Cell atlases: types, subtypes, states... and outliers?

Fabio Zanini
UNSW Sydney

http://fabilab.org
I acknowledge the traditional custodians of the land I am streaming from, the Gadigal and Bedigal people of the Eora nation, and their connection to land, waters, and culture.

I also pay my respects to Elders past and present.
My lab: single cell data → new hypotheses

New methods: CS, math, eng

Fascinating topics: bio, medicine

http://fabilab.org
Cell atlases aim to answer: **WHO does it and HOW?**
The human genome lesson

2001: "how"? First human genome sequenced.

The biophysical source of cellular behaviour is not anymore an obscure, almost mythical entity: it's a finite and well defined catalogue of genes.

(Only problem: it’s very long and full of junk?)

Nonetheless, there's a list of all 20,000 genes.

That roughly settles the “how”.

2021: what about the “who”?

That’s what single cell sequencing is about.

Now, how does it actually work?
Single cell RNA-Seq (scRNA-Seq)

Gene 1 Gene 2

9 2

+19,998 more genes (approx)
Single cell RNA-Seq (scRNA-Seq)

Gene 1 Gene 2

9 2

+19,998 more genes (approx)

x 10,000 cells (approx)

Cells

<table>
<thead>
<tr>
<th>Genes</th>
<th>Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 2 0</td>
<td>0 0 5</td>
</tr>
<tr>
<td></td>
<td>8 1 0</td>
</tr>
</tbody>
</table>
**Technical note: Embeddings and clustering**

**Embedding** algorithm locates cells with similar profiles close to one another in this abstract plane.

1 cell

Cells also colored by "cluster", what the computer thinks is a cell type based on a group of similar molecular profiles.

**NB**: usually transcriptomes, but could be proteomes or else, *does not really matter*:

i.e. the RNA in scRNA-Seq is a "dummy molecule" that is only used as a proxy for "molecular profile". You can also mix molecule types (e.g. RNA + proteins + epigenetic marks), same soup.
Cell atlases via scRNA-Seq

Like physics has done long ago, we are building the **standard model of organismal biology**:  

1. **What cell types** make up complex organisms?  
2. **What interactions** connect them?
The human genome lesson

2001: "how"? First human genome sequenced.

The biophysical source of cellular behaviour is not anymore an obscure, almost mythical entity: it’s a finite and well defined catalogue of genes.

(Only problem: it’s very long and full of junk?)

Nonetheless, there’s a list of all 20,000 genes.

That roughly settles the “how”.

2021: what about the “who”?

That’s what single cell sequencing is about.

So now we’ll have a list of all ~2,000 cell types?

Well...
Technical note: Embeddings and clustering

1 cell
Technical note: Embeddings and clustering
Technical note: Embeddings and clustering
Summary (spoiler alert!)

Outsourcing the easy part → machine learning on graphs:
1. igraph: Foundational library for graph analysis, please reach out if interested!
2. northstar: Classify single cell data guided by an atlas, but finds new cell types too

The lung neonatal cell atlas → a real world example:
1. New cell types: Macrophage, myofibroblast precursors
2. The blurry line of cell types and states: monocytes, arterial cells, capillaries

Outlier cells → a few examples from our work:
1. Male hyperoxia fibroblasts: a male-specific pathogenic cell type?
2. Venezuelan equine encephalitis virus and "superproducer cells"
3. Antibodies against complex viruses: need a "hero cell"
1. Machine learning on graphs
igraph: fast library for graph algorithms

R interface
Mathematica interface
Python interface
C core
Design principles

Help welcome!

Passionate about graphs? Reach out and contribute!

igraph.org

Jupyter notebook/script
scanpy
Leiden
Biological cell types
northstar: clustering or classifying??

Problems with clustering:
1. **Identity** of each cluster
2. **Resolution**
northstar: clustering or classifying??

Problems with clustering:
1. **Identity** of each cluster
2. Resolution

The problem with supervised classification:
1. Missing cell types (e.g. cancer, developmental progenitors)
northstar: clustering or classifying??

1. Create compressed atlas
2. Merge new cell with comp. atlas
3a. Select features from atlas markers
3b. Select features from new data
4. Build similarity graph
5. Cluster around compressed atlas

Training data is embedded in new dataset

import northstar
sa = northstar.Averages(
    atlas='Darmanis_2015',
)
sa.fit(new_data)
types = sa.membership

Zanini et al. Scientific Reports (2020)
northstar: Glioblastoma (brain cancer)

What are these cells?

