RESEARCH ARTICLE

Toblerone: detecting exon deletion events in cancer using RNA-seq [version 1; peer review: 2 approved]

Andrew Lonsdale[1-3], Andreas Halman[1,3], Lauren Brown[2,4,5], Hansen Kosasih[2], Paul Ekert[1-5], Alicia Oshlack[1,3,6]

[1]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, VIC, 3010, Australia
[2]Murdoch Children's Research Institute, Parkville, VIC, 3052, Australia
[3]Peter MacCallum Cancer Centre, Parkville, VIC, 3052, Australia
[4]School of Women's and Children's Health, UNSW Sydney, Sydney, NSW, 2052, Australia
[5]Children's Cancer Institute Australia, Sydney, NSW, 2052, Australia
[6]School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia

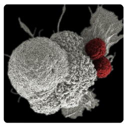
v1 **First published:** 03 Feb 2023, **12**:130
https://doi.org/10.12688/f1000research.129490.1

Latest published: 03 Feb 2023, **12**:130
https://doi.org/10.12688/f1000research.129490.1

Abstract

Cancer is driven by mutations of the genome that can result in the activation of oncogenes or repression of tumour suppressor genes. In acute lymphoblastic leukemia (ALL) focal deletions in IKAROS family zinc finger 1 (IKZF1) result in the loss of zinc-finger DNA-binding domains and a dominant negative isoform that is associated with higher rates of relapse and poorer patient outcomes. Clinically, the presence of IKZF1 deletions informs prognosis and treatment options. In this work we developed a method for detecting exon deletions in genes using RNA-seq with application to IKZF1. We developed a pipeline that first uses a custom transcriptome reference consisting of transcripts with exon deletions. Next, RNA-seq reads are mapped using a pseudoalignment algorithm to identify reads that uniquely support deletions. These are then evaluated for evidence of the deletion with respect to gene expression and other samples. We applied the algorithm, named Toblerone, to a cohort of 99 B-ALL paediatric samples including validated IKZF1 deletions. Furthermore, we developed a graphical desktop app for non-bioinformatics users that can quickly and easily identify and report deletions in IKZF1 from RNA-seq data with informative graphical outputs.

Keywords
cancer, RNA-seq



This article is included in the Oncology gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 03 Feb 2023	✓ view	✓ view

1. **Katherine Pillman**, University of South Australia, Adelaide, Australia

2. **Matt Field**, James Cook University, Cairns, Australia

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Bioinformatics gateway.

Corresponding author: Alicia Oshlack (alicia.oshlack@petermac.org)

Author roles: Lonsdale A: Conceptualization, Formal Analysis, Investigation, Software, Validation, Writing – Original Draft Preparation; **Halman A**: Software, Visualization, Writing – Review & Editing; **Brown L**: Investigation, Writing – Review & Editing; **Kosasih H**: Investigation, Validation, Writing – Review & Editing; **Ekert P**: Data Curation, Investigation, Writing – Review & Editing; **Oshlack A**: Conceptualization, Investigation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: NHMRC project grant GNT1140626

How to cite this article: Lonsdale A, Halman A, Brown L *et al.* **Toblerone: detecting exon deletion events in cancer using RNA-seq [version 1; peer review: 2 approved]** F1000Research 2023, **12**:130 https://doi.org/10.12688/f1000research.129490.1

First published: 03 Feb 2023, **12**:130 https://doi.org/10.12688/f1000research.129490.1

Introduction

B-cell precursor ALL (BCP-ALL, or B-ALL) is the most common childhood cancer. High risk subtypes of B-ALL include Ph+ (presence of BCR-ABL1 fusion) and Ph-like (similar expression profile to Ph+ without BCR-ABL1 fusion) (Roberts *et al.* 2014). These subtypes are often characterised by additional focal deletions in IKAROS family zinc finger 1 (IKZF1) gene. IKZF1 consists of 8 exons and encodes a DNA-binding protein and B-cell transcription factor, IKAROS (Mullighan *et al.* 2009; Boer *et al.* 2016). The most common IKZF1 alterations that occur in B-ALL are whole gene deletions or deletions of exons 4-7. The focal deletion results in the production of a dominant-negative isoform, IK6, lacking the N-terminal DNA-binding domains of IKAROS (Mullighan *et al.* 2009; Dörge *et al.* 2013; van der Veer *et al.* 2013). Previously, we and others have showed that IK6 could be detected using RNA-seq (Brown *et al.* 2020; Tran *et al.* 2022; Rehn *et al.* 2022). In our work, we did so by adding the deletion transcript to the reference transcriptome and measuring the transcripts per million (TPM) (Brown *et al.* 2020). Given IK6 lacks exons 4-7 (del4_7) of the canonical IKZF1 transcript, this isoform includes a novel splice junction between exon 3 and exon 8. In our cohort, we found that the splice junction indicating an exon deletion was rarely detected in the cohort in IK6-negative samples, and was increasingly expressed as the purity of the tumour increased as verified by real-time quantitative polymerase chain reaction (RQ-PCR) (Brown *et al.* 2020). Other isoforms of IKZF1 deletions in the cohort were not as reliably detected. Only one sample was known to have a deletion of exons 2-7 (del2_7) and our previous predictions of increased expression of the del2_7 transcript were unable to be validated. Previous candidates of deletions of exons 2-8 (del2_8) and exons 4-8 (del4_8) were also unable to be reliably validated. Here we develop an improved method for detecting focal deletions that can be applied more broadly to other gene deletions. Prior knowledge of which exon deletions are of interest is not required.

The key idea in our method is to generate a set of reference transcripts that correspond to all the possible exon deletions in a gene. This idea represents something of a midpoint between using annotated transcripts as the reference transcriptome and using assembly, by pre-defining potential transcripts in the class of whole exon skipping. Essentially, we generate a reference index of possible transcripts based solely on exon deletions of annotated exons, balancing the speed of a reference-based approach while allowing for a certain class of unknown events. We extract reads that support these deletions, test for sufficient coverage (expression) of a deletion, and then infer the relative abundance of deletion reads within a gene.

We name this method Toblerone and apply it to a B-ALL cohort from the Royal Children's Hospital (RCH), previously described in Brown *et al*. We validate the method on the IKZF1 gene and identify additional deletions with high relative expression of the deletion transcript. Toblerone can be applied in both a test and discovery context. Firstly, it can be used to test for known deletion events such as IK6 (exons 4-7 deletion) in IKZF1 or secondly, it can be used for discovery of new genes with exon deleted transcripts, by applying it to any gene with three or more exons. Toblerone is a tool that can quickly and accurately identify skipped exons that may indicate the presence of sub-gene deletions in a cancer sample.

Methods

Toblerone method overview

Toblerone begins by creating a custom reference. For any candidate gene, e.g., IKZF1, we take the canonical transcript and generate a new transcriptome reference that consists of the original transcript plus a set of deletion transcripts. Deletion transcripts consist of removing combinations of consecutive internal exons, excluding edge exons (first and last). This avoids the lack of splicing events at the ends of genes. A schematic of deletion transcripts for a five-exon gene is shown in Figure 1A, resulting in six deletion transcripts. These deletion transcripts along with the canonical transcript are used as the complete reference for the gene. The new custom reference is then indexed using a De Bruijn graph for pseudoalignment of RNA-seq reads.

