Cancer Science
The official journal of the Japanese Cancer Association

High-throughput resequencing of target-captured cDNA in cancer cells

Toshihide Ueno,[1,5] Yoshihiro Yamashita,[1,5] Manabu Soda,[1] Kazutaka Fukumura,[2] Mizuo Ando,[2] Azusa Yamato,[1] Masahito Kawazu,[2] Young Lim Choi[1,2] and Hiroyuki Mano[1,2,3,4]

[1]Division of Functional Genomics, Jichi Medical University, Tochigi; [2]Department of Medical Genomics, Graduate School of Medicine, University of Tokyo, Tokyo; [3]CREST Japan Science and Technology Agency, Saitama, Japan

(Received May 30, 2011/Revised September 7, 2011/Accepted September 14, 2011/Accepted manuscript online September 20, 2011/Article first published online October 13, 2011)

The recent advent of whole exon (exome)-capture technology, coupled with second-generation sequencers, has made it possible to readily detect genomic alterations that affect encoded proteins in cancer cells. Such target resequencing of the cancer genome, however, fails to detect most clinically-relevant gene fusions, given that such oncogenic fusion genes are often generated through intron-to-intron ligation. To develop a resequencing platform that simultaneously captures point mutations, insertions–deletions (indels), and gene fusions in the cancer genome, we chose cDNA as the input for target capture and extensive resequencing, and we describe the versatility of such a cDNA-capture system. As a test case, we constructed a custom target-capture system for 913 cancer-related genes, and we purified cDNA fragments for the target gene set from five cell lines of CML. Our target gene set included Abelson murine leukemia viral oncogene homolog 1 (*ABL1*), but it did not include breakpoint cluster region (*BCR*); however, the sequence output faithfully detected reads spanning the fusion points of these two genes in all cell lines, confirming the ability of cDNA capture to detect gene fusions. Furthermore, computational analysis of the sequence dataset successfully identified non-synonymous mutations and indels, including those of tumor protein p53 (*TP53*). Our data might thus support the feasibility of a cDNA-capture system coupled with massively parallel sequencing as a simple platform for the detection of a variety of anomalies in protein-coding genes among hundreds of cancer specimens. (*Cancer Sci* 2012; 103: 131–135)

Cancer is thought to result from various alterations of the genome, including point mutations, insertions–deletions (indels), and genomic rearrangements.[1] Whereas comprehensive sequencing of the cancer genome, or "cancer genome resequencing", is a promising approach to the identification of such anomalies, and to provide a basis for the development of effective treatment strategies for cancer, determination of the nucleotide sequence of the entire human genome with conventional Sanger sequencers remains a highly demanding task. However, the recent advent of massively parallel sequencing systems, or second-generation sequencers, has rendered such projects manageable in private laboratories[2] and triggered the formation of large-scale consortia, such as The Cancer Genome Atlas and International Cancer Genome Consortium,[3] to undertake cancer genome resequencing for hundreds of specimens. Cancer genome resequencing with massively parallel sequencers has already provided a wealth of information on genome-wide mutation status for melanoma,[4] acute myeloid leukemia,[5] hepatocellular carcinoma,[6] and other cancers.

Even with the current massively parallel sequencers, however, the determination and compilation of the full genome sequence for a given sample might still take almost 1 month. Comparison of the cancer genome among many specimens thus remains time-consuming and labor intensive. Anomalies in protein-coding genes likely play a major role in carcinogenesis. Given that exonic regions occupy only ~1.3% of the human genome, sequencing such targeted regions would be expected to markedly facilitate the discovery of proteins that are activated or inactivated specifically in cancer cells. Indeed, target-capture strategies, coupled with massively parallel sequencers, have revealed important genetic changes in cancer,[7] as well as in hereditary disorders.[8,9]

One important drawback of such target-capture approaches, however, is their inability to detect gene fusions. Most cancer-associated gene fusion events occur within introns (resulting in exon-to-exon ligation in the corresponding mRNA), and exon capture does not reveal breakage and ligation of intronic regions. Recurrent gene fusions were once thought to be rare in epithelial tumors compared with hematologic malignancies and sarcomas;[10] however, our recent discovery of the echinoderm microtubule associated protein like-4 (EML4)-anaplastic lymphoma kinase (ALK) fusion gene in lung cancer and the discovery by others of rearrangements in loci for the v-ets avian erythroblastosis virus E26 oncogene homolog (ETS) family of transcription factors in prostate cancer have led to a revision of this notion.[11,12] It would thus be desirable to develop a resequencing platform that is able to capture, within a reasonable timeframe, all gene fusions, point mutations, and indels in the cancer genome. In pursuit of this goal, we have now examined the efficacy of high-throughput sequencing of captured cDNA for the identification of such cancer genome anomalies.

Materials and Methods

Cell lines. Cell lines established from the blast crisis stage of CML, including MEG-01s, KCL-22-SR, K562, NCO2, and KU812,[13,14] were obtained from the Japanese Collection of Research Bioresources (Osaka, Japan) and were maintained in RPMI-1640 medium (Invitrogen, Carlsbad, CA, USA) supplemented with 10% FBS (Invitrogen). Total RNA was isolated from each cell line with the use of an RNeasy mini kit (Qiagen, Valencia, CA, USA) and was subjected to cDNA synthesis with an oligo(dT) primer.

Gene expression profiling. The cDNA prepared from total poly(A)-RNA of KCL-22-SR cells was subjected to hybridization with the HGU95Av2 microarray (Affymetrix, Santa Clara, CA, USA), as described previously.[15] The expression intensity of each test gene on the array was normalized by the 50th percentile value.

cDNA-capture methods. RNA probes of 120 bases were designed to cover (with a 60-base overlap) cDNA of 913 human protein-coding genes (Table S1), and were synthesized by Agilent Technologies (Santa Clara, CA, USA). During the design of the probes, the Repeat Masker dataset (http://www.repeatmasker.org) was used to remove probes corresponding to

[4]To whom correspondence should be addressed. E-mail: hmano@jichi.ac.jp
[5]These authors contributed equally to this work.

doi: 10.1111/j.1349-7006.2011.02105.x

repetitive sequences in the human genome. Hybridization of DNA fragments to the RNA probes was performed according to the protocols recommended for the SureSelect Target Enrichment system (Agilent). We also used the SureSelect Human X Chromosome Demo kit (Agilent) to examine purification efficiency. Purified DNA fragments were then subjected to sequencing with a Genome Analyzer IIx (GAIIx; Illumina, San Diego, CA, USA) for 76 bases from both ends by the paired-end sequencing system.

Computational pipeline. Raw read data were quality filtered on the basis of the presence of the Illumina adaptor sequences and a *Q*-value of ≥20. The resulting read sequences were then subjected to an in-house computational pipeline to identify various mutations (Fig. S1). In brief, read sequences were matched with the Bowtie algorithm[16] to the cDNA sequences of the 913 genes used to construct our custom-made SureSelect system. The matched reads were then examined for the presence of non-synonymous mutations and single nucleotide polymorphisms (SNP) deposited in dbSNP (build 132, http://www.ncbi.nlm.nih.gov/projects/SNP/index.html). The remaining reads were further matched to the cDNA sequences with Burrows-Wheeler Aligner (BWA) and Basic Local Alignment Search Tool (BLAST) algorithms to search for indels and multiple mutations.[17,18] Candidates for non-synonymous mutations were identified only when ≥20% of reads correspond to the mutations at positions with ≥50 coverage.

For the selection of reads corresponding to possible fusion cDNA, nucleotide sequences of 20 bp were obtained from both ends of each read and were separately matched to RefSeq mRNA (http://www.ncbi.nlm.nih.gov), KnownGeneMrna,[19] and the human genome sequence (GRCh37, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/?build=37). Reads were considered to be derived from fusion genes if the ends of a given read matched to different genes within the 913-gene group, or one end matched to a single gene within the 913-gene group and the other end matched to a sequence in RefSeq, KnownGeneMrna, or the human genome sequence that did not correspond to the 913 genes. Candidates for fusion genes were identified only when four or more reads were mapped to possible fusion points.

RT–PCR. To confirm the presence of an alternatively-spliced mixed-lineage leukemia (*MLL*) mRNA, we subjected oligo(dT)-primed cDNA of KU812 cells to PCR with the combination of the F-1 primer (5′-ACCTCGTGGGAGACCTAGAAGTGG-3′) and the R primer (5′-AGTCATTGGAAGCTTGTCTGCCTG-3′), or with the combination of the F-2 primer (5′-CCTGTGGGTA-GGGTTTCCAAAGAG-3′) and the R primer.

Results

Efficiency of cDNA-capture sequencing. Paired-end sequencing of target-captured cDNA was briefly described in a previous study;[20] however, how the efficiency of target purification with cDNA compares with that with genomic DNA remains unclear. We therefore attempted to optimize the conditions for cDNA purification with the SureSelect system. Oligo(dT)-primed cDNA of KCL-22-SR cells were fragmented to a mean size of 500 or 200 bp and then subjected to purification with the use of the SureSelect Human X Chromosome Demo kit, which is designed to capture genomic sequences derived from the human X chromosome. Genomic DNA of KCL-22-SR cells was similarly processed and hybridized with the X Chromosome Demo kit. The purified fragments at either 4 or 8 pM were then sequenced by the GAIIx system.

The X chromosome-mapped cDNA reads occupied 62.1%, 81.6%, 62.4%, and 82.2% of quality filter-passed reads for the experiments with 4 pM of 500-bp fragments, 4 pM of 200-bp fragments, 8 pM of 500-bp fragments, and 8 pM of 200-bp fragments, respectively (Fig. 1). Thus, these results suggested that the shorter cDNA fragments were captured more efficiently than the longer ones. Furthermore, the purification efficiency for genomic DNA fragments was not higher than that for cDNA, irrespective of DNA concentration and fragmentation size (Fig. 1), supporting the feasibility of cDNA-capture approaches.

The ability to detect breakpoint cluster region (*BCR*)-Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) fusion reads was reduced for the cDNA sheared to ~200 bp compared with that for those of ~500 bp (see below). The former cDNA detected 83.7% or 76% of the fusion reads detected by the latter cDNA at input concentrations of 4 and 8 pM, respectively. This result is in line with our computational bootstrap trial (n = 10 000) showing that the number of randomly-fragmented, 200-bp reads encompassing the *BCR–ABL1* fusion point is ~2.5 times higher than that of 500-bp reads (data not shown). However, given that the total number of high-quality reads was much higher in the data for the 200-bp cDNA than in those for the 500-bp cDNA (Fig. 1), we chose to use 8 pM of cDNA with a mean size of 200 bp for further experiments.

Custom cDNA-capture system. We also tested whether extensive sequencing of cDNA generated from total poly(A)-RNA (unselected cDNA) might serve to identify gene fusions, point mutations, and indels. For this purpose, unselected cDNA were prepared from KCL22-SR cells, and subjected to GAIIx sequencing, yielding 34.1 million reads, which mapped to 36 128 RefSeq entries (data not shown). The distribution of read number per transcript in the data is shown in Figure 2a. Among the 36 128 entries, only 200 (0.55%) accounted for ~20% of total reads, and 4.55% accounted for ~50% of reads. Thus, as expected, resequencing data for unselected cDNA consist mostly of reads corresponding to a limited number of highly-abundant transcripts.

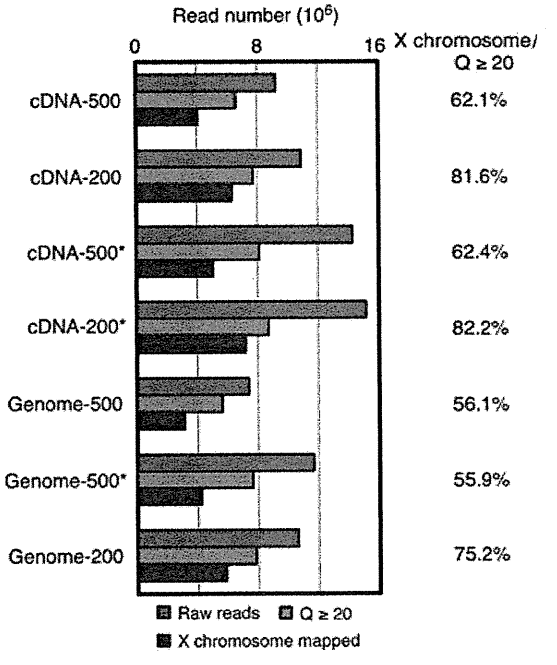

Fig. 1. Comparison of capture efficiency between cDNA and genomic DNA. Genomic DNA or cDNA of KCL-22-SR cells was fragmented to a mean size of 200 or 500 bp, and then subjected to purification with the SureSelect Human X Chromosome Demo kit, followed by GAIIx sequencing at a concentration of 4 or 8 pM (the latter indicated by an asterisk). Numbers of raw reads, reads with a *Q*-value of ≥20 ($Q \geq 20$), and reads mapped to the human X chromosome are shown for each experiment. Percentage of X chromosome-mapped reads among the reads with a *Q*-value of ≥20 is shown on the right.

doi: 10.1111/j.1349-7006.2011.02105.x

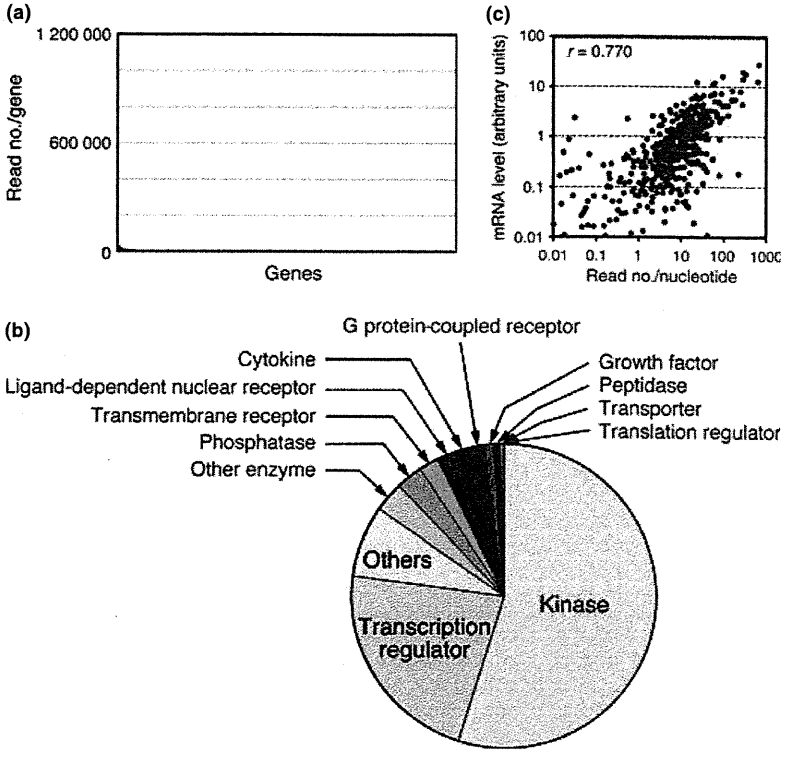

Fig. 2. Capture of a selected set of cDNA. (a) Read number for each gene was calculated from the sequencing data for the unselected cDNA of KCL-22-SR cells. Genes were sorted according to their read number. A small number of genes accounted for most of the sequence reads. (b) Functional annotation for the encoded proteins of our target cDNA (n = 913). (c) Read number per nucleotide for each captured cDNA in KCL-22-SR cells is compared with the expression intensity (arbitrary units) of the same cDNA examined with an HGU95Av2 microarray. Pearson's correlation coefficient (r) for the comparison is also demonstrated.

