In a fit of pique : Analyzing microbial ChIP-Seq data with pique

Russell Y. Neches[1,3,*], Elizabeth G. Wilbanks[1,3], Phillip M. Seitzer[2,3], and Marc 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**
[*]**to whom correspondence should be addressed**

ABSTRACT

While numerous effective peak finders have been developed for eukaryotic systems, we have found that the approaches used can be error prone when run on high coverage bacterial and archaeal ChIP-Seq datasets. We have developed Pique, an easy to use ChIP-Seq peak finding application for bacterial and archaeal ChIP-Seq experiments. The software is cross-platform and Open Source, and based on only freely licensed dependencies. Output is provided in standardized file formats, and may be easily imported by the Gaggle Genome Browser Bare et al. (2010) for manual curation and data exploration, or into statistical and graphics software such as R R Core Team (2013) for further analysis. The software is available under the BSD-3 license, and tutorial and test data are included with the documentation.
`http://github.com/ryneches/pique`

Keywords: ChIP-seq, genomics, bacteria, archaea, bioinformatics

INTRODUCTION

Next generation sequencing coupled with chromatin immunoprecipitation (ChIP-Seq) is a powerful technique for studying protein-DNA interactions. The growing popularity of ChIP-Seq has spurred the development of over thirty peak picking algorithms (an extensive survey of these packages was conducted by Wilbanks *et al.* Wilbanks and Facciotti (2010)). The relative performance of representative peak detection algorithms has been recently reviewed by several authors Pepke et al. (2009); Laajala et al. (2009); Szalkowski and Schmid (2010); Feng et al. (2011); Rye et al. (2011).

While ChIP-Seq has been predominantly used to interrogate protein-DNA interactions in eukaryotic systems, it is an especially powerful tool for studying microbes. The small genomes and rapid growth rates, as well as the extensive repertoire of experimental genetic tools available for microbial systems permit ChIP-Seq to provide a particularly clear picture of the state of a cell's transcriptional regulatory machinery.

However, only one ChIP-Seq analysis tool, CSDeconv Lun et al. (2009), has been explicitly developed for microbial data. This MATLAB package successfully finds peaks in microbial ChIP-Seq data, but its application is limited by its dependency on proprietary software, lack of support for manual curation, and difficult operation (users must modify application source code to load their own data). Wilbanks and Facciotti (2010) 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. Pique is distributed under the BSD license, and all dependencies are freely licensed.

1 APPROACH

Because bacterial and archaeal genome sizes are typically two or three orders of magnitude smaller than eukaryotic genomes, ChIP-Seq in bacteria and archaea may cost-effectively yield coverage several orders of magnitude larger than in eukaryotic systems. This results in continuous coverage rather than the sparse coverage typically present in eukaryotic ChIP-Seq data. This feature of microbial ChIP-Seq experiments

permits simpler, faster algorithms to be used. In this manuscript, we present an approach based on classic noise reduction techniques from signal processing.

Pique is designed for use in systems that have genomic complexities such as insertion sequence (IS) elements, gene dosage polymorphisms and accessory genomes that cause coverage variations unrelated to ChIP, or in cases where the organism under study is not identical to the reference genome. The enrichment "pedestals" and "holes" that result from these effects may be be problematic for detecting peaks and calculating enrichment levels. If the user provides a map of these features, the software will automatically perform a segmented analysis to avoid these problems.

The wide variety of microbial systems, target proteins, protocols, and experimental conditions calls for tailored statistical approaches to ChIP-Seq. Rather than attempting to anticipate each of these (and their combinations) with a very large number of statistical and heuristic parameters, we have chosen to focus on the aspects of the analysis that are common to all ChIP-Seq experiments; finding putative peaks, estimating binding coordinates and binding affinities. The determination of statistical significance is typically straightforward for any particular experiment, but is quite difficult to robustly generalize.

Pique allows users to create high-quality peak lists in two ways. First, for each peak reported we report metrics that can be used to ascertain which peaks are statistically significant (usually, this involves little more than sorting the table and choosing a cutoff). Second, we provide integrated support for curation using the Gaggle Genome Browser. Bare et al. (2010) This permits interactive curation of the peak list and analysis of the ChIP-Seq data in the context of other Gaggle-enabled resources. Bare et al. (2007); Shannon et al. (2006) Interactive curation of a microbial ChIP-Seq data set can typically be completed in a few minutes.

