From 210621 to 210625 :

* I re-wrote the code using Argparse module so I can maintain the code and re-use it easily. It is easier now to execute on several files, and the outputs will be automatically saved in the appropriate folder.
* Performed the dimensionality reduction techniques on the dataset limited to the histone variants and the chaperones, the PCA doesn’t perform well as previously, the T-sne however still does a good job even if it is a little bit worse.
* What raises questions however is the T-sne’s ability to separate samples even when the genes were randomly selected, so may be an in depth look to the technique would explain this.
* Performed the hierarchical clustering on the normalized dataset (~ 22000 genes), the samples are clustering by tissue, and it is very clear.
* I have got the same results when used only the top 1000 expressed genes.
* However, when using the dataset limited to the chaperones and histone variants the clustering wasn’t very clear, except for the brain, blood and some other few tissues, in which the expression is not really high relative to the other tissues, but it is high enough to let them cluster together.
* I am currently running the code to cluster the randomly selected 76 genes, but from what I tested previously for one random set, they do cluster by tissue for most of the tissues and they are mixed for some, brain and blood are always clustering in a really good way.
* For the gene-by-gene clustering, I still don’t have the results yet because the code kept stopping, but from what I saw when I was testing the replicative histones cluster together and the chaperones together, this was noticed also in the clustering of samples by genes.
* So, for the clustermaps of samples by genes:
  + For the top 1000 expressed genes the samples are clustering by tissue, and it is very clear, the genes also present some clustering, but I can’t tell which genes, for that I will need to see how to get the gene names, if necessary, overall, there are 5 clusters.
  + For the chaperones and histones dataset, the tissues are not clustering very well, although there seems to be two clusters if we focus on the lowly expressed genes.
  + As for genes, we can observe that the replicative histones are either expressed at a low or medium level, the fact that they are lowly expressed might be due to the library construction protocol (True Seq). However, there are some chaperones that correlate with this set of histones, and they are the chaperones that are responsible for the replicative histones.
  + CENP-A and HJURP are always clustering together, in the lowly expressed region and we can see that for some tissues the expression level is higher, this could be in relation to the transformed cells present in the dataset.
  + What is also interesting is the fact that H3F3B and H3F3A don’t cluster together.
* When checking the GTEx expression profiles for these genes, they do fall in the same line with what I am observing, but from ENSEMBL there are variation between the datasets.
* I am currently running the code to cluster not only the randomly selected 76 genes (10 datasets) but also the top 76 expressed genes and the dataset without the replicative histones.
* I also generated an unweighted network for genes that have a correlation over 70% but it wasn’t looking good, so I will need to figure out how to better represent it using an interactive package.

This week’s plans:

* Finish the clustering of the previously mentioned datasets.
* Check how the different metrics perform the clustering to understand the difference they deliver.
* Continue investigating the genes expression on GTEx (I checked the chaperones, now I will check the histones).
* Check the ENSEMBL arrangement of HIST1.
* See if there is any detectable batch effect in the dataset and try to correct it using BatchQC (if possible), because this could be the reason why PCA is not working well.
* Code a better representative plot for the co-expression network
* Check how to use WGCNA with R because it is the only available package, so I can validate my work against it or use what it gives if necessary.