Supplementary Material: TriageTools: tools for partitioning and prioritizing analysis of high-throughput sequencing data

Danai Fimereli[1], Vincent Detours[1,2], and Tomasz Konopka[1]

[1] *IRIBHM, Université Libre de Bruxelles, Brussels, Belgium*
[2] *Welbio, Université Libre de Bruxelles, Brussels, Belgium*

In this supplementary material, we elaborate on some topics related to extraction of reads from high-throughput sequencing data. In section A, we present a case study where the target region is non-coding. In section B, we compare our extraction method to an alternative approach using a traditional aligner (Bowtie [1]) and custom indexes.

A Selection of non-coding, intronic, and inter-genic regions

As discussed in the main text, extraction of reads from non-coding target regions should be expected to be less specific than from exonic regions. We here illustrate this effect using a target region of around 20kb on chromosome six (hg19 genome, chr6:139,777,423-139,797,012). This regions contains the non-coding gene LOC645434 with three introns and more than 10kb of intergenic sequence. In the full alignment of our RNA-seq test sample, the region is covered relatively weakly (216 unique read-ids), with some reads having mapping quality equal to zero.

We attempted to extract reads on this target region using the triage algorithm following the same protocol as described in the main text. Generally, the classification performance corresponds to our expectations (Figure S1). Given that relatively few reads actually originate from the target region, misclassification of a read changes the apparent true positive rate (TPR) more strongly than in our other examples. What is more, the TPR seems to be non-monotonic with the hits threshold. This is a result of the low-mappability of the region and the fact that Bowtie/Tophat utilize some randomization in such situations (this effect was also observed in the other test cases, but is more clearly visible here due to smaller overall number of reads; c.f. Discussion in main text). Also, because of the low mappability, the triage procedure picks up many reads that have similar sequence but actually map elsewhere in the genome. The false positive rate (FPR) is higher than in the single-gene extraction example even though the target regions are of similar size.

B Alternative approach to read selection using custom Bowtie indexes

Instead of using our custom algorithm for targeted extraction, one could ask if a similar result could be obtained using traditional aligners in non-canonical ways. In particular, one could try the following approach: create a custom index for a traditional aligner consisting only of the target sequence, align all reads onto this custom genome, identify reads that are successfully aligned and ignore those that are not. The output of this procedure would be comparable in form to the output from the triage procedure. Both outputs would then be re-aligned onto the full reference genome to obtain proper coordinates and mapping scores. We here investigate this alternative approach and show that it performs similarly to the triage method in some situations, but also fails in others, particularly with RNA-seq samples.

We studied the same samples described in Table 1 in the main text: a paired-end (2x50bp) RNA-seq sample, a single-end (1x75bp) RNA-seq sample, a synthetic exome (2x100bp), and a synthetic whole genome (2x100bp). We attempted to select for a single gene (NOTCH1) as well as for a set of cancer genes (cancer). We extracted sequences of these target regions together with their flanking areas. We then concatenated the sequences and created custom Bowtie2 indexes.

For the exome and whole genome samples, we aligned the raw data onto the custom reference sequence and then extracted the aligned reads using a combination of samtools [2], a custom script, and a tool from the triagetools package (tools like Picard's SamToFastq crashed on several samples producing truncated output; we thus needed a different approach). After selection, we re-aligned the selected reads

