The Journal of Molecular Diagnostics, Vol. 22, No. 3, March 2020

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

Junction Location Identifier (JuLI)

Accurate Detection of DNA Fusions in Clinical Sequencing for Precision Oncology

Hyun-Tae Shin,*† Nayoung K.D. Kim,*‡ Jae Won Yun,*§ Boram Lee,*¶ Sungkyu Kyung,*‡ Ki-Wook Lee,*‖ Daeun Ryu,* Jinho Kim,* Joon Seol Bae,* Donghyun Park,*‡ Yoon-La Choi,** Se-Hoon Lee,†† Myung-Ju Ahn,†† Keunchil Park,†† and Woong-Yang Park*‡¶‖‡‡

From the Samsung Genome Institute, Samsung Medical Center, Seoul; the Veterans Medical Research Institute,† Veterans Health Service Medical Center, Seoul; Geninus Inc.,‡ Seoul; the Departments of Laboratory Medicine and Genetics,§ Pathology and Translational Genomics,** Hematology and Oncology,†† and Molecular Cell Biology,‡‡ Sungkyunkwan University School of Medicine, Seoul; the Departments of Health Sciences and Technology¶ and Digital Health,‖ Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea*

Accepted for publication October 25, 2019.

Address correspondence to Woong-Yang Park, M.D., Ph.D., Samsung Genome Institute, Samsung Medical Center, 81, Ilwon-ro, Gangnam-gu, Seoul, Republic of Korea. E-mail: woongyang.park@samsung.com.

Accurate detection of genomic fusions by high-throughput sequencing in clinical samples with inadequate tumor purity and formalin-fixed, paraffin-embedded tissue is an essential task in precise oncology. We developed the fusion detection algorithm Junction Location Identifier (JuLI) for optimization of high-depth clinical sequencing. Novel filtering steps were implemented to minimize false positives in the clinical setting. The algorithm was comprehensively validated using high-depth sequencing data from cancer cell lines and clinical samples and genome sequencing data from NA12878. JuLI showed improved performance mainly in positive predictive value over state-of-the-art fusion callers in cases with high-depth clinical sequencing and rescued a driver fusion from false negative in plasma cell-free DNA using joint calling. *(J Mol Diagn 2020, 22: 304—318; https://doi.org/10.1016/j.jmoldx.2019.10.015)*

High-throughput sequencing is becoming increasingly prevalent in precision cancer medicine worldwide. In the Republic of Korea and the United States of America, assays using high-throughput sequencing have received regulatory approval as companion diagnostic tests for personalized care.[1,2] Most assays use sequencing technology to identify clinically actionable single-nucleotide variants and small insertions/deletions because they are relatively easy to detect and interpret. However, some cancers, such as ALK tyrosine kinase receptor (*ALK*)—rearranged non—small-cell lung cancers (NSCLCs) and *BCR-ABL* rearranged chronic myeloid leukemias, are driven by somatic genomic fusions that cannot be detected by these methods for single-nucleotide variants/insertions/deletions. Patients with these oncogenic fusions respond to tyrosine kinase inhibitors, and such genomic changes are now key therapeutic targets.[3,4]

Several factors are prerequisite for accurate detection of genomic fusions in the clinical setting. First, obtaining a representative specimen that provides an adequate amount of tumor sample for genome profiling is an ongoing challenge. Our previous study has shown that numerous important variants are present at a low allelic fraction.[5]

Supported by the Korean Health Technology Research and Development Project, Ministry of Health and Welfare, Republic of Korea, grant HI13C2096 (W.-Y.P.); the Korean Health Technology Research & Development Project via the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea, grant HI15C3224 (J.K.); the National Research Foundation of Korea, funded by the Korean Government (Ministry of Science and ICT), grants 2018M3C9A6017315 (D.P.) and 2019R1F1A1042379 (N.K.D.K.); and VHS Medical Center Research grant, Republic of Korea, VHSMC 19042 (H.-T.S.).

H.-T.S., N.K.D.K., and J.W.Y. contributed equally to this work.

Disclosures: This study was part of a doctoral dissertation by H.-T.S.

https://doi.org/10.1016/j.jmoldx.2019.10.015

Unlike tissues used for research, tissues from clinical procedures, such as biopsies, tend to have inadequate tumor purity. Recently, cell-free DNA (cfDNA) testing by ultra-deep sequencing has been introduced for genotyping primary cancers and monitoring of post-treatment recurrence in oncology, and this test aims to detect approximately 0.1% of allele fractions.[6–8] Furthermore, considering the heterogeneity of individual tumors, complete profiling of a tumor may require multiple samplings from different regions, which is not clinically feasible. To capture these low fraction variants, sufficient sequencing coverage and specialized algorithms are imperative for a clinical assay. Second, it is important to obtain a sufficient quality of formalin-fixed, paraffin-embedded (FFPE) specimens for genome profiling. FFPE is preferred for most molecular analyses of clinical pathologies because of its advantages in collection and storage. However, formalin fixation results in DNA and RNA damage, which is affected by various preanalytical factors, such as duration of storage, formalin fixation, and ischemic time.[9–11] These fragmented nucleic acids act as noise and may make it difficult to detect oncogenic fusions. The detection of genomic fusions in clinical samples tends to be challenging because of the above-mentioned problems.

As the importance of detecting genomic fusions in clinical decision making continues to increase, a critical area for improvement is currently the accuracy of detecting actionable fusions for the realization of precision cancer medicine. Herein, we focused on improving the reliability of detecting somatic actionable fusions in cancer using high-depth DNA sequencing. To address the above problems, a fusion detection algorithm optimized for clinical purposes was developed and was validated using cancer cell lines with known driver fusions, 459 NSCLC samples with known *ALK* fusion and/or proto-oncogene tyrosine-protein kinase receptor Ret (*RET*) fusion status, and 46 prostate cancer samples with known *TMPRSS2* fusion status (Supplemental Table S1).

Materials and Methods

Study Design

The Institutional Review Board of Samsung Medical Center (SMC; Seoul, Republic of Korea) approved this study. Clinical samples were obtained at SMC between March 2014 and February 2017, with informed consent from some patients, whereas consent was waived by the Institutional Review Board for others. The inclusion criteria for samples in this study were as follows: sample was profiled using CancerSCAN[5] or LiquidSCAN,[12] the custom sequencing platforms of SMC; and clinical information of the patient was stored in the clinical data warehouse of SMC.

Panel Design for Fusion Detection

Samples were prepared and analyzed using CancerSCAN or LiquidSCAN, targeted-sequencing platforms designed at SMC.[5,12] To identify fusions using a targeted panel, introns that contain well-known breakpoints of a set of clinically relevant fusions were tiled across the hotspot. Introns of five genes from an 83-gene panel (CancerSCAN version 1 and LiquidSCAN version 1) and introns of 22 genes from a 381-gene panel (CancerSCAN version 2) were densely covered with capture probes. All panels targeted hotspot introns of *ALK*, *RET*, and *TMPRSS2*. The average DNA fragment size of the platform was approximately 180 bp, and the read length was 100 bp, thus indicating that most fragments were fully sequenced. The other specific details of the panels have been previously reported.[5,12]

Cell Line Mix Experiment

Four cell lines (H2228, BHP10-3, U118MG, and SK-NEP-1) known to harbor specific fusions were used (Supplemental Table S2). The cell lines were cultured in our laboratory. Before extraction of DNA, the cells were washed two times with phosphate-buffered saline. When the samples were pooled, the value from the Qubit HS assay (Life Technologies, Carlsbad, CA) was used, and DNAs were mixed equally to a total amount of 500 ng. The sequencing data of cell lines have been deposited in National Center for Biotechnology Information sequence read archive (*https://www.ncbi.nlm.nih.gov/sra*; accession number PRJNA514104).

PCR Validation of Fusions

The reference sequence of a target gene and breakpoint region was retrieved from the University of California, Santa Cruz genome browser. A target-specific primer was designed using Primer3 for PCR on the basis of the reference sequence and was confirmed using Primer-BLAST (NIH, Bethesda, MD) (Table 1). The translocation target gene was amplified by PCR using specific primers. The cycling conditions were as follows: 94°C for 5 minutes, followed by 44 cycles of denaturation (94°C for 30 seconds), annealing (60°C for 1 minute), and extension (72°C for 1 minute), with final extension at 72°C for 10 minutes. The reactions were performed using HelixAmp TM Ready-2X-Go Hot-Taq (Nanohelix, Daejeon, Republic of Korea). Sequences of the PCR products were determined by an automated method (ABI Prism 3730) using the Big Dye Terminator Kit (Applied Biosystems, Foster City, CA). Translocation breakpoint region sequences were verified by means of BLAST (NIH) and DNAstar (Lasergene, Madison, WI).

