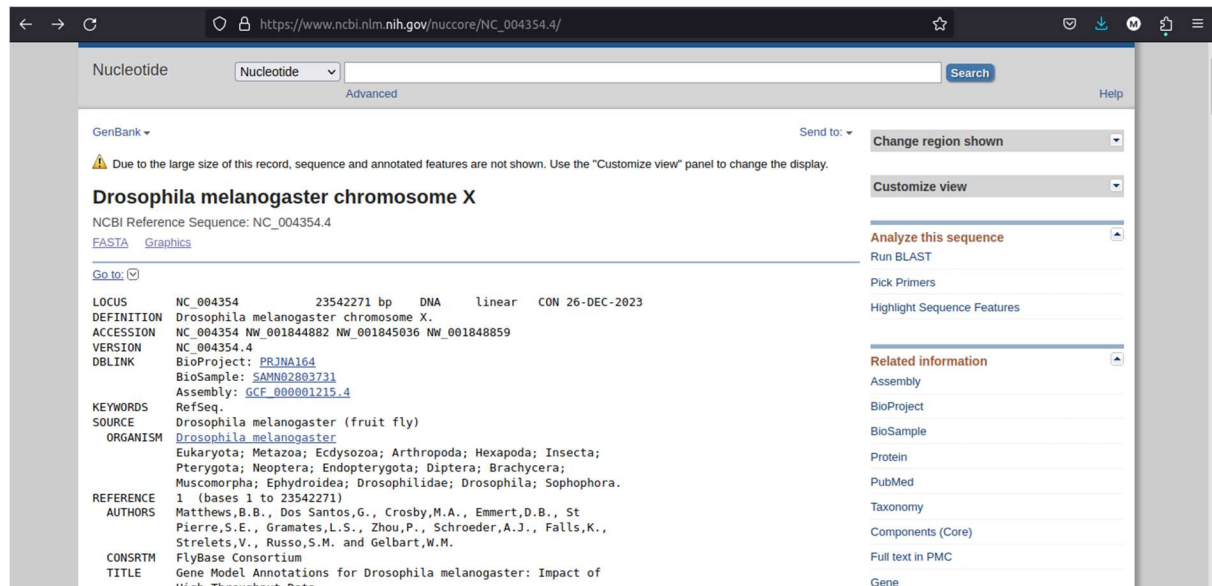


Загрузите fastq-файлы под идентификаторами SRR1663608, SRR1663609, SRR1663610, SRR1663611 из баз данных NCBI или ENA, проведите контроль качества, тримминг, повторите контроль качества.

Я так поняла, что изначальные данные уже без адаптеров, поэтому на всякий случай провела тримминг с помощью cutadapt -quality-cutoff=20 для удаления низкокачественных баз. Их не оказалось. В принципе, все показатели в пределах нормы, кроме Sequence Length Distribution и Overrepresented sequences.

Найдите соответствующий референсный геном в базах данных NCBI.

Подготовьте референсный геном.



The screenshot shows the NCBI Nucleotide database page for the Drosophila melanogaster chromosome X. The page includes a search bar at the top, a navigation menu on the left, and a main content area with the following information:

- Nucleotide** (dropdown menu)
- Advanced** (button)
- Search** (button)
- Help** (link)
- GenBank** (dropdown menu)
- Send to:** (dropdown menu)
- Change region shown** (dropdown menu)
- Customize view** (dropdown menu)
- Analyze this sequence** (button)
- Run BLAST** (button)
- Pick Primers** (button)
- Highlight Sequence Features** (button)
- Related information** (button)
- Assembly** (button)
- BioProject** (button)
- BioSample** (button)
- Protein** (button)
- PubMed** (button)
- Taxonomy** (button)
- Components (Core)** (button)
- Full text in PMC** (button)
- Gene** (button)

Due to the large size of this record, sequence and annotated features are not shown. Use the "Customize view" panel to change the display.

Drosophila melanogaster chromosome X

NCBI Reference Sequence: NC_004354.4

[FASTA](#) [Graphics](#)

[Go to:](#) (dropdown menu)

LOCUS NC_004354 23542271 bp DNA linear CON 26-DEC-2023

DEFINITION Drosophila melanogaster chromosome X.

ACCESSION NC_004354 NW_001844882 NW_001845036 NW_001848859

VERSION NC_004354.4

DBLINK BioProject: [PRJNA164](#)
BioSample: [SAMN02803731](#)
Assembly: [GCF_000001215.4](#)

KEYWORDS RefSeq.

