The Journal of Open Source Software

seabreeze: A Pipeline for Analyzing Structural Variation Between Bacterial Genome Assemblies

Ira Zibbu [1,2], **Claus O. Wilke** [3], **and Jeffrey E. Barrick** [1¶]

1 Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas, United States of America **2** School of Biology, Indian Institute of Science Education and Research, Thiruvananthapuram, Kerala, India **3** Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America ¶ Corresponding author

DOI: 10.21105/joss.08065

Software

- Review
- Repository
- Archive

Editor: Abhishek Tiwari

Reviewers:

- @JeanMainguy
- @dr-joe-wirth

Submitted: 03 February 2025
Published: 24 July 2025

Summary

Structural mutations—such as large insertions, deletions, duplications, inversions, and translocations—play a unique and important role in bacterial evolution. Recent advances in long-read sequencing have made accurate, high-throughput predictions of structural mutations possible. *seabreeze* is a tool for comprehensively analyzing genetic variation among bacterial genomes caused by structural mutations. It manages a workflow that combines existing packages and custom scripts to automate and unite several analyses into a single, easy-to-use pipeline. For specified pairs of bacterial genomes, *seabreeze* predicts and visualizes structural mutations, and annotates the affected genes. It also uses information about transposons at the boundaries of structural mutations to predict their involvement in generating the mutations. Finally, *seabreeze* provides information about size differences between the two genomes and changes in replichore balance within circular chromosomes. Although *seabreeze* was developed to characterize structural mutations that evolved in laboratory experiments, it can be used to analyze any sufficiently closely related genomes from strains of the same bacterial species.

Statement of Need

Evolution experiments are one of many approaches that are used to understand the processes that shape microbial genomes. In these experiments, populations of microbes are propagated under controlled conditions for a sufficient period of time to observe evolution in action. Then, evolved genomes are resequenced and compared to the ancestral genome. We created *seabreeze*, a Snakemake pipeline that automates comparing genome assemblies to predict and analyze structural variation that emerges in these experiments. *seabreeze* fulfills two core needs. First, although a wide range of software programs exist to identify structural variation from high-throughput sequencing data (Ahsan et al., 2023), they make assumptions that are only appropriate for eukaryotic genomes and/or are limited by how they compare reads to a reference genome. In contrast, *seabreeze* is explicitly tailored for bacterial genome analysis and takes advantage of the benefits of comparing genome assemblies. Second, *seabreeze* unites several standalone open-source tools and new custom scripts into a single easy-to-use pipeline for comprehensive bacterial genome analysis. Other notable tools for bacterial genome analysis such as Artemis (Carver et al., 2005) and Mauve (Darling et al., 2010) can detect and visualize rearrangements in bacterial genomes but lack other functionality, such as predicting the mechanisms of structural mutations and annotating what genes they affect.

Implementation

The latest release of *seabreeze* can be downloaded from the official GitHub repository. The online documentation details installation, usage, and output. It contains a tutorial to demonstrate how it can be used. *seabreeze* uses Snakemake (Köster & Rahmann, 2012) to manage the pipeline and automatically install and manage dependencies for the workflow from `conda-forge` (Conda-Forge Community, 2015) and `bioconda` repositories (The Bioconda Team et al., 2018). Users supply fully assembled sequences in FASTA format for each reference-query pair, and specify which pairs of sequences to compare in a CSV file to allow for batch-processing. Suitable input files can be generated from long-read DNA sequencing datasets with assemblers such as Flye (Kolmogorov et al., 2019), Canu (Koren et al., 2017) or Raven (Vaser & Šikić, 2021), or from consensus assembly tools like Trycycler (Wick et al., 2021).

seabreeze has several subcommands to perform specific tasks. It begins by accounting for the circular nature of bacterial chromosomes and plasmids by rotating the reference-query pairs to a common start coordinate. Next, it uses the following packages to perform various analysis steps for each reference-query pair: (1) compute size difference between genomes with `biopython` (Cock et al., 2009); (2) predict the locations of transposons with `ISEScan` (Xie & Tang, 2017); (3) align genomes with `mummer4` (Marçais et al., 2018); (4) predict structural mutations with `SyRI` (Goel et al., 2019) and filter false-positive calls with a custom script; (5) generate intuitive synteny plots to visualize structural variation with a customized version of `plotsr` (Figure 1) (Goel & Schneeberger, 2022); (6) annotate the genes contained in the mutated regions with `prokka` (Seemann, 2014) and `breseq` (Deatherage & Barrick, 2014).

