ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Introduction to gene expression

Konstantin Zaitsev

March 22th, 2021

About the course

About the course

Authors of this course:

1. 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

Link to the Zoom meetings:
https://us02web.zoom.us/j/89812669912?pwd=ckJSZWk5bm5CWi9weWlabVgvaDl3dz09

Most important link:
https://drive.google.com/drive/folders/1DHdQIAzXjxVK47v1qF5ZozlZe5c5LG6_?usp=sharing

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 (?))

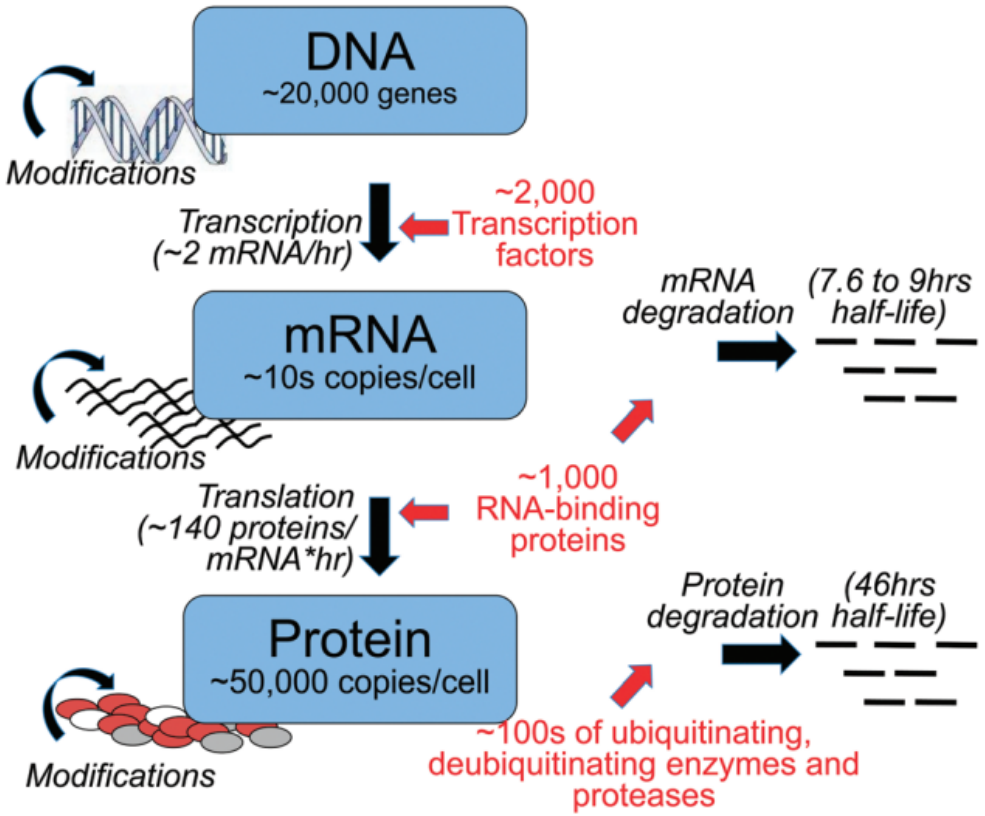
https://drive.google.com/drive/folders/1DHdQIAzXjxVK47v1qF5ZozlZe5c5LG6_?usp=sharing

About the course: syllabus

- Central dogma of molecular biology, structure of gene, types of RNA, structure of RNA, transcription, reverse transcription, 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



Taken from https://www.ncbi.nlm.nih.gov/pubmed/26259698

Why we measure RNA abundance

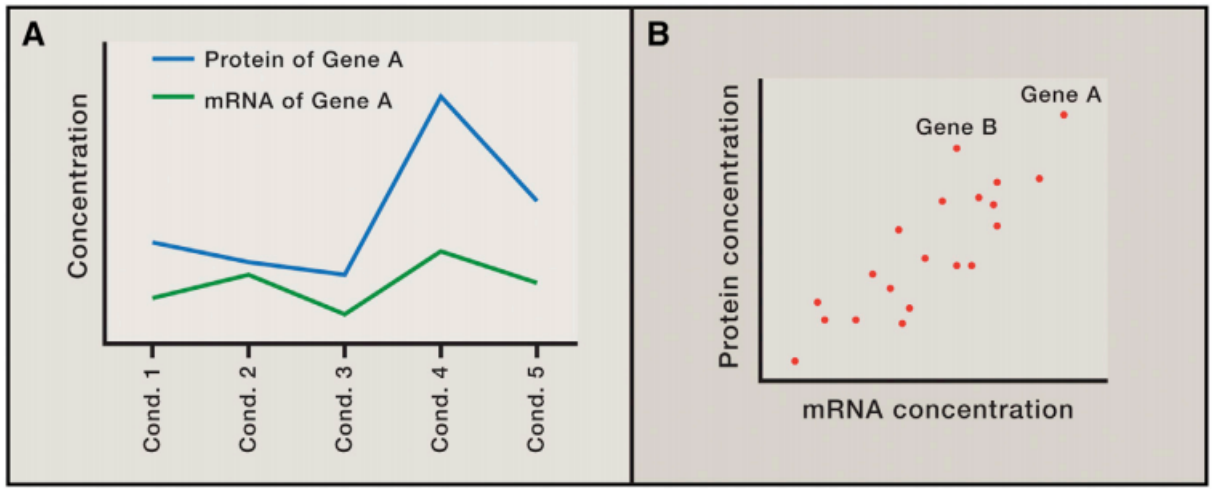


Figure 1. Different Types of Correlations between Protein and mRNA Levels Need to Be Distinguished

