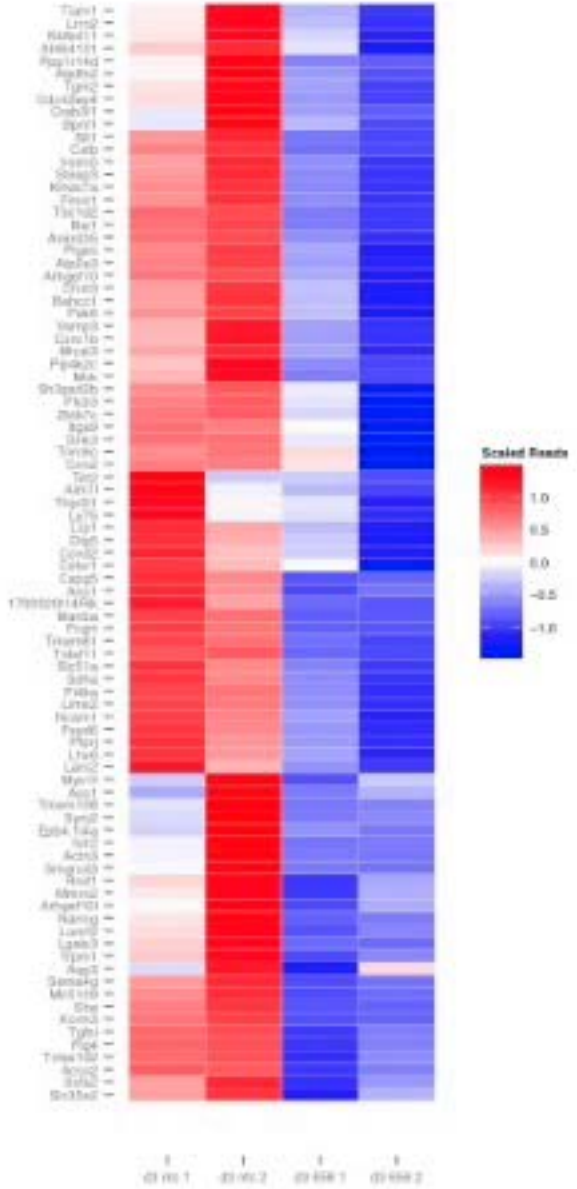
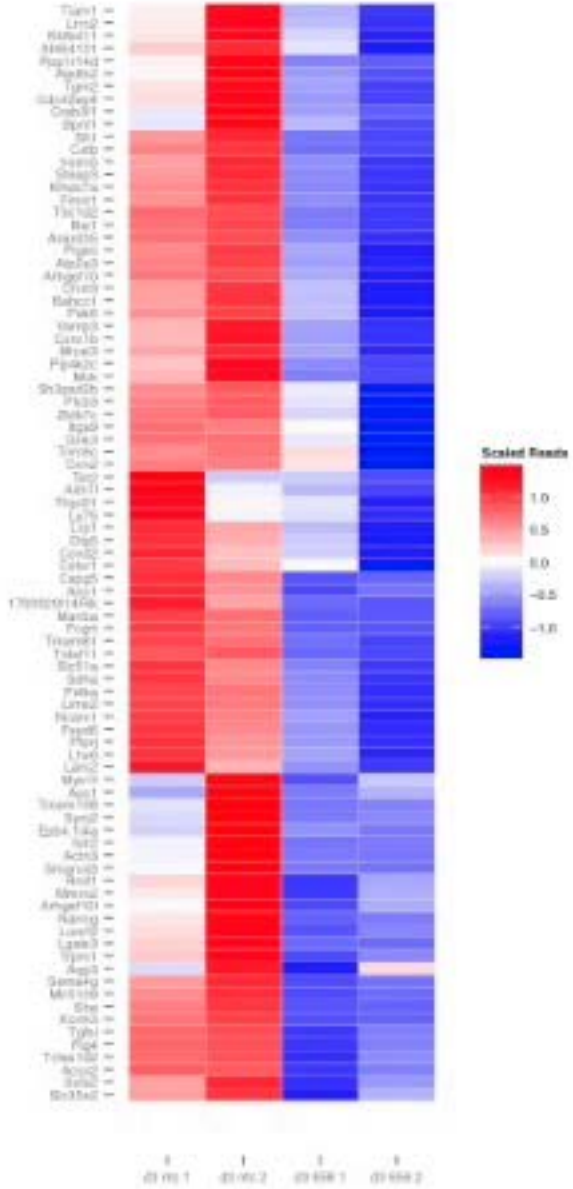


Here's a heatmap!

The **rows** are **genes**.

The **columns** are **RNA-seq samples**.

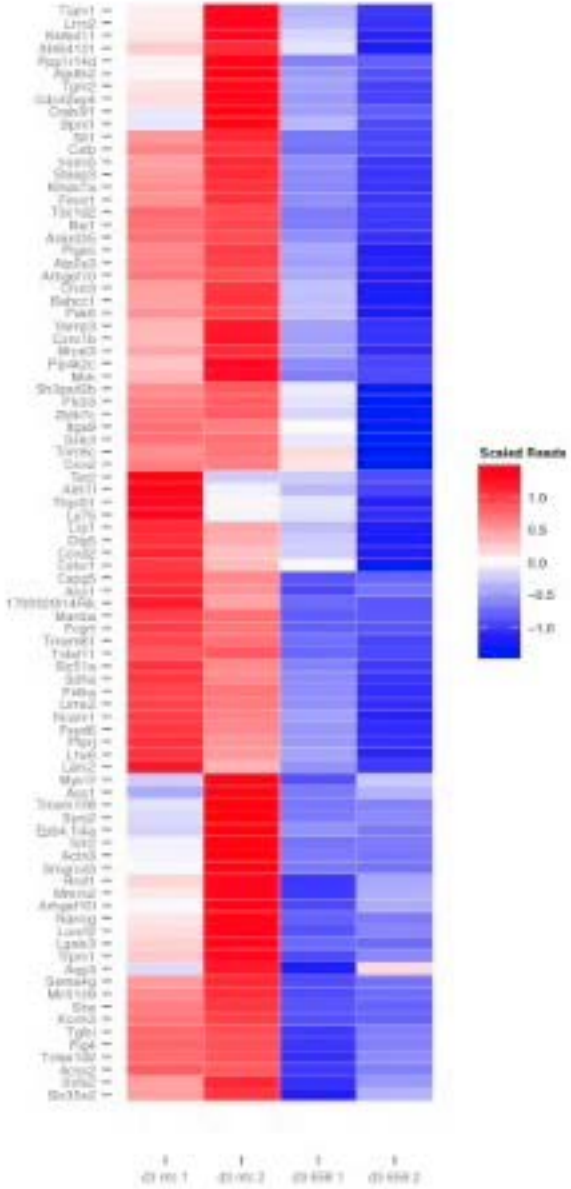


Here's a heatmap!

The **rows** are **genes**.

The **columns** are **RNA-seq samples**.

This data has been modified in 2 ways so that we can gain some insights from it.



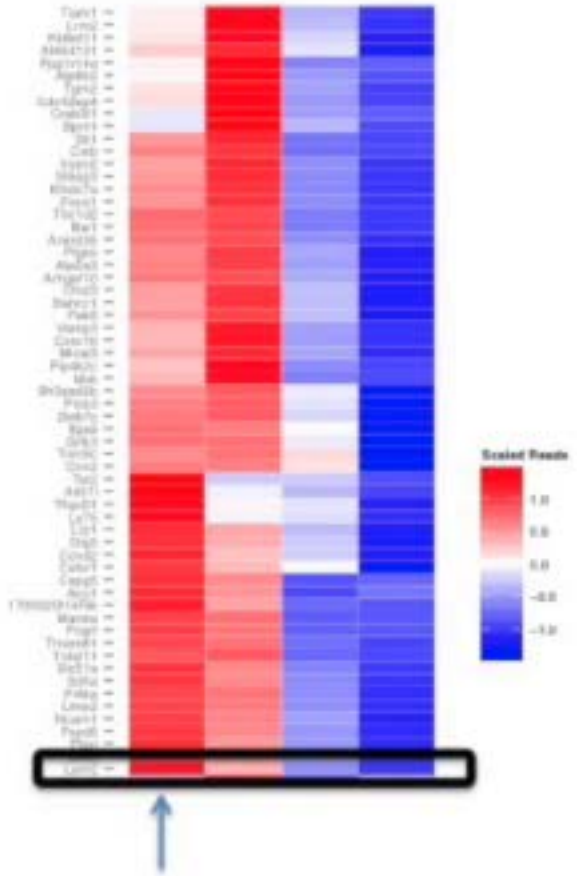
Here's a heatmap!

The **rows** are **genes**.

The **columns** are **RNA-seq samples**.

This data has been modified in 2 ways so that we can gain some insights from it.

1) The **relative abundances** have been scaled. In this case, this was done on per gene basis (other heatmaps scale all the genes at once). This makes it easy to see that sample X has more/less of gene Y than sample Z.



It's easy to see that Sample 1 expresses this gene more than the others.

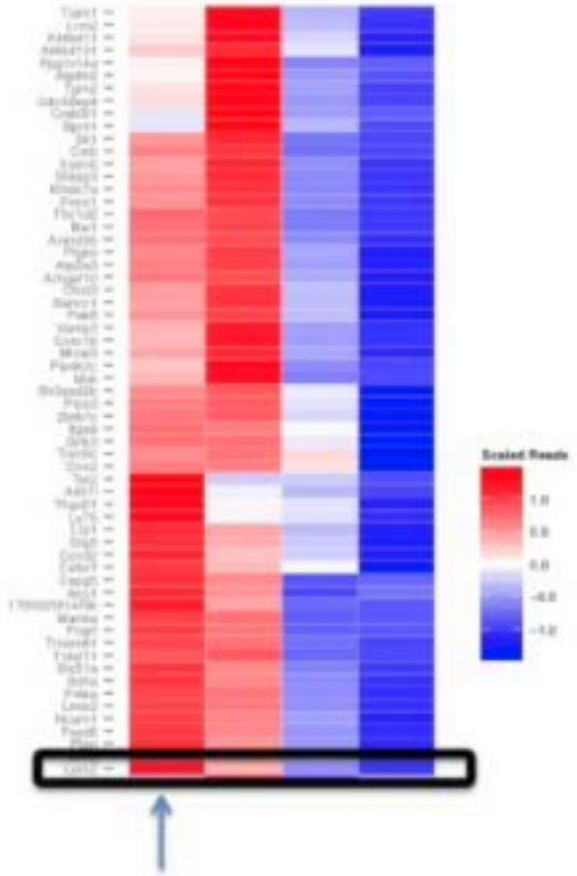
Here's a heatmap!

The **rows** are **genes**.

The **columns** are **RNA-seq samples**.

This data has been modified in 2 ways so that we can gain some insights from it.

1) The **relative abundances** have been scaled. In this case, this was done on per gene basis (other heatmaps scale all the genes at once). This makes it easy to see that sample X has more/less of gene Y than sample Z.



Here's a heatmap!

The **rows** are **genes**.

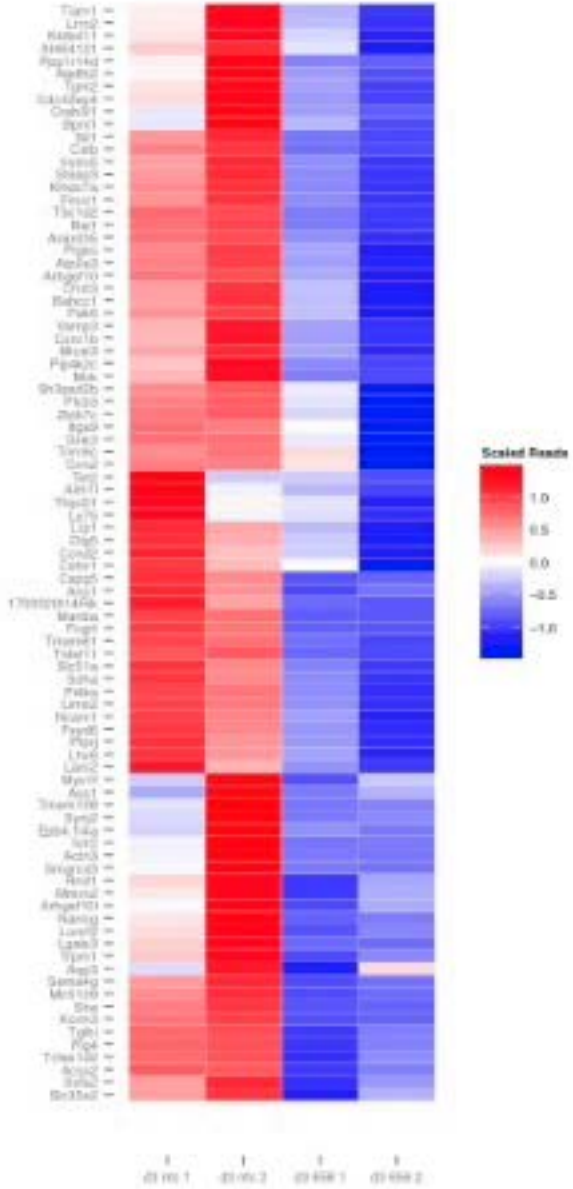
The **columns** are **RNA-seq samples**.

This data has been modified in 2 ways so that we can gain some insights from it.

1) The **relative abundances** have been scaled. In this case, this was done on per gene basis (other heatmaps scale all the genes at once). This makes it easy to see that sample X has more/less of gene Y than sample Z.

It's easy to see that Sample 1 expresses this gene more than the others.

However, this specific scaling means we can't compare across genes. The dark red bar in the Sample 1 for this gene doesn't mean that Sample 1 transcribes it more than other genes, just other samples.



Here's a heatmap!

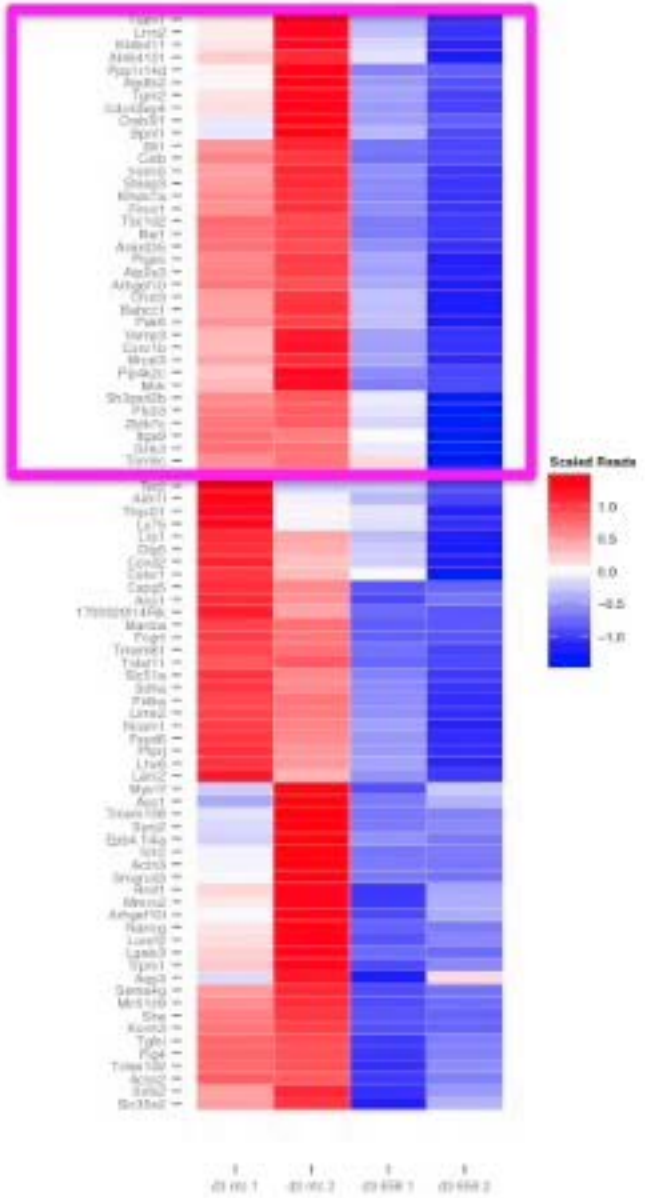
The **rows** are **genes**.

The **columns** are **RNA-seq samples**.

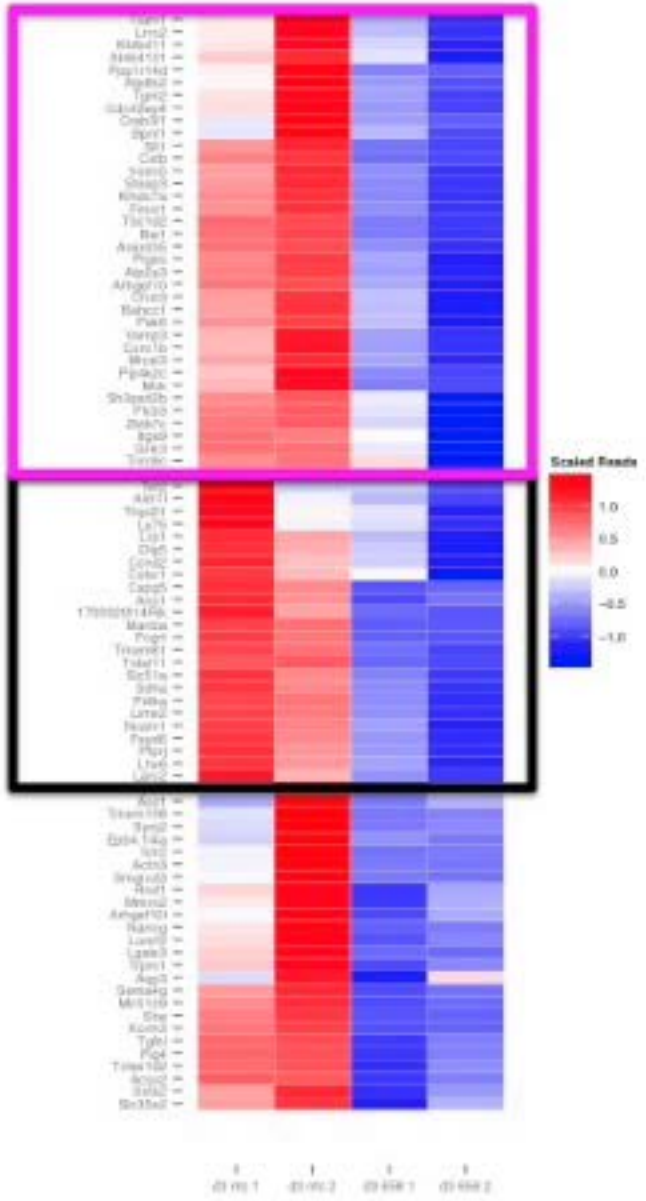
This data has been modified in 2 ways so that we can gain some insights from it.

1) The **relative abundances** have been scaled. In this case, this was done on per gene basis (other heatmaps scale all the genes at once). This makes it easy to see that sample X has more/less of gene Y than sample Z.

2) The rows/genes have been grouped according to "similarity".

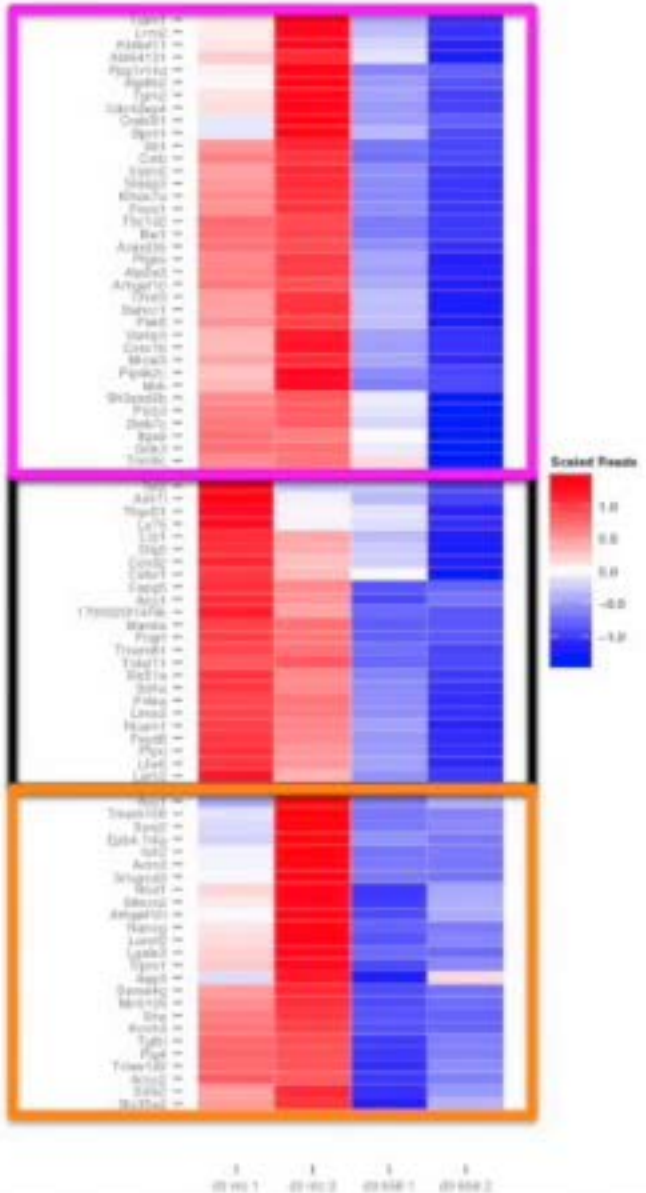


These genes are transcribed most in the 2nd sample (and least in the 4th sample).