We therefore attempted to construct a custom SureSelect system to capture cDNA for cancer-related genes. For this purpose, we selected 913 genes that yielded 56 892 hybridization probes corresponding to ~3.77 Mbp of total capture capacity. The target genes encoded human protein kinases (all members in the human genome), transcription regulators, phosphatases, and other proteins (Fig. 2b; Table S1).

To compare the information provided by the sequence data from unselected and captured cDNA, we purified target cDNA from KCL-22-SR cells with the use of our custom SureSelect system, and determined their nucleotide sequences with GAIIx. A comparable amount of filter-passed reads (39.2 million) to that of unselected cDNA were thus obtained. We found that 88% of the captured cDNA were mapped to the target genes in our SureSelect system, while only 6.6% of the unselected cDNA were mapped to the 913 targets (data not shown). The read number obtained for each gene in the captured cDNA dataset is shown in Figure S2, with the distribution being markedly different from that obtained by sequencing of the unselected cDNA (Fig. 2a). As expected, the read number per nucleotide in each cDNA for the captured dataset was highly correlated to the expression intensity of the same gene quantified with the HGU95Av2 GeneChip expression array (Pearson's correlation coefficient = 0.770, $P < 2.2 \times 10^{-16}$) (Fig. 2c).

We further isolated target cDNA from other CML cell lines, including K562, KU812, MEG-01s, and NCO2, and the purified cDNA fragments were subjected to GAIIx sequencing. As in the case for KCL-22-SR, 86–88% of the obtained reads were successfully mapped to the target cDNA in each cell line (Table S2).

Screening of fusion cDNA. Our target set of 913 genes did not include *BCR*, but it did contain *ABL1*. Thus, if we were able to isolate sequence reads encompassing the fusion point of *BCR–ABL1*, cDNA-capture approaches for a given gene set would likely be able to detect gene fusions to unknown partners. In fact, we detected 45 sequence reads for KCL-22-SR cells that covered the *BCR–ABL1* fusion point (Fig. 3a). Likewise, the sequence datasets for K562, KU812, MEG-01s, and NCO2 cells contained 53, 8, 11, and 10 such fusion reads, respectively (data not shown). Furthermore, our sequence data faithfully recapitulated two variants of *BCR–ABL1* cDNA in these cell lines; a fusion variant between exon 13 of *BCR* and exon 2 of *ABL1* was detected in KCL-22-SR, MEG-01s, and NCO2 cells, whereas a fusion variant between exon 14 of *BCR* and exon 2 of *ABL1* was detected in K562 and KU812 cells.[14]

In addition to *BCR–ABL1*, we identified 72 independent candidates for fusion cDNA (including fusions to non-coding RNA) from the CML cell lines. Surprisingly, however, the screening of fusion genes among the unselected cDNA of KCL-22-SR with our rather non-stringent threshold (≥4 reads mapped to a candidate fusion point) failed to isolate *BCR–ABL1* cDNA. We could not even detect any fusion candidates (involving one of our target genes in either or both ends of fusion events) from this dataset, while a total of nine candidates (including *BCR–ABL1*) were isolated from the captured cDNA of the same cell line.

Our Bowtie mapping of both ends of each read to human mRNA or genome databases (Fig. S1) resulted in the detection of not only *BCR–ABL1* fusions, but also a large number of alternatively-spliced messages. From the captured cDNA of KCL-22-SR, for instance, we could detect 79 alternatively-spliced transcripts for 72 independent genes (data not shown). In contrast, from the unselected cDNA of the same cell line, only three independent, alternatively-spliced transcripts were identified among three genes within the 913 targets.

One such example of alternatively-spliced message was *MLL* (ensemble accession no.: ENST00000389506) in KU812, MEG-01s, and K562 cells. In addition to a set of reads that completely matched exon 3 of *MLL*, we obtained reads that lacked an internal 2193-bp sequence in exon 3 (Fig. 3b). Such in-frame truncation would be expected to generate an MLL protein lacking amino acids 276–1006 of the wild-type protein. To confirm the presence of such transcripts, we performed RT–PCR analysis with total RNA from KU812 cells, and PCR primers designed as in Figure 3b. The combination of the F-1 and R primers would be expected to yield both the wild-type (2536 bp) and truncated

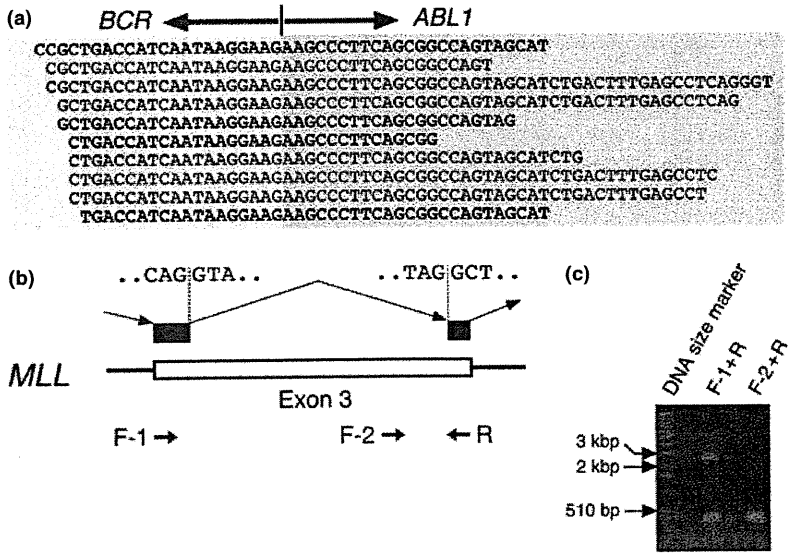

Fig. 3. Detection of gene fusions and alternative mRNA splicing in CML cells. (a) Our computational pipeline yielded 45 reads for KCL-22-SR cells that encompassed the fusion point of breakpoint cluster region (*BCR*)-Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) cDNA, some of which are shown aligned. Reads in the sense or antisense strand are designated in black and blue letters, respectively, and the *BCR* and *ABL1* portions of the sequences are shaded differentially. (b) Some of the reads that mapped to exon 3 of mixed-lineage leukemia (*MLL*) skipped a 2193-bp region within this exon. Nucleotide sequences of the cryptic splicing sites are shown, as are the positions of PCR primers used to confirm the alternative splicing. (c) Gel electrophoresis of the RT–PCR products obtained with total RNA isolated from KU812 cells and with either the F-1 and R primer pair or the F-2 and R primer pair. A 1-kb ladder of DNA size markers was also included.

(343 bp) products, whereas that of the F-2 and R primers would yield only the wild-type product of 339 bp. Gel electrophoresis of the RT–PCR products confirmed the presence of the truncated mRNA (Fig. 3c). Given that the donor and acceptor sites for this alternative splicing harbor the consensus sequences for mRNA splicing (Fig. 3b), some CML cells likely make use of such cryptic splicing sites after *MLL* transcription.

Other variants. From the captured cDNA for KCL-22-SR, NCO2, MEG-01s, K562, and KU812 cells, we detected 156, 18, 28, 23, and 21 non-synonymous mutations among the 913 target genes, respectively. An analysis of the unselected cDNA from KCL-22-SR, however, identified only 19 mutations within the target genes, 16 of which were discovered in the captured cDNA as well. Comparison of the read sequences from the unselected KCL-22-SR cDNA to all RefSeq exonic sequences discovered a total of 597 non-synonymous mutations.

Furthermore, 19, eight, four, 11, and two indels were detected with the captured cDNA of KCL-22-SR, NCO2, MEG-01s, K562, and KU812, respectively. Most of the detected indels were only 1 bp in length, whereas the others were either 2 or 3 bp (Fig. S3). Detailed analysis of these nucleotide changes will be described elsewhere (Toshihide Ueno and Yoshihiro Yamashita, personal communication).

One of the most frequent genetic changes in the blast crisis of CML is point mutation or loss (or both) of *TP53*.[21] Indeed, our sequence data for this gene revealed non-synonymous point mutations in NCO2 and KU812 cells, a 1-bp insertion in K562 cells, a 1-bp deletion in KCL-22-SR cells, and a 3-bp deletion in MEG-01s cells (Fig. 4; Fig. S4; Table S3), all of which were confirmed by Sanger sequencing (data not shown). In NCO2 cells, for instance, 100% of *TP53* reads harbored a G-to-C substitution at nucleotide position 993 of *TP53* mRNA (GenBank accession no.: NM_000546), resulting in a glycine-to-arginine amino acid change (Fig. 4a). The data were also indicative of loss of heterozygosity for *TP53* in NCO2 cells. Similarly, 75% or 78% of *TP53* reads contained a C insertion or a CAC deletion in K562 (Fig. 4b) or MEG-01s (Fig. S4) cells, respectively.

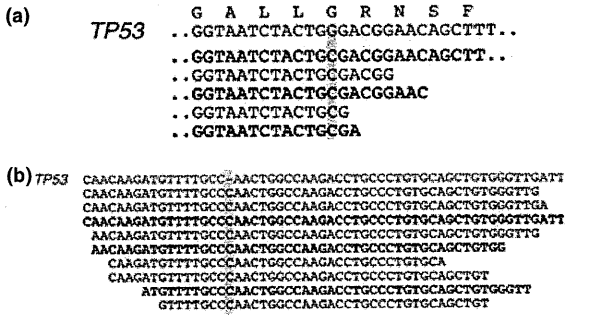

Fig. 4. Anomalies in *TP53* in CML cell lines. (a) Read sequences for NCO2 cells are shown aligned with the reference nucleotide and predicted amino acid sequences (red letters) for *TP53*, revealing a G-to-C substitution in all the reads. Sense or antisense strands are denoted in black and blue letters, respectively. (b) Alignment of the read sequences for K562 cells with the cDNA sequence of *TP53* as in (a), revealing a C insertion.

Discussion

We have shown that a cDNA-capture system, coupled with massively parallel sequencing, is a feasible and relatively simple approach to the simultaneous detection of point mutations, indels, and gene fusions in target cDNA. There are, however, both advantages and disadvantages of this technique compared with the conventional exon-capture system for genomic DNA.

The ability to detect gene fusions, in addition to other mutations with a single sequencing reaction, is one of the most important benefits of the cDNA-capture approach. Furthermore, the efficiency of exon capture with genomic DNA is dependent on the sequence context of each exon. The mean exon size for the human genome is only <200 bp, and the efficiency of exon purification is markedly affected by GC content and sequence complexity.[22] In contrast, even exons with a high GC content might be well isolated by the cDNA-capture system if adjacent exons have a normal GC content and are efficiently targeted by hybridization probes.

Levin *et al.*[20] conducted deep sequencing of captured cDNA for K562 cells, and identified five candidates for fusion genes in addition to *BCR–ABL1*. However, we could not detect any of the five candidates through our analysis with K562, probably because our 913 target genes did not contain those involved in the gene fusions in their report, other than nascent polypeptide-associated complex alpha subunit (*NACA*). While Levin *et al.* discovered primase, DNA, polypeptide 1 (*PRIM1*)–*NACA* fusion transcripts, the low expression level of *PRIM1–NACA* in K562 (only 2.5% of that of *BCR–ABL1* in their dataset)[20] might account for the failure in our analysis.

doi: 10.1111/j.1349-7006.2011.02105.x

However, for experiments based on capture of genomic DNA, sequencing a paired normal specimen allows the efficient subtraction of rare SNP not present in the current databases from the dataset of cancer tissue. This is not always the case, however, for the cDNA-capture approach, given that gene expression profiles differ markedly among samples (even among those obtained from the same individual). Genes with sequence alterations in the cancer specimen might not be expressed in a given normal specimen, and it is not possible to readily determine whether such alterations are germ-line polymorphisms, while algorithms to predict the effect on protein functions for a given amino acid change are currently available[23] and synonymous-to-non-synonymous ratio of nucleotide alterations for a given gene/dataset might provide clues as to how such changes are selected in tumor cells.[24]

In addition, the cDNA-capture system cannot obtain a sufficient number of reads for genes expressed at a low level, and the overall sensitivity of cDNA capture is dependent on the total read number provided by sequencers. We are able to run only two samples per flow cell of the GAIIx system, whereas up to eight samples can be run in a single flow cell for whole exome sequencing of human genomic DNA.

Despite such limitations, our study shows that cDNA capture is an efficient process, and extensive sequencing of such purified cDNA is a straightforward approach to interrogate the target cDNA for various genetic changes in a single platform. Large-scale resequencing of hundreds of cancer specimens might thus become within the scope of private laboratories with the adoption of the cDNA-capture approach.

Acknowledgments

This study was supported in part by grants for Research on Human Genome Tailor-Made and for Third-Term Comprehensive Control Research for Cancer from the Ministry of Health, Labor, and Welfare of Japan; Grants-in-Aid for Scientific Research (B) and for Young Scientists (A) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan; and by grants from the Japan Society for the Promotion of Science, Takeda Science Foundation, the Naito Foundation, Sankyo Foundation of Life Science, The Sagawa Foundation for Promotion of Cancer Research, the Yasuda Medical Foundation, the Mitsubishi Foundation, and Kobayashi Foundation for Cancer Research.

Disclosure Statement

K. Fukumura, M. Ando, M. Kawazu, Y.L. Choi and H. Mano belong to the Department of Medical Genomics, Graduate School of Medicine, University of Tokyo, which receives research funding from Illumina Inc.

References

1 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719–24.
2 Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011; **470**: 198–203.
3 Ledford H. Big science: the cancer genome challenge. *Nature* 2010; **464**: 972–4.
4 Pleasance ED, Cheetham RK, Stephens PJ *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010; **463**: 191–6.
5 Ley TJ, Mardis ER, Ding L *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.
6 Totoki Y, Tatsuno K, Yamamoto S *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* 2011; **43**: 464–9.
7 Wei X, Walia V, Lin JC *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* 2011; **43**: 442–6.
8 Otto EA, Hurd TW, Airik R *et al.* Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet* 2010; **42**: 840–50.
9 Bilguvar K, Ozturk AK, Louvi A *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010; **467**: 207–10.
10 Mitelman F. Recurrent chromosome aberrations in cancer. *Mutat Res* 2000; **462**: 247–53.
11 Soda M, Choi YL, Enomoto M *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* 2007; **448**: 561–6.
12 Tomlins SA, Rhodes DR, Perner S *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; **310**: 644–8.
13 Ohmine K, Nagai T, Tarumoto T *et al.* Analysis of gene expression profiles in an imatinib-resistant cell line, KCL22/SR. *Stem Cells* 2003; **21**: 315–21.
14 Drexler HG, MacLeod RA, Uphoff CC. Leukemia cell lines: *in vitro* models for the study of Philadelphia chromosome-positive leukemia. *Leuk Res* 1999; **23**: 207–15.
15 Choi YL, Tsukasaki K, O'Neill MC *et al.* A genomic analysis of adult T-cell leukemia. *Oncogene* 2007; **26**: 1245–55.
16 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**: R25.
17 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–10.
18 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**: 1754–60.
19 Fujita PA, Rhead B, Zweig AS *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2011; **39**: D876–82.
20 Levin JZ, Berger MF, Adiconis X *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* 2009; **10**: R115.
21 Calabretta B, Perrotti D. The biology of CML blast crisis. *Blood* 2004; **103**: 4010–22.
22 Shen P, Wang W, Krishnakumar S *et al.* High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci USA* 2011; **108**: 6549–54.
23 Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–9.
24 Babenko VN, Basu MK, Kondrashov FA, Rogozin IB, Koonin EV. Signs of positive selection of somatic mutations in human cancers detected by EST sequence analysis. *BMC Cancer* 2006; **6**: 36.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Algorithm of the computational pipeline.

