САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО
Кафедра прикладной математики

Построение алгоритмов для анализа и улучшения качества геномных сборок

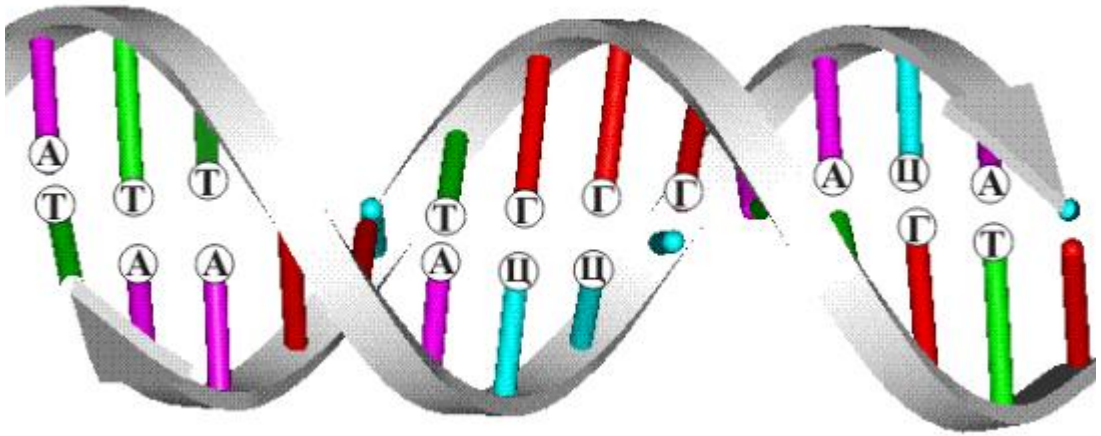
диссертация на соискание степени магистра

Выполнила студентка гр. 63601/3: Шелудько Е.Э.
Руководитель: Мясникова Е.М.

2015 г.

Основные понятия

Геном – это полная совокупность наследственной информации организма. Закодирована в молекулах ДНК.