These genes are transcribed most in the 2nd sample (and least in the 4th sample).

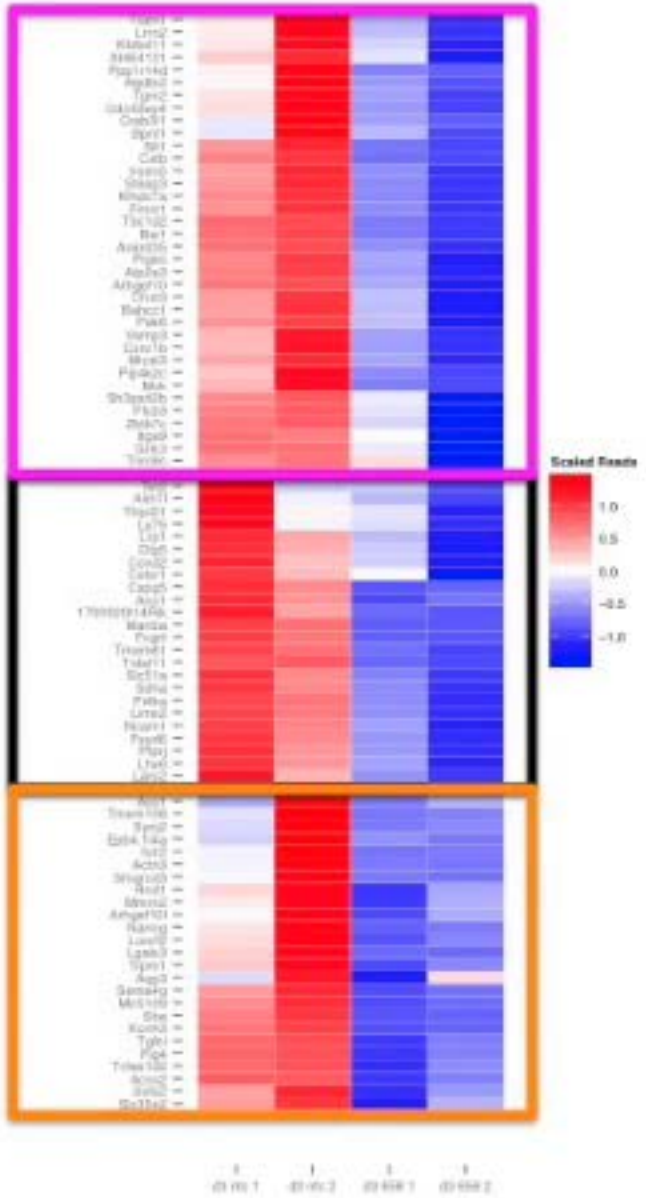
These genes are transcribed most in the 1st sample (and least in the 4th sample).



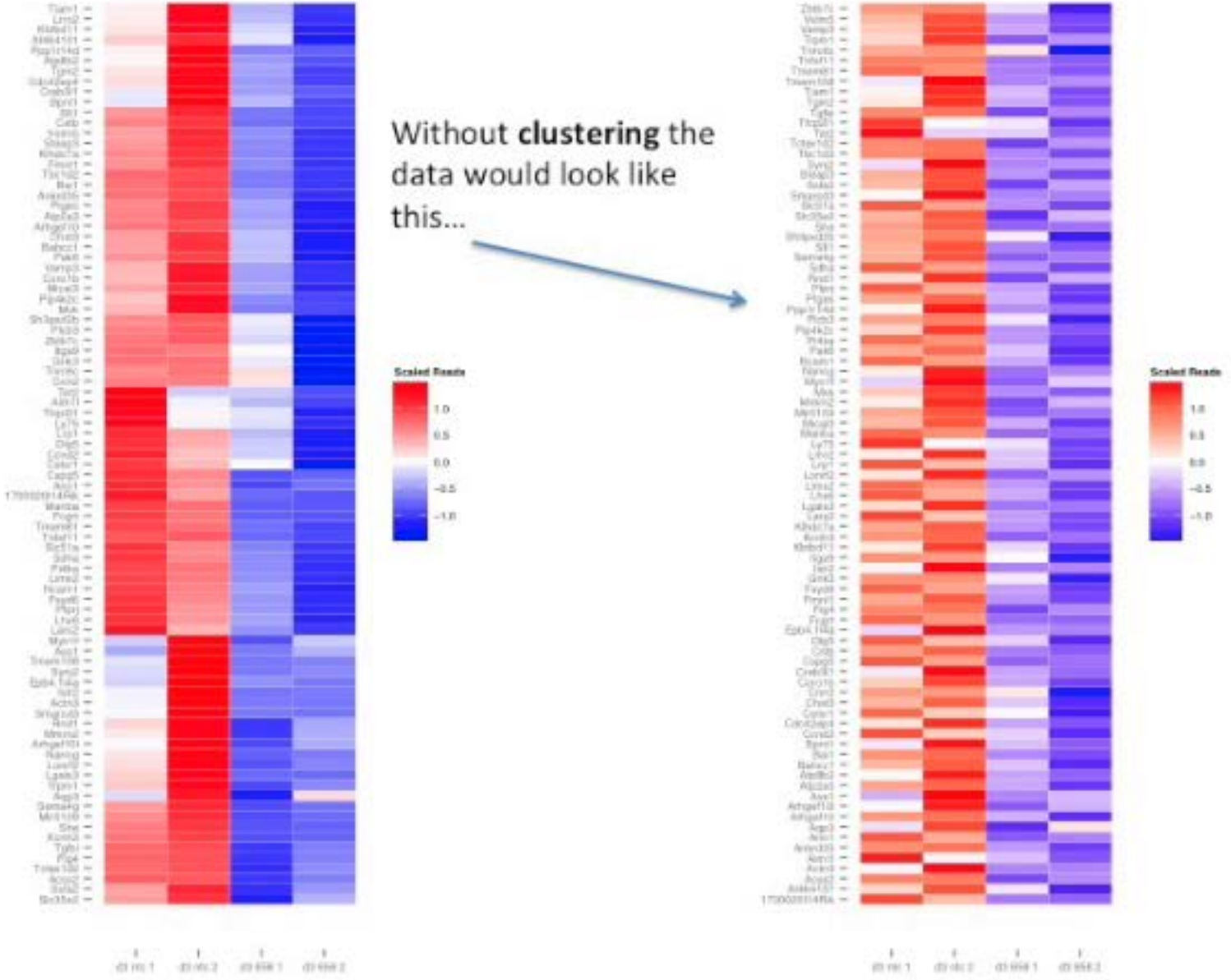
These genes are transcribed most in the 2nd sample (and least in the 4th sample).

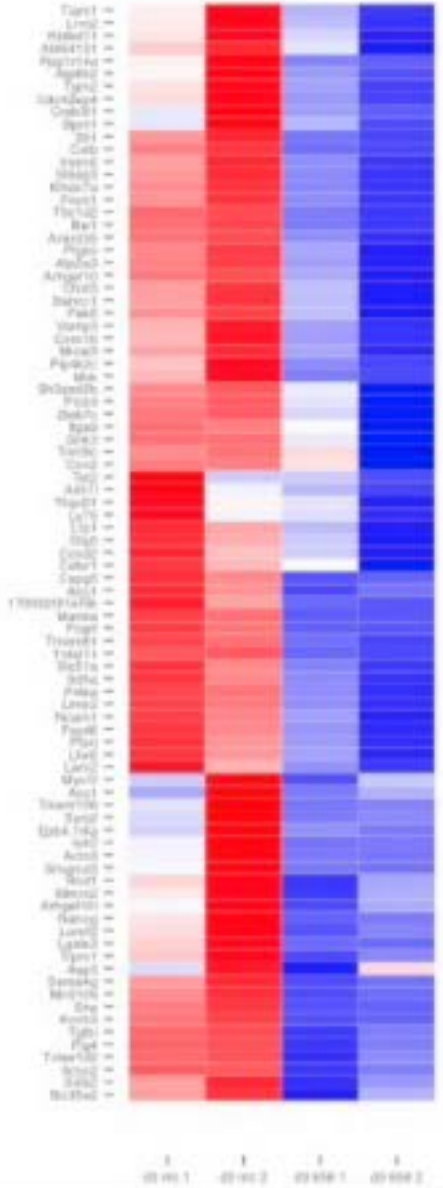
These genes are transcribed most in the 1st sample (and least in the 4th sample).

These genes are transcribe most in the 2nd sample (and least in the 3rd sample).

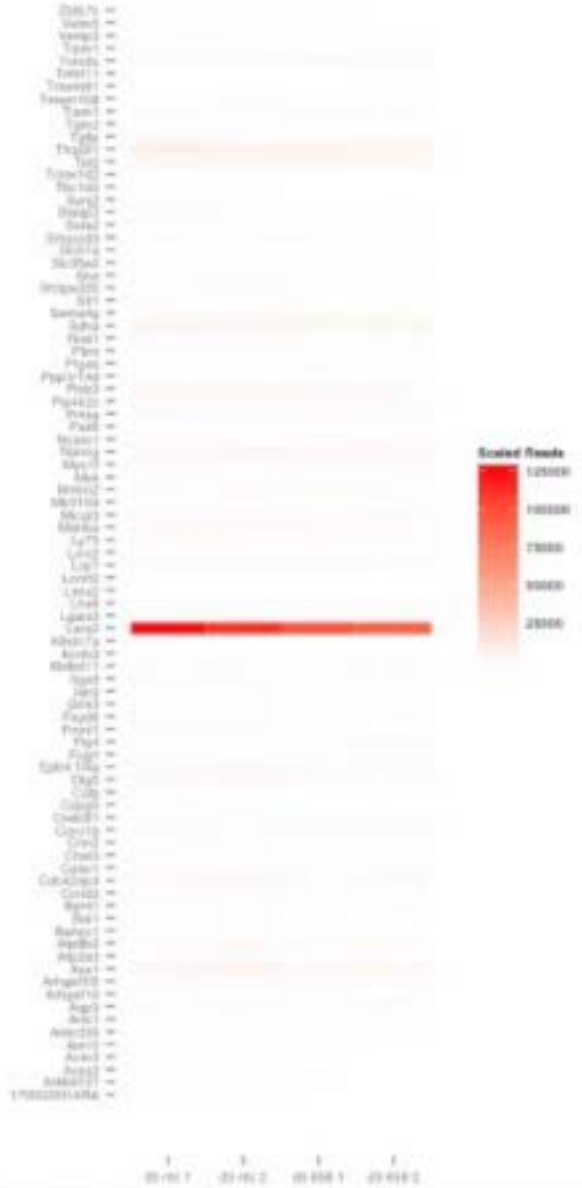


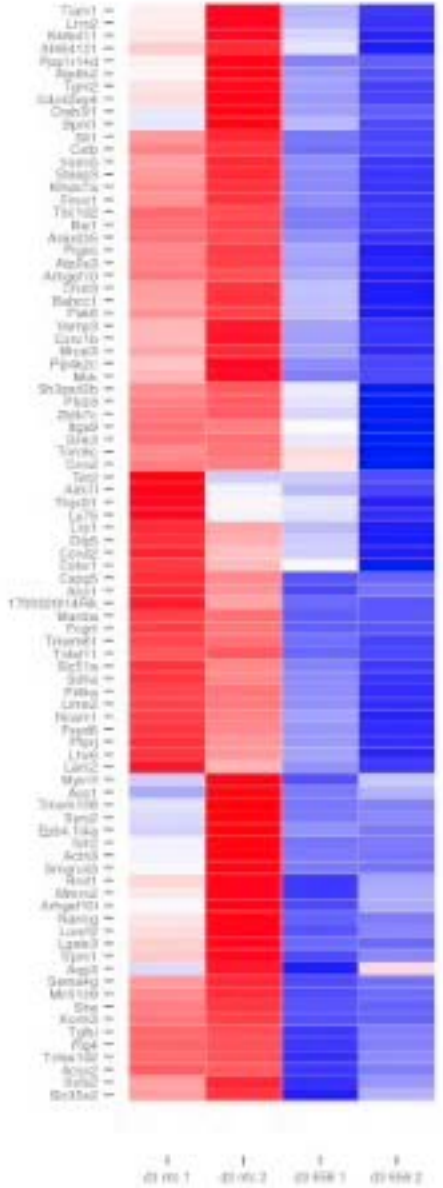
The “clustering” isn’t by chance, but due to a computer program that tries to put “similar” things close together.


Without **clustering** the data would look like this...
Scaled Reads
1.0
0.5
0.0
−0.5
−1.0
Scaled Reads
1.0
0.5
0.0
−0.5
−1.0



Without **clustering** or **scaling**, the data would look like this!!!!



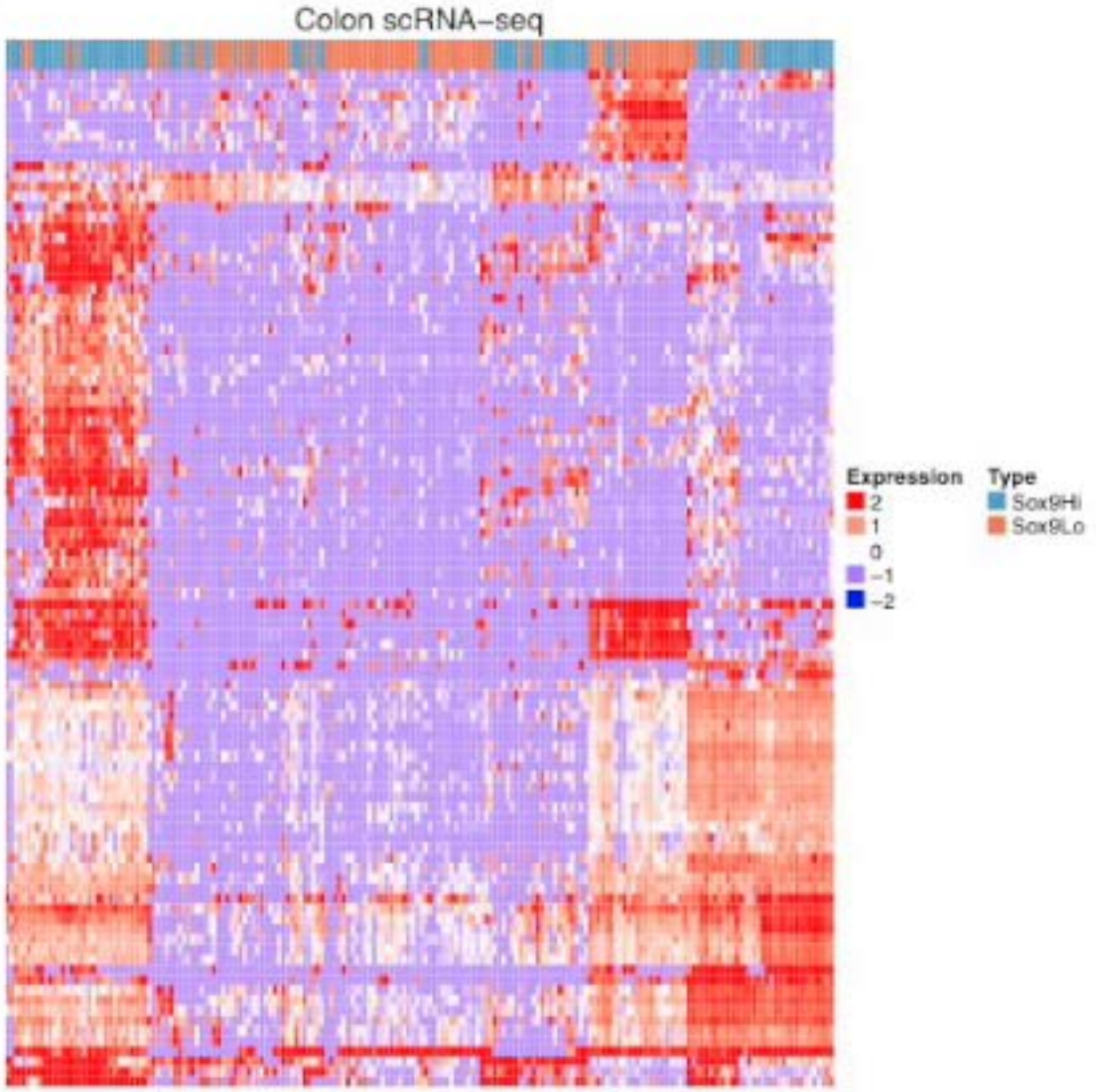


Without **clustering** or **scaling**, the data would look like this!!!!

Notice that one gene is highly transcribed compared to the others.

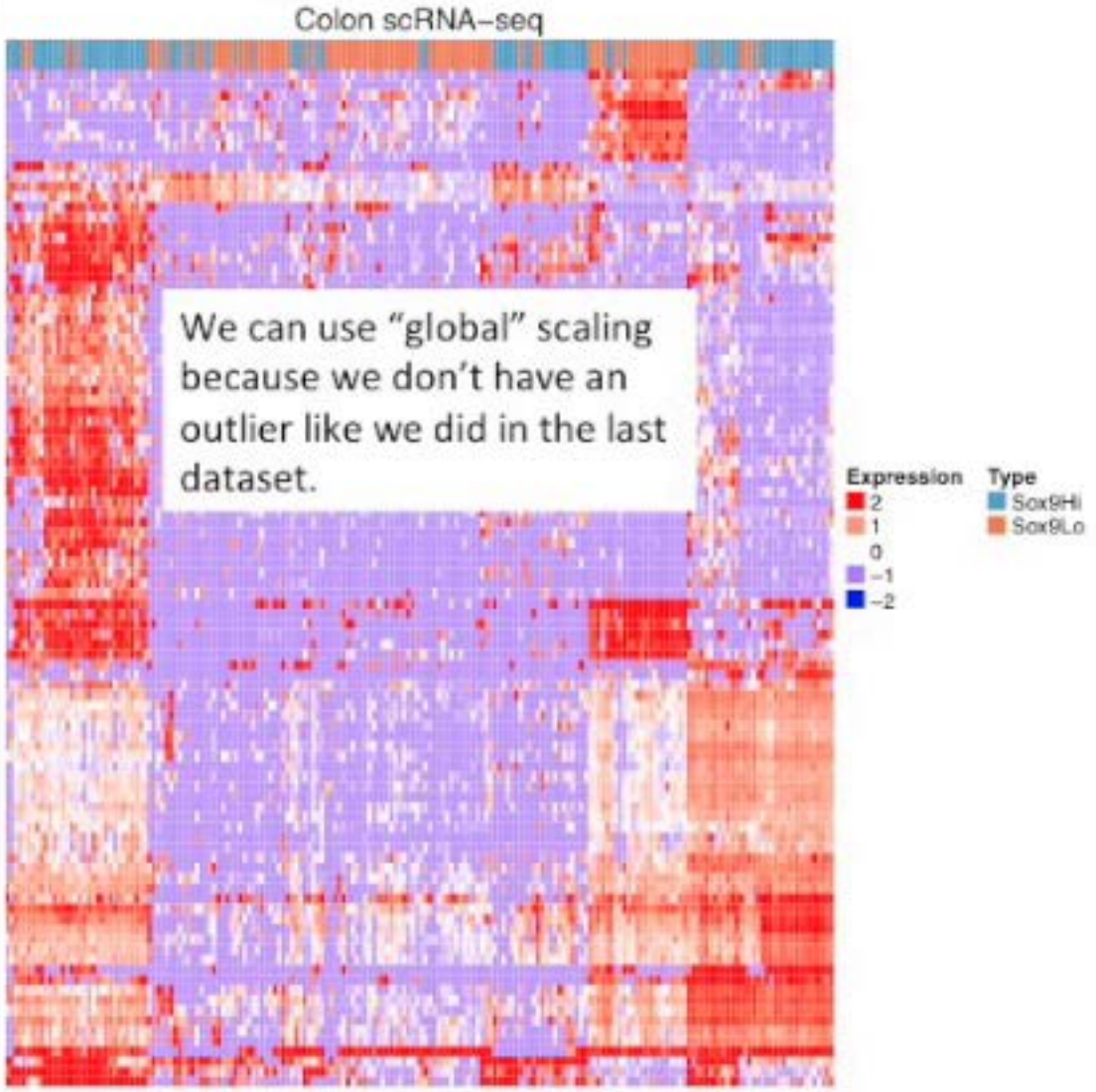
It's an outlier…

Another example…



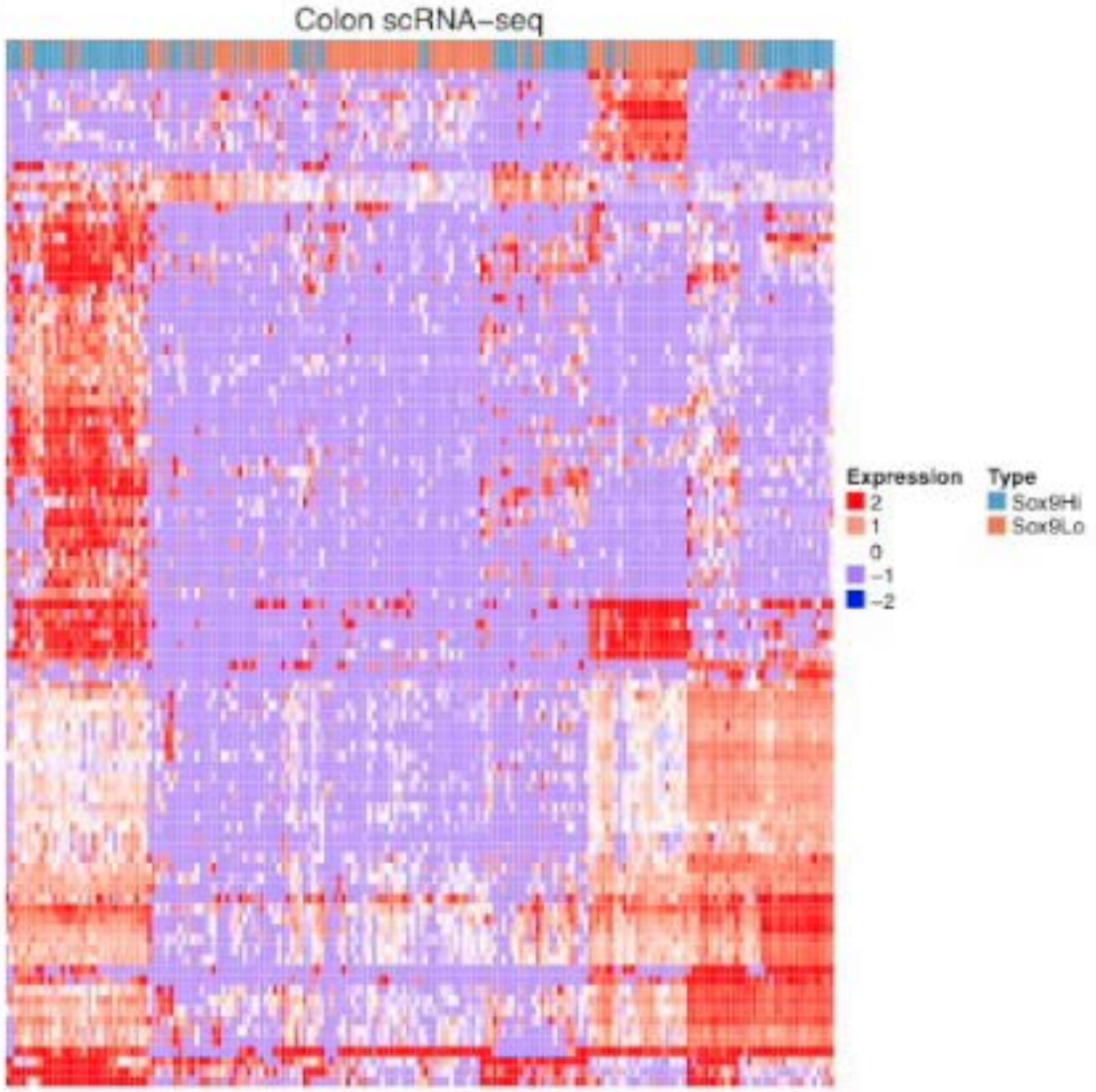
This heatmap has been **scaled** and **clustered**.

