nature methods

BRIEF COMMUNICATION

https://doi.org/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] and Yi Xing[1,2,3,7*]

A major limitation of RNA sequencing (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, thereby 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 (for example, ENCODE[3], Roadmap Epigenomics[4]) provides unprecedented resources for characterizing alternative splicing 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 genes with low expression[6].

Motivated by recent successes in the use of 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 used 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 on the basis of 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)) for 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 messenger RNA (mRNA) levels of trans RNA-binding proteins (RBPs) in two conditions (Fig. 1b and 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 used 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 short-hairpin RNA (shRNA) in both cell lines, corresponding to 408 knockdown-versus-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 for independent evaluation of 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[1] and rMATS[2] (Supplementary Figs. 2 and 3). From the high-confidence differentially spliced versus unchanged exons called by DARTS BHT(flat) (Supplementary Table 2), we used 90% of labeled events for training and fivefold cross-validation, and the remaining 10% of 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 with 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 fivefold cross-validation in each leave-out dataset. Additionally, we implemented another alternative baseline method by predicting sample-specific exon-inclusion

[1]Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [3]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA. [4]Department of Molecular and Medical Pharmacology, University of California, Los Angeles, Los Angeles, CA, USA. [5]Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA. [7]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA. [8]These authors contributed equally: Zijun Zhang, Zhicheng Pan. *e-mail: XINGYI@email.chop.edu

