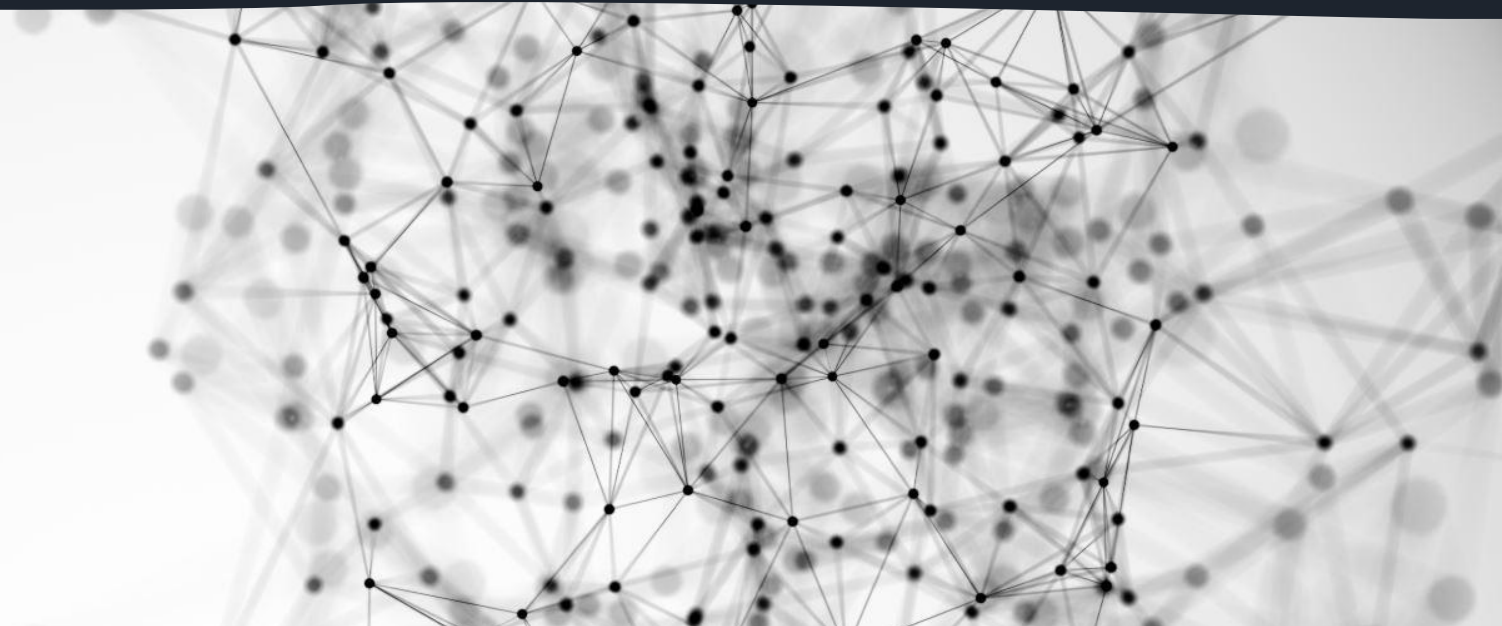


Single Cell RNA Sequencing on the Cloud

Arash Yunesi

Department of Statistics and Probability

Computational Mathematics, Science, and Engineering



A little about myself:

- ✓ **Undergraduate in Physics and Mathematics from Sharif University of Tech. (Iran)**
- ✓ **PhD in Theoretical High Energy Physics from Florida State University**
- ✓ **My previous research was on Dark Matter, Baryogenesis, General Relativity, and ...**
- ✓ **Currently dual PhD at CMSE and department of Statistics and Probability**
- ✓ **Working with Professor Yuying Xie**
- ✓ **Research on Single Cell RNA Sequencing**
- ✓ **Developing New Algorithms in Numerical Linear Algebra**

The general aim of my research:

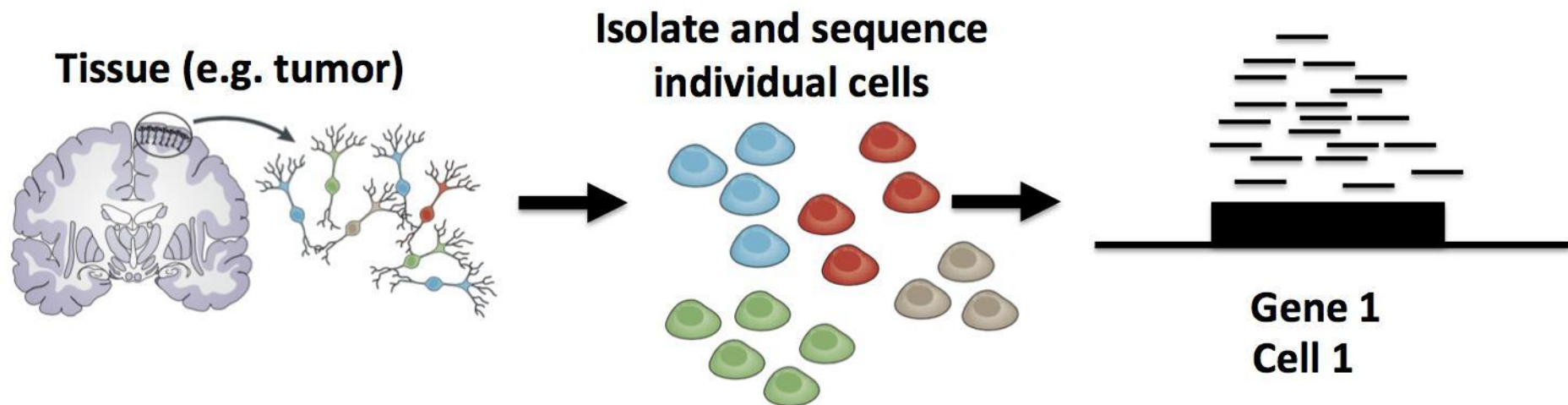
- Many disease have a strong genetic component:

Cancer, Alzheimer's disease, Autoimmune diseases, etc.

- Cancer is the next frontier of medicine.
- Immune system shuts down inside and around tumors!
- Two breast cancers are not the same!
- Many genes play an important role in development of cancer.
- Many genes are involved in the immune system failure.
- We need to find the genes that are responsible.
- Produce a drug specific to each patient.

We have a plan!

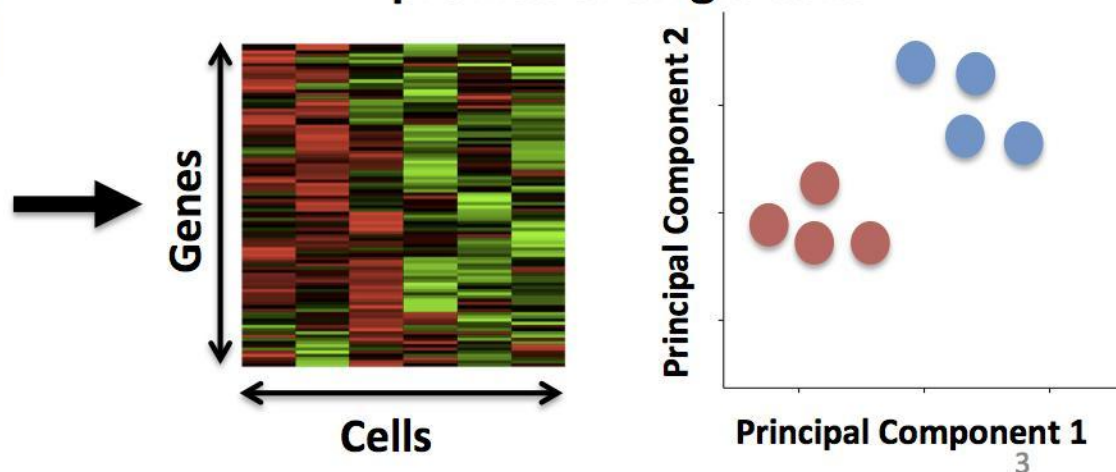
Single-cell RNA-Seq (scRNA-Seq)



Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Compare gene expression profiles of single cells



