ПРАКТИКУМ 15. ЗАДАНИЯ ПО СБОРКЕ DE NOVO

Выполнила Неверова-Симчит Елена

Подготовка к выполнению практикума

1. Подготовка чтений программой trimmomatic
2. Подготовка k-меров длины k=31
3. Сборка на основе k-меров
4. Анализ программой megablast

    Самый длинный контиг

    Второй по длине контиг (ID 5)

    Третий по длине контиг (ID 4)

6. Запуск SPAdes

Подготовка к выполнению практикума

ID моих чтений проекта по секвенированию бактерии *Buchnera aphidicola* str. Tuc7 согласно таблице – SRR4240360. Это короткие (в моем случае длины 36) одноконцевые чтения, полученные в проекте. Ссылка, по которой доступны эти чтения.

Создаем рабочую поддиректорию pr15 (/mnt/scratch/NGS/aliserana/pr15), переходим в нее и далее работаем там.

Скачиваем архивированный файл с чтениями:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/000/SRR4240360/SRR4240360.fastq.gz

Получился файл SRR4240360.fastq.gz, размер файла – 194 MB. Еще мне захотелось посмотреть на качество данных чтений:

fastqc SRR4240360.fastq.gz

Получены файлы SRR4240360_fastqc.html, SRR4240360_fastqc.zip. Html-файл также скопирован в public_html. Получается, что длина всех моих чтений – 36, всего их в файле содержится 8254632 штук. Качество чтений в целом хорошее, спорные моменты возникают только во второй трети (начиная с 25 нуклеотида – желтая зона для «усов»), вероятнее всего, при чистке для многих чтений удалятся нуклеотиды с 30 позиции (красная зона для «усов») – рисунок 1. Также, мне показалось интересным, что здесь GC состав снижен (30%), распределение каждого из нуклеотидов по позициям в чтениях допустимое – рисунок 1.

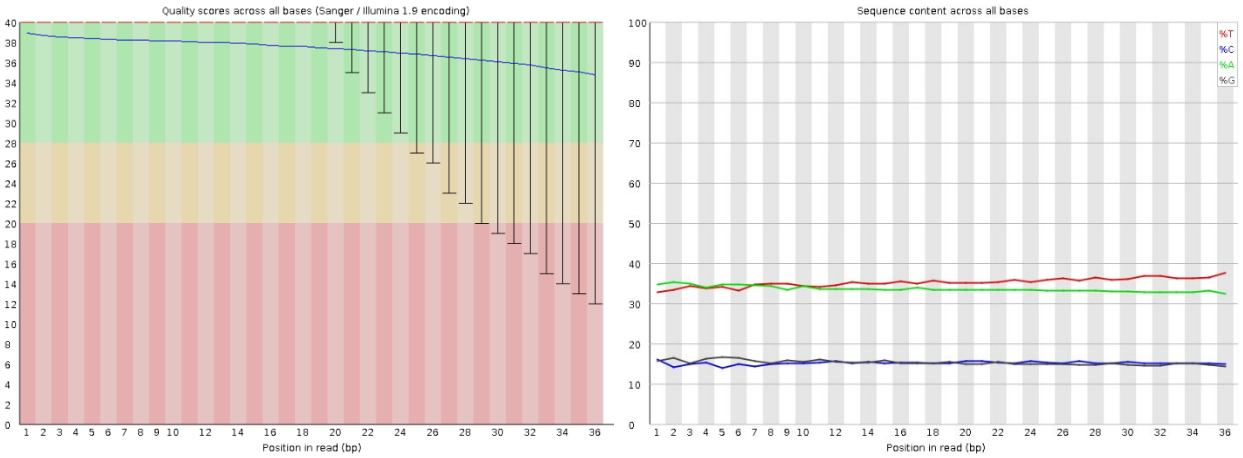

Рисунок 1. Данные о качестве чтений SRR4240360. Слева - распределение качества в позициях, справа - представленность нуклеотидов по позициям в чтениях.

В дальнейшем нам понадобится информация об адаптерах для Illumina, изначально она хранится в /mnt/scratch/NGS/adapters/ в 6 разных файлах. Объединим только те, которые относятся к одноконцевым чтениям в один – adapters.fasta (cat /mnt/scratch/NGS/adapters/*SE.fa > adapters.fasta).

