UCLA

Statistical Methods for Bulk and Single-cell RNA Sequencing Data

Jingyi Jessica Li

Department of Statistics
University of California, Los Angeles

http://jsb.ucla.edu

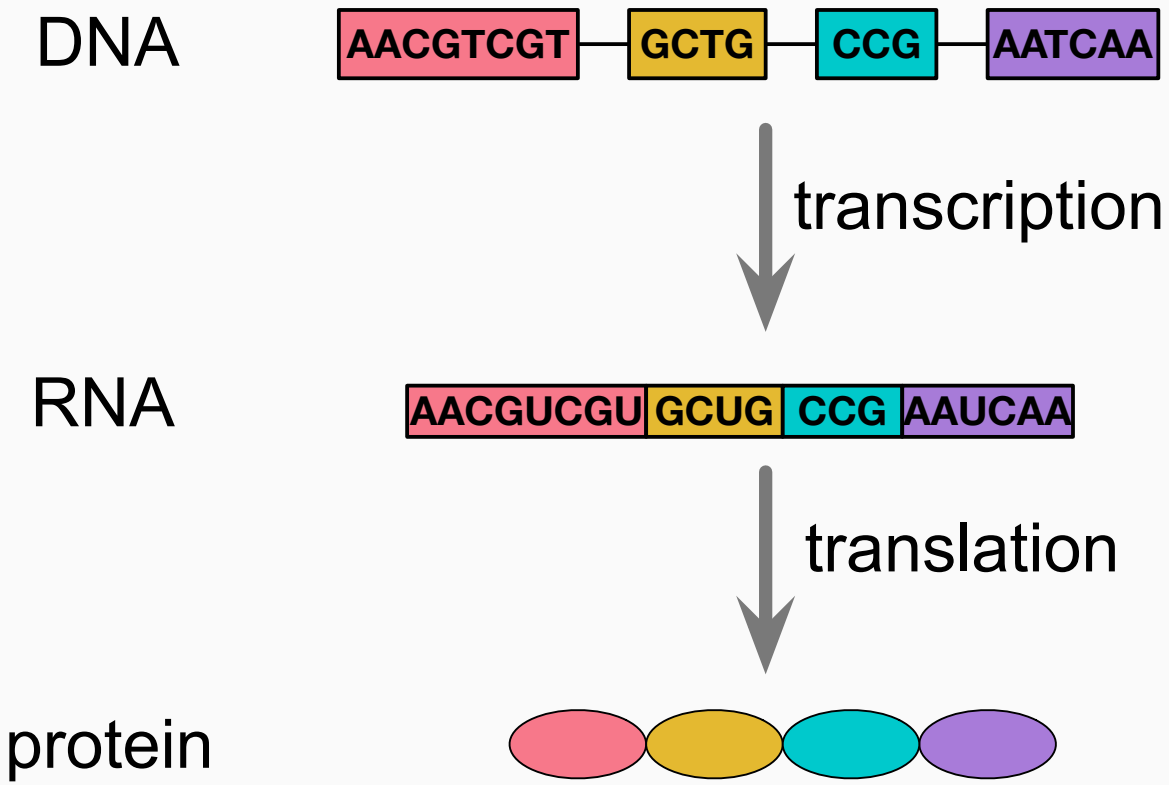
The central dogma of molecular biology

2018 marks the 60th anniversary of the central dogma:
DNA makes RNA makes proteins.

Francis Crick speaking at the 1963 CSH Symposium [Cobb, *PLoS Biology*, 2017]

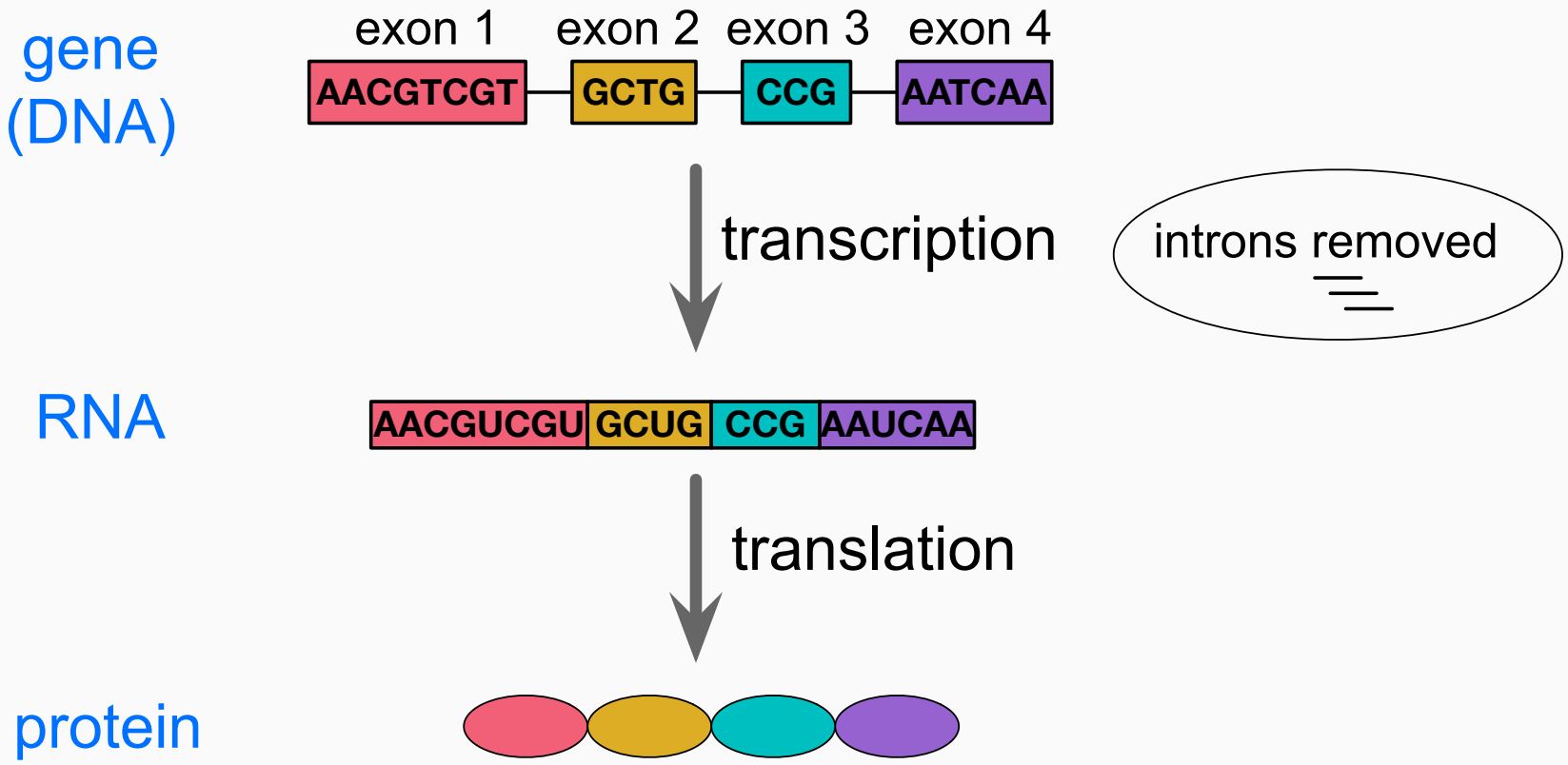
The central dogma of molecular biology

The central dogma of molecular biology:
DNA makes RNA makes proteins.

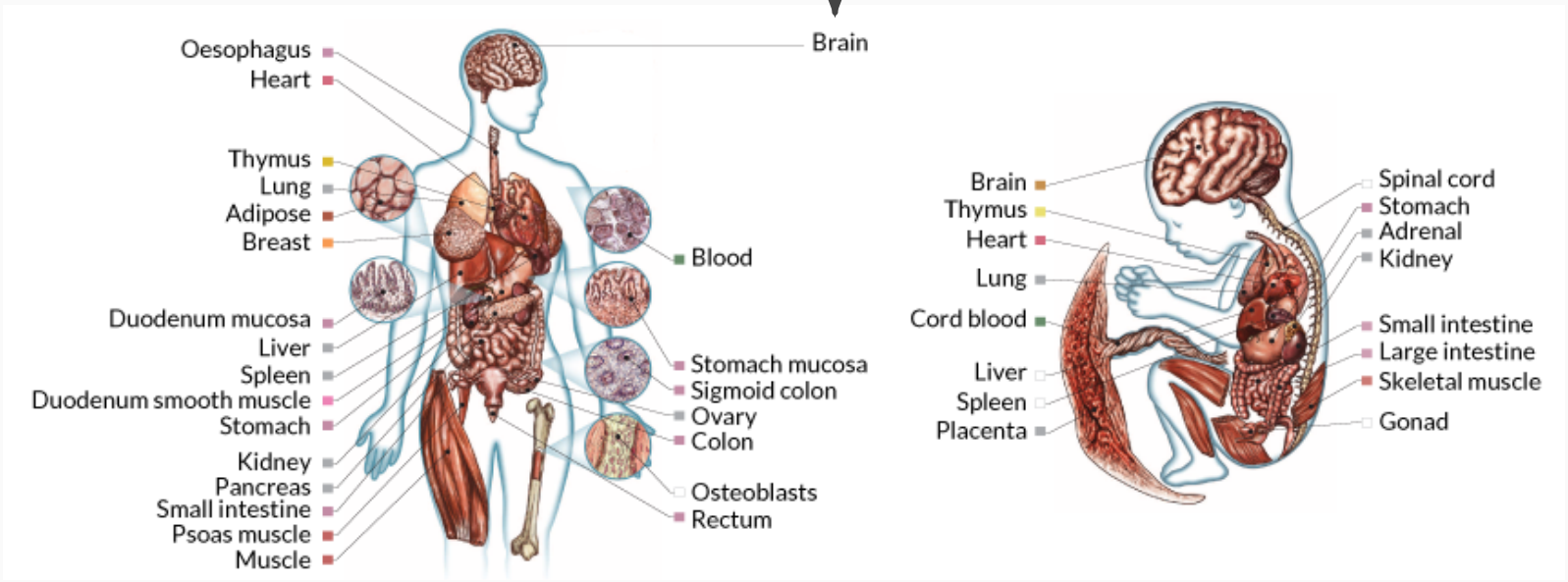


The central dogma of molecular biology

In transcription, a particular segment of DNA (combinations of exons) is copied into RNA segments.



Understanding genome functions



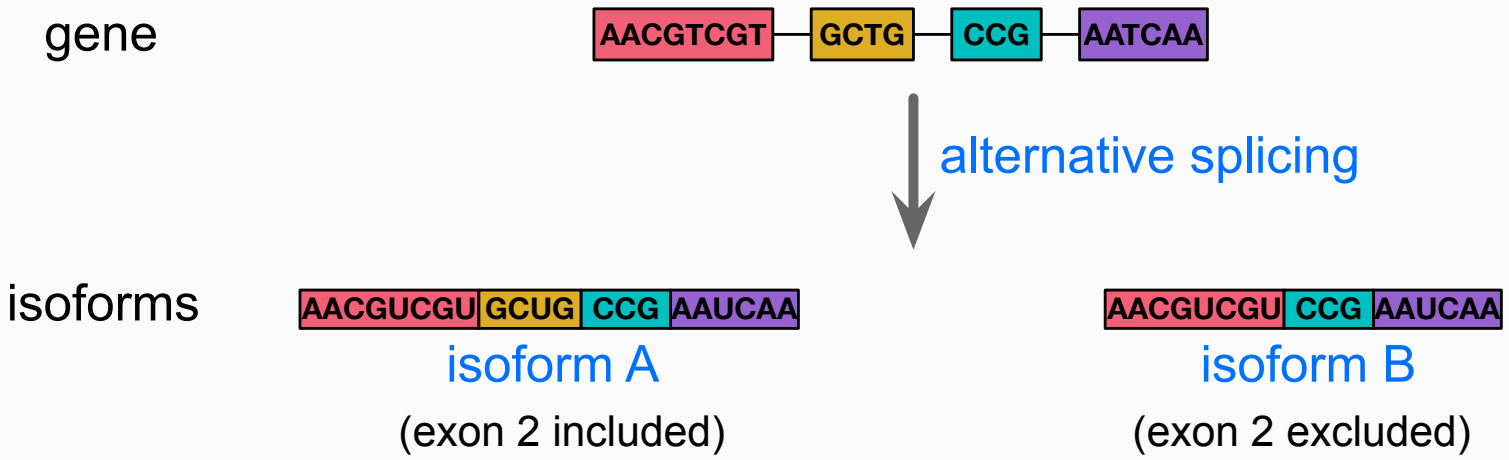
[Kundaje et al., *Nature*, 2015]

Understanding genome functions

?

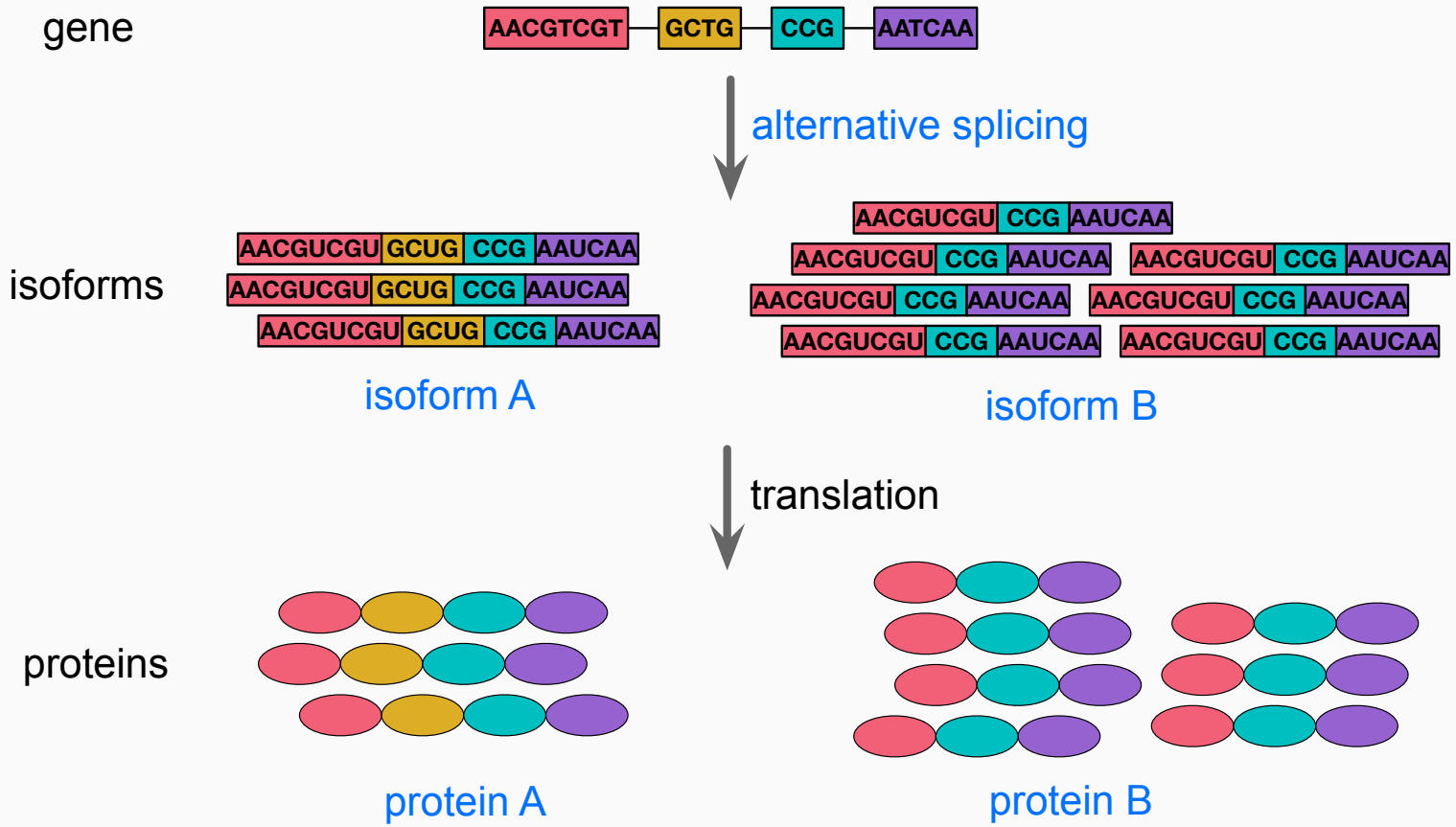
Alternative splicing

In alternative splicing, particular exons of a gene may be included into or excluded from a mature RNA isoform [Chow et al., *Cell*, 1977].



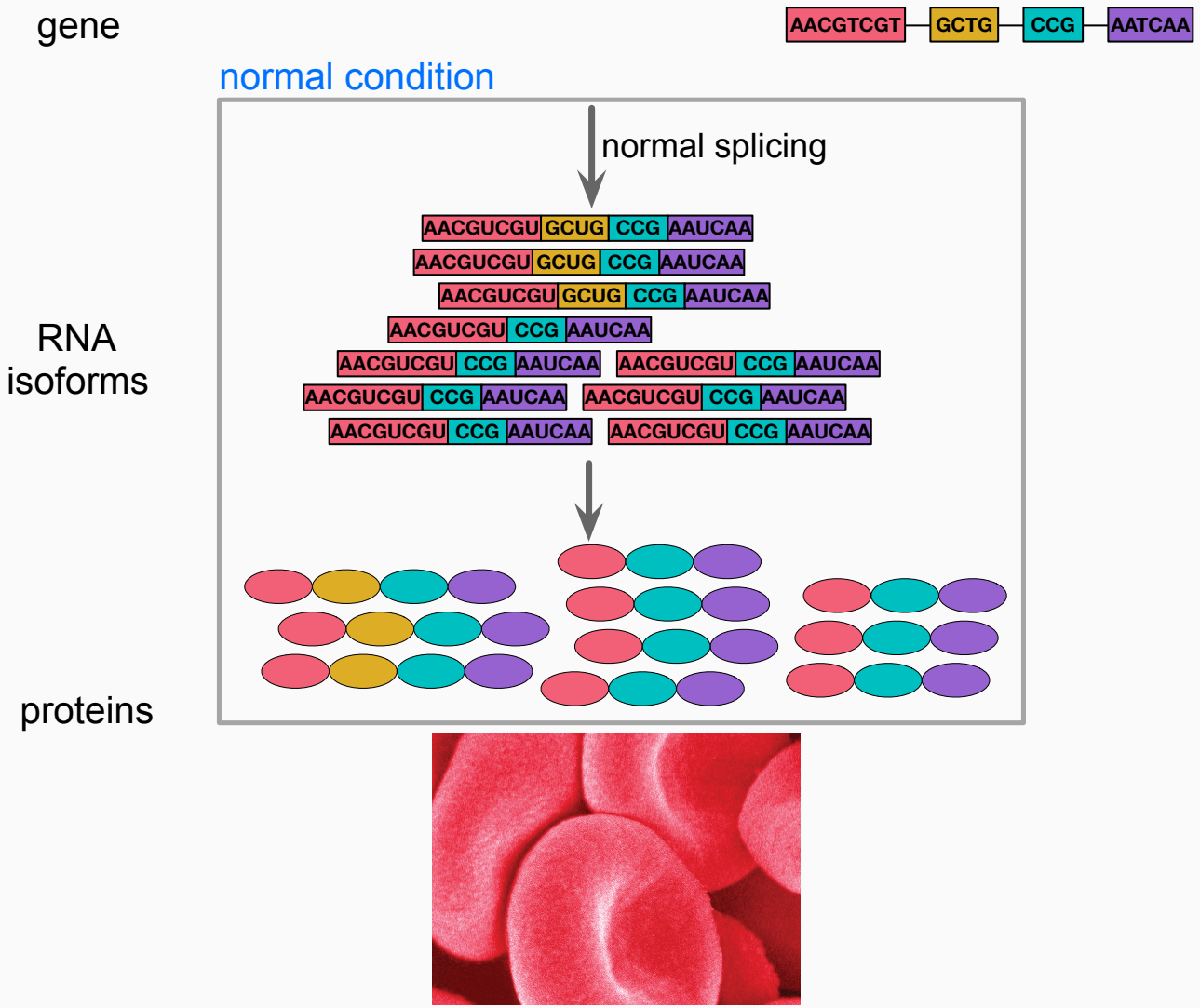
Alternative splicing

In alternative splicing, particular exons of a gene may be included into or excluded from a mature RNA isoform [Chow et al., *Cell*, 1977].



