Detection of Internal Tandem Duplications in the *FLT3* gene using PiVAT Software

Sean M. Polvino, Yue Ke, Nicholas J. Lodato, and Zhaohui Wang

Pillar Biosciences Inc., Natick MA USA

Abstract

Introduction: The *FLT3* gene is one of the most important genes in myeloid cancers and also one of the most difficult to sequence accurately due to the presence of internal tandem duplications (ITDs). Detection of ITDs is challenging for many next generation sequencing (NGS) bioinformatics pipelines due to limited read lengths. Here we developed a novel method to reliably and accurately detect ITDs and incorporated it into the Pillar Variant Analysis Toolkit (PiVAT).

Methods: To detect ITDs we utilized several components within PiVAT including a proprietary paired-end assembly method, local realignment, and processing of SA tag information. To assess the performance of PiVAT, publicly available data from 208 clinically detected ITDs identified in a 664-patient acute myelogenous leukemia (AML) study was used to simulate paired-end read NGS data for analysis. The data contained ITDs ranging in length from 3bp to 201bp with a median length of 36bp. Data was simulated using multiple read lengths from 150bp to 275bp in 25bp increments. To verify the synthetic data results, cell lines and reference samples with previously characterized ITDs of varying lengths in the *FLT3* gene were sequenced. Libraries were constructed using Pillar Biosciences' ONCO/Reveal Myeloid Panel and then sequenced on an Illumina MiSeq instrument. Secondary analysis was conducted using Pillar Biosciences' PiVAT software. ITDs with lengths of 21bp, 30bp, 33bp, 42bp, and 126bp were present in the samples and utilized to evaluate the detection capabilities of the PiVAT software.

Results: The simulation data demonstrated that with a read length of 150bp, PiVAT was able to detect ITD lengths up to 81bp, which represented >94% of the 208 ITDs that were in the cohort. Increasing the read length to 175bp allowed further detection of ITDs up to 117bp enabling detection of > 98% of the ITDs present in the study. Analysis of the experimental sequencing data from the real samples demonstrated strong agreement with the synthetic results. PiVAT was able to detect all but the 126bp ITD at a read length of 150bp and by increasing the read length to 175bp PiVAT was able to detect all of the ITDs. This indicated the ability to improve detection capability by increasing read length for even larger ITDs if desired.

Conclusions: The results demonstrated that the PiVAT software is capable of reliably detecting ITDs of up to 81bp using a read length of 150bp. In addition, by sequencing at a longer read length of 175 bp or greater PiVAT is able to detect ITDs of 126bp or longer.

Simulated Data and Analysis Methods

Simulated Data Generation:

- Data from Metzeler KH, Herold T, Rothenberg-thurley M, *et al.*[1] was used to generate simulated reads with the insertions as specified.
- This represents 208 clinically detected ITDs in a 664 patient AML study. The distribution of the ITD sizes can be seen in Figure 1.
- Each ITD described was used to generate synthetic fastq reads for analysis which would have a VAF of ~20%. The data was first used to create paired end reads at 275bp long with the desired variant. The variant was allowed to exist in all of the tiled amplicons that cover the *FLT3* exon 14-15 region in question.
- Real data obtained from the Pillar ONCO/Reveal Myeloid Panel using the genome in a bottle sample, NA12878, was used as a base background sample in which each synthetic fastq data was then individually spiked into to create variants with a random noise background, mimicking data similar to a real sample. This resulted in 208 samples with a PE read length of 275bp.
- In order to simulate shorter read lengths the data was trimmed in 25bp increments down to 150bp.
- Data from all variants and all read lengths was then analyzed using PiVAT 2019.3.1 to ascertain the detection capabilities as a function of read length and also of variant.

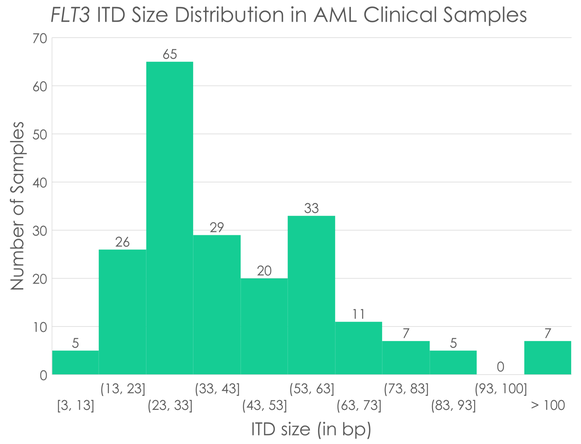

Figure 1 – ITD size distribution in AML clinical samples use in analysis

ITD Detection Method:

- PiVAT has a special built method for detecting ITDs based on the SA tag information that is generated using the BWA aligner
- The software examines the data in the tag and is able to realign quickly based on this information by modifying the cigar
- By employing this cigar realignment method it is significantly faster than performing an additional processing step to use another alignment algorithm such as Smith-Waterman local realignment.
- For shorter ITDs that do not generate a secondary alignment tag, these can be detected as well by using the existing local realignment algorithm in PiVAT

