RNAmut: robust identification of somatic mutations in acute myeloid leukemia using RNA-sequencing

Muxin Gu,[1,2] Maximillian Zwiebel,[1,3] Swee Hoe Ong,[4] Nick Boughton,[5] Josep Nomdedeu,[1,2,6] Faisal Basheer,[1,2,7] Yasuhito Nannya,[8] Pedro M. Quiros,[1,2] Seishi Ogawa,[8] Mario Cazzola,[9] Roland Rad,[3] Adam P. Butler,[4] MS Vijayabaskar[1,2] and George S. Vassiliou[1,2,7]

[1]Haematological Cancer Genetics, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [2]Wellcome Trust–MRC Stem Cell Institute, Cambridge Biomedical Campus, University of Cambridge, Cambridge, UK; [3]German Consortium for Translational Cancer Research (DKTK), Partnering Site, Munich, Germany; [4]Cancer Ageing and Somatic Mutation, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [5]Core Software Services, Wellcome Sanger Institute, Hinxton, Cambridge, UK; [6]Hospital de la Santa Creu I Sant Pau, Barcelona, Spain; [7]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, UK; [8]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan and [9]Fondazione IRCCS Policlinico San Matteo and University of Pavia, Pavia, Italy

Correspondence: GEORGE VASSILIOU - gsv20@sanger.ac.uk
MS VIJAYABASKAR - vm11@sanger.ac.uk

doi:10.3324/haematol.2019.230821

Supplementary Appendix

Contents

List of Figures

List of Tables

1 Algorithm Design

1.1 Transcript Indexing

To boost the alignment speed, transcript sequences of the 33 clinically relevant genes (Table 1) to AML diagnosis were indexed prior to read alignment. Transcript sequences of GRCh38 v93 were downloaded from the Ensembl database [1]. Non-coding transcripts and the ones without Consensus CDS annotations were excluded. For multiple protein-coding transcripts that only differ in untranslated regions (UTRs), only the one with longest UTR was retained. Along each transcript (including both sense and antisense sequences), a sliding window of k-mers were used to compute the hash function that maps each k-mer sequence to the isoform(s) and locus/loci that the k-mer belongs to (Supplementary Figure 1A). The hash function allows for fast retrieval of the genic loci for any given k-mer sequence with a time complexity of *O(1)*. The optimal k-mer size $k = 10$ was chosen, considering the balance between memory usage and alignment speed (Supplementary Table 1).

***k*-mer Size**	**Probability of Match by Chance**	**Minimum RAM (MB)**	**Alignment Speed (sec/million reads)**
9	0.64	75	18.3
10	0.16	279	8.6
11	0.040	1022	7.2
12	0.010	3362	7.0

Supplementary Table 1: Performance of index using different *k*-mer size. Memory usage was calculated on a 64-bit operating system. Alignment speed were tested on a laptop with Intel Core i7 1.8-4.0 GHz Processor and solid state hard drive and the average speed of 151 samples was shown. RAM = Random access memory.

1.2 RNA-seq alignment

Single-end and paired-end reads were aligned with slight differences. Prior to alignment, unknown nucleotides (N in FASTQ files) were trimmed from 5' and 3' ends of sequenced reads. Trimmed reads that are shorter than 40 bp were discarded. Each trimmed read was divided into consecutive k-mers. If the read length is not a multiple of k-mer length, an overlapping k-mer is added to the 3' end. Using the pre-built index, the transcript location(s) of each k-mer were retrieved and assessed. A reads is considered unaligned if

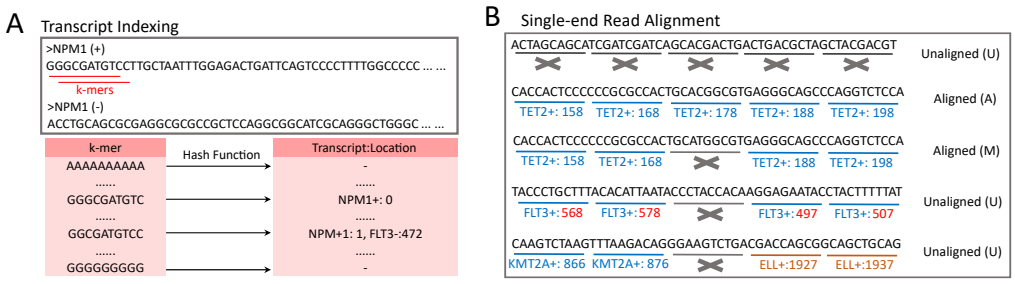
fewer than N of the k-mers are mapped to the same transcript, where:

$$N = Max(2, [\frac{1}{2} \cdot \frac{L}{k}]) \tag{1}$$

in which L is the length of the read and k is k-mer length which means that at least half of the k-mers must be mapped to the same transcript (Supplementary Figure 1). Poly-A or poly-T k-mers were considered as mapped to the end of the transcript. In addition, every pair of mapped k-mers must be in the correct order within the tolerated range of 30 bp of insertion or deletion:

$$|(T_i - T_j) - (R_i - R_j)| \leqslant 10 \quad \forall\, i, j \; and \; i \neq j \tag{2}$$

where i and j are a pair of k-mers, T is the location of the k-mer on the transcript and R is its location on the read. Reads that satisfy both criteria were considered as aligned. An additional requirement must be met for paired-end alignment - both 5' and 3' reads must be aligned to the same transcript and the outer distance (i.e. fragment length) must be less or equal to 1000 bp, which corresponds to the maximum DNA length in a typical sequencing library.



Supplementary Figure 1: Transcript indexing. (A) Index was built by constructing the hash function that maps every k-mer within the transcript sequences of the 33 target genes, including both sense (+) and antisense (-) strands, to its location on the transcript. It is possible that a k-mer maps to multiple locations. **(B)** Examples of read alignment by k-mer mapping. Reads were classified into unaligned (type U), aligned with mismatches (type M) and perfectly aligned (type A).

1.3 Detection of substitutions and small indels

From reads aligned to transcripts, the ones with imperfect alignment (i.e. containing unmapped k-mer(s)) were selected for detection of substitutions

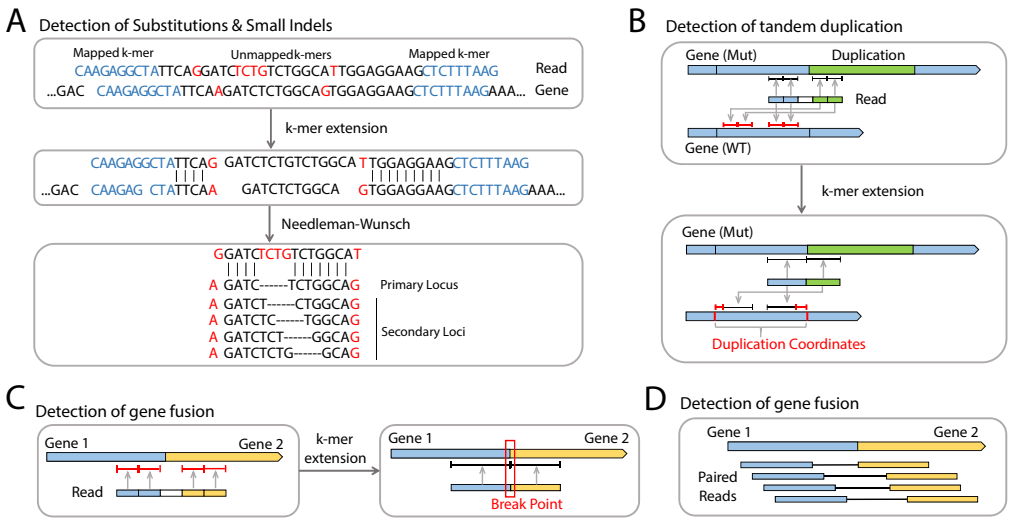
and small indels. For each unmapped region, the pair of mapped k-mer flanking this region were used as anchors and matching bases between the read and transcript were extended from the anchors into the unmapped region, resulting in N_T bases remaining on the transcript and N_R bases remaining on the read. If the unmapped region is located at the 5' or 3' end of read, the extending process was done from one anchor only. The type of the mutation could be identified by comparing N_T and N_R:

$$\begin{cases} N_T = 0 \; and \; N_R > 1 & Insertion \\ N_T > 1 \; and \; N_R = 0 & Deletion \\ N_T = 1 \; and \; N_R = 1 & Substitution \\ N_T > 1 \; and \; N_R > 1 & Multiple \end{cases} \tag{3}$$

For the case that more than 1 bases are remaining in both the transcript and the read, indicating the possibility of multiple mutations, the Needleman-Wunsch algorithm (match score = 10, mismatch score = -8, gap open penalty = -9, gap extension penalty = -2) was applied to find out the location mutations (Supplementary Figure 2A). Reads that contain $\geqslant$ 12 mismatches or $\geqslant$ 3 indels were discarded as unaligned. The locations of mutations on transcripts were converted to genomic coordinates. If multiple genomic coordinates can represent the same mutation, the one with the lowest coordinate was used as the primary locus (Supplementary Figure 2A). A number of criteria must be satisfied for mutation calling. The RNA-seq quality (Phred Score) of the mutated base of substitutions, the average score of inserted bases, or the average score of the two bases flanking the deletion site must be $\geqslant$ 20 (i.e. error $\leqslant$ 0.01). At least 5 reads with unique sequences have to be covering the mutation site and the variant allele frequency (VAF), which is defined as:

$$VAF = \frac{N_{mut}}{N_{mut} + N_{wt}} \tag{4}$$

where N_{mut} and N_{wt} are the number of mutated and normal reads respectively, must be $\geqslant$ 0.05 for the mutation to be called in the initial round. However, mutation calling was not attempted for mismatches within the first and last 10 bp of reads to avoid false discovery because for example, it is impossible to distinguish whether a 2-bp mismatch at the beginning of a read is due to an insertion or two substitutions. Instead, these reads were retained for realignment and VAF correction.



Supplementary Figure 2: Strategies for detection of various types of mutations. **(A)** Substitutions and small indels, **(B)** tandem duplications, **(C)** gene fusion using RNA-seq reads that span the breakpoint or chimeric reads, and **(D)** gene fusion from paired-end reads that align to each of the fusion partners.

1.4 Detection of tandem duplication

From the pool of unaligned reads, the ones with both 5'- and 3'-end k-mers mapped to the same transcript and the 3' k-mer is mapped upstream of the 5' k-mer, were flagged for potential tandem duplication. To reduce false positive rate, at least one of the 5' or 3' end must contain $\geqslant 2$ consecutively mapped k-mers for the call of tandem duplication to be attempted. From the mapped 5' k-mer(s) towards downstream and 3' k-mer(s) towards upstream, each matching bases between the read and transcript was used to extend the matched region. The extension process terminates if either of the two conditions is met. Firstly, if all bases in the read were covered, then the coordinates of the duplicated region on the transcript could be obtained by the first and last matched base on the transcript (Supplementary Figure 2B). Secondly, if both 5' and 3' extension reached a mismatching base and nucleotides remained in the read, then this indicated that the remaining nucleotides were inserted between the two duplicated regions. At least 5 reads with unique sequences must cover a putative duplication site and the VAF must be >0.05 for the tandem duplication to be called initially. Unaligned reads containing at least 2 consecutive mapped k-mers only at one end were retained for realignment and VAF correction.

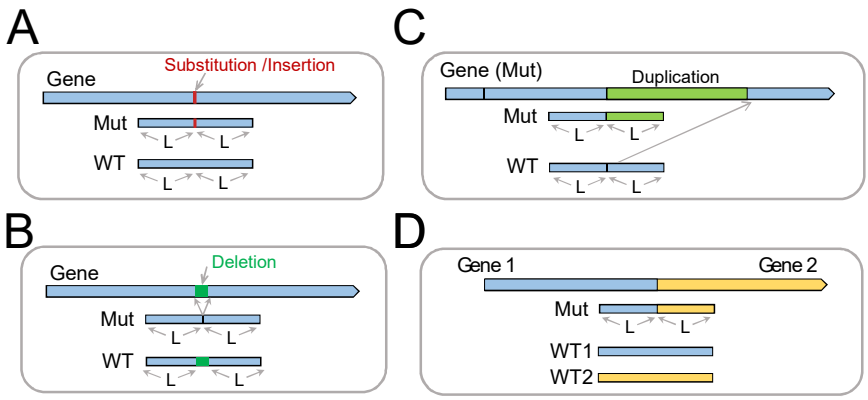
1.5 Detection of gene fusion

Reads whose 5' and 3' k-mers were mapped to each of the partners in a fusion pair were flagged as putative chimeric reads. At least one of the 5' or 3' end must contain $\geqslant 2$ consecutively mapped k-mers. Reads with only one mapped k-mer at both 5' and 3' ends were retained for realignment. Matching bases to the corresponding transcript were extended from both ends until no bases remain in the read. Read with $\geqslant 1$ bases remaining were discarded. The coordinate of the breakpoint on each transcript was obtained from the ends of the extension of 5' and 3' k-mers (Supplementary Figure 2C). At least 3 reads with unique sequences must cover the breakpoint for a fusion to be called. For paired-end RNA-seq specifically, an independent strategy was used by extracting the pairs in which each read were aligned to one of the fusion partners (Supplementary Figure 2D). However, paired-end read spanning the fusion partners do not reveal the coordinates of the breakpoint.