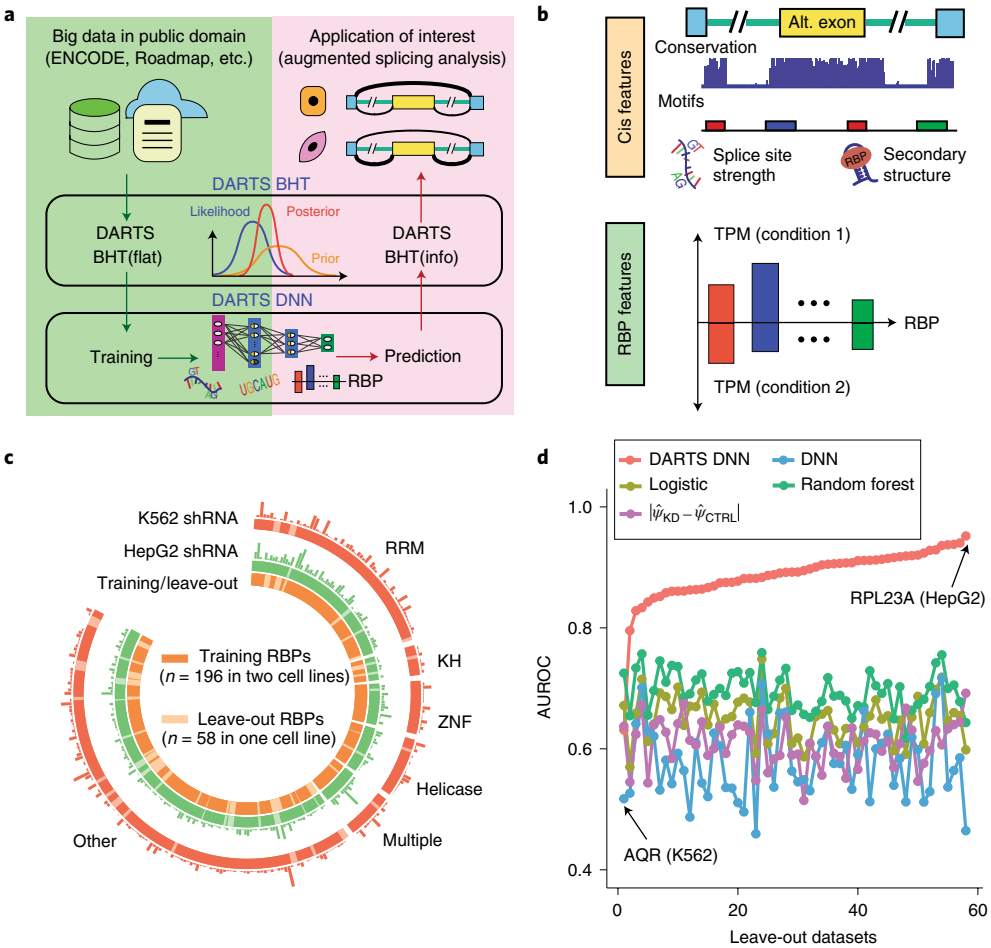

Fig. 1 | The DARTS computational framework. a, Overall workflow of DARTS. **b**, Schematic of the DARTS DNN features, including cis sequence features and trans RBP features. **c**, Overview of training and leave-out RBPs, and the number of significant differential splicing events called by DARTS BHT(flat) on the ENCODE data (illustrated by bar charts above the outer and middle circles). We used 196 RBPs knocked down in both the K562 and HepG2 cell lines for training (orange), while the remaining 58 RBPs knocked down in only one cell line were leave-out data (light orange) (illustrated in the inner circle). RRM, RNA recognition motif; KH, K homology; ZNF, zinc finger. **d**, Comparison of the DARTS DNN with baseline methods in leave-out datasets. KD, knockdown; CTRL, control; RPL23A, ribosomal protein L23a; AQR, aquarius intron-binding spliceosomal factor.

levels (PSI values; percent spliced in, or ψ)[1,10] and 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 out of 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 on the basis of cis sequence features and expression levels of trans RBPs; and the likelihood ratio, determined by modeling of the biological variation and estimation uncertainty of the splice isoform ratio based on observed RNA-seq read counts. Simulation studies demonstrated that the informative prior improves the inference when the observed data are limited, for instance, because of low levels of gene expression 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 the 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 $\rho = -0.69$, $P < 2.2 \times 10^{-16}$), 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 used RNA-seq data from diverse cell types generated by the Roadmap Epigenomics consortium[4]. We performed 253 pairwise comparisons of Roadmap samples

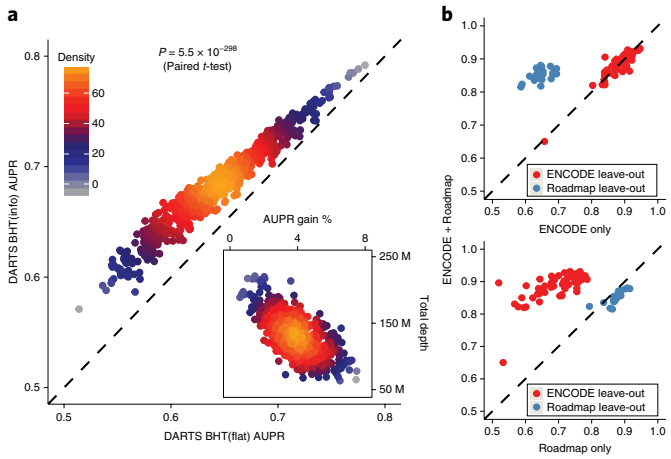

Fig. 2 | Performance evaluation of the DARTS BHT framework, and the influence of training datasets on the performance of the DARTS DNN. **a**, The performance of DARTS BHT(info) versus DARTS BHT(flat) in the cell-type-specific differential splicing analysis of HepG2 and K562 (two-sided paired *t*-test; $n = 672$ 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 both ENCODE and Roadmap data, ENCODE data only, or Roadmap data only when applied to ENCODE or Roadmap leave-out data.

