

From:

<https://github.com/shendurelab/cfDNA>

PCR duplicate removal: eliminate duplicate or redundant DNA fragments that result from the Polymerase Chain Reaction (PCR) amplification of DNA.

special tool (SAMtools API) to find the positions where DNA fragments start and end.

"fragment coverage," which means all the positions between the start and end points of a DNA fragment, including the start and end points.

Windowed Protection Score (WPS) of a window of size  $k$  as the number of molecules spanning the window minus those with an endpoint within the window.

Savitzky-Golay filter

#### **NOTES FROM MEETING 9/25/2023 :**

- Our epigenetic data relies on the biopsy data(source of data).

### **How do we compare with ground truth?**

We can use the normal cell's information and its components to compare that with a sick cell. Sometimes, we can use the amount of the healthy DNA copy numbers as a reference and infer what is a reasonable boundary for an infected cell's DNA copy number.

Cancer - compared with healthy data

- Using filters ???

## **Data**

Data cleaning- for some of them the variation are removed, but for others maybe we don't know it's a huge variation

- Data cleaning is both easy and hard. Since the obvious outliers are already spotted when the data is first collected, the second-hand data we will receive is already cleaned. This means the left over data may have some hidden

Our first priority is to set up a numerical matrix.

Each column of the matrix would be the levels of signals that belong to different organs. Instead of using 0 for no signal, and 1 for some signal. We should use a more precise approach: The stronger the organ signal, the higher the number.

## **Nucleosomes**

For nucleosomes data, we could use the 0 or 1 approach to construct a binary matrix. Or we can approximate the amount of occupancy of the nucleosomes .

Nucleosome maps are closely related to epigenetic signals. Nucleosome maps can vary between different cell types and organs due to tissue-specific gene expression patterns and regulatory mechanisms.

epigenetic modifications, such as DNA methylation and histone modifications, alter the structure of chromatin, influencing how genes are accessed and regulated. Nucleosomes are fundamental units of chromatin, consisting of DNA wrapped around histone proteins. Changes in nucleosome positioning and modifications of histone proteins are key components of epigenetic regulation.

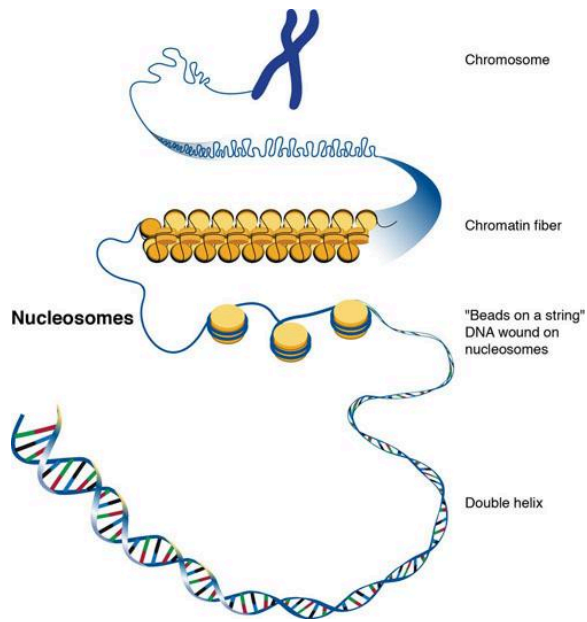
**Nucleosome Positioning:** The precise positioning of nucleosomes along the DNA strand can influence gene expression. Nucleosome mapping techniques reveal where nucleosomes are located within the genome. Regions with tightly packed nucleosomes are often associated with gene repression, as they can block the access of transcription factors and other regulatory proteins to the DNA. Conversely, nucleosome-depleted regions are often found near active genes, making the DNA accessible for transcriptional machinery.

Mapping nucleosomes is important because their positioning plays a crucial role in regulating various biological processes, such as gene expression, DNA replication, and repair.

Do we have some sample data of the nucleosomes mapping for different organs? Or how else do we compare the signals?

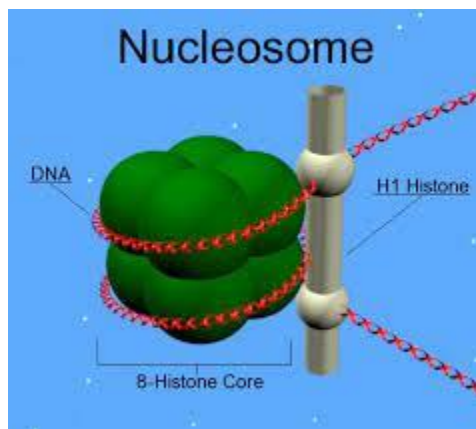
How do we extract chromatin from DNA?

**Immunoprecipitation:** Antibodies specific to the protein of interest (in this case, histones or nucleosome-associated proteins) are added to the cell lysate. These antibodies will selectively bind to the target proteins, forming antibody-protein complexes. These complexes are then pulled out of the solution using a solid support, such as protein A/G beads, that binds specifically to the antibodies. This process is called immunoprecipitation



Immunoprecipitation is a laboratory technique used to isolate a specific protein or protein complex, along with the DNA or RNA molecules it is associated with, from a mixture of biomolecules.

In the context of nucleosome mapping, immunoprecipitation is used to selectively isolate DNA fragments that are bound to histone proteins or other nucleosome-associated proteins.



### Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq):

Question: How do we handle big amount of data?

Meeting notes 10/2/2023

We need to create a reference matrix after we get 2 piece of data.

The first piece is about cell fragments.

The second piece of data is about transcription starts site. (TFs stands for a window of important locations which contains nucleosomes. Nucleosomes free positions would not be shown in the TFs)

The transcription starts

We need to intersect those data to calculate the reference matrix.



Nucleosomes has some tails that branches out and not directly from the original strands of DNA. When the transcription factors reads it, it knows which part of the information on the DNA would be more important and which parts can be skipped. This is indeed how the location of nucleosomes can affect different gene expression if not all parts of DNA are being transcribed.





From EE87917.frag.tsv from DIR06cGc

What does each column means?

```
chr1 10000 10150 1
chr1 10002 10144 1
chr1 10003 10139 1
chr1 10008 10156 1
chr1 10012 10138 1
```

What does each folder means

 **Debajyoti Kabiraj** > **Trainee02\_Peter** 

	Name ▾	M
	DIR_05cGC.tar.gz	Oc
	DIR_06crHL_sub.tar.gz	6 c
	DIR_08cOC.tar.gz	Oc

2 data

First data:

TFs

Each row is different sample

Do we know if the sample is healthy or cancer or cancer from which organs?

Second data:

Sample

Is everything unknown?

Data:

Col: chromosomes name

- Different chromosomes at different location, it's different and not comparable. Comes from different part of genomes
- Only same chromosomes are comparable for test and the training sample

Start of fragment

End of fragment

Count of fragment (ignore for now)

Study genome density

Calculate density for a particular window

For each window count, how many fragments are mapped there

Files:

1 healthy - male and female?? For female, no Y chromosomes. Healthy x-average copy number is 1, female copy number is 2.

2 unhealthy - sample

22 regular 2 sex chromosomes

1-22

23-24 chromosomes - CHR X OR Y

OVARIAN CANCER + GASTRIC CANCER

What does each chromosome number mean?

Gene is under shadowed as it should be -> false analysis

Mountains and valleys have different meanings.

Take 100 base pair window

Count how many reads

Making graph to interval

If data has start 120 ends 190, then 100-200 window read count +1  
If start 190 ends 240, then both window could +1 or make it fraction +0.5

100 - 200 : reading count

200 - 300 : reading count

[[read count from 0-100], [read count from 100-200], [read count from 200-300]]

How to incorporate Epigenetic signal is Yet to check.

