Computational proteogenomic identification and functional interpretation of translated fusions and micro structural variations in cancer: Supplementary Materials

Yen-Yi Lin[1,†], Alex Gawronski[1,†], Faraz Hach[1,3,†], Sujun Li[2], Ibrahim Numanagić[1], Iman Sarrafi[1,3], Swati Mishra[5], Andrew McPherson[1], Colin Collins[3,4], Milan Radovich[5], Haixu Tang[2], and S. Cenk Sahinalp[1,2,3,*]

[1]*School of Computing Science, Simon Fraser University, Burnaby, BC, Canada*
[2]*School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[3]*Vancouver Prostate Centre, Vancouver, BC, Canada*
[4]*Dept. of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada*
[5]*Department of Surgery, Indiana University, School of Medicine, Indianapolis, IN, USA*
[†]*The authors wish it to be known that the first three authors should be regarded as joint first authors.*
[*]*To whom the correspondence should be addressed*

November 24, 2017

Contents

1 Advantage of a Unified Algorithm for Detecting Structual Variants **2**

2 Methods of Detecting Aberrations in RNA-Seq and WGS datasets **3**
2.1 Fusion Detection in Transcriptomic Data . . . 3
2.2 MiStrVar: Detection of microSVs on Matching WGS and RNA-Seq Data . . . 3
2.2.1 MiStrVar Dynamic Programming Formulation . . . 6
2.2.2 Full Formulation . . . 7
2.3 Identification of Translated Sequence Aberrations . . . 10
2.4 Class-Specific Peptide-Level FDR in ProTIE . . . 11
2.5 Identification of Transcriptomic Evidence for Genomic Aberrations . . . 12

3 Simulation and Cell line Results of MiStrVar **12**
3.1 MicroSV Detection Performance of MiStrVar on Simulated Data . . . 12
3.2 MicroSV Detection Results in HCC1143 Cell Line . . . 14
3.2.1 Chromatogram interpretation for hetrozygous microSVs . . . 14
3.2.2 RNA-Seq Support for microSVs in HCC1143 Cell Line . . . 16

3.3 Further Support and Evaluation of the Detected microSVs in the HCC1143 Cell Line 16

4 Overview of CPTAC datasets 17

5 High-Confidence deFuse Calls in CPTAC Datasets 17
5.1 Proteomics Support of High-Confidence deFuse Calls 19

6 Summary of microSVs in CPTAC datasets 19
6.1 MicroSVs Detection Results is WGS datasets 19
6.2 RNA-Seq support for microSVs in CPTAC Breast Cancer Patients 20

7 Mechanistic and Functional Interpretation of microSV and Fusion Peptides Detected in TCGA/CPTAC BRCA Dataset 23
7.1 Fusions Peptides 23
7.2 Genomic MicroSVs Peptides 23

1 Advantage of a Unified Algorithm for Detecting Structual Variants

MiStrVar uses a unified dynamic programming formulation, superior to tools that identify each type of variant individually, especially because these tools misinterpret certain variants, such as inversions, as a combination of other variants, as illustrated in Figure 1; here a single inversion event shown in A can also be interpreted as a deletion and an insertion with several mismatches, as shown in B. It is clear that A is the more likely/parsimonious scenario involving just one inversion; however, each of these interpretations may receive similar score by a standard variant caller. This happens more often than thought: Figure 1C illustrates an inversion event in SLC3A1 3'UTR discovered by MiStrVar that has been misinterpreted in dbSNP. Our unified formulation ensures to identify the single most parsimonious microSV supported by a given contig, reducing the number of false positives.

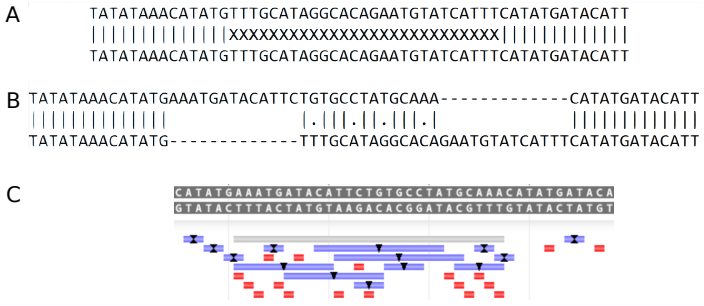

Figure 1: Example to illustrate the advantage of a unified algorithm for detecting structural variants. **(A)** Optimal alignment of two sequences using the unified algorithm, in this case a 27bp microinversion in SLC3A1 3'UTR. **(B)** Optimal Smith-Waterman alignment of the same two sequences.**(C)** Cluster of dbSNP entries in the inverted region. Blue lines are indels, red lines are single nucleotide polymorphisms and the grey line is a multiple nucleotide polymorphism. Most, if not all, are incorrect.

2 Methods of Detecting Aberrations in RNA-Seq and WGS datasets

2.1 Fusion Detection in Transcriptomic Data

We use deFuse [1] as our primary fusion detection tool. deFuse detects candidate fusion events and constructs the fusion sequences spanning the junction of the fusion breakpoints from discordantly mapped RNA-Seq reads based on maximum parsimony. deFuse maps paired-end reads to the reference genome. All reads that can be mapped to the reference within expected distance range and correct orientation (i.e., concordantly) are discarded. Among the remaining reads, discordantly mapped read-pairs and one-end anchors are considered for fusion transcripts. More precisely, (1) discordantly mapped *spanning reads*, where the two ends map to two different genes, anchor the search for fusion breakpoints, and (2) *split reads*, where one end maps to a single gene and the other end spans the junction of a fusion breakpoint, help pinpoint the junctions.

As a first step, deFuse identifies potential fusion events by clustering spanning reads. Many reads can be mapped to multiple locations due to homologous genes and splice variants; the vast majority of available tools either discard these ambiguous mappings or randomly select one mapping location among them. The unique feature of deFuse is that it resolves mapping ambiguity by jointly considering all mapping loci of each of the reads and iteratively identifying the one that is shared by the maximum possible number of reads as their true mapping locus. As a second step deFuse identifies split reads corresponding to each potential fusion event using a dynamic programming formulation, to exactly identify the fusion breakpoints at nucleotide level resolution.

2.2 MiStrVar: Detection of microSVs on Matching WGS and RNA-Seq Data

We consider as microSV those genomic sequence alterations shorter than a few hundred bps and at least 5bp. Aberrations of length <5bp are many times too short to differentiate from single nucleotide variants (SNVs). We are interested in microSVs that exist "in" or "near" exonic regions that may lead to translation into a novel peptide and contribute to the cancer phenotype. For this reason MiStrVar uses a masked reference genome primarily comprised of regions of potential transcription.

MiStrVar works in three major steps: in step A it collects all one-end anchors (see below) from WGS data, and cluster them according to the mapping loci. In step B, unmapped ends of the reads in each cluster are assembled into a contig representing the genomic sequence alteration implied by a microSV. In step C, the exact nature of each microSV is determined by aligning the associated contigs with the reference genome. (See Figure 2 for an overview.) Below we describe each of these three steps in detail.

In **step (A)**, MiStrVar identifies all one-end anchors (OEA) in the read data: an OEA is a paired-end-read for which only one end maps to the reference genome within a user defined error threshold. Once all reads are (multiply) mapped to a reference genome using mrsFAST-ultra [2, 3], and all OEAs are extracted, the mapped ends of OEAs are clustered based on the mapping loci. MiStrVar provides the user two options for cluster identification, each satisfying one of the following distinct goals.

1. For applications where sensitivity is of high priority, MiStrVar employs a sweeping algorithm for OEA mapping loci (introduced for VariationHunter [4]) which identifies *every possible read mapping cluster* that can support a microSV in $O(n+\#clust)$ time where n is

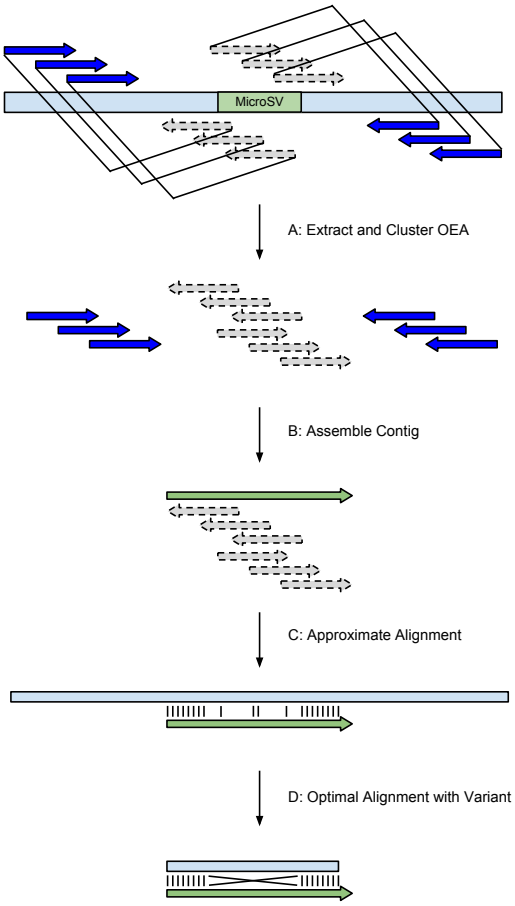

Figure 2: A sketch of our computational framework for detecting microSVs in tumor samples. **A.** All one-end-anchors (OEA) are extracted from the mapping file and clustered based on the mapped mates. **B.** The unmapped mates are then assembled into a contig. **C.** The contig is aligned to the reference within 1Kb from the mapped mates. **D.** The reference is clipped and the optimal alignment is found using a dynamic programming formulation allowing for a structural alteration event.

the total number of read mapping loci and $\#clust$ is the total number of clusters obtained (i.e. the output size). Note that each read mapping locus can be a part of several OEA clusters.

2. For applications where running time is of high priority, MiStrVar employs an iterative greedy strategy, which "anchors" the first cluster with the "leftmost" mapping locus of an OEA on the genome and extends the cluster to include any other OEA mapping that overlaps with the anchor OEA mapping. Once all such OEAs mappings are added to the cluster, the iterative strategy greedily anchors the next cluster with the first OEA mapping not included in the previous cluster. Note that in this strategy each OEA mapping can be a part of a single cluster.[1]

In **step (B)**, for each OEA cluster identified in step (A), MiStrVar assembles the unmapped end of the reads to form contigs (of length <400bp in practice) by aiming to solve the **dominant superstring (DSS)** problem defined as follows.
Given a collection of length k strings $S = \{S_1, \ldots, S_n\}$ from the standard DNA alphabet, identify a subset $S' = \{S'_1, \ldots, S'_m\}$ of S and its shortest superstring S^* (so that each S'_i is a substring of S^*) for which $Obj(S^*) = |S^*|/m$ is as small as possible.[2]

As per the well known shortest superstring problem, DSS is NP-hard [5]. However a simple greedy strategy provides a constant factor approximation to the shortest superstring problem - which is MAX-SNP-hard and thus a polynomial time approximation scheme is unlikely [6]. MiStrVar thus employ a similar greedy strategy to solve the DSS problem as follows. It starts with $S^* = \{S_i, S_j\}$, a pair of strings from S whose suffix-prefix overlap is maximum possible. In each successive iteration, it greedily identifies the string S_ℓ from $S - S^*$ whose inclusion in S^* improves $Obj(S^*)$ the most, until it can not be improved further. As a result, MiStrVar identifies a dominant superstring for the unmapped ends of each cluster of OEAs as a contig.
In **step (C)**, each contig associated to an OEA cluster is aligned to a region (of length several kilobases long) surrounding the OEA mapping loci, first through a simple *local-to-global* sequence alignment algorithm, that does not consider any structural alteration. (The reverse complement of the contig is also aligned to the same region.) The start and end position of this first, crude alignment is used to determine the approximate locus and length of the potential microSV implied by the contig. The exact microSV breakpoints are obtained in the next step through a more sophisticated alignment that considers structural alterations, which is applied to the portion of the reference genome restricted by the first alignment. The dynamic programming formulation for this alignment is an extension of the Schöniger-Waterman algorithm [7] which was designed to capture inversions in the alignment. We give details of this extension below, which enables the user to

1. discover the single best optimal event, rather than an arbitrary number of events [3],
2. handle gaps extending over breakpoints (in cases of missing contig sequence), and,
3. simultaneously predict duplications, insertions, deletions and SNVs in addition to inversions.

[1] All experimental results in this paper are based on the second clustering algorithm due to running time constraints.
[2] Since this formulation ignores reads that include read errors it requires the read coverage to be "reasonably" high and the read errors to be comparatively rare.
[3] In rare cases where two microSVs occur in close proximity that they would fall within the same contig, one of the microSVs would not be reported. More commonly, if another signature exists in a contig, it is not a true SV and this approach improves precision.

2.2.1 MiStrVar Dynamic Programming Formulation

The basic recurrence used in the Schöniger-Waterman alignment algorithm is as follows.

$$U(i,j) = max\{U(i-1,j) + \beta, W(i-1,j) + \alpha + \beta\} \quad (1)$$

$$V(i,j) = max\{V(i,j-1) + \beta, W(i,j-1) + \alpha + \beta\} \quad (2)$$

$$W(i,j) = max\{\underset{g,h}{max}\{W(g-1,h-1) + Z(g,h,i,j) + \gamma\},$$
$$W(i-1,j-1) + s(a_i,b_j), U(i,j), V(i,j), 0\} \quad (3)$$

$$Z(g,h,i,j) = W_I(\overline{S_{h,j}}, T_{g,i}) \quad (4)$$

This recurrence is itself an extension of the Smith-Waterman formulation for local alignment. Equations 1 and 2 correspond to typical affine gap matrices with gap opening (α) and gap extension (β) penalties. The recurrence includes an additional score matrix Z (Equation 4) which produces the optimal global alignment of a subsequence from g to i with the reverse compliment of subsequence from h to j. This score is added to the score of $W(g-1,h-1)$ as a fifth choice for function $W(i,j)$ (Equation 3) with an additional penalty γ. This can be interpreted as a "correction" to one of the sequences. It should also be noted, since Z is computed with a standard global alignment, it allows for mismatches and gaps to occur within the inversion. Therefore, germline inversions with SNPs or indels would still be detectable.

In order to limit this formulation to an optimal, single inversion, we introduce three additional matrices (W', U', V'), all computed with the basic Smith-Waterman algorithm. The term $W(g-1,h-1)$ in Equation 3 is replaced with $W'(g-1,h-1)$. This ensures that at any position (i,j), an alignment with a single inversion ($W'(g-1,h-1) + Z(g,h,i,j)$) is compared to an alignment with at most one inversion ($W(i,j)$). This approach can be easily generalized to k optimal inversions by introducing additional matrices, at a computational cost of $O(k)$.

Since the original formulation adds the score of $Z(g,h,i,j)$ only to the value of $W(g-1,h-1)$, the affine gap matrices computed by $U(g-1,h)$ and $V(g,h-1)$ are ignored. Furthermore the affine gap scores of the global alignment W_I are not considered. Therefore a gap cannot cross the boundary without incurring an additional gap opening penalty. This may be a rare case in general, but in our application if the inversion is greater than the read length it is possible for a contig to not include the entire inversion, which requires a gap over the breakpoint in the alignment. To solve this problem , all the affine gap scores of the global, inversion alignment are added to the corresponding affine gap functions U and V (and *vice versa*) with the appropriate extension penalties. See Section 2.2.2 for the full formulation.

$$W(i,j) = max\{\underset{g,h}{max}\{W(g-1,h-1) + Z(g,h,i,j) + \gamma\},$$
$$\underset{g}{max}\{W(g-1,j) + D(g,i) + \theta\},$$
$$W(i-1,j-1) + s(a_i,b_j), U(i,j), V(i,j), 0\} \quad (5)$$

$$D(g,i) = \underset{i',g'}{max}\{(W_T(i,i') - W_T(g,g') - (i-g)\lambda)/2)\} \quad (6)$$

where W_T is the self-alignment of T with $\alpha = \beta = -\infty$

To include duplications, we substitute Equation 3 with Equation 5 above. This modification adds a sixth case where the score of function $D(g,i)$ is added to $(g-1,j)$ with an additional error θ. $D(g,i)$ (Equation 6) is the error of the best alignment of input string T (in this context, the contig sequence, indexed by g and i) and itself (indexed by g' and i'). Note that this formulation is not restricted to tandem duplications only; specifically, (g,i) does not need to be adjacent to

(g', i'). This allows for the detection of "interspersed duplications", where the "source" sequence can be located anywhere on the contig. This error is calculated by subtracting the alignment score from the maximum score for a sequence of length $i - g$, defined by multiplying the length by the match score λ. Since gaps are disallowed, and the mismatch score is $-\lambda$, dividing this number by 2 yields the final mismatch error. Again, this allows for the detection of SNPs within a duplication. Indels can also be detected with an additional computational cost through an alternate error computation. Similar to inversions, this "corrects" the contig sequence, but in this case the score is computed as if the duplicated sequence were removed. The modifications to the recurrence for the purposes of selecting a single optimal duplication and handling affine gap scores are very similar to those we described for inversions.

If neither including an inversion or a duplication improves the alignment score over the basic Smith-Waterman alignment that allows insertions, deletions and mismatches, these events are extracted through traceback and reported.

