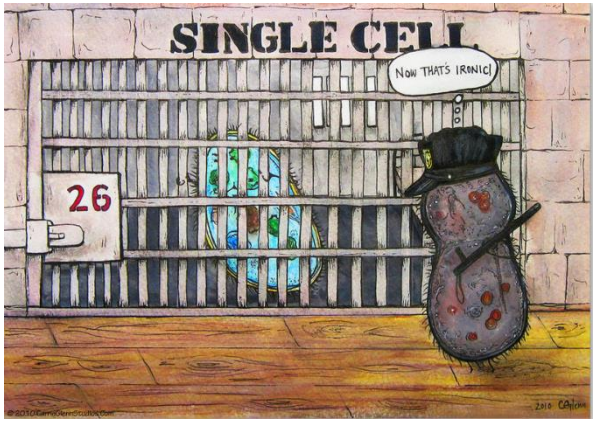
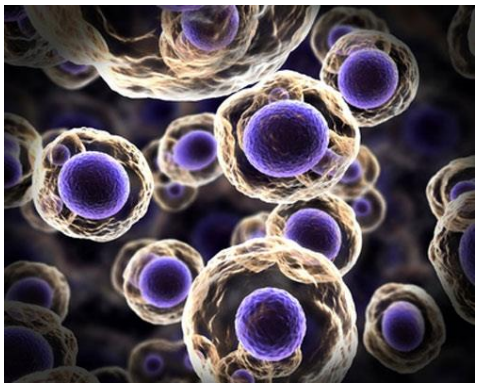
Single cell transcriptomics



Ester Feldmesser

Introduction to Deep Sequencing Analysis- June 2015

Single cell applications



- Single-cell genome sequencing and structure
- Single-cell methylation analysis
- Single-cell transcriptomics (and proteomics)
- Combined approaches

Overview

- Motivation for single cell transcriptomics
- Protocols
- Pitfalls and how to overcome them
- One specific protocol: Quantitative single-cell RNA-seq with unique molecular identifiers
- Some examples of results
- *In situ* single cell transcriptomics

Reference papers

REVIEWS

SINGLE-CELL OMICS

Computational and analytical challenges in single-cell transcriptomics

Oliver Stegle[1], Sarah A. Teichmann[1,2] and John C. Marioni[1,2]

Abstract | The development of high-throughput RNA sequencing (RNA-seq) at the single-cell level has already led to profound new discoveries in biology, ranging from the identification of novel cell types to the study of global patterns of stochastic gene expression. Alongside the technological breakthroughs that have facilitated the large-scale generation of single-cell transcriptomic data, it is important to consider the specific computational and analytical challenges that still have to be overcome. Although some tools for analysing RNA-seq data from bulk cell populations can be readil applied to single-cell RNA-seq data, many new computational strategies are required to fully exploit this data type and to enable a comprehensive yet detailed study of gene expression at the single-cell level.

doi:10.1038/nrg3833

Quantitative single-cell RNA-seq with unique molecular identifiers

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg & Sten Linnarsson

Affiliations | Contributions | Corresponding author

Nature Methods **11**, 163–166 (2014) | doi:10.1038/nmeth.2772
Received 27 September 2013 | Accepted 25 November 2013 | Published online 22 December 2013

Single-cell RNA sequencing (RNA-seq) is a powerful tool to reveal cellular heterogeneity, discover new cell types and characterize tumor microevolution. However, losses in cDNA synthesis and bias in cDNA amplification lead to severe quantitative errors. We show that molecular labels—random sequences that label individual molecules—can nearly

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Single-cell sequencing-based technologies will revolutionize whole-organism science

Ehud Shapiro[1,2], Tamir Biezuner[1,2] and Sten Linnarsson[3]

Abstract | The unabated progress in next-generation sequencing technologies is fostering a wave of new genomics, epigenomics, transcriptomics and proteomics technologies. These sequencing-based technologies are increasingly being targeted to individual cells, which will allow many new and longstanding questions to be addressed. For example, single-cell genomics will help to uncover cell lineage relationships; single-cell transcriptomics will supplant the coarse notion of marker-based cell types; and single-cell epigenomics and proteomics will allow the functional states of individual cells to be analysed. These technologies will become integrated within a decade or so, enabling high-throughput, multi-dimensional analyses of individual cells that will produce detailed knowledge of the cell lineage trees of higher organisms, including humans. Such studies will have important implications for both basic biological research and medicine.

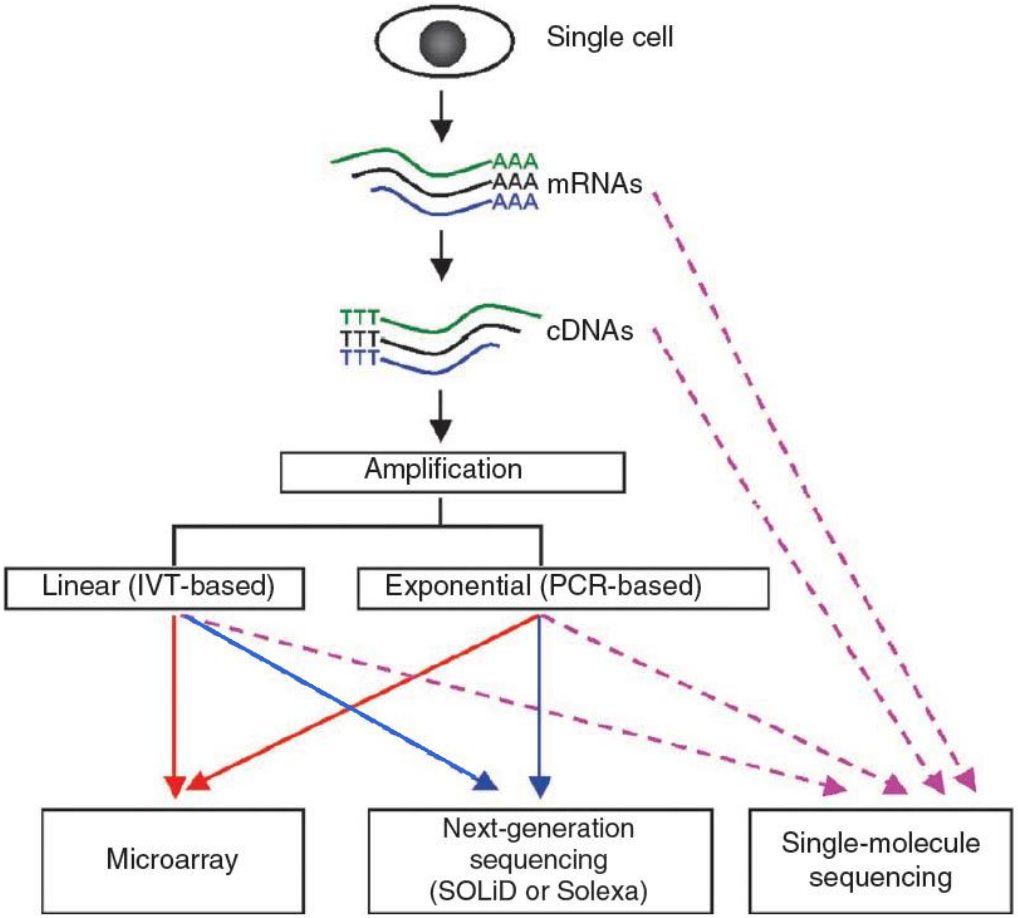
Motivation

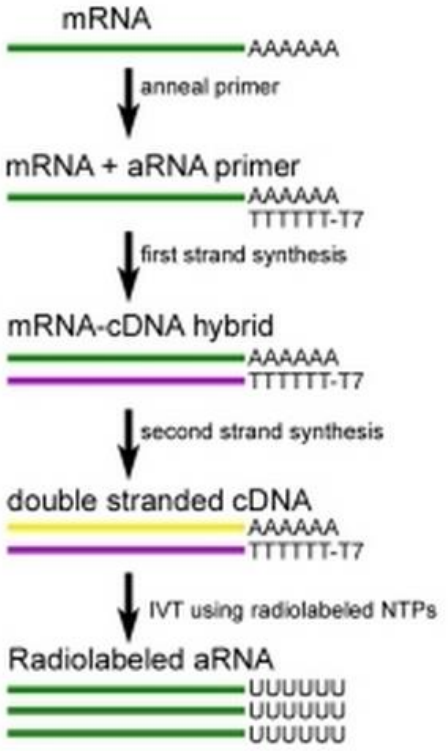
- Tissues are complex and not homogenous
- Different cell types gene expression can be masked
- Identification of cell types and cell states
- Learn about noise in gene expression
- Ideally would like to combine transcriptional profiling with spatial information

Protocols for single cell

- Isolate individual cells:
  - Manually by micromanipulation
  - Manually by fluorescence-activated cell sorting (FACS)
  - Exploiting a microfluidics based system
  - Laser capture microdissection
- Capture the poly-adenylated fraction of mRNA molecules
- Obtain cDNA by reverse transcription