Infiltrating immune cells

Neoplastic (cancer) cells

2. A real-world atlas
Cell atlas of the neonatal lung

Q: How does birth – exposure to oxygen and pathogens – affect lung development?

**Results summary**

- **15+ types** of immune cells
- Around 1/3 of them were basically undescribed before, especially macrophage progenitors
- Heterogeneous **monocytes**
- **T cells** of unknown function

Domingo-Gonzales R*, Zanini F* et al eLife (2020)
Blurry cell types: alveolar macrophages

Three lung-specific macrophages:
- I: novel embryonic population (E18.5)
- II: novel transitional population (P1-7)
- III: adult alveolar macrophages

Domingo-Gonzales R*, Zanini F* et al eLife (2020)
The early phenotype is novel and specific to perinatal development.
Endothelial cells (blood vessels)

... I'll spare you this one, but we have written an entire paper just on these cells

Zanini F et al unpublished (2021)
3. Outliers
Male hyperoxia fibroblast: killer or victim?

These cells are only found in neonatal mice, only in males, only if they are "born prematurely" (i.e. they are exposed to high-oxygen atmosphere for a week).

How do you deal with this kind of "cell type"?

Zanini F et al. unpublished (2021 n.2)
Venezuelan equine encephalitis virus: "superproducer cells"

How can these cells produce so much virus so quickly?

Yao Z*, Zanini F* et al. PLOS NTD (2021)
Neutralizing antibodies: the "hero cell"

This cell contained the most potent broadly neutralising antibody against dengue virus in the world, and it defined a whole new class of antibodies against dengue virus.

We found it by a combination of extreme luck and extreme perseverance (as always).

Summary

Outsourcing the easy part → machine learning on graphs:
1. igraph: Foundational library for graph analysis, please reach out if interested!
2. northstar: Classify single cell data guided by an atlas, but finds new cell types too

The lung neonatal cell atlas → a real world example:
1. New cell types: Macrophage, myofibroblast precursors
2. The blurry line of cell types and states: monocytes, arterial cells, capillaries

Outlier cells → a few examples from our work:
1. Male hyperoxia fibroblasts: a male-specific pathogenic cell type?
2. Venezuelan equine encephalitis virus and "superproducer cells"
3. Antibodies against complex viruses: need a "hero cell"
Setty et al. 2019 (Dana Pe’er and friends – classic example of clever single cell marketing)

Seven transcription factors (heptad) tightly regulate hematopoiesis via highly interconnected network of enhancers.

The same enhancers appear heterogeneous in acute myeloid leukemic (AML) samples.

Corces et al. 2016

Master regulators of leukemic cell state

How are these 7 transcription factors expressed in a leukemic cell line (ME-1)?

They are heterogeneous correlated

They are

How are key hematopoietic transcription factors expressed in a leukemic cell line?

1. They are expressed
2. They are heterogeneous
3. They are correlated
4. The core triad GATA2, TAL1, ERG aligns with the stem cell-red blood cell transition (pink → light green)

**Master regulators of leukemic cell state**
0. The single cell graph (or network)

Each dot (vertex) is a cell
Each line (edge) is connecting similar cells

Measurements: gene expression, chromatin state, etc.

Embedding: the scatter plots you see everywhere
Clustering: definition of biological cell types
Pseudotime: gradual changes in cell phenotypes

Equivalent to your social network of friends online
Macrophage proliferation reduced at birth
The rise of lymphocytes

None at birth
Most abundant population by P21

Function might be quite different from our expectations

B cells

Naive, quiescent

Most T cells CD4- CD8-
for both mRNA and protein

Domingo-Gonzales R*, Zanini F* et al eLife (2020)
northstar is used to find a **common embedding** between:

- Tabula Muris (10X droplet data) which is adult only
- Our data set (SmartSeq2 plate data) which includes **novel developmental cell types**
0. The single cell graph (or network)

Measurements:
gene expression, chromatin state, etc.

Cell graph or network

Each dot (vertex) is a cell

Each line (edge) is connecting similar cells

Embedding: the scatter plots you see everywhere

Clustering: definition of biological cell types

Pseudotime: gradual changes in cell phenotypes
0. The single cell graph (or network)

- **Measurements**: gene expression, chromatin state, etc.
- **Cell graph or network**
  - Each dot (vertex) is a cell
  - Each line (edge) is connecting similar cells

- **Embedding**: the scatter plots you see everywhere
- **Clustering**: definition of biological cell types
- **Pseudotime**: gradual changes in cell phenotypes