SOURCE Drosophila melanogaster (fruit fly)

ORGANISM [Drosophila melanogaster](#)
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.

REFERENCE 1 (bases 1 to 23542271)

AUTHORS Matthews,B.B., Dos Santos,G., Crosby,M.A., Emmert,D.B., St
Pierre,S.E., Gramates,L.S., Zhou,P., Schroeder,A.J., Falls,K.,
Strelets,V., Russo,S.M. and Gelbart,W.M.

CONSRTH FlyBase Consortium

TITLE Gene Model Annotations for Drosophila melanogaster: Impact of
[Link Throughout Data](#)

Идентификаторы соответствуют виду Drosophila melanogaster, поэтому я нашла на NCBI Reference Sequence: NC_004354.4

Постройте выравнивание при помощи bwa.

Приветствие - hw6 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

ПРОВΟДНИК

- hw6 (РАБОЧАЯ ОБЛАСТЬ)
 - hw6
 - adapters.fasta
 - hw6.code-workspace
 - sequence.fasta
 - sequence.fasta.amb
 - sequence.fasta.ann
 - sequence.fasta.bwt
 - sequence.fasta.pac
 - sequence.fasta.sa
 - SRR1663608_fastqc.html
 - SRR1663608_fastqc.zip
 - SRR1663608_trimmed_fastqc.html
 - SRR1663608_trimmed_fastqc.zip
 - SRR1663608_trimmed.fastq
 - SRR1663608.fastq
 - SRR1663608.fastq.gz
 - SRR1663608.sam
 - SRR1663609_fastqc.html
 - SRR1663609_fastqc.zip
 - SRR1663609_trimmed_fastqc.html
 - SRR1663609_trimmed_fastqc.zip
 - SRR1663609_trimmed.fastq
 - SRR1663609.fastq
 - SRR1663609.fastq.gz
 - SRR1663609.sam

СТРУКТУРА

ВРЕМЕННАЯ ШКАЛА

ТЕРМИНАЛ

```
[M::process] read 100154 sequences (10000160 bp)...
[M::mem_process_seqs] Processed 100112 reads in 10.612 CPU sec, 10.406 real se
c
[M::process] read 100114 sequences (10000047 bp)...
[M::mem_process_seqs] Processed 100154 reads in 14.266 CPU sec, 14.069 real se
c
[M::process] read 100206 sequences (10000040 bp)...
[M::mem_process_seqs] Processed 100114 reads in 26.662 CPU sec, 26.470 real se
c
[M::process] read 100304 sequences (10000086 bp)...
[M::mem_process_seqs] Processed 100206 reads in 20.652 CPU sec, 20.440 real se
c
[M::process] read 100380 sequences (10000010 bp)...
[M::mem_process_seqs] Processed 100304 reads in 8.700 CPU sec, 8.489 real sec
[M::process] read 100242 sequences (10000086 bp)...
[M::mem_process_seqs] Processed 100380 reads in 9.360 CPU sec, 9.167 real sec
[M::process] read 100254 sequences (10000194 bp)...
[M::mem_process_seqs] Processed 100242 reads in 8.223 CPU sec, 8.024 real sec
[M::process] read 100250 sequences (10000083 bp)...
[M::mem_process_seqs] Processed 100254 reads in 8.884 CPU sec, 8.677 real sec
[M::process] read 100420 sequences (10000177 bp)...
[M::mem_process_seqs] Processed 100250 reads in 8.676 CPU sec, 8.448 real sec
[M::process] read 100450 sequences (10000023 bp)...
[M::mem_process_seqs] Processed 100420 reads in 9.494 CPU sec, 9.303 real sec
[M::process] read 100354 sequences (10000004 bp)...
[M::mem_process_seqs] Processed 100450 reads in 9.598 CPU sec, 9.395 real sec
[M::process] read 100358 sequences (10000047 bp)...
[M::mem_process_seqs] Processed 100354 reads in 8.202 CPU sec, 8.013 real sec
[M::process] read 100514 sequences (10000019 bp)...
[M::mem_process_seqs] Processed 100358 reads in 8.789 CPU sec, 8.574 real sec
[M::process] read 32284 sequences (3212212 bp)...
[M::mem_process_seqs] Processed 100514 reads in 9.525 CPU sec, 9.377 real sec
[M::mem_process_seqs] Processed 32284 reads in 3.230 CPU sec, 3.105 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem sequence.fasta SRR1663611.fastq
[main] Real time: 1577.796 sec; CPU: 1613.330 sec
nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/nasty/Deskt
op/Магистратура/биоинформатика/hw6$
```

Примените команду SortSam пакета GATK.

