NMFP: a non-negative matrix factorization based preselection method to increase accuracy of identifying mRNA isoforms from RNA-seq data

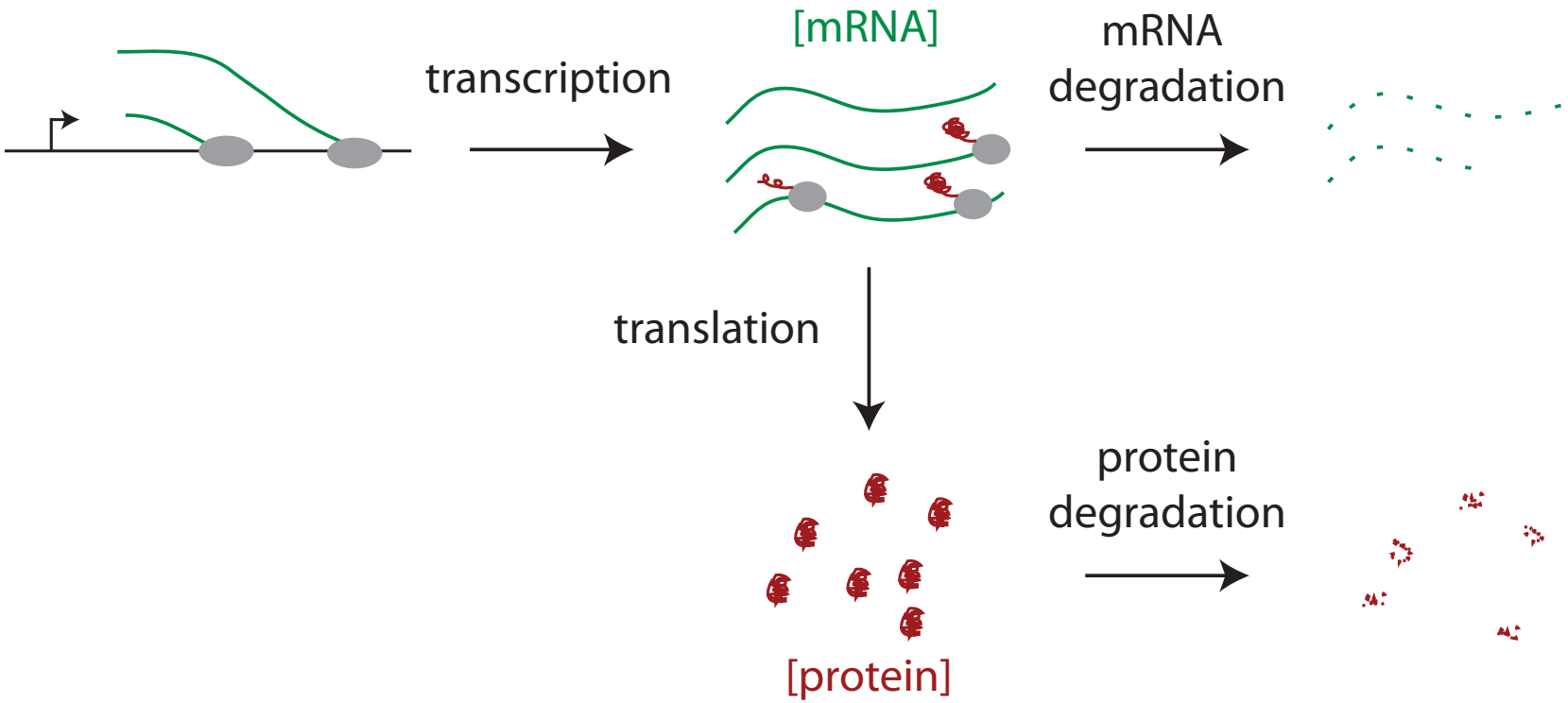
Yuting Ye; Jingyi Jessica Li

Division of Biostatistics, University of California at Berkeley;
Department of Statistics, University of California at Los Angeles

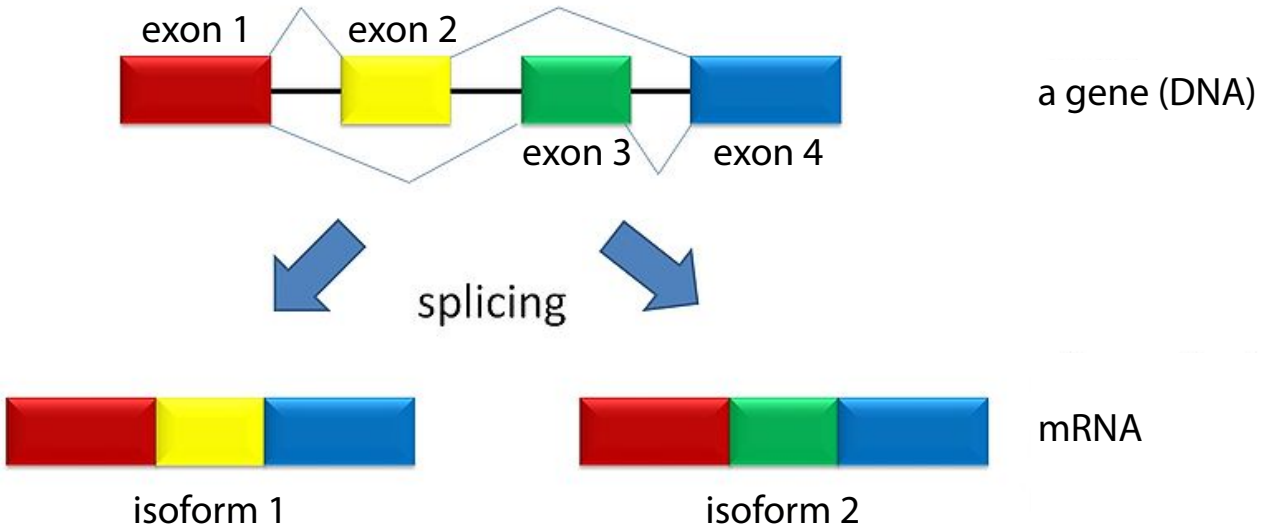
July 20, 2016

Central Dogma

The Central Dogma is the foundation of molecular biology.



Alternative Splicing



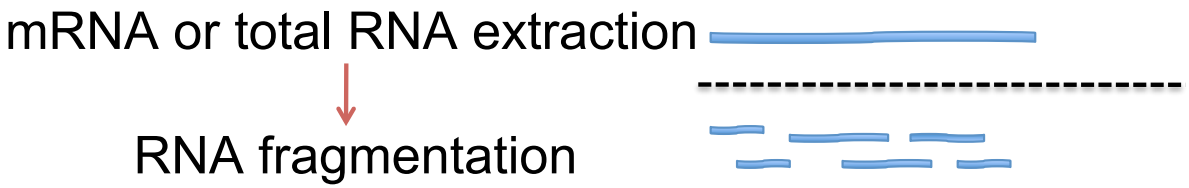
$$4 \text{ exons} \Rightarrow 2^4 - 1 \text{ possible isoforms.}$$

- Alternative splicing contributes to the diversity of mRNA isoforms and hence proteins.
- Aberrant structures or abundance of mRNA isoforms can cause various human diseases.

