Геномные Анализы Индивидуальных Клеток

Пётр Харченко
Department of Biomedical Informatics,
Harvard Medical School

July 20th, 2015.

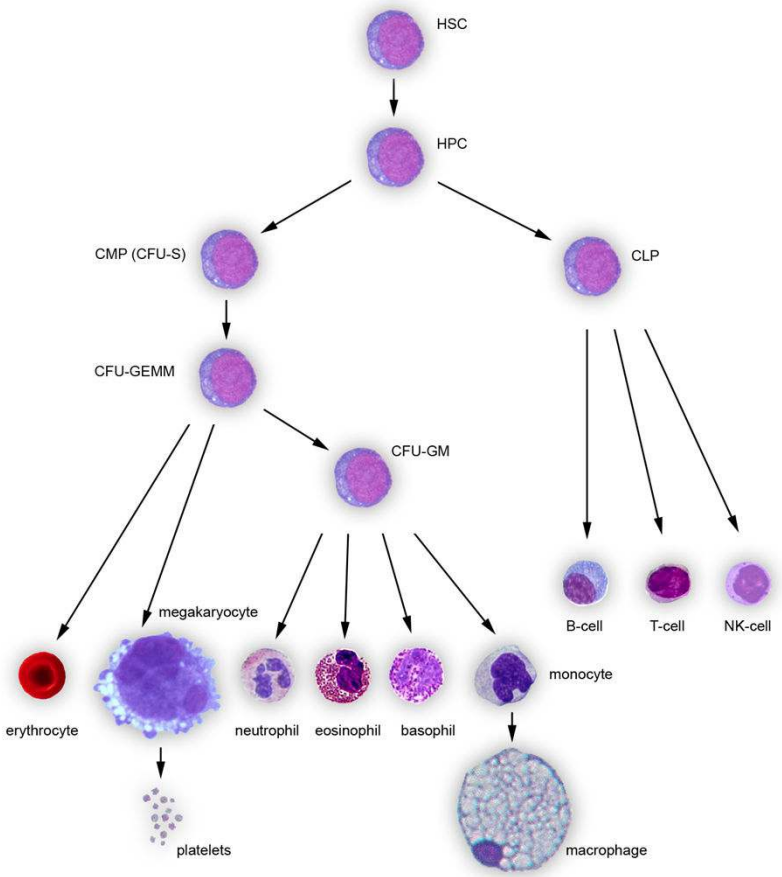
cbmi
Center for Biomedical Informatics

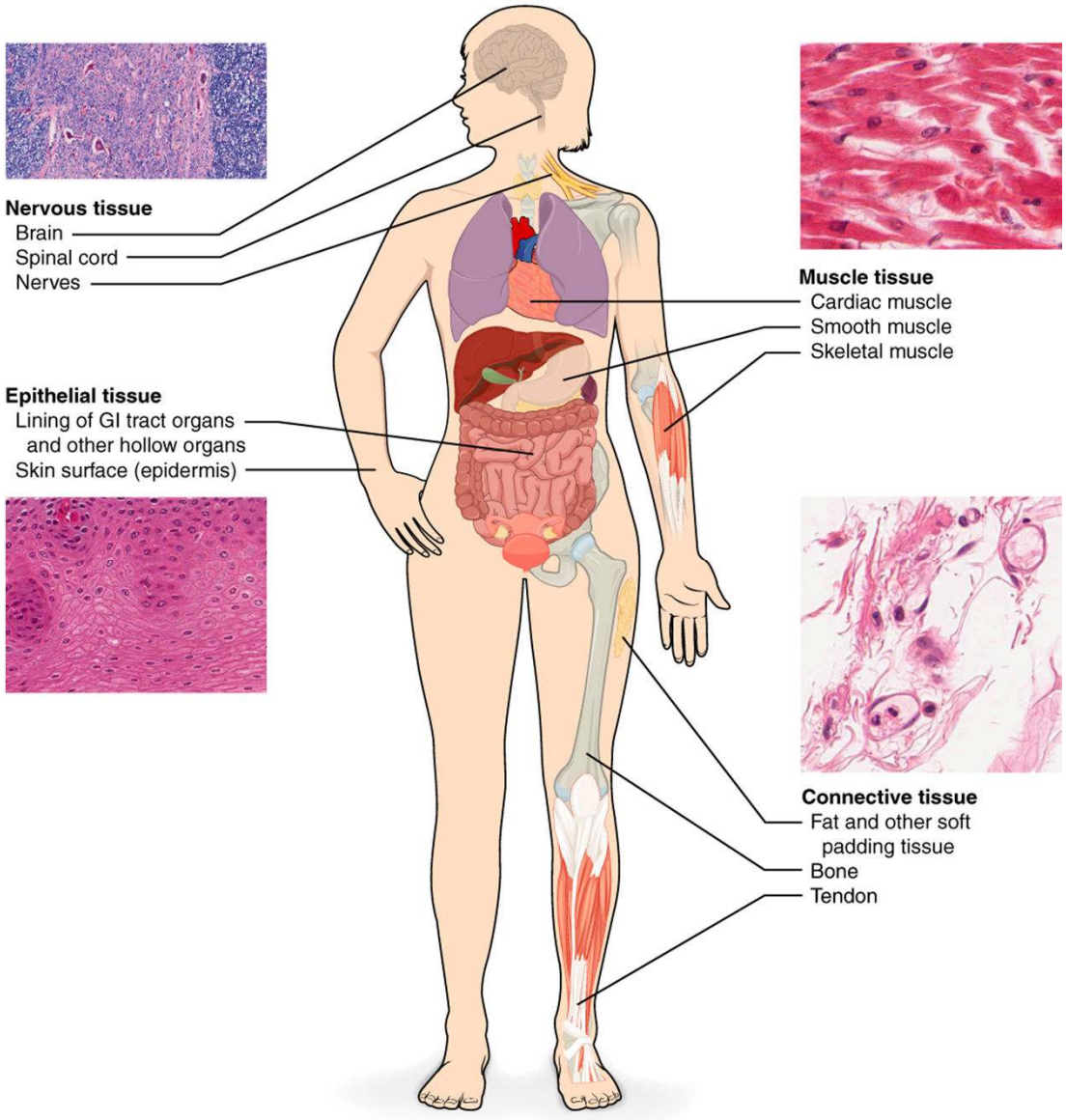
HSCI
HARVARD STEM CELL INSTITUTE

VE RI TAS

Клетки, Клеточные Типы и Ткани

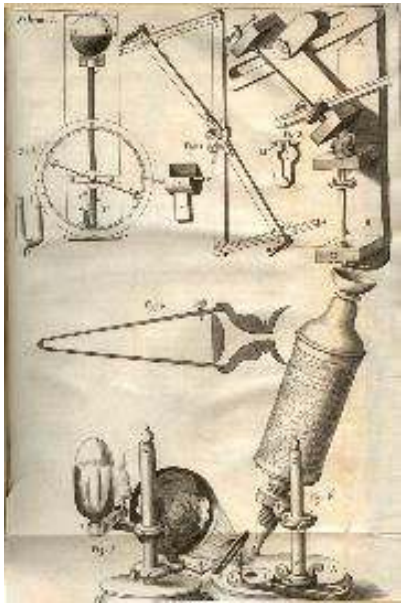
- Averages of ‘bulk’ samples
  - ~10^7 cells
- Complex tissue organization

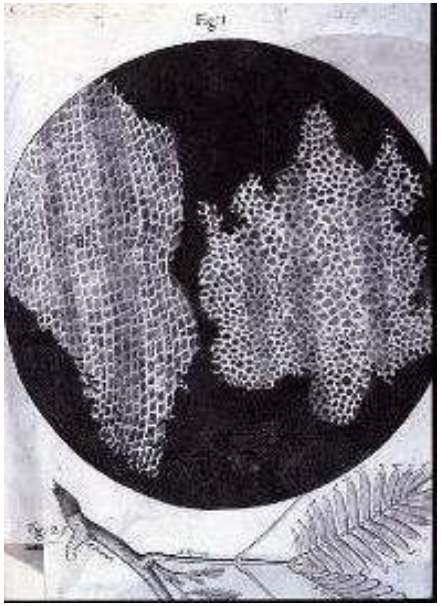


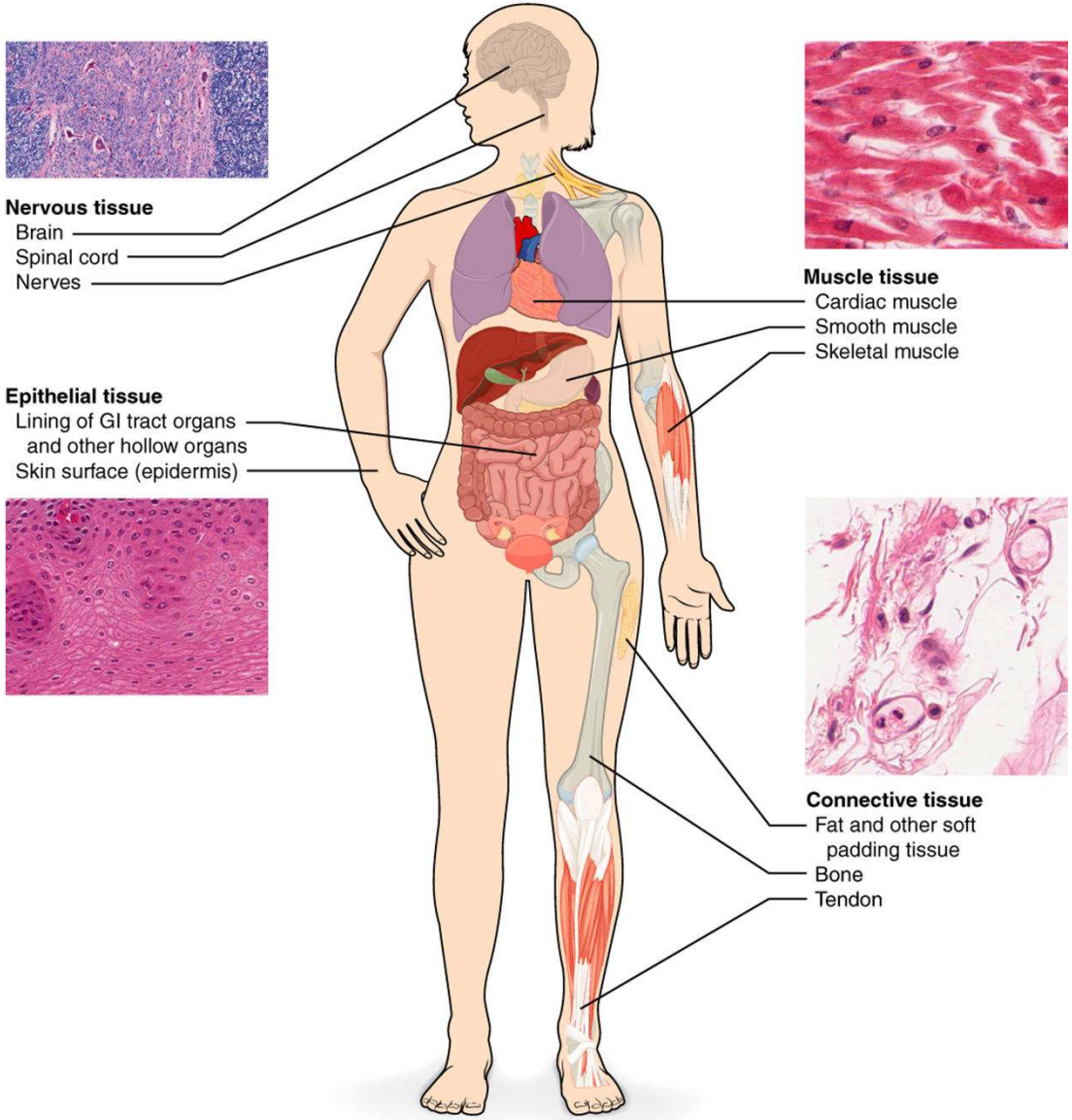


Клетки, Клеточные Типы и Ткани

- Averages of ‘bulk’ samples
  - ~10^7 cells
- Complex tissue organization
- Cell Types
  - Early Microscopy
  - Robert Hooke (1635-1703)

