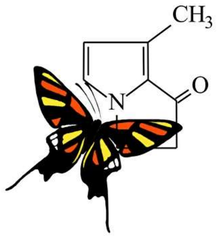
Разработка компьютерного метода классификации днРНК *A. thaliana* L

•

Выполнил: Москаленко Н. О., студент 17410 группы ФЕН НГУ, кафедра информационной биологии

Руководитель: Генаев М. А. к.б.н., н.с. лаборатории эволюционной биоинформатики и теоретической генетики

Новосибирский государственный университет
*НАСТОЯЩАЯ НАУКА



ИЦГ

2021

Длинные некодирующие РНК

ДнРНК – это транскрипты длиной более 200 нуклеотидов неспособные кодировать белки.

Основные характеристики: подвергаются полиаденилированию и сплайсингу, альтернативному 3′-концевому процессингу; слабоконсервативны; локализуются в ядре, цитоплазме, менее митохондриях; транскрибуются у растений РНК-полимеразой IV и V;

Классификация днРНК по расположению в геноме:

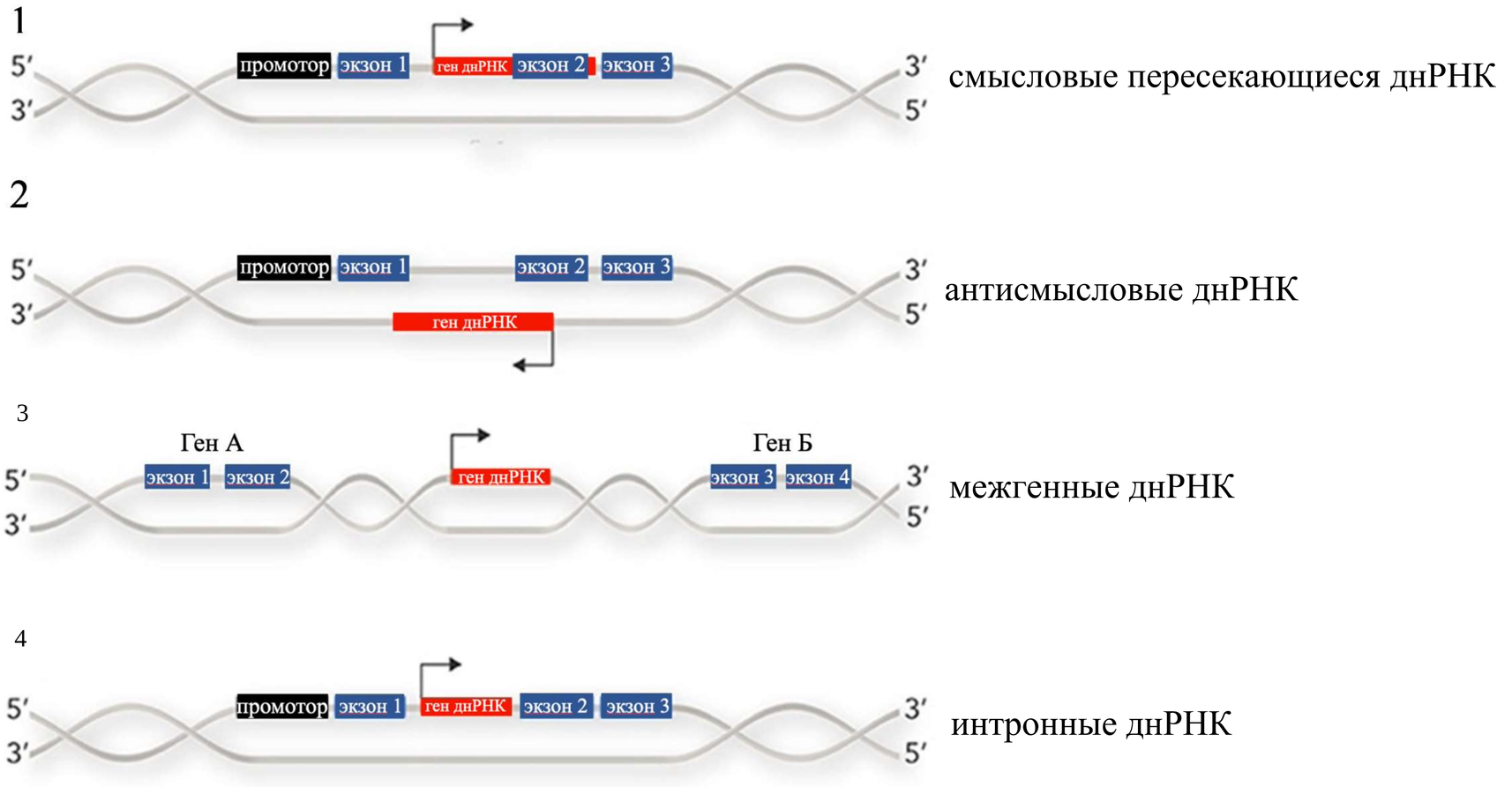


Рисунок 1 — Классификация днРНК

Функции днРНК:

