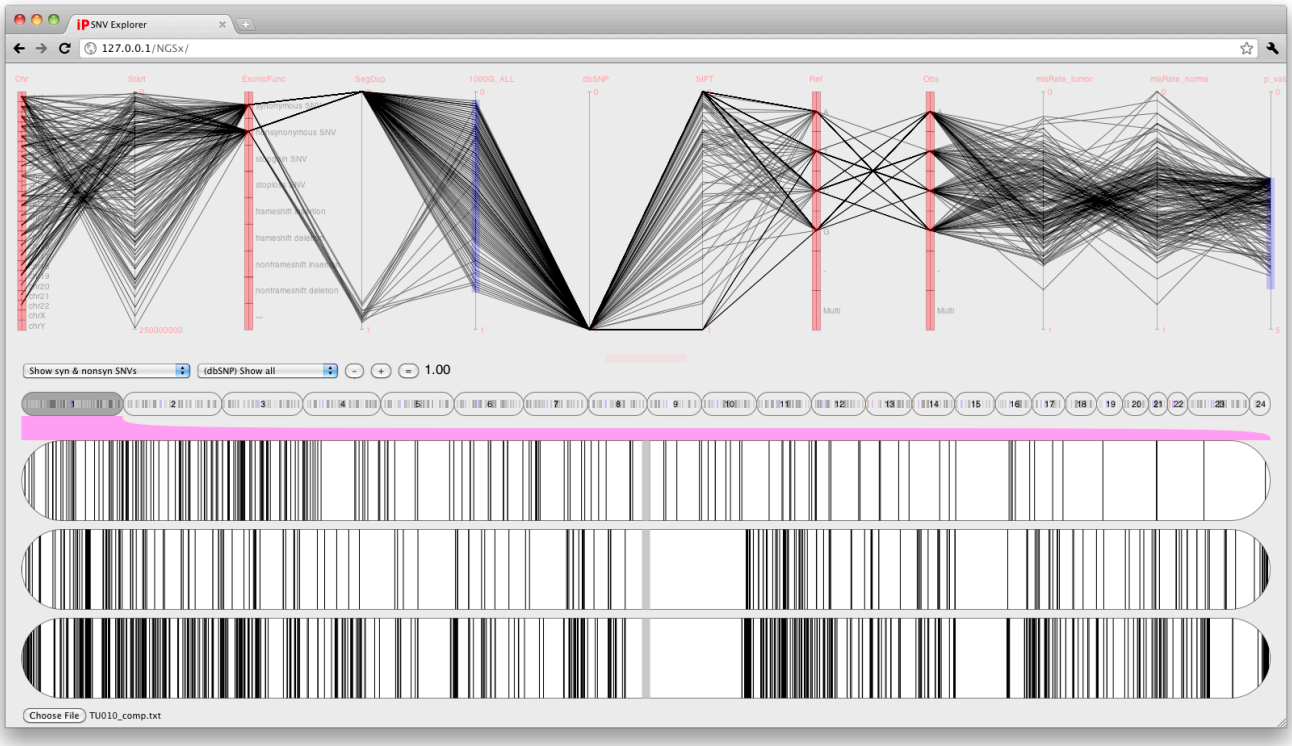
NGS Explorer: An Application for Visually Contextualizing and Interrogating Multivariate Omics Data

Georg Tremmel*, Atsushi Niida, Yuichi Shiraishi, Masao Nagasaki and Satoru Miyano

Human Genome Center, Institute of Medial Science, The University of Tokyo, Japan



Abstract

NGS Explorer is a HTML5, browser-based application for visualizing and interrogating multi-dimensional omics data. The application uses Parallel Coordinates combined with a genome track view to contextualize the genomic information. The intended user is a biologist, who wants to manually and visually down-select already processed NGS (Next Generation Sequencing) data in order to reveal mutations that satisfy criteria selected by the biologist.

By enabling the application to run within web browsers we also lower the barrier of usage and allow the application to be used across platforms.

Keywords: Next-Generation-Sequencing Data, Parallel Coordinates, Information Visualization, Visual Analysis

Human Genome Center, Institute of Medical Science
University of Tokyo, 4-6-1 Shirokanedai, Minatoku
Tokyo 108-8639, Japan
*e-mail: tremmel@hgc.jp

Index Terms: J.3 [Computer Applications]: Life and Medical Sciences—Biology and Genetics; H.3.5 [Information Storage and Retrieval]: Online Information Services—Web-based services

1 Introduction

Next-generation sequencing (NGS) has become commonplace and affordable, with the consequence, that the bottleneck of NGS Analysis is no longer the costs of sequencing, but the costs of analysis. At the same time, sequencing machines have found their way into biological labs, and biologists are facing the challenge of dealing with the output of NGS sequencers and NGS pipelines.

Here we introduce NGS Explorer, a HTML5 browser-based application for contextualizing and interrogating multi-variate omics data. NGS Explorer is combining the visualization strength of Parallel Coordinates with the context-sensitive information available in the linear genome track view.

The aim of this application is to help the biologist make sense of processed NGS data, by enabling him to visually investigate the data. NGS Explorer assists in finding relevant and biologically interesting mutations, which require further attention and investigation.