(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

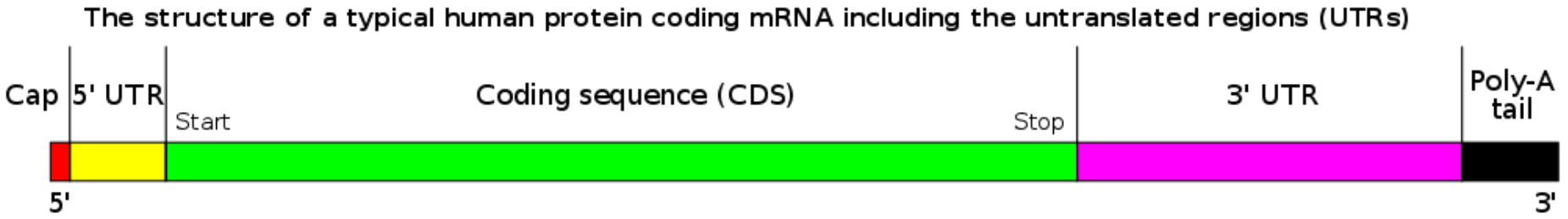
Taken from https://www.cell.com/cell/pdf/S0092-8674(16)30270-7.pdf

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



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129258/ and https://bionumbers.hms.harvard.edu/bionumber.aspx?id=111220

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

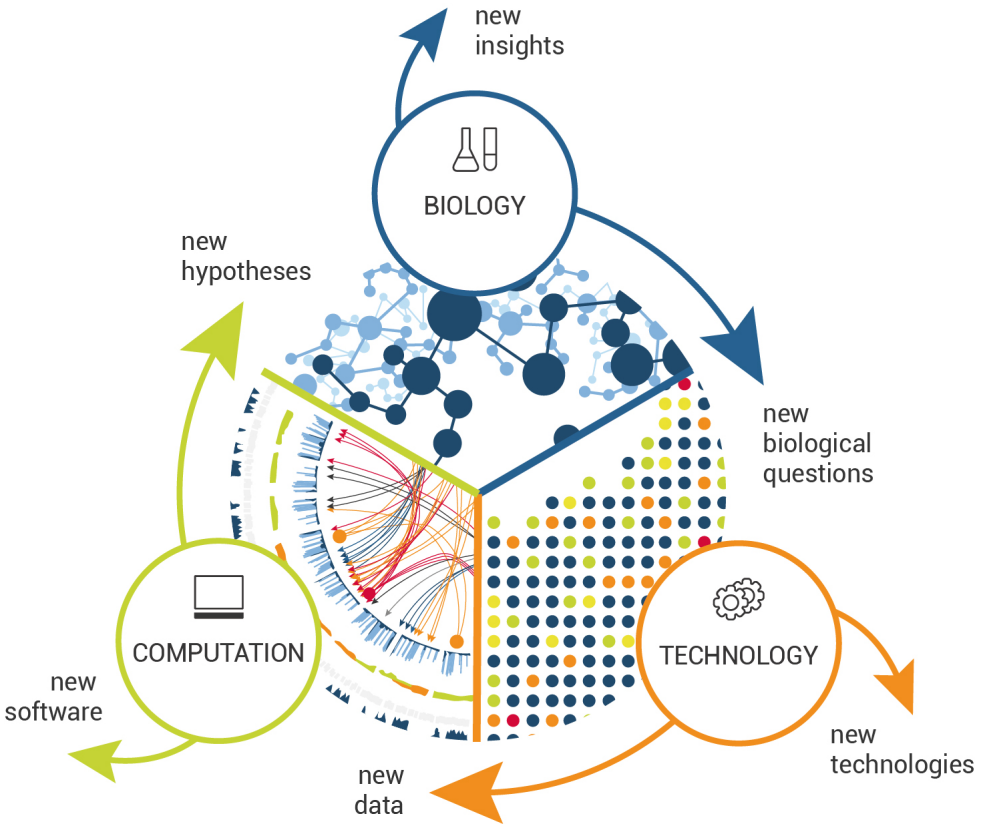
Strategy	Type of RNA	Ribosomal RNA content	Unprocessed RNA content	Genomic DNA content	Isolation method
Total RNA	All	High	High	High	None
PolyA selection	Coding	Low	Low	Low	Hybridization with poly(dT) oligomers
rRNA depletion	Coding, noncoding	Low	High	High	Removal of oligomers complementary to rRNA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

Systems biology approach



https://www.omicscouts.com/en/disease-and-systems-biology.html

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

Berry et al, 2010

LETTERS

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

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

"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."

