CS262 Winter 2016

Single Cell Sequencing
Background

**NATURE METHODS | METHAGORA**

Significant increase in publications and data in the last two years.
Background

Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing

Mei-Chong Wendy Leeb,1, Fernando J. Lopez-Diazb,1, Shahid Yar Khanb,2, Muhammad Akram Taric,3, Yelena Daync, Charles Joseph Vasked, Amie J. Radenbaughd, Hyunsung John Kin, Beverly M. Emersonb,4, and Nader Pourmanda,4

Focus

Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation

Aleksandra A. Kolodziejczyka,1,5 Jong Kyoung Kim,1,5 Jason C.H. Tsang,2 Tomislav Ilicic,1,2, Johan Henriksson,1 Kedar N. Natarajan,1,2 Alex C. Tuck,1,5 Xuefei Gao,2 Marc Bühler,3 Pentao Liu,2 John C. Marioni1,2,4, and Sarah A. Teichmann1,2

Cell Reports

Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations

Nicola K. Wilson,1,9 David G. Kent,1,9 Florian Buettner,2,9 Mona Shehata,7 Iain C. Macaulay,6 Manuel Sánchez Castillol, Caroline A. Oedekoven,1 Evangelia Diamanti,1 Reiner Schult,1 Thierry Voet,3,6 Carlos Caldas,7 John Stingl,7 Anthony R. Green,1 Fabian J. Theis,2,6 and

Article

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

Amit Zeisel,1† Ana B. Muñoz-Manchado,1† Simone Codeluppi,1 Peter Lönnerberg,1 Gioele La Manno,1 Anna Juréus,1 Sueli Marques,1 Hermanny Munguba,1 Lique He,2 Christer Betsholtz,2,3 Charlotte Rolny,4 Gonçalo Castelo-Branco,1 Jens Hjerling-Leffler,1† Sten Linnarsson1†
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology
How do animals grow and develop from a single cell?
Developmental Biology

- **G2**: The cell "double checks" the duplicated chromosomes for error, making any needed repairs.
- **Mitosis**: Each of the 46 chromosomes is duplicated by the cell.
- **G1**: Cellular contents, excluding the chromosomes, are duplicated.
- **G0**: Cell cycle arrest.
- **Self-renewal**: Pluripotent Stem Cells
- **Differentiation**: Tissue Stem/Progenitor Cells

Various cell types
We need single-cell resolution to:

• Discover more complicated mechanisms in cellular development

• Confirm the distinct gene expression signatures across different cell types

• Identify functional differences among the same cell cell type
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology
Cancer Biology

Tumors are composed of genetically and phenotypically heterogeneous clones.

Deep (bulk) sequencing can only capture 1% of the cell population (excluding some types such as circulating tumor cells).
We need single-cell resolution to:

- Find evidence for models of cancer
- Infer timing of mutations and the drivers
- Evaluate effectiveness of targeted therapy
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology
Microbiology
The Human Microbiome Project says the human body has 100 trillion microscopic life forms living in it.

You call this living?
We need single-cell resolution to:

- Discover low-abundance species that are difficult to culture in vitro
- Monitor transcriptional gene activation mechanisms for functional annotation
Applications

- Developmental Biology
- Cancer Biology
- Microbiology
- Neurology
Neurology
We need single-cell resolution to:

- Study the mosaic genomes of individual neurons and compositions in the brain.
- Follow genetic variations during fetal development.
- Develop targeted therapy for neurological diseases for specific cell types.
One Cell at a Time

Bo Wang, Jason Zhu

Now THAT'S IRONIC!
**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**

- 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**

- 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.
Single-Cell Technologies

(i) isolate single cells
(ii) amplify genome efficiently
(iii) sequence DNA

There are many options available to isolate a single cell from a heterogeneous population of cells. In addition to well-established methods, such as fluorescence activated cell sorting (FACS), microscopy, and the use of antibody capture for cell separation, there is an increasingly ingenious armamentarium of modern methods to isolate single cells with ever greater accuracy and specificity. The most critical part of single-cell analysis is sample preparation, where specific cells need to be accurately isolated from a heterogeneous population. This section highlights commonly used and novel techniques for isolation of single cells from suspension or tissues.

