Visualizing Chromosomal Data for Cancer Detection

Zihe Zhang
Professor Subhajyoti De

Introduction

Cancer is fundamentally rooted in genetic mutations. By carefully examining chromosomal data, researchers can trace the origins of these mutations, potentially unveiling critical insights into cancer development. Chromosomes, the carriers of genetic material, undergo various mutations that directly influence cellular behavior and disease progression. Understanding the nuances of these changes is crucial, yet the complexity of genomic data poses significant visualization challenges.

Visualizing genomic data before proceeding with advanced analytical techniques like machine learning is vital. It allows researchers to identify and interpret patterns and anomalies within the genetic sequences effectively. This preliminary step ensures that subsequent analyses are based on accurate and clearly understood genetic information. The complexity arises from the vast amount of data and its intricate relationships, which are not immediately apparent without sophisticated graphical representations.

Visualizing genomic data is challenging due to its high dimensionality and massive volume, with each dataset containing potentially gigabytes of diverse data types such as DNA sequences and gene expressions [6]. These attributes require specialized visualization techniques to represent the data clearly and effectively. Further complicating visualization are the long sequences in genomic data, dynamic interactions among genetic elements, and inherent noise and variability. These factors need advanced, interactive tools that can accommodate multi-scale data exploration and provide clarity for researchers with varying levels of expertise in genomics.

There are primarily two types of genetic alterations relevant in the context of cancer: gene amplification and gene deletion. Gene amplification refers to the increase in the number of copies of a gene, potentially leading to overexpression. In contrast, gene deletion involves the loss of segments of DNA, possibly resulting in the absence of essential protein production. Both types of alterations can disrupt normal cellular functions and are key indicators of potential oncogenic processes. We will focus on specifically these two types of mutations.

Despite the availability of various tools and methods for analyzing genomic data, there remains a significant gap in standardizing these processes for efficient and clear visualization.

The final problem statement we have formulated is: "Developing a standard pipeline approach to visualize chromosomal data to better understand the biological components for non-invasive cancer detection." The objective is to find new techniques to handle and store large volumes of complex genomic data, preprocess the CSV formatted data into BED file format, and effectively utilize Circos plots to visualize ovarian chromosomal data. By achieving these goals, we aim to enhance the clarity and efficiency of cancer research, ultimately contributing to the development of non-invasive diagnostic tools.

Related Work

To understand the key benefits and trade-offs when using Circos plot to graph chromosomal data, we must reference the paper of “Circos : An information aesthetic for comparative genomics.” Although this paper is not lengthy, it has provided several key ideas: utility in genomics and comparative genomics, enhanced data interpretation, customization and flexibility, and applications beyond genomics [1]. Circos excels in representing genomic data due to its circular layout which is naturally suited for displaying the cyclical nature of genomes. This layout is particularly advantageous for comparative genomics, as it facilitates the visualization of connections between different genomic regions or between genomes of different species. This makes it easier for researchers to detect similarities and variations that are crucial for studies on evolution, species diversity, and genetic disorders [1].

The circular layout of Circos not only handles large genomic datasets effectively but also highlights relationships and patterns that might be missed in linear presentations. This feature significantly enhances data interpretation, making complex genomic data more accessible and understandable. Circos offers extensive customization options, allowing users to tailor the visualization to fit specific datasets and research needs [1]. This flexibility extends to the integration of various data types, such as gene expression levels and mutation data, providing a comprehensive view of the genomic information. While designed for genomic data, Circos's application extends to other fields requiring the visualization of complex relationships in circular formats, such as in network analysis and social sciences. This cross-disciplinary potential broadens the impact of Circos, making it a versatile tool for various scientific communities.

Another intriguing paper, titled “STRATAG: An R package for manipulating, summarizing, and analyzing population genetic data,” explores the use of S4 objects in R to efficiently manage complex genetic information. While the paper primarily introduces the STRATAG package rather than delving into our main focus of Circos plots, it provides valuable insights on alternative

methods for storing genetic data in R beyond the conventional data frame approach. This knowledge can inspire further innovations in handling large-scale genetic datasets effectively. The STRATAG package is designed for manipulating, summarizing, and analyzing population genetic data [2].