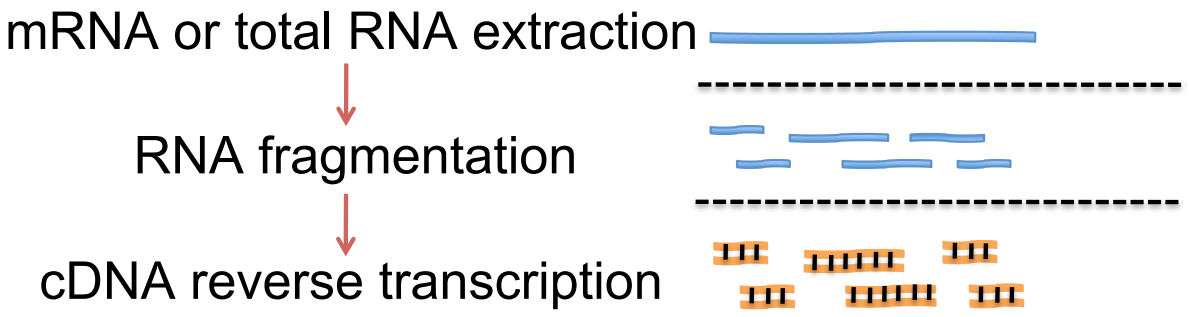
RNA-seq Reads Generation Mechanism

mRNA or total RNA extraction

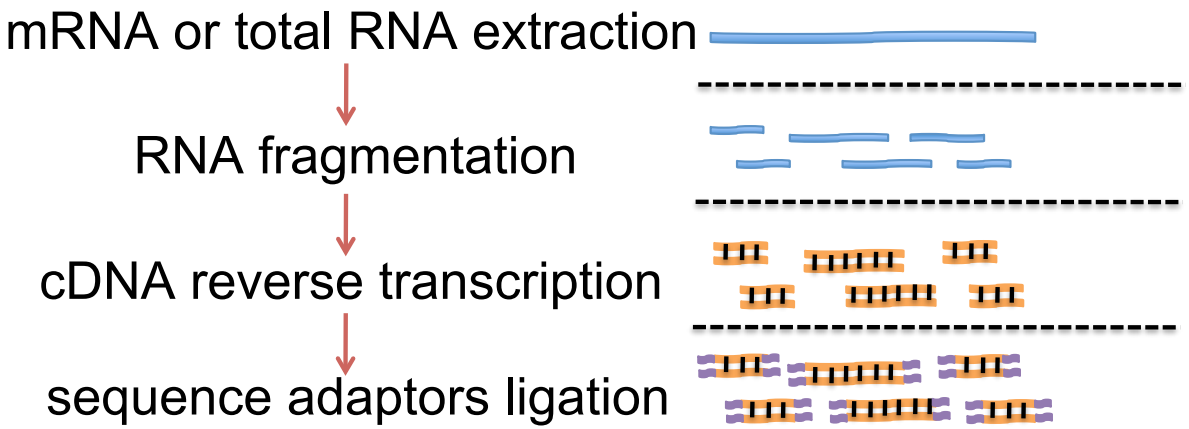
RNA-seq Reads Generation Mechanism



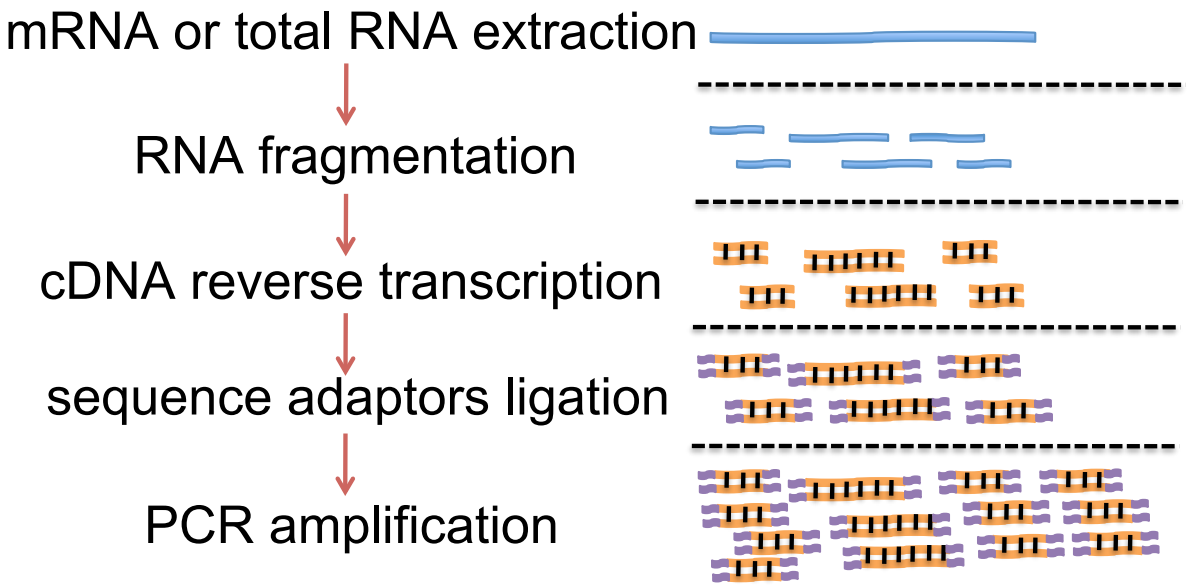
RNA-seq Reads Generation Mechanism



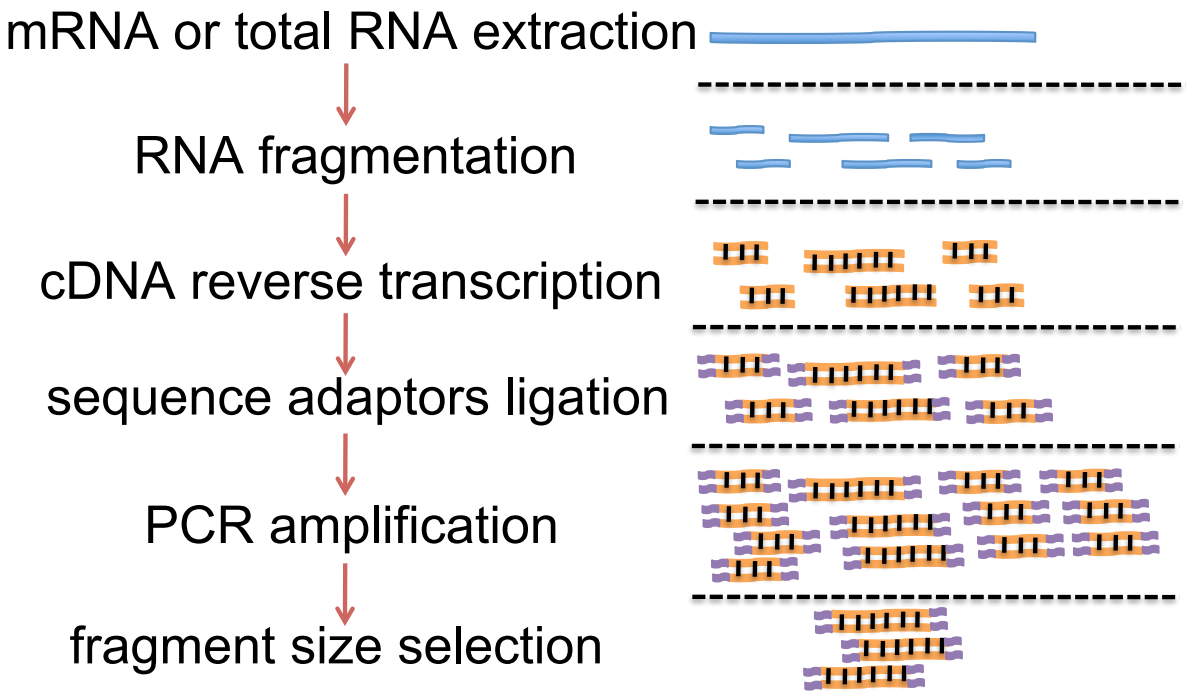
RNA-seq Reads Generation Mechanism



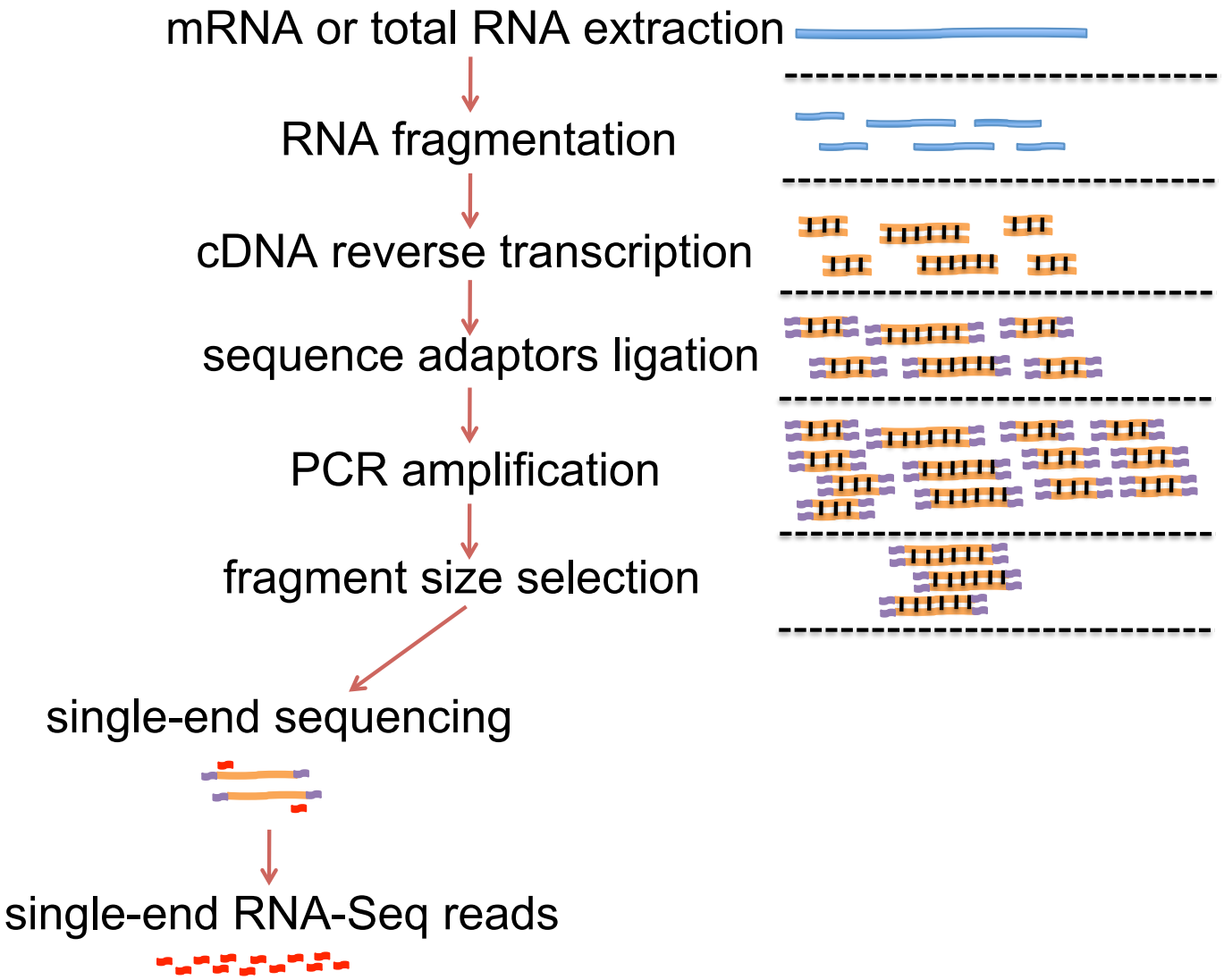
