Prioritisation of structural variant calls in cancer genomes

Miika J Ahdesmäki [Corresp., 1] , Brad A Chapman [2] , Pablo Cingolani [3] , Oliver Hofmann [4] , Aleksandr Sidoruk [5, 6] , Zhongwu Lai [7] , Gennadii Zakharov [5, 8] , Mikhail Rodichenko [5] , Mikhail Alperovich [5] , David Jenkins [9] , T. Hedley Carr [1] , Daniel Stetson [7] , Brian Dougherty [7] , J. Carl Barrett [7] , Justin H Johnson [7]

[1] Innovative Medicines and Early Development, Oncology, AstraZeneca, Cambridge, United Kingdom

[2] Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States

[3] Kew Inc., Cambridge, Massachusetts, United States

[4] Centre for Cancer Research, University of Melbourne, Melbourne, Australia

[5] EPAM Systems Inc., Newtown, Pennsylvania, United States

[6] Department of software engineering, St. Petersburg State University, St. Petersburg, Russia

[7] Innovative Medicines and Early Development, Oncology, AstraZeneca, Waltham, Massachusetts, United States

[8] Pavlov Institute of Physiology, Russian Academy of Sciences, St. Petersburg, Russia

[9] Boston University, Boston, Massachusetts, United States

Corresponding Author: Miika J Ahdesmäki
Email address: miika.ahdesmaki@astrazeneca.com

Sensitivity of short read DNA-sequencing for gene fusion detection is improving, but is hampered by the significant amount of noise composed of uninteresting or false positive hits in the data. In this paper we describe a tiered prioritisation approach to extract high impact gene fusion events from existing structural variant calls. Using cell line and patient DNA sequence data we improve the annotation and interpretation of structural variant calls to best highlight likely cancer driving fusions. We also considerably improve on the automated visualisation of the high impact structural variants to highlight the effects of the variants on the resulting transcripts. The resulting framework greatly improves on readily detecting clinically actionable structural variants.

Prioritisation of Structural Variant Calls in Cancer Genomes

Miika Ahdesmäki[1], **Brad Chapman**[2], **Pablo Cingolani**[3], **Oliver Hofmann**[4], **Aleksandr Sidoruk**[5,8], **Zhongwu Lai**[6], **Gennadii Zakharov**[5,9], **Mikhail Rodichenko**[5], **Mikhail Alperovich**[5], **David Jenkins**[7], **T. Hedley Carr**[1], **Daniel Stetson**[6], **Brian Dougherty**[6], **J. Carl Barrett**[6], and **Justin Johnson**[6]

[1]**Innovative Medicines and Early Development, Oncology, AstraZeneca, Cambridge, United Kingdom**
[2]**Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States**
[3]**Kew Inc., Cambridge, Massachusetts, United States**
[4]**Centre for Cancer Research, University of Melbourne, Melbourne, Australia**
[5]**EPAM Systems Inc., Newtown, Pennsylvania, United States**
[6]**Innovative Medicines and Early Development, Oncology, Waltham, Massachusetts, United States**
[7]**Boston University, Boston, Massachusetts, United States**
[8]**Department of software engineering, St. Petersburg State University, St. Petersburg, Russia**
[9]**Pavlov Institute of Physiology, Russian Academy of Sciences, St. Petersburg, Russia**

Corresponding author:
Miika Ahdesmäki[1]

Email address: miika.ahdesmaki@astrazeneca.com

ABSTRACT

Sensitivity of short read DNA-sequencing for gene fusion detection is improving, but is hampered by the significant amount of noise composed of uninteresting or false positive hits in the data. In this paper we describe a tiered prioritisation approach to extract high impact gene fusion events from existing structural variant calls. Using cell line and patient DNA sequence data we improve the annotation and interpretation of structural variant calls to best highlight likely cancer driving fusions. We also considerably improve on the automated visualisation of the high impact structural variants to highlight the effects of the variants on the resulting transcripts. The resulting framework greatly improves on readily detecting clinically actionable structural variants.

INTRODUCTION

Structural variants (SVs) such as inversions, tandem duplications, large deletions and more complex chromosomal rearrangements are implicated as driver events in multiple cancers (Latysheva and Babu, 2016). Clinical detection of SVs in Mendelian diseases has been considered by e.g. Noll et al. (2016) but to our knowledge no prioritisation approach for oncology is publicly available. The mechanisms for oncogenic driver generation include activating fusions combining the coding frames (quite often in the intronic regions) of two genes, as well as truncating mutations in tumor suppressor genes or whole exon losses. Some well understood examples include TMPRSS2-ERG in prostate cancer (Tomlins et al., 2008), FGFR1,3-TACC1,3 in bladder and other cancers (The Cancer Genome Atlas Research Network, 2014), EGFRv3 deletion in glioblastoma and other tumours (Sugawa et al., 1990) and EML4-ALK in lung cancer (Soda et al., 2007).

The accurate calling of these complex, structural variants in short read DNA sequencing data is complicated by noise, manifested as false positives and lack of specificity. In many cases, the number of whole genome SV calls, including complex breakends, can be in the tens of thousands. While long

46 read sequencing is likely to improve the calling of structural variants especially in germline DNA, tumour
47 DNA from formalin-fixed paraffin-embedded (FFPE) tissue and circulating tumour DNA samples is
48 inherently limited to short DNA fragment size. Utilising the currently available large amount of short
49 read sequencing data to the full is therefore well motivated. It is also imperative to promptly pinpoint any
50 clinically important structural variants when present in data.

