Text Report on

Classification of cells using single cell transcriptomics data and machine learning

Reporter: Jiahui Zhong

Student Number: 20120196

Supervisor: Prof. Vladimir Brusic

Dr. Heshan Du

Dr. Xiaoling Liu

Content

Abstract

In recent years, single cell transcriptomics (SCT) becomes much popular research method instead of bulk sequencing technology. It can detect heterogeneous genetic information which is not obtained by mixed sample multicellular sequencing. This leads the whole field of genetics into a new dimension. The main steps of SCT are a) single cell capturing/ sorting, b) reverse transcription/ PCR amplification, c) library building and sequencing, d) biological information analysis. Downstream analysis to single cell experimental data with supervised machine learning can build classification of cells, detect rare subtype of blood cells, refine the ontology of immune cells and conduct diseases diagnosis and health prediction. This project focus on classification of cells using SCT data and machine learning. The objective of this project is to develop and implement an applied system that can use the updating single cell database, perform supervised machine learning, and validate the performance of the system on the follow-up new data sets. The project includes the following steps: data collection, feature extraction, model structure building and training, classification system training, testing and validation, and comparative analysis. This system will be used to determine the cell type and the sorting of one single cell.

Aims and objectives

1. Organize data from single cell transcriptome (SCT) data files (10X technology).

2. Develop ontology of human cell types.

3. Develop AI based classification models for various cell types and their states.

4. Validate those models with experimental data reported elsewhere.

5. Implement a system for assessment of anonymous file for each individual cell.

6. Explore the use of this system for potential blood testing.

7. Deliverables objectives: 5 journal papers and 1 conference paper.

Background

1 What is single cell sequencing?

Single cell sequencing is a sequencing technique for obtaining genetic information of a single cell. Single-cell sequencing is to answer different types of questions at the single-cell level: a) DNA sequence: ATCG sequence and abundance of each sequence; b) epigenetic modification of DNA: methylation, hydroxymethylation, and various modifications of histone; c) RNA sequence: AUCG sequence and abundance of each sequence; d) epigenetic modification of RNA: for example, m6A modification, which has been very popular in recent research; e) chromatin structure: 3C, 4C, 5C, etc.; f) other applications: DNA damage location, protein-protein interaction, etc.

2 Why use single cell sequencing?

In the field of life sciences, single cell sequencing has solved many biological problems that have not been clarified for a long time because of the bottleneck of biotechnology.[1] Single cell sequencing technology can interpret many gene level problems with high cost-effectiveness. The research and application of single cell sequencing technology have increased exponentially in recent years.[2]

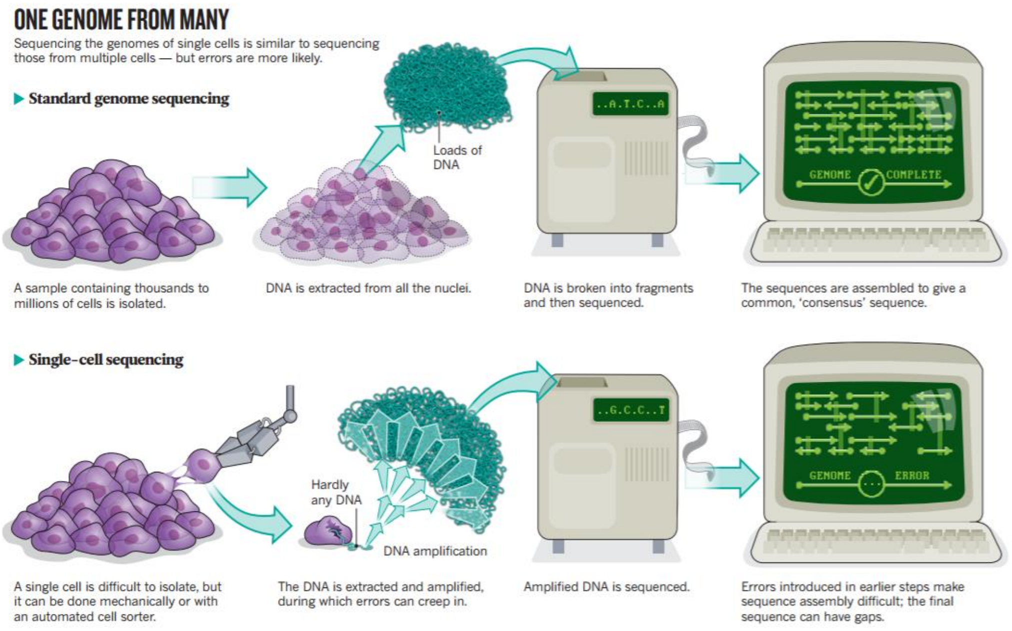

ONE GENOME FROM MANY
Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.
Standard genome sequencing
Loads of DNA
..A.T.C..A
GENOME COMPLETE
A sample containing thousands to millions of cells is isolated.
DNA is extracted from all the nuclei.
DNA is broken into fragments and then sequenced.
The sequences are assembled to give a common, 'consensus' sequence.
Single-cell sequencing
Hardly any DNA
DNA amplification
..G.C.C..T
GENOME ERROR
A single cell is difficult to isolate, but it can be done mechanically or with an automated cell sorter.
The DNA is extracted and amplified, during which errors can creep in.
Amplified DNA is sequenced.
Errors introduced in earlier steps make sequence assembly difficult; the final sequence can have gaps.

Fig. 1 Standard genome sequencing and single cell sequencing workflow. [3]

2. 1 The division of fine tissue

In many cases, the target sample for transcriptome sequencing is not necessarily the entire organ (e.g. the whole pancreas or leaf), but a more specific local tissue. At this time, the traditional transcriptome will face limitations. For example, embryo research is much involved in the field of developmental biology. However, almost all species have very small individuals at embryonic stage. When studying precise sections, it is easy to attach other tissue parts to the captured sample (e.g. when pulling off the skin tissue of early embryo as sample, it is difficult to ensure that other tissues, such as muscle, are not involved).[4] Another similar example is the research on intestine of fruit flies. It is too difficult to separate the whole tissue, which is most probably brought into other tissues together (even the whole individual). In this case, the target tissue only accounts for a small part of the whole sequencing sample, and sequencing samples have a much high probability of not presenting the concerned information.

2. 2 Cellular heterogeneity

For multicellular organisms, there are differences between cells and cells. Oosperm begins to divide from one cell and gradually forms blastocyst. When it eventually develops into individuals, there will be more and more differences between cells: some differentiate into neurons, some differentiate into skeletal muscles, each expressing different genetic information and assuming different physiological functions. For example, in cancer tissues, genetic information such as genome and transcriptome of cells at different locations is different. The genetic information of the cells in the center of the mass, the cells around the mass, the cells in the lymphatic metastasis, and the cells in the distal metastasis are different.[5] This difference can determine whether a certain treatment is effective for the tumor in clinical practice. This is the heterogeneity of genetic information. Traditional research methods are carried out at the multicellular level. Conventional transcriptome essentially conducts detection to a mixture of different kinds of cells in tissue. The final signal value of those is actually the average of multiple cells, which loses innumerable information of heterogeneity. Single cell sequencing can detect heterogeneous information which cannot be obtained by mixed sample sequencing. Single cell analysis is much used in diagnostic application of cancer, multiple sclerosis, diabetes, and infection. [6-9]

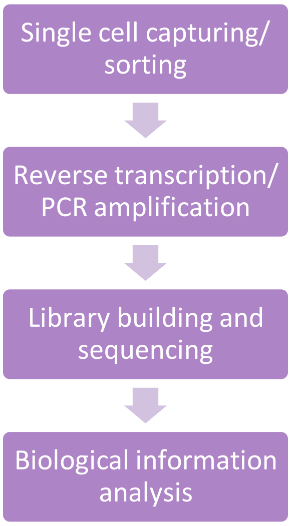

Single cell capturing/ sorting
Reverse transcription/ PCR amplification
Library building and sequencing
Biological information analysis

Fig. 2 The main steps of the single cell transcriptome.

3 How to realize single cell sequencing?

3. 1 Separate individual cell

Currently there are two strategies to achieve single-cell sequencing.

The first strategy is to separate individual cells, construct sequencing libraries independently and finally conduct sequencing routes. This process can be produced by flow cytometry (including microfluidic chips) or laser capture microdissection (LCM). Flow cytometry is mainly used in cell samples. For tissue slice samples, single cells are obtained mainly through LCM method. The principle can be seen in the following schematic diagram.

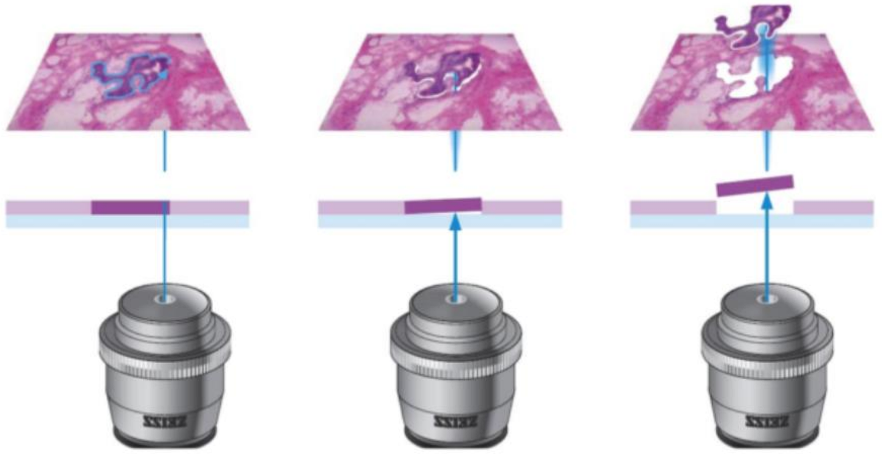


Fig. 3 Laser capture microdissection (LCM) for tissue sample catching. (Guillot et al, 2015)

3. 2 Single cell recognition based on label - barcode

However, the throughput is very low when single cells are separated and built library one by one and then sequenced separately, which is mainly limited by the cost. As the number of cells to be tested increases, the cost of sequencing increases almost linearly as well. The cost of sequencing would be very high even if a dozen or twenty cells are tested. Nonetheless, these dozens of cells are not enough to explain one investigation. In order to overcome this difficulty, the second strategy has been adopted in recent years: single cell recognition based on label - barcode. Its main idea is to add a unique DNA sequence to each cell so that when sequencing, the sequence carrying the same barcode is considered to come from the same cell. This strategy can measure information from thousands of single cells by just building one library at one time.

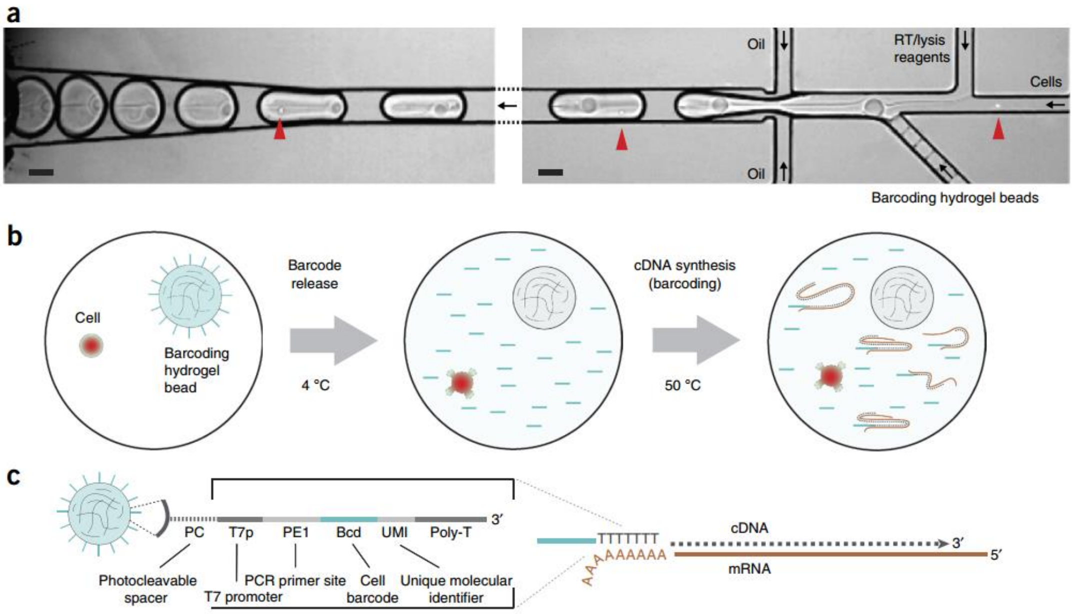


Fig. 4 Single-cell sequencing and barcoding with droplet microfluidics. [10]

For RNA (transcriptome mRNA), the scheme of adding label to cells is to add barcode to the 5' end of poly T primer during reverse transcription process before sequencing of mRNA. See the diagram above. First, single cell suspension samples and barcode hydrogel beads are wrapped in an oil droplet through microfluidic chips. After reverse transcription in oil droplets, the cDNA library of each single cell carries the unique barcode (blue part). Finally, cDNA libraries of all single cells are mixed and sequenced together. Then the barcode can be recognized by specific program and each single cell can be distinguished and profiled separately.

