ONLINE METHODS

DNA preparation, DNA capture and sequencing. The tissues and clinical information used in this study were obtained under informed consent and approval of the institutional review boards of each institute. DNA was extracted from liver cancer tissue and matched non-cancerous liver tissues or blood using a general protocol for genome sequencing. Exome capture was carried out using the SureSelect Human All Exon V3 or V4 plus kit depending on the samples (**Supplementary Table 28**). Preparation of sequencing libraries, DNA capture methods and Illumina sequencing were carried out as described in the **Supplementary Note**.

Mutation calling. *Mutation calling (National Cancer Center Research Institute).* Paired-end reads were aligned to the human reference genome (GRCh37) using the Burrows-Wheeler Aligner (BWA)[40] for both tumor and normal samples. Probable PCR duplications, for which paired-end reads aligned to the same genomic position, were removed, and pileup files were generated using SAMtools[41] and a program developed in house. Details on our filtering conditions are provided in **Supplementary Tables 29** and **30**.

Mutation calling (Research Center for Advanced Science and Technology). Next-generation sequencing reads were mapped to the human genome (hg19) using BWA and Novoalign independently. Reads with a minimal editing distance to the reference genome were taken to represent optimal alignments. Then, bam files were locally realigned with SRMA. Normal-tumor pair bam files were processed using an in-house genotyper (karkinos), with the variants further filtered to remove all variants observed fewer than four times or present at an allele frequency of less than 0.12 after adjustment for tumor sample purity. The variants also had to have a score of greater than Q20 (representing the root mean square of mapping quality). In addition, reads harboring the variant had to be observed in both forward and reverse orientation. If a variant was present in reads of only one orientation, we checked for strand bias using a *t* test comparing these reads to the reads without the variant, and variants with a *P* value of <0.03 for strand bias were rejected. Variants also had to be called in different sequence cycles and have at least one call that was outside of 3% of read ends. Variants could not be located within 5 bp of an indel call, and variants where the mean base quality of the supporting reads was lower than 10 on the Phred scale were removed. Germline variants having an allelic frequency of greater than 0.1 were collected for 50 normal liver exome samples and used as the panel of normal variants. Any variant that was observed in this panel with a population frequency of greater than 2% was filtered out. Finally, variants also observed in the paired normal sample with an allelic frequency of greater than 3% and sites registered in dbSNP Build 134 with validated status were removed.

Mutation calling (Baylor College of Medicine). Initial sequence analysis was performed using the Human Genome Sequencing Center (HGSC) Mercury analysis pipeline. First, the primary analysis software on the instrument produced bcl files that were transferred off the instrument to the HGSC analysis infrastructure by the HiSeq Real-Time Analysis module. Once each run was complete and all bcl files were transferred, Mercury ran the vendor's primary analysis software (CASAVA), which demultiplexed pooled samples and generated sequence reads and base call confidence values (qualities). In the next step, reads were mapped to the GRCh37 human reference genome using BWA (BWA2), producing a bam3 (binary alignment/map) file. The third step involved quality recalibration (using GATK4) and, where necessary, the merging of bam files for separate sequence events into a single sample-level bam file. Sorting of bam files, duplicate read marking and realignment to improve indel discovery all occurred at this step.

Processing the significantly mutated genes. The significantly mutated genes for this study were identified through three separate tests as described below (an aggregated somatic alteration method, MutSigCV[42] and an inactivation bias method), and the resulting gene lists were combined in a final table of significantly mutated genes (**Supplementary Table 13**).We also developed two tests to detect bias in the mutation list that could be a source of artifact (K.R.C., E.S., L.A.D. and D.A.W., unpublished data). One of these tests examined sequencing center bias, and the other examined bias in mutation allele fraction, which if consistently low would suggest that a gene was a passenger rather than a driver. Genes in the final combined table that failed these bias tests were removed from the final list of significantly mutated genes. Data from each process are shown in **Supplementary Tables 7–12**, and the steps are shown schematically in **Supplementary Figure 16**.

Aggregated somatic alteration method. We identified significantly altered genes by aggregating somatic substitutions, short indels, homozygous deletions and focal amplifications. We initially estimated the expected number of each alteration in each gene as follows.

First, the substitution rate was estimated by dividing the number of synonymous mutations in a sample by the number of synonymous sites in the genome. For each gene, the expected number of substitutions was calculated by multiplying the substitution rate by the number of nonsynonymous sites and splice sites in the gene. Because the substitution rate at CpG sites was much higher than that in other regions, the substitution rates and expected numbers of substitutions at CpG and non-CpG sites were estimated separately using the following equation:

$$\mathrm{EN} = \sum_{i=1}^{n}\left(\frac{M_{\mathrm{CG}_i} \times N_{\mathrm{CG}}}{S_{\mathrm{CG}} \times C_i} + \frac{M_{\mathrm{NCG}_i} \times N_{\mathrm{NCG}}}{S_{\mathrm{NCG}} \times C_i}\right)$$

where n is the number of samples, M_{CG_i} is the number of synonymous mutations at CpG sites in the ith sample, M_{NCG_i} is the number of synonymous mutations in non-CpG sites in the ith sample, S_{CG} is the number of synonymous sites at CpG sites in the genome, S_{NCG} is the number of synonymous sites at non-CpG sites in the genome, N_{CG} is the number of nonsynonymous sites and splice sites at CpG sites in a gene, N_{NCG} is the number of nonsynonymous sites and splice sites at non-CpG sites in a gene, C_i is the fraction of sequence coverage in the genome in the ith sample (usually the fraction of coding regions that have more than 20× sequence depth for whole-exome sequencing) and EN is the expected number of nonsynonymous and splice-site substitutions in a gene.

Second, the coding indel rate was estimated by dividing the number of coding indels in a sample by the number of coding sites in the genome. For each gene, the expected number was calculated by multiplying the coding indel rate by the coding length of a gene as follows:

$$\mathrm{EI} = \sum_{i=1}^{n}\frac{I_i \times L}{S \times C_i}$$

where I_i is the number of coding indels in the ith sample, S is the number of coding sites in the genome, L is the coding length of the gene and EI is the expected number of coding indels in a gene.

Third, as regions of focal amplification and homozygous deletion are much broader than gene regions, the number of focal amplifications and homozygous deletions affecting a gene in a sample is 0 or 1 and is not influenced by gene length. Therefore, the expected number of these events is the same for all genes. The expected numbers of focal amplifications and homozygous deletions were estimated separately by dividing the total length of the focal amplification or homozygous deletion region in a sample by the length of the genome as follows:

$$\mathrm{EA} = \sum_{i=1}^{n}\frac{A_i}{G \times C_i}$$

$$\mathrm{ED} = \sum_{i=1}^{n}\frac{D_i}{G \times C_i}$$

where A_i is the total length of focal amplifications in the ith sample, D_i is the total length of homozygous deletions in the ith sample, G is the length of the genome, EA is the expected number of focal amplifications in the gene and ED is the expected number of homozygous deletions in the gene.

