Snail Sequences?

I was looking over the FastQC statistics for your Pokeweed and 4o’clock Illumina libraries, with help from Scott and Dhivya. FastQC reports “Over-represented sequences”, which is a list of the top most-frequent reads taken from a sample of the library. Unsurprisingly, one of those over-represented sequences was adapter. But the rest didn’t have any hits from FastQC. I BLASTed them against nt, and the top hit for all of them was: “Lymnaea stagnalis NMDA-type glutamate receptor mRNA, complete cds”. Lymnaea are snails. There were other animal hits as well, but L. stagnalis was top of all of them. So we started to wonder whether your samples had animal, particularly snail, sequences in the mix!

Looking at the non-adapter over-represented sequences, Scott noticed that they were all very similar to the first two; it looked like they were all variants of two particular sequences. I took the first 25bp of the top two non-adapter sequences, to serve as representative sequences of the two classes of over-represented sequences. I then started searching your samples for those 25-mers to get a sense of how common they are. Actually, to speed things up, I just took the first 250,000 reads of the samples and checked those, to speed things up. Here are the counts we got:

**SA12001 sequencing run:**
- `'AGTGGACTATCCATGAACGCAAAGC'` 4o'clock_top250k_R1.fq : 34314
- `'AGTGGACTATCCATGAACGCAAAGC'` 4o'clock_top250k_R2.fq : 43667
- `'CTGCGTTGATACCTGCTTTGCGT'` 4o'clock_top250k_R1.fq : 20631
- `'CTGCGTTGATACCTGCTTTGCGT'` 4o'clock_top250k_R2.fq : 12410
- `'AGTGGACTATCCATGAACGCAAAGC'` Pokeweed_top250k_R1.fq : 42244
- `'AGTGGACTATCCATGAACGCAAAGC'` Pokeweed_top250k_R2.fq : 61
- `'CTGCGTTGATACCTGCTTTGCGT'` Pokeweed_top250k_R1.fq : 22986
- `'CTGCGTTGATACCTGCTTTGCGT'` Pokeweed_top250k_R2.fq : 9363

**SA12028 sequencing run:**
- `'AGTGGACTATCCATGAACGCAAAGC'` Spinach_top250k_R1.fq : 27695
- `'AGTGGACTATCCATGAACGCAAAGC'` Spinach_top250k_R2.fq : 29407
- `'CTGCGTTGATACCTGCTTTGCGT'` Spinach_top250k_R1.fq : 90
- `'CTGCGTTGATACCTGCTTTGCGT'` Spinach_top250k_R2.fq : 27
- `'AGTGGACTATCCATGAACGCAAAGC'` Molluga_top250k_R1.fq : 24809
- `'AGTGGACTATCCATGAACGCAAAGC'` Molluga_top250k_R2.fq : 26372
- `'CTGCGTTGATACCTGCTTTGCGT'` Molluga_top250k_R1.fq : 75
- `'CTGCGTTGATACCTGCTTTGCGT'` Molluga_top250k_R2.fq : 41

I search for occurrence of the same 25-mers in other samples in the same sequencing runs. Unfortunately, I am having trouble getting access to the SA12001 sequencing run (something is goofy with the iRODS server, but it’s being worked on). But I got some sample data from the SA12028 run. I used Mpenn (a rodent) from Lauren O’Connell that was in the same lane as the Spinach, and I used Copenhagen_Lung_1 (whatever that is) that was in lane 6, between Molluga (lane 5) and Spinach (lane 7).

- `'AGTGGACTATCCATGAACGCAAAGC'` Mpenn_top250k_R1.fq : 22
- `'AGTGGACTATCCATGAACGCAAAGC'` Mpenn_top250k_R2.fq : 30
- `'CTGCGTTGATACCTGCTTTGCGT'` Mpenn_top250k_R1.fq : 0
- `'CTGCGTTGATACCTGCTTTGCGT'` Mpenn_top250k_R2.fq : 0
- `'AGTGGACTATCCATGAACGCAAAGC'` Copenhagen_Lung1_top250k_R1.fq : 0
- `'AGTGGACTATCCATGAACGCAAAGC'` Copenhagen_Lung1_top250k_R2.fq : 0
- `'CTGCGTTGATACCTGCTTTGCGT'` Copenhagen_Lung1_top250k_R1.fq : 0
- `'CTGCGTTGATACCTGCTTTGCGT'` Copenhagen_Lung1_top250k_R2.fq : 0

The numbers are WAY lower (well, they’re zero across the board for Copenhagen!) for the samples that aren’t yours in this run.

It looks like these 25-mers are far more common in your samples than nearby sample in the same sequencing run, which indicates that if they represent contamination, it’s probably on the sample side, not the sequencing side.