bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Splice Expression Variation Analysis (SEVA) for Inter-tumor Heterogeneity of Gene Isoform Usage in Cancer

Bahman Afsari [1,+] Theresa Guo [2,+] Michael Considine [1]
Liliana Florea [3] Dylan Kelley [2] Emily Flam [2] Patrick K. Ha [4]
Donald Geman [5] Michael F. Ochs [6] Joseph A. Califano [7]
Daria A. Gaykalova [2,*] Alexander V. Favorov [1,8,9*] Elana J. Fertig [1,*]

[1]Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, 21205, USA
[2]Department of Otolaryngology - Head and Neck Surgery, Johns Hopkins University, Baltimore, MD, 21205, USA
[3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, 21205, USA
[4]Department of Otolaryngology – Head and Neck Surgery, University of California, San Francisco, CA, 94158, USA
[5]Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD, 21218, USA
[6]Department of Mathematics & Statistics, The College of New Jersey, Ewing, NJ, 08628, USA
[7]Division of Otolaryngology, Department of Surgery, University of California, San Diego, CA, 92093, USA
[8]Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, RAS, Moscow, 119333, Russia and
[9]Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 117545, Russia

[+]Co-first authors
[*]To whom correspondence should be addressed.

Abstract

Motivation: Alternative splicing events (ASE) cause expression of a variable repertoire of potential protein products that are critical to carcinogenesis. Current methods to detect ASEs in tumor samples compare mean expression of gene isoforms relative to that of normal samples. However, these comparisons may not account for heterogeneous gene isoform usage that is common in tumors.
Results: Therefore, we introduce Splice Expression Variability Analysis (SEVA) to detect differential splice variation, which accounts for tumor heterogeneity. This algorithm compares the degree of variability of junction expression profiles within a population of normal samples relative to that in tumor samples using a rank-based multivariate statistic that models the biological structure of ASEs. Simulated data show that SEVA is more robust to tumor heterogeneity and its candidates are more independent of differential expression than EBSeq, DiffSplice, and rMATS. SEVA analysis of head and neck tumors identified differential gene isoform usage robust in cross-study validation. The algorithm observes substantial differences in gene isoform usage between head and neck tumor subtypes, with greater inter-tumor heterogeneity in HPV-negative tumors with alterations to genes that regulate RNA splice machinery. Thus, SEVA is well suited for differential ASE analysis and assessment of ASE inter-tumor heterogeneity in RNA-seq data from primary tumor samples.
Availability: SEVA is implemented in the R/Bioconductor package GSReg.
Contact: bahman@jhu.edu, dgaykal1@jhmi.edu, favorov@sensi.org, ejfertig@jhmi.edu

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

Cancer is a disease of genetic disruption. Integrated analyses of DNA and RNA sequencing data identify clusters of tumor samples with common gene expression changes but lack consistent DNA alterations (Mo *et al.*, 2013). It is essential to find the hidden sources of molecular alterations that drive gene expression changes in heterogeneous tumor populations. Alternative splicing events (ASE) results in expression of different transcript isoforms and consequently a more variable repertoire of potential protein products (Lim *et al.*, 2011). ASEs are a significant component of expression alterations in cancer, and have been demonstrated to be critical to the development of malignant phenotypes in a variety of solid and liquid tumors (Chen and Weiss, 2015). Expression of alternative gene isoforms, even in a small set of genes grouped into common pathways, represents a relatively unexplored source of tumor-driving alterations.

Recent bioinformatics tools have demonstrated the ability to identify expressed gene isoforms from RNA-seq reads for a single sample (Song *et al.*, 2016; Canzar *et al.*, 2016; Pertea *et al.*, 2015; Guttman *et al.*, 2010; Li *et al.*, 2011; Wang *et al.*, 2010). These tools can systematically evaluate the gene isoforms that are expressed in a sample. Other techniques have been developed to classify cancer based upon isoform expression (Sebestyen *et al.*, 2015). Nonetheless, it is essential to perform analysis to identify splice variants that are uniquely expressed in RNA-seq data from tumor samples compared to normal tissues in order to characterize the landscape of splicing events specific to cancer. Most reported techniques to define differential ASE expression rely on comparing mean expression values to determine differences in ASE expression between sample groups, such as normal and tumor samples (Li *et al.*, 2011; Anders *et al.*, 2012; Hu *et al.*, 2013; Shen *et al.*, 2012, 2014).

Although there are a breadth of available ASE algorithms, few have been validated in tumor samples (Feng *et al.*, 2013). In cancer, splice variant patterns may be variable within tumors of the same subtype while ultimately having the same impact on the function of a gene or common pathway. A similar concept is observed in DNA mutations, where individual tumors can harbor differing mutations that are mutually exclusive and act within common pathways (Jones *et al.*, 2008; Thomas *et al.*, 2007). Therefore, altered splicing patterns seen in tumors may result in a variable gene isoform profiles across tumor samples. Current algorithms rely on statistics that identify differential mean expression in the isoform of a gene between sets. These algorithms may be appropriate in cases where tumors have homogeneous gene isoform usage relative to normal samples. However, these methods are insufficient to account for inter-tumor heterogeneity of gene isoform usage. Therefore, new algorithms are needed to compare variations in the isoforms that are expressed independent of their mean expression in tumor and normal samples

In this paper, we develop a novel algorithm called Splice Expression Variability Analysis (SEVA) for differential gene isoform usage in cancer from RNA-seq data implemented in the R/Bioconductor package GSReg. SEVA simultaneously account for tumor heterogeneity and mitigate confounding of ASEs with differentially expressed genes. This algorithm uses a multivariate, non-parametric variability statistic to compare the heterogeneity of expression profiles for gene isoforms in tumor relative to normal samples. The performance of SEVA was compared with three existing algorithms designed for differential splice variant expression analysis, EBSeq (Anders *et al.*, 2012), DiffSplice (Hu *et al.*, 2013), and rMATS (Shen *et al.*, 2014), in simulated RNA-seq data generated with Polyester (Frazee *et al.*, 2015b). We show that SEVA had the most robust performance in heterogeneous test samples, which are representative of primary tumor samples. SEVA was unique in identifying alternative splicing events independent of overall gene expression differences when there is heterogeneity in simulated cancer samples. Additional validation was performed using cross-study validation with publicly available RNA-seq data for primary tumor data from HPV+ head and neck squamous cell carcinoma (HNSCC) tumors and normal samples. In addition to finding differential alternative splicing events, SEVA can also uniquely quantify the relative heterogeneity of gene isoform usage for each phenotype among those genes. SEVA finds greater inter-tumor heterogeneity in ASEs specific to HPV+ HNSCC tumor samples that include experimentally validated splice variants in HNSCC from a previously microarray study (Li *et al.*, 2014). SEVA does not observe differences of heterogeneity in gene isoform usage between the dominant HNSCC subtypes: HPV+ and HPV-. Nonetheless, it observes greater inter-tumor heterogeneity in gene isoform usage among HPV- HNSCC tumors that have genetic alterations in genes that regulate RNA splicing machinery (Ebert and Bernard, 2011). Therefore, SEVA represents a robust algorithm that is well suited to find inter-tumor heterogeneity in gene isoform usage

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in cancer samples relative to samples from a control group.