Single-cell transcriptome





Nature Methods 8, S6–S11 (2011) doi:10.1038/nmeth.1557

Overview

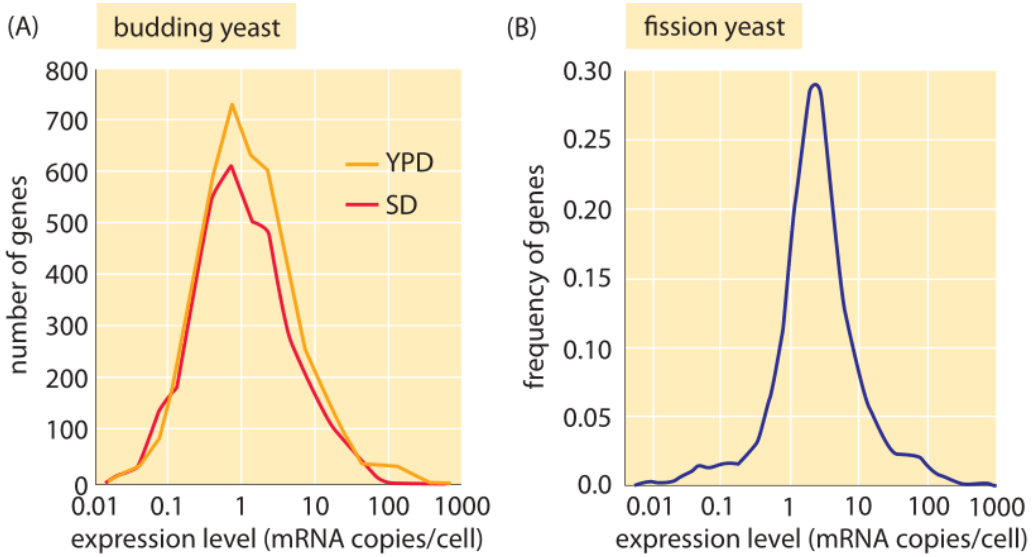
- Motivation for single cell transcriptomics
- Protocols
- Pitfalls and how to overcome them
- One specific protocol: Quantitative single-cell RNA-seq with unique molecular identifiers
- Some examples of results
- In situ single cell transcriptomics

How many mRNAs are there in a cell?

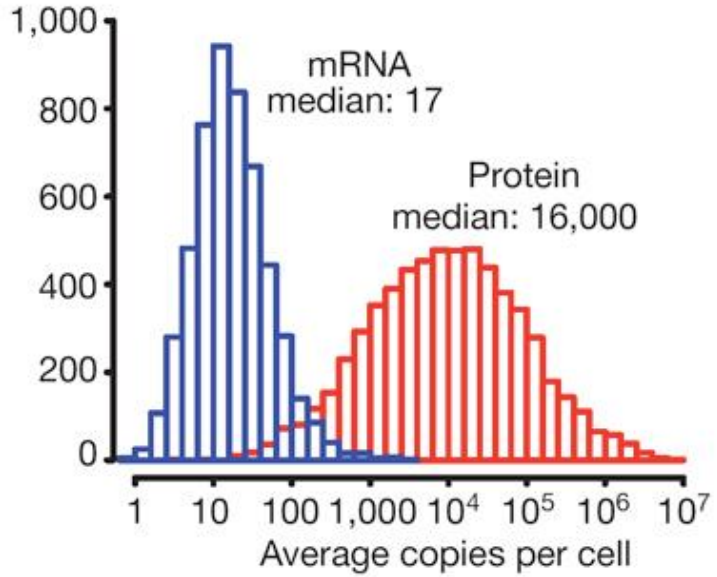
- Approximately 360,000 mRNA molecules in a single mammalian cell, made up of approximately 12,000 different transcripts

- Some mRNAs comprise 3% of the mRNA pool whereas others account for less than 0.1%. These rare or low-abundance mRNAs may have a copy number of only 5-15 molecules per cell.

mRNAs numbers in a cell



Mammalian



http://book.bionumbers.org/how-many-mrnas-are-in-a-cell/
https://www.mdc-berlin.de/36010975/en/highlights/archive/2011/selbach

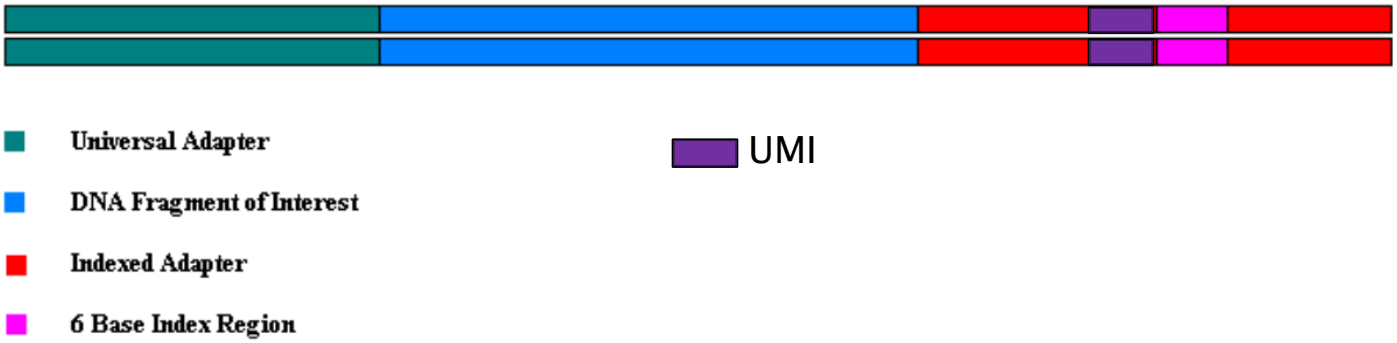
Incorporating quantitative standards

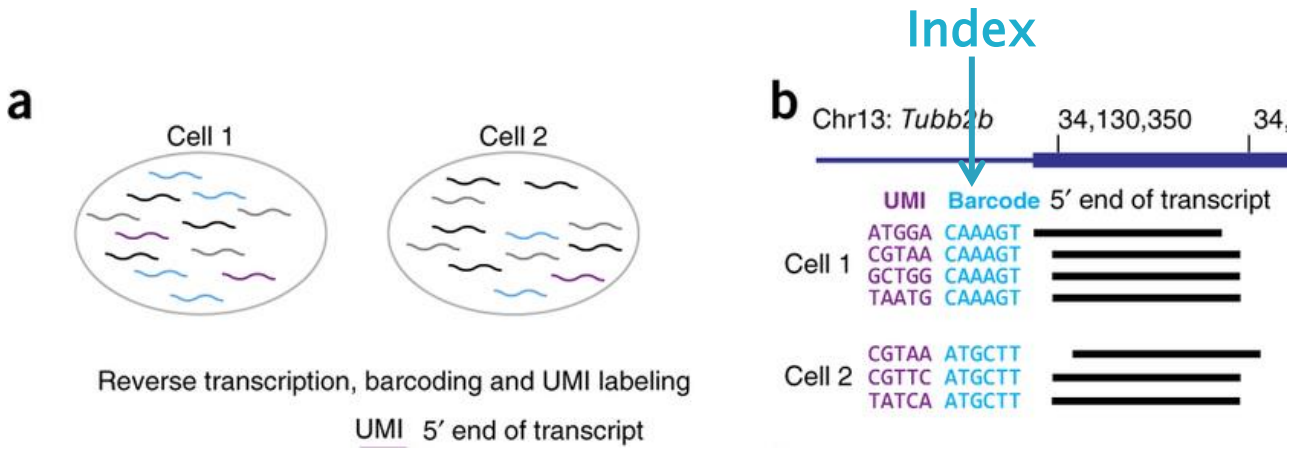
- Use extrinsic spike-in molecules: artificial spike-in mix is the External RNA Control Consortium (ERCC) set of 92 synthetic spikes based on bacterial sequences

- Unique molecular identifiers (UMIs) have been used to barcode individual molecules, when sequencing the 3′ or 5′ end of the amplified transcript.

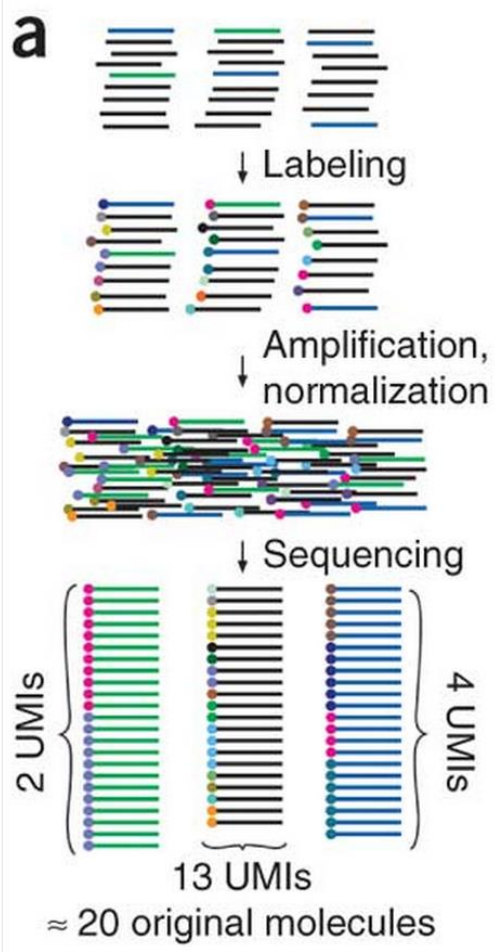
Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

Unique molecular identifier (UMI)