Library Generation of Cell Line Data:

- The Seracare Myeloid sample and cell lines MV4, PL21, and MOLM13 containing *FLT3* ITD of 21bp, 30bp, 33bp, 42bp, and 126bp were used as real test samples in order to validate the efficacy of the PiVAT detection methods.
- Data from several cell lines was sequenced using a MiSeq instrument at 2 x 275bp. Additional sequencing at 2 x 150bp was also performing using a Miseq instrument
- The 2 x 275bp data was trimmed in 25bp increments down to 150bp, and analyzed using PiVAT with the native 2 x 150bp data.

Results and Conclusions

Synthetic Data Results:

- Overall, including all variants, at a sequencing depth of 2 x 150bp, PiVAT detected 88% of all variants as seen in Table 1, and 81bp was the maximum length ITD that was detected.
- Including all ITDs 81bp or shorter, at 2 x 150bp sequencing PiVAT successfully identified 94% of all variants.
- Above 150bp read length, there is little improvement between 175bp and 275bp
- 275bp was able to detect up to 174bp
- The true detection edge for intermediate lengths is difficult to determine as the dataset has a large gap with no ITDs in between 117bp and 174bp
- Examination of the missed ITDs when running 2 x 150bp, when the insert is too close to the end of the read as to allow a complete secondary alignment

Range	Detection Rate					
	150bp	175bp	200bp	225bp	250bp	275bp
Called	184	202	202	204	204	205
Missed	24	6	6	4	4	3
Overall	**88%**	**97%**	**97%**	**98%**	**98%**	**99%**
≤81 (only)	**94%**	**99%**	**99%**	**100%**	**100%**	**100%**
≤21	97%	100%	100%	100%	100%	100%
22 – 42	93%	100%	100%	100%	100%	100%
43 – 63	94%	98%	96%	100%	100%	100%
64 – 84	89%	100%	100%	100%	100%	100%
>100	0%	55%	64%	64%	64%	73%

Table1: PiVAT detection efficiency of 208 clinically detected ITDs. Overall values, including only ITDs ≤81, and organized by increments

Sample	Expected ITD	Variant Call from PiVAT	Limits of Detection
Seracare Myeloid Mutation DNA mix	c.1759_1800dup 42bp insertion	c.1759_1800dup	Detected at all sequencing read lengths
Seracare Myeloid Mutation DNA mix	ITD + 5bp insertion 33bp insertion	c.1806_1807insGGGGCTTTCAGAGAATATGAATATGATCTCAAA	Detected at all sequencing read lengths
DSMZ Cell Line PL-21	ITD + 7bp insertion 126bp insertion	c.1837+15_1837+16insTCAAAACGGTACAGGTGACCGGCTCCTCAGATAATGAGTACTTCTACGTTGATTTCAGAGAATATGAATATGATCTCAAATGGGAGTTTCCAAGAGAAAATTTAGAGTTTGGTAAGAATGGAATGT	Not detectable at 2 x 150bp
DSMZ Cell Line MOLM-13	21bp duplication	c.1775_1795dup	Detected at all sequencing read lengths
DSMZ Cell Line MV4-11	30bp duplication	c.1772_1801dup	Detected at all sequencing read lengths

Table 2: Expected ITDs detected in real samples; variant call for each ITD detected and any known limitations on the variant call.

Experimental Data Results:

- PiVAT was able to detected all ITDs correctly at all lengths except for 126bp at 150bp read length (table 2).
- Variant call is consistent and identical across read lengths as well as reproducible across sequencers.

ITD Detection using PiVAT and the ONCO/Reveal Myeloid Panel Summary:

- The Pillar Variant Analysis Toolkit, PiVAT uses a specially defined algorithm to detect ITDs which is able to detect ITDs up to 81bp in length when used in conjunction with 150bp read length in paired end sequencing.
- The ability to detect longer ITDs is possible but requires the use of longer read lengths.
- The experimental results agree with the synthetic results, and indicate an ability to detect ITDs up to 174bp when run with a 275bp read length in paired end sequencing.

Future Work and Improvements:

- The ability to detect a given ITD using PiVAT and the Pillar ONCO/Reveal Myeloid Panel is dependent on 3 main factors: sequencing read length, ITD location (relative to amplicon positions), and ITD length.
- A current limitation of the algorithm requires the entire ITD to be seen in order to be able to detect it, i.e. at least 1 full copy of the ITD must be sequenced through
- It is possible in some cases, however that a partial copy of the ITD can be seen. Currently these cases are not handled. Handling these special edge cases would allow the detection of long ITDs, however they would not necessarily be able to be specifically identified, but be more used as a way to mark that an ITD is present.

[1]:Metzeler KH, Herold T, Rothenberg-thurley M, *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood. 2016;128(5):686-98.