Fourth, the expected number of protein-altering mutations was calculated by aggregating the expected numbers of nonsynonymous and splice-site substitutions in CpG and non-CpG sites, coding indels, focal amplifications and homozygous deletions as follows:

$$\mathrm{E} = \mathrm{EN} + \mathrm{EI} + \mathrm{EA} + \mathrm{ED}$$

where E is the expected number of protein-altering mutations in a gene.

Fifth, tests of the significance of each gene were performed by assuming a Poisson distribution of mutation number. Adjustment for multiple testing was performed using the Benjamini-Hochberg method[8].

Inactivation bias method. The number of missense mutations was compared to the number of inactivating mutations (nonsense, frameshift and splice site) using a χ^2 test.

Analysis of sequencing center bias. Because multiple centers participated in this study, we sought to control for the influence of differences in mutation calling strategy, which might promote a gene to significance merely because of a bias in the variant callers used. Many studies do not use multiple callers and therefore have no way to control for these biases. For each gene with more than five variants, we counted the number of subjects for whom the gene was called for each center. These counts were compared to the total number of subjects using the χ^2 test. The results of the analysis for center bias are presented in **Supplementary Table 11**.

Analysis of subclone bias. Oncogenic driver events in a given tumor should exhibit allele fractions that are roughly the same as the mean allele fraction for the entire sample for any given subject. We separated oncogenic (driver) events from recurrent passenger events by comparing the allele fraction of mutations in candidate genes to the matched mean allele fraction of the sample, across all samples in the cohort. First, the mean somatic allele fraction was calculated for each subject (AF*s*). Next, for each variant in each gene, the allele fraction for the variant (AF*g*) was compared to the AF*s* in the respective subject. We calculated the fraction of events where AF*g* was less than AF*s* and generated a *P* value using a one-sided pairwise Wilcoxon test where the alternative hypothesis was that AF*g* was less than AF*s* (always with respect to the relevant subject). The histogram of all allele fraction biases (sum(AF*g* < AF*s*)/*n*, where *n* is variant count) is shown in **Supplementary Figure 30**. Selected significantly mutated genes are plotted individually to show how known drivers are distributed by this test. Note that several tumor-suppressor genes exhibited enrichment above the average allele fractions (for example, *RB1* and *TP53*). In these cases, the genes were typically both mutated and underwent loss of heterozygosity (LOH) for the wild-type allele. The results of subclone bias testing for all genes with more than five mutations are presented in **Supplementary Table 12**.

Copy number analysis, tumor purity and adjustment of mutated allele frequency. Initial copy number estimates were obtained by comparing read depth information for tumor and normal samples using VarScan2 (ref. 43). Depth estimates were then segmented using circular binary segmentation (CBS) as implemented in the DNAcopy package in R[44]. We used the JISTIC[45] program to generate a combined copy number matrix file. The VCRome2.1 probe locations were used as marker positions for copy number analysis. We then used JISTIC to calculate the significance for copy number gains and losses. Focal amplification at the *TERT* locus was determined using the average read depth of each captured target region.

Evaluation of tumor ploidy and purity. Using bam files from normal and tumor samples, read depth was calculated for each captured target region. After normalization by the number of total reads and GC content using regression analysis, the tumor/normal depth ratio was calculated, and values were smoothed using the moving average. Copy number peaks were then estimated using wavelet analysis, and each peak was approximated using Gaussian models. Hidden Markov models (HMMs) with the calculated Gaussian peaks were constructed, and copy number peaks were linked to genomic regions. The allelic imbalance for each copy number peak was calculated on the basis of heterozygous SNPs within the assigned region, and imbalance information and peak distances were further analyzed by model fitting where the optimal solution for a copy number peak was determined using vector matching, yielding estimated copy number and tumor purity and ploidy data simultaneously. Detailed algorithms will be described elsewhere (H.U., S.Y., K.T. and H.A., unpublished data).

HBV integration analysis. *HBV integration detection.* Viral genomes (HBV, NC_003977.1; HPV-16, NC_001526; HPV-18, NC_001357; HTLV-1, NC_001436) were downloaded from NCBI and included in the reference files when reads were mapped by BWA. No read was mapped to a virus other than HBV. To achieve more precise HBV mapping, we mapped all reads to the HBV reference sequence using the *q*-gram and Smith-Waterman method. An 11-mer *q*-gram was first applied to both strands of the HBV reference, and reads with 15 or more hits were subjected to Smith-Waterman alignment. The other end of each read was mapped to the hg19 human sequence using BWA. Finally, HBV integration sites were clustered by genomic position with a window size of 300 bp (approximately equal to the library fragment size), and sites with more than three supporting reads were used in the analysis.

Randomization test of HBV integration and copy number breakpoints. The 7,891 copy number breakpoints and 1,039 HBV integration sites were detected in 70 HBV-positive samples. Coexistence of the copy number breakpoints and HBV integration sites was examined using a 500-kb window size. To show statistical significance, we performed a randomization test by switching the position of the HBV integration sites to the same number of integration sites observed in the normal sample of other cases. We repeated this switching 100,000 times to yield distributions and estimated the *P* value.

Verification of single-nucleotide variation. We validated our mutation calls for frequently mutated genes (**Supplementary Table 31**) by resequencing samples using the Ion Proton sequencer (Life Technologies). Details are provided in the **Supplementary Note**.

Sanger sequencing of the *TERT* promoter. Bidirectional sequencing of the *TERT* promoter region was completed for 519 HCC samples. PCR runs were set up using 20 ng of genomic DNA, 10 µM manually designed primers (**Supplementary Table 32**) and KAPA HiFi DNA polymerase (Kapa Biosystems, KK2612). Touchdown PCR was performed with the following parameters: an initial denaturation at 98 °C for 5 min followed by 10 cycles of 98 °C for 30 s, 72 °C for 30 s and 72 °C for 1 min (decreasing the annealing temperature by 1 °C per cycle). The reaction then continued with 30 cycles of 98 °C for 30 s, 63 °C for 30 s and 72 °C for 1 min followed by a final extension at 72 °C for 5 min. The PCR products were purified with a 1:15 dilution of Exo-SAP, diluted by 0.6× and cycle sequenced for 25 cycles using a 1:64 dilution of BigDye Terminator v3.1 reaction mix (Applied Biosystems, 4337456). Finally, reactions were precipitated with ethanol, resuspended in 0.1 mM EDTA and analyzed on ABI 3730xl sequencing instruments using the Rapid36 run module and 3xx base-caller. SNPs were identified using SNP Detector software and were validated visually with Consed.

Analysis of mutation patterns and signatures. Mutation patterns for cases with hypermutation and IHCC cases were distinct from those for HCC cases (**Supplementary Figs. 4** and **21**), and cases with a small number of mutations cannot accurately represent the frequency of mutational patterns; therefore, cases with hypermutation, IHCC cases and cases with fewer than 40 mutations were excluded from further mutation pattern analysis.