Подготовка чтений программой trimmomatic

Как показано выше, качество некоторых чтений в конечных позициях неудовлетворительно, поэтому применим trimmomatic 0.22 для одноконцевых чтений. Возможно, это остатки адаптеров, поэтому удалим их:

```
TrimmomaticSE SRR4240360.fastq.gz SRR4240360_trimmed0.fastq.gz ILLUMINACLIP:adapters.fasta:2:7:7
```

Получаем файл SRR4240360_trimmed0.fastq.gz. Проведем анализ качества полученных чтений:

```
fastqc SRR4240360_trimmed0.fastq.gz
```

Получены файлы SRR4240360_trimmed0_fastqc.html, SRR4240360_trimmed0_fastqc.zip. Html-файл также скопирован в public_html. Осталось 8213351 чтений, таким образом с остатками адаптеров были 0,5% (41281) чтений.

Затем удалим все нуклеотиды с качеством ниже 20 с конца и оставим только те чтения, длина которых не меньше 32 нуклеотидов:

```
TrimmomaticSE SRR4240360_trimmed0.fastq.gz SRR4240360_trimmed1.fastq.gz TRAILING:20 MINLEN:32
```

Получаем файл SRR4240360_trimmed1.fastq.gz. Размер файла в итоге уменьшился немного – до 184 MB (на 10 MB). Далее работаем именно с этим файлом.

Снова проведем анализ качества полученных чтений:

```
fastqc SRR4240360_trimmed1.fastq.gz
```

Получены файлы SRR4240360_trimmed1_fastqc.html, SRR4240360_trimmed1_fastqc.zip. Html-файл также скопирован в public_html. Осталось 7921744 чтений – 95,97% от исходного числа. Качество чтений улучшилось, особенно в последних позициях – рисунок 2.

Соберем все результаты в единый файл при помощи multiqc, version 1.15:

```
multiqc ../
```

Найдены были все 3 отчетных файла и получены на выходе папка multiqc_data и файл multiqc_report.html (по аналогии скопирован в public_html). Получаем, что качество чтений улучшилось, но незначительно. Удаление адаптеров не особо повлияло на среднее качество по позициям, по сравнению с удалением «плохих концов» – рисунок 2.

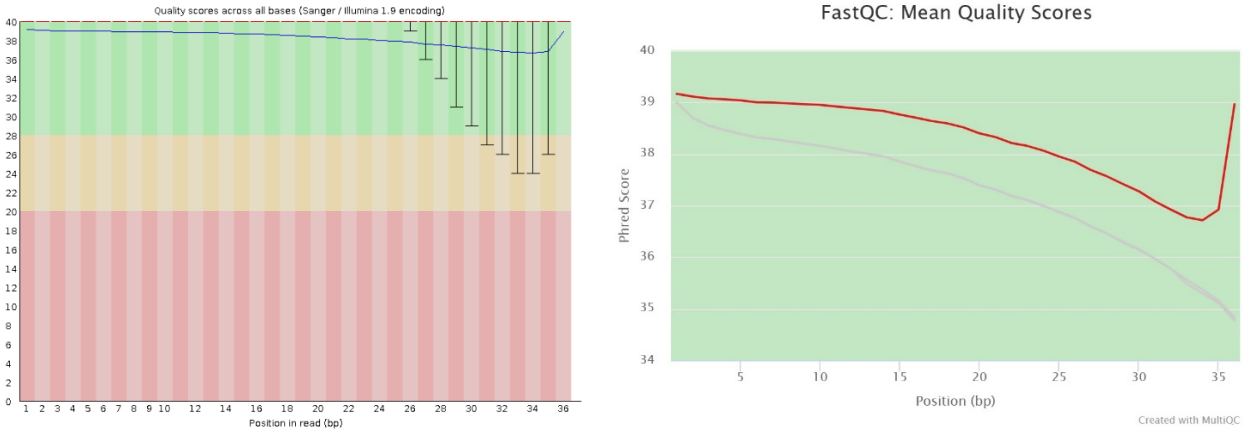

Рисунок 2. Качество чтений после триммирования. Слева - распределение качества по позициям в итоге (SRR4240360_trimmed1.fastq.gz). Справа - распределение средних значений качества в каждой позиции для всех файлов, красный – SRR4240360_trimmed1.fastq.gz, бледно-красный - исходный и промежуточный файлы.

