ОТЧЕТ
о научных исследованиях, проведенных в 2017 году с использованием вычислительной системы и деятельности ИВЦ НГУ

Работа была выполнена в рамках договор между Министерством образования и науки Российской Федерации, Федеральным государственным бюджетным учреждением науки Институтом цитологии и генетики Сибирского отделения Российской академии наук, и Рогаевым Евгением Ивановичем о выделении гранта Правительства Российской Федерации для государственной поддержки научных исследований, проводимых под руководством ведущих ученых в российских образовательных организациях высшего образования, научных учреждениях, подведомственных Федеральному агентству научных организаций, и государственных научных центрах Российской Федерации

от 28 июля 2013 г. № 14.B25.31.0033, доп. соглашение №1 от 18.02.2014 г., доп. соглашение №2 от 15.05.2015 г.

Область наук Биология

Направление научного исследования Идентификация генов, ответственных за функции мозга и патологии, на основе экспериментального исследования и биоинформатической реконструкции генных сетей нейробиологических процессов.

Целью работ, проводимых в рамках проекта, является идентификация генов, ответственных за функции мозга, связанные с различными типами поведения, а также патологиями центральной нервной системы (ЦНС), на основе экспериментального исследования и биоинформатической реконструкции генных сетей нейробиологических процессов у людей и экспериментальных животных.

Для достижения прославленной цели мы решали задачи секевенирования и сборки denovo генома лисы.

Тема работы:

Улучшение сборки скаффолдов генома лисицы

Состав научного коллектива задействованный при выполнении данной работы:

Генаев Михаил Александрович, ИЦиГ СО РАН

Ершов Никита Игоревич, ИЦиГ СО РАН

Афонников Дмитрий Аркадьевич, ИЦиГ СО РАН, НГУ.

Ранее в рамках проекта нами были определены полные геномные последовательности лис из группы животных, не подвергавшихся отбору на конкретный поведенческий фенотип. В 2017 году нами была проведена работа по улучшению сборки генома лисы за счет более точной сборки скаффолдов из контигов, полученных на первом этапе выполнения проекта. Сборка контигов в скаффолды осуществлялась за счет использования информации о

транскриптоме (библиотеки RNA-seq), парных чтениях с короткими (библиотеки Paired-End) и длинными (библиотеки Mate-Pair) вставками, а также синтении с близкородственным геномом (*Canis familiaris*). Блок-схема анализа приведена на рисунке 1.