The number of each of 96 possible somatic substitution types, C>A/G>T, C>G/G>C, C>T/G>A, T>A/A>T, T>C/A>G and T>G/A>C with the bases immediately 5′ and 3′ to each substitution in coding regions, was counted for each sample. The frequency of each of these substitutions was determined by dividing each count by the total number of substitutions, and the resulting frequencies were used for principal-component analysis. Principal-component analysis was implemented using the R command prcomp with the scaling option on. We used Wilks' λ test to evaluate the significance of the mean vector differences in different populations. We applied NMF to the 96-substitution pattern using published software[13], running 1,000 iterations of NMF with each NMF run iterated until convergence was achieved (10,000 iterations without change) or until the maximum number of 1,000,000 iterations was reached. We used another published software package[14] for model selection in NMF (selecting the input number of mutational signatures). Details on model selection for our NMF analysis are provided in the **Supplementary Note** and **Supplementary Figures 31–35**.

Pathway analysis. We used gene sets from MsigDB C2.all as pathway data sets. To assign *P* values representing the enrichment of mutations in pathways, we first checked whether a gene had at least one non-silent mutation or overlapped with focal CNAs for each sample in a given pathway (gene set). If so, we referred to such a gene as a 'mutated gene' for a sample. We then computed a population frequency for pathways with at least one mutated gene in the given pathway and divided the frequency by the total length of the unioned exons of all genes in the pathway to correct for the greater number of mutations in longer genes. This quotient was used as a test statistic. We used a bootstrapping approach to calculate *P* values. In the bootstrapping approach, we randomly selected as many genes as in the given pathway from all genes in the genome and then calculated the statistic. We repeated this sampling 2,000 times, calculating a fraction corresponding to the number of sampling results in which a statistic value was greater than or equal to the value in the observed data. This fraction was used as a *P* value.

To find intensively mutated gene modules in liver cancer tissue using the identified significantly mutated gene sets from MsigDB analysis, we used Pathway Commons[15] data for the whole unbiased human gene network and integrated the gene sets into this network. All pairs of gene relationships were weighted by how many mutated genes were shared by the two genes (shared ratio). These gene relationships constituted the gene network. The whole network was split into one large connected network and some isolated small networks. To extract gene modules, we recursively eliminated edges with low shared ratio values and distinguished into the smaller modules. Although the recursive edge elimination procedure gradually clarifies tightly connected gene modules, gene modules were rarely isolated from the whole network. Using this compression process and some additional manual curation, we finally selected ten representative modules that were intensively mutated in liver cancer tissues. We took essentially the same approach as described above to calculate *P* values for mutation enrichment and mutual exclusivity for a gene pair or combination of modules. For mutation enrichment, we used all genes in a pair of modules. For mutual exclusivity, if a module had at least one mutated gene, we referred to such a module as an 'impaired module' and computed a frequency of impaired modules for each sample.

Outcome analysis from non-negative matrix factorization signatures. NMF signature values were merged with annotated clinical data for our cohort. We performed calculations using standardized signature values to control for differences in mutational rate among the subjects. For the standardized data, the contributions of each signature within a subject summed to 1. We performed Cox proportional hazards analysis[46] using the R[44] survival package, factoring in all three signature components (signature A, signature B and signature C), age at diagnosis and histological tumor grade.

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
42. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
43. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
44. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2010).
45. Sanchez-Garcia, F., Akavia, U.D., Mozes, E. & Pe'er, D. JISTIC: identification of significant targets in cancer. *BMC Bioinformatics* **11**, 189 (2010).
46. Cox, D.R. & Oakes, D. *Analysis of Survival Data* (Chapman & Hall/CRC, Boca Raton, FL, 1984).

doi:10.1038/ng.3126

Molecular Mechanisms Underlying Oncogenic *RET* Fusion in Lung Adenocarcinoma

Tatsuji Mizukami, MD,†[*][†] Kouya Shiraishi, PhD,[*] Yoko Shimada, MFSc,[*] Hideaki Ogiwara, PhD,[*] Koji Tsuta, MD, PhD,[‡] Hitoshi Ichikawa, PhD,[§] Hiromi Sakamoto, PhD,[§] Mamoru Kato, PhD,[¶] Tatsuhiro Shibata, MD, PhD,[¶] Takashi Nakano, MD, PhD,[†] and Takashi Kohno, PhD[*]*

Background: Oncogenic *RET* fusion, caused by an inversion in chromosome 10, was recently identified as a driver mutation for the development of lung adenocarcinoma (LADC). Nevertheless, the molecular mechanism(s) underlying the rearrangement of the *RET* locus during lung carcinogenesis are unknown.

Methods: Genomic segments containing breakpoint junctions for *RET* fusions were cloned and analyzed by genomic polymerase chain reaction and genome capture sequencing using a next-generation sequencer to identify the mechanisms involved in DNA strand breaks and illegitimate joining of DNA ends. Of the 18 cases studied, 16 were identified by screening 671 LADC cases and two were previously published.

Results: Almost all (17 of 18, 94%) of the breakpoints in *RET* were located within a 2.0-kb region spanning exon 11 to intron 11 and no breakpoint occurred within 4 bp of any other. This suggested that as in papillary thyroid carcinoma, DNA strand breaks formed at nonspecific sites within this region trigger *RET* fusion. Just over half of the *RET* fusions in LADC (10 of 18, 56%) were caused by simple reciprocal inversion, and two DNA-repair mechanisms, namely nonhomologous end joining and break-induced replication, were deduced to have contributed to the illegitimate joining of the DNA ends.

Conclusions: Oncogenic *RET* fusion in LADC occurs through multiple pathways and involves the illegitimate repair of DNA strand breaks through mechanisms different from those identified in papillary thyroid carcinoma, where *RET* fusion also functions as a driver mutation.

Key Words: Lung adenocarcinoma, Molecular target therapy, Personalized medicine, *RET*, Gene fusion, DNA strand break.

