Index

Abstract

The paper begins by giving a brief background of what genomes are, describes their structure, and what a genome sequence is. It then talks about visualising these sequences – why it is done, and what such a visualization would show. It then dives into the different types of visualizations that exist, while classifying them based on their layout. The advantages and disadvantages of each visualization kind are discussed, along with examples.

Keywords

Data visualization, genetics, genome, sequencing

Introduction

Genomic sequencing has boundless uses in the world of science, especially in the field of medicine. Geneticists have by now managed to map out the entire human genome, along with genomes and genes from a vast variety of other organisms. This genetic data is computed and analysed and pored over by scientists to draw results, a lot of which is easier done by making visualizations out of it. Before talking about the actual visualizations and their kinds, the paper attempts briefly to explain what genomes and genetic data are, and how one gets to visualizing them.

Background

What is a genome?

A genome refers to all the genetic information stored in an organism. For most eukaryotic organisms, this is the sum total of all the DNA in the organism - both nuclear DNA as well as mitochondrial DNA. This genetic information defines everything about the organism - its physical structure, its biological processes, and its information passage to its progeny. The genome is akin to an instruction manual, defining the rules for each cell, on how to grow, live, and reproduce.

The size of an organism's genome is consistent across the species, but there is a stark contrast between different species. Some, such as small eukaryotic bacteria, have a single circular strand of DNA. Others, like Ophioglossum, a fern, can have 720 pairs of chromosomes [5] making up the genome.

While a genome *is* simply stored information, its medium of storage is quite different from what we as humans may be used to. This information is not carved into rocks, nor ink on paper. Neither is it stored in CDs or pen drives. Biology uses the most ingenious way of storing data; storing massive, massive amounts of information at a size even the smallest intel chip could never reach. On Earth, lifeforms store their genetic information in the form of DNA or RNA. These are polymeric chains of acid

molecules, stored in a highly coiled structure. This structure is further explained in the next section.

Understanding the structure of DNA

The structure of DNA (deoxyribonucleic acid) is actually remarkably simple, for a structure that stores such complex and copious amounts of data. At the most basic level, DNA is a polymer of deoxyribonucleic acids. Each monomer consists of a deoxyribose sugar, attached to a phosphate group and one of the four nucleotide bases: adenine, thymine, cytosine, and guanine (A, T, C, G). The former two form the backbone of a single DNA 'strand', while the bases form hydrogen bonds with their complementary bases (A with T, C with G) in a reverse DNA strand. Together, the two strands are curved into a double helix structure, the 'twisted ladder' image of DNA that's most common. This long strand of DNA is further coiled and looped, wrapped around proteins called histones, giving a 'pearls on a necklace' sort of appearance. This structure is called chromatin. Chromatins are further condensed, coiled into structures called chromosomes. Together, all the chromosomes are present in the nuclei of every cell of an organism.

Sequences and Sequencing

DNA hence consists of a long sequence of nucleotide base pairs. This sequence of ATCG is what is referred to as the DNA sequence. DNA sequences are what provide answers to so many questions – they help identify relations between organisms, provide evolutionary history, find sources and cures for diseases, understand behaviours, and so much more.

How a DNA sequence provides all these answers can be understood by how it functions. The entire DNA sequence is made up of several genes. Each gene consists of exons, which are the coding segments, and introns, which do not code for anything. The exons essentially code for all the proteins that the organism's body is made up of. The exons consist of multiple codons – triplets of nucleotides (for example, AAA, AGT, CGC, etc.). Each codon codes for a particular amino acid, the building blocks of proteins. In this way, an entire protein can be made, by assembling together the different amino acids which the gene encodes.

The process of 'sequencing' is used to determine the sequence of the DNA. The first successful DNA sequencing was done in the 1970s, using a process called two-dimensional chromatography. Today, easier, more automated, and far quicker methods are used for the same. In one method, entire genomes are cut into smaller fragments, which are then sequenced parallelly in DNA sequencers using next-generation sequencing methods. Another method called large-scale sequencing or de novo sequencing uses viral vectors to sequence long DNA sequences, like entire chromosomes. These are both together termed high throughput methods,

How are these sequences stored?

