Paper Reading

Design and Analysis of Single-Cell Sequencing Experiments

Kuei-Yueh (Clint) Ko

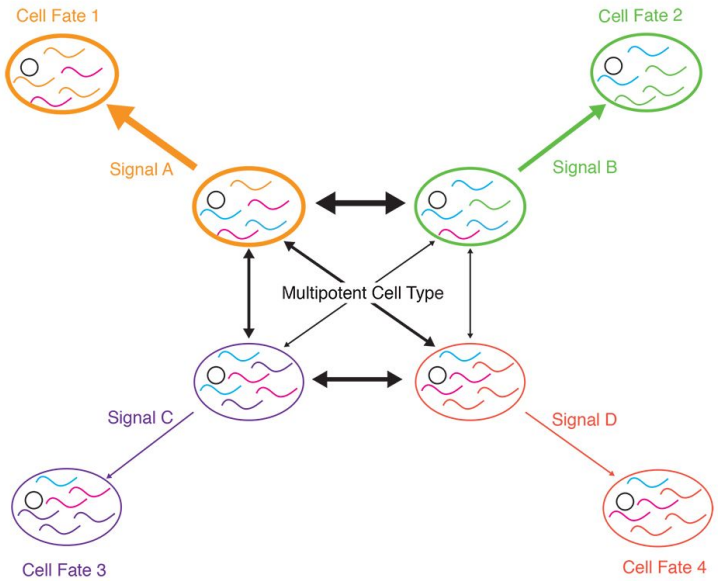
12.06.2017

Authors

Dominic Grun

Research Interests

- Quantitative single cell biology
- Regulation of gene expression during cellular differentiation
- The role of biological gene expression noise during cellular differentiation

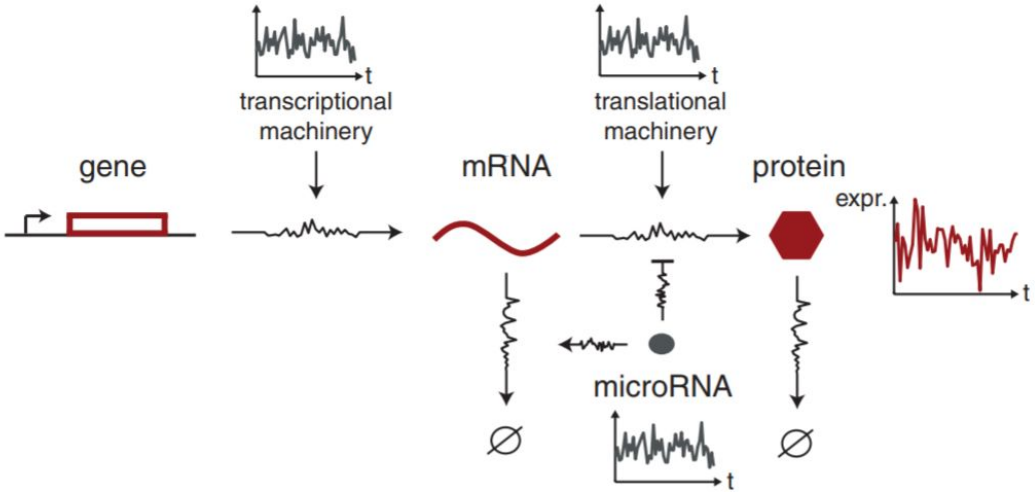


Authors

[Alexander van Oudenaarden]

Research Interests

- Stochastic gene expression
- Developing novel tools to quantify gene expression in single cells
- MicroRNAs



Contents

Contents

Experimental Technique

Data Processing

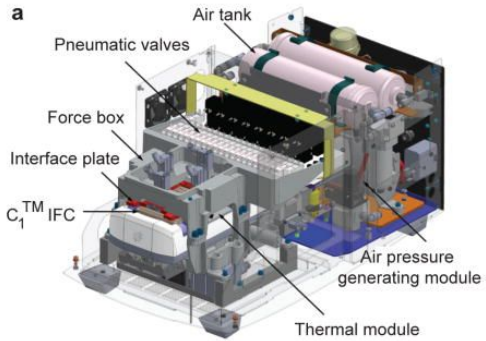
Biological Applications

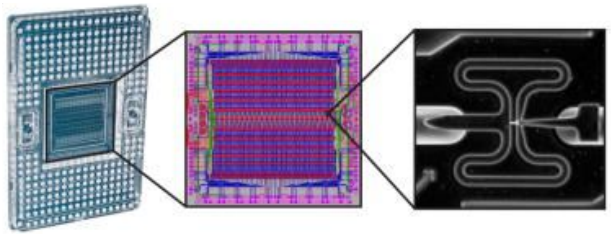
Isolating Single Cells for Sequencing

Isolating Single Cells for Sequencing

- **FACS**
- **Micromanipulation**
  - Glass micropipette
- **Microfluidic devices**
  - Islam et al., 2014
  - Pollen et al., 2014

Fluidigm C1 autoprep system





Average capture: 72 ± 5 single cells per chip

Pollen et al., 2014

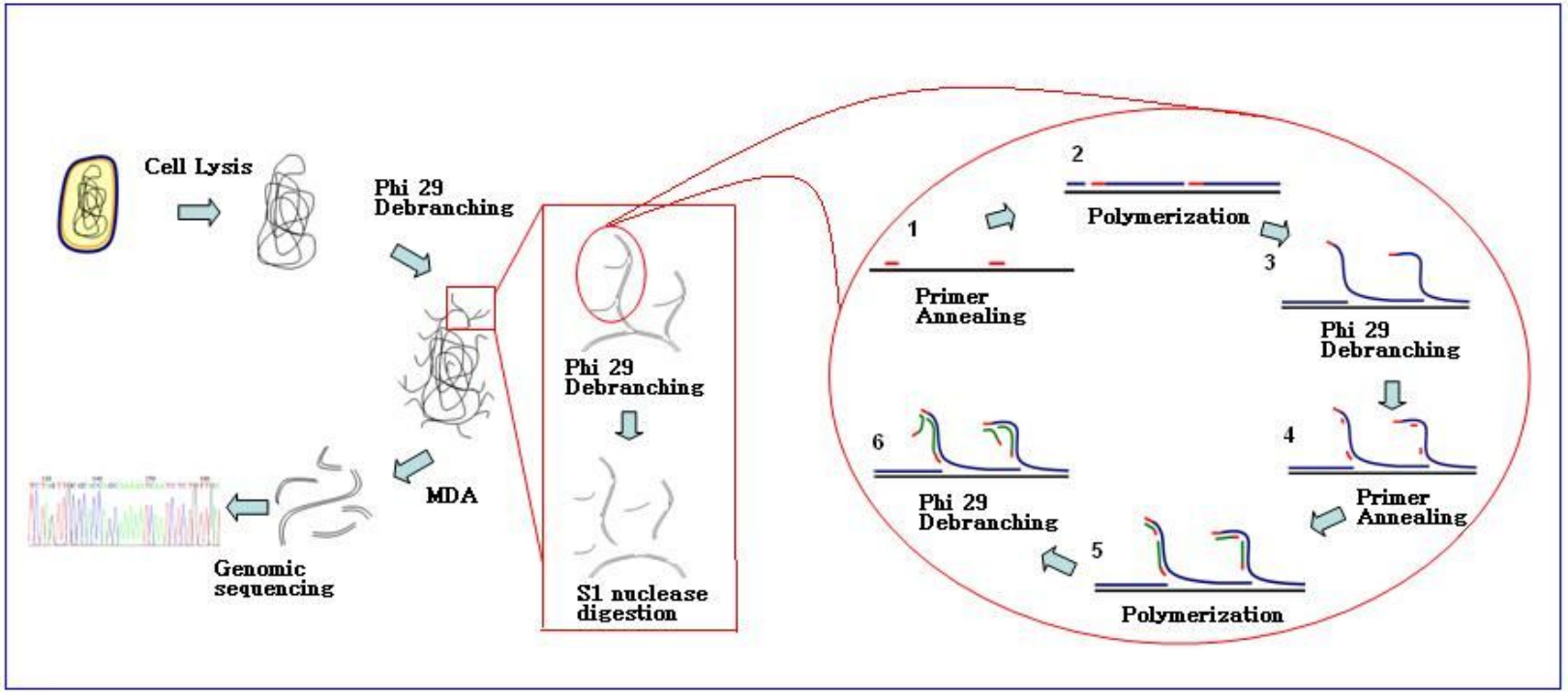
Comparison of Whole-Genome Amplification Techniques

Comparison of Whole-Genome Amplification Techniques

- **Single cell -> extract DNA**
  - If DNA < 10 pg:
    - Perform whole-genome amplification
- **Whole-genome amplification**
  - PCR
  - **Multiple displacement amplification (MDA)**
- **Multiple annealing and looping based amplification cycles (MALBAC)**