Alignment and Preprocessing

Paired-end reads were aligned using BWA-MEM version 0.7.5 at its default settings[13] with the human reference genome (hg19). Aligned reads with mapping quality <20 were filtered out, and the remaining reads were sorted using SAMtools version 0.1.18.[14] To prepare appropriate input Binary Alignment Map (BAM) files for other callers, MarkDuplicates of Picard version 1.93 (Broad Institute,

Table 1 List of Sanger Sequencing Primers in Validation Samples

Identifier	Diagnosis	Tissue	Fusion genes	Read count	Translocation primer	PCR product size, bp
CS_VAL_00001	Lung cancer	Fresh	*EML4-ALK*	37	F: 5′-TTCACAGCAGCCGTATTGTC-3′ R: 5′-TACACTGCAGGTGGGTGGT-3′	190
CS_VAL_00002	Lung cancer	Fresh	*KIF5B-RET*	33	F: 5′-CCAGCTGTGGAGGTGACAG-3′ R: 5′-TGGGTTTGGTGACAAGTTTTT-3′	239
CS_VAL_00003	Lung cancer	FFPE	*EML4-ALK*	73	F: 5′-GAGCTTCTCTGAGTAAGGCATTT-3′ R: 5′-GTTGGGAGCTTCCGTTTTG-3′	245
CS_VAL_00004	Lung cancer	FFPE	*KIF5B-RET*	21	F: 5′-TGGCAATTAATGAACAAAGCTG-3′ R: 5′-GGGGAAAAGTGTGAGGATGA-3′	154
CS_VAL_00005	Prostate cancer	FFPE	*TMPRSS2-ERG*	20	F: 5′-AGACCCTGGTTCCCACACT-3′ R: 5′-GGTAAACTCTCCCTGCCACA-3′	213
CS_VAL_00006	Prostate cancer	FFPE	*TMPRSS2-SLC43A2*	6	F: 5′-CCTGGACAATTTGTTGCTCA-3′ R: 5′-TACAGGTGCTGCCTACGTGA-3′	131
CS_VAL_00007	Prostate cancer	FFPE	*TMPRSS2-ERG*	29	F: 5′-TGTCTGTGTTCACGGCTGTC-3′ R: 5′-AAACCAGAGGCATGAGGATG-3′	245
CS_VAL_00008	Prostate cancer	FFPE	*TMPRSS2-ERG*	28	F: 5′-AATGGCATCATAGCATCCAA-3′ R: 5′-TCCCCACCTTTACTGAGTGC-3′	240
CS_VAL_00009	Prostate cancer	FFPE	*TMPRSS2-ERG*	10	F: 5′-AAATACACGTTTTAGGAGCAAACA-3′ R: 5′-CTGATCCCAGAGGAGAGTGC-3′	208
CS_VAL_00010	Prostate cancer	FFPE	*TMPRSS2-ERG*	7	F: 5′-CCATCAGCATGACTGAAAGGT-3′ R: 5′-TATGCCAGTAACCACCACCA-3′	246
CS_VAL_00011	Prostate cancer	FFPE	*TMPRSS2-GPR98*	21	F: 5′-ACCCCAGATCTTGGCAGAG-3′ R: 5′-TGAGGAAATGCCTGTTTTGA-3′	239
CS_VAL_00012	Prostate cancer	FFPE	*TMPRSS2-ERG*	17	F: 5′-TGAGGAAAACTGAGACTCCAAA-3′ R: 5′-TTTCTGCTGAACAGCCACTG-3′	244
CS_VAL_00013	Prostate cancer	FFPE	*TMPRSS2-ERG*	36	F: 5′-CAGATGGGAATCGATGTGAA-3′ R: 5′-ATTTGCTCAGGAAGGTGCAT-3′	250
CS_VAL_00014	Sarcoma	FFPE	*EWSR1-WT1*	22	F: 5′-TGACTCCTTGTTCTTGCATCA-3′ R: 5′-GCCCTAAGAAACCTGGCTCT-3′	239
CS_VAL_00015	Sarcoma	FFPE	*EWSR1-FLI1*	29	F: 5′-ATCGTTTTTGGCCTCCCTAT-3′ R: 5′-GCAAAACCCAGTGACAGTGA-3′	241
CS_VAL_00016	Sarcoma	FFPE	*EWSR1-FLI1*	40	F: 5′-CCCACCTTCGGTAAATTGAG-3′ R: 5′-CATGGAAATGTCATCTTTGTGG-3′	242
CS_VAL_00017	Sarcoma	FFPE	*EWSR1-FLI1*	22	F: 5′-TGCAGGCCACTATGATTTTG-3′ R: 5′-AGTCCCTCTGGAAAGCCAAT-3′	218
CS_VAL_00018	Sarcoma	FFPE	*EWSR1-FLI1*	29	F: 5′-GCAATGAAAAAGGGCATGTT-3′ R: 5′-GGCTCCTACAGACCCTGTGA-3′	242
CS_VAL_00019	Lung cancer	FFPE	*EML4-ALK*	8	F: 5′-GGAGGTGTTGGGATCATTCA-3′ R: 5′-GAAGGTTGGGTGGAAGCAC-3′	204
CS_VAL_00020	Lung cancer	FFPE	*EML4-ALK*	26	F: 5′-GTGCTGCTGTGAACCTGAGA-3′ R: 5′-GGAAGAGTGGGCTAGTGCAT-3′	235
CS_VAL_00021	Lung cancer	FFPE	*EML4-ALK*	17	F: 5′-ATCTGTTTCCCCCAACATCA-3′ R: 5′-TGTAAATTGCCGAGCACGTA-3′	344
CS_VAL_00022	Lung cancer	FFPE	*EML4-ALK*	180	F: 5′-CTAATGAACAGGCTGCATGG-3′ R: 5′-ATCTGTCCTGGGCATGTCTC-3′	227
CS_VAL_00023	Lung cancer	FFPE	*EML4-ALK*	9	F: 5′-CACTTAAAGAAAAACGGGAAGG-3′ R: 5′-GCTGGGCTTTACACACAGAA-3′	242
CS_VAL_00024	Lung cancer	FFPE	*EML4-ALK*	44	F: 5′-TTGGGAGAAGCTGAAAATTCC-3′ R: 5′-CTCAAGAGCCTTTCCCTCTG-3′	211
CS_VAL_00025	Lung cancer	FFPE	*EML4-ALK*	39	F: 5′-AGGCTGCATGGAATCTGAAT-3′ R: 5′-GCACTACACAGGCCACTTCC-3′	217
CS_VAL_00026	Lung cancer	FFPE	*EML4-ALK*	9	F: 5′-CTCCTCCAAATCAAGCAAGC-3′ R: 5′-GCACTACACAGGCCACTTCC-3′	224
CS_VAL_00027	Lung cancer	FFPE	*EML4-ALK*	76	F: 5′-GTACACTGCAGGTGGGTGGT-3′ R: 5′-TGACCATGCACAGGGAAATA-3′	232
CS_VAL_00028	Lung cancer	FFPE	*EML4-ALK*	24	F: 5′-TGGAGACACATACTTAATTCTAAAACC-3′ R: 5′-AGCTCTGAACCTTTCCATCA-3′	400

(*table continues*)

Table 1 (*continued*)

Identifier	Diagnosis	Tissue	Fusion genes	Read count	Translocation primer	PCR product size, bp
CS_VAL_00029	Lung cancer	FFPE	*KIF5B-RET*	118	F: 5′-TCCCACTTTGGATCCTCCTA-3′ R: 5′-CATGTGTAGGCTGAGCATGG-3′	237
CS_VAL_00030	Lung cancer	FFPE	*KIF5B-RET*	79	F: 5′-CCTGGAGGCTCTGAGTAGCA-3′ R: 5′-CTGTTCTTTGCCAGCACTGT-3′	200
CS_VAL_00031	Lung cancer	FFPE	*CCDC6-RET*	33	F: 5′-TGGCTGATTTTGGGGAAATA-3′ R: 5′-AACCCACAGTCAAGGTCAGTG-3′	233
CS_VAL_00032	Lung cancer	FFPE	*KIF5B-RET*	35	F: 5′-TTGGACCTCATGTTTGATTCTTT-3′ R: 5′-GGAGGGCAGGGGATCTTC-3′	246
CS_VAL_00033	Breast cancer	Fresh	*NBEA-BRCA2*	58	F: 5′-TTAGTCTGTTACAAAAATTTTCATTTG-3′ R: 5′-CATTGAAACAACAGAATCATGACA-3′	347
CS_VAL_00034	Breast cancer	Fresh	*ERBB2-STARD3*	29	F: 5′-AGCAGTGTCTGTGTGCTTGG-3′ R: 5′-ATAGACACCAACCGCTCTCG-3′	436
CS_VAL_00035	Sarcoma	FFPE	*COL3A1-ERBB2*	156	F: 5′-GCAGTGCAGCTCACAGATG-3′ R: 5′-TGTGAATCATGCCCTACTGG-3′	188
CS_VAL_00036	Lung cancer	Fresh	*CMIP-ROS1*	283	F: 5′-GTGTCAGAGGACCAGGAAGG-3′ R: 5′-TCCAGCCTGTGCTTCAACTA-3′	500
CS_VAL_00037	Lung cancer	Fresh	*CD74-ROS1*	203	F: 5′-ATCAGCCACCCCCTTAATTC-3′ R: 5′-TCCCAGCCCTTGAATCTAGT-3′	386

F, forward; FFPE, formalin fixed, paraffin embedded; R, reverse.

Cambridge, MA), which is commonly used for marking and removal of duplicate reads,[15] was employed.

Workflow for Fusion Identification

Fusion Detection Algorithm

To identify genomic fusions for clinical applications, an algorithm called Junction Location Identifier (JuLI) was developed with the aim of reducing the number of false positives generated while maintaining sensitivity. The R package of JuLI is available online at GitHub (*https://github.com/sgilab/JuLI*, last accessed October 28, 2019). Initially, basic statistics of the BAM files, such as read length and median insert size, are calculated and used for further steps. Candidate breaks are then defined using two or more clipped reads, including at least one soft-clipped read, against the genome reference. If a matched normal sample is available as a control, breaks with twice the cutoff value of the clipped reads are scanned in the normal sample, and candidate breaks that overlapped with the breaks in the normal sample are excluded. If a set of normal samples is available, a control panel can be generated using a function in JuLI, which incorporates the breakpoints in multiple samples. All the samples in the present study were processed without matched normal or control panel filtering. The algorithm then involves two separate parts (ie, discordant and split read analyses). The user can set all parameters of each step.