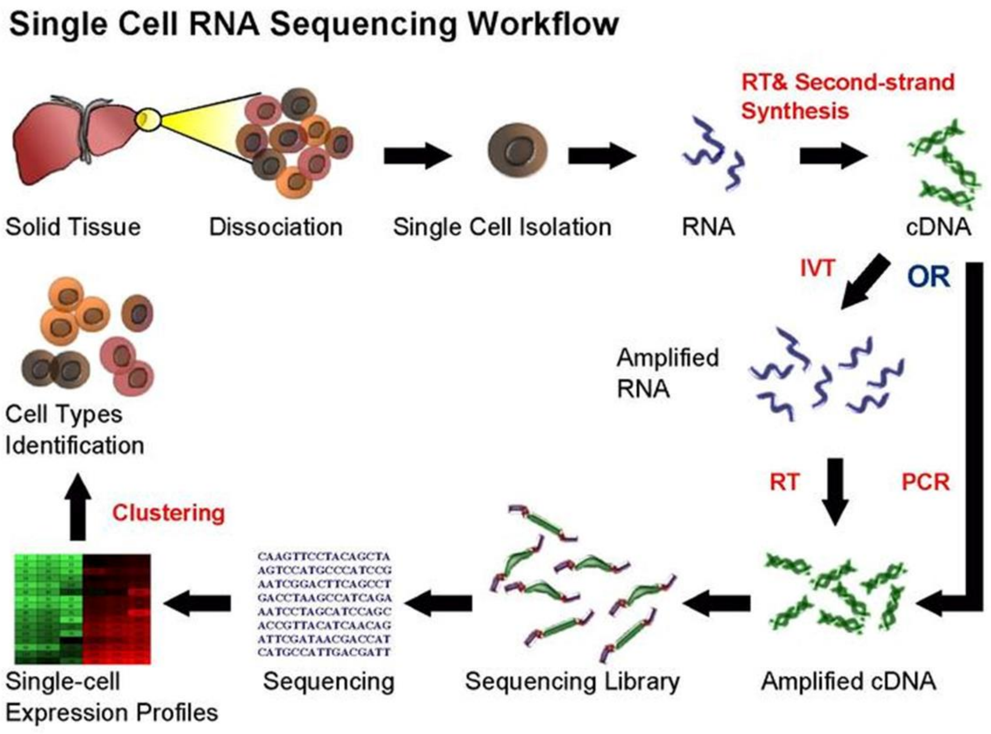


Fig. 5 Single cell RNA sequencing workflow. (Image source: Wikipedia)

4 Computational analysis of SCT data

Deep learning algorithms extract original features from large annotated SCT data sets (such as images or genomes) and use them to create a prediction tool based on hidden patterns. Once the training is completed, the algorithm can apply this training to analyze other data.

The challenges of SCT sequencing technology are isolating single cells, DNA amplification and data processing (quality control, error correction and bias removal). Issues of resulting SCT data are bias in captured cells, non-uniform amplification, mixing data from different protocols, amplification of errors, allele dropout, sampling bias of DNA fragments. Computational methods for processing and analyzing single-cell sequencing data begin from quantitative standards (spike-ins and unique molecular identifiers (UMIs)) and quality control, such as base quality, contamination, sample mix-up, batch effects and reproducibility based on replicates. Computational analysis of SCT data involves dimension reduction and clustering (e.g. t-SNE, PCA, ICA, Spectral t-SNE), hierarchical clustering, consensus clustering, mapping cell types/clusters across time, pseudo-time trajectories, spatial mapping of single cells, etc.

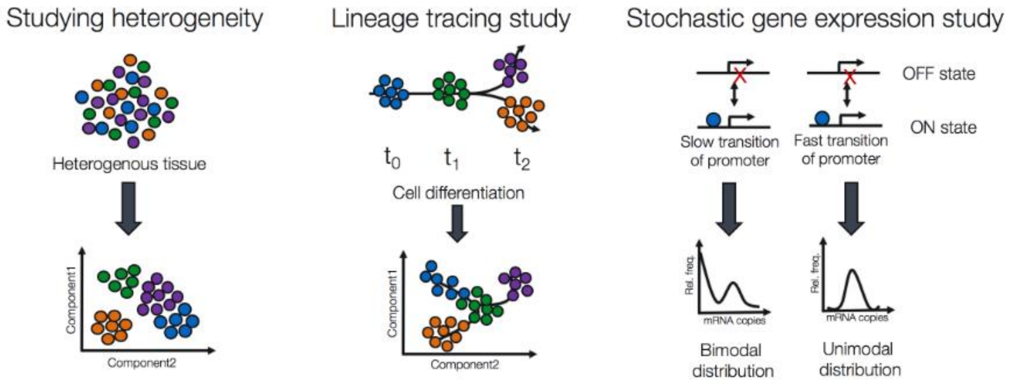


Fig. 6 Popular methods to address common investigations. [11]

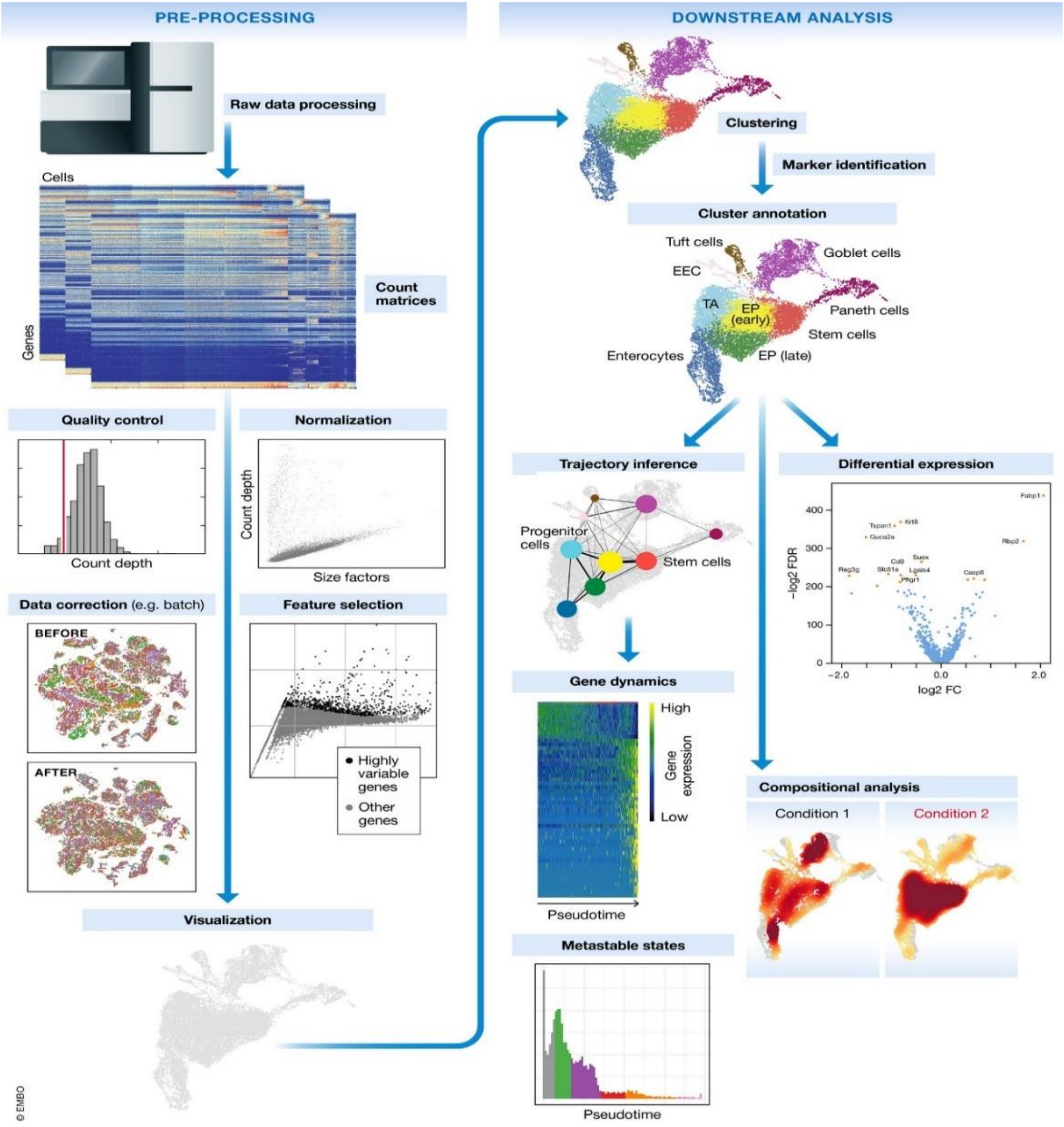


Fig. 7 Schematic on workflow for single cell RNA transcriptomics analyzing. [12]

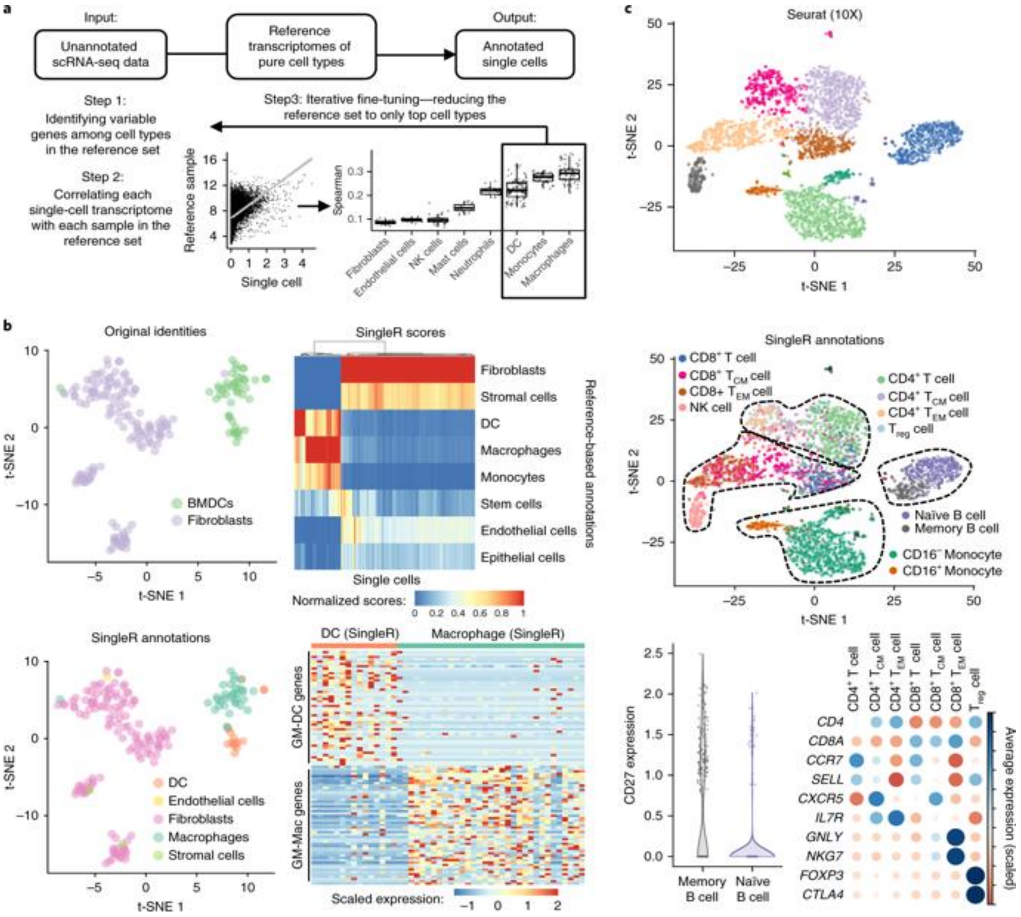


Fig. 8 Visualization analysis of single cell RNA transcriptomics data. [13]

Big data analysis of scRNA-seq can be used to explore which cell types are present in a tissue, identify unknown/rare cell types or states, elucidate the changes in gene expression during differentiation processes or across time or states, identify genes that are differentially expressed in particular cell types between conditions (e.g. treatment or disease), explore changes in expression among a cell type while incorporating spatial, regulatory, and/or protein information (Fig. 5). The pre-processing upstream analysis and the downstream analysis of single cell RNA transcriptomics can be shown as follows (Fig. 6, 7).