The **scaling** is "global" – not per row/gene – but for all rows/genes.



This heatmap has been **scaled** and **clustered**.

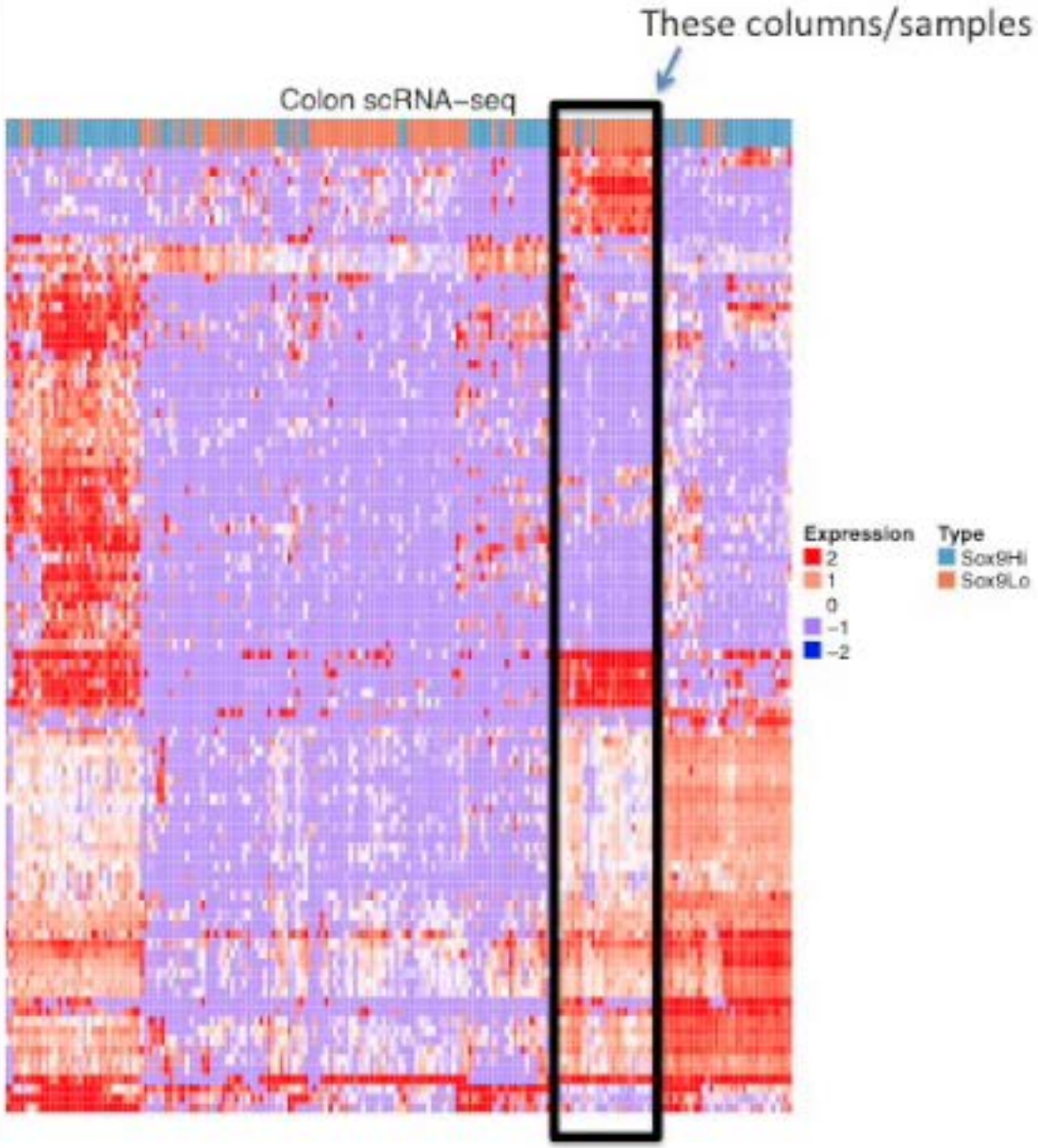
The **scaling** is "global" – not per row/gene – but for all rows/genes.



This heatmap has been **scaled** and **clustered**.

The **scaling** is "global" – not per row/gene – but for all rows/genes.

The **clustering** is by **column**/sample **AND** by **row**/gene.

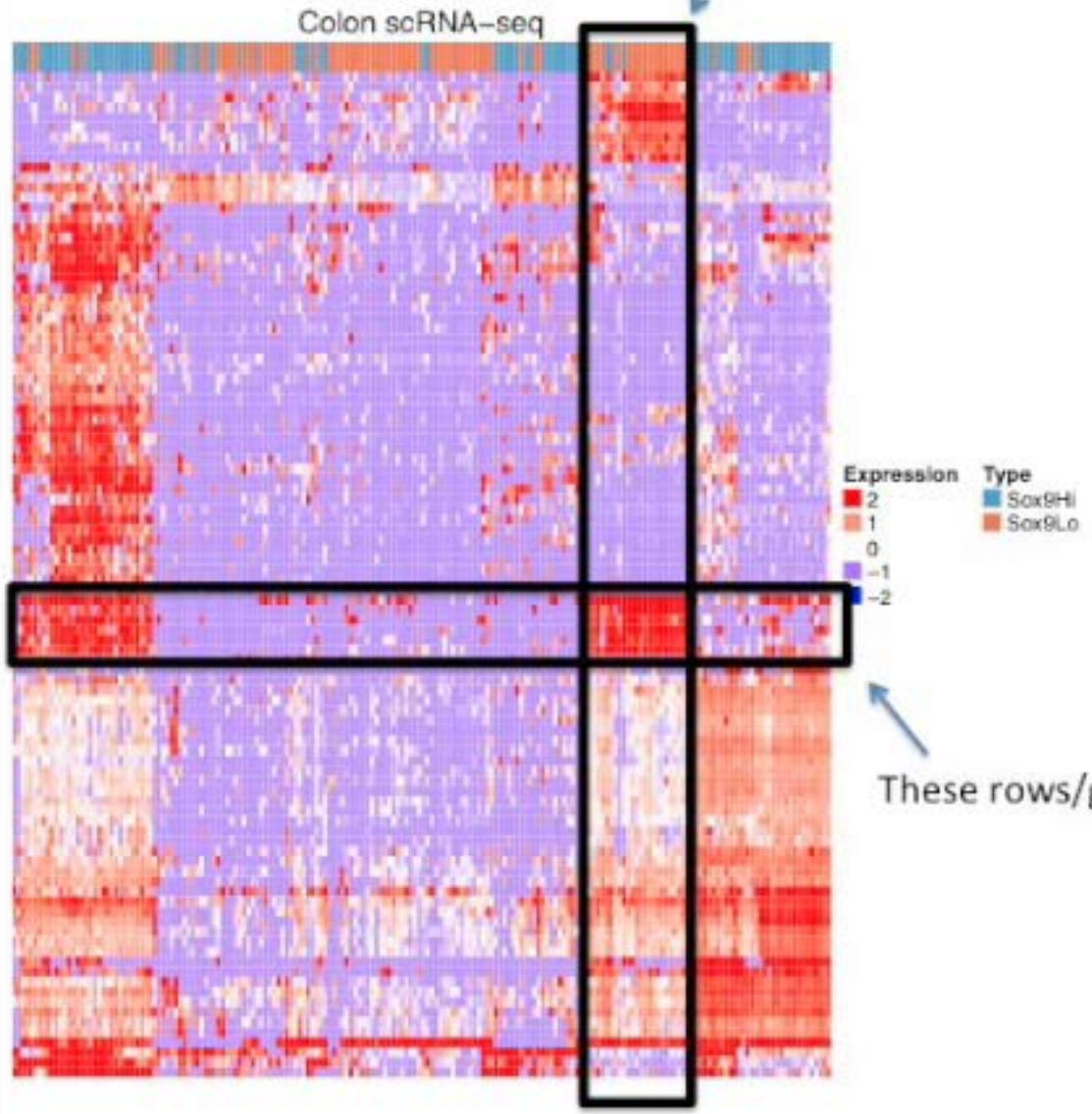


This heatmap has been **scaled** and **clustered**.

The **scaling** is "global" – not per row/gene – but for all rows/genes.

The **clustering** is by **column**/sample **AND** by **row**/gene.

These columns/samples cluster together.

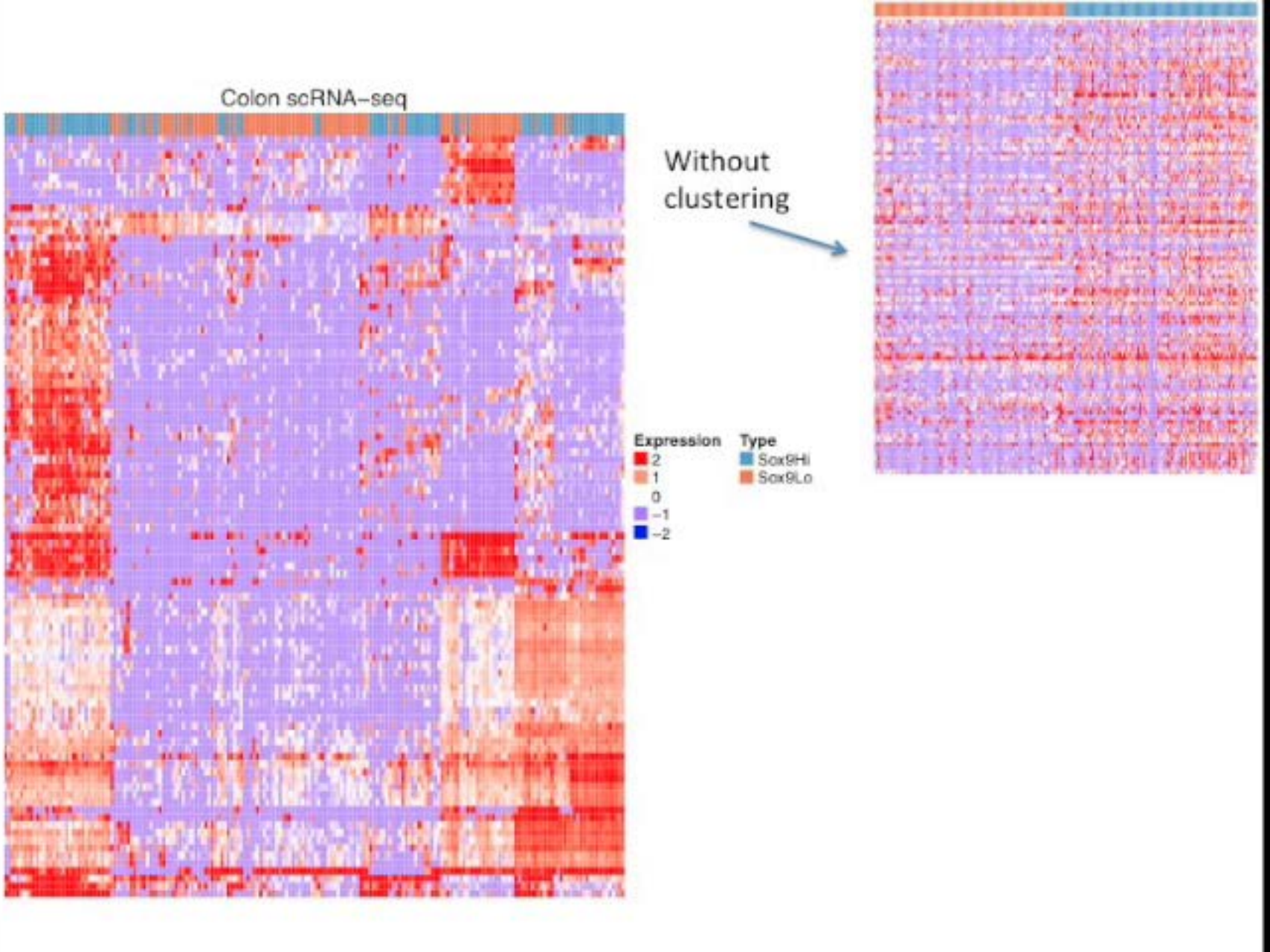


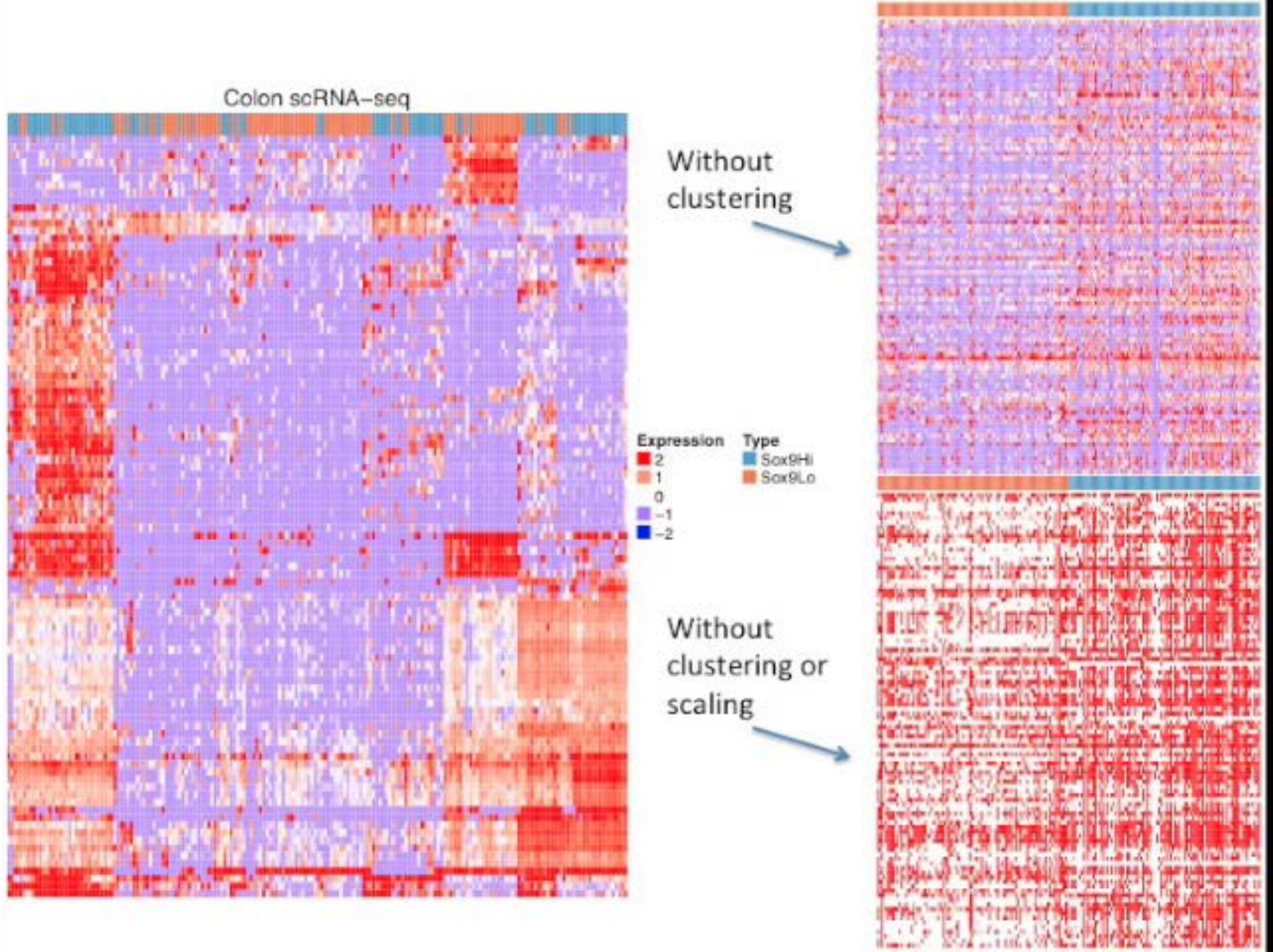
This heatmap has been **scaled** and **clustered**.

The **scaling** is "global" – not per row/gene – but for all rows/genes.

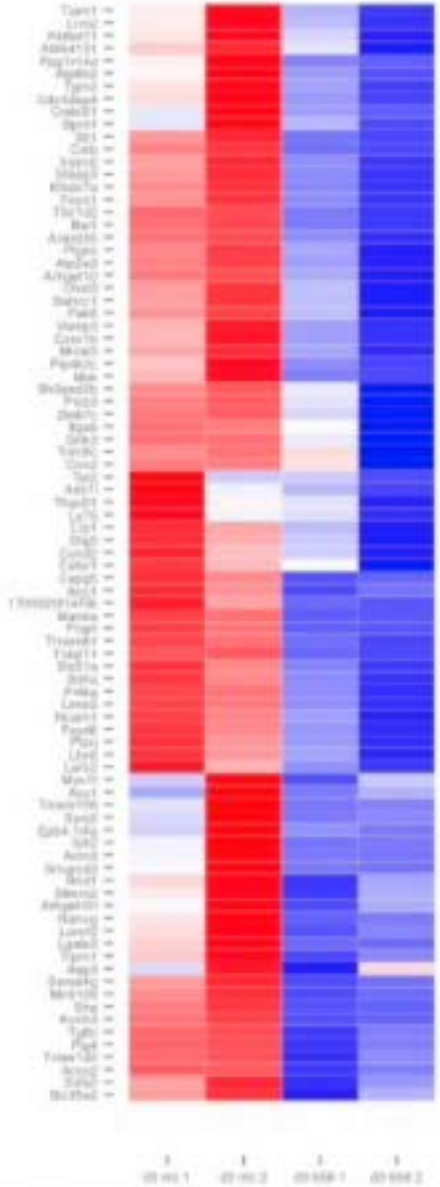
The **clustering** is by **column**/sample **AND** by **row**/gene.

These rows/genes cluster together.