51 In this paper we propose a tiered prioritisation approach to extract structural variants most likely to
52 contribute to cancer proliferation and enable validation and follow up for a subset of high priority events.
53 The prioritisation is based on greatly improved structural variant annotation in the variant annotation
54 tool SnpEff (Cingolani et al., 2012) and can be applied to the output of any state of the art SV caller.
55 Similar prioritisation work has been published in the domain of small variants, see for example Carr
56 et al. (2016); Munz et al. (2015). The important aspect of easy and automated visualisation of the effects
57 of structural variants on genes and coding exons is often overlooked with focus on structural variant
58 calling algorithm performance. We thus further implement interactive structural variant visualisations
59 in the New Genome Browser (`https://github.com/epam/NGB`). We show the full utility of
60 the improved prioritisation and visualisation approaches in samples with structural variants leading to
61 oncogenic gene fusions. The calling and filtering of RNA-seq based expressed, typically gain of function,
62 fusions is well established and could be used to complement the DNA-focus of our approach, see e.g.
63 `https://github.com/ndaniel/fusioncatcher` (Nicorici et al., 2014).

METHODS

65 The prioritisation approach proposed here bins SVs into tiers on predicted effect and builds on the output
66 of any readily available SV calling algorithm. No new SV calling algorithm is proposed. In brief, most
67 short read SV calling pipelines start with alignment of the DNA data to the human reference using an
68 aligner like bwa-mem (Li, 2013). This is followed by the chosen SV caller integrating evidence from split
69 and discordant reads, and potentially coverage (Alkan et al., 2011), to make structural variant calls for
70 deletions (DEL), tandem duplications (DUP), inversions (INV) and other more complex variants (BND).
71 An example of these events is visualised in figure 1 in Tattini et al. (2015). The prioritisation is indifferent
72 to the type of SV and it is the expected effect, such as a gene fusion that is of primary focus.

73 For SV calling we utilised two freely available SV callers that integrate evidence from split and
74 discordant reads, Manta (Chen et al., 2016) and Lumpy (Layer et al., 2014). Both benchmark well
75 in synthetic somatic data sets (see ICGC-TCGA DREAM Mutation Calling challenge leaderboards
76 `https://www.synapse.org/#!Synapse:syn312572/wiki/247695`) as well as germline
77 reference standards (Genome in a Bottle NA12878). Any structural variant caller (such as BRASS,
78 `https://github.com/cancerit/BRASS`) producing vcf files compliant with the vcf specifi-
79 cation (`https://samtools.github.io/hts-specs/VCFv4.3.pdf`) and compatible with
80 SnpEff could equally well be used with our proposed methodology, provided they also quantify the
81 numbers of discordant and split reads supporting at least the alternative allele.

82 To practically facilitate the prioritisation we improved annotations in SnpEff 4.3 to ease interpretation
83 of fusion events, adding the Sequence Ontology (Eilbeck et al., 2005) annotation type *gene_fusion* for
84 events where the open reading frames are in the same direction. Further, *bidirectional_gene_fusion*
85 was introduced for where the frames of the putatively fused genes are opposing and therefore unlikely
86 to be functional and *frameshift_variant* when the coding of the resulting fusion is out of frame, thus
87 likely resulting in a truncated protein. The last two types are very important and interesting for loss
88 of function of e.g. tumour suppressors. Other annotation improvements in SnpEff 4.3 include: *chro-
89 mosome_number_variation*, *duplication* and *inversion*, which refer to large chromosomal deletions,
90 duplications and inversions respectively (involving a whole exon, transcript, gene or even larger genomic
91 regions), *exon_loss_variant* (whole or significant part of the exon was deleted) and *feature_ablation* (whole
92 gene deleted).

93 We built a three tier prioritisation system (`https://github.com/AstraZeneca-NGS/simple_`
94 `sv_annotation`, Ahdesmaki (2016)) using fusion and exon loss annotations. Given a list of genes of
95 interest (GOI) we assign priorities, given in parentheses, as follows:

96 • Gene fusion
97   Fusion affecting two genes based on SnpEff annotation
98   on list of known pairs from FusionCatcher (Nicorici et al., 2014) (1)
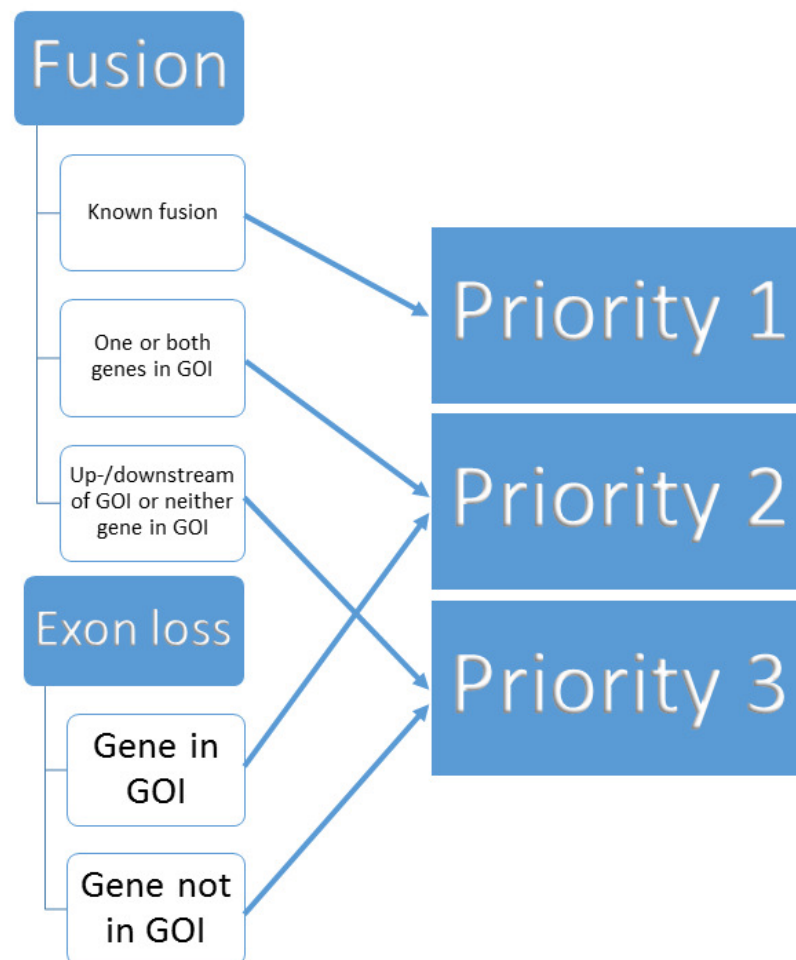
Figure 1. Binning of structural variants into 3 priorities.