Подготовка k-меров длины k=31

Воспользуемся программой velveth 1.2.09 для получения списка k-меров длины k=31 среди наших дважды тримированных чтений – руководство, также помогает опция -help. С учетом того, что чтения в нашем случае короткие и непарные (short):

```
velveth assem 31 -fastq.gz -short SRR4240360_trimmed1.fastq.gz
```

На выходе получена папка assem, содержащая файлы Log, Roadmaps, Sequences.

Сборка на основе k-меров

Запустим программу velvetg 1.2.10+dfsg1 для получения сборки генома на основе наших k-меров (строим граф де Брёйна):

```
velvetg assem
```

На выходе получены в папке assem следующие файлы: contigs.fa (содержит контиги длиною более чем 2k, то есть длиннее 62 нуклеотидов), Graph, LastGraph (вся информация об итоговом графе), PreGraph, stats.txt (табулированный, позволяет определить нужное покрытие). Полученный граф состоял из 601 вершины, то есть всего у нас столько контигов. N50 = 43070 (информация из Log).

Узнаем длины трёх самых длинных контигов и их покрытие:

```
cut -f 2,6,7 assem/stats.txt | sort -n -r > numbers
```

Также полезно:

```
grep '^>' assem/contigs.fa > con
```

Искомые результаты представлены в таблице 1. Также удалось выяснить, что L50 = 5.

Таблица 1. Длина и покрытие трёх самых длинных контигов

ID	Длина контига	Покрытие
1	113474	33.534396
5	91818	33.497430
4	64155	35.869924

Откроем con в excel. Среднее покрытие контигов – 121,14 чтений, медиана покрытий контигов – 0,66 чтений, при этом нижний квартиль 0,5 чтений, верхний – 1 чтение, дециль 0,1 = 0,42, дециль 0,9 = 9,3. Поэтому «типичными» я решила считать контиги с покрытием менее, чем в 50 чтений. Данные о контигах с аномально большими покрытиями представлены в таблице 2.

Таблица 2. Контиги с аномально большим покрытием (более чем в 5 раз отличающимся от "типичного")

ID	Длина контига	Покрытие
499	34	58946
500	34	257
496	35	109.5
497	35	90.5
482	65	81.09375

Анализ программой megablast

Сравним программой megablast каждый из трёх самых длинных контигов с хромосомой *Buchnera aphidicola* (для GenBank/EMBL AC: CP009253): ссылка.

Файл с координатами участков выравнивания для каждого из контигов.

