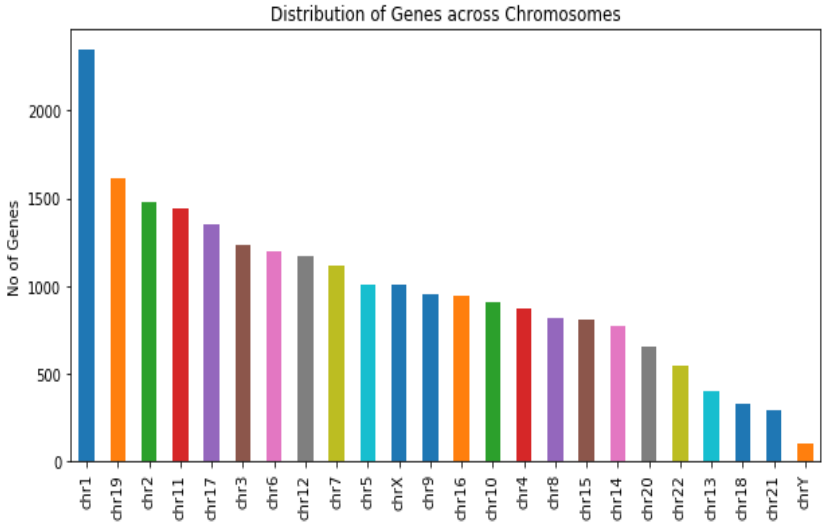
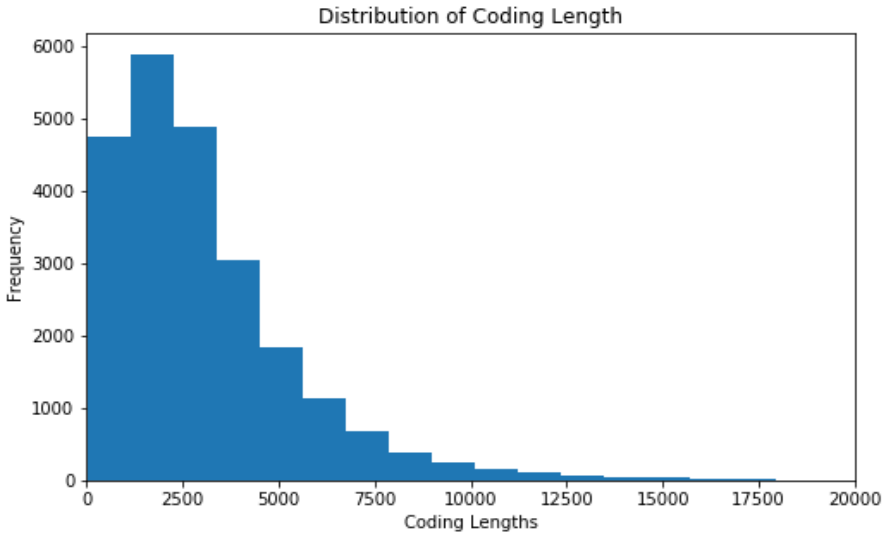
Mini Project 2: Unsupervised Single-Cell Analysis in Triple-Negative Breast Cancer

CS 498: Data Science & Analytics (Spring 2019)

Task 0 – Getting Started

1. Need for multiple cells
   - to observe the effect of metformin on the genes
   - to determine the causal relationship of metformin on different cells
   - to determine the causal relationship of metformin on different cells
2. No of sequenced cells: **192**
3. No of sequenced genes: **23346**
4. Are genes equally distributed across chromosomes: **No**
5. Distribution of coding length
   - Looks like Exponential distribution





Task 1.1: Bayesian Network for Quality Control

a. Factorization of the joint probability distribution.

$$P(Temp, Time, Viab, Qual) = P(Qual \mid Time, Viab)P(Viab \mid Temp, Time)P(Temp)P(Time)$$

b. Number of parameters needed to define the conditional probability distribution

$$P(\text{Qual}|\text{Time},\text{Time}) = \sum_{\text{Viab}} P(\text{Qual}|\text{Viab},\text{Time})P(\text{Viab}|\text{Temp},\text{Time})$$

No of parameters = 4+8 = **12**

c. Conditional probability tables

P(Time)	
Time = long	0.0964
Time = short	0.9036

P(Temp)	
Temp = Hot	0.0958
Temp = Warm	0.3044
Temp = Cool	0.4948
Temp = Cold	0.105

P(Qual|Viab,Time)

viab	time	qual = good	qual = bad
high	short	0.894852	0.105148
high	long	0.522293	0.477707
low	short	0.517321	0.482679
low	long	0.047619	0.952381