The core data structure used in the STRATAG package is the S4 object type known as gtypes [2]. This object class is specifically structured to manage genetic data effectively. The gtypes object contains multiple slots for storing genotypes, stratification schemes, and other optional data like sequences and descriptions. The gtypes S4 objects are compatible with several other popular genetic data analysis packages in R. There are functions provided within STRATAG to convert gtypes objects to other data types used by packages like ADEGENET, PEGAS, and PHANGORN [2]. This interoperability facilitates seamless transitions between different analyses without needing to reformat the data .Within the STRATAG package, the gtypes objects can be manipulated using standard R functions to accommodate various analytical needs. This includes subsetting, indexing, and altering stratification schemes using built-in methods designed for S4 objects. These functionalities underscore the flexibility and adaptability of the S4 object structure in handling complex genetic datasets.

The design of the gtypes S4 objects also focuses on efficient memory management, which is crucial for handling large genetic datasets [7]. The structure allows for efficient calculations and data manipulations, reflecting the advantage of using S4 objects in resource-intensive computational environments.

Progress

Understanding the data

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Chromosome**	**Wide peak start**	**Wide peak end**	**amp/del**	**2q3**	**Wide Peak Limits**	**Peak Limits**	**Region Lim**	**q values**	**Residual q**	**Description with number**	**Description**
2	**chr19**	34934008	35059410	Amp	19q12	chr19:34934008-35(	chr19:34981905-:	chr19:23143	1.89E-102	2.57E-96	**CCNE1 [1]**	CCNE1
3	**chr3**	170150840	170280412	Amp	3q26.2	chr3:170150840-17(	chr3:170249939-	chr3:149545	8.00E-63	3.28E-56	**MECOM [0]**	MECOM
4	**chr8**	128754508	129388953	Amp	8q24.21	chr8:128754508-12!	chr8:128809381-	chr8:100035	1.86E-79	3.70E-56	**MYC [2]**	MYC
5	**chr11**	77509837	77679169	Amp	11q14.1	chr11:77509837-77(	chr11:77579027-:	chr11:68349	9.28E-28	9.28E-28	**ALG8 [4]**	ALG8
6	**chr12**	25059809	25394729	Amp	12p12.1	chr12:25059809-25:	chr12:25311849-:	chr12:17822	6.39E-25	2.66E-23	**KRAS [4]**	KRAS
7	**chr19**	44455609	45045812	Amp	19q13.2	chr19:44455609-45(	chr19:44828320-:	chr19:23143	5.54E-29	7.57E-23	**19q13.2 [24]**	19q13.2
8	**chr8**	144180457	146072839	Amp	8q24.3	chr8:144180457-14(	chr8:145628324-	chr8:100035	1.57E-41	4.34E-20	**8q24.3 [75]**	8q24.3
9	**chr19**	12983113	14103663	Amp	19p13.13	chr19:12983113-141	chr19:13749627-	chr19:87065	5.66E-26	6.50E-18	**19p13.13 [29]**	19p13.13
10	**chr1**	39109153	41523062	Amp	1p34.2	chr1:39109153-415:	chr1:40147908-4(	chr1:350309	1.28E-21	6.50E-16	**1p34.2(MYCL1) [41]**	1p34.2(MYCL
11	**chr1**	148646436	149486905	Amp	1q21.2	chr1:148646436-14!	chr1:148763510-	chr1:120326	5.19E-18	1.45E-13	**1q21.2(MCL1) [28]**	1q21.2(MCL1)

Figure 1. The first few rows of the ovarian cancer patient's chromosomal data

Initially, our ovarian cancer data has multiple columns (Figure 1). It might seem overwhelming, but for our research, we will only be focusing on a few of the selected important features rather than all the information.

The dataset consists of several key features crucial for understanding genomic alterations and their implications in gene expression (Figure 1). The first feature is the “Chromosome Number” (Ex: chr1), which identifies the chromosome where the gene is located. This information is vital as it helps map the gene's exact location within the genome, which is essential for genetic linkage studies and understanding chromosome-specific diseases. The “Start” and “End” of Gene’s Relative Location indicate the starting and ending positions of the gene on the chromosome, respectively. They are integers. These positions are critical for precisely identifying the segment of the DNA where the gene begins and ends, facilitating genomic sequencing and annotation.