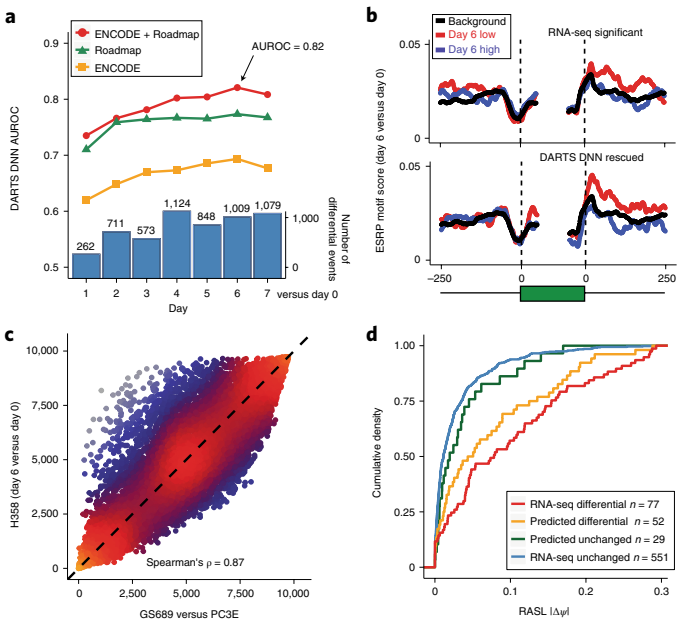

Fig. 3 | DARTS analysis of alternative splicing during the EMT. **a**, The performance of the DARTS DNN on the time-course RNA-seq data of an inducible H358 lung cancer cell line model of the EMT. The numbers of differential splicing events called by DARTS BHT(flat) are shown as bar plots at the bottom. **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 comparison of day 6 versus day 0. **c**, DARTS DNN predictions for the H358 EMT time course (day 6 versus day 0) and in GS689 versus 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 cumulative density functions of 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).

(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, alternative 3′ splice sites, and retained introns. 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 reanalyzed 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 versus unchanged splicing events during the EMT. The DARTS DNN trained on combined ENCODE and Roadmap data had 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 combined ENCODE and Roadmap data for the comparison of day 6 versus day 0. 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 comparison of day 6 versus day 0 (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 epithelial splicing regulatory proteins 1 and 2 (ESRP1 and ESRP2; ref. [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 incorporation of the informative prior were not enriched for the ESRP motif (Supplementary Fig. 8). ESRPs are epithelial-specific splicing factors, the downregulation of which 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 carried out paired-end RNA-seq of the PC3E and GS689 prostate-cancer cell lines, which have contrasting epithelial versus mesenchymal characteristics[2,16]. The DARTS DNN scores of these two EMT systems were highly correlated (Spearman's $\rho = 0.87$, $P < 2.2 \times 10^{-16}$; Fig. 3c), suggesting that the DARTS DNN can capture a core EMT splicing signature.

To assess whether DARTS can uncover bona fide differential splicing events from genes expressed at low levels, we carried out

targeted splicing profiling using the RNA-mediated oligonucleotide annealing, selection, and ligation with next-generation sequencing (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 where RASL – |ΔPSI| < 0.3. As expected, alternative splicing events called as differential or unchanged by RNA-seq data alone (by DARTS BHT(flat)) had 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 and low DARTS DNN scores (false positive rate or FPR < 5% and > 80%), respectively (Supplementary Table 6). DARTS-DNN-predicted differential events had significantly greater RASL – |ΔPSI| values than DARTS-DNN-predicted unchanged events ($P = 0.035$, 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 ($P = 0.001$, two-sided Wilcoxon test) and had significantly lower RNA-seq coverage ($P = 2.1 \times 10^{-7}$, two-sided Wilcoxon test) compared with 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 with 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, rMATS[2] and SUPPA2[18] that adopted alignment-based versus alignment-free strategies for quantifying splicing with 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 genes with low expression and expand the findings beyond a conventional RNA-seq splicing analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0351-9.