1.6 Realignment and VAF correction

Realignment improves the VAF calculation by taking previously discarded or ignored reads, which contain true positives, and aligning them to the wild-type (WT) and mutated sequences around the mutation site. Realignments were attempted for each of the substitutions, small indels, tandem duplications and gene fusions detected from chimeric reads but not for gene fusions detected by paired-end reads. The maximum read length L was obtained during the read alignment step. For substitutions and small indels, mutated and WT sequences were constructed from the subsequence $\pm L$ bp around mutation spot, including and excluding the mutated region respectively (Supplementary Figure 3A, B). For tandem duplication, the mutated sequence was constructed by joining L bp at the end and L bp at the beginning of the duplicated region while the WT was constructed by $\pm L$ bp around the end of the duplicated region (Supplementary Figure 3C). For gene fusion, mutated sequence was constructed by joining L bp of the 5' transcript upstream of the breakpoint and L bp of the 3' transcript downstream of the breakpoint. Two WT sequences were constructed for gene fusion, each composed of $\pm L$ bp around the breakpoint of their transcript (Supplementary Figure 3D). For each mutation, a pair of new indices were built for the mutated sequence and WT sequence(s) using the same algorithm as transcript indexing (Supplementary Figure 1A). The new indices were used to realign the reads that were retained for realignment (explained in previous sections).

Realignment tolerates $\leqslant 2$ mismatches, no insertion or deletion and no mismatch within the mutated spot for both WT and mutated sequences. A read will be marked as mutated or WT if it is exclusively aligned to the mutated sequence or to the WT sequence respectively. The VAF values were updated using the new N_{mut} and N_{wt} after realignment.



Supplementary Figure 3: Construction of mutated and *WT* sequence for realignment. L denotes the maximum length of RNA-seq reads. **(A)**Substitution and small insertion, **(B)** small deletion. **(C)** tandem duplication and **(D)** gene fusion.

1.7 Flagging sequencing artefact

Insertions and deletions that are single-nucleotide long or consisting of homopolymers were check for possible sequencing artefacts. The sequence surrounding the insertion or deletion sites were extracted. If the surrounding sequence consists of $\geqslant$4 nucleotides that are the same as the inserted or deleted nucleotide(s), it will be flagged as a potential sequencing artefact. For example, an insertion of A in a regions of AAAAAA will be flagged as artefact. Sequencing artefacts will be reported but not checked for oncogenicity unless explicitly stated in the oncogenicity filter.

1.8 Oncogenicity filter

After obtaining the mutations in an RNA-seq dataset, the next step was to identify the oncogenic ones that are potential drivers and remove irrelevant

mutations such as single nucleotide polymorphisms (SNPs) and benign mutations. We used the same criteria used by Papaemmanuil *et al.* 2016 [2], which select for known hotspots, recurrent mutations in public databases [3, 4] and mutations in functional domains. Gene fusions, *FLT3*-ITD and *MLL*-PTD were always retained. The criteria are summarized in Supplementary Table 2.

Gene	Included Mutations
NPM1	W288fs, W290fs
IDH1	R132
IDH2	R140, R172
FLT3	D835, D839, Y842C, N841,R834Q, V592, Y572,
DNMT3A	R882, Frameshift, Stop-codon gain, F909, P904, W893, Q886, N879, E863D, P849L, Q842H, T835M, K829, R803, N797K, W795C, R792H, P777R, E774D, L773, R771Q, S770, F752, F751L, R749C, R736, R729, P718R, V716D, S714, L713F, G707, I705T, D702V, G699D, D686, S669F, A662G, R635Q, W581, L547, G543, C497, K468
TP53	Frameshift, Stop-codon gain, R290H, E286, R283H, D281N, P278S, A276P, C275, R273, V272M, R267, G266R, L265P, G262V, E258A, I254V, R248, G245D, C242S, S241, N239D, C238Y, M237, Y234, P223S, Y220, S215, Y205D, V203E, P196Q, I195, H193, H179R, C176, R175H, V173L, H168P, K164E, Y163C, G154D, V143M, T140N, F134I, K132R, L130V, R110L, F109C, F54L, E11K
CEBPA	Frameshift, Stop-codon gain, A.A.Insertion 300-302, L338P, N321, V314A, Q312K, Q311K, E309, R306P, A303P, D301, R300, R297, A295E, N293S,
TET2	Frameshift, Stop-codon gain, R1896T, T1884A, H1881R, R1868Y, A1512V, R1467K, Q1445R, V1417F, H1417R, H1380, G1370E, F1368V, R1359C, L1322Q, C1298W, G1282C, C1273, A1264, R1262, R1261C, C1221, N1102, H949R, A665D, S460F, E154V
RUNX1	Frameshift, Stop-codon gain, M267I, N260K, R250C, S226, R207P, R204Q, R201, P200S, D198, K171N, G168R, S167N, R166, G165, R162, D160Y, A149, S141, P113, R107, S100F, S94I
WT1	Frameshift, Stop-codon gain,H465, D464, R462, R458P, R434, R370P, R369G, G351R
BCOR	Frameshift, Stop-codon gain,L1550, R1131L
ASXL1	Frameshift, Stop-codon gain, G646fs, K85R, P511S, A530V, A772T, R786K, T787N, E801, V1060D
U2AF1	Frameshift, Stop-codon gain, R188H, Q157, R156, R35, S34, R28
SRSF2	P95, A.A.Deletion 90-110, F57Y, Y44H
SF3B1	K700, K666, Frameshift, Stop-codon gain, A.A.Deletion 690-710, D799G, D781G, E776D, R775L, A749T, G742D, G740, I704N, V701F, A672V, H662, N626, R625, E622, S611F, G605D,

Supplementary Table 2: Selection criteria for oncogenic mutations. Known hotspots are colored in red.

2 Supplementary Methods

2.1 Data acquisition

BAM files of the RNA-seq data of the 151 AML samples were downloaded from the TCGA portal [5]. FASTQ files of the 437 RNA-seq data from the Leucegene datasets [6, 7, 8, 9] were downloaded from Gene Expression Omnibus [10] using the fastq-dump.2 in the SRA Toolkits [11]. RNA-seq data and genotypes of the MDS cohort were obtained from the Ogawa and Cazzola groups (unpublished data). Gene annotation, coding/non-coding transcripts and CCDS sequences of the human assembly GRCh38 version 93 were downloaded from Ensembl database [1].

2.2 Bioinformatics analyses

RNA-seq data were mapped to the GRCh38 version 93 by STAR v2.7.0d [12] using parameters `--outFilterMismatchNoverReadLmax 0.05 --alignIntronMax 500000`. Alignments were visualized by Interactive Genomics Viewer IGV [13].

For benchmarking read alignment, sequences of the 33 genes relevant to AML diagnosis were indexed using BWA 0.7-17 [14] and RNA-seq samples were aligned using BWA-MEM under default parameters. Reads with BWA score higher than 95 (out of 100) were marked as aligned. The sequences were also indexed by Salmon v0.13.1 [15] using -k 31 and aligned using default parameters.

For quantification of Variant Allele Frequencies (VAFs), the pileup function of Samtools 1.9 [16] was used to calculate VAFs of substitutions, allowing only bases with Phred33 score >20. Varscan v2.4.3 [17] was used with Samtools mpileup to identify and calculate the VAFs of small indels and substitutions, using default parameters. To detect gene fusions, a fusion index was first built using FuSeq [18] with the default SQLite database of GRCh37 version 75. FuSeq was run under default parameters to identify gene fusions.

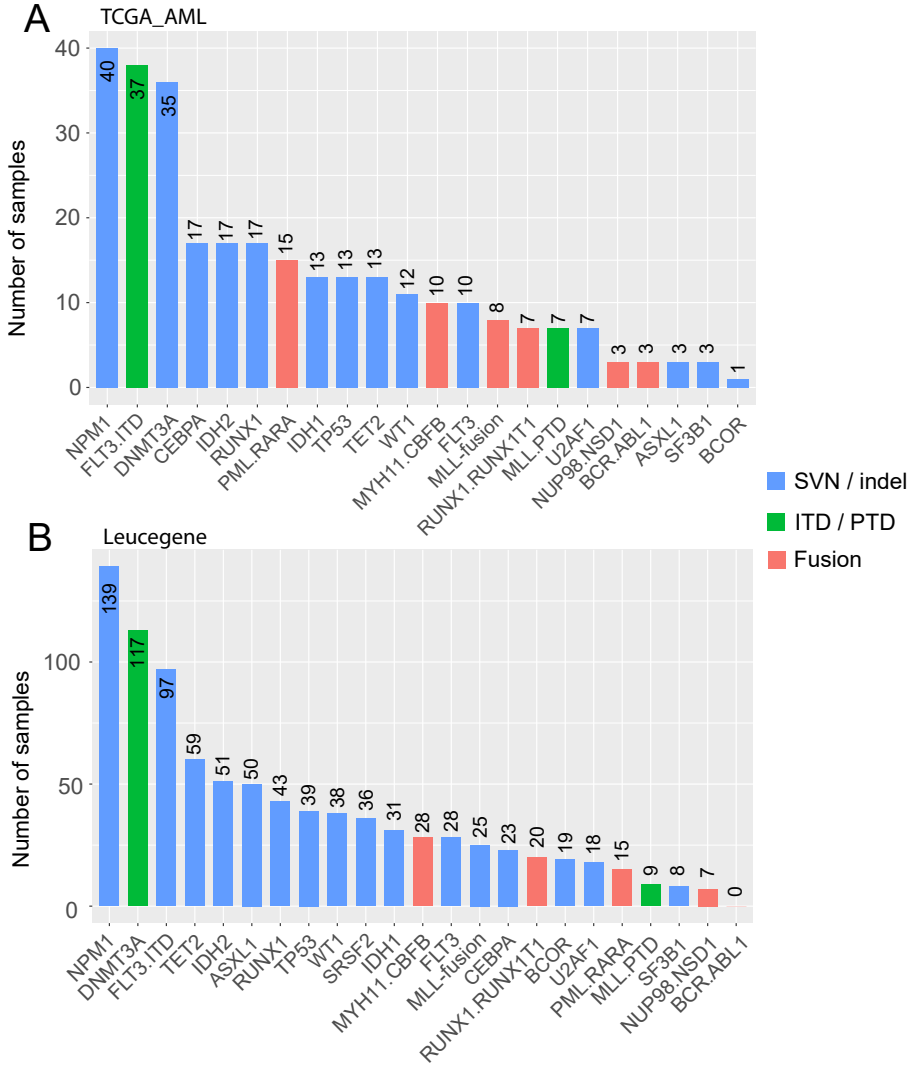
3 Supplementary Results

3.1 Mutations detected by our software

To test the performance our software, we acquired the RNA-seq data from three independent cohorts − the TCGA AML cohort of 151 patient samples, the Leucegene AML datasets of 437 patients [8, 6, 9, 7], and an MDS cohort of 164 patients [19].

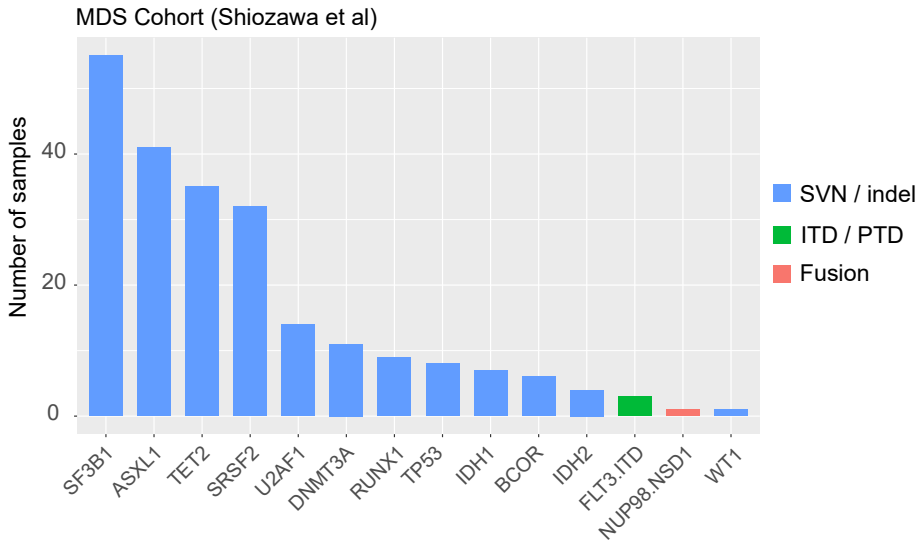
We performed analyses on these 3 cohorts with the 33-gene panel and called mutations within these genes. Mutation landscapes are summarized in this section.