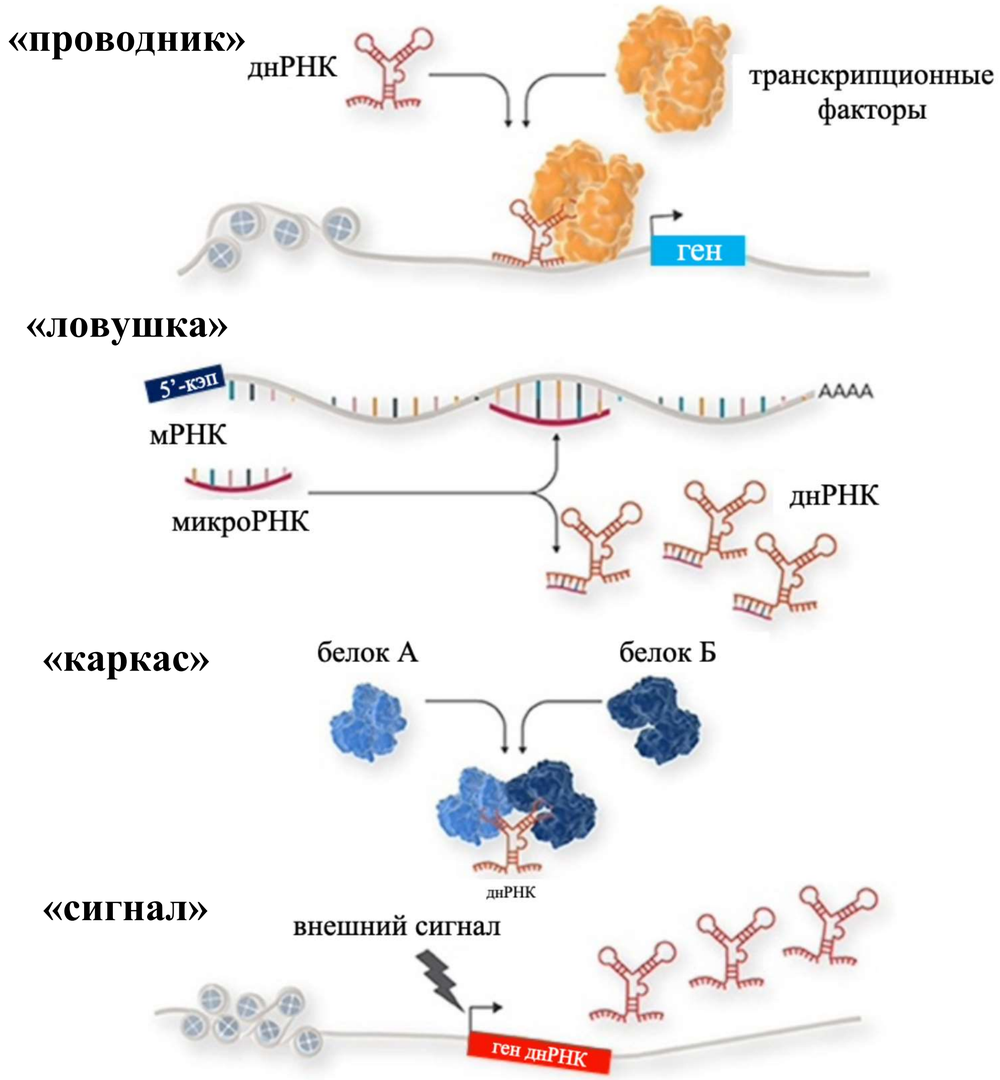


Рисунок 2 — Молекулярные функции днРНК

днРНК участвуют в процессах:

- подавление экспрессии генов,
- органогенез в корнях растений,
- фосфатный гомеостаз,
- развитие ответной реакции на стресс,
- регуляция цветения, фотоморфогенез, размножение.

Модельный организм — *Arabidopsis thaliana* L.

Основные характеристики: отдел Покрытосеменные, класс Двудольные, семейство Капустные, не имеет агрономического значения; длина генома около 157 млн пар нуклеотидов, 5 хромосом, около 27416 генов, кодирующих примерно 35000 белков;

Рисунок 3 — Внешний вид *A. thaliana*.

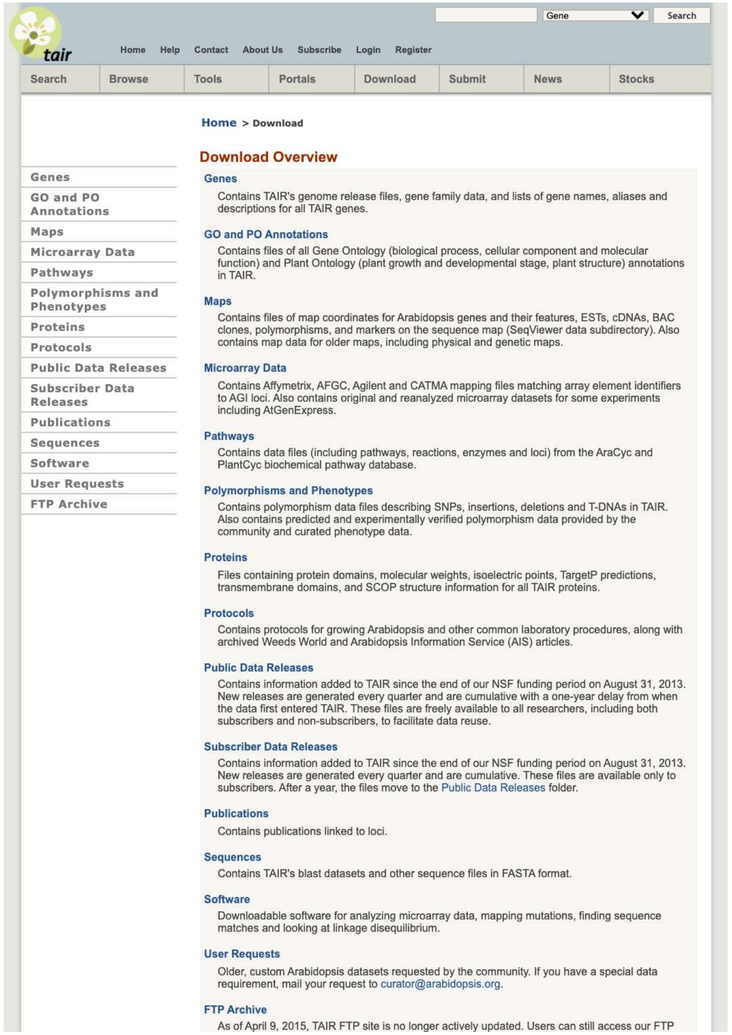


Рисунок 4 — The Arabidopsis Information Resource (www.arabidopsis.org).