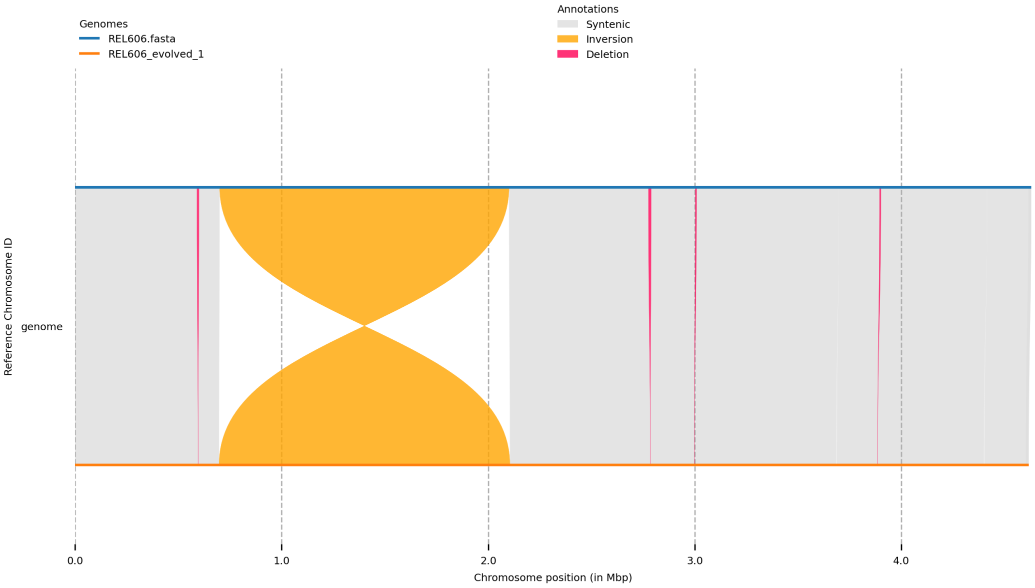

Figure 1: Synteny plot generated by seabreeze. This plot compares an ancestor genome (top, in blue) to its simulated evolved genome (bottom, in orange). Grey regions are syntenic (i.e., gene presence and order is preserved between both genomes). A single large inversion (orange ribbon) and several deletions (red ribbons) are visible.

Most bacterial chromosomes are circular and contain a single origin and terminus of replication, which are used to define an origin-terminus axis that divides the genome into two halves called replichores (Rocha, 2004). *seabreeze* introduces new scripts to analyze the placement of inversions relative to the origin-terminus axis and how they affect the symmetry of the two replichores. Most structural mutations occur through recombination between homologous sequences, and in particular, bacterial genomes often contain multiple copies of simple transposons (also known as insertion sequences) that serve as sites for recombination (Achaz et al., 2003). *seabreeze* uses the locations of transposons to predict whether they were involved in generating inversions and deletions. Highly diverged genomes may have successive structural mutations or low sequence homology in syntenic regions that complicate these inferences, and *seabreeze* may be unable to identify the individual mutational steps that led to complex structural variation in these cases.

Overall, the growing use of long-read sequencing will increasingly make it possible to compare evolved bacterial genomes to their ancestors at the assembly level versus by mapping reads to a reference genome. *seabreeze* will be of utility to researchers investigating how and why the organization of bacterial genomes evolves.

Acknowledgments

Development of *seabreeze* was supported by the National Science Foundation (DEB-1951307). We acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing computing resources. Ira Zibbu acknowledges additional support from the Khorana Scholars Program and the DST INSPIRE Fellowship. Claus Wilke acknowledges support from the Blumberg Centennial Professorship in Molecular Evolution at the University of Texas at Austin. We thank the developers of `plotsr` for making their code available under the MIT license.

References

Achaz, G., Coissac, E., Netter, P., & Rocha, E. P. C. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics*, *164*(4), 1279–1289. https://doi.org/10.1093/genetics/164.4.1279

Ahsan, M. U., Liu, Q., Perdomo, J. E., Fang, L., & Wang, K. (2023). A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nature Methods*, *20*(8), 1143–1158. https://doi.org/10.1038/s41592-023-01932-w

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G., & Parkhill, J. (2005). ACT: The artemis comparison tool. *Bioinformatics*, *21*(16), 3422–3423. https://doi.org/10.1093/bioinformatics/bti553

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Conda-Forge Community. (2015). *The conda-forge project: Community-based software distribution built on the conda package format and ecosystem*. Zenodo. https://doi.org/10.5281/ZENODO.4774216

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, *5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. In L. Sun & W. Shou (Eds.), *Engineering and analyzing multicellular systems* (Vol. 1151, pp. 165–188). Springer New York. https://doi.org/10.1007/978-1-4939-0554-6_12

Goel, M., & Schneeberger, K. (2022). Plotsr: Visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*, *38*(10), 2922–2926. https://doi.org/10.1093/bioinformatics/btac196

Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, *20*(1), 277. https://doi.org/10.1186/s13059-019-1911-0

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Rocha, E. P. C. (2004). The replication-related organization of bacterial genomes. *Microbiology*, *150*(6), 1609–1627. https://doi.org/10.1099/mic.0.26974-0

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

The Bioconda Team, Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Vaser, R., & Šikić, M. (2021). Time- and memory-efficient genome assembly with raven. *Nature Computational Science*, *1*(5), 332–336. https://doi.org/10.1038/s43588-021-00073-4

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., Wyres, K. L., & Holt, K. E. (2021). Trycycler: Consensus long-read assemblies for bacterial genomes. *Genome Biology*, *22*(1), 266. https://doi.org/10.1186/s13059-021-02483-z

Xie, Z., & Tang, H. (2017). ISEScan: Automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*, *33*(21), 3340–3347. https://doi.org/10.1093/bioinformatics/btx433