99   one or both genes on GOI list (2)
100   neither on known pairs or GOI list (3)
101   Fusion affecting one gene
102   on GOI list (2)
103   not on GOI list (3)
104 • Whole exon loss based on SnpEff annotation
105   affecting a gene on GOI list (2)
106   not affecting a gene on GOI list (3)
107 • Upstream or downstream of GOI list genes based on SnpEff annotation (3)
108 • Other variant (REJECT)
109 • Missing ANN or SVTYPE in variant call file (REJECT)

110   The process is visualised in Figure 1.

111   An example of a priority one gene fusion is given below, where the ANN field is the annotation
112 provided by SnpEff. The duplication is interpreted by SnpEff as a gene fusion affecting FGFR3 and
113 TACC3. This is a gene fusion on the list of known and published fusions and therefore is given priority
114 one.

```
115  chr4      1727831 947      N        <DUP>    0.0      PASS      SVTYPE=DUP;
116    SVLEN=79207;END=1807038;STRANDS=−+:9072;CIPOS=0,0;CIEND=0,0;
117    CIPOS95=0,0;CIEND95=0,0;SU=9072;PE=4054;SR=5018;AC=0;AN=0;ANN=<
118    DUP>|gene_fusion|HIGH|FGFR3&TACC3|ENSG00000068078&ENSG00000013810
```

3/13

```
119   |gene_variant|ENSG00000013810|||||||||   GT:SU:PE:SR:GQ:SQ:GL:DP:
120   RO:AO:QR:QA:RS:AS:RP:AP:AB
121   ./.:9072:4054:5018:.:.:−13017,−1748,−453:11177:3218:
122   7958:3217:7957:1516:4654:1701:3303:0.71
```

For the GOI, as a supplement to the prioritisation implementation we have provided a list of 300+ genes commonly associated with cancer, including genes involved in the MAPK and PI3K pathways (including receptor tyrosine kinase genes), DNA damage response, immuno-oncology and others. Alternatively, the user can provide their own gene lists in the implementation. The proposed prioritisation approach can be applied to variants from both paired (tumour/normal) and tumour only data, depending only on the structural variant callers' capabilities to handle paired samples. We confirmed the approach using TCGA data with known gene fusions.

RESULTS AND DISCUSSION

Prioritisation efficiently identifies clinically relevant gene fusions

We estimated the ability of our prioritisation approach to retain known mutations while reducing false positive events using samples with known structural variants (in a background of less well characterised SVs). Synthetic datasets from the ICGC-TCGA DREAM Mutation Calling challenge have also known artificial structural variants spiked in. These artificial SVs are useful, but do not however necessarily represent a realistic quantity of somatic SVs since they are not generated from a biological model, so we focused on known events in real sequenced samples. We collected sequencing data for seven samples with known SVs from cell lines, a patient derived xenograft and a clinical sample (Table 1). Although whole genome sequencing data or targeted capture including introns is preferred, any hybrid capture data can yield meaningful results if the breakpoints are close to captured regions or there are off target reads.

Sample	Panel or WGS	Manta call	Lumpy call	Fusion(s)
HDC134P rep. 1	Panel with intronic probes	INV	INV	EML4-ALK
HDC134P rep. 2	Panel with intronic probes	INV	INV	EML4-ALK
HDC134P rep. 3	Panel with intronic probes	INV	BND	EML4-ALK
HDC140P rep. 1	Panel with intronic probes	INV	BND	CCDC6-RET
HDC140P rep. 2	Panel with intronic probes	INV	BND	CCDC6-RET
HDC140P rep. 3	Panel with intronic probes	INV	BND	CCDC6-RET
HDC141P rep. 1	Panel with intronic probes	BND	BND	ROS1-SLC34A2
HDC141P rep. 2	Panel with intronic probes	BND	BND	ROS1-SLC34A2
HDC141P rep. 3	Panel with intronic probes	BND	BND	ROS1-SLC34A2
MCF7	WGS	DUP	DUP	ESR1-CCDC170
RT4	Panel with intronic probes	DUP	DUP	TACC3-FGFR3
PDX model	WES	DUP	DUP	TACC3-FGFR3
Prostate cancer patient sample	FMI Panel with intronic probes	DEL	DEL	TMPRSS2-ERG

Table 1. Collection of structural variants leading to oncogenic fusions in different sample types. All events are ranked into the highest category (1) by the prioritisation scheme.

Following bwa-mem alignment to hg38, Lumpy and Manta both call the breakpoints and event types for the structural variants in Table 1, with slightly different interpretations for some like CCDC6-RET. As part of the updates to SnpEff we ensured that all these different types of SV events (INV, DUP, DEL) affecting two genes were correctly annotated as gene fusions.

The total number of calls for the samples in Table 1 as well as the numbers of variants falling into the tiers are shown in Table 2. The percentages of events falling into tier 1 ranged from between 0.2% (Manta MCF7 calls) to 40% (Manta HDC134P rep. 1). Focusing on the whole genome sequenced (WGS) MCF7 sample, the concordance of tier 1 events between Manta and Lumpy was fourteen SVs, with Manta having two private tandem duplications and Lumpy one private tandem duplication in addition to six private BND events. This followed the general trend of Lumpy calling more BND events than Manta. The read support for the private SVs did not differ in any great manner from the shared SVs and therefore could still be true positives.