5 Supervised and unsupervised machine learning to SCT data

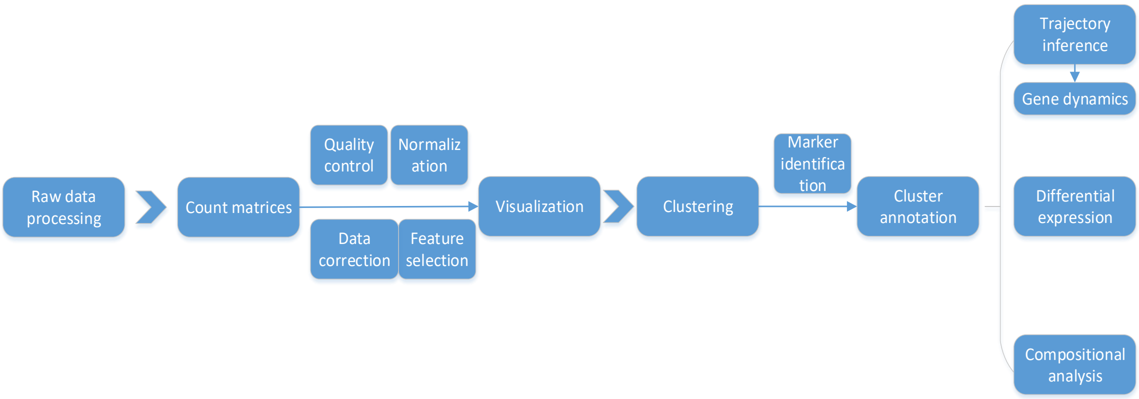


Fig. 9 A typical single-cell RNA-seq analysis workflow using unsupervised machine learning methods for one study at a time

Previous analysis methods of single cell RNA-seq has been supported by unsupervised learning methods, such as principal component analysis or clustering because repositories of known cases are lacking. The problem with unsupervised method is that number of classes is unknown. Data used in supervised learning is known and labeled (the class is known), while unsupervised learning it is not. Other problems are the accuracy of unsupervised learning is low and it has low sensitivity for highly dimensional data. However, we need to know the exact class of data subset and have the ontology of classes. Supervised learning such as artificial neural networks (ANN) can be used for further diagnostic purposes. [14-16]

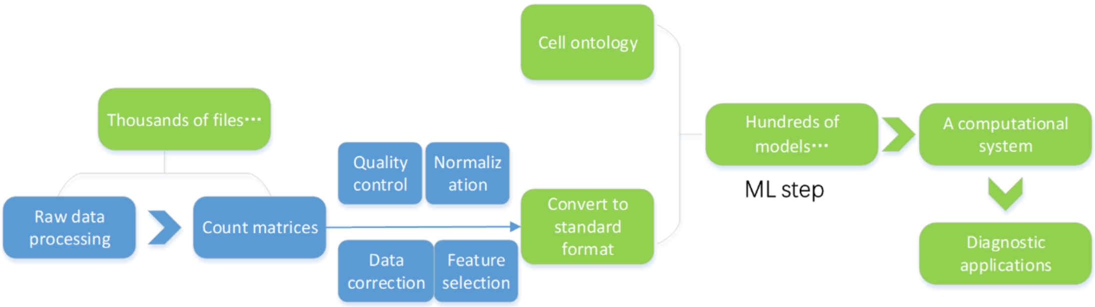


Fig. 10 Our single-cell RNA-seq analysis workflow using supervised machine learning methods.

Methodology

The project focuses on the development and implementation of supervised machine methods to do classification of single cells by the gene expression and utilization of this knowledge for improved molecular diagnostics of disease. About 250 data sets from human, mouse and other organisms single cell studies are available. These data have been collected from GEO database and 10X GemCode company demonstration supply [17, 18]. The data sets contain between 500 and 12,000 cells along with the metadata that describe the experiment and sample conditions. A computational system will be generated for recognizing critical features to distinguish different cell types with subsequent optimized machine learning model structures and parameters.

1. Raw SCT data collection and pre-processing.

2. Metadata construction and reference genome assembly building.

3. Feature selection/extraction.

4. Machine learning (ANN) training model building and optimization of model structure and parameters.

5. Case studies on different samples.

6. Set up a computational system.

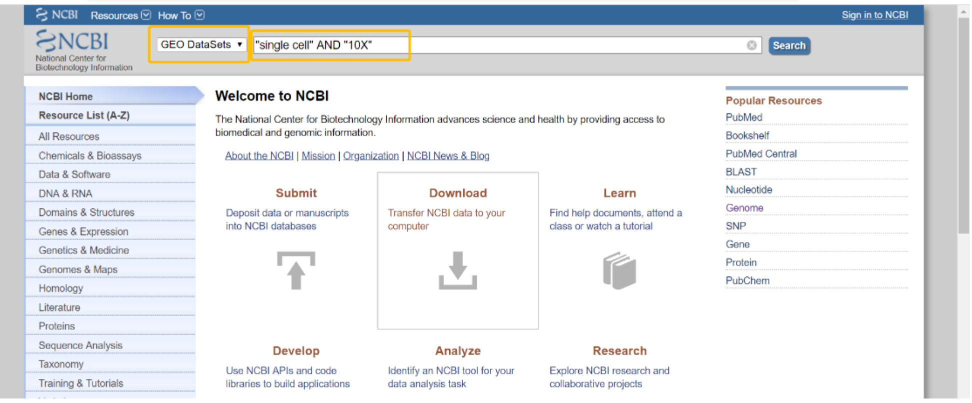
7. Testing with new updating SCT data sets.

8. Diagnostic applications.

Progress

1 Raw data collection (over 1,500 files)

The 10X single cell transcriptome sequencing data of the relevant articles published by 13th July 2019 were searched for using computer with the keywords - “single cell” AND “10X” in “GEO Datasets” (GEO (Gene Expression Omnibus) database of NCBI) (https://www.ncbi.nlm.nih.gov/). Raw data (matrix.mtx, barcodes.tsv, genes.tsv) of 10X related studies on GEO database by 13th July 2019 and single cell gene expression datasets on 10X genomics company website by 16th September 2019 have both been collected.



The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Deposit data or manuscripts into NCBI databases

Transfer NCBI data to your computer

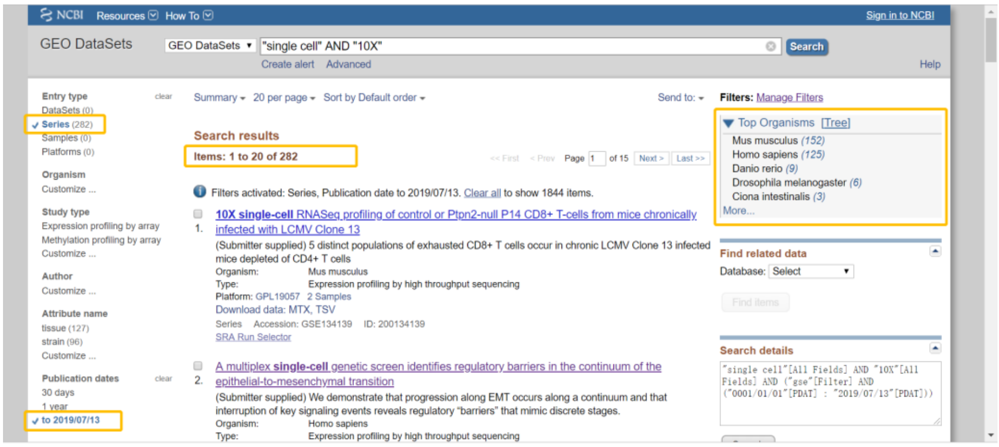
Find help documents, attend a class or watch a tutorial

Use NCBI APIs and code libraries to build applications

Identify an NCBI tool for your data analysis task

Explore NCBI research and collaborative projects

Fig. 11 GEO database searched with keywords “single cell” AND “10X”.



GEO DataSets "single cell" AND "10X"

Entry type (clear)
DataSets (0)
✓ Series (282)
Samples (0)
Platforms (0)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (127)
strain (96)
Customize ...

Publication dates (clear)
30 days
1 year
✓ to 2019/07/13

Search results

Items: 1 to 20 of 282

Filters activated: Series, Publication date to 2019/07/13. Clear all to show 1844 items.

1. 10X single-cell RNASeq profiling of control or Ptpn2-null P14 CD8+ T-cells from mice chronically infected with LCMV Clone 13
(Submitter supplied) 5 distinct populations of exhausted CD8+ T cells occur in chronic LCMV Clone 13 infected mice depleted of CD4+ T cells

Organism:	Mus musculus
Type:	Expression profiling by high throughput sequencing

Platform: GPL19057 2 Samples
Download data: MTX, TSV
Series Accession: GSE134139 ID: 200134139
SRA Run Selector

2. A multiplex single-cell genetic screen identifies regulatory barriers in the continuum of the epithelial-to-mesenchymal transition
(Submitter supplied) We demonstrate that progression along EMT occurs along a continuum and that interruption of key signaling events reveals regulatory "barriers" that mimic discrete stages.

Organism:	Homo sapiens
Type:	Expression profiling by high throughput sequencing

Filters: Manage Filters

Top Organisms [Tree]
Mus musculus (152)
Homo sapiens (125)
Danio rerio (9)
Drosophila melanogaster (6)
Ciona intestinalis (3)
More...

Find related data
Database: Select

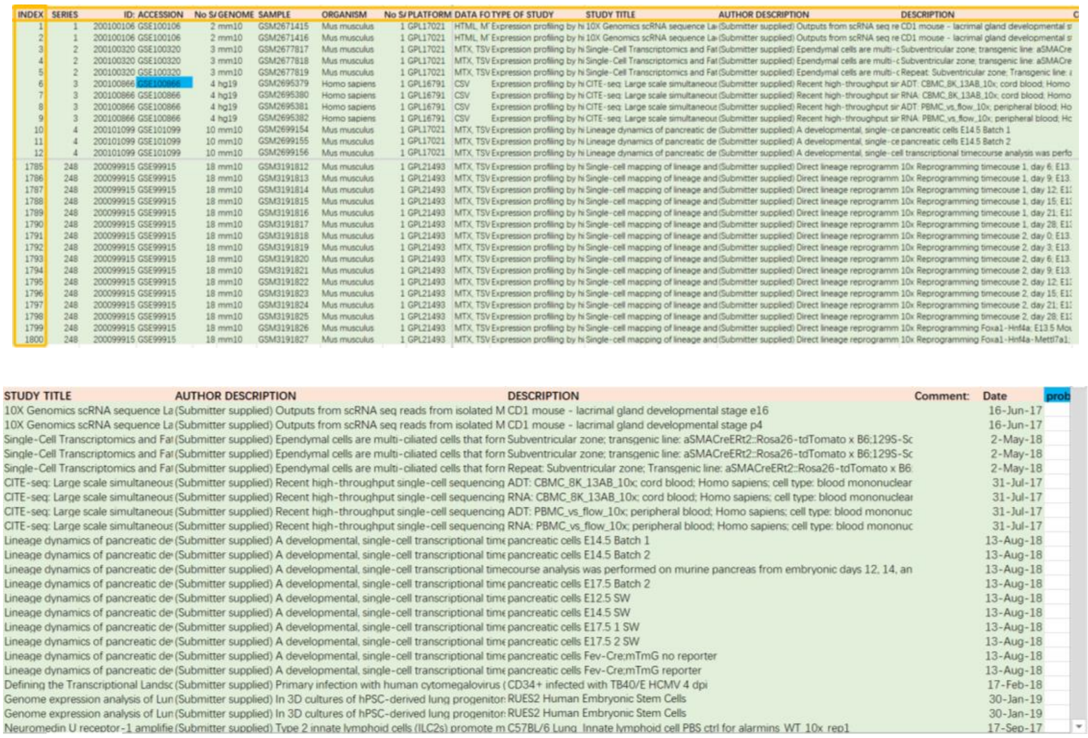
Search details
```
"single cell"[All Fields] AND "10X"[All Fields] AND ("gse"[Filter] AND ("0001/01/01"[PDAT] : "2019/07/13"[PDAT]))
```

Fig. 12 Data collected from relevant articles published by July 13th, 2019.