Another crucial column is the “Type of Alteration”, which specifies whether a genetic alteration is an amplification (amp) or a deletion (del). This distinction is important because different alterations can have varying biological impacts, such as influencing gene expression levels, which may contribute to disease pathogenesis, including cancer.

The “Q Value” is a statistical measure indicating the False Discovery Rate (FDR), representing the probability that a particular gene expression result is a false discovery. This value is derived from the p-value but adjusted for multiple comparisons in experiments. A lower Q value increases the confidence in the results, crucial for reducing false positives in large-scale genomic studies involving numerous tests.

Lastly, the “Gene’s Description” provides a brief overview of the gene's function, structure, or other relevant biological information, linking genetic alterations to potential phenotypic changes. This information aids in understanding the broader implications of genetic research and disease correlation.

Handling the Q value

As described earlier, the Q value indicates the statistical significance of observed gene expression. In our dataset, the Q values are extremely small; for example, some values are 1.89×10^{-102}, 8×10^{-63}, and 1.86×10^{-79}. To effectively compare these values, we calculate the negative logarithm of the Q value. A higher negative log Q value suggests increased confidence that the observed gene expression is significant and not due to chance. Depending on the type of Circos graph we aim to create, we have options for handling these values. We can normalize the processed data to have a constant mean and standard deviation, or we can set a maximum cap for the negative log Q value. Any value exceeding this cap will be fixed at the maximum.

These adjustments help improve the visualization by restricting the spread of the Q values, facilitating clearer and more effective graphical representation.

Storing information in BED file

#chrom	chromStart	chromEnd	name
chr1	39109153	41523062	-48.107426874943435
chr1	148646436	149486905	-39.79979797671503
chr1	232526850	233617929	-13.958226860165869
chr1	154088696	154329131	-19.36743068331123
chr1	36338557	36480225	-24.577019934250583
chr1	178780208	178854243	-6.5274028716890475
chr1	241440102	247249719	-7.603886908583042
chr1	1095521	1467046	-4.071253230998703
chr2	189030591	192069072	-9.57850969534065
chr2	28497864	28879745	-2.9718399811343885
chr2	113658200	113809678	-2.5262662888131637
chr2	176544739	179048787	-7.23607880669681
chr2	41813425	68291342	-2.486832503053504
chr3	170150840	170280412	-142.98341931694503

Figure 2. An example of BED filed created from the original chromosomal data

In genomic studies, such as our analysis of original ovarian cancer chromosome data, BED and CSV files are used distinctly, aligning with their specific strengths. BED files, tailored for genomic data, come with a predefined format that includes set columns such as chromosome, start, and end positions (Figure 2). This structured approach ensures efficient handling of genomic intervals but also means that the columns in a BED file cannot be customized or expanded beyond their initial specification. This makes BED files especially useful for tasks that require quick and straightforward access to genomic coordinates, such as mapping and sequencing.

Conversely, CSV files are general-purpose and flexible, allowing any number of columns to be added or adjusted as needed. While this makes CSV files less specialized for genomic data, their adaptability makes them useful for broader data analysis tasks that require integration with standard data processing tools.

Now, we need to extract the chromosome names, start location, end location, and negative log Q value or comments, and save them into BED files. After doing so, we can finally begin the graphing portion of the research.

Circos Graph with SRPlot

In our research, we utilized the SRPlot website, a pre-configured bioinformatic graphing platform, to create Circos plots. SRPlot offers an easy-to-use interface that simplifies the process of generating complex genomic visualizations, such as the Circos plot, which are essential for understanding the genomic architecture in our study of ovarian cancer[3]. While the platform streamlines plot creation and is accessible even to those with limited computational expertise, it does come with some limitations. The primary drawbacks include a lack of customization options for chart types and the addition of more tracks, which could enhance the depth of our analysis. Additionally, the inability to annotate genes directly within the plots restricts the detailed presentation of gene-related findings. Moreover, SRPlot is not open-source, which limits further customization and development by the research community, potentially affecting the adaptability of the tool to specific research needs.

After using the site, we created the following chart (Figure 3).

