Understanding large and complex biological data sets using visualization

Stephen Taylor (Radcliffe Department Medicine, MRC Weatherall Institute of Molecular Medicine, UK)

Molecular biology experiments are generating an unprecedented amount of information from a variety of different experimental modalities. DNA sequencing machines, proteomics mass cytometry and microscopes generate huge amounts of data every day. Not only is the data large, but it is also multidimensional. Understanding trends and getting actionable insights from these data requires techniques that allow comprehension at a high level but also insight into what underlies these trends. Lots of small errors or poor summarization can lead to false results and reproducibility issues in large data sets. Hence it is essential we do not cherry-pick results to suit a hypothesis but instead examine all data and publish accurate insights in a data-driven way. This article will give an overview of some of the problems faced by the researcher in understanding epigenetic changes (which are related to changes in the physical structure of DNA) when presented with raw analysis results using visualization methods. We will also discuss the new challenges faced by using machine learning which can be helped by visualization.

Introduction

Modern science institutes generate gigabytes of data. For example, a single human genome can now be sequenced in a day, generating 1.5 GB of raw DNA information, and it is becoming increasingly easy to generate data from multi-omics experiments. For example, RNA Seq (sequencing the transcriptome), chromatin immunoprecipitation (ChIP)-Seq (sequencing to understand epigenetics), proteomics and metabolomics experiments are becoming routine. Imaging experiments are also adding to this data accumulation – in the fields of spatial transcriptomics or proteomics, generating several gigabytes per square millimetre of sample image analysed. Data collection is now not an issue. What is becoming difficult is how to sensibly process and integrate all the data and make it comprehensible to understand the underlying mechanisms and function.

To gain insight from these large data sets, having the right tools to process and visualize the data is essential. Imagine trying to understand a protein structure from raw X-ray crystallography output or understand a gene regulatory network without being able to draw a graph. These views don't represent true reality but are abstractions of something more complex to allow the scientist to build a mental model of the complex biological process.

The focus of this article will be why visualization is important in understanding complex information generated as part of bioinformatics analysis. To illustrate this we will look at some of the steps that are used in studying the epigenetics and chromatin structure highlighting the pros and cons of high-throughput bioinformatics analysis and where visualization can help.

Understanding the non-coding genome

Where do proteins bind to DNA?

Gene expression is modulated by epigenetic processes that are not yet fully understood. Broadly, proteins bind to the non-coding elements of DNA which causes different regions to interact with each other in 3D space, and in a tissue-specific manner to control gene expression.

Trying to understand the complex structure of elements that bind to these regions on the genome is a complex task. Genome browsers are a great example of visualizing huge amounts of complex data and are commonly used as a first step in visualizing genomic data. For example, a chromosome browser such as the UCSC genome browser is essentially a 2D representation of a complex 3D, 2-metre-long molecule of tightly packed DNA (see Figure 1). This makes such information-rich data palatable for the molecular biologist who wants a basic understanding of the genome.

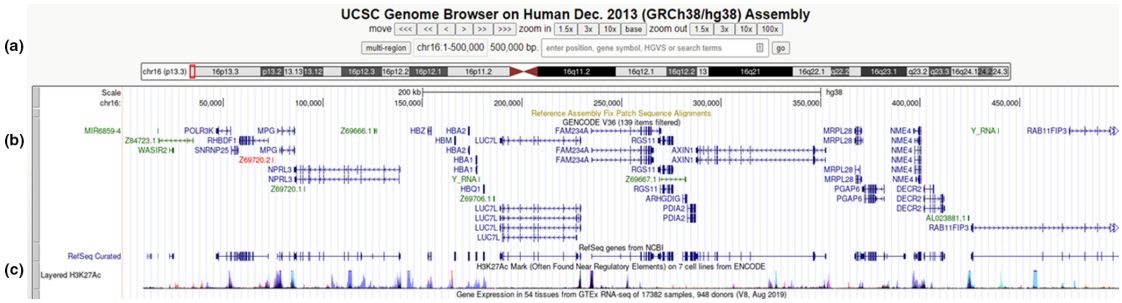

Figure 1. The UCSC Genome Browser shows a subset of the genome in a web browser and serves as a reference map for visualizing the results from a variety of experiments generated around the world. Here you can see a region of chromosome 16 from 1 to 500,000,000 bp which is associated with α-globin formation in the blood. (a) Overview of chromosome. (b) Gene locations and direction of transcription. (c) ChIP-Seq data shown as series of peaks where proteins bind.

