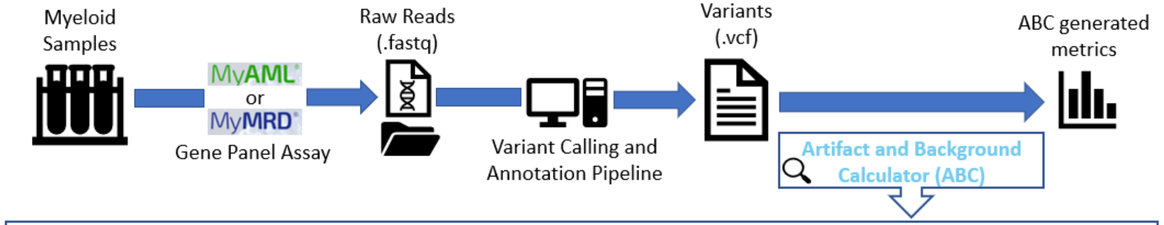
Poster # I008 Artifact and Background Calculator Identifies Patterns of False Positive Variant Calls in Clinical NGS Panels

Joshua D. Wemmer, Jillian Burke and Julian D'Angelo | Invivoscribe Inc.

INTRODUCTION

- Next-generation sequencing (NGS) based targeted gene panels require high recall and precision to reliably detect clinically significant variants for acute myeloid leukemia (AML); however technical artifacts can arise and lead to artifactual variant calls that are not easily distinguished from real variants. For NGS assays that report at low variant read frequencies (VRFs), filtering false positives can be a bottleneck in clinical lab workflows.
- To automate the filtering process, we developed/ Artifact and Background Calculator (ABC), a pipeline that identifies artifactual variant calls and applies a genomic position-specific background error rate (BER) normalization.
- A meta-analysis was conducted to elucidate the most common patterns distinguishing artifacts from real variants, where they occur in the genome, and their likely causes.



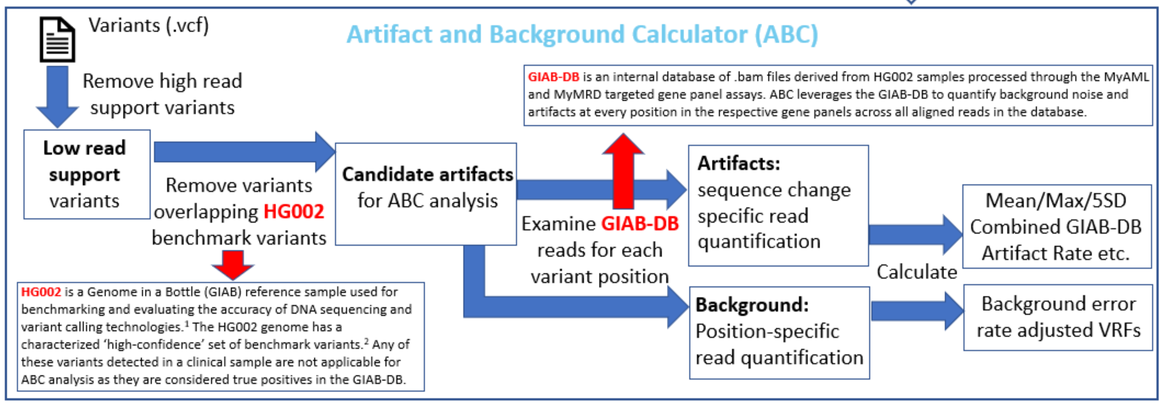


Fig 1. Clinical Workflow with ABC. Clinical samples processed through MyMRD or MyAML Gene Panel Assays undergo variant calling and annotation followed by ABC analysis. ABC takes in Variant Call Format (.vcf) files as input. High support variants are removed from analysis and GIAB benchmark variants are also removed to avoid processing true positive variants in the reference samples.[2] The remaining variants are examined across the GIAB-DB to quantify artifact likelihood and background error rates. ABC generated metrics along with other variant-related tabulated data are then uploaded to an IVS internal database for variant curation and reporting.