RNA-seq Reads Generation Mechanism



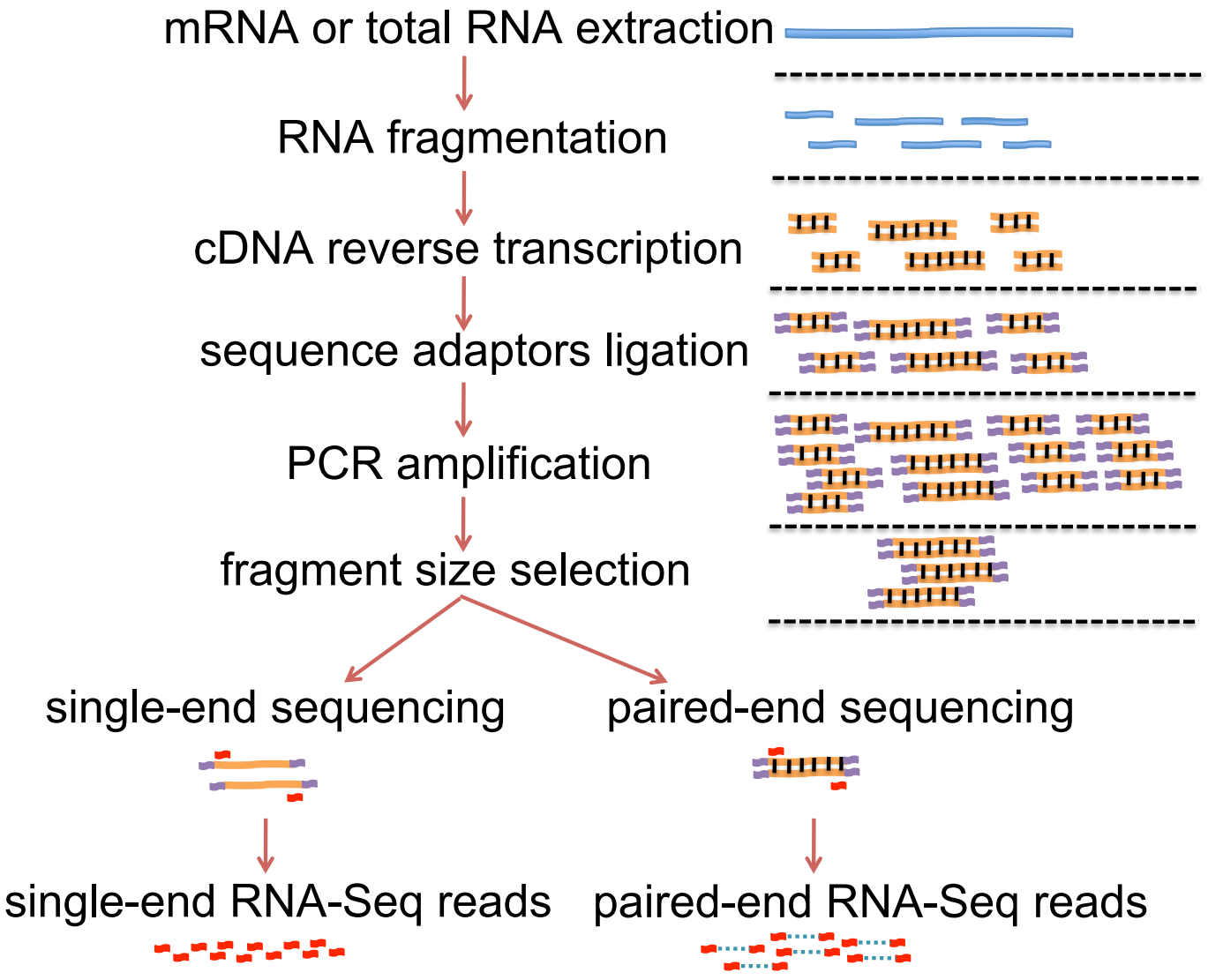
RNA-seq Reads Generation Mechanism



RNA-seq Reads Generation Mechanism

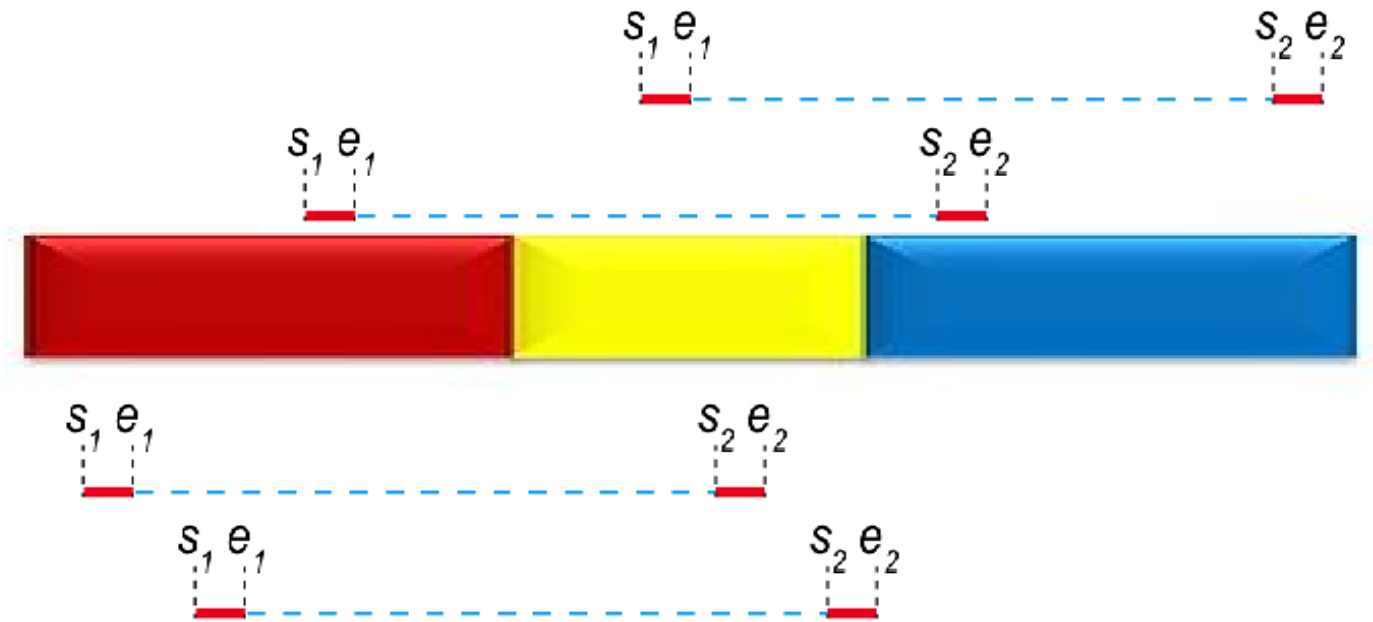


RNA-seq Reads Generation Mechanism



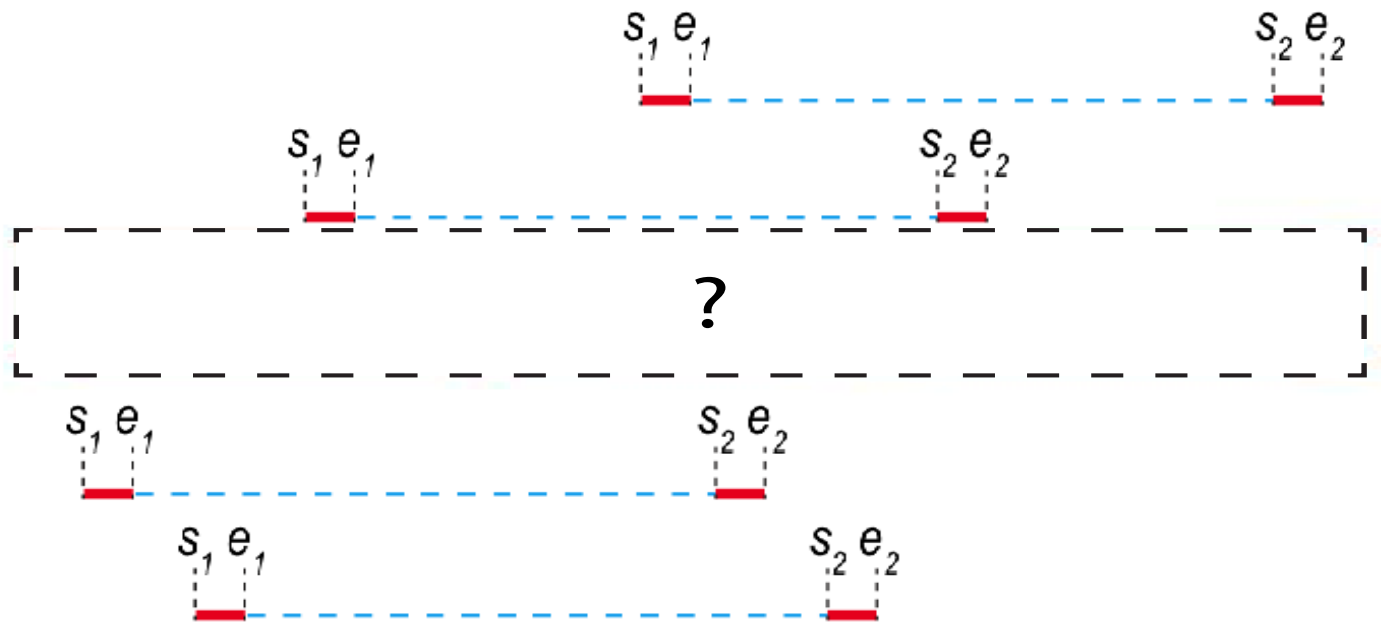
Isoform Discovery

- **Goal**: to discern the mRNA isoforms mostly likely to exist in a sample.