2 METHODS

Pique requires BAM files as inputLi et al. (2009). Therefore, prior to using Pique, reads should first be filtered, trimmed, and aligned to a reference genome (ideally, all contigs of the reference genome should be used as the mapping target).

By default, Pique treats each contig as a single analysis region, but the user may designate regions within a contig for separate analysis. This may be useful when coverage levels are systematically altered over large regions. Pique supports three features types: analysis regions, masking regions, and normalization regions. Analysis regions are segments of sequence data that are to be processed together. By default, the software treats each contig in the BAM file as a single analysis region, but the user may choose to split these into smaller segments to compensate for large duplicated features. Masking regions are simply removed from their respective analysis regions, and are useful for removing coverage variation due to repetitive DNA. Normalization regions selected within an analysis region are used to compensate for total coverage discrepancies between the background and ChIP alignments.

The user launches the analysis by providing alignment an file for the ChIP data, an alignment file for the control data, and a coverage feature map. The primary analysis proceeds as follows :

- The alignment files are digested into numeric coverage tracks, and the analysis regions are initialized in memory. Masking regions are applied.
- High-k noise is removed using a Wiener-Kolmogorov filter. Wiener (1942) The filter delay α is chosen to approximate to the expected footprint size of the targeted protein. The choice of filter implies the existence of two inputs; a "true" signal, and a noise source. Both are assumed to be stationary stochastic processes combined additively.
- A Blackman window of a diameter equal to the read length is convolved with the filtered coverage track to remove features smaller than one read. This reduces the effect of fragmentation position bias, and may be especially useful when transposon-based library construction is used.
- The noise threshold in the ChIP coverage track is measured by comparing the coverage distribution in the ChIP track to the control track within user-annotated non-peak regions. Features that exceed the noise threshold are identified.
- Because read orientations are constrained by the fragment size, binding events cause an offset enrichment between the forward and reverse strands. Pique exploits this by requiring that the stop

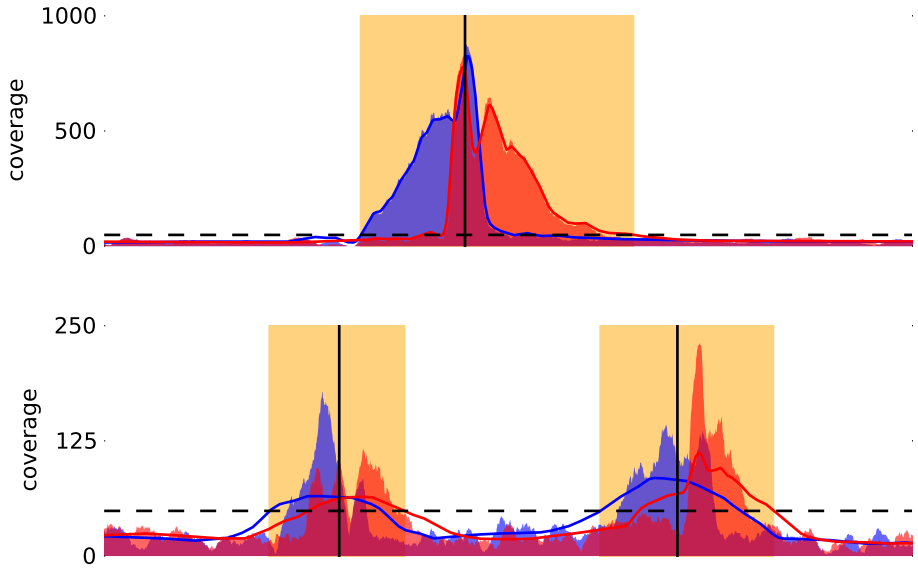

Figure 1. Peaks found in the included sample dataset, derived from ChIP-seq of tfbD in *Halobacterium salinarum* sp. NRC1. Blue and red shading indicate coverage of reads aligned from the ChIP-derived data to the forward and reverse strands, respectively. Blue and red lines represent the filtered coverage levels for the forward and reverse strands, respectively. The dashed line is the detected noise threshold for the region. Detected peaks are indicated in orange boxes.

coordinate of the forward strand enrichment envelope must fall between the coordinates of the reverse strand enrichment envelope, and that the start coordinate of the reverse strand enrichment envelope fall between the coordinates of the forward strand enrichment envelope. (We call this the overlap criterion.)

For each putative peak, Pique calculates the enrichment ratio of the ChIP alignment to the control alignment, the binding coordinate, and the enrichment normalization factor for that analysis region.