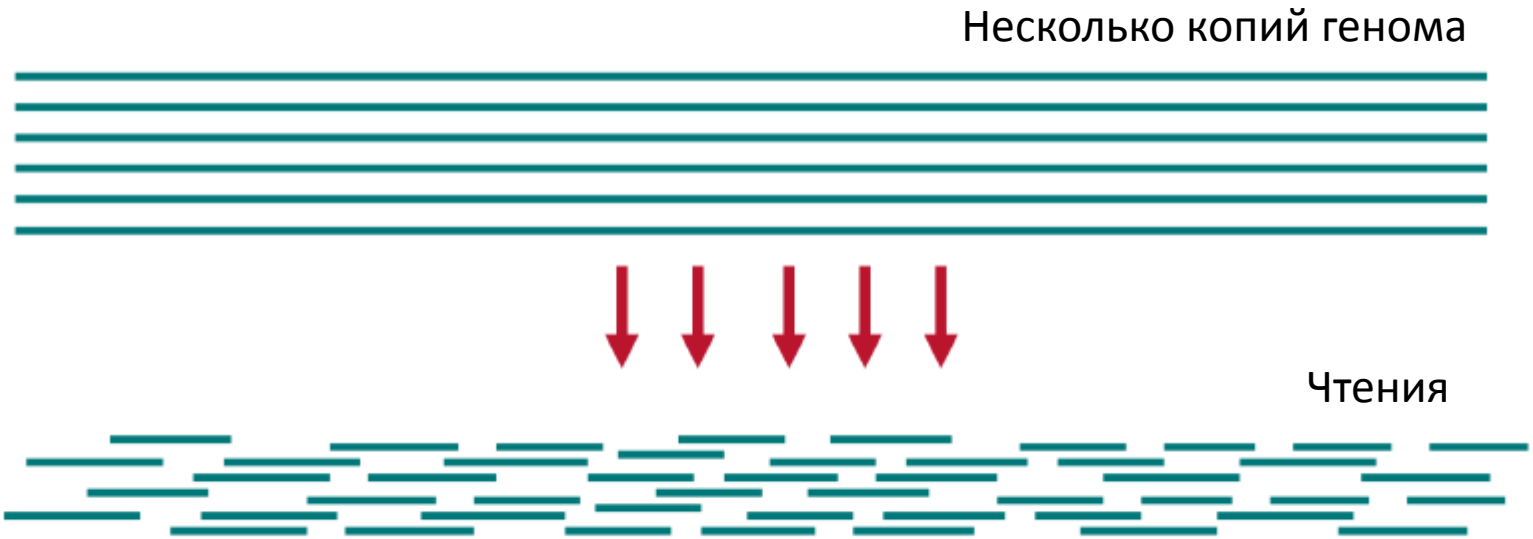


Нуклеотиды:

- аденин (A)
- цитозин (C)
- гуанин (G)
- тимин (T)

Комплементарная парность оснований: A - T, C - G.

Секвенирование



Секвенатор компании Illumina: вычленяет случайно расположенный в геноме фрагмент длинной около 500 нуклеотидов, а затем происходит считывание двух последовательностей с его концов (длиной примерно по 100 нуклеотидов каждая). Эти последовательности называются *парными чтениями*.

Сборка генома

Сборка генома - это процесс восстановления генома из чтений.

Чтения (риды, reads) – фрагменты ДНК, которые выдает секвенатор.

Контиги (contigs) – максимальные последовательности генома, которые удалось восстановить.

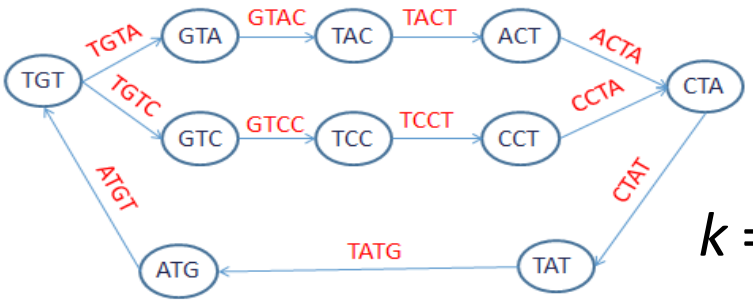
Существует две различные задачи сборки:
• сборка при наличии ранее восстановленного (референсного) генома особи того же вида, что и исследуемая;
• *de novo* сборка, когда отсутствует восстановленный геном.

Цели и задачи

- Анализ и выявление особенностей современных сборщиков на геномах кишечной палочки (E. coli), дрозофилы и туберкулезной палочки;
- Оценки результатов качества сборок;
- Выбор параметров сборщика, обеспечивающих хорошее качество сборки:
  - Моделирование последовательности нуклеотидов цепями Маркова разных порядков;
  - Исследования влияния разного GC-контента на качество сборки генома;
  - Анализ влияния выбора *k*-мер на качество сборки.

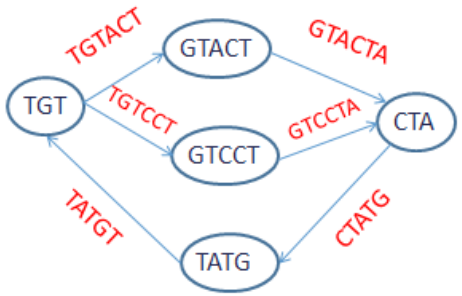
Граф де Брюйна (*de Bruijn*)

Вершинами этого графа являются *k*-меры, то есть строки длины *k*.

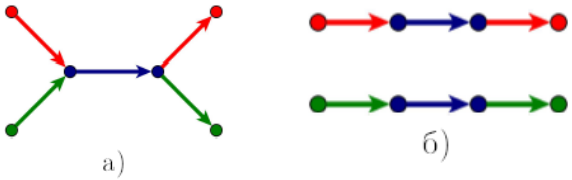


$k = 3$

Сжатый граф:



Повтор – последовательность нуклеотидов длины ≥(*k*-1), присутствующие в геноме более одного раза.



