**BMI 5330 Project Update**

Timothy Bi

Database curation:

At present, I have downloaded .vcf files and related files (reference GRCh38) for dbSNP, dbVar, and clinVar.

* **dbSNP:** For GRCh38, dbSNP stores .vcf files with tabix indexing for convenience. Based on RefSeq, the location for NPM1 is NC\_000005.10 (171387116..171410900). I accessed all the NPM1 variants and stored them into a separate file using the following command:  
    
  tabix GCF\_000001405.40.gz NC\_000005.10:171387116-171410900 > NPM1\_snvs.tsv

We can make a similar file for FLT3 using the same command and the coordinates from RefSeq (NC\_000013.11 (28003274..28100576, complement)). It bears noting that FLT3 is transcribed on the reverse strand.

Next, I plan to read these .tsvs into R for further cleanup (e.g. ensure that all variants are for NPM1, parse the “info” column of the vcf). I will then classify each mutation based on its type (silent, missense, nonsense, frameshift, etc.), as well as produce salient graphs. This will give us a better idea of the mutational landscapes found in NPM1 and FLT3. Then, I will stratify dbsnp mutations by their pathogenicity (as documented by ClinVar) and report any trends that I observe.

* **dbVar:** After further consideration and experimentation with dbVar data, I have decided to not use the dbVar database in my project and will instead will focus on dbSNP, clinVar, and COSMIC. The first reason is that structural variants change much more than just the function of a single gene, and thus many deletions / duplications are not relevant to the scope of this project. The second is that certain common structural variants, especially translocations, are already used in prognostic criteria separately from NPM1 and FLT3 mutation status, which affects our ability to link structural variants and pathogenicity to an NPM1- or FLT3-specific effect. We can use COSMIC instead to identify structural variants in NPM1 / FLT3 (such as persistently activating translocations).
* **ClinVar:** To be explored more thoroughly. Right now, my plan is to build a table as a reference and compare it with dbSNP and COSMIC (see below for more details). dbSNP contains clinVar IDs for each mutation if they exist, making it easy to identify other pertinent clinical information not found in the dbSNP record.
* **COSMIC:** This database contains a large variety of information on expertly-curated cancer mutations, which may provide greater specificity than ClinVar. After exploring the files available, I will focus on the “census genes mutations” file. Other files, such as for fusions or copy number analysis, may also be useful to study in lieu of dbVar for structural variations that specifically involve NPM1 and/or FLT3 (especially since their functions are reviewed by experts and should be high quality).  
    
  To isolate oncogenic NPM1 and FLT3 mutations from the consensus mutations file, I used the command (replacing NMT1 with FLT3 for the latter):

awk -F"\t" '{if ($1=="NPM1") {print}}' Cosmic\_MutantCensus\_v98\_GRCh38.tsv > NPM1\_consensus\_mutations.tsv

From a cursory parse through the output for NPM1, it is evident that a four-nucleotide frameshift insertion (TCTG) makes up the bulk of the samples. COSMIC documents mutations based on unique samples (i.e. there are over 3000 samples that all document this TCTG insertion, so there are 3000 entries for the same mutation). Therefore, I will reduce this table in R to contain only unique mutations (COSMIC also provides a unique mutation ID which will be helpful in this regard). I will also cross-compare annotated COSMIC mutations with clinVar mutations to observe the concordancy of the two databases and their pathogenicity classification.

Finally, I have briefly explored the copy number alteration (CNA) and fusion files. Notably, NPM1 has a number of fusion products, while FLT3 does not. I use a similar awk command as above to parse the CNA and fusion files. Further analysis of CNA will include identification of putative “hotspot” regions for duplication and / or deletion, while further analysis of fusion products will include a literature search for the fusion product and its implications towards AML / other diseases.

Single cell analysis:

I have identified a dataset of interest from this paper (doi 10.1016/j.cell.2019.01.031) which contains a cohort of 16 AML patients with matched whole exome sequencing to identify their mutational burden. Of these, 4 patients have both NPM1 and FLT3 mutations, as well as one patient with NPM1 and one patient with FLT3 mutation. The strategy for analysis is to first harmonize all samples using Harmony, then identify leukemic clusters for further analysis. Previously, I had proposed using a known leukemic score to determine likely leukemic clusters; however, upon closer examination, the authors of this article mention using a machine-learning algorithm to classify leukemic and non-leukemic cells. Therefore, if possible, I will use the annotations provided by the authors. The authors have uploaded both raw and annotated to the Gene Expression Omnibus database (GSE 116256).

Upon identification of I will perform differential gene expression between leukemic cells in samples. If NPM1m/FLT3m, NPM1m/FLT3wt, and NPM1wt/FLT3m leukemic cells primarily populate different clusters, I will run differential expression between said clusters to identify transcriptional differences between the different categories. Otherwise, I will run differential expression between leukemic cells of different NPM1/FLT3 mutational status. This will give us an idea of common downstream targets that may be modulated by NPM1 and/or FLT3 and explain the molecular interactions which may help to explain the prognostic effects of said mutations.