Supplementary Materials for "A non-negative matrix factorization based preselection procedure for more accurate isoform discovery from RNA-seq data"

Yuting Ye*

Division of Biostatistics, University of California, Berkeley

Jingyi Jessica Li†

Department of Statistics, University of California, Los Angeles

1 Subexons and contradicting bins

1.1 Definition of subexons

Exons are not the minimal splicing units. In some types of alternative splicing, such as alternative 5' ends and alternative 3' ends, splicing can occur inside an exon. Also, there can be differences between the exon boundaries from annotations and those from de novo assembles. Hence to capture slight differences among isoform structures, we split exons into *subexons*, the minimal splicing units. Subexons are defined as non-overlapping transcribed regions between adjacent splicing sites. Every exon in the input annotation or de novo assembly can be fully recovered by a set of subexons. For illustration of subexons, please see Figure 1 extracted from the SLIDE paper[1].

1.2 Contradicting bins

We define *bins* as two-dimensional vectors that describe the exon indices of the starting and ending positions of mapped reads (single-ended reads or paired-end reads decomposed into two ends). For example, Bin $(4, 4)$ contains reads whose starting and ending positions

*Email: yeyt@berkeley.edu

†Email: jli@stat.ucla.edu; Corresponding author

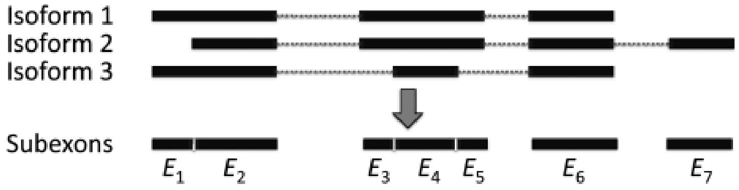


Figure 1: **Definition of the subexon**

are both in Subexon 4. For reads that cannot originate from the same transcript, their corresponding bins are mutually exclusive. We call them *contradicting bins*. For example, Bins $(4, 4)$ and $(3, 5)$ are contradicting bins, because Bin $(4, 4)$ indicates the existence of Subexon 4 but Bin $(3, 5)$ indicates the skipping of Subexon 4.

1.3 Decomposing isoforms candidates containing contradicting bins

After non-negative matrix factorization (NMF) is completed, a basis matrix W would be obtained, and each column of W represents an isoform candidate (See the main text). However, isoform candidates may contain contradicting bins, and such candidates cannot be true isoforms. To resolve this issue without losing possibly true isoforms, we decompose an isoform candidate with two contradicting bins into two isoform candidates, each containing one of the two bins. We use **Figure 1** as an example. Suppose an isoform candidate contains contradicting Bins $(4, 4)$ and $(3, 5)$, which indicate contradicting status of Subexon 4. Suppose all the other bins are non-contradicting and indicate the existence of Subexons 1, 2, 3, 5, 6, and 7. Then we decompose the isoform candidate into two candidates: 1111111 and 1110111, where the former contains all subexons and supports Bin $(4, 4)$ while the latter excludes Subexon 4 and supports Bin $(3, 5)$. This procedure is to reduce our chance of missing true isoforms.

2 K-means and gap statistic

2.1 Motivation

With objective function $\min_{W\geq 0, H\geq 0} ||V - WH||_F$ and additional orthogonality constraint on H, i.e., $H^T H = I$, NMF can be regarded as one type of K-means clustering on the bins (rows of V) with non-negativity constraint. The reason is that the purpose of NMF is to cluster bins into *bin groups*, which are sub-structures of isoforms and can form into multiple isoforms including the true ones. This motivated us to use the *gap statistic*, a method for choosing the number of cluster K in K-means clustering, to select the rank of NMF. Gap statistic was proposed by Tibshirani et al. [2] and has since been a widely used metric for choosing K in K-means clustering because of its good performance in estimating the number of well separated clusters.