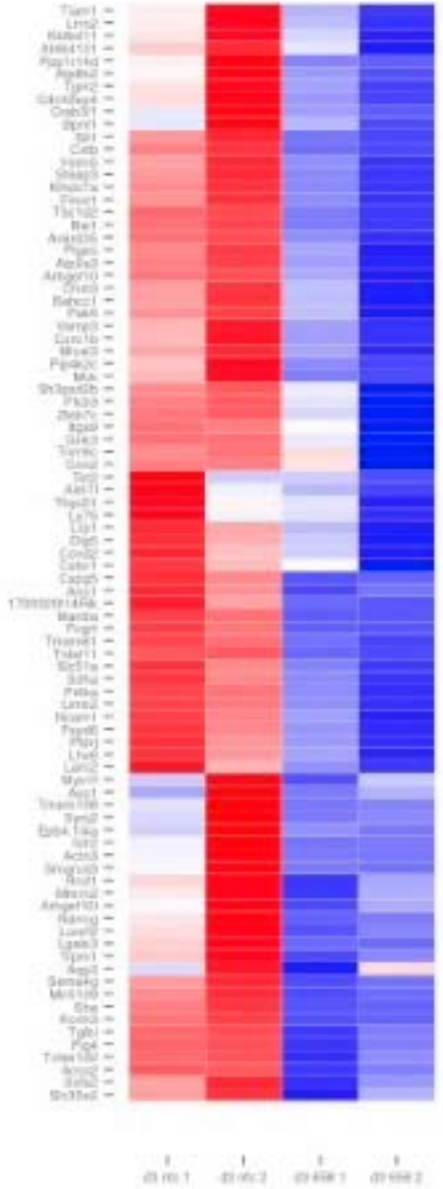

Colon scRNA-seq
Without clustering
Expression
2
1
0
−1
−2
Type
Sox9Hi
Sox9Lo


Colon scRNA–seq
Expression
2
1
0
−1
−2
Type
Sox9Hi
Sox9Lo
Without clustering
Without clustering or scaling

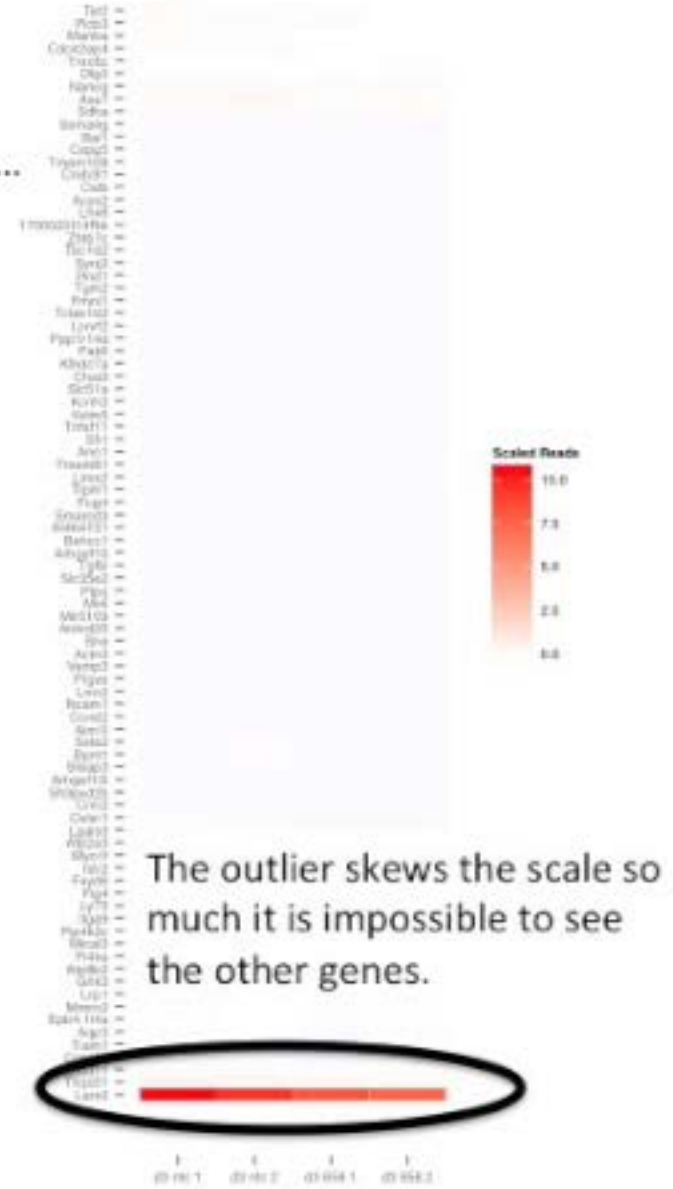
A quick aside....



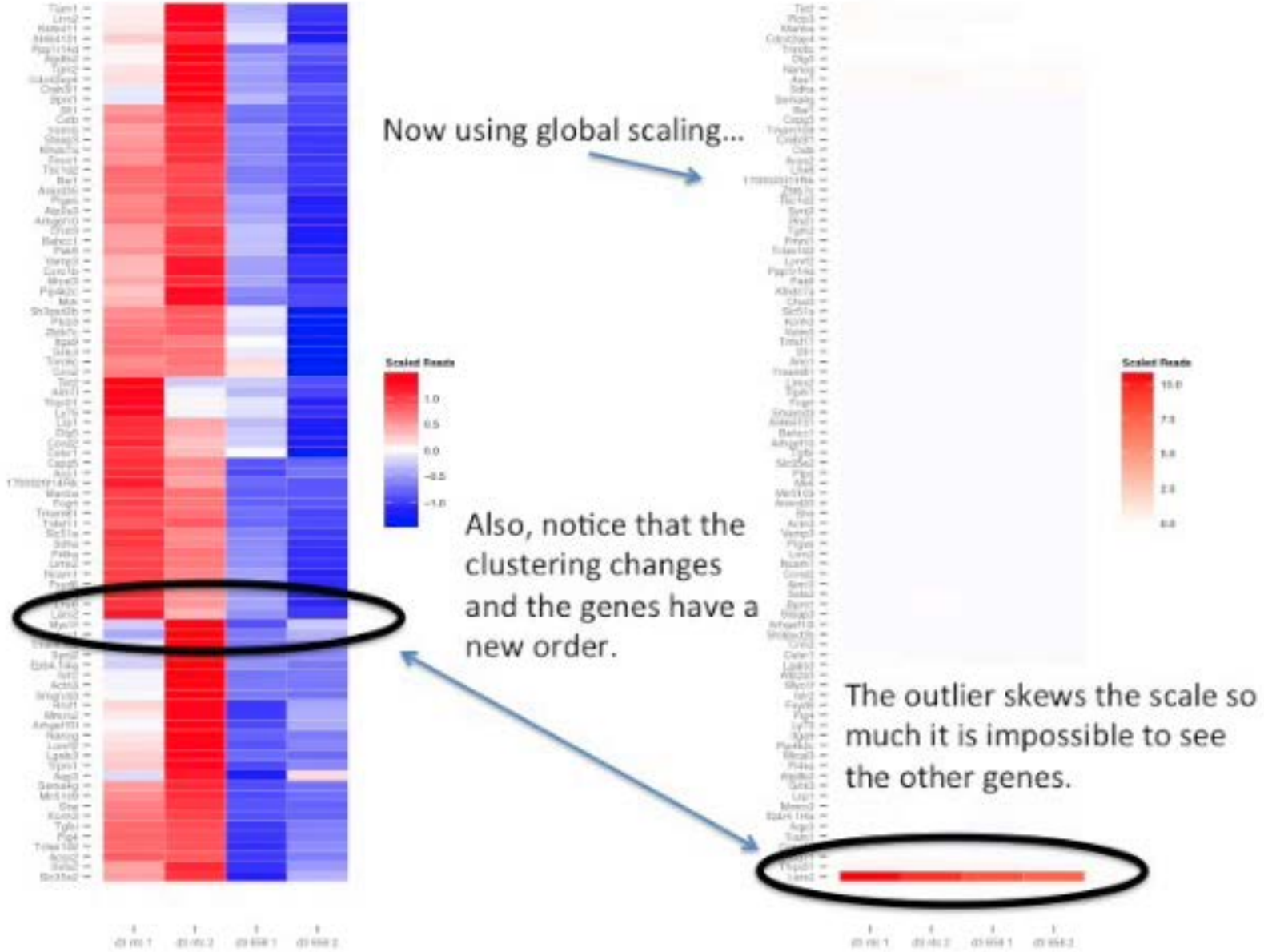
What if we had used global scaling with the first heatmap?



Now using global scaling…



The outlier skews the scale so much it is impossible to see the other genes.


Now using global scaling…
Also, notice that the clustering changes and the genes have a new order.
The outlier skews the scale so much it is impossible to see the other genes.


Scaling can affect two things:
1) How brightly colored the genes are and whether you can compare between them.
2) The clustering.
Scaled Reads
1.0
0.5
0.0
−0.5
−1.0
Also, notice that the clustering changes and the genes have a new order.
Scaled Reads
The outlier skews the scale so much it is impossible to see the other genes.

... now back to the action.

How to scale data…

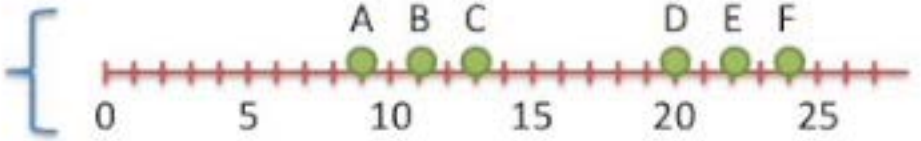
- Regardless of whether you do it by gene or globally, the most common method is…

nameless!

I hate to coin a new term, but let's call it "Z-Score Scaling" because, technically, it converts the data to "Z-scores"

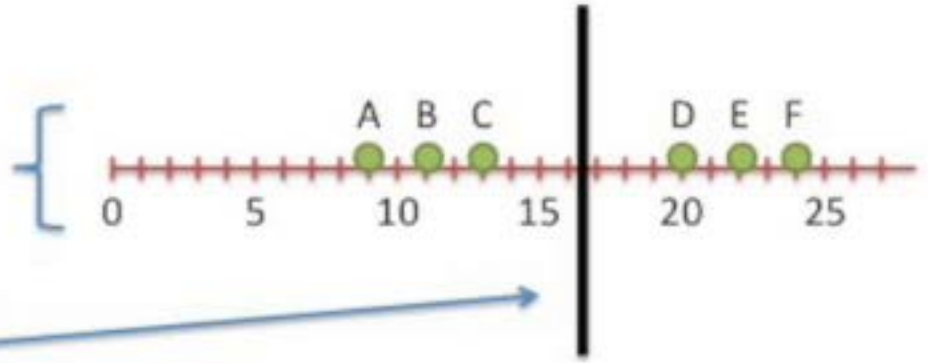
Converting to Z-Scores (i.e. Z-score scaling)

RNA-seq read counts from 6 samples.



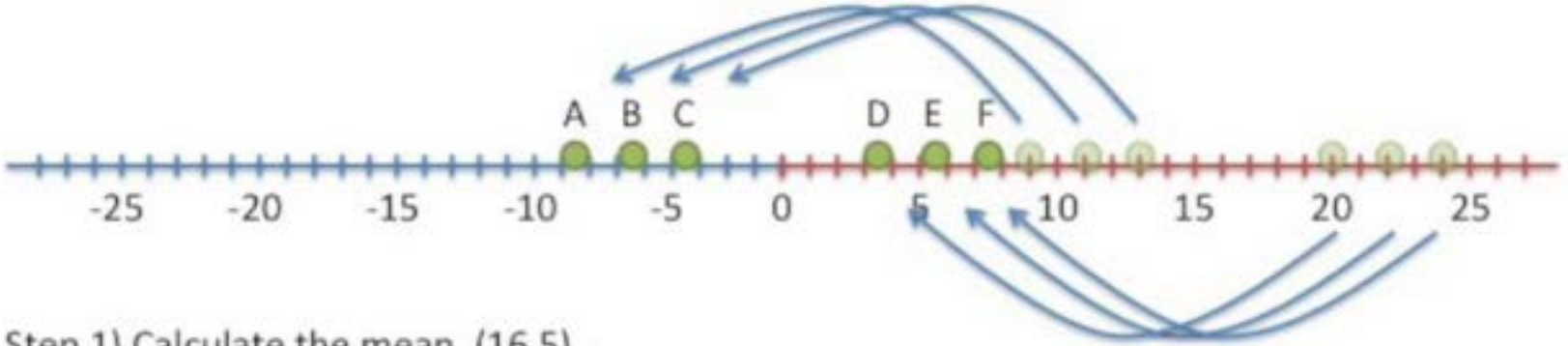
Converting to Z-Scores (i.e. Z-score scaling)

RNA-seq read counts from 6 samples.



Step 1) Calculate the mean (16.5)

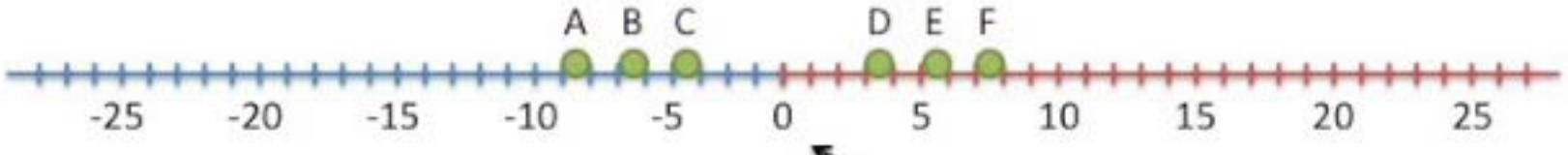
Converting to Z-Scores (i.e. Z-score scaling)



Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

Converting to Z-Scores (i.e. Z-score scaling)

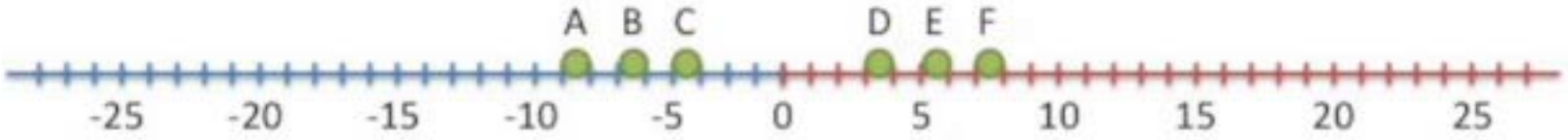


This centers the data around 0.

Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

Converting to Z-Scores (i.e. Z-score scaling)



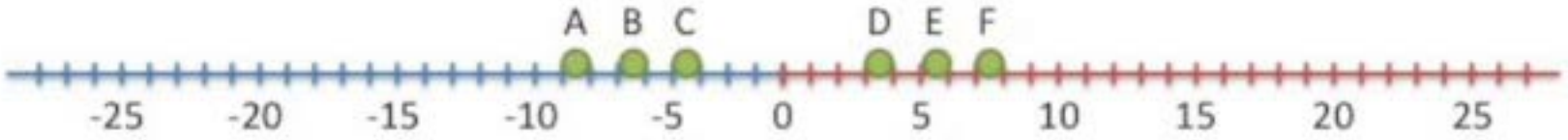
Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

This centers the data around 0.

Samples with relatively high transcription get **positive** values.

Converting to Z-Scores (i.e. Z-score scaling)



Step 1) Calculate the mean (16.5)

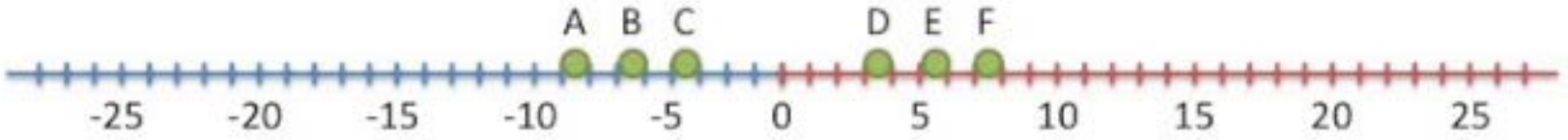
Step 2) Subtract the mean from each value

This centers the data around 0.

Samples with relatively high transcription get **positive** values.

Samples with relatively low transcription get **negative** values.

Converting to Z-Scores (i.e. Z-score scaling)

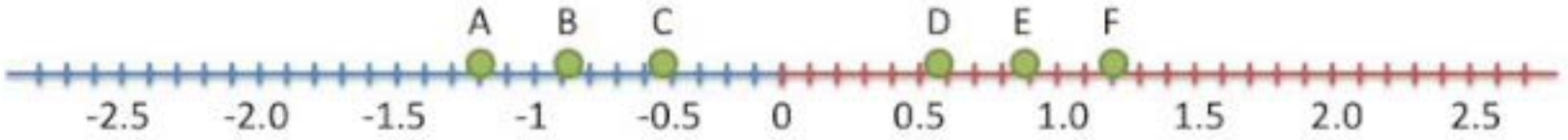


Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

Step 3) Calculate the standard deviation (6.28)

Converting to Z-Scores (i.e. Z-score scaling)



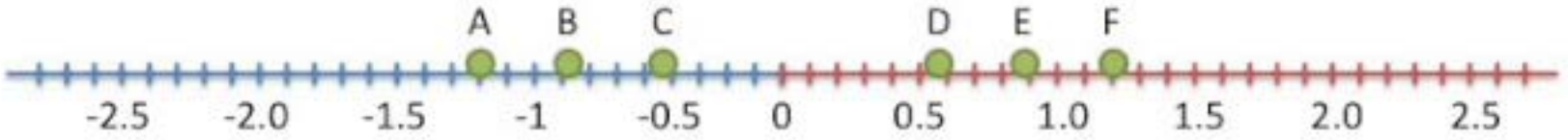
Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

Step 3) Calculate the standard deviation (6.28)

Step 4) Divide by the standard deviation (notice, the scale on the axis has changed)

Converting to Z-Scores (i.e. Z-score scaling)



Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

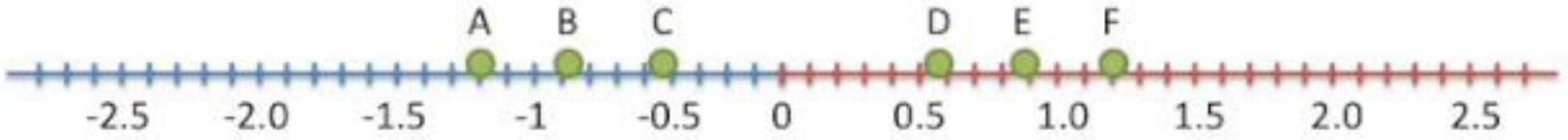
Step 3) Calculate the standard deviation (6.28)

Step 4) Divide by the standard deviation (notice, the scale on the axis has changed)

The data used to be spread from -8 to +8.

Now it is between -1.2 and 1.2

Converting to Z-Scores (i.e. Z-score scaling)



Step 1) Calculate the mean (16.5)

Step 2) Subtract the mean from each value

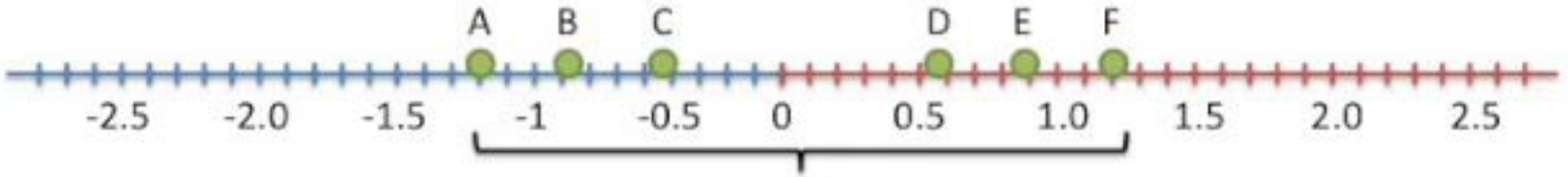
Step 3) Calculate the standard deviation (6.28)

Step 4) Divide by the standard deviation (notice, the scale on the axis has changed)

The formula for Z-score scaling

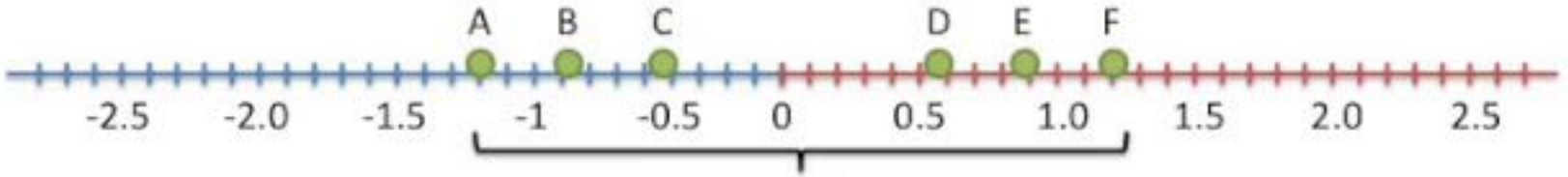
$$\frac{\text{sample value} - \text{the mean}}{\text{the standard deviation}} \quad \text{a.k.a.} \quad \frac{s_i - \mu}{\sigma}$$

Converting to Z-Scores (i.e. Z-score scaling)



Regardless of the variation in the original data, dividing by the standard deviation ensures that it's tightly grouped.

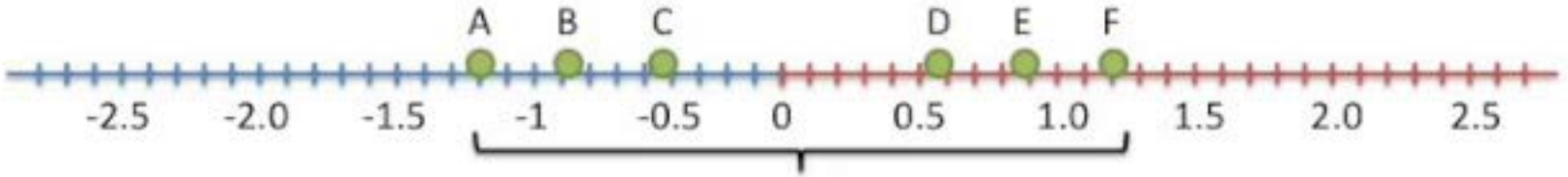
Converting to Z-Scores (i.e. Z-score scaling)



Why do we need to ensure that the data is tightly grouped?