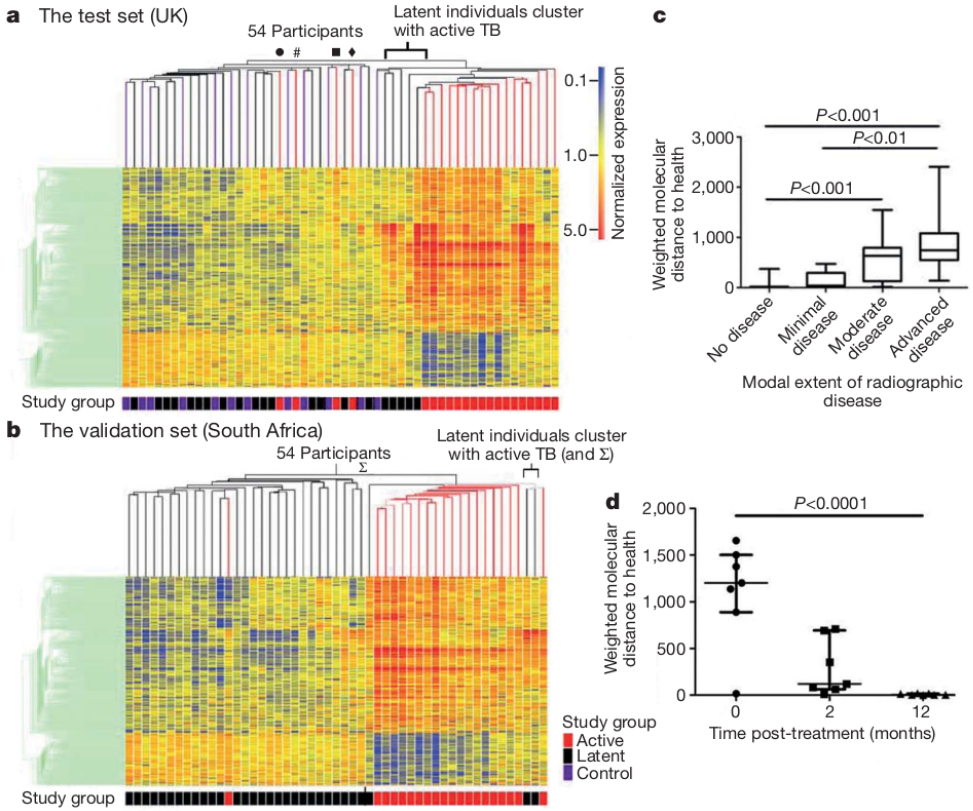
Taken from https://www.nature.com/articles/nature09247

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

Taken from https://www.nature.com/articles/nature09247

Berry et al, 2010



Taken from https://www.nature.com/articles/nature09247

Berry et al, 2010

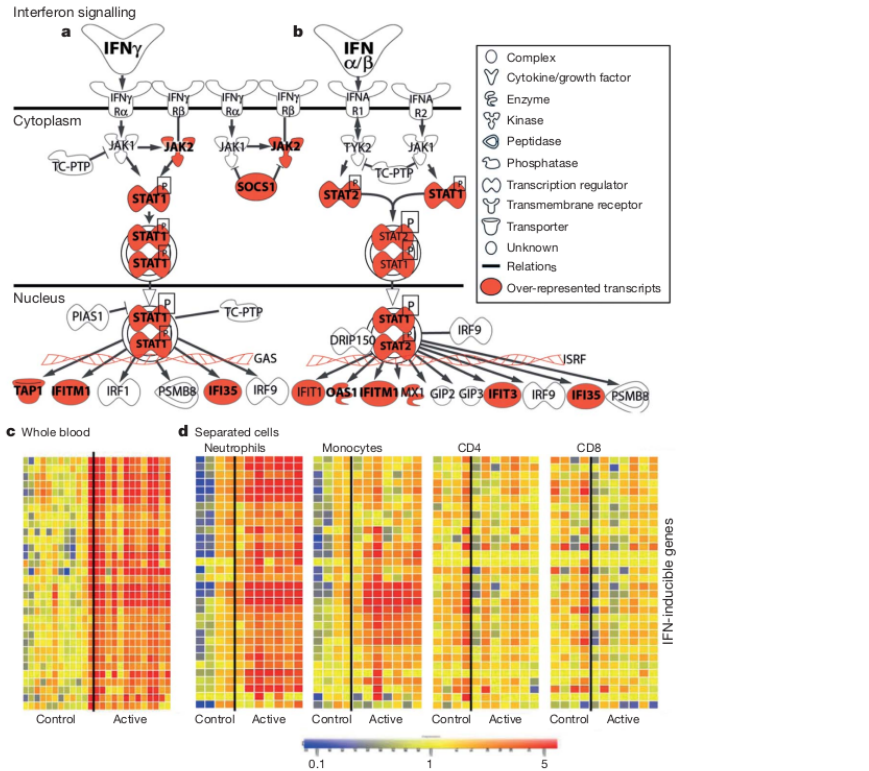
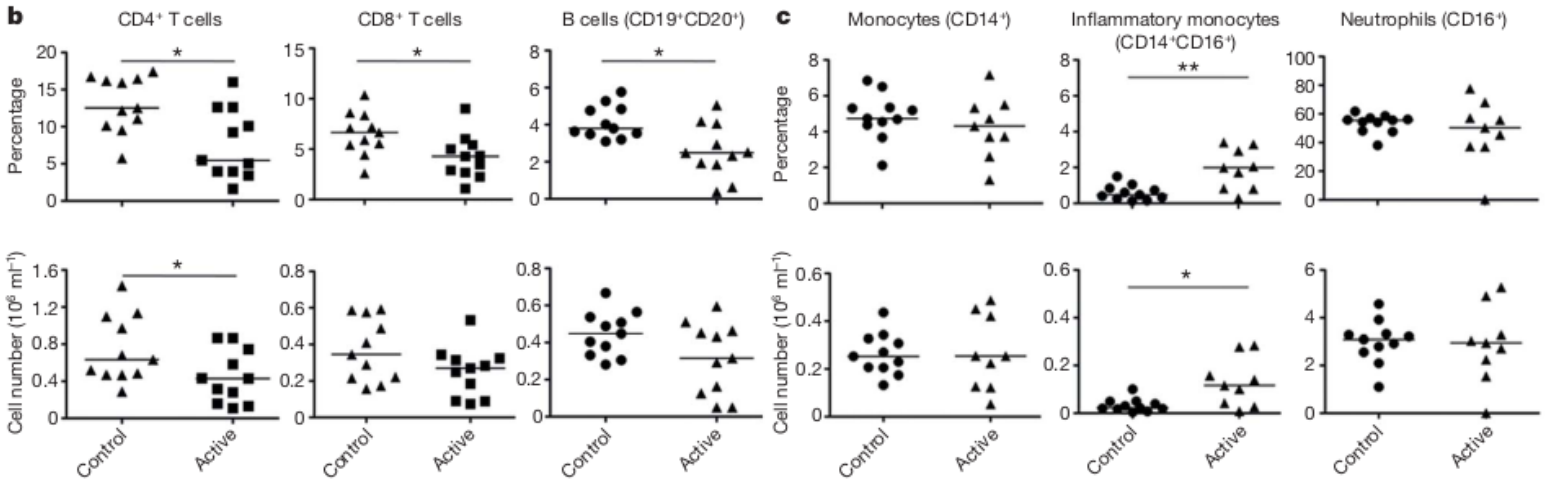


