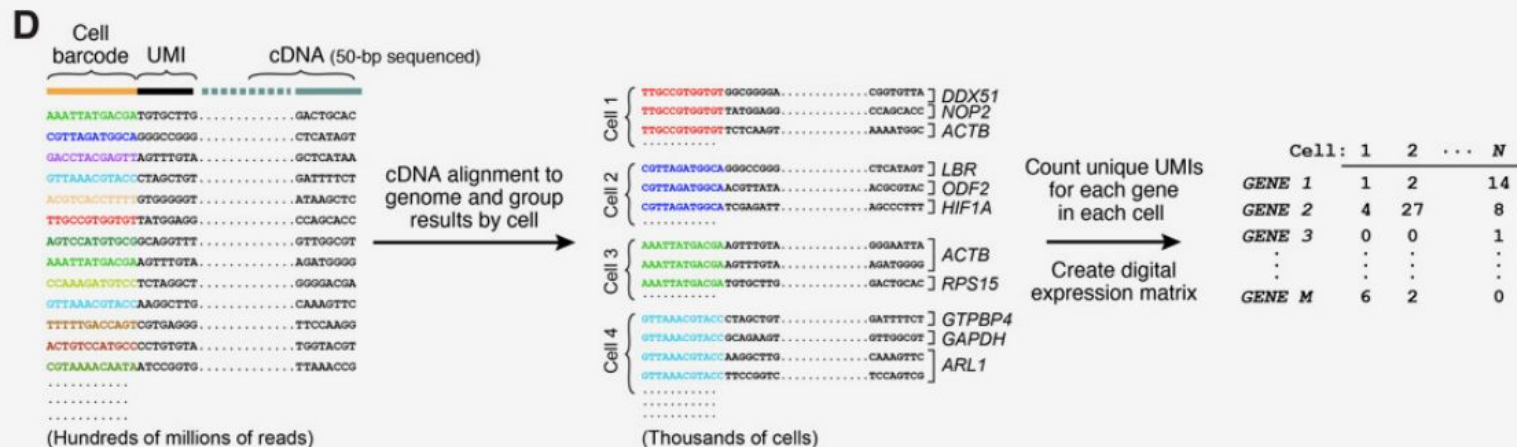




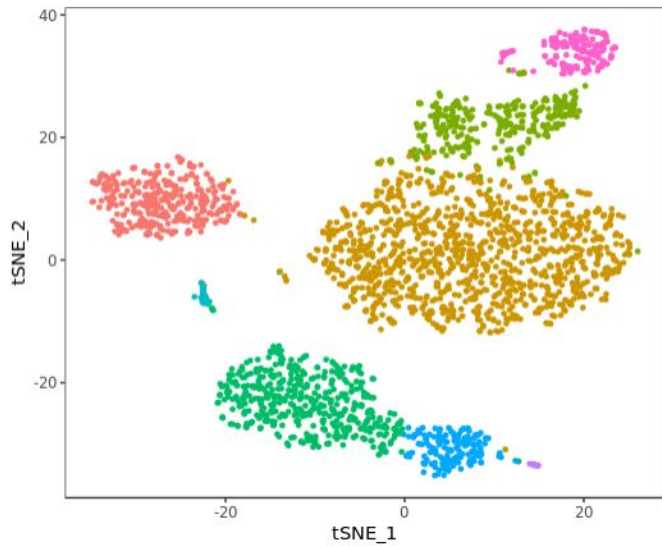
# Automated pathway annotation for single-cell RNA-seq