Single-Cell Technologies

Flow cytometry e.g. FACS

Microdissection e.g. laser capture

Micromanipulation e.g. mechanical aspirator

Microfluidics e.g. chamber or droplet

Single-cell separation/isolation

Cell-lysat analysis

Live-cell analysis

qRT-PCR/RNA-seq: transcriptomics
Hi-C: 3D chromosome conformation
RSMA/bisulfite seq: DNA methylation
locus-specific nucleosome mapping

Fluorescent reporter for gene expression
CiA: induction of chromatin modifications
Fabs/Hisac: global distribution of chromatin modifications
m6A-Tracer: tracking of protein-DNA interactions
FLIM-FRET: chromatin compaction

ChIP-qPCR/seq: histone modifications
DamID: chromatin-associated factors
 Genome-wide nucleosome mapping

TRENDS in Cell Biology
Cell Sorting

Manual

dissociated cells

fluorescence microscope

~100 cells

Cell Sorting

FACS: fluorescence activated cell sorting

1. cell suspension
2. staining of membrane, cytoplasm, and nucleus
3. cell sorting
4. cell culture, analysis, etc.

FACS machine
Laser
nozzle
flow cytometer
sort electronics

- Laser
- dissociated cells
- >10 000 cells

http://www.flowlab-childrens-harvard.com/yahoo_site_admin/assets/images/principle123.285181420_std.gif
Cell Sorting

**LCM: laser capture microdissection**

- **LCM**
  - cryosectioned tissue
  - UV laser
  - IR laser
  - >1000 cells

- **LCM**: laser capture microdissection

- **Table 2**

  - LCM: laser capture microdissection

  - **Figure 3**

  - http://www.genomemedicine.com/content/figures/gm247-2-l.jpg

http://www.genomemedicine.com/content/figures/gm247-2-l.jpg
Cell Sorting

Microfluidics: can isolate rare circulating cells

Cell Sorting

High-throughput (~100,000 cells)

Drop-seq

inDrop

http://www.cell.com/abstract/S0092-8674(15)00549-8

http://www.cell.com/cell/abstract/S0092-8674(15)00500-0
(i) isolate single cells

(ii) amplify genome efficiently

(iii) sequence DNA
Amplification and Sequencing

Review: Next Generation Sequencing (NGS)
Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

library preparation

- Bisulfite DNA
- dsDNA Conversion
- End Repair
- dA Tailing
- Adaptor Ligation
- Size Selection
- Amplification/Purification

NGS (Illumina)

https://www.illumina.com

©2011, Illumina Inc. All rights reserved.
Amplification and Sequencing

Review: Next Generation Sequencing (NGS)

https://www.illumina.com
Review: Next Generation Sequencing (NGS)
Review: RNA-Sequencing

http://www.yourgenome.org/sites/default/files/images/illustrations/central_dogma-03.png
Single-cell Amplification

Review: RNA-sequencing

library preparation

http://www.yourgenome.org/sites/default/files/images/illustrations/central_dogma-03.png
Amplification and Sequencing

Review: RNA-Sequencing

Samples of interest
- Condition 1 (normal colon)
- Condition 2 (colon tumor)

Isolate RNAs
- Generate cDNA, fragment, size select, add linkers

Sequence ends

Map to genome, transcriptome, and predicted exon junctions

Downstream analysis

100s of millions of paired reads
10s of billions bases of sequence

Amplification and Sequencing

Review: RNA-Sequencing

![RNA-Sequencing Flowchart]

1. Total RNA
2. Oligo dT enrichment
3. mRNA
4. Fragmentation
5. Random hexamer primed cDNA synthesis
6. HiSeq™ 2000 sequencing
7. Mapping to gene
8. Reference gene
9. Gene function analysis
Amplification and Sequencing

Review: RNA-Sequencing

![Diagram of RNA-sequencing process]

**Total RNA**
- Oligo dT enrichment
- Fragmentation
- Random hexamer primed cDNA synthesis
- HiSeq™ 2000 sequencing
- Mapping to gene

**Gene function analysis**

![Heatmap of expression change]

- **Expression change**
  - Gained weight
  - Weight loss < 5%
  - Weight loss 5–10%
  - Weight loss > 10%

![Box plots for gene expression]

- **Candidate genes**
  - NR4A3
  - RIMS2
  - AQP3
Single-cell Amplification

**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**
  - 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**
  - 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.
Single-cell Amplification

Digital Expression Matrix: counting unique molecules

http://www.nature.com/nmeth/journal/v9/n3/images_article/nmeth.f.355-F1.jpg
DNA sequencing:
- new amplification methods other than PCR
- statistical methods for SNPs/CNV calling

RNA sequencing:
- standards created for quality control
- can achieve high sequencing depth
- high cell throughput methods arising
Downstream Analysis

Supervised Analysis

Cell Population Identification

Unsupervised Analysis

a Obtain an unbiased sample of single cells

b Generate single-cell expression profiles

c Identify cell types by clustering

Figure 3 | Cell-type discovery by unbiased sampling and transcriptome profiling of single cells.

Downstream Analysis

How do cell types differ from each other?

