Published in final edited form as:

IEEE Life Sci Lett. 2015 August ; 1(2): 22–25. doi:10.1109/LLS.2015.2465870.

SNAPR: a bioinformatics pipeline for efficient and accurate RNA-seq alignment and analysis

Andrew T. Magis, **Cory C. Funk**, and **Nathan D. Price**[§]

Institute for Systems Biology, Seattle, WA 98109

Abstract

The process of converting raw RNA sequencing data to interpretable results can be circuitous and time consuming, requiring multiple steps. We present an RNA-seq mapping algorithm that streamlines this process. Our algorithm utilizes a hash table approach to leverage the availability and power of high memory machines. SNAPR, which can be run on a single library or thousands of libraries, can take compressed or uncompressed FASTQ and BAM files as inputs, and can output a sorted BAM file, individual read counts, gene fusions and identify exogenous RNA species in a single step. SNAPR also does native Phred score filtering of reads. SNAPR is also well suited for future sequencing platforms that generate longer reads. Using SNAPR, we show how we can analyze data from hundreds of TCGA samples in a matter of hours, while identifying gene fusions and viral events at the same time. With the references genome and transcriptome undergoing periodic updates, and the need for uniform parameters when integrating multiple data sets, there is great need for a streamlined process for RNA-seq analysis. We demonstrate how SNAPR does this efficiently and accurately, with the high-throughput capacity needed to do high-volume analyses.

I. Introduction

RNA sequencing (RNA-seq) is the primary technology used to measure genome-wide gene expression and transcriptome variation in biological samples. As of 2015, the NCBI Sequence Read Archive contained over 3.5 petabases of sequencing data, with a projected doubling time of 22.3 months. The explosive growth of next-generation sequence data now exceeds the growth rate of storage capacity [1]. The complexity of the transcriptome presents particular challenges for RNA-seq alignment algorithms. Pseudogenes, paralogs with high sequence similarity, and low complexity/repetitive regions can contribute to misaligned reads. The GENCODE project estimates over 14,000 pseudogenes exist in the human genome [2]. Compounding the alignment challenge is the fact that new species of RNA continue to be discovered, including novel fusion genes and trans-splicing events. While DNA sequencing is more commonly used to identify genomic rearrangements, RNA-seq used for this purpose can more easily identify functionally aberrant species with a role in pathology [3]. Researchers' ability to process and analyze RNA-seq data depends upon

Corresponding Author: nprice@systemsbiology.org.

bioinformatics tools that are fast, accurate, and easy to use—especially when applied to large data sets or when additional analyses beyond standard mapping are required. Here we present a new RNA-seq alignment algorithm based on the DNA-focused Scalable Nucleotide Alignment Program (SNAP) [4]. We call this new alignment algorithm SNAPR (SNAP for RNA, pronounced “snapper”).

II. Methods

SNAPR categorizes each putative paired-end alignment based on the transcriptome annotation that is provided as an input. All valid alignments for each paired-end read are compared to the list of annotated transcripts. If both mates align within the boundaries of an annotated gene, that alignment is categorized as *intra-gene*. Putative alignments that are not intra-gene but occur on the same chromosome are *intra-chromosomal*. Finally, putative alignments that cross chromosome boundaries are categorized as *inter-chromosomal*. Each paired-end read can have valid alignments that occur in one or more of these categories. SNAPR prioritizes alignments in the following order: intra-gene, intra-chromosomal, inter-chromosomal. For example, if a read generates an alignment that is categorized as intra-gene, no alignments in any of the other categories are considered valid, even if those alignments contain fewer mismatches with the reference genome. This categorization is designed to leverage the likelihood of alignments occurring within annotated genes while biasing the algorithm against inter-chromosomal alignments. As a result, any intra- or inter-chromosomal alignments that do pass through these filters are subsequently more likely to result from real biological events. A detailed description of the methods can be found at https://price.systemsbiology.net/SNAPR.