Next reads are pseudoaligned to the custom reference and quantified at the equivalence class (EC) level based on which of the transcripts they are compatible with. Due to the mutually exclusive construction of the deletion transcripts, there is one EC that is uniquely compatible with each deletion. These correspond to the reads that have splice junctions across the deleted exons. Rather than using the transcript-compatibility count (TCC) (Ntranos *et al.* 2016) of a deletion transcript, derived from all EC that would support a deletion transcript, the Toblerone approach uses only the count from the EC uniquely supporting that deletion (Figure 1B). We refer to the EC counts for this subset of EC as the unique count (UC) of each deletion transcript. The reads for the remainder of EC, those that support the original canonical transcript for a gene, or support more than one deletion transcript, are used in the sum total of EC counts as an expression measure for the gene. We refer to this as the total EC count for a gene, or simply the total count (TC). Deletion detection requires coverage across the junction including overhang of the exon-exon boundary by at least 5 bp (by default). Any read that does not meet this threshold is not included as a UC, only in the TC. Deletion UC as a proportion of the TC is calculated and scaled by the length of the canonical transcript and the read length to account for the increased number of reads overlapping a

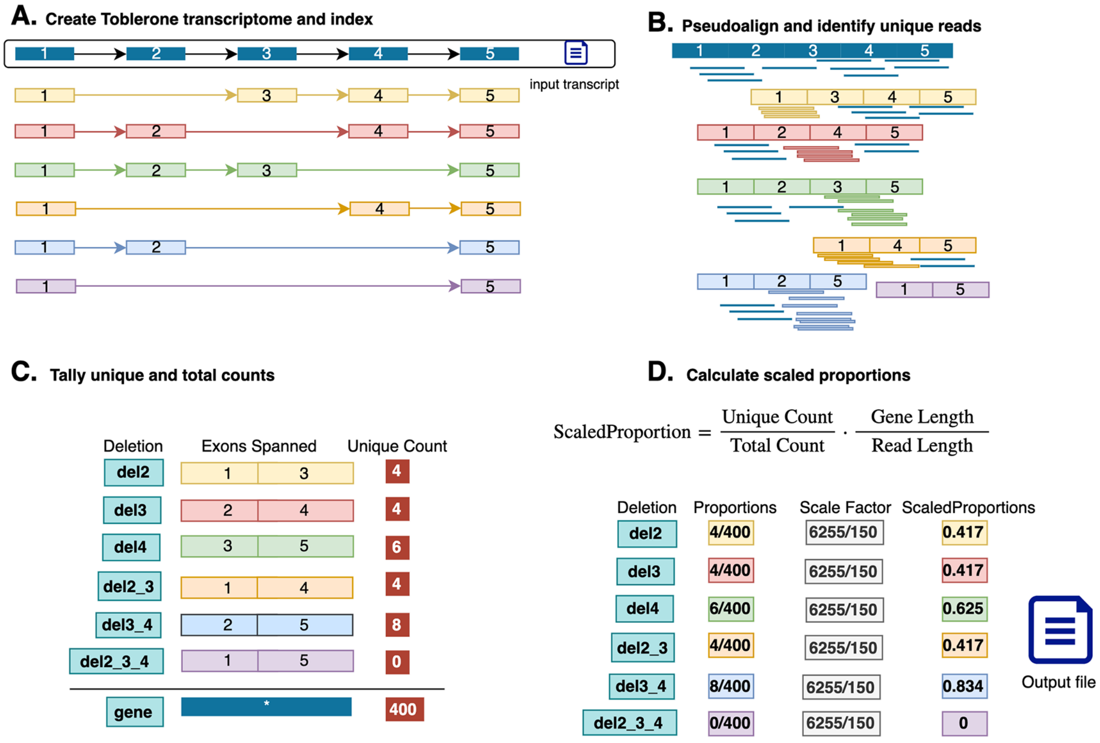

A. Create Toblerone transcriptome and index
1 2 3 4 5
input transcript
1 3 4 5
1 2 4 5
1 2 3 5
1 4 5
1 2 5
1 5
B. Pseudoalign and identify unique reads
1 2 3 4 5
1 3 4 5
1 2 4 5
1 2 3 5
1 4 5
1 2 5
1 5
C. Tally unique and total counts
Deletion
Exons Spanned
Unique Count
del2 1 3 4
del3 2 4 4
del4 3 5 6
del2_3 1 4 4
del3_4 2 5 8
del2_3_4 1 5 0
gene * 400
D. Calculate scaled proportions
ScaledProportion = Unique Count / Total Count · Gene Length / Read Length
Deletion
Proportions
Scale Factor
ScaledProportions
del2 4/400 6255/150 0.417
del3 4/400 6255/150 0.417
del4 6/400 6255/150 0.625
del2_3 4/400 6255/150 0.417
del3_4 8/400 6255/150 0.834
del2_3_4 0/400 6255/150 0
Output file

Figure 1. Overview of the Toblerone method. A) Using a canonical input transcript, deletion transcripts are generated and indexed into a custom reference. B) Reads are then pseudoaligned to the custom reference. C) reads uniquely supporting a deletion transcript are aggregated into the unique count (UC) for each transcript. All other reads are added to gene total. D) Scaled proportions of unique counts (UC) to totals counts (TC) are calculated for each deletion and saved to the output file.

splice junction with increased read length. A command line tool is available to generate and index the deletion transcriptome, and perform pseudoalignment.

Toblerone index

The Toblerone transcriptome is generated in several steps, starting from a BED12 definition of the canonical gene(s) of interest that is given to the provided “create_bedfiles.py” Python script. New BED definitions for each Toblerone deletion transcript are then created. For each BED file entry of a transcript of length N exons, the number of deletions transcripts added for a given transcript can be determined using binomial theorem. For N exons, initially N-2 is used to calculate the internal combinations of exon deletions. The method requires continued runs of exons, coincidentally forming a pattern of triangular numbers. These can also be calculated binomially and equal to the binomial coefficient of t+1 and 2 for each t triangle number. Substituting N-2 for t leads to a binomial coefficient of *N*-1 and 2 for calculating the added Toblerone deletion transcripts. The results of this script are multiple BED files, which are merged, and along with a FASTA file of the canonical transcripts corresponding to the BED12 input, passed to the bedtools (Quinlan 2014) program to create the Toblerone transcriptome FASTA file for indexing.

The core Toblerone program is an adaptation of a pseudoaligner written in the Rust programming language (2018 edition v1.31.0). This simple pseudoaligner, modified from 10XGenomics, implements concepts from several notable pseudoalignment and RNA-seq publications (Bray *et al.* 2016; Srivastava *et al.* 2016, 2019; Ntranos *et al.* 2016; Limasset *et al.* 2017; Orenstein *et al.* 2017; Li and Yan 2015). Two main modes of operation are available: index and map. Index is used to create the de Bruijn index of the Toblerone transcriptome, providing a FASTA file of a generated transcriptome and an output file to store the index.

