Сборка генома с помощью облаков ридов

Толстоганов Иван Николаевич
научный руководитель: Банкевич Антон Викторович

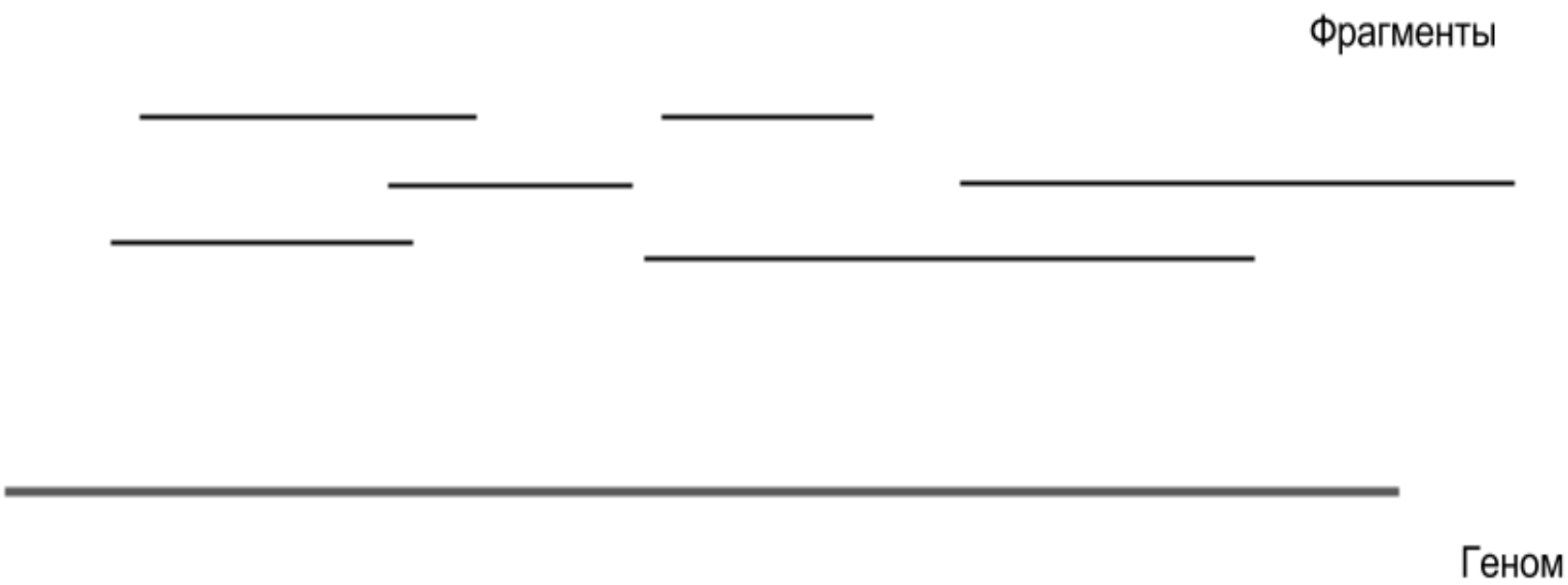
СПб АУ НОЦНТ РАН

15 июня 2017 г.