SNAPR is freely available via the Apache 2.0 license. System requirements for SNAPR include >40 GB of RAM memory for both the creation of indices and alignment. SNAPR was run on Amazon EC2 *cr1.8xlarge* instances. An Amazon AMI containing SNAPR is publically available and can also be found at the Price Lab website (above). It is recommended that SNAPR be run on a multi-core machine, preferably containing a solid-state hard drive for faster index loading. Simulated RNA sequencing data was generated using Mason (https://www.seqan.de/projects/mason/) using the Venter genome [5]. SNAPR indices were generated using the GRCh37 genome assembly and the Ensembl v68 human genome annotation. All sequenced viral genomes (1376 genomes), all sequenced fungal genomes (35 genomes), and all sequenced bacterial genomes (2646 genomes) were downloaded from the NCBI FTP site (ftp.ncbi.nih.gov/genomes). A custom Python script was used to select one bacterial genome from each genus to be part of the contamination database, for a total of 1145 genomes. 312 RNA-seq samples of stomach adenocarcinoma, 159 glioblastoma multiforme and 117 acute myeloid leukemia and 808 ovarian serous cystadenocarcinoma were identified and downloaded from the UCSC Cancer Genomics Hub (https://cghub.ucsc.edu).

III. Results

A. Performance of SNAPR

In order to test performance and accuracy on the identification of genes from pseudogenes where ground truth is known, we simulated 2.5×10^6 paired-end reads of varying lengths derived from the published Venter genome [5] using Mason [6]. Each data set contained standard Illumina™ sequencing error rates as well as homozygous and heterozygous variants and indels present in the Venter genome. In the first dataset, 80% of the reads were generated from the Ensembl v68 transcriptome (e.g. crossing splice junctions) and 20% were generated directly from the genome. For the second dataset, 80% of the reads were generated from the transcriptome as before, but the remaining 20% were generated only from annotated pseudogenes. This dataset was used to estimate mapping and variant-calling accuracies in an exaggerated mapping challenge. We chose three of the most recently published and most used aligners with which to compare: Tophat2/Bowtie2 [7], STAR [8], and Subjunc [9]. The Genome Analysis Toolkit v2.5 [10] was used for variant calling in order to estimate read mapping accuracies; incorrectly mapped reads should generate spurious variants while missing real variants. The Receiver Operating Characteristic (ROC) curves for variant calling are shown in Figure 1A. SNAPR was the most accurate aligner on both datasets, most notably in the presence of excessive pseudogene reads (Figure 1B). SNAPR was also the fastest aligner when BAM conversion was taken into account: nearly twice as fast as STAR and 25× faster than Tophat2/Bowtie2 (Figure 1C).

We generated a third RNA-seq dataset with Mason that contained 2×10^6 paired-end reads with standard Illumina™ sequencing error rates, again using the Venter genome. This dataset was processed using all four aligners and gene read counts were generated. SNAPR and Subjunc are able to generate gene read counts internally, while read counts for STAR and Tophat2/Bowtie2 were generated using htseq-count. Computed gene read counts were compared to the exact gene read counts generated by Mason. The scatterplot comparisons are presented in Figure 2. Among the aligners tested, SNAPR had the best R^2 value (0.985) and appeared to have the most even distribution of read count variation. STAR had the fewest instances where it under-estimated read counts, with an apparent bias in read count estimation towards higher read counts. Both Subjunc and TopHat/BowTie appeared to have a similar, though less pronounced bias. While gene length is not explicitly part of the metric, there appears to be some differences in the distribution of error related to gene counts for the different algorithms.

B. Features of SNAPR

SNAPR incorporates the annotation directly into the alignment process to improve alignment accuracy. Each mate (one for single-end, two for paired-end) is aligned to both the genome and the transcriptome independently, creating a set of unique putative alignment positions for mate(s) to both indices. The final alignment position and subsequent mapping score is determined based on several criteria (see methods). Aligning to both the transcriptome and the genome simultaneously serves a dual purpose: it ensures that all possible alignment positions for a mate are considered, and it allows paired end reads to cross transcriptome/genome boundaries. While an annotation provides critical prior

knowledge about the likelihood of alignment positions, gene boundaries are often truncated at the 5' and 3' ends. SNAPR can also align paired end reads for which one mate occurs in an intron and the other in an exon, a situation resulting from unannotated exons or sequenced pre-mRNA. The current version of SNAPR is designed for genomes with curated annotations, and does not attempt to find novel splice junctions. To maximize efficiency of disk usage, which is a critical issue in the analysis of growing amounts of RNA-seq data, SNAPR is capable of natively reading from and writing to BAM format without requiring any external software packages, as well as standard FASTQ and SAM formats. This can be of benefit for instances where the raw FASTQ files aren't made available and for instances when one wishes to combine multiple data sets. SNAPR enables the mapping to be done under the same parameters under a single implementation. Sample quality is of paramount importance in RNA-seq analysis, including filtering of low-quality reads generated by the sequencer and identification of sample contaminants. SNAPR performs quality filtering of input reads automatically, with the default setting requiring >80% of the read to have a Phred score of 20 or better.