3 RESULTS

Performance of Pique was benchmarked against CSDeconv using a ChIP-Seq experiment targeting TfbD (transcription initiation factor IIB 4) DNA-binding events in late exponential phase of *Halobacterium salinarum* sp. NRC1 described in Wilbanks *et al.* Wilbanks et al. (2012). Peak lists for both Pique and CSdeconv were compared with experimentally verified transcription start sites collected by Koide *et al* Koide et al. (2009) and scanned for the presence of putative binding motifs using MASTBailey and Gribskov (1998). An *ab initio* search for putative binding motifs in each peak list was conducted, and the recovered motifs compared using STAMP Mahony et al. (2007).

3.1 TfbD in *Halobacterium salinarum* sp. NRC1

The putative TfbD binding motif for this organism is analogous to *Pyrococcus furiosus* a similar archaeon's TFB protein in photocrosslinking study Renfrow et al. (2004), and is thought to consist of a TFB recognition element (BRE), a TATA box, a proximal promoter element (PPE) and a transcription start site (TSS). This putative promoter motif was scanned against each peak region using MAST Bailey and Gribskov (1998) with an *e*-value cutoff of 100 and the default motif *p*-value cutoff of 10^{-4}. Peaks found by each software package were also evaluated for presence of experimentally determined transcript start sites (TSS) Koide et al. (2009) (Table 1).

Enrichment ratios (the ratio of the integrated coverage to the background) for each peak were calculated, and peaks from each software package for each category were then placed in rank order (Fig.

Found by	Peaks	TSS	Motif	TSS & Motif
Pique or CSDeconv	610	332	197	69
Pique	417	225	128	50
CSDeconv	449	213	138	50
Pique & CSDeconv	257	106	69	31
Pique only	160	119	59	19
CSDeconv only	192	107	69	19

Table 1. Putative binding motifs and transcript start sites detected in peaks recovered by Pique and CSDeconv.

Peak list	Motif *e*-value
CSDeconv all	3.3×10^{-144}
CSDeconv only	5.2×10^{-35}
Pique & CSDeconv	6.9×10^{-77}
Pique all	4.6×10^{-130}
Pique only	4.9×10^{-82}

Table 2. Statistical support for motifs computed from lists of peaks found by Pique and CSDeconv.

2). It was found that Pique recovered more peaks than CSDeconv, and more peaks with experimentally verified transcript start sites. It was found that Pique and CSDeconv recovered peaks containing the full putative binding motif (BRE-TATA-PPE-TSS) at the same rate.

In order to ascertain the quality of the predicted peak regions uniquely predicted by each software package, an *ab initio* search for the putative binding motif was conducted for each list of peaks.

Sequence entries were created by extracting 100-bp stretches of sequence centered at each respective called peak center. Each sequence data set was subjected to an iterative MC MAST/MEME MotifCatcher search [3] with 100 seeds. This search was conducted for five groups of sequences corresponding to all peak regions found by Pique, regions found exclusively by Pique, all peak regions found by CSDeconv, regions found by CSDeconv exclusively, and peak regions found by both software packages. Motifs could be discovered on the forward and reverse strand, and could be anywhere from 20 to 40 nucleotides in length. A putative binding motif was discovered in all five sets of peaks.

The pairwise distances between motifs was computed by the average log likelihood ratio (ALLR), and clustered using UPGMA (program defaults). All motifs that were significantly similar to the putative canonical TFB motif were clustered, and for cases where more than two significant motif matches were discovered in a given data set, motifs were clustered and familial profiles were computed. Membership clustering thresholds were tried from 0 to 1.0 in increments of 0.10, and the familial profile motif with the lowest *e*-value was taken from each dataset as its "representative" motif profile. For cases where two or fewer significant motif matches were determined, the motif match with the lower *e*-value was taken as the representative motif profile of its respective data set. Representative motifs were compared using STAMP Mahony et al. (2007) with the ALLR motif comparison measure and UPGMA clustering (Fig. 3).

It was found that the motifs computed from the list of peak regions recovered by Pique and CSDeconv clustered closely with the canonical binding motif, as did motifs computed from the list of peak regions in common to both software packages. However, motifs computed from the list of peak regions recovered *exclusively* by Pique clustered with the canonical motif, but the motifs computed from the list of peak regions recovered *exclusively* by CSDeconv did not (Fig. 3). Furthermore, the *e*-value of the best motif computed from the CSDeconv-unique peak regions is the highest among the five, suggesting that a higher proportion of false positives may be among this list.

