СБМ

НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
СИСТЕМНОЙ БИОЛОГИИ И МЕДИЦИНЫ
РОСПОТРЕБНАДЗОРА

Введение в Linux

Данил Вадимович КРИВОНОС
Лаб. математической биологии и биоинформатики
НИИ СБМ Роспотребнадзора

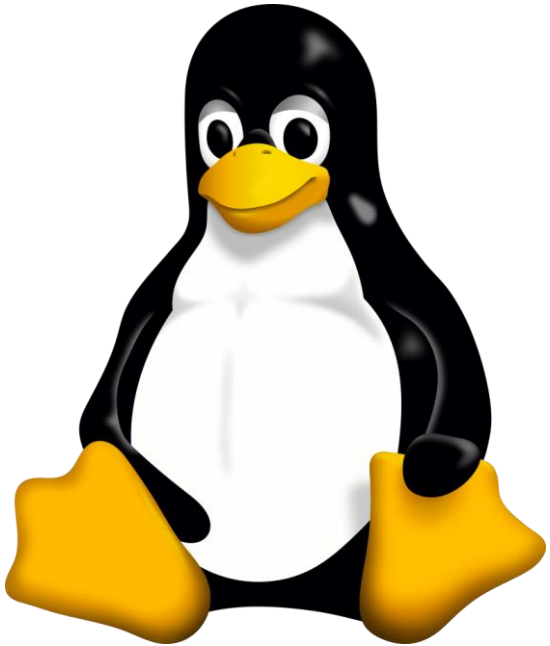
WWW.SYSBIOMED.RU

О чем мы сегодня поговорим и что будем делать?

1. Что такое Linux и где мы будем работать?
2. Зачем нам это и почему?
3. Что нужно уметь, чтобы было понятно работать?
4. Попрактикуемся 😳

Что такое Linux?

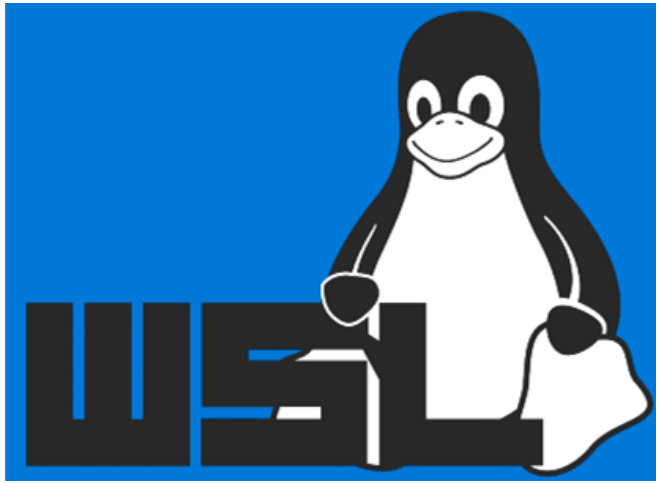
Linux — это семейство **операционных систем** (ОС), работающих на основе **одноименного ядра**. Нет одной операционной системы Linux, как, например, Windows или MacOS. Есть множество дистрибутивов (набор файлов, необходимых для установки ПО), выполняющих конкретные задачи.



Самой известной ОС на ядре Linux является Ubuntu.

Где мы будем работать?

Windows Subsystem for Linux (WSL) — слой совместимости для запуска Linux-приложений в ОС Windows 10 и выше. В рамках сотрудничества компаний Microsoft и Canonical стало возможным использовать оригинальный образ ОС **Ubuntu 14.04** для непосредственного запуска поверх WSL множества инструментов и утилит из этой ОС без какой-либо виртуализации.



bash ? Что это?

BASH — Bourne-Again SHell (что может переводится как «перерожденный шел», или «Снова шел Борна(создатель sh)»), самый популярный **командный интерпретатор** в юниксоподобных системах, в особенности в GNU/Linux.

Список базовых команд

Что нам нужно знать:

cd - команда для перехода между директориями.

ls - команда для просмотра файлов в папке.

cat - вывод содержимого файла на экран.

rm - удалить файл.

cp -скопировать файл.

mv - переместить файл.

mkdir - создать пустую директорию.

Что полезно знать:

less - чтение текста фрагментами (выйти из less “q”).

head - посмотреть заголовок текстового файла.

tail - посмотреть окончание текстового файла.

man - вывод справочной информации.

rmdir - удалить пустую директорию.

Полезно, но скорее всего, не пригодится

grep - поиск по регулярному выражению в текстовом файле.

wc - подсчет строк, слов и байт (выдает информацию в таком порядке).

zless - аналог less, но для упакованных текстовых файлов.

zcat - аналог less, но для упакованных текстовых файлов.


