## Обмен файлами с сервером

Под линуксом удобно копировать файлы, например, используя scp:
*scp [-r] [-P port] <source> <destination>*
-r -- копировать директории рекурсивно;
-P -- номер порта;
<sorce>, <destination> -- что и куда буем копировать файлы. Можно копировать как на сервер, так и с него. Путь на сервере указывается как
*user@194.85.238.21:/acestorage2/students/.../*
Например, залить локальную папку на сервер можно командой
scp -r -P 23 ~/mylocaldir/ *user@194.85.238.21:/acestorage2/students/user/*

Можно использовать, например, программу FileZilla (http://filezilla-project.org/). Она есть и для линукса, и для Windows. Пользоваться ей достаточно просто. В качестве хоста вводим
*sftp://194.85.238.21*
Далее вводим логин, пароль, порт (23) и подключаемся. Копировать файлы можно просто перетаскивая их, или правой кнопкой по нужным файлам и Upload/ Download.

FileZilla далеко не единственный клиент. Например, в Windows можно пользоваться привычным Total Commander.

Напоминаю, домашние задания должны быть в директории /labnas/students/. Все данные для заданий будут в /labnas/NGS/.

## Bowtie

Сайт: http://bowtie-bio.sourceforge.net/index.shtml
Статья: http://genomebiology.com/content/pdf/gb-2009-10-3-r25.pdf
Программа находится в /labnas/NGS/2/bowtie-0.12.7/.

Для того, чтобы приложить риды необходимо сперва построить индекс генома (или контигов), к которому будем прикладывать. Это делается командой:
*bowtie-build <genome file> <index name>*
Индекс строиться один раз для конкретного генома. Для больших геномов (например человека) индекс можно скачать с сайта bowtie.

Риды можно прикладывать в непарном и парном режимах.

Общие опции:
-q/-f -- формат ридов (fastq/fasta)
-p <thread number> -- количество потоков (много не используйте)
--al <filename> -- вывести все приложившиеся риды в указанный файл

--un <filename> -- вывести все не приложившиеся риды в указанный файл
-m <int> -- игнорировать риды, приложившиеся больше, чем заданное число раз
-a -- выводить все приложения рида (по умолчанию выводится первое найденное)
--best -- выводить лучшее возможное приложение для рида
Последние три опции могут замедлить работу bowtie.

Запуск в одиночном режиме:
*bowtie -q -p 4 --best  <index name> <fastq file with reads>  > 1.log  2> 1.err*
В файле 1.log будет вся информация о приложившихся ридах -- стренд, позиция и т.д. В файле 1.err будет краткая статистика.

Опции парного режима:
--minins <int> -- минимальное расстояние вставки (insert size)
--maxins <int> -- максимальное расстояние вставки
--fr/--rf/--ff -- ориентация ридов

Запуск в парном режиме:
*bowtie -q -p 4 --fr --minins 0 --maxins 500  <index name>  -1 <left fastq file>  -2 <right fastq file>  > 1.log  2> 1.err*

Полный список опций:
bowtie -h

Различные ориентации пар ридов:
FR (forward - reverse)  —-->  <-----
RF (reverse - forward)  <-----  —-->
FF (forward - forward)  —-->  —-->
Расстояние вставки во всех случаях определяется как расстояние между дальними концами ридов.

## Второе домашнее задание (до 23:59 18.03.12)

0. Посмотреть на статистику ридов, разобраться в выводе Bowtie.

1. По выводу Bowtie построить график покрытия генома, определить среднее покрытие и долю покрытой области генома. Покрытие одного нуклеотида есть количество ридов, приложившихся так, что их концы находятся по разные стороны от нуклеотида. График можно строить усредняя, например, по 1000 нуклеотидов. Доля покрытой области генома определяется как процент нуклеотидов с ненулевым покрытием по отношению ко всей длине генома.

2. По выводу Bowtie построить график распределения расстояния вставки, определить среднее расстояние вставки и его среднеквадратичное отклонение.

По оси X -- расстояние вставки, по оси Y -- количество ридов в заданным расстоянием вставки.

3*. Задание, которое можно сделать вместо 1 и 2.
Получить парную статистику по запускам Bowtie в одиночном режиме. У нас есть данные о том как приложились левые и правые риды по одиночке. Хочется узнать процент пар с каждой из ориентаций, процент пар с одним приложившимся ридом и процент пар без приложений. Для приложившихся пар хочется узнать среднее расстояние вставки и его СКО.

Данные для первых двух заданий.
Если ваша фамилия начинается на А-И:
Папка: /2/E.coli/
Датасет №1: 05.reads.left.corrected.fastq, 05.reads.right.corrected.fastq
Датасет №2: s_6_1.fastq, s_6_2.fastq
Геном: MG1655-K12.fasta

Если ваша фамилия начинается на К-Я:
Папка: /2/P.stipitis/
Датасет №1: HTC10499_s_8_1.fastq, HTC10499_s_8_2.fastq
Датасет №2: HTC10508_s_8_1.fastq, HTC10508_s_8_2.fastq
Геном: P.stipitis.fasta

Третье задание можно протестировать на любых датасетах.