GENOME ANALYSIS

StainedGlass: Making colorful dot-plots of genomic sequence

Mitchell R. Vollger[1,*], Peter Kerpedjiev[2], Adam M. Phillippy[3,*], and Evan E. Eichler[1,4,*]

[1]Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
[2]Reservoir Genomics LLC, Oakland, CA
[3]Genome Informatics Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
[4]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA.

*To whom correspondence should be addressed.

Abstract

Summary: Visualization of genomic repeats is often accomplished through the use of dot plots; however, the emergence of telomere-to-telomere assemblies with multi-megabase repeats requires new visualization strategies. Here, we introduce StainedGlass which can generate publication quality figures that communicate the identity and orientation of multi-megabase repeats while scaling to entire genomes.
Availability and implementation: StainedGlass is implemented using snakemake and is available open source under the MIT license at mrvollger.github.io/StainedGlass/.
Contact: mvollger@uw.edu

1 Introduction

Dot plots are a powerful way to show sequence similarity that often reveal the underlying structures of complex repeats.

However, with increasingly contiguous assemblies of reference genomes (VGP) and complete human chromosomes (chr8, chrX, T2T) repeat structures including centromeres and other hererochromatic arrays are now for the first time available for analysis. The size and complexity of these structures, often many megabase pairs in humans, elude traditional dot plots for two reasons: 1) current visualization methods are largely based on perfect k-mer matches which do not lend themselves to the expected gaps and mismatches between large repeats, and 2) for tandem arrays of consisting of megabases of sequence dot plots are often just black squares that relay little information other the the size and presence of sequence similarity.

In order to examine the centromere in human chr8 a colored dot plot based on sequence alignment rather than small k-mers was designed which allowed the authors to make a model for centomere evolution.

In this work, we present StainedGlass, which generalizes the idea of colored dot plots based on sequence alignment and provides an easy, scale-able, and customize-able workflow so that it can be applied to new genomes.

2 Usage and examples

Bofelli *et al.*, 2000 example cite
The tool is made available using snakemake which allows for reproducible and scalable data analyses. Additionally, stability of new changes is automatically tested with each new change with continuous integration via github actions.

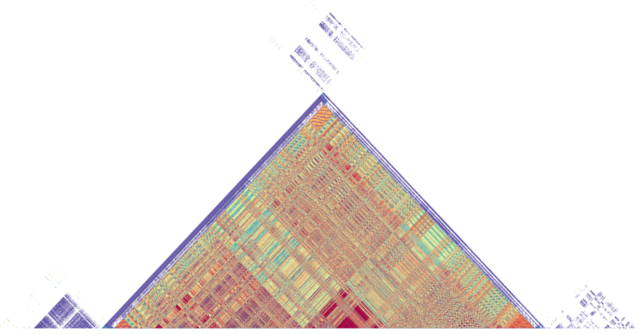

Fig. 1. Caption, caption.

Figure 1

3 Conclusion

StainedGlass is a visualization tool for large genomic repeats and building on snakemake makes StainedGlass both reproducible and scalable at the whole genome level. The output visualizations produced by StainedGlass are publication ready while also providing an option for interactive data exploration through the use of HiGlass.

Acknowledgements

The authors thank T. Brown for help in editing this manuscript and G. A. Logsdon for aesthetic suggestions.

Funding

This work was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (A.M.P.) and grants from the U.S. National Institutes of Health (NIH grants 5R01HG002385 to E.E.E.; 5U01HG010971 to E.E.E.; and 1U01HG010973 to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

References

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.