BayesPeak—an R package for analysing ChIP-seq data

Jonathan Cairns1,*, Christiana Spyrou2, Rory Stark3, Mike L. Smith1, Andy G. Lynch1 and Simon Tavare1

1Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, 2MRC Clinical Sciences Centre, Imperial College, Faculty of Medicine, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN and 3Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Identification of genomic regions of interest in ChIP-seq data, commonly referred to as peak-calling, aims to find the locations of transcription factor binding sites, modified histones or nucleosomes. The BayesPeak algorithm was developed to model the data structure using Bayesian statistical techniques and was shown to be a reliable method, but did not have a full-genome implementation.

Results: In this note we present BayesPeak, an R package for genome-wide peak-calling that provides a flexible implementation of the BayesPeak algorithm and is compatible with downstream BioConductor packages. The BayesPeak package introduces a new method for summarizing posterior probability output, along with methods for handling overfitting and support for parallel processing.

Availability: Available as part of BioConductor version 2.6. URL: http://bioconductor.org/packages/release/bioc/html/BayesPeak.html
Contact: jonathan.cairns@cancer.org.uk

Received on June 24, 2010; revised on December 7, 2010; accepted on December 9, 2010

1 INTRODUCTION

Chromatin Immunoprecipitation (ChIP) experiments produce short DNA fragments, preferentially selected to identify the locations of protein binding sites, histone modifications or nucleosome positions. In the ChIP-seq protocol, as described in Robertson et al. (2007), the 5′-end of one strand of each fragment is sequenced, obtaining a ‘read’, and then aligned to a reference genome. These aligned reads form ‘peaks’—localized regions of high read density—along the genome. Determining the locations and magnitudes of these peaks is an active area of research, and a number of tools exist for the so-called ‘peak-calling’, using a variety of methodologies.

The algorithm described in Spyrou et al. (2009) takes a Bayesian approach to modelling aligned reads from ChIP-seq data. Many peak-callers model read counts with the Poisson distribution, and each job region is expanded by 2 Kb in each direction (to allow peaks falling on the boundary between two jobs to be called).

Within a job, the region is divided into small bins (each of length 100 bases, by default), and reads are aggregated by the bin in which they start and the strand on which they lie. A state-of-the-art HMM is fitted to these aggregate counts. The HMM’s hidden states correspond to enrichment or unenrichment for sites of interest. A hidden state produces two negative binomial emissions, each corresponding to a bin count (one on each strand), with enriched states tending to emit larger values. The HMM is fitted through MCMC techniques that sample from the posterior distributions of the parameters. The analysis is performed a second time on the same job region, but with all bins offset by half their width (the ‘offset’ analysis) as illustrated in Figure 1. Further details can be found in Spyrou et al. (2009).

The output of each job is the posterior probability (PP) of each site being enriched. The data are summarized to form the final peaks as follows. All bins with PP values greater than a user-specified threshold (by default, 0.5)
Then, for each output. For example, direct analysis of the peaks in any downstream package compatible with \texttt{boa} permitting convergence tests such as the Geweke diagnostic in version 1.1.3, MCMC samples of several key parameters are also present, (excluding half of the draws as burn-in). As of \texttt{boa} version 0.13-5, a lower bound for the probability of enrichment in at least one of the \textit{n} bins is \( F(i) = 1 - \prod_{i \in S_n} q_i \).

The ‘best’ (highest) lower bound for the probability of peak enrichment is therefore the maximum of this quantity, \( \sup_{i \in S_n} F(i) = 1 - Q(n), \) where \( Q(n) = \inf_{i \in S_n} \prod_{i \in S_n} q_i \).

We can find \( Q(n) \) by dynamic programming since, by conditioning on whether \( i \in I \), we have \( Q(i) = \min(Q(i-1), q_i Q(i-2)) \).

The advantage of using this method over taking the maximum PP value is that it can give an appropriate score to sustained regions of only moderately large PP values, which will be undervalued when taking the maximum.

We tested \texttt{BayesPeak} on the NRSF/REST ChIP-seq dataset from (Johnson et al., 2007), in which a small subset of regions have been experimentally validated, and we compared the findings against other common peak callers.

### 3 RESULTS

We present the peak-caller comparison results in the Supplementary Material. \texttt{BayesPeak} demonstrated a competitive sensitivity and specificity on the genome-wide scale and showed a substantial overlap with other peak-callers. The over-fitting correction greatly improved the enrichment for true binding sites in \texttt{BayesPeak}’s data, as did subsequent filtering by PP value.

In its raw output, \texttt{BayesPeak} returns PP values for each bin and, for each job, the posterior mean of each estimated parameter (excluding half of the draws as burn-in). As of \texttt{BayesPeak} version 1.1.3, MCMC samples of several key parameters are also present, permitting convergence tests such as the Geweke diagnostic in the \texttt{boa} (Smith, 2007) or \texttt{coda} (Plummer et al., 2010) packages.

Since the summarized output is in RangedData format, this allows direct analysis of the peaks in any downstream package compatible with \texttt{IRanges} (Pages et al., 2010), including those in BioConductor. For example, \texttt{ChIPpeakAnno} (Zhu et al., 2010) can annotate the output.

We have observed some phenomena that occur with lower quality data. For example, over-fitting can occur as follows: the model assumes that for each job there are both enriched and unenriched states. As such, when there are no peaks in a job or when the peaks are extremely weak, these two states are used to explain the natural variance present in the unenriched background. We identify over-fit jobs from their low \( \lambda_1 \) values (where \( \lambda_1 \) is the expected number of counts in an enriched bin), and from their PP values being spread out over \([0, 1]\) rather than tending to be 0 or 1. \texttt{BayesPeak} supports the identification and removal of jobs that have encountered this effect (Supplementary Table S2 and Fig. S1).

### 4 DISCUSSION

\texttt{BayesPeak} provides a Bayesian analysis, with advantages including allowance for overdispersion in read counts and a competitive genome-wide specificity and sensitivity. By anticipating peak structure, \texttt{BayesPeak} does not call peaks based on sheer numbers of reads without appropriate read formation.

Careful selection of job regions may improve the analysis. For example, we can use prior knowledge to partition jobs in a manner that avoids analysing the centromeres and telomeres, which usually contain no reads. This will prevent unnecessary computation, and may also improve results in the surrounding regions.

There is scope for adapting the \texttt{BayesPeak} approach to other forms of peak-calling. For example, some histone mark data consist of regions of enrichment containing many peaks and, in BrDU-seq data, peaks are much broader than those in transcription factor data.

### ACKNOWLEDGEMENT

We are grateful to Dr Jason Carroll’s group and Dr Duncan Odom’s group for access to test datasets, and to Dr Mark Dunning for beta testing.

\textbf{Funding:} J.C. is funded by a Medical Research Council grant. We acknowledge the support of the University of Cambridge, Cancer Research UK, and Hutchinson-Whampoa.

\textbf{Conflict of Interest:} none declared.

### REFERENCES

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 5, R80.

Johnson,D.S. et al. (2007) Genome-wide mapping of in vivo protein-DNA interactions. Nature, 447, 736--740.

Pages,J. et al. (2010) IRanges: Infrastructure for Manipulating Intervals on Sequences. R package version 1.6.6.

Plummer,M. et al. (2010) Coda: Output Analysis and Diagnostics for MCMC. R package version 0.13-5.

Robertson,G. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Rev Genet., 8, 209.

Uhrina,S. (2009) Multicore - parallel processing in R on machines with multiple cores or CPUs, version 0.1-4.

Zhu,J.X. et al. (2010) ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics, 11, 237.