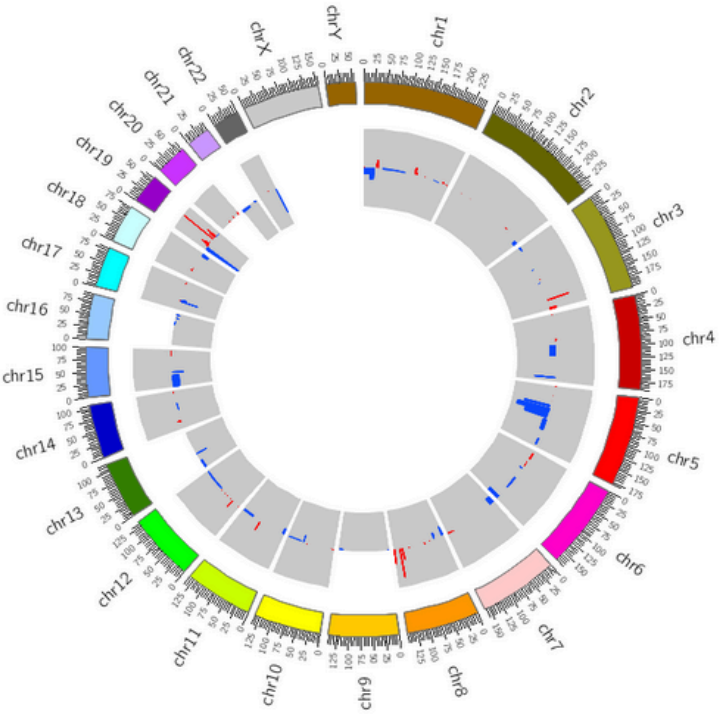

Figure 3. Circos plot created with SRPlot with ovarian cancer's chromosomal data

The blue bars or intervals represent the negative log Q values of gene deletions, while the red bars or intervals represent the negative log Q values of gene amplifications. One key observation is that for some chromosomes, there is no gray area to plot anything, just pure white space, indicating that we do not have any data associated with those particular chromosomes. With more comprehensive chromosomal data, we could fully graph the entire genome; however, with the current graph, we can still observe some intriguing patterns. For instance, on chromosome 2, we see almost no colored intervals from either amplifications or deletions. This might suggest that chromosome 2's gene alterations are not severe and have relatively less association with the mutations that caused ovarian cancer. On chromosome 5, we can clearly see a cluster of blue intervals, indicating a strong confidence that there is almost certainly gene deletion occurring around these intervals. Chromosome 19 could also be problematic; it has high Q values for both gene amplification and deletion, making it the most severely mutated chromosome in our dataset. We should pay more attention to this chromosome when making predictions or conducting other numerical analyses. Another point to note is that the maximum Y-value, the maximum height of the bars in each histogram, is uniformly set to be the highest negative log Q value of the entire population. This configuration might differ in later plots when using PyCirclize.

Circos Graph with PyCirclize

This time, instead of using a website, we are utilizing PyCirclize, a Python library specifically designed for creating Circos plots [4]. PyCirclize excels in customization and flexibility, offering robust support for multiple types of Circos plots, which is essential for visualizing complex genomic interactions and alterations [4]. Its extensive documentation makes it particularly user-friendly for researchers, facilitating the development of highly tailored visual representations of genomic data. This adaptability and comprehensive support enable precise and informative genomic visualizations, enhancing our understanding of the intricate genomic landscapes involved in our research.

First, we graphed the gene amplification and gene deletion individually to study the pattern within themselves. Next, we will put them together for comparison. Lastly, we will try to add comments and annotations to the gene expressions.

