

COMPOSITIONS AND METHODS FOR IDENTIFICATION OF GENE FUSIONS

FIELD OF INVENTION

[0001] The present invention relates to compositions and methods for identification of gene fusions. More specifically, the present invention relates to compositions and methods for ligating single stranded polynucleotides of unknown sequence and a method for identifying from a sequencing library of such polynucleotides any target gene with atypical exon configuration, or target gene which has been fused with any other DNA sequence such as a partner gene in a fusion configuration.

BACKGROUND OF THE INVENTION

[0002] Malignancies are often driven by mutations, referred to as pathogenic variants, in the genomic DNA of cells. The most common types of pathogenic variants are single nucleotide variants within the coding regions of genes, however many other types of mutations can drive oncogenesis. Approximately 5% of solid tumors are driven by structural rearrangement of chromosomes which disrupt the linear order of the chromosome resulting in fusion of normally non-adjacent regions within or between chromosomes. This can result in gene fusions at the DNA level that are transcribed into RNA transcripts and translated into a chimeric protein with oncogenic properties.

[0003] Gene fusions, or translocations, is a class of genomic structural variants, and are estimated to contribute to disease in 16% of cancers and are the primary drivers of 1% of cancers ¹. Gene fusion prevalence, driver gene identities, and partner gene repertoires vary extensively between cancer types creating a need for partner-agnostic pan-cancer gene fusion testing ¹⁻³. Gene fusion testing enables the use of targeted therapeutics available for treatment of kinase fusions ⁴. Multi-gene partner-agnostic fusion gene detection tests that have high sensitivity and specificity at low cost are essential for driving uptake of gene fusion testing in countries with high cancer rates and low reimbursement ⁵.

[0004] Such oncogenic fusions can involve fusion of a "driver gene" which provides oncogenic function with any of a large assortment of "partner genes". A variable number of exons from the driver gene (often constituting a functional domain) is linked with a variable number of exons from the partner gene (zero exons are sometimes observed if only the promoter is required). The result of this fusion can be

a chimeric protein with oncogenic function due to, for example, abnormal expression levels, abnormal cellular localization, or abnormal activity.

[0005] Gene fusions involving one or two genes commonly involve two genes with a truncated driver gene, often a kinase, that is fused with one of many partner gene promoters^{6,7}. The orientation of gene fusions can vary with the driver gene at the 5' or 3' position, driving oncogenesis through abnormal expression levels, abnormal cellular localization, or abnormal activity, ultimately affecting pathways that affect tumor growth^{1,8}. In the case of fusions involving a kinase, gene breakpoints can be located within intron regions that link a functional gene promoter with a driver kinase domain to produce a functional protein⁹. Other clinically significant structural changes include the loss of one or more exons within a single gene *e.g.* *EGFR* and *MET* exon 14 skipping events.

[0006] Molecular testing results guide the clinical management of cancer patients and the usage of targeted therapeutics available for treatment of gene fusion driven cancers. While gene fusions involve many functional classes of proteins, small molecule therapy development for oncogenic fusion genes has focused on kinase fusions^{10,11}. Tyrosine kinase inhibitors (TKIs) have been developed over the last two decades as treatments for 5' (*FGFR* and *EGFR*) and 3' (*ALK*, *CSF1*, *MET*, *NTRK*, *RET*, *ROS1*) receptor tyrosine kinase (RTK) gene fusions^{12,13}. Gene fusions involving the ligand *NRG1* also result in constitutive activity of *ERBB* family RTKs¹⁴. Similarly, MEK inhibitors show promise for *BRAF* and *RAF1* gene fusions encoding serine/threonine receptor kinases^{15,16}. The combination of fusion detection and the availability of tissue specific inhibitors has been transformative for patient outcomes with gene fusion driven cancers including prostate (*ERG*), bladder (*FGFR3*), lung (*ALK*, *ROS1*, *NRG1*), thyroid (*RET*), and bile duct (*FGFR2*)¹⁷. A new wave of tissue agnostic drugs such as those targeting *NTRK* gene fusions offers further hope for patients who previously required off-label therapeutic treatment^{13,17,18}.

[0007] Current gene fusion detection methods include next generation sequencing (NGS) methods that can interrogate many targets for aberrations in combination with any partner gene exons, while whole transcriptome sequencing can interrogate the whole exome for gene fusions but can identify many chimeric reads or passenger gene fusions^{1,2}. Therefore, classification methods can interrogate features that differentiate therapeutically relevant fusions from artifacts including driver gene identity and orientation, kinase domain retention, in-frame coding regions, and complete retention of fused exons³⁰.

[0008] The detection of gene fusions is routinely performed on formalin fixed and paraffin embedded (FFPE) tissue using single gene histochemical methods including immunohistochemistry (IHC) and Fluorescence in-situ Hybridization (FISH) ¹⁹. FFPE RNA may be fragmented which prevents ribo-depletion, or polyA tail based methods, and conversion into cDNA and amplification in general is inhibited by formalin-induced RNA modifications ^{31–33}. Additionally, the low fraction of degraded fusion transcripts in total RNA and the limited fraction of transcripts that are successfully extracted and converted into cDNA can be a challenge ^{24,34}. A hybrid-capture approach for low quality FFPE RNA samples has been described but requires higher input template amounts ³⁵. FISH provides an indication of large-scale translocations that may result in a gene fusion but does not provide information about gene fusion identity or expression. IHC enables detection of aberrant expression of protein variants but can be challenging to interpret. Neither provides the base-level resolution produced by sequencing methods that enables pan-fusion detection and functional interpretation. Consequently, many clinical labs now utilize multi-gene panel amplicon and hybrid-capture methods with next generation sequencing (NGS) for improved sensitivity or lower costs and higher throughput ^{6,20,21}. Both amplicon and hybrid capture assay technologies have unique challenges, however it is important for clinical assays cover the combinatorial diversity of partner and driver genes to guide clinical intervention ^{3,13,22}. Therefore, assays that incorporate partner gene agnostic detection methods to detect any driver gene fusion are useful for clinical testing ^{23–25}, for example, for targets like *ALK* and *NTRK* that are promiscuous in terms of 5' partners, have large introns, and are unpredictable in terms of recurrent breakpoints ¹⁸. RNA-based panels enable partner-agnostic pan-fusion assessment due to the absence of introns which greatly reduces probe tiling, sequencing depth, or multiplexing requirements in addition to providing functional evidence of transcription ²³.

[0009] Polymerase chain reaction (PCR)-based methods have been used to exponentially amplify variants for identification or characterization, however such methods require some knowledge of the variant location and adjacent sequence to design flanking sequences that can be used as primers for targeted amplification. In the case of mutations involving structural rearrangements, the break site location is only loosely constrained in sequence space within both the partner and the driver genes resulting in a large potential sequence search space.

[0010] Partner-agnostic Reverse-Transcriptase (RT)-PCR-based methods provide an elegant strategy for fusion gene identification. First, cDNA (generated from total RNA) is used as a template in order to exclude large intronic regions thereby reducing sequence search space to only the shorter exons. Second, single primer extension from driver gene exons across exon junctions followed by amplification from a ligated universal adapter sequence removes the need for knowledge of partner gene sequence or identity. As a result, the sequence search space can be drastically reduced to only that of a select number of targeted driver gene exons. Subsequent PCR-amplification and sequencing of amplicons permits identification of atypical driver gene exon order or structural rearrangements involving fusion of driver gene exons with any other partner gene sequence.

[0011] Partner-agnostic PCR-based methods use a TA ligation method to add such an adapter to the ends of all DNA fragments from a patient sample to permit amplification from within the driver-gene region. Because TA ligation relies on a single nucleotide overhang this strategy has limited efficiency and conversion rates. As a result, sensitivity is decreased, and fusions that are poorly expressed, poorly represented, or have a limited number of amplifiable molecules can be difficult to identify. An alternative approach to improve ligation efficiency is to use longer overhangs that stabilize the interaction between ligation donor and acceptor molecules. This however requires *a priori* knowledge of the donor and acceptor sequences and limits ligation to between a single acceptor and donor molecule which makes this approach impractical for adapter ligation to a mixed population of fragments with unknown sequence as is the case in patient-derived cDNA fragments. It is also useful to analyse and process sequencing data from the PCR amplified adapter-ligated DNA molecules to identify specific genomic rearrangements or fusion events which may be clinically actionable. Therefore, a computer implemented method is useful to process the voluminous quantity of data so that results can be obtained in time to assist the patient.

SUMMARY OF THE INVENTION

[0012] The present invention relates to compositions and methods for identification of gene fusions. More specifically, the present invention relates to compositions and methods for ligating single stranded polynucleotides of unknown sequence. The resulting sequencing library, with inserts derived from the set of PCR amplified adapter-ligated DNA molecules are then subjected to computational analysis.

[0013] In one aspect, the present invention provides a method for ligating single stranded DNA molecules by:

- i) providing a first single stranded DNA molecule comprising a known sequence of at least 10 nucleotides in length;
- ii) providing a second single stranded DNA molecule comprising an unknown sequence of at least 10 nucleotides in length;
- iii) providing a single stranded RNA molecule comprising a 5' end substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end comprising a random sequence of at least 10 ribonucleotides in length;
- iv) providing a DNA ligase; and
- v) combining the first single stranded DNA molecule, the second single stranded DNA molecule, the single stranded RNA molecule and the DNA ligase,

where the first single stranded DNA molecule and the second single stranded DNA molecule are ligated to provide a ligated single stranded nucleic acid molecule when the unknown sequence of the first single stranded DNA molecule is substantially complementary to the random sequence of the single stranded RNA molecule.

[0014] In some embodiments, the first single stranded DNA molecule is blocked at the 3' end.

[0015] In some embodiments, the first single stranded DNA molecule further includes a spacer sequence adjacent to the 3' end of the known sequence.

[0016] In some embodiments, the known sequence of the first single stranded DNA molecule is from ten to about 100 nucleotides in length.

[0017] In some embodiments, the unknown sequence of the second single stranded DNA molecule is from ten to about 10,000 nucleotides in length.

[0018] In some embodiments, the single stranded RNA splint molecule is about 20 nucleotides in length and the 10 nucleotides at the 5' end are substantially complementary to the known sequence of the first single stranded DNA molecule and the random sequence is substantially complementary to the unknown sequence of second single stranded DNA molecule.

[0019] In some embodiments, the first single stranded DNA molecule, the second single stranded DNA molecule, and the single stranded RNA molecule are combined under conditions suitable for annealing of the first single stranded DNA molecule and the second single stranded DNA molecule to the single stranded RNA molecule.

[0020] In some embodiments, the method further comprises use of an antiprimer molecule, where the first single stranded DNA molecule is combined with the single stranded RNA molecule and the second single stranded DNA molecule is combined with the antiprimer single stranded DNA molecule prior to combination of the first single stranded DNA molecule and the single stranded RNA molecule duplex and the second single stranded DNA molecule and the antiprimer single stranded DNA molecule duplex.

[0021] In some embodiments, the DNA ligase is a *Chlorella* virus DNA ligase.

[0022] In some embodiments, the second single stranded DNA molecule is *in vitro* ssDNA or single stranded cDNA.

[0023] In some embodiments, the second single stranded DNA molecule further includes a second known sequence 5' to the unknown sequence.

[0024] In some embodiments, the second single stranded DNA molecule further includes a primer sequence at the 5' end.

[0025] In some embodiments, the method further comprises amplifying the ligated single stranded nucleic acid molecule.

[0026] In some embodiments, the amplifying is performed by PCR.

[0027] In some embodiments, the PCR is linear PCR or nested PCR.

[0028] In an alternative aspect, the present invention provides a kit including:

- i) a first single stranded DNA molecule including a known sequence of at least 10 nucleotides in length;
- ii) optionally, a second single stranded DNA molecule including an unknown sequence of least 10 nucleotides in length; and

iii) a single stranded RNA *splint* molecule including a 5' end substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end including a random sequence of least 10 ribonucleotides in length.

[0029] In some embodiments, the kit further includes a DNA ligase.

[0030] In some embodiments, the DNA ligase is a splint ligase.

[0031] In some embodiments, the kit further includes instructions for use.

[0032] In an alternative aspect, the present invention provides a computer implemented method for identifying from a plurality of read pairs R1 and R2 of a sequenced sample with inserts derived from a set of PCR amplified adapter ligated DNA molecules, any target gene fused with any other DNA sequence which are potentially clinically actionable, including:

generating a k-mer index of predetermined length k from a reference genome or a transcriptome;

mapping each sequencing read from the plurality of read pairs to exons of the reference genome with the k-mer index to identify corresponding exon regions of the sequencing read, and generating for each sequencing read a mapping vector including one or more blocks, where each block is either an exon block defined by

a range of consecutive k-mer matches and their corresponding locations on an exon of the reference genome, or a mismatch block defined by

a length of consecutive k-mer mis-matches;

identifying exon blocks with overlapping consecutive k-mer matches in each of the mapping vectors, and merging the same exon blocks with adjustment of mapping coordinates or retaining the exon block having predetermined high priority criteria to resolve mapping ambiguities;

combining both the mapping vectors corresponding to each read pair and sorting by genomic coordinates to provide intermediate mapping vectors, and merging any exon blocks with the same respective exons with adjustment of mapping coordinates to generate merged mapping vectors;

identifying fusion junction candidates in the merged mapping vectors by searching for a pair of consecutive exon blocks from different genes, called fusion

partner genes, with a mismatch block in between, and recording the corresponding read pairs for each of the identified fusion junction candidates;

clustering all the fusion junction candidates based on genomic coordinates of junction points of the pairs of consecutive exons from different genes where junctions closer than a predetermined number of base pairs are considered to be the same fusion junction candidate, and calculating a first predetermined set of metrics based on the read pairs including read depth;

assembling a fusion transcript for each fusion junction candidate cluster using the corresponding recorded read pairs to generate assembled transcripts, and calculating a second predetermined set of metrics from the assembled transcripts;

mapping the plurality of read pairs to the fusion transcripts, and calculating a third predetermined set of metrics based on the mapping output including a number of mapped reads; and

categorizing any fusion transcript as being a high confidence reportable fusion when

the orientation of both fusion partners is concordant,

exactly one of the partners is a gene targeted by the assay,

minimum thresholds for specific metrics of the first, second and third predetermined set of metrics are met, and

the number of mapped reads and the read depth exceed predetermined thresholds.

[0033] In some embodiments, exon blocks having less than a minimum mapped bps are filtered out.

[0034] In some embodiments, the minimum mapped bps is 24 bps.

[0035] In some embodiments, each of the mapping vectors is represented by the expression (e_n, read_st→read_end, map_st→map_end) for each exon block and (miss, miss_st→miss_end, len) for each mismatch block, where e_n is the exon unique identifier, read_st and miss_st are the starting base pair on the sequencing read, read_end and miss_end are the ending base pair on the sequencing read, map_st is the starting mapped region on the exon for read_st, map_end is the ending

mapped region on the exon for read_end, miss is a status identifier for a mis-match, and len is the bps length of the mismatch block.

[0036] In some embodiments, the method further includes error correcting the mapping vectors to remove mismatch blocks interposed between two exon blocks with the same exon.

[0037] In some embodiments, error correcting includes removing the length of consecutive k-mer mis-matches of the mismatch block interposed between a first range of consecutive k-mer matches of a first exon block and a second range of consecutive k-mer matches of a second exon block of the same exon, when a continuation of the first range by the length of consecutive k-mer is sequentially followed by a beginning of the second range.

[0038] In some embodiments, error correcting to remove a mismatch block between two consecutive exon blocks with the same exon includes determining when $\text{read_end} + \text{len} + 1$ for the first exon block = read_st for the subsequent exon block to remove the mismatch block.

[0039] In some embodiments, error correcting to remove the length of consecutive k-mer mis-matches includes determining when $\text{map_end} + \text{len} + 1$ for a first exon = map_st for a subsequent exon to remove the length of consecutive k-mer mis-matches.

[0040] In some embodiments, identifying fusion junction candidates includes executing an error correction applied on k-mer gaps g_1 , g_2 and g_k of the fusion junction candidates when $g_k > 0$, where

g_1 is a distance between map_end of the first exon block to the corresponding exon extremity,

g_2 is a distance between map_st of the second exon block to the corresponding exon extremity,

g_k is the difference between the k-mer length k and the mismatch block len ,

and the k-mer gap correction steps include

while $g_k > 0$

a. if $g_1 > 0$ then

- i. $g_k = g_k - 1$
- ii. $g_1 = g_1 - 1$
- b. else If $g_2 > 0$ then
 - i. $g_k = g_k - 1$
 - ii. $g_2 = g_2 - 1$

[0041] In some embodiments, the predetermined high priority criteria include

an exon block with a larger mapping range than the other,

an exon block where its exon comes from an assay targeted gene,

an exon block matching a list of preferred candidates, or

when both exons are from the same gene, selecting the exon with highest priority, calculated from sorting the corresponding reference genome transcripts for each exon in descending order of length.

[0042] In some embodiments, where if none of the predetermined high priority criteria is met, then remove the exon block having the largest exon gap, where an exon gap is a distance between map_end of a first exon and map_st of a subsequent exon to their respective exon extremities.

[0043] In some embodiments, the predetermined number of base pairs to cluster junction candidates is 50 or less.

[0044] In some embodiments, the method further includes determining a unique amplicon end by retrieving the last exon block mapping coordinate map_end and adding to it the length of the last mismatch block for each of the R2 reads.

[0045] In some embodiments, identifying fusion junction candidates further includes identifying normal junction candidates in the merged mapping vector by searching for pairs of consecutive exon blocks from a same gene with a mismatch block in between, and recording the corresponding read pairs for each of the identified normal junction candidates.

[0046] In some embodiments, the first predetermined set of metrics includes a number of unique amplicon ends, an estimated length of the fusion transcript, and a ratio between fusion junction candidates and normal junction candidates.

[0047] In some embodiments, the estimated length of the fusion transcript is determined by the average mapped length of all the read pairs in the corresponding junction candidate.

[0048] In some embodiments, assembling the fusion transcript includes the following steps for each fusion junction candidate cluster

- a. constructing a de-Bruijn graph with the k-mers of all cluster reads,
- b. executing bubble correction and tip removal algorithms, to correct sequencing errors,
- c. determining a number of contigs by traversing the de-Bruijn graph starting with a node that has the k-mer of the gene-specific primer that captured the fusion, and
- d. determining a longest contig as being a true assembled fusion transcript.

[0049] In some embodiments, the second predetermined set of metrics includes an assembled fusion transcript length.

[0050] In some embodiments, the third predetermined set of metrics includes the number of unique transcript ends determined from the mapping output.

[0051] In some embodiments, categorizing further includes categorizing any assembled fusion transcript as being an artifact if either the orientation of the fusion partners is not concordant or both partners are genes not targeted by the assay.

[0052] In some embodiments, categorizing further includes categorizing any assembled fusion transcript as being filtered if minimum thresholds for any one of the specific metrics is not met.

[0053] In some embodiments, categorizing further includes categorizing any assembled fusion transcript as being a supporting fusion if another assembled fusion with more reads is found on the same partner genes.

[0054] In some embodiments, categorizing further includes calculating a fusion score for each assembled fusion transcript based on any number of metrics of the first, second and third predetermined set of metrics, and the scoring function is calculated

using the expression
$$s(j) = \sum_i^n w_i \cdot \frac{m_{ij} - \mu_i}{|\mu_i|}, \text{ for } i \in \text{metrics}$$
, where m_{ij} is the value of the metric i for the assembled fusion transcript j , w_i is the weight associated with metric i , and μ_i is the *expected mean* of metric i , calculated from a control dataset with known fusions,

comparing the fusion score to a first predetermined threshold and a second predetermined threshold, and

categorizing the assembled fusion transcript as medium confidence if the fusion score is greater than the first predetermined threshold,

categorizing the assembled fusion transcript as low confidence if the fusion score is less than the first predetermined threshold but greater than the second predetermined threshold, or

categorizing the assembled fusion transcript as filtered if the fusion score is less than the second predetermined threshold, and

outputting a result where low and medium confidence are subjected to orthogonal testing.

[0055] In some embodiments, the metrics used for the scoring function include the read depth, the number of unique reads, the assembled fusion transcript length, and a ratio of read depth between fusion junction candidates and normal junction candidates with the same gene as the fusion target gene.

[0056] This summary of the invention does not necessarily describe all features of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0057] These and other features of the invention will become more apparent from the following description in which reference is made to the appended drawings wherein:

[0058] **FIGURES 1A-D** show a schematic representation of an DNA:RNA duplex formed by the the ssRNA splint, ssDNA phosphodonor (adapter) and ssDNA phosphoacceptor (amplicon) molecules. The following elements are represented: defined ribonucleotides (5' ssRNA splint), random synthesized ribonucleotides (3' ssRNA splint), defined deoxyribonucleotides (ssDNA phosphodonor and ssDNA phosphoacceptor), undefined (unknown) deoxyribonucleotides (ssDNA phosphoacceptor), C3 blocked deoxyribonucleotide (ssDNA phosphodonor), phosphorylated deoxyribonucleotide (ssDNA phosphodonor), defined deoxyribonucleotides used as primer sequence (ssDNA phosphoacceptor), defined deoxyribonucleotide (ssDNA antiprimer). **A.** The annealed phosphodonor hybrid duplex consisting of ssRNA splint and ssDNA adapter. **B.** The annealed phosphoacceptor duplex consisting of ssDNA amplicon and ssDNA antiprimer. **C.** The annealed ligation complex consisting of adapter and amplicon coordinated by RNA splint. **D.** The final ligation product after ligation by DNA ligase.

[0059] **FIGURES 2A-C** show oligonucleotide sequences used in adapter ligation. **A.** Nucleotide sequence of the ssDNA phosphodonor oligonucleotide. **B.** Ribonucleotide sequence of the ssRNA splint oligonucleotide. **C.** The ssDNA phosphodonor oligonucleotide and ssRNA splint oligonucleotide in annealed configuration showing base pairing. Randomly incorporated nucleotides are shown as italic *N*. ssDNA phosphodonor modified bases (5' phosphorylation, 3' C3 spacer) are shown as underlined nucleotides A and C

[0060] **FIGURE 3** shows a schematic representation of sequencing read orientation with respect to produced sequencing library molecules. P5 and P7 represent added Illumina sequencing platforms. Grey represents a library insert including ligated phosphodonor and phosphoacceptor molecules. The phosphoacceptor sequence range includes a central unknown sequence that can be determined by sequencing and a flanking known sequence used to orientate the sequencing data.

[0061] **FIGURE 4** is a flowchart showing the computer implemented method for identifying target genes fused with any other DNA sequence which are potentially clinically actionable, according to a present embodiment

[0062] **FIGURES 5A-B** show examples of k-mer indexing and alignment. **A.** k-mer indexing. **B.** k-mer alignment.

[0063] **FIGURE 6** shows an example of error-correction after a pseudo-alignment step according to a present embodiment;

[0064] **FIGURE 7** shows an example of merged mapping of an R1 and R2 read, according to a present embodiment;

[0065] **FIGURE 8** shows a junction signature example with $k=5$;

[0066] **FIGURE 9** shows examples of alterations on the fusion gap g in the presence of sequencing errors/SNVs;

[0067] **FIGURE 10** is a flowchart showing the fusion confidence generation method, according to a present embodiment; and.

[0068] **FIGURE 11** shows example scores for fusion calls made on a control dataset with known fusions.

[0069] **FIGURES 12A-G** show ligation induced shift of dsDNA geneblock derived linear amplicon. (A) Anti-primers; (B-G) Visualization of 1ng of components as follows: (B) Linear ssDNA template (C) With ligation and antiprimer (D) With ligation without antiprimer (E) Without ligation (no adapter:splint) (F) Without ligation (no ligase) (G) dsDNA template. The dsDNA ladder does not accurately size migration of non-dsDNA molecules.

[0070] **FIGURES 13A-E** show schematic representations of a diverse population of transcripts that can be converted into Illumina NGS libraries via ligation and amplification. (A) dsDNA template is amplified by PCR with a single gene specific primer (GSP). (B) The GSP is extended 5'→3' by linear PCR to generate a ssDNA amplicon and the randomly located amplicon 3' end is ligated to an adapter using the ligation method. (C) Successfully ligated amplicon is used as template for PCR amplification with adapter-specific primer and a second set of internal gene specific primers. (D) The internal GSP and adapter specific primers have Nextera tails that are used to add Illumina NGS library P5 and P7 sequences by PCR. (E) The final NGS library includes the entire population of initial amplicons with variable lengths and unknown 3' regions.

[0071] **FIGURES 14A-D** show that geneblock ligation product derived NGS library is amplified from by PCR. (A) Library yields from template negative ligations. (B) Library yields from template positive ligations. (C-D) Visualization of 1ng of components with Agilent Bioanalyzer High Sensitivity DNA Chip electrophoresis as

electropherogram and gel image composite (C) Specific product in ligation positive samples at 450bp. (D) Non-specific products present in ligation negative samples due to lack of template causing non-specific primer-primer interactions.

[0072] **FIGURES 15A-D** show that RNA transcript derived NGS library can be amplified from ligation product by PCR. (A) library yields from RNA transcript derived linear amplicon. (B-D) Visualization of 1ng of components with Agilent Bioanalyzer High Sensitivity DNA Chip electrophoresis as electropherogram and gel image composite. (B) Non-specific products from template negative ligation. (C) Non-specific products from ligase negative ligation. (D) Specific product in ligation positive sample.

[0073] **FIGURE 16** shows the diversity of *TUBB* gene amplicon template molecule lengths as determined by paired end read sequencing of library fragments. The template molecule length is defined as the total length of the *TUBB* transcript between the start of sequencing read 1 (R1) and the end of sequencing read 2 (R2). For each unique sequence length, the count of unique sequences of that length is also shown.

[0074] **FIGURE 17** is a flowchart showing an overview of the assay protocol. A. RNA is extracted and converted into cDNA. B. The cDNA is enriched for target gene exons by linear multiplex PCR, adapter ligated by random splint mediated ligation, and amplified for target gene exons by nested multiplex PCR. C. Libraries are generated by PCR and paired-end sequenced. D. Fusion calling pipeline flowchart. Reads from demultiplexed sample fastQ data files are filtered for quality, indexed using a *k-mer* sliding window and aligned against a reference exome reference to identify on-target reads. Candidate gene fusions are identified, evaluated, and assigned a call status. Starting at the top of the flowchart in part D of FIG. 17, the box of the first row represents an input file provided to the computing system. The boxes in the second row, and of the fourth to seventh rows represent processing steps executed by the computing system, and the boxes in the third and eighth rows represent output files provided by the computing system.

[0075] **FIGURE 18** shows the *TUBB* gene exons 3 and 4. Enrichment and amplification primers are situated within exon 4, proximal to the exon junction, to enrich and amplify the adjacent exon, which in the normal *TUBB* transcript is exon 3.

[0076] **FIGURE 19** shows a schematic representation of targeted molecule adapter ligation. The enrichment (white) and adapter (black) molecules are coordinated in a duplex by an RNA splint molecule (grey) with a random 10 nt 3' extension. The enrichment molecule is duplexed at the 5' end with a blocking oligo to prevent off-target ligation.

[0077] **FIGURE 20** shows a schematic representation of the final composition of the target amplified and indexed library molecule. The target amplification molecule is generated by amplification with target amplification and adapter primers with Nextera (Illumina) 5'-tails which are converted into UDI indexed sequencing platforms using Nextera indexing primers (Illumina). The final library molecule contains the amplification primer site, exon junction site, and adjacent unknown sequence representing the expected adjacent exon in normal reads or an unexpected exon in gene fusions.

[0078] **FIGURE 21** shows the read coverage of the *TUBB* gene transcript. Target-specific enrichment was directed by an enrichment primer. After adapter-ligation, targeted amplification was performed using tailed amplification and adapter primers. Library conversion uses Nextera tails which permits paired-end sequencing into the insert from both directions. Unique reads are established based on their adapter-read start site positions relative to the aligned transcript.

[0079] **FIGURE 22** shows on-target unique reads from three control samples in five runs: No template control (NTC), fusion-negative sample, and 18 fusion-positive sample.

[0080] **FIGURE 23** shows counts of total unique fusion reads for each of the 18 fusion-positive sample fusions from ten observations for each scale of total template input (12.5ng, 25ng, 50ng, 100ng) plotted against manufacturer provided ddPCR expected copies for each fusion.

[0081] **FIGURE 24** shows on-target unique fusion reads (median) for ten observations of the 18 fusion-positive sample with increasing input mass.

[0082] **FIGURE 25** shows on-target unique fusion supporting read counts relative to increasing template input mass for 5 FFPE cell lines, and 15 FFPE archival clinical specimens (7: <2 years old, 8: 2-5 years old).

[0083] **FIGURE 26** shows on-target unique fusion supporting read counts relative to qPCR CT value for template input mass groups of 20 FFPE samples. Samples with no detectable amplification were assigned a qPCR value of 40.

DETAILED DESCRIPTION

[0084] The present disclosure provides, in part, methods for ligating single stranded polynucleotides that may be of unknown sequence, and a method for identifying from a sequencing library of such polynucleotides any target gene with atypical exon configuration, or target gene which has been fused with any other DNA sequence such as a partner gene in a fusion configuration.

[0085] The methods described herein permit sequence overhangs of at least 10 nucleotides. Accordingly, the length of the sequence overhangs employed in the methods described herein provide greater stability to the interaction between ligation donor and acceptor molecules.

[0086] In some embodiments, the methods described herein do not require *a priori* knowledge of acceptor sequences. Accordingly, the methods described herein may be used in applications involving unknown sequences, for example, in adapter ligation to a mixed population of fragments with unknown sequence such as patient-derived cDNA fragments.

[0087] In some embodiments, the methods described herein are not limited to ligation to between a single acceptor and donor molecule. Accordingly, the methods described herein may be used in applications where multiple sequences can be ligated in a single assay, for example for ligation of a large population of unknown DNA molecules using a large population of random RNA splint molecules.

[0088] In one embodiment, there is provided a method for ligating single stranded deoxyribonucleic (DNA) molecules including:

- i) providing a first single stranded DNA molecule including a known sequence of at least 10 nucleotides in length;
- ii) providing a second single stranded DNA molecule including an unknown sequence of at least 10 nucleotides in length;

iii) providing a single stranded RNA molecule including a 5' sequence substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end including a random sequence of least 10 ribonucleotides in length;

iv) providing a DNA ligase; and

v) combining the first single stranded DNA molecule, the second single stranded DNA molecule, the single stranded RNA molecule and the DNA ligase,

where the first single stranded DNA molecule and the second single stranded DNA molecule are ligated to provide a ligated single stranded nucleic acid molecule when the unknown sequence of the first single stranded DNA molecule is substantially complementary to the random sequence of the single stranded RNA molecule.

[0089] As used herein, the first single stranded DNA molecule may be referred to as the "adapter," "ligation phosphodonor," "ligation donor," "phosphodonor" or "donor" molecule, and include a phosphate group at the 5' end. The known sequence of the first single stranded DNA molecule may be a single defined sequence, for example, a sequence that can include one or both of the following characteristics: the two bases at the 5' terminus may independently be adenine or thymine (A or T); the sequence may have a guanine or cytosine (G,C) ratio of approximately 50%. In some embodiments, the known sequence of the first single stranded DNA molecule may be 5'-ATCTGTCTCTTATACACATCTCCGAGCCCACGAGAC-3' (SEQ ID NO: 1). It is to be understood that multiple adapter sequences can be used in a single reaction, provided that the reaction includes single stranded RNA molecules including a sequence complementary to the sequence of each adapter.

[0090] In general, the known sequence of the first single stranded DNA molecule is of a length sufficient to allow annealing to the substantially complementary 5' sequence of the single stranded RNA molecule to, for example, form a stable duplex. It is to be understood that amount of overlap between the molecules determines the stability and specificity of the ligation duplex, as known in the art. For example, shorter or longer overlaps affect the temperature at which the ligation needs to be performed as shorter overlaps require lower temperatures for stabilization of short duplexes while longer overlaps require higher temperatures to achieve acceptable specificity. It is also to be understood that melting temperatures for DNA:RNA hybrid duplexes are more stable for a given length of complementarity than DNA:DNA. By

“substantially complementary” is meant a sequence that is capable of annealing/hybridization under conditions, for example, conditions as described herein or known in the art to form a stable duplex. A “stable duplex” is a molecule that has at least a 4 base pair overlap between the single stranded RNA molecule and each of the first and second single stranded DNA molecules. In some embodiments, a “stable duplex” has at least a 10 base pair overlap between the single stranded RNA molecule and each of the first and second single stranded DNA molecules *i.e.*, for a total of 20 base pair overlap.

[0091] In some embodiments, the known sequence of the first single stranded DNA molecule is of a length suitable for use as a primer.

[0092] The known sequence of the first single stranded DNA molecule may be at least 10 nucleotides in length. In some embodiments, the known sequence of the first single stranded DNA molecule may be 10 nucleotides to about 100 nucleotides in length, or any value in between and inclusive of this range, for example, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or 100 nucleotides in length.

[0093] In some embodiments, the known sequence of the first single stranded DNA molecule may be blocked at the 3' end. A DNA sequence may be “blocked at the 3' end” if it includes a modification that prevents 3' extension, for example, during PCR (polymerase chain reaction). Examples of such 3' modifications include, without limitation, phosphorylation, spacers, modified bases, phosphorothioate bonds, fluorophores, such as C3 Spacer phosphoramidite, Dideoxycytidine (ddC), or 3' Inverted dT.

[0094] In some embodiments, the known sequence of the first single stranded DNA molecule may include a spacer sequence adjacent to the 3' end of the known sequence *i.e.*, between the known sequence and the 3' block. The spacer sequence may be a low GC sequence. By “low GC” sequence is meant a sequence with fewer than 50% of the constituent nucleobases being guanine or cytosine (G,C). In some embodiments, the low GC sequence may be two nucleobases, of which one or both may be adenine or thymine (A, T).

[0095] As used herein, the second single stranded DNA molecule may be referred to as the “template,” “ligation phosphoacceptor,” “ligation acceptor,” “phosphoacceptor” or “acceptor” molecule, and include a hydroxyl group at the 3' end. The unknown sequence of the second single stranded DNA molecule may be obtained from any

suitable source or sample. For example, the unknown sequence of the second single stranded DNA molecule may be any *in vitro* ssDNA or single stranded cDNA. In some embodiments, the unknown sequence of the second single stranded DNA molecule may be obtained from a patient known to have, or suspected of having a cancer, for example, of blood, lung, breast, or colorectal cancer. In some embodiments, the unknown sequence of the second single stranded DNA molecule may be obtained from a patient known to have, or suspected of having a cancer with an oncogenic structural rearrangement in a target gene, such as a kinase gene fusion. In some embodiments, the gene fusion may be, without limitation, a 5' (for example, FGFR or EGFR) or a 3' (for example, ALK, CSF1, MET, NTRK, RET, or ROS1) receptor tyrosine kinase (RTK) gene fusion; a BRAF or RAF1 gene fusion encoding a serine/threonine receptor kinase; a gene fusion involving the ligand NRG1; a gene fusion driven cancer including prostate (ERG), bladder (FGFR3), lung (ALK, ROS1, NRG1), thyroid (RET), or bile duct (FGFR2); or a NTRK gene fusion. In some embodiments, the unknown sequence of the second single stranded DNA molecule may be obtained from a patient known to have, or suspected of having, a cancer that is a solid tumour. In some embodiments, the cancer may be one where there are therapeutic drugs available or in development and the methods of described herein may be used to determine appropriate therapies for the patient. In some embodiments, the therapeutic drug may be, without limitation, a tyrosine kinase inhibitor (TKI) for the treatment of 5' (FGFR or EGFR) and 3' (ALK, CSF1, MET, NTRK, RET, ROS1) receptor tyrosine kinase (RTK) gene fusions, a MEK inhibitor for the treatment of BRAF or RAF1 gene fusions encoding serine/threonine receptor kinases, *etc.* In some embodiments, the patient may be a human.

[0096] As used herein, a "target gene" is a gene known to have, or suspected of having, a structural rearrangement of chromosomes which disrupt the linear order of the chromosome resulting in fusion of normally non-adjacent regions within or between chromosomes that can result in gene fusions at the DNA level that are transcribed into RNA transcripts and translated into a chimeric protein with oncogenic properties. Accordingly, a target gene may include an "oncogenic fusion" which involves fusion of a "driver gene" which provides oncogenic function with any of a large assortment of "partner genes". In an oncogenic fusion, a variable number of exons from the driver gene (which may include a functional domain) can be linked with a variable number of exons from the partner gene or, alternatively, with no exons from the partner gene if, for example, only the promoter is required, resulting in a chimeric protein.

[0097] In some embodiments, the target gene may be, without limitation, kinase encoding genes involved in solid tumour oncogenesis or genes involved in fusions in blood cancers. In some embodiments, the target gene may be, without limitation, selected from the genes listed in **Table 1**.

[0098] In general, the unknown sequence of the second single stranded DNA molecule is of a length sufficient to allow annealing to the random 3' sequence of the single stranded RNA molecule to, for example, form a stable duplex, as described herein or known in the art.

[0099] In some embodiments, the unknown sequence of the second single stranded DNA molecule may be at least 10 nucleotides in length. In some embodiments, the unknown sequence of the second single stranded DNA molecule may be 10 nucleotides to about 10,000 nucleotides in length, or any value in between and inclusive of this range, for example, 10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, or 10000 nucleotides in length.

[00100] In some embodiments, the second single stranded DNA molecule may further include a second known sequence 5' to the unknown sequence. The second known sequence of the second single stranded DNA molecule may be obtained from a target gene, as described herein.

[00101] In some embodiments, the second single stranded DNA molecule may include a primer sequence 5' to the second known sequence. In some embodiments, the primer sequence may be provided at the 5' end of the second known sequence. In some embodiments, the second known sequence may be the primer sequence. In some embodiments, the primer sequence may be obtained from a target gene, as described herein, for example, in **Table 3 (Outer Primers)**.

[00102] In some embodiments, the second single stranded DNA molecule including the unknown sequence and the second known sequence, and the primer sequence may be referred to as an "amplicon."

[00103] Single stranded DNA molecules may be synthesized *in vitro* ssDNA oligonucleotides, cDNA molecules generated by reverse transcription of RNA, or generated by linear amplification from double stranded DNA, such as cDNA. It is to

be understood that any suitable method for generating single stranded DNA may be used.

[00104] As used herein, the single stranded RNA molecule may be referred to as the “splint” molecule. In general, the 5' sequence of the single stranded RNA molecule may be substantially complementary to the known sequence of the first single stranded DNA molecule and of a length sufficient to allow annealing to the known sequence of the first single stranded DNA molecule to, for example, form a stable duplex as described herein or known in the art.

[00105] In some embodiments, the 5' sequence of the single stranded RNA molecule may be at least 10 nucleotides in length. In some embodiments, the 5' sequence of the single stranded RNA molecule may be 10 nucleotides to about 100 nucleotides in length, or any value in between and inclusive of this range, for example, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or 100 nucleotides in length.

[00106] By “random sequence” is meant a DNA sequence that contains the DNA nucleotides in no particular order. The random sequence of the single stranded RNA molecule may be generated by any suitable means such that all possible permutations of the four nucleotides (A, T, G, C) are enabled. In some embodiments, 4^{10} different permutations of the 3' random sequence region are contemplated.

[00107] In general, the random sequence of the single stranded RNA molecule may be of a length sufficient to allow annealing to the unknown sequence of the second single stranded DNA molecule to, for example, form a stable duplex as described herein or known in the art. In some embodiments, the random sequence of the single stranded RNA molecule may be at least 10 nucleotides in length. In some embodiments, the random sequence of the single stranded RNA molecule may be 10 nucleotides to about 20 nucleotides in length or any value in between and inclusive of this range, for example, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides in length.

[00108] In some embodiments, the single stranded RNA molecule may be about 20 nucleotides in length where 10 nucleotides at the 5' end are substantially complementary to the known sequence of the first single stranded DNA molecule and 10 random nucleotides at the 3' end are substantially complementary to the unknown sequence of the second single stranded DNA molecule.