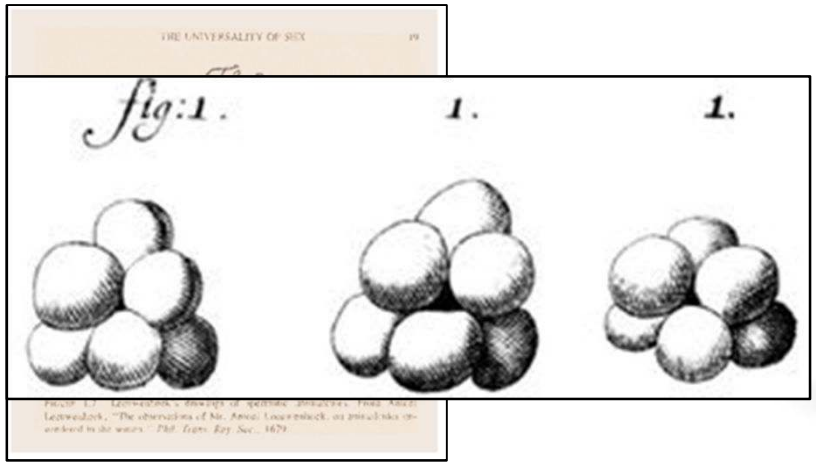


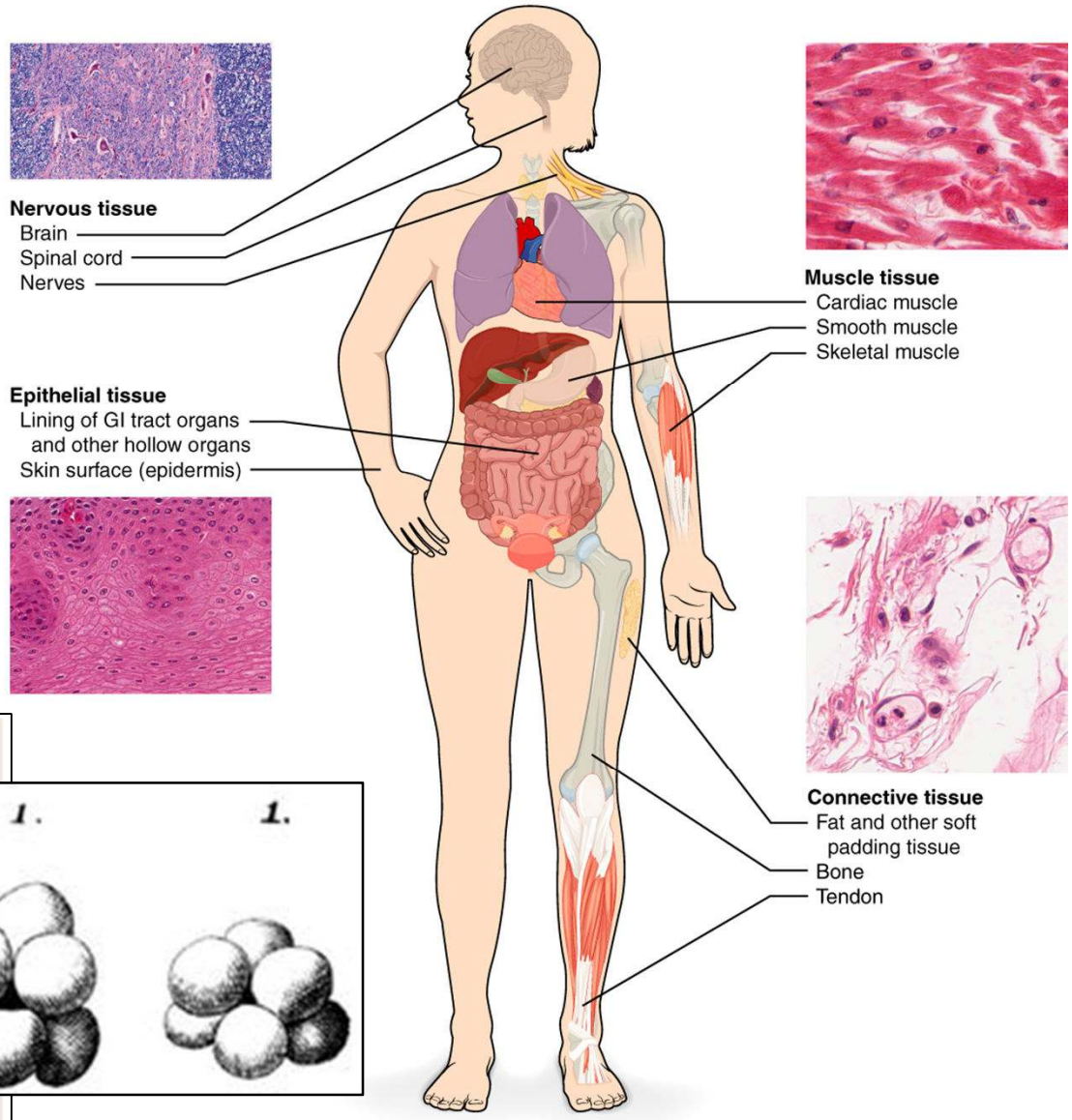




Клетки, Клеточные Типы и Ткани

- Averages of ‘bulk’ samples
  - ~10^7 cells
- Complex tissue organization
- Cell Types
  - Early Microscopy
  - Anton van Leeuwenhoek (1632-1723)

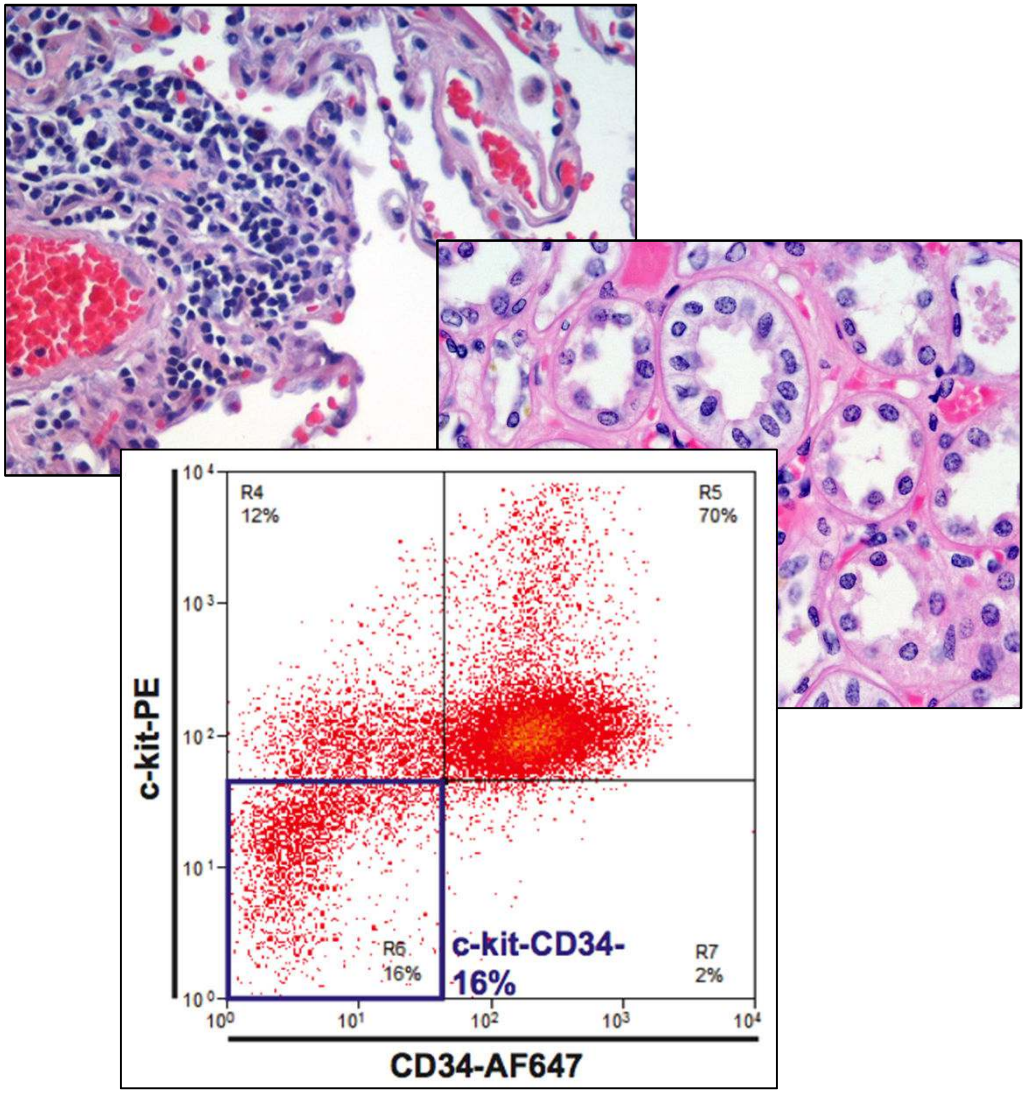




Клетки, Клеточные Типы и Ткани

- Averages of ‘bulk’ samples
  - ~10^7 cells
- Complex tissue organization
- Cell Types
  - Early Microscopy
  - Anton van Leeuwenhoek (1632-1723)

