



ITMO UNIVERSITY

CT Lab  
ITMO UNIVERSITY

# Introduction to gene expression

Konstantin Zaitsev

March 16<sup>th</sup>, 2019

# About the course

# About the course

Authors of this course:

1. Konstantin Zaitsev,

[https://www.researchgate.net/profile/Konstantin\\_Zaitsev](https://www.researchgate.net/profile/Konstantin_Zaitsev),

<https://stepik.org/course/512/>

2. Alexander Tkachenko,

<https://publons.com/researcher/3041889/alexander-tkachenko/>,

<https://stepik.org/course/94/>

# About the course

- Course is scientific-oriented
- By “gene expression” we almost always mean “RNA expression” if not said otherwise
- Course will take place in Zoom

Most important link:

<https://drive.google.com/drive/folders/1XvE5leRSwm1HEU6HP8MVakkj48Bobi>

# About the course: prerequisites

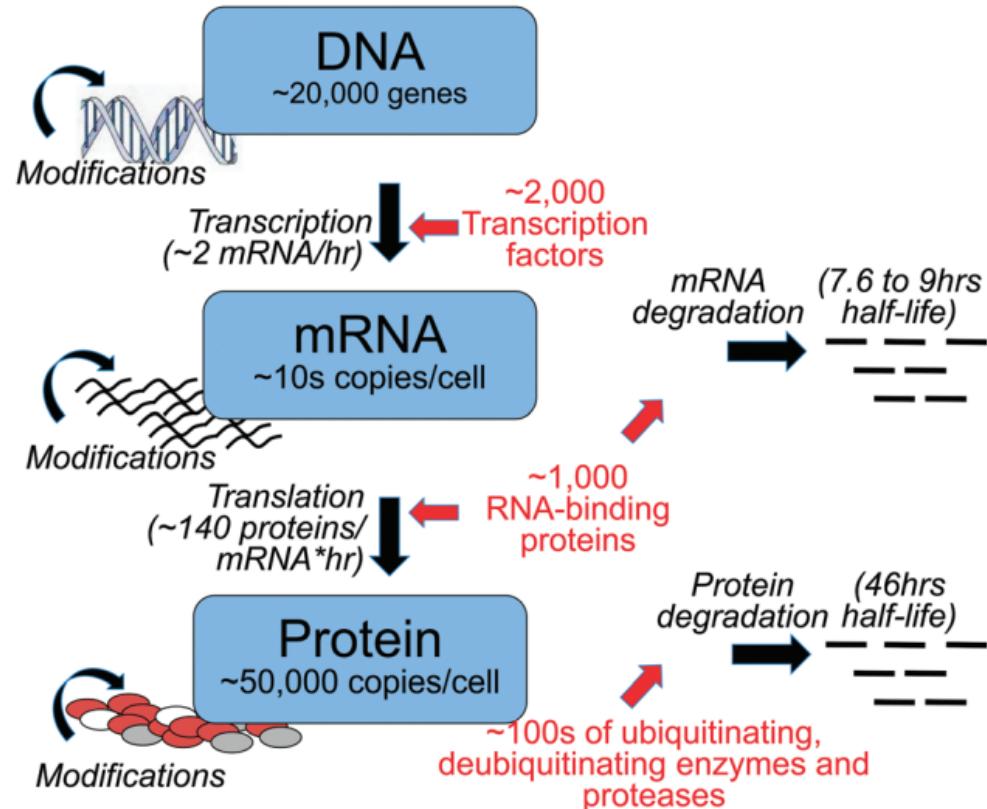
- You are expected to be able to use google
- You are expected to be able to read documentation / papers
- You will be given access to a machine with all the packages installed
- 4 homeworks which you will have to defend ( in person (?) )

# About the course: syllabus

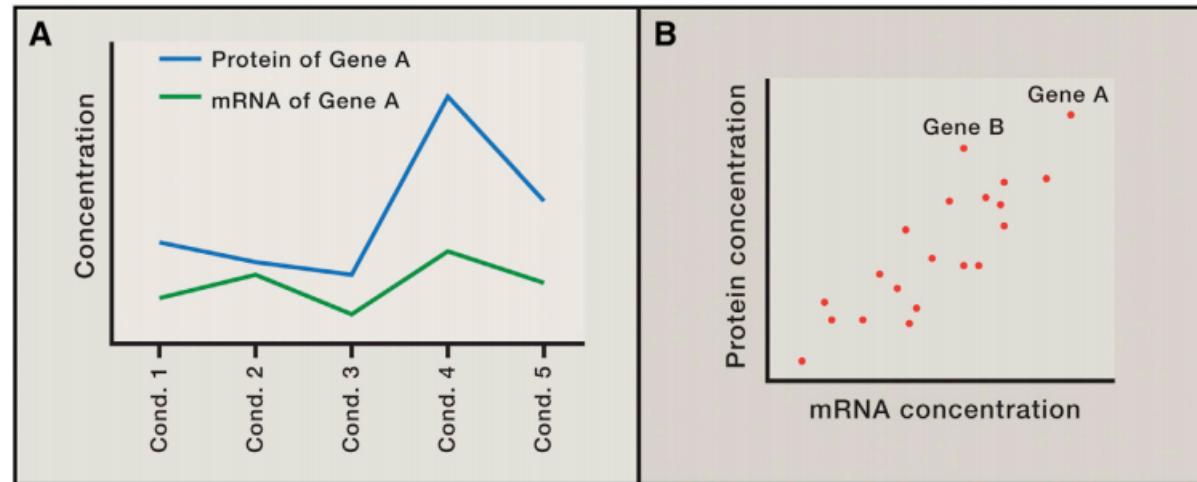
- Central dogma of molecular biology, structure of gene, types of RNA, structure of RNA, transcription, reverse transcription, qPCR, FISH, FACS
- Microarray: quantification, normalization, basic analysis
- RNA-seq: alignment, quantification, QC, normalization, basic analysis
- Overall quality control: PCA, clustering, outlier detection
- Overall quality control: batch correction
- Differential expression (DE): limma for microarray, Deseq2 for RNA-seq
- Downstream analysis: pathway/gene set enrichment analysis
- Downstream analysis: gene expression deconvolution
- Transcriptome assembly, functional annotation
- Single-cell transcriptomics: Seurat basic analysis
- Single-cell transcriptomics: Trajectory analysis, RNA velocity, optimal transport
- Visual data exploration: phantasus, JBR genome browser
- Experimental design of gene expression study

# Measuring RNA abundance

# Why we measure RNA abundance



# Why we measure RNA abundance



**Figure 1. Different Types of Correlations between Protein and mRNA Levels Need to Be Distinguished**

(A) Variation of one protein can be correlated with the variation of its coding mRNA across different conditions, tissues, individuals, or time points. This type of analysis addresses the question to what extent variation of protein levels is determined by variation of the corresponding mRNA for a specific gene.

