bamSliceR: cross-cohort variant and allelic bias analysis for rare variants and rare diseases

Yizhou Peter Huang[a,b] (https://orcid.org/0009-0004-5689-0193), Lauren Harmon[b] (https://orcid.org/0000-0002-2248-060), Eve Gardner[b](https://orcid.org/0009-0009-4133-8343), Xiaotu Ma[c] , Josiah Harsh[b], Zhaoyu Xue[b] (https://orcid.org/0000-0001-7288-8780), Hong Wen[b] (https://orcid.org/0000-0001-8739-4572), Marcel Ramos[d] (https://orcid.org/0000-0002-3242-0582), Sean Davis[e] (https://orcid.org/0000-0002-8991-6458), Timothy J. Triche, Jr.[b,a](https://orcid.org/0000-0001-5665-946X)

[a] Michigan State University, East Lansing, MI, US
[b] Van Andel Institute, Grand Rapids, MI, US
[c] St. Jude Children's Research Hospital, Memphis, TN, US
[d] Roswell Park Cancer Institute: Buffalo, NY, US
[e] University of Colorado Anschutz Medical Campus: Aurora, CO, US

Abstract

Rare diseases and conditions create unique challenges for genetic epidemiologists precisely because cases and samples are scarce. In recent years, whole-genome and whole-transcriptome sequencing (WGS /WTS) have eased the study of rare genetic variants. Paired WGS and WTS data are ideal, but logistical and financial constraints often preclude generating paired WGS and WTS data. Thus, many databases contain a patchwork of specimens with either WGS or WTS data, but only a minority of samples have both. The NCI Genomic Data Commons facilitates controlled access to genomic and transcriptomic data for thousands of subjects, many with unpaired sequencing results. Local reanalysis of expressed variants across whole transcriptomes requires significant data storage, compute, and expertise. We developed the ***bamSliceR*** package to facilitate swift transition from aligned sequence reads to expressed variant characterization. ***bamSliceR*** leverages the NCI Genomic Data Commons API to query genomic sub-regions of aligned sequence reads from specimens identified through the robust Bioconductor ecosystem. We demonstrate how population-scale targeted genomic analysis can be completed using orders of magnitude fewer resources in this fashion, with minimal compute burden. We demonstrate pilot results from ***bamSliceR*** for the TARGET pediatric AML and BEAT-AML projects, where identification of rare but recurrent somatic variants directly yields biologically testable hypotheses. ***bamSliceR*** and its documentation are freely available on GitHub at https://github.com/trichelab/bamSliceR .

Introduction

RNA sequencing (RNA-seq) captures transcriptomes with allelic resolution, and thereby can implicate functional genomic variants in disease. The underlying technology is now sufficiently mature to be a staple in research and clinical settings. A recent surge in RNA-seq data from clinical trials has overtaken whole genome sequencing (WGS) results in many settings as a primary tool for interrogating genome function and directly assessing expressed genomic variants[1]. For example, multiple Children's Oncology Group trials have collected RNA-seq data as a biological correlative, often outpacing genome sequencing in the same populations[2]. With continued accrual, the number of patients with RNA-seq data has nearly tripled, without a concomitant increase in WGS data[3]. Combining results from WGS and WTS can maximize variant yield for rare diseases.

In 2019, the Leucegene project recognized this opportunity, and developed a local assembly approach to identify mutations in RNA-seq data[4]. This method is computationally efficient relative to prior whole genome variant calling algorithms, but it involves a re-alignment process and does not take full advantage of the standardized Binary Alignment and Mapping (BAM) format[5] supported by cloud-based repositories such as the NCI Genomic Data Commons (GDC). The GDC is an important resource of harmonized sequencing data from large translational research consortia, including The Cancer Genome Atlas[6] (TCGA), TARGET[2], and BEAT-AML[7]. The GDC Application Programming Interface (API) allows authorized users to query aligned data via genomic ranges or slices[8]. This provides an efficient basis for pipelines and tools facilitating rapid exploration, annotation, and experimental validation of genetic lesions in rare diseases.

Here, we present an R/Bioconductor software package (***bamSliceR***) for automatically and efficiently extracting coordinate- or range-based BAM reads from targeted genomic regions, and from large RNA-seq cohorts. bamSliceR bypasses the historically tedious variant calling workflow, thereby reducing time, space, and computational burden by orders of magnitude relative to standard methods. Users can quickly transition from raw alignment data to variant characterization, and perform population-scale targeted genomic inspections. We leverage RNA-seq data from more than 3,000 subjects to demonstrate how this lightweight workflow can expand sample size for rare variants and diseases, increasing scale and scope for discovery and validation.

Methods

The ***bamSliceR*** package has four basic functions, as illustrated in Figure 1. The first function integrates the GDC API within the ***bamSliceR*** R statistical programming environment, which enables users to query databases, remotely BAM-slice genomic regions of interest, and automatically import those sequences for local alignment. *bamSliceR* generates a metadata frame for each BAM file, which enables users to custom-filter the data based on sample type(s), sequencing type(s) (RNA-seq/WGS), or other criteria. This enhances the user's ability to tailor the analysis to their specific needs. Next, ***bamSliceR*** uses gmapR[9] and VariantTools[10] to tally coverage and counts of variant alleles. ***bamSliceR*** then estimates the Variant Allele Fraction (VAF) for each mutation, and applies VariantAnnotation[11] to predict associated amino acid changes. This allows parallel computing and accelerated analysis time for users with access to a High-Performance Computing (HPC) cluster or cloud. Annotation of predicted variant consequence is delegated to the Ensembl Variant Effect Predictor (VEP). Annotated variants can be exported to VCF, VRanges objects, or Mutation Annotation Format (MAF)[12]. ***bamSliceR*** also incorporates hooks for downstream analysis and visualization (co-occurrence, oncoplots, lollipop diagrams for protein hotspots, and survival analysis with Kaplan-meier plots).