Fig. S2. Read number distribution of all poly(A)-RNA data.

Fig. S3. Numbers of 1-, 2-, or 3-bp indels for the entire dataset.

Fig. S4. A CAG-deletion in the *TP53* message in MEG-01s cells.

Table S1. Gene list for the custom cDNA-capture system.

Table S2. Purification of the target cDNA in CML cell lines.

Table S3. TP53 mutation status in CML cell lines.

nature medicine

RET, ROS1 and ALK fusions in lung cancer

Kengo Takeuchi[1,2], Manabu Soda[3], Yuki Togashi[1,2], Ritsuro Suzuki[4], Seiji Sakata[1], Satoko Hatano[1], Reimi Asaka[1,2], Wakako Hamanaka[2], Hironori Ninomiya[2], Hirofumi Uehara[5], Young Lim Choi[6], Yukitoshi Satoh[5,7], Sakae Okumura[5], Ken Nakagawa[5], Hiroyuki Mano[3,6] & Yuichi Ishikawa[2]

Through an integrated molecular- and histopathology-based screening system, we performed a screening for fusions of anaplastic lymphoma kinase (ALK) and c-ros oncogene 1, receptor tyrosine kinase (ROS1) in 1,529 lung cancers and identified 44 ALK-fusion–positive and 13 ROS1-fusion–positive adenocarcinomas, including for unidentified fusion partners for ROS1. In addition, we discovered previously unidentified kinase fusions that may be promising for molecular-targeted therapy, kinesin family member 5B (KIF5B)-ret proto-oncogene (RET) and coiled-coil domain containing 6 (CCDC6)-RET, in 14 adenocarcinomas. A multivariate analysis of 1,116 adenocarcinomas containing these 71 kinase-fusion–positive adenocarcinomas identified four independent factors that are indicators of poor prognosis: age ≥50 years, male sex, high pathological stage and negative kinase-fusion status.

Echinoderm microtubule associated protein like 4 (EML4)-ALK was the first targetable fusion oncokinase to be identified in non–small cell lung cancer (NSCLC)[1]. This fusion is found in approximately 4–6% of lung adenocarcinomas[2,3]. ROS1 is another receptor tyrosine kinase that forms fusions in NSCLC[4]. Solute carrier family 34 (sodium phosphate), member 2 (SLC34A2)-ROS1 and CD74 molecule, major histocompatibility complex, class II invariant chain (CD74)-ROS1 were identified in 1 out of 41 NSCLC cell lines and 1 out of 150 lung cancer samples, respectively[4]. However, the oncogenic ability of these ROS1 fusion proteins and the incidence of ROS1 fusions in lung cancers are still unclear.

We screened for known and unknown kinase fusions in lung cancers using a histopathology-based system with tissue microarrays of 1,528 surgically removed tissues (**Supplementary Methods** and **Supplementary Appendix**). Immunohistochemistry of antibodies to ALK using the intercalated antibody-enhanced polymer method[2,3,5–7] detected 45 tumors with ALK kinase domain expression (**Supplementary Fig. 1**). In 44 adenocarcinomas, multiplex RT-PCR[2,3] identified 41 *EML4-ALK*–positive and 3 *KIF5B-ALK*–positive adenocarcinomas, including a previously unidentified KIF5B-ALK fusion variant, K17;A20 (**Supplementary Table 1**). Further, we used fluorescence *in situ* hybridization (FISH) for split and fusion assays to confirm the presence of ALK fusions[2,3,8]. The FISH results for the *ALK* split assay, the *EML4-ALK* fusion assay and the *KIF5B-ALK* fusion assay in the 44 adenocarcinomas were all consistent with the presence of the corresponding fusion gene (**Supplementary Figs. 2** and **3**). The remaining tumor that was positive for antibodies to ALK as determined by immunohistochemistry (a large-cell neuroendocrine carcinoma) was negative in the FISH assays and expressed wild-type ALK. ALK fusions existed in 3.0% (44 out of 1,485) of the NSCLCs and 3.9% (44 out of 1,121) of the adenocarcinomas. We included 20 previously reported ALK-fusion–positive and 304 ALK-fusion–negative tumors, all of which were screened with multiplex RT-PCR. Because specimens of these 324 patients were collected consecutively during the period of tissue collection, they served as positive and negative controls, respectively[1–3,8,9]. The immunohistochemistry results using the intercalated antibody-enhanced polymer method were complete matches in the 20 fusion-positive and the 304 fusion-negative tumors.

We used split FISH assays for the screening for *ROS1* gene rearrangement (**Fig. 1**). In 11 of the 13 *ROS1* split FISH-positive tumors (**Fig. 1a**), 5′ rapid amplification of complementary DNA ends (5′ RACE) identified two known and three unknown fusion partners for *ROS1*: *TPM3*, *SDC4*, *SLC34A2*, *CD74* and *EZR* (**Fig. 1b** and **Supplementary Table 1**); RT-PCR confirmed this finding (**Fig. 1c**). In a 5′-RACE–negative tumor (ROS#12) (again, where split FISH is used to detect candidate fusion genes of interest by the presence of rearrangements and RACE is used for the identification of fusion partners), each fusion-specific RT-PCR (using a common reverse primer) amplified the same band, which contained an *LRIG3* sequence. This tumor was proven fusion-positive in RT-PCR specific to *LRIG3-ROS1*, an unidentified fusion. Fusion FISH results confirmed that all 12 cases harbored the corresponding fusion (**Fig. 1a**). All fusion FISH assays for these six ROS1 fusions were negative for the tumor ROS#13 (the frozen material had been consumed), indicating an unknown fusion partner for *ROS1*. *ROS1* split FISH screening failed for nine NSCLCs, including five adenocarcinomas. We identified ROS1 fusions in 0.9% (13 out of 1,476) of the NSCLCs and 1.2% (13 out of 1,116) of the adenocarcinomas.

We performed *KIF5B* split FISH to discover new fusion kinases, as we previously identified KIF5B-ALK fusions in lung cancer[3]. As such, we hypothesized that KIF5B might be rearranged in lung cancer. In 24 *KIF5B* split FISH-positive tumors, 3′ RACE identified an in-frame fusion between *KIF5B* exon 23 and *RET* exon 12

[1]Pathology Project for Molecular Targets, the Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [2]Division of Pathology, the Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan. [3]Division of Functional Genomics, Jichi Medical University, Tochigi, Japan. [4]Department of Hematopoietic Stem Cell Transplantation Data Management and Biostatistics, Nagoya University Graduate School of Medicine, Nagoya, Japan. [5]Department of Thoracic Surgical Oncology, Thoracic Center, the Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan. [6]Department of Medical Genomics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [7]Present address: Department of Thoracic Surgery, Kitasato University School of Medicine, Kanagawa, Japan. Correspondence should be addressed to K.T. (kentakeuchi-tky@umin.net).

Received 28 September 2011; accepted 3 January 2012; published online 12 February 2012; doi:10.1038/nm.2658

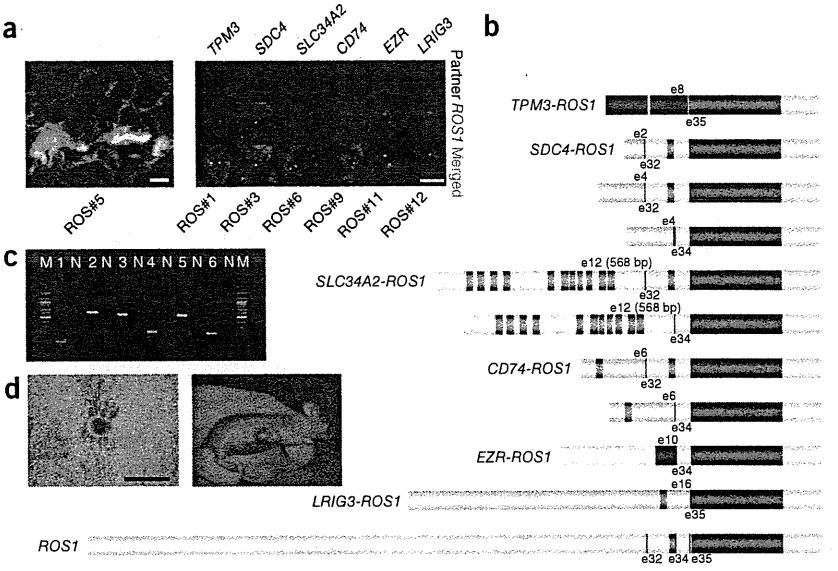

Figure 1 Identification of ROS1 fusions. (**a**) *ROS1* split (left) and fusion (right) FISH assay data (scale bars, 20 μm). In the split assay, multiple tumor cells harbored individual 3′ side signals (green), indicating the presence of a *ROS1* rearrangement. In the fusion assay, a fusion signal (yellow) was observed in the representative tumor cell of each subject, which is consistent with the presence of t(1;6)(q21.2;q22) for *TPM3-ROS1*, t(6;20)(q22;q12) for *SDC4-ROS1*, t(4;6)(q15.2;q22) for *SLC34A2-ROS1*, t(5;6)(q32;q22) for *CD74-ROS1*, inv(6)(q22q25.3) for *EZR-ROS1* or t(6;12)(q22;q14.1) for *LRIG3-ROS1*. (**b**) The break points of ROS1 are exons 32, 34 and 35. All of the break points allow the resulting fusion to harbor the kinase domain of ROS1 (red), and the exon 32 break point allows the resulting fusion to harbor the transmembrane domain of ROS1 (orange). In the fusion partners, dark blue and orange represent coiled-coil and transmembrane domains, respectively. Coiled-coil domains may contribute to homodimelization, but only TPM3 and EZR contained these domains. In contrast to ALK and RET fusions, the role of the fusion partner's coiled-coil domain is unknown in ROS1 fusions. e, exon. (**c**) Results for fusion-specific RT-PCR for tumors ROS#1 (lane 1, TPM3-ROS1, T8;R35, predicted product size of 119 bp), ROS#3 (lane 2, SDC4-ROS1, S2;R32, 596 bp), ROS#6 (lane 3, SLC34A2-ROS1, S13del2046;R32 and S13del2046;R34, 544 bp and 235 bp, respectively), ROS#8 (lane 4, CD74-ROS1, C6;R34, 230 bp), ROS#10 (lane 5, EZR-ROS1, E10;R34, 527 bp), and ROS#12 (lane 6, LRIG3-ROS1, L16;R35, 218 bp). M and N represent the size marker (100-bp ladder) and the non-template control, respectively. (**d**) The transforming potential of the ROS1 fusion. Mouse 3T3 fibroblasts infected with a retrovirus encoding SDC4-ROS1 derived from tumor ROS#4 formed multiple foci (scale bar, 1 mm). All of the four nude mice injected with the corresponding 3T3 cells developed a subcutaneous tumor (right).

(tumor RET#11). *RET* split FISH on the tissue arrays identified 22 fusion-positive tumors in 1,528 lung cancers (**Fig. 2a**), from which a multiplex RT-PCR system that captures all possible *KIF5B-RET* fusions detected 12 fusion-positive tumors: eight tumors with the fusion of *KIF5B* exon 15 and *RET* exon 12 (K15;R12) and one tumor each with the K16;R12, K22;R12, K23;R12 and K24;R11 fusions (**Fig. 2b** and **Supplementary Table 1**). The *KIF5B-RET* fusion FISH results were consistent with the presence of inv(10)(p11.22q11.2) in all 12 of these tumors (**Fig. 2a**).

In a routine histopathological diagnosis, we encountered an adenocarcinoma that showed a mucinous cribriform pattern (**Fig. 2c**) that was previously reported as a histopathological marker for the presence of EML4-ALK (**Supplementary Fig. 4**)[9–11]. Notably, this adenocarcinoma (tumor RET#14) was negative for ALK fusion and was positive for *CCDC6-RET*, as determined by FISH and inverse RT-PCR; the latter fusion gene was first described in thyroid cancer[12]. RT-PCR identified another tumor positive for the *CCDC6-RET* fusion (RET#13) in the remaining 10 tumors. The 14 RET-positive tumors (out of the total 1,528 tumors tested, with one additional tumor (RET#14) found through routine pathology diagnostic service) were also positive in the revised multiplex RT-PCR that captured *EML4-ALK*, *KIF5B-ALK*, *KIF5B-RET* and *CCDC6-RET* simultaneously (**Fig. 2d**). The RET kinase domain expression using real-time RT-PCR was weak or undetectable for the remaining nine tumors determined to be positive in the *RET* split FISH screening. Perhaps the genomic rearrangement occurred downstream of the *RET* break points. *RET* split FISH screening failed in three NSCLCs, including two adenocarcinomas. RET#14 was the index case found in routine pathology diagnostic service but not in the 1,528 cohort. RET fusions existed in 0.9% (13 out of 1,482) of the NSCLCs and 1.2% (13 out of 1,119) of the adenocarcinomas. The 14 RET fusion-positive subjects did not receive vandetanib.

We concluded that the rearrangements described above are somatic without using any matched normal tissues. Our histopathology-based screening method preserves the samples' histological architecture. This allows observers to confirm that internal non-tumor cells, for example, epithelial cells, inflammatory cells or fibroblasts, are negative in a test of interest.

All 71 kinase-fusion–positive (44 ALK, 13 ROS1 and 14 RET fusions) lung cancers were exclusively adenocarcinomas (6% of all adenocarcinomas in the present study), were positive for antibodies to TTF1, which is regarded as a marker for lung adenocarcinoma, as determined by immunohistochemistry (excluding two ALK-positive tumors) and were negative for *EGFR* and *KRAS* mutations. Thirteen of the 44 ALK-positive tumors (30%) were weakly positive for p63 expression (were weakly positive for a squamous cell carcinoma marker, p63) (**Supplementary Table 1**). Thirty-three tumors showed a mucinous cribriform pattern in at least 5% of their area; 22 tumors had this pattern in >25% of their area (**Fig. 2c**, **Supplementary Table 1** and **Supplementary Fig. 4**). The frequency of mucinous cribriform carcinoma was significantly higher in the kinase-fusion–positive group of tumors than in the 77 fusion-negative adenocarcinomas (22 out of 71 compared to 7 out of 77, respectively; $P = 0.00088$). Notably, we observed this pattern preferentially in EML4-ALK–positive tumors (70%, 29 out of 41); all three CD74-ROS1–positive tumors also showed this pattern. Recognizing this pattern in routine pathology diagnoses led to the identification of the CCDC6-RET fusion (tumor RET#14). In organs other than the lung, secretory breast carcinoma, which is characterized by a cribriform pattern with abundant secretory material, harbors the ets variant 6 (ETV6)-neurotrophic tyrosine kinase, receptor, type 3 (NTRK3) fusion (ref. 13). We identified an ALK-fusion–positive renal cell carcinoma that showed a mucinous cribriform pattern[7]. This pattern may be linked to the presence of particular kinase fusions[10], and this possibility warrants further study.

