Published online 29 June 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 20* ***6985–6996***
doi:10.1093/nar/gkq579

Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer

Juliann Chmielecki[1], Martin Peifer[2], Peilin Jia[3], Nicholas D. Socci[4], Katherine Hutchinson[5], Agnes Viale[6], Zhongming Zhao[3], Roman K. Thomas[2,7,8] and William Pao[5,*]

[1]Weill Graduate School of Medical Sciences, Cornell University, New York, NY 10021, USA, [2]Max Planck Institute for Neurological Research with Klaus Joachim Zülch Laboratories of the Max Planck Society and the Medical Faculty, University of Cologne, Cologne, Germany, [3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, [4]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, [5]Department of Medicine, Vanderbilt-Ingram Cancer Center, Nashville, TN 37232, [6]Genomics Core Laboratory, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA, [7]Department I of Internal Medicine and Center of Integrated Oncology Köln – Bonn, University of Cologne and [8]Chemical Genomics Center of the Max Planck Society, Dortmund, Germany

Received March 8, 2010; Revised May 18, 2010; Accepted June 9, 2010

ABSTRACT

Tyrosine kinase (TK) fusions are attractive drug targets in cancers. However, rapid identification of these lesions has been hampered by experimental limitations. Our *in silico* analysis of known cancer-derived TK fusions revealed that most breakpoints occur within a defined region upstream of a conserved GXGXXG kinase motif. We therefore designed a novel DNA-based targeted sequencing approach to screen systematically for fusions within the 90 human TKs; it should detect 92% of known TK fusions. We deliberately paired 'in-solution' DNA capture with 454 sequencing to minimize starting material requirements, take advantage of long sequence reads, and facilitate mapping of fusions. To validate this platform, we analyzed genomic DNA from thyroid cancer cells (TPC-1) and leukemia cells (KG-1) with fusions known only at the mRNA level. We readily identified for the first time the genomic fusion sequences of *CCDC6-RET* in TPC-1 cells and *FGFR1OP2-FGFR1* in KG-1 cells. These data demonstrate the feasibility of this approach to identify TK fusions across multiple human cancers in a high-throughput, unbiased manner. This method is distinct from other similar efforts, because it focuses specifically on targets with therapeutic potential, uses only 1.5 μg of DNA, and circumvents the need for complex computational sequence analysis.

INTRODUCTION

Tyrosine kinases (TKs) are tightly regulated signaling enzymes that control multiple cellular processes. When TK signaling becomes deregulated due to mutations or rearrangements involving the kinase domain, the resultant sustained activity can lead to cancer. Because constitutive kinase activity can also be required for tumor maintenance, aberrant TKs serve as attractive therapeutic targets (1). The best example of this concept involves the BCR-ABL (BCR - breakpoint cluster region gene; ABL - Abelson murine leukemia viral oncogene homolog 1 gene) TK fusion protein in patients with chronic myelogenous leukemia (CML) (2). As a consequence of the fusion, the ABL kinase is constitutively activated (3,4). CML cells are dependent upon signaling from BCR-ABL and die upon treatment with the kinase inhibitor, imatinib (Gleevec). Clinically, the drug has revolutionized treatment of the disease.

Thus far, only a limited number of TK fusions have been found in cancers. The majority of TK fusions have been identified in hematopoietic malignancies as opposed to solid tumors, because the latter are difficult to karyotype, harbor multiple genomic aberrations, and are often

*To whom correspondence should be addressed. Tel: +1 615 936 3831; Fax: +1 615 343 7602; Email: william.pao@vanderbilt.edu

clonally heterogenous (5). Nevertheless, fusion proteins do exist in epithelial cancers. TK fusions involving *RET*, *NTRK1, -3*, and the serine/threonine kinase, *BRAF*, have been found in radiation-induced thyroid cancers (6–8). Recently, a subset of non-small cell lung cancers (NSCLCs) was discovered to harbor gain-of-function *ALK* and *ROS* fusions (9–11). Importantly, only 2 years later, an ALK inhibitor has already demonstrated promising activity in patients with *EML4-ALK*-fusion-positive lung tumors (12). However, the identification of these fusions involved highly laborious techniques not amenable to rapid screening.

The recent application of high throughput next generation sequencing technologies to whole genomes and whole transcriptomes has facilitated the discovery of multiple translocation events in cancer cells (13–19). These efforts have been enhanced further by re-sequencing of selectively captured regions of interest (20). Capture technologies utilize complementary oligonucleotide 'baits' either immobilized on a chip (solid capture) or in-solution (liquid capture) (21–24). Traditionally, chip based arrays have been coupled with long read 454 sequencing (21,23,24), while 'in-solution' capture was has been paired with short read sequencing (i.e. Illumina GA and AB SOLiD) (22). However, these methods require large amounts of starting material and generate vast numbers of sequences. Furthermore, these methods have not focused specifically on identifying TK fusions.

Here, we present the development of a 'rationally designed' DNA capture strategy that overcomes many of the limitations of current fusion discovery platforms. Our strategy focuses specifically on the discovery of novel TK fusions because of their clinical significance and inherent druggability. We hypothesized that tumors contain as yet unidentified TK fusions whose discovery has been hindered by experimental limitations. In contrast to other targeted approaches, our design was based upon unique conserved genomic properties of existing TK fusions. Importantly, our capture-sequence approach is feasible, rapid, and requires minimal amounts of starting tumor cell DNA, making it amenable for high-throughput screens to identify systematically TK fusions in any cancer.

MATERIALS AND METHODS

Genomic coordinate mapping

The nucleotides encoding the GXGXXG motif for all 90 TK kinases (25) and four serine/threonine kinases (BRAF, AKT-1, -2, -3) were mapped using MapBack (hg18), a locally created database that maps protein residues to corresponding genomic coordinates (http://cbio.mskcc.org/Public/products/human_mapped/Mapback; A. Lash and C. Byrne, manuscript in preparation). From these regions, genomic coordinates for the three preceding introns and the two preceding exons were mapped using ENSEMBL (hg18; http://www.ensembl.org). All exons/introns were labeled according to ENSEMBL numbering. In cases where the GXGXXG motif was encoded by more than one exon, only the first exon was included in capture, as breaks are likely to occur upstream of this motif. For *ABL1*, capture included intron 1-2 to exon 4 (GXGXXG motif), and for *ROS1*, capture included intron 31-32 to exon 36 (GXGXXG motif). An extra exon and intron upstream of the target region were also included for *ABL2*. Coordinates were submitted to Agilent Technologies (Santa Clara, CA, USA) for custom bait design. Repetitive elements, as identified by the UCSC genome browser (26) (http://www.genome.ucsc.edu), were excluded from bait design.

Samples and cell lines

The human cell lines KG-1 and TPC-1 have been characterized previously (27,28). KG-1 cells and TPC-1 cells were kindly provided by R. Levine and J. Fagin (MSKCC), respectively. KG-1 cells were cultured in RPMI media (American Type Tissue Collection, ATCC) supplemented with 10% fetal bovine serum (Gemini Bio Products) and pen-strep solution (Gemini Bio Products; final concentration 100 U/ml penicillin, 100 μg/ml streptomycin). TPC-1 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with glucose (4.5 g/l), 5% fetal bovine serum, pen-strep solution (final concentration 100 U/ml penicillin, 100 μg/ml streptomycin) and 2 mM glutamine. All cells were grown in a humidified incubator with 5% CO_2 at 37°C.