Discordant Read Analysis

As JuLI does not remove duplicate reads as a part of the algorithm, counting supporting reads is important to reduce the number of false-positive calls. JuLI first uses information, including the genomic positions of both paired reads, Concise Idiosyncratic Gapped Alignment Report strings, and the QNAME of sequencing reads in the BAM file, to reduce redundant duplicated or noise signals. Candidate breaks with fewer than three unique discordant reads are filtered. Next, consensus contigs from the matched and clipped side of each candidate break are generated. The average number of pairwise differences, representing nucleotide diversity (π), between the reads and the consensus contig on both sides of the candidate break is calculated as follows:

$$\pi = \frac{N}{N-1}\sum_i p_i \pi_i \quad (1)$$

where N is the number of reads across the break, ρ_i is the frequency of the ith read across the break, and π_i is the proportion of bases that differ between the read and consensus contig truncated to the read length. If the normalized nucleotide diversity of either the clipped or matched side is >2.0, the break is excluded from further processing. The normalized nucleotide diversity is calculated using the following formula:

$$normalized\ \pi = \frac{\pi - \overline{\pi}_{matched\ side}}{S_{matched\ side}} \quad (2)$$

where $\overline{\pi}_{matched\ side}$ is the mean of nucleotide diversity of matched sides, and $S_{matched\ side}$ denotes the SD of nucleotide diversity of the matched sides. Candidate breaks that pass the filters described above are paired with each other using the pair information, and split side contigs of each pair are aligned to the matched side contigs of their partners. If one of the two pairs matches >70% of the split contig length and is >10 bp, the pair is called a fusion event. If there are

no candidate pairs that passed the filters, a fusion event is defined if more than six discordant reads formed a cluster and >70% and >20 bp of the split contig are mapped to the reference sequence of the cluster region.

Split Read Analysis

Split read and discordant read analyses are conducted similarly. Candidate breaks with fewer than three split reads are filtered and then subjected to the following filtering steps, including nucleotide diversity analysis and pairwise local alignment. As JuLI is based on split information, fusions with a length less than half the read length are not considered. In split read analysis, if both pairs matched ≥70% of the length of the split contigs, the pair is considered a fusion event.

Joint Call Analysis

Genotyping using additional information is one of the common strategies for detection of variants. JuLI's joint call function is also able to detect fusions with a small number of supporting reads using merged additional information from BAM files.[16,17] The joint call combines information from multiple BAM files in each analysis step and separates the numbers of each supporting read in the final step to produce individual results of the BAM files. If some fusion events have been previously defined in other BAM files, the fusions can be efficiently detected by specifying the target area using the Browser Extensible Data format. This is useful for the case in cfDNA analysis as acquired serial samples for cfDNA may not have enough supporting reads, which makes it difficult to detect the events (see *Discussion*).

Settings of Algorithms

The documentation for each algorithm was carefully studied to determine and apply parameters that could be optimized in the clinical sample data.

JuLI

All analyses were performed using JuLI version 0.1.3 with the default parameters. Fusion events in the University of California, Santa Cruz gap database were excluded from further analysis.

SvABA

All analyses were performed using SvABA version 134.[18] A -M flag was applied so that the number of weird reads was not limited in highly fragmented FFPEs. Sorted, indexed, and duplication-free BAMs were employed for SvABA. The command line for the analysis was as follows: $svaba run −t $INPUT.bam −p 1 -G $reference.fa −a sample_id −M 100000.

Delly

Delly version 0.7.8[19] was used for all analyses with the default parameters. BAM files were preprocessed as recommended by the developers (sorting, indexing, and duplicate marking). The exclusion regions of the hg19 reference included in the Delly source code were applied. The Delly command line for the analysis was as follows: $delly call -x human.hg19.excl.tsv -o $OUTPUT.bcf -g $reference.fa $INPUT.bam.

The output with binary variant call format was converted to variant call format using BCFtools, which was included as a submodule in Delly. The results of variant call format that passed the quality filter for all analyses were selected.

Manta

All analyses were performed using Manta version 1.2.2.[20] All high-depth filters were disabled by applying the –exome flag during configuration for high-depth sequencing data. Manta was used to analyze sorted, indexed, and duplication-free BAMs. The command line for configuring was as follows: $configManta.py –tumorBam $INPUT.bam –referenceFasta $reference.fa –runDir $OUTPUT_DIR –exome.

Next, a workflow run script with a single node was launched using the following command line for execution: $OUTPUT_DIR/runWorkflow.py -m local -j 1.

The results of variant call format that passed the quality filter for all analyses were selected. The high-depth filter parameter was applied for whole-genome sequencing (WGS) analysis.

LUMPY

LUMPY version 0.2.8 was used for all analyses.[21] LUMPY was used to analyze sorted, indexed, and duplication-free BAM files. The BAM files were split into paired-end and split-read files using SAMtools version 0.1.19[14] with the recommended parameters, and statistical analysis of the library sizes was performed by means of a script in LUMPY. Then, the following LUMPY command line for fusion detection was executed: $lumpy -mw 4 -tt 0.0 − pe bam_file:$INPUT.discordant.pe.bam,histo_file:$INPUT.pe.histo,mean:$MEAN,stdev:$STDEV,read_length:100,min_non_overlap:100,discordant_z:4,back_distance:20,weight:1,id:1,min_mapping_threshold:20 -sr bam_file:$INPUT.sr.bam,back_distance:20,weight:1,id:2,min_mapping_threshold:20 > $OUTPUT.pesr.bedpe.

novoBreak

All analyses were performed using novoBreak version 1.1.[22] Sorted, indexed, and duplication-free BAMs were employed for novoBreak. A control BAM file was simulated using

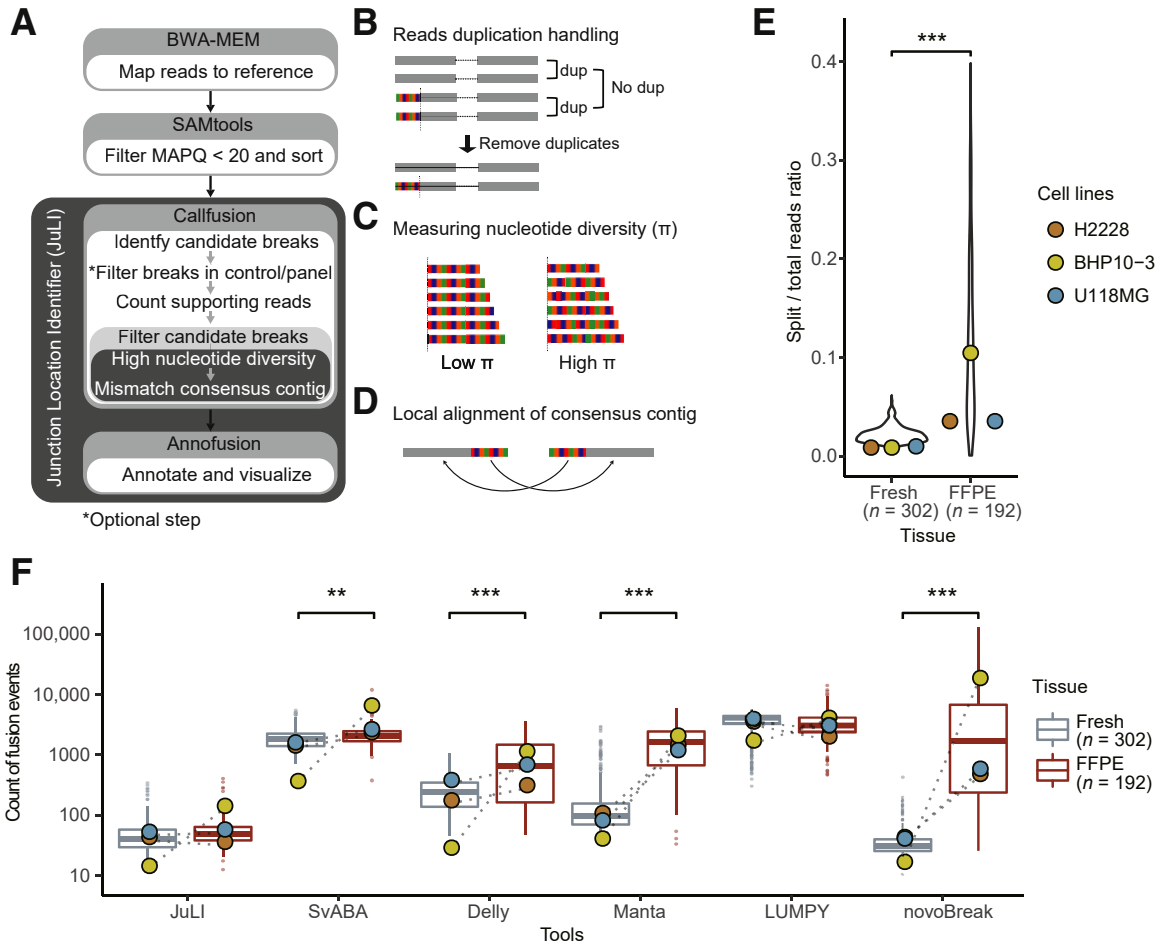

Figure 1 The fusion detection algorithm for clinical sequencing. **A:** The scheme of Junction Location Identifier (JuLI). JuLI implements novel filtering steps to reduce the number of false positives while maintaining sensitivity by fine-tuning the counting of supporting reads without duplicate removal. **B:** JuLI uses information, including the genomic positions, Concise Idiosyncratic Gapped Alignment Report strings, and read names in the Binary Alignment Map (BAM) file, to reduce redundant duplicated or noise signals. **C:** After measuring the nucleotide diversity of the breaks, JuLI filters breaks with high nucleotide diversity for the analysis. **D:** The candidate breaks are paired with each other using pair information, and split side contigs of each pair are aligned to the matched side contigs of their partners. **E:** The ratios of split read/total read counts for formalin-fixed, paraffin-embedded (FFPE), fresh clinical tissue samples, and pair cell lines tested by CancerSCAN. For FFPE samples, split reads and variability of split reads increases significantly (*t*-test, $P < 10^{-22}$). **F:** Variant counts obtained from the callers in patient samples and pair cell lines. Some callers show increasing variant counts in FFPE tissues. This phenomenon was due to the low quality of FFPE samples because of DNA degradation or damage. $n = 494$ (**E** and **F**, FFPE and fresh clinical tissue samples); $n = 3$ (**E** and **F**, pair cell lines). $^{**}P < 0.01$, $^{***}P < 0.001$. dup, duplication; MAPQ, mapping quality.

wgsim (GitHub, *http://github.com/lh3/wgsim*, last accessed October 28, 2019), and the output was used as control input to novoBreak. The command line for the analysis was as follows: $run_novoBreak.sh $novoBreak_exe_dir $reference.fa $INPUT.bam $CONTROL.bam 1 $OUTPUT_DIR.