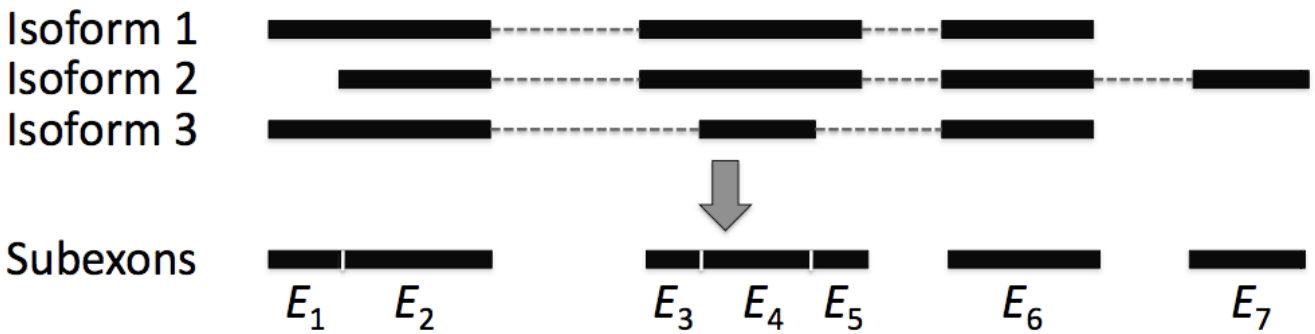
Isoform Discovery

- **Goal**: to discern the mRNA isoforms mostly likely to exist in a sample.



Subexons & Possible Isoforms

A **subexon** is defined as a transcribed region between adjacent splicing sites in any annotated mRNA isoforms.



For a gene with n subexons, there are a total of $2^n - 1$ possible isoforms.

Bins

Key information in paired-end RNA-Seq data:

We categorize paired-end reads into paired-end **bins**: Bin (i, j, k, l) contains reads whose s_1, e_1, s_2 and e_2 are in Subexons i, j, k and l respectively.

Bin count vector:

$$\mathbf{B} = (B_{(1,1,1,1)}, ..., B_{(n,n,n,n)})^T,$$

where $B_{(a,b,c,d)} = \#$ of reads mapped into Bin (a, b, c, d).

Isoform discovery methods for Microarray data

A model for probe intensity and isoform abundance (Li and Wong, PNAS 2001)

$$v = a \cdot t + e,$$

where

- v: probe intensity
- a: probe affinity
- t: isoform abundance
- e: error term

if the probe is in the isoform.

Isoform discovery methods for Microarray data

SPACE (Anton et al, Genome Biol. 2008)

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{T} + \boldsymbol{\epsilon},$$

where

- $\mathbf{V}$: a $p \times m$ matrix representing the intensities of p probes in m samples
- $\mathbf{A}$: a $p \times p$ diagonal matrix representing the affinity of p probes
- $\mathbf{G}$: a $p \times s$ indicator matrix with binary elements indicating the existence of the p probes in s isoforms
- $\mathbf{T}$: an $s \times m$ matrix representing the abundance of s isoforms in m samples
- $\boldsymbol{\epsilon}$: a $p \times m$ error matrix

Isoform discovery methods for Microarray data

SPACE (Anton et al, Genome Biol. 2008)

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{T} + \epsilon.$$

- SPACE uses non-negative matrix factorization (NMF) to decompose $\mathbf{V}$ into $\mathbf{W} = \mathbf{A} \cdot \mathbf{G}$ and $\mathbf{T}$.
- NMF has the advantage of estimating $\mathbf{W}$ and $\mathbf{T}$ as sparse and non-negative matrices, which have interpretable biological meanings.
- Given the NMF estimate $\hat{\mathbf{W}} = (\hat{W}_{ik})$, SPACE estimates $\mathbf{A}$ as

$$\hat{\mathbf{A}} = diag(\hat{A}_{11}, \cdots, \hat{A}_{pp}), \text{ where } \hat{A}_{ii} = \max_{k=1,\cdots,s} (\hat{W}_{ik}).$$

- With $\hat{\mathbf{A}}$ and $\hat{\mathbf{W}}$, SPACE estimates $\mathbf{G}$ as

$$\hat{\mathbf{G}} = (\hat{G}_{ij}), \text{ with } \hat{G}_{ij} = \mathbb{I}\left((\hat{\mathbf{A}}^1 \cdot \hat{\mathbf{W}})_{ij} > c\right).$$

Isoform discovery methods for Microarray data

SPACE (Anton et al, Genome Biol. 2008)

$$\mathbf{V} = \mathbf{A} \cdot \mathbf{G} \cdot \mathbf{T} + \epsilon.$$

Issues:

- How to determine the rank of NMF, i.e., $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{T})$ remains an open question.
- NMF results are often not unique.
- Isoform structures learned by NMF can be biologically invalid, i.e., some probes cannot coexist in one isoform.

Microarray $\rightarrow$ RNA-seq?

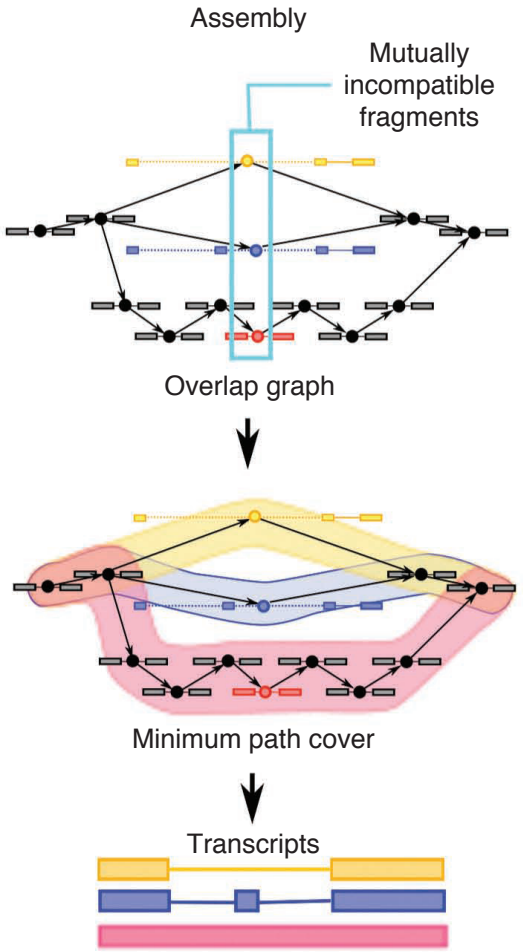
Isoform discovery methods for RNA-seq data

Cufflinks (Trapnell et al, Nat Biotechnol. 2010)

- Cufflinks constructs a De Brujin graph to assemble RNA-seq reads into mRNA isoforms, which are identified as maximal paths in the graph.
- Cufflinks is annotation free and robust to RNA-seq data noise.

Issues:

- Cufflinks cannot find overlapping isoforms.
- Complex gene structures lead to large numbers of nodes and edges, making Cufflinks difficult to find isoforms as



Isoform discovery methods for RNA-seq data

SLIDE (Li et al, PNAS 2011)

- SLIDE constructs a search space of all possible isoforms and then identifies isoforms based on a regularized linear model.

$$b_i = \sum_{j=1}^{2^n-1} F_{ij} t_j + \epsilon_i, \text{ for } i = 1, \cdots, \#\text{ bins}.$$

  - $b_i = B_i/||\mathbf{B}||$: proportion of reads in the i-th bin "observed"
  - $F_{ij} = \Pr($ the read in the i-th bin $\mid j$-th isoform $)$ "needed"
  - $t_j = P($a read is from the j-th isoform$)$ "parameter of interest"
  - ϵ_j: error term with mean 0

- SLIDE estimates $\mathbf{t}$ as

$$\hat{\mathbf{t}} = \arg\min_{\mathbf{t}} \sum_{i=1}^{\#\text{ bins}} (b_i - \mathbf{F}_i \mathbf{t})^2 + \lambda \sum_{j=1}^{2^n-1} \frac{|t_j|}{n_j},$$

Isoform discovery methods for RNA-seq data

SLIDE (Li et al, PNAS 2011)

- SLIDE uses annotation information and can find overlapping isoforms.