Toblerone mapping

Map mode is used with either single or paired end reads as input, along with an index, to produce output. For optimal performance, only reads that are likely to match to genes of interest in the input transcriptome should be provided, such as by extracting reads from a prepared genome alignment. Toblerone diverges from the traditional pseudoalignment and template code in several critical ways during the mapping stage. Firstly, for any read, the number of transcripts in the equivalence class is checked; any read matching an EC unique to one transcript is treated differently due to the mutually exclusive nature of the input transcriptome. The coverage of these reads over the input transcript must be complete, excepting the number of allowed mismatches (default: 2). These reads with a unique EC (if trim is enabled, default: 5) are checked for equality between the EC and the EC of a shorter trimmed version of the read.

Secondly, uniqueness is favoured when resolving ambiguity. When comparing the equivalence classes of paired-end reads, the matching EC with fewer transcripts is preferred in such cases ensuring that the read supporting a deletion is used in further processing. Thirdly, only unique EC (UC) are tallied individually in memory against a specific transcript and written to output, with all other reads compatible with more than one transcript assigned only to a gene-level total count (TC). For each deletion, the UC and TC are written to file along with the gene length, nominated read length. The proportion of UC to TC for each deletion-gene combination is also included, as a raw value and with a scaling factor equal to the gene length divided by the read length (paired end reads weighted double). The scaled proportions are then calculated by multiplying the scale factor by the original proportion (Figure 1). These results are written to a CSV file, or standard output, as directed by the user.

Toblerone app

A graphical Toblerone app is written in Python (v3) and JavaScript programming languages, and is compatible with Linux, Mac and Windows (using Windows Subsystem for Linux) operating systems. This interface collects the IKZF1 index, core Toblerone program, RCH reference values for deletions and functions for extracting relevant reads into a convenient package. This facilitates single sample analysis *via* a genome aligned BAM file. Once downloaded, it can be easily run as a desktop app. A gene and cohort must be selected (currently only the IKZF1 gene and the RCH B-ALL cohort) along with an indexed genome aligned RNA-seq BAM file. The relevant IKZF1 reads are extracted and mapped with Toblerone to the IKZF1 deletion transcriptome. Finally, the results are displayed in an interactive format along with the high, medium and low confidence thresholds for clinical deletions, as well as outlier information for those without absolute values of reference. Further cohorts can be added to the app in future.

Ethical Considerations

Ethical approval for the cohort was obtained from the Royal Children's Hospital Human Research Ethics Committee (HREC 34127) as described previously (Brown *et al.* 2020). The selected B-ALL cohort consisted of children treated at the Royal Children's Hospital. The Children's Cancer Centre Tissue Bank undertook biobanking and informed consent for sequencing analysis.

Results

Application to paediatric acute lymphoblastic leukaemia cohort

We applied Toblerone to the IKZF1 gene from RNA-seq data from a cohort of 99 B-ALL patients (Figure 2), which included annotations for validated IKZF1 deletion status from microarray and RQ-PCR, as well as other clinical data including fusion status and karyotyping result (Brown *et al.* 2020). The scaled proportions for each of the 21 IKZF1 deletion transcripts are shown in Figure 2A, and illustrate the different frequencies of exon deletions. The majority of deletions are rarely observed in the cohort, with several notable exceptions. Deletions of exon 4 and exon 6 are commonly observed and are consistent with IKZF1 annotated alternative splicing transcripts attributed to hg38 reference transcripts (ENST00000343574, ENST00000359197, ENST00000439701). The relative expression of these transcripts is variable across samples but does not usually account for more than half of the transcripts based on the calculated proportions. For IK6, the validated del4_7 deletions have proportions that make them clearly identifiable visually, a result consistent with the previous methods applied to this cohort. In order to test for outliers in the proportion of reads indicating a deletion, we transformed the scaled proportion into a Z-score for each deletion. The distribution of Z-scores for deletion transcript with at least one sample with Z-score greater than or equal to three is shown, along with its outliers (Figure 2B). Since low gene expression can lead to high Z-score and therefore outliers with only a small number of counts, we exclude deletions that had outliers due to samples with UC counts of 10 or less. These are however included in Supplementary Figure 1. The validated del4-8 sample is also observed as an outlier of exon 4 deletions. In this case, the signal is indirect as loss of half the exons alters the proportion of splicing events of exon4 relative to the gene expression. Outlier detection identifies two other known IKZF1 deletions; the single known del2_7 deletion in the cohort and a new result of a del3_7 deletion previously only predicted from a low resolution microarray.

Although outliers can easily identify four of the six IK6 deletions, identifying the remaining two clonal, as well as a further five sub-clonal deletions presents additional challenges. In Figure 2C we show the calculated del4_7 proportions for every sample in the cohort on a log scale. We again observe the high proportion of IK6 UC as outliers and can empirically establish thresholds for high, medium and low confidence calls for the IKZF1 del4_7 deletion. We define

