Yeast
Yeast 2014; **31**: 375–391.
Published online 28 July 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/yea.3031

Research Article

bPeaks: a bioinformatics tool to detect transcription factor binding sites from ChIPseq data in yeasts and other organisms with small genomes

Jawad Merhej[1,2], Amandine Frigo[3], Stéphane Le Crom[3,4,5], Jean-Michel Camadro[6], Frédéric Devaux[1,2] and Gaëlle Lelandais[6]*

[1]*Sorbonne Universités, UPMC University of Paris 06, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France*
[2]*CNRS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France*
[3]*Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), Inserm U1024 and CNRS UMR 8197, Paris, France*
[4]*Sorbonne Universités, UPMC University of Paris 06, UMR 7622, Laboratoire de Biologie du Développement, Paris, France*
[5]*CNRS, UMR 7622, Laboratoire de Biologie du Développement, Paris, France*
[6]*Institut Jacques Monod, CNRS UMR 7592, University of Paris Diderot, Paris, France*

**Correspondence to:*
G. Lelandais, Institut Jacques Monod, CNRS UMR 7592, University of Paris Diderot, Paris, France.
E-mail: gaelle.lelandais@univ-paris-diderot.fr

Abstract

Peak calling is a critical step in ChIPseq data analysis. Choosing the correct algorithm as well as optimized parameters for a specific biological system is an essential task. In this article, we present an original peak-calling method (bPeaks) specifically designed to detect transcription factor (TF) binding sites in small eukaryotic genomes, such as in yeasts. As TF interactions with DNA are strong and generate high binding signals, bPeaks uses simple parameters to compare the sequences (reads) obtained from the immunoprecipitation (IP) with those from the control DNA (input). Because yeasts have small genomes (<20 Mb), our program has the advantage of using ChIPseq information at the single nucleotide level and can explore, in a reasonable computational time, results obtained with different sets of parameter values. Graphical outputs and text files are provided to rapidly assess the relevance of the detected peaks. Taking advantage of the simple promoter structure in yeasts, additional functions were implemented in bPeaks to automatically assign the peaks to promoter regions and retrieve peak coordinates on the DNA sequence for further predictions of regulatory motifs, enriched in the list of peaks. Applications of the bPeaks program to three different ChIPseq datasets from *Saccharomyces cerevisiae*, *Candida albicans* and *Candida glabrata* are presented. Each time, bPeaks allowed us to correctly predict the DNA binding sequence of the studied TF and provided relevant lists of peaks. The bioinformatics tool bPeaks is freely distributed to academic users. Supplementary data, together with detailed tutorials, are available online: http://bpeaks.gene-networks.net. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: **ChIPseq; bioinformatics; peak-calling; yeasts; transcription factors; regulatory motifs**

Received: 11 March 2014
Accepted: 3 July 2014

Introduction

Transcriptional regulations are mediated by DNA-binding of transcription factors (TFs) that recognize specific DNA elements which, in yeasts, are located upstream of their regulated genes (target genes). Identification of transcription factor binding sites is an important preliminary to: (a) discovering DNA-regulatory motifs, which are recognized by a particular transcription factor; and (b) identifying potential target genes of this factor, which is the first step to building genome-wide transcriptional regulatory networks (Harbison *et al.*, 2004). Chromatin immunoprecipitation (ChIP), followed by high-throughput sequencing (ChIPseq), is a powerful technique for the genome-wide detection of protein–DNA-binding sites (Johnson *et al.*, 2007). The ChIP experimental procedure consists in treating living

cells with a crosslinking agent that attaches proteins to their DNA substrates. After DNA extraction and fragmentation, DNA sequences associated with a particular protein of interest are isolated using a specific antibody against the protein (for review of ChIP-based methods, see Kim and Ren, 2006). In ChIPseq experiments, the DNA fragments of interest are directly sequenced using high-throughput sequencing technologies. Compared to ChIP on chip – the microarray based technology – ChIPseq offers higher resolution, greater sequence coverage and increased signal:noise ratio (Park, 2009). Applied to the detection of transcription factor binding to genomic DNA, ChIPseq allows the determination of two to four times more binding sites than previous methods (Robertson *et al.*, 2007).

