**BMI 5330 Final Project Proposal**

**Introduction and Background**

Acute myeloid leukemia (AML) is a cancer that is derived from hematopoietic stem cells in the bone marrow. It is characterized by a differentiation block in the myeloid-lineage pathway of hematopoiesis, resulting in the accumulation of immature leukemic blasts1. Over time, these immature blasts crowd out the bone marrow space, resulting in impaired hematopoiesis. As such, patients commonly present with non-specific symptoms that are related to low mature blood cell counts (aka pancytopenia), such as fatigue, weakness, pallor, frequent infections, and/or hemorrhagic symptoms2,3. Eventually, patients tend to succumb to infection-related or hemorrhagic complications due to underlying neutropenia or anemia4.

A number of important driver mutations have been implicated in the pathogenesis of AML5. These mutations affect a wide range of targets, from transcription factors (such as RUNX1, CEBPA, and GATA2) to epigenetic modifiers (such as DNMT3a, TET2, IDH1/2, ASXL1, etc.) to non-leukemia-specific tumor-suppressor genes (such as TP53 and WT1). This in turn contributes to the complex molecular landscape of AML, the wide variety of interventions available, and highly variable relapse rates6. Of these known driver mutations, two are of particular interest because of their unique relationship in determining AML prognosis: NPM1 and FLT3.

NPM1 is a protein that participates in a wide range of biological functions related to cell cycle integrity, including ribosome biogenesis, histone chaperoning, centrosome duplication, maintaining nucleolar structure, and DNA repair / tumor suppression7. Its role requires it to effectively shuttle between the nucleoplasm, the nucleolus, and cytoplasm (mutations commonly sequester NPM1 to the cytoplasm). It is the most commonly mutated protein in AML and is considered a licensing mutation. FLT3, on the other hand, is a cell-surface receptor tyrosine kinase that binds to FLT3-ligand (FLT3L) to dimerize and activate, stimulating cellular proliferation and inhibiting apoptosis and differentiation8,9. Mutations to FLT3 cause it to become constitutively activated, releasing it from dependence on FLT3L.

Current guidelines on stratifying AML risk categories based on mutations to NPM1 and/or FLT3 depends on the mutational status and expression levels of both proteins1,7, 10-11. Contrary to what one might intuitively expect, mutated NPM1 confers a protective benefit (lower risk category). For instance, wild-type NPM1 with high levels of FLT3-ITD is considered adverse risk, while mutated NPM1 with FLT3-ITDhigh is only considered intermediate risk. While the prognostic relationship between NPM1 and FLT3 is well-established, the exact molecular mechanism of this relationship is not well understood. Are these mutations affecting two biologically independent pathways? Or are there common upstream / downstream targets that are affected by NPM1 and FLT3 mutations? We will seek to answer these questions throughout the course of this project.

**Aims**

Aim 1: Identify common mutations in NPM1 and FLT3 and their frequencies in adult AMLs, and their clinical significance.

To develop a holistic understanding of the landscape of NPM1 and FLT3 mutations in AML, we will need to characterize the types and frequencies of known mutations. Specifically, we will use existing databases to identify three types of mutations: SNPs, indels, and structural variations. For SNPs, we will further characterize them as being silent, missense, or nonsense mutations. Missense mutations will be characterized by the nature of the amino acid change, and nonsense mutations will be characterized by amount of truncation. For indels, we will further characterize whether mutations are frameshift mutations. Where available, we will incorporate allele frequencies into our analysis to determine the frequency of mutations in the population at large.

Next, we will identify mutations that are clinically pathogenic based on database annotation. If based on our previous analyses we identify common mutations that do not have clinical annotation, we will perform a careful literature search to confirm their clinical status. Then, we will compare the frequencies of various types of mutations between (presumed) benign and pathogenic mutations to determine which kinds of mutations are more likely to be pathogenic.

**A diagram of a gene expression

Description automatically generated with medium confidenceA diagram of a gene expression

Description automatically generated with medium confidence**Aim 2: Identify potential shared upstream drivers and/or downstream targets of NPM1 and FLT3

To elucidate the molecular mechanisms behind the possible interaction between NPM1 and FLT3, we will analyze single-cell RNA-seq datasets from patients with AML with known mutational profiles. Specifically, we will choose a dataset with clinical annotations and patients who have mutations in only NPM1, only in FLT3, or in both. We will correlate patient survival with mutational profile and compare with current risk classification algorithms. Differential gene expression profiles will be obtained to identify genes of interest that may play a role in driving differences in prognosis. Finally, we will attempt to confirm our findings with existing gene regulon databases or a literature search.

