Title
Deep-learning augmented RNA-seq analysis of transcript splicing

Journal
Nature Methods, 16(4)

ISSN
1548-7091

Authors
Zhang, Zijun
Pan, Zhicheng
Ying, Yi
et al.

Publication Date
2019-04-01

DOI
10.1038/s41592-019-0351-9

Peer reviewed

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2020 November 02.

Published in final edited form as:

Nat Methods. 2019 April ; 16(4): 307–310. doi:10.1038/s41592-019-0351-9.

Deep-learning augmented RNA-seq analysis of transcript splicing

Zijun Zhang[1,8], **Zhicheng Pan**[1,8], **Yi Ying**[2], **Zhijie Xie**[2], **Samir Adhikari**[2,3], **John Phillips**[4], **Russ P. Carstens**[5], **Douglas L. Black**[2], **Yingnian Wu**[6], **Yi Xing**[1,2,3,7]

[1]Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, Los Angeles, USA.

[2]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, USA.

[3]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, USA.

[4]Department of Molecular and Medical Pharmacology, University of California, Los Angeles, Los Angeles, USA.

[5]Department of Medicine, University of Pennsylvania, Philadelphia, USA.

[6]Department of Statistics, University of California, Los Angeles, Los Angeles, USA.

[7]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, USA.

[8]These authors contributed equally to this work.

Abstract

A major limitation for RNA-seq analysis of alternative splicing is its reliance on high sequencing coverage. We report DARTS (https://github.com/Xinglab/DARTS), a computational framework that integrates deep learning-based predictions with empirical RNA-seq evidence to infer differential alternative splicing between biological samples. DARTS leverages public RNA-seq big data to provide a knowledge base of splicing regulation via deep learning, helping researchers better characterize alternative splicing using RNA-seq datasets even with modest coverage.

RNA sequencing (RNA-seq) enables transcriptome-wide profiling of alternative splicing[1, 2]. The rapid accumulation of RNA-seq data in public repositories (e.g. ENCODE[3], Roadmap Epigenomics[4]) provides unprecedented resources for characterizing alternative splicing

Author Contributions

Z.Z. and Y.X. conceived the study; Z.Z., Y.N.W., and Y.X. designed the research; Z.Z., Z.P., Y.Y., S.A., and J.P. performed the research; Z.X., R.P.C., and D.L.B contributed analytic tools; Z.Z. and Y.X. analyzed data; and Z.Z. and Y.X. wrote the paper with input from all authors.

Competing Financial Interests

Y.X. and D.L.B are scientific cofounders of Panorama Medicine Inc. Z.Z. and Y.X. are in the process of filing a patent application for DARTS.

across diverse biological states. However, an inherent limitation of RNA-seq is that it is restricted by sequencing depth[5], and cannot reliably quantify splicing in lowly expressed genes[6].

Motivated by recent successes in using machine learning to predict exon inclusion/skipping levels in bulk tissues or single cells[7–10], we hypothesized that large-scale RNA-seq resources can be utilized to construct a deep learning model of differential alternative splicing. To test this hypothesis, we developed DARTS (Deep-learning Augmented RNA-seq analysis of Transcript Splicing). DARTS consists of two core components: a Deep Neural Network (DNN) model that predicts differential alternative splicing between two conditions based on exon-specific sequence features and sample-specific regulatory features; and a Bayesian Hypothesis Testing (BHT) statistical model that infers differential alternative splicing by integrating empirical evidence in a specific RNA-seq dataset with prior probability of differential alternative splicing (Fig. 1a). During training, large-scale RNA-seq data are analyzed by the DARTS BHT with an uninformative prior (DARTS BHT(flat), with only RNA-seq data used for the inference) to generate training labels of high-confidence differential or unchanged splicing events between conditions, which are then used to train the DARTS DNN. During application, the trained DARTS DNN is used to predict differential alternative splicing in a user-specific dataset. This prediction is then incorporated as an informative prior with the observed RNA-seq read counts by the DARTS BHT (DARTS BHT(info)) to perform deep learning augmented splicing analysis.

Unlike methods that use *cis* sequence features to predict exon splicing patterns in specific samples[7–10], the DARTS DNN predicts differential alternative splicing by incorporating both *cis* sequence features and mRNA levels of *trans* RNA binding proteins (RBPs) in two conditions (Fig. 1b, Supplementary Fig. 1). This design allows the DARTS DNN to consider how altered expression of RBPs affects splicing. We initially focused on exon skipping, the most frequent alternative splicing pattern in animals[6]. We compiled 2,926 *cis* sequence features and 1,498 annotated RBPs[11] whose mRNA levels were treated as *trans* RBP features (Supplementary Table 1).

To train the DARTS DNN, we utilized large-scale RBP-depletion RNA-seq data in two human cell lines (K562 and HepG2) generated by the ENCODE consortium[12] (Fig. 1c). We used RNA-seq data of 196 RBPs depleted by shRNA in both cell lines, corresponding to 408 knockdown vs. control pairwise comparisons (Fig. 1c). The remaining ENCODE data, corresponding to 58 RBPs depleted in only one cell line, were excluded from training and used as leave-out data to independently evaluate the DARTS DNN (Fig. 1c). To generate training labels, we applied DARTS BHT(flat) to calculate the probability of an exon being differentially spliced or unchanged in each pairwise comparison. DARTS BHT(flat) was benchmarked using simulation datasets and compared favorably to two state-of-the-art statistical models for differential splicing inference MISO and rMATS (Supplementary Fig. 2 and 3). From the high-confidence differentially spliced vs. unchanged exons called by DARTS BHT(flat) (Supplementary Table 2), we used 90% labelled events for training and 5-fold cross validation, and the remaining 10% events for testing (Methods). The performance of the DARTS DNN increased as training progressed, reaching a maximum Area Under the

Receiver Operating Characteristic curve (AUROC) of 0.97 during cross-validation and 0.86 during testing (Supplementary Fig. 4).

