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

Bahman Afsari[+], Theresa Guo[+], Michael Considine, Liliana Florea, Dylan Kelley, Emily Flam, Patrick K. Ha, Donald Geman, Michael F. Ochs, Joseph A. Califano, Daria A. Gaykalova[*], Alexander V. Favorov[*], Elana J. Fertig[*]
+ Co-first * Co-corresponding: dgaykal1@jhmi.edu, favorov@sensi.org, ejfertig@jhmi.edu

ABSTRACT

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 in tumors. 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.

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[1]. 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[2]. 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[3]. 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[4-9]. These tools can systematically evaluate the gene isoforms that are expressed in a sample. Other techniques have been developed to utilize isoforms as cancer biomarkers[10]. Nonetheless, in order to characterize the landscape of splicing events specific to cancer, it is essential to perform analysis to identify splice variants that are uniquely expressed in RNA-seq data from tumor samples compared to normal tissue. Most reported techniques to define differential ASE expression rely on comparing mean expression values to determine differences in ASE expression between clinical variables, such as normal and tumor samples[8,11-14]. In spite of the breadth of available ASE algorithms, few have been validated in primary tumor samples[15].

In primary tumors, 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 a common pathway[16,17]. 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.

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[11], DiffSplice[12], and rMATS, in simulated RNA-seq data generated with Polyester[18]. 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 publically 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 OPSCC tumor samples that include experimentally validated splice variants in HNSCC from a previously microarray study[19]. 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[20]. Therefore, SEVA represents a robust algorithm that is well suited to find inter-tumor heterogeneity in gene isoform usage relative to normal samples.

MATERIAL AND METHODS

Splice Expression Variation Analysis (SEVA)

Expression of alternative splice variants in a cancer sample can alter the expression pattern of all the isoforms of that gene. Since the ASE variants can be specific to individual tumors, expression of ASEs can be be more variable in tumors than normal samples. We call a gene with such differential variability in exon junction expression Differentially Spliced (DS).

Recently, a novel statistical method, EVA, was introduced for differential variability analysis of gene expression profiles[21]. This current study adapts the two inputs to EVA to account for the multivariate changes of gene isoform expression patterns between phenotypes resulting in a new algorithm called SEVA. In the case of ASEs, expression in junctions between exons or from exons to a retained intro measured with RNA-seq data provides direct evidence of gene isoform expression in each sample (Fig 1a). Therefore, SEVA takes the profile of junction expression of each gene as input. 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[22] (Fig 1b, Supplemental File 1).

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[4,12,23]. Therefore, SEVA computes the dissimilarity using expression profiles of the sets of "competing" junctions within a gene (Fig 1a). Previously, we showed that the Kendall-τ dissimilarity, a ranked-based metric, can quantify the relative variability of the multivariate distribution of a profile of gene expression for such dysregulation to model inter-tumor heterogeneity[21]. SEVA calculates a gene-level dissimilarity measure for each phenotype by summing the measures obtained for each junction within a gene (shown for normal samples in Fig 1c). Using a rank-based dissimilarity normalizes differences in junction expression from overexpression of a gene, and makes the analysis blind to whether the gene is differentially expressed.

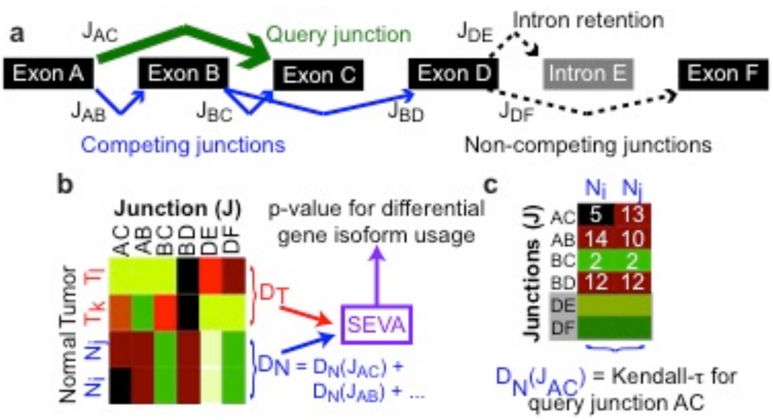

Fig 1 SEVA schematic. (a) Exons, retained introns, and junctions for gene with alternative isoform usage. (b) SEVA compares the expected dissimilarity of the expression for J_{AC} to J_{DF} in normal samples (blue) to that in tumor samples (red). (c) SEVA dissimilarity is the sum of Kendall-τ dissimilarity 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).