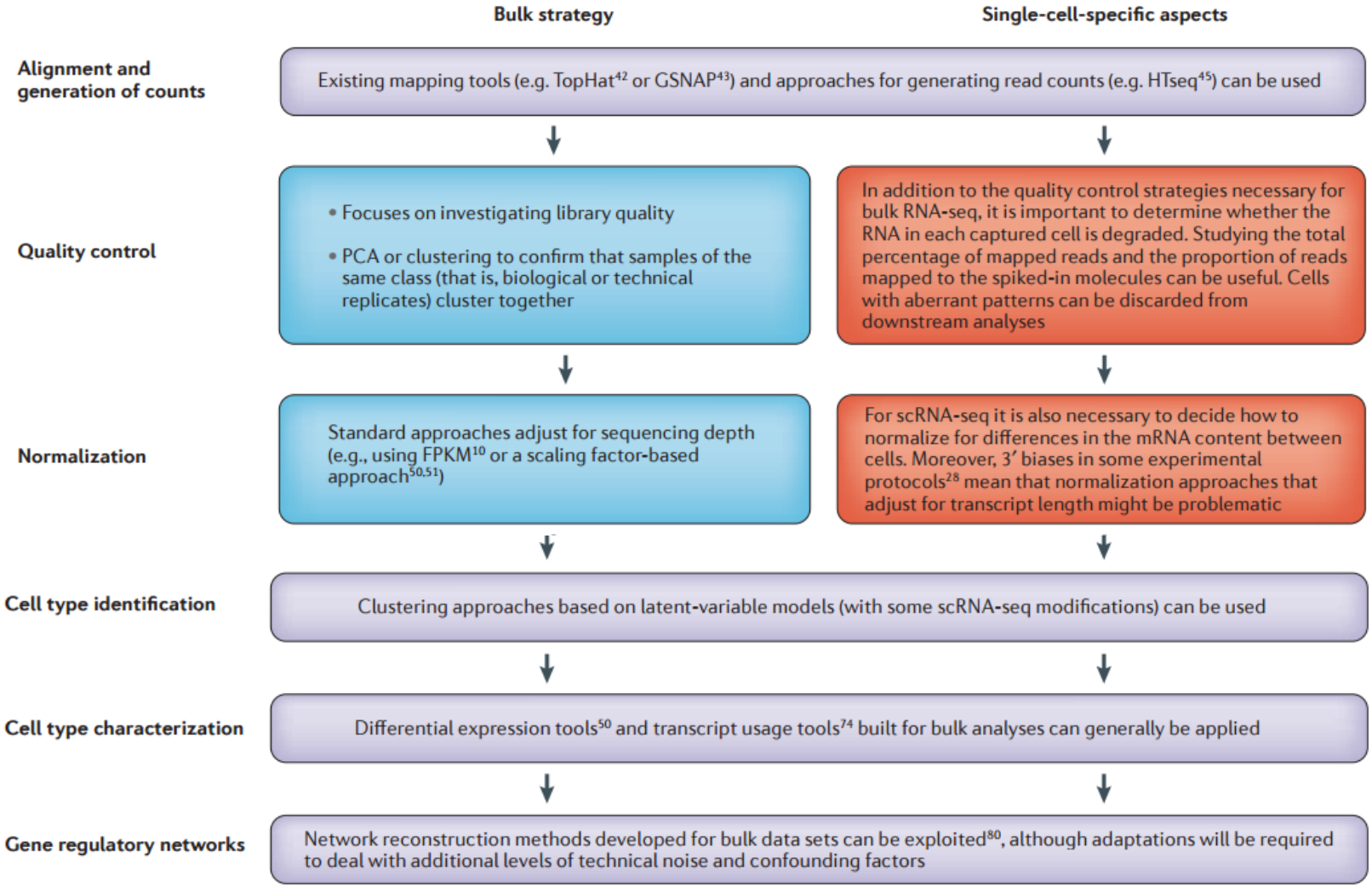
UMIs can be generated by adding oligonucleotide labels



Nature Methods 9, 72–74 (2012) doi:10.1038/nmeth.1778

Transcript quantification and quality control

Comparison of bulk and scRNA-seq analytical strategies

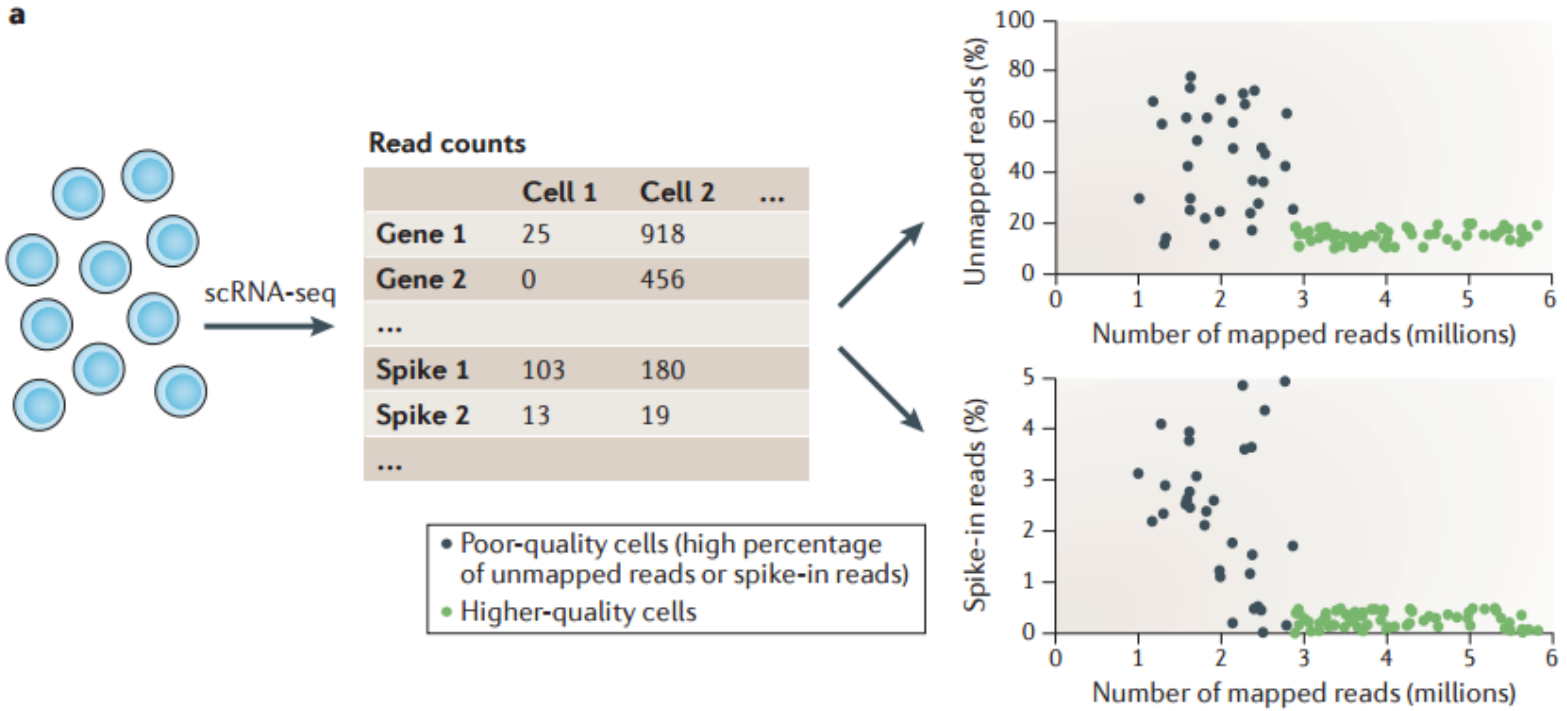


Quality control

- Cells captured may contain degraded RNA: for example, because the cell is stressed because of extraction and isolation
- Check:
  - ✓ Mapping to the genome
  - ✓ Ratio of reads mapped to the genome versus the number of reads mapped to the extrinsic spike-ins
  - ✓ Principal component analysis
- Discard outlier cells

Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

Quality control (2)



b. PCA

The expectation when applying PCA is that good-quality cells cluster together and poor-quality cells are outliers.

Poor-quality cells may also form a second distinct cluster.

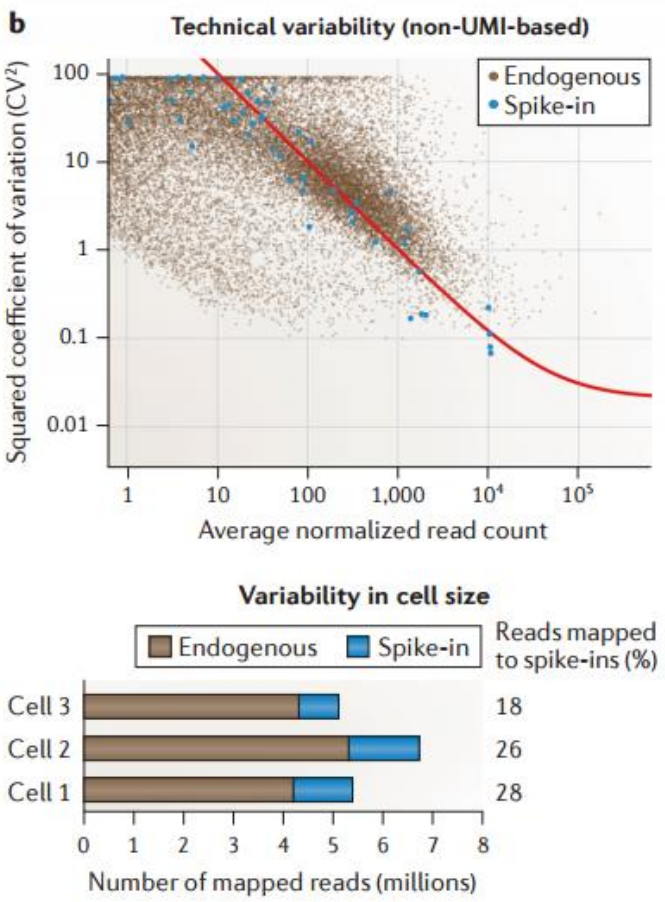
Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