To test the general applicability of the DARTS DNN, we used the leave-out data, corresponding to 58 RBPs that had never been seen during training (Fig. 1c). The trained DARTS DNN showed a high accuracy (average AUROC=0.87) on the leave-out data (Supplementary Table 3). We used the leave-out data to compare the DARTS DNN to three alternative baseline methods: the identical DNN structure trained on individual leave-out datasets (DNN), logistic regression with L2 penalty (Logistic), and Random Forest. We trained these baseline methods using 5-fold cross-validation in each leave-out dataset. Additionally, we implemented another alternative baseline method, by predicting sample-specific exon inclusion levels (PSI values)[1, 10] then taking the absolute difference of the predicted PSI values between two conditions as the metric for differential splicing ($|\hat{\psi}_{KD} - \hat{\psi}_{CTRL}|$). The DARTS DNN trained on the large-scale ENCODE data outperformed baseline methods by a large margin in 57/58 experiments (Fig. 1d). The DARTS DNN model trained on individual leave-out datasets was the worst performer, illustrating the importance of training the DARTS DNN on large-scale data comprising diverse perturbation experiments.

Next, we evaluated the ability of the DARTS framework to infer differential splicing from a specific RNA-seq dataset, by incorporating the DARTS DNN predictions as the informative prior and observed RNA-seq read counts as the likelihood (DARTS BHT(info)). The posterior ratio of differential splicing consists of two components: the prior ratio, generated by the DARTS DNN based on *cis* sequence features and *trans* RBP expression levels; and the likelihood ratio, determined by modelling the biological variation and estimation uncertainty of splice isoform ratio based on observed RNA-seq read counts. Simulation studies demonstrated that the informative prior improves the inference when the observed data is limited, for instance due to low gene expression levels or limited RNA-seq depth, but does not overwhelm the evidence in the observed data (Supplementary Fig. 5).

We used DARTS BHT(info) and DARTS BHT(flat) to infer cell-type-specific splicing events between HepG2 and K562 cell lines. To obtain high-confidence differential and unchanged splicing events between the two cell types, we aggregated all 24 or 28 RNA-seq replicates of HepG2 or K562 from ENCODE and applied DARTS BHT(flat) to this ultra-deep RNA-seq dataset. Next, we applied DARTS BHT(info) and DARTS BHT(flat) to all possible (24×28) pairwise comparisons between individual replicates of HepG2 and K562, and computed the Area Under Precision Recall Curve (AUPR) for the two methods (Supplementary Table 4). DARTS BHT(info) outperformed DARTS BHT(flat) in all pairwise comparisons, and the performance gain was negatively correlated with the RNA-seq depth of individual replicates (Spearman's rho=−0.69, p-value<2.2e-16), with the largest gain coming from comparisons involving low-coverage RNA-seq samples (Fig. 2a). Thus, incorporating the DNN prediction as prior information improves the detection of cell-type-specific splicing events from low-coverage RNA-seq data.

Next, we determined whether the DARTS DNN can be extended to additional cell types, and how the choice of training datasets influences its performance. We utilized RNA-seq data

from diverse cell types generated by the Roadmap Epigenomics consortium[4]. We performed 253 pairwise comparisons of Roadmap samples (Supplementary Table 5) by DARTS BHT(flat) to generate training data for the DARTS DNN. We excluded all pairwise comparisons involving the thymus tissue from training to use as leave-out data for independent evaluation. We trained three DARTS DNN models, using ENCODE data only, Roadmap data only, or both (Fig. 2b). The DARTS DNN trained on ENCODE data exhibited high predictive power for leave-out ENCODE data but modest predictive power for leave-out Roadmap data. Conversely, the DARTS DNN trained on Roadmap data had high predictive power for leave-out Roadmap data but modest predictive power for leave-out ENCODE data. The DARTS DNN trained on combined ENCODE and Roadmap data had the best performance (Fig. 2b).

We extended the DARTS DNN beyond exon skipping to predict other types of alternative splicing patterns. We compiled *cis* sequence features (Supplementary Table 1) and trained three DNN models for predicting differential alternative 5' splice sites (A5SS), alternative 3' splice sites (A3SS), and retained introns (RI). Trained on ENCODE and Roadmap data, these DNN models exhibited a high prediction accuracy in independent leave-out datasets (Supplementary Fig. 6).

Finally, we applied DARTS to study alternative splicing during the epithelial-mesenchymal transition (EMT), a key process in embryonic development and cancer metastasis[13]. We re-analyzed our published time-course RNA-seq data on an inducible H358 lung cancer cell line model of the EMT[14]. We used DARTS BHT(flat) to compare each day to Day 0, then assessed the ability of the DARTS DNN to predict high-confidence differential vs. unchanged splicing events during the EMT. The DARTS DNN trained on ENCODE +Roadmap data displayed the best performance, followed closely by the DARTS DNN trained on Roadmap data, whereas the DARTS DNN trained on ENCODE data performed least well (Fig. 3a). This was expected, given that the Roadmap data cover epithelial and mesenchymal cell types. The best prediction accuracy (AUROC=0.82) was achieved by the DARTS DNN trained on ENCODE+Roadmap for the Day 6 versus Day 0 comparison. As an example, the DARTS DNN predicted the EMT-associated alternative splicing change in *PLEKHA1* (Supplementary Fig. 7).