(B) Concentrations of several proteins measured under the same condition can be correlated with their coding mRNAs. Here, each dot represents a single protein-mRNA pair. This type of analysis addresses the question to what extent concentration differences between transcripts from different genes show up at the level of proteins.

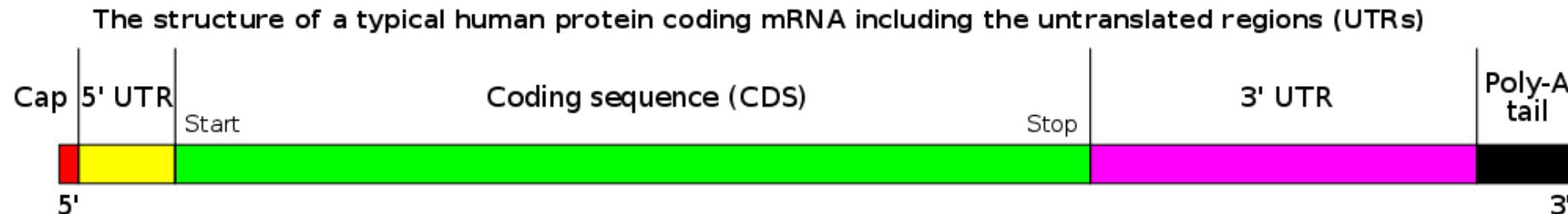
- A is very often the case
- B is almost never the case

# Why we measure RNA abundance

- While proteins are the key players in biological processes, RNA abundance is a good enough approximation for protein abundance
- Measuring RNA abundance is **much easier**
- Given several conditions and a gene: for **the same gene** fold change in RNA abundance will inform us about fold change in protein level

# Structure of mRNA

- We are mostly interested in mRNA (messenger RNA), because it is protein coding RNA
- Estimated  $10^5$  to  $10^6$  mRNA molecules per animal cell with high dynamic range for genes: from several copies to  $10^4$



# Types of RNA

- rRNA - ribosomal RNA: 80% of total cell RNA
- tRNA - transfer RNA: 15% of the cell RNA
- mRNA - messenger RNA for protein-coding genes
- Others RNAs: miRNA, lncRNA

# Strategies to capture RNA

Strategy	Type of RNA	Ribosomal RNA content	Unprocessed RNA content	Genomic DNA content	Isolation method
Total RNA	All	High	High	High	None
PolyA selection	Coding	Low	Low	Low	Hybridization with poly(dT) oligomers
rRNA depletion	Coding, noncoding	Low	High	High	Removal of oligomers complementary to rRNA
RNA capture	Targeted	Low	Moderate	Low	Hybridization with probes complementary to desired transcripts

# Why we measure RNA abundance

What can be a sample? RNA isolated from pretty much anywhere

- Blood draw
- Tissue / organ
- Specific isolated cell type

# Gene expression studies

- Usually you have the phenotype, but you don't know why it happens
- You take samples from several conditions related to your phenotype
- You try to figure out what is different in your phenotype from the control (or other condition)

Berry et al, 2010

LETTERS

---

## An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis

Matthew P. R. Berry<sup>1</sup>, Christine M. Graham<sup>1\*</sup>, Finlay W. McNab<sup>1\*</sup>, Zhaohui Xu<sup>6</sup>, Susannah A. A. Bloch<sup>3</sup>, Tolu Oni<sup>4,5</sup>, Katalin A. Wilkinson<sup>2,4</sup>, Romain Banchereau<sup>9</sup>, Jason Skinner<sup>6</sup>, Robert J. Wilkinson<sup>2,4,5</sup>, Charles Quinn<sup>6</sup>, Derek Blankenship<sup>7</sup>, Ranju Dhawan<sup>8</sup>, John J. Cush<sup>6</sup>, Asuncion Mejias<sup>10</sup>, Octavio Ramilo<sup>10</sup>, Onn M. Kon<sup>3</sup>, Virginia Pascual<sup>6</sup>, Jacques Banchereau<sup>6</sup>, Damien Chaussabel<sup>6</sup> & Anne O'Garra<sup>1</sup>

# Berry et al, 2010

LETTERS

## An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis

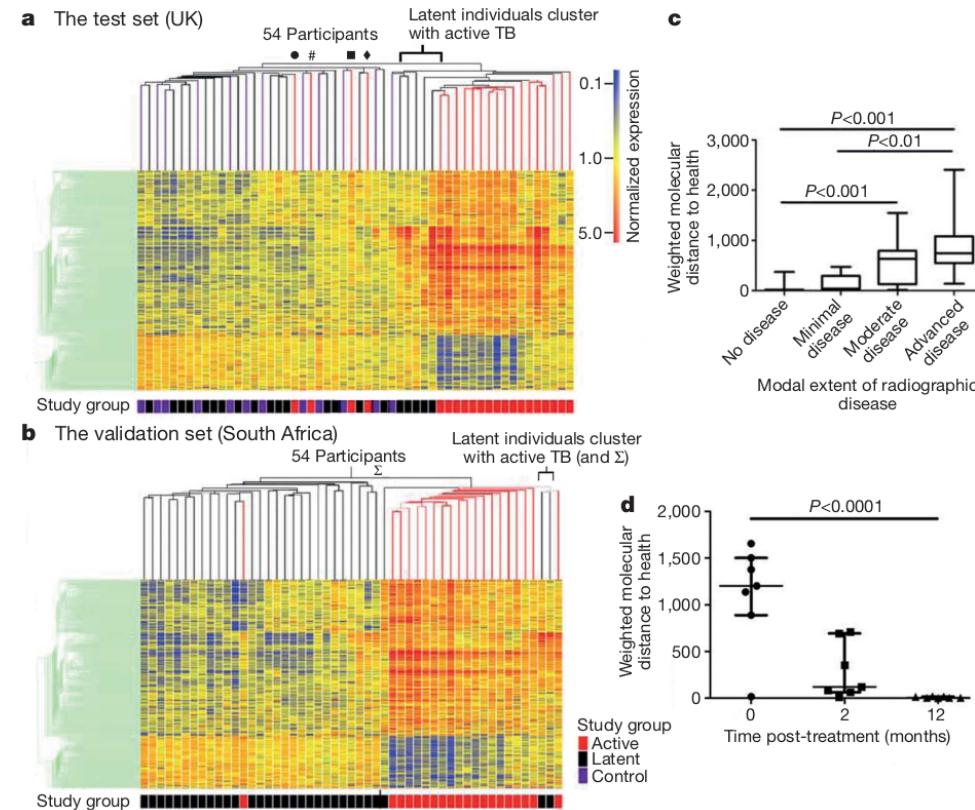
Matthew P. R. Berry<sup>1</sup>, Christine M. Graham<sup>1\*</sup>, Finlay W. McNab<sup>1\*</sup>, Zhaohui Xu<sup>6</sup>, Susannah A. A. Bloch<sup>3</sup>, Tolu Oni<sup>4,5</sup>, Katalin A. Wilkinson<sup>3,4</sup>, Romain Banchereau<sup>6</sup>, Jason Skinner<sup>6</sup>, Robert J. Wilkinson<sup>2,4,5</sup>, Charles Quinn<sup>6</sup>, Derek Blankenship<sup>7</sup>, Ranju Dhawan<sup>8</sup>, John J. Cush<sup>6</sup>, Asuncion Mejias<sup>10</sup>, Octavio Ramilo<sup>10</sup>, Onn M. Kon<sup>3</sup>, Virginia Pascual<sup>6</sup>, Jacques Banchereau<sup>6</sup>, Damien Chaussabel<sup>6</sup> & Anne O'Garra<sup>1</sup>