Figure 4 | Interferon-inducible gene expression in active TB. Canonical pathway of Ingenuity pathways analysis for interferon signalling; symbol indicates gene function (legend on right). Transcripts over-represented in test set patients with active TB shaded red. **a**, Type II IFN-γ. **b**, Type I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

Taken from https://www.nature.com/articles/nature09247

Berry et al, 2010



Taken from https://www.nature.com/articles/nature09247

How do we know what's diffent

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 levels in these two conditions are 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

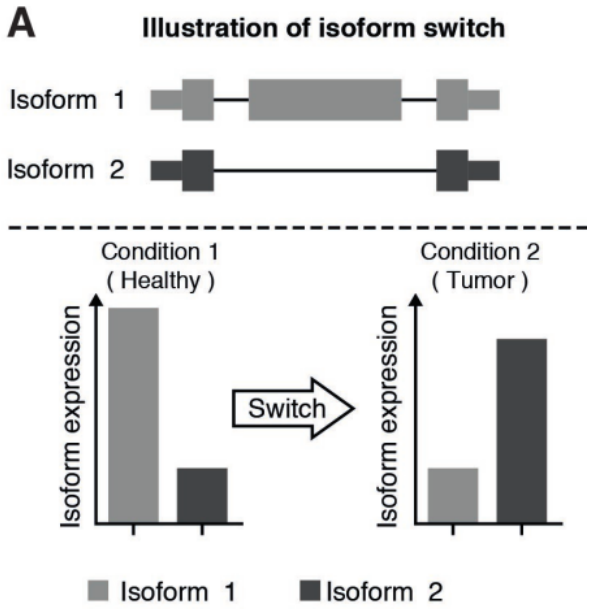
Alternative splicing

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



Taken from https://mcr.aacrjournals.org/content/15/9/1206.figures-only

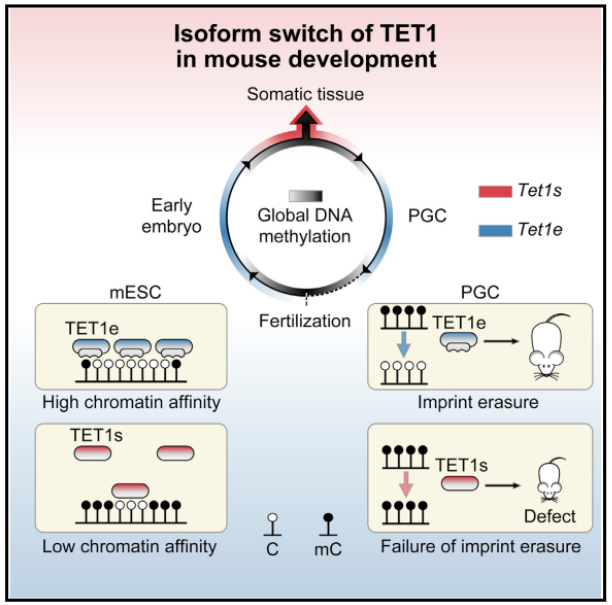
Isoform switch

Article

Molecular Cell

Isoform Switch of TET1 Regulates DNA Demethylation and Mouse Development

Graphical Abstract



Authors

Wenhao Zhang, Weikun Xia, Qiujun Wang, ..., Shaorong Gao, Yong-hui Jiang, Wei Xie