To further assess the DARTS DNN predictions, we compiled 449 "DARTS DNN rescued" events from the Day 6 vs. Day 0 comparison (Methods). A subset of these "DARTS DNN rescued" events had significantly reduced exon inclusion during the EMT, and their downstream intronic regions were enriched for the consensus motif of the splicing factors ESRP1/2[15] (Fig. 3b). A similar pattern of ESRP motif enrichment was observed for differential splicing events called by DARTS BHT(flat) using RNA-seq data alone (Fig. 3b). By contrast, events that were called significant by DARTS BHT(flat) but fell below the significance threshold (posterior probability<0.9) after incorporating the informative prior were not enriched for the ESRP motif (Supplementary Fig. 8). ESRPs are epithelial-specific splicing factors whose downregulation is a major driver of alternative splicing during the EMT[14]. This observed pattern of ESRP motif enrichment is consistent with ESRP binding downstream of alternative exons enhancing exon inclusion[13].

To extend our DARTS analysis of the H358 EMT system, we performed paired-end RNA-seq of the PC3E and GS689 prostate cancer cell lines, which have contrasting epithelial vs. mesenchymal characteristics[2, 16]. The DARTS DNN scores of these two EMT systems were highly correlated (Spearman's rho=0.87, p-value<2.2e-16; Fig. 3c), suggesting that the DARTS DNN can capture a core EMT splicing signature.

To assess if DARTS can uncover *bona fide* differential splicing events from lowly expressed genes, we performed targeted splicing profiling using the RASL-seq technology[17] and estimated the absolute difference of PSI values (RASL-|ΔPSI|) for 1,058 alternative splicing events between PC3E and GS689 (Methods). We restricted our further analysis to events with RASL-|ΔPSI| value <0.3. As expected, alternative splicing events called as differential or unchanged using RNA-seq data alone (by DARTS BHT(flat)) displayed the highest or lowest RASL-|ΔPSI| values, respectively (Fig. 3d). For the remaining events called as inconclusive by DARTS BHT(flat), we compiled DARTS DNN-predicted differential events and unchanged events, with high (FPR<5%) and low (FPR>80%) DARTS DNN scores respectively (Supplementary Table 6). DARTS DNN-predicted differential events had significantly larger RASL-|ΔPSI| values than DARTS DNN-predicted unchanged events (p-value=0.035, one-sided Wilcoxon test), with the former group similar to the RNA-seq differential events and the latter group similar to the RNA-seq unchanged events (Fig. 3d). DARTS DNN-predicted differential events were in genes with significantly lower expression levels (p-value=0.001, Wilcoxon test) and had significantly lower RNA-seq coverage (p-value=2.1e-7, Wilcoxon test) compared to differential events called by DARTS BHT(flat) (Supplementary Fig. 9a,b). Collectively, among the events analyzed by RASL-seq, DARTS DNN predicted 52 additional differential splicing events, beyond the 77 events called using RNA-seq data alone. Moreover, on RNA-seq inconclusive events with high or low DARTS DNN scores, we used RASL-seq to define the ground truth with RASL-|ΔPSI|>5% as differential and RASL-|ΔPSI|<1% as unchanged. We benchmarked the performance of DARTS BHT(info), DARTS BHT(flat), DARTS DNN, as well as rMATS[2] and SUPPA2[18] that adopted alignment-based vs alignment-free strategies for quantifying splicing using RNA-seq data. DARTS BHT(info) consistently outperformed baseline methods that use RNA-seq data alone to call differential splicing (Supplementary Fig. 9c-d). These data suggest that by combining deep learning predictions with empirical evidence in user-specific RNA-seq data, DARTS can uncover alternative splicing changes in lowly expressed genes and expand the findings beyond a conventional RNA-seq splicing analysis.

Methods

DARTS Bayesian hypothesis testing (BHT) framework

We developed DARTS BHT, a Bayesian statistical framework to determine the statistical significance of differential splicing events or unchanged splicing events between RNA-seq data of two biological conditions. The DARTS BHT framework was designed to integrate deep learning based prediction as prior and empirical evidence in a specific RNA-seq dataset as likelihood. We start by modelling the simplest case, i.e. testing the difference in exon inclusion levels (PSI values) between two conditions without replicates, i.e. one sample per condition (for model with replicates, see Supplementary Notes):

$$I_{ij}|\psi_{ij} \sim \text{Binomial}\left(n = I_{ij} + S_{ij}, p = f_i(\psi_{ij})\right)$$

$$\psi_{i1} = \mu_i$$

$$\psi_{i2} = \mu_i + \delta_i$$

$$\mu_i \sim \text{Unif}(0, 1)$$

$$\delta_i \sim N\left(0, \tau^2\right)$$

Where I_{ij}, S_{ij} and ψ_{ij} are the exon inclusion read count, the exon skipping read count, and the exon inclusion level for exon i in sample group $j \in \{1, 2\}$, respectively; f_i is the length normalization function for exon i that accounts for the effective lengths of the exon inclusion and skipping isoforms[2]; μ_i is the baseline inclusion level for exon i, and δ_i is the expected difference of the exon inclusion levels between the two conditions. The goal of the differential splicing analysis is to test whether the difference in exon inclusion levels between the two conditions δ_i exceeds a user-defined threshold c (e.g. 5%) with a high probability, i.e.

$$P\left(|\delta_i| > c|I_{ij}, S_{ij}\right) \approx 1$$

In Bayesian statistics, this test can be approached by assuming a “spike-and-slab” prior for the parameter of interest δ. The spike-and-slab prior is a two-component mixture prior distribution, with the “spike” component depicting the probability of the model parameter δ being constrained around zero, and the “slab” component depicting the unconstrained distribution of the model parameter δ.