Regardless of the variation in the original data, dividing by the standard deviation ensures that it's tightly grouped.

Converting to Z-Scores (i.e. Z-score scaling)



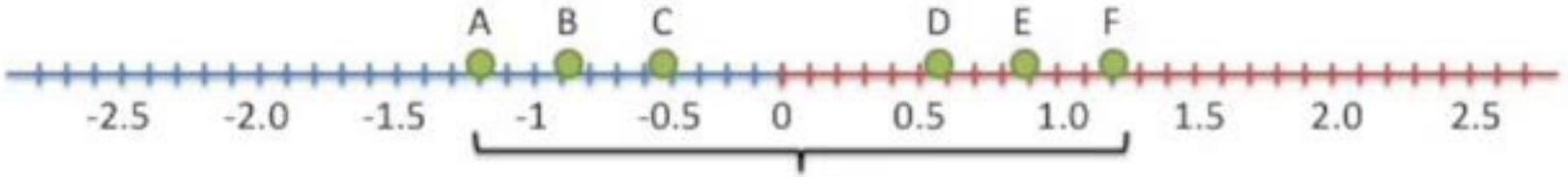
Why do we need to ensure that the data is tightly grouped?

Regardless of the variation in the original data, dividing by the standard deviation ensures that it's tightly grouped.

Because we can only discern so many shades of colors.

The wider the range, the more subtle the difference in the shades.

Converting to Z-Scores (i.e. Z-score scaling)



Why do we need to ensure that the data is tightly grouped?

Regardless of the variation in the original data, dividing by the standard deviation ensures that it's tightly grouped.

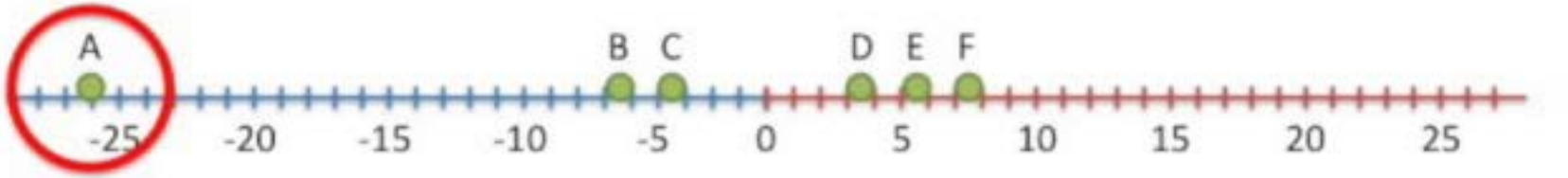
Because we can only discern so many shades of colors.

The wider the range, the more subtle the difference in the shades.

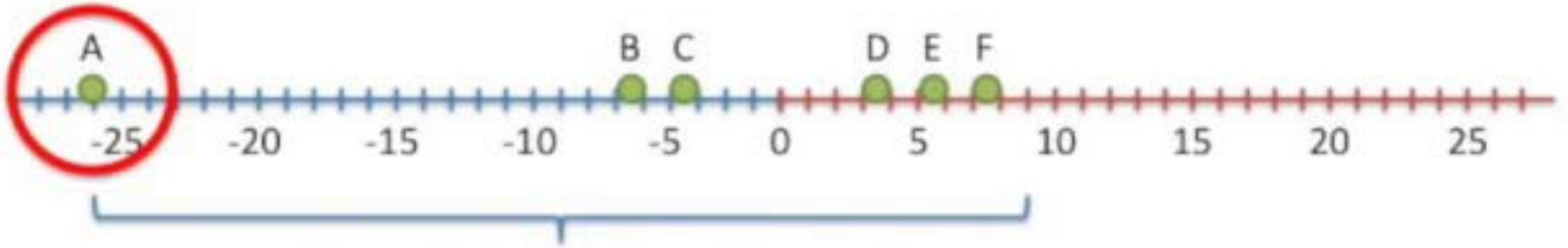
By tightly grouping the data, we use fewer shades and it is easier to see, "Sample 1 has more transcription than Sample 2..."

A brief aside... What if there is an outlier?

A brief aside... What if there is an outlier?

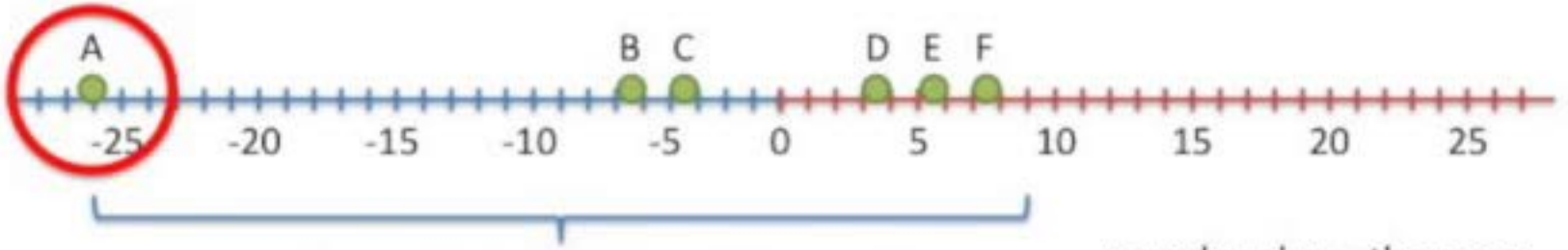


A brief aside... What if there is an outlier?



The standard deviation will be much larger.

A brief aside... What if there is an outlier?

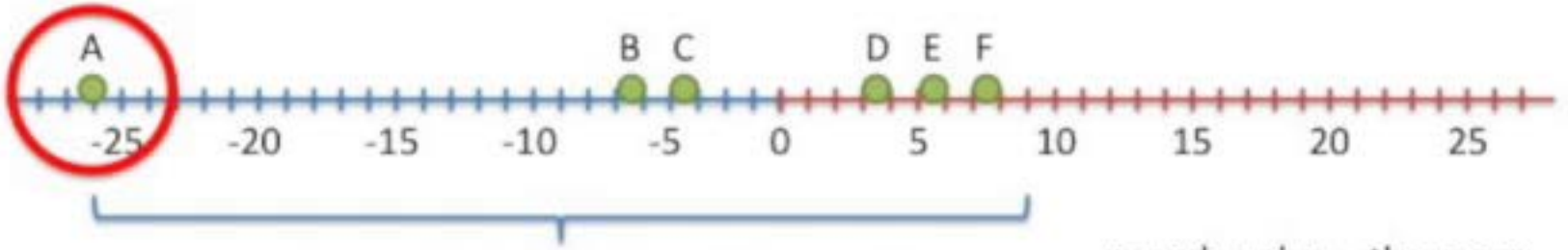


The standard deviation will be much larger.

$$\frac{\text{sample value} - \text{the mean}}{\text{the standard deviation}}$$

That is to say, the denominator will be larger.

A brief aside... What if there is an outlier?

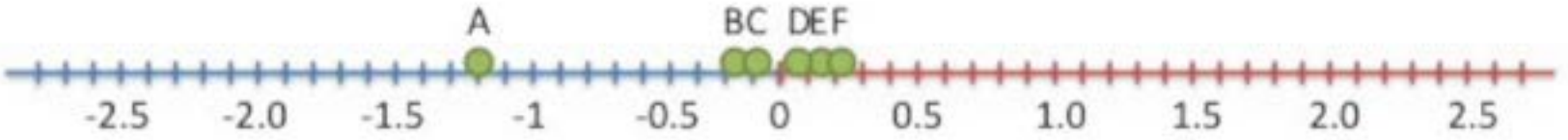


The standard deviation will be much larger.

$$\frac{\text{sample value} - \text{the mean}}{\text{the standard deviation}}$$

That is to say, the denominator will be larger.

And the values near zero will get compressed a lot and it will be hard to separate them with only a few shades.



When we did “global scaling” on the dataset with the outlier, we saw what happens with an outlier.

One gene is clearly highly expressed, but we can’t see any differences in the other genes.

Clustering – The fun part!

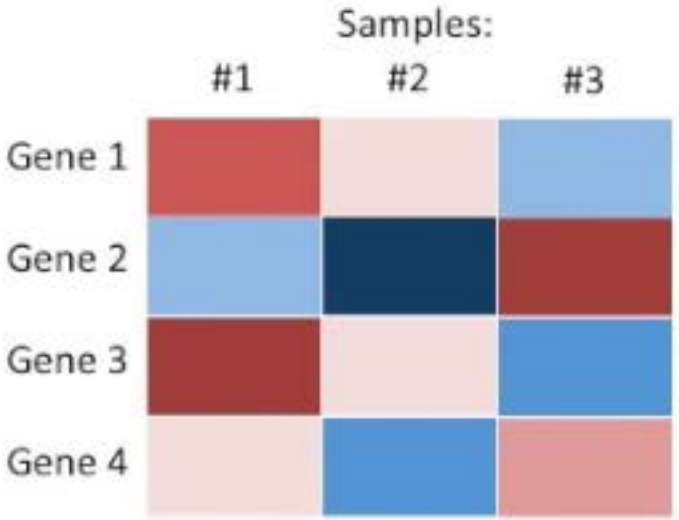
Clustering – The fun part!

- There are two main types of clustering:
  - Hierarchical
  - K-means

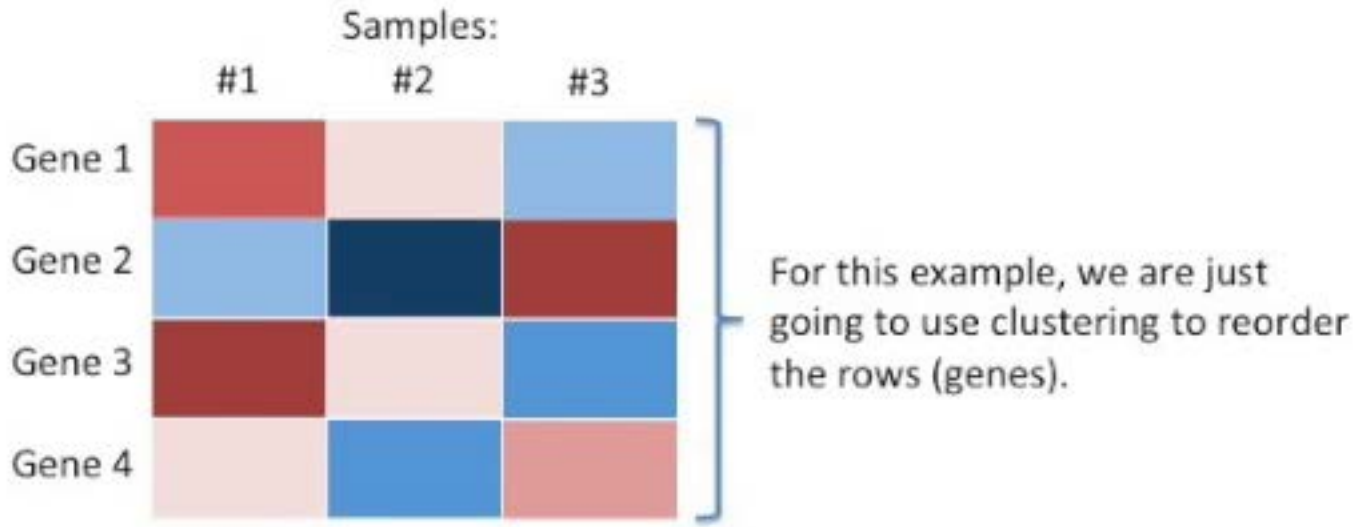
Clustering – The fun part!

- There are two main types of clustering:
  - Hierarchical
  - K-means
- We'll focus on hierarchical clustering for now…

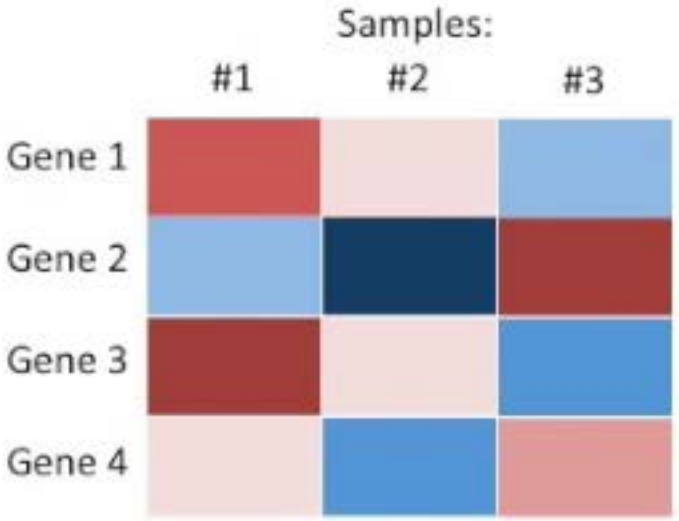
Hierarchical Clustering



Hierarchical Clustering



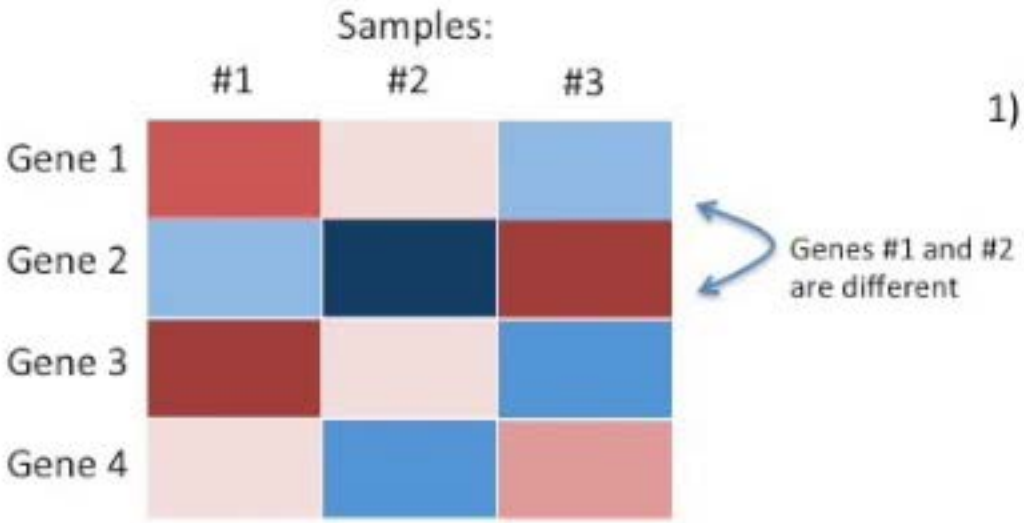
Hierarchical Clustering



Conceptually…

1) Figure out which gene is most similar to gene #1.

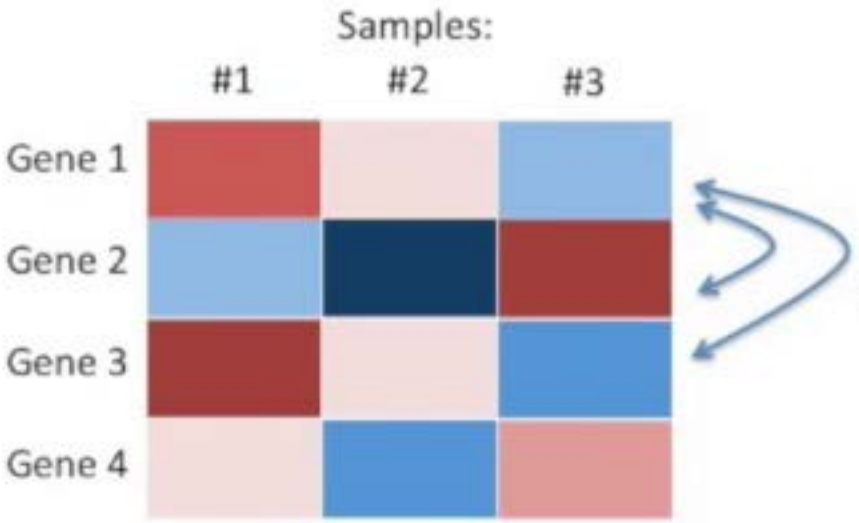
Hierarchical Clustering



Conceptually…

1) Figure out which gene is most similar to gene #1.

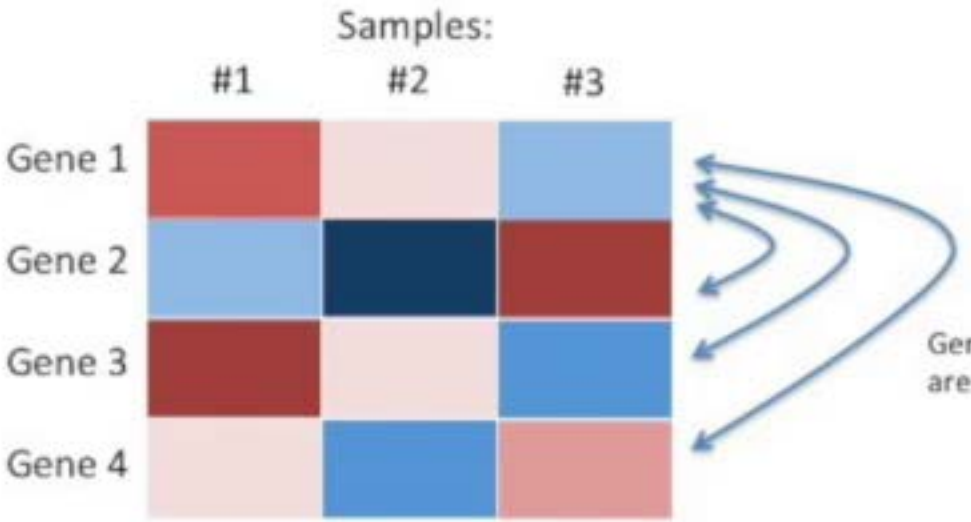
Hierarchical Clustering



Conceptually…

1) Figure out which gene is most similar to gene #1.

Hierarchical Clustering

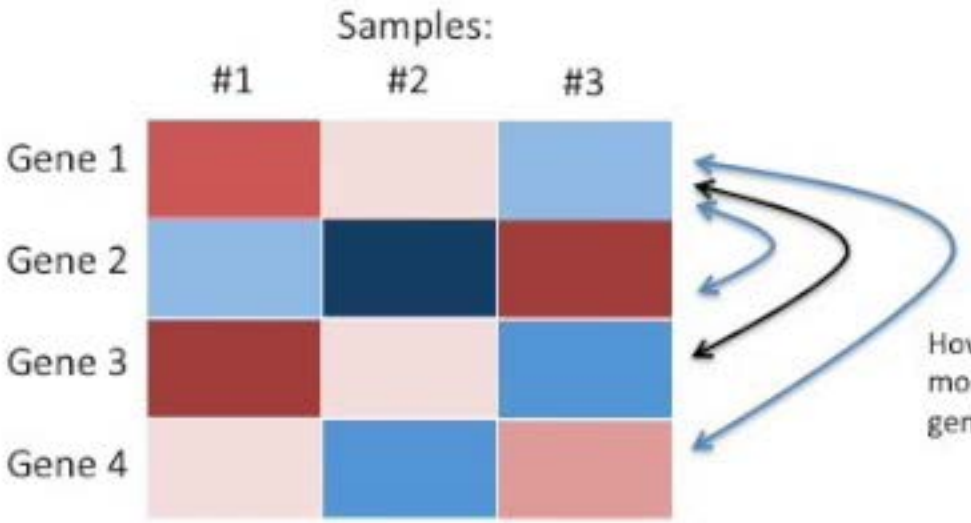


Conceptually…

1) Figure out which gene is most similar to gene #1.

Genes #1 and #4 are similar.

Hierarchical Clustering

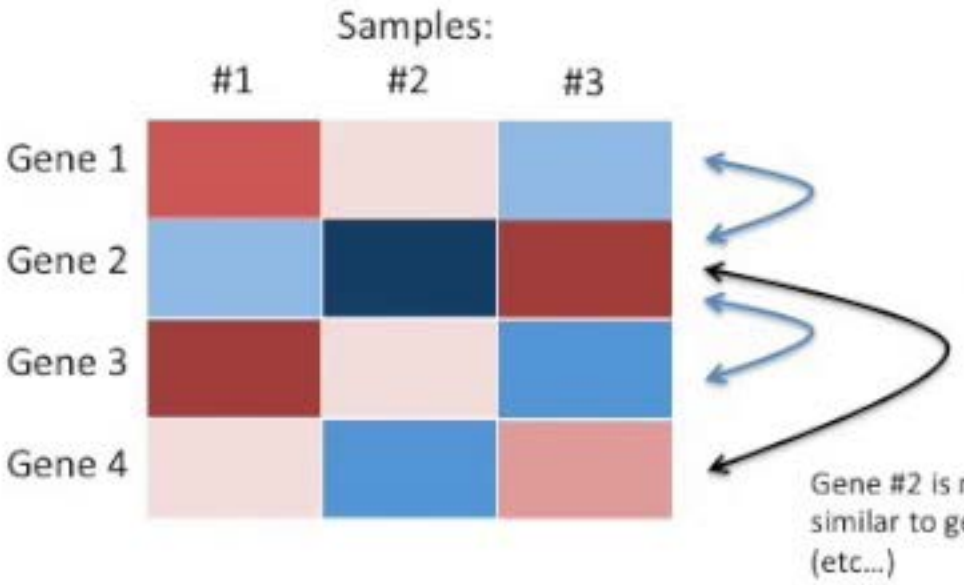


Conceptually…

1) Figure out which gene is most similar to gene #1.

