BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–2

Analyzing ChIP-seq data with Pique

R.Y. Neches [1]*, E.G. Wilbanks [1] and M.T. Facciotti [2]

[1]Microbiology Graduate Group, University of California, Davis.
[2]Department of Biomedical Engineering, University of California, Davis.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:

It was found that most peak finders designed for ChIP-seq experiments are designed for ChIP-seq in eukaryotes. To make cost-effective use of current sequencing capacity, they must be cleverly optimized to work with sparse-coverage data, and must take into account the effect of chromatin structure on the variation in background coverage. For ChIP-seq in bacteria and archaea, the differences in genome structure, organization and size make the models used for identifying coverage enrichment in eukaryotic poorly suited for use in bacteria and archaea. For example, on this data, CSDeconv has a running time on the order of days for a single ChIP-seq mapping.

Fortunately, many of the statistical challenges for ChIP-seq in eukaryotes are simply not present in experiments using bacterial and archaeal models; this is due in part to higher genome coverage – typically in inverse proportion to genome size – and in part to the absence of non-random coverage variation due to highly structured chromatin.

Results:

Availability:

Contact: ryneches@ucdavis.edu

1 INTRODUCTION

2 APPROACH

The 40-bp reads were quality filtered and quality trimmed and aligned to the *Halobacterium salinarum sp. NRC1* reverence genome using bowtie. Reads mapped in the forward and reverse orientation were separated, and used to calculate distinct coverage tracks.

ChIP-seq in archaea yields coverage several orders of magnitude larger than in eukaryotic systems, and so we designed a system to take advantage of this fact. Given the importance of manual curation and settings optimization, Pique provides output suitable for use in the Gaggle Genome Browser. This permits convenient interactive curation of the peak list.

3 METHODS

For simplicity of design, we have employed several standard algorithms and design principles from signal processing.

Fig. 1. Caption, caption.

- Raw data is normalized with respect to the background. The user selects one or more regions which are representative of the un-enriched background. Resequenced genomes often have coverage artifacts caused by features such as repetitive regions and gene dosage variation. For this reason, we advise the use of curated background regions. This operation is carried out by `piquify.py`.
- A mask is applied to the ChIP track to remove regions with ambiguous read mapping. For example, it is impossible to map reads to unique loci in highly repetitive or palindromic regions, such as IS elements. As a result, the coverage is impossible to measure unambiguously, and the regions must be excluded from downstream analysis. (`pique.py`)
- The "DC" component is removed using linear detrending (scipy.detrend). This removes effects due to coverage variation features larger than about 100Kb. (`pique.py`)
- High-k noise in coverage is removed using a Wiener-Kolmogorov filter. The filter delay α is chosen to approximate to the expected footprint size of the targeted protein. (`pique.py`)
- A coverage amplitude cutoff calculated from the detrended background track such that any given locus is equally likely to be above or below the cutoff. Enrichment features are defined with respect to this coverage level (`pique.py`)
- A sliding window moving average is used to identify regions whose coverage level deviates from the background. Peaks usually contain gaps in coverage that with widths on the order of the experimentally selected fragment size; the window width is chosen to correspond to this size. (`pique.py`)

These steps yield simple rectangular envelopes around putative regions of enrichment. To determine if these enriched regions correspond to binding events, we apply a very simple statistical model :

- Coordinates of enriched regions in a peak are offset between strands, with the forward strand enriched upstream of the reverse strand. The first condition of the model is that the envelops must be overlapping rectangles; the end coordinate of the forward strand envelope must fall within the reverse strand envelope, and the start coordinate of the reverse strand envelope must fall within the forward strand envelope.
- Enrichment that are thought to represent binding events produce a characteristic shape envelope, which we model using a sum over set of Gaussians.

*to whom correspondence should be addressed

4 DISCUSSION

If the putative peak passes all of the above tests, this means that the peak "looks" like a peak. To make sure that we are not finding horsies by gazing at clouds, we also require that the integral of the coverage in the raw data within the putative peak region exceeds the integral of the coverage in the background by a margin set by the user. (Other tests for statistical significance may also work, be more shiny, et cetera. For example, Monte Carlo simulations of random subsamples of the ChIP track and the background track until a coalescent is found.)

5 CONCLUSION

The choice of filter implies some specific assumptions about the nature of the coverage noise. The Wiener-Kolmogorov filter was the first and simplest statistical signal filter, first published by Norbert Wiener in 1949, and independently derived in discrete-time form by Andrey Kolmogorov in 1941. The approach assumes the existence of two inputs; a "true" signal, and a noise source. Both are assumed to be stationary stochastic processes combined additively.

ACKNOWLEDGEMENT

Funding:

REFERENCES