Optimization of an SQL Database Towards Selective Targeting of Acute Myeloid Leukemia Cells

A Technical Report submitted to the Department of Biomedical Engineering

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Casey Evans

Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Shannon Barker, PhD, Department of Biomedical Engineering

Optimization of an SQL Database Towards Selective Targeting of Acute Myeloid Leukemia Cells

Abstract

Acute Myeloid Leukemia (AML) occurs when there is an accumulation of poorly differentiated cells in the blood and bone marrow (Meyers et al., 2013). AML is the most prevalent form of leukemia for adults, representing roughly 80% of cases, and follows a rapid clinical progression. In 2015, there were over 20,000 recorded cases of AML and over 10,000 recorded deaths in the US (De Kouchkovsky & Abdul-Hay, 2016). Despite advances in treatment, it is estimated that up to 70% of AML patients 65 years or older will die as a result of the disease within 1 year of their diagnosis (Döhner et al., 2015). There is a critical unmet need to identify new targets in order to improve standard-of-care therapy and increase overall survival. ZielBio's novel drug discovery platform has been used to successfully identify targets that have been overlooked by comparable screening techniques. The platform integrates phage display technology in order to reverse engineer the discovery process, allowing unbiased screening to be coupled with bioinformatics and drive target identification (Brinton et al., 2016). By performing screening within the native context of the cell, the platform is able to identify mislocalized targets. Thus far, it has yet to be applied to hematological cancers. This capstone project aims to identify AML-selective targets for the future development of an anti-cancer targeted therapeutic. In addition, the database is examined and refined in order to increase the efficiency of its search processes.

Keywords: Phage Display, Acute Myeloid Leukemia, Drug Discovery

Introduction

Background on AML

Acute Myeloid Leukemia (AML) is a clonal disorder characterized by a lack of cell differentiation in the myeloid tissue and an accumulation of underdeveloped progenitor cells in bone marrow (**Figure 1)**, that leads to hematopoietic (bone marrow) failure (Pollyea et al., 2011). AML is the most common form of leukemia in adults, with an age-adjusted annual incidence of 4.3 per 100,000. While making up only a small portion of all cancer diagnoses, AML is very deadly, with a 5-year survival rate of roughly 24%. With a median age of diagnosis of 68 years, and incidence increasing with age, AML presents the greatest threat to the elderly (Shallis et al., 2019). Despite advances in treatment, it is estimated that up to 70% of AML patients 65 years or older will die as a result of the disease within 1 year of their diagnosis (Dohner et al., 2015). This elderly population is often less fit for strong chemotherapeutic regimens. These realities illustrate why there is a critical unmet need for new and effective therapies for AML in order to improve the standard-of-care therapy, reduce toxic side effects, and increase overall survival.

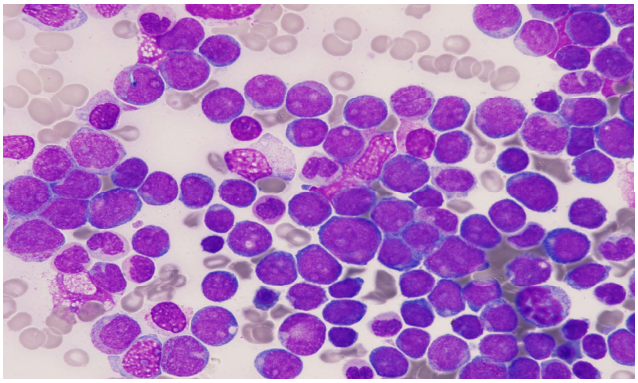

Figure 1. AML cells. Bone marrow aspirate acquired from a 63-year-old female with Acute Myeloid Leukemia. From Bacova et al, 2022.

Thus far, the focus in AML exploratory research has largely surrounded genomics, since this is particularly accessible for hematologic malignancies. The resultant FLT3 inhibitors and other targeted therapies represent forward progress in treatment. Yet, many patients have AML clones that do not harbor a genomic alteration for which these targeted therapies apply, and rapid onset of resistance excludes even more patients from benefitting. The targets identified by genetic sequencing are valuable and while continued sequencing may occasionally produce a new target, it is more likely to continue to converge upon the same targets.

Background on ZielBio

ZielBio is a clinical stage biotechnology company located in Charlottesville, Virginia. ZielBio has a drug discovery platform that integrates phage display technology in order to reverse engineer the discovery process, allowing unbiased screening to be coupled with bioinformatics and drive target identification (Brinton et al., 2016). By performing screening within the native context of the cell, the platform is able to identify mislocalized targets. Thus far, the platform has yet to be applied in hematological cancers. This capstone project aims to harness the platform to identify AML-selective targets for future development of an anti-cancer targeted therapeutic.

