Genomics Homework 4

Xiaojun Gao

We use the following command:

```
krakenuniq --report-file krakenuniq.report —db <> --threads 4 --preload
--preload-size 100G —paired <> <>
```

Replace <> with actual file names.

For sample 1, the species is likely Human alphaherpesvirus 1. The genus name is Simplexvirus. The number of reads is 223, while the k-mers is 5607. It is also the species with the highest percentage, with 40.4%, while the secondary species is only 13.59%. I also looked up the species and see on Google that it is an infectious virus, though not directly related to oscular, that might have cause the oscular infection.

For sample 2, the species is likely Cryptococcus neoformans. The genus name is Cryptococcus. The number of read is 4238, while the k-mers is 348417. It's percentage is 96.43%, which is really a high percentage that almost certainly indicates this is the pathogen. By looking up on Google, it is a fungal that might be the cause of oscular infection.

For sample 3, the species is likely Toxoplasma gondii. The genus name is Toxoplasma. The number of reads is 2400, while the k-mers is 143993. It is also the species with the highest percentage, with 97.28%, which also almost certainly indicates this is the pathogen. By looking up on Google, it is a human parasite that might be the cause of oscular infection.

For sample 4, there is likely no infection since there is no likely pathogen from result of krakenUniq by looking up the pathogens on Google. The top non-human species is Sordaria macrospora, with read number 28 and k-mers 31. Its genus name is Sordaria.

For sample 5, there is likely no infection since there is no likely pathogen from result of krakenUniq by looking up the pathogens on Google. The top non-human species is Ochrobactrum anthropi, with read number 330 and k-mers 5205. Its genus name is Ochrobactrum.

For sample 6L and 6R, the species is likely Rubella virus. Its genus name is Rubivirus. For sample 6L the k-mers is 210 while the read number is 5. For sample 6R the k-mers is 1031 while the read number is 290.

The determination process is somewhat tricky.

For report by 6R, the top non-human species sorted by percentage is Rubella virus, while this species has a comparatively low percentage in report by 6L. Using BLAST on multiple sequences from both 6L and 6R, I am able to say that this is the actual species and that krakenUniq and BLAST identified it to be the same species.

However, I kept on looking for a better match. The second non-human species sorted by percentage for 6R is Ochrobactrum anthropi, which has a comparatively higher percentage in 6L than Rubella virus. Despite this, when I BLAST the corresponding sequences, BLAST showed it to be another species with much higher possibility based on E-value when I tried blasting the corresponding sequences in both 6L and 6R. The species that is likely is Brucella anthropi.

I then looked for at report 6L for the non-human species based on percentage, there are 3 species that have a higher percentage in 5L report than Rubella that are also found in 6R, but they have a very low percentage in 6R so they are not likely the target. As for species with even lower percentage than Rubella virus for 6L, they are not likely since they do not appear in report 6R.

That is, the result should be between Rubella virus and Brucella anthropi. It is difficult to decide. However, Rubella virus has a higher likelihood since it is very prominent in one eye. Also, by looking up on NCBI, perhaps as a way of cheating to get the results, the paper related to it says it is Rubella infection.