The time and space requirements of our algorithm is $O(n^4)$ and $O(n^2)$ respectively, which is identical to the original complexity of Schöniger-Waterman method. Although the running time may seem impractical on a genome-wide scale, two restrictions can improve it drastically. The first is the range of variant lengths, defined by the difference between the minimum and maximum length ($max(i - g) - min(i - g)$ or $max(j - h) - min(j - h)$). This is accurately estimated for both inversions and duplications from the basic $O(n^2)$ alignments, and the largest of the two ranges is selected. The second is the number of insertions and deletions within the variant, or more specifically the difference in length between the sub-sequence (g, i) and (h, j). Indels within variants are highly unlikely to occur in real data, especially in somatic calls, so we disallow any insertions and deletions in the variants. These two restrictions reduce the running time to $O(n^2 r)$, where $r \ll n$ is the range estimate. The restriction on indels does reduce the scope of possible structural variants the algorithm can detect. However, for detection of somatic calls this would likely be negligible since two mutations would need to occur one after the other at the same location. If the first mutation leads to a phenotype, there would be no further selection for another mutation. Indeed in the context of germline event detection, cases such as this may occur more frequently and would be missed when using this optimization.

2.2.2 Full Formulation

This formulation uses $O(n^4)$ space for clarity of presentation. In the implementation, the Z matrices are computed iteratively for a fixed j and g and reused for each h and i, reducing the memory complexity. Furthermore we use global alignment [8] in practice rather than local alignment originally used by Schöniger and Waterman. This is because we only align to a portion of the reference genome restricted by the first alignment (as outlined above) which is approximately the same length as the contig.

$$U(0,0) = V(0,0) = W(0,0) = U_I(0,0) = V_I(0,0) = W_I(0,0) = W_T(0,0) = 0 \quad (7)$$

$$U(0,y) = U_I(0,y) = \alpha + y\beta \quad (8)$$
$$V(0,y) = V_I(0,y) = -\infty \quad (9)$$
$$W(0,y) = W_I(0,y) = \alpha + y\beta \quad (10)$$
$$W_T(0,y) = 0 \quad (11)$$
$$U(x,0) = U_I(x,0) = -\infty \quad (12)$$
$$V(x,0) = V_I(x,0) = \alpha + y\beta \quad (13)$$
$$W(x,0) = W_I(x,0) = \alpha + y\beta \quad (14)$$
$$W_T(x,0) = 0 \quad (15)$$

$$U_{Zk}(g,h,g,h) = V_{Zk}(g,h,g,h) = W_{Zk}(g,h,g,h) = 0 \quad (16)$$

$$U_{Z1}(g,h,g+1,y) = \alpha + y\beta \quad (17)$$
$$V_{Z1}(g,h,g+1,y) = -\infty \quad (18)$$
$$W_{Z1}(g,h,g+1,y) = \alpha + y\beta \quad (19)$$
$$U_{Z2}(g,h,g+1,y) = \alpha + y\beta \quad (20)$$
$$V_{Z2}(g,h,g+1,y) = -\infty \quad (21)$$
$$W_{Z2}(g,h,g+1,y) = \alpha + y\beta \quad (22)$$
$$U_{Z3}(g,h,g+1,y) = y\beta \quad (23)$$
$$V_{Z3}(g,h,g+1,y) = -\infty \quad (24)$$
$$W_{Z3}(g,h,g+1,y) = y\beta \quad (25)$$

$$U_{Z1}(g,h,x,h+1) = \alpha + y\beta \quad (26)$$
$$V_{Z1}(g,h,x,h+1) = \alpha + y\beta \quad (27)$$
$$W_{Z1}(g,h,x,h+1) = -\infty \quad (28)$$
$$U_{Z2}(g,h,x,h+1) = y\beta \quad (29)$$
$$V_{Z2}(g,h,x,h+1) = y\beta \quad (30)$$
$$W_{Z2}(g,h,x,h+1) = -\infty \quad (31)$$
$$U_{Z3}(g,h,x,h+1) = \alpha + y\beta \quad (32)$$
$$V_{Z3}(g,h,x,h+1) = \alpha + y\beta \quad (33)$$
$$W_{Z3}(g,h,x,h+1) = -\infty \quad (34)$$

where
$0 < g \leq \mid S \mid,$
$0 < h \leq \mid T \mid,$
$0 < x \leq \mid S \mid,$
$0 < y \leq \mid T \mid,$
$0 < k \leq 3$

Figure 3: Matrix initialisation.

$$U_I(i,j) = max\{U_I(i-1,j) + \beta, W_I(i-1,j) + \alpha + \beta, Z_U, D_U\} \tag{35}$$

$$Z_U = \begin{cases} \max\limits_k\{Z(i,j,k)\} + \gamma, & \text{if } U_{Zk}(g,h,i,j) > max\{V_{Zk}(g,h,i,j), W_{Zk}(g,h,i,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{36}$$

$$D_U = \begin{cases} \max\limits_g\{U(g-1,j) + D(g,i) + \theta\}, & \text{if } U(g-1,j) > max\{V(g-1,j), W(g-1,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{37}$$

$$V_I(i,j) = max\{V_I(i,j-1) + \beta, W_I(i,j-1) + \alpha + \beta, Z_V, D_V\} \tag{38}$$

$$Z_V = \begin{cases} \max\limits_k\{Z(i,j,k)\} + \gamma, & \text{if } V_{Zk}(g,h,i,j) > max\{U_{Zk}(g,h,i,j), W_{Zk}(g,h,i,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{39}$$

$$D_V = \begin{cases} \max\limits_g\{V(g-1,j) + D(g,i) + \theta\}, & \text{if } V(g-1,j) > max\{U(g-1,j), W(g-1,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{40}$$

$$W_I(i,j) = max\{ \\ max\{U_I(i-1,j-1), V_I(i-1,j-1), W_I(i-1,j-1)\} + s(a_i,b_j), Z_W, D_W\} \tag{41}$$

$$Z_W = \begin{cases} \max\limits_k\{Z(i,j,k)\} + \gamma, & \text{if } W_{Zk}(g,h,i,j) > max\{U_{Zk}(g,h,i,j), V_{Zk}(g,h,i,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{42}$$

$$D_W = \begin{cases} \max\limits_g\{W(g-1,j) + D(g,i) + \theta\}, & \text{if } W(g-1,j) > max\{U(g-1,j), V(g-1,j)\} \\ -\infty, & \text{otherwise} \end{cases} \tag{43}$$

Figure 4: Main alignment matrices.

$$U(i,j) = max\{U(i-1,j) + \beta, W(i-1,j) + \alpha + \beta\} \tag{44}$$

$$V(i,j) = max\{V(i,j-1) + \beta, W(i,j-1) + \alpha + \beta\} \tag{45}$$

$$W(i,j) = max\{max\{U(i-1,j-1), V(i-1,j-1), W(i-1,j-1)\} + s(a_i,b_j)\} \tag{46}$$

Figure 5: Global alignment without SV.

$$W_T(i,j) = W_T(i-1,j-1) + s(a_i,b_j) \tag{47}$$

Figure 6: Ungapped local self-alignment.

$$U_{Zk}(g,h,i,j) = max\{U_{Zk}(g,h,i-1,j)+\beta, W_{Zk}(g,h,i-1,j)+\alpha+\beta\} \quad (48)$$

$$V_{Zk}(g,h,i,j) = max\{V_{Zk}(g,h,i,j-1)+\beta, W_{Zk}(g,h,i,j-1)+\alpha+\beta\} \quad (49)$$

$$W_{Zk}(g,h,i,j) = max\{U_{Zk}(g,h,i-1,j-1), V_{Zk}(g,h,i-1,j-1), W_{Zk}(g,h,i-1,j-1)\} + \overline{s}(a_i,b_j) \quad (50)$$

$$Z(i,j,1) = \max_{g,h}\{U_{Z1}(g,h,i,j), V_{Z1}(g,h,i,j), W_{Z1}(g,h,i,j)\} + U(g,h) \quad (51)$$

$$Z(i,j,2) = \max_{g,h}\{U_{Z2}(g,h,i,j), V_{Z2}(g,h,i,j), W_{Z2}(g,h,i,j)\} + V(g,h) \quad (52)$$

$$Z(i,j,3) = \max_{g,h}\{U_{Z3}(g,h,i,j), V_{Z3}(g,h,i,j), W_{Z3}(g,h,i,j)\} + W(g,h) \quad (53)$$

$$D(g,i) = \max_{i',g'}\{(W_T(i,i') - W_T(g,g') - (i-g)\lambda)/2)\} \quad (54)$$

Figure 7: SV alignments.

where

$0 < g' < i' < g < i < g' < i' \leq \mid S \mid$

$0 < h < j \leq \mid T \mid$

$0 < \ k \leq 3$

$\alpha = \beta = -\infty$ for W_T which is equivalent to W but a self-alignment of T

$$s(a_i,b_j) = \begin{cases} 1, & \text{if } T_i = S_j \\ -1, & \text{otherwise} \end{cases}$$

$$\overline{s}(a_i,b_j) = \begin{cases} 1, & \text{if } T_i = \overline{S}_j \\ -1, & \text{otherwise} \end{cases}$$

2.3 Identification of Translated Sequence Aberrations

ProTIE provides the ability to detect translated aberrations by searching mass spectra against an aberrant peptide database. More specifically, given transcriptomic breakpoints pointing to fusions or microSVs, ProTIE identifies respective aberrant peptides from proteomic data by first generating a peptide database, and then identifying aberrant peptides based on mass spectrometry search results.

The database is a combination of known (wildtype) human peptides and either the fusion peptides (used for fusion discovery), derived from the fusion breakpoints suggested by deFuse (and/or Comrad/nFuse), or microSVs breakpoint peptides, derived from the breakpoints suggested by MiStrVar. For each fusion or microSV breakpoint, six different reading frames (both forward and backward reading frames) are considered - until a stop codon. Potential peptides (of residue length five or more) resulting from each breakpoint junction, as well as downstream peptides that result from a shift in the reading frame, are included in the peptide database allowing at most one miscleavage site (i.e. consisting of at most two amino acids of K or R for

trypsin specificity used in the CPTAC data) as aberrant peptides.[4]

ProTIE uses Ensembl human protein database GRCh37.70 (Ensembl, `ftp://ftp.ensembl.org/pub/release-70/fasta/homo_sapiens/pep/Homo_sapiens.GRCh37.70.pep.all.fa.gz`) to derive the known (wildtype) human peptides. Note that the Ensembl database includes 104,785 peptide sequences among which 75,994 are annotated as known peptides, 10,449 are annotated as novel peptides and an additional 18,342 are annotated as putative peptides. ProTIE includes only the set of known peptides in the primary peptide database it establishes; however it also provides information for the mass spectra that do not match known or aberrant peptides, but can be matched to novel or putative sequences.[5]

ProTIE conducts peptide identification by searching tandem mass spectra (MS/MS) against the peptide database it sets up, as described above. For that, it first converts raw files into Mascot Generic Format (MGF) and uses MS-GF+ [9] engine to perform the search. We adopted the search parameters recommended by the CPTAC Common Data Analysis Pipeline (CDAP) [10] as follows. (1) The precursor mass tolerance is set to 20ppm. (2) The Fragment method is set as 3 for HCD. (3) Instrument is set as 3 for Q-Exactive. (4) Number of tolerable termini is 1. (5) Maximum length of peptide is 50. (6) Modifications include: Carbamidomythyl is fixed in Cystine, Oxidation is set as variable modification in M, iTRAQ 4plex is fixed at N-termini and any Lysine(K) residue.

To obtain a subset of high confidence matches, ProTIE selects only the spectra where the top 20 peaks in the PSMs have matched fragmentation ions. If fewer than 20 peaks exist, all peaks must match. The major fragmentation ions annotated are: b-, b-neutral loss ions, y-, y-neutral loss ions.

We apply 1% spectrum-level FDR control on the identification as suggested in CDAP. Based on the search result, we keep a spectra in ProTIE if its best PSM (in terms of lowest q-value) can not match any known peptides in Ensembl annotation or decoy sequences (i.e. false positives). Spectra that match novel or putative peptides are still kept with special remarks for further analysis.

2.4 Class-Specific Peptide-Level FDR in ProTIE

It has been argued in the literature that stringent class-specific peptide-level FDR estimates may be necessary for reporting novel peptides in proteogenomics studies [11]. In order to address this issue, for any search result provided from MS-GF+, we first cluster all peptide-spectra matches into known or novel categories based on their peptide sequences: a PSM is assigned to the known class if the peptide is a known peptide or the decoy sequence of a known peptide; otherwise it will be assigned to the novel class. We then recalibrate FDR for records in the novel class using original E-value from MS-GF+: a peptide p is assigned the best spectral E-value $E(p)$ it can get from any records in the novel class. Given a PSM M with E-value s, we collect all PSMs in the novel class whose E-value $\leq s$, and calculate the ratio of records containing decoy sequences as the new peptide-level FDR for M. In table

[4]Peptides shorter than five residues are discarded due to mass spectrometry detection range limit.

[5]We establish a single database for the breakpoints identified in all patients, however we maintain the patient for each potential aberrant peptide in order to make sure that mass spectra from a particular patient (or a set of patients) can only match an aberrant peptide from the same patient.

2.5 Identification of Transcriptomic Evidence for Genomic Aberrations

Our pipeline provides the user with the additional ability to jointly analyze matching WGS and RNA-Seq data for identifying transcribed genomic (in fact genetic) microSVs. Given a set of genomic microSVs, along with their breakpoints detected by MiStrVar, our pipeline generates corresponding aberrant transcripts. It then maps RNA-Seq reads to the collection of these aberrant transcripts using mrsFAST-ultra (error threshold, 6%). An RNA-Seq read mapping is said to provide evidence for the transcription of a microSV in two ways: either (i) an RNA-Seq read is (uniquely) mapped across a breakpoint - providing evidence for the transcription of the associated microSV (both in the form of inversions and duplications) ; or (ii) a paired end RNA-Seq read is mapped to the reference transcriptome discordantly (due to change of mapping orientation) - again providing evidence for the transcription of the associated microSV (relevant for inversions only). Note that no read that can be mapped concordantly to a known isoform or potential novel spliceform (through the use of the splice-aware mapper STAR [12]) is considered to be a supportive evidence of a transcribed microSV.

3 Simulation and Cell line Results of MiStrVar

3.1 MicroSV Detection Performance of MiStrVar on Simulated Data

In order to get a sense of its true positive rate, we first applied MiStrVar to the dataset generated by simulating reads from a modified version of the Chromosome 1 of the Venter genome. For this modification, we randomly implanted 205 inversions and 115 duplications, satisfying the following. (1) Minimum distance between consecutive events would be at least 10,000 bp. (2) Inversion lengths vary between 5 and 400 bp. (3) Duplicated segment lengths vary between 5 and 50 bp and the distance between the original and duplicate copies would be no more than 50 bp. (4) Each SV is implanted in a sequence segment (of length at least 20bps longer than the SV itself on both directions) which is uniquely mappable, i.e. reads originating from this segment can only be mapped to the correct locus on the human reference genome hg19 (GRCh37).

On the modified Venter genome, we generated a dataset of 40M error-free paired end reads of length 2x100 bp (providing a coverage of roughly 30x) via wgsim (`https://github.com/lh3/wgsim`). The average insert length and the standard deviation were 450bp and 50bp respectively. These reads were mapped to GRCh37.75 using BWA aln with default parameters and sorted by coordinate using samtools. All compared tools used this BAM file as input.

We generated another dataset from chromosome 1 of the unmodified Venter genome with high coverage (160M paired end reads of length 2x100bp, with a coverage of roughly 120X). This dataset was used to identify microSVs already present in the Venter genome (with respect to the human reference genome).

MiStrVar performs very well on microinversions with a precision and recall of roughly 90% for inversions of length at most 100bp (see Table 1). For longer inversions (101-400bp) recall is even higher, very close to 100% without a significant drop in precision.

In order to compare MiStrVar's performance against available SV detection tools, we used the top five tools based on their performance as assessed by two recent reviews [13, 14]. These reviews evaluated the recall and precision of SV discovery tools across different variant length categories. The first two of these tools are Pindel [15] and SoftSV [14], which exhibited good performance for both inversions and tandem duplications for the shortest set of variants. In addition, Delly [16] and Breakdancer [17] tested well for inversions and ITDetector [18] tested well for duplications only. Note that ITDetector is designed particularly for for finding short

Table 1: Comparison of precision, recall, false discovery rate (FDR) and false negative rate (FNR) of MiStrVar against other SV discovery tools. All tools were run with default parameters and the calls for each microSV type (we only considered the calls made by each tool for that microSV) were called true or false based on the metrics provided by the tools (quality, identity or support, if they exist). The threshold values for each metric were chosen to maximize the F-score (Supplementary Table 2). Only inversions of length ≤400bp were considered in the calculations. If a tool does not provide precise breakpoints, breakpoints falling within a provided range are counted as true positives. Known insertion SNPs were filtered for all duplication results.

		5-100 bp				101-400 bp			
SV Type	**Tool**	**Precision**	**Recall**	**FDR**	**FNR**	**Precision**	**Recall**	**FDR**	**FNR**
Inversions	**MiStrVar**	91.20%	92.68%	8.80%	7.32%	93.10%	98.78%	6.90%	1.22%
	Breakdancer	66.67%	1.63%	33.33%	98.37%	59.00%	95.00%	41.35%	4.88%
	Delly	67.00%	1.63%	33.00%	98.37%	61.98%	91.46%	38.02%	8.54%
	Pindel	82.64%	81.30%	17.36%	18.70%	88.51%	93.90%	11.49%	6.10%
	SoftSV	0.00%	0.00%	100.00%	100.00%	93.75%	18.29%	6.25%	81.71%
All Duplications	**MiStrVar**	30.85%	53.91%	69.15%	46.09%	N/A	N/A	N/A	N/A
	ITDetector	13.54%	40.87%	86.46%	59.13%	N/A	N/A	N/A	N/A
	Pindel	5.00%	15.65%	95.00%	84.35%	N/A	N/A	N/A	N/A
	SoftSV	16.24%	16.52%	83.76%	83.48%	N/A	N/A	N/A	N/A
Tandem Duplications	**MiStrVar**	100.00%	86.67%	0.00%	13.33%	N/A	N/A	N/A	N/A
	ITDetector	3.17%	80.00%	96.83%	20.00%	N/A	N/A	N/A	N/A
	Pindel	0.00%	0.00%	100.00%	100.00%	N/A	N/A	N/A	N/A
	SoftSV	6.67%	46.67%	93.33%	53.33%	N/A	N/A	N/A	N/A

Table 2: Summary of cutoffs for metrics used to define the true set for each tool which maximize the F-score. Those marked as "default" showed no improvement from using more stringent cutoffs and therefore the entire set of results was used for precision/recall calculations.