Student: Фирулёва Мария  
Tutor: Зайцев Константин

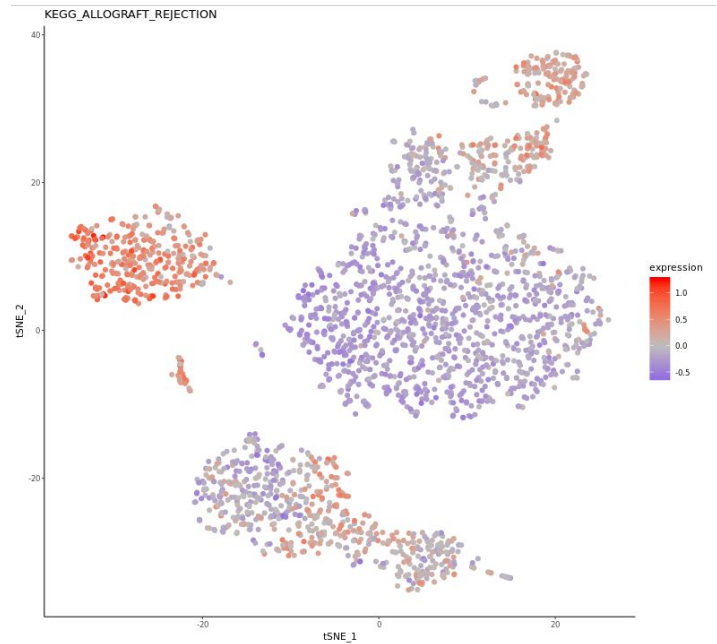
# DropSeq



# Seurat R toolkit



- B cells
- CD4 T cells
- CD8 T cells
- CD14+ Monocytes
- Dendritic cells
- FCGR3A+ Monocytes
- Megakaryocytes
- NK cells





**Input pathway contains gene "B", gene "D" and gene "E"**

|          | Cell 1 | Cell 2 |
|----------|--------|--------|
| Gene "A" | 25     | 5      |
| Gene "B" | 18     | 7      |
| Gene "C" | 95     | 8      |
| Gene "D" | 11     | 18     |
| Gene "E" | 10     | 23     |
| Gene "G" | 5      | 5      |



|          | Cell 1 | Cell 2 |
|----------|--------|--------|
| Gene "B" | 18     | 7      |
| Gene "D" | 11     | 18     |
| Gene "E" | 10     | 23     |



|           | Cell 1 | Cell 2 |
|-----------|--------|--------|
| Target PW | 13     | 16     |



# Compare reference and “random” vectors

Reference vector

|           | Cell 1 | Cell 2 |
|-----------|--------|--------|
| Target PW | 13     | 16     |

| <u>1000</u>    | Cell 1 | Cell 2 |
|----------------|--------|--------|
| 3 random genes | 11     | 2      |

Vector after random sample



Result vector

|           | Cell 1 | Cell 2 |
|-----------|--------|--------|
| Target PW | 0,7    | 0,002  |



# Statistical analysis

- **Hypergeometric distribution**
  - Number of cells is the population size
  - Number of cells with p-value  $< 0.01$  is the number of success states in the population
  - Number of cells in the cluster is the number of draws
  - Number of cells in the cluster with p-value  $< 0.01$  is the number of observed successes
- **Bonferroni correction**


## The longest pathway for cumulative matrix

|              | Cell 1 | Cell 2 |
|--------------|--------|--------|
| Rand. gene 1 | 5      | 7      |
| Rand. gene 2 | 5      | 5      |
| Rand. gene 3 | 1      | 111    |
| Rand. gene 4 | 7      | 34     |
| ...          | ...    | ...    |
| Rand. gene N | 13     | 28     |