[00109] In some embodiments, the sequence of the single stranded RNA molecule may be 5'-
rGrUrCrUrCrGrUrGrGrCrUrCrGrGrArGrArUrGrUrGrUrArArGrArGrArCrArGrAr
U rNrNrNrNrNrNrNrNrN-3' (SEQ ID NO: 2). It is to be understood that the
sequence of the single stranded RNA molecule is unrestricted, as long as it has
complementarity with the phosphodonor adapter sequence, for example, of at least
ten nucleotides. The randomly incorporated ribonucleotides ('rN') are at the 3' end
and include at least 4 nucleotides.

[00110] The first and second single stranded DNA molecules may be mixed
with the single stranded RNA molecule under conditions suitable for annealing of the
first and second single stranded DNA molecules to the single stranded RNA
molecule, as described herein or known in the art.

[00111] In some embodiments, the first single stranded DNA molecule may be
mixed with the single stranded RNA molecule under conditions suitable for annealing
of the first single stranded DNA molecule to the single stranded RNA molecule, and
second single stranded DNA molecule may be mixed with an antiprimer molecule
under conditions suitable for annealing of the second single stranded DNA molecule
to the antiprimer molecule. The first single stranded DNA molecule/ single stranded
RNA molecule duplex may then be mixed with the second single stranded DNA
molecule/antiprimer molecule under conditions suitable for annealing of the second
single stranded DNA molecules to the single stranded RNA molecule. As used
herein, "antiprimer" refers to a sequence that is the reverse complement of at least a
portion of the second known sequence of the second single stranded DNA molecule.
In general, the antiprimer molecule may be of a length sufficient to allow annealing to
the second known sequence of the second single stranded DNA molecule to, for
example, form a stable duplex as described herein or known in the art.

[00112] The first and second single stranded DNA molecules, when annealed
to the single stranded RNA molecule, are positioned such that they are compatible
for end-to-end ligation by a "DNA ligase," which is defined herein as a ligase capable
of such end-to-end ligation of two single stranded DNA molecules coordinated in
positioned by a RNA splint molecule by way of hybrid duplex base-pairing, such as a
PBCV-1 DNA Ligase (*Chlorella* virus DNA Ligase or SplintR® ligase). Ligation
conditions may be determined based on, for example, the length and composition of
the DNA and RNA molecules (e.g., the first and/or second single stranded DNA

molecules, the single stranded RNA molecule) that would provide sufficient annealing. Exemplary ligation conditions are described herein.

[00113] The ligated molecule can be amplified using a primer complementary to the adapter sequence, such as the Nextera i7-R (5'-GTCTCGTGGGCTCGGAG-3'; SEQ ID NO: 3) and a primer internal to the primer used in the amplicon sequence. The internal primer sequence may be obtained from a target gene, as described herein. The internal primer sequence may further include a common 5' sequence, for example, the Nextera XT universal amplification sequence (5'-TCGTCCGCGTCAGATGTGTATAAGAGACAG-3' (SEQ ID NO: 4)). It is to be understood that any suitable platform can be used, for example, the Fluidigm CS1/CS2 platform and associated sequences. In some embodiments, the internal primer sequences may be those set forth in **Table 6**.

[00114] Accordingly, in some embodiments, the first set of target specific primers used in linear amplification to generate the phosphoacceptor molecule is termed the "outer" primer set. Once the phosphoacceptor molecules are ligated to the phosphodonor adapter, then in combination with adapter-specific primer, an "inner" primer set is used to amplify target regions internal to the initial target-specific primer sites. For the initial linear amplicon, the first primer set used is on the outside, while the second primer set used is on the inside with respect to the molecule. The outer set is used for linear amplification, the inner set is used for ligation product PCR amplification. For example:

Linear amplicon: 5'-'OUTER GSP'-'INNER GSP'-unknown-3'

Ligation product: 5'-'OUTER GSP'-'INNER GSP'-unknown-adapter-3'

PCR amplicon: 5'-'INNER GSP'-unknown-adapter-3'

[00115] A ligation complex can be generated by annealing the products of two annealing processes, where

A is the adapter phosphodonor molecule

B is the RNA splint molecule

C is the amplicon phosphoaccepter molecule

D is the antiprimer molecule

Annealing 1: $A + B \rightarrow AB$

Annealing 2: $C + D \rightarrow CD$

Annealing 3 and ligation: $AB + CD + \text{ligase} \rightarrow ABCD$,

where A and C are covalently linked by the ligase through the condensation reaction of a phosphodiester bond. B and D interact with the newly created AC molecule through non-covalent hydrogen bonding. Molecules B and D are degraded, lost, or removed, in subsequent manipulation by denaturation or purification and molecule AC is used for subsequent manipulation such as PCR.

[00116] Accordingly, the final ligation involves mixing an annealed hybrid duplex with an annealed DNA duplex, each of which have single stranded extensions (the RNA splint random sequence and the linear amplicon unknown sequence). The extensions of these two molecules can base pair to coordinate ligation by the DNA ligase, e.g., splint ligase. The splint molecule 3' ssRNA extension base pairs with 3' ssDNA extension of amplicon to form hybrid duplex positioning 3' terminal hydroxyl residue of amplicon DNA adjacent to 5' terminal phosphate residue of adapter DNA. Before phosphodiester bond is formed these two ends are approximately centrally coordinated by splinting by the RNA molecule.

[00117] Antiprimer molecules may be used to block excess outer primer from the linear amplification step from participating in the ligation and to reduce the amount of non-specific artifact generated as a result of undesired ligation of excess primer. Accordingly, the addition of a limiting amount of antiprimer can result in base pairing and formation of a duplex between antiprimer and outer primer to form dsDNA duplexes which are not ligation templates. The linear amplicon is the result of linear extension of the outer primers. For these molecules, the 5' sequence corresponding to the outer primer also participates in base pairing with the antiprimers, however the 3' linear extension remains in ssDNA configuration that is available for ligation.

[00118] In some embodiments, amplification of the ligated molecule can be used to generate sequencing libraries, as described herein.

[00119] In some embodiments, the present disclosure provides a composition comprising:

i) a first single stranded DNA molecule including a known sequence of at least 10 nucleotides in length;

ii) a second single stranded DNA molecule including an unknown sequence of least 10 nucleotides in length;

iii) a single stranded RNA molecule including a 5' sequence substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end including a random sequence of least 10 ribonucleotides in length.

[00120] In some embodiments, the present disclosure provides a composition comprising two or more of the sequences set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the composition comprises ten or more of the sequences set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the composition comprises twenty or more of the sequences set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the composition comprises thirty or more of the sequences set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the composition comprises forty or more of the sequences set out in each of **Tables 3, 5** or **6**. In some embodiments, the composition comprises all of the sequences set out in each of **Tables 3, 5** and/or **6**.

[00121] In some embodiments, the present disclosure provides a kit comprising the composition, as described herein, and a DNA ligase, such as a splint ligase, optionally with instructions for use.

[00122] In some embodiments, the present disclosure provides a kit comprising two or more of the primers set out in each of **Tables** set out in **3, 5** and/or **6**, as described herein, optionally with instructions for use. In some embodiments, the kit comprises ten or more of the primers set out in each of **Tables 3, 5** or **6**. In some embodiments, the kit comprises twenty or more of the primers set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the kit comprises thirty or more of the primers set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the kit comprises forty or more of the primers set out in each of **Tables 3, 5** and/or **6**. In some embodiments, the kit comprises all of the primers set out in each of **Tables 3, 5** and/or **6**.

[00123] The ligated single stranded polynucleotide (including a primer sequence) can be used in amplification methods such as PCR, for example linear PCR and/or nested PCR.

[00124] The amplified products may be analysed using computational methods, as generally described in **Figure 4**, which sets out steps in the fusions calling pipeline.

[00125] In some embodiments, the methods described herein may be useful in analysing a sample containing poor quality RNA for example, a Formalin-Fixed Paraffin-Embedded (FFPE) tissue specimen obtained from, for example, a clinical patient. In some embodiments, the methods described herein may be useful in analysing a sample with low cellularity, for example, a small biopsy or fine needle aspirate, that yield low amounts of RNA. In some embodiments, the methods described herein may be useful in analysing a sample with low tumor purity, for example, a sample with very low fusion to normal read ratios, as described herein. In some embodiments, in low tumor purity cases, macrodissection of the FFPE section after evaluation by a pathologist or deeper sequencing may be performed. In some embodiments, in addition to qPCR CT values, the methods described herein provide multiple metrics that enable contextualization of low confidence or negative fusion calls including, for example, on-target reads, unique transcript reads, total and unique fusion reads, and fusion-to-normal read ratios.

[00126] The present invention will be further illustrated in the following examples.

[00127] **EXAMPLES**

[00128] **Example 1**

[00129] ssDNA ligation template phosphoacceptor oligonucleotides were generated by linear amplification from double stranded cDNA.

[00130] This mixed population of ligation acceptor molecules with unknown 3' sequence is then mixed with an adapter molecule with a single defined sequence (ligation donor molecule). The donor and acceptor molecules are then coordinated end-to-end in a ligation compatible orientation with a splinting RNA oligo through formation of a DNA:RNA duplex depending on sequence-based complementarity (**Figure 1**).

[00131] A synthesized mixture of RNA oligos with sequence matching the reverse complement of the adaptor molecule followed by a 10nt undefined stretch of nucleotides with 4¹⁰ sequence permutations is used. The RNA oligo therefore forms

an RNA:DNA duplex with the adaptor leaving a 10nt ssRNA extension for interaction with a sequence compatible ligation acceptor molecule. This design permits any ssDNA molecule with any 10nt 3' sequence to find a RNA oligo with compatible sequence allowing for formation of a DNA:RNA duplex and ultimately ligation of the phosphoacceptor ssDNA and phosphodonor ssDNA adapter molecule ends which are coordinated in to position by the ssRNA splint for formation of the phosphodiester bond by the DNA ligase.

[00132] The adapter ligated ssDNA molecules can then be amplified by traditional PCR, converted into a library for sequencing, sequenced, and interrogated for sequence and identification of any fusion genes (**Figure 2**).

[00133] **Methods:**

[00134] **Gene and Exon Selection:** Genes were selected to be included in targeted content if they contribute to established oncogenic structural rearrangements in lung, breast, or colorectal solid tumours and if there are available therapeutic drugs available or in development. Exons for these genes were selected to be included in targeted content if they had more than one reported incidence in the literature or in large public databases (for example, The Cancer Genome Atlas Program (TCGA), and the Catalogue of Somatic Mutations in Cancer (COSMIC), *etc.*).

[00135] **Primer and Oligo Design** Three sets of multiplexed oligos were used in this assay design, in addition to a random RNA splint pool, a universal adapter DNA oligo, and a DNA oligo for universal adapter amplification. The universal adapter and amplification oligo sequences employ Nextera XT amplification sequences for compatibility with existing library amplification methods. The RNA splint was designed to be the reverse complement of the adapter oligo with the addition of a low GC spacer sequence and a random 10mer sequence at the 3' end. For the multiplexed primer pools, in general, the inner primers were designed to be as close as possible to 20bp downstream of the exon junction, and the paired outer primer designed to be as close as possible to the inner primer without overlapping. The anti-primer set represented the reverse complement sequences of all the outer primers. After design of the inner primers a Nextera XT universal amplification sequence was appended to the 5' end of the oligo to provide a universal sequence platform for library amplification steps.

[00136] More specifically, targeted exons were amplified using two sets of primers – an “outer” primer distal to the expected fusion junction and an “inner” primer proximal to the expected fusion junction. All the primers for each set were combined to form an “outer” multiplexed pool and an “inner” multiplexed pool.

[00137] The “outer” multiplexed pool was used for targeted enrichment of cDNA template by primer-directed linear amplification (there is no primer partner primer for the alternate strand and therefore amplification is linear rather than exponential). After ligation of an adapter to the target-enriched cDNA, the “inner” multiplexed pool was used with an adapter-sequence specific general amplification primer for nested amplification of adapter-ligated target-enriched cDNA. Each ligation product involving a target transcript was amplified from one direction by the target primer and from the other direction by a common primer with sequence complementarity to the adapter added to target amplicon by ligation.

[00138] While the “outer” primer included only the gene-specific sequence, the “inner” primers included a 3’ gene-specific sequence in addition to a common 5’ tail which adds the Nextera XT universal amplification sequence for universal amplification. The target amplicon ligation product therefore possesses at each flanking end, universal amplification sequences (one added by the adapter and one added by the “inner” primer) which were used for amplicon indexing and addition of Illumina sequencing p5/p7 platforms by PCR using the IDT for Illumina Nextera Unique Dual indexing primers.

[00139] To design the multiplexed oligos a custom software was developed which performs the following steps: 1. Gene, transcript, and exon annotations are imported from Ensembl database via Biomart. 2. Single Nucleotide Polymorphisms (SNPs) are imported from dbSNP. 3. An existing primer panel listing exon targets and any primers, if already designed, is imported. 4. Primers are designed for all desired exon targets using primer3 with parameters including exon identity, panel context file, proximity to exon junction, GC content, oligo length, SNP filter, and undesired characteristic filter. Parameters are progressively relaxed to achieve a minimum number of candidate primers. 5. From all the candidate primers, inner and outer pools are assembled with a semi-supervised method that minimizes undesired primer-primer interactions by way of calculating all possible Gibb’s free energy values for possible pools and selecting the pool that minimizes the total sum of these values.

[00140] **RNA Extraction:** RNA was extracted from FFPE samples using a Maxwell RSC instrument (Promega) and RNA FFPE kit (Promega) according to manufacturer instructions with elution in 40uL EB.

[00141] **RNA Quality Assessment:** RNA was assessed using the Bioanalyzer RNA Pico assay (Agilent) and a DV200 ratio representing the relative fraction of molecules > 200nt was calculated.

[00142] **cDNA Synthesis:** First strand cDNA was generated from total extracted RNA with random hexamers using the Superscript IV system (Thermofisher) according to manufacturer instructions. Second strand cDNA synthesis was generated using the NEB second strand synthesis module according to manufacturer instructions. cDNA synthesis product was purified with a 1.8X Ampure XP SPRI purification and elution in 25uL EB.

[00143] **cDNA Quality Assessment:** High quality RNA was converted with 100% efficiency to DNA resulting in a 1:1 conversion efficiency which is monitored. As a result of modification of RNA molecules and the presence of inhibitors, RNA derived from FFPE material frequently has a conversion below 10%. Converted cDNA was assessed for amplifiability with a real-time PCR based assay according to manufacturer instructions with in-house validated threshold values (0.1) and cycle threshold (CT) values.

[00144] **Target Enrichment:** The assay employed a panel consisting of 19 target genes and 2 control genes (**Table 1**) covering 150 exons (**Table 2**).

Table 1: Target Genes

Gene	Transcript	Exons Covered	Total Exons	Orientation
ALK	ENST00000389048	2,10,16-22	7	Reverse
BRAF	ENST00000646891	2,6-12,15	9	Reverse
BRAF	ENST00000496384	19	1	Reverse
CSF1	ENST00000329608.11	6-8	3	Reverse
KRAS	ENST00000256078.8	2	1	Reverse (variant)
MET	ENST00000318493.11	2-6,13-17,21	11	Reverse
NRG1	ENST00000519301.6	1	1	Reverse
NRG1	ENST00000287842.7	1-8	8	Reverse
NTRK1	ENST00000530298.5	2	1	Reverse
NTRK1	ENST00000524377.5	2,7-13	8	Reverse

NTRK2	ENST00000277120.7	10-17	8	Reverse
NTRK3	ENST00000394480.6	7,12-18	8	Reverse
PDGFB	ENST00000331163.10	2	1	Reverse
PPARG	ENST00000397010	2	1	Reverse
PPARG	ENST00000287820.10	2,7	2	Reverse
RAF1	ENST00000251849.8	2,8-10	4	Reverse
RET	ENST00000355710.8	2-4,7-13	10	Reverse
ROS1	ENST00000368508.7	31-37	7	Reverse
ERG	ENST00000398919.6	2-11	10	Reverse
EGFR	ENST00000275493.7	8,23	2	Reverse
EGFR	ENST00000275493.7	1-3,7-9,11,13,16,20-21,24-25,27	14	Forward
FGFR1	ENST00000447712.6	1-10,14,17	12	Forward
FGFR2	ENST00000457416.6	2,8-10,13,16-17	7	Forward
FGFR3	ENST00000440486.7	2-4,8-10,13-14,16-18	11	Forward
TUBB	ENST00000327892.13	4	1	Reverse (HouseKeeping)
VCP	ENST00000358901.11	4,8	2	Reverse (HouseKeeping)

[00145] Table 2: Target Gene Exons

Gene Name	Gene ID	Exon Rank	Exon	Orientation
ALK	ENSG00000171094	2	ENSE00001313023	Reverse
ALK	ENSG00000171094	10	ENSE00001255635	Reverse
ALK	ENSG00000171094	16	ENSE00001255591	Reverse
ALK	ENSG00000171094	17	ENSE00001154420	Reverse
ALK	ENSG00000171094	18	ENSE00001154417	Reverse
ALK	ENSG00000171094	19	ENSE00001154411	Reverse
ALK	ENSG00000171094	20	ENSE00001154407	Reverse
ALK	ENSG00000171094	21	ENSE00001154403	Reverse
ALK	ENSG00000171094	22	ENSE00001154399	Reverse
BRAF	ENSG00000157764	2	ENSE00003603715	Reverse
BRAF	ENSG00000157764	6	ENSE00003687908	Reverse
BRAF	ENSG00000157764	7	ENSE00003487759	Reverse

BRAF	ENSG00000157764	8	ENSE00003569635	Reverse
BRAF	ENSG00000157764	9	ENSE00003680515	Reverse
BRAF	ENSG00000157764	10	ENSE00003685923	Reverse
BRAF	ENSG00000157764	11	ENSE00003559218	Reverse
BRAF	ENSG00000157764	12	ENSE00003521664	Reverse
BRAF	ENSG00000157764	15	ENSE00003485507	Reverse
BRAF	ENSG00000157764	19	ENSE00001847804	Reverse
CSF1	ENSG00000184371	6	ENSE00001321883	Reverse
CSF1	ENSG00000184371	7	ENSE00001291595	Reverse
CSF1	ENSG00000184371	8	ENSE00001310214	Reverse
KRAS	ENSG00000133703	2	ENSE00000936617	Reverse
MET	ENSG00000105976	2	ENSE00001238915	Reverse
MET	ENSG00000105976	3	ENSE00003565669	Reverse
MET	ENSG00000105976	4	ENSE00003622384	Reverse
MET	ENSG00000105976	5	ENSE00003549605	Reverse
MET	ENSG00000105976	6	ENSE00000717728	Reverse
MET	ENSG00000105976	13	ENSE00000717811	Reverse
MET	ENSG00000105976	14	ENSE00000717833	Reverse
MET	ENSG00000105976	15	ENSE00000717861	Reverse
MET	ENSG00000105976	16	ENSE00003596200	Reverse
MET	ENSG00000105976	17	ENSE00003663921	Reverse
MET	ENSG00000105976	21	ENSE00001898661	Reverse
NRG1	ENSG00000157168	1	ENSE00003844941	Reverse
NRG1	ENSG00000157168	1	ENSE00003752907	Reverse
NRG1	ENSG00000157168	2	ENSE00003732872	Reverse
NRG1	ENSG00000157168	3	ENSE00003749649	Reverse
NRG1	ENSG00000157168	4	ENSE00003743434	Reverse
NRG1	ENSG00000157168	5	ENSE00003644235	Reverse
NRG1	ENSG00000157168	6	ENSE00003573549	Reverse
NRG1	ENSG00000157168	7	ENSE00003622757	Reverse
NRG1	ENSG00000157168	8	ENSE00001700249	Reverse

NTRK1	ENSG00000198400	2	ENSE00002198227	Reverse
NTRK1	ENSG00000198400	2	ENSE00003567933	Reverse
NTRK1	ENSG00000198400	7	ENSE00003593627	Reverse
NTRK1	ENSG00000198400	8	ENSE00003653319	Reverse
NTRK1	ENSG00000198400	9	ENSE00002402062	Reverse
NTRK1	ENSG00000198400	10	ENSE00003559053	Reverse
NTRK1	ENSG00000198400	11	ENSE00003538368	Reverse
NTRK1	ENSG00000198400	12	ENSE00003626699	Reverse
NTRK1	ENSG00000198400	13	ENSE00003547826	Reverse
NTRK2	ENSG00000148053	10	ENSE00001025534	Reverse
NTRK2	ENSG00000148053	11	ENSE00001264498	Reverse
NTRK2	ENSG00000148053	12	ENSE00000982960	Reverse
NTRK2	ENSG00000148053	13	ENSE00001197359	Reverse
NTRK2	ENSG00000148053	14	ENSE00001090519	Reverse
NTRK2	ENSG00000148053	15	ENSE00001090518	Reverse
NTRK2	ENSG00000148053	16	ENSE00001090522	Reverse
NTRK2	ENSG00000148053	17	ENSE00001090521	Reverse
NTRK3	ENSG00000140538	7	ENSE00002504927	Reverse
NTRK3	ENSG00000140538	12	ENSE00003313953	Reverse
NTRK3	ENSG00000140538	13	ENSE00003330291	Reverse
NTRK3	ENSG00000140538	14	ENSE00003656026	Reverse
NTRK3	ENSG00000140538	15	ENSE00001134154	Reverse
NTRK3	ENSG00000140538	16	ENSE00003541167	Reverse
NTRK3	ENSG00000140538	17	ENSE00003448289	Reverse
NTRK3	ENSG00000140538	18	ENSE00001766468	Reverse
PDGFB	ENSG00000100311	2	ENSE00003510240	Reverse
PPARG	ENSG00000132170	2	ENSE00001209410	Reverse
PPARG	ENSG00000132170	2	ENSE00003553746	Reverse
PPARG	ENSG00000132170	7	ENSE00001527068	Reverse
RAF1	ENSG00000132155	2	ENSE00003585647	Reverse
RAF1	ENSG00000132155	8	ENSE00003669646	Reverse

RAF1	ENSG00000132155	9	ENSE00003614346	Reverse
RAF1	ENSG00000132155	10	ENSE00003602339	Reverse
RET	ENSG00000165731	2	ENSE00001365997	Reverse
RET	ENSG00000165731	3	ENSE00001095923	Reverse
RET	ENSG00000165731	4	ENSE00001095946	Reverse
RET	ENSG00000165731	7	ENSE00001312192	Reverse
RET	ENSG00000165731	8	ENSE00001164816	Reverse
RET	ENSG00000165731	9	ENSE00001164807	Reverse
RET	ENSG00000165731	10	ENSE00001164798	Reverse
RET	ENSG00000165731	11	ENSE00001164787	Reverse
RET	ENSG00000165731	12	ENSE00001095944	Reverse
RET	ENSG00000165731	13	ENSE00001164770	Reverse
ROS1	ENSG00000047936	31	ENSE00000762708	Reverse
ROS1	ENSG00000047936	32	ENSE00000762706	Reverse
ROS1	ENSG00000047936	33	ENSE00000762704	Reverse
ROS1	ENSG00000047936	34	ENSE00000762702	Reverse
ROS1	ENSG00000047936	35	ENSE00000442460	Reverse
ROS1	ENSG00000047936	36	ENSE00003478828	Reverse
ROS1	ENSG00000047936	37	ENSE00000762695	Reverse
ERG	ENSG00000157554	2	ENSE00001418790	Reverse
ERG	ENSG00000157554	3	ENSE00003661463	Reverse
ERG	ENSG00000157554	4	ENSE00003712731	Reverse
ERG	ENSG00000157554	5	ENSE00003508916	Reverse
ERG	ENSG00000157554	6	ENSE00003536380	Reverse
ERG	ENSG00000157554	7	ENSE00003462848	Reverse
ERG	ENSG00000157554	8	ENSE00003486142	Reverse
ERG	ENSG00000157554	9	ENSE00003491473	Reverse
ERG	ENSG00000157554	10	ENSE00003585332	Reverse
ERG	ENSG00000157554	11	ENSE00001033342	Reverse
EGFR	ENSG00000146648	8	ENSE00001751179	Reverse
EGFR	ENSG00000146648	23	ENSE00003625684	Reverse

TUBB	ENSG00000232575.10	4	ENSE00001715725	Reverse
VCP	ENSG00000165280.16	4	ENSE00003587441	Reverse
VCP	ENSG00000165280.16	8	ENSE00001091832	Reverse
FGFR1	ENSG00000077782	1	ENSE00001527464	Forward
FGFR1	ENSG00000077782	2	ENSE00003610788	Forward
FGFR1	ENSG00000077782	3	ENSE00001316315	Forward
FGFR1	ENSG00000077782	4	ENSE00003551643	Forward
FGFR1	ENSG00000077782	5	ENSE00003614250	Forward
FGFR1	ENSG00000077782	6	ENSE00003683948	Forward
FGFR1	ENSG00000077782	7	ENSE00003501167	Forward
FGFR1	ENSG00000077782	8	ENSE00003510418	Forward
FGFR1	ENSG00000077782	9	ENSE00003528389	Forward
FGFR1	ENSG00000077782	10	ENSE00003489763	Forward
FGFR1	ENSG00000077782	14	ENSE00003576378	Forward
FGFR1	ENSG00000077782	17	ENSE00003611337	Forward
FGFR2	ENSG00000066468	2	ENSE00003567716	Forward
FGFR2	ENSG00000066468	8	ENSE00003605080	Forward
FGFR2	ENSG00000066468	9	ENSE00003629514	Forward
FGFR2	ENSG00000066468	10	ENSE00003553724	Forward
FGFR2	ENSG00000066468	13	ENSE00003525972	Forward
FGFR2	ENSG00000066468	16	ENSE00003680367	Forward
FGFR2	ENSG00000066468	17	ENSE00003790431	Forward
FGFR3	ENSG00000068078	2	ENSE00001596390	Forward
FGFR3	ENSG00000068078	3	ENSE00001164902	Forward
FGFR3	ENSG00000068078	4	ENSE00000778155	Forward
FGFR3	ENSG00000068078	8	ENSE00000778159	Forward
FGFR3	ENSG00000068078	9	ENSE00000778160	Forward
FGFR3	ENSG00000068078	10	ENSE00000843002	Forward
FGFR3	ENSG00000068078	16	ENSE00003575180	Forward
FGFR3	ENSG00000068078	17	ENSE00003518612	Forward
FGFR3	ENSG00000068078	18	ENSE00002246626	Forward

EGFR	ENSG00000146648	1	ENSE00001841347	Forward
EGFR	ENSG00000146648	2	ENSE00003541288	Forward
EGFR	ENSG00000146648	3	ENSE00001704157	Forward
EGFR	ENSG00000146648	7	ENSE00001623732	Forward
EGFR	ENSG00000146648	8	ENSE00001751179	Forward
EGFR	ENSG00000146648	9	ENSE00001084929	Forward
EGFR	ENSG00000146648	11	ENSE00001084926	Forward
EGFR	ENSG00000146648	13	ENSE00001084939	Forward
EGFR	ENSG00000146648	16	ENSE00001768076	Forward
EGFR	ENSG00000146648	20	ENSE00001601336	Forward
EGFR	ENSG00000146648	21	ENSE00001681524	Forward
EGFR	ENSG00000146648	24	ENSE00001790701	Forward
EGFR	ENSG00000146648	25	ENSE00001801208	Forward
EGFR	ENSG00000146648	27	ENSE00001795780	Forward

[00146] The 19 target genes listed in **Table 1** represent oncogenic driver genes frequently observed in solid tumour structural rearrangements that have FDA approved drugs available for treatment (referred to as actionable). The selected 61 target gene exons are those that have been reported to be the junctional driver gene exon in such fusion genes in the literature and in clinical databases. 152 primers (**Table 3**) were designed to cover these targets and are used in a multiplexed primer pool to amplify target exons from ds cDNA with anchored linear PCR using the Qiagen Multiplex PCR kit.

[00147] **Table 3: Sequences of Outer Primers presented in 5' to 3' orientation.**

SEQ ID NO:	Sequence Name	Sequence
5	CG005.O.v2.0.1	TGATCAGGGCTTCCATGAGG
6	CG005.O.v2.0.2	CTTGGGTCGTTGGGCATTCC
7	CG005.O.v2.0.3	TGCTCAGCTTGTACTCAGGG
8	CG005.O.v2.0.4	GCAGAGGTCACCACAGAGAG
9	CG005.O.v2.0.5	AGGGTCCATGTGACATTCGTC
10	CG005.O.v2.0.6	GGAAACCCCATCTTCCCAT
11	CG005.O.v2.0.7	TGGCACCTCCTGCAAAG
12	CG005.O.v2.0.8	CCAGGCTGATGGAGATATTGTCA
13	CG005.O.v2.0.9	CAGGAAAGGAAGGAGCTTTCA

SEQ ID NO:	Sequence Name	Sequence
14	CG005.O.v2.0.10	AGGAACAGAACTGTGTTTTGATGT
15	CG005.O.v2.0.11	TCTGGTGCCATCCACAAAATG
16	CG005.O.v2.0.12	GCTAAGGACGCCTCTTCCTG
17	CG005.O.v2.0.13	GGCTTGTAAGTGTGAGGTG
18	CG005.O.v2.0.14	TGCCCATCAGGAATCTCCC
19	CG005.O.v2.0.15	CTGGTCGGAAGGGCTGTG
20	CG005.O.v2.0.16	TGATGGTGGATTATGCTCCC
21	CG005.O.v2.0.17	GGGTAGCAGACAAACCTGTG
22	CG005.O.v2.0.18	CTCGCTGAGGTCCTGGAGA
23	CG005.O.v2.0.19	GCGGAAACCCTGGAAAAGC
24	CG005.O.v2.0.20	CCTGTCTGTATCCTGAGTCAG
25	CG005.O.v2.0.21	GAGACTGTTCTGTGCGTCCA
26	CG005.O.v2.0.22	CCGGGTCAGTCTAGGGATG
27	CG005.O.v2.0.23	TTGCTGATTCAGGCTTGGC
28	CG005.O.v2.0.24	ACTACTCTCCATAAATGCTACGA
29	CG005.O.v2.0.25	TCTCTTGCCGGAATGTCAGC
30	CG005.O.v2.0.26	TGTCCGGAACACAAAGACA
31	CG005.O.v2.0.27	ACCACGTACCAGATGGATGTG
32	CG005.O.v2.0.28	AGATTTTGGGCTGGCCAAAC
33	CG005.O.v2.0.29	ACCTGCAGATCATCAGAGGA
34	CG005.O.v2.0.30	CGCCGTCTTCTCCATCTCA
35	CG005.O.v2.0.31	GGCCGACAGCTATGAGATGG
36	CG005.O.v2.0.32	CCTGTGCCATCCAAACTGCA
37	CG005.O.v2.0.33	TAGTCGCCCCAAAGTTCCGTG
38	CG005.O.v2.0.34	GCTTCTTGACGCGATACAGC
39	CG005.O.v2.0.35	ATGCCGACGAGTACCTCATC
40	CG005.O.v2.0.36	AGCGATGCGACCCTCC
41	CG005.O.v2.0.37	AACTGTGAGGTGGTCCTTGG
42	CG005.O.v2.0.38	GGATGGAGGAGATCTCGCTG
43	CG005.O.v2.0.39	GAGGAACTGCCAAAGCTGGA
44	CG005.O.v2.0.40	TCTCTGTGAGTCATTTGTCTTGC
45	CG005.O.v2.0.41	AGCATGCATTAACCGTGGAGA
46	CG005.O.v2.0.42	GGGCTCATATGGTAAATCTGGC
47	CG005.O.v2.0.43	TGAGCAGCTTTCGACTGGG
48	CG005.O.v2.0.44	TTCATCCCAACGGTGTCTGG
49	CG005.O.v2.0.45	TGTTGACGTCTGGAAGGCC
50	CG005.O.v2.0.46	GGATCTAACTGAGGACGCTGG
51	CG005.O.v2.0.47	TCCGGGACAGTCTGAATCATG
52	CG005.O.v2.0.48	TCTCTGTCTTAGCCAGGTGTG
53	CG005.O.v2.0.49	GCCTCTATGCTTGCGTAACC
54	CG005.O.v2.0.50	GCCAGGACCCGAACAGAG
55	CG005.O.v2.0.51	CTCTCCTCCAGTGGGACTCC
56	CG005.O.v2.0.52	ACATCCAGTGGCTAAAGCACA
57	CG005.O.v2.0.53	GGAGTATACGTGCTTGCGCGG
58	CG005.O.v2.0.54	TGACTTCCACAGCCAGATGG

SEQ ID NO:	Sequence Name	Sequence
59	CG005.O.v2.0.55	CCTCGGAGGATGATGATGATGA
60	CG005.O.v2.0.56	ATAGCAGACTTTGGCCTCGC
61	CG005.O.v2.0.57	GAAGTGCCTCCTCTTCTGGG
62	CG005.O.v2.0.58	CATGCAGTGCCCTCACAGAG
63	CG005.O.v2.0.59	CCACACTGCGCTGGTTGA
64	CG005.O.v2.0.60	GCAACTACACCTGCATTGTGG
65	CG005.O.v2.0.61	TGACCTTCAAGGACTTGGTGTC
66	CG005.O.v2.0.62	GCAGGGGTCTCCGAGTATG
67	CG005.O.v2.0.63	TGGTCACCATGGCAACCTTG
68	CG005.O.v2.0.64	GTTCAATGTGACCGAGGCG
69	CG005.O.v2.0.65	AAGCCAGACTTCAGCAGCCA
70	CG005.O.v2.0.66	GGGATTCCCGTGGAGGA
71	CG005.O.v2.0.67	CCAGAGACCAACGTTCAAGC
72	CG005.O.v2.0.68	GACAGACGCTCCATCCTCG
73	CG005.O.v2.0.69	GTCACCTTTGAGGACGCCG
74	CG005.O.v2.0.70	CAAGAAAGGCCTGGGCTCC
75	CG005.O.v2.0.71	TGGCCAATGTCTCCGAGCTC
76	CG005.O.v2.0.72	GCTGCAGGTGCTGAATGC
77	CG005.O.v2.0.73	GCGCCTCCTCGGAGTC
78	CG005.O.v2.0.74	ACAACGTGATGAAGATCGCAG
79	CG005.O.v2.0.75	GCTCACCTTCAAGGACCTGG
80	CG005.O.v2.0.76	CCAGAGGTGTCACCCAAACC
81	CG005.O.v2.0.77	GGCCCACTTCAAGCAG
82	CG005.O.v2.0.78	TCCCTGTGGAGGAGCTCTTC
83	CG005.O.v2.0.80	CCTCCGAAATCCAAAGTCCCA
84	CG005.O.v2.0.81	AGACAACACCAGCAATCAATCC
85	CG005.O.v2.0.82	AGCCTATCCAAATGAGGAGTGTG
86	CG005.O.v2.0.83	GAGTCCCCACTAGTTAGGATGG
87	CG005.O.v2.0.84	ATTGCTCCTCTGCACCAAGG
88	CG005.O.v2.0.85	ATCCGGGACACCAGTTCAGA
89	CG005.O.v2.0.86	GGGCAGAGAGGCATTGACTG
90	CG005.O.v2.0.87	GCTGCAAAGCTGTGGTAAAC
91	CG005.O.v2.0.88	AGGATTTACAGCACAGTGA
92	CG005.O.v2.0.89	GGAGACACTGGATGGGAGTC
93	CG005.O.v2.0.90	GAGCGAGAGGACATTGGGATG
94	CG005.O.v2.0.93	TGCAGATGCCGGTTATGGTC
95	CG005.O.v2.0.94	GCACTCCCCTCCATTACAC
96	CG005.O.v2.0.95	ATGATGCCGACCACAAGGAG
97	CG005.O.v2.0.96	TCAATCCCAAGATGCTTGTAGA
98	CG005.O.v2.0.97	AAGATGGCTTGTCCCAGTGG
99	CG005.O.v2.0.98	AGGACTAGTTTGGAACCTGCAG
100	CG005.O.v2.0.99	TGGGAGACTCTGAAGACACA
101	CG005.O.v2.0.100	AATTTGCTGATCACTTTCACA
102	CG005.O.v2.0.101	GGAGACGGAGCGCTCTAC
103	CG005.O.v2.0.102	AACAGGCTGCAGTCCGAGTG

SEQ ID NO:	Sequence Name	Sequence
104	CG005.O.v2.0.103	TTCGTCCTTCTTCTCCACCG
105	CG005.O.v2.0.104	CCGTCCACATTTGTTGAGCA
106	CG005.O.v2.0.105	GCACTCAGCAAGGAAGACC
107	CG005.O.v2.0.106	GAGACCCCAAAAGGTGTTTCG
108	CG005.O.v2.0.107	GCCACGAAACGGAGACCA
109	CG005.O.v2.0.108	AGCACGTACGTTCTTCCTG
110	CG005.O.v2.0.109	GCTGCCACCCAATGTCATGA
111	CG005.O.v2.0.110	CTGCCCATCCACAGAGAAGG
112	CG005.O.v2.0.111	GTGTCTTGCCAACTTAAGCAGA
113	CG005.O.v2.0.112	ACTTCTGTTCGTGGTGTCCC
114	CG005.O.v2.0.113	GCTAGGAACACTTTTCCAAAGGC
115	CG005.O.v2.0.114	TCCCATTGGAGATGTGATGGAG
116	CG005.O.v2.0.115	CTGCGTCAGTTCCGTGGG
117	CG005.O.v2.0.116	CTCATGCTGGAGGTTGGTCA
118	CG005.O.v2.0.117	TCATTGCTGATAACGGAGGC
119	CG005.O.v2.0.118	ATCAGTGACGTCTGTGGAAGG
120	CG005.O.v2.0.119	TTTCCAAAGGCTCCCTCACC
121	CG005.O.v2.0.120	ACTCTCTGTAGTGAACCTCCGG
122	CG005.O.v2.0.121	CTCCTGCCAGAGCTGCATCC
123	CG005.O.v2.0.122	GTGTCTTCTTCTGGTTTGTGGG
124	CG005.O.v2.0.123	TCATGACGAAGAGAACCACCA
125	CG005.O.v2.0.124	TGGAGCATTTGGGAGAGCC
126	CG005.O.v2.0.125	CTCATGCTGCAGGTTGGTGA
127	CG005.O.v2.0.126	TGGTTGATGTGGTGCACTGG
128	CG005.O.v2.0.127	TGGAGATCATCAAAGGAGCGG
129	CG005.O.v2.0.128	ACCATGGTCATTTCTGTTAAAGGC
130	CG005.O.v2.0.129	TGGAGCTCCAGGGCTTGTAG
131	CG005.O.v2.0.130	CCATTACGGAGAGATCCACGG
132	CG005.O.v2.0.131	ACCATTGCTGATCGTCTTCCA
133	CG005.O.v2.0.132	CGAGTGGACAGCATCACTTCA
134	CG005.O.v2.0.133	CTGTGACCAGCCTGTTGGG
135	CG005.O.v2.0.134	GTTGGGCTCAGATTGTTGGG
136	CG005.O.v2.0.135	CTCACGAAGGGATGTGGGTG
137	CG005.O.v2.0.136	TCCAAATTCGCCTTCTCCTAGA
138	CG005.O.v2.0.137	ACTTCTCCCGCTGCTCGC
139	CG005.O.v2.0.138	TGACATGTGGGTGGTTGACC
140	CG005.O.v2.0.139	CAGCACCGAGACGATGAAGG
141	CG005.O.v2.0.140	AGTTGCAGGTGCCATAGCC
142	CG005.O.v2.0.141	GTCTCCACAACATCGCAGTG
143	CG005.O.v2.0.142	CCACACTCCTCACACTCCAG
144	CG005.O.v2.0.143	TTGGCACCAGAGGAATGCAG
145	CG005.O.v2.0.144	GGTCCACATACAGCTTCTCCC
146	CG005.O.v2.0.145	AGCTCAGCCAACCTTTGTCT
147	CG005.O.v2.0.146	TCAGATGTGCCTCCTTCAGG
148	CG005.O.v2.0.147	TCTTCTATGCCAGACAAAGGTCA

SEQ ID NO:	Sequence Name	Sequence
149	CG005.O.v2.0.148	GCAAACACTACTGCAGGATCC
150	CG005.O.v2.0.149	TCAGCTTTCTCCCACTGTATTGA
151	CG005.O.v2.0.150	ACATAAGCAGGACCTTGGCTG
152	CG005.O.v2.0.151	CTTCCCAGCAAGAGACGCAG
153	CG005.O.v2.0.152	CGCCCTCTGTGTAGTGGC
154	CG005.O.v2.0.153	TCCTCAAAGGCTTTACGAAGGT
155	CG005.O.v2.0.154	CCACTGTGTCATCAATGGGC
156	CG005v1.1.25	CTCTATTGTTGGATCATATTCGTCC

[00148] The multiplex primer mix covering these exons was made at 1uM and a final concentration of 0.1uM was used in a 50ul reaction with 11.25ul of template cDNA, and 30ul of the Qiagen Multiplex 2x mastermix. Thermalcycler program: 95C 15 minutes, 8 cycles of (94C 30 seconds, 60C 3 minutes, 72C 30 seconds), 4C hold. Enrichment product was purified with a 1.8X Ampure XP SPRI purification and elution in 20uL EB.