ZielBio's platform has been successfully used to perform more than 300 screens. The sequences from peptides discovered through this platform are stored in an internal database. Currently, this database's search processes have yet to be optimized, and there is limited capacity. By restructuring the database and incorporating SQL, the databases search processes can be improved.

Materials and Methods

Cell Culture

MOLM-13 cells were acquired from DSMZ (Leibniz Insitute) and maintained in culture using RPMI media with 10% FBS and L-glutamine (Matsuo, 1997). Cells were maintained at <90% confluency and passaged twice weekly. Cell line identity was authenticated via STR analysis (ATCC). Cells were routinely tested for mycoplasma using the Mycoprobe mycoplasma detection kit (R&D).

Biopanning

Cells were established with greater than four passages prior to screening. The cells were combined with 10^{11} phage from New England Biolabs' PhD7 library and incubated for 4 hours. Subsequently, unbound phage were removed by washing several times before a glycine elution buffer was used to collect phage that had bound to the cells. The cells were then discarded and the eluted phage separated for amplification. Amplification was performed by combining phage with ER2738 E. coli in Luria Broth (LB) Media and allowing for an incubation period of 4.5 hours. Phage were then precipitated using a PEG (polyethylene glycol) solution to remove the bacterial cells. Phage were titered on agar plates in order to quantify the number of phage that were present by counting plaques. From there, the biopanning was repeated, with the amplified phage sample being used in in subsequent rounds of as the input. The process was repeated four times in order

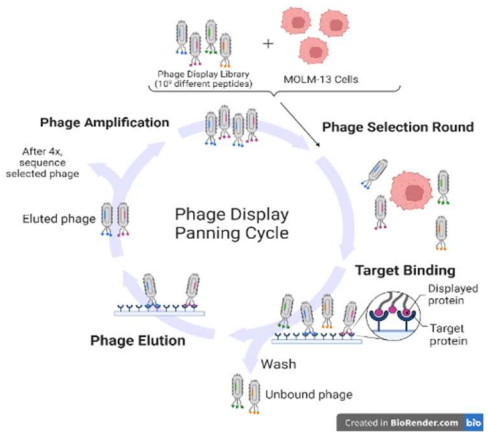

Figure 2. Phage Display Screening process. Phage that are not exhibiting specific binding are gradually filtered out in favor of phage that are exhibiting specific binding.

to ensure that the phage present were exhibiting higher afifnity binding to the MOLM-13 cells (**Figure 2)**.

Titering

LB media was inoculated in a baffle bottom flask with ER2738 E. coli and incubated for 1.5-3 hours. The optical density (OD_{600}) was measured every 30 minutes using a nano spectrometer until reaching roughly 0.5. In parallel, top agarose was boiled in water and then the temperature lowered to 120°C. Serial dilutions of the unamplified and amplified eluate were prepared such that only 50-400 plaques would result per plate (making them "countable"). Once the OD_{600} reached 0.5, ER2738 culture was dispensed into the tubes of serially diluted phage. IPTG-X-gal was then combined with liquid top agarose, which turns the phage plaques blue so they can be distinguished from wild type phages. After a short incubation, bacteria infected with phage was added to the agarose and IPTG-X-gal, vortexed, and then poured onto a pre-warmed agar plate. Plates were left for 10 minutes to solidify before inverting and incubating overnight at 37°C.

Sanger Sequencing of Individual Plaques

During each round and post-amplification, plaques from titer plates were picked for sequencing. Phage DNA was PCR-amplified around the region of the displayed peptide insertion. A DNA gel was then run in order to verify the PCR product was the correct size in each sample, and samples were sent off to be Sanger sequenced (Eurofins). When the results were received, many of the nucleotides were unidentifiable, as seen in **Figure 3**.

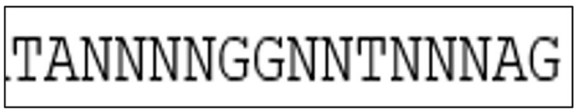

Figure 3. Nucleotide sequences acquired via Sanger sequencing. An "N" indicates that the nucleotide in question was unable to be correctly called by the instrument (likely from multiple peaks) and therefore could not be determined conclusively.