Correspondence

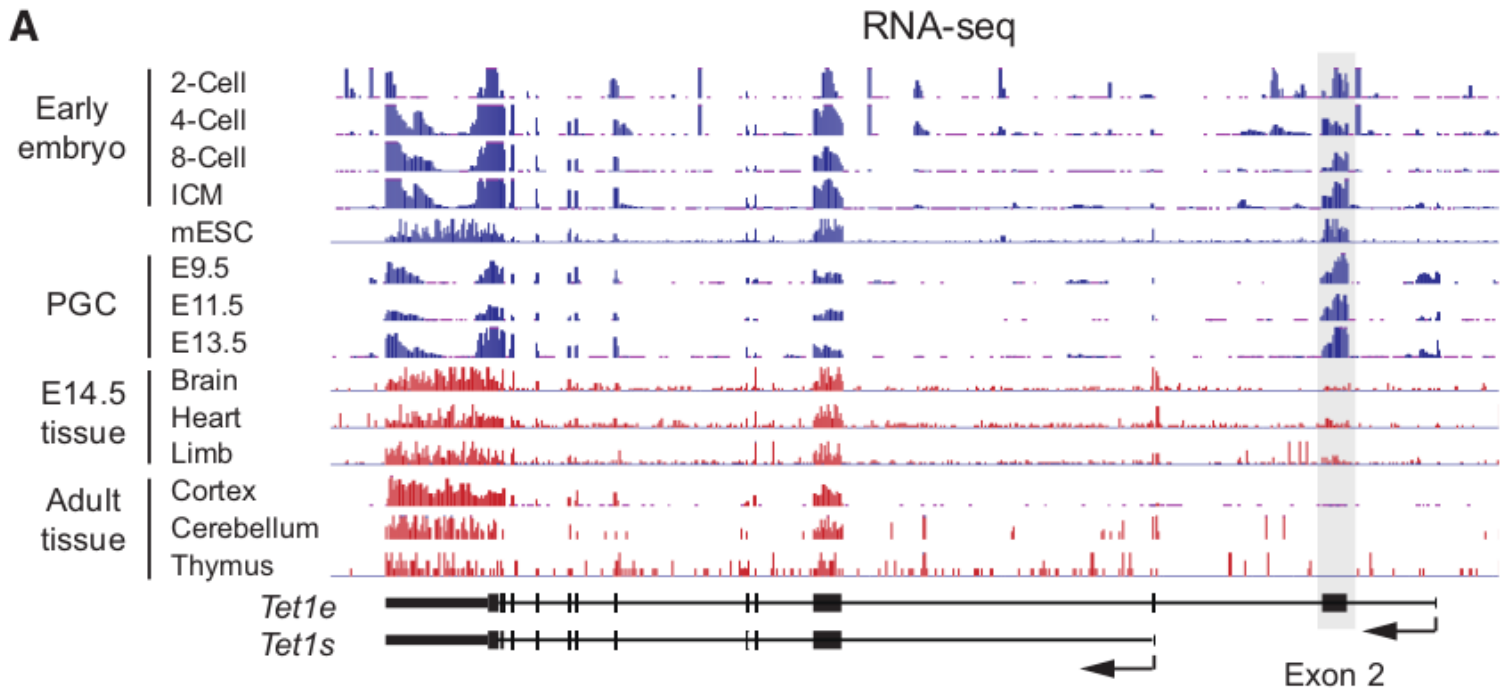
xiewei121@tsinghua.edu.cn

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.

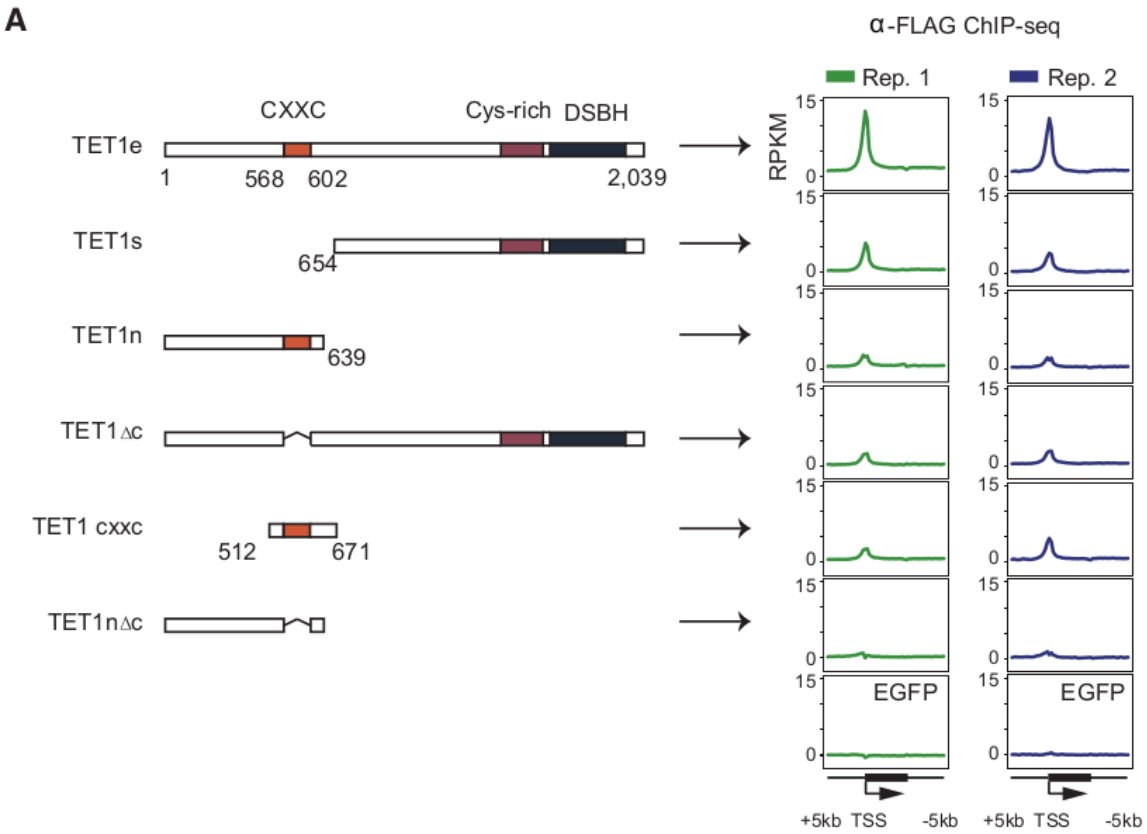
Taken from https://www.ncbi.nlm.nih.gov/pubmed/27916660