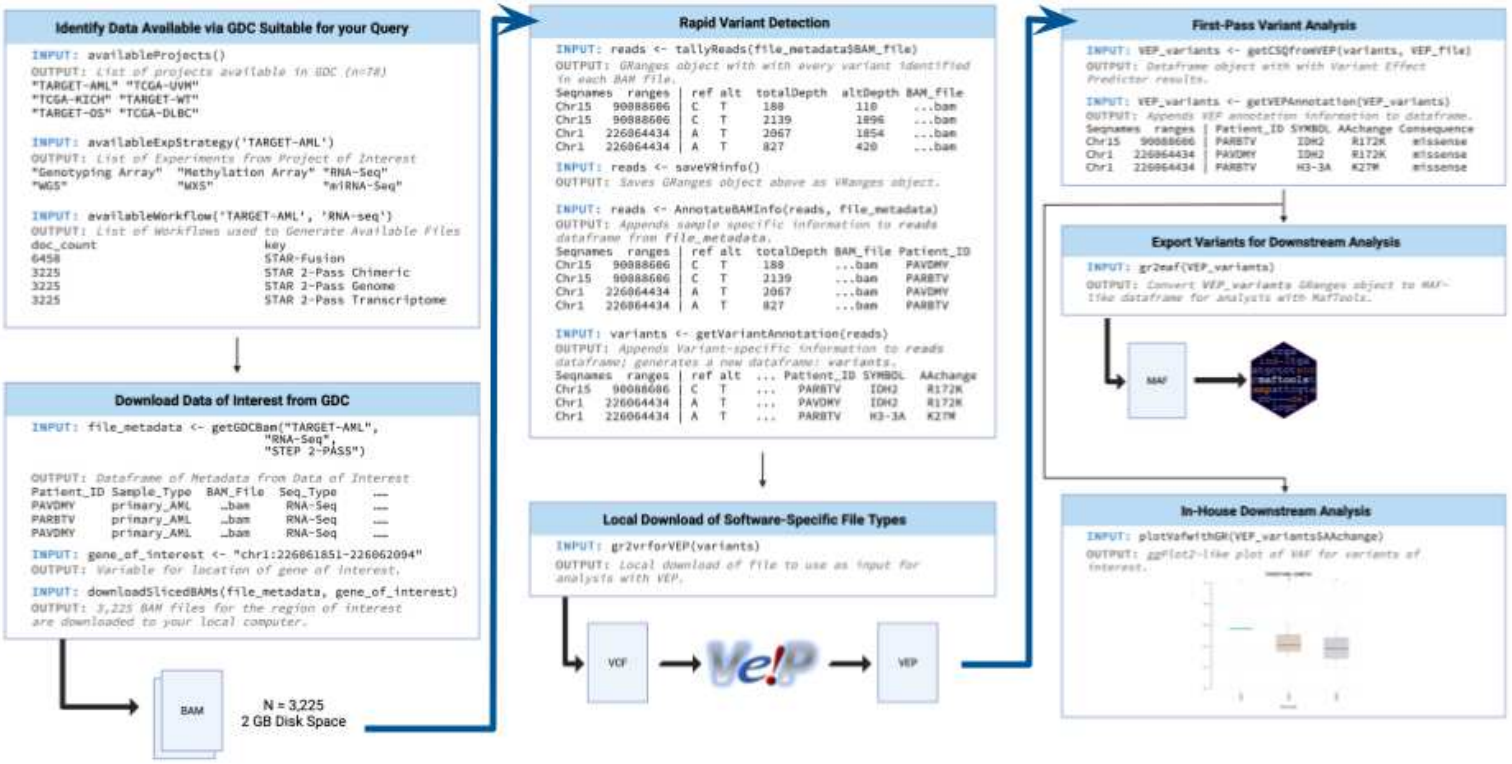


Figure 1. bamSliceR Workflow and Functionality for querying, downloading, and annotating BAM files.

Results

Oncohistone variants in pediatric leukemia

Pediatric acute myeloid leukemia (AML) is a genetically heterogeneous and often lethal disease[13]. Most cases reveal few if any actionable sequence variants, with a remarkably low mutation burden and a preponderance of diverse structural variants[14]. Nevertheless, molecular features can drive treatment decisions. For example, most patients with *DEK::NUP214* fusions harbor co-occurring *FLT3* internal tandem duplications. Targeting this aberration revealed the immunogenicity of the fusion, and when combined with stem cell transplantation, has improved 5-year survival from under 10% to over 85% in young patients[15]. Unfortunately, identifying actionable sequence variants from WGS is challenging both technically and statistically, as the population of AML patients is small relative to common diseases. Characterized pediatric cases with WGS number in the low hundreds. The analysis of expressed sequence variants from RNA-seq reads is therefore quite attractive.

