Практикум 15

Сборка de novo

В этом практикуме мы работали с бактерией *Buchnera aphidicola str. Tuc7* (AC проекта: SRR4240356). Секвенирование проводилось по технологии Illumina.

Чтобы скачать архив с чтениями, воспользовались командой wget:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/006/SRR4240356/SRR4240356.fastq.gz

Далее мы, используя программу fastqc визуализировали качество наших прочтений:

fastqc SRR4240356.fastq.gz

Из директории /mnt/scratch/NGS/adapters были скопированы адаптеры (их оказалось два) для одиночных чтений в один файл adapters-SE.fasta:
cat TruSeq2-SE.fa > adapters-SE.fasta
cat TruSeq3-SE.fa >> adapters-SE.fasta

Затем была выполнена фильтрация чтения при помощи программы trimmomatic. Мы удалили адаптеры, нуклеотиды с правого конца с качеством ниже 20 и оставили чтения длиной не меньше 32):
java -jar /usr/share/java/trimmomatic.jar SE -phred33 SRR4240356.fastq.gz SRR4240356_noadapters.fq.gz ILLUMINACLIP:adapters-SE.fasta:2:7:7
java -jar /usr/share/java/trimmomatic.jar SE -phred33 SRR4240356_noadapters.fastq.gz SRR4240356_out.fq.gz TRAILING:20 MINLEN:32

Снова использовали fastqc для визуализации качества чтений:
fastqc SRR4240356_out.fq.gz

В результате после фильтрации чтений выяснилось, что 152380 последовательностей (2.03%) являлись остатками адаптеров.

Таблица 1. Результаты до и после тримирования

Параметры	До тримирования	После удаления адаптеров	После второго тримирования
Размер файла (AC.fastq.gz), в байтах	174262033 байт (174,26 Mb)	171398665 байт (171,4 Mb)	162162102 байт (162,22 Mb)
Количество чтений	7511529	7359149	7059570
Длина чтений	36	1-36	32-36

Далее была запущена программа velveth с параметрами для коротких одноконцевых чтений, параметр hash_length, подготовка k-меров k=31
velveth velveth 31 -short -fastq SRR4240356_out.fq.gz

Запуск программы velvetg со следующими параметрами для сборки *de novo* на основе заданных k-меров:
velvetg velveth

По итогу мы получили информацию о N50 (65554), а еще файлы с контигами и их характеристиками.

Далее использовались следующие команды:
sort -n -r -k 2 stats.txt | head - информация о самых больших контигах (таблица 2)
sort -n -k 6 -r stats.txt | head -n 3 - информация о контигах с аномально большим покрытием (таблица 3)
sort -n -k 6 stats.txt | head -n 3 - информация о контигах с аномально малым покрытием (таблица 4)

Таблица 2. Информация о трех самых длинных контигах

ID	Длина	Покрытие	Файл с последовательностью
8	111962	38.668870	contig8
6	107488	34.195585	contig6
10	80939	37.546325	contig10

Таблица 3. Информация о контигах с аномально большим покрытием

ID	Длина	Покрытие	Файл с последовательностью
64	1	266957.0	-
129	1	1134.0	-
28	282	458.432624	contig28

Из таблицы можно увидеть, что встречаются контиги с длиной всего 1, т.е. они имеют лишь 31 нуклеотид. При этом у них аномально большое покрытие. Может быть, что это шум.

Таблица 4. Информация о контигах с аномально малым покрытием

ID	Длина	Покрытие	Файл с последовательностью
251	3	1.0	-
253	2	1.0	-
274	1	1.0	-

Если говорить о контигах с малым покрытием, то у нас получился один с покрытием 1 и длиной 1, контиги, с длинами 2 и 3 также имеют аномально малое покрытие.

Файлы с контигами:
.\contig6.txt contig6
.\contig8.txt contig8
.\contig10.txt contig10
.\contig28.txt contig28

Анализ (Megablast)

На сайте NCBI запускаем уже знакомый нам megablast. Вводим две последовательности: наш контиг и геном бактерии *Buchnera aphidicola* (GenBank/EMBL AC — CP009253). Сделаем анализ каждого выравнивания. Сравнение процента идентичности представлено в таблице 5. Сравнения характеристик выравниваний для каждого контига (количество гэпов, однонуклеотидных замен, участок генома бактерии и участок нашего контига, вошедших в выравнивание) представлены в таблицах 6-8, соответственно.