In the DARTS BHT statistical framework, we impose a spike prior H_0 with a small variance $\tau = \tau_0$, such that the probability of δ concentrates around 0, to account for random biological or technical fluctuations in PSI values between two biological conditions for unchanged splicing events. We impose a slab prior H_1 with a much larger variance $\tau = \tau_1$ to model the difference in PSI values between two conditions for differential splicing events. We set $\tau_0 = 0.03$, corresponding to 90% density constrained within $\delta \in [-0.05, 0.05]$, and $\tau_1 = 0.3$; we note that the final inference is robust to choice of τ values (for more details, see Supplementary Notes and Supplementary Fig. 10). The posterior probability of a splicing event being generated by the two models can be written as:

$$P(H_1|I_{ij}, S_{ij}) = \frac{1}{Z} P(H_1) \cdot P(I_{ij}, S_{ij} \Big| H_1)$$

$$P(I_{ij}, S_{ij}|H_1) = \int_{\delta} \int_{\mu} P(I_{ij}, S_{ij}|\mu_i, \delta_i) \cdot P(\mu_i, \delta_i|H_1) d\mu_i d\delta_i$$

$$P(H_0|I_{ij}, S_{ij}) = \frac{1}{Z} P(H_0) \cdot P(I_{ij}, S_{ij}|H_0)$$

$$P(I_{ij}, S_{ij}|H_0) = \int_{\delta} \int_{\mu} P(I_{ij}, S_{ij}|\mu_i, \delta_i) \cdot P(\mu_i, \delta_i|H_0) d\mu_i d\delta_i$$

Where $P(H_1)$ is the prior probability of exon i being differentially spliced, determined by exon-specific *cis* features and sample-specific *trans* RBP expression levels in the two biological conditions, which is independent of the observed RNA-seq read counts. $P(H_0) = 1 - P(H_1)$ is the prior probability of exon i being unchanged. $P(I_{ij}, S_{ij}|H_1)$ and $P(I_{ij}, S_{ij}|H_0)$ represent the likelihoods under the model of differential splicing or unchanged splicing respectively. Z is a normalizing constant.

Since we are comparing only two models, we can further re-write the above equation as a factorization of the ratios between prior and likelihood:

$$\frac{P(H_1 \Big| I_{ij}, S_{ij})}{P(H_0 \Big| I_{ij}, S_{ij})} = \frac{P(H_1)}{P(H_0)} \cdot \frac{P(I_{ij}, S_{ij} \Big| H_1)}{P(I_{ij}, S_{ij} \Big| H_0)}$$

Note that when the prior distribution is flat, i.e. $P(H_0) = P(H_1) = 0.5$, the above equation is equivalent to a likelihood ratio test, which we refer to as DARTS BHT(flat). When $P(H_0)$ and $P(H_1)$ incorporate an informative prior based on exon- and sample-specific predictive features, we refer to this DARTS BHT model as DARTS BHT(info).

Finally, using the equation above, we can derive the marginal posterior probability $P(\delta_i | I_{ij}, S_{ij})$ for the parameter of interest δ_i as a mixture of the posterior conditioned on the two models:

$$P(\delta_i|I_{ij}, S_{ij}) = P(\delta_i|H_1, I_{ij}, S_{ij}) \cdot P(H_1|I_{ij}, S_{ij}) + P(\delta_i|H_0, I_{ij}, S_{ij}) \cdot P(H_0 \Big| I_{ij}, S_{ij})$$

Hence, the final inference is performed on the probability $P(|\delta_i| > c | I_{ij}, S_{ij})$. In our analysis, we set c=0.05 (i.e. a 5% change in exon inclusion level) and call events with $P(|\delta_i| > 0.05 | I_{ij}, S_{ij}) > 0.9$ as significant differential splicing events and

$P(|\delta_i| > 0.05 \mid I_{ij}, S_{ij}) < 0.1$ as significant unchanged splicing events. Events with $0.1 \leq P(|\delta_i| > 0.05 \mid I_{ij}, S_{ij}) \leq 0.9$ are deemed as inconclusive. In the following text, we omit the subscripts and use $P(|\delta_i| > c \mid I_{ij}, S_{ij})$ and $P(|\Delta\psi| > c)$ interchangeably.

DARTS deep neural network (DNN) model for predicting differential alternative splicing

A core component of the DARTS BHT framework is a deep neural network (DNN) model that generates a probability of differential splicing between two biological conditions using exon- and sample-specific predictive features. We designed the DARTS DNN to predict differential splicing of a given exon based on exon-specific *cis* sequence features and sample-specific *trans* RBP expression levels in two biological conditions.