P(Viab|Temp, Time)

temp	time	viab = high	viab = low
cool	short	0.948268	0.0517319
cool	long	0.792829	0.207171
hot	short	0.212815	0.787185
hot	long	0.0714286	0.928571
warm	short	0.906657	0.093343
warm	long	0.707143	0.292857
cold	short	0.415966	0.584034
cold	long	0.265306	0.734694

Task 1.1: Bayesian Network for Quality Control

d. P (Qual | Temp, Time)

$$P(\text{Qual}|\text{Temp},\text{Time}) = \sum_{\text{Viab}} P(\text{Qual}|\text{Viab},\text{Time})P(\text{Viab}|\text{Temp},\text{Time})$$

e. Quality of the sequenced cells

- **Criteria for Bad Quality**

  Cell = Bad Quality if
  P (Qual = **bad** | Temp, Time) > 0.5

- Observation: Cells which have **long time**: are bad quality cells

- Searched cells matching this criteria in BayesInferenceBase.csv, BayesInferenceMetf.csv

- Dropped the cells for further analysis

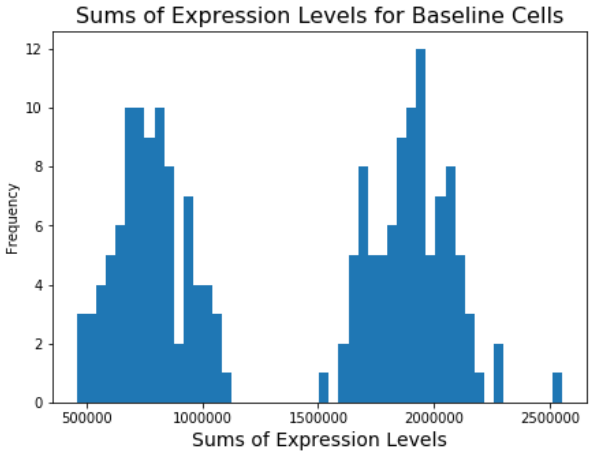
P (Qual | Temp, Time)

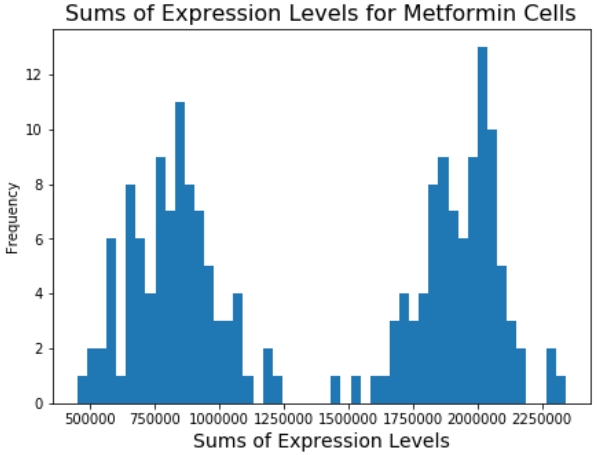
temp	time	qual = good	qual = bad
cool	short	0.875322	0.124678
cool	long	0.423954	0.576046
hot	short	0.597665	0.402335
hot	long	0.0815243	0.918476
warm	short	0.859612	0.140388
warm	long	0.383281	0.616719
cold	short	0.674361	0.325639
cold	long	0.173553	0.826447

Task 1.2: Data Standardization

a. Sum of Expression Levels

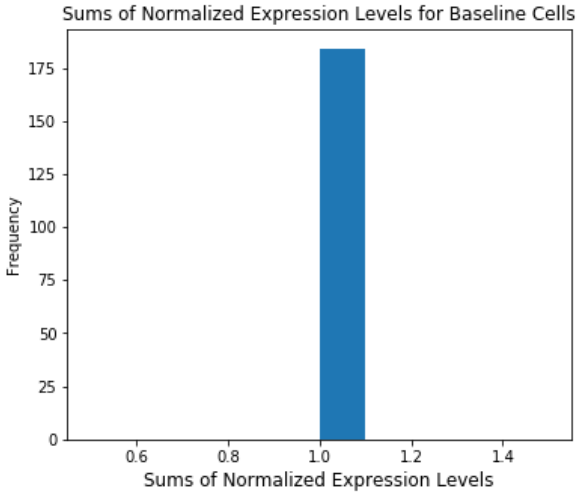
Observation: It seems like there are two brackets in which the sum of Expression levels are divided and the ones with sum expressions between around 500000 and 1250000 got spread after Metformin and the ones with sum of expressions between 1500000 and 2500000 got shifted a little towards right

