# R21 project sequencing results / June 2020

In this project, monkeys were infected with a Mac251 SIV strain.

We have 69 sequenced samples from 11 different monkeys. In addition, we have two samples from the stock virus.

Two main issues make sequence data analysis for these samples difficult.

1. The virus is not very virulent in the monkeys and the viral loads are low.
2. The stock virus was not diverse at all, so the infections were with almost clonal virus.

The result of both is that we have very little diversity to work with. There is very little signal in the data in general.



This plot shows average genetic diversity for each sample. The red dots are from monkeys that were infected with TB and SIV. The blue dots are from monkeys infected with SIV only.

Overall, genetic diversity is very low. We don’t see a clear difference between plasma (full circles) and non-plasma (open circles).

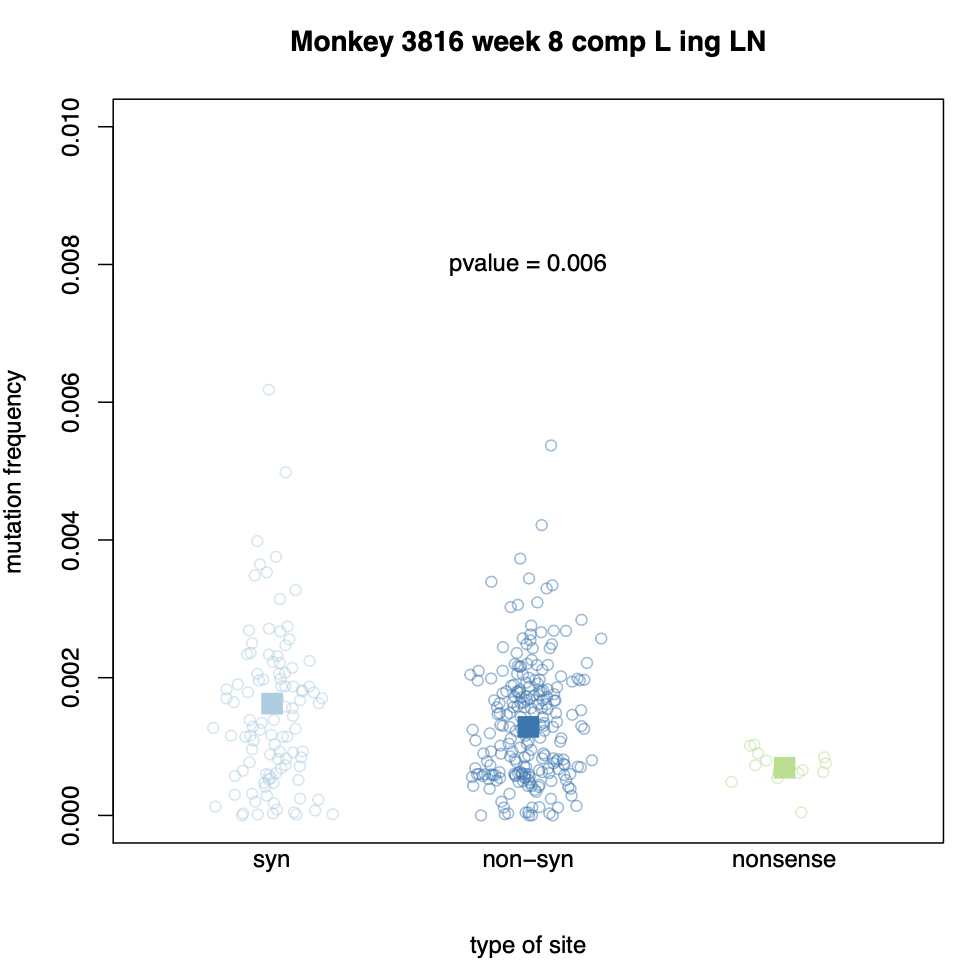
## Frequency of synonymous vs non-synonymous transition mutations.

To see if there is any signal in the data, postdoc Senay Yitbarek and I decided to compare the mutation frequency of synonymous vs non-synonymous sites.

We focused on transition mutations (A<->G and C<->T) because they are more common than transversion mutations in viruses.

We found that yes, there is signal.

In most samples (though not all), the non-synonymous and nonsense transition mutations have lower frequency (according to non-parametric Wilcox test) than synonymous mutations.



What we see in this plot is that in a sample from animal 3816, non-synonymous mutations have, on average a lower mutation frequency than synonymous mutations. The p-value is for a one-sided Wilcoxon test to determine if synonymous mutations have higher frequencies than non-synonymous and nonsense mutations combined.

We find that for plasma samples, 14 out of 19 have a significant difference between synonymous and non-synonymous/nonsense mutations (74%).

We find that for non-plasma samples, 25 out of 44 have a significant difference between synonymous and non-synonymous/nonsense mutations (57%).   
For 6 samples, there are not enough data (positions with at least 6000 read depth) to do the test.

This result, that frequencies of synonymous sites are higher than for non-synonymous sites shows

1. That there is signal in the data. What we see is real diversity, and not just sequencing noise.
2. That there is selection occurring on non-synonymous and nonsense mutations in the macaques. Selection makes these sites less frequent.

Next steps:

Now that it is clear that there is signal in the data – we can look at other effects that can be studied in low-diversity samples. The first thing we’ll look at is CpG-creating mutations (see Theys et al 2018 Plos Genetics).