	Metric	**5-100bp Inversions**	**101-400bp Inversions**	**All Duplications**	**Tandem Duplications**
MiStrVar	Min. Identity	99%	99%	99%	100%
	Min. Read Support	default	default	30	default
Breakdancer	Score (0-99)	default	default	N/A	N/A
Delly	Min. Read Support	10	10	N/A	N/A
Pindel	Min. Read Support	6	10	3	default
SoftSV	Min. Read Support	default	default	30	default
ITDetector	Grade (A,B,C)	N/A	N/A	default	default

tandem duplications; Delly and Breakdancer, can predict long duplications quite well, however both perform relatively poorly on shorter events.

The comparative performance of MiStrVar against these methods are presented in Table 1. Note that in this analysis a call is considered to be a true positive if the actual breakpoint(s) was(were) within 5bps of the predicted breakpoint(s). The documentation for each tool was examined in order to determine whether any parameters could be changed to improve the results for short structural variant discovery. Other than parameters to disable prediction of event types not within the simulated data (i.e. translocations), we could not find any such parameters. Therefore, we have run all tools with default parameters and defined the true calls based on thresholds for the provided metrics (quality, identity or support, if they exists). We used the metric cutoffs which maximize the F-score, shown in Table 2. Only inversions of length ≤400bp were considered in the calculations. We also filtered known germline insertions from all duplication results.

For short inversions (5-100bp), MiStrVar outperformed the available alternatives by a sizeable margin: Delly and Breakdancer were able to identify only 2 inversions (of length ∼90bp), while SoftSV produced only false positives. The performance of Pindel was the closest to MiStrVar, still ∼10% lower in both precision and recall. This is a considerable improvement when you

Table 3: Running time comparison of all the SV detection tools on simulated data.

Tool	CPU Time
MiStrVar	38m19s
BreakDancer	77m23s
Delly	10m35s
Pindel	40m4s
SoftSV	2m54s
ITDetector	138m39s

consider that the false discovery rate and false negative rate is more than halved with MiStrVar. For longer inversions (>100bp), the performance of Breakdancer, Delly and Pindel were roughly the same, still weaker than MiStrVar in both precision and recall. SoftSV had a very poor recall performance, but was the most precise in this category.

For duplications, all tools we tested suffer from a high number of false positives. MiStrVar was able to identify all duplications at low precision, the recall was down to ∼54% when the precision was ∼31%. Even so, MiStrVar significantly outperformed all the other tools. To ensure that our performance was not due to non-tandem duplications only, we also computed the precision and recall values specific to tandem duplications. As shown in Table 1, with the exception of MiStrVar, the relative performance of the tools did not change dramatically, improving the recall and worsening the precision. In contrast, MiStrVar was able to find all the tandem duplications with high precision. Notably, Pindel was unable to find any of the duplications in this category. MiStrVar, on the other hand, improved significantly in recall with only a small drop in precision.

We also compared unfiltered microSV calls from all the five tools against that of MiStrVar. Among these tools, Pindel was the most successful for inversions, correctly returning 147 of the 205 inversion breakpoints; unfortunately it could not identify any of the duplication breakpoints correctly. SoftSV on the other was the best for duplication breakpoints, identifying 10 correct duplication breakpoint, while also returning 11 correct inversion breakpoints. In contrast, the unfiltered results returned by MiStrVar are much more precise, with 162 correct inversion breakpoints and and 98 correct duplications breakpoints.

Finally, the tools were compared in terms of running times. All tools were executed with the same parameters used for the simulation predictions on the same system (AMD FX-9590, 16GB RAM). If multiple stages were required, we report the total CPU time for all stages. BAM file preparation was not considered, but it should be noted that MiStrVar will accept fastq, BAM or SAM, and does not require the input to be sorted which may save users a great deal of time. The CPU time used by each tool is summarized in Table 3. SoftSV and Delly had the best running times (which is not surprising as they have the worst sensitivity), followed by MiStrVar and Pindel. BreakDancer and ITDetector had relatively poor running times compared to the other tools.

3.2 MicroSV Detection Results in HCC1143 Cell Line

3.2.1 Chromatogram interpretation for hetrozygous microSVs

Four of the inversions had amplicons with some nucleotides matching the reverse genomic strand and some matching the forward strand. This occurred in the amplicons from all four normal samples and two of the tumor samples. To resolve this discrepancy, the chromatogram corresponding to each amplicon was examined, first for the four normal samples, for which each of the inversion locations had either one or two peaks. In locations with two peaks, the bases always

Table 4: Summary of microinverions (Inv.) and microduplications (Dup.) predicted by MiStrVar in HCC1143 Cell line WGS datasets. **Coding Region Calls** are those predictions that overlap with known coding regions defined in the Ensembl gene annotation for human reference genome version GRCh37.70. Values in brackets indicate calls unique to the sample. These may be somatic (tumor sample), or germline calls that have been lost in tumor by a large deletion (normal sample).

Type	**# Reads (Million)**	**# Calls**			**# Coding Region Calls**		
		Inv.	**Dup.**	**Total**	**Inv.**	**Dup.**	**Total**
Tumor	11,975.30	4,028	37,023	41,051	63 (53)	124 (54)	187 (107)
Normal	13,131.73	4,650	43,383	48,033	63 (53)	143 (73)	206 (126)

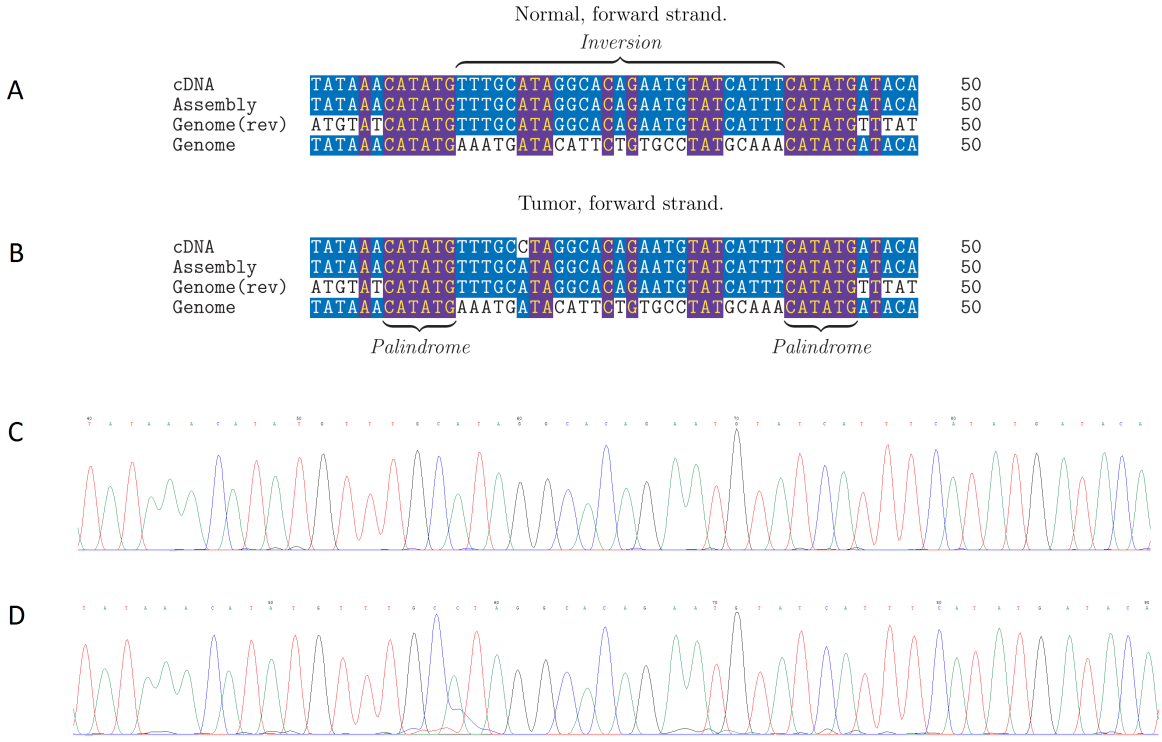

Figure 8: Sanger sequencing validation of the microinversion in the SLC3A1 gene we discovered in the HCC1143 cell line. (A/B) Sequence of the cDNA produced by PCR, the assembled contig, and the forward/reverse reference for that locus is given above. The cDNA is available for both the normal (A) and tumor (B) samples from the HCC1143 cell line. Blue color indicates matching sequences while purple color indicates palindromic sequences. (C/D) The corresponding chromatogram for the normal (C) and tumor (D) cDNA is also provided above. Red, green, blue and black peaks indicate T, A, C and G, respectively.

matched either the forward or reverse strand, exhibiting a classical case of heterozygous inversion that only occurs on one allele. For the final two inversion predictions, the amplicons for BOK and UBP1 corresponding to the tumor sample, only matched the forward genomic strand, which indicates no inversion at these locations. The amplicon corresponding to the normal sample of

UBP1 contained many N bases in the sequence. Not enough information could be drawn from the chromatogram to conclusively say whether the amplicon supports an inversion.

Three of the duplication chromatograms showed two peaks at the insertion site and immediately downstream. One of the two peaks support the reference and the other the inserted sequence and the shifted reference, indicating that these calls are heterozygous. This was observed in both normal and tumor samples for GPRIN2 and only in normal for PALM2-AKAP2 and PRSS48. The final amplicon for ADAMTS7 showed only reference sequence at the insertion site, indicating that there is no duplication.

3.2.2 RNA-Seq Support for microSVs in HCC1143 Cell Line

In addition to matching exonic microSV calls with unique proteomic signatures, we looked for RNA-seq level support for microinversions and microduplications detected by MiStrVar on the HCC1143 cell line. RNA-Seq data for HCC1143 cell line was composed of 81.73M paired-end reads of length 2x101bp each. Among these, ProTIE uses 8.75M (∼10%) reads that cannot be mapped concordantly to known transcripts or potential spliceforms by splice-aware mapper STAR, to validate transcribed microSVs identified at the genomic level. In short, there are 24 such microSVs with supporting RNA-Seq reads, and all but one of them are microduplications. This is in agreement with the WGS data whose analysis revealed no no high confidence, exonic microinversions. In fact, the single microinversion with RNA-Seq support also has low WGS read support and contains an error.

3.3 Further Support and Evaluation of the Detected microSVs in the HCC1143 Cell Line

Note that with the exception of one microinversion and one microduplication, all microSVs discovered in the HCC1143 cell line have corresponding entries in The Database of Short Genetic Variation (dbSNP) [19]. The microinversions are annotated as "multiple nucleotide polymorphisms" (MNP) microduplications are annotated as "insertion/deletions". This is an additional indication that MiStrVar predicts real events. Furthermore, this observation points to the much needed differentiation of germline MNPs and insertions from inversions and duplications, currently missing in dbSNP.

It is interesting to add here that, two of the likely somatic calls made by MiStrVar also have entries in dbSNP. The first one, the microduplication in FAM20C mentioned above, is missed by Sanger Sequencing on the normal sample; it can still be a germline event since it is supported by 22 WGS reads. The second microduplication in gene KIAA1009 is even more likely to be somatic since it is not supported by WGS or Sanger Sequencing in the normal sample. Two additional calls, a microinversion in BOK and a microduplication in ADAMTS7, could not be validated by Sanger Sequencing in either sample, but have dbSNP entries. Both of these calls are observed across several TCGA samples in addition to the HCC1143 cell line so it is likely that Sanger Sequencing failed to detect them.

Another interesting observation are the calls in the genes PALM2-AKAP2 and PRSS48. These calls were only validated in the normal sample, and in one allele. The event PALM2-AKAP2 is likely to be present and yet missed by Sanger Sequencing in the tumor sample since it was observed with high WGS read support. In contrast, the call in PRSS48 has no supporting reads in the tumor sample in either WGS or RNA-Seq. Upon further investigation of this region, we noticed that the read coverage is approximately half in the tumor sample in comparison to the normal sample and all the reads are wildtype. This implies that this is a germline event where the allele containing the call has be deleted in the tumor. This observation highlights the

possibility that due to their small size microSVs could be deleted or amplified through larger events.

4 Overview of CPTAC datasets

Table 5 provides some details on samples that were used in our analysis. Since we applied MiStrVar and ProTIE to the complete set of 22 TCGA breast cancer patients for which matching WGS, RNA-Seq and CPTAC Mass Spectrometry data were all available, we also list their information in Table 6.

The mass spectrometry datasets released by CPTAC were selected from all four major breast cancer intrinsic subtypes (Luminal A, Luminal B, Basal-like/triple-negative, HER2-enriched). Each iTRAQ experiment included three TCGA samples and one common internal reference control sample. A single mixture consists of 25 proteome and 13 phosphor-proteome data files, in total 500 GB data. Our data analysis indicates that a two-dimensional reversed-phase liquid chromatographytandem mass spectrometric (2D-LC/MS/MS) sample comprises of about 0.87 million MS/MS spectra (per mixture). When we search them against Ensembl Human protein database, about 0.38 million MS/MS spectra in a mixture are matched to at least one peptide under 1% false discovery rate. These spectra lead to 59,387 proteins (42,840 known, 6,250 novel, 10,026 putative) with some peptides being covered by at least one spectra. The remaining 0.49 million spectra ($\approx 56\%$ of the whole set) do not match to any protein in the Ensembl database.

ProTIE obtains the intersection between these (0.49 million) unidentified spectra and the aforementioned set of fusions with missed cleaved polypeptides, to obtain 3,150,502 potential fusion peptides from 105 breast cancer patients [6] [7] ProTIE uses a similar workflow to identify potential microSV peptides; for this case 635,125 potential microSV peptides were obtained from 22 patients.

Table 5: Available omics data for TCGA/CPTAC breast cancer samples.

WGS			**RNA-Seq**		**Mass Spec.**	
Solid Tumor	**Blood Normal**	**Solid Normal**	**Solid Tumor**	**Solid Normal**	**Mixture**	**Number of Patients**
✓	✓	✓	✓	✓	✓	**2**
✓	✓		✓	✓	✓	**1**
✓		✓	✓	✓	✓	**3**
✓	✓		✓		✓	**16**
			✓	✓	✓	**10**
			✓		✓	**73**
					Total:	**105**

5 High-Confidence deFuse Calls in CPTAC Datasets

The first part in Table

[6] Each breakpoint is associated with six reading frames and thus can result in (one of) six distinct proteins, and each such potential protein can lead to multiple potential peptides according to the number of K/R in the sequence. (see Figure

[7] Note that a reversed database was also appended here to control the false discovery rate.

Table 6: General information on all 22 TCGA breast cancer patients with both tumor/normal WGS and tumor RNA-Seq data. **(A) Clinical Data.** The PAM50 mRMA cancer subtypes and AJCC stage for each patient. **(B) Data Source.** *Tissue Source* indicates the medical facility the sample and the relevant clinical data originates from; *Sequencing Center* indicates the location of actual sequencing: **WUSM** indicates Washington University School of Medicine, and **HMS** indicates Harvard Medical School. **(C) Number of Reads.** The BAM files corresponding to the majority of the samples contain paired-end reads of length 2x100bp (data from WUSM) or 2x51bp (data from HMS). There are only two exceptions: the solid tumor of patient **A09I** contains additional 206 million paired-end reads with respective lengths of 100bp and 44bp; the solid tumor of patient A0CM contains additional 579 million single end reads. These two inconsistent data sets are not used in our analysis. Note that all RNA-Seq datasets are from UNC (University of North Carolina Medical School), and on average include 76M paired-end reads of length 2x50bp.

Patient	Cancer Subtypes	AJCC Stage	Tissue Source	Sequencing Center	Number of WGS Paired-End Reads (Millions)		
					Solid Tumor	**Blood Normal**	**Solid Normal**
A09I	Basal-like	IIA	Indivumed	WUSM	687.13	584.10	
A0AV	Basal-like	IIIC	U of Pittsburgh	WUSM	954.38	558.24	
A0CE	Basal-like	IIA	Christiana Healthcare	WUSM	628.67	552.64	691.68
A0CM	Basal-like	IIA	Walter Reed	WUSM	784.96	540.58	
A0D0	Basal-like	IIA	Walter Reed	WUSM	788.15	516.83	
A0D1	Basal-like	IIB	Walter Reed	WUSM	1015.35	573.74	
A0D2	Basal-like	IIIA	Walter Reed	WUSM	689.42	646.66	
A0DG	Basal-like	I	U of Pittsburgh	WUSM	893.64		522.71
A0E0	HER2-enriched	IB	U of Pittsburgh	WUSM	686.41	521.02	793.65
A0EY	HER2-enriched	IIA	Walter Reed	WUSM	907.43	579.46	
A0HK	HER2-enriched	II	U of Pittsburgh	HMS	193.03	180.00	
A0J6	HER2-enriched	IIA	MSKCC	WUSM	592.48	661.74	
A0JJ	HER2-enriched	IIIA	MSKCC	HMS	214.75	208.31	
A0JL	HER2-enriched	IIIA	MSKCC	HMS	220.37	217.78	
A0JM	Luminal A	IIB	MSKCC	WUSM	1126.03	671.57	
A0TX	Luminal A	IIB	Mayo	WUSM	1018.72	642.28	
A0YG	Luminal A	IIA	Walter Reed	WUSM	872.51	514.68	
A12L	Luminal B	IIIA	ILSBio	WUSM	1031.04	650.09	
A12Q	Luminal B	IIIC	ILSBio	WUSM	1011.50	640.57	
A130	Luminal B	IIB	ILSBio	WUSM	813.37	654.17	
A18R	Luminal B	IIB	U of Pittsburgh	WUSM	1002.25		594.58
A18U	Luminal B	IIA	U of Pittsburgh	WUSM	906.82		605.45

5.1 Proteomics Support of High-Confidence deFuse Calls

Among the remaining fusions, two stand out with respect to peptide-spectrum matching quality, respectively observed in patients A08G and A15A. The corresponding PSMs generated by pFind Studio [20, 21] can be found in figs. 9 to 11.