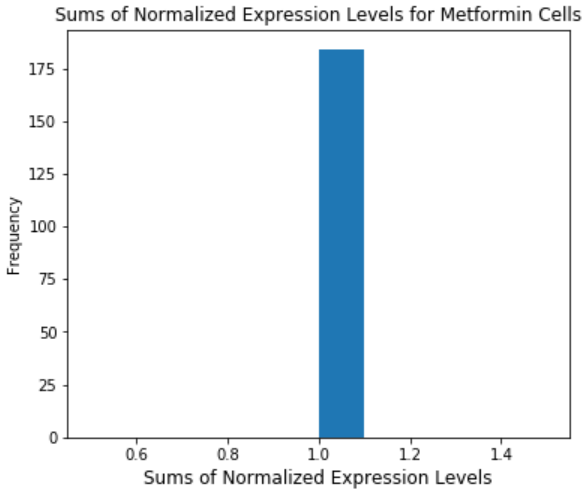




b. Sum of Normalized Expression Levels

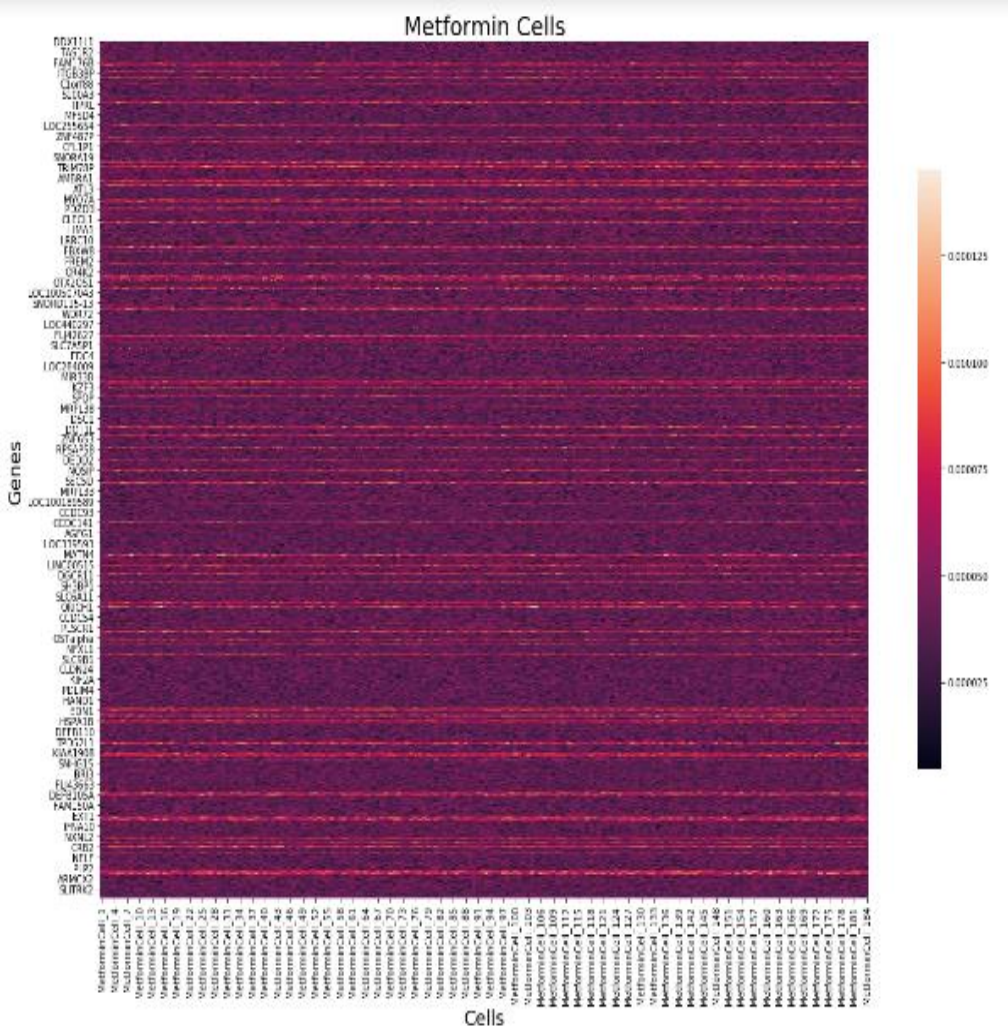
Observation:
All cells will fall into one bin in the histogram



