Visual Microscope for Massive Genomics Datasets, Expanded Perception and Interaction

Dominic Branchaud
UNSW Art & Design
Sydney, NSW, Australia

Walter Muskovic
Children's Cancer Institute
Sydney, NSW, Australia

Maria Kavallaris
Children's Cancer Institute
Sydney, NSW, Australia

Daniel Filonik
UNSW Art & Design
Sydney, NSW, Australia

Tomasz Bednarz
UNSW Art & Design / CSIRO Data61
Sydney, NSW, Australia

Figure 1: Visual Microscope being used to analyse gene expression in the EPICylinder with the Microsoft HoloLens

ABSTRACT

An innovative fully interactive and ultra-high resolution navigation tool has been developed to browse and analyze gene expression levels from human cancer cells, acting as a visual microscope on data. The tool uses high-performance visualization and computer graphics technology to enable genome scientists to observe the evolution of regulatory elements across time and gain valuable insights from their dataset as never before.

CCS CONCEPTS

• **Human-centered computing → Visual analytics**;

KEYWORDS

visualization, interaction design, big data, augmented reality, high-performance graphics

ACM Reference Format:
Dominic Branchaud, Walter Muskovic, Maria Kavallaris, Daniel Filonik, and Tomasz Bednarz. 2018. Visual Microscope for Massive Genomics Datasets, Expanded Perception and Interaction. In *Proceedings of SIGGRAPH '18 Posters.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3230744.3230745

SIGGRAPH '18 Posters, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5817-0/18/08.
https://doi.org/10.1145/3230744.3230745

1 INTRODUCTION

Following the Human Genome Project [Chial 2008], extensive effort was dedicated to building a comprehensive annotation of the functional elements of the genome, led by the Roadmap Epigenomics and ENCODE [Stanford-University 2018]. More recently, studies have identified these regulatory elements to produce small RNA molecules called enhancer RNAs (eRNAs), that are readily detectable through modern RNA-sequencing technologies. Computational approaches to make sense of eRNAs have relied on static snapshots from a broad range of cell types. By taking repeated measurements of a single cell-type across time, it is possible to identify changes in eRNAs to correspond closely with changes in the activity of specific classes of genes. Interpretation of these complex datasets is hindered by the scarcity of tools for their visualization and interpretation. A unique challenge is presented by the dynamic nature of the sequencing data and the vast differences in scale between gene size and the intervening spaces between genes and regulatory elements.

2 CURRENT GENOME BROWSER TOOLS

Current genome browser tools with the capability of visualizing RNA-sequencing data alongside relevant biological annotations, such as the Broad Institute's Integrative Genomics Viewer (IGV) [Thorvaldsdottir et al. 2013], are restricted by several key limitations. Existing browsers were designed for visual examination and comparison of a small number of samples at a time. Vertical stacking of dozens of samples makes meaningful interpretation difficult. Visualization is further complicated by the large file sizes of each sample, with comparison of large genomic regions quickly

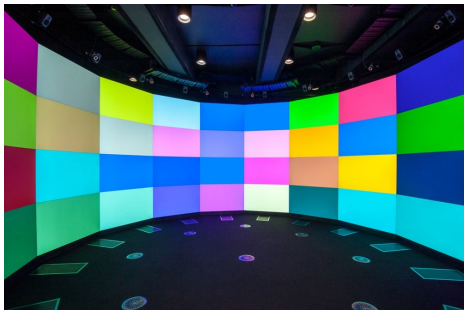

Figure 2: EPICylinder

exceeding available memory. Popular web-based genome browsers, the most popular being Ensembl, suffer from similar limitations. These browsers were designed to provide a graphical interface to curated repositories of consortium-generated genome projects. The rapid increase in user-generated genomic data, and correspondingly increasing data traffic, reduce the efficiency of these web-based browsers for visualization of externally produced data. Finally, established tools do not provide any mechanism for the visual interpretation of sequencing data in a manner appropriate to the dynamic nature of time series data.