Still it is important to consider several technical aspects to obtain high-quality ChIPseq data (Kidder *et al.*, 2011). This includes the quality of antibodies, the mandatory use of controls, the quality of the DNA library construction, the sequencing procedure and, finally, the bioinformatics and statistical analyses applied to the experimental results. It is critical to use relevant computational methods for processing sequencing data and thus ensuring the inference of biologically meaningful information (Diaz *et al.*, 2012). The classic workflow for ChIPseq analysis can be divided into three main steps: (a) quality controls and filtering of low quality sequences; (b) mapping of the remaining sequences (or reads) to the reference genome; and (c) peak finding (or 'peak calling') to detect protein–DNA interactions over the whole genome (for more details, see the reviews of Kim and Ren, 2006; Pepke *et al.*, 2009; Kidder *et al.*, 2011; Bailey *et al.*, 2013). Peak calling is clearly the most challenging part of ChIPseq data analysis. It consists in using a computational procedure to identify genomic regions with a significant enrichment of reads in ChIP sample relative to background noise. To correctly estimate enrichments, it is necessary to use a control sample in which, for instance, genomic DNA was sequenced without antibody enrichment (total INPUT) (Liang and Keles, 2012). In control samples, distributions of reads are far from being uniform (Rozowsky *et al.*, 2009) and hence regions with high read counts do not necessarily represent DNA-binding sites for proteins. The main challenge of peak calling is to distinguish real binding events from intrinsic variability in the sequencing depth.

In the last few years, a burst of peak-calling methods has been developed. More than 30 analytical programs are currently available (for method reviews and comparisons, see e.g. Fejes *et al.*, 2008; Kharchenko *et al.*, 2008; Rozowsky *et al.*, 2009; Xu *et al.*, 2010; Cheng *et al.*, 2011; Boeva *et al.*, 2012; Wang *et al.*, 2013; Wilbanks and Facciotti, 2010; Malone *et al.*, 2011; Bailey *et al.*, 2013). Importantly, the method and its parameter values should be chosen to correctly fit the characteristics of the genomic regions to be identified, which depend on the type of immunoprecipitated protein and of the protocol used (Pepke *et al.*, 2009). Proteins such as RNA polymerases, general transcription factors or histones, generally bind to DNA in broad regions, yielding to numerous and relatively large peaks with low read density (Wang *et al.*, 2013). In the case of specific transcription factors, the ChIPseq signals are expected to be sharper and specific to short DNA sequences located upstream of the transcriptional start site of a relatively limited number of genes (Robertson *et al.*, 2007). Also, handling ChIPseq data for species with large genome sizes (e.g. human, mouse) addresses very different challenges compared to species with small genome sizes, such as yeasts (<20 Mb). Sequencing depth is the major concern for people who perform deep sequencing analyses in organisms with large genomes (Landt *et al.*, 2012). This explains constant improvements in sequencers to obtain more and more reads in a single run. Dealing with this immense quantity of sequence data (hundreds of millions of reads) requires computational skills and suitable hardware resources (Nagasaki *et al.*, 2013). For species with small genome sizes, the situation is very different. Bioinformatics simulations based on the yeast genome estimated that only 260 000 unique mapped reads are enough to saturate the binding sites of a particular TF (with a five-fold enrichment) (Lefrancois *et al.*, 2009), whereas 12 million mapped reads are typically used with the human genome (Rozowsky *et al.*, 2009).

Working with a reasonable number of reads in genomes with an average complexity represents an important computational advantage. Peak calling can be performed in a few min with a desktop workstation, allowing for iterative procedures and more systematic analyses of the data. In that respect, we developed a simple and robust bioinformatics tool called bPeaks ('basic Peaks') for the detection of TF binding sites from ChIPseq data in small eukaryotic genomes. The

DOI: 10.1002/yea

program bPeaks performs a high-resolution analysis (at the nucleotide scale) of ChIPseq results. It uses a sliding window that scans every position of a genome and compares the read number obtained from the immunoprecipitation (IP) sample with those obtained from a control sample. To define a genomic region as a 'peak', four criteria have to be satisfied: (a) a high number of reads in the IP sample (criterion 1 or C_1); (b) a low number of reads in the control sample (C_2); (c) a high value of log fold change (or logFC) between numbers of reads in IP and control samples (C_3); and (d) a good sequencing coverage in both IP and control samples (C_4). Peaks are therefore defined as the genomic regions reaching four threshold values (T_1, T_2, T_3 and T_4) corresponding to each of the four criteria mentioned above. After the peak calling, bPeaks generates output files and graphics to help users to choose combinations of threshold values and to define lists of relevant peaks for further analyses. In this paper, we explain the bPeaks program and present several applications to ChIPseq data obtained in yeasts *Saccharomyces cerevisiae*, *Candida albicans* and *Candida glabrata*.

