BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–2

Analyzing high coverage microbial ChIP-seq data with Pique

R.Y. Neches [1,3]*, E.G. Wilbanks [1,3] and M.T. Facciotti [2,3]

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:
Most ChIP-Seq peak finders are designed to protein-DNA binding events in eukaryotic datasets. To make cost-effective use of current sequencing capacity, the peak finders 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. While numerous effective peak finders have been developed for eukaryotic data, these algorithmic approaches can be suprisingly error prone in our hands when run on high-coverage bacterial and archaeal ChIP-Seq datasets.

Results:
Fortunately, many of the statistical challenges for peak detection inherent in eukaryotic ChIP-seq data are not present in bacterial and archaeal datasets; 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. In response, we have developed Pique, a conceptually simple, easy to run ChIP-Seq peak finding pipeline for bacterial and archaeal ChIP-Seq data. The software is cross-platform and Open Source, and based on Open Source dependencies. Output is easily imported into the Gaggle Genome Browser for manual curation of peaks and the exploration of the dataset in the context of Gaggle-enabled resources.

Availability:
The software is available under the BSD-3 license at
http://github.com/ryneches/pique.
A tutorial and test data are included with the documentation.

Contact: ryneches@ucdavis.edu

1 INTRODUCTION

Next generation sequencing coupled with chromatin immunoprecipitation (ChIP-Seq) is revolutionizing our abilty to genomically map protein-DNA associations for a wide variety of proteins. The growing popularity of ChIP-Seq has spurred the development of numerous peak picking algorithms. All but one, (CSDeconv (REF)) of the peak picking algorithms have been designed to perform well on eukaryotic datasets. To make most cost-effective use of current sequencing abilities the peak picking methods have employed a number of sophisticated strategies to detect peaks in the typically sparcely covered eukaryotic datasets for which they are designed. The problem of finding peaks in such data is apparently so challenging that over 30 different methods have been published since 2007. The relative performance of 11 represetative peak detection algorithms on eukatyoric data has been recently reviewed by Wilbanks and Facciotti (REF).

While ChIP-Seq has been predominantly used to interrogate protein-DNA interactions in eukaryotic systems (REFS) there are clear advantages to adopting this technology for studying microbial systems that are largely associated with the relatively small sizes of microbial genomes (the genome of *E. coli* is ≈2000 times smaller than the human genome). In our hands, peak detection algorithms designed for use in eukaryotic ChIP-Seq experiments are very prone to errors on microbial ChIP-Seq data. This is surprising; eukaryotic ChIP-Seq necessarily involves more challenging biochemical and statistical approaches than microbial ChIP-Seq.

To our knowledge, only one other peak finding package, CSDeconv (REF) has been explicitly developed for finding peaks in microbial ChIP-Seq data. This MATLAB package does a good job (WE THINK) of finding peaks in microbial ChIP-Seq data but it's broad applicaiton is limited by the requirement for the proprietary software package MATLAB, very slow performance and lack of support for manual curation.

Herein we describe Pique, a conceptually simple, Python-based peak finding package that enables easy and rapid peak finding in bacterial and archaeal ChIP-Seq datasets. The output is easily imported into the Gaggle Genome Browser (REF) to enable rapid manual peak list curation and analysis of ChIP-Seq data in the context of other Gaggle-enabled (REF) resources. Furthermore, many interesting organisms have genomic complexities, such as IS elements, gene dosage polymorphisms and accessory genomes that cause variations in sequence coverage unrelated to ChIP. This may also occur if the organism under study is not identical to the reference genome. The resulting enrichment "pedestals" and "holes" can confound peak calling, and in most cases must be processed separately. Pique allows the user to optionally supply a map of these features as a GFF file, and will automatically perform a segmented analysis.

*to whom correspondence should be addressed

2 APPROACH

ChIP-seq in bacteria and archaea yields coverage several orders of magnitude larger than in eukaryotic systems. This generates data with near-continuous signal across the microbial chromosome rather than the sparse coverage typically present in eukaryotic ChIP-Seq data. This feature of microbial ChIP-Seq data permits simpler, faster algorithms to be used. We have based our algorithm on classic noise reduction techniques from signal processing.

Additionally, we find that

Finally, 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 and analysis of the ChIP-Seq data in the context of other Gaggle-enabled resources.

3 METHODS

Prior to running Pique, 40-bp Illumina (Solexa) reads should be quality filtered, quality trimmed, and aligned to a reference genome using the user's preferred short-read sequence aligner. Pique requires a BAM file as input [1]. We suggest using all contigs of the reference genome as the mapping target, but the user may prefer to proceed with one contig at a time if desired.

- 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.

Fig. 1. HOW ABOUT A FIGURE OF A PEAK with rectangular envelope and subsequent shape envelope

- Enrichment that are thought to represent binding events produce a characteristic shape envelope, which we model using a sum over set of Gaussians.

Fig. 2. GAGGLE GENOME BROWSER

4 DISCUSSION

If a putative peak passes all of the tests above, Pique concludes 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. YES - WHAT/HOW IS THE CURRENT BG SELECTED?

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. (SOME OF THIS COULD PERHAPS GO ABOVE?) 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. We note that Pique should also work well with eukaryotic datasets provided they are gathered with greater coverage than has been previously reported.

NEED SOME SUMMARY SHOWING THAT THE PIQUE WORKS

NEED SOME RATIONALE FOR SELECTION OF FILTERS

ACKNOWLEDGEMENT

Funding: This project was funded by UC Davis startup funds to MTF, NSF graduate fellowship AWARD NUMBER to EGW and DARPA AWARD NUMBER TO RN.

REFERENCES

[1]Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**(16), 2078–2079.