2 Methods

2.1 *In silico* data

To simulate isoform expression, we used the expression of isoforms from the HPV+ RNA-seq data from Guo *et al.*, 2016. For efficiency of the simulations, we selected genes in chromosome 1 whose expression 4 to 9 in log2 scale in normal samples (600 genes). We generated a dataset of 25 tumor and 25 normal simulated samples. To simulate normal samples, we calculated the average isoform expression for these genes in normal samples and input these values to Polyester with default values to simulate inter-sample variation (Frazee *et al.*, 2015b). Genes in tumor samples are simulated as differentially spliced by exchanging junction expression from the normal samples randomly between junctions in isoforms of the gene, analogous to previous differential gene isoform simulation studies (Alamancos *et al.*, 2015; Liu *et al.*, 2014). Genes were simulated as differentially expressed by introducing a log fold change of 1 to all isoforms relative to the values in normal samples. In the simulation, 150 genes are differentially spliced, 150 differentially expressed, 150 both, and 150 neither. The number of tumor samples with alternative splice variant usage and differential expression varies to model inter-tumor heterogeneity.

2.2 HPV+ HNSCC and normal RNA-seq datasets

We use RNA-seq data for 46 HPV+ HNSCC and 25 independent normal samples from uvulopalatopharyngoplasty previously described in Guo *et al.*, 2016 and 44 HPV+ HNSCC and 16 matched normal tissues from the freeze set for TCGA HNSCC (Cancer Genome Atlas Network, 2015).

2.3 RNA-sequencing data normalization and mutation calls

All *in silico* and real RNA-seq data are normalized with the RNA-seq v2 pipeline from TCGA (Cancer Genome Atlas Network, 2015). Junction expression is obtained directly from the MapSplice (Wang *et al.*, 2010) output for each sample, setting expression values to zero for junctions that are not detected from MapSplice in a given sample. Simulated data is also aligned with TopHat2 version 2.1.0 (Kim *et al.*, 2013). Gene and isoform expression data from TCGA were obtained from the level 3 normalized data, but junction expression was obtained by rerunning MapSplice to perform *de novo* junction identification to compare with the training RNA-seq data from (Guo *et al.*, 2016). Comparisons between HPV-positive and HPV-negative HNSCC are made for level 3 junction data for previously annotated junctions for TCGA samples available on FireBrowse. TCGA samples with mutations or copy number alterations any of the SF3B1, SF1, SF3A1, SRSF2, U2AF1, U2AF2, ZRSR2, or PRPF40B genes in cBioPortal (Gao *et al.*, 2013) are said to have altered RNA splice machinery based upon annotations in Ebert and Bernard, 2011.

2.4 Implementation and software

SEVA is implemented in the R/Bioconductor package GSReg (Afsari *et al.*, 2014a). The analyses presented in this study remove genes with less than three junctions from analysis. Additional filtering criterion are described in the vignette, but not used. The SEVA analysis of junction expression is computationally efficient. All of the SEVA computations for simulated data completed on a Lenovo Thinkpad with Core (TM) i7-3720QM Intel CPU @2.6 GHz in less than an hour. Genes with Bonferroni adjusted p-values below 1% are statistically significant. All code for the SEVA analyses is available from https://github.com/FertigLab/SEVA.

EBSeq is performed with the R/Bioconductor package EBSeq version 3.3 (Anders *et al.*, 2012). Isoform expression for all genes is the input in the EBSeq analysis. Isoforms with posterior probability above 99% were called significantly differentially spliced. EBSeq was also applied to gene expression values, and genes with a posterior probability above 99% were significantly differentially expressed. DiffSplice 0.1.2 beta version (Hu *et al.*, 2013) is run directly on aligned RNA-seq data obtained from the MapSplice alignment. Default parameters were used, with a false discovery rate of 0.01. Because DiffSplice requires equal numbers of samples in each group, we select a random subset of 14 HPV+ HNSCC and 14 normal samples from the dataset in (Guo *et al.*, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

rMATS version 3.2.5 (Shen *et al.*, 2014) is run on TopHat2 (Kim *et al.*, 2013) aligned simulated data. To produce read counts to use as reference annotation for rMATS, we constructed a merged gene annotation set by running cuffmerge v2.2.1 on the transcript predictions from the individual samples, produced with CLASS2 v.2.1.6 citepclass2, and the hg19 gene annotations used in the TCGA RSEM v2 pipeline (Cancer Genome Atlas Network, 2015). Read counts were generated with Ballgown (Frazee *et al.*, 2015a).

We perform cross study validation by comparing whether statistics in one cohort are significantly enriched in the other using the function wilcoxGST in LIMMA version 3.24.15 (Ritchie *et al.*, 2015).

3 Algorithm

3.1 Splice Expression Variation Analysis (SEVA)

A gene can have multiple isoforms. Each isoform has a unique pattern of junctions between exons or retained introns. In the case of ASEs, expression in junctions between exons or from exons to a retained intron measured with RNA-seq data provides direct evidence of gene isoform expression in each sample (Fig 1a). Because isoforms span multiple junctions, changes to isoform expression in a sample will impact the distribution of RNA-seq reads for multiple junctions simultaneously. Since the ASE variants can be specific to individual tumors, expression of ASEs can be be more variable in tumors than in normal samples. In these cases, the multivariate distribution of the set of junction expression can quantify gene-level dysregulation that can be associated with an ASE. This study aims to develop a new algorithm to detect differentially spliced genes that have such differential variability in the expression profile of all junctions.