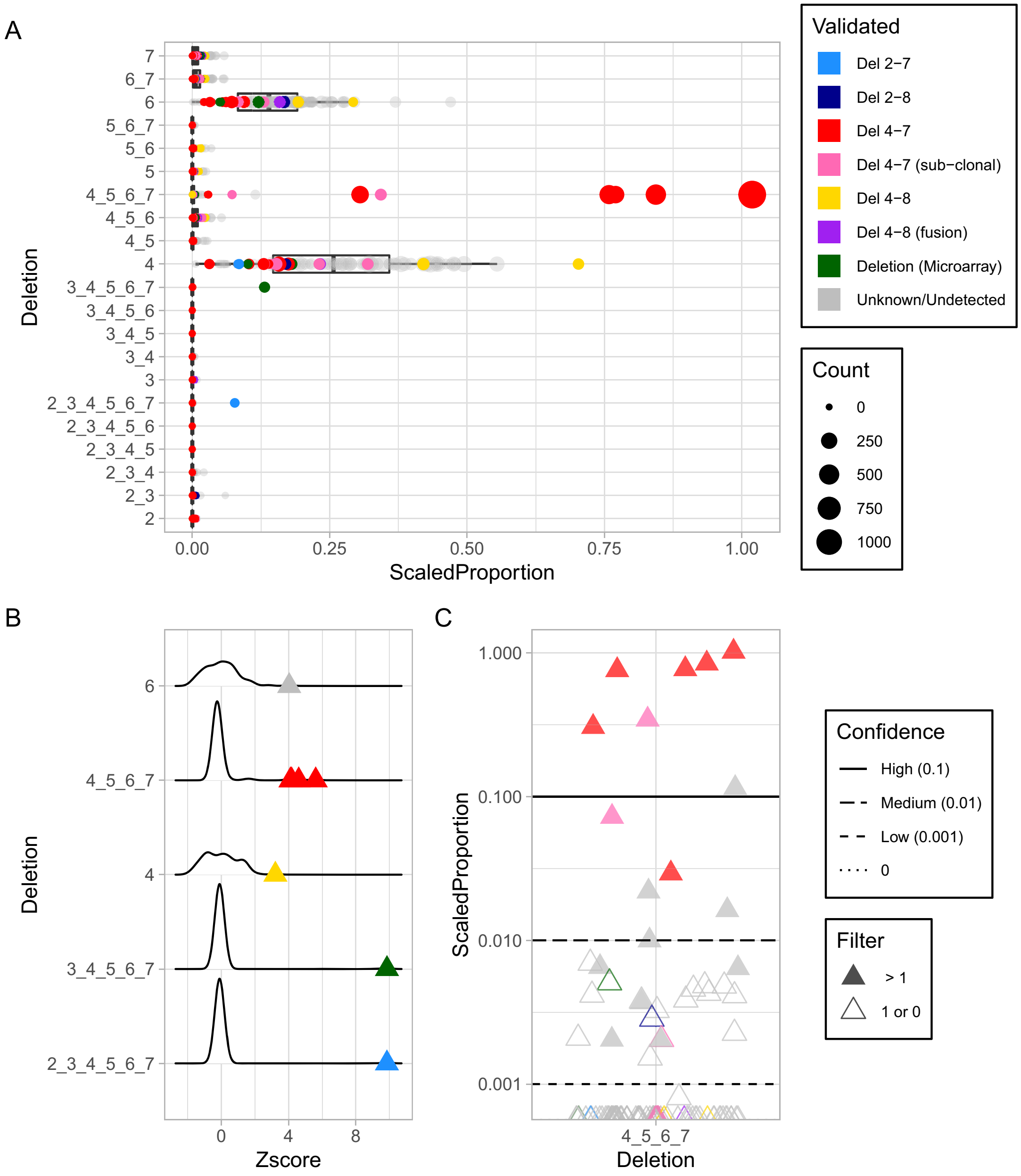


Figure 2. Application of Toblerone to RCH cohort. A) Box plots of scaled proportions of unique counts (UC) for each transcript of IKZF1 exon deletions. Each row is labelled with the exon numbers that are deleted in that transcript. Each sample is a dot with a size proportional to the UC and coloured by known IKZF1 deletion status (light blue: del2_7, dark blue: del2_8, red: del4_7, pink: del4_7 sub-clonal, yellow: del4_8, purple: del4_8 fusion, green: non-specific deletion from DNA microarray, grey: deletion unknown or undetected). B) Distributions of the scaled proportions transformed into a Z-score within each deletion. Only the IKZF1 deletion transcripts with at least one Z-score outlier ≥ 3 and minimum UC count of 10 are shown, with outliers shown as triangles coloured as in A. C) Scaled proportion scores of IKZF1 del4_7 for the cohort with high, medium and low confidence lines at 0.1 (solid), 0.01 (longdash) and 0.001 (dashed). Samples with two or more counts of the IK6 UC are filled triangles while zero or one count is empty. Many samples including sub-clonal true positives have 0 counts (dotted line).

high confidence at 0.1 scaled proportion, as most values exceeding this are known IK6 deletions; medium is set at a scaled proportion of 0.01 and a low confidence threshold of 0.001 isolates a cluster of samples that include some sub-clonal IK6 or unresolved microarray deletions as well as many unvalidated samples. The choices of these thresholds are informed by a ROC curve of validated true positives and negatives (Supplementary Figure 2). The area under the ROC of 0.859 confirms that scaled proportions effectively predict true positive del 4-7 samples, and the thresholds to identify when false-positives may be introduced. In the low confidence band especially, many samples have a single read supporting the del4_7 deletion. Low or singular counts are insufficient to identify deletions, and two validated sub-clonal IK6 samples lacked any reads supporting deletion and had counts of 0. Inclusion of the low confidence results may lead to increased false positives due to the possibility of technological or biological artefact, and so would not typically be recommended.

Effect of parameters on results

EC counts for Toblerone deletions can be heavily influenced by the alignment of reads across a junction, which is controlled by parameters set at runtime, specifically mismatch and trim. The mismatch parameter controls how many single nucleotides each k-mer can differ from the reference in order to match. Since reads at exon-exon boundaries form the signal of Toblerone, reads that overhang one exon by a small number of bases, equal to the number of allowed mismatches, may be erroneously assigned as supporting a deletion. Conversely, disallowing mismatches can reduce the signal from genuine deletion by excluding any reads that contain genomic variants or mutations. To balance the effects of allowing mismatches, Toblerone includes a trim operator which tests the robustness of read support. For each read that uniquely supports a deletion, nucleotides are removed from the ends of the read and the support checked again. Reads that are still unique and only support the deletion pass the trim test. Otherwise, these reads contribute only to the gene total. The trim operation essentially prevents soft-clipped reads contributing to the UC total, and reduces noise by removing reads with a small overhang of bases at the exon-exon boundary, ensuring that any mismatches are not the sole evidence a read supports a deletion. The trim value must equal or exceed the mismatch value. The effect of Toblerone's trim and mismatch on the counts, and the effect of the changes in these counts to the scaled proportion, are illustrated in Supplementary figure 3A and 3B, respectively.

