Работа на сервере ace

Для работе на сервере придется использовать командную строчку. (для работы под Windows понадобится, например, putty.)

Подключение из АУ: *ssh username@192.168.222.223*
Из дома (через порт 23): *ssh username@194.85.238.21 -p 23*
Уже с сервера, используя те же логин и пароль, подключаемся к одному из узлов кластера:
ssh username@ant<номер кластера>

Чтобы работать на кластере, создайте свою папку в /tmp. Перед работой данные **НУЖНО СКОПИРОВАТЬ** (да, это действительно нужно):
cp *<путь к файлу на ace>* /tmp/*<куда копировать>*

Копирование данных с ace на локальный компьютер:
scp -P 23 prjbel@194.85.238.21:<путь к файлу на ace> <куда копировать>

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, но мне кажется проще использовать python. Главное, чтобы программу можно было бы запустить на сервере и я бы смог её понять. Хотелось бы видеть 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/

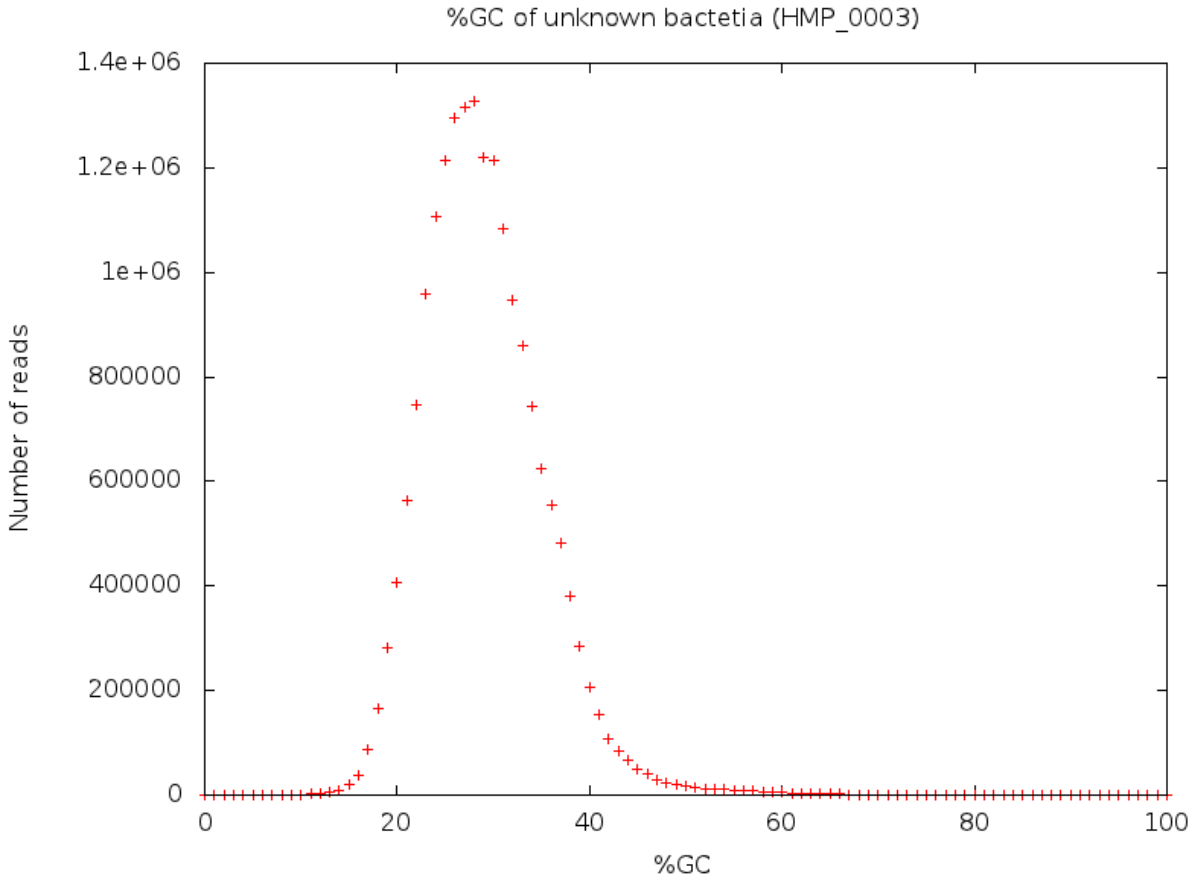
Перевод: http://ru.wikibooks.org/wiki/%D0%A3%D1%87%D0%B5%D0%B1%D0%BD%D0%B8%D0%BA_Python_2.6
Можно пользоваться и документацией: http://docs.python.org/reference/

Домашние задания, а именно код и графики оставляйте в папке /storage/labnas/NGS/students/<ваша фамилия>/<номер задания>/<номер задачки внутри задания>/
Дедлайн к выполнению -- 23:59 в субботу через одно занятие. То есть в обычном режиме -- ровно 4 недели.

Первое домашнее задание

1. GC состав.
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%:


%GC of unknown bactetia (HMP_0003)
Number of reads
1.4e+06
1.2e+06
1e+06
800000
600000
400000
200000
0
0
20
40
60
80
100
%GC

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
- 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/

Тестовые файлы: /storage/labnas/NGS/1/test.fastq /storage/labnas/NGS/1/test3.fastq
Входной файл: /storage/acestorage/data/input/E.coli/sc_lane_1/ ecoli_mda_lane1.fastq

2. Распределение качества.
По FASTQ файлу построить распределение вероятности ошибки в зависимости от позиции нуклеотида.

Тестовые файлы: /storage/labnas/NGS/1/test.fastq /storage/labnas/NGS/1/test3.fastq
Входной файл: /storage/acestorage/data/input/E.coli/sc_lane_1/ ecoli_mda_lane1.fastq