DNA capture and 454 sequencing

Genomic DNA from all samples was extracted using standard phenol extraction protocols. 1.5 μg was sheared with a Roche Nebulizer to 300–500 bp fragments. Fragment size was confirmed on a BioAnalyser, DNA 7500 assay (Agilent). 454 adaptors (Roche) were ligated according to the manufacturer's instructions. Ligated products were size selected on an agarose gel, purified using the AMpure kit (Agencourt), and PCR amplified for 15 cycles. The PCR products were purified with a mini-elute PCR purification kit (QIAGEN). Capture was performed at Agilent Technologies (Santa Clara, CA, USA) using their SureSelect Target Enrichment System. Subsequently, 2–4 μl of eluted single stranded DNA was used for emulsion PCR with emPCRkit I (Roche). Approximately 300 000 beads/sample/run were used for sequencing on a 454 FLX sequencer (Roche).

Computational analysis

Two independent BLAT-based methods were used for 454 sequence analysis. The first method aligned 454 reads to a custom library of known genes derived from BLAT (29). Only TK-containing sequences (and *BRAF* and *AKT-1,-2,-3*-containing sequences) were considered for further evaluation (Supplementary Figure S2). To minimize recovery of repetitive elements and low complexity sequences, candidate fusion sequences then had to meet the following criteria: (i) the entire length of the sequence had to map to $\leq$3 targets within the genome; (ii) the sequence overlap between the targets could not exceed 5 bps and (iii) any sequence gaps between the targets could not exceed 5 bps.

The second method mapped the 454 reads to the entire human genome using BLAT and considered only those

reads that hit one kinase target and one other region in order along the query sequence. The candidate sequences in the high-scoring segment pairs (HSPs) of the BLAT output then had to meet the following criteria: (i) the distance between the targets could not exceed 5 bps; (ii) no more than one mismatch was allowed, and gaps were $\leq$2 bps; (iii) alignment to the two targets had to account for $\geq$95% of the query sequence; and (iv) no more than 5 bases could be removed from either end of the sequence.

Additionally, sequences from candidate fusions identified by either method had to be recovered from at least two independent 454 sequences. A combined list of TK fusion candidates was generated from sequences that met these criteria. These sequences were then aligned back to the entire genome using BLAT, and those with $\geq$98% identity to a single repetitive element were eliminated. The remaining candidate fusion sequences were validated by PCR using fusion point-spanning primers and appropriate genomic DNA. Upon PCR confirmation, the genomic fusion sequence was queried against all the 454 reads from the same sample to find additional fusion sequences that may have been missed by automated methods. In particular, this step allowed us to recover fragments where the length of sequence from one fusion partner was <25 bps, the minimum length required for BLAT searches.

Mapping of 454 sequences and enrichment calculation

454 sequences were mapped initially to the human genome (hg18) using BLAT (29) with parameters that would allow for partial mapping of reads, as would be the case if fusions were present. Average read length was determined from all reads so as not to exclude any fusion sequences mapping to multiple targets.

The capture efficiency was calculated from the fraction of mapped sequences that overlapped with bait target regions. To put this number in proper context, and to account for the small portion of the genome used for capture, we computed a 'normalized' enrichment factor defined as:

$$\text{Enrichment} = \frac{\text{No. of reads hitting targets/total number of reads}}{\text{total size of targets in base pairs/total size of genome (in bp)}}$$

This equation corrects for cases where a small genomic region was targeted with baits.

The mapping outputs from each sample were then converted to and loaded as .bed files into the UCSC genome browser (see Supplementary Table S3 for sample barcodes) in addition to the genomic coordinates for each of the custom baits provided by Agilent Technologies. We also loaded the target sequences used for bait design as reference. Genomic coordinates for the individual baits are available upon request. Fusion sequences were mapped to the region with highest homology. The mapping of the sequences and coordinates across the entire genome is publicly available at: http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&hgt.customText=http://cbio.mskcc.org/~socci/JC/groupA.bed.gz.

Fusion point confirmation

PCR amplification of the candidate fusion genomic breakpoints and wild-type sequences was performed with M13-tagged primers (Supplementary Table S3) using HotStarTaq Master Mix (QIAGEN) and standard cycling conditions (95°C for 15 m, 35 cycles of 94°C for 30 s, 60°C for 30 s, 72°C 1 min and final extension at 72°C for 10 m). Normal male DNA (Promega) was used as a negative control. PCR products were separated by agarose gel electrophoresis. Excess primers and dNTPs were removed with ExoSAP-IT (USB Corporation), as per the manufacturer's instructions, prior to direct dideoxynucleotide sequencing at the Vanderbilt DNA Sequencing Facility.

5′-Rapid amplification of cDNA ends

Total RNA was extracted from KG-1 cells with TriZol reagent (Invitrogen). Extracted RNA was treated with DNase I (Sigma-Aldrich) and precipitated using standard protocols. Rapid amplification of cDNA ends (RACE) was done with a 5′-RACE system (Invitrogen), as per the manufacturer's instructions. Five micrograms of RNA was used for the initial cDNA reaction with FGFR1.GSP1 (ACGGTTGGGTTTGTCCTTGT). Following dC-tailing, the cDNA was amplified with FGFR1.GSP2 (TCAGAGACCCCTGCTAGCAT) and the provided abridged anchor primer. PCR products were confirmed by agarose gel electrophoresis, and cloned into pCR.II-TOPO using a TOPO-TA cloning kit (Invitrogen). The presence of a PCR product insert was confirmed by Eco-RI digestion (New England BioLabs), and inserts were sequenced using the T7 tag within the plasmid (Vanderbilt DNA Sequencing Facility).

Long-range PCR

Long-range PCR was performed with the LongRange PCR Kit (QIAGEN), as per the manufacturer's instructions. An *FGFR1OP2*-F primer (AGATGATCCGGGTATAATAA) within exon 4 and an *FGFR1*-R primer (AGAAGAACCCCAGAGTTCAT) within exon 10 were used to amplify the genomic fusion sequence. The products were separated by agarose gel electrophoresis, and excess dNTPs and primers were removed with ExoSAP-IT (USB Corporation). The product was sequenced in steps using the primers FGFR1.fus-R1 (TCCAAAGACCATGGTAGGCC), FGFR1.fus-R2 (CACCTCTTCCAGCTTGACAT), FGFR1.fus-R3 (CGGTCATTCTTGCACACACC) and FGFR1.fus-R4 (ATGGGAGGGACCTGGTAGGA) at the Vanderbilt DNA Sequencing Facility.

RESULTS

In silico analysis of TK fusions

Using an *in silico* approach, we analyzed the protein sequences from known cancer-derived TK rearrangements

($n = 59$; Supplementary Table S1). Strikingly, all TK alterations identified to date contain an intact conserved GXGXXG motif (Figure 1A), which is essential for kinase activity (30). We also found that fusion points within the TK protein sequence usually occur within ~200 amino acids upstream of this motif, with few exceptions (e.g. JAK2 and ABL1).

This result prompted us to examine corresponding genomic fusion points at the DNA level. Although fewer DNA fusion sequences have been mapped, these also occur within a defined region (Figure 1B). In most instances, the GXGXXG motif is encoded by a single exon. The distance from the GXGXXG motif to the fusion point was variable due to the differences in intron and exon sizes. However, 80% of the TK fusions we analyzed had a fusion point within three introns upstream from the GXGXXG-encoding exon (Figure 1C). Breaks within *ABL1*, *-2* and *ROS* occurred outside of this pattern. Assuming that novel TK fusions will follow a similar pattern, it should be possible to search systematically in a non-biased manner for fusions involving breaks within these regions in any of the 90 TKs in the human genome (25) using genomic DNA.

Strategy for systematic discovery of TK fusions

Our strategy for TK fusion discovery employs existing 'DNA capture' technology (21–23) followed by 'next-generation' sequencing (31) of recovered sequences. SureSelect technology (Agilent) captures specific DNA regions of interest ('catch') using 120-mer RNA 'baits' in-solution, allowing for enrichment of target sequences compared to unselected DNA (22). For this project, we paired SureSelect Target Enrichment technology with 454 sequencing (Figure 2) because the amount of starting template needed for SureSelect is the least (1.5 μg) of any of the existing capture platforms, and the 454 platform delivers the longest read-lengths per sequence among next-generation sequencing platforms. To date, pairing 'in-solution' capture with long read sequencers (e.g. 454) has not been reported. We reasoned that longer reads would allow us to directly find fusion points without the need for intensive bioinformatic mapping algorithms.