Сборка метагенома: мотивация

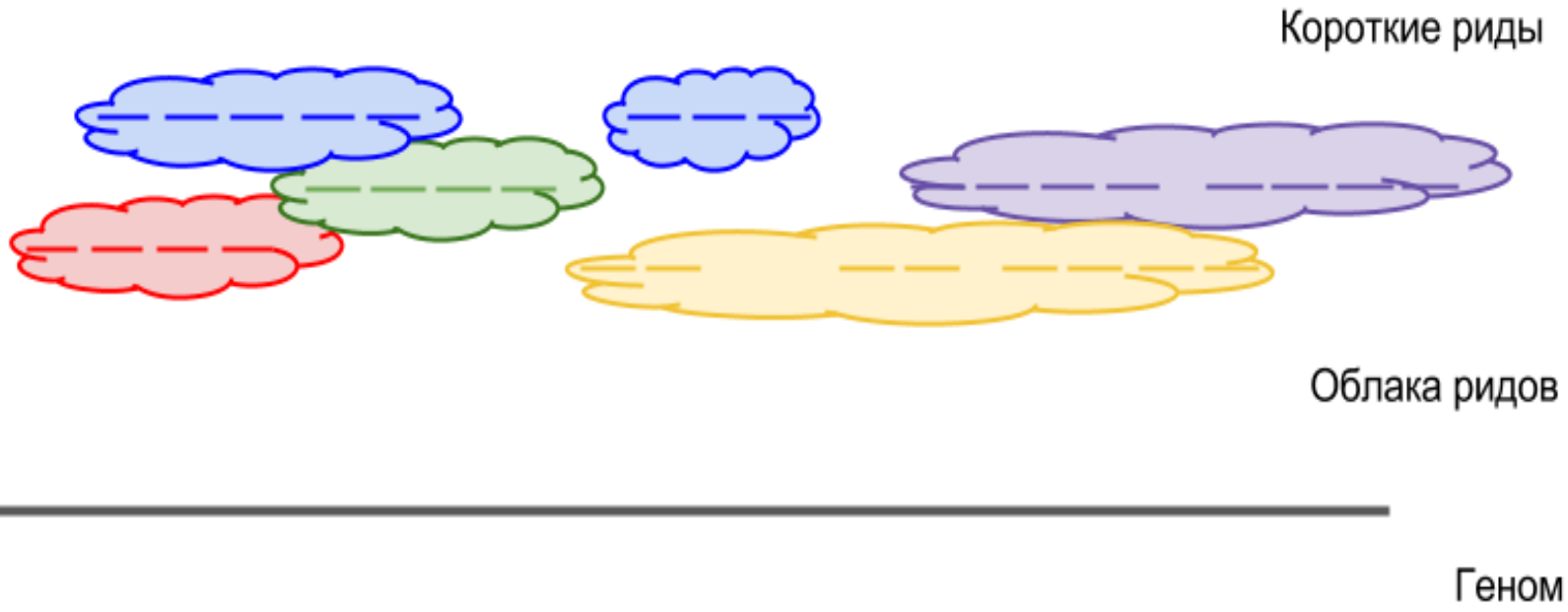
- Поиск новых генов
- Обнаружение сложных вариаций
- Предсказание функции белка

Облака ридов: общее описание



- ДНК разделяется на длинные фрагменты (до 50 килобаз)

Облака ридов: общее описание



- ДНК разделяется на длинные фрагменты (до 50 килобаз)
- Фрагменты секвенируются и баркодируются
- Из одного фрагмента формируется облако коротких ридов