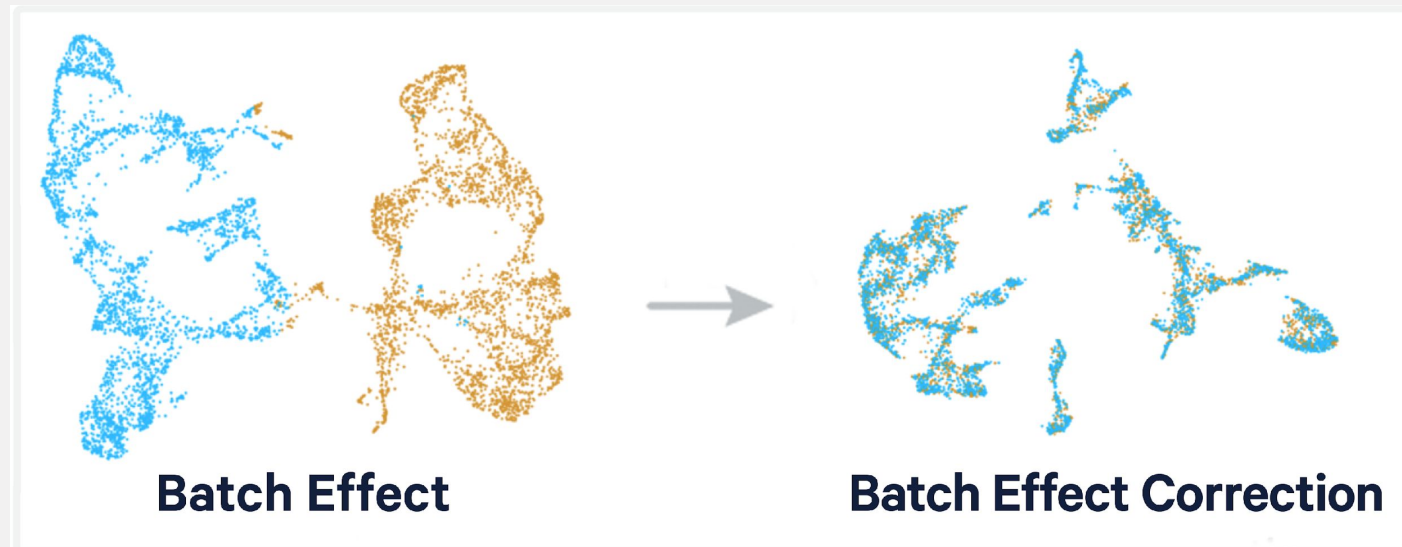
1. Thousands of cells per sample.
2. ~30k genes per cell.
3. Complicated gene mutations.
4. Complicated gene relationships.
5. Statistical error.
6. Cancer stem cells.

What I need from the cloud?

- ❑ Many studies require dozens to thousands of samples (millions of cells).
- ❑ Large scales of data sets is a challenge.
- ❑ Conventional HPC centers are limited by structure and user traffic.
- ❑ In future we will be working with patient data.
- ❑ Most conventional HPC centers are not HIPAA compliant.
- ✓ Scalability of the cloud structure (VM Scale Set)
- ✓ In near future most cloud services might offer HIPAA compliant resources.
- ✗ No cloud service with HIPAA compliance currently.

My project on the cloud:

- ❑ Small statistical error in each gene measurement causes big misalignment in clustering.
- ❑ I am developing a novel technique for removing batch effect.
- ❑ For this project I have used Data Science VM.
- ❑ Required software are Matlab, R, Rstudio, and Python.
- ❑ Monthly budget of \$250-\$400.
- ❑ Public data used, requires no HIPAA compliance.



HIPAA compliance:

- ✓ **For near future we will analyze patient data.**
- ✓ **No HIPAA compliant cloud service.**
- ✓ **All major cloud services are pushing for FDA approval.**
- ✓ **Big pharma or hospitals might host HIPAA compliant cloud infrastructure.**
- ✓ **Patient data can be analyzed with our pretrained models on the cloud.**

My future use of the cloud:

- ✓ Continue building statistical models
- ✓ Improve the numerical algorithms
- ✓ Expand to deep learning (GPU extensive)
- ✓ Azure has several families of GPU VMs
- ✓ Nvidia and AMD based GPUs
- ✓ NDm A100 v4 series: single VM and eight NVIDIA Ampere A100 80GB Tensor Core GPUs ~\$6000
- ✓ Requires funding from collaborators
- ✓ I would like to explore AWS

Ash Or Durbatuluk

What I learned working with cloud services:

- 1. Creating new resources**
- 2. Managing storage**
- 3. Different methods of interaction with the resources**
- 4. GPU-enabled resources**
- 5. Structure of the cloud**
- 6. Troubleshooting and seeking help (just ask ChatGPT!)**
- 7. Last but not least, budgeting 101**

Thank you

- ✓ Pat and Mahmoud for organizing and teaching.
- ✓ Professor O'Shea for giving me this chance.
- ✓ Fellows of the cloud for their companionship.
- ✓ Everyone else who made this great experience possible.
- ✓ My advisor, Yuying Xie for his continued help.

Questions?

Comments?