Sequenced samples may contain products of viral infections and/or bacterial or fungal species, either expected or not. SNAPR allows users the option of providing a secondary alignment database of their choosing to which unaligned reads are tested, automatically writing a list of all alternative alignments and corresponding read sequences. All analyses for this paper were performed using our'contaminant database' containing all sequenced viral genomes (1376), fungal genomes (35), and one genome from each sequenced bacterial genus (1145). This database, along with another database containing all sequenced human pathogens derived from PATRIC [11], is available from our download page. To demonstrate the utility of such a contaminant database we analyzed 312 RNA-seq samples of stomach adenocarcinoma, where the presence of Epstein-Barr Virus (EBV) has been previously reported [12]. We identified 70 samples with detectable (>10) and 24 samples with appreciable (>1000) numbers of reads mapping to the EBV type 1 genome. Of these, 17 samples contained EBV type 1 as the strongest identifiable externally mapping target in the report. All of these 17 samples also contained detectable levels of EBV type 2. Additionally, 41 samples contained detectable levels of cytomegalovirus (CMV), with one sample containing CMV as the strongest identifiable externally mapping target. The distribution of read counts for EBV is shown in Figure 1D. We note that among the 1295 glioblastoma multiforme (GBM), stomach adenocarcinoma (STAD), acute myeloid leukemia (LAML), and ovarian serous cystadenocarcinoma (OV) samples processed for this paper, only stomach adenocarcinoma exhibited any appreciable amounts of EBV, as expected. Also, SNAPR identified ten LAML samples with very high read counts mapping to genus Acinetobacter (~6×10^6 reads), species of which are commonly associated with hospital-acquired infections in immunocompromised patients.

RNA-seq is commonly used for the analysis of differentially expressed genes. The most widely accepted method for estimating statistically significant variation in read counts is based on the negative binomial distribution. This method has been adopted by the R packages DESeq [13]and edgeR [14], among others. The process of generating read counts can add substantial processing time per sample. SNAPR automatically reports all read counts for immediate statistical analysis, with no running time penalty and no additional

software required. SNAPR also reports spliced read counts normalized by gene expression for all annotated splice junctions to enable alternative splicing detection.

For the identification of fusion genes or trans-splicing events, SNAPR makes use of both mate pair information and spliced reads to filter out false positives (Figure 3A). SNAPR automatically reports all putative fusion events in a sorted report as well as GTF format for easy visualization. All read sequences participating in putative fusions are also output automatically. A novel gene fusion was recently identified between the genes *FGFR3* and *TACC3* in approximately 3% of the studied glioblastoma multiforme (GBM) samples [15]. We applied SNAPR to the fusion samples (TCGA-27-1835,TCGA-76-4925) identified in Singh et al., and it reported the *TACC3-FGFR3* fusion as the top intra-chromosomal fusion candidate for both samples, with over 20,000 evidentiary reads in sample TCGA-27-1835 and nearly 600 evidentiary reads in sample TCGA-76-4925. We next applied SNAPR to a chronic myelogenous leukemia (CML) sample from dbGaP (SRR607562), identifying the canonical *BCR-ABL1* fusion gene product as the top inter-chromosomal fusion event with nearly 100 evidentiary reads (Figure 3B). Recently Frattini et al. reported the landscape of gene fusions in 58 glioblastoma RNA-seq samples from TCGA [16]. We ran SNAPR on the same 58 samples, partially or completely identifying 93% of the fusions reported in Frattini et al. (Figure 3C).