For MCF7 we provide a text file of the tier 1 and tier 2 events at `https://github.com/AstraZeneca-NGS/publication_data`, in addition to providing the vcf files.

In Table 2, the primary priority (known fusions) column lists the one true fusion known to be present in the HDC, RT4 cell lines and the patient sample. Manta reported two close but different breakpoints for the EML4-ALK fusion in the HDC134P replicates. The background of the PDX model is not fully characterised and therefore in the list of 5 fusions there may be false positives or SVs of unknown significance besides the FGFR3-TACC3 fusion. MCF7 has been previously characterised to at least contain the activating ESR1-CCDC170 fusion (Veeraraghavan et al., 2014) and be ESR1 driven. The remaining > 10 fusions may be false positives or of less importance. This highlights that for the thousands of SV calls the prioritisation correctly draws the attention to the true positive events in the the primary priority. If no events of interest are found in the primary priority bin the secondary priority can contain novel fusions or exon loss events of high interest as well. The tertiary priority (upstream, downstream events in genes of interest and fusions in genes of uninterest) is a catch-all category that should receive less attention.

To visualise the prioritised SV calls in the three replicate samples (HDC134P, HDC140P, HDC141P) run in triplicates, we utilised the UpSet package (Lex et al., 2014) to show con- and discordance in the prioritised calls. The plots in Figure 2 (a), (b) and (c) show the concordance histograms for each of the triplicates. The known fusions are detected by both Lumpy and Manta in all triplicates. There is additionally one structural variant (RET fused with chromosome 13) in HDC140P detected by all the algorithms in all the replicates. All the rest of the calls are private to one caller and one replicate (noise) but the number is small. This shows that the true events are very confidently called by both algorithms but there is a varying amount of false positives with Lumpy producing slightly more.

To show that the proposed approach correctly identifies the true events also in data not part of the sample set in Table 1, we applied the prioritisation to the TCGA bladder cancer cohort (The Cancer Genome Atlas Research Network, 2014). The FGFR3-TACC3 fusion in the RT4 cell line and patient derived xenograft (PDX) model in Table 1 is found in several TCGA samples, e.g. TCGA-CF-A3MF, TCGA-CF-A3MG and TCGA-CF-A3MH. Typically the breakpoints of gene fusions in the intronic regions, however in two of the TCGA samples (TCGA-CF-A3MH, TCGA-CF-A3MF), the FGFR3 breakpoint is in the last exon. The events were correctly annotated in these samples by our approach.

Visualising gene fusions resulting from structural variants

Visualisation of structural variants to highlight the breakpoints and affected exons in a putative fusion transcript is an area of active development with no one tool currently being the industry standard. We initially identified Svviz (Spies et al., 2015), one of the earlier tools, to examine the validated fusion variants highlighted by prioritisation. The FGFR3-TACC3 tandem duplication (RT4 cell line) is shown in Figure 3; TACC3 is not captured by the panel used and therefore no reads in support of the reference allele for TACC3 are shown. Svviz reassembles the reads around the putative breakpoints in its analysis and requires an amount of manual intervention.

We next decided to implement a variant call based gene fusion visualisation scheme in the open source New Genome Browser (NGB, `https://github.com/epam/NGB`). NGB takes the variant breakpoints annotated by SnpEff and uses Ensembl and UniProt based annotation to visualise the fusion product in both reference as well as the actual sequence context. The resulting plots highlight the fused exons of the affected genes. NGB uses vcf files as input and allows filtering based on vcf properties, including based on the priority from the prioritisation scheme to whittle down the putative variants.

The NGB visualisation of the FGFR3-TACC3 fusion is shown in Figure 4 (a). Unlike the Svviz plot (Figure 3), the visualisation is fully interactive html5 in the browser. Red highlighting is used to show the breakpoints relative to the coding regions in the alternative allele view and the red line shows the fusion points in the reference allele view. As NGB is a full feature genome browser, viewing both the read evidence as well as the fusion effects is simple. Figure 4 (b) shows the read level evidence side-by-side from the two breakpoints. Soft clipping of the reads around the breakpoint are shown by the coloured base tails of the reads.

In Figure 5 an interchromosomal translocation resulting in a fusion between ROS1 and SLC34A2 is shown. If multiple genes are overlapping the breakpoints NGB allows choosing the most relevant gene for the researcher. Another example for the EML4-ALK fusion that results from an inversion is shown if Figure 6.

Sample	Algorithm	DUP, DEL, INV	BND	Primary priority	Secondary priority	Tertiary priority
HDC134P rep. 1	Manta	5	0	2	2	0
	Lumpy	43	9	1	0	1
HDC134P rep. 2	Manta	2	0	2	0	0
	Lumpy	41	4	1	0	0
HDC134P rep. 3	Manta	3	0	2	0	0
	Lumpy	41	2	1	2	1
HDC140P rep. 1	Manta	1	1	1	1	0
	Lumpy	48	20	1	2	0
HDC140P rep. 2	Manta	1	1	1	1	0
	Lumpy	30	12	1	1	0
HDC140P rep. 3	Manta	1	1	1	1	0
	Lumpy	56	9	1	3	0
HDC141P rep. 1	Manta	0	2	2	0	0
	Lumpy	24	7	1	0	0
HDC141P rep. 2	Manta	0	2	2	0	0
	Lumpy	17	1	1	0	0
HDC141P rep. 3	Manta	0	2	2	0	0
	Lumpy	25	6	1	1	1
MCF7 (WGS)	Manta	8239	1990	16	48	2814
	Lumpy	4277	2750	21	38	1683
RT4	Manta	169	158	1	20	135
	Lumpy	1509	14659	1	302	10300
PDX model	Manta	248	65	5	2	164
	Lumpy	143	862	5	8	292
Patient sample	Manta	30	51	1	6	37
	Lumpy	1034	3621	1	13	177

Table 2. Raw SV call numbers for Manta and Lumpy are given in the DUP, DEL, INV and BND columns. The prioritised calls are shown in the last three columns. The Primary priority column corresponds to the number of detected fusions reported previously in the literature. All samples are from small hybrid capture panels except for the MCF7 sample, thus the relatively low numbers of calls per sample.