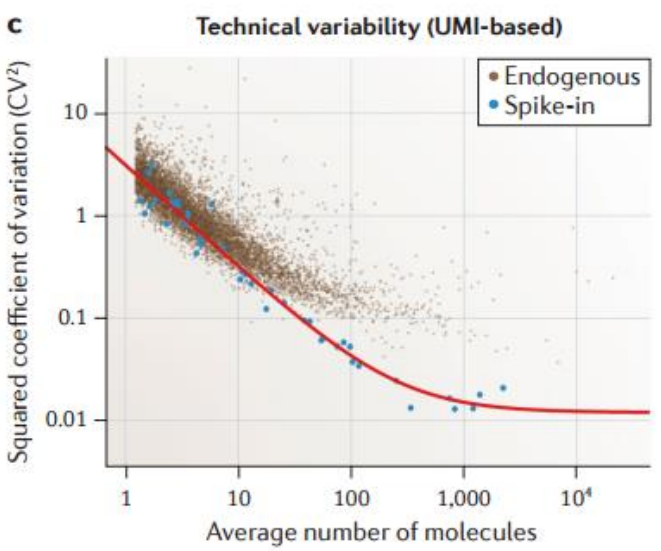
Normalization method

- Common approaches for normalizing bulk RNA-seq data make an implicit assumption: that the total amount of RNA processed in each sample is approximately the same or that the variation is technical.
- This assumption has been shown to be misleading (for example, upregulation of *MYC* leads to a two-fold increase in the number of transcripts)

Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

Normalization





Capture efficiency

Cell	Number of UMIs associated with spike-in molecules	Initial number of spiked in molecules	Capture efficiency (%)
1	1,436	10,000	14.36
2	2,598	10,000	25.98
3	987	10,000	9.87

Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

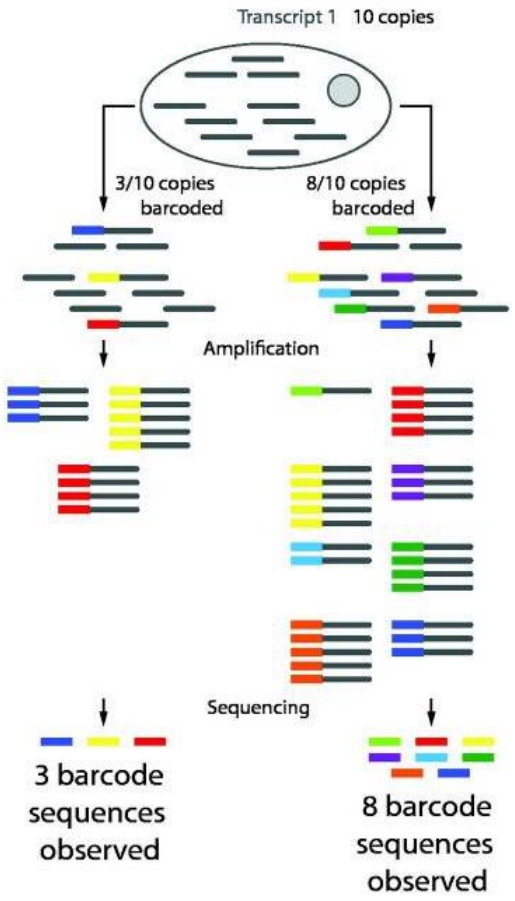
Normalization issues

- ERCC have short poly-A tails (efficiency RT)
- ERCC spikes are short (500 – 2000)
- 5′-to-3′ length bias is inherent to many scRNA-seq protocols (problems normalizing)
- Normalizing for transcript length is challenging with current scRNA-seq protocols (3' biases)
- Differences in the efficiency of the reverse transcription reaction (UMI)

Nature Reviews Genetics 16, 133-145 (2015) doi:10.1038/nrg3833

Differences in the efficiency of the reverse transcription reaction (UMI)

UMI principle and the output depending on reaction efficiency



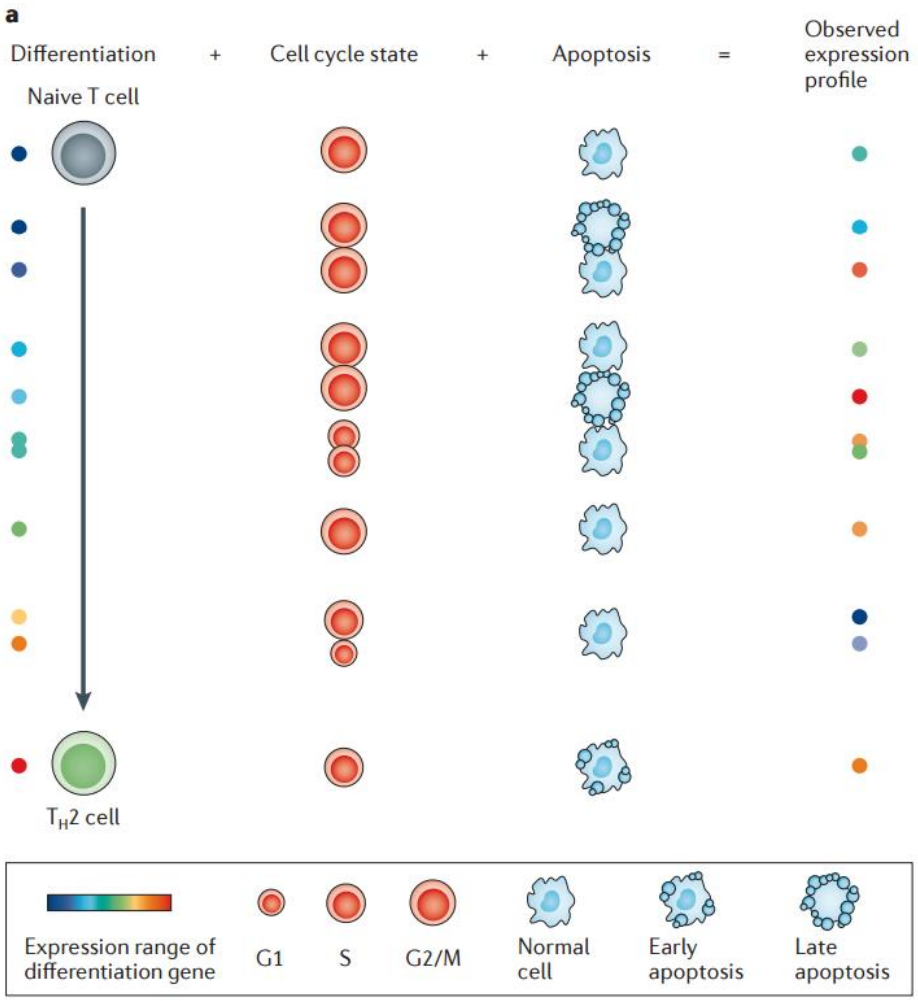
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3877366/

scRNA-seq experimental design

- Number of cells sequenced
- Depth (# reads) for each individual cell
  - Jaitin et al generated an average of 22,000 aligned sequence reads from 1,536 cells
  - Mahata et al. generated 12-20 million sequence reads from each of 93 cells
- Batch effects
- Biological factors: cell cycle, physical condition of the cell

Nature Reviews Genetics 16, 133-145 (2015) doi:10.1038/nrg3833

Confounding variables



Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

Overview

- Motivation for single cell transcriptomics
- Protocols
- Pitfalls and how to overcome them
- One specific protocol: Quantitative single-cell RNA-seq with unique molecular identifiers
- Some examples of results
- *In situ* single cell transcriptomics

NATURE METHODS | BRIEF COMMUNICATION

Quantitative single-cell RNA-seq with unique molecular identifiers

Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg & Sten Linnarsson

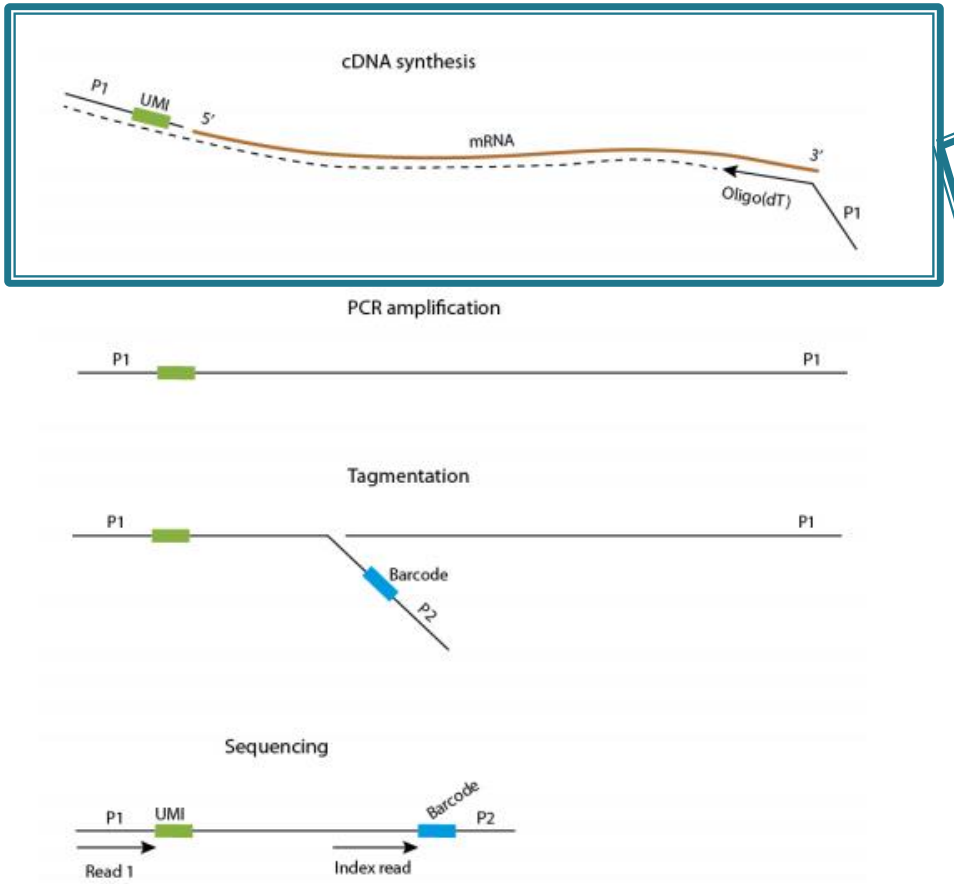
Affiliations | Contributions | Corresponding author

