Все задания практикума 15 выполнялись в папке /mnt/scratch/NGS/yaz008/pr15

Практикум 15 – Сборка de novo

Триммирование

Я скачал архив с ридами по своему коду доступа SRR4240356

wget “ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR424/001/SRR4240356/SRR4240356.fastq.gz”

Далее я объединил все адаптеры в один файл

cat /mnt/scratch/NGS/adapters/*fa > adapters.fa

Затем при помощи программы TrimmomaticSE я убрал остатки адаптеров

TrimmomaticSE SRR4240356.fastq.gz trimmed_no_adapters.fq.gz ILLUMINACLIP:adapters.fa:2:7:7

Выходной файл содержит 7701762 рида

Далее я обрезал некоторые риды, удалив с их правых концов нуклеотиды с качеством ниже 20 (риды с длиной меньше 32 при этом удалялись полностью)

TrimmomaticSE trimmed_no_adapters.fq.gz trimmed.fastq.gz TRAILING:20 MINLEN:32

В итоге осталось 7376028 ридов, размер файла сократился с 174262033 байт до 162017494 байт (-7.56%)

Сборка

При помощи программы velveth я сделал k-меры (k=31)

velveth . 31 -short -fastq.gz ../reads/trimmed.fastq.gz

Затем при помощи программы velveth я собрал контиги
velvetg velv/ &> both.log

N50 = 65554

Далее я при помощи данной команды sort -nk2 stats.txt | tail -3 я получил 3 самых длинных контига:

ID	Length	Coverage
10	80939	37.524173
6	107488	34.174029
8	111962	38.660197

Затем для поиска контигов с аномальным покрытием я использовал команду

```
sort -nk6 stats.txt | less
```

Я обнаружил, что покрытия возрастают достаточно плавно от 1 до 458, однако 2 последних контига сильно выбивались

ID	Length	Coverage
127	1	1134
64	1	266951

Анализ

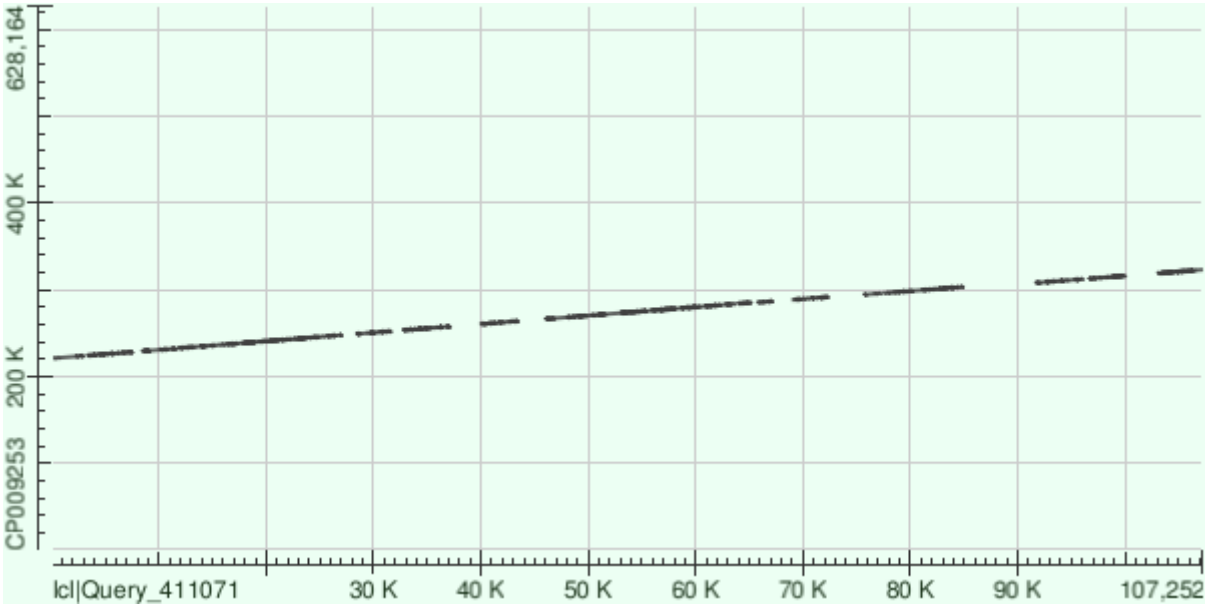
Я зашёл на сайт NCBI и мегабластировал 3 самых длинных контига на геном бактерии Buchnera aphidicola (GenBank/EMBL AC — CP009253)

ID	E-value	Per. Identity
6	0	78.76%
8	0	81.46%
10	0	74.88%

Между участками всех контигов, которые нормально выровнялись на наш геном, произошли несколько делеций.