Примените команду MarkDuplicates пакета GATK.

Приветствие - hw6 (рабочая область) - Visual Studio Code

Файл Правка Выделение Вид Переход Выполнить Терминал Справка

ПРОВΟДНИК

- hw6 (РАБОЧАЯ ОБЛАСТЬ)
 - hw6
 - adapters.fasta
 - hw6.code-workspace
 - picard.jar
 - sequence.fasta
 - sequence.fasta.amb
 - sequence.fasta.ann
 - sequence.fasta.bwt
 - sequence.fasta.pac
 - sequence.fasta.sa
 - SRR1663608_fastqc.html
 - SRR1663608_fastqc.zip
 - SRR1663608_trimmed_fastqc.html
 - SRR1663608_trimmed_fastqc.zip
 - SRR1663608_trimmed.fastq
 - SRR1663608.bam
 - SRR1663608.dedup.bam
 - SRR1663608.fastq
 - SRR1663608.fastq.gz
 - SRR1663608.metrics.txt
 - SRR1663608.sam
 - SRR1663608.sorted.bam
 - SRR1663608.sorted.bam.bai
 - SRR1663609_fastqc.html
 - SRR1663609_fastqc.zip

СТРУКТУРА

ВРЕМЕННАЯ ШКАЛА

ТЕРМИНАЛ

```
freeMemory: 311009816; totalMemory: 770703360; maxMemory: 2059403264
INFO 2024-04-19 22:54:54 MarkDuplicates Will retain up to 64356352 dup
licate indices before spilling to disk.
INFO 2024-04-19 22:54:54 MarkDuplicates Traversing read pair informati
on and detecting duplicates.
INFO 2024-04-19 22:54:54 MarkDuplicates Traversing fragment informatio
n and detecting duplicates.
INFO 2024-04-19 22:54:56 MarkDuplicates Sorting list of duplicate reco
rds.
INFO 2024-04-19 22:54:56 MarkDuplicates After generateDuplicateIndexes
freeMemory: 763222376; totalMemory: 1286602752; maxMemory: 2059403264
INFO 2024-04-19 22:54:56 MarkDuplicates Marking 1903394 records as dup
licates.
INFO 2024-04-19 22:54:56 MarkDuplicates Found 0 optical duplicate clus
ters.
INFO 2024-04-19 22:54:56 MarkDuplicates Reads are assumed to be orde
red by: coordinate
INFO 2024-04-19 22:55:57 MarkDuplicates Written 10 000 000 records.
Elapsed time: 00:01:00s. Time for last 10 000 000: 60s. Last read positi
on: */*
INFO 2024-04-19 22:56:49 MarkDuplicates Written 20 000 000 records.
Elapsed time: 00:01:53s. Time for last 10 000 000: 52s. Last read positi
on: */*
INFO 2024-04-19 22:56:51 MarkDuplicates Writing complete. Closing inpu
t iterator.
INFO 2024-04-19 22:56:51 MarkDuplicates Duplicate Index cleanup.
INFO 2024-04-19 22:56:51 MarkDuplicates Getting Memory Stats.
INFO 2024-04-19 22:56:51 MarkDuplicates Before output close freeMemory
: 23830440; totalMemory: 31457280; maxMemory: 2059403264
INFO 2024-04-19 22:56:51 MarkDuplicates Closed outputs. Getting more M
emory Stats.
INFO 2024-04-19 22:56:51 MarkDuplicates After output close freeMemory:
23830440; totalMemory: 31457280; maxMemory: 2059403264
[Fri Apr 19 22:56:51 MSK 2024] picard.sam.markduplicates.MarkDuplicates done.
Elapsed time: 2,34 minutes.
Runtime.totalMemory()=31457280
(gatk env) nastyslav@nastyslav-In-555:/media/nastyslav/F29E58A99E5867DD/Users/
nasty/Desktop/Магистратура/биоинформатика/hw6$
```

Пришлите результаты команды:

```
samtools flagstat [полученный bam-файл]
```

И опишите их.

```
nasty/Desktop/Магистратура/биоинформатика/hw6$ samtools flagstat  
SRR1663608.dedup.bam
```