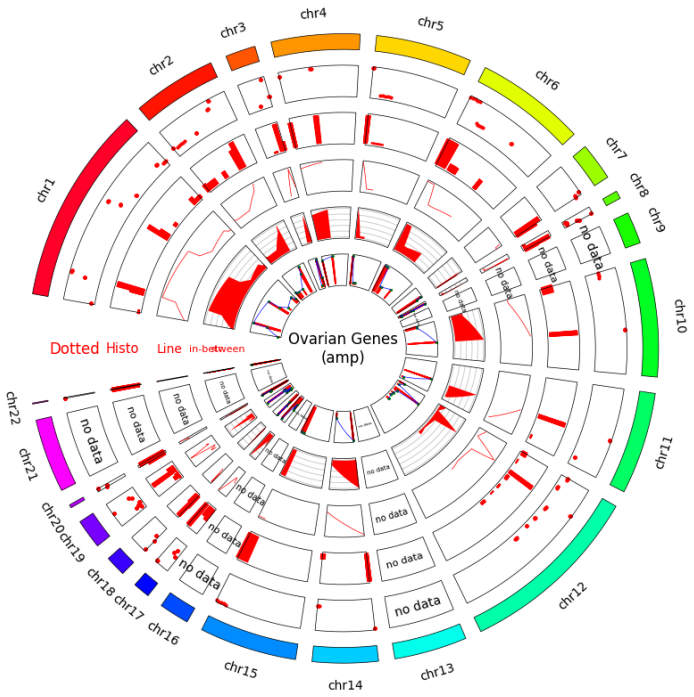

Figure 4. The amplification gene Circos plot created by PyCirclize

There are multiple tracks, and they are more visually appealing than the plot created using SRPlot earlier (Figure 4). Each track contains a different type of graph, but all ultimately depict the same data. For example, the first outer track is a scatter plot, followed by a histogram. Subsequent tracks include a line chart and a fill-in-between chart.

Let's focus on the histogram, which is the second outer track for now. The maximum height of the histograms is not the same for each sector of the chart. The maximum height of each sector is set to be the maximum value of the negative log Q value belonging to that specific chromosome. This setting helps us visualize the amplification with respect to each individual chromosome. In the previous SRPlot, there were almost no signs of amplification in chromosome 2, but here the graph still shows parts of chromosome 2 that might have a relatively higher chance of mutation, based on the location of the high red bars. Another observation is at chromosome 12, where almost the entire first half of the genome shows signs of amplification; however, the only high Q value part covers only a small portion of that interval.