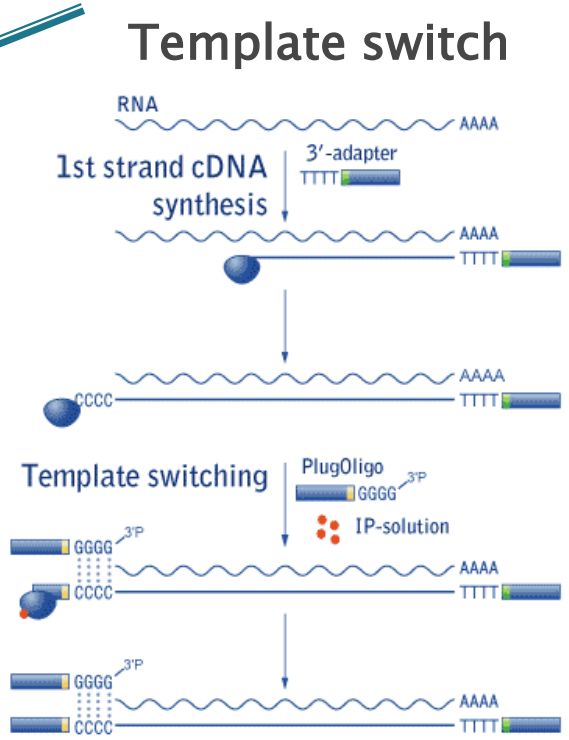
Nature Methods **11**, 163–166 (2014) | doi:10.1038/nmeth.2772
Received 27 September 2013 | Accepted 25 November 2013 | Published online 22 December 2013

Single-cell RNA sequencing (RNA-seq) is a powerful tool to reveal cellular heterogeneity, discover new cell types and characterize tumor microevolution. However, losses in cDNA synthesis and bias in cDNA amplification lead to severe quantitative errors. We show that molecular labels—random sequences that label individual molecules—can nearly eliminate amplification noise, and that microfluidic sample preparation and optimized reagents produce a fivefold improvement in mRNA capture efficiency.

- Experiment performed on ES cells
- Unique molecular identifiers – reduce amplification noise
- Microfluidic sample preparation – improves mRNA capture

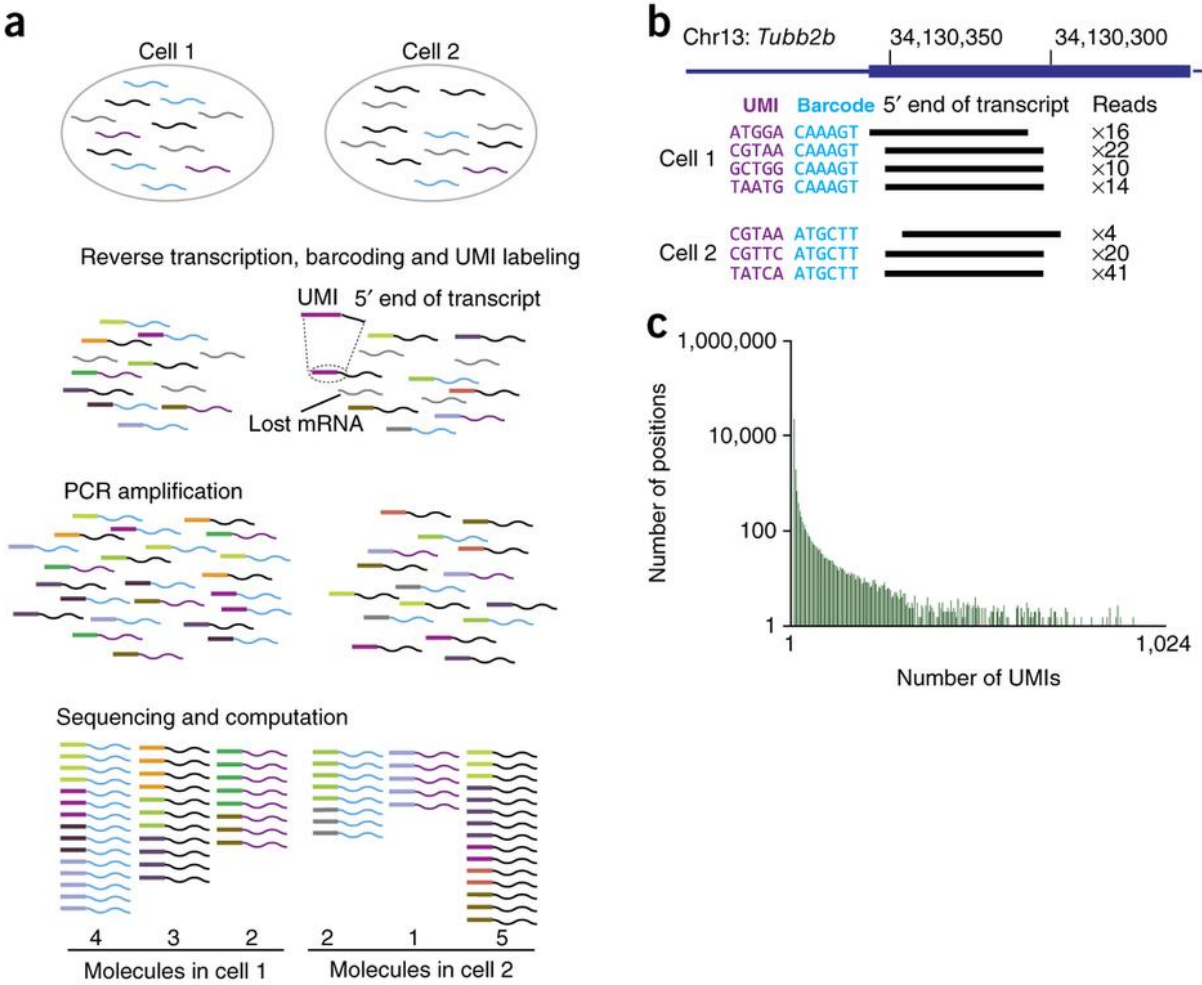
Unique molecular identifier (UMI) method





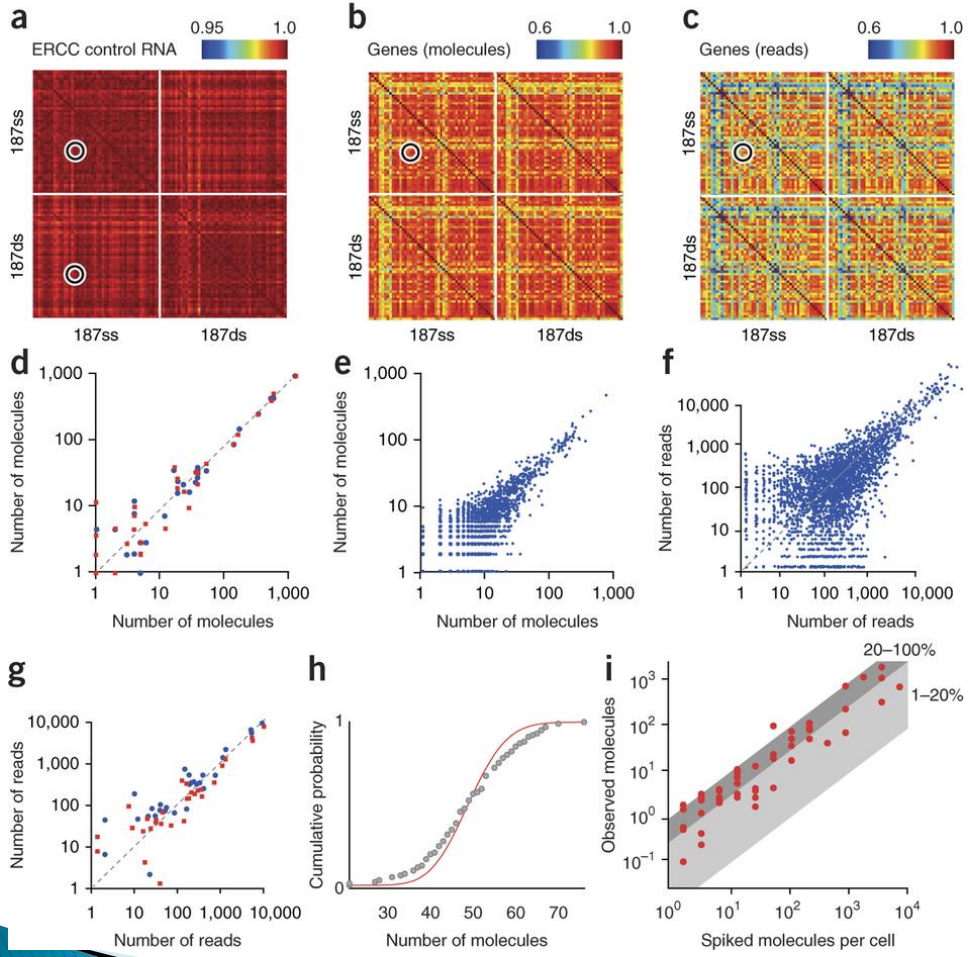
Nature Methods 11, 163–166 (2014) doi:10.1038/nmeth.2772
http://www.evrogen.com/img/Mint-Uni-scheme-large.png