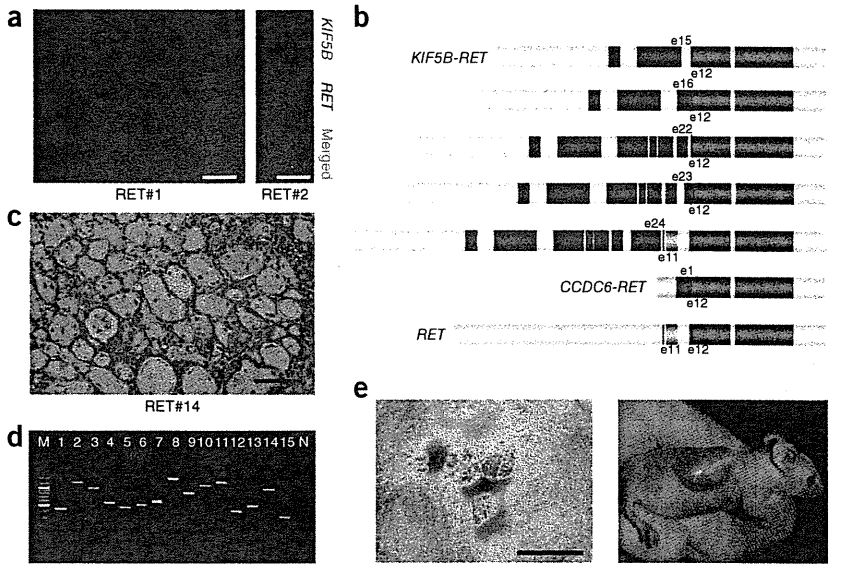

Figure 2 Discovery of RET fusions. (**a**) *RET* split (left) and fusion (right) FISH assay data (scale bars, 20 μm). In the split assay, multiple tumor cells harbored individual 3′ side signals (green), indicating the presence of *RET* rearrangement. In the fusion assay, a fusion signal (yellow) was observed in the representative tumor cell of subject RET#2, which is consistent with the presence of inv(10)(p11.22q11.2). (**b**) The break points of RET are exons 11 and 12. Both of the break points allow the resulting fusion to harbor the kinase domain of RET (red), and the exon 11 break point allows the resulting fusion to harbor the transmembrane domain of RET (orange). In the fusion partners, dark blue represents a coiled-coil domain, which probably contributes to the homodimelization of the fusion. Only the longer isoforms of RET and the RET fusions are shown. (**c**) Subject RET#14 showed the representative histopathology of mucinous cribriform carcinoma (scale bar, 100 μm). (**d**) The results for fusion-specific RT-PCR for subjects ALK#10 (lane 1, EML4-ALK, E13;A20, predicted product size of 432 bp), ALK#16 (lane 2, EML4-ALK, E20;A20, 1185 bp), ALK#26 (lane 3, EML4-ALK, E6;A20, 913 bp), ALK#38 (lane 4, EML4-ALK, E14;ins11del49A20, 546 bp), ALK#39 (lane 5, EML4-ALK, E2;A20, 454 bp), ALK#40 (lane 6, EML4-ALK, E13;ins69A20, 501 bp), ALK#41 (lane 7, EML4-ALK, E14;del14A20, 570 bp), ALK#42 (lane 8, KIF5B-ALK, K17;A20, 1,483 bp), ALK#44 (lane 9, KIF5B-ALK, K24;A20, 814 bp), RET#6 (lane 10, KIF5B-RET, K15;R12, 1,104 bp), RET#9 (lane 11, KIF5B-RET, K16;R12, 1,293 bp), RET#10 (lane 12, KIF5B-RET, K22;R12, 420 bp), RET#11 (lane 13, KIF5B-RET, K23;R12, 525 bp), RET#12 (lane 14, KIF5B-RET, K24;R11, 999 bp) and RET#13 (lane 15, CCDC6-RET, C1;R12, 352 bp). M and N represent the size marker (100-bp ladder) and non-template control, respectively. (**e**) The transforming potential of the KIF5B-RET fusion. Mouse 3T3 fibroblasts infected with a retrovirus encoding K15;R12L derived from tumor RET#7 fomed multiple foci (scale bar, 1 mm). All of the four nude mice injected with the corresponding 3T3 cells developed a subcutaneous tumor (right).

Supplementary Tables 1–4 summarize the clinicopathological features of the subjects. Briefly, young age, low smoking index and small tumor size characterized the kinase-fusion–positive group of subjects (**Supplementary Table 2**). A multivariate analysis of the adenocarcinomas revealed four independent factors that were indicators of poor prognosis: age ≥50 years, male sex, high pathological stage and negative kinase-fusion status (**Supplementary Table 3**). There was no significant difference in overall survival between the kinase-positive and epidermal growth factor receptor (EGFR)-mutant groups (P = 0.32). **Supplementary Table 4** shows the clinicopathological features of the subjects stratified by each fusion.

The transforming ability of CCDC6-RET and all of the ALK fusions, excluding K17;A20, was shown previously[1–3,8,12]. 3T3 cells infected with a virus expressing K17;A20, tropomyosin 3 (TPM3)-ROS1, syndecan 4 (SDC4)-ROS1, SLC34A2-ROS1, CD74-ROS1, ezrin (EZR)-ROS1, leucine-rich repeats and immunoglobulin-like domains 3 (LRIG3) (transcript variant 2)-ROS1 or KIF5B-RET (with both the longer (RET51) and shorter (RET9) RET isoforms) led to multiple transformed foci formation in culture and in subcutaneous tumors in a nude mouse tumorigenicity assay (**Figs. 1d, 2e** and **Supplementary Fig. 5**).

To test whether vandetanib, an inhibitor of vascular endothelial growth factor receptor (VEGFR-2), VEGFR-3, EGFR and RET[14], might be effective for the treatment of RET-fusion–positive tumors, we induced Flag-tagged EML4-ALK (E13;A20) or KIF5B-RET (K15;R12L and K15;R12S) in Ba/F3 cells, which are dependent on interleukin-3 (IL-3) for growth. All transfected cells, including those without any kinase fusion, proliferated in the presence of IL-3, but only cells expressing E13;A20 or K15;R12L grew in the absence of IL-3 (**Supplementary Fig. 6a**). In the absence of IL-3, vandetanib inhibited the proliferation of cells expressing K15;R12L (**Supplementary Fig. 6c**) but not the proliferation of cells expressing E13;A20 (**Supplementary Fig. 6d**). Crizotinib was not effective in inhibiting the proliferation of Ba/F3 cells expressing K15;R12L (**Supplementary Fig. 7**).

In 1985, a 3T3 assay identified *RET* as a rearranged transforming gene[15]. RET fusions have been identified exclusively in papillary thyroid carcinoma and are more frequently observed in radiation-associated thyroid cancers (for example, in survivors of the Chernobyl accident[16], atomic bomb survivors[17] and post-radiation therapy patients[18]). Therefore, a retrospective comparison of RET fusions in individuals with lung cancer with and without a history of radiation exposure warrants further study. If a positive association is found between RET fusion and radiation exposure in these studies, it might be desirable for individuals with internal or therapeutic exposure to irradiation (for example, those individuals involved in the Fukushima accident) to be monitored prospectively for lung cancer as well as thyroid cancer.

In Japan, more than 40% of lung adenocarcinomas in younger individuals harbor EGFR mutations[19]. In this study, 16% (17 out of 107) of younger individuals (≤50 years of age) with adenocarcinoma harbored a kinase fusion. Collectively, as long as molecular target diagnoses are properly performed, >50% of the individuals with lung adenocarcinoma in this generation may benefit from treatment with corresponding kinase inhibitors. Integrated pathology-based screening techniques can also be used for the selection of individuals to receive this treatment[20]. The results of our study will facilitate the development of a molecular classification of lung adenocarcinomas that is closely related to both the pathogenesis and the treatment of disease. This study was approved by the Institutional Review Board of the Cancer Institute Hospital, and all subjects provided informed consent.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemedicine/.

Note: Supplementary information is available on the Nature Medicine website.

ACKNOWLEDGMENTS

We thank M. Iwakoshi, K. Shiozawa, T. Kakita, H. Nagano and K. Nomura for their technical assistance and S. Sengoku for providing administrative assistance. This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan, as well as by grants from the Japan Society for the Promotion of Science; the Ministry of Health, Labor and Welfare of Japan; the Vehicle Racing Commemorative Foundation of Japan; the Princess Takamatsu Cancer Research Fund; and the Uehara Memorial Foundation.

AUTHOR CONTRIBUTIONS

K.T. conceived of and led the entire project, designed the FISH probes, screened samples using FISH and immunohistochemistry, performed histopathological analyses, generated figures and tables and wrote the manuscript. M.S. performed functional analyses and generated the figures. Y.T. performed inverse RT-PCR and RACE experiments and their corresponding analyses. R.S. conducted statistical analyses. S.S. performed FISH and histopathological analyses. S.H. processed and analyzed the tissue microarrays and FISH screening and generated figures. R.A. processed the FISH probe library. W.H. made and analyzed the database and processed tissue microarrays. H.N., H.U., Y.S., S.O. and K.N. collected specimens and clinical information and were involved in planning the project. Y.L.C. conducted functional analyses. H.M. supervised the functional analyses and planned the project. Y.I. performed histopathological analyses and collected specimens. All authors participated in the discussion and interpretation of the data and the results.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemedicine/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Soda, M. *et al. Nature* **448**, 561–566 (2007).
2. Takeuchi, K. *et al. Clin. Cancer Res.* **14**, 6618–6624 (2008).
3. Takeuchi, K. *et al. Clin. Cancer Res.* **15**, 3143–3149 (2009).
4. Rikova, K. *et al. Cell* **131**, 1190–1203 (2007).
5. Takeuchi, K. *et al. Haematologica* **96**, 464–467 (2011).
6. Takeuchi, K. *et al. Clin. Cancer Res.* **17**, 3341–3348 (2011).
7. Sugawara, E. *et al. Cancer* published online, doi:10.1002/cncr.27391 (17 January 2012).
8. Choi, Y.L. *et al. Cancer Res.* **68**, 4971–4976 (2008).
9. Inamura, K. *et al. J. Thorac. Oncol.* **3**, 13–17 (2008).
10. Takeuchi, K. *Pathol. and Clin. Med.* **28**, 139–144 (2010).
11. Jokoji, R. *et al. J. Clin. Pathol.* **63**, 1066–1070 (2010).
12. Grieco, M. *et al. Cell* **60**, 557–563 (1990).
13. Tognon, C. *et al. Cancer Cell* **2**, 367–376 (2002).
14. Flanigan, J., Deshpande, H. & Gettinger, S. *Biologics* **4**, 237–243 (2010).
15. Takahashi, M., Ritz, J. & Cooper, G.M. *Cell* **42**, 581–588 (1985).
16. Ito, T. *et al. Lancet* **344**, 259 (1994).
17. Hamatani, K. *et al. Cancer Res.* **68**, 7176–7182 (2008).
18. Bounacer, A. *et al. Oncogene* **15**, 1263–1273 (1997).
19. Kosaka, T. *et al. Cancer Res.* **64**, 8919–8923 (2004).
20. Han, B. *et al. Cancer Res.* **68**, 7629–7637 (2008).

Cancer Science
The official journal of the Japanese Cancer Association JCA

High-throughput resequencing of target-captured cDNA in cancer cells

Toshihide Ueno,[1,5] Yoshihiro Yamashita,[1,5] Manabu Soda,[1] Kazutaka Fukumura,[2] Mizuo Ando,[2] Azusa Yamato,[1] Masahito Kawazu,[2] Young Lim Choi[1,2] and Hiroyuki Mano[1,2,3,4]

[1]Division of Functional Genomics, Jichi Medical University, Tochigi; [2]Department of Medical Genomics, Graduate School of Medicine, University of Tokyo, Tokyo; [3]CREST Japan Science and Technology Agency, Saitama, Japan

(Received May 30, 2011/Revised September 7, 2011/Accepted September 14, 2011/Accepted manuscript online September 20, 2011/Article first published online October 13, 2011)

The recent advent of whole exon (exome)-capture technology, coupled with second-generation sequencers, has made it possible to readily detect genomic alterations that affect encoded proteins in cancer cells. Such target resequencing of the cancer genome, however, fails to detect most clinically-relevant gene fusions, given that such oncogenic fusion genes are often generated through intron-to-intron ligation. To develop a resequencing platform that simultaneously captures point mutations, insertions–deletions (indels), and gene fusions in the cancer genome, we chose cDNA as the input for target capture and extensive resequencing, and we describe the versatility of such a cDNA-capture system. As a test case, we constructed a custom target-capture system for 913 cancer-related genes, and we purified cDNA fragments for the target gene set from five cell lines of CML. Our target gene set included Abelson murine leukemia viral oncogene homolog 1 (*ABL1*), but it did not include breakpoint cluster region (*BCR*); however, the sequence output faithfully detected reads spanning the fusion points of these two genes in all cell lines, confirming the ability of cDNA capture to detect gene fusions. Furthermore, computational analysis of the sequence dataset successfully identified non-synonymous mutations and indels, including those of tumor protein p53 (*TP53*). Our data might thus support the feasibility of a cDNA-capture system coupled with massively parallel sequencing as a simple platform for the detection of a variety of anomalies in protein-coding genes among hundreds of cancer specimens. (*Cancer Sci* 2012; 103: 131–135)

Cancer is thought to result from various alterations of the genome, including point mutations, insertions–deletions (indels), and genomic rearrangements.[1] Whereas comprehensive sequencing of the cancer genome, or "cancer genome resequencing", is a promising approach to the identification of such anomalies, and to provide a basis for the development of effective treatment strategies for cancer, determination of the nucleotide sequence of the entire human genome with conventional Sanger sequencers remains a highly demanding task. However, the recent advent of massively parallel sequencing systems, or second-generation sequencers, has rendered such projects manageable in private laboratories[2] and triggered the formation of large-scale consortia, such as The Cancer Genome Atlas and International Cancer Genome Consortium,[3] to undertake cancer genome resequencing for hundreds of specimens. Cancer genome resequencing with massively parallel sequencers has already provided a wealth of information on genome-wide mutation status for melanoma,[4] acute myeloid leukemia,[5] hepatocellular carcinoma,[6] and other cancers.

Even with the current massively parallel sequencers, however, the determination and compilation of the full genome sequence for a given sample might still take almost 1 month. Comparison of the cancer genome among many specimens thus remains time-consuming and labor intensive. Anomalies in protein-coding genes likely play a major role in carcinogenesis. Given that exonic regions occupy only ~1.3% of the human genome, sequencing such targeted regions would be expected to markedly facilitate the discovery of proteins that are activated or inactivated specifically in cancer cells. Indeed, target-capture strategies, coupled with massively parallel sequencers, have revealed important genetic changes in cancer,[7] as well as in hereditary disorders.[8,9]

One important drawback of such target-capture approaches, however, is their inability to detect gene fusions. Most cancer-associated gene fusion events occur within introns (resulting in exon-to-exon ligation in the corresponding mRNA), and exon capture does not reveal breakage and ligation of intronic regions. Recurrent gene fusions were once thought to be rare in epithelial tumors compared with hematologic malignancies and sarcomas;[10] however, our recent discovery of the echinoderm microtubule associated protein like-4 (EML4)-anaplastic lymphoma kinase (ALK) fusion gene in lung cancer and the discovery by others of rearrangements in loci for the v-ets avian erythroblastosis virus E26 oncogene homolog (ETS) family of transcription factors in prostate cancer have led to a revision of this notion.[11,12] It would thus be desirable to develop a resequencing platform that is able to capture, within a reasonable timeframe, all gene fusions, point mutations, and indels in the cancer genome. In pursuit of this goal, we have now examined the efficacy of high-throughput sequencing of captured cDNA for the identification of such cancer genome anomalies.