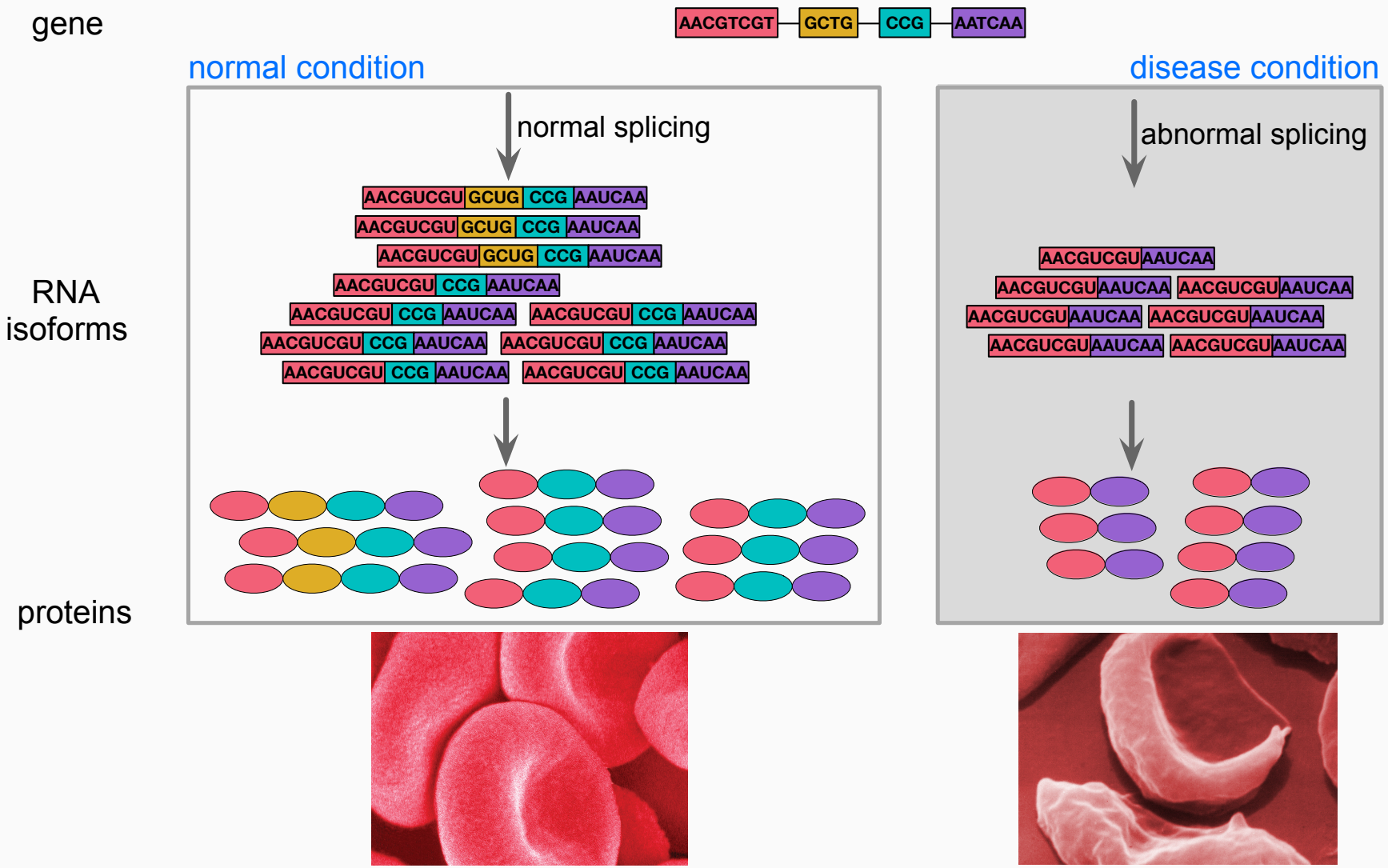
Diversity in RNA isoform structures

Abnormal splicing can lead to genetic diseases.

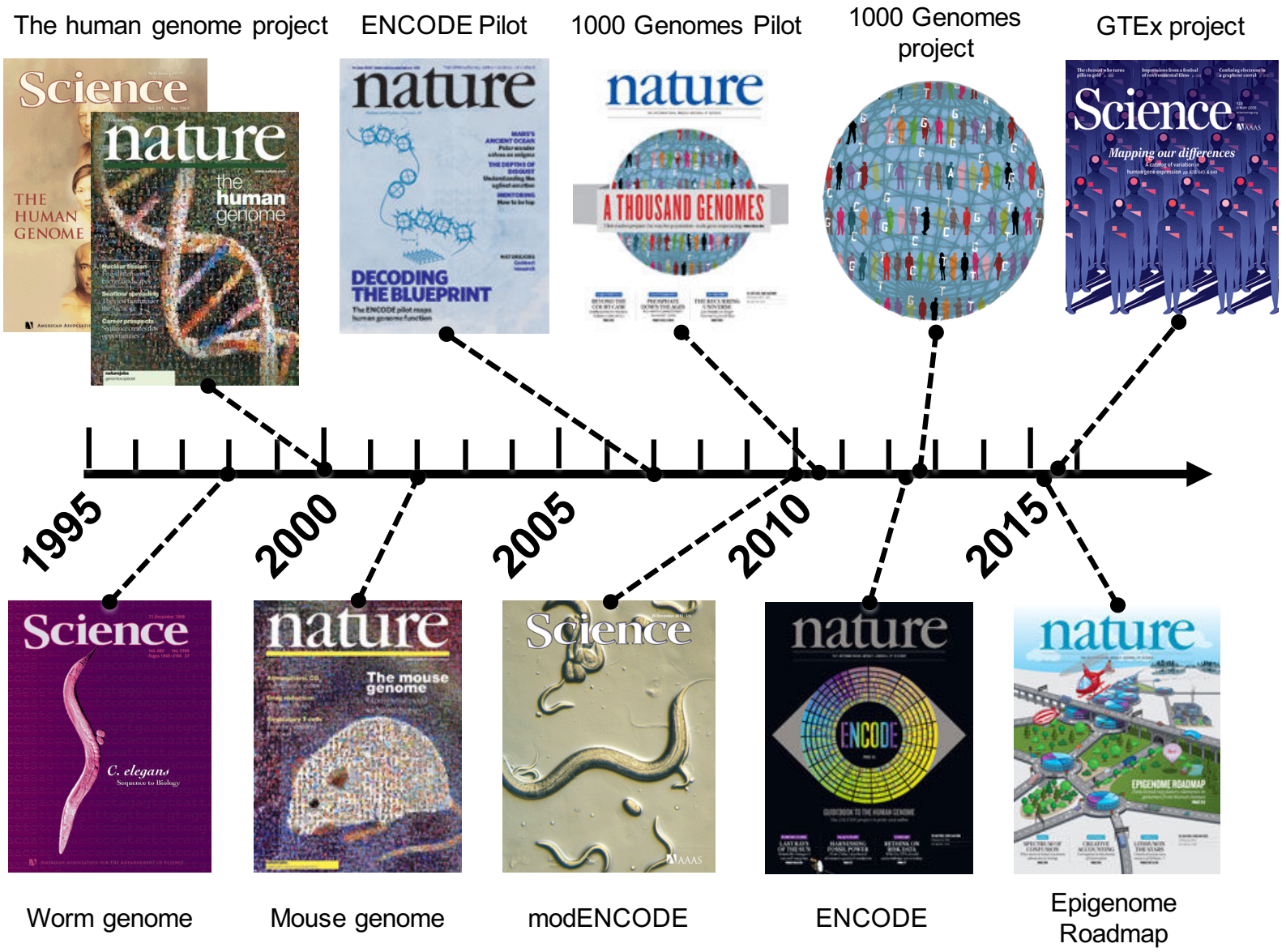


Diversity in RNA isoform structures

Abnormal splicing can lead to genetic diseases.

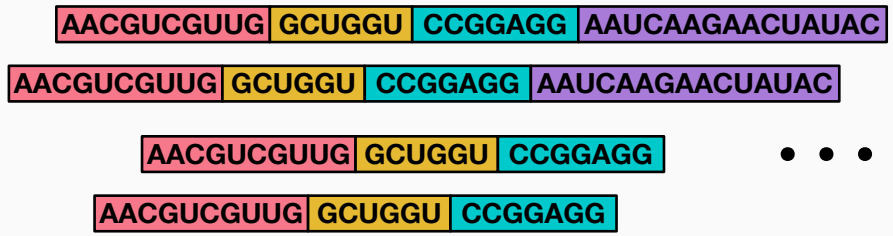


Understanding genome functions



RNA sequencing (RNA-seq) technology

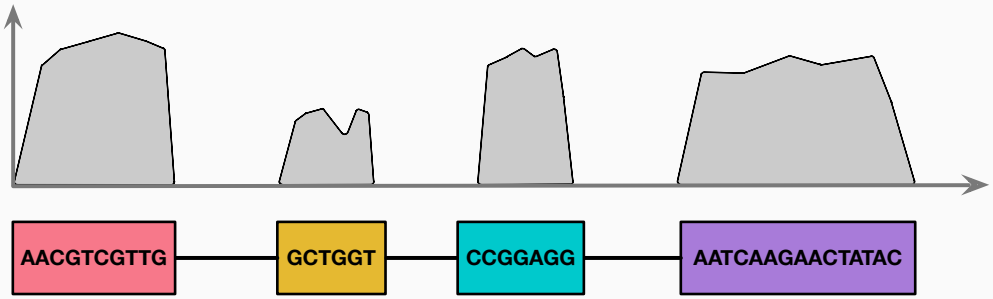
full length RNA isoforms
(unknown)



RNA-seq
experiments

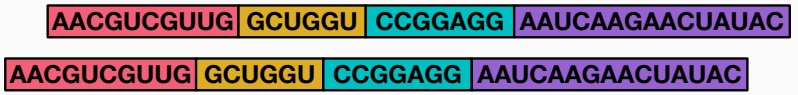
statistical
inference

RNA-seq data
(observed)

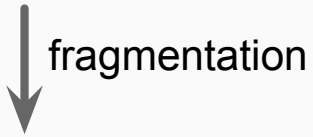


RNA sequencing (RNA-seq) experiment

full length RNA isoforms
(1712 bp on average)



• • •

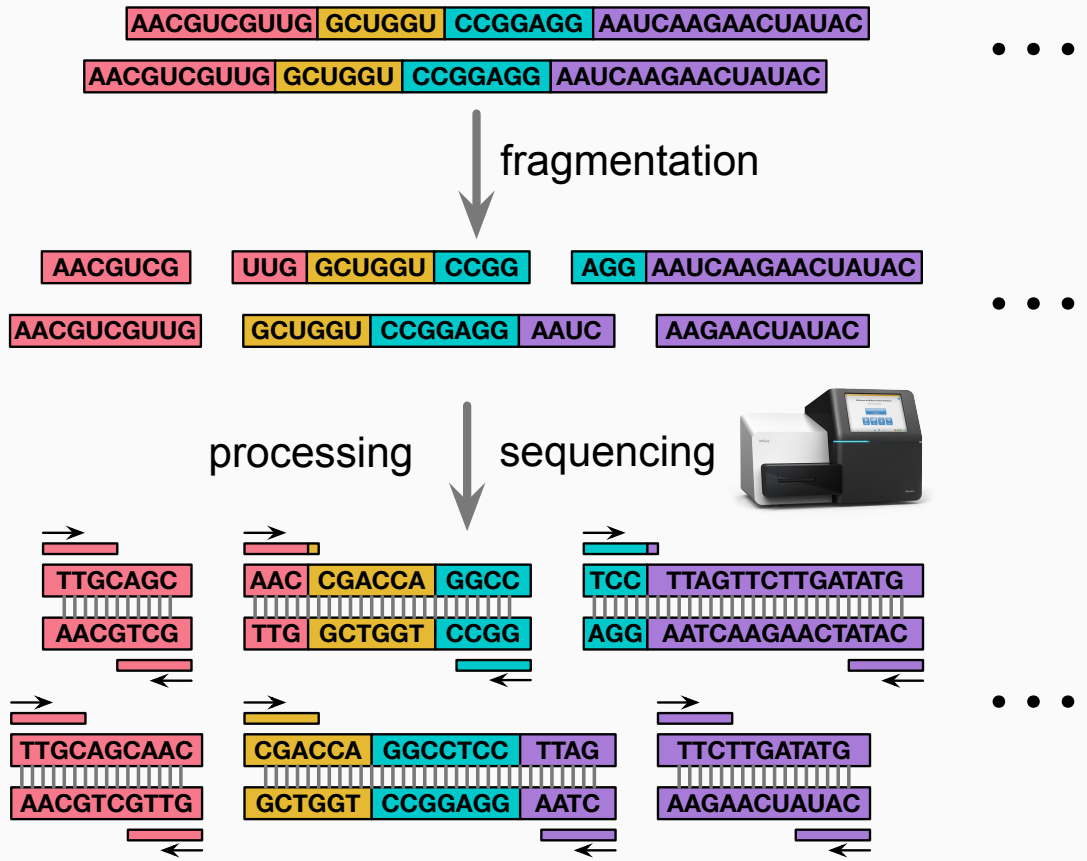


RNA fragments
(< 600 bp)

• • •

RNA sequencing (RNA-seq) experiment

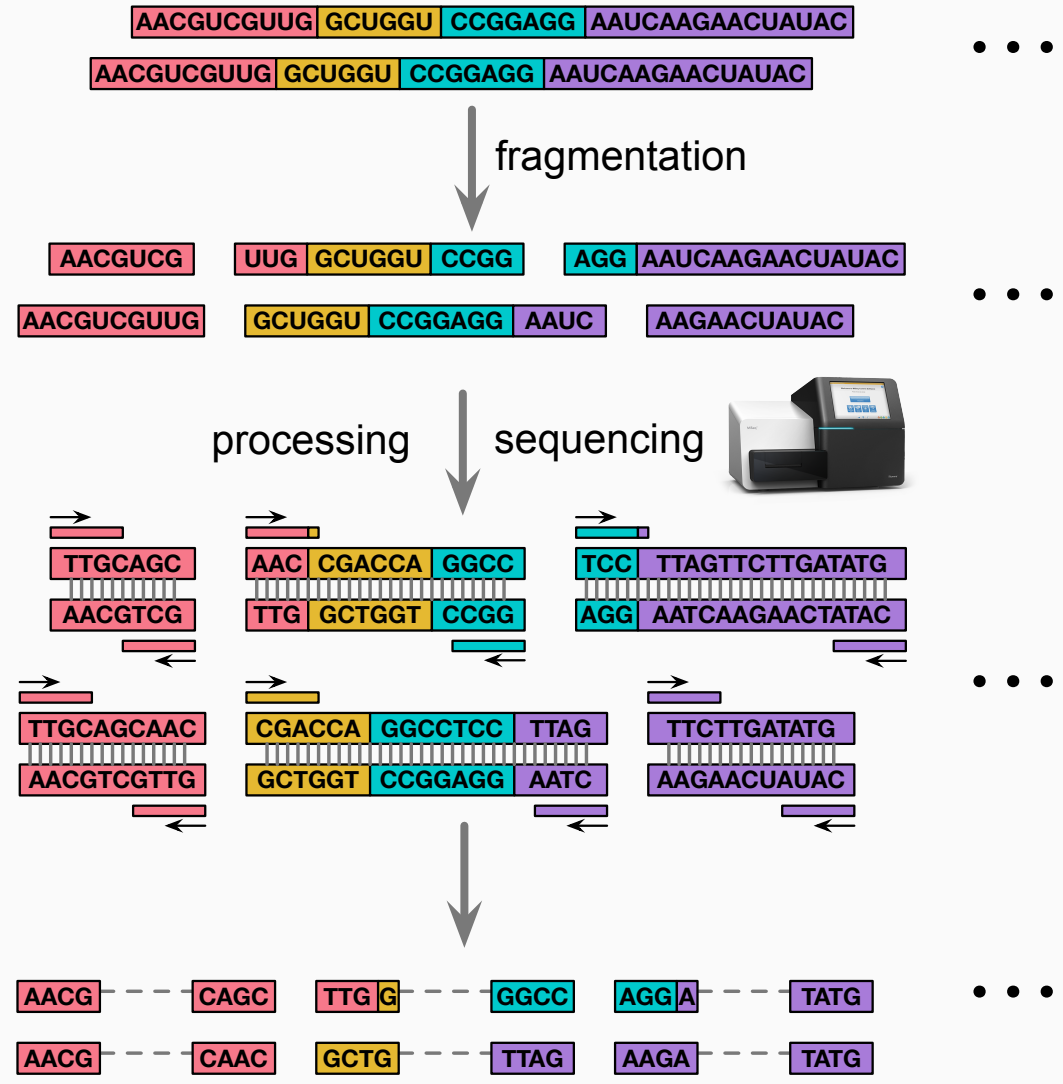
full length RNA isoforms
(1712 bp on average)



RNA fragments
(< 600 bp)

RNA sequencing (RNA-seq) experiment

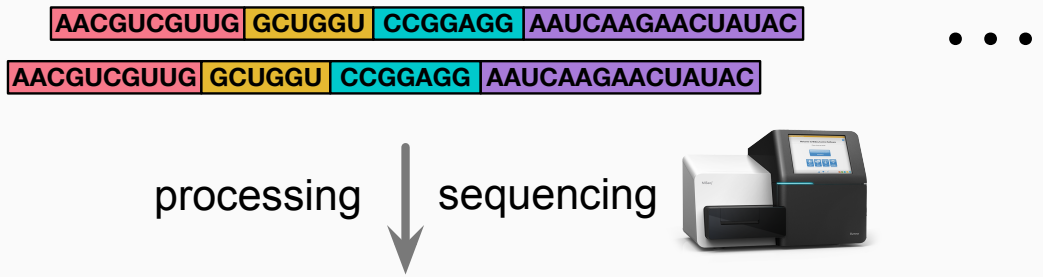
full length RNA isoforms
(1712 bp on average)