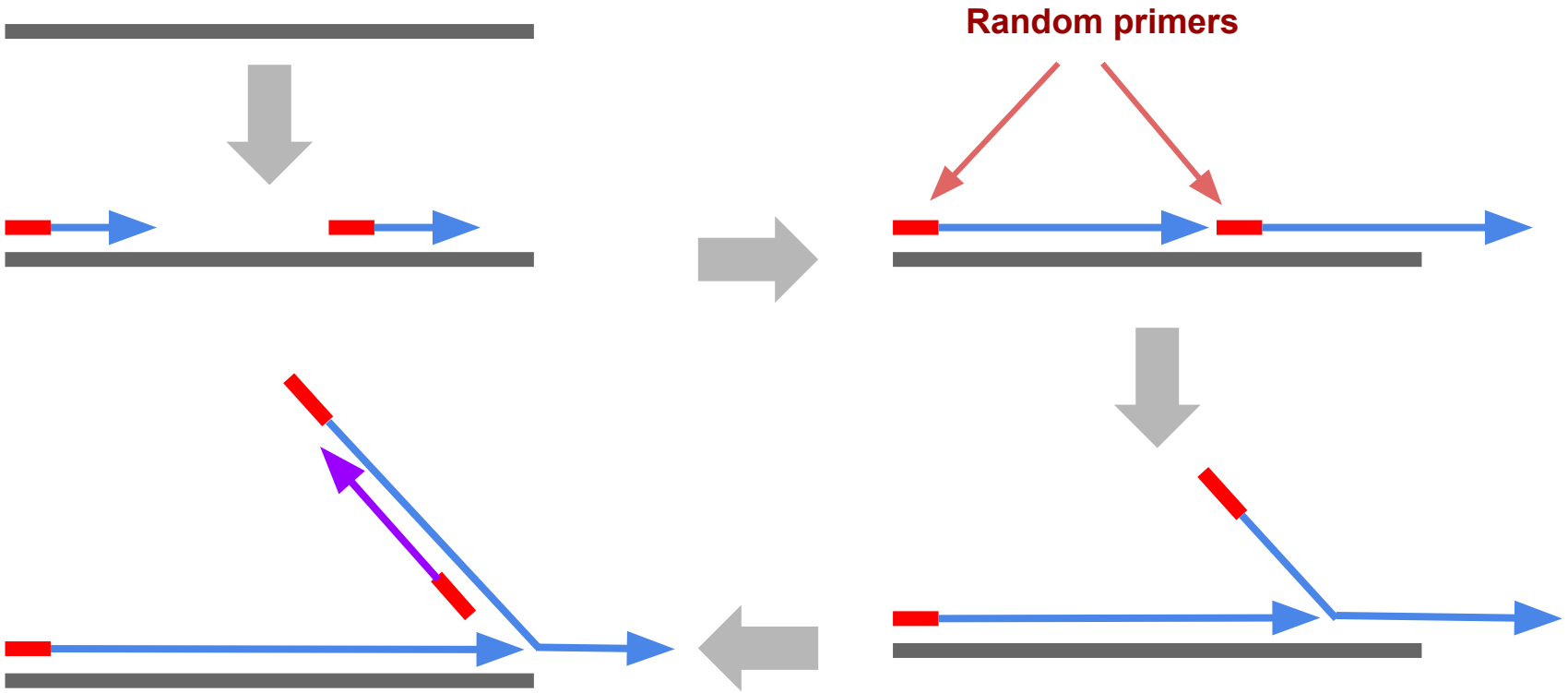
Comparison of Whole-Genome Amplification Techniques

Multiple displacement amplification (MDA)



Comparison of Whole-Genome Amplification Techniques

Multiple displacement amplification (MDA)



Comparison of Whole-Genome Amplification Techniques

PCR vs MDA

	PCR	**MDA**
Polymerase	Taq Polymerase	Phi 29 polymerase
Process	● Denature (94-95℃) ● Anneal (50-56℃) ● Extend (72℃)	Reaction of phi 29 polymerase carried out at 30℃

Products of MDA have lower error rate and larger sizes compared to PCR based Taq amplification.

Comparison of Whole-Genome Amplification Techniques

Allele Dropout

amplification failure of one of the two alleles at a given locus

Blais et al., 2015

Factors of Allele Dropout

“failure of amplification of one of two alleles at any single target locus may be due to either **sequence independent factors** or **allele-specific sequence variations**”

Blais et al., 2015

sequence independent factors or **allele-specific sequence variations**

Comparison of Whole-Genome Amplification Techniques

Factors of Allele Dropout

"failure of amplification of one of two alleles at any single target locus may be due to either **sequence independent factors** or **allele-specific sequence variations**"

Sequence independent factors

- variations in DNA extraction quantity or quality
- presence of PCR inhibitors
- variations in pipetting volumes of reagents or templates
- imprecisions in thermocycler temperatures

allele-specific sequence variations

- DNA Secondary structure (ex: GC rich sequences)

Blais et al., 2015
Frey et al., 2008

Comparison of Whole-Genome Amplification Techniques

Allele Dropout Rate

PCR
MDA (~31%-65%)
MALBAC (~1%)

- Therefore, MALBAC shows higher detection efficiency for SNPs and CNVs

Note: I haven't fully understand the process of MALBAC.

Blais et al., 2015
Frey et al., 2008

Analysis of Single-Cell Genome Sequencing Data

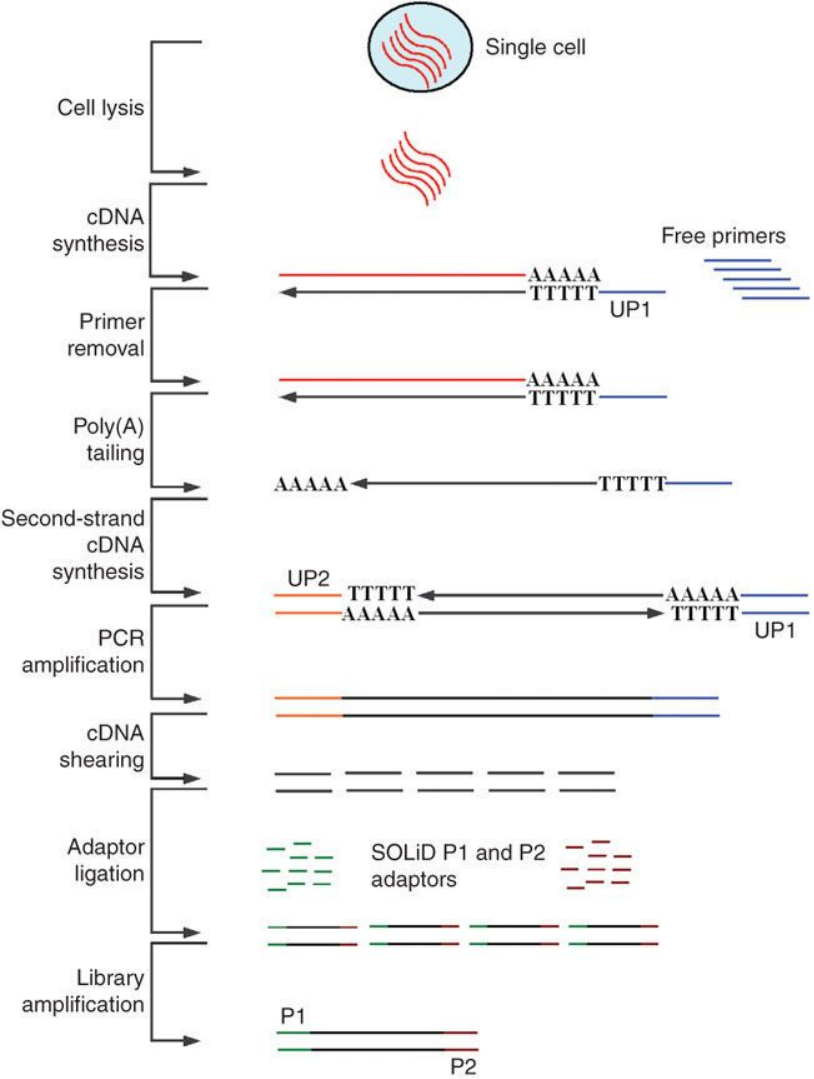
Analysis of Single-Cell Genome Sequencing Data

- First step: Mapping to a ref genome
  - Ex: UCSC genome browser
- Prior to mapping
  - It is advisable to inspect the read quality and trim low-quality bases as well as remaining
  - If the remaining read length is too short, reads should be discarded
- After reading is performed, reads that map to more than a single locus should be discards or counted with reduced uiform weight
- CNVs
- Circular binary alsegmentation algorithm
  - T-stat with a permutation reference distribution to infer p values for break points
  -

Analysis of Single-Cell Genome Sequencing Data

- First step: Mapping to a ref genome
  - Ex: UCSC genome browser
- Prior to mapping
  - It is advisable to inspect the read quality and trim low-quality bases as well as remaining
  - If the remaining read length is too short, reads should be discarded
- After reading is performed, reads that map to more than a single locus should be discards or counted with reduced uiform weight
- CNVs
- Circular binary alsegmentation algorithm
  - T-stat with a permutation reference distribution to infer p values for break points
  -

Comparison of Single-Cell Transcriptome Sequencing Technique

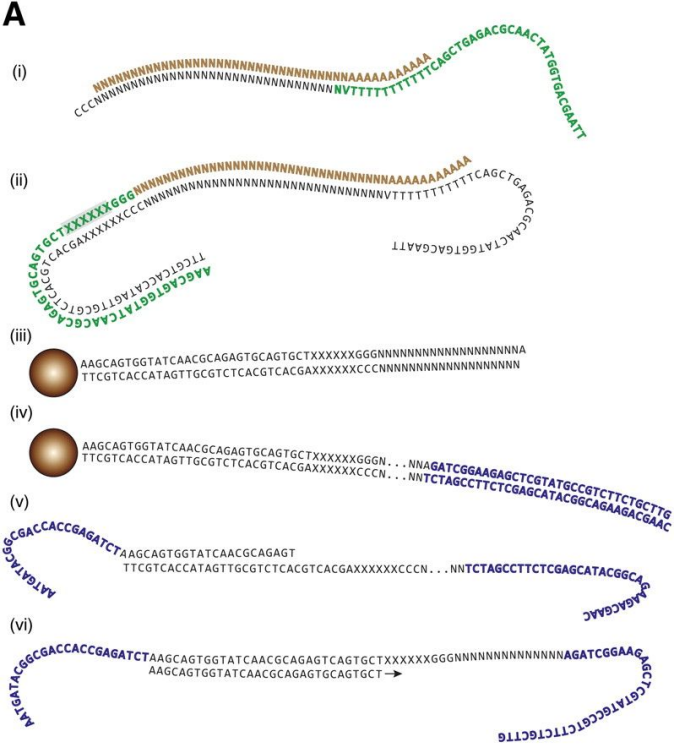


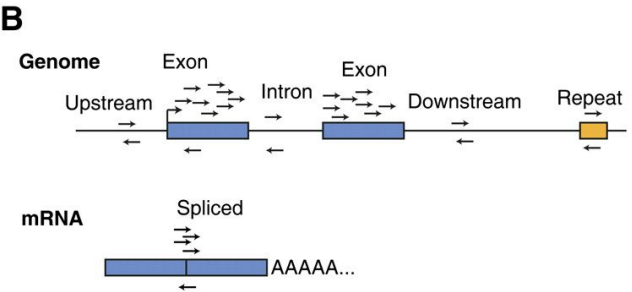
Comparison of Single-Cell Transcriptome Sequencing Technique