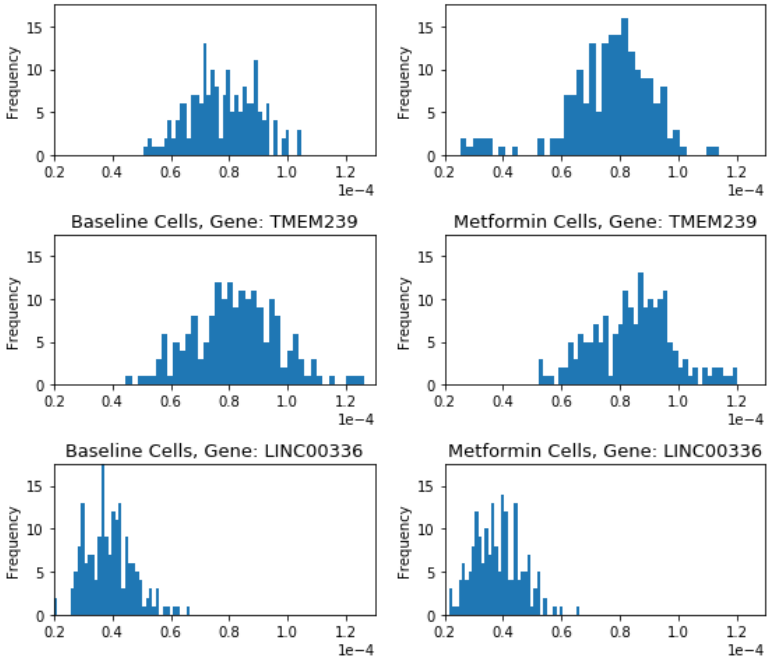


Task 1.3: Visual Inspection

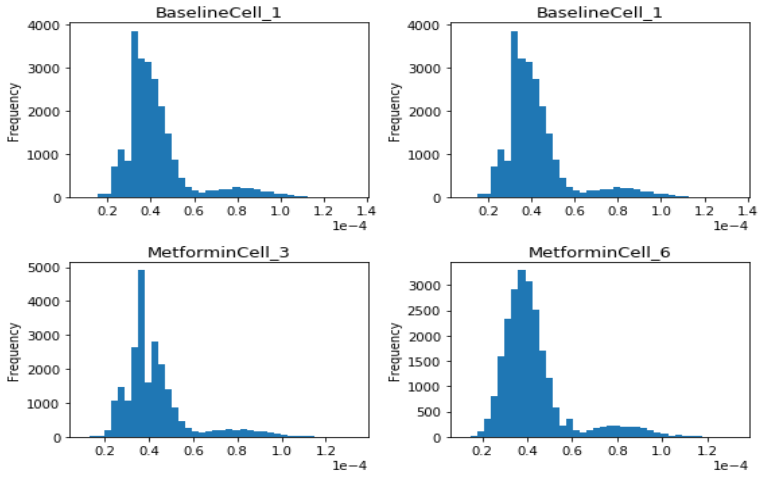
Heat Map



distribution of gene expression across cells



distribution of gene expression across genes



Task 2.1: Kolmogorov–Smirnov (KS) Test

a. KS Test is a **non-parametric** test. These are used when the underlying distribution is not known or the data does not satisfies the assumptions of parametric tests.

b. **p-value** of a two-sample KS test

```
p_values.head()
```

```
0    0.888064
1    0.472197
2    0.133963
3    0.888064
4    0.732819
dtype: float64
```

c. **Null Hypothesis**: The gene's expression is not altered which means the difference observed in its expression level between baseline cells and metformin-treated cells is not statistically significant.

d. # of genes with significantly altered expression

	alpha	count
0	0.100	2301.0
1	0.050	1385.0
2	0.010	325.0
3	0.005	170.0
4	0.001	42.0

Task 2.2: Multiple Testing

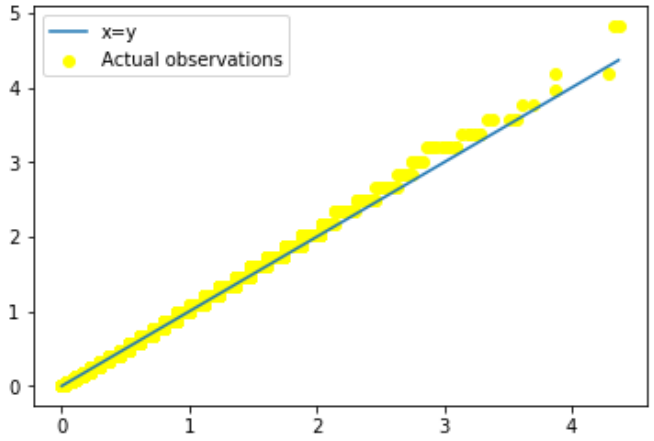
a. **P-value** in our context is the threshold of whether gene's expression is altered , which means the difference observed in its expression level between baseline cells and metformin-treated cells is statistically significant.