To capture genomic regions of interest, we mapped the nucleotides encoding the GXGXXG motif for all 90 TKs in the human genome (Supplementary Table S2). We also included *AKT-1*, *-2* and *-3*, and *BRAF*; these serine-threonine kinases have been implicated in cancer, and *BRAF* is rearranged in a subset of thyroid cancers (32). These regions were extended to include the entire GXGXXG-encoding exon, two preceding exons, and three preceding introns (Figure 1C). For TKs shown repeatedly to break outside this pattern (e.g. *ABL1*, *-2* and *ROS*), the capture region was increased to include those areas where previous breaks had been observed. Based on our *in silico* analysis, capture of these mapped regions should detect 92% of known fusion points. The collective genomic coordinates were then submitted for custom bait design with 2x coverage for all capture regions. Coordinates for *AATK* were inadvertently left out of bait design. This focused capture strategy targets the regions where fusion points are most likely to occur and should reduce recovery of 'diluting' wild-type sequences. To decrease the amount of low complexity DNA, repeating regions (as identified by the UCSC genome browser, hg18) were excluded from bait design. Bait tiling averaged 73% across all targets (range: 34–100%, excluding *AATK*; Supplementary Table S2). Therefore, evenness of bait coverage across the target regions was dependent on the presence or absence of repetitive elements.

Analysis of enriched sequences following capture

The assay was validated using DNA from two human cancer cell lines with known TK fusions: the thyroid papillary carcinoma cell line, TPC-1, known to harbor a *CCDC6-RET* fusion (27), and the acute myeloid leukemia cell line, KG-1, known to contain an *FGFR1OP2-FGFR1* fusion (28). For both cell lines, fusions had been previously identified only at the mRNA level. Genomic DNA was sheared and ligated to 454 sequencing adaptors (Figure 2). The length of the sheared DNA (300–500 nt) was significantly longer than the baits (120-mers), allowing for capture of fusion sequences with only a short TK-containing portion.

Following DNA capture, ~60 000–100 000 reads per sample were generated using the 454 FLX platform (Table 1). In total, the length of recovered reads averaged 193 bp (Supplementary Figure S1), and sequences corresponding to TK 'baited' regions were enriched ~776-fold, indicating the efficiency of our capture. Twenty two per cent of the 'catch' mapped to bait regions, and the average enrichment across all kinases ranged from 80- to 3180-fold (excluding *AATK*; Supplementary Table S2). To visualize the 'bait' and 'catch' coverage across the genome, these files were loaded as custom tracks into the UCSC genome browser (publicly available, 'Materials and Methods' section). The average bait coverage for TPC-1 sequences was 0.90x (range 0–17.0x) and 1.51x for KG-1 sequences (range 0–29.4x). Of the 13 972 baits used for capture, 2552 (18.3%) were covered by >2 TPC-1 sequences. 3 697 baits (26.5%) had >2x coverage by KG-1 sequences. The differences in coverage between the two cell lines can be explained partly by the greater number of sequences recovered from capture of KG-1 DNA (Table 1).

The recovered sequences were analyzed for fusions using two novel independently derived computational algorithms (Supplementary Figure S2). Both BLAT-based methods separated target-containing sequences from non-target containing sequences as an initial filter. This was achieved by aligning the sequences to a library of known human genes or the entire human genome. Each algorithm generated a list of potential candidate fusions from the target-containing sequences using slightly different alignment and stringency criteria ('Materials and Methods' section). An entire sequence read had to be completely 'mappable' with small (2–5 bp) gaps or overlapping regions. To reduce the number of false positives, fusion

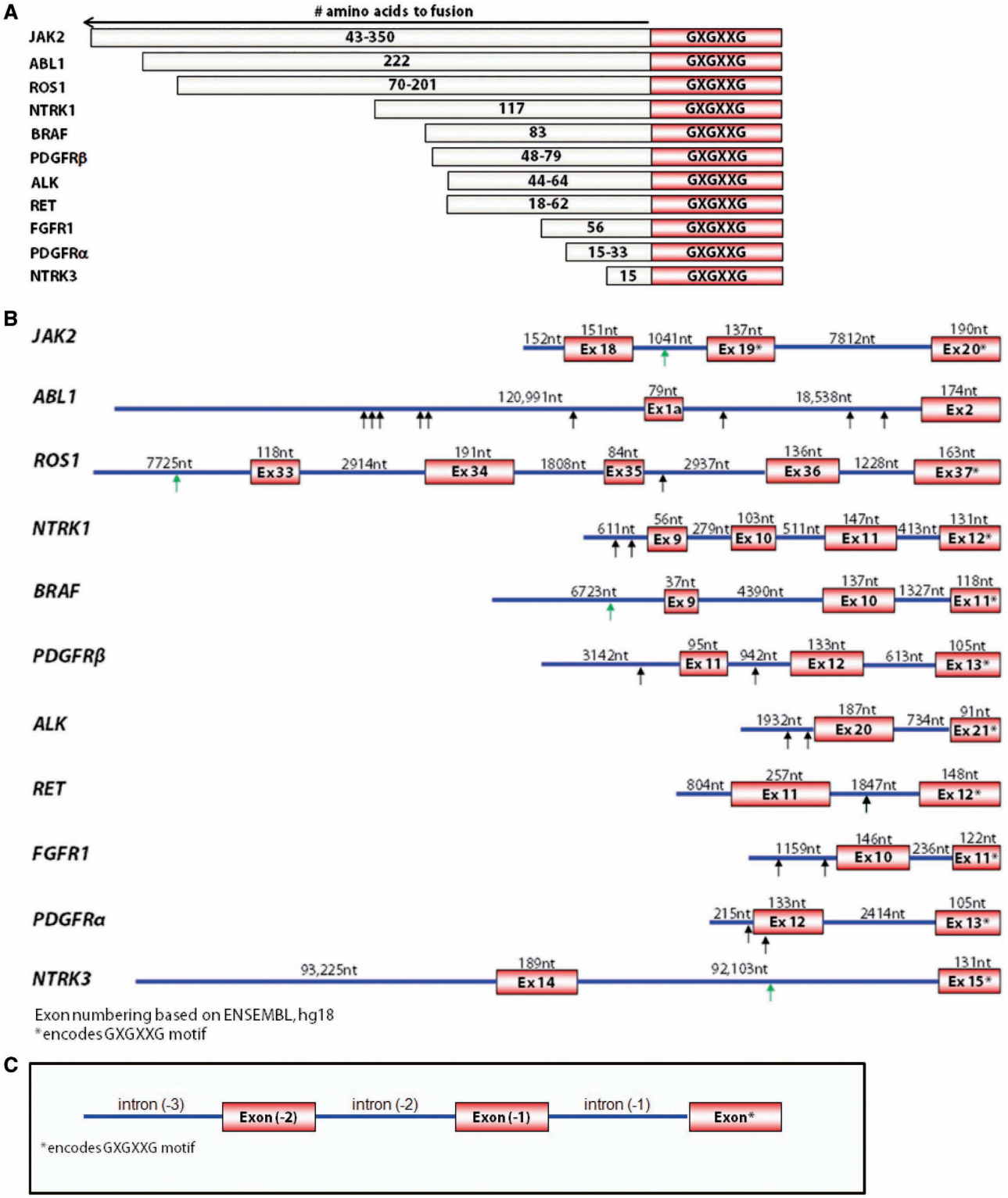

Figure 1. Conserved fusion point patterns from cancer-derived TK fusions. (**A**) All TK fusions identified to date contain an intact GXGXXG motif essential for kinase function. Fusion points occur within a defined region (~200 amino acids) upstream of this motif, with few exceptions (e.g. JAK2, ABL1). Fusion points are usually conserved within the same kinase, regardless of the upstream partner. (**B**) At the genomic level, the GXGXXG motif is usually encoded by a single exon (asterisked). Genomic breaks usually occur within the three introns preceding the GXGXXG-encoding exon. The GXGXXG-encoding exon for *ABL1* (exon 4) is not pictured. The TK order is the same as that in (A). Green arrows indicate known genomic fusion points; black arrows indicate suspected fusion points that have not been mapped at the genomic level. (**C**) Based on this analysis (A and B), we targeted the region upstream of the GXGXXG-encoding exon for DNA capture. The genomic coordinates for the GXGXXG-encoding exon, the two preceding exons, and the three preceding introns were mapped for all 90 human TKs. Four serine/threonine kinases (*BRAF*, *AKT-1*, *-2*, -3) were also included. For kinases that repeatedly break outside this pattern (e.g. *ROS1*, *ABL1*), additional regions were included in the capture design.