Issues:

- Compared to Cufflinks, SLIDE is sensitive to RNA-seq data noise.
- A gene with n exons will give rise to a total of $2^n - 1$ possible isoforms, placing great difficulty on the regularized linear model to find the correct isoforms.

Preselection based on NMF

Large search space is a big issue for all methods. Let's first see in the field of statistics, what to do if it comes to high dimension, or $p >> n$?

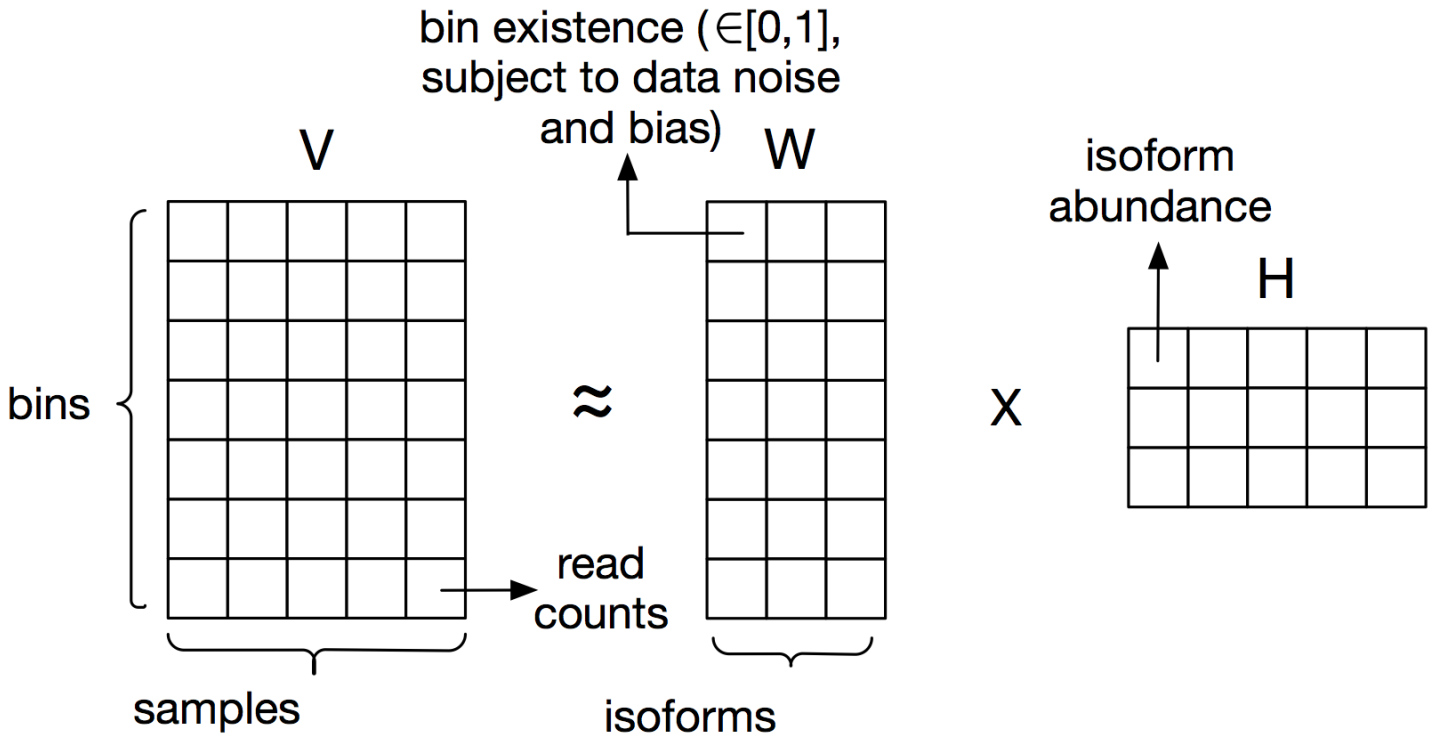
- The main stream is to preselect features before applying any models.
- Fan and Lv in 2007 proposed Sure Independence Screening (SIS) for preselection. It just chooses features with high correlations with the response. Very simple, but very effective.
- Motivated by such idea, we're going to develop a preselection method which can provide a small pool of candidates that other methods can make use of.

Cont: Preselection based on NMF

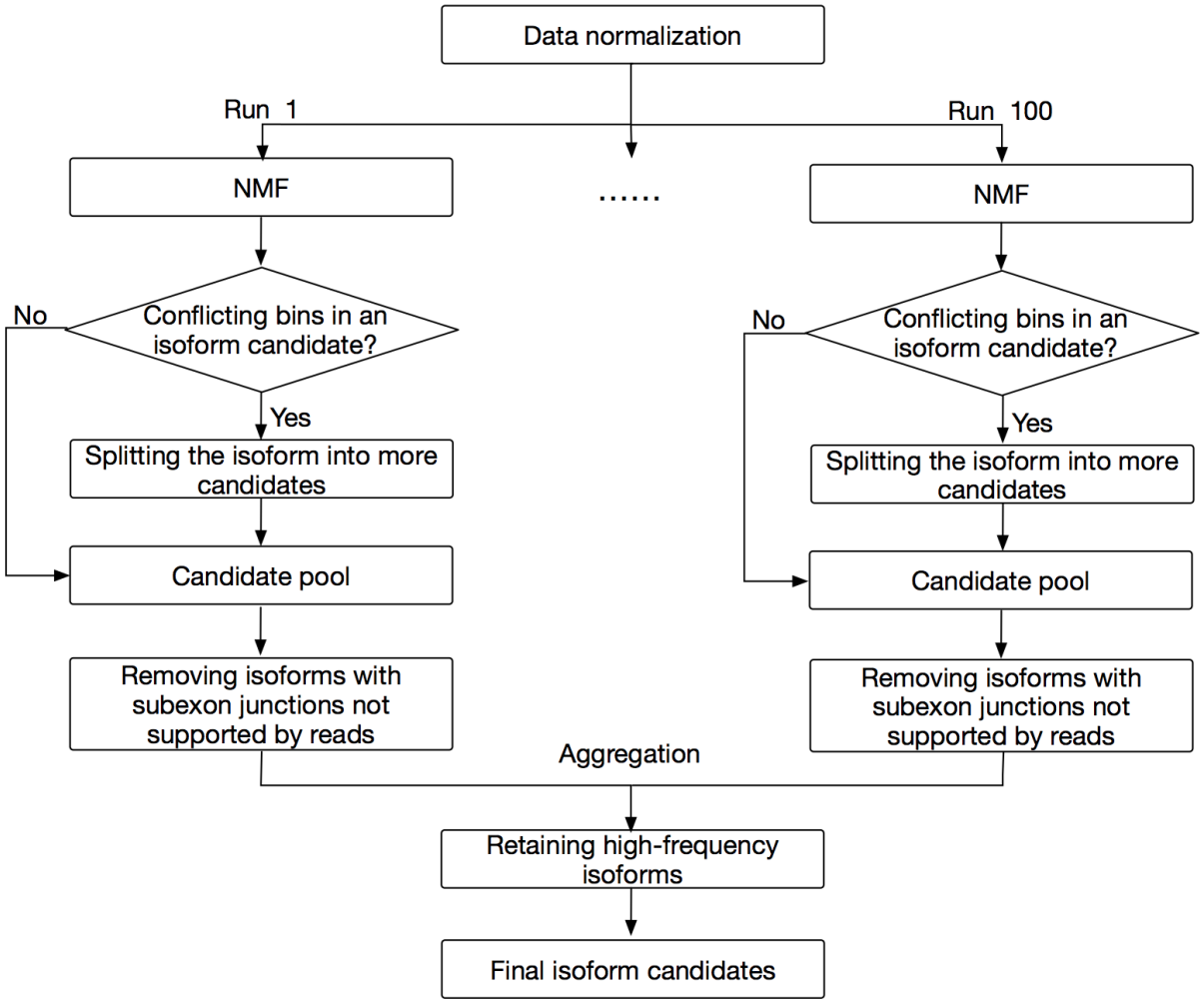
- Non-negative matrix factorization (NMF) embraces good properties including the interpretability and sparsity of decomposition results.
- However, how to determine the NMF rank and the problem of non-unique factorizations make NMF not directly applicable for isoform discovery.

We propose to use NMF to shrink the search space by aggregating mRNA isoforms found by NMF over multiple runs.