b. p-values follow a uniform distribution if null hypothesis is true

(c) p-values if no gene expression was altered

	alpha	Expected Count	Actual Count
0	0.100	2334.600	2301.0
1	0.050	1167.300	1385.0
2	0.010	233.460	325.0
3	0.005	116.730	170.0
4	0.001	23.346	42.0

(d) Q-Q plot

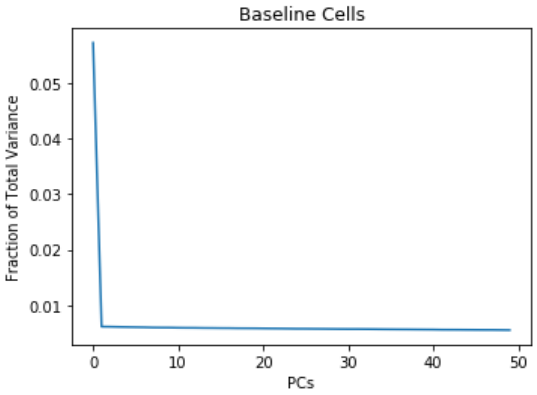


e. (i) After taking -log10(pvalues) the ones with the highest values of -log10(pvalues) are the ones in the "tail"of the p-value distribution and are more visible

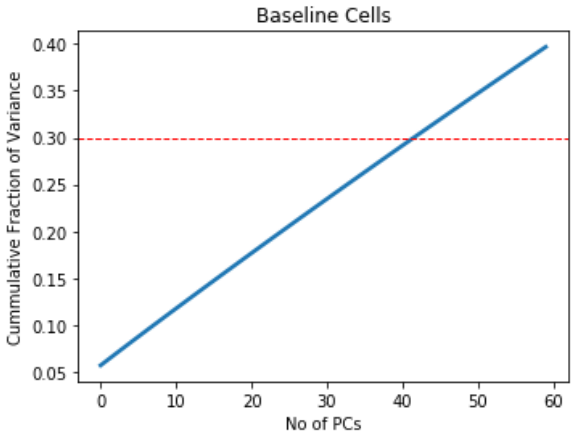
(ii) The Q-Q plot approximately aligns with the x=y which means that the p-values are uniformly distributed implying that the null hypothesis is true