(*J Thorac Oncol.* 2014;9: 622–630)

Oncogenic fusion of *RET* (rearranged during transfection) tyrosine kinase gene partnered with *KIF5B* (kinesin family member 5B) and *CCDC6* (coiled-coil domain containing 6) was identified as a novel druggable driver mutation in a small subset (1–2%) of patients with lung adenocarcinoma (LADC).[1–4] Vandetanib (ZD6474) and cabozantinib (XL184), two U.S. Food and Drug Administration–approved inhibitors of the *RET* tyrosine kinase showed therapeutic responses in a few patients with *RET* fusion-positive LADC.[5,6] Several clinical trials are currently underway to examine the therapeutic effects of RET tyrosine kinase inhibitors, including these two drugs.[7,8] *RET* fusions are generated by pericentric (includes the centromere, with a breakpoint in each arm) and paracentric (not including the centromere, with both breaks in the same arm) inversions of chromosome 10 (Fig. 1*A*). As the majority of *RET* fusion-positive patients are never-smokers,[3,9,10] cigarette smoking does not cause a predisposition. Therefore, the mechanism(s) responsible for the rearrangement of the *RET* locus are unknown. Elucidation of such a mechanism(s) may help to identify risk factors that can be modified or other preventive methods that can reduce the incidence of LADC; however, no such mechanism has been identified.[8]

Analyzing the breakpoints and structural aberrations in cancer cell genomes is a powerful method of identifying the underlying molecular mechanism(s) responsible, as the breakpoints retain "traces" of the DNA strand breaks and the illegitimate joining of DNA ends.[11–13] In fact, several studies have characterized the structure of the breakpoints responsible for the *ELE1* (also known as *RFG, NCOA4*, and *ARA70*)-*RET* oncogenic fusion in cases of papillary thyroid cancer (PTC), including post-Chernobyl irradiation-induced cases, to elucidate the mechanism underlying chromosome 10 inversion generating this fusion (Fig. 1*A*).[14–17]

Here, we examined the molecular processes underlying chromosome inversions that generate oncogenic *RET* fusions in LADC by cloning genomic segments containing breakpoint junctions and by comparing their structures with those identified in PTC. The results will increase our understanding of how *RET* fusions are generated and will also have implications for diagnosis of *RET* fusion-positive LADCs.

PATIENTS AND METHODS

Patient Samples

Fourteen frozen tissues (13 surgical specimens and a pleural effusion) and two methanol-fixed paraffin-embedded tissues from surgical specimens were obtained from the

*Division of Genome Biology, National Cancer Center Research Institute, Tokyo, Japan; †Department of Radiation Oncology, Gunma University Graduate School of Medicine, Gunma, Japan; ‡Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, Tokyo, Japan; §Division of Genetics, National Cancer Center Research Institute, Tokyo, Japan; and ¶Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Takashi Kohno, PhD, Division of Genome Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan. E-mail: tkkohno@ncc.go.jp

ISSN: 1556-0864/14/0905-0622

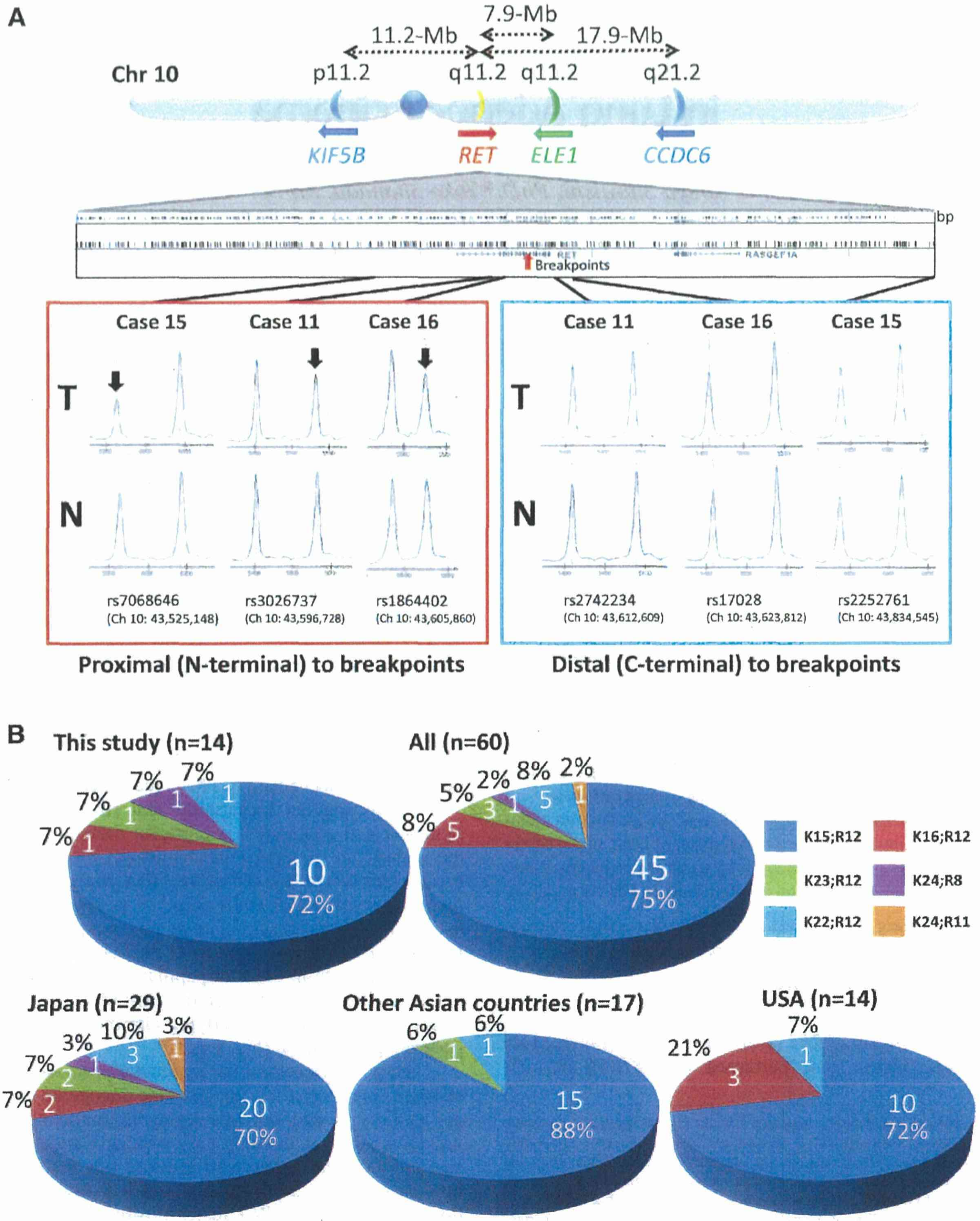

FIGURE 1. *RET* fusions. *A, Upper:* location of the *RET* oncogene and its fusion-partner genes *KIF5B, CCDC6,* and *ELE1* on chromosome 10. The *KIF5B-RET* fusion is generated in LADC, whereas the *CCDC6-RET* fusion is generated in LADC and PTC. The *ELE1-RET* fusion is frequent in radiation-induced PTC. *Lower:* LOH analysis. Allelic imbalance at SNP sites proximal and distal to the breakpoints were examined by MassArray analysis in three LADC cases with putative nonreciprocal inversions. Cases 11, 15, and 16 exhibited allelic imbalance (23%, 41%, and 29%, as indicated by arrows) at SNP loci proximal to the breakpoints, consistent with the fact that these samples have 20% to 40% tumor content. *B,* Fractions of *KIF5B-RET* fusion variants in LADCs. Fractions comprise the cohort from this study and eight published cohorts. Fractions in patients from Japan, other Asian countries (Korea and China), and the United States are shown below. LADC, lung adenocarcinoma; PTC, papillary thyroid carcinoma; LOH, loss of heterozygosity.

