

# Segzoo: a turnkey system that summarizes genome annotations

Mickaël Mendez<sup>1,2, </sup>, Yushan Liu<sup>2,3,4, </sup>, Marc Asenjo Ponce de León<sup>2,5, </sup>, Michael M. Hoffman<sup>1,2,6,7,\*, </sup>

<sup>1</sup>Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada

<sup>2</sup>Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G 1L7, Canada

<sup>3</sup>Department of Mathematics, University of Toronto, Toronto, ON M5S 2E4, Canada

<sup>4</sup>Department of Statistical Sciences, University of Toronto, Toronto, ON M5G 1Z5, Canada

<sup>5</sup>Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain

<sup>6</sup>Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 1L7, Canada

<sup>7</sup>Vector Institute, Toronto, ON M5G 0C6, Canada

\*Corresponding author. Princess Margaret Cancer Centre, University Health Network, Princess Margaret Cancer Research Tower 11-311, 101 College St, Toronto, ON M5G 1L7, Canada. E-mail: michael.hoffman@utoronto.ca

Associate Editor: Peter Robinson

## Abstract

**Motivation:** Segmentation and automated genome annotation (SAGA) techniques, such as Segway and ChromHMM, assign labels to every part of the genome, identifying similar patterns across multiple genomic input signals. Inferring biological meaning in these patterns remains challenging. Doing so requires a time-consuming process of manually downloading reference data, running multiple analysis methods, and interpreting many individual results.

**Results:** To simplify these tasks, we developed the turnkey system Segzoo. As input, Segzoo only requires a genome annotation file in browser extensible data (BED) format. It automatically downloads the rest of the data required for comparisons. Segzoo performs analyses using these data and summarizes results in a single visualization.

**Availability and implementation:** The source code for Python  $\geq 3.7$  on Linux is freely available for download at <https://github.com/hoffmangroup/segzoo> under the GNU General Public License (GPL) version 2. Segzoo is also available in the Bioconda package segzoo: <https://anaconda.org/bioconda/segzoo>. We have deposited in Zenodo the version of the Segzoo source which produced the results in this article (<https://doi.org/10.5281/zenodo.10988775>), other code and data used to produce the results (<https://doi.org/10.5281/zenodo.10477083>), and the results (<https://doi.org/10.5281/zenodo.10477106>).

## 1 Introduction

Segmentation and automated genome annotation (SAGA) methods (Libbrecht et al. 2021), such as Segway (Hoffman et al. 2012) and ChromHMM (Ernst and Kellis 2012), partition the genome into segmentation made of non-overlapping segments. The methods assign a label to each segment such that the segments sharing the same label exhibit similarities in the input data. As the input data can consist of complex combinations of chromatin immunoprecipitation sequencing (ChIP-seq) data (Barski et al. 2007, Johnson et al. 2007, Robertson et al. 2007) or RNA sequencing (RNA-seq) data, inferring a biological role for a label requires comparing the segmentation against multiple reference datasets. This requires a non-trivial amount of toil (Rau 2016), including downloading reference data. It also involves orchestrating the use of these data with a variety of tools, such as Segtools (Buske et al. 2011) and BEDTools (Quinlan and Hall 2010), potentially through the development of custom scripts. Interpretation requires simultaneous consideration of multiple visualizations and analyses.

To help assign biological hypotheses to SAGA labels, we designed the turnkey system Segzoo. Segzoo requires only a genome segmentation file as input to generate all the visualizations. Segzoo uses Go Get Data (GGD; Cormier et al. 2021) to automatically download all required data for these analyses and produces an easy to interpret figure which reveals patterns of segmented regions. Here, we demonstrate the use of Segzoo to quickly identify promoter features from a Segway annotation of the H1 human embryonic stem cell line.

## 2 Methods

Segzoo summarizes genome annotation through a four-step workflow. First, it uses GGD to download reference data for the genome assembly of the segmentation. Second, Segzoo generates segmentation-centric summary statistics using Segtools and BEDTools (Quinlan and Hall 2010). Third, Segzoo parses the

Received: 6 October 2023. Revised: 27 February 2024. Accepted: 19 March 2024

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

output of each tool and summarizes them into one table. Fourth, Segzoo plots the table as a compact figure to allow rapid biological hypothesis assignment to SAGA labels.

## 2.1 Dependencies

In addition to a segmentation, Segzoo needs a gene annotation and a reference genome to perform analyses. Segzoo uses GGD to automatically download the data dependencies and Conda (Conda contributors 2022; <https://docs.conda.io/projects/conda/>) to automatically install the software dependencies. Segzoo uses the Snakemake (Köster and Rahmann 2012) workflow manager to automatically perform the necessary steps. Using Snakemake also allows taking advantage of multiprocessing to accelerate the analyses.