The first protocol for single-cell sequencing was published in 2009 by the Surani laboratory (Tang et al., 2009)

Tang et al., 2009

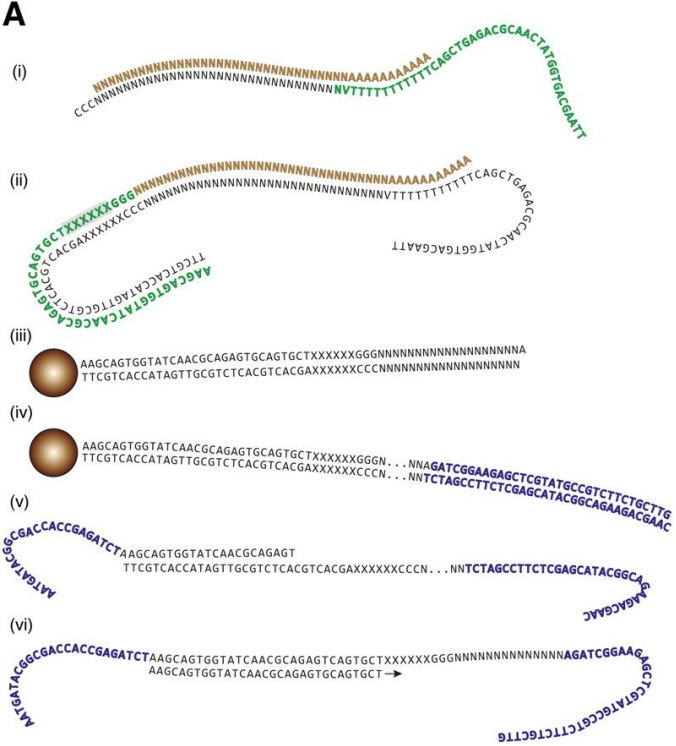
Comparison of Single-Cell Transcriptome Sequencing Technique





Single cell tagged reverse transcription (STRT seq)

Islam et al., 2011

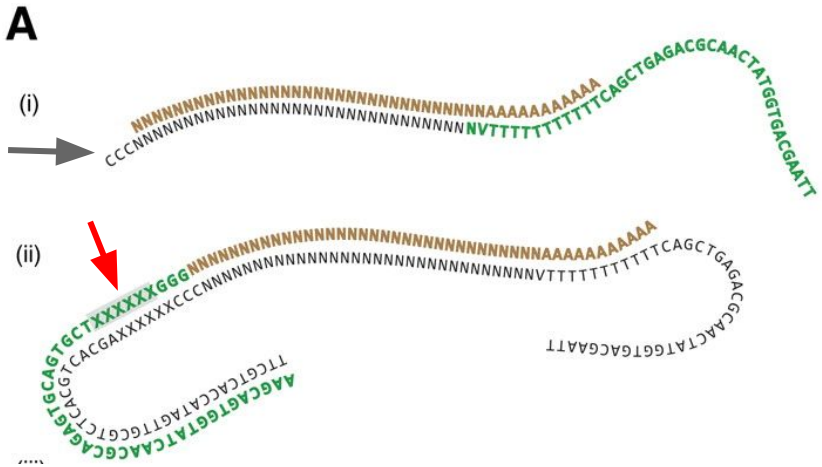


(i) Reverse Transcription

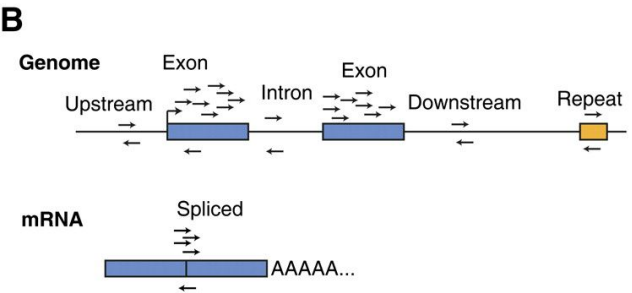
mRNA (Brown)

Tailed oligo-dT primer (green)

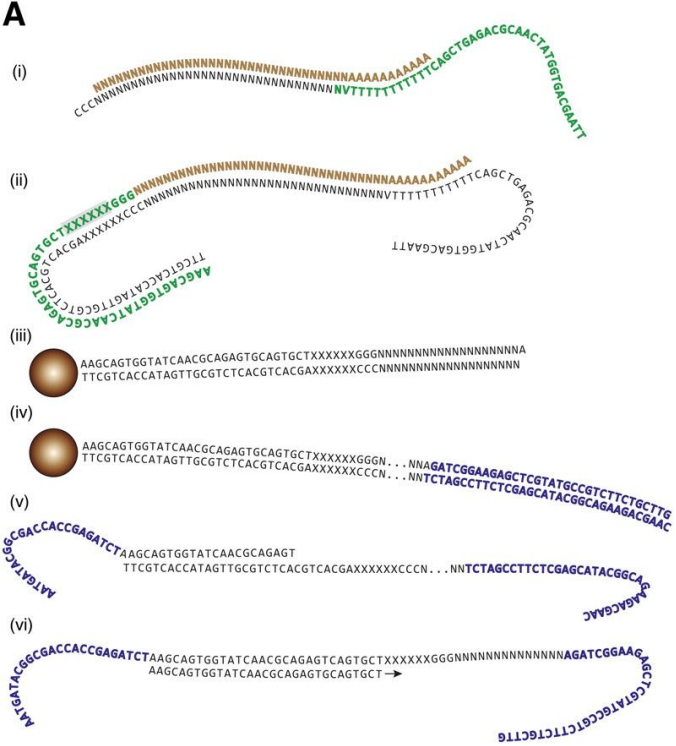
3-6 added cytosines

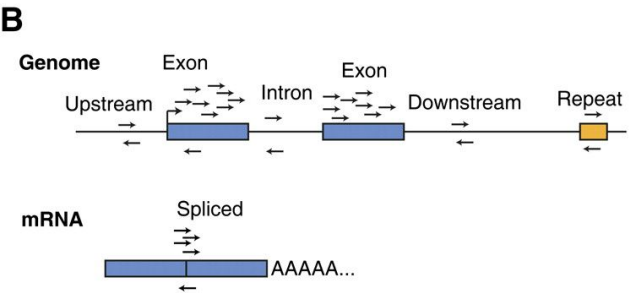


(ii) a helper oligo (green) causes template-switching and thereby introduces a barcode (shaded) and a primer sequence into the cDNA