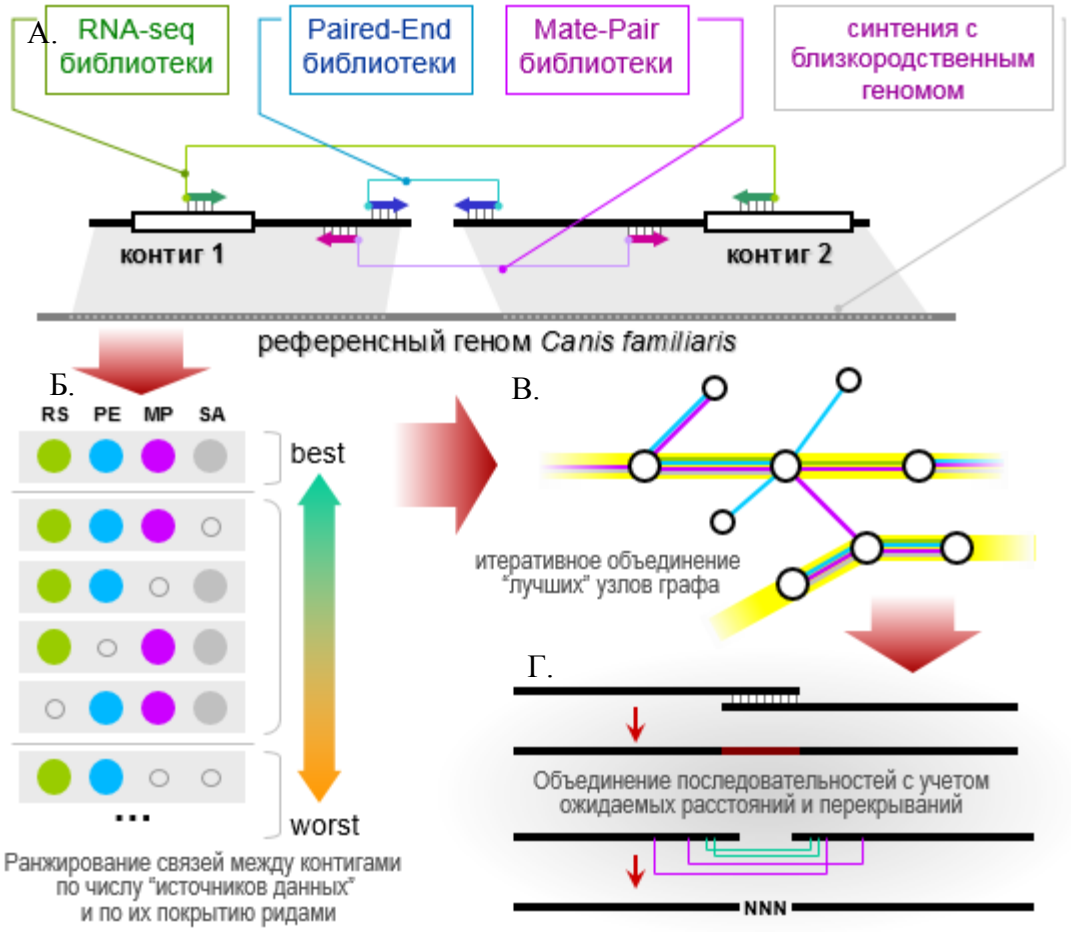

Рисунок 1. Блок-схема первого этапа улучшения качества сборки скаффолдов за счет использование библиотек с длинными вставками, последовательностей транскриптома и выравнивания последовательностей контигов с геномными последовательностями млекопитающих.

Алгоритм улучшения содержал два этапа. На первом этапе использовались геномные и транскриптомные библиотеки лис, полученные в результате выполнения проекта.

(1) Картирование последовательностей транскриптомов и геномных библиотек на контиги генома дикой лисы (рисунок 1А).

Было осуществлено выравнивание последовательностей транскриптов, полученных из библиотек транскриптома лис и получен список контигов, которые содержали выравнивания фрагментов на общие последовательности транскриптов лисы (для выравнивания использовалась программа blat (Kent, 2002)). Было осуществлено картирование фрагментов из библиотек с короткой вставкой (Paired-End) и определены контиги, на которые картируются парные риды из одного фрагмента ДНК. Было осуществлено картирование фрагментов из библиотек с длинной вставкой (Mate-Pair) и определены контиги, на которые картируются

парные риды из одного фрагмента ДНК. Картирование фрагментов ДНК-библиотек осуществлялось при помощи программы bowtie2 (Langmead & Salzberg, 2012)).

(2) Вычисление веса связи между парами контигов на основе частот перекрываний с фрагментами ДНК и РНК, полученными на этапе 1 (рисунок 1 Б).

(3) Построение и анализ графа Де-Бройна для контигов на основе весов, полученных в п. 2 (рисунок 1 В).

(4) Объединение контигов в скаффолды на основе графа связей (рисунок 1 Г). Работа алгоритма на этапах анализа 2-4 была реализована с помощью программ на языках Perl и C++, разработанных участниками проекта.

На втором этапе для улучшения сборки контигов нами был использован алгоритм учета синтениии между геномами лисы и геномами собаки (Kirkness et al., 2003), хорька (Peng et al., 2014), кошки (Pontius et al., 2007) и панды (Li et al., 2010). Ниже описаны основные шаги этого этапа улучшения геномной сборки.

(1) Выравнивание геномных последовательностей для двух пар организмов. На этом этапе происходит выравнивание всех видов, включая лису на геном собаки. Для выравнивания использовалась программа NUCmer, которая является частью программного пакета MUMmer 3.1 (Delcher et al., 2003). Выравнивание осуществлялось только на консервативные районы генома собаки, для этого использовалась опция –mumreference. Результатом выполнения этого шага являются файлы парных выравниваний геномных последовательностей в формате *.delta: dog_vs_fox.delta, dog_vs_cat.delta, dog_vs_ferret.delta, dog_vs_panda.delta в формате delta.