Results

Development of a Fusion Detection Algorithm for Clinical Sequencing

Since 2014, a custom-designed panel (CancerSCAN) was used for precision oncology that covers up to 381 cancer-related genes, including introns containing frequent break-points in selected fusion genes.[5] To obtain high detection rates, a mean sequencing coverage of approximately 1200× and a target insert size of approximately 180 bp in the initial alignment were ensured. Several algorithms have been developed to improve the accuracy of our platform. For fusion detection, herein, JuLI, which was optimized for high-depth sequencing, was developed (Figure 1, A–D). JuLI uses information from both discordant and proper pair reads to detect a wide range of structural variations, including duplications, deletions, inversions, and inter-chromosomal translocations, at single-nucleotide resolution. When the fusion pipeline was tested for repeatability (three times) and reproducibility (three batches), the repeatability and the reproducibility was observed as 100% (3/3) and 100% (3/3), respectively. Generally, it is preferable to conduct high-depth sequencing with relatively short insert sizes (150 to 200 bp) to achieve high sensitivity of

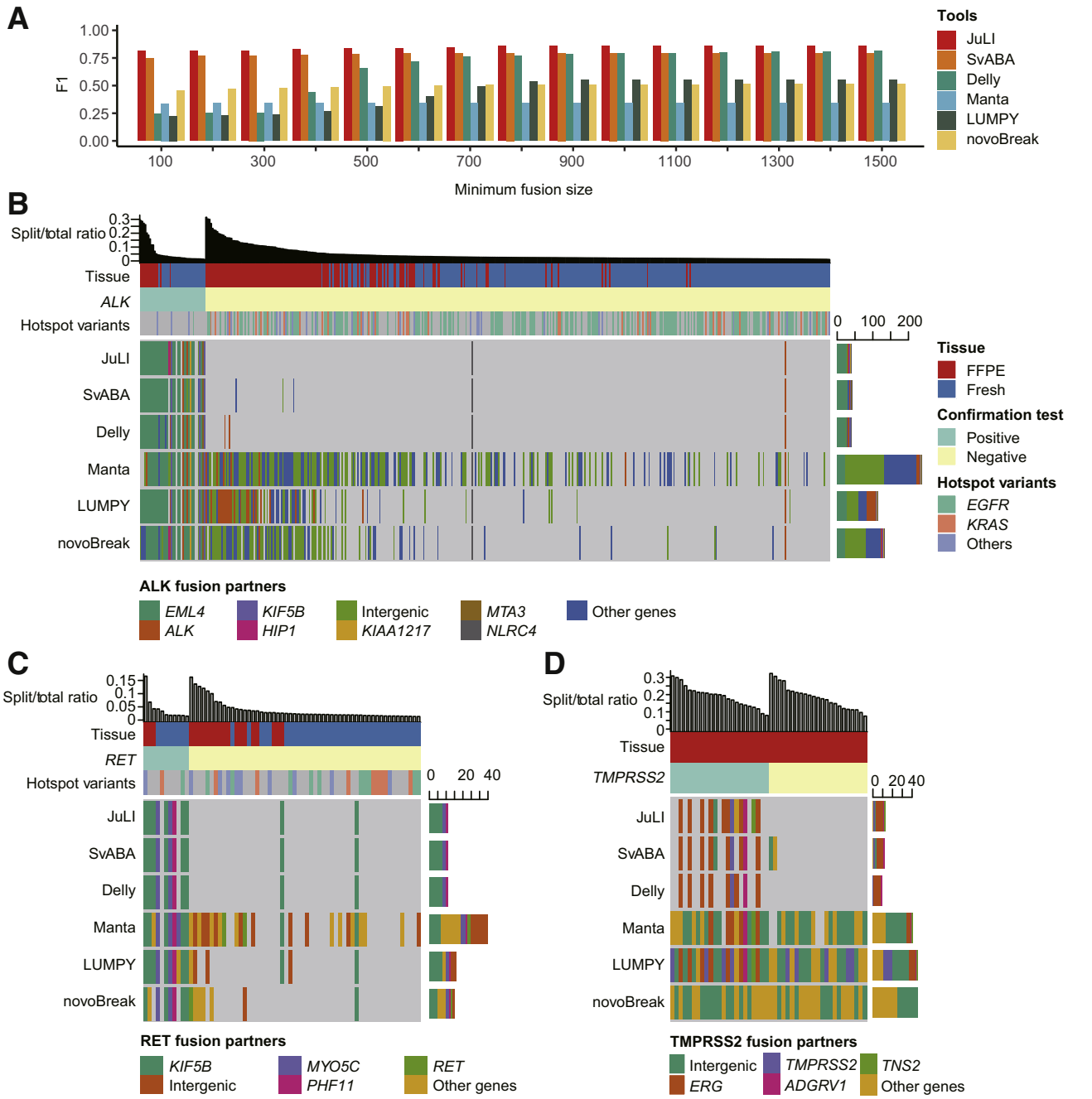

Figure 2 Validation on high-depth clinical samples. **A:** The F1 score [the harmonic average of positive predictive value (alias precision) and sensitivity (alias recall)] of the callers according to the minimum fusion length in 494 clinical samples examined by CancerSCAN. **B:** Validation on 441 non–small-cell lung cancer (NSCLC) samples with known *ALK* fusion status via immunohistochemistry (IHC) and/or fluorescence *in situ* hybridization (FISH) analyses. **C:** Validation on 67 NSCLC samples with known *RET* fusion status via IHC and/or FISH analyses. **D:** Validation on 46 prostate cancer samples with known *ERG* fusion status via IHC and/or FISH analyses. All samples were sequenced using Illumina Hi-Seq with high coverage (approximately 1300×). When two or more events in the fusion gene were detected, fusion was defined on the basis of the most supportive read counts. Hotspot mutations of NSCLCs were in *EGFR* (L858R or exon 19 insertion/deletion) or *KRAS* (G12, G13, or Q61). FFPE, formalin fixed, paraffin embedded; JuLI, Junction Location Identifier.

target-enriched sequencing in various platforms, including panel-based platforms. However, PCR duplicates generated during preprocessing for sequencing may result in overestimation of variants, and this situation may cause false-positive results that could be even worse with short insert sizes.[23] To avoid this problem, identifying duplicates using Picard[15] or SAMtools[14] is a necessary step in general bioinformatics analysis. However, because this process uses only limited information on sequence alignment map files, it is possible to unintentionally remove reads with evidence of rearrangement (Supplemental Figure S1), which may, thus, affect the sensitivity of detecting lower tumor cell content.

The reads supporting candidate breaks were carefully counted by determining duplicate fragments using Concise Idiosyncratic Gapped Alignment Report and pair locations without applying a general deduplication step (Figure 1B) (see *Materials and Methods*). Next, the candidates with sufficient supporting reads were subjected to the following two filtering steps. First, nucleotide diversity (π), which is the average number of pairwise differences between the reads and the consensus contig, was measured and the breaks with high nucleotide diversity were excluded from further processing (Figure 1C) (see *Materials and Methods*). Second, the candidate break and partner breaks were paired via pair information and compared by pairwise local alignments (Figure 1D) (see *Materials and Methods*). Through these filtering steps, fusions could be accurately detected by reducing the number of false positives.

Effects of Damaged DNA in FFPE Tissues

As mentioned above, one of the challenges in analyzing clinical samples is that FFPE tissues usually contain degraded DNA and smaller fragment sizes.[10] As a consequence, the ratio of split/total reads is substantially higher in FFPE samples than that in fresh samples (t-test, $P < 10^{-22}$) (Figure 1E). To eliminate the differences between individual samples, three pairs of fresh and routinely processed FFPE cancer cell lines were chosen for sequencing to compare tissue effects. Furthermore, differences in the split/total read ratio were also observed (Figure 1E). An increase in the numbers of split reads could affect noise in fusion analyses and may cause numerous false-positive events. To compare FFPE effects and for further analysis, several state-of-the-art fusion callers, including SvABA,[18] Delly,[19] Manta,[20] LUMPY,[21] and novoBreak,[22] that use split and discordant read information, similar to JuLI, were chosen. In the comparison of fusion event count, a significant increase in count of fusion events in FFPE tissues was observed when using SvABA, Delly, Manta, and novoBreak (Figure 1F). The count of fusion events of JuLI and LUMPY was not affected by the tissue type, but the count of LUMPY was 10 times higher than that of JuLI, regardless of the tissue type (Figure 1F). Analysis of three paired fresh and routine FFPE cancer cell line specimens revealed differences in counts of fusion events between the FFPE and fresh specimens. BHP10-3 revealed the highest change in the split/total read ratio (Figure 1E) and showed the highest difference in fusion counts using most callers (Figure 1F). Numerous split reads were observed in the FFPE specimen of BHP10-3, probably because of DNA damage during sample preparation (Supplemental Figure S2). For the tools affected by the FFPE tissues, the count of fusion events was positively correlated with split/total reads ratio (Supplemental Figure S3). However, JuLI showed the least increase in the number of fusion events with increasing split/total reads ratio (Supplemental Figure S3). Low quality of FFPE tissue can cause numerous false-positive results with most callers, but such quality issues did not significantly affect the results yielded by JuLI.