207 Other programs in development to better visualise the read level data include for example Genome
208 Ribbon, see `http://genomeribbon.com/`.

CONCLUSION

210 Here we presented a highly effective scheme for structural variant calling algorithms to prioritise for
211 known fusion events as well as aberrations in a panel of cancer related genes.

212 This method prioritises based on biological information such as genes of interest and can be used in
213 combination with orthogonal discovery based approaches (Ganel et al., 2016). In their approach, Ganel et
214 al. produce in silico SV impact predictions that can be useful when whittling down the number of SVs of
215 unknown significance and narrowing down to the likely most pathogenic ones; if the more hypothesis
216 driven prioritisation described here does not yield satisfactory results, the approach by Ganel et al. might
217 uncover additional novel variants or be used to provide a relative scoring for within-tier SVs. Future work
218 also includes further incorporating protein domain information and which domains are retained in the
219 suspected gene fusion.

220 We further developed a visualisation framework in the New Genome Browser to illustrate the effects
221 of the structural variants on genes in a user friendly, simple manner. We expect these visualisations to be
222 extremely helpful for scientists in quickly producing publication ready gene fusion figures.

223 We look forward to suggestions from other groups to further improve structural variant calling
224 interpretation and visualisation in cancer. This could be in the form of providing lists of genes of interest,
225 suggesting alternative tiers to the prioritisation or adding support for other structural variant callers.
226 The approach could also readily be extended to other fusion driven diseases by replacing the gene lists
227 accordingly.

AVAILABILITY OF DATA AND ALGORITHMS

229 The structural variant call (vcf) level data will be made available for all the samples except for the PDX
230 model at `https://github.com/AstraZeneca-NGS/publication_data`. The proposed pri-
231 oritisation framework is fully integrated along with all the tools required to produce the results from
232 raw sequencing data in bcbio (`https://github.com/chapmanb/bcbio-nextgen`). SnpEff
233 4.3 and later are available at `http://snpeff.sourceforge.net/`. The prioritisation code for
234 structural variants is accessible at `https://github.com/AstraZeneca-NGS/simple_sv_`
235 `annotation`. The New Genome Browser is available at `https://github.com/epam/NGB`. All
236 software used herein is freely available under open source licences.

ACKNOWLEDGMENTS

238 The following centres are kindly acknowledged for sequencing the three replicates of the samples
239 HDC134P, HDC140P and HDC141P.

240 • West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust,
241 Mindelsohn Way, Edgbaston, Birmingham B15 2TG, UK, `http://www.bwnft.nhs.uk/`
242 `healthcare-professional/west-midlands-regional-genetics-laboratory`
243 • Molecular Diagnostics, The Centre for Molecular Pathology, The Royal Marsden, 15 Cotswold
244 Road, Sutton, Surrey SM2 5NG, UK, `http://www.icr.ac.uk`
245 • All Wales Medical Genetics Service, University Hospital of Wales, Heath Park, Cardiff CF14 4XW,
246 UK

REFERENCES

248 Ahdesmaki, M. (2016). Simple SV annotation. doi: 10.5281/zenodo.167572.
249 Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping.
250 *Nat. Rev. Genet.*, 12(5):363–376.
251 Carr, T. H., McEwen, R., Dougherty, B., Johnson, J. H., Dry, J. R., Lai, Z., Ghazoui, Z., Laing, N. M.,
252 Hodgson, D. R., Cruzalegui, F., Hollingsworth, S. J., and Barrett, J. C. (2016). Defining actionable
253 mutations for oncology therapeutic development. *Nat. Rev. Cancer*, 16(5):319–329.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A. J., Kruglyak, S., and Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222.

Cingolani, P., Platts, A., Wang, l. e. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, 6(5):R44.

Ganel, L., Abel, H. J., FinMetSeq Consortium, and Hall, I. M. (2016). SVScore: An Impact Prediction Tool For Structural Variation. *Bioinformatics*, doi: 10.1093/bioinformatics/btw789.

Latysheva, N. S. and Babu, M. M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.*, 44(10):4487–4503.

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, 15(6):R84.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph*, 20(12):1983–1992.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *bioRxiv*, arXiv:1303.3997.

Munz, M., Ruark, E., Renwick, A., Ramsay, E., Clarke, M., Mahamdallie, S., Cloke, V., Seal, S., Strydom, A., Lunter, G., and Rahman, N. (2015). CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. *Genome Med*, 7:76.

Nicorici, D., Satalan, M., Edgren, H., Kangaspeska, S., Murumagi, A., Kallioniemi, O., Virtanen, S., and Kilkku, O. (2014). FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*, doi: 10.1101/011650.

Noll, A. C., Miller, N. A., Smith, L. D., Yoo, B., Fiedler, S., Cooley, L. D., Willig, L. K., Petrikin, J. E., Cakici, J., Lesko, J., Newton, A., Detherage, K., Thiffault, I., Saunders, C. J., Farrow, E. G., and Kingsmore, S. F. (2016). Clinical detection of deletion structural variants in whole-genome sequences. *Nature Genomic Medicine*, 1(16026).

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y., and Mano, H. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153):561–566.

Spies, N., Zook, J. M., Salit, M., and Sidow, A. (2015). svviz: a read viewer for validating structural variants. *Bioinformatics*, 31(24):3994–3996.

