

# Сборка и аннотирование геномов

Антон Александров

Практическая школа по биоинформатике

МНЛ «Компьютерные технологии»

19.02.2014

# Входные данные

- Чтения (возможно, парные):
  - fastq
  - sff

Несколько копий генома



Чтения



# Выходные данные

- Контиги или скэффолды
  - fasta
- Аннотированные контиги
  - Список генов
  - Филогенетические данные

# ITMO assembler

- <http://genome.ifmo.ru/assembler>
- Консольная версия
  - Больше параметров
- Графическая версия
- Парные чтения
- Секвенатор-Который-Нельзя-Называть
- Ion Torrent
- Java => кроссплатформенность

# ITMO assembler. Базовые параметры

- `./itmo-assembler.sh -h`
- `-a` – длина якоря (по умолчанию: 19)
- `-k` – длина k-мера (по умолчанию: 19)
- `-i` – входные файлы
- `-w` – рабочая директория (по умолчанию: `workDir`)
- `-l`, `-L` – минимальный и максимальный размеры инсерта (по умолчанию: 0, 1000)

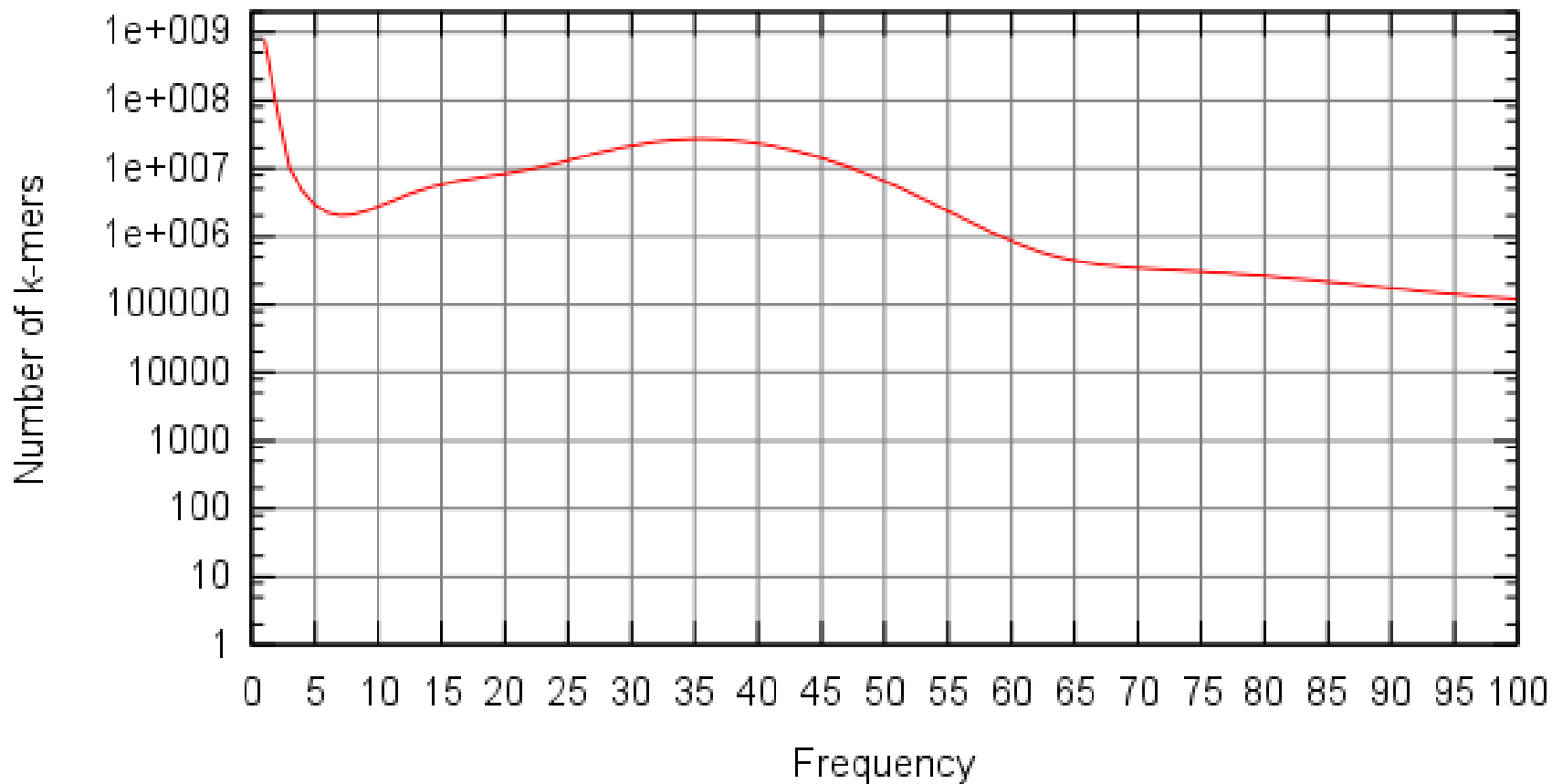
# Давайте запустим сборщик

- `./itmo-assembler.sh -i  
work/buchnera*fastq -w work/assembly`

# ITMO assembler. Дополнительные параметры

- -M – запустить микросборку (сложно)
- --orientations – список взаимных ориентаций парных чтений ([FR]) через пробел
- -b – максимальная частота ошибочного k-мера
- -u – порог качества для обрезания концов чтений
- -g – порог минимальная частота k-мера, используемого в сборке квазиконтигов
- -p – максимальное число потоков выполнения
- -m – максимальное количество используемой памяти (должен быть 1-ым параметром)

# ITMO assembler. Распределение частот k-меров





# ITMO assembler

 English  Русский

## 1. Choose input files

Add files

Change file path

Remove selected files

## 2. Set assembly options and parameters

Anchor length in error correction

19

Maximal errors