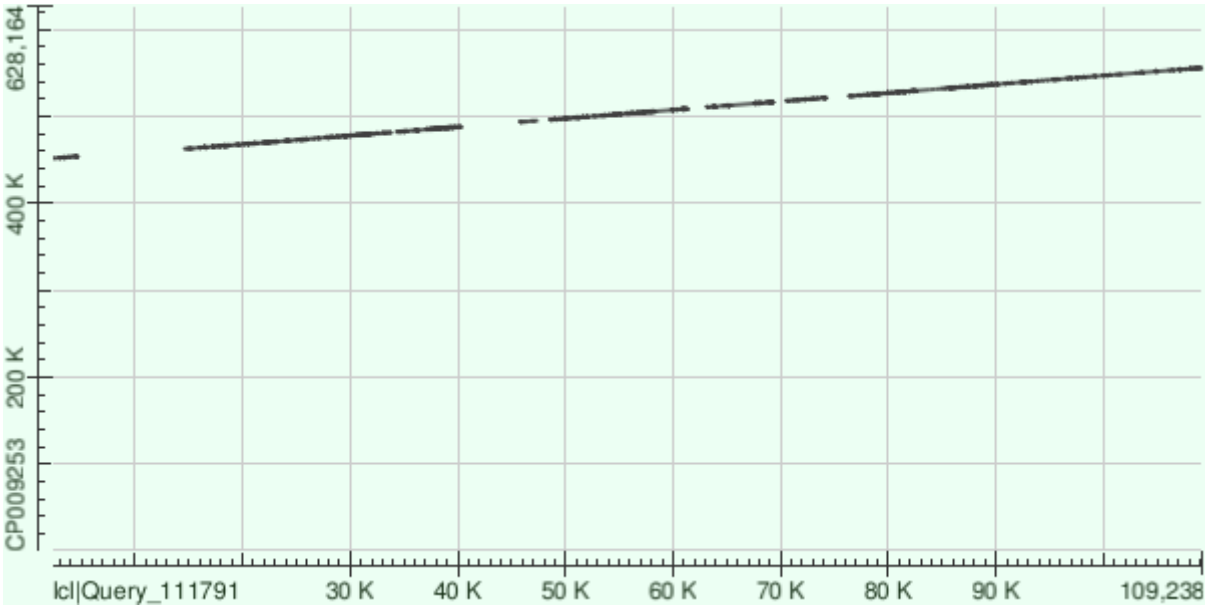
Также можно заметить, что на Dot Plot для 10 контига график выравнивания направлен вниз – это обзначает, что контиг записан в другом порядке.

Contig6



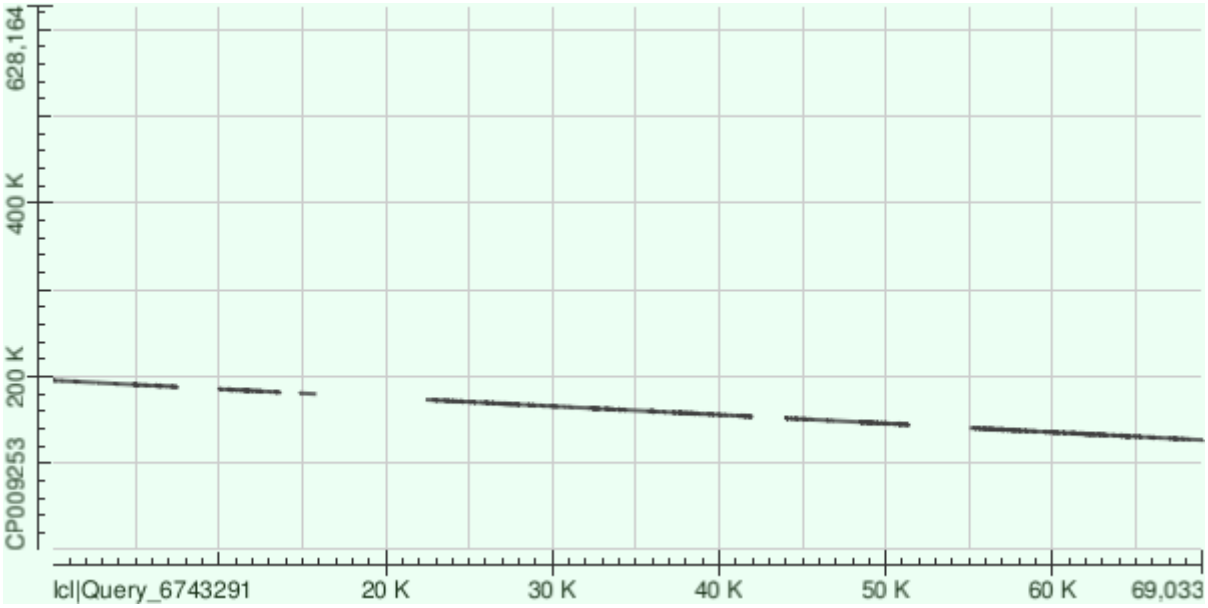
Координаты участка генома	Координаты участка контига	Gaps	Однонуклеотидные различия
220869-223720	146-2996	19	483
224057-228137	3385-7496	163	799
228994-232057	8396-11516	97	573
232358-236859	11665-16194	130	885
236918-247569	16292-26990	390	2272
248967-252161	28467-31669	94	625
253244-257546	32780-37082	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

Contig8



Координаты участка генома	Координаты участка контига	Gaps	Однонуклеотидные различия
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	154	914
496111-500325	48567-52845	351	1750
500370-508806	52961-61406	187	1150
510438-516539	63097-69275	99	763
517766-521500	70536-74265	207	1109
523105-528679	76268-81855	545	3211
528794-550219	81925-103395	133	950
550361-555905	103601-109238	55	488

Contig10



Координаты участка генома	Координаты участка контига	Gaps	Однонуклеотидные различия
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	112	144
181712-185328	10021-13675	99	774
187938-192665	2708-7482	13	859
192777-193984	1427-2632	11	222
194042-195400	37-1400	544	1121