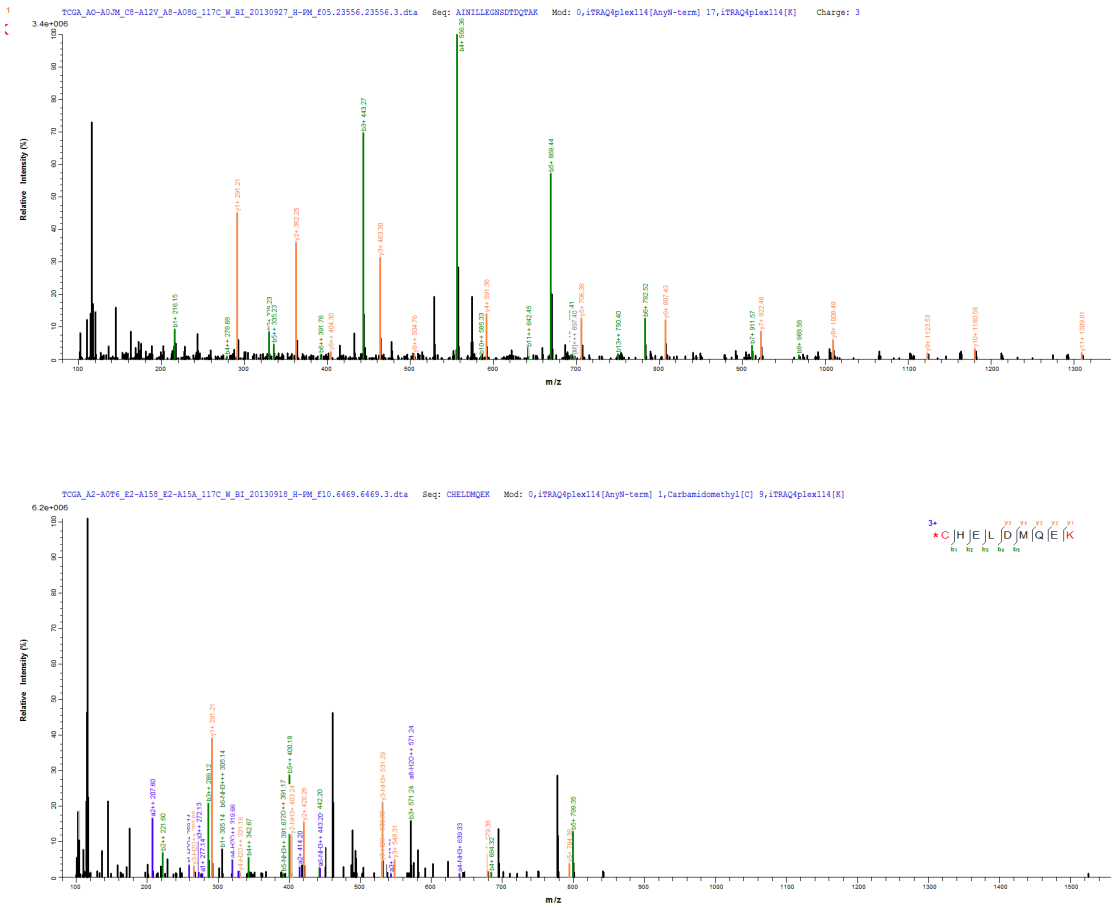

Figure 10: A PSM supporting a fusion between genes HOOK3 and CTA-392C11.1 in patient A15A (Luminal B, Stage IIIC). The peptide crosses the fusion breakpoint predicted from RNA-Seq data at amino acid M.

6 Summary of microSVs in CPTAC datasets

6.1 MicroSVs Detection Results is WGS datasets

We applied MiStrVar to WGS datasets of 22 CPTAC patients. The number of detected microSVs varies between samples, ranging from one to several thousand. See Table 8 for details. In addition, Table 9 and Table 10 provide information of high confident microinversions and microduplications among 22 patients.

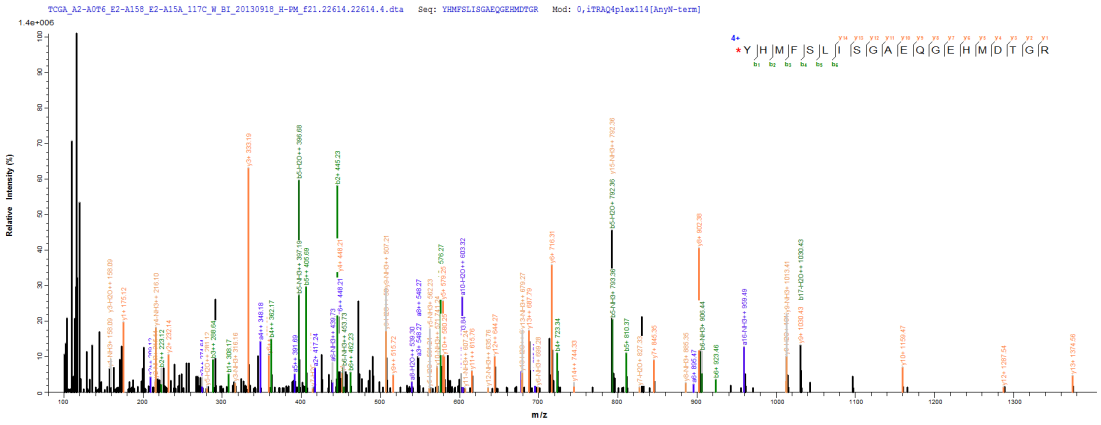

Figure 11: Another PSM supporting the fusion in figure 10 between genes HOOK3 and CTA-392C11.1 in the patient A15A (Luminal B, Stage IIIC). The peptide is located downstream of the breakpoint.

Patient	Cinical Info	Gene 1	Gene 2	deFuse Score	Breakpoint Location	Peptide Sequence	# of Spectra	BP	Str FDR
Additional Fusions with Multiple Supporting Spectra									
A06Z	LB, IIB	RAB15	TMEM98	0.01	coding, utr5p	QIWDTAGQENR	2	✓	
A0C1	LA, IIIA	RPL14	FAM155A	0.02	coding, coding	ASAAAAAAAAK	2	✓	✓
A0D2	BL, IIB	ACTG1	ACTB	0.52	coding, coding	HHGIVTNWDDMEK	4	✓	
A0E0	BL, IIIC	PEA15	CPEB2	0.05	coding, coding	YPGTLLQDLTNNITLEDLEQLK	2	✓	✓
A0EX	LA, IIB	RAB6B	CFL1	0.02	utr3p, coding	EAGVAVSDGVIK	3	✓	
A0TR	LA, II	ZNF587	TMEM163	0.12	utr3p, intron	QSETLSQNKK	2	✓	
A12D	H2, IIA	**RPL19**	**CALR**	0.04	coding, coding	PAGQGVFPASSPGMDGEWEPPVIQNPEYK	5	✓	✓
A12D	H2, IIA	SCGB2A2	EEF1A1P5	0.05	coding, pseudogene	ATAFIDQMASSGGLARIYVSNDDNATTNAIDELK	2	✓	
A12D	H2, IIA	EIF4A1	ABL2	0.16	utr3p, intron	SLNKCHFLR	3	✓	
A12U	LB, IB	NME1	RP11-111A21.1	0.01	coding, downstream	SVMLGETNPADSKPGTIR	2	✓	✓
A12W	LB, IIIB	CTNNA3	CEP120	0.007	intron, intron	LALDIEIATYKT	2	✓	
A13F	LB, IIIA	RPL14	S100A16	0.17	coding, utr3p	SAAAAAAAAAK	2	✓	✓
A142	BL, IIB	HSP90AB1	AC096579.7	0.01	coding, ncRNA	FEINPDHPIVETLR	4	✓	✓
A150	BL, IIA	HSPA8	RP11-537H15.3	0.03	coding, intron	(*)HVAMNPTNTVFDAK	2	✓	✓
A159	BL, IIA	DLG4	**VIM**	0.36	intron, coding	SYVTTSTRR	2	✓	
A15A	LB, IIIC	WASH4P	ABC7-42389800N19.1	0.07	coding, pseudogene	PKSGSGGEGVMEPPR	2	✓	
A18Q	BL, IIB	MGP	EEF1A1P5	0.01	coding, pseudogene	FFFFPQSHLVTFAPVNVTTEVK	5	✓	✓
A18U	LB, IIIA	ZNF354A	RP11-383H13.1	0.47	coding, intron	DGSGVSSLGVTPESR	2	✓	✓
A1AQ	BL, II	**CDKN2A**	LINC00486	0.01	coding, intron	GGGGGGGGCCPR	2	✓	
Additional Cancer-related Genes in Fusions									
A0BZ	LB, IIIA	**MDM2**	ZC4H2	0.69	utr3p, downstream	ISFFLEVLQALFGVDNTSATTK	1	✓	
A0C1	LA, IIIA	USP42	**CD44**	0.19	intron, utr3p	YEKENWSGFFFFFLK	1	✓	
A0EQ	H2, IIA	**ANKRD30A**	BLOC1S6/NEDD4	0.85 0.76	coding, utr3p coding, intron	ISGKLEELEK	1	✓	✓
A09I	LB, IIA	YARS/ZNFX1	**GRB7**	0.08 0.02	intron/utr3p upstream/intron	GQEFKTSLTNMAK	1	✓	
A09I	LB, IIA	**ERBB2**	NME2P1	0.01	coding, pseudogene	IQHYIDLK	1	✓	

6.2 RNA-Seq support for microSVs in CPTAC Breast Cancer Patients

Among 22 breast cancer patients with matching WGS and RNA-Seq data, MiStrVar detected 69,876 exonic microinversions and microduplications. Out of these microSVs, 905 microinversions and 1310 microduplications (among which 33 are tandem), i.e. 2,215 calls overall, are supported by at least one RNA-Seq read (as determined by the splice-aware mapper STAR) which can not be concordantly mapped to any other known transcript.

Table 8: Number of exonic microinversion and microduplication calls made by MiStrVar in each of the 22 TCGA breast cancer patients with both tumor/normal WGS and tumor RNA-Seq data. Even though few of these patients, especially A0CM, have microSV profiles with a high number of inversions, the sequencing quality is consistent with others. Additionally their copy number profiles are also similar to others.

Patients	Cancer Subypes	AJCC Stage	Microinversions			Microduplications		
			Solid Tumor	Blood Normal	Solid Normal	Solid Tumor	Blood Normal	Solid Normal
A09I	Basal-like	IIA	62	1071		89	169	
A0AV	Basal-like	IIIC	79	56		151	122	
A0CE	Basal-like	IIA	559	151	58	156	54	79
A0CM	Basal-like	IIA	36402	274		2867	74	
A0D0	Basal-like	IIA	429	1335		135	169	
A0D1	Basal-like	IIB	765	1373		150	163	
A0D2	Basal-like	IIIA	2031	552		484	109	
A0DG	Basal-like	I	64		35	128		96
A0E0	HER2-enriched	IB	281	67	991	100	73	134
A0EY	HER2-enriched	IIA	2075	1762		194	168	
A0HK	HER2-enriched	II	5	0		10	6	
A0J6	HER2-enriched	IIA	7305	53		1220	58	
A0JJ	HER2-enriched	IIIA	2	3		9	10	
A0JL	HER2-enriched	IIIA	2	0		10	13	
A0JM	Luminal A	IIB	502	80		128	82	
A0TX	Luminal A	IIB	134	293		90	84	
A0YG	Luminal A	IIA	105	696		70	73	
A12L	Luminal B	IIIA	397	109		151	88	
A12Q	Luminal B	IIIC	107	57		112	86	
A130	Luminal B	IIB	198	57		79	60	
A18R	Luminal B	IIB	428		47	99		75
A18U	Luminal B	IIA	142		52	81		72

Table 9: The top microinversions found in introns (WGS read support > 40, sequence identity 100%) and UTRs (WGS read support > 10, sequence identity 100%). Boldface microinversions are also observed in the HCC1143 cell line, and all except BOK are validated. The "Pali." column provides the length of flanking palindromic sequences. The "Support" column indicates the minimum per base coverage of the inverted region, in the sample with the highest support for the call. The dbSNP ID refers to the multiple nucleotide polymorphism (MNP) corresponding to the microinversion. If more than one ID is shown, these are SNPs with equal allele frequency that can be explained by the inversion.

						# of Samples		WGS Support		RNA-Seq	
Chr.	**Location**	**Len.**	**Pali.**	**Gene**	**Region**	**Tumor**	**Normal**	**Tumor**	**Normal**	Support	**dbSNP ID**
2	44545739	27	6	**SLC3A1**	3'UTR	18	22	88	36	✓	rs71416108
7	24745614	21	5	DFNA5	3'UTR	2	2	16	9		rs386711358
1	226259222	7	3	H3F3A	3'UTR	17	20	33	24		
6	170859029	26	0	PSMB1	3'UTR	1	2	17	33	✓	
17	416906	87	5	VPS53	3'UTR	1	0	21	0		
13	49000665	1267	0	LPAR6	5'UTR	1	0	14	0		
3	170821851	26	3	**TNIK**	Intron	20	23	78	47	N/A	rs781523247
7	117357036	29	3	**CTTNBP2**	Intron	22	24	48	39	N/A	rs386717124
19	56389843	32	2	**NLRP4**	Intron	20	23	61	43	N/A	rs386811126
22	31291523	23	2	**OSBP2**	Intron	15	18	67	40	N/A	rs67147751
9	28014540	29	3	**LINGO2**	Intron	10	13	49	40	N/A	rs386733960
2	242500549	12	4	**BOK**	Intron	15	14	55	40	N/A	rs386657165
3	85078096	6	31	CADM2	Intron	17	19	57	44	N/A	rs71616888
1	223947597	8	11	CAPN2	Intron	5	7	73	25	N/A	rs386639771
8	3904131	21	5	CSMD1	Intron	4	4	46	46	N/A	rs768996207
17	72953704	5	3	HID1	Intron	7	7	47	12	N/A	rs374377884
12	75573538	5	19	KCNC2	Intron	10	10	61	45	N/A	rs201249335,rs201655437,rs200375416, rs201997536,rs201455075
6	151077062	23	6	PLEKHG1	Intron	6	5	43	25	N/A	rs12662499,rs71570234,rs71570235,rs71570235 rs71570236,rs56028508,rs56028508
13	49002453	868	0	RB1	Intron	1	0	45	0	N/A	
9	92084244	35	3	SEMA4D	Intron	9	11	52	36	N/A	rs71497306
5	179256683	22	3	SQSTM1	Intron	18	19	61	38	N/A	rs71577407
2	64117422	25	2	UGP2	Intron	4	5	45	37	N/A	rs543356344,rs543356344,rs529034848, rs540886904,rs559363558,rs139485198
8	100686027	5	4	VPS13B	Intron	5	7	84	26	N/A	rs386728165
5	15840451	30	2	FBXL7	Intron	14	14	63	34	N/A	rs386685795
6	108787955	12	23	LACE1	Intron	13	16	42	25	N/A	rs71553768
3	197731404	85	0	LMLN	Intron	1	1	29	58	N/A	
2	214485778	5	13	SPAG16	Intron	5	6	59	44	N/A	rs74181305

Table 10: The top exonic microduplications (WGS read support > 40, sequence identity 100%) and somatic microduplications (WGS read support > 10, sequence identity 100%). Boldface microduplications are also observed in the HCC1143 cell line. The "Type" column indicates if the inversion is tandem or interspersed. The "Support" column indicates the minimum per base coverage of the duplicated region, in the sample with the highest support for the call. The dbSNP ID refers to the indel corresponding to the microduplication.

