Improving low allelic variant detection using error corrected sequencing coupled with a binomial error rate

Karl R. Franke[1], Ben Huang[2,3], Jeff Myers[1], E. Anders Kolb[1], and Erin L. Crowgey[1]

[1] Research Department, Nemours Children's Hospital Delaware, Wilmington, DE

[2] Department of Pediatrics, University of California San Francisco, San Francisco, CA.

[3] Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA.

Introduction

Next generation sequencing (NGS) techniques have been extensively utilized in the medical field to help understand the association of genomic mutations and disease onset and progression. A tremendous amount of effort has been directed towards improving bioinformatic techniques for detecting genomic alterations, with a major focus on improving the types of variants detected, such as structural variants and insertions and deletions. Our team focuses on improving the detection of low allelic events via a technique called error corrected sequencing, or ECS. This technique involves the use of unique molecular indexes, UMI, that enables the bar-coding of individual molecules being prepared for NGS. The UMIs, coupled with downstream bioinformatics, allows for the ability to correct for library preparation "errors" that are introduced during the low level amplification steps, required in most library preparation techniques. We further link these results with a binomial reduction method to further remove sequencing errors.

Dataset and NGS

The Children's Oncology Group (COG) AAML1031 Phase III trial enrolled AML patients from 2011 to 2016. From the 1,097 patients eligible for analysis, 1,782 samples were provided from: diagnosis (n=824), remission (n= 820) and relapse (n= 138). DNA was extracted from bone marrow samples and prepared for NGS using Anchored Multi-plex PCR (AMP) technology within the Archer chemistry. The library preparation included unique molecular indexes (UMIs) for ECS techniques.

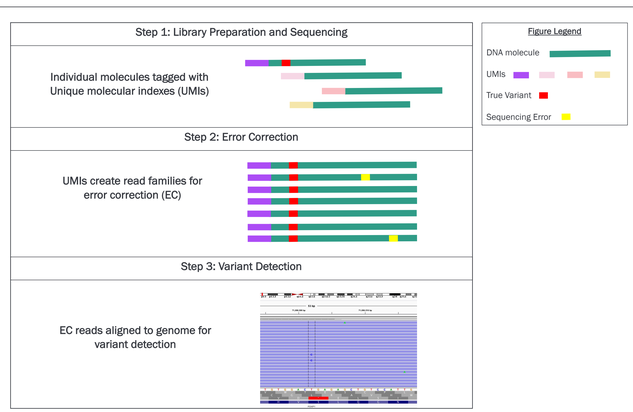

Figure 1. Overview of error corrected sequencing techniques
By tagging each DNA molecule with a unique molecular index (**Step 1**), read families can be generated, by aligning reads with the same UMI, and used to determine sequencing errors versus low allelic variants (**Step2**). Error-corrected sequences are then aligned to a reference genome and used to calculate variants (**Step 3**)

For each of the DNA libraries sequenced, an average molecular bin depth was calculated as well as the number of unique molecules per gene specific primer..

Binomial reduction and library statistics

Using the UMIs, it was determined that the libraries consisted of high complexity DNA, as the average deduplication ratio was 2 and the average unique molecule per GSP was 800.

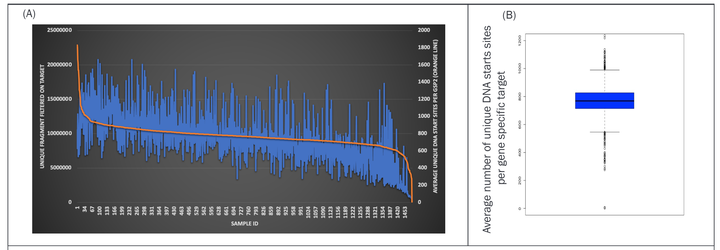

Figure 2. Bin depth versus unique molecules per gene specific primer
Panel A: Unique fragments (blue) versus the average unique DNA start sites per GSP2 (orange)
Panel B: Distribution of unique molecules per GSP

After variant calling, a binomial reduction calculation was implemented to remove low allelic calls that were due to sequencing errors.