Цель и задачи

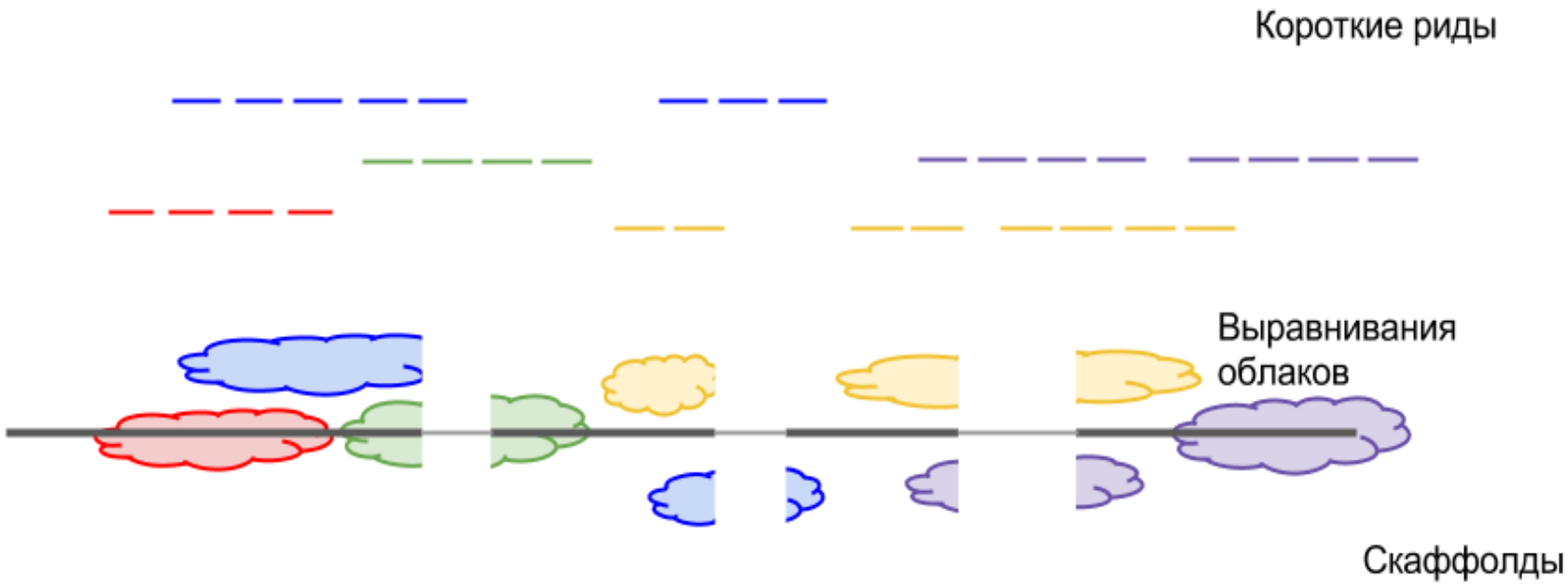
Цель

Разработка алгоритма для сборки метагенома с помощью облаков ридов

Задачи

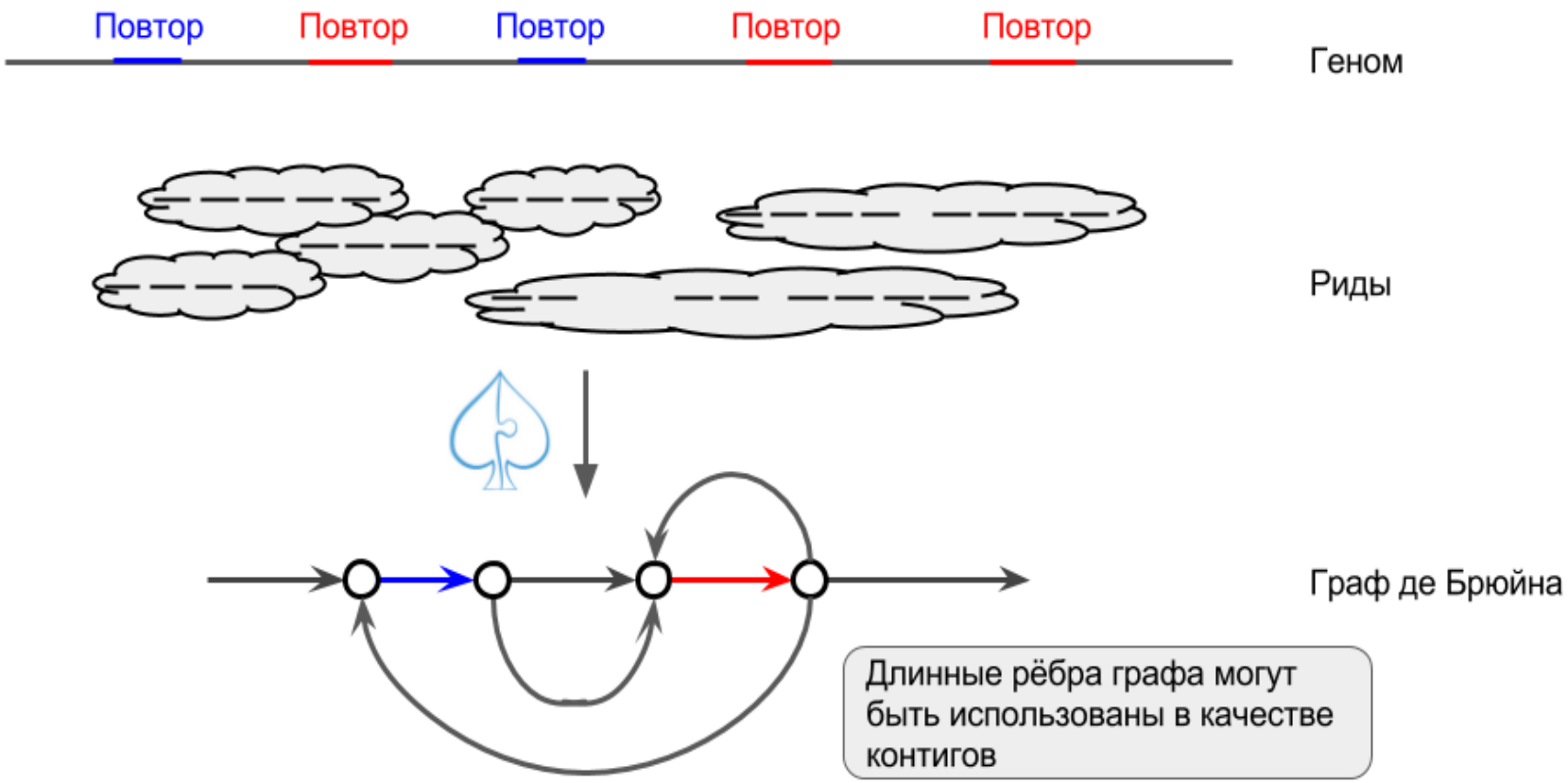
- Реализовать алгоритм в качестве дополнительного модуля к ассемблеру SPAdes
- Поддержать два наиболее распространённых протокола создания облаков ридов
- Сравнить эффективность алгоритма с альтернативными подходами на экспериментальных данных