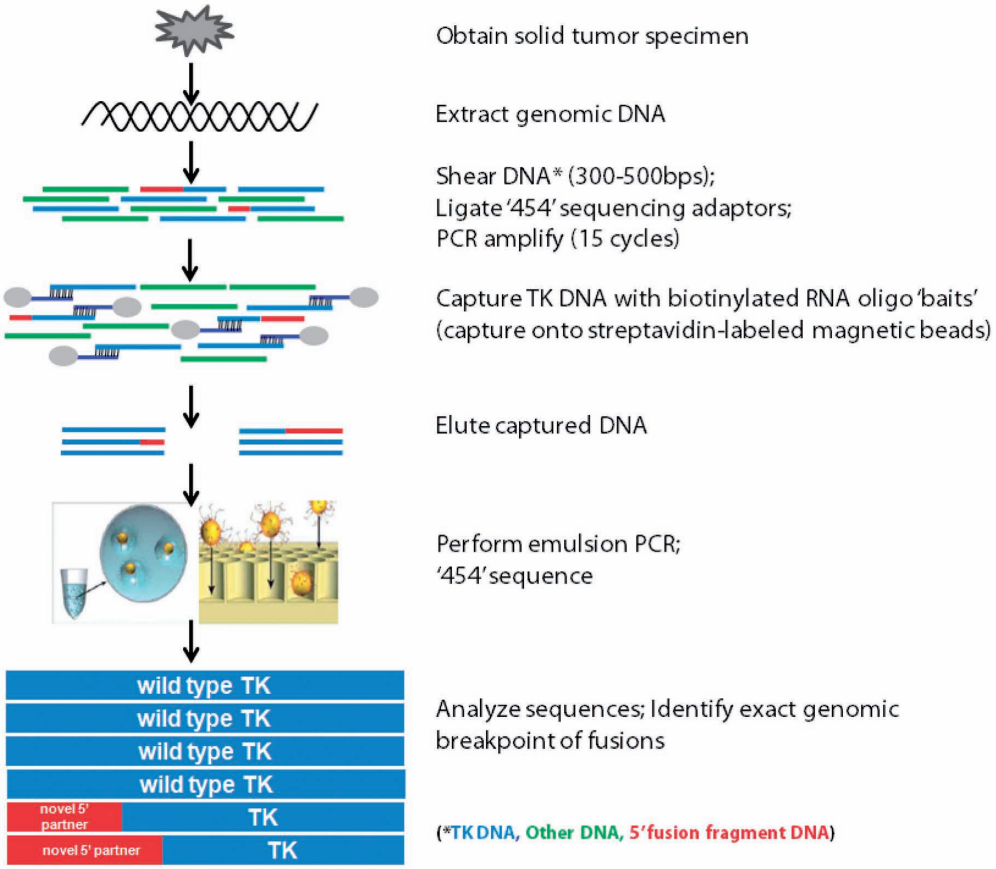

Figure 2. Schematic of experimental workflow. 454 depiction adapted from Margulies *et al.* 2005.

Table 1. 454 sequencing and bait enrichment statistics

Sample	TK status	Total number of reads	Number of bait hits	Hit (%)	Fold enrichment
TPC-1	CCDC6-RET	58 216	12 564	22	772.32
KG-1	FGFR1OP2-FGFR1	95 393	21 161	22	793.84

The total number of mappable reads was calculated for each sample. These reads were then mapped against the genomic coordinates of the custom designed baits within each target region (Figure 1C) to determine how many sequences 'hit' the baits. Because repeating regions were excluded from bait design, the genomic regions against which baits were designed were smaller than the target regions originally outlined. The fold enrichment represents the increased probability of recovering kinase sequences after capture compared to their recovery by chance alone among sequences from the entire genome

sequences also had to be recovered at least twice. Both methods produced similar results, and a combined list of fusion candidates was compiled for each sample (Supplementary Table S3).

Identification of the *CCDC6-RET* genomic fusion sequence in TPC-1 cells

Approximately 60 000 reads were recovered from captured TPC-1 DNA (Table 1). Around 22% of sequences mapped to baited target regions, translating to a 772-fold enrichment of kinase-containing DNA over other genomic sequences. Computational analyses identified 12 potential kinase rearrangements (Supplementary Table S3), including a fusion sequence recovered twice that mapped to intronic regions from *CCDC6* and *RET* (Figure 3A). This *CCDC6-RET* fusion was the only candidate validated by PCR and direct sequencing (Figure 3B–D). The 12 nt sequence surrounding the fusion point was subsequently queried against the entire pool of TPC-1 454 sequences. One additional fusion sequence was found; it was likely missed by the automated algorithms, because it contained only a short (15 bp) fragment of *CCDC6*. In total, all three sequences contained the same fusion point (Figure 3A). These data demonstrate the feasibility of our platform to detect TK fusions from genomic DNA.

We next investigated the 'bait' and 'catch' coverage across *RET* kinase by simultaneously mapping these coordinates within the UCSC browser (26) (Figure 3E). Approximately 97% of *RET* was covered with baits, and *RET*-containing sequences covered most of the target region (1053-fold enrichment; 3.6-fold average coverage; Supplementary Table S2). Areas with few or no recovered sequences contained mostly repetitive regions, which did not have corresponding baits. The *CCDC6-RET* fusion sequences could have been captured by four separate baits within intron 11-12 of *RET*.