Парные чтения помогают разрешать повторы.

Сборщики

Сборщики, основанные на использовании графах де Брюйна:

- ABySS - использует распределенное хранение несжатого графа де Брюйна, при котором информация о различных *k*-мерах распределена по различным физическим машинам.

- Velvet - использует оригинальное представление сжатого графа, в котором однозначные участки графа сжимаются не в длинные ребра, а длинные вершины.

- SPAdes - собирает контиги с разным уровнем покрытия, выполняя последовательные сборки с повышающимся значением размера *k*-меры.

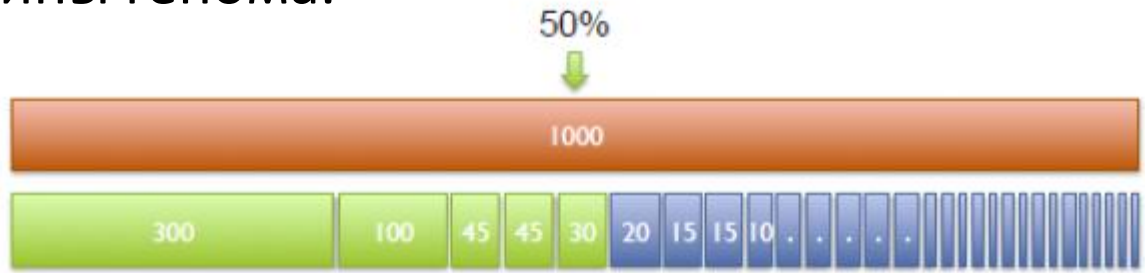
Параметры оценки качества сборки

- **# contigs** - количество контигов;
- **Total length** – общее количество нуклеотидов в сборке;
- **GC (%)** – общее количество G и C нуклеотидов в сборке, деленное на общую длину сборки;
- **Reference GC (%)** – общее количество G и C нуклеотидов в исходной последовательности;
- **N50** – значение, при котором контиги длинее этого значения покрывают половину собранного генома;
- **# misassemblies** – количество неправильных соединений геномных последовательностей;
- **Unaligned length** – общая длина всех невыровненных нуклеотидов по исходной последовательности;
- **Genome fraction (%)** – доля выровненных нуклеотидов в исходной последовательности.

Параметр N50

N50 – важная характеристика качества сборки.

Контиги упорядочивают по убыванию длины.
Суммируют их длины, начиная с самого большого контига.
N50 – длина контига, на котором достигается покрытие половины генома.



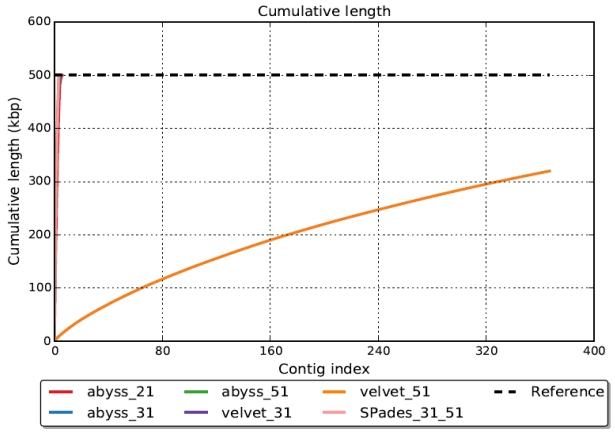
N50 =30
(300+100+45+45+30=520 > 500)