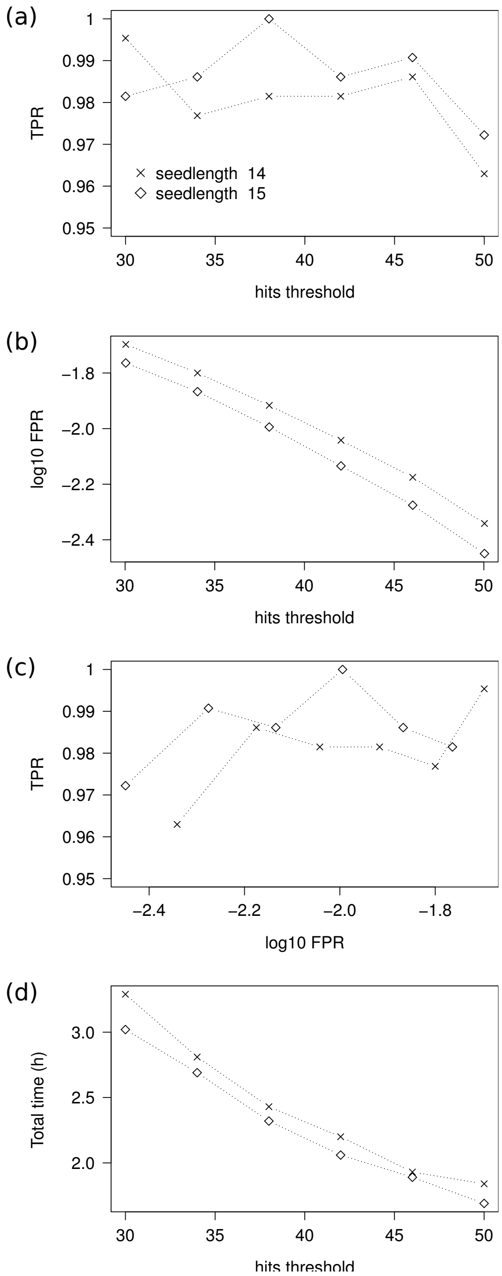

Figure S1: Analogous to figures in main text, but note different scales on some horizontal and vertical axes. Here, extraction was carried out a non-coding, intronic, and intergenic region of length around 20kb (chr6:139,777,423-139,797,012). (a) True positive rate of classification decreases with the hits parameter. Movement in TPR arise as a result of non-deterministic alignment by Bowtie/Tophat. (b) Proportion of reads that are actually off-target but pass through the classification procedure. As expected, this false positive rate decreases with s and H and is higher that for selection of exonic regions. (c) A ROC-style representation of the panels (a) and (b). To make the points distinguishable, the vertical scale only shows a small fraction of the full TPR range and the horizontal scale is logarithmic. The series of points in each seedlength group represent different hits thresholds. (d) Running time of classification and mapping. For comparison, the running for the full alignment was about 42 hours.

onto the full hg19 genome. For the RNA-seq samples, we tried two distinct approaches. In the first, we performed selection using Bowtie2 ignoring the fact that the sample originated from transcripts and not from raw DNA sequence. In the second, we ran selection using Tophat2, thus allowing the software to handle reads containing splicing events. We then compared the results with selection carried out by the triage protocol described in the main text (it is the same for DNA- and RNA-seq).

For all test cases, we measured the total running time and the number of reads in the final alignment falling in the target region. In the non-triage approaches, our custom script to identify aligned reads had nontrivial running time and so we subtracted the time for this step from the total. We performed all tests providing access to a single core as well as four processing cores. In the latter case, all invocations of Bowtie2 and Tophat2 were given access to all available cores; the triage method always executed on a single core.

While running experiments using RNA-seq samples, the extraction step using Tophat2 crashed due to memory limitations. The software attempted to keep reads in memory if they were not mapped onto the custom genome in a first pass in view of re-processing them differently in a second pass. As the majority of reads in the full sample were not mappable to the custom reference, this step became prohibitively expensive and ultimately led to out-of-memory failure. In what follows, we thus present results only for extraction using Bowtie2.

Speedups of the successful methods over full alignment are summarized in Figure S2. For small target regions (NOTCH1), the triage procedure consistently gave better performance, especially when using a single processor core. When multiple cores were available, the margin decreased because the aligner-extraction was allowed to take advantage of the additional resources while the triage approach was always ran on a single core. Despite this handicap, the triage method produced comparable or slightly better performance (measurements based on a single run, performance may change slightly from run to run, machine used, etc.). In any case, the speedup factors were typically higher when running in single-processor mode.

