Анализ последовательностей полного генома *Drosophila melanogaster*

Руководитель: Геннадий Захаров (EPAM Systems, Lifesciencs department)

Исполнитель: Анна Намятова

Введение

- Модельный объект
- Результаты изучения генома дрозофилы важны для понимания работы схожих генов у млекопитающих, в том числе, человека
- Линия agnostic потенциально может использоваться для изучения синдрома Вильямса у человека

Исходные данные

Три линии дикого типа:

- Canton-S (CS)
- Berlin (Ber)
- Oregon-R (Or)

Две линии с нарушениями в строении и функционировании нервной системы

- $agnostic^{ts3}$(ts3)
- $agnostic^{x1}$(X1)

Цель и задачи

Цель: Проверить геномы на наличие мутаций, ответственных за формирование мутантного фенотипа

Задачи (план работы):

- проверить качество ридов
- сделать пайплайн для нахождение мутаций в Snakemake
- сравнить геномы дикого и мутантного типов
- проверить, есть ли мутации, ответственные за формирование определенных признаков

Все стадии включаются в общий пайплайн в виде файла Snakefile

Snakefile

```
REFERENCE = "Drosophila_melanogaster.BDGP6.dna_sm.toplevel.fa"
WILDSTRAINS = ["GER1800140_Cs_LP180329001.g", "GER1800136_Ber_LP180329001.g", "GER1800138_OR_LP180329001.g"]


rule mergewt:
    input:
      expand('{sample}.vcf.gz', sample=WILDSTRAINS)
    output:
      "wildstrains.vcf.merged.gz"
    shell:
      #"/mnt/c/Users/Anya/Documents/Miniprojects/vcftools/bin/vcf-merge {input} | bgzip > {output}"
      "vcf-merge {input} | bgzip > {output}"

rule gatk:
    input:
        ref="Drosophila_melanogaster.BDGP6.dna_sm.toplevel.fa",
        bam="{sample}_picard.bam",
        bai="{sample}_picard.bam.bai"
    output:
        "{sample}.g.vcf.gz"
    threads: 2
    shell:
        #"gatk --java-options -Xmx3g HaplotypeCaller -R {input.ref} -I {input.bam} -O {output} -ERC GVCF"
        "gatk --java-options -Xmx3g HaplotypeCaller -R {input.ref} -I {input.bam} -O {output}"

rule picard:
    input:
        "{sample}_sorted.bam"
    output:
        "{sample}_picard.bam"
    threads: 4
    shell:
        #"java -jar /mnt/c/Users/Anya/Documents/picard.jar AddOrReplaceReadGroups I={input} O={output} RGID=1 RGLB=libraryname RGPL=illumina RGPU=unitna
        "picardtools AddOrReplaceReadGroups I={input} O={output} RGID=1 RGLB=libraryname RGPL=illumina RGPU=unitname RGSM=20"
```

FastQC - проверка последовательностей

Sanger/Illumina 1.9 35-76 bp

Линия	Количество ридов
Berlin	28494973
Oregon	30649092
Canton-S	26603406
agnostic ts3	29565269
agnostic X1	28110768

FastQC - проверка последовательностей

FastQC Report

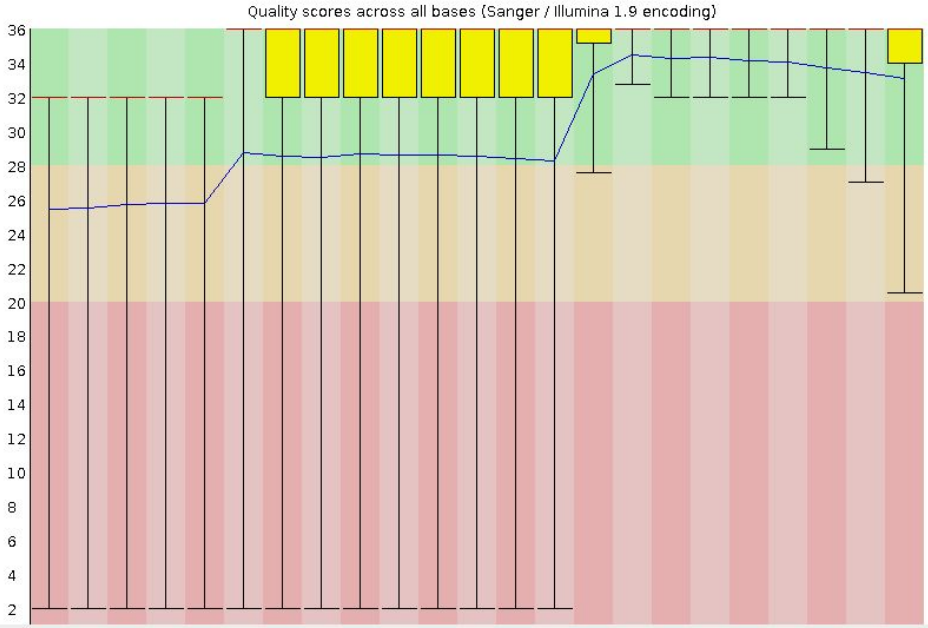
Линия Berlin

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Filename	GER1800136_Ber_LP180329001_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28494973
Filtered Sequences	0
Sequence length	35-76
%GC	41