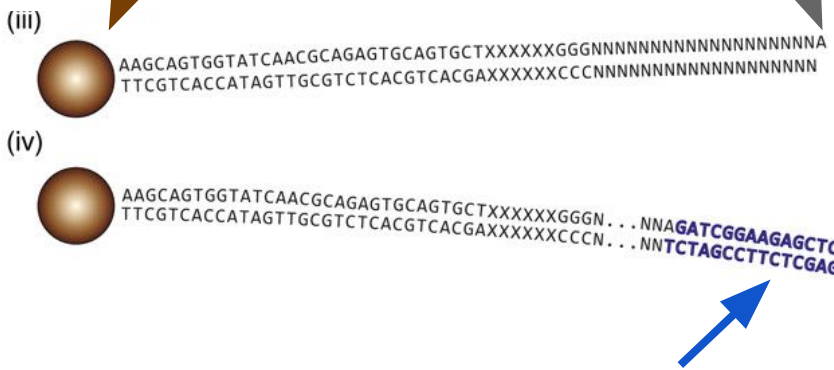
Islam et al., 2011





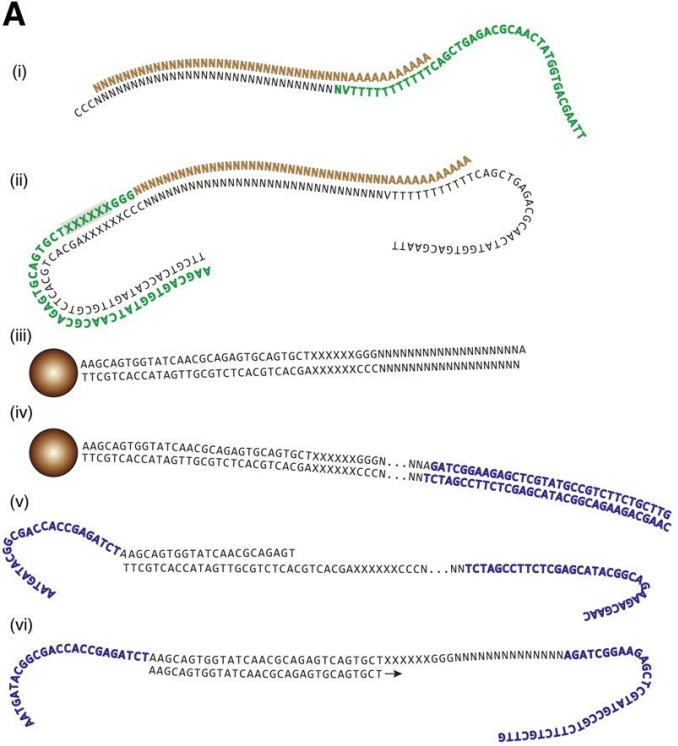
fragmented, and A-tailed

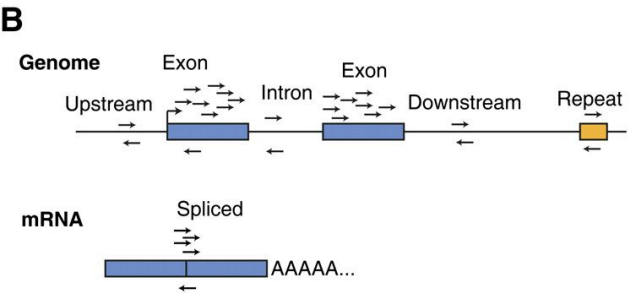
immobilized on beads



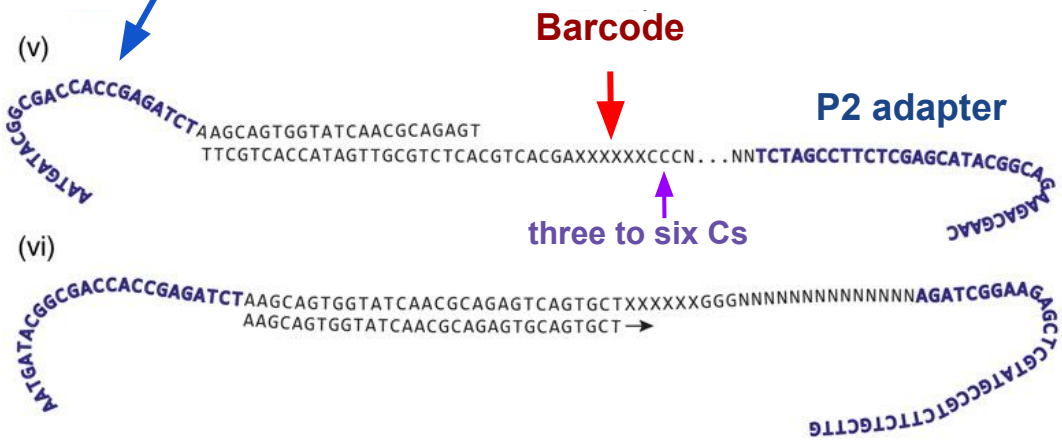
P2 adapter (blue) is ligated to the free end

Islam et al., 2011





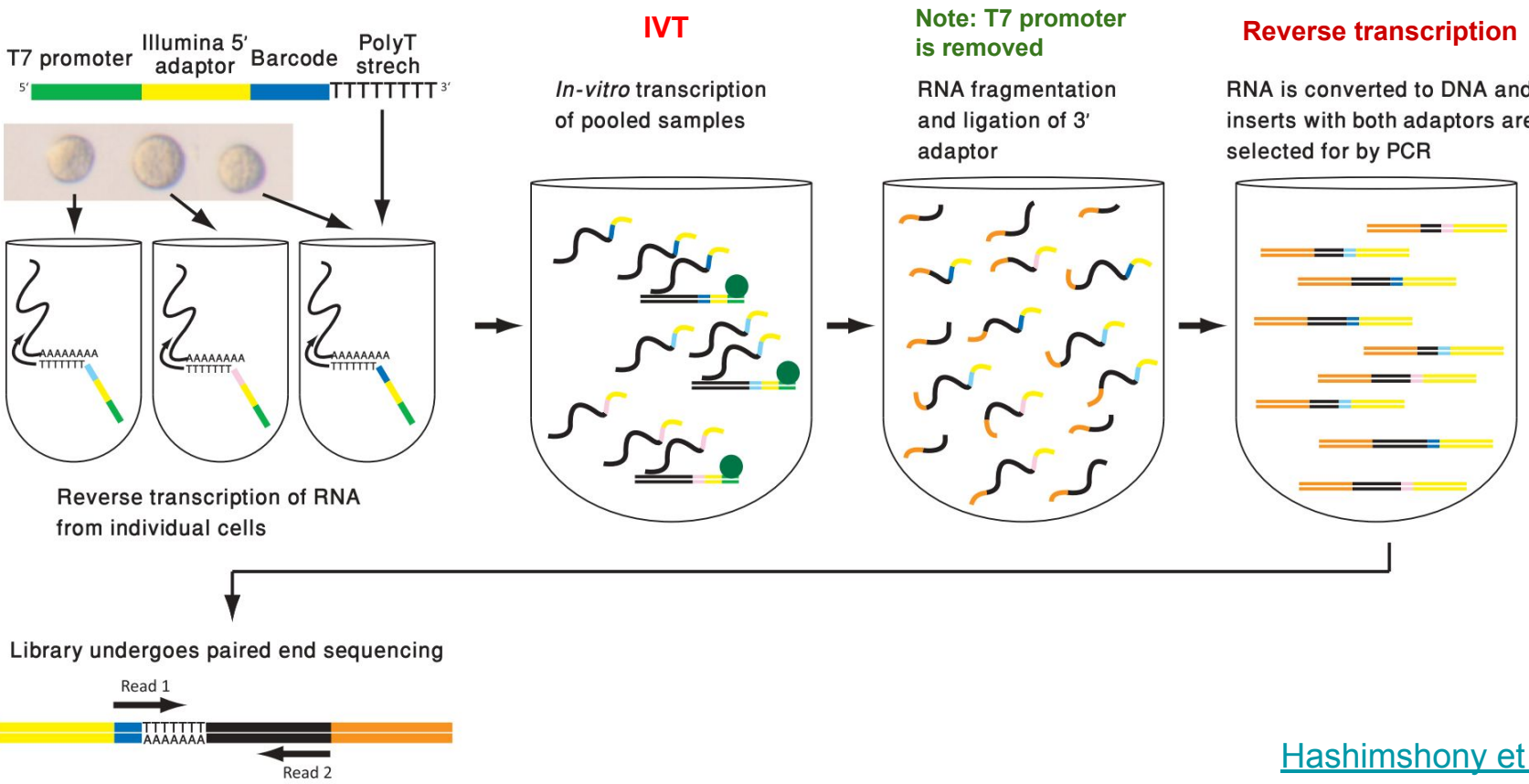
the P1 adapter is introduced in the library PCR step



Final library is sequenced from the P1 side using a custom primer. Each read (arrow) begins by the barcode, followed by three to six Cs, followed by the mRNA insert

Islam et al., 2011

Comparison of Single-Cell Transcriptome Sequencing Technique



Hashimshony et al., 2012

Comparison of Single-Cell Transcriptome Sequencing Technique

- The first protocol for single-cell sequencing was published in 2009 by the Surani laboratory (Tang et al., 2009)
  - Trace the derivation of mouse embryonic stem cells from the inner cell mass with single-cell resolution (Tang et al., 2010)
- Single cell tagged reverse transcription (STRT seq)
  - Islam et al., 2011
  - Template-switching property of the reverse transcriptase to tag the 5’ end of polyadenylated mRNA molecules
- Cell expression by linear amplification and sequencing (CEL-seq)
- Smart-seq and Smart-seq2 methods are a more recent alternative
  - Nextera technology
  - Tn5 transposes simulatneously fragments the cDNA adn ligates seq adaptors to all fragments
- Quartz-seq method

Data Analysis of Single-Cell Transcriptome Data Preprocessing and Read Mapping

Data Analysis of Single-Cell Transcriptome Data Preprocessing and Read Mapping

- Data processing and filtering steps -> reduce the impact of technical noise
- First analysis step: quality filtering or trimming of the sequencing reads prior to mapping the reads to a reference database
  - Standard tools
    - Fastqc
    - Standard mapping tools: bwa
      - Trimming of low-quality bases from the end of the reads
    - Mapping
      - Garber et al., 2011
- Due to the low read coverage of the gene body in single-cell sequencing experiments, isoform quantification with standard methods such as Cufflinks can be problematic
  - If isoform info is not essential for the study

Data Analysis of Single-Cell Transcriptome Data Preprocessing and Read Mapping

Expression Quantification and Filtering

Expression Quantification and Filtering

Barcode (UMI)

Transcripts per one million reads (TPM)