Облака ридов: скафолдинг

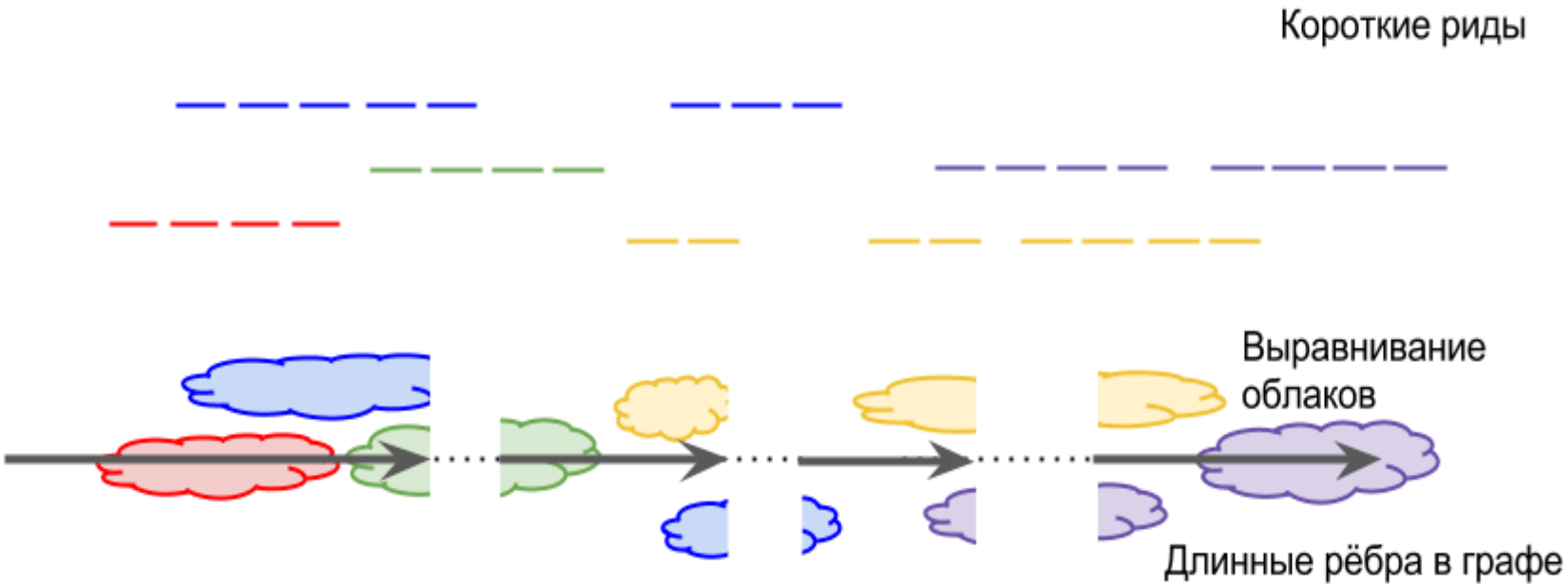


- Облако может быть восстановлено путем выравнивания на контиг
- Выравнивание используется для скаффолдинга

Граф де Брюйна



Облака ридов на графе де Брюйна



- Облака могут быть приложены и к рёбрам графа
- Для каждого длинного ребра мы хотим найти следующее в геномном пути
- Для этого используются облака ридов и структура графа

Закрытие пробелов

- Длинные рёбра в геномном пути могут быть расположены далеко друг от друга
- Пробелы могут быть закрыты после нахождения правильного порядка