National Cancer Center (NCC) Biobank. These samples were from patients with LADC who received therapy at the NCC Hospital (Tokyo, Japan) between 1997 and 2012. All frozen samples were confirmed to be positive for *KIF5B-RET* fusion by reverse-transcriptase polymerase chain reaction (PCR) analysis, according to a previously described method.[3] *CCDC6-RET* fusion was detected by fusion fluorescence in situ hybridization (FISH) analysis of paraffin-embedded tissues using *RET*- and *CCDC6*-specific probes (Chromosome Science Labo Inc., Sapporo, Japan). This study was approved by the Institutional Review Board of the NCC.

Cloning and Sequencing of DNAs Containing Breakpoint Junctions

Genomic DNAs were extracted from cancer and noncancerous tissues using the QIAamp DNA Mini Kit or the QIAamp DNA Micro Kit (Qiagen, Hilden, Germany). Genomic DNA fragments containing breakpoint junctions were amplified by genomic PCR using primers that hybridized within the *KIF5B* and *RET* loci. PCR products specifically amplified in samples of interest were subjected to direct Sanger sequencing. The primers used are listed in Supplementary Table 1 (Supplementary Digital Content 1, http://links.lww.com/JTO/A541).

Genome-Capture Deep Sequencing Using a Next-Generation Speed Sequencer

Nucleotide sequences of *CCDC6-RET* fusion breakpoints were examined by targeted genome capture and massively parallel sequencing using an Ion Torrent Personal Genome Machine (Ion Torrent PGM) sequencing system and the Ion TargetSeq Custom Enrichment Kit (Life Technologies, Carlsbad, CA). One microgram of genomic DNA was subjected to enrichment using the probes listed in Supplementary Table 2 (Supplementary Digital Content 1, http://links.lww.com/JTO/A541). The mean depth of sequencing was approximately 1000.

Analysis of Sequence Reads Obtained by a Second-Generation Sequencer

Sequence reads were analyzed using a program developed by the authors. Briefly, reads were mapped to sequences of the *RET* and *CCDC6* genes using the Burrows-Wheeler Aligner, Smith-Waterman alignment (BWA-SW) software[18] to detect reads that mapped to both the *RET* and *CCDC6* genes. Breakpoints were extracted from the local alignment results of BWA-SW. The detailed procedure is described in Supplementary Notes (Supplementary Digital Content 2, http://links.lww.com/JTO/A542). Structures of breakpoint junctions were verified by Sanger sequencing of genomic PCR products.

Loss of Heterozygosity Analysis

Genomic DNAs obtained from cancerous and noncancerous tissues were subjected to single nucleotide polymorphism (SNP) genotyping using the Illumina HumanOmni1 2.5M Chip (Illumina, San Diego, CA). Based on the B-allele frequencies obtained using the Illumina GenomeStudio software, loss of heterozygosity (LOH) regions in *RET* and surrounding regions were deduced. Representative SNP loci were subjected to analysis of allelic imbalance using the Sequenom MassARRAY system (Sequenom, San Diego, CA).

Analysis of Nucleotide Sequences

Nucleotide sequence analysis, including search for sequence homology, was performed using the Genetyx-SV/RC Ver 8.0.1. software (Genetyx, Tokyo, Japan). Information about the distribution of repetitive elements, GC contents, conservation, DNA methylation, DNase sensitivity, and histone modification within the *RET* gene was obtained using the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

RESULTS

KIF5B-RET Fusion Variations in LADC

In our previous study, six of 319 LADC cases (1.9%) carried *KIF5B-RET* fusions.[3] In this study, we examined *KIF5B-RET* fusion by reverse-transcriptase PCR in a further 352 LADC cases and found eight additional *KIF5B-RET* fusion-positive cases. In total, 14 of 671 cases (2.1%) were positive for *KIF5B-RET* fusions (cases 1–4 and 7–16 in Table 1 and Supplementary Table 3, Supplementary Digital Content 1, http://links.lww.com/JTO/A541), and this frequency was consistent with those reported for other cohorts.[9,10,19]

Among those 14 cases, 10 (71%) contained a fusion of *KIF5B* exon 15 to *RET* exon 12 (K15;R12), whereas the remaining four each contained other variants. Thus, K15;R12 is the most frequent variant (Fig. 1*B*). The prevalence of the K15;R12 variant (45 of 60, 75%) was verified in a total of 60 cases, including 46 cases from eight other cohorts published to date[1–4,9,10,19,20] (Fig. 1*B*, Supplementary Table 4, Supplementary Digital Content 1, http://links.lww.com/JTO/A541). This preference was similar among cohorts from Japan, other Asian countries, and the United States ($p > 0.05$ by Fisher's exact test).

Distribution of Breakpoints in the *RET* and *KIF5B* Genes

To explore the molecular processes underlying *RET* fusion in LADC, we examined the location (clustering) of the breakpoints and the structure of the breakpoint junctions; information about the former enabled us to deduce the genomic or chromosomal features that make DNA susceptible to strand breaks, whereas information about the latter enabled us to deduce the mechanism underlying the illegitimate joining of DNA ends by DNA repair pathways.

The locations of the 28 breakpoints in the 14 *KIF5B-RET* fusion-positive cases mentioned above were identified by Sanger sequencing analysis of genomic PCR products and mapped (yellow arrowheads in Fig. 2*A* and *B*). The breakpoints in a single Korean case from another study were also identified and mapped (orange arrowheads in Fig. 2*A*; case 17 in Table 1). Consistent with the predominance of K15;R12 variants, most of the breakpoints were mapped to intron 11 of *RET* and intron 15 of *KIF5B* (Fig. 2, detailed information in Supplementary Table 5, Supplementary Digital Content 1, http://links.lww.com/JTO/A541).