For sequenced DNA to be of use, it needs to be stored and shared. This is done through databases called genome browsers. A genome browser is any graphical interface that stores and displays genomic data. Popular browsers include NCBI (National Centre for Biotechnology Information), UCSC (University of California Santa Cruz) and Ensembl, a joint project between the European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL), and the Wellcome Trust Sanger Institute. Such browsers not only store the data in plain text formats like .jsonl, but also usually visualize the data. These visualizations usually depict different 'tracks' or features of the genetic data, which can be turned on and off as per the user's needs. Some of the features you can see include the actual genes, genetic variations, protein alignments, etc.`

Example of visualization, by genome browser Ensembl [10]

Visualising DNA sequences

As such, several methods of visualization of genetic data exist. Each has its own purpose, its own benefits and shortcomings, which are discussed in this section. But before diving into the different kinds of data visualizations, it may be beneficial to understand why we need to visualize the data in the first place, and what the visualizations are trying to show.

Why visualize

Abstracting and visualizing data, to understand it, to make a record of it, or to share it with others, is an age-old practice – even cave paintings were visualizations of familial data, or of dangers one might encounter ahead. Today, when we think of data, we think differently. We think of strings of numbers, of code. We think of statistics and studying trends. That is not necessarily all that data is, nor are bar graphs and pie charts the only kinds of visualizations that exist.

Regardless of what the data is, visualizing it can help in so many ways. Some visualizations tell a story, explain a narrative. They help the person visualizing convey what they want, in a medium that is more easily understood. Visualizations also help order and structure data, making it not just neater and easier to understand, but also make it possible to identify patterns, and create meaning. Existing data helps create projective visualizations, enabling us to talk about the uncertain with a

certain predictive confidence. Further, humans are visual creatures. It is far easier for us to understand a visual representation of information, than to think of it otherwise. Moreover, a lot of the data we deal with today is simply unfathomable, and visualizations help us as humans to not only perceive it, but also make sense of it.

For DNA in particular, visualizations are especially helpful. Sure, one could argue that even the string of ATCGGTAGACTAAACGATCGGGACT is a visualization in itself – it is abstracted, we are representing the nucleotide bases with lines and curves that make more meaning to us than seeing the actual bases. However, further visualizations of these sequences is what the paper is referring to. These visualizations are of use in a variety of fields, like medicine, biotechnology, forensics, virology, etc. Genome visualizations help us understand the causes of diseases, as well as find cures. They help establish relations between different organisms, based on commonalities in their genomes. They help understand better the function of DNA itself, help us unravel the mysteries of how it functions, why it is so long, what the non-coding parts do, and so on. Genomics is a field with massive uses in today's world, and visualizing the genomes helps in every aspect, making every analysis easier and more efficient.

What does a visualization show?

Genome sequence visualizations show genomic data, ie, DNA sequences, at different focal lengths, for different purposes. Linear visualizations like those in genome browsers show different lengths of sequences, depending on their zoom level. They can show up to the highest detail, going as far as showing the ATCG sequence itself. A little zoomed out, they can show the different areas in a sequence, allowing us to identify and distinguish between coding and non-coding areas, look at different genes and their distribution across the chromosome, compare lengths of different genes or chromosomes, contrast and compare different tracks of a gene, look at mutations, identify promoter regions/VNTRs/other parts of the sequence, etc. On the other hand, 'genome maps' or circular visualizations of genomes usually consist of the entire genome of an organism (usually a prokaryote with a small, circular genome). Despite being smaller than eukaryotic genomes, these sequences are still atrociously large, and need to be aggregated or segmented to be viewed. These visualizations also show multiple tracks in parallel, such as tracks showing gene annotations, epigenetic signals, or gene expression. Some visualizations are better for showing biological information, like untranslated regions, binding sites, promoters, etc. Others help show interconnectivity, and establish relations between different parts of a sequence. Another common kind of genetic visualizations are heatmaps - which show really only one track, but are highly successful in showing different concentrations of the feature across the DNA sequence. Ultimately, what a visualization shows depends on various factors, like its level of detail or focus, its layout, and its function.

