Continuous Chromatin State Feature Annotation of the Human Epigenome

SIMON FRASER UNIVERSITY

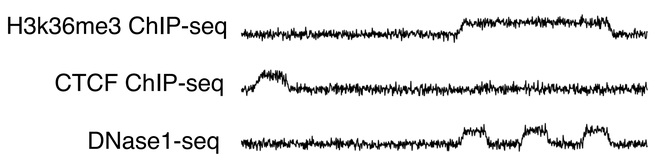
Habib Daneshpajouh*, Bowen Chen*, Neda Shokraneh, Shohre Masoumi, Kay C. Wiese, Maxwell W Libbrecht

* Equal contribution

Semi-automated genome annotation (SAGA)

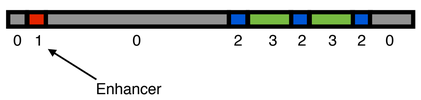
Input: Real-valued functional genomics data tracks defined over the genome, from a single cell type.

Output: Annotation that summarizes the regulatory activity at each base pair, in that cell type.



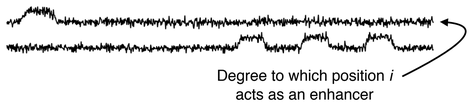
Previous annotations: Discrete chromatin state labels.

- Each position receives a single discrete label.
- Examples: HMMSeg, ChromHMM, Segway.

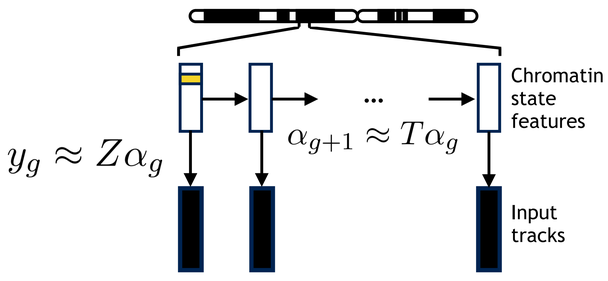


Proposed annotations: Continuous chromatin state features.

- Each position is represented by a vector of features representing the strength of multiple types of activities.

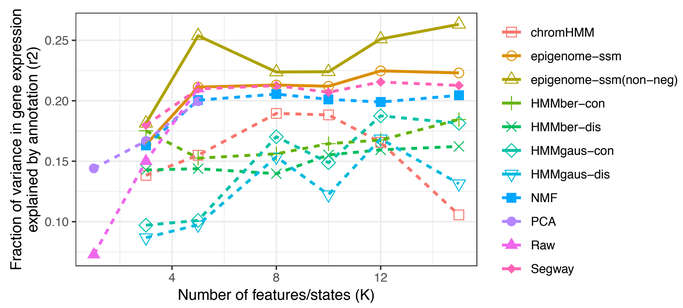


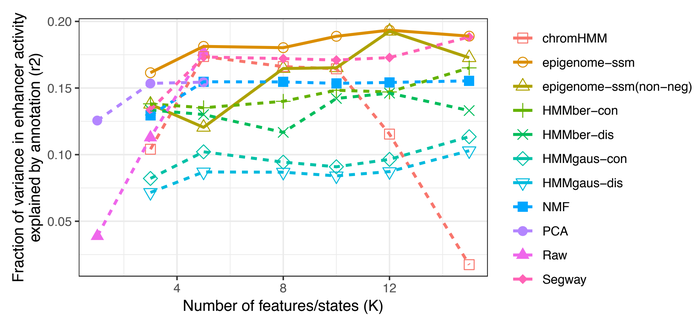
Method: Nonnegative Kalman filter state space model (epigenome-ssm)



We learn the epigenome-ssm model from data using an EM-like message-passing algorithm.