Recently, a novel statistical method, EVA, was introduced for differential variability analysis of gene expression profiles (Afsari *et al.*, 2014a). This current study adapts the statistical framework from EVA to a new algorithm Splice Expression Variability Analysis (SEVA) that accounts for the multivariate changes of gene isoform expression patterns between phenotypes resulting in a new algorithm called SEVA. Briefly, SEVA quantifies the relative dissimilarity between profiles of junction expression in samples from the same phenotype by computing the average dissimilarity between all pairs of samples (denoted by D). Then, given two phenotypes, the algorithm tests whether there is a statistically significant difference in the level of variability of the expression profiles between the two phenotypes using an approximation from U-Statistics theory (Vaart, 1998) (Fig 1b, Supplemental File 1). The statistics from SEVA then depend upon the metric used to quantify the dissimilarity in the analysis.

The dissimilarity measure used for SEVA must account for the structure of gene isoforms. In alternative splicing, the multivariate distribution of the set of junction expression can quantify gene-level dysregulation that can be associated with an ASE. Moreover, junctions between exons form a splice graph that delineates all feasible gene isoforms (Song *et al.*, 2016; Hu *et al.*, 2013; Xing *et al.*, 2006). A gene isoform is a single path on this graph. Splicing regulation events changes the relative expression of the isoform, and, in turn, the relative coverage of all the junctions that overlap on the graph. In other words, overlapping junctions compete for read coverage from RNA-seq because an isoform can only go through one set of junctions on a single path on the splice graph. Therefore, SEVA computes the dissimilarity using expression profiles of the sets of "competing" junctions within a gene (Fig 1a).

SEVA defines a "splice dissimilarity" measurement for this analysis. The splice dissimilarity is based upon the Kendall-τ dissimilarity, a ranked-based metric, to quantify the relative variability of the multivariate distribution of a profile of junction expression for such dysregulation to model inter-class heterogeneity (Afsari *et al.*, 2014a). Specifically, the algorithm first computes dissimilarities for each junction (called a query junction). The dissimilarity for the query junction is calculated by computing the Kendall-τ dissimilarity for the expression profiles for all junctions that compete with the query junction (Fig 1a). The splice dissimilarity metric is a gene-level dissimilarity measure obtained by summing the Kendall-τ dissimilarities obtained for every junction in the gene.

The splice dissimilarity measurement is computed separately for all pairs of samples from each class that is compared (e.g., tumor and normal samples and shown for normal samples in Fig 1c). The expected dissimilarity of the profiles of junction expression in each phenotype provide a metric to quantify the relative heterogeneity of gene isoform usage in that phenotype, and serve as input to the comparison with U-Statistics theory described above, Moreover, basing the dissimilarity

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

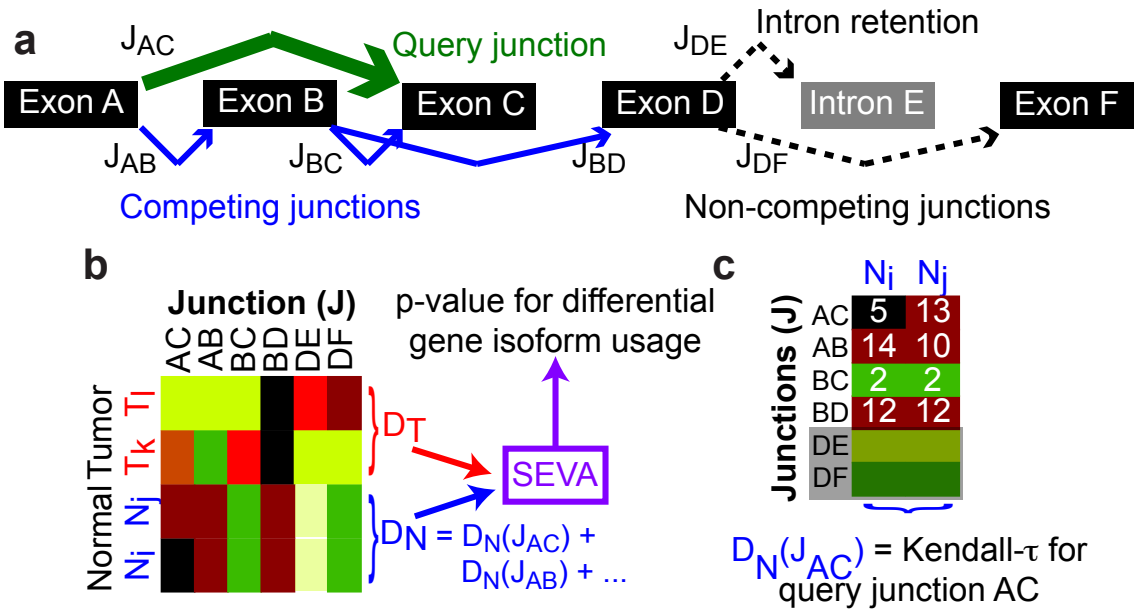

Figure 1: **SEVA schematic.** (a) Exons, retained introns, and junctions for gene with alternative isoform usage. (b) SEVA compares the expected dissimilarity of the expression for junction AC (J_{AC}) to junction DF (J_{DF}) in normal samples (blue) to that in tumor samples (red). (c) SEVA dissimilarity is the sum of Kendall-τ dissimilarity for each query junction (green in a) using a set of competing junctions (blue in a), excluding non-competing junctions (black in a and greyed in heatmap).

on the a rank-based Kendall-τ dissimilarity normalizes differences in junction expression from overexpression of a gene. Specifically, if the expression of a gene is higher in one sample than another, all its junctions will have higher expression but their ranks will not change. As a result, the analysis is blind to whether the gene is differentially expressed. Moreover, because the ranks are compared inside groups of competing junctions rather than between all the junctions in a gene, the measure is blind even for relative changes in coverage between parts of the same gene if they are not annotated as splicing alternatives.

4 Results

4.1 SEVA has greater accuracy than DiffSplice, EBSeq, or rMATS in identifying differential ASE candidates in simulated gene expression data with inter-tumor heterogeneity

We generate in silico RNA-seq data with known gene isoform usage to benchmark the performance of SEVA relative to EBSeq (Anders *et al.*, 2012), DiffSplice (Hu *et al.*, 2013), and rMATS (Shen *et al.*, 2014) in detecting true differential alternative splice events in populations of tumor and normal samples. In total, four simulated datasets are created with 10, 15, 20, or 25 of the total 25 tumor samples containing the differentially expressed and / or differentially spliced genes to test the recall of the algorithms to inter-tumor heterogeneity of gene isoform usage. We use these results to estimate the precision (positive predictive value) and recall (sensitivity) of different algorithms (Fig 2).