"Most people infected with *M. tuberculosis* remain asymptomatic, termed latent TB, with a 10% lifetime risk of developing active TB disease. Current tests, however, cannot identify which individuals will develop disease."

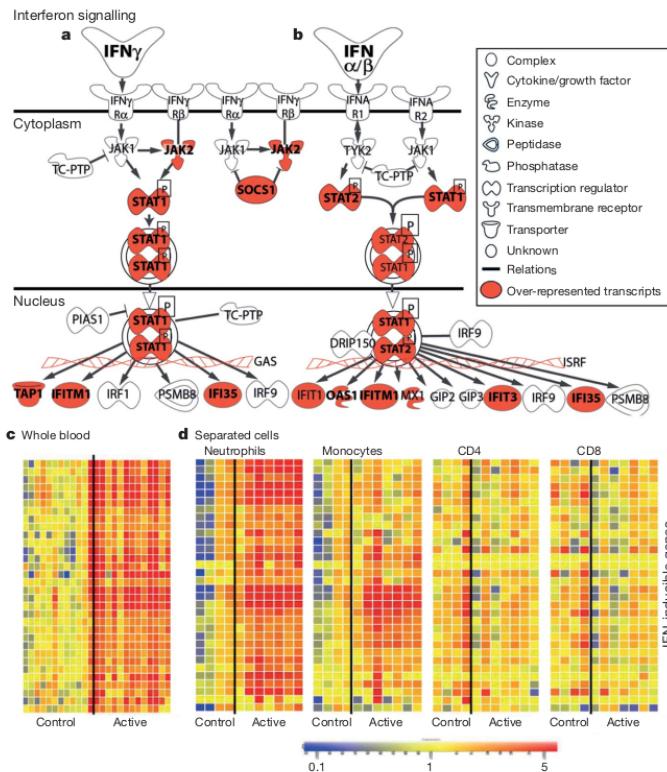
# Berry et al, 2010

- Here we identify a whole-blood 393 transcript signature for active TB in intermediate and high-burden settings, correlating with radiological extent of disease and reverting to that of healthy controls after treatment
- A subset of patients with latent TB had signatures similar to those in patients with active TB
- We also identify a specific 86-transcript signature that discriminates active TB from other inflammatory and infectious diseases

# Berry et al, 2010



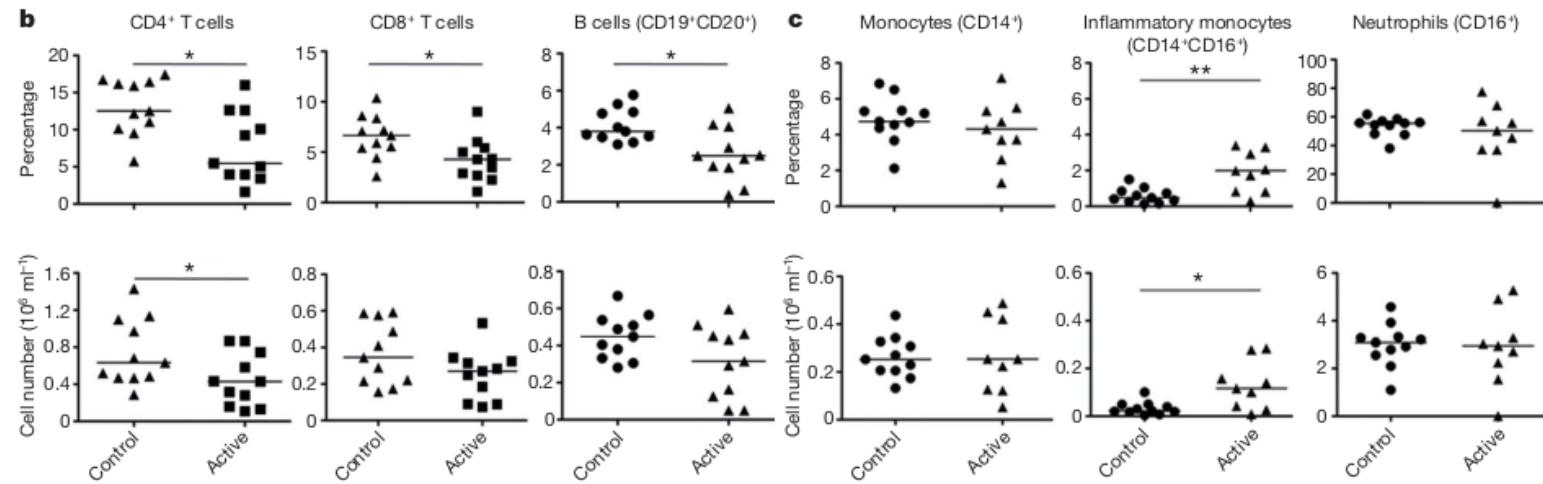
# Berry et al, 2010



**Figure 4 | Interferon-inducible gene expression in active TB.** Canonical pathway of Ingenuity pathways analysis for interferon signalling; symbol indicates gene function (legend on right). Transcripts over-represented in test set patients with active TB shaded red. **a**, Type II IFN- $\gamma$ . **b**, Type I IFN- $\alpha/\beta$

signalling. Transcript abundance of representative IFN-inducible genes in active TB from **(c)** whole blood and **(d)** separated blood leucocyte population. Transcript abundance/expression is normalized to the median of the healthy controls.