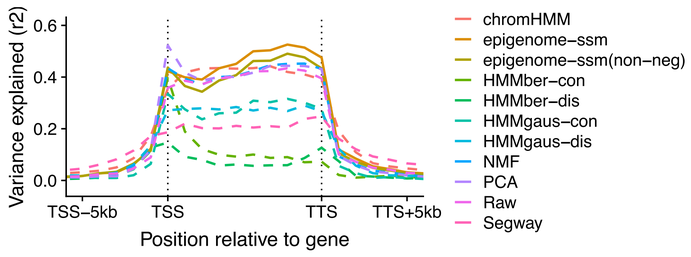
Chromatin state features are predictive of known genomic phenomena





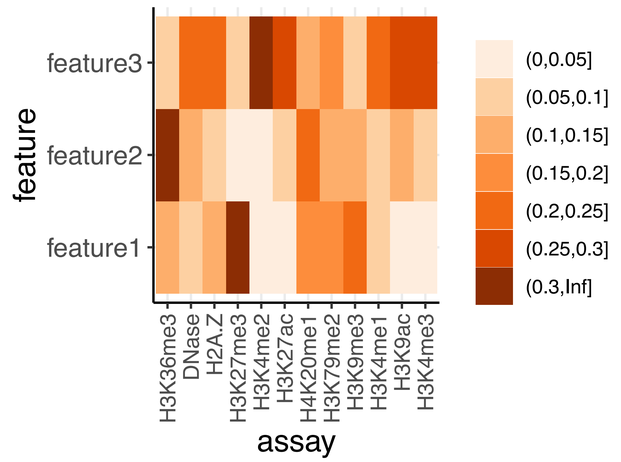
Alternative methods:

- Segway/chromHMM: Well-known annotation methods
- HMM<gaus/ber>-<dis/con>: HMM model.
  - gaus: Takes continuous tracks as input (similar to Segway).
  - ber: Takes binarized tracks as input (similar to ChromHMM).
  - dis: Discrete chromatin state labels as output.
  - con: Continuous probability tracks as output.
- PCA/NMF: Dimensionality reduction methods. Principal component analysis and non-negative matrix factorization.

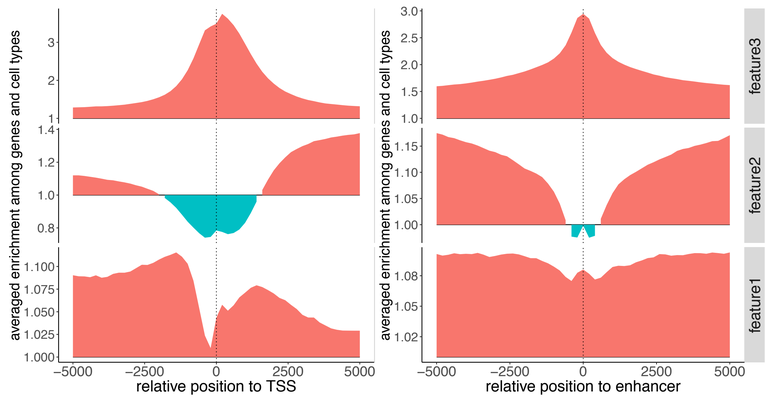


Chromatin state features recapitulate known genome biology

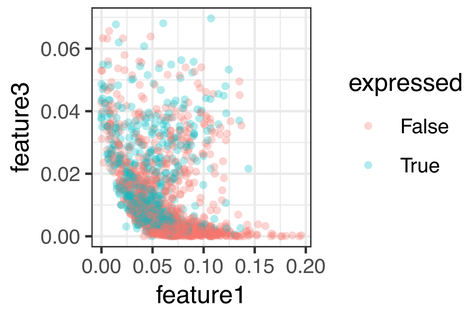
- Feature 1: Repression.
- Feature 2: Transcription-specific activity.
- Feature 3: General regulatory activity.



Promoters and enhancers are marked by distinctive chromatin state feature patterns.



Continuous chromatin state features enable expressive visualizations.



Summary

- Semi-automated genome annotation (SAGA) methods summarize a set of epigenomics assays (such as ChIP-seq).
- Existing SAGA methods output an annotation of the genome that assigns a chromatin state label to each genomic position.
- We propose an annotation strategy (epigenome-ssm) that instead outputs a vector of chromatin state features at each position rather than a single discrete label.
- Advantages of continuous chromatin state features: They (1) can capture varying strength of elements; (2) can capture combinatorial patterns of activity (such as intronic enhancers); and (3) preserve the underlying continuous nature of the input data.

Chen et al. (2018). Continuous chromatin state feature annotation of the human epigenome. bioRxiv, 473017.