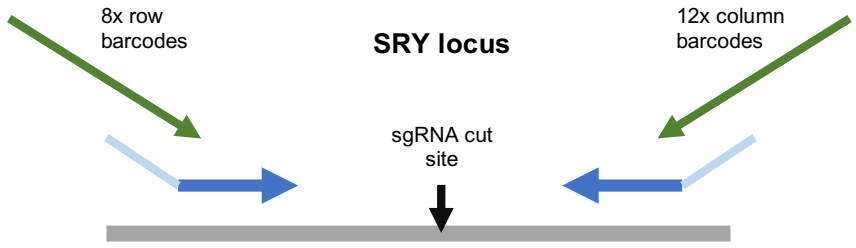
CRISPR Genotyping with Illumina Sequencing

STEMREM 201B

Kevin Parker

Objectives: The goal of this project is to get you a bit of practice working on the server, and to have you interact with .fastq files.

Project: Imagine that you are trying to create mutations in a gene of interest (in this case, SRY). You've picked 96 colonies, and now you need to screen them to see which ones have the wild-type sequence, and which ones have indels. There are several ways to genotype colonies (e.g., see http://blog.addgene.org/crispr-101-validating-your-genome-edit), but we've elected to use the MiSeq to screen all 96 colonies at once. To do this, we designed primers that amplify a 150bp region surrounding our targeted cut site on SRY, and amplified genomic DNA from all 96 samples using PCR. We then prepared our plate for Illumina sequencing – you can look on Illumina's website for more information on libraries are prepared for sequencing, and we'll discuss it a bit in class as well – but essentially, you can do this by performing additional PCR using primers containing the required Illumina adapter sequences and unique barcodes. This means that you can pool your entire plate of samples, run them together, and then the MiSeq will output an individual .fastq file containing the reads originating from each well.



(look at https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-03.pdf for more information - this is known as "dual indexing")

Workflow: I have simulated .fastq read files for 6 wells. Each well (theoretically) has three possibilities: two wild-type alleles, one wild-type and one mutant allele, or two mutant alleles. Your goal is to report back, for each colony, what allele(s) are present – was there a cut, and if so, what is the cut?

Useful information:

- Our forward primer is 5'-GTCCAGCTGTGCAAGAGAAT-3'
- Our reverse primer is 5'-TGAATGCGTTCATGGGTCGC-3'
- Files are located at /home/krparker/stemrem/crispr/fastq on rice.stanford.edu
- Feel free to work either on your own computer (if you have a Mac) or on the server – use "scp" if you are copying from server to own computer, or "cp" (or "scp") if you are copying within the server.
- When you first log in to the server, you may need to type "bash" (followed by enter), otherwise certain commands may not work – there's no harm in running it even if you don't need to.
- I would recommend downloading a good text editor (e.g., Sublime Text), which will automatically highlight bash commands/comments, making it easier to read.

Tutorial (broad overview, with useful commands):

1. Make a new folder called "crispr" in your home directory
    a. mkdir; cd
2. Copy the files to your directory (or local computer)
    a. scp
3. Unzip the files
    a. gunzip
4. Look at the files
    a. head; cat
5. How many reads are in each file?
    a. wc
6. How many times do we see the WT allele in each file?
    a. grep
7. What are the most commonly seen reads in each file?
    a. sort, uniq

See the file 'CRISPR_commands.txt' for more information:

- Lines beginning with a "#" are commented out – i.e., they won't run
- You should be able to copy/type other lines into the terminal (of course, sometimes you will need to change the name of the file)

Challenge: Using a for loop, examine the .fastq files in /home/krparker/stemrem/crispr/96_wells. Which ones are WT? Which ones are heterozygous mutant, and which are homozygous mutant? Output the results in a single file, which has the well number, number of total reads, number of wild-type reads, and the number of reads corresponding to the two most common mutant reads as well as their sequences. As an example, your output might look something like the file sample_output.txt. See challenge_notes.txt for more information.