						# of Samples		WGS Support		RNA-Seq Support		
Chr.	**Location**	**Len.**	**Gene**	**Region**	**Type**	**Tumor**	**Normal**	**Tumor**	**Normal**	**Tumor**	**Normal**	**dbSNP ID**
1	186365852	8	AL596220.1	Exon	Tandem	**9**	10	42	20	0	0	rs145764138
1	203186950	24	CHIT1	Exon	Tandem	**2**	2	54	14	0	N/A	rs386369359
19	51857874	6	ETFB	Exon	Tandem	**2**	2	41	20	0	N/A	rs61361626
12	121434630	8	HNF1A	Exon	Tandem	**18**	23	50	33	0	0	rs58371019
10	91497902	6	KIF20B	Exon	Inter.	**12**	13	43	25	24	7	rs144593231
16	2185524	7	PKD1	Exon	Tandem	**5**	3	75	31	0	0	rs3072277
4	152201018	5	**PRSS48**	Exon	Tandem	**7**	8	47	20	0	0	rs71901196
22	20458100	6	RIMBP3	Exon	Tandem	2	3	45	29	0	0	rs374606390
1	180199692	24	LHX4	Exon	Tandem	1	0	12	N/A	0	N/A	-
18	34366699	5	TPGS2	3'UTR	Tandem	1	0	11	N/A	2	N/A	-
7	102222821	12	RASA4	3'UTR	Tandem	1	0	10	N/A	3	N/A	-
12	76443235	6	NAP1L1	3'UTR	Tandem	1	0	13	N/A	0	N/A	-

7 Mechanistic and Functional Interpretation of microSV and Fusion Peptides Detected in TCGA/CPTAC BRCA Dataset

7.1 Fusions Peptides

Many of the fused genes with detected novel peptides (each typically observed in a single patient) are associated with breast cancer. A selection of these fusions are listed in Table

The remaining fusions associated with highlighted genes in Table

In addition to fused tumor suppressors, we also detected peptide evidence for fused oncogenes. The discovered fused oncogenes are: ANKRD30A, also known as NY-BR-1, a breast differentiation antigen observed in many breast cancer cells [22]; GRB7, a breast cancer driver gene which participates in Development ERBB-family signaling pathway [23, 24]; ERBB2, a well known breast cancer oncogene and biomarker [25] as well as the coexpressed gene Ribosomal protein L19 (RPL19); CALR, a gene highly expressed in approximately 5% of breast cancer cells and associated with metastasis [26]; and finally VIM, a protein involved in the epithelial to mesenchymal transition which drives metastasis [27]. The fusions involving ANKRD30A, RPL19 and CALR meet our stringent FDR criteria, while the others do not. In a number of cases, we can not pinpoint its fusion partners based on RNA-Seq data alone. The proteogenomics results help to increase our confidence of these fusions, and reduce the number of fusion partner candidates in the corresponding patients. The ERBB2 fusion is particularly interesting since ERBB2 is amplified in 15% of breast cancers and targeted with a variety of FDA approved drugs, making it a possible target for clinical analysis.

In the final list of 295 candidate fusions, 107 of the involved genes are also reported to be involved in a fusion according to TCGA Fusion gene Data Portal [8]. 58 of these genes have records in breast cancer (BRCA), and among them 19 genes are reported in the breast cancer database alone.

Among the ten cancer-related fusion genes in Table

7.2 Genomic MicroSVs Peptides

Analysis of cell line and TCGA data allowed us to investigate the prevalence of microSVs in cancer. In particular, we present here the very first analysis of transcribed and translated microinversions in any cancer dataset. The only study in the literature on microinversions focus exclusively on a small number of "easier to detect"' intronic and intergenic events, for phylogenetic purposes [28, 29, 30]. Our validation results on the HCC1143 breast cancer cell line confirmed the presence of several high-confidence, germline calls in intronic regions, in agreement with this study. Interestingly, many of these inversions are flanked with short palindromic sequences. This appears to support the microinversion mechanism proposed earlier by Kelchner et al. [31] where these palindromes form stems through base pairing. We note here though that the palindromic sequences we observe typically include two to six nucleotides, shorter than those with eleven or more nucleotides observed by Kelchner et al.

The importance of small duplications has already been well established in AML [32], however little work has been done on other types of cancer. Our Sanger sequencing based validation results in the HCC1143 cell line confirms the presence and expression of microduplications in breast cancer. One of the microduplications we discovered in gene FAM20C has been shown [33]

[8]Note that results in this database are based on 10431 calls from 2961 TCGA patients, which contains much broader scope than 105 breast cancer patients selected by CPTAC.

to have a dramatic effect on cell adhesion, migration, and invasion of breast cancer cells. This could be of interest because the HCC1143 cell line that includes the aberrant FAM20C gene was derived from a non-metastatic tumor.

We observed that the number of microinversion calls varies between samples, ranging from two calls to several thousand. Much of this variation can be explained by the number of reads in the sample (e.g. A0HK, A0JJ and A0JL have the smallest WGS datasets); see Table 6). Other cases are explained by the variation in read quality (e.g. A0J6 has the highest number of low quality reads). However one patient, A0CM-01A, is a clear outlier and cannot be explained by either of these factors, e.g., the proportion of high quality calls for this patient is similar to that of other samples so the calls here are not inflated by false positives. It remains unclear what differentiates this patient from the others from a clinical point of view [34]. In addition, the number of SNVs, indels and fusions observed in this sample are in the normal range (Supplementary Table 7) even though the number of duplications are on the higher end. This may indicate that microinversions may have a dominant role in cancer progression under certain circumstances, even though further research is needed to identify the specific microinversion mechanism underlying these events.

The number of exonic microduplications in the TCGA patient samples were similar to that we observed for microinversions. The variance in the number of calls is also similar to that of microinversions. The patient A0CM again has the highest number of calls, but not drastically so.

As per our results from the HCC1143 cell line, most of the germline calls MiStrVar returned for the TCGA samples have corresponding dbSNP entries. This is particularly evident among high confidence intron and UTR microinversions (Table 9). Out of these calls, three microinversions found in the genes KCNC2, PLEKHG1 and UGP2 are especially interesting. These calls have multiple SNPs between the two associated breakpoints. Most (sometimes all) of the remaining inversion sequence are palindromic, i.e. the reverse complement is identical to the reference. The minor allele frequency of these SNPs is identical, implying that they are a consequence of a single event. It is likely that these SNPs are miscategorized in dbSNP and are actually part of germline microinversions. Several other microinversions that actually have a single MNP entry also have several redundant SNPs found in the same location with the same allelic frequency. This appears to be a common occurrence and may warrant an additional variant class in dbSNP as well as the removal of potentially incorrect and/or redundant entries. Many of the microinversion calls validated in the HCC1143 cell line appear in more than half of the TCGA samples (Table 9). The single validated UTR call in gene SLC3A1 also has RNA-Seq support in five samples. In addition, an entirely novel microinversion in the 3'UTR of PSMB1, with RNA-Seq support, was discovered. Other novel microinversions include those found in genes DFNA5 and RB1, both known tumor suppressors in breast cancer [35, 36]. Again, we often see short palindromic sequences in the flanking regions of the microinversions. However the high confidence call in PSMB1 demonstrates that some microinversions lack any flanking palindromic sequences.

The vast majority of microduplications appear to be germline events. In fact, all of the microduplications under our strictest filtering at the genomic level are germline events (Table 10). Unfortunately the majority of these do not have RNA-Seq or proteomic support. Interestingly, the only non-tandem microduplication in this subset does have RNA-Seq support in multiple samples. It occurs in KIF20B, a known oncogene for liver, bladder and pancreatic cancer, [37, 38, 39]. Note that this call would likely be missed by available duplication callers, or would be miscategorized as an indel. Although somatic calls are in the minority, we observed some high confidence calls in exons and UTR (Table 10). One of these calls is in NAP1L1, a tumor suppressor gene in neuroendocrine tumors and small intestine, ovarian and liver cancer [40]. Additionally, two UTR microduplications in TPGS2 and RASA4 genes have RNA-seq support. Among them, RASA4 methylation has been associated with poor prognosis in juvenile myelomonocytic leukemia [41].

From our list of high confidence microSV calls (Table

Deletions, translocations and allele loss at the genomic loci containing RPL14 have been observed in variety of cancers [42], including breast cancer [43]. This may be the case within patients AOCE, A18R (deletion) and A0JM (LOH). The unusually long case in patient A18U may lead to protein instability, causing the same phenotype as a deletion. Polyalanine tract lengths have been shown to be associated with cancer risk in other genes, such as androgen receptor in prostate cancer [44].

Another interesting example, RBBP8 is a tumor suppressor specifically related to breast cancer. We have observed through inspecting geneMania [45] that RBBP8 is associated with the recombinational repair pathway ($p < 1.27 \times 10^{-9}$) (Supplementary Figure 12). RBBP8 is also known to modulate the important tumor suppressor BRCA1 [46] and act as a tumor suppressor itself through binding with the MRE11-RAD50-NBS1 (MRN) complex [47] or replication protein A (RPA) [48].[9]

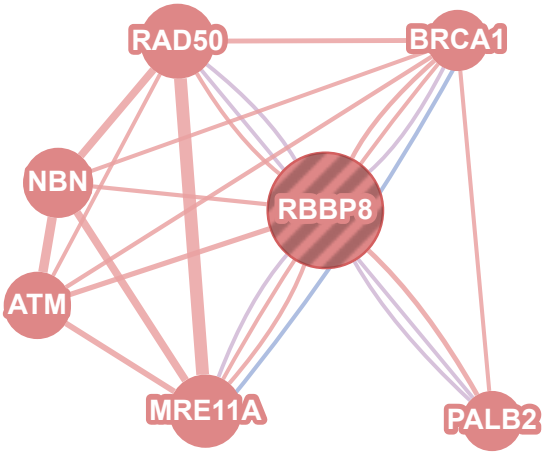

Figure 12: Functional analysis graph from GeneMania. Red lines indicate direct physical interaction, purple lines indicate co-expression and blue lines co-localisation. The thickness of the line represents the combined weights of the interaction across all analysed networks of that type. The diameter of the circles is inversely proportional to the rank of the gene in a list sorted by functional relatedness to the striped gene. This graph contains all genes interacting with RBBP8 in the recombinational repair pathway ($p < 1.27 \times 10^{-9}$). RBBP8 is closely associated with BRCA1, an important tumor suppressor gene in breast cancer.

[9]Binding of MRN and RPA occur through a domain at the N-terminus of the RBBP8 protein, which overlaps with the predicted microinversion. We hypothesize that the microinversion in this gene leads to the production of an aberrant peptide which is unable to bind to MRN or RPA, disrupting double stranded break repair and contributing to the cancer.

Appendix - Chromatograms

The following figures are excerpts of the chromatograms produced from Sanger sequencing validation of our top microSV calls in the HCC1143 cell line WGS data. The subsequences captured in these images include the SV (for validated cases) or the wildtype sequence (for invalidated cases) and some flanking sequence. Each call was validated using forward and reverse primers for both the normal and tumor sample. If a specific case is omitted below it is because neither the variant or the wildtype could be identified in the chromatogram (inconclusive case).

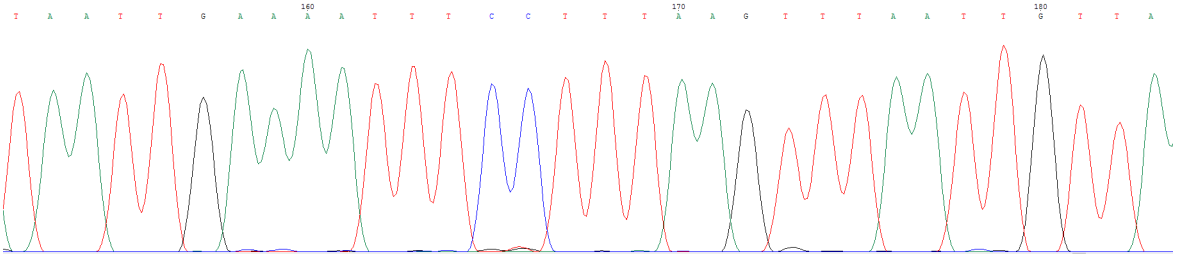

Figure 13: Chromatogram for cDNA produced when using forward primers capturing the microinversion in TNIK in the normal sample.

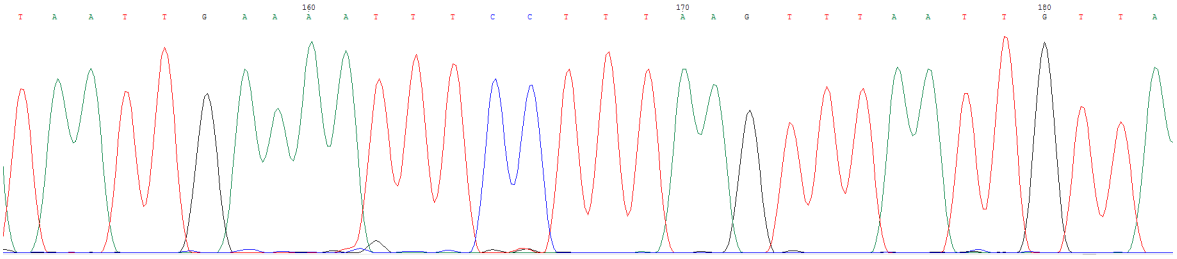

Figure 14: Chromatogram for cDNA produced when using forward primers capturing the microinversion in TNIK in the tumor sample.

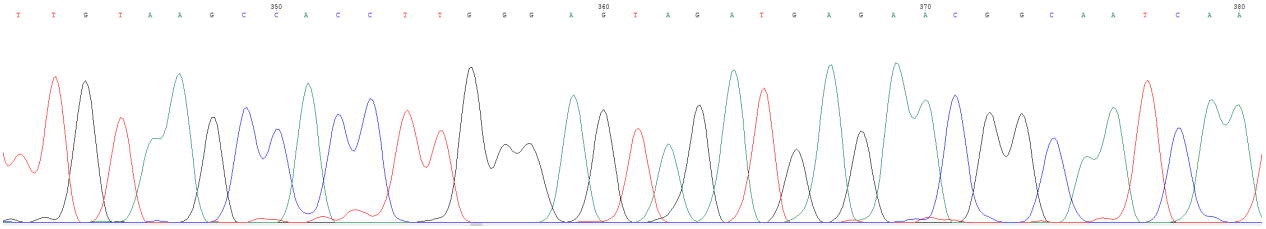

Figure 15: Chromatogram for cDNA produced when using forward primers capturing the microinversion in CTTNBP2 in the normal sample.

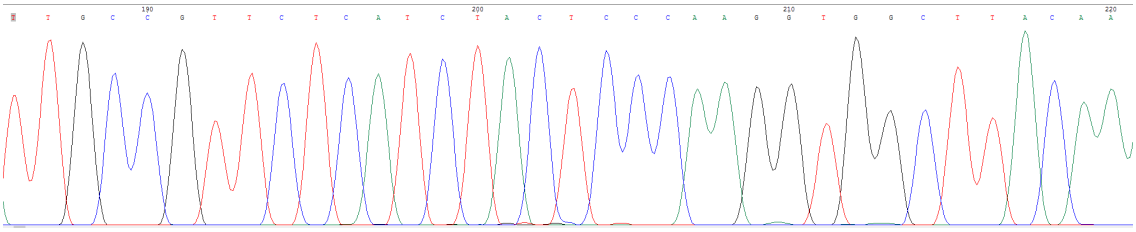

Figure 16: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in CTTNBP2 in the normal sample.

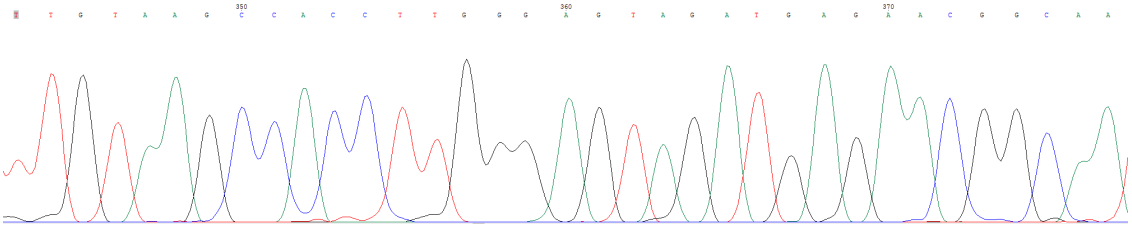

Figure 17: Chromatogram for cDNA produced when using forward primers capturing the microinversion in CTTNBP2 in the tumor sample.

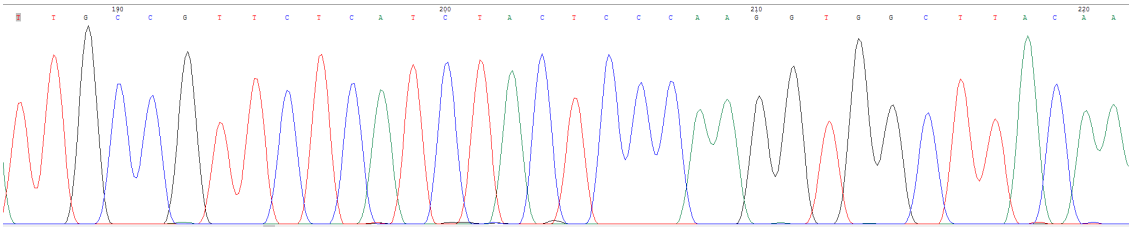

Figure 18: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in CTTNBP2 in the tumor sample.

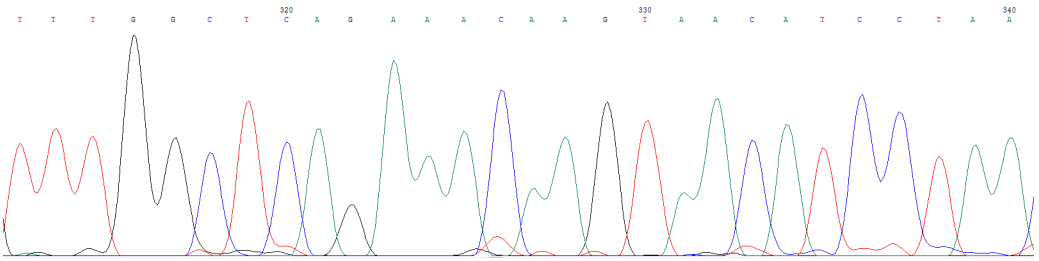

Figure 19: Chromatogram for cDNA produced when using forward primers capturing the microinversion in PFKP in the normal sample.