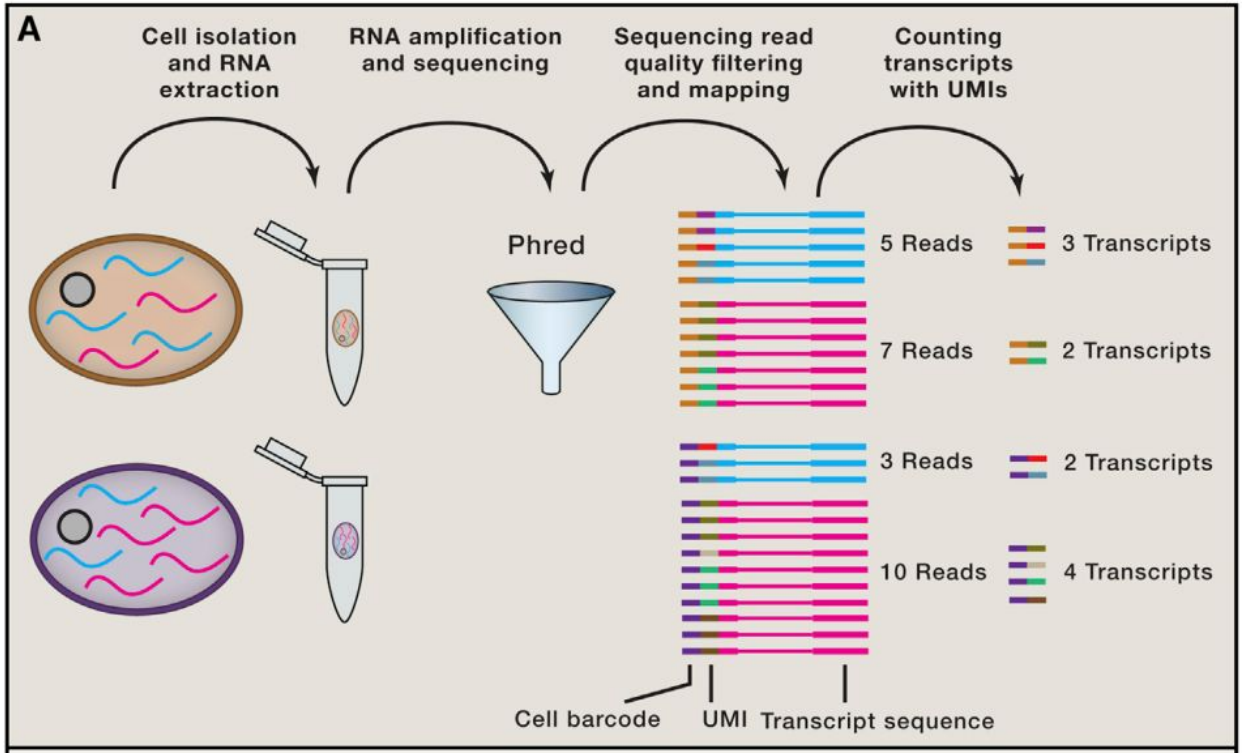
RPKM

Spike-in

Expression Quantification and Filtering

Sequenced cell barcode

unique molecular identifiers



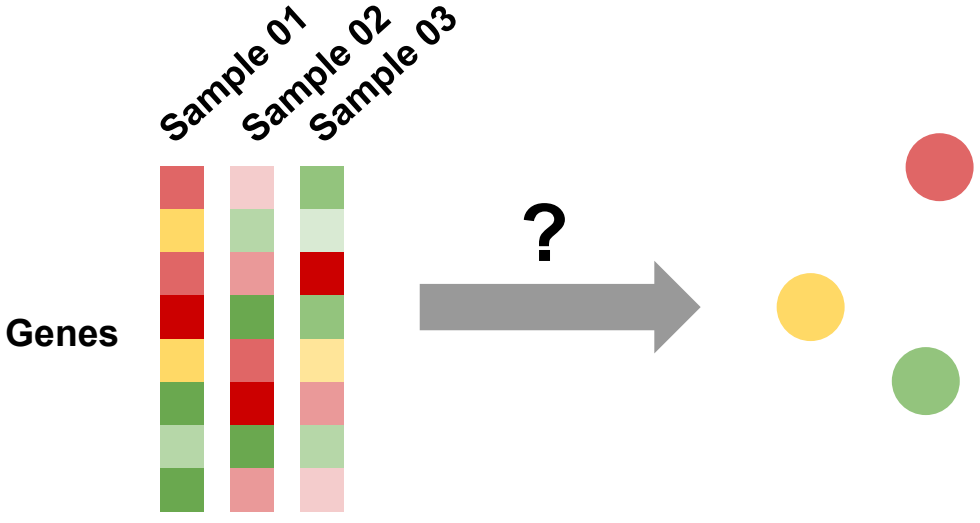
Data Normalization

Data Normalization

Biological Insights from Single-Cell Transcriptome Data Identification of Cell types

Biological Insights from Single-Cell Transcriptome Data Identification of Cell types

- Most important application of single-cell mRNA sequencing
  - **Identification of cell types in a complex mixture**



Biological Insights from Single-Cell Transcriptome Data Identification of Cell types

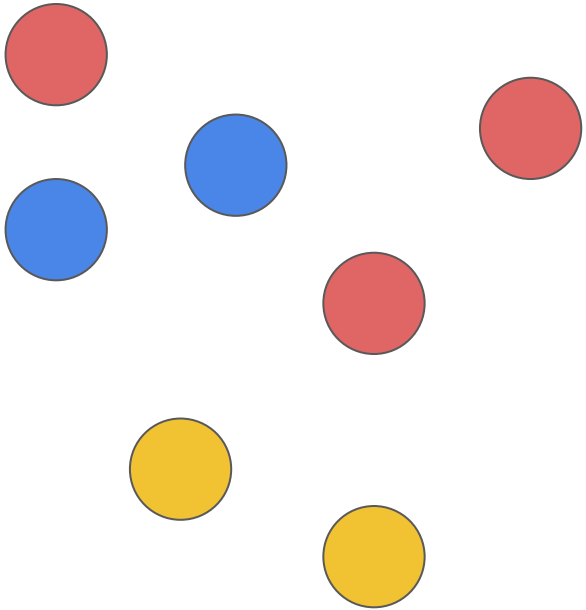
- Examples
  - **Spleen** (Jaitin et al., 2014)
  - **Lung Epithelium** (Treutlein et al., 2014)
  - **Retina** (Macosko et al., 2015)
  - **Mouse hippocampus (also uncovered novel cell types)** (Zeisel et al., 2015)

Biological Insights from Single-Cell Transcriptome Data Identification of Cell types

- One general problem
  - Confounding factors
    - technical variability
    - biological variability
      - Ex: cell-to-cell differences in the cell cycle phase
  - **=> batch effects**

Identification of Marker Genes

Identification of Marker Genes



Marker genes

- Cell surface markers
- Fluorescent reporter genes

Question: How to identify marker genes?

- Identification of differentially expressed genes
  - DESeq

Identification of Marker Genes

DESeq

Assumption

- Most genes are not differentially expressed.

Probabilistic Model

- Negative Binomial Distribution

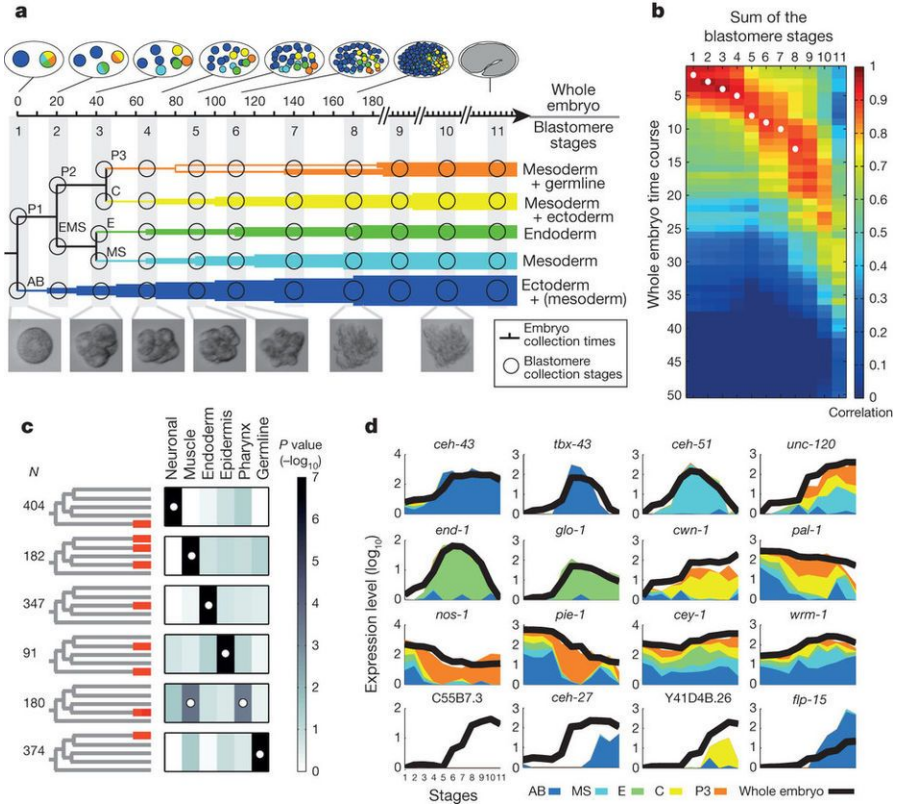
$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2),$$

the number of reads in sample j that are assigned to gene i can be modeled by a negative binomial (NB) distribution

Anders & Huber, 2010

Inference of Differentiation Dynamics

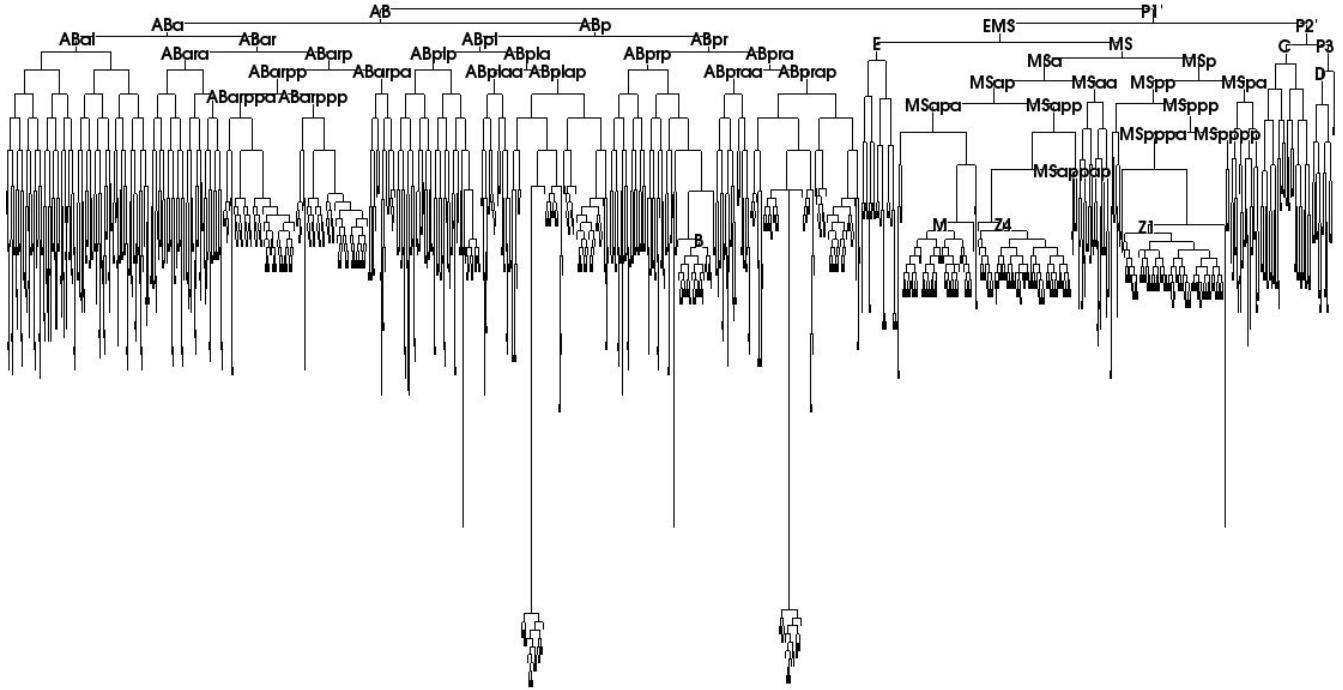
Inference of Differentiation Dynamics



[Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer]