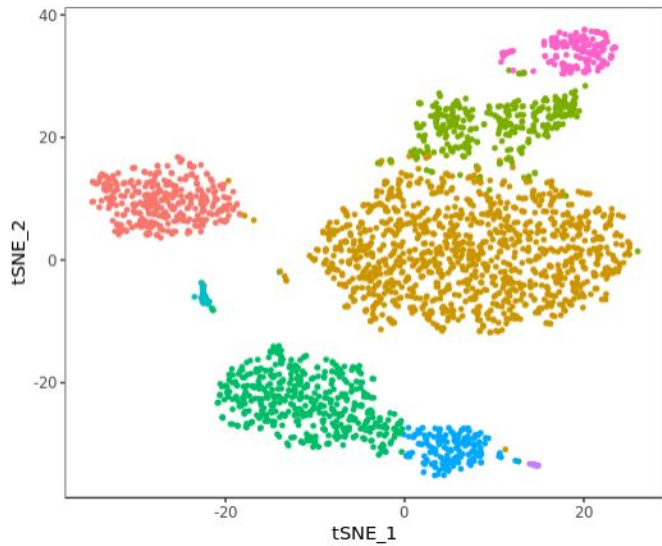
|                     | Cell 1    | Cell 2     |
|---------------------|-----------|------------|
| PW (1 gene)         | 5         | 7          |
| PW (2 genes)        | 10        | 12         |
| <b>PW (3 genes)</b> | <b>11</b> | <b>123</b> |
| PW (4 genes)        | 18        | 157        |
| ...                 | ...       | ...        |
| PW (N genes)        | 1948      | 2765       |

# Result

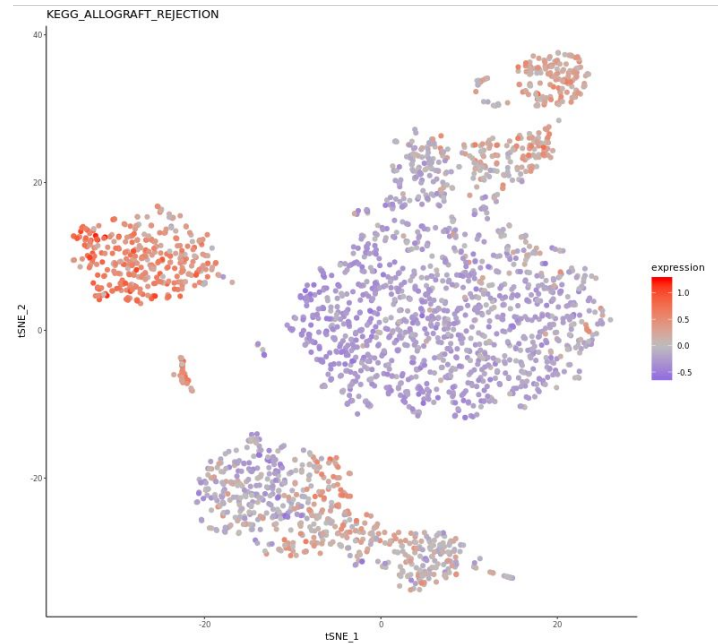
|   | B cells       | CD4 T cells  | CD8 T cells | CD14+ Monocytes | Dendritic cells | FCGR3A+ Monocytes | Megakaryocytes | NK cells     |
|---|---------------|--------------|-------------|-----------------|-----------------|-------------------|----------------|--------------|
| <b>KLEIN_PRIMARY Effusion_Lymphoma_DN</b>   | 1.037160e-255 | 1.000000e+00 | 1           | 1.000000e+00    | 1.000000e+00    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>KEGG_Asthma</b>  | 2.995296e-188 | 1.000000e+00 | 1           | 1.000000e+00    | 3.117624e-19    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>KEGG_Intestinal_Immune_Network_For_IG...</b>   | 5.290776e-163 | 1.000000e+00 | 1           | 1.000000e+00    | 7.287582e-05    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>KEGG_Autoimmune_Thyroid_Disease</b>  | 3.186094e-138 | 1.000000e+00 | 1           | 1.000000e+00    | 1.348480e-06    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>HADDAD_B_Lymphocyte_Progenitor</b>   | 9.531863e-136 | 1.000000e+00 | 1           | 1.000000e+00    | 1.000000e+00    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>YU_MYC_Targets_DN</b>  | 1.394662e-133 | 1.000000e+00 | 1           | 1.000000e+00    | 1.000000e+00    | 1                 | 1.000000e+00   | 1.000000e+00 |
|  <b>KEGG_Allograft_Rejection</b> | 2.023047e-124 | 1.000000e+00 | 1           | 1.000000e+00    | 7.342853e-08    | 1                 | 1.000000e+00   | 1.000000e+00 |
| <b>KEGG_Type_I_Diabetes_Mellitus</b>  | 5.239544e-87  | 1.000000e+00 | 1           | 1.000000e+00    | 8.264055e-08    | 1                 | 1.000000e+00   | 1.000000e+00 |