Paper codebase video

All sources - due 10/16

-authors

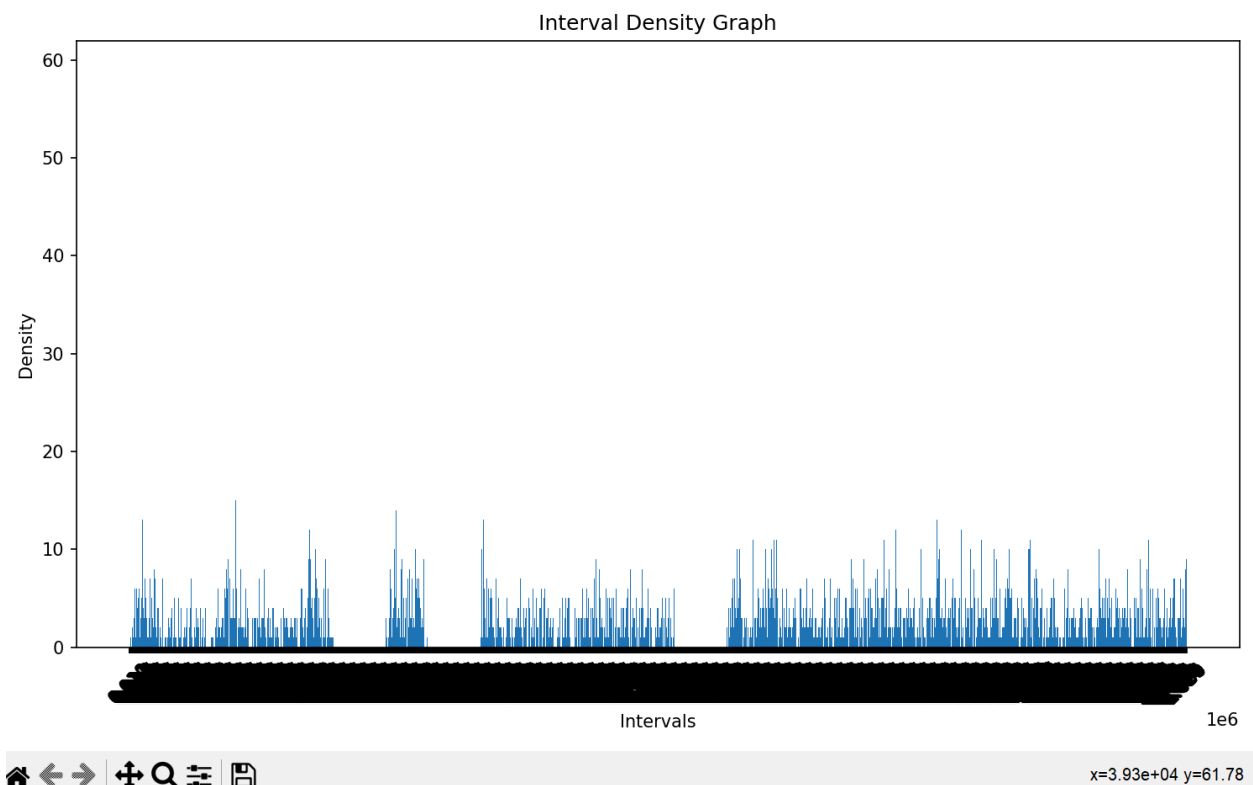
-title

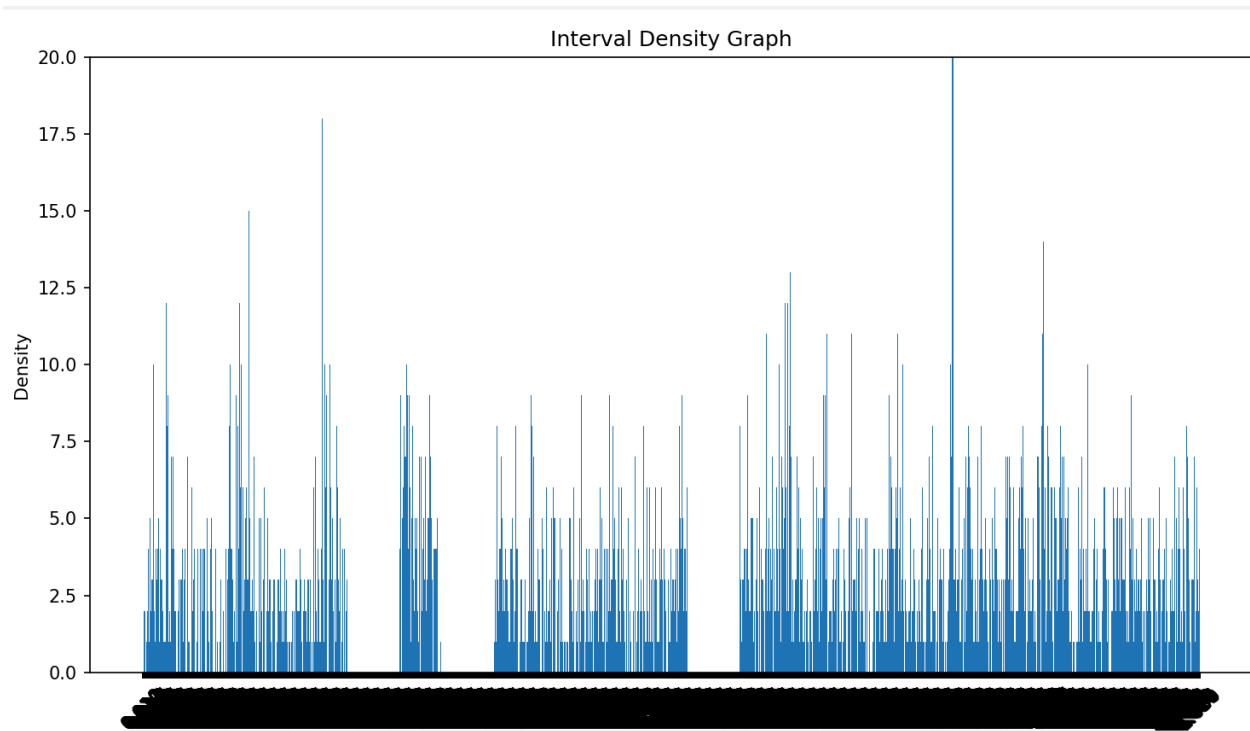
-2-3 sentence description

Second stage - due 10/23

Oct/25 present on just literature review

First 10000 in GC





Increase interval size makes calculation less  
 200 has more noise than 100 base pairs  
 Lack of density might mean lack of fragment copy number

The mountains and valleys means nucleosome occupancy

Tune algo to find intermediate scale (where the irregular numbers of copy occurs)  
 -1kb or 10kb  
 There might be biases in the data.

GC constant, some parts are harder and easier to sequence which results in alteration

If DNA is tightly bound to something else, other part would get discarded. Fragile piece might be broken. Those can help explain the mountains and valleys.

False sense of security for under represented regions

1. central mire: Center of chromosomes is hard to sequence

Check data with primer tumor data, there should be some similarity or overlap with the copy numbers.

y=  
 Overall expression

Gcm?

Predict y given x

Predict epigenome using red count

Run program with whole data, some beads are biologically more meaningful.

TSS

Overlay TSS, GC content (feature 1,2, 3... )

Score shows dna fragment:

More concentrated different color

Matrix col: TSS (color)

Row: different samples

N x n matrix , non negative matrix factorization (NMF)

Without y training VS with more info training model

4 col

Gene name , chr, start, end

-2200 bp

Read count

Change the scale

Fragment:

	0-100	100-200
S1:	1	1
S2:	2	3

In our second piece of data: TSS(transcription start sites), We have 4 columns total. They are chromosome types, start of transcription sites, end of the transcription sites, name.

We would want to create a new matrix which each column stands for TSS1, TSS2, and so on. Each row of the new matrix would be different blood sample.

Each cells means the amount of intervals of blood sample that lies within the interval of that specific TSS. For example, in the 1st row and 1st column, that cell would record how many intervals of sample1 would fell into interval of TSS1.

Matrix:

	TSS1	TSS2	TSS3
--	------	------	------



Sample1:	?	?	?
Sample2:	?	?	?
Sample3:	?	?	?

MLR

[https://docs.google.com/forms/d/e/1FAIpQLSeiRQI5QhMf0MqOMd\\_JTaCap54C9-WfktZBWwshsFrG79cJoA/viewform](https://docs.google.com/forms/d/e/1FAIpQLSeiRQI5QhMf0MqOMd_JTaCap54C9-WfktZBWwshsFrG79cJoA/viewform)

How many windows of fragments fell into the TSS?

Color things by cancer types

After check check Derivative of read count, WPS

Results should be similar to the TSS approach.

Across group variation difference

Subgroups of our data and signal can identify, maybe we don't need all of the data of x and y

NMF approach:

$$M = W * H$$

Here, W will have dimensions (number of blood samples x 5) and H will have dimensions (5 x number of TSS intervals). The number 5 here represents the number of latent features/components.

- **W (Matrix):** Represents the association of each blood sample with the 5 latent features/components. Each row of W corresponds to a blood sample, and the 5 columns represent the weights/strengths of the 5 latent features for that sample.
- **H (Matrix):** Represents how each of the 5 latent features/components is associated with the TSS intervals. Each row of H corresponds to one of the latent features, and the columns represent the weights/strengths of that feature for each TSS interval.

To gain insights from W and H:

- Examine the rows of W to see which blood samples have strong associations with the latent features.
- Examine the rows of H to see which TSS intervals are strongly associated with each latent feature.

**To choose N:**

**Elbow Method:** Similar to K-means clustering, one can plot the reconstruction error as a function of the number of components. At some point, increasing the number of components

results in diminishing returns in terms of reducing the error. The point where the decrease in error slows down (the "elbow") can be considered an appropriate number of components.

Around TSS3180 results appears

What next? Classification algorithm?

Do we need column of Y?

How to choose n-num of component?

Results for 1 complete file of blood sample:

```
PS C:\Users\pzxia\OneDrive\桌面\data> & C:/Users/pzxia/AppData/Local/Programs/Python/Python38/python.exe c:/Users/pzxia/OneDrive/桌面/data/createReferenceMatrix.py
TSS1 TSS2 TSS3 TSS4 TSS5 ... TSS9160 TSS9161 TSS9162 TSS9163 TSS9164
0 414 439 413 476 373 ... 508 582 564 555 563

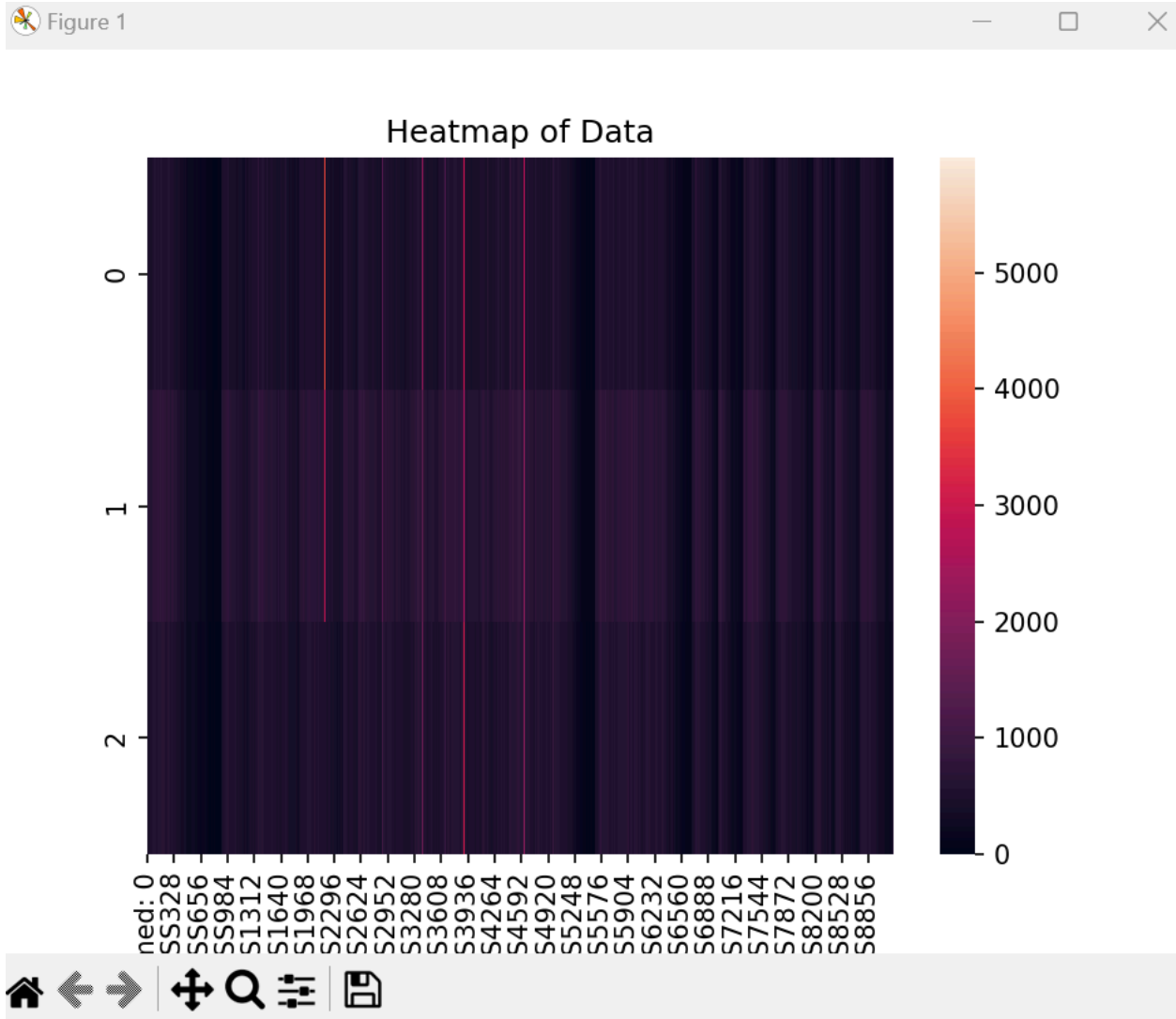
[1 rows x 9164 columns]
M: [[19.24450381 17.11666464 11.19776762 5.84819417 3.54407404]]

H: [[4.08649617e+00 1.65100120e+01 6.10165725e+00 ... 1.61187720e+01
2.08226333e+01 9.92159250e+00]
[8.71487752e+00 5.92870703e-01 1.16389874e+01 ... 5.55108667e+00
3.77779325e+00 5.63612905e+00]
[1.08100553e+01 2.22965767e+00 2.59392251e-01 ... 2.08355641e-01
2.16987813e-02 1.61375425e+01]
[4.60912662e+00 9.00613117e+00 1.33278767e+01 ... 1.13559680e+01
2.12224632e+00 1.56382395e+01]
```