(A) Formula for binomial reduction:

x = total # of reads supporting the alt in this sample
a = total # of reads at this location in this sample
b = background

Background =
(sum of all reads supporting the alt in all samples)/
(sum of all reads at this location in all samples)

pbinom(x-1, size=a, prob=b, lower.tail=FALSE)

(B)

Bonferroni correct alpha for filtering

0.05 / (# positions in panel *3)

DNA Panel 1.42316*10^-7

Figure 3. Overview of binomial reduction calculation
Panel A: formula used for the binomial reduction to improve detection of low allelic variant calls.
Panel B: Bonferroni correct alpha used for filtering

Single Nucleotide Variant (SNV) analysis

To benchmark these results, SNVs from 55 genes were compared against the National Cancer Institute's established TARGET AML dataset. Nine genes were identified to have a difference in % prevalence between the pediatric AML cohorts.

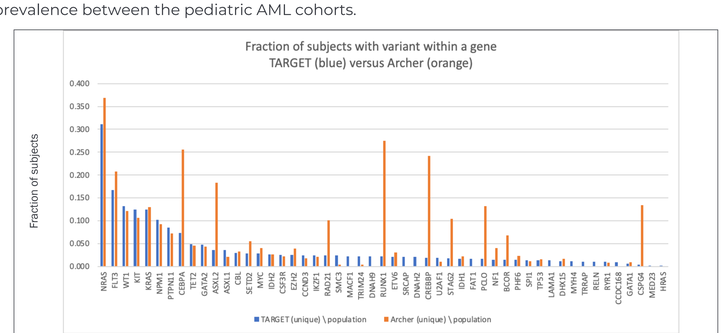

Figure 4. Summary of comparison between TARGET and Archer
Fraction of subjects with variant within a gene for TARGET (blue) versus Archer (orange)

Variant allele frequency (VAF) distributions

The distributions between SNVs detected via ECS technique (Archer; orange) versus both Archer and TARGET (blue) were compared. There was a statistical difference in the VAFs between variants that were detected with both methods versus just Archer. It is important to note that these are not necessarily biologically relevant and much work is needed to understand their clinical significance.

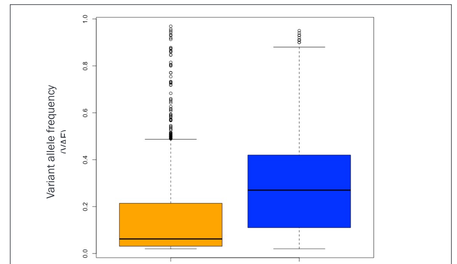

Figure 5. SNV VAF comparison: Archer vs TARGET
VAFs calculated with Archer UMI data (orange) were statistically significantly lower than those calculated via TARGET data (blue) for variants found in both datasets.

Summary

- UMIs offer the ability to correct for library sequencing mistakes, and also provide value in calculating library complexity
- UMIs enable the precise calculation of unique molecules per target region
- Binomial reduction calculation can reduce background sequencing noise that complicates the ability to determine signal from noise for low allelic variants
- Benchmarking our findings against a similar cohort sequenced without UMI indicated strong consistency across 46 genes out of 55, with 9 being discordant in regards to % prevalence
- Genes that were discordant had variants with lower VAFs compared to variants detected by both Archer and TARGET

References

1. Crowgey EL, Mahajan N, Wong WH, Gopalakrishnapillai A, Barwe SP, Kolb EA, et al. Error-corrected sequencing strategies enable comprehensive detection of leukemic mutations relevant for diagnosis and minimal residual disease monitoring. BMC Medical Genomics. 2020 Mar 4;13(1):32.
2. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, et al. Anchored multiplex PCR for targeted next-generation sequencing. Nat Med. 2014 Dec;20(12):1479–84.
3. Lam SW, Cleton-Jansen A-M, Cleven AHG, Ruano D, van Wezel T, Szuhai K, et al. Molecular Analysis of Gene Fusions in Bone and Soft Tissue Tumors by Anchored Multiplex PCR-Based Targeted Next-Generation Sequencing. J Mol Diagn. 2018 Sep;20(5):653–63.
4. Bolouri H, Farrar JE, Triche T, Ries RE, Lim EL, Alonzo TA, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med. 2018 Jan;24(1):103–12.

This work was supported by an NIH NCI R01CA211711 (PI Druley) and the Lisa Dean Moseley Foundation (PI Crowgey).