RegTools: Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer

Yang-Yang Feng[1], Avinash Ramu[2], Kelsy C. Cotto[1], Zachary L. Skidmore[1], Jason Kunisaki[1], Donald F. Conrad[2], Yiing Lin[3], William Chapman[3], Ravindra Uppaluri[4], Ramaswamy Govindan[5], Obi L. Griffith[1,2,5,6,*], Malachi Griffith[1,2,5,6,*]

AFFILIATIONS

[1]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA
[2]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA
[3]Department of Surgery, Washington University School of Medicine, St. Louis, Missouri, USA
[4]Department of Head and Neck Surgical Oncology, Dana-Farber Cancer Institute, Brigham & Women's Hospital, Boston, Massachusetts, USA.
[5]Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, USA
[6]Siteman Cancer Center, Washington University School of Medicine, St. Louis, Missouri, USA
[*]To whom correspondence should be addressed: obigriffith@wustl.edu, mgriffit@wustl.edu

ABSTRACT

The interpretation of variants in cancer is frequently focused on direct protein coding alterations. However, this analysis strategy excludes somatic mutations in non-coding regions of the genome and even exonic mutations may have unidentified non-coding consequences. Here we present RegTools, a software package designed to integrate analysis of somatic variant calls from genomic data with splice junctions extracted from transcriptomic data in order to efficiently identify variants that may cause aberrant splicing in tumors. RegTools is open source (MIT license) and freely available as source code or as a Docker image (http://regtools.org/).

1 INTRODUCTION

Alternative splicing of messenger RNA is a biological process that allows a single gene to encode multiple gene products, increasing a cell's functional diversity and regulatory precision. However, splicing malfunction can lead to imbalances in transcriptional output or even the presence of oncogenic novel transcripts (Chabot and Shkreta, 2016). The interpretation of variants in cancer is frequently focused on direct protein coding alterations (Vogelstein *et al.*, 2013). However, most somatic mutations arise in intronic and intergenic regions, and exonic mutations may also have unidentified consequences. For example, mutations can affect splicing either in trans, by acting on splicing effectors, or in cis, by altering the splicing signals located on the transcripts themselves (Climente-González *et al.*, 2017). However, our understanding of the landscape of these variants is limited and few tools exist for their discovery. Here we present RegTools, a free, open-source software package designed to efficiently identify potential cis-acting, splicing-relevant variants in tumors (regtools.org).