Illustration of NMF



Flowchart of NMFP



Ambiguous Isoforms

1. Some factorization isoforms may be ambiguous or biologically invalid. For example, an candidate isoform with both Bin $(1, 3, 4, 4)$ and Bin $(2, 2, 4, 4)$ can't exist since Bin $(1, 3, 4, 4)$ implies the splicing of Subexon 2 while Bin $(2, 2, 4, 4)$ indicates the existence of Subexon 2.
2. To avoid ambiguous candidate isoforms as much as possible, a new object function is proposed for NMF:

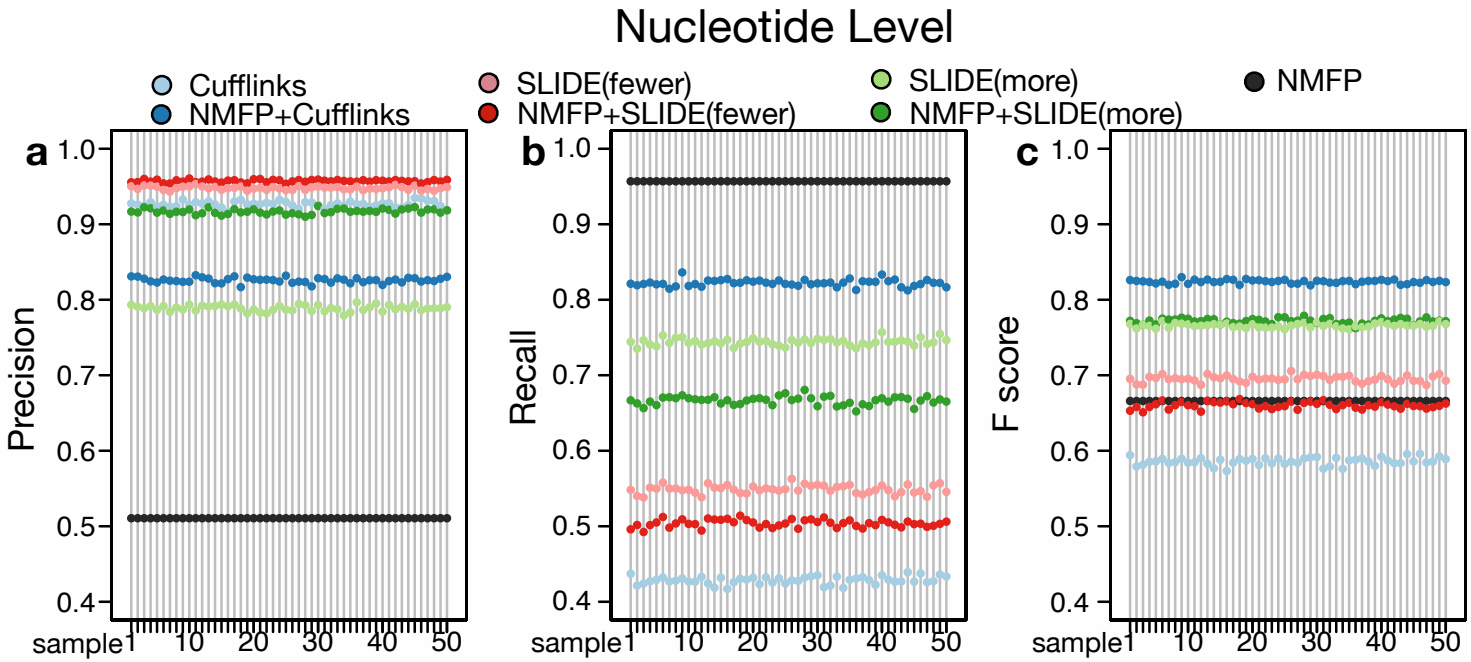
$$D(\mathbf{V}, \mathbf{WH}) = \sum_{\mathbf{i,j}} \left(\mathbf{V_{ij}} \log \frac{\mathbf{V_{ij}}}{(\mathbf{WH})_{\mathbf{ij}}} - \mathbf{V_{ij}} + (\mathbf{WH})_{\mathbf{ij}} \right) + \alpha \sum_{\mathbf{i} \perp \mathbf{j}} (\mathbf{WW^T})_{\mathbf{ij}}$$

3. Ambiguous candidate isoforms are split into biologically valid ones.

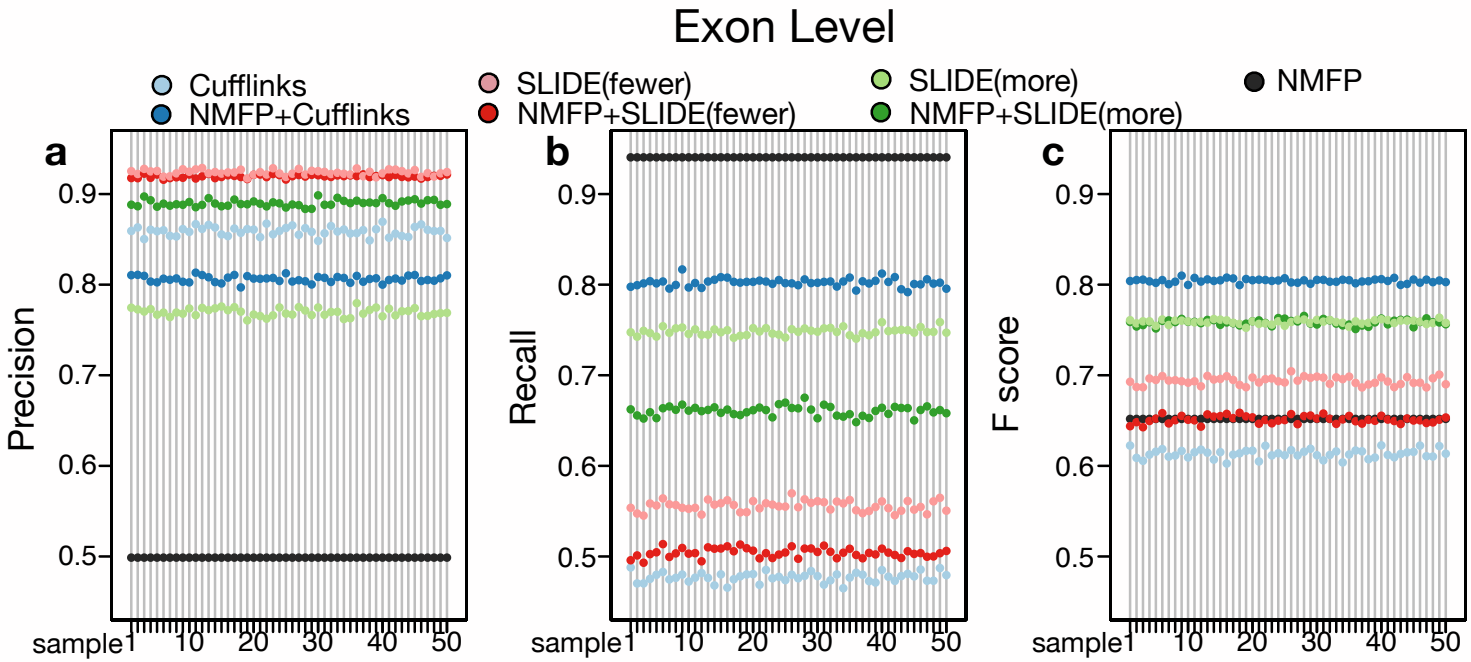
Simulation in D. melanogaster

- Flux Simulator is used to simulate the data from chr3R of *D.melanogaster* with reference genome dm6 and 363 Ensembl annotation BDGP6 of release.
- 50 samples are simulated, each with $10,000,000$ RNA molecules and $50,000,000$ paired-end reads with length 2×76 bp, from these genes isoforms in the annotation.
- The isoform abundance is randomly assigned by Flux Simulator.
- 51.7% (2132) genes contain 3 to 10 subexons. Among these genes, 44.6% (951) have more than one isoforms in the annotation. NMFP is applied to these 951 genes.