Modern Histology

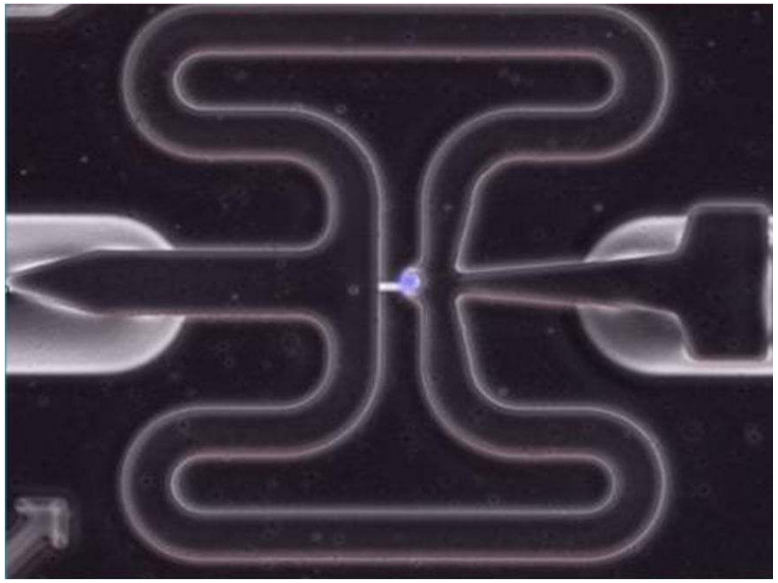


Single-Cell Genomics

- Цели
  - Классификация: клеточные типы, подтипы, состояния
  - Механизмы: молекулярные отличия, динамика
- Измерения
  - Транскрипционное состояние (single-cell RNA-seq, FISSEQ)
  - Епигенетическое состояние (DNA methylation, accessibility)
  - Геномная последовательность (somatic mutations)
- Технологии
  - Миниатюризированные протоколы
  - Микрофлюидика: клапаны, пузырьки
  - Микроскопия: in-situ labeling, sequencing

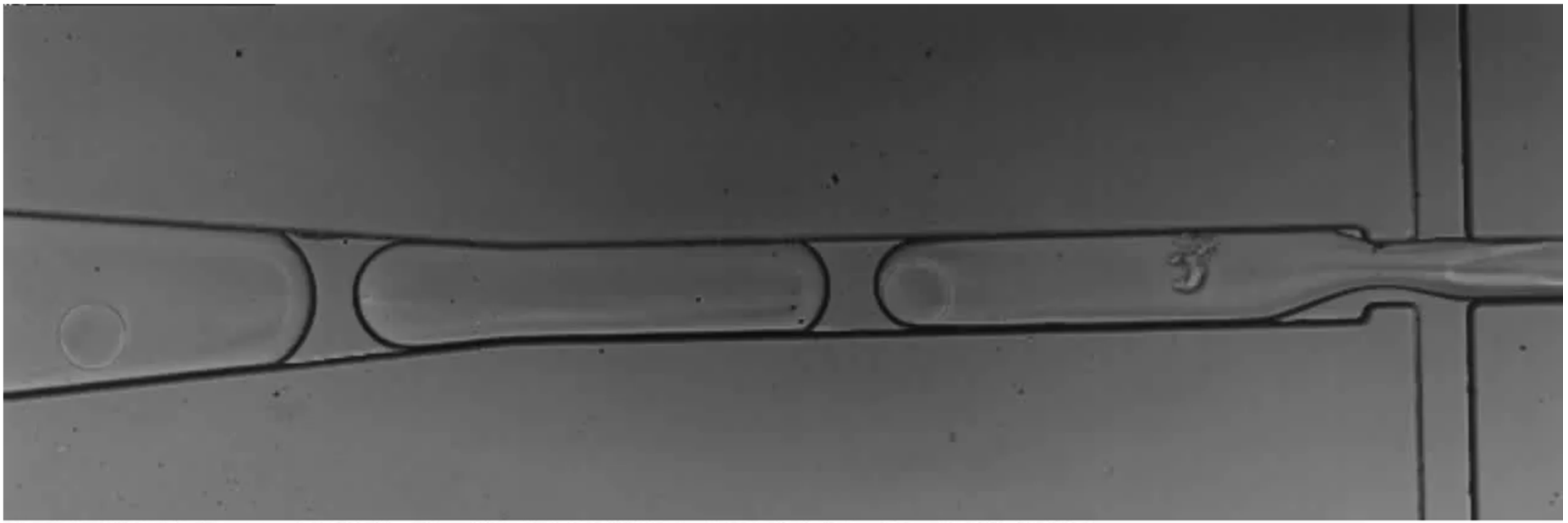
Транскриптомика:
Single-Cell RNA Sequencing

- Разделение и манипулирование клетками
  - Commercial C1 platform from Fluidigm



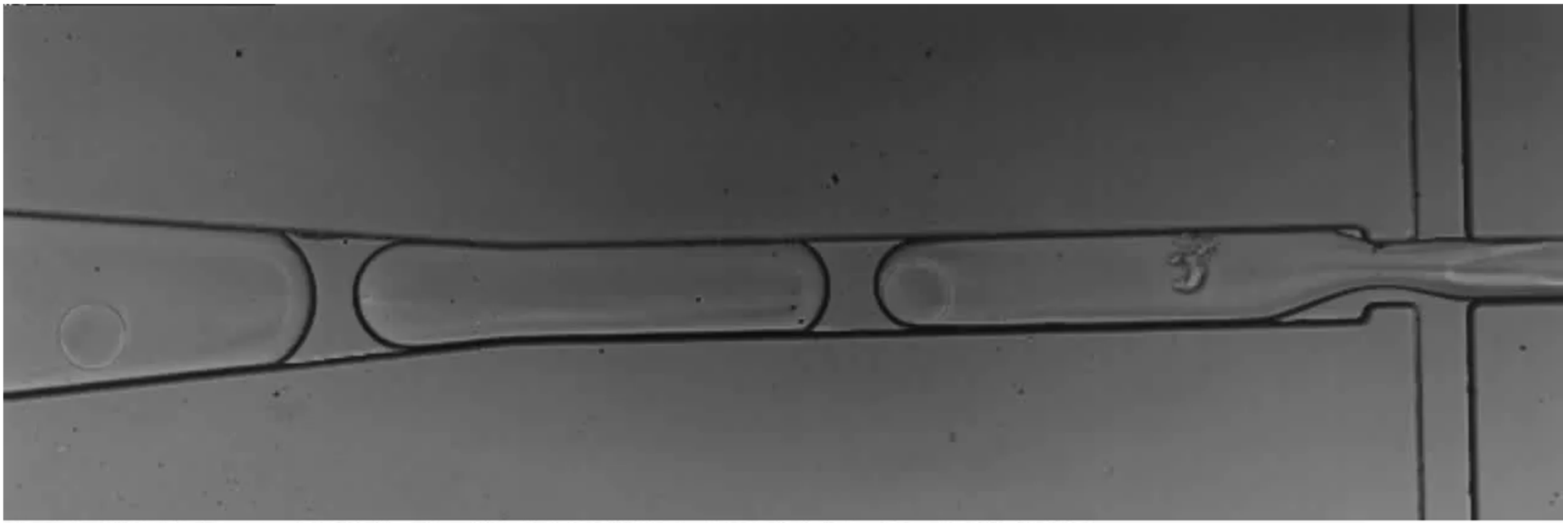
Транскриптомика: Single-Cell RNA Sequencing

- Разделение и манипулирование клетками
  - Commercial C1 platform from Fluidigm
  - Using Cell Sorters to place cells
    - 384-well plates
  - Droplet microfluidics (~10K cells/run)



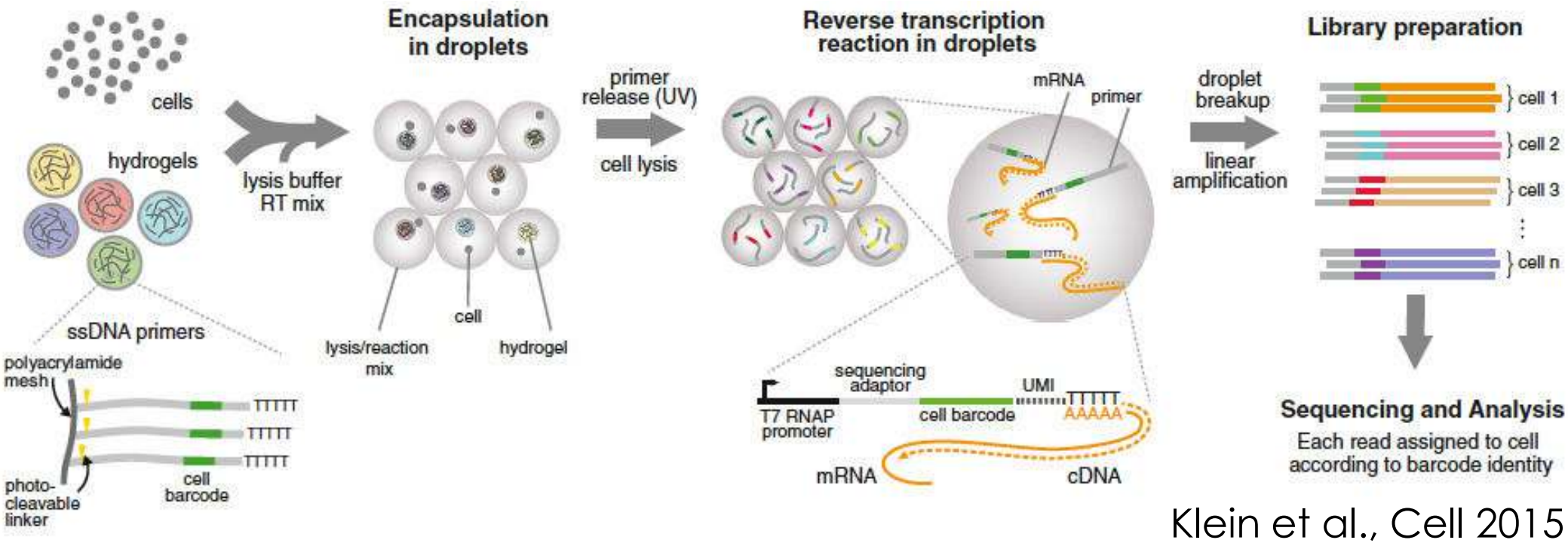
Транскриптомика: Single-Cell RNA Sequencing

- Разделение и манипулирование клетками
  - Commercial C1 platform from Fluidigm
  - Using Cell Sorters to place cells
    - 384-well plates
  - Droplet microfluidics (~10K cells/run)



Транскриптомика: Single-Cell RNA Sequencing

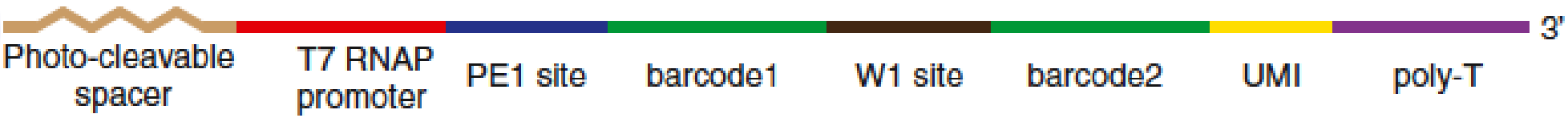
- Разделение и манипулирование клетками
  - Commercial C1 platform from Fluidigm
  - Using Cell Sorters to place cells
    - 384-well plates
  - Droplet microfluidics (~10K cells/run)