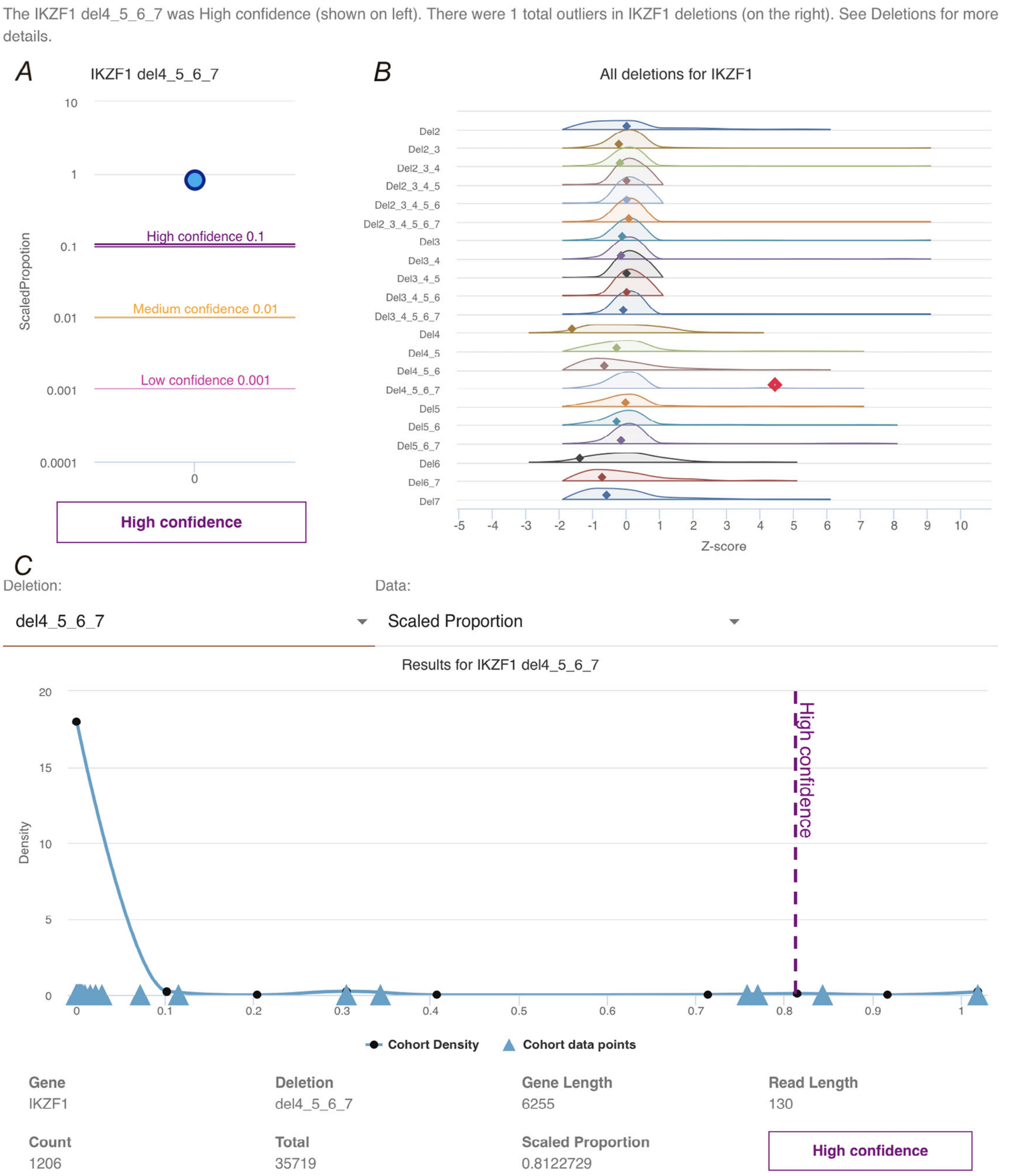
The total number of samples with any read support for IKZF1 del4_7 is shown in Supplementary Figure 3A for combinations of trim and mismatch. With default parameters, 34 samples could be detected with one or more reads. If both trim and mismatch are set to 0, this reduces to 31 samples. The number of samples with observed del4_7 deletion reads are typically reduced by higher trimming values, and increased by allowing more mismatched bases. Although the change in the number of samples is modest, the effect of these parameters is observable in the calculated scaled proportions in some samples (Supplementary Figure 3B). Though the majority of samples are relatively unchanged, notably two validated deletions have an increased proportional expression when including trimming and mismatches.

Toblerone app for single sample analysis

In addition to a command line tool, a Toblerone graphical desktop app for non-bioinformatics users was created (Figure 3A). This enhances the command line tool by allowing a user to provide either an hg19 or hg38 genome aligned BAM file. Relevant reads for IKZF1 are extracted and then pseudoaligned against the Toblerone IKZF1 transcriptome using the core Toblerone command line tool that has been integrated into the app. The results are then displayed and compared against the background cohort. The app uses the RCH cohort explored above, as a background, and the same parameters are used as defaults when comparing to new data. The high, medium and low thresholds established for this cohort are also included in the app. We demonstrate the app using a relapse sample from a patient in the RCH cohort (B_ALL16-4). The primary sample was validated as a sub-clonal IKZF1 del4_7 deletion and included in the cohort. The relapse sample was subsequently validated as a full IKZF1 deletion, though not included in the cohort data in Toblerone. The result of submitting this RNA-seq sample within the Toblerone app is shown in Figure 3. Firstly, the app indicates that the del4_7 deletion has been detected in the sample, and assigns it within the high confidence band (Figure 3A) with a scaled proportion of 0.949. Z-scores for this sample for all deletions in IKZF1 are shown in Figure 3B, and outliers are indicated in red. The app also gives details for each deletion in a separate tab, with the background deletion distribution able to be inspected. This is shown for IK6 and the relapse sample in Figure 3C, where it is clear the scaled proportion of the relapse sample is greater than any value seen in the background cohort. This plot can also be viewed as Z-scores if selected by the user.

Application to CD22 resistance in CAR T-cell therapy

While Toblerone has been developed and tested on the IKZF1 genes across a cohort of samples, it can also be used with different genes and to compare between samples in paired study design. As a proof of concept, we use data from a recent publication exploring acquired resistance when using CD22 as an immunotherapeutic target for CAR T-cell therapy (Zheng *et al.* 2022). In that work, RNA-seq data were taken before and after for B-ALL patients undergoing inotuzumab treatment. Notably for the patient with specimen identifier PAVDRV, a decrease in CD22 protein expression after treatment occurred without down regulation of the CD22 transcript. To compare Zheng's findings with a Toblerone approach, a CD22 index was created and Toblerone was applied to the samples pro- and post- treatment (no replicates). We were able to confirm the study's result that a transcript with loss of exon 2, specifically the deletion of exons 2-6, was increased after treatment with inotuzumab (Figure 4). Additionally, we ranked and visualised the Toblerone differences in deletion proportions of all transcripts in the Toblerone reference (Supplementary figure 3). This showed



The IKZF1 del4_5_6_7 was High confidence (shown on left). There were 1 total outliers in IKZF1 deletions (on the right). See Deletions for more details.

Deletion: del4_5_6_7

Data: Scaled Proportion

Gene	Deletion	Gene Length	Read Length
IKZF1	del4_5_6_7	6255	130
Count	**Total**	**Scaled Proportion**	**High confidence**
1206	35719	0.8122729	

Figure 3. Overview of Toblerone app applied to a new sample. A) Scaled proportion of the IKZF1 del4-7 for selected ample (blue dot) compared to the previously calculated confidence thresholds and B) the Z-score distributions for all IKZF1 deletion transcripts with the sample indicated dots. Any outliers are shown as red dots; C) detailed view of the del4-7 background distribution in the RCH cohort, with gaussian density estimate line in blue and each cohort value as triangles along the x-axis. The B_ALL16_4 sample is indicated by vertical dotted line, coloured by the confidence value.

that post-inotuzumab treatment, there is an increase in the expression of the transcript with an exon 12 deletion. While the original publication noted this as a known deletion observed in the B-ALL samples of the TARGET cohort, it was not attributed to patient PAVDRV as a possible cause of the loss of CD22 protein.