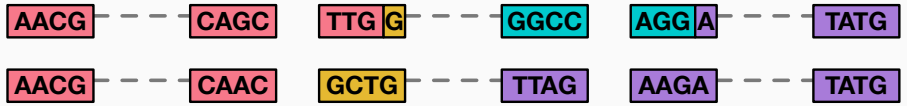
RNA-seq reads $\propto$ isoform abundance $\times$ isoform length

Mapping RNA-seq reads to the reference genome

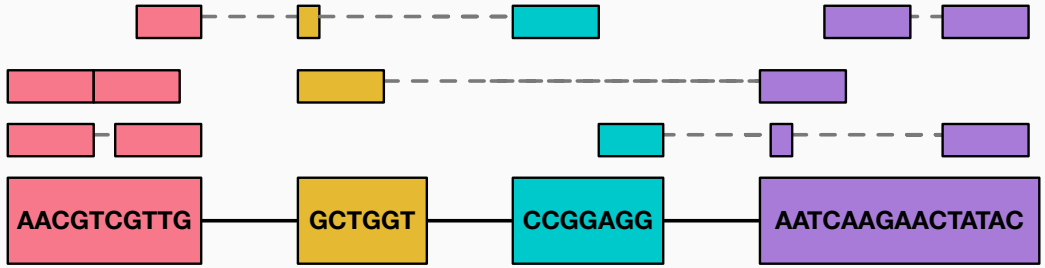
full length RNA isoforms
(1712 bp on average)



RNA-seq reads
(< 300 bp)

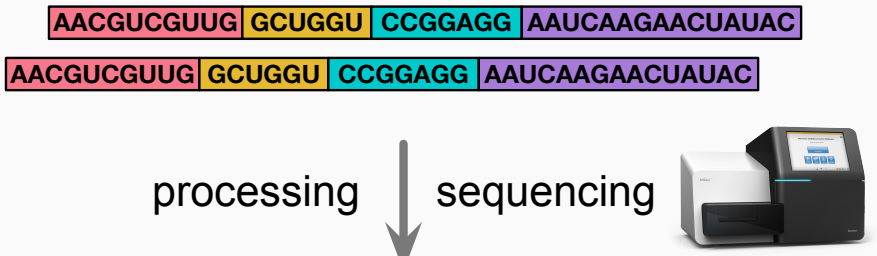


RNA-seq reads
aligned to genome

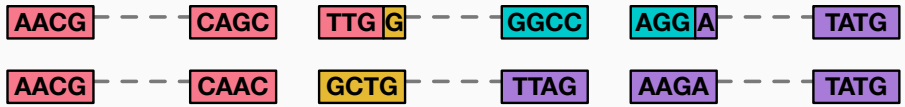


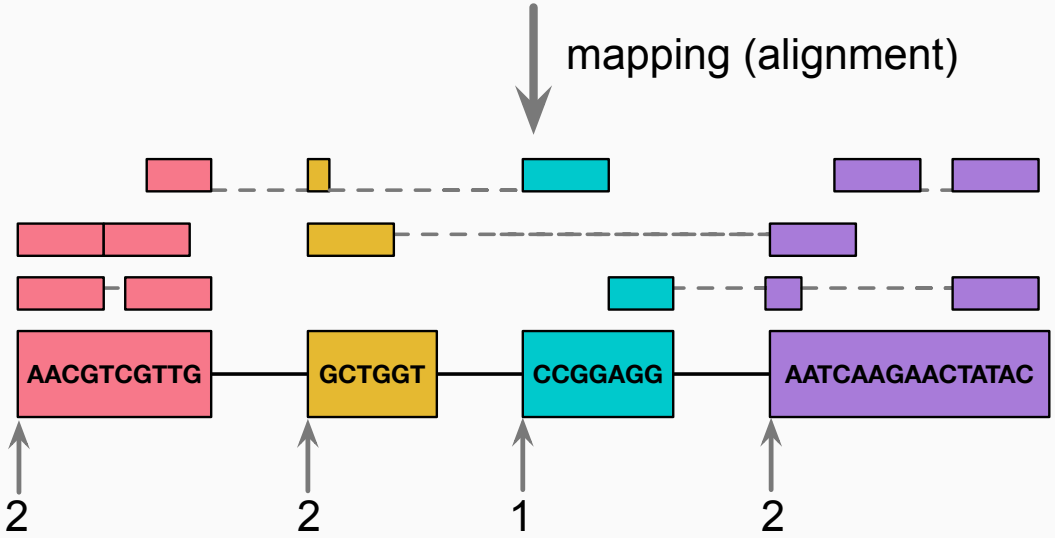
Mapping RNA-seq reads to the reference genome

full length mRNA transcript
(1712 bp on average)



RNA-seq reads
(< 300 bp)

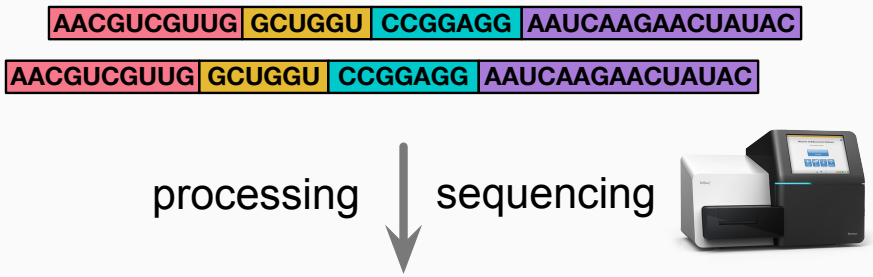




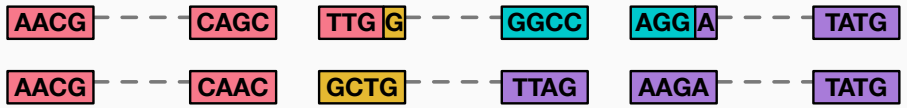
RNA-seq reads
aligned to genome

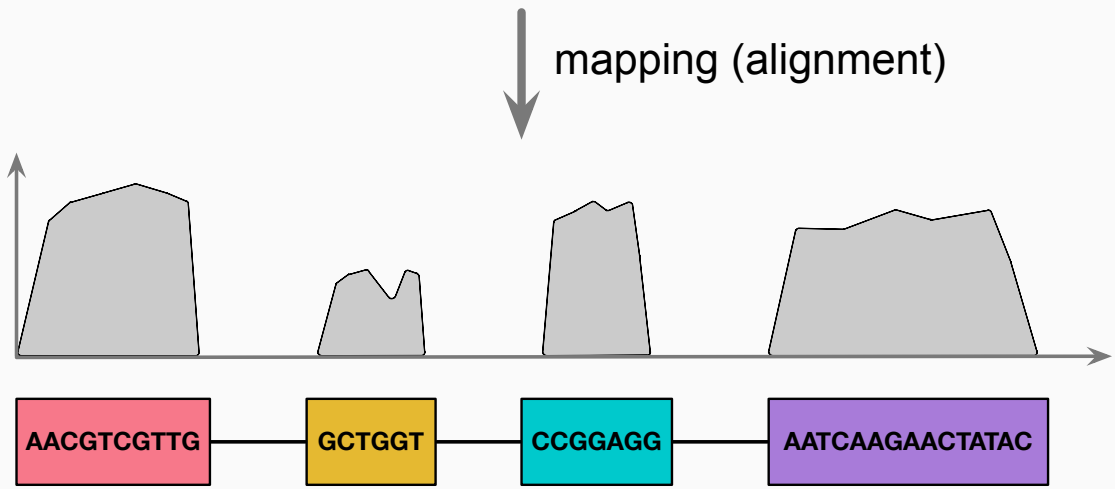
Mapping RNA-seq reads to the reference genome

full length RNA isoforms
(1712 bp on average)



RNA-seq reads
(< 300 bp)

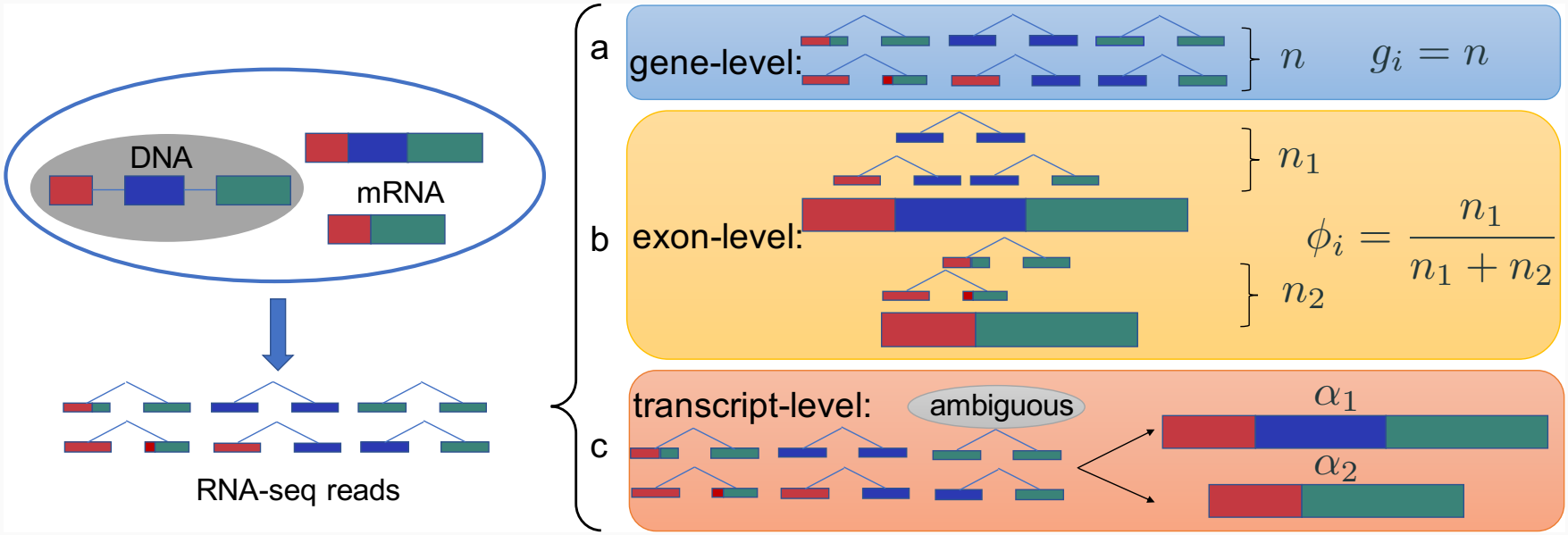




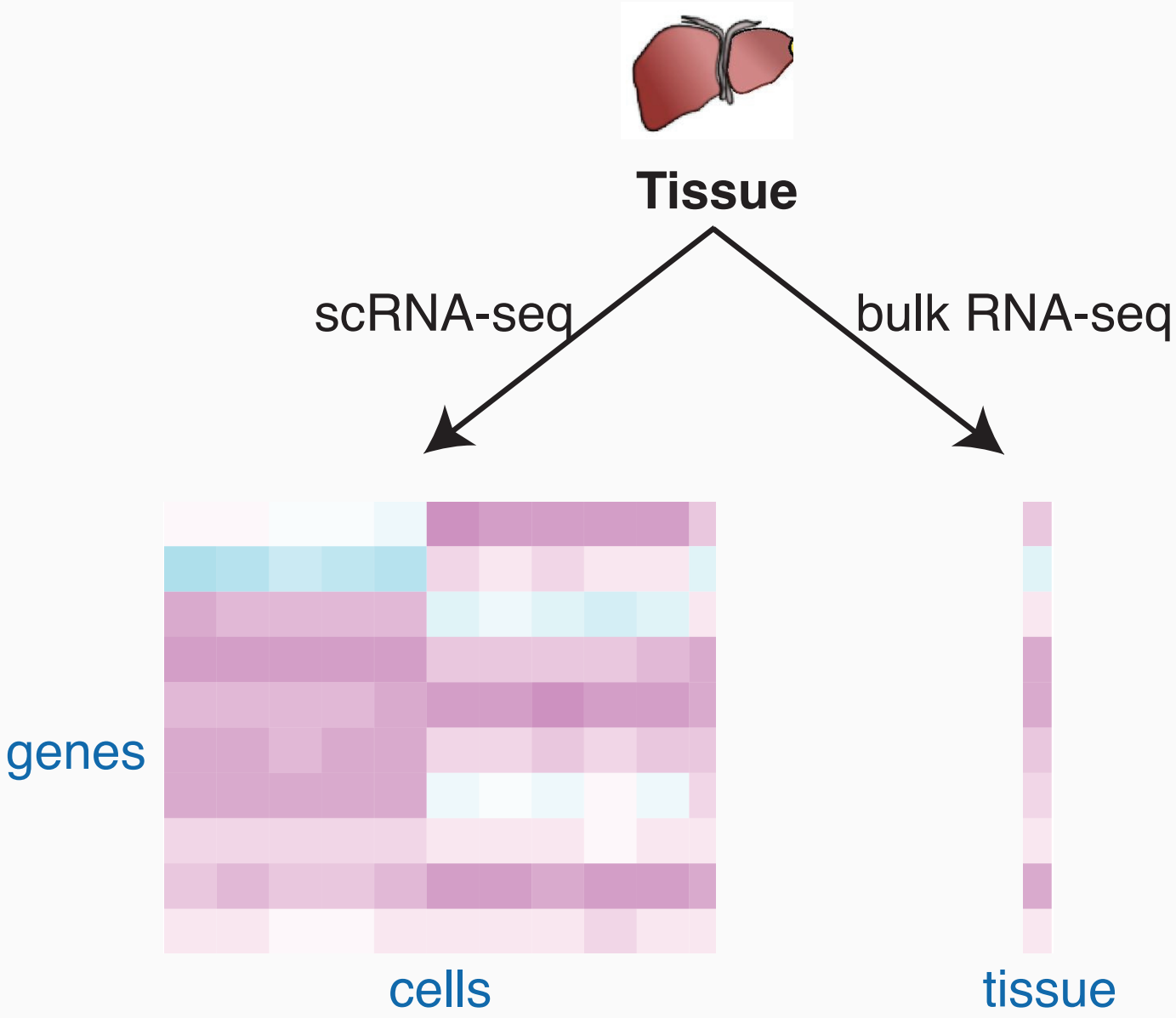
histogram of
RNA-seq read counts

Reference-based RNA-seq data analysis

1. Align RNA-seq reads to a reference genome
2. Analyze aligned reads at three levels

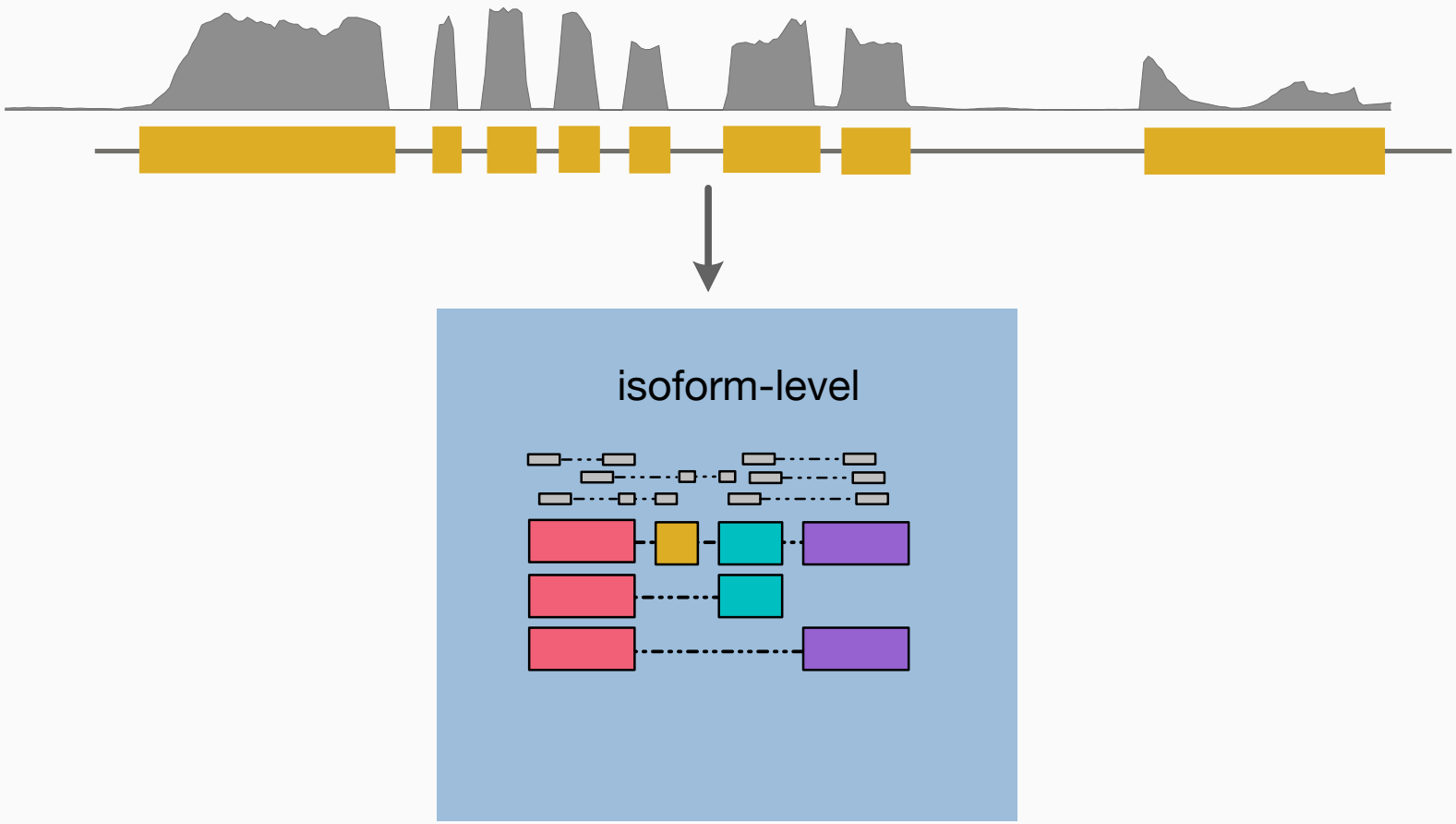


Single-cell (sc) vs. bulk RNA-seq at the gene level



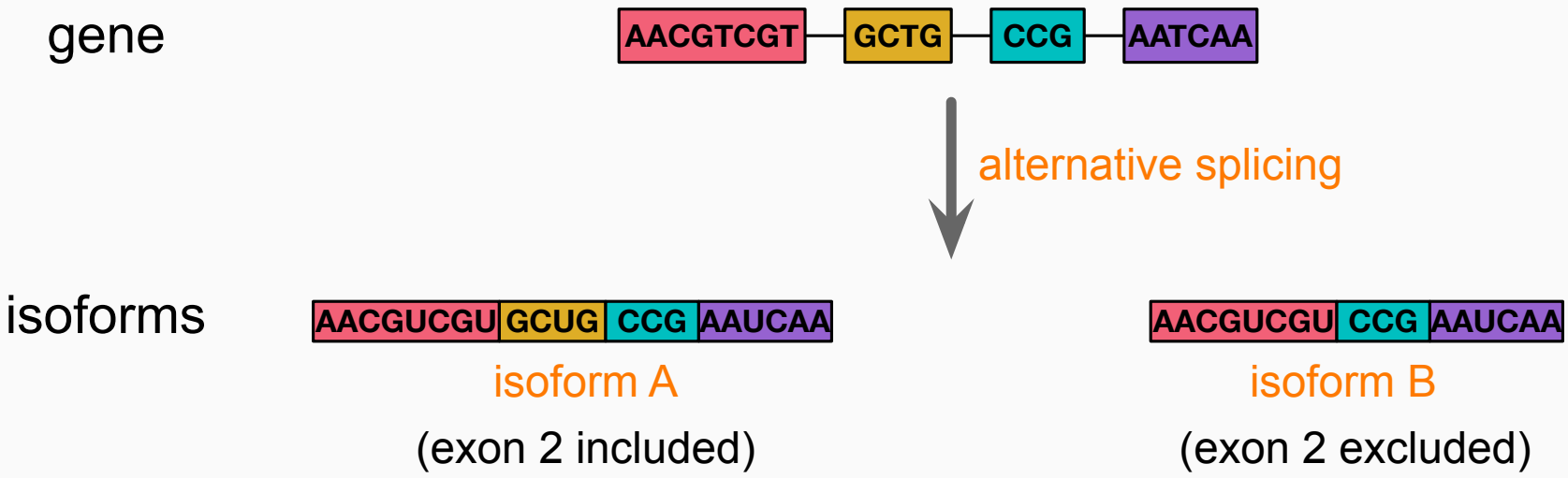
Bulk RNA-seq: transcript/isoform discovery & quantification