Troubleshooting Sanger sequencing

To determine the cause of the sequencing issue and attempt to produce valid results, the process was restarted. During biopanning, especially in the first couple of rounds, each displayed peptide may only be present as ten copies. Therefore, it is important not to sample from the eluted phage. Thus, two separate samples were run in parallel; one sample (S1) was left untouched to avoid loss of phage and the other (S2) was used to check the progression of the biopanning. The cycle was repeated from the beginning S1 also went through additional Sanger sequencing after the completion of each round in order to identified any mistakes in the process.

There was a possibility that the LB Media that was used during the initial process had become contaminated, and that this was contributing to the invalid sequencing results. LB Media is prone to bacterial contamination due to its makeup of yeast, tryptone, and sodium chloride. To compensate for this, additional steps were taken to avoid contamination of the LB media, such as avoiding direct contact of the serological pipette to the stock solution through pouring it into a 50 ml tube. Furthermore, the LB media was routinely inspected visually, and was disposed of and replaced if any evidence of contamination could be seen.

Some successful Sanger sequences were obtained during the screening process, suggesting that the biopanning was proceeding normally. Unfortunately, some of the poor Sanger sequencing results persisted, and these are believed to be related to the way samples are being shipped (still under investigation). Given that the Sanger sequencing issues did not appear to indicate an issue with the screen, the final round of samples were prepared for deep sequencing.

Illumina Sequencing

The amplified sample from the 4th round of biopanning was PEG-precipitated and centrifuged to pellet the phage. DNA was extracted from the phage pellet by suspending it in Iodide buffer and adding ethanol before an incubation of 10 minutes at room temperature. The solution was then microcentrifuged for 10 minutes at 4℃, and washed with 70% ethanol. The pellet was left to dry for roughly 2 hours and then suspended in sterile water. The absorbance of the sample was then measured at 260 nm in order to determine the concentration of DNA within the sample. 1ng of DNA was then added to a solution containing 25 uL of MyTaq HS Red Mix, 2 uL of forward primer, 2 uL of reverse primer, and 20 uL of H_2O. The sample then underwent PCR, with 1 minute at 95 degrees C, 20 repetitions of 95, 60, and 72 degrees C in 30 second intervals, followed by a 4°C hold overnight. PCR product was cleaned up with QIAquick PCR purification kit (Qiagen) and quantitated via Qubit. The sample was submitted to the UVA Biomolecular Research Core Facility core facility. Quality was confirmed via tapestation and then run on an Illumina MiSeq Sequencer with a MiSeq 150 V3 kit.

Processing of Deep Sequencing Results

The fastq files from Illumina contain millions of reads. The file was processed via skewer (Jiang et al., 2014) to remove pieces of the reads corresponding to native phage DNA that flank the displayed region. Post-processing, the 21 remaining nucleotides code for each 7aa displayed peptide, and frequencies obtained for each sequence indicate the degree to which that peptide bound the cells. As previously published, sequences are compared to the database of previous screens to determine selectivity (Brinton, et al., 2016).

Results

Biopanning isolated phage that bound MOLM-13 cells

The screening protocol was successfully modified to identify phage interacting in solution to a suspension cell line (MOLM-13). Phage were successful collected and input into subsequent rounds (**Figure 4**).

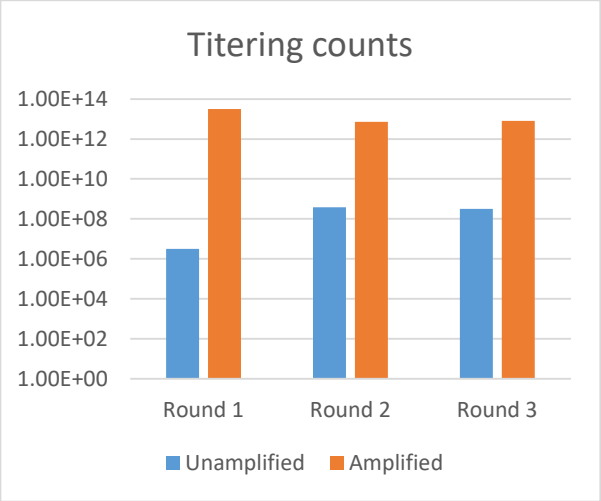

Figure 4. Titering counts. The number of phage output after each round or amplification.

While many issues were encountered with the Sanger sequencing, some plaques were successfully sequenced (**Figure 5**).

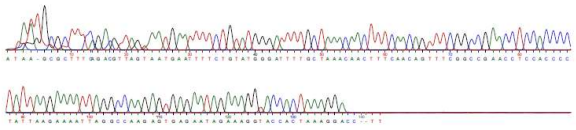

Figure 5. Successful Sanger sequencing trace. The flanking region ACCTCCACC is captured along with the displayed peptide sequence.