	A	B	C	D	E	F	G	H	I	
1		TSS1	TSS2	TSS3	TSS4	TSS5	TSS6	TSS7	TSS8	T
2	0	414	439	413	476	373	451	465	399	
3	1	637	547	536	632	544	571	585	568	
4	2	396	394	394	402	411	396	412	361	

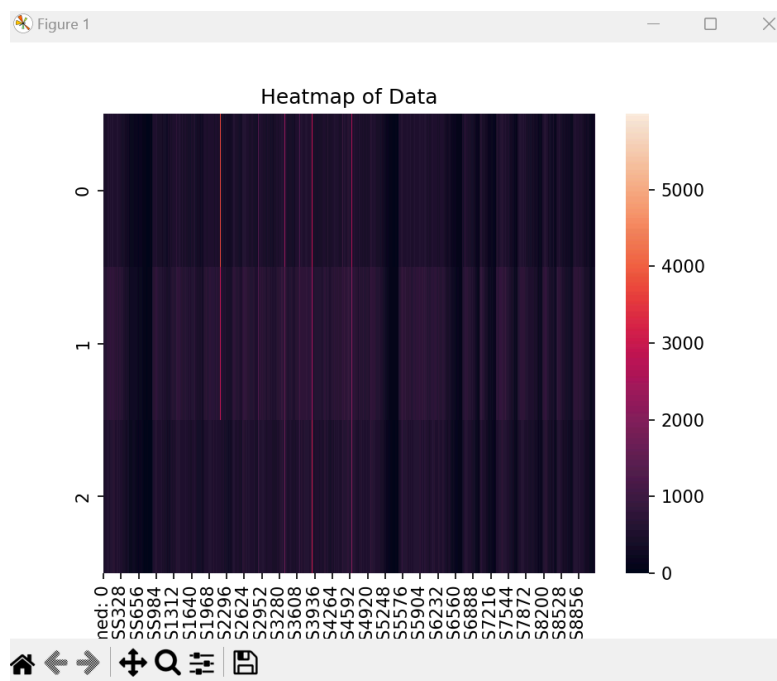
Meeting 10/30/2023:

1. Doing heat map
2. instead of checking interval, we check if the mean of blood sample falls in the certain intervals
3. clustering based on row or cols
4. Use ameral to create reference matrix using mid interval

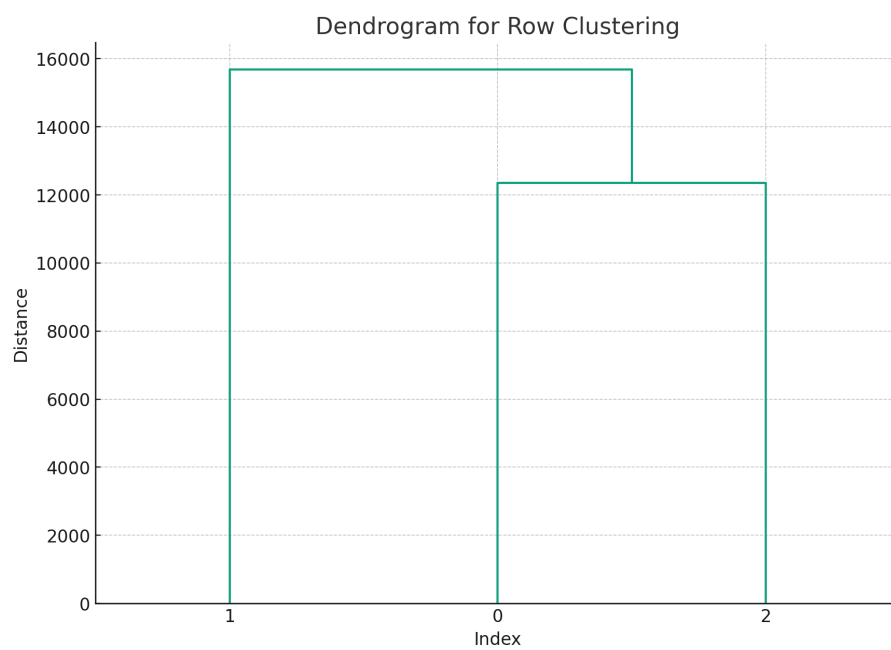


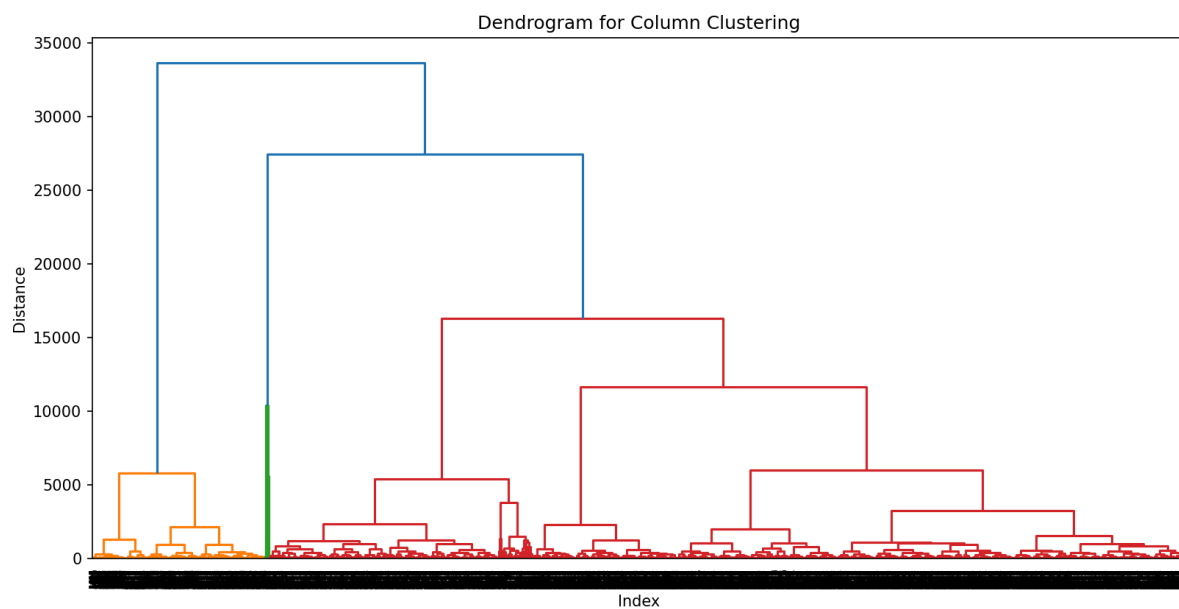
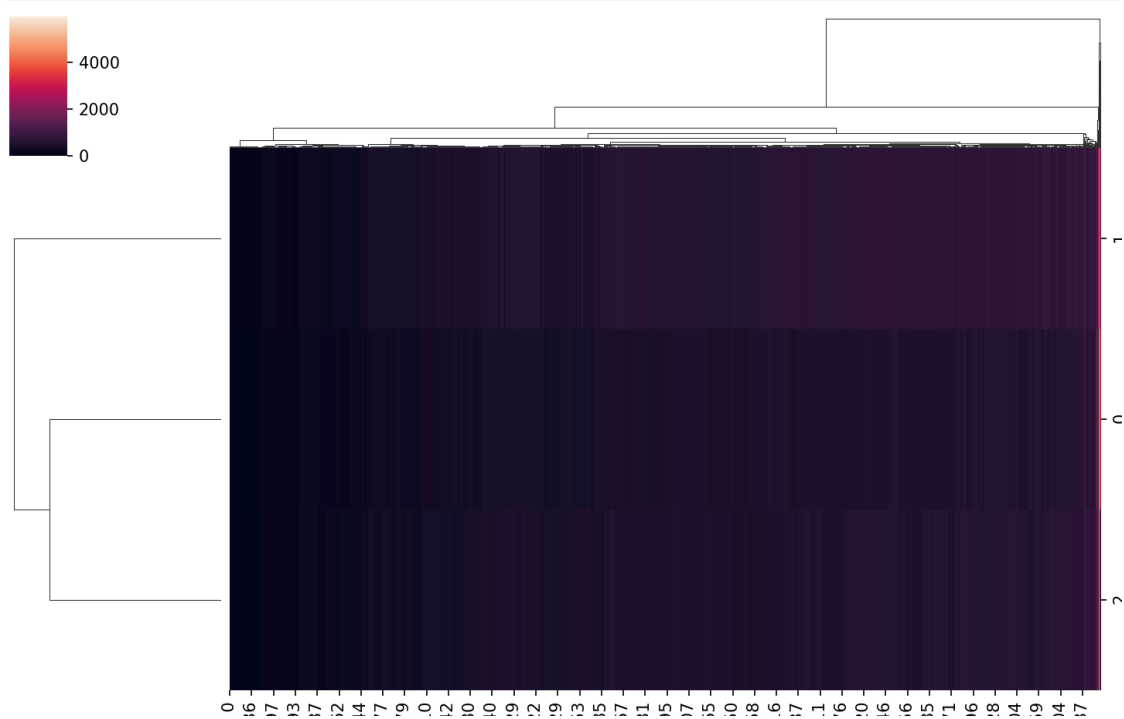
Use NMF on matrix to get signatures  $M = W * H$   
 $k=?$