In total, we have processed 312 stomach adenocarcinoma (STAD), 58 glioblastoma multiforme (GBM), 117 acute myeloid leukemia (LAML), and 808 ovarian serous cystadenocarcinoma (OV) samples, reading directly from BAM format, quality filtering all input reads, generating read counts for immediate statistical analysis, identifying putative fusion events and contaminants, and finally writing the new alignments directly back to BAM format. No step of this analysis required any external software package and was completed using a single command for each sample. As discussed in Frattini et al., we find evidence for fusions in GBM involving *EGFR, LANCL2*, and *SEPT14*. In addition to the viral and bacterial infections identified in STAD, SNAPR finds evidence for fusions with *ERBB2* [17] and *IGF2* [18], which are commonly amplified and overexpressed in gastric cancers. Few cancer types are more associated with translocations than leukemia, and SNAPR identifies the canonical fusions in the LAML samples: *BCR-ABL1, CBFB-MYH11, RUNX1-RUNX1T1*, and *PML-RARA*. Finally, in OV samples SNAPR finds evidence for fusions involving *IGF2* [19], *H19* [19], *GPX3* [20], *MUC16* [21], and *WNT7A* [22].

IV. Conclusion

The promise of discovery through RNA sequencing is bottlenecked by our ability to analyze the data, and the rate of data generation continues to accelerate. Databases such as The Cancer Genome Atlas and the NCBI Sequence Read Archive are accumulating petabytes of data. SNAPR provides a fast, accurate, and integrated package that streamlines basic RNAseq alignment and analysis for large scale processing. SNAPR enables researchers to leverage hundreds to thousands of samples easily to accurately identify statistically rare patterns of gene expression or other transcriptomic perturbations, while also providing

automatic additional detection of contaminants and fusion events. The code and additional detailed documentation can be found at http://price.systemsbiology.net.

Acknowledgments

We thank Ravi Pandya, Bill Boloski, and Taylor Sittler for work on SNAP and comments on this manuscript. This work was supported by NIH grants U01AG046139, R01AI084914, Camille Dreyfus Teacher-Scholar program and BD2K 1U54EB020406-01.

References

1. Kodama Y, Shumway M, Leinonen R, et al. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. Jan; 2012 40(Database issue):D54–6. [PubMed: 22009675]
2. Pei B, Sisu C, Frankish A, et al. The GENCODE pseudogene resource. Genome Biol. 2012; 13(9):R51. [PubMed: 22951037]
3. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. Feb; 2011 12(2):87–98. [PubMed: 21191423]
4. Zaharia M, Bolosky WJ, Curtis K, et al. Faster and More Accurate Sequence Alignment with SNAP. ArXiv. Nov.2011 1111
5. Levy S, Sutton G, Ng P, et al. The diploid genome sequence of an individual human. Plos Biol. 2007; 5(10):e254. [PubMed: 17803354]
6. Holtgrewe M. Mason–a read simulator for second generation sequencing data. Technical Report FU Berlin. 2010
7. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. Apr.2013 14(4):R36. [PubMed: 23618408]
8. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. Jan; 2013 29(1):15–21. [PubMed: 23104886]
9. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. May.2013 41(10):e108. [PubMed: 23558742]
10. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. May; 2011 43(5):491–498. [PubMed: 21478889]
11. Gillespie JJ, Wattam AR, Cammer SA, et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect. Immun. Nov; 2011 79(11):4286–4298. [PubMed: 21896772]
12. Comprehensive molecular characterization of gastric adenocarcinoma. Sep; 2014 513(7517):202–209.
13. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10):R106. [PubMed: 20979621]
14. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. Jan; 2010 26(1):139–140. [PubMed: 19910308]
15. Singh D, Chan JM, Zoppoli P, et al. Transforming fusions of FGFR and TACC genes in human glioblastoma. Science. Sep; 2012 337(6099):1231–1235. [PubMed: 22837387]
16. Frattini V, Trifonov V, Chan JM, et al. The integrated landscape of driver genomic alterations in glioblastoma. Nat. Genet. Aug.2013 45:1141–1149. [PubMed: 23917401]
17. Gravalos C, Jimeno A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. Ann. Oncol. Sep; 2008 19(9):1523–1529. [PubMed: 18441328]
18. Wu MS, Wang HP, Lin CC, et al. Loss of imprinting and overexpression of IGF2 gene in gastric adenocarcinoma. Cancer Lett. Nov; 1997 120(1):9–14. [PubMed: 9570380]
19. Murphy SK, Huang Z, Wen Y, et al. Frequent IGF2/H19 domain epigenetic alterations and elevated IGF2 expression in epithelial ovarian cancer. Mol. Cancer Res. Apr; 2006 4(4):283–292. [PubMed: 16603642]

20. Hough CD, Cho KR, Zonderman AB, et al. Coordinately up-regulated genes in ovarian cancer. Cancer Res. May; 2001 61(10):3869–3876. [PubMed: 11358798]

