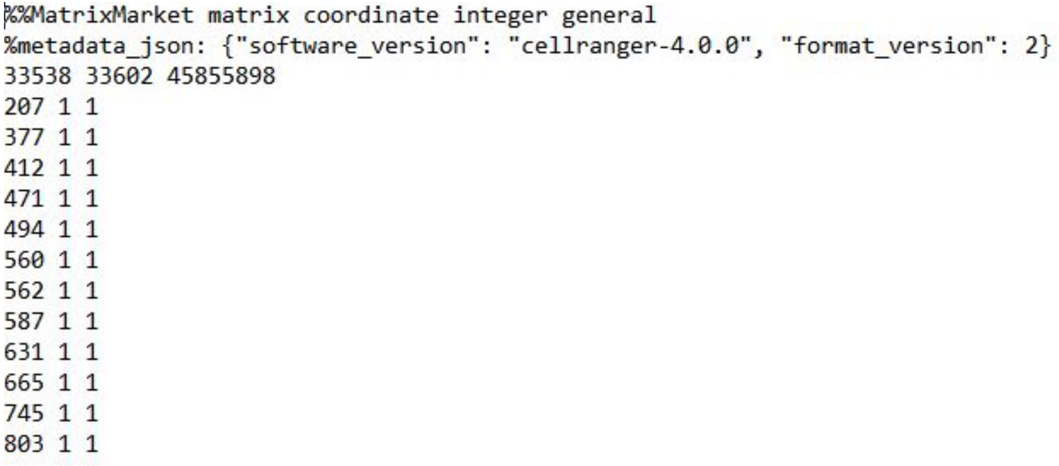
scRNA Compression

Spencer Jenkins and Niko Zhang

Gene Expression Data

Can we compress the information stored in MTX files at a high level?

- No bit vectors, Huffman encoding, etc.

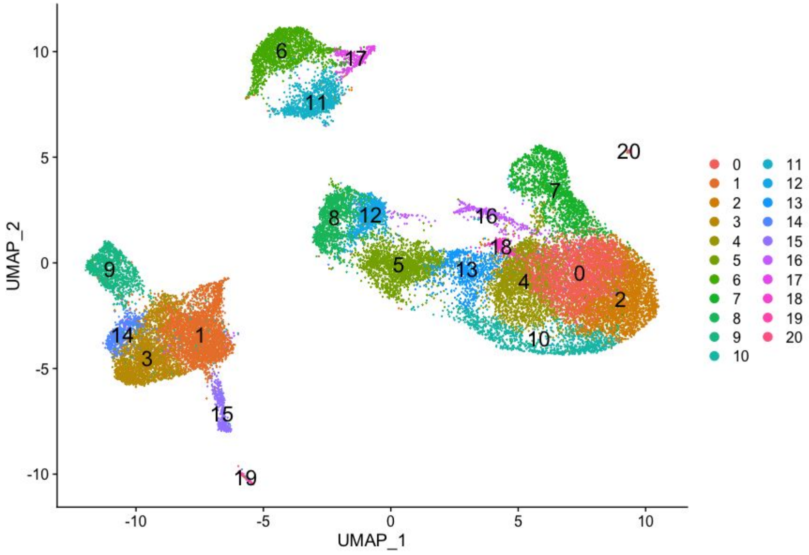


```
%%MatrixMarket matrix coordinate integer general
%metadata_json: {"software_version": "cellranger-4.0.0", "format_version": 2}
33538 33602 45855898
207 1 1
377 1 1
412 1 1
471 1 1
494 1 1
560 1 1
562 1 1
587 1 1
631 1 1
665 1 1
745 1 1
803 1 1
```