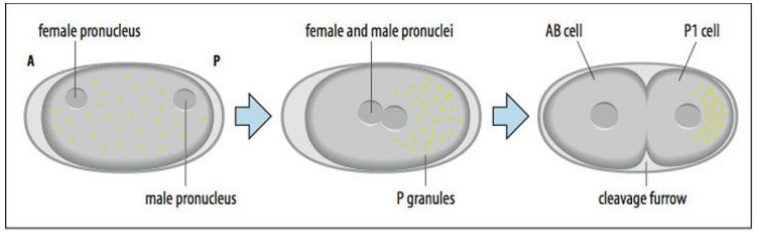
Inference of Differentiation Dynamics

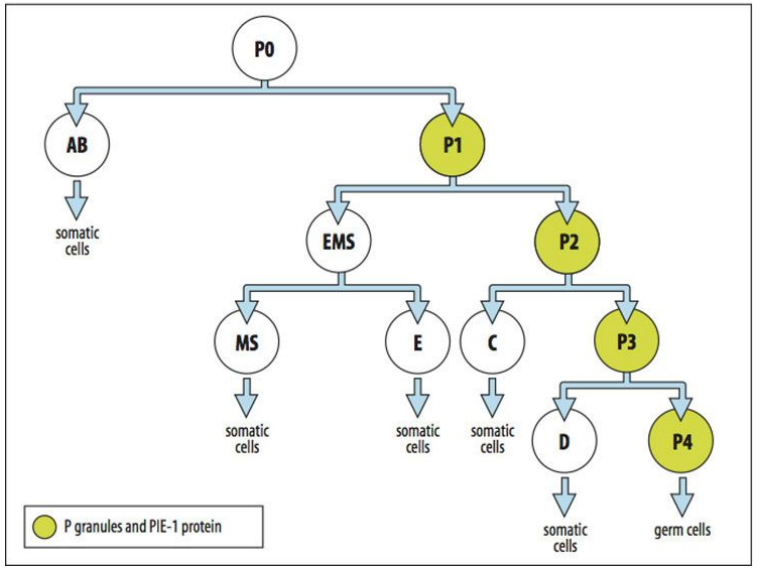
Example: Cell differentiation of *C. elegans*



Inference of Differentiation Dynamics

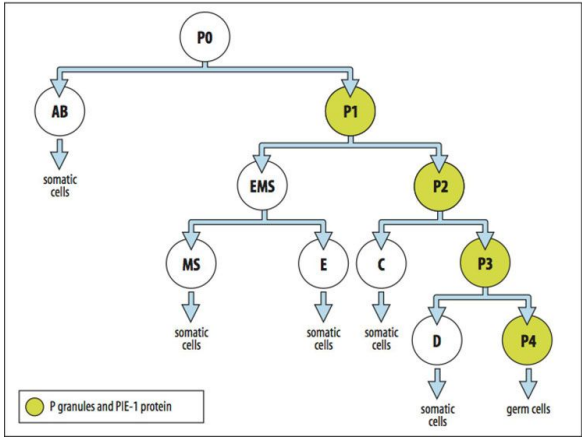
Example: Cell differentiation of *C. elegans*

