Splice Expression Variation Analysis (SEVA) for Differential Gene Isoform Usage in Cancer

Bahman Afsari[1,†], Theresa Guo[2,†], Michael Considine[1], Liliana Florea[3], Dylan Kelly[2], Emily Flam[2], Patrick K. Ha[4], Donald Geman[5,6], Michael F. Ochs[7], Joseph A. Califano[8,9], Daria A. Gaykalova[2,*], Alexander V. Favorov[1,10,11,*], Elana J. Fertig[1,4,*]

1. Department of Oncology, The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, 1550 Orleans Street, Baltimore, MD 21231
2. Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD USA
3. McKusick-Nathans Institute of Genetic Medicine, Department of Medicine, Johns Hopkins University, Baltimore, MD, USA
4. Department of Otolaryngology – Head and Neck Surgery (OHNS), University of California, San Francisco, CA, USA
5. Institute for Computational Medicine, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA
6. Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD, USA
7. Department of Mathematics and Statistics, The College of New Jersey, Ewing Township, NJ USA
8. Head and Neck Cancer Center, Moores Cancer Center, University of California, San Diego, 3855 Health Sciences Dr., MC 0803, La Jolla, CA, USA
9. Division of Otolaryngology-Head and Neck Surgery, Department of Surgery, 3855 Health Sciences Dr., MC 0803, La Jolla, CA, USA
10. Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia
11. Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia

E-mail addresses: Bahman Afsari: bahman@jhu.edu, Theresa Guo: tguo5@jhmi.edu, Michael Considine: mconsid3@jhmi.edu, Liliana Florea: florea@jhu.edu, Dylan Kelly:dkelle21@jhu.edu, Emily Flam: elflam@email.wm.edu, Patrick K. Ha: Patrick.Ha@ucsf.edu, Donald Geman: geman@jhu.edu, Michael F. Ochs: ochsm@tcnj.edu, Joseph A. Califano: jcalifano@ucsd.edu, Daria A. Gaykalova dgaykal1@jhmi.edu, Alexander V. Favorov: favorov@sensi.org, Elana J. Fertig: ejfertig@jhmi.edu

†These authors contributed equally

*Corresponding and Senior authors: Daria A. Gaykalova, Alexander V Favorov, and Elana J. Fertig

Keywords: bioinformatics, RNA-sequencing, gene isoforms, splice variation, tumor heterogeneity

Abstract

Current differential splice variant analysis algorithms do not account for heterogeneous gene isoform usage in tumors. We introduce Splice Expression Variability Analysis (SEVA) to detect differential splice variation usage in tumor and normal samples, accounting for tumor heterogeneity. SEVA compares the degree of variability of junction expression within a population of normal samples relative to that in tumor samples. SEVA is robust and computationally efficient relative to EBSeq and DiffSplice in simulated data. SEVA identified differential gene isoform usage robust in cross-study validation in head and neck tumors. Therefore, SEVA can identify alternative gene isoform usage in heterogeneous tumor samples.

Background

Cancer is often considered a disease of genetic disruption. In some tumors, DNA alterations in a small set of genes within a common pathway can drive cancer development [1]. However, individual tumors of the same subtype may have a diverse landscape of DNA alterations [2, 3]. Gene expression analyses of high throughput RNA-sequencing of tumors in large consortia studies (reviewed in [4]) have highlighted the pervasive and heterogeneous gene expression changes in individual tumors relative to normal samples [3, 5-10]. However, these integrated analyses of DNA and RNA sequencing data identify clusters of tumor samples with common gene expression changes but lack consistent DNA alterations [11]. This incomplete explanation of gene expression changes occurs even in tumors with established drivers such as adenoid cystic carcinoma [12], and is even more pervasive in cancers lacking clear genetic drivers such as head and neck squamous cell carcinoma [13, 14]. Therefore, 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 [15]. ASEs are a significant component of expression alterations in cancer, and have been demonstrated to be critically important in the development of malignant phenotypes in a variety of solid and liquid tumors [16]. 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-sequencing reads for a single sample [17-22]. These tools offer the opportunity to systematically evaluate the gene isoforms that are expressed in a sample. 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 [21, 23-25]. In spite of the breadth of available ASE algorithms, few have been validated in primary tumor samples [26].

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 [27, 28]. Therefore, altered splicing patterns seen in tumors may result in a variable gene isoform profile across tumor samples, rather than a uniform expression pattern that is more likely to be observed across normal tissue 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 will not identify alterations in splicing that occur only within a subset of samples. Moreover, relying on mean estimates of isoform expression for identifying differential expression of isoforms may confound genes with alternative splicing and

genes with differential overall expression. While the confounding of ASEs with differential expression may be mitigated by data normalization [23] or network-based algorithms [24], these approaches are still insufficient to account for tumor heterogeneity.

In this paper, we develop a novel algorithm called Splice Expression Variability Analysis (SEVA) to 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. Because these isoform variants are tumor-specific, ASEs specific to tumors are expected to have more variable exon junction expression than normal samples. Previously, we developed a multivariate algorithm to compare the variability of expression profiles between tumor and normal samples called, Expression Variation Analysis (EVA). EVA was implemented in the R/Bioconductor package GSReg [10] and applied to compare the distribution of gene expression patterns in sets of genes in pathways. Compared to traditional gene enrichment analyses that find consistent changes in expression values based upon univariate differential expression statistics, EVA is unique in its ability to determine dysregulated pathways [10]. SEVA adapts the multivariate statistics from the EVA pathway analysis to compare interactions between the exons in a gene, similar to how dysregulated signaling networks alter the interactions between pathway genes. SEVA also develops a new statistics from EVA to compare the gene isoform profiles based upon the biological knowledge about formation of isoforms in order to improve SEVA's detection power.

In this study, we describe the SEVA algorithm and its implementation in the R/Bioconductor package, GSReg. The performance of SEVA was compared with two existing algorithms designed for differential splice variant expression analysis, EBSeq and DiffSplice, in simulated RNA-sequencing data generated with Polyester [29]. We show that SEVA had the most robust performance in heterogeneous test samples, which are representative of primary tumor samples. In contrast to EBSeq and DiffSplice, SEVA was able to identify alternative splicing events independent of overall gene expression differences when there is heterogeneity in simulated cancer samples. Finally, additional validation was performed using publically available RNA-sequencing data for primary tumor data from HPV-positive oropharynx squamous cell carcinoma (OPSCC) tumors and normal samples from both The Cancer Genome Atlas (TCGA) [30] (n=44 tumors, n=16 normals) and from our recent publication (n=46 tumors and n=25 normals) [31]. In the tumor samples from these cohorts, SEVA finds a list of cancer-specific ASEs with manageable number of hits for experimental validation and include experimentally validated splice variants in HNSCC from a previously microarray study [32]. Based on performance in both simulated and real data, SEVA represents a robust algorithm that is well suited for differential ASE analysis, particularly in RNA-sequencing data from heterogeneous primary tumor samples.

