Using DNA methylation for acute myeloid leukemia subtyping

PI: Prof. Runsheng Li Co-I: Prof. Liang Zhang

Background of research

Acute myeloid leukemia (AML) is one of the most lethal cancers worldwide. In Hong Kong, it occurs in about 300 patients each year. Overall treatment outcome is unsatisfactory and only 30-40% of patients can achieve long-term remission. There is still an unmet need to develop a personalized and effective treatment for AML based on their unique genetic and epigenetic characteristics.

The recommendations for diagnosis and management from European LeukemiaNet (ELN) use both the cytogenetic and genetic information to classify the AML into different subtypes [1]. By the subtyping with cytogenetic information, around 49% of AML patients are **C**ytogenetically **N**ormal (CN-AML). This group represents patients with a very broad clinical outcomes, and the overall risk level for the CN-AML patients are "intermediate". The genetic variations are currently the only markers used to further subtyping the CN-AML patients. Some mutations for essential genes, such as FLT3-ITD, CEBPA, NPM1 and IDH2, can indicate different prognosis in CN-AML. However, there are still around 36% CN-AML patients who do not hold a significant mutation [2]. Also, the effectiveness of genetic mutation could be low in prognostic prediction in some CN-AML patients (Fig.1A). The overall allele burden does not significantly correlate with leukaemia free time as the epiallele burden does [3]. A novel layer of markers including the DNA modifications, can provide additional information to classify CN-AML patients [3].

Although the aberrant methylation in genomic DNA is a common feature for leukemia cells, the whole epigenetic landscape for CN-AML is still unknown. The comparison between the whole genome for 5-Methylcytosine (5mC) pattern of the CN-AML bone marrow cells and DNMT3A mutated cell lines shows exclusive divergence. The change of CN-AML cells occurs in non-CGI regions and DNMT3A mutations confer a pattern of global hypomethylation that specifically targets HOX genes [4]. This finding indicates that the representative methylation change in the leukemia cells could be independent to the known methylation drivers like TETs and DNMTs. The recent published large integrated databases can also provide some insight for the CpG methylation for 5mC in CN-AML, ie. the TCGA-AML and Target-AML project. The screening of

5mC in primary AML samples showed that over 70% of all hypermethylation sites from different mutant subtypes (IDH1, 2 and TET2 mutates) are shared. These shared sites are likely to be the 'generic', AML-associated changes. With this concept, some 5mC modifications from Illumina 450K methylation array, which is based on bisulfite sequencing (BS), has been chosen as markers to differ ALL from AML samples using TCGA-AML dataset [5].

The DNA modifications could be a good marker layer for CN-AML subtyping. In our analysis using the CN-AML patient information from both TCGA-AML (n=91) and TAGET-AML (n=111) studies, over two hundred CpG islands in the promoter region show strong correlation (FDR <0.05) with overall survival time (OS) in CN-AML samples (Fig. 1A). The positive correlation between the DNA methylation level at the promoter of LZTS2 and NR6A1 genes and the OS in CN-AML patients from SWOG trials (n=72) [6] can also be validated in this analysis. Moreover, the high epiallele burden is also significantly linked to poor OS (Fig. 1C).

The BS method has some disadvantages in detecting DNA modifications in clinical cancer samples. Firstly, for the high throughput BS-based method, the BS-induced DNA degradation leads to depletion of genomic regions enriched for unmethylated cytosines. As a result, the biased estimation of the m5C within hypomethylated regions is expected. Secondly, the WGBS can only detect the 5mC. The other types of modifications, including the N6-methyladenine (6mA), 5-formylcytosine (5fC) and 5-Hydroxymethylcytosine (5hmC), would be missed by using only this technology. Thirdly, compared with cytogenetic and genetic information, the feasibility to obtain the epigenetic information from DNA has hindered the use of these epigenetic changes in clinical diagnostics. The traditional BS method to detect the specific change in a unique site would require sophisticated laboratory bench work , which could result in long turnover time and reduce the feasibility to use this technology clinically.

Nanopore sequencing, provided by Oxford Nanopore Technology (ONT), is currently the most effective way to get DNA modification information. The native DNA is used for the sequencing, so the amplification bias was no longer an issue for this platform. After loading the library, the DNA sequences are basecalled by the recording of current signal changes when the native DNA passes through the Nanopore. The current signal for the “ATCG” could be different from the modified ones, and these modifications can be precisely retained by bioinformatic analysis. Theoretically, this technology can detect any type of methylation. By now, the 5mC modifications can be predicted in very high accuracy (>99%) and reach 95% consistency with BS covered regions using pre-trained models with the state-to-art deep learning methods [7, 8]. For the other type of modifications aside from 5mC, like 5fC, 5hmC, the accuracy is still in doubt. We have tried to train our own models to call these specific modification types, and developed a framework pipeline for Long Read Modifications Finding (named as “nanoCEM”). This pipeline can be sued to prepare the datesets used to call different types of DNA or RNA modification by comparing with a modification free sample, which can be whole genome DNA amplification (WGA) data (Fig. 2A). Moreover, because of the real-time basecalling and modification calling, the results can be fetched right after the sequencing begins. The turnover time for this technology could be as short as the library preparation time, which takes only two hours.