3.1.1 TCGA and Leucegene show similar distributions of mutations



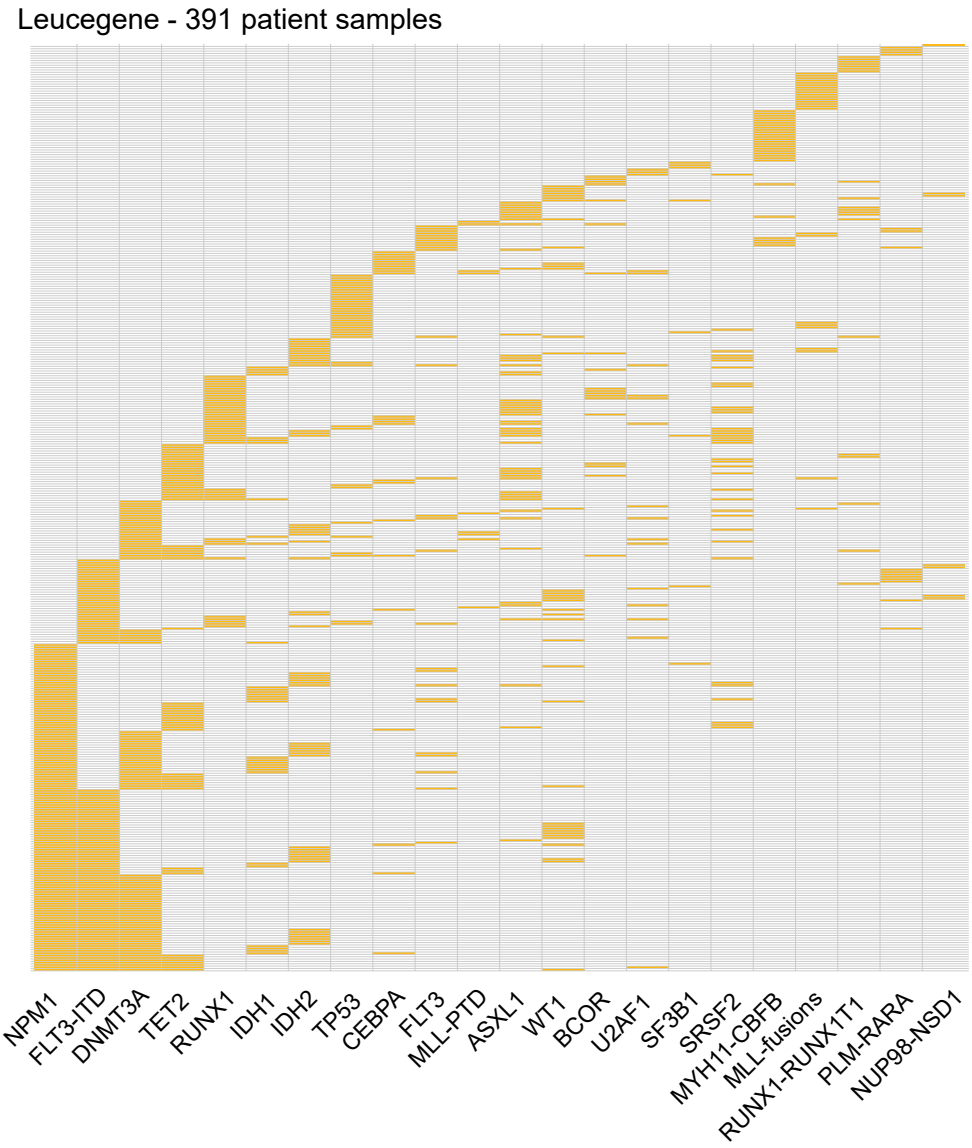
Supplementary Figure 4: Mutations detected by our software in TCGA and Leucegene datasets. Our software identified similar distributions of mutations in **(A)** the AML cohort of TCGA and **(B)** the Leucegene datases.

3.1.2 Number of mutations in the MDS cohort



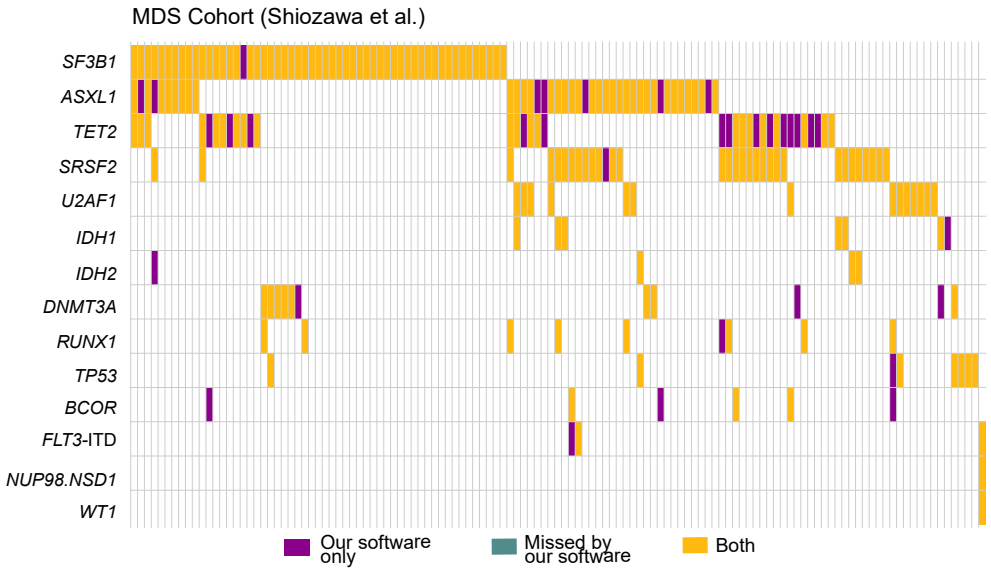
Supplementary Figure 5: Mutations detected by our software in the MDS dataset. For many gene fusions, no mutation were detected in the MDS cohord and are therefore not plotted.

3.1.3 Mutational landscape in Leucegene datasets



Supplementary Figure 6: Landscape of mutations in the Leucegene datasets. Only 391 of the 437 samples showing at least one mutations in the tested genes are shown.

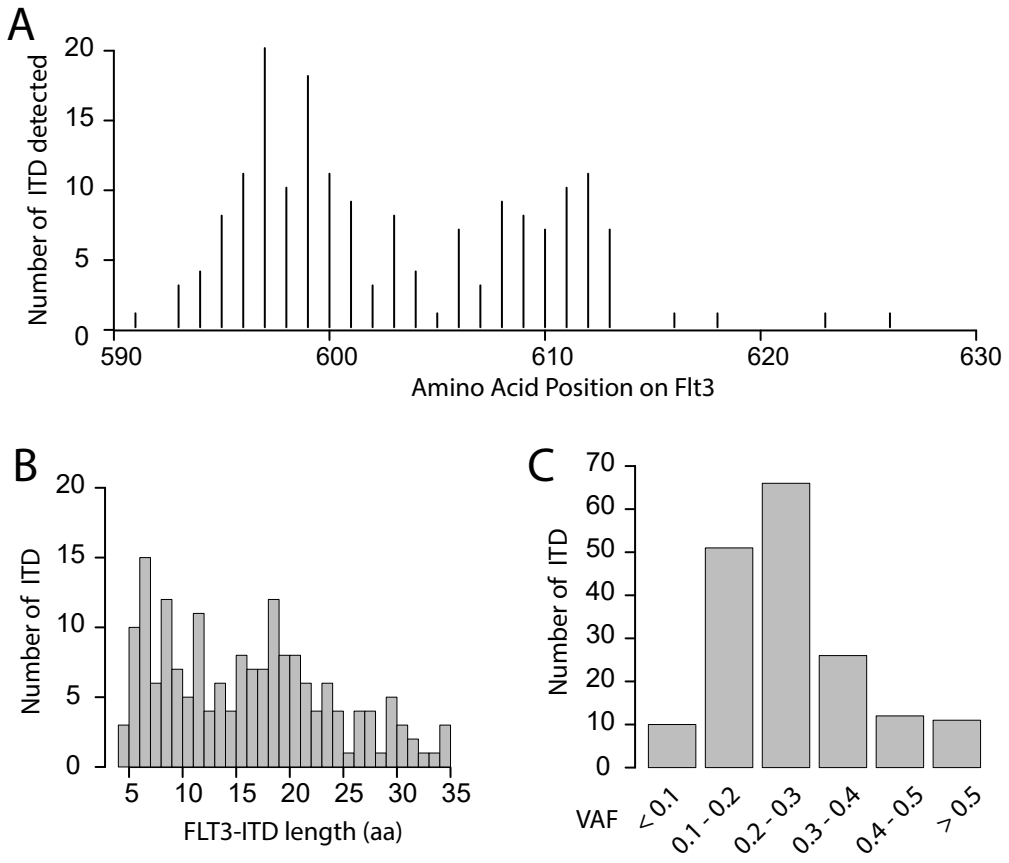
3.1.4 Mutational landscape in the MDS dataset



Supplementary Figure 7: Landscape of mutations in the MDS dataset. Mutations detected by both our software and by Shiozawa *et al.* are depicted in yellow, additional mutations detected only by our software in purple. Our software detected all annotated mutations. Details of the mutations in individual samples are given in Supplementary Data.

3.1.5 Summary of detected *FLT3*-ITDs in TCGA

All the *FLT3*-ITD detected by our software are located between amino acid 590 and 630 (Supplementary Figure 8A) with length from 4 amino acid upto 35 (Supplementary Figure 8B), which ares consistent with COSMIC database. Our software also reports the allelic frequency of ITDs (Supplementary Figure 8C), which could be useful for prognostic prediction.



Supplementary Figure 8: Summary of detected *FLT3*-ITDs. (A) Number of *FLT3*-ITD detected at each amino-acid position. **(B)** Distribution of lengths of ITDs. **(C)** Distribution of VAFs (indication of ITD-to-WT ratio).

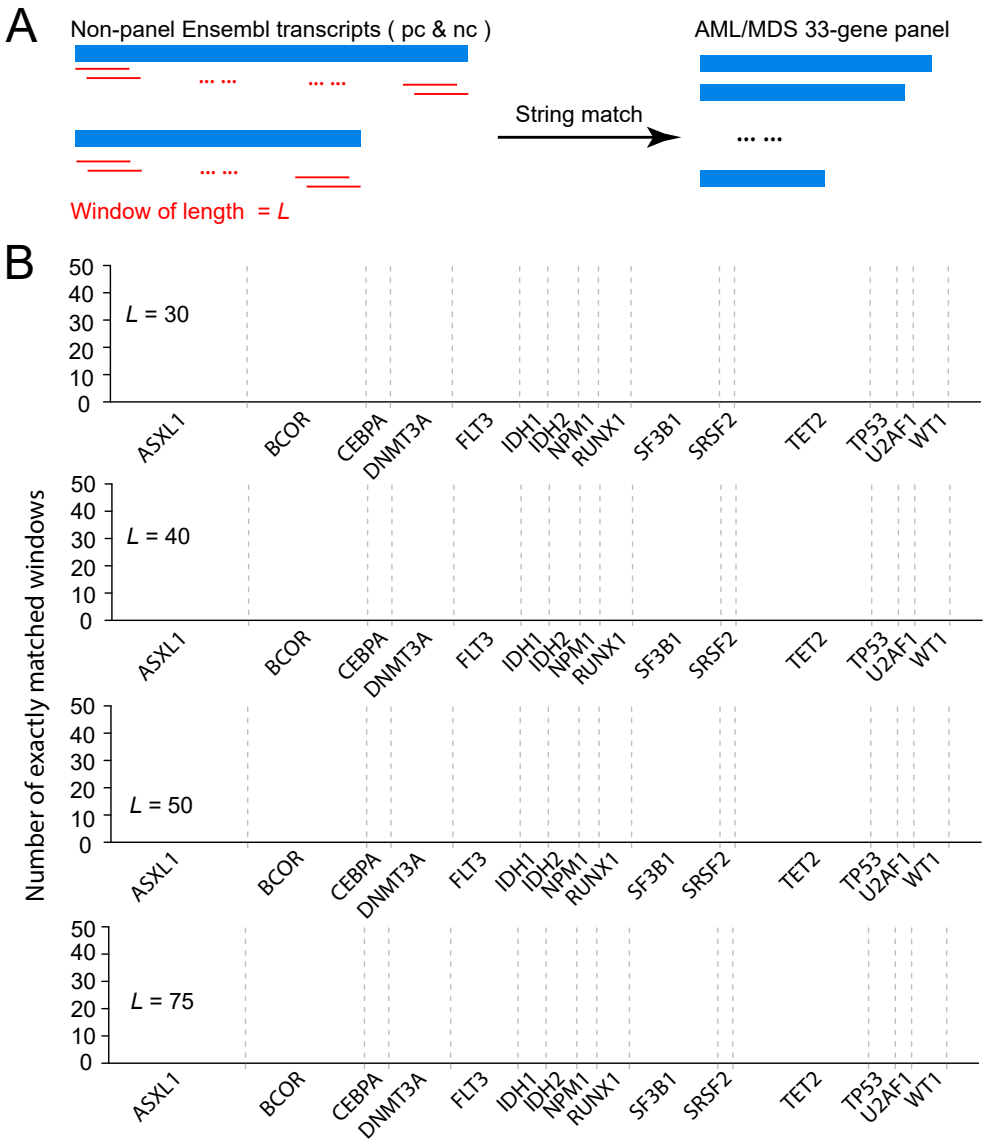
3.2 Check for multiple mapping

3.2.1 Check for multiple mapping by exact match

Since our algorithm only focuses on 33 genes and ignores the rest of the transcriptome, it is essential to confirm that RNA-seq reads produced from the transcripts outside the 33 genes do not fortuitously align to our panel genes. To test for this, we first generated the sequences of every L-bp windows (for $L = 30$, 40, 50 and 75) on all the Ensembl coding and non-coding transcripts other than these 33 genes (Supplementary Figure 9A). Genes that are antisense to panel genes, such as *MFSD11*, *NDE1* and *CDC42SE1* were excluded. We carried out string-matches for 30, 40, 50 and 75-bp windows against the coding regions of 33 panel genes (i.e. excluding UTRs). For all sizes of windows, we observed no sequence identity to the coding regions of our 33-gene panel (Supplementary Figure 9B).