Mutations of histone H3.3 lysine 27 (H3K27) to methionine (M, H3K27M) were originally documented in high-grade midline glioma,[16] where they frequently accompany *TP53* mutations. More recently, mutations of H3.1 K27 to isoleucine (I) or methionine (H3K27I/M) were documented in adult AML patients[17] and in pre-leukemic stem cells from patients who went on to develop secondary AML[18]. However, the genetic context, age groups (histone mutations have not previously been reported in pediatric AML patients), mechanism, and clinical impact of H3K27 mutations in myeloid leukemogenesis remains poorly understood at best.

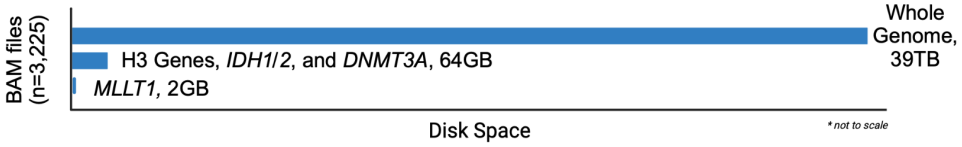


Figure 2. Reduction of Disk Space Required for Analysis by BamSlicing. The 3,225 RNAseq BAM files obtained from 2,281 subjects in TARGET-AML are approximately 39TB. Using bamSliceR to slicing the reads covering exons of 11 Histone 3 genes (*H3F3A, HIST1H3A, HIST1H3H, HIST1H3I, HIST1H3J, HIST1H3B, HIST1H3C, HIST1H3D, HIST1H3E, HIST1H3F, HIST1H3G*) and genes encoding epigenetic factors (*IDH1/2, DNMT3A, RUNX1, ASXL1/2, TET1/2*) that collectively span 152,026 bp of non-contiguous genomic space reduced the total size of BAM files to 62GB. Furthermore, slicing the reads covering a single Human *ENL* gene span of 73,594 bp only requires 2GB of total disk space for 3,225 BAM files.

To investigate the genetic landscape of H3K27M in AML across age groups, we examined 11 histone 3 genes as well as genes encoding epigenetic factors that are frequently mutated in AML patients (*IDH1/2*, *DNMT3A*, *RUNX1*, *ASXL1/2*, and *TET1/2*). These genes collectively span 152,026 bp of non-contiguous genomic space. Using the *bamSliceR* pipeline, we automatically processed 3,225 and 735 RNA-seq BAM files obtained from 2,281 and 653 subjects in the TARGET-AML (pediatric) and BEAT-AML (adult) cohorts, respectively. This step alone reduced the total size of the BAM files of the TARGET-AML cohort from 39TB to 62GB (Fig. 2), while still retaining all the essential genetic information required to perform an epidemiological study of these rare AML mutations. We identified 9 pAML and 7 adult AML patients that harbored a K27M mutation, with multiple lines of evidence based on Variant Allele Frequency (VAF >0.15), total read depth (>8), and WGS data where available (Supplementary Table S1). We found H3K27M mutations on both replication-coupled H3.1 and replication-independent H3.3, with most mutations occurring in the H3.1 gene (Supplementary Table S1). The incidence of H3K27M in pediatric AML is lower (~0.3 %) than in adult AML (~0.8%; Supplementary Table S2), and we confirmed the existence of H3K27 variants in normal karyotype induction failure patients (Figure 4).

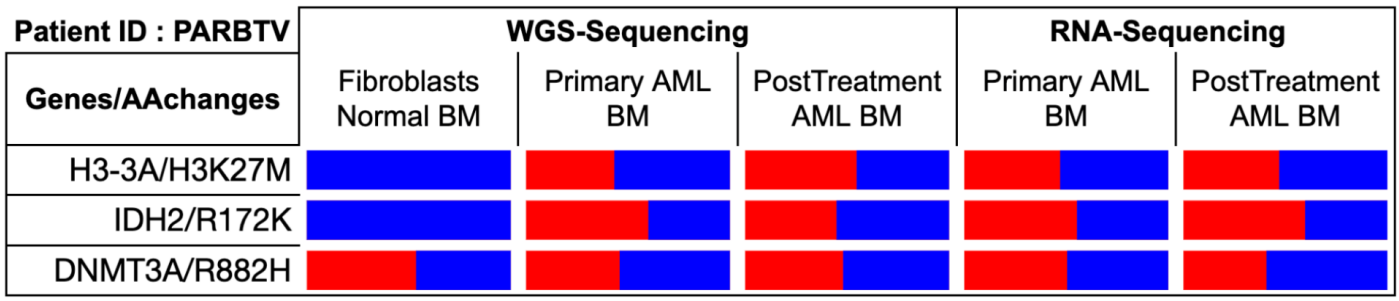


Figure 4. Clonal mutations observed in TARGET AML subject PARBTV from WGS and RNAseq Data. Variant allele frequency of mutations *H3F3A K27M, IDH2 R172K* (somatic), and *DNMT3A R882H* (germline) of a pediatric patient from the TARGET-AML cohort with both RNAseq and WGS data from two timepoints (diagnosis and after treatment) (BLUE: allele frequency of WT; RED: allele frequency of mutant).

