Data and scripts for this Tutorial and Lectures are available at https://github.com/hyphaltip/CSHL_2012_NGS.

Preparation Steps

1. Download Broad IGV viewer at http://www.broadinstitute.org/igv/

2. Download the Saccharomyces genome from SGD site. Uncompress this and get the .fsa file which is the genome. Copy it and run rename_seq.pl on it to fix the chromosome names so they match the GFF file: https://github.com/hyphaltip/CSHL_2012_NGS/blob/master/data/rename_seq.pl

- You may need to fix this GFF file so it doesn't have any sequence
- Do a grep to find where the '>' lines are where the sequence as fasta is in there and find the first one
- grep -n ">" saccharomyces_cerevisiae_R64-1-1_20110208.gff
- read in the first N lines using head
- head -n 16425 saccharomyces_cerevisiae_R64-1-1_20110208.gff > saccharomyces_cerevisiae_R64-1-1_20110208.noseq.gff
- Use this saccharomyces_cerevisiae_R64-1-1_20110208.noseq.gff for GFF file later needs.

3. Downlod the read data from SRA and convert it -- This step already done for you, folder is on server or you can run the download script when you get home (rerequires curl, sra toolkit, and for download speedup Aspera client

    - For Asperga, get the web client and find the ascp binary and install in your path. For example on OSX it installs in "/Applications/Aspera Connect.app/Contents/Resources/ascp".

    - Download script to obtain all the data is here https://github.com/hyphaltip/CSHL_2012_NGS/blob/master/data/download.sh

Tutorial

1. Trim FASTQ data for quality using sickle - run 'sickle pe' to see how to run PE options

2. Compare the FASTQC quality report for one of the files (_1 or _2) files both before and after trimming. Set this up in the background so you can run it and do other things in the meantime.
    - Run fastqc -h to get help

3. Align reads to the genome using BWA. This requires you to also build and index for the genome

4. Call SNPs with SAMTools - refer to the SAMtools manpage on mpileup for more details. http://samtools.sourceforge.net/

5. Call SNPs with GATK

6. Run Filtering steps on GATK output SNPs to remove potential biased or low-quality ones

7. Calculate the total number of remaining SNPs.

8. For advanced users, intersect this list of SNPs (in the VCF file) with the GFF for the genome to determine which SNPs are in coding regions. Read up on BEDTools. The genome annotatio in GFF is available in the folder where the genome was downloaded from SGD.

9. Open the genome file for Saccharomces in IGV. Then add the GFF file as annotation track. Then BAM file, and VCF file in IGV to vi

Feel free to try this also with your own favorite organism. Many datasets exist in the SRA from genome resequencing. To extend the problem, ddownload more than 4 strains so you can apply comparisons between individuals instead of just between one individual and the reference.