Discussion

Toblerone is a targeted technique for identifying internal exon deletions in specific genes from RNA-seq. Inspired by the detection of deletions in the IKZF1 gene in a study by Brown *et al.*, we have tailored an algorithm to specifically look for

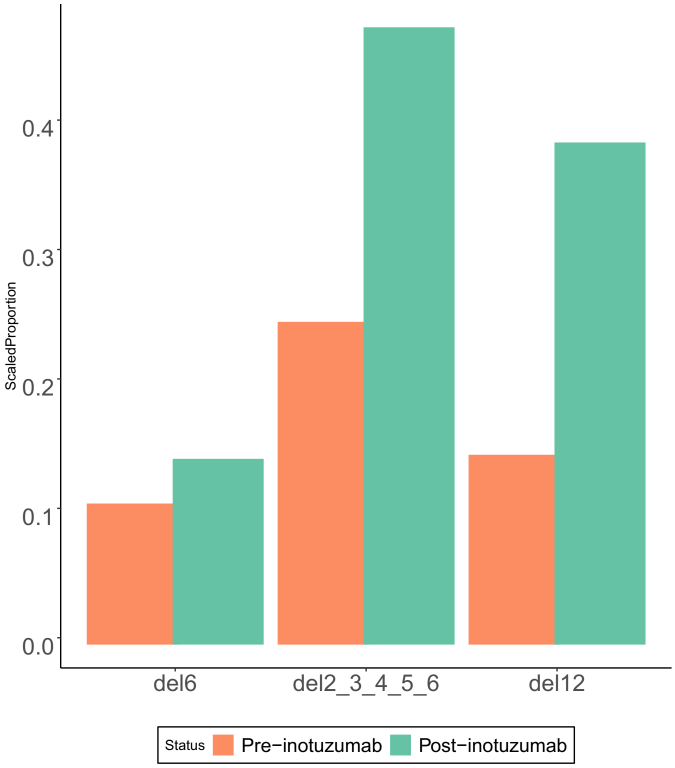


Figure 4. Differences in scaled proportions of the 3 most highly expressed CD22 deletion transcripts for patient identifier PAVDRV. Coloured bars represent the patient sample taken from peripheral blood before (orange) and after (green) treatment with inotuzumab. Selected deletions showed most deviation from pre-treatment results (Supplementary Figure 3).

these novel isoforms in RNA-seq. A recent approach that performs more comprehensive RNA-seq analysis specific to ALL is RasCALL (Rehn *et al.* 2022). Toblerone is not intended to be a comprehensive structural variant detection method however, it complements existing RNA-seq analysis tools in cancer such as fusion detection (e.g. JAFFA (Davidson, Majewski, and Oshlack 2015), Aribba (Uhrig *et al.* 2021), STAR-fusion etc) and SV detection (e.g., MINTIE (Cmero *et al.* 2020)). These and other tools may be more suitable for broader detection of unusual fusions, transcripts or more complex deletions, especially in situations where atypical events may be missed by standard of care molecular diagnostics (Nardi *et al.* 2022). Similarity, Toblerone is not intended for comprehensive analysis of alternative splicing from RNA-seq. LeafCutter (Li *et al.* 2018) and MAJIQ (Vaquero-Garcia *et al.* 2016) may be more suitable when considering these changes at a whole transcriptome level.

The Toblerone analysis has been clearly demonstrated for the detection of deletions in IKZF1 including the most commonly seen deletion involving exons 4 to 7 as well as rarer deletions with other combinations of exons. The algorithm can also be extended to discover novel exon deletions in other genes, with some caveats. Toblerone is only applicable to genes with three or more exons where the deletion does not involve the loss of the terminal exons. Loss of terminal exons could potentially be seen in differential expression analysis at the exon or gene level. If these conditions are met, then parallel indexes for additional genes of interest for a given cohort could be executed, e.g., PAX5 in B-ALL (Mullighan *et al.* 2009). Toblerone indexes can also be created for other organisms, following the same index creation process with substituted reference files and gene definitions. In IKZF1 research, mouse models of Ikzf1 can be used to understand its role in immunity and diseases related to Ikaros (Boast *et al.* 2021). Example indexes for PAX5 (hg38) and Ikzf1 (mm10) are available with the Toblerone source code.

When a deletion transcript is lowly expressed, it is difficult to detect with RNA-seq. Toblerone uses junction read counts to infer the novel junctions produced by a deletion. Even with checks to ensure that reads supporting deletions are not due to minor overlaps at exon boundaries, singular reads for deletions were found in many samples without validated deletions. Conversely, we were not able to detect any supporting reads in several subclonal del4_7 deletions. Determining whether these low counts are a biological or technical artefact remains an open question. With this in mind, we proposed thresholds for high, low and medium confidence deletions to provide a guide to interpreting Toblerone results, as well as establishing the level of background noise for the method. Extending to other genes and cohorts may require validated samples to establish appropriate thresholds between validated samples and remainder of the cohort samples. However, as shown with acquired resistance with CD22 case study, albeit without sufficient replicates for statistical analysis, Toblerone can also be used in case-control studies to compare exon deletions proportions between two groups.

The generation of the Toblerone transcriptome and pseudoalignment is targeted towards single gene analysis and therefore, is not optimised for a complete transcriptome or competitive pseudoalignment between multiple genes. Results presented here have used extracted reads from genome alignments, and as such, the reads have already been aligned by more conventional approaches such as STAR (Dobin *et al.* 2013). Conceptually, it is possible to use raw reads and implement the Toblerone approach as part of a competitive pseudoalignment, for single or multiple genes of interest. The Toblerone custom deletion transcriptome can be used with any pseudoaligner that divulges EC counts prior to abundance estimation. So conceptually, a complete pseudoalignment from raw FASTQ data can be performed in programs such as Kallisto (Bray *et al.* 2016) and Salmon (Patro *et al.* 2017) with a mixture of reference transcripts and additional deletions transcripts. This approach however stores EC counts compatible with more than one transcript, which is superfluous for the needs of the Toblerone algorithm, and this additional overhead increases with each gene considered and prevents efficient computation across multiple samples. The Toblerone pseudoaligner is designed to only compute the necessary counts, and is optimised to identify reads that support deletions. Future work on Toblerone however, can address these limitations for multi-gene analysis in a competitive alignment, through improvements to the Toblerone pseudoaligner or by reprocessing conventional pseudoaligner results to reduce redundant counts, and removing reads that do not robustly support a deletion.

The fundamental idea of Toblerone is to modify the transcriptome reference to contain transcripts with all variations for a sequence event of interest. By generating a transcriptome that consists of transcripts with deleted exons, reads that uniquely support that transcript can be identified, inspected and compared. Here, we have modified the reference to represent exon deletions, but there are other possible variants which may be detected. For example, intron retention is a natural extension of the existing algorithm, and exon duplications or inversions could also be identified. Toblerone adds insight into the clinically relevant transcriptional consequences of mutations by extending the analysis of RNA-seq, which is being applied more generally in many malignancies, notably lymphoblastic leukaemia. It can quantify the relative expression of known clinically relevant exon deletions in cancer, and aid in the discovery of new ones.

Code availability

The original template Rust pseudoaligner from 10XGenomics is on GitHub: https://github.com/10XGenomics/rust-pseudoaligner. Source code for the Toblerone (v0.0.9 DOI: 10.5281/zenodo.7563716) adaption is available at: https://github.com/oshlack/toblerone along with instructions for compilation and usage.

Source code for the Toblerone App (v1.05 DOI: 10.5281/zenodo.7563747) is available at: https://github.com/Oshlack/TobleroneApp and binaries can be downloaded directly from http://oshlacklab.com/TobleroneApp/. Software is available under the MIT license.

Data availability

Underlying data

Unprocessed RNA-seq data for 99 primary RCH cohort samples and selected relapse samples is available from the European Genome-Phenome Archive (Freeberg *et al.* 2022) (accession number EGAS00001004212). Raw RNA-seq data from the published CD22 resistance patients is publicly available in the Short Read Archive, BioProject ID PRJNA764243.