Sequence identity was observed in the UTRs of *IDH2* and *SF3B1* for 30-mers (unpublished data), which is due to repetitive sequences in the UTRs. However, mutations in UTRs do not cause changes in protein product and no recurrent diagnostically or prognostically important UTR mutations have been reported in AML or MDS.

It is worth mentioning that the exact match does not simulate the RNA-mut algorithm. Firstly, it only takes single-end reads as input, which are much more likely to be mistakenly aligned than paired-end reads and hence an overestimation of potential errors. On the other hand, this method tolerates no mismatches, which is an underestimation of real error rate.

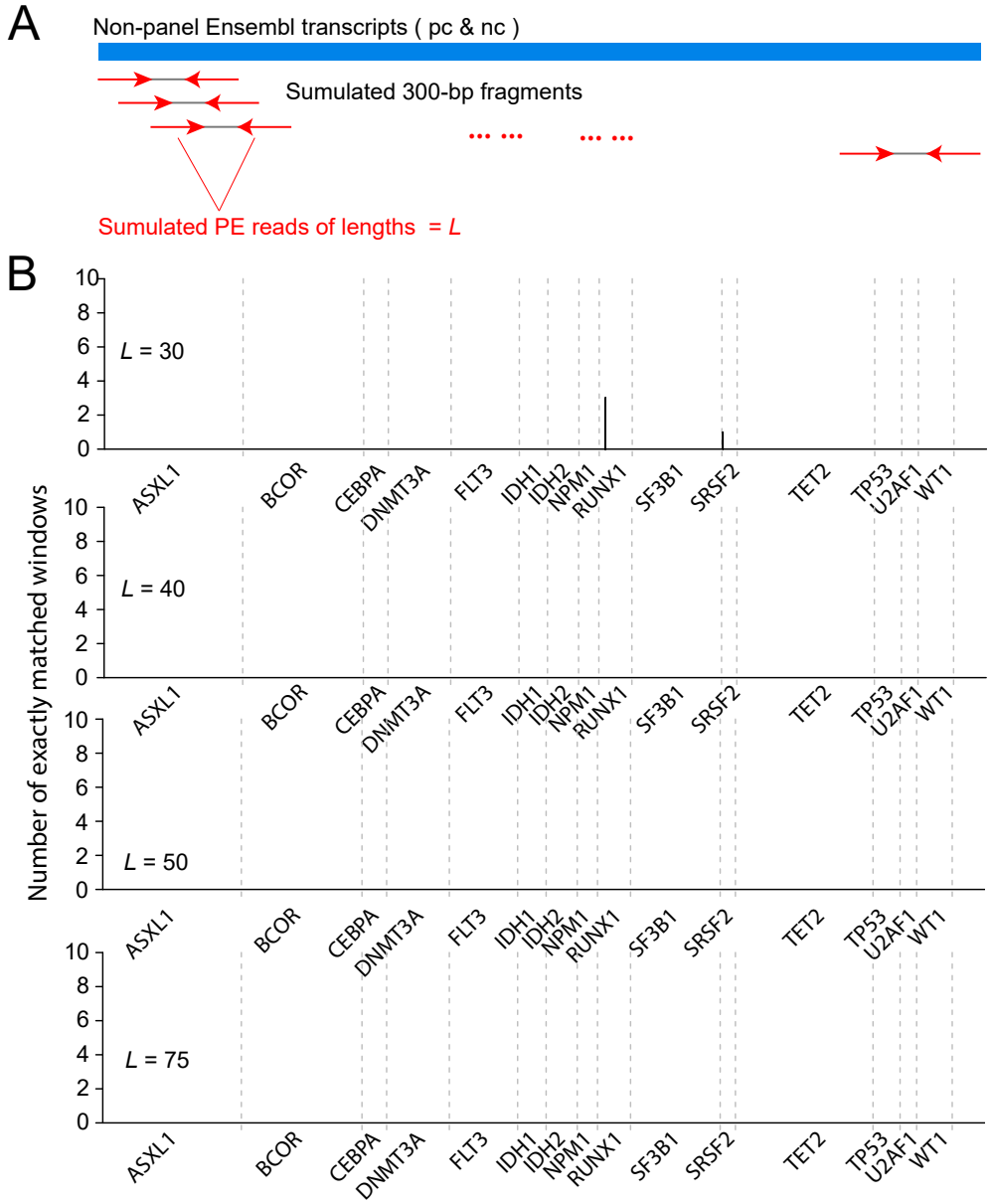


Supplementary Figure 9: Uniqueness of mapping for the subsequences of panel genes. **(A)** Ensembl transcripts other than the 33 panel genes were obtained. Subsequences were generated using a sliding window of 30, 40, 50 and 75 bps moving by 1 bp at a time. **(B)** Each window was searched against the coding sequence of the 33 panel genes and the number of matches at each genic position were plotted.

3.2.2 Check for multiple mapping by simulation

To better simulate real RNA-seq data, we took all Ensembl transcripts other than the 33 panel genes and also excluding antisense gene *MFSD11*, *NDE1* and *CDC42SE1*. We first generated 300-bp fragments (i.e. typical size of sequencing libraries) by sliding 1 bp at a time on the transcript. For each fragment, we produced a pair of reads of a L bp from the two ends of the fragment. This process was repeated with $L = 30$, 40, 50 and 75, producing 4 sets of simulated paired-end reads. Each set of simulated reads were passed to the RNAmut alignment algorithm, using the default parameters.

For 30 bp, we observed reads from 1-3 non-panel genes mistakenly aligned to *SRSF2* and *RUNX1*, whereas for 40, 50 and 75-bp simulation, no reads were aligned to the coding regions of panel genes. This indicates that a minimum threshold of 40 bp of paired-end reads after end-clipping could be a good choice of parameter.

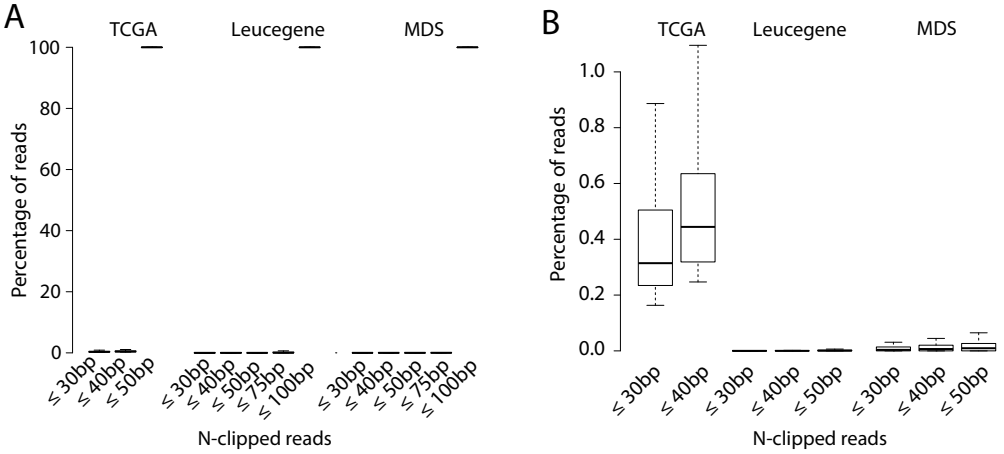


Supplementary Figure 10: Estimation of mapping errors by simulation. **(A)** Generation of simulated reads. Protein-coding and non-coding transcripts from the Ensembl database (other than the panel genes) were used. 300-bp fragments were first generated using sliding windows and then paired-end reads were generated from the two ends of the fragments. **(B)** Simulated reads were aligned using RNAmut algorithm with default parameters. Numbers of simulated reads aligned to the panel genes were plotted against coding-region positions.

3.2.3 Distribution of read lengths

To establish the lengths or end-clipped reads in real data, we examined the raw sequencing reads in TCGA, Leucegene and MDS cohorts. TCGA datasets was sequenced before 2013 and the read lengths are typically 50 bp whereas the other two cohorts are more recent and both of which are 100 bp. The unknown nucleoties (N) were clipped from both ends of reads and the lengths of the remaining reads were examined.

We observed that less than 1 percent of the reads were shorter than 40 bp in the TCGA dataset, suggesting very little impact on the mutation-calling results due to the small fraction (Supplementary Figure 11A, B). For modern sequencing data, the effect is much less significant or non-existent.



Supplementary Figure 11: Distribution of end-clipped read lengths. (A) Distribution of the percentages or reads within 30, 40, 50, 75 and 100 bp in each dataset. **(B)** Same as (A) but zoomed in on the low values.

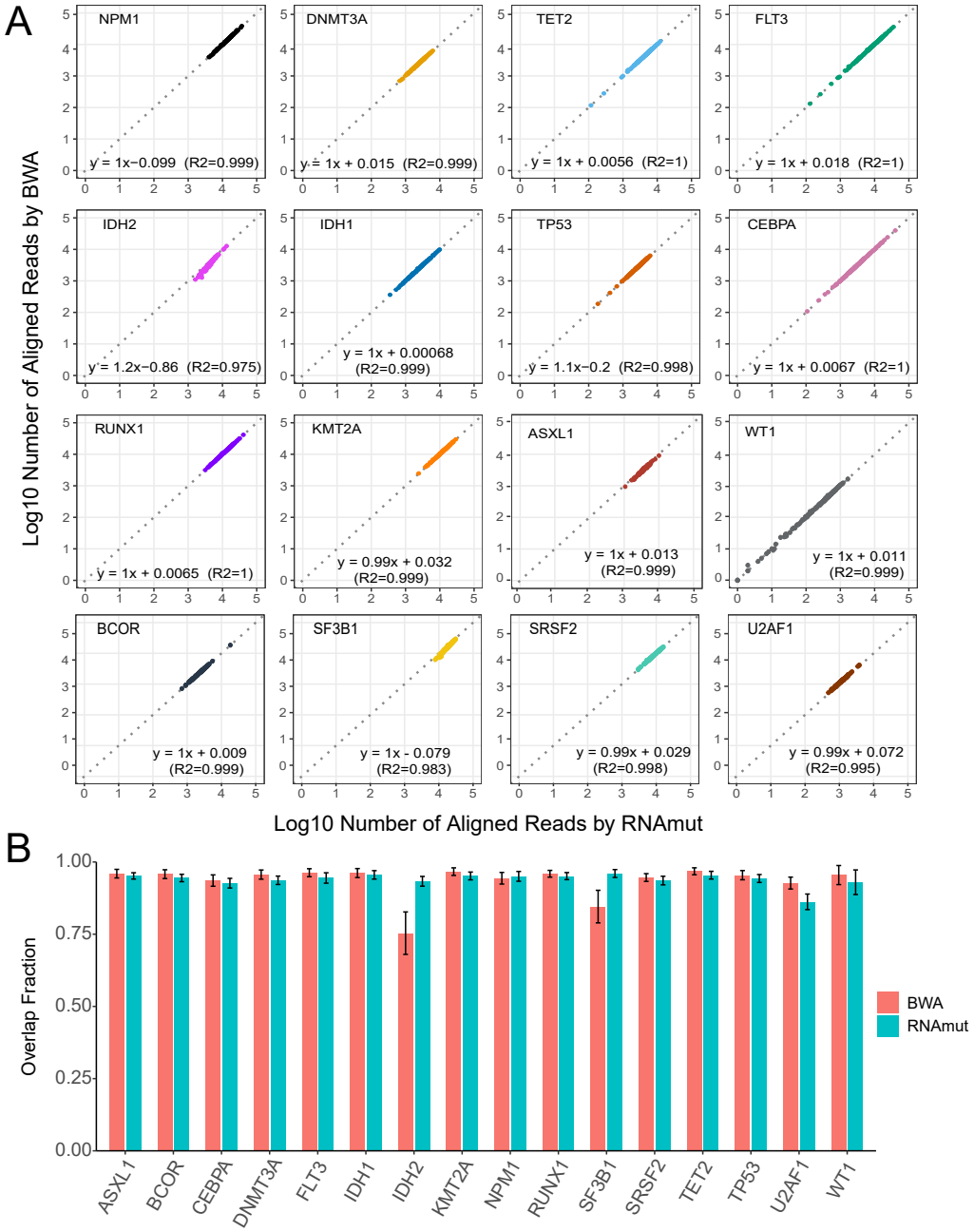
3.3 Software benchmarks

To validate each step of our pipeline, we compared our results with existing bioinformatic tools. Read alignment was compared with the commonly used aligners BWA [14], STAR [12] and Salmon [15]. For the purpose of benchmarking, we chose BWA and Salmon for the panel-gene alignment (instead of genome alignment), for closest resemblance to the RNAmut algorithm. Furthermore, we also benchmarked the alignment of RNAmut to the whole-genome alignment by STAR.

RNAmut's quantification of single nucleotide variants (SNVs) was compared with Samtools [16] and Varscan [17], indel detection was compared with Varscan and gene fusion was compared with FuSeq [18].