[00149] **Adapter Ligation:** Resulting ssDNA amplicons were used as ligation phosphoacceptor template molecules in a ligation with a ssDNA phosphodonor adapter coordinated by a splint RNA in a DNA:RNA hybrid duplex (**Figure 3**) to permit PCR-based amplification of the linear amplicons using NEB SplintR ligase. The ligation process involved three steps and oligos listed in **Table 4**.

[00150] **Table 4: Sequences of Ligation Oligos presented in 5' to 3' orientation.** '5Phos/' refers to a 5' phosphate modification. '3SpC3' refers to a 3' C3 spacer modification. 'r' indicates subsequent base is a ribonucleotide while 'N' indicates random incorporation of any of the four standard bases at the given position.

SEQ ID NO.	Sequence Name	Sequence
1	FUSION-V2-ADAPTER	/5Phos/ AT CTGTCTCTTATACACATCT CCGAGCCCACGAGAC /3SpC3/
2	FUSION-V2-SPLINT	rGrUrCrUrCrGrUrGrGrCrUrCrGrArGrArUrGrUrGrUrArUrArArGrArGrArCrArGrArU rNrNrNrNrNrNrNrNrNrN

[00151] **Reaction 1:** Annealing of ssDNA phosphodonor adapter and RNA splint (**Figure 3**): 95ul of 100uM adapter and 95ul of 100uM splint were combined with 10ul of annealing buffer (10mM Tris pH 8, 1mM EDTA, 1M NaCl). Thermalcycler program: 85C 2 minutes, 80 cycles of 30 seconds with a reduction of -1C per cycle, 4C hold.

[00152] **Reaction 2:** Annealing of ssDNA phosphoacceptor amplicon with anti-primers pool (Table 5) to prevent adapter ligation to any unextended outer primer not removed by SPRI clean-up: 19ul of enriched amplicon was combined with, 0.125uL of 2uM antiprimers (0.25pmol), 1.15uL annealing buffer, and 2.725ul of water. Thermalcycler program: 95C for 2 minutes, 70 cycles of 30 seconds with a reduction of -1C per cycle, 25C hold.

[00153] **Table 5: Sequences of Antiprimers presented in 5' to 3' orientation.**

SEQ ID NO:	Sequence Name	Sequence
157	CG005.A.v2.0.1	CCTCATGGAAGCCCTGATCA
158	CG005.A.v2.0.2	GGAATGCCCAACGACCCAAG
159	CG005.A.v2.0.3	CCCTGAGTACAAGCTGAGCA
160	CG005.A.v2.0.4	CTCTCTGTGGTGACCTCTGC
161	CG005.A.v2.0.5	GACGAATGTCACATGGACCCT
162	CG005.A.v2.0.6	ATGGGGAAGATGGGGTTTCC
163	CG005.A.v2.0.7	CTTTGCAGGAGGGTGCCA
164	CG005.A.v2.0.8	TGACAATATCTCCATCAGCCTGG
165	CG005.A.v2.0.9	TGAAAGACTCCTTCCCTTTCCTG
166	CG005.A.v2.0.10	ACATCAAAACACAGTTCTGTTCCCT
167	CG005.A.v2.0.11	CATTTTGTGGATGGCACCAGA
168	CG005.A.v2.0.12	CAGGAAGAGGCGTCCTTAGC
169	CG005.A.v2.0.13	CACCTCAGCAGTTACAAGCC
170	CG005.A.v2.0.14	GGGAGATTCTGATGGGCA
171	CG005.A.v2.0.15	CACAGCCCTTCCGACCAAG
172	CG005.A.v2.0.16	GGGAGCATAATCCACCATCA
173	CG005.A.v2.0.17	CACAGGTTTGTCTGCTACCC
174	CG005.A.v2.0.18	TCTCCAGGACCTCAGCGAG
175	CG005.A.v2.0.19	GCTTTTCCAGGGTTTCCGC
176	CG005.A.v2.0.20	CTGACTCAGGATGACAGACAGG
177	CG005.A.v2.0.21	TGGACGCACAGAACAGTCTC
178	CG005.A.v2.0.22	CATCCCTAGCAGTGACCCGG
179	CG005.A.v2.0.23	GCCAAGCCTGAATCAGCAA
180	CG005.A.v2.0.24	TCGTAGCATTTATGGAGAGTGAGT
181	CG005.A.v2.0.25	GCTGACATTCCGGCAAGAGA
182	CG005.A.v2.0.26	TGTCTTTGTGTTCCCGGACA
183	CG005.A.v2.0.27	CACATCCATCTGGTACGTGGT
184	CG005.A.v2.0.28	GTTTGGCCAGCCAAAATCT
185	CG005.A.v2.0.29	TCCTCTGATGATCTGCAGGT
186	CG005.A.v2.0.30	TGAGATGGAGGAAGACGGCG
187	CG005.A.v2.0.31	CCATCTCATAGCTGTCGGCC
188	CG005.A.v2.0.32	TGCAGTTTGGATGGCACAGG
189	CG005.A.v2.0.33	CACGGAACCTTTGGGCGACTA

SEQ ID NO:	Sequence Name	Sequence
190	CG005.A.v2.0.34	GCTGTATCGCTGCAAGAAGC
191	CG005.A.v2.0.35	GATGAGGTACTCGTCGGCAT
192	CG005.A.v2.0.36	GGAGGGTCGCATCGCT
193	CG005.A.v2.0.37	CCAAGGACCACCTCACAGTT
194	CG005.A.v2.0.38	CAGCGAGATCTCCTCCATCC
195	CG005.A.v2.0.39	TCCAGCTTTGGCAGTTCCTC
196	CG005.A.v2.0.40	GCAAGACAAATGACTCACAGAGA
197	CG005.A.v2.0.41	TCTCCACGGTTAATGCATGCT
198	CG005.A.v2.0.42	GCCAGATTTACCATATGAGCCC
199	CG005.A.v2.0.43	CCCAGTCGAAAGCTGCTCA
200	CG005.A.v2.0.44	CCAGACACCGTTGGGATGAA
201	CG005.A.v2.0.45	GGCCTTCCAGACGTCAACA
202	CG005.A.v2.0.46	CCAGCGTCCTCAGTTAGATCC
203	CG005.A.v2.0.47	CATGATTCAGACTGTCCCGGA
204	CG005.A.v2.0.48	CACACCTGGCTAAGACAGAGA
205	CG005.A.v2.0.49	GGTTACGCAAGCATAGAGGC
206	CG005.A.v2.0.50	CTCTGTTCCGGGTCCTGGC
207	CG005.A.v2.0.51	GGAGTCCCCTGGAGGAGAG
208	CG005.A.v2.0.52	TGTGCTTTAGCCACTGGATGT
209	CG005.A.v2.0.53	CCGCCAAGCACGTATACTCC
210	CG005.A.v2.0.54	CCATCTGGCTGTGGAAGTCA
211	CG005.A.v2.0.55	TCATCATCATCATCTCCGAGG
212	CG005.A.v2.0.56	GCGAGGCCAAAGTCTGCTAT
213	CG005.A.v2.0.57	CCCAGAAGAGGAGGCACTTC
214	CG005.A.v2.0.58	CTCTGTGAGGGCACTGCATG
215	CG005.A.v2.0.59	TCAACCAGCGCAGTGTGG
216	CG005.A.v2.0.60	CCACAATGCAGGTGTAGTTGC
217	CG005.A.v2.0.61	GACACCAAGTCCTTGAAGGTCA
218	CG005.A.v2.0.62	CATACTCGGAGACCCCTGC
219	CG005.A.v2.0.63	CAAGGTTGCCATGGTGACCA
220	CG005.A.v2.0.64	CGCCTCGGTCACATTGAAC
221	CG005.A.v2.0.65	TGGCTGCTGAAGTCTGGCTT
222	CG005.A.v2.0.66	TCCTCCACGGGAATCCC
223	CG005.A.v2.0.67	GCTTGAACGTTGGTCTCTGG
224	CG005.A.v2.0.68	CGAGGATGGAGCGTCTGTC
225	CG005.A.v2.0.69	CGGCGTCCTCAAAGGTGAC
226	CG005.A.v2.0.70	GGAGCCCAGGCCTTTCTTG
227	CG005.A.v2.0.71	GAGCTCGGAGACATTGGCCA
228	CG005.A.v2.0.72	GCATTACGACCTGCAGC
229	CG005.A.v2.0.73	GACTCCGAGGAGGCGC
230	CG005.A.v2.0.74	CTGCGATCTTCATCACGTTGT
231	CG005.A.v2.0.75	CCAGGTCCTTGAAGGTGAGC
232	CG005.A.v2.0.76	GGTTTGGGTGACACCTCTGG
233	CG005.A.v2.0.77	CTGCTTGAAGGTGGGCC
234	CG005.A.v2.0.78	GAAGAGCTCCTCCACAGGGA

SEQ ID NO:	Sequence Name	Sequence
235	CG005.A.v2.0.80	TGGGACTTTGGATTTCGGAGG
236	CG005.A.v2.0.81	GGATTGATTGCTGGTGTGTCT
237	CG005.A.v2.0.82	CACACTCCTCATTTGGATAGGCT
238	CG005.A.v2.0.83	CCATCCTAACTAGTGGGGACTC
239	CG005.A.v2.0.84	CCTTGGTGCAGAGGAGCAAT
240	CG005.A.v2.0.85	TCTGAACTGGTGTCCCGGAT
241	CG005.A.v2.0.86	CAGTCAATGCCTCTCTGCCC
242	CG005.A.v2.0.87	GTTTACCACAGCTTTGCAGC
243	CG005.A.v2.0.88	TCACTGTGCTGTGAAATCCT
244	CG005.A.v2.0.89	GACTCCCATCCAGTGTCTCC
245	CG005.A.v2.0.90	CATCCCAATGTCCTCTCGCTC
246	CG005.A.v2.0.93	GACCATAACCGGCATCTGCA
247	CG005.A.v2.0.94	GTGTGAATGGAGGGGAGTGC
248	CG005.A.v2.0.95	CTCCTTGTGGTCGGCATCAT
249	CG005.A.v2.0.96	TCTACAAGCATCTTGGGATTGA
250	CG005.A.v2.0.97	CCACTGGGACAAGCCATCTT
251	CG005.A.v2.0.98	CTGCAGGTTCCAACTAGTCCT
252	CG005.A.v2.0.99	TGTGTCTTCAGAGTCTCCCA
253	CG005.A.v2.0.100	TGTGCAAAGTGATCAGCAAATT
254	CG005.A.v2.0.101	GTAGAGCGCTCCGTCTCC
255	CG005.A.v2.0.102	CACTCGGACTGCAGCCTGTT
256	CG005.A.v2.0.103	CGGTGGAGAAGAAGGACGAA
257	CG005.A.v2.0.104	TGCTCAACAAATGTGGACGG
258	CG005.A.v2.0.105	GGTCTTCCTTGCTGAGTGC
259	CG005.A.v2.0.106	CGAAACACCTTTTGGGGTCTC
260	CG005.A.v2.0.107	TGGTCTCCGTTTCGTGGC
261	CG005.A.v2.0.108	CAGGAAGAACGTGACGTGCT
262	CG005.A.v2.0.109	TCATGACATTGGGTGGCAGC
263	CG005.A.v2.0.110	CCTTCTCTGTGGATGGGCAG
264	CG005.A.v2.0.111	TCTGCTTAAGTTGGCAAGACAC
265	CG005.A.v2.0.112	GGGACACCACGAACAGAAGT
266	CG005.A.v2.0.113	GCCTTTGGAAAAGTGTTCCTAGC
267	CG005.A.v2.0.114	CTCCATCACATCTCCAATGGGA
268	CG005.A.v2.0.115	CCCACGGAAGTACGCGAG
269	CG005.A.v2.0.116	TGACCAACCTCCAGCATGAG
270	CG005.A.v2.0.117	GCCTCCGTTATCAGCAATGA
271	CG005.A.v2.0.118	CCTTCCACAGACGTCACTGAT
272	CG005.A.v2.0.119	GGTGAGGGAGCCTTTGGAAA
273	CG005.A.v2.0.120	CCGGAAGTTCACTACAGAGAGT
274	CG005.A.v2.0.121	GGATGCAGCTCTGGCAGGAG
275	CG005.A.v2.0.122	CCCACAAACCAGAAGAAGACAC
276	CG005.A.v2.0.123	TGGTGGTTCTTTCGTATGA
277	CG005.A.v2.0.124	GGCTCTCCCAAATGCTCCA
278	CG005.A.v2.0.125	TCACCAACCTGCAGCATGAG
279	CG005.A.v2.0.126	CCACTGCACCACATCAACCA

SEQ ID NO:	Sequence Name	Sequence
280	CG005.A.v2.0.127	CCGCTCCTTTGATGATCTCCA
281	CG005.A.v2.0.128	GCCTTTAACGAAATGACCATGGT
282	CG005.A.v2.0.129	CTACAAGCCCTGGAGCTCCA
283	CG005.A.v2.0.130	CCGTGGATCTCTCCGTAATGG
284	CG005.A.v2.0.131	TGGAAGACGATCAGCAATGGT
285	CG005.A.v2.0.132	TGAAGTGATGCTGTCCACTCG
286	CG005.A.v2.0.133	CCCAACAGGCTGGTCACAG
287	CG005.A.v2.0.134	CCCAACAATCTGAGCCCCAAC
288	CG005.A.v2.0.135	CACCCACATCCCTTCGTGAG
289	CG005.A.v2.0.136	TCTAGGAGAAGGCGAATTTGGA
290	CG005.A.v2.0.137	GCGAGCAGCGGGAGAAGT
291	CG005.A.v2.0.138	GGTCAACCACCCACATGTCA
292	CG005.A.v2.0.139	CCTTCATCGTCTCGGTGCTG
293	CG005.A.v2.0.140	GGCTATGGCACCTGCAACT
294	CG005.A.v2.0.141	CACTGCGATGTTGTGGAGAC
295	CG005.A.v2.0.142	CTGGAGTGTGAGGAGTGTGG
296	CG005.A.v2.0.143	CTGCATTCCTCTGGTGCCAA
297	CG005.A.v2.0.144	GGGAGAAGCTGTATGTGGACC
298	CG005.A.v2.0.145	AGACAAAGAGTTGGCTGAGCT
299	CG005.A.v2.0.146	CCTGAAGGAGGCACATCTGA
300	CG005.A.v2.0.147	TGACCTTTGTCTGGCATAGAAGA
301	CG005.A.v2.0.148	GGATCCTGCAGTAGTGTTTGC
302	CG005.A.v2.0.149	TCAATACAGTGGGAGAAAGCTGA
303	CG005.A.v2.0.150	CAGCCAAGGTCCTGCTTATGT
304	CG005.A.v2.0.151	CTGCGTCTCTTGCTGGGAAG
305	CG005.A.v2.0.152	GCCACTACACAGAGGGCG
306	CG005.A.v2.0.153	ACCTTCGTAAAGCCTTTGAGGA
307	CG005.A.v2.0.154	GCCCATTGATGACACAGTGG
308	CG005v1.2.25	GGACGAATATGATCCAACAATAGAG

[00154] **Combination of reactions 1 and 2:** The entire reaction 2 product (23uL) is combined with 7uL of a ligation mastermix (1uL of 50uM annealed splint:adapter, 3uL of 10x NEB splintR reaction buffer, 2.25uL EB, 0.75uL NEB splintR ligase). Thermalcycler program: 25C 8 hours, 9 cycles of (6 minutes, -1C per cycle), 16C 30 minutes, 65C 2 minutes, 4C hold.

[00155] Ligation product was purified with a 1.8X Ampure XP SPRI purification and elution in 20uL EB.

[00156] **Nested Target PCR Amplification:** The ligation product was amplified using a general amplification primer i7R with sequence complementarity to the ligated adaptor molecule (Nextera i7_R: GTCTCGTGGGCTCGGAG, SEQ ID

NO: 3), and a nested gene-specific primer pool which included 150 gene specific primers (GSP) (Table 6) that are internal to the first gene specific linear amplification primer using KAPA HiFi hotstart readymix.

[00157] **Table 6: Sequences of Inner Primers presented in 5' to 3' orientation. Primers consist of a common 5' Nextera XT universal amplification sequence followed by a target exon specific GSP sequence.**

SEQ ID NO:	Sequence Name	Sequence
309	CG005.l.v2.0.1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCAGTTCGTCCTGTTTCA
310	CG005.l.v2.0.2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CACCTGGCCTTCATACACCTC
311	CG005.l.v2.0.3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGCTCTGCAGCTCCATCTG
312	CG005.l.v2.0.4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GAGAGGATCAGCGAGAGTGG
313	CG005.l.v2.0.5	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGTCTACCTCACAGTGAAGTGC
314	CG005.l.v2.0.6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCCATCCATTTGCGGGGTCAT
315	CG005.l.v2.0.8	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATGGCTCTGGATCCTTGTCC
316	CG005.l.v2.0.10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGATGCTGCCATGATGGTGG
317	CG005.l.v2.0.12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGTGGTATTGGGTGGTGTTC
318	CG005.l.v2.0.13	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTGTAGGTGCTGTACATTC
319	CG005.l.v2.0.14	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCTCCCAATCATCACTCGAGTC
320	CG005.l.v2.0.15	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAATGGATTTTGAAGGAGACGG
321	CG005.l.v2.0.16	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTCCCCACCAAATTTGTCC
322	CG005.l.v2.0.18	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCTTTACGTTAGTTAGTGAGC
323	CG005.l.v2.0.19	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCGACAAAAGTCACAAAATGCT
324	CG005.l.v2.0.20	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCTCTGGTTGCTCCAAGGGA
325	CG005.l.v2.0.21	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CACTGGCAGTTCCACCTGTC
326	CG005.l.v2.0.22	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTACAGGCAGTTGCAATCAG
327	CG005.l.v2.0.23	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCATGCCCTTTGAGAACCTAGA
328	CG005.l.v2.0.24	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTCCATCAGTGGCGATCTCC
329	CG005.l.v2.0.25	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCAGGGAATGCGTGGACAAG
330	CG005.l.v2.0.26	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTCCCACTACCTGCTCAAC
331	CG005.l.v2.0.27	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

SEQ ID NO:	Sequence Name	Sequence
		TACAGCTTTGGTGCCACCTG
332	CG005.l.v2.0.28	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTGGGTGCGGAAGAGAAAGA
333	CG005.l.v2.0.29	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AAACCGGACTGAAGGAGCTG
334	CG005.l.v2.0.30	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGCTCGGACGCACGA
335	CG005.l.v2.0.31	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTCCGCAAGTGTAAGAAGTGC
336	CG005.l.v2.0.32	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTGCACCTACGGATGCAC
337	CG005.l.v2.0.34	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGA CTGAGGACAGCATAGACG
338	CG005.l.v2.0.35	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCACAGCAGGGCTTCTTCAG
339	CG005.l.v2.0.37	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGGA AATTACCTATGTGCAGAGG
340	CG005.l.v2.0.38	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCTTGGATCCAAAGGTCATC
341	CG005.l.v2.0.39	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCGGCTACTTGTGGTCCA
342	CG005.l.v2.0.40	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTTGCTTTTGGTCAACACGG
343	CG005.l.v2.0.41	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGCTTTATCAACATCATCTGAAG
344	CG005.l.v2.0.42	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGGCCTAGTTGTAATTCTTTGCG
345	CG005.l.v2.0.43	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGACCGGTCCAGGCTGATC
346	CG005.l.v2.0.44	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCC TTTGGCCACACTGCA
347	CG005.l.v2.0.45	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCACTGCCGCACATGGTC
348	CG005.l.v2.0.46	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGGTCTTCAGTTTTGGGCACT
349	CG005.l.v2.0.47	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGTAAGCCAGCCCATCTACC
350	CG005.l.v2.0.48	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCACACTCAAACAACGACTG
351	CG005.l.v2.0.49	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCAGTGACACCACCTACTTC
352	CG005.l.v2.0.51	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATGCTAGCAGGGGTCTCTGA
353	CG005.l.v2.0.52	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGCACATCGAGGTGAATGGG
354	CG005.l.v2.0.53	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGACTCTCCCATCACTCTGC
355	CG005.l.v2.0.54	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACAAGCTGGCCAAGAGCATC
356	CG005.l.v2.0.55	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GAGGAGAAAGAAACAGATAACACC
357	CG005.l.v2.0.56	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGGGACATTCAACACATCGA
358	CG005.l.v2.0.57	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCACAGCCCACTCTGCACC
359	CG005.l.v2.0.58	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGAGACCCACCTTCAAGCA

SEQ ID NO:	Sequence Name	Sequence
360	CG005.l.v2.0.59	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCAAAGAATTCAAACCTGACCA
361	CG005.l.v2.0.60	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCAGCATCAACCACACATAC
362	CG005.l.v2.0.61	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCATGCACCTACCAGCTGGC
363	CG005.l.v2.0.63	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCCCTCCTTCAGTTTAGTTGA
364	CG005.l.v2.0.64	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGTCTGCCTGGCTCACTGTC
365	CG005.l.v2.0.65	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTGCACAAGCTGACCAAACG
366	CG005.l.v2.0.66	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAAGGACACAGAATGGATAAGC
367	CG005.l.v2.0.67	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGTTGGTAGAAGACTTGGATCGA
368	CG005.l.v2.0.69	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTTCTCATCACTCTGCGTGG
369	CG005.l.v2.0.70	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CACCGTGCACAAGATCTCCC
370	CG005.l.v2.0.71	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCTGCCGACCCCAAATGG
371	CG005.l.v2.0.72	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACTCCGGGGCCTACAGCT
372	CG005.l.v2.0.73	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCTTGGGGACGGAGCAG
373	CG005.l.v2.0.76	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGTTAACCACGACGAAC
374	CG005.l.v2.0.78	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTTCAAGCTGCTGAAGGAGG
375	CG005.l.v2.0.80	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACATATGGTCAGCCTTGTCCC
376	CG005.l.v2.0.81	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCTGTGAAATTCTGATCTGGTTG
377	CG005.l.v2.0.82	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGTA CTCTTG CATCGTAGCGA
378	CG005.l.v2.0.84	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTGCAAGCACAGCGGG
379	CG005.l.v2.0.85	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGCATTTCGGCTTTAGGGTG
380	CG005.l.v2.0.86	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCAGGACTGGAAATGTCTGC
381	CG005.l.v2.0.87	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGTTGATATTTCATCACGGCG
382	CG005.l.v2.0.88	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCATCATTGTCCAACAAAGTCCC
383	CG005.l.v2.0.89	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTCACATGAGGGGTTGATGG
384	CG005.l.v2.0.90	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCTCGGTCAGAAATTGGGAAAC
385	CG005.l.v2.0.93	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTCAGCACTCTCTTCTGGTAC
386	CG005.l.v2.0.94	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCTCCGCACATTTTACAAGATGG
387	CG005.l.v2.0.96	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGTAGTTTTGGCAGCGATCA
388	CG005.l.v2.0.97	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

SEQ ID NO:	Sequence Name	Sequence
		AGAAGTATTTGCTCCTTCTGTGGA
389	CG005.l.v2.0.98	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCGATTCCCTGGCTTTTCATCTC
390	CG005.l.v2.0.99	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TATGCTCCTTCAGTTGAGGC
391	CG005.l.v2.0.100	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACTCTCCAGAATCAGCCAGTG
392	CG005.l.v2.0.101	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGGCGAGTTTGGTCCAAGG
393	CG005.l.v2.0.102	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCTGCGGAGGCTTACTG
394	CG005.l.v2.0.103	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGTCTCCAGATGTGCTGTTAG
395	CG005.l.v2.0.104	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAGCAGCGTAGAAAGGAAGA
396	CG005.l.v2.0.105	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCCACTTGAGCACGATGTC
397	CG005.l.v2.0.107	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGACCACTCTTCACGATGGTG
398	CG005.l.v2.0.108	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCTGTTGAGGTCACTGGTG
399	CG005.l.v2.0.109	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CATGAAATGCAGGGACATGGC
400	CG005.l.v2.0.111	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AAATCCCAACACAGACGCAA
401	CG005.l.v2.0.112	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TGTCATTGCTGCAGTTCCA
402	CG005.l.v2.0.114	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCTGGCAGAGTCATCATCA
403	CG005.l.v2.0.115	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGTTGCCCTCAGCCATC
404	CG005.l.v2.0.116	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAAGTCCTTGCGTGCATTG
405	CG005.l.v2.0.118	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACTTCTGTTGCTGGTGTCCC
406	CG005.l.v2.0.119	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTCACCCAGTTCTCGCTTCA
407	CG005.l.v2.0.121	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGATGTCACAGCTGCAGTTG
408	CG005.l.v2.0.122	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTGGGTACAGTGATAGGAGG
409	CG005.l.v2.0.123	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGGACACAGGCAAAAGCAG
410	CG005.l.v2.0.124	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCGTGGCTGTCCATCCAC
411	CG005.l.v2.0.125	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCCTCCCTCTGGAAATCCTTC
412	CG005.l.v2.0.126	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTGGCTGAGTCCTCCTCACC
413	CG005.l.v2.0.129	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCAGGTTGTCTTGAATGTCTTCA
414	CG005.l.v2.0.131	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCCAAGCTCCCTGTATGTGC
415	CG005.l.v2.0.132	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACTTCACTGGCTTCTATTTCCCA
416	CG005.l.v2.0.133	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCAGATTGTTGGGGCTACTGG

SEQ ID NO:	Sequence Name	Sequence
417	CG005.l.v2.0.134	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTACTGGACAGGGCTGAAG
418	CG005.l.v2.0.135	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGTGACAGGAAGACCTTGAG
419	CG005.l.v2.0.136	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GAACCAAGTTCTTCCGAGGGA
420	CG005.l.v2.0.137	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACCTCCAGGCTGTCCG
421	CG005.l.v2.0.138	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTTCAGGACGTTGAACTCTG
422	CG005.l.v2.0.139	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GAAGGAGAAGAGGACAGCCG
423	CG005.l.v2.0.141	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGTCTTGCTGCTGGGAGAG
424	CG005.l.v2.0.143	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTTGTACTGGACGTTGATGC
425	CG005.l.v2.0.144	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCTCCAGTAAGCATCCCTC
426	CG005.l.v2.0.146	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGAATTCAATCTTCTCCTGGTCTG
427	CG005.l.v2.0.147	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTCAGTGGGATTGTAACAACC
428	CG005.l.v2.0.148	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCATCTTCCACCTTAAATTCTGGT
429	CG005.l.v2.0.150	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGCTGCATGAAGTTTTAACATGG
430	CG005.l.v2.0.151	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCGAGGGAAGGCAGGAAGA
431	CG005.l.v2.0.152	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAGTTGTTACCTGCCCCAGA
432	CG005.l.v2.0.153	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTTGCTCTCAGACTCACCA
433	CG005.l.v2.0.154	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CACATGGATACGTTTGCCG
434	CG005v1.0.4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CCCAGAGCAAGGAAGTGTTATC
435	CG005v1.0.6	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGAGAAGGCATGTTTGTGG
436	CG005v1.0.18	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GCTCGCACAGCCACG
437	CG005v1.0.22	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTCGGGAGATGACGAAGACG
438	CG005v1.0.25	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGCTGTATCGTCAAGGCAC
439	CG005v1.0.26	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGATACTGCACTTGTCTGGC
440	CG005v1.0.30	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTCCATGAGCTGTTTGGC
441	CG005v1.0.32	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTTCGTCTTCTTCTCCACCG
442	CG005v1.0.38	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTCGCCTAGCTCCCTTTTCAG
443	CG005v1.0.40	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGGCCAACACCTTGTCTTG
444	CG005v1.0.46	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTCACCTCAGCATCTCATAAAGC
445	CG005v1.0.47	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

SEQ ID NO:	Sequence Name	Sequence
		TTGAGAAAATGGCCTTGTTGT
446	CG005v1.0.52	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGGGAAGCAGTTGCAGGTG
447	CG005v1.0.57	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATAAGCACTGTCACCCCTTCC
448	CG005v1.0.66	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GATCCAGACAACCTGTTCAAACCTG
449	CG005v1.0.68	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AACCTGTGGTTGATCCTCCATC
450	CG005v1.0.82	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CTGCCCCATCCACAGAGAAGG
451	CG005v1.0.85	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GGTACATGATGCTTTCAGGAGG
452	CG005v1.0.87	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CAAAGTTGGTGGGCCAGAATG
453	CG005v1.0.90	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGCCGTCTCTTGCTGACTG
454	CG005v1.0.95	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGTAATTTGGGAATGCCTGGT
455	CG005v1.0.127	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGTCGGGCTCTGGAGGAAA
456	CG005v1.0.128	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCTCCAAAATGGCCCGAGAC
457	CG005v1.0.132	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG CGTGTCTTACCGTGACGTC
458	CG005v1.0.133	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ACTTCCAGAGGACCCAAAATGG

[00158] The nested primer mix was made at 2uM, and a final concentration of 0.1uM nested primer mix and 0.1uM Nextera i7 R primer was used in a 50ul reaction with 19uL ligation product. Thermalcycler program: 95C 15 minutes, 8 cycles of (94C 30 seconds, 60C 3 minutes, 72C 30 seconds), 4C hold. Enrichment product is purified with a 1.8X Ampure XP SPRI purification and elution in 20uL EB.

[00159] **Sequencing Library Construction:** A sequencing library is generated by the PCR-based addition of indexing and clustering NGS tails using the Nextera XT system (Illumina) and IDT for Illumina UDI adapters (Illumina) according to manufacturer instructions. 7.5ul amplification product is combined with 12.5ul KAPA HiFi PCR mix, and 5ul adapter mix in a 25ul reaction. Thermalcycler program: 95C 30 seconds, 15 cycles (95C 30 seconds, 55C 30 seconds, 72C 30 seconds), hold 4C. Library is purified with a 0.7X Ampure XP SPRI purification and elution in 30uL EB.

[00160] **Sequencing Library Normalization, Dilution, and Sequencing:** The libraries were quantitated with a qubit BR DNA kit (Thermofisher) according to manufacturer instructions and individually normalized to 20nM. 5uL of each library is

then pooled, and the pool was diluted to 4nM before denaturing and diluting according to Illumina MiSeq protocol. The final library was sequenced at 12.5pM with 5% PhiX using an Illumina MiSeq instrument with a 300 cycle cartridge.

[00161] As a demonstration, we performed ligation in the absence of individual components to demonstrate molecular shifts in size of the apparent 400bp ssDNA template generated from a pool of 78 synthetic dsDNA "geneblocks" (Table 7) with an average length of 500bp and with each construct encoding a fusion gene construct involving one of the targeted gene exons (Figure 12).

[00162] **Table 7: Sequences of Geneblocks presented in 5' to 3' orientation.**

SEQ ID NO:	Geneblock Name	Geneblock Sequence
459	GB_TCGA_EML4_941_ALK_4124	CAAATTCGAGCATCACCTTCTCCCCAGCCCTCTTCACAACCTC TCCAAATACACAGACAACTCCAGAAAGCAAGAATGCTACTCC CACCAAAAGCATAAAACGACCATCACCAGCTGAAAAGTCACAT AATTCTTGGGAAAATTTCAGATGATAGCCGTAATAAATTGTCGAA AATACCTTCAACACCCAAATTAATACCAAAAGTTACCAAACTG CAGACAAGCATAAAGATGTCATCATCAACCAAGTGACCGCCG GAAGCACCAGGAGCTGCAAGCCATGCAGATGGAGCTGCAGAG CCCTGAGTACAAGCTGAGCAAGCTCCGCACCTCGACCATCAT GACCGACTACAACCCCACTACTGCTTTGCTGGCAAGACCTCC TCCATCAGTGACCTGAAGGAGGTGCCGCGGAAAAACATCACC CTCATTGCGGGTCTGGGCCATGGCGCCTTTGGGGAGGTGTAT GAAGGCCAGGTGTCCGGAATGCCAACGA
460	GB_TCGA_STRN_421_ALK_4124	TTCCTGCAGCACGAGTGGGCCCGCTTCGAGGTGGAGAGAGCC CAGTGGGAGGTGGAGCGGGCGGAGCTGCAGGCCAGATTGC CTTCCTGCAGGGAGAAAGGAAGGGCCAAGAAAATTTGAAGAA GGATCTTGTGAGGAGGATCAAAATGTTGGAGTATGCTCTTAA CAGGAAAGAGCCAAATACCAAGTTGAAATACGGGACAGAAAT TGAATCAGGGAGATATGAAGCCTCCAAGCTATGATTCTGTGTA CCGCCGGAAGCACCAGGAGCTGCAAGCCATGCAGATGGAGCT GCAGAGCCCTGAGTACAAGCTGAGCAAGCTCCGCACCTCGAC CATCATGACCGACTACAACCCCACTACTGCTTTGCTGGCAAG ACCTCCTCCATCAGTGACCTGAAGGAGGTGCCGCGGAAAAAC ATCACCTCATTGCGGGTCTGGGCCATGGCGCCTTTGGGGAG GTGTATGAAGGCCAGGTGTCCGGAATGCCAACGA
461	GB_TCGA_CDK2_1032_ALK_4019	CACAGCTGTGGACATCTGGAGCCTGGGCTGCATCTTTGCTGA GATGGTGACTCGCCGGGCCCTATTCCCTGGAGATTCTGAGATT GACCAGCTCTTCCGGATCTTTCGGAATCTGGGGACCCAGAT GAGGTGGTGTGGCCAGGAGTTACTTCTATGCCTGATTACAAGC CAAGTTTCCCCAAGTGGGCCCGGCAAGATTTTAGTAAAGTTGT ACCTCCCCTGGATGAAGATGGACGGAGCTTGTTATCGTGTCTAC CCACCCCGGAGCCACACCTGCCACTCTCGCTGATCCTCTCTG TGGTGACCTCTGCCCTCGTGGCCGCCCTGGTCCTGGCTTTCT CCGGCATCATGATTGTGTACCGCCGGAAGCACCAGGAGCTGC AAGCCATGCAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGA GCAAGCTCCGCACCTCGACCATCATGACCGACTACAACCCCA ACTACTGCTTTGCTGGCAAGACCTCCTCCATCAG
462	GB_TCGA_	TCACAACCTCTCCAAATACACAGACAACTCCAGAAAGCAAGA

	EML4_976_ALK_3866	ATGCTACTCCCACCAAAAGCATAAAACGACCATCACCAGCTGA AAAGTCACACAATTTCATGGGAAAATTCAGATGATAGCCGTAATA AATTGTGCGAAAAACCTTCAACACCCAAATTAATACCAAAAGTT ACCAAAAATACAGACAAGCATAAAGATGTCATCATCAACCAAG CAAAAATGTCAACTCGCGAAAAAACAGCCAAGTGATGGAAGG CCACGGGGAAGTGAATATTAAGCATTATCTAAACTGCAGTCAC TGTGAGGTAGACGAATGTCACATGGACCCTGAAAGCCACAAG GTCATCTGCTTCTGTGACCACGGGACGGTGCTGGCTGAGGAT GGCGTCTCCTGCATTGTGTCACCCACCCCGGAGCCACACCTG CCACTCTCGCTGATCCTCTCTGTGGTGACCTCTGCCCTCGTGG CCGCCCTGGTCCTGGCTTCTCCGGCAT
463	GB_NCBI_K IF5B_2771_ALK_2771	AATCTTGAACAGCTCACTAAAGTGCACAAACAGTTGGTACGTG ATAATGCAGATCTCCGCTGTGAACCTCCTAAGTTGGAAAAGCG ACTTCGAGCTACAGCTGAGAGAGTGAAAGCTTTGGAATCAGCA CTGAAAGAAGCTAAAGAAAATGCATCTCGTGATCGCAAACGCT ATCAGCAAGAAGTAGATCGCATAAAGGAAGCAGTCAGGTCAAA GAATATGGCCAGAAGAGGGCATTCTGCACAGATTGTGTACCG CCGGAAGCACCAGGAGCTGCAAGCCATGCAGATGGAGCTGCA GAGCCCTGAGTACAAGCTGAGCAAGCTCCGCACCTCGACCAT CATGACCGACTACAACCCCACTACTGCTTTGCTGGCAAGACC TCCTCCATCAGTGACCTGAAGGAGGTGCCGCGGAAAAACATC ACCTCATTCGGGGTCTGGGCCATGGCGCCTTTGGGGAGGTG TATGAAGGCCAGGTGTCCGGAATGCCCAACGA
464	GB_NCBI_T FG_637_AL K_637	AACTTCGAAAATAAAGTGAATCGTTTATTGGATAGCTTGGAAACCA CCTGGAGAACCAGGACCTTCCACCAATATTCCTGAAAATGATA CTGTGGATGGTAGGGAAGAAAAGTCTGCTTCTGATTCTTCTGG AAAACAGTCTACTCAGGTTATGGCAGCAAGTATGTCTGCTTTT GATCCTTTAAAAAACCAAGATGAAATCAATAAAAAATGTTATGTC AGCGTTTGGCTTAACAGATGATCAGGTTTCAGTGATACCGCCGG AAGCACCAGGAGCTGCAAGCCATGCAGATGGAGCTGCAGAGC CCTGAGTACAAGCTGAGCAAGCTCCGCACCTCGACCATCATG ACCGACTACAACCCCACTACTGCTTTGCTGGCAAGACCTCCT CCATCAGTGACCTGAAGGAGGTGCCGCGGAAAAACATCACCC TCATTCGGGGTCTGGGCCATGGCGCCTTTGGGGAGGTGTATG AAGGCCAGGTGTCCGGAATGCCCAACGAC
465	GB_TCGA_ TMPRSS2_ 372_BRAF_ 1375	GATAACAGCAAGATGGCTTTGAACCTCAGGGTCAACCACAGCTA TTGGACCTTACTATGAAAACCATGGATACCAACCGGAAAAACCC CTATCCCGCACAGCCCACTGTGGTCCCCACTGTCTACGAGGT GCATCCGGCTCAGTACTACCCGTCCTCCCGTGCCCCAGTACGC CCCGAGGGTCTGACGCAGGCTTCCAACCCCGTCGTCTGCAC GCAGCCCCAAATCCCCATCCGGGACAGTGTCACCTCAAAAAA CACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCTCTG ATGGGCAGATTACAGTGGGACAAAGAATTGGATCTGGATCATT TGGAACAGTCTACAAGGGAAAGTGGCATGGTGATGTGGCAGT GAAAATGTTGAATGTGACAGCACCTACACCTCAGCAGTTACAA GCCTTCAAAAAATGAAGTAGGAGTACTCAGGAAAAACACGACATG TGAATATCCTACTCTTCATGGGCTATTCCACAA
466	GB_TCGA_ DSTYK_329_ BRAF_123 8	AATGATCCGCGAGCTGTGCCGGGGCTTCGGCCGCTACCGCC GCTACCTGGGACGGCTGCGACAGAACCTGCGCGAGACCCAG AAGTTCTCCGCGACATCAAGTGCTCCCAACCAACCACTTGTG TCTCCTCCCTCACGGGCGGCGGGGGCCGAGCGCGGCCCT GCAGGCGATGTGCGCGAAACCGGGCTGCAGGCGGGATCAAC CACAGGTTTGTCTGCTACCCCCCTGCCTCATTACCTGGCTCA CTAACTAAGTGAAAGCCTTACAGAAATCTCCAGGACCTCAGC GAGAAAGGAAGTCATCTTCATCCTCAGAAGACAGGAATCGAAT GAAAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATT CCTGATGGGCAGATTACAGTGGGACAAAGAATTGGATCTGGAT CATTTGGAACAGTCTACAAGGGAAAGTG
467	GB_TCGA_ MACF1_112	TCCCACAGGGGAACAGATACCCCAAGTTTCAGCAGAGACAGAA GGAATTAAGAAGGAGGTGATGGAGCACAGGCTGGTGTGGGA