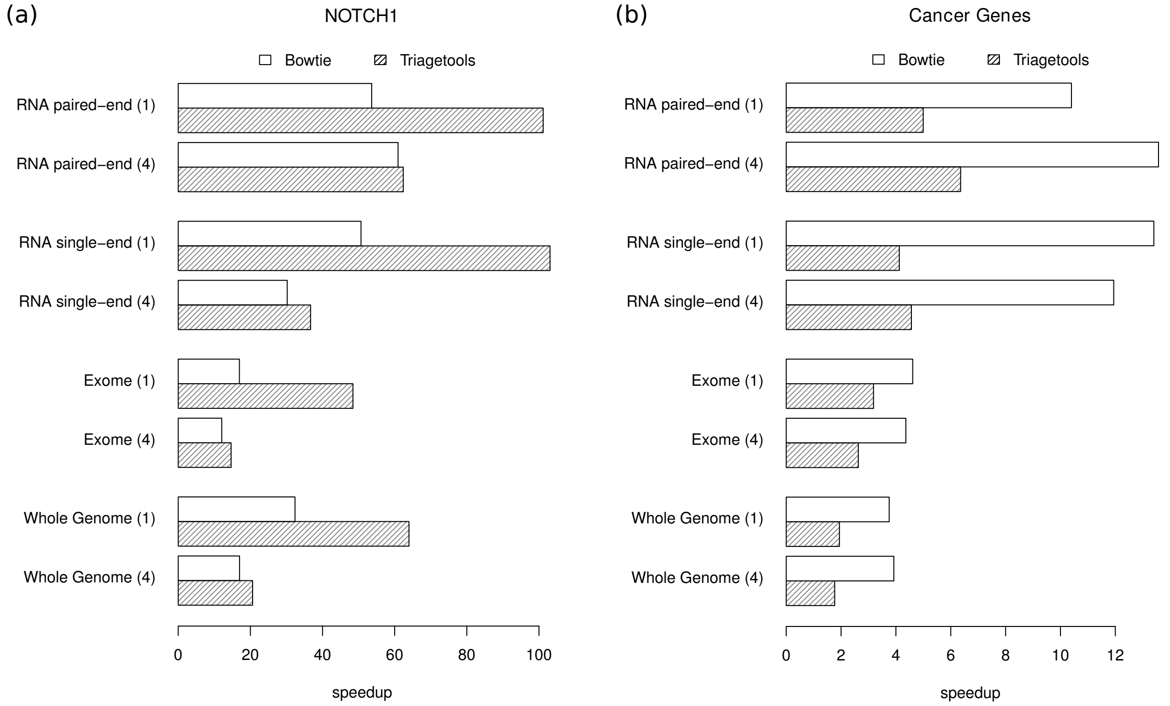

Figure S2: Speedup factors of two extraction/alignment methods compared to full alignment. Unfilled bars indicate results for extraction using Bowtie with custom index followed by alignment with either Bowtie (DNA-seq) or Tophat (RNA-seq). Filled bars indicate results for extraction carried out with triagetools followed by alignment with either Bowtie or Tophat. Numbers in parentheses in sample names indicate number of cores used for extraction and alignment; speedups in each case are computed by dividing the running time of full alignment using the corresponding number of cores by the running time of the selection plus alignment pipeline. (a) Results using a single gene as the target region. (b) Results using a large set of cancer genes as the target region.

For large target regions (cancer genes), the triage method was less effective in terms of speedup. One reason for this was that the triage procedure produced higher false positive rates (data not shown), which led to longer alignment times and hence lower speedups. This opens the question of further performance tuning of the triage method and also reaffirms the message that it is designed for small target regions.