Таблица 5. Сравнение процента идентичности

ID контига	E-value	Per.Identity	Контиг
8	0.0	81.46%	contig8
6	0.0	78.76%	contig6
10	0.0	74.88%	contig10

Contig 6

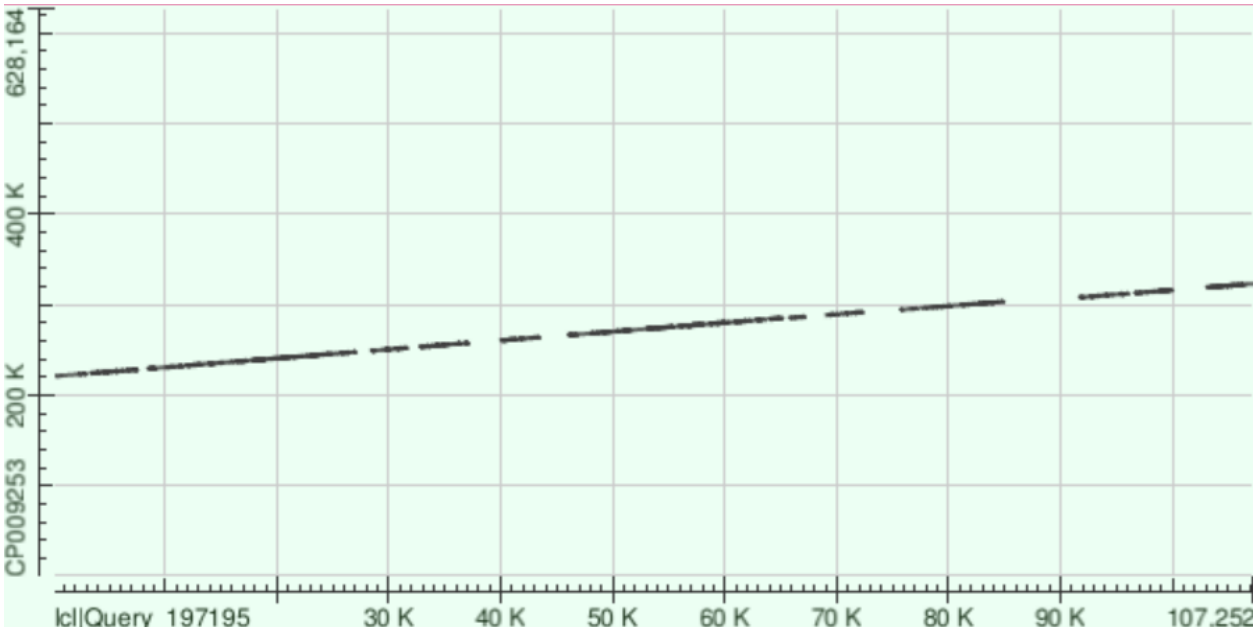

Таблица 6. Характеристика выравнивания для contig6

Координаты участка генома	Координаты участка контига	Число гэпов	Число однонуклеотидных различий
220869-223720	146-2996	19	483
224057-228137	3385-7496	163	799
228944-232057	8396-11516	97	573
232358-236859	11665-16194	130	985
236918-247596	16292-26990	390	2272
248967-252161	28467-31669	94	625
253244-257546	32780-37082	192	978
260224-263784	39869-43440	111	717
266073-275551	45989-55468	363	1689
275566-283706	55527-63756	421	1579
283963-285070	64004-65113	46	422
285200-286535	65810-67144	27	295
288181-291560	68934-72299	98	671
294227-295755	75721-77247	14	279
295935-303252	77556-84909	186	1547
307878-312179	91741-96052	120	889
312679-315982	96698-100006	89	681
318826-323043	103039-107252	174	950

Заметим, что между участками контига, которые выравнялись на наш геном произошло несколько делеций. Аналогичная ситуация, впоследствии, будет заметна и с двумя другими контигами (8, 10).