Materials and Methods

Cell lines. Cell lines established from the blast crisis stage of CML, including MEG-01s, KCL-22-SR, K562, NCO2, and KU812,[13,14] were obtained from the Japanese Collection of Research Bioresources (Osaka, Japan) and were maintained in RPMI-1640 medium (Invitrogen, Carlsbad, CA, USA) supplemented with 10% FBS (Invitrogen). Total RNA was isolated from each cell line with the use of an RNeasy mini kit (Qiagen, Valencia, CA, USA) and was subjected to cDNA synthesis with an oligo(dT) primer.

Gene expression profiling. The cDNA prepared from total poly(A)-RNA of KCL-22-SR cells was subjected to hybridization with the HGU95Av2 microarray (Affymetrix, Santa Clara, CA, USA), as described previously.[15] The expression intensity of each test gene on the array was normalized by the 50th percentile value.

cDNA-capture methods. RNA probes of 120 bases were designed to cover (with a 60-base overlap) cDNA of 913 human protein-coding genes (Table S1), and were synthesized by Agilent Technologies (Santa Clara, CA, USA). During the design of the probes, the Repeat Masker dataset (http://www.repeatmasker.org) was used to remove probes corresponding to

[4]To whom correspondence should be addressed. E-mail: hmano@jichi.ac.jp
[5]These authors contributed equally to this work.

doi: 10.1111/j.1349-7006.2011.02105.x

repetitive sequences in the human genome. Hybridization of DNA fragments to the RNA probes was performed according to the protocols recommended for the SureSelect Target Enrichment system (Agilent). We also used the SureSelect Human X Chromosome Demo kit (Agilent) to examine purification efficiency. Purified DNA fragments were then subjected to sequencing with a Genome Analyzer IIx (GAIIx; Illumina, San Diego, CA, USA) for 76 bases from both ends by the paired-end sequencing system.

Computational pipeline. Raw read data were quality filtered on the basis of the presence of the Illumina adaptor sequences and a *Q*-value of ≥20. The resulting read sequences were then subjected to an in-house computational pipeline to identify various mutations (Fig. S1). In brief, read sequences were matched with the Bowtie algorithm[16] to the cDNA sequences of the 913 genes used to construct our custom-made SureSelect system. The matched reads were then examined for the presence of non-synonymous mutations and single nucleotide polymorphisms (SNP) deposited in dbSNP (build 132, http://www.ncbi.nlm.nih.gov/projects/SNP/index.html). The remaining reads were further matched to the cDNA sequences with Burrows-Wheeler Aligner (BWA) and Basic Local Alignment Search Tool (BLAST) algorithms to search for indels and multiple mutations.[17,18] Candidates for non-synonymous mutations were identified only when ≥20% of reads correspond to the mutations at positions with ≥50 coverage.

For the selection of reads corresponding to possible fusion cDNA, nucleotide sequences of 20 bp were obtained from both ends of each read and were separately matched to RefSeq mRNA (http://www.ncbi.nlm.nih.gov), KnownGeneMrna,[19] and the human genome sequence (GRCh37, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/?build=37). Reads were considered to be derived from fusion genes if the ends of a given read matched to different genes within the 913-gene group, or one end matched to a single gene within the 913-gene group and the other end matched to a sequence in RefSeq, KnownGeneMrna, or the human genome sequence that did not correspond to the 913 genes. Candidates for fusion genes were identified only when four or more reads were mapped to possible fusion points.

RT-PCR. To confirm the presence of an alternatively-spliced mixed-lineage leukemia (*MLL*) mRNA, we subjected oligo(dT)-primed cDNA of KU812 cells to PCR with the combination of the F-1 primer (5′-ACCTCGTGGGAGACCTAGAAGTGG-3′) and the R primer (5′-AGTCATTGGAAGCTTGTCTGCCTG-3′), or with the combination of the F-2 primer (5′-CCTGTGGGTAGGGTTTCCAAAGAG-3′) and the R primer.

Results

Efficiency of cDNA-capture sequencing. Paired-end sequencing of target-captured cDNA was briefly described in a previous study;[20] however, how the efficiency of target purification with cDNA compares with that with genomic DNA remains unclear. We therefore attempted to optimize the conditions for cDNA purification with the SureSelect system. Oligo(dT)-primed cDNA of KCL-22-SR cells were fragmented to a mean size of 500 or 200 bp and then subjected to purification with the use of the SureSelect Human X Chromosome Demo kit, which is designed to capture genomic sequences derived from the human X chromosome. Genomic DNA of KCL-22-SR cells was similarly processed and hybridized with the X Chromosome Demo kit. The purified fragments at either 4 or 8 pM were then sequenced by the GAIIx system.

The X chromosome-mapped cDNA reads occupied 62.1%, 81.6%, 62.4%, and 82.2% of quality filter-passed reads for the experiments with 4 pM of 500-bp fragments, 4 pM of 200-bp fragments, 8 pM of 500-bp fragments, and 8 pM of 200-bp fragments, respectively (Fig. 1). Thus, these results suggested that the shorter cDNA fragments were captured more efficiently than the longer ones. Furthermore, the purification efficiency for genomic DNA fragments was not higher than that for cDNA, irrespective of DNA concentration and fragmentation size (Fig. 1), supporting the feasibility of cDNA-capture approaches.

The ability to detect breakpoint cluster region (*BCR*)-Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) fusion reads was reduced for the cDNA sheared to ~200 bp compared with that for those of ~500 bp (see below). The former cDNA detected 83.7% or 76% of the fusion reads detected by the latter cDNA at input concentrations of 4 and 8 pM, respectively. This result is in line with our computational bootstrap trial (n = 10 000) showing that the number of randomly-fragmented, 200-bp reads encompassing the *BCR–ABL1* fusion point is ~2.5 times higher than that of 500-bp reads (data not shown). However, given that the total number of high-quality reads was much higher in the data for the 200-bp cDNA than in those for the 500-bp cDNA (Fig. 1), we chose to use 8 pM of cDNA with a mean size of 200 bp for further experiments.

Custom cDNA-capture system. We also tested whether extensive sequencing of cDNA generated from total poly(A)-RNA (unselected cDNA) might serve to identify gene fusions, point mutations, and indels. For this purpose, unselected cDNA were prepared from KCL22-SR cells, and subjected to GAIIx sequencing, yielding 34.1 million reads, which mapped to 36 128 RefSeq entries (data not shown). The distribution of read number per transcript in the data is shown in Figure 2a. Among the 36 128 entries, only 200 (0.55%) accounted for ~20% of total reads, and 4.55% accounted for ~50% of reads. Thus, as expected, resequencing data for unselected cDNA consist mostly of reads corresponding to a limited number of highly-abundant transcripts.

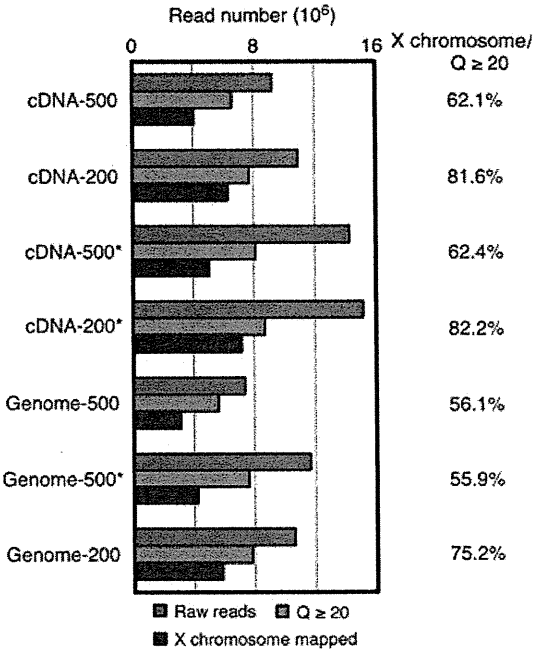

Fig. 1. Comparison of capture efficiency between cDNA and genomic DNA. Genomic DNA or cDNA of KCL-22-SR cells was fragmented to a mean size of 200 or 500 bp, and then subjected to purification with the SureSelect Human X Chromosome Demo kit, followed by GAIIx sequencing at a concentration of 4 or 8 pM (the latter indicated by an asterisk). Numbers of raw reads, reads with a *Q*-value of ≥20 ($Q \geq 20$), and reads mapped to the human X chromosome are shown for each experiment. Percentage of X chromosome-mapped reads among the reads with a *Q*-value of ≥20 is shown on the right.

doi: 10.1111/j.1349-7006.2011.02105.x

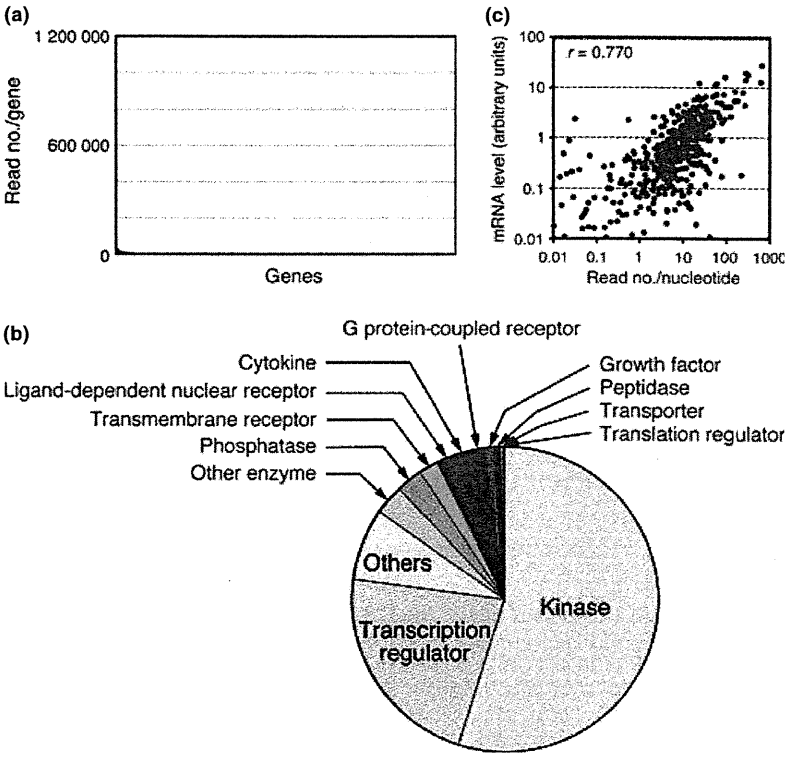

Fig. 2. Capture of a selected set of cDNA. (a) Read number for each gene was calculated from the sequencing data for the unselected cDNA of KCL-22-SR cells. Genes were sorted according to their read number. A small number of genes accounted for most of the sequence reads. (b) Functional annotation for the encoded proteins of our target cDNA (n = 913). (c) Read number per nucleotide for each captured cDNA in KCL-22-SR cells is compared with the expression intensity (arbitrary units) of the same cDNA examined with an HGU95Av2 microarray. Pearson's correlation coefficient (r) for the comparison is also demonstrated.

We therefore attempted to construct a custom SureSelect system to capture cDNA for cancer-related genes. For this purpose, we selected 913 genes that yielded 56 892 hybridization probes corresponding to ~3.77 Mbp of total capture capacity. The target genes encoded human protein kinases (all members in the human genome), transcription regulators, phosphatases, and other proteins (Fig. 2b; Table S1).

To compare the information provided by the sequence data from unselected and captured cDNA, we purified target cDNA from KCL-22-SR cells with the use of our custom SureSelect system, and determined their nucleotide sequences with GAIIx. A comparable amount of filter-passed reads (39.2 million) to that of unselected cDNA were thus obtained. We found that 88% of the captured cDNA were mapped to the target genes in our SureSelect system, while only 6.6% of the unselected cDNA were mapped to the 913 targets (data not shown). The read number obtained for each gene in the captured cDNA dataset is shown in Figure S2, with the distribution being markedly different from that obtained by sequencing of the unselected cDNA (Fig. 2a). As expected, the read number per nucleotide in each cDNA for the captured dataset was highly correlated to the expression intensity of the same gene quantified with the HGU95Av2 GeneChip expression array (Pearson's correlation coefficient = 0.770, $P < 2.2 \times 10^{-16}$) (Fig. 2c).

We further isolated target cDNA from other CML cell lines, including K562, KU812, MEG-01s, and NCO2, and the purified cDNA fragments were subjected to GAIIx sequencing. As in the case for KCL-22-SR, 86–88% of the obtained reads were successfully mapped to the target cDNA in each cell line (Table S2).

Screening of fusion cDNA. Our target set of 913 genes did not include *BCR*, but it did contain *ABL1*. Thus, if we were able to isolate sequence reads encompassing the fusion point of *BCR–ABL1*, cDNA-capture approaches for a given gene set would likely be able to detect gene fusions to unknown partners. In fact, we detected 45 sequence reads for KCL-22-SR cells that covered the *BCR–ABL1* fusion point (Fig. 3a). Likewise, the sequence datasets for K562, KU812, MEG-01s, and NCO2 cells contained 53, 8, 11, and 10 such fusion reads, respectively (data not shown). Furthermore, our sequence data faithfully recapitulated two variants of *BCR–ABL1* cDNA in these cell lines; a fusion variant between exon 13 of *BCR* and exon 2 of *ABL1* was detected in KCL-22-SR, MEG-01s, and NCO2 cells, whereas a fusion variant between exon 14 of *BCR* and exon 2 of *ABL1* was detected in K562 and KU812 cells.[14]

In addition to *BCR–ABL1*, we identified 72 independent candidates for fusion cDNA (including fusions to non-coding RNA) from the CML cell lines. Surprisingly, however, the screening of fusion genes among the unselected cDNA of KCL-22-SR with our rather non-stringent threshold (≥4 reads mapped to a candidate fusion point) failed to isolate *BCR–ABL1* cDNA. We could not even detect any fusion candidates (involving one of our target genes in either or both ends of fusion events) from this dataset, while a total of nine candidates (including *BCR–ABL1*) were isolated from the captured cDNA of the same cell line.

Our Bowtie mapping of both ends of each read to human mRNA or genome databases (Fig. S1) resulted in the detection of not only *BCR–ABL1* fusions, but also a large number of alternatively-spliced messages. From the captured cDNA of KCL-22-SR, for instance, we could detect 79 alternatively-spliced transcripts for 72 independent genes (data not shown). In contrast, from the unselected cDNA of the same cell line, only three independent, alternatively-spliced transcripts were identified among three genes within the 913 targets.

