compreheNGSive: A Tool for Exploring Next-Gen Sequencing Variants

Alex Bigelow *
University of Utah
Scientific Computing and
Imaging Institute

Miriah Meyer
University of Utah
Scientific Computing and
Imaging Institute

Nicola J. Camp
University of Utah
Division of Genetic
Epidemiology

Index Terms: next-gen sequencing, biovis, design study

1 Introduction

Next-generation sequencing data is notoriously difficult to process, store, and summarize, let alone analyze with any degree of comprehensiveness. This type of data epitomizes the difficulties that accompany today's data deluge: its size is massive, and the cost to acquire it is declining sharply, further increasing its abundance [1]. Raw next-generation sequencing (NGS) data must pass through several layers of nontrivial post processing [2]. First, the raw data is assembled into a genome. Next, a variety of attributes are assigned to genomic locations describing the quality of the assembly. And finally, the genome is *annotated* to assign biologically meaningful concepts to locations in the genome. Currently, these steps require significant competence with command-line utilities and scripting, as well as reading a great deal of documentation that is often insufficient. To add to the complexity most steps, especially annotation, involve their own specific file formats. The time and effort to integrate data from multiple formats and sources can be overwhelming — even a seemingly simple task of assigning individual genomes to meaningful groups becomes difficult. We observed that much of the data is never analyzed because of these complexities.

Existing tools for analyzing NGS data are similarly limiting in the amount of data that can be displayed at once. Tools for looking at the data in genomic coordinates, such as the UCSC Genome Browser [3], restrict analysis to comparisons that can be made locally around specific genomic features. The other commonly used method involves filtering the data in a spreadsheet program such as Excel, where scientists must select a small subset of interesting features based on a handful of attributes. With data sets that potentially have millions of points, experts analyze a very small fraction of the data. Though this kind of extreme filtering is necessary for human-driven analysis, no tool provides a global overview of the data to inform this selection.

In this design study we are collaborating with a group of genetic epidemiology experts who are using NGS data to study breast cancer [4]. Our work with this group is focused on overcoming many of the challenges for analyzing NGS data, specifically for the biological questions they are tackling. Based on a careful analysis of the problem and needs of the group we are developing an interactive visualization tool, called *compreheNGSive*, to support the integration and exploration of post processed NGS data. Our contributions are an articulation of a workflow and set of visualization tasks needed by our collaborators, and an early prototype of our system *compreheNGSive* that supports this workflow and these tasks.

2 Data and Tasks

To understand our collaborators' NGS data and analysis needs we worked with this group for a year, spending four days a week in their lab while conducting numerous interviews — one of the scientists is a co-author of this abstract. Based on these interviews we developed over ten paper prototypes and five software prototypes and acquired feedback on the prototypes from the scientists. By analyzing this feedback we uncovered several necessary visualization tasks and identified a workflow applicable to their data exploration.

2.1 Data

The data our collaborators study is a set of NGS data sets from a population of fifty people. Our collaborators are interested in *variants*, which are specific locations in the genome that are different from person to person. If we represent the genome as a series of letters, then variants will be either a letter change, missing letters, or inserted letters. Specifically, the biologists are looking for sets of variants that change in similar ways across a subset of the population. In their initial analysis stage, the biologists are looking at a small, targeted part of the genome which has been implicated with breast cancer, containing 2919 variants.

Each variant has a set of attributes made up of a variety of quality scores and annotations — the scientists currently work with approximately eight attributes, but we anticipate that the number could go as high as several dozen attributes. Annotations are particularly problematic as many of the variants are missing data due to the specificity of existing automatic annotation tools. For example, one of the annotation tools used by our collaborators, called VAAST [5], only supports annotating regions of the genome responsible for coding proteins. Analyzing missing data is difficult to do in traditional NGS analysis tools, but was specifically requested by our collaborators. We discuss our solution to this problem in Section 3.2.

2.2 Tasks

The most challenging aspect of this design study has been the articulation of the required analysis tasks — a well-known hurdle in collaborative, problem-driven visualization research [6]. At the start of our collaboration the scientists could only articulate that they wanted to explore *interesting* variants. It took more than a year of embedded work with these scientists and multiple prototypes to understand the meaning of *interesting*. In particular, our software prototypes caused the scientists rethink their analysis multiple times as they dug into the data for the first time.

The workflow we identified involves four stages: first, the myriad of data files are gathered following processing of the NGS data; second, each individual in the population under study is assigned to a group (such as control versus test cases); third, variants of interest are explored and prioritized between the groups and across the attributes; and fourth, the highest priority variants are studied in more detail, which many entail follow-up experiments.

Our tool *compreheNGSive* specifically aids the second and third stages of the workflow by supporting the following analysis tasks:

1. Assign each individual in the population to a group.
2. Select a subset of variants based on the correlation of user-selected attributes.
3. Filter a subset of variants across the entire set of attributes.
4. View the genomic location of multiple subsets of variants.
5. Create lists of high-priority variants.