3.3.1 Comparison between our software and BWA

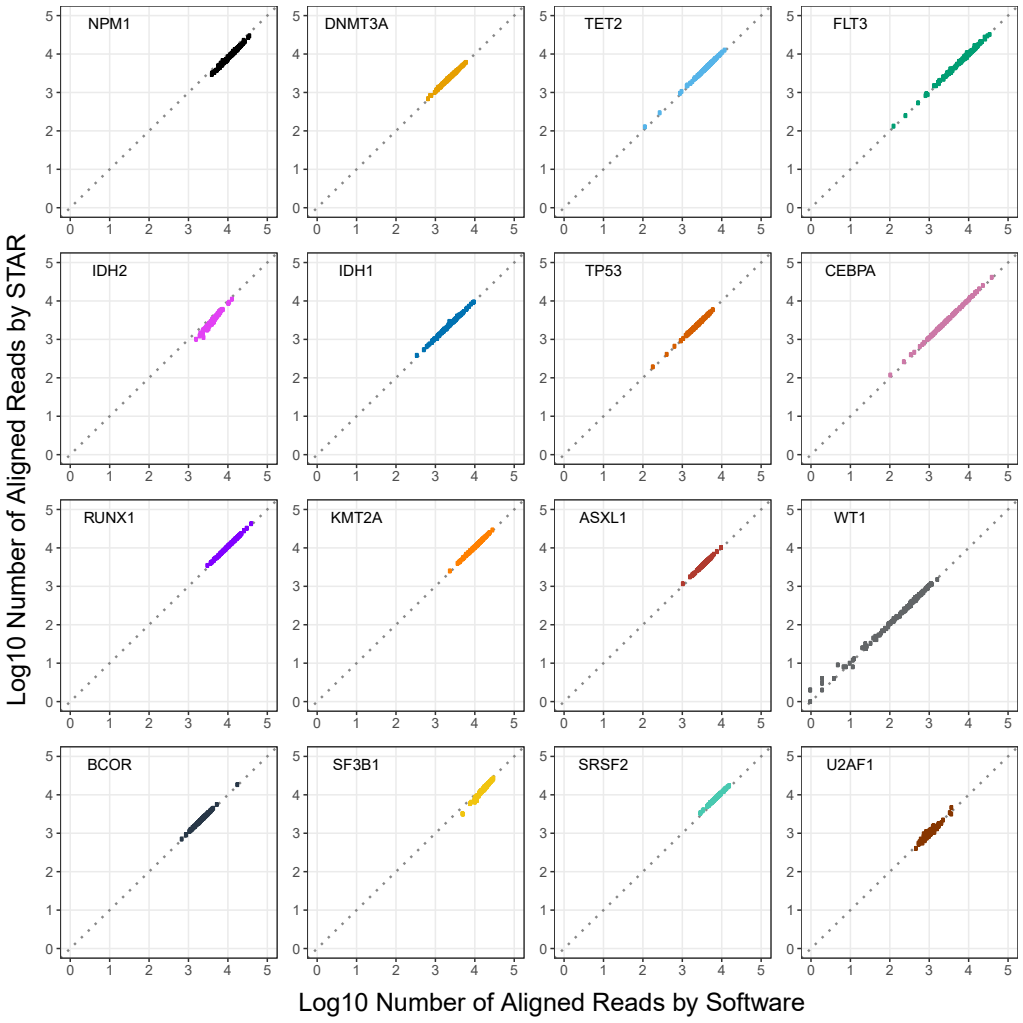
Since our alignment is based on transcript sequence, the closest resemblance is to the non-spliced aligner BWA. We constructed BWA index using transcript sequences of 33 panel genes and aligned the RNA-seq reads to the transcripts. The number of reads aligned by our software show very good agreement with BWA (Supplementary Figure 12A). The common set of reads aligned by both our software and BWA comprises approximately 95% of the reads aligned by any one software (Supplementary Figure 12B), with the exception of *IDH2* where the common set is ∼75% of the reads aligned by BWA, which is due to our software aligned more reads than BWA in *IDH2*.



Supplementary Figure 12: Comparison between read alignment by our software and BWA. (A) Scatter plot of number of reads aligned to each gene by our software versus those aligned by BWA for the 151 RNA-seq samples in TCGA AML. Genes that are fusion partners were not included. **(B)** Fractions of reads aligned by both (i.e. overlap) in reads aligned by BWA or our software. Bars represent the mean of fractions of 151 samples and error bars show the standard deviation.

3.3.2 Comparison between our software and STAR

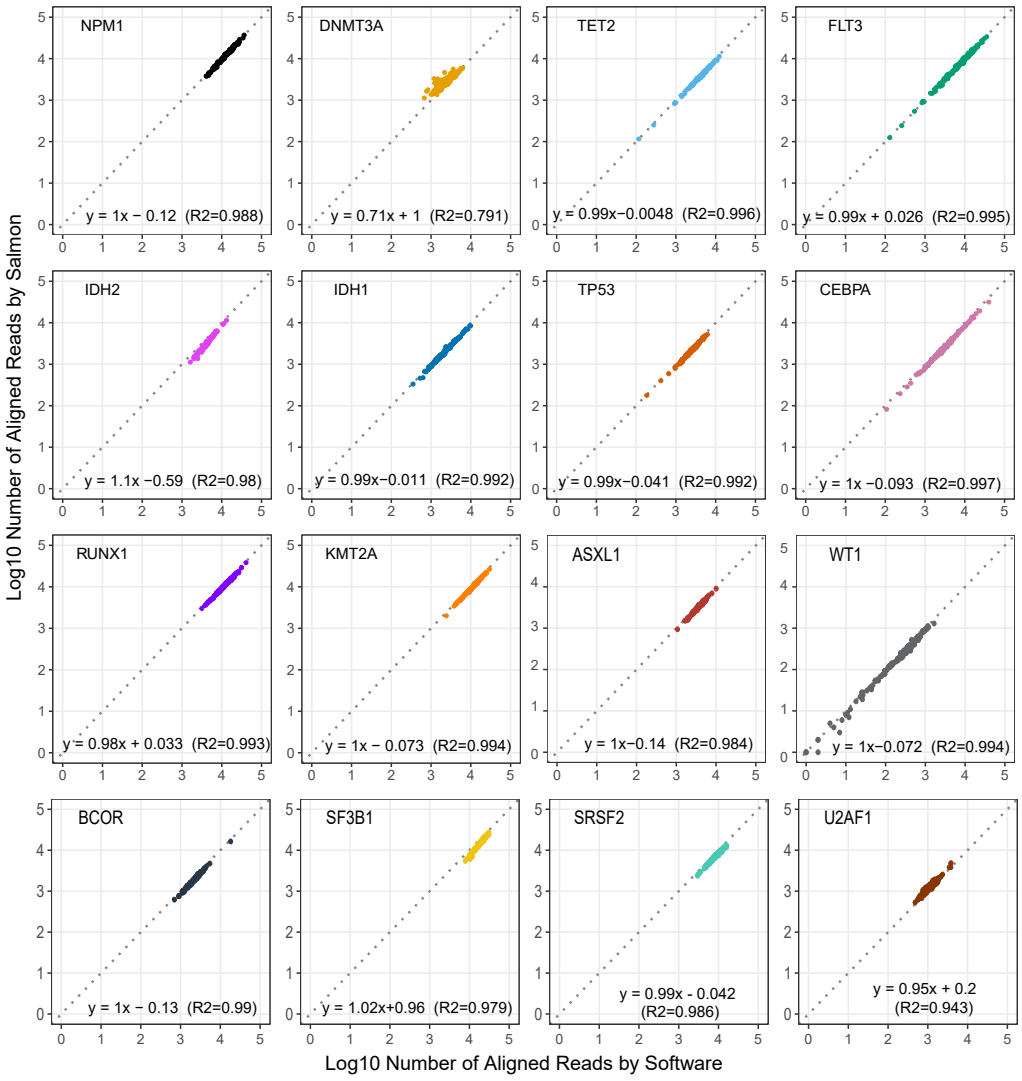
The spliced aligner STAR is the standard aligner for RNA-seq data. To perform STAR alignments, we first aligned RNA-seq reads to the entire genome. Reads aligned to the panel genes were extracted using Samtools. Reads aligned to the intronic regions were removed with customized scripts. We also observed very good correlations between the number of reads aligned by our software and STAR (Supplementary Figure 13).



Supplementary Figure 13: Comparison between read alignment by our software and STAR. Similar to Supplementary Figure 12A.

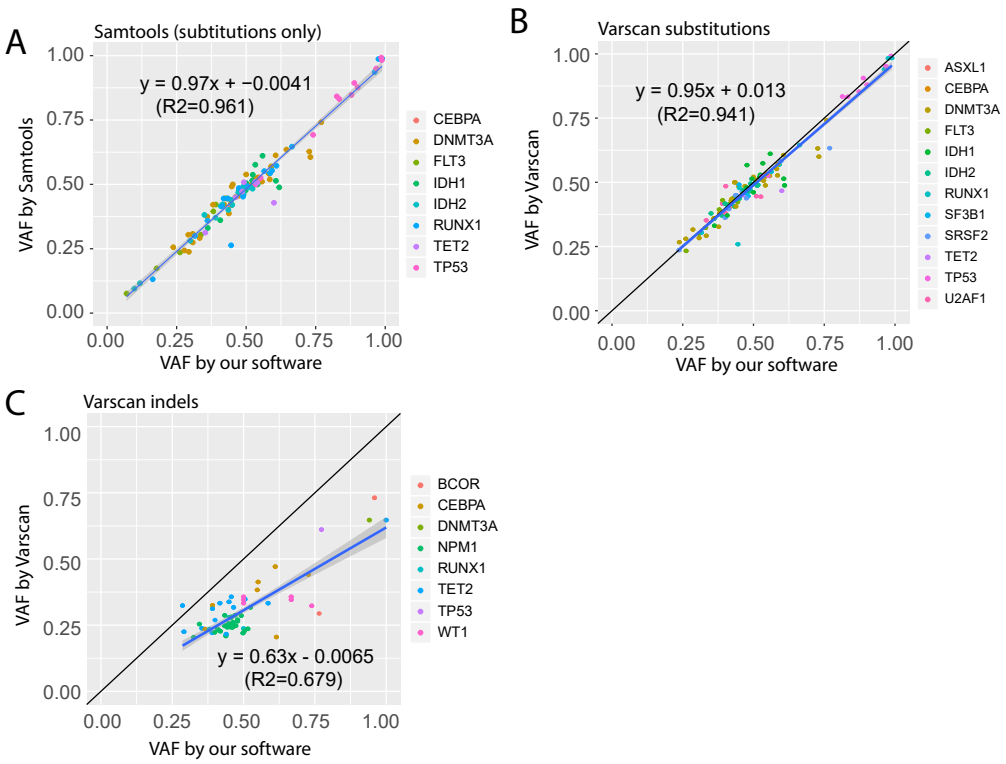
3.3.3 Comparison between our software and Salmon

Salmon is another widely used aligner which aligns RNA-seq reads directly to transcripts, which is also similar to RNAmut's alignment in nature. To compare with Salmon, we built the gene index using transcript sequences and quantified number of reads aligned to each gene. The comparison shows very good correlation between our software and Salmon (Supplementary Figure 15).



Supplementary Figure 14: Comparison between read alignment by our software and Salmon. Similar to Supplementary Figure 12A.

3.3.4 Our VAF calculation agrees with Samtools and Varscan



Supplementary Figure 15: Comparison between VAFs calculated by our software and (A) Samtools pileup, which only detects substitutions, **(B)** Varscan for substitutions and **(C)** Varscan for indels.

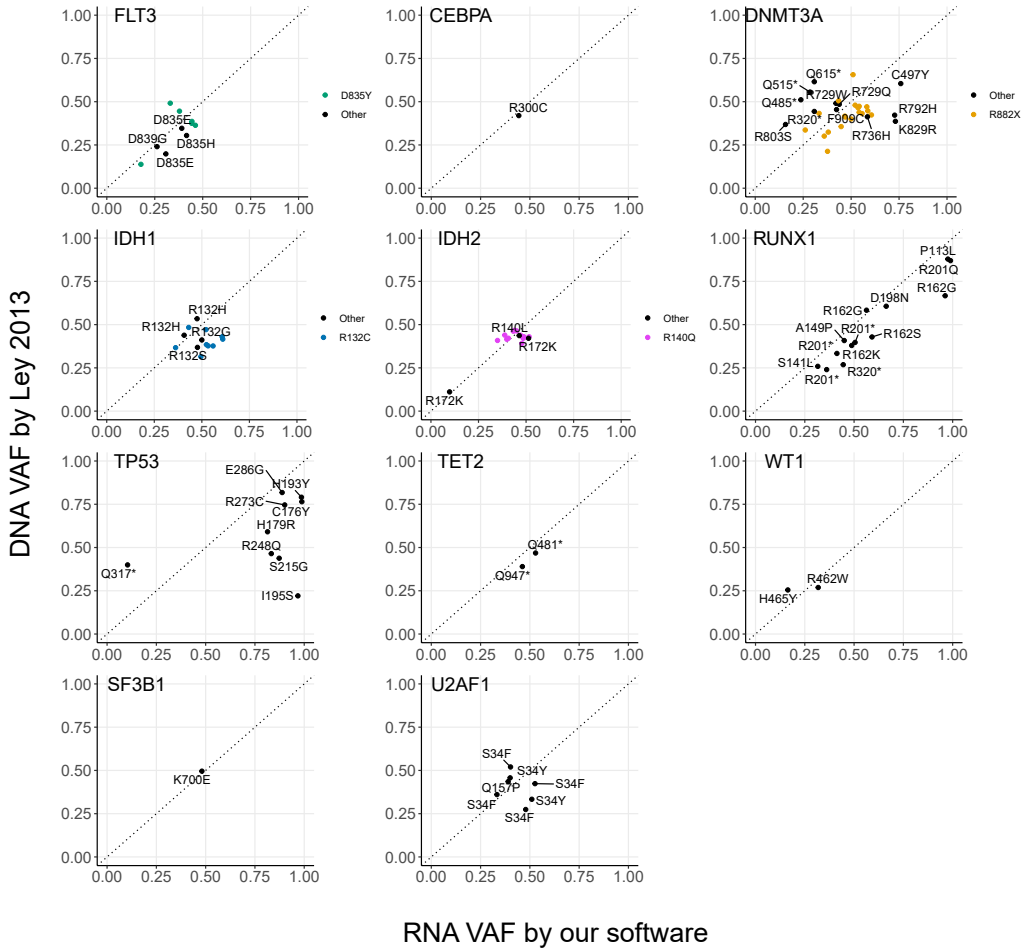
3.3.5 Our fusion detection agrees with and out-performs FuSeq for MLL fusions