We used ***bamSliceR*** to generate VAF distribution plots for each mutation and estimate their clonal status. For example, we see that *IDH2* ^{R172K} and *DNMT3A*^{R132H} mutations are persistent clonal events (VAF ~50%) in the pAML cohort, that were previously believed to only occur in adult AML patients (Fig 3C; Supplementary Fig S1). We discerned that the K27M mutation in *H3C2*, *H3C3*, *H3C4*, *H3C11* (H3.1) and *H3F3A* (H3.3) genes are always clonal (Supplementary Fig S2), consistent with their occurrence in pediatric high-grade glioma. Oncoplots and mutual exclusivity analysis showed that *H3F3A* K27M (H3.3) and *IDH2* mutations always co-occur ($p<0.05$). Interestingly, two pAML patients harboring *H3F3A* K27M and *IDH2*^{R172K} mutations failed induction therapy, suggesting that the two mutations may synergize to cause chemoresistance. For pAML patients where RNA-seq and WGS data are both available, we used bamSliceR to confirm that the somatic *H3F3A* K27M and *IDH2*^{R172K} mutations are constantly expressed at high levels, with low allelic bias, consistent with DNA sequencing results (Supplementary Table S1; Fig 4F). To facilitate in-depth studies of patients with samples at multiple disease stages, ***bamSliceR*** includes functionality to identify and subset the matched subjects and data files. Using this function, we found that one pAML patient with somatic *H3F3A* K27M and *IDH2*^{R172K} mutations also harbored a germline *DNMT3A*^{R882C} mutation throughout disease progression (Fig 4).

Taken together, these data suggest that H3K27M is a clonal mutation that may synergize with metabolic and epigenetic variants (e.g. *IDH2*^{R172K} and *DNMT3A*^{R882C}) to drive aggressive and refractory pAML. *IDH* mutations have been thought mutually exclusive with *H3K27* variants. Not only do they co-occur, *IDH2* mutations are in fact enriched for *H3K27* mutant cases. By using RNA-seq data to expand our sample size and automating the alignment and variant calling process in ***bamSliceR***, we document statistically and clinically significant co-occurrence of oncometabolic *IDH2* variants with high-risk *H3K27* pAML mutant cases, yielding testable hypotheses and new translational avenues[19] for a subset of AML patients at high risk of treatment failure.

MLLT1 YEATS domain indels in pediatric tumors

The ENL protein (encoded by the *MLLT1* gene) is a subunit of the super elongation complex (SEC) involved in transcriptional elongation during early development. Small in-frame insertion-deletion (indel) mutations within the YEATS domain of *MLLT1* were first identified in Wilms tumors[20]. Further work revealed that indels in the YEATS domain alter chromatin states, dysregulating cell-fate control and driving tumorigenesis[21]. These same indels can transform hematopoietic cells by mitigating polycomb silencing, but only one case of pediatric AML that harbored this type of mutation had previously been documented (http://cancer.sanger.ac.uk/cosmic).

As a second case study for the efficacy of our package, we used ***bamSliceR*** to analyze the TARGET RNA-seq and WGS data for any evidence of ENL (*MLLT1*) YEATS domain mutations. Using our pipeline, we discovered three pAML patients carrying mutations in *ENL*-YEATS, with one pAML patient exhibiting allelic bias similar to that observed in Wilms tumor patients (Fig 5). We also identified putative indel mutations near the YEATS domain that have not been detected in Wilms tumor patients (Supplementary Table S3). These new results lead to the biologically testable hypothesis that *ENL*-YEATS mutations primarily upregulate *HOXA* gene expression in pAML, similar to their role and function in favorable histology Wilms tumors[22]

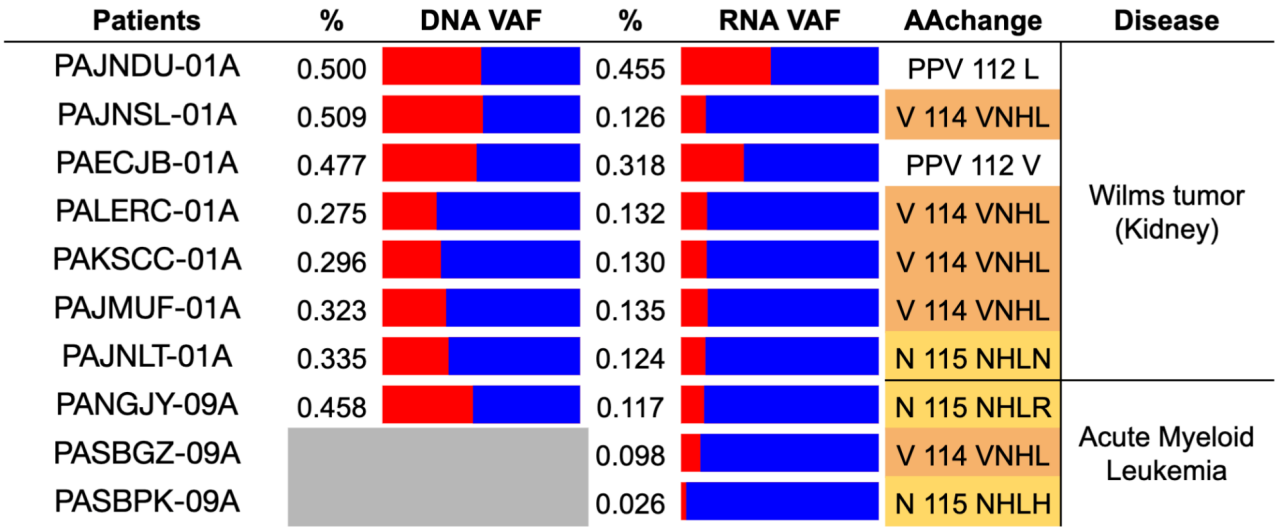


Figure 5. Analysis of Rare *ENL* YEATS Domain Mutations in pAML patients.
Allelic bias based on Variant Allele Frequency of in-frame insertion mutations in Human *ENL* gene of Wilms Tumor and AML patients with both RNAseq and WGS data.