Task 3.1-2: PCA & t-SNE

Scree Plot

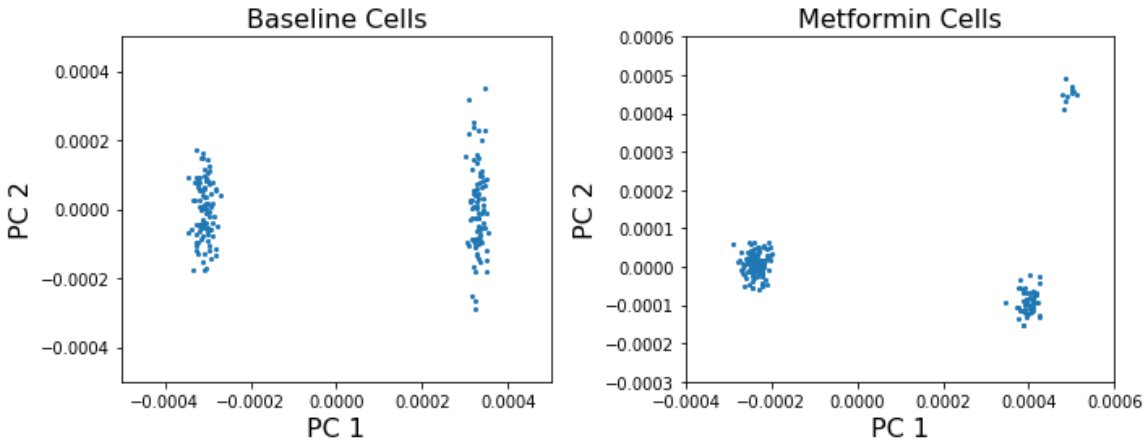


Cumulative Variance

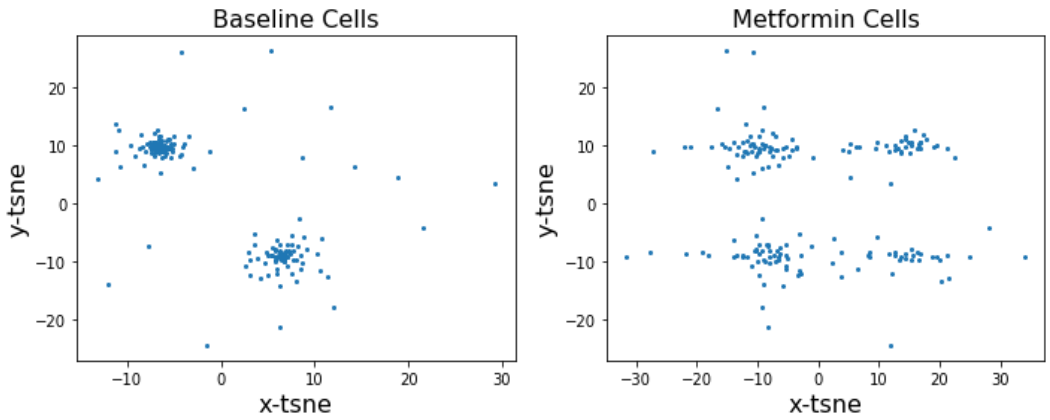


No of PCs needed to explain 30% of total variance**: 42 PCs**

PCA Results



t-SNE Results



Comparison: PCA forms good clusters whereas t-SNE does not result in good clusters