AIDE: annotation-assisted isoform discovery

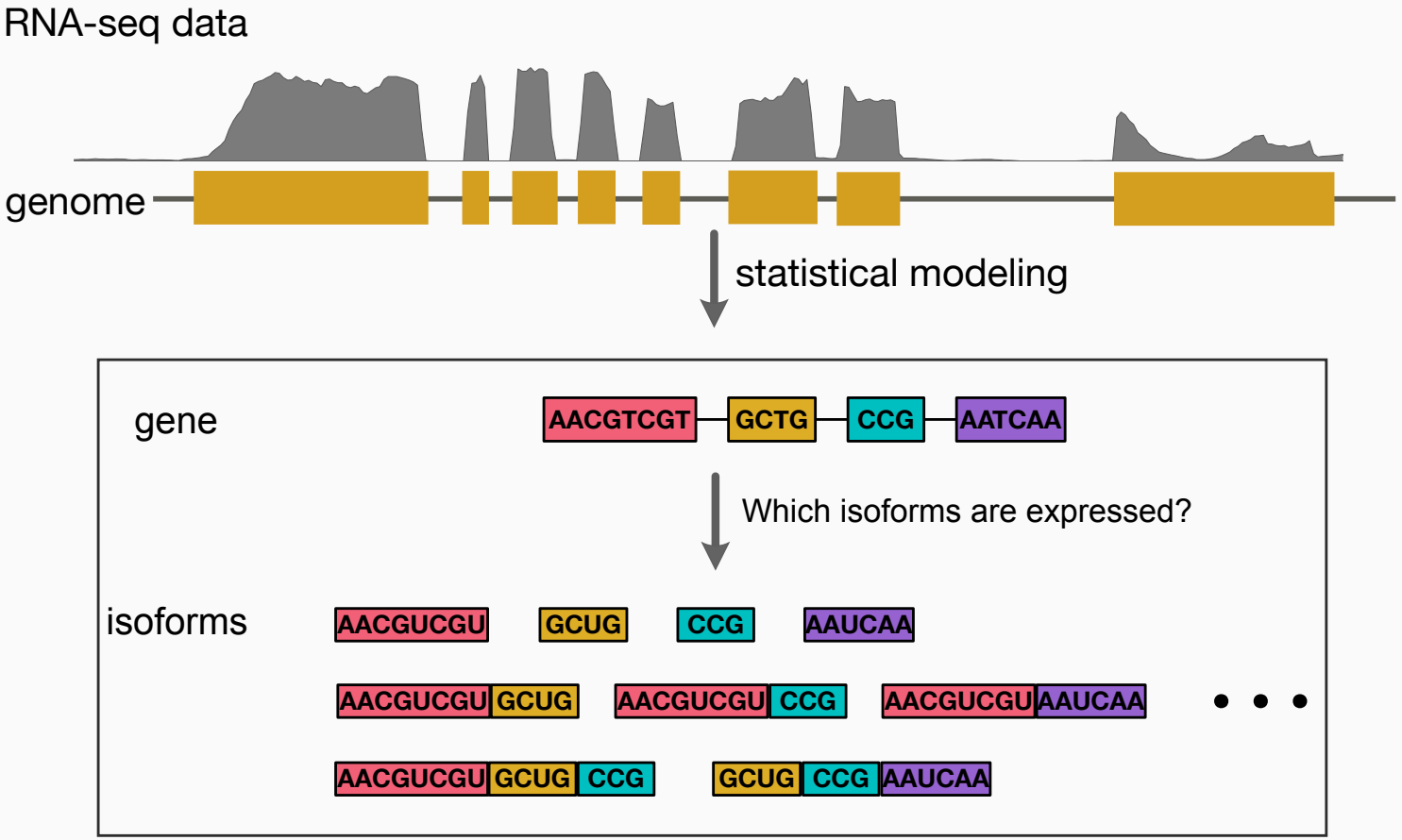


Isoform discovery: which isoforms are expressed?

- More than 90% genes undergo alternative splicing in mammals [Hooper, *Human Genomics*, 2014].
- At least 35% genetic diseases involve abnormal splicing [Manning et al., *Nature Reviews Mol. Cell Biol.* 2017].

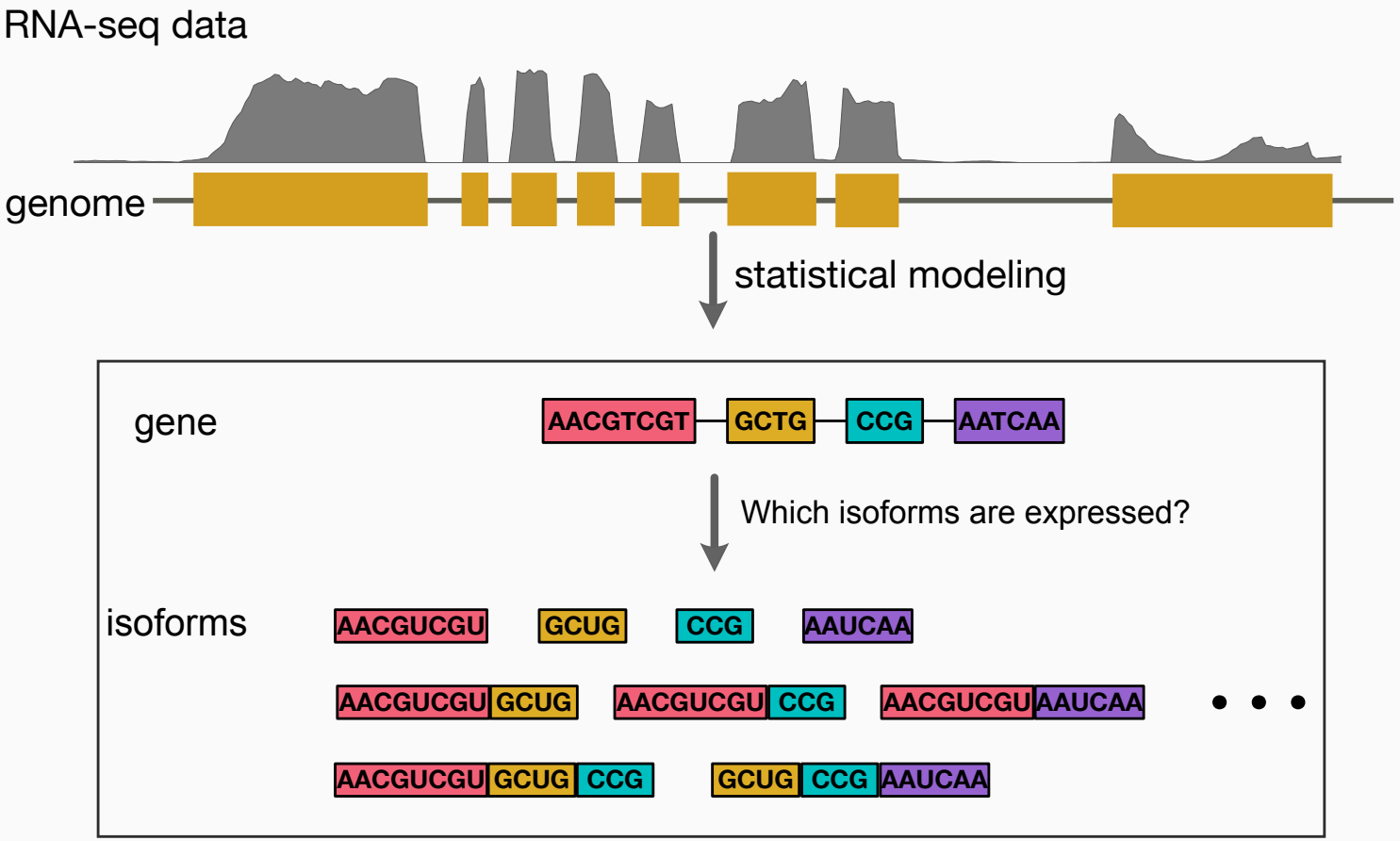


Isoform discovery: which isoforms are expressed?



Challenge 1: large number of candidate isoforms

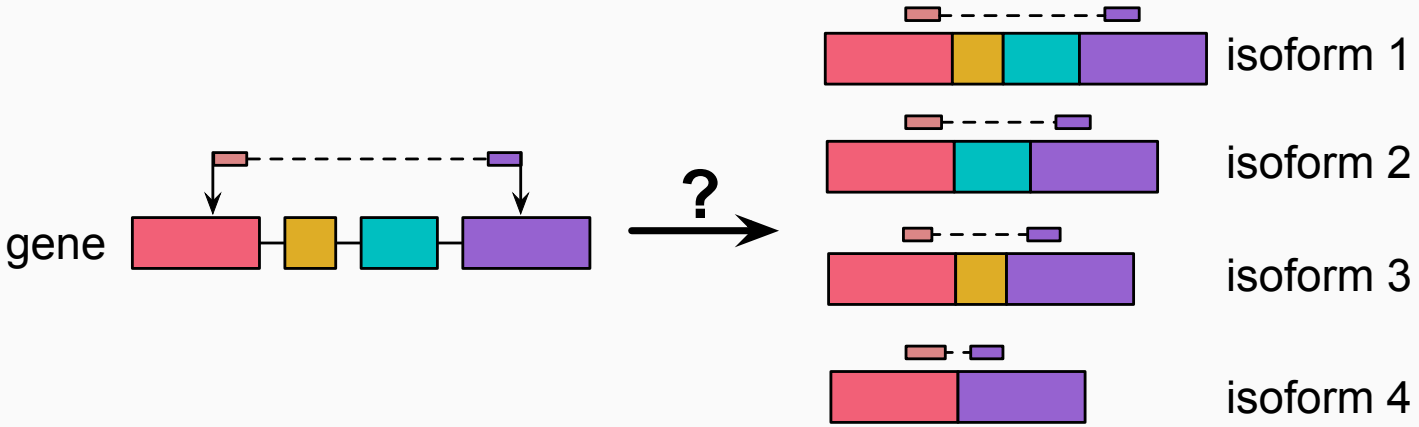
Variable size (# of candidate isoforms) $= 2^{\#\text{ of exons}} - 1$



For this 4-exon gene, $2^4 - 1 = 15$ candidate isoforms

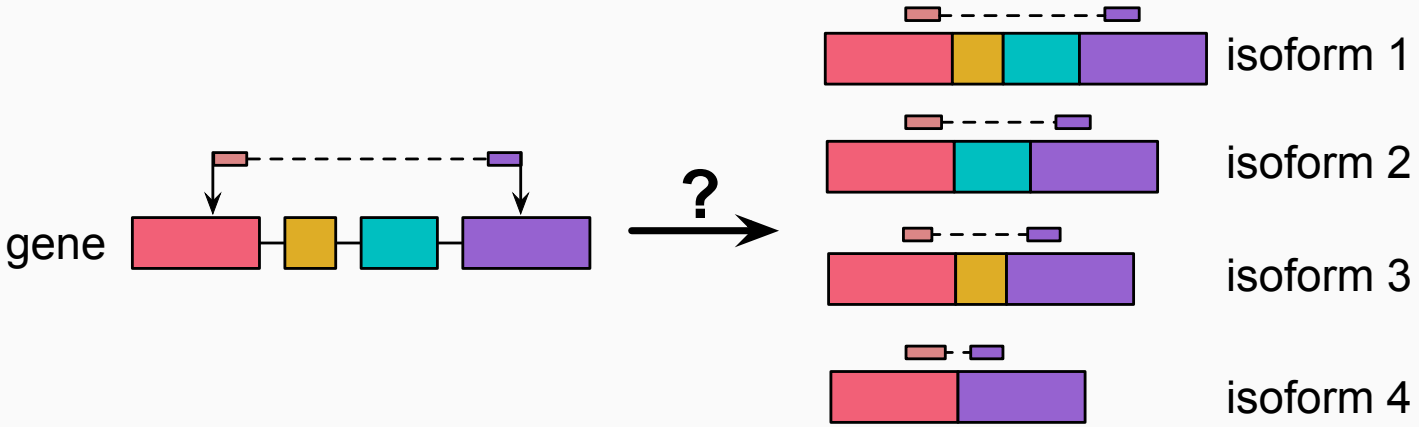
Challenge 2: great information loss

- RNA-seq reads are very short compared with full-length isoforms.
- Most RNA-seq reads do not uniquely map to a single isoform.



Challenge 2: great information loss

- RNA-seq reads are very short compared with full-length isoforms.
- Most RNA-seq reads do not uniquely map to a single isoform.



- Technical biases introduced into RNA-seq experiments.

Existing isoform discovery methods

State-of-the-art methods for isoform discovery:

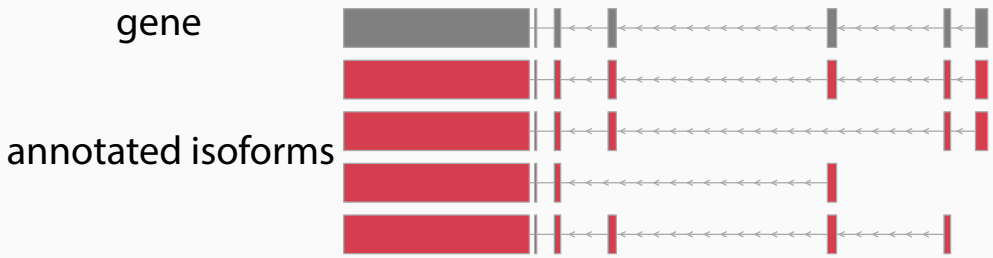
- SIIER [Jiang et al., *Bioinformatics*, 2009]
- Cufflinks [Trapnell et al., *Nature Biotechnology*, 2010]
- SLIDE [Li et al., *Proc. Natl. Acad. Sci.* 2011]
- StringTie [Pertea et al., *Nature Biotechnology*, 2015]
- ...

Limitations:

1. Low accuracy for genes with complex splicing structures.
2. Difficult to improve isoform-level performance.
   [Kanitz et al., *Genome Biology*, 2015]
3. Usage of annotations results in false positives.

Usage of annotations results in false positives

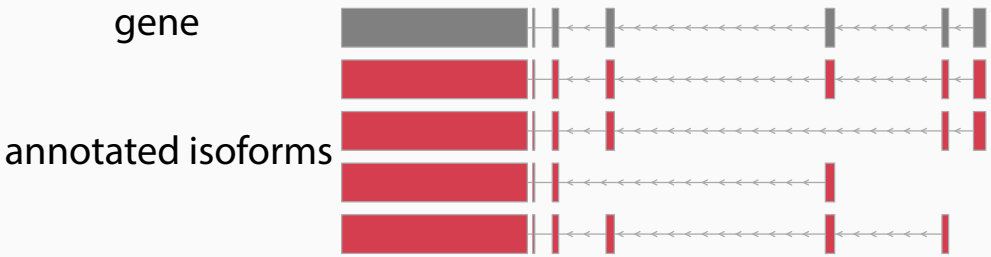
Annotated isoforms are experimentally validated:



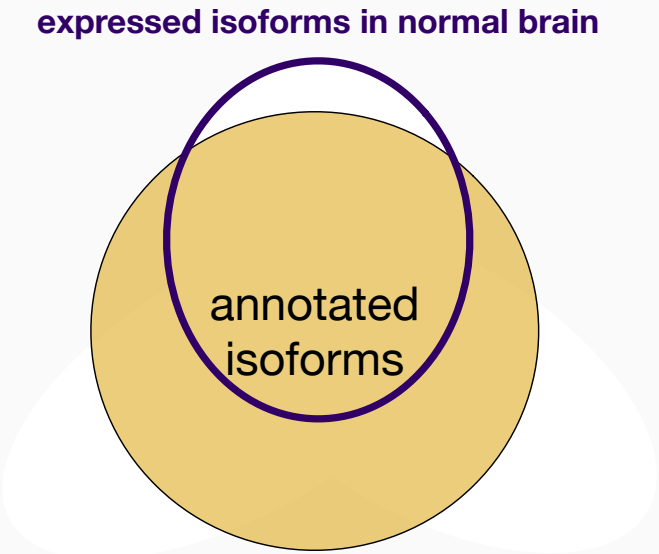
- *Ensembl* database: $203,903$ isoforms
  [Zerbino et al., *Nucleic Acids Research*, 2017]

Usage of annotations results in false positives

Annotated isoforms are experimentally validated:

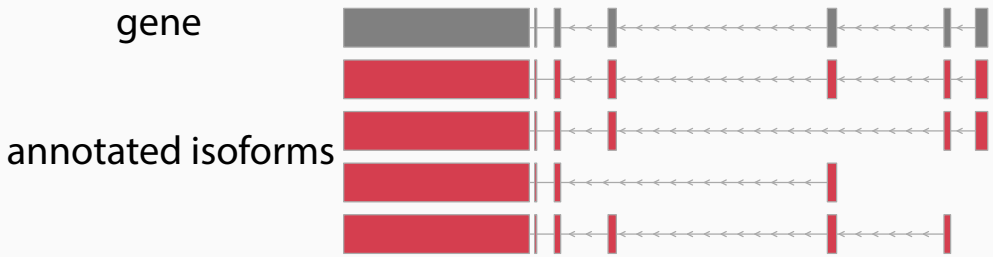


- *Ensembl* database: 203,903 isoforms
  [Zerbino et al., *Nucleic Acids Research*, 2017]