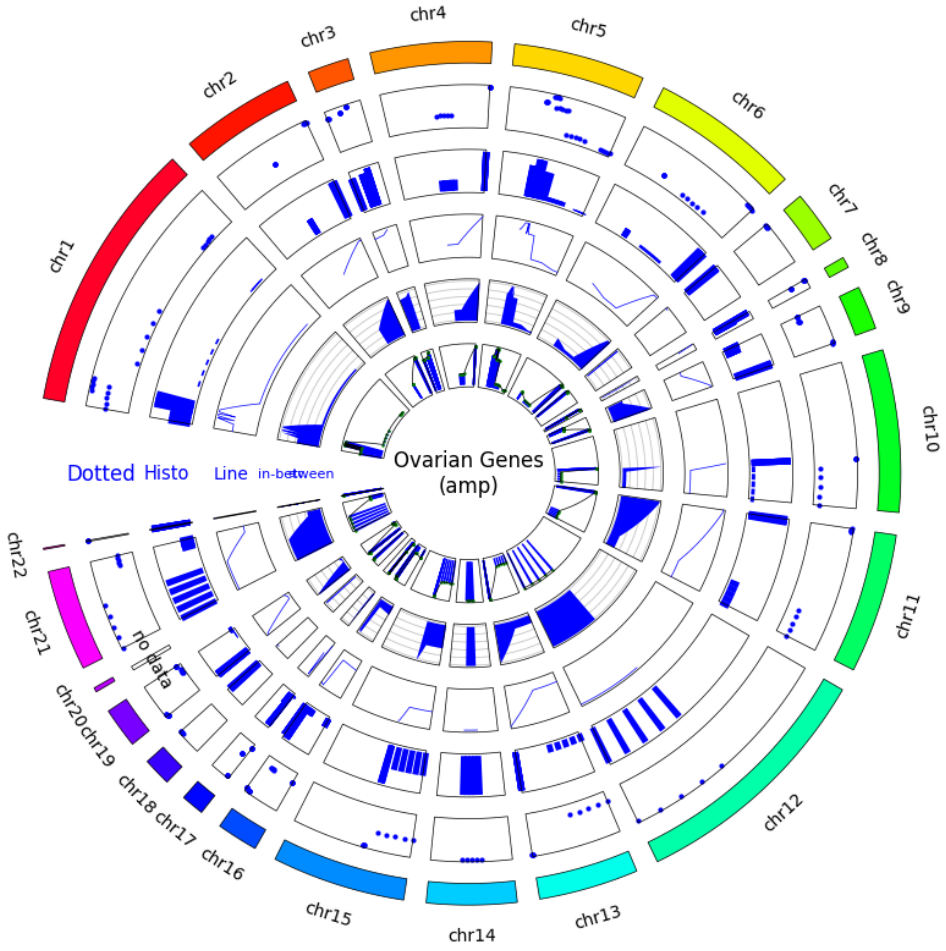

Figure 5. The deletion gene Circos plot created by PyCirclize

This Circos plot graphs only the genes with deletions, omitting any with amplifications (Figure 5). We have to pay attention to chromosomes 12, 15, and 21, all of which show continuously high Q values for more than half of the interval of the given chromosomes. For chromosome 1, there are only signs of deletion at the beginning of the genome; it is relatively safe for the later parts.