TABLE 1. Structure of Breakpoint Junctions of *RET* Fusions in Lung Adenocarcinoma

				Deletion in the Joining		DNA Segment Duplication by Inversion		Nucleotide Overlap at Junction		Nucleotide Insertion at Junction				
No.	Sample Name	Fusion Partner	Reciprocal/ Nonreciprocal	*RET*	Partner	*RET*	Partner	Partner -*RET*	*RET*-Partner	Partner -*RET*	*RET*-Partner	Mode of DNA End Joining	LOH Proximal to *RET*	Smoking
1	BR0020	*KIF5B*	Reciprocal	—	—	—	—	—	—	—	—	NHEJ	NT	No
2	L07K201T	*KIF5B*	Reciprocal	14 bp	19 bp	—	—	C	—	—	ATA	NHEJ	NT	Yes
3	349T	*KIF5B*	Reciprocal	1 bp	7 bp	—	—	—	—	A	A	NHEJ	NT	Yes
4	AD08-341T	*KIF5B*	Reciprocal	16 bp	26 bp	—	—	—	—	—	—	NHEJ	NT	No
5	RET-030	*CCDC6*	Reciprocal	52 bp	1021 bp		—	— —		—	—	NHEJ	NT	No
6	RET-024	*CCDC6*	Reciprocal	14 bp	2 bp	—	—	—	—	—	—	NHEJ	NT	Yes
7	AD12-106T	*KIF5B*	Reciprocal	—	573 bp	490 bp	—	—	—	—	—	BIR	NT	Yes
8	BR0030	*KIF5B*	Reciprocal	—	—	—	211 bp	—	—	—	—	BIR	NT	No
9	442T	*KIF5B*	Reciprocal	269 bp	—	—	232 bp	—	—	—	—	BIR	NT	No
10	AD08-144T	*KIF5B*	Reciprocal	5 bp	—	—	33 bp	—	—	—	—	BIR	NT	No
11	BR1001	*KIF5B*	Nonreciprocal					—		AGT		NHEJ	+	No
12	AD09-369T	*KIF5B*	Nonreciprocal					CTC		—		NHEJ (alternative end joining)	NT	No
13	BR1002	*KIF5B*	Nonreciprocal					A		—		NHEJ	NT	No
14	AD12-001T	*KIF5B*	Nonreciprocal					—		—		NHEJ	NT	Yes
15	BR1003	*KIF5B*	Nonreciprocal					—		CTTT		NHEJ	+	No
16	BR1004	*KIF5B*	Nonreciprocal					—		RET exon 7 to intron 7 (359 bp)		Complex rearrange	+	No
17	AK55[a]	*KIF5B*	Nonreciprocal					—		GT		NHEJ	NT	No
18	LC-2/ad[b]	*CCDC6*	Nonreciprocal					—		—		NHEJ	NT	Unknown

[a] Ju et al.[4]
[b] Suzuki et al.[21]
LOH, loss of heterozygosity; NHEJ, nonhomologous end joining; NT, not tested; BIR, break-induced replication; blank, not applicable.

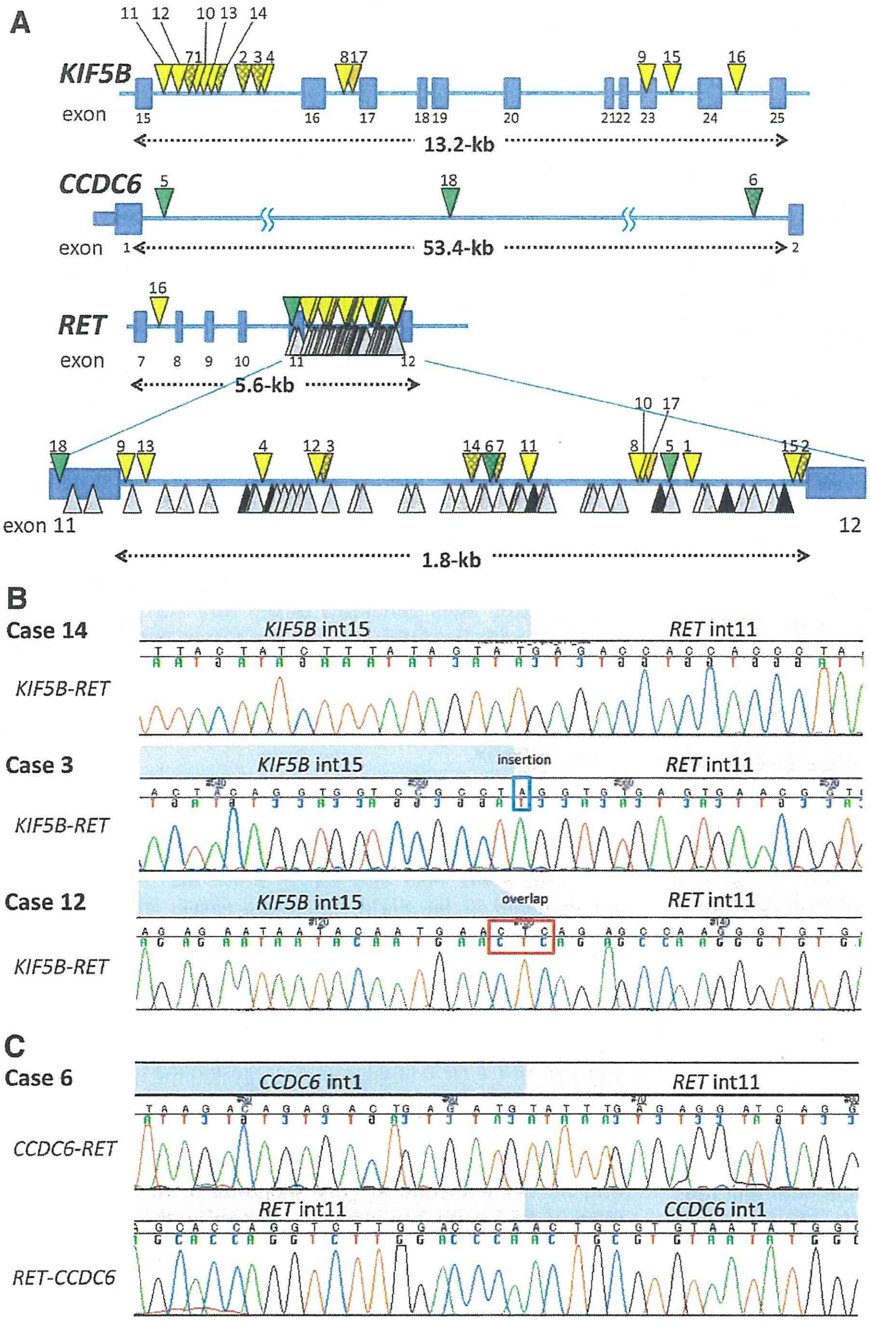

FIGURE 2. Breakpoint analysis. *A*, Distribution of breakpoints in the *CCDC6*, *KIF5B*, and *RET* genes. Yellow arrowheads indicate the locations of breakpoints for *KIF5B-RET* fusions in Japanese cases (cases 1–4 and 7–16 in Table 1), whereas the orange arrowhead indicates the breakpoints in a single Korean case (case 17). Green arrowheads indicate the locations of breakpoints of *CCDC6-RET* fusions in three Japanese cases (cases 5, 6, and 18). Arrowheads for ever-smoker LADC cases are hatched. Gray and black arrowheads indicate breakpoints of *RET-ELE1* fusion in 38 radiation-induced post-Chernobyl PTCs and six sporadic PTCs, respectively.[14–17] *B*, Electropherograms for Sanger sequencing of genomic fragments encompassing *KIF5B-RET* breakpoint junctions. PCR products were directly sequenced. Examples of three fusion patterns (joined without any nucleotide insertions or overlaps, joined with a nucleotide insertion, and joined with three nucleotide overlap) are shown. Inserted and overlapping nucleotides at breakpoint junctions are indicated, respectively, by the blue and red boxes. *C*, Electropherogram for Sanger sequencing of genomic fragments encompassing *CCDC6-RET* and *RET-CCDC6* breakpoint junctions. LDAC, lung adenocarcinoma; PCR, polymerase chain reaction; PTC, papillary thyroid carcinoma.