However, gene #1 is most similar to gene #3.

Hierarchical Clustering

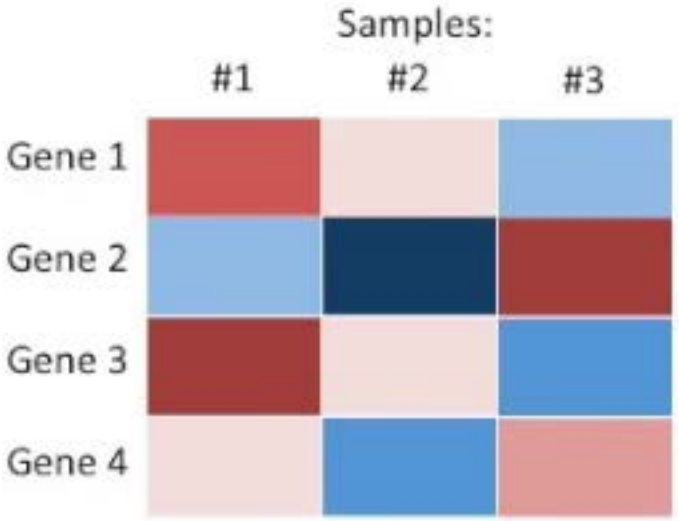


Conceptually...

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2... (and then #3 and then #4).

Hierarchical Clustering



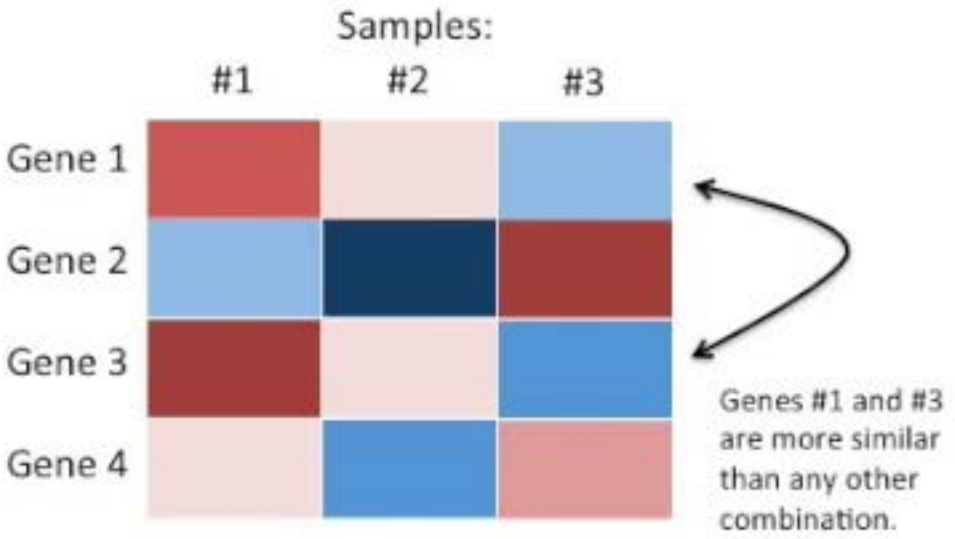
Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge them into a cluster.

Hierarchical Clustering



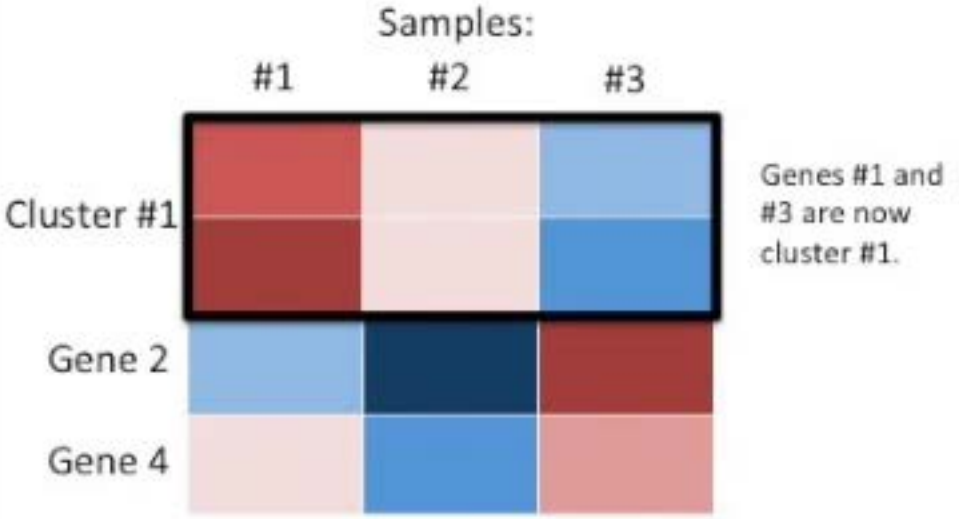
Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge them into a cluster.

Hierarchical Clustering



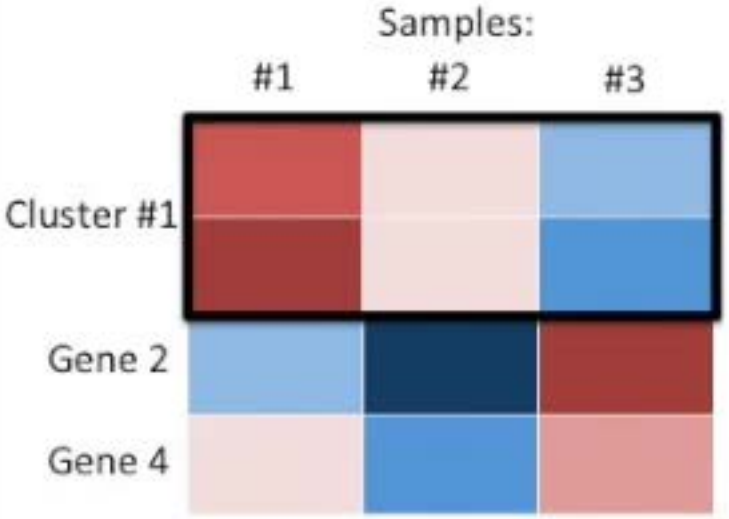
Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge them into a cluster.

Hierarchical Clustering



Conceptually…

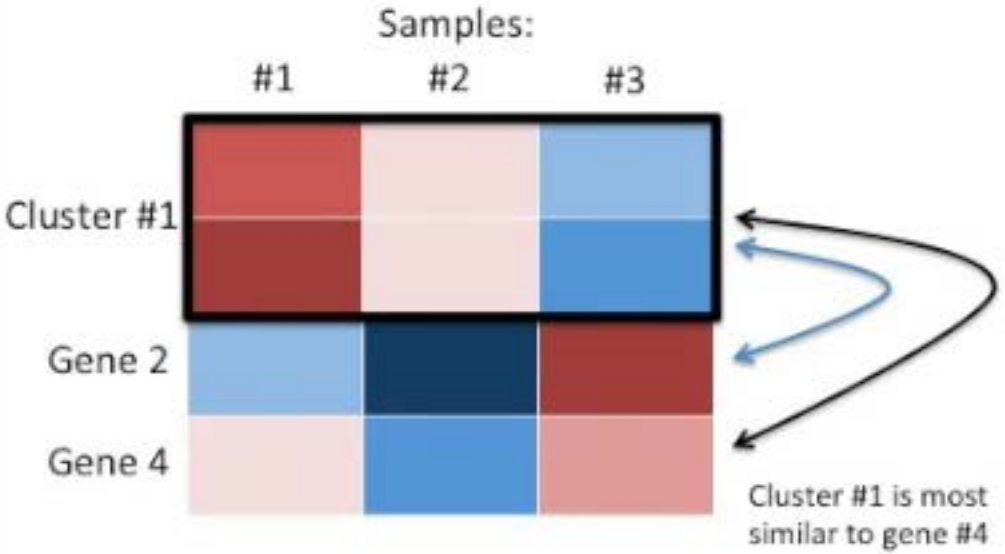
1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge those into a cluster.

4) Go back to step 1, but now treat the new cluster like it's a single gene.

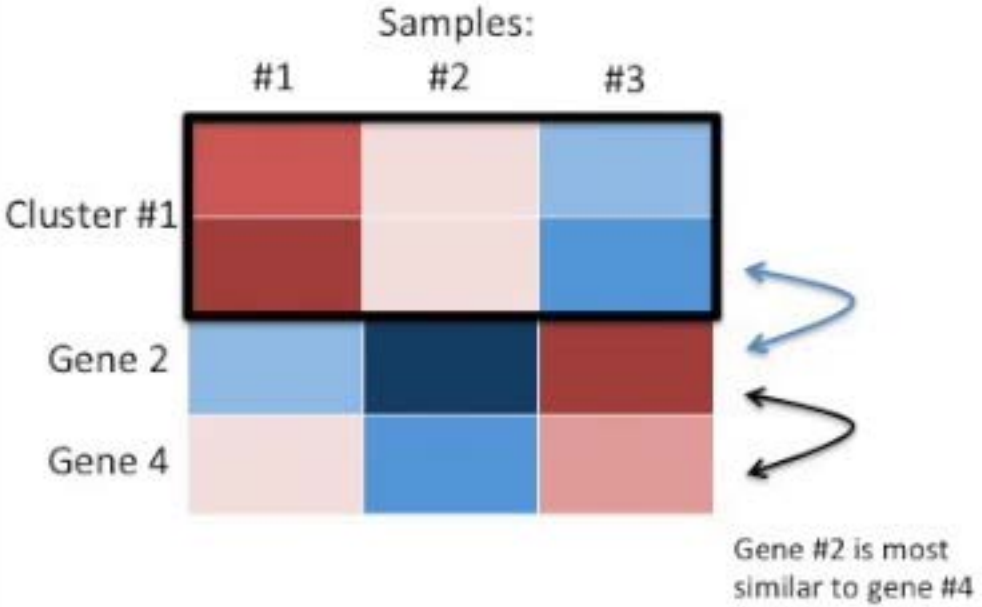
Hierarchical Clustering



Conceptually…

1) Figure out which gene is most similar to gene #1/cluster #1
2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).
3) Of the different combinations, figures out which two genes are the most similar. Merge those into a cluster.
4) Go back to step 1, but now treat the new cluster like it's a single gene.

Hierarchical Clustering



Conceptually…

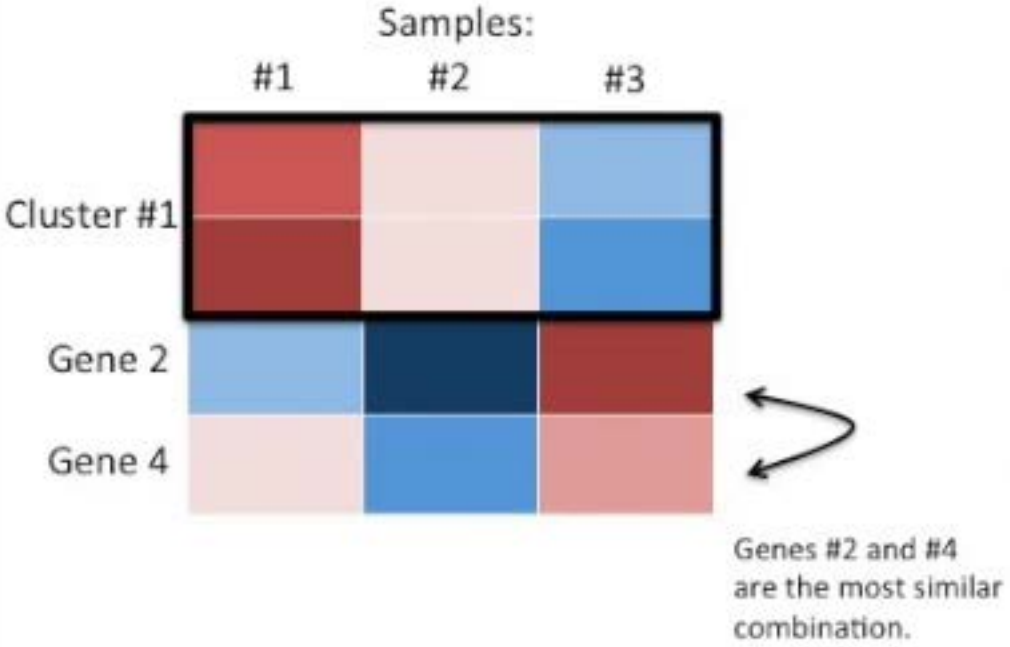
1) Figure out which gene is most similar to gene #1/cluster #1

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge those into a cluster.

4) Go back to step 1, but now treat the new cluster like it's a single gene.

Hierarchical Clustering



Conceptually…

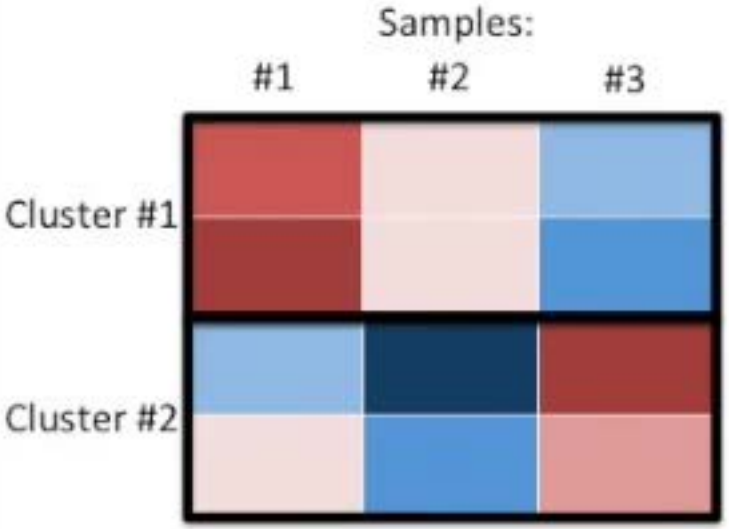
1) Figure out which gene is most similar to gene #1/cluster #1

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge those into a cluster.

4) Go back to step 1, but now treat the new cluster like it's a single gene.

Hierarchical Clustering



Done!

Conceptually…

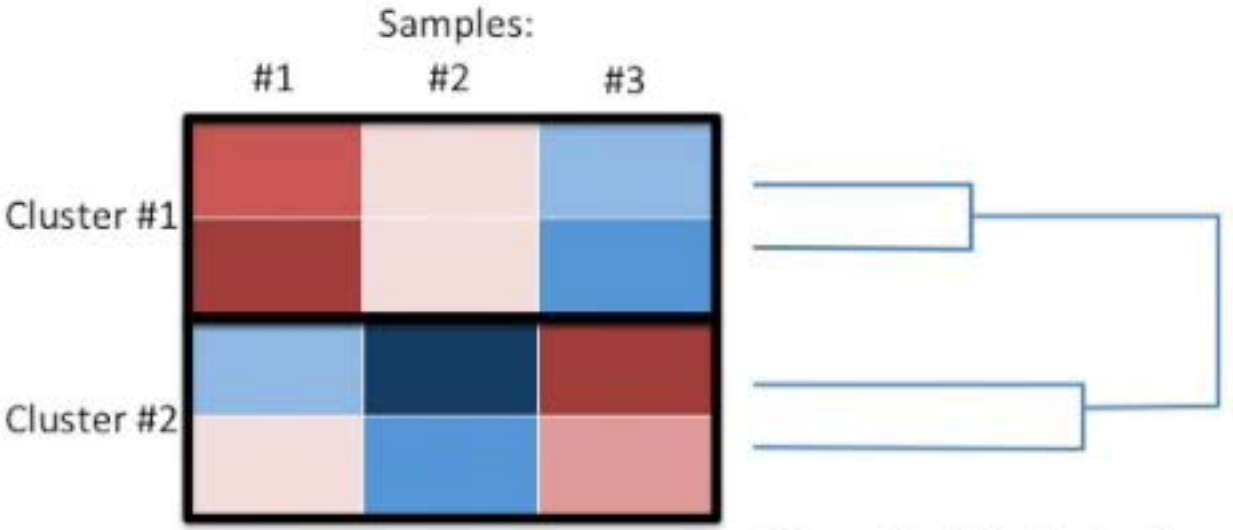
1) Figure out which gene is most similar to gene #1/cluster #1

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figures out which two genes are the most similar. Merge those into a cluster.

4) Go back to step 1, but now treat the new cluster like it's a single gene.

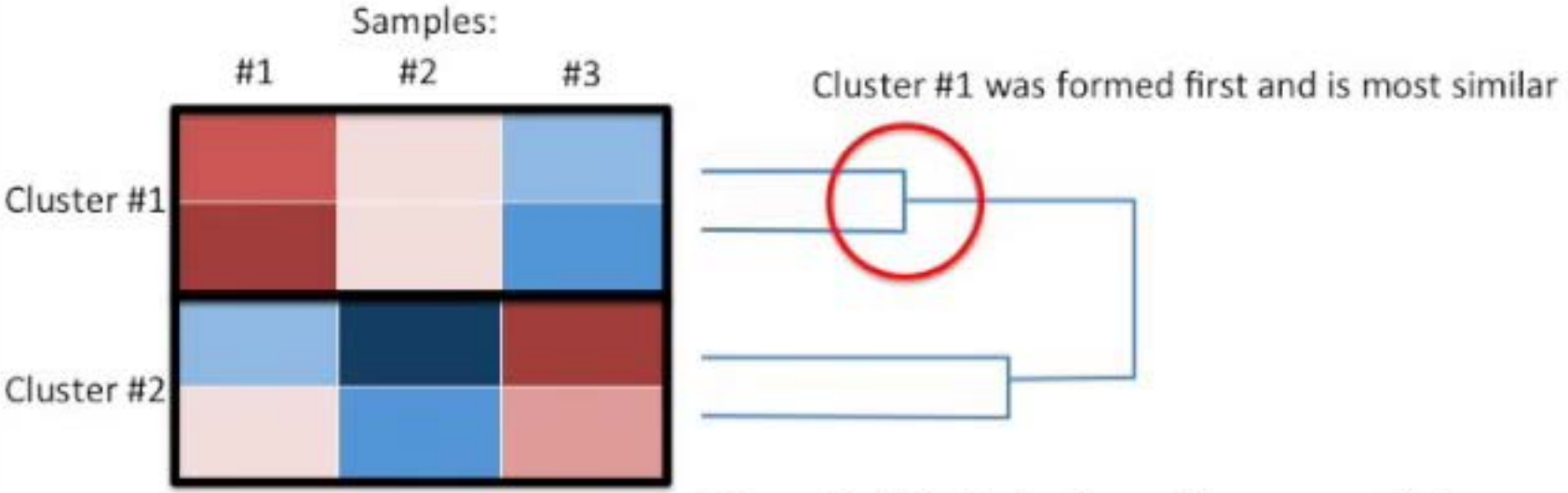
Hierarchical Clustering



Hierarchical clustering is usually accompanied by a "dendrogram".

It indicates both the similarity and the order that the clusters were formed.

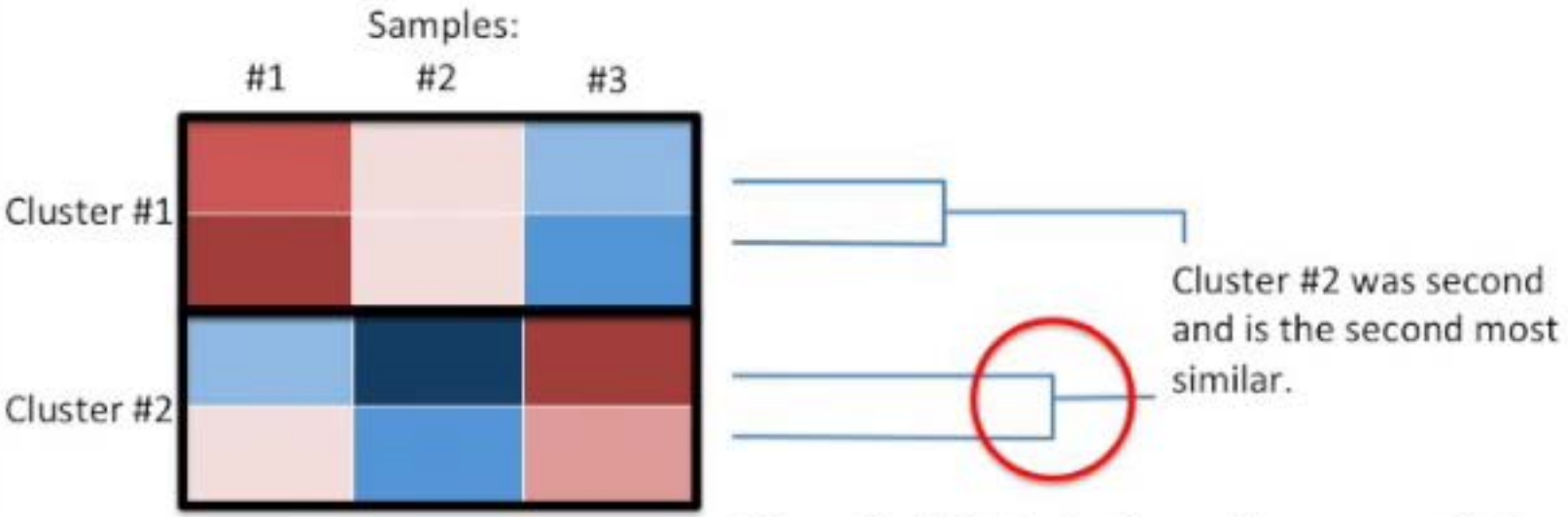
Hierarchical Clustering



Hierarchical clustering is usually accompanied by a "dendrogram".