Prepping for 11/13.2023 meetings:  
3 Gastric cancer sample heat map of TSS



Clustering:

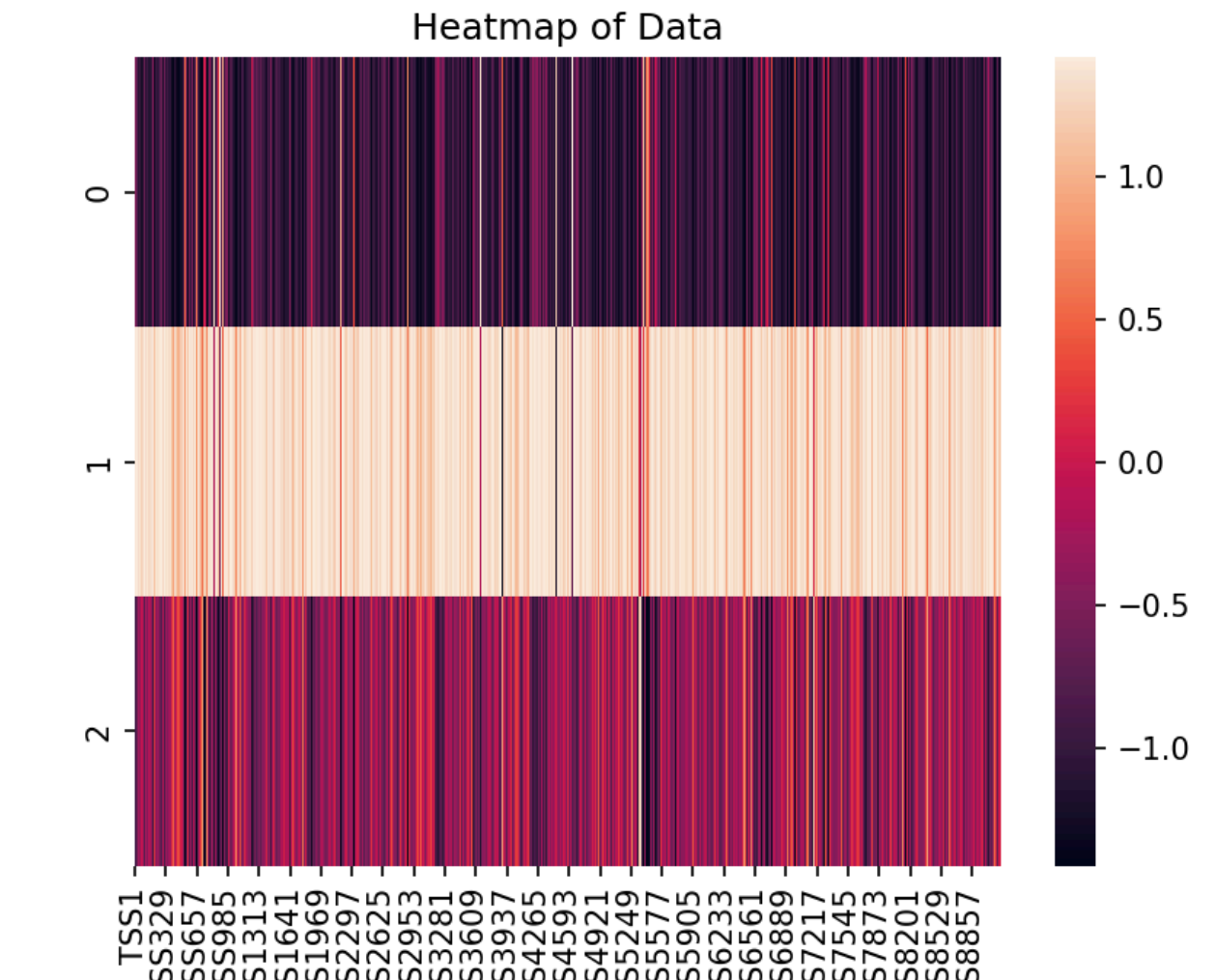




Combine small sets of signature together.

Do we choose the N for NMF based on the col clustering graph?

Data-normalization: what does normalization mean? How do i do it?  
After normalized through z-score of col(mean = 0, sd=1)



Meeting notes 11/13/2023:

We should normalize by row and not by columns, Mean =0 for the row, sd =1 for the row

After asking debajoyti for the complete sets of TSS matrix's reference matrix,

TFBS of sample matrix

Cell free DNA count matrix

TSS size in the samples

, we then do the same clustering NMF,

For 11/27/2023 meeting:

Questions regarding given matrix:

blood_vessel_h16	adrenal_gland_h16	bone_element_h16	Bronchus_h16
0	0	0	0
1	0	0	0
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

EE87893In1	EE87894In1	EE87895In1
0.00820292521372716	0.00760121151974732	0.006642369176726
0.00650066442998882	0.0042914542526098	0.005760225877229
0.00410146260686358	0.00858290850521961	0.006086494674061
0.00820292521372716	0.0062868668454457	0.00496332262944
0.0106021270368524	0	0.004458276765745
0.00410146260686358	0.0042914542526098	0.003043247337030
0.0130013288599776	0.00760121151974732	0.006086494674061

For the EE part:

Each row means different TFBS, but for where and which organ?

Each col means different sample

For the specific genomes part:

Are they the y-label for the first matrix? Given similar TFBS data, should we be able to predict the signal of signature in different parts? How to use this part of data in collaboration with the first part?

Overall:

Do I have to normalize the data again?

Afterward, NMF, PCA, clustering?

Timeline for project?

- benching marking methods, checking accuracy of results

- part 1: develop method and comparing method

- part 2 check if our results make sense if results is good

1.Reduce to less fraction num

AA y1

X overall

Analysis in 2 part

Only take X

NMF on the X, reduce dimensions , seeing similarity with organs

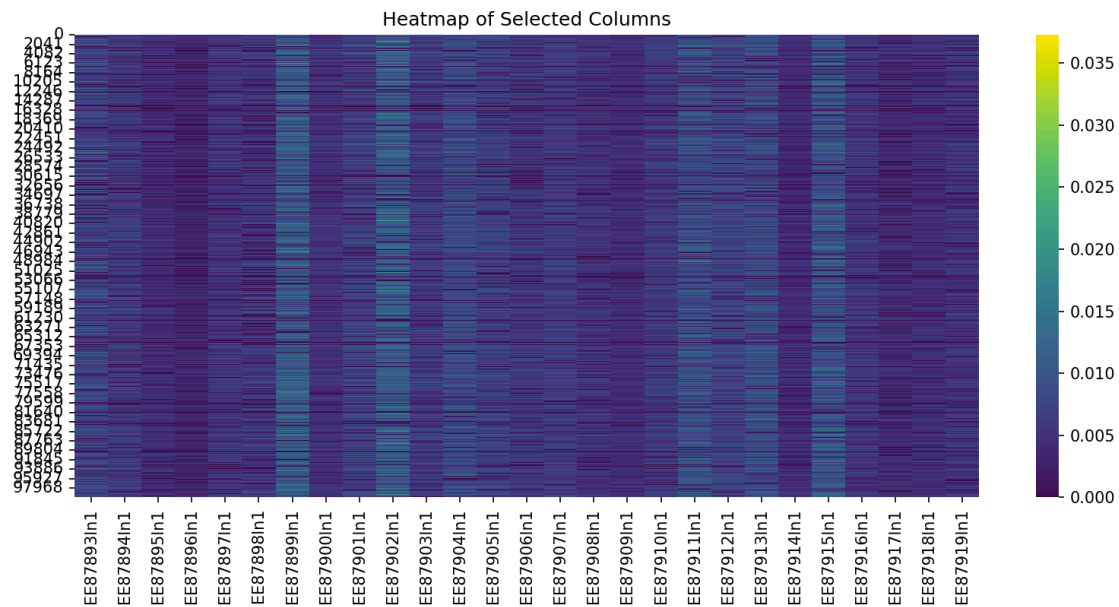
Y1 = weighted combination

Regression y = function of x

Finding coefficient

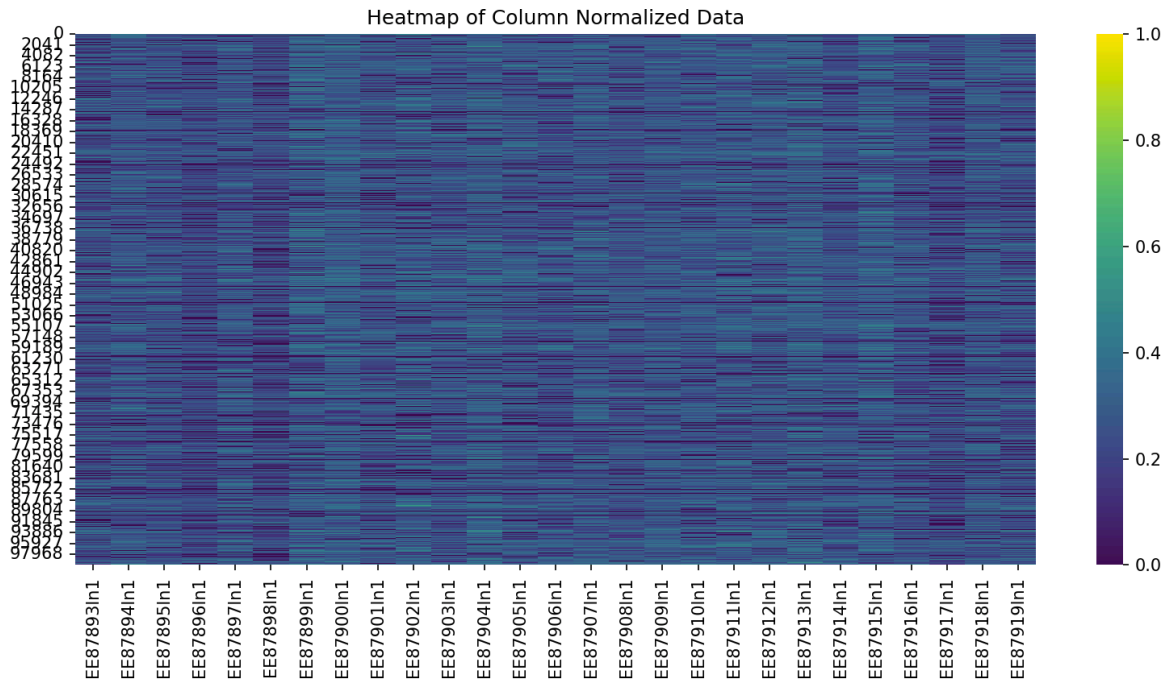
Training coefficient with non - coefficient

