученной картинки непосредственно предложенным методом. Этот подход показал очень слабую предсказательную мощность. Второй же подход требовал модификации кода для

одномерного случая и его применения непосредственно к последовательности. Предварительные тесты на геноме *Mycobacterium tuberculosis* показывают, что предсказание нуклеотидов обученной сетью на всем геноме сразу незначительно отличается от случайного.

Мы планируем провести полноценное тестирование второго подхода для разных классов участков генома по отдельности — кодирующих, некодирующих, интронах, экзонах, псевдогенах и т. п., а также проанализировать зависимость результатов от параметров обучения и архитектуры нейросети.

[1] D. Ulyanov, A. Vedaldi and V. Lempitsky, «Deep Image Prior» CVPR, 2018.

## Предсказание консервативности нуклеотидов в геноме кишечной палочки по контексту

***Юлия Преображенская*, *Зоя Червонцева***

Если что-то консервативно, скорее всего, оно важно. На том или ином эволюционном расстоянии консервативны гены, белковые домены, отдельные аминокислоты, в функциональных центрах ферментов. Регуляторные последовательности тоже более консервативны, чем их нефункциональное окружение. Фактически именно трактовке консервативности посвящена вся сравнительная геномика, когда из наблюдения о консервативности тех или иных участков последовательности делаются предсказания о биологических процессах. В этой работе мы попробуем предсказывать консервативность по последовательности, не используя родственные геномы, а затем будем интерпретировать результаты с точки зрения биологических процессов.

В идеальном случае, если бы мы знали всё обо всех биологических процессах, происходящих в каждом организме, мы, вероятно, смогли бы очень хорошо предсказывать консервативность каждого конкретного нуклеотида на любом эволюционном расстоянии. И наоборот, если бы кому-то удалось предъявить “черный ящик”, идеально предсказывающий степень консервативности для каждого нуклеотида, глядя только на последовательность одного никогда им не виденного генома, мы бы заключили, что этот ящик многое знает о биологических процессах. Мы хотим построить такой “черный ящик” при помощи нейронных сетей, а потом попытаться понять, на основании чего он делает свои предсказания.

В качестве обучающей выборки мы возьмем три сотни геномов кишечной палочки и будем учить сетку предсказывать для каждого нуклеотида степень его консервативности внутри вида. Степень консервативности мы определяем при помощи блочного множественного выравнивания геномов, а на вход сетке будем подавать последовательность в некотором окне вокруг интересующего нуклеотида. Мы планируем использовать сверточные нейросети, так как их просто интерпретировать.

## Предсказание вторичной структуры тРНК с помощью глубокого обучения

***З.С. Червонцева, Е.И. Григорашвили, М.С. Гельфанд***

РНК – одна из ключевых макромолекул, необходимых для существования живых организмов. Предсказание вторичной структуры РНК – важная задача биоинформатики, поскольку структура РНК определяет её функцию.

Для предсказания вторичной структуры РНК существует ряд методов, основанных на статистических и термодинамических моделях, однако их точность недостаточно высока. Улучшение существующих методов и создание новых – весьма перспективная и актуальная задача. Использование глубокого обучения может позволить выделить признаки в последовательности, неочевидно детерминирующие структуру, а следовательно – повысить точность предсказания.

Таким образом, задача проекта заключается в том, чтобы создать глубокую нейронную сеть, которая может находить такие признаки. Поскольку для обучения глубоких нейронных сетей нужны большие тренировочные выборки, для начала предполагается использовать последовательности тРНК с известной структурой.

Однако, если посмотреть глубже, эту задачу можно сформулировать по-разному.

1) Для корректного предсказания вторичной структуры могут быть важны третичные взаимодействия в молекуле РНК. Для тРНК эти взаимодействия известны, и, учитывая их, можно обучить нейросеть определять в последовательности характерные признаки, и «переводить» последовательность нуклеотидов во вторичную структуру тРНК, т.е. предсказывать её вторичную структуру.
2) Нейросеть может быть обучена разбивать последовательности на два класса: «тРНК» и «не-тРНК», т.е. предсказывать, определяет последовательность структуру, характерную для тРНК или нет.

В проекте запланирована работа над обеими задачами, однако конкретный подход определен для первой из них. Из литературы известно, что использование механизма внимания в нейронных сетях показало хорошие результаты для перевода текстов. Поскольку наша задача напоминает задачу перевода текста с одного языка на другой, представляется логичным использовать глубокую сеть с механизмом внимания для решения поставленной задачи.