4 DISCUSSION

Pique does not attempt to filter peaks that are statistically insignificant. We have found that this part of the analysis is usually specific to the data and to the experiment, and can be highly idiosyncratic. Pique is designed to achieve a low false-negative rate. This allows Pique to work without modification on many

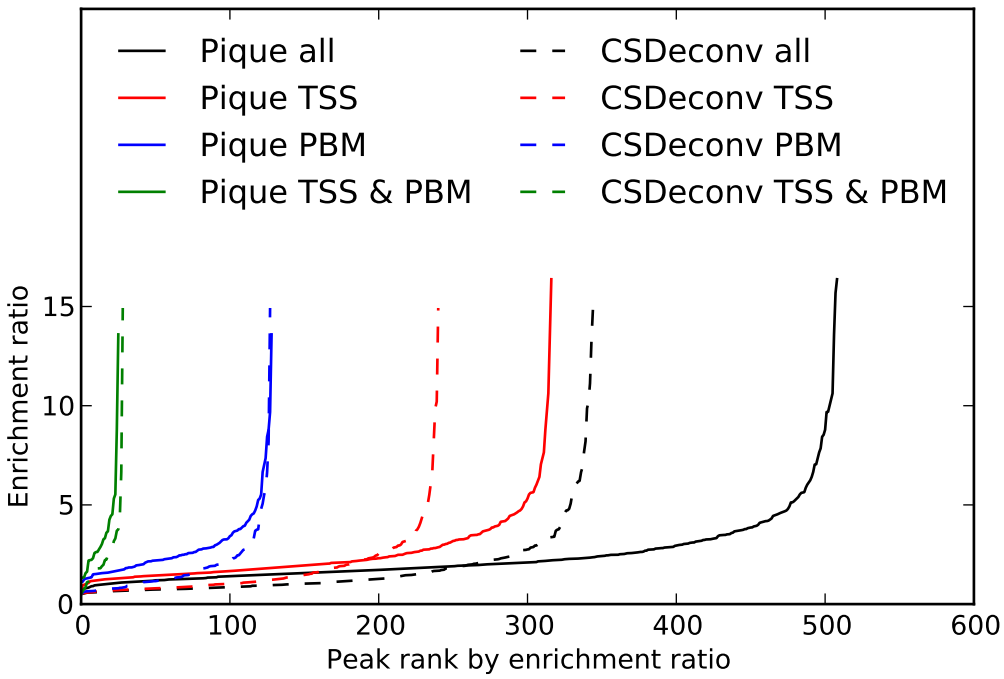

Figure 2. Performance of Pique and CSDeconv on a ChIP-seq experiment (*Halobacterium salinarum* sp. NRC1 transcription initiation factor TFB tfbD in late stationary phase) was studied by comparing lists of peaks. Peak lists are shown here ranked by enrichment ratio. Pique recovers more peaks than CSDeconv, and more of these peak regions contain an experimentally determined transcription start site (TSS). Both software packages recovered the same number of peaks containing the predicted binding motif (PBM). Both software packages recovered the same number of peaks in which a transcript start site was found less than 50bp downstream from a binding motif. TSS coordinates were experimentally verified (Koide *et al.*) Koide et al. (2009) and motifs were predicted using MotifCatcher. Seitzer et al. (2012)

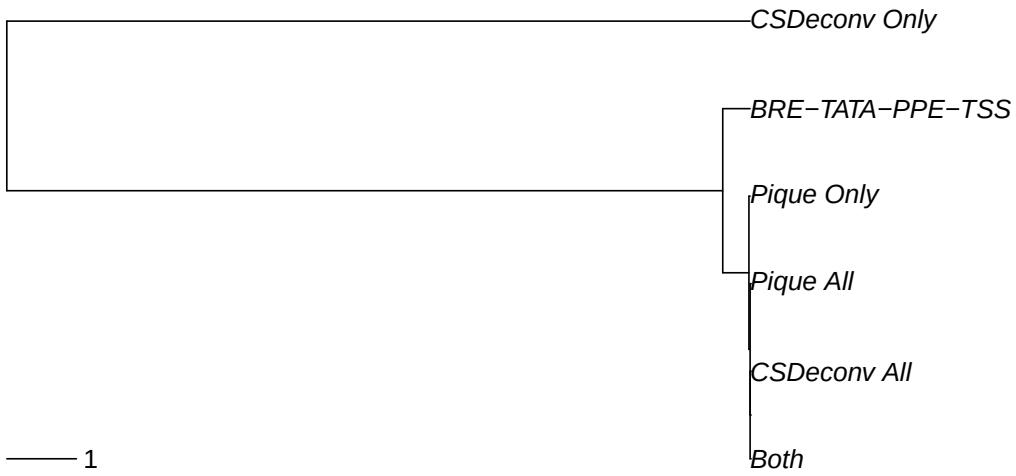

Figure 3. Motifs recovered by Pique and CSDeconv are qualitatively different.

different kinds of experiments, but at the cost of some post-filtering. Pique provides the user with output that can be used to support a variety of such statistical tests.