2.2 K-means clustering

Suppose there are n p-dimensional data points, $X_1, X_2, \dots, X_n \in \mathbb{R}^p$ and the goal is to cluster them into K clusters $C_1, \dots, C_K$. Given K, K-means clustering would assign the n data points to K clusters, i.e., find the cluster memberships $C_1, \dots, C_K$ by minimizing the following objective function

$$\arg\min_{C_1,\dots,C_k} \sum_{r=1}^{K} \sum_{i \in C_r} ||X_i - \mu_r||, \tag{1}$$

where μ_r is the mean of cluster C_r, which is a subset of the n data points. Formula (1) is equivalent to

$$\arg\min_{C_1,\dots,C_k} \sum_{r=1}^{K} \frac{1}{2n_r} \sum_{i,j \in C_r} d_{ij} \tag{2}$$

n_r is the number of points in cluster C_r and d_{ij} is the distance between X_i and X_j, i.e. $||X_i - X_j||$. There are many choices for the distance metric, such as the Euclidean distance. The objective funciton $W_K = \sum_{r=1}^{K} \frac{1}{2n_r} \sum_{i,j \in C_r} d_{ij}$ is the within-cluster variance, which is a basic statistic for determining K.

2.3 Gap statistic

Gap statistic is defined as $Gap_n(k) = E_k^*[\log(W_k)] - \log(W_k)$. The first term is the expected W_k under a reference distribution with no clusters, and the second term is the observed W_k. The idea is to choose the number of clusters as the value of k that leads to the largest $Gap_n(k)$. To estimate $E_k^*[\log(W_k)]$, the simplest reference distribution is the uniform distribution in all the p dimensions over the range of the observed data. The gap statistic algorithm is sketched as follows.

1. Vary the number of clusters $k = 1, \dots, T$, and cluster the data $X_1, \dots, X_n$ by K-means clustering into k clusters, resulting in W_k, $k = 1, \dots, T$, where T is the upper bound on k.

2. Generate B reference data sets from the specified reference distribution (e.g. uniform distribution). Then we cluster each data set into k clusters, resulting in W_{kb}^*, $k = 1, \dots, T$; $b = 1, \dots, B$.

3. Let $\bar{w} = \frac{1}{B} \sum_{b=1}^{B} \log(W_{kb}^*)$, $sd_k = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\log(W_{kb}^*) - \bar{w})^2}$, $s_k = sd_k \sqrt{1 + \frac{1}{B}}$.

4. Estimate the gap statistic as $\hat{Gap}_n(k) = \bar{w} - log(W_k)$, for $k = 1, \dots, T$.
5. Choose the number of clusters as $\hat{K}$ = smallest k s.t. $\hat{Gap}_n(k) \geq \hat{Gap}_n(k+1) - s_{k+1}$.

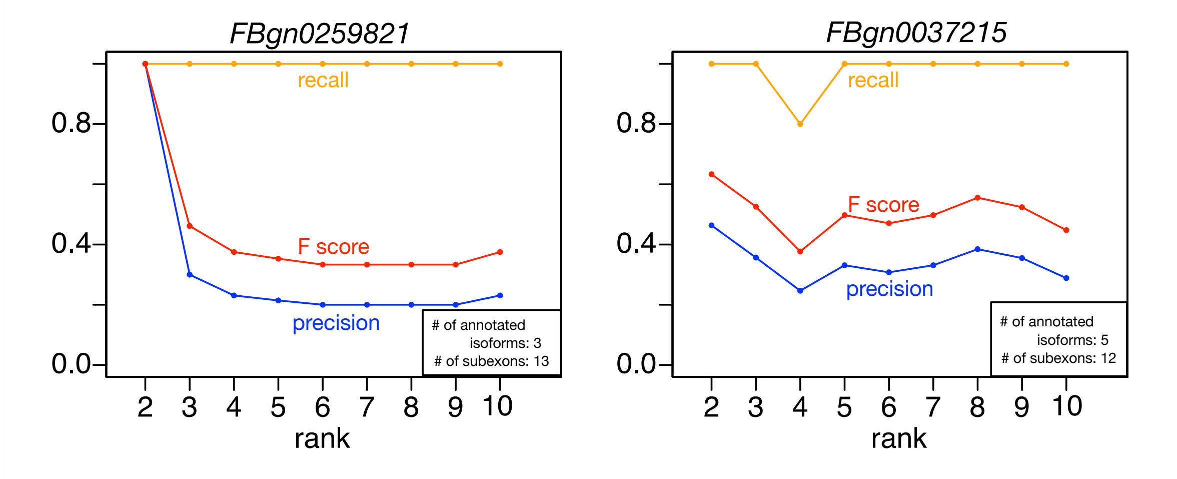
2.4 Application of gap statistic to NMF rank determination