Практика …

СБМ

НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ
СИСТЕМНОЙ БИОЛОГИИ И МЕДИЦИНЫ
РОСПОТРЕБНАДЗОРА

Сборка геномов

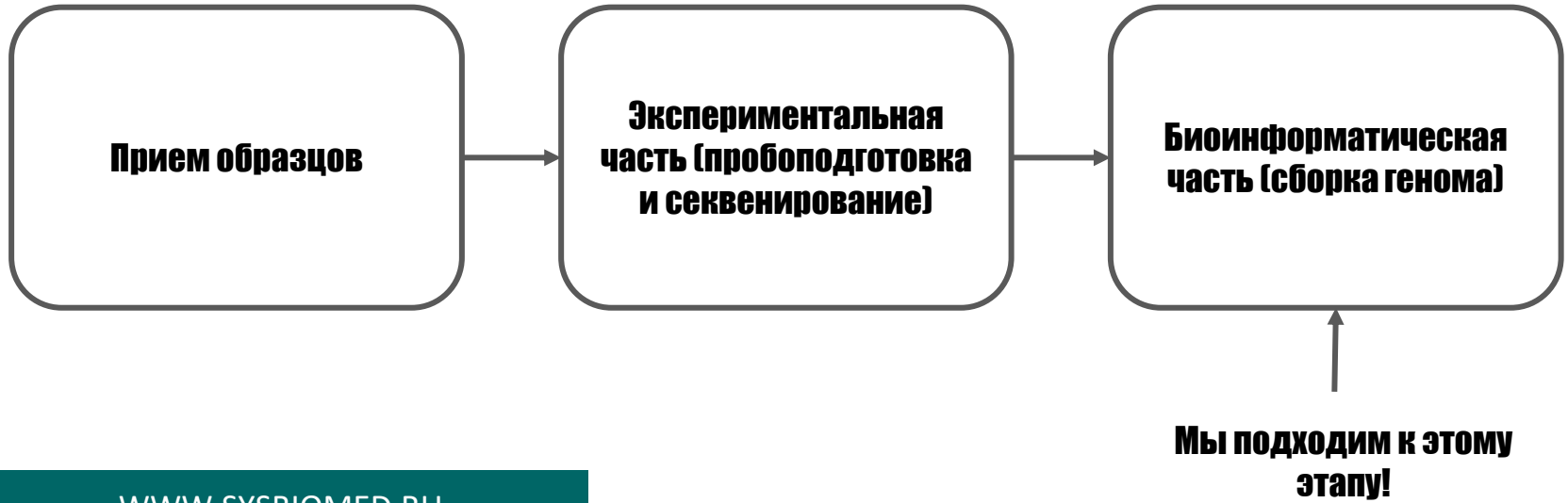
Д.В. Кривонос

О чем мы сегодня поговорим и что будем делать?

1. Повторим теорию.
2. Ура, собираем геномы 😊
3. Посмотрим, что у нас получилось.

Что? Зачем?

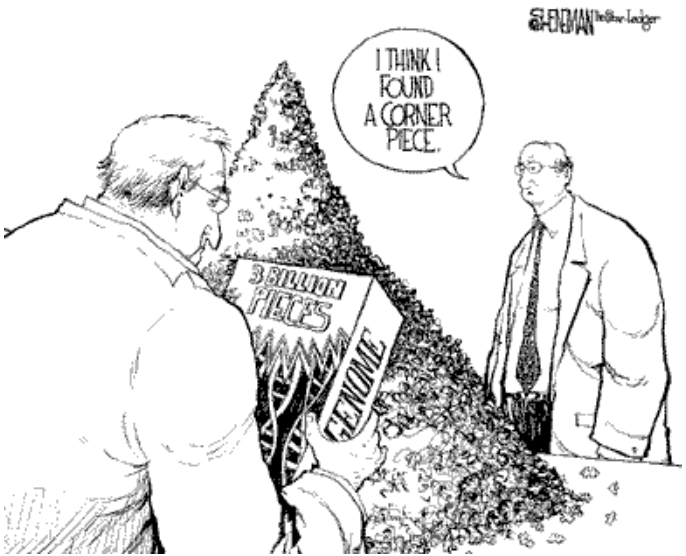
Биоинформатический анализ является неотъемлемой частью секвенирования нового поколения, по сути своей сводя все приложенные ранее усилия к конечному результату.




Биоинформатика

Что такое сборка генома?