Apart from speedup, an important issue in targeted extraction is quality of the output. Generally, both the triage and the aligner-based method gave very high true positive rates for the exome and whole genome samples; each approach gave marginally better results than the other in at least one case (data not shown). However, the two approaches gave significantly different results for the RNA-seq samples (Figure S3). In the single-end sample, the triage procedure was able to identify and extract reads that eventually aligned across splice junction, while the traditional aligner-based approach mistakenly eliminated the majority of such reads. In the paired-end sample, this effect was less pronounced, but all evidence for one splice event was completely missing in the NOTCH1 gene and many others were also missing in genes from the cancer set. Thus, for the RNA-seq samples, the high speedups obtained via aligner-based extraction were in part a result of a sacrifice of ability to perform downstream analyses.

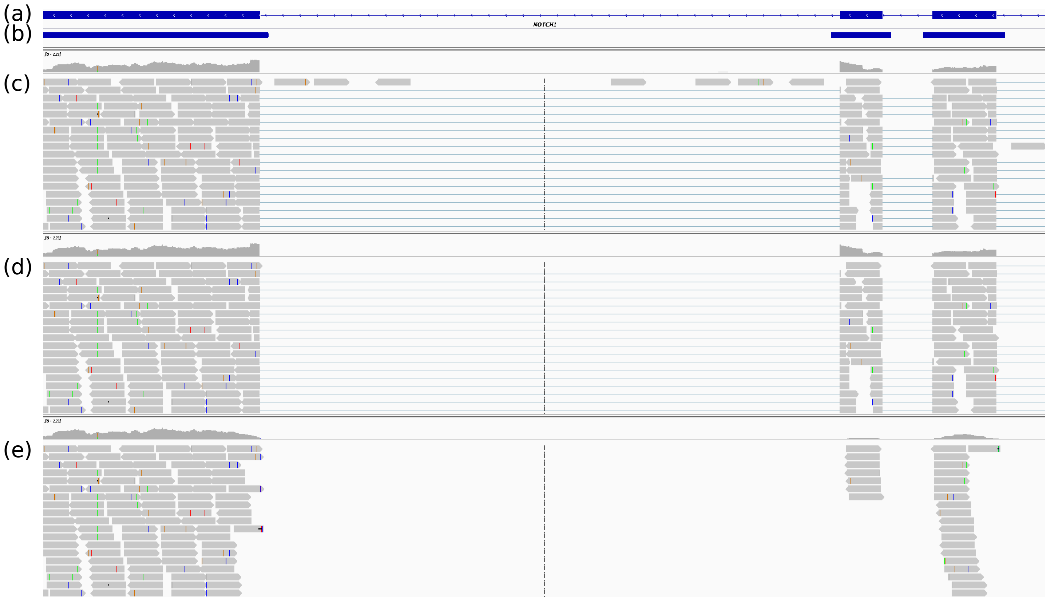

Figure S3: Read selection strategies can have significant impact on final alignment in the target region, visualized using the Integrative Genomics Viewer [3]. (a) The gene annotation track shows exonic and intronic regions for a gene of interest (NOTCH1). (b) Target region for read extraction consists of exonic and 20bp flanking regions. (c) Full alignment of a single-end RNA-seq sample with 75bp reads shows most coverage lies in exonic regions. Some reads are spliced and align across neighboring exons. (d) Alignment obtained from output of the triage read extraction procedure shows the same coverage patterns on the target region as the full alignment. Reads from the intronic regions are missing as these regions are not included in the target. (e) When the read extraction is performed using a traditional aligner, reads matching the reference sequence are picked out successfully, but reads spanning exon junctions are often discarded.

Of course, the aligner-based extraction pipeline could be modified by building a custom reference genome using exon junction information. However, this would make that approach annotation-dependent and make it more complicated to maintain, generalize, and execute. In contrast, the triage approach is tolerant to splicing events event without annotations and allows one to perform the extraction with a single command.

References

[1] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009) *Genome Biology* **10**, R25.

[2] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009) *Bioinformatics* **25**, 2078–9.

[3] Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., and Mesirov, J. (2011) *Nature biotechnology* **29(1)**, 24–26.