	03_BRAF_1 201	CACAGTGAATGAGGTGAGCCGTGCTCTCTTAGAGCTGGTGCC CTGGAGAGCCAGAGAAGGGCTGGATAAACTTGTGTCCGATGC TAACGAGCAGTACAACTAGTCAGTGACACTATTGGACAAAGG GTGGATGAAATTGATGCTGCTATTACAGAGATCACAACAGGACT TGATTAGAGACCAAGGATTTCTGTGGTGATGGAGGATCAACCAC AGGTTTGTCTGCTACCCCCCTGCCTCATTACCTGGCTCACTA ACTAACGTGAAAGCCTTACAGAAATCTCCAGGACCTCAGCGAG AAAGGAAGTCATCTTCATCCTCAGAAGACAGGAATCGAATGAA AACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCTT GATGGGCAGATTACAGTGGGACAAAGAATTG
468	GB_TCGA_ CUX1_834_ BRAF_1041	TAAAGAGAAAAATCCGAGAATATGAACAGACACTGAAGAACCA AGCCGAAACCATAGCTCTTGAGAAGGAACAGAAAGTTACAGAAT GACTTTGCAGAAAAGGAGAGAAAAGCTGCAGGAGACACAGATG TCCACCACCTCAAAGCTGGAGGAAGCTGAGCATAAGGTTTCAG AGCCTACAAACAGCCCTGGAAAAAACTCGAACAGAAATTATTG ACCTGAAAACCAAATACGATGAAGAACTACTGCAAAGCCCCA AATTCTCACCAGTCCGTCTCCTTCAAATCCATTCCAATTCCAC AGCCCTTCCGACCAGCAGATGAAGATCATCGAAATCAATTG GCAACGAGACCGATCCTCATCAGCTCCCAATGTGCATATAAC ACAATAGAACCTGTCAATATTGATGACTTGATTAGAGACCAAG GATTTCTGTGGTGATGGAGGATCAACCACAGGTTTGTCTGCTAC CCCCCTGCCTCATTACCTGGCTCACTA
469	GB_NCBI_K IAA1549_13 47_BRAF_1 347	GCGACACCGGGTGATCCGCAGGGCACAGATGCAGATCGACAA GATCCTGGACCCACAGGCCAGCGTGCCCTCCGTGTTTCATAGA GCCAGGAAGAGCTCACGGATAAAACGTTCTCCCAAGCCTCG CCGGAAACACCAGGTCAACGGCTGTCCTGCCGACGCTGAGAA GGACCGGCTCATCACACAGACAGCGATGGCACCTACAGGAG GCCCCCGGGCGTCCACAACCTCAGCCTACATCGGATGCCGAGA CTTGATTAGAGACCAAGGATTTCTGTGGTGATGGAGGATCAACC ACAGGTTTGTCTGCTACCCCCCTGCCTCATTACCTGGCTCAC TAATAACGTGAAAGCCTTACAGAAATCTCCAGGACCTCAGCG AGAAAGGAAGTCATCTTCATCCTCAGAAGACAGGAATCGAATG AAAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTG CTGATGGGCAGATTACAGTGGGACAAAGAATTG
470	GB_NCBI_K IAA1549_16 66_BRAF_1 666	CCCTTCCCGGCCTCCCAGTACATCCCACCCAGCCGTCCATC GAGGAGGACGCCAGACCATGCACTCCCTCCTGGACGACGCC TTTGCCCTCGTGGCCCCAGCAGCCAGCCTGCCAGCACCGCA GGTGATAGGCCCGGAGTCCCACCCGGCCTGCCCGCAAAACAG CACCCCTTCCCAGGAAGAGAGGGCAGCCACCCAGTGGGGGT CCTTCTACAGCCCAGCCCAGACGGCCAACAATCCCTGCAGTA AAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCC TGATGGGCAGATTACAGTGGGACAAAGAATTGGATCTGGATCA TTTGGAACAGTCTACAAGGGAAAGTGGCATGGTGATGTGGCA GTGAAAATGTTGAATGTGACAGCACCTACACCTCAGCAGTTAC AAGCCTTCAAAAATGAAGTAGGAGTACTCAGGAAAAACACGACA TGTGAATATCCTACTCTTCATGGGCTATTCCACAAA
471	GB_TCGA_ PAPD7_257 1_RAF1_14 05	CCAGCGCTTGACTTTGTGCTGTCCAGCCCCAGCTCCTGTCT TCAGGCTCCTCGGCCCTTCTGTGTCTTCACTTTCTGGGAGTG ACGTTGATTACAGACACACCGCCCTGCACAACGCCCACTGTTTA CCAGTTCACTCTGCAAGCGCCAGCTCCTCTCATGGCCGGCTT ACCCACCGCCTTGCCAATGCCAGTGGCAAACCTCAGCCCCAC CACTTCCAGAACACTGATCATGACAACCAACAATCAGAGGCCT CGTGGACAGAGAGATTCAAGCTATTATTGGGAAATAGAAGCCA GTGAAGTGATGCTGTCCACTCGGATTGGGTCAGGCTCTTTTG AACTGTTTATAAGGGTAAATGGCACGGAGATGTTGCAGTAAAG ATCCTAAAGTTGTGACCCCAACCCAGAGCAATTCCAGGCCT TCAGGAATGAGGTGGCTGTTCTGCGCAAAACACGGCATGTGA ACATTCTGCTTTTCATGGGGTACATGACAA
472	GB_TCGA_ AGGF1_125	TAGTTATGATGAAAAATACTGGACTGTATTTTGACCACAGCACTG GTTTCTATTATGATTCTGAAAAATCAACTCTATTATGATCCTTCCA

	2_RAF1_12 49	CTGGAATTTATTACTATTGTGATGTGGAAAGTGGTCGTTATCAG TTTCATTCTCGAGTAGATTTGCAACCTTATCCGACTTCTAGCAC AAAACAAAGTAAAGATAAAAAATTGAAGAAGAAAAGAAAGATC CAGATTCTTCTGCAACAAATGAGGAAAAGGATGCAATTCGAAG TCACAGCGAATCAGCCTCACCTTCAGCCCTGTCCAGTAGCCCC AACAACTGAGCCCAACAGGCTGGTCACAGCCGAAAAACCCCC GTGCCAGCACAAAGAGAGCGGGCACCAGTATCTGGGACCCAG GAGAAAAACAAATTAGGCCTCGTGGACAGAGAGATTCAAGCT ATTATTGGGAAATAGAAGCCAGTGAAGTGATGCTGTCCACTCG GATTGGGTCAGGCTCTTTTGGA
473	GB_TCGA_ CMTM8_44 1_RAF1_38 9	ACGATGGAGGAGCCGACGCGCCCGCTCGCACACAGTCAC CACCACCGCCAGCTCCTTCGCAGAGAACTTCTCCACCAGCAG CAGCAGCTTCGCCTACGACCGGGAGTTCTCCGACCCCTGCC CGGCTTCCTCATCGTGGCCGAGATCGTTTAAGAATTGTTTAA CTGCATCAATGGAGCACATACAGGGAGCTTGGAAAGACGATCA GCAATGGTTTGGATTCAAAGATGCCGTGTTTGATGGCTCCAG CTGCATCTCTCTACAATAGTTCAGCAGTTTGGCTATCAGCGC CGGGCATCAGATGATGGCAAACACAGATCCTTCTAAGACAA GCAACACTATCCGTGTTTTCTTGCCGAACAAGCAAAGAACAGT GGTCAATGTGCGAAATGG
474	GB_TCGA_ GPRC5A_6 43_NRG1_6 17	TTTCATAACAGCATGAAGTGCCGTGGAAGTGAATAGGCGTGT CCTCTCCCTCGACCCCTCCCCCTCCTTGTCCTCTGCTCACCCC TCGCTCGTTCCCTCCCTCCGGCGAGGGCCGCTTTATAACAA CTGCTCAGAGTGCGAGGGCGGGATAGCTGTCCAAGGTCTCCC CCAGCACTGAGGAGCTCGCCTGCTGCCCTCTTGCGCGCGGGA AGCAGCACCAAGTTCACGGCCAACGCCTTGGCACTAGGCCTT GCCTCCCCGATTGAAAGAGATGAAAAGCCAGGAATCGGCTGC AGGTTCCAAACTAGTCTTCCGTGTGAAACCAAGTTCTGAATAC TCCTCTCTCAGATTCAAGTGGTCAAGAATGGGAATGAATTGA ATCGAAAAACAAACCACAAAATATCAAGATACAAAAAAGCCA GGGAAGTCAGAACTTCGCATTAACAAAGCATCACTGGCTGATT CTGGAGAGTATATGTGCAAGTGATCAGCAA
475	GB_TCGA_ SDC4_485_ NRG1_1019	CTGGATGACTTGGAAGACTCCATGATCGGGCCCTGAAGTTGTCC ATCCCTTGGTGCCCTAGATAACCATATCCCTGAGAGGGCAGG GTCTGGGAGCCAAGTCCCCACCGAACCCAAGAACTAGAGGA GAATGAGGTTATCCCCAAGAGAATCTCACCCGTTGAAGAGAGT GAGGATGTGTCCAACAAGGTGTCAATGTCCAGCACTGTGCGAG GGCAGCAACATCTTTGAGAGAACGGAGGTCTGGCAGCTACA TCTACATCCACCACTGGGACAAGCCATCTTGTAATGTGCGG AGAAGGAGAAAACCTTCTGTGTGAATGGAGGGGAGTGCTTCAT GGTGAAGAGACCTTTCAAACCCCTCGAGATACTTGTCAGTGC CCAAATGAGTTTACTGGTGATCGCTGCCAAAACACGTAATGG CCAGCTTCTACAAGGCGGAGGAGCTGTACCAGAAGAGAGTGC TGACCATAACCGGCATCTGCATCGCCCTCCT
476	GB_other_C D74_812_N RG1_1019	GGGCCATGCAGAAATGCCACCAAGTATGGCAACATGACAGAG GACCATGTGATGCACCTGCTCCAGAATGCTGACCCCTGAAG GTGTACCCGCCACTGAAGGGGAGCTTCCCGGAGAACCTGAGA CACCTTAAGAACACCATGGAGACCATAGACTGGAAGGTCTTTG AGAGCTGGATGCACCATTTGGCTCCTGTTTGAATGAGCAGGCA CTCCTTGGAGCAAAAGCCCACTGACGCTCCACCGAAAGCTAC ATCTACATCCACCACTGGGACAAGCCATCTTGTAATGTGCG GAGAAGGAGAAAACCTTCTGTGTGAATGGAGGGGAGTGCTTC ATGGTGAAAGACCTTTCAAACCCCTCGAGATACTTGTCAGT GCCCAAATGAGTTTACTGGTGATCGCTGCCAAAACACGTAAT GGCCAGCTTCTACAAGGCGGAGGAGCTGTACCAGAAGAGAGT GCTGACCATAACCGGCATCTGCATCGCCCTCCT
477	GB_TCGA_ CEL_920_N TRK1_906	GGAGGTGCCAGCGTCTCTCTGCAGACCCCTCTCCCCCTACAAC AAGGGCTCATCCGGCGAGCCATCAGCCAGAGCGGCGTGGC CCTGAGTCCCTGGGTATCCAGAAAAACCACTCTTCTGGGCC AAAAAGGTGGCTGAGAAGGTGGGTGGCCTGTGGGTGATGCC

		GCCAGGATGGCCCACTGTCTGAAGGTTACTGATCCCCGAGCC CTGACGCTGGCCTATAAGGTGCCGCTGGCAGGCCTGGAGTTC CCGGCCAGTGTGCAGCTGCACACGGCGGTGGAGATGCACCA CTGGTGCATCCCCCTTCTGTGGATGGGCAGCCGGCACCCTC TCTGCGCTGGCTCTTCAATGGCTCCGTGCTCAATGAGACCAGC TTCATCTTCACTGAGTTCCTGGAGCCGGCAGCCAATGAGACCG TGCGGCACGGGTGTCTGCGCCTCAACCAGCCCACCCACGTCA ACAACGGCAACTACACGCTGCTGGCTGCCAACCCTT
478	GB_TCGA_ SQSTM1_8 49_NTRK1_ 1251	CCCAGCCCCCTTCGGGCACCTGTCTGAGGGCTTCTCGCACAGC CGCTGGCTCCGGAAGGTGAAACACGGACACTTCGGGTGGCCA GGATGGGAAATGGGTCCACCAGGAACTGGAGCCCACGTCTCT CCTCGTGCAGGGGAGGCCCGCCCTGGCCCCACGGCAGAATC AGCTTCTGGTCCATCGGAGGATCCGAGTGTGAATTCCTGAAG AACGTTGGGGAGAGTGTGGCAGCTGCCCTTAGCCCTCTGGAC ACTAACAGCACATCTGGAGACCCGGTGGAGAAGAAGGACGAA ACACCTTTTGGGGTCTCGGTGGCTGTGGGCCTGGCCGTCTTT GCCTGCCTCTTCTTCTACGCTGCTCCTTGTGCTCAACAAAT GTGGACGGAGAAACAAGTTTGGGATCAACCGCCCGGCTGTGC TGGCTCCAGAGGATGGGCTGGCCATGTCCCTGCATTTTCATGA CATTGGGTGGCAGCTCCCTGTCCCCCACCAGGGGCAA
479	GB_TCGA_ LMNA_762 NTRK1_130 7	CCGCAAGACCCTTGACTCAGTAGCCAAGGAGCGCGCCCGCCT GCAGCTGGAGCTGAGCAAAGTGCGTGAGGAGTTTAAGGAGCT GAAAGCGCGCAATACCAAGAAGGAGGGTGACCTGATAGCTGC TCAGGCTCGGCTGAAGGACCTGGAGGGCTCTGCTGAACCTCAA GGAGGCCGCACTGAGCACTGCTCTCAGTGAGAAGCGCACGCT GGAGGGCGAGCTGCATGATCTGCGGGGCCAGGTGGCCAAGG TCTCGGTGGCTGTGGGCCTGGCCGTCTTGCCTGCCTCTTCC TTTCTACGCTGCTCCTTGTGCTCAACAAATGTGGACGGAGAAA CAAGTTTGGGATCAACCGCCCGGCTGTGCTGGCTCCAGAGGA TGGGCTGGCCATGTCCCTGCATTTTCATGACATTGGGTGGCAG CTCCCTGTCCCCCACCAGGGGCAAAGGCTCTGGGCTCCAAAG CCACATCATCGAGAACCCACAATACTTCAGTGATGCCT
480	GB_TCGA_ TPM3_892 NTRK1_141 0	GACTTGGAACGCACAGAGGAACGAGCTGAGCTGGCAGAGTCT AAGTGTCTGAGCTGGAGGAGGAGCTGAAGAATGTCACCAAC AACCTCAAGTCTCTTGAGGCTCAGGCGGAGAACTCTCAAA AAGAAGATAAATATGAGGAAGAAATCAAGATTCTTACTGATAAA CTCAAGGAGGCAGAGACCCGTGCTGAGTTTGTGAGAGATCG GTAGCCAAGCTGGAAGACAAATTGATGACCTGGAAGGCCCCG GCTGTGCTGGCTCCAGAGGATGGGCTGGCCATGTCCCTGCAT TTCATGACATTGGGTGGCAGCTCCCTGTCCCCCACCAGGGGC AAAGGCTCTGGGCTCCAAGGCCACATCATCGAGAACCCACAA TACTTCAGTGATGCCTGTGTTACCCACATCAAGCGCCGGGACA TCGTGCTCAAGTGGGAGCTGGGGGAGGGCGCCTTTGGGAAG GTCTTCTTGTGCTGAGTGCCACAACCTCCTGCCTGA
481	GB_NCBI_ MPRIIP_304 8_NTRK1_3 048	GTCTTATTGCGGGTAAAGGAATCGGAAATACAGTACCTGAAAC AGGAGATTAGCTCCCTCAAGGATGAGCTGCAGACGGCACTGC GGGACAAGAAGTACGCAAGTGACAAGTACAAAGACATCTACAC AGAGCTCAGCATCGCGAAGGCTAAGGCTGACTGTGACATCAG CAGGTTGAAGGAGCAGCTCAAGGCTGCAACGGAAGCACTGGG GGAGAAGTCCCCTGACAGTGCCACGGTGTCCGGATATGGCCC GGCTGTGCTGGCTCCAGAGGATGGGCTGGCCATGTCCCTGCA TTTCATGACATTGGGTGGCAGCTCCCTGTCCCCCACCAGGGG CAAAGGCTCTGGGCTCCAAGGCCACATCATCGAGAACCCACA ATACTTCAGTGATGCCTGTGTTACCCACATCAAGCGCCGGGAC ATCGTGCTCAAGTGGGAGCTGGGGGAGGGCGCCTTTGGGAA GGTCTTCTTGTGCTGAGTGCCACAACCTCCTGCCTGA
482	GB_NCBI_C D74_871_N TRK1_871	GCTCCAGAATGCTGACCCCTGAAGGTGTACCCGCCACTGAA GGGGAGCTTCCCGGAGAACCTGAGACACCTTAAGAACACCAT GGAGACCATAGACTGGAAGGTCTTTGAGAGCTGGATGGACCA TTGGCTCCTGTTTGAATGAGCAGGCACTCCTTGGAGCAAAAG

		CCCACTGACGCTCCACCGAAAGAGTCACTGGAAGTGGAGGAC CCGTCTTCTGGGCTGGGTGTGACCAAGCAGGATCTGGGCCCA GACACTAACAGCACATCTGGAGACCCGGTGGAGAAGAAGGAC GAAACACCTTTTGGGGTCTCGGTGGCTGTGGGCCCTGGCCGTC TTTGCCTGCCTCTTCCTTTCTACGCTGCTCCTTGTGCTCAACAA ATGTGGACGGAGAAACAAAGTTTGGGATCAACCGCCGGCTGT GCTGGCTCCAGAGGATGGGCTGGCCATGTCCCTGCATTTTCAT GACATTGGGTGGCAGCTCCCTGTCCCCACCGAGG
483	GB_TCGA_ ETV6_1283_ NTRK3_18 91	AAGCCATCCAGCCCCCGGCAGGAGAGCACACGCGTGATCCAG CTGATGCCCAGCCCCATCATGCACCCTCTGATCCTGAACCCCC GGCACTCCGTGGATTTCAAACAGTCCAGGCTCTCCGAGGACG GGCTGCATAGGGAAGGGAAGCCCATCAACCTCTCTCATCGGG AAGACCTGGCTTACATGAACCACATCATGGTCTCTGTCTCCCC GCCTGAAGAGCACGCCATGCCATTGGGAGAATAGCAGATGT GCAGCACATTAAGAGGAGAGACATCGTGTGAAGCGAGAACT GGGTGAGGGAGCCTTTGAAAGGTCTTCTGGCCGAGTGCTA CAACCTCAGCCCGACCAAGGACAAGATGCTTGTGGCTGTGAA GGCCCTGAAGGATCCCACCTGGCTGCCCGGAAGGATTTCCA GAGGGAGCCGAGCTGCTACCAACCTGCAGCATGACACAT TGTCAGTTCTATGGAGTGTGCGGCGATGGGGACCC
484	GB_TCGA_ ETV6_737_ NTRK3_170 2	TGGGCTGAAAATGAGTTTTCTTTAAGGCCAATTGACAGCAACA CGTTTGAAATGAATGGCAAAGCTCTCCTGCTGCTGACCAAAGA GGACTTTCGCTATCGATCTCCTCATTACGGTGATGTGCTCTAT GAACTCCTTCAGCATATTCTGAAGCAGAGGAAACCTCGGATTC TTTTTTCACCATTTCTCCACCCTGGAAACTCTATACACACACAG CCGGAGGTGCATCTGCATCAGAACCATGAAGAAAGTCCCGTG GCTGTCTCAGTGGTGAGGAGGACTCAGCCAGCCCACTGCAC CACATCAACCACGGCATCACACGCCCCTCGTCACTGGATGCC GGGCCCCGACACTGTGGTCATTGGCATGACTCGCATCCCTGTC ATTGAGAACCCCCAGTACTTCCGTCAGGGACACAACCTGCCACA AGCCCGGACACGTATGTGCAGCACATTAAGAGGAGAGACATCG TGCTGAAGCGAGAACTGGGTGAGGGAGCCTT
485	GB_TCGA_ TMPRSS2_ 149_ERG_5 29	GAGGGCGAGGGGCGGGGAGCGCCGCTGGAGCGCGGCAGG TCATATTGAACATTCCAGATACCTATCATTACTCGATGCTGTTG ATAACAGCAAGATGGCTTTGAACCTCAGAACTCTCCTGATGAAT GCAGTGTGGCCAAAGGCGGGAAGATGGTGGGCAGCCAGAC ACCGTTGGGATGAACCTACGGCAGCTACATGGAGGAGAAGCAC ATGCCACCCCCAAACATGACCACGAACGAGCGCAGAGTTATC GTGCCAGCAGATCCTACGCTATGGAGTACAGACCATGTGCGGG CAGTGGCTGGAGTGGGCGGTGAAAGAATATGGCCTTCCAGAC GTCAACATCTTGTATTCCAGAAC
486	GB_TCGA_ MX1_802_E RG_311	AGCCATCGCCGTCATCGGGGACCAGAGCTCGGGCAAGAGCTC CGTGTGGAGGCACTGTCAGGAGTTGCCCTTCCCAGAGGCAG CGGGATCGTGACCAGATGCCCGCTGGTGCTGAAACTGAAGAA ACTTGTGAACGAAGATAAGTGGAGAGGCAAGGTCAGTTACCA GGACTACGAGATTGAGATTTCCGGATGCTTCAGAGGTAGAAAAG GAAATTAATAAAGGAAGCCTTATCAGTTGTGAGTGAGGACCAG TCGTTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAGA CAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTT CCAAGATGAGCCACGCGTCCCTCAGCAGGATTGGCTGTCTC AACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAG CCAGGTGAATGGCTCAAGGAACTCTCCTGATGAATGCAGTGTG GCCAAAG
487	GB_TCGA_ NDRG1_28 8_ERG_311	GCCCGGACGTAAACAAACCTCGCCTGGCTCCCAGCTGGTGCT GAAGCTCGTCAGTTCACCATCCGCCCTCGGCTTCCGCGGGGGC GCTGGGCGCCAGCCTCGGCACCGTCCTTTCTTTCTCCCTC GCGTTAGGCAGGTGACAGCAGGGACATGTCTCGGGAGATGCA GGATGTAGACCTCGCTGAGGTGAAGCCTTTGGTGGAGAAAGG GGAGACCATCACCGGCCTCCTGCAAGAGTTTGATGTCAGGA AGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTGAGTGT

		GCCTACGGAACGCCACACCTGGCTAAGACAGAGATGACCGCG TCCTCCTCCAGCGACTATGGACAGACTTCCAAGATGAGCCAC GCGTCCCTCAGCAGGATTGGCTGTCTCAACCCCCAGCCAGGG TCACCATCAAATGGAATGTAACCCCTAGCCAGGTGAATGGCTC AAGGAACCTCTCCTGATGAATGCAGTGTGGCCAAAG
488	GB_TCGA_ TMPRSS2_ 149_ERG_1 23	GAGGGCGAGGGGCGGGGAGCGCCGCTGGAGCGCGCAGG TCATATTGAACATTCCAGATACCTATCATTACTCGATGCTGTTG ATAACAGCAAGATGGCTTTGAACTCAGTTATTCCAGGATCTTTG GAGACCCGAGGAAAGCCGTGTTGACCAAAAGCAAGACAAATG ACTCACAGAGAAAAAGATGGCAGAACCAAGGGCAACTAAAG CCGTCAGGTTCTGAACAGCTGGTAGATGGGCTGGCTTACTGA AGGACATGATTCACTGTCCCGGACCCAGCAGCTCATATCAA GGAAGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTGAG TGTGCCTACGGAACGCCACA
489	GB_TCGA_ PAX8_1253 PPARG_8 2	CCTTCCAACACGCCACTGGGGCGCAACCTCTCGACTCACCAG ACCTACCCCGTGGTGGCAGATCCTCACTCACCCCTTCGCCATAA AGCAGGAAACCCCCGAGGTGTCCAGTTCTAGCTCCACCCCTT CCTCTTTATCTAGCTCCGCTTTTTGGATCTGCAGCAAGTCGG CTCCGGGTCCCGCCCTTCAATGCCTTTCCCATGCTGCCTC CGTGACGGGCAGTTCACGGGCCAGGCCCTCCTCTCAGaaatg accatggtgacacagagatgccattctggccaccaactttgggatcagctcgtggtatctct cogtaatggaagaccactcccactcttgatataagccctcactactgtgactctccagc atttactccacattaogaagacattccattcacaagaacagatccagtggtgagattaca agtatgacctgaaactcaagagtaccaaagtcaatcaaagtggagcctgc
490	GB_TCGA_ CCDC6_535 RET_2368	TGGGAAGTCGGGGGCGATTGTCATCTCGCCGTTCCGCTGGA GGAGCTCACCAACCGCCTGGCCTCGCTGCAGCAAGAGAACAA GGTGCTGAAGATAGAGCTGGAGACCTACAACTGAAGTGCAA GGCACTGCAGGAGGAGAACCAGCAGCTGCGCAAAGCCAGCG TGACCATCGAGGATCCAAAGTGGGAATTCCTCGGAAGAACTT GGTTCTTGGAAAACTCTAGGAGAAGGCGAATTTGAAAAAGTG GTCAAGGCAACGGCCTTCCATCTGAAAGGCAGAGCAGGGTAC ACCACGGTGGCCGTGAAGATGCTGAAAGAGAACGCCTCCCCG AGTGAGCTGCGAGACCTGCTGTCAGAGTTCAACGCTGGAAG CAGGTCAACCACCCACATGTCATCAAATTGTATGGGGCCTGCA GCC
491	GB_TCGA_ NCOA4_101 4_RET_236 8	GTCTCTCAAAACCATTCAAATTCCTGAGCACTTGATGGCTCATG CTAGTTACAGCAAATATTGGGCCCTTCTGGAGAAGAGAGGCTG TATCTCCATGCCAGAGCAGAAGTCAGCATCCGGTATTGTAGCT GTCCCTTTCAGCGAATGGCTCCTTGAAGCAAACCTGCCAGTG GTTATCAAGCTCCTTACATACCCAGCACCGACCCCCAGGACTG GCTTACCCAAAAGCAGACCTTGGAGAACAGTCAGGAGGATCC AAAGTGGGAATTCCTCGGAAGAATTGGTTCTTGGAAAACT CTAGGAGAAGGCGAATTTGAAAAGTGGTCAAGGCAACGGCC TTCCATCTGAAAGGCAGAGCAGGGTACACCACGGTGGCCGTG AAGATGCTGAAAGAGAACGCCTCCCCGAGTGAGCTGCGAGAC CTGCTGTCAGAGTTCAACGTCTGAAGCAGGTCAACCACCCAC ATGTCATCAAATTGTATGGGGCCTGCAGCC
492	GB_TCGA_ TRIM33_21 45_RET_23 68	GCACAGCGGCCCTCAATATTCATGATGCAGCCACACCTCCAA AGACAACACTCAAACCCAGGGCATGCTGGACCCTTTCCCGTA GTATCGGTACACAACACCACAATCAACCAACGAGCCCTACTA CAGCAACTATGGCAAATGCAAACCGAGGTCCCACCCAGCCCAT CTGTTACAGCAATAGAGCTAATCCCCTCAGTTACCAATCCAGA AAACCTTCCATCGCTGCCAGATATTCACCCATACAGGAGGAT CCAAAGTGGGAATTCCTCGGAAGAATTGGTTCTTGGAAAAA CTCTAGGAGAAGGCGAATTTGAAAAAGTGGTCAAGGCAACGG CCTTCCATCTGAAAGGCAGAGCAGGGTACACCACGGTGGCCG TGAAGATGCTGAAAGAGAACGCCTCCCCGAGTGAGCTGCGAG ACCTGCTGTCAGAGTTCAACGTCTGAAGCAGGTCAACCACCC ACATGTCATCAAATTGTATGGGGCCTGCAGCC
493	GB_COSMI	GAAAGAAGTTTTACAGGCCCTAGAAGAATTGCTGTCAATTAT

	C_KIF5B_2 183_RET_2 326	GATCAGAAGTCTCAGGAAGTTGAAGACAAAACCTAAGGAATATG AATTGCTTAGTGATGAATTGAATCAGAAATCGGCAACTTTAGCG AGTATAGATGCTGAGCTTCAGAACTTAAGGAAATGACCAACC ACCAGAAAAACGAGCAGCTGAGATGATGGCATCTTTACTAAA AGACCTTGCAGAAATAGGAATTGCTGTGGGAAATCCATGGAG AACCAGGTCTCCGTGGATGCCTTCAAGATCCTGGAGGATCCAA AGTGGGAATTCCCTCGGAAGAACTTGTTCTTGAAAACTCT AGGAGAAGGCGAATTTGGAAAAGTGGTCAAGGCAACGGCCTT CCATCTGAAAGGCAGAGCAGGGTACACCACGGTGGCCGTGAA GATGCTGAAAGAGAACGCCTCCCCGAGTGAGCTGCGAGACCT GCTGTCAGAGTTCAACGTCTGAAGCAGG
494	GB_TCGA_ CEP85L_23 33_ROS1_5 976	AACAAAACAAAGAAAAGCAGAGAAGAATTGAGACCTTGAAAA GTATCTGGCTGATCTTCCAACCTAGATGATGTACAGAGTCAG AGTCTACAGCTGCAGATTCTGGAAGAAAAAATAAGAAATTTACA AGAGGCTTTGATAGATACAGAAAAAACTTGAAGAGATCAAAA AAGCAGTGTCAAGATAAAGAGACACAGTTAATATGCCAGAAAA AGAAAGAAAAGGAGTTAGTAACCTACCGTTGAGAGTACTCTTCC AACCCAAGAGGAGATTGAAAATCTTCTGCCTTCCCTCGGGAA AACTGACTCTGCGTCTCTTGCTGGGAAGTGGAGCCTTTGGAG AAGTGTATGAAGGAACAGCAGTGGACATCTTAGGAGTTGGAAG TGGAGAAATCAAAGTAGCAGTGAAGACTTTGAAGAAGGGTTCC ACAGACCAGGAGAAGATTGAATTCCTGAAGGAGGCACATCTGA TGAGCAAATTTAATCATCCCAACATT
495	GB_TCGA_ CLTC_5346 _ROS1_584 0	GAACCTCTGCAGTGGTTTTTGCAGGAAGAAAAAGAGAGTGCT TTGGAGCTTGTCTGTTTACCTGTTACGATCTTTTAAGGCCAGAT GTGCTCCTAGAACTGCATGGAGGCACAATATCATGGATTTTG CCATGCCCTATTTTCATCCAGGTCATGAAGGAGTACTTGACAAA GGTGGATAAAATTAGATGCTTCAGAACTACTGAGAAAAAGAA GAACAAGCTACAGAGACACAACCCATTGTTTATGTCTGGCATA GAAGATTAAAGATCAAAAAAGTGCCAAGGAAGGGGTGACAGT GCTTATAAACGAAGACAAAGAGTTGGCTGAGCTGCGAGGTCT GGCAGCCGGAGTAGGCCTGGCTAATGCCTGCTATGCAATACA TACTCTTCCAACCCAAGAGGAGATTGAAAATCTTCTGCTTTC CCTCGGGAAAACTGACTCTGCGTCTCTTGCTGGGAAGTGGA GCCTTTGGAGAAGTGATGAAGGAACAGC
496	GB_TCGA_ CD74_812 ROS1_5756	GGGCCCATGCAGAATGCCACCAAGTATGGCAACATGACAGAG GACCATGTGATGCACCTGCTCCAGAATGCTGACCCCTGAAG GTGTACCCGCCACTGAAGGGGAGCTTCCCGGAGAACCTGAGA CACCTTAAGAACACCATGGAGACCATAGACTGGAAGGTCTTTG AGAGCTGGATGCACCATTTGGCTCCTGTTTGAATGAGCAGGCA CTCCTTGGAGCAAAAAGCCCACTGACGCTCCACCGAAAGATGAT TTTTGGATACCAGAAACAAGTTTCATACTTACTATTATAGTTGG AATATTTCTGGTTGTTACAATCCCACTGACCTTTGTCTGGCATA GAAGATTAAAGATCAAAAAAGTGCCAAGGAAGGGGTGACAGT GCTTATAAACGAAGACAAAGAGTTGGCTGAGCTGCGAGGTCT GGCAGCCGGAGTAGGCCTGGCTAATGCCTGCTATGCAATACA TACTCTTCCAACCCAAGAGGAGATTGAAAA
497	GB_TCGA_ SLC34A2_2 107_ROS1_ 5447	CTGCCGCTGTGGATGCGCTCGCTGAAGCCCTGGGATGCCGTC GTCTCCAAGTTACCCGGCTGCTTCCAGATGCGCTGCTGCTGCT GCTGCCGCGTGTGCTGCCGCGCGTGCTGCTTGTGTGTGGCT GCCCCAAGTGCTGCCGCTGCAGCAAGTGCTGCGAGGACTTGG AGGAGGCGCAGGAGGGGCGAGGATGTCCCTGTCAAGGCTCCT GAGACCTTTGATAACATAACCATTAGCAGAGAGGCTCAGGCTG GAGTCCCAAATAAACCAGGCATTCCCAAATTAAGAGGGAG TAAAAATTCAATACAGTGGGAGAAAGCTGAAGATAATGGATGT AGAATTACATACTATATCCTTGAGATAAGAAAGAGCACTTCAAA TAATTTACAGAACCAGAAATTAAGGTGGAAGATGACATTTAATG GATCCTGCAGTAGTGTGTCACATGGAAGTCCAAAAACCTGAA AGGAATATTTCAAGTTCAGAGTAGTAGCTGC
498	GB_TCGA_	CGCCAACGCCACAACCACCGCGCACGGCCCCCTGACTCCGTC

	EGFR_345_ CHODL_47 0	CAGTATTGATCGGGAGAGCCGGAGCGAGCTCTTCGGGGAGCA GCGATGCGACCCTCCGGGACGGCCGGGGCAGCGCTCCTGGC GCTGCTGGCTGCGCTCTGCCCGGCGAGTCGGGCTCTGGAGG AAAAGAAAGGCCAAAAGGTGTGTTTTGCTGACTTCAAGCATCC CTGCTACAAAATGGCCTACTTCCATGAACTGTCCAGCCGAGTG AGCTTTCAGGAGGCACGCCTGGCTTGTGAGAGTGAGGGAGGA GTCCTCCTCAGCCTTGAGAATGAAGCAGAACAGAAGTTAATAG AGAGCATGTTGCAAAACCTGACAAAACCCGGGACAGGGATTTC TGATGGTGATTTCTGGATAGGGCTTTGGAGGAATGGAGATGG GCA
499	GB_TCGA_ EGFR_3203 _SEPT14_1 236	CTACGGGGTGACTGTTTTGGGAGTTGATGACCTTTGGATCCAAG CCATATGACGGAATCCCTGCCAGCGAGATCTCCTCCATCCTGG AGAAAGGAGAACGCCTCCCTCAGCCACCCATATGTACCATCGA TGCTACATGATCATGGTCAAGTGCTGGATGATAGACGCAGAT AGTCGCCCAAAGTTCCGTGAGTTGATCATCGAATTCCTCCAAA TGGCCCGAGACCCCCAGCGCTACCTTGTCAATTCAGCTGCAGG ACAAGTTCGAGCATCTTAAAATGATTCAACAGGAGGAGATAAG GAAGCTCGAGGAAGAGAAAAACAACCTGGAAGGAGAAATCATA GATTTTTATAAAATGAAAGCTGCCTCCGAAGCACTGCAGACTC AGCTGAGCACCGATACAAAGAAAGACAAACATCGTAAGAAATA ATAGTTTCTCTTACTATTCTGAGAGCCCTATCATTCTACATCGC AACTTCCTGTGAGATTGTCTTTGTAGC
500	GB_TCGA_ EGFR_3528 _VSTM2A_1 162	GACGACGTGGTGGATGCCGACGAGTACCTCATCCACAGCAG GGCTTCTTCAGCAGCCCTCCACGTACGGACTCCCTCCTG AGCTCTCTGAGTGCAACCAGCAACAATCCACCGTGGCTTGCA TTGATAGAAAATGGGCTGCAAAGCTGTCCCATCAAGGAAGACAG CTTCTTGACGCGATACAGCTCAGACCCACAGGCGCCTTGACT GAGGACAGCATAGACGACACCTTCTCCAGTGCCTGCAAG AGCAAATCGCCTGTAAAATCTACGGAGCGGACAGCAAAGTTGA CCCTAAACTCCAAGCACCACCCTGCACCCACTGTACTCTAATT CACTACACAAGGAGCGCCTGCTTCCGGAAGCATAAATGAAGA GGCTATCACATGCTTTGTTGATCATATTTTCTTTGGCAAAACAC TGATCTTTTATTTAAGAGAATTAACGTGAAGTGATAGAACGTT TTCTAATAGCAAGATCTATTTTTTCCCTT
501	GB_TCGA_ SEC61G_29 2_EGFR_34 5	AGCTCCAGTGCTACGTGTCCCTGGCATTITTAGGTGTGGTTGG GTAGGCAGTCATGGATCAGGTAATGCAGTTTGTGAGCCAAGT CGGCAGTTTGTAAAGGACTCCATTCCGGCTGGTTAAAAGATGCA CTAAACCTGTAGAAAAGAATTCCAGAAGATTGCCATGGCAAC AGCAATAGGATTGCTATAATGGGATTCAATTGGCTTCTTTGTGA AATTGATCCATATTCCTATTAATAACATCATTGTTTTGCCAAGG CACGAGTAACAAGCTCACGCAGTTGGGCACTTTTGAAGATCAT TTTCTCAGCCTCCAGAGGATGTTCAATAACTGTGAGGTGGTCC TTGGGAATTTGGAAATTACCTATGTGCAGAGGAATTATGATCTT TCCTTCTTAAAGACCATCCAGGAGGTGGCTGGTTATGTCTCA TTGCCCTCAACACAGTGGAGCGAATTCCCTTTGAAAAACCTGCA GATCATCAGAGGAAATATGTACTA
502	GB_TCGA_ FGFR1_103 3_ADAM18_ 187	CGCACAAAGCCACGGCGGACTCTCCCGAGGCGGAACCTCCAC GCCGAGCGAGGGTCAGTTTGAAAAGGAGGATCGAGCTCACTG TGGAGTATCCATGGAGATGTGGAGCCTTGTACCAACCTCTAA CTGCAGAACTGGGATGTGGAGCTGGAAGTGCCCTCCTCTTCTG GGCTGTGCTGGTCACAGCCACACTCTGCACCGCTAGGCCGTC CCCGACCTTGCCCTGAACAAGATGATTTACATCATTACAATTGAT GGACAACCTTACACTCTACATCTCGGAAAACAATCATTCTTACC CCAGAACTTTTGTTTTATACATATAATGAACTGGATCTTTGC ATTCTGTGTCTCCATATTTTATGATGCATTGCCATTACCAAGGA TATGCTGCCGAATTTCCAAATTCATTTGTGACACTCAGTATATG TTCTGGTCTCAGGGGATTTCTCCAGTTTGAAAATATCAGTTATG GAATTG
503	GB_TCGA_ FGFR1_854	TCGCACAAGCCACGGCGGACTCTCCCGAGGCGGAACCTCCAC GCCGAGCGAGGTCCCTATGTAACTACTTGAGACAAAAATAAA