Methods

Performing a peak calling analysis with bPeaks

The program bPeaks compares IP and control signals (Figure 1A), i.e. numbers of mapped reads obtained from the IP and the control samples. It uses overlapping sliding windows to scan the genome (Figure 1A). A combination of four thresholds is applied to identify positive windows, i.e. genomic regions with: (a) a high number of reads in the IP sample (threshold T_1); (b) a low number of reads in the control sample (threshold T_2); (c) a high value of logFC between IP and control (threshold T_3); and (d) a good sequencing coverage (IP and control samples, threshold T_4) (Figure 1B). The program bPeaks is written in R programming language. It is freely available from the CRAN website (http://cran.r-project.org/web/packages/bPeaks). Tutorials are also available online (http://bpeaks.gene-networks.net).

Parameter calculations

Global parameters (G_{IP} and $G_{Control}$) are calculated to quantify the sequencing coverage in IP and control samples, as follows:

$$G_{IP} = \frac{1}{n}\sum_{i=1}^{n} x_{IP_i} \quad (1)$$

and

$$G_{Control} = \frac{1}{n}\sum_{i=1}^{n} x_{Control_i} \quad (2)$$

where x_{IP_i} and $x_{Control_i}$, respectively, denote the number of sequences mapped at position i of the reference genome in IP and control samples, and n the total number of nucleotides in the genome.

Local parameters ($C_{1,w}$, $C_{2,w}$, $C_{3,w}$ and $C_{4,w}$) in the sliding windows are also calculated. For each window (w), we have:

$$C_{1,w} = \frac{1}{n_w}\sum_{i=1}^{n_w} x_{IP_{i,w}} \quad (3)$$

$$C_{2,w} = \frac{1}{n_w}\sum_{i=1}^{n_w} x_{Control_{i,w}} \quad (4)$$

$$C_{3,w} = \frac{1}{n_w}\sum_{i=1}^{n_w} \log\left(\frac{x_{IP_{i,w}}}{x_{Control_{i,w}}}\right) \quad (5)$$

$$C_{4,w} = \frac{1}{n_w}\sum_{i=1}^{n_w} \frac{\log\left(x_{IP_{i,w}}\right) + \log\left(x_{Control_{i,w}}\right)}{2} \quad (6)$$

where n_w denotes the number of nucleotides that belong to window w. These parameters quantify important properties regarding peak detection: local IP signal ($C_{i,w}$), local control signal ($C_{2,w}$), logFC between IP and control signals ($C_{3,w}$) and density of reads ($C_{4,w}$).

Peak detection

Genomic regions are identified combining global and local parameters, with four thresholds denoted T_1, T_2, T_3 and T_4. Positive windows are selected if: $C_{1,w} \geq T_1 \times G_{IP}$, $C_{2,w} \leq T_2 \times G_{Control}$, $C_{3,w} \geq T_3$ and $C_{4,w} \geq Quantile(C_4, T_4)$ (Figure 1B). Typically, T_1 and T_2 values lie between 1 and 6, as they represent multiplicative values associated with the sequencing coverage in IP and control samples; T_3 lies around 2 or 3, as it represents a

