STARRPeaker: Uniform processing and accurate identification of STARR-seq active regions

Donghoon Lee[1,2], Manman Shi[3], Jennifer Moran[3], Martha Wall[3], Jing Zhang[1,2], Jason Liu[2], Dominic Fitzgerald[3], Yasuhiro Kyono[3], Lijia Ma[3,4], Kevin P White[3,5]*, Mark Gerstein[1,2,6,7]*

[1] Program in Computational Biology and Bioinformatics [2] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[3] Institute for Genomics and System Biology, University of Chicago, Chicago, IL, 60637, USA [4] School of Life Sciences, Westlake University, Hangzhou, 310024, China
[5] Tempus Labs, Inc. Chicago IL 60654, USA [6] Department of Computer Science [7] Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

Background: High-throughput reporter assays, such as self-transcribing active regulatory region sequencing (STARR-seq), allow for unbiased and quantitative assessment of enhancers at a genome-wide level. Recent advances in STARR-seq technology have employed progressively more complex genomic libraries and increased sequencing depths, to assay larger sized regions, up to the entire human genome. These advances necessitate a reliable processing pipeline and peak-calling algorithm.

Methods and Results: Most STARR-seq studies have relied on chromatin immunoprecipitation sequencing (ChIP-seq) processing pipeline to identify peaks. However, there are key differences in STARR-seq versus ChIP-seq data: STARR-seq uses transcribed RNA to measure enhancer activity, making determining the basal transcription rate important. Furthermore, STARR-seq output coverage is non-uniform, overdispersed (**Fig 1**), and often confounded by sequencing biases such as GC content and mappability. Moreover, here, we observed a clear correlation between RNA thermodynamic stability and STARR-seq readout, suggesting that STARR-seq might be sensitive to RNA secondary structure and stability (**Fig 2**). Considering these findings, we developed STARRPeaker: a negative binomial regression framework for uniformly processing STARR-seq data. We applied STARRPeaker to two whole human genome STARR-seq experiments; HepG2 and K562.