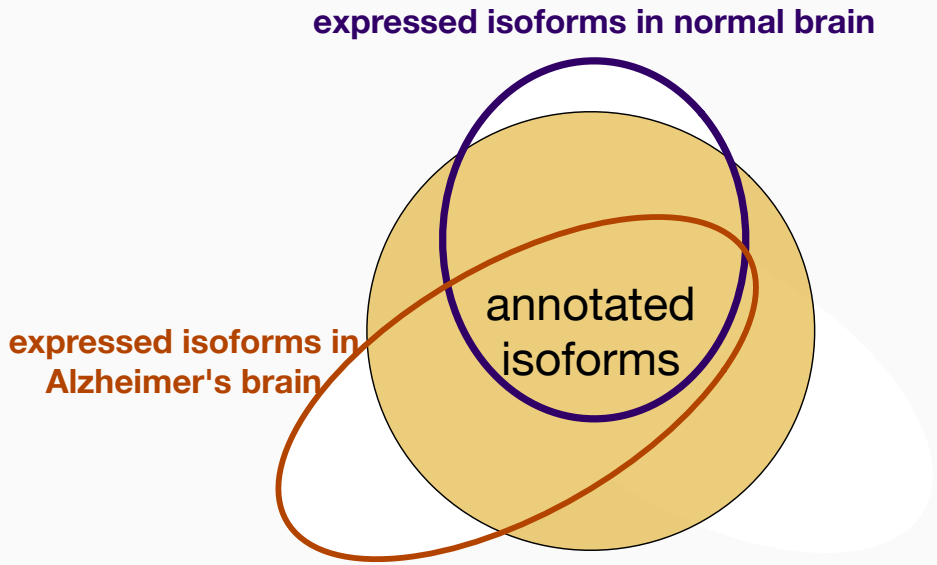


Usage of annotations results in false positives

Annotated isoforms are experimentally validated:

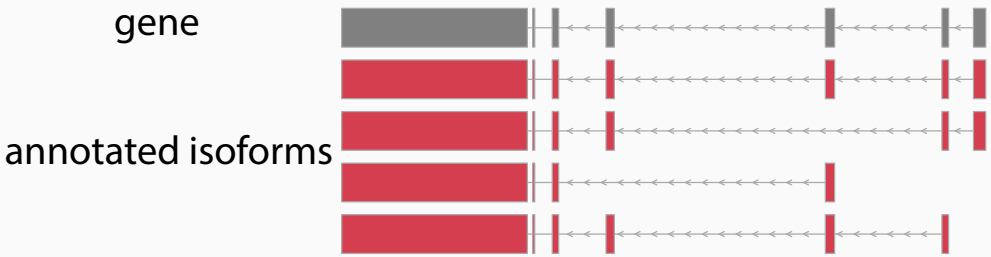


- *Ensembl* database: 203,903 isoforms
  [Zerbino et al., *Nucleic Acids Research*, 2017]

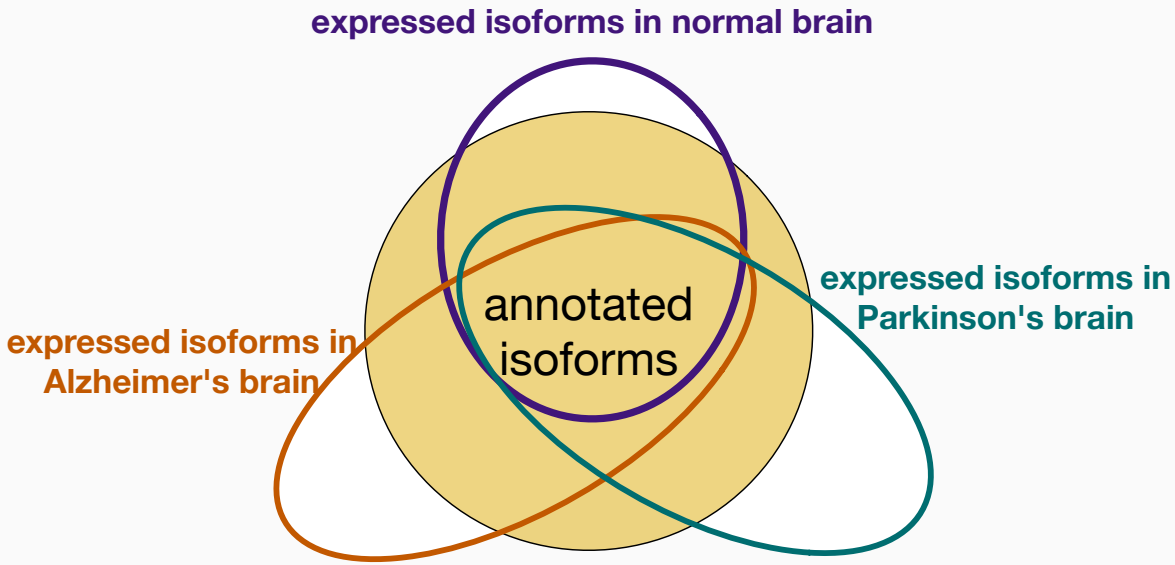


Usage of annotations results in false positives

Annotated isoforms are experimentally validated:



- *Ensembl* database: 203, 903 isoforms
  [Zerbino et al., *Nucleic Acids Research*, 2017]



False positives $\rightarrow$ false discoveries

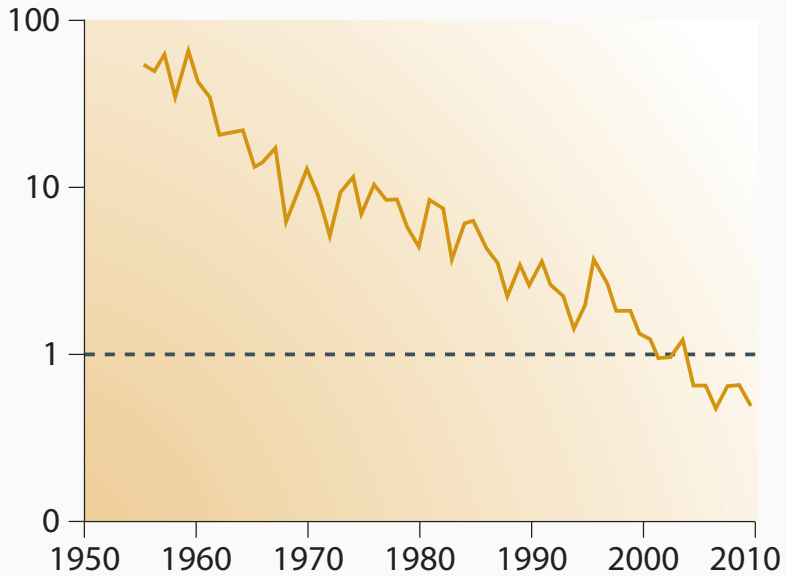
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

68,836 Save	3,644 Citation
2,622,757 View	10,479 Share

Number of drugs per billion US$ R&D spending



[Scannell et al., *Nat. Rev. Drug Discov.* 2012]

Highlights of the AIDE method

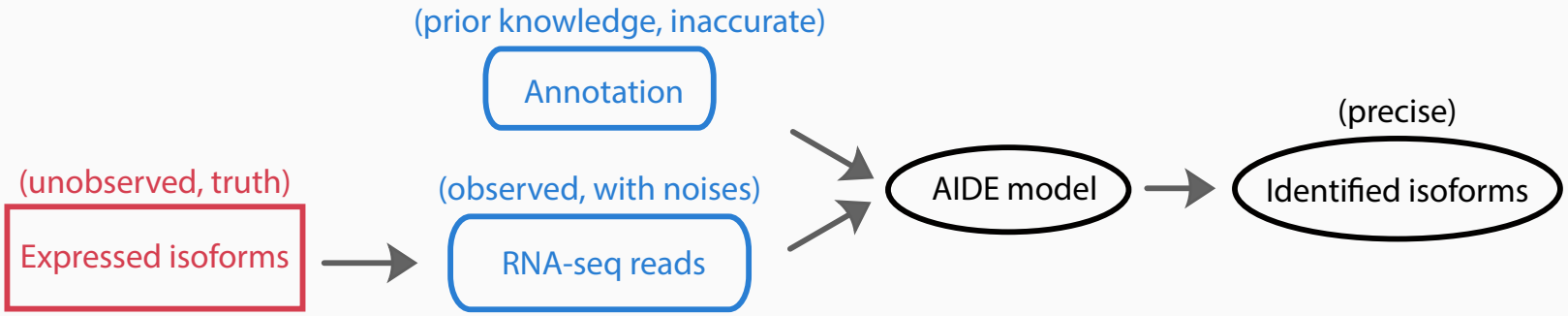
1. Selectively leverage annotation information to increase the precision and robustness of isoform discovery.

Highlights of the AIDE method

1. Selectively leverage annotation information to increase the precision and robustness of isoform discovery.
2. Practical probabilistic model to account for technical biases.
3. Conservatively identify isoforms that make statistically significant contributions to explaining the observed RNA-seq reads.

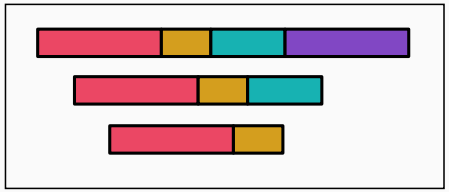
Highlights of the AIDE method

1. Selectively leverage annotation information to increase the precision and robustness of isoform discovery.
2. Practical probabilistic model to account for technical biases.
3. Conservatively identify isoforms that make statistically significant contributions to explaining the observed RNA-seq reads.
4. First method to control false discoveries by employing a statistical testing procedure.

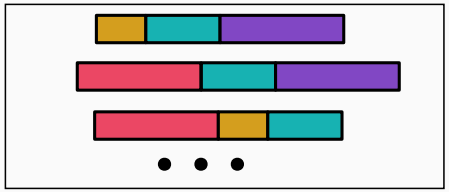


The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:

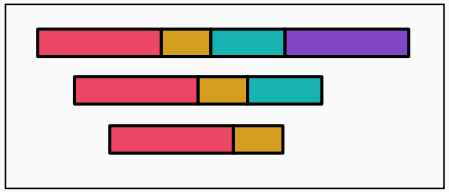


Stage 1: candidates are annotated isoforms only

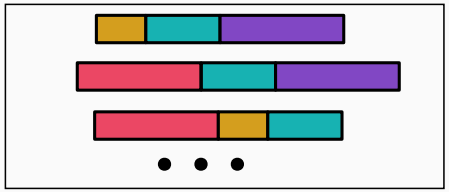
Initialization → Forward step ⇄ Backward step

The stepwise selection in AIDE: two stages

annotated isoforms:

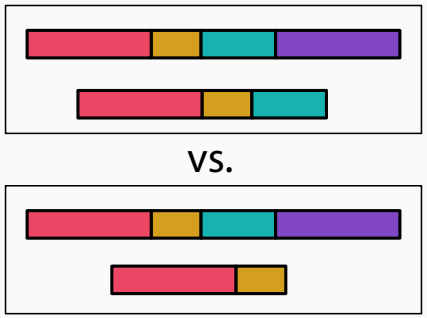


non-annotated isoforms:



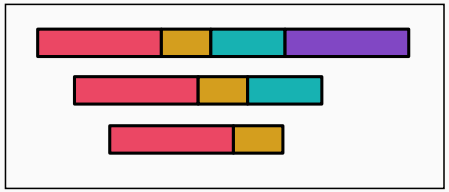
Stage 1: candidates are annotated isoforms only

Initialization → Forward step Backward step

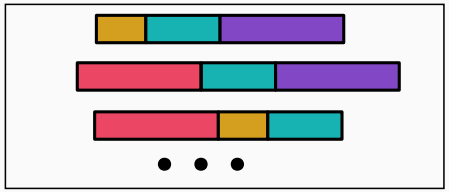


The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:



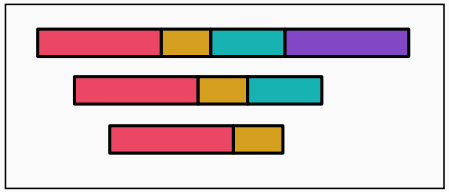
Stage 1: candidates are annotated isoforms only

Initialization → Forward step → Backward step

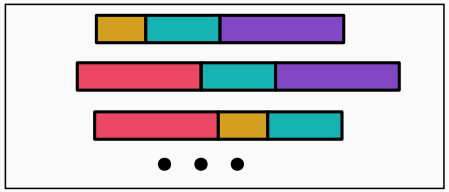
selected based on MLE

The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:



Stage 1: candidates are annotated isoforms only

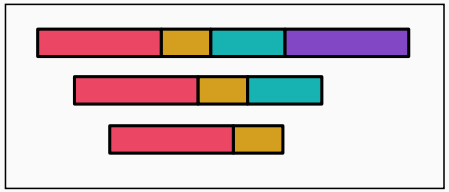
Initialization → Forward step

vs.
LRT

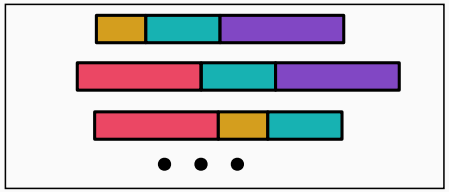
Backward step

The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:

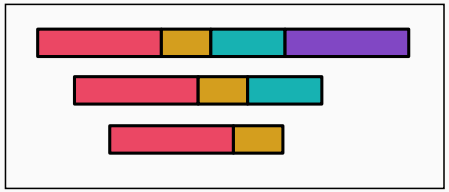


Stage 1: candidates are annotated isoforms only

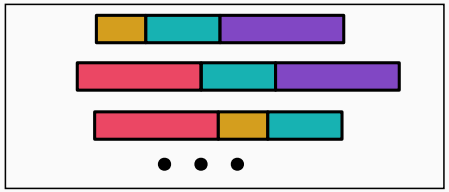
Initialization → Forward step Backward step

The stepwise selection in AIDE: two stages

annotated isoforms:



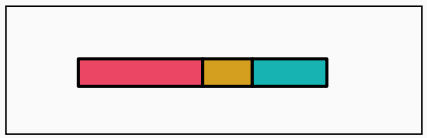
non-annotated isoforms:



Stage 1: candidates are annotated isoforms only

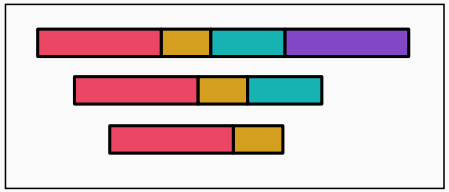
Initialization → Forward step → Backward step

vs.

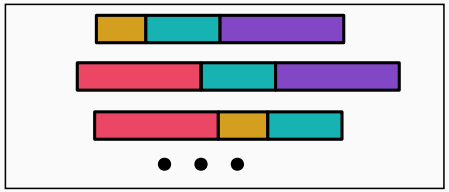


The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:



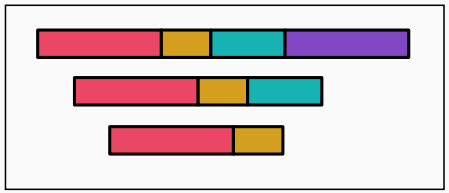
Stage 1: candidates are annotated isoforms only

Initialization → Forward step → Backward step

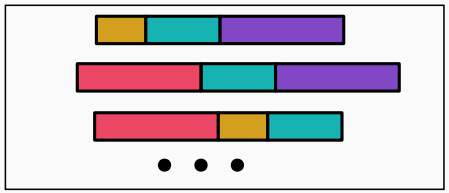
selected based on MLE

The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:



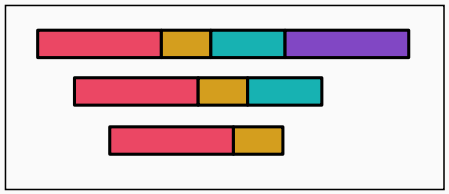
Stage 1: candidates are annotated isoforms only

Initialization → Forward step → Backward step

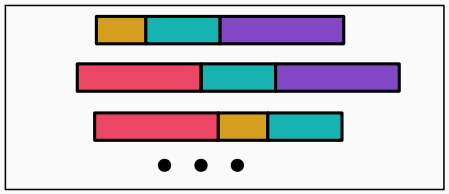
vs.
LRT

The stepwise selection in AIDE: two stages

annotated isoforms:



non-annotated isoforms:



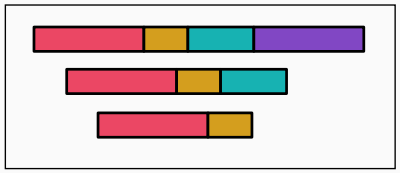
Stage 1: candidates are annotated isoforms only

Initialization → Forward step ⇄ Backward step

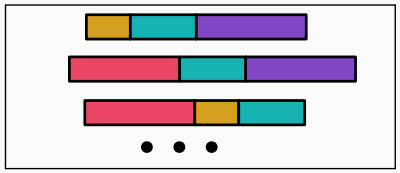
output

The stepwise selection in AIDE: two stages

annotated isoforms:



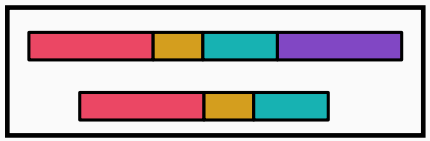
non-annotated isoforms:



Stage 1: candidates are annotated isoforms only

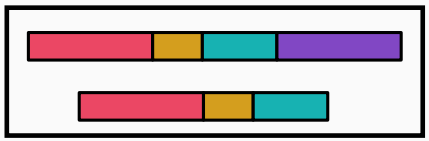
Initialization → Forward step ⇄ Backward step

output



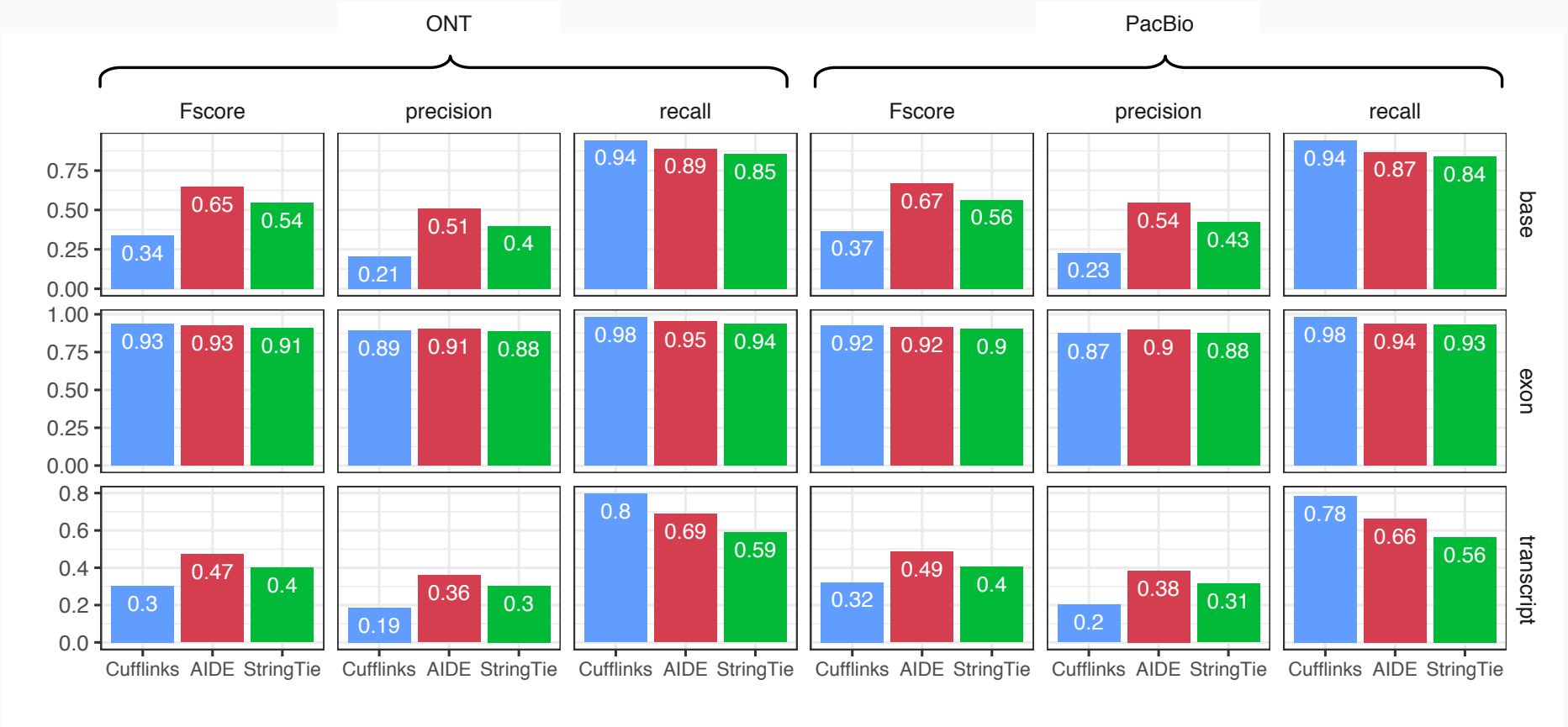
Stage 2: candidates are all possible isoforms