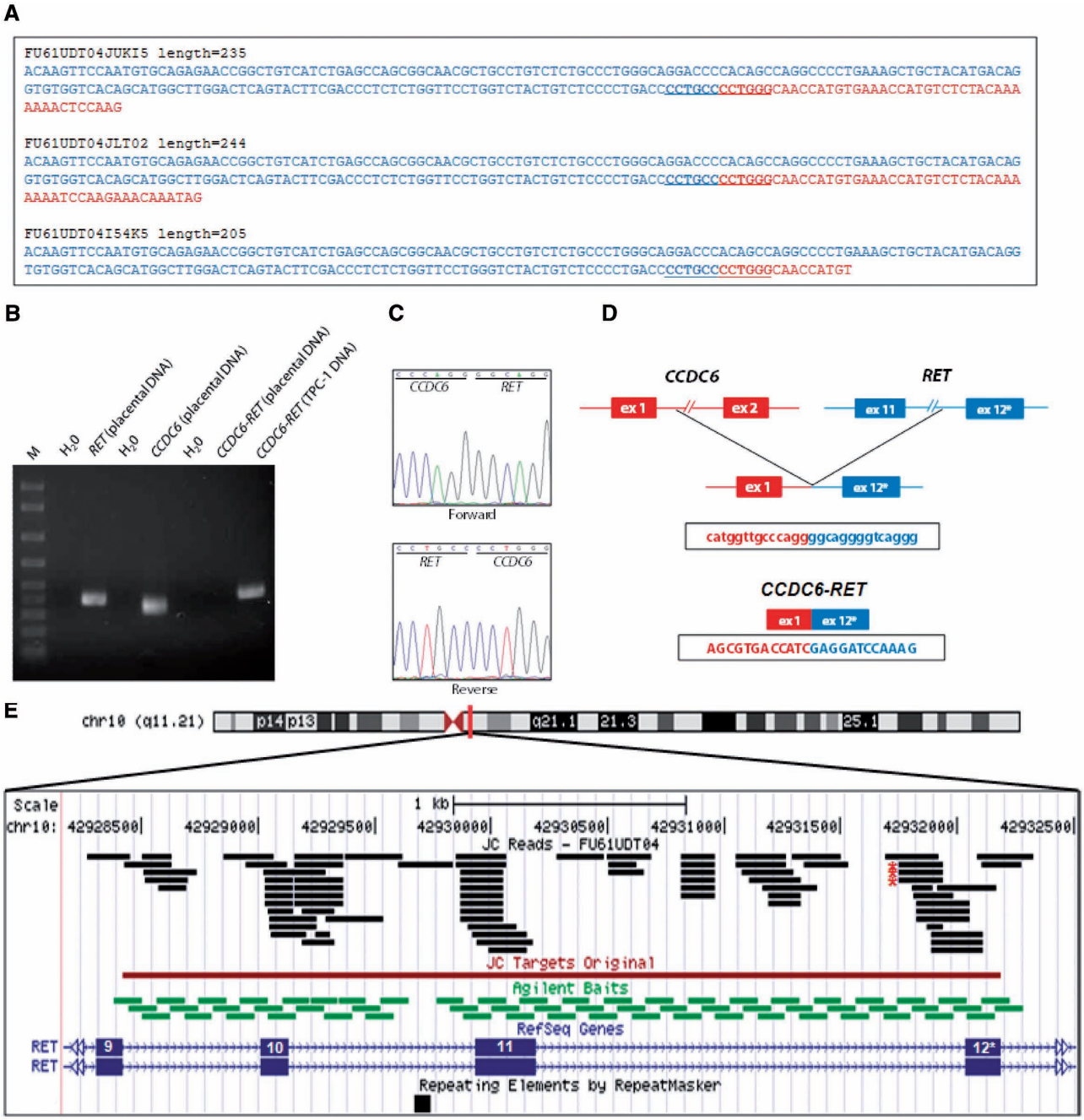

Figure 3. Identification of the genomic fusion point of *CCDC6-RET* in TPC-1 thyroid cancer cells. (**A**) Using the capture-sequence method, three different candidate sequences (labeled with 'FU61UDT04' barcodes) were recovered that mapped to the intron between *RET* exons 11 and 12 (blue) and the intron between *CCDC6* exons 1 and 2 (red). (**B**) PCR using primers specific for wild-type *RET*, wild-type *CCDC6*, and the identified *CCDC6-RET* fusion were used to confirm presence of the fusion only in TPC-1 cells. (**C**) Direct sequencing chromatograms of the PCR product containing the fusion. The corresponding 12 nt are underlined in A. (**D**) Schematic of the genomic structure of the fusion point. The GXGXXG-encoding exon is marked with an asterisk. (**E**) 454 sequences from captured TPC-1 DNA (black) and the custom 120-mer baits (green) designed for capture were mapped to the genome. The target capture region for *RET* (red) is shown as a reference. The location of the gene within the chromosome is denoted by a vertical red bar. Repeating elements (black box) were excluded from bait design. The three *CCDC6-RET* fusion sequences are starred.

Mapping the genomic breakpoint involving *FGFR1OP2-FGFR1* in KG-1 cells

Reads recovered from captured KG-1 DNA were enriched ~794-fold for TK-containing sequences (Table 1). From the 22% of sequences that mapped to baited target regions, we identified 30 potential TK alterations, one of which involved two non-contiguous portions of *FGFR1* (Figure 4A, Supplementary Table S3). Only the fusion involving *FGFR1* was confirmed by PCR using fusion-specific primers and direct sequencing of the PCR products (Figure 4B and C). No additional reads were found after querying the *FGFR1* fusion sequence against