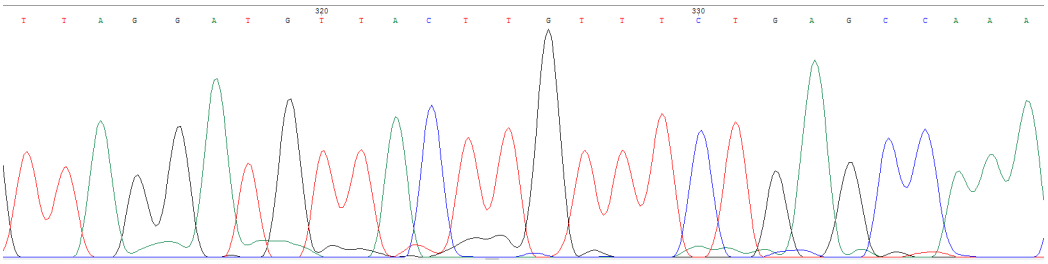

Figure 20: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in PFKP in the normal sample.

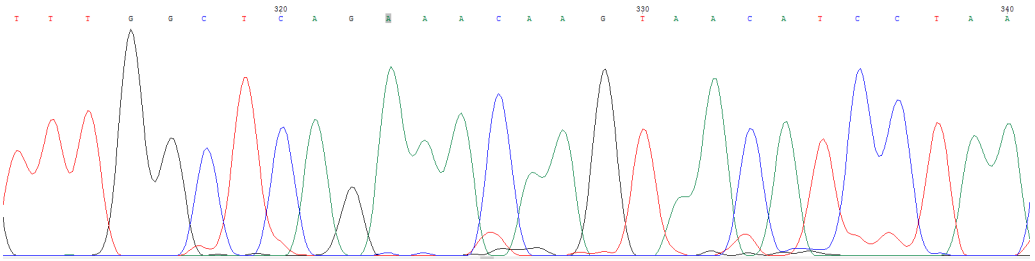

Figure 21: Chromatogram for cDNA produced when using forward primers capturing the microinversion in PFKP in the tumor sample.

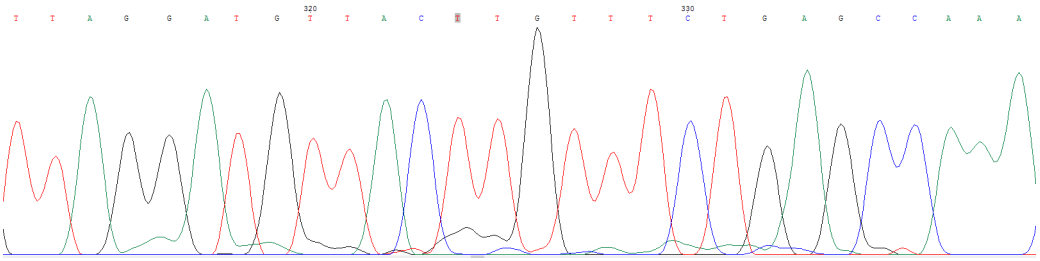

Figure 22: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in PFKP in the tumor sample.

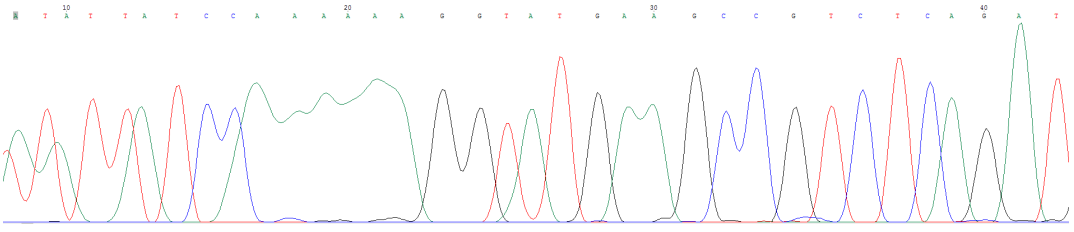

Figure 23: Chromatogram for cDNA produced when using forward primers capturing the microinversion in NLRP4 in the normal sample.

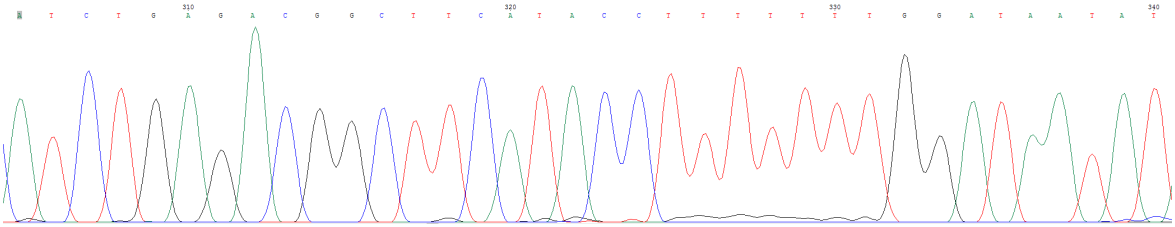

Figure 24: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in NLRP4 in the normal sample.

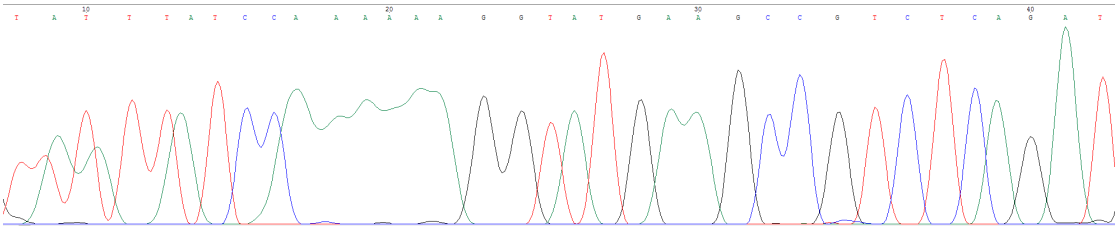

Figure 25: Chromatogram for cDNA produced when using forward primers capturing the microinversion in NLRP4 in the tumor sample.

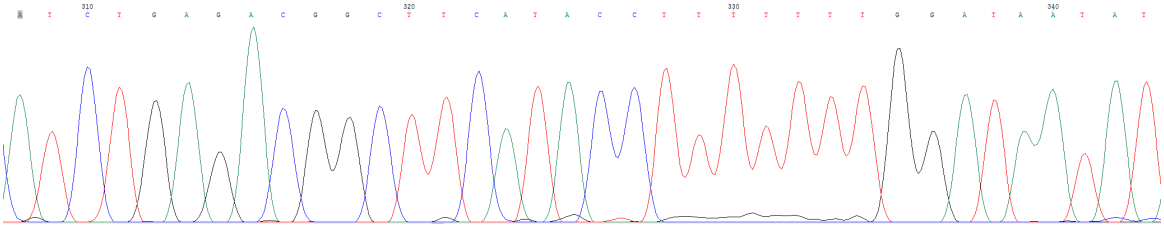

Figure 26: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in NLRP4 in the tumor sample.

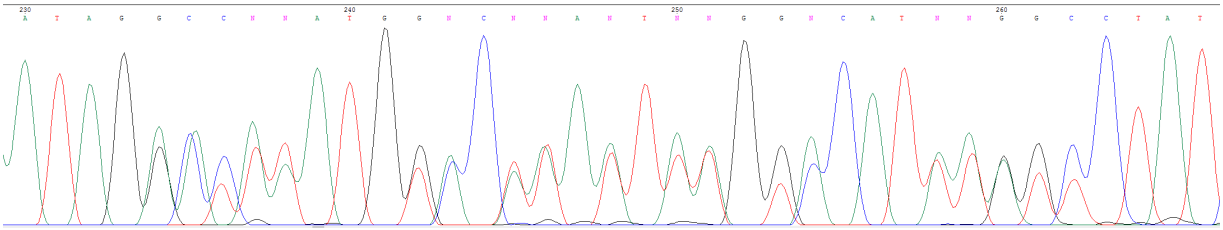

Figure 27: Chromatogram for cDNA produced when using forward primers capturing the microinversion in ZNF57 in the normal sample.

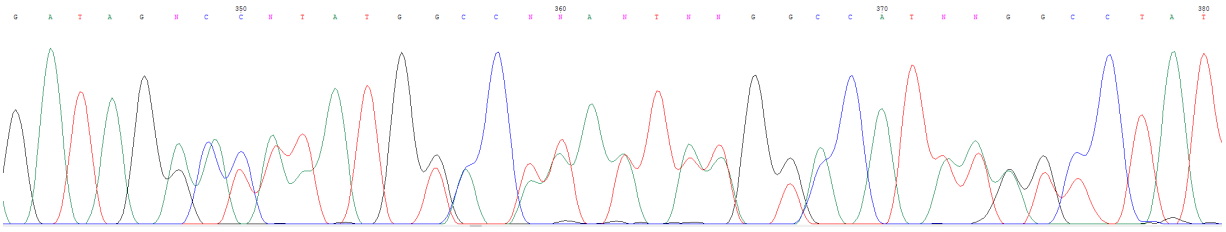

Figure 28: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in ZNF57 in the normal sample.

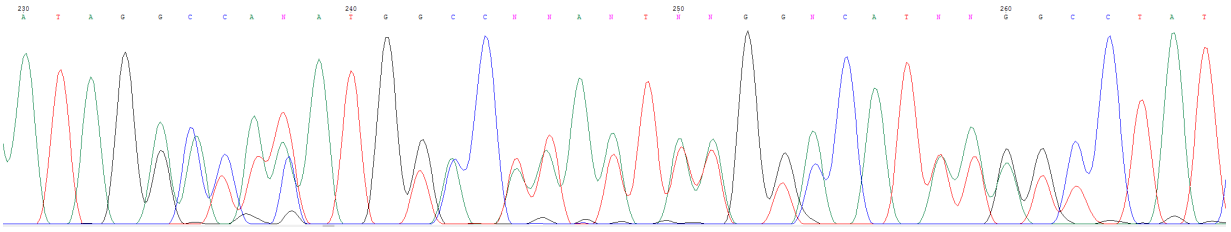

Figure 29: Chromatogram for cDNA produced when using forward primers capturing the microinversion in ZNF57 in the tumor sample.

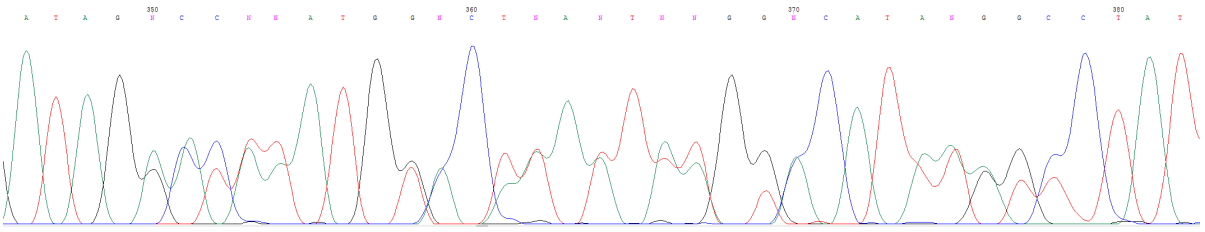

Figure 30: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in ZNF57 in the tumor sample.

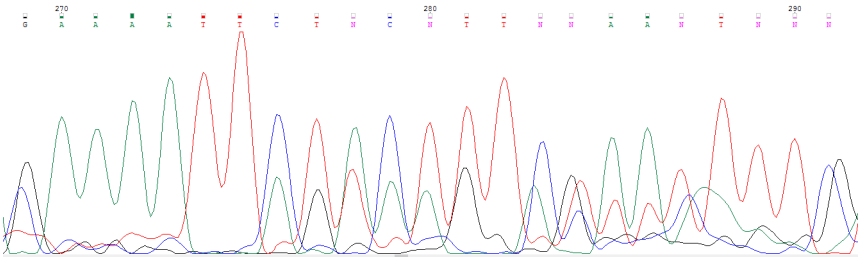

Figure 31: Chromatogram for cDNA produced when using forward primers capturing the microinversion in OSBP2 in the normal sample.

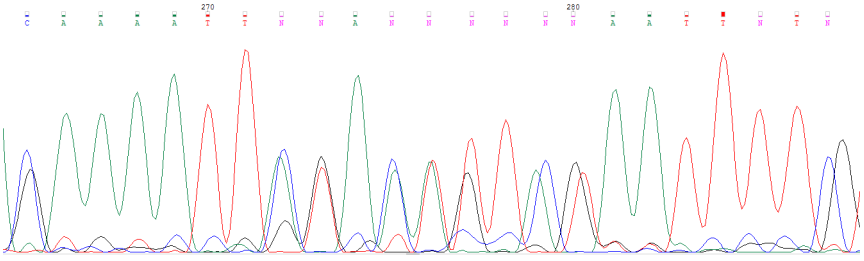

Figure 32: Chromatogram for cDNA produced when using forward primers capturing the microinversion in OSBP2 in the tumor sample.

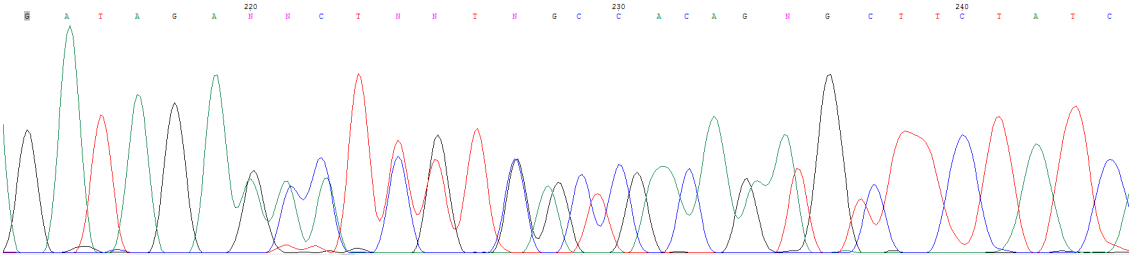

Figure 33: Chromatogram for cDNA produced when using forward primers capturing the microinversion in GNG12-AS1 in the normal sample.

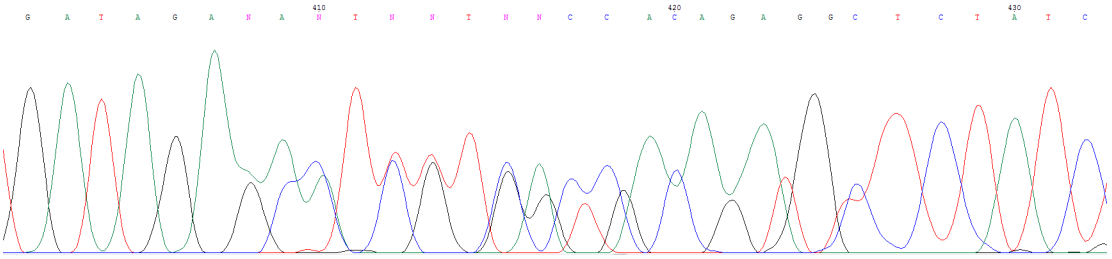

Figure 34: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in GNG12-AS1 in the normal sample.

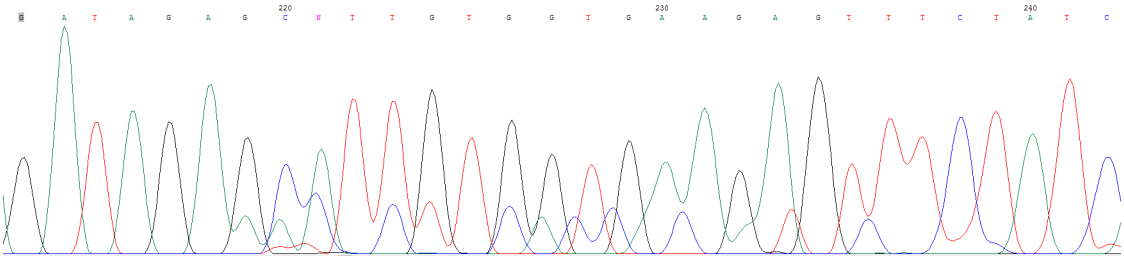

Figure 35: Chromatogram for cDNA produced when using forward primers capturing the microinversion in GNG12-AS1 in the tumor sample.

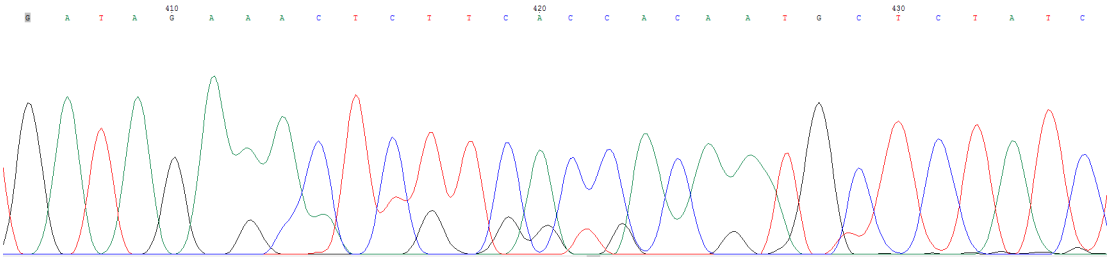

Figure 36: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in GNG12-AS1 in the tumor sample.