Validation of Analytical Sensitivity on Cancer Cell Lines and Patient Samples

To evaluate the accuracy of the algorithm over a wide range of tumor purity, experimental schemes designed by Frampton et al[24] were adopted. To simulate different tumor purity levels, four cancer cell lines harboring known fusions and a normal sample were manually mixed at different ratios, generating a range of expected tumor purity levels (5% to 100%) (Supplemental Table S2). All cell line specimens were profiled using CancerSCAN version 1, which targeted 83 genes. The mixed fraction of the fusions showed a high correlation [correlation coefficient (r) = 0.95] with the relative value of the normalized supporting reads (Supplemental Figure S4). JuLI, SvABA, Delly, Manta, and LUMPY achieved 100% sensitivity (32/32), but novoBreak missed one large deletion between *GOPC* and *ROS1* with 5% mix fraction in this experiment (Supplemental Table S3). In addition, 37 fusions in patients' tissues detected by JuLI with a wide range of supporting reads (range, 6 to 283) were validated by PCR to verify the estimated fusion breakpoints. The locations of all fusion sequences at the estimated breakpoints were confirmed (Table 1).

Performance Validation Using Clinical Samples

Because of the differences in the performance between callers depending on the range of fusion length,[18] the F1 score [the harmonic average of positive predictive value (PPV; alias precision) and sensitivity (alias recall)] of the callers was measured according to the minimum fusion length in 494 clinical samples examined by CancerSCAN (Figure 2A). Fusion results that are shorter than the minimum length in each caller were excluded from the comparison, and the performance comparison criteria are described in the following paragraph. In all ranges of minimum fusion size, JuLI showed better performance, particularly in PPV compared with other callers. Although JuLI and SvABA were less affected by performance over the range of fusion sizes, Delly exhibited increased performance at a relatively long length of fusion. Manta, LUMPY, and novoBreak tended to have lower PPV compared with sensitivity and a decrease in performance in predominantly FFPE tissues compared with that in fresh tissues (Supplemental Table S4). The minimum length of F1 score saturation for each caller was 800 bp for JuLI and SvABA, 1500 bp for Delly, 1900 bp for Manta, 1200 bp for LUMPY, and 1300 bp for novoBreak. To compare, except for the regions with different performance, the results were compared except for the fusions with the length <1250 bp, which is the median value of the performance saturation length of each caller.

Activation of kinase gene by chromosomal rearrangement has been identified as a recurrent driver event in NSCLCs.[25,26] *ALK* rearrangement acts as an oncogenic driver in 4% to 6% of NSCLCs.[26] In *ALK*-rearranged NSCLCs, ALK inhibitor demonstrates therapeutic efficacy in terms of improved survival, and the *EML4-ALK* variants and *ALK*-fusion partners may affect sensitivity to ALK inhibitors.[27–29] *RET* rearrangements have been identified in 1% to 2% of NSCLCs and are the potential therapeutic targets of multitargeted kinase inhibitors.[25,30] Therefore, accurate detection of an oncogenic fusion is important for clinical decision making. Over the last 4 years, CancerSCAN has been used at the oncology clinic of SMC. Performance validation was conducted in a prospective cohort of 448 patients with NSCLC, and ALK and/or RET status was profiled by immunohistochemistry and/or fluorescence *in situ* hybridization. Of the 441 patients tested for ALK, 9.5% (42/441) were positive; and 67 patients were tested for RET, of whom 16.4% (11/67) were positive (Supplemental Table S1). No patient was both ALK and RET positive, and the results of the immunohistochemistry/fluorescence *in situ* hybridization of ALK and other hotspot mutations in *EGFR* (L858R or exon 19 insertion/deletion) or *KRAS* (G12, G13, or Q61) showed a mutually exclusive pattern (Fisher exact test, $P < 10^{-11}$). A total of 79 patients were profiled using CancerSCAN version 1, which targeted 83 genes, whereas the rest were profiled using version 2, which targeted 381 genes.[5] Both V1 and V2 panels covered the same hotspot introns involved in *ALK* and *RET* rearrangement (introns of *ALK* between exons 19 and 21 and of *RET* between exons 6 and 12).

As mentioned above, fusion events that were ≥1250 bp in size were considered, and one or more break was found in the analysis of *ALK* and *RET* region. Most *ALK* and *RET* activation cases involved the rearrangement or activating mutations that activate the kinase domain; in case of NSCLC, *ALK*, and *RET* are primarily activated by fusion with various partners.[27,30,31] Therefore, intragenic rearrangements were assumed in *ALK* and *RET* as a false positive. The respective sensitivity and PPV of *ALK* fusions were as follows: JuLI, 90.5% (38/42 samples; 95% CI, 77.4%–97.3%) and 95.0% (38/40; 95% CI, 83.1%–99.4%); SvABA, 88.1% (37/42; 95% CI, 74.4%–96.0%) and 88.1% (37/42; 95% CI, 74.4%–96.0%); Delly, 88.1% (37/42; 95% CI, 74.4%–96.0%) and 90.2% (37/41; 95% CI, 76.9%–97.3%); Manta, 83.3% (35/42; 95% CI, 68.6%–93.0%) and 14.7% (35/238; 95% CI, 10.5%–19.9%); LUMPY, 88.1% (37/42; 95% CI, 74.4%–96.0%) and 32.2% (37/115; 95% CI, 23.8%–41.5%); and novoBreak, 90.5% (38/42; 95% CI, 77.4%–97.3%) and 28.6% (38/133; 95% CI, 21.1%–37.0%) (Figure 2B). For *RET* fusions, JuLI, SvABA, and Delly achieved same sensitivity and PPV [81.8% (9/11; 95% CI, 48.2%–97.7%) and 81.8% (9/11; 95% CI, 48.2%–97.7%), respectively]. The respective sensitivity and PPV of remaining callers were as follows: Manta, 90.9% (10/11; 95% CI, 58.7%–99.8%) and 28.6% (10/35; 95% CI, 14.6%–46.3%); LUMPY, 90.9% (10/11; 95% CI, 58.7%–99.8%) and 62.5% (10/16; 95% CI, 35.4%–84.8%); and novoBreak, 72.7% (8/11; 95% CI, 39.0%–94.0%) and 53.3% (8/15; 95% CI, 26.6%–78.7%) (Figure 2C). Six samples that yielded false-negative results of *ALK* and *RET* in JuLI analysis also tested negative in most callers, and the tumor purity of these samples was significantly lower than that of the test-positive samples (Supplemental Figure S5). Therefore, some false negatives may be due to low tumor purity. Four false positives of *ALK* and *RET* identified in JuLI results were observed in all other callers, and the fusions were clearly identified in browser view (Supplemental Figure S6).

To further compare other clinically significant fusions, 46 archived prostate cancer samples were retrospectively collected and analysis of *ERG* fusion status was performed by immunohistochemistry and/or fluorescence *in situ* hybridization. Of the 46 patients, 23 (50.0%) were *ERG* fusion positive (Supplemental Table S1). All patients with prostate cancer were profiled using CancerSCAN version 1, and the panel covered the hotspot introns between exons 1 and 6 of *TMPRSS2*, the most common fusion partner of *ERG* fusion.[32] The performance of the callers with the same criteria as those of NSCLC was measured. The respective sensitivity and PPV of *ERG* fusions were as follows: JuLI, 56.5% (13/23; 95% CI, 34.5%–76.8%) and 100.0% (13/13; 95% CI, 75.3%–100%); SvABA, 43.5% (10/23; 95% CI, 23.2%–65.5%) and 83.3% (10/12; 95% CI, 51.6%–97.9%); Delly, 39.1% (9/23; 95% CI, 19.7%–61.5%) and 100.0% (9/9; 95% CI, 66.4%–100%); Manta, 95.7% (22/23; 95% CI, 78.1%–99.9%) and 53.7% (22/41; 95% CI, 37.4%–69.3%); LUMPY, 100.0% (23/23; 95% CI, 85.2%–100%) and 50.0% (23/46; 95% CI, 34.9%–65.1%); and novoBreak, 100.0% (23/23; 95% CI, 85.2%–100%) and 50.0% (23/46; 95% CI, 34.9%–65.1%) (Figure 2D). There was no difference in purity distribution between true positive and false negative of JuLI. The relatively low sensitivity of this retrospective set may be due to other partners of *ERG* that were not targeted.[33] Overall, the number of false calls occurred as the split/total read ratio increased, but this issue had less effect in JuLI.