Закрытие пробелов

- Длинные рёбра в геномном пути могут быть расположены далеко друг от друга
- Пробелы могут быть закрыты после нахождения правильного порядка
- Для закрытия используются пути из коротких ребер

Оценка качества

- Используется простой метагеном из 10 организмов
- Референсы недоступны, вместо них используются:
  - 23 последовательности рибосомальных оперонов
  - Отдельные сборки каждого организма

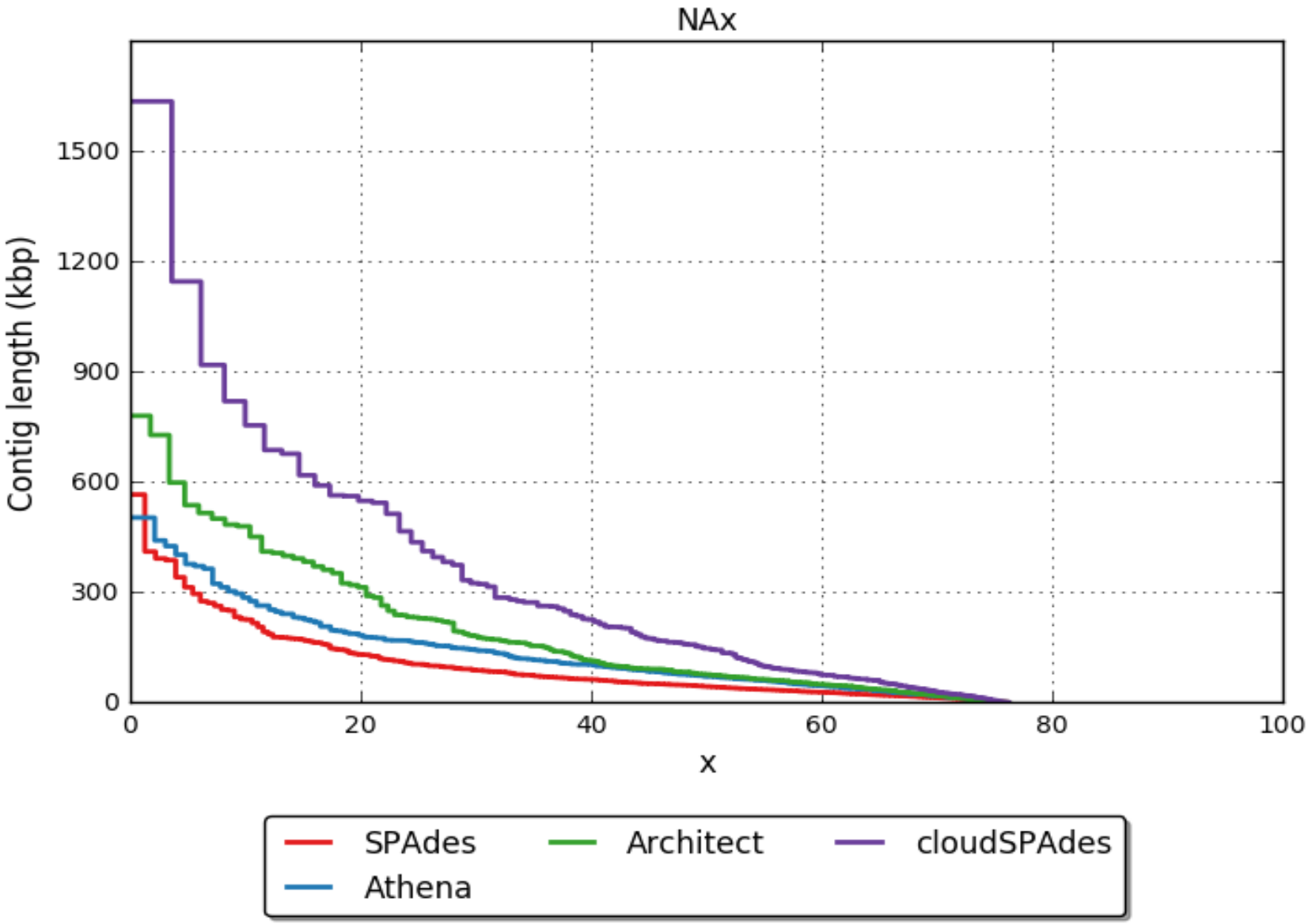
Сборки были переданы в качестве референсов инструменту metaQUAST, предназначенному для оценки качества метагеномной сборки