Types of visualizations

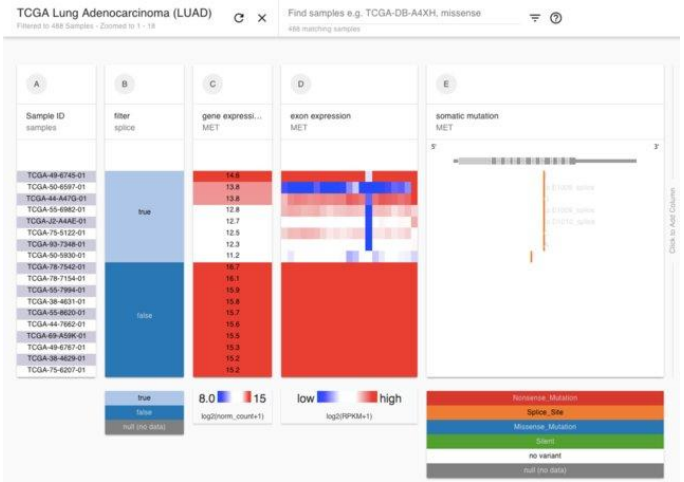
The vast variety of existing genomic visualizations can be classified in several different ways. Purposes overlap, complexities can vary even within different kinds of visualizations, and chronology is not particularly helpful. Therefore, in this paper, the visualizations are categorized simply by their layout – one of the first things one notices when they see the visualizations. The two major categories in this are linear layouts and circular layouts. Two other fringe categories also exist, which are space-filling curves and spatial arrangements.

Linear

As the name suggests, these kinds of visualizations are laid out linearly. They may have one or more axes. These visualizations are the most common kind, popular enough to be the only kind used in most genome browsers. These are also the easiest and most intuitive for us as humans to read, as they generally flow linearly from left to right, the same way that most cultures read. Linear visualizations may show a single track, or depict multiple parallel tracks, depending on the purpose. Following are some linear genetic visualizations:

Genome Browsers

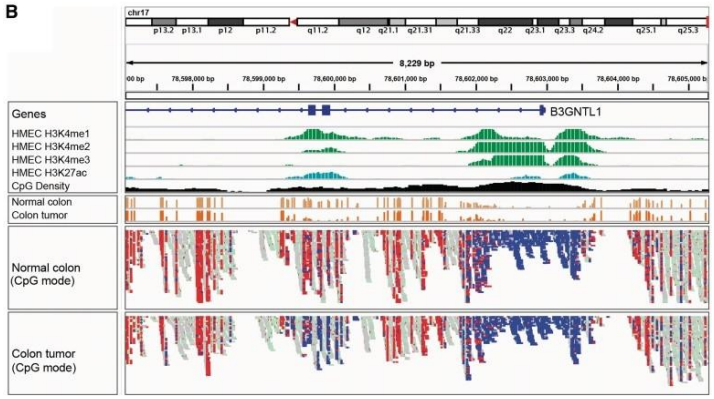
Genome browsers depict the DNA sequence being looked at in their own style, each particular to different browsers. Most of these have several commonalities. They generally display parallel tracks of information of the selected sequence. These could include the sequence of the DNA itself, the different genes it codes for, the phenotypes it is associated with, the proteins it can make, parts of the sequence like exons/introns/binding sites etc. These also often show a lot of metadata, more than other visualization kinds. For example:



A screenshot from the USCS Genome Browser [3]

The above image from the USCS Genome Browser shows 5 different tracks for a particular length of DNA. The first shows the different genes associated with TCGA lung adenocarcinoma. The second returns a true/false value for whether the genes are spliced. The third shows the level of gene expression, while the fourth shows levels of expression of different exons within each gene. The last column shows the mutations found within that length of DNA per gene. The linear layout along a horixontalk

axis enables us to see 5 different degrees of information for each of the genes mentioned, associated with the particular DNA.