None of the *RET* and *KIF5B* breakpoints mapped at the same position, and no breakpoint was within 6 bp of another. To further investigate the breakpoint clustering, we mapped breakpoints in three cases of *CCDC6-RET* fusion, a minor fusion variant (cases 5, 6, and 18 in Table 1 and Supplementary Table 3, Supplementary Digital Content 1, http://links.lww.com/JTO/A541). Two of these cases were primary tumors, diagnosed by break apart and fusion *FISH*, and their breakpoints were determined by genome-capture deep sequencing of genomic DNAs using a second-generation sequencer. The remaining case was a LADC cell line from a Japanese patient, for which the breakpoints had previously been determined by the same method.[21] Two breakpoints and one breakpoint in the *RET* gene were mapped to intron 11 and exon 11, respectively (green arrowheads in Fig. 2), and no breakpoint was located within 5 bp of another. In total, a 2.0-kb region spanning exon 11 to intron 11 of *RET* and a 5.6-kb region spanning intron 15 of *KIF5B* (10 of 15, 75%) contained the majority of breakpoints (17 of 18 [94%] and 10 of 15 [75%], respectively), and these breakpoints were at least 5 bp from each other. Breakpoints within exon 11 to intron 11 of *RET* and intron 15 of *KIF5B* were not distributed in an evidently biased manner, nor did they exhibit any particular nucleotide sequence or composition (Supplementary Table 5, Supplementary Digital Content 1, http://links.lww.com/JTO/A541). Therefore, DNA strand breaks triggering oncogenic *RET* fusions in LADC occur preferentially in a few defined regions, but at nonspecific sites within those regions.

Reciprocal and Nonreciprocal Inversions Causing *RET* Fusions

To explore the modes of DNA end joining that give rise to *RET* fusion, we investigated the structures of *RET* fusion breakpoint junctions. To address whether chromosome inversion events were reciprocal, we cloned genomic segments containing reciprocal breakpoint junctions (i.e., *RET-KIF5B* and *RET-CCDC6*) from 17 Japanese cases (Table 1). Ten of the 17 cases, consisting of eight *KIF5B-RET* and two *CCDC6-RET* cases, allowed amplification of reciprocal genomic segments using PCR primers set 1 kb away from the identified *KIF5B-RET* or *CCDC6-RET* breakpoints. This indicated that these fusions were the results of simple reciprocal inversions (cases 1–10 in Table 1, Fig. 2*C*). On the other hand, the remaining seven cases did not allow amplification of genomic segments encompassing the reciprocal breakpoint junctions (cases 11–16 and 18 in Table 1). Three of these seven cases, for which corresponding noncancerous DNA was available, were subjected to LOH analysis at the *RET* locus. LOH was detected at a region proximal (N-terminal) to the breakpoints in all three cases (cases 11, 15, and 16 in Table 1, Fig. 1*A*), indicating nonreciprocal inversion associated with deletion of a copy of the region proximal to the breakpoints. In addition, the inversion in the aforementioned Korean case (case 17) is also nonreciprocal.[4] These data suggested that only a fraction of *RET* fusions (10 of 18, 56%) are caused by simple reciprocal inversions.

Modes of DNA End Joining That Give Rise to Reciprocal Inversions

Two major types of DNA repair pathways cause structural variations.[11,12] The first type is nonhomologous end joining (NHEJ) of DNA double strand breaks (DSBs). which requires very short (a few base pairs) or no homology, and often inserts a few nucleotides at breakpoint junctions.[8,22,23] NHEJ has canonical and noncanonical forms; in the latter, called alternative end joining, DNA ends are joined using microhomology of a few nucleotides at breakpoints.[24] The second type includes repair pathways that use long (>10 bp) homology at DNA ends, such as break-induced replication (BIR) and nonallelic homologous recombination.[12,25]

Sequence analysis of breakpoint-containing genomic segments in 10 reciprocal cases revealed that deletions frequently (8 of 10, 80%) occur in *RET* and/or its partner locus (i.e., *KIF5B* or *CCDC6*) upon DNA end joining (Table 1). This analysis also enabled us to deduce that both types of repair pathways described above are involved in these joining events. In six of the cases (cases 1–6 in Table 1), four DNA ends were joined, and in two cases, insertions were observed (representative cases in Supplementary Fig. 1, Supplementary Digital Content 3, http://links.lww.com/JTO/A543). The lack of significant homology between the sequences of the *RET* and *KIF5B/CCDC6* breakpoints led us to deduce that DNA end joining was mediated by NHEJ in these six cases: two DSBs formed, one each in *RET* and its partner locus, and the four resultant DNA ends were illegitimately joined by canonical or noncanonical NHEJ (Fig. 3*A*).

The remaining four cases (cases 7–10 in Table 1) had a distinctive feature. DNA segments of 33 to 490 bp from either the *RET* or *KIF5B* locus were retained at both the *KIF5B-RET* and *RET-KIF5B* breakpoints, resulting in duplication of these segments. Notably, two regions encompassing the breakpoint in a locus exhibited sequence homology to the duplicated segment of the other locus (representative cases in Supplementary Fig. 2, Supplementary Digital Content 3, http://links.lww.com/JTO/A543). This feature led us to deduce that these joining events were mediated by BIR, using both DNA ends generated by DNA single-strand breaks at the *RET* or fusion-partner locus (Fig. 3*B*). Specifically, two DNA broken ends generated at the *RET* (or partner locus) annealed with the DSB sites of the fusion-partner (or *RET*) locus through sequence homology and were then subjected to ectopic DNA replication. This process left the same DNA segment at both breakpoint junctions, resulting in duplication of the segment.