Klein et al., Cell 2015

Клеточные и Молекулярные Штрих-Коды

- Read structure



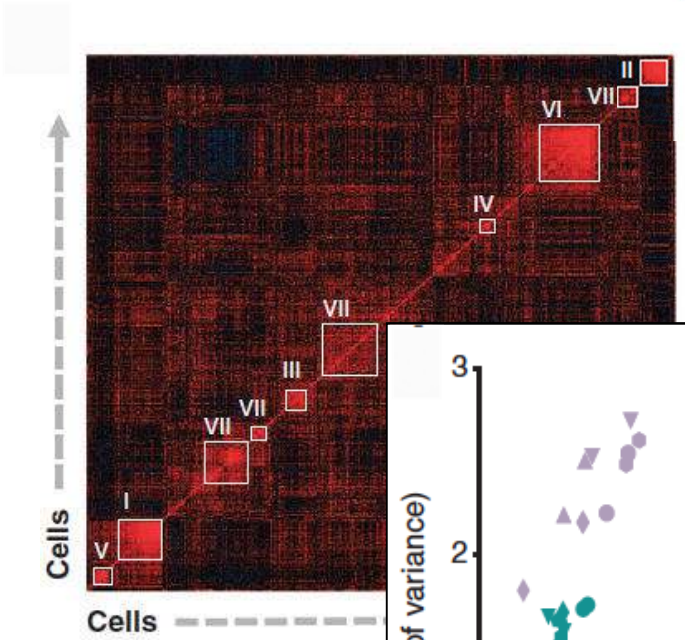
- Barcodes have to be extracted prior to alignment
- Molecular barcode tabulation, collisions
  - Expected number of colliding cells: N barcodes; k cells

$$k_d = k\left(1-(1-1/N)^{k-1}\right) \quad f_d = k_d/(k-k_d) \sim 1\%$$

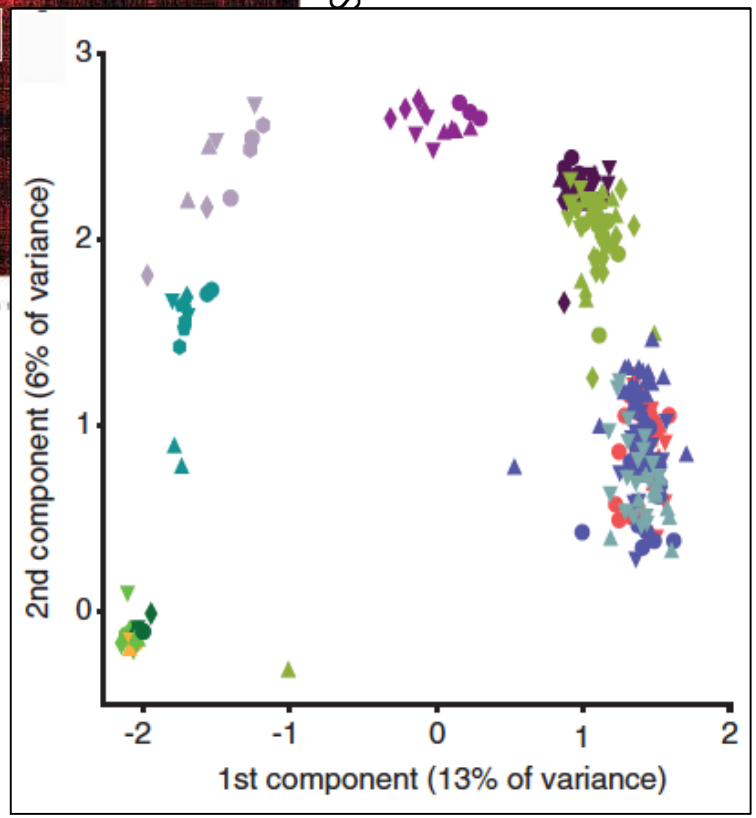
- UMI reduction
  - Count the number of unique molecular barcodes per transcript

Поиск транскрипционно-различных подгрупп

- Clustering
  - K-means, affinity
  - Challenging topology
  - Single, hard, partition



Jaitin *et al.* S

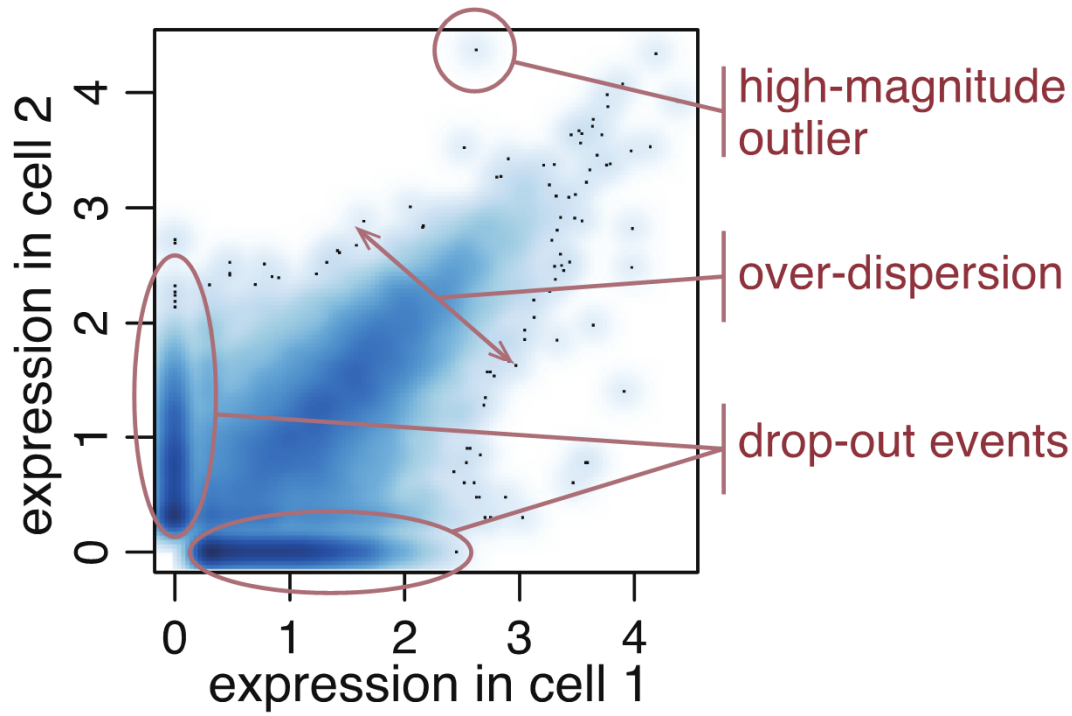


Deng *et al.* Science 2014

- Dimensionality Reduction
  - PCA, ICA
  - Can detect gradual differences
  - Sensitivity, significance

Variability in single-cell RNA-seq data

- Differences between individual cells (of the same type)
- Overdispersion
- Measurement failures
- Cells vary in "quality"
- Problems for PCA
  - Non-Gaussian
  - Noisy cells form outliers
  - Drop-out events can mask true covariance structure

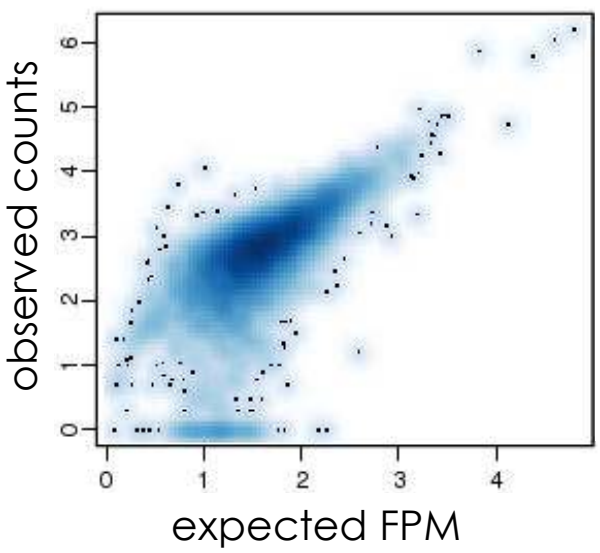


- Biological and technical
  - Control for the technical variability
  - Focus on the biological variability

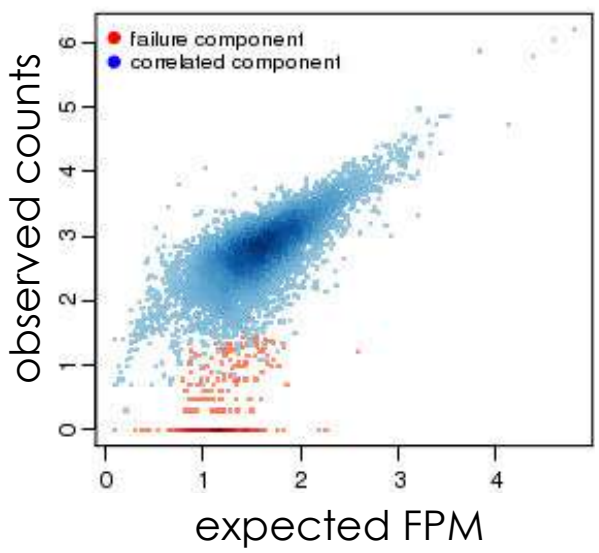
Модель Ошибок в Отдельной Клетке

- Mixture: may amplify or drop-out, depending on expression level
- $count_i \sim NegativeBinomial(\ M_i\ )$ | $count_i \sim Poisson()$
  - M_i – expected expression magnitude for gene i
    (based on consensus of non-drop-out measurements within a group)
- Mixing between the two options depends on the magnitude itself
  - probability of drop-out is modeled using logistic regression