Initialization → Forward step ⇄ Backward step



AIDE outperforms state-of-the-art methods

- Human embryonic stem cells
- Input: Illumina RNA-seq data
- Evaluation: PacBio and Nanopore ONT RNA-seq data

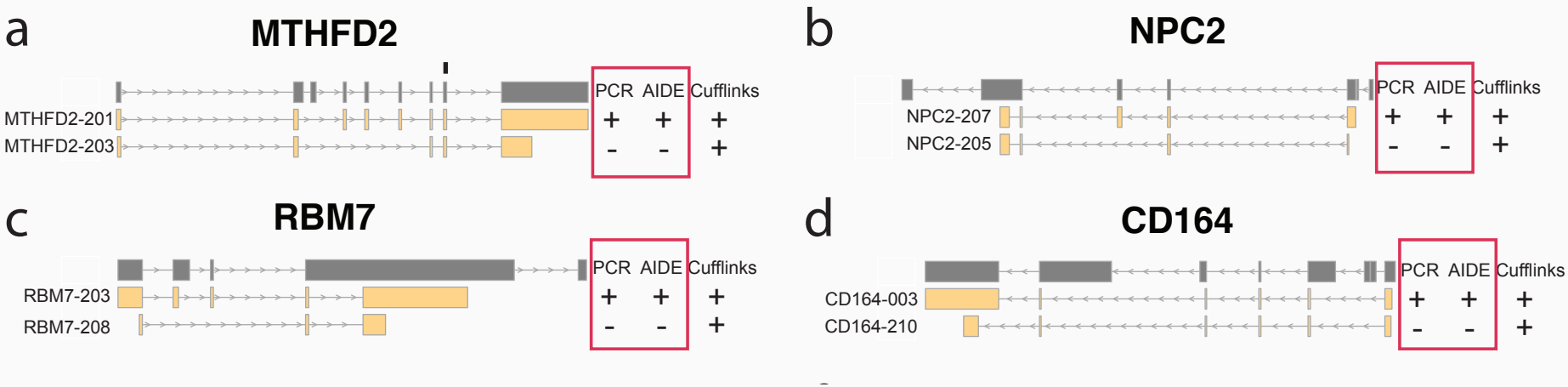


AIDE effectively reduces false discoveries in real data

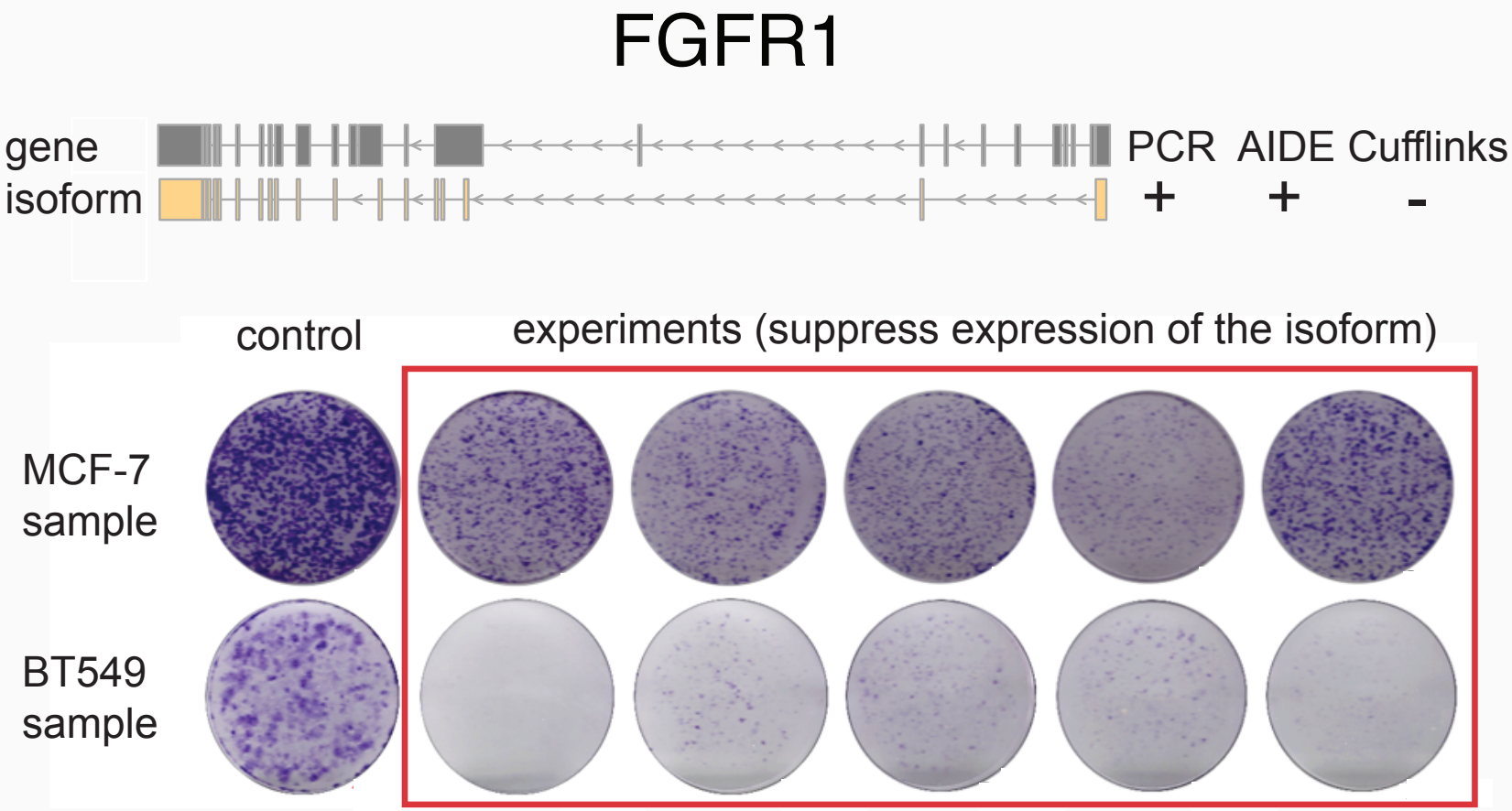
- Data: breast cancer RNA-seq samples
- Six genes:
  - isoforms identified only by Cufflinks but not by AIDE
  - experimental validation (PCR)

AIDE effectively reduces false discoveries in real data

- Data: breast cancer RNA-seq samples
- Six genes:
  - isoforms identified only by Cufflinks but not by AIDE
  - experimental validation (PCR)
- Four genes:
  the isoforms uniquely predicted by Cufflinks were false positives

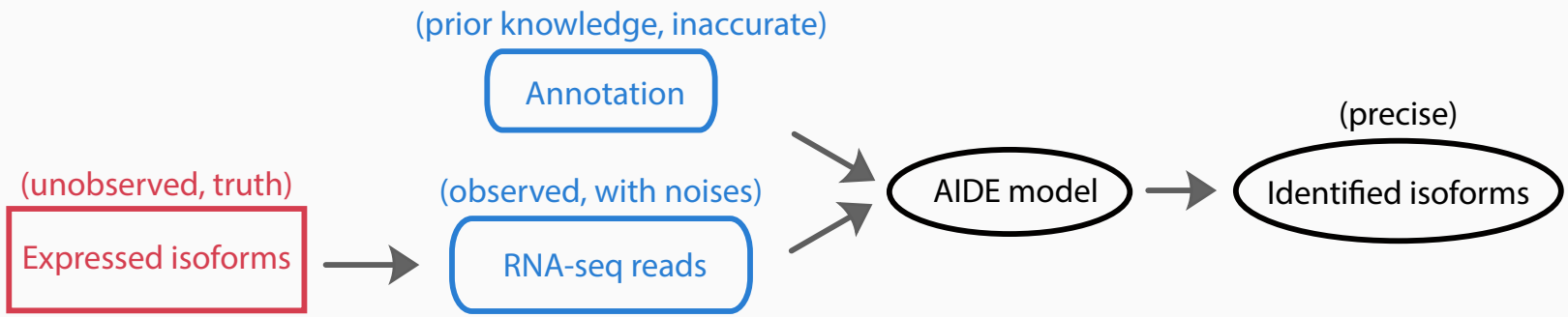


AIDE discovers isoforms with biological significance



Summary of the AIDE method

- The first isoform discovery method that directly controls false discoveries by implementing the statistical model selection principle.

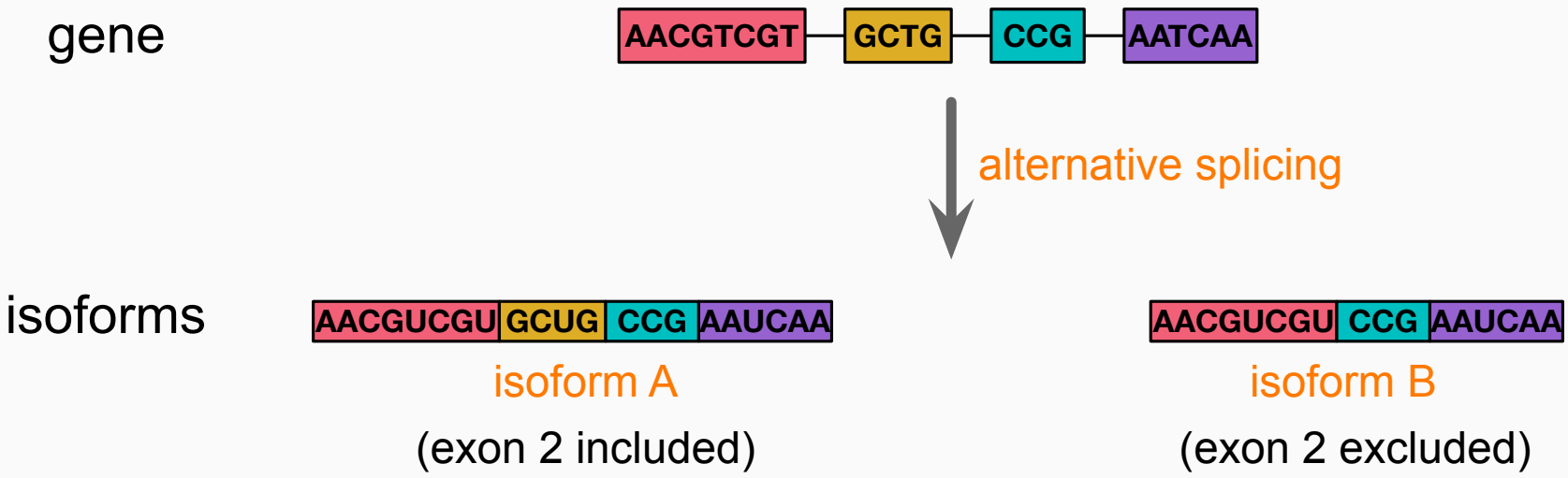


- Software: `https://github.com/Vivianstats/AIDE`
- Manuscript:

Under review at *Genome Research*.

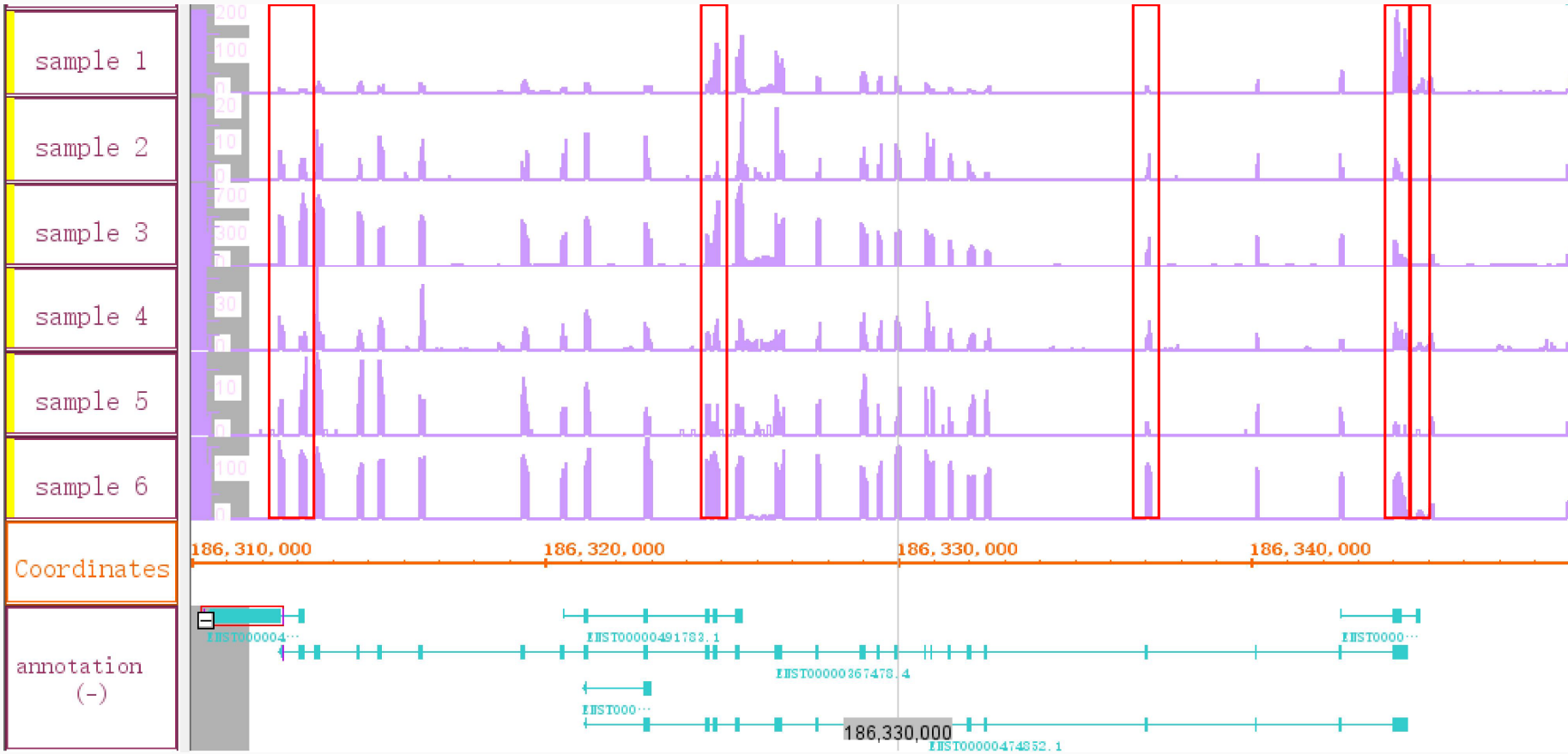
Isoform quantification: what are the isoform expression levels?

- More than 90% genes undergo alternative splicing in mammals [Hooper, *Human Genomics*, 2014].
- At least 35% genetic diseases involve abnormal splicing [Manning et al., *Nature Reviews Mol. Cell Biol.* 2017].



Motivation: multiple human ESC RNA-seq samples

chr1; gene: *TPR*



How to combine multiple RNA-seq samples?

Given D RNA-Seq (technical or biological) replicate samples and gene annotations, how to estimate the abundance of each annotated isoform for every gene?

How to combine multiple RNA-seq samples?

Given D RNA-Seq (technical or biological) replicate samples and gene annotations, how to estimate the abundance of each annotated isoform for every gene?

- Apply a single-sample method to each sample separately and then average the estimated isoform abundance across multiple samples?

How to combine multiple RNA-seq samples?

Given D RNA-Seq (technical or biological) replicate samples and gene annotations, how to estimate the abundance of each annotated isoform for every gene?

- Apply a single-sample method to each sample separately and then average the estimated isoform abundance across multiple samples?
  - This does not fully use the multi-sample information to reduce the variance in estimating isoform abundance

How to combine multiple RNA-seq samples?

Given D RNA-Seq (technical or biological) replicate samples and gene annotations, how to estimate the abundance of each annotated isoform for every gene?

- Apply a single-sample method to each sample separately and then average the estimated isoform abundance across multiple samples?
  - This does not fully use the multi-sample information to reduce the variance in estimating isoform abundance

- Apply a single-sample method to a pooled sample from the D samples?

How to combine multiple RNA-seq samples?

Given D RNA-Seq (technical or biological) replicate samples and gene annotations, how to estimate the abundance of each annotated isoform for every gene?