Sample ID	MLL-fusion Our software	MLL-fusion FuSeq	PML-RARA Our software	PML-RARA FuSeq	MYH11-CBFB Our software	MYH11-CBFB FuSeq	RUNX1-RUNX1T1 Our software	RUNX1-RUNX1T1 FuSeq	BCR-ABL1 Our software	BCR-ABL1 FuSeq	NUP98-NSD1 Our software	NUP98-NSD1 FuSeq
2844	ELL											
2834	ELL	ELL										
2883	MLLT4											
2842	MLLT10	MLLT10										
2893	MLLT4											
2894	MLLT3											
2911	ELL	ELL										
2956	MLLT3	MLLT3										
2823			Y	Y								
2840			Y	Y								
2841			Y	Y								
2862			Y	Y								
2872			Y	Y								
2897			Y	Y								
2980			Y	Y								
2982			Y	Y								
2991			Y	Y								
2994			Y	Y								
2998			Y	Y								
2999			Y	Y								
3001			Y	Y								
3007			Y	Y								
3012			Y	Y								
2815					Y	Y						
2828					Y	Y						
2846					Y	Y						
2870					Y	Y						
2881					Y	Y						
2888					Y	Y						
2889					Y	Y						
2892					Y	Y						
2914					Y	Y						
2942					Y	Y						
2806							Y	Y				
2819							Y	Y				
2858							Y	Y				
2875							Y	Y				
2886							Y	Y				
2937							Y	Y				
2950							Y	Y				
2817									Y	Y		
2901									Y	Y		
2941									Y	Y		
2856											Y	Y
2918											Y	Y
2930											Y	Y

Supplementary Table 3: Comparison between our software and FuSeq for detecting gene fusions. Samples where a fusion is detected are indicated as Y. For *MLL*-fusions, the fusion partner is shown in the box.

3.4 RNA and DNA VAFs

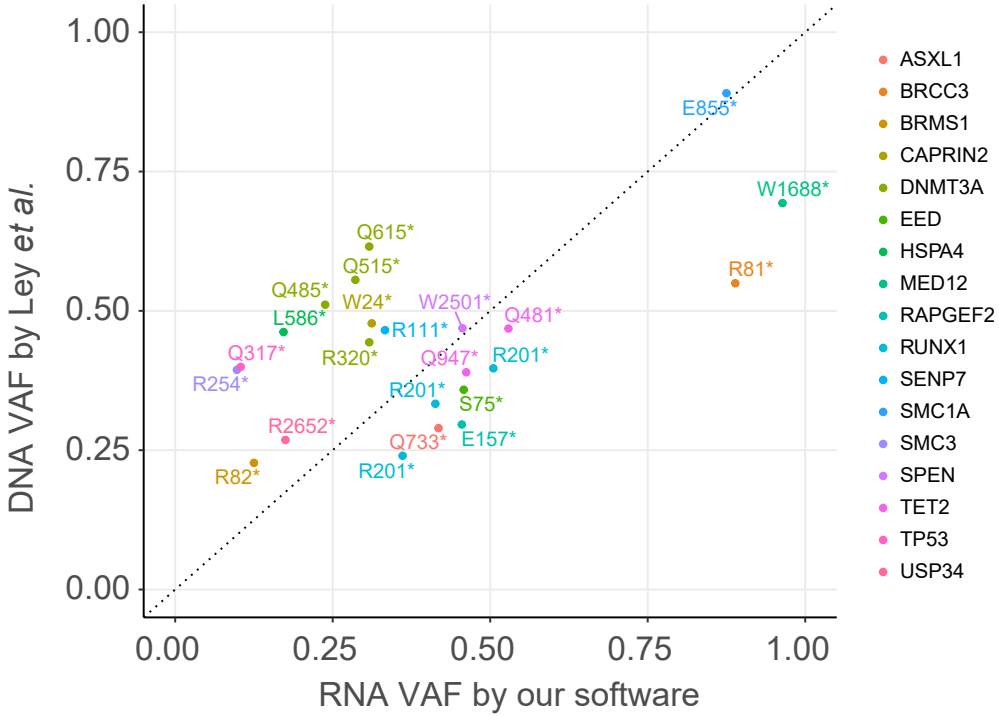
3.4.1 Comparision between DNA and RNA VAFs



Supplementary Figure 16: Comparison of RNA VAF and DNA VAF RNA VAFs were calculated by our software and DNA VAFs were obtained from whole exome data by Ley *et al.* 2013 [20]. Only VAFs for substitutions are shown since the software Ley et al. used did not report VAFs for indels or ITDs.

3.4.2 VAFs of Putative Non-sense Mediated Decay Mutations

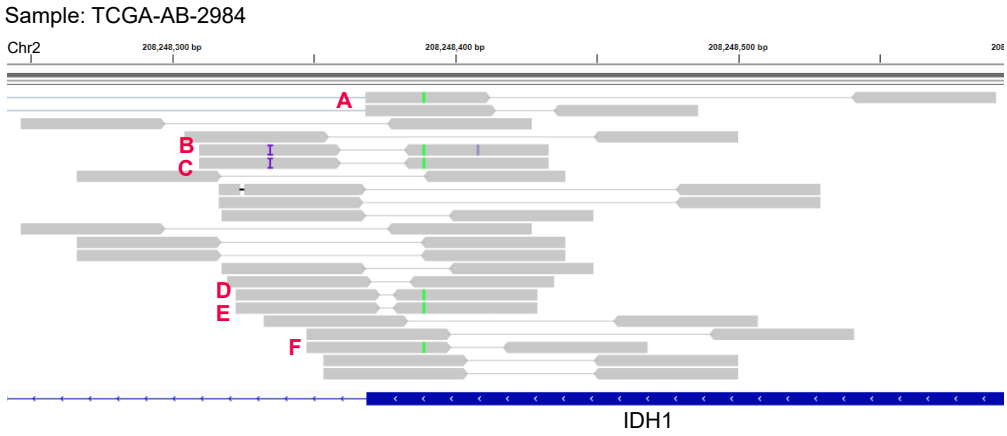
To test whether non-sense mediated decay can lead to RNAmut missing mutations from RNA-seq data, we assessed the differences between DNA and RNA VAFs. Since Ley *et al.* only reported DNA VAFs for gain-of-stop-codon, but not for frameshift mutations, while our software reports both, we only checked the correlation for stop-codon gains in 17 genes identified in the TCGA dataset by Ley *et al.* 2013.
We constructed a new index for all these genes (as not all were included in our 33-gene panel) and then called mutations using RNAmut. RNAmut reported 23 mutations of stop-codon gains across these 17 genes. Comparison between RNA and DNA VAFs shows roughly similar levels of RNA VAFs compared to DNA VAFs (Supplementary Figure 17) and there was no evidence for consistently lower values derived from RNA data. Pertinently, even in instances with lower RNA VAFs mutations were easily detectable from RNA-seq data.



Supplementary Figure 17: VAFs of potential non-sense mediated Decay. RNA VAFs quantified by our software is plotted against DNA VAFs quantified by Ley *et al.*

3.5 The IDH1 mutation not detected by RNA-seq

In sample TCGA-AB-2984, our software failed to detect the *IDH1* R132C mutation that has been annotated from whole exome sequencing. We inspected the RNA-seq reads around the hotspot (chr2:208248389) and found only one mutated reads aligned to the hotspot (Supplementary Figure 18).

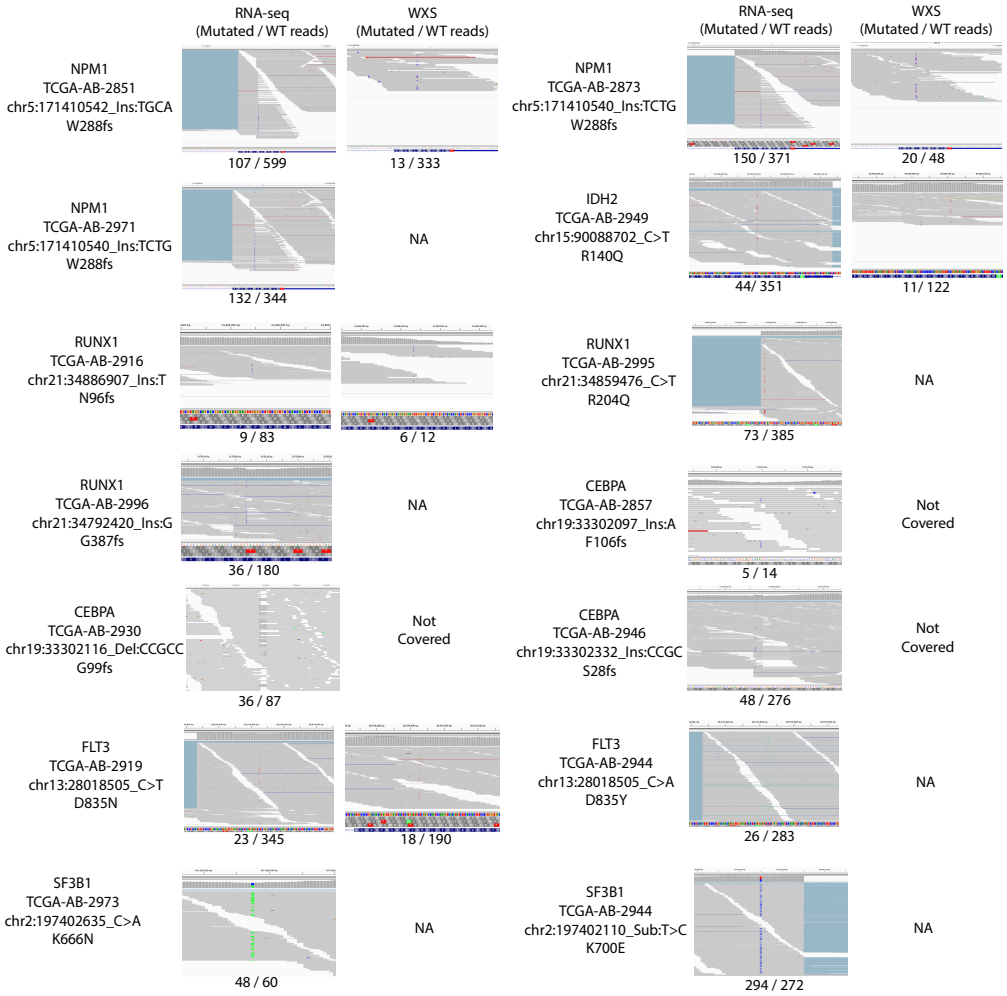


Supplementary Figure 18: RNA-seq reads of the sample TCGA-AB-2984 around the *IDH1* R132C hotspot. Only one read (A) with mutation is found at this hotspot. Other reads (B-F) contain sequences outside the exon of *IDH1*, which were not captured by our software. These reads may come from genomic contamination of the RNA-seq library or RNA from retained introns.

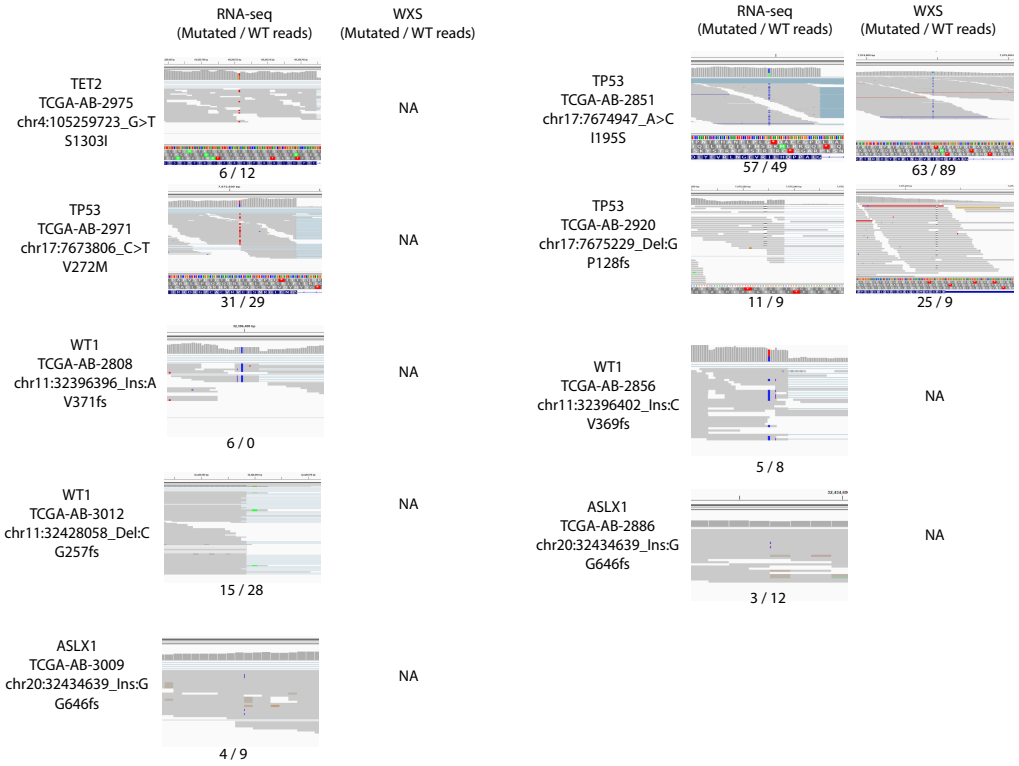
3.6 Evidence for novel detections by our software

Our software detected mutations in 29 samples that were not previously detected (23 SNVs/indels and 6 ITDs). In this section we provide evidence of mutated reads in RNA-seq and where available whole exome sequencing (WXS) data. Reads containing substitutions and small indels were visualized in the IGV browser [13]. However, reads from tandem duplications cannot be visualized by IGV because they are unaligned to the genome. Instead, we listed all these reads in relation to the duplication junction to demonstrate that our discoveries are true positives.