*e-mail: alex.bigelow@utah.edu

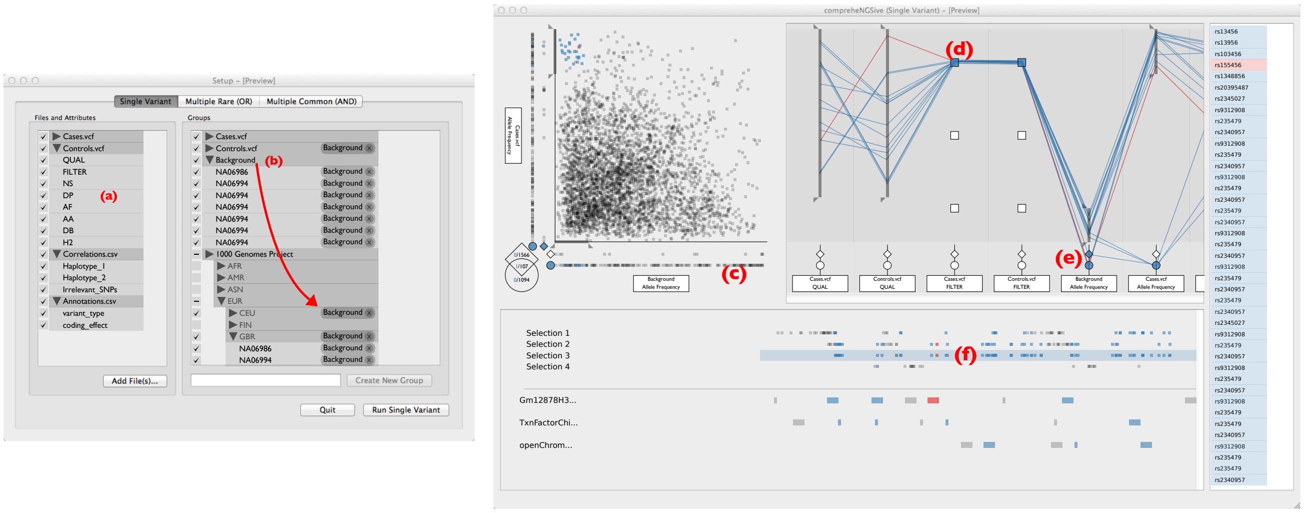

Figure 1: Screenshots from the prototype of *compreheNGSive*. (left) Interactive interface for assigning individuals to groups. (right) Multiple-linked views for selecting and prioritizing variants.

3 *compreheNGSive*

Usage of the tool is divided into two phases: file and group management, and exploration, selection, and prioritization of variants. Figures 1(left) and 1(right) show the interfaces for these two phases.

3.1 File loading and group management

The first phase contains two views. The first view (Figure 1(a)) is for loading files, with checkboxes for including/excluding specific variant attributes in the resulting visualization. *compreheNGSive* supports numerous industry standard NGS file formats, incorporating any scores from quality control and variant calling steps. It also supports generic tabular files, simplifying bringing values from custom annotation and association analyses into the same framework.

The other view (Figure 1(b)) is used to define, manage, and manipulate groups of individuals, supporting Task 1. For example, data from the 1000 Genomes Project[7] has natural, nested subdivisions based on ethnicity. Typical group assignments include selecting specific individuals as cases, controls, or custom background groups that control for population stratification. *compreheNGSive* allows for quick, arbitrary assignments via drag-and-drop interactions that would be difficult to perform in a command-line interface.

3.2 Variant exploration, selection, and prioritization

The second phase includes three linked views [8] including a scatterplot (Figure 1(c)), a parallel coordinates view (Figure 1(d)), and a genome view (Figure 1(f)). Each view is linked by a set of selected variants; all selected variants are shown in each view highlighted in blue. Mousing over a single variant in any view highlights that variant in all of the views. The user can add a variant to a list of high-priority variants from any of these views, supporting Task 5.

The scatterplot view is a two-dimensional, global representation of *all* variants for any two user-selected attributes, supporting Task 2. The attributes can be selected from the parallel coordinates view. Missing data for the selected attributes are represented along one-dimensions plots parallel to the attributes' axes in the scatterplot. Conversely, the parallel coordinates view shows a user-selected subset of variants across *all* of the attributes, supporting Task 3. The axes can be re-ordered, turned on and off, and also used to filter the set of selected variants with sliders. The axes support both quantitative and categorical data, and missing data is represented as check boxes below each axis (Figure 1(e)). These mechanisms allow the user to both visualize and filter based on virtually *any* value, without hiding variants that have missing annotations.

The genome view allows for viewing the location of variants within the genome. The selected subset of variants, along with lists of high-priority variants are shown along the one-dimensional genomic coordinate system — this coordinate system is augmented with the user-defined regions of interest which serve as landmarks and provide context. The genome view provides a familiar frame of reference and an interface that allows analysis of any genomic feature set in the context of the variants, supporting Task 4.

4 Conclusion

Unlike many existing tools, *compreheNGSive* supports exploration of NGS data by integrating and leveraging all aspects of the data. Furthermore, it assumes no biological significance to any particular variant; it simply displays all of the data. As such, *compreheNGSive* simplifies many of the tasks for analysts working with NGS data while allowing full, easy access to *all* of the data from a variety of sources. We are continuing to develop *compreheNGSive* and gathering feedback about its features from our collaborators.

Acknowledgements

This work was funded in part by a grant from the Avon Foundation.

References

[1] Kris Wetterstrand. Dna sequencing costs: Data from the nhgri large-scale genome sequencing program.

[2] Mark A DePristo, Eric Banks, Ryan Poplin, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nat Genet*, 43(5):491–498, 05 2011.

[3] W. James Kent, Charles W. Sugnet, Terrence S. Furey, et al. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[4] Nicola J Camp, Marina Parry, Stacey Knight, et al. Fine-mapping casp8 risk variants in breast cancer. *Cancer Epidemiol Biomarkers Prev*, 21(1):176–181, Jan 2012.

[5] Mark Yandell, Chad D Huff, Hao Hu, et al. A probabilistic disease-gene finder for personal genomes. *Genome Research*, 2011.

[6] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *to appear in IEEE Trans. on Visualization and Computer Graphics (Proceedings of InfoVis)*, 2012.

[7] A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 10 2010.

[8] Jonathan C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. Intl. Conf. on Coordinated and Multiple Views in Exploratory Visualization (CMV)*, pages 61–71. IEEE Computer Society, 2007.