Extended data

Zenodo: Extended data and images for “Toblerone: detecting exon deletion events in cancer using RNA-seq” (https://doi.org/10.5281/zenodo.7574657) (Lonsdale *et al.* 2023).

This repository contains the following extended data:

- Supplementary Figure 1.
- Supplementary Figure 2.
- Supplementary Figure 3.
- Supplementary Figure 4.

- Processed IKZF1 data for the RCH cohort by Toblerone.

- STROBE reporting guidelines for the RCH cohort re-analysis.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license.

Acknowledgements

Tumour samples were made available by the Children's Cancer Centre Tissue Bank at the Murdoch Children's Research Institute and The Royal Children's Hospital (www.mcri.edu.au/childrenscancercentretissuebank), made possible through generous support by CIKA (Cancer In Kids @ RCH), The Royal Children's Hospital Foundation and the Murdoch Children's Research Institute.

References

Boast B, de Jesus C, Nunes-Santos HS, *et al.*: **Ikaros-Associated Diseases: From Mice to Humans and Back Again.** *Front. Pediatr.* 2021; **9**(July): 705497.
PubMed Abstract | Publisher Full Text | Free Full Text

Boer JM, van der Veer A, Rizopoulos D, *et al.*: **Prognostic Value of Rare IKZF1 Deletion in Childhood B-Cell Precursor Acute Lymphoblastic Leukemia: An International Collaborative Study.** *Leukemia.* 2016; **30**(1): 32–38.
PubMed Abstract | Publisher Full Text

Bray NL, Pimentel H, Melsted P, *et al.*: **Near-Optimal Probabilistic RNA-Seq Quantification.** *Nat. Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text

Brown LM, Lonsdale A, Zhu A, *et al.*: **The Application of RNA Sequencing for the Diagnosis and Genomic Classification of Pediatric Acute Lymphoblastic Leukemia.** *Blood Adv.* 2020; **4**(5): 930–942.
PubMed Abstract | Publisher Full Text | Free Full Text

Cmero M, Schmidt B, Majewski IJ, *et al.*: **MINTIE: Identifying Novel Structural and Splice Variants in Transcriptomes Using RNA-Seq Data.** *Cold Spring Harbor Laboratory.* 2020.
Publisher Full Text

Davidson NM, Majewski IJ, Oshlack A: **JAFFA: High Sensitivity Transcriptome-Focused Fusion Gene Detection.** *Genome Med.* 2015; **7**(1): 43.
PubMed Abstract | Publisher Full Text | Free Full Text

Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: Ultrafast Universal RNA-Seq Aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

Dörge P, Meissner B, Zimmermann M, *et al.*: **IKZF1 Deletion Is an Independent Predictor of Outcome in Pediatric Acute Lymphoblastic Leukemia Treated according to the ALL-BFM 2000 Protocol.** *Haematologica.* 2013; **98**(3): 428–432.
PubMed Abstract | Publisher Full Text | Free Full Text

Freeberg MA, Fromont LA, D'Altri T, *et al.*: **The European Genome-Phenome Archive in 2021.** *Nucleic Acids Res.* 2022; **50**(D1): D980–D987.
PubMed Abstract | Publisher Full Text | Free Full Text

Limasset A, Rizk G, Chikhi R, *et al.*: **Fast and Scalable Minimal Perfect Hashing for Massive Key Sets.** *arXiv [cs. DS]. arXiv.* 2017.
Reference Source

Li YI, Knowles DA, Humphrey J, *et al.*: **Annotation-Free Quantification of RNA Splicing Using LeafCutter.** *Nat. Genet.* 2018; **50**(1): 151–158.
PubMed Abstract | Publisher Full Text | Free Full Text

Li Y, Yan X: **MSPKmerCounter: A Fast and Memory Efficient Approach for K-Mer Counting.** *arXiv [q-bio.GN]. arXiv.* 2015.
Reference Source

Lonsdale A, Halman A, Brown LM, *et al.*: **Toblerone: detecting exon deletion events in cancer using RNA-seq.** 2023.
Publisher Full Text

Mullighan CG, Xiaoping S, Zhang J, *et al.*: **Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia.** *N. Engl. J. Med.* 2009; **360**(5): 470–480.
PubMed Abstract | Publisher Full Text | Free Full Text

Nardi V, McAfee SL, Dal Cin P, *et al.*: **Chemotherapy Resistance in B-ALL with Cryptic NUP214-ABL1 Is Amenable to Kinase Inhibition and Immunotherapy.** *Oncologist.* 2022; **27**(2): 82–86.
PubMed Abstract | Publisher Full Text | Free Full Text

Ntranos V, Kamath GM, Zhang JM, *et al.*: **Fast and Accurate Single-Cell RNA-Seq Analysis by Clustering of Transcript-Compatibility Counts.** *Genome Biol.* 2016; **17**(1): 112.
PubMed Abstract | Publisher Full Text | Free Full Text

Orenstein Y, Pellow D, Marçais G, *et al.*: **Designing Small Universal K-Mer Hitting Sets for Improved Analysis of High-Throughput Sequencing.** *PLoS Comput. Biol.* 2017; **13**(10): e1005777.
PubMed Abstract | Publisher Full Text | Free Full Text

Patro R, Duggal G, Love MI, *et al.*: **Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.** *Nat. Methods.* 2017; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text

Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Current Protocols in Bioinformatics/Editoral Board, Andreas D. Baxevanis ... [et al.].* 2014; **47** (September): 11.12.1–34.

Rehn J, Mayoh C, Heatley SL, *et al.*: **Rascall: Rapid (Ra) Screening (Sc) of RNA-Seq Data for Prognostically Significant Genomic Alterations in Acute Lymphoblastic Leukaemia (ALL).** *PLoS Genet.* 2022; **18**(10): e1010300.
PubMed Abstract | Publisher Full Text | Free Full Text

Roberts KG, Pei D, Campana D, *et al.*: **Outcomes of Children with BCR-ABL1–like Acute Lymphoblastic Leukemia Treated with Risk-Directed Therapy Based on the Levels of Minimal Residual Disease.** *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 2014; **32**(27): 3012–3020.
PubMed Abstract | Publisher Full Text | Free Full Text

Srivastava A, Malik L, Smith T, *et al.*: **Alevin Efficiently Estimates Accurate Gene Abundances from dscRNA-Seq Data.** *Genome Biol.* 2019; **20**(1): 65.
PubMed Abstract | Publisher Full Text | Free Full Text

Srivastava A, Sarkar H, Gupta N, *et al.*: **RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-Seq Reads to Transcriptomes.** *Bioinformatics.* 2016; **32**(12): i192–i200.
PubMed Abstract | Publisher Full Text | Free Full Text

Tran TH, Langlois S, Meloche C, *et al.*: **Whole-Transcriptome Analysis in Acute Lymphoblastic Leukemia: A Report from the DFCI ALL Consortium Protocol 16-001.** *Blood Adv.* 2022; **6**(4): 1329–1341.
PubMed Abstract | Free Full Text

