Continued development of the OncoMiner pipeline PROVEAN for scoring exonic variant effects in pediatric acute myeloid leukemia

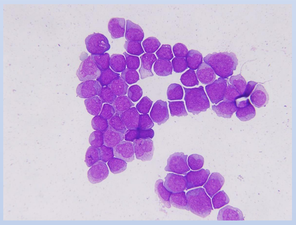
Angelica Marquez,[1] Jonathon Mohl,[2,3,4] and Ming-Ying Leung[2,3,4,5]

[1]Department of Computer Science, [2]Bioinformatics Program, [3]Border Biomedical Research Center,
[4]Computational Science Program, [5]Department of Mathematical Sciences, The University of Texas at El Paso

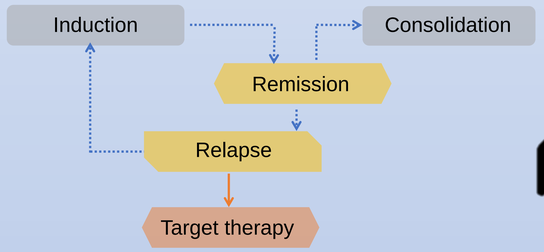
INTRODUCTION

Acute myeloid leukemia (AML) – cancer produced at the bone marrow, where myeloid stem cells cannot mature into healthy myeloblast white cells.
Whole exome sequencing (WES) -- captures protein-coding exons by solution hybridization, result of Next-generation sequencer (NGS) technology. Data collected using platform Illumina Hi-Seq 2000
Single-nucleotide polymorphisms (SNPs) – Changes in the DNA sequence by a single nucleotide substitution, insertion, or deletion.

The objective of the current study is to identify exonic DNA variants from AML patients at diagnose, in remission following induction therapy, and after relapse.



Histophathological image of acute myeloid leukemia. *Washington University. https://cancergenome.nih.gov*



Flowchart of standard treatment from AML patients by the American Cancer Society

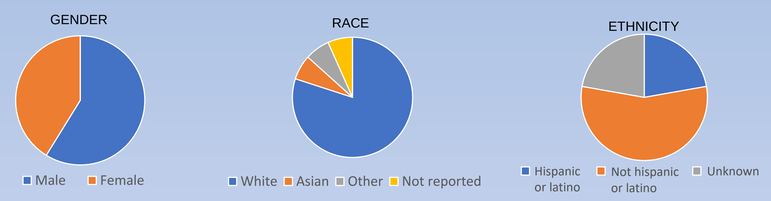
The treatment of pediatric AML been based on adult AML studies, with the assumption of similar biology. Molecular profiling of AML in pediatric and adult studies has started to characterize AML as a disease with distinct age-dependent alterations. [1]

MATERIALS AND METHODS

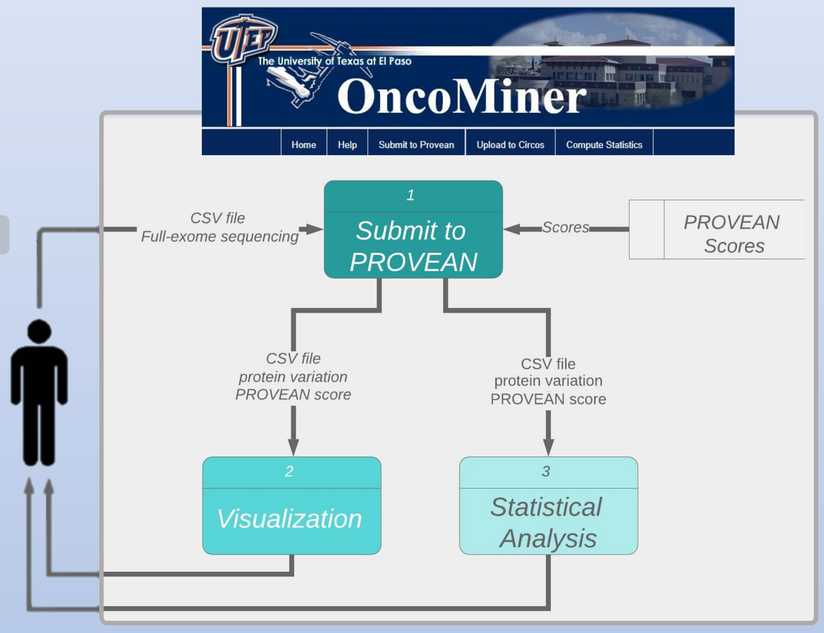
Data retrieved from The Cancer Genome Atlas (TCGA), project TARGET-AML. BAM (Binary Alignment Map) files range from 17 to 31 GB

18 patients between one and 17 years old, where each individual has three WES datasets at three phases:

- Primary tumor sample collected at diagnosis[2]
- Case-matched tissue sample collected at remission (<5% blasts detected following standard induction therapy)[2]
- Relapsed tumor sample[2]



OncoMiner (OncoMiner.utep.edu) – Computational pipeline identifies possible cancer associated variants from a list of single-nucleotide substitutions ad small insertions/deletions reported as the result of WES. Variant scoring, visualization, and statistical comparison.[2]



Flowchart of OncoMiner pipeline, including data flow of inputs and outputs

PROVEAN stands for **Pro**tein **V**ariation **E**ffect **A**nalyzer.
It provides predictions of the functional effects of protein sequence variations, including single or multiple substitutions and insertions and deletions (indels) in amino acid sequences. [3]

$$\Delta(Q, v, S) = A(Q', S) - A(Q, S)$$

Q Query sequence
Q' Query mutated by variant
v variant
S database sequence S
A Alignment score