Overview of tagging single mRNA molecules



Nature Methods 11, 163–166 (2014) doi:10.1038/nmeth.2772

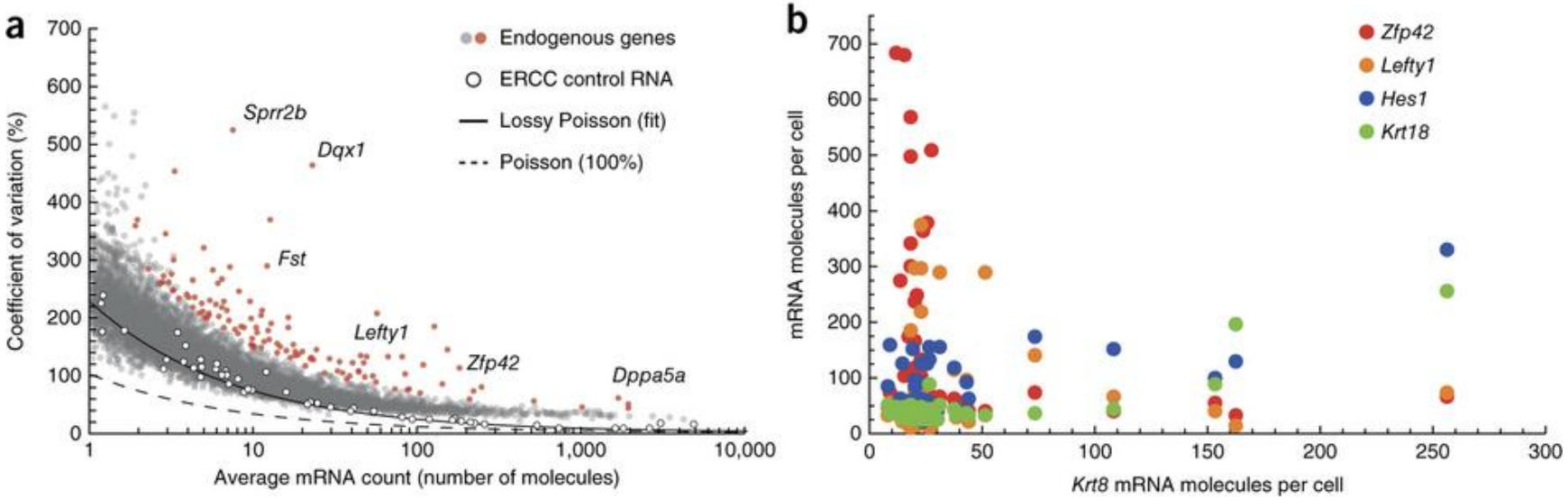
Reproducibility of molecule counting



(**a**,**b**,**c**) Pairwise correlation coefficients calculated for ERCC, endogenous genes using molecule counts or counting reads instead of molecules
(**d**) Pairwise comparison of two wells indicated in **a**. Red squares and blue dots show comparisons within and between libraries, respectively.
(**e**,**f**) Scatterplot showing the two cells indicated in **b** or **e**, based on molecule counts or read counts respectively.
(**g**) Scatterplot as in **d** but using reads instead of molecules. (**h**) Distribution of molecule counts for a single ERCC spike-in transcript (gray dots) compared with the cumulative density function of the Poisson distribution (red line). (**i**) mRNA capture efficiency shown as observed molecule counts versus number of spiked-in molecules for ERCC control RNA transcripts. The shaded bands indicate efficiencies above (dark gray) and below (light gray) 20%. Each red dot represents the average of a single ERCC RNA across 96 wells. Similar results were obtained in one replicate experiment.

Nature Methods 11, 163–166 (2014) doi:10.1038/nmeth.2772

Transcriptional noise in ES cells



(a) The coefficient of variation as a function of the mean number of mRNA molecules detected, for genes expressed in ES cells ($n = 4$) Genes with significant ($\alpha = 0.05$) excess noise are shown as red dots

(b) Coexpression of noisy genes. The expression of four genes, across 41 ES cells, is plotted against the expression of *Krt8*. Two branches were observed, interpreted as 'epiblast-like' (high *Krt8*, *Krt18* and *Hes1*) and 'pluripotent-like' (high *Zfp42* and *Lefty1*)

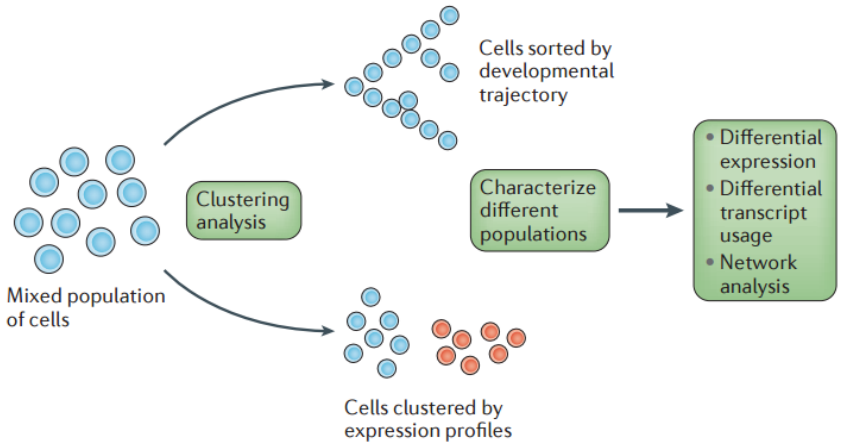
Nature Methods 11, 163–166 (2014) doi:10.1038/nmeth.2772

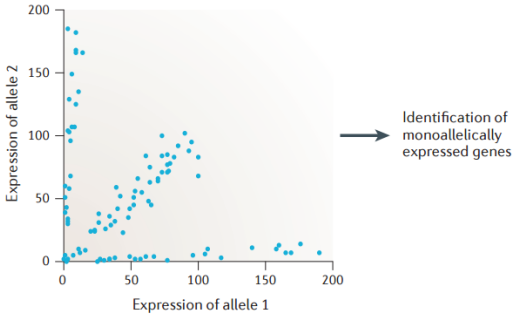
Overview

- Motivation for single cell transcriptomics
- Protocols
- Pitfalls and how to overcome them
- One specific protocol: Quantitative single-cell RNA-seq with unique molecular identifiers
- Some examples of results
- *In situ* single cell transcriptomics

What information can be obtained from scRNA-seq?

- Identification of cell type and cellular state
- Differential expression at the gene level
- Differential expression at the transcript-isoform level?
- Identification of highly variable genes
- Kinetics of gene expression
- Detection of allele-specific expression