Isoform switch



Taken from https://www.ncbi.nlm.nih.gov/pubmed/27916660

Isoform switch



Taken from https://www.ncbi.nlm.nih.gov/pubmed/27916660

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

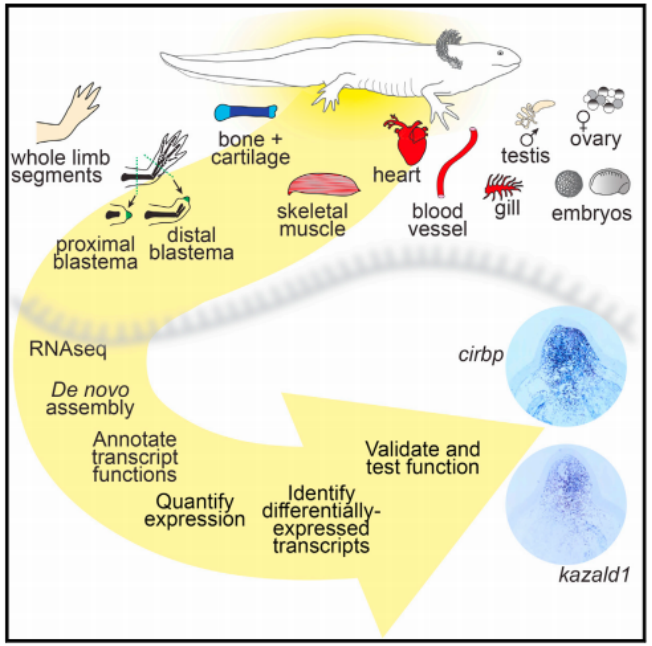
- Predicting gene/transcript function
- Homology search to known sequence data
- Protein domain identification
- Protein signal peptide and transmembrane domain prediction