2 Metadata construction

The aggregated data annotation of the studies has been arrayed into a Metadata chart form, which is designed with "INDEX", "SERIES", "ACCESSION", "GENOME", "ORGANISM", etc. as the captions of each column in Microsoft Excel 2016. The metadata is sorted by "ACCESSION", that is the number name of series (e.g. GSE119561). ACCESSION is arranged in order from small to large, from top to bottom. This is very important to the follow-up work, because it can be found that many related data sets have very similar series numbers. Only 10X relevant research is included in metadata, other research with other single cell transcriptomics technologies (e.g. Drop-seq, SMART-seq, inDrop, etc.) of the same super series is not involved in. Sample number (e.g. GSM3377671) is unique for each 10X study.

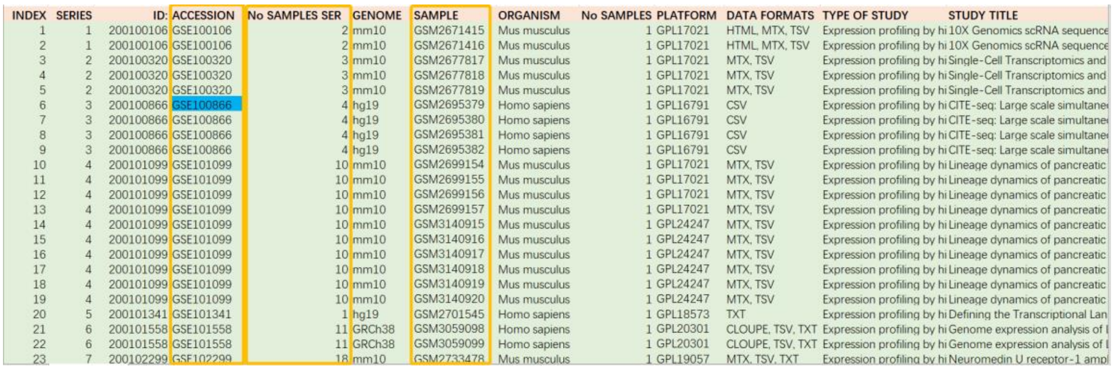
The metadata totally has 15 kinds of organisms – "Mus musculus", "Homo sapiens", "Danio rerio", "Ambystoma mexicanum", "Gallus gallus", etc. The metadata has 248 series mapped with 1800 samples, the description of each study is involved in the metadata and some of them has additional comments.



INDEX	SERIES	ID:	ACCESSION	No S/	GENOME	SAMPLE	ORGANISM	No S/	PLATFORM	DATA FO	TYPE OF STUDY	STUDY TITLE	AUTHOR DESCRIPTION	DESCRIPTION	C
1	1	200100106	GSE100106	2	mm10	GSM2671415	Mus musculus	1	GPL17021	HTML, M	Expression profiling by hi	10X Genomics scRNA sequence La	(Submitter supplied) Outputs from scRNA seq re	CD1 mouse - lacrimal gland developmental s	
2	1	200100106	GSE100106	2	mm10	GSM2671416	Mus musculus	1	GPL17021	HTML, M	Expression profiling by hi	10X Genomics scRNA sequence La	(Submitter supplied) Outputs from scRNA seq re	CD1 mouse - lacrimal gland developmental s	
3	2	200100320	GSE100320	3	mm10	GSM2677817	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-c	Subventricular zone; transgenic line: aSMACre	
4	2	200100320	GSE100320	3	mm10	GSM2677818	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-c	Subventricular zone; transgenic line: aSMACre	
5	2	200100320	GSE100320	3	mm10	GSM2677819	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-c	Repeat: Subventricular zone; Transgenic line: a	
6	3	200100866	GSE100866	4	hg19	GSM2695379	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultaneou	(Submitter supplied) Recent high-throughput sir	ADT: CBMC_8K_13AB_10x; cord blood; Homo	
7	3	200100866	GSE100866	4	hg19	GSM2695380	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultaneou	(Submitter supplied) Recent high-throughput sir	RNA: CBMC_8K_13AB_10x; cord blood; Homo	
8	3	200100866	GSE100866	4	hg19	GSM2695381	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultaneou	(Submitter supplied) Recent high-throughput sir	ADT: PBMC_vs_flow_10x; peripheral blood; Ho	
9	3	200100866	GSE100866	4	hg19	GSM2695382	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultaneou	(Submitter supplied) Recent high-throughput sir	RNA: PBMC_vs_flow_10x; peripheral blood; Hc	
10	4	200101099	GSE101099	10	mm10	GSM2699154	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-ce	pancreatic cells E14.5 Batch 1	
11	4	200101099	GSE101099	10	mm10	GSM2699155	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-ce	pancreatic cells E14.5 Batch 2	
12	4	200101099	GSE101099	10	mm10	GSM2699156	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional timecourse analysis was perfo		
1785	248	200099915	GSE99915	18	mm10	GSM3191812	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 6; E13.	
1786	248	200099915	GSE99915	18	mm10	GSM3191813	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 9; E13.	
1787	248	200099915	GSE99915	18	mm10	GSM3191814	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 12; E1	
1788	248	200099915	GSE99915	18	mm10	GSM3191815	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 15; E1	
1789	248	200099915	GSE99915	18	mm10	GSM3191816	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 21; E1	
1790	248	200099915	GSE99915	18	mm10	GSM3191817	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 1, day 28; E1	
1791	248	200099915	GSE99915	18	mm10	GSM3191818	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 0; E13.	
1792	248	200099915	GSE99915	18	mm10	GSM3191819	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 3; E13.	
1793	248	200099915	GSE99915	18	mm10	GSM3191820	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 6; E13.	
1794	248	200099915	GSE99915	18	mm10	GSM3191821	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 9; E13.	
1795	248	200099915	GSE99915	18	mm10	GSM3191822	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 12; E1	
1796	248	200099915	GSE99915	18	mm10	GSM3191823	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 15; E1	
1797	248	200099915	GSE99915	18	mm10	GSM3191824	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 21; E1	
1798	248	200099915	GSE99915	18	mm10	GSM3191825	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming timecouse 2, day 28; E1	
1799	248	200099915	GSE99915	18	mm10	GSM3191826	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming Foxa1-Hnf4a; E13.5 Mo	
1800	248	200099915	GSE99915	18	mm10	GSM3191827	Mus musculus	1	GPL21493	MTX, TSV	Expression profiling by hi	Single-cell mapping of lineage and	(Submitter supplied) Direct lineage reprogramm	10x Reprogramming Foxa1-Hnf4a-Mettl7a1;	

STUDY TITLE	AUTHOR DESCRIPTION	DESCRIPTION	Comment:	Date	prob
10X Genomics scRNA sequence La	(Submitter supplied) Outputs from scRNA seq reads from isolated M	CD1 mouse - lacrimal gland developmental stage e16		16-Jun-17	
10X Genomics scRNA sequence La	(Submitter supplied) Outputs from scRNA seq reads from isolated M	CD1 mouse - lacrimal gland developmental stage p4		16-Jun-17	
Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-ciliated cells that forn	Subventricular zone; transgenic line: aSMACreERt2::Rosa26-tdTomato x B6;129S-Sc		2-May-18	
Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-ciliated cells that forn	Subventricular zone; transgenic line: aSMACreERt2::Rosa26-tdTomato x B6;129S-Sc		2-May-18	
Single-Cell Transcriptomics and Fat	(Submitter supplied) Ependymal cells are multi-ciliated cells that forn	Repeat: Subventricular zone; Transgenic line: aSMACreERt2::Rosa26-tdTomato x B6		2-May-18	
CITE-seq: Large scale simultaneous	(Submitter supplied) Recent high-throughput single-cell sequencing	ADT: CBMC_8K_13AB_10x; cord blood; Homo sapiens; cell type: blood mononuclear		31-Jul-17	
CITE-seq: Large scale simultaneous	(Submitter supplied) Recent high-throughput single-cell sequencing	RNA: CBMC_8K_13AB_10x; cord blood; Homo sapiens; cell type: blood mononuclear		31-Jul-17	
CITE-seq: Large scale simultaneous	(Submitter supplied) Recent high-throughput single-cell sequencing	ADT: PBMC_vs_flow_10x; peripheral blood; Homo sapiens; cell type: blood mononuc		31-Jul-17	
CITE-seq: Large scale simultaneous	(Submitter supplied) Recent high-throughput single-cell sequencing	RNA: PBMC_vs_flow_10x; peripheral blood; Homo sapiens; cell type: blood mononuc		31-Jul-17	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E14.5 Batch 1		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E14.5 Batch 2		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional timecourse analysis was performed on murine pancreas from embryonic days 12, 14, an			13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E17.5 Batch 2		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E12.5 SW		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E14.5 SW		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E17.5 1 SW		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells E17.5 2 SW		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells Fev-Cre;mTmG no reporter		13-Aug-18	
Lineage dynamics of pancreatic de	(Submitter supplied) A developmental, single-cell transcriptional time	pancreatic cells Fev-Cre;mTmG reporter		13-Aug-18	
Defining the Transcriptional Landsc	(Submitter supplied) Primary infection with human cytomegalovirus (	CD34+ infected with TB40/E HCMV 4 dpi		17-Feb-18	
Genome expression analysis of Lun	(Submitter supplied) In 3D cultures of hPSC-derived lung progenitor	RUES2 Human Embryonic Stem Cells		30-Jan-19	
Genome expression analysis of Lun	(Submitter supplied) In 3D cultures of hPSC-derived lung progenitor	RUES2 Human Embryonic Stem Cells		30-Jan-19	
Neuromedin U receptor-1 amplifie	(Submitter supplied) Type 2 innate lymphoid cells (ILC2s) promote m	C57BL/6 Lung Innate lymphoid cell PBS ctrl for alarmins WT 10x rep1		17-Sep-17	

Fig. 13 The whole picture of Metadata chart form.