<u>Удобный модельный организм, так как:</u>

- имеет полностью секвенированный и наиболее аннотированный геном,
- существует большое количество баз данных и информационных ресурсов с геномными последовательностями.

TAIR (The Arabidopsis Information Resource) — информационный ресурс по арабидопсису, поддерживающий базу данных генетических и молекулярно-биологических данных, а именно: полная последовательность генома, структура генов, экспрессия генов, информацию о ДНК, белках и другие. Данные обновляются каждую неделю на основе последних опубликованных исследований.

Актуальность

На текущий момент в базе данных EnsemblPlants для *A. thaliana* идентифицировано 3481 последовательность днРНК, но кроме этого о них больше ничего не известно, они попрежнему не классифицированы и следовательно функционально не аннотированы.

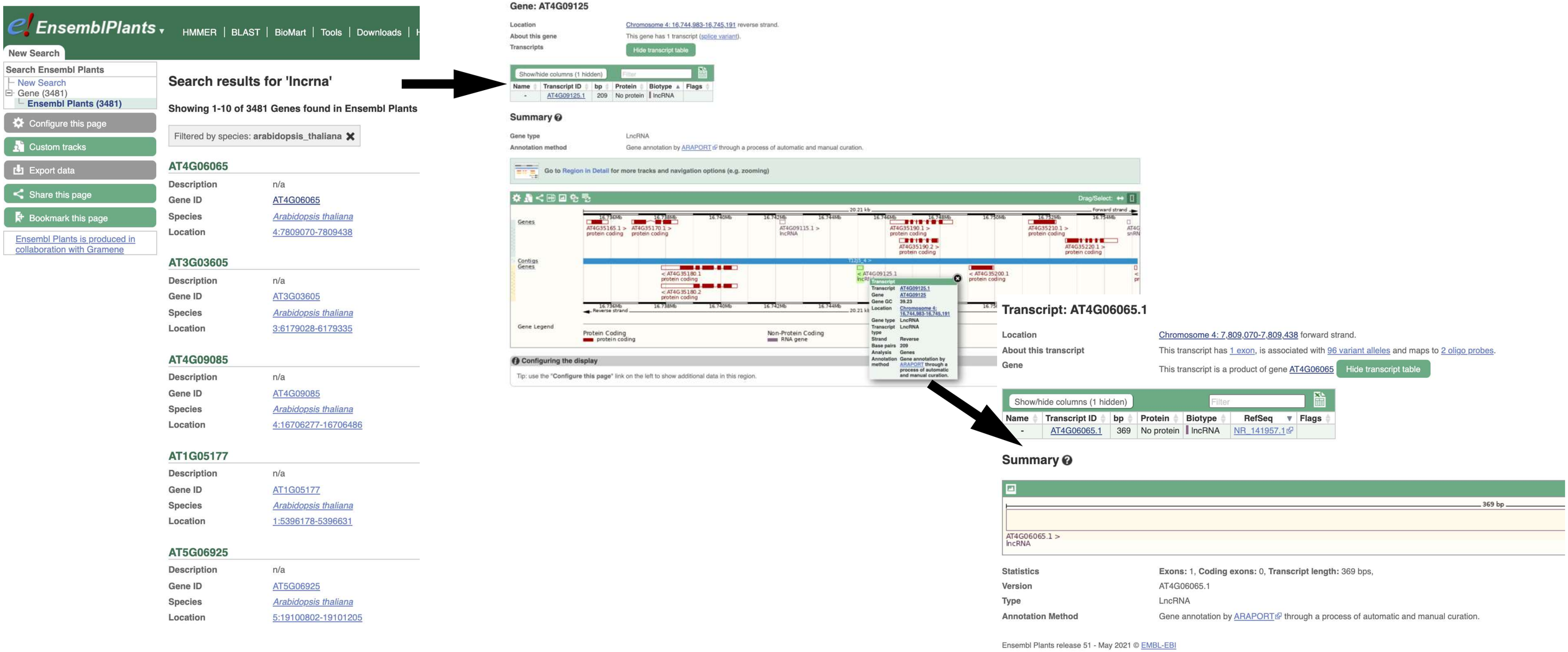


Рисунок 5 — Доступная информация в EnsemblPlants о днРНК *A. thaliana*.

Цель и задачи

Цель работы: Создание компьютерного метода классификации днРНК и аннотация днРНК *A. thaliana* L.

Задачи:

1. На основе анализа литературы охарактеризовать основные структурные и функциональные особенности основных классов днРНК
2. Составить биоинформатический конвейер по классификации днРНК
3. Разработка метода классификации днРНК по локализации их в геноме
4. Выявление особенностей выравнивания днРНК на белок кодирующие гены
5. Статистический анализ особенностей геномной локализации днРНК