In silico data

To simulate isoform expression, we used the expression of isoforms from the HPV+ RNA-seq data from [24]. For efficiency of the simulations, we selected genes in chromosome 1 whose expression 4 to 9 in log 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[18]. Genes in tumor samples are simulated as DS by exchanging junction expression from the normal samples randomly between junctions in isoforms of the gene, analogous to previous differential gene isoform simulation studies[25,26]. 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.

HPV+ HNSCC and normal RNA-seq datasets

We use RNA-seq data for 46 HPV+ OPSCC and 25 independent normal samples from uvulopalatopharyngoplasty previously described in [24] and 44 HPV+ HNSCC and 16 matched normal tissues from the freeze set for TCGA HNSCC[27].

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[27]. Junction expression is obtained directly from the MapSplice[9] 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[28] and junction data is obtained similarly to MapSplice. 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 [24]. 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[29] are said to have altered RNA splice machinery based upon annotations in [20].

Implementation and software

SEVA is implemented in the R/Bioconductor package GSReg[21]. 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.

EBSeq is performed with the R/Bioconductor package EBSeq version 3.3[11]. 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[12] 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 of each group, we select a random subset of 14 HPV+ OPSCC and 14 normal samples from the dataset in [24]. rMATS version 3.2.5[14] is run on TopHat2[28] aligned simulated data.

We perform cross study validation by comparing whether statistics in on cohort are significantly enriched in the other using the function wilcoxGST in LIMMA version 3.24.15.

RESULTS

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

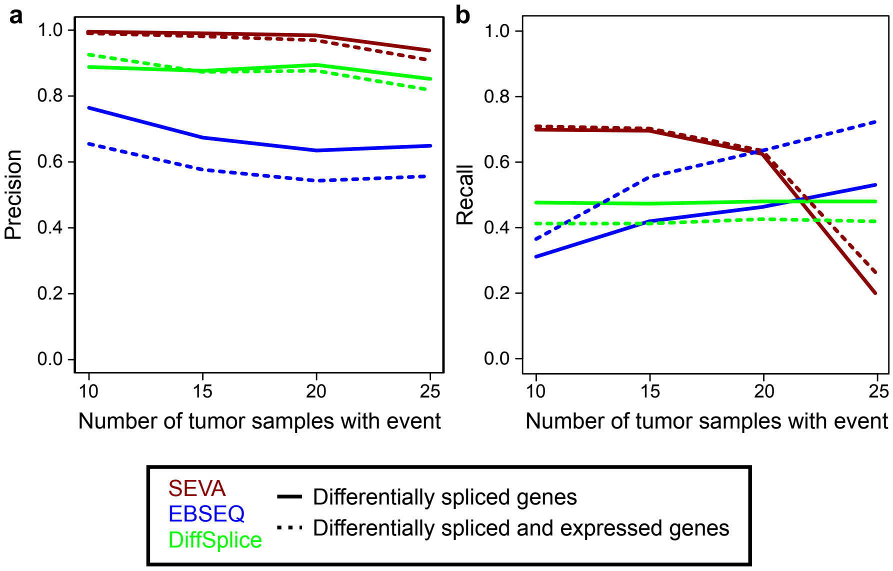

Fig 2 Algorithm comparison in simulated RNA-Seq data. (a) Precision for algorithms in legend. Varying numbers of the total tumor samples with alterations (x-axis), with inter-tumor heterogeneity decreasing along the x-axis. Precision computed separately for differentially spliced genes not differentially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

We generate *in silico* RNA-seq data with known gene isoform usage to benchmark the performance of SEVA relative to EBSeq[11], DiffSplice[12], and rMATS[14] in detecting true differential alterative splice events in populations of tumor and normal samples. We select these algorithms for analysis comparison because they can be run on MapSplice[9] aligned data in TCGA[27], and therefore do not introduce the alignment algorithm as an additional variable in the comparison study. 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.