3.6.1 Evidence for substitutions and small indels



Supplementary Figure 19: Evidence for substitutions and small indels detected by our software. IGV browser tracks showing evidence of mutated reads for mutations detected by our software but not by previous studies. WXS did not cover the N-terminal domain of *CEBPA* and is hence not shown.



Supplementary Figure 20: Evidence for substitutions and small indels detected by our software (continued).

3.6.2 Evidence verifying newly detected *FLT3*-ITDs

DupEnd / (Insertion) / DupStart	Read ID
Sample: TCGA-AB-2823.ITD: chr13:28033991-28034136	
TCTCAAATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATG	SOLEXA2_0122:5:78:8853:3082/2_rev
CAAATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAAT	SOLEXA2_0122:5:75:6211:20213/2_rev
AAATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATA	SOLEXA3_0140:2:120:15472:12796/1_rev
AAATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATA	SOLEXA2_0122:5:85:11060:10706/2
AATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATAT	SOLEXA2_0122:5:93:3465:16053/2_rev
AATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATAT	SOLEXA2_0122:5:75:10984:15839/2
ATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATG	SOLEXA3_0140:2:24:18022:18128/2
ATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATG	SOLEXA2_0122:5:26:11889:3003/2_rev
ATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATG	SOLEXA3_0140:2:103:19572:1917/1
GGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGAT	SOLEXA3_0140:2:106:3407:10950/2_rev
GGGAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGAT	SOLEXA3_0140:2:54:11231:11636/2_rev
GAGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCT	SOLEXA2_0122:5:88:14467:5430/1
AGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTC	SOLEXA3_0140:2:49:1961:3351/1
AGTTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTC	SOLEXA2_0122:5:31:6745:19023/2
TTTCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAA	SOLEXA2_0122:5:66:13720:3765/2_rev
TCCAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAAT	SOLEXA3_0140:2:109:11783:5741/1_rev
CAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGG	SOLEXA3_0140:2:55:12081:14503/2_rev
CAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGG	SOLEXA2_0122:5:63:3615:10142/2
CAAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGG	SOLEXA3_0140:2:105:13005:19611/2
AAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGG	SOLEXA3_0140:2:45:7162:14941/2
AAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGG	SOLEXA2_0122:5:31:15148:18558/1_rev
AAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGG	SOLEXA3_0140:2:78:4949:1702/1_rev
AAGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGG	SOLEXA3_0140:2:96:6686:11949/1_rev
AGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGA	SOLEXA3_0140:2:58:17139:14951/2
AGAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGA	SOLEXA2_0122:5:44:6116:16994/2_rev
GAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAG	SOLEXA2_0122:5:114:14378:6841/1_rev
GAGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAG	SOLEXA3_0140:2:107:10471:8120/2_rev
AGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGT	SOLEXA2_0122:5:15:9986:2873/1_rev
AGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGT	SOLEXA2_0122:5:83:9208:2728/1
AGAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGT	SOLEXA3_0140:2:95:11038:8907/2
GAAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTT	SOLEXA2_0122:5:65:10632:17647/2_rev
AAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTT	SOLEXA3_0140:2:1:2713:12414/1_rev
AAAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTT	SOLEXA2_0122:5:30:6131:8594/2
AAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTC	SOLEXA2_0122:5:118:5265:3880/1
AAATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTC	SOLEXA2_0122:5:114:2478:8090/2
ATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTCCA	SOLEXA2_0122:5:44:13892:15381/1
ATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTCCA	SOLEXA2_0122:5:106:13030:19235/1
ATTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTCCA	SOLEXA3_0140:2:61:15216:6070/1_rev
TTTAGAGTTTGG T AGAGAATATGAATATGATCTCAAATGGGAGTTTCCAA	SOLEXA2_0122:5:115:4161:17442/1
Sample: TCGA-AB-2823. ITD: chr13:28034110-28034181	
CTACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCG	SOLEXA2_0122:5:80:6414:11108/2
CTACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCG	SOLEXA2_0122:5:61:4349:8630/1
CTACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCG	SOLEXA2_0122:5:11:7640:16497/2_rev
CTACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCG	SOLEXA3_0140:2:16:11203:18013/2_rev
ACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGC	SOLEXA2_0122:5:96:17043:6482/1_rev
ACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGC	SOLEXA2_0122:5:94:4136:8911/2_rev
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCT	SOLEXA3_0140:2:33:12050:1299/1_rev
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCT	SOLEXA3_0140:2:11:6690:6934/1_rev
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCT	SOLEXA2_0122:5:78:4403:10981/2
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCT	SOLEXA3_0140:2:8:17467:5048/1
GTTGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTC	SOLEXA2_0122:5:112:7686:4283/2
TGATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCT	SOLEXA2_0122:5:66:12294:12490/2
GATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTC	SOLEXA3_0140:2:16:7145:9583/1
ATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCA	SOLEXA3_0140:2:2:10228:13571/1
ATTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCA	SOLEXA2_0122:5:10:16344:8444/2_rev
TTTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAG	SOLEXA3_0140:2:96:14504:13549/2_rev
TTCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGA	SOLEXA2_0122:5:69:14521:4399/1
TCAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGAT	SOLEXA3_0140:2:50:17297:17194/2
CAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATA	SOLEXA3_0140:2:68:8131:7955/2_rev
CAGAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATA	SOLEXA2_0122:5:19:10173:20348/1_rev
GAGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAAT	SOLEXA2_0122:5:69:19022:13367/2
AGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATG	SOLEXA3_0140:2:95:17951:11266/2
AGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATG	SOLEXA2_0122:5:8:12908:14475/2
AGAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATG	SOLEXA3_0140:2:107:16869:21047/1_rev
GAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGA	SOLEXA2_0122:5:102:13360:13586/1
GAATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGA	SOLEXA2_0122:5:62:8922:11669/2
AATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAG	SOLEXA3_0140:2:83:1483:10977/2
ATATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGT	SOLEXA2_0122:5:49:12200:16639/2_rev
TATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTA	SOLEXA3_0140:2:25:15935:20306/2_rev
ATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTAC	SOLEXA2_0122:5:66:4385:2883/2_rev
ATGAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTAC	SOLEXA2_0122:5:29:16750:18669/1

GAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTT	SOLEXA2_0122:5:94:17431:11642/2_rev
GAATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTT	SOLEXA2_0122:5:21:7896:1352/1
AATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTC	SOLEXA2_0122:5:30:19116:10569/1_rev
ATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCT	SOLEXA2_0122:5:106:18556:11097/2
ATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCT	SOLEXA2_0122:5:20:15511:18616/1
ATATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCT	SOLEXA2_0122:5:98:18561:19697/2
TATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTA	SOLEXA3_0140:2:25:11526:7594/1_rev
TATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTA	SOLEXA2_0122:5:117:17000:2379/1_rev
TATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTA	SOLEXA3_0140:2:20:4195:20209/1
ATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTAC	SOLEXA2_0122:5:103:3790:12210/1_rev
ATGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTAC	SOLEXA2_0122:5:75:17171:16837/2_rev
TGATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACG	SOLEXA2_0122:5:104:9498:20453/1_rev
ATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTT	SOLEXA2_0122:5:74:10757:4143/1
ATCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTT	SOLEXA2_0122:5:97:15064:7222/1_rev
TCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTTG	SOLEXA3_0140:2:11:12395:7414/2
TCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTTG	SOLEXA2_0122:5:89:13262:13107/2_rev
TCTCAAATGGCAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTTG	SOLEXA2_0122:5:77:5473:11938/1_rev
Sample: TCGA-AB-2862. ITD: chr13:28034092-28034160	
AATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTAC	SOLEXA11_36:3:27:18519:17314/2_rev
AATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTAC	SOLEXA3_1:3:53:17417:16056/2
ATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACT	SOLEXA11_36:3:39:4909:1819/1
ATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACT	SOLEXA3_1:3:71:5826:12224/1_rev
ATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACT	SOLEXA3_1:3:35:6804:14524/2
TATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTT	SOLEXA3_1:3:53:10971:21048/2_rev
TATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTT	SOLEXA3_1:3:110:16954:18768/2
TGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCT	SOLEXA11_36:3:106:9357:17783/2
TGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCT	SOLEXA11_36:3:16:7662:7097/2
TGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCT	SOLEXA3_1:3:22:10902:17961/2
GAATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTA	SOLEXA3_1:3:86:12306:4162/2
AATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTAC	SOLEXA3_1:3:51:15104:5838/1_rev
ATATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACG	SOLEXA11_36:3:113:19754:8438/1
TATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGT	SOLEXA11_36:3:120:3459:7168/1
TATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGT	SOLEXA3_1:3:53:14102:20288/2
ATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTT	SOLEXA11_36:3:49:16159:15010/2
ATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTT	SOLEXA3_1:3:91:2159:9380/1_rev
ATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTT	SOLEXA3_1:3:11:3455:14590/1_rev
ATGATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTT	SOLEXA11_36:3:99:11129:4928/2_rev
GATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGA	SOLEXA3_1:3:49:10730:17663/1
GATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGA	SOLEXA3_1:3:24:18866:2661/1
GATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGA	SOLEXA3_1:3:117:14611:7288/2_rev
ATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGAT	SOLEXA3_1:3:66:14854:13036/2
ATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGAT	SOLEXA3_1:3:88:12347:18143/1
ATCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGAT	SOLEXA3_1:3:116:6612:10612/2_rev
TCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATT	SOLEXA3_1:3:32:7949:5753/2
TCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATT	SOLEXA3_1:3:25:17864:17692/2
TCTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATT	SOLEXA3_1:3:91:2659:18446/2_rev
CTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTT	SOLEXA3_1:3:17:18954:15359/1_rev
CTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTT	SOLEXA11_36:3:78:5408:15527/1_rev
CTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTT	SOLEXA11_36:3:45:9883:2754/1
CTCAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTT	SOLEXA11_36:3:80:5563:17900/2
CAAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCA	SOLEXA11_36:3:77:9073:5114/2
AAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAG	SOLEXA11_36:3:115:3710:18448/2
AAATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAG	SOLEXA11_36:3:10:11854:3359/1_rev
AATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGA	SOLEXA3_1:3:13:7141:11843/2
AATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGA	SOLEXA3_1:3:69:18702:19224/2_rev
AATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGA	SOLEXA3_1:3:104:19559:15512/2_rev
ATGGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAG	SOLEXA11_36:3:23:12139:6894/2
GGGAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAA	SOLEXA3_1:3:72:15152:18501/2
GAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATA	SOLEXA3_1:3:58:11596:19239/2_rev
GAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATA	SOLEXA3_1:3:83:16234:9463/2_rev
GAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATA	SOLEXA3_1:3:26:17972:3509/2_rev
GAGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATA	SOLEXA3_1:3:47:14127:12110/1_rev
AGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATAT	SOLEXA3_1:3:21:3247:11838/1
AGTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATAT	SOLEXA3_1:3:43:5396:13133/2
GTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATG	SOLEXA3_1:3:51:18644:16202/1
GTTTCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATG	SOLEXA11_36:3:105:18088:15980/2_rev
TCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAAT	SOLEXA3_1:3:1:19757:11326/1
TCCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAAT	SOLEXA3_1:3:93:18580:15300/1
CCAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATA	SOLEXA3_1:3:102:1391:13581/1_rev
CAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATAT	SOLEXA3_1:3:45:12308:20158/2
CAAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATAT	SOLEXA3_1:3:87:12038:8680/1
AAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATG	SOLEXA3_1:3:99:17451:14629/2
AAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATG	SOLEXA3_1:3:93:13743:7527/1_rev
AAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATG	SOLEXA11_36:3:3:13637:15990/1_rev
AAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATG	SOLEXA11_36:3:88:8606:4535/2_rev
AAGAGAAAAT AATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATG	SOLEXA11_36:3:86:18747:9194/1