Performance Based on Sequencing Coverage

For panel-based high-throughput sequencing in clinical practice, test performance must at least be comparable to conventional molecular tests. The factors that constitute sufficient sequencing depth are influenced by tumor purity and clonality of variants as well as other characteristics of a patient's tumor sample, including tissue preparation methods and sequencing platforms used. Lowering tumor purity reduces detection sensitivity by proportionally reducing the effective range of mutant alleles in tumor cells. In contrast to research samples, requirements for sufficient tumor purity for clinical specimens may not be met; therefore, it is important to have adequate coverage.[5]

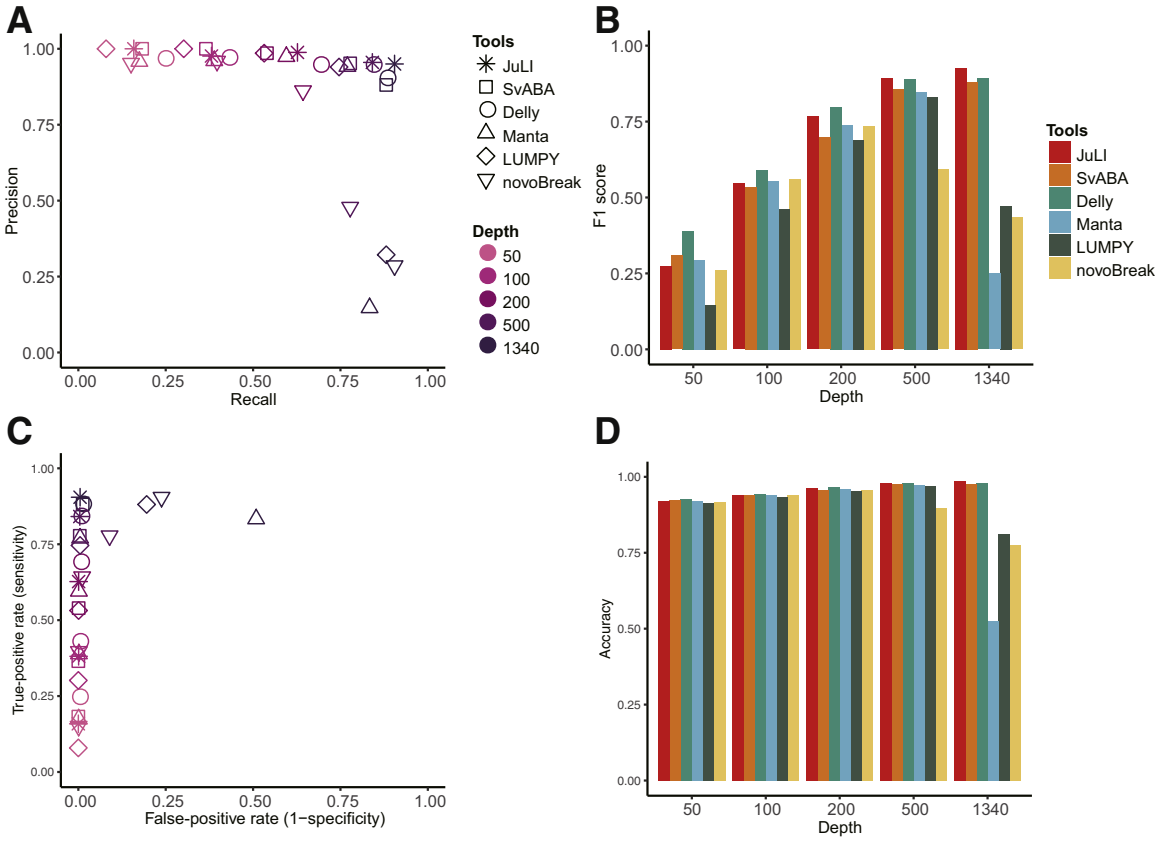

Figure 3 The effect of depth on fusion detection in clinical samples. **A:** Positive predictive value (PPV; alias precision) and sensitivity (alias recall) based on *in silico* down-sampling experiments. A total of 441 non–small-cell lung cancer (NSCLC) samples were down sampled from the original depth (1340×), and the average performance was measured at each depth (three iterations). **B:** The F1 score, which is the harmonic average of PPV and sensitivity, on the basis of the coverage change. **C:** The receiver operating characteristic curve. **D:** Accuracy, which is the proportion of true results (both true positives and true negatives), among all the results. JuLI, Junction Location Identifier.

In silico down-sampling experiments (three iterations) were conducted using the *ALK* set of NSCLCs as an alternative method for investigating the effect of sequencing depth on performance. In down-sampling experiments, the average sensitivity of all of the callers improved with increased coverage (Figure 3, A and C). Although sensitivity increased at high coverage, PPV and specificity decreased in Manta, LUMPY, and novoBreak. In contrast, JuLI, SvABA, and Delly were less affected by coverage. The F1 score improved in the 50× to 200× range, which is the typical range used in WGS or whole-exome sequencing in all callers (Figure 3B). By contrast, the F1 score worsened at high depth (1340×) in Manta, LUMPY, and novoBreak, but not in others. Accuracy, which is the proportion of true results (both true positives and true negatives) among all of the results, was also maintained at a high-depth range in JuLI, SvABA, and Delly in contrast to the other callers (Figure 3D). Thus, it was confirmed that sensitivity improved with increased sequencing depth; however, accuracy may decrease owing to increased noise levels above the threshold in some algorithms. To effectively apply high-throughput sequencing in a clinical setting, it is necessary to use software optimized to reduce such noise.

Sensitivity Validation Using WGS Data

To estimate the sensitivity based on real WGS data, raw FASTQ data of NA12878 (*ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147.fastq.gz*) were downloaded from European Nucleotide Archive (*https://www.ebi.ac.uk/ena/data/view/ERA172924*, last accessed August 8, 2019). These data represent approximately 50× coverage, which has been widely used by tools for the estimation of a variety of variation tools. The results made by each tool were compared with the truth set by Layer et al,[21] who developed LUMPY in 2014. They provided a truth set containing 4095 deletions detected by at least one tool in the 50× data set that were validated by split-read mapping analysis of independent long-read sequencing data from PacBio or Illumina platforms. In this comparison, LUMPY (47.4%; 1942/4095; 95% CI, 45.9%–49.0%) was the most sensitive, followed by JuLI (41.9%; 1717/4095; 95% CI, 40.4%–43.5%), SvABA (41.9%; 1716/4095; 95% CI, 40.4%–43.4%), Delly (38.5%; 1575/4095; 95% CI, 37.0%–40.0%), Manta (37.9%; 1552/4095; 95% CI, 36.4%–39.4%), and novoBreak (37.7%; 1542/4095; 95% CI, 36.2%–39.2%). The performance of JuLI was maintained as much as other

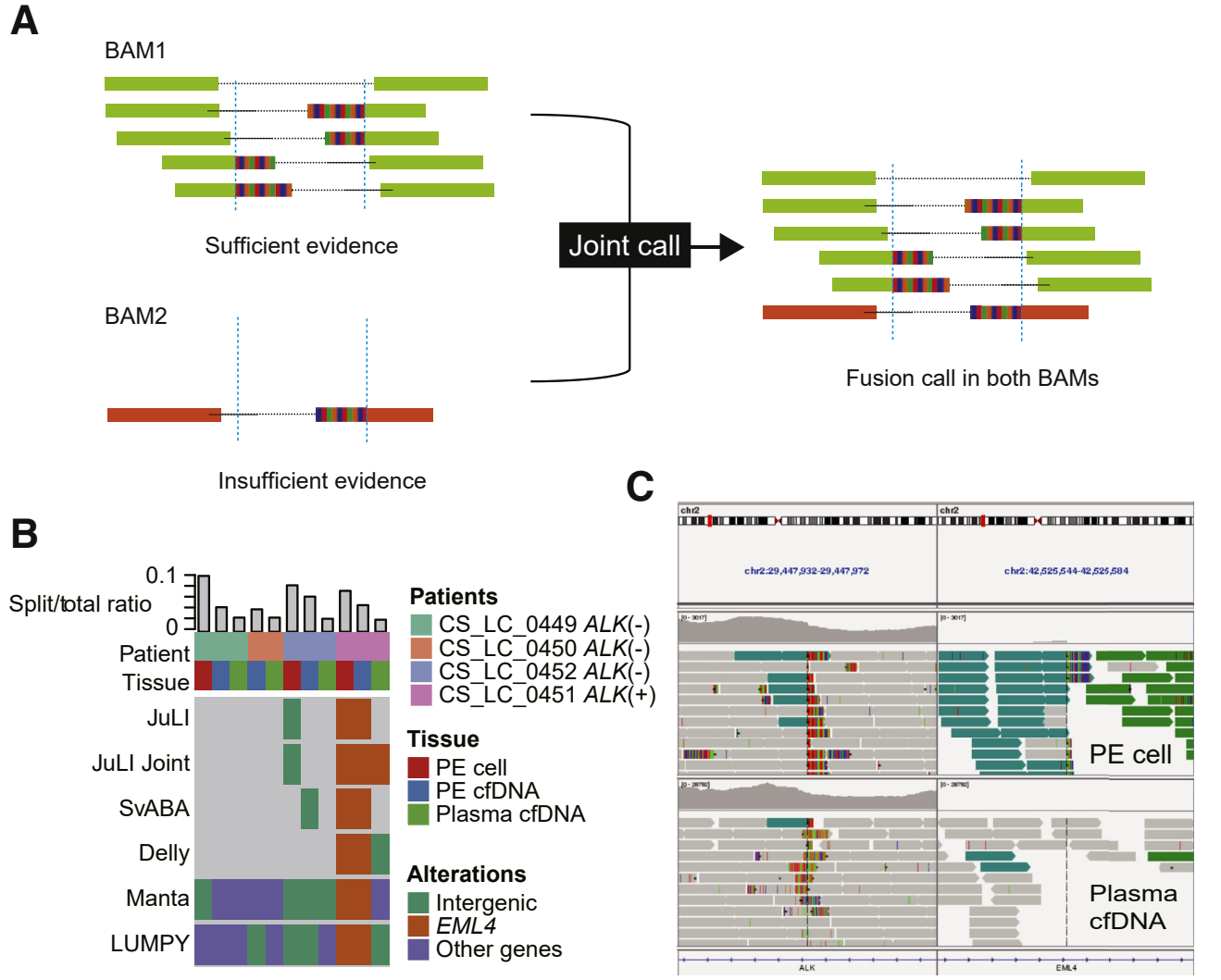

Figure 4 The joint call function to detect fusions with low supporting evidence in serial/multiregion sampling tissues. **A:** The joint call function combines information from multiple Binary Alignment Map (BAM) files and produces the individual result of the BAM files. **B:** Junction Location Identifier (JuLI) with the joint call function rescues the *EML4-ALK* fusion of CS_LC_0451 from false negative in plasma cell-free DNA (cfDNA). **C:** Only two discordant reads supporting the *ALK* fusion are observed in plasma cfDNA of CS_LC_0451. chr., chromosome; PE, pleural effusion.

callers, even at a low depth, such as WGS; however, specificity or PPV, representing the frequency of false-positive calls, was not evaluated because of the lack of true-negative reference.

