Primer check on public data over time

1 Summary

We want to see if primer identity varies of time / isolates / region / etc.

1. Make an example primer (primer blast)
2. Download all public nucleotide data.
3. Blast primer against fasta.
4. Tidy and interpret results.

2 Primer make

As a test, I made a random example primer using Primer Blast on the reference nucleotide sequence of G from strain B1 (AF013254.1) and picked the first primer. (Input used GenBank: AF013254.1 and run "get primers" with default settings)

- Ref gene used AF013254.1:
  - https://www.ncbi.nlm.nih.gov/nuccore/AF013254.1?from=4690&to=5589&report=gbwithparts
- Primer blast:
  - https://www.ncbi.nlm.nih.gov/tools/primer-blast/primertool.cgi

2.1 Results:

- Forward_AF013254.1: TGCCTATGGTTCAGGGCAAG
- Reverse_AF013254.1: TCCTGGTTTCTTGGCGTACC

3 Download public sequence data

Next we download viral sequence data from NCBI.

https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType%2F_s=Nucleotide&VirusLineage%2F_ss=Human%20orthopneumovirus,%20taxid:11250&SeqType_s=Nucleotide&HostLineage_ss=Homo%20(humans),%20taxid:9605&VirusLineage_ss=Human%20orthopneumovirus%20(HRSV),%20taxid:11250&ProtNames_ss=attachment%20glycoprotein&CollectionDate_dr=2018-01-01T00:00:00.00Z%20TO%202022-01-07T23:59:59.00Z

3.1 Download settings

- Virus: Human orthopneumovirus (HRSV), taxid:11250
- Proteins: attachment glycoprotein
- Host: Homo (humans), taxid:9605
- Collection Date: From Jan 1, 2018 To Jan 8, 2022

3.2 Results:

- Download 1: Sequence data (FASTA Format) Nucleotide [sequences.fasta]
- Download 2: Current table view result CSV format, all columns, meta data. [sequences.csv]

4 Blast

Then blast each primer against all fasta.

https://blast.ncbi.nlm.nih.gov/Blast.cgi

- Enter Query Sequence:
  - > Forward_AF013254.1
  - TGCCTATGGTTCAGGGCAAG
- Enter Subject Sequence:
  - sequences.fasta
- Program selection:
  - Highly similar sequences (megablast)
- Results:
  - Download as "Hit table (text)"

5 Convert table into tsv

Careful to edit this file. Convert header from comments (#) to tabbed spaces. This (results of blast match) is then used in the R script and merged with sequence meta data (date, geo location, etc). Downloaded Hit table (text):

- XGYK6RVT114-Alignment.txt Modified Hit table (text):
- XGYK6RVT114-Alignment.tsv

6 Analyse primer blast

The bulk of work is in deciding how to best interpret the results. There are severl criteria for confirming primer efficiency, but basically we want 100 match to the majority of sequences and be able to see when there is a deviation.

- Method R script: primer_check.R
- Final output of html plot:
  - **Forward_primer_match_TGCCTATGGTTCAGGGCAAG.html**

This output shows the first primer (forward) which has the same binding to all samples:

- 1 mismatch (19 of 20 nt)
- 95% identity
- bit score 32.2
- alignment length 20 (full primer match)

Note: there are other measures that can be included. And more external tools could be applied to confirm PCR settings; Tm (°C), %CG content. I like to use https://www.thermofisher.com/ch/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html.