1. Анализ современных сборщиков

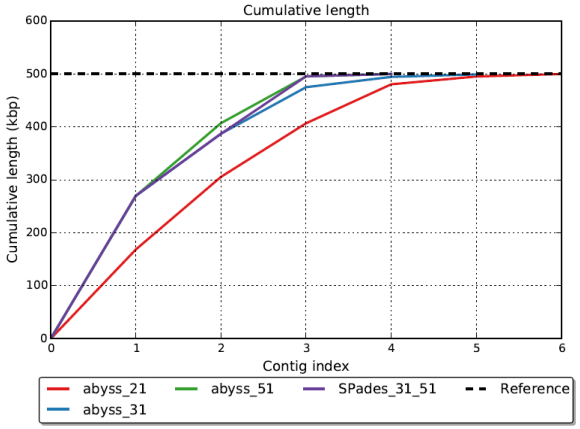
Результаты сборки генома E. coli

	Abyss *k*=21	Abyss *k*=31	Abyss *k*=51	Velvet *k*=31	Velvet *k*=51	Spades *k*=31,51
# contigs	6	5	**4**	367	367	**4**
Total length	500046	499283	**500240**	320126	320126	**499990**
GC (%)	50.95	50.94	**50.95**	50.55	50.55	**50.93**
Reference GC (%)	50.92	50.92	**50.92**	50.92	50.92	**50.92**
N50	137150	269818	**269795**	902	902	**270153**
# misassemblies	0	0	**0**	0	0	**0**
Unaligned length	0	0	**0**	0	0	**0**
Genome fraction (%)	99.745	99.655	**99.840**	63.689	63.689	**99.978**

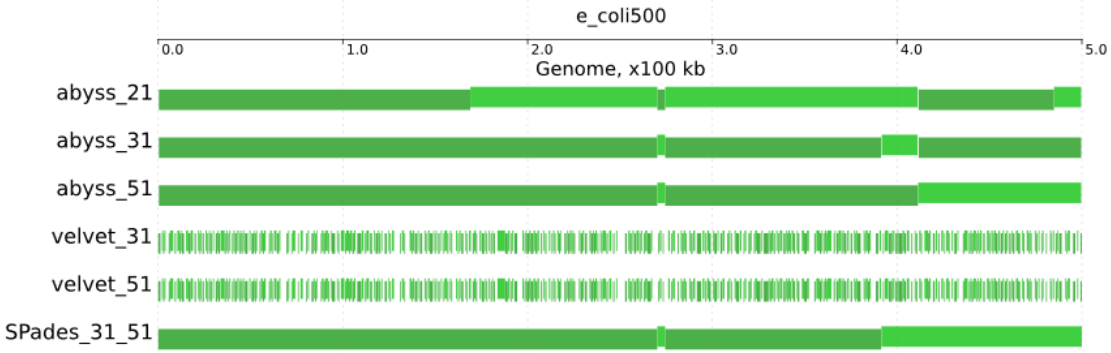
Результаты сборки генома E. coli



Сравнения кумулятивного покрытия контигами для всех сборщиков



Сравнения кумулятивного покрытия контигами для сборщиков ABySS и SPAdes



Выравнивание контигов на референсный геном

Выводы: лучшие результаты показали сборщики SPAdes и ABySS при k=51.

2. Моделирование генома цепями Маркова

Цепь Маркова (ЦМ)

Последовательность дискретных случайных величин $\{X_n\}$ образует однородную цепь Маркова, если

$$P(X_n = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, X_{n-1} = i) =$$

$$= P(X_n = j \mid X_{n-1} = i)$$

Цепь Маркова порядка l :

$$P(X_n = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{n-2} = k_{n-2}, X_{n-1} = i) =$$

$$= P(X_n = j \mid X_{n-l} = k_{n-l}, \ldots, X_{n-1} = i)$$

Матрица переходных вероятностей на n-м шаге:

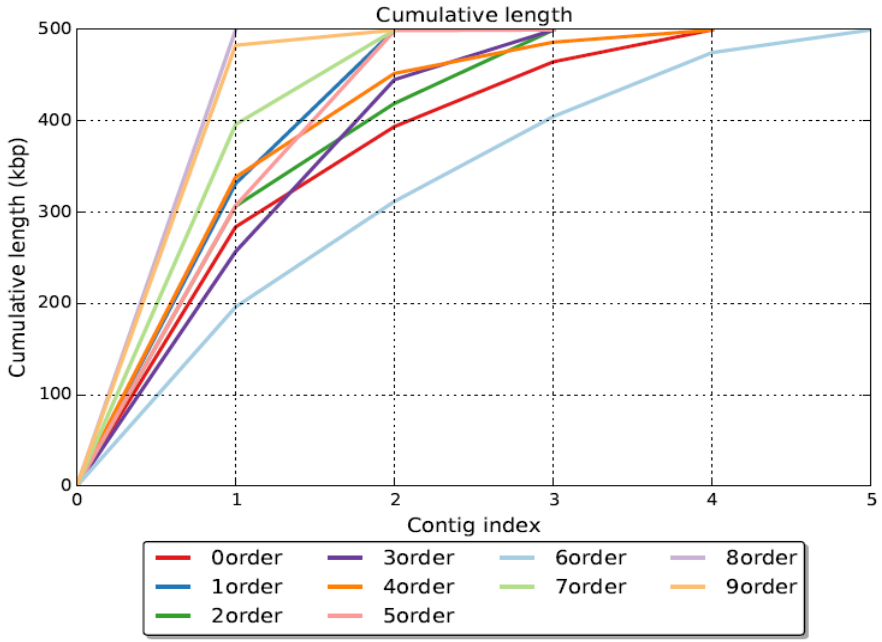
$$P_{ij}(n) \equiv P(X_n = j \mid X_{n-1} = i).$$

Результаты сборки последовательностей – ЦМ разных порядков

	0order	1order	2order	3order	4order	5order	6order	7order	8order	9order
# contigs	4	2	3	3	4	3	5	2	**1**	**2**
Total length	499466	499407	499357	499424	499363	499394	499495	499269	**499272**	**499302**
GC (%)	50.57	50.61	50.59	50.61	50.57	50.59	50.59	50.54	**50.61**	**50.50**
N50	283746	331334	306503	256709	338184	307028	115526	395925	**499272**	**482628**

Сборщик ABySS
k = 51

Результаты сборки последовательностей – ЦМ разных порядков



Сравнения кумулятивного покрытия контигами для ЦМ разных порядков

Выводы: лучшее качество сборки обеспечивается на последовательностях, в основе которых лежат цепи Маркова 8 и 9 порядков.

3. Моделирование разного GC-контента

Результаты сборки последовательностей с разным GC-контентом

Сборщик ABySS
k = 51

	0order_N1	0order_N3	0order_N5	0order_N4	0order_N2
# contigs	2	4	2	**1**	2
Total length	499428	499425	499278	**499243**	499342
GC (%)	10.69	11.23	42.95	**59.64**	89.63
N50	461338	285575	301476	**499243**	370915

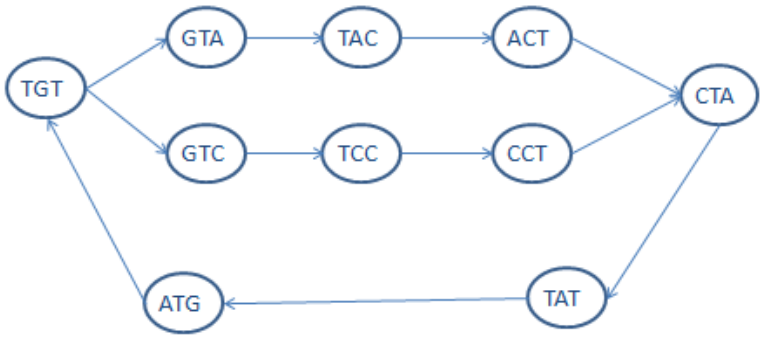
Сборщик ABySS
k = 51

	1order_N1	1order_N2	1order_N3	1order_N5	1order_N4
# contigs	2	5	**2**	5	4
Total length	499376	499598	**499419**	499516	499429
GC (%)	44.38	45.44	**51.44**	51.44	62.61
N50	307504	274082	**446825**	87384	271013

Выводы: явная зависимость качества сборки от GC-контента не обнаружена.