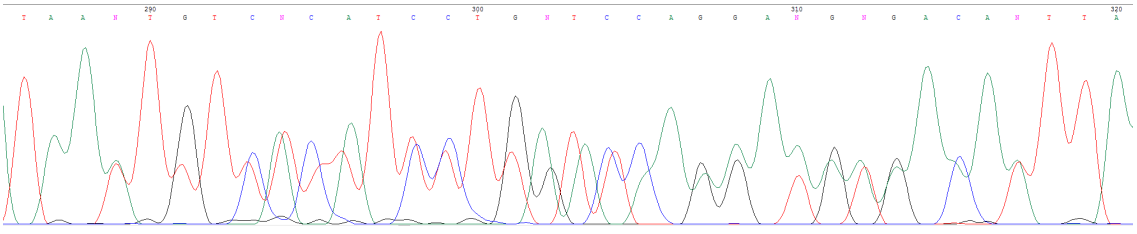

Figure 37: Chromatogram for cDNA produced when using forward primers capturing the microinversion in LINGO2 in the normal sample.

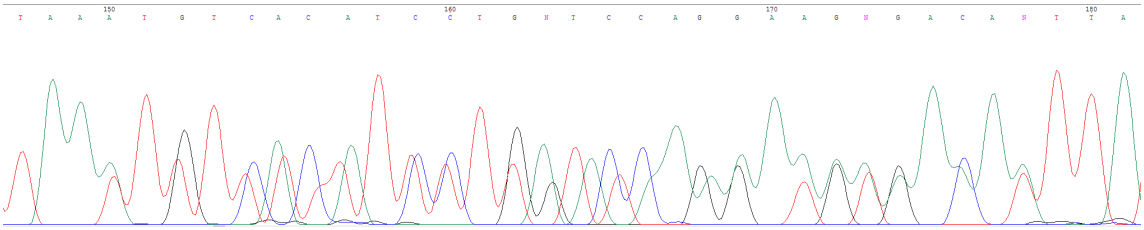

Figure 38: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in LINGO2 in the normal sample.

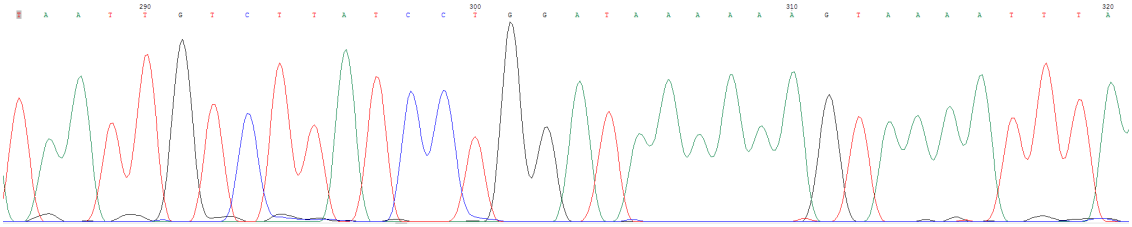

Figure 39: Chromatogram for cDNA produced when using forward primers capturing the microinversion in LINGO2 in the tumor sample.

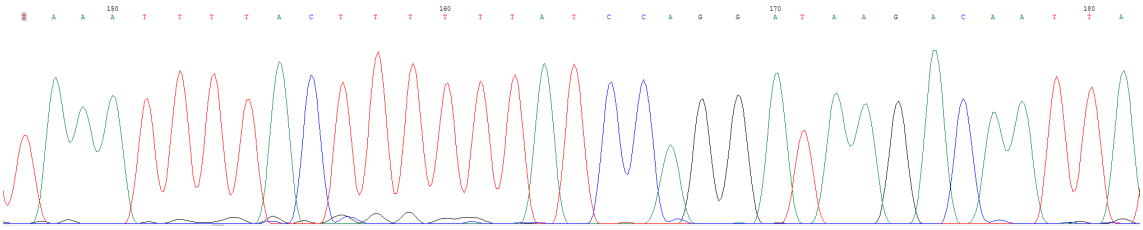

Figure 40: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in LINGO2 in the tumor sample.

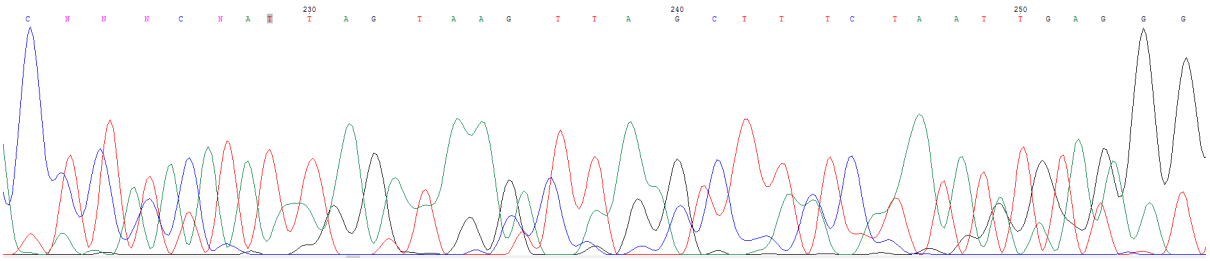

Figure 41: Chromatogram for cDNA produced when using forward primers capturing the microinversion in UBP1 in the normal sample.

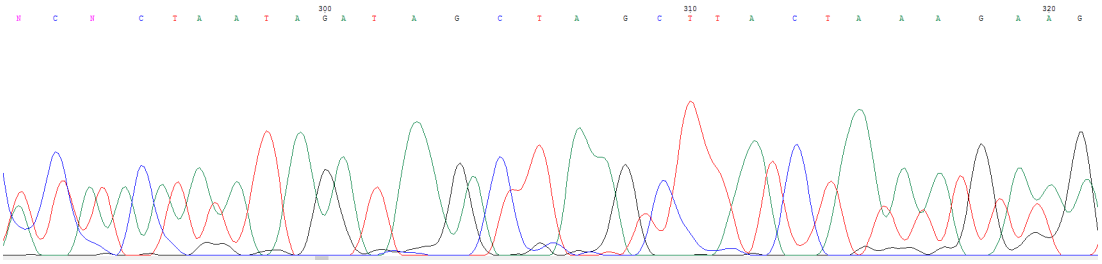

Figure 42: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in UBP1 in the normal sample.

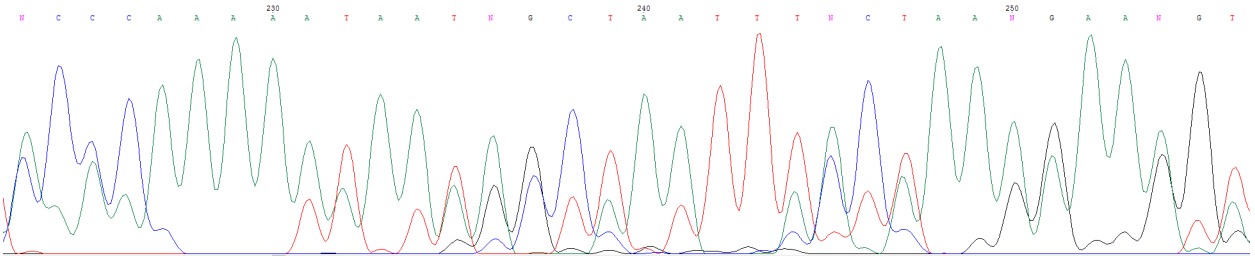

Figure 43: Chromatogram for cDNA produced when using forward primers capturing the microinversion in UBP1 in the tumor sample.

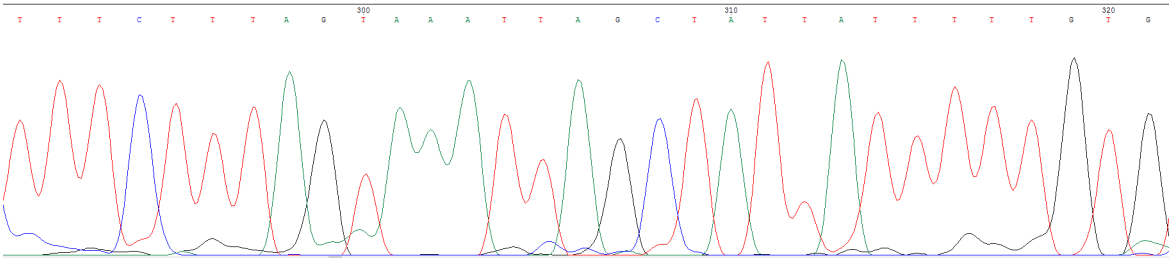

Figure 44: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in UBP1 in the tumor sample.

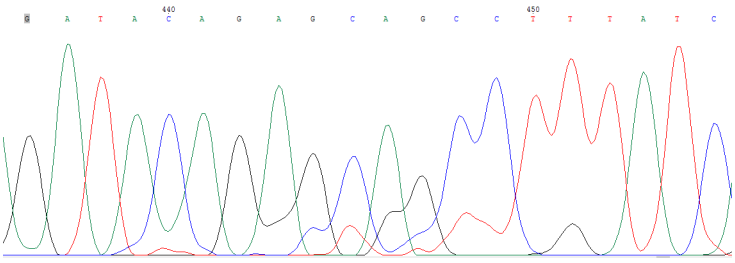

Figure 45: Chromatogram for cDNA produced when using forward primers capturing the microinversion in BOK in the normal sample.

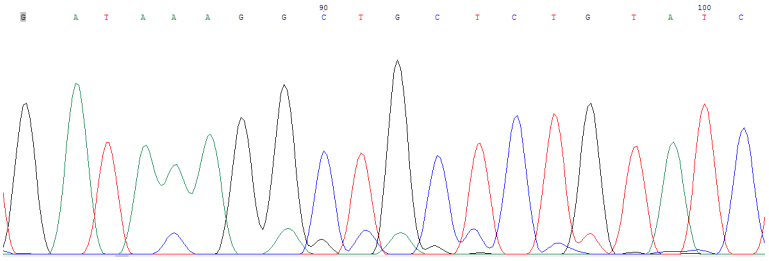

Figure 46: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in BOK in the normal sample.

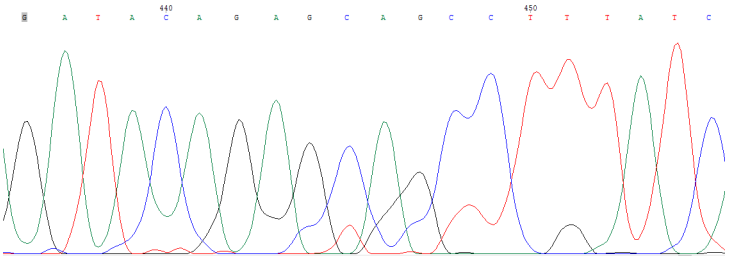

Figure 47: Chromatogram for cDNA produced when using forward primers capturing the microinversion in BOK in the tumor sample.

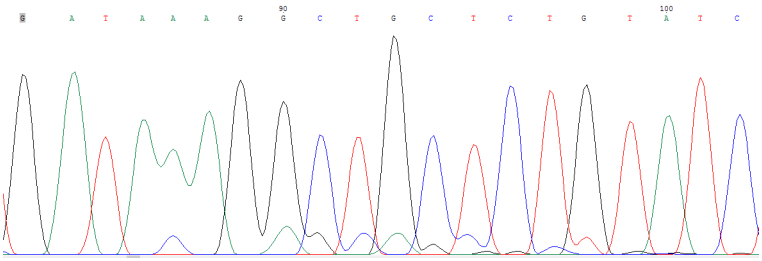

Figure 48: Chromatogram for cDNA produced when using reverse primers capturing the microinversion in BOK in the tumor sample.

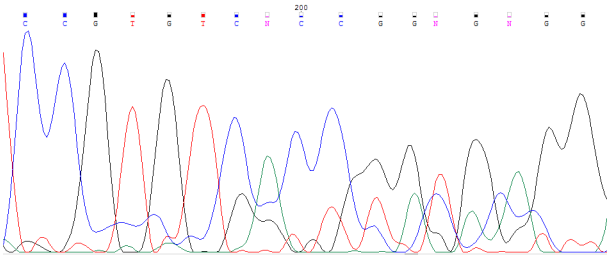

Figure 49: Chromatogram for cDNA produced when using forward primers capturing the microduplication in GTPBP6 in the normal sample.

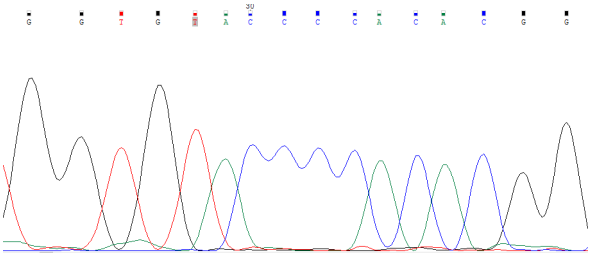

Figure 50: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in GTPBP6 in the normal sample.

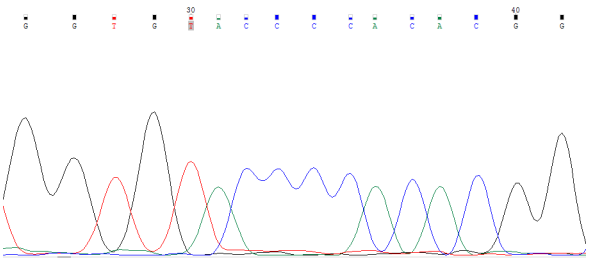

Figure 51: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in GTPBP6 in the tumor sample.

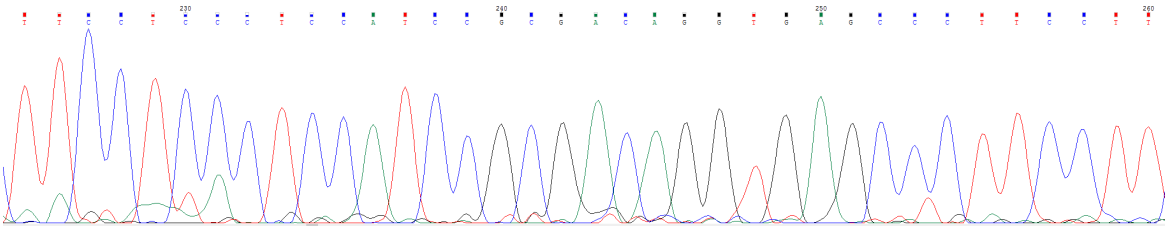

Figure 52: Chromatogram for cDNA produced when using forward primers capturing the microduplication in FAM20C in the tumor sample.

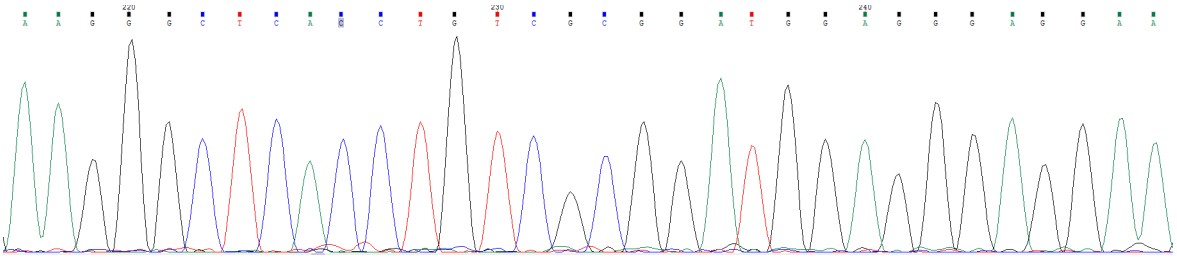

Figure 53: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in FAM20C in the tumor sample.

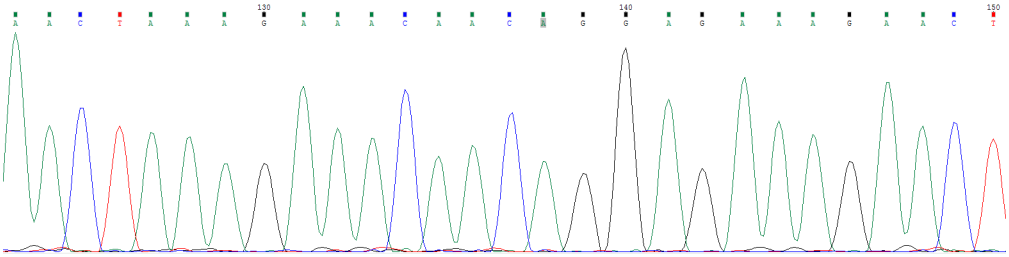

Figure 54: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in KIAA1009 in the tumor sample.

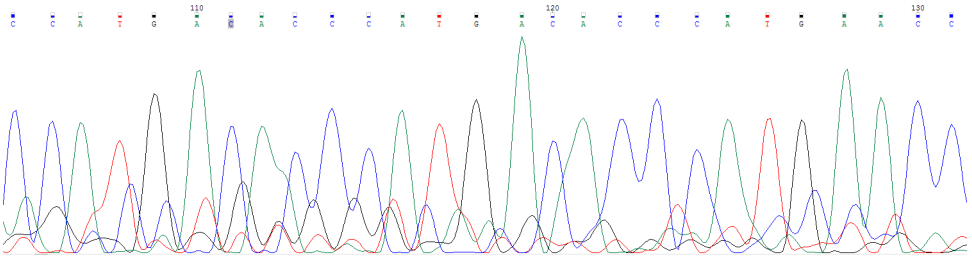

Figure 55: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in BAIAP2L2 in the normal sample.