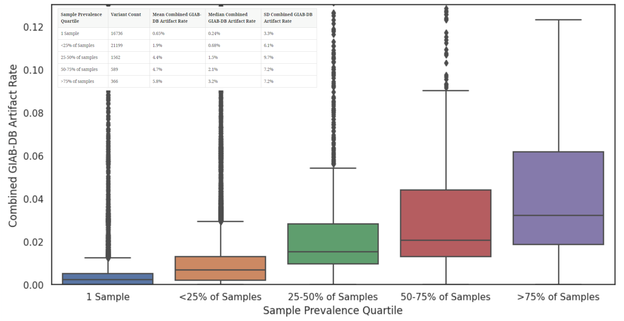
METHODS

- 122 AML clinical patient samples were processed through the MyAML® Gene Panel Assay, which can detect somatic mutations that are present at as low as 1% allelic frequency. Sample libraries were sequenced on an Illumina NovaSeq™ 6000. Variants were detected using an in-house variant calling pipeline, MyAML® v2.0 Software.
- Variant Call Format (.vcf) files were then passed to the ABC software, where variants identified in the clinical samples with VRF >25% and indels >2bp in size were filtered out from analysis. Additionally, any variant that overlapped an HG002 benchmark variant was filtered out as they are considered true positives in the reference samples[2].
- Leveraging the GIAB-DB, the ABC software generated artifact and background error rate metrics for each candidate variant.
  - Note: The GIAB-DB contains assay and sequencer specific sample sets. The MyAML-Novaseq specific GIAB-DB consisted of 17 HG002 replicate samples sequenced with the MyAML® Gene Panel Assay and Illumina NovaSeq™ 6000. The final database was prepared by pre-processing and aligning reads to the hg19 human reference genome to obtain binary alignment map (.bam) files for each GIAB replicate sample.
- To study artifact associated patterns, several variables were examined (described below in Glossary):

Glossary:

- **Sample Prevalence**: The number of clinical samples the candidate artifact was observed in (out of 122 total patient samples).
- **Sample Prevalence Quartiles**: Candidate artifacts that recurred in multiple samples were binned into the following groups: <25% of samples, 25-50% of samples, 50-75% of samples and >75% of samples.
- **Gene Symbols:** HUGO Gene Nomenclature Committee gene symbols that were most represented from the candidate artifacts were analyzed.
- **Mean GIAB-DB Read Depth**: Mean total depth across all BAM files in the GIAB-DB at a given genomic position.
- **Combined GIAB-DB Artifact Rate**: Primary metric used to analyze artifact patterns, calculated as the cumulative artifact reads in GIAB-DB divided by the cumulative read depth at the given position.

RESULTS



Sample Prevalence Quartile	Variant Count	Mean Combined GIAB-DB Artifact Rate	Median Combined GIAB-DB Artifact Rate	SD Combined GIAB-DB Artifact Rate
1 Sample	19730	0.65%	0.24%	3.3%
<25% of Samples	21199	1.9%	0.68%	6.1%
25-50% of Samples	1562	6.4%	1.5%	9.7%
50-75% of Samples	389	6.7%	2.0%	7.2%
>75% of Samples	366	5.8%	3.2%	7.2%

Fig 2. Boxplot representations of the Combined GIAB-DB Artifact Rate Distributions Across Sample Prevalence Quartiles. A total of 44842 unique variants across 122 samples were binned based on their sample prevalence quartile (see definition in Methods section). An increase in the combined GIAB-DB artifact rate is associated with an increase in the sample prevalence quartile. A table summarizing the distribution metrics for each boxplot is shown in the upper left-hand corner of the figure.

Sample Prevalence Quartile	1 Sample	<25% of samples	25-50% of samples	50-75% of samples	>75% of samples
1 Sample	1.000	0.000	0.000	7.036800e-283	9.527431e-229
<25% of samples	0.000	1.000	1.726533e-146	2.756110e-86	1.203049e-83
25-50% of samples	0.000	1.726533e-146	1.000	0.02033859	1.722983e-08
50-75% of samples	7.036800e-283	2.756110e-86	0.02033859	1.000	0.02387834
>75% of samples	9.527431e-229	1.203049e-83	1.722983e-08	0.02387834	1.000

Table 1. Dunn's Tests FDR adjusted P-values for Pair-wise Comparisons of Combined GIAB-DB Artifact Rate Medians Across Sample Prevalence Quartiles. Kruskal-Wallis and Post-hoc Dunn's tests identified statistically significant differences between medians for all pairwise comparisons of sample prevalence quartiles except for comparisons: 25-50% vs 50-75% and 50-75% vs >75%.

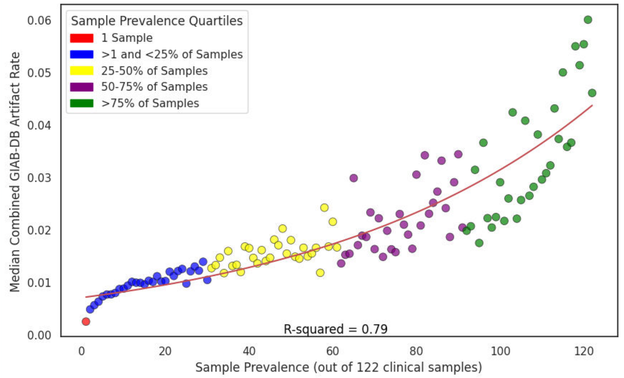


Fig 3. Scatterplot of the Median Combined GIAB-DB Artifact Rate Distributions by Sample Prevalence. Median combined GIAB-DB artifact rates were calculated and plotted across sample prevalence (1 to 122 samples). Sample prevalence quartiles are represented as different colored dots. An exponential model was fitted to the data ($y=a*e^{b*x}$, a= 0.0071, b=0.0149) and the R^2 value was calculated to see goodness of fit (R^2=0.79). An increase in sample prevalence is correlated to an exponential increase in the combined GIAB-DB artifact rate.