DOI: 10.1002/yea

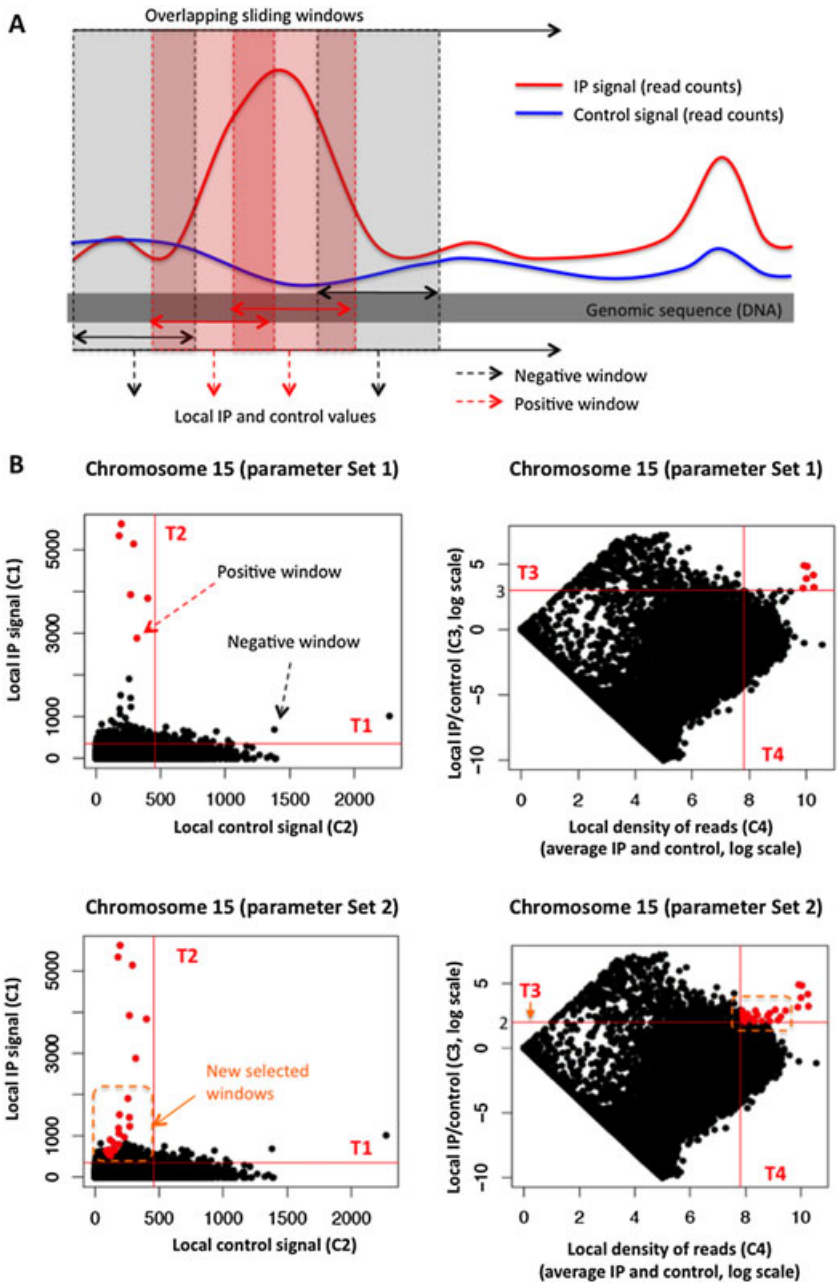

Figure 1. General principle of the bPeaks method. (A) The bPeaks program uses a sliding windows to scan the entire genomic sequence. For each window, the program calculates local IP and control values (see Methods). Positive windows exhibit a high value in IP signal, a low value in control signal, a high value of logFC between IP and control signals and a sequencing coverage high enough to ensure good confidence in IP and control signals. (B) Graphical representations of local parameters calculated with bPeaks on chromosome 15 of yeast *S. cerevisiae*, analysing Pdr1p ChIPseq data (see Results). To apply bPeaks, the user specifies values for the four thresholds (T_1, T_2, T_3 and T_4). Positive windows exhibit IP signal $> T_1$, control signal $< T_2$, logFC $> T_3$ and density of reads $> T_4$. Results associated with parameter sets 1 and 2 (see Table 1) are shown here. Note that difference between the two sets of parameters only relies on T_3 (lower in parameter set 2 than in parameter set 1; see orange arrow). Decreasing *T*3 allows the selection of additional windows (surrounded by orange broken line). This explains why S1 results are necessarily included in S3 results (Figure 3A)

DOI: 10.1002/yea

logFC (base 2); and T_4 lies between 0 and 1, as it is a threshold associated with the cumulative distribution function of $C_{4,w}$ values on each chromosome. Successive positive windows are merged to define larger genomic regions, referred to as 'basic peaks' (bPeaks).

Input and output data files

An overview of bPeaks input and output files is shown Figure 2. Sequencing results should be first converted to data files, with the number of sequences mapped on each nucleotide in the reference genome (Figure 2A). This can be performed with other bioinformatics programs, such as BEDTOOLS (Quinlan and Hall, 2010). After parameter selection and peak detection (Figure 2B), bPeaks generates several output files, PDF files for graphics and BED files for peak locations (Figure 2C). These files can be used for further investigations with, for instance, extraction of DNA sequences of peaks (FASTA file), detection of regulatory motifs or calculation of the proportion of peaks in promoters (Figure 2D).