3 HIGH-PERFORMANCE VISUAL ANALYTICS SOLUTION

In this project, we took into account all the above-mentioned limitations, and have developed a groundbreaking tool to extend the ways genomics datasets can be viewed and analysed. To achieve that, we have employed the EPICylinder located in the Expanded Perception and Interaction Centre (epicentre.matters.today). The EPICylinder constitutes nearly 120 million pixels in stereo 3D, resolution 26,880×4,320 synchronized at 120Hz. The system contains uniquely designed 56× of 60" display cubes, assembled in a 4×14 matrix with 1-2mm edge-to-edge bezels, currently constituting the highest resolution cylinder in the world. Visualizations are driven by a 28 node cluster equipped with Xeon E5-2650 and 28x NVIDIA Quadro M6000. The system also includes 12× IR VICON tracking cameras able to locate trackers in 3D physical space, directly translating position of the user or interactive controllers to cluster-based Virtual Reality system developed using Unity3D game engine. For extended exploration of the big genome datasets we have also utilized Microsoft HoloLens allowing us to use another layer of interactivity, not only visual but also voice activated commands: select, zoom in, zoom out, load and save bookmark.

4 VISUAL ANALYTICS WORKFLOW

4.1 Data Source and Pre-Processing

The data was provided by the Children's Cancer Institute in BedGraph format, combined with the GTF open source annotation file extracted from the GENCODE project. The dataset consisted of measures of RNA expression for the full human genome, with both forward (positive) and reverse (negative) strands separated. Time point were taken 40 times at 10 minutes intervals, on by harvesting synchronized brain cancer cells across time. In total 54 GB of data were generated, with 41 time slices and a total of 1.36 billion lines of expression values.

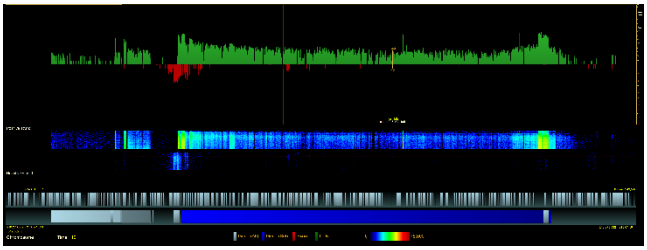

Figure 3: Truncated overview of RNA expression in Chromosome 1 of a DNA sample of synchronised glioblastoma cells

4.2 Data Pre-Processing

Before data could be visualized using our High End Visualization system it was preprocessed by splitting it into chromosomes (chunking/classifying), stripping all irrelevant data (filtering and denoising), reducing size by converting text to binary formats and pre-generating view outputs. In total, we have produced 3906 views per chromosome, giving us 58097 genes indexed in annotation files, with 159653 total number of pre-generated views, totaling over 89GB of data that was then distributed across the cluster nodes (caching).

4.3 Visualization Modes and Interaction

The dataset can be displayed as 2D and 3D signals. In 2D mode, one time slice is visualised at a time, with horizontal axis being genome coordinates, and vertical axis gene expression level. In 3D mode, all time slices are rendered at once as a 2D heatmap (horizonal axis: coordinate, vertical axis: time and color describing expression level). There is also the full chromosome map displayed with the location of every annotated gene. A search bar allows the user to search and jump directly to a particular gene of interest. Hovering the cursor over the gene area triggers a real-time display of gene annotations. Our viewer also allows the user to view the level of RNA expression at a particular coordinate inside the DNA sequence, and also the locations of Exons/Introns/UTRs when zoomed on a gene. The basic interactions are usually carried out using a game controller that is being tracked in the space by our tracking system (VICON). The application also can use Microsoft HoloLens, with voice/gesture recognition substantially speeding up browsing time. The user can select a chromosome of interest, zoom in or out on any region of interest within a chromosome, zoom in on a specific gene, save/load a bookmark for the selected region of interest, highlight significant expression levels by changing scale between linear and logarithmic, and also changing heat map color.

REFERENCES

Heidi Chial. 2008. DNA sequencing technologies key to the Human Genome Project. *Nature Education* 1 (2008), 219.

Stanford-University. 2018. ENCODE: Encyclopedia of DNA Elements. www.encodeproject.org

Helga Thorvaldsdottir, James T. Robinson, and Jill P. Mesirov. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14 (2013), 178–192.