	_LETM2_12 8	TATGGCCTTCTACAGTTATAATTCAAGTTCTGGCTATTGCTCGAA CAAGATTCCCATTTAAATAAGACATGTATGAAGAACTATGAGAG CAAGAAGTACTCGGATCCTAGTCAGCCAGGCAATACAGTACTT CAGCCAGGAAGTACTAGTAATAACAAAAGCTACACACATCCACTT GCTGGCTGCAAGAAGTTCTGGCAAACCTCAGCTGGAGCAAG C
504	GB_TCGA_ BAG4_660_ FGFR1_156 3	CCACCTGGCTGGGAGAAGGCGGAGGAGGCGATGGCTACTAT CCCTCGGGAGGCGCCTGGCCAGAGCCTGGTCGAGCCGGAGG AAGCCACCAGGAGCAGCCACCATATCCTAGCTACAATTCTAAC TATTGGAATTCTACTGCGAGATCTAGGGCTCCTTACCCAAGTA CATATCCTGTAAGACCAGAATTGCAAGGCCAGGTCCGTTATGC CACCTGGAGCATCATAATGGACTCTGTGGTGGCCTCTGACAAG GGCAACTACACCTGCATTGTGGAGAATGAGTACGGCAGCATC AACCACACATACCAGCTGGATGTCGTGGAGCGGTCCCCCTCAC CGGCCCCATCCTGCAAGCAGGGTTGCCCGCCAACAAACAGTG GCCCTGGGTAGCAACGTGGAGTTCATGTGTAAGGTGTACAGT GACCCGCGAGCCGCACATCCAGTGGCTAA
505	GB_TCGA_ FGFR2_294 8_CCAR2_5 81	TGATGTCTGGTCCCTTCGGGGTGTTAATGTGGGAGATCTTCACT TTAGGGGGCTCGCCCTACCCAGGGATTCCCGTGAGGAACTT TTTAAGCTGCTGAAGGAAGGACACAGAATGGATAAGCCAGCCA ACTGCACCAACGAAGTGTACATGATGATGAGGGACTGTTGGCA TGCAAGTGGCCTCCAGAGACCAACGTTCAAGCAGTTGGTAGA AGACTTGGATCGAATTCTCACTCTCACAACCAATGAGGGTGGG GAGAAACAGCGGGTCTTCACTGGTATTGTTACCAGCTTGCATG ACTACTTTGGGGTTGTGGATGAAGAGGTCTTTTTTCAGCTAAG TGTGGTGAAGGGCCGTCTGCCCCAGCTGGGTGAGAAGGTGCT GGTGAAGGTGCATACAACCCAGGCCAGGCAGTGGCCTGGAA TGCTGTCAAGGTGCAACCGCTCTCCAACCGCCCTACTGAAG TCCCCAGCACCTCCTCTTCTGCATGTAGCAG
506	GB_TCGA_ FGFR2_208 6_FRK_791	TGTGCCGAATGAAGAACAGGACCAAGAAGCCAGACTTCAGCA GCCAGCCGGCTGTGCACAAGCTGACCAAAACGTATCCCCCTGC GGAGACAGGTAACAGTTTCGGCTGAGTCCAGCTCCTCCATGA ACTCCAACACCCCGCTGGTGAGGATAACAACACGCGCTCTCTC AACGGCAGACACCCCCATGCTGGCAGGGGTCTCCGAGTATGA ACTTCCAGAGGACCCAAAATGGGAGTTTCCAAGAGATAAGTGG TTCTTTGGAGCAATCGGAAGATCAGATGCAGAGAAACAATAT TATATTCAAGAAACAAGACCGGTTCTTTCTAATCAGAGAAAGT GAAAGCCAAAAGGAGAATTCTCTCTTTTCAGTTTTAGATGGAG CAGTTGTAAACACTACAGAATTAAGAGACTGGATGAAGGGGG ATTTTTTCTCACGCGAAGAAGAATCTTTTCAACACTGAACGAAT TTGTGAGCCACTACACCAAGACAAGTGAC
507	GB_TCGA_ FGFR3_635 _TACC3_46 0	GGCCAGCAGGAGCAGTTGGTCTTCGGCAGCGGGGATGCTGT GGAGCTGAGCTGTCCCCCGCCCGGGGTGGTCCCATGGGGC CCACTGTCTGGGTCAAGGATGGCAGAGGGCTGGTGCCCTCGG AGCGTGTCTGGTGGGGCCCCAGCGGCTGCAGGTGCTGAAT GCCTCCACAGGAGTCCGGGGCCTACAGCTGCCGGCAGCG GCTCACGCAGCGCTACTGTGCCACTTCAGTGTGCGGGTGAC AGCCAACAGCTCATCAAGGAAGTGGATGCCAAAATACTCATG GAATTCTACAGAAACAGTGGAGGCTGACACCGACCTCCTGG GGGATGCAAGCCAGCCTTTGGGAGTGGCAGCTCCAGCGAGT CTGGCCAGGTGCCCTGGCTGACCTGGACTGCTCAAGCTCTT CCCAGAGCCAGGAAGTTCTGAGAACCAATGGTGTCTCCAG GAAAAGTGTCTGGCAGCCCTGAGCAAGCCGTGGAGGAAAAC
508	GB_TCGA_ FGFR3_701 _TACC3_23 78	GGTGGTCCCATGGGGCCCACTGTCTGGGTCAAGGATGGCACA GGGCTGGTGCCCTCGGAGCGTGTCTGGTGGGGCCCCAGCG GCTGCAGGTGCTGAATGCCTCCACAGGAGTCCGGGGCCTA CAGCTGCCGGCAGCGGCTCACGCAGCGCTACTGTGCCACTT CAGTGTGCGGGTGACAGACGCTCCATCCTCGGGAGATGACGA AGACGGGGAGGACGAGGCTGAGGACACAGGTGTGGACACAG AACGAAGAGTCACTGAAGAAGTGCGTGGAGGATTACCTGGCA

		AGGATCACCCAGGAGGGCCAGAGGTACCAAGCCCTGAAGGC CCACGCGGAGGAGAAGCTGCAGCTGGCAAACGAGGAGATCG CCCAGGTCCGGAGCAAGGCCAGGCGGAAGCGTTGGCCCTC CAGGCCAGCCTGAGGAAGGAGCAGATGCGCATCCAGTCGCTG GAGAAGACAGTGGAGCAGAAGACTAAAGAGAACGAGGAGCTG A
509	GB_TCGA_ FGFR3_253 0_TACC3_2 096	TGACGTCTGGTCCTTTGGGGTCTGCTCTGGGAGATCTTCACG CTGGGGGGCTCCCCGTACCCCGGCATCCCTGTGGAGGAGCT CTTCAAGCTGCTGAAGGAGGGCCACCGCATGGACAAGCCCGC CAACTGCACACACGACCTGTACATGATCATGCGGGAGTGTCTG GCATGCCGCGCCCTCCAGAGGGCCACCTTCAAGCAGCTGGT GGAGGACCTGGACCGTGTCTTACCGTGACGTCCACCGACGT AAAGGCGACACAGGAGGAGAACCAGGGAGCTGAGGAGCAGGT GTGAGGAGCTCCACGGGAAGAACCTGGAAGTGGGGAAGATCA TGGACAGGTTTGAAGAGGTTGTGTACCAGGCCATGGAGGAAG TTCAGAAGCAGAAGGAACCTTCCAAAGCTGAAATCCAGAAAGT TCTAAAAGAAAAAGACCAACTTACCACAGATCTGAACTCCATG GAGAAGTCCTTCTCCGACCTCTTCAAGCGTTTGTAGA
510	GB_Serase q_NTRK_AF AP1_NTRK 2	TCTCCTGCTGACAGAAAGGCCTTAGGCTGTCTGCTGACAAGC TGTCCTCTAACCATTACAAATACCCTGCCTCCGCTCAGTCTGT CACTAATACCTCTTCTGTGGGGAGGGCGTCTCTCGGGCTCAA CTCGCAGCTCAAGGGTAAAAAGCCCCCGTGGCGTCTAATGG GGTCACAGGAAAAGGGAAGACTCTGAGCAGTCAGCCAAAGAA AGCGGATCCCGCGGCTGTTGTGAAAAGGACGGGTTTCGAGTGC AAACCCAAATTATCCTGATGTAATTTATGAAGATTATGGAAGT CAGCGAATGACATCGGGGACACCACGAACAGAAGTAATGAAA TCCCTTCCACAGACGTCACCTGATAAAACCGGTCGGGAACATCT CTCGGTCTATGCTGTGGTGGTGATTGCGTCTGTGTGGGATTT TGCTTTTGGTAATGCTGTTTCTGCTTAAGTTGGCAAGACACTC CAAGTTTGGCATGAAAGATTTCTCATGGTT
511	GB_Serase q_NTRK_N ACC2_NTR K2	CTCTACTCGGAAGGGGACCCCGGGGAGAAGCTGGAGCTGGT GGCAGGCTCTGGGGTCTACATCACACGGGCCAGCTGATGAA CTGCCACCTGTGTGCAGGGGTGAAGCATAAGGTCTTGCTGCG GAGGCTCCTGGCCACCTTCTTTGACAGGAACACGCTGGCCAA CAGCTGCGGGACTGGCATCCGCTCGTCCACCAGCGACCCACAG CCGGAAGCCGCTGGACAGCCGGGTCTGAACGCTGTGAAATA TTATGGAACTGCAGCGAATGACATCGGGGACACCACGAACAG AAGTAATGAAATCCCTTCCACAGACGTCACCTGATAAAACCGGT CGGGAACATCTCTCGGTCTATGCTGTGGTGGTGATTGCGTCTG TGGTGGGATTTTGCCTTTTGGTAATGCTGTTTCTGCTTAAGTTG GCAAGACACTCCAAGTTTGGCATGAAAGATTTCTCATGGTTTG GATTTGGGAAAGTAAAATCAAGACAAGGTGTTGG
512	GB_Serase q_NTRK_Q KI_NTRK2	ATTACTGGGCCTGCGCCGTTCTCCCACCAGCTGCCCTGCGT ACTCCTACGCCAGCTGGCCCTACCATAATGCCTTTGATCAGAC AAATACAGACCGCTGTCATGCCAAACGGAACCTCCTCACCCAAC TGCTGCAATAGTTCCCTCCAGGGCCCGAAGCTGGTTTAATCTAT ACACCCTATGAGTACCCCTACACATTGGCACCAGCTACATCAA TCCTTGAGTATCCTATTGAACCTAGTGGTGTATTAGGCCACAGC CTCCGTTATCAGCAATGATGATGACTCTGCCAGCCCACTCCAT CACATCTCCAATGGGAGTAACACTCCATCTTCTTCGGAAGGTG GCCAGATGCTGTCATTATTGGAATGACCAAGATCCCTGTCAT TGAAAATCCCCAGTACTTTGGCATCACCAACAGTCAGCTCAAG CCAGACACATTTGTTTCAGCACATCAAGCGACATAACATTGTTCT GAAAAGGGAGCTAGGCGAAGGAGCCTT
513	GB_Serase q_NTRK_T RIM24_NTR K2	CAAGCCATAAAACAGTGGCAGATCAGCAGTGGACAGGGAACC CCATCAACTACCAACAGCACATCCTCTACTCCTTCCAGCCCCA CGATTACTAGTGCAGCAGGATATGATGGAAAGGCTTTTGGTTT ACCTATGATCGATTTGAGCTCACCAGTGGGAGGGTCTTATAAT CTTCCCTCTCTCCGGATATTGACTGTTCAAGTACTATTATGCT GGACAATATTGTGAGGAAAGATACTAATATAGATCATTCTCAT

		GGTTTGGATTGGGAAAGTAAATCAAGACAAGGTGTTGGCCC AGCCTCCGTTATCAGCAATGATGATGACTCTGCCAGCCCACTC CATCACATCTCCAATGGGAGTAACACTCCATCTTCTTCGGAAG GTGGCCCCAGATGCTGTCATTATTGGAATGACCAAGATCCCTGT CATTGAAATCCCCAGTACTTTGGCATCACCAACAGTCAGCTC AAGCCAGACACATTTGTTGAGCACAT
514	GB_Serase q_NTRK_P AN3_NTRK 2	AGGCAGTAGCGGGGACTCGATGGACCGCGGCTGGCAATTC CAGGAATGGATGGAGGTGCTTTAACTGATACAAGTCTTGTTCA GCACATCAAGCGACATAACATTGTTCTGAAAAGGGAGCTAGGC GAAGGAGCCTTTGGAAAAGTGTTCCTAGCTGAATGCTATAACC TCTGTCTGAGCAGGACAAGATCTTGGTGGCAGTGAAAGACCC TGAAGGATGCCAGTGACAATGCACGCAAGGACTTCCACCGTG AGGCCGAGCTCCTGACCAACCTCCAGCATGAGCACATCGTCA AGTTCTATGGCGTCTGCGTGGAGGGCGACCC
515	GB_TCGA_ MALAT1- ALK_16	tggggagaaaaatgttttttctaagattttccacagatgctatagtactattgacaaactgggtt agagaaggaggtaccgctgtgctgttggcacgaacaccttcagggactggagctgctttat ccttgaagagtattcccagtgagctgaaaagtagcacagtgagcttgggtcatattc agtcattcaggagaacttcagaagagcttgagtaggccaatgttgaGACACTATG GTGGTGGAGGTGGCTGGAATGATAACACTTCCCTGCTTGGG CCGGAAAATCTTTGCAGGAGGGTGCCACCGGAGGACATTTCCT GCCCCAGGCCATGAAGAAGTGGGGGTGGGAGACAAGAGGG GGTTTCGGAGGGGGTGGAGGGGGGTGCTCCTCAGGTGGAGG AGGCGGAGGATATATAGTGATGGAAGGCCACGGGGAAGTGAA TATTAAGCATTATCTAAACTGCAGTCACTGTGAGGTAGACGA
516	GB_TCGA_ KCNQ5_1- ALK_10- shortened	ATGAGCCTGCTGGGGAAGCCGCTCTCTTACACGAGTAGCCAG AGCTGCCGGCGCAACGTCAAGTACCGGCGGGTGCAGAACTAC CTGTACAACGTGCTGGAGAGACCCCGCGGCTGGGCGTTTCATC TACCACGCTTTCTGTGACACTATGGTTCTGGCTGCAGATGGTCG CATGGTGGGGACAAGGATCCAGAGCCATCGTGGCTTTTGACA ATATCTCCATCAGCCTGGACTGCTACCTCACCATTAGCGGAGA GGACAAGATCCTGCAGAAACAGCACCCAAATCAAGAAACCTG TTTGAGAGAAACCCAAACAAGGAGCTGAAACCCGGGGAAAAT CACCAGACAGACCCCATCTTTGACCCTACAGTTTCATTGGCT GTTCAACCATGTGGG
517	GB_TCGA_ PPP4R3B_1 1-ALK_2	AAAACCGAAAAAGTGAATTTCTAAATTTTTCTACAACCATTTGT ATGCATGTTCTCACAGCACCCTTTTGACCAATACTTCAGAAGA CAAATGTGAAAAGGATTTTTTTTTTAAACATTACAGATAGTTG GAGTTTCATATGTACCCCTTCACATTCCCATTCCCATTCTACCC CCTCTTCTCCATCTCTCAAGATAATATAGTTGGATCAAACAAA AACAACACAATTTGTCCCGGACACTATGGTCATAGCTCCTTGG AATCACCACAAACATGCCTTCTCCTTCTCCTGATTATTTTACA TGGAATCTCACCTGGATAATGAAAGACTCCTTCCCTTCTCTGTC TCATCGCAGCCGATATGGTCTGGAGTGCAGCTTTGACTTCCCC TGTGAGCTGGAGTATTTCCCTCCACTGCATGACCTCAGGAACC AGAGCTGGTCTGGCGCCGCATCCCTCCGAGGAGGCCCTCC CAGATGGACTTGCTGGATGG
518	GB_variant_ BRAF_15_V 600E	TTCCACAAAGCCACAACCTGGCTATTGTTACCCAGTGGTGTGAG GGCTCCAGCTTGTATCACCATCTCCATATCATTGAGACCAAATT TGAGATGATCAAACCTATAGATATTGCACGACAGACTGCACAG GGCATGGATTACTTACACGCCAAGTCAATCATCCACAGAGACC TCAAGAGTAATAATATATTTCTTCATGAAGACCTCACAGTAAAA ATAGGTGATTTTGGTCTAGCTACAGACACTATGGaGAAATCTC GATGGAGTGGGTCCCATCAGTTTGAACAGTTGTCTGGATCCAT TTTGTGGATGGCACCAGAAGTCATCAGAATGCAAGATAAAAAAT CCATACAGCTTTCAGTCAGATGTATATGCATTTGGAATTGTTCT GTATGAATTGATGACTGGACAGTTACCTTATTCAAACATCAACA ACAGGGACCAGATAATTTTTATGGTGGGACGAGGATACCTGTC TCCAGATCTCAGTAAGGTACGGA
519	GB_TCGA_ STRIP1_10-	TCACGCCCCCTCACTTTCCCATGAAGAAAGTTCTCTTGCTGCTC TGGAAGACAGTATTGTGCACGCTAGGCGGCTTTGAGGAGCTG

	CSF1_6	CAGAGCATGAAGGCTGAGAAGCGCAGCATCCTGGGCCTCCCC CCGCTTCCTGAGGACAGCATCAAAGTGATTTCGCAACATGAGAG CAGCCTCTCCACCAGCATCTGCTTCAGACTTGATTGAGCAGCA GCAGAAACGGGGCCGCCGAGAGCACAAGGACACTATGATGTG GTGACCAAGCCTGATTGCAACTGCCTGTACCCCAAAGCCATCC CTAGCAGTGACCCGGCCTCTGTCTCCCCTCATCAGCCCCCTCG CCCCCTCCATGGCCCCCTGTGGCTGGCTTGACCTGGGAGGACT CTGAGGGAAGTGAAGGACAGCTCCCTCTTGCTGGTGAGCAGC CCCTGCACACAGTGGATCCAGGCAGTGCCAAGCAGCGGCCAC CCAGGAGCACCTGCCAGAGCTTTGAGCCGCCAGA
520	GB_variant_ EGFR_20_T 790M_L792 H_G796R_ C797S	AACTGAATTCAAAAAGATCAAAGTGCTGGGCTCCGGTGCGTTC GGCAGGTGTATAAGGGACTCTGGATCCCAGAAGGTGAGAAA GTTAAATTTCCCGTCGCTATCAAGGAATTAAGAGAAGCAACAT CTCCGAAAGCCAACAAGGAAATCCTCGATGAAGCCTACGTGAT GGCCAGCGTGGACAACCCCCACGTGTGCCGCTGCTGGGCA TCTGCCTCACCTCCACCGTGCAGCTCATCGACACTATGaTgcag cAcatgcccttcCgctCcCTCCTGGACTATGTCCGGGAACACAAAGAC AATATTGGCTCCCAAGTACCTGCTCAACTGGTGTGTGCAGATCG CAAAGGGCATGAACTACTTGGAGGACCGTCGCTTGGTGACCC GCGACCTGGCAGCCAGGAACGTACTGGTGAAAACACCGCAGC ATGTCAAGATCACAGATTTTGGGCTGGCCAACTGCTGGGTGC GGAAGAGAAAAGAATACCATGCAGAAG
521	GB_variant_ EGFR_21_L 858R	GCTGGGCATCTGCCTCACCTCCACCGTGCAGCTCATCACGCA GCTCATGCCCTTCGGCTGCCTCCTGGACTATGTCCGGGAACA CAAAGACAATATTGGCTCCCAGTACCTGCTCAACTGGTGTGTG CAGATCGCAAAGGGCATGAACTACTTGGAGGACCGTCGCTTG GTGCACCGCGACCTGGCAGCCAGGAACGTACTGGTGAAAACA CCGCAGCATGTCAAGATCACAGATTTTGGGGACACTATGCGGG CCAACTGCTGGGTGCGGAAGAGAAAAGAATACCATGCAGAAG GAGGCAAAGTGCCTATCAAGTGGATGGCATTGGAATCAATTTT ACACAGAATCTATACCCACCAGAGTGATGTCTGGAGCTACGGG GTGACTGTTTGGGAGTTGATGACCTTTGGATCCAAGCCATATG ACGGAATCCCTGCCAGCGAGATCTCCTCCATCCTGGAGAAAG GAGAACGCCTCCCTCAGCCACCCATATGTACCA
522	GB_TCGA_ FGFR1_1- LETM2_2sh ortened	TCGCACAAGCCACGGCGGACTCTCCCAGGGCGGAACCTCCAC GCCGAGCGAGGACACTATGGTCCCCTATGTTAACTACTTGAGAC AAAAATAAATATGGCCTTCTACAGTTATAATTCAGTTCTGGCTA TTGCTCGAACAAGATTCCCATTTAAATAAGACATGTATGAAGAA CTATGAGAGCAAGAACTACTCGGATCCTAGTCAGCCAGGCAAT ACAGTACTTCACCCAGGAAGTACTAATAACAAAGCTACACA CATCCACTTGCTGGCTGCAAGAAGTTCTGGCAAACCTCAGCT GGAGCAAGC
523	GB_variant_ KRAS_2_G 12D_G13D- shortened	AGAGGCTCAGCGGCTCCCAGGTGCGGGAGAGAGGCCTGCTG AAAATGACTGAATATAAACTTGTGGTAGTTGGAGCTGACACTAT GGATGACGTAGGCAAGAGTGCCTTGACGATACAGCTAATTCAG AATCATTTTGTGGACGAATATGATCCAACAATAGAGGATTCCCTA CAGGAAGCAAGTAGTAATTGATGGAGAAACCTGTCTCTTGGAT ATTCTCGACACAGCAGGTCAAGAGGAGTACAGTGCAATGAGG GACCAGTACATGAGGACTGGGGAGGGCTTTCTTTGTGATTTG CCATAAATAATACTAAATCATTTGAAGATATTACC
524	GB_variant_ MET_14_del	GAGAATATACACTTACATTCTGAAGCCGTTTTATGCACGGTCC CCAATGACCTGCTGAAATTGAACAGCGAGCTAAATATAGAGTG GAAGCAAGCAATTTCTTCAACCGTCCCTGGAAAAAGTAATGTTT AACCAGATCAGAATTTACAGGATTGATTGCTGGTGTCTC AATATCAACAGCACTGTTATTACTACTTGGGTTTTCTGTGGC TGAAAAAGAGAAAGCAAATTAAGGACACTATGATCAGTTTCT AATTCATCTCAGAACGGTTTCATGCCGACAAGTGCAATCCTC TGACAGACATGTCCCCATCCTAACTAGTGGGAGCTCTGATAT ATCCAGTCCATTACTGCAAAATCTGTCCACATTGACCTCAGTG CTCTAAATCCAGAGCTGGTCCAGGCAGTGACAGCATGTAGTGAT

		TGGGCCCAGTAGCCTGATTGTGCATTTCAATGAAGTCATAGGA AGAGGGCATTTTGGTTGTGTATA
525	GB_TCGA_ PTPRZ1_1- MET_2- shortened	TACGCACGCACGATCTCACTTCGATCTATACACTGGAGGATTA AAACAAACAAACAAAAAAACATTTCTTCGCTCCCCCTCCCTC TCCACTCTGAGAAGCAGAGGAGCCGCACGGCGAGGGGCCGC AGACCGTCTGGAAATGCGAATCCTAAAGCGTTTCTCGCTTGC ATTCAGCTCCTCTGTGTTTGCCGCCTGGGACACTATGATAAAC CTCTCATAATGAAGGCCCCCGCTGTGCTTGACCTGGCATCCT CGTGCTCCTGTTTACCTTGGTGCAGAGGAGCAATGGGGAGTG TAAAGAGGCACTAGCAAAGTCCGAGATGAATGTGAATATGAAG TATCAGCTTCCCAACTTCACCGCGGAAACACCCATCCAGAATG TCATTCTACATGAGCATCACATTTTCTTGGTGCCACTAACTAC ATTTATGTTTTAAATGAGGAAGACCTTCA
526	GB_TCGA_ ZBTB7B_1- NTRK1_2- shortened	ACGCGCTTCTTCCACACTGTGAGCCTCAGCAGCTCCAGCCA GCGGACCCGACGGCTGAGAGGACACTATGCCCTCAACAGAAA GTGCGAGAAGATGCCAAACAGCTCATGGACCAGGGGAGGAGC TTCCACAGTGCGTCCACACTGGGCCAAGGGCAAGGTCCAAAG TGTCTACTCCAGGACTGCCCCCTTGGTTGGACATCTAAGACCTG GCCGCCCTCATCTGCCTGGCACCTCTGTACCCCTGCTTG ACGGTGAAGTCTGGGACACCATGCAGTTGCGGGCTGCTAGA TCTCGGTGCACAACTTGTGGCAGC
527	GB_TCGA_ ARHGEF25 _10- NTRK1_7	CAAACACGAGCGCCGGCTGCATATGTATGTGGTGTACTGTCA GAATAAGCCCAAGTCAGAGCATGTGGTGTGACAGTTTGGGGA CAGCTACTTTGAGGAGCTCCGGCAGCAGCTGGGGCACCGCCT GCAGCTGAACGACCTCCTCATCAAACCTGTGCAGCGGATCAT GAAATACCAGCTGCTGCTCAAGGATTTTCTCAAGTATTACAATA GAGCTGGGATGGATACTGCAGACCTAGAGGACACTATGAAAT CTGGGGGTCTGCCATCCCTGGGGCTGACCCTGGCCAATGTCA CCAGTGACCTCAACAGGAAGAACGTGACGTGCTGGGCAGAGA ACGATGTGGGCCGGGCAGAGGTCTCTGTTCAAGTCAACGTCT CCTTCCCGGCCAGTGTGCAGCTGCACACGGCGGTGGAGATGC ACCACTGGTGCATCCCCTTCTCTGTGGATGGGCAGCCGGCAC CGTCTCTGCGCTGGCTCTTCAATGGCTCCGTGCTCA
528	GB_TCGA_ VSP18_4- NTRK3_18	CGTGAACGTGCTGGGGGAGACTGAGCAGGCCATCCACAATA CCTGCTGTCACTGTATGCCCGTGGCCGGCCGGACTCACTACT GGCCTATCTGGAGCAGGCTGGGGCCAGCCCCACCGGGTGC ATTACGACCTCAAGTATGCGCTGCGGCTCTGCGCCAGCATG GCCACCACCGCGCTTGTGTCCATGTCTACAAGGTCTTAGAGCT GTATGAGGAGGCCGTGGACCTGGCCCTGCAGGACACTATGGT GGGAGGACACACCATGCTCCCCATTGCTGGATGCCTCCTGA AAGCATCATGTACCGGAAGTTCACTACAGAGAGTGATGTATGG AGCTTCGGGGTGATCCTCTGGGAGATCTTACCTATGGAAAGC AGCCATGGTTCCAACTCTCAAACACGGAGGTCAATTGAGTGCAT TACCCAAGGTGCTGTTTTGGAGCGGCCCGAGTCTGCCCAA AGAGGTGTACGATGTCATGCTGGGGTCTGCGCAGA
529	GB_TCGA_ FAT1_2- NTRK3_7	CCAGTTTCTCTGTCTTCTACTTGCTATGTTGAAGTTGAGGTGGT TGATGTGAATGAGAACCTGCACCCACCCGTGTTTTCCAGCTTT GTGGAAAAGGGGACAGTGAAAGAGATGCACCTGTTGGTTCA TTGGTAATGACGGTGTGCGCTCATGATGAGGACGCCAGAAGA GATGGGGAGATCCGATACTCCATTAGAGATGGCTCTGGCGTT GGTGTTTTCAAATAGGTGAAGAGACAGGACACTATGGCAGTT GGAGCAGAACTTTTTCAACTGCAGCTGTGACATCCGCTGGATG CAGCTCTGGCAGGAGCAGGGGGAGGCCAAGCTCAACAGCCA GAACCTCTACTGCATCAACGCTGATGGCTCCAGCTTCTCTCT TTCCGCATGAACATCAGTCAGTGTGACCTTCTGAGATCAGCG TGAGCCACGTCAACCTGACCGTACGAGAGGGTGACAACGCTG TTATCACTTGCAATGGCTCTGGATCACCCCTT
530	GB_COSMI C_COL1A1_ 25-	TGAACGTGGTTCTCCTGGCCCTGCTGGCCCCAAAGGATCTCC TGGTGAAGCTGGTCTGCCCGTGAAGCTGGTCTGCTGGTGC CAAGGGTCTGACTGGAAGCCCTGGCAGCCCTGGTCTGATGG

	PDGFB_2	CAAACTGGCCCCCTGGTCCCGCCGGTCAAGATGGTCGCCC CGGACCCCCAGGCCACCTGGTGCCCGTGGTCAGGCTGGTG TGATGGGATTCCCTGGACCTAAAGGTGCTGCTGACACTATGG GGGACCCCATTCCTCGAGGAGCTTTATGAGATGCTGAGTGACC ACTCGATCCGCTCCTTTGATGATCTCCAACGCCTGCTGCACGG AGACCCCGGAGAGGAAGATGGGGCCGAGTTGGACCTGAACAT GACCCGCTCCCACTCTGGAGGCGAGCTGGAGAGCTTGGCTCG TGGAAGAAGGAGCCTGGGTTCCCTGACCATTGCTGAGCCGGC CATGATCGCCGAGTGCAAGACGCGCACCCGAGGTGTTCTG
531	GB_TCGA_ MKRN2_1- PPARG_2- shortened	ACGACGACGGTCCCTCAGCCCAGCCACCATGAGCACCAAGCA GATCACTTGCAGGACACTATGATTTGAAAGAAGCCAACACTAA ACCACAAATATACAACAAGGCCATTTTCTCAAACGAGAGTCAG CCTTTAACGAAATGACCATGGTTGACACAGAGATGCCATTCTG GCCCCACCAACTTTGGGATCAGCTCCGTGGATCTCTCCGTAATG GAAGACCACTCCCACTCCTTTGATATCAAGCCCTTCACTACTG TTGACTTCTCCAGCATTTCTACTCCACATTACGAAGACATTCCA TTCACAAGAACA
532	GB_TCGA_ KIF5B_15- RET_11	AGAAGAACTTGCTGTCATTATGATCAGAAAGTCTCAGGAAGTT GAAGACAAAATAAGGAATATGAATTGCTTAGTGATGAATTGAA TCAGAAATCGGCAACTTTAGCGAGTATAGATGCTGAGCTTCAG AACTTAAGGAAATGACCAACCACCAGAAAAACGAGCAGCTG AGATGATGGCATCTTTACTAAAGACCTTGCAGAAATAGGAATT GCTGTGGGAAATAATGATGTAAAGGACACTATGATCCACTGTG CGACGAGCTGTGCCGACCGGTGATCGCAGCCGCTGTCTCTCT CTCCTTCATCGTCTCGGTGCTGCTGTCTGCCTTCTGCATCCAC TGCTACCACAAGTTTGCCACAAGCCACCCATCTCCTCAGCTG AGATGACCTTCCGGAGGCCCGCCAGGCCTTCCCGGTGACGT ACTCCTCTCCGCTGCCCGCCGCCCCCTCGCTGGACTCCATGG AGAACCAGGTCTCCGTGGATGCCTTCAA
533	GB_TCGA_ KIF5B_24- RET_10	CAGCTCACTAAAGTGCACAAACAGTTGGTACGTGATAATGCAG ATCTCCGCTGTGAACCTTCTAAGTTGGAAGGCGACTTCGAGC TACAGCTGAGAGAGTGAAAGCTTTGGAATCAGCACTGAAAGAA GCTAAAGAAAAATGCATCTCGTGATCGCAAACGCTATCAGCAAG AAGTAGATCGCATAAAGGAAGCAGTCAGGTCAAAGAATATGGC CAGAAGAGGGCATTCTGCACAGATTGGACACTATGGGGGCGAG CATTGTTGGGGGACACGAGCCTGGGGAGCCCCGGGGGATTA AAGCTGGCTATGGCACCTGCAACTGCTTCCCTGAGGAGGAGA AGTGCTTCTGCGAGCCCCGAAGACATCCAGGATCCACTGTGCG ACGAGCTGTGCCGACCGGTGATCGCAGCCGCTGTCTCTTCT CCTTCATCGTCTCGGTGCTGCTGTCTGCCTTCTGCATCCACTG CTACCACAAGTTTGCCACAAGCCACCCATCTC
534	GB_TCGA_ FKBP15_22- RET_8	GAACCGGAAACAACCTGGAACCTCAAGGTGACATCCCTGGAGGA GGAACCTGACTGACCTTCGAGTTGAGAAGGAGTCCTTGAAAA GAACCTCTCAGAAAGGAAAAAGATCAGCTCAAGAGCGTTCT CAGGCCGAGGAGGAGATAGATGAAATTGCAAGTCATACCAAG GAGGAATTGGACAACTTCGACAGCTCTTGAAAAAGACTCGAG TGTCACACAGACCAAGCAGCTGCAGAGCAGGACACTATGATGT GGCCGAGGAGGCGGGCTGCCCCCTGTCTGTGCAGTCAGCA AGAGACGGCTGGAGTGTGAGGAGTGTGGCGGCCCTGGGCTCC CCAACAGGCAGGTGTGAGTGGAGGCAAGGAGATGGCAAAGG GATCACCAGGAACCTTCTCCACCTGCTCTCCAGCACCAAGACC TGCCCCGACGGCCACTGCGATGTTGTGGAGACCCAAGACATC AACATTGCCCCCTCAGGACTGCCTCCGGGGCAGCATTGT
535	GB_TCGA_ KIF5B_24- RET_7	CAGCTCACTAAAGTGCACAAACAGTTGGTACGTGATAATGCAG ATCTCCGCTGTGAACCTTCTAAGTTGGAAGGCGACTTCGAGC TACAGCTGAGAGAGTGAAAGCTTTGGAATCAGCACTGAAAGAA GCTAAAGAAAAATGCATCTCGTGATCGCAAACGCTATCAGCAAG AAGTAGATCGCATAAAGGAAGCAGTCAGGTCAAAGAATATGGC CAGAAGAGGGCATTCTGCACAGATTGGACACTATGATCGGGA AAGTCTGTGTGGAAACTGCCAGGCATTCAAGTGGCATCAACGT

		CCAGTACAAGCTGCATTCTCTGGTGCCAACTGCAGCAGCGTA GGGGTGGTCACCTCAGCCGAGGACACCTCGGGGATCCTGTTT GTGAATGACACCAAGGCCCTGCGGCGGCCCAAGTGTGCCGAA CTTCACTACATGGTGGTGGCCACCGACCAGCAGACCTCTAGG CAGGCCCAGGCCAGCTGCTTGTAACAGTGG
536	GB_TCGA_ CCDC6_1- RET_2- shortened	ATTGTCATCTCGCCGTTCCGCCTGGAGGAGCTCACC AACCGC CTGGCCTCGCTGCAGCAAGAGAACAAGGTGCTGAAGATAGAG CTGGAGACCTACAACTGAAGTGCAAGGCACTGCAGGAGGAG AACCGCGACCTGCGCAAAGCCAGCGTGACCATCGACACTATG TGGCATTGGGCTCTACTTCTCGAGGGATGCTTACTGGGAGAA GCTGTATGTGGACCAGGCAGCCGGCACGCCCTTGCTGTACGT CCATGCCCTGCGGGACGCCCTGAGGAGGTGCCAGCTTCC GCCTGGGCCAGCATCTCTACGGCACGTACCGCACACGGCTGC ATGAGAACAACCTGGATCTGCATCCAGGAGGACACCGGCCTCC TCTACCTTAACCGGAGCCTGGACCATAGCTCCTGGGAGAA

[00163] As indicated in **Figure 12A**, antiprimers (reverse complement oligos of the GSP molecules) were included to convert excess unextended gene specific primer into dsDNA duplexes not available for ligation. **Figures 12B-G** show visualization of 1ng of components with Agilent Bioanalyzer High Sensitivity DNA Chip, as follows (B) Linear ssDNA template (C) With ligation and antiprimer (D) With ligation without antiprimer (E) Without ligation (no adapter:splint) (F) Without ligation (no ligase) (G) dsDNA template. electrophoresis as electropherogram and gel image composite. Ligation was expected for **Figures 12C** and **12D** and not for **12E** and **12F**.

[00164] Because the bioanalyzer system uses a dsDNA ladder and ligation products are ssDNA, partial duplexes, hybrid duplexes and dsDNA duplexes, migration sizing was not accurate. We observed the initial ssDNA template and two shifted populations representing ligation negative and ligation positive reactions. In the absence of splint:adapter hybrid duplex or ligase (ligation negative) we observed one shift. In the presence of all components and in the absence of antiprimers, which are annealed to the linear amplicon to prevent ligation of unextended primer, we observed a second shift (ligation positive). This is as expected as both the splint:adapter and ligase are necessary for ligation while the antiprimers are not and therefore their absence does not affect ligation at the phosphoacceptor 3' end however because they form partial duplexes with the phosphoacceptor molecules at their 5' ends they do affect migration.

[00165] Following ligation of phosphoacceptor linear ssDNA amplicon product with ligase to phosphodonor adapter via random splinting, next generation sequencing libraries were generated from ligation product by performing PCR with a primer specific to the phosphodonor adapter (Nextera i7_R: 5'-

GTCTCGTGGGCTCGGAG-3', SEQ ID NO: 3) and with nested primers internal to the first GSP primer set used to generate the ssDNA ligation template (**Table 6**) (**Figure 13**). These primers have Illumina Nextera platform tails that were subsequently used to add indexes and add Illumina p5 and p7 platforms.

[00166] Following ligation of geneblock derived linear amplicon, next generation sequencing libraries were generated from ligation product (**Figure 14**). When libraries were generated from reactions lacking initial dsDNA template, we observed no difference in library mass yield when ligation components were all present or individually omitted. When geneblock template was used, we observed a 15-35x difference between ligation negative and ligation positive reactions. In ligation negative reactions non-specific amplification was observed while in ligation positive reactions specific product was observed at 450bp.

[00167] To demonstrate the applicability of this method to an extensively diverse population of endogenous cellular RNA, an Illumina NGS library was generated from universal normal human RNA (**Figure 15**). Total RNA was randomly converted into double stranded cDNA which was then subjected to linear amplification by PCR to generate linear ssDNA template for ligation. When the ligation product was amplified by PCR, non-specific product < 300bp was observed for a library from ligation product made without template as well as from product in the absence of ligase. Specific library of size 300-1000bp was generated in the presence of ligase indicative of a diverse population of amplicon with variable lengths. Library yield was 3fold that of the ligation negative and template negative libraries.

[00168] When ligation product was converted into Illumina NGS library and sequenced via Illumina paired end 300 cycle miseq sequencing with ~1M reads per sample, the diversity of individual template molecules could be assessed by the counting of unique amplicon end positions represented in the library since template RNA molecules exist as a randomly fragmented population (**Figure 16**). When sequenced, the end position is the start of the illumina P7 read while the GSP sequence is the start of the illumina P5 read. To assess diversity, we examined *TUBB* gene amplicons observed in the library generated from universal normal human RNA. The *TUBB* cDNA transcript was amplified with a single gene specific primer 608bp from the start of the *TUBB* transcript meaning that amplicons of up to 608nt could be theoretically observed. Due to degradation of the 5' end due to RNA instability and the inability of random cDNA priming to capture the entire 5' end of the

transcript observation of the entire transcript is however not expected. Consequently, we observed end sites of up to 529nt from the start of the GSP and saw that longer transcripts are more relatively more rare than short transcripts as predicted due to RNA instability.