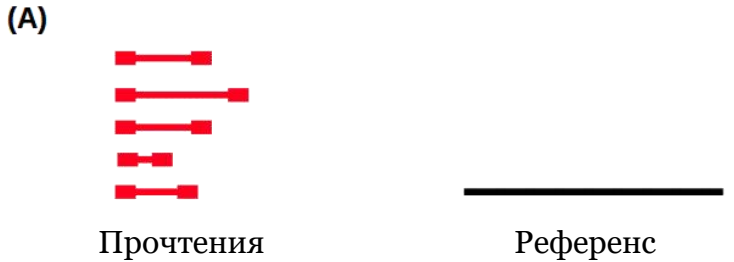
Сборка генома - объединения коротких фрагментов ДНК (прочтений, ридов) в одну или несколько длинных последовательностей в целях восстановления последовательностей ДНК. **Сборка осуществляется комплексными математическими алгоритмами, реализованные в специализированный программе-сборщике.**



Что такое сборка генома?

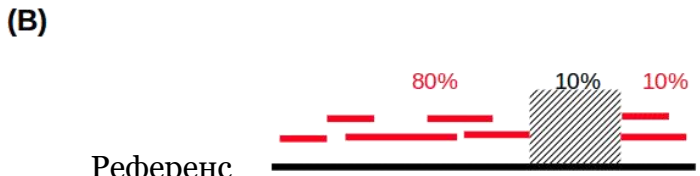
Критерии сборки

Глубина зависит от того, как много пришлось прочтений пришлось на участок



$$C = \frac{\text{Основания, полученные в результате секвенирования}}{\text{Основания референса}}$$

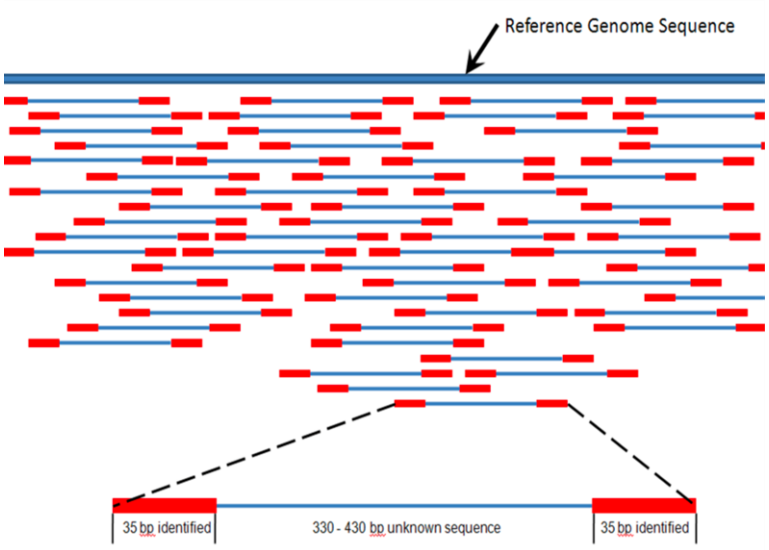
Ширина зависит от того, как равномерно прочтения легли на геном



$$C = \frac{\text{Область покрытая прочтениями}}{\text{Длина генома}}$$

Особенности сборки с референсом

Сборка с референсом – сборка с использованием генома генетически близкого организма, который берется за **эталон** (референс). В ходе такой сборки прочтения накладываются на референс, после чего формируется консенсусная последовательность.



В случае **SARS-CoV-2** в качестве референса берется последовательность уханьского штамма Wuhan-Hu-1 (MN908947.3).

Какие файлы нам интересны?

fastq(.fastq) - файлы, содержащие “сырые” данные (прочтения).

```
@88876672-b9bc-49a6-8873-7220690e61f4 runid=05b2438e98ded676842dc340ab2aa51c60e3bae5
read=26 ch=1064 start_time=2022-09-02T11:15:03.580637+00:00 flow_cell_id=PAI77680
protocol_group_id=SARS-CoV2 sample_id=Plate_2022_02_22-03_02_8-6 barcode=barcode01
barcode_alias=barcode01 parent_read_id=88876672-b9bc-49a6-8873-7220690e61f4
basecall_model_version_id=2021-05-05_dna_r9.4.1_promethion_384_dd219f32
GCCGTTCGATTTTCAGATGGTGTTTAACGAAAGTTGTCAGTATTTTGTGGTTTTCATTTGTTATCGTGAAGCATTCGCCGTTTTTT
CACGCCGCTTCTGGTTCAGAGCGAGGCATGAGGTGGGTGCAATGAGACAGTGAACACAGGGCCGGGAGTAGAGAAGGATCCTCTGT
GACCGCTTCCTTCCAGACTTAGCTTTGAAAAC
+
$$%&(01+'&'''$##$##$&.00+)(*'''*+./AC?)(('''(+;9521001./('''(''&+-
)(''(%%%%(&%%%$$),110/)&&(0011-
,((%%%$%$%()*4899:;>1/**++35<><:<;;<B>>==9::::?633554333442/,*.5/)766;::::;777442556>1
:<.)'))('(+''+,-'&&'&$
```

Индекс прочтения

Нуклеотидная последовательность

Строка качества

Какие файлы нам интересны?

fasta (.fasta) - файлы, содержащие собранные нуклеотидные последовательности.

```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete
genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTA
GGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTG
TTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTG
TCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAATTGGCAAAGAA
```

Название последовательности

Нуклеотидная последовательность

Что мы будем использовать?