# Berry et al, 2010



# How do we know what's different

Main word behind comparative gene expression studies is **differential expression** (DE).

- Given two conditions  $A$  and  $B$
- We try to find genes for which we can statistically confirm that their average expression level in these two conditions is different

Briefly speaking, differential expression tests are statistical frameworks to identify genes of possible interest.

We say that gene is differentially expressed, when they have different average expression levels (between conditions).

# What else

- Previous example was about gene expression on a **gene level**
- We can go to **transcript level** and detect different isoforms

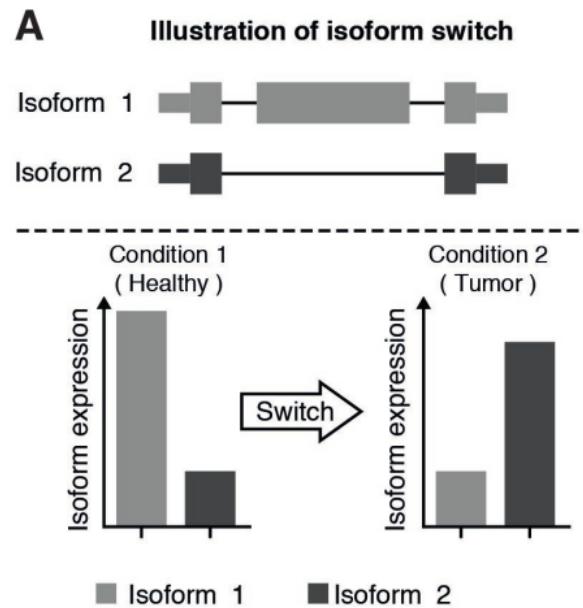
# Alternative splicing

# Terminology

- Alternative splicing, alternative transcription start- and termination sites
- All of the above we wall alternative splicing
- All of the above events expend RNA repertoire of most human genes

# Isoform switch

- Isoform switch is one of alternative splicing scenarios
- Under certain condition, gene is spliced differently



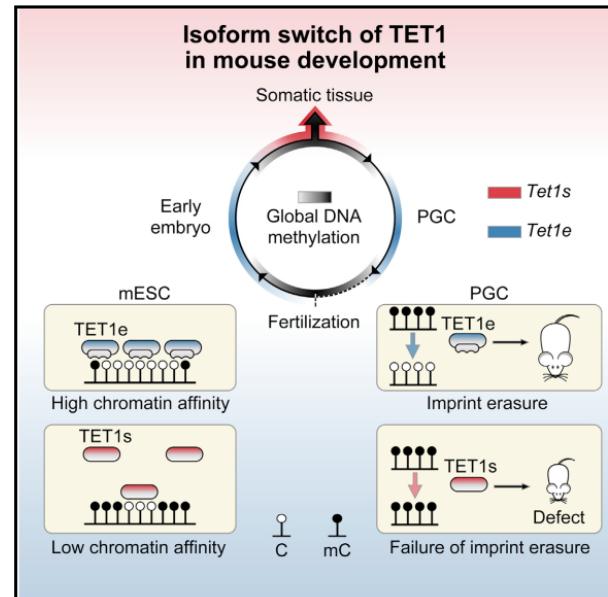
# Isoform switch

Article

## Molecular Cell

### Isoform Switch of TET1 Regulates DNA Demethylation and Mouse Development

#### Graphical Abstract



#### Authors

Wen-hao Zhang, Wei-kun Xia,  
Qiu-jun Wang, ..., Shaorong Gao,  
Yong-hui Jiang, Wei Xie