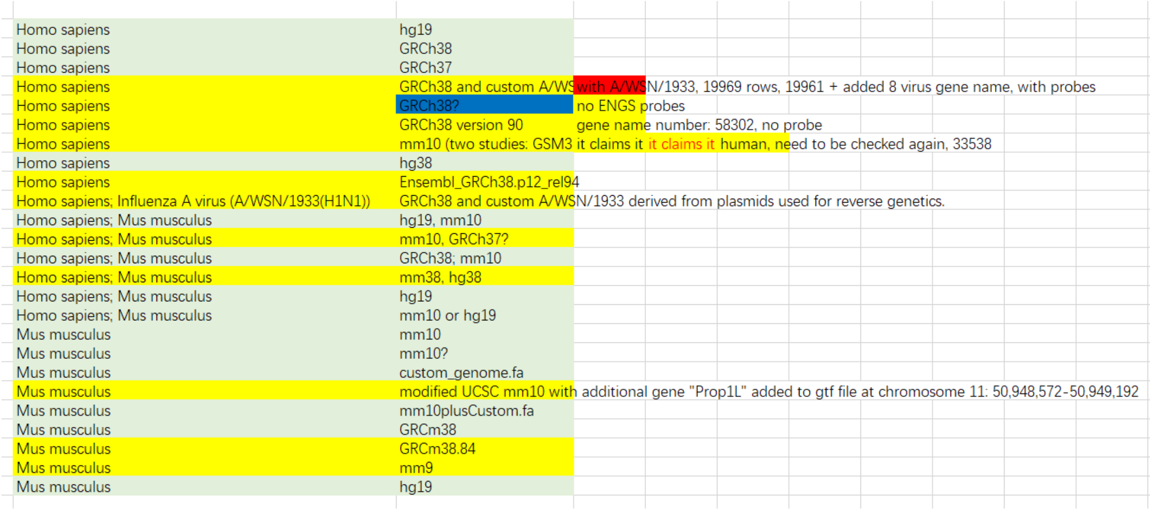


INDEX	SERIES	ID:	ACCESSION	No SAMPLES SER	GENOME	SAMPLE	ORGANISM	No SAMPLES	PLATFORM	DATA FORMATS	TYPE OF STUDY	STUDY TITLE
1	1	200100106	GSE100106	2	mm10	GSM2671415	Mus musculus	1	GPL17021	HTML, MTX, TSV	Expression profiling by hi	10X Genomics scRNA sequence
2	1	200100106	GSE100106	2	mm10	GSM2671416	Mus musculus	1	GPL17021	HTML, MTX, TSV	Expression profiling by hi	10X Genomics scRNA sequence
3	2	200100320	GSE100320	3	mm10	GSM2677817	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and
4	2	200100320	GSE100320	3	mm10	GSM2677818	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and
5	2	200100320	GSE100320	3	mm10	GSM2677819	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Single-Cell Transcriptomics and
6	3	200100866	GSE100866	4	hg19	GSM2695379	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultane
7	3	200100866	GSE100866	4	hg19	GSM2695380	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultane
8	3	200100866	GSE100866	4	hg19	GSM2695381	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultane
9	3	200100866	GSE100866	4	hg19	GSM2695382	Homo sapiens	1	GPL16791	CSV	Expression profiling by hi	CITE-seq: Large scale simultane
10	4	200101099	GSE101099	10	mm10	GSM2699154	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
11	4	200101099	GSE101099	10	mm10	GSM2699155	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
12	4	200101099	GSE101099	10	mm10	GSM2699156	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
13	4	200101099	GSE101099	10	mm10	GSM2699157	Mus musculus	1	GPL17021	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
14	4	200101099	GSE101099	10	mm10	GSM3140915	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
15	4	200101099	GSE101099	10	mm10	GSM3140916	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
16	4	200101099	GSE101099	10	mm10	GSM3140917	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
17	4	200101099	GSE101099	10	mm10	GSM3140918	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
18	4	200101099	GSE101099	10	mm10	GSM3140919	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
19	4	200101099	GSE101099	10	mm10	GSM3140920	Mus musculus	1	GPL24247	MTX, TSV	Expression profiling by hi	Lineage dynamics of pancreatic
20	5	200101341	GSE101341	1	hg19	GSM2701545	Homo sapiens	1	GPL18573	TXT	Expression profiling by hi	Defining the Transcriptional Lan
21	6	200101558	GSE101558	11	GRCh38	GSM3059098	Homo sapiens	1	GPL20301	CLOUPE, TSV, TXT	Expression profiling by hi	Genome expression analysis of l
22	6	200101558	GSE101558	11	GRCh38	GSM3059099	Homo sapiens	1	GPL20301	CLOUPE, TSV, TXT	Expression profiling by hi	Genome expression analysis of l
23	7	200102299	GSE102299	18	mm10	GSM2733478	Mus musculus	1	GPL19057	MTX, TSV, TXT	Expression profiling by hi	Neuromedin U receptor-1 ampl

Fig. 14 Components of Metadata.

Kingdom	Phylum	Organism	Genome	Number of sample in metadata
Virus	Riboviria (Influenza virus)	Canis lupus familiaris; Influenza A virus (A/WSN/1933(H1N1)); Homo sapiens	GRCh38, CanFam3.1, Influenza A virus (A/WSN/1933(H1N1))	
Fungi	Ascomycota	Saccharomyces cerevisiae	Saccharomyces_cerevisiae R64 ?	
			sacCer3	
Plantae	Angiosperms	Arabidopsis thaliana	TAIR10	
Animalia	Nematoda	Caenorhabditis elegans	WS260	
	Platyhelminthes	Schmidtea mediterranea	GSE72389?	
	Arthropoda	Drosophila melanogaster	dm6	
			dm3	
			Drosophila melanogaster Release 6	
			BDGP6 version 87	
	Chordata	Ciona intestinalis	KH2012	
		Ambystoma mexicanum	Am_2.2	
		Danio rerio	dr82?	
			dr82	
			GRCz10	
			GRCz11/danRer11	
			GRCz11	
		Gallus gallus	ENSEMBL Gallus gallus 5.0	
		Sus scrofa	Sscrofa11 Release 91	
		Rattus norvegicus	Rnor6 Release 92	
		Mus musculus	mm10	
			mm10?	
			mm10 (GRCm38)	
			custom_genome.fa	2
			GRCm38	
			GRCm38.84	
			mm9	
			mm10 or hg19	
			hg19	
		Homo sapiens; Mus musculus	hg19, mm10	
			mm10, GRCh37?	
			GRCh38; mm10	
			mm38, hg38	
			hg19	
			mm10 or hg19	
		Homo sapiens	hg19	
			GRCh38	
			GRCh37	
			GRCh38?	
			mm10	
			hg38	
			hg19?	
			Ensembl_GRCh38.p12_rel94	

Fig. 15 Total organisms and versions of genome sorted out from metadata.



Homo sapiens	hg19	
Homo sapiens	GRCh38	
Homo sapiens	GRCh37	
Homo sapiens	GRCh38 and custom A/WS	with A/WSN/1933, 19969 rows, 19961 + added 8 virus gene name, with probes
Homo sapiens	GRCh38?	no ENGS probes
Homo sapiens	GRCh38 version 90	gene name number: 58302, no probe
Homo sapiens	mm10 (two studies: GSM3	it claims it it claims it human, need to be checked again, 33538
Homo sapiens	hg38	
Homo sapiens	Ensembl_GRCh38.p12_rel94	
Homo sapiens; Influenza A virus (A/WSN/1933(H1N1))	GRCh38 and custom A/WSN/1933 derived from plasmids used for reverse genetics.	
Homo sapiens; Mus musculus	hg19, mm10	
Homo sapiens; Mus musculus	mm10, GRCh37?	
Homo sapiens; Mus musculus	GRCh38; mm10	
Homo sapiens; Mus musculus	mm38, hg38	
Homo sapiens; Mus musculus	hg19	
Homo sapiens; Mus musculus	mm10 or hg19	
Mus musculus	mm10	
Mus musculus	mm10?	
Mus musculus	custom_genome.fa	
Mus musculus	modified UCSC mm10 with additional gene "Prop1L" added to gtf file at chromosome 11: 50,948,572-50,949,192	
Mus musculus	mm10plusCustom.fa	
Mus musculus	GRCm38	
Mus musculus	GRCm38.84	
Mus musculus	mm9	
Mus musculus	hg19	

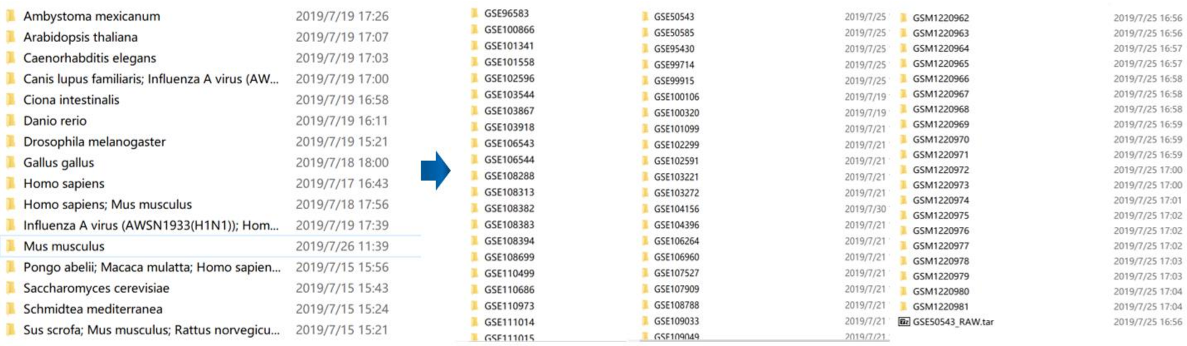
Fig. 16 Comparative analysis to different alternative genome of human and mouse.

3 Quality control

Studies which are not related but filtered by GEO browser with the key words are excluded (e.g. 10X Hank's salt solution). Series who have an inconsistent description of their studies are excluded out as well.

4 Data pre-processing (Data cleaning, filtering and standardization)

All relevant different types of data files have been downloaded ("download.py") according to series from GEO website and each of them has been cleaned and corrected one by one. Data files of super series studies have been split up. The compressed data from each downloaded file of each series is captured/read and uncompressed by the program "compress.py". Different formats of the matrix file (e.g. .h5, .csv, .tsv, .txt, .mtx (tsv)) of each sample has been converted (program "convert.py") into CSV file (.csv) for visualization, with cell barcodes/cell numbers as the horizontal heading, gene names as the vertical heading and gene expression numbers as the digital matrix.



Ambystoma mexicanum	2019/7/19 17:26
Arabidopsis thaliana	2019/7/19 17:07
Caenorhabditis elegans	2019/7/19 17:03
Canis lupus familiaris; Influenza A virus (AW...	2019/7/19 17:00
Ciona intestinalis	2019/7/19 16:58
Danio rerio	2019/7/19 16:11
Drosophila melanogaster	2019/7/19 15:21
Gallus gallus	2019/7/18 18:00
Homo sapiens	2019/7/17 16:43
Homo sapiens; Mus musculus	2019/7/18 17:56
Influenza A virus (AWSN1933(H1N1)); Hom...	2019/7/19 17:39
Mus musculus	2019/7/26 11:39
Pongo abelii; Macaca mulatta; Homo sapien...	2019/7/15 15:56
Saccharomyces cerevisiae	2019/7/15 15:43
Schmidtea mediterranea	2019/7/15 15:24
Sus scrofa; Mus musculus; Rattus norvegicu...	2019/7/15 15:21

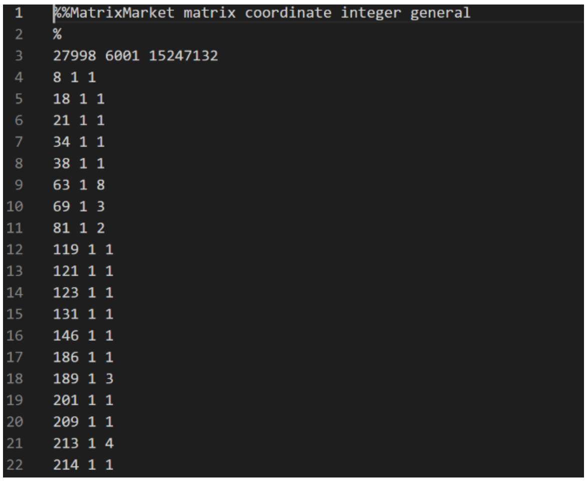
GSE96583
GSE100866
GSE101341
GSE101558
GSE102596
GSE103544
GSE103867
GSE103918
GSE106543
GSE106544
GSE108288
GSE108313
GSE108382
GSE108383
GSE108394
GSE108699
GSE110499
GSE110686
GSE110973
GSE111014
GSE111015

GSE50543	2019/7/25
GSE50585	2019/7/25
GSE95430	2019/7/25
GSE99714	2019/7/25
GSE99915	2019/7/25
GSE100106	2019/7/19
GSE100320	2019/7/19
GSE101099	2019/7/21
GSE102299	2019/7/21
GSE102591	2019/7/21
GSE103221	2019/7/21
GSE103272	2019/7/21
GSE104156	2019/7/30
GSE104396	2019/7/21
GSE106264	2019/7/21
GSE106960	2019/7/21
GSE107527	2019/7/21
GSE107909	2019/7/21
GSE108788	2019/7/21
GSE109033	2019/7/21
GSE109049	2019/7/21

GSM1220962	2019/7/25 16:56
GSM1220963	2019/7/25 16:56
GSM1220964	2019/7/25 16:57
GSM1220965	2019/7/25 16:57
GSM1220966	2019/7/25 16:58
GSM1220967	2019/7/25 16:58
GSM1220968	2019/7/25 16:58
GSM1220969	2019/7/25 16:59
GSM1220970	2019/7/25 16:59
GSM1220971	2019/7/25 16:59
GSM1220972	2019/7/25 17:00
GSM1220973	2019/7/25 17:00
GSM1220974	2019/7/25 17:01
GSM1220975	2019/7/25 17:02
GSM1220976	2019/7/25 17:02
GSM1220977	2019/7/25 17:02
GSM1220978	2019/7/25 17:03
GSM1220979	2019/7/25 17:03
GSM1220980	2019/7/25 17:04
GSM1220981	2019/7/25 17:04
GSE50543_RAW.tar	2019/7/25 16:56

Fig. 17 Thousands of data files downloaded and cleaned.