Our Idea - delta encoding via clusters

Cluster cells based on their gene profile

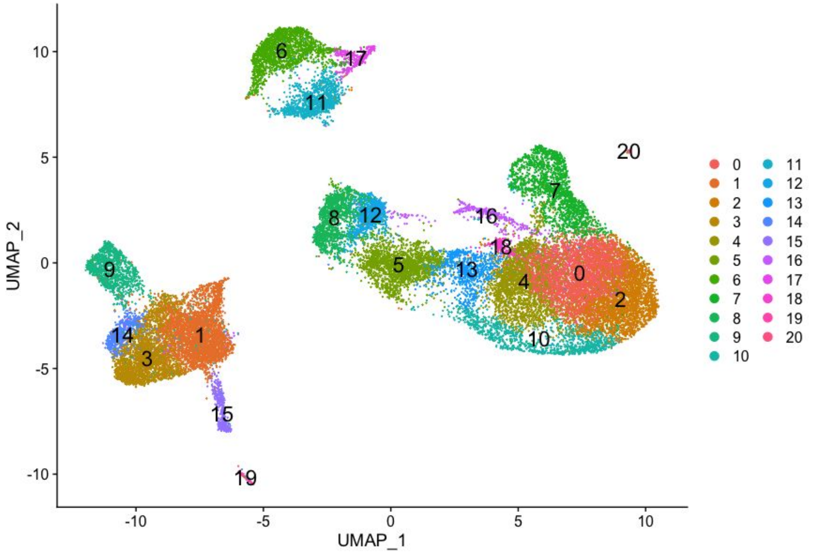
- Each cluster will have a set of common genes



Our Idea - delta encoding via clusters

Store the common genes for each cluster in cluster_genes.txt

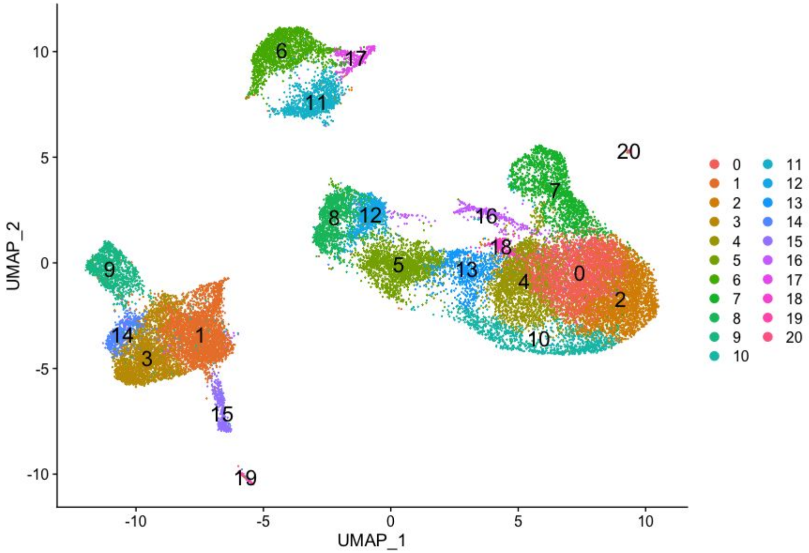
Store the genes not present in the corresponding cluster for each cell in deltas.txt



Our Idea - delta encoding via clusters

For now, ignore count information

Amount of compression is determined by the **number of clusters** and the **clustering algorithm** used.



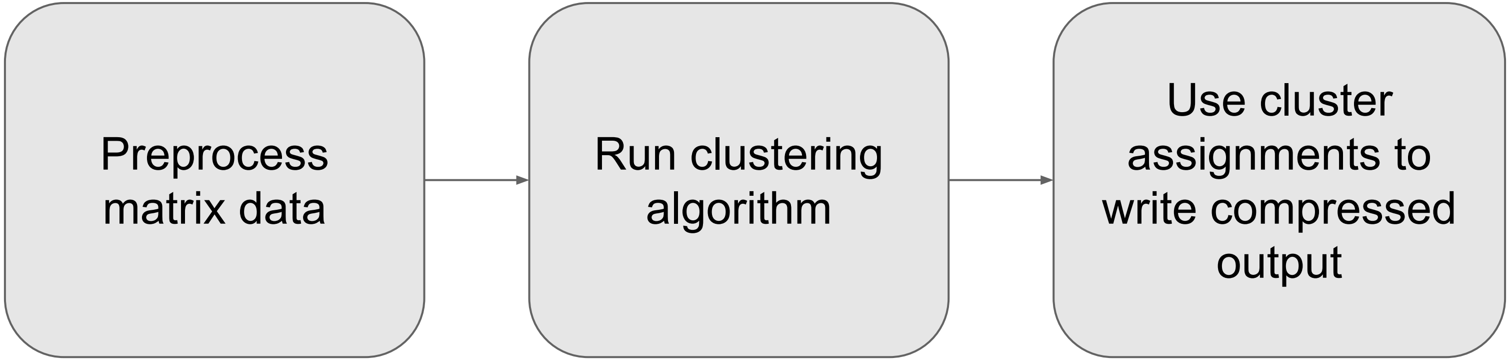
Compression Approaches

High-Level

- Bioinformatics-tailored
- Context-aware
- Examples: Clustering/delta encoding
- **(Our principal approach)**

