penalized importance sampling. The specific methods are as follows.


do not hallucinate.
transitions between those bins. If a boundary corresponds to the beginning of a domain, the boundary score is the posterior probability of a background to foreground transition and vice versa.

Next an empirical distribution of posterior transition probabilities is constructed by computing posterior transition probabilities from a dataset of randomly permuted bins with the same HMM parameters. Those bins whose posterior transition probabilities have significant empirical P-values are kept and consecutive significant bins are joined as being one boundary. We score each boundary with the posterior probability that a single transition occurs in this boundary. The peak of a boundary is set to the start of the bin with the largest transition probability (see Supplementary Section S3 for details).

Incorporating a control sample: ChIP-Seq experiments are influenced by background noises, contamination and other possible sources of error, and researchers have begun to realize the necessity of generating experimental controls in ChIP-Seq experiments. Two common forms of control exist: a non-specific antibody such as IgG to control the immunoprecipitation, and sequencing of whole cell extract to control for contamination and other possible sources of error. With the availability of a control sample, we use a similar two-state HMM with the novel NBDiff distribution to describe the relationship between the read counts in the two samples. Analogous to the Skellam distribution (Skellam, 1946), the NBDiff distribution describes the difference of two independent negative binomial random variables (see Supplementary Section S1.2 for details).

Simultaneously segmenting two modifications: the simultaneous analysis of two histone modification marks may reveal more accurate information about the status of genomic regions. It helps to understand the functions of different histone modification marks. It is also of interest to compare samples from different cells types because histone modification patterns are dynamic and subject to change during cell differentiation. We use the NBDiff distribution to model the read count difference between the two samples, and employ three-state HMM: where the basal state means these two signals are similar, the second state represents the signal in test sample A is greater than that in the test sample B and the third state represents the opposite case (details given in Supplementary Section S2.1).

3 EVALUATION AND APPLICATIONS

We simulated H3K36me3 ChIP-Seq data and compared RSEG, SICER (Zang et al., 2009) and HPeak (Qin et al., 2010). In terms of domain identification, RSEG outperforms SICER and HPeak for both single-sample analysis and yields comparable results to SICER for analysis with control samples (Supplementary Section S4.1 and 4.2). We applied RSEG to H3K36me3 ChIP-Seq dataset from (Barski et al., 2007) and found a strong association between H3K36me3 domain boundaries with TSS and transcription termination site (TTS), which supports that RSEG can find high-quality domain boundaries (Supplementary Section S4.3).

We applied RSEG to four histone modification marks (H3K9me3, H3K27me3, H3K36me3 and H3K79me2) from two separate studies (Barski et al., 2007; Mikkelsen et al., 2007) (Supplementary Section S5.1). In particular, we discovered an interesting relationship between the two gene-overlapping marks H3K36me3 and H3K79me2 through boundary analysis. H3K79me2 tends to associate with 3′-ends of genes, while H3K36me3 associates with 3′-ends. About 41% of gene-overlapping K79 domains cover TSS in contrast to 11% of K36 domains. On the other hand, 84% of K36 domains cover TTS in contrast to 23% of K79 domains (Table 1). In those genes with both H3K36me3 and H3K79me2 signals, H3K79me2 domains tend to precede H3K36me3 domains, for example the DFP2 gene (Fig. 1) (see Supplementary Section S5.2 for more information). This novel discovery demonstrates the usefulness of boundary analysis for dispersed histone modification marks.

Finally we applied our three-state HMM to simultaneously analyze H3K36me3 and H3K79me2 (Supplementary Section S5.4). The result agrees with the above observations. The application of our three-state HMM to find differentially histone modification regions is given in Supplementary Section S5.3.

Conflict of Interest: none declared.

REFERENCES

Barks,A. et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell, 129, 825–837.
Kouzarides,T. (2007) Chromatin modifications and their function. Cell, 129, 603–705.
Mikkelsen,T.S. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 448, 553–560.
Pepke,S. et al. (2009) Computation for ChIP-seq and RNA-seq studies. Nat. Meth., 6, 822–832.
Qin,Z. et al. (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinformatics, 11, 369.
Rubin,E.L. (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77, 257–286.
Skellam,J.G. (1946) The frequency distribution of the difference between two poison variates belonging to different populations. J. R. Stat. Soc. Ser. A, 109, 296.
Zang,C. et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics, 25, 1952–1958.

Table 1. Location of H3K36me3 and H3K79me2 domain boundaries relative to genes

| Boundaries (5′ → 3′) | K79 (%) | K36 (%) |
|----------------------|---------|---------|
| Upstream TSS → Inside Gene | 31 | 3 |
| Upstream TSS → Downstream TTS | 10 | 8 |
| Inside Gene → Inside Gene | 46 | 13 |
| Inside Gene → Downstream TTS | 13 | 76 |

Fig. 1. The H3K36me3 and H3K79me2 domains and their boundaries at DFP2 (chr11:64,854,646–6,880,304).