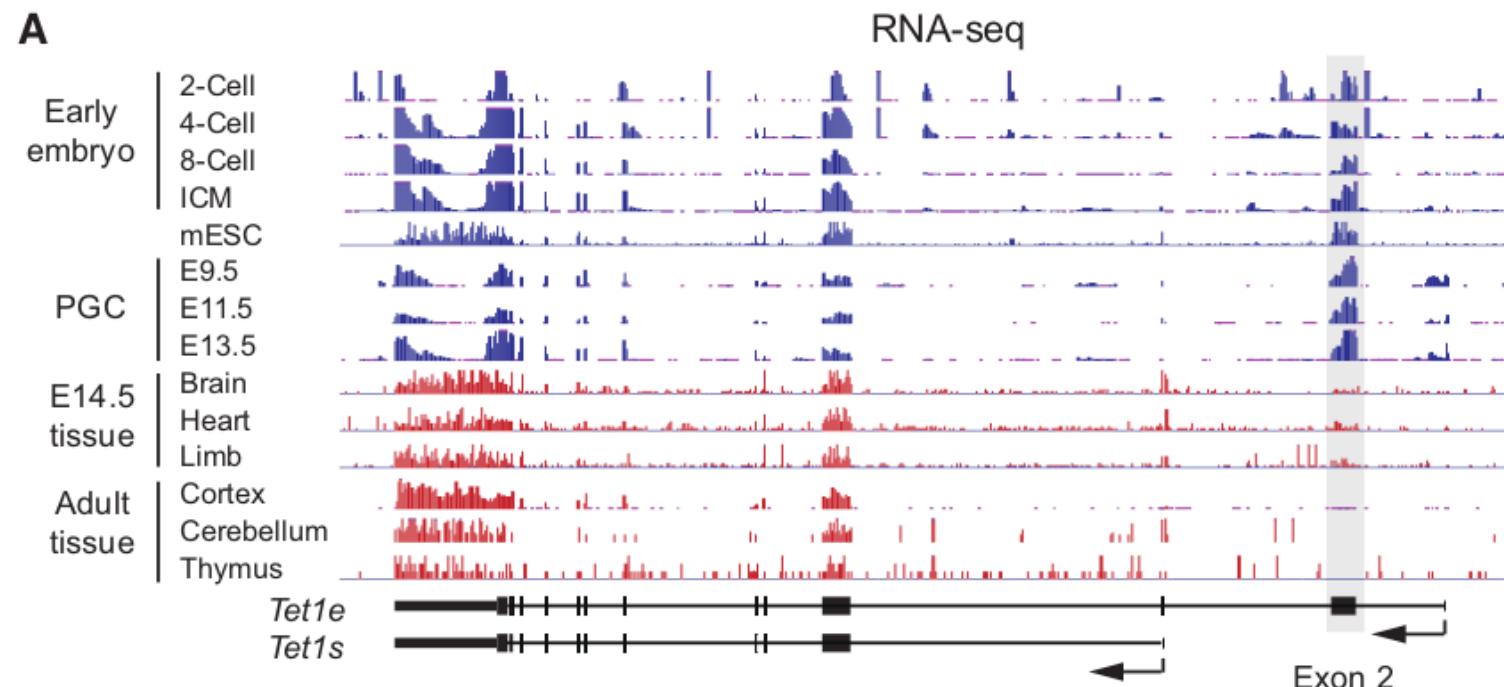
#### Correspondence

xiewei121@tsinghua.edu.cn

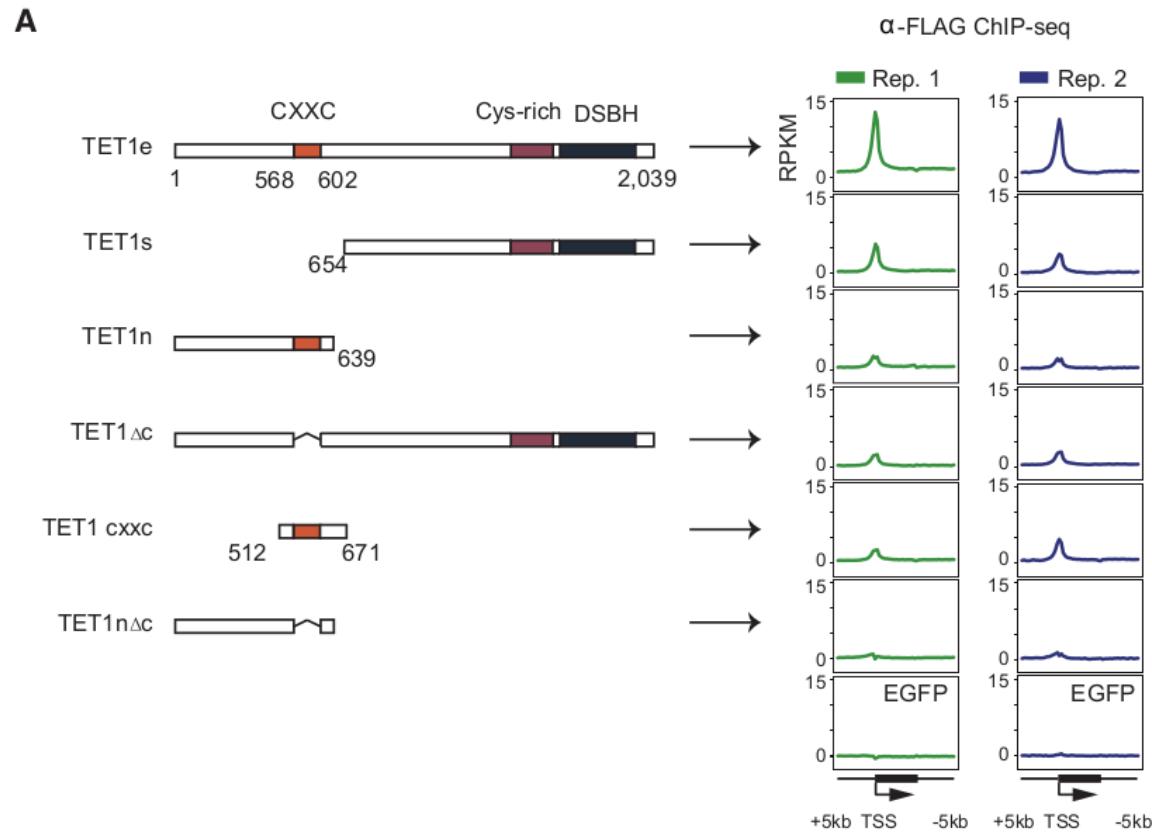
#### In Brief

TET proteins regulate DNA demethylation and development. Zhang et al. found two distinct isoforms of TET1 in mouse early embryos and somatic tissues. Exclusive expression of the short isoform showed defects in chromatin binding and imprint erasure, indicating that epigenetic memory can be regulated by a simple isoform switch.

# Isoform switch



# Isoform switch



# What else

- Mouse and human (I mostly work with) are very well annotated
- What if the organism of interest is not well-studied?

# Transcriptome assembly and functional annotation

# Transcriptome assembly

- RNA-Seq Assembly is an identification of all expressed isoforms from RNA-seq
- Reference-guided Transcriptome assembly
- De novo RNA-seq Assembly

# Functional annotation

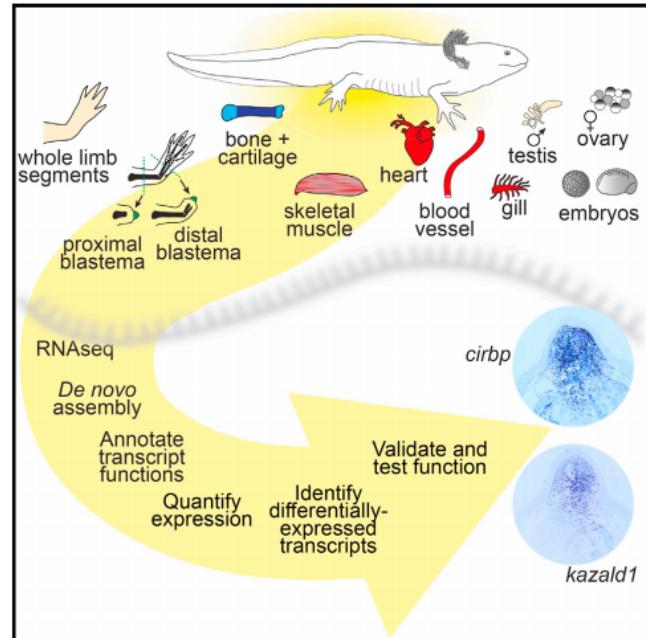
- Predicting gene/transcript function
- Homology search to known sequence data
- Protein domain identification
- Protein signal peptide and transmembrane domain prediction