21. Thériault C, Pinard M, Comamala M, et al. MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis. Gynecol. Oncol. Jun; 2011 121(3):434–443. [PubMed: 21421261]

22. Yoshioka S, King ML, Ran S, et al. WNT7A regulates tumor growth and progression in ovarian cancer through the WNT/β-catenin pathway. Mol. Cancer Res. Mar; 2012 10(3):469–482. [PubMed: 22232518]

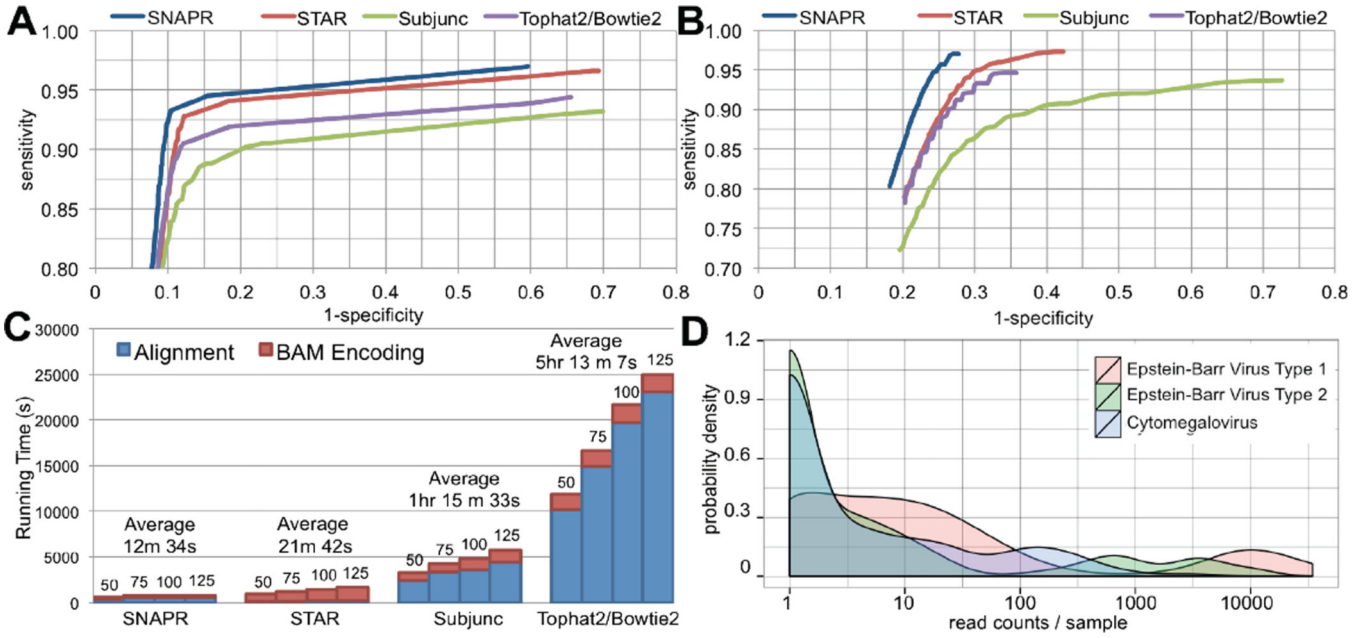

Figure 1.

(A) ROC curve for variant calling with GATK using 2.5×10^7 100bp paired-end reads generated by Mason with Illumina™ sequencing errors using the Venter genome. 80% of reads were generated from the Ensembl v68 annotation, while 20% were generated directly from the genome. **(B)** ROC curve for variant calling with GATK using 2.5×10^7 100bp paired-end reads generated by Mason with Illumina™ sequencing errors using the Venter genome. 80% of reads were generated from the Ensembl v68 annotation, while 20% were generated from annotated pseudogenes. **(C)** Running times for all four aligners on 2.5×10^7 paired-reads of varying lengths (50, 75, 100, 125bp). All processing was performed on 16 cores using an Amazon EC2 *cr1.8xlarge* instance. **(D)** Distribution of reads aligning to Epstein-Barr virus Type 1, Epstein-Barr virus Type 2, and cytomegalovirus genomes from 312 processed TCGA stomach adenocarcinoma samples by SNAPR. Read counts were automatically generated using the contamination database functionality implemented in SNAPR. Samples with zero counts are not shown. Note the x-axis is on a logarithmic scale.