Received: 5 September 2018; Accepted: 1 February 2019;
Published online: 25 March 2019

References

1. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. *Nat. Methods* **7**, 1009–1015 (2010).
2. Shen, S. et al. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
3. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
4. Kundaje, A. et al. *Nature* **518**, 317–330 (2015).
5. Cieślik, M. & Chinnaiyan, A. M. *Nat. Rev. Genet.* **19**, 93–109 (2018).
6. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. *Am. J. Hum. Genet.* **102**, 11–26 (2018).
7. Xiong, H. Y. et al. *Science* **347**, 1254806 (2015).
8. Barash, Y. et al. *Nature* **465**, 53–59 (2010).
9. Leung, M. K., Xiong, H. Y., Lee, L. J. & Frey, B. J. *Bioinformatics* **30**, i121–i129 (2014).
10. Huang, Y. & Sanguinetti, G. *Genome. Biol.* **18**, 123 (2017).
11. Gerstberger, S., Hafner, M. & Tuschl, T. *Nat. Rev. Genet.* **15**, 829–845 (2014).
12. Van Nostrand, E. L. et al. Preprint at bioRxiv https://www.biorxiv.org/content/10.1101/179648v1?versioned=true (2017).
13. Warzecha, C. C. et al. *EMBO J.* **29**, 3286–3300 (2010).
14. Yang, Y. et al. *Mol. Cell. Biol.* **36**, 1704–1719 (2016).
15. Dittmar, K. A. et al. *Mol. Cell. Biol.* **32**, 1468–1482 (2012).
16. Lu, Z. X. et al. *Mol. Cancer Res.* **13**, 305–318 (2015).
17. Li, H., Qiu, J. & Fu, X.-D. *Curr. Protoc. Mol. Biol.* **98**, 13.1–4.13.9 (2012).
18. Trincado, J. L. et al. *Genome. Biol.* **19**, 40 (2018).

Acknowledgements

We thank X.-D. 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.

Author contributions

Z.Z. and Y.X. conceived the study; Z.Z., Y.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 other authors.

Competing interests

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

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0351-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Y.X.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

DARTS Bayesian hypothesis testing 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 modeling the simplest case, that is, testing the difference in exon-inclusion levels (PSI values) between two conditions without replicates (one sample per condition; for model with replicates, see the Supplementary Notes):

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

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

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

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

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

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 (for example, 5%) with a high probability:

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

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 (further details are provided in the Supplementary Notes and Supplementary Fig. 10). The posterior probability of a splicing event being generated by the two models can be written as

$$p(H_1 \mid I_{ij}, S_{ij}) = \frac{1}{Z} p(H_1) \times p(I_{ij}, S_{ij} | H_1)$$

$$p(I_{ij}, S_{ij} \mid H_1) = \int_{\delta} \int_{\mu} p(I_{ij}, S_{ij} \mid \mu_i, \delta_i) \times p(\mu_i, \delta_i \mid H_1)\, d\mu_i d\delta_i$$

$$p(H_0 \mid I_{ij}, S_{ij}) = \frac{1}{Z} p(H_0) \times p(I_{ij}, S_{ij} \mid H_0)$$

$$p(I_{ij}, S_{ij} \mid H_0) = \int_{\delta} \int_{\mu} p(I_{ij}, S_{ij} \mid \mu_i, \delta_i) \times p(\mu_i, \delta_i \mid H_0)\, d\mu_i d\delta_i$$

where $p(H_1)$ is the prior probability of exon i being differentially spliced, determined by exon-specific cis features and sample-specific expression levels of trans RBPs 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 and unchanged splicing, respectively. Z is a normalizing constant.

Because we are comparing only two models, we can further rewrite the above equation as a factorization of the ratios between prior and likelihood:

$$\frac{p(H_1|I_{ij}, S_{ij})}{p(H_0|I_{ij}, S_{ij})} = \frac{p(H_1)}{p(H_0)} \times \frac{p(I_{ij}, S_{ij}|H_1)}{p(I_{ij}, S_{ij}|H_0)}$$

Note that when the prior distribution is flat, that is, $p(H_0) = p(H_1) = 0.5$, 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:

$$p(\delta_i \mid I_{ij}, S_{ij}) = p(\delta_i \mid H_1, I_{ij}, S_{ij}) \times p(H_1 \mid I_{ij}, S_{ij}) + p(\delta_i \mid H_0, I_{ij}, S_{ij}) \times p(H_0 | I_{ij}, S_{ij})$$