As noted above, a useful feature of the DARTS BHT framework is its capability to determine the statistical significance of both differential splicing events and unchanged splicing events. Specifically, for a splicing event i in the comparison k between RNA-seq datasets from two distinct biological conditions ($j \in \{1, 2\}$), let $Y_{ik} = 1$ if this event is differentially spliced (i.e. H_1 is true); and $Y_{ik} = 0$ if H_0 is true as labels for differential or unchanged splicing events respectively. The task of predicting differential splicing can be formulated as:

$$P(Y_{ik} = 1) = F(Y_{ik}; E_i, G_k)$$

Where Y_{ik} is the label for event i in the comparison k; E_i is a vector of 2,926, 2,973, 2,971, and 1,748 *cis* sequence features for event i, including evolutionary conservation, splice site strength, regulatory motif composition, and RNA secondary structure for skipped exons, alternative 5' splice sites, alternative 3' splice sites, and retained introns, respectively. G_k is a vector of 2,996 (=1,498×2) normalized gene expression levels of 1,498 RBPs in the two conditions. See Supplementary Table 1 for a full list of the features. The prediction of $P(Y_{ik} = 1)$ based on the features from any specific RNA-seq dataset can then be incorporated as an informative prior for $P(H_1)$ in the DARTS BHT framework.

We implemented a deep learning model (DARTS DNN) to learn the unknown function F that maps the predictive features to splicing profiles (differential vs. unchanged). We designed the DARTS DNN with 4 hidden layers and 7,923,402 parameters. The configuration of the DNN was: an input layer with 5922(=2926+1498*2) variables; 4 fully-connected hidden layers with 1200, 500, 300, 200 variables and the ReLU activation function; and an output layer with 2 variables and the Softmax activation function. We implemented the DARTS DNN using Keras (https://github.com/keras-team/keras) with the Theano backend.

To mitigate potential overfitting of the DARTS DNN, we added a drop-out probability[19] for connections between hidden layers. Specifically, the variables in the four hidden layers were randomly turned off during the training process with probability 0.6, 0.5, 0.3, and 0.1, respectively. We also added batch-normalization layers[20] for all hidden layers to help the model converge and generalize. Finally, we used the RMSprop optimizer to adaptively

adjust for the magnitudes of the components of the gradient in this deep architecture and chose 1000 labelled alternative splicing events as one mini-batch to obtain a more stable gradient. In each mini-batch we balanced the composition of positive and negative labels by adding more positive events in the mini-batch such that positive : negative = 1:3 in the mini-batch. Since there were significantly more negative (unchanged) events compared to positive (differential) events, such a balanced composition will provide a gradient for learning the positive events in different mini-batches.

To monitor the training loss and validation loss, we computed the loss every 10 mini-batches and saved the current model parameters if the validation loss was lower than the previous best model. We trained the DARTS DNN on Tesla K40m.

Processing of ENCODE RNA-seq data and training of the DARTS DNN model

We used a comprehensive RNA-seq dataset from the ENCODE consortium to train the DARTS DNN. The ENCODE investigators have performed systematic shRNA knockdown of over 250 RBPs in two human cell lines HepG2 and K562. We downloaded all available (as of May 2017) RNA-seq alignments (ENCODE processing pipeline on the human genome version hg19) for shRNA knockdown and control samples from the ENCODE data portal (https://www.encodeproject.org/).

We processed the RNA-seq alignments (bam files) using rMATS[2] (v4.0.1). Given RNA-seq alignment files, rMATS constructs splicing graphs, detects annotated and novel alternative splicing events, and counts the number of RNA-seq reads for each exon and splice junction. Given the modest depth of the ENCODE RNA-seq data (32 million read pairs per replicate on average), the read counts from the two replicates were pooled together for downstream analyses.

We processed the raw RNA-seq reads with Kallisto[21] (v0.43.0) to quantify gene expression levels using Gencode[22] (v19) protein coding transcripts as the index. For each of the two biological conditions in a given comparison (i.e. shRNA knockdown vs. control), we extracted the Kallisto derived gene-level TPM values of 1,498 known RBPs[11]. The TPM value of each RBP was normalized by dividing by its maximum observed TPM value of all comparisons, then used as RBP expression features by the DARTS DNN.

To generate training labels for the DARTS DNN, DARTS BHT(flat) was applied to the ENCODE RNA-seq data. Events with posterior probability $P(|\Psi| > 0.05) > 0.9$ were called positive (Y=1). Events with posterior probability $P(|\Psi| > 0.05) < 0.1$ were called negative (Y=0). We defined these significant differential splicing events and significant unchanged splicing events as labelled events and used them to train the DARTS DNN.

The vast majority of the RBPs (n=196) in the ENCODE data were knocked-down by at least one shRNA in both HepG2 and K562 cell lines, corresponding to a total of 408 comparisons between knockdown and control. We set aside 10% of the labelled positive events and the same number of labelled negative events in each comparison as the testing data for estimating the generalization error of the trained DNN model. For the remaining 90% of the labelled events, we further split them into 5-fold cross-validation subsets for the purposes of

training, monitoring overfitting, and early-stopping. We also collected ENCODE RBP knockdown experiments performed in only one cell line (either HepG2 or K562, n=58) as leave-out datasets. All labelled events in these leave-out datasets were only utilized for evaluating the trained DARTS DNN and were never used during training.

We randomly drew 4 RBPs without replacement for a training batch, and iterated through all 196 RBPs as an epoch. The performance of the DARTS DNN was measured by Area Under the Receiver Operating Characteristics curve (AUROC). The model with the best performance during training and cross-validation was selected, and subsequently benchmarked using the testing data and leave-out data.