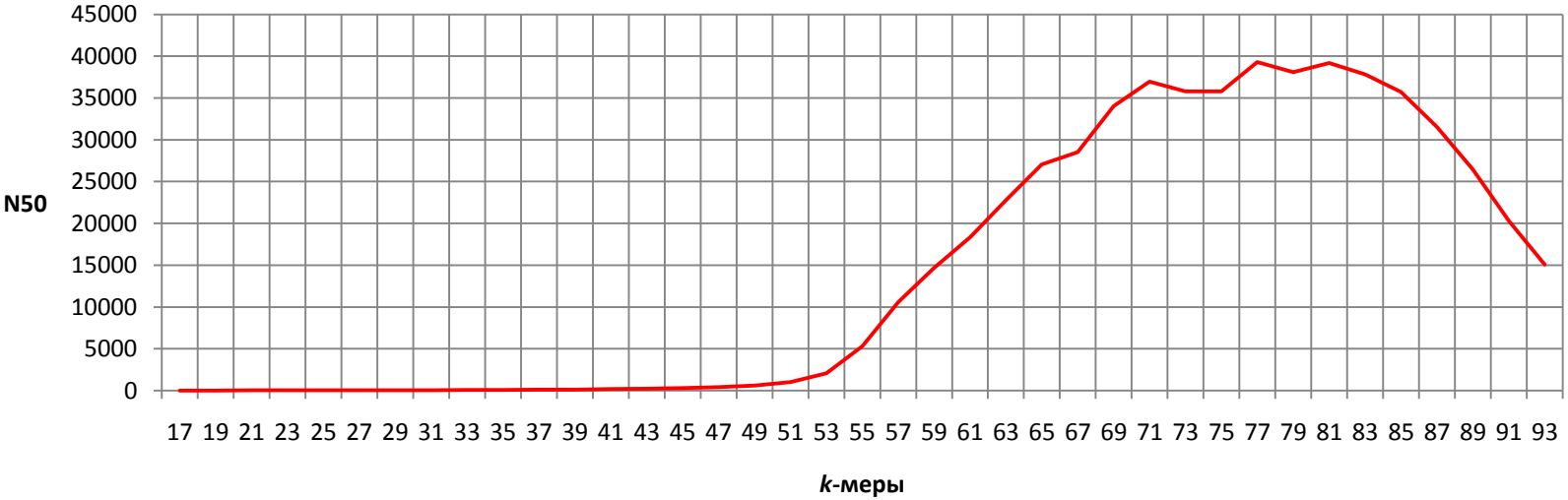
4. Анализ влияния выбора *k*-мер на качество сборки

Результаты анализа *k*-мер



$k = 3$

Изменения оценки N50 для туберкулезной палочки:



Выводы: существует значение *k*-меры, обеспечивающее наилучшее качество сборки.

Результаты

- Исследованы различные сборщики при различных значениях *k*-мер на разных данных.
- Смоделированы последовательности, в основе которых лежали цепи Маркова до 9 порядка, которые были исследованы на качество сборки.
- Произведено исследования влияния разного GC-контента на качество сборки генома.
- Исследовано влияние выбора значений *k*-меры на сборщик.

Общие выводы:

- Применения предложенной методики позволяет выбрать параметры сборки так, чтобы она давала хорошие результаты.
- Эта работа важна для понимания качества полученной *de novo* сборки, что может оказаться критичным при секвенировании организмов из ранее неисследованных семейств.
- Параметры сборщиков нужно выбирать учитывая свойства и характеристики конкретного генома.

Благодарность:

Урошлеву Леониду Андреевичу (ИОГен РАН)
Мясниковой Екатерине Марковне
Кадыровой Наталье Олеговне