Материалы

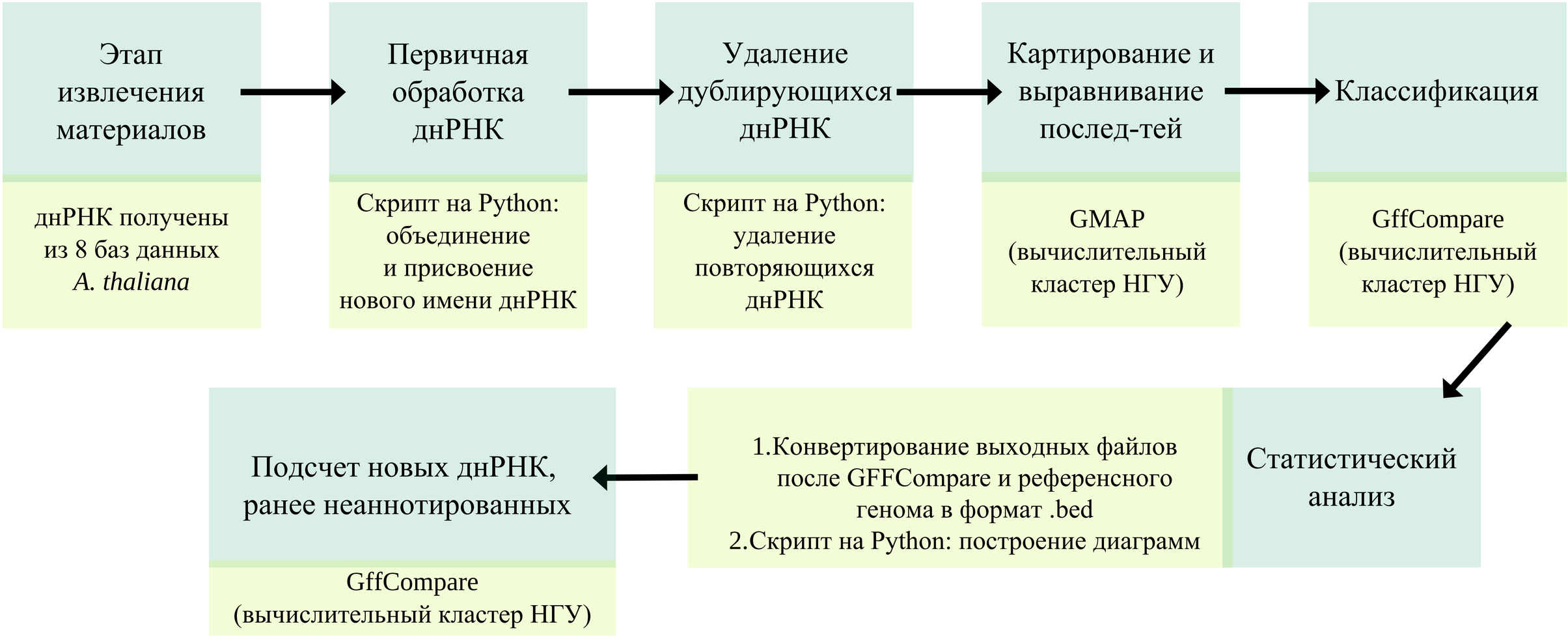
- Референсный геном — (TAIR10), содержит 27655 генов, кодирующих белки; 5178 генов, некодирующих белки; 3481 генов днРНК; размер 119688634 п.н. (135 Mb); дата сборки 15.03.2018
- ДнРНК из открытых баз данных в общем количестве 35390.

Название базы данных	**Информация о ресурсе**	**Количество днРНК *A. thaliana***
PlancDB	Комплексная функциональная база данных днРНК растений. Содержит 1246371 последовательности, идентифицированных с помощью RNA-seq, а также другую информацию об органоспецифической экспрессии и другие биологические данные.	13455
GREENC	Содержит информацию об 120000 последовательностях днРНК 37 видов растений, в том числе их координаты в геноме, кодирующий потенциал и т.д	3008
PNRD	Содержит записи не только о днРНК, но и о других РНК 150 видов растений. Имеет собственный геномный браузер.	2597
Cantata db 2.0	содержит > 45 000 днРНК десяти модельных видов растений и информацию о тканеспецифической экспрессии, кодирующем потенциале, также последовательность оценивается на основе потенциальной роли в регуляции сплайсинга.	4373
EVLncRNA	Высококачественная база данных, в которой вручную собраны опубликованные и экспериментально подтвержденных днРНК. Всего 1543 последовательности для 77 видов.	144
RNA central	Интегрирующая база данных, использует последовательности нескольких баз и обеспечивает текстовый поиск.	3884
NCBI	Национальный центр биотехнологической информации, универсальный портал с большим набором данных и множеством ресурсов.	3883
NONCODE	Включает информацию о > 500 000 днРНК 16 видов. Arabidopsis - единственный вид растений, представленный в данной базе данных, так как NONCODE фокусируется в основном на днРНК человека и мышей.	4046

Таблица 1 — Список использованных баз данных днРНК

Методы

Биоинформатический конвейер



Результаты

Удаление дубликаций / GMAP

Количество дублирующихся последовательностей днРНК равно 12716

Название базы данных	До фильтрации	После фильтрации
PlancDB	13455	9552
GREENC	3008	1852
PNRD	2597	2307
Cantata db 2.0	4373	2231
EVLncRNA	144	86
RNA central	3884	3884
NCBI	3883	0
NONCODE	4046	2762
Суммарное количество днРНК	**35390**	**22674**

GMAP: из 22674 последовательностей на референсный геном TAIR10 было выровнено 22673 (кроме одной днРНК), это можно объяснить погрешностью

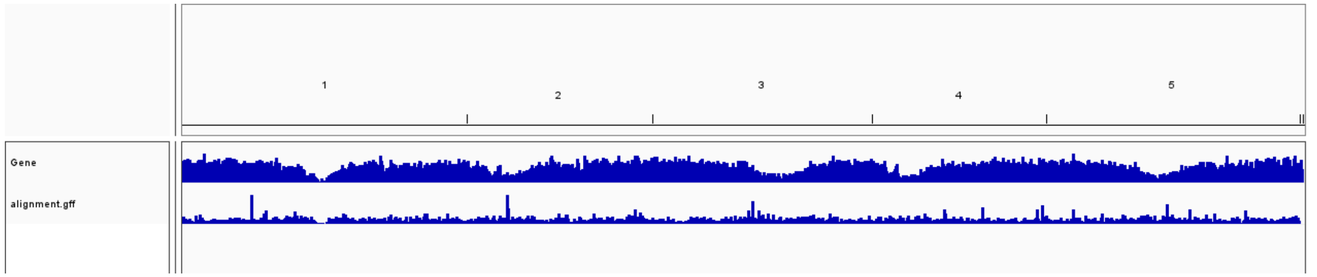
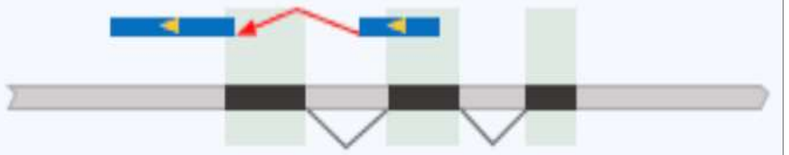


Рисунок 5 — Визуализация выравнивания при помощи IGV-браузера

GffCompare

Использовались: файл после выравнивания в GMAP в формате .gff и референсный геном в формате .gff, с предварительно были удаленными днРНК, посредством Python.