**Fig 1. Single cell analysis pipeline for AML cohort with known mutational status.** Patients with mutations in NPM1, FLT3, or both will be analyzed using scRNA-seq before differential gene expression and survival analysis are performed.

**Significance and Rationale**  
As mentioned previously, the prognostic significance of various mutations and cytogenetic abnormalities in AML is well-established. Nevertheless, the underlying molecular biology that drives risk classification of these mutations is not well understood. And while the functional effects of mutations in NPM1 and FLT3 have been studied individually, we have very few mechanistic insights into how these mutations can work in tandem to affect AML progression. Evidently, there is a knowledge gap which could contain new therapeutic possibilities or insights as to how we can improve current methods of treating AML. For instance, we may gain insight into how to better leverage the protective effect of NPM1, or how to better target high FLT3 levels (in both instances perhaps by modulating the other).

Furthermore, while much work has been done to characterize the complex microenvironment and spectrum of malignant cell types that can be found in AML using single-cell genomics12,13, few have focused their efforts specifically on teasing apart the dynamics of possible interactions between NPM1 and FLT3. Nevertheless, the utility of single-cell genomics experiments lies in their adaptability to a variety of scientific questions. Therefore, we can easily use existing datasets of AML patients with known mutational profiles that include NPM1 and/or FLT3 to more closely interrogate gene expression profiles associated with either mutation, or both. Computational advances and the development of new analysis tools will allow us to further refine our analyses. This will allow us to generate a list of putative candidates for further study *in vitro* or even *in vivo*.

**Proposed Research Activities**

Aim 1: Identify common mutations in NPM1 and FLT3 and their frequencies in adult AMLs, and their clinical significance.

Relevant mutations for NPM1 and FLT3 will be identified using dbSNP14 and dbVar15, both of which are publicly available databases for genetic variation that are maintained by the NIH. As before, we will assign these mutations into three categories: SNPs, indels, and structural variations. For SNPs, we will classify each as being silent, missense (single amino acid substitution), or nonsense (early stop codon). For indels, we will classify each as being either an in-frame or a frameshift insertion or deletion. We will then visualize the frequencies and types of mutations using R.

While dbSNP and dbVar can give us an overarching view of the general allelic variation across a large population, it contains many germline SNPs that are present at high frequencies and have no association with AML. Therefore, we can use ClinVar16 and the Catalog of Somatic Mutations in Cancer (COSMIC)17 to assign pathogenic significance. ClinVar associations are already noted in dbSNP and dbVar. COSMIC, on the other hand, is a manually curated database containing somatic mutations related to human cancers. As such, it offers an orthogonal approach to determining the pathogenicity of identified mutations. We will use mutations identified both by ClinVar and by COSMIC to be cancer-related and define those mutants as “pathogenic”, while mutations identified by neither ClinVar nor COSMIC will be defined as “benign”. We will then compare the frequencies of different types of mutations between pathogenic and benign variants to determine whether a certain type of mutation is more frequently associated with cancer.

Aim 2: Identify potential shared upstream drivers and/or downstream targets of NPM1 and FLT3

Next-generation sequencing and single cell genomics technologies have improved dramatically over the past decade. Increasing throughput and decreasing sequencing costs have allowed researchers to sequence large cohorts of patients and generate databases that others can use to answer their own scientific questions. Unfortunately, large databases generated for this purpose (such as BEAT-AML or TCGA-LAML) have restricted access to sequencing data due to genetic information privacy concerns. Instead, we will use aforementioned scRNA-seq datasets12 that have been uploaded to the Gene Expression Omnibus (GEO) database. Although the cohorts in these datasets are typically smaller in size, which may limit the statistical power of our observations, they will nevertheless serve as a valuable source of data from which we can infer the molecular mechanisms of NPM1 and FLT3 mutations.

**Single-cell preprocessing pipeline.** The single-cell genomics analysis pipeline is no longer in its infancy. While computational tools for downstream analyses are still being developed and benchmarked, researchers have generally arrived at a consensus for best practices in single-cell analysis18. We will preprocess raw data in R using Seurat19 to remove low-quality cells; then, we will use SCTransform20 to normalize and transform the data. We will then use Harmony21 to mitigate batch effects, after which we will perform dimensionality reduction for visualization using UMAP22, and cluster using the Louvain method, which is included in the Seurat workflow.

