Probabilistic Inference for ChIP-seq by

Xuekui Zhang Department of Statistics, University of British Columbia, Vancouver, BC, Canada
xzhang@stat.ubc.ca

Abstract

ChIP-seq, which combines chromatin immunoprecipitation with massively parallel short-read sequencing, can profile in vivo genome-wide transcription factor-DNA association with higher sensitivity, specificity and spatial resolution than ChIP-chip. While it presents new opportunities for research, ChIP-seq poses new challenges for statistical analysis that derive from the complexity of the biological systems characterized and the variability and biases in its digital sequence data. We propose a method called PICS (Probabilistic Inference for ChIP-seq) for extracting information from ChIP-seq aligned-read data in order to identify regions bound by transcription factors. PICS identifies enriched regions by modeling local concentrations of directional reads, and uses DNA fragment length prior information to discriminate closely adjacent binding events via a Bayesian hierarchical t-mixture model. Its per-event fragment length estimates also allow it to remove from analysis regi! ons that have atypical lengths. PICS uses pre-calculated, whole-genome read mappability profiles and a truncated t-distribution to adjust binding event models for reads that are missing due to local genome repetitiveness. It estimates uncertainties in model parameters that can be used to define confidence regions on binding event locations and to filter estimates. Finally, PICS calculates a per-event enrichment score relative to a control sample, and can use a control sample to estimate a false discovery rate. We compared PICS to the alternative methods MACS, QuEST, and CisGenome, using published GABP and FOXA1 data sets from human cell lines, and found that PICS' predicted binding sites were more consistent with computationally predicted binding motifs.

This is joint work with Gordon Robertson, Martin Krzywinski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo.

Keywords: Bayesian hierarchical model; ChIP-seq; EM algorithm; High-throughput Sequencing; Missing values; Mixture model; Transcription factor; Truncated data; t-distribution.