We apply SEVA, EBSeq, DiffSplice, and rMATS to detect differential splice status of genes in each simulated dataset. SEVA's precision remains around 95% while that of DiffSplice fluctuates around 90% and the range of EBSeq's precision is 60%-80%. The precision for rMATS is around 90% for differentially spliced genes that are not differentially expressed and ranges from 75% to 90% for differentially spliced genes that are also differentially expressed. These results are independent of the number of cancer samples containing the alternative gene isoform expression (Fig 2a). The precision of both SEVA and DiffSplice is independent of whether the gene is differentially expressed in addition to differentially spliced. EBSeq has lower precision for detecting differential splice status

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

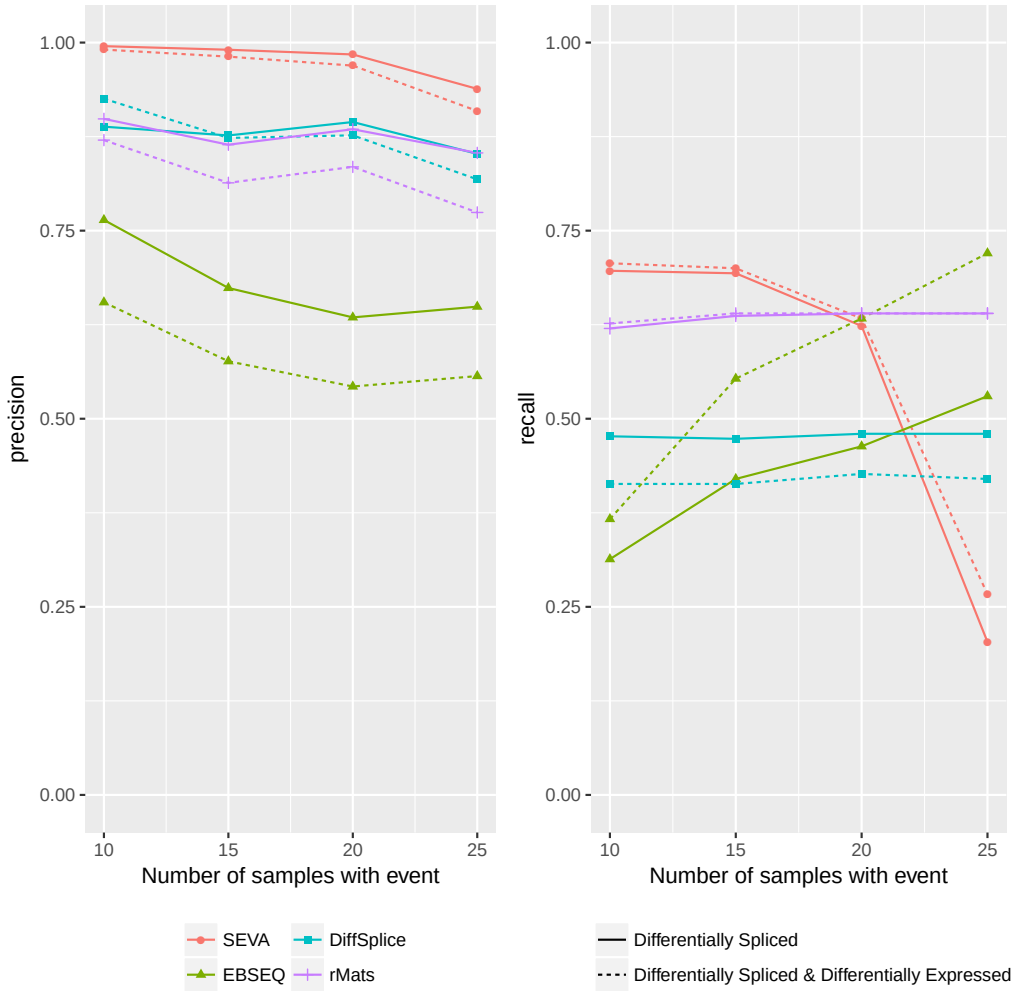

Figure 2: **Algorithm comparison in simulated RNA-Seq data.** (a) Precision for algorithms in legend. Varying numbers of the total tumor samples with alterations (x-axis), with inter-tumor heterogeneity decreasing along the x-axis. Precision computed separately for differentially spliced genes not differentially expressed (solid) and genes that are (dashed). (b) Recall of simulated data, as in (a).

among differentially expressed genes compared that among a mixed pool of differentially and non-differentially expressed genes.

SEVA has the highest recall when alternative gene isoform expression occurs in fewer than 20 of the tumor samples, but drops sharply in the more homogeneous population of 25 tumor samples all containing the same gene isoform usage (Fig 2b). This performance is consistent with the construction of SEVA to specifically identify genes with high relative heterogeneity of gene isoform usage between sample phenotypes, in contrast to other algorithms that seek genes with homogeneous differential gene isoform usage. The recall for EBSeq remains consistently higher among differentially expressed genes than among a mixed pool of genes, and increases with the number of tumor samples containing the alternative isoform usage. On the other hand, EBSeq has lower recall for differentially spliced genes among the mixed pool than that among genes only differentially expressed genes, and remains below 50% regardless of the number of tumor samples with alternative splice events. The recall for rMATS remains between 60% and 65%, independent of the sample size and differential expression status of the genes. Both DiffSplice (below 50% for non-differentially expressed genes and approximately 40% for differentially expressed) and rMATS (between 60% and 65% for all genes) have modest recall independent of tumor heterogeneity in the simulations. Taken together, the simulations suggest that the performance of SEVA is particularly robust to heterogeneity in gene isoform usage in cancer samples.

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.2 SEVA identifies a robust set of ASEs in non-differentially expressed genes from RNA-sequencing data for HPV+ HNSCC tumors and normal samples

We use RNA-seq data for 46 HPV+ HNSCC and 25 normal samples from Guo *et al.*, 2016 as a benchmark for empirical analysis of SEVA in real sequencing data. SEVA identified 274 genes as having significant alternative gene isoform usage in cancer. In addition to identifying the altered gene isoforms in each class, the statistics underlying SEVA enable quantification of relative variation in isoform usage for each gene in each of the phenotypes that are compared in the analysis. We plot these statistics to compare variation of isoform usage in all genes and the genes that SEVA calls statistically significant to test our central hypothesis that gene isoform usage is more variable in tumor than normal samples (Fig 3a). Consistent with our hypothesis, the variation in all genes is shifted towards higher variation in tumor samples. Moreover, more of these significant genes have more variable gene isoform usage in tumor samples than in normal samples (Fig 3a).