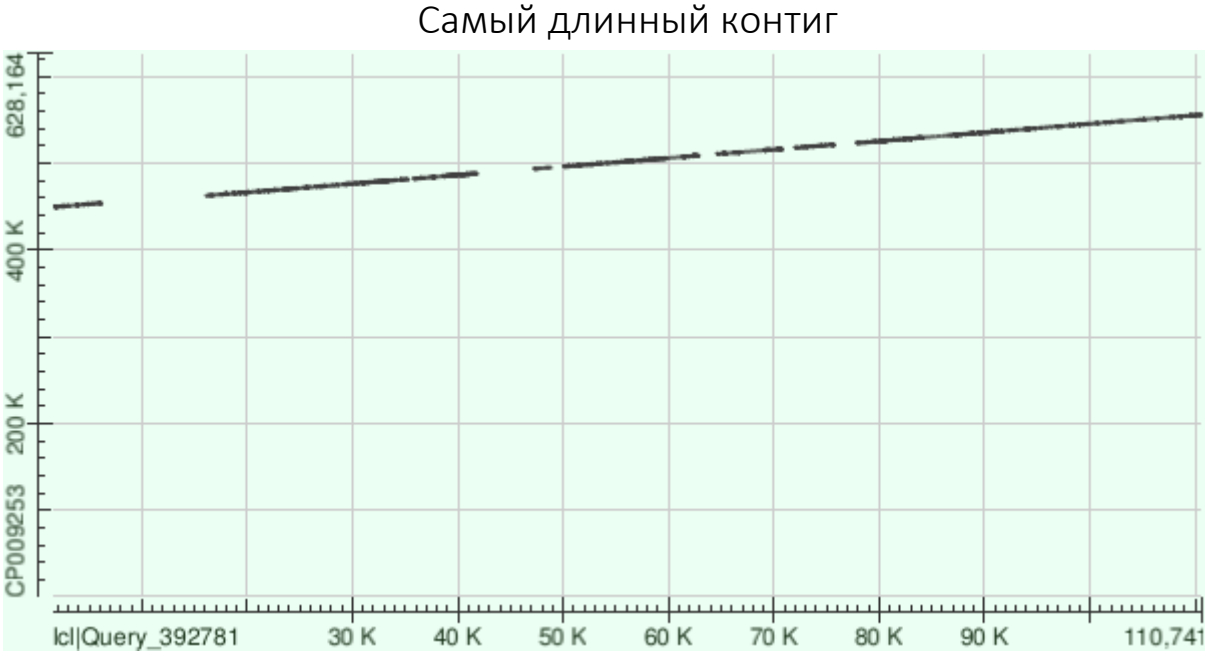

Рисунок 3. Dot-plot выравнивания 1 контига (query) на хромосому (subject)

Самый длинный контиг ложится на вторую половину генома – рисунок 3. Присутствуют небольшие негомологичные участки, возможно, они вызваны неточностью сборки последовательности самого контига. Таблица 1 показывает, что качество всех участков выравнивания высокое (e-value).

Таблица 3. Характеристики участков выравнивания для 1 контига.

№	Координаты хромосомы (участок выравнивания)		Score в битах	E-value	Длина	Идентичных нуклеотидов		Гэпы		Направление последовательности	
	Начало	Конец				Число	%	Число	%	Контиг	Референс
1	449411	454069	2167	0.0	4732	3571	75	152	3	Plus	Plus
2	462496	467421	2724	0.0	5015	3862	77	162	3	Plus	Plus
3	467412	474667	4047	0.0	7389	5691	77	208	3	Plus	Plus
4	474844	480660	2237	0.0	5971	4426	74	250	4	Plus	Plus
5	480874	481545	573	2e-162	686	564	82	20	3	Plus	Plus
6	481997	488106	2278	0.0	6238	4621	74	308	5	Plus	Plus
7	493487	494864	1014	0.0	1384	1108	80	13	1	Plus	Plus
8	495033	495148	145	1e-33	120	107	89	5	4	Plus	Plus
9	496111	500325	1914	0.0	4323	3253	75	153	4	Plus	Plus
10	500370	508806	3949	0.0	8614	6513	76	345	4	Plus	Plus
11	510438	516539	3895	0.0	6238	4894	78	194	3	Plus	Plus
12	517766	521500	2128	0.0	3782	2922	77	99	3	Plus	Plus
13	523105	528679	3029	0.0	5687	4373	77	210	4	Plus	Plus
14	528794	550219	17265	0.0	21721	17688	81	545	3	Plus	Plus
15	550361	555905	4331	0.0	5655	4573	81	127	2	Plus	Plus

Рисунок 4. Dot-plot выравнивания 5 контига (query) на хромосому (subject)

Второй по длине контиг ложится на геном в первой его половине и тоже с наличием негомологичных участков – рисунок 4. Этот разрыв нельзя назвать делецией или вставкой, потому что несовпадающий участок одинаковый по длине выделяется и в контиге, и в геноме (таблица 4).

Таблица 4. Характеристики участков выравнивания для 5 контига.

№	Координаты хромосомы (участок выравнивания)		Score в битах	E-value	Длина	Идентичных нуклеотидов		Гэпы		Направление последовательности	
	Начало	Конец				Число	%	Число	%	Контиг	Референс
1	88200	93683	2462	0.0	5607	4223	75	243	4	Plus	Plus
2	93821	98092	2518	0.0	4345	3372	78	125	3	Plus	Plus
3	98408	99303	713	0.0	901	731	81	9	1	Plus	Plus
4	101712	108876	3777	0.0	7274	5567	77	215	3	Plus	Plus
5	126623	127815	1123	0.0	1199	1004	84	11	1	Plus	Plus
6	127825	140555	5465	0.0	13010	9571	74	548	4	Plus	Plus
7	144368	151796	4401	0.0	7536	5859	78	243	3	Plus	Plus
8	153752	161738	4769	0.0	8168	6355	78	264	3	Plus	Plus
9	161898	166752	3415	0.0	4914	3911	80	112	2	Plus	Plus
10	166750	173180	3301	0.0	6517	4967	76	159	2	Plus	Plus

Рисунок 5. Dot-plot выравнивания 4 контига (query) на хромосому (subject)

Третий по длине контиг лег «по-интересному». Для него условная точка начала генома попала на середину последовательности, поэтому на рисунке 5 мы видим разрыв линии. Снова присутствуют участки, где не удалось выявить гомологию. То, что этот контиг короче двух прошлых повлияло и на сокращение длин участков выравнивания – таблица 5.