A negative delta score would imply that the variant causes the query sequence to become less similar to S. Significantly negative average delta scores can suggest that the variant v has a deleterious effect. [3]

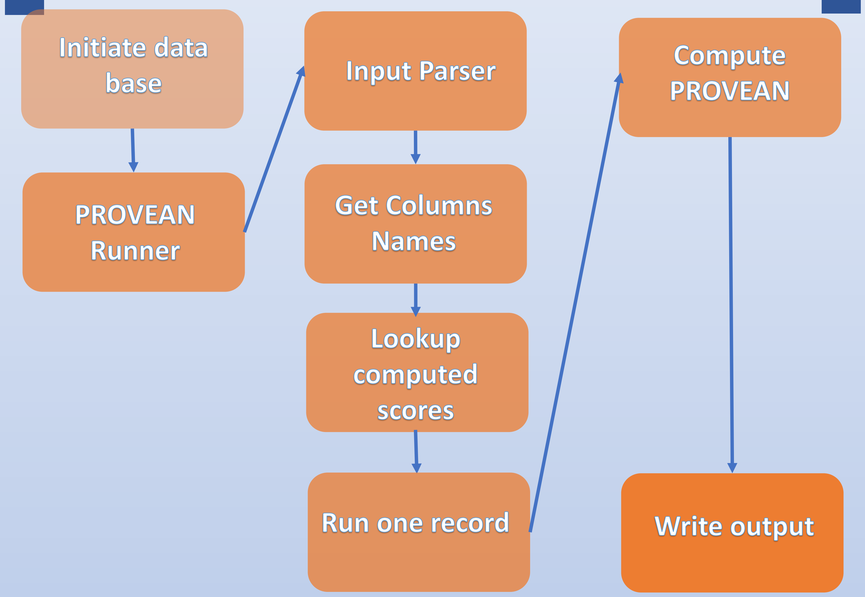
var_index	chrom	gene_nam	left	right	ref_seq	var_seq1	var_seq2	count	var_score
1977096	chr10	IDI1	1.28E+08	128115711	T		T	14	29.35714
2009630	chr4	ZNF141	1.9E+08	189953115	A	T	T	31	28.87097
594084	chr11	SIGIRR	26565186	26565187	T	T	T	6	29.66667
959029	chr13	ZMYM2	1.13E+08	113197969	A	G	G	70	29.6
2296082	chr12	NINJ2	1.24E+08	123606268	G	A	A	13	27.69231

OncoMiner Input (OMI) File

var_index	chrom	gene_name	left	right	ref_seq	var_seq1	var_seq2	count	var_score	where	change_type1	ref_aa	var_aa	ref_codon	var_codon	AA.variation	PROVEAN.score
2285847	chr12	SBNO1	123309736	123309737	T		T	11	26	CDS	Non-synonymous	E	V	GAA	GTA	I1348_A1392del	-62.385
8469	chr11	NLRP6	280816	280817	A	T	T	63	27.80952	CDS	Synonymous	V	V	GTA	GTT	Y361F	2.534
8469	chr11	NLRP6	280816	280817	A	T	T	63	27.80952	CDS	Synonymous	V	V	GTA	GTT	Y361F	2.534
1402	chr1	OR4F5	69968	69969	A	G	G	10	26.2	CDS	Synonymous	T	T	ACA	ACG	Q293R	2.659
1229	chr1	OR4F5	69728	69729	T	C	C	15	25.26667	CDS	Synonymous	Y	Y	TAT	TAC	I213T	-3.179
854	chr1	OR4F5	69127	69128	T	C	C	9	26.77778	CDS	Non-synonymous	F	S	TTC	TCC	S13P	-3.754
1207	chr1	OR4F5	69679	69680	G	T	T	9	29.66667	CDS	Non-synonymous	W	L	TGG	TTG	G197C	-8.864

Expected output PROVEAN csv file

RESULTS



Flowchart of OncoMiner pipeline, including data flow of inputs and outputs

DISCUSSION

- There is a necessity of target therapy for pediatric acute leukemia.
- The communication diagram supported component level testing for "Submit to Provean" troubleshooting.
- Further problems are still being identified, most of them have been in the pre-Provean computation.
- Additional corrections are necessary to analyze larger data sets using OncoMiner pipeline.
- NGS data analysis is time consuming. Efficient parallelized computational approaches on high performance computers are desirable for processing such large datasets.

REFERENCES

1. Farrar, J. E., Schuback, H. L., Ries, R. E., Wai, D., Hampton, O. A., Trevino, L. R., . . . Meshinchi, S. (2016). Genomic Profiling of Pediatric Acute Myeloid Leukemia Reveals a Changing Mutational Landscape from Disease Diagnosis to Relapse. Cancer Research, 76(8), 2197-2205. doi:10.1158/0008-5472.can-15-1015
2. GenomeOC. (2017, August 18). Acute Myeloid Leukemia. Retrieved from https://ocg.cancer.gov/programs/target/acute-myeloid-leukemia
3. Leung MY., Knapka J.A., Wagler A.E., Rodriguez G., Kirken R.A. (2016) OncoMiner: A Pipeline for Bioinformatics Analysis of Exonic Sequence Variants in Cancer. In: Wong KC. (eds) Big Data Analytics in Genomics. Springer, Cham

ACKNOWLEDGEMENTS

Research reported in this poster was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under linked Award Numbers RL5GM118969, TL4GM118971, and UL1GM118970. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Equipment in the Bioinformatics Core Facility used for this project is supported in part by the NIH Grant #5G12MD007592 from the National Center for Research Resources to the UTEP Border Biomedical Research Center.