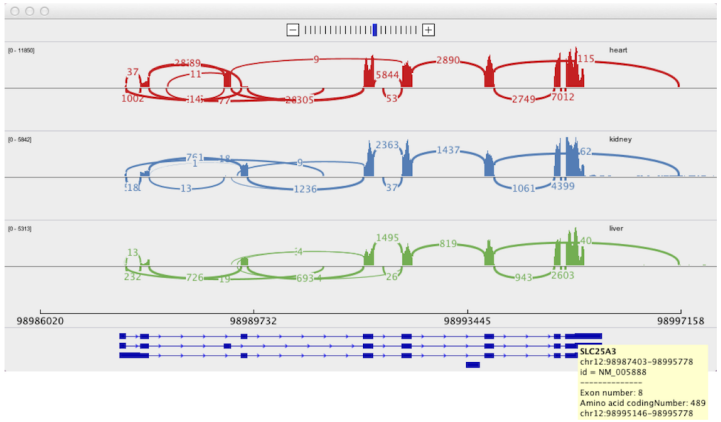


A screenshot from a Whole Genome Bisulfite Sequencing attempt using IGV [13]

In this image, contrary to the last, different tracks are laid out vertically, as opposed to horizontally. At the very top, we see bands on the 17th chromosome, providing not only metadata on what we are looking at, but also an idea of where the different genes are located on that particular length of DNA. Next, we see quantitative data on exactly how many base pairs are present on different lengths of the above DNA. Further below, we see expression for different HMEC genes present in this DNA. Next, we see a comparison of expression between the same length f DNA, between a normal colon cell and a tumorous colon cell. Finally, we see the difference in methylation of these two. This visualization is cancer-specific. Oncological research forms a large section on genomic research.

Sashimi Plot

In genomics, splicing is the process of removing introns from a segment of genetic material, keeping only the coding exons. Sashimi plots are a form of linear visualizations that use spliced DNA to show interconnections between different exons.



A Sashimi Plot [2]

Sashimi plots are generally used to display isoforms of the same gene. They use different colours to differentiate between the isoforms, without the colours having any semantic value. The sashimi plot of each isoform consists of 3 parts: a horizontal line, histograms over this line, and arcs connecting different parts of the line. The line itself is representative of the gene. The histograms are placed over the exons, the length of the bar representing the read frequency of the exon (read frequency being the number of times an exon comes up when the DNA is sequenced). The arcs correspond to the junction reads between

exons. These arcs have a labelling number, which tells the junction read frequency of the exons connected by the ends of the arc. The stroke width of the arc is also proportional to the number of junction reads that span these exons [4].

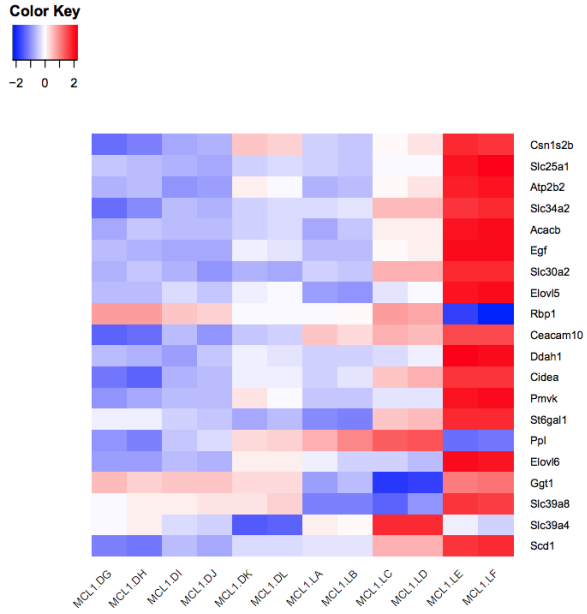
In the above image, the isoforms of the gene SLC25A3, found on the shown segment of chromosome 12, are plotted. One can see the read frequencies of the different coding exons of the gene In this plot, there are also three lines at the bottom with the positions of the exons marked, in dark blue.