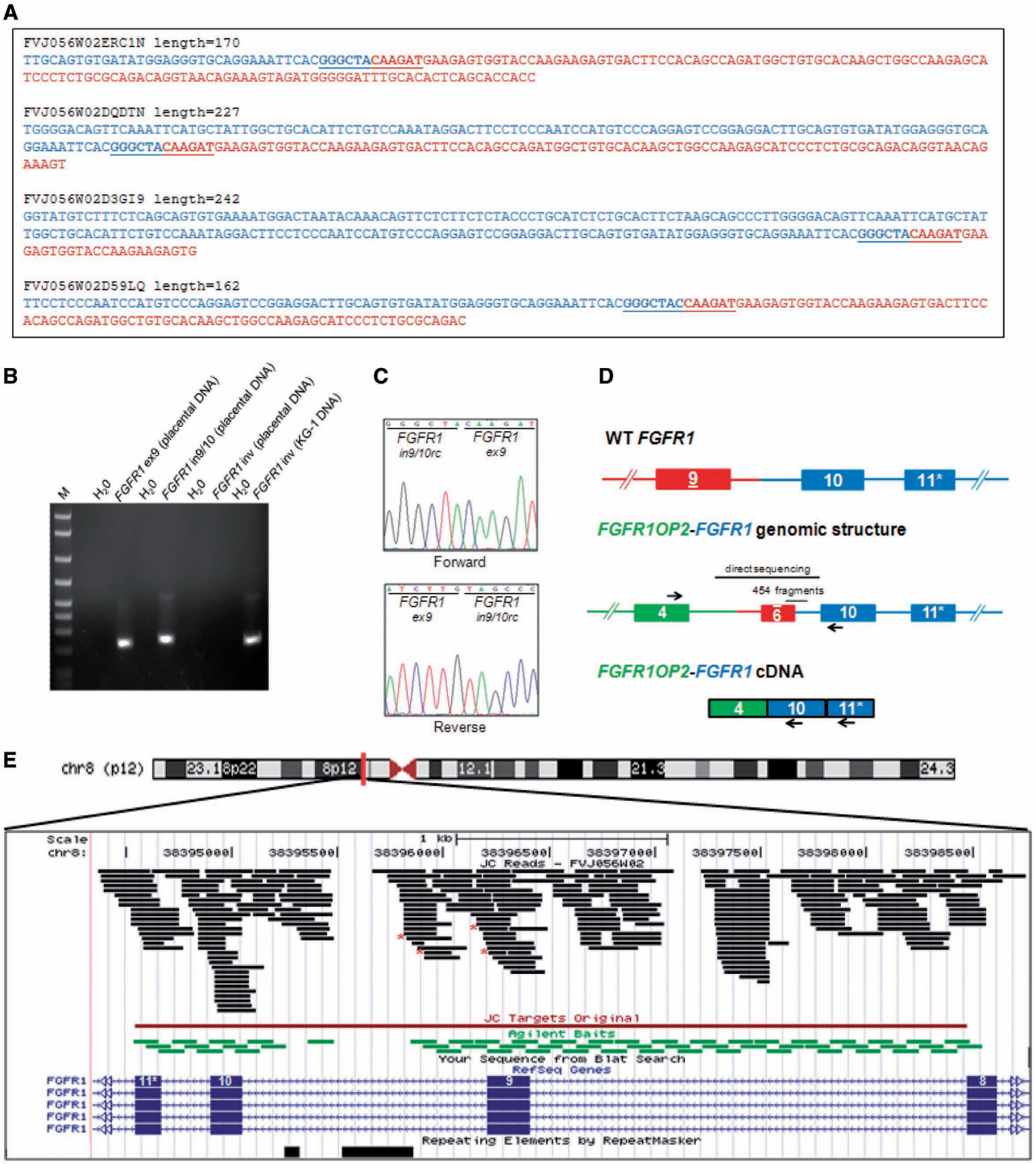

Figure 4. Genomic structure of *FGFR1OP2-FGFR1* in KG-1 cells. (**A**) Four sequences (labeled with 'FVJ056W02' barcodes) were recovered that mapped to the forward strand of exon 9 (red) and the reverse complement (rc) of the intron between exons 9 and 10 (blue) within *FGFR1*, indicating a possible fusion. The fourth sequence contains an extra C across the fusion point. This may be due to a 454 sequencing artifact, as it was not detected by direct sequencing. (**B**) The fusion was confirmed only in KG-1 cells by PCR using breakpoint-spanning primers. The *FGFR1* regions from exon 9 (ex9) and intron 9-10 (in9/10) correspond to the areas disrupted as a consequence of the fusion. (**C**) Direct sequencing chromatograms of the PCR product containing the fusion. The corresponding 12 nt are underlined in **A**. (**D**) Schematic of the genomic structure of *FGFR1OP2-FGFR1*. Primers used for long-range PCR and 5′RACE are indicated with arrows. The GXGXXG-encoding exon is marked with an asterisk. (**E**) 454 sequences from captured KG-1 DNA (black) were mapped to the genome with the custom 120-mer baits (green), and the target region for *FGFR1* (red). The location of the gene within the chromosome is denoted by a vertical red bar. Repeating elements (black boxes) were excluded from bait design. The four sequences containing the *FGFR1* rearrangement (starred) were mapped to the region with highest homology.

the pool of KG-1 454 sequences. The four sequences containing this fusion mapped to the forward strand of a portion of exon 9 and to the reverse complement of a portion of the intron between exons 9 and 10, suggesting an *FGFR1* rearrangement occurring 5′ to exon 10 (Figure 4D). However, from the 454 reads alone, we were not able to deduce the exact upstream partner. Subsequent application of 5′–RACE using primers specific to exons 10 and 11 of *FGFR1* found only the *FGFR2OP1-FGFR1* fusion previously reported (data not shown).

To elucidate further how the recovered 454 fusion sequences were related to the *FGFR2OP1-FGFR1* alteration, we performed sequencing of long-range PCR products obtained from amplification of the DNA sequence between exon 4 of *FGFR1OP2* and exon 10 of *FGFR1*. This ~5 kb region contained sequences that matched 100% to our 454 fusion reads, but we found that the genomic structure was much more complex, involving elements from intron 4-5 of *FGFR1OP2,* the inverted truncated exon 9, and intron 9-10 of *FGFR1* (Figure 4D). Collectively, these data illustrate two points. First, data from our capture-sequence approach suggested a breakpoint occurring at a specific region in the kinase and were sufficient to allow us to find readily the upstream partner using 5′-RACE. Second, fusion breakpoints can be more complex than just the juxtaposition of two intronic elements from two different genes.

Approximately 88% of the *FGFR1* capture region was covered with baits (Figure 4E; Supplementary Table S2), and *FGFR1*-containing KG-1 sequences were enriched 1922-fold with an 11-fold average coverage of the target region. Simultaneous mapping of the baits and sequences revealed a similar pattern to that observed for TPC-1 reads (Figure 4E). A large repeating region between exons 9 and 10 contained no sequence coverage due to the lack of baits. The four *FGFR1* sequences containing the rearrangement could have been captured by 8 baits.

Analysis of NSCLC control cell lines

Three well-characterized lung cancer cell lines without known fusions were also included as controls: NCI-H820, harboring mutant *EGFR* (exon 19 deletion, T790M) and amplification of *MET* (33), NCI-H1703, with amplification of *PDGFRβ* (9) and NCI-H3255, containing amplified mutant *EGFR* (L858R) (34). Analysis of the sequences from H820, H3255 and H1703 identified 8, 12 and 24 candidate fusions, respectively (data not shown). None of the fusions were confirmed by PCR with breakpoint spanning primers. Enrichment of TK-containing sequences for these lines was similar to that of the fusion lines (data not shown). These data are consistent with a modest false positive rate also observed with other fusion discovery efforts that have used 454 sequencing (13,17). We hypothesize that the false positive sequences are likely an artifact of the ligation step used to add sequencing adaptors onto DNA fragments. We expect this number to decrease in future iterations of the method as refinements are made to the sequencing preparation steps.

DISCUSSION

The success of the ABL TK inhibitor, imatinib, in CML patients with BCR-ABL translocations and of the new ALK TK inhibitor in lung cancer patients with EML4-ALK fusions illustrates that cancer-driving TK fusions serve as excellent therapeutic targets. Exactly how many TK fusions exist in cancers, though, is currently unknown, because identification of TK fusions has traditionally required highly laborious techniques. As an example, BCR-ABL rearrangements were identified using conventional karyotyping, which first revealed the Philadelphia chromosome (35), followed by identification of the *t*(9;22) translocation (36), and eventual cloning of the ABL translocation (37). EML4-ALK fusions were discovered by (i) application of a tumor-derived cDNA expression library to a mouse 3T3 fibroblast focus formation assay (10) and (ii) immunoaffinity phosphoproteomic profiling by mass spectrometry (9).

Based upon an *in silico* analysis, we found that known TK rearrangements display conserved fusion point properties that might make them amenable to systematic screening using new technologies. This analysis led us to design a novel DNA-based approach to identify TK fusions in a selectively targeted high-throughput manner. Regions of genomic DNA likely to be involved in TK fusion events are captured by hybridization to custom designed RNA baits. Captured DNA is eluted and sequenced using next-generation 454 sequencing. We chose 454 sequencing over other deep sequencing platforms specifically because of its ability to achieve longer read lengths (~200 nt; Table 1, Supplementary Figure S1), which could facilitate direct identification of fusion points that lie far upstream from the captured region (Figure 3A). Novel computational algorithms then allow for rapid identification of candidate fusions, which are validated by simple direct PCR or 5′-RACE methods. As proof that this approach is feasible and robust, we used it to map previously unknown genomic breakpoints in two human cancer cell lines (including both solid and hematologic malignancies) harboring fusions with *RET* and *FGFR1,* respectively, using only 1.5 μg of DNA from each line. In both cases, we identified novel genomic fusion sequences and structures of the breakpoints. Importantly, having established workflow for the platform, identification of candidate fusion sequences in a given tumor sample could theoretically be completed in ~2–3 weeks, encompassing DNA isolation and preparation, DNA capture, 454 sequencing and computational analysis.

The advent of high throughput next-generation sequencing technologies has recently facilitated the discovery of multiple types of translocation events in cancer cells (Table 2). One approach involves whole genome sequencing, which requires adequate starting material and large numbers of sequences for adequate coverage. As a result, genome-wide next-generation sequencing efforts are often coupled with copy number and karyotype data to prioritize regions where translocations may have occurred (15,16). This strategy requires complex computational algorithms that integrate sequencing and chromosomal analyses.