Rank-transformation of the DARTS informative prior

In a typical RNA-seq study, the number of unchanged splicing events can be orders of magnitude larger than differential splicing events, and machine learning algorithms may be biased to the majority class. To mitigate this potential bias, we used an unsupervised rank-transformation to rescale DARTS DNN scores to derive the informative prior for the DARTS BHT framework. Specifically, we first fit a two-component Gaussian mixture model for all the DARTS DNN scores to derive the mean and variance of the two mixed Gaussian components as well as the posterior probability λ of each DARTS DNN score belonging to a specific component. Setting the new mean and variance of the two Gaussian components to μ_0 and μ_1, σ_0 and σ_1, respectively, each DARTS DNN score was rank-transformed to the new Gaussian components and then averaged by the weight parameter λ. Finally, to maintain a valid prior probability, the transformed DARTS DNN scores were rescaled to $[\alpha,\ 1-\alpha]$, where $\alpha \in [0, 0.5)$ sets the desired prior strength for the DARTS BHT framework and a smaller α value corresponds to a stronger strength of the informative prior. Using this rescaling scheme, the entire ranks of the DARTS DNN scores are preserved while the potential bias for negative over positive events is reduced. In practice, we set $\mu_0 = 0.05$, $\mu_1 = 0.95$, $\sigma_0 = \sigma_1 = 0.1$, and $\alpha = 0.05$.

Generalization of the DARTS framework to diverse tissues and cell types

We generalized the DARTS framework to incorporate diverse tissues and cell types by utilizing RNA-seq resources from the Roadmap Epigenomics project[4]. The Roadmap data was processed following the same protocol as for the ENCODE data. We took all Roadmap data with 101bp x 2 or 100bp x 2 paired-end RNA-seq, and truncated reads from the 101bp x 2 datasets to 100bp for rMATS. In total, this represented 23 distinct tissues or cell types. All possible pairwise comparisons (n=253) between these 23 RNA-seq samples were performed. Comparisons involving thymus were held out as Roadmap leave-out data, and all remaining comparisons were used as training datasets.

We trained three DARTS DNN models using different training datasets: i) ENCODE data only, ii) Roadmap data only, and iii) the combination of ENCODE+Roadmap data. The performances of the three models were subsequently benchmarked by using ENCODE or Roadmap leave-out datasets.

DARTS splicing analyses of EMT-associated RNA-seq datasets

We applied the trained DARTS model to study EMT-associated alternative splicing events in two distinct human cell culture systems: H358 lung cancer cell line induced to undergo EMT through a 7-day time course[14], and PC3E/GS689 prostate cancer cell lines that had contrasting epithelial versus mesenchymal characteristics[2, 16].

For the H358 time-course RNA-seq data (GSE75492), we used DARTS BHT(flat) to compare RNA-seq data from Day 1 to Day 7 against Day 0. Splicing events that displayed a high DARTS DNN score of differential splicing (FPR<5%) and a non-trivial splicing change (over 10% difference in exon inclusion level), but did not pass the significance threshold by DARTS BHT(flat) using observed RNA-seq read counts alone were defined as DARTS DNN rescued events. Motif analysis was performed by calculating the average percentage of nucleotides covered by any of the top 12 ESRP SELEX-seq hexamer motifs[15] in a 45bp sliding window. Background sequences were significant unchanged events by DARTS BHT(flat). For the PC3E and GS689 cell lines, we conducted RASL-seq[17] and RNA-seq experiments on the same batch of RNA samples, each with 3 replicates and on average 125 million read pairs per replicate (raw data deposited as GSE112037). RASL-seq reads were aligned to the pool of target splice junctions in the RASL-seq library using Blat[23]. RASL-PSI values were calculated as $\frac{I}{I+S}$, where I is the number of exon inclusion splice junction reads and S is the number of exon skipping splice junction reads. Alternative splicing events with total RASL-seq read counts larger than 5 in every replicate were used for downstream analyses. Gene expression levels of RBPs in the two datasets were quantified using Kallisto v0.43.0.

RASL-seq library preparation and sequencing

RASL-seq was performed as described[24] with some modifications. Total RNA from PC3E and GS689 cell lines were extracted with Trizol (Thermo Fisher Scientific). RASL-seq oligonucleotides (a gift from Xiang-Dong Fu, UCSD) were annealed to 1 μg of total RNA, followed by selection by oligo-dT beads. Paired probes templated by polyA$^+$ RNA were ligated and then eluted. 5 μl of the eluted ligated oligos were used for 8 cycles of PCR amplification using primers F1: 5'-CCGAGATCTACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3' and R1: 5'-GTGACTGGAGTTCAGACGTGTGCGCTGATGCTACGACCACAGG-3'. One third of the resulting PCR products were used in the second round of PCR amplification (9 cycles) using primers F2: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG-3' and R2: 5'-CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTTCAGACGTGTGC-3'; indexes used in this study were Illumina indexes D701-D706. The indexed PCR products were pooled and sequenced on a Miseq with a custom sequencing primer 5'-ACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3' and a custom index sequencing primer 5'-TAGCATCAGCGCACACGTCTGAACTCCAGTCAC-3'.

Data Availability

The RNA-seq data that support the findings of the deep learning models are available from the ENCODE project (https://www.encodeproject.org/) and the Roadmap Epigenomics project (http://www.roadmapepigenomics.org/). The H358 time-course RNA-seq data were downloaded from GEO with accession ID GSE75492. The PC3E-GS689 RNA-seq data and RASL-seq data can be accessed from GEO with accession ID GSE112037.

