**Convert SRA files to FASTQ files:**

1. Download the SRA toolkit from <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>
2. Decompress:

Tar –zxvf sratoolkit.2.3.4-2-centos\_linux64.tar.gz

1. Convert SRA to fasta by fastq-dump:

./sratoolkit.2.3.4-2-centos\_linux64/bin/fastq-dump /home/pjflaherty/flahertylab/freeze/baker\_yeast/\*.sra

**Fastq to fasta: using fastx toolkit**

<http://hannonlab.cshl.edu/fastx_toolkit/commandline.html#fastq_to_fasta_usage>

**Map reads to reference:**

<http://icb.med.cornell.edu/wiki/index.php/Elementolab/BWA_tutorial>

**How to install the BWA**

<http://icb.med.cornell.edu/wiki/index.php/Elementolab/BWA_tutorial>

export PAHT=$PATH://)))

How to map

<http://ged.msu.edu/angus/tutorials-2011/bwa_tutorial.html>

Now there are several steps involved in mapping our sequence reads and getting the output into a usable form. First we need to tell bwa to make an index of the reference genome; this will take a few minutes, so we’ve already got the index already generated in the data directory, but if you want to try it yourself, you can run (but see the note below first!!):

%% bwa index dmel-all-chromosome-r5.37.fasta

Note: This step takes several minutes. If you run it, it will overwrite the index files we have already made, so don’t run the above line exactly; instead, create a copy of the reference genome and then index the copy instead, so that we can preserve our pre-computed reference index:

%% cp dmel-all-chromosome-r5.37.fasta dmel-all-chromosome-r5.37.copy.fasta

%% bwa index dmel-all-chromosome-r5.37.copy.fasta

Next, we do the actual mapping. These were paired-end reads, which means that for each DNA fragment, we have sequence data from both ends. The sequences are therefore stored in two separate files (one for the data from each end), so we have two mapping steps to perform. For now, we’ll use bwa’s default settings. The files you’ll be running this on are datasets that have been trimmed down to just the first 1 million sequence reads to speed things up, but at the end you’ll be able to work with the final product from an analysis of the full dataset that we ran earlier (some of these steps take upwards of an hour on the full dataset, but just a couple minutes on the trimmed dataset). Run:

%% bwa aln dmel-all-chromosome-r5.37.fasta RAL357\_1.fastq > RAL357\_1.sai

%% bwa aln dmel-all-chromosome-r5.37.fasta RAL357\_2.fastq > RAL357\_2.sai

These .sai files aren’t very useful to us, so we need to convert them into SAM files. In this step, bwa takes the information from the two separate ends of each sequence and combines everything together. Here’s how you do it (this may take around 10 minutes):

%% bwa sampe dmel-all-chromosome-r5.37.fasta RAL357\_1.sai RAL357\_2.sai RAL357\_1.fastq RAL357\_2.fastq > RAL357\_bwa.sam

The SAM file is technically human-readable; take a look at it with:

%% more RAL357\_bwa.sam