Table 2. Comparison of deep sequencing-based fusion discovery efforts

Method	Focus	Sequencer	Starting material	Length of reads	Total no. of reads
Genome wide paired end	Somatic rearrangements	Illumina GA	5 μg gDNA	29–36 nt	36.2 M
Transcriptome sequencing	All genes fusions	Illumina GA with 454	50 μg total RNA	36 nt–230 nt	66.9 M/500–800 K
Transcriptome paired end	All genes fusions	Illumina GA	1 μg mRNA	50–100 nt	16.9–25 M
Targeted RNA-seq	Cancer-related genes	Illumina GA	3 μg mRNA	76 nt	8
TK capture sequence	**TK fusions**	**454**	**1.5 μg gDNA**	**220 nt**	**60–100 K**

Multiple studies have focused on fusion discovery within cancer samples and cell lines. While the recovery of known and unknown fusions has been highest in transcriptome-based approaches, the amount of starting material is often limiting for most patient samples. Furthermore, the number of reads required to detect confidently fusions presents financial and computational challenges. This study (bold) outlines a strategy that uses minimal amounts of genomic DNA and requires analysis of fewer sequence reads. However, sequencing strategies are evolving rapidly, and additional improvements are expected in starting material requirements and computational algorithms for the analysis of large quantities of sequences. See Discussion text for references. gDNA = genomic DNA; nt = nucleotides; M = million; K = thousand

A number of RNA-based approaches (RNA-seq) also have been used to detect multiple types of fusion events (Table 2). These include whole transcriptome sequencing (13,14,18), and screening for disparate levels of expression between 5′ and 3′-ends of kinase genes on exon arrays (with expression higher in the 3′-end) (38). Whole transcriptome sequencing provides meaningful expression level data while eliminating background from non-coding genomic elements. However, this approach also requires analysis of a large number of sequences due to the abundance of housekeeping, ribosomal and mitochondrial transcripts, and is often coupled with copy number data as well. Recently, selective capture of 467 cancer related genes was applied to a cDNA library, and analysis of the subsequent massively parallel sequencing found multiple fusion events, demonstrating the use of this capture-sequence approach to detect translocations (20). However, the sequencing analysis demands are high. Most importantly, the sample requirements (20 μg DNA, up to 100 μg of total RNA, or 1 μg of mRNA) for the above DNA- and RNA-based platforms are prohibitive for most standard patient samples, thus limiting the broad utility of these approaches. A DNA-based capture platform avoids the use of RNA, which is inherently less stable than DNA and more difficult to extract without degradation from clinical specimens.

While the fusion events identified through other methods are important for understanding the pathogenesis of cancer cells, their therapeutic potential remains unknown. Therefore, we focused specifically on identification of TK fusions which can be potentially targeted with specific kinase inhibitors. Based on the genomic properties of known TK fusions, we designed a highly focused capture platform that should detect 92% of known TK fusions. From ~160 000 sequences, we recovered only one sequence that fell within the housekeeping gene category. Furthermore, high density SNP analysis (Affymetrix 6.0 arrays) of DNA from both TPC-1 and KG-1 cells revealed that only the *CCDC6-RET* fusion was apparent from copy number studies (data not shown). Although this approach was validated using only cell line DNA, we do not anticipate problems using DNA from more heterogeneous tumor samples. Given the increased efficiency of capture methods developed by Agilent since our pilot study, we expect a greater percentage of the ~100 000 recovered sequences from each run to map back to our target regions. Even from this pilot study, we know that fusions are still detectable with as few as ~13 000 sequences (the number of TPC-1 sequences that mapped back to kinase targets; Table 1).

While our capture-sequence platform has distinct advantages, there are some limitations. First, if fusions occur outside of the targeted regions or within large repetitive elements, they would not be identified with this method. However, based on our *in silico* analysis, the majority of fusions should occur within the targeted regions, and our original design attempted to balance feasibility with the effort required to sift through hundreds of thousands of wild-type sequences. Other groups have been unable to detect a conserved motif at translocation breakpoints (16), indicating that fusions may occur in broadly defined regions versus at specific sequence motifs. Many of the genomic fusion sequences that we have analyzed occur outside of repeat elements, and those that occur within masked regions are still amenable to capture due to the length of the recovered sequences extending into regions without baits (Chmielecki and Pao, unpublished observations). Second, we detected a modest number of false positive fusion events (~0.025% of total sequences). However, this is a common problem for all fusion discovery platforms (with DNA and RNA), and the number of candidate fusions identified with our method is comparable to other similar methods. Third, the percentage of total sequences mapping to our baits is much lower than similar capture reports (22). However, our study used an early 'beta' version of the SureSelect technology, and capture has improved significantly since the official launch of the product. Additionally, SureSelect was not optimized for 454 sequencing at the time of this study. Fourth, other types of fusion events (e.g. ETS gene fusions) were omitted from our design, because targeted therapies for these types of translocations have yet to be developed. However, in theory, a similar capture platform could be designed based on the genomic properties of transcription factor fusions. Fifth, gaps in sequence coverage may be unavoidable due to deletions in tumor cell DNA, GC-richness, and regions not amenable to PCR

amplification. This fact is demonstrated in our pilot project where sequence gaps across regions are not the same for each sample (see custom UCSC Genome Browser). Finally, we note that strategies for fusion identification (Table 2) are rapidly evolving, and this platform represents just one approach towards kinase fusion discovery. The capture strategy we designed could theoretically be paired with any sequencing technology (e.g. SOLiD and Illumina GA, for which SureSelect is already optimized). As computational algorithms are refined further for such applications, these different sequencing approaches may represent alternatives that offer greater cost-effective analysis.

In the future, we plan to increase bait coverage from 2x to 5x, allowing for greater enrichment of our target regions. Baits with poor capture efficiency are also being redesigned to better capture some genomic regions. One redesign strategy includes creating baits against the opposite DNA strand in the event that the opposite sequence is more amenable to capture. Recovered sequences will then be sequenced using the 454 Titanium platform, which allows for much longer reads (~500 bp). This longer read length will enhance sequence coverage across the target regions and allow for sequencing into areas not covered directly with baits. Additionally, the improved bait design paired with longer sequencing reads should further decrease the number of false positive candidate fusions by mapping more precisely the recovered sequences.

In summary, we have devised a DNA-based platform using highly focused targeted capture and 454 sequencing for rapid and systematic discovery of TK fusions in cancers. Our data demonstrate that this novel method is feasible, rapid and applicable to routine tumor samples. Using this platform, we identified for the first time genomic fusion sequences of two TK fusion proteins (*CCDC6-RET* and *FGFR1OP2-FGFR1*), including the unique genomic structure of *FGFR1OP2-FGFR1*. We now plan to screen tumor sets for novel kinase rearrangements. To extend further its utility, we will also determine if this method can be used on DNA extracted from formalin-fixed paraffin-embedded tissue. Ideally, we hope to accelerate discovery of multiple novel TK fusions and facilitate clinical development of targeted anti-cancer therapies.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Juan Li of the MSKCC Genomics Core Laboratory for assistance with sample preparation and 454 sequencing; Ross Levine and Jim Fagin (MSKCC) for providing KG-1 and TPC-1 cells, respectively; Alex Lash and Caitlin Byrne of the MSKCC computational biology core for assistance with genomic mapping; Paula Woods (VICC) for technical assistance and Jennifer Pietenpol (VICC) for critical reading of the manuscript.

FUNDING

National Institutes of Health National Cancer Institute (NCI) [grants R01-CA121210 (to W.P.); P01-CA129243 (to W.P.)]; Stand Up To Cancer-American Association for Cancer Research Innovative Research Grant [SU2C-AACR-IR60109 (to W.P.)]. The MSKCC Genomics core is supported by an NCI CCSG award to MSKCC (P30-CA008748). W.P. received additional support from Vanderbilt's Specialized Program of Research Excellence in Lung Cancer grant (CA90949) and the VICC Cancer Center Core grant (P30-CA68485). R.K.T. is supported by the German Federal Ministry of Science and Education (BMBF) as part of the German National Genome Research Network (NGFNplus) program (grant 01GS08100). Funding for open access charge: SU2C-AACR Innovative Research Grant (SU2C-AACR-IR60109).

Conflict of interest statement. None declared.