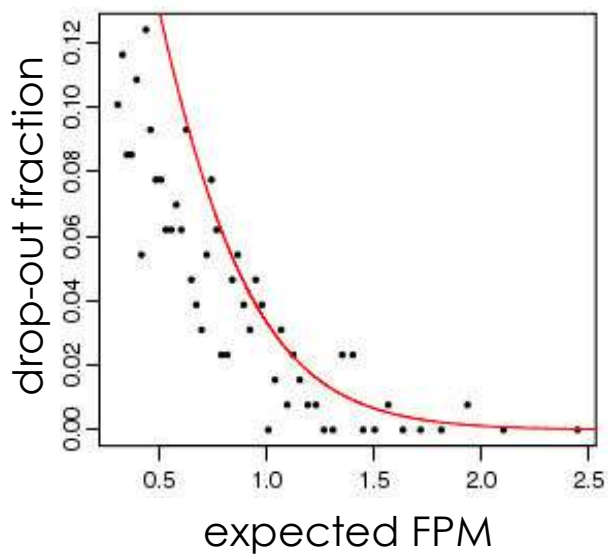
observed vs. expected



mixture fit



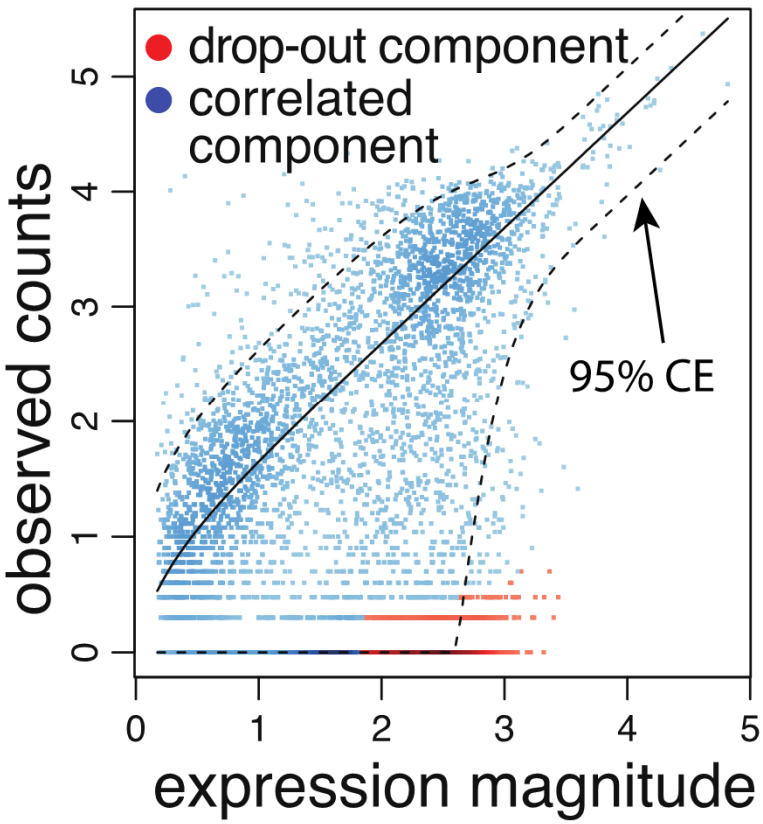
drop-out dependency

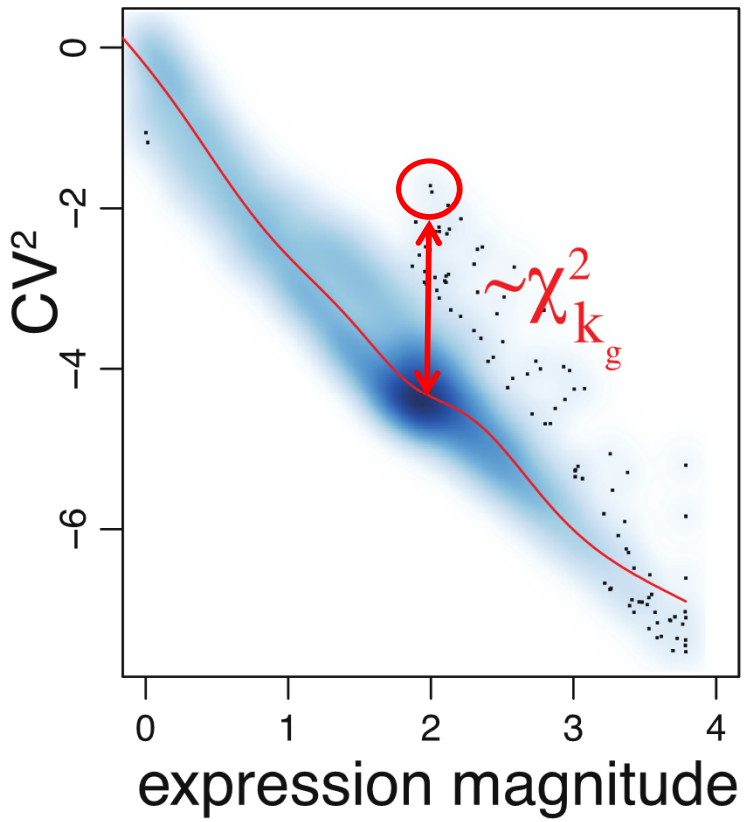


Kharchenko *et al.* Nature Methods 2014

Нормализация Дисперсии

- Account for cell- and gene-specific uncertainty
  - Cell-specific error models
  - Translate into χ^2 statistic
- Account for expression magnitude dependency
  - Adjusting with local regression fit



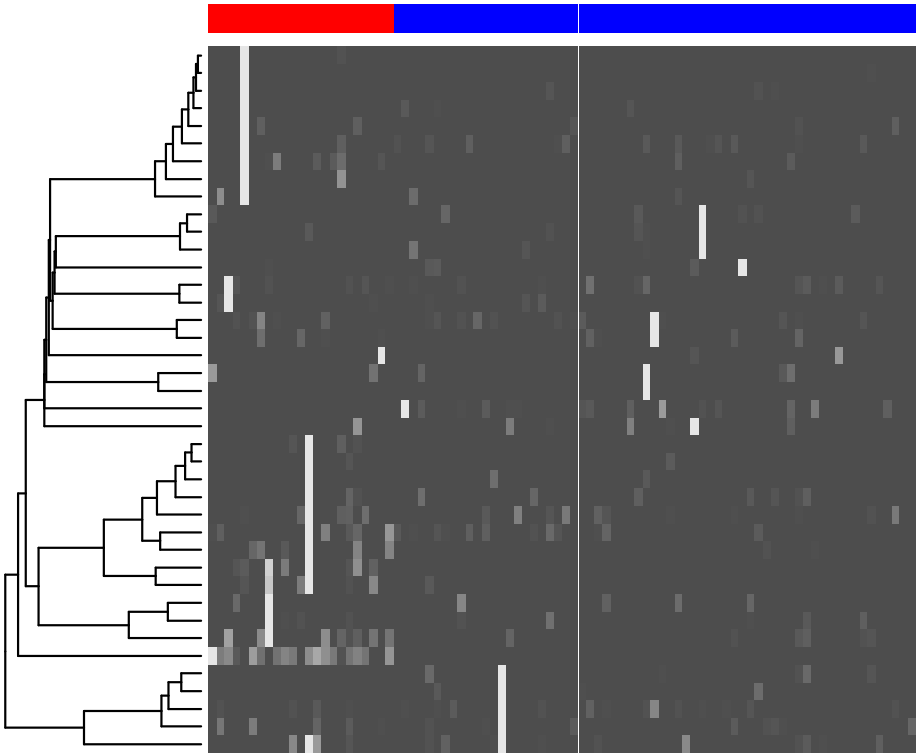


Kharchenko *et al.* Nature Methods 2014

Определение Высоко-Дисперсных Генов

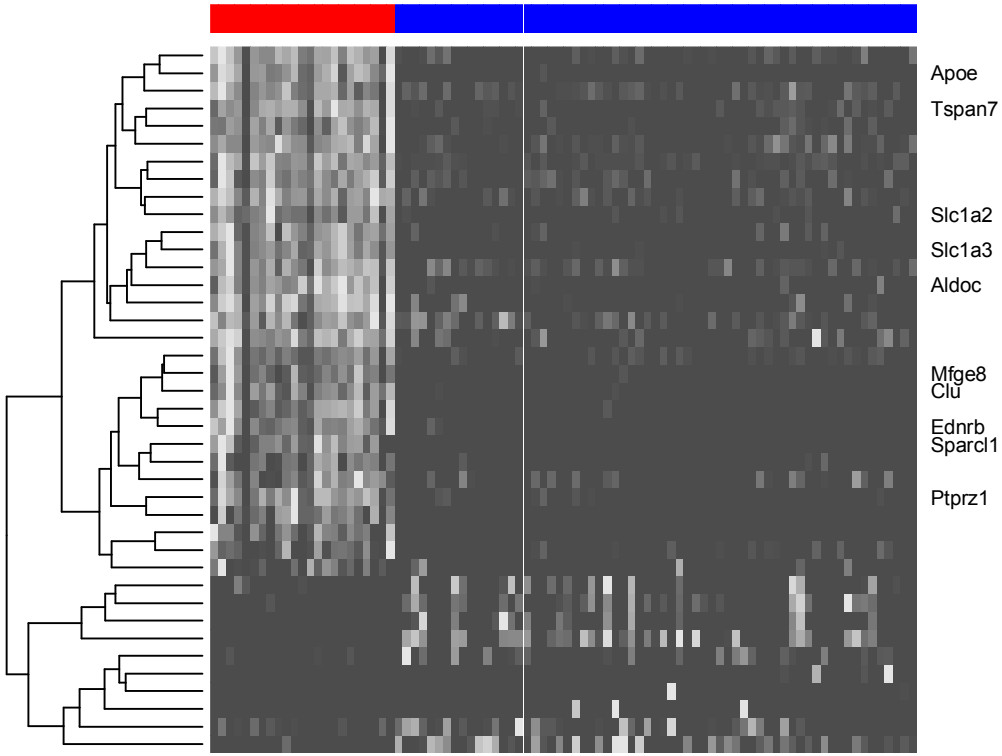
Brennecke et al. Nat. Methods 2013

Astrocytes **NPCs**



Cell-specific models (PAGODA)

Astrocytes **NPCs**



Stochastic neighbor embedding (SNE)

- Convert the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

Gaussian:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- probability that x_i would pick x_j as its neighbor conditional on neighboring being picked in proportion to their probability density under a Gaussian centered at x_i with variance σ_i

Maaten L, Hinton G. Visualizing Data using t-SNE. 9(Nov):2579--2605, 2008.

Stochastic neighbor embedding (SNE)

- Define a similar conditional probability / similarity metric for the low-dimensional counterparts y_i and y_j

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

- Note: If the map points y_i and y_j correctly model the similarity between the high-dimensional datapoints x_i and x_j, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal
  - Probability of picking each other as neighbors is same regardless of dimension

Maaten L, Hinton G. Visualizing Data using t-SNE. 9(Nov):2579--2605, 2008.