# Visualize result

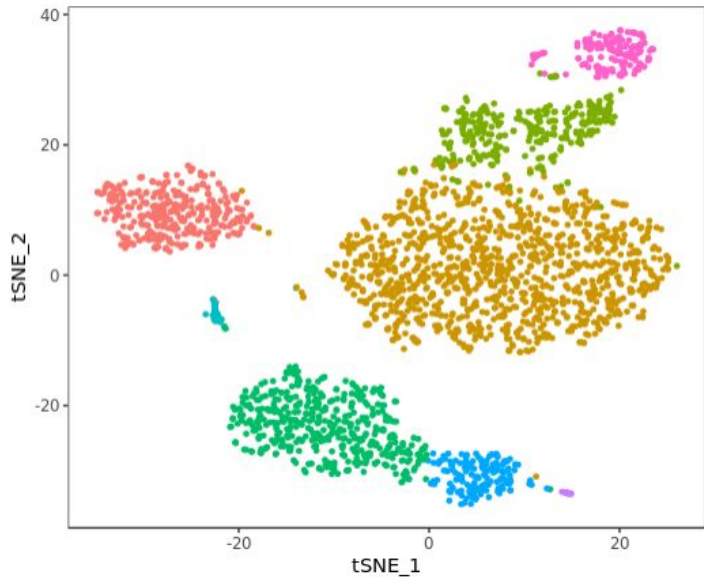


- B cells
- CD4 T cells
- CD8 T cells
- CD14+ Monocytes
- Dendritic cells
- FCGR3A+ Monocytes
- Megakaryocytes
- NK cells

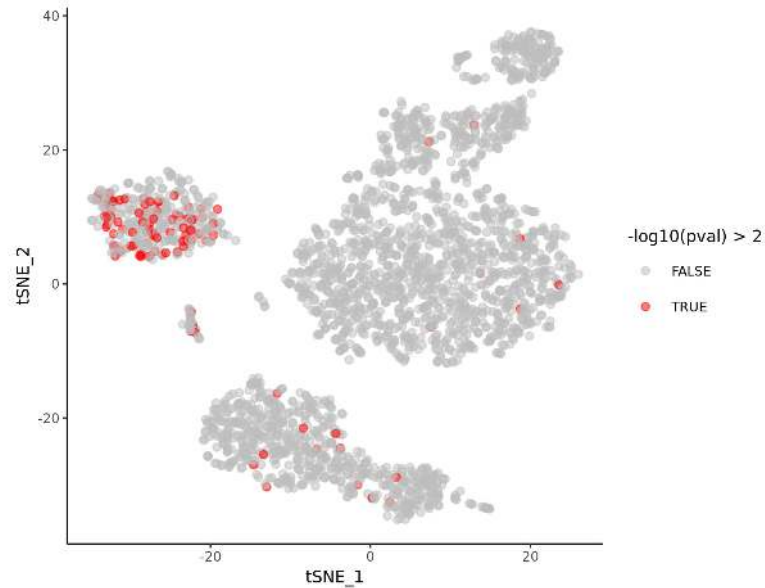




# tsNE: p-values



- B cells
- CD4 T cells
- CD8 T cells
- CD14+ Monocytes
- Dendritic cells
- FCGR3A+ Monocytes
- Megakaryocytes
- NK cells





# Summary

- Lists of pathway genes as input
- Function was implemented in C++
- Function: calculate p-values for every cell from seurat gene expression matrix, list of pathway genes, number of random generations;  $p < 0.01$  (after multiple hypothesis correction)
- After optimization 10 000 as a sample number not time-consuming
- Visualize results for each clusters



**Thanks for your attention!**  
**Questions?**