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

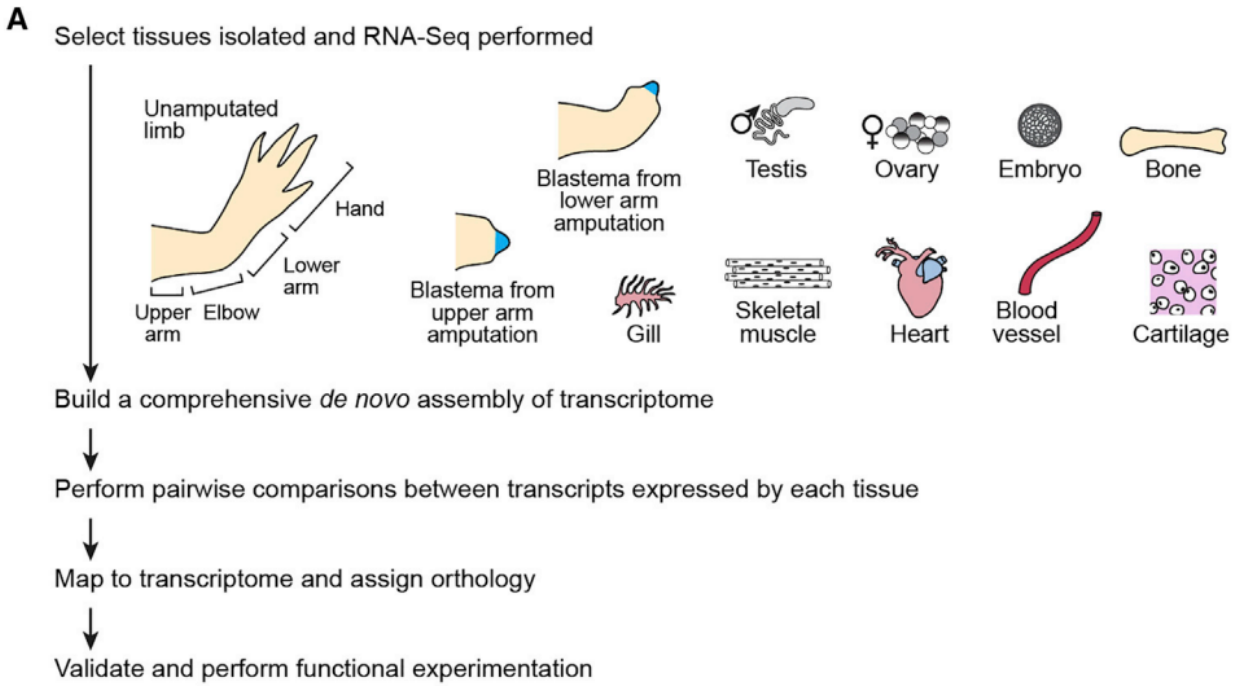
Correspondence

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

In Brief

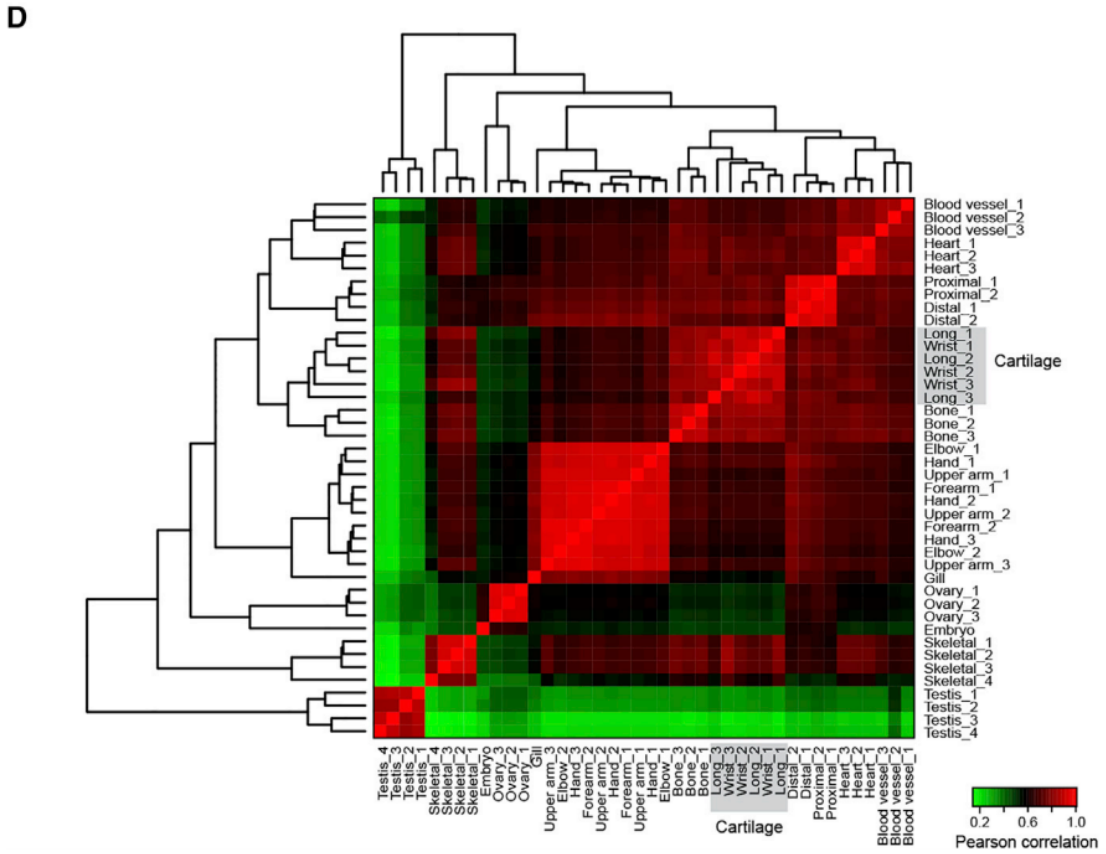
Discovery of genes driving axolotl limb regeneration has been challenging, due to limited genomic resources. Bryant et al. have created a transcriptome with near-complete sequence information for most axolotl genes, identified transcriptional profiles that distinguish blastemas from differentiated limb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

Example study



Taken from https://www.cell.com/action/showPdf?pii=S2211-1247%2816%2931770-3

Example study



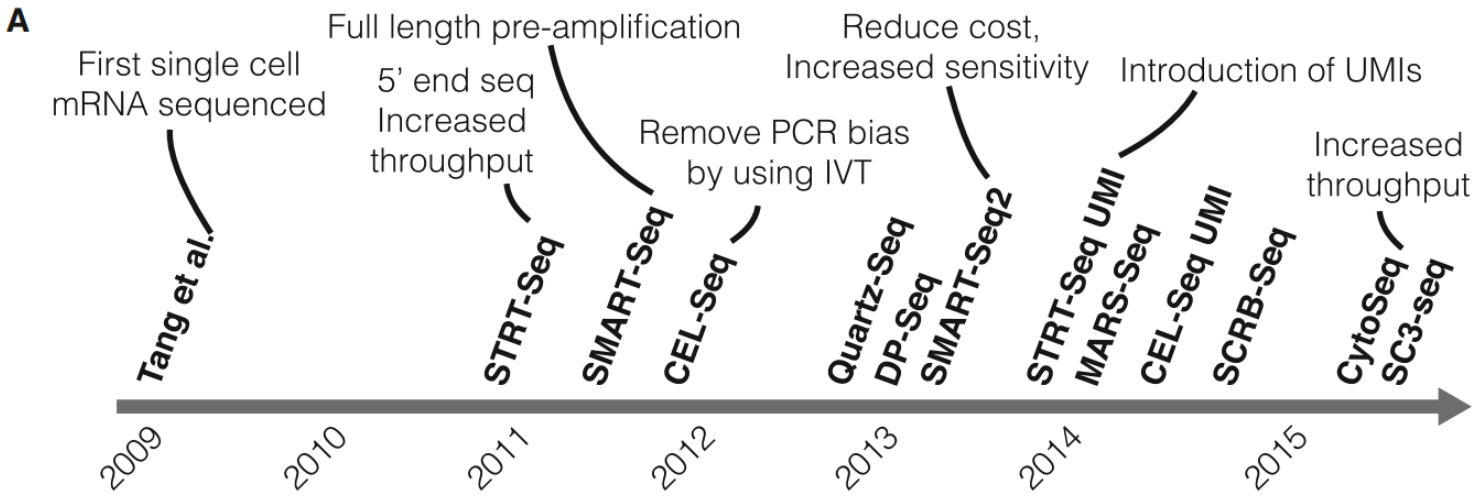
Taken from https://www.cell.com/action/showPdf?pii=S2211-1247%2816%2931770-3

Single-cell technologies

Single-cell RNA-seq

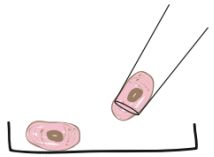
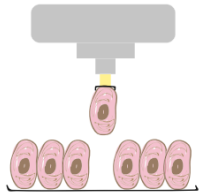
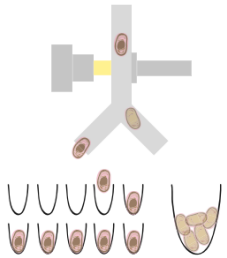
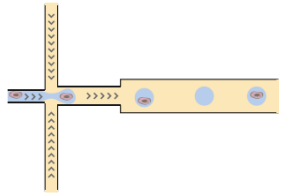
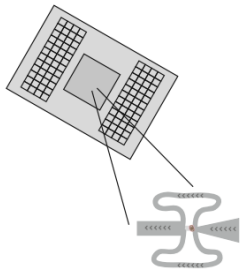
- These days we can measure RNA abundance within single cells
- Better resolution: more samples (cells), better understanding where changes are coming from
- Worse resolution: we have much less RNA within a single cell than within a sample. We can capture even less. We can sequence even less