Results

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 expected to 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 such differential variability analysis of gene expression profiles [10]. Briefly, EVA quantifies the relative dissimilarity between gene expression sample profiles from the same phenotype by computing the average dissimilarity between all pairs of samples (denoted by D). Then, given two phenotypes (called N for normal and T for tumor with corresponding average dissimilarity D_N and D_T), the algorithm tests whether there is a statistically significant difference in the level of variability of the expression profiles between the two phenotypes (Fig 1a). In order to do this, EVA uses an approximation from U-Statistics theory [33] to test the null hypothesis that $D_N - D_T = 0$ described in detail in [34] and summarized in Supplemental File 1. The statistics from EVA depend upon: (1) the input gene expression profile and (2) the dissimilarity measure D.

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 exon junctions measured with RNA-seq data provides direct evidence of gene isoform expression in each sample. Therefore, SEVA takes the profile of exon junction expression of each gene as input. In the case of alternative splicing, the multivariate distribution of the set of exon 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 [17, 24, 35]. Whereas EVA uses Kendall-τ dissimilarity to consider all possible comparisons of expression of genes within a pathway, SEVA computes the dissimilarity

using expression profiles of the sets of "competing" junctions within a gene (Fig 1b). 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 [10]. SEVA calculates a gene-level dissimilarity measure for each phenotype by summing the measures obtained for each junction within a gene (Fig 1c and 1d). Using a rank-based dissimilarity for this application will normalize differences in exon junction expression that arise from overexpression of a gene, making the analysis of alternatively spliced genes blind to whether that gene is differentially expressed. SEVA is implemented in the R/Bioconductor package GSReg as of version 1.9.2.

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

We generate *in silico* RNA-sequencing data from the R/Bioconductor package Polyester [29] with known gene isoform usage to benchmark the performance of SEVA relative to EBSeq [23] and DiffSplice [24] 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 [22] aligned data in TCGA [14], and therefore do not introduce the alignment algorithm as an additional variable in the comparison study. The simulated data contains 600 total genes from chromosome 1 with 25 tumor and 25 normal samples. Among the genes, 150 have no change between tumor and normal samples (Supplemental Fig 1a), 150 are differentially spliced (Supplemental Fig 1b), 150 are differentially expressed (Supplemental Fig 1c), and 150

are both differentially expressed and differentially spliced (Supplemental Fig 1d). 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 applied SEVA, EBSeq and DiffSplice 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. On the other hand, 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. EBSeq also shows higher false positive rate than SEVA, with more false positives in among differentially expressed genes than non-differentially expressed genes (Supplemental Fig 2). False positives cannot be computed for DiffSplice because the software does not return statistics for all negative controls in the simulation.

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). 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 and is independent of the differentially expressed status of genes relative to DiffSplice and EBSeq when there is heterogeneity in cancer (Supplemental File 3).

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

We use RNA-sequencing data for 46 HPV-positive OPSCC and 25 normal samples from [31] 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. As expected, 217 of genes show higher variation in cancer than 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 not known in this real dataset, we cannot assess the true independence between differentially spliced and differentially expressed genes in contrast to the *in silico* data (Supplemental File 2).

SEVA analysis of RNA-sequencing data confirms previously validated HNSCC-specific splice variants TP63, LAMA3, and DST

Previously, alternative splice events in six genes (VEGFC, DST, LAMA3, SDHA, TP63, and RASIP1) were observed as being unique to HNSCC samples from microarray data and these genes were experimentally tested and validated in an additional robust, independent cohort of samples [32]. We analyze these six genes to determine if SEVA, EBSeq, or DiffSplice are also identify them (Table 1). As shown in Table 1, SEVA identified four out of the six genes as DS genes. EBSeq only identifies one of them (VEGFC) and DiffSplice identified none.

Table 1 SEVA and EBSeq p-values for experimentally validated, HNSCC-specific splice variant candidates from [32]. The EBSeq algorithm automatically adjusts for multiple hypothesis testing. SEVA p-values are not adjusted, but will not be impacted by Bonferroni adjustment used by SEVA because there are only 6 tests. DiffSplice does not identify the candidates and therefore no p-values are reported.

Genes	**SEVA p-value**	**EBSeq p-value**
VEGFC	0.22	2.65e-6
DST	2.88e-10	0.27
LAMA3	1.03e-5	0.46
SDHA	0.85	0.18
TP63	5.78e-10	0.11
RASIP1	4.76e-7	0.13

In Fig 4, we create multi-dimensional scaling (MDS) plots of the modified Kendall-τ distance of the significant genes in SEVA and EBSeq, with corresponding heatmaps of these dissimilarities in Supplemental Figure 4. The closer two samples in this MDS plot, the less variable their gene 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. These four independently identified differentially spliced genes 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 (Supplemental Figure 5) and was not detected by SEVA. The detection by EBSeq indicates that VEGFC has a

coordinated change of expression from normal to tumor without significant change in variability.

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

We also apply SEVA to independent RNA-sequencing data for 44 HPV-positive HNSCC and 16 normal samples in The Cancer Genome Atlas (TCGA) [14] to cross-validate the ASE candidates in the training data from [31]. 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 (Fig 5). 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.

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 either EBSeq [36] or DiffSplice [24] in simulated datasets that reflect this tumor heterogeneity. The precision of both SEVA and DiffSplice 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 [23]. 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 shows almost constant recall (around 40%). 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. More specifically, in our simulations, we only consider (Supplemental Fig 1) two patterns for cancer samples: one profile for non-disrupted samples, resembling the same profile as normal samples; and another profile for disrupted samples, whose expression profile differ from that of normal but that is similar among the cases on the disrupted profile. We hypothesize that SEVA will have lower

recall than techniques based upon differential isoform expression in populations with homogeneous isoform usage. However, as many studies have shown, such as [5, 6, 10], cancer samples are more heterogeneous and encompasses a bigger spectrum of subtypes. 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 splice variant candidates expressed in subsets of tumors.