12

k-mer size

19

More

Working directory

workDir

Choose

Memory to use

5585M

Set

More

## 3. Run assembler

Start assembly

Stop assembly

## 4. Assembling status

Running stage:

Elapsed time:

Remaining time:

Overall progress:

Log (workDir/log):

# mira-assembler

- Другой сборщик
- Manifest-file
  - parameter = value
  - # comment
- Linux, MacOS

# mira-assembler. Общие параметры сборки

- project = <project name>
- job =
  - **genome/est**
  - **denovo/mapping**
  - **accurate/draft**
- parameters =
  - -NW:somrnl=0
  - -GE:not=3

# mira-assembler. Параметры библиотеки

- readgroup = id
- data = file1.fastq file2.fastq
- technology =
  - iontor
  - 454
  - solexa
  - sanger

# mira-assembler. Параметры библиотеки

- `template_size = min max exclusion_criterion`  
`autorefine`
- `segment_placing = ---> <---`
- *autopairing*

# Давайте запустим mira-assembler

- Создадим директорию `work/mira_assembly`
- Поместим в нее файл `manifest`
- Перейдем в нее
- Запустим `mira manifest`

# Другие сборщики

- SPAdes
  - <http://bioinf.spbau.ru/spades>
- CLC-bio
  - <http://www.clcbio.com/>
- ABySS
  - <http://www.bcgsc.ca/platform/bioinfo/software/abyss>

# Что делать с контигами?

- Оценка качества
- Аннотирование



# Оценка качества. Quast

- <http://quast.bioinf.spbau.ru>
- `quast.py -o dir -R /data/buchnera.fasta /data/buchnera_contigs.fasta`
- `nano dir/report.txt`

# Оценка качества. Quast.

- N50
- Maximal contig length
- # misassemblies
- Mismatches per 100kbp
- Indels per 100kbp
- Unaligned contigs
- Genome fraction

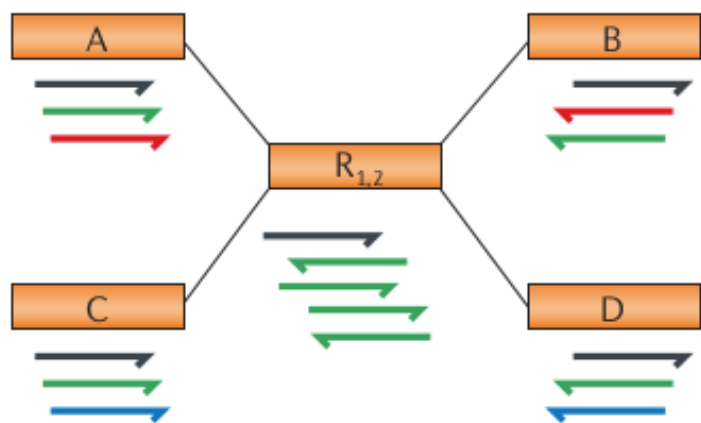
# Оценка качества. N50

- N«число» (NG«число») – такая длина контига, что контиги длины не меньше её составляют не меньше «число»% сборки (генома)

