Картирование и анализ геномных вариаций

Антон Александров

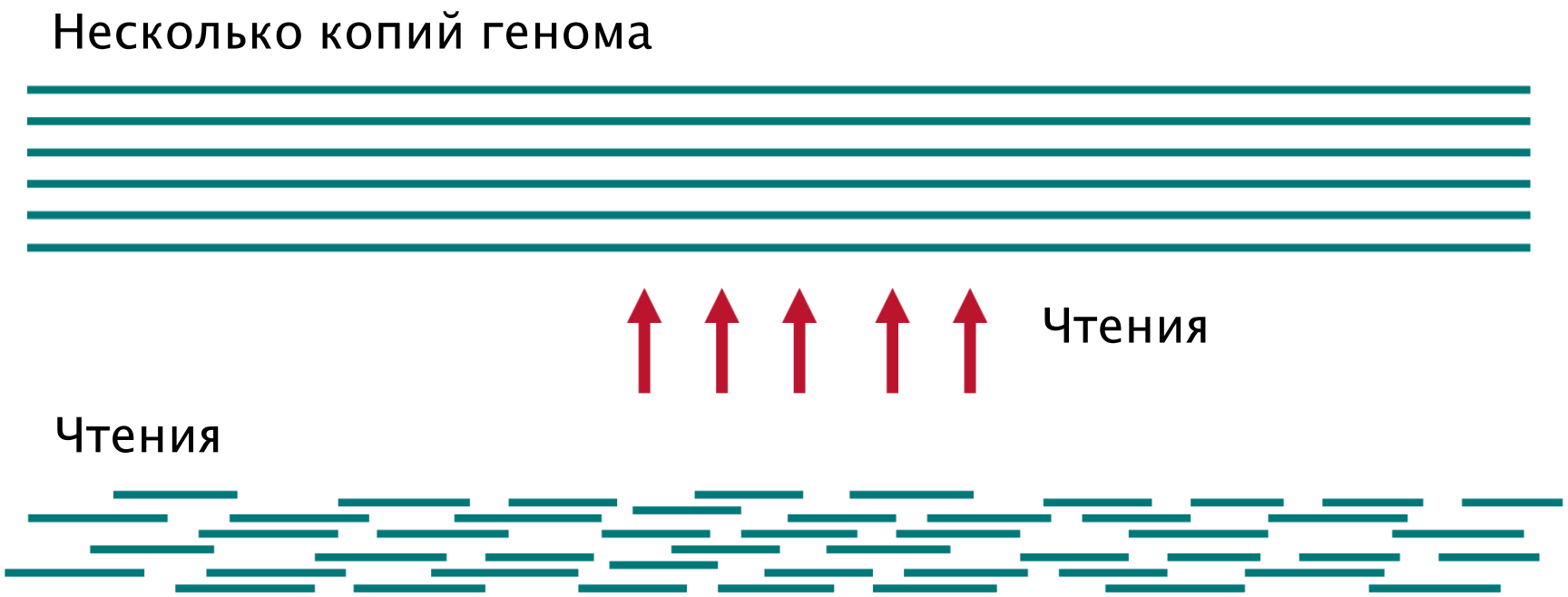
Входные данные (1)

- Чтения
  - fastq
  - sff
- Референсный геном
  - Контиги или скэффолды
  - fasta

Входные данные (2)

- Аннотация генома
  - Позиции и длины генов
  - gff3
  - Bed
  - gtf

Картирование чтений на геном (1)



Картирование чтений на геном (2)

- Не та особь
  - Геномные вариации
    - SNPs, короткие инделы
    - Структурные вариации
- Ошибки секвенирования
  - Ошибки вставки
  - Ошибки удаления
- Ошибки сборки
  - Мизассемблы

Картирование чтений на геном (3)

- bowtie, bowtie2
- bwa
- Blast
- TMAP
- …

Индекс генома

- Индексный файл
- bowtie2

Давайте построим индекс

- `bowtie2-build buchnera.fasta prefix`

Bowtie2. Базовые параметры

- bowtie2
  - -p threads
  - -I min_insert_size –X max_insert_size
  - -x index_prefix
  - -1 reads_1.fastq -2 reads_2.fastq | -U reads.fastq
  - -S reads.sam

