Project #2. “Why did I get the flu?”. Deep sequencing, error control, p-value, viral evolution..

New bioinformatics skills covered: deep sequence data, awk one-liners

This year, you prudently got the flu vaccine. So when your roommate, who forgot to get vaccinated, came down with the flu, you weren’t worried. But somehow, several days later, you started feeling feverish, weak, and all-around-awful. You knew it was the flu, but how could this be?!

Suspecting that this season’s vaccine wasn’t a match for the flu virus that infected your roommate, you had some friends in the med school run a hemagglutination inhibition (HI) assay on virus samples from your roommate.

The results showed that your roommate’s virus closely matched the HI profile for an H3N2 strain called A/Hong Kong/4801/2014 (H3N2). *(Find out what strains were in this season’s vaccine. Was that one of the flu strains covered by this vaccine?)*

You’ve heard of viral quasispecies, and you suspect that maybe a small portion of the virus population mutated and evolved while replicating inside your roommate’s cells, which could explain how it was able to infect you. To find out, you have your friends set up a targeted deep sequencing experiment to analyze the HA genes in your roommate’s viral sample. They set up an Illumina single-end sequencing run. When they send you the results, you start analyzing your roommate’s sequence data right away.

1. Inspect the data from your roommate

Sequencing results are usually stored in the NCBI Sequence Read Archive, or SRA (it was previously known as the Short Read Archive, but with arrival of the new long-read NGS technologies, the name was appropriately updated). Your roommate’s data was published there and labeled SRR1705851, so you can download it from the SRA FTP server and unpack it on your machine:
http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/

2. Align your roommate’s data to the reference sequence

To look for mutations, we will need the reference sequence for the influenza hemagglutinin gene. You can download it here (just copy and paste, or select Send to -> File -> FASTA format).

For advanced users:

> *The reference sequence for the influenza hemagglutinin gene is in the public file and is available from NCBI GenBank. You can download data from GenBank automatically using the EntrezDirect utility (AKA eutils). You can install it and use the EntrezDirect command 'efetch' to download the reference sequence, which has the NCBI id number: KF848938.1*
>
> *You must specify which NCBI database to use, the id or ascension number of the sequence, and the format you would like. You can redirect the output into a file with the ".fasta" extension.*
>
> *There are versions of eutils for R and Python, see some examples here.*

Index the reference file, align your roommate's viral data to the reference sequence and make an mpileup. Since we learned how this works last time, you can do most of it in one line with pipes. This will also avoid making all those intermediate files. See example for bwa mem (remember, you can use any modern aligner and variant caller of your choice)

```
bwa mem reference.fasta roommate.fastq | samtools view -S -b - |
samtools sort - roommate
```

Samtools regards a "-" as the standard input (stdin). For more details see samtools documentation.

It's good to check that we selected correct reference - you can count percentage of reads that mapped (*samtools view -f4*)

3. Look for common variants with VarScan

Index the bam file, and make an mpileup (consult with your lab notebook from Project #1). To save computing power, the default behavior of samtools is to stop piling up the base calls at each position when it gets to 8000 calls. Since our variants may be quite rare, set that depth limit to something we know is higher than our coverage with the -d flag.

(What is the desired value of d if we want to keep all possible variants, given the length of the reference sequence, the number of reads, and the average read length?)

Run VarScan on the mpileup. First, look for positions where most of the viruses infecting your roommate differ from the reference. Maybe there is a common mutation that wouldn't have shown up in the HI test. Use a high minimum variant frequency cut-off (N) to find only those mutants present in most (95% or more = 0.95) of the viral DNA molecules.
(How many variants are reported back?)

You can pull out the variants in a convenient format using awk or any parsing tool of your choice. Awk excels at parsing delimited data where you have a lot of fields, like in a vcf file. The simplest form of an awk command is (all one line):

```
awk '/search_pattern/ {actiontotakeonmatches; another_action;}' file_to_awk
```

Awk has a lot of cool features, and it is worth learning more. But for now, we just need to extract some particular columns from our file. awk script below only output the fields you are interested in (reference, position, alternate).

```
cat filename.vcf | awk 'NR>24 {print $1, $2, $4, $5}'
```

NR stands for number of read lines, so it's saying 'only if the line number is >24, do what's next'. Awk uses the dollar sign to mark fields (columns), so what's next is 'print columns 1,2, 4 and 5".

What do these mutations do? Could they be what allowed your roommate's virus to escape the antibodies in your body from the flu vaccine? Since we are only looking at a single gene, we will do this manually.

You can open your reference and vcf file in IGV. Use the codon table to check if these mutations can affect the protein.

4. Look for rare variants with VarScan

Now try looking for rare variants. Set the minimum variant frequency to 0.001 (0.1%) and run the scan again on the same mpileup file.

How many variants are reported back now, and how abundant are they? There is a value in the FREQ field buried deeply in the 10th column of vcf file - you can eyeball it or, again, pull it out with *awk* (better).

Wow! You take your data back to your friends at the med school and excitedly show them all of the rare variants you found. They don't seem nearly as excited as you, and when you ask why, they point out that, at these frequencies, it's very difficult to tell the difference between real, rare mutants in the viral population, and errors introduced in the sequencing and amplification process.