We apply SEVA, EBSeq, DiffSplice, and rMATS to detect DS status of genes in each simulated dataset. SEVA's precision remains around 95% while that of DiffSplice fluctuates around 90% and the precision of EBSeq's ranges is 60%-80%. 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 DS status among differentially expressed genes compared that among a mixed pool of differentially and non-differentially expressed genes. rMATS has low precision for both non-differentially expressed (6%) and differentially expressed genes (5%).

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. Taken together, the simulations suggest that the performance of SEVA is particularly robust to heterogeneity in gene isoform usage in cancer samples. Both DiffSplice (below 50% for non-differentially expressed genes and approximately 40% for differentially expressed) and rMATS (60% for non-differentially expressed genes and 50% differentially expressed) have modest recall independent of tumor heterogeneity in the simulations.

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

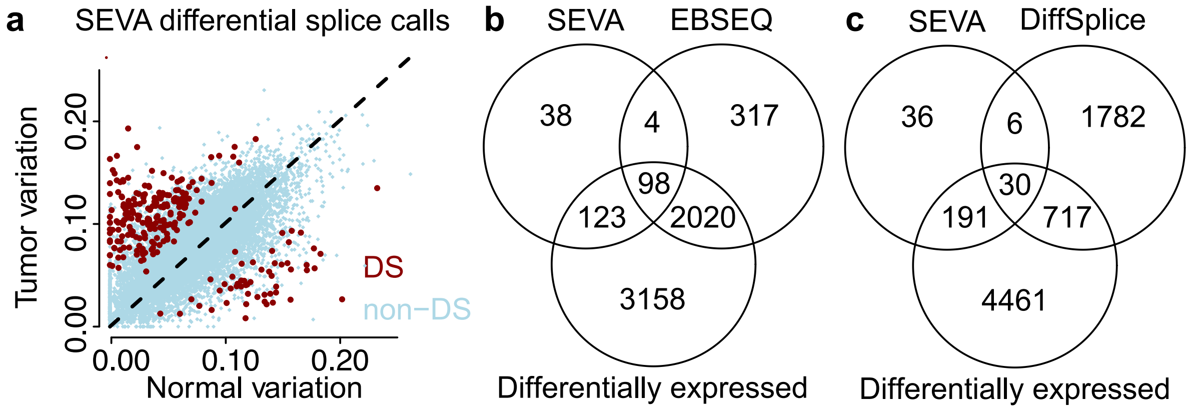

Fig 3. Comparison of differential gene isoform usage algorithms in real HPV+ HNSCC RNA-seq data. (a) Variability of junction expression profiles in corresponding to gene isoforms. Each point represents a gene, x-axis and y-axis its variability computed for SEVA in normal vs cancer, respectively. The red points represent significantly differentially (DS) spliced 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).

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 isform 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). 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-positives 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.