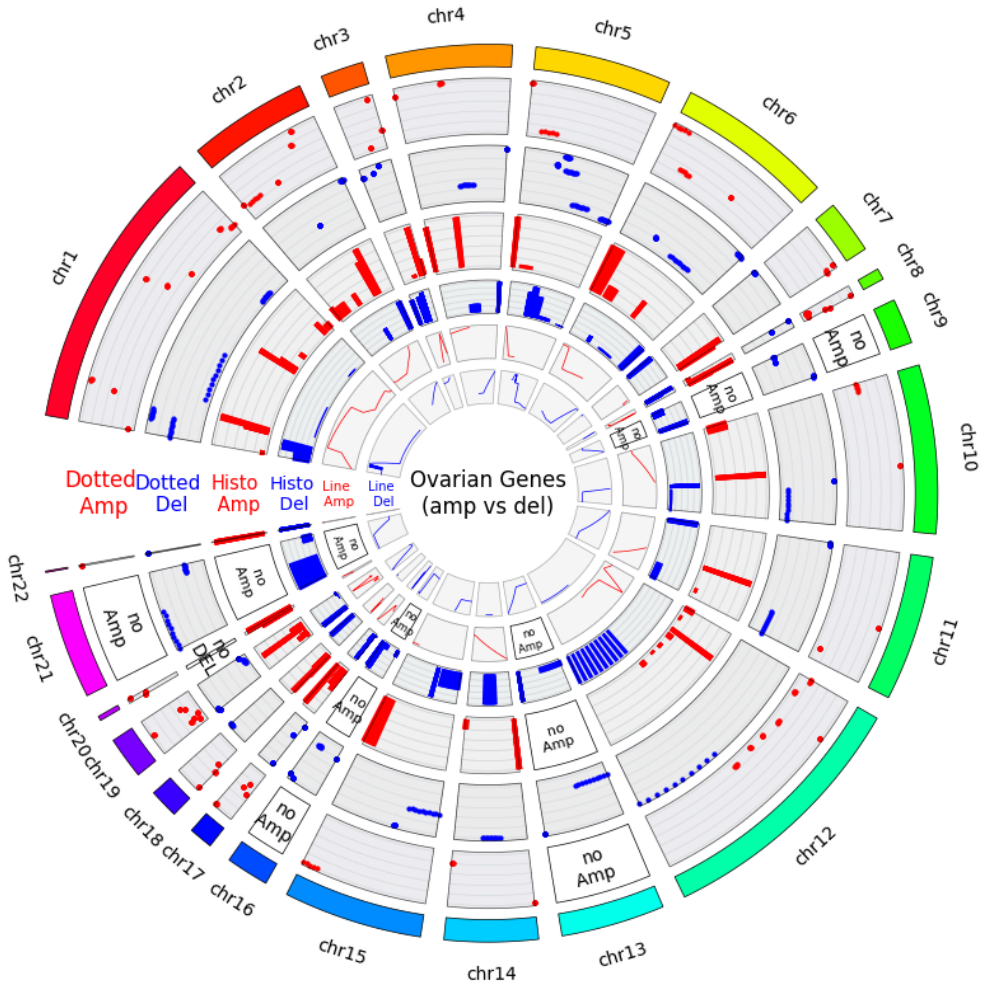

Figure 6. The amplification and deletion gene Circos plot created by PyCirclize

Remember that the red bars represent the genes with amplifications and the blue ones indicate deletions. From this comparison plot, we can easily compare the peaks of Q values for each chromosome (Figure 6). We will focus only on the histograms. For chromosome 3, there is relatively high confidence in observing both amplification and deletion. We need to conduct a more in-depth analysis of this chromosome. Although in the SRPlot the Q values (for chromosome 3) are not extremely high across all data, they are relatively high in the local area. For chromosome 12, the first half shows medium Q value confidence for amplification, and the second half has high confidence in deletions, pointing to another problematic chromosome. There are many other useful observations that could help us determine which specific chromosomes or pairs of chromosomes we want to investigate further. This could lead us on an easier path to uncovering the real cause of the cancer and the root of the illness.

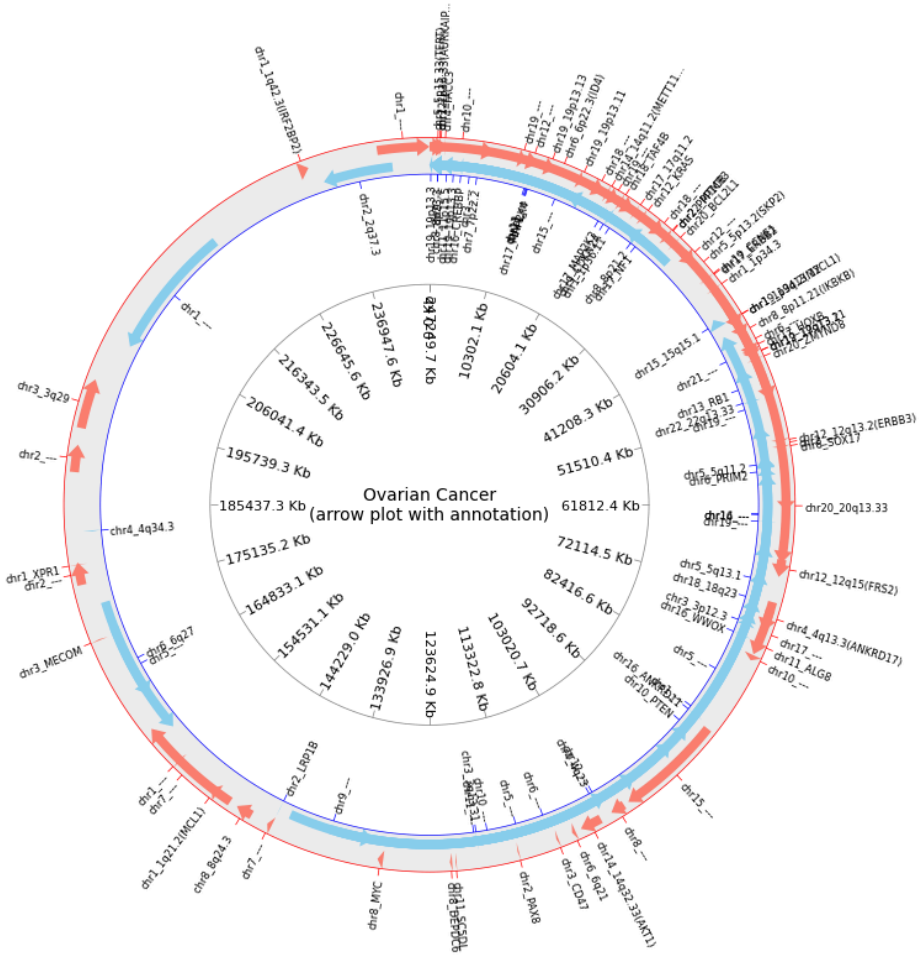

Figure 7. Annotated Circos plot for both gene amplification and deletion

This Circos plot represents our entire genome in a complete circle (Figure 7). The arrows indicate the intervals of particular gene expressions. Light blue signifies deletions, and red indicates amplifications. We noticed that some labels are left blank, for example, "chr15_." Not all genes are annotated in our data, but we can still derive some useful context. For annotations that include a number with a "q" and other integers, like "chr3_3q29," the "3q29" refers to a specific region on chromosome 3. For genes with actual genomic names and labels, such as "chr8_MYC," it is often difficult for us to understand what "MYC" means without consulting the genomic database. Having additional tools to help us gain more context would be greatly beneficial.