After the fourth and final round of screening, DNA was isolated from the phage pool and quantified via a nanospectometer, finding a concentration of 370 ng/ml. The genomic region containing the displayed peptide was PCR amplified and barcoded. The resultant PCR product was check on a Qubit and contained 54,600 ng/mL. Tapestation indicated there was sufficient product at the correct size to proceed with deep sequencing;

however, some higher molecular contamination was identified (**Figure 6**).

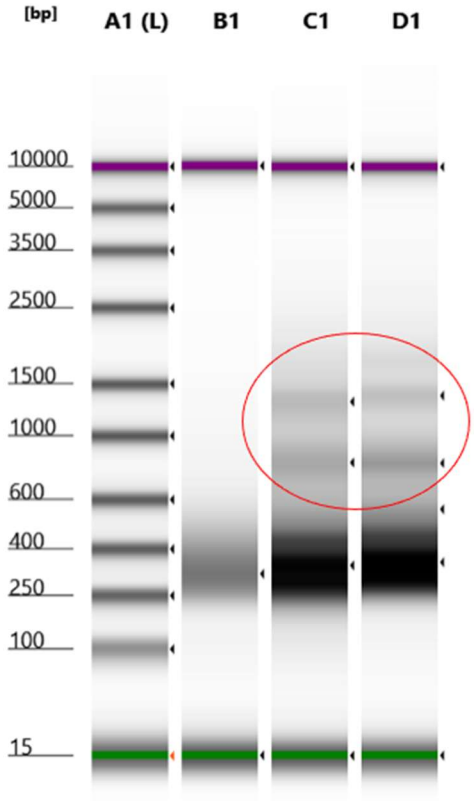

Figure 6. Tapestation. Strong band at expected size as well as some higher molecular weight contamination.

Deep sequencing reveals multiple high binding peptide sequences

Metric	Value/Result
Max error rate	0.1
Max indel rate	0.03
Head Trim	
Reads processed	2,536,030
Short reads filtered out	2
Reads lacking head flanking region	82,621 (3.26%)
Tail Trim	
Reads processed	2,536,028
Short reads filtered out	204 (0.01%)
Reads lacking tail flanking region	356,896 (11.2%)
Total correct reads	**2,178,928**

Table1. Processing Metrics.

Fastq files obtained from the deep sequencing contained 2.5 million reads. **Table 1** shows the metrics from the processing of the sequences. Future work will validate the top hits and determine what their binding partners are on the cell surface.

Database Framework

Due to the many issues that occurred during the sequencing process, the troubleshooting that was required, and the time constraints of this project, little progress was made in terms of optimizing the ZielBio Database. However, there a few steps that could be taken to simplify this process in the future. The main issue that makes redesigning the database difficult is the fastq files, which are the file type that store the DNA sequences. Converting the fastq files to ".csv" or Excel files would allow them to be imported into an SQL IDE (integrated development environment) and make the process of making an effective database much simpler. Early in the stages of the project, there was a brief attempt to convert the fastq files into Excel. However, this attempt was relatively unsuccessful and the Excel files were undecipherable. By finding an effective method of converting the fastq files, the DNA sequences can be imported into SQL and then organized into an effective database through the use of DDL techniques (database definition language).

Discussion

Sequencing Issues

Throughout this project, there were many issues regarding Sanger sequencing of the DNA that is extracted from the phage. When samples were sent off to be Sanger sequenced through Eurofins genomics, they often came back with a significant number of errors and unreadable nucleotides.

There are many possible explanations for the sequencing issues that were experienced over the course of this project. One possibility is that despite the steps that were taken to avoid it, contamination still occurred, leading to a combination of selectively-binding phage and wild-type phage that caused sequencing to fail. A second possibility is that the DNA was being improperly acquired from the plaques generated from phage

(**Figure 7A**). The protocol followed instructed for the plaques to be "stabbed" with a 1 ml pipette tip in order to minimize the amount of agar that inevitably surrounds the plaque. The plaques were then placed into deionized (DI) water and pipetted repeatedly to break apart the agar and allow the water to have direct contact with the plaque. It is possible that the plaques were repeatedly not broken up to a sufficient degree, leading to insufficient DNA and faulty sequencing. However, due to the repeated DNA gels that were run, which verified that the plaque water mixture contained DNA, this explanation appears unlikely (**Figure 7B**). Furthermore, ZielBio's experienced research associates observed the process of breaking up the plaques on several occasions, and noted nothing substantially different from previous successful screens.