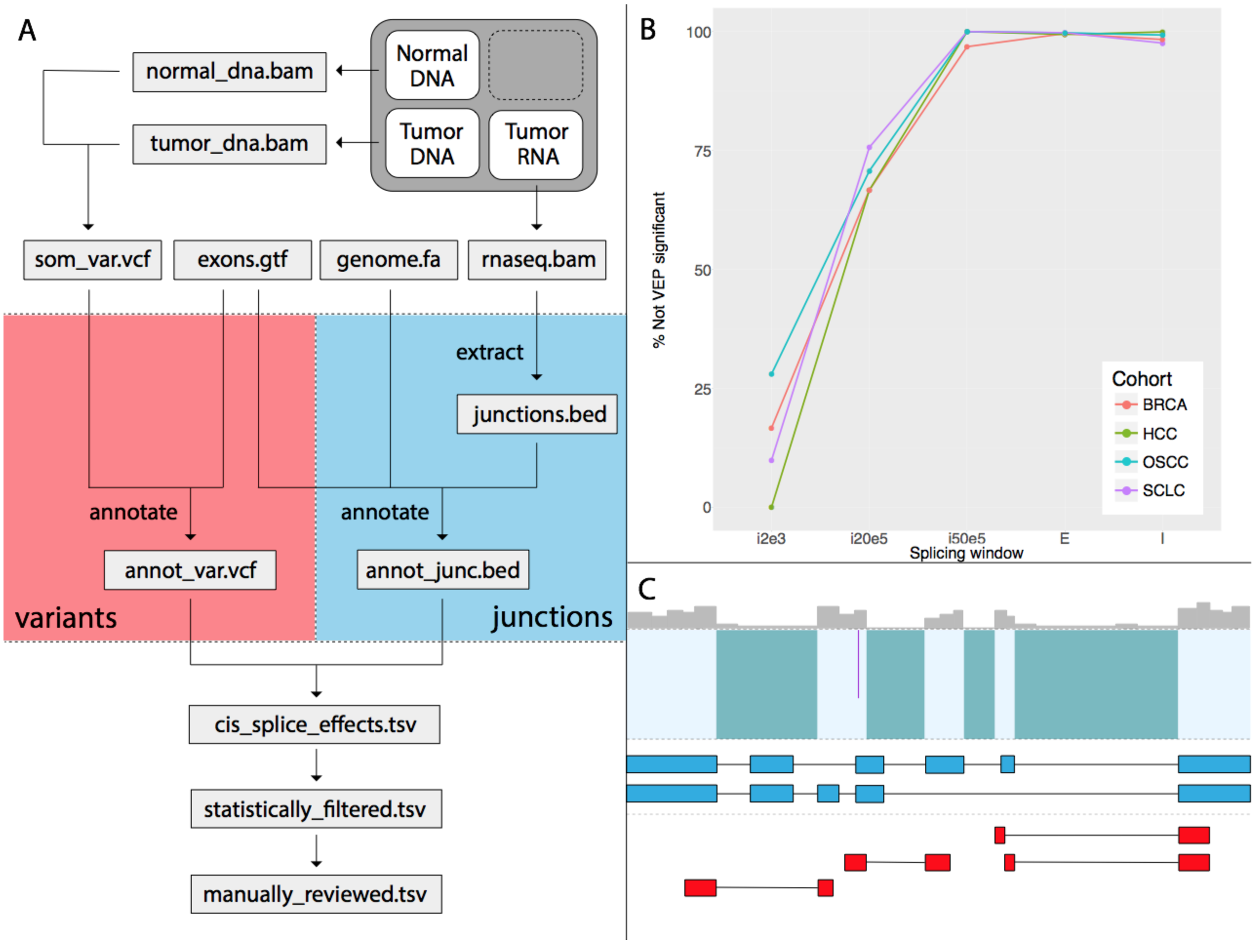

Figure 1: ***Regtools cis-splice-effects identify*** **workflow and demonstrated application to patient data.** RegTools is a suite of tools designed to enable flexible, streamlined discovery of variants which cause aberrant splicing effects that result in novel exon-exon junctions. (A) The *cis-splice-effects identify* workflow takes variant calls and RNA-seq alignments along with genome and transcriptome references and outputs information about novel junctions and potential associated cis splice-altering sequence variants. RegTools is agnostic to downstream research goals and its output can be filtered through user-specific methods and thus can be applied to a broad set of scientific questions. (B) Our application of the *cis-splice-effects identify* workflow to patient samples demonstrates its utility in identifying potentially important DNA variants that would be missed without consideration of the functional RNA-seq output (e.g. by direct annotation of the DNA variants with tools such as VEP). We identified significant variants in 5 different splicing windows (‘i2e3’, ‘i20e5’, ‘i50e5’, ‘E’, and ‘I’) that considered increasing

distances from exon edges (i.e. known splice sites) and compared our results to VEP annotations (**Supplementary Methods)**. (C) Panel C provides an illustrated Integrative Genomics Viewer (IGV) session of a hypothetical cis splice-altering variant which could be identified by our workflow and then manually reviewed (Robinson *et al.*, 2011). Coverage at each base position is shown at the top in gray. The next track shows individual read alignments with the variant position highlighted in purple for supporting reads. In the bottom two tracks, reference exons from a GTF file are depicted in blue and identified junctions from a BED file are shown in red.

2 USAGE AND FEATURES

RegTools is a suite of tools designed to aid users in a broad range of splicing-related analyses and is efficiently implemented in C++. The *variants annotate*, *junctions extract*, and *junctions annotate* commands can be performed separately or combined with the *cis-splice-effects identify* command (**Supplemental Methods**). *Cis-splice-effects identify* requires the following as input: a VCF file of variant calls, a BAM file of aligned RNA-sequencing reads, a reference genome FASTA file, and a reference transcriptome GTF file (**Figure 1**). First, variants within a “splicing window” (based on position relative to known exons) are annotated as splicing-relevant. The default window extends 2 bp into the intron and 3 bp into the exon, from each exon edge (i2e3). Next, the pipeline scans the transcriptomic data in regions local to splicing-relevant variants (“variant regions”) for exon-exon junctions, which are stored in a BED file for later analysis. By default, each variant region extends from the 5’ end of the exon directly upstream of the variant-associated exon to the 3’ end of the exon directly downstream of it. To enable the discovery of an arbitrarily expansive set of variants and junctions, RegTools allows the user to customize the size of splicing windows and variant regions (**Supplementary Figure 1**). Based on the user-supplied reference genome FASTA and transcriptome GTF, each extracted junction is annotated with the splice-site motif, the overlapping transcripts and genes, the number of acceptors/exons/donors skipped, and whether the acceptor/donor sites are known to the reference. Finally, each junction is annotated with its associated variant and the type of splicing aberration that has occurred (“junction type”: ‘DA’ for known donor and acceptor; ‘NDA’ for novel combination of known donor and acceptor; ‘D’ for known donor only; ‘A’ for known acceptor only; ‘N’ for unknown donor and acceptor) (**Supplementary Figure 2**; **Supplementary Methods**).