(2) Фильтрация выравниваний. Для анализа синтении геномных последовательностей были выбраны выравнивания длиной больше 300 п.н. имеющие свыше 60% совпадений символов. Результатом работы этого шага стал набор файлов локализации протяженных выравниваний в парах геномов в формате *.bed: dog_vs_fox.bed, dog_vs_cat.bed, dog_vs_ferret.bed, dog_vs_panda.bed в формате bed.

(3) Идентификация парных синтенных фрагментов в геноме лисы относительно остальных геномов. На этом шаге осуществляется поиск пересечений выравниваний между последовательностями сборки генома дикой лисы и последовательностью генома любого другого (или с несколькими) вида. Таким образом определялись консервативные фрагменты геномных последовательнгостей, присутствующие одновременно в нескольких геномах и перекрывающиеся с контигами лисы. Поиск осуществлялся с помощью программы bedtools v2.24.0 (Quinlan et al., 2010). На этом этапе при помощи разработанных нами скриптов формировались цепочки контигов лисы вида:

ctg1(ctg_len1) <distance> ctg2(ctg_len2),

где ctg1,ctg2 – идентификаторы контигов лисы; ctg_len1,ctg_len2 – длины контигов лисы; distance – расстояние между контигами, вычисленное на основе выравнивания (отбирались пары контигов с расстояниями между ними не более 10000 п.н.). Результатом этапа является формирование файла chain.txt.

(4) Объединение контигов в скаффолды. На этом этапе проводили построение ориентированного взвешенного графа, из цепочек выравненных контигов с учетом их взаимного расположения в скаффолдах. В этом графе узлами являлись контиги, а вес ребра

определялся параметром distance. Далее для этого графа производился поиск максимального пути, состоящего из наибольшего количества ребер (в случае, если число ребер одинаково выбирается путь, в котором суммарная длина контигов больше). Результатом этапа является файл paths.txt в который записываются цепочки контигов для объединения и расстояние между контигами, полученное на основании выравнивания. Если расстояние между контигами положительное, то между контигами вставляется последовательность символов N, причем длина последовательности равна этому расстоянию. Если расстояние между контигами отризательное (перекрывание выравниваний), то между ними вставляется единичный N символ.

1.3 Анализ улучшенной сборки генома лисы

Анализ сборки генома был проведен с помощью программы quast (Gurevich et al., 2013). Результаты представлены в Таблице 1.

Таблица 1. Характеристики сборок генома дикой лисы после выполнения улучшения с использованием геномов других организмов.

Сборка	Контиги	Скаффолды (RNA-seq, PE, MP)	Скаффолды (синтения)
Число последовательностей	367672	134684	101343
Число последовательностей более 1000 п.н.	273180	66977	53810
Полная длина	2572959176	2777174993	2794078319
Полная длина последовательностей более 1000 п.н.	2509224869	2730080474	2761300908
Длина максимального фрагмента	315380	5494015	7345328
GC (%)	41.18	41.18	41.18
N50	17944	474350	607908
N75	7757	172122	234409
L50	36581	1400	1145
L75	91057	3824	2993
Число N на 100 тыс. п.н.	0.00	7361.91	7952.35

Для полученной сборки генома был проведен поиск генов с помощью программы

FGENESH+ (Solovyev et al., 2006). Всего нами было идентифицировано 70576 предсказанных генов в последовательностях контигов. Это превышает почти в два раза число генов, известных для генома собаки. Следует отметить, что часть предсказанных генов оказались фрагментированы.

С помощью программы BLAST (Altschul et al., 1997) мы провели поиск ближайших гомологов в аминокислотных последовательностях генома собаки (аннотация БД Ensembl, http://www.ensembl.org/Canis_familiaris/Info/Index) для предсказанных генов в геноме лисы. Всего нами было найдены гомологи для 52849 предсказанных генов.

Заключение

Таким образом, в результате выполнения проекта нами получена улучшенная сборка генома лисы, имеющая характеристику N50 для контигов 17944, для скаффолдов 607908. Проведено предсказание генов в этой сборке и установлены гомологи этих генов для генома собаки. Полученные результаты позволили в дальнейшем использовать эту сборку для геномных исследований и анализа транскриптомов лис.