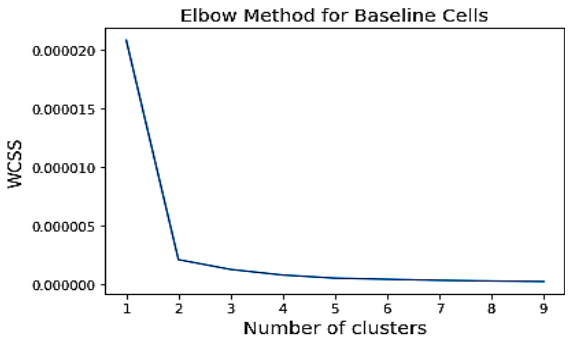
Task 3.3: Clustering

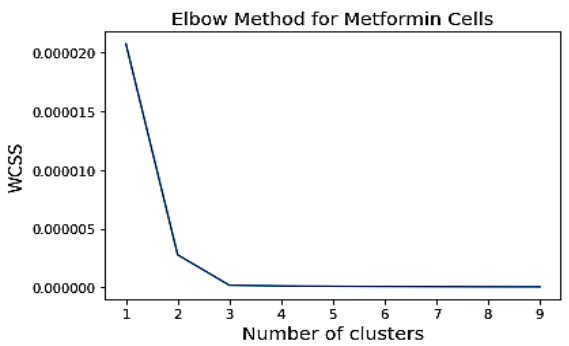
(a) k-means clustering

- **Elbow Method**

i. Baseline Cells: **2** clusters

ii. Metformin Cells: **3** clusters



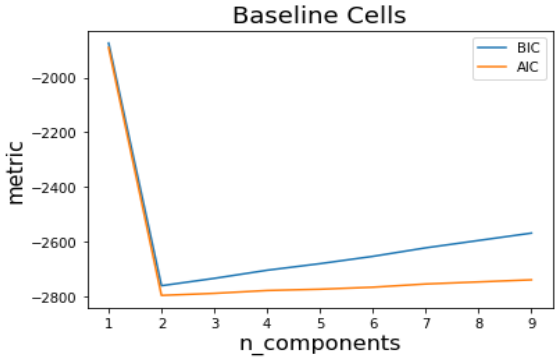


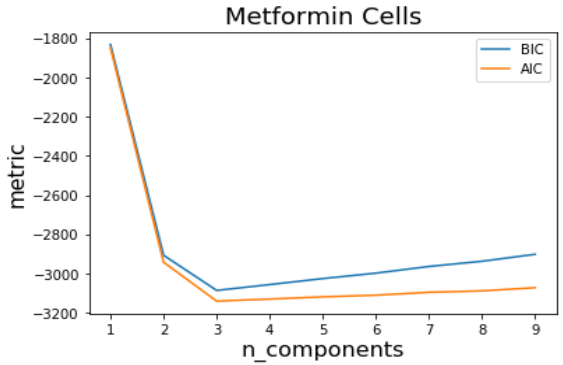
(b) GMM clustering

- **Elbow Method**

i. Baseline Cells: **2** clusters

ii. Metformin Cells: **3** clusters



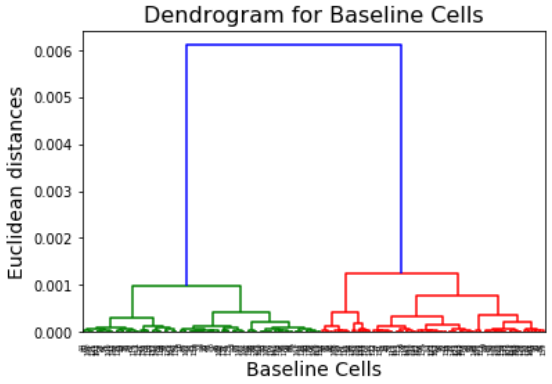


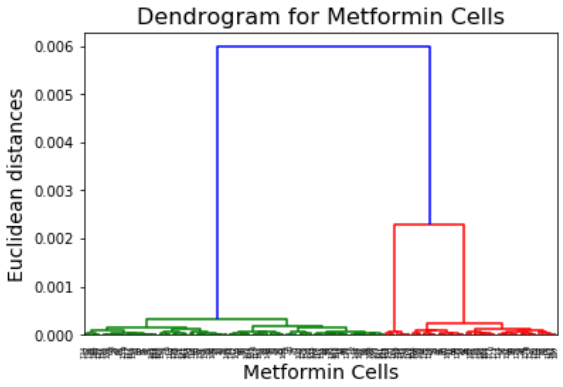
(c) Hierarchical clustering

- **Dendrogram Method**

i. Baseline Cells: **2** clusters

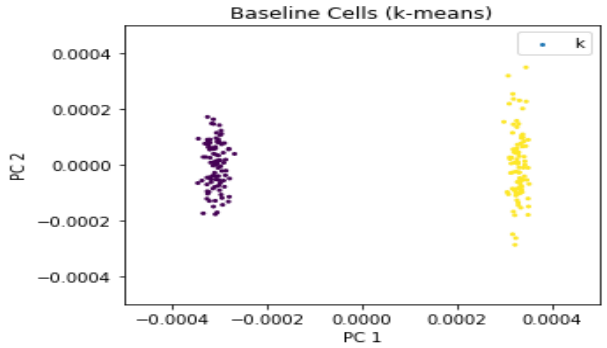
ii. Metformin Cells: **3** clusters

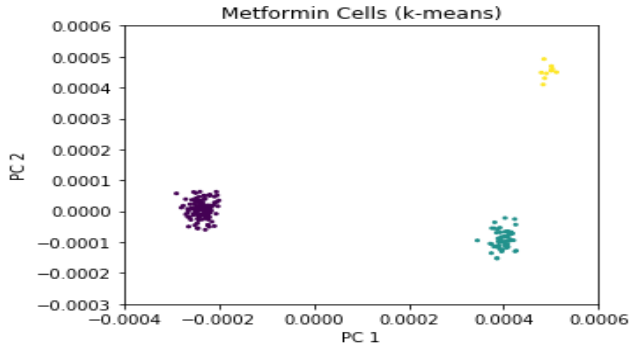




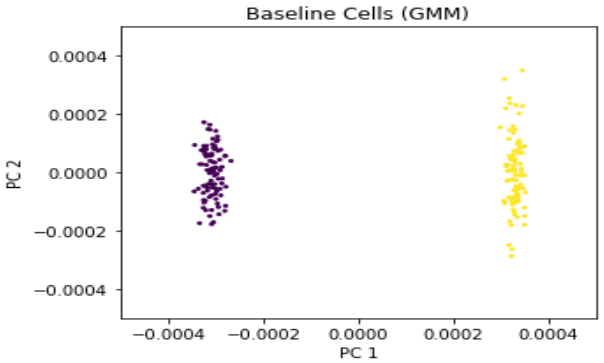
Task 3.3: Clustering

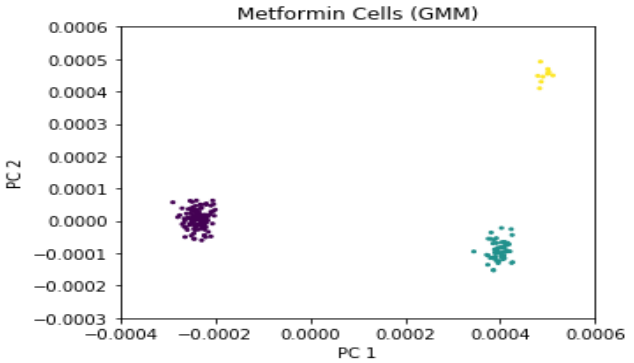
(a) k-means clustering



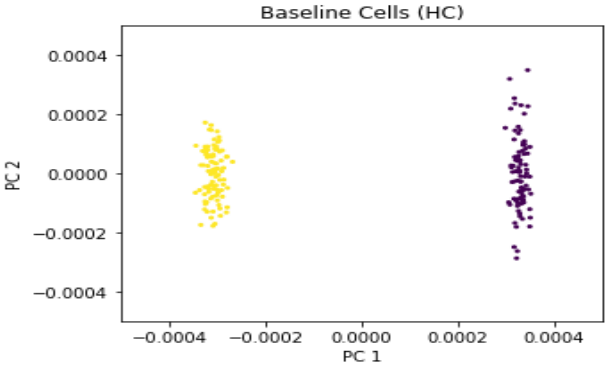


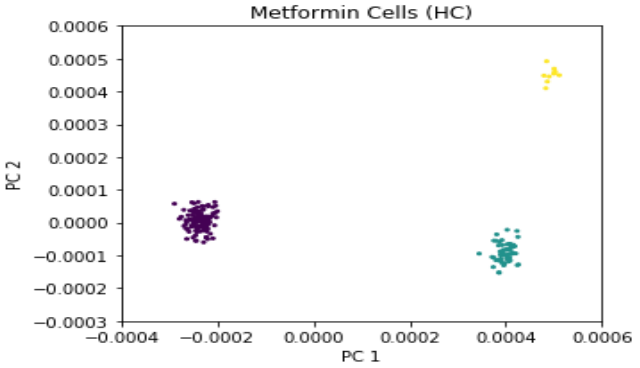
(b) GMM clustering





(c) Hierarchical Clustering





Similar results were obtained. We decided to use **k-means**

Task 4.1: Identify altered genes

(a) Use mean of GMM to compare and all Mi are not identical to one of the baseline subpopulations.

(b) Use Euclidean Distance between mean of each cluster.

- M0 is most similar to baseline subpopulation cluster 0
- M1 is most similar to baseline subpopulation cluster 1