Discussion

We developed ***bamSliceR*** to address two practical challenges: resource-sparing identification of candidate subjects, and variant detection from aligned sequence reads, across thousands of controlled-access subjects. The GDC BAM Slicing API (https://docs.gdc.cancer.gov/API/Users_Guide/BAM_Slicing/) is a practical and well documented REST API for this purpose. Yet, perhaps due to the challenge of the former task, we find relatively little published work referencing usage of the GDC API, even in studies where specific candidate variants are evaluated. Instead, reliance upon variant calls from existing studies, or transfer and recall of variants from raw sequence data, is often documented. The former assumes a single best method for variant detection fits all experimental designs (an assumption contradicted by many benchmarks[*cite*]). The latter is grossly inefficient.

In clinical genetic analysis, direct evaluation of fragment-level evidence for a candidate variant is routine, regardless of the confidence level a variant calling pipeline may assign to a putative genetic variant. The same raw material is available via controlled access across many population-scale projects representing billions of dollars in public funding. The relatively simple toolkit we provide here extends this practice in an efficient and user-friendly way to the Bioconductor ecosystem, significantly expanding the pool of potential subjects for rare

variant detection in rare diseases. Importantly, the same well-documented API implemented by the GDC is feasible for Common Fund datasets, such as the Gabriella Miller Kids First! (GMKF) and INCLUDE projects, and can be further extended to transcriptome-indexed reads as well as open access resources such as SRA.

Support for efficient retrieval of range-based queries is a key feature of the SAM/BAM format and htslib/htsget implementations. Authentication and authorization create challenges for straightforward usage in controlled access data, but as the GDC API and its downstream users illustrate, this challenge can be overcome, and is increasingly important as projects with whole-genome sequence data from population-scale biobanks emerge. We present bamSliceR as a concrete example of what can be accomplished by wedding clinical and genomic data management processes to an efficient, standardized API. Our hope is that it will enable users to perform previously challenging evaluation of raw data evidence for genetic and genomic variants at scale, and that user uptake will spur further expansion of support for coordinate-based queries of sequence databases.

Acknowledgements

The research in this paper was supported in part by NIAID R01 AI171984 to TJT, NCI P50 CA254897 to TJT, the Michelle Lunn Hope Foundation, and the Van Andel Institute. Computation for the work described in this paper was supported by the High Performance Cluster and Cloud Computing (HPC3) Resource at the Van Andel Research Institute. We thank Darrell Chandler for his expertise and insight throughout all aspects of the study and for his assistance in copy-editing the manuscript.