Code Availability

The DARTS program, trained model parameters, and predictive features are provided at GitHub (https://github.com/Xinglab/DARTS).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Xiang-Dong Fu (UCSD) for the RASL oligos and advice on RASL-seq. This study is supported by National Institutes of Health grants (R01GM088342, R01GM117624, U01HG007912, and U01CA233074 to Y.X.). Z.Z. is partially supported by a UCLA Dissertation Year Fellowship.

References

1. Katz Y, Wang ET, Airoldi EM & Burge CB Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7, 1009–1015 (2010). [PubMed: 21057496]
2. 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). [PubMed: 25480548]
3. Consortium EP An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
4. Roadmap Epigenomics C et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]
5. Cieslik M & Chinnaiyan AM Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet 19, 93–109 (2018). [PubMed: 29279605]
6. Park E, Pan Z, Zhang Z, Lin L & Xing Y The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am J Hum Genet 102, 11–26 (2018). [PubMed: 29304370]
7. Xiong HY et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806 (2015). [PubMed: 25525159]
8. Barash Y et al. Deciphering the splicing code. Nature 465, 53–59 (2010). [PubMed: 20445623]
9. Leung MK, Xiong HY, Lee LJ & Frey BJ Deep learning of the tissue-regulated splicing code. Bioinformatics 30, i121–129 (2014). [PubMed: 24931975]
10. Huang Y & Sanguinetti G BRIE: transcriptome-wide splicing quantification in single cells. Genome biology 18, 123 (2017). [PubMed: 28655331]
11. Gerstberger S, Hafner M & Tuschl T A census of human RNA-binding proteins. Nat Rev Genet 15, 829–845 (2014). [PubMed: 25365966]
12. Van Nostrand EL et al. A large-scale binding and functional map of human RNA binding proteins. bioRxiv, 179648 (2017).
13. Warzecha CC et al. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. EMBO J 29, 3286–3300 (2010). [PubMed: 20711167]
14. Yang Y et al. Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. Mol Cell Biol 36, 1704–1719 (2016). [PubMed: 27044866]

15. Dittmar KA et al. Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. Mol Cell Biol 32, 1468–1482 (2012). [PubMed: 22354987]
16. Lu ZX et al. Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. Mol Cancer Res 13, 305–318 (2015). [PubMed: 25274489]
17. Li H, Qiu J & Fu XD RASL-seq for massively parallel and quantitative analysis of gene expression. Curr Protoc Mol Biol Chapter 4, Unit 4 13 11–19 (2012).
18. Trincado JL et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol 19, 40 (2018). [PubMed: 29571299]
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I & Salakhutdinov R Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958 (2014).
20. Ioffe S & Szegedy C in International conference on machine learning 448–456 (2015).
21. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527 (2016). [PubMed: 27043002]
22. Harrow J et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1, S4 1–9 (2006).
23. Kent WJ BLAT--the BLAST-like alignment tool. Genome Res 12, 656–664 (2002). [PubMed: 11932250]
24. Ying Y et al. Splicing Activation by Rbfox Requires Self-Aggregation through Its Tyrosine-Rich Domain. Cell 170, 312–323 e310 (2017). [PubMed: 28708999]

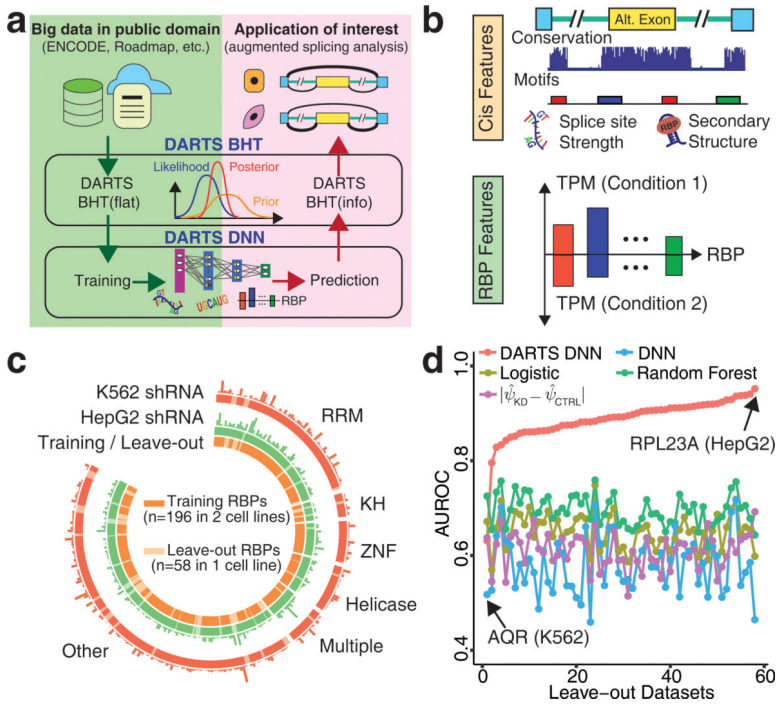

Figure 1.