Hence, the final inference is performed on the probability $p(|\delta_i| > c|I_{ij}, S_{ij})$. In our analysis, we set $c = 0.05$ (that is, 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 $p(|\delta_i| > 0.05|I_{ij}, S_{ij}) < 0.1$ as significant unchanged splicing events. Events with $0.1 \le p(|\delta_i| > 0.05|I_{ij}, S_{ij}) \le 0.9$ are deemed inconclusive. In the following text, we omit the subscripts and use $p(|\delta_i| > c|I_{ij}, S_{ij})$ and $p(|\Delta\psi| > c)$ interchangeably.

DARTS DNN model for predicting differential alternative splicing. A core component of the DARTS BHT framework is a 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 (that is, H_1 is true) and $Y_{ik} = 0$ if H_0 is true as labels for differential and unchanged splicing events, respectively. The task of predicting differential splicing can be formulated as

$$p(Y_{ik} = 1) = \text{F}(Y_{ik}; E_i, G_k)$$

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; and G_k is a vector of 2,996 (that is, $1{,}498 \times 2$) 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 versus unchanged). For skipped exons, we designed the DARTS DNN with four hidden layers and 7,923,402 parameters. The configuration of the DNN was as follows: an input layer with 5,922 (that is, $2{,}926 + 1{,}498 \times 2$) variables; four fully connected hidden layers with 1,200, 500, 300, and 200 variables and the ReLU activation function; and an output layer with two variables and the Softmax activation function. We implemented the DARTS DNN using Keras (https://github.com/keras-team/keras) with the Theano back-end.

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 probabilities of 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 1,000 labeled 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. Because there were significantly more negative (unchanged) events than 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 ten mini-batches and saved the current model parameters if the validation loss was lower than that of 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 more than 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] (v.4.0.1). Starting with 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] (v.0.43.0) to quantify gene expression levels using Gencode[22] (v.19) protein-coding transcripts as the index. For each of the two biological conditions in a given comparison (that is, shRNA knockdown and control), we extracted the Kallisto-derived gene-level transcripts per kilobase million (TPM) values of 1,498 known RBPs[11]. We normalized the TPM value of each RBP by dividing by its maximum observed TPM value of all comparisons, then used that as the RBP expression feature in the DARTS DNN.

To generate training labels for the DARTS DNN, we applied DARTS BHT(flat) to the ENCODE RNA-seq data. Events with posterior probability $p(|\Delta\psi| > 0.05) > 0.9$ were called positive ($Y = 1$). Events with posterior probability $p(|\Delta\psi| > 0.05) < 0.1$ were called negative ($Y = 0$). We defined these significant differential splicing events and significant unchanged splicing events as labeled 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 labeled positive events and the same number of labeled 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 labeled events, we further split them into fivefold 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 labeled events in these leave-out datasets were used only to evaluate the trained DARTS DNN and never 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 on the basis of the 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 the number of 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. With the new mean and variance of the two Gaussian components set at 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, we rescaled the transformed DARTS DNN scores 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. With 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 using RNA-seq resources from the Roadmap Epigenomics project[4]. The Roadmap data were processed via the same protocol used for the ENCODE data. We took all Roadmap data with 101 bp × 2 or 100 bp × 2 paired-end RNA-seq, and truncated reads from the 101 bp × 2 datasets to 100 bp 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 made. 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: (1) ENCODE data only, (2) Roadmap data only, and (3) the combination of ENCODE and Roadmap data. We subsequently benchmarked the performances of the three models 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 seven-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 that for day 0. Splicing events that displayed a high DARTS DNN score of differential splicing (FPR < 5%) and a non-trivial splicing change (more than 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. We carried out motif analysis by calculating the average percentage of nucleotides covered by any of the top 12 ESRP SELEX-seq hexamer motifs[15] in a 45-bp 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 three replicates and on average 125 million read pairs per RNA-seq 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 $I/(I+S)$, 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 greater than five in every replicate were used for downstream analyses. Gene expression levels of RBPs in the two datasets were quantified with Kallisto v.0.43.0.