Индекс	Значение	Интерпритация	Схематичное изображение
s	19	Антисмысловые интронные днРНК	
x	5392	Антисмысловые экзонные днРНК	
i	166	Полностью содержится в референсном интроне	
u	10956	Межгенные днРНК	

Межгенные днРНК (индекс «u») — 10956 последовательностей; 43,3%,

Антисмысловые днРНК — (индексы «x» и «s») — 5411 последовательностей; 23,9%,

Интронные днРНК — (индекс «i») — 166 последовательностей; 0,73%;

Таблица 3 — Количество и значение индексов в GffCompare (черным — референсная последовательность, синим — классифицируемые последовательности, закрашенная область — повторяющиеся области в геноме)

Статистический анализ

Для каждой из 22673 днРНК была посчитана длина экзонов и интронов, составлено распределение, визуализированное при помощи следующих графиков:

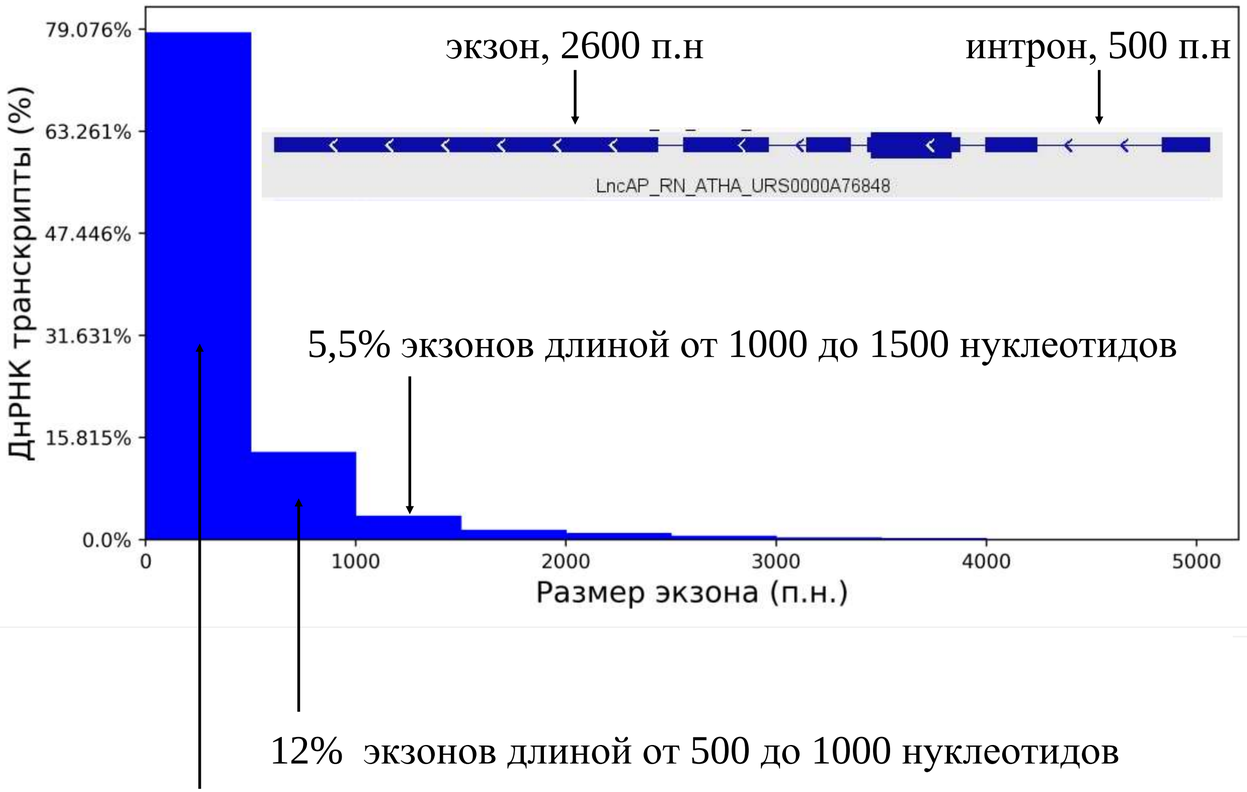


Рисунок 5 — Распределение днРНК по размеру экзона

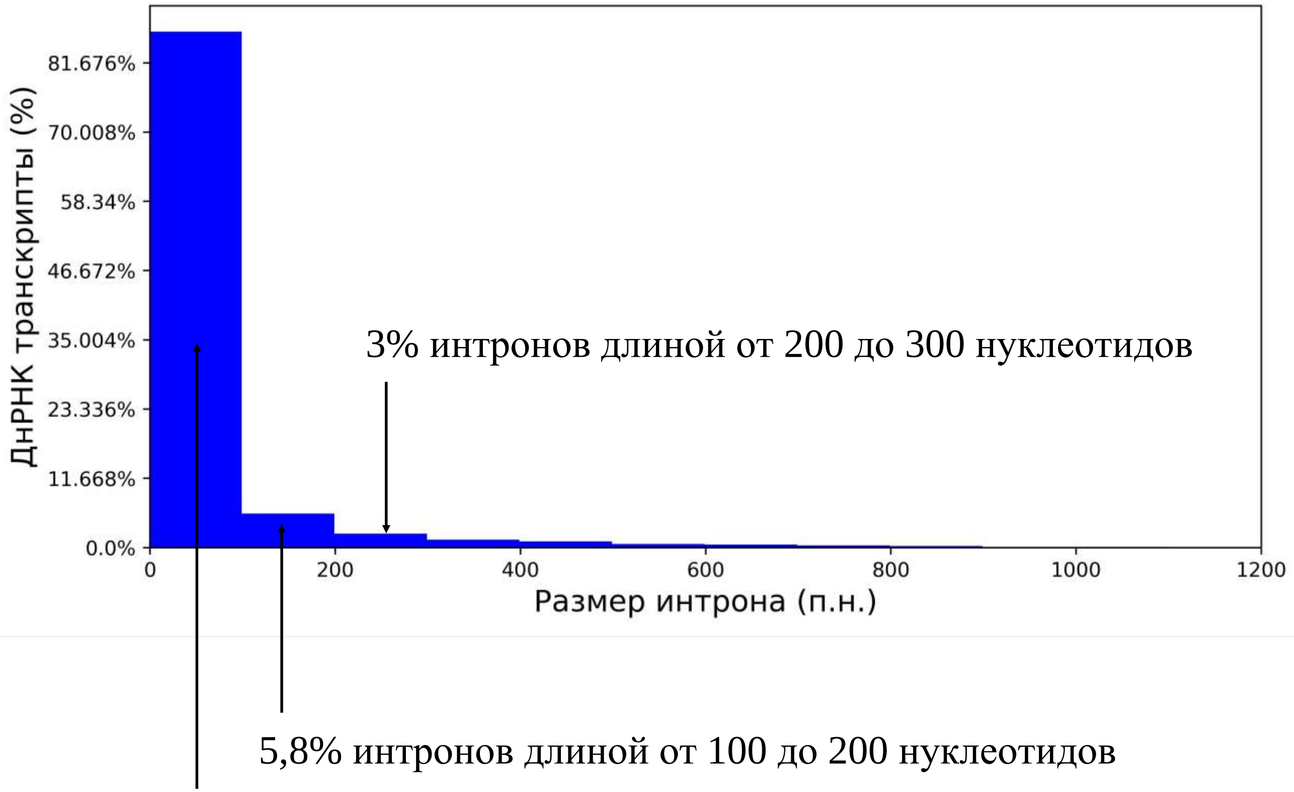


Рисунок 6 — Распределение днРНК по размеру интрона

Для каждой из 22673 днРНК была посчитана количество экзонов и составлено распределени;

Также было определено количество днРНК для каждой из хромосом и посчитана соответствующая плотность, в зависимости от длины.