Speculated Mode of DNA End Joining Giving Rise to Nonreciprocal Inversion

Our study also speculated about the modes of joining involved in the eight remaining cases, which were not likely to have been subjected to simple reciprocal inversion and are therefore defined here as nonreciprocal (cases 11–18 in Table 1). Due to the lack of sequence information from breakpoints in reciprocal counterparts, deletions could not be assessed. The lack of significant homology between the *RET* and *KIF5B/CCDC6* breakpoints suggested the involvement of NHEJ. Consistent with this idea, insertion of a few nucleotides, a common trace of NHEJ, was observed in three cases (cases 11, 15, and 17). A single case (case 16) had an insertion of 349 nucleotides, corresponding to the inverted segment of *RET* exon 7 to intron 7, suggesting the occurrence of an unspecified complex rearrangement mediated by a process other than NHEJ, such as fork stalling and template switching (Lee et al., 2007). These results suggest that the predominant molecular process is illegitimate NHEJ repair, in which two DSBs are formed both in the *RET* and partner loci, and one end of the partner locus (the N-terminal part of *KIF5B* or *CCDC6*) and one end of the *RET* locus (the C-terminal part) are joined by NHEJ. Nevertheless, the remaining two DNA ends were not joined in a simple manner. DNA segments within the DNA ends were either lost or joined with DNA ends other than those at the *RET*, *KIF5B*, and *CCDC6* loci, consistent with the observations of LOH at regions proximal to breakpoints in *RET* (Table 1). In fact, in case 17, the 3′ part of the *KIF5B* gene was fused to the *KIAA1462* gene, 2.0 Mb away from *KIF5B*.[4]

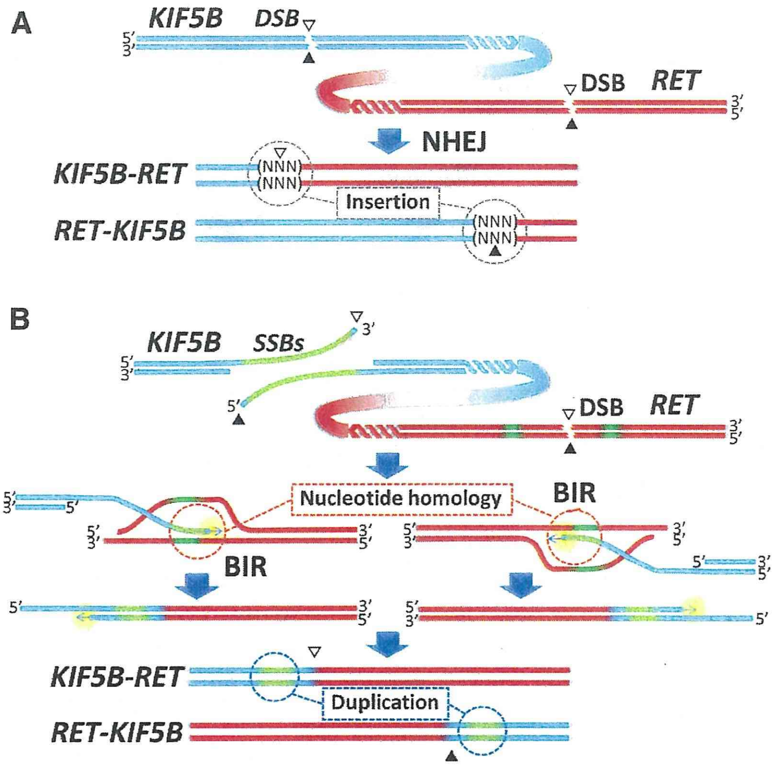

FIGURE 3. Deduced processes of reciprocal inversion by NHEJ and BIR. *A*, NHEJ. Four DNA ends generated by DSBs at *RET* and a partner locus were directly joined. Often, insertions of nucleotides (NNN) at breakpoint junctions are observed. *B*, BIR. Here, DNA single-strand breaks (SSBs) occur in the *KIF5B* locus and a DSB occurs in the *RET* locus. The two SSBs at the *KIF5B* locus trigger BIR by annealing at two homologous sites in the *RET* locus. BIR results in duplication of a *KIF5B* segment. As a result, the *RET* breakpoints in the *KIF5B-RET* and *RET-KIF5B* fusions are located at the same position (a DSB site), whereas the *KIF5B* breakpoints in these fusions are located at different positions (two SSB sites). ▽, breakpoints for partner-*RET* fusion; ▲, breakpoints for *RET*-partner fusion. NHEJ, nonhomologous end joining; BIR, break-induced replication.

DISCUSSION

In this study, we investigated the molecular mechanisms underlying oncogenic *RET* fusion in LADCs. Distribution of breakpoints made us consider a 2.0-kb segment spanning *RET* exon 11 to intron 11 (and also a 5.6-kb segment spanning *KIF5B* intron 15) as a breakpoint cluster region(s). The breakpoints in these regions were dispersed at intervals larger than 4 bp. The inferred breakpoints do not necessarily indicate the sites of actual DNA breaks because resection of nucleotides from DNA ends sometimes occurs during the DNA repair.[23] In fact, we observed nucleotide deletions in eight of 10 LADC cases with reciprocal *KIF5B*/*CCDC6-RET* inversions. Nevertheless, when the locations of putative breakpoints before DNA end resection were included, the breakpoint distribution remained scattered. These data strongly suggested that the majority of DNA breaks triggering *RET* fusions occur at nonspecific sites in defined regions of a few kb in size. Furthermore, this seems to hold true irrespective of etiology and tumor type: the distribution of breakpoints was not significantly different between ever- and never-smokers, and *RET* exon 11 to intron 11 was also defined as a breakpoint cluster region for *RET* fusions in PTCs, as previously reported.[14–17] The cases shown in Figure 2 (gray and black arrowheads) include PTCs induced by post-Chernobyl irradiation, in which DNA breaks were presumably caused exclusively by irradiation; the random breakpoint distributions in these PTCs were similar to those of the LADCs we analyzed.

We investigated the DNA end-joining pathways that gave rise to *RET* fusions by analyzing the structures of breakpoint junctions. NHEJ was found to be one of the major pathways of DNA end joining. We and others also showed that NHEJ is also prominently involved in interstitial deletions that inactivate tumor-suppressor genes, such as *CDKN2A/p16* and *STK11/LKB1*, in lung cancer.[13,26,27] Thus, NHEJ contributes to the occurrence of driver mutations in both tumor-suppressor genes and oncogenes during lung carcinogenesis. Our data also reveal a possible contribution of BIR in DNA end joining to generate reciprocal inversions. We deduced that BIR occurred from DNA ends, probably generated by DNA single-strand breaks, in the *RET* or partner locus, beginning with annealing with the other locus through nucleotide homologies of tens to hundreds of base pairs. This process resulted in duplication of breakpoint-flanking DNA segments of tens to hundreds of base pairs. BIR has recently been implicated in oncogenic *RAF* fusions in pediatric brain tumors.[28] In those cases, the sequence homology used for annealing of DNA ends was on the order of a few base pairs. Thus, BIR might generate oncogenic fusions frequently, although the detailed process may differ according to tumor type.

Irrespective of the similarities in breakpoint distribution, several processes involved in *RET* fusions differed between LADC and PTC (Fig. 4). Reciprocal inversion was unlikely to have occurred by BIR in PTC because none of the PTC cases exhibited the duplication of DNA segments that were observed in LADC; therefore, the joining of DNA ends in PTC was likely to have been mediated exclusively by NHEJ.[17] This is plausible because *RET* fusions preferentially occur in PTCs in patients suffering from high-dose radiation exposure, suggesting that DSBs generated at the *RET* or partner loci triggered the chromosome rearrangements that generated *RET* fusions.[29] Repetitive NHEJ repair of abundant