Stochastic neighbor embedding (SNE)

- SNE finds the set of y_i and y_j s that minimize the sum of Kullback-Leibler divergences between conditional probabilities $p_{j|i}$ and $q_{j|i}$ over all datapoints by gradient descent

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

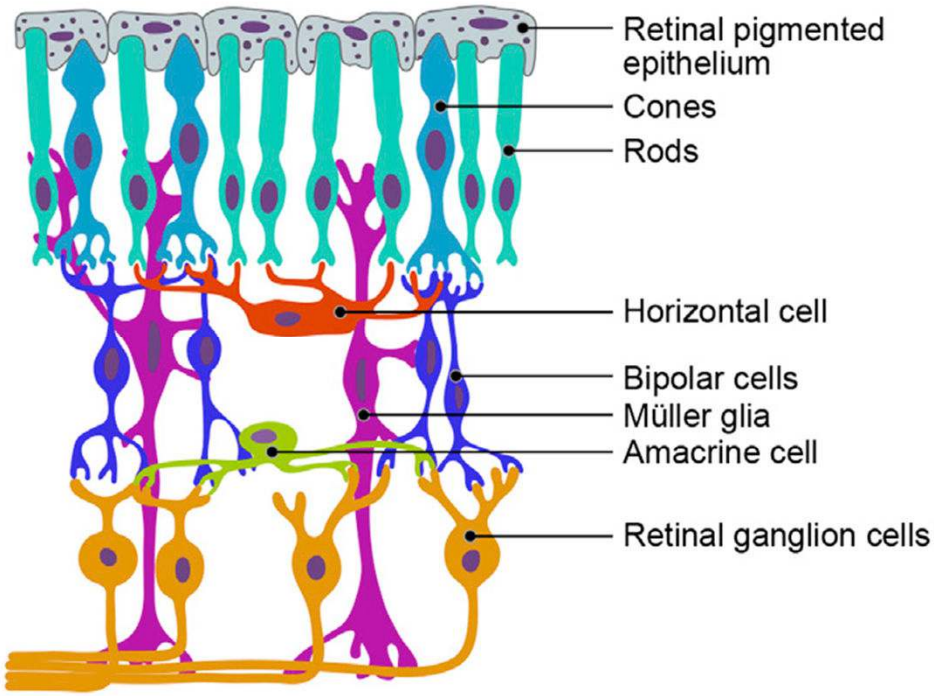
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

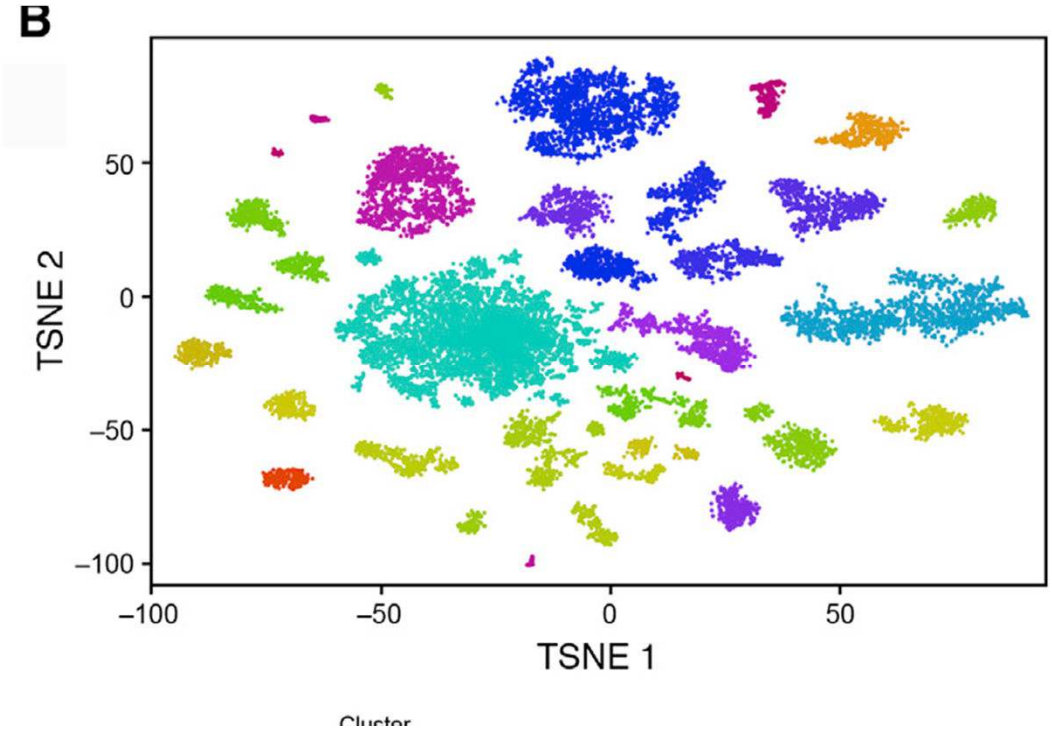
$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Maaten L, Hinton G. Visualizing Data using t-SNE. 9(Nov):2579--2605, 2008.

Single-Cell RNA-seq: tissue composition

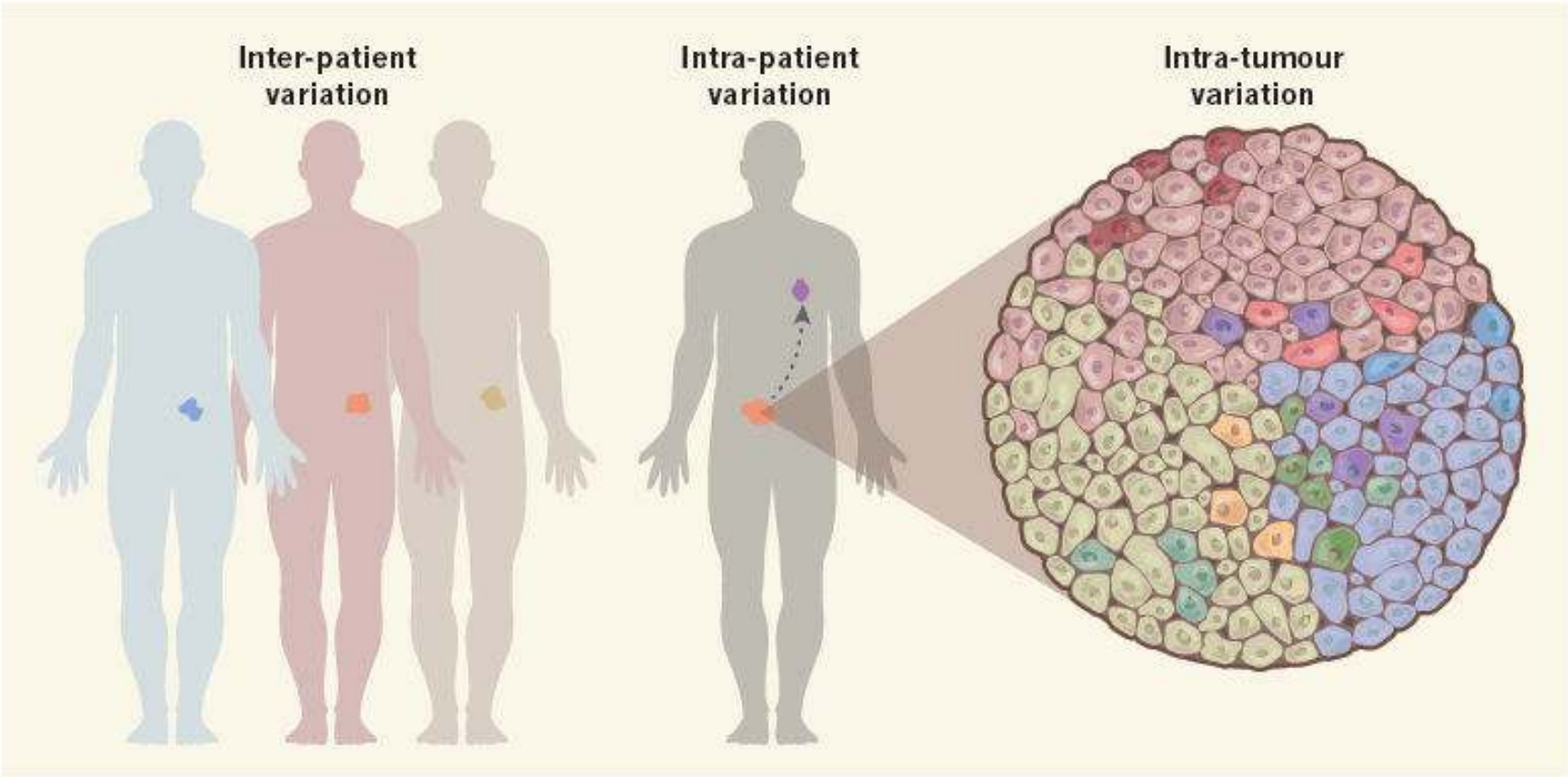
- Mouse Retina: Macosco *et al.* Cell 2015



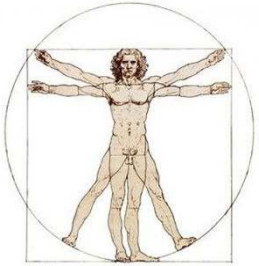


- 50 thousand cells collected using droplet method
- Recovers most known subtypes of cells