20297349 + 0 in total (QC-passed reads + QC-failed reads) **Общее количество прочтений, прошедших контроль качества (QC-passed reads) и не прошедших контроль качества (QC-failed reads).**

20189612 + 0 primary – **количество основных прочтений**

0 + 0 secondary **Количество вторичных выравниваний, которые могут быть получены для прочтения, имеющего несколько выравниваний.**

107737 + 0 supplementary **Количество дополнительных выравниваний, которые используются для представления разрывных выравниваний или очень длинных выравниваний.**

1903394 + 0 duplicates **Количество дубликатов, обнаруженных и удаленных программой Picard MarkDuplicates.**

1903394 + 0 primary duplicates **Количество дубликатов среди первичных выравниваний.**

5545894 + 0 mapped (27.32% : N/A) **Количество прочтений, которые были успешно выровнены на референсный геном (mapped), в процентном соотношении ко всем прочтениям**

5438157 + 0 primary mapped (26.94% : N/A) **Количество первичных выравниваний, которые были успешно выровнены на референсный геном, в процентном соотношении ко всем первичным выравниваниям.**

0 + 0 paired in sequencing **Количество пар прочтений, обработанных программой для секвенирования.**

0 + 0 read1

0 + 0 read2

0 + 0 properly paired (N/A : N/A) **Количество пар прочтений, которые правильно выровнены и ориентированы относительно друг друга.**

0 + 0 with itself and mate mapped **Количество пар прочтений, в которых оба члена пары выровнены на референсный геном.**

0 + 0 singletons (N/A : N/A) **Количество прочтений, которые выровнены без партнера в паре.**

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5) **Количество пар прочтений, в которых партнеры выровнены на разные хромосомы.**

```
nasty/Desktop/Магистратура/биоинформатика/hw6$ samtools flagstat  
SRR1663609.de  
dup.bam
```

20384968 + 0 in total (QC-passed reads + QC-failed reads)

20262940 + 0 primary

0 + 0 secondary

122028 + 0 supplementary

2092353 + 0 duplicates

```
2092353 + 0 primary duplicates
5819769 + 0 mapped (28.55% : N/A)
5697741 + 0 primary mapped (28.12% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
nasty/Desktop/Магистратура/биоинформатика/hw6$ samtools flagstat
SRR1663610.de
```

```
dup.bam
20471652 + 0 in total (QC-passed reads + QC-failed reads)
20377462 + 0 primary
0 + 0 secondary
94190 + 0 supplementary
1915057 + 0 duplicates
1915057 + 0 primary duplicates
5651108 + 0 mapped (27.60% : N/A)
5556918 + 0 primary mapped (27.27% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

```
nasty/Desktop/Магистратура/биоинформатика/hw6$ samtools flagstat
SRR1663611.de
```

```
dup.bam
19574893 + 0 in total (QC-passed reads + QC-failed reads)
19474128 + 0 primary
0 + 0 secondary
100765 + 0 supplementary
1782365 + 0 duplicates
1782365 + 0 primary duplicates
5459713 + 0 mapped (27.89% : N/A)
5358948 + 0 primary mapped (27.52% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
```

```
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Подумала, что вдруг надо было команду применить на другой файл .bam, поэтому вот еще вариант:

```
nasty/Desktop/Магистратура/биоинформатика/hw6$ samtools flagstat
SRR1663608.sorted.bam
20297349 + 0 in total (QC-passed reads + QC-failed reads)
20189612 + 0 primary
0 + 0 secondary
107737 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
5545894 + 0 mapped (27.32% : N/A)
5438157 + 0 primary mapped (26.94% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Самостоятельно изучите команды BaseRecalibrator и ApplyBQSR. Каких файлов Вам не хватает для того, чтобы применить их к исследуемым файлам? Можете ли Вы найти их в Сети? Можете ли Вы найти аналогичные файлы для Homo sapiens? Приложите ссылки к ним в отчет (сами файлы велики, их не следует загружать без надобности).

Не хватает файлов с данными о качестве баз (Base Quality Score Recalibration Files). Эти файлы используются для обучения модели и калибровки базовых качеств. Они включают в себя файлы с известными вариантами (Known Sites) и файлы, созданные BaseRecalibrator'ом во время предыдущего этапа анализа.

Пример ссылок для скачивания файлов Known Sites для Homo sapiens:

(<https://console.cloud.google.com/storage/browser/gatk-best-practices/somatic-hg38>)

<https://gnomad.broadinstitute.org/downloads>

https://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/