Some recommended filtering might include eliminating peaks that are significantly narrower than the size range of the sequencing library, peaks with a normalized enrichment ratio below unity, or peaks that have predicted binding sites that are very skewed from the center of the enriched region. Depending on how many peaks are recovered, the user may wish to try one or all of these, perhaps with clustering. However, if a "perfect" peak list is required, we have found that heuristic filtering is inadequate regardless of the software used. To facilitate manual curation, Pique outputs a track file of the coverage, a quantitative positional data of the estimated binding sites, and a bookmark file annotating the peaks. These files are simple to process by a variety of tools, and can be loaded directly into the Gaggle Genome Browser. False positives are easy to recognize visually, and can be easily deleted.

5 CONCLUSION

We conclude that Pique provides a rapid, open source platform for the sensitive detection of transcription factor binding sites in bacterial and archaeal ChIP-seq experiments. We leverage standard signal processing algorithms to rapidly identify binding sites. Downstream analysis is supported via integration with statistical and graphics software such as R, and curation via integration with the user-friendly Gaggle Genome Browser and the suite of Gaggle tools.

We note that Pique should also work well with eukaryotic datasets provided they are gathered with greater coverage than has been previously reported.

ACKNOWLEDGMENTS

This project was funded by UC Davis startup funds to M.T.F., NSF Graduate Research Fellowship awarded to E.G.W. and DARPA award number HR0011-05-1-0057 to R.Y.N.

REFERENCES

Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54.

Bare, J. C., Koide, T., Reiss, D. J., Tenenbaum, D., and Baliga, N. S. (2010). Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics*, 11(1):382.

Bare, J. C., Shannon, P., Schmid, A., and Baliga, N. (2007). The Firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications. *BMC Bioinformatics*, 8(1):456+.

Feng, X., Grossman, R., and Stein, L. (2011). PeakRanger: A cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, 12(1):139+.

Koide, T., Reiss, D. J., Bare, J. C., Pang, W. L., Facciotti, M. T., Schmid, A. K., Pan, M., Marzolf, B., Van, P. T., Lo, F.-Y., Pratap, A., Deutsch, E. W., Peterson, A., Martin, D., and Baliga, N. S. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. *Molecular Systems Biology*, 5(1).

Laajala, T., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., and Elo, L. (2009). A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, 10(1):618+.

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.

Lun, D., Sherrid, A., Weiner, B., Sherman, D., and Galagan, J. (2009). A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biology*, 10(12):R142+.

Mahony, S., Auron, P. E., and Benos, P. V. (2007). DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS computational biology*, 3(3):e61+.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11s):S22–S32.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renfrow, M. B., Naryshkin, N., Lewis, L. M., Chen, H.-T. T., Ebright, R. H., and Scott, R. A. (2004). Transcription factor B contacts promoter DNA near the transcription start site of the archaeal transcription initiation complex. *The Journal of biological chemistry*, 279(4):2825–2831.

Rye, M. B., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Research*, 39(4):e25.

Seitzer, P., Wilbanks, E. G., Larsen, D. J., and Facciotti, M. T. (2012). A monte carlo-based framework enhances the discovery and interpretation of regulatory sequence motifs. *BMC Bioinformatics*, 13(1):317.

Shannon, P., Reiss, D., Bonneau, R., and Baliga, N. (2006). The Gaggle: An open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, 7(1):176+.

Szalkowski, A. M. and Schmid, C. D. (2010). Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Briefings in Bioinformatics*.

Wiener, N. (1942). The interpolation, extrapolation and smoothing of stationary time series. Technical report, Office of Scientific Research and Development.

Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS ONE*, 5(7):e11471+.

Wilbanks, E. G., Larsen, D. J., Neches, R. Y., Yao, A. I., Wu, C.-Y., Kjolby, R. A. S., and Facciotti, M. T. (2012). A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. *Nucleic Acids Research*, pages gks063+.