Heatmap for unnormalized TFBS



Normalized data heatmap(by column by sample):





### Questions:

What does each coefficient mean?

If i want to predict our Y, using a made up X, then our Y-value would be a singular value rather than a vector or array? Are we measuring prediction(y) for the transcription start site 'EE87899ln1', given the provided biological signatures(x)?

```

32
33 # Accessing the coefficients for 'EE87899ln1' from the results
34 coefficients_ee87899ln1 = individual_results_varX['EE87899ln1']
35 print(coefficients_ee87899ln1[0])
36 print(len(coefficients_ee87899ln1[0]))
37

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```

6ln1', 'EE87917ln1', 'EE87918ln1', 'EE87919ln1']
[-2.09061768e-04 -8.25792007e-05  4.46174024e-06  5.17150023e-05
 1.18968380e-04 -7.90498054e-05 -3.22506779e-05 -1.05941376e-04
 4.55343340e-05  2.03571104e-05 -5.33028024e-05 -1.59767718e-04
 4.99353532e-05  3.73477423e-04  1.33358167e-04 -8.39906922e-05
 6.71929435e-06 -2.71500974e-05 -1.28237071e-04 -2.61854096e-05
 9.35375809e-06 -8.42775446e-05  8.14973434e-05  3.28093113e-05]
24
PS C:\Users\pzyia\OneDrive\桌面\data\com>

```

Genes alterations

Make app shows timeline for upload data

Knowing genes and location

We can map on the actual genomes and show it

3 types of info

List of epigenetic signals that were altered

Then we can visualize it

Next week to spring break visualization

Circos plot in R

Genome data

Ways to see how cancer is formed

1 way is to take tumor out and get the genes

But if cancer is spread out it is hard to remove tumor and study them

From cfDNA in blood, maybe we can find the similar pattern. Certain signals are reinforced

2/27/2024

Delphi score - combined score

if delphi score predict

DNA copy number gain and loss

More DNA copy there more fragment and copy = amplify

Qvalue

$\log(\text{Qvalue}) = \text{copy number score}$

Use circos plot first

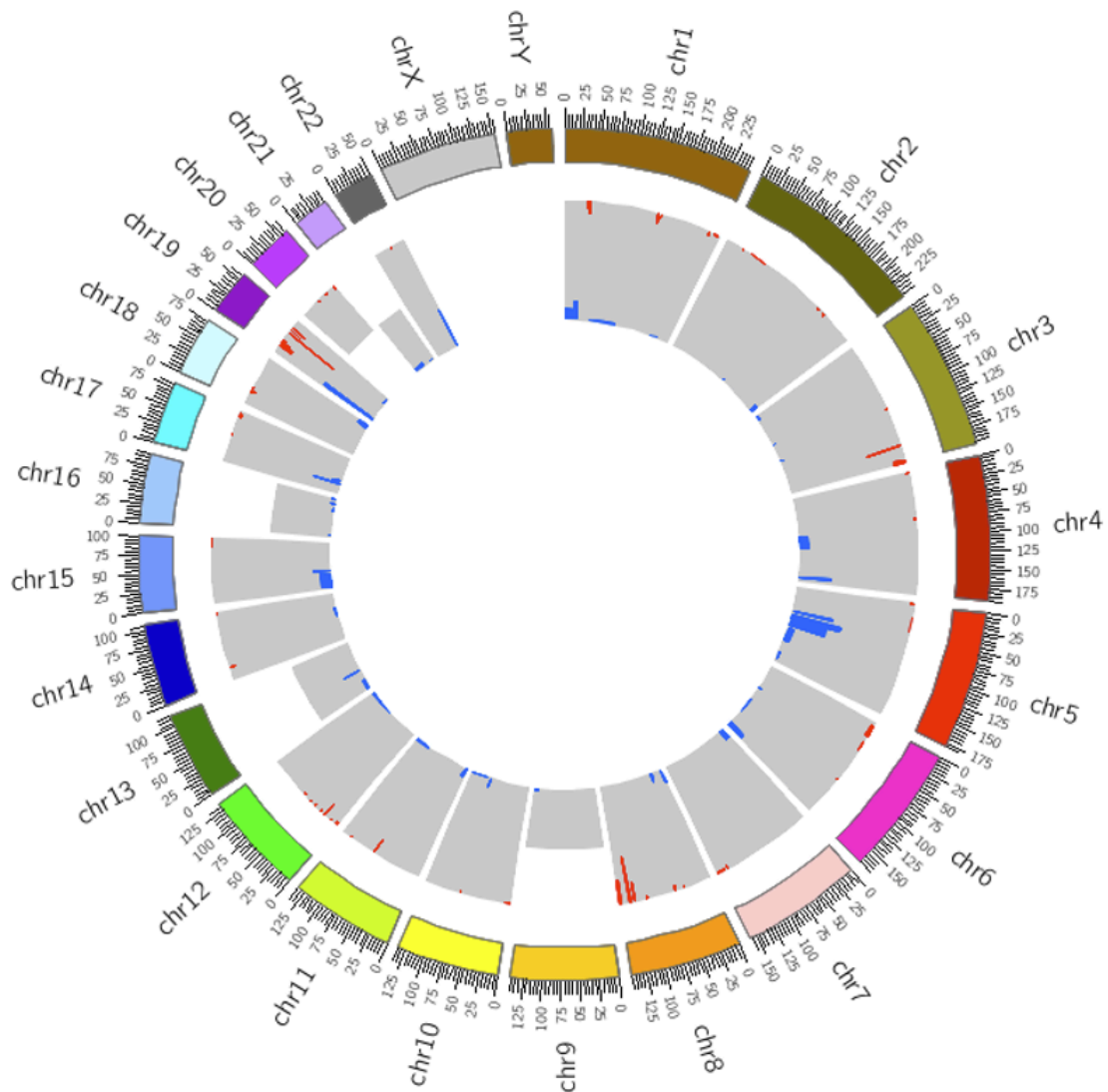
Link to last study, ovarian cancer

Some fragment is amplified or deleted

Data analysis

1. **Chromosome (Chromosome):** Indicates the chromosome number where the genetic alteration is located.
2. **Wide Peak Start (Wide peak start):** The starting position of the wide peak on the chromosome, which likely indicates a region of genomic alteration.
3. **Wide Peak End (Wide peak end):** The ending position of the wide peak, defining the other boundary of the genomic alteration region.
4. **Amp/Del (amp/del):** Specifies whether the region is amplified (Amp) or deleted (Del).