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

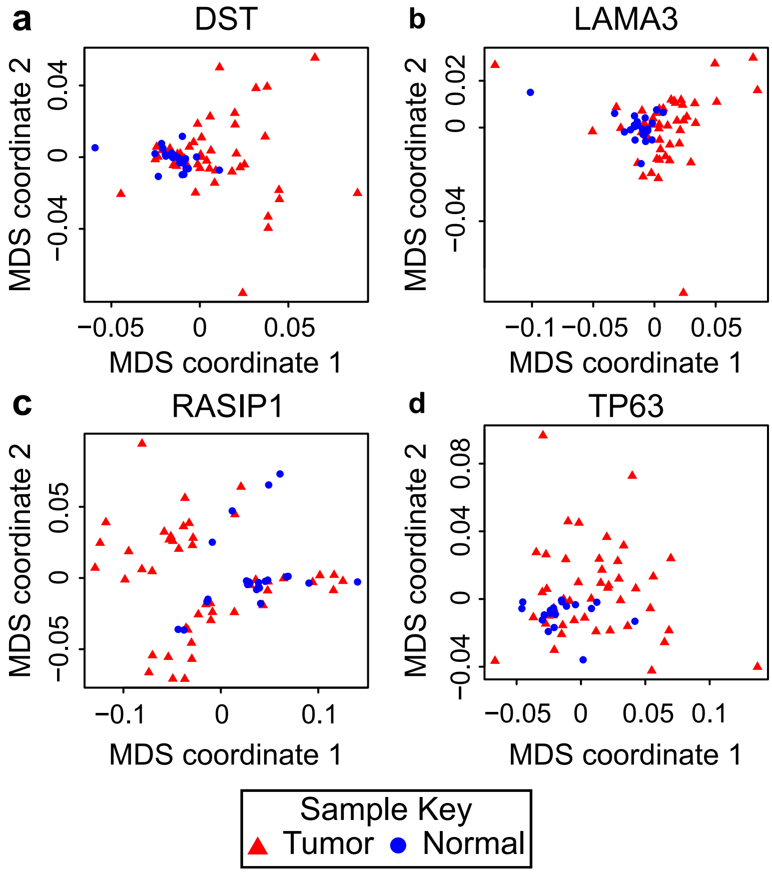

Fig 4. Multidimensional scaling (MDS) plot of modified Kendall-τ distances in real HPV+ HNSCC junction expression from RNA-seq for (a) DST, (b) LAMA3, (c) RASIP1, and (d) TP63. Relative spread of samples in the MDS plots indicates their relative variability in normal samples (blue circles) and tumor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

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 HNSCC samples from microarray data[19]. 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[19]. SEVA identified all three validated *DST* (p-value $3x10^{-10}$), LAMA3 (p-value $1x10^{-10}$), and *TP63* (p-value $6x10^{-10}$) genes, as well as *RASIP1* ($5x10^{-7}$) 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 3x10^{-6}) that was not validated and DiffSplice identified none.

In Fig 4, we create multi-dimensional scaling (MDS) plots of the modified Kendall-τ distance of the significant genes in SEVA. The closer two samples in this 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 normal samples are closer to each other than cancer samples to each other, 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. Therefore, SEVA is ideally suited to detect genes with more variable gene isoform usage in tumor relative to normal samples as hypothesized.

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[27] to cross-validate the ASE candidates in the training data from [24]. 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 2.2e-12.

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

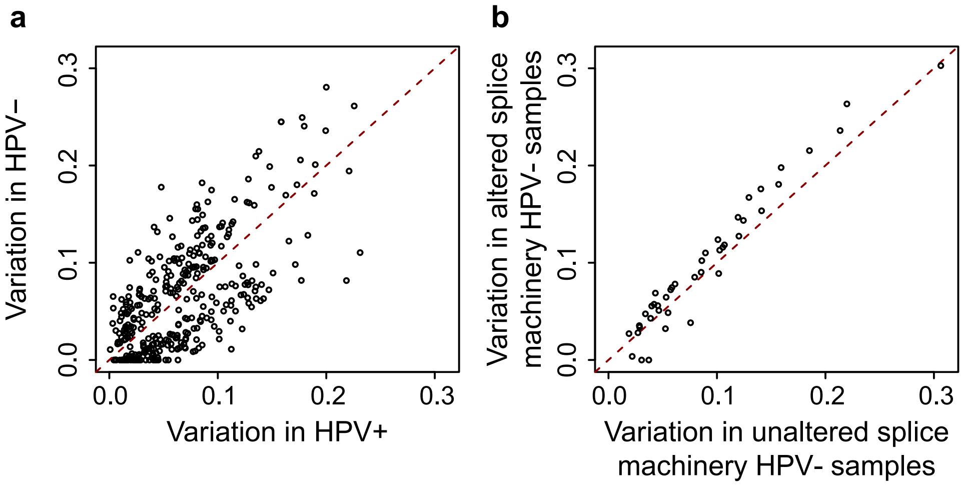

Fig 5. Comparison of differential gene isoform in TCGA HNSCC RNA-seq data. (a) Variability of junction expression profiles for genes significantly DS from SEVA in HPV+ vs HPV-, respectively. (b) As for (a) comparing HPV- samples with and without alterations in RNA splice machinery genes.

HNSCC tumors have two predominant subtypes: HPV+ and HPV-. HPV- tumors have greater genomic heterogeneity than HPV+[27,30]. 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 the introduction of heterogeneous splice variant machinery arise from differences in tumorigenesis in the two subtypes, and can be enhanced within a subtype by gene alterations to the splice variant machinery.