To uncover where proteins bind to DNA, ChIP assays are used to find the so-called chromatin marks. ChIP-Seq involves using an antibody that is specific for the protein of interest. In the assay the protein is cross-linked to the DNA it binds to and then the DNA is sheared. The antibody can be used to pull out just the regions where the protein binds and the other non-bound regions discarded. The sequences in the bound regions can be sequenced and then mapped back to the genome, and these can be visualized on a genome browser as an area of enrichment. On a chromosome browser this is visualized as a peak whose size is proportional to the number of reads mapped (see Figure 1c). This is an example of taking semi-structured data – the DNA sequence – and making it structured by mapping it to a standard coordinate system so it can be visualized and queried.

There are many different chromatin marks that can be assayed, all giving orthogonal views to precise binding site locations of enhancers, promoters, open chromatin regions and CTCF sites, the latter mediating chromatin loops in selected tissues. These are extremely useful in understanding the non-coding regions of the genome that modulate gene activity. They potentially offer new ways of understanding the fine tuning of gene expression and hence are also potential drug targets.

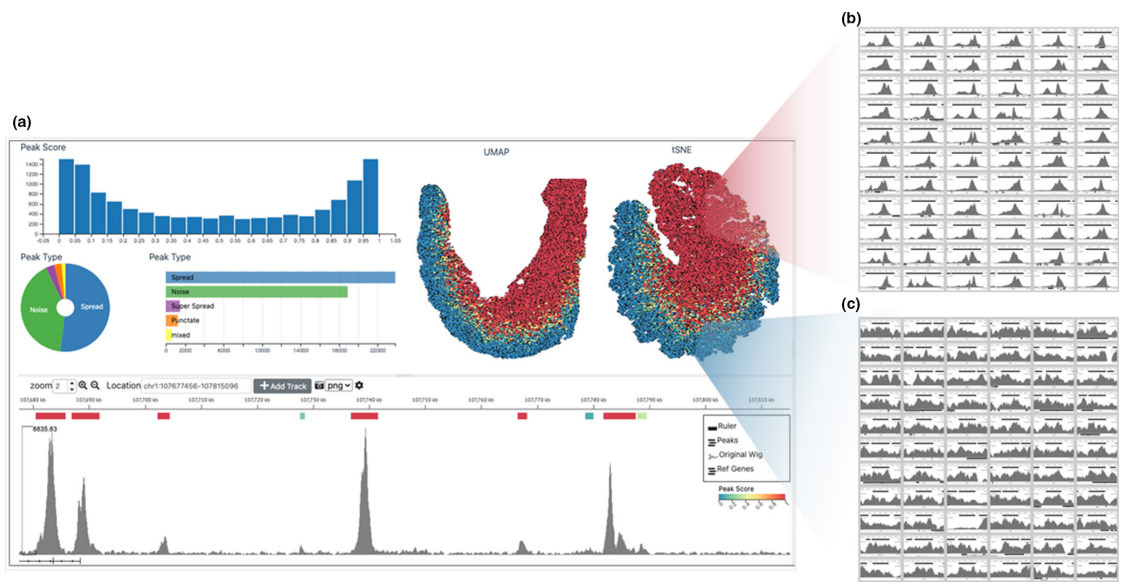

Figure 2. Example of Lanceotron output visualized in MLV showing thousands of items from an experiment. (a) Peak calls are visualized with an interactive bed file, charts, clustering and linked genome browser. Filtering can be applied using MLV or any other criteria based on a column in the interactive bed file. (b and c) Peak calls retrieved from ENCODE and converted into a feature vector. Every peak is then compared to every other peak which creates an *n*-dimensional matrix. This is visualized using UMAP and tSNE which are both techniques to reduce dimensionality of big matrices to a 2D graph coloured by the score generated by Lanceotron. These are shown as red (high peak score) to blue (low peak store). The image thumbnail panel for **b** shows high and **c** low scoring regions. The peaks in **b** are visibly much clearer than **c**.