Orthogonal Plots

This section describes two different kinds of orthogonal plots that one may encounter in genomic research. While the sashimi plots mentioned earlier did have an orthogonal component (the histograms), the whole plot ran along one axis. In these orthogonal plots, the presence of two right-angled axes forming a matrix is quintessential to the function of the plots. Following are the two different kinds:

Heatmaps

Heatmaps use colour-coding to depict interactions. They generally use different saturation levels of the same colour. Generally, one axis represents the genetic information - could be different chromosomes, different genes, or different parts of a gene. The other axis sometimes represents another value, like different samples, different organisms, etc. Other times, the perpendicular axis also shows the same genetic material, and the plot studies a feature of that gene, like its expression [11]. For the latter kind, heatmap plots are specifically useful because of the sheer vastness of the raw data – this data can be clustered into bins when being represented in this plot. A sought-after feature in heatmaps is the ability to choose scale, as it enables one to look at the interactions on different levels, from the chromosome level down to its nucleotide bases.

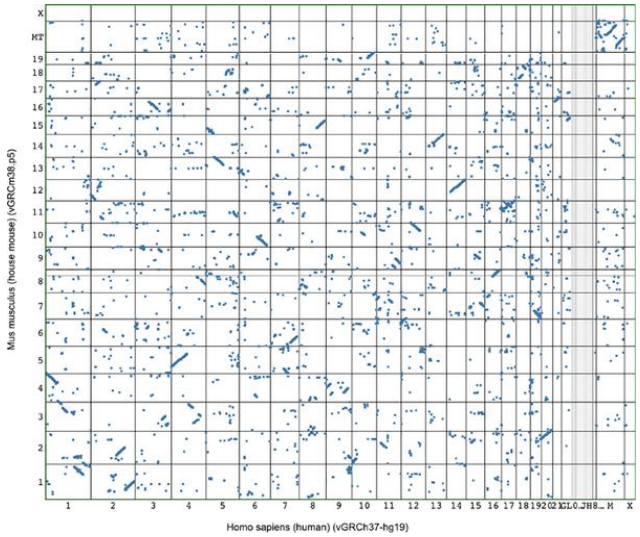


A heatmap comparing expression of genes in lactating mice vs pregnant mice [12]

The x-axis in this has the codes for each of the mice that the sample was taken from, from which the genes were extracted. Each column is hence representative of one mouse. The codes MCL1.DX are for the pregnant mice, while MCL1.LX are for the

lactating mice. On the y-axis, we have the different genes of which we are studying the expression. Bright red being a very high level of expression, while bright blue being a low level of expression. From the resultant matrix, we can clearly see that the genes in question are expressed much more in lactating mice than in pregnant ones.

Scatter Plots



[image: scatter plot] A scatter plot showing synteny between the human genome and the mouse genome [8]

Scatter plots are fairly common in multiple domains. Even within genetics, they may be used for different purposes. One of the purposes is to study synteny across different genomes [1]. Two or more genomic regions are considered ‘syntenic’ if they are derived from an ancestral genomic region. To compare, two different genomes are plotted, one on each perpendicular axis. Dots mark regions of commonalities across the board, that is, each dot refers to a gene the chromosomes of the different genomes have in common. Diagonal lines formed by co-linear dots mark regions of synteny.