A previous study [32] identified and experimentally validated six genes with ASE in HNSCC. SEVA identifies four of those six genes in HPV+ vs normal samples. EBSeq identifies only VEGFC, while DiffSplice does not identify any of these genes. The previous study [32] did not account for the association of HPV status with the splice events in the HNSCC samples, which may account for the two missing genes from the SEVA analysis. The relevance of SEVA for alternative gene isoform usage is further confirmed with the robust cross-study validation of candidates from the training set with SEVA analysis of alternative splice events in HPV-positive OPSCC and normal TCGA samples.

SEVA uses competition among the set of junctions to assess a holistic pattern change from RNA-sequencing data between normal and tumor tissues. The competing junctions are those overlapping regions found in the genome. Because they overlap, reads that are aligned to competing junctions must correspond to different isoforms, i.e. if two mRNAs contain two competing junctions, the mRNAs correspond to different splice variants of the same gene. Therefore, a severe change in the profile and the ranking of a junction among its competitors represents a drastic alteration in the

expression profile of the isoforms. By definition, the modified Kendall-τ quantifies such extreme changes in junction expression and, consequently, it indirectly quantifies the alteration in the isoform pattern of expression. We observe that this dissimilarity metric accurately characterizes relative heterogeneity of gene isoform usage in the confirmed HNSCC-specific ASEs DST, LAMA3, TP63, and RASIP1.

In this study, SEVA, EBSeq, and DiffSplice were all applied to RNA-sequencing data [31] normalized with MapSplice [22] to enable cross-study validation with the TCGA normalized data. While there are numerous other algorithms for such differential splice analysis, many rely on data obtained from distinct alignment and normalization pipelines [17, 23, 25]. 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-sequencing 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 [37, 38].

The SEVA algorithm has been implemented in the R/Bioconductor package GSReg for differential variability analysis [10] as of version 1.9.2. In future studies, SEVA can be applied to different types of cancer and be validated on different tumor types. Also, SEVA can be adapted to detect simultaneously the variability change and mean change from normal to tumor. To boost the signal to noise ratio in these analyses,

it will be necessary to further filter genes and junctions for analysis to ensure biological relevance. Moreover, one obvious improvement to EVA and SEVA in future studies will be to stabilize the variance estimation using empirical Bayes methods. Although not necessary in this study, filters on junction expression changes relative to background and lowly expressed genes have been implemented in the SEVA functions in GSReg. While SEVA prioritizes genes with alternative isoform usage in tumor samples, additional statistics are needed to prioritize the precise isoforms used in each sample. These statistics will enable additional modifications of the Kendall-τ dissimilarity measure to integrate mutation, copy number, and DNA methylation data to determine whether the ASE candidates are part of a larger of pool of gene alterations that drive the heterogeneous transcriptional changes in cancer.

Conclusion

In this study, we introduce Splice Expression Variation Analysis algorithm that identifies Alternative Splicing Events in cancer samples by variability analysis of junction expression profiles. 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-positive OPSCC from [31] were significantly enriched in cross-study validation on RNA-sequencing data for HPV-positive HNSCC samples in TCGA. Therefore, SEVA is particularly adept at inferring ASEs in tumor samples with heterogeneous gene isoform

usage relative to normal samples with more homogeneous gene isoform usage that are independent of the differential expression status of the gene.

Methods

HPV-positive HNSCC and normal RNA-sequencing datasets

We use RNA-sequencing data for 46 HPV-positive OPSCC and 25 independent normal samples from uvulopalatopharyngoplasty previously described in [31]. Raw data is in process at dbGAP. Normalized gene, isoform, and junction expression values will be released on GEO. In addition, we obtained independent RNA-sequencing data from 44 HPV-positive HNSCC and 16 matched normal tissues from the freeze set of 281 samples used in the TCGA HNSCC marker paper [14].

In silico data

To simulate isoform expression, we used the expression of isoforms from the HPV-positive RNA-sequencing data provided in [31]. Then, we selected genes in chromosome 1 whose expression was in a medium range in normal samples, i.e. between 4 to 9 in log scale. The resulting gene selection left 600 genes, from which we generated a dataset of 25 tumor and 25 normal simulated samples. For these genes, we calculated the average isoform expression for normal samples and input these values in Polyester [29] to obtain simulated fastq files each normal sample. To simulate tumor heterogeneity, a subset of the total 25 tumor samples are set to match the distribution of gene isoform values in the normal samples and an additional subset have disrupted gene isoform usage. These disrupted cancer samples were simulated by dividing the genes into four sets of 150 genes, with mean isoform expression calculated

according to the following conditions and input to Polyester to simulate the following scenarios:

1) Neutral genes: no change in average expression of isoforms from normal to disrupted cancer samples (Supplemental Fig 1a). Average isoform expression values from the normal samples are used as the input to Polyester.
2) Differentially spliced only: the average isoforms expression corresponding to a gene for the normal samples are permuted in disrupted cancer samples. Hence, the total isoform expression for a gene remains the same but their isoform expression changes. (Supplemental Fig 1b) depicts an example for samples indexed by 36-50.
3) Differentially expressed only: all isoforms are differentially expressed with log fold change of 1 in the disrupted cancer samples (Supplemental Fig 1c). Genes are randomly selected to be over or underexpressed in the tumor samples relative to normal samples.
4) Differentially spliced and differentially expressed: Both processes described in steps 2 and 3 are applied to disrupted tumor samples (Supplemental Fig 1d).

In these simulations, all other parameters were set to default in the Polyester package.

RNA-sequencing data normalization

All *in silico* and real RNA-sequencing datasets are normalized with the RNA-seq v2 pipeline from TCGA [14]. Specifically, they are aligned to hg19 with MapSplice [22] multithread version 2.0.1.9. Junction expression is obtained directly from the MapSplice output for each sample, setting expression values to zero for junctions that are not detected from MapSplice in a given sample. Gene and isoform expression are quantified with RSEM [39] and upper quartile normalized [40]. 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 [31].