NMF is a way of K-means clustering that clusters the bins with similar expression levels into the bin groups, i.e., splicing structures that can be reconstructed into isoforms. In most cases, the number of bin groups is close to the number of isoforms. For exmple, assume there is a 5-subexon gene with 3 isoform, 11111, 11011 and 11101. The relative abundance of the three isoforms are 50%, 35% and 15% respectively. Then the subexons have relative expression levels as 100%, 100%, 65%, 85% and 100% sequentially. Therefore, Subexons 1, 2 and 5 will be clustered into one bin group, while Subexon 3 and Subexon 4 will each be clustered as one bin group respectively. In this example, both the number of isoforms and the number of bin groups are 3. For genes with more complicated splicing structures, the number of bin groups may be more than the number of isoforms. In such cases, our estimated number of bin groups, $\hat{K}$ from gap statistic, could be larger than the number of true isoforms. However, from our simulation results, we observed that NMFP is not sensitive to the NMF rank choice and performs reasonably well as long as the rank is no less than the number of annotated isoforms. (See the section **Low sensitivity of NMFP to ranks** in the main text.) Combined with the fact that gap statistic tends to be conservative [2], the NMF rank should be better chosen as larger than $\hat{K}$, the number of clusters chosen by the gap statistic on V. In our results, we chose the NMF rank as $\hat{K} + 1$.

3 More results

3.1 Low sensitivity of NMFP to ranks (More results)

Continued from the main text, here we attach two more simulation examples to illustrate that NMFP is not sensitive to the choice of NMF rank. In **Figure 2(a)**, Gene *FBgn0259821* has three annotated isoforms (Ensemble BDGP6 of release 80) with 13 subexons. NMFP is able to capture all the annotated isoforms (recall rate = 1) regardless of the rank choices. The precision rate of NMFP is 1 when the rank equals 2. Although it decreases when the rank increases to 3 because higher ranks would lead to more isoform candidates, it becomes relatively stable after rank equals 4. In **Figure 2(b)**, Gene *FBgn0037215* has 5 annotated isoforms with 12 subexons. NMFP has stable performance across all the rank choices.



(a) Gene *FBgn0259821*

(b) Gene *FBgn0037215*

Figure 2: **The performance of NMFP in terms of the change of ranks** The orange line represents recall curve, the red line F score curve and the blue line represents precision curve.

3.2 Detailed information of genes on chromosome chr19 of *Mus musculus*

In the section **Simulation results in *Mus musculus*** in the main text, we did another simulation to demonstrate the performance of NMFP on mouse transcriptome. Apart from what has been already stated in the main text, some supplementary detail (**Table 1**) is provided here about the genes we selected to work on from chromosome chr19 of *Mus musculus* (reference genome mm10 and annotation GRCm38 of release 81).