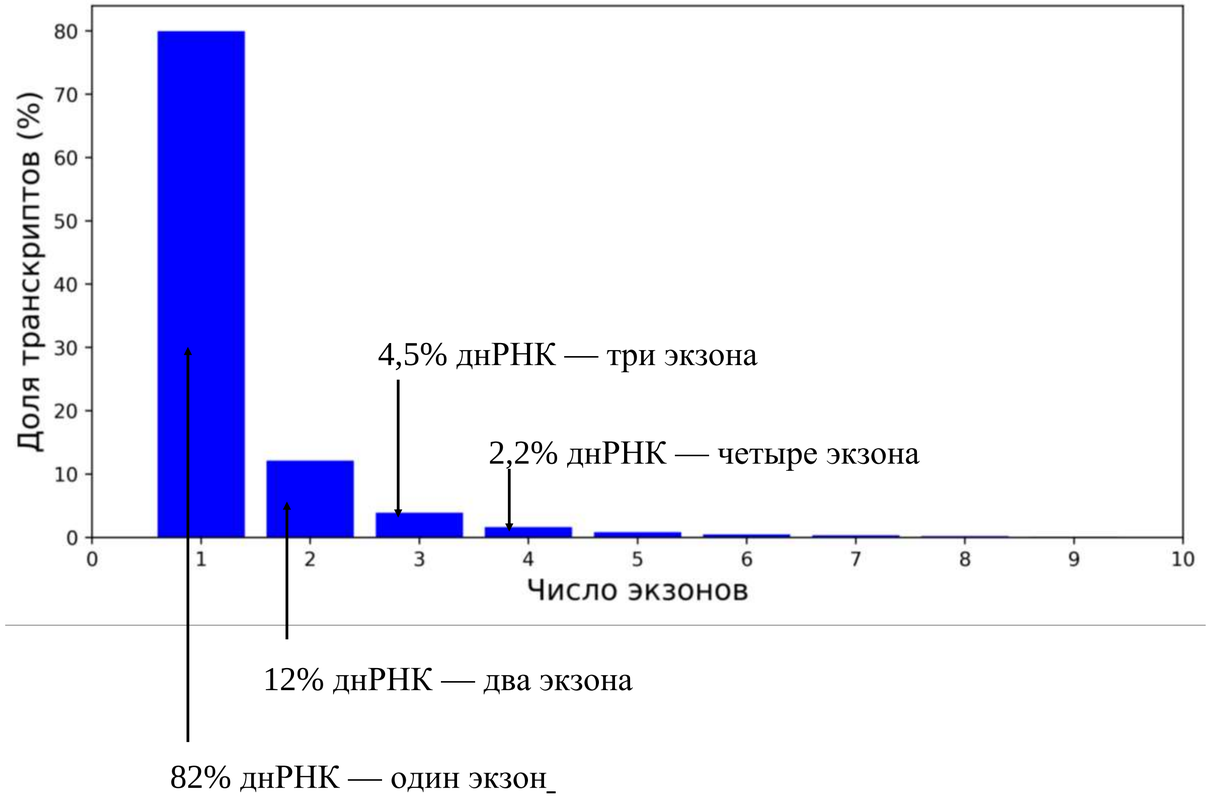


Рисунок 7 — Доля днРНК с определенным количеством экзонов

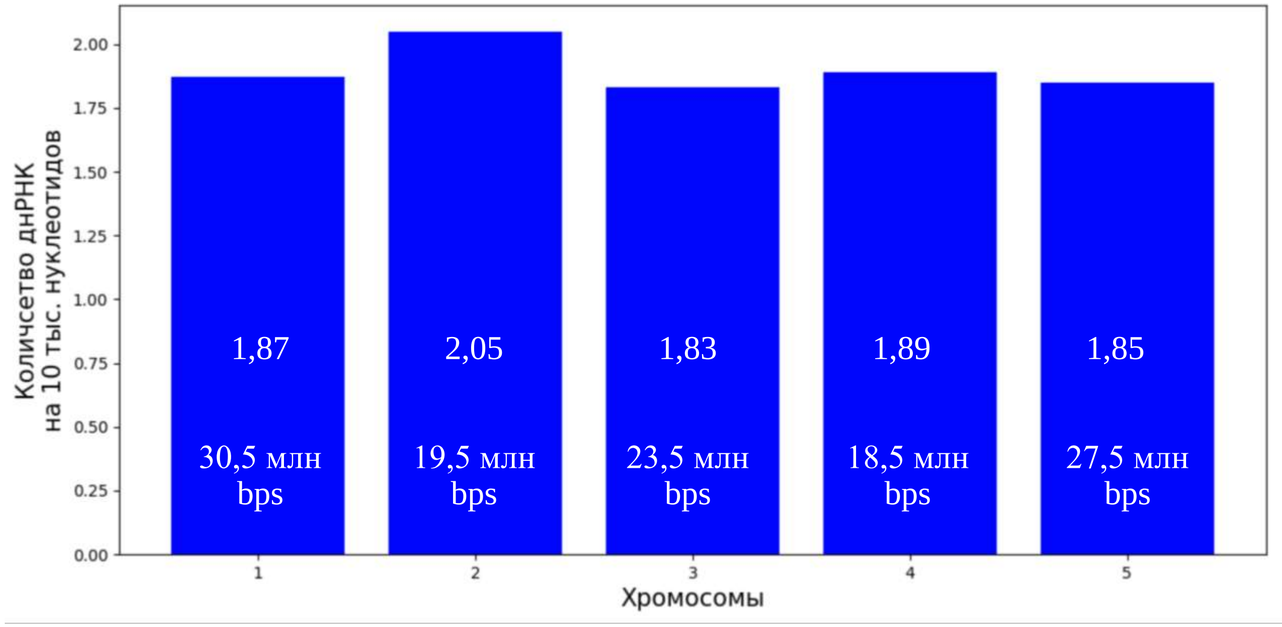


Рисунок 8 — Плотность днРНК на каждой хромосоме

На данном графике показаны только 33 антисмысловых днРНК, которые полностью совпадают с каким-либо экзоном на референсном геноме