## 2.2 Analysis tools

Segzoo uses Segtools to perform a series of analyses. It combines the results of `segtools-aggregation`, `segtools-gmtk-parameters`, `segtools-overlap`, and `segtools-length-distribution` into a single plot. Segzoo also uses BEDTools's (Quinlan and Hall 2010) `bedtools nuc` command to calculate the nucleotide frequency and G + C content for each segmentation label.

## 2.3 Input data

Here, we demonstrate the use of Segzoo by creating a Segway annotation of the H1 human embryonic stem cell line. To do this, we used ENCODE ChIP-seq data for CTCF (ENCFF332TNJ), H3K27me3 (ENCFF417VQQ), H3K4me1 (ENCFF584AVI), H3K27ac (ENCFF919FBG), and H3K4me3 (ENCFF760NUN). We trained the 10-label Segway model with the following parameters: `segway --num-instances=10 --resolution=10 --minibatch-fraction=0.01 --num-labels=10`. We set the environment variable `SEGWAY_RANDOM_SEED=22426492` to make the results reproducible.

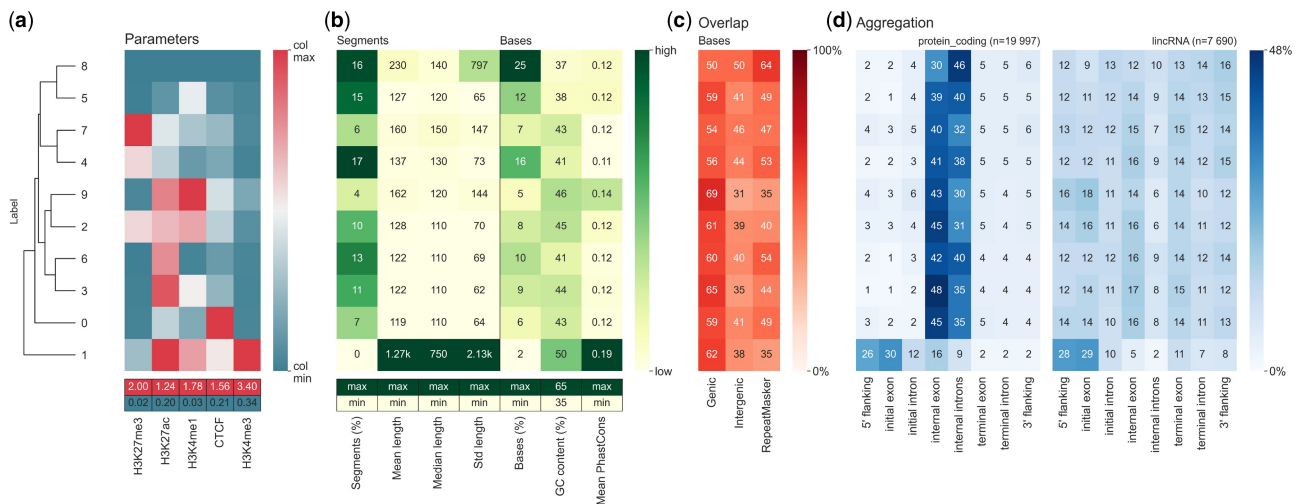
We used Segzoo to generate a detailed summary figure (Fig. 1) with the following single command:

```
segzoo --parameters traindir/params/params.params \
--dendrogram annotdir/segway.bed.gz
```

To plot the Parameters heatmap (Fig. 1a), we used the `--parameter` option with Segway's trained parameter file, `params.params`. The file contains the mean of the Gaussian distribution learned by Segway for each pair of dataset and label. The other heatmaps (Fig. 1b–d) used the output of `segway annotate, segway.bed.gz`.

## 3 Results

The Segzoo visualization consists of an array of heatmaps, with rows representing SAGA labels and columns representing different quantitative descriptions of those labels (Fig. 1). The first



**Figure 1** Segzoo summary of Segway H1 human embryonic stem cell line annotation. Numeric values in each heatmap represent various summary statistics for a label. (a) *Top*: Gaussian mean parameters from Segway trained model. Column-normalized values, with minimum and maximum values mapped to the range 0 to 100. *Left*: Dendrogram indicating hierarchical clustering on rows from weighted pair group method with arithmetic average (WPGMA; Sokal and Michener 1958). *Bottom*: Maximum and minimum values of each column. (b) *Top*: Heatmap of statistics summarizing segments. Grouped by segment: fraction of segments annotated with the label (“segments”), mean segment length (“mean length”), median segment length (“median length”), standard deviation of segment length (“std length”). Grouped by base: fraction of bases annotated with the label (“base pairs”), G + C content, mean PhastCons (Siepel et al. 2005) conservation score (“mean PhastCons”). The letter “k” indicates thousands. *Bottom*: Indication of maximum and minimum values of the color scale for each column from one of: column maximum value (“max”), column minimum value (“min”), other value (a number). (c) Fractional overlap of each label with genic, intergenic, and RepeatMasker (Smit et al. 2010; <https://www.repeatmasker.org/>) regions. (d) Fractional overlap between segments and components and flanking regions of GENCODE (Harrow et al. 2012) protein-coding genes (left) and long intergenic non-coding RNA (lincRNA) genes (right). Row-normalized values per table, with values summing up to 100.