**Identification of leukemic stem cells (LSCs).** In order to identify relevant differential gene expression profiles between LSCs with NPM1 and/or FLT3 mutations, we need to be able to accurately identify said LSCs. However, accurate identification of LSCs from patient samples is still an area of ongoing research. Some approaches, such as CloneTracer, attempt to identify LSCs based on their distinct mutational profile (inferred from SNVs, CNVs, and mtDNA mutations)23. For the purposes of our project, we will identify clusters with high similarity to LSCs based on a well-established LSC gene score developed by Ng et al.24 We will then run differential gene expression using DESeq225 to identify genes of interest that are apparently modulated by NPM1 and FLT3. Finally, we will confirm their functional significance with a review of the literature and propose plausible *in vitro* and *in vivo* studies for future validation.

**Expected Outcomes**

Firstly, we will expect to generate a comprehensive catalog of all mutations and structural variations, as well as high-level visualizations of the mutational landscape in NPM1 and FLT3. In this process we will likely trim out low-confidence mutations. We also anticipate generating useful comparisons in terms of mutational types between pathogenic and benign variants, with certain types of mutations potentially being more common in pathogenic variants. We hope that future researchers will be able to use our findings to aid in predicting the pathogenicity of future mutations.

Analysis of single-cell data will bring out the transcriptional heterogeneity of AML across patients with a variety of mutations. In particular, we expect to identify leukemic clusters, which we can use to compare gene expression profiles of leukemic cells in patients with a variety of mutations. Of course, our focus in this project is the interplay between NPM1 and FLT3, but we anticipate that others will be able to use the framework we develop to perform similar analyses of their own, perhaps even on the same dataset. Finally, we expect to identify genes of interest that are differentially expressed in leukemic cells with NPM1 versus FLT3 mutations, which we will confirm with the literature.