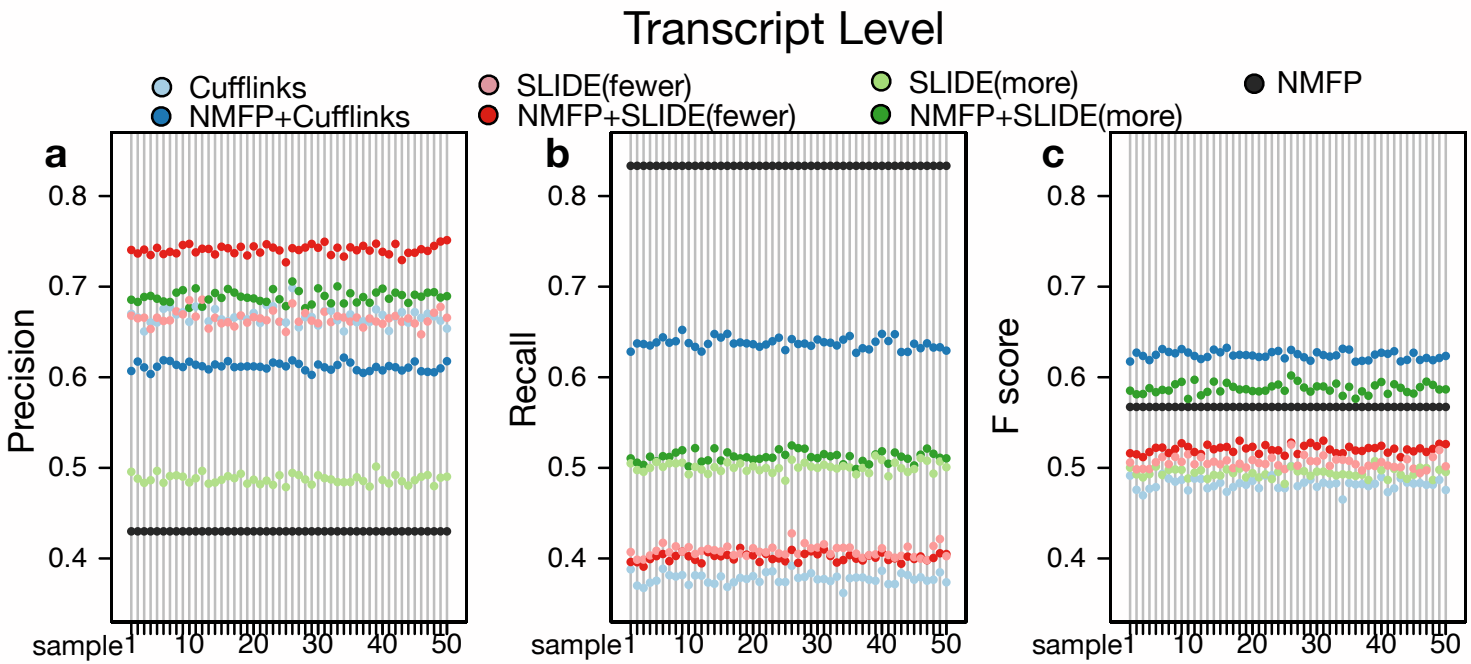
Results at Nucleotide Level



Results at Exon Level

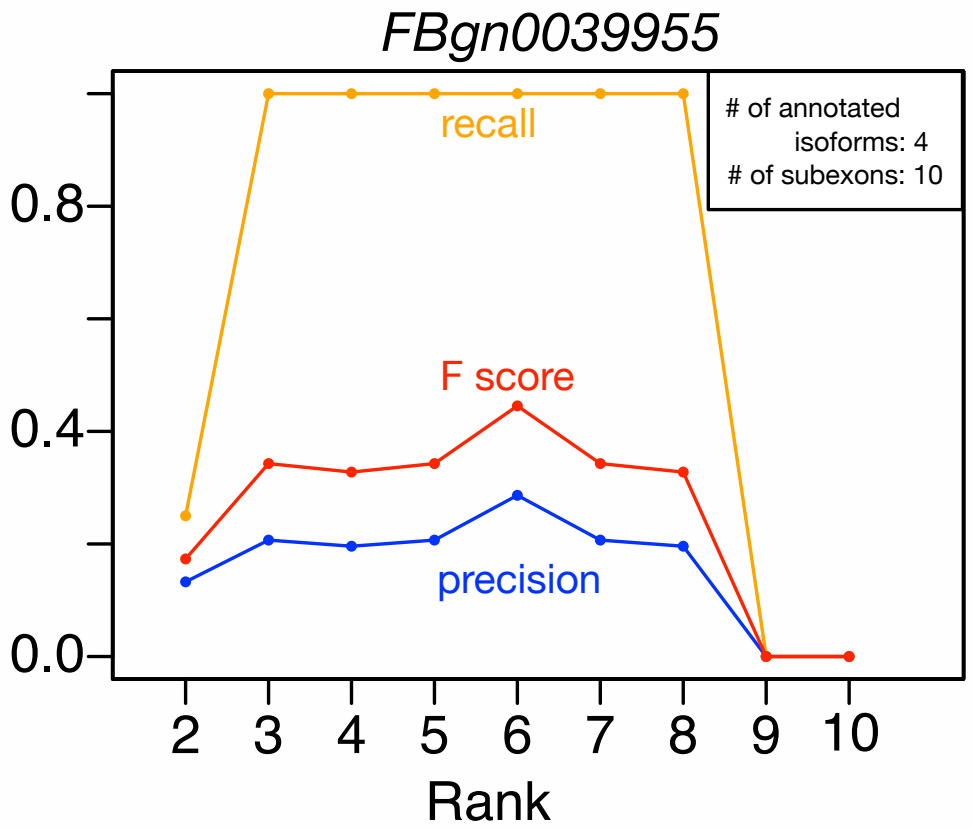


Results at Transcript Level



Robust to the choices of NMF rank

Gene FBgn0039955 is selected to display that NMFP is robust to the choices of NMF rank.

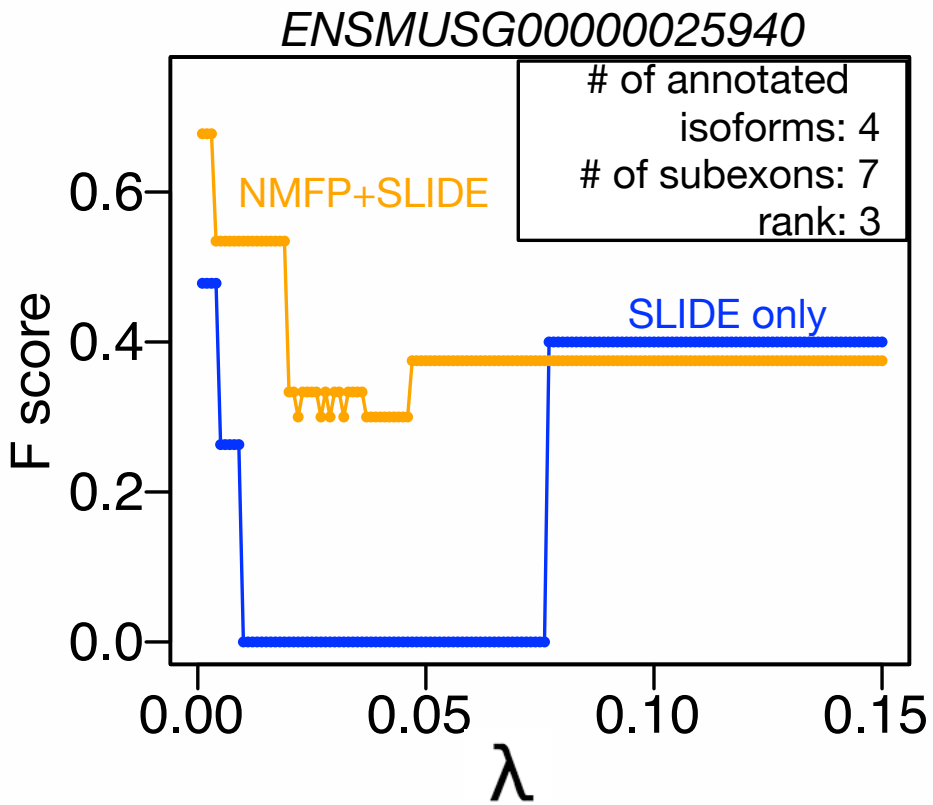


Simulation in M. musculus

- Flux Simulator is used to simulate the data from chr1 of *M.musculus* with reference genome mm10 and annotation GRCm38 of release 81.
- 100 RNA-seq samples are simulated with paired-end reads of length 2×76 bp.
  - The 100 samples have 10 different numbers of RNA molecules and 10 different numbers of reads. The numbers of RNA molecules range from $4,200,000$ to $6,000,000$, increasing in steps of $200,000$. The number of reads range from $11,000,000$ to $20,000,000$, increasing in steps of $1,000,000$.
- On chr1, there are 3432 genes, among which 852 genes are of interest with $3-10$ subexons and $2-17$ annotated isoforms.

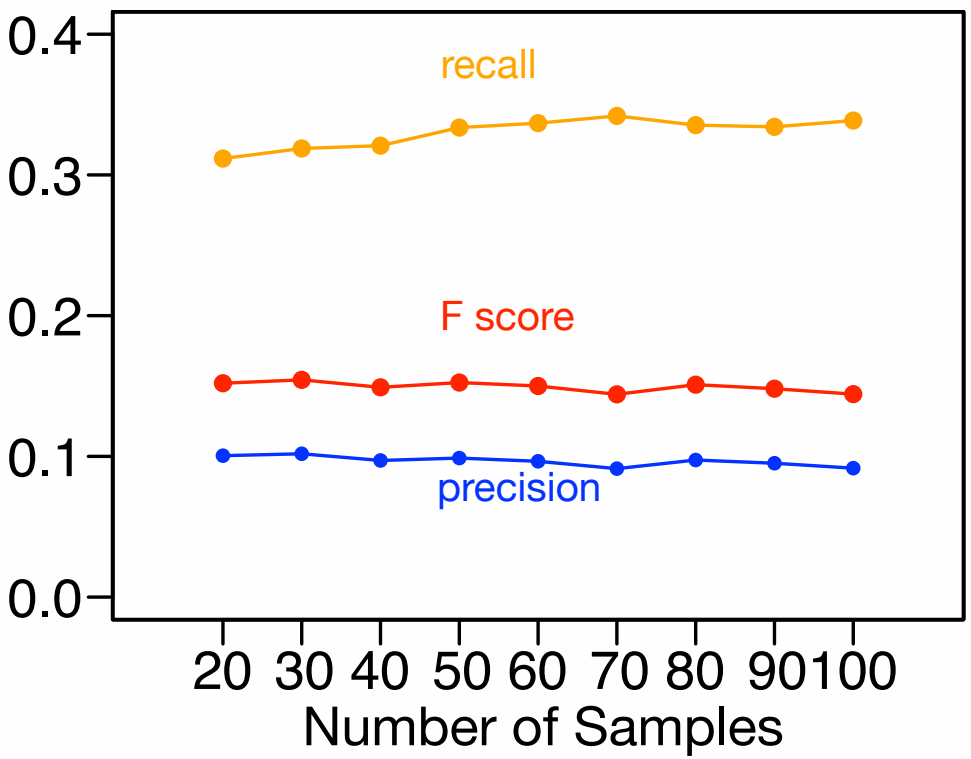
Robustness of NMFP+SLIDE to the choices of λ

Gene ENSMUSG00000025940 is selected to display that SLIDE combined with NMFP is robust to the choices of λ, the coeffiecient of the L_1 norm penalty.