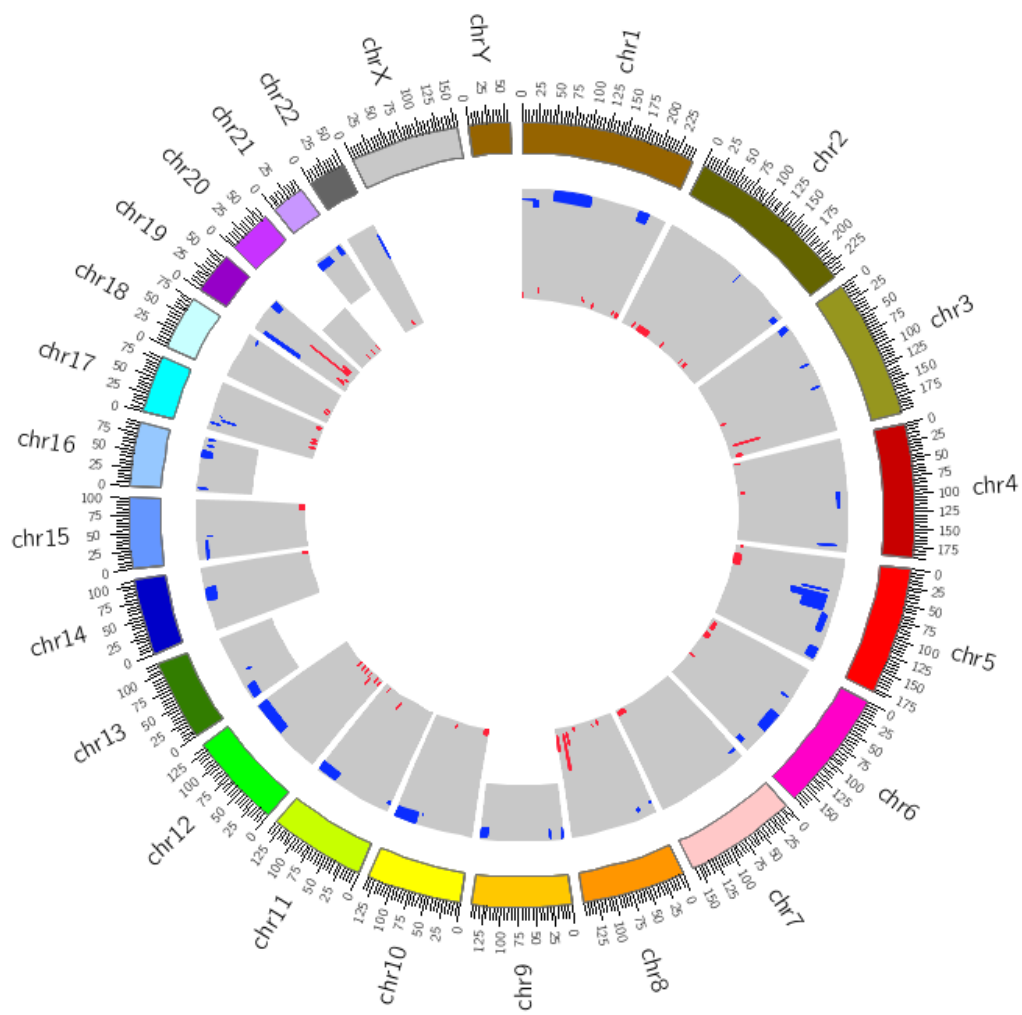
5. **Cytoband (2q3, 19q12, etc.):** The specific location on the chromosome, typically used in cytogenetic or molecular cytogenetic analysis to identify particular parts of chromosomes.
6. **Wide Peak Limits:** A more detailed notation of the wide peak region, specifying the chromosome, start and end positions, and the probes involved.
7. **Peak Limits:** Similar to Wide Peak Limits but might represent a more focused region within the wide peak.
8. **Region Limits:** Represents an even broader region than the wide peak, possibly encompassing the entire area of interest or alteration.
9. **q values:** Statistical values indicating the significance of the observed genomic alteration.
10. **Residual q values after removing segments shared with higher peaks:** Adjusted q values after accounting for overlapping regions with other significant peaks.
11. **Description with number/Description:** Provides the gene or genomic region's name, sometimes with a number indicating the count of something, possibly related observations or genes.
12. **Comment:** Any additional notes or comments on the specific data row.
13. **Deep Del, Shallow Del, Neutral, Low Gain, High Gain:** These columns likely represent the number of observations or samples where a deep deletion, shallow deletion, neutral change, low gain, or high gain was observed in the genomic region.
14. **Percentage (% Deep Del, % Shallow Del, % Neutral, % Low Gain, % High Gain):** The percentages corresponding to the absolute numbers in the preceding columns, representing the distribution of different types of alterations across samples.
15. **n\_genes:** The number of genes found within the specified region.
16. **Genes:** Lists the genes located in the alteration region, sometimes with additional data such as the relative expression level or a specific metric related to the study.



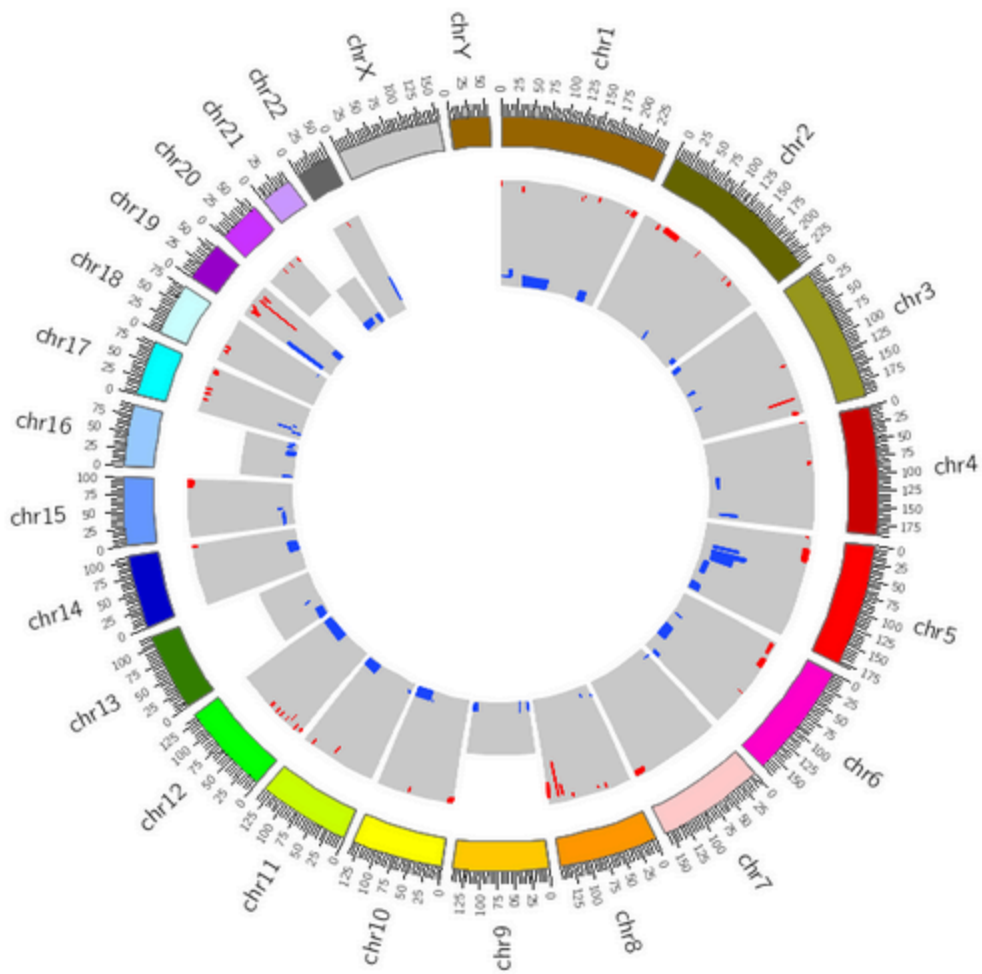
[https://www.bioinformatics.com.cn/plot\\_basic\\_2\\_tracks\\_circos\\_histogram\\_plot\\_031\\_en](https://www.bioinformatics.com.cn/plot_basic_2_tracks_circos_histogram_plot_031_en)

By using outer track AMP(red), inner track DEL(blue)

Below is using DEL on the outer track

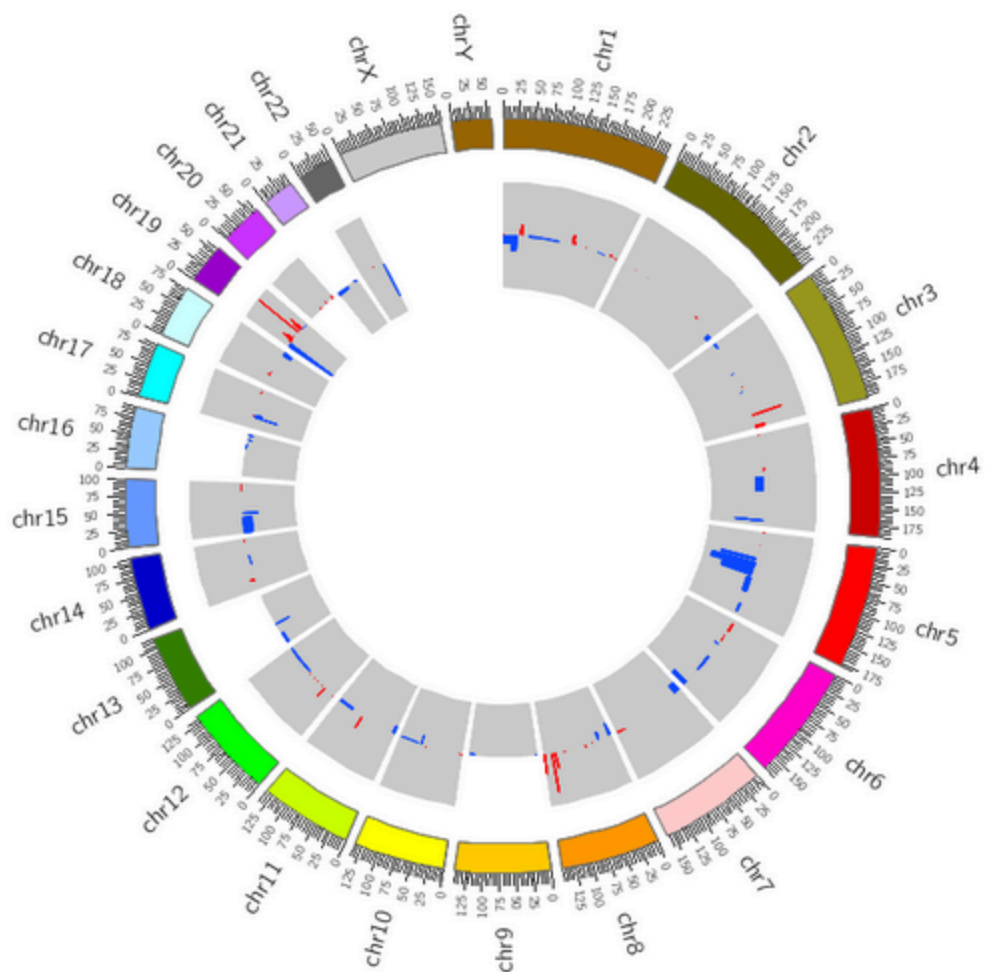


Below is normalized all q value for all AMP and del with amp(red) outer track and del(inner)

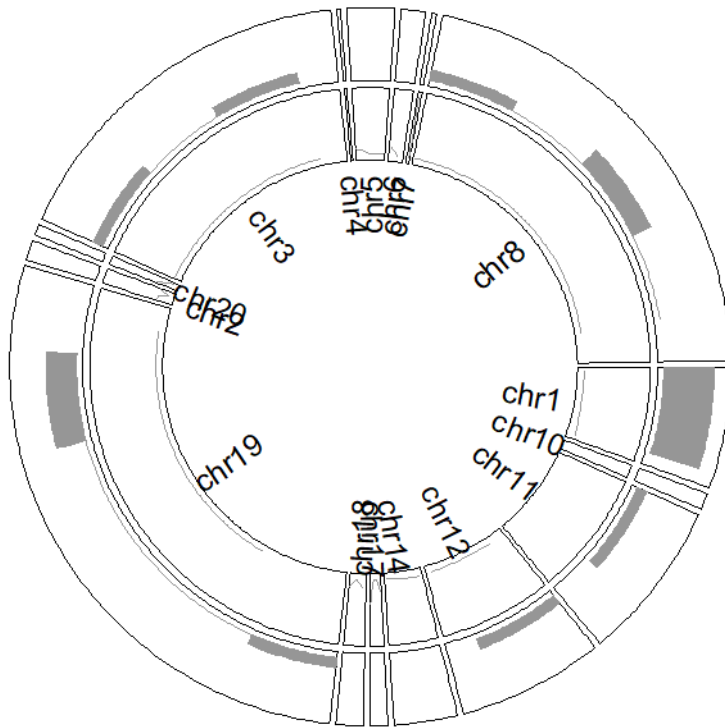


Observation:

If the value is negative then the starting point of the histogram is not from the 0-axis but from the corresponding inner and outer track starting line.