We also apply EBSeq and DiffSplice to these same data to compare with EVA (Fig 3b and c). We select these algorithms for analysis comparison because they can be run on MapSplice (Wang *et al.*, 2010) aligned data in TCGA (Cancer Genome Atlas Network, 2015), and therefore do not introduce the alignment algorithm as an additional variable in the comparison study. EBSeq and DiffSplice methods both identify far more genes with alternative isoform expression than SEVA (n=2439 and n=2535), which makes them more prone to false positives and hinders candidate selection for further experimental validation. Moreover, EBSeq identifies higher portion of differentially expressed genes as differentially spliced (40%) than either SEVA (5%) or DiffSplice (13%), indicating the potential for more false-positive hits for alternative splicing events. However, the proportion of differential expressed genes among the identified genes are similar between EBSeq (87%) and SEVA (84%) and lower in DiffSplice (30%). Since the ground truth is unknown in this real data, we cannot assess the true independence of differentially spliced and differentially expressed genes.

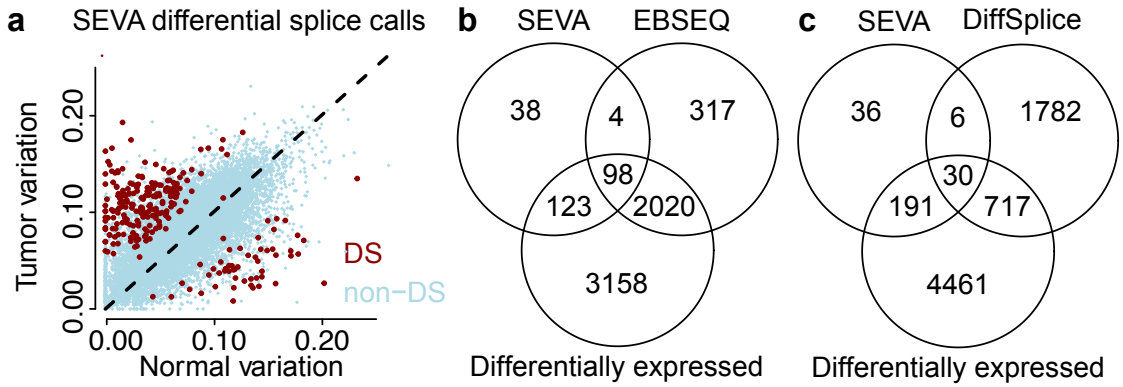

Figure 3: **Comparison of differential gene isoform usage algorithms in real HPV+ HNSCC RNA-seq data.** (a) Variability of junction expression profiles in corresponding to gene isoforms. Each point represents a gene, x-axis and y-axis its variability computed for SEVA in normal vs cancer, respectively. The red points represent significantly differentially spliced (DS) genes identified with SEVA, and blue genes that were not significantly spliced (non-DS). (b) Venn diagram comparing differentially spliced genes identified by SEVA and EBSeq, as well as differential expression status of each gene. (c) Comparison of SEVA and DiffSplice as described in (b).

4.3 SEVA analysis finds greater variation in tumor than normal samples in previously validated HNSCC-specific splice variants TP63, LAMA3, and DST

Recent data suggests that the majority of ASE in HNSCC (39%) are classified as alternative start sites (manuscript under review), which can be recognized by ASE-detection algorithms as insertion and/or deletion alternative splicing events. Indeed, alternative start site splice events in six genes (VEGFC, DST, LAMA3, SDHA, TP63, and RASIP1) were recently observed as being unique to

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

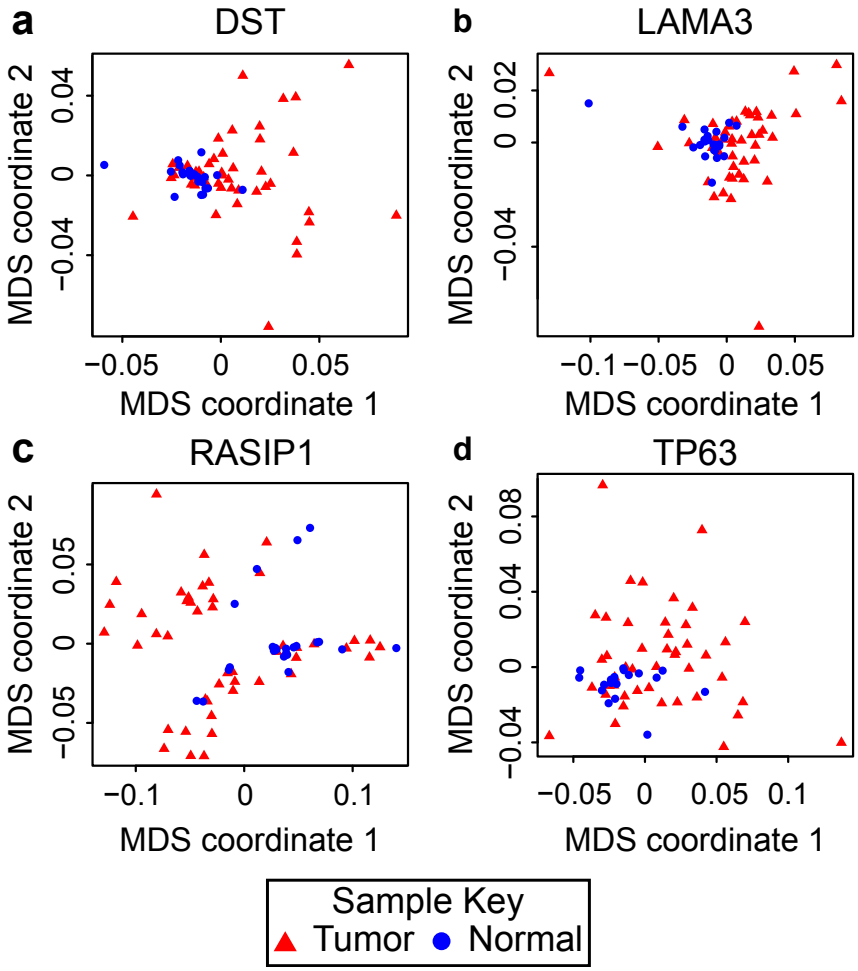

Figure 4: **Multidimensional scaling (MDS) plot of splice dissimilarity measures in real HPV+ HNSCC junction expression from RNA-seq** for (a) DST, (b) LAMA3, (c) RASIP1, and (d) TP63. Relative spread of samples in the MDS plots indicates their relative variability in normal samples (blue circles) and tumor samples (red triangles).