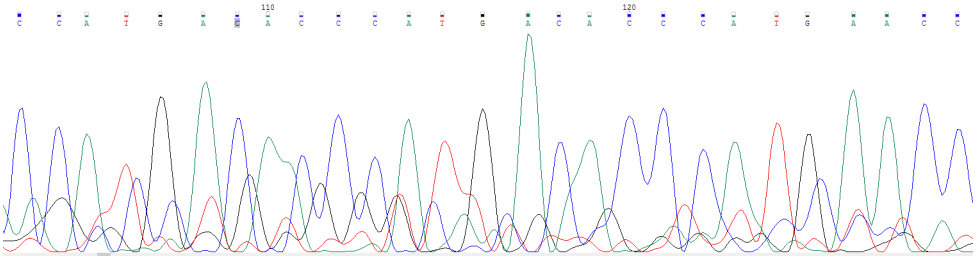

Figure 56: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in BAIAP2L2 in the tumor sample.

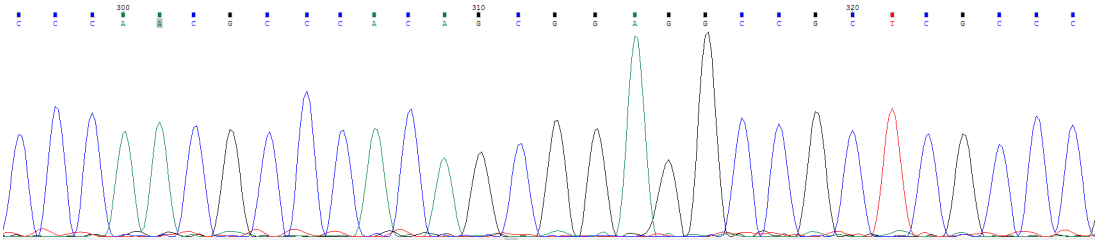

Figure 57: Chromatogram for cDNA produced when using forward primers capturing the microduplication in RBMXL3 in the normal sample.

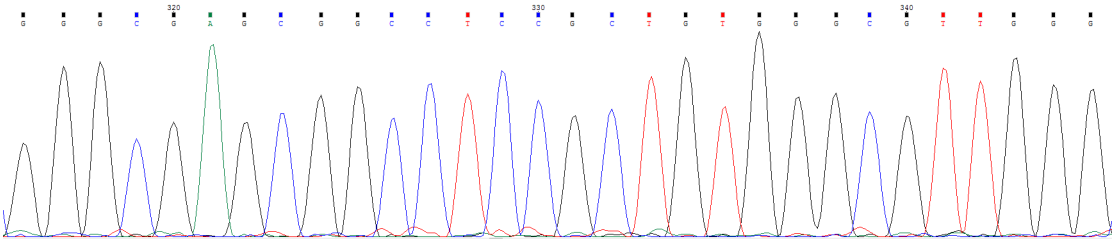

Figure 58: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in RBMXL3 in the normal sample.

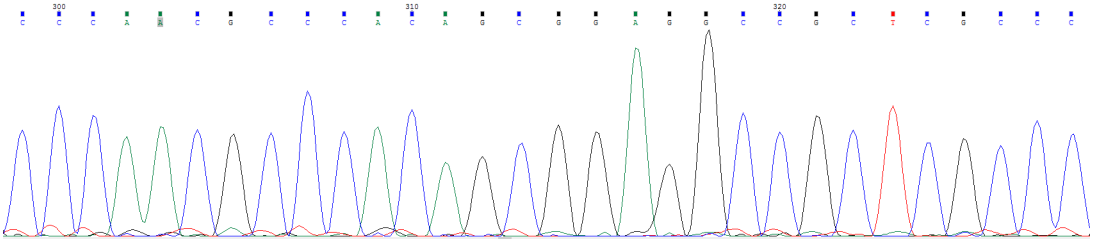

Figure 59: Chromatogram for cDNA produced when using forward primers capturing the microduplication in RBMXL3 in the tumor sample.

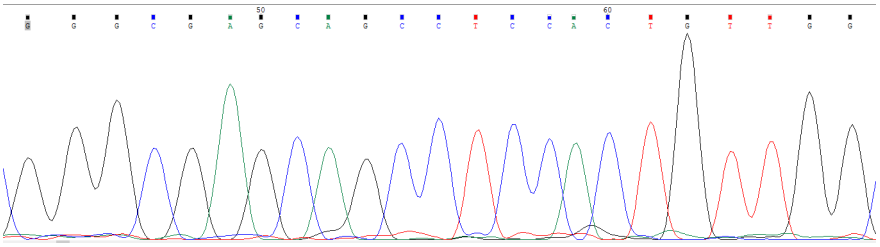

Figure 60: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in RBMXL3 in the tumor sample.

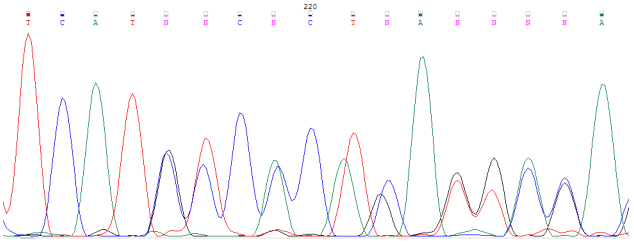

Figure 61: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in GPRIN2 in the normal sample.

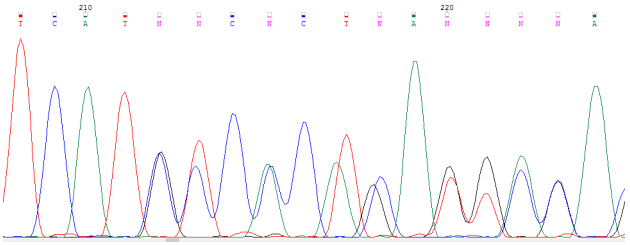

Figure 62: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in GPRIN2 in the tumor sample.

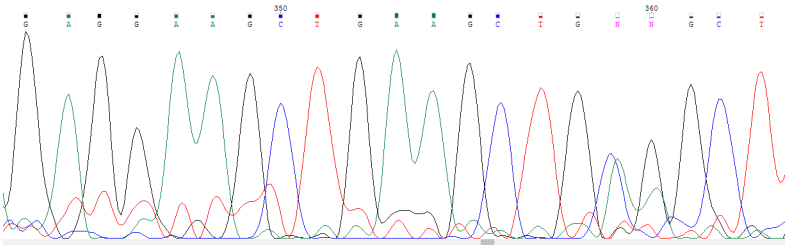

Figure 63: Chromatogram for cDNA produced when using forward primers capturing the microduplication in PALM2-AKAP2 in the normal sample.

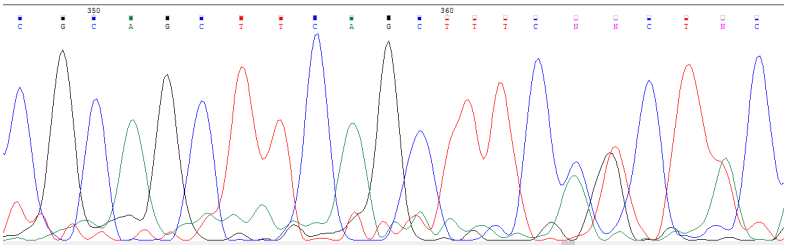

Figure 64: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in PALM2-AKAP2 in the normal sample.

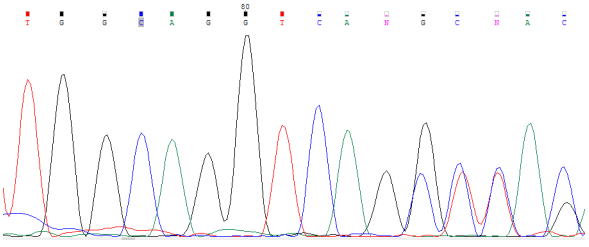

Figure 65: Chromatogram for cDNA produced when using forward primers capturing the microduplication in PRSS48 in the normal sample.

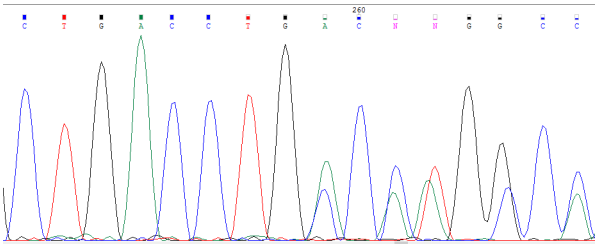

Figure 66: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in PRSS48 in the normal sample.

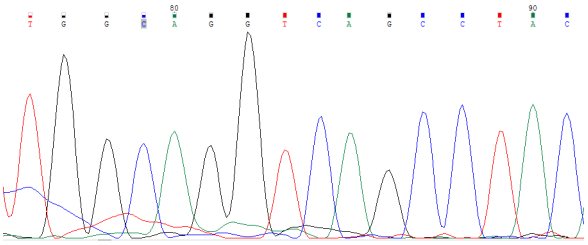

Figure 67: Chromatogram for cDNA produced when using forward primers capturing the microduplication in PRSS48 in the tumor sample.

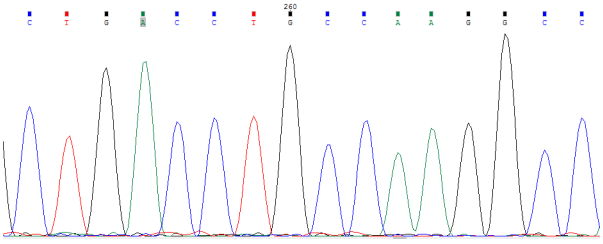

Figure 68: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in PRSS48 in the tumor sample.

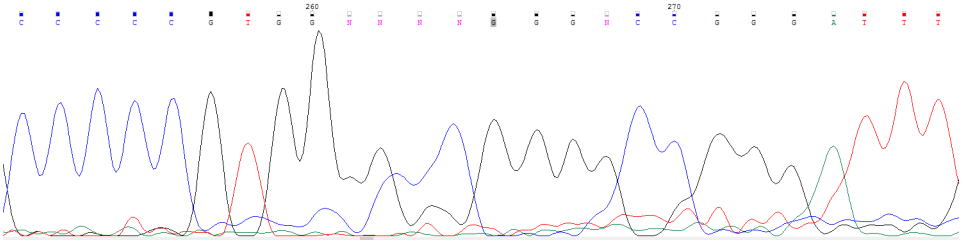

Figure 69: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in ADAMTS19 in the tumor sample.

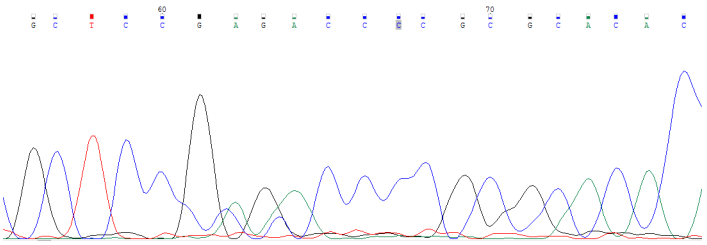

Figure 70: Chromatogram for cDNA produced when using forward primers capturing the microduplication in CIDEA in the tumor sample.

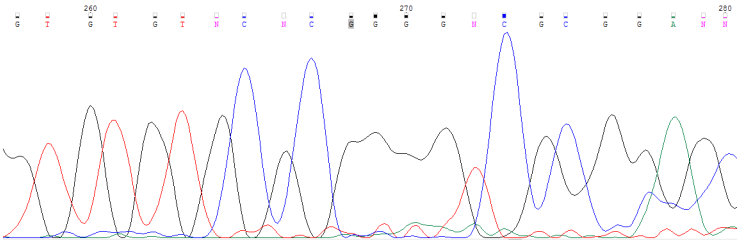

Figure 71: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in CIDEA in the tumor sample.

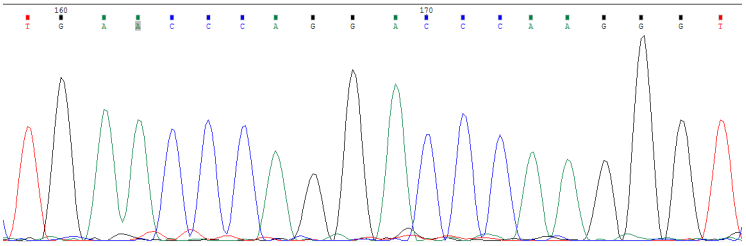

Figure 72: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in ADAMTS7 in the normal sample.

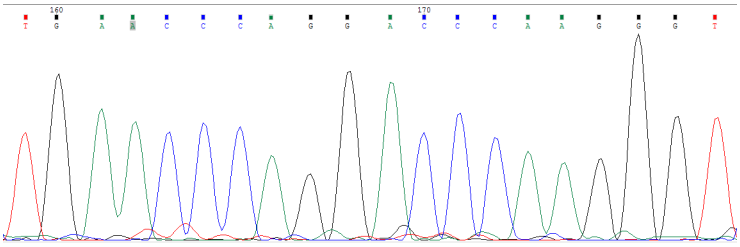

Figure 73: Chromatogram for cDNA produced when using reverse primers capturing the microduplication in ADAMTS7 in the tumor sample.

References

[1] McPherson, A. *et al. PLoS Comput Biol* **7**, e1001138 (2011).

[2] Hach, F. *et al. Nature methods* **7**, 576–577 (2010).

[3] Hach, F. *et al. Nucleic acids research* (2014).

[4] Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. *Genome Research* **19**, 1270–1278 (2009).

[5] Gallant, J., Maier, D. & Astorer, J. *Journal of Computer and System Sciences* **20**, 50 – 58 (1980).

[6] Blum, A., Jiang, T., Li, M., Tromp, J. & Yannakakis, M. *J. ACM* **41**, 630–647 (1994).

[7] Schöniger, M. & Waterman, M. S. *Bulletin of mathematical biology* **54**, 521–536 (1992).

[8] Needleman, S. B. & Wunsch, C. D. *J. Mol. Biol.* **48**, 443–453 (1970).

[9] Kim, S. & Pevzner, P. A. *Nature Communications* **5**, 5277+ (2014).

[10] Rudnick, P. A. *et al. J. Proteome Res.* **15**, 1023–1032 (2016). URL `http://dx.doi.org/10.1021/acs.jproteome.5b01091`.

[11] Nesvizhskii, A. I. *Nat Meth* **11**, 1114–1125 (2014).

[12] Dobin, A. *et al. Bioinformatics (Oxford, England)* **29**, 15–21 (2013). URL `http://dx.doi.org/10.1093/bioinformatics/bts635`.

[13] Schroder, J. *et al. Bioinformatics* (2014).

[14] Bartenhagen, C. & Dugas, M. *Brief. Bioinformatics* (2015).

[15] Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. *Bioinformatics* **25**, 2865–2871 (2009).

[16] Rausch, T. *et al. Bioinformatics* **28**, i333–i339 (2012).

[17] Fan, X., Abbott, T. E., Larson, D. & Chen, K. *Curr Protoc Bioinformatics* **2014** (2014).

[18] Chiba, K. *et al. Bioinformatics* **31**, 116–118 (2015).

[19] Sherry, S. T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).

[20] Li, D. *et al. Bioinformatics* **21**, 3049–3050 (2005).

[21] Wang, L.-H. H. *et al. Rapid communications in mass spectrometry : RCM* **21**, 2985–2991 (2007).

[22] Balafoutas, D. *et al. BMC cancer* **13**, 271+ (2013).

[23] Futreal, P. A. *et al. Nature reviews. Cancer* **4**, 177–183 (2004).

[24] Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. & Cooper, C. S. *Nat Rev Cancer* **10**, 59–64 (2010).

[25] Yu, D. & Hung, M.-C. *Oncogene* **19**, 6115–6121 (2000).

[26] Lwin, Z.-M. *et al. Modern Pathology* **23**, 1559–1566 (2010).

[27] Mendez, M. G., Kojima, S.-I. & Goldman, R. D. *FASEB journal* **24**, 1838–1851 (2010).

[28] Hara, Y. & Imanishi, T. *BMC Evol. Biol.* **11**, 308 (2011).

[29] Szamalek, J. M. *et al. Hum. Genet.* **119**, 103–112 (2006).

[30] Ma, J. *et al. Genome Res.* **16**, 1557–1565 (2006).

[31] Kelchner, S. A. & Wendel, J. F. *Curr. Genet.* **30**, 259–262 (1996).

[32] Nakao, M. *et al. Leukemia* **10**, 1911–1918 (1996).

[33] Tagliabracci, V. S. *et al. Cell* **161**, 1619–1632 (2015).

[34] Koboldt, D. C. *et al. Nature* **490**, 61–70 (2012).

[35] Kim, M. S. *et al. Biochem. Biophys. Res. Commun.* **370**, 38–43 (2008).

[36] Robinson, T. J. *et al. PLoS ONE* **8**, e78641 (2013).

[37] Liu, X. *et al. Cancer Res.* **74**, 6623–6634 (2014).

[38] Kanehira, M. *et al. Cancer Res.* **67**, 3276–3285 (2007).

[39] Ansari, D. *et al. J. Cancer Res. Clin. Oncol.* **141**, 369–380 (2015).

[40] Cirillo, D. *et al. Genome Biol.* **15**, R13 (2014).

[41] Poetsch, A. R. *et al. Epigenetics* **9**, 1252–1260 (2014).

[42] Shriver, S. P. *et al. Mutat. Res.* **406**, 9–23 (1998).

[43] Chen, L. C. *et al. Cancer Res.* **54**, 3021–3024 (1994).

[44] Stanford, J. L. *et al. Cancer Res.* **57**, 1194–1198 (1997).

[45] Warde-Farley, D. *et al. Nucleic Acids Res.* **38**, W214–220 (2010).

[46] Soria-Bretones, I., Saez, C., Ruiz-Borrego, M., Japon, M. A. & Huertas, P. *Cancer Med* **2**, 774–783 (2013).

[47] Yuan, J. & Chen, J. *J. Biol. Chem.* **285**, 1097–1104 (2010).

[48] Sartori, A. A. *et al. Nature* **450**, 509–514 (2007).