Оценка качества

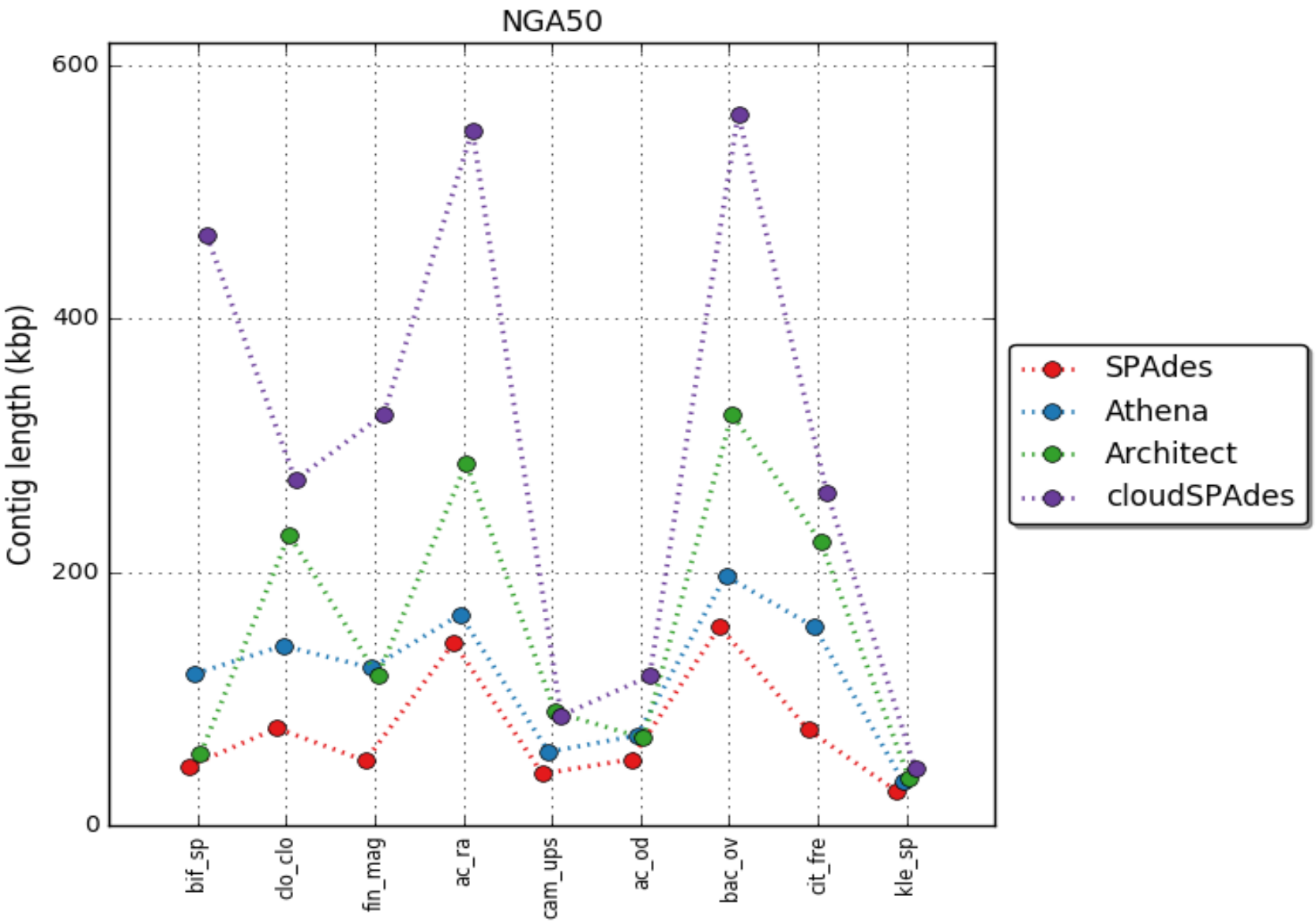
Assembly	SPAdes	Athena	Architect	cloudSPAdes
# contigs	4333	3700	3989	3607
Total length	44454653	**47352283**	44467866	45282462
N50	70276	114821	153964	**270900**
# misassemblies	**21**	25	29	39
Genome fraction (%)	98.744	98.923	98.778	**99.061**
# N's per 100 kbp	**0.00**	**0.00**	?	127.14
Largest alignment	565835	503302	780028	**1634993**
NA50	44044	71261	76715	**147448**