Давайте запустим bowtie2

- `bowtie2 -p 3 -X 1000 -x prefix -1 ~/work/buchnera_1.fastq -2 ~/work/buchnera_2.fastq -S reads.sam`

Bowtie2. Дополнительные параметры

- bowtie2
  - --no-mixed
  - --no-discordant

Bowtie2. Дополнительные параметры

- bowtie2
  - --ma – match bonus (2)
  - --mp – mismatch penalty (6)
  - --rdg – read gap open, extend penalty (5, 3)
  - --rfg – reference gap open, extend penalty (5, 3)
  - --score-min (L, -0.6, -0.6)

SAM-файл

- Для каждого чтения:
  - Нуклеотидная последовательность
  - Качество для каждого нуклеотида
  - Позиция картирования
  - Качество картирования

BAM-файл

- Бинарный аналог SAM-файла
- Создан для ускорения обработки

samtools. Базовые параметры

- samtools команда аргументы
  - view
  - sort
  - index
  - tview
  - faidx
  - mpileup

samtools view

- samtools view выведет список параметров
- -S: input is SAM
- -u: uncompressed BAM output
- -q: minimum mapping quality
- `samtools view –Su reads.sam > reads.bam`

samtools sort

- samtools sort выведет список параметров
- `mv reads.bam reads.unsorted.bam`
- `samtools sort reads.unsorted.bam reads`
- Появится файл reads.bam

samtools index

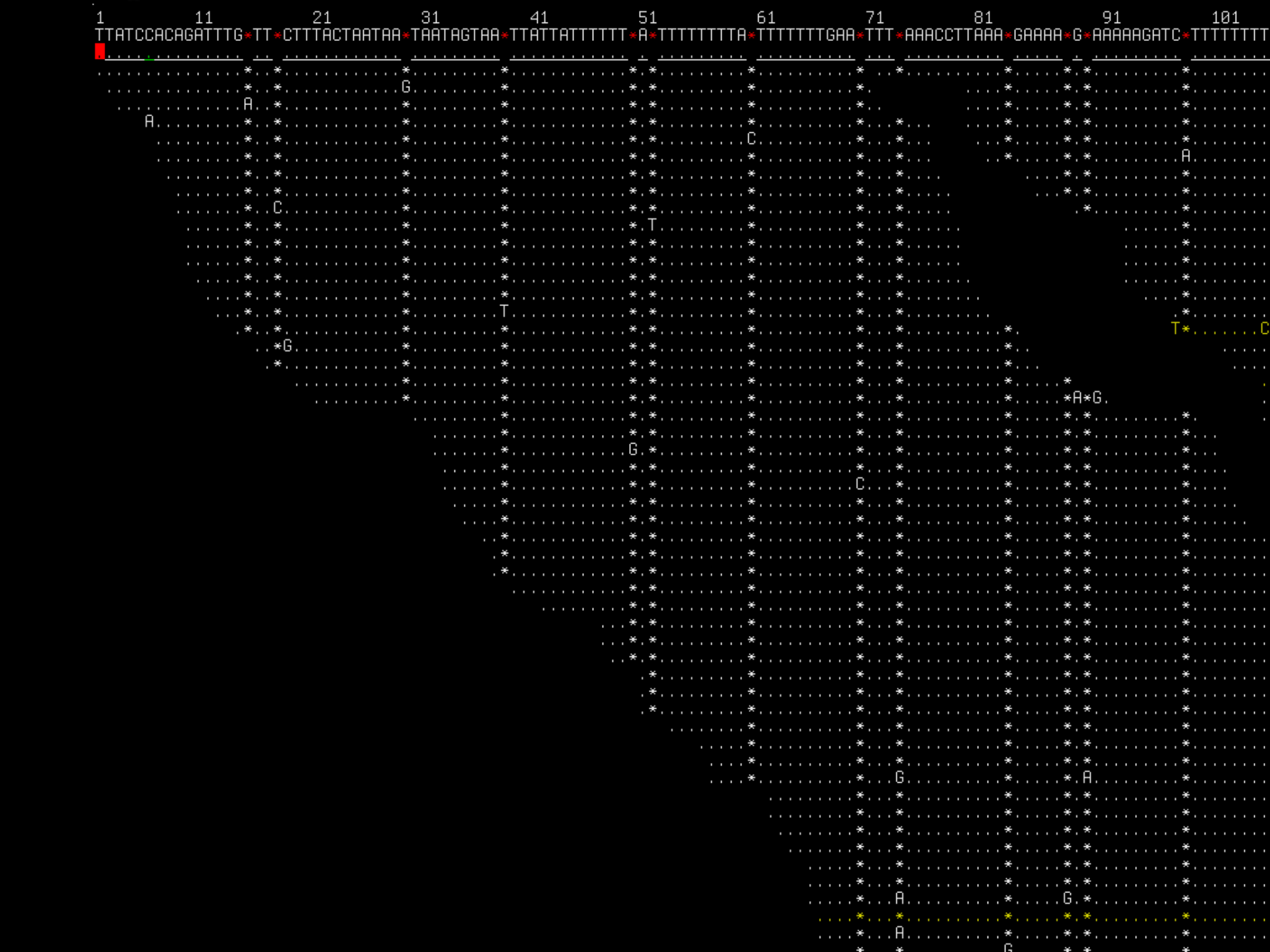
- samtools index выведет список параметров
- `samtools index reads.bam`
- Появится файл reads.bam.bai