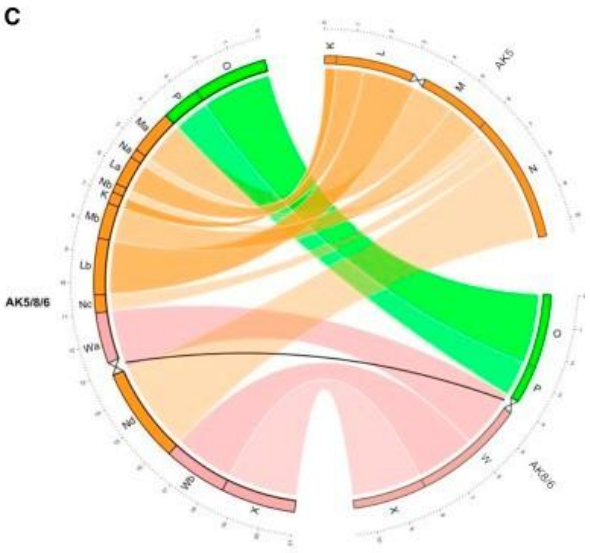
Circular

Circular plots show the genetic material arranged in a circle, along the circumference. These differ from linear plots in mainly the zoomability. While linear layouts are often interactive and can be zoomed into to see higher levels of detail, circular plots do not offer this option. They also tend to show more of an overview, often even show entire genomes.

To show multiple tracks in circular plots, one method is when the circumference is divided into arcs, each depicting one track (series arrangement). The other method is parallel arrangement, in which concentric circles are used, each circle coding for a particular track. In circular plots, it is common to see inter-relations depicted by connecting different parts of the circumference with arcs within the circle.

In the image, the authors take homeologues AK5 and AK8/6 from one set of plants, and compare it to the chromosome AK5/8/6 from another set of plants. The circular plot has the two smaller homeologues on one side of the circumference, while the larger chromosome forms the other side of the circumference. The bands within the circle then show interconnections between

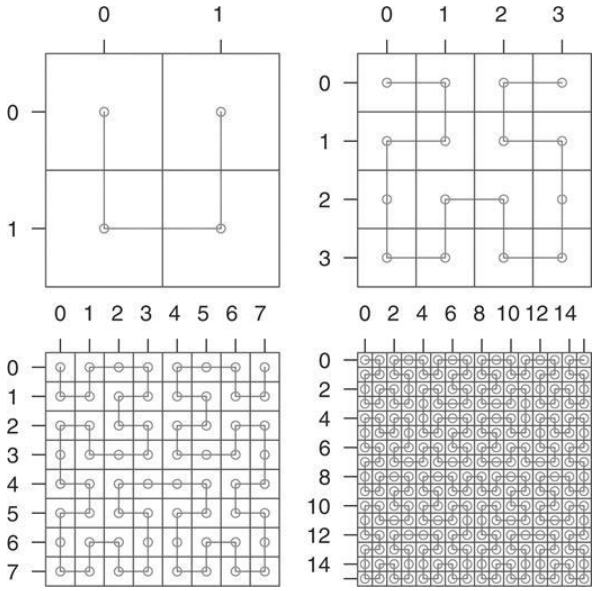
these chromosomes, highlighting parts which are the same/similar. The overall visualization tries to show AK5/8/6 as a combination of AK5 and AK8/6, while also highlighting certain structural differences between them.



A circular plot showing collinearity between homoeologs in Cardamine amara and Cardamine rivularis and a chromosome in the Cardamine pratensis and in C. × schulzii [6]

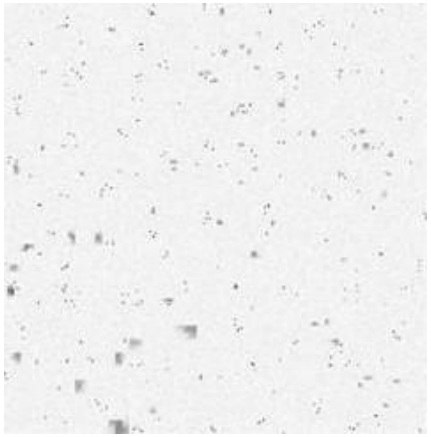
Space-filling Curves

In stark contrast to linear or circular layouts are space-filling layouts, like visualizations that follow the Hilbert curve. The advantage these have over the traditional linear layouts is primarily that they show sequences that are close in real life, as close in the visualization (understandable from the given figure showing how a Hilbert curve is structured). This is not possible otherwise, because the sequences are so long, that they have to be stacked parallely. The disadvantage of such curves, however, is that showing more than one feature set is very hard, since these arrangements do not allow parallel stacking.