heatmap, “Parameters”, indicates the mean parameters of the Gaussian distributions learned for each label on each input dataset (Fig. 1a). The other heatmaps summarize other aspects of the output segmentation, including summary statistics of the segments (Fig. 1b), proportion of bases in genic regions (Fig. 1c), and overlap between segments and components of GENCODE genes—both protein-coding genes (Fig. 1d, left), and lincRNAs (Fig. 1d, right). To simplify the interpretation of labels with similar patterns, Segzoo automatically performs row-wise and column-wise hierarchical clustering on the Parameters table. All the heatmaps share the same row order to simplify the interpretation of all the descriptions for a given label.

Here, Segzoo allows us to quickly see that label 1 has several promoter features (Fig. 1). First, label 1 has the highest mean H3K4me3 signal (Fig. 1a). This signal often indicates promoter regions (Liang et al. 2004). Second, label 1 has the highest G + C content (Fig. 1b), another feature of promoter regions (Akan and Deloukas 2008). Label 1 has the highest mean length and median length. We don't expect this for a promoter region. Label 1, however, has the highest length standard deviation and the fewest segments. This suggests that Segway annotated not only promoter regions, but also long genomic regions with the label 1. Third, the Overlap table shows the segments with the label 1 mostly annotate genic regions (Fig. 1c). Fourth, the Aggregation table shows an enrichment of label 1 at the 5' end of both protein coding genes and lincRNAs (Fig. 1d). The four heatmaps together help generate the hypothesis that label 1 annotates promoter regions. Interestingly, we can also see easily that label 9 has promoter features such as high H3K4me3 signal, high G + C content, and frequent overlap at lincRNA 5' ends.

## 4 Conclusion

Segzoo simplifies the process of assigning a biological hypothesis to SAGA labels. To simplify the process, it generates a compact visualization that summarizes the overlap between a SAGA annotation and multiple reference datasets. Segzoo automatically download the reference datasets via standardized recipes with GGD, which makes it easier to attain higher degrees of reproducibility. For example, this work meets the silver standard of reproducibility (Heil et al. 2021).

## Acknowledgments

We thank Carl Virtanen and Zhibin Lu at the University Health Network High-Performance Computing Centre and Bioinformatics Core for technical assistance.

## Author contributions

Mickaël Mendez (Conceptualization [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [supporting], Software [lead], Supervision [supporting], Validation [lead], Visualization [lead], Writing—original draft [lead]), Yushan Liu (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Software [supporting], Validation [supporting], Visualization [supporting], Writing—

original draft [supporting]), Marc Asenjo Ponce de León (Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Software [supporting], Validation [supporting], Visualization [supporting], Writing—original draft [supporting]), and Michael M. Hoffman (Conceptualization [lead], Funding acquisition [lead], Project administration [lead], Supervision [lead], Writing—review & editing [lead])

## Conflict of interest

None declared.

## Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H.).

## References

- Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. *Gene* 2008; **410**:165–76.
- Barski A, Cuddapah S, Cui K et al. High-resolution profiling of histone methylations in the human genome. *Cell* 2007; **129**:823–37.
- Buske OJ, Hoffman MM, Ponts N et al. Exploratory analysis of genomic segmentations with segtools. *BMC Bioinformatics* 2011; **12**:415.
- Conda contributors. Conda, 2022.
- Cormier MJ, Belyeu JR, Pedersen BS et al. Go Get Data (GGD) is a framework that facilitates reproducible access to genomic data. *Nat Commun* 2021; **12**:2151. <https://doi.org/10.1038/s41467-021-22381-z>
- Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012; **9**:215–6.
- Harrow J, Frankish A, Gonzalez J et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res* 2012; **22**:1760–74.
- Heil BJ, Hoffman MM, Markowitz F et al. Reproducibility standards for machine learning in the life sciences. *Nat Methods* 2021; **18**:1132–5.
- Hoffman MM, Buske OJ, Wang J et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012; **9**:473–6.
- Johnson DS, Mortazavi A, Myers RM et al. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007; **316**:1497–502.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012; **28**:2520–2.
- Liang G, Lin JC, Wei V et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci USA* 2004; **101**:7357–62.
- Libbrecht MW, Chan RC, Hoffman MM. Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput Biol* 2021; **17**:e1009423.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**:841–2.

- Rau V. Eliminating toil. In: Beyer B, Jones C, Petoff J *et al.*, editors., *Site Reliability Engineering: How Google Runs Production Systems*, chapter 5. O'Reilly, 2016.
- Robertson G, Hirst M, Bainbridge M *et al.* Genome-wide profiles of *STAT1* DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;**4**:651–7.
- Siepel A, Bejerano G, Pedersen JS *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50.
- Smit A, Hubley R, Green P. RepeatMasker Open-4.0, 2010.
- Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 1958;**38**:1409.