Per base sequence quality



Produced by FastQC (version 0.10.1)

FastQC - проверка последовательностей

Overrepresented sequences: Линия Berlin

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Sequence Duplication Level

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNN

Trimmomatic: удаление некачественных сайтов

FastQC Report

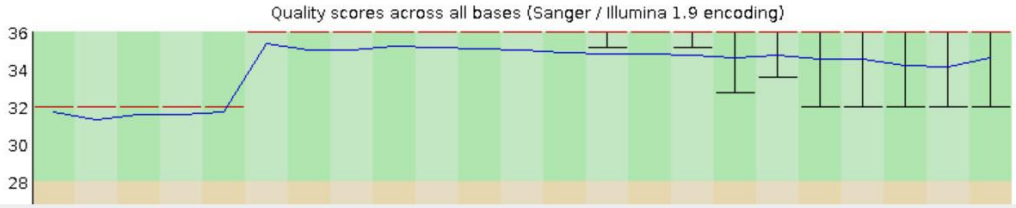
Thu 25 Oct 2018
GER1800136_Ber_LP180329001_R1_output_paired.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Measure	Value
Filename	GER1800136_Ber_LP180329001_R1_output_paired.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22651471
Filtered Sequences	0
Sequence length	20-76
%GC	41

Per base sequence quality



Produced by FastQC (version 0.10.1)

Выравнивание и индексирование

bwa - сборка ридов в формате .sam

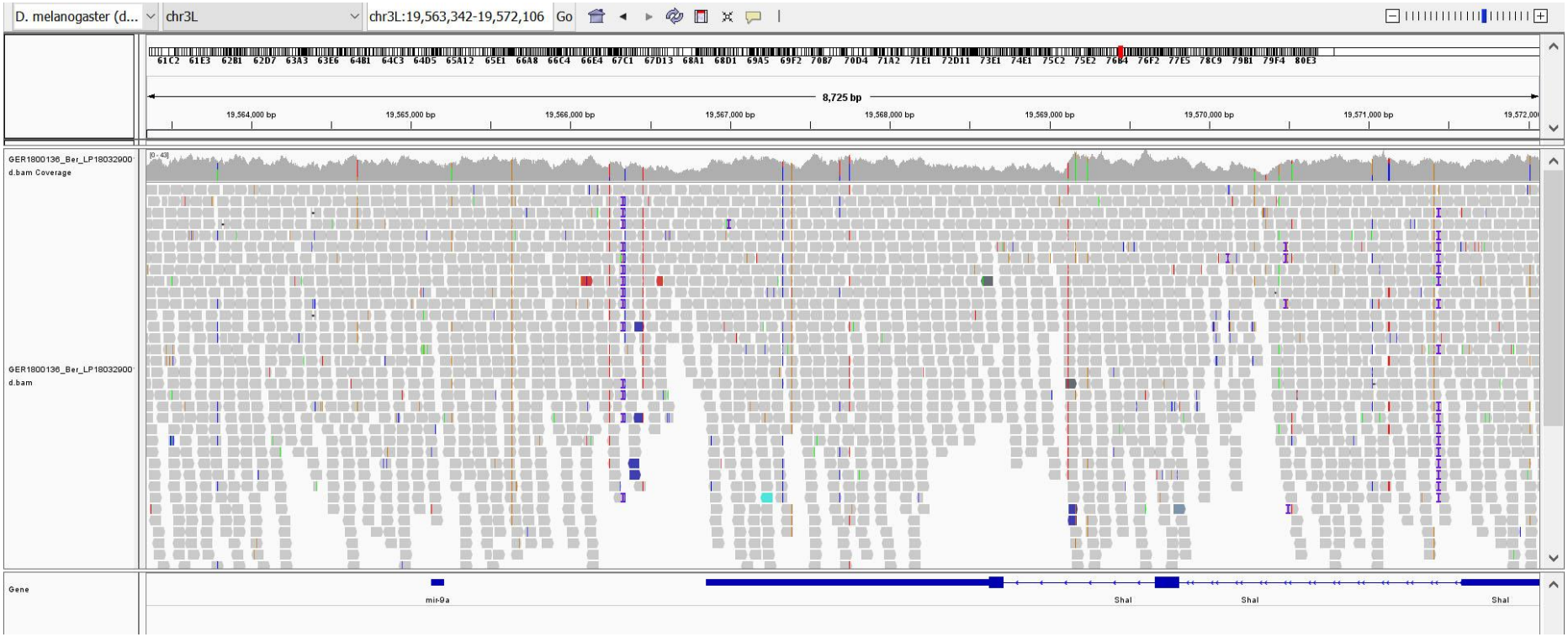
samtools - создание сжатого файла в формате .bam

samtools sort - сортировка .bam файла

picard - преобразование .bam файла - добавление ReadGroup

samtools index - индексирование .bam файла

IGV: Визуализация геномов



Линия Ber

Поиск вариаций

gatk - поиск вариаций для каждой линии

vcf-merge - слияние вариаций диких линий

Поиск вариаций

Подсчет вариаций в файле .vcf.gz с помощью команды
zcat {lineage}.g.vcf.gz | grep -v '#' | wc -l