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[31], DiffSplice[12], or rMATS[14] in simulated datasets that reflect this 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. In addition, SEVA finds candidates that are independent of the differential expression status of the gene in contrast to EBSeq or DiffSplice in simulated data. 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. 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[11]. Moreover, this property assures that SEVA has a lower false positive rate (i.e., a higher specificity) reducing the number of candidates for biological validation of alternative splice events.

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. 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, many studies have shown cancer samples are more heterogeneous and encompasses a bigger spectrum of subtypes[21,32,33]. In practice, we hypothesize that differentially spliced genes show multiple patterns of isoform expression in tumors in multiple different cancer subtypes. Therefore, based upon the simulated data, we hypothesize that SEVA is uniquely suited to identify inter-tumor heterogeneity in gene isoform usage in tumors and their subtypes.

SEVA, EBSeq, and DiffSplice were all applied to RNA-seq data[24] normalized with MapSplice[9] 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. It is also directly applicable to estimates of percent spliced[25] 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[4,11,13]. 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[34,35].

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 that cancer samples have higher degree of heterogeneity compared to normal samples. As further validation, genes with alternative splicing events in HPV+ OPSCC from [24] were significantly enriched in cross-study validation on RNA-seq data for HPV+ HNSCC samples in TCGA[27]. 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[19]. 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 lead to the development of the algorithm. In addition, 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[27,30]. 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[20] 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.

REFERENCES

1 Mo, Q. *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences* **110**, 4245-4250, 2013.

2 Lim, K. H. *et al.* Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proceedings of the National Academy of Sciences* **108**, 11093-11098, 2011.

3 Chen, J. & Weiss, W. Alternative splicing in cancer: implications for biology and therapy. *Oncogene* **34**, 1-14, 2015.

4 Song, L. *et al.* CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic acids research*, gkw158, 2016.

5 Canzar, S. *et al.* CIDANE: Comprehensive isoform discovery and abundance estimation. *Genome biology* **17**, 1, 2016.

6 Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295, 2015.

7 Guttman, M. *et al.* Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. *Nature biotechnology* **28**, 503, 2010.

8 Li, J. J. *et al.* 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**, 19867-19872, 2011.

9 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178-e178, 2010.

10 Sebestyen, E. *et al.* Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res* **43**, 1345-1356, 2015.

11 Anders, S. *et al.* Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017, 2012.

12 Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic acids research* **41**, e39-e39, 2013.

13 Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic acids research*, gkr1291, 2012.

14 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593-5601, 2014.

15 Feng, H. *et al.* Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer letters* **340**, 179-191, 2013.

16 Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-1806, 2008.

17 Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nat Genet* **39**, 347-351, 2007.

18 Frazee, A. C. *et al.* Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* **31**, 2778-2784, 2015.

19 Li, R. *et al.* Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. *PloS one* **9**, e91263, 2014.

20 Ebert, B. & Bernard, O. A. Mutations in RNA splicing machinery in human cancers. *N Engl J Med* **365**, 2534-2535, 2011.

21 Afsari, B. *et al.* Learning dysregulated pathways in cancers from differential variability analysis. *Cancer Inform* **13**, 61-67, 2014.

22 Vaart, A. W. v. d. *Asymptotic statistics*. (Cambridge University Press, 1998).

23 Xing, Y. *et al.* An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic acids research* **34**, 3150-3160, 2006.

24 Guo, T. *et al.* Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. *International journal of cancer* **139**, 373-382, 2016.

25 Alamancos, G. P. *et al.* Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**, 1521-1531, 2015.

26 Liu, R. *et al.* Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics* **15**, 364, 2014.

27 Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582, 2015.

28 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, 2013.

29 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1, 2013.

30 Mroz, E. A. *et al.* Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas. *PLoS Med* **12**, e1001786, 2015.

31 Leng, N. *et al.* EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035-1043, 2013.

32 Corrada Bravo, H. *et al.* Gene expression anti-profiles as a basis for accurate universal cancer signatures. *BMC Bioinformatics* **13**, 1-11, 2012.