SEVA algorithm and implementation

SEVA extends the pathway dysregulation algorithm EVA to perform differential gene isoform analysis as described in the results. Mathematical details about the algorithm and modified Kendall-τ dissimilarity measure are in Supplemental File 1. We previously implemented EVA in the R/Bioconductor package GSReg [10]. We have modified the package as of version 1.6.1 to implement SEVA algorithm for this present study. The function SEVA.meangenemaxjunc.Filter implements SEVA with inputs of junction. Total gene expression values are used for filtering, described in detail in the GSReg package vignette. The default gene mapping used to define sets of competing junctions for each query junction is hg19 (Fig 1b). Users can modify this function to filter junctions and update the gene coordinates using newer libraries. In the analyses presented in this study, we used the EVA's default filter to remove genes with less than three junctions from analysis. The SEVA analysis of junction expression is computationally efficient. All of the computations in this manuscript performed on a Lenovo Thinkpad with Core (TM) i7-3720QM Intel CPU @2.6 GHz in less than an hour with code from Supplemental File 3. Most of the computational time is used to find the overlaps of the junction expression. In this paper, genes with Bonferroni adjusted p-values below 1% are called statistically significant.

EBSeq analysis

EBSeq is performed with the R/Bioconductor package EBSeq version 3.3 [36]. We normalize isoform expression for all genes as the input in the EBSeq analysis. Isoforms with posterior probability above 99% were called significantly differentially spliced. EBSeq was also applied to normalized total gene expression values, and genes with a posterior probability above 99% were called significantly differentially expressed.

DiffSplice analysis

DiffSplice 0.1.2 beta version [24] 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 phenotype, we select a random subset of 14 HPV-positive OPSCC and 14 normal samples from the training dataset for analysis [31].

Cross-study validation

Since the number of normal samples in TCGA is small (16), we chose it as the test set. As described in the previous section, we apply SEVA to the training set and apply it to TCGA for HPV+ samples and normal samples. The histogram of the p-values on training and test set are shown in Fig 5a and Fig 5c. Then, we use the z-scores of the genes identified on the training using the TCGA data. We test if these z-scores are enriched among all TCGA genes using the function wilcoxGST in the limma package version 3.24.15.

Figures

Fig 1. Schematic of the differential gene isoform usage algorithm Splice-Expression Variability Analysis (SEVA). (a) Genes G_1 to G_8 are in a set that are annotated to a single pathway. The EVA pathway dysregulation algorithm upon which SEVA is based compares the expected dissimilarity of the expression profiles for genes G_1 to G_8 in normal samples (D_N; samples N_i and N_j; blue) to the expected dissimilarity these expression profiles in tumor samples (D_T; samples T_k and T_l; red). On the heatmap red indicates higher expression and green lower expression. (b) Schematic of exons and exon junctions for a gene being assessed for differential isoform usage. SEVA inputs comparisons of each junction in a gene (called the query junction; green) overlapping junctions (competing junctions; blue) and not non-overlapping junctions (non-competing junctions; black dashed). (c) The Kendall-τ dissimilarity is calculated for each query junction using a set of junctions defined by that query junction and competing junctions, excluding non-competing junctions (greyed in the heatmap) by computing the number of disagreeing comparisons between relative junction expression in each sample. (d) SEVA uses the sums over the Kendall-τ dissimilarity for each query junction and generates a p-value comparing relative isoform usage in sample groups with the same statistics as EVA.

Fig 2. **Comparison of differential gene isoform usage algorithms in simulated RNA-Seq data.** (a) Precision for EBSeq (blue), DiffSplice (green), and SEVA (red). In this simulation, varying numbers of the total tumor samples have both differentially expressed (DE) and differentially spliced (DS) genes (x-axis). Solid lines represent the precision for differentially spliced, but not differentially expressed genes and dashed lines the precision for both differentially spliced and differentially expressed genes. (b) Recall of simulated data, plotted as in (a).

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 in SEVA sense in normal vs cancer, respectively. The red points represent significantly differentially (DS) spliced genes identified with EVA, 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).

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).

Fig 5. Cross-study validation of SEVA differential gene isoform candidates from real HPV+ HNSCC RNA-Seq data in the training cohort with test RNA-Seq data in TCGA. (a) Histogram of SEVA p-values in the training set. (b) p-value histogram on TCGA data. (c) Histogram of p-value calculated on the test TCGA data for candidates identified as significantly differentially spliced in training data. (d) Comparison of p-values of candidates on training versus test datasets.

Supplementary Material

Supplemental Fig 1. Example of the expression for four isoforms comprising a single gene in the simulated data. Each of the four conditions that a gene is (a) not changed, (b) differentially spliced, (c) differentially expressed, and (d) both differentially spliced and differentially expressed in tumor samples relative to normal samples. These examples represent the case in which the gene has the differential expression or differential splicing event in only half of the tumor samples.

Supplemental Fig 2. (a) Precision-Recall and (b) ROC curve comparing differential gene isoform candidates from SEVA and EBSeq. EBSeq is plotted in blue and SEVA in red. Solid lines represent differentially spliced, but not differentially expressed genes and dashed lines both differentially spliced and differentially expressed genes.

Supplemental Fig 3. Heatmaps of modified Kendall-τ distances in real HPV+ HNSCC junction expression from RNA-seq for (a) DST, (b) LAMA3, (c) RASIP1, and (d) TP63. Coloring along rows and columns is blue for normal samples and red for tumor samples.

Supplemental Fig 4. Distribution of modified Kendall-τ distances in real HPV+ HNSCC junction expression from RNA-seq for VEGFC as an MDS plot (a) and heatmap (b). In both plots, normal samples are indicated in blue and tumor samples in red.

Supplemental File 1. Detailed mathematical description of the Kendall-τ distances and U-theory statistics used for SEVA.

Supplemental File 2. Venn diagrams testing for the dependence status of differentially expressed genes (DE) and algorithms hits on the simulated data. Each Venn diagram contains the DE genes, DS genes, and hits by the algorithm (i.e. SEVA, EBSeq or DiffSplice). "Number of perturbed samples" represents the heterogeneity in the cancer, as described in the *in silico* data, with heterogeneity decreasing as this number increase. The p-value is calculated for the Fisher test of the dependence between DE and the hits by the algorithm (the larger the better).

Supplemental File 3. Vignette containing R code for all analyses performed in this manuscript.

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not applicable

Availability of data and materials

The dataset from [41] is pending submission to dbGAP. TCGA raw data are available from dbGAP (phs000178.v1.p1) and total gene expression values from the Level 3 data in the BROAD Firehose.