ARTIC SARS-CoV-2 Workflow – специализированная программа-сборщик, которая сильно упрощает жизнь при сборке SARS-CoV-2. Программа сама фильтрует прочтения по качесве, срезает праймеры, картирует прочтения и формирует консенсус, а также определяет линию генноварианта .

```
nextflow run epi2me-labs/wf-artic --fastq <path_to_fastq> --scheme_version <primer_scheme> --out_dir <path_to_out>
```

Для того, чтобы запустить программу нужно будет указать **путь к папке с баркодами** <path_to_fastq>, **схему праймеров** <primer_scheme> и **путь к выходным файлам** <path_to_out>.


Практика …

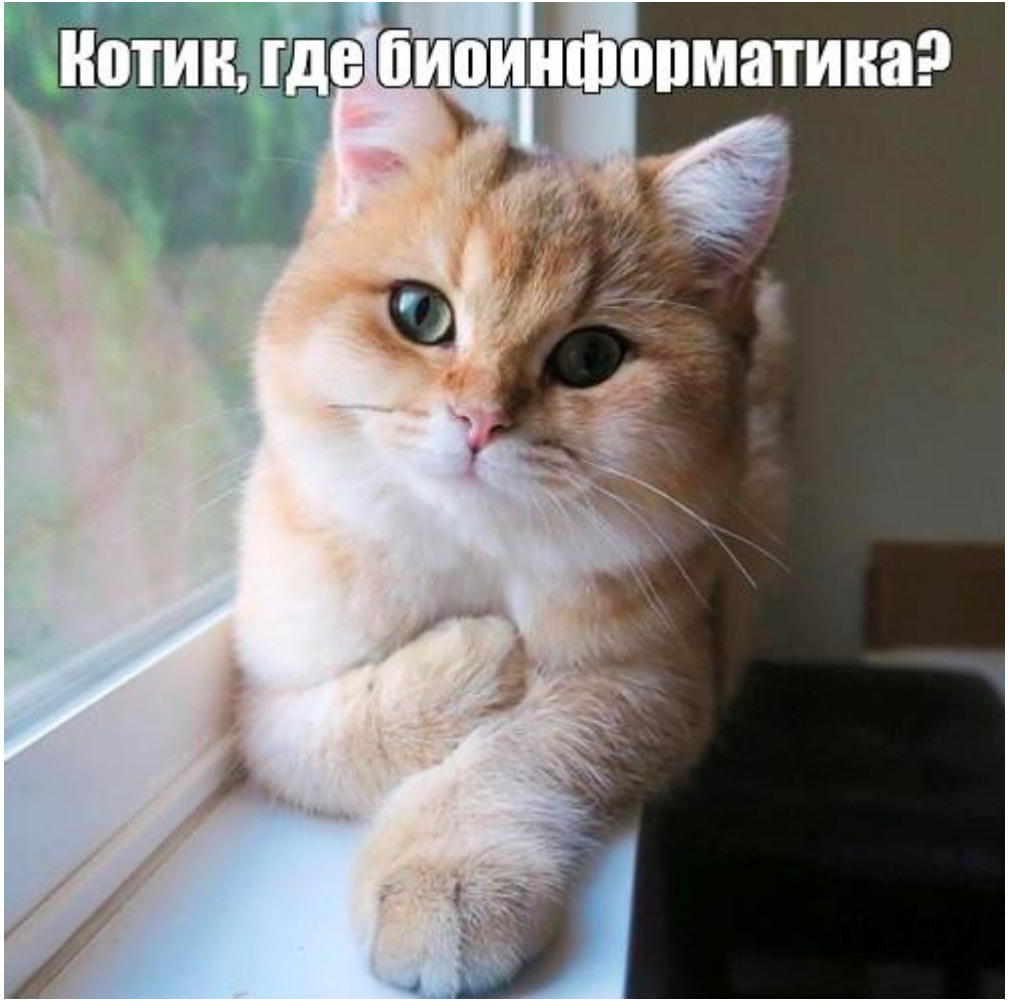
Эх, просто руками делать не хочется … 😩

```
awk '/^>/ {out = substr($1, 2) ".fasta"; print > out} !/^>/ {print >> out}' <YOUR FASTA>
```

<YOUR FASTA> - название файла сборки

Пример:

```
awk '/^>/ {out = substr($1, 2) ".fasta"; print > out} !/^>/ {print >> out}' all_consensus.fasta
```


Котик, где биоинформатика?