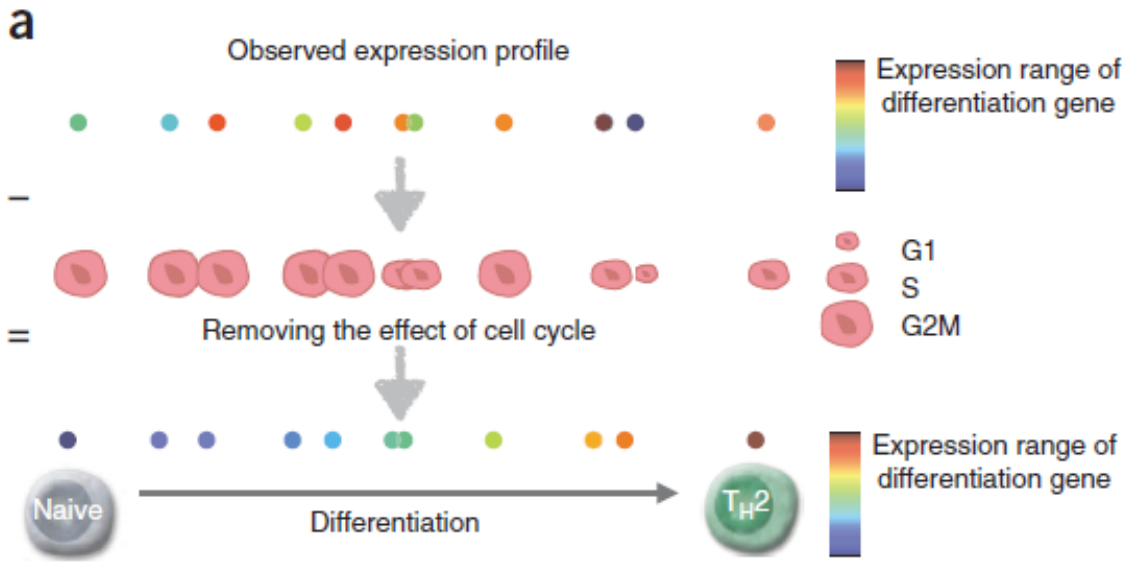


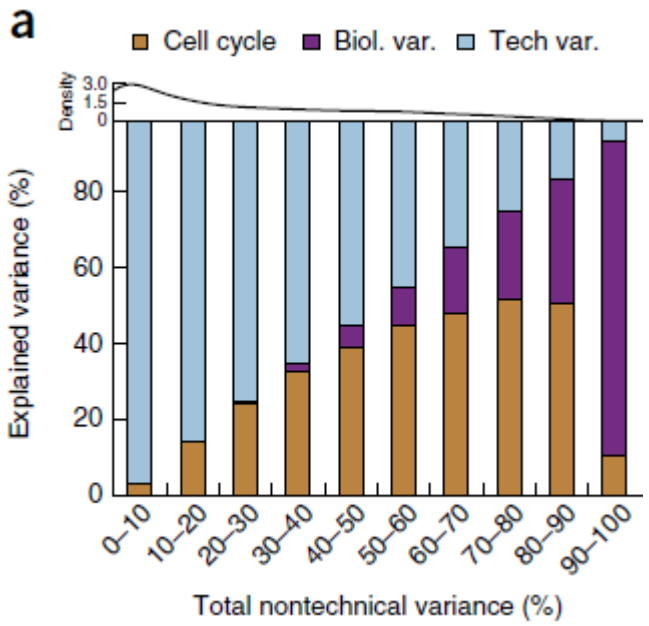


Nature Reviews Genetics 16, 133–145 (2015) doi:10.1038/nrg3833

Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Florian Buettner[1,2,5], Kedar N Natarajan[2,3,5], F Paolo Casale[2], Valentina Proserpio[2,3], Antonio Scialdone[2,3], Fabian J Theis[1,4], Sarah A Teichmann[2,3], John C Marioni[2,3] & Oliver Stegle[2]





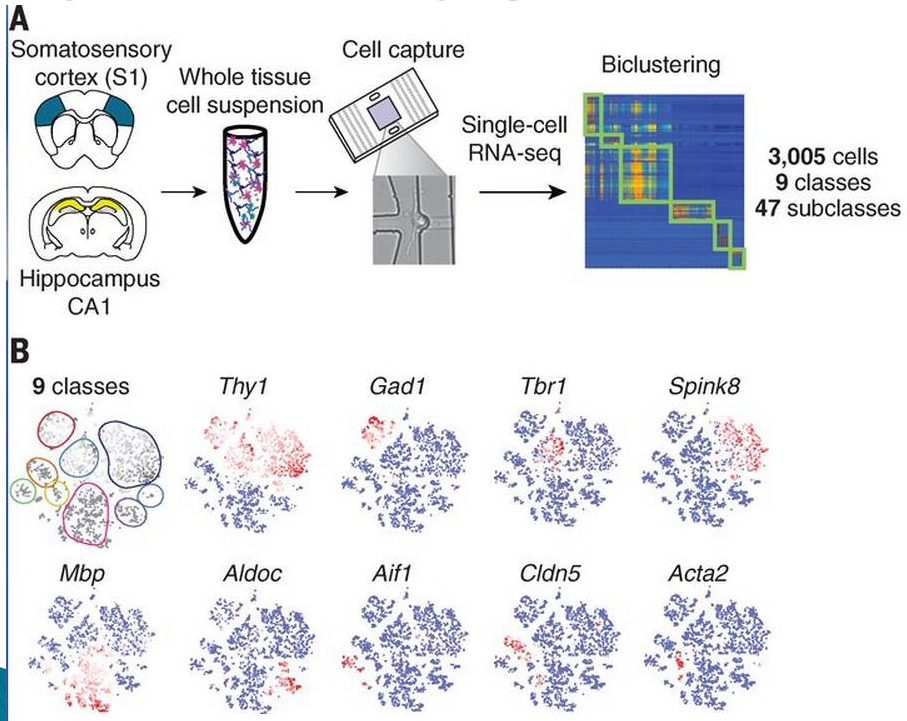
- Use a latent variable model approach to regress out variation attributable to the cell-cycle
- Technique can be used to remove variation for any given set or sets of genes

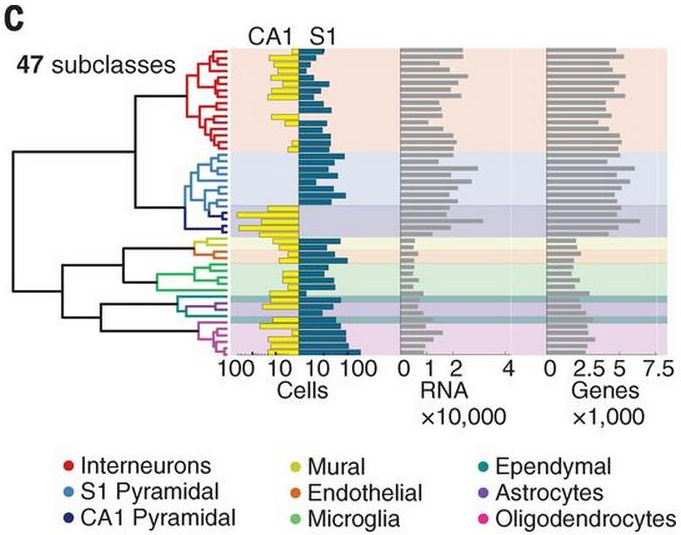
Published Online February 19 2015
Science 6 March 2015:
Vol. 347 no. 6226 pp. 1138-1142
DOI: 10.1126/science.aaa1934

REPORT

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

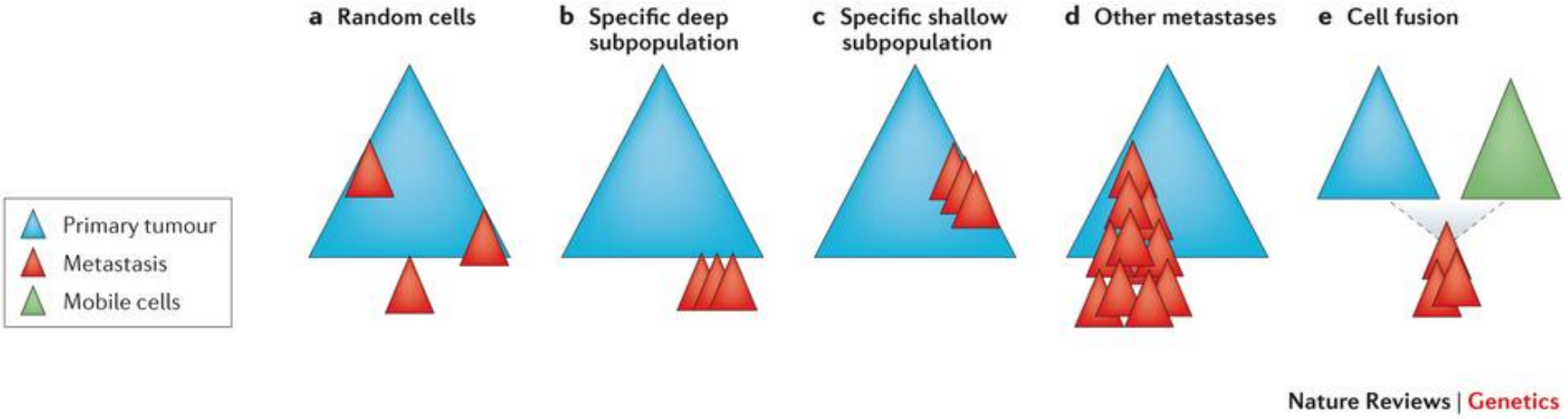
Amit Zeisel[1,*], Ana B. Muñoz-Manchado[1,*], Simone Codeluppi[1], Peter Lönnerberg[1], Gioele La Manno[1], Anna Juréus[1], Sueli Marques[1], Hermany Munguba[1], Liqun He[2], Christer Betsholtz[2,3], Charlotte Rolny[4], Gonçalo Castelo-Branco[1], Jens Hjerling-Leffler[1,†], Sten Linnarsson[1,†]





Clinical use of single cell genomics/transcriptomics

Cell lineage reconstruction of cancer will elucidate its development

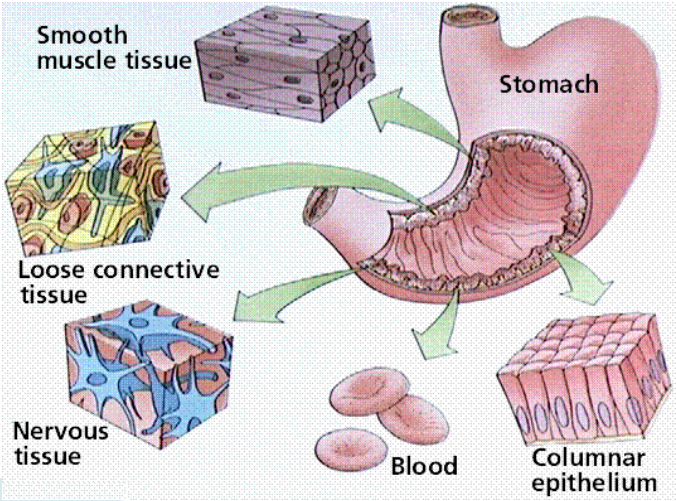


Alternative hypotheses on the origin of metastases.

Nature Reviews Genetics 14, 618–630 (2013) doi:10.1038/nrg3542

Seeing is believing:
new methods for *in situ* single-cell transcriptomics

- Spatially defined single cells in live tissue
- Artifacts when separating by pipetting or using laser capture
- There is also a protocol for fixed cells in the tissue.



Genome Biology 2014 15:110.