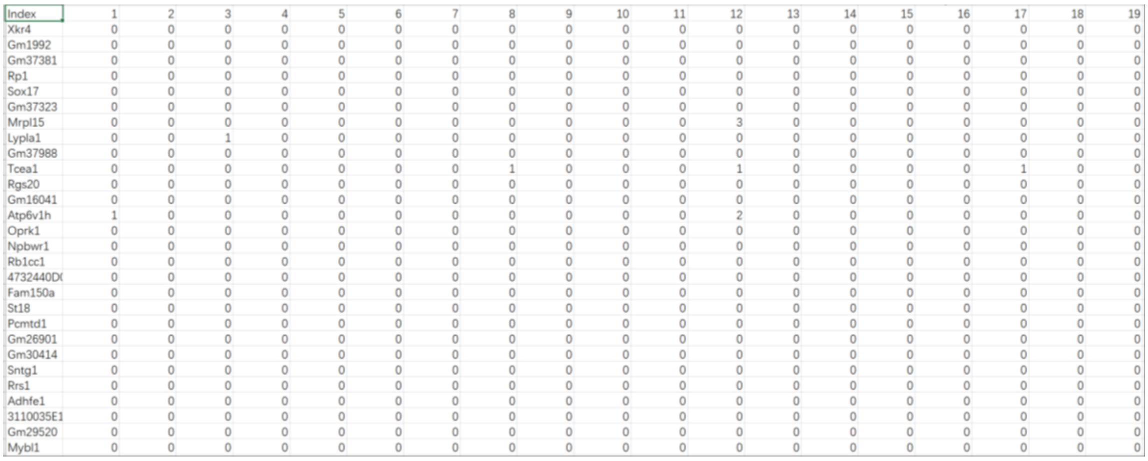
Data filtering (program “filter_data.py”) has been done by the exclusion of null data, which can be caused by the count of empty plates in the experimental protocol of single cell transcriptomics – the cell capture rate is zero at this situation. For example, in the raw data of the study sample GSMXXXXXXX, the cell barcode is 70,0000, however the actual gene expression is only ______, which means it calculates lots of empty gene expression, so we filtered output the actual meaningful data by removing the null data in each matrix of each raw data file.



```
%%MatrixMarket matrix coordinate integer general
%
27998 6001 15247132
8 1 1
18 1 1
21 1 1
34 1 1
38 1 1
63 1 8
69 1 3
81 1 2
119 1 1
121 1 1
123 1 1
131 1 1
146 1 1
186 1 1
189 1 3
201 1 1
209 1 1
213 1 4
214 1 1
```

Fig. 18 MTX file needs to be converted to CSV file for better visualization.

Data conversion has been conducted by transforming the produced CSV file into four different kinds of standard file formats - .h5, .csv, .npz, .mtx (tsv), which have been selected and decided to be the common, unified and standardized output format for various purpose of use, for example, visualization or statistical utilization. The program (“convert.py”) has been prepared for this process.



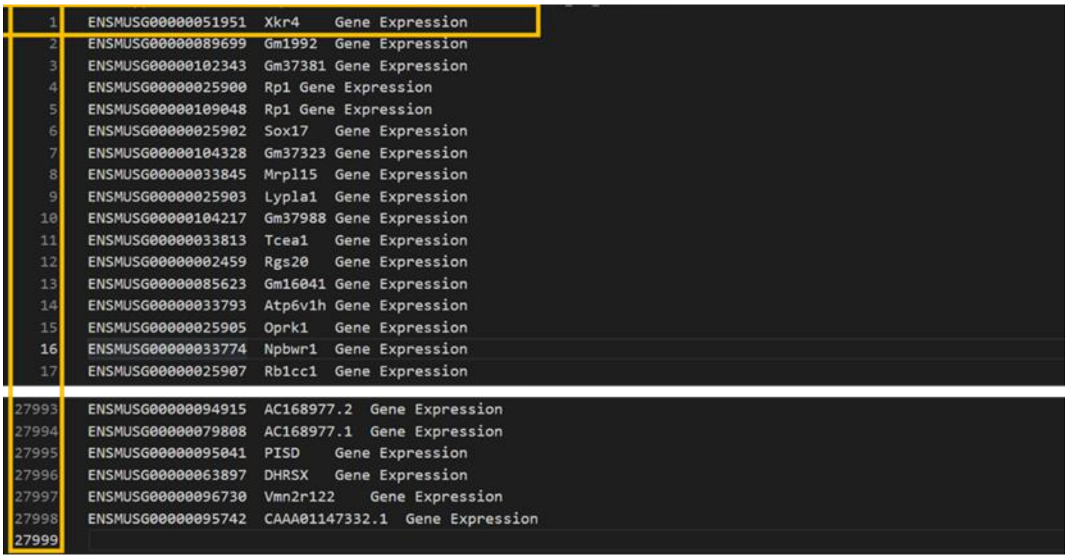
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Xkr4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm1992	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm37381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rp1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sox17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm37323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mrpl15	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Lypla1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm37988	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tcea1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0
Rgs20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm16041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Atp6v1h	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Oprk1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Npbwr1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rb1cc1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4732440D(	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fam150a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
St18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pcmtd1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm26901	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm30414	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sntg1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Rrs1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Adhfe1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3110035E1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gm29520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mybl1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 19 Data count matrix for one sample as an example.

5 Reference genome assembly selected and built

Genome assembly is the gene name database comprises the names and IDs of all known genes so far, it is wanted as the annotation tracks available. Different genome version is used in different studies.

The alteration of genomic versions and the lack of uniform naming standards have led to very complex confusion. One gene name can have several different probes name, this is not comparable between two different genomes of one same organism. High quality and highly accurate control would be decided to exclude some of the studies which only supply gene name list without probes number or those only have probes number list without gene name list. Sometimes one probe corresponds to different gene names (synonym or alias). Clear lists of fullest gene names and common gene names should be made as the reference genome for each organism, with the probes mapping to the genes. Reference genome assembly needs to be found and serves in the follow-up machine learning section, which contains the common list of gene names appeared in every study of one same species.



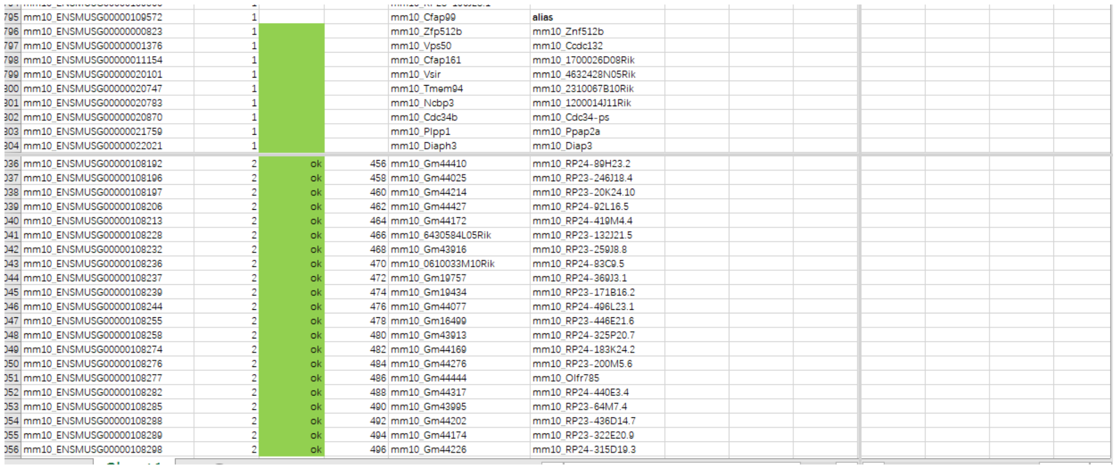
```
    1  ENSMUSG00000051951  Xkr4    Gene Expression
    2  ENSMUSG00000089699  Gm1992  Gene Expression
    3  ENSMUSG00000102343  Gm37381 Gene Expression
    4  ENSMUSG00000025900  Rp1 Gene Expression
    5  ENSMUSG00000109048  Rp1 Gene Expression
    6  ENSMUSG00000025902  Sox17   Gene Expression
    7  ENSMUSG00000104328  Gm37323 Gene Expression
    8  ENSMUSG00000033845  Mrpl15  Gene Expression
    9  ENSMUSG00000025903  Lypla1  Gene Expression
   10  ENSMUSG00000104217  Gm37988 Gene Expression
   11  ENSMUSG00000033813  Tcea1   Gene Expression
   12  ENSMUSG00000002459  Rgs20   Gene Expression
   13  ENSMUSG00000085623  Gm16041 Gene Expression
   14  ENSMUSG00000033793  Atp6v1h Gene Expression
   15  ENSMUSG00000025905  Oprk1   Gene Expression
   16  ENSMUSG00000033774  Npbwr1  Gene Expression
   17  ENSMUSG00000025907  Rb1cc1  Gene Expression

27993  ENSMUSG00000094915  AC168977.2  Gene Expression
27994  ENSMUSG00000079808  AC168977.1  Gene Expression
27995  ENSMUSG00000095041  PISD    Gene Expression
27996  ENSMUSG00000063897  DHRSX   Gene Expression
27997  ENSMUSG00000096730  Vmn2r122    Gene Expression
27998  ENSMUSG00000095742  CAAA01147332.1  Gene Expression
27999
```

Fig. 20 Genome assembly used in GSM3937878 as an example.

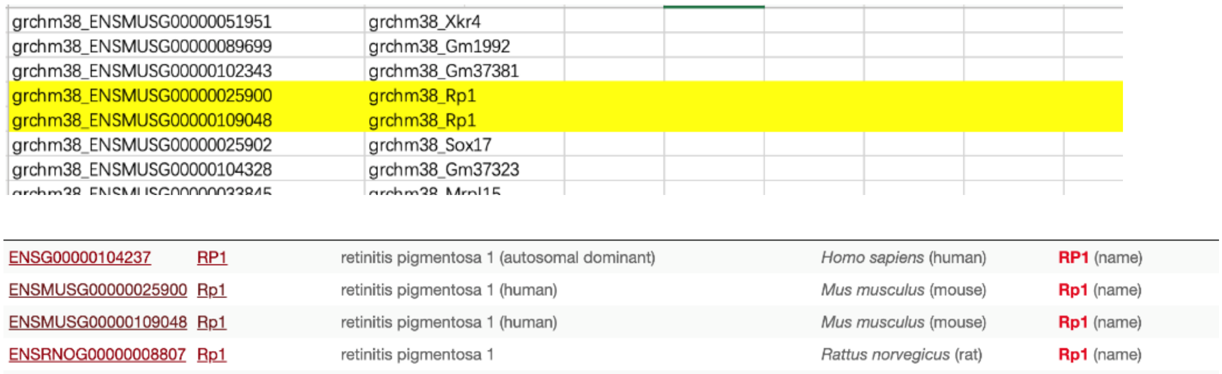
NCBI, ENSEMBL and UCSC provide three most complete genome databases as the three well-known genome browsers retrieving genomic information. The reference genome assembly of different organism has been searched from ENSEMBL genome browser as the current version, with the reason that the number probes of ENSEMBL genome assembly are updated continuously and regularly.

Reference genome assembly is created for each different organism separately. It is created with two columns in Excel using the gene file (.tsv) supplied in the sample webpage, one is Ensembl Gene ID and one is Gene Name. It is used to map Ensembl Gene ID (e.g. ENSG00000210049) to Gene Name (e.g. MT-TF) for each organism. When it comes to the organisms which have enormous amount of studies, like Homo sapiens and Mus musculus, etc., five genome builds have been selected randomly in different samples of different series and have been compared and merged to make the final reference assembly, which is named as “common list”.