33 Eddy, J. A. *et al.* Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). *PLoS Comput Biol* **6**, e1000792, 2010.

34 Afsari, B. *et al.* Rank Discriminants for Predicting Phenotypes from RNA Expression. *Annals of Applied Statistics* **8**, 1469-1491, 2014.

35 Bolstad, B. M. *et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193, 2003.

SUPPLEMENTAL FILE 1: Supplemental Methods

Kendall-τ and modified Kendall-τ distances

Kendall-τ distance (also known as "bubble distance") was introduced by Maurice Kendall and is a measure of the disparity between the orderings of two lists of numbers. If $A=(a_1,a_2,...,a_m)$ and $B=(b_1,b_2,...,b_m)$ are two profiles, then Kendall-τ distance between them is:

$$D_{\text{Kendall-}\tau}(A,B) = \frac{1}{\binom{m}{2}} \sum_{i<j} I\Big(I(a_i < a_j) \neq I(b_i < b_j)\Big).$$

In this equation, I() is the indicator function that has a value of 1 when the argument is true and zero otherwise. Therefore, $I\Big(I(a_i < a_j) \neq I(b_i < b_j)\Big) = 1$ if either i) $a_i < a_j$ and $b_i > b_j$ or ii) $a_i > a_j$ and $b_i < b_j$.

In the modified version of Kendall-τ used for the SEVA algorithm in this paper, the Kendall-τ distance is computed between the set of junctions annotated to a gene. In this modified distance metric, the only comparisons considered are those between the set of competing junctions C. If C={(i,j)|i,j compete, i<j}, then the modified Kendall-τ becomes:

$$D_{\text{modified-Kendall-}\tau}(A,B) = \frac{1}{|C|} \sum_{(i,j)\in C} I\Big(I(a_i < a_j) \neq I(b_i < b_j)\Big).$$

EVA and SEVA Variance Estimation

The EVA statistic compares the average dissimilarity or distance (denoted by D) between any two random samples from a normal population and the same quantity for a tumor population. The calculation of this statistic relies on the two corresponding sample-based empirical average distances, denoted by D_N and D_T. Fortunately, U-Statistics provides a well-established mathematical theory to compare such quantities. Perhaps the main result of the U-statistics theory for application to EVA is that D_N and D_T converge asymptotically to normal distributions as the sample sizes increase. This convergence assumes that D is bounded, which occurs for the Kendall-τ distance and modified Kendall-τ distance used in EVA and SEVA, respectively. More precisely, if n_N is the number of normal samples and $E(D_N)$ is the mean of the distance between independent normal samples $\sqrt{n_N}(D_N\text{-}E(D_N))$ converges to a mean zero normal distribution with variance σ_N^2 as n_N grows to infinity. The same holds true for the tumor population: $\sqrt{n_T}(D_T\text{-}E(D_T))$ is asymptotically normal with mean zero and some variance σ_T^2, described in further detail below. A simple calculation reveals that under the null hypothesis that the variability of the distance between two normal and two cancer samples are equal, i.e. $E(D_N) = E(D_T)$, the sample adjusted difference $\sqrt{n_N + n_T}(D_N\text{-}D_T)$ converges to zero mean normal with variance $\sigma^2 = \frac{\sigma_N^2}{\lambda_N} + \frac{\sigma_T^2}{\lambda_T}$ where $\lambda_N = \frac{n_N}{n_N+n_T}$, the fraction of normal samples, and $\lambda_T = \frac{n_T}{n_N+n_T}$, the fraction of tumor samples.

Hence, the core of the EVA and SEVA algorithms is estimating the parameters σ_N^2 and σ_T^2. The theory of U-Statistics provides an asymptotical approximation for $\sigma_N^2 \to 4\mathrm{Cov}(D(X,X'), D(X,X''))$ where X, X' and X'' are independent and identically distributed profiles from normal population. Applying a similar formula to profiles from the tumor population approximates σ_T^2. In practice, we estimate the above covariance terms from our data. Clearly, estimating these covariances which involve three random variables, is more difficult than in the standard case of variance estimation for a single random variable. To reduce the variance of our estimates, we use a modified estimator which avoids certain degeneracy which appears in U-Statistics.