Available @ github.com/gersteinlab/starrpeaker

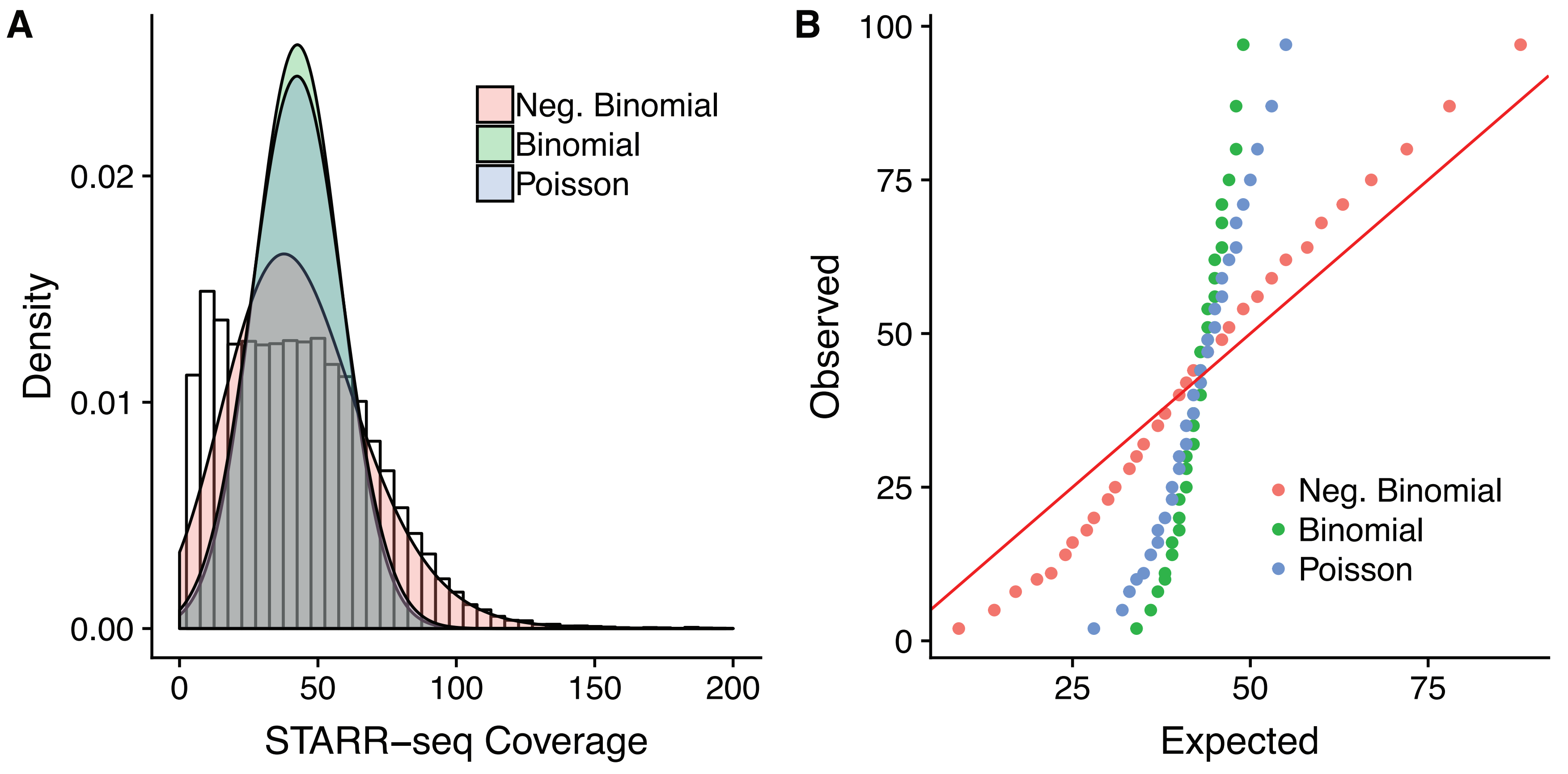

Figure 1. Modeling of STARR-seq fragment coverage

$$Y_i \sim NB(\mu_i, \theta)$$

expected count   dispersion

$$\ln \mu_i = (\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki})$$

$$\mu_i = \exp\underbrace{(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki})}_{\text{covariates: controls for confounding variables}}$$

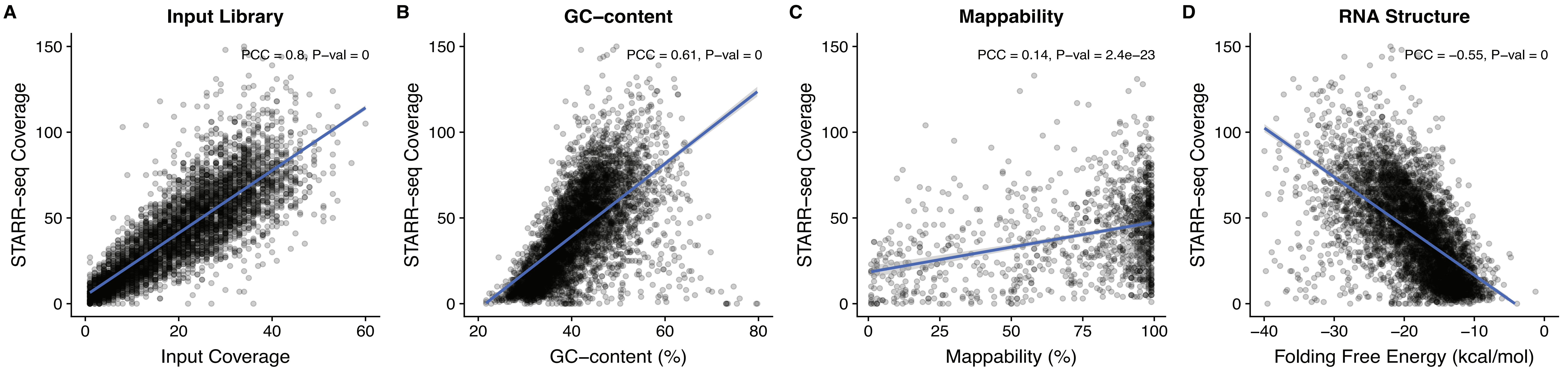

Figure 2. Confounding factors in the STARR-seq assay

Conclusion: Our method identifies highly reproducible and epigenetically active enhancers across replicates (**Fig 3**). Moreover, STARRPeaker outperforms other peak callers in terms of identifying known enhancers. Thus, our framework optimized for processing STARR-seq data accurately characterizes cell-type-specific enhancers, while addressing potential confounders.

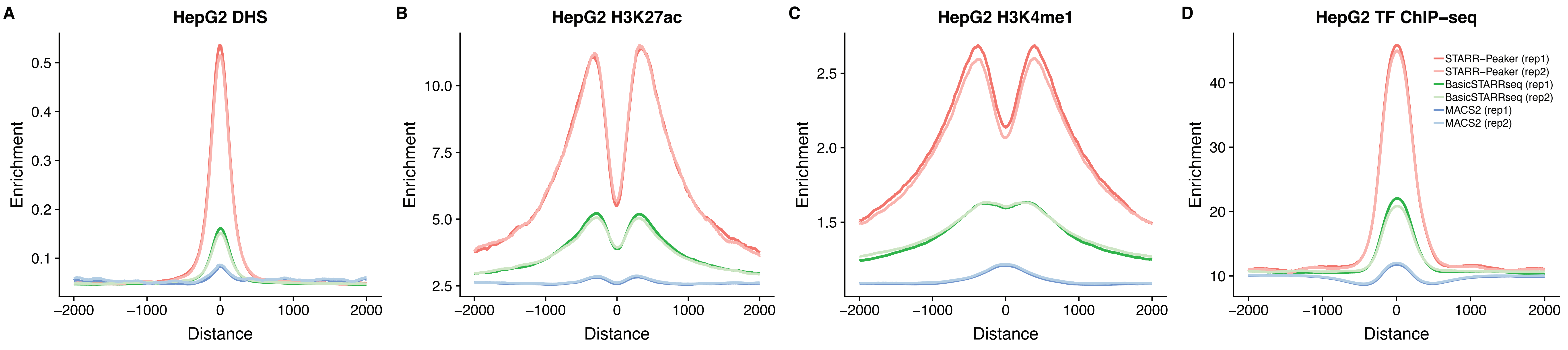

Figure 3. Enrichment of epigenetic signals around STARRPeaker peaks

Yale Gerstein Lab

Post-doctoral position in bioinformatics available. For more info: jobs.gersteinlab.org