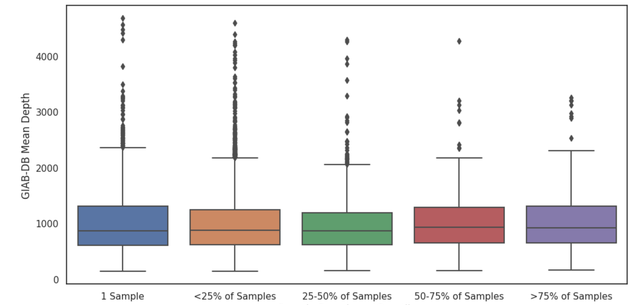


Fig 4. Boxplots of the GIAB-DB Mean Depth Distributions Across Sample Prevalence Quartiles. GIAB-DB mean depth distributions are consistent and don't explain the differences between the combined GIAB-DB artifact rates across sample prevalence quartiles.

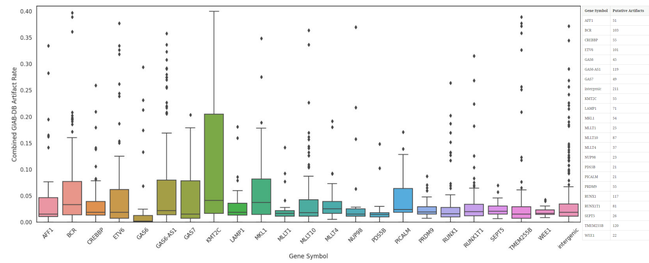


Fig 5a. Boxplots of the Combined GIAB-DB Artifact Rate of Putative Artifacts Across Top Represented MyAML Genes. Candidate artifacts found in >25% of clinical samples analyzed was used as a heuristic to determine putative artifacts. Boxplot distributions of the combined GIAB-DB artifact rates were plotted for each gene that contained at least 20 putative artifacts. A total of 1349 artifacts were analyzed across 22 genes and intergenic regions (mean artifacts per gene=67, min=21, max=211, SD=44).

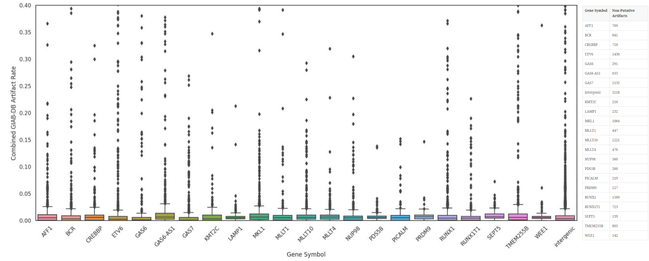


Fig 5b. Boxplots of the Combined GIAB-DB Artifact Rate of Non-Putative Artifacts Across Top Represented MyAML Genes. Candidate artifacts found in <25% of clinical samples analyzed was used as a heuristic to determine non-putative artifacts. Boxplot distributions of the combined GIAB-DB artifact rate were plotted for each gene that contained at least 20 non-putative artifacts. A total of 19237 artifacts were analyzed across 22 genes and intergenic regions (mean artifacts per gene=836, min=127, max=5218, SD=1022). Comparing same genes between Fig 4a and 4b shows higher combined GIAB-DB artifact rates in putative artifacts.

CONCLUSIONS

- Applications that rely on detecting variants at low VRFs are most impacted by normalizing error rates. Additionally, false positives add time and complexity to variant interpretation.
- The use of ABC in our clinical lab has significantly improved the turnaround times for annotation and variant reporting.
- Sample prevalence is characteristic of artifacts because identical wet-lab and computational parameters can consistently produce the same artifacts across independent samples. Our results support this hypothesis and ABC-generated metrics will allow for in depth analysis of artifacts for improved filtering.
- Lastly, the identification of patterns and genes associated with artifacts and regions with high error rates can guide future assay development across sequencing platforms and variant detection methodologies.

REFERENCES

1. Zook, Justin M., et al. "Extensive sequencing of seven human genomes to characterize benchmark reference materials." *Scientific Data*, vol. 3, no. 1, 2016, https://doi.org/10.1038/sdata.2016.25.
2. Krusche, Peter, et al. "Best practices for benchmarking germline small-variant calls in human genomes." *Nature Biotechnology*, vol. 37, no. 5, 2019, pp. 555–560, https://doi.org/10.1038/s41587-019-0054-x.