Sample: TCGA-AB-2896. ITD: chr13:28033977-28034129	
AGTTTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATAT	SOLEXA12_58:1:114:12269:8993/1
AGTTTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATAT	SOLEXA4_149:1:117:5999:8336/2
AGTTTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATAT	SOLEXA12_58:1:90:18506:12471/2_rev
GTTTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATG	SOLEXA12_58:1:103:6068:16212/2_rev
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA4_149:1:109:19604:8542/1_rev
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA4_149:1:109:11537:7087/2_rev
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA4_149:1:28:15392:16387/2
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA12_58:1:35:6385:9888/1_rev
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA4_149:1:107:15576:16617/2
TTCCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGAT	SOLEXA4_149:1:85:2343:15374/2_rev
CCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCT	SOLEXA4_149:1:107:15931:15766/2_rev
CCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCT	SOLEXA12_58:1:75:16524:7340/2_rev
CCAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCT	SOLEXA12_58:1:63:16102:15939/2_rev
CAAGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTC	SOLEXA4_149:1:39:5807:13905/1
AGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAA	SOLEXA12_58:1:9:11588:12896/2
AGAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAA	SOLEXA4_149:1:48:4380:14677/2
GAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAA	SOLEXA4_149:1:88:13271:12734/1
GAGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAA	SOLEXA12_58:1:84:4925:4650/2
AGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAAT	SOLEXA12_58:1:50:4393:9289/2
AGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAAT	SOLEXA12_58:1:67:6852:18637/2_rev
AGAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAAT	SOLEXA4_149:1:2:10667:3835/1
GAAAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATG	SOLEXA4_149:1:5:1716:14595/1_rev
AAATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGG	SOLEXA4_149:1:7:18239:18652/1_rev
AATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGA	SOLEXA4_149:1:11:15809:14905/1
AATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGA	SOLEXA12_58:1:99:7464:8444/2
ATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAG	SOLEXA4_149:1:14:15013:13519/2
ATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAG	SOLEXA12_58:1:57:16667:1054/1_rev
ATTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAG	SOLEXA12_58:1:25:14579:20820/1
TTTAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGT	SOLEXA12_58:1:78:15478:5622/1_rev
TAGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTT	SOLEXA4_149:1:57:8153:7113/2_rev
AGAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTC	SOLEXA12_58:1:28:6811:19346/1
GAGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCC	SOLEXA12_58:1:6:16396:4124/2_rev
AGTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCA	SOLEXA12_58:1:72:7172:1822/1
GTTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAA	SOLEXA4_149:1:62:5210:9561/2_rev
TTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAG	SOLEXA12_58:1:23:15037:18044/1_rev
TTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAG	SOLEXA4_149:1:84:2758:1953/2_rev
TTTGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAG	SOLEXA12_58:1:114:9933:5703/2_rev
TGGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAG	SOLEXA12_58:1:20:13587:18020/1
GGGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGA	SOLEXA12_58:1:59:10497:18090/1
GGAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAA	SOLEXA12_58:1:119:7873:4064/2_rev
GAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAA	SOLEXA12_58:1:75:7636:20621/1_rev
GAAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAA	SOLEXA4_149:1:33:6060:18719/2
AAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAA	SOLEXA4_149:1:118:1962:12296/2
AAGGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAA	SOLEXA12_58:1:94:18233:1164/2_rev
GGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATT	SOLEXA12_58:1:84:10613:21191/1
GGTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATT	SOLEXA12_58:1:120:8684:17277/1
GTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTT	SOLEXA4_149:1:59:9742:9916/1
GTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTT	SOLEXA4_149:1:106:5233:15507/2
GTACTAGGAT ATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTT	SOLEXA4_149:1:43:3986:3447/1_rev
Sample: TCGA-AB-2919. ITD: chr13:28034089-28034181	
TATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGAC	SOLEXA5_0118:1:40:18858:5550/2_rev
GAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGG	SOLEXA5_0118:1:35:4531:1282/2_rev
GAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGG	SOLEXA12_0040:1:52:3014:15735/2_rev
AATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGC	SOLEXA12_0040:1:45:4733:20061/2_rev
ATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCT	SOLEXA5_0118:1:6:2243:11816/2
ATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCA	SOLEXA12_0040:1:113:4260:13180/1
ATCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCA	SOLEXA5_0118:1:42:9391:1276/1_rev
TCTCAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAG	SOLEXA12_0040:1:94:12111:11907/2
CAAATGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATA	SOLEXA5_0118:1:53:1989:3388/1
TGGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGA	SOLEXA5_0118:1:92:9541:20109/2
GGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAG	SOLEXA5_0118:1:36:15359:19269/1_rev
GGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAG	SOLEXA5_0118:1:109:13373:4132/2_rev
GGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAG	SOLEXA5_0118:1:27:16351:9243/2
GGGAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAG	SOLEXA12_0040:1:67:2587:18062/1_rev
GAGTTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTA	SOLEXA12_0040:1:94:17983:12467/1
TTTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTT	SOLEXA12_0040:1:42:4106:18929/1_rev
TTCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTC	SOLEXA12_0040:1:52:12860:2241/1
TCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCT	SOLEXA12_0040:1:88:13237:11394/1_rev
TCCAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCT	SOLEXA12_0040:1:110:11565:14616/2_rev
CAAGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCTAC	SOLEXA5_0118:1:8:4272:16930/2_rev
AGAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGT	SOLEXA5_0118:1:113:17446:16533/1
GAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTT	SOLEXA5_0118:1:75:7417:12288/2
GAGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTT	SOLEXA12_0040:1:107:11445:14961/2_rev
AGAAAATTTACAG GTGACCGGCTCCTCAGATAATGAGTACTTCTACGTTG	SOLEXA5_0118:1:39:10777:17140/1

Sample: TCGA-AB-2949. ITD: chr13:28034107-28034133	
TACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAA	SOLEXA9_0091:4:98:3784:8637/1
ACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAAC	SOLEXA5_0130:5:101:14036:6855/2_rev
ACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAAC	SOLEXA9_0091:4:43:17511:5629/1_rev
ACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAAC	SOLEXA5_0130:5:39:8643:2114/2_rev
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACG	SOLEXA9_0091:4:81:14696:6484/1
CGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACG	SOLEXA9_0091:4:65:5733:4103/2
GTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGG	SOLEXA9_0091:4:61:10276:6831/1
GTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGG	SOLEXA9_0091:4:1:6967:6956/1
GTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGG	SOLEXA5_0130:5:69:15019:12956/1
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:108:16681:9419/1_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA5_0130:5:119:8419:21145/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:52:17085:3583/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:75:9093:6798/1_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:95:18577:18319/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA5_0130:5:20:8565:11445/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:72:5326:18262/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA5_0130:5:9:7432:3537/2_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA9_0091:4:114:4644:4715/1_rev
TTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG	SOLEXA5_0130:5:30:19752:10644/1_rev
TGATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG G	SOLEXA9_0091:4:71:18543:3122/2_rev
GATTTCAGAGAATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GA	SOLEXA5_0130:5:39:12664:4695/1_rev
AATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATAT	SOLEXA5_0130:5:119:10052:1170/1_rev
ATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATG	SOLEXA9_0091:4:17:6852:8672/1_rev
ATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATG	SOLEXA9_0091:4:81:18701:4349/2
ATATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATG	SOLEXA9_0091:4:89:15952:5951/1_rev
TATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGA	SOLEXA5_0130:5:90:3701:3700/1
TATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGA	SOLEXA5_0130:5:4:6102:15149/1
ATGAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGAT	SOLEXA9_0091:4:108:9803:3480/1
GAATATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCT	SOLEXA9_0091:4:80:5806:9294/1_rev
ATGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCTCAAA	SOLEXA5_0130:5:51:11432:9755/1_rev
TGATCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCTCAAAT	SOLEXA9_0091:4:28:11884:5506/1
TCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCTCAAATGGG	SOLEXA5_0130:5:34:19443:2860/2
TCTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCTCAAATGGG	SOLEXA5_0130:5:14:4020:17847/2_rev
CTCAAATGGGAGGTTC TAAACGGG GAATATGAATATGATCTCAAATGGGA	SOLEXA9_0091:4:116:16189:16806/2

Supplementary Table 4: RNA-seq reads originated from the duplication junctions are listed for each *FLT3*-ITD detected by our software. Junctions consist of the end of duplicated sequence (DupEnd) followed by the start of the duplicated sequence (DupStart). In some cases, one or more nucleotides are inserted in between.

4 Supplementary data

Excel sheets for all detected mutations in the TCGA AML and MDS cohorts, and Leucegene datasets are provided in a separate file.

References

[1] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. Ensembl 2018. *Nucleic Acids Res*, 46(D1):D754–D761, 2018.

[2] E. Papaemmanuil, M. Gerstung, L. Bullinger, V. I. Gaidzik, P. Paschka, N. D. Roberts, N. E. Potter, M. Heuser, F. Thol, N. Bolli, G. Gundem, P. Van Loo, I. Martincorena, P. Ganly, L. Mudie, S. McLaren, S. O'Meara, K. Raine, D. R. Jones, J. W. Teague, A. P. Butler, M. F. Greaves, A. Ganser, K. Dohner, R. F. Schlenk, H. Dohner, and P. J. Campbell. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med*, 374(23):2209–2221, 2016.

[3] S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton. The catalogue of somatic mutations in cancer (cosmic). *Curr Protoc Hum Genet*, Chapter 10:Unit 10 11, 2008.

[4] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.

[5] Network Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–20, 2013.

[6] C. Simon, J. Chagraoui, J. Krosl, P. Gendron, B. Wilhelm, S. Lemieux, G. Boucher, P. Chagnon, S. Drouin, R. Lambert, C. Rondeau, A. Bilodeau, S. Lavallee, M. Sauvageau, J. Hebert, and G. Sauvageau. A key role for ezh2 and associated genes in mouse and human adult t-cell acute leukemia. *Genes Dev*, 26(7):651–6, 2012.

[7] C. Pabst, A. Bergeron, V. P. Lavallee, J. Yeh, P. Gendron, G. L. Norddahl, J. Krosl, I. Boivin, E. Deneault, J. Simard, S. Imren, G. Boucher, K. Eppert, T. Herold, S. K. Bohlander, K. Humphries, S. Lemieux, J. Hebert, G. Sauvageau, and F. Barabe. Gpr56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood*, 127(16):2018–27, 2016.

[8] T. Macrae, T. Sargeant, S. Lemieux, J. Hebert, E. Deneault, and G. Sauvageau. Rna-seq reveals spliceosome and proteasome genes as most consistent transcripts in human cancer cells. *PLoS One*, 8(9):e72884, 2013.

[9] V. P. Lavallee, S. Lemieux, G. Boucher, P. Gendron, I. Boivin, R. N. Armstrong, G. Sauvageau, and J. Hebert. Rna-sequencing analysis of core binding factor aml identifies recurrent zbtb7a mutations and defines runx1-cbfa2t3 fusion signature. *Blood*, 127(20):2498–501, 2016.

[10] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, 2002.

[11] R. Leinonen, H. Sugawara, M. Shumway, and Collaboration International Nucleotide Sequence Database. The sequence read archive. *Nucleic Acids Res*, 39(Database issue):D19–21, 2011.

[12] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[13] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178–92, 2013.

[14] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[15] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, 14(4):417–419, 2017.

[16] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project

Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[17] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, 22(3):568–76, 2012.

[18] T. N. Vu, W. Deng, Q. T. Trac, S. Calza, W. Hwang, and Y. Pawitan. A fast detection of fusion genes from paired-end rna-seq data. *BMC Genomics*, 19(1):786, 2018.

[19] Y. Shiozawa, L. Malcovati, A. Galli, A. Sato-Otsubo, K. Kataoka, Y. Sato, Y. Watatani, H. Suzuki, T. Yoshizato, K. Yoshida, M. Sanada, H. Makishima, Y. Shiraishi, K. Chiba, E. Hellstrom-Lindberg, S. Miyano, S. Ogawa, and M. Cazzola. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat Commun*, 9(1):3649, 09 2018.

[20] T. J. Ley, C. Miller, L. Ding, B. J. Raphael, A. J. Mungall, A. Robertson, K. Hoadley, Jr. Triche, T. J., P. W. Laird, J. D. Baty, L. L. Fulton, R. Fulton, S. E. Heath, J. Kalicki-Veizer, C. Kandoth, J. M. Klco, D. C. Koboldt, K. L. Kanchi, S. Kulkarni, T. L. Lamprecht, D. E. Larson, L. Lin, C. Lu, M. D. McLellan, J. F. McMichael, J. Payton, H. Schmidt, D. H. Spencer, M. H. Tomasson, J. W. Wallis, L. D. Wartman, M. A. Watson, J. Welch, M. C. Wendl, A. Ally, M. Balasundaram, I. Birol, Y. Butterfield, R. Chiu, A. Chu, E. Chuah, H. J. Chun, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R. A. Holt, S. Jones, A. Karsan, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, K. Mungall, J. Parker, E. Pleasance, P. Plettner, J. Schein, D. Stoll, L. Swanson, A. Tam, N. Thiessen, R. Varhol, N. Wye, Y. Zhao, S. Gabriel, G. Getz, C. Sougnez, L. Zou, M. D. Leiserson, F. Vandin, H. T. Wu, F. Applebaum, S. B. Baylin, R. Akbani, B. M. Broom, K. Chen, T. C. Motter, K. Nguyen, J. N. Weinstein, N. Zhang, M. L. Ferguson, C. Adams, A. Black, J. Bowen, J. Gastier-Foster, T. Grossman, T. Lichtenberg, L. Wise, T. Davidsen, J. A. Demchok, K. R. Shaw, M. Sheth, H. J. Sofia, L. Yang, J. R. Downing, and G. Eley. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–74, 2013.