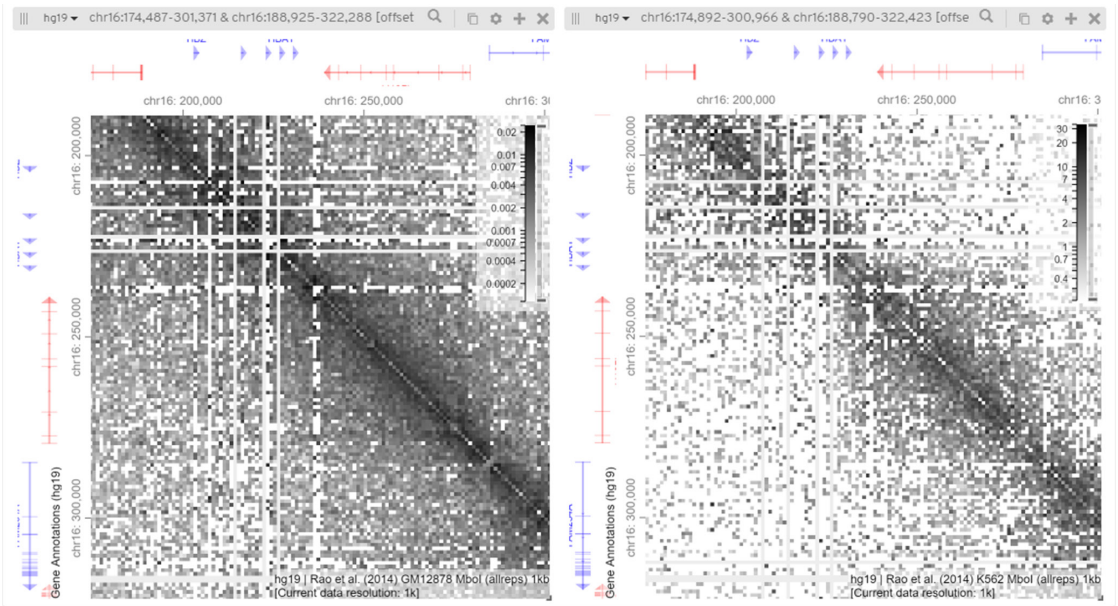

Figure 3. Hi-Glass Contact Map visualization of (a) B-cells (white blood cells) vs (b) K562 lymphoblastoid cells (red progenitor blood cells) in the α-globin region. The darker, more dense display shows the interaction frequencies of chromosome regions (shown on *x*- and *y*-axes) occurring in the genome in white blood cells (left) vs red progenitor blood cells (right). This is interpreted as α-globin is highly expressed in (b) and the chromatin is in a more open configuration compared with (a) so the transcriptional machinery may access it.

Depending on antibody affinity, experimental conditions, experiment design and depth of sequencing chosen, there may be thousands of potential candidate peaks that can be assessed computationally to determine whether they are true binding regions or background noise. Traditionally bioinformatics tools, called 'peak callers', are used to find areas of enrichment. These peaks are compared to background levels using statistical models and if they pass a threshold they are deemed a biological signal. Peak callers often overcall however, and many peaks may in fact be false positives. In addition, protein binding regions can often occur in open chromatin where the background signal may be higher – in many cases this results in false negatives with the peak caller ignoring these peaks.

It has been observed that humans are adept at classifying true peaks based on their shape. This is an example where we have a mixture of structured data (positions on the chromosome) but there is semi-structured data associated with it (data that represents the abstract shape based on how the reads map to the genome). Given there are many thousands of peaks to assess, this can prove laborious using a genome browser to examine each peak one by one. Using visualization tools to view and cluster large numbers of such 'peak calls' can save time, produce more accurate results and can help understand trends across the data.

In our lab, this has led to the development of tools such as Multi Locus View (MLV) which allow visualization of thousands of peaks, keeping the human in the loop to set better thresholds for getting true positive peaks and therefore to find more accurate binding sites. With all these peaks in the database we can convert the peak to a vector and use techniques such as dimensionality reduction to group together similarly shaped peaks. In Figure 2 we show an ENCODE data set (the current gold standard for epigenetics data) looking for H3K27ac (a repressive epigenetic mark) in a prostate cancer cell line.