- Apply a single-sample method to each sample separately and then average the estimated isoform abundance across multiple samples?
  - This does not fully use the multi-sample information to reduce the variance in estimating isoform abundance
- Apply a single-sample method to a pooled sample from the D samples?
  - The estimated isoform abundance may be biased by outlier samples

MSIQ

Joint Modeling of **M**ultiple RNA-seq **S**amples for Accurate **I**soform **Q**uantification

Summary

- It is necessary to consider the heterogeneity of different samples to make robust isoform quantification

Summary

- It is necessary to consider the heterogeneity of different samples to make robust isoform quantification
- MSIQ is able to identify a consistent group of samples that are most representative of the biological condition

Summary

- It is necessary to consider the heterogeneity of different samples to make robust isoform quantification
- MSIQ is able to identify a consistent group of samples that are most representative of the biological condition
- MSIQ increases the accuracy of isoform quantification by incorporating the information from multiple samples

Summary

- It is necessary to consider the heterogeneity of different samples to make robust isoform quantification
- MSIQ is able to identify a consistent group of samples that are most representative of the biological condition
- MSIQ increases the accuracy of isoform quantification by incorporating the information from multiple samples
- Our proposed hierarchical model is an umbrella framework that are generalizable to incorporate more delicate consideration of read generating mechanisms

Paper and Software

MSIQ: joint modeling of multiple RNA-seq samples for accurate isoform quantification

by Wei Vivian Li, Anqi Zhao, Shihua Zhang, and Jingyi Jessica Li

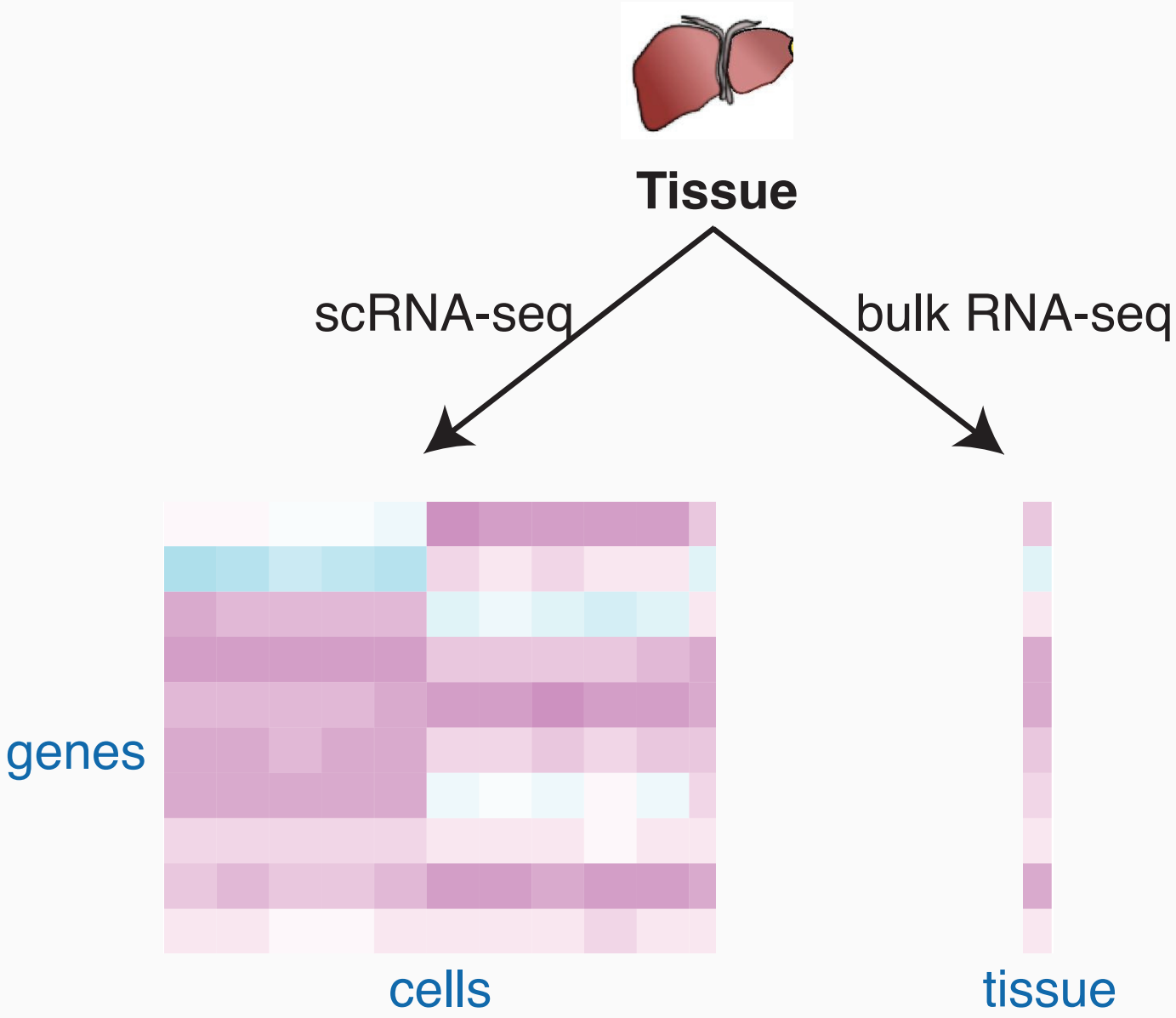
Annals of Applied Statistics 12(1):510–539

R package `MSIQ`

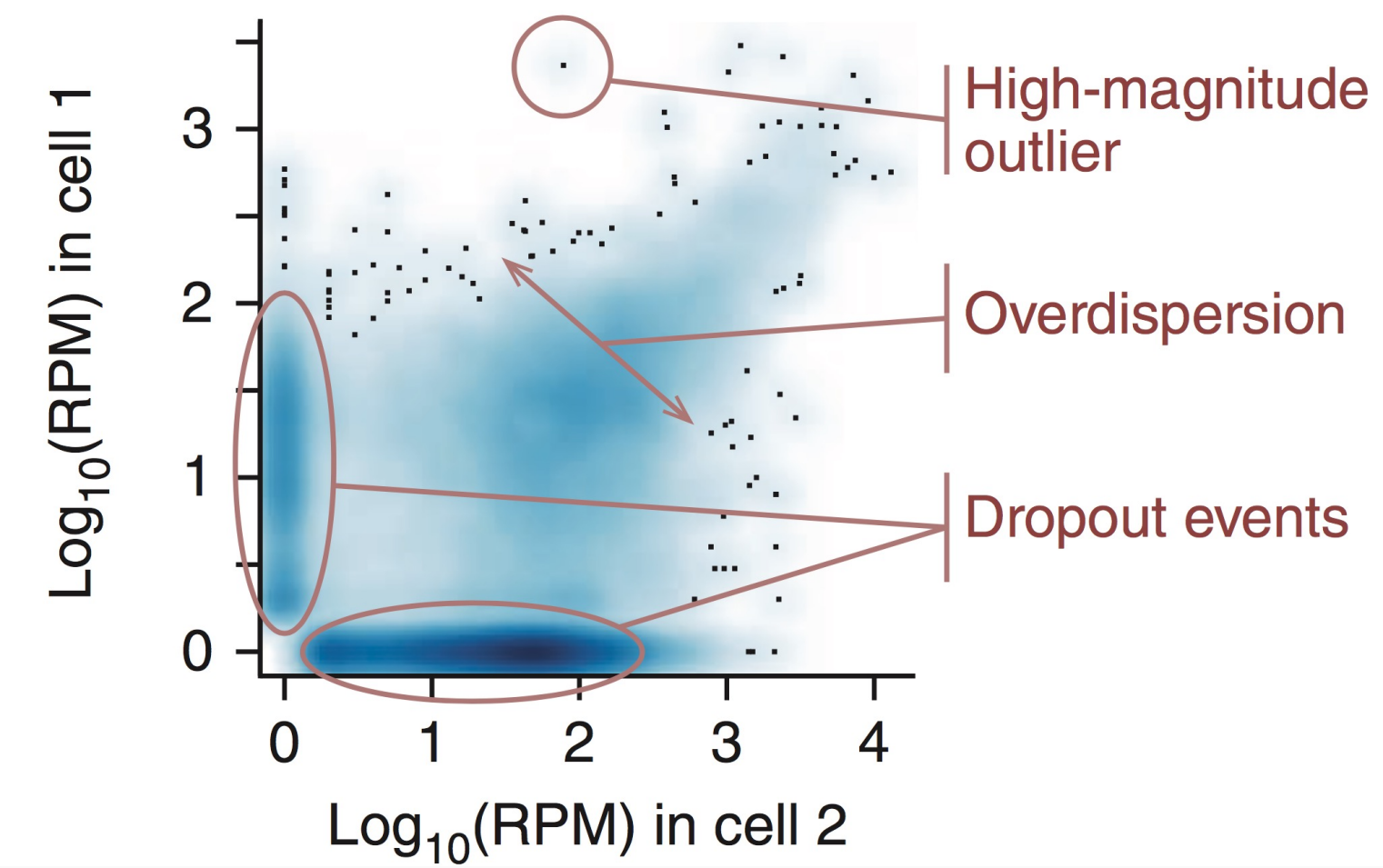
`http://github.com/Vivianstats/MSIQ`

Single-cell RNA-seq: dropout imputation

scRNA-seq vs. bulk RNA-seq at the gene level



Dropout events in scRNA-seq



from [Kharchenko et al., *Nature methods*, 2014]

Dropout events in scRNA-seq

- A dropout event occurs when a transcript is expressed in a cell but is entirely undetected in its mRNA profile

- Dropout events occur due to low amounts of mRNA in individual cells

- The frequency of dropout events depends on scRNA-seq protocols
  - Fluidigm C1 platform: ~ 100 cells, ~ 1 million reads per cell
  - Droplet microfluidics: $\sim 10,000$ cells, ~ 100K reads per cell [Zilionis et al., *Nature Protocols*, 2017]

- **Trade-off**: given the same budget, more cells, more dropouts

Statistical methods for scRNA-seq data analysis

- Clustering / cell type identification
  - **SNN-Cliq** [Xu et al., *Bioinformatics*, 2015]: uses the ranking of genes to construct a graph and learn cell clusters
  - **CIDR** [Lin et al., *Genome Biology*, 2017]: incorporates implicit imputation of dropout values
- Cell relationship reconstruction
  - **Seurat** [Satija et al., *Nature biotechnology*, 2015]: infers the spatial origins of cells from their scRNA-seq data and a spatial reference map of landmark genes, whose expressions are imputed based on highly variable genes
- Dimension reduction
  - **ZIFA** [Pierson et al., *Genome biology*, 2015]: accounts for dropout events based on an empirical observation: dropout rate of a gene depends on its mean expression level in the population

Genome-wide explicit imputation for dropouts

Why do we need genome-wide explicit imputation methods?

Downstream analyses relying on the accuracy of gene expression measurements:

- differential gene expression analysis
- identification of cell-type-specific genes
- reconstruction of differentiation trajectory

It is important to adjust/correct the false zero expression values due to dropouts

Genome-wide imputation methods for scRNA-seq

MAGIC [Dijk et al., *Cell*, 2018]:

- the first method for explicit and genome-wide imputation of scRNA-seq gene expression data
- imputes missing expression values by sharing information across similar cells
- creates a Markov transition matrix, which determines the weights of the cells

SAVER [Huang et al., *Nature Methods*, 2018]:

- borrows information across genes using a Bayesian approach

DrImpute [Kwak et al., *bioRxiv*, 2017]:

- borrows information across cells by averaging multiple imputation results

and several other recent methods available on bioRxiv

Genome-wide imputation methods for scRNA-seq

Limitations of aforementioned methods:

- It is not ideal to impute all gene expressions
  - imputing expressions unaffected by dropout would introduce new bias
  - could also eliminate meaningful biological variation

- It is inappropriate to treat all zero expressions as missing values
  - some zero expressions may reflect true biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

Genome-wide imputation methods for scRNA-seq

Limitations of aforementioned methods:

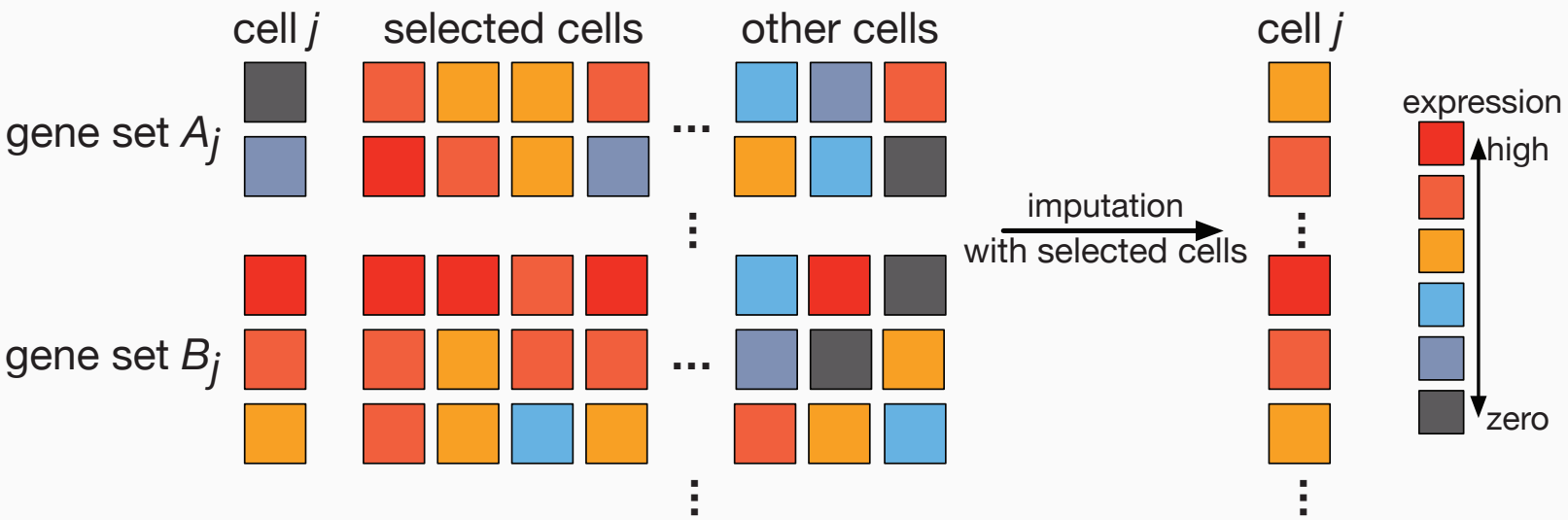
- It is not ideal to impute all gene expressions
  - imputing expressions unaffected by dropout would introduce new bias
  - could also eliminate meaningful biological variation

- It is inappropriate to treat all zero expressions as missing values
  - some zero expressions may reflect true biological non-expression
  - zero expressions can be resulted from gene expression stochasticity

How to determine which values are affected by the dropout events?

Our method: scImpute

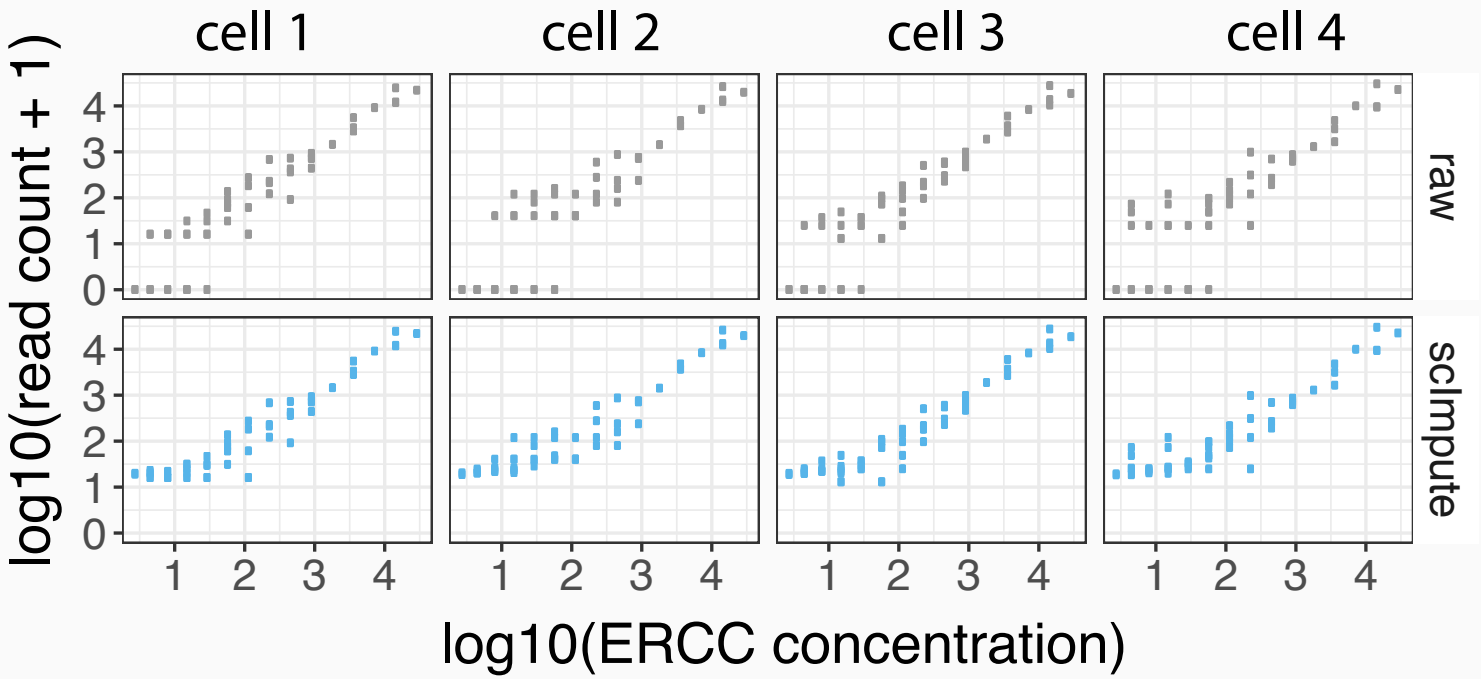
1. For each gene, to determine which expression values are most likely affected by dropout events
2. For each cell, to impute the highly likely dropout values by borrowing information from the same genes' expression in similar cells



Example 1: ERCC spike-ins

scImpute recovers the true expression of the ERCC spike-in transcripts, especially low abundance transcripts that are impacted by dropout events

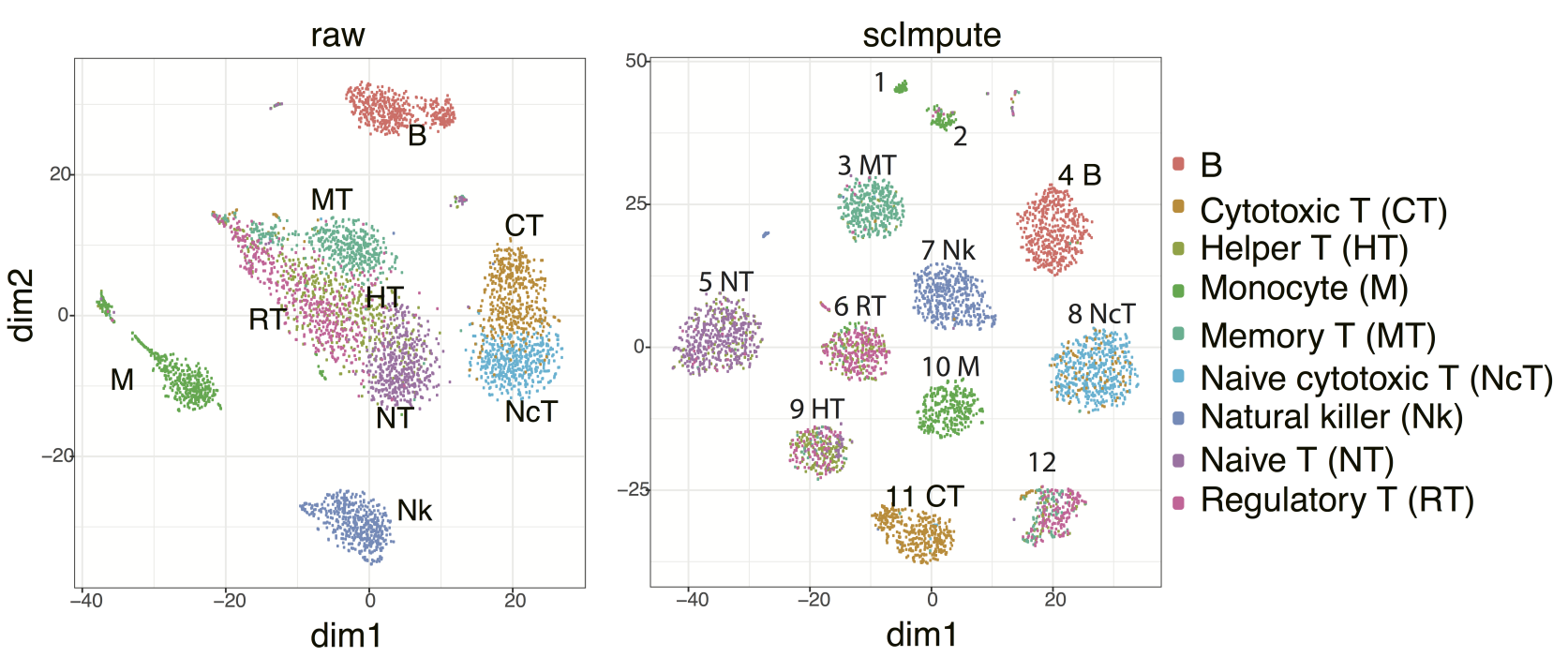
- 3,005 cells from the mouse somatosensory cortex region
- 57 ERCC transcripts