The SEVA algorithm is available from the GSReg package (as of version 1.9.2) and analyses from this paper can be reproduced using scripts from Supplemental File 2.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors' research was supported by: National Institutes of Health: National Cancer Institute P30 CA006973 (BA, MC, AVF, and EJF) and R01 CA177669 (EJF), National Institute on Deafness and Other Communication Disorders T32DC000027-26 (TG), National Institute of Dental and Craniofacial Research R21 DE025398 (DAG), R01 DE023347 (JAC), P50 DE019032 pilot funding (DAG and EJF), R01 DE023227 (PH), and National Library of Medicine R01LM011000 (MFO). Additional funding was provided from the Russian Foundation for Basic Research 14-04-01872 (AVF), The National Science Foundation ABI-1356078 (LF), and The Adenoid Cystic Carcinoma Research Foundation (EJF and PH).

Author contributions

JAC, MFO, DAG, and EJF formulated the study design. JAC and DAG created the RNA-Seq data from the training cohort of HPV+ HNSCC tumors and normal samples used in this study. BA, TG, AVF, PH, DG, MFO, JAC, DAG, and EJF contributed to algorithm development, lead by BA. BA and EJF implemented the algorithm. BA,TG, MC, LF, JAC, DAG, and EJF performed data analysis. All authors were involved in writing and approving the final manuscript.

Acknowledgements

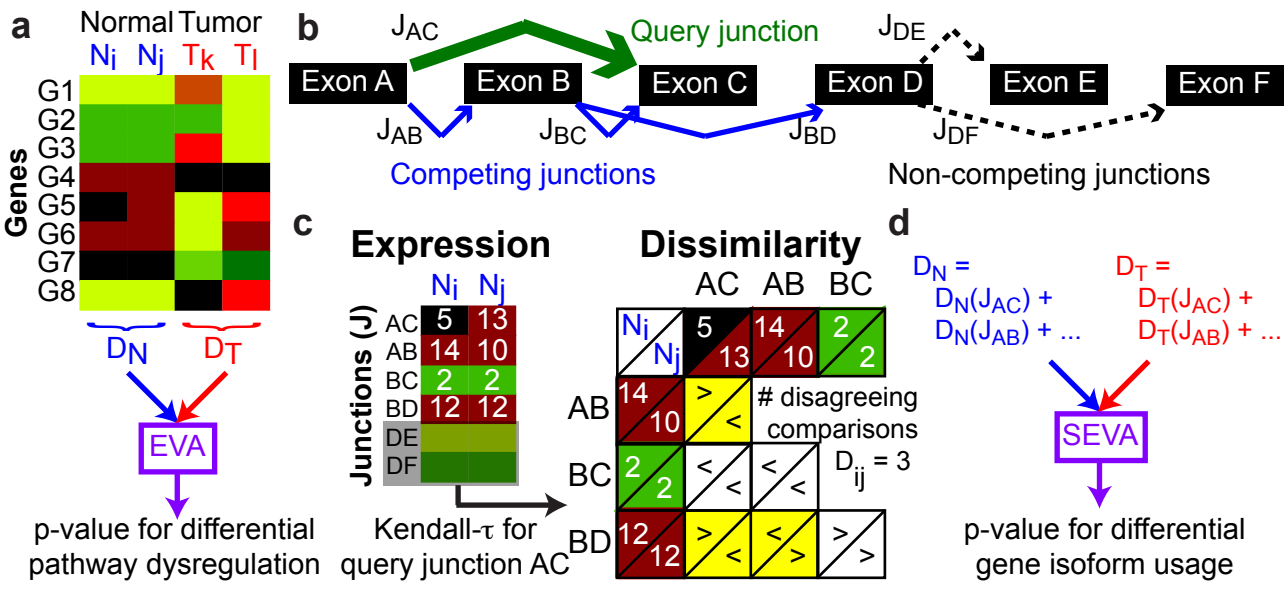
We thank Leslie Cope, Sarah Wheelan, and Ludmila V Danilova for advice on analyses, data preprocessing, and data access.