More recently machine learning (ML) offers a promising way of identifying true signals accurately and rapidly above other methods. We can treat finding a characteristic peak as a problem in computer vision, where – just by looking at the shape alone – we can find peaks that are likely to have arisen from a true protein-binding event and ignore peaks that are simply due to noise from non-specific binding events in the data (which we class as 'poor peaks'). By training a model to learn what are robust peaks and what are poor peaks means better accuracy and specificity can be achieved over existing tools. The current statistical models are too simplistic to represent the underlying biology and ML offers a way to improve the detection since it can be trained with a variety of different peak types. The HSK27ac data set above was assessed by ENCODE as having approximately 45,000 peaks. When we assessed these using the ML peak caller 'Lanceotron' however, it showed ~15,000 of these may not be real peaks at

all. This highlights the importance of looking at the complete dataset where possible and not just high-level trends. The combination of being able to combine structured (the positions) with unstructured data (the peak image) is very powerful.

What is the structure of the folded DNA and proteins

Understanding how chromatin (DNA and proteins) is folded in the nucleus is fundamental to understanding its function. As mentioned earlier, gene expression has been found to be driven by the way the genome is folded in the nucleus. In the last few years techniques have become available to define where chromatin interacts and thereby understand its 3D structure. Chromatin Conformation Capture (3C) techniques allow a chromosome-wide contact map to be studied in a particular tissue or cell type showing where chromatin regions most frequently interact with other regions, and in some cases, other chromosomes. Using sophisticated modelling techniques we are beginning to piece together the model of the 3D structure of the genome. There are several caveats with these techniques, not least that there are no robust single-cell 3C techniques. Information is gathered as an aggregate across several thousand cells, and within each of the cells the chromatin may be in different conformational states. Similarly the quality of the data can depend on which type of 3C method you use, how much DNA has been used to make the DNA library and whether the apparent interaction is in fact due to background noise. Nonetheless this is a burgeoning and exciting field and thus has many opportunities where visualization can help make sense of this complex data. Tools such as Hi-Glass (see Figure 3) allow the user to 'zoom' into the contact maps and compare structures between different tissues. In areas of gene expression there are clear differences between cell or tissue types demonstrating different chromatin conformation.

The 2D representation of these data is a convenient way of comparing structures but we know these are actually 3D structures. There are bioinformatics tools that have been developed to use contact map data as input to generate a 3D structure, potentially the ultimate way to understand the data. The best methods use high-resolution fluorescent imaging markers as ground truth. These techniques are on the edge of what currently can be achieved based on resolution and the number of fluorescent markers that can be visualized. It is made more challenging by the fact that these are extremely dynamic structures which cannot be captured easily. Despite limitations they do offer a snapshot about where chromatin and loci are positioned in 3D space. There are a large number of simulation and modelling tools that have been developed to infer 3D structure based on molecular dynamics principles, traditionally used in areas such as protein folding.

We developed a dynamic modelling tool called CSynth (see Figure 4) which can dynamically assemble 3D models based on contact data. It also allows comparison of models generated from data in different tissues/cell states as well as the results of third-party 3D modelling outputs. CSynth encourages interaction with the modelling parameters – on-screen or in virtual reality – allowing experimentation to see how these affect the 3D model.

With any visualization of 3D genome data, caution should be used because the generated structures