Sugawa, N., Ekstrand, A. J., James, C. D., and Collins, V. P. (1990). Identical splicing of aberrant epidermal growth factor receptor transcripts from amplified rearranged genes in human glioblastomas. *Proc. Natl. Acad. Sci. U.S.A.*, 87(21):8602–8606.

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*, 3:92.

The Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322.

Tomlins, S. A., Laxman, B., Varambally, S., Cao, X., Yu, J., Helgeson, B. E., Cao, Q., Prensner, J. R., Rubin, M. A., Shah, R. B., Mehra, R., and Chinnaiyan, A. M. (2008). Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia*, 10(2):177–188.

Veeraraghavan, J., Tan, Y., Cao, X., Kim, J. A., Wang, X., Chamness, G. C., Maiti, S. N., Cooper, L. J. N., Edwards, D. P., Contreras, A., Hilsenbeck, S. G., Chang, E. C., Schiff, R., and Wang, X. (2014). Recurrent ESR1CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nature Communications*, 5(4577).
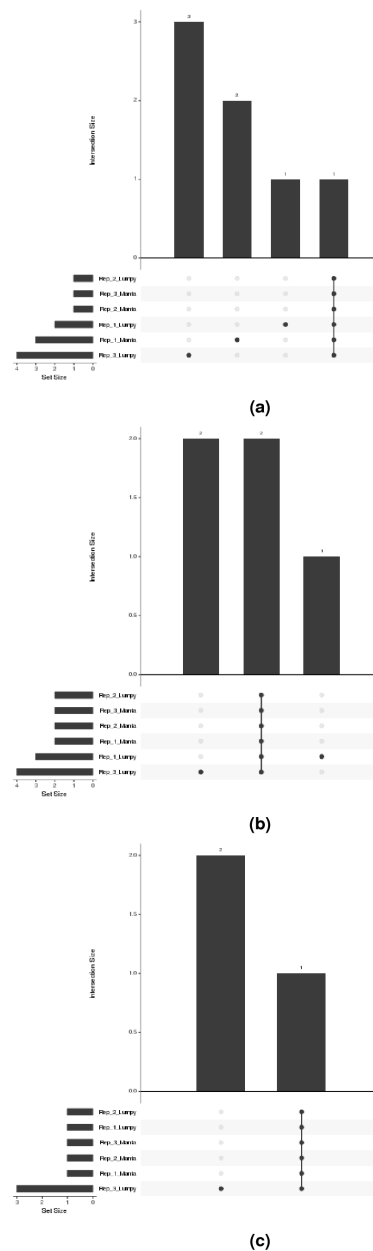
Figure 2. Prioritised SV call concordance. The true positives are concordantly detected in addition to private (non-replicable) false positives. (a), (b) and (c) correspond to HDC134P (EML4-ALK), HDC140P (CCDC6-RET and RET-chr13), and HDC141P (SLC34A2-ROS1), respectively.
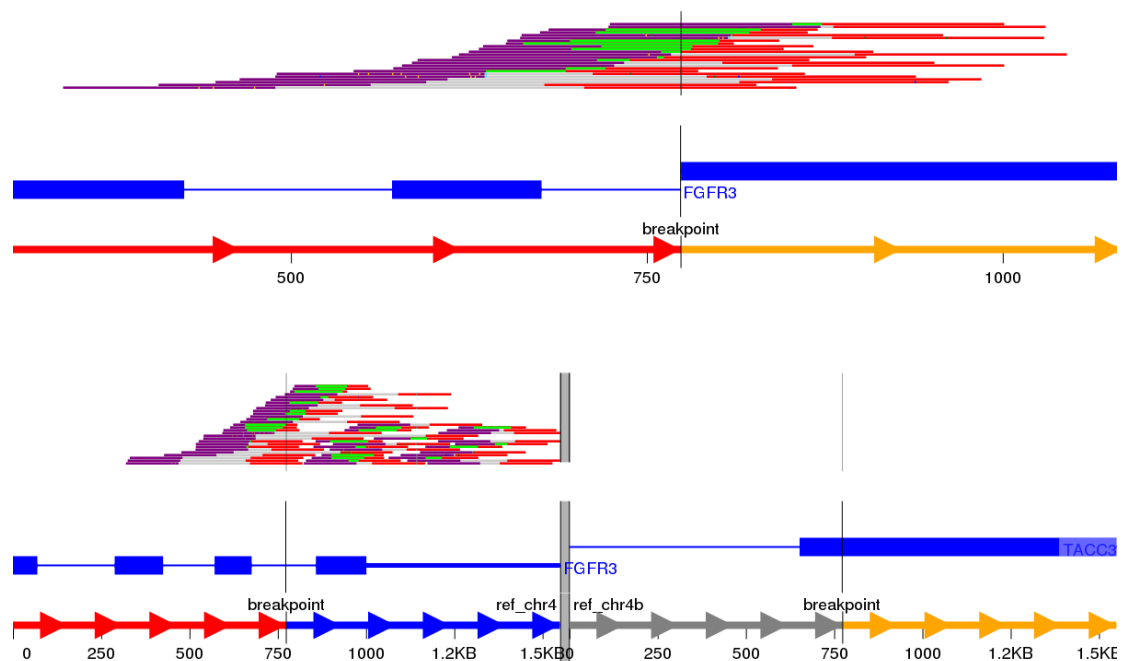
Figure 3. Svviz output for the FGFR3-TACC3 fusion (tandem duplication) in the RT4 cell line. Read evidence is shown for both how the last intron of FGFR3 is fused to an exon of TACC3 as well as for the reference alleles.
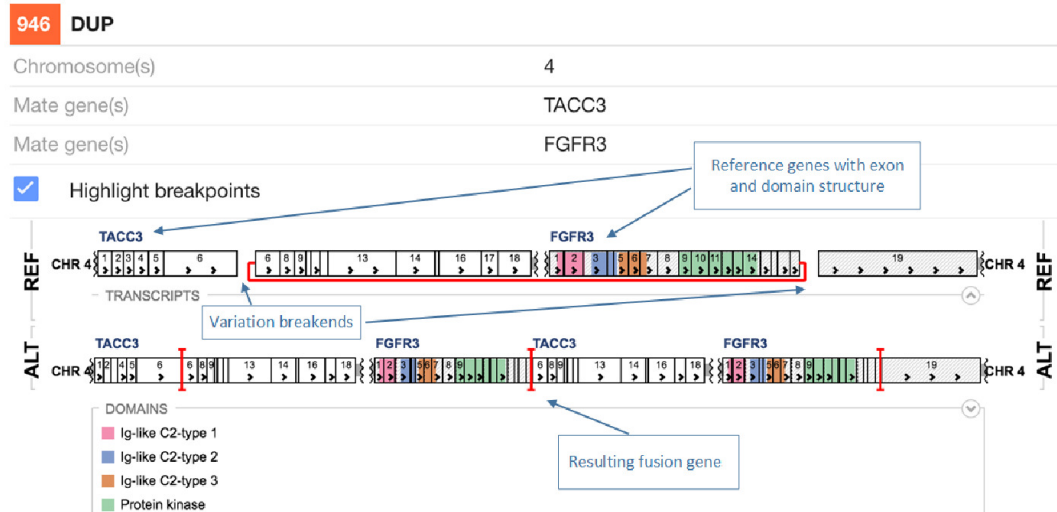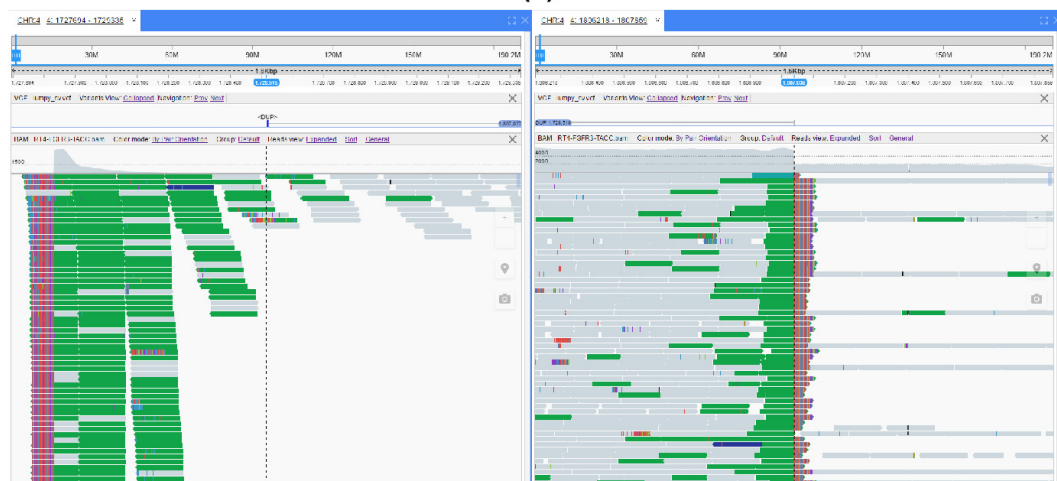
Figure 4. FGFR3-TACC3 tandem duplication fusion exon level visualisation in the New Genome Browser. Protein domains and exons affected by the structural variant are highlighted in colours. (a) shows the effect of the fusion and (b) the read evidence for the event at both breakpoints.