Aside from the whole genome screening of the methylation change, there is a specific target for tumorigenesis, the ribosomal DNA (rDNA). The change of rDNA is favoured to provide infinity proliferation ability for cancer cells. And the copy number variation (CNV) for rDNAs has been found in most cancers. In the human genome, the 18S-5.8S-28S clusters with 43 Kb repetitive unit are dispersed near the centromere or telomeres on five separate chromosomes (i.e., chromosomes 13, 14, 15, 21 and 22). Although the ~400 copies of rDNA copies have identical sequence, their methylation status in different cells could be very divergent, varies from 0 to 100%. Thus, the DNA methylation status in rDNA can server as a good indicator for overall epiallelic burden in leukaemia cells [9]. The traditional high throughput BS and immunoprecipitation methods for 5mC modification falls short in highly repetitive regions, only the long read based Nanopore reads can provide a reliable solution [7]. With this specific marker, the overall epiallelic burden can be quickly identified using targeted sequencing instead using the average of multiple loci.

Work done by us.

The PI has been working on the methodology development for Nanopore sequencing, especially for modification findings. As one of the pioneer labs to use Nanopore technology in Hong Kong, the PA team has developed several pipelines used for Nanopore DNA and RNA sequencing [10, 11], and used this technique to solve the "hard to solve" questions in genome research, including the structural variation finding on mitochondrial DNA and assembly of gapped rDNA cluster in non-model species. We have developed the modification finding pipeline for DNA and RNA using Nanopore sequencing ("nanoCEM", https://github.com/lrslab/nanoCEM). This pipeline consists of a framework used to visualize and compare different samples, which could be used to prepare the input for further machine learning process.

We have sequenced eight CN-AML patients' bone marrow samples with DNMT3a mutation, and eight other CN-AML patients' sample from other patients using Nanopore R9.4.1 platform. These sequencing data together with the DNA methylation profiling can be used as a start point to test our selected sites.

The classification for AML is fast evolving since high-throughput sequencing was introduced to provide the layer of genetic changes. The DNA modifications were not yet included in the ENL classification of hematopoietic neoplasms. With the mature of DNA modification detection using third-generation sequencing, we believed this is the right time to add another layer of information, the modifications, to better reflect the prognosis of CN-AML patients.

Aims and Hypotheses to be Tested:

The proposal arises from an unmet clinical need to subtype CN-AML patients. Based on our preliminary analysis, we aim to select the genomic loci with unique DNA modification pattern related to prognosis. These sites can be further used to build a screen panel with Nanopore sequencing platform for CN-AML subtyping in further studies.