Спасибо за внимание

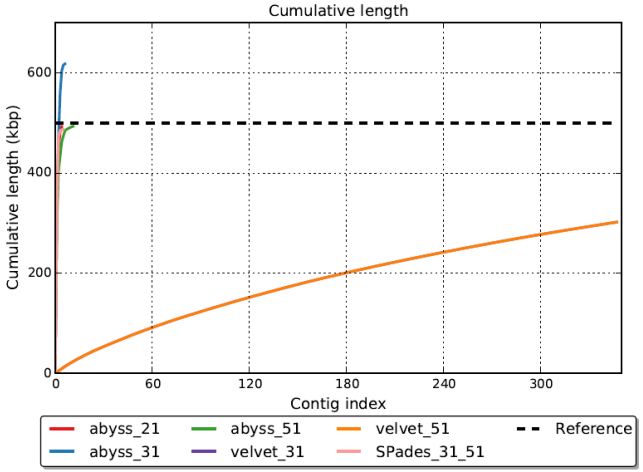
Проверка независимости нуклеотидов

Длина по x нуклеотидов, где x =	Квантиль H	Число степеней свободы	P(x,y)	Отвергается /принимается гипотеза
2	16.92	9	130 463.46	отвергается
3	61.66	45	270 287.17	отвергается
4	222.08	189	357 945.11	отвергается
5	830.46	765	409 735.84	отвергается
6	3 198.99	3069	454 729.52	отвергается
7	12 543.96	12285	499 730.50	отвергается
8	49 665.84	49149	566 328.30	отвергается
9	197 384.90	196353	762 772.14	отвергается
10	776 396.10	774348	1 430 472.58	отвергается

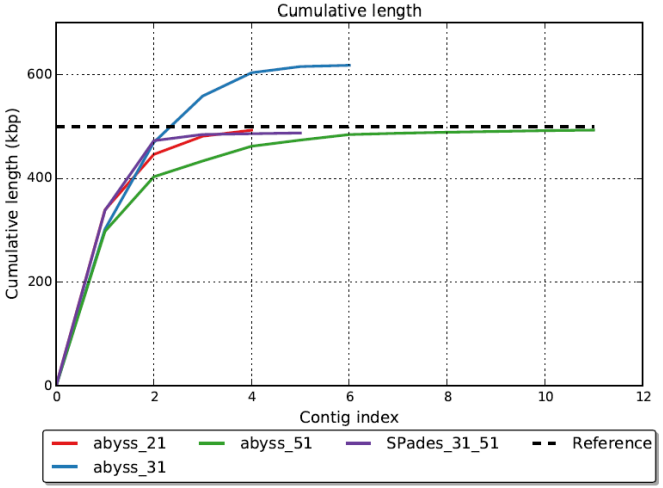
Результаты сборки генома дрозофилы

	Abyss *k*=21	Abyss *k*=31	Abyss *k*=51	Velvet *k*=31	Velvet *k*=51	Spades *k*=31,51
# contigs	4	6	**11**	347	347	**5**
Total length	492740	617700	**492879**	301941	301941	**487357**
GC (%)	42.86	42.34	**42.73**	42.80	42.80	**42.84**
Reference GC (%)	42.60	42.60	**42.60**	42.60	42.60	**42.60**
N50	339256	167145	**297910**	901	901	**338341**
# misassemblies	0	0	**1**	0	0	**0**
Unaligned length	0	0	**0**	0	0	**0**
Genome fraction (%)	96.919	97.617	**97.659**	60.058	60.058	**97.463**

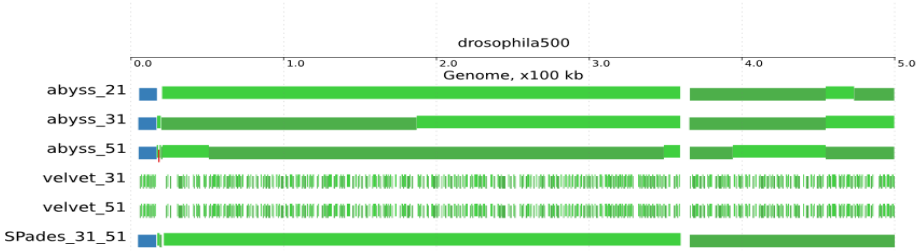
Результаты сборки генома дрозофилы



Сравнения роста длин контигов для всех сборщиков



Сравнения роста длин контигов для сборщиков ABySS и SPAdes



Выравнивание контигов на исходный геном