Assigning detected peaks to genes with bPeaks

Annotations of gene positions for 10 different yeast species are directly available in the bPeaks tool: *Saccharomyces cerevisiae*, *Candida albicans*, *Candida glabrata*, *Debaryomyces hansenii*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Pichia sorbitophila*, *Saccharomyces kluyveri*, *Yarrowia lipolytica* and *Zygosaccharomyces rouxii.*

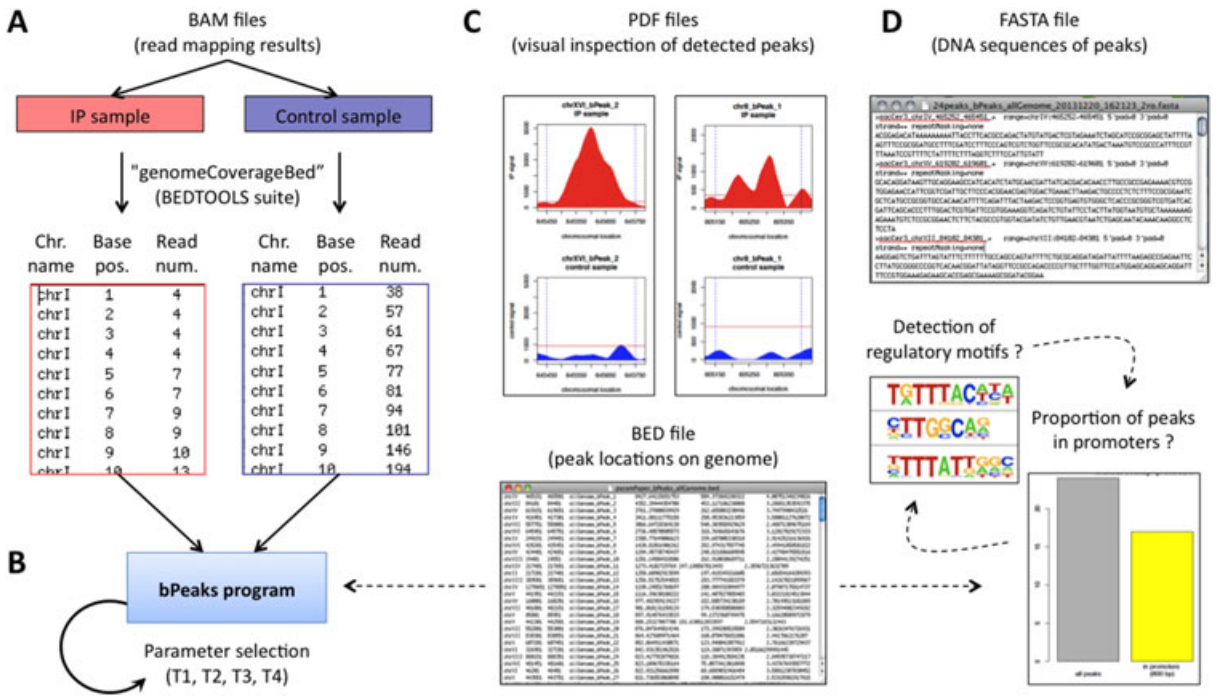

Figure 2. Overview of the bPeaks protocol used to analyse ChIPseq data. (A) Input files for bPeaks program (see Methods). The sequencing results should be converted to data files with the number of sequences mapped on each nucleotide in the reference genome. (B) Use of the bPeaks program requires the selection of values for thresholds *T*1, *T*2, *T*3 and *T*4. A procedure was implemented in the bPeaks program to automatically test several combination of parameters (see Table 1 for an example on the Pdr1p dataset). Supplementary parameter combinations can be chosen to explore the bPeaks parameter space more deeply. (C) Output files provided by the bPeaks program. These files allow the biological relevance of the genomic regions identified with bPeaks to be assessed. The user can either modify bPeaks parameters to increase the peak calling stringency of sensitivity (see Table 1) or use other tools to deeply analyse the identified genomic regions. (D) Illustration of further analyses that can be performed to assess the relevance of the detected peaks. Peak assignment to promoters is a functionality that is available in bPeaks tools for 10 yeast species (see Methods)

DOI: 10.1002/yea

This allows bPeaks to automatically calculate the proportion of peaks in promoter regions of genes (default value is 800 bp before start codon ATG). Annotations were collected from the *Saccharomyces* Genome Database (SGD) (Cherry *et al.*, 2012), the *Candida* Genome Database (Inglis *et al.*, 2012) and the Genolevures database (Sherman *et al.*, 2009). For other organisms, the user can specify the boundaries of any genomic element (gene, promoter, non-coding elements, etc.) and use bPeaks to identify peaks that fall in each category.

