Studying gene expression and sequence variability for the same individuals

\*Please provide as automated scripts as possible\*

\*You are free to split in teams of 2 people, or work alone\*

In this work, we will study, in conjunction, the expression levels of 10 individuals from the 1000Genomes Project, as well as their nucleotide diversity.

In fact, we will just produce the polymorphism data and the RNA (count) data. We will not answer more biological questions using these results. However, we can think about possible such questions

Data download

Go to the: <https://www.internationalgenome.org/category/rnaseq/>

Follow the link for the “RNAseq on 60 CEU individuals”. This will bring you to a web page for the expression data of 60 Individuals that exist in the 1000Genomes project, of European origin, for which we also have expression data.

Follow the link that refers to the “Detailed sample information and links to data “ and then Download the fastq files (paired-end) for **5 individuals (the first five in the list)** (please pay attention that each individual refers to two files \_1.fastq.gz and \_2.fastq.gz since it’s paired-end).

Perform all the RNA-seq analysis up to the counts data.

For **the same individuals** download the DNA-sequencing data as follows:

* go to the ftp site (and use the phase3 data)  
  <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/>
* The problem is that each sample has been sequenced \*multiple\* times. We will NOT download all these multiple fastq files for each sample. Instead, consult the   
  <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/20130502.phase3.analysis.sequence.index> file. It contains information for each file located in the ftp server.   
    
  For example,the first row:  
  data/NA20505/sequence\_read/ERR005686\_2.filt.fastq.gz 958161180bc95293a1cc29b1082dd685 ERR005686 SRP000540 1000 Genomes Toscan population sequencing SC ERA000103 2009-08-28 00:00:00 SRS001673 NA20505 TSI ERX000613 ILLUMINA Illumina Genome Analyzer II g1k-sc-NA20505-TOS-1 g1k-sc-NA20505-TOS-1-NA20505-76-2\_1 175 PAIRED data/NA20505/sequence\_read/ERR005686\_1.filt.fastq.gz 0 14586711 1108590036 low coverage  
    
  i.e. the sample located at the (relative path) data/NA20505/sequence\_read/ERR005686\_2.filt.fastq.gz (this is the second mate of the paired-end file, the first ends in \_1.filt.gastq.gz) contains 14586711 reads and 1108590036 bp (i.e. the one-before-last field provides the number of bp and the two-before-last field the number of reads).
* Get the \*paired-end\* files that have the greatest number of bp
* Download \*only\* these fastq files (for the 5 samples that you have RNA-seq data).
* Perform all the analysis up to the point that you have a VCF file

Warning: Mate records missing for 13169 records; first such record: <SAM\_Alignment object: Paired-end read 'ERR009126.553853' aligned to NC\_000001.11:[14489,14525)/+>.

Warning: Mate records missing for 14362 records; first such record: <SAM\_Alignment object: Paired-end read 'ERR009160.3421547' aligned to NC\_000001.11:[258944,258980)/->.

6221775 SAM alignment pairs processed.

./rna-seq.sh: line 161: 1113 Segmentation fault STAR --genomeDir ./ --readFilesIn ${fastqdir}trimmed\_${map\_2}\_1.fastq ${fastqdir}trimmed\_${map\_2}\_2.fastq --runThreadN 8 --outSAMtyp

e BAM SortedByCoordinate --quantMode GeneCounts --outFileNamePrefix star\_count\_${map\_2}

Grep -ie “mapped” \* -- look for mapped %

cat ENS

grep -iE "gene=" Homo\_sapiens.gff > fixed\_Homo\_sapiens.gff