Is there any addition diversity in the same cell type?

http://www.people.vcu.edu/~mreimers/OGMDA/gene_expression_matrix.gif
Dimension Reduction

Principle Component Analysis (PCA)

e.g., visualizing the samples in a smaller subspace

http://www.nlpca.org/fig_pca_principal_component_analysis.png
Variance / Standard Deviation: measure of the spread of the data
(Calculation: average distance from the mean of the data)

Covariance: measure of how much each of the dimensions vary from the mean with respect to each other; measured between 2 dimensions to see if there is a relationship between the 2 dimensions

* The covariance between one dimension and itself is the variance.
E.g. for 3 dimensions, consider random vector \((x,y,z)\):

\[
C = \begin{bmatrix}
\text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\
\text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\
\text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z)
\end{bmatrix}
\]

Diagonal is the variances of \(x\), \(y\) and \(z\)

\(\text{cov}(x,y) = \text{cov}(y,x)\) hence matrix is symmetrical about the diagonal

N-dimensional data will result in \(n \times n\) covariance matrix
The eigenvalue problem is any problem having the following form:

\[ A \cdot v = \lambda \cdot v \]

- \( A \): \( n \times n \) matrix
- \( v \): \( n \times 1 \) non-zero vector
- \( \lambda \): scalar

Any value of \( \lambda \) for which this equation has a solution is called the eigenvalue of \( A \) and vector \( v \) which corresponds to this value is called the eigenvector of \( A \).
Dimension Reduction

Principle Component Analysis (PCA)

http://www.nlpca.org/fig_pca_principal_component_analysis.png
Principal component analysis (PCA) converts a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components.

The first principal component is the projection of the data into a single dimension that has as high a variance as possible (that is, accounts for as much of the variability in the data as possible); each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components.

Therefore the PCs provide a view on the structure of the data that best explains its variance.

cf. Wikipedia: Principal component analysis
The example data is two-dimensional, but most of the information is contained along a dimension shown here by the red vector. We could thus restrict our analysis to a projection along that vector.
PCA process – STEP 1

• Subtract the mean
  from each of the data dimensions. All the x values have x subtracted and y values have y subtracted from them. This produces a data set whose mean is zero.

Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The variance and co-variance values are not affected by the mean value.
# PCA

## PCA process –STEP 1

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>1.9</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.6</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.69</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>-1.31</td>
<td>-1.21</td>
</tr>
<tr>
<td></td>
<td>.39</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.09</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>1.29</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>.49</td>
<td>.79</td>
</tr>
<tr>
<td></td>
<td>.19</td>
<td>-.31</td>
</tr>
<tr>
<td></td>
<td>-.81</td>
<td>-.81</td>
</tr>
<tr>
<td></td>
<td>-.31</td>
<td>-.31</td>
</tr>
<tr>
<td></td>
<td>-.71</td>
<td>-.101</td>
</tr>
</tbody>
</table>
PCA process –STEP 2

• Calculate the covariance matrix
  
  \[
  \text{cov} = \begin{pmatrix}
  .616555556 & .615444444 \\
  .615444444 & .716555556
  \end{pmatrix}
  \]

• since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together.
PCA process –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

\[
\text{eigenvalues} = \begin{bmatrix}
0.0490833989 \\
1.28402771
\end{bmatrix}
\]

\[
\text{eigenvectors} = \begin{bmatrix}
-0.735178656 & -0.677873399 \\
0.677873399 & -0.735178656
\end{bmatrix}
\]
PCA process –STEP 4

• Reduce dimensionality and form *feature vector* the eigenvector with the *highest* eigenvalue is the *principle component* of the data set.

• In our example, the eigenvector with the larges eigenvalue was the one that pointed down the middle of the data.

• Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance.
PCA process –STEP 4

• Now, if you like, you can decide to *ignore* the components of lesser significance

• You do lose some information, but if the eigenvalues are small, you don’t lose much

  • n dimensions in your data
  • calculate n eigenvectors and eigenvalues
  • choose only the first p eigenvectors
  • final data set has only p dimensions.
Dimension Reduction

Principle Component Analysis (PCA)

• linear multivariate statistical analysis
• understand underlying data structures
• identify bias, experimental errors, batch effects
• visualize the samples in a smaller subspace (dimension reduction)
• visualize the relationship between variables (correlation analysis)
t-SNE

2D projection of the swissroll

Unrolled manifold

t-SNE

original data space

PCA

component space
t-SNE

Key quantities

$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Kullback–Leibler divergence (to be minimized)

high-dimensional joint distribution

low-dimensional joint distribution
Cluster Analysis

Cluster: a collection of data objects
  Similar to the objects in the same cluster (Intraclass similarity)
  Dissimilar to the objects in other clusters (Interclass dissimilarity)
Cluster analysis
  Statistical method for grouping a set of data objects into clusters
  A good clustering method produces high quality clusters with high
    intraclass similarity and low interclass similarity
Clustering is an unsupervised classification method
Can be a stand-alone tool or as a preprocessing step for other algorithms
Cluster Analysis