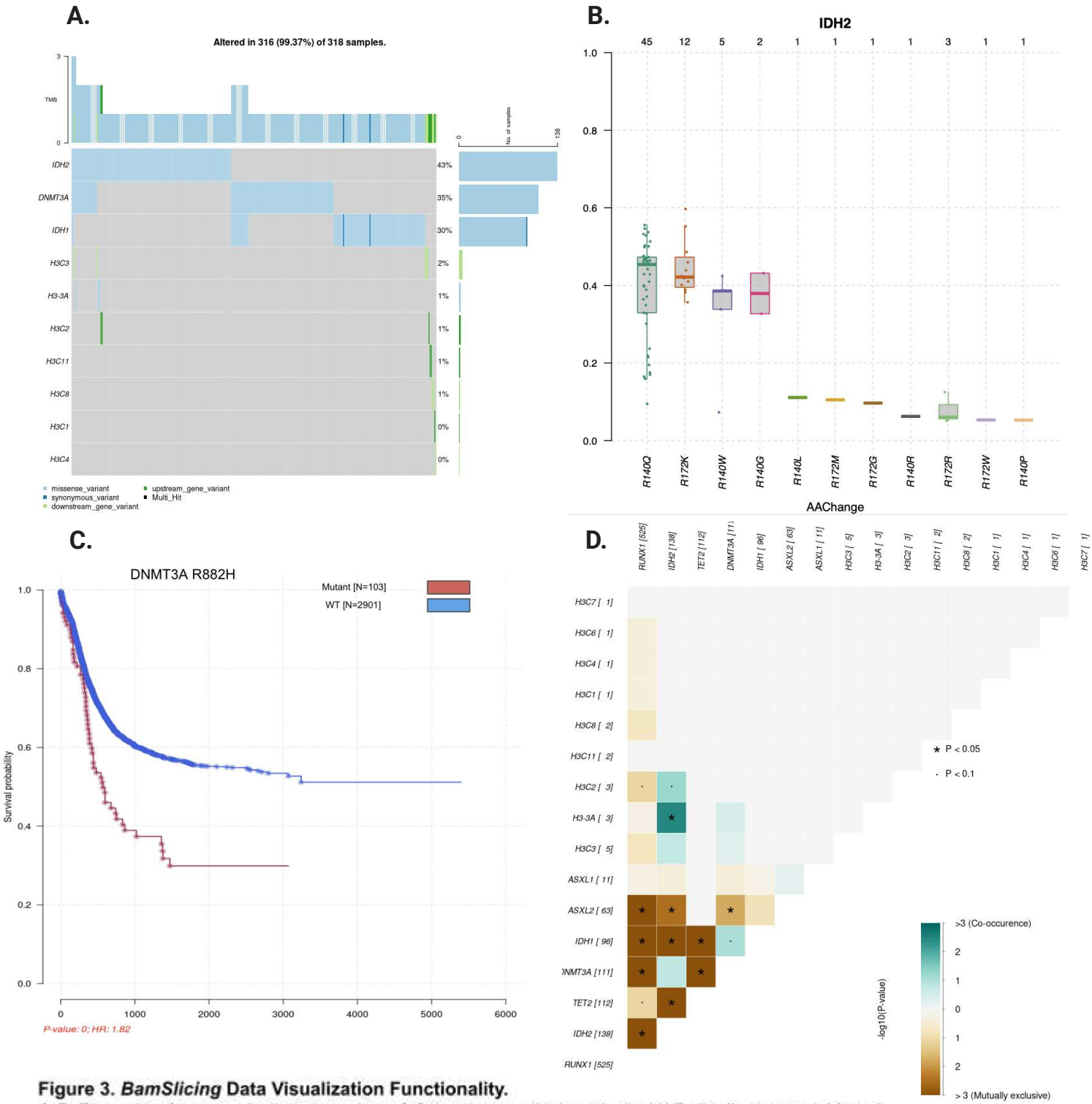


Figure 3. *BamSlicing* Data Visualization Functionality.
A-D Example of automatically generation of *Oncoplot*, *survival analysis*, *VAF distribution* and *Mutual Exclusivity analysis*. **B.** VAF plotting of different mutations of *IDH2* at either *R140* or *R172* ordering by median of VAF. *R140Q* and *R172K* are most prevalent and *R172K* mutation are always clonal which are usually have mean allele frequency around ~50% assuming pure sample (VAF plotting of *H3K27M* and *DNMT3A* are in Supplementary Figure S1 and S2). **C.** Kaplan meier curve by grouping samples based on mutation status (WT vs. DNMT3A) in both TARGET and BEAT AML cohorts (n = 2934). **D.** Mutual exclusivity plot by pair-wise Fisher's Exact test detected *H3-3A K27M* and *IDH2 R172K* are co-occuring mutations ($p < 0.0032$).

Supplemental data

(Each table should be a separate excel file.)

Table S1

Co-occurence status of *H3K27M* and *IDH2* mutations in TARGET-AML and BEAT-AML.
Identification of 9 pAML patients that harboared K27M mutation on Histone 3 genes based on evidence of VAF > 0.15 and total read depth > 8 and whether the mutation can be captured by both RNA-seq and WGS data. Co-occurence of *H3F3A* (H3.3) *K27M* and *IDH2 R172K* are shown in 2 pAML patients. Co-occurence of *H3K27M* (H3.3 & H3.1) and *IDH2 R140Q/R172K* are shown in 5 adult AML patients.

		H3K27M					IDH2				
	CASE_ID	SYMBOL	DNA-vaf-T1	DNA-vaf-T2	RNA-vaf-T1	RNA-vaf-T2	SYMBOL	DNA-vaf-T1	DNA-vaf-T2	RNA-vaf-T1	RNA-vaf-T2
H3.3	PARBTV	H3F3A	47%	48%	47%%	47%	R172K	38%	48%	56%	59%%
	PAVDMY	H3F3A			49%	47%/47%	R172K			39%	40%/48%
H3.1	PAPVGE	HIST1H3C			58%%		-			-	
	PAUZTH	HIST1H3J(K27I)			0%	38%%	-			-	-
	PAUUPR	HIST1H3I			0%	35%%	-			-	-
	PAKWCU	HIST1H3D			17%%		-			-	
	PAXFAG	HIST1H3C			54%%		-			-	
	PAXKAL	HIST1H3I			50%		-			-	
	PATFGK	HIST1H3C			50%		-			-	

		H3K27M				IDH2			
	CASE_ID	SYMBOL	DNA-vaf	RNA-vaf	TS-vaf	SYMBOL	DNA-vaf	RNA-vaf	TS-vaf
H3.3	2148	H3F3A	29%	47%		R140Q	40%	52%	
H3.1	2354	HIST1H3C	34%/33%	low-exp		R172K	36%/31%	47%	
	2429	HIST1H3B	33%/47%	35.29%/100%		-	-	-	
	2498	HIST1H3C	37%	38%		R140Q	48%	47%	
	2530	HIST1H3	40%	low-exp		R172K	46%	49%	

		B							
	2611	HIST1H3D		58%	39%	-		-	-
	2721	HIST1H3B			36%	R172K			36%

Table S2

***H3K27* variants: published work (n=1049) and TARGET/BEAT-AML cohorts (n = 2934, via bamSliceR).** Lehnertz et al. *Blood 2017* documented 2 adult AML patients. Boileau et al. *Nat Commun 2019* documented 4 adult AML patients. We documented 16 AML patients from TARGET and BEAT AML cohorts.

Cohort	Cohort Size (n)	WGS/WXS (n)	RNAseq (n)	DNA Methylation (n)	H3K27M/I (%/n)
Lehnertz et al. *Blood* (2017)					
Leucegene	415		415		0.48%/2
Boileau et al. *Nat Commun* (2019)					
Toronto	312	312			0.64%/2
Lebanon	122	122			0.8%/1
TCGA	200	200	200	200	0.5%/1
Identification of H3K27M in BEAT-AML and TARGET-AML					
TARGET 20/21	2045	365	**2281**	2000	0.4%/9
Beat-AML	826	798	**653**		0.8%/7

Table S3

Human *MLLT1* YEATS domain insertion/deletion variants identified in the pan-TARGET cohort.