Analyses of ChIPseq data in this study

Transcription factor Pdr1p in S. cerevisiae

A *S. cerevisiae* myc-tagged PDR1 strain (Fardeau *et al.*, 2007) was used to perform chromatin immunoprecipitation (ChIP). Overnight culture was harvested at OD 0.6–0.8 and fixed with 1% formaldehyde for 15 min at room temperature with casual agitation. The crosslinking was stopped by adding glycine to a final concentration of 340 mM and incubating for 5 min at room temperature. Cells were disrupted using a FastPrep®-24 instrument (MP Biomedial). Cell extracts were sonicated using a Bioruptor® standard sonication device (Diagenode), leading to DNA fragments of around 300 bp. Cell debris was then eliminated by centrifugation. A portion corresponding to 1% of total soluble fraction was retained for further DNA extraction (control sample or INPUT). The remaining fraction was used to immunoprecipitate the myc-tagged Pdr1p protein with anti-c-myc antibody (Roche Applied Science) bound to Dynabeads® magnetic beads (Invitrogen). After overnight incubation with gentle shaking at 4 °C, the IP complexes were washed and eluted from the beads by heating the samples for 20 min at 65 °C with shaking at 1200 rpm in elution buffer containing 0.5% SDS. The crosslinking of the IP (immunoprecipitated chromatin) and INPUT (whole chromatin) was then reversed by heating the samples at 65 °C overnight. Reversed chromatin was then digested with proteinase K (Roche Applied Science) at 37 °C for 2 h. DNA was extracted using a standard phenol/chloroform extraction protocol, treated with RNAse (Fermentas) to totally eliminate any residual RNA, and purified using QIAquick® PCR Purification Kit (Qiagen). The collected DNA samples were used to construct libraries using the NEXTflex ChIPseq Kit (Illumina), following the supplier's instructions. Sequencing was performed using a HiSeq sequencing instrument (Illumina technology available at the transcriptome platform at the Ecole Normale Supérieure: http://www.transcriptome.ens.fr/sgdb/, Paris, France). After quality controls and filtering of low quality bases, around 30 million sequences (IP sample) and 88 million sequences (control sample) were mapped on the *S. cerevisiae* genome, using the bowtie algorithm (Langmead *et al.*, 2009). Output files (SAM format) were converted into BAM files and indexed using the SAMTOOLS suite (Li *et al.*, 2009). Numbers of sequences mapped on each nucleotide in the reference genome were finally calculated using the 'genomeCoverageBed' tool, available from the BEDTOOLS suite (Quinlan and Hall, 2010) and stored in two data files (one for the IP sample and one for the control sample). These files can be downloaded from the bPeaks website (http://bpeaks.gene-networks.net). Note that all sequencing data (Pdr1p, Sfl1p and CgAp1p) used in this article are 'single read' datasets.

Transcription factor Sfl1p in C. albicans

FASTQ files were collected from the SRA database (http://www.ncbi.nlm.nih.gov/sra) under Accession No. SRP017529. Between 14 and 22 million reads were available for each sample. Read mappings and file conversions for peak calling with the bPeaks program were performed as described for Pdr1p ChIPseq data.

Transcription factor CgAp1p in Candida glabrata

The same ChIP protocol described above for Pdr1p was applied to immunoprecipitate *C. glabrata* CgAP1 myc-tagged protein after 10 min of exposure to 1 mM selenite (stress condition) and in optimal growth conditions (non-stress condition). After library constructions and the sequencing procedure, 22 million reads (IP sample, stress condition), 21 million reads (IP sample, non-stress condition) and 23 million reads (associated INPUT controls) were analysed using the same procedure as described for Pdr1p ChIPseq data.