Contig 8

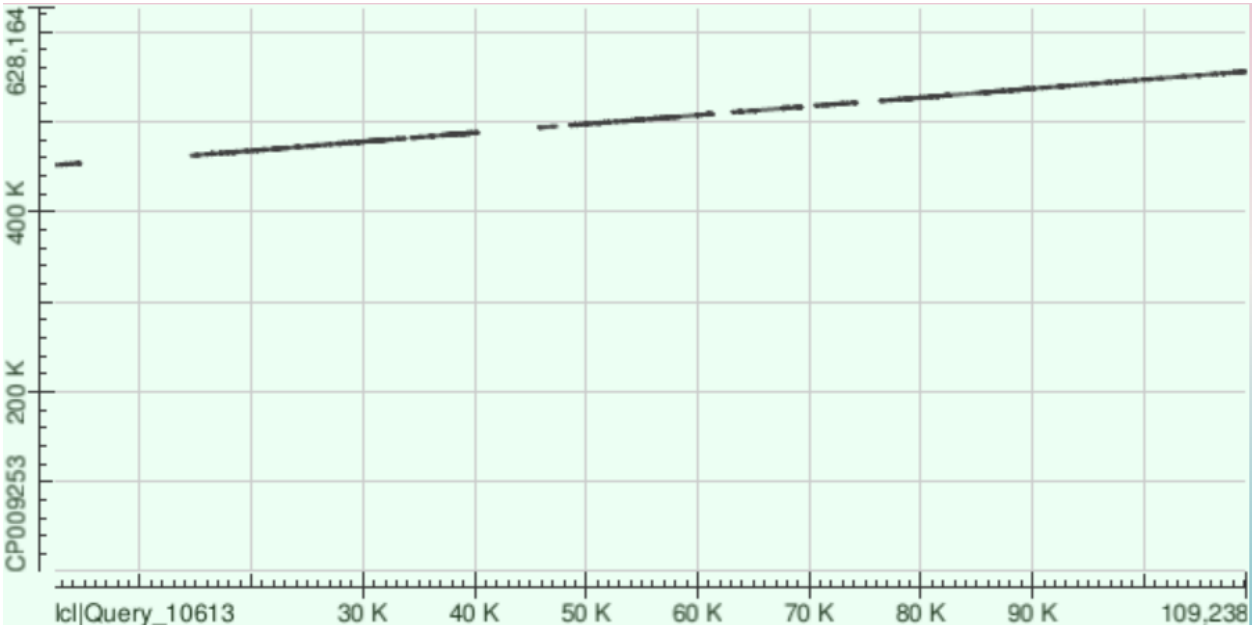

Таблица 7. Характеристика выравнивания для contig8

Координаты участка генома	Координаты участка контига	Число гэпов	Число однонуклеотидных различий
451729-454069	2390-4733	55	488
462496-467421	14624-19565	162	992
467412-474667	19595-26906	208	1489
474844-480660	27009-32884	255	1288
480874-481545	33090-33769	20	102
481997-488106	34243-40300	308	1309
493487-494864	45773-47149	13	262
495033-495148	47283-47401	5	7
496111-500325	48567-52845	154	914
500370-508806	52961-61406	351	1750
510438-516539	63097-69275	187	1150
517766-521500	70536-74265	99	763
523105-528679	76268-81855	207	1109
528794-550219	81925-103395	545	3211
550361-555905	103601-109238	133	950

Опять же, на DotPlot видно делеции.

Contig 10

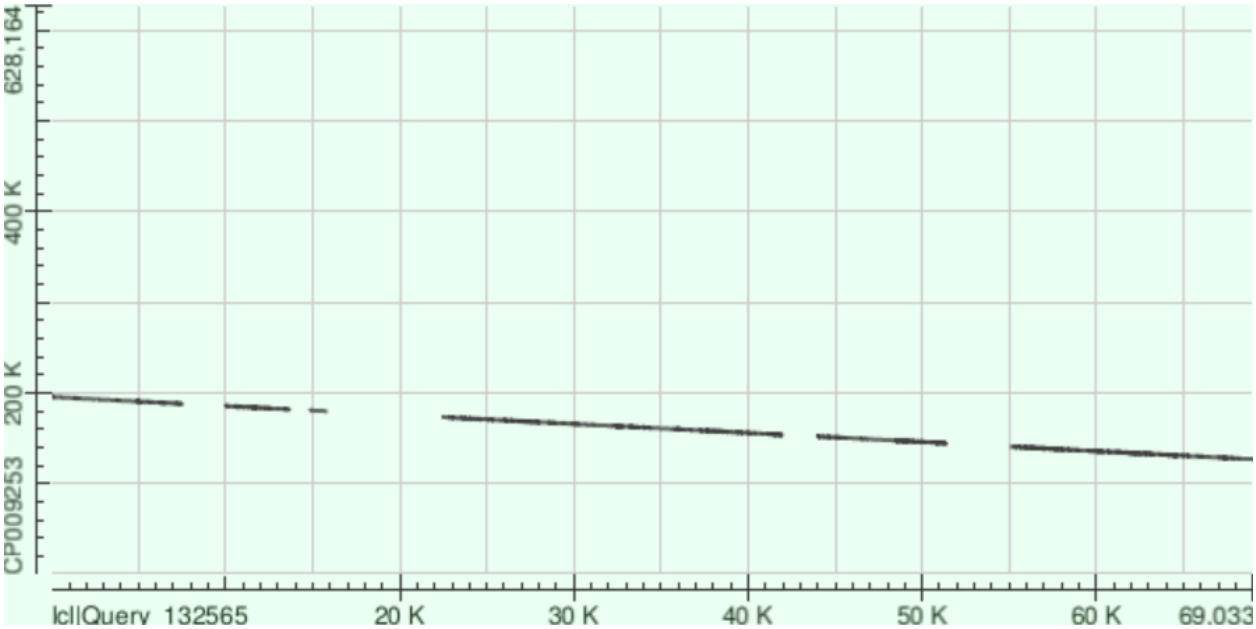

Таблица 8. Характеристика выравнивания для contig10

Координаты участка генома	Координаты участка контига	Число гэпов	Число однонуклеотидных различий
126623-127815	67840-69033	11	184
127825-140555	55035-67775	544	2723
144368-151796	43997-51396	243	1430
153752-161738	33933-42017	266	1557
161898-166752	28867-33727	108	894
166750-173180	22393-28836	159	1393
179654-180620	14869-15834	1	144
181712-185328	10021-13675	112	774
187938-192665	2708-7482	99	859
192777-193984	1427-2632	4	222
194042-195400	37-1400	13	1121

На DotPlot снова видны делеции, а еще прямая идет в обратном направлении, т.е. можно сказать, что наш контиг записан в обратном порядке.