# Example study

## Cell Reports

### A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors

#### Graphical Abstract



#### Authors

Donald M. Bryant, Kimberly Johnson, Tia DiTommaso, ..., Aviv Regev, Brian J. Haas, Jessica L. Whited

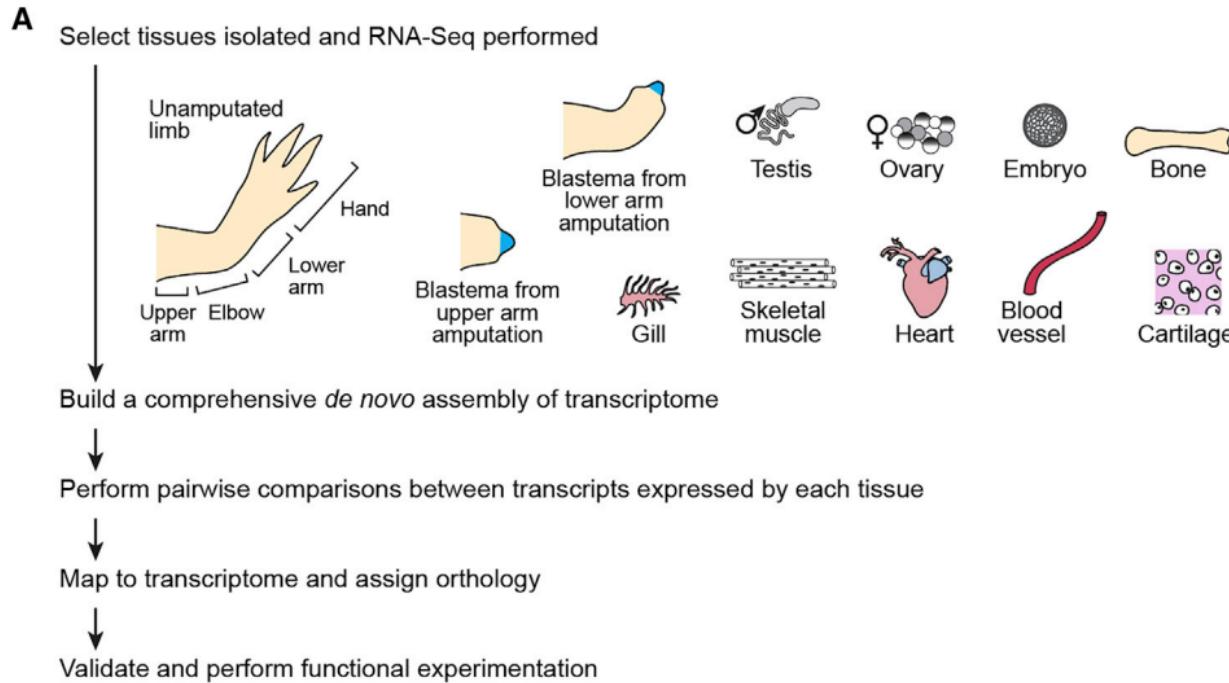
#### Correspondence

bhaas@broadinstitute.org (B.J.H.), jwhited@bwh.harvard.edu (J.L.W.)

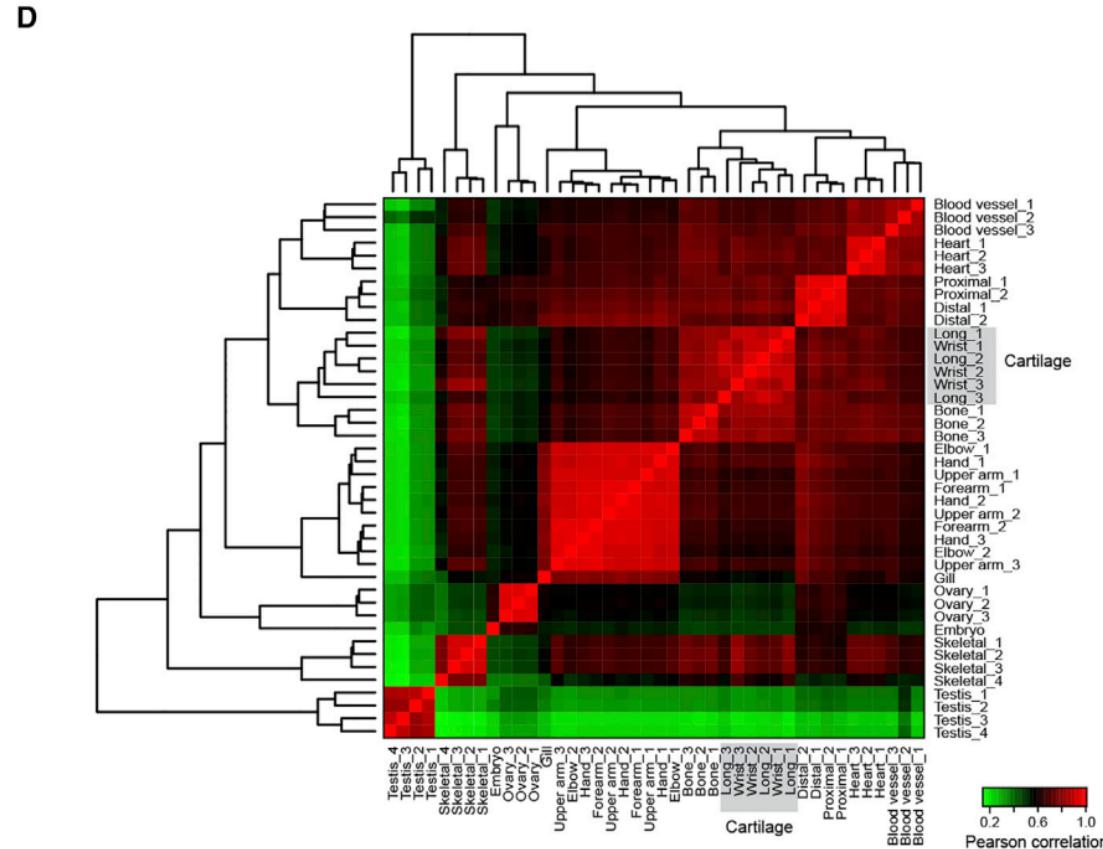
#### In Brief

Discovery of genes driving axolotl limb regeneration has been challenging, due to limited genomic resources. Bryant et al. have created a transcriptome with near-complete sequence information for most axolotl genes, identified transcriptional profiles that distinguish blastemas from differentiated limb tissues, and uncovered functional roles for *cirbp* and *kazald1* in limb regeneration.