[00169] To quantify the diversity of initial template molecules incorporated into the library, NGS libraries were prepared from multiple templates: no template control (NTC), amsbio universal normal human RNA, seraseq v4 RNA fusion mix, and a blend of amsbio universal normal human RNA and synthetic geneblocks (**Table 8**).

[00170] **Table 8: Read count and template diversity from ligation product libraries**

Sample Description	RNA (ng)	Library (ng)	On-target reads	On-target Unique Ends	Genes (21)	Exons (150)
NTC	0	2.9	2	2	1	2
Amsbio	500	32.7	302915	11985	21	148
Seraseq	25	11.5	244954	6224	20	103
10pg Geneblocks + Amsbio	25	45	533337	15086	21	137

[00171] Libraries were sequenced via Illumina paired end 300 cycle miseq sequencing with ~1M reads per sample. Gene specific primers were used in the linear amplification process to amplify regions from 150 exons within 21 genes. As expected, no significant number of on-target reads were observed when library was generated without template. In contrast, 244,954-533,337 on-target reads were generated from template positive samples. This corresponded to 6224-15,086 unique on-target ends suggesting at least this many unique template molecules were incorporated into the library. These reads were associated with 20-21 of the 21 expected genes and 103-148 of the expected 150 exons. Coverage is not expected for all targets in all samples due to differences in gene expression and input amounts.

[00172] Accordingly, application of the fusion detection methods permitted identification of structural rearrangements in reads from the previously described sample libraries after Illumina paired end 300 cycle miseq sequencing with ~1M reads per sample (**Table 9**).

[00173] **Table 9: Read count, template diversity, fusion calls, fusion read depths, and fusion unique ends from ligation product libraries**

Sample Description	RNA (ng)	Library (ng)	On-target reads	On-target Unique Ends	Genes (21)	Exons (150)	Fusions	Read depth (min-max)	Unique ends (min-max)
NTC	0	2.9	2	2	1	2	0	NA	NA
Ambio	500	32.7	352915	11985	21	148	0	NA	NA
Seraseq	25	11.5	244954	8224	20	103	18	31-3326	12-75
10pg GeneBlocks + Ambio	25	45	533337	15088	21	137	56	66-39918	9-292

[00174] The NTC sample is template negative and fusion negative and consequently no significant number of reads and no fusions were detected. The ambio universal normal human RNA sample is template positive with no fusions expected and this was observed. The seraseq v4 RNA fusion mix is template positive and contains 18 expected fusions which were all detected. The ambio universal normal human RNA sample spiked with 10pg of 78 fusion geneblocks resulted in detection of all detectable 56 fusions based on the target GSP sequences used.

[00175] The resulting sequencing library, with inserts derived from the set of PCR amplified adapter-ligated DNA molecules, is then subjected to computational analysis that includes filtering and error correction steps for the purposes of identifying among this set, any target gene with atypical exon configuration, or target gene which has been fused with any other DNA sequence such as a partner gene in a fusion configuration. Such genomic rearrangements or fusion events are clinically actionable with targeted inhibitors. In the case of cancer for example, targeted inhibitors can be administered to cancer patients when such fusion events are identified in patient samples. This combinational analysis is now described.

[00176] The above-mentioned computational analysis is referred to as a fusion pipeline which is programmed and executed locally on one or more computer systems, or in the cloud.

[00177] **Input Files**

[00178] Input files stored on some medium are provided to the computing system. These input files can include data stored or presented in any conventionally known format. Alternatively, if conventional formats are insufficient for the fusion pipeline, then custom formats can be developed with the fusion pipeline algorithm

configured to recognize these custom formats. The required files for the fusion pipeline of the present example include:

[00179] - A *reference genome* FASTA file, assembly GRCh38.p13, openly available at the Ensembl FTP server
https://uswest.ensembl.org/Homo_sapiens/Info/Index.

[00180] - An annotations file in TSV format, with all the CDS exons and corresponding transcript annotations (Ensembl 99, obtainable at Ensembl Biomart - <https://uswest.ensembl.org/biomart/martview>).

[00181] The assay manifest file, with information about all gene targets and corresponding gene-specific primers. This can be the information previously described in the target enrichment section above.

[00182] - The test samples FASTQ files, the output of the Illumina MiSeq sequencer as described in the Laboratory Methods section, containing the cDNA sequences where fusion events will be called. This information is obtained after using the adapter/amplicon/primer PCR methods that were previously described. For each sample, the sequencer outputs two files, one with the R1 reads, corresponding to reads starting at the gene-specific primer, and another with the R2 reads, starting from the adapter end. R1 and R2 reads are referred to as read 1 and read 2 on Figure 2, in the Laboratory Methods section. Each R1 read has a corresponding R2 read and are called a read pair.

[00183] **Read Filtering**

[00184] The first step in the pipeline is to process reads from the FASTQ files, filtering out-of-target/misprimed reads and also generating QC metrics.

[00185] For each R1/R2 read pair, since it is known that the R1 read starts at a gene-specific primer on a driver gene (the known sequence, Figure 2), the first step is to identify which primer is present at the start of the R1 read, with maximum hamming distance of 1 from all the primer sequences in the assay manifest file. If there is a primer satisfying this distance condition, the read pair is labelled with the corresponding primer, otherwise the read pair is removed.

[00186] Then, for all labelled read pairs, the start of the R1 read is aligned with the expected target sequence (primer + 20 bps, up to exon boundary, again the

known sequence of Figure 2). The alignment distance d_{ij} between a read i and the expected target sequence j is defined as

$$d_{ij} = \frac{2l - G_{ij}}{2}$$

where l is the length and G_{ij} is the global alignment score of the compared sequences, with score 2 for match, -1 for mismatch and no penalty for gaps, obtained with a dynamic programming alignment using the open-source Bio.pairwise2 module from Biopython. If the primer binds to the known sequence region, as expected, this distance should usually be zero, so larger values point to an *off-target sequence* (when the primer binds to a different region) and should be filtered out. But sequencing errors can also cause the distance to increase, even when the primer correctly amplified the known region, so the challenge is to select a threshold that is small enough such that most off-target sequences are filtered out, without removing good sequences that have sequencing error. After several tests, the distance threshold with best results was found to be 1. Therefore, all read pairs where $d_{ij} > 1$ are filtered out, and the remaining are saved in an intermediate file to be used as an input in the fusion calling step.

[00187] **Fusion Calling**

[00188] The fusion calling step takes as an input the filtered reads file generated in the read filtering step above, in addition to the reference genome and annotations file. Each of the phases of this step are described in the sections below.

[00189] **K-mer Indexing**

[00190] In order to identify the origin of a particular cDNA sequence, a *mapping* step is performed, where a given read sequence from a test sample is *mapped* to a reference sequence (in this case, the human genome), so the step of finding genome mutations can be done. This is a common step in any variant calling pipeline, and different methods have been proposed. In the described pipeline according to the present embodiments, a k-mer mapping algorithm is used, and a few definitions will be needed to describe this mapping step. It should be noted that this fusion pipeline and any of its described constituent steps are not limited to human genome analysis. In alternate embodiments, the fusion pipeline can be configured to analyze any animal genome, provided a reference sequence is available.

[00191] A k-mer is a substring of length k of a given DNA sequence. For a given k-mer, the canonical representation, or simply canonical k-mer, is the lexicographically smaller of the k-mer itself and its reverse complement. For instance, for the 5-mer TAGCT, the reverse complement is AGCTA and therefore AGCTA is the canonical k-mer. This means that a k-mer and its reverse complement have the same canonical representation.

[00192] A k-mer index is a structure where information about a given k-mer from a (usually very large) set of k-mers can be accessed. This information is commonly the list of sequences and locations where this k-mer is present, such a reference genome or transcriptome.

[00193] In this pipeline, the k-mer index *I* is a simple hash table where the keys are all canonical k-mers found in the reference genome, and the information stored is the set of tuples, with exons that have the k-mer and the coordinate of the k-mer hit within the exon. The default k-mer length is 17, however the entire process can be repeated with a different value of k. The default value of 17 was determined based on tests with different k-mer sizes. In similar projects, either a default is selected based on performance tests, or a range of different values can be used, and all results are compared. In one test case, a k value of 17 was optimal, but k of 19 or 21 are also usable with similar results. $I[k]$ is defined as the k-mer hits of a k-mer *k*:

$$I[k] = \{(exon_1, pos_1), (exon_2, pos_2), \dots\}$$

[00194] For instance, in the example from **Figure 5** (a), $I[GCAGGTT] = \{(ENSE001, 1)\}$ and $I[TGGCAGG] = \{(ENSE001, 12)\}$. ENSE001 is an example exon identifier, based on the Ensembl nomenclature, where all the exon identifiers have the "ENSE" prefix.

[00195] Alternative structures have been proposed in the literature to store k-mer indexes in a more efficient way, especially with respect to memory usage reduction and might be used on future versions of the k-mer index.

[00196] The k-mer index is created on step 2 on Figure 4. It is kept in memory during the execution, but other approaches could be used in order to minimize the memory usage, such as using a fast database such as ZODB (<http://www.zodb.org>) that would offer a trade-off between speed and memory.

[00197] After its creation, the k-mer index is stored on a network or some removable storage medium in a serialized form using *pickle* (<https://docs.python.org/2/library/pickle.html>), which means that it does not need to be created for each new run of the pipeline, if there are no changes in the reference genome or annotation file, and can instead be directly loaded from the *pickle* file.

[00198] Read Pseudo-alignment

[00199] This step (**Figure 4, step 3**) uses the k-mer index to map reads to reference genome exons. The reference genome is the linear DNA sequence with all chromosomes. Then with the k-mer index the system can find the genome location of a read by analyzing the k-mer patterns of the read. The aim is to identify the corresponding exon regions of the sequencing read (**Figure 5 (b)**), to then find reads that map to two different genes (the driver gene in the known sequence, and a different partner gene in the unknown sequence, **Figure 2**), indicating a possible fusion supporting read. The following subsections describe each step in the pseudo-alignment phase.

[00200] K-mer mapping and finding exon candidates

[00201] Given a k-mer from a sequencing read, a k-mer hit is the position of this k-mer in the read and the mapping information of this k-mer from the k-mer index. If a k-mer is not present in the index, the corresponding mapping information is called a *miss*.

[00202] For instance, the k-mer hit of a k-mer k_1 on position 1 of a read is a tuple $(1, I[k_1])$, where $I[k_1]$ is the k-mer index information of this k-mer (or miss if the k-mer is not present in the index). Therefore, repeating this process for all k-mers in a read results in a vector of k-mer hits, representing the original read:

$$(1, I[k_1]), (2, I[k_2]), \dots, (n, I[k_n])$$

Where n is the number of k-mers on the read, and $k_1, \dots, k_n$ are all the k-mer in the read.

[00203] Consecutive hits for the same unique exon match are then aggregated into a single element called exon candidate e_i , which is defined as a tuple of three elements: the exon, the range of the hit on the read, and the range of the mapped region on the exon: $(e_i, \text{read_st} \rightarrow \text{read_end}, \text{map_st} \rightarrow \text{map_end})$. Consecutive misses are also merged, but instead of a mapped range, the length is stored instead.

The combined misses are called a missed block. Exon candidates with small range are filtered, that is, if $\text{read_end} - \text{read_st} < \text{min_threshold}$. The default min_threshold is $24 - k + 1$, resulting in a minimum mapped requirement of 24 bps in the present embodiment.

[00204] At the end of the process, a read is then represented as a *mapping vector* r_i composed of consecutive exon candidates and misses.

[00205] For instance, $r_i = [(e_1, 1 \rightarrow 41, 50 \rightarrow 90), (\text{miss}, 42 \rightarrow 61, 20), (e_2, 62 \rightarrow 85, 1 \rightarrow 24)]$ means that from the base pairs 1 to 41, the read maps to region 50-90 on exon e_1 , then the next 20 base pairs on the range 42-61 do not map to anything on the k-mer index, and finally the base pairs from 62 to 85 map to region 1 to 24 on exon e_2 .

[00206] **Sequencing error/SNP correction**

[00207] This error correcting step is optional if sequencing errors are at a minimum or nonexistent. This error correcting step is applied to any read vector from the previous step that has a missed block, such that the corrected version is retained in the pool of exon candidates. Read errors or SNPs might cause the k-mer mapping step to miss some hits. When this happens, it is possible to correct the resulting mapping if there are exon candidates “around” the missed k-mers, if the mapped coordinates agree. Specifically, consider a missed block with exon candidate block with the same exon before and after it:

$$(e_1, r_1 \rightarrow r_2, m_1 \rightarrow m_2), (\text{miss}, \text{len}), (e_1, r_3 \rightarrow r_4, m_3 \rightarrow m_4)$$

If $r_2 + \text{len} + 1 = r_3$ (or similarly, $m_2 + \text{len} + 1 = m_3$), then we can assume that the missed block was caused by an error since the surrounding areas agree.

[00208] For instance, in the example shown on **Figure 6**, the initial mapping results in $r_i = [(e_1, 1 \rightarrow 3, 51 \rightarrow 53), (\text{miss}, 4 \rightarrow 10, 7), (e_1, 11 \rightarrow 12, 61 \rightarrow 62)]$. Since $53 + 7 + 1 = 61$, the miss can be removed and the two exon candidates are merged, resulting in $r_i = (e_1, 1 \rightarrow 12, 51 \rightarrow 62)$.

[00209] **Resolving mapping ambiguities**

[00210] This step is applied against all the valid exon candidates resulting from the previous steps above. After the k-mer mapping step, some mapped exon candidates might be overlapping, either from mismatched regions due to sequence

similarity between different regions, or different exon annotations for the same genomic region, among other reasons. Those ambiguities are resolved, resulting in a mapping without overlapping exon candidates. Following is an example of how a mapped ambiguity is corrected.

[00211] For instance, if $r_i = [(e_1, 1 \rightarrow 21, 51 \rightarrow 71), (e_2, 7 \rightarrow 15, 62 \rightarrow 70)]$, meaning that region 1-21 of this read mapped to e_1 from 51-71, but also region 7-15 mapped to a different exon e_2 , from 62-70. Therefore, there is an overlap (in this case the mapping of e_2 is completely contained in e_1 , but not always). The e_2 mapping is removed, since it is smaller than e_1 .

[00212] For each overlapping pair of exon candidates, the following checks are made until one of the conditions is met:

1. If the exon is the same on both candidates, they are merged, adjusting the mapping coordinates accordingly.
2. If one of the candidates has a much smaller mapping range than the other, it is removed. This is because a larger mapped area indicates a higher confidence in the accuracy of the mapping. The chance of a spurious match decreases rapidly with increasing mapped length.
3. If one of the candidates has an exon from an assay targeted gene, the other candidate is removed. It is assumed that sequences come from the targeted genes, so if there is ambiguity in a match, because the sequence is similar to different parts of the genome, this is resolved in favour of the targeted gene match, since it is the expected sequence to see.
4. If one of the candidates matches a list of preferred candidates, the other candidate is removed. This is useful when mapping the R2 read, using the exon candidates from R1 as the preferred candidate list.
5. If both candidate exons are from the same gene, choose the candidate with highest priority, calculated from sorting the corresponding transcripts for each exon in descending order of length.
6. If all previous tests do not solve the ambiguity, delete the candidate with the largest exon gap, and in case of a tie, keep the candidate with lexicographically smaller ID, to be consistent.

[00213] **Amplicon End Estimation**

[00214] Since the assay is based on an anchored linear amplification step, different sequenced amplicons will have different lengths, and therefore different endpoints (or starting R2 sequences). This step is a data extraction of information of the various amplicons to determine their uniqueness based on its endpoint and length. The genomic coordinate of the amplicon end of the current read pair being processed can be estimated by looking at the R2 read mapping. In this mapping, the end coordinate estimate is obtained by adding the last exon candidate mapping coordinate to the length of a missed region (if present, to account for errors at the end of the read). On the other hand, if there is no missed region then nothing is added. Using the previous example, if $r_i = [(e_1, 1 \rightarrow 41, 50 \rightarrow 90), (miss, 42 \rightarrow 61, 20), (e_2, 62 \rightarrow 85, 1 \rightarrow 24)]$, there is no missed region at the end, so nothing is added, and 24 is the end coordinate. But suppose that $r_i = [(e_1, 1 \rightarrow 41, 50 \rightarrow 90), (miss, 42 \rightarrow 61, 20), (e_2, 62 \rightarrow 85, 1 \rightarrow 24), (miss, 86 \rightarrow 90, 5)]$. Then 5 is added to 24, giving 29 as the end coordinate.

[00215] This will be important to estimate the number of unique amplicons on the list of supporting reads for all fusions. It is noted that all amplicons have the same starting point, since they all start at primer. However, the endpoint is different due to the application process.

[00216] **Mate-pair Merging**

[00217] After the k-mer mapping steps described above have been performed on both reads of a read pair, a merging algorithm tries to combine both reads into a single mapping vector. The algorithm merges all exon candidates from both mapping into a single vector (excluding mismatches) and sort it by mapping coordinates (After that, apply the algorithm for solving mapping ambiguities presented in the section above. This will merge concordant exons on both mapping and fix possible discrepancies. The result will be a single mapping vector, per read pair, that will then be used to find fusion events.

[00218] **Figure 7** graphically shows an example of this process, for a read pair mapping to three different exons, e_1 , e_2 and e_3 . The resulting mapping for both reads are

$r_1 = (e_1, 1 \rightarrow 20, 21 \rightarrow 40), (miss, 21 \rightarrow 40, 20), (e_2, 41 \rightarrow 80, 1 \rightarrow 40), (miss, 81 \rightarrow 100, 20), (e_3, 101 \rightarrow 120, 1 \rightarrow 20)$ and

$r_2 = (e_2, 1 \rightarrow 30, 11 \rightarrow 40), (miss, 31 \rightarrow 50, 20), (e_3, 51 \rightarrow 90, 1 \rightarrow 40), (miss, 91 \rightarrow 120, 30)$.

After combining and sorting by genomic coordinates, the result is the mapping

$r_s = (e_1, 1 \rightarrow 20, 21 \rightarrow 40), (e_2, 41 \rightarrow 80, 1 \rightarrow 40), (e_2, 1 \rightarrow 30, 11 \rightarrow 40), (e_3, 101 \rightarrow 120, 1 \rightarrow 20), (e_3, 51 \rightarrow 91, 1 \rightarrow 40).$

Then, the merging process is applied, resulting in $r_m = (e_1, 1 \rightarrow 20, 21 \rightarrow 40), (e_2, 1 \rightarrow 80, 1 \rightarrow 40), (e_3, 51 \rightarrow 120, 1 \rightarrow 40).$

[00219] **Junction Candidates**

[00220] After the mate-pair merging, the next step is to find possible fusion junctions on the resulting mapping vector.

[00221] A *junction candidate* is defined as a pair of consecutive exon candidates on a given read mapping, with a *miss* region in between. If each exon is from a different gene (a known driver gene and an unknown partner gene), it will be called a *fusion junction candidate*, otherwise it is a *normal junction candidate*. To find junction candidates, the algorithm scans the mapping for consecutive exon candidates, with some additional conditions described below.

[00222] Consider that for a “perfect” fusion, where the entire sequence of the driver gene exon is immediately followed by the entire sequence of a exon from a different partner gene, a particular “fusion signature” in a mapping vector is expected as shown in **Figure 8**. In Figure 8, the 5-mers of exon 1 (shown below the box labelled Exon 1, representing the driver gene) are followed by four 5-mer misses (shown above the boxes labelled Exon 1 and Exon 2), then by the 5-mers of Exon 2 (shown below the box labelled Exon 2, representing the partner gene).

$$r_1 = (e_1, r_1 \rightarrow r_2, m_1 \rightarrow m_2), (miss, (r_2+1) \rightarrow (r_3-1), k-1), (e_2, r_3 \rightarrow r_4, m_3 \rightarrow m_4)$$

[00223] where the length of the missed block is exactly $k-1$, and k is the k-mer length. This happens because the k-mers that cover the junction of the two exons are not naturally present in the DNA and therefore are not also in the k-mer index (except for spurious hits that are most likely filtered in the previous steps). There is exactly $k-1$ k-mers covering the junction, which results in the missed block with length $k-1$ in the mapping vector.

[00224] In addition, if the exons are completely present, the junction point happens exactly between the extremities closest to each other of both exons. This is referred to as a perfect signature as shown in Figure 9a. In this example, m_2 is either exon e_1 length or 1 (the latter meaning that the mapping is in reverse

orientation), and similarly m_3 should also be either exon e_2 length or 1. The exon gaps g_1 and g_2 is defined as the distance between m_2 and m_3 to their respective exon extremities. Also, the k-mer gap g_k is defined as the difference between the length of the miss block and $(k-1)$, in the junction signature. The junction signature gap (g) is then defined as the tuple (g_1, g_k, g_2) . In a “perfect” signature, $g = (0, 0, 0)$.

[00225] **Junction signature gap in the presence of sequencing errors**

[00226] If sequencing errors or SNVs happen close to the junction point, the previously discussed error correction techniques will not work, since there are no exon candidates for the same exon on both sides of the error. More specifically, because the error correction is based on having good hits around the error, on both sides. If the error is too close to the junction (less than the k-mer size), then there will be no good hits on one of the sides, closest to the junction point. Also, the k-mer misses caused by the error will result in changes in the fusion signature gap, as can be seen on **Figure 9a-9d**. Figure 9a is an example showing no errors, or a perfect signature, with $g = (0, 0, 0)$. Figure 9b and 9c shows examples of one error 2 bps from the junction point, causing two k-mers to miss, increasing both the exon gap g_1 and the k-mer gap by two. Figure 9d is an example showing errors on both exons, causing $g_1 = 5$, $g_2 = 2$, and the k-mer gap $g_k = g_1 + g_2 = 7$.

[00227] In the presence of errors, the exon gaps g_1 and g_2 increase by the same amount as the k-mer gap g_k , leading to the identity $g_k = g_1 + g_2$. This fact points to a gap correction algorithm, where g_k can be reduced by one, also reducing either of the exon gaps (the largest), until either g_k is zero, or both g_1 and g_2 is zero. The gap correction algorithm is expressed as follows below, where sub-parts a and b are iterative until the listed conditions are met.

Gap correction algorithm:

1. While $g_k > 0$
 - a. If $g_1 > 0$ then
 - i. $g_k = g_k - 1$
 - ii. $g_1 = g_1 - 1$
 - b. Else If $g_2 > 0$ then
 - i. $g_k = g_k - 1$
 - ii. $g_2 = g_2 - 1$

[00228] **Clustering fusion junctions and recruiting supporting reads**

[00229] For all junction candidates found in a read pair, if there is a single *fusion* junction candidate where the k-mer gap is smaller than a given threshold (default is $k + 1$), then the read pair will be added to a list of supporting reads for this fusion junction.

[00230] It should be noted that the previously described gap correction reduces the gap in some cases where there are errors close to the junction, but it does not remove all gaps. Gaps can still be large if there are different sections of the read mapping to different parts of the genomes, and such gaps cannot be corrected, but if they are still large it mostly means a mapping error or random mapping and will not be considered a fusion. This mostly happens due to sequence similarities/repeats on different parts of the genome, meaning that a read can map to different regions but does not necessarily means that the read comes from that region, is just a mismatch due to seq similarity.

[00231] If a read has no fusion junction candidates but it has normal junction candidates, it will be counted as a normal junction for that particular gene. This is useful to estimate the ratio between the number of fusion supporting reads and normal reads for all fusion calls.

[00232] After all read pairs are processed, all fusion junction candidates are clustered based on the genomic coordinates of the junction point, with a tolerance of 50 bps, meaning that junctions that are less than 50 bps apart are considered to be the same fusion.

[00233] **Fusion Candidate Calling**

[00234] After the junction clustering step, a list of fusion candidates for the sample results, each with a number of supporting read pairs. For each fusion candidate, several metrics are calculated based on the supporting read pairs, such as the read depth, the number of unique amplicon ends, the estimated length of the fusion transcript, the ratio between fusion and normal reads, among other metrics. The estimated length is based on the average length of the read mapping length of all supporting reads for the fusion.

[00235] **Local assembly of fusion transcripts**

[00236] For each called fusion, all supporting reads for that fusion are assembled using *de-Bruijn graph-based* assembly. After the assembly, *all reads* of the sample are mapped against all assembled fusion transcripts. This is done as follows:

1. For each fusion:
 - a. a de-Bruijn graph (https://en.wikipedia.org/wiki/De_Bruijn_graph) with the k-mers of all supporting reads is built.
 - b. The graph is simplified with bubble correction and tip removal algorithms, to account for sequencing errors.
 - c. The graph is traversed, starting with the node that has the k-mer of the gene-specific primer that captured the fusion, and a number of contigs is extracted.
 - d. The fusion calling algorithm is rerun on the output contigs. The longest contig that has the expected fusion junction is considered the correct assembled fusion transcript.
2. A reference file with all fusion transcripts is created, and the original FASTQ files from the sample are mapped to this reference file using BWA (<http://bio-bwa.sourceforge.net>).
3. New metrics are added to the fusion calls, such as the assembled transcript length, the number of total mapped reads and unique mapped reads, and the estimated number of unique transcript ends from the mapping output.

[00237] **Confidence Score Assignment**

[00238] The fusion calling algorithm is lenient since it looks for reads with mapping regions in two different genes. Although there are some filtering steps, there is at this point a high number of artifact calls. Therefore, all called fusions will be scored based on the corresponding metrics and depending on the resulting score will be either filtered and discarded, or receive a confidence label, from Low, Medium or High. An overview of the confidence scoring method can be seen on **Figure 10**.

1. Artifact filter: The first step checks if the orientation of both fusion partners is concordant, and if exactly one of the partners is a gene targeted by the assay. If either of the conditions is not satisfied, the fusion is filtered as an **artifact**. Fusion junction candidates in the merged mapping vectors having a pair of consecutive exon blocks from different genes are called fusion partner genes.

2. Supporting fusions: Even in a normal cell, because of alternative splicing, the same gene can have different transcripts (isoforms), meaning different variants of the same gene that skip/include different exons. This means that the same fusion, caused by a breakpoint in the genome, can also have different variants with the same two partner genes, but between different exons. For example, if it is seen in the same patient a fusion EML4;7 – ALK;20 (between EML4 at exon 7 and ALK exon 20) and a fusion EML4;6 – ALK; 20. Instead of considering two different fusions, the one with the most supporting reads is considered the main fusion, and all others are supporting fusions, meaning that the supporting fusion reads are added to the main fusion. This in turn can improve the main fusion score.
3. Metrics filter: Several metrics are checked each with a minimum threshold value that the fusion has to pass. Failing on any of the metrics will label the call as **filtered**.
4. High confidence: If a fusion call passed the previous steps and has read depth and number of unique reads higher than a given threshold, it will be labelled **high confidence**.
5. The remaining fusions will be processed by a scoring function that will be explained below. Depending on the score, a fusion can either be considered **medium** or **low confidence**, or be **filtered**.

[00239] **Fusion scoring function**

[00240] For a given fusion j , and a set of metrics, the fusion score $s(j)$ is given by

$$s(j) = \sum_i^n w_i \cdot \frac{m_{ij} - \mu_i}{|\mu_i|}, \text{ for } i \in \text{metrics}$$

where m_{ij} is the value of the metric i for the fusion j , w_i is the weight associated with metric i , and μ_i is the *expected mean* of metric i , calculated from a control dataset with known fusions.

[00241] The set of metrics can be chosen from several different metrics, with varying results. Best results in our tests were obtained by using four metrics: read depth, number of unique reads, fusion transcript length, and the fusion/normal read

depth ratio. The fusion/normal read depth ratio is the total number of normal junctions for the same gene as the target gene for the fusion.

[00242] By looking at the scores of fusions in a training set, two thresholds t_1 and t_2 are determined, such that fusions with score $> t_1$ are labelled **medium confidence**, while fusions with score $< t_2$ are **filtered** and discarded. The remaining fusions receive a **low confidence** label. **Figure 11** shows the scores for fusion calls made on a control dataset with known fusions. The thresholds were defined as $t_1 = 0$ and $t_2 = -1$. Scores that are categorized as medium and low confidence are subject to review for further consideration of actions to be taken, including additional analysis. For each confidence level, different performance metrics can be determined, in terms of sensitivity and specificity. For example, low confidence means that there is a significant chance that the fusion might be an artifact, so an orthogonal validation method should be used for confirmation. Conversely, high confidence calls are $>99\%$ guaranteed to be correct, so no validation should be needed. Medium is an in-between confidence, so it's up to the discretion of who is analyzing the results to determine what action to take. If the result is expected, such as a common fusion for a specific cancer type, then no confirmation might be needed.

[00243] Following is a summary of the fusion pipeline method. In a first aspect, the fusion pipeline method is a computer implemented method for identifying from a plurality of read pairs R1 and R2 of a sequenced sample with inserts derived from a set of PCR amplified adapter ligated DNA molecules, any target gene fused with any other DNA sequence which are potentially clinically actionable. The method includes

generating a k-mer index of predetermined length k from a reference genome or a transcriptome;

mapping each sequencing read from the plurality of read pairs to exons of the reference genome with the k-mer index to identify corresponding exon regions of the sequencing read, and generating for each sequencing read a mapping vector including one or more blocks, where each block is either an exon block defined by

a range of consecutive k-mer matches and their corresponding locations on an exon of the reference genome, or a mismatch block defined by

a length of consecutive k-mer mis-matches;

identifying exon blocks with overlapping consecutive k-mer matches in each of the mapping vectors, and merging the same exon blocks with adjustment of mapping coordinates or retaining the exon block having predetermined high priority criteria to resolve mapping ambiguities;

combining both the mapping vectors corresponding to each read pair and sorting by genomic coordinates to provide intermediate mapping vectors, and merging any exon blocks with the same respective exons with adjustment of mapping coordinates to generate merged mapping vectors;

identifying fusion junction candidates in the merged mapping vectors by searching for a pair of consecutive exon blocks from different genes, called fusion partner genes, with a mismatch block in between, and recording the corresponding read pairs for each of the identified fusion junction candidates;

clustering all the fusion junction candidates based on genomic coordinates of junction points of the pairs of consecutive exons from different genes where junctions closer than a predetermined number of base pairs are considered to be the same fusion junction candidate, and calculating a first predetermined set of metrics based on the read pairs including read depth;

assembling a fusion transcript for each fusion junction candidate cluster using the corresponding recorded read pairs to generate assembled transcripts, and calculating a second predetermined set of metrics from the assembled transcripts;

mapping the plurality of read pairs to the fusion transcripts, and calculating a third predetermined set of metrics based on the mapping output including a number of mapped reads; and

categorizing any fusion transcript as being a high confidence reportable fusion when

the orientation of both fusion partners is concordant,

exactly one of the partners is a gene targeted by the assay,

minimum thresholds for specific metrics of the first, second and third predetermined set of metrics are met, and

the number of mapped reads and the read depth exceed predetermined thresholds.

[00244] In a first embodiment of the first aspect, exon blocks having less than a minimum mapped bps are filtered out, and the minimum mapped bps is 24 bps.

[00245] In a second embodiment of the first aspect, each of the mapping vectors is represented by the expression (e_n, read_st→read_end, map_st→map_end) for each exon block and (miss, miss_st→miss_end, len) for each mismatch block, where e_n is the exon unique identifier, read_st and miss_st are the starting base pair on the sequencing read, read_end and miss_end are the ending base pair on the sequencing read, map_st is the starting mapped region on the exon for read_st, map_end is the ending mapped region on the exon for read_end, miss is a status identifier for a mis-match, and len is the bps length of the mismatch block.

[00246] In a third embodiment of the first aspect, the method can include error correcting the mapping vectors to remove mismatch blocks interposed between two exon blocks with the same exon. In this embodiment, error correcting includes removing the length of consecutive k-mer mis-matches of the mismatch block interposed between a first range of consecutive k-mer matches of a first exon block and a second range of consecutive k-mer matches of a second exon block of the same exon, when a continuation of the first range by the length of consecutive k-mer is sequentially followed by a beginning of the second range. More specifically, error correcting to remove a mismatch block between two consecutive exon blocks with the same exon includes determining when read_end + len + 1 for the first exon block = read_st for the subsequent exon block to remove the mismatch block. Alternately, error correcting to remove the length of consecutive k-mer mis-matches includes determining when map_end + len + 1 for a first exon = map_st for a subsequent exon to remove the length of consecutive k-mer mis-matches.

[00247] In the second embodiment, identifying fusion junction candidates includes executing an error correction applied on k-mer gaps g_1, g_2 and g_k of the fusion junction candidates when g_k > 0, where

g_1 is a distance between map_end of the first exon block to the corresponding exon extremity,

g₂ is a distance between map_{st} of the second exon block to the corresponding exon extremity,
 g_k is the difference between the k-mer length k and the mismatch block len,
 and the k-mer gap correction steps include

```

while gk > 0
  a.   if g1 > 0 then
        i.   gk = gk - 1
        ii.  g1 = g1 - 1
  b.   else If g2 > 0 then
        i.   gk = gk - 1
        ii.  g2 = g2 - 1
  
```

[00248] Alternately for the second embodiment, the predetermined high priority criteria include

an exon block with a larger mapping range than the other,
 an exon block where its exon comes from an assay targeted gene,
 an exon block matching a list of preferred candidates, or
 when both exons are from the same gene, selecting the exon with highest priority, calculated from sorting the corresponding reference genome transcripts for each exon in descending order of length.

[00249] If none of the predetermined high priority criteria is met, then the exon block having the largest exon gap is removed, where an exon gap is a distance between map_{end} of a first exon and map_{st} of a subsequent exon to their respective exon extremities.

[00250] In this embodiment, assembling the fusion transcript includes the following steps for each fusion junction candidate cluster:

- a. constructing a de-Bruijn graph with the k-mers of all cluster reads,
- b. executing bubble correction and tip removal algorithms, to correct sequencing errors,
- c. determining a number of contigs by traversing the de-Bruijn graph starting with a node that has the k-mer of the gene-specific primer that captured the fusion, and
- d. determining a longest contig as being a true assembled fusion transcript.

[00251] Here, the second predetermined set of metrics includes an assembled fusion transcript length, and the third predetermined set of metrics includes the number of unique transcript ends determined from the mapping output.

[00252] In this embodiment, categorizing further includes

calculating a fusion score for each assembled fusion transcript based on any number of metrics of the first, second and third predetermined set of metrics, and the scoring function is calculated using the expression

$$s(j) = \sum_i^n w_i \cdot \frac{m_{ij} - \mu_i}{|\mu_i|}, \text{ for } i \in \text{metrics}$$

, where m_{ij} is the value of the metric i for the assembled fusion transcript j , w_i is the weight associated with metric i , and μ_i is the *expected mean* of metric i , calculated from a control dataset with known fusions,

comparing the fusion score to a first predetermined threshold and a second predetermined threshold, and

categorizing the assembled fusion transcript as medium confidence if the fusion score is greater than the first predetermined threshold,

categorizing the assembled fusion transcript as low confidence if the fusion score is less than the first predetermined threshold but greater than the second predetermined threshold, or

categorizing the assembled fusion transcript as filtered if the fusion score is less than the second predetermined threshold, and

outputting a result where low and medium confidence are subjected to orthogonal testing.

[00253] The metrics used for the scoring function include the read depth, the number of unique reads, the assembled fusion transcript length, and a ratio of read depth between fusion junction candidates and normal junction candidates with the same gene as the fusion target gene.

[00254] In a fourth embodiment of the first aspect, the predetermined number of base pairs to cluster junction candidates is 50 or less.

[00255] In a fifth embodiment of the first aspect, the method further includes determining a unique amplicon end by retrieving the last exon block mapping coordinate `map_end` and adding to it the length of the last mismatch block for each of

the R2 reads. In this embodiment, identifying fusion junction candidates further includes identifying normal junction candidates in the merged mapping vector by searching for pairs of consecutive exon blocks from a same gene with a mismatch block in between, and recording the corresponding read pairs for each of the identified normal junction candidates. In this embodiment, the first predetermined set of metrics includes a number of unique amplicon ends, an estimated length of the fusion transcript, and a ratio between fusion junction candidates and normal junction candidates. The estimated length of the fusion transcript is determined by the average mapped length of all the read pairs in the corresponding junction candidate.

[00256] In a sixth embodiment of the first aspect, categorizing further includes categorizing any assembled fusion transcript as being an artifact if either the orientation of the fusion partners is not concordant or both partners are genes not targeted by the assay, or categorizing any assembled fusion transcript as being a supporting fusion if another assembled fusion with more reads is found on the same partner genes.

[00257] **Example 2**

[00258] An RNA-based amplicon assay for the sensitive and specific partner agnostic detection of fusions in poor quality samples involving 18 clinically relevant solid tumor driver genes is described. This assay performs well on fragmented FFPE RNA from clinical specimens with no prior knowledge of the fusion partner gene at low sequencing depths. Target enrichment is performed first to minimize loss of template information during adapter ligation. Then a random population of RNA splints act as scaffolds for annealing a universal adapter. After a second round of amplification and indexing, up to 21 clinical sample libraries are sequenced in multiplex. The actionable gene fusion content, cloud-based analysis, and integrative evidence algorithm provides accessible and focused results to ordering clinicians. These results are presented in a clinical reporting platform that enables rapid interpretation and reporting of clinical findings.

[00259] To address challenges associated with poor quality RNA, pre-sequencing QC steps were employed. Imposing a qPCR CT threshold value of 30 for a housekeeping gene was found to be effective at screening samples for assay performance. Excluding low confidence fusion calls also improved specificity. With these filters, a sample rejection rate for archival clinical FFPE specimens with inputs as low as 10 ng of 32% (32 of 101) was observed. Despite having qPCR CT values

above 30, nine filtered samples still yielded correct fusion calls. Sample rejection was minimized by testing all samples and performing reflex testing with an alternative orthogonal DNA or antibody-based method on positive fusion calls in sub-threshold samples. This is advantageous due to the multi-target nature of NGS testing compared with immunohistochemical methods that are often single-target.

[00260] Successful performance was observed in archival FFPE specimens less than two years old down to 20 ng of input RNA, an amount obtainable from a single FFPE section. In an older cohort of up to five years of age, successful detection was observed with 100 ng of template. In this example, increasing template input beyond 100 ng did not improve detection of gene fusions in low quality samples.

[00261] **Materials and Methods**

[00262] **Gene and exon target selection:**

[00263] 18 driver genes (*ALK*, *BRAF*, *CSF1*, *EGFR*, *ERG*, *FGFR1*, *FGFR2*, *FGFR3*, *MET*, *NRG1*, *NTRK1*, *NTRK2*, *NTRK3*, *PDGFB*, *PPARG*, *RAF1*, *RET*, *ROS1*) were targeted based on their presence and clinical actionability in solid tumors (**Table 10**).

Table 10: Target genes associated with cancer type along with a non-limiting list of known/reported partner genes.