HNSCC samples from microarray data (Li *et al.*, 2014). Three of these genes (DST, LAMA3 and TP63) were also confirmed as differentially spliced in HNSCC tumors with experimental validation in an independent cohort of samples, while the other three genes (VEGFC, SDHA, and RASIP1) were not confirmed (Li *et al.*, 2014). SEVA identified all three validated DST (p-value 3×10^{-10}), LAMA3 (p-value 1×10^{-10}), and TP63 (p-value 6×10^{-10}) genes, as well as RASIP1 (5×10^{-7}) as differentially spliced genes. Splicing of SDHA and VEGFC were found non-significant in agreement with experiments. Notably, EBSeq only identifies VEGFC to be differentially spliced (p-value 3×10^{-6}) that was not validated. DiffSplice did not identify significant alternative splicing any any of these genes.

Calculation of the splice dissimilarity measure in SEVA makes this algorithm unique in being able to both quantify and visualize the relative heterogeneity of gene isoform usage in each phenotype. In Fig 4, we create multi-dimensional scaling (MDS) plots of the splice dissimilarity measure of the significant genes in SEVA. The closer two samples in the MDS plot, the less variable their junction expression profiles. As a result, these figures enable us to visually test the hypothesis that the modified Kendall-τ distance enables SEVA to identify more variable gene isoform usage in tumor than normal samples. The differentially spliced genes identified with SEVA (DST, LAMA3, RASIP1, and TP63) confirm that the variation of gene isoform expression in normal samples is lower than the variation in tumor samples, as hypothesized, and therefore not significant in EBSeq (Fig 4). On the other hand, VEGFC has consistent variability in cancer and normal samples and was not detected by SEVA (Supplemental Fig 2). Therefore, SEVA is ideally suited to detect genes with more variable gene isoform usage in tumor relative to normal samples as hypothesized.

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.4 SEVA candidates in the training set are significantly enriched in cross-study validation on TCGA data

We also apply SEVA to independent RNA-seq data for 44 HPV+ HNSCC and 16 normal samples in TCGA (Cancer Genome Atlas Network, 2015) to cross-validate the ASE candidates in the training data from Guo *et al.*, 2016. 32% (70 out of 220 the gene candidates) of the hits are statistically significant in the SEVA analysis of the TCGA data. 43 of the genes identified on the training set were not expressed on the TCGA set. To test the significance of the list of genes and consistency of SEVA across two data sets, we check whether the ASE candidates from training set are significantly enriched on the TCGA data. To do so, we calculate the SEVA p-values for all genes on the TCGA test set. A mean-rank gene set analysis indicates that the candidate genes identified on training are enriched among all genes with p-value $< 2 \times 10^{-12}$.

4.5 SEVA identifies more heterogeneity in splice variant usage in HPV-HNSCC samples with mutations in splice machinery genes

HNSCC tumors have two predominant subtypes: HPV+ and HPV-. HPV- tumors have greater genomic heterogeneity than HPV+ (Cancer Genome Atlas Network, 2015; Mroz *et al.*, 2015). Therefore, we apply SEVA to test whether there is higher inter-tumor heterogeneity in splice variant usage in these tumor subgroups. SEVA observes many genes as having alternative gene isoform usage between the two HNSCC subtypes, but no difference in relative heterogeneity (Fig 5a). This finding occurs although a larger number of samples HPV- tumors (44 of 243, 18%) have alterations to genes in the splice variant machinery in contrast to HPV+ (3 of 36 with sequencing data, 8%). We further apply SEVA to compare inter-tumor heterogeneity of gene isoform usage in HPV- tumors with and without genetic alterations to the splice machinery. We observe far fewer significant genes than in this comparison (Fig 5b). Nonetheless, the significant genes from this analysis have greater variability in samples with alterations in the splice variant machinery. These findings suggest that although the mechanisms of tumorigenesis are vastly different in HPV+ and HPV- tumors, both have similar heterogeneity in gene isoform usage. The mechanisms that cause mutations to the genes that encode for the splice variant machinery further increase the heterogeneity of splice variant usage in HPV- HNSCC.

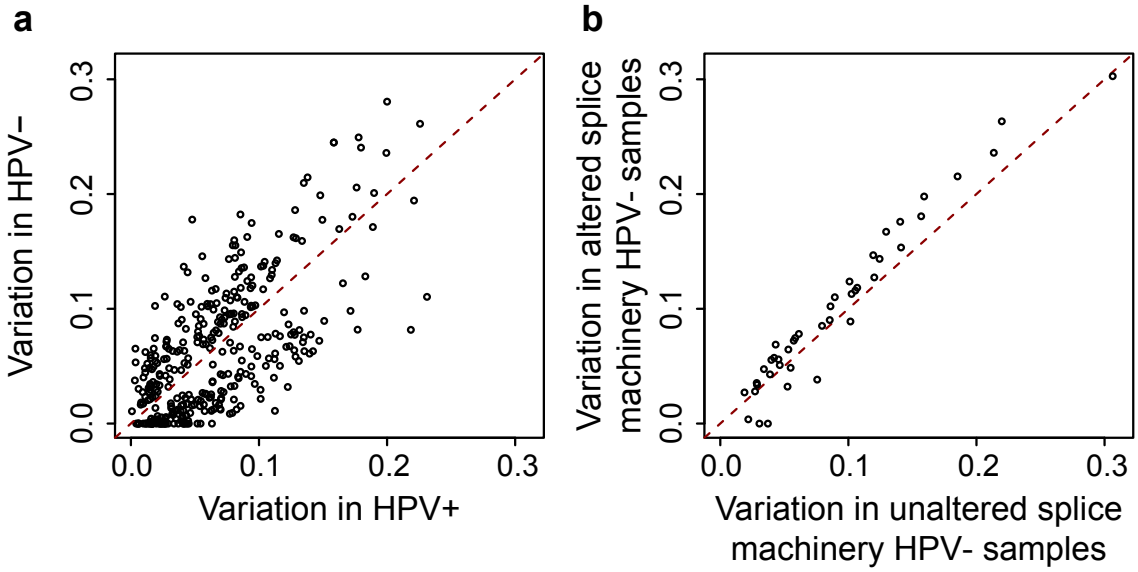

Figure 5: **Comparison of differential gene isoform in TCGA HNSCC RNA-seq data.** (a) Variability of junction expression profiles for genes significantly DS from SEVA in HPV+ vs HPV- HNSCC, respectively. (b) As for (a) comparing HPV- samples with and without alterations in RNA splice machinery genes.