Future combine Q val from adjacent blocks - observation independence, adjusted blocks not independent of each others, maybe correction and integration needed



4/10/2024  
Questions

3. For the basic pycirclize, we have to set the beginning and end based on the maxim distances between cell fragment is that ok?

The outer track start and end just use the base hg38 ones. We just plot the interval with dots

4. should i normalize the log q column?

Dont need to, but we can put on a max cap.

5. q value is how much a gene express itself?

For the graph produced by the online website, see if i can get the code underneath and configure it. Also fix my axis problem. They should start and end from same axis of 0.

For 4/17/2024

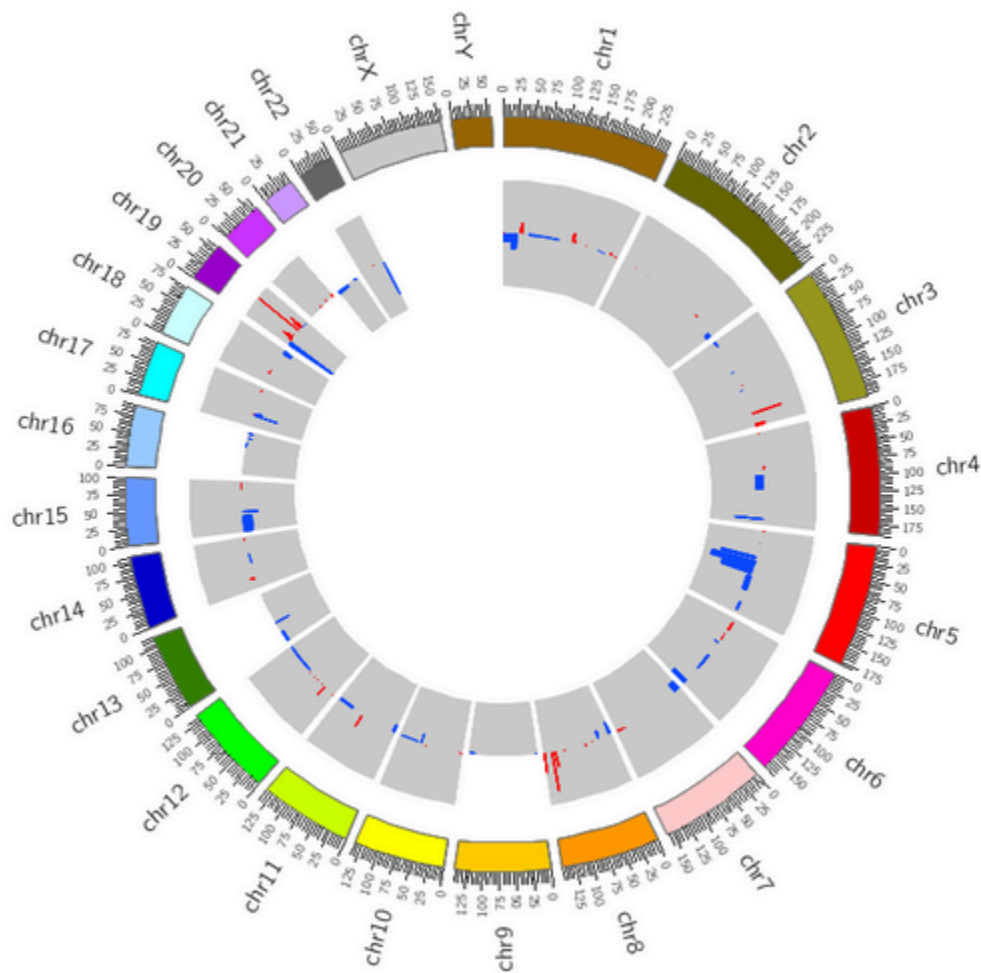
I used negative log Q value for the input for SRplot website.