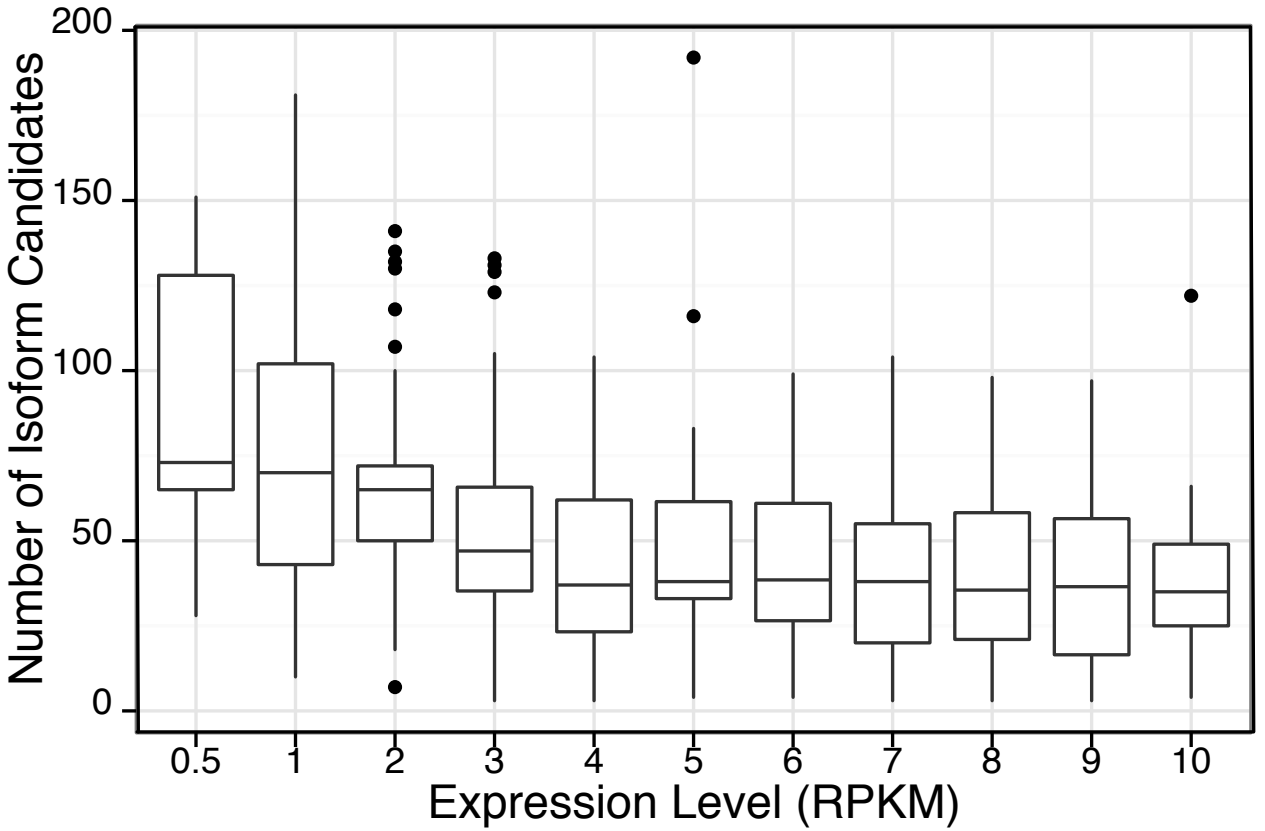
Robustness of NMFP to the number of input RNA-seq samples

NMFP is applied to different numbers of simulated mouse samples, from 20 to 100 samples.



Improvement on lowly expressed sample by NMFP

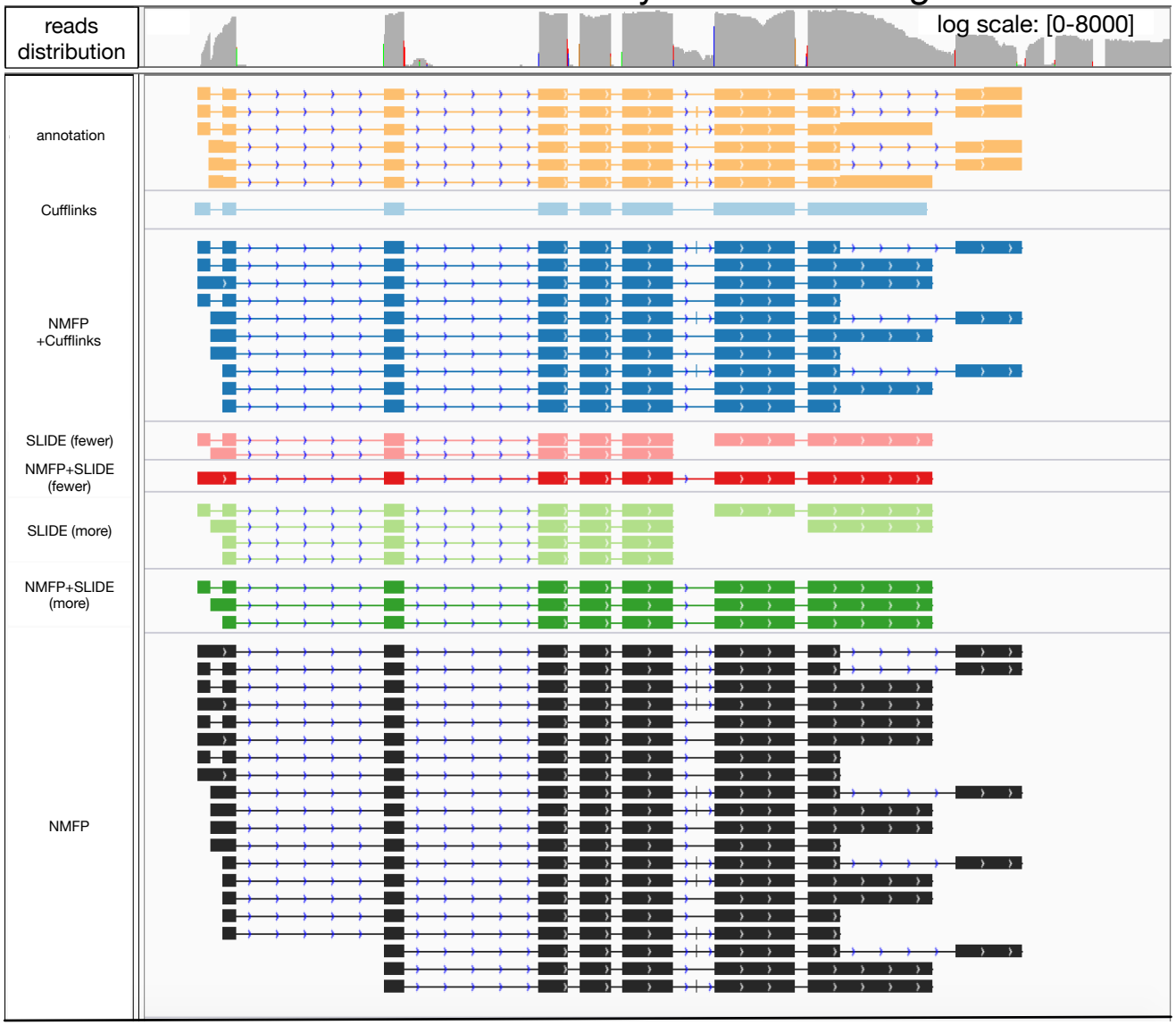
NMFP is capable of increasing the isoform discovery accuracy for a lowly expressed gene in a sample by leveraging other samples.



Real Data

- NMFP is further applied on 74 real RNA-seq data sets of *D. melanogaster* (reference genome dm5 and annota- tion BDGP5 of release 66). The results are interpreted by studying a few genes in Integrative Genomics Viewer (IGV).
- For *D. melanogaster* gene *FBgn0037643*, there are 11 subexons and 6 annotated isoforms.

Reads Distribution and Assembly Results for *FBgn0037643*



Conclusions

- NMFP can effectively shrink the isoform search space to improve the performance downstream isoform discovery methods.
- Two remaining issues with NMFP
  - Parallelization is important to increase the computational efficiency of NMFP.
  - Proper normalization is necessary for aggregating data from different sources.
- The NMFP source codes and examples are available at `http://www.stat.ucla.edu/~jingyi.li/packages/NMFP.zip`.

Contact Info: yeyt@berkeley.edu

Thank you!