The browser-based approach is used to lower the barrier of usage for the application. Creating the application with HTML5

and JavaScript also has the additional advantage of speed of development and ease of deployment.

In the next section we will details the visualization methods used, in section 3 we will summarize the technological background, section 4 discusses the example data and shows actual uses of the Application. The final section presents conclusion and outlooks.

2 Visualization Methods

Parallel Coordinates are a common yet powerful method for displaying and interactively interrogating multivariate data. Combined with a linear genome track view we can show the genomic data within context. We are distinguishing in the Parallel Coordinates selection method between a continuous, range selection and a discrete, button-like selection. A chromosome overview is present in both the Parallel Coordinate view as well as in the genome track view, allowing for situation-depended selection. It is also possible to make multiple independent selections on the same coordinate, allowing – for example - for the simultaneous selection of low and high values, but not of median ones.

Standard features of parallel coordinates like re-ordering, scaling and rotation of the dimension axis are also available.

Presets allow the current values of all Parallel Coordinates selections to be saved and later applied to a different set of similar data.

2.1 Pseudo-Continuous Display of Discrete Values

While Parallel Coordinates can create a rich, gradient display if the data is continuous, they convene less information if the data is discrete and only consists of few values. A dimension, consisting of whether an SNV (Single-nucleotide variation) is found in a database or not, communicates only binary information. The data-ink ratio[1] is improved by overloading the discrete values and displaying them in a pseudo-continuous way. Instead of the two binary values occupying points on the opposite ends of a coordinate, they are given a range on the coordinate. The height of these ranges reflects the amount of data points, thus giving a visual overview of the distribution of the data points at this particular coordinate.

Inselberg's definition of Parallel Coordinates include that each variable should be treated uniformly[2]. While we are violating this definition, we can show that the transformation from data → pictures gains additional information, which would not be present, if a more stringent interpretation of Parallel Coordinates would be applied.

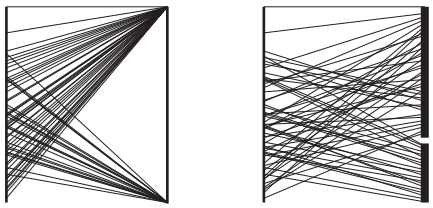


Figure 1: Discrete (left) and Pseudo-continuous display

The introduction of the deviation forces us to consider the selection behavior and the borders between the discrete values. Selection behavior of the coordinate remains discrete, either one or the other value can be selected. We also marked the border between the discrete values more clearly, to create a visible border between the pseudo-continuous discrete ranges.

2.2 Genome Track Matrix

The genome track view displays the genomic information in a linear contextualized fashion, in which different samples occupy the y-axis. In the case of positional punctual SNV information, the tracks show large gaps between the SNVs, therefore reducing the convened information of the view[3][4].

We offer the option of reducing the genome view to an Exon-Only view and an SNV-Only view, essentially removing the position, which do not carry information, which creates a matrix display. As it is relevant for biologists to see the relations amongst SNVs across different samples, we offer visual clues links between similar variants.

3 Technological Considerations

The main reason for developing Java-based applications has traditionally been the ability to deploy the same application across various platforms. The current generation browsers and the introduction of new standards like HTML5, CSS3, ECMAScript 5 (JavaScript 1.8) and WebGL together with the increase of the speed of JavaScript engines, make it possible to develop browser-based cross platform applications. This browser-based approach also has the advantage that the end-user does not have to install additional software on the computer, thus lowering the barrier of interacting and increasing the potential usage the application.

Data can be imported as TSV (Tab-Separated Value) text files and subsequently parsed into Javascript, or directly imported as JSON (JavaScript Object Notation) files. Selections can also be exported either as TSV or JSON text files.

The eloquence, expressiveness and elegance of Javascript also allows for rapid prototyping and implementation of visualization ideas, while at the same time guaranteeing cross-platform and cross-browser compatibility.

4 Discussion / Example

The example data consists multivariate SNV data, the processed output of NGS Exon-seq pipeline. Currently we are working with 50 samples, each sample is made up of around 2000 SNVs, for a total of about 100 000 SNV data points.

Each SNV data point consists of information covering the name of the gene, the type of amino acid change and positional and structural information such as the number of the chromosome, the position on the chromosome, the start and end points, segmentation duplication, as well as the reference and observed amino acids. Not all the information can be readily displayed with Parallel Coordinates; we are working on transforming the data and on fitting the view according to the data.

5 Conclusion and Outlook

We showed, that HTML5 and Javascript provide an environment for applications that can rival traditional desktop applications. Although we only show SNV data in this example, the flexibility of displaying the data with Parallel Coordinates, combined with a linear genome view, allows us to incorporate other biological data and meta-data.

The web-based nature of the application makes it also possible to easily use, share and communicate the visual analysis.

References

[1] E. R. Tufte. The Visual Display of Quantitative Information. *Graphics Press*, Cheshire, Connecticut, 2001.

[2] A. Inselberg, The plane with parallel coordinates, *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985.

[3] H. Siirtola, "Combining Parallel Coordinates with the Reorderable Matrix," in International Conference on Coordinated and Multiple Views in Exploratory Visualization - CMV 2003 -, 2003, pp. 63–74.