3.3 Sensitivity of SLIDE and NMFP+SLIDE to the parameter λ (More results)

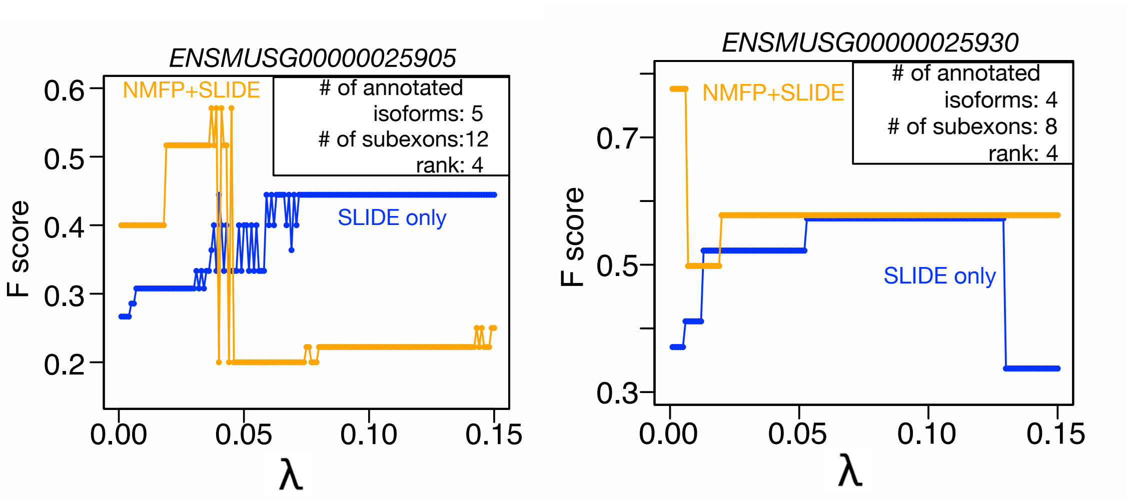
Continued from the main text, here we include two more simulation results to show that NMFP can help SLIDE achieve better isoform discovery accuracy at lower values of λ, the regularization parameter used in the LASSO step in SLIDE. Hence, the choice of a proper value for λ becomes an easier task for SLIDE+NMFP than for SLIDE. In **Figure 3 (a)**, Gene *ENSMUSG00000025905* has 5 annotated isoforms with 12 subexons. The rank is set as 4. NMFP+SLIDE has much higher F scores than SLIDE for $\lambda < 0.04$. In **Figure 3 (b)**, Gene *ENSMUSG00000025930* has 4 annotated isoforms with 8 subexons. The NMF rank is set as 4. We also observe that NMFP+SLIDE has better performance than SLIDE especially when $\lambda < 0.015$. Since NMFP can largely reduce the isoform candidate pool for SLIDE, it is recommended to use a small λ value for NMFP+SLIDE.

3.4 Real data case study (More results)

Continued from the main text, we use another two cases to show that NMFP has good performance on real data. In **Figure 4**, RNA-seq reads for gene *FBgn0019936* were generated by the modENCODE consortium [3] from the heads of mated female *D.melanogaste* after 1 day of eclosion (SRA accession: SRR070434, SRR070435 andSRR100279; see the Supplemental Material "Updated Table S2.xlsx" in [4] for more detail). This gene has 1 annotated isoform (shown in orange), which is well supported by the RNA-seq reads (shown in gray). Cufflinks alone connected the latter three exons together with the introns in between into one piece (shown in light blue). NMFP+Cufflinks accurately assembled the annotated isoform and recovered another isoform (shown in dark blue), which reflects the low read counts of the Exon 3. Similarly, NMFP+SLIDE at $\lambda = 0.2$ ("more", shown in dark green) and $\lambda = 0.01$ ("fewer", shown in dark red) achieved better isoform discovery results than their SLIDE counterparts (shown in light green and light red). In **Figure 5**, RNA-seq reads for gene *FBgn0038145* were also generated by the modENCODE corsortium from *D. melanogaster* L3 stage larvae and 12 hours post-molt (SRA accession: SRS004682; see the Supplemental Material "Updated Table S2.xlsx" in [4] for more detail). *FBgn0038145* has a complicated splicing structure and 5 annotated isoforms (shown in orange). Cufflinks alone assembled one transcript (shown in light blue) similar to the first annotated one except that part of Exon 1 is missed. NMFP+Cufflinks identified 4 isoforms (shown in dark blue) among which 2 are annotated. NMFP also improved the performance of SLIDE at both $\lambda = 0.2$ ("more", shown in light and dark green) and $\lambda = 0.01$ ("fewer", shown in light and dark red). One significant contribution of NMFP to Cufflinks and SLIDE is capturing Exon 2, which is missed by Cufflinks and SLIDE alone because of its low read coverage compared to the other exons.