(a)

(b)
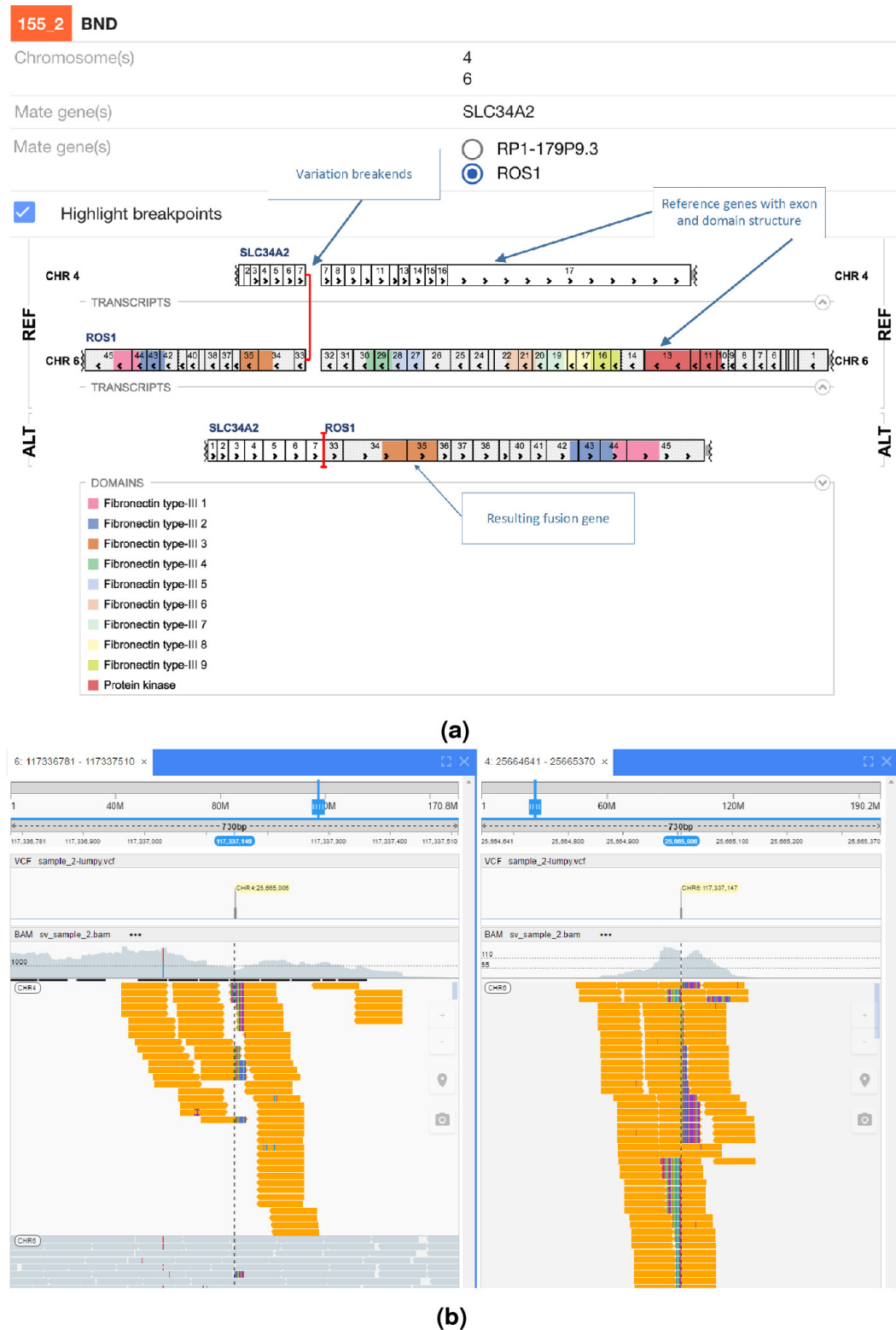
Figure 5. ROS1-SLC34A2 interchromosomal translocation fusion. (a) shows the effect of the fusion and (b) the read evidence for the event at both breakpoints.