795	mm10_ENSMUSG00000109572	1			mm10_Cfap99	**alias**
796	mm10_ENSMUSG00000000823	1			mm10_Zfp512b	mm10_Znf512b
797	mm10_ENSMUSG00000001376	1			mm10_Vps50	mm10_Ccdc132
798	mm10_ENSMUSG00000011154	1			mm10_Cfap161	mm10_1700026D08Rik
799	mm10_ENSMUSG00000020101	1			mm10_Vsir	mm10_4632428N05Rik
800	mm10_ENSMUSG00000020747	1			mm10_Tmem94	mm10_2310067B10Rik
801	mm10_ENSMUSG00000020783	1			mm10_Ncbp3	mm10_1200014J11Rik
802	mm10_ENSMUSG00000020870	1			mm10_Cdc34b	mm10_Cdc34-ps
803	mm10_ENSMUSG00000021759	1			mm10_Plpp1	mm10_Ppap2a
804	mm10_ENSMUSG00000022021	1			mm10_Diaph3	mm10_Diap3
036	mm10_ENSMUSG00000108192	2	ok	456	mm10_Gm44410	mm10_RP24-89H23.2
037	mm10_ENSMUSG00000108196	2	ok	458	mm10_Gm44025	mm10_RP23-246J18.4
038	mm10_ENSMUSG00000108197	2	ok	460	mm10_Gm44214	mm10_RP23-20K24.10
039	mm10_ENSMUSG00000108206	2	ok	462	mm10_Gm44427	mm10_RP24-92L16.5
040	mm10_ENSMUSG00000108213	2	ok	464	mm10_Gm44172	mm10_RP24-419M4.4
041	mm10_ENSMUSG00000108228	2	ok	466	mm10_6430584L05Rik	mm10_RP23-132J21.5
042	mm10_ENSMUSG00000108232	2	ok	468	mm10_Gm43916	mm10_RP23-259J8.8
043	mm10_ENSMUSG00000108236	2	ok	470	mm10_0610033M10Rik	mm10_RP24-83C9.5
044	mm10_ENSMUSG00000108237	2	ok	472	mm10_Gm19757	mm10_RP24-369J3.1
045	mm10_ENSMUSG00000108239	2	ok	474	mm10_Gm19434	mm10_RP23-171B16.2
046	mm10_ENSMUSG00000108244	2	ok	476	mm10_Gm44077	mm10_RP24-496L23.1
047	mm10_ENSMUSG00000108255	2	ok	478	mm10_Gm16499	mm10_RP23-446E21.6
048	mm10_ENSMUSG00000108258	2	ok	480	mm10_Gm43913	mm10_RP24-325P20.7
049	mm10_ENSMUSG00000108274	2	ok	482	mm10_Gm44169	mm10_RP24-183K24.2
050	mm10_ENSMUSG00000108276	2	ok	484	mm10_Gm44276	mm10_RP23-200M5.6
051	mm10_ENSMUSG00000108277	2	ok	486	mm10_Gm44444	mm10_Olfr785
052	mm10_ENSMUSG00000108282	2	ok	488	mm10_Gm44317	mm10_RP24-440E3.4
053	mm10_ENSMUSG00000108285	2	ok	490	mm10_Gm43995	mm10_RP23-64M7.4
054	mm10_ENSMUSG00000108288	2	ok	492	mm10_Gm44202	mm10_RP23-436D14.7
055	mm10_ENSMUSG00000108289	2	ok	494	mm10_Gm44174	mm10_RP23-322E20.9
056	mm10_ENSMUSG00000108298	2	ok	496	mm10_Gm44226	mm10_RP24-315D19.3

Fig. 21 Genome assembly with its alias in Mouse genome mm10 as an example.



grchm38_ENSMUSG00000051951	grchm38_Xkr4
grchm38_ENSMUSG00000089699	grchm38_Gm1992
grchm38_ENSMUSG00000102343	grchm38_Gm37381
grchm38_ENSMUSG00000025900	grchm38_Rp1
grchm38_ENSMUSG00000109048	grchm38_Rp1
grchm38_ENSMUSG00000025902	grchm38_Sox17
grchm38_ENSMUSG00000104328	grchm38_Gm37323

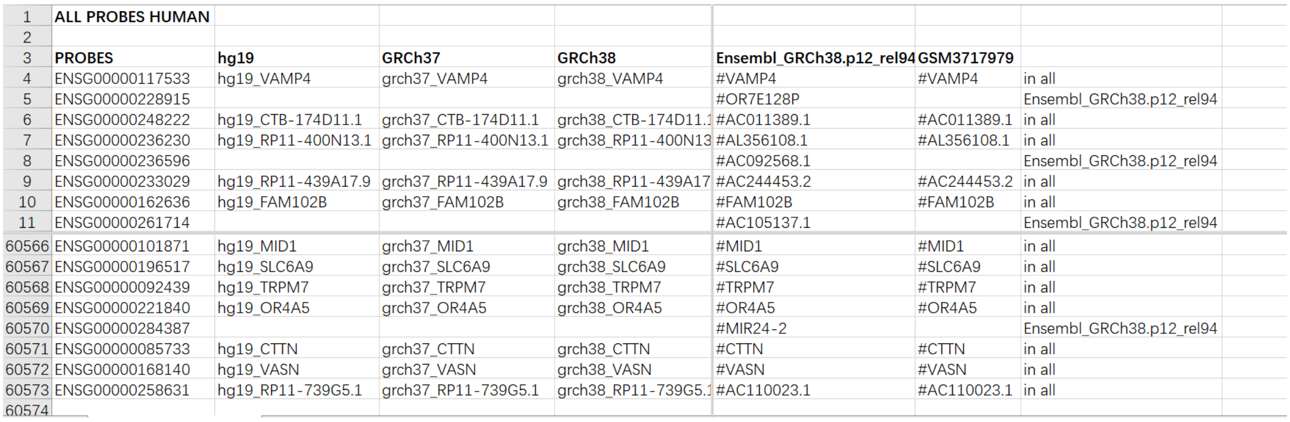
ENSG00000104237	RP1	retinitis pigmentosa 1 (autosomal dominant)	*Homo sapiens* (human)	**RP1** (name)
ENSMUSG00000025900	Rp1	retinitis pigmentosa 1 (human)	*Mus musculus* (mouse)	**Rp1** (name)
ENSMUSG00000109048	Rp1	retinitis pigmentosa 1 (human)	*Mus musculus* (mouse)	**Rp1** (name)
ENSRNOG00000008807	Rp1	retinitis pigmentosa 1	*Rattus norvegicus* (rat)	**Rp1** (name)

Fig. 22 An example for the same gene name with different probe numbers in one genome version (GRCm 38).

Correction has been made when the genomes adopted in some studies show the wrong data format, the decimal point in probe, the space keys, confused/mixed genome version and the incorrect naming. Corrected and cleaned genome file is saved with the format ".txt" or ".tsv" but not ".csv", in case of Excel date format confusion. Also, the gene names and the gene IDs are renamed with adding organisms' names in front of them to distinguish properly. There are some genome related file supplied in sample webpage is showed in ".H5" file format, a short code ("h5_to_csv.py") is created to open ".H5" files properly. These cleared versions of genome mapping assemblies for each organism are used as the references for the follow-up machine learning section.

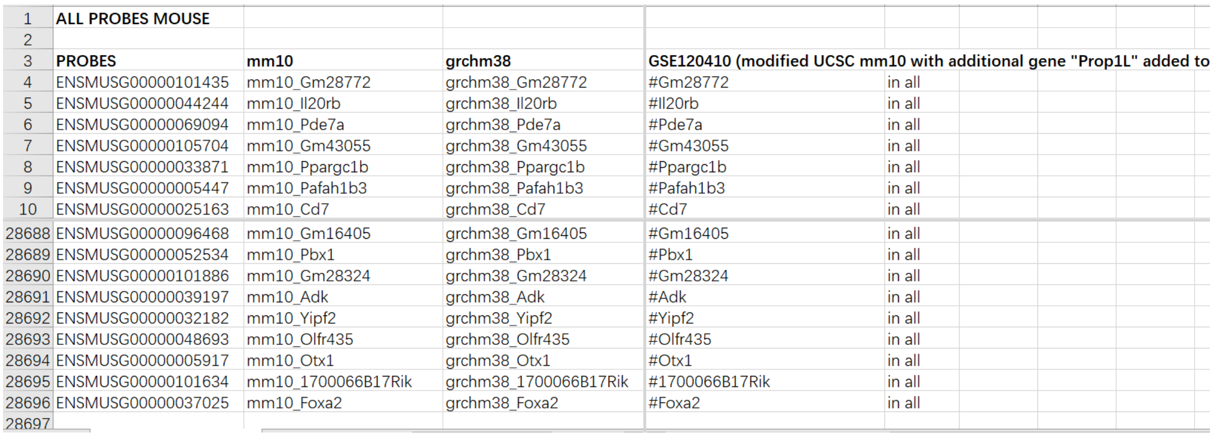
	Component	Common list probes number	Full list probes number	Note
Mouse	"grchm38_GSE132199.tsv" "mm10_2_GSM2928504_GSE109049.tsv" "mm10_3_GSM3272966_GSE117176.tsv" "mm10_4_GSM3937878_GSE134139.tsv" "mm10_5_GSM3537044_GSE124577.tsv" … … (special genome)	27998	28693	"mm10_1_GSM2671415_GSE100106.tsv" Deleted. (Not clear version.)
Human	"grch37_1_GSM3073089_GSE112570.tsv" "grch38_1_GSE117403.tsv" "grch38_2_GSM3375767_GSE119506.tsv" "grch38_3_GSM3478791_GSE122703.tsv" "grch38_4_GSM3543618_GSE124703.tsv" "grch38_5_GSM3813936_GSE131685.tsv" "hg19_2_GSM3430548_GSE121267.tsv" "hg19_3_GSM3635372_GSE127471.tsv" "hg19_4_GSM2897333_GSE108394.tsv" … … (special genome)	30710	60570	"hg19_1_GSM2867931_GSE106245.tsv" "hg19_5_GSM3143601_GSE114530.tsv" Deleted. (Decimal point, date format error, version error.)

Tab. 23 Components and the number of gene probes of common list (reference genome assembly) and full list for Mouse and Human.



1	**ALL PROBES HUMAN**						
2							
3	**PROBES**	**hg19**	**GRCh37**	**GRCh38**	**Ensembl_GRCh38.p12_rel94**	**GSM3717979**	
4	ENSG00000117533	hg19_VAMP4	grch37_VAMP4	grch38_VAMP4	#VAMP4	#VAMP4	in all
5	ENSG00000228915				#OR7E128P		Ensembl_GRCh38.p12_rel94
6	ENSG00000248222	hg19_CTB-174D11.1	grch37_CTB-174D11.1	grch38_CTB-174D11.1	#AC011389.1	#AC011389.1	in all
7	ENSG00000236230	hg19_RP11-400N13.1	grch37_RP11-400N13.1	grch38_RP11-400N13	#AL356108.1	#AL356108.1	in all
8	ENSG00000236596				#AC092568.1		Ensembl_GRCh38.p12_rel94
9	ENSG00000233029	hg19_RP11-439A17.9	grch37_RP11-439A17.9	grch38_RP11-439A17	#AC244453.2	#AC244453.2	in all
10	ENSG00000162636	hg19_FAM102B	grch37_FAM102B	grch38_FAM102B	#FAM102B	#FAM102B	in all
11	ENSG00000261714				#AC105137.1		Ensembl_GRCh38.p12_rel94
60566	ENSG00000101871	hg19_MID1	grch37_MID1	grch38_MID1	#MID1	#MID1	in all
60567	ENSG00000196517	hg19_SLC6A9	grch37_SLC6A9	grch38_SLC6A9	#SLC6A9	#SLC6A9	in all
60568	ENSG00000092439	hg19_TRPM7	grch37_TRPM7	grch38_TRPM7	#TRPM7	#TRPM7	in all
60569	ENSG00000221840	hg19_OR4A5	grch37_OR4A5	grch38_OR4A5	#OR4A5	#OR4A5	in all
60570	ENSG00000284387				#MIR24-2		Ensembl_GRCh38.p12_rel94
60571	ENSG00000085733	hg19_CTTN	grch37_CTTN	grch38_CTTN	#CTTN	#CTTN	in all
60572	ENSG00000168140	hg19_VASN	grch37_VASN	grch38_VASN	#VASN	#VASN	in all
60573	ENSG00000258631	hg19_RP11-739G5.1	grch37_RP11-739G5.1	grch38_RP11-739G5.1	#AC110023.1	#AC110023.1	in all
60574							

Fig. 24 Full list of Human genome, Ensembl probes with the mapping to different genome versions.



1	**ALL PROBES MOUSE**				
2					
3	**PROBES**	**mm10**	**grchm38**	**GSE120410 (modified UCSC mm10 with additional gene "Prop1L" added to**	
4	ENSMUSG00000101435	mm10_Gm28772	grchm38_Gm28772	#Gm28772	in all
5	ENSMUSG00000044244	mm10_Il20rb	grchm38_Il20rb	#Il20rb	in all
6	ENSMUSG00000069094	mm10_Pde7a	grchm38_Pde7a	#Pde7a	in all
7	ENSMUSG00000105704	mm10_Gm43055	grchm38_Gm43055	#Gm43055	in all
8	ENSMUSG00000033871	mm10_Ppargc1b	grchm38_Ppargc1b	#Ppargc1b	in all
9	ENSMUSG00000005447	mm10_Pafah1b3	grchm38_Pafah1b3	#Pafah1b3	in all
10	ENSMUSG00000025163	mm10_Cd7	grchm38_Cd7	#Cd7	in all
28688	ENSMUSG00000096468	mm10_Gm16405	grchm38_Gm16405	#Gm16405	in all
28689	ENSMUSG00000052534	mm10_Pbx1	grchm38_Pbx1	#Pbx1	in all
28690	ENSMUSG00000101886	mm10_Gm28324	grchm38_Gm28324	#Gm28324	in all
28691	ENSMUSG00000039197	mm10_Adk	grchm38_Adk	#Adk	in all
28692	ENSMUSG00000032182	mm10_Yipf2	grchm38_Yipf2	#Yipf2	in all
28693	ENSMUSG00000048693	mm10_Olfr435	grchm38_Olfr435	#Olfr435	in all
28694	ENSMUSG00000005917	mm10_Otx1	grchm38_Otx1	#Otx1	in all
28695	ENSMUSG00000101634	mm10_1700066B17Rik	grchm38_1700066B17Rik	#1700066B17Rik	in all
28696	ENSMUSG00000037025	mm10_Foxa2	grchm38_Foxa2	#Foxa2	in all
28697					

Fig. 25 Full list of Mouse genome, Ensembl probes with the mapping to different genome versions.