They suggest a control reaction in which they will sequence an isogenic viral sample (all virus particles genetically identical) derived from a virus clone that matches the reference sequence. By looking at the errors in this reference and comparing them to the mutations in your roommate's sample, you hope you'll be able to figure out which variants are real.

You asked for help, and your friends took the isogenic (100% pure) sample of the standard (reference) H3N2 influenza virus, PCR amplified, and subcloned into a plasmid. They sequenced it three times on an Illumina machine.

Any “mutations” you detect in the control samples which don’t contain any true genetic variants must be due to errors. You can use the frequency of the errors from the control to help figure out what’s an error and what’s a true variant in the data from your roommate.

5. Inspect and align the control sample sequencing data

You can download fastq data for the three controls (from sequencing of isogenic reference samples) from SRA FTP:
SRR1705858: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz
SRR1705859: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz
SRR1705860: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz

Calculate how many reads are in each file, and take a rough estimate of the coverage in your samples. Remember, coverage is the number of reads corresponding to each position in the reference, so you also need to know how many base pairs long the reference sequence is.

Align each control fastq file to the reference (KF848938.1.fasta), convert it to a bam file, and sort it. You need to make a separate bam alignment for each control sample, but you do not have to index the reference sequence again. Be sure to give each bam file a unique name.

Use samtools to index each of the new control alignment (bam) files.

6. Use VarScan to look for rare variants in the reference files.

Run VarScan with a minimum variant frequency of 0.001 (0.1%) on each of the reference alignments. Be sure to tell VarScan to only output variants and to format the output in the vcf format.

Parse each vcf file so that you get three lists containing the reference base, position, alternative base, and frequency. Copy those lists into a spreadsheet of your choice (e.g. Excel, Google Sheets, R, etc.) and, possibly, into your lab notebook.

7. Compare the control results to your roommate’s results

Examine the data in the spreadsheet. Are there any positions in your roommate’s sample (aside from the common high-frequency ones) that stick out as possibly being more than just sequencing error? Calculate the average and standard deviation of the frequencies reported within each list (you should have 3 averages and 3 standard deviations, and put them in the same units). (*Should we be using common (high-frequency, >90%) variants for this calculation?*)

Did VarScan report rare mutations in your roommate's file with frequencies that are more than 3 standard deviations away from the averages in the reference files? If there are, use IGV to identify the original amino acid at each position, position number in the protein sequence, and the amino acid resulting from the mutation.

8. Epitope mapping

Use the epitope locations listed in Munoz et al to determine if any of the high confidence (greater than 3 standard deviations away from reference error rate) mutations from your roommate's flu infection are located in an epitope region of hemagglutinin (epitopes are the parts of the protein structure recognized by antibodies). If so, list which epitope regions are mutated.

For your lab report

For the abstract, remember that your goal is to figure out how you got the flu even though the HI test said your roommate's flu strain was covered by the vaccine.

For the introduction, briefly cover how the flu vaccine works, and the idea of antigenic drift and viral quasispecies. Also provide some background on targeted deep sequencing for studying mixed populations, and the sources of error in next generation sequencing that this lab project tried to correct.

In your methods section, briefly introduce the sequencing data (what are the samples, how many cycles was the sequencing run, was the data pre-processed?). Describe how you used VarScan to examine both your roommate's sample and the reference samples. Be sure to describe the parameters you used if they were not the default options.

In your results section, include the number of reads you started with and the number of reads that mapped, for each of the 4 data sets. Report the average and standard deviation of the frequencies from each reference sample. For your roommates data, make tables or a list containing the common mutations, and any rare mutations you think are truly represented in the viral population. List these mutations both in terms of the DNA base, and the protein amino acid changes, and state whether they are non-synonymous or synonymous. You do not need to list all of the mutations reported by VarScan in your lab report, but include them in your notebook.

State whether any of the mutations affect the epitope regions of the hemagglutinin protein and if so, report which epitopes they affect. In the discussion, be sure to explain how you decided which mutations were most likely to be real.

Given your results, explain how you think you were able to get the flu from your roommate, even though you had received the flu vaccine. Also in the discussion section, propose at least one additional way to control for error in deep sequencing experiments like this, and explain why error control is important for accurately identifying and quantitating rare variants. Our approach was pretty quick and simple; there are many more sophisticated methods out there. You can suggest laboratory steps to minimize errors in the first place, bioinformatics steps you could implement on our data, or existing software. For each suggestion, include a sentence or two explaining how it would reduce error.

***Optional Extra-Credit Challenge Question**

For 1 bonus point on this lab report each.

1) How would you calculate the ACTUAL average coverage per position for one of our data sets, only for mapped reads, and taking into consideration the fact that the reads can be not all the same length? You can use a script or software that someone else wrote - in this case please explain how it works, and how you would call it at the command line. Include your approach, and your answer, if you found one, at the end of your lab report, after the discussion.

2) Using positions reported by VarScan in all 3 of the reference sequences, can you distinguish PCR ("upstream") and sequencing ("during") errors? Provide average and standard deviation for both types of error.

3) If you are familiar with the PDB database, you can try to explore VMD, PyMOL, Jmol, RasMol, or some other PDB-viewing application to provide an image of the H3N2 hemagglutinin molecule and highlight amino acid changes you've found.