Uhrig S, Ellermann J, Walther T, *et al.* **Accurate and Efficient Detection of Gene Fusions from RNA Sequencing Data.** *Genome Res.* 2021 January; **31**: 448–460
PubMed Abstract | Publisher Full Text | Free Full Text

Vaquero-Garcia J, Barrera A, Gazzara MR, *et al.*: **A New View of Transcriptome Complexity and Regulation through the Lens of Local Splicing Variations.** *elife.* 2016; **5**(February): e11752.
PubMed Abstract | Publisher Full Text | Free Full Text

van der Veer A, Waanders E, Pieters R, *et al.*: **Independent Prognostic Value of BCR-ABL1-like Signature and IKZF1 Deletion, but Not High CRLF2 Expression, in Children with B-Cell Precursor ALL.** *Blood.* 2013; **122**(15): 2622–2629.
PubMed Abstract | Publisher Full Text | Free Full Text

Zheng S, Gillespie E, Naqvi AS, *et al.*: **Modulation of CD22 Protein Expression in Childhood Leukemia by Pervasive Splicing Aberrations: Implications for CD22-Directed Immunotherapies.** *Blood Cancer Discovery.* 2022; **3**: 103–115.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 07 September 2023

https://doi.org/10.5256/f1000research.142177.r192103

✓

Matt Field
James Cook University, Cairns, Australia

General comments:

Here Lonsdale et al describe Toblerone, software designed to detect exon deletions in known genes from RNA-Seq. Overall this tool is useful for targeted applications and would likely be of interest to the broader research/clinical community. The manuscript is well written and encouragingly all relevant code and supporting documentation is available enabling other research/clinicians to utilise the framework. My comments below are relevant to the specific sections:

Methods: "Overview":

- Define equivalence class.
- While the total counts is a useful measure in some contexts, did you think about comparing/reporting the read count supporting the deletion to only the reads that definitively do not support the deletion? For example in a deletion skipping exon 2, supporting reads would span exon 1-3 while non-support reads would be the sum of exons spanning boundaries 1-2 and 2-3.

Methods: Index

- In some instances, non-canonical forms can be important. Is it possible to input multiple isoforms for the same gene for indexing?
- Is there any handling for aberrant read alignments that don't clearly skip canonical exons? For example supporting reads for a deletions that span a canonical exon boundary and a non-canonical exon boundary might be important to the patient treatment.

Methods: "app"

- The app is a great idea and should be useful in targeted cases. Are there any plans to allow users to request additional applications for other relevant genes? It would be great to see a bit more documentation in the app README on github.

Results: ALL

- Just to clarify, the exon 4 and 6 deletions are not related to ALL? Might be useful to display the most common non-ALL and ALL-specific isoforms.

Results: Parameters effect

- Do deletions being in frame and out of frame have any impact on the algorithm?

- What, if any impact does the length of each alignment segment have on the call? For example, is there any difference between one read that has 140 bases aligned to one exon and only 10 bases aligned across the skipped exon whereas another read has 75 bases aligned to each. Similarly is the mapping score and the uniqueness of each alignment segment factored in?

Results: CD22

- Nice example of another application. While the paper largely describes the application for IKZF1, it's uptake will largely be determined by how generalisable the framework is. For example, determining the cutoffs in Fig 2C and the placement of the new sample relative to others in Fig 3 is specific to IKZF1 but it would be useful to describe what specifically is required for any new proposed use case.

- Do you suspect each application (i.e. gene) will require custom cutoffs/filters?

Discussion:

- Do you suspect each application (i.e. gene) will require custom cutoffs/filters?

- A specific application would be identifying somatic events in paired tumour / normal samples. Is there any thought to further developments/optimisations around this common use case?

- I could imagine scenarios where it is desirable to interrogate a small panel of commonly mutated genes specific to type of cancer (similar to targeted gene panel sequencing). Are there are limitations that might present challenges to this level of scaling?

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 25 April 2023

https://doi.org/10.5256/f1000research.142177.r163943

✓

Katherine Pillman

Centre for Cancer Biology, an Alliance between SA Pathology, University of South Australia, Adelaide, Australia

Article Summary
The article describes ‘Toblerone’, an improved targeted method of identifying and quantifying focal deletions of whole internal exons in one or a few genes. The method works without a priori knowledge of which exons will be deleted and was developed to improve the accuracy and sensitivity of detection of exonic deletions for a clinically important gene, IKZF1.

Overall, I think this is a clearly written and well-explained approach that describes the design and implementation of a solution to the clinically important problem identified in a previous study by the group (Brown et al. 2007). The provided app provides an easy visual way of assessing the results and could also be used as a template when expanding the method to other genes. I think that the approach of creating individual transcripts defining each potential set of exonic deletions is a good one; it deals with some tricky edge cases for mapping reads across splice junctions in a reasonable and transparent way. For example, other solutions wherein one maps reads to only the two junction-flanking exons can raise problems when one of these exons is much shorter than the read length.

Major comments
None.

Minor comments
Including a brief explanation of any of the following terms would make the paper more widely accessible:

- Focal deletion
- Pseudoalignment
- Equivalence Class Level

In the calculation of Scaled Proportion, it seems to me that the trim value should be part of the equation as it affects the proportion of reads that would be expected to be detected over a junction for a given sequencing depth. Might the solution be to subtract the trim length (=5 by default) from the read length prior to calculation, to account for the loss of the use of the last 5 bases? Would it be correct to think that the length of the part of the read available for counting is what the data should be normalised by?

I could not find figure legends for the Supp Figures, can these please be explicitly described somewhere.

If it is possible, a simple explanation of the meaning of the Scaled Proportion metric would be a great help to the user intuitively understanding their results. Is it “the fraction of reads across a deletion junction relative to what would be expected for a given expression level and read length”? Meaning it is an estimate of the true fraction of transcripts present in Sample X which contain that deletion, with a theoretical maximum of 1.0 (though only in a perfect world where read coverage is not stochastic or biased across a transcript)?

On the last line of Page 7, "Supplementary figure 3" should be changed to "Supplementary figure 4".

In Fig3C, I found it confusing that the purple dashed line represents the single sample position, whereas similar lines in other figures (purple lines in Fig3A, dashed lines in Fig2C) were used to describe the thresholds. I think it would help if this line can replaced with a different coloured symbol and/or changed to a new colour.

I also thought that if a genome-mapped bam was used as the first step, it is important that the mapper used allows soft clipping (like STAR in default mode). A mapper that does not soft clip might discard reads which had a smallish number of bases mapping across a non-canonical junction, which would reduce the sensitivity of Toblerone. This could be worth noting in the text.

A suggestion: a great use of Toblerone would be in conjunction with an in-clinic capture panel, with probes tiled on the gene specifically so as to avoid capture bias over exon-exon junctions. The huge sequencing depth one gets from capture panels would solve the issues described in determining whether the samples with 0 counts lack the genomic fusion or simply have low sequencing coverage.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptome bioinformatics, gene regulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.