Structure of a Hilbert curve [9]

Hilbert curve visualizations usually use a single colour or shades of a single colour to show peaks along the curve. The data itself is binned first, and then coloured pixels represent the peaks. Peaks close to each other in the sequence means that those peaks are also closer to each other in the visualization. The saturation or darkness of the colour represents the height of the peak.



Genomic visualization along a Hilbert curve layout [9]

In the given visualization, we can see peaks of data represented by dark grey along the plot. We can see that some peaks occur close together, where the dark pixels are clustered, while some are randomly distributed. A higher occurrence of dark pixels in the bottom left quadrant shows that the second quarter of the genome has the most peaks.

Conclusion

The paper discusses the different types of data visualization kinds that already exist for genome sequences. The paper also shows an amateur attempt at classification of these visualizations, on the basis of their layout. Different visualization kinds are described with examples of each. It is clear from this paper that a vast variety of visualization kinds exist for genomic sequences – with none being particularly better or worse than the other – which kind one uses can be decided based on the advantages and disadvantages of the visualizations. Genomic research is making leaps and bounds, tackling new problems every day, and visualizing the computed data helps infinitely in making it more understandable.

References

1. Asher Haug-Baltzell, Sean A Stephens, Sean Davey, Carlos E Scheidegger, Eric Lyons, SynMap2 and SynMap3D: web-based whole-genome synteny browsers, Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages 2197–2198, https://doi.org/10.1093/bioinformatics/btx144
2. Broad Institute and the Regents of the University of California. (n.d.). Sashimi plot. Sashimi Plot | Integrative Genomics Viewer. Retrieved March 31, 2022, from https://software.broadinstitute.org/software/igv/Sashimi
3. Goldman, Mary, et al. "The UCSC Xena platform for public and private cancer genomics data visualization and interpretation." biorxiv (2019): 326470.
4. Katz, Y., Wang, E. T., Silterra, J., Schwartz, S., Wong, B., Mesirov, J. P., ... & Burge, C. B. (2013). Sashimi plots: Quantitative visualization of RNA sequencing read alignments. arXiv preprint arXiv:1306.3466.
5. Khandelwal S (March 1990). "Chromosome evolution in the genus Ophioglossum L.". Botanical Journal of the Linnean Society. 102 (3): 205–17. doi:10.1111/j.1095-8339.1990.tb01876.x.
6. Mandáková T, Kovarík A, Zozomová-Lihová J, et al. The more the merrier: recent hybridization and polyploidy in cardamine. Plant Cell. 2013;25(9):3280-3295. doi:10.1105/tpc.113.114405
7. Nielsen, C., Cantor, M., Dubchak, I. et al. Visualizing genomes: techniques and challenges. Nat Methods 7, S5–S15 (2010). https://doi.org/10.1038/nmeth.1422
8. Nusrat, S et al. “Tasks, Techniques, and Tools for Genomic Data Visualization.” Computer graphics forum : journal of the European Association for Computer Graphics vol. 38,3 (2019): 781-805. doi:10.1111/cgf.13727
9. Simon Anders, Visualization of genomic data with the Hilbert curve, Bioinformatics, Volume 25, Issue 10, 15 May 2009, Pages 1231–1235, https://doi.org/10.1093/bioinformatics/btp152
10. The Public Engagement team at the Wellcome Genome Campus. (2016, January 25). How are sequenced genomes stored and shared? YourGenome.org. Retrieved March 31, 2022, from https://www.yourgenome.org/facts/how-are-sequenced-genomes-stored-and-shared
11. Zuguang Gu, Roland Eils, Matthias Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, Bioinformatics, Volume 32, Issue 18, 15 September 2016, Pages 2847–2849, https://doi.org/10.1093/bioinformatics/btw313
12. Maria Doyle, 2021 Visualization of RNA-Seq results with heatmap2 (Galaxy Training Materials) https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-heatmap2/tutorial.html

13. Lee, I. et al. (2019). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. bioRxiv. doi:10.1101/504993v2

‘