# R21 project sequencing results / June 2020

**Abstract**

In this project, 11 monkeys were infected with a Mac251 SIV strain.

7 of them TB and SIV, 4 of them only SIV.

We found no evidence of selective sweeps (except in one monkey at one site). This is likely because the virus didn’t replicate well in the macaques and because the monkeys were only infected for 8-9 weeks (compare Ita et al 2018). We find evidence for mutation selection balance with non-synonymous and nonsense mutations being present at lower frequencies than synonymous mutations.

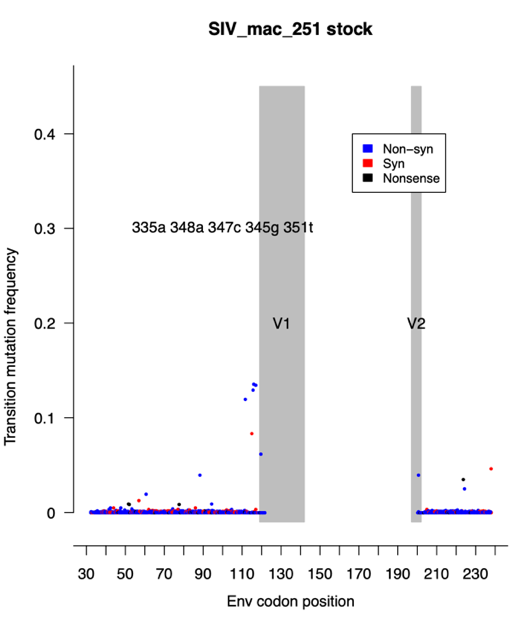
We have 69 sequenced samples from 11 different monkeys. Four of the monkeys were infected with SIV only, while the other 7 monkeys were infected with SIV and TB. In addition, we have one sample from the stock virus.

Three 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.
3. The monkeys were only infected for 8-9 weeks.

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

## Diversity in stock virus

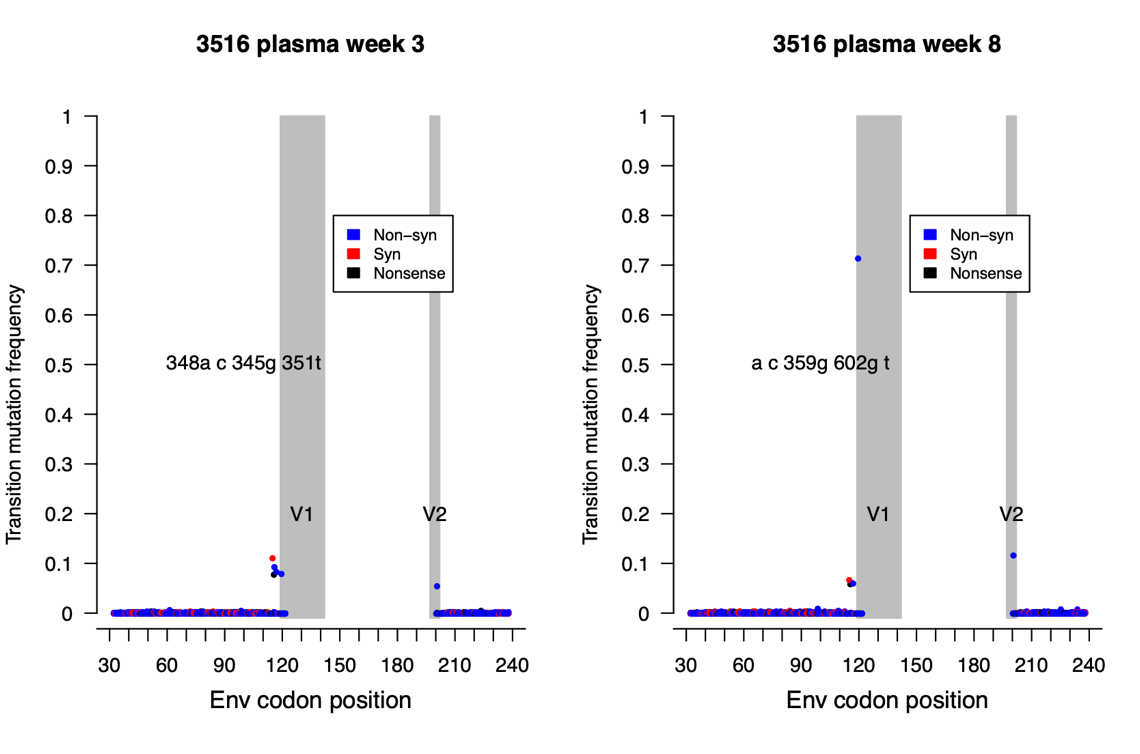


Diversity in the stockvirus is low – lower than in an otherwise similar study (Ita 2018).