REFERENCES

1. Sawyers,C. (2004) Targeted cancer therapy. *Nature*, **432**, 294–297.
2. Druker,B.J. (2002) STI571 (Gleevec) as a paradigm for cancer therapy. *Trends Mol. Med.*, **8**, S14–18.
3. Hantschel,O., Nagar,B., Guettler,S., Kretzschmar,J., Dorey,K., Kuriyan,J. and Superti-Furga,G. (2003) A myristoyl/phosphotyrosine switch regulates c-Abl. *Cell*, **112**, 845–857.
4. Nagar,B., Hantschel,O., Young,M.A., Scheffzek,K., Veach,D., Bornmann,W., Clarkson,B., Superti-Furga,G. and Kuriyan,J. (2003) Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell*, **112**, 859–871.
5. Kumar-Sinha,C., Tomlins,S.A. and Chinnaiyan,A.M. (2006) Evidence of recurrent gene fusions in common epithelial tumors. *Trends Mol. Med.*, **12**, 529–536.
6. Santoro,M., Rosati,R., Grieco,M., Berlingieri,M.T., D'Amato,G.L., de Franciscis,V. and Fusco,A. (1990) The ret proto-oncogene is consistently expressed in human pheochromocytomas and thyroid medullary carcinomas. *Oncogene*, **5**, 1595–1598.
7. Ciampi,R., Knauf,J.A., Kerler,R., Gandhi,M., Zhu,Z., Nikiforova,M.N., Rabes,H.M., Fagin,J.A. and Nikiforov,Y.E. (2005) Oncogenic AKAP9-BRAF fusion is a novel mechanism of MAPK pathway activation in thyroid cancer. *J. Clin. Invest.*, **115**, 94–101.
8. Rabes,H.M., Demidchik,E.P., Sidorow,J.D., Lengfelder,E., Beimfohr,C., Hoelzel,D. and Klugbauer,S. (2000) Pattern of radiation-induced RET and NTRK1 rearrangements in 191 post-chernobyl papillary thyroid carcinomas: biological, phenotypic, and clinical implications. *Clin. Cancer Res.*, **6**, 1093–1103.
9. Rikova,K., Guo,A., Zeng,Q., Possemato,A., Yu,J., Haack,H., Nardone,J., Lee,K., Reeves,C., Li,Y. *et al.* (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, **131**, 1190–1203.
10. Soda,M., Choi,Y.L., Enomoto,M., Takada,S., Yamashita,Y., Ishikawa,S., Fujiwara,S., Watanabe,H., Kurashina,K., Hatanaka,H. *et al.* (2007) Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, **448**, 561–566.
11. Takeuchi,K., Choi,Y.L., Togashi,Y., Soda,M., Hatano,S., Inamura,K., Takada,S., Ueno,T., Yamashita,Y., Satoh,Y. *et al.* (2009) KIF5B-ALK, a novel fusion oncokinase identified by an immunohistochemistry-based diagnostic system for ALK-positive lung cancer. *Clin. Cancer Res.*, **15**, 3143–3149.
12. Horn,L. and Pao,W. (2009) EML4-ALK: Honing in on a new target in non-small-cell lung cancer. *J. Clin. Oncol.*, **27**, 4232–4235.

13. Maher,C.A., Kumar-Sinha,C., Cao,X., Kalyana-Sundaram,S., Han,B., Jing,X., Sam,L., Barrette,T., Palanisamy,N. and Chinnaiyan,A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.
14. Maher,C.A., Palanisamy,N., Brenner,J.C., Cao,X., Kalyana-Sundaram,S., Luo,S., Khrebtukova,I., Barrette,T.R., Grasso,C., Yu,J. *et al.* (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.
15. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
16. Stephens,P.J., McBride,D.J., Lin,M.L., Varela,I., Pleasance,E.D., Simpson,J.T., Stebbings,L.A., Leroy,C., Edkins,S., Mudie,L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.
17. Zhao,Q., Caballero,O.L., Levy,S., Stevenson,B.J., Iseli,C., de Souza,S.J., Galante,P.A., Busam,D., Leversha,M.A., Chadalavada,K. *et al.* (2009) Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc. Natl Acad. Sci. USA*, **106**, 1886–1891.
18. Berger,M.F., Levin,J.Z., Vijayendran,K., Sivachenko,A., Adiconis,X., Maguire,J., Johnson,L.A., Robinson,J., Verhaak,R.G., Sougnez,C. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res.*
19. Leary,R.J., Kinde,I., Diehl,F., Schmidt,K., Clouser,C., Duncan,C., Antipova,A., Lee,C., McKernan,K., De La Vega,F.M. *et al.* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Science Translational Medicine*, **2**.
20. Levin,J.Z., Berger,M.F., Adiconis,X., Rogov,P., Melnikov,A., Fennell,T., Nusbaum,C., Garraway,L.A. and Gnirke,A. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.*, **10**, R115.
21. Albert,T.J., Molla,M.N., Muzny,D.M., Nazareth,L., Wheeler,D., Song,X., Richmond,T.A., Middle,C.M., Rodesch,M.J., Packard,C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.
22. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
23. Okou,D.T., Steinberg,K.M., Middle,C., Cutler,D.J., Albert,T.J. and Zwick,M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.
24. Hodges,E., Xuan,Z., Balija,V., Kramer,M., Molla,M.N., Smith,S.W., Middle,C.M., Rodesch,M.J., Albert,T.J., Hannon,G.J. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–1527.
25. Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
26. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.
27. Jossart,G.H., Greulich,K.M., Siperstein,A.E., Duh,Q., Clark,O.H. and Weier,H.U. (1995) Molecular and cytogenetic characterization of a t(1;10;21) translocation in the human papillary thyroid cancer cell line TPC-1 expressing the ret/H4 chimeric transcript. *Surgery*, **118**, 1018–1023.
28. Gu,T.L., Goss,V.L., Reeves,C., Popova,L., Nardone,J., Macneill,J., Walters,D.K., Wang,Y., Rush,J., Comb,M.J. *et al.* (2006) Phosphotyrosine profiling identifies the KG-1 cell line as a model for the study of FGFR1 fusions in acute myeloid leukemia. *Blood*, **108**, 4202–4204.
29. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
30. Hanks,S.K., Quinn,A.M. and Hunter,T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, **241**, 42–52.
31. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
32. Ciampi,R., Knauf,J.A., Rabes,H.M., Fagin,J.A. and Nikiforov,Y.E. (2005) BRAF kinase activation via chromosomal rearrangement in radiation-induced and sporadic thyroid cancer. *Cell Cycle*, **4**, 547–548.
33. Bean,J., Brennan,C., Shih,J.Y., Riely,G., Viale,A., Wang,L., Chitale,D., Motoi,N., Szoke,J., Broderick,S. *et al.* (2007) MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. *Proc. Natl Acad. Sci. USA*, **104**, 20932–20937.
34. Tracy,S., Mukohara,T., Hansen,M., Meyerson,M., Johnson,B.E. and Janne,P.A. (2004) Gefitinib induces apoptosis in the EGFRL858R non-small-cell lung cancer cell line H3255. *Cancer Res.*, **64**, 7241–7244.
35. Nowell,P.C. and Hungerford,D.A. (1960) Chromosome studies on normal and leukemic human leukocytes. *J. Natl Cancer Inst.*, **25**, 85–109.
36. Rowley,J.D. (1973) Letter: a new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, **243**, 290–293.
37. Ben-Neriah,Y., Daley,G.Q., Mes-Masson,A.M., Witte,O.N. and Baltimore,D. (1986) The chronic myelogenous leukemia-specific P210 protein is the product of the bcr/abl hybrid gene. *Science*, **233**, 212–214.
38. Koivunen,J.P., Mermel,C., Zejnullahu,K., Murphy,C., Lifshits,E., Holmes,A.J., Choi,H.G., Kim,J., Chiang,D., Thomas,R. *et al.* (2008) EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin. Cancer Res.*, **14**, 4275–4283.