DOI: 10.1002/yea

Running bPeaks

To apply the bPeaks program (Pdr1p ChIPseq data), the following R code can be used:

```
library(bPeaks)
## get CDS annotations for yeasts
data(yeastCDS)
## read the sequencing result files
pdr1Data = dataReading('IPdata.txt', 'controlData.
txt', yeastSpecies = yeastCDS$Saccharomyces.
cerevisiae)
## bPeaks analyses with different parameter
associations
bPeaksAnalysis(IPdata = pdr1Data$IPdata,
controlData = pdr1Data$controlData,
IPcoeff = c(6, 4, 2), controlCoeff = c(2, 4, 6),
log2FC = c(3, 2), averageQuartile = c(0.9, 0.7) )
```

As an illustration here, values 6, 4 and 2 for 'IPcoeff' (threshold T_1); 2, 4 and 6 for 'controlCoeff' (threshold T_2); 3 and 2 for 'log2FC' (threshold T_3); and 0.9 and 0.7 for 'averageQuartile' (threshold T_4) were combined to perform peak calling with different parameter associations ($3 \times 3 \times 2 \times 2 = 36$ combinations), from the most stringent combination ($T_1 = 6$, $T_2 = 2$, $T_3 = 3$ and $T_4 = 0.9$) to the less stringent one ($T_1 = 2$, $T_2 = 6$, $T_3 = 2$ and $T_4 = 0.7$). The parameter associations and the compositions of the detected lists of peaks are discussed in Results. Note that the user can easily specify additional parameter associations in order to explore the parameter space more deeply.

Retrieving DNA sequences of peaks and searching for regulatory motifs

Starting from the genomic locations of the peaks detected with bPeaks (BED files), DNA sequences of interesting peaks were retrieved using the 'getfasta' function from the BEDTOOLS suite (Quinlan and Hall, 2010). These genomic sequences were used as inputs for the 'peak-motif' tool (http://rsat.ulb.ac.be/) to search for regulatory motifs (Thomas-Chollier *et al.*, 2012).

Technical information

Memory requirements to use bPeaks, in terms of computational resources, are proportional to the size of the genome and the sequencing coverage (number of reads). As an illustration, analysis of Pdr1p ChIPseq data lasts around 5 min (one set of parameters) using the HP Z820 Workstation [Intel Xeon E5-2609 2.4 Ghz CPU and 16 GB DDR3-1600 (8×2 GB) RAM].

Results

A case study of the Pdr1p transcription factor in *Saccharomyces cerevisiae*

To assess the relevance of bPeaks, we performed ChIPseq experiments (IP and control samples) of the transcription factor (TF) Pdr1p, in the model yeast *S. cerevisiae* (see Methods). Pdr1p belongs to the GAL4 family of yeast TFs, characterized by the Zn_2Cys_6 DNA-binding motif (Schjerling and Holmberg, 1996). It plays a central role in the regulation of pleiotropic drug resistance through transcriptional controls of about 30 genes (Kolaczkowska and Goffeau, 1999; DeRisi *et al.*, 2000; Fardeau *et al.*, 2007). Pdr1p was chosen to benchmark the bPeaks program for three main reasons. First, Pdr1p is a promoter-resident regulator, which, in contrast to other stress-responsive TFs, does not need a particular environmental stimulation to bind DNA (Fardeau *et al.*, 2007). This property greatly simplified our ChIPseq analyses, as no particular treatment was required to observe Pdr1p binding to its target genes. Second, the DNA consensus sequence recognized by Pdr1p, called the pleiotropic drug-response element (PDRE; 5′-TCCGCGGA-3′), has been characterized without ambiguity (Mamnun *et al.*, 2002). Third, several groups have studied the genome-wide binding patterns of Pdr1p using ChIP on chip technology (DeRisi *et al.*, 2000; Devaux *et al.*, 2001; Fardeau *et al.*, 2007). The set of genes regulated by Pdr1p has thus been extensively described in the literature (DeRisi *et al.*, 2000; Devaux *et al.*, 2001; Fardeau *et al.*, 2007).

Influence of parameter values on bPeaks results

Parameter choice represents a key step in using every peak-calling program. In bPeaks, the detection of peaks relies on four thresholds, T_1, T_2, T_3 and T_4 (see Methods). To evaluate the influence of parameter values on final results, we applied bPeaks with

DOI: 10.1002/yea

36 different combinations of thresholds associated to different levels of peak calling specificity and sensitivity (see R code, Methods). Table 1 summarizes critical properties regarding the detected peaks. Combinations of thresholds were ordered according to the average IP