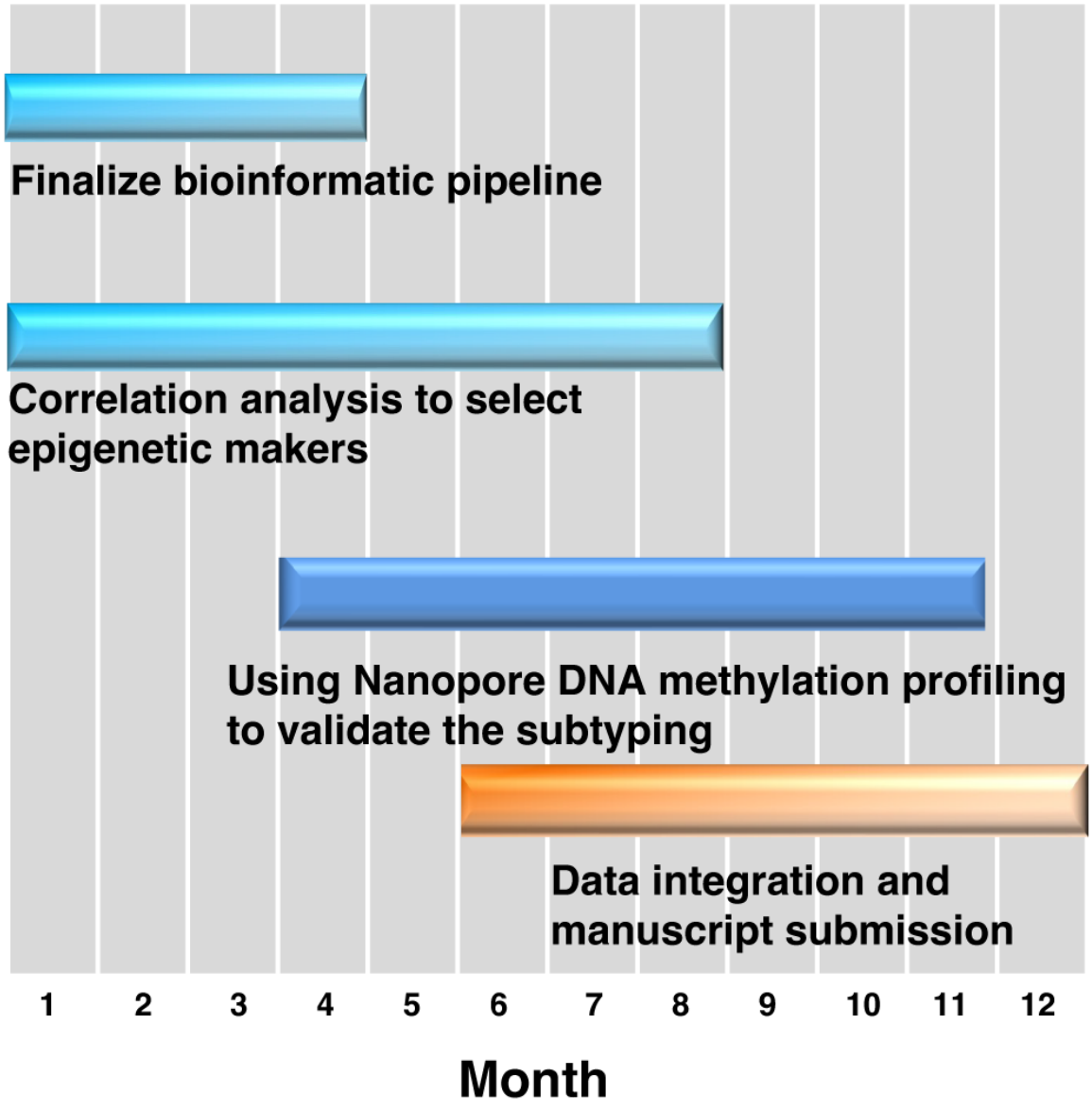
Objective:

Identify the DNA modifications related to the prognosis of CN-AML patients by mining public databases and validate these sites in obtained Nanopore sequencing data.

References:

1. Döhner, H., et al., *Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel.* Blood, 2017. **129**(4): p. 424-447.
2. Tsui, S.P., et al., *A Mutation Pentad Defined Outcome of De Novo and Cytogenetically Normal Acute Myeloid Leukaemia in Young Adults*. 2019, American Society of Hematology Washington, DC.
3. Li, S., et al., *Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia.* Nature medicine, 2016. **22**(7): p. 792-799.
4. Qu, Y., et al., *Differential methylation in CN-AML preferentially targets non-CGI regions and is dictated by DNMT3A mutational status and associated with predominant hypomethylation of HOX genes.* Epigenetics, 2014. **9**(8): p. 1108-19.
5. Jiang, H., et al., *DNA methylation markers in the diagnosis and prognosis of common leukemias.* Signal Transduct Target Ther, 2020. **5**(1): p. 3.
6. Qu, X., et al., *Prognostic methylation markers for overall survival in cytogenetically normal patients with acute myeloid leukemia treated on SWOG trials.* Cancer, 2017. **123**(13): p. 2472-2481.

7. Giesselmann, P., et al., *Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing.* Nature Biotechnology, 2019. **37**(12): p. 1478-1481.
8. Wick, R.R., L.M. Judd, and K.E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing.* Genome biology, 2019. **20**(1): p. 129.
9. Raval, A., et al., *Reduced rRNA expression and increased rDNA promoter methylation in CD34+ cells of patients with myelodysplastic syndromes.* Blood, 2012. **120**(24): p. 4812-4818.
10. Zhang, S., et al., *New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs.* Nucleic Acids Res, 2020. **48**(14): p. 7700-7711.
11. Li, R., et al., *Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development.* Genome Res, 2020. **30**(2): p. 287-298.
12. Au, C.H., et al., *Rapid detection of chromosomal translocation and precise breakpoint characterization in acute myeloid leukemia by nanopore long-read sequencing.* Cancer Genet, 2019. **239**: p. 22-25.