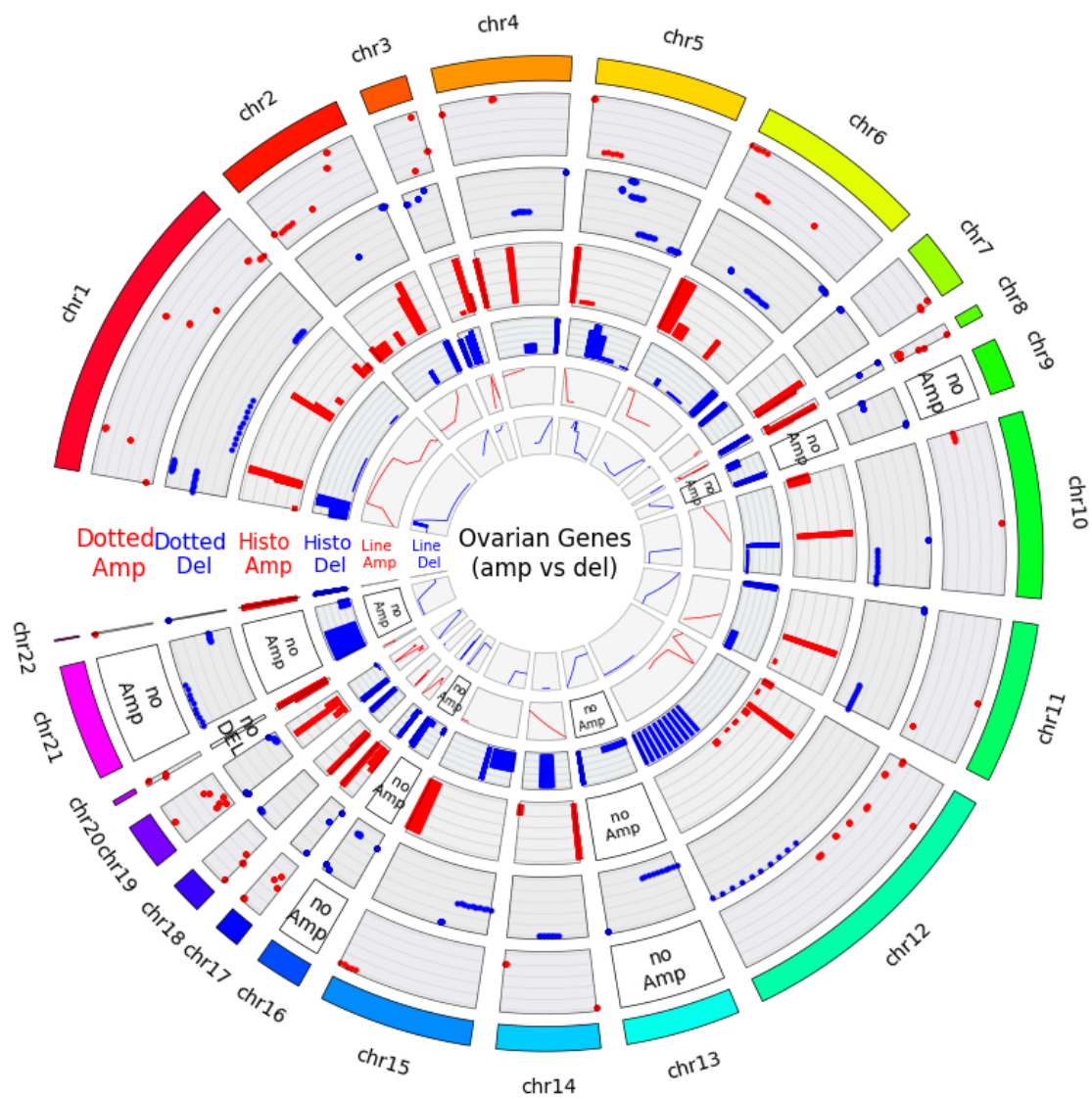
Questions

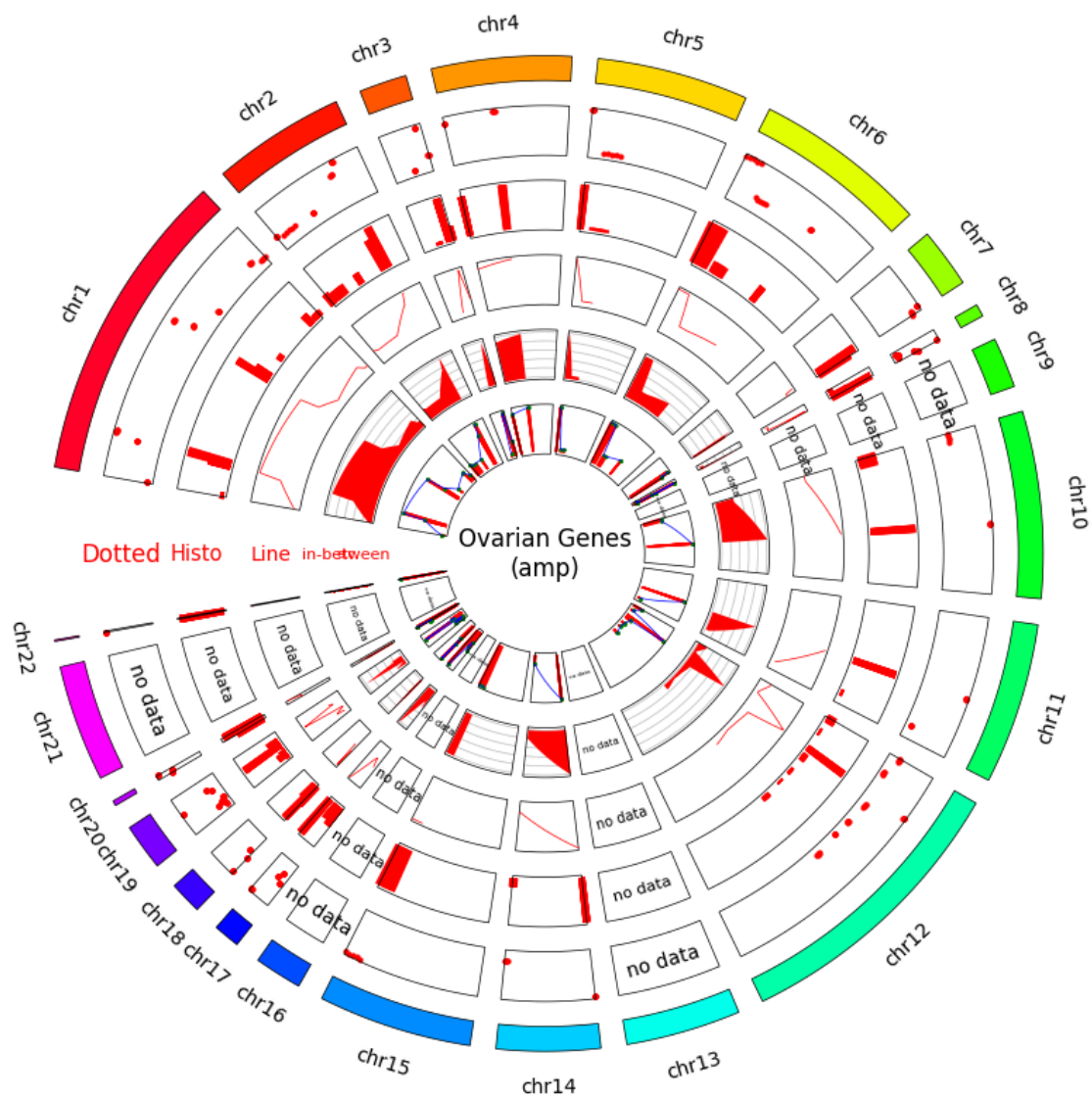


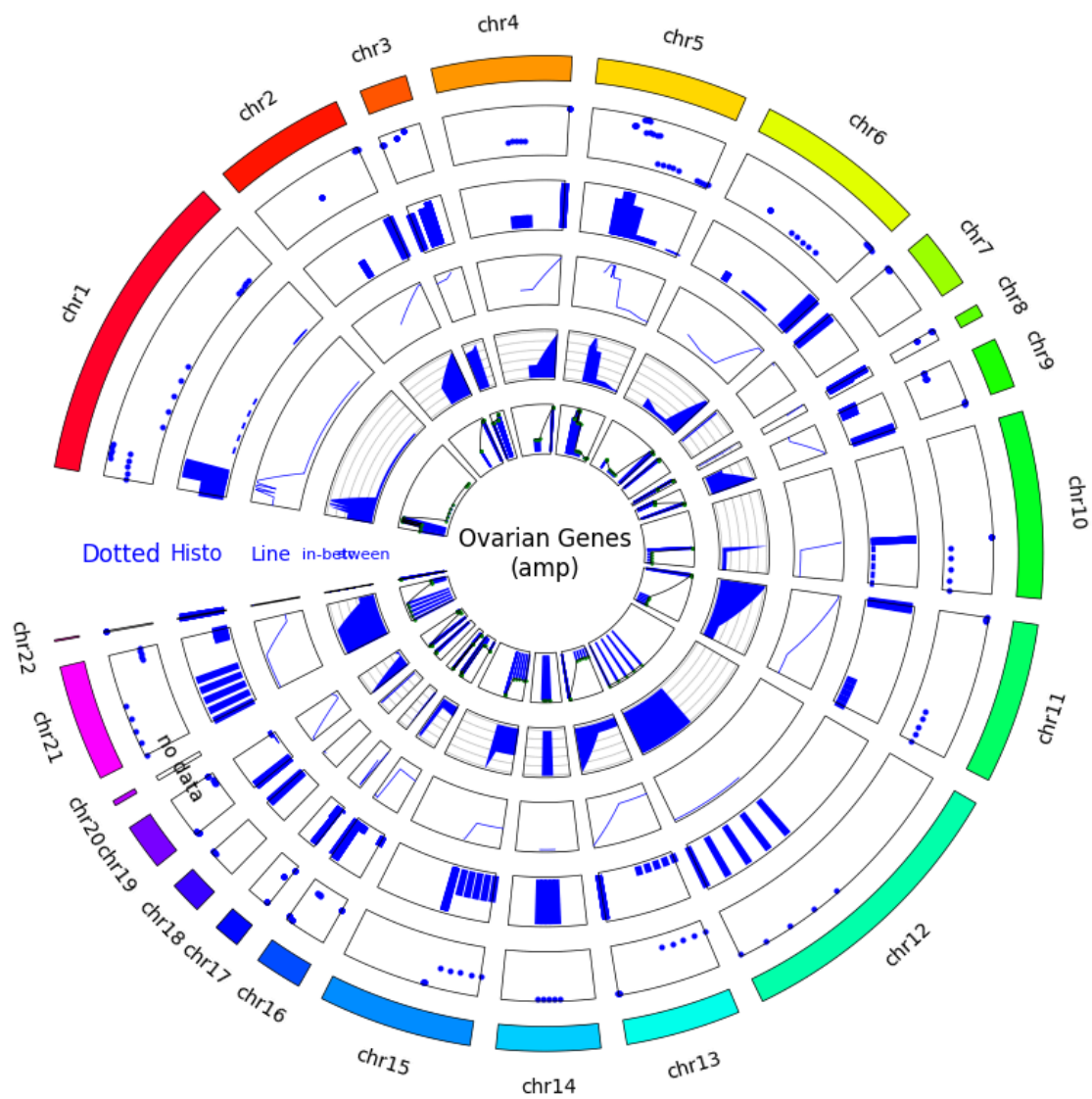
1. What exactly does Q represent? Higher is better or lower? What value do I use for the plot? Is  $-\log(Q)$  OK?  
P-value adjust repeating experiments  
P-value usually lower but adjust for others factors
2. How to analyze the plot? Higher  $-\log(Q)$  means more significant?

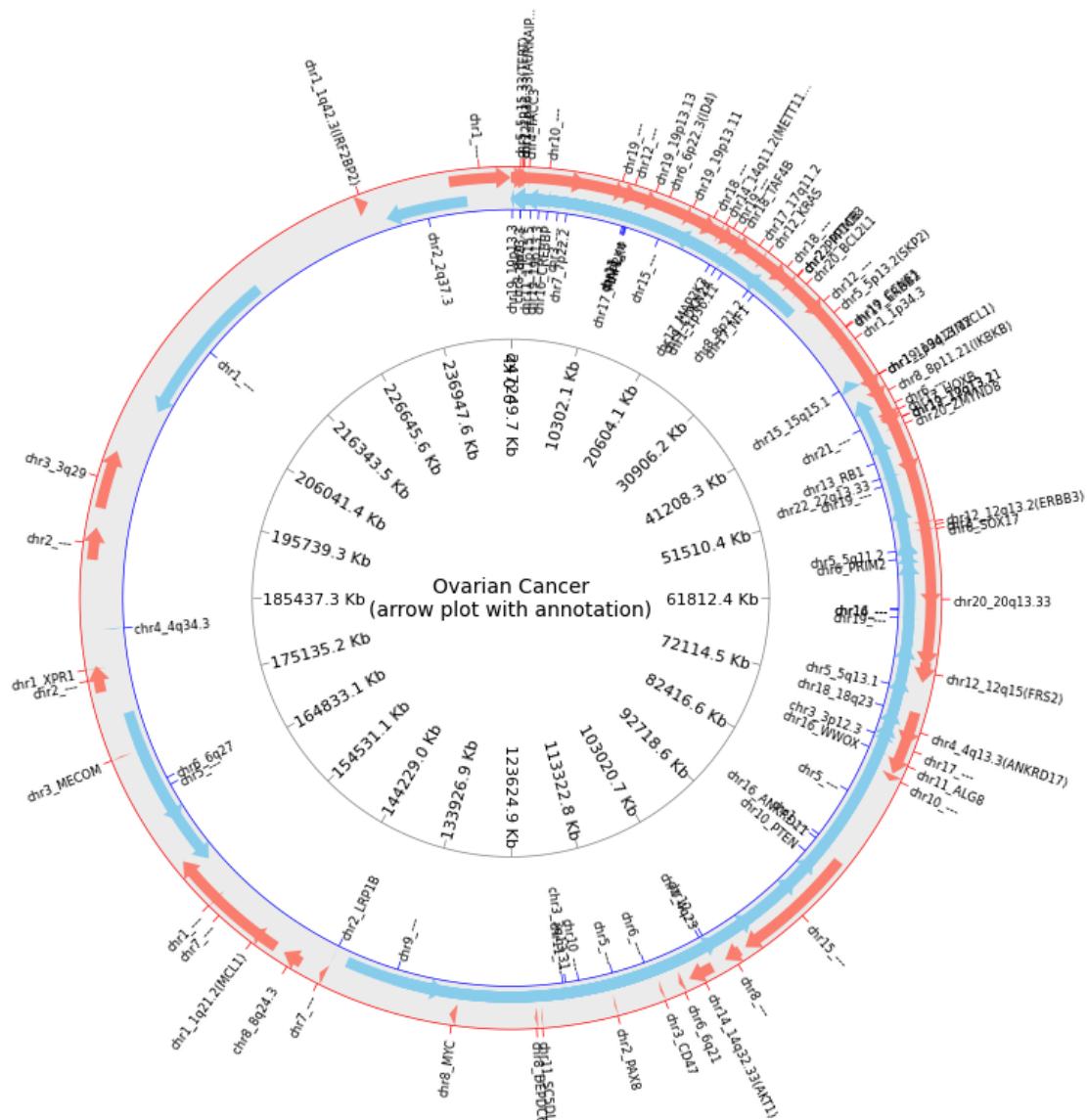
Final plots:











Represent and explain genetic information from cancer

- Right data structure
- right visualization

S3 s4(with functions, safer structure ) objects in R

- Q values are adjusted p-values that account for the false discovery rate (FDR) in multiple testing.

<https://www.google.com/search?client=firefox-b-1-d&q=r+S4>

<https://www.google.com/search?client=firefox-b-1-d&q=p+value+and+q+value>