- cloudSPAdes: Реализация алгоритма сборки с помощью облаков ридов на основе SPAdes
- Athena, Architect: Альтернативные инструменты для метагеномной сборки

NAx


NAx
Contig length (kbp)
1500
1200
900
600
300
0
0
20
40
60
80
100
x
SPAdes
Athena
Architect
cloudSPAdes

NGA50 для каждого организма



Оценка качества: итоги

Преимущества cloudSPAdes:

- Длинные контиги
- Восстановлено 22 оперона из 23, нашлось 9 новых

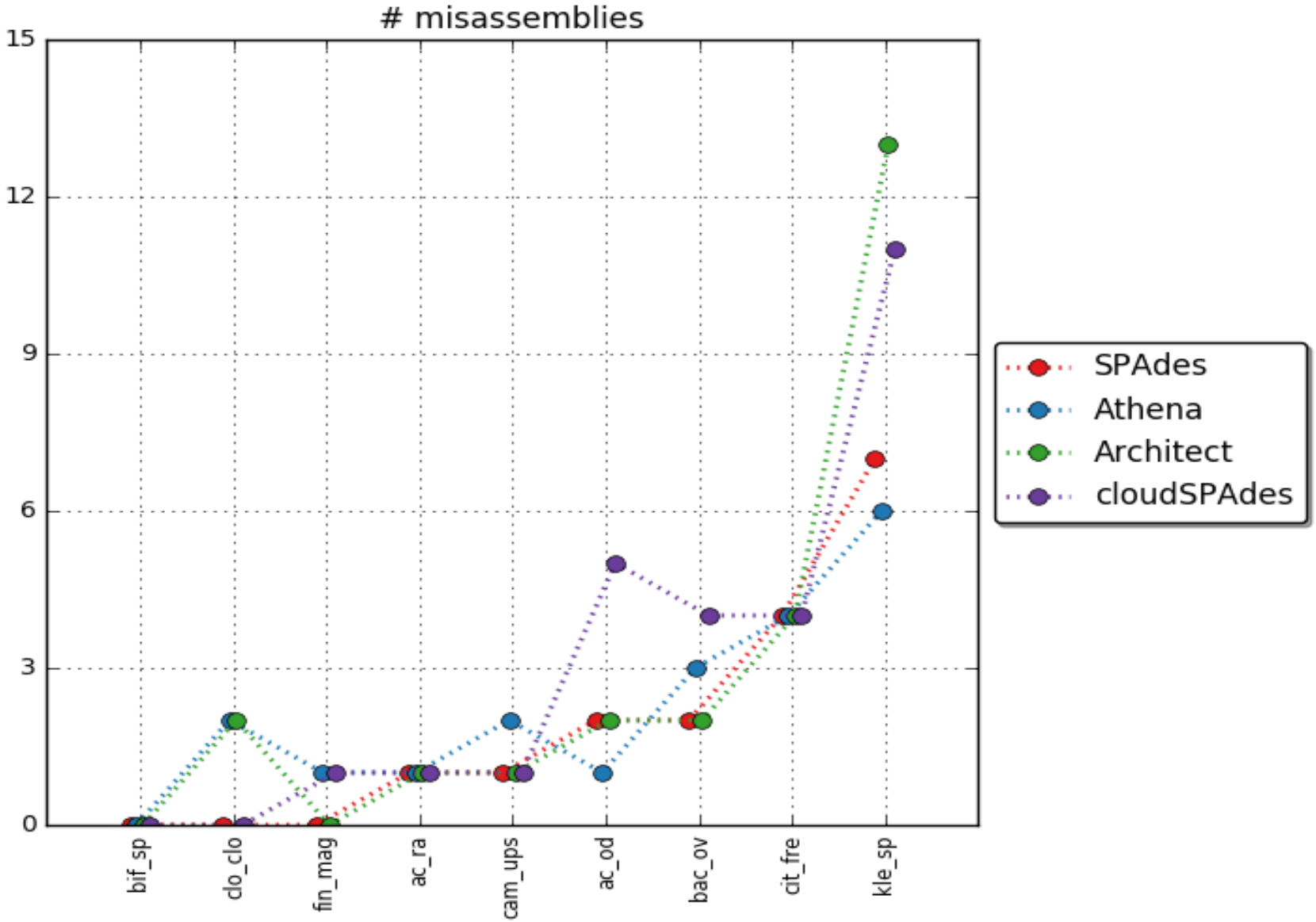
Недостатки:

- Ошибки сборки в низкопокрытых геномах

Результаты

- Разработан алгоритм сборки генома с помощью облаков ридов
- Алгоритм встроен в ассемблер SPAdes в качестве дополнительного модуля разрешения повторов
- Поддержаны протоколы создания облаков ридов TruSeq Synthetic Long Reads и 10XGenomics GemCode

Ошибки сборки



Полные результаты на GemCode

Assembly	SPAdes	cloudSPAdes	Architect	ARCS	Athena	FragScaff
# contigs	4333	3607	3989	4234	3700	4704
Largest contig	565922	**1635422**	1163531	565922	506140	1064532
Total length	44454653	45282462	44467866	44455643	47352283	**60294921**
N50	70276	**270900**	153964	87970	114821	87521
# misassemblies	**21**	40	29	62	27	365
Genome fraction (%)	98.745	**99.061**	98.776	98.750	98.964	98.839
# N's per 100 kbp	**0.00**	127.14	?	2.23	**0.00**	?
Largest alignment	565835	**1634993**	780028	565835	503302	618353
NA50	44044	**147327**	76715	47089	71261	18146

Результаты на TSLR

Assembly	Baseline	cloudSPAdes	Architect	FragScaff
# contigs	3657	3372	3407	3293
Total length	46637846	46894699	46683828	**52366979**
N50	78631	143747	**191269**	100994
# misassemblies	**57**	75	186	305
Genome fraction (%)	42.747	42.727	**42.786**	42.765
# N's per 100 kbp	**9.40**	624.58	18.31	10755.30
Largest alignment	739469	**1311221**	1240613	739469
NA50	77174	122883	**126089**	59381