Таблица 5. Характеристики участков выравнивания для 4 контига.

№	Координаты хромосомы (участок выравнивания)		Score в битах	E-value	Длина	Идентичных нуклеотидов		Гэпы		Направление последовательности	
	Начало	Конец				Число	%	Число	%	Контиг	Референс
1	599832	604795	3068	0.0	5046	3946	78	170	3	Plus	Plus
2	611229	611524	209	3e-53	297	236	79	2	1	Plus	Plus
3	611633	613671	1238	0.0	2086	1625	78	66	3	Plus	Plus
4	613658	620926	4959	0.0	7379	5485	74	184	2	Plus	Plus
5	621055	627014	2889	0.0	6173	4678	76	248	4	Plus	Plus
6	2004	11103	5749	0.0	9223	7229	78	256	3	Plus	Plus
7	13994	14465	403	1e-111	478	393	82	9	2	Plus	Plus
8	14727	17919	1583	0.0	3226	2451	76	88	3	Plus	Plus
9	17962	20182	2270	0.0	2231	1902	85	30	1	Plus	Plus
10	20358	22183	1476	0.0	1851	1509	82	51	3	Plus	Plus
11	23067	28363	2772	0.0	5433	4159	77	219	4	Plus	Plus
12	30013	32745	1578	0.0	2777	2150	77	84	3	Plus	Plus

Доли гэпов в участках выравнивания и идентичных нуклеотидов похожие для контигов. Данные контиги направлены в одинаковую сторону и совпадают с направлением референса. Что примечательно, судя по координатам, три самых длинных контига при выравнивании на геном не пересекаются.

Запуск SPAdes

Попробуем запустить SPAdes после чтения его мануала. На вход подаем триммированные ранее и нам нужна только сборка:

```
spades -s SRR4240360_trimmed1.fastq.gz --only-assembler -o spades
```

Получаем папку spades и огромное число файлов в ней. Сравним эти результаты с полученными ранее результатами программы velvet.

Заметно еще при запуске, что здесь длина k-мера может достигать значительно больших чисел (<128).

```
grep '^>' spades/contigs.fasta | wc -l
```

Здесь мы получили 502 контига (velvet выдававал 601).

Изучим полученные в spades файлы подробнее:

contigs.fasta – содержит сборку контигов, scaffolds.fasta – содержит сборку скэффолдов (такого velvet не давал), contigs.paths – пути графа, в соответствии с полученными контигами (такого тоже velvet не давал), scaffolds.paths – пути графа, в соответствии с полученными контигами (такого тем более velvet не давал), assembly_graph.fastg – граф

сборки, assembly_graph_with_scaffolds.gfa – этот же граф в архивированном формате (этого тоже при velvet не было). Но файла-аналога stats.txt нет!

Полученный граф состоял из 601 вершины, то есть всего у нас столько контигов. N50 = 43070 (информация из Log).

Узнаем длины трёх самых длинных контигов и их покрытие:

```
grep '^>' spades/contigs.fasta > con
```

Файл con откроем в excel и получим нужные данные – таблица 6.

Таблица 6. Длина и покрытие трёх самых длинных контигов для SPAdes.

ID	Длина	Покрытие
1	174223	17.202819
2	104293	16.110857
3	91713	15.638111

Видим, что, по сравнению с velvet, длина полученных SPAdes контигов выросла.

Применим megablast для каждого из контигов и рассмотрим полученные dotplot. Во всех случаях возросло количество участков выравнивания – 28, 15, 17, соответственно.