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 X.-D. Fu) were annealed to 1 µg of total RNA and then subjected to selection by oligo-dT beads. Paired probes templated by poly(A)$^+$ RNA were ligated and then eluted. We used 5 µl of the eluted ligated oligos for eight 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 (nine cycles) with 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 custom sequencing primer 5′-ACACTCTTTCCCTACACGACGGCGACCACCGAGAT-3′ and custom index sequencing primer 5′-TAGCATCAGCGCACACGTCTGAACTCCAGTCAC-3′.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

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

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 accession GSE75492. The PC3E-GS689 RNA-seq data and RASL-seq data can be accessed from GEO under accession GSE112037.

References

19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
20. Ioffe, S. & Szegedy, C. In *Proc. 32nd International Conference on Machine Learning* (eds Bach, F. & Blei, D.) 448–456 (PMLR/Microtome Publishing, Brookline, MA, USA, 2015).
21. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. *Nat. Biotechnol.* **34**, 525–527 (2016).
22. Harrow, J. et al. *Genome. Biol.* **7**, 1–S4.9 (2006).
23. Kent, W. J. *Genome Res.* **12**, 656–664 (2002).
24. Ying, Y. et al. *Cell* **170**, 312–323 (2017).

nature research

Corresponding author(s): Yi Xing

Last updated by author(s): Jan 23, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Publicly available data were downloaded from ENCODE/Roadmap data portal or from GEO using corresponding GEO accession IDs as stated in the manuscript. Data download was performed by the unix program wget(v1.12).
Data analysis	The DARTS software code, predictive features and trained model parameters are freely available in GitHub as stated in the manuscript. DARTS was developed and tested in Python(v2.7) and R(v3.4.3). DARTS deep learning model was implemented in Keras(v2.0.6). RNA-seq raw data was processed by STAR(v2.5.2a), Kallisto(v0.43.0), rMATS(v4.0.1), MISO(v0.5.3). Simulated RNA-seq data was generated by Flux-simulator(v1.2.1). RASL-seq raw data was processed by Blat(v36x1). Predictive features by prediction tools were generated by NuPoP(1.24.0), Maxent(20-Apr-2004), RNAfold(2.2.10).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

High-throughput sequencing data of PC3E and GS689 cell lines that support the findings in this study were deposited into GEO with the accession ID GSE112037. Publicly available data were downloaded from ENCODE/Roadmap data portal or from GEO using corresponding GEO accession IDs as stated in the manuscript. Computer programs and software are available in GitHub https://github.com/Xinglab/DARTS.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Publicly available data in ENCODE/Roadmap download portal were used (as of May 2017).
Data exclusions	No data were excluded.
Replication	Three biological replicates of RNA-seq experiments were performed on PC3E and GS689 cell lines. No replication failed.
Randomization	This is not relevant to our study, because our study does not involve the assignment of test subjects or treatments.
Blinding	This is not relevant to our study, because our study does not involve the assignment of test subjects or treatments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	PC3E and GS689 cell lines were derived from the parental PC3 cell line (ATCC CRL-1435). PC3E was derived from PC3 cells by fluorescence-activated cell sorting for E-cadherin-positive cells, and GS689 was recovered from a secondary metastatic liver tumor after intravenous injection of PC3 cells into mouse (Mol Cancer Res. 2015 13(2):305-18).
Authentication	Authentication for all cell lines used in this study was performed by Laragen, Inc using the Promega powerplex16 System recommended by American Type Culture Collection. The STR alleles were searched either on ATCC or DSMZ databases depending on availability of the cell lines in the databases.
Mycoplasma contamination	Cell lines were not tested for Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Not applicable.