Primer design notes

- Multiplex primer scheme generated for rabies virus (RABV), targeting RABV lineages found circulating in Southeast Asia

Steps

Getting the data:

- Used RABV-GLUE to filter publically available rabies virus sequences
    - Go to http://rabv.glue.cvr.ac.uk/#/home
    - Choose *sequence data* tab, then *all NCBI sequences*
    - Click on *Filters* box to apply filters

Choosing representative sequences:

- My filters:
    - Sequence length > 11800
    - Global region matches South-eastern Asia
- Produced around 20 sequences (this was in March 2019)
- Downloaded sequence and metadata for these 20 sequences by choosing in the download box. Box appears to choose a file name, press ok, then you must click on the file in the *ready to download* box in order to download to your computer (if you just press ok it won't download)
- I then manually filtered through the metadata file to choose the best representative sequences
    - I choose sequences from difference countries (to capture geographical diversity)
    - Tried to include the most recent sequences but also older sequences (to also capture temporal diversity)
    - I also used the sequence data (the downloaded fasta file of sequences) to filter data based on sequence identity. If sequences had >99% identity to other genomes in the file only one representative was kept. Keeping very similar sequences would bias the primer selection so best to remove near-identical sequences.

Fixing the sequence data:

- Some of the sequences contained IUPAC ambiguity codes. (For example, at a particular position there may have been a mixture of potential basecalls (maybe 60% of seq reads had an A, and 40% had a T) and sometimes this is assigned an ambiguity code to represent the mixture. For A or T the IUPAC code is W.). However, Primal Scheme doesn't accept ambiguity codes in the sequences so these positions have to be masked, i.e. changed to an 'N'. I used my a command in the command line to do this:

```
perl -pe 's/[^AGTC\n]/N/gi unless m/>/;' old.fa > new.fa
```

where old.fa is replaced with your original fasta file and new.fa is replaced with what you want to call the new file (with ambiguity codes converted to N

- Sequence gaps (e.g. "-" in the sequence data) have to be masked by N

Prioritising the sequences

- It matters what order the sequences are in the fasta file
- Have to ensure the first genome in the file is the most recent, complete genome for a given geographical region/lineage

- I wanted to prioritise sequences from the Philippines so I put these first in the fasta file
- However, there was some missing data at the beginning and end of the sequence so I spliced in sequence data from another genome. Because the ends of rabies virus genome are highly conserved I decided that it was ok to use sequence data from another more complete genome to fill in this missing data. (But it is not ok to do this for other part of the genome as there may be variation that is important, and may not be the same for other genomes!)

Extra information

- There was only a small number of whole genome sequences available for the Philippines (~5 genomes published), but there are a large number of partial genome sequences available
- In order to incorporate this information in the primer design I downloaded all the partial genome information (for the N gene), made a 99% majority consensus sequence from these data, and added it as an additional representative sequence. The rest of the genome was masked with N bases
- I did this to incorporate any additional known diversity for rabies circulating in the Philippines, as it may improve primer design to capture Philippines RABV lineages

Final set of sequences

- I ended up with a set of 11 sequences (10 whole genomes and 1 partial genome)
- The final set is in the file called "SEasia_selectionPlusNconsensus_aln_masked_upper_spliced.fasta"
- The metadata associated with these sequences is in the file "SEasia_wgs_RepresentativeSeq". This file shows all the whole genome sequences that were available, and the yellow highlighted rows are the sequences I chose as representatives

Primal Scheme

- The final fasta file was submitted to Primal Scheme for primal design, choosing an amplicon length of 400bp with a 50bp overlap
- This generated the scheme shown in the multiplexPrimerScheme folder. I ordered these primers as "budget oligos" from ThermoFisher (25N, desalted oligos)

Note: This was my logic to design primers for rabies in the Philippines/SE Asia but there may have been a better logic!!!