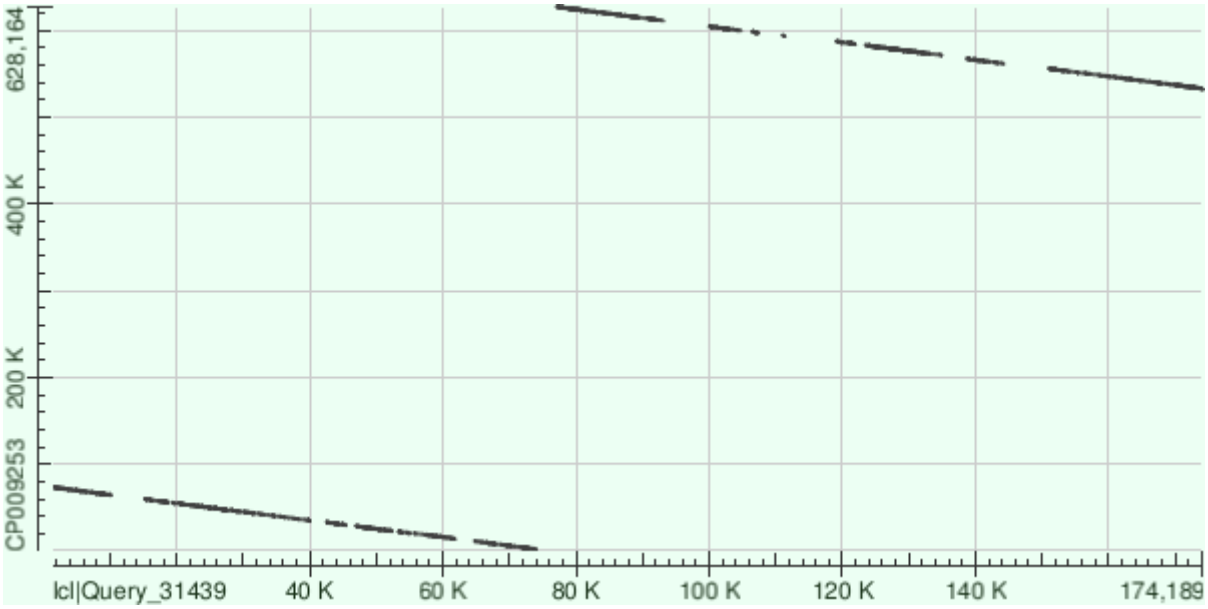

Рисунок 6. Dot-plot выравнивания 1 контига (query) на хромосому (subject)

Самый интересный результат выравнивания для 1 контига. По рисунку 6 видно, что последовательность контига оказалась инвертированной по сравнению с последовательностью референса. Также если сравнивать рисунки 5 и 6, то можно заметить, что они описывают схожие участки хромосом. Поэтому, на мой взгляд, первый контиг SPAdes другой вариант сборки контига с ID 4, полученного velvet.