Low-Level

- More generalized
- Treats data in binary
- Examples: gzip, bitvectors
- **(To be explored)**



Preprocess matrix data

Run clustering algorithm

Use cluster assignments to write compressed output

Our (Lossless) File Format

Format for `cluster_genes.txt`

Columns separated by commas. Each row is not necessarily the same length

	Gene ID 1	Gene ID 2	Gene ID 3	Gene ID 4	Gene ID 5
Cluster 1	9740	36609	36610	2059	
Cluster 2	2059	22634	16186	26143	5765
Cluster 3	24095	11755	23988		

Format for `deltas.txt`

Columns separated by commas. Each row is not necessarily the same length.

	Cluster ID	Gene ID 2	Gene ID 3	Gene ID 4	Gene ID 5
Cell 1	4	8197	6150		
Cell 2	7	4099	4100	8197	34822
Cell 3	1				

Cell 3 is empty because it belongs to a single-cell cluster (outlier)

Dataset used

Database: **Gene Expression Omnibus** (GEO)

- public functional genomics data repository

Dataset: scRNA transcripts collected from white blood cells of patients suffering from breast cancer

We focused on compressing

`GSE294399_WBC_020823_matrix.mtx`

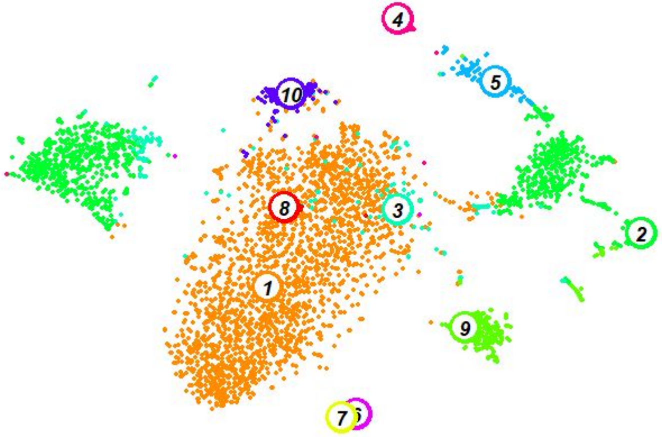
- 36630 genes
- 4137 cells
- 43.2 MB (excluding counts)

- 97.2% sparse

Results for Different Clustering Methods

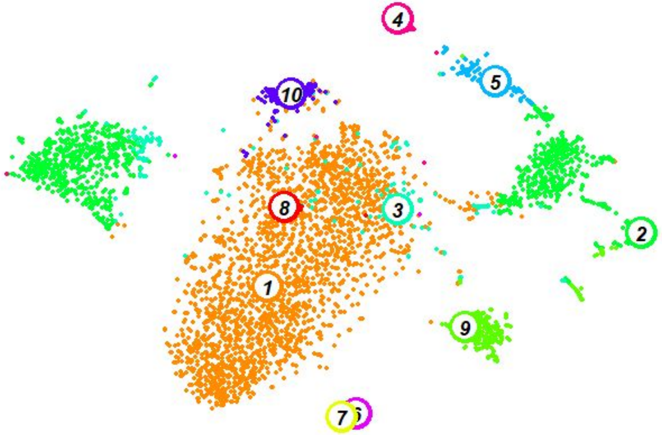
RaceID clustering library

1. Pre-process data
   a. Filter out cells with low transcript count
   b. Normalize
   c. Compute pairwise distance matrix
2. Find clusters using *k*-medoids
3. Post-process results
   a. Reduce outliers
   b. Refine using random forest



RaceID clustering library

- Number of clusters: 10
- Distance metric: Pearson
- Filtering count: 1 (no filtering)
- Execution time: 15 min.