Joint Call to Detect Fusions with Insufficient Evidence

Recent reports have shown that the detection of *ALK* fusion in cfDNA is feasible in clinic setting.[34,35] Serial tumor sampling on progression has been helpful in determining the optimal subsequent treatment decision making for patients. However, this is often complicated by insufficient tumor purity for molecular analysis and tumor heterogeneity.[36] If a more sensitive detection is possible in a series of samples with insufficient supporting reads, slightly earlier decision making can be made for precise medicine. Genotyping known events in one set of samples and placing higher confidence in a genotyped event that was already detected in a related set of samples are a common strategy for variant calling.[16,17] To achieve more sensitive fusion detection in clinical sequencing, the joint call strategy in JuLI that can detect fusions with low supporting evidence in serial/multiregion sampling tissues was also implemented (Figure 4A). To verify the performance of this function, *in silico* down-sampling experiments (mean coverage: 1×, 5×, and 10×, with 100 iterations) were performed on mixed cell lines with relatively low cell ratios of 5% to 40% (Supplemental Table S3). In this simulation, most callers showed up to 1% to 2% sensitivity at 10×, and joint call showed 40.6% sensitivity at 10× and could detect 7.3% even at 1× (Supplemental Table S5).

To confirm the utility of the joint call function of JuLI in clinic, it was applied to *ALK* detection in cfDNA of NSCLCs. Pleural effusion and peripheral plasma were

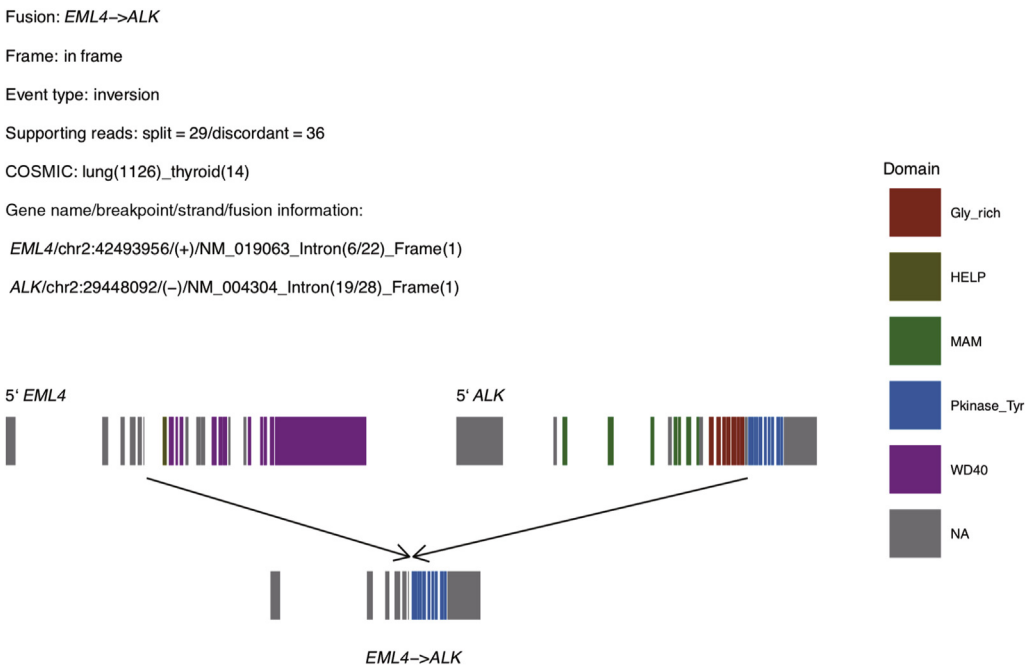

Figure 5 A representation showing Junction Location Identifier output with annotation and visualization. Annotation of fusions and a graphically visualized fusion diagram with the domain status in PDF format. COSMIC, Catalogue of Somatic Mutations in Cancer; HELP, hydrophobic echinoderm microtubule—associated protein—like protein; Gly_rich, glycine rich protein; MAM, meprin, A-5 protein, and receptor protein-tyrosine phosphatase mu; NA, not available; Pkinase_Tyr, protein tyrosine kinase; WD40, Trp-Asp 40 (*WD40*) repeats.

collected from four patients with NSCLCs, whose ALK status was confirmed in primary tissue (one positive and three negative) (Supplemental Table S1). DNA of cells in pleural effusion, cfDNA of pleural effusion, and cfDNA of plasma were processed using LiquidSCAN (average coverage, approximately 4300×) and analyzed using the fusion callers. novoBreak was excluded from this analysis because some samples did not show any results in the ultra high-depth data. In this analysis, the callers, except for JuLI with joint calling, did not detect the *EML4-ALK* fusion in plasma cfDNA of the *ALK*-positive patient (CS_LC_0451); JuLI with the joint calling option was able to identify the fusion (Figure 4B). There were only two discordant reads supporting the *ALK* fusion in plasma cfDNA of CS_LC_0451 (Figure 4C).

Annotation and Visualization

Accurate annotation of rearrangements is critical for clinical decision making. JuLI annotates functional consequences of genomic fusions that are identified using high-throughput sequencing data in a strand-specific manner (Supplemental Figure S7). Even with breaks at the same location, this annotation approach allows the user to easily distinguish between positive and negative strand events. Moreover, JuLI provides three useful pieces of information. First, JuLI predicts whether the fusion transcript is in frame or out of frame by means of the University of California, Santa Cruz database.[37] Second, JuLI provides the frequency of fusion events based on cancer types in Catalogue of Somatic Mutations in Cancer (COSMIC).[38] Third, JuLI annotates chimera protein domains via the UniProt[39] and Pfam databases.[40] Graphically visualized fusion diagrams were automatically generated in PDF format, showing all annotation results (Figure 5).

Comparison of Running Time between Paired FFPE and Fresh Tissues

Sequencing data were generated from paired FFPE and fresh cancer cell lines using CancerSCAN V1; and of these cell lines, BHP10-3 showed the highest change in the split/total read ratio (Figure 1E). To compare elapsed time, BHP10-3 pair analysis time of each caller was measured with 10 iterations (Supplemental Figure S8). A 1.1- to 29.3-fold increase was observed in analysis time in low-quality FFPE tissue with these callers. Although JuLI was relatively slow in low-quality FFPE tissue because it implements several steps to improve accuracy, the speed can be increased through parallel processing across multiple cores.

Discussion

To implement precision medicine at SMC, JuLI, a novel fusion detection algorithm optimized for clinical application, was developed. Using the assembled sequences and calling fusions with additional information are one of the common strategies among various variant callers. JuLI

implements novel filtering steps along with this strategy to accurately detect somatic fusions in clinical samples. The tool was validated on four cancer cell lines and on 505 clinical tumor specimens. JuLI has several characteristics. First, with the implementation of the noise reduction algorithm to minimize false-positive calls, it maintains good analytical specificity without loss of sensitivity, particularly in noisy samples, such as FFPE samples. Second, JuLI can detect fusions with insufficient evidence in serial or multiregion sequencing samples by using the joint call function. Third, JuLI is easy to use with the provided R package, which is available via GitHub and supports comprehensive annotation and visualization of structural variations.

An intriguing point is that the joint call strategy can be used for monitoring cancer in specimens such as cfDNA, blood samples of minimal residual leukemic cell follow-up, and follow-up biopsy specimen without ideal tumor purity.[5] Split reads originating from the primary tumor have high specificity in the location of fusion junction and adjacent DNA sequences, with uniqueness of split portion. Sensitive calling for these predefined fusion signals in follow-up specimens provided a good chance for early detection of relapsing cancer with good specificity.

Clinically, quantitative evaluation of fusion transcript is emphasized in follow-up of some cancers, particularly for chronic myeloid leukemia. Discontinuation of tyrosine kinase inhibitors is suggested in recent National Comprehensive Cancer Network guidelines for chronic myeloid leukemia, and the practice is performed on the basis of quantitative evaluation of the *BCR-ABL1* transcript. Despite the normalization, RNA assay does not provide direct information on the number of malignant clones. However, ultrahigh coverage next-generation sequencing for DNA fusion could provide direct information on the number of remnant malignant clones in future precision medicine.

For accurate detection of tumor-driving fusions, it is assumed to be necessary to detect fusions in both RNA and DNA. RNA is a suitable material for directly detecting chimeric transcripts; however, quality may be compromised because of long storage time or degradation during FFPE preparation.[41] Detection sensitivity for fusions may be maximized by simultaneously performing DNA and RNA assays. Furthermore, combined fusion analysis for DNA and RNA can help identify loss of function of a tumor suppressor gene via fusion or complex fusions involving noncoding regions.

The limitation of this study is that a limited number of fusion events were tested for performance validation. Most callers used in this study for comparison reported their results through a genome-wide comparison of several samples in their articles. However, quality (ie, condition)-wide comparison of 505 patient samples with three clinically important fusion events and four cancer cell lines with known fusions was performed. In a clinical setting, it may be inevitable to examine tissues with inappropriate quality (low tumor purity or poor-quality FFPE). Therefore, these results could provide useful information to select callers in a clinical setting.

In clinical setting, although sensitivity is important, maintaining PPV is also essential to reduce the number of false positives. In particular, if the prevalence is relatively low, such as that of the *ALK* fusion in NSCLC (2% to 7%),[29] several wrong decisions can be made when PPV is not guaranteed. Clinical decisions based on false test results are risky and may lead to inappropriate treatment strategies. Therefore, if the PPV cannot provide a sufficiently high confidence level, it will be difficult to use the method for diagnostic purposes. Because JuLI has better PPV relative to the existing algorithms, it is likely to deliver accurate fusion profiling data to help clinicians to make optimal therapeutic decisions.