**References:**

1. Pelcovits A, Niroula R. Acute Myeloid Leukemia: A Review. R I Med J (2013). 2020 Apr 1;103(3):38-40. PMID: 32236160.
2. Meyers, C.A., Albitar, M. and Estey, E. (2005), Cognitive impairment, fatigue, and cytokine levels in patients with acute myelogenous leukemia or myelodysplastic syndrome. Cancer, 104: 788-793. <https://doi.org/10.1002/cncr.21234>
3. Nebgen DR, Rhodes HE, Hartman C, Munsell MF, Lu KH. Abnormal Uterine Bleeding as the Presenting Symptom of Hematologic Cancer. Obstet Gynecol. 2016 Aug;128(2):357-363. doi: 10.1097/AOG.0000000000001529. PMID: 27400009; PMCID: PMC4961605.
4. Klepin HD. Myelodysplastic Syndromes and Acute Myeloid Leukemia in the Elderly. Clin Geriatr Med. 2016 Feb;32(1):155-73. doi: 10.1016/j.cger.2015.08.010. PMID: 26614866; PMCID: PMC4664149.
5. Kishtagari A, Levine RL, Viny AD. Driver mutations in acute myeloid leukemia. Curr Opin Hematol. 2020 Mar;27(2):49-57. doi: 10.1097/MOH.0000000000000567. PMID: 31972687.
6. Esther N. Oliva, Jacob Franek, Dipen Patel, Omer Zaidi, Salem Abi Nehme, Antonio M. Almeida; The Real-World Incidence of Relapse in Acute Myeloid Leukemia (AML): A Systematic Literature Review (SLR). Blood 2018; 132 (Supplement 1): 5188. doi: https://doi.org/10.1182/blood-2018-99-111839
7. Hindley A, Catherwood MA, McMullin MF, Mills KI. Significance of NPM1 Gene Mutations in AML. Int J Mol Sci. 2021 Sep 17;22(18):10040. doi: 10.3390/ijms221810040. PMID: 34576201; PMCID: PMC8467861.
8. Gilliland DG, Griffin JD. The roles of FLT3 in hematopoiesis and leukemia. Blood. 2002 Sep 1;100(5):1532-42. doi: 10.1182/blood-2002-02-0492. PMID: 12176867.
9. Kennedy VE, Smith CC. *FLT3* Mutations in Acute Myeloid Leukemia: Key Concepts and Emerging Controversies. Front Oncol. 2020 Dec 23;10:612880. doi: 10.3389/fonc.2020.612880. PMID: 33425766; PMCID: PMC7787101.
10. Kim YK, Kim HN, Lee SR, Ahn JS, Yang DH, Lee JJ, Lee IK, Shin MG, Kim HJ. Prognostic significance of nucleophosmin mutations and FLT3 internal tandem duplication in adult patients with cytogenetically normal acute myeloid leukemia. Korean J Hematol. 2010 Mar;45(1):36-45. doi: 10.5045/kjh.2010.45.1.36. Epub 2010 Mar 31. PMID: 21120161; PMCID: PMC2983002.
11. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, Dombret H, Ebert BL, Fenaux P, Larson RA, Levine RL, Lo-Coco F, Naoe T, Niederwieser D, Ossenkoppele GJ, Sanz M, Sierra J, Tallman MS, Tien HF, Wei AH, Löwenberg B, Bloomfield CD. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017 Jan 26;129(4):424-447. doi: 10.1182/blood-2016-08-733196. Epub 2016 Nov 28. PMID: 27895058; PMCID: PMC5291965.
12. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, Pinkus GS, Pozdnyakova O, Galinsky I, Stone RM, Graubert TA, Shalek AK, Aster JC, Lane AA, Bernstein BE. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019 Mar 7;176(6):1265-1281.e24. doi: 10.1016/j.cell.2019.01.031. Epub 2019 Feb 28. PMID: 30827681; PMCID: PMC6515904.
13. Morita K, Wang F, Jahn K, Hu T, Tanaka T, Sasaki Y, Kuipers J, Loghavi S, Wang SA, Yan Y, Furudate K, Matthews J, Little L, Gumbs C, Zhang J, Song X, Thompson E, Patel KP, Bueso-Ramos CE, DiNardo CD, Ravandi F, Jabbour E, Andreeff M, Cortes J, Bhalla K, Garcia-Manero G, Kantarjian H, Konopleva M, Nakada D, Navin N, Beerenwinkel N, Futreal PA, Takahashi K. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nat Commun. 2020 Oct 21;11(1):5327. doi: 10.1038/s41467-020-19119-8. Erratum in: Nat Commun. 2020 Nov 19;11(1):5996. Erratum in: Nat Commun. 2021 May 10;12(1):2823. PMID: 33087716; PMCID: PMC7577981.
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11. doi: 10.1093/nar/29.1.308. PMID: 11125122; PMCID: PMC29783.
15. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013 Jan;41(Database issue):D936-41. doi: 10.1093/nar/gks1213. Epub 2012 Nov 27. PMID: 23193291; PMCID: PMC3531204.
16. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062-D1067. doi: 10.1093/nar/gkx1153. PMID: 29165669; PMCID: PMC5753237.
17. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941-D947. doi: 10.1093/nar/gky1015. PMID: 30371878; PMCID: PMC6323903.
18. Heumos, L., Schaar, A.C., Lance, C. et al. Best practices for single-cell analysis across modalities. Nat Rev Genet (2023). https://doi.org/10.1038/s41576-023-00586-w
19. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi: 10.1016/j.cell.2021.04.048. Epub 2021 May 31. PMID: 34062119; PMCID: PMC8238499.
20. Hafemeister C, Satija R (2019). “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.” Genome Biology, **20**, 296. doi:10.1186/s13059-019-1874-1
21. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Dec;16(12):1289-1296. doi: 10.1038/s41592-019-0619-0. Epub 2019 Nov 18. PMID: 31740819; PMCID: PMC6884693.
22. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Dec 3. doi: 10.1038/nbt.4314. Epub ahead of print. PMID: 30531897.
23. Beneyto-Calabuig S, Merbach AK, Kniffka JA, Antes M, Szu-Tu C, Rohde C, Waclawiczek A, Stelmach P, Gräßle S, Pervan P, Janssen M, Landry JJM, Benes V, Jauch A, Brough M, Bauer M, Besenbeck B, Felden J, Bäumer S, Hundemer M, Sauer T, Pabst C, Wickenhauser C, Angenendt L, Schliemann C, Trumpp A, Haas S, Scherer M, Raffel S, Müller-Tidow C, Velten L. Clonally resolved single-cell multi-omics identifies routes of cellular differentiation in acute myeloid leukemia. Cell Stem Cell. 2023 May 4;30(5):706-721.e8. doi: 10.1016/j.stem.2023.04.001. Epub 2023 Apr 24. PMID: 37098346.
24. Ng SW, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD, Schuh AC, Yee KW, Bullinger L, Herold T, Görlich D, Büchner T, Hiddemann W, Berdel WE, Wörmann B, Cheok M, Preudhomme C, Dombret H, Metzeler K, Buske C, Löwenberg B, Valk PJ, Zandstra PW, Minden MD, Dick JE, Wang JC. A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature. 2016 Dec 15;540(7633):433-437. doi: 10.1038/nature20598. Epub 2016 Dec 7. PMID: 27926740.
25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.