Example 2: cell clustering

4,500 peripheral blood mononuclear cells (PBMCs) from high-throughput droplet-based system 10x genomics [Zheng et al., *Nature communications*, 2017]

Proportion of zero expression is 92.6%



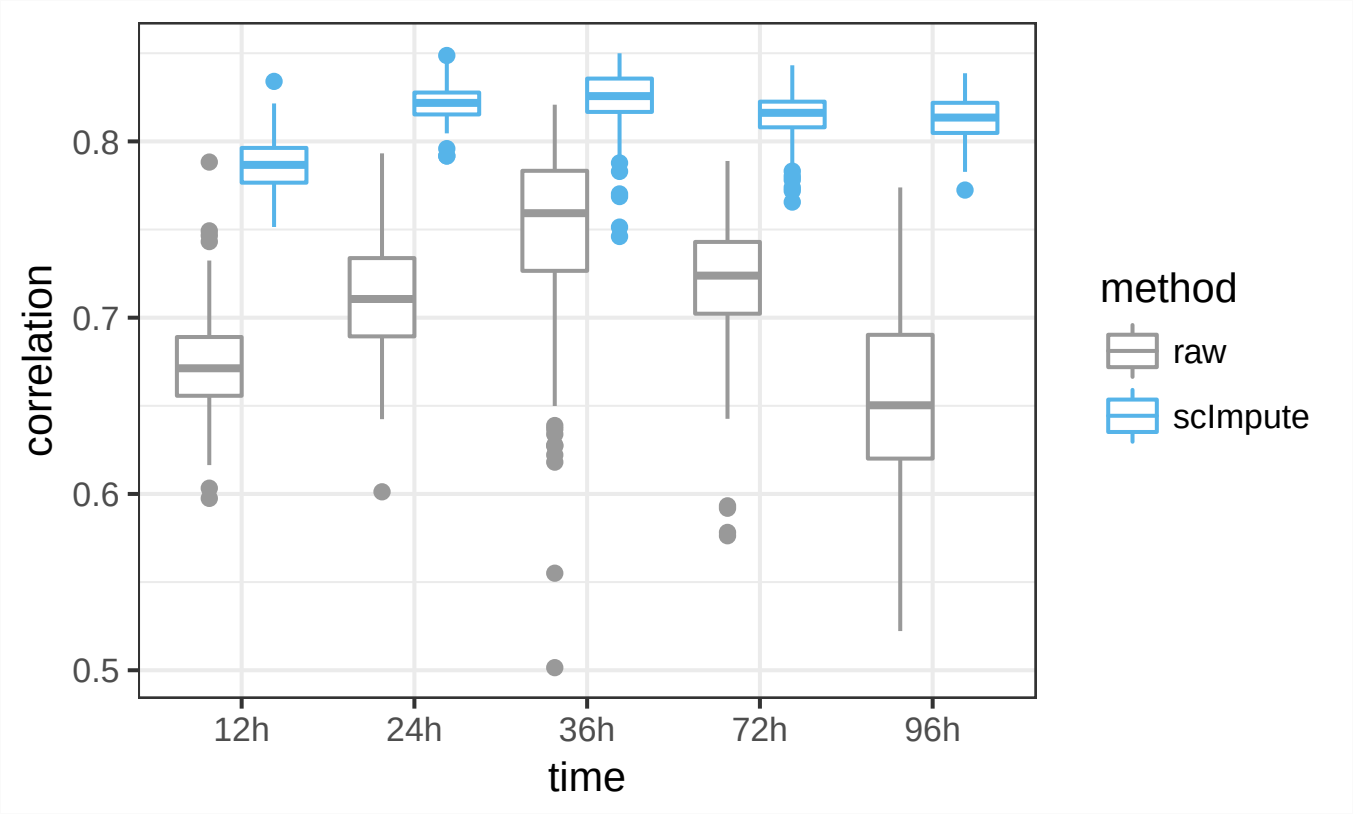
Example 3: gene expression dynamics

Bulk and single-cell time-course RNA-seq data profiled at 0, 12, 24, 36, 72, and 96 h of the differentiation of embryonic stem cells into definitive endorderm cells [Chu et al., *Genome biology*, 2016]

time point	00h	12h	24h	36h	72h	96h	total
scRNA-seq (cells)	92	102	66	172	138	188	758
bulk RNA-seq (replicates)	0	3	3	3	3	3	15

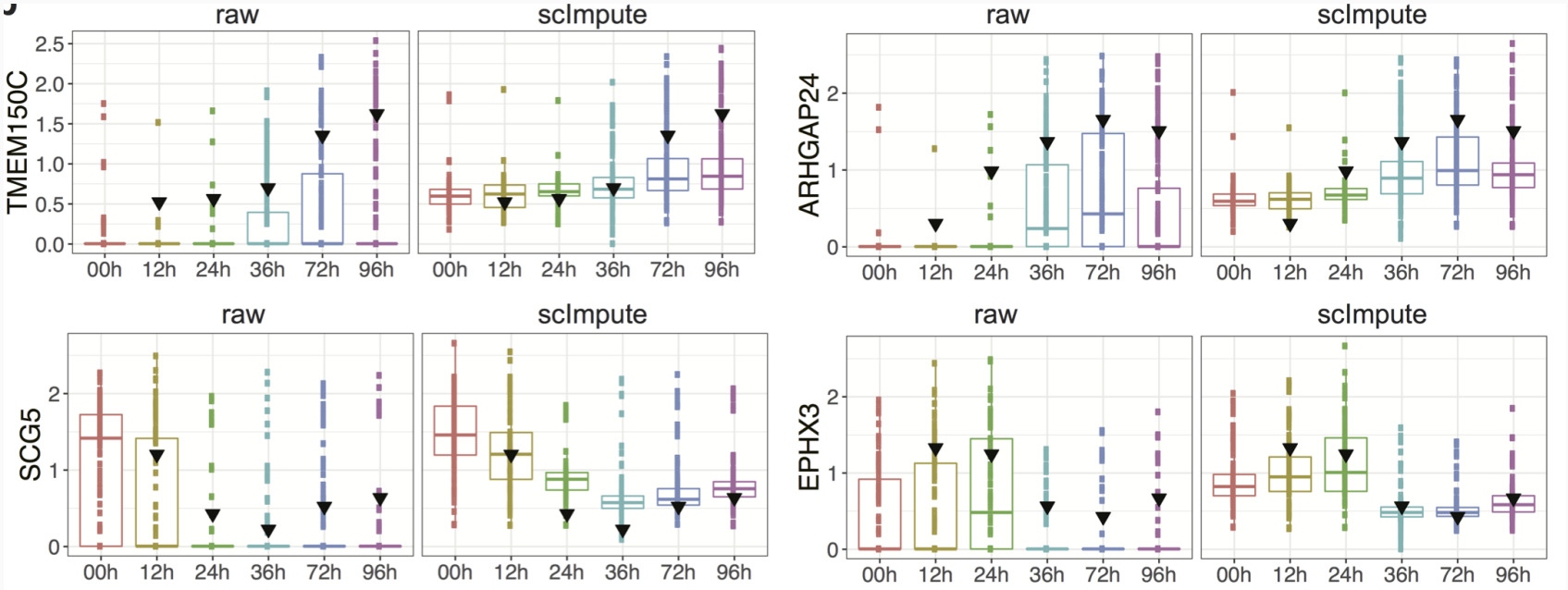
Example 3: gene expression dynamics

Correlation between gene expression in single-cell and bulk data



Example 3: gene expression dynamics

Imputed read counts reflect more accurate gene expression dynamics along the time course



Conclusions

- scImpute is a flexible and easily interpretable statistical method that addresses the dropout events prevalent in scRNA-seq data
- scImpute focuses on imputing the missing expression values of dropout genes, while retaining the expression levels of genes that are largely unaffected by dropout events
- scImpute is compatible with existing pipelines or downstream analysis of scRNA-seq data, such as normalization, differential expression analysis, clustering and classification
- scImpute scales up well when the number of cells increases

Paper and Software

An accurate and robust imputation method scImpute for single-cell RNA-seq data

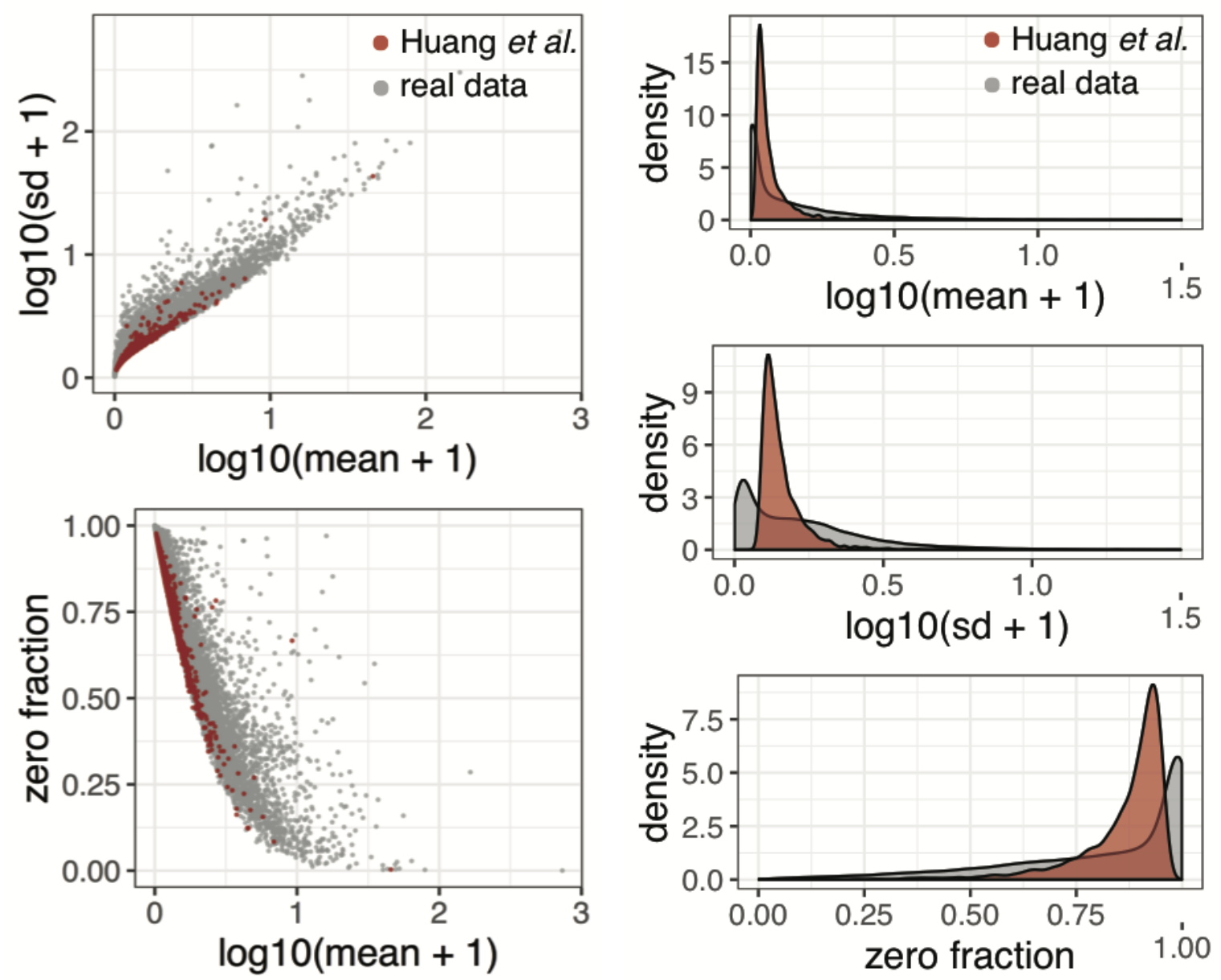
by Wei Vivian Li and Jingyi Jessica Li

Nature Communications 9:997

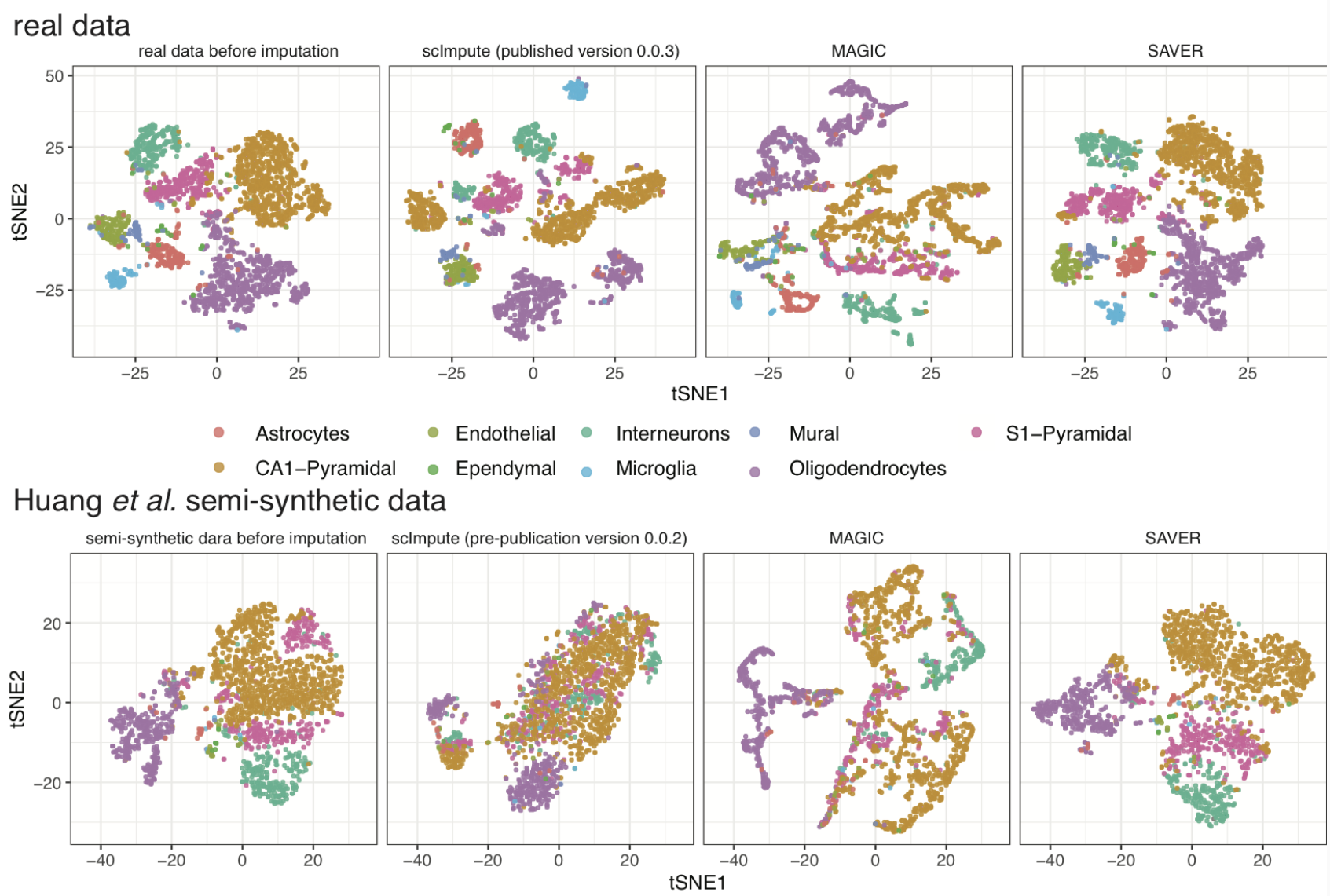
R package `scImpute`

`https://github.com/Vivianstats/scImpute`

Real vs. semi-synthetic data



Real vs. semi-synthetic data



Benchmark standard

		labels used in Huang *et al* .						
		0	**1**	**2**	**3**	**4**	**5**	**6**
labels reported in Zeisel *et al* .	**CA1-Pyramidal**	442	20	289	1	4	42	40
	S1-Pyramidal	2	273	1	1	0	32	11
	Oligodendrocytes	0	0	0	282	0	62	2
	Interneurons	5	7	2	0	220	6	1
	Endothelial	0	0	0	0	1	0	14
	Microglia	0	0	0	0	0	0	6
	Mural	0	1	0	0	0	0	0
	Ependymal	0	0	0	0	0	0	7
	Astrocytes	0	1	0	2	0	1	20

Acknowledgements

Wei Vivian Li (PhD student, UCLA)

Collaborators:

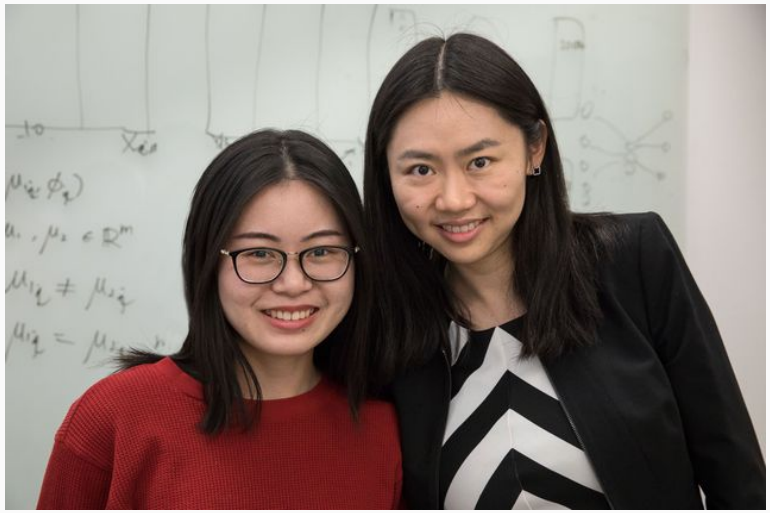
Prof. Alexander Hoffmann (UCLA)

Prof. Hubing Shi (Sichuan University)

Prof. Xin Tong (USC)

Prof. Shihua Zhang (CAS)

Dr. Anqi Zhao (Harvard)



Website: http://jsb.ucla.edu

Email: jli@stat.ucla.edu