RaceID clustering library - results

`cluster_genes.txt` - 0.003 MB

`deltas.txt` - 23.3 MB

Total - 23.3 MB

46.1%

Compared to

MTX file (counts excluded) - 43.2 MB

Clustering using a neural network

We implemented a neural network that generates clusters based on the cells’ gene profiles

- Gene counts used as input (4137 x 36630 matrix)

Pipeline of neural network

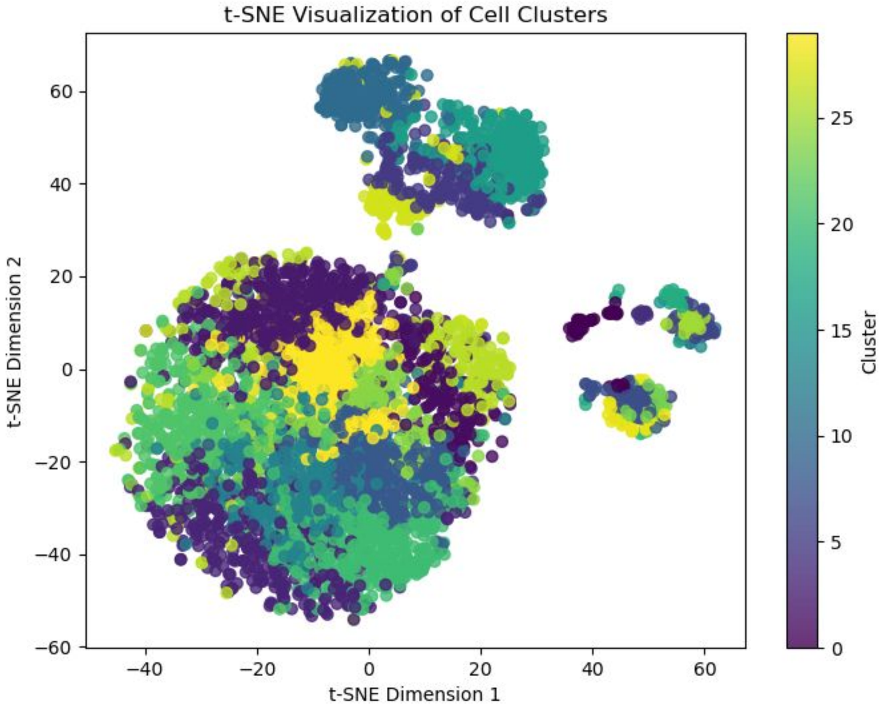
1. Preprocessing
    a. Log transform and scale
2. Represent 36630 genes in 32-dimension latent space
    a. Pretrain autoencoder on given data to encode each gene profile as a vector of length 32
    b. 100 epochs
3. Initialize 30 cluster centers using K-means
4. Iteratively improve cluster centers using the KL divergence between the probabilistic cluster assignment distribution and a more confident target distribution
    a. 20 epochs

Neural network - results

30 clusters found

- 9 outliers
- 21 actual clusters

Clustering took ~ 4.5 minutes



Neural network - results

`cluster_genes.txt` - 0.28 MB

`deltas.txt` - 22.71 MB

Total - 22.99 MB

46.8%

Compared to

MTX file (counts excluded) - 43.2 MB

Results - summary

MTX file (no counts) - 43.2 MB

	RaceID	Neural Network
Running time	~15 minutes	~4.5 minutes
Clusters	10	30
Storage	23.3 MB	22.99 MB

Future work

1. **Collect more compression data from more datasets**
2. **Compare to other high-level compression methods (CSR, CSC)**
3. **Measure the effect of the number/quality of clusters on compression**
4. More clustering techniques (low priority due to time)
5. **Add information about counts as a third file**
6. Low-level compression (bit vector, etc.) (low priority due to time)

Preview: Tar GZ

MTX file (no counts) - 43.2 MB

Tar GZ applied to uncompressed MTX - 11.4 MB

	RaceID	Neural Network
Storage	23.3 MB	22.99 MB
Tar GZ Storage	8.33 MB	8.24 MB
Compression (wrt original MTX file)	80.8%	80.9%