Table 1: **Summary of the genes used in the section "Simulation results in *Mus musculus*" in the main text.** The table lists the numbers of the genes that have 3-30 subexons and 2-17 annotated isoforms.

# of subexons n	$3 \leq n \leq 6$	$7 \leq n \leq 10$	$11 \leq n \leq 14$	$15 \leq n \leq 18$	$19 \leq n \leq 22$	$n \geq 23$
	155	185	163	134	89	126
# of isoforms q	$2 \leq q \leq 3$	$4 \leq q \leq 5$	$6 \leq q \leq 7$	$8 \leq q \leq 9$	$10 \leq q \leq 11$	$q \geq 12$
	382	206	133	81	22	28



(a) Gene *ENSMUSG00000025905*

(b) Gene *ENSMUSG00000025930*

Figure 3: **The performance of SLIDE with NMFP and SLIDE alone at various λ values.** The orange line represents the F scores of NMFP+SLIDE, while the blue line represents the F scores of SLIDE alone.

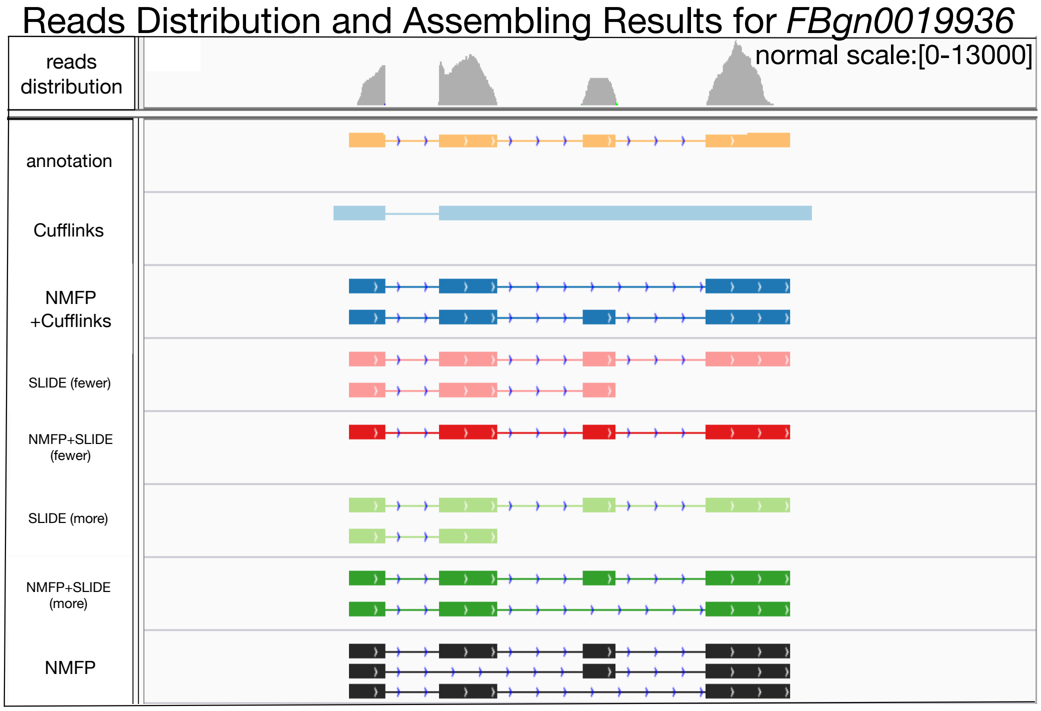


Figure 4: **Real data results for Gene *FBgn0019936***

References

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

[2] Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(2), 411–423 (2001)

[3] Gerstein, M.B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J., *et al.*: Comparative analysis of the transcriptome across distant species. Nature **512**(7515), 445–448 (2014)

[4] Li, J.J., Huang, H., Bickel, P.J., Brenner, S.E.: Comparison of d. melanogaster and c. elegans developmental stages, tissues, and cells by modencode rna-seq data. Genome research **24**(7), 1086–1101 (2014)