In situ single-cell transcriptomics

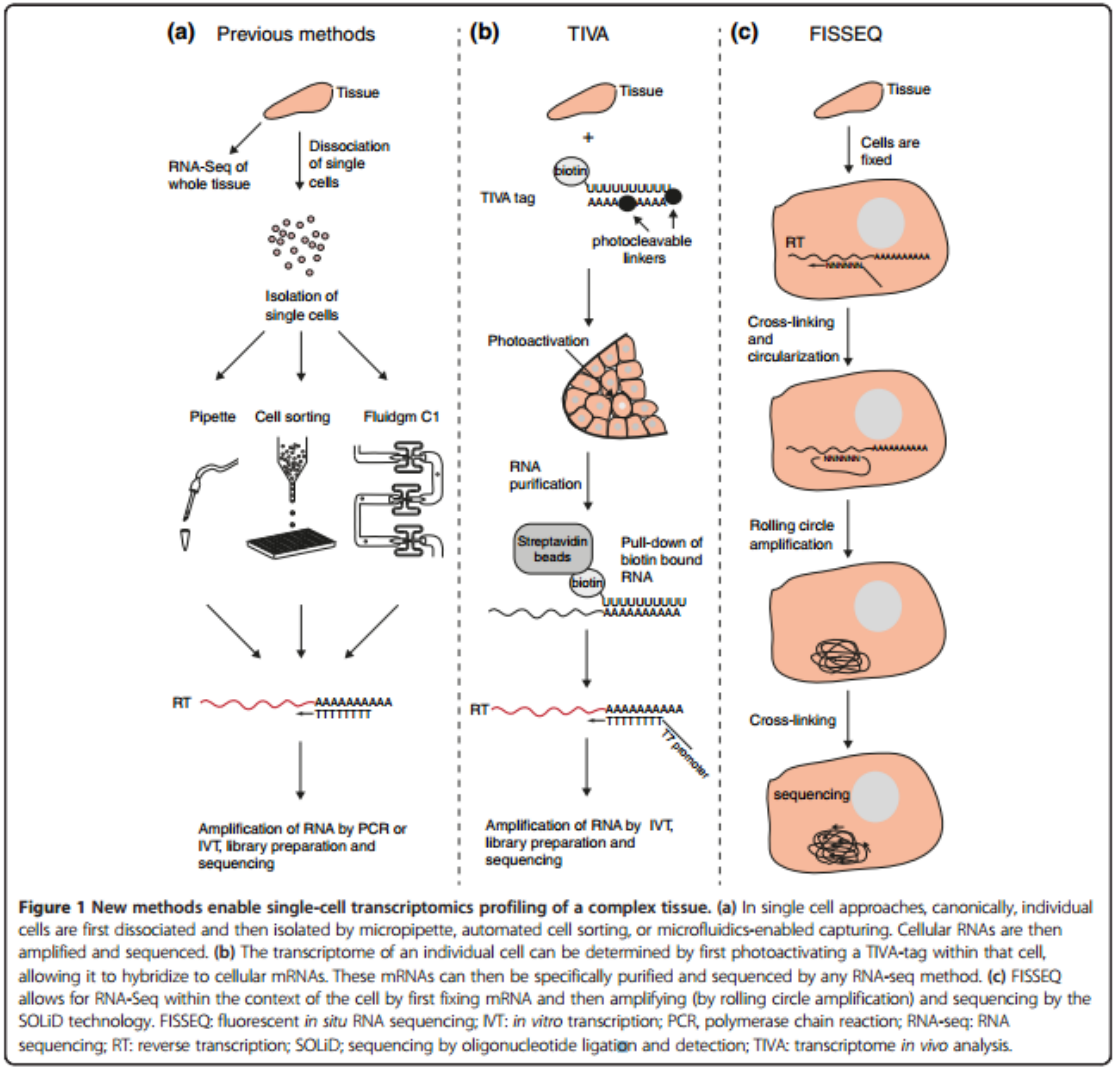
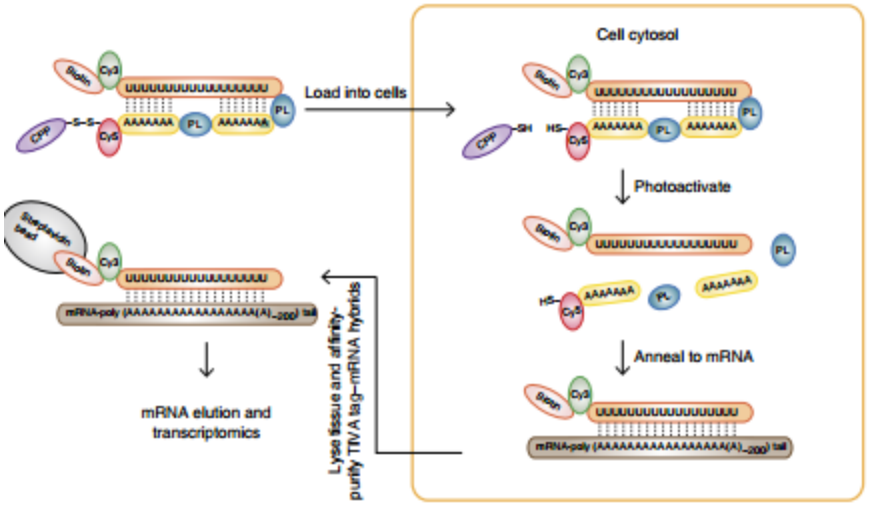


Figure 1 New methods enable single-cell transcriptomics profiling of a complex tissue. (a) In single cell approaches, canonically, individual cells are first dissociated and then isolated by micropipette, automated cell sorting, or microfluidics-enabled capturing. Cellular RNAs are then amplified and sequenced. **(b)** The transcriptome of an individual cell can be determined by first photoactivating a TIVA-tag within that cell, allowing it to hybridize to cellular mRNAs. These mRNAs can then be specifically purified and sequenced by any RNA-seq method. **(c)** FISSEQ allows for RNA-Seq within the context of the cell by first fixing mRNA and then amplifying (by rolling circle amplification) and sequencing by the SOLiD technology. FISSEQ: fluorescent *in situ* RNA sequencing; IVT: *in vitro* transcription; PCR, polymerase chain reaction; RNA-seq: RNA sequencing; RT: reverse transcription; SOLiD; sequencing by oligonucleotide ligation and detection; TIVA: transcriptome *in vivo* analysis.

Genome Biology 2014 15:110.

Transcriptome *in vivo* analysis (TIVA)



TIVA tag:

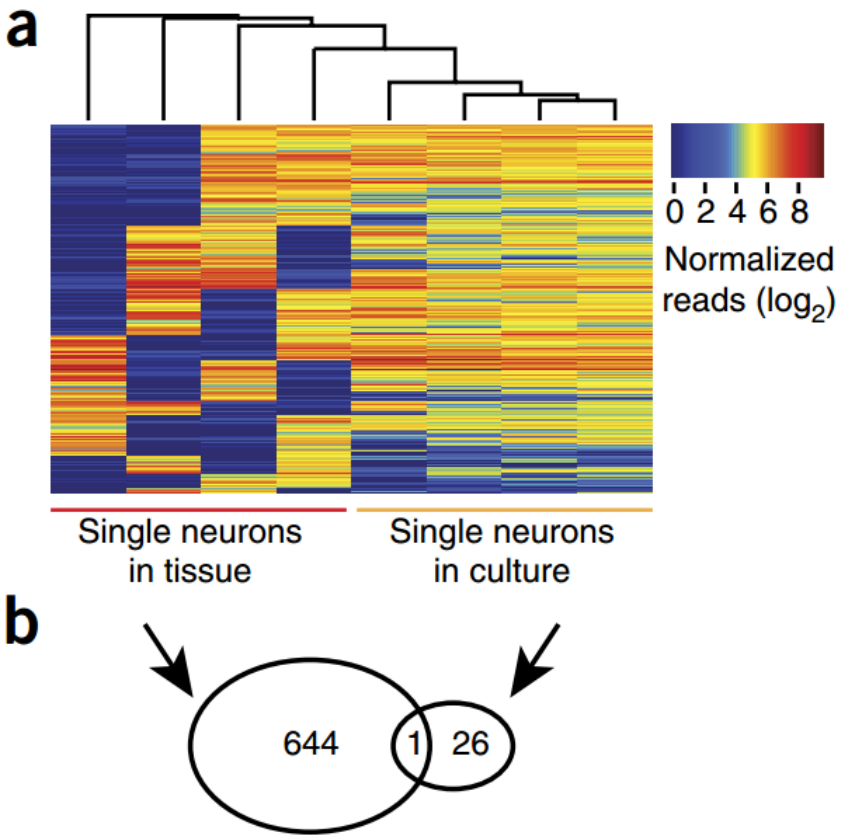
- a cell penetrating peptide,
- a photocleavable linker
- the fluorophores Cy3 and Cy5
- a poly(U) oligonucleotide
- biotin

TIVA tag:

- can permeate into the cells of a tissue the peptide dissociates from the TIVA tag
- laser photoactivation on a particular cell causes its TIVA tags to come undone
- tags can then anneal to mRNAs in the light-selected cell
- the desired mRNA can be pulled out by using streptavidin beads and then be sequenced

Nature Methods 11, 190–196 (2014) doi:10.1038/nmeth.2804

Bimodal transcripts in single hippocampal neurons in tissue



(a) Heatmaps shows clustering of 645 bimodally expressed genes. Bimodally expressed genes were defined as having a gap in expression of at least four log units in two samples.
(b) Overlap between bimodal genes in single neurons from tissue and from culture (4 cells in each group).

Nature Methods 11, 190–196 (2014) doi:10.1038/nmeth.2804

Thanks for your attention!

Any questions?