Chromosome	POS	SYMBOL	AAchange	REFCODON	VARCODON	REFAA	VARAA	alt_count	totla_count	VAF	patient_id
chr19	6230649	MLLT1	V114VNHL	GTG	GTGAACCACCTG	V	VNHL	5	51	0.09804	PASBGZ
chr19	6230642	MLLT1	H116HLRP	CAC	CACCTGCGCCCC	H	HLRP	2	460	0.00435	PANGJY
chr19	6230645	MLLT1	N115NPLR	AAC	AACCCCCTGCGC	N	NPLR	4	470	0.00851	PANGJY
chr19	6230645	MLLT1	N115NHLR	AAC	AACCACCTGCGC	N	NHLR	55	470	0.11702	PANGJY
chr19	6230640	MLLT1	HL116L	CACCTG	CTG	HL	L	2	134	0.01493	PALHVV
chr19	6230646	MLLT1	N115NHLH	AAC	AACCACCTGCAC	N	NHLH	2	76	0.02632	PASBPK
chr19	6230582	MLLT1	LL135L	CTCCTG	CTG	LL	L	2	175	0.01143	PAUMUZ
chr19	6230611	MLLT1	NP126P	AACCCC	CCC	NP	P	2	370	0.00541	PAUHGM
chr19	6230616	MLLT1	TFN123N	ACCTTCAAC	AAC	TFN	N	2	48	0.04167	PAWVPZ
chr19	6230620	MLLT1	TF123F	ACCTTC	TTC	TF	F	2	690	0.00290	PAVDXR
chr19	6230574	MLLT1	AG138G	GCCGGC	GGC	AG	G	2	175	0.01143	PABYYR
chr19	6230570	MLLT1	GG139G	GGCGGG	GGG	GG	G	2	1047	0.00191	PAUWZR

Figure S1

VAF distribution of DNMT3A variants.

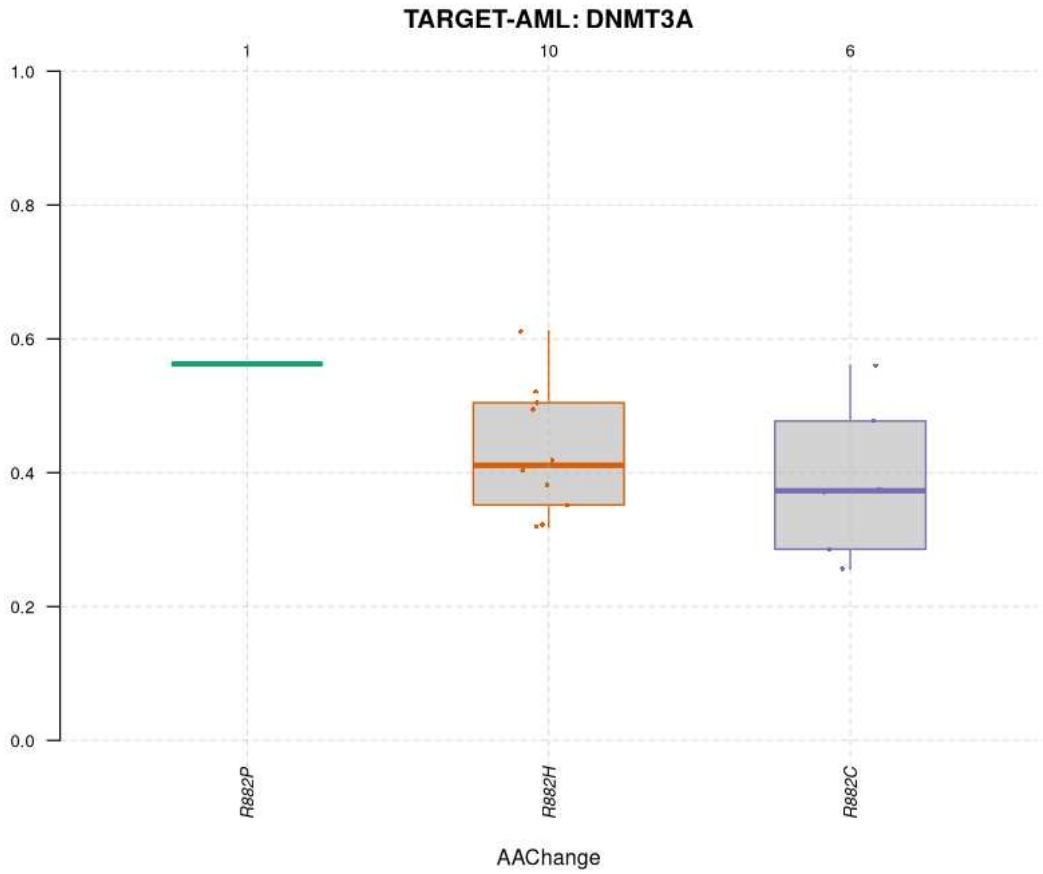
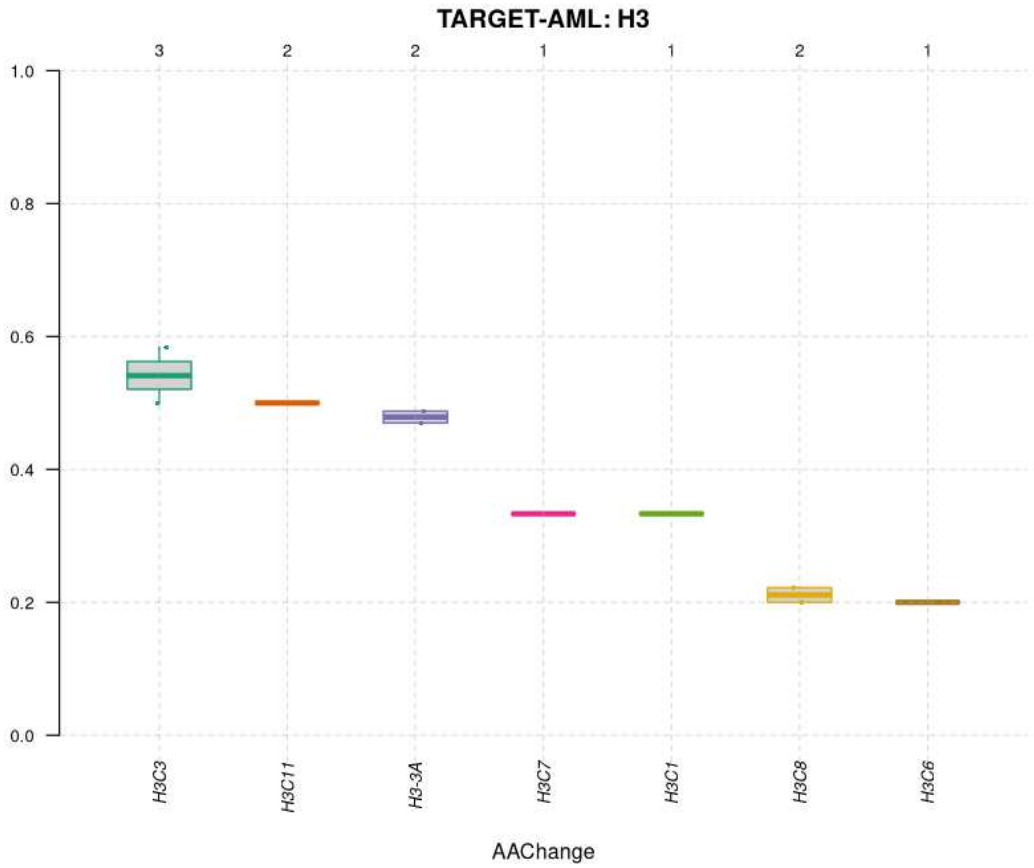


