**Easy questions we can answer relatively quickly:**

1. Revisit the properties of the CDR3 based on amino acids now that the data set is much larger. This would mean rerunning a lot of the same analysis that were already setup for the first set of samples.
2. What is the Shannon diversity at the level of particular gamma chains and not the whole population?
   1. For instance, do we find that Vg4 is more or less diverse than Vg3 etc?
   2. Does this segregate with tissue?
   3. Does this segregate with disease?
3. Can we do an unbiased clustering of the groups by gamma chain based on properties of the gamma chains
   1. For instance will the Vg4 and Vg9 cluster together since they lack positive charges in the CDR loops?
      1. We would need to assign properties to the amino acids
      2. Would could do it based only on the CDR1,2,3 sequences
4. What is the usage of Histidine in the CDR3 data set?
   1. By group
   2. By tissue
   3. By delta or gamma chain
   4. Is it enriched in particular gamma chains?
   5. Since I know we will find it enriched in Actives, we should test to see how it segregates with the two groups of Active CD based on diversity index (Figure 4B). Split the diversity into the top vs bottom 50% by patient and then ask in those groups what is the representation of Histidine in the CDR3
      1. We could also try looking at the framework regions independently or in addition to the CDR loops

**The more difficult question but most important from the biological perspective. Question 4 above is related to the logic below.**

1. The most critical question we want to answer is whether there are there particular amino acids/motifs that are enriched in the CDR3 in Active CD vs. Controls (GFD are also a part of this analysis but the more interesting question is when comparing Actives to GFD)
   1. Look across groups
   2. Look across tissues
   3. Look at both delta and gamma chains
   4. If you find an enriched amino acid in Actives for example, test if it is more enriched in the top 25% of clones as opposed to the bottom 25% of clones. This would suggest antigen drive
   5. Once these amino acids have been identified then we can ask if they are germline or non-germline encoded (N vs. P). This is where we will have to use the Vquest files. I would again focus on the gamma chain first since it is easeier.
   6. One thing we can consider here is to filter the sequences by J usage so that the J doesn’t drown out the signal at the tail end of the CDR3 when doing the alignment based algorithms