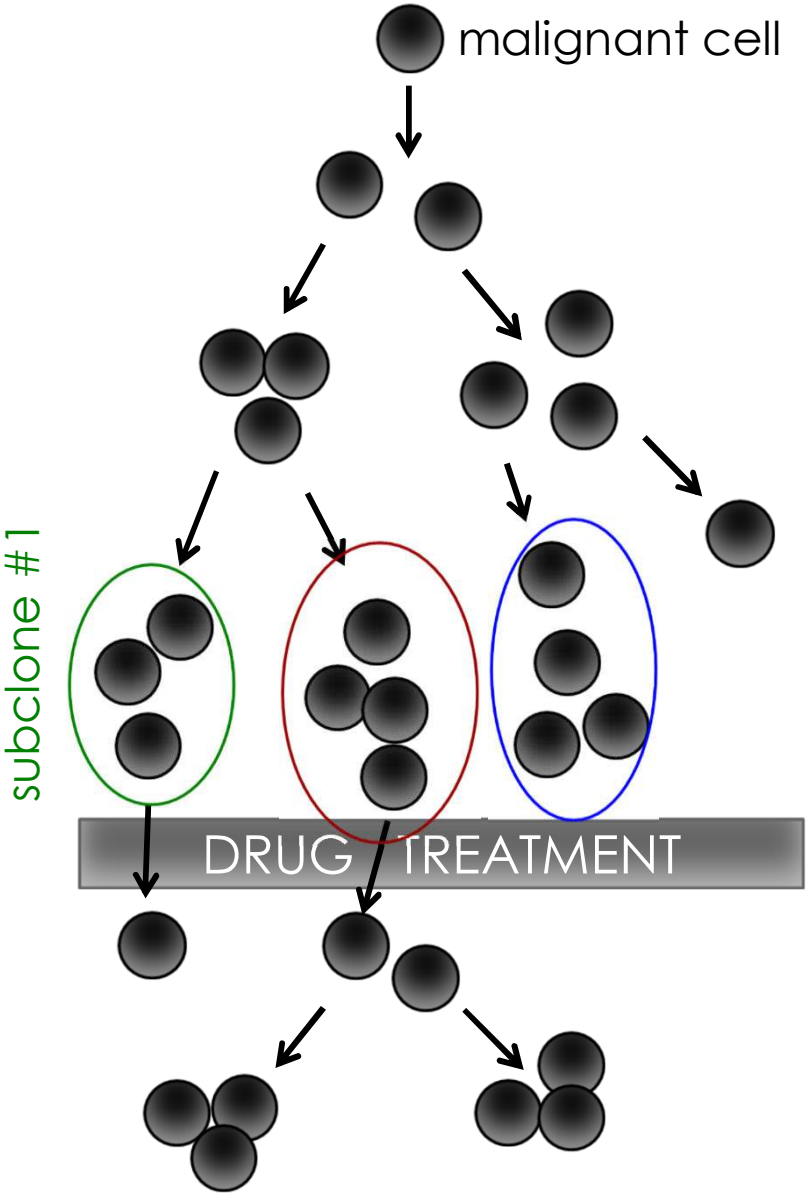
Гетерогенность Раковых Тканей



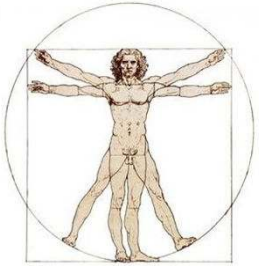
Гетерогенность Раковых Тканей



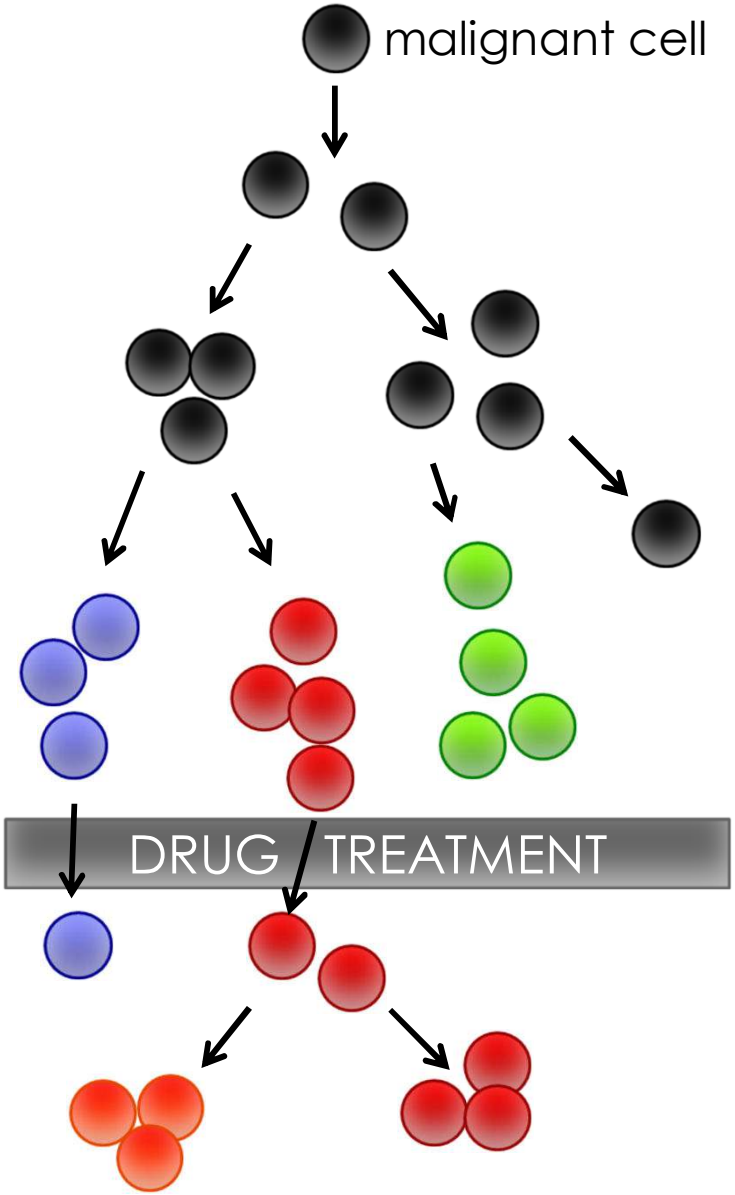
- Clinical relevance
  - Progression
  - Drug Resistance
- Genetic Heterogeneity
  - Clonal evolution
  - Genome sequencing



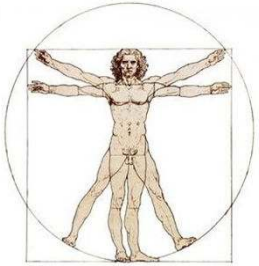
Гетерогенность Раковых Тканей



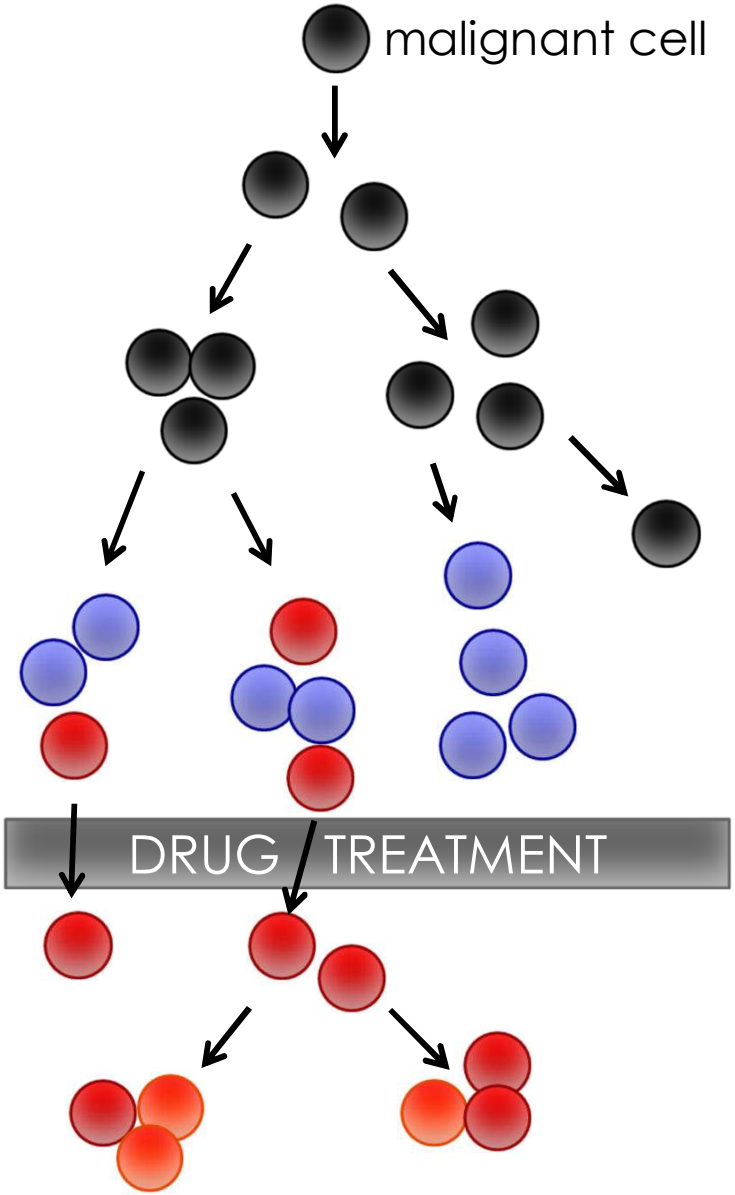
- Clinical relevance
  - Progression
  - Drug Resistance
- Genetic Heterogeneity
  - Clonal evolution
  - Genome sequencing
- Transcriptional Heterogeneity
  - Transient vs. stable states



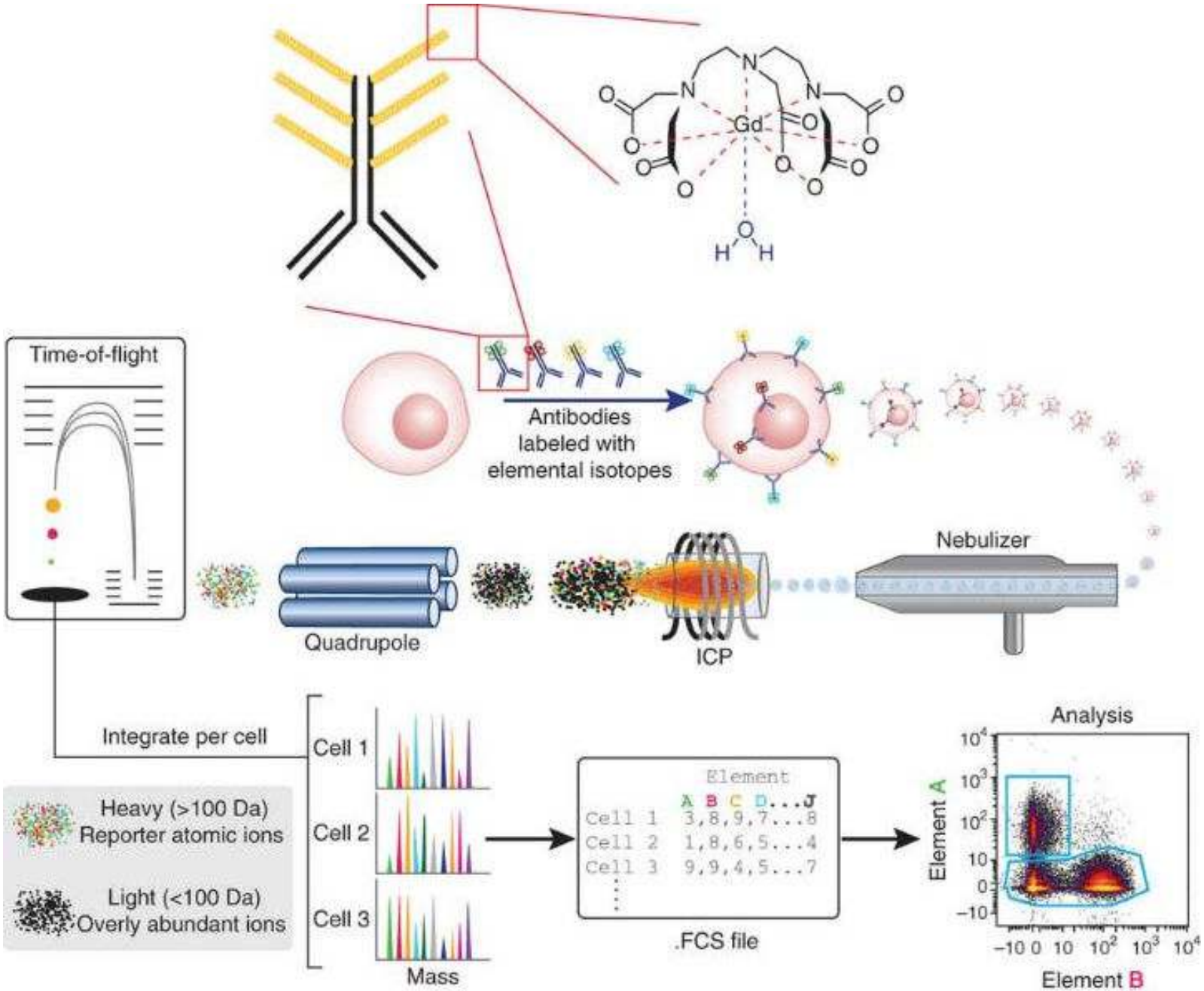
Гетерогенность Раковых Тканей



- Clinical relevance
  - Progression
  - Drug Resistance
- Genetic Heterogeneity
  - Clonal evolution
  - Genome sequencing
- Transcriptional Heterogeneity
  - Transient vs. stable states