- M2 is most similar to baseline subpopulation cluster 1

(c) altered genes:

['TRNP1', 'OMA1', 'LRRC8D', 'LOC653513', 'LOC339529', 'OR2AK2', 'PHYH', 'LOC100507605', 'SPOCK2', 'LOC728190', 'TCTN3', 'PIK3AP1', 'MIR608', 'MIR4295', 'HPS5', 'LINC00301', 'B3GNT1', 'SRSF8', 'C11orf92', 'CDKN1B', 'DHH', 'RDH16', 'CCDC59', 'RMST', 'SDS', 'LOC116437', 'RNF6', 'GUCY1B2', 'SLC15A1', 'MRPL52', 'MIR409', 'LOC283663', 'MYO1E', 'SNORD18A', 'IDH3A', 'PGP', 'VASN', 'CEP112', 'SNORA37', 'ZNF57', 'NRTN', 'TRAPPC5', 'LOC284395', 'CYP2A6', 'ERCC2', 'C19orf63', 'NAB1', 'PRKAG3', 'SSTR4', 'HNF4A', 'LINC00310', 'DSCAM-AS1', 'MIR3928', 'BTD', 'TCTA', 'GNAI2', 'NPHP3', 'GBA3', 'RANBP9', 'IL17F', 'TMEM244', 'OSTCP1', 'SKAP2', 'AUTS2', 'KCND2', 'FAM71F1', 'CREB3L2', 'SORBS3', 'CCDC25', 'SGK196', 'MYC', 'LINC00051', 'GML', 'MELK', 'LOC100507244', 'FAM47A', 'PABPC1L2A']

Task 4.2: Gene Set Characterization

(a) The most common pathways: to get the pathways with most altered genes.

	way	gene	times
175	Wnt signaling pathway	[AADACL4, ADAD1, ADAM17, ADCY4, ADH1B, ADRA2B,...	31
115	Notch signaling pathway	[AAMP, ABI2, ABRA, ACBD3, ACTR1A, ADCYAP1, AFG...	20

(b) Novel genes:

['LOC339529', 'C11orf92', 'SDS', 'GUCY1B2', 'VASN', 'CEP112', 'CYP2A6', 'PRKAG3', 'NPHP3', 'GBA3', 'SKAP2', 'KCND2', 'CREB3L2', 'LOC100507244', 'FAM47A']