Group objects according to their similarity

**Cluster:**
a set of objects that are similar to each other and separated from the other objects.

Example: green/red data points were generated from two different normal distributions.
Hierarchical Clustering

• This produces a binary tree or **dendrogram**

• The final cluster is the root and each data item is a leaf

• The height of the bars indicate how close the items are
Hierarchical Clustering

Start with every data point in a separate cluster
Keep merging the most similar pairs of data points/clusters until we have one big cluster left

*This is called a bottom-up or agglomerative method*
Hierarchical Clustering

Levels of Clustering

a) Six Clusters
b) Four Clusters
c) Three Clusters
d) Two Clusters
e) One Cluster
We already know about distance measures between data items, but what about between a data item and a cluster or between two clusters? We just treat a data point as a cluster with a single item, so our only problem is to define a linkage method between clusters. As usual, there are lots of choices...
Hierarchical Clustering

Average Linkage

• Definition
  • Each cluster $c_i$ is associated with a mean vector $\mu_i$ which is the mean of all the data items in the cluster
  • The distance between two clusters $c_i$ and $c_j$ is then just $d(\mu_i, \mu_j)$

• This is somewhat non-standard – this method is usually referred to as centroid linkage and average linkage is defined as the average of all pairwise distances between points in the two clusters
Hierarchical Clustering

Single Linkage

• The minimum of all pairwise distances between points in the two clusters
• Tends to produce long, “loose” clusters
Hierarchical Clustering

Complete Linkage

- The maximum of all pairwise distances between points in the two clusters
- Tends to produce very tight clusters
Hierarchical Clustering

Distances between clusters (summary)

- Calculation of the distance between two clusters is based on the pairwise distances between members of the clusters.
  - **Complete linkage**: largest distance between points
  - **Average linkage**: average distance between pairs of points
  - **Single linkage**: smallest distance between points
  - **Centroid**: distance between centroids

Complete linkage gives preference to compact/spherical clusters. Single linkage can produce long stretched clusters.
Hierarchical Clustering

• Major advantage
  • Conceptually very simple
  • Easy to implement → most commonly used technique

• Major weakness of agglomerative clustering methods
  • do not scale well: time complexity of at least $O(n^2)$, where $n$ is the number of total objects
  • can never undo what was done previously → high likelihood of getting stuck in local minima
Other Challenges
Batch Effects Occur

the batch effect represents the **systematic technical differences** when samples are processed and measured in different batches and which are **unrelated to any biological variation** recorded.
Batch Effects Occur

As shown in section 4.2.3, gPCA is not sensitive to filtering, so filtering can be used to reduce the data dimension and facilitate implementing gPCA by reducing the analysis time without worry.

http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=4179&context=etd


Biological Effects

• Cancer: cell lineage
• Metagenomics: cis/trans mechanisms
• Stem Cells: cellular phenotypes
• Immunology: cell type identification
• Neurology: somatic mutations
Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells

Florian Buettner¹,²,⁵, Kedar N Natarajan²,³,⁵, F Paolo Casale², Valentina Proserpio²,³, Antonio Scialdone²,³, Fabian J Theis¹,⁴, Sarah A Teichmann²,³, John C Marioni²,³ & Oliver Stegle²

---

Example

Observed expression profile

Removing the effect of cell cycle

Expression range of differentiation gene

Expression range of differentiation gene

Genes annotated to cell-cycle

Genes not annotated to cell-cycle
Full example: DropSeq

1. Cells from suspension
2. Microparticle and lysis buffer
3. Oil
4. Cell lysis (in seconds)
5. RNA hybridization
6. Break droplets
7. Reverse transcription with template switching
8. PCR (STAMPs as template)
9. Sequencing and analysis

- Each mRNA is mapped to its cell-of-origin and gene-of-origin
- Each cell's pool of mRNA can be analyzed

Cell barcode | UMI | cDNA (50-bp sequenced)
---|---|---

<table>
<thead>
<tr>
<th>Cell</th>
<th>1</th>
<th>2</th>
<th>\ldots</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE 1</td>
<td>1</td>
<td>2</td>
<td>\ldots</td>
<td>14</td>
</tr>
<tr>
<td>GENE 2</td>
<td>4</td>
<td>27</td>
<td>\ldots</td>
<td>8</td>
</tr>
<tr>
<td>GENE 3</td>
<td>0</td>
<td>0</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>GENE M</td>
<td>6</td>
<td>2</td>
<td>\ldots</td>
<td>0</td>
</tr>
</tbody>
</table>

Hundreds of millions of reads

Thousands of cells