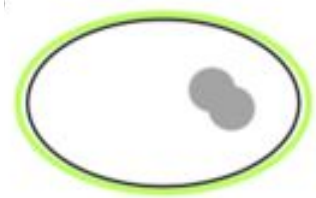




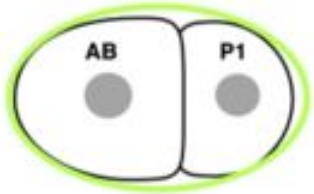
Inference of Differentiation Dynamics

Example: Cell differentiation of *C. elegans*

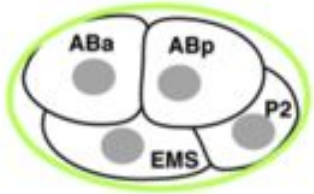




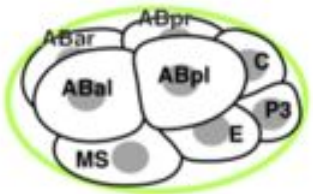
N = 1



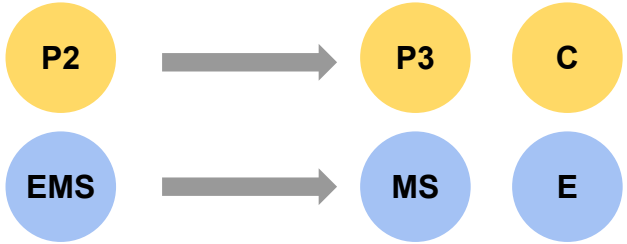
N = 2



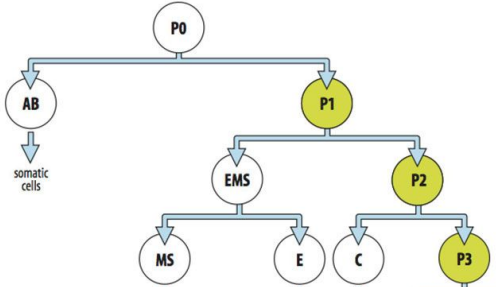
N = 4

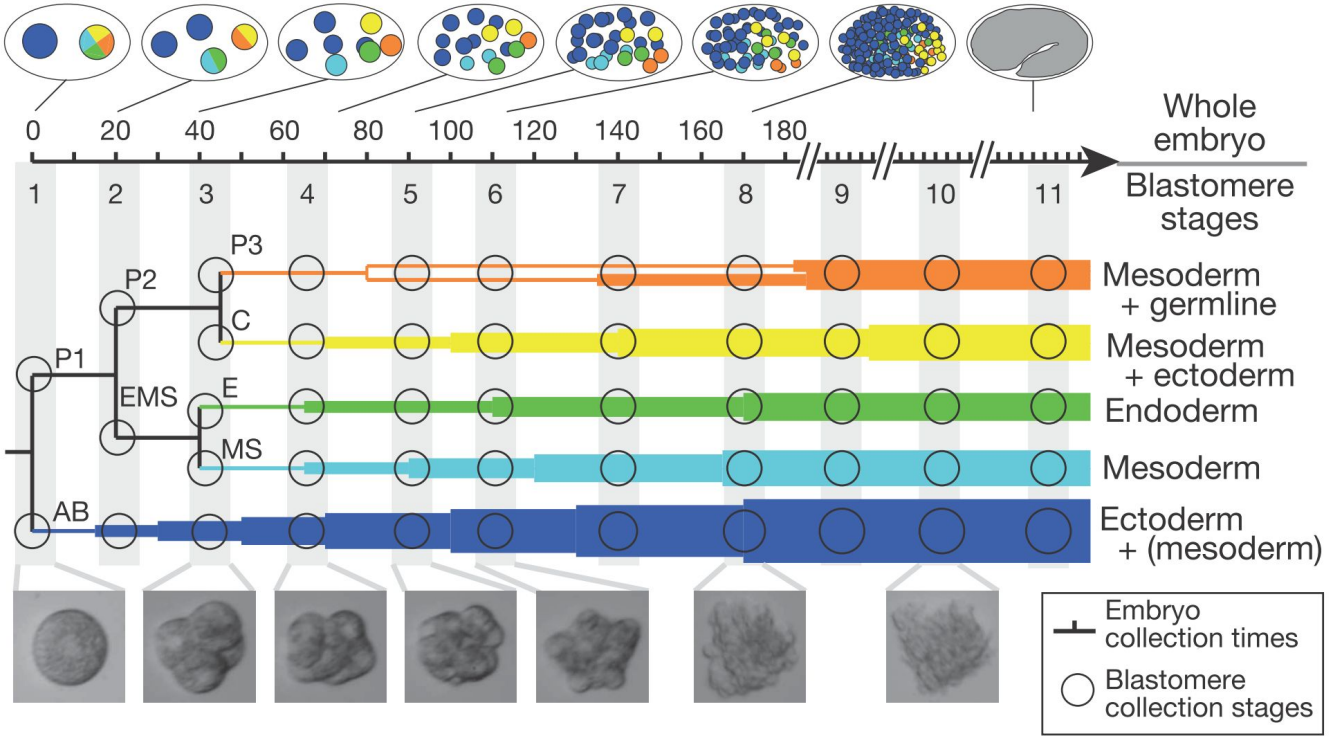


N = 8

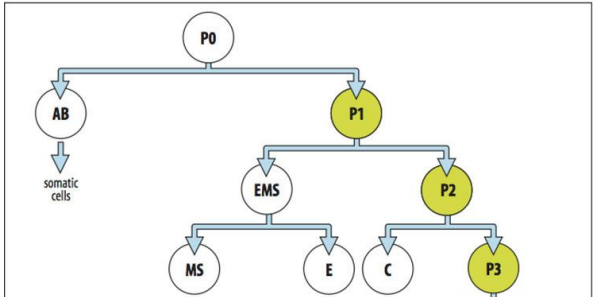


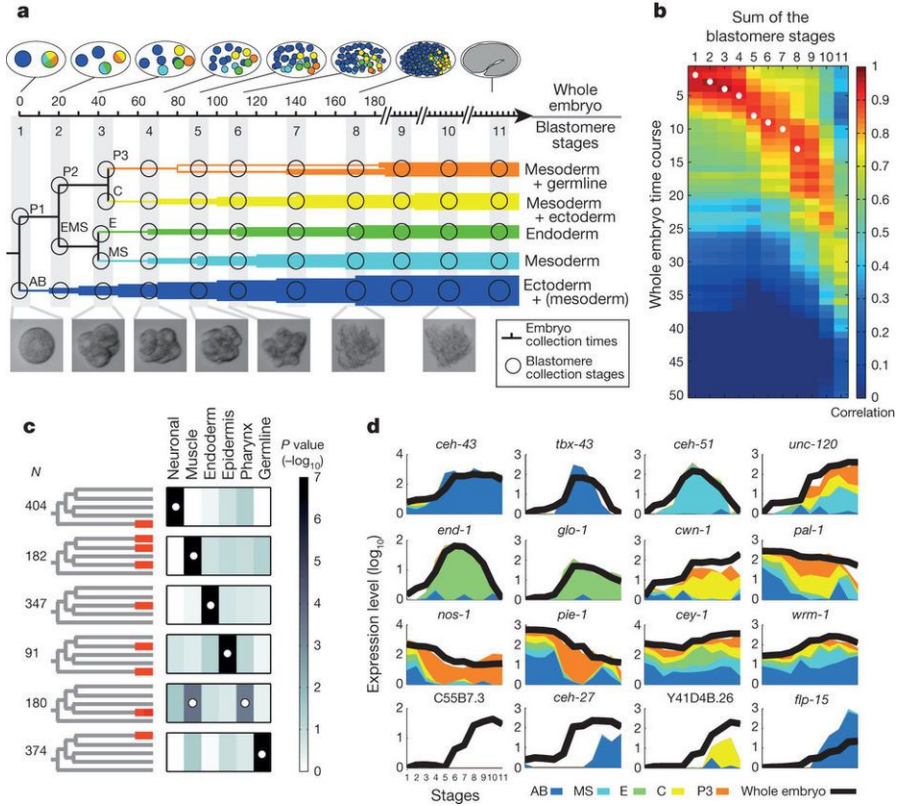
Inference of Differentiation Dynamics





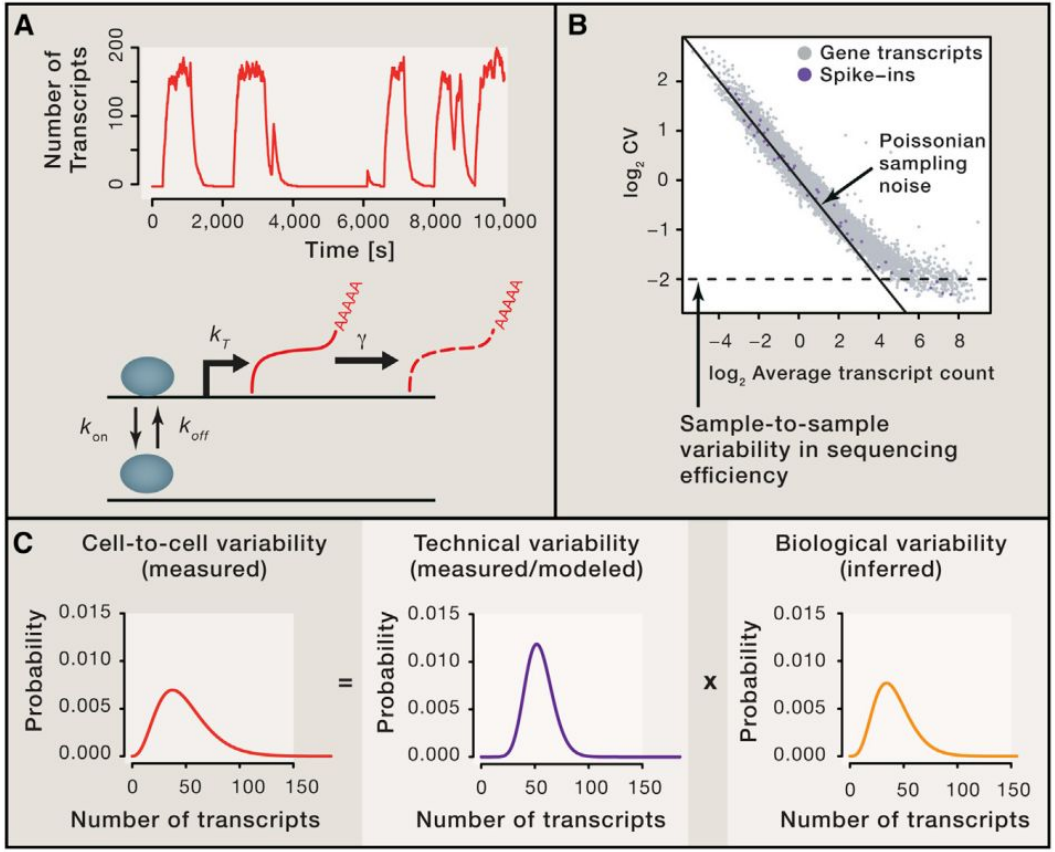
Inference of Differentiation Dynamics





Measuring Gene Expression Noise

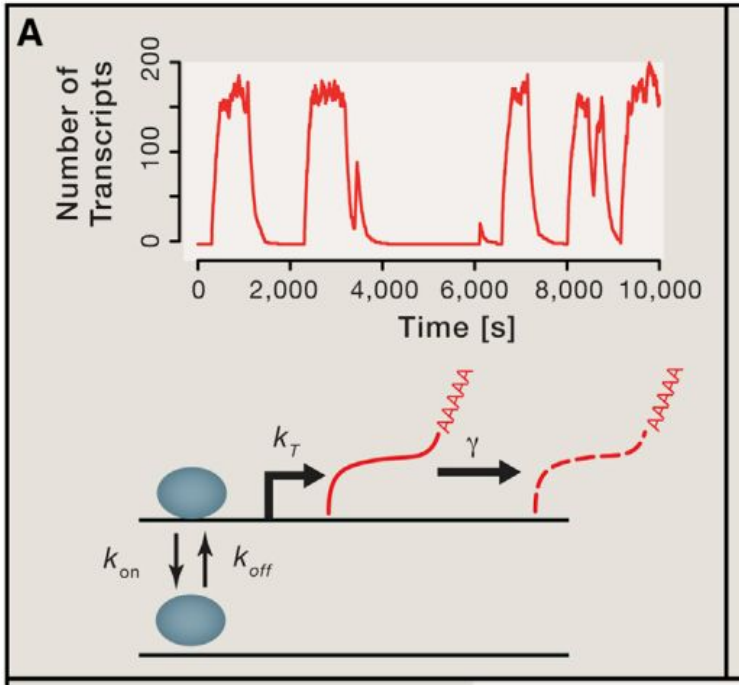
Measuring Gene Expression Noise

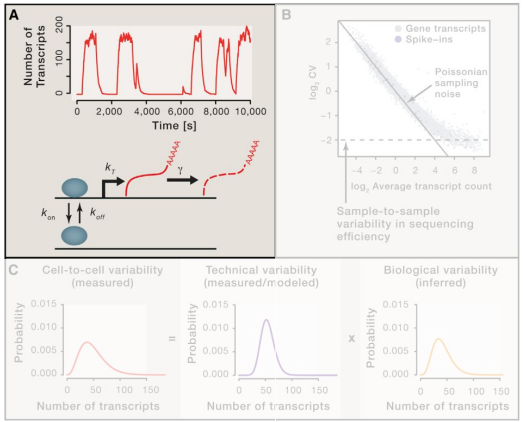


Measuring Gene Expression Noise

- **Gene Activation**
- Sequencing Noise
- Noise Components

- **Gene Activation**





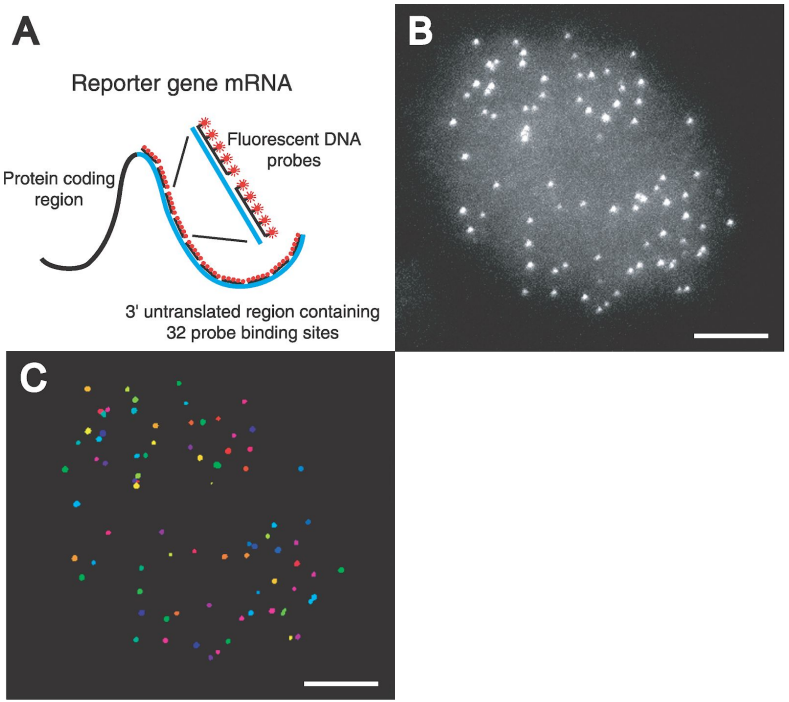
Measuring Gene Expression Noise

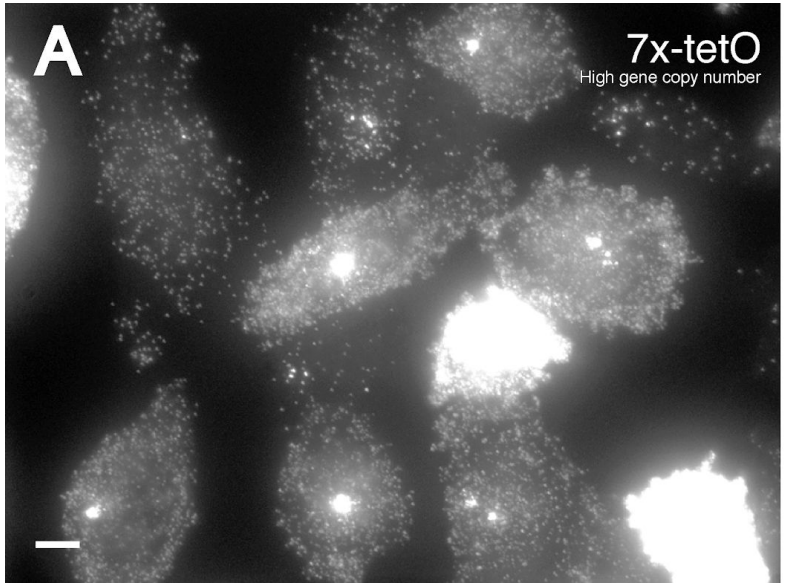
- **Gene Activation**

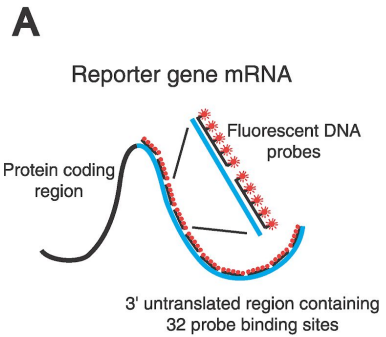
