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 (Rhie *et al.*, 2021) and complete human chromosomes (Miga *et al.*, 2020; Logsdon *et al.*, 2021; Nurk *et al.*, 2021) repeat structures including centromeres and other heterochromatic 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 or 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 than the size and presence of sequence similarity.

In order to examine the centromere of human chromosome eight a colored dot plot based on sequence alignment rather than small k-mers was designed which allowed the authors to make a model for centromere evolution (Logsdon *et al.*, 2021). In this work, we present StainedGlass, which generalizes the idea of colored dot plots based on sequence alignment and provides an easy, scalable, and customizable workflow so that it can be applied to new genomes.

2 Methods

To generate pairwise sequence identity dot-plots for StainedGlass the input sequence is fragmented into windows of a preset size (default 5 kbp) and then all possible pairwise alignments between the fragments are calculated using minimap2 (Li, 2018) in an all by all alignment mode. The color used in the dot-plot is determined by the sequence identity of the alignment which is calculated as:

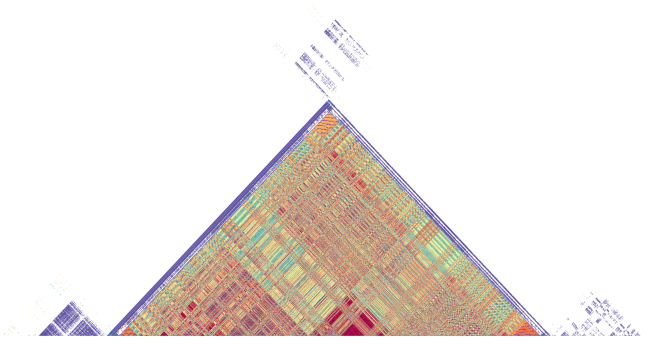

Fig. 1. Output from StainedGlass showing the sequence identity within human chromosome 8.

$$ID = 100\left(\frac{M}{M+X+I+D}\right)$$

where ID is the percent sequence identity, M the number of matches, X the number of mismatches, I the number of insertion events, and D the number of deletion events. When there are multiple alignments between the same two fragments of sequences all alignments other than the one with the most matches are filtered out regardless of their sequence identity.

The resulting matrix of percent identity scores can then be visualized using either static figures, or with HiGlass (Kerpedjiev *et al.*, 2018) which allows for interactive data exploration. The static figures are more appropriate for visualization of relatively small regions (30 Mbp or less) at publication quality (Figure 1) while the HiGlass visualization is better for data exploration of whole genome alignments.

The tool is made available using snakemake (Köster and Rahmann, 2012, 2018; Mölder *et al.*, 2021) which allows for reproducible and scalable data analyses. The stability of new changes are automatically tested with each new change using continuous integration via github actions. Finally, StainedGlass is snakemake standard compliant so it has automated usage documentation.

3 Usage and examples

To generate the alignments for StainedGlass you can execute the workflow as follows.

```
snakemake --use-conda --cores 4
```

To generate the static figures you can add make_figures to the command.

```
snakemake --use-conda --cores 4 make_figures
```

StainedGlass can also be used to make cooler files that can be loaded into HiGlass for whole genome visualizations.

```
snakemake --use-conda --cores 4 cooler
```

An example of this interactive browser for the telomere to telomere assembly of CHM13 can be found at resgen.io.

4 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

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J. M., Ouellette, S. B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B. H., Pfister, H., Mirny, L. A., Park, P. J., and Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.*, **19**(1), 125.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**(19), 2520–2522.

Köster, J. and Rahmann, S. (2018). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **34**(20), 3600.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P. C., Rhie, A., de Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., Hoekzema, K., Murali, S. C., Munson, K. M., Baker, C., Sorensen, M., Lewis, A. M., Surti, U., Gerton, J. L., Larionov, V., Ventura, M., Miga, K. H., Phillippy, A. M., and Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, **593**(7857), 101–107.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., Markovic, C., Maduro, V., Dutra, A., Bouffard, G. G., Chang, A. M., Hansen, N. F., Wilfert, A. B., Thibaud-Nissen, F., Schmitt, A. D., Belton, J.-M., Selvaraj, S., Dennis, M. Y., Soto, D. C., Sahasrabudhe, R., Kaya, G., Quick, J., Loman, N. J., Holmes, N., Loose, M., Surti, U., Risques, R. A., Graves Lindsay, T. A., Fulton, R., Hall, I., Paten, B., Howe, K., Timp, W., Young, A., Mullikin, J. C., Pevzner, P. A., Gerton, J. L., Sullivan, B. A., Eichler, E. E., and Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**(7823), 79–84.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Köster, J. (2021). Sustainable data analysis with snakemake. *F1000Res.*, **10**(33), 33.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sovic, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2021). The complete sequence of a human genome.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Vernes, S. C., Lama, T. M., Grutzner, F., Warren, W. C., Balakrishnan, C. N., Burt, D., George, J. M., Biegler, M. T., Iorns, D., Digby, A., Eason, D., Robertson, B., Edwards, T., Wilkinson, M., Turner, G., Meyer, A., Kautt, A. F., Franchini, P., William Detrich, H., Svardal, H., Wagner, M., Naylor, G. J. P., Pippel, M., Malinsky, M., Mooney, M., Simbirsky, M., Hannigan, B. T., Pesout, T., Houck, M., Misuraca, A., Kingan, S. B., Hall, R., Kronenberg, Z., Sović, I., Dunn, C., Ning, Z., Hastie, A., Lee, J., Selvaraj, S., Green, R. E., Putnam, N. H., Gut, I., Ghurye, J., Garrison, E., Sims, Y., Collins, J., Pelan, S., Torrance, J., Tracey, A., Wood, J., Dagnew, R. E., Guan, D., London, S. E., Clayton, D. F., Mello, C. V., Friedrich, S. R., Lovell, P. V., Osipova, E., Al-Ajli, F. O., Secomandi, S., Kim, H., Theofanopoulou, C., Hiller, M., Zhou, Y., Harris, R. S., Makova, K. D., Medvedev, P., Hoffman, J., Masterson, P., Clark, K., Martin, F., Howe, K., Flicek, P., Walenz, B. P., Kwak, W., Clawson, H., Diekhans, M., Nassar, L., Paten, B., Kraus, R. H. S., Crawford, A. J., Gilbert, M. T. P., Zhang, G., Venkatesh, B., Murphy, R. W., Koepfli, K.-P., Shapiro, B., Johnson, W. E., Di Palma, F., Marques-Bonet, T., Teeling, E. C., Warnow, T., Graves, J. M., Ryder, O. A., Haussler, D., O'Brien, S. J., Korlach, J., Lewin, H. A., Howe, K., Myers, E. W., Durbin, R., Phillippy, A. M., and Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**(7856), 737–746.