210614 to 210618

* Applied PCA on the top 1000 genes using:
  + Raw counts
  + Logarithmic scale log2(raw counts + 1)
  + Standardized counts on a logarithmic scale
* Implemented the mean-variance plot in R and in Python:
  + In python I colored the samples per tissue type
* Ran the top 1000 genes by Gene Ontology
* Applied the T-sne method on the top 1000 genes and the complete dataset after filtering and normalization.
  + In this case I only used the standardized counts on a logarithmic scale
* Performed hierarchical clustering using the data for the top 1000 genes:
  + I tried to plot it interactively using dash-bio but it was too big for the browser to load the 9662 samples at once, so I plotted the clustermap using seaborn.
  + The correlation matrix was calculated based on log2(counts +1).
  + I tried the different methods to perform the clustering and the ‘single’ method gave the best results.
* Finally, I tried to optimize the code so it could be easily maintained and re-executed (I was thinking for the TCGA dataset), however it started behaving differently therefore I stopped, I will try to re-write whenever I am free.

210621 plans:

* Add the ensemble gene identifications for the list of histone variants that I compiled (I used the Gene database but the data I have uses Ensembl)
* Perform the dimensionality reduction techniques on this list.
* Perform the hierarchical clustering
* Compare the histone variant and the histone chaperone expression patterns.
* Apply the previous work on TCGA dataset (depending on the computation time I am not sure that I could apply everything this week).