## Распознавание участков H-DNA сверточными нейронными сетями (CNN)

***Бочкарева Мария, бакалавриат, ФКН ВШЭ***

Целью данной работы является построение сверточных нейронных сетей (CNN) для распознавания участков H-DNA в геноме человека, а также сравнение CNN с моделями машинного обучения, основанными на характеристиках как последовательности, так и структуры.

## Распознавание квадруплексов сверточными нейронными сетями (CNN)

***Латышев Павел, бакалавриат, ФКН ВШЭ***

Были построены модели машинного обучения, распознающие квадруплексы на основе информации о последовательности. Предсказательная сила таких моделей не превышает 80%. Целью данной работы является построение сверточных нейронных сетей (CNN) для распознавания квадруплексов в геноме человека, а также сравнение CNN с моделями машинного обучения, основанными на характеристиках как последовательности, так и структуры.

# Транспозоны

## Поиск активной генетической рекомбинации, основанной на транспозонах семейства RAG, в геномах моллюсков

***Исаев С. В.*, *Панчин Ю. В.***

V(D)J-рекомбинация — это важный механизм соматической рекомбинации ДНК, лежащий в основе адаптивного иммунитета позвоночных животных. Этот механизм обеспечивается благодаря работе белков RAG1 и RAG2, произошедших, предположительно, из транспозонов, обнаруженных как у ланцетников (*Huang et al.*, 2016), так и у более эволюционно далёких групп (*Kapitonov and Jurka*, 2005). В ранних работах, посвящённых исследованию мобильных элементов семейства RAG у моллюска *Aplysia californica* (*Panchin and Moroz*, 2008), были обнаружены их сайты рекомбинации и оценено количество таких транспозонов, однако в связи с отсутствием полного генома изучаемого организма до конца неясно, является ли представленная информация исчерпывающей.

Сейчас накопилось много геномных данных об аплизии, что позволяет более детально изучить это явление, опираясь на данные секвенирования. Путём анализа сырых прочтений можно обнаружить мотив сайтов рекомбинации, а также попытаться установить активность этих транспозонов в клетках моллюска. После проработки пайплайна описанной выше работы мы планируем применить его на остальных представителях Metazoa, у которых обнаружены белки с доменом RAG. Основной целью работы является поиск ситуаций, в которых мобильным элементом является участок ДНК, не содержащий в себе ORF, соответствующий транспозазе. Параллельной задачей в ходе работы является поиск тканеспецифичной активности изучаемой транспозиции. Для этого на модельных организмах с опубликованными результатами секвенирования РНК будет сравнен уровень экспрессии транспозаз с доменом RAG в различных тканях.

Результаты работы позволят судить о наличии или отсутствии механизмов рекомбинации, которые могут вносить разнообразие в генетический репертуар клеток беспозвоночных животных, подобно системе V(D)J-рекомбинации позвоночных.

## Модели машинного обучения для распознавания структур стебель-петля на 3’-концах транспозонов L1 и Alu в геноме человека

***Антон Заикин[1], Александр Шеин[2]***
*1 - Аспирант ФКН ВШЭ*
*2- Магистр 2 года, программа больших данных, ФБМ, ВШЭ*

Механизмы транспозиции L1 и Alu в геноме человека остаются неизвестными. Существует гипотеза, что структура стебель-петля играет существенную роль в ретротранспозиции L1 и Alu в геноме человека. Мы построили и исследовали два типа моделей, основанных на информации о последовательностях и структурных свойствах, для распознавания структур типа стебель-петля на 3`-концах L1 и Alu человека. Было построено 6 вариантов моделей попарного сравнения, распознающих структуры на конце L1 и Alu как по отдельности, так и как совместную выборку, а также отличающие 3’-концевые структуры от структур на 5’-концах транспозонов. Были обнаружили параметры, дающие наибольший вклад в распознавание: Shift, Tilt, Rise и гидрофильность.

## Модели машинного обучения для распознавания 3’-конца псевдогенов и транспозонов человека

*Воронкова Анастасия, бакалавр, ФКН ВШЭ*

Известно, что 3’-конец играет важную роль в распознавании ретротранспозонов белками LINE. Существует гипотеза, что псевдогены появляются в геноме в результате ретротранспозиции. Целью данной работы является изучение механизмов копирования псевдогенов. В данной работе поставлены задачи проверить, обладают ли псевдогены 3’-концевой шпилькой и для тех псевдогенов, которые имеют на своем 3’-конце шпильку, построить модели машинного обучения, распознающие 3’-концевую шпильку как отдельно, так и совместно со шпильками транспозонов L1 и Alu.