All code involved in this article can be found on https://github.com/SingleCellAnalysis/SingleCellAnalysis.

6 Case study

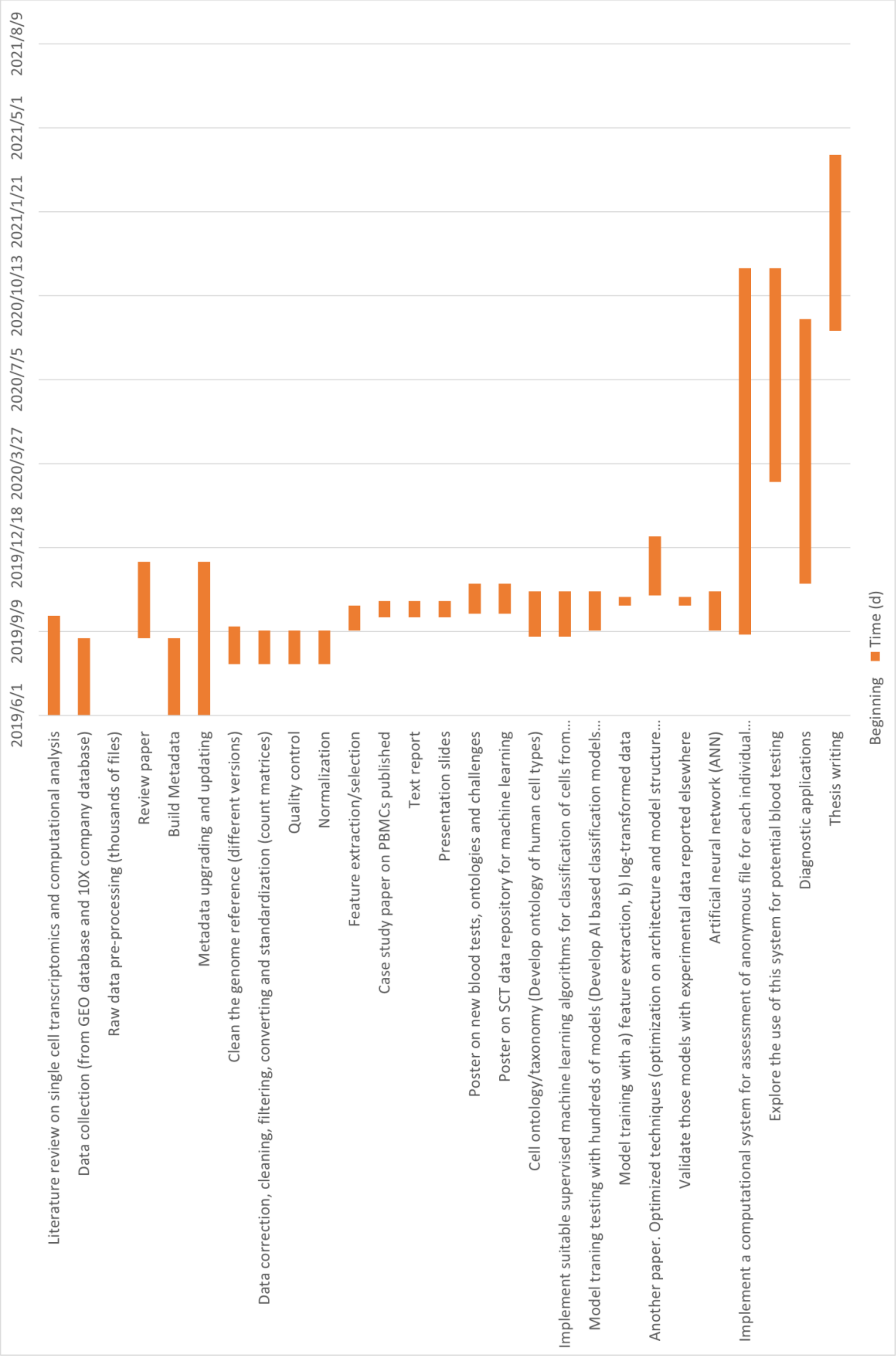
A case study on classification of five cell types from PBMC samples have been submitted to BIBM.

Conclusion

Downstream big data analysis of single cell transcriptomics data can contribute a lot to human life science knowledgement. It can support further precise classification of single cell, detect of rare cell subtypes and conduct diseases diagnosis, detection and prediction. In our project, around 2,000 SCT data files have been collected (from GEO database and other qualified resources), cleaned and standardized, and the description information of each sample study has been mapped into the metadata. Reference genome assembly has been built for machine learning model training. Refined cell ontology serves as annotation labeling in supervised machine learning still needs to be cleared. In the case study of classification of PBMC blood sample, the accuracy of cell recognition and classification is 90%, much more accuracy of classification still needs to be achieved for realistic clinical implementation. Optimization feature extraction methods and training model structure and parameters needs to be found. The overall goal of this project is to develop a system that will enable users to perform automatized analysis of SCT data, which can highly increase the promotion in cell biology, developmental biology and the preclinical and prognostic diagnosis and treatment of diseases.

Project Plan

Task	Responsible	Beginning	Finish date	Time (d)	Status
Literature review on single cell transcriptomics and computational analysis	Jiahui ZHONG	2019/6/1	2019/9/28	119	Complete
Data collection (from GEO database and 10X company database)	Jiahui ZHONG	2019/6/1	2019/9/1	92	Complete
Raw data pre-processing (thousands of files)	Jiahui ZHONG			0	Complete
Review paper	Jiahui ZHONG	2019/9/1	2019/12/1	91	40%
Build Metadata	Jiahui ZHONG	2019/6/1	2019/9/1	92	Complete
Metadata upgrading and updating	Jiahui ZHONG	2019/6/1	2019/12/1	183	98%
Clean the genome reference (different versions)	Jiahui ZHONG	2019/8/1	2019/9/15	45	Complete
Data correction, cleaning, filtering, converting and standardization (count matrices)	Jiahui ZHONG	2019/8/1	2019/9/10	40	50% (only 1
Quality control	Jiahui ZHONG	2019/8/1	2019/9/10	40	Complete
Normalization	Jiahui ZHONG	2019/8/1	2019/9/10	40	Complete
Feature extraction/selection	Jiahui ZHONG	2019/9/10	2019/10/10	30	Complete
Case study paper on PBMCs published	Jiahui ZHONG	2019/9/26	2019/10/15	19	Complete
Text report	Jiahui ZHONG	2019/9/26	2019/10/15	19	30%
Presentation slides	Jiahui ZHONG	2019/9/26	2019/10/15	19	90%
Poster on new blood tests, ontologies and challenges	Jiahui ZHONG	2019/9/30	2019/11/5	36	0%
Poster on SCT data repository for machine learning	Jiahui ZHONG	2019/9/30	2019/11/5	36	0%
Cell ontology/taxonomy (Develop ontology of human cell types)	Jiahui ZHONG	2019/9/3	2019/10/27	54	20%
Implement suitable supervised machine learning algorithms for classification of cells from SCT data	Jiahui ZHONG	2019/9/3	2019/10/27	54	90%
Model traning testing with hundreds of models (Develop AI based classification models for various cell types and their states)	Jiahui ZHONG	2019/9/10	2019/10/27	47	90%
Model training with a) feature extraction, b) log-transformed data	Jiahui ZHONG	2019/10/10	2019/10/20	10	0%
Another paper. Optimized techniques (optimization on architecture and model structure and parameters). New data sets with higher quality. On the base of the published PBMC paper.	Jiahui ZHONG	2019/10/22	2019/12/31	70	5%
Validate those models with experimental data reported elsewhere	Jiahui ZHONG	2019/10/10	2019/10/20	10	0%
Artificial neural network (ANN)	Jiahui ZHONG	2019/9/10	2019/10/27	47	80%
Implement a computational system for assessment of anonymous file for each individual cell	Jiahui ZHONG	2019/9/5	2020/11/15	437	40%
Explore the use of this system for potential blood testing	Jiahui ZHONG	2020/3/5	2020/11/15	255	0%
Diagnostic applications	Jiahui ZHONG	2019/11/5	2020/9/15	315	30%
Thesis writing	Jiahui ZHONG	2020/9/1	2021/3/30	210	2%


2019/6/1
2019/9/9
2019/12/18
2020/3/27
2020/7/5
2020/10/13
2021/1/21
2021/5/1
2021/8/9
Literature review on single cell transcriptomics and computational analysis
Data collection (from GEO database and 10X company database)
Raw data pre-processing (thousands of files)
Review paper
Build Metadata
Metadata upgrading and updating
Clean the genome reference (different versions)
Data correction, cleaning, filtering, converting and standardization (count matrices)
Quality control
Normalization
Feature extraction/selection
Case study paper on PBMCs published
Text report
Presentation slides
Poster on new blood tests, ontologies and challenges
Poster on SCT data repository for machine learning
Cell ontology/taxonomy (Develop ontology of human cell types)
Implement suitable supervised machine learning algorithms for classification of cells from...
Model traning testing with hundreds of models (Develop AI based classification models...
Model training with a) feature extraction, b) log-transformed data
Another paper. Optimized techniques (optimization on architecture and model structure...
Validate those models with experimental data reported elsewhere
Artificial neural network (ANN)
Implement a computational system for assessment of anonymous file for each individual...
Explore the use of this system for potential blood testing
Diagnostic applications
Thesis writing
Beginning
■ Time (d)

Reference

1. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nature Reviews Genetics. 2016 Mar;17(3):175.

2. Todd R, Margolin DH. Challenges of single-cell diagnostics: analysis of gene expression. Trends in molecular medicine. 2002 Jun 1;8(6):254-7.

3. Owens B. Genomics: The single life. Nature News. 2012 Nov 1;491(7422):27.

4. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May 21;161(5):1187-201.

5. Powell AA, Talasaz AH, Zhang H, Coram MA, Reddy A, Deng G, Telli ML, Advani RH, Carlson RW, Mollick JA, Sheth S. Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. PloS one. 2012 May 7;7(5):e33788.

6. Babbe H, Roers A, Waisman A, Lassmann H, Goebels N, Hohlfeld R, Friese M, Schröder R, Deckert M, Schmidt S, Ravid R. Clonal expansions of CD8+ T cells dominate the T cell infiltrate in active multiple sclerosis lesions as shown by micromanipulation and single cell polymerase chain reaction. Journal of Experimental Medicine. 2000 Aug 7;192(3):393-404.

7. Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell metabolism. 2016 Oct 11;24(4):593-607. A

8. Baxter AE, Niessl J, Fromentin R, Richard J, Porichis F, Charlebois R, Massanella M, Brassard N, Alsahafi N, Delgado GG, Routy JP. Single-cell characterization of viral translation-competent reservoirs in HIV-infected individuals. Cell host & microbe. 2016 Sep 14;20(3):368-80.

9. Heldt FS, Kupke SY, Dorl S, Reichl U, Frensing T. Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection. Nature communications. 2015 Nov 20;6:8938. A

10. Zilionis R, Nainys J, Veres A, et al. Single-cell barcoding and sequencing using droplet microfluidics[J]. Nature protocols, 2017, 12(1): 44.

11. Junker J P, van Oudenaarden A. Every cell is special: genome-wide studies add a new dimension to single-cell biology[J]. Cell, 2014, 157(1): 8-11.

12. Luecken M D, Theis F J. Current best practices in single‐cell RNA‐seq analysis: a tutorial[J]. Molecular systems biology, 2019, 15(6).

13. Aran D, Looney A P, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage[J]. Nature immunology, 2019, 20(2): 163.

14. Hsu DF, Chung YS, Kristal BS. Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. In Advanced data mining technologies in bioinformatics 2006 (pp. 32-62).

IGI Global.

15. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications. 1998 Jul;13(4):18-28.

16. Hopfield JJ. Artificial neural networks. IEEE Circuits and Devices Magazine. 1988 Sep;4(5):3-10.

17. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012 Nov 26;41(D1):D991-5.

18. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017 Jan 16;8:14049.

Appendix 1