Выделение вариаций (анонс)

- `samtools faidx buchnera.fasta`
- `samtools mpileup -uf buchnera.fasta reads.bam | bcftools view -cvg - > var.vcf`

samtools tview

- samtools tview выведет список параметров
- `samtools tview reads.bam buchnera.fasta`


1 11 21 31 41 51 61 71 81 91 101
TTATCCACAGATTTG*TT*CTTTACTAATAA*TAATAGTAA*TTATTATTTTTT*A*TTTTTTTA*TTTTTTTGAA*TTT*AAACCTTAAA*GAAAA*G*AAAAAGATC*TTTTTTTT

Integrative Genomics Viewer

- IGV
- http://www.broadinstitute.org/igv/
- Winscp
  - Копируем buchnera.fasta, reads.bam, reads.bam.bai
- Запускаем IGV

Integrative Genomics Viewer

Выделение вариаций (variant calling)

- samtools/bcftools
- GATK
  - http://www.broadinstitute.org/gatk/
- Ion Torrent Variant Caller

Индексация генома

- samtools faidx reference.fasta
- Создаст файл reference.fasta.fai

Выделение вариаций

- samtools mpileup –uf reference.fasta reads.bam | bcftools view –cvg - > var.vcf

Фильтрация вариаций (1)

- False positives vs false negatives
- gatk best practices
  - http://www.broadinstitute.org/gatk/guide/best-practices
- vcfutils.pl varFilter

Фильтрация вариаций (2)

- vcfutils.pl varFilter
  - -Q: minimum Root-Mean-Square quality for SNP
  - -d: minimum read depth
  - -D: maximum read depth
  - …

Аннотация вариаций

- annovar
- gene-talk.de – для человека
- …

Аннотация вариаций. Annovar

- Что такое аннотирование?
  - Добавление биологических знаний
- Не поддерживает явно vcf
- Предоставляет конвертер в свой формат

Конвертирование вариаций

- convert2annovar.pl -format vcf4 var.vcf > var.annovar

Аннотация. GFF

- ## comment
- seqname source feature start end …
- `head /data/tuc.gff`

Annovar

- mkdir db
- cp /data/tuc.gff db/
- annotate_variation.pl -regionanno
  -dbtype gff3 -gff3dbfile tuc.gff
  var.annovar db

Аннотированные мутации

- `head var.annovar.hg18_gff3`

Результаты (1)

- Мы научились:
  - Запускать команды
  - Смотреть по ним справку
  - Манипулировать различными типами файлов
  - Писать скрипты

Результаты (2)

- Мы можем:
  - Собирать и анализировать геномы
  - Получать и анализировать геномные вариации
  - **Создавать скрипты, делающие все это автоматически**