Single-cell RNA-seq



Taken from https://www.cell.com/action/showPdf?pii=S1097-2765%2815%2900261-0

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

Taken from https://www.cell.com/action/showPdf?pii=S1097-2765%2815%2900261-0

Single-cell RNA-seq

RESEARCH ARTICLE

IMMUNOGENOMICS

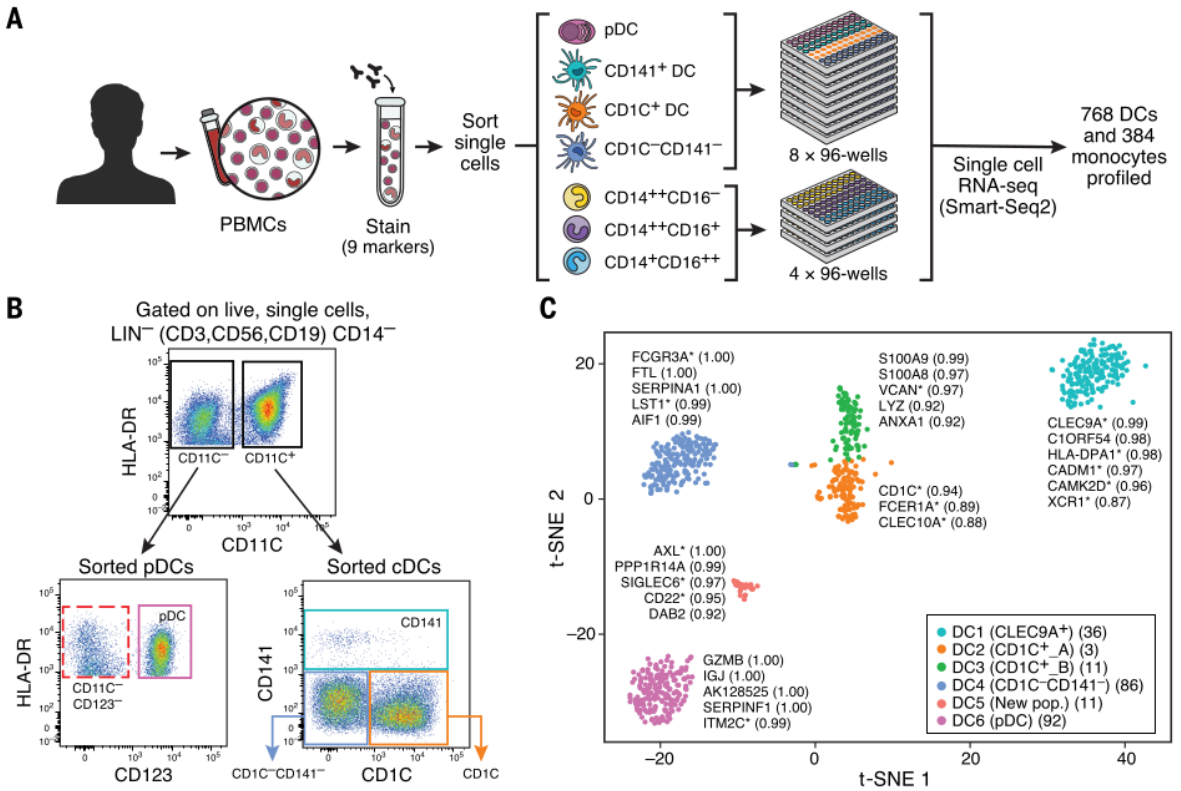
Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors

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

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$^+$ 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.

Taken from https://science.sciencemag.org/content/356/6335/eaah4573

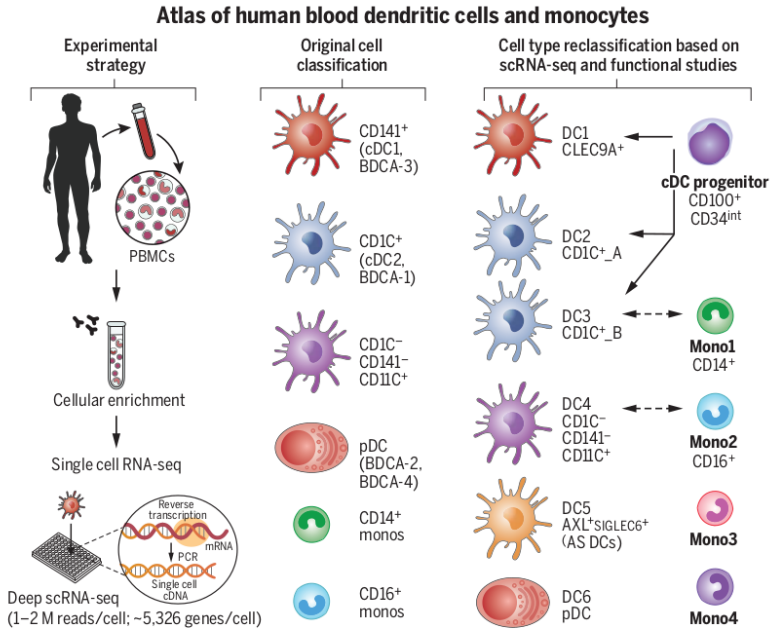
Single-cell RNA-seq



Taken from https://science.sciencemag.org/content/356/6335/eaah4573

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$^-$ HLA-DR$^+$ 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.

Taken from https://science.sciencemag.org/content/356/6335/eaah4573

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

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

Any questions ?