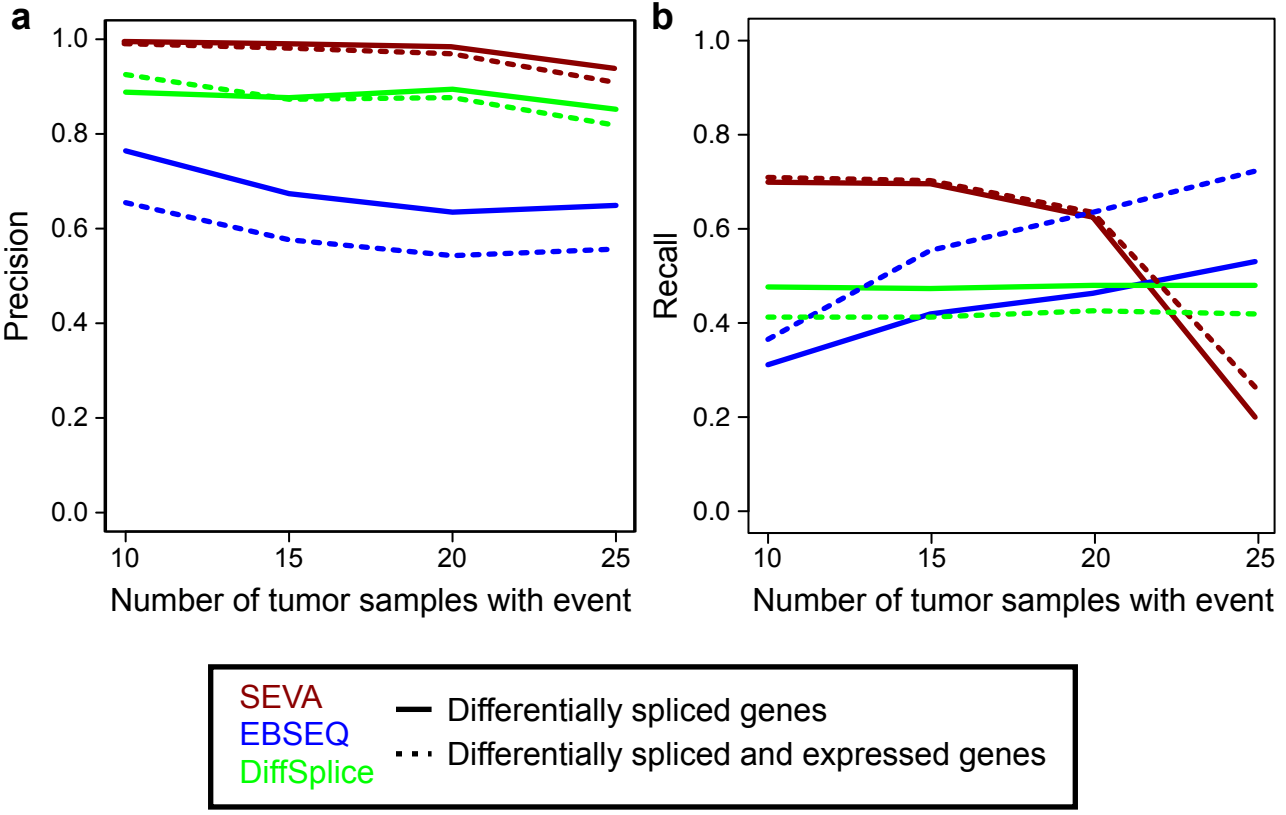
References

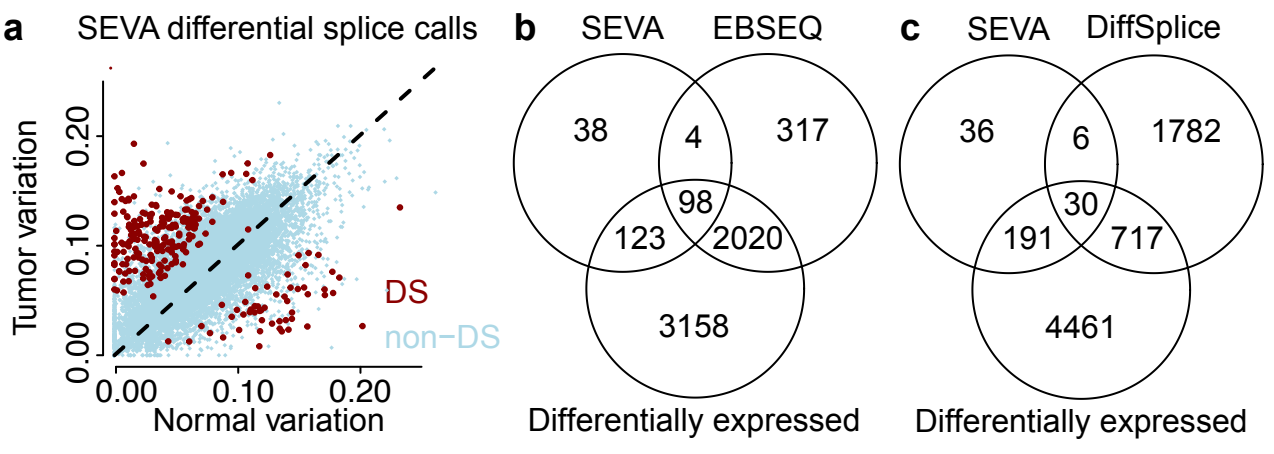
1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA and Kinzler KW. Cancer genome landscapes. science. 2013; 339(6127):1546-1558.
2. Izumchenko E, Chang X, Brait M, Fertig E, Kagohara LT, Bedi A, Marchionni L, Agrawal N, Ravi R and Jones S. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. Nature communications. 2015; 6.
3. Mroz EA and Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral oncology. 2013; 49(3):211-215.
4. Kannan L, Ramos M, Re A, El-Hachem N, Safikhani Z, Gendoo DMA, Davis S, Gomez-Cabrero D, Castelo R, Hansen KD, Carey VJ, Morgan M, Culhane AC, Haibe-Kains B and Waldron L. Public data and open source tools for multi-assay genomic investigation of disease. Briefings in Bioinformatics. 2016; 17(4):603-615.
5. Eddy JA, Hood L, Price ND and Geman D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Comput Biol. 2010; 6(5):e1000792.
6. Corrada Bravo H, Pihur V, McCall M, Irizarry RA and Leek JT. Gene expression anti-profiles as a basis for accurate universal cancer signatures. BMC Bioinformatics. 2012; 13(1):1-11.
7. Ochs MF, Farrar JE, Considine M, Wei Y, Meschinchi S and Arceci RJ. (2013). Outlier Gene Set Analysis Combined with Top Scoring Pair Provides Robust Biomarkers of Pathway Activity. In: Ngom A, Formenti E, Hao J-K, Zhao X-M and van Laarhoven T, eds. Pattern Recognition in Bioinformatics: 8th IAPR International Conference, PRIB 2013, Nice, France, June 17-20, 2013 Proceedings. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 47-58.
8. MacDonald JW and Ghosh D. COPA--cancer outlier profile analysis. Bioinformatics. 2006; 22(23):2950-2951.
9. Liu Y, Koyuturk M, Barnholtz-Sloan JS and Chance MR. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. BMC Syst Biol. 2012; 6:65.
10. Afsari B, Geman D and Fertig EJ. Learning dysregulated pathways in cancers from differential variability analysis. Cancer Inform. 2014; 13(Suppl 5):61-67.
11. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M and Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences. 2013; 110(11):4245-4250.
12. Rettig EM, Talbot CC, Jr., Sausen M, Jones S, Bishop JA, Wood LD, Tokheim C, Niknafs N, Karchin R, Fertig EJ, Wheelan SJ, Marchionni L, Considine M, Fakhry C, Papadopoulos N, Kinzler KW, et al. Whole-Genome Sequencing of Salivary Gland Adenoid Cystic Carcinoma. Cancer Prev Res (Phila). 2016; 9(4):265-274.
13. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, Zhang N, El-Naggar AK, Jasser SA, Weinstein JN, Trevino L, Drummond JA, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011; 333(6046):1154-1157.