Conclusion

In conclusion, the Circos plot stands out as a powerful visualization tool for discerning patterns in genomic data. It effectively highlights the peaks and lows within zones of gene amplification and deletion, providing a preliminary indication of potential chromosomal issues through the contrast of amplification and deletion zones. Nevertheless, a more comprehensive analysis still necessitates the integration of numerical and machine learning prediction models to deepen our understanding.

While there are numerous Circos libraries available that support various programming languages, the journey from initial genomic data to effective visualization can be difficult. I have encountered challenges in navigating these early steps due to a lack of intuitive guidance. Therefore, there is a pressing need for more detailed user guides and documentation that helps specifically the new users who are not familiar with the biological data. Such resources would significantly lower the barriers to participate in this complex but crucial field of genomic research, ensuring that more researchers can leverage the full potential of Circos plots in their work.

This enhancement in support and resources would not only improve the learning process but also empower researchers to conduct more precise and insightful genomic analyses, advancing the field of genomics further.

Future work

In the future, my initial focus will be on mastering the creation of Circos plots using R. Given that R is a standardized software for processing complex genomic data, its use will substantially ease the integration of other standard genomic data processing modules if required. Beyond the decision of which platform to employ for graphing, we plan to enhance our graphing capabilities through integration with a dynamic web application. Developing an easy-to-navigate website will enable more workers in the biological field to access this graphing tool [8]. Users will have the capability to dynamically upload their chromosomal data. Perhaps, we could develop a site similar to SRplot—the previously used site for Circos plots—but with increased customization options, such as the ability to add multiple tracks, and manage colors and chart types, as well as to apply self-assigned annotations. Increasing flexibility and customizability in a genomic graphing tool could prove to be highly influential.

Furthermore, it is important to note that annotations pertaining to gene expression are typically presented in academic and medical terminology. These terms can be confusing and misleading without proper reference to the extensive genomic databases. The development of a tool that can translate these medical terms into more understandable language would greatly simplify the interpretation of genetic data. A large language model, when trained with the appropriate medical data, could perform exceptionally well [5]. Integrating the Circos plot with a large language model's text generation capabilities to explain gene expression could be highly beneficial [5].

Bibliography

[1] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. Genome Research, 19(9), 1639-1645. https://doi.org/10.1101/gr.092759.109

[2] Archer, F.I., Adams, P.E. and Schneiders, B.B. (2017), stratag: An r package for manipulating, summarizing and analysing population genetic data. Mol Ecol Resour, 17: 5-11. https://doi.org/10.1111/1755-0998.12559

[3] Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, et al. (2023) SRplot: A free online platform for data visualization and graphing. PLoS ONE 18(11): e0294236. https://doi.org/10.1371/journal.pone.0294236

[4] Moshi4. (2023). pyCirclize: Python library for creating circos plots (Version 1.4.0). GitHub. https://github.com/moshi4/pyCirclize

[5] Tariq, A., Luo, M., Urooj, A., Das, A., Jeong, J., Trivedi, S., Patel, B., & Banerjee, I. (2024). Domain-specific LLM Development and Evaluation – A Case-study for Prostate Cancer. *medRxiv*, 2024.03.15.24304362. https://doi.org/10.1101/2024.03.15.24304362

[6] Karen Y. Oróstica, Ricardo A. Verdugo. (2016). chromPlot: visualization of genomic data in chromosomal context, *Bioinformatics*, Volume 32, Issue 15, August 2016, Pages 2366–2368, https://doi.org/10.1093/bioinformatics/btw137

[7] Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*, 6. https://doi.org/10.3389/fgene.2015.00208

[8] Martín-Payo, R., Martínez-Urquijo, A., Zabaleta-del-Olmo, E. et al. (2023). Use a web-app to improve breast cancer risk factors and symptoms knowledge and adherence to healthy diet and physical activity in women without breast cancer diagnosis (Precam project). *Cancer Causes Control 34*, 113–122. https://doi.org/10.1007/s10552-022-01647-x