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5 Discussion

In this study, we develop SEVA to identify holistic and multivariate changes in isoform expression in tumor samples with heterogeneous gene isoform usage. Consistent with its formulation, we observe that SEVA has higher precision than EBSeq (Leng *et al.*, 2013), DiffSplice (Hu *et al.*, 2013), or rMATS (Shen *et al.*, 2014) in simulated datasets that reflect tumor heterogeneity. The precision of SEVA, DiffSplice, and rMATS remain independent of the heterogeneity of gene isoform usage in the tumor samples, whereas that of EBSeq decreases with increasing homogeneity in gene isoform usage in the tumor samples. While SEVA retains a lower false positive rate in the simulated data, the recall depends on the heterogeneity of gene isoform usage. In our simulations, as the ratio of disrupted samples in the cancer batch increases, the recall of SEVA reduces dramatically (from 0.7 to 0.2). DiffSplice and rMATS show almost constant recall (around 40-50% and 50-60%, respectively). While EBSeq recall increases with the homogeneity in gene isoform usage, SEVA loses its recall when gene isoform usage is greater than 80%. SEVA performs relatively best in the case of high heterogeneity of junction expression in the tumor population. Notably, as the number of cancers with an ASE increases the junction expression profiles are more homogeneous and therefore not accurately detected with SEVA. We hypothesize that SEVA will have lower recall than techniques based upon differential isoform expression in populations with homogeneous isoform usage. However, cancer samples are more heterogeneous and encompasses a bigger spectrum of subtypes (Afsari *et al.*, 2014a; Corrada Bravo *et al.*, 2012; Eddy *et al.*, 2010). In practice, we hypothesize that differentially spliced genes show multiple patterns of isoform expression in tumors in multiple different cancer subtypes. Therefore, in practice we anticipate far less than 80% homogeneity in driver splice events in cancer. Based upon the simulated data and pervasive genomic heterogeneity in tumors, we hypothesize that SEVA is uniquely suited to identify clinically relevant gene isoform usage in tumors and their subtypes. Moreover, it also the only algorithm that is able to quantify the extent of such heterogeneity of isoform usage for each gene.

Whereas the other algorithms compare mean expression, the ranked-based nature of the modified Kendall-τ in SEVA is blind to such coordinated changes without further normalization of gene expression values (Anders *et al.*, 2012). SEVA finds candidates that are independent of the differential expression status of the gene in contrast to EBSeq or DiffSplice. Therefore, we hypothesize that ASE candidates from SEVA are uniquely independent of differential expression status of genes when the independence between these events is the ground truth. As a result, SEVA is also uniquely suited to quantify differential gene isoform usage between sample groups that have numerous differentially expressed genes, such as tumor and normal samples. The algorithm compares differences in variation in isoform usage between the two groups and does not rely on samples of either group as a reference. Therefore, it can also be applied to compare gene isoform usage in samples from two distinct tumor subtypes.

SEVA, EBSeq, and DiffSplice were all applied to RNA-seq data (Guo *et al.*, 2016) normalized with MapSplice (Wang *et al.*, 2010) to enable cross-study validation with the TCGA normalized data. SEVA inputs junction expression to use direct evidence of alternative splice usage and intron retention. Therefore, the algorithm can readily be applied to junction expression obtained from other aligners. It is also directly applicable to estimates of percent spliced (Alamancos *et al.*, 2015) in place of junction expression, which also be compared in future studies. While there are numerous other algorithms for such differential splice analysis, many rely on data obtained from distinct alignment and normalization pipelines (Song *et al.*, 2016; Anders *et al.*, 2012; Shen *et al.*, 2012). These preprocessing techniques may introduce additional variables into the differential splice variant analysis, complicating the direct comparisons of gene candidates on *in silico* and RNA-seq datasets presented in this paper. Therefore, future studies are needed to compare the performance of such differential splice variant algorithms across normalization pipelines on real biological datasets with known ground truth of gene isoform usage. Nonetheless, the SEVA algorithm is applicable for differential splice variant analysis from junction expression from any alignment algorithm and its rank-based statistics make it likely to be independent of the normalization procedure (Afsari *et al.*, 2014b; Bolstad *et al.*, 2003).

SEVA has uniformly high precision relative to EBSeq and DiffSplice in detecting ASEs from *in silico* data. Our simulations suggest that SEVA performs better in scenarios in which cancer samples have higher degree of heterogeneity compared to normal samples. As further validation, genes with alternative splicing events in HPV+ HNSCC from Guo *et al.*, 2016 were significantly enriched in cross-study validation on RNA-seq data for HPV+ HNSCC samples in TCGA (Cancer

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genome Atlas Network, 2015). Moreover, the modified Kendall-τ dissimilarity metric used in SEVA also accurately characterizes higher heterogeneity of gene isoform usage in tumors relative to normal in the confirmed HNSCC-specific ASEs DST, LAMA3, and TP63, and also RASIP1 identified in previous microarray analysis (Li *et al.*, 2014). Therefore, SEVA is a novel algorithm adept at inferring ASEs in tumor samples with heterogeneous gene isoform usage relative to normal samples.

SEVA is also unique in its ability to quantify which phenotype has more variable gene isoform usage. The algorithm observes higher inter-tumor heterogeneity in splice variant usage in HPV+ HNSCC tumors relative to normal samples, consistent with the hypothesis of inter-tumor heterogeneity that led to the development of the algorithm. HNSCC is divided into two primary subtypes (HPV- and HPV+). Of these, HPV- tumors are established as having more variable genetic alterations than HPV+ tumors (Cancer Genome Atlas Network, 2015; Mroz *et al.*, 2015). Nonetheless, SEVA analysis identifies no difference in the heterogeneity of splice variant usage between the tumor types. This similarity is observed in spite of different samples sizes (44 HPV+ and 235 HPV-), suggesting that the SEVA statistics are robust to imbalanced study design. Nonetheless, HPV- samples have a greater rate of genetic alterations to genes regulating the splice variant machinery. SEVA identifies that HPV- HNSCC tumors with alterations in these genes have greater variation in isoform usage than those that do not. Together, these analyses suggest that the mechanisms of tumorigenesis introduce substantial inter-tumor heterogeneity in splice variant usage specific to each cancer subtype. Variation in splice variant usage is further enhanced by genetic alterations to the splice variant machinery (Ebert and Bernard, 2011) within a specific cancer subtype. Further pan-cancer and pan-genomics analyses are essential to distinguish the relative impact of tumorigenesis and alterations to splice variant machinery on tumor-specific alternative gene isoform usage in cancer and the functional impact of these splice variants on tumor progression and therapeutic response.

Acknowledgements

We thank Simina Boca, Leslie Cope, Ludmila V Danilova, and Sarah Wheelan for advice on analyses, data preprocessing, data access, and suggestions.