# Example study



# Example study

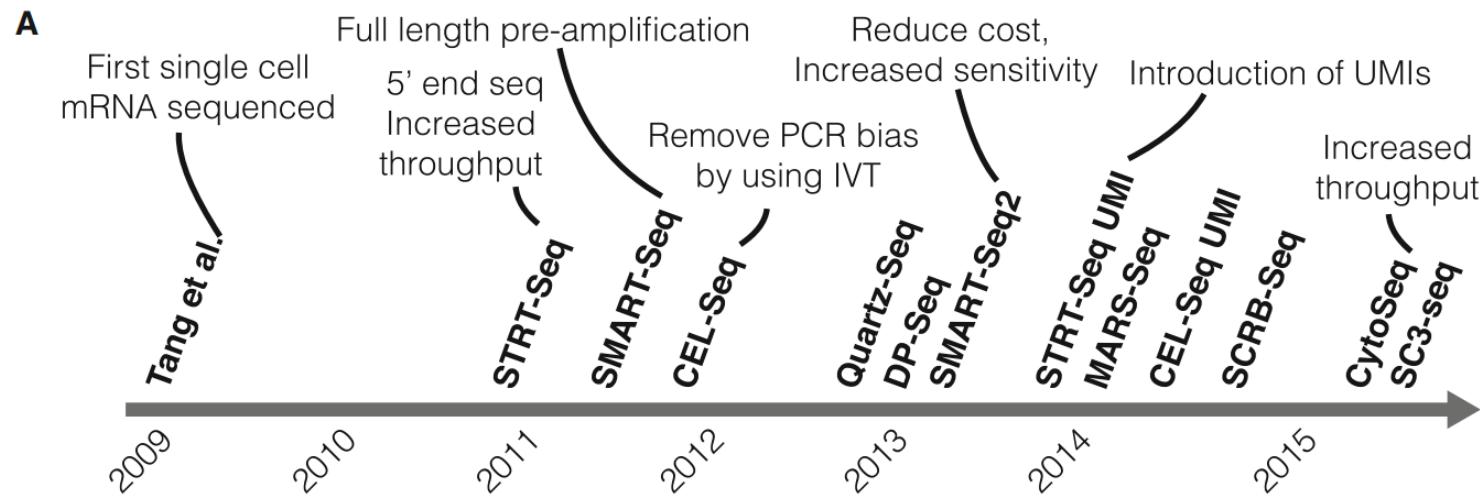


# Single-cell technologies

# Single-cell RNA-seq

- These days we can measure RNA abundance within single cells
- Better resolution: more samples (cells), better understanding where changes are coming from
- Worse resolution: we have much less RNA within a single cell than within a sample. We can capture even less. We can sequence even less

# Single-cell RNA-seq



# Single-cell RNA-seq

MICROPIPETTING MICROMANIPULATION	LASER CAPTURE MICRODISSECTION	FACS	MICRODROPLETS	MICROFLUIDICS e.g. FLUIDIGM C1
low number of cells	low number of cells	hundreds of cells	large number of cells	hundreds of cells
any tissue	any tissue	dissociated cells	dissociated cells	dissociated cells
enables selection of cells based on morphology or fluorescent markers	enables selection of cells based on morphology or fluorescent markers	enables selection of cells based on size or fluorescent markers	no selection of cells (can presort with FACS)	no selection of cells (can presort with FACS)
visualisation of cells	visualisation of cells	fluorescence and light scattering measurements	fluorescence detection	visualisation of cells
time consuming	time consuming	fast	fast	fast
reaction in microliter volumes	reaction in microliter volumes	reaction in microliter volumes	reaction in nanoliter volumes	reaction in nanoliter volumes