(a)

(b)
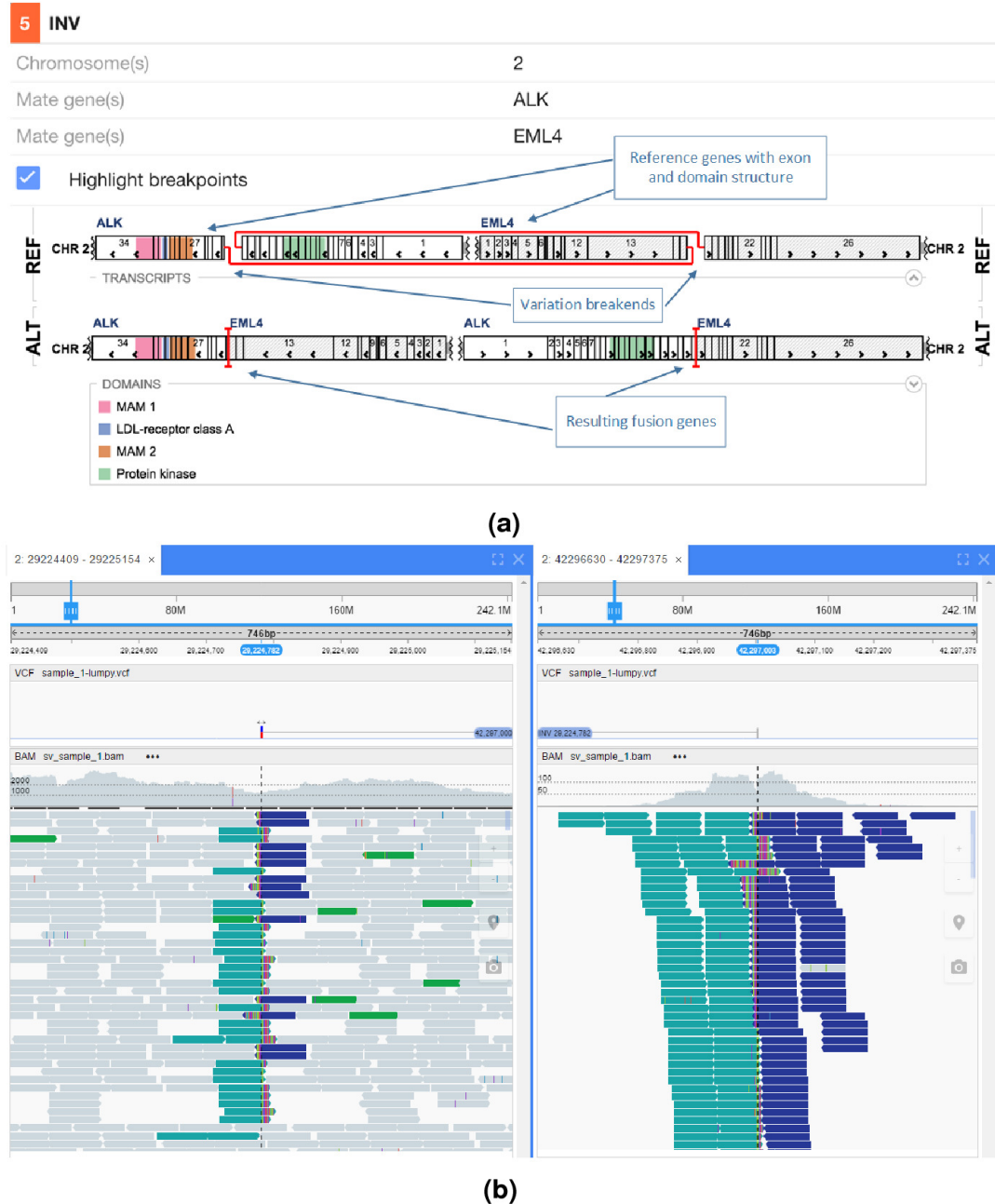
Figure 6. EML4-ALK inversion fusion. (a) shows the effect of the fusion and (b) the read evidence for the event at both breakpoints.