Diversity in the monkeys is even lower than in the stock. Likely, there was a severe bottleneck at infection, and not enough replication to generate a lot of diversity in the 8-9 weeks after infection.

Also, it is likely that we are missing diversity in the V1 loop because most of the V1 loop is not sequenced due to the primers being too far away from the loop.

In one of the monkeys there is evidence for a soft sweep (starting from standing genetic variation) at position 359g (K120R), moving from 10% freq to 70% freq.



## Diversity higher in plasma than tissue, and lower in SIV-only animals



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. According to a non-parametric test, plasma samples have higher diversity (0.69% on average) than tissue samples (0.58% on average) (p = 0.014, two-sided Wilcoxon test). When we focus on the plasma samples, the SIV-only animals have lower diversity (0.63% on average) than the co-infected animals (0.71% on average) (p = 0.02, one-sided Wilcoxon test).

## No difference between genetic drift in plasma, lymph nodes and lung tissue.

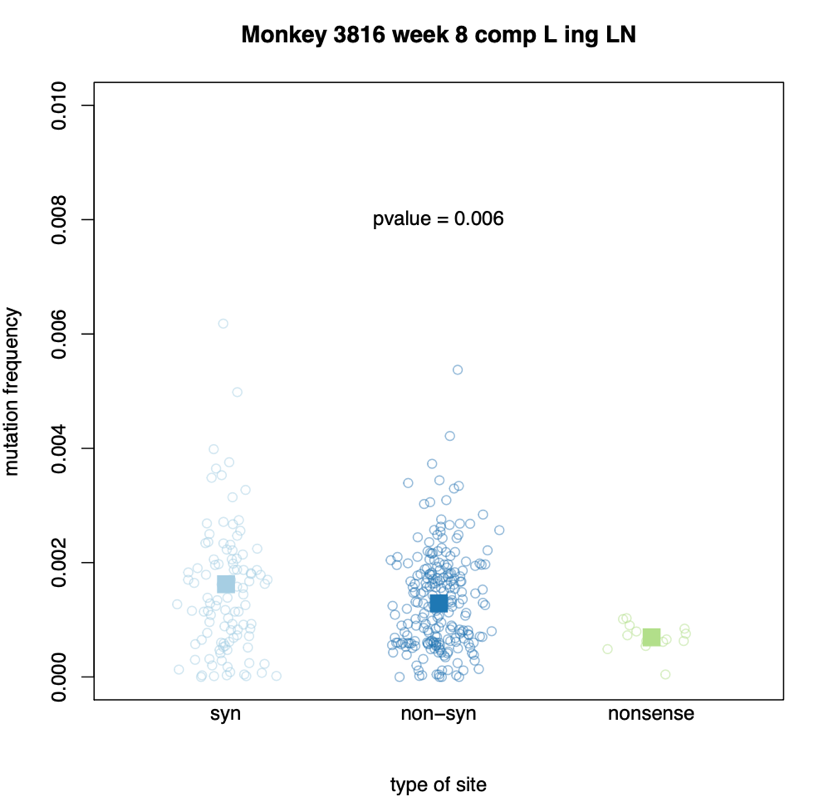
We were interested to find out whether the virus behaved differently in tissue versus plasma. Particularly, we predicted that virus may undergo more genetic drift in the tissues than in the blood plasma. For this analysis we focused on sites that had a fairly high frequency in the stock virus (17 sites with a mutation at 0.8% frequency or higher). For these high frequency sites, we determined their frequency in the early plasma sample (at 2-5 weeks) and compared that baseline frequency to the frequency of the same mutation in (1) the late plasma sample (8 or 9 weeks), (2) the lymph node samples and (3) the lung tissue samples. In general, the frequency changes were modest. We found that the mean absolute frequency change between the plasma samples was 1.7% of the early plasma frequency, between the lymph node samples and the early plasma samples it was also 1.7% of the early plasma frequency and between the lung samples and the early plasma samples it is 2.3% of the early plasma frequency. A non-parametric test (Wilcoxon) for all comparisons was not significant.

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

We decided to compare the mutation frequency of synonymous vs non-synonymous sites. Here 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, 16 out of 19 have a significant difference between synonymous and non-synonymous/nonsense mutations (84%).

We find that for tissue samples, 26 out of 50 have a significant difference between synonymous and non-synonymous/nonsense mutations (52%).

This difference may be due to better sample quality of the plasma samples or due to higher effective population sizes in the blood.

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.