It indicates both the similarity and the order that the clusters were formed.

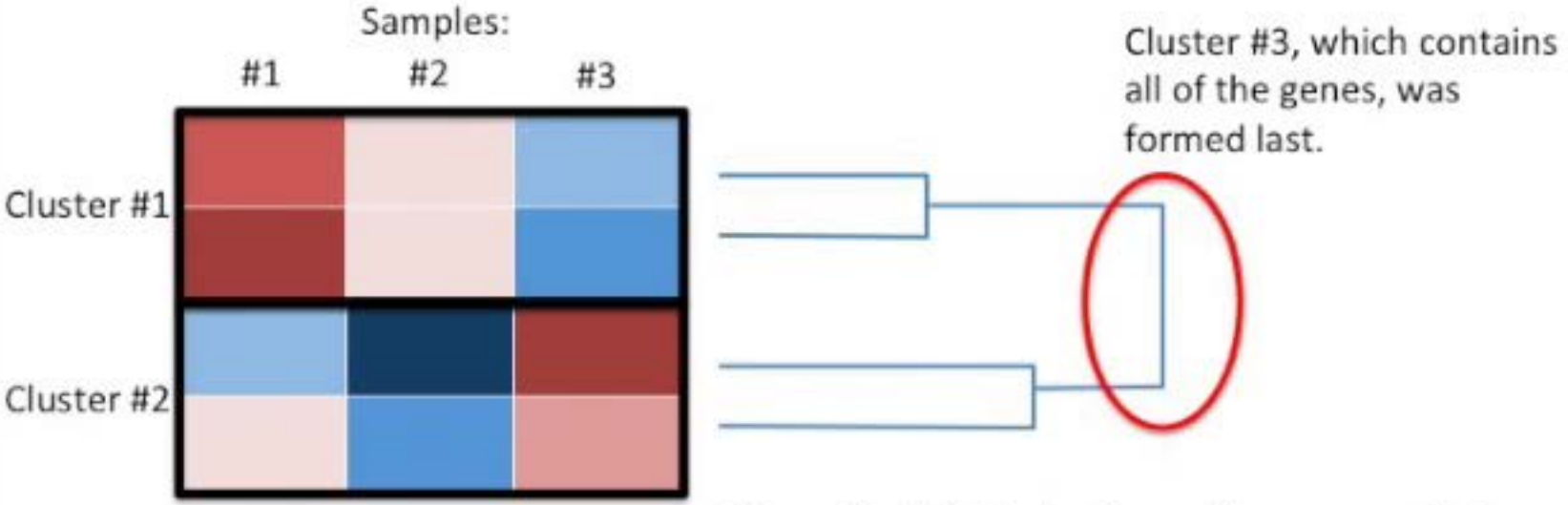
Hierarchical Clustering



Hierarchical clustering is usually accompanied by a "dendrogram".

It indicates both the similarity and the order that the clusters were formed.

Hierarchical Clustering



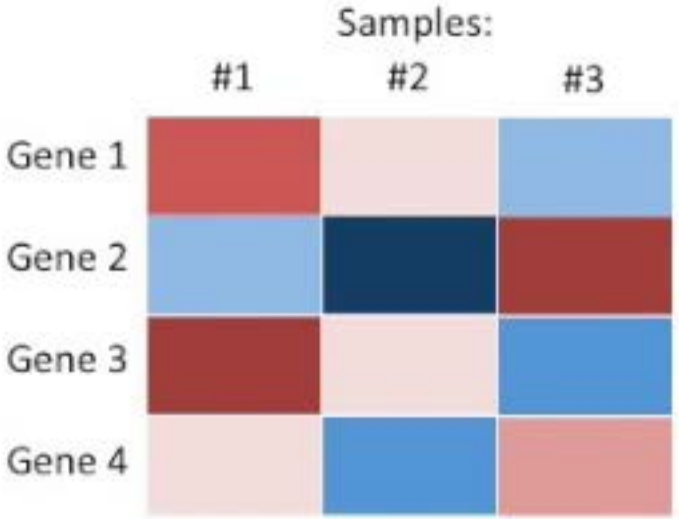
Cluster #3, which contains all of the genes, was formed last.

Hierarchical clustering is usually accompanied by a "dendrogram".

It indicates both the similarity and the order that the clusters were formed.

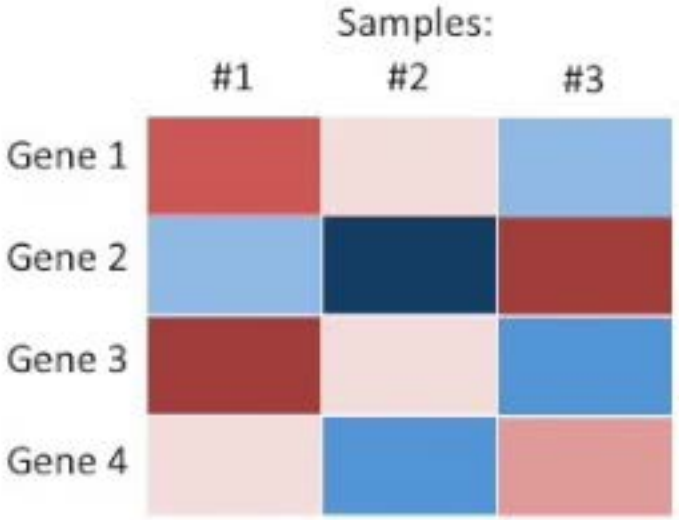
Hierarchical Clustering – a few nit-picky details

Hierarchical Clustering – a few nit-picky details



1) Figure out which gene is **most similar** to gene #1.

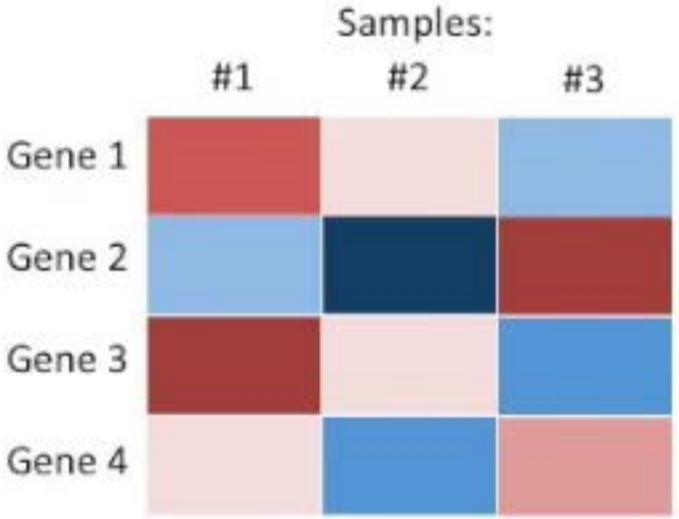
Hierarchical Clustering – a few nit-picky details



1) Figure out which gene is **most similar** to gene #1.

We have to define what "**most similar**" means!

Hierarchical Clustering – a few nit-picky details

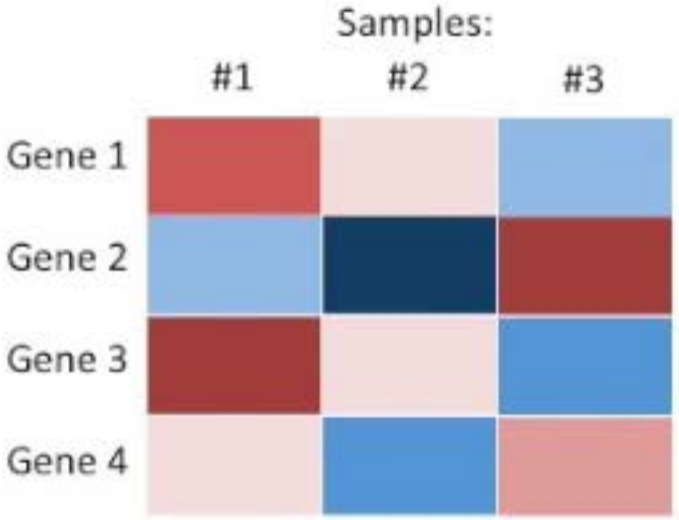


1) Figure out which gene is **most similar** to gene #1.

The method for determining similarity is arbitrarily chosen. However, there are some common practices.

1) Euclidian distance between genes:

Hierarchical Clustering – a few nit-picky details



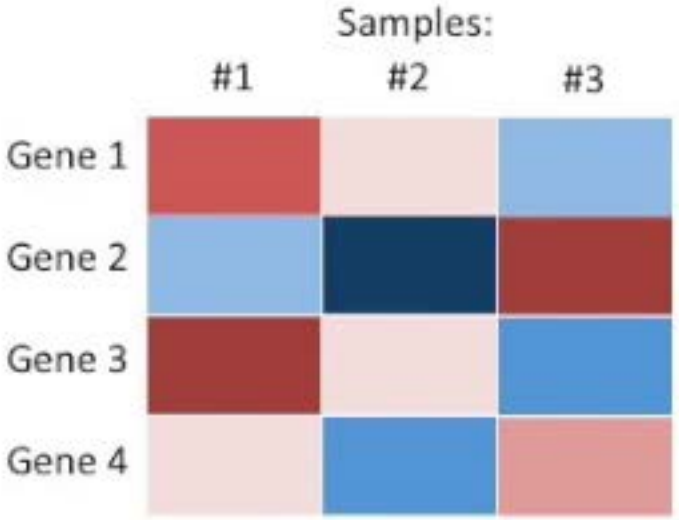
1) Figure out which gene is **most similar** to gene #1.

The method for determining similarity is arbitrarily chosen. However, there are some common practices.

1) Euclidian distance between genes:

$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2 + (\text{difference in sample...})^2}$$

Hierarchical Clustering – a few nit-picky details



1) Figure out which gene is **most similar** to gene #1.

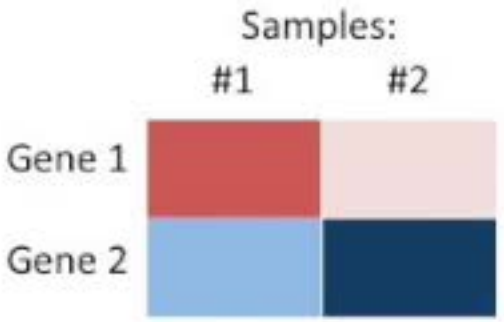
The method for determining similarity is arbitrarily chosen. However, there are some common practices.

1) Euclidian distance between genes:

$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2 + (\text{difference in sample...})^2}$$

To see the Euclidian distance in action, let's assume there are only two samples and two genes.

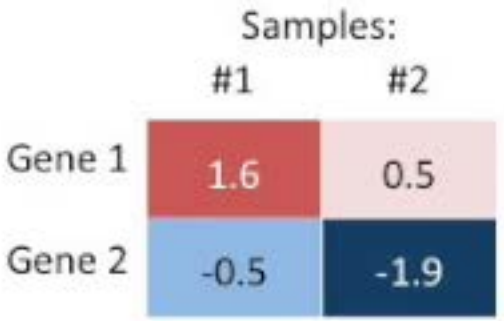
Hierarchical Clustering – a few nit-picky details



$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – a few nit-picky details



$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – a few nit-picky details

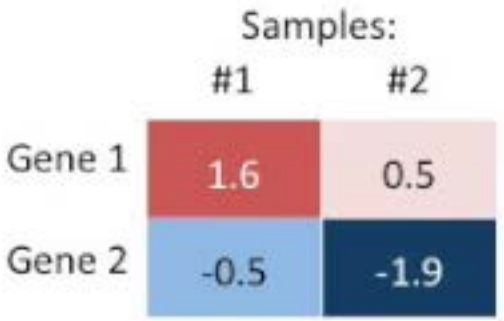
	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

$$\sqrt{(1.6-(-0.5))^2+(0.5-(-1.9))^2}$$

$$\sqrt{(\text{difference in sample \#1})^2+(\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – a few nit-picky details



$$\sqrt{(1.6-(-0.5))^2+(0.5-(-1.9))^2}$$

Sample #1: the difference between genes #1 and #2

$$\sqrt{(\text{difference in sample \#1})^2+(\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – a few nit-picky details

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

$$\sqrt{(1.6 - (-0.5))^2 + (0.5 - (-1.9))^2}$$

Sample #1: the difference between genes #1 and #2

Sample #2: The difference between genes #1 and #2

$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – a few nit-picky details

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

$$\sqrt{(1.6-(-0.5))^2+(0.5-(-1.9))^2}$$

$$\sqrt{(2.1)^2+(2.4)^2}$$

This is the "distance" between genes #1 and #2.

$$\sqrt{(\text{difference in sample \#1})^2+(\text{difference in sample \#2})^2}$$

You might recognize this as the Pythagorean Theorem.

Hierarchical Clustering – distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - Manhattan
  - Canberra
  - etc.

Hierarchical Clustering – distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - Manhattan
  - Canberra
  - etc.

For example, the Manhattan distance is just the absolute value of the differences….

|difference in sample #1|+ |difference in sample #2| + |difference in gene …|

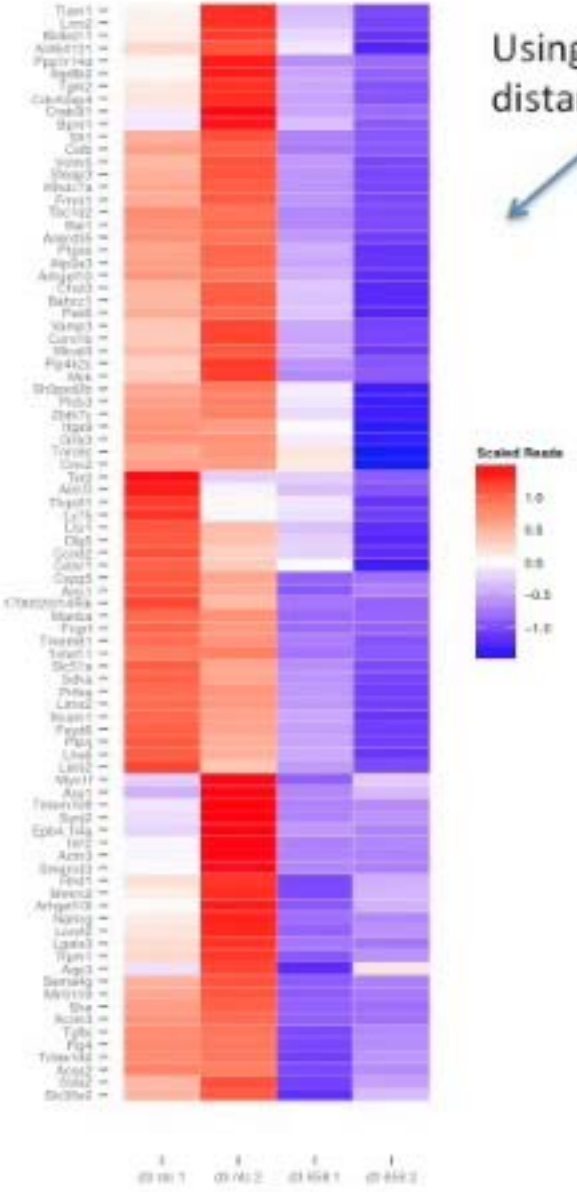
Hierarchical Clustering – distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - Manhattan
  - Canberra
  - etc.

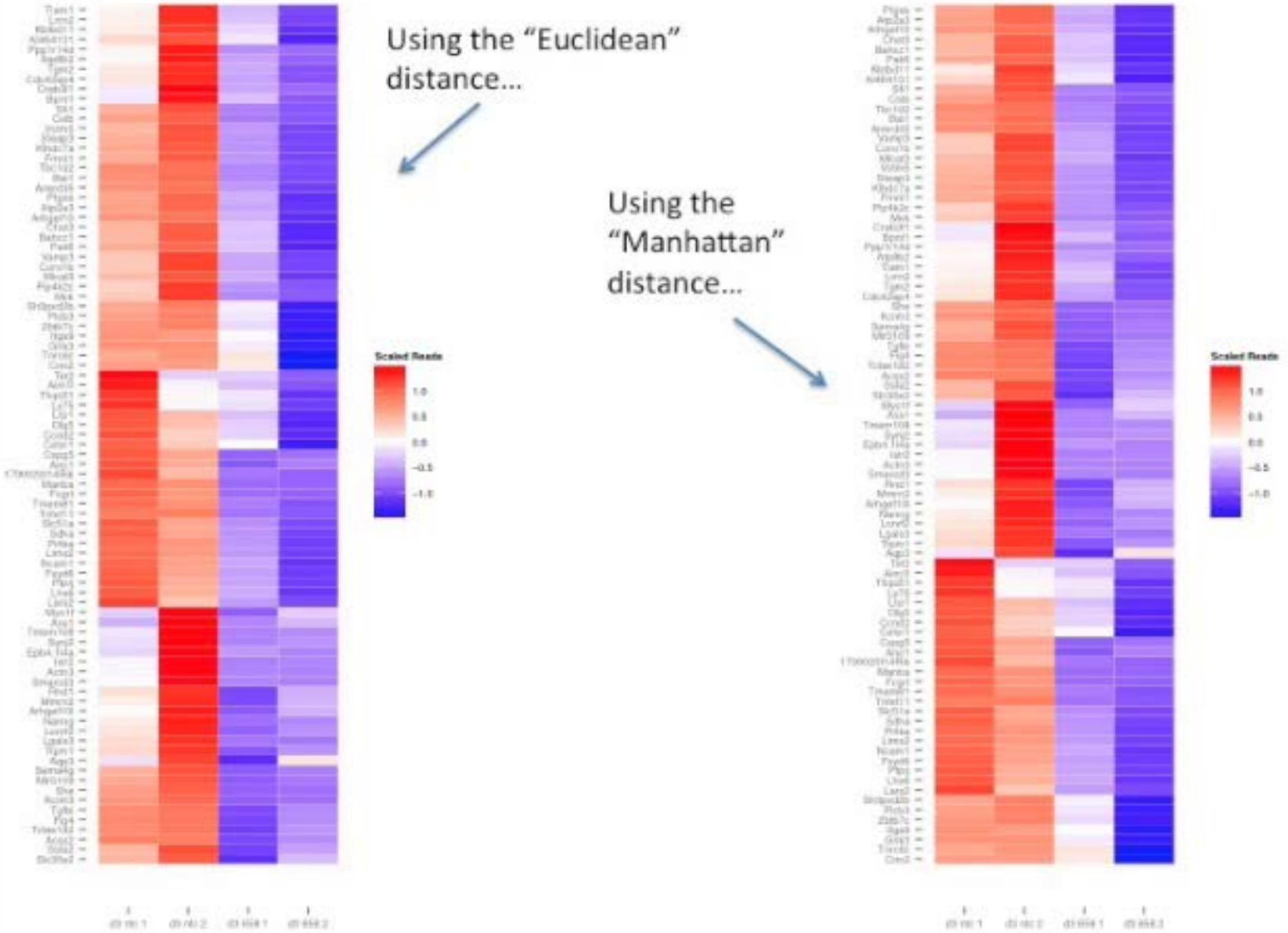
For example, the Manhattan distance is just the absolute value of the differences….

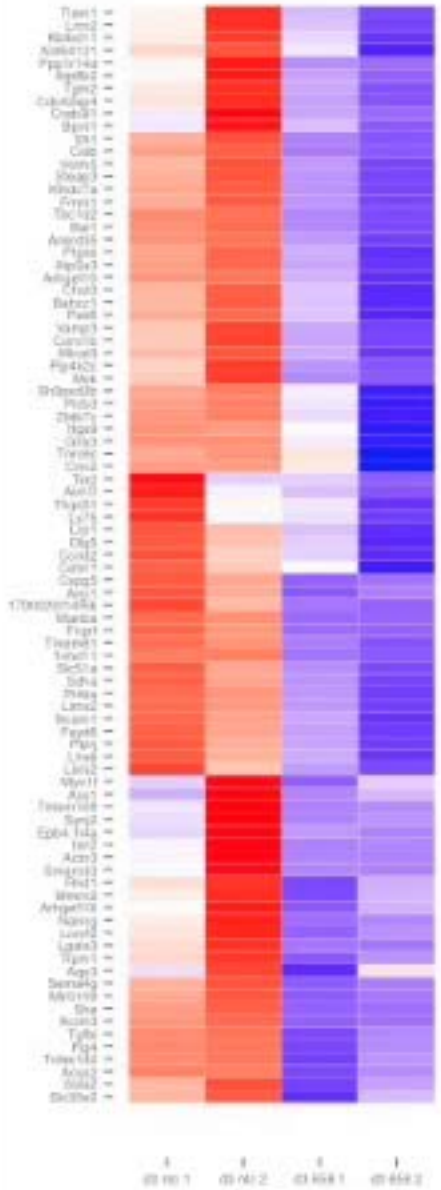
|difference in sample #1|+ |difference in sample #2| + |difference in gene …|

- Yes, it makes a difference.