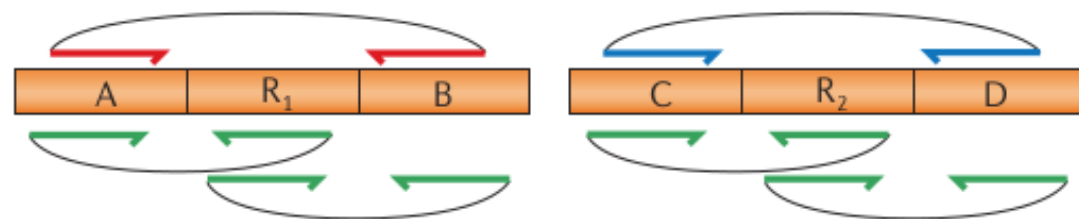
# Оценка качества. Misassemblies

- Нарушение парности чтений

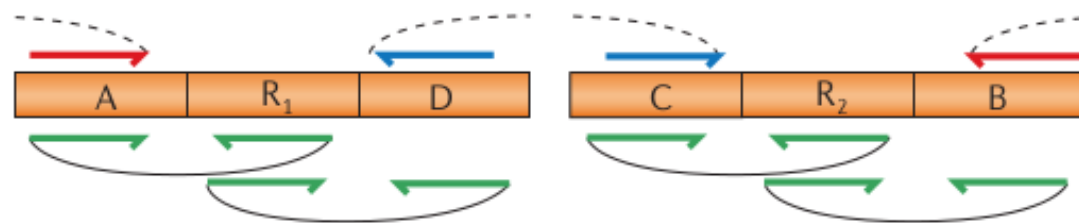
**Aa** Assembly graph



**Ab** Correct assembly



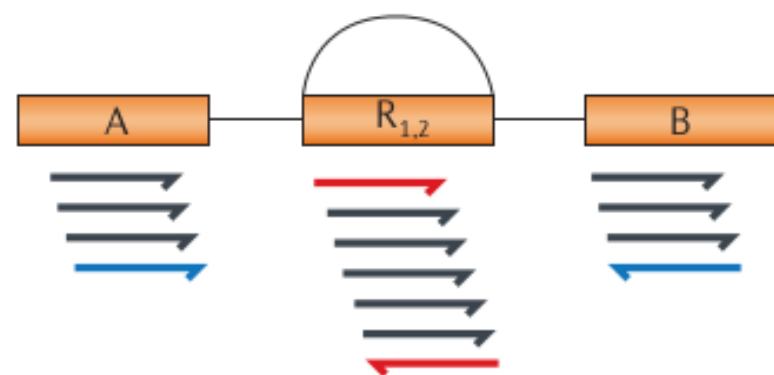
**Ac** Misassembly



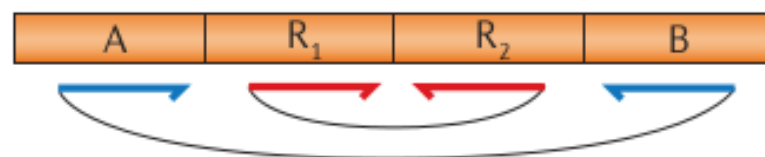
# Оценка качества. Misassemblies

- Нарушение ограничений на длину инсерта

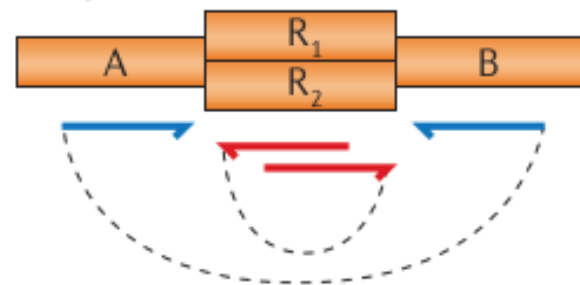
**Ba** Assembly graph



**Bb** Correct assembly



**Bc** Misassembly



# Аннотирование

- myRAST
- Входные данные – fasta-файл с контигами



# Аннотирование. myRAST

Feature ID: fig|6666667.3895.peg.1

Genome: 6666667.3895

Function: Dihydroorotate dehydrogenase (EC 1.3.3.1)

Region count: 10

Region size: 5000

[Compute genbank comparison](#)

[Show genbank comparison](#)

[Export data](#)

<Contig<

<<<

<<

<

>

>>

>>>

>Contig>

**6666667.3895**

**Buchnera aphidicola str. Sg (Schizaphis graminum)**



**Buchnera aphidicola str. APS (Acyrthosiphon pisum)**



**Buchnera aphidicola str. Sg (Schizaphis graminum)**



**Buchnera aphidicola str. 5A (Acyrthosiphon pisum)**



**Buchnera aphidicola str. Tuc7 (Acyrthosiphon pisum)**



**Buchnera aphidicola str. APS (Acyrthosiphon pisum)**