One such example of alternatively-spliced message was *MLL* (ensemble accession no.: ENST00000389506) in KU812, MEG-01s, and K562 cells. In addition to a set of reads that completely matched exon 3 of *MLL*, we obtained reads that lacked an internal 2193-bp sequence in exon 3 (Fig. 3b). Such in-frame truncation would be expected to generate an MLL protein lacking amino acids 276–1006 of the wild-type protein. To confirm the presence of such transcripts, we performed RT–PCR analysis with total RNA from KU812 cells, and PCR primers designed as in Figure 3b. The combination of the F-1 and R primers would be expected to yield both the wild-type (2536 bp) and truncated

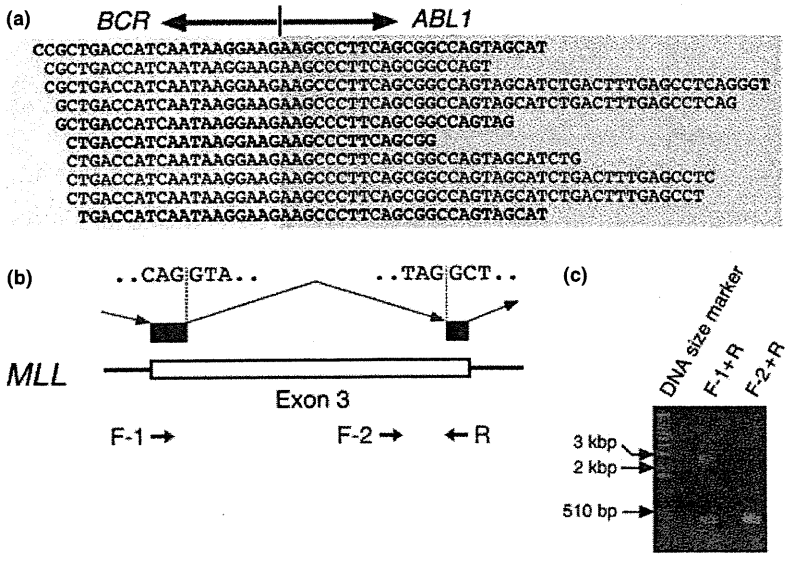

Fig. 3. Detection of gene fusions and alternative mRNA splicing in CML cells. (a) Our computational pipeline yielded 45 reads for KCL-22-SR cells that encompassed the fusion point of breakpoint cluster region (*BCR*)-Abelson murine leukemia viral oncogene homolog 1 (*ABL1*) cDNA, some of which are shown aligned. Reads in the sense or antisense strand are designated in black and blue letters, respectively, and the *BCR* and *ABL1* portions of the sequences are shaded differentially. (b) Some of the reads that mapped to exon 3 of mixed-lineage leukemia (*MLL*) skipped a 2193-bp region within this exon. Nucleotide sequences of the cryptic splicing sites are shown, as are the positions of PCR primers used to confirm the alternative splicing. (c) Gel electrophoresis of the RT–PCR products obtained with total RNA isolated from KU812 cells and with either the F-1 and R primer pair or the F-2 and R primer pair. A 1-kb ladder of DNA size markers was also included.

(343 bp) products, whereas that of the F-2 and R primers would yield only the wild-type product of 339 bp. Gel electrophoresis of the RT–PCR products confirmed the presence of the truncated mRNA (Fig. 3c). Given that the donor and acceptor sites for this alternative splicing harbor the consensus sequences for mRNA splicing (Fig. 3b), some CML cells likely make use of such cryptic splicing sites after *MLL* transcription.

Other variants. From the captured cDNA for KCL-22-SR, NCO2, MEG-01s, K562, and KU812 cells, we detected 156, 18, 28, 23, and 21 non-synonymous mutations among the 913 target genes, respectively. An analysis of the unselected cDNA from KCL-22-SR, however, identified only 19 mutations within the target genes, 16 of which were discovered in the captured cDNA as well. Comparison of the read sequences from the unselected KCL-22-SR cDNA to all RefSeq exonic sequences discovered a total of 597 non-synonymous mutations.

Furthermore, 19, eight, four, 11, and two indels were detected with the captured cDNA of KCL-22-SR, NCO2, MEG-01s, K562, and KU812, respectively. Most of the detected indels were only 1 bp in length, whereas the others were either 2 or 3 bp (Fig. S3). Detailed analysis of these nucleotide changes will be described elsewhere (Toshihide Ueno and Yoshihiro Yamashita, personal communication).

One of the most frequent genetic changes in the blast crisis of CML is point mutation or loss (or both) of *TP53*.[21] Indeed, our sequence data for this gene revealed non-synonymous point mutations in NCO2 and KU812 cells, a 1-bp insertion in K562 cells, a 1-bp deletion in KCL-22-SR cells, and a 3-bp deletion in MEG-01s cells (Fig. 4; Fig. S4; Table S3), all of which were confirmed by Sanger sequencing (data not shown). In NCO2 cells, for instance, 100% of *TP53* reads harbored a G-to-C substitution at nucleotide position 993 of *TP53* mRNA (GenBank accession no.: NM_000546), resulting in a glycine-to-arginine amino acid change (Fig. 4a). The data were also indicative of loss of heterozygosity for *TP53* in NCO2 cells. Similarly, 75% or 78% of *TP53* reads contained a C insertion or a CAC deletion in K562 (Fig. 4b) or MEG-01s (Fig. S4) cells, respectively.

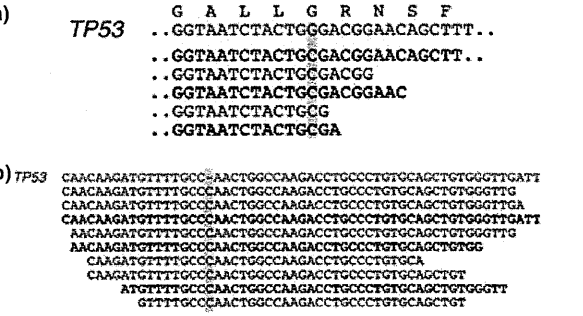

Fig. 4. Anomalies in *TP53* in CML cell lines. (a) Read sequences for NCO2 cells are shown aligned with the reference nucleotide and predicted amino acid sequences (red letters) for *TP53*, revealing a G-to-C substitution in all the reads. Sense or antisense strands are denoted in black and blue letters, respectively. (b) Alignment of the read sequences for K562 cells with the cDNA sequence of *TP53* as in (a), revealing a C insertion.

Discussion

We have shown that a cDNA-capture system, coupled with massively parallel sequencing, is a feasible and relatively simple approach to the simultaneous detection of point mutations, indels, and gene fusions in target cDNA. There are, however, both advantages and disadvantages of this technique compared with the conventional exon-capture system for genomic DNA.

The ability to detect gene fusions, in addition to other mutations with a single sequencing reaction, is one of the most important benefits of the cDNA-capture approach. Furthermore, the efficiency of exon capture with genomic DNA is dependent on the sequence context of each exon. The mean exon size for the human genome is only <200 bp, and the efficiency of exon purification is markedly affected by GC content and sequence complexity.[22] In contrast, even exons with a high GC content might be well isolated by the cDNA-capture system if adjacent exons have a normal GC content and are efficiently targeted by hybridization probes.

Levin *et al.*[20] conducted deep sequencing of captured cDNA for K562 cells, and identified five candidates for fusion genes in addition to *BCR–ABL1*. However, we could not detect any of the five candidates through our analysis with K562, probably because our 913 target genes did not contain those involved in the gene fusions in their report, other than nascent polypeptide-associated complex alpha subunit (*NACA*). While Levin *et al.* discovered primase, DNA, polypeptide 1 (*PRIM1*)–*NACA* fusion transcripts, the low expression level of *PRIM1–NACA* in K562 (only 2.5% of that of *BCR–ABL1* in their dataset)[20] might account for the failure in our analysis.

doi: 10.1111/j.1349-7006.2011.02105.x

However, for experiments based on capture of genomic DNA, sequencing a paired normal specimen allows the efficient subtraction of rare SNP not present in the current databases from the dataset of cancer tissue. This is not always the case, however, for the cDNA-capture approach, given that gene expression profiles differ markedly among samples (even among those obtained from the same individual). Genes with sequence alterations in the cancer specimen might not be expressed in a given normal specimen, and it is not possible to readily determine whether such alterations are germ-line polymorphisms, while algorithms to predict the effect on protein functions for a given amino acid change are currently available[23] and synonymous-to-non-synonymous ratio of nucleotide alterations for a given gene/dataset might provide clues as to how such changes are selected in tumor cells.[24]

In addition, the cDNA-capture system cannot obtain a sufficient number of reads for genes expressed at a low level, and the overall sensitivity of cDNA capture is dependent on the total read number provided by sequencers. We are able to run only two samples per flow cell of the GAIIx system, whereas up to eight samples can be run in a single flow cell for whole exome sequencing of human genomic DNA.

Despite such limitations, our study shows that cDNA capture is an efficient process, and extensive sequencing of such purified cDNA is a straightforward approach to interrogate the target cDNA for various genetic changes in a single platform. Large-scale resequencing of hundreds of cancer specimens might thus become within the scope of private laboratories with the adoption of the cDNA-capture approach.

Acknowledgments

This study was supported in part by grants for Research on Human Genome Tailor-Made and for Third-Term Comprehensive Control Research for Cancer from the Ministry of Health, Labor, and Welfare of Japan; Grants-in-Aid for Scientific Research (B) and for Young Scientists (A) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan; and by grants from the Japan Society for the Promotion of Science, Takeda Science Foundation, the Naito Foundation, Sankyo Foundation of Life Science, The Sagawa Foundation for Promotion of Cancer Research, the Yasuda Medical Foundation, the Mitsubishi Foundation, and Kobayashi Foundation for Cancer Research.

Disclosure Statement

K. Fukumura, M. Ando, M. Kawazu, Y.L. Choi and H. Mano belong to the Department of Medical Genomics, Graduate School of Medicine, University of Tokyo, which receives research funding from Illumina Inc.

References

1 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719–24.
2 Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011; **470**: 198–203.
3 Ledford H. Big science: the cancer genome challenge. *Nature* 2010; **464**: 972–4.
4 Pleasance ED, Cheetham RK, Stephens PJ *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010; **463**: 191–6.
5 Ley TJ, Mardis ER, Ding L *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008; **456**: 66–72.
6 Totoki Y, Tatsuno K, Yamamoto S *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet* 2011; **43**: 464–9.
7 Wei X, Walia V, Lin JC *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* 2011; **43**: 442–6.
8 Otto EA, Hurd TW, Airik R *et al.* Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet* 2010; **42**: 840–50.
9 Bilguvar K, Ozturk AK, Louvi A *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* 2010; **467**: 207–10.
10 Mitelman F. Recurrent chromosome aberrations in cancer. *Mutat Res* 2000; **462**: 247–53.
11 Soda M, Choi YL, Enomoto M *et al.* Identification of the transforming *EML4–ALK* fusion gene in non-small-cell lung cancer. *Nature* 2007; **448**: 561–6.
12 Tomlins SA, Rhodes DR, Perner S *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; **310**: 644–8.
13 Ohmine K, Nagai T, Tarumoto T *et al.* Analysis of gene expression profiles in an imatinib-resistant cell line, KCL22/SR. *Stem Cells* 2003; **21**: 315–21.
14 Drexler HG, MacLeod RA, Uphoff CC. Leukemia cell lines: *in vitro* models for the study of Philadelphia chromosome-positive leukemia. *Leuk Res* 1999; **23**: 207–15.
15 Choi YL, Tsukasaki K, O'Neill MC *et al.* A genomic analysis of adult T-cell leukemia. *Oncogene* 2007; **26**: 1245–55.
16 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**: R25.
17 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–10.
18 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**: 1754–60.
19 Fujita PA, Rhead B, Zweig AS *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2011; **39**: D876–82.
20 Levin JZ, Berger MF, Adiconis X *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol* 2009; **10**: R115.
21 Calabretta B, Perrotti D. The biology of CML blast crisis. *Blood* 2004; **103**: 4010–22.
22 Shen P, Wang W, Krishnakumar S *et al.* High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci USA* 2011; **108**: 6549–54.
23 Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–9.
24 Babenko VN, Basu MK, Kondrashov FA, Rogozin IB, Koonin EV. Signs of positive selection of somatic mutations in human cancers detected by EST sequence analysis. *BMC Cancer* 2006; **6**: 36.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Algorithm of the computational pipeline.

Fig. S2. Read number distribution of all poly(A)-RNA data.

Fig. S3. Numbers of 1-, 2-, or 3-bp indels for the entire dataset.

Fig. S4. A CAG-deletion in the *TP53* message in MEG-01s cells.

Table S1. Gene list for the custom cDNA-capture system.

Table S2. Purification of the target cDNA in CML cell lines.

Table S3. TP53 mutation status in CML cell lines.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Molecular Cell
Article

MCPIP1 Ribonuclease Antagonizes Dicer and Terminates MicroRNA Biogenesis through Precursor MicroRNA Degradation

Hiroshi I. Suzuki,[1] Mayu Arase,[1] Hironori Matsuyama,[1] Young Lim Choi,[2] Toshihide Ueno,[3] Hiroyuki Mano,[2,3] Koichi Sugimoto,[4] and Kohei Miyazono[1,*]
[1]Department of Molecular Pathology
[2]Department of Medical Genomics
Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3]Division of Functional Genomics, Jichi Medical University, 3311-1 Yakushiji, Shimotsukeshi, Tochigi 329-0498, Japan
[4]Division of Hematology, Department of Internal Medicine, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan
*Correspondence: miyazono@m.u-tokyo.ac.jp
DOI 10.1016/j.molcel.2011.09.012

SUMMARY

MicroRNAs (miRNAs) are versatile regulators of gene expression and undergo complex maturation processes. However, the mechanism(s) stabilizing or reducing these small RNAs remains poorly understood. Here we identify mammalian immune regulator MCPIP1 (Zc3h12a) ribonuclease as a broad suppressor of miRNA activity and biogenesis, which counteracts Dicer, a central ribonuclease in miRNA processing. MCPIP1 suppresses miRNA biosynthesis via cleavage of the terminal loops of precursor miRNAs (pre-miRNAs). MCPIP1 also carries a vertebrate-specific oligomerization domain important for pre-miRNA recognition, indicating its recent evolution. Furthermore, we observed potential antagonism between MCPIP1 and Dicer function in human cancer and found a regulatory role of MCPIP1 in the signaling axis comprising miR-155 and its target c-Maf. These results collectively suggest that the balance between processing and destroying ribonucleases modulates miRNA biogenesis and potentially affects pathological miRNA dysregulation. The presence of this abortive processing machinery and diversity of MCPIP1-related genes may imply a dynamic evolutional transition of the RNA silencing system.

INTRODUCTION

miRNAs constitute a large family of short endogenous, noncoding RNA molecules that regulate posttranscriptional gene silencing. miRNAs operate in many important aspects of diverse biological programs, including differentiation, proliferation, and cell death, and their deregulation is associated with an expanding number of disease processes such as cancer and cardiovascular diseases. The canonical miRNA biogenesis pathway is composed of several steps in mammalian cells (Davis and Hata, 2009; Siomi and Siomi, 2010). The primary transcripts of miRNA genes, primary miRNAs (pri-miRNAs), are transcribed and endonucleolytically cleaved to precursor miRNAs (pre-miRNAs) by the nuclear RNase III Drosha. The pre-miRNAs are exported into the cytoplasm by exportin-5 (XPO5) and further cleaved to a double-stranded miRNA duplex by another RNaseIII, Dicer. The mature miRNA strand is subsequently incorporated into the RNA-induced silencing complex (RISC). While it has been recently shown that multiple steps of miRNA biogenesis are regulated by several RNA binding factors such as Lin-28, hnRNP A1, and KSRP, and crosstalk with intracellular signaling networks (Suzuki and Miyazono, 2011; Suzuki et al., 2009), *cis*-regulatory elements and *trans*-acting factors responsible for stabilization or reduction of miRNAs have been largely unknown.