Target gene	List of known fusion partner genes	Tumour types
ALK	EML4, TPM3, STRN, KIF5B, TFG, HIP1, DCTN1, VCL, KLC1, C2orf44, CDK2	Lung, Thyroid, Skin, Kidney, Large intestine
BRAF	SLC45A3, TRIM24, KIAA1549, AKAP9, SND1, RNF130, TMPRSS2, DSTYK, MACF1, CUX1, DLG1, AGAP3	CNS, Thyroid, Pancreas, Prostate, Colon
CSF1	COL6A3	Connective tissue
EGFR	SEPT14, CHODL, VSTM2A	Lung, CNS
ERG	TMPRSS2, SLC45A3, NDRG1, EWSR1, FUS	Prostate, Bone
FGFR1	TACC1, PLAG1, ZNF703, ADAM18, LETM2, BAG4	CNS, Lung, Thyroid
FGFR2	CCAR2, FRK	CNS, Lung, Thyroid
FGFR3	TACC3, BAIAP2L1	CNS, Lung, Thyroid, Urothelial
MET	CAPZA2, KIF5B, LAMA3, MIR548F1, PTPRZ1, ST7, TFG, DYNC111, FAM3C, TFG, WNT2	Lung, CNS, Breast, Thyroid
NRG1	GPRC5A, SDC4, CD74	Lung, Stomach
NTRK1	TPM3, LMNA, TFG, TP53, CEL, SQSTM1, MPRIP, CD74, TPR, ARHGEF2, IRF2BP2, RFWD2, NFASC, BCAN, TP53, CTSC, RABGAP1L, CHTOP, AFAP1, IGFBP7, GRPA1, LRRC71, SSBP2, MRPL24, PLEKHA6, PDE4DIP, MIR548F1, SCYL3	Thyroid, Breast, Kidney, Large Intestine, Salivary gland, Pancreatic, Skin
NTRK2	NACC2, QKI, TRIM24, PAN3, SQSTM1, TRAF3, STRN, ETV6, BCR, TLE4, DAB2IP, VCL, AGBL4, AFAP1	Thyroid, Breast, Kidney, Large intestine, Salivary gland
NTRK3	ETV6, EML4, TPM4, TFG, MYO5A, BTBD1, LYN, RBPMS,	Thyroid, Breast, Pancreatic,

	UBE2R2, HNRNPA2B1	Colon
PDGFB	COL1A1	Soft tissue
PPARG	PAX8	Thyroid
RAF1	ESRP1, HAC11, SRGAP3	Prostate, Pancreas, CNS
RET	CCDC6, KIF5B, NCOA4, PRKAR1A, GOLGA5, ERC1, KTN1, TRIM24, TRIM33, RRB1	Lung, Thyroid, Skin, Colon
ROS1	SLC34A2, CD74, EZR, SDC4, LRIG3, TPM3, GOPC, CEP85L, CLTC	Lung, Stomach, CNS

[00264] In total, 146 exons were selected based on biologically relevant configurations found in the literature and fusion databases ^{1,3,26} (Table 11).

Table 11: Target gene exon list. Target gene exons are listed for each reference transcript. Some exons described in the literature are present only in alternative transcripts and therefore are listed separately. The housekeeping gene *TUBB* is included for control analyses.

Gene	Transcript ID (hg19)	Target Exon	Orientation
<i>ALK</i>	ENST00000389048	2, 10, 16-22	3'-fusion
<i>BRAF</i>	ENST00000646891	2, 6-12, 15	3'-fusion
<i>BRAF</i>	ENST00000496384	19	3'-fusion
<i>CSF1</i>	ENST00000329608	6-8	3'-fusion
<i>EGFR</i>	ENST00000275493	1-3, 7-9, 11, 13, 16, 20-21, 24-25, 27	5'-fusion
<i>EGFR</i>	ENST00000275493	8, 23	3'-fusion
<i>ERG</i>	ENST00000398919	2-11	3'-fusion
<i>FGFR1</i>	ENST00000447712	1-10, 14, 17	5'-fusion
<i>FGFR2</i>	ENST00000457416	2, 8-10, 13, 16-17	5'-fusion
<i>FGFR3</i>	ENST00000440486	2-4, 8-10, 16-18	5'-fusion
<i>MET</i>	ENST00000318493	2-6, 13-17, 21	3'-fusion
<i>NRG1</i>	ENST00000519301	1	3'-fusion
<i>NRG1</i>	ENST00000287842	1-8	3'-fusion
<i>NTRK1</i>	ENST00000530298	2	3'-fusion
<i>NTRK1</i>	ENST00000524377	2, 7-13	3'-fusion
<i>NTRK2</i>	ENST00000277120	10-17	3'-fusion
<i>NTRK3</i>	ENST00000394480	7, 12-18	3'-fusion
<i>PDGFB</i>	ENST00000331163	2	3'-fusion
<i>PPARG</i>	ENST00000397010	2	3'-fusion
<i>PPARG</i>	ENST00000287820	2, 7	3'-fusion
<i>RAF1</i>	ENST00000251849	2, 8-10	3'-fusion
<i>RET</i>	ENST00000355710	2-4, 7-13	3'-fusion
<i>ROS1</i>	ENST00000368508	31-37	3'-fusion
<i>TUBB</i>	ENST00000327892	4	N/A

[00265] The tubulin gene *TUBB* was also included as a non-fusion housekeeping gene target to assess assay performance in all sample types.

[00266] **Specimens and samples:**

[00267] Universal normal human RNA (catalog R4234565-1, Amsbio LLC, Abingdon, United Kingdom) was used as a fusion-negative analytical sample. The 18 fusion RNA mix (SeraSeq catalog 0710-0497, LGC Clinical Diagnostics/SeraCare, Milford, Massachusetts, USA) was selected as a fusion-positive analytical sample. Double-stranded 500 bp geneblocks (gBlocks) (Integrated DNA Technologies (IDT), Coralville, Iowa, USA) were designed and centered around exonic junctions were also used as positive testing controls ^{1,26}. A 10 pg blend was spiked into 1 mg of the fusion-negative sample and used in runs (referred to as “74 Fusion-Positive”). Testing and validation specimens included FFPE cell pellets (catalog numbers: 3020-0130 H2228, 3020-1130 H596, 3020-1430 HCC78, 3130-0130 U118MG, 3060-1130 A431, Amsbio LLC, Abingdon, United Kingdom) and FFPE clinical samples sourced from commercial vendors with sample collection dates ranging from 2013-2019 (Reprocell, Beltsville, Maryland, Accio Biobank Online/Tissue for Research, United Kingdom, iSpecimen, Lexington, Massachusetts, USA, Trans-Hit Biomarkers Inc., Azena, USA, Amsbio LLC, Abingdon, United Kingdom).

[00268] **RNA Extraction and electrophoresis:**

[00269] RNA was extracted from FFPE samples using a Maxwell RSC instrument and RNA FFPE kit (Promega, Madison, Wisconsin, USA). RNA yields were quantitated using the Qubit RNA kit (Invitrogen, Waltham, Massachusetts, USA). RNA DV200 was calculated using the Bioanalyzer RNA Pico assay (Agilent, Santa Clara, California, USA) as per manufacturer protocols.

[00270] **cDNA Synthesis and quantitative PCR:**

[00271] First strand cDNA was generated with random hexamers using the Superscript IV system (Invitrogen, Waltham, Massachusetts, USA). Second strand cDNA synthesis was generated using the New England BioLabs (NEB) ultra second strand synthesis module (NEB, Ipswich, Massachusetts, USA). The double stranded cDNA allows reverse or forward enrichment primers to be designed to enrich for 5'-partner-target-3' or 5'-target-partner-3' configurations. Using 1.8X AMPure XP beads (Beckman Coulter, Pasadena, California, USA), the purified cDNA was assessed for amplifiability with a PreSeq RNA QC assay as per manufacturer's protocols (Archer Dx, Inc., Boulder, Colorado, USA).

[00272] **Oligonucleotide design and dilution:**

[00273] Nested enrichment primers and amplification primers were designed with Primer3²⁷. One enrichment and one amplification primer per exon was selected using a custom method based for multiplex optimization based on minimization of *in silico* free energy. Blocking oligos are the reverse-complement of enrichment primers. Enrichment primers, amplification primers, and blocking oligos were synthesized (Integrated DNA Technologies (IDT), Coralville, Iowa, USA) and pooled at 2 uM individual final concentration. The adapter and amplification primers contain 5' Nextera XT extension sequences for creation of a Nextera XT sequencing library (Illumina, San Diego, California, USA).

[00274] **Target Enrichment:**

[00275] Double stranded cDNA was amplified with linear PCR using the Qiagen Multiplex PCR kit (Qiagen, Hilden, Germany) and a 2 uM multiplex enrichment primer mix in a 25 ul reaction. Thermal cycler program: 95°C 15 minutes, 8 cycles of (94°C 30 seconds, 60°C 3 minutes, 72°C 30 seconds), 4°C hold. Enrichment product is purified with 1.8X AMPure XP beads (Beckman Coulter, Pasadena, California, USA).

[00276] **Adapter ligation:**

[00277] Adapter and RNA splint pool were hybridized in annealing buffer (10mM Tris pH 8, 1mM EDTA, 1M NaCl) with thermal cycler program: 85°C 2 minutes, 80 cycles of 30 seconds with a reduction of -1°C per cycle, 4°C hold. Enrichment products and blocking oligos were hybridized in annealing buffer with thermal cycler program: 95°C for 2 minutes, 70 cycles of 30 seconds with a reduction of -1°C per cycle, 25°C hold. The hybridized duplexes were ligated using SplintR Ligase (New England BioLabs, Ipswich, Massachusetts, USA) with thermal cycler program: 25°C 8 hours, 9 cycles of (6 minutes, -1°C per cycle), 16°C 30 minutes, 65°C 2 minutes, 4°C hold. Ligation product was purified with 1.8X AMPure XP beads (Beckman Coulter, Pasadena, California, USA).

[00278] **Target Amplification:**

[00279] The ligation product was amplified in KAPA HiFi hotstart readymix reaction (Roche, Basel, Switzerland) with a multiplex amplification primer pool, Nextera i7 R primer (targeting adapter) with thermal cycler program: 95°C 15 minutes, 8 cycles of (94°C 30 seconds, 60°C 3 minutes, 72°C 30 seconds), 4°C hold.

Amplification product was purified with 1.8X Ampure XP beads (Beckman Coulter, Pasadena, California, USA).

[00280] **Library Construction, Normalization and Sequencing:**

[00281] A sequencing library was generated by PCR using the Nextera XT system (Illumina, San Diego, California, USA) using IDT for Illumina UDI adapters (Illumina, San Diego, California, USA) with KAPA HiFi PCR mix (Roche, Basel, Switzerland) using thermal cycler program: 95°C 30 seconds, 15 cycles (95°C 30 seconds, 55°C 30 seconds, 72°C 30 seconds), hold 4°C. Library product was purified with 0.7X AMPure XP beads (Beckman Coulter, Pasadena, California, USA). The libraries were quantified with a Qubit BR DNA kit (ThermoFisher Scientific, Waltham, Massachusetts, USA) and normalized to 20 nM. A pooled 4nM library is denatured, diluted, and sequenced with 5% PhiX using 2x150 cycle paired-end sequencing on a MiSeq™ instrument (Illumina, San Diego, California, USA).

[00282] **Computational Methods:**

[00283] Normal and cancer samples contain many chimeric reads that must be evaluated as biologically relevant fusions, artifacts, or alternative transcripts^{1,3}. The sliding window *k-mer* based approach with two stage alignment for detection of atypical gene exons fusions paired with a classifier is executed on the input sequence data, which can be provided in FASTQ files. The following steps of the method correspond to the steps discussed earlier.

[00284] **Read Filtering:**

[00285] Reads are loaded from the FASTQ files, filtering out-of-target/misprimed reads by labeling read pairs with primer identity based on the presence of the expected primer sequence at the start of the R1 read with maximum *hamming distance* of 1 from all the primer sequences in the assay manifest file. Labeled read pairs are aligned with the expected target sequence up to exon boundary. Alignments result in read pairs being assessed as on-target or off-target (when the primer binds to a different region) with tolerance for sequencing errors. Off-target reads are filtered and the remaining are used as an input for k-mer indexing.

[00286] **K-mer indexing and read pseudo-alignment:**

[00287] On-target read sequences are mapped to the reference human genome hg38 using a k-mer mapping algorithm. A k-mer index is created from on-

target reads where a k-mer is a substring of length k of a given DNA sequence and is a structure where information about a given k-mer from a set of k-mers can be accessed (see **Figure 5**). The aim is to identify the corresponding exon regions of the sequencing read, to then find reads that map to two different genes (the driver gene in the known sequence, and a partner gene in the unknown sequence, see **Figure 8**), indicating a possible fusion supporting read. For a “perfect” fusion, where the entire sequence of the driver gene exon is immediately followed by the entire sequence of a exon from a different partner gene, we expect exactly k-1 k-mers covering the junction and with no discrepancies (gaps) between the mapped and expected exon lengths.

[00288] **K-mer mapping and finding exon candidates**

[00289] Given a k-mer from a sequencing read, a k-mer hit is the position of this k-mer in the read and the mapping information of this k-mer from the k-mer index. If a k-mer is not present in the index, the corresponding mapping information is called a miss. Repeating this process for all k-mers in a read results in a vector of k-mer hits, representing the original read. Consecutive hits for the same unique exon match are then aggregated. Consecutive misses are also merged. The combined misses are called a missed block. Exon candidates with a small range mapped are filtered, with a minimum mapped requirement of 24 bps. At the end of the process, a read is then represented as a mapping vector composed of consecutive exon candidates and misses.

[00290] After this step, additional steps of sequencing error/SNP correction and resolving mapping ambiguities as previously described are executed.

[00291] **Unique amplicon determination:**

[00292] This step is similar to the previously described step of amplicon end estimation. To estimate the number of unique amplicons on the list of supporting reads for all fusions, the aligned endpoints of the sequencing read R2 are used. Since the assay is based on an anchored linear amplification step, different sequenced amplicons will have different lengths, and therefore different endpoints (or starting R2 sequences). From the R2 read mapping, the end coordinate estimate is obtained by adding the last exon candidate mapping coordinate to the length of a missed region (if present, to account for errors at the end of the read).

[00293] **Mate-pair Merging**

[00294] After the k-mer mapping steps described above have been performed on both reads of a read pair, a merging algorithm combines exon candidates from both reads into a single mapping vector sorted by mapping coordinates. The end result is a single mapping vector, per read pair, used to find fusion events.

[00295] Junction Candidates

[00296] The merged mapping vectors are searched for junction candidates, defined as a pair of consecutive exon candidates on a given read mapping, with a miss region in between. If each exon is from a different gene (a known driver gene and an unknown partner gene), it will be called a fusion junction candidate, otherwise it is a normal junction candidate. Due to possible sequencing errors or SNVs close to the junction point, the previously described gap correction algorithm can be executed.

[00297] Clustering fusion junctions and recruiting supporting reads

[00298] For all junction candidates found in a read pair, if there is a single *fusion* junction candidate where the k-mer gap is smaller than a given threshold (default is $k + 1$), then the read pair will be added to a list of supporting reads for this fusion junction. If a read has no fusion junction candidates but it has normal junction candidates, it will be counted as a normal junction for that particular gene. This is useful to estimate the ratio between the number of fusion supporting reads and normal reads for all fusion calls. After all read pairs are processed, all fusion junction candidates are clustered based on the genomic coordinates of the junction point, with a tolerance of 50 bps, meaning that junctions that are less than 50 bps apart are considered to be the same fusion.

[00299] **Fusion Candidate Calling:**

[00300] After the junction clustering step, for each fusion candidate, several metrics are calculated based on the supporting read pairs, such as the read depth, the number of unique amplicon ends, the estimated length of the fusion transcript, the ratio between fusion and normal reads, among other metrics.

[00301] **Local assembly of fusion transcripts:**

[00302] For each fusion, all its supporting reads are assembled using a *de-Bruijn graph-based* assembly. After the assembly, *all reads* of the sample are

mapped against all assembled fusion transcripts to identify additional supporting reads that were filtered during the k-mer matching process.

[00303] **Fusion scoring function**

[00304] For a given fusion j , and a set of metrics, the fusion score $s(j)$ is given by

$$s(j) = \sum_i^n w_i \cdot \frac{m_{ij} - \mu_i}{|\mu_i|}, \text{ for } i \in \text{metrics}$$

[00305] where m_{ij} is the value of the metric i for the fusion j , w_i is the weight associated with metric i , and μ_i is the *expected mean* of metric i , calculated from a control dataset with known fusions. The set of metrics selected were read depth, number of unique reads, fusion transcript length, and the fusion/normal read depth ratio.

[00306] **Confidence Score Assignment**

[00307] The fusion calling algorithm looks for reads with mapping regions in two different genes and must differentiate true fusions from a high number of artifact calls. All called fusions are scored based on a set of calculated metrics, and depending on the resulting score are either filtered, or receive a confidence label, from low, medium or high (see **Figure 10**). The fusion score is calculated using a set of metrics, weights, and thresholds calculated from a control dataset with known fusions with values defined as $t_1 = 0$ and $t_2 = -1$ (see **Figure 11**). The set of metrics selected were read depth, number of unique reads, fusion transcript length, and the fusion/normal read depth ratio. Fusions with score $> t_1$ are labeled medium confidence, while fusions with score $< t_2$ are filtered. The remaining fusions receive a low confidence label.

1. Artifact filter: The first step checks if the orientation of the fusion partners is concordant, and if one of the partners is a gene targeted by the assay. If either of the conditions is not satisfied, the fusion is filtered as an **artifact**.
2. Supporting fusions: secondary fusion calls with the same exon junction between the same two genes are gathered as supporting of a primary fusion call with the highest support.

3. Metrics filter: Several metrics are checked each with a minimum threshold value that the fusion has to pass. Failing on any of the metrics will label the call as **filtered**.
4. High confidence: If a fusion call passed the previous steps and has read depth and number of unique reads higher than a given threshold, it is labeled **high confidence**.
5. The remaining fusions are processed by a *scoring function* for reporting purposes. Depending on the score, a fusion is considered **medium** or **low confidence**, or be **filtered**.

[00308] **Results**

[00309] **A method for enrichment and amplification of target gene exons**

[00310] This example demonstrates a partner-agnostic NGS method and algorithm for targeted exon enrichment and detection of gene fusions in clinical FFPE tumors as shown in **Figure 17**, where the right side flowchart of Figure 17 “D: Fusion Analysis” is the same as Figure 4. Starting with FFPE tumor RNA template, the assay interrogated 146 exons in 18 driver genes for gene fusions. The assay employed random hexamer priming of cDNA to support fragmented and degraded FFPE RNA template. Double stranded cDNA was enriched for target exons using multiplex linear amplification from exon junction adjacent primers (**Figure 18**). The assay enabled target amplification without knowledge of the gene fusion partner by specifically ligating a universal adapter to the 3' ends of enriched amplicons using a population of RNA splints with an adapter specific region and a random 10mer region along with the efficient Chlorella virus PBCV-1 DNA ligase²⁸ (**Figure 19**). Target amplification was performed using multiplex PCR using an adapter specific primer and an internally nested exon-specific amplification primer. Libraries were then generated, normalized, pooled, and sequenced on an Illumina MiSeq™ using paired-end 2 x 150 bps, yielding target exon sequence and junction coverage from the P5 read direction, and partner gene coverage from the P7 read direction (**Figure 20**).

[00311] One feature of the assay is the ability to detect unique template molecules based on the template cDNA molecule sequence since reads originating from the adapter have variable start site positions. To assess this, a normal human universal RNA template (fusion-negative) was used and the read coverage of the housekeeping gene *TUBB* (tubulin) was determined. *TUBB* exons 1-3 were targeted for amplification from within exon 4. Exons 1-3 had an average read depth of 4,279X

while the downstream regions (609-2,670) had an average read depth of 3X indicating effective and specific enrichment (**Figure 21**). From a sequencing run of 785,777 normal RNA derived reads, there were 12,588 *TUBB* gene transcript aligned reads, with 303 unique start sites.

[00312] **Detection of gene fusions in reference samples**

[00313] A custom analysis pipeline to identify on-target reads and the subset of candidate fusion reads was developed. On-target reads were collapsed into on-target unique reads based on adapter-read start site position relative to the aligned transcript. On-target reads with atypical exon organizations were classified as candidate fusion reads and scored. To establish specificity of read classification, a no-template control (NTC), a fusion-negative universal tissue normal human RNA control, and an 18 fusion positive control were assayed. The NTC contained no on-target reads, in contrast with the fusion-negative and the fusion-negative samples (**Figure 22**). On-target unique fusion reads were only identified in the 18 fusion-positive sample (median 12% from five runs), demonstrating enrichment of the initial ~3,000 copies per fusion target.

[00314] To establish the detection of fusions involving all 18 targeted genes 74 synthetic 500 bp dsDNA fusion templates ('geneblocks') were synthesized (**Table 12**).

Table 12: Geneblock details. Called fusions all had junctions between end of upstream exon and start of downstream exon. Geneblocks not called had junctions between at least one exon starting within the exon or were reciprocal fusions that are not expected to be amplified.

Type	Detection Status	5'-gene	5'-exon	3'-gene	3'-exon
Fusion	Yes	<i>CDK2</i>	exon 7	<i>ALK</i>	exon 19
Fusion	Yes	<i>EML4</i>	exon 7	<i>ALK</i>	exon 18
Fusion	Yes	<i>EML4</i>	exon 6	<i>ALK</i>	exon 20
Fusion	Yes	<i>KCNQ5</i>	exon 1	<i>ALK</i>	exon 10
Fusion	Yes	<i>KIF5B</i>	exon 24	<i>ALK</i>	exon 20
Fusion	Yes	<i>PPP4R3B</i>	exon 11	<i>ALK</i>	exon 20
Fusion	Yes	<i>STRN</i>	exon 3	<i>ALK</i>	exon 20
Fusion	Yes	<i>TFG</i>	exon 5	<i>ALK</i>	exon 20
Fusion	Yes	<i>CUX1</i>	exon 8	<i>BRAF</i>	exon 8
Fusion	Yes	<i>DSTYK</i>	exon 11	<i>BRAF</i>	exon 10
Fusion	Yes	<i>KIAA1549</i>	exon 16	<i>BRAF</i>	exon 11
Fusion	Yes	<i>MACF1</i>	exon 66	<i>BRAF</i>	exon 9
Fusion	Yes	<i>TMPRSS2</i>	exon 3	<i>BRAF</i>	exon 11
Fusion	Yes	<i>STRIP1</i>	exon 9	<i>CSF1</i>	exon 6
Fusion	Yes	<i>MX1</i>	exon 9	<i>ERG</i>	exon 4
Fusion	Yes	<i>NDRG1</i>	exon 3	<i>ERG</i>	exon 4
Fusion	Yes	<i>TMPRSS2</i>	exon 2	<i>ERG</i>	exon 5
Fusion	Yes	<i>TMPRSS2</i>	exon 2	<i>ERG</i>	exon 2

Fusion	Yes	<i>PTPRZ1</i>	exon 13	<i>MET</i>	exon 2
Fusion	Yes	<i>CD74</i>	exon 6	<i>NRG1</i>	exon 6
Fusion	Yes	<i>GPRC5A</i>	exon 13	<i>NRG1</i>	exon 22
Fusion	Yes	<i>SDC4</i>	exon 4	<i>NRG1</i>	exon 6
Fusion	Yes	<i>ARHGEF25</i>	exon 11	<i>NTRK1</i>	exon 8
Fusion	Yes	<i>CD74</i>	exon 8	<i>NTRK1</i>	exon 10
Fusion	Yes	<i>CEL</i>	exon 7	<i>NTRK1</i>	exon 9
Fusion	Yes	<i>LMNA</i>	exon 3	<i>NTRK1</i>	exon 11
Fusion	Yes	<i>MPRIIP</i>	exon 22	<i>NTRK1</i>	exon 12
Fusion	Yes	<i>SQSTM1</i>	exon 6	<i>NTRK1</i>	exon 10
Fusion	Yes	<i>TPM3</i>	exon 8	<i>NTRK1</i>	exon 12
Fusion	Yes	<i>ZBTB7B</i>	exon 13	<i>NTRK1</i>	exon 2
Fusion	Yes	<i>AFAP1</i>	exon 14	<i>NTRK2</i>	exon 12
Fusion	Yes	<i>NACC2</i>	exon 4	<i>NTRK2</i>	exon 13
Fusion	Yes	<i>QKI</i>	exon 6	<i>NTRK2</i>	exon 16
Fusion	Yes	<i>ETV6</i>	exon 5	<i>NTRK3</i>	exon 15
Fusion	Yes	<i>ETV6</i>	exon 4	<i>NTRK3</i>	exon 14
Fusion	Yes	<i>FAT1</i>	exon 2	<i>NTRK3</i>	exon 7
Fusion	Yes	<i>VSP18</i>	exon 4	<i>NTRK3</i>	exon 19
Fusion	Yes	<i>COL1A1</i>	exon 25	<i>PDGFB</i>	exon 2
Fusion	Yes	<i>MKRN2</i>	exon 1	<i>PPARG</i>	exon 4
Fusion	Yes	<i>PAX8</i>	exon 9	<i>PPARG</i>	exon 2
Fusion	Yes	<i>AGGF1</i>	exon 5	<i>RAF1</i>	exon 9
Fusion	Yes	<i>CMTM8</i>	exon 1	<i>RAF1</i>	exon 2
Fusion	Yes	<i>PAPD7</i>	exon 11	<i>RAF1</i>	exon 11
Fusion	Yes	<i>CCDC6</i>	exon 1	<i>RET</i>	exon 12
Fusion	Yes	<i>CCDC6</i>	exon 1	<i>RET</i>	exon 2
Fusion	Yes	<i>FKBP15</i>	exon 22	<i>RET</i>	exon 8
Fusion	Yes	<i>KIF5B</i>	exon 24	<i>RET</i>	exon 7
Fusion	Yes	<i>KIF5B</i>	exon 24	<i>RET</i>	exon 10
Fusion	Yes	<i>KIF5B</i>	exon 15	<i>RET</i>	exon 11
Fusion	Yes	<i>NCOA4</i>	exon 8	<i>RET</i>	exon 12
Fusion	Yes	<i>TRIM33</i>	exon 11	<i>RET</i>	exon 12
Fusion	Yes	<i>CD74</i>	exon 6	<i>ROS1</i>	exon 35
Fusion	Yes	<i>CEP85L</i>	exon 9	<i>ROS1</i>	exon 37
Fusion	Yes	<i>CLTC</i>	exon 31	<i>ROS1</i>	exon 36
Fusion	Yes	<i>EGFR</i>	exon 1	<i>CHODL</i>	exon 3
Fusion	Yes	<i>EGFR</i>	exon 24	<i>SEPT14</i>	exon 10
Fusion	Yes	<i>EGFR</i>	exon 27	<i>VSTM2A</i>	exon 6
Fusion	Yes	<i>FGFR1</i>	exon 1	<i>LETM2</i>	exon 2
Fusion	Yes	<i>FGFR1</i>	exon 1	<i>LETM2</i>	exon 2
Fusion	Yes	<i>FGFR1</i>	exon 3	<i>ADAM18</i>	exon 4
Fusion	Yes	<i>FGFR2</i>	exon 10	<i>FRK</i>	exon 4
Fusion	Yes	<i>FGFR2</i>	exon 17	<i>CCAR2</i>	exon 4
Fusion	Yes	<i>FGFR3</i>	exon 4	<i>TACC3</i>	exon 14
Fusion	Yes	<i>FGFR3</i>	exon 3	<i>TACC3</i>	exon 4
Fusion	Yes	<i>FGFR3</i>	exon 17	<i>TACC3</i>	exon 11
Exon skipping	Yes	<i>MET</i>	exon 13	<i>MET</i>	exon 15
mid-exon-fusion	No	<i>MALAT1</i>	mid exon 3	<i>ALK</i>	exon 16
mid-exon-fusion	Filtered	<i>KIAA1549</i>	exon 15	<i>BRAF</i>	exon 9 then 11
mid-exon-fusion	Filtered	<i>PAN3</i>	mid-exon-2	<i>NTRK2</i>	exon 15
mid-exon-fusion	Filtered	<i>TRIM24</i>	mid-exon-12	<i>NTRK2</i>	exon 13
mid-exon-fusion	Filtered	<i>KIF5B</i>	exon 15	<i>RET</i>	mid exon 11
mid-exon-fusion	Filtered	<i>SLC34A2</i>	mid-exon-13	<i>ROS1</i>	exon 33
reciprocal fusion	No	<i>SLC34A2</i>	exon 3	<i>EGFR</i>	exon 2
reciprocal fusion	No	<i>BAG4</i>	exon 2	<i>FGFR1</i>	exon 6

[00315] Geneblocks encode biologically relevant fusions (65 two genes, 1 exon skipping) while 8 encoded non-biologically relevant fusions (2 reciprocal orientation, 6 mid-exon). 10 pg of the dsDNA geneblock pool spiked into the fusion-negative normal cDNA background were tested, and all 66 detectable geneblocks across 18 genes and 137 exons with the non-biologically relevant geneblock fusions filtered were identified. None of the geneblock fusions were detected in the fusion-negative samples.

[00316] Assessment of analytical sensitivity and specificity was performed using a dilution series experiment using the 18 fusion-positive sample. Three runs were performed representing ten replicates of the template mass input series: 12.5 ng, 25 ng, 50 ng, and 100 ng (750-12,000 copies per fusion). Unique reads supporting both normal and fusion transcripts increased with increasing input mass and fusion copies as provided by the manufacturer (**Figure 23**). Similar numbers of unique fusion reads were observed for each of the 18 fusion targets (**Figure 24**). Sensitivity of detection of fusions increased from 90.6% at 12.5ng (750 copies) to 98.3-99.4% for 25-100 ng (1,500-12,000 copies) with a minimum PPV of 99.4% (**Table 13**).

Table 13: Sensitivity and PPV for the 18 fusion-positive sample with increasing input mass. True positives (TP), false negatives (FN), false positives (FP) are indicated. Sensitivity is defined as $\text{Sensitivity} = (\text{TP})/(\text{TP}+\text{FN})$. PPV is defined as $\text{PPV} = (\text{TP})/(\text{TP}+\text{FP})$.

RNA Input	ddPCR Copies (Vendor)	TP	FN	FP	Sensitivity	PPV
12.5ng	753	163	17	0	90.6%	100.0%
25ng	1,507	177	3	0	98.3%	100.0%
50ng	3,014	178	2	1	98.9%	99.4%
100ng	6,028	179	1	0	99.4%	100.0%

[00317] Sensitivity of detection of SLC45A3-*BRAF* and *EGFR-SEPT14* gene fusions in the contrived 18 fusion-positive sample was lowest (**Table 14**) with corresponding low fusion to normal read ratios (0.00-0.01 and 0.01-0.08) (**Table 15**).

Table 14: Sensitivity = (TP)/(TP+FN) for expected fusions in the 18 fusion-positive sample with increasing template input mass. Assay of each input level was replicated for a total of 10 possible observations.

Fusion	12.5ng	25ng	50ng	100ng
<i>CCDC6-RET (1;12)</i>	80%	100%	100%	100%
<i>CD74-ROS1 (6;35)</i>	100%	100%	100%	100%
<i>EGFR-EGFR (1;8)</i>	100%	100%	100%	100%
<i>EGFR-SEPT14 (24;10)</i>	50%	90%	90%	100%

<i>EML4-ALK (14;20)</i>	70%	100%	100%	100%
<i>ETV6-NTRK3 (5;15)</i>	100%	100%	100%	100%
<i>FGFR3-BAIAP2L1 (17;2)</i>	100%	100%	100%	100%
<i>FGFR3-TACC3 (17;11)</i>	100%	100%	100%	100%
<i>KIF5B-RET (24;11)</i>	100%	100%	100%	100%
<i>LMNA-NTRK1 (3;11)</i>	100%	100%	100%	100%
<i>MET-MET (13;15)</i>	90%	100%	100%	100%
<i>NCOA4-RET (8;12)</i>	100%	100%	100%	100%
<i>PAX8-PPARG (9;2)</i>	100%	100%	100%	100%
<i>SLC34A2-ROS1 (4;35)</i>	100%	100%	100%	100%
<i>SLC45A3-BRAF (1;8)</i>	60%	80%	90%	90%
<i>TFG-NTRK1 (5;10)</i>	100%	100%	100%	100%
<i>TPR52-ERG (1;2)</i>	80%	100%	100%	100%
<i>TPM3-NTRK1 (8;10)</i>	100%	100%	100%	100%

Table 15: Median on-target read counts for target genes (includes any normal transcript or fusion transcript derived reads) for the NTC 'no template control', the fusion-negative sample, and 18 fusion-positive sample. Detected on-target fusion to normal read ratio range for observations is indicated for each fusion.

Target Gene	NTC on-target reads	Fusion-Negative Sample on-target reads	18 Fusion-Positive Sample on-target reads	Fusion to normal read ratio
<i>BRAF</i>	0	17386	60664	0.00-0.01 (<i>SLC45A3-BRAF</i>)
<i>TUBB</i>	0	5479	52515	N/A
<i>RAF1</i>	1	24970	42128	N/A
<i>FGFR1</i>	0	11669	36072	N/A
<i>EGFR</i>	1	164717	10008	0.01-0.08 (<i>EGFR-SEPT14</i>), 0.05-0.51 (<i>EGFR ex1-EGFR ex8</i>)
<i>NTRK1</i>	0	126	5873	0.01-0.26 (<i>TFG-NTRK1</i>), 0.03-0.25 (<i>LMNA-NTRK1</i>), 0.08-0.44 (<i>TPM3-NTRK1</i>)
<i>ALK</i>	0	49	3327	0.04-1.16 (<i>EML4-ALK</i>)
<i>FGFR3</i>	0	9655	3565	1.79-7.26 (<i>FGFR3-TACC3</i>), 0.73-3.97 (<i>FGFR3-BAIAP2L1</i>)
<i>CSF1</i>	0	2282	2680	N/A
<i>ERG</i>	0	21022	2304	0.05-108.00 (<i>TPR52-ERG</i>)
<i>RET</i>	0	1776	1723	0.02-2.92 (<i>KIF5B-RET</i>), 0.04-1.29 (<i>CCDC6-RET</i>), 0.13-3.88 (<i>NCOA4-RET</i>)
<i>NTRK3</i>	0	1317	840	0.64-32.07 (<i>ETV6-NTRK3</i>)
<i>ROS1</i>	0	314	717	0.01-1.42 (<i>CD74-ROS1</i>), 0.00-0.40 (<i>SLC34A2-ROS1</i>)
<i>NTRK2</i>	0	314	606	N/A
<i>PPARG</i>	0	2121	463	22.86-14,572.00 (<i>PAX8-PPARG</i>)
<i>MET</i>	0	29864	447	0.13-5.00 (<i>MET ex13-MET ex15</i>)
<i>PDGFB</i>	0	17	32	N/A
<i>FGFR2</i>	0	4219	1	N/A
<i>NRG1</i>	0	2135	1	N/A

[00318] At 25 ng (~1500 copies per fusion) or higher, 16 of 18 fusions had 100% detection. The fusions *SLC45A3-BRAF* and *EGFR-SEPT14* were not detected in 6 of 60 observations at these input levels. These targets had high normal transcript reads and low fusion to normal read ratios of < 1% and 1-8% respectively. The

fusions *BRAF-SLC34A3*, *LMNA-NTRK1*, *TFG-NTRK1*, and *EGFR-SEPT14* were detected with 1-4% fusion to normal transcript read ratios but had lower numbers of normal transcripts. One false positive medium confidence artifact in the 18 fusion-positive sample dilution experiments for a single replicate (*NFRKB-PDGFB*) was observed. This fusion had two magnitudes lower read support compared to other fusions detected in this sample with 3 total and 1 unique supporting fusion read.

[00319] **Detection of gene fusions in clinical FFPE specimens**

[00320] To evaluate gene fusion detection from fusion-positive FFPE specimens of varying quality we performed a dilution series experiment on FFPE material from 5 treated cell lines (< 1 year old), and 15 archival clinical samples (7: < 2 years old, 8: 2-5 years old). RNA quality was assessed by DV200, and cDNA quality assessed using qPCR. RNA DV200 values ranged from 86-91% (median 90%) for the cell-lines and 30-86% (median 69%) for the archival specimens. Median qPCR CT values decreased with increasing template mass (20 ng, 100 ng, 250 ng, 500 ng), ranging from 26.6-23.3 CT value for cell lines and 33.2-27.8 CT value for archival specimens (Table 16).

Table 16: Table of 20 FFPE testing samples and results.

Sample ID	Type	Sample	Block Date	Fusion	Input (ng)	DV200	qPCR	Unique normal reads	Unique Fusion reads
57	FFPE Cell Line	3020-0130 H2228	2019	<i>EML4-ALK</i>	20	90%	25.85	5264	44
57	FFPE Cell Line	3020-0130 H2228	2019	<i>EML4-ALK</i>	100	90%	23.49	9332	83
57	FFPE Cell Line	3020-0130 H2228	2019	<i>EML4-ALK</i>	250	90%	22.61	10411	114
57	FFPE Cell Line	3020-0130 H2228	2019	<i>EML4-ALK</i>	500	90%	22.63	10514	135
58	FFPE Cell Line	3020-1130 H596	2019	<i>del MET 14</i>	20	86%	26.68	4311	46
58	FFPE Cell Line	3020-1130 H596	2019	<i>del MET 14</i>	100	86%	24.31	7695	86
58	FFPE Cell Line	3020-1130 H596	2019	<i>del MET 14</i>	250	86%	23.3	8396	92
58	FFPE Cell Line	3020-1130 H596	2019	<i>del MET 14</i>	500	86%	23.34	8194	85
59	FFPE Cell Line	3020-1430 HCC78	2019	<i>SLC34A2-ROS1</i>	20	88%	27.4	3842	126
59	FFPE Cell Line	3020-1430 HCC78	2019	<i>SLC34A2-ROS1</i>	100	88%	24.79	6516	180
59	FFPE Cell Line	3020-1430 HCC78	2019	<i>SLC34A2-ROS1</i>	250	88%	24.28	7258	188
59	FFPE Cell Line	3020-1430 HCC78	2019	<i>SLC34A2-ROS1</i>	500	88%	23.86	8917	221
60	FFPE Cell Line	3130-0130 U118MG	2019	<i>GOPC-ROS1</i>	20	91%	26.61	5154	35
60	FFPE Cell Line	3130-0130 U118MG	2019	<i>GOPC-ROS1</i>	100	91%	24.47	8948	81
60	FFPE Cell Line	3130-0130 U118MG	2019	<i>GOPC-ROS1</i>	250	91%	23.44	9947	89
60	FFPE Cell Line	3130-0130 U118MG	2019	<i>GOPC-ROS1</i>	500	91%	23.52	8860	83
61	FFPE Cell	3060-1130 A431	2019	<i>EGFR-PPARG</i>	20	90%	26.41	3959	29

	Line								
61	FFPE Cell Line	3060-1130 A431	2019	EGFR-PPARG	100	90%	24.26	6654	72
61	FFPE Cell Line	3060-1130 A431	2019	EGFR-PPARG	250	90%	23.31	8109	106
61	FFPE Cell Line	3060-1130 A431	2019	EGFR-PPARG	500	90%	22.86	6431	67
62	FFPE Specimen	Lung	2018	KIF5B-RET	20	>70%	27.19	1448	7
62	FFPE Specimen	Lung	2018	KIF5B-RET	100	>70%	24.61	3885	29
62	FFPE Specimen	Lung	2018	KIF5B-RET	250	>70%	24.21	4534	23
62	FFPE Specimen	Lung	2018	KIF5B-RET	500	>70%	23.89	3998	28
63	FFPE Specimen	Lung	2018	CD74-ROS1	20	30-50%	29.25	1628	7
63	FFPE Specimen	Lung	2018	CD74-ROS1	100	30-50%	27.4	3095	19
63	FFPE Specimen	Lung	2018	CD74-ROS1	250	30-50%	26.81	2730	19
63	FFPE Specimen	Lung	2018	CD74-ROS1	500	30-50%	26.56	4001	38
64	FFPE Specimen	Colon	2018	TPM3-NTRK1	20	<30%	33.75	81	5
64	FFPE Specimen	Colon	2018	TPM3-NTRK1	100	<30%	32.71	135	8
64	FFPE Specimen	Colon	2018	TPM3-NTRK1	250	<30%	31.46	185	11
64	FFPE Specimen	Colon	2018	TPM3-NTRK1	500	<30%	30.38	200	20
65	FFPE Specimen	Unknown	2018	ETV6-NTRK3	20	59%	29.93	569	13
65	FFPE Specimen	Unknown	2018	ETV6-NTRK3	100	59%	27.92	1786	22
65	FFPE Specimen	Unknown	2018	ETV6-NTRK3	250	59%	26.68	1868	39
65	FFPE Specimen	Unknown	2018	ETV6-NTRK3	500	59%	26.62	2219	38
66	FFPE Specimen	Unknown	2018	ETV6-NTRK3	20	38%	ND	67	0
66	FFPE Specimen	Unknown	2018	ETV6-NTRK3	100	38%	31.4	155	3
66	FFPE Specimen	Unknown	2018	ETV6-NTRK3	250	38%	31.39	268	3
66	FFPE Specimen	Unknown	2018	ETV6-NTRK3	500	38%	30.27	263	0
67	FFPE Specimen	Lung	2018	EML4-ALK	20	86%	29.94	1073	7
67	FFPE Specimen	Lung	2018	EML4-ALK	100	86%	26.43	2679	37
67	FFPE Specimen	Lung	2018	EML4-ALK	250	86%	23.92	3288	46
67	FFPE Specimen	Lung	2018	EML4-ALK	500	86%	23.3	3699	55
68	FFPE Specimen	Lung	2018	EML4-ALK	20	63%	33.17	434	5
68	FFPE Specimen	Lung	2018	EML4-ALK	100	63%	ND	909	16
68	FFPE Specimen	Lung	2018	EML4-ALK	250	63%	ND	715	12
68	FFPE Specimen	Lung	2018	EML4-ALK	500	63%	26.88	2153	36
69	FFPE Specimen	Lung	2015	EML4-ALK	20	61%	30.85	456	2
69	FFPE Specimen	Lung	2015	EML4-ALK	100	61%	29.15	1227	7
69	FFPE Specimen	Lung	2015	EML4-ALK	250	61%	28.18	1649	8
69	FFPE Specimen	Lung	2015	EML4-ALK	500	61%	27.85	2057	16
70	FFPE Specimen	Lung	2015	EML4-ALK	20	69%	ND	15	0
70	FFPE Specimen	Lung	2015	EML4-ALK	100	69%	ND	48	0

70	FFPE Specimen	Lung	2015	EML4-ALK	250	69%	ND	121	0
70	FFPE Specimen	Lung	2015	EML4-ALK	500	69%	ND	123	0
71	FFPE Specimen	Lung	2015	EML4-ALK	20	66%	ND	13	0
71	FFPE Specimen	Lung	2015	EML4-ALK	100	66%	ND	28	0
71	FFPE Specimen	Lung	2015	EML4-ALK	250	66%	34.98	51	0
71	FFPE Specimen	Lung	2015	EML4-ALK	500	66%	32.58	75	1
72	FFPE Specimen	Lung	2015	EML4-ALK	20	85%	30.9	1448	0
72	FFPE Specimen	Lung	2015	EML4-ALK	100	85%	29.59	3885	16
72	FFPE Specimen	Lung	2015	EML4-ALK	250	85%	27.4	4534	34
72	FFPE Specimen	Lung	2015	EML4-ALK	500	85%	27.44	3998	43
73	FFPE Specimen	Lung	2015	EML4-ALK	20	72%	31.29	1628	2
73	FFPE Specimen	Lung	2015	EML4-ALK	100	72%	30.75	3095	4
73	FFPE Specimen	Lung	2015	EML4-ALK	250	72%	28.41	2730	8
73	FFPE Specimen	Lung	2015	EML4-ALK	500	72%	27.55	4001	18
74	FFPE Specimen	Lung	2014	EML4-ALK	20	70%	33.24	81	0
74	FFPE Specimen	Lung	2014	EML4-ALK	100	70%	32.14	135	3
74	FFPE Specimen	Lung	2014	EML4-ALK	250	70%	33.6	185	5
74	FFPE Specimen	Lung	2014	EML4-ALK	500	70%	29.97	200	7
75	FFPE Specimen	Lung	2013	EML4-ALK	20	>70%	ND	569	0
75	FFPE Specimen	Lung	2013	EML4-ALK	100	>70%	32.7	1786	3
75	FFPE Specimen	Lung	2013	EML4-ALK	250	>70%	29.61	1868	6
75	FFPE Specimen	Lung	2013	EML4-ALK	500	>70%	29.59	2219	8
76	FFPE Specimen	Lung	2015	EML4-ALK	20	63%	36.98	67	0
76	FFPE Specimen	Lung	2015	EML4-ALK	100	63%	ND	155	0
76	FFPE Specimen	Lung	2015	EML4-ALK	250	63%	ND	268	0
76	FFPE Specimen	Lung	2015	EML4-ALK	500	63%	33.27	263	0

[00321] Gene fusions were detected at all input levels for the FFPE cell lines and specimens less than 2 years old, while gene fusions were not detected at 20 ng inputs in the specimens that were older than 2 years (**Figure 25**). Increasing the input mass of RNA did not compensate for decreased quality.