3 APPLICATION TO TUMOR SAMPLES

We applied RegTools to four patient cohorts: 28 hepatocellular carcinoma (HCC) samples (dbgap accession: phs001106), 21 small cell lung cancer (SCLC) samples (phs001049), 106 breast carcinoma (BRCA) samples (phs000178), and 33 oral squamous cell carcinoma (OSCC) samples (phs001623). DNA and RNA alignment and variant calling were performed as previously described using the Genome Modeling System (GMS) (Griffith, Griffith, *et al.*, 2015; Griffith, Miller, *et al.*, 2015; **Supplementary Methods)**. Our measurement of effect size was complicated by the fact that few samples share any particular (recurring) somatic variant. To prioritize candidates, we ignored DA and N junctions and required > 5 reads of support. As *cis-splice-effects identify* results merely reflect the proximity of junctions to potential splice variants, we performed statistical analyses to filter out false positives. We initially tested for significantly increased levels of a novel junction in the presence of a particular variant using both an "outlier" and a "ratio" method. Since matched normal samples of the same tissue type were available for HCC, we also initially considered three additional analyses for this cohort (**Supplementary Methods**). We decided to proceed using the simple outlier method alone, as its results were either comparable to or nearly a strict subset of the results of the other methods, indicating potentially higher quality calls and more efficiently prioritizing results for downstream analysis and manual review (**Supplementary Table 1**; **Supplementary Methods**).

We completed the above workflow for 5 different splicing window sizes: 'i2e3' , 'i20e5', 'i50e5', 'E' (entire exon), and 'I' (entire intron) (**Supplementary Table 2**; **Supplementary Methods**). Each successively broader analysis identified additional variants in each cohort (**Supplementary Table 2**; **Supplementary Figure 3**). In smaller windows, NDA junctions constituted the majority of junctions while A and D junctions remained fairly even. As window size increased, the proportion of NDA junctions decreased (**Supplementary Table 3**; **Supplementary Figure 4**). This might be explained by the fact that in larger windows, even the ostensibly exonic-only "E" window, variants are more likely to lie in intronic regions and therefore less likely to cause skipping through the disruption of splicing machinery on canonical exons. All identified junctions are listed in **Supplementary Tables 5-8**, with particular examples shown in **Supplementary Figures 5 - 10.**

To compare our results against existing approaches, we annotated all variants identified by *cis-splice-effects identify* with Ensembl's Variant Effect Predictor (VEP) in the "per_gene" and "pick" modes (McLaren *et al.*, 2016). We considered any variant with at least one splicing-related annotation to be "VEP significant". Most splicing-unrelated annotations were 'intronic', 'missense', 'upstream gene', 'non-coding transcript', 'synonymous', and 'UTR' (**Supplementary Figure 11**). In small windows (i2e3 and i50e5), a large percentage of outlier significant variants were VEP significant. This percentage dropped steeply to ~1% in the i50e5, E, and I windows (**Supplementary Table 4**; **Supplementary Figure 12**). Interestingly, the proportion of VEP significant variants was consistently higher in the set of outlier significant splice variants versus unfiltered RegTools splice variants, suggesting that our approach identified true positives while also detecting splice variants which VEP missed (**Supplementary Table 4**; **Supplementary Figure 13**).

4 DISCUSSION

Few tools are directed at linking aberrant splicing to variants in cis, and most are tailored to the authors' particular aims (Jayasinghe *et al.*, 2018; Pertea *et al.*, 2001; Jung *et al.*, 2015). RegTools is designed for broad applicability and computational efficiency (**Supplementary Figure 14)**. By relying on well-established standards for sequence alignments, annotation files, and variant calls and by remaining agnostic to downstream statistical methods and comparisons, our tool can be applied to a wide set of scientific queries and datasets. In our analysis, we showed that RegTools combined with downstream filtering identifies splice variants that the field standard VEP misses by not accounting for sample-specific transcriptomic information. Importantly, RegTools can be integrated with existing utilities such as SUPPA2 to focus on functional splicing alterations (Trincado *et al.*, 2018). As such, this flexible and robust tool could be applied to various large-scale pan-cancer datasets to elucidate the role of splice variants in cancer. The exploration of novel tumor-specific junctions will undoubtedly lead to translational applications, from discovering novel tumor drivers, diagnostic and prognostic biomarkers, and drug targets, to perhaps even identifying a previously untapped source of neoantigens for personalized immunotherapy.

FUNDING

A.R. was supported by the ‘Burroughs Wellcome Fund Institutional Program Unifying Population and Laboratory Based Sciences Award’ at Washington University. K.C.C. was supported by Siteman Cancer Center under fund number #3477-92400. O.L.G was supported by grants from the National Cancer Institute (NCI) of the NIH under award numbers K22CA188163 and U01CA209936. This work was supported by a grant to M.G. from the National Human Genome Research Institute (NHGRI) of the NIH under award number R00HG007940.

REFERENCES

Chabot,B. and Shkreta,L. (2016) Defective control of pre-messenger RNA splicing in human disease. *J. Cell Biol.*, **212**, 13–27.

Climente-González,H. *et al.* (2017) The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.*, **20**, 2215–2226.

Griffith,M., Griffith,O.L., *et al.* (2015) Genome Modeling System: A Knowledge Management Platform for Genomics. *PLoS Comput. Biol.*, **11**, e1004274.

Griffith,M., Miller,C.A., *et al.* (2015) Optimizing cancer genome sequencing and analysis. *Cell Syst*, **1**, 210–223.

Jayasinghe,R.G. *et al.* (2018) Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.*, **23**, 270–281.e3.

Jung,H. *et al.* (2015) Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.*, **47**, 1242–1248.

McLaren,W. *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.

Pertea,M. *et al.* (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Trincado,J.L. *et al.* (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.