Team Members: Nathan Vallapureddy, Kaushik Srinivasan

In our project proposal, we outlined how we are going to classify SNP-genes pairs and its associated tissue specificity. We are planning to use deep learning to recognize patterns in our SNP-gene pairs. Since our proposal submission, we have made respectable progress. Our PhD supervisor for our project, Yuan He, has provided us with SNP-gene genetic coordinates for brain and non-brain tissues.

Kaushik initially attempted to use samtools to obtain genetic code online, but changed to using UCSC's (Univ. California Santa Cruz) genomic database. UCSC returns the sequence formatted as XML, compared to samtools' FASTA format, it should be easier to incorporate into our python environment. As our data corresponds to HG19 sequences, there should not be significant discrepancies between the two genomic banks. Using BeautifulSoup (a web parser), Kaushik is writing a function to streamline obtaining genetic code given parameters.

Meanwhile, Nathan has been testing various methods encoding for the genetic data to input into the neural network. He has been reading papers on deep learning on SNP genes, especially a couple by Anshul Kundaje, from Stanford University, who works extensively on deep learning for genomic data. He is also setting up a framework to analyze some of the metadata associated with the SNPs (e.g. chromosome number), and evaluating different loss functions to account for an imbalance in the data (~2,000 SNPs in the brain vs ~20,000 non-brain SNPs)

We predict that Kaushik should be able to finish the function for getting genetic code by Wednesday, and we would have feasible results by Friday of this week. In the following week, we would be testing various parameters, and we would find the optimal number of k-mers. Nathan will have his framework completed by Thursday, and we predict that Nathan will be able to incorporate Kaushik's genetic code by Sunday.