In mammalian systems, previous studies have established the close relationship between the immune system and the small RNA regulation. miRNA expression is dynamically regulated during the immune response (Moschos et al., 2007; O'Connell et al., 2010), and miRNA biogenesis also seems actively regulated in this setting; KSRP has been shown to regulate miRNA maturation in macrophage activation (Ruggiero et al., 2009).

In the present study, we identify the mammalian immune regulator, MCPIP1 (monocyte chemoattractant protein [MCP]-1-induced protein 1, also known as Zc3h12a [Zinc-finger CCCH-type containing 12A]), as a broad suppressor of the miRNA pathway. MCPIP1 with NYN nuclease domain preferentially cleaves the terminal loops of pre-miRNAs and counteracts Dicer, leading to the inhibition of de novo miRNA synthesis. Inverse correlation between MCPIP1 and Dicer function was seen in human lung cancer, in which low Dicer expression is associated with poor prognosis. These results suggest that the balance of productive and abortive ribonucleases modulates miRNA biogenesis and potentially affects pathological miRNA dysregulation. The presence of this abortive processing machinery and diversity of MCPIP1-related genes also implies a dynamic evolutional transition of the RNA silencing system.

RESULTS

Silencing of miRNA Activity by a CCCH Zinc-Finger Protein, MCPIP1

We screened regulator(s) of the miRNA pathway employing luciferase reporter constructs with the complementary sequence to several miRNAs in the 3′ UTR. Based on a close relationship between small RNA and immune system (O'Connell et al., 2010), we particularly focused on the immune response-associated genes harboring a potential RNA binding domain. In this process, we found that, although the ectopic expression of short fragments of pri-miRNAs caused a 5- to 10-fold reduction of luciferase activity in this assay, human MCPIP1 (also known as Zc3h12a), one of the CCCH-type zinc-finger proteins with RNA-binding potential (Liang et al., 2008a), remarkably attenuated the RNAi activity mediated by introduction of pri-miR-122 (Figure 1A). We also confirmed similar results using various mammalian cell lines such as HeLa, HepG2, and Cos-7 cells to exclude the possibility of cell-type-specific phenomena.

miRNA silencing activity of MCPIP1 was extended to other miRNAs, such as miR-21 and miR-135b (Figures 1B and 1C) and recapitulated in the experiments using miR-135b and its target, *adenomatous polyposis coli* (*APC*) 3′ UTR (Figure 1D) (Nagel et al., 2008). In addition, siRNA-mediated MCPIP1 knockdown consistently enhanced endogenous miRNA activity (Figure 1E) in HepG2 hepatocellular carcinoma cells, which express MCPIP1 at the basal level. These results suggested that MCPIP1 suppresses miRNA functions in gene silencing.

Inhibition of miRNA Biogenesis by MCPIP1

About 60 CCCH-type zinc-finger proteins have been identified so far in mouse and human (Liang et al., 2008a). MCPIP1 exerts the most potent suppressive activity to miRNAs among the MCPIP family (MCPIP1/2/3/4) and other tested CCCH-type zinc-finger proteins such as Cpsf4, Cpsf4l, Zfp36, Rc3h1, and Rc3h2 (Figure 1A), which prompted us to focus on the MCPIP1 function.

Because some CCCH proteins, including Zfp36 and Rc3h1, are involved in RNA metabolism (Liang et al., 2008a) and MCPIP1 appears to have an influence on broad miRNA activities, we examined a potential impact of MCPIP1 on the intracellular fates of miRNAs and their precursors. Quantitative RT-PCR analysis showed that MCPIP1 remarkably suppressed mature miRNA production from ectopically expressed pri-miRNAs of various miRNAs, such as miR-135b, -146a, -21, -155, -143, and -145 (Figures 1F, 1G, and S1A–S1D). Interestingly, MCPIP1 decreased the expression levels of precursor forms (pre-miRNAs) of these miRNAs, but the pri-miRNAs showed no significant change (Figures 1F, 1G, and S1A–S1D). Northern blot analyses confirmed these observations (Figures 1H, 1I, and S1E–S1H). MCPIP1 overexpression also decreased the mature miRNA levels of endogenously expressed miRNAs by 20%–30%.

Two major steps of mammalian miRNA biogenesis are executed by nuclear Drosha and cytoplasmic Dicer. We then asked in which cellular compartment MCPIP1 localizes and functions. Immunocytochemical analysis showed that MCPIP1 localized mainly in the cytoplasm as dot-like structures (Figure 2A). GFP-tagged MCPIP1 also partly colocalized with HA-tagged GW182 or Dcp1a, suggesting that MCPIP1 localizes in GW bodies or P bodies (Figure 2B). Colocalization frequency of GFP-MCPIP1 with HA-GW182 was higher than that with HA-Dcp1a, while GFP-MCPIP1 was often adjacent to Dcp1a-positive foci. In addition, MCPIP1 strongly reduced miRNA expression and activity under pri-miRNA overexpression (Figure 1), but showed no significant effect on miRNA activity upon miRNA duplex introduction (Figures 2C and 2D) or the stability of introduced miRNA duplexes (Figures 2E and 2F).

Next, we confirmed that MCPIP1 knockdown elevated the intrinsic levels of several mature miRNAs, such as miR-21, -26a, -107, -182, -146a, -17-5p, and -135b in HepG2 cells, without concomitant changes of the corresponding pri-miRNAs (Figure 2G) and enhanced the miRNA maturation efficacy from ectopically expressed pri-miRNAs (Figures S2A–S2D). Microarray analysis of miRNA expression profile demonstrated that wide proportion of miRNAs tended to be upregulated by MCPIP1 suppression (Figure 2H). These results thus proposed that MCPIP1 acts on pre-miRNAs in the cytoplasm to inhibit de novo maturation of a broad range of miRNA species, without deteriorating the major molecular components of the miRNA pathway (Figure S2E). Immunoblot analysis also confirmed endogenous expression of Dicer and MCPIP1 in the cell lines used in this study (Figure S2F).

Cleavage of Pre-miRNAs by MCPIP1 Ribonuclease

The N terminus of MCPIP1 (residues 133–300) is well conserved among other related proteins including MCPIP2, MCPIP3, and MCPIP4 and has been previously referred to as *N*edd4-BP1, bacterial *Yac*P *N*uclease (NYN) domain (Anantharaman and Aravind, 2006) (Figures 3A and S3A). The NYN domain shares a similar protein fold with two characterized nuclease domains, the PilT N-terminal (PIN) domain and FLAP nuclease domain, and is thus predicted to exert nuclease activity (Figures 3A and S3A–S3B) (Anantharaman and Aravind, 2006). In fact, it has been recently shown that MCPIP1/Zc3h12a is involved in mRNA decay of several genes as a ribonuclease (Matsushita et al., 2009).

We therefore examined whether pre-miRNAs can be cleaved by MCPIP1 through in vitro cleavage assay using the radiolabeled pre-miRNAs and FLAG-tagged MCPIP1. When the radiolabeled pre-miRNAs were incubated with immunoprecipitated FLAG-tagged MCPIP1 in vitro, the resultant products migrated as a broad and fuzzy band around 25–35 nt in length on electrophoresis, while FLAG-Dicer generated ~22 nt mature miRNA duplex (Figures 3B and 3C), demonstrating that MCPIP1 targets pre-miRNAs as a ribonuclease. We observed that both Dicer and MCPIP1 cleaved pre-miRNAs in a time- and dose-dependent manner (Figures S3C and S3D). The NYN domain of MCPIP1 has the canonical four conserved acidic residues (Asp 141, Asp 195, Asp 226, Asp 244) by analogy with other PIN domain RNases (Figure S3A) (Anantharaman and Aravind, 2006). These acidic residues have been reported to be crucial for metal ion binding and nuclease activity (Figure S3B). In accordance with these findings, mutant MCPIP1 with D141N mutation in the NYN domain, but not C306R mutation in the CCCH motif, failed to cleave pre-miRNAs in vitro (Figure 3D),

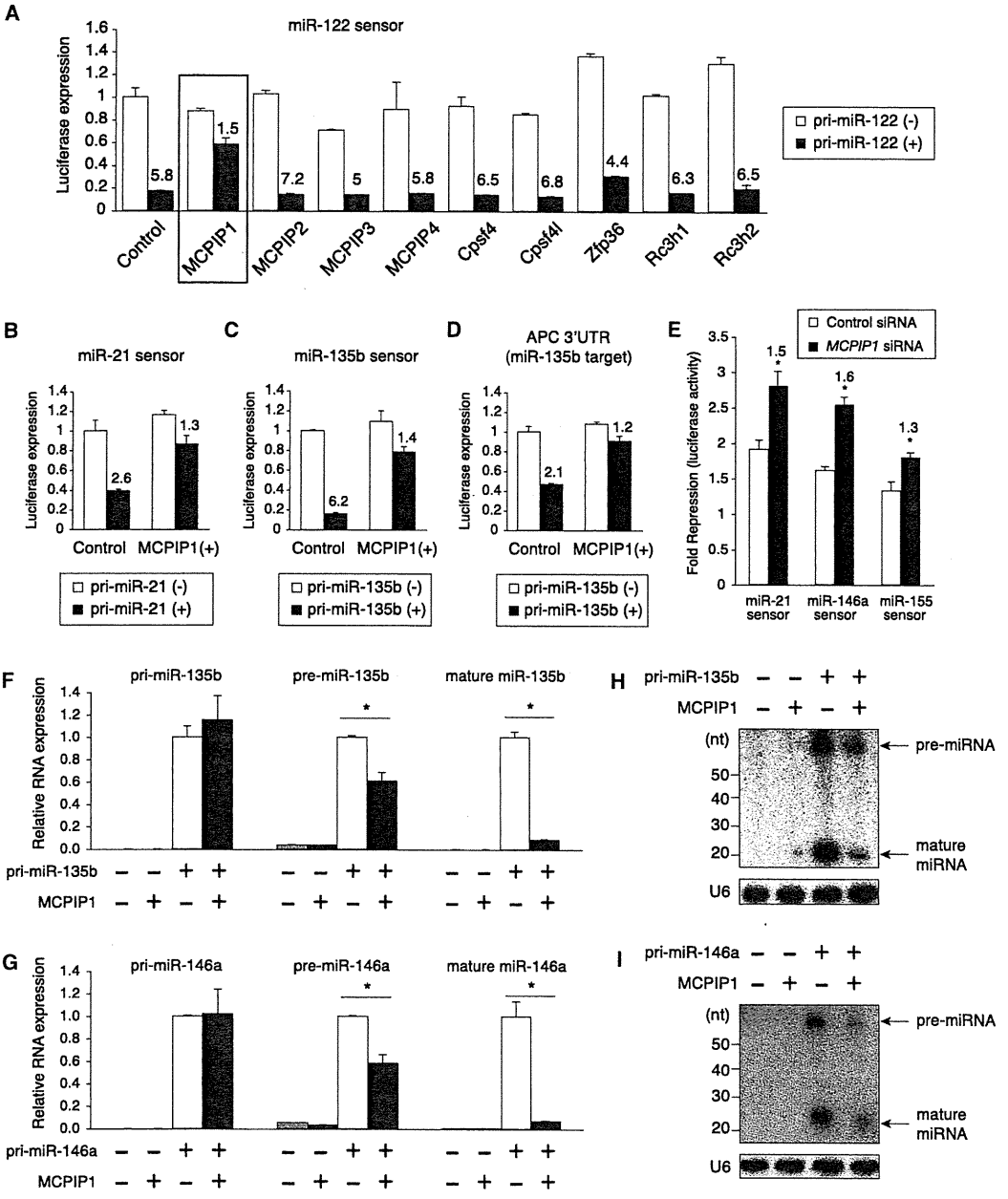

Figure 1. Suppression of miRNA Activity and Biogenesis by a CCCH-Type Zinc-Finger Gene, MCPIP1

(A) Effects of CCCH zinc-finger genes on miRNA activity. HEK293T cells were cotransfected with miR-122 sensor vector, pri-miR-122 expression vector, and the expression vectors of indicated genes, and applied to luciferase reporter assay at 48 hr after transfection. Numbers indicate the ratio of fold repression.

(B–D) Suppression of a range of miRNA activities by MCPIP1. miR-21 sensor (B), miR-135b sensor (C), or APC 3′ UTR luciferase vector (D), and the corresponding pri-miRNA expression vectors were ectopically expressed with/without MCPIP1 plasmids in HEK293T cells and subjected to luciferase reporter assay as shown in (A). Numbers indicate the ratio of fold repression.

(E) Enhancement of endogenous miRNA activity by MCPIP1 knockdown. HepG2 cells were transfected with indicated miRNA sensor vector and control siRNA or siRNA for *MCPIP1*. miRNA activities were measured by the ratio of the luciferase expression of miRNA sensor vector to that of control sensor vector (*$p < 0.05$, compared to Control siRNA; $n = 3$). Numbers indicate the ratio between control siRNA and *MCPIP1* siRNA.

(F–I) Inhibition of de novo miRNA biosynthesis by MCPIP1 overexpression. Quantitative RT-PCR (qRT-PCR) (F and G) and northern blot (H and I) analyses of RNA samples from HEK293T cells transfected with pri-miR-135b (F and H), -146a (G, I), and MCPIP1 (*$p < 0.05$, compared to empty vector; $n = 3$). See also Figure S1. Error bars represent SEM.

indicating the indispensable role of NYN domain for MCPIP1 ribonuclease function. Recombinant MCPIP1 proteins furthermore cleaved pre-miRNAs in a NYN domain-dependent manner in vitro (Figure 3E). On the other hand, a mutation in the CCCH motif (C306R), which had no significant effect on in vitro cleavage activity (Figures 3D and 3E), attenuated the miRNA suppressor activity of MCPIP1 in vivo (Figures 3F and 3G). This result suggests that the CCCH motif is additionally important for the in vivo function of MCPIP1, as investigated in Figure 7 (see below).

Functional Antagonism between MCPIP1 Ribonuclease and Dicer and Its Implication in Human Cancer

We then investigated whether MCPIP1 and its ribonuclease function can compete with Dicer function by in vitro competition analysis. As shown in Figures 4A and 4B, MCPIP1 antagonized Dicer-mediated pre-miRNA processing in vitro, depending on the intact NYN domain but not the CCCH motif, in a dose-dependent manner. Sequential in vitro cleavage analysis showed that MCPIP1-cleaved fragments were inefficient substrates for Dicer processing (Figure S4A). These results thus suggested that MCPIP1-mediated pre-miRNA cleavage could interfere with Dicer activity. In addition, absence of the degraded products in northern blot analyses suggested that MCPIP1-cleaved fragments may be unstable and subjected to a rapid subsequent degradation in vivo (Figure 1G, 1I, and S1E–S1H).