[00322] Evaluation of sample quality with respect to assay performance was more effective with qPCR than DV200. At the lowest input of 20 ng of RNA, the sensitivity of detection was 65% without taking into account any quality assessment by DV200 and qPCR, 75% with DV200, and 100% with qPCR quality assessment (**Table 17**).

Table 17: Sensitivity of the detection of 20 fusion-positive archival FFPE samples at different template input mass levels with and without screening by DV200 or qPCR.

Filter	Samples	20ng RNA	100ng RNA	250ng RNA	500ng RNA
All samples	20	65% (13)	85% (17)	85% (17)	85% (17)
DV200 > 70%	11	73% (8)	100% (11)	100% (11)	100% (11)
qPCR CT < 30	9	100% (9)	100% (9)	100% (9)	100% (9)

[00323] A threshold CT value equal to 30 was used to reduce the inclusion of poor quality samples that can result in false negatives, however gene fusions were still detectable in some of these samples. There was no separation of data based on total unique reads and qPCR values with respect to template input amount showing that sample quality (qPCR) is more predictive of outcome than input mass (**Figure 26**).

[00324] The assay performance was next validated using 56 unique archival clinical FFPE samples that were represented across 101 observations (**Table 18**; four observations of a reference standard were omitted).

Table 18: Table of validation samples and results. TN = True Negative, TP = True Positive, FP = False Positive, FN = False Negative.

Sample ID	Tissue	Orthogonal Status	qPCR	Unique Reads	RNA Input (ng)	Status	Fusion to normal reads
1	Lung	Negative	Pass (26.54)	Pass (1142)	399	TN	N/A
2	Lung	Negative	Pass (22.76)	Pass (6885)	306	TN	N/A
3	Lung	Negative	Pass (24.14)	Pass (2856)	500	TN (low <i>PHF21A-NTRK2</i>)	N/A
4	Lung	Negative	Pass (24.78)	Pass (2584)	334	TN	N/A
4	Lung	Negative	Pass (23.31)	Pass (3801)	416	TN	N/A
5	Lung	Negative	Pass (25.93)	Pass (5221)	95	TN (low <i>SYN2-PPARG</i>)	N/A
5	Lung	Negative	Pass (25.93)	Pass (4716)	98	TN	N/A
5	Lung	Negative	Pass (25.08)	Pass (8443)	104	TN	N/A
5	Lung	Negative	Pass (25.11)	Pass (8943)	66	TN	N/A
6	Lung	Negative	Pass (24.23)	Pass (7825)	110	TN (low <i>SYN2-PPARG</i>)	N/A
6	Lung	Negative	Pass (24.23)	Pass (7302)	110	TN (medium <i>SYN2-PPARG</i>)	N/A
6	Lung	Negative	Pass (24.28)	Pass (9274)	121	TN	N/A
6	Lung	Negative	Pass (24.28)	Pass (9090)	121	TN	N/A
7	Lung	ALK positive	Fail (32.15)	Pass (532)	255	TP-High (<i>MTA3-ALK</i>)	0.24
7	Lung	ALK positive	Fail (32.15)	Pass (354)	255	TP-High (<i>MTA3-ALK</i>)	1.00
8	Lung	Negative	Fail (34.8)	Pass (1770)	246	TN	N/A
8	Lung	Negative	Fail (34.8)	Pass (1758)	246	TN	N/A

8	Lung	ALK positive	Fail (30)	Pass (4682)	479	FN	N/A
9	Lung	ALK positive	Fail (30)	Pass (953)	138	TP-High (EML4-ALK)	2.48
9	Lung	ALK positive	Fail (30)	Pass (775)	138	TP-High (EML4-ALK)	0.64
9	Lung	ALK positive	Fail (30.77)	Pass (1928)	1637	TP-High (EML4-ALK)	0.75
9	Lung	ALK positive	Fail (30.77)	Pass (1861)	1637	TP-High (EML4-ALK)	0.47
10	Lung	ALK positive	Fail (30.66)	Pass (1169)	270	FN	N/A
10	Lung	ALK positive	Fail (32.85)	Pass (2756)	480	FN	N/A
10	Lung	ALK positive	Fail (32.85)	Pass (2419)	480	FN	N/A
11	Lung	ALK positive	Fail (31.11)	Pass (487)	270	TP-High (EML4-ALK)	2.07
11	Lung	ALK positive	Fail (31.11)	Pass (343)	270	TP-High (EML4-ALK)	1.91
12	Lung	ALK positive	Pass (24.19)	Pass (3017)	170	TP-High (EML4-ALK)	0.09
13	Lung	ALK positive	Pass (28.16)	Pass (4256)	69	TP-Low (SP100-ALK)	0.03
14	Lung	ROS1 positive	Pass (29.7)	Pass (1309)	246	TP-High (EZR-ROS1)	0.13
14	Lung	ROS1 positive	Pass (29.7)	Pass (1222)	246	TP-High (EZR-ROS1)	0.13
14	Lung	ROS1 positive	Pass (28.55)	Pass (1092)	250	TP-High (EZR-ROS1)	0.08
15	Pleural Fluid	Negative	Fail (31.29)	Fail (278)	441	TN	N/A
16	Lung	ALK positive	Pass (25.79)	Pass (3312)	15	TP-High (EML4-ALK)	2.43
17	Lung	ALK positive	Pass (28.84)	Pass (1872)	83	TP-High (EML4-ALK)	0.54
17	Lung	ALK positive	Pass (28.84)	Pass (1774)	46	TP-High (EML4-ALK)	0.30
17	Lung	ALK positive	Pass (27.73)	Pass (4249)	83	TP-High (EML4-ALK)	0.28
17	Lung	ALK positive	Pass (27.73)	Pass (3757)	83	TP-High (EML4-ALK)	0.31
18	Lung	ALK positive	Pass (28.59)	Pass (1931)	246	TP-High (EML4-ALK)	0.62
18	Lung	ALK positive	Pass (28.59)	Pass (2203)	246	TP-High (EML4-ALK)	0.44
18	Lung	ALK positive	Pass (26.43)	Pass (4251)	227	TP-High (EML4-ALK)	1.08
18	Lung	ALK positive	Pass (26.43)	Pass (3799)	227	TP-High (EML4-ALK)	0.55
19	Lung	ALK positive	Pass (26.85)	Pass (2993)	119	TP-High (EML4-ALK)	1.26
19	Lung	ALK positive	Pass (26.85)	Pass (3205)	119	TP-High (EML4-ALK)	1.21
19	Lung	ALK positive	Pass (26.58)	Pass (4394)	131	TP-High (EML4-ALK)	0.62
19	Lung	ALK positive	Pass (26.58)	Pass (3980)	131	TP-High (EML4-ALK)	0.85
20	Lung	ALK positive	Pass (24.97)	Pass (5757)	1501	TP-High (EML4-ALK)	4.83
21	Lung	ALK positive	Fail (33.14)	Fail (102)	100	FN	N/A
21	Lung	ALK positive	Fail (33.14)	Fail (67)	100	TP-Low (EML4-ALK)	1.5
21	Lung	ALK positive	Fail (33.38)	Fail (149)	132	TP-Low (EML4-ALK)	0.63

21	Lung	ALK positive	Fail (33.38)	Fail (150)	132	FN	N/A
21	Lung	ALK positive	Fail (32.22)	Fail (273)	132	TP-Low (EML4-ALK)	0.17
22	Lung	ALK positive	Fail (32.4)	Fail (62)	227	TP-Medium (EML4-ALK)	0.33
22	Lung	ALK positive	Fail (32.4)	Fail (84)	227	TP-Medium (EML4-ALK)	3.22
22	Lung	ALK positive	Fail (32.37)	Fail (127)	250	FN	N/A
23	Lung	ALK positive	Fail (32.14)	Fail (30)	250	FN	N/A
23	Lung	ALK positive	Fail (32.14)	Fail (86)	388	TP-Medium (EML4-ALK)	2.38
23	Lung	ALK positive	Fail (32.14)	Fail (79)	388	FN	N/A
24	Lung	ALK positive	Pass (28.5)	Pass (969)	375	TP-Low (EML4-ALK)	0.24
24	Lung	ALK positive	Pass (28.5)	Pass (760)	250	TP-Low (SP100-ALK)	0.24
25	Brain	NTRK2 positive	Pass (27.31)	Pass (3694)	156	TP-High (STRN3-NTRK2)	0.01
26	Brain	NTRK2 positive	Pass (28.97)	Pass (757)	270	TP-High (WNK2-NTRK2)	4.83
27	Lung	ROS1 positive	Pass (29.76)	Fail (64)	219	FN	N/A
28	Lung	ROS1 positive	Fail (32.31)	Fail (56)	256	FN	N/A
28	Lung	ROS1 positive	Fail (30.88)	Fail (146)	525	FN	N/A
29	Lung	RET positive	Fail (30.1)	Fail (234)	332	TP-Medium (KIF5B-RET)	0.23
30	Lung	RET positive	Pass (28.14)	Pass (503)	332	TP-High (KIF5B-RET)	5.11
31	Lung	Negative	Pass (25.91)	Pass (3283)	270	TN	N/A
31	Lung	Negative	Pass (25.91)	Pass (2782)	246	TN	N/A
32	Lung	Negative	Pass (24.96)	Pass (5718)	270	TN	N/A
32	Lung	Negative	Pass (24.96)	Pass (4441)	270	TN	N/A
33	Lung	RET positive	Pass (27.65)	Pass (3414)	270	TP-High (CCDC6-RET)	0.01
33	Lung	RET positive	Pass (25.68)	Pass (4501)	1366	TP-High (CCDC6-RET)	0.02
34	Lung	Negative	Pass (25.96)	Pass (4690)	270	FP-Low (TMBIM4-PDGFB)	0.33
34	Lung	Negative	Pass (25.96)	Pass (2633)	270	TN	N/A
35	Lung	Negative	Pass (23.29)	Pass (6807)	1009	TN (SYN2-PPARG)	N/A
36	Lung	Met del 14	Pass (23.44)	Pass (7746)	250	TP-High (Met del 14), (low NSG1-PPARG, low SYN2-PPARG)	0.04
37	Lung	Negative	Pass (29.25)	Pass (789)	250	TN	N/A
38	Lung	Negative	Pass (28.94)	Pass (1006)	250	TN	N/A
39	Bone	ALK positive	Pass (29.61)	Pass (564)	270	TP-Medium (CCDC88A-ALK)	0.09
39	Bone	ALK positive	Pass (29.61)	Pass (543)	270	TP-Medium (CCDC88A-ALK)	0.10
40	Thyroid	NTRK positive	Fail (36)	Fail (82)	270	FN	N/A
40	Thyroid	NTRK positive	Fail (36)	Fail (107)	861	FN	N/A
41	Lymph Node	Met del 14	Pass (26.89)	Pass (3180)	109	TP-High (Met del 14)	0.04
42	Lung	Met del 14	Pass (24.29)	Pass (10257)	52	TP-High (Met del 14)	0.03
43	Pleural Fluid	Met del 14	Pass (27.39)	Pass (3915)	15	TP-High (Met del 14)	0.03

44	Lung	<i>Met</i> del 14	Pass (26.29)	Pass (1488)	199	TP-Medium (<i>Met</i> del 14)	0.01
45	Brain	<i>Met</i> del 14	Pass (27.84)	Pass (1829)	106	TP-High (<i>Met</i> del 14)	0.03
46	Lung	<i>Met</i> del 14	Pass (25.78)	Pass (5215)	200	TP-High (<i>Met</i> del 14)	0.02
47	Lung	<i>Met</i> del 14	Pass (24.82)	Pass (6905)	222	TP-High (<i>Met</i> del 14)	0.04
48	Lung	<i>Met</i> del 14	Pass (26.45)	Pass (3482)	19	TP-High (<i>Met</i> del 14)	0.08
49	Lung	<i>ALK</i> positive	Pass (26.48)	Pass (3992)	197	TP-High (<i>KLC1-ALK</i>)	0.38
50	Lung	<i>ROS1</i> positive	Pass (28.21)	Pass (2007)	250	TP-High (<i>SLC34A2-ROS1</i>)	0.02
51	Lung	<i>ROS1</i> positive	Fail (31.14)	Pass (1172)	10	TP-High (<i>SLC34A2-ROS1</i>)	0.03
52	Lung	Negative	Pass (24.77)	Pass (6857)	122	TN	N/A
52	Lung	Negative	Pass (24.77)	Pass (5892)	122	TN	N/A
53	Lymph Node	Negative	Pass (28.33)	Pass (2044)	21	TN	N/A
53	Lymph Node	Negative	Pass (28.33)	Pass (1355)	21	TN	N/A
54	Lung	Negative	Pass (28.94)	Pass (970)	250	TN	N/A
55	Colon	<i>NTRK3</i> positive	Pass (29.54)	Pass (802)	270	TP-High (<i>KANK1-NTRK3</i>)	2.23
56	Lung	<i>Met</i> del 14	Pass (28.65)	Pass (3712)	250	TP-Medium (<i>Met</i> del 14)	0.01

[00325] Five validation sequencing experiments were performed by two different technologists on two instruments and on different days. Each sequencing run consisted of 3 controls and 21 samples. Samples were clinical FFPE specimens derived from patients with cancers of the lung, brain, lymph node, thyroid, and bone, with orthogonal molecular results including NGS testing, Immunohistochemistry (IHC), and Fluorescence In-situ Hybridization (FISH). When screened for minimum quality requirements (qPCR CT < 30), 69 observations from 44 unique samples were considered. There was no difference in the mass of template RNA used for low quality samples with qPCR CT > 30 (median 256 ng, minimum 11 ng) and high-quality samples with qPCR CT < 30 (median 223 ng, minimum 15 ng). Samples were classified as true positive, true negative, false positive, and false negative according to the assay relative to their orthogonal status.

[00326] Summary performance metrics were calculated with and without screening quality by qPCR as well as with and without low confidence fusion calls (Table 19).

Table 19: Assay validation performance metrics. Classifications at the sample level. Passing screen was defined as samples with qPCR CT values < 30 and/or

where fusions were high or medium confidence as indicated. Accuracy of the assay where $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$. Sensitivity of the assay where $\text{Sensitivity} = (\text{TP}) / (\text{TP} + \text{FN})$. Specificity of the assay where $\text{Specificity} = (\text{TN}) / (\text{TN} + \text{FP})$. Positive predictive value of the assay where $\text{PPV} = (\text{TP}) / (\text{TP} + \text{FP})$. Negative predictive value of the assay where $\text{NPV} = (\text{TN}) / (\text{TN} + \text{FN})$.

Metric	All Samples	qPCR < 30	Exclude low confidence calls	qPCR < 30 Exclude low confidence calls
Unique Samples	56	44	56	44
Total Samples	101	69	101	69
True Positive	57	41	51	38
True Negative	29	26	30	27
False Positive	1	1	0	0
False Negative	14	1	20	4
Accuracy	85.2%	97.1%	80.2%	94.2%
Sensitivity	80.3%	97.6%	71.8%	90.5%
Specificity	96.7%	96.3%	100.0%	100.0%
PPV	98.3%	97.6%	100.0%	100.0%
NPV	67.4%	96.3%	60.0%	87.1%

[00327] When all 101 sample observations from 56 samples were considered, accuracy was 85%, sensitivity was 80%, specificity was 97%, PPV was 98%, and NPV was 67%. Applying qPCR thresholds ($\text{CT} < 30$) decreased false negatives (improving NPV and sensitivity) while excluding low confidence fusion calls removed false positives (improving PPV and specificity). When both were considered, all metrics improved (accuracy = 94%, sensitivity = 90%, specificity = 100%, PPV = 100%, NPV = 87%).

[00328] There were four false negative observations after filtering by qPCR and fusion call confidence. One expected *ROS1* fusion was not detected in a sample with low amplifiability near cut-off (qPCR CT 29.76) and had low unique read start-site diversity suggestive of low template quality. Samples with qPCR CT values below 30 typically have more than 300 unique read start-sites. Three additional samples had fusions called with low confidence and thus classified as fusion-negative. All three samples had low amplifiability (qPCR CT 28.16, 28.5, 28.5) and low fusion to normal read ratios for *ALK* (3%, 24%, 24%) which are typically high due to low normal transcript expression (average 92%, median 55% for all *ALK* fusion calls).

[00329] The predominant and most frequent unexpected fusion call observed in 5 clinical samples was *SYN2-PPARG*. These were observed in high quality (qPCR CT 23-25), high depth samples (5,221-7,746 on-target unique reads) and believed to represent a low level of read-through transcription between *PPARG* and adjacent

genes since false positive fusions involving the downstream genes *TSEN2* and *MKRN2* were also observed in other samples. Consequently, fusion calls between *PPARG* and *SYN2*, *TSEN2*, or *MKRN2* are filtered. Detection of a *SYN2-PPARG* fusion has been reported in the literature without further investigation of clinical relevance ²⁹.

[00330] **References**

- [00331] 1. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227-238.e3 (2018).
- [00332] 2. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).
- [00333] 3. Jang, Y. E. *et al.* ChimerDB 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res.* **48**, D817–D824 (2020).
- [00334] 4. Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat. Commun.* **5**, 4846 (2014).
- [00335] 5. Chambers, P. *et al.* Understanding Molecular Testing Uptake Across Tumor Types in Eight Countries: Results From a Multinational Cross-Sectional Survey. *JCO Oncol. Pract.* **16**, e770–e778 (2020).
- [00336] 6. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
- [00337] 7. Yun, J. W. *et al.* Dysregulation of cancer genes by recurrent intergenic fusions. *Genome Biol.* **21**, 166 (2020).
- [00338] 8. Li, J. *et al.* A functional genomic approach to actionable gene fusions for precision oncology. *Sci. Adv.* **8**, eabm2382 (2022).
- [00339] 9. Du, Z. & Lovly, C. M. Mechanisms of receptor tyrosine kinase activation in cancer. *Mol. Cancer* **17**, 58 (2018).
- [00340] 10. Kumar-Sinha, C., Kalyana-Sundaram, S. & Chinnaiyan, A. M. Landscape of gene fusions in epithelial cancers: seq and ye shall find. *Genome Med.* **7**, 129 (2015).
- [00341] 11. Yamaoka, T., Kusumoto, S., Ando, K., Ohba, M. & Ohmori, T. Receptor Tyrosine Kinase-Targeted Cancer Therapy. *Int. J. Mol. Sci.* **19**, E3491 (2018).
- [00342] 12. Cohen, P., Cross, D. & Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov.* **20**, 551–569 (2021).
- [00343] 13. Schram, A. M., Chang, M. T., Jonsson, P. & Drilon, A. Fusions in solid tumours: diagnostic strategies, targeted therapy, and acquired resistance. *Nat. Rev. Clin. Oncol.* **14**, 735–748 (2017).

- [00344] 14. Laskin, J. *et al.* NRG1 fusion-driven tumors: biology, detection, and the therapeutic role of afatinib and other ErbB-targeting agents. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **31**, 1693–1703 (2020).
- [00345] 15. Ross, J. S. *et al.* The distribution of BRAF gene fusions in solid tumors and response to targeted therapy. *Int. J. Cancer* **138**, 881–890 (2016).
- [00346] 16. Zanwar, S. *et al.* Clinical and therapeutic implications of BRAF fusions in histiocytic disorders. *Blood Cancer J.* **12**, 97 (2022).
- [00347] 17. Nikanjam, M., Okamura, R., Barkauskas, D. A. & Kurzrock, R. Targeting fusions for improved outcomes in oncology treatment. *Cancer* **126**, 1315–1321 (2020).
- [00348] 18. Weiss, L. M. & Funari, V. A. NTRK fusions and Trk proteins: what are they and how to test for them. *Hum. Pathol.* **112**, 59–69 (2021).
- [00349] 19. Schmitt, F., Di Lorito, A. & Vielh, P. Molecular Testing on Cytology for Gene Fusion Detection. *Front. Med.* **8**, 643113 (2021).
- [00350] 20. Tachon, G. *et al.* Targeted RNA-sequencing assays: a step forward compared to FISH and IHC techniques? *Cancer Med.* **8**, 7556–7566 (2019).
- [00351] 21. Wu, Y.-C. *et al.* Comparison of IHC, FISH and RT-PCR Methods for Detection of ALK Rearrangements in 312 Non-Small Cell Lung Cancer Patients in Taiwan. *PLOS ONE* **8**, e70839 (2013).
- [00352] 22. Somaschini, A. *et al.* Mining potentially actionable kinase gene fusions in cancer cell lines with the KuNG FU database. *Sci. Data* **7**, 420 (2020).
- [00353] 23. Bruno, R. & Fontanini, G. Next Generation Sequencing for Gene Fusion Analysis in Lung Cancer: A Literature Review. *Diagnostics* **10**, 521 (2020).
- [00354] 24. Heydt, C. *et al.* Detection of gene fusions using targeted next-generation sequencing: a comparative evaluation. *BMC Med. Genomics* **14**, 62 (2021).
- [00355] 25. Schröder, J., Kumar, A. & Wong, S. Q. Overview of Fusion Detection Strategies Using Next-Generation Sequencing. *Methods Mol. Biol. Clifton NJ* **1908**, 125–138 (2019).
- [00356] 26. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
- [00357] 27. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
- [00358] 28. Lohman, G. J. S., Zhang, Y., Zhelkovsky, A. M., Cantor, E. J. & Evans, T. C. Efficient DNA ligation in DNA-RNA hybrid helices by *Chlorella* virus DNA ligase. *Nucleic Acids Res.* **42**, 1831–1844 (2014).
- [00359] 29. Soon, G. S. T., Chang, K. T. E., Kuick, C. H. & Petersson, F. A case of nasal low-grade non-intestinal-type adenocarcinoma with aberrant CDX2

expression and a novel SYN2-PPARG gene fusion in a 13-year-old girl. *Virchows Arch. Int. J. Pathol.* **474**, 619–623 (2019).

[00360] 30. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.* **44**, 4487–4503 (2016).

[00361] 31. Groelz, D., Viertler, C., Pabst, D., Dettmann, N. & Zatloukal, K. Impact of storage conditions on the quality of nucleic acids in paraffin embedded tissues. *PLoS ONE* **13**, e0203608 (2018).

[00362] 32. Malnar, M. & Režen, T. Factors affecting RNA quantification from tissue long-term stored in formalin. *J. Pharmacol. Toxicol. Methods* **96**, 61–66 (2019).

[00363] 33. von Ahlfen, S., Missel, A., Bendrat, K. & Schlumpberger, M. Determinants of RNA Quality from FFPE Samples. *PLoS ONE* **2**, e1261 (2007).

[00364] 34. Kresse, S. H. *et al.* Evaluation of commercial DNA and RNA extraction methods for high-throughput sequencing of FFPE samples. *PLoS ONE* **13**, e0197456 (2018).

[00365] 35. Cieslik, M. *et al.* The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).

[00366] All citations are hereby incorporated by reference.

[00367] The present invention has been described with regard to one or more embodiments. However, it will be apparent to persons skilled in the art that a number of variations and modifications can be made without departing from the scope of the invention as defined in the claims. Therefore, although various embodiments of the invention are disclosed herein, many adaptations and modifications may be made within the scope of the invention in accordance with the common general knowledge of those skilled in this art. Such modifications include the substitution of known equivalents for any aspect of the invention in order to achieve the same result in substantially the same way. Numeric ranges are inclusive of the numbers defining the range. In the specification, the word “comprising” is used as an open-ended term, substantially equivalent to the phrase “including, but not limited to,” and the word “comprises” has a corresponding meaning. It is to be however understood that, where the words “comprising” or “comprises,” or a variation having the same root, are used herein, variation or modification to “consisting” or “consists,” which excludes any element, step, or ingredient not specified, or to “consisting essentially of” or “consists essentially of,” which limits to the specified materials or recited steps together with those that do not materially affect the basic and novel characteristics of the claimed invention, is also contemplated. Citation of references herein shall not be

construed as an admission that such references are prior art to the present invention. All publications are incorporated herein by reference as if each individual publication was specifically and individually indicated to be incorporated by reference herein and as though fully set forth herein. The invention includes all embodiments and variations substantially as hereinbefore described and with reference to the examples and drawings.

WHAT IS CLAIMED IS:

1. A method for ligating single stranded DNA molecules comprising:

- i) providing a first single stranded DNA molecule comprising a known sequence of at least 10 nucleotides in length;
- ii) providing a second single stranded DNA molecule comprising an unknown sequence of least 10 nucleotides in length;
- iii) providing a single stranded RNA molecule comprising a 5' end substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end comprising a random sequence of least 10 ribonucleotides in length;
- iv) providing a DNA ligase; and
- v) combining the first single stranded DNA molecule, the second single stranded DNA molecule, the single stranded RNA molecule and the DNA ligase,

wherein the first single stranded DNA molecule and the second single stranded DNA molecule are ligated to provide a ligated single stranded nucleic acid molecule when the unknown sequence of the first single stranded DNA molecule is substantially complementary to the random sequence of the single stranded RNA molecule.

- 2. The method of claim 1 wherein the first single stranded DNA molecule is blocked at the 3' end.
- 3. The method of claim 1 or 2 wherein the first single stranded DNA molecule further comprises a spacer sequence adjacent to the 3' end of the known sequence.
- 4. The method of any one of claims 1 to 3 wherein the known sequence of the first single stranded DNA molecule is from ten to 100 nucleotides in length.
- 5. The method of any one of claims 1 to 4 wherein the unknown sequence of the second single stranded DNA molecule is from ten to 10,000 nucleotides in length.
- 6. The method of any one of claims 1 to 5 wherein the single stranded RNA splint molecule is about 20 nucleotides in length and wherein 10 nucleotides at the 5' end are substantially complementary to the known sequence of the first single

stranded DNA molecule and the random sequence is substantially complementary to the unknown sequence of second single stranded DNA molecule.

7. The method of any one of claims 1 to 6 wherein the first single stranded DNA molecule, the second single stranded DNA molecule, and the single stranded RNA molecule are combined under conditions suitable for annealing of the first single stranded DNA molecule and the second single stranded DNA molecule to the single stranded RNA molecule.

8. The method of any one of claims 1 to 7 further comprising an antiprimer molecule, wherein the first single stranded DNA molecule is combined with the single stranded RNA molecule and the second single stranded DNA molecule is combined with the antiprimer sequence prior to combination of the first single stranded DNA molecule and the single stranded RNA molecule duplex and the second single stranded DNA molecule and antiprimer duplex.

9. The method of any one of claims 1 to 8 wherein the DNA ligase is a *Chlorella* virus DNA ligase.

10. The method of any one of claims 1 to 9 wherein the second single stranded DNA molecule is *in vitro* ssDNA or single stranded cDNA.

11. The method of any one of claims 1 to 10 wherein the second single stranded DNA molecule further comprises a second known sequence 5' to the unknown sequence.

12. The method of any one of claims 1 to 11 wherein the second single stranded DNA molecule further comprises a primer sequence at the 5' end.

13. The method of claim 12, further comprising amplifying the ligated single stranded nucleic acid molecule.

14. The method of claim 13 wherein the amplifying is performed by PCR.

15. The method of claim 14 wherein the PCR is linear PCR or nested PCR.

16. A kit comprising:

i) a first single stranded DNA molecule comprising a known sequence of at least 10 nucleotides in length;

ii) a second single stranded DNA molecule comprising an unknown sequence of least 10 nucleotides in length; and

iii) a single stranded RNA *splint* molecule comprising a 5' end substantially complementary to the known sequence of the first single stranded DNA molecule and a 3' end comprising a random sequence of least 10 ribonucleotides in length.

17. The kit of claim 16 further comprising a DNA ligase.

18. The kit of claim 17 wherein the DNA ligase is a splint ligase.

19. The kit of any one of claims 16 to 18 further comprising instructions for use.

20. A computer implemented method for identifying from a plurality of read pairs R1 and R2 of a sequenced sample with inserts derived from a set of PCR amplified adapter ligated DNA molecules, any target gene fused with any other DNA sequence which are potentially clinically actionable, comprising:

generating a k-mer index of predetermined length k from a reference genome or a transcriptome;

mapping each sequencing read from the plurality of read pairs to exons of the reference genome with the k-mer index to identify corresponding exon regions of the sequencing read, and generating for each sequencing read a mapping vector including one or more blocks, where each block is either an exon block defined by a range of consecutive k-mer matches and their corresponding locations on an exon of the reference genome, or a mismatch block defined by a length of consecutive k-mer mis-matches;

identifying exon blocks with overlapping consecutive k-mer matches in each of the mapping vectors, and merging the same exon blocks with adjustment of mapping coordinates or retaining the exon block having predetermined high priority criteria to resolve mapping ambiguities;

combining both the mapping vectors corresponding to each read pair and sorting by genomic coordinates to provide intermediate mapping vectors, and merging any exon blocks with the same respective exons with adjustment of mapping

coordinates to generate merged mapping vectors;

identifying fusion junction candidates in the merged mapping vectors by searching for a pair of consecutive exon blocks from different genes, called fusion partner genes, with a mismatch block in between, and recording the corresponding read pairs for each of the identified fusion junction candidates;

clustering all the fusion junction candidates based on genomic coordinates of junction points of the pairs of consecutive exons from different genes where junctions closer than a predetermined number of base pairs are considered to be the same fusion junction candidate, and calculating a first predetermined set of metrics based on the read pairs including read depth;

assembling a fusion transcript for each fusion junction candidate cluster using the corresponding recorded read pairs to generate assembled transcripts, and calculating a second predetermined set of metrics from the assembled transcripts;

mapping the plurality of read pairs to the fusion transcripts, and calculating a third predetermined set of metrics based on the mapping output including a number of mapped reads; and

categorizing any fusion transcript as being a high confidence reportable fusion when

the orientation of both fusion partners is concordant,
exactly one of the partners is a gene targeted by the assay,
minimum thresholds for specific metrics of the first, second and third
predetermined set of metrics are met, and
the number of mapped reads and the read depth exceed
predetermined thresholds.

21. The method of claim 20, wherein exon blocks having less than a minimum mapped bps are filtered out.
22. The method of claim 21, wherein the minimum mapped bps is 24 bps.
23. The method of claim 20, wherein each of the mapping vectors is represented

by the expression (e_n, read_st→read_end, map_st→map_end) for each exon block and (miss, miss_st→miss_end, len) for each mismatch block, where e_n is the exon unique identifier, read_st and miss_st are the starting base pair on the sequencing read, read_end and miss_end are the ending base pair on the sequencing read, map_st is the starting mapped region on the exon for read_st, map_end is the ending mapped region on the exon for read_end, miss is a status identifier for a mis-match, and len is the bps length of the mismatch block.

24. The method of claim 20, further including error correcting the mapping vectors to remove mismatch blocks interposed between two exon blocks with the same exon.

25. The method of claim 24 wherein error correcting includes removing the length of consecutive k-mer mis-matches of the mismatch block interposed between a first range of consecutive k-mer matches of a first exon block and a second range of consecutive k-mer matches of a second exon block of the same exon, when a continuation of the first range by the length of consecutive k-mer is sequentially followed by a beginning of the second range.

26. The method of claim 25, wherein error correcting to remove a mismatch block between two consecutive exon blocks with the same exon includes determining when read_end + len + 1 for the first exon block = read_st for the subsequent exon block to remove the mismatch block.

27. The method of claim 25, wherein error correcting to remove the length of consecutive k-mer mis-matches includes determining when map_end + len + 1 for a first exon = map_st for a subsequent exon to remove the length of consecutive k-mer mis-matches.

28. The method of claim 23, wherein identifying fusion junction candidates includes executing an error correction applied on k-mer gaps g_1, g_2 and g_k of the fusion junction candidates when g_k > 0, where

g_1 is a distance between map_end of the first exon block to the corresponding exon extremity,

g_2 is a distance between map_st of the second exon block to the corresponding exon extremity,

g_k is the difference between the k-mer length k and the mismatch block len, and the k-mer gap correction steps include

while g_k > 0

- a. if $g_1 > 0$ then
 - i. $g_k = g_k - 1$
 - ii. $g_1 = g_1 - 1$
- b. else If $g_2 > 0$ then
 - i. $g_k = g_k - 1$
 - ii. $g_2 = g_2 - 1$

29. The method of claim 23, wherein the predetermined high priority criteria include

an exon block with a larger mapping range than the other,
an exon block where its exon comes from an assay targeted gene,
an exon block matching a list of preferred candidates, or
when both exons are from the same gene, selecting the exon with highest priority, calculated from sorting the corresponding reference genome transcripts for each exon in descending order of length.

30. The method of claim 29, wherein if none of the predetermined high priority criteria is met, then remove the exon block having the largest exon gap, where an exon gap is a distance between map_end of a first exon and map_st of a subsequent exon to their respective exon extremities.

31. The method of claim 20, wherein the predetermined number of base pairs to cluster junction candidates is 50 or less.

32. The method of claim 20, further including determining a unique amplicon end by retrieving the last exon block mapping coordinate map_end and adding to it the length of the last mismatch block for each of the R2 reads.

33. The method of claim 32, wherein identifying fusion junction candidates further includes identifying normal junction candidates in the merged mapping vector by searching for pairs of consecutive exon blocks from a same gene with a mismatch block in between, and recording the corresponding read pairs for each of the identified normal junction candidates.

34. The method of claim 33, wherein the first predetermined set of metrics includes a number of unique amplicon ends, an estimated length of the fusion

transcript, and a ratio between fusion junction candidates and normal junction candidates.

35. The method of claim 34, wherein the estimated length of the fusion transcript is determined by the average mapped length of all the read pairs in the corresponding junction candidate.

36. The method of claim 30, wherein assembling the fusion transcript includes the following steps for each fusion junction candidate cluster

- a. constructing a de-Bruijn graph with the k-mers of all cluster reads,
- b. executing bubble correction and tip removal algorithms, to correct sequencing errors,
- c. determining a number of contigs by traversing the de-Bruijn graph starting with a node that has the k-mer of the gene-specific primer that captured the fusion, and
- d. determining a longest contig as being a true assembled fusion transcript.

37. The method of claim 36, wherein the second predetermined set of metrics includes an assembled fusion transcript length.

38. The method of claim 37, wherein, the third predetermined set of metrics includes the number of unique transcript ends determined from the mapping output.

39. The method of claim 20, wherein categorizing further includes categorizing any assembled fusion transcript as being an artifact if either the orientation of the fusion partners is not concordant or both partners are genes not targeted by the assay.

40. The method of claim 20, wherein categorizing further includes categorizing any assembled fusion transcript as being filtered if minimum thresholds for any one of the specific metrics is not met.

41. The method of claim 20, wherein categorizing further includes categorizing any assembled fusion transcript as being a supporting fusion if another assembled fusion with more reads is found on the same partner genes.

42. The method of claim 38, wherein categorizing further includes

calculating a fusion score for each assembled fusion transcript based on any number of metrics of the first, second and third predetermined set of metrics, and the scoring function is calculated using the expression

$$s(j) = \sum_i^n w_i \cdot \frac{m_{ij} - \mu_i}{|\mu_i|}, \text{ for } i \in \text{metrics}$$

, where m_{ij} is the value of the metric i for the assembled fusion transcript j , w_i is the weight associated with metric i , and μ_i is the *expected mean* of metric i , calculated from a control dataset with known fusions,

comparing the fusion score to a first predetermined threshold and a second predetermined threshold, and

categorizing the assembled fusion transcript as medium confidence if the fusion score is greater than the first predetermined threshold,

categorizing the assembled fusion transcript as low confidence if the fusion score is less than the first predetermined threshold but greater than the second predetermined threshold, or

categorizing the assembled fusion transcript as filtered if the fusion score is less than the second predetermined threshold, and

outputting a result where low and medium confidence are subjected to orthogonal testing.

43. The method of claim 42, wherein the metrics used for the scoring function include the read depth, the number of unique reads, the assembled fusion transcript length, and a ratio of read depth between fusion junction candidates and normal junction candidates with the same gene as the fusion target gene.