SNPS and more SNPs

On this assignment remember that you are encouraged to work with other students if you have problems finding the data you need.

A) You will need to get a vm at https://vm-manage.oit.duke.edu

B) I would recommend using the LAMP image as a starting point.
You will need to run
sudo apt-get update
sudo apt-get install bwa
sudo apt-get install samtools

to install the basic tools needed for this assignment.
You will also need to get the sratools from the NCBI web site, choosing the precompiled ubuntu linux version.

Assignment
Using the S288c S. cerevisiae reference sequence, available from NCBI
1) Write a **script** that will download the following data, run bwa and samtools, and pull out a list of SNPs relative to the above reference sequence, and identify among the snps which are strain specific, and which are shared by two or more strains.

Note that the assignment is writing a script that does the above, not doing it by hand typing in each command to get the result.

And the following raw genomic fastq files from the SRA (short read archive), which you will need to get using sratoolkit (hint prefetch then fastq-dump also check that you should get pairs of 101 base pari long reads, not single 202 base pair long reads)
Another hint – The oit vm you will be using has limited space, so as your script is running, you probably should delete stuff as it is no longer needed. i.e. it is probably better to have it download one sequence at a time, extract the SNPs, delete the now unneeded files, and then move on to the next one.

2) The script should then display the results on the web page of your vm.
You can either do this through PHP or perl-cgi or python-cgi as a script launched from the web site, or you can write a stand-alone script that generates a results file in the appropriate html format that you can then link to your web page.

3) provide a link on your web site to your scripts source code .

4) The web site should include some sort of summary of the SNPs, as well as a link to the long list of SNPs. The summary could be something like:

Chromosome	yjm984				yjm969 …	
SNPS:	private number	private density	shared number	shared density	private	shared …
1	682	1/1293bp	484	1/2321 bp		
2						
…						
16						
mitochondria						

The list of SNPs should be in some reasonable format where each SNP is labeled by the reference (S288c) the strain, the chromosome, the position, the base change.

5) A brief summary page on the web site, including the density of SNPs observed, how this compares to the density of SNPs in humans, and whether the distribution of common SNPs suggests a history of genetic exchange among the ancestors of these strains.

5) Email a link to your web page – I will grade it from there.

Another hint – these are large files. While developing and debugging your script it might be good to use a limited amount of data so that it runs faster. For instance try only 100kb from the reference set, and perhaps only the first 500,000 lines from the fastq files.

Strains to look at:
yjm984
yjm969
yjm1190
yjm975
yjm1439
yjm689
yjm1199
yjm193