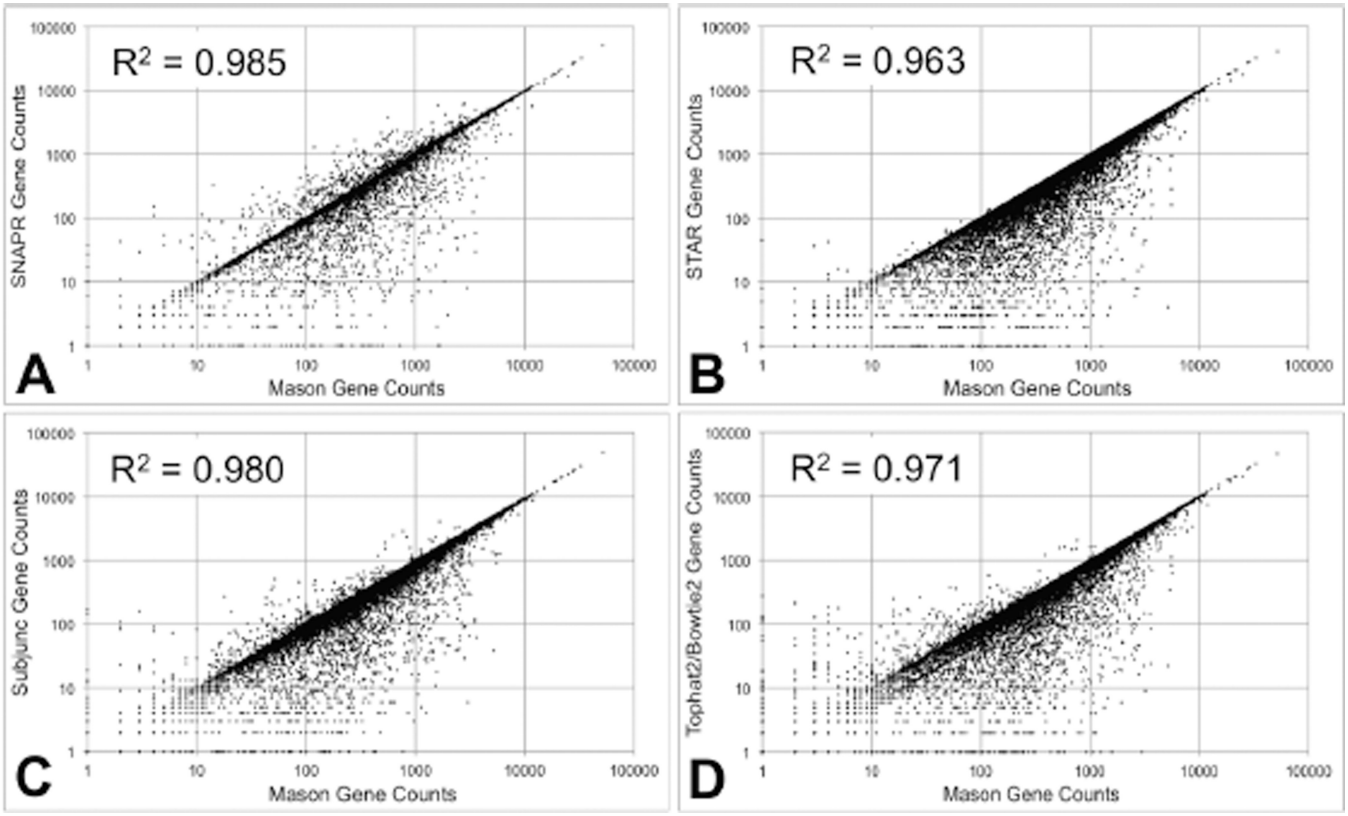

Figure 2.

Aligned gene read counts for 2×10^7 paired-end reads generated by Mason with Illumina™ sequencing errors using the Venter genome.100% of reads were generated from the Ensembl v68 annotation, and aligned using each of the four tested aligners. 'Correct' read counts (x-axis)for each gene were generated from the Mason SAM file. **(A)** SNAPR automatically generates gene read counts. **(B)** Read counts for STAR alignments were generated using htseq-count. **(C)** Read counts for Subjunc alignments were generated using the featureCounts functionality. **(D)** Read counts for Tophat2/Bowtie2 alignments were generated using htseq-count.