# Single-cell RNA-seq

## RESEARCH ARTICLE

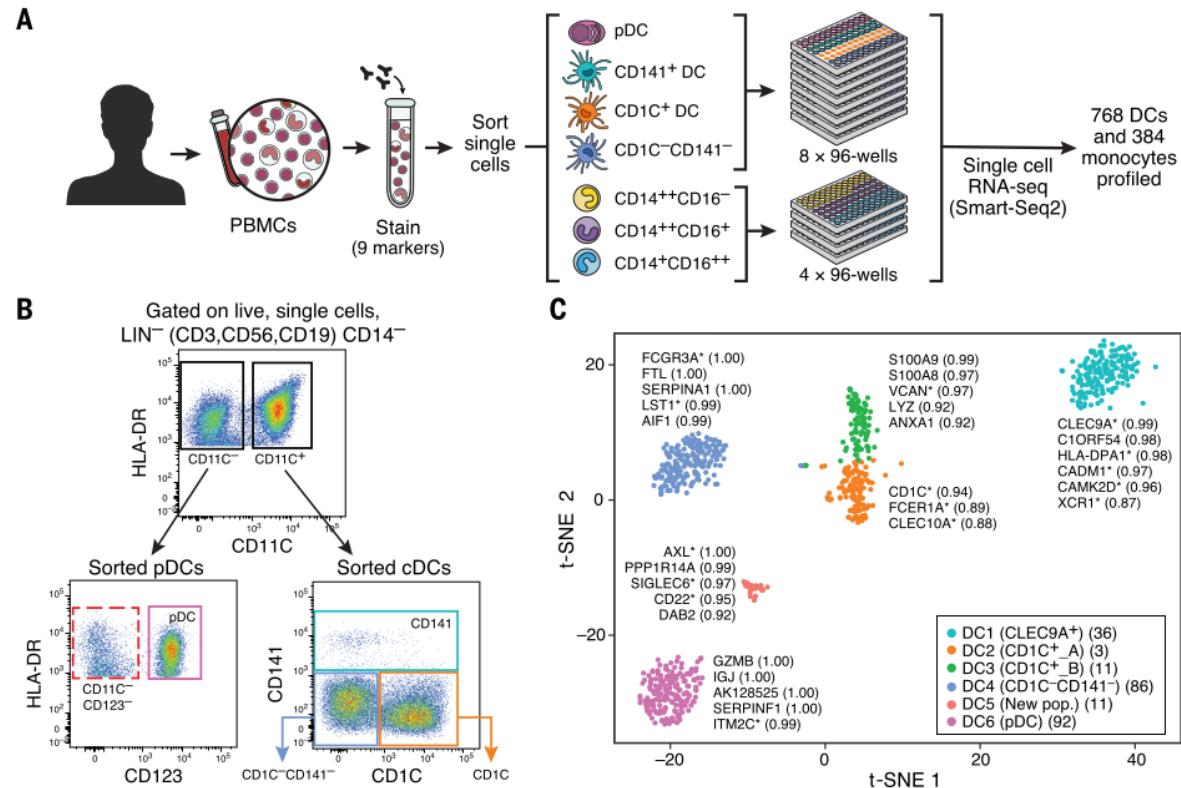
### IMMUNOGENOMICS

# Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors

Alexandra-Chloé Villani,<sup>1,2\*</sup>† Rahul Satija,<sup>1,3,4\*</sup> Gary Reynolds,<sup>5</sup> Siranush Sarkizova,<sup>1</sup> Karthik Shekhar,<sup>1</sup> James Fletcher,<sup>5</sup> Morgane Griesbeck,<sup>6</sup> Andrew Butler,<sup>3,4</sup> Shiwei Zheng,<sup>3,4</sup> Suzan Lazo,<sup>7</sup> Laura Jardine,<sup>5</sup> David Dixon,<sup>5</sup> Emily Stephenson,<sup>5</sup> Emil Nilsson,<sup>8</sup> Ida Grundberg,<sup>8</sup> David McDonald,<sup>5</sup> Andrew Filby,<sup>5</sup> Weibo Li,<sup>1,2</sup> Philip L. De Jager,<sup>1,9</sup> Orit Rozenblatt-Rosen,<sup>1</sup> Andrew A. Lane,<sup>1,7</sup> Muzlifah Haniffa,<sup>5,10</sup>† Aviv Regev,<sup>1,11,12</sup>† Nir Hacohen<sup>1,2</sup>†

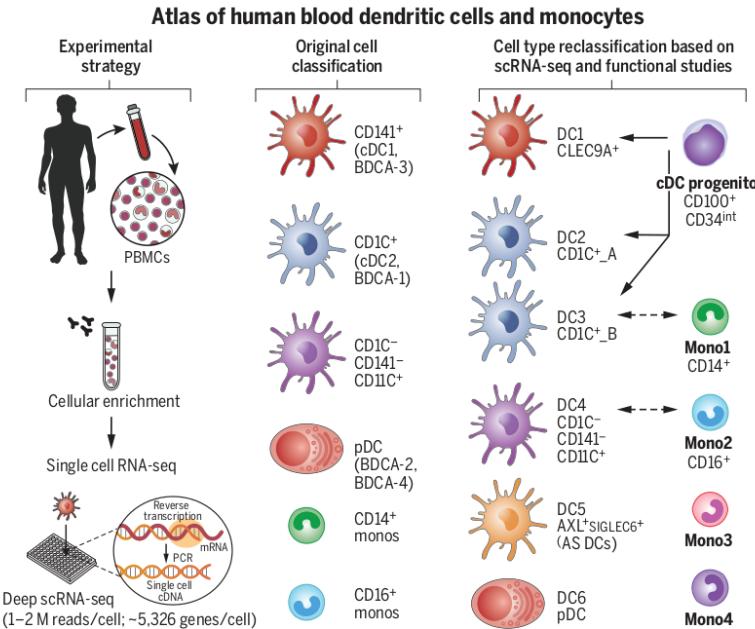
Dendritic cells (DCs) and monocytes play a central role in pathogen sensing, phagocytosis, and antigen presentation and consist of multiple specialized subtypes. However, their identities and interrelationships are not fully understood. Using unbiased single-cell RNA sequencing (RNA-seq) of ~2400 cells, we identified six human DCs and four monocyte subtypes in human blood. Our study reveals a new DC subset that shares properties with plasmacytoid DCs (pDCs) but potently activates T cells, thus redefining pDCs; a new subdivision within the CD1C<sup>+</sup> subset of DCs; the relationship between blastic plasmacytoid DC neoplasia cells and healthy DCs; and circulating progenitor of conventional DCs (cDCs). Our revised taxonomy will enable more accurate functional and developmental analyses as well as immune monitoring in health and disease.

# Single-cell RNA-seq



# Single-cell RNA-seq

- Performed scRNA-seq on myeloid blood cells
- Compared existed classification with their findings
- Identified new subsets



**Establishing a human blood monocyte and dendritic cell atlas.** We isolated ~2400 cells enriched from the healthy human blood lineage<sup>-</sup> HLA-DR<sup>+</sup> compartment and subjected them to single-cell RNA sequencing. This strategy, together with follow-up profiling and functional and phenotypic characterization, led us to update the original cell classification to include six DCs, four monocyte subtypes, and one conventional DC progenitor.

# Structure of the course

# Day 1: Transcription and regulation of transcription

- **Alexander Tkachenko** will walk you through transcription and remind you the biology behind transcription and how transcription can be regulated.

# Day 2: Microarray and gene expression studies

- Microarrays and how they measure RNA abundance
- You will be introduced to basic concepts and ideas in gene expression studies
- Secondary analysis of gene expression datasets
- **Homework 1:** Analysis of microarray dataset

## Day 3: RNA-seq

- Sequencing of RNA
- Alignment / quantification
- Normalization and differential expression
- **Homework 2:** Alignment and quantification of RNA-seq dataset with further downstream analysis

# Day 4: Different topics

- Non-model species: transcriptome assembly
- Gene expression deconvolution
- Experimental design of RNA-seq studies
- **Homework 3:** Transcriptome assembly

# Day 5: Single-cell transcriptomics

- Single-cell RNA-seq
- Methods in scRNA-seq
- Visual data analysis
- **Homework 4:** Analysis of scRNA-seq dataset

# Why this way?

- We will cover things mostly in chronological order
- Basic models and frameworks will get more complicated as we get through

# Grading the course

Homeworks (during this week):

- 4 homeworks (50 points)

Examination is somewhere in May:

- Dataset processing as a part of examination (25 points)
- Oral examination (25 points)

# Grading the course

- $\geq 50$  is E (all homeworks will give you E)
- $\geq 60$  is D
- $\geq 70$  is C
- $\geq 80$  is B
- $\geq 90$  is A (that's where you aim)

# Any questions ?