Figure S2

VAF distribution of H3K27 variants.



References

1. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *Int. J. Mol. Sci.* **18**, (2017).
2. Bolouri, H. *et al.* The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2018).
3. Farrar, J. E. *et al.* Long Noncoding RNA Expression Independently Predicts Outcome in Pediatric Acute Myeloid Leukemia. *J. Clin. Oncol.* JCO2201114 (2023).
4. Audemard, E. O. *et al.* Targeted variant detection using unaligned RNA-Seq reads. *Life Sci Alliance* **2**, (2019).
5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
6. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
7. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
8. Wilson, S. *et al.* Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API. *Cancer Res.* **77**, e15–e18 (2017).
9. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
10. Lawrence, M. & Gentleman, R. VariantTools: an extensible framework for developing and testing variant callers. *Bioinformatics* **33**, 3311–3313 (2017).

11. Obenchain, V. *et al.* VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
12. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
13. Gruber, T. A. *et al.* An Inv(16)(p13.3q24.3)-encoded CBFA2T3-GLIS2 fusion protein defines an aggressive subtype of pediatric acute megakaryoblastic leukemia. *Cancer Cell* **22**, 683–697 (2012).
14. Bolouri, H. *et al.* A B-cell developmental gene regulatory network is activated in infant AML. *PLoS One* **16**, e0259197 (2021).
15. Tarlock, K. *et al.* Significant Improvements in Survival for Patients with t(6;9)(p23;q34)/DEK-NUP214 in Contemporary Trials with Intensification of Therapy: A Report from the Children's Oncology Group. *Blood* **138**, 519 (2021).
16. Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226–231 (2012).
17. Lehnertz, B. *et al.* H3 K27M/I mutations promote context-dependent transformation in acute myeloid leukemia with RUNX1 alterations. *Blood* **130**, 2204–2214 (2017).
18. Boileau, M. *et al.* Mutant H3 histones drive human pre-leukemic hematopoietic stem cell expansion and promote leukemic aggressiveness. *Nat. Commun.* **10**, 1–12 (2019).
19. Thomas, D. *et al.* Dysregulated Lipid Synthesis by Oncogenic IDH1 Mutation Is a Targetable Synthetic Lethal Vulnerability. *Cancer Discov.* **13**, 496–515 (2023).
20. Perlman, E. J. *et al.* MLLT1 YEATS domain mutations in clinically distinctive Favourable

Histology Wilms tumours. *Nat. Commun.* **6**, 1–10 (2015).

21. Wan, L. *et al.* Impaired cell fate through gain-of-function mutations in a chromatin reader. *Nature* **577**, 121–126 (2020).

22. Gadd, S. *et al.* A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat. Genet.* **49**, 1487–1494 (2017).