Acknowledgments

We thank Jungwook Park and Geunhan Jeong for critical advice on the tool development.

Author Contributions

H.-T.S., S.K., and K.-W.L. performed bioinformatic analysis, with guidance from W.-Y.P.; H.-T.S., N.K.D.K., and J.W.Y. wrote the manuscript, with substantial input from J.K., J.S.B., and D.P.; H.-T.S., N.K.D.K., B.L., and D.R. developed the tool.; Y.-L.C., S.-H.L., M.-J.A., and K.P. provided clinical data.; and all authors reviewed the manuscript.

Supplemental Data

Supplemental material for this article can be found at *https://doi.org/10.1016/j.jmoldx.2019.10.015*.

References

1. Lee SH, Lee B, Shim JH, Lee KW, Yun JW, Kim SY, Kim TY, Kim YH, Ko YH, Chung HC, Yu CS, Lee J, Rha SY, Kim TW, Jung KH, Im SA, Moon HG, Cho S, Kang JH, Kim J, Kim SK, Ryu HS, Ha SY, Kim JI, Chung YJ, Kim C, Kim HL, Park WY, Noh DY, Park K: Landscape of actionable genetic alterations profiled from 1,071 tumor samples in Korean cancer patients. Cancer Res Treat 2019, 51:211−222
2. Allegretti M, Fabi A, Buglioni S, Martayan A, Conti L, Pescarmona E, Ciliberto G, Giacomini P: Tearing down the walls: FDA approves next generation sequencing (NGS) assays for actionable cancer genomic aberrations. J Exp Clin Cancer Res 2018, 37:47
3. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL: Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001, 344: 1031−1037
4. Awad MM, Shaw AT: ALK inhibitors in non-small cell lung cancer: crizotinib and beyond. Clin Adv Hematol Oncol 2014, 12:429−439
5. Shin HT, Choi YL, Yun JW, Kim NKD, Kim SY, Jeon HJ, Nam JY, Lee C, Ryu D, Kim SC, Park K, Lee E, Bae JS, Son DS, Joung JG,

Lee J, Kim ST, Ahn MJ, Lee SH, Ahn JS, Lee WY, Oh BY, Park YH, Lee JE, Lee KH, Kim HC, Kim KM, Im YH, Park K, Park PJ, Park WY: Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun 2017, 8:1377

6. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al: Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017, 9:eaan2415
7. Christensen E, Nordentoft I, Vang S, Birkenkamp-Demtroder K, Jensen JB, Agerbaek M, Pedersen JS, Dyrskjot L: Optimized targeted sequencing of cell-free plasma DNA from bladder cancer patients. Sci Rep 2018, 8:1917
8. Oellerich M, Schutz E, Beck J, Kanzow P, Plowman PN, Weiss GJ, Walson PD: Using circulating cell-free DNA to monitor personalized cancer therapy. Crit Rev Clin Lab Sci 2017, 54:205−218
9. Araujo LH, Timmers C, Shilo K, Zhao W, Zhang J, Yu L, Natarajan TG, Miller CJ, Yilmaz AS, Liu T, Amann J, Lapa ESJR, Ferreira CG, Carbone DP: Impact of pre-analytical variables on cancer targeted gene sequencing efficiency. PLoS One 2015, 10: e0143092
10. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ: Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn 2013, 15:623−633
11. Evers DL, He J, Kim YH, Mason JT, O'Leary TJ: Paraffin embedding contributes to RNA aggregation, reduced RNA yield, and low RNA quality. J Mol Diagn 2011, 13:687−694
12. Park G, Park JK, Son DS, Shin SH, Kim YJ, Jeon HJ, Lee J, Park WY, Lee KH, Park D: Utility of targeted deep sequencing for detecting circulating tumor DNA in pancreatic cancer patients. Sci Rep 2018, 8:11631
13. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754−1760
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; Genome Project Data Processing Subgroup: The sequence alignment/map format and SAMtools. Bioinformatics 2009, 25:2078−2079
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297−1303
16. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E: Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 2018. https://doi.org/10.1101/201178
17. Cameron DL, Schroder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT: GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 2017, 27:2050−2060
18. Wala JA, Bandopadhayay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, Schumacher S, Li Y, Weischenfeldt J, Yao X, Nusbaum C, Campbell P, Getz G, Meyerson M, Zhang CZ, Imielinski M, Beroukhim R: SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res 2018, 28:581−591
19. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012, 28:i333−i339
20. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT: Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 2016, 32:1220−1222
21. Layer RM, Chiang C, Quinlan AR, Hall IM: LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 2014, 15: R84
22. Chong Z, Ruan J, Gao M, Zhou W, Chen T, Fan X, Ding L, Lee AY, Boutros P, Chen J, Chen K: novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods 2017, 14: 65−67
23. Zhou W, Chen T, Zhao H, Eterovic AK, Meric-Bernstam F, Mills GB, Chen K: Bias from removing read duplication in ultra-deep sequencing experiments. Bioinformatics 2014, 30:1073−1080
24. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al: Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013, 31:1023−1031
25. Pan Y, Zhang Y, Li Y, Hu H, Wang L, Li H, Wang R, Ye T, Luo X, Zhang Y, Li B, Cai D, Shen L, Sun Y, Chen H: ALK, ROS1 and RET fusions in 1139 lung adenocarcinomas: a comprehensive study of common and fusion pattern-specific clinicopathologic, histologic and cytologic features. Lung Cancer 2014, 84:121−126
26. Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, Lim Choi Y, Satoh Y, Okumura S, Nakagawa K, Mano H, Ishikawa Y: RET, ROS1 and ALK fusions in lung cancer. Nat Med 2012, 18: 378−381
27. Noh KW, Lee MS, Lee SE, Song JY, Shin HT, Kim YJ, Oh DY, Jung K, Sung M, Kim M, An S, Han J, Shim YM, Zo JI, Kim J, Park WY, Lee SH, Choi YL: Molecular breakdown: a comprehensive view of anaplastic lymphoma kinase (ALK)-rearranged non-small cell lung cancer. J Pathol 2017, 243:307−319
28. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou SH, Dezube BJ, Janne PA, Costa DB, Varella-Garcia M, Kim WH, Lynch TJ, Fidias P, Stubbs H, Engelman JA, Sequist LV, Tan W, Gandhi L, Mino-Kenudson M, Wei GC, Shreeve SM, Ratain MJ, Settleman J, Christensen JG, Haber DA, Wilner K, Salgia R, Shapiro GI, Clark JW, Iafrate AJ: Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010, 363:1693−1703
29. Shaw AT, Kim DW, Nakagawa K, Seto T, Crino L, Ahn MJ, De Pas T, Besse B, Solomon BJ, Blackhall F, Wu YL, Thomas M, O'Byrne KJ, Moro-Sibilot D, Camidge DR, Mok T, Hirsh V, Riely GJ, Iyer S, Tassell V, Polli A, Wilner KD, Janne PA: Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. N Engl J Med 2013, 368:2385−2394
30. Lee SE, Lee B, Hong M, Song JY, Jung K, Lira ME, Mao M, Han J, Kim J, Choi YL: Comprehensive analysis of RET and ROS1 rearrangement in lung adenocarcinoma. Mod Pathol 2015, 28:468−479
31. Hallberg B, Palmer RH: Mechanistic insight into ALK receptor tyrosine kinase in human cancer biology. Nat Rev Cancer 2013, 13: 685−700
32. Barros-Silva JD, Paulo P, Bakken AC, Cerveira N, Lovf M, Henrique R, Jeronimo C, Lothe RA, Skotheim RI, Teixeira MR: Novel 5' fusion partners of ETV1 and ETV4 in prostate cancer. Neoplasia 2013, 15:720−726
33. Cancer Genome Atlas Research Network: The molecular taxonomy of primary prostate cancer. Cell 2015, 163:1011−1025
34. Thompson JC, Yee SS, Troxel AB, Savitch SL, Fan R, Balli D, Lieberman DB, Morrissette JD, Evans TL, Bauml J, Aggarwal C, Kosteva JA, Alley E, Ciunci C, Cohen RB, Bagley S, Stonehouse-Lee S, Sherry VE, Gilbert E, Langer C, Vachani A, Carpenter EL: Detection of therapeutically targetable driver and resistance mutations in lung cancer patients by next-generation sequencing of cell-free circulating tumor DNA. Clin Cancer Res 2016, 22: 5772−5782
35. Paweletz CP, Sacher AG, Raymond CK, Alden RS, O'Connell A, Mach SL, Kuang Y, Gandhi L, Kirschmeier P, English JM, Lim LP, Janne PA, Oxnard GR: Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients. Clin Cancer Res 2016, 22:915−922

36. Dagogo-Jack I, Brannon AR, Ferris LA, Campbell CD, Lin JJ, Schultz KR, Ackil J, Stevens S, Dardaei L, Yoda S, Hubbeling H, Digumarthy SR, Riester M, Hata AN, Sequist LV, Lennes IT, Iafrate AJ, Heist RS, Azzoli CG, Farago AF, Engelman JA, Lennerz JK, Benes CH, Leary RJ, Shaw AT, Gainor JF: Tracking the evolution of resistance to ALK tyrosine kinase inhibitors through longitudinal analysis of circulating tumor DNA. JCO Precis Oncol 2018, 2018
37. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. Genome Res 2002, 12:996−1006
38. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015, 43:D805−D811
39. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004, 32:D115−D119
40. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: The Pfam protein families database. Nucleic Acids Res 2010, 38:D211−D222
41. Ludyga N, Grunwald B, Azimzadeh O, Englert S, Hofler H, Tapio S, Aubele M: Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. Virchows Arch 2012, 460: 131−140