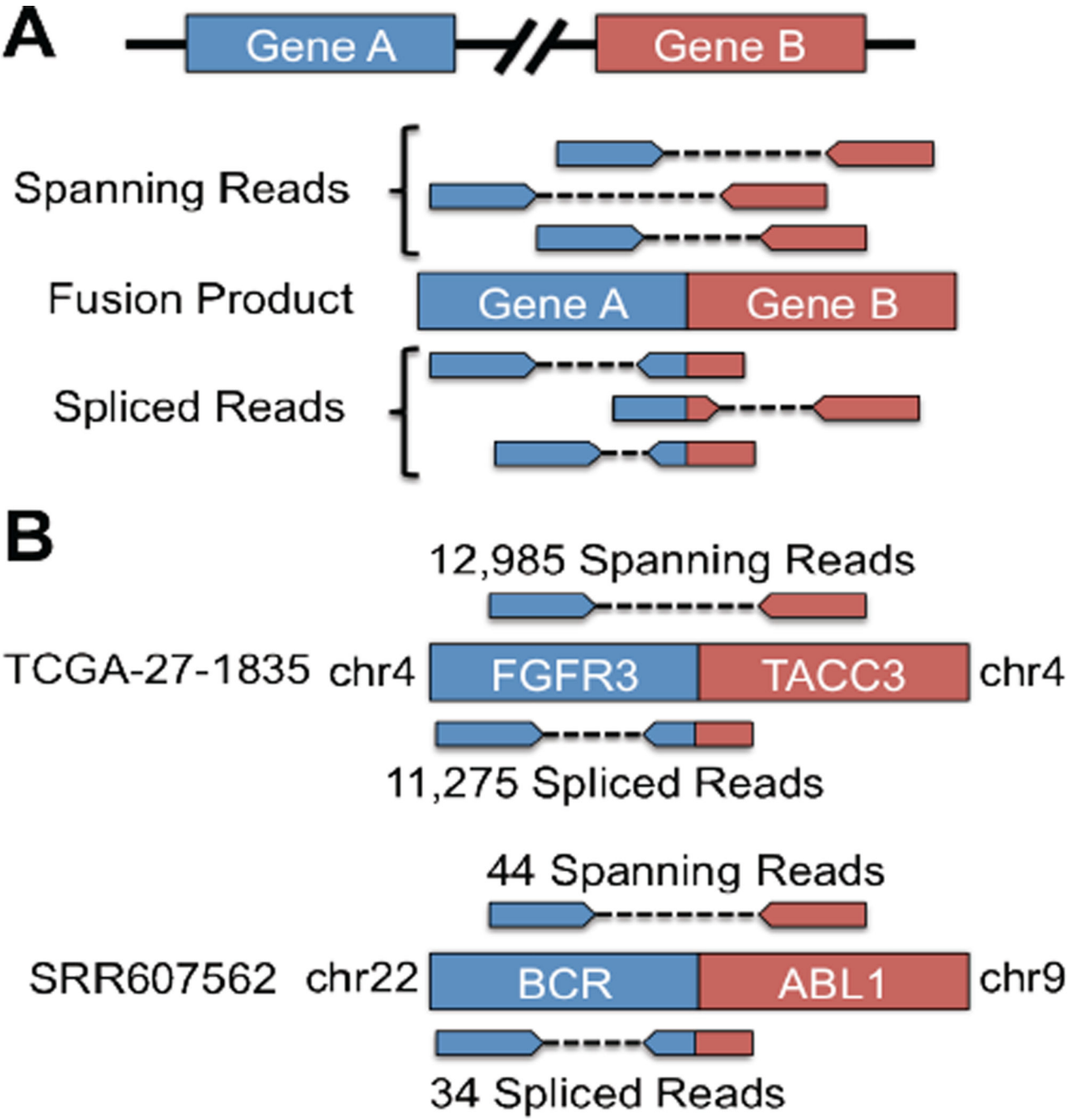

Figure 3.

(A) SNAPR only considers fusion events that are supported by spanning reads as well as spliced reads. **(B)** SNAPR finds the *FGFR3-TACC3* fusion event in GBM sample TCGA-27-1835 with thousands of evidentiary reads, and the *BCR-ABL1* fusion in the CML sample SRR607562 with nearly 100 evidentiary reads.