Общее количество вариаций для диких линий:
1209056

Количество вариаций для отдельных диких линий:
Ber: 792356
Or: 926503
Cs: 954491

Количество вариаций для мутантных линий:
X1: 802664
ts: 807566

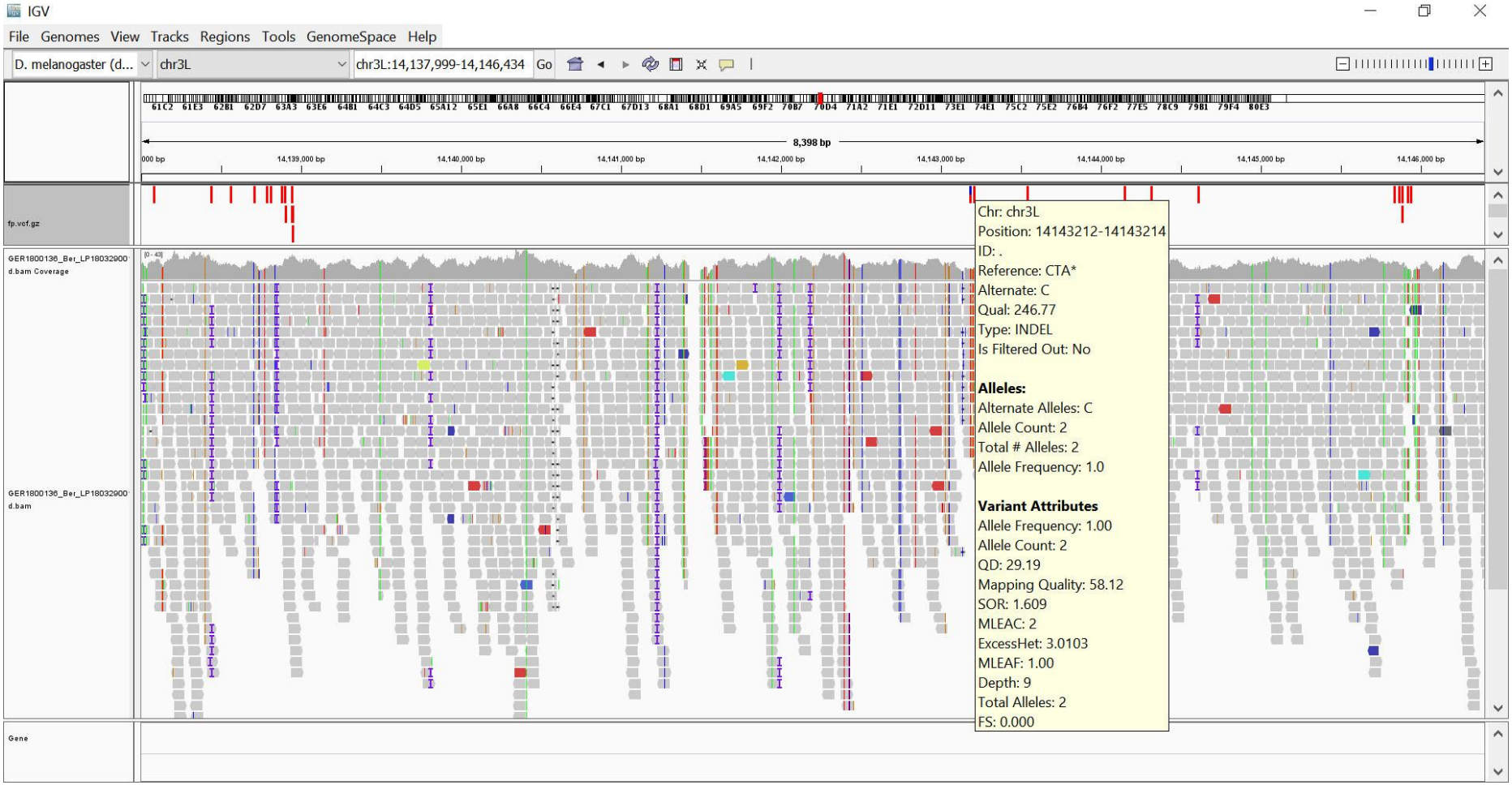
Сравнение диких и мутантных линий

rtg vcfeval - сравнение мутаций найденных в диких линия с мутациями из каждой мутантной линии (ts3, X1)

Результаты для ts3
195743 уникальных мутаций

Результаты для X1
175467 уникальных мутаций

Уникальные мутации для диких линий



Мутации для ts3

Что дальше?

snpeff, snpsift - аннотирование мутаций, поиск незначащих мутаций, приписывание биологического эффекта