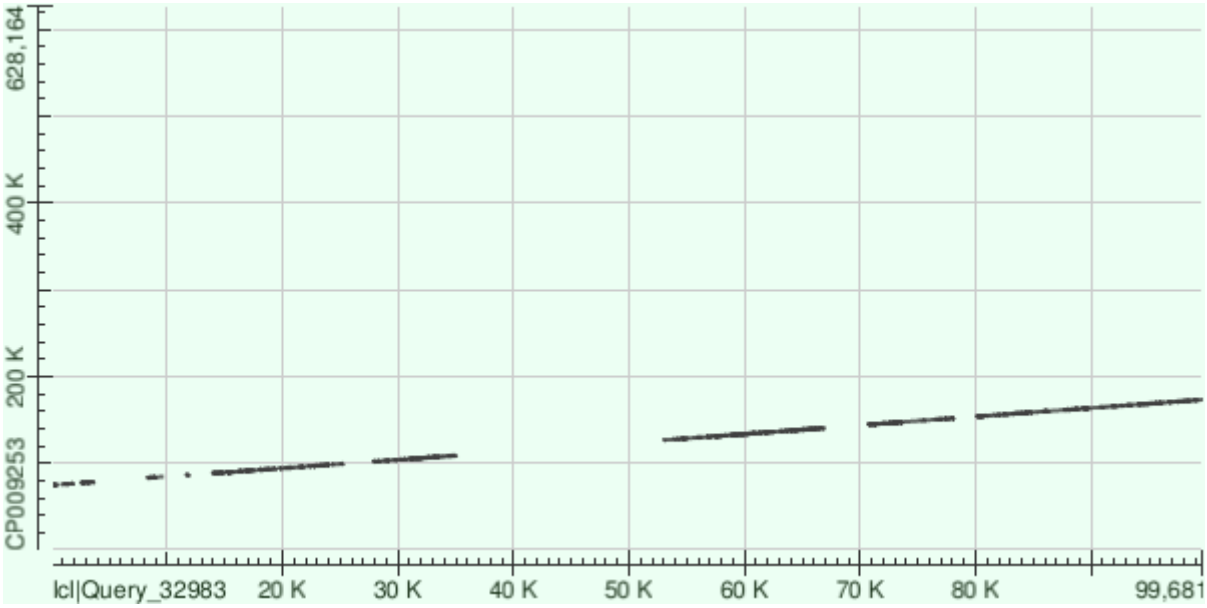

Рисунок 7. Dot-plot выравнивания 2 контига (query) на хромосому (subject)

Вторые по длине контиги обеих программ описывают приблизительно тот же участок хромосомы – рисунки 4 и 7. Центральный разрыв в сборке SPAdes уменьшился, но зато в целом число областей, для которых не была обнаружена схожесть, увеличилось.

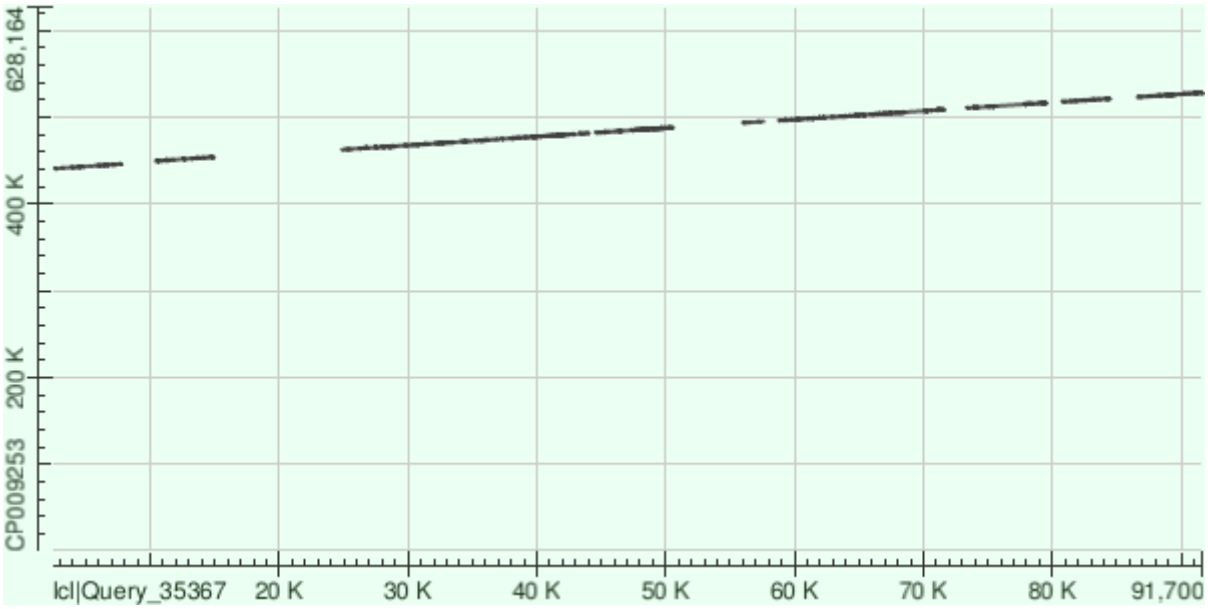

Рисунок 8. Dot-plot выравнивания 3 контига (query) на хромосому (subject)

Третий контиг у SPAdes получился хуже – он аналогичен самому длинному контигу сборки velvet, но короче него и по качеству тоже уступает – области, которых не удалось выравнять увеличились.

Данные контиги почему-то уже не все направлены в одинаковую сторону и самый длинный не совпадает с направлением референса. Но зато три самых длинных контига при выравнивании на геном снова не пересекаются. В целом, обе программы работают хорошо, каждая со своими недостатками и преимуществами.