The DARTS computational framework for deep learning-augmented RNA-seq analysis of transcript splicing. **(a)** Overall workflow of DARTS. (**b**) Schematic illustration of the DARTS DNN features, including *cis* sequence features and *trans* RBP features. **(c)** Overview of training and leave-out RBPs, and the number of significant differential splicing events called by DARTS BHT(flat) on the ENCODE data (illustrated by bar charts above the outer and middle circles). 196 RBPs knocked-down in both the K562 and HepG2 cell lines are used for training (orange), while the remaining 58 RBPs knocked-down in only one cell line are leave-out data (light orange) (illustrated in the inner circle). **(d)** Comparison of the DARTS DNN with baseline methods in leave-out datasets.

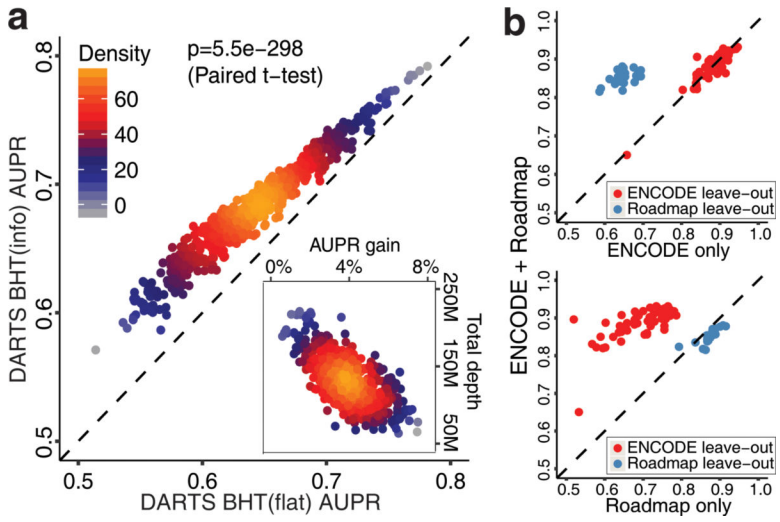

Figure 2.

The performance evaluation of the DARTS Bayesian Hypothesis Testing (BHT) framework and the influence of training datasets on the performance of the DARTS DNN. **(a)** The performance of DARTS BHT(info) vs. DARTS BHT(flat) in the cell-type-specific differential splicing analysis of HepG2 and K562 (two-sided paired t-test, n=672 pairwise comparisons). The performance gain by DARTS BHT(info) is plotted against the RNA-seq depth in pairwise comparisons of individual replicates (inset). **(b)** AUROC values of the DARTS DNN trained on ENCODE+Roadmap data, ENCODE data only, or Roadmap data only when applied to ENCODE or Roadmap leave-out data.

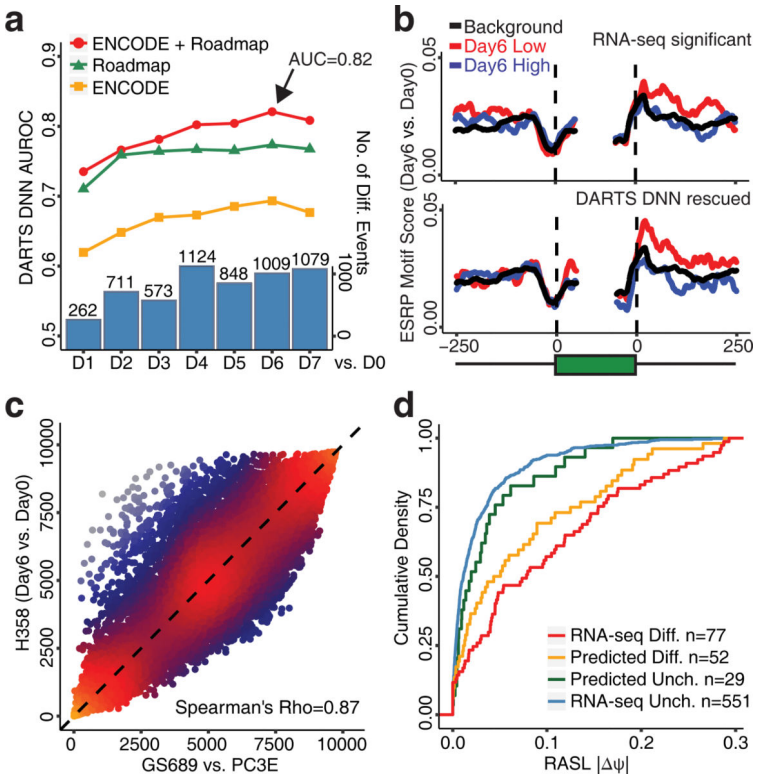

Figure 3.

DARTS analysis of alternative splicing during the EMT. (**a**) The performance of the DARTS DNN on the time-course RNA-seq data of an inducible H358 lung cancer cell line model of the EMT. The numbers of differential splicing events called by DARTS BHT(flat) are shown as bar plots at the bottom. (**b**) Meta-exon motif analysis of the ESRP motif for RNA-seq differential events called by DARTS BHT(flat) and DARTS DNN rescued events in the Day 6 vs. Day 0 comparison. (**c**) DARTS DNN predictions for the H358 EMT time course (Day 6 vs. Day 0) and in GS689 vs. PC3E. Plotted are the ranks of predicted DARTS DNN scores. (**d**) RASL-seq validation of RNA-seq called events and DARTS DNN predicted events. Plotted are the RASL-|ΔPSI| values of RNA-seq inconclusive events with high DARTS DNN scores (FPR<5%; n=52 events) (orange line) and RNA-seq inconclusive events with low DARTS DNN scores (FPR>80%; n=29 events) (green line).