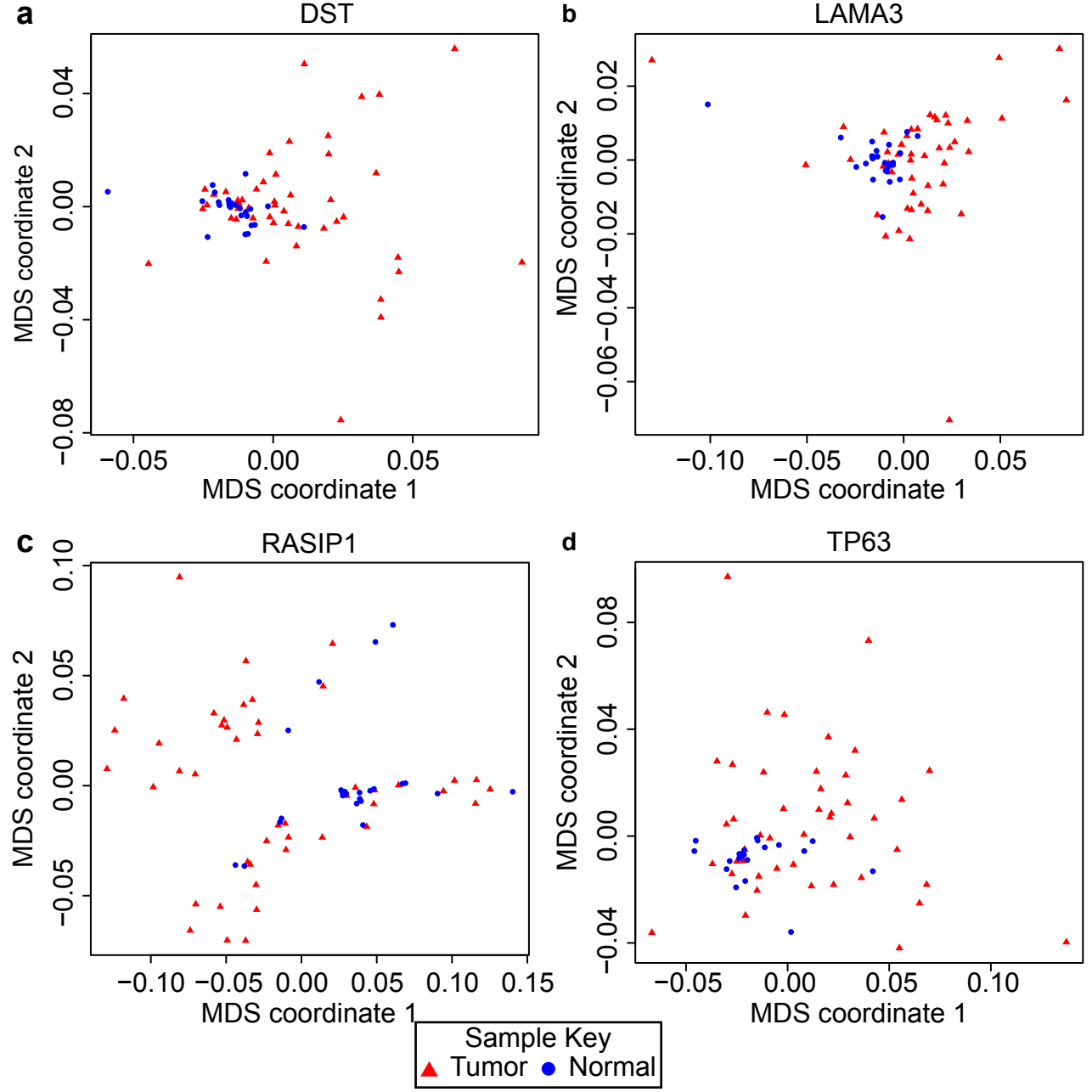
14. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015; 517(7536):576-582.
15. Lim KH, Ferraris L, Filloux ME, Raphael BJ and Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proceedings of the National Academy of Sciences. 2011; 108(27):11093-11098.
16. Chen J and Weiss W. Alternative splicing in cancer: implications for biology and therapy. Oncogene. 2015; 34(1):1-14.
17. Song L, Sabunciyan S and Florea L. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. Nucleic acids research. 2016:gkw158.
18. Canzar S, Andreotti S, Weese D, Reinert K and Klau GW. CIDANE: Comprehensive isoform discovery and abundance estimation. Genome biology. 2016; 17(1):1.
19. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015; 33(3):290-295.
20. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A and Nusbaum C. Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs. Nature biotechnology. 2010; 28(5):503.
21. Li JJ, Jiang C-R, Brown JB, Huang H and Bickel PJ. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. Proceedings of the National Academy of Sciences. 2011; 108(50):19867-19872.
22. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA and Perou CM. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research. 2010; 38(18):e178-e178.
23. Anders S, Reyes A and Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012; 22(10):2008-2017.
24. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan P-F, Hammond SM and Makowski L. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic acids research. 2013; 41(2):e39-e39.
25. Shen S, Park JW, Huang J, Dittmar KA, Lu Z-x, Zhou Q, Carstens RP and Xing Y. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic acids research. 2012:gkr1291.
26. Feng H, Qin Z and Zhang X. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. Cancer letters. 2013; 340(2):179-191.
27. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008; 321(5897):1801-1806.
28. Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, MacConaill L, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, et al. High-throughput oncogene mutation profiling in human cancer. Nat Genet. 2007; 39(3):347-351.