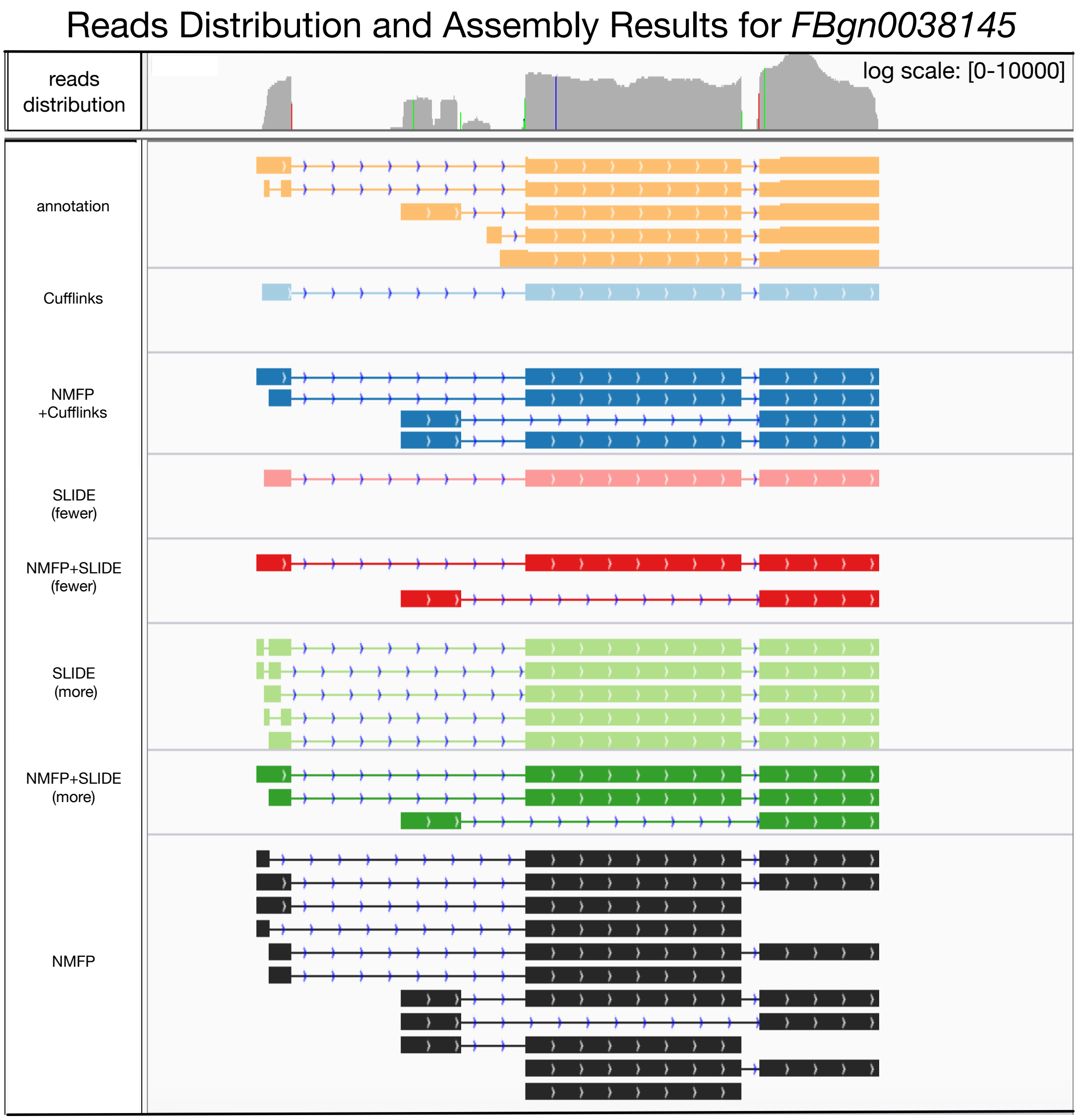


Figure 5: **Real data results for Gene *FBgn0038145***