While individual miRNAs can behave as tumor suppressors or oncogenes (Suzuki and Miyazono, 2010), global miRNA downregulation has been shown to be a general trait of human cancers in several reports (Kumar et al., 2007; Lu et al., 2005; Ozen et al., 2008). Several mechanisms including downregulation of miRNA processing factors, such as Dicer and Drosha, and mutations of TRBP and XPO5 have been reported (Karube et al., 2005; Martello et al., 2010; Melo et al., 2010). Recent reports also revealed that certain oncogenesis-related molecules, such as p53, Smad, and estrogen receptor, are coupled with miRNA biogenesis (Suzuki and Miyazono, 2011). As evidence demonstrating the clinical relevance of these phenomena, reduced expression of Dicer is associated with poor survival in lung, ovarian, and breast cancer patients (Karube et al., 2005; Martello et al., 2010; Merritt et al., 2008).

Based on these observations, we investigated potential functional crosstalk between MCPIP1 and Dicer using several public gene expression datasets containing molecular and associated clinical details. We extracted gene sets associated negatively or positively with Dicer expression status ("Dicer High/Low Downregulated or Upregulated genes") from several human cancer cohorts, and performed gene set enrichment analysis (GSEA) (Subramanian et al., 2005) to examine whether these gene sets are overexpressed or underexpressed along the expression status of MCPIP1. During these trials, we found a potential antagonistic relationship between MCPIP1 and Dicer in human lung cancer. In the dataset of lung adenocarcinoma patients (Bild et al., 2006), GSEA demonstrated that "Dicer High/Low Downregulated genes" and "Dicer High/Low Upregulated genes" are enriched in patients with high MCPIP1 expression ("MCPIP1 High Case") and low MCPIP1 expression ("MCPIP1 Low Case"), respectively (Figure 4C). GSEA using the sets of potential miRNA target genes also demonstrated that a large proportion of miRNA target gene sets are upregulated in the MCPIP1-High group against the MCPIP1-Low group, while the opposite phenomenon was seen in the comparison between Dicer-High and -Low group (Figure 4D). Furthermore, we observed that high MCPIP1 levels are associated with poor survival in lung adenocarcinoma patients (Figure 4E), in an opposite manner to the association between low Dicer levels and poor prognosis (Karube et al., 2005; Merritt et al., 2008). Similar results were also obtained in another dataset of lung adenocarcinoma patients (Shedden et al., 2008) and a dataset of lung squamous cell carcinoma patients (Larsen et al., 2007) (Figures S4B–S4E). Together with in vitro antagonism, these findings demonstrate an inverse correlation between MCPIP1 expression and Dicer function in human cancer and suggest that MCPIP1 could drive opposite effects against Dicer on the transcriptome.

Modulation of miR-155/c-Maf axis by MCPIP1

We next investigated the impact of MCPIP1 on the function of endogenous miRNAs. MCPIP1 expression dynamically changes during the inflammatory response (Liang et al., 2008a; Liang et al., 2008b; Matsushita et al., 2009), and the importance of MCPIP1 in the immune system has recently emerged from a manifestation of profound autoimmune phenotype in mice lacking MCPIP1 (Matsushita et al., 2009). MCPIP1 has been shown to promote the mRNA decay of several inflammatory genes such as IL-6 as an endoRNase (Matsushita et al., 2009). The severity of autoimmune phenotype in *MCPIP1*$^{-/-}$ mice suggests the existence of additional targets of MCPIP1 other than IL-6 (Matsushita et al., 2009), and it might thus be partly explained by miRNA deregulation. In support of this notion, it has been shown that transgenic mice overexpressing several miRNAs, such as miR-17-92 and miR-155, exhibit autoimmune and/or lymphoproliferative diseases (Costinean et al., 2009; Xiao et al., 2008). Interestingly, increased and decreased expression of Th2 cytokine, IL-4, has been observed in miR-155$^{-/-}$ and *MCPIP1*$^{-/-}$ T cells, respectively (Matsushita et al., 2009; Rodriguez et al., 2007).

We then examined the potential involvement of MCPIP1 in regulation of miR-155 function in Jurkat T cells. In accordance with our previous results, MCPIP1 knockdown, in fact, caused miR-155 upregulation in Jurkat T cells (Figure 5A). MCPIP1 knockdown further suppressed the expression of the miR-155 target c-Maf and the induction of its transcriptional target IL-4 by phorbol myristate acetate (PMA) and ionomycin in a miR-155-dependent fashion (Rodriguez et al., 2007) (Figures 5B–5D), while IL-6 showed no significant effect on this axis (Figure S5A). These results are consistent with increased and decreased expression of IL-4 in miR-155$^{-/-}$ and *MCPIP1*$^{-/-}$ T cells (Matsushita et al., 2009; Rodriguez et al., 2007), reinforcing the physiological relevance of MCPIP1-mediated miRNA regulation. In addition, we observed that MCPIP1 is induced by lipopolysaccharide (LPS) and involved in the restriction of miR-155 upregulation under innate immune response and miR-16 downregulation during macrophage differentiation, while Dicer expression is not affected by LPS (Li et al., 2010) (Figures 5E, S5B, and S5C). These findings underscore active MCPIP1 involvement in miRNA biogenesis, together with the inverse

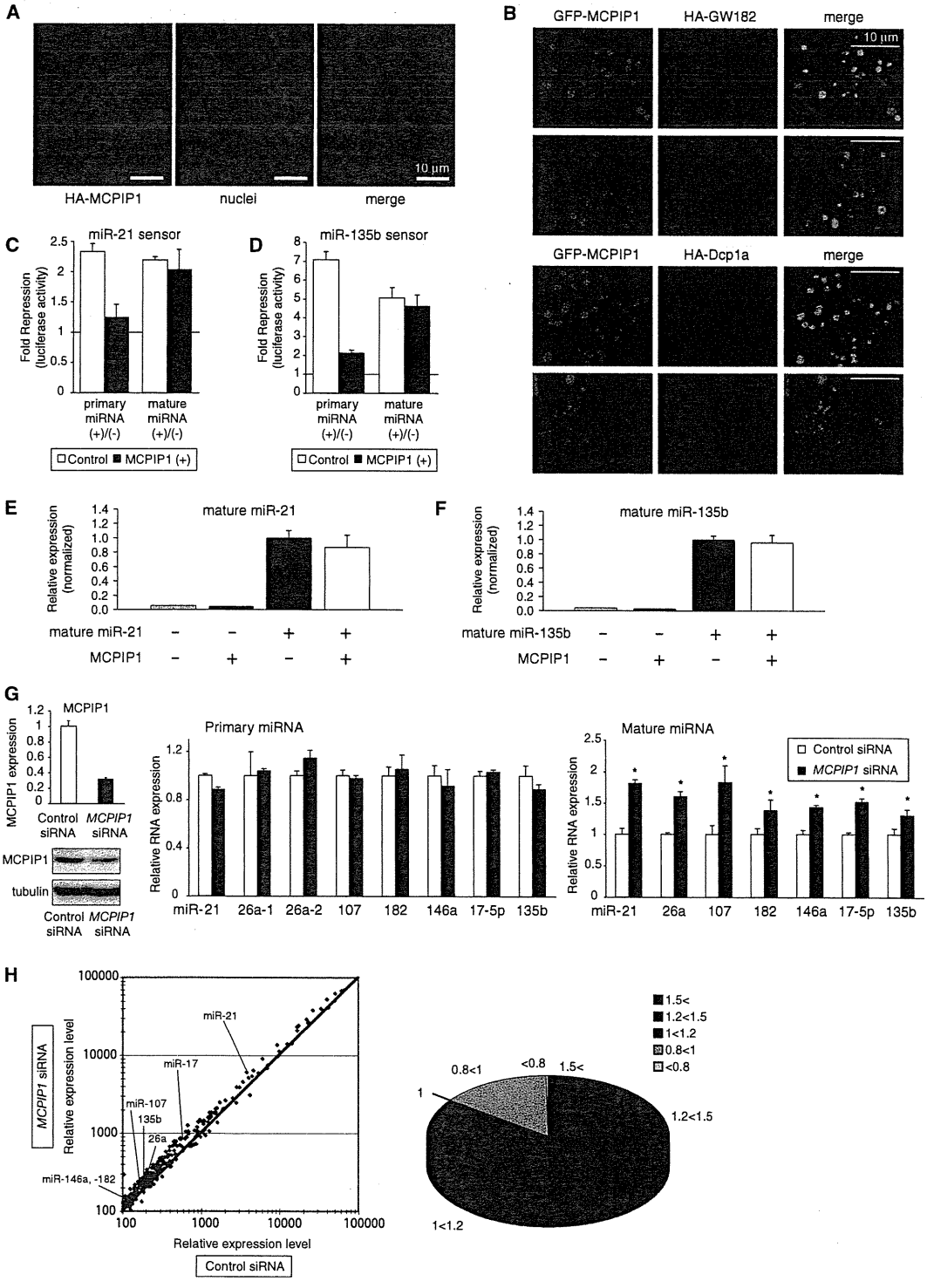

Figure 2. Cytoplasmic Localization of MCPIP1 and Enhancement of miRNA Maturation by MCPIP1 Depletion

(A) Immunocytochemical analysis of overexpressed HA-MCPIP1. Scale bar: 10 μm.

(B) Localization of GFP-MCPIP1, HA-GW182 and HA-Dcp1a.

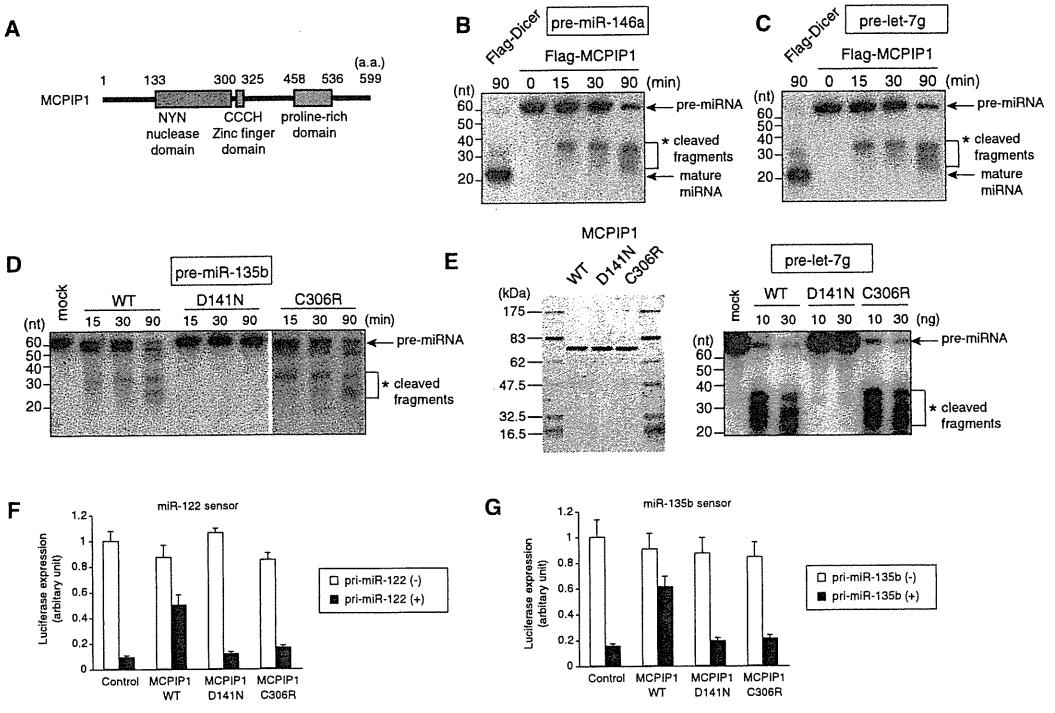

Figure 3. Cleavage of Pre-miRNAs by MCPIP1 Ribonuclease

(A) Domain organization of MCPIP1. NYN, *N*edd4-BP1, bacterial *Yac*P *N*uclease domain; CCCH, CCCH zinc-finger domain; PRD, proline-rich domain. See also Figure S3.

(B and C) In vitro cleavage of pre-miRNAs by MCPIP1. Radiolabeled pre-miR-146a (B) or pre-let-7g (C) were incubated with immunoprecipitated FLAG-Dicer or FLAG-MCPIP1 from HEK293T cells for indicated periods. The RNA was then extracted and analyzed by denaturing PAGE and autoradiography. 'nt' denotes nucleotides.

(D and E) Requirement of an intact NYN domain for in vitro RNase activity of MCPIP1. Radiolabeled pre-miRNAs were incubated with immunoprecipitated FLAG-MCPIP1 proteins (wild-type [WT], D141N, and C306R) (D) or recombinant MCPIP1 proteins (E).

(F and G) Effects of mutations in the NYN domain and CCCH domain on in vivo miRNA silencing activities. miR-122 sensor (F) or miR-135b sensor (G) and the corresponding pri-miRNA were coexpressed with/without MCPIP1 into HEK293T cells, and subjected to luciferase reporter assay (n = 3). Error bars represent SEM.

correlation between MCPIP1 and Dicer function in human cancer (Figure 4).

Targeting of the Terminal Loops of Pre-miRNAs by MCPIP1 Ribonuclease

The results shown in Figures 3 and 4 demonstrate that MCPIP1 cleaves pre-miRNAs and antagonizes Dicer function. We further analyzed the mode of MCPIP1 action on pre-miRNAs by direct cloning of pre-miRNA fragments cleaved by MCPIP1. Various 5′ and 3′ miRNA strands accompanied with additional bases from the terminal loop were recovered as the cleaved fragments of pre-miR-146a, pre-let-7g, pre-miR-135b, pre-miR-143 and pre-miR-16-1 (Figures 6A, 6B, and S6A–S6C), indicating that the primary cleavage sites of MCPIP1 were dispersed in the terminal loop of pre-miRNAs. A comparison with the predicted secondary structures of pre-miRNAs (Zuker, 2003) shows that MCPIP1 preferentially cleaves the unpaired regions around the terminal loops as an endoribonuclease (Figure 6C). Furthermore, addition of Lin-28b, an RNA-binding protein interacting with the terminal-loop of pre-let-7 (Heo et al., 2008), abolished the

(C and D) Differential effects of MCPIP1 on miRNA activities under introduction of pri-miRNAs and miRNA duplexes. miRNA sensor vectors (C: miR-21, D: miR-135b) and the corresponding pri-miRNA expression vectors or synthetic miRNA duplexes were cotransfected with/without MCPIP1 plasmids into HEK293T cells, and applied to luciferase reporter assay.

(E and F) Effect of MCPIP1 on the stability of introduced miRNAs duplexes. MCPIP1 plasmid was cotransfected with synthetic miRNA duplexes (10 nM) (E: miR-21, F: miR-135b) in HEK293T cells. At 48 hr posttransfection, miRNA expression levels were analyzed with qRT-PCR.

(G) Upregulation of mature miRNA levels, but not pri-miRNA levels, of several miRNAs by MCPIP1 depletion. The levels of MCPIP1, pri-miRNAs, and mature miRNAs were compared by qRT-PCR analyses in HepG2 cells transfected with control siRNA or siRNA for *MCPIP1* (*$p < 0.05$, compared to Control siRNA; n = 3).

(H) Global effects of MCPIP1 knockdown on endogenous miRNA expression, analyzed by miRNA microarray analysis in HepG2 cells. The proportion of miRNAs at different fold change levels is shown in the right panel. See also Figure S2. Error bars represent SEM.