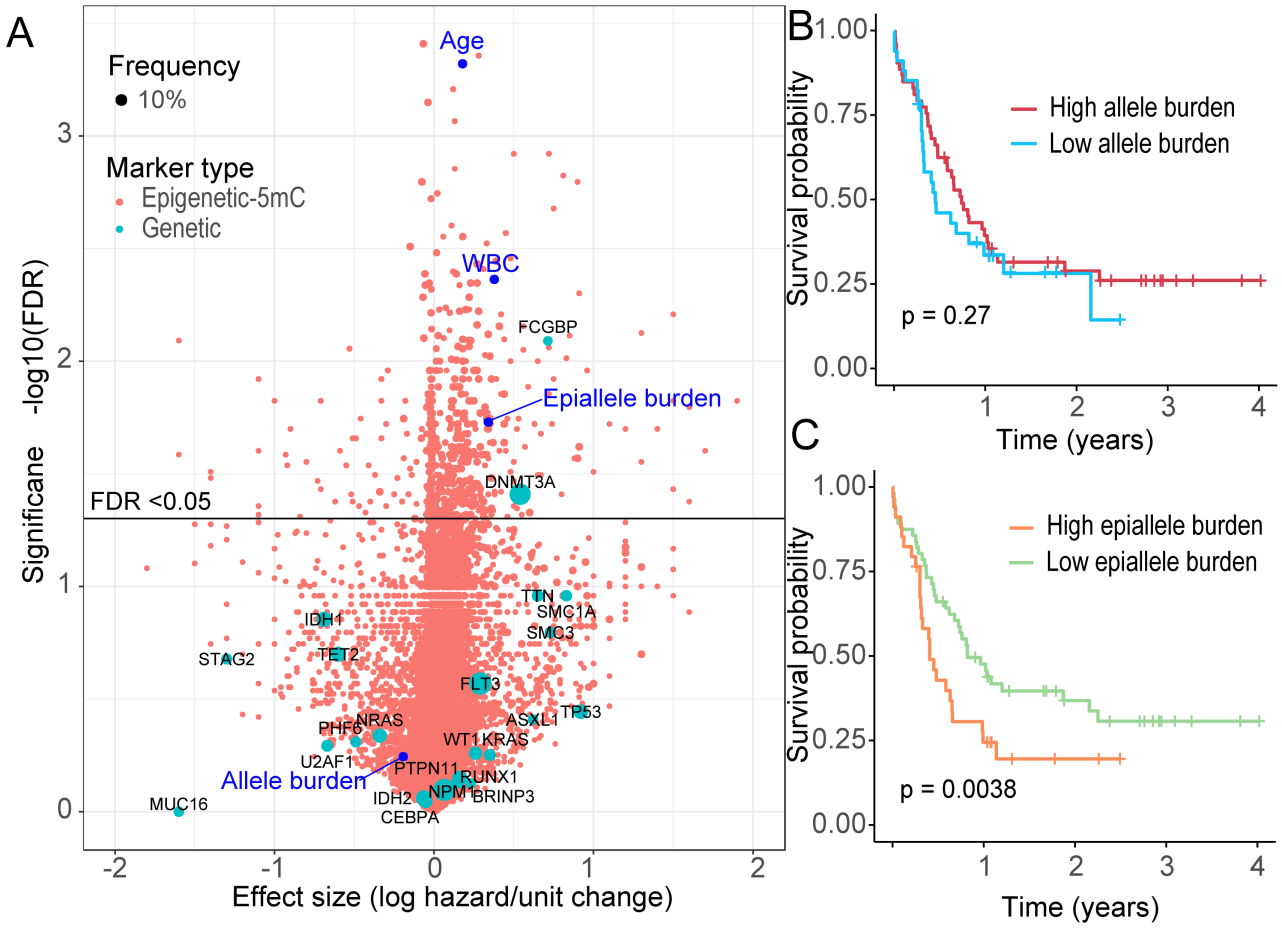


Fig 1. The epigenetic changes are better prognostic markers than genetic mutation. (A) volcano plot shows the relative contribution to prognosis (expressed as the logarithmic hazard on the x-axis; positive values indicate a worsening effect) vs false discovery rate (FDR) values (expressed on an inverted logarithmic scale on the y-axis) for each variables included in their random-effects model. The size of circles for the genetic mutations represent the frequency of this mutation, only mutations with >5% frequency are included. The age and white blood cell count (WBC) are shown for reference. (B) Overall survival analysis for patients with high or low allele burden assessed from somatic mutation burden inferred from the VAF values from whole exon sequencing. (C) Overall survival analysis for patients with high or low epiallele burden assessed from promoter 5mC methylation inferred from the BS sequencing data. The samples with overall allele/epiallele burden higher than average are marked as "high". All three analysis was carried out with n = 202, combined TCGA-AML and TARGET-AML public databases.