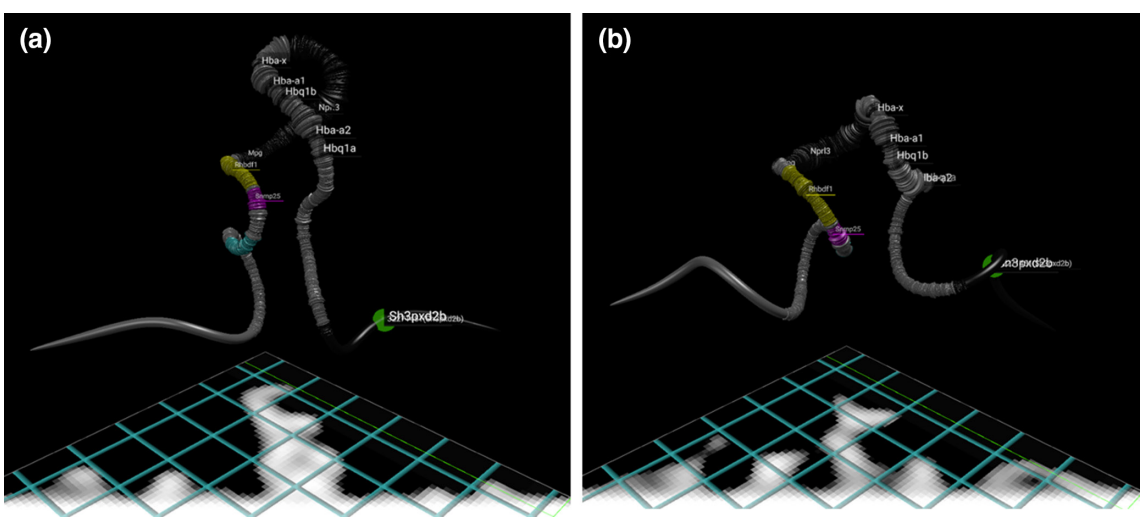

Figure 4. 3D modelling the genome using CSynth which takes the contact map data and uses forces to seek conformations that best satisfy the known interaction frequencies. It shows simulated structure in the mouse α-globin region in (a) red blood cell progenitors and (b) white blood cells. The 2D contact map projection is shown on the triangular grid below the 3D models as a further aid to visualize the differences between structures.

need to be verified by real data. This can be difficult because even state-of-the-art experimental techniques cannot capture the dynamics and resolution required to get fully accurate genome models. However, with an increasing number of chromatin structural alterations linked to human diseases and development abnormalities understanding the 3D structure holds promise to understand these complex biological mechanisms.

Summary

This article has shown a number of visualization methods that can be used when trying to understand some of the more complex aspects of gene expression. The tools shown in this article are not exhaustive, but highlight important issues in big data visualization in this field.

Visualization is becoming increasingly important as data volumes increase in size and complexity, but care should be taken to verify these assumptions by checking any underlying data to ensure beautiful visualizations are not hiding ugly data! ■

Acknowledgements

Thank you Simon McGowan for proof reading and discussions about this article.

Further reading

- Grigoryev, Y. (2012) How much information is stored in the human genome? *BiteSize Bio* https://bitesizebio.com/8378/how-much-information-is-stored-in-the-human-genome/ [Accessed 6 September 2021]
- ChIP sequencing. https://en.wikipedia.org/wiki/ChIP_sequencing. https://www.utsouthwestern.edu/labs/bioinformatics-lab/analysis/chip-seq/
- Sergeant, M.J., Hughes, J.R., Hentges, L. et al. (2021) Multi Locus View: an extensible web-based tool for the analysis of genomic data. *Commun. Biol.* **25**,623. DOI: 10.1038/s42003-021-02097-y
- Hentges, L.D., Sergeant, M.J., Downes, D.J. et al. (2021) LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq. *bioRxiv* DOI: 10.1101/2021.01.25.428108
- Oskolkov, N. (2020) tSNE vs UMAP: Global Structure. *towards data science.* https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17 [Accessed 6 September 2021]
- Nature Portfolio collection. *The 3D genome.* https://www.nature.com/collections/rsxlmsyslk [Accessed 6 September 2021]
- Todd, S., Todd, P., McGowan, S.J. et al. (2021) CSynth: an interactive modelling and visualization tool for 3D chromatin structure. *Bioinformatics*. **37**, 951–955. DOI: 10.1093/bioinformatics/btaa757

Stephen Taylor is a group leader at Oxford University, working on new ways to analyse and integrate biological research and clinical data. They are developing new ways of using visualization to handle massive data sets using techniques from machine learning to immersive technologies such as virtual reality. His current areas of research include spatial proteomics, multiplex imaging, epigenetics and 3D genome modelling.