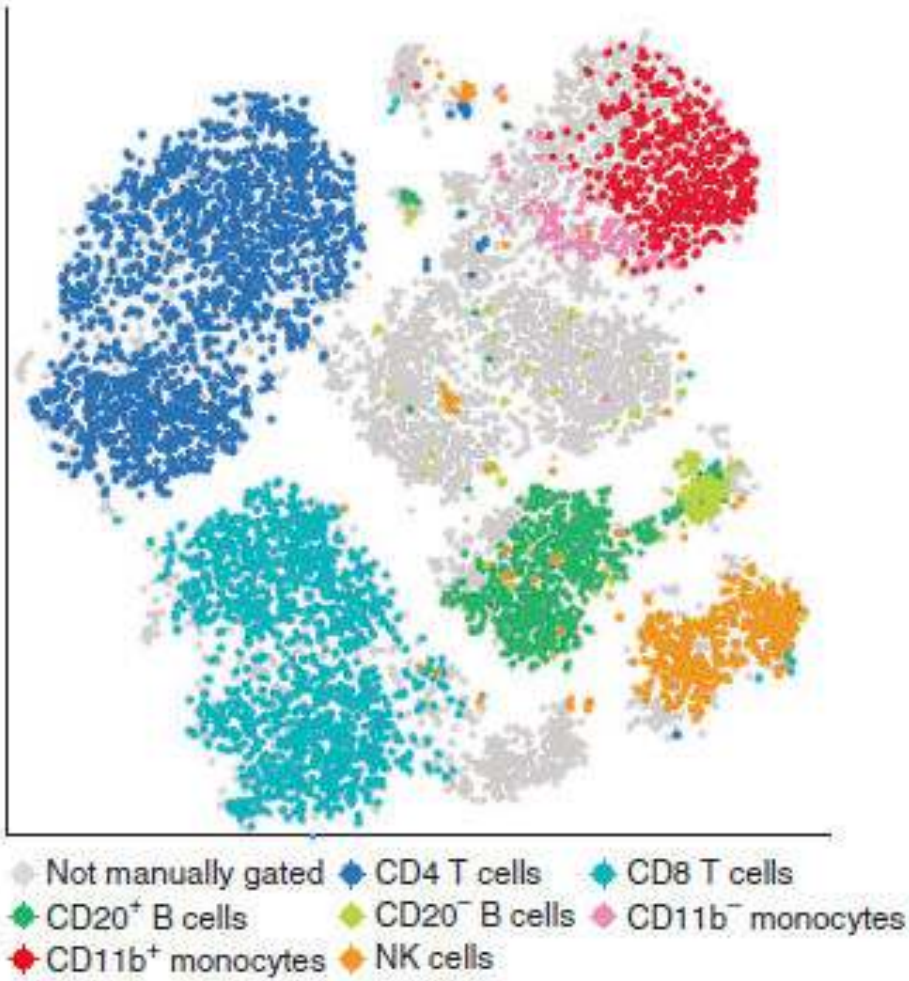


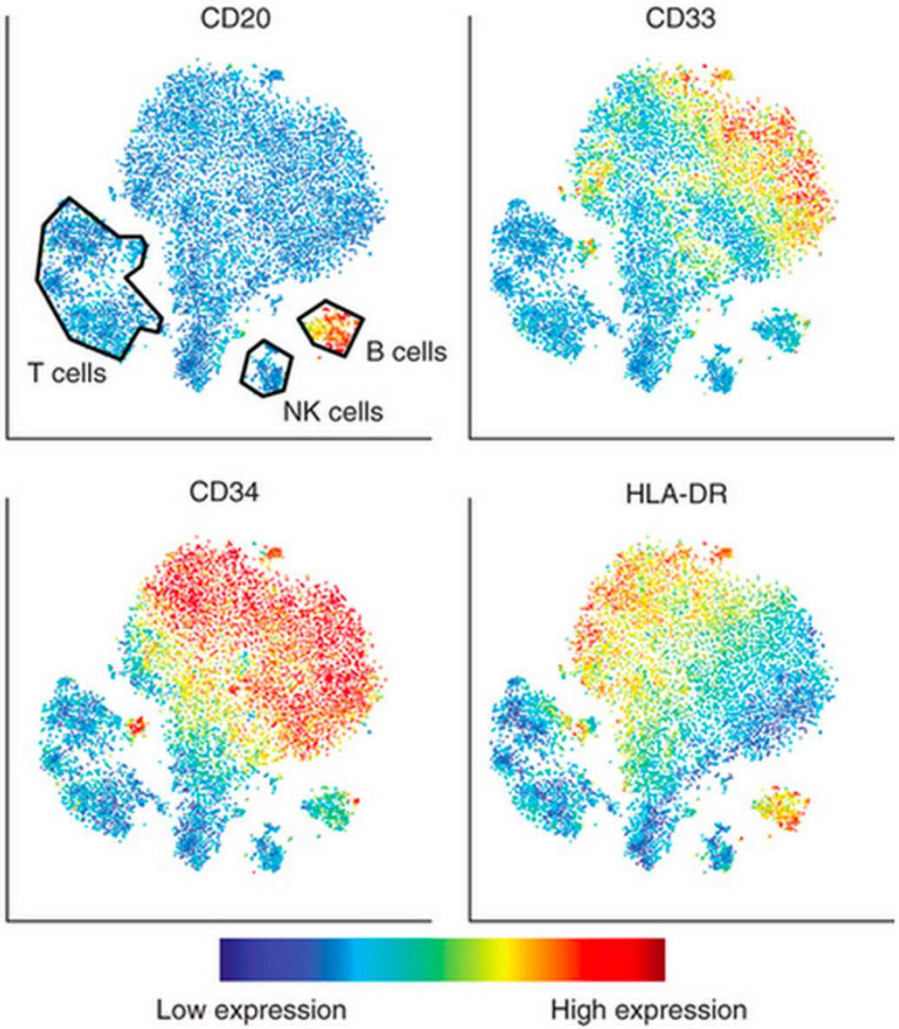
Mass Cytometry: CyTOF



viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

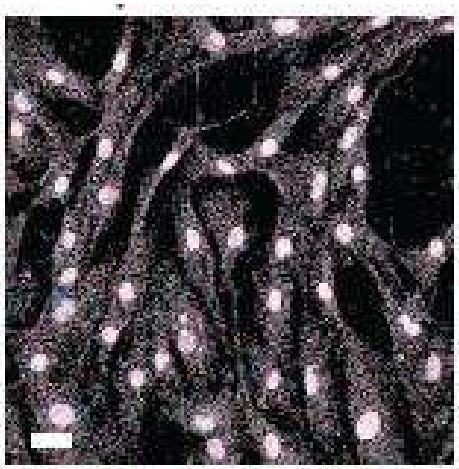
El-ad David Amir[1], Kara L Davis[2,3], Michelle D Tadmor[1,3], Erin F Simonds[2,3], Jacob H Levine[1,3], Sean C Bendall[2,3], Daniel K Shenfeld[1,3], Smita Krishnaswamy[1], Garry P Nolan[2,4] & Dana Pe'er[1,4]

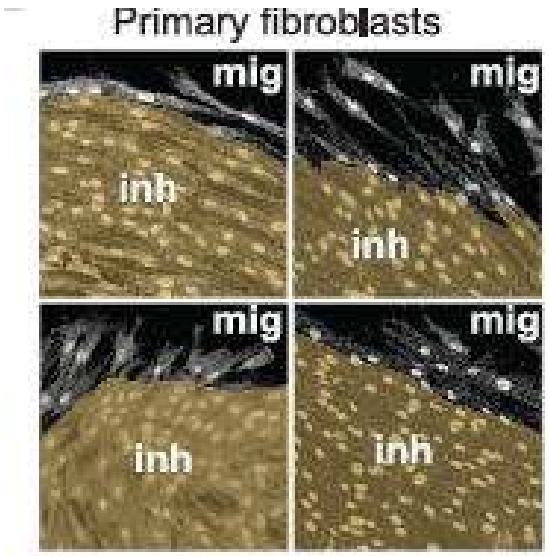


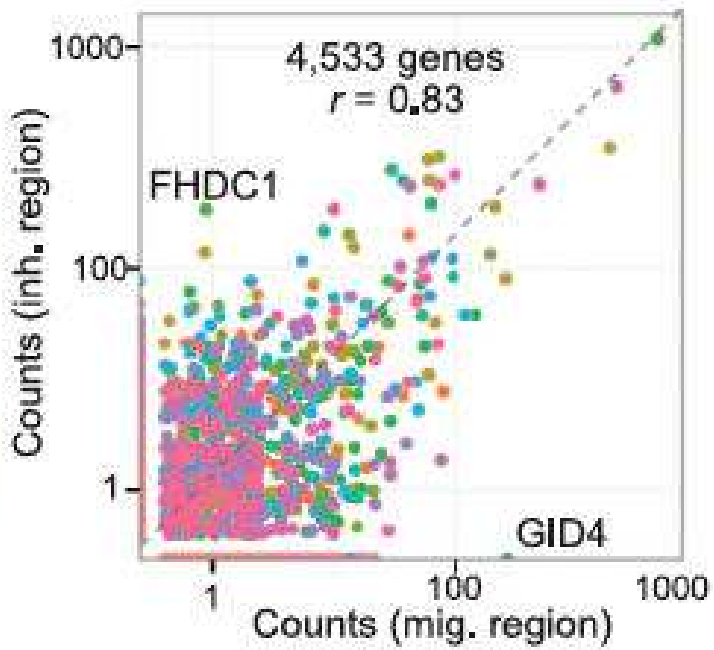


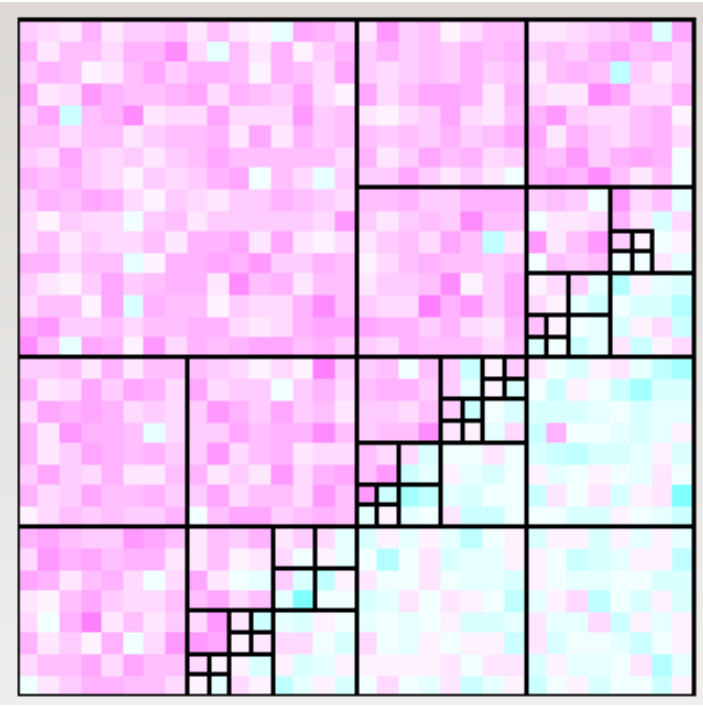
In situ RNA sequencing: FISSEQ

- Lee et al. Science 2014









Multi-scale Poisson models