Paper
[Stochastic mRNA Synthesis in Mammalian Cells]

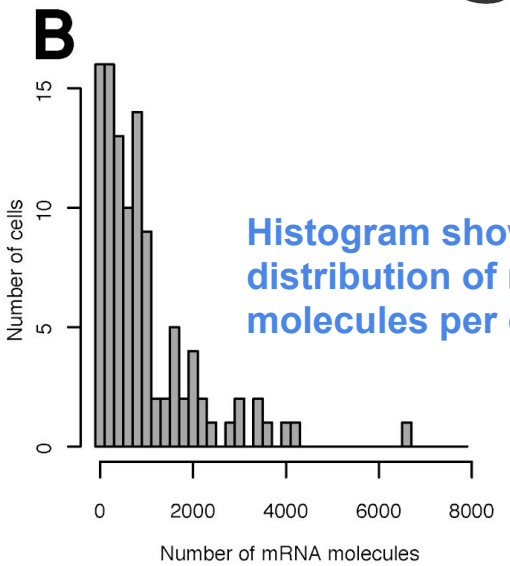
Aim
cell-to-cell variation in gene expression in mammalian cells by accurately counting single molecules of mRNA through the use of fluorescence in situ hybridization (FISH)

Material
Chinese Hamster Ovary (CHO) Cells
- high growth rate
- high protein productivity







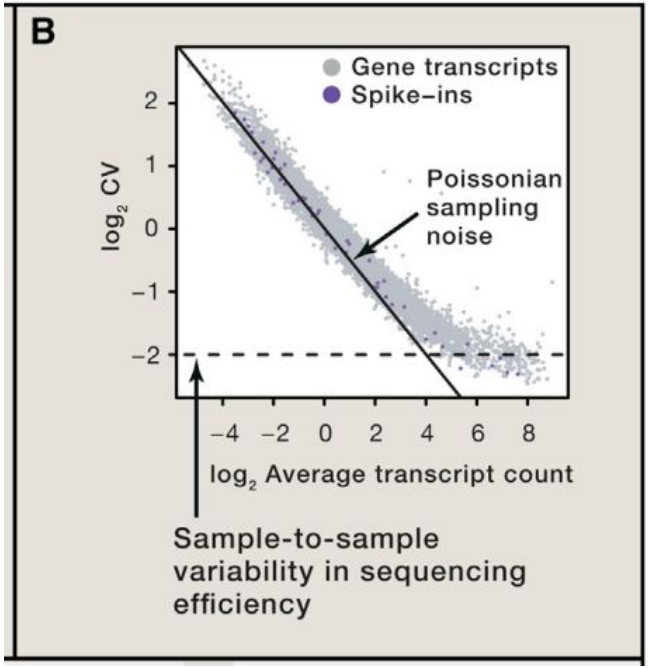


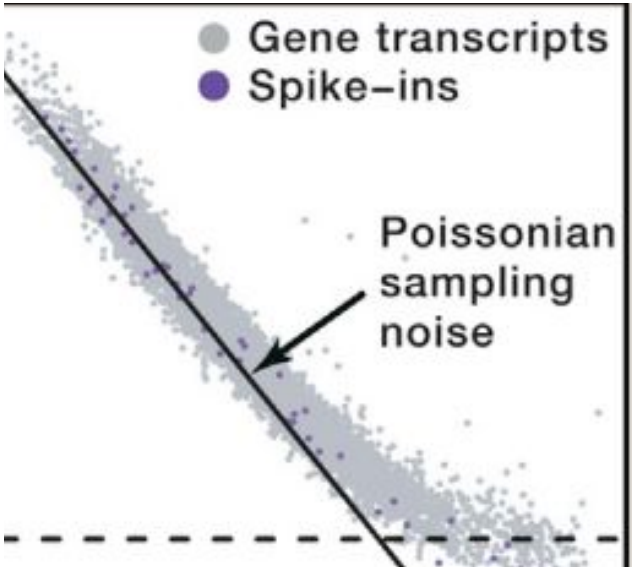
Histogram showing the distribution of mRNA molecules per cell

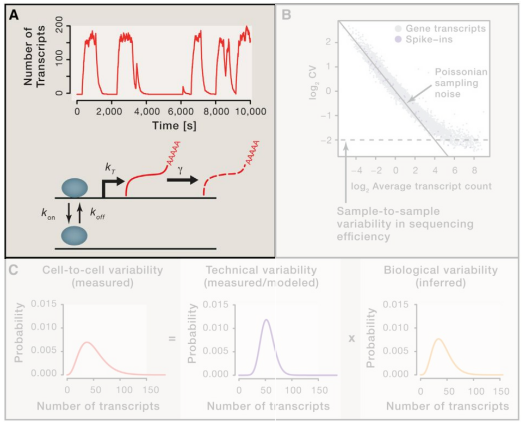
- Gene Activation
- **Sequencing Noise**
- Noise Components

Measuring Gene Expression Noise

- **Sequencing Noise**



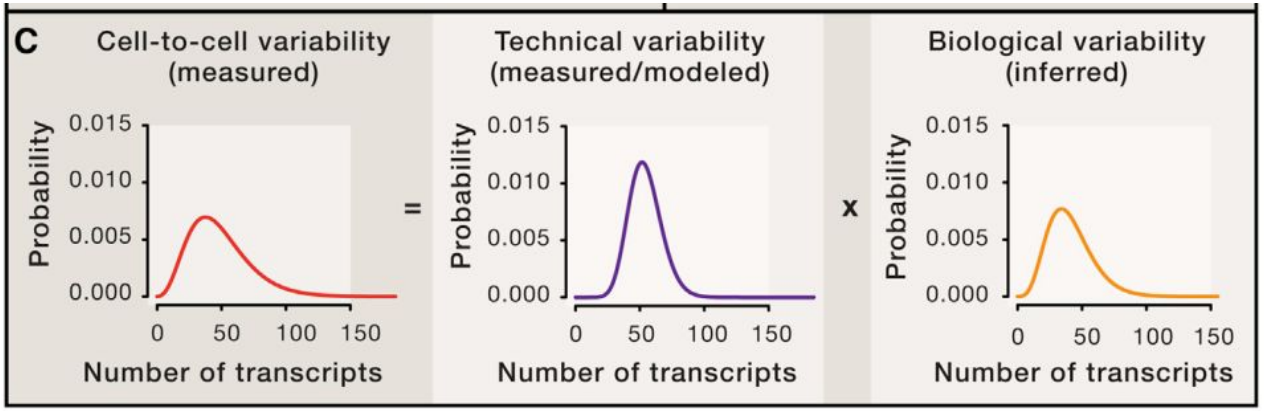


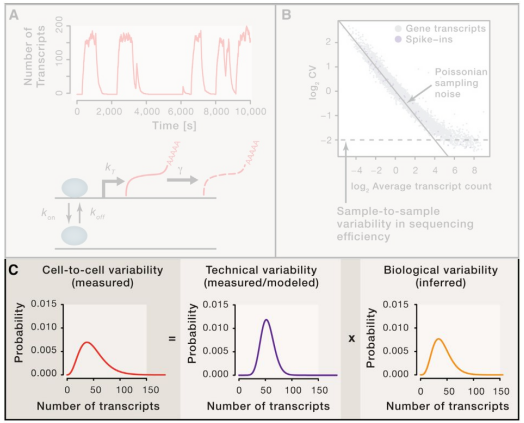


Measuring Gene Expression Noise

- Gene Activation
- Sequencing Noise
- **Noise Components**

- **Noise Components**





Thank you

Investigating Allelic Expression

Investigating Allelic Expression

Investigating Allelic Expression

Example: parental X chromosome

Concluding Remarks