Preparation Steps

1. Checkout my github repo for this course.

Using Git on the command line (or install Git for Mac)

```
$ git clone git://github.com/hyphaltip/CSHL_NGS.git
```

1. The data you need are available from here. There are directions at the bottom if you want to see how the data are obtained but not necessary for this tutorial.

2. We will use some data from a strain of yeast, W303 which has some drug resistance properties and mixed ancestry from the reference strain S288C. A project sequenced the genome and analyzed it to show amino acid alterations in ~800 genes. Ralser et al 2012. We will attempt to replicate some aspects of this analysis (though note they did an assembly of the read not simply mapping them to the refeence).

3. These data are available in the data/example/W303_chrII_1.fastq.bz2 and W303_chrII_2.fastq.bz2 – you will need to uncompress with bunzip2

4. The genome is available in data/genome/Saccharomyces.fa.gz – you need to uncompress this with gunzip data/genome/Saccharomyces.fa.gz

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.

`fastqc -h` to get help

1. Align reads to the genome using BWA. This requires you to also build and index for the genome. See the lecture notes. (bwa index)

2. Realign reads with Picard and GATK based on lecture. (picard

3. Fix the Read groups see this slide (picard)

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

5. Call SNPs with GATK, using example from the lecture

6. Run Filtering steps on GATK output SNPs to remove potential biased or low-quality ones using options provided in lecture.

7. Calculate the total number of remaining SNPs. Count the lines or use vcftools.

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 annotation in GFF is available in the folder where the genome was downloaded from SGD.

9. Open the genome file for Saccharomces in IGV. http://www.broadinstitute.org/igv/ is here.

Then add the GFF file as annotation track. Then BAM file, and VCF file in IGV to view the SNPs in context of the gene annotation and the read-depth

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

For example, here is the Drosophila reference panel which included sequencing 192 individuals. Or find something smaller (10 C.elegans for example).

Obtaining the datasets from the public archives for this tutorial

1. Download the Saccharomyces genome from SGD genome release (2011). Uncompress this and get the .fsa file which is the genome.

You could do this like

```
 $ curl -O
http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_Curre

$ tar zxf S288C_reference_genome_Current_Release.tgz
```

Run this script to fix the chromosome names in the download file so the will match the GFF file.

```
perl CSHL_NGS/data/rename_seq.pl S288C_reference_genome_R64-1-1_20110203/S288C_reference_sequence_R64-
```

- You need to fix this GFF file so it doesn't have any sequence, to do this a grep to find where the '>' lines are where the sequence as fasta is in there and find the first one.

Commands to run

```
grep -n ">" S288C_reference_genome_R64-1-1_20110203/saccharomyces_cerevisiae_R64-1-1_20110208.gff
# note the number for the first '>' - there are a few lines before that we want to drop so
# and the last position in the file we want is 16425 so we can use head
$ head -n 16425 S288C_reference_genome_R64-1-1_20110203/saccharomyces_cerevisiae_R64-1-1_20110208.gff
```

- Use this saccharomyces_cerevisiae_R64-1-1_20110208.noseq.gff for GFF file later needs.
- SRA data are available to download here. There is a script and table listing the sources for some strain data. You can use the Aspera pluging for fast downloading or wget/curl will work. The fastq-dump script as part of the sratoolkit is needed.
    - The download script to obtain all the data is here https://github.com/hyphaltip/CSHL_NGS/blob/master/data/download.sh or in the github repo you checked out – `CSHL_NGS/data/download.sh`