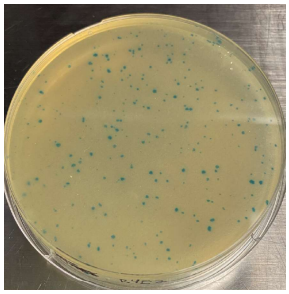

Figure 7A. Plaques on Agar plate. Plaques were generated from phage through titering.

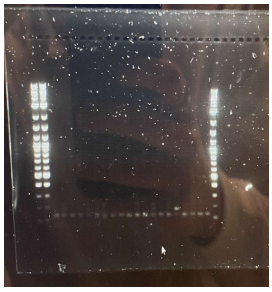

Figure 7B. A DNA gel image. A consistent low band is s confirming that DNA is present in the sample.

The current hypothesis is that there may be a process breakdown in the way that the samples are shipped to the contract research organization that runs the Sanger sequencing. This is currently under investigation and a pilot run doing full plasmid sequencing at a different company has been initiated to circumvent this issue.

After the conclusion of the screening process for the second time, Illumina sequencing was performed. However, when the results came in it was determined that there was strong chemical interference. This was likely due to a lack of purity in the sample. After performing an additional purification step, the DNA was successfully sequenced in a second Illumina sequencing run.

While the obstacles that were encountered led to many delays, overall the project resulted in successful identification of sequencing motifs that bind to AML cells. Further validation and target identification efforts present an exciting avenue for discovery of new therapies to help AML patients.

End Matter

Author Contributions and Notes

The author delares no conflict of interest.

Acknowledgents

I would like to thank Lindsey Brinton, Kim Kelly, Samanha Perez, Brian Murphy, Sarah Hall, and Abby Colvin of ZielBio their help on this project. Thank you to Yongde Bao for kindly running the Illumina sequencing at the UVA Biomolecular Research Core Facility. Additionally, I would like to thank Noah Perry, Shannon Barker, Timothy Allen, Natasha Claxton, and Zehra Demir of the BME Capstone teaching team for all of their support during the course of this project.

References

Bacova, B., Sobotka, J., Kacirkova, P., Rivnacova, V., Karlova/Zubata, I., & Novak, J. (2022). Acute myeloid leukemia with variant t(8;10;21). *Leukemia research reports*, *18*, 100350. https://doi.org/10.1016/j.lrr.2022.100350

Brinton, L. T., Bauknight, D. K., Dasa, S. S. K., & Kelly, K. A. (2016). PHASTpep: Analysis Software for Discovery of Cell-Selective Peptides via Phage Display and Next-Generation Sequencing. *PLOS ONE*, *11*(5), 1–22. https://doi.org/10.1371/journal.pone.0155244

De Kouchkovsky, I., & Abdul-Hay, M. (2016). 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood cancer journal*, *6*(7), e441. https://doi.org/10.1038/bcj.2016.50

Döhner, H., Weisdorf, D. J., & Bloomfield, C. D. (2015). Acute Myeloid Leukemia. *New England Journal of Medicine*, *373*(12), 1136–1152. https://doi.org/10.1056/NEJMra1406184

Jiang, H., Lei, R., Ding, S.-W., & Zhu, S. (2014). Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, *15*(1), 182. https://doi.org/10.1186/1471-2105-15-182

Matsuo, Y., MacLeod, R., Uphoff, C., Drexler, H., Nishizaki, C., Katayama, Y., Kimura, G., Fujii, N., Omoto, E., Harada, M., & Orita, K. (1997). Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). *Leukemia*, *11*(9), 1469–1477. https://doi.org/10.1038/sj.leu.2400768

Meyers, J., Yu, Y., Kaye, J. A., & Davis, K. L. (2013). Medicare fee-for-service enrollees with primary acute myeloid leukemia: an analysis of treatment patterns, survival, and healthcare resource utilization and costs. *Applied health economics and health policy*, *11*(3), 275–286. https://doi.org/10.1007/s40258-013-0032-2

Pollyea, D. A., Kohrt, H. E., & Medeiros, B. C. (2011). Acute myeloid leukaemia in the elderly: a review. *British journal of haematology*, *152*(5), 524–542. https://doi.org/10.1111/j.1365-2141.2010.08470.x

Shallis, R. M., Wang, R., Davidoff, A., Ma, X., & Zeidan, A. M. (2019). Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. *Blood reviews*, *36*, 70–87. https://doi.org/10.1016/j.blre.2019.04.