Using the “Euclidean” distance…

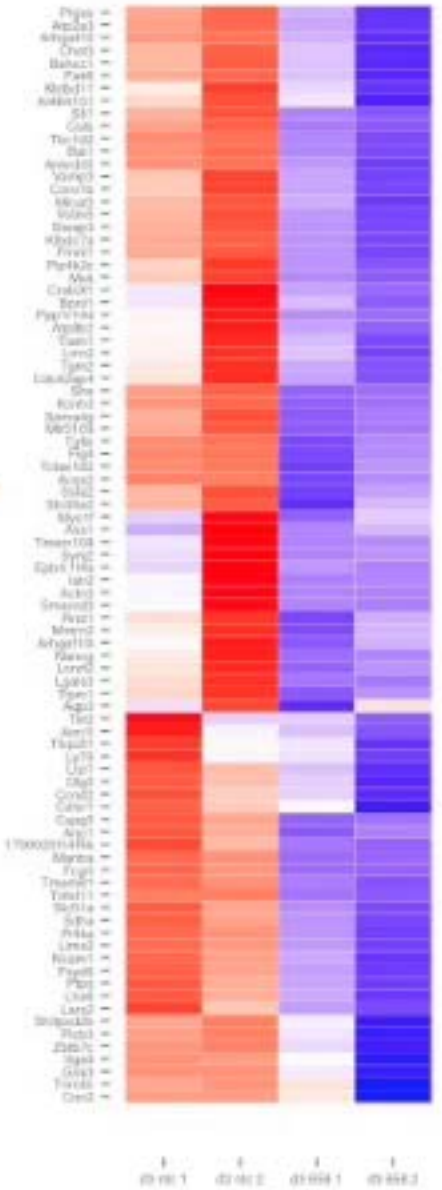

Using the "Euclidean" distance…
Using the "Manhattan" distance…
Scaled Reads
1.0
0.5
0.0
−0.5
−1.0



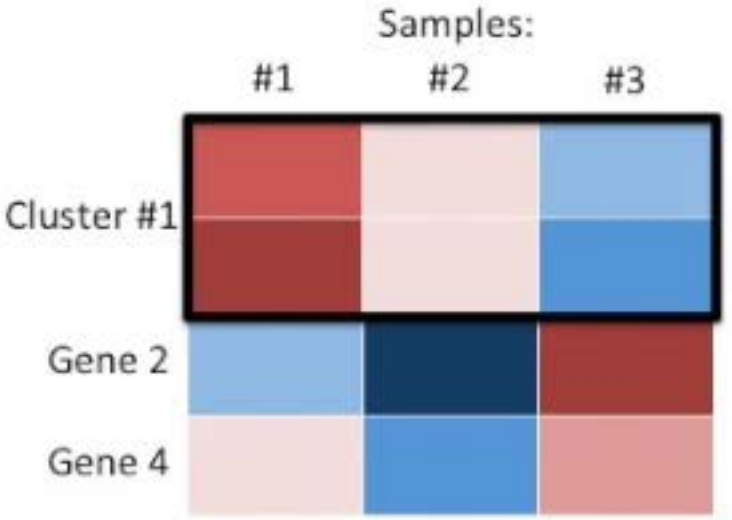
Using the "Euclidean" distance…

Using the "Manhattan" distance…

But the choice is arbitrary…

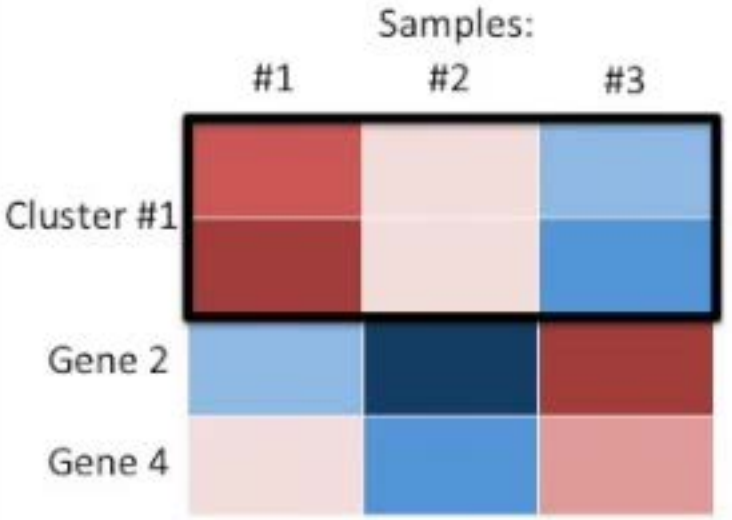


Hierarchical Clustering – more nit-picky details



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

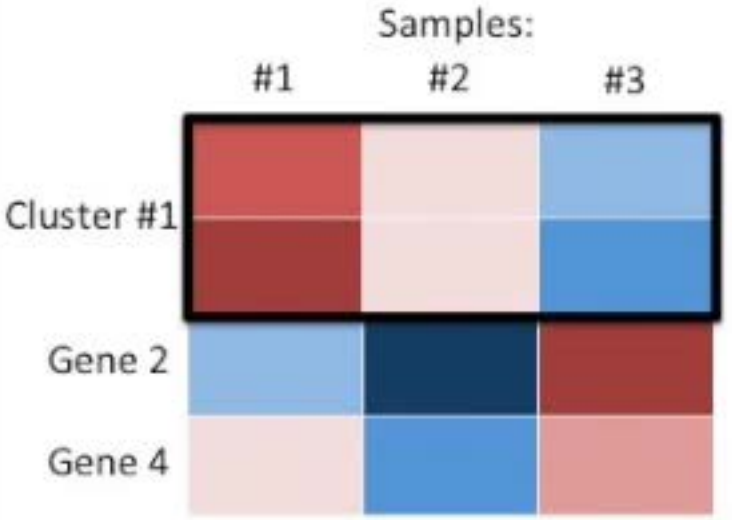
Hierarchical Clustering – more nit-picky details



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

Well, there are different ways to do that, too.

Hierarchical Clustering – more nit-picky details



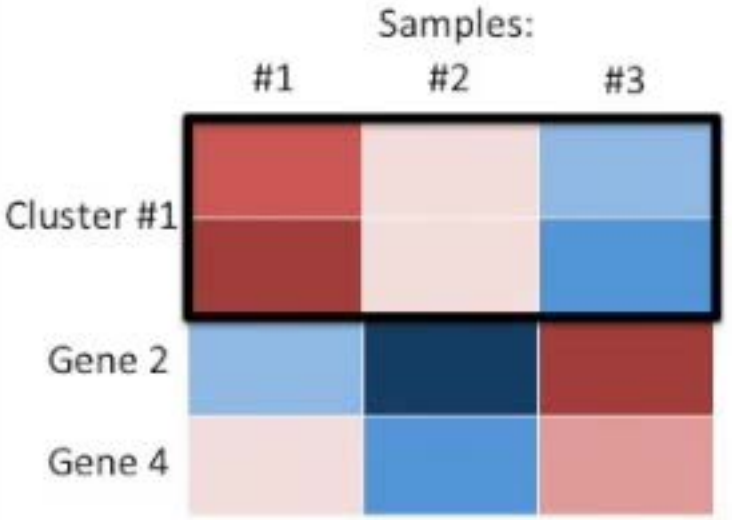
Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

Well, there are different ways to do that, too.

One simple idea is to compare other genes to the average of the measurements from each sample.

But there are lots more.

Hierarchical Clustering – more nit-picky details



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

Well, there are different ways to do that, too.

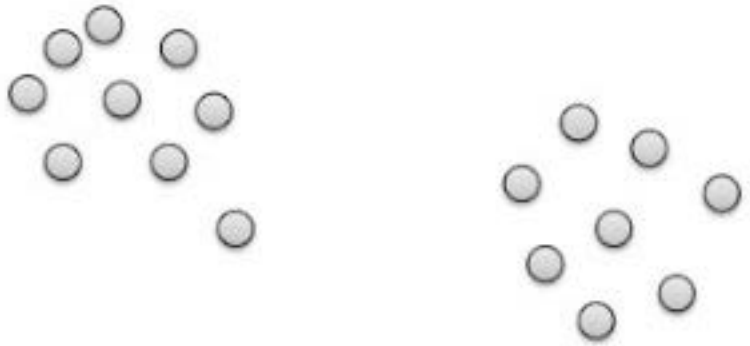
One simple idea is to compare other genes to the average of the measurements from each sample.

But there are lots more.

And these effect clustering as well...

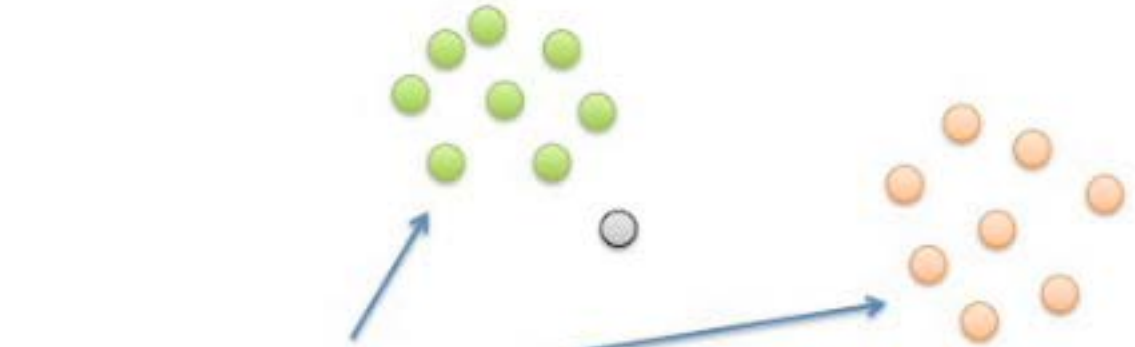
Different Ways To Compare To Clusters

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



Different Ways To Compare To Clusters

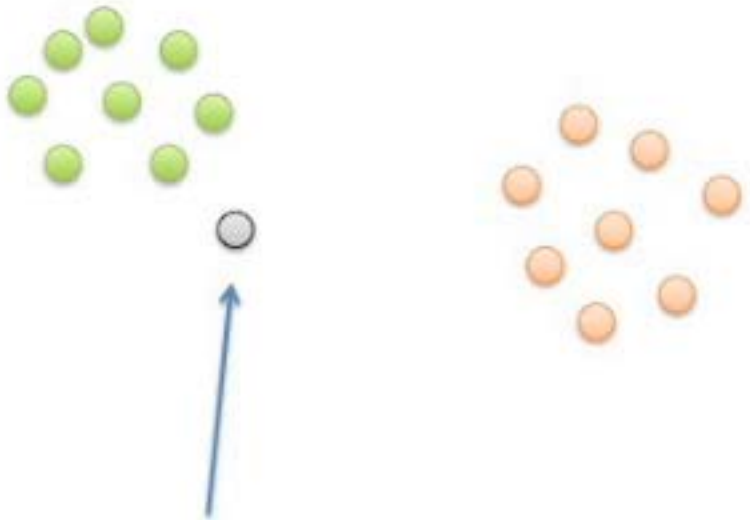
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



Now imagine that we have already formed these two clusters...

Different Ways To Compare To Clusters

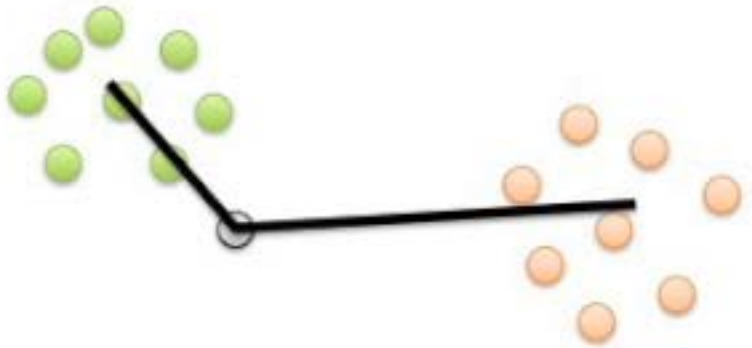
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



... and we just want to figure out which cluster this last point belongs to.

Different Ways To Compare To Clusters

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to...

1) The average

Different Ways To Compare To Clusters

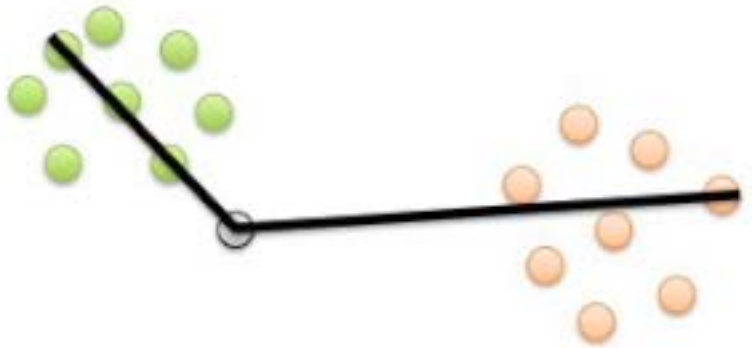
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.

We can compare that point to...

1) The average
2) The closest point

Different Ways To Compare To Clusters

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.

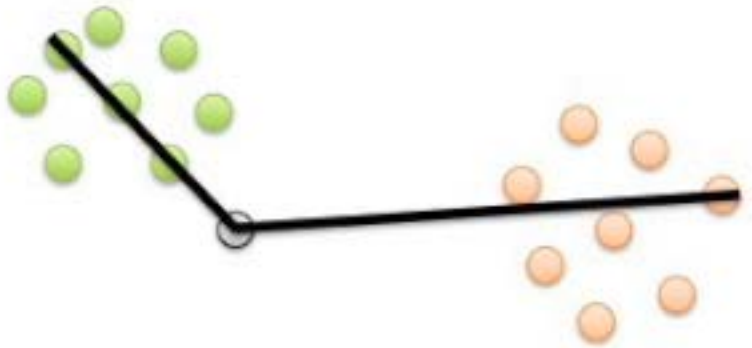


We can compare that point to...

1) The average
2) The closest point
3) The furthest point

Different Ways To Compare To Clusters

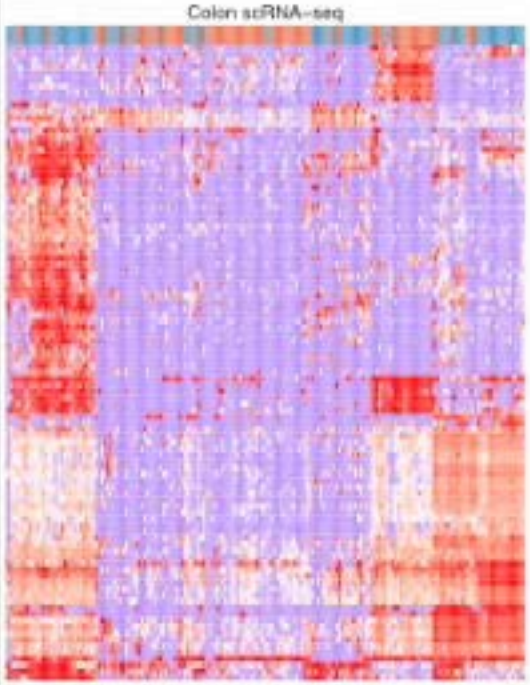
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to...

1) The average
2) The closest point
3) The furthest point
4) etc.

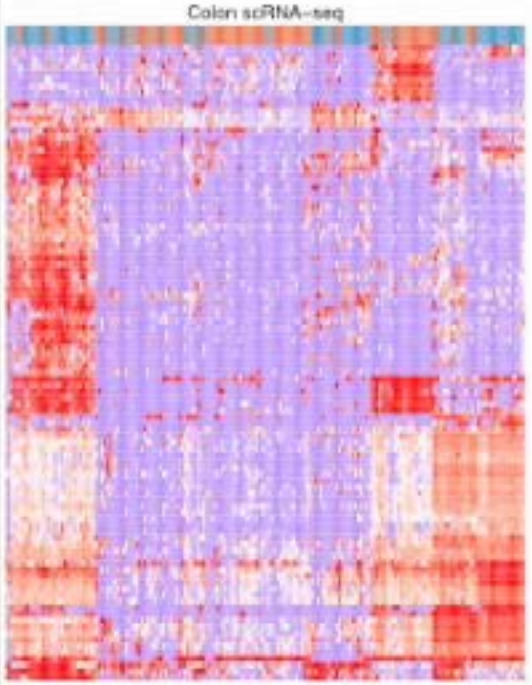
Some examples...



Compare points to the **furthest** in the cluster.

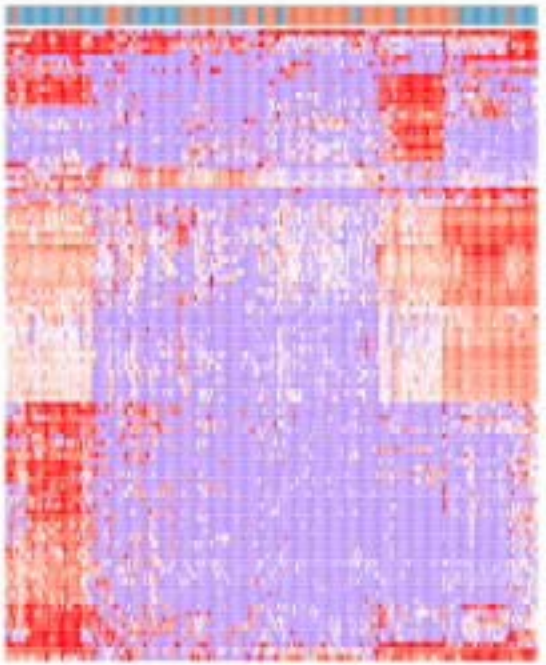
NOTE: This is the default for clustering in R.

Some examples…



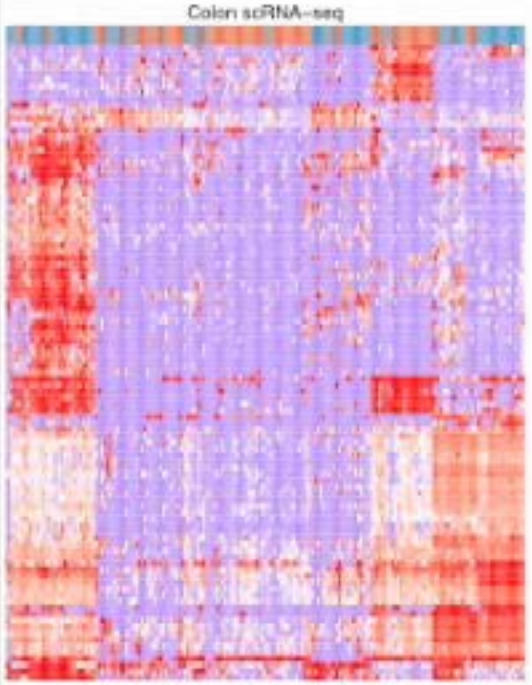
Compare points to the **furthest** in the cluster.

NOTE: This is the default for clustering in R.



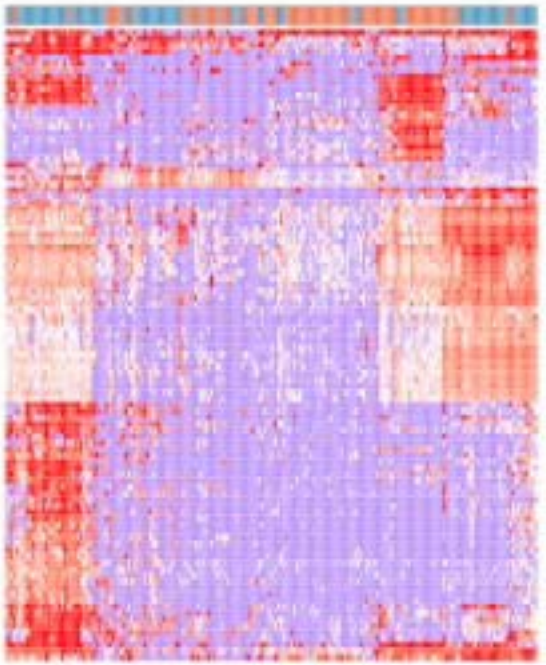
Compare points to the cluster **average**

Some examples…

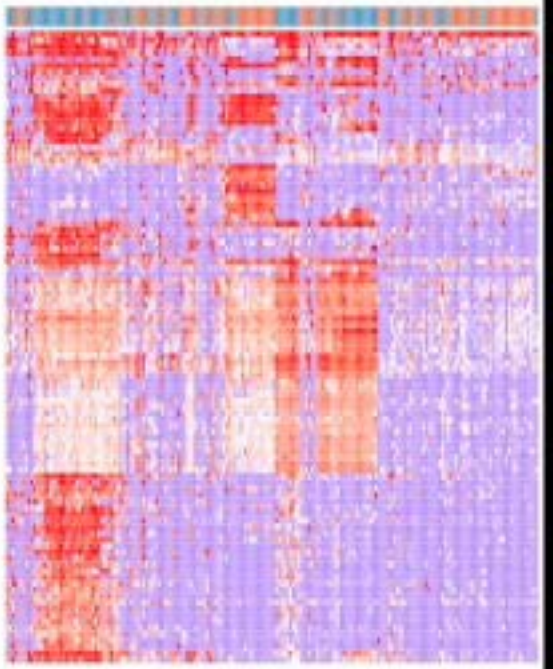


Compare points to the **furthest** in the cluster.

NOTE: This is the default for clustering in R.



Compare points to the cluster **average**



Compare points to the **closest** in the cluster.

In summary, to make a heatmap you:

In summary, to make a heatmap you:

- Scale the data (either per gene, or globally).

In summary, to make a heatmap you:

- Scale the data (either per gene, per sample, or globally).
- Cluster the data (either by gene, or sample, or both gene and sample)