Funding

The authors' research was supported by: National Institutes of Health: National Cancer Institute [P30 CA006973 to BA, MC, AVF, and EJF, R01 CA177669 to EJF], National Institute on Deafness and Other Communication Disorders [T32DC000027-26 to TG], National Institute of Dental and Craniofacial Research [R21 DE025398 to DAG, R01 DE023347 to JAC, P50 DE019032 pilot funding to DAG and EJF, R01 DE023227 to PH], National Library of Medicine [R01LM011000 to MFO], The National Science Foundation [ABI-1356078 to LF], and The Adenoid Cystic Carcinoma Research Foundation [to EJF, PH]. Funding for open access charge: National Institute of Dental and Craniofacial Research.

References

Afsari, B., Geman, D., and Fertig, E. J. (2014a). Learning dysregulated pathways in cancers from differential variability analysis. *Cancer Inform*, **13**(Suppl 5), 61–7.

Afsari, B., Neto, U. B., and Geman, D. (2014b). Rank discriminants for predicting phenotypes from rna expression. *Annals of Applied Statistics*, **8**(3), 1469–1491.

Alamancos, G. P., Pages, A., Trincado, J. L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**(9), 1521–31.

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Genome Res*, **22**(10), 2008–17.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–93.

Cancer Genome Atlas Network (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, **517**(7536), 576–582.

Canzar, S., Andreotti, S., Weese, D., Reinert, K., and Klau, G. W. (2016). Cidane: Comprehensive isoform discovery and abundance estimation. *Genome biology*, **17**(1), 1.

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Chen, J. and Weiss, W. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, **34**(1), 1–14.

Corrada Bravo, H., Pihur, V., McCall, M., Irizarry, R. A., and Leek, J. T. (2012). Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics*, **13**(1), 1–11.

Ebert, B. and Bernard, O. A. (2011). Mutations in rna splicing machinery in human cancers. *N Engl J Med*, **365**(26), 2534–5.

Eddy, J. A., Hood, L., Price, N. D., and Geman, D. (2010). Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac). *PLoS Comput Biol*, **6**(5), e1000792.

Feng, H., Qin, Z., and Zhang, X. (2013). Opportunities and methods for studying alternative splicing in cancer with rna-seq. *Cancer letters*, **340**(2), 179–191.

Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015a). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature biotechnology*, **33**(3), 243–246.

Frazee, A. C., Jaffe, A. E., Langmead, B., and Leek, J. T. (2015b). Polyester: simulating rna-seq datasets with differential transcript expression. *Bioinformatics*, **31**(17), 2778–2784.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci Signal*, **6**(269), pl1.

Guo, T., Gaykalova, D. A., Considine, M., Wheelan, S., Pallavajjala, A., Bishop, J. A., Westra, W. H., Ideker, T., Koch, W. M., and Khan, Z. (2016). Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. *International journal of cancer*, **139**(2), 373–382.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., and Nusbaum, C. (2010). Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincrnas. *Nature biotechnology*, **28**(5), 503.

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., Monroy, A., Kuan, P.-F., Hammond, S. M., and Makowski, L. (2013). Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic acids research*, **41**(2), e39–e39.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S. M., Fu, B., Lin, M. T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E., and Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**(5897), 1801–6.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**(4), R36.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, **29**(8), 1035–1043.

Li, J. J., Jiang, C.-R., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Sparse linear modeling of next-generation mrna sequencing (rna-seq) data for isoform discovery and abundance estimation. *Proceedings of the National Academy of Sciences*, **108**(50), 19867–19872.

Li, R., Ochs, M. F., Ahn, S. M., Hennessey, P., Tan, M., Soudry, E., Gaykalova, D. A., Uemura, M., Brait, M., and Shao, C. (2014). Expression microarray analysis reveals alternative splicing of lama3 and dst genes in head and neck squamous cell carcinoma. *PloS one*, **9**(3), e91263.

Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., and Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mrna processing defects in human genes. *Proceedings of the National Academy of Sciences*, **108**(27), 11093–11098.

Liu, R., Loraine, A. E., and Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using rna-seq in plant systems. *BMC Bioinformatics*, **15**, 364.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, **110**(11), 4245–4250.

Mroz, E. A., Tward, A. D., Hammon, R. J., Ren, Y., and Rocco, J. W. (2015). Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the cancer genome atlas. *PLoS Med*, **12**(2), e1001786.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, **33**(3), 290–295.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

Sebestyen, E., Zawisza, M., and Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res*, **43**(3), 1345–56.

bioRxiv preprint doi: https://doi.org/10.1101/091637; this version posted June 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P., and Xing, Y. (2012). Mats: a bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic acids research*, page gkr1291.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rmats: robust and flexible detection of differential alternative splicing from replicate rna-seq data. *Proc Natl Acad Sci U S A*, **111**(51), E5593–601.

Song, L., Sabunciyan, S., and Florea, L. (2016). Class2: accurate and efficient splice variant annotation from rna-seq reads. *Nucleic acids research*, page gkw158.

Thomas, R. K., Baker, A. C., Debiasi, R. M., Winckler, W., Laframboise, T., Lin, W. M., Wang, M., Feng, W., Zander, T., MacConaill, L., Lee, J. C., Nicoletti, R., Hatton, C., Goyette, M., Girard, L., Majmudar, K., Ziaugra, L., Wong, K. K., Gabriel, S., Beroukhim, R., Peyton, M., Barretina, J., Dutt, A., Emery, C., Greulich, H., Shah, K., Sasaki, H., Gazdar, A., Minna, J., Armstrong, S. A., Mellinghoff, I. K., Hodi, F. S., Dranoff, G., Mischel, P. S., Cloughesy, T. F., Nelson, S. F., Liau, L. M., Mertz, K., Rubin, M. A., Moch, H., Loda, M., Catalona, W., Fletcher, J., Signoretti, S., Kaye, F., Anderson, K. C., Demetri, G. D., Dummer, R., Wagner, S., Herlyn, M., Sellers, W. R., Meyerson, M., and Garraway, L. A. (2007). High-throughput oncogene mutation profiling in human cancer. *Nat Genet*, **39**(3), 347–51.

Vaart, A. W. v. d. (1998). *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK ; New York, NY, USA.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., and Perou, C. M. (2010). Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic acids research*, **38**(18), e178–e178.

Xing, Y., Yu, T., Wu, Y. N., Roy, M., Kim, J., and Lee, C. (2006). An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic acids research*, **34**(10), 3150–3160.