## Работа на сервере

Подключение из АУ: *ssh username@192.168.222.223*
Из дома (через порт 23): *ssh username@194.85.238.21 -p 23*
Для работы под Windows понадобится putty.

В /labnas/students создайте себе папку и перенесите туда свои файлы. Чтобы мне и вам было удобнее создавайте папки с номером занятия. Пожалуйста, работайте только в этих папках, а не в /home/username -- там мало места. Избегайте копирования больших файлов, лучше создавать на них ссылки командой
*ln -s <path>*

Для удобства можно пользоваться программой tmux (но совсем не обязательно). Это программа автоматически сохраняет вашу текущую сессию и позволяет работать на сервере сразу в окнах. Для запуска просто наберите
*tmux*
При каждом следующем заходе на сервер набирайте
*tmux attach*
Для выхода набирайте
Ctrl+B, затем D.
Все команды в tmux начинаются с Ctrl+B. Чтобы увидеть полный список можно нажать Ctrl+B, затем ? или почитать мануал http://www.openbsd.org/cgi-bin/man.cgi?query=tmux&sektion=1 (KEY BINDINGS).

## Общие моменты

Язык программирования можете выбрать любой. Главное, чтобы программу на нем можно было бы запустить на сервере и я бы смог её понять. Python предпочтителен, так как он простой, и в курсе скорее всего будет рассказ про Biopython.

Программы на питоне обычно имеют разрешение .py и запускаются командой
*python <foo.py> <arguments>*
Хотелось бы видеть понятный код с комментариями там, где они нужны. Хорошо, если эти комментарии будут еще и нести какой-нибудь смысл.

Официальный учебник по питону: http://docs.python.org/tutorial/
Перевод: http://ru.wikibooks.org/wiki/%D0%A3%D1%87%D0%B5%D0%B1%D0%BD%D0%B8%D0%BA_Python_2.6
Можно пользоваться и документацией: http://docs.python.org/reference/

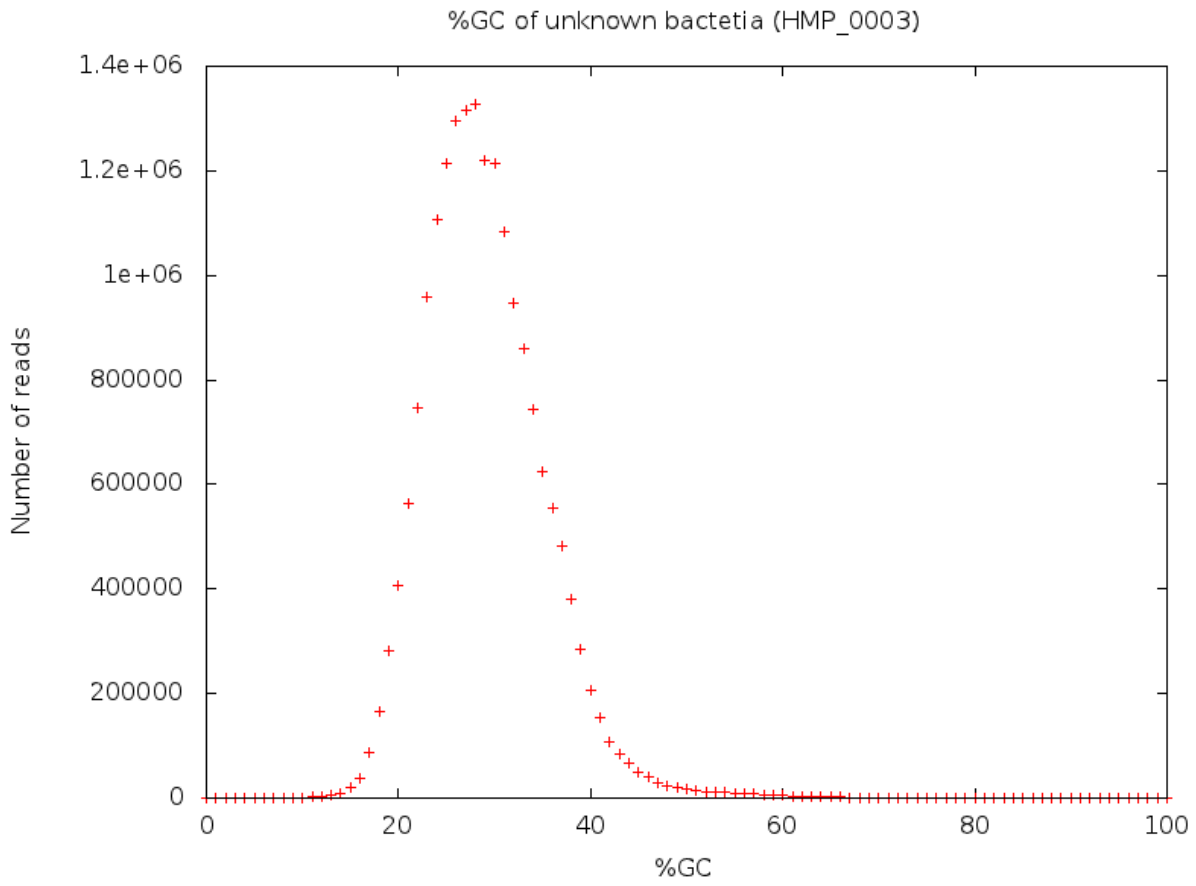
Домашние задания (в том числе графики) оставляйте в своих папках с понятными

названиями. Например, 1_gc.py. Дедлайн -- 23:59 в день перед занятием. В первый раз дедлайна не будет.

## Первое домашнее задание

1. По FASTQ (http://en.wikipedia.org/wiki/FASTQ_format) файлу определить GC-состав ридов и его распределение. GC состав определяется как доля нуклеотидов G и C по отношению к общему числу нуклеотидов, обычно измеряется в процентах. Распределение GC состава есть график, на котором по оси x отложен GC состав, а по оси y -- его частота, то есть количество ридов с соответствующим составом.

Пример графика GC распределения неизвестной бактерии с GC составом 30%:



Как и во многих заданиях на анализ данных, меня интересует больше не код, а скорее метод в общем и результат. Для улучшения результата добавьте в программу следующее:

- Не учитывать нуклеотиды с плохим качеством (качество определять автоматически)
- Не учитывать риды, в которых мало хороших нуклеотидов
- Возможно, что-нибудь еще на ваш выбор

График можно строить в любой удобной программе. В питоне есть библиотека для построения графиков: http://matplotlib.sourceforge.net/

2.1. Разбить один FASTQ файл на 4: левые (/1) риды с парой, правые (/2) с парой, левые без пары, правые без пары. Если у рида есть пара, то они идут подряд, причем всегда левый идет первым.

2.2. Разбить несколько (можно 2) FASTQ файлов на такие же 4 файла. Расположение ридов в исходных файлах произвольное.