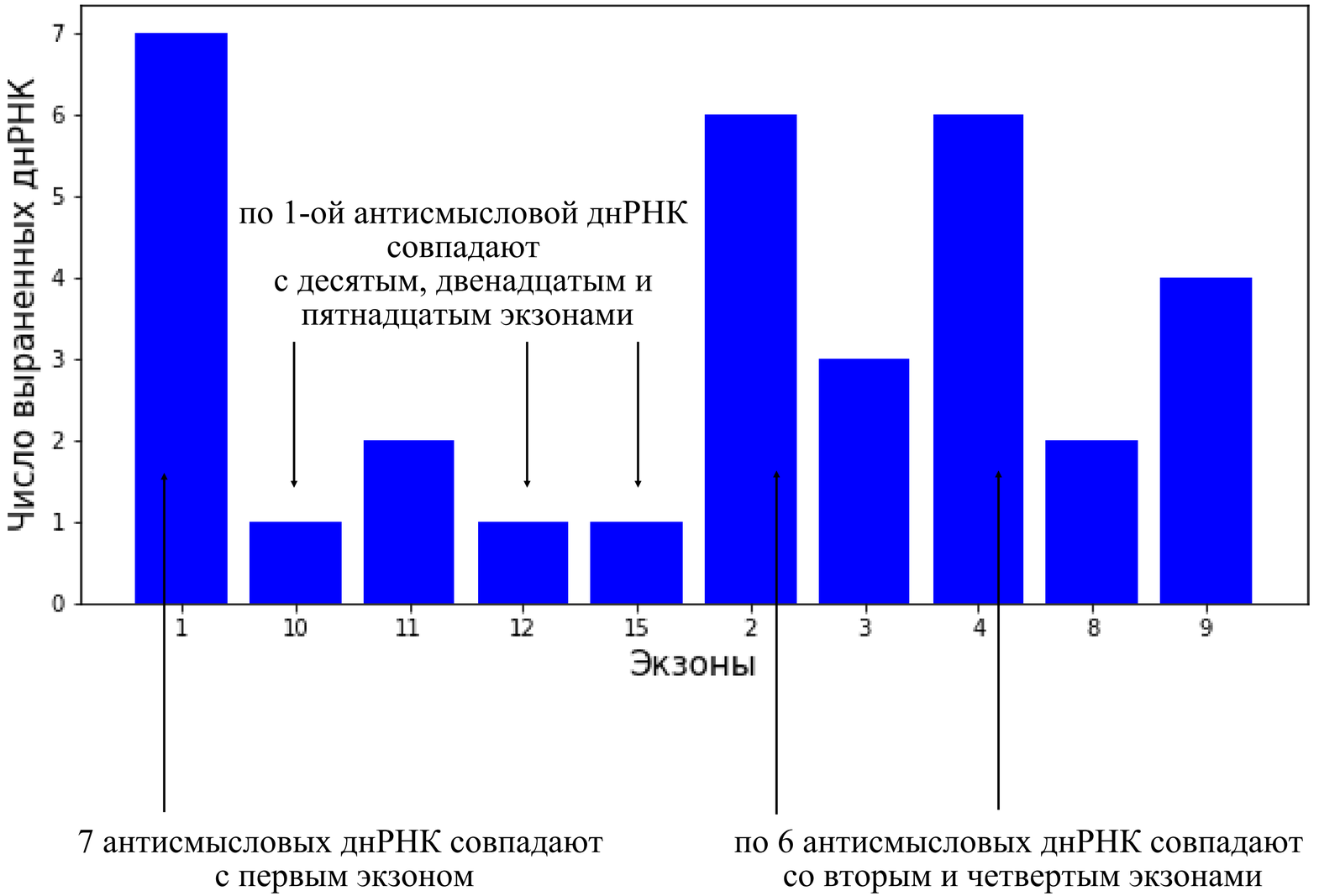
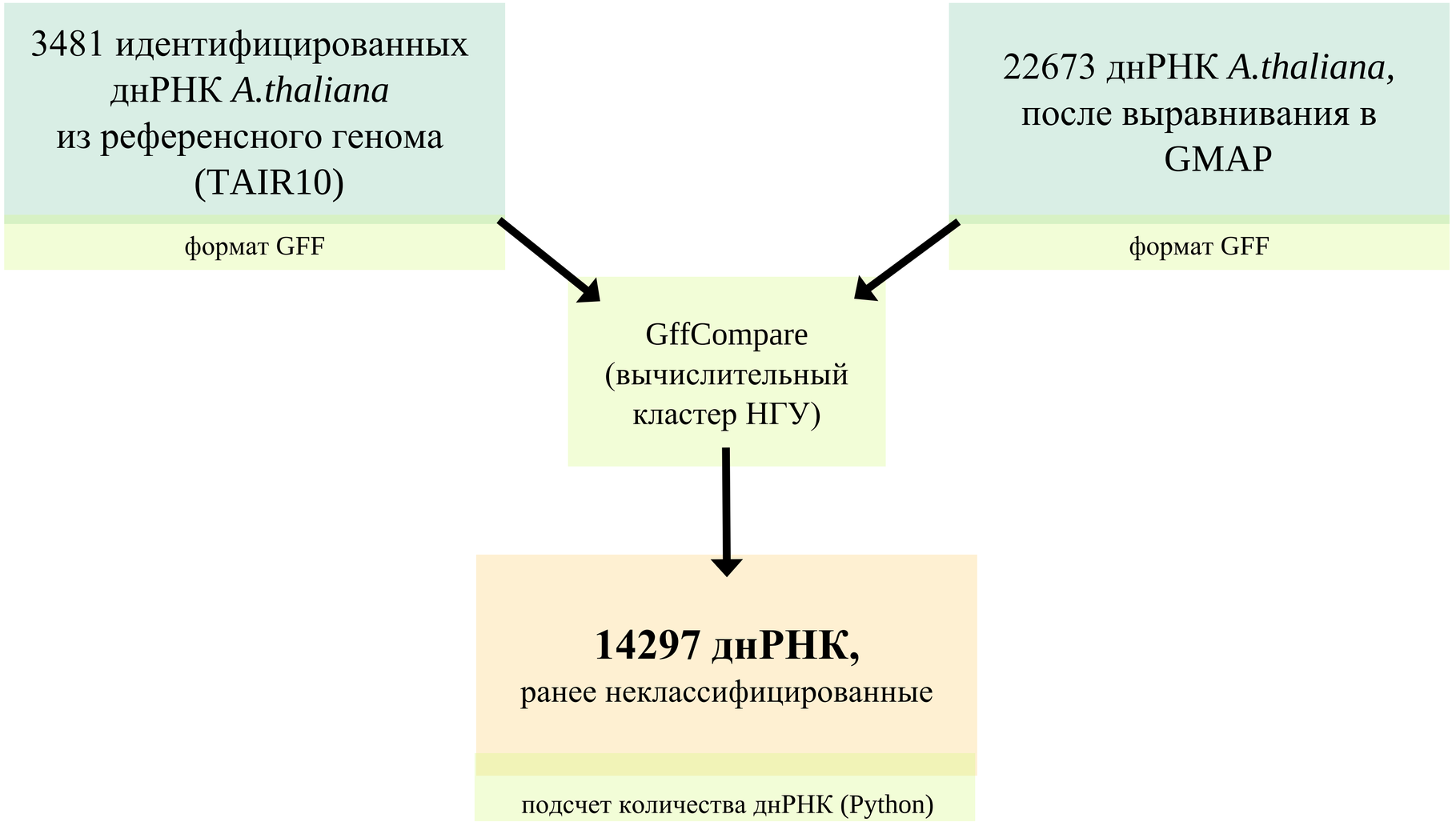


Рисунок 9 — Доля днРНК полностью выровненных на определенное количество экзонов

Подсчет ранее неклассифицированных днРНК



Выводы

1.Сформирован биоинформатический конвейер по классификации днРНК на основе их расположения в геноме.

2.Сформирована выборка выборка днРНК *A.thaliana*, представленных в восьми базах данных, включающая 22674 последовательностей.

3.Классификация днРНК позволила выявить основные классы днРНК:

- 10956 межгенных (43%),
- 5411 антисмысловых (23,9%),
- 166 интронных (0,73%).

4.Среди днРНК преобладают последовательности с одним экзоном (82%), 78% имеют длину экзона менее 500 нуклеотидов, большинство днРНК имеют интроны длиной до 100 нуклеотидов (82%) .

5.Анализ позволил выявить 14297 последовательности днРНК ранее неклассифицированные в референсном геноме.

Спасибо за внимание!