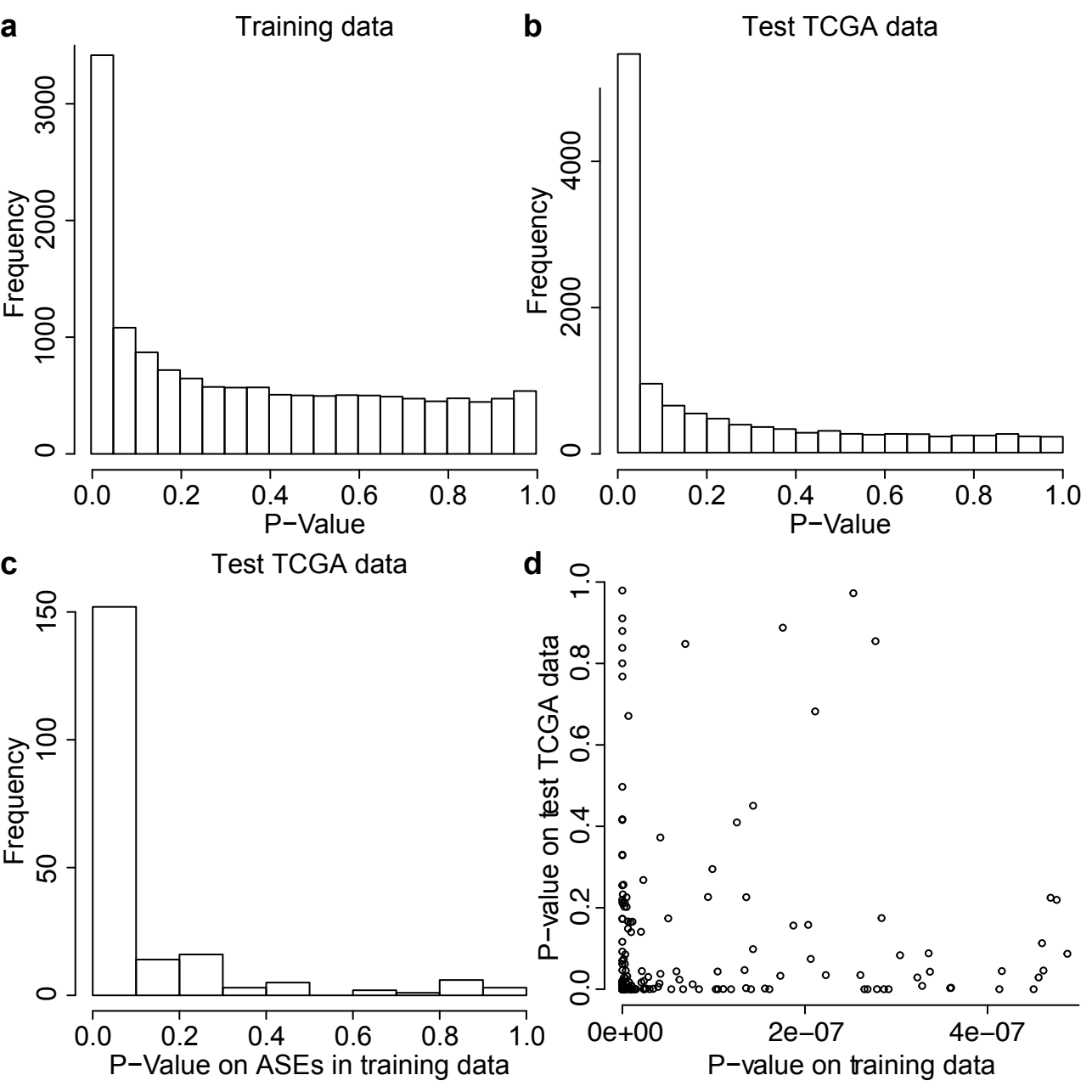
29. Frazee AC, Jaffe AE, Langmead B and Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015; 31(17):2778-2784.
30. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489(7417):519-525.
31. Guo T, Gaykalova DA, Considine M, Wheelan S, Pallavajjala A, Bishop JA, Westra WH, Ideker T, Koch WM and Khan Z. Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. International journal of cancer. 2016; 139(2):373-382.
32. Li R, Ochs MF, Ahn SM, Hennessey P, Tan M, Soudry E, Gaykalova DA, Uemura M, Brait M and Shao C. Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. PloS one. 2014; 9(3):e91263.
33. Vaart AWvd. (1998). Asymptotic statistics. (Cambridge, UK ; New York, NY, USA: Cambridge University Press).
34. Afsari B. (2013). Modeling cancer phenotypes with order statistics of transcript data. Electrical and Computer Engineering: The Johns Hopkins University).
35. Xing Y, Yu T, Wu YN, Roy M, Kim J and Lee C. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic acids research. 2006; 34(10):3150-3160.
36. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM and Kendziorski C. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013; 29(8):1035-1043.
37. Afsari B, Neto UB and Geman D. Rank Discriminants for Predicting Phenotypes from RNA Expression. Annals of Applied Statistics. 2014; 8(3):1469-1491.
38. Bolstad BM, Irizarry RA, Astrand M and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003; 19(2):185-193.
39. Li B and Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323.
40. Bullard JH, Purdom E, Hansen KD and Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94.
41. Guo T, Gaykalova DA, Considine M, Wheelan S, Pallavajjala A, Bishop JA, Westra WH, Ideker T, Koch WM, Khan Z, Fertig EJ and Califano JA. Characterization of functionally active gene fusions in human papillomavirus related oropharyngeal squamous cell carcinoma. Int J Cancer. 2016; 139(2):373-382.


a
Normal Tumor
N_i N_j T_k T_l
Genes
G1
G2
G3
G4
G5
G6
G7
G8
D_N
D_T
EVA
p-value for differential pathway dysregulation
b
J_{AC}
Query junction
J_{DE}
Exon A
Exon B
Exon C
Exon D
Exon E
Exon F
J_{AB}
J_{BC}
J_{BD}
J_{DF}
Competing junctions
Non-competing junctions
c
Expression
Junctions (J)
N_i N_j
AC 5 13
AB 14 10
BC 2 2
BD 12 12
DE
DF
Kendall-τ for query junction AC
Dissimilarity
AC AB BC
AB
BC
BD
disagreeing comparisons
D_{ij} = 3
d
D_N = $D_N(J_{AC})$ + $D_N(J_{AB})$ + ...
D_T = $D_T(J_{AC})$ + $D_T(J_{AB})$ + ...
SEVA
p-value for differential gene isoform usage


a
b
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
10
15
20
25
Number of tumor samples with event
SEVA
EBSEQ
DiffSplice
Differentially spliced genes
Differentially spliced and expressed genes


a
SEVA differential splice calls
Tumor variation
0.00
0.10
0.20
Normal variation
0.00
0.10
0.20
DS
non−DS
b
SEVA
EBSEQ
38
4
317
98
123
2020
3158
Differentially expressed
c
SEVA
DiffSplice
36
6
1782
30
191
717
4461
Differentially expressed


a
DST
MDS coordinate 2
MDS coordinate 1
b
LAMA3
MDS coordinate 2
MDS coordinate 1
c
RASIP1
MDS coordinate 2
MDS coordinate 1
d
TP63
MDS coordinate 2
MDS coordinate 1
Sample Key
Tumor
Normal


a
Training data
Frequency
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
P−Value
b
Test TCGA data
Frequency
0
2000
4000
0.0
0.2
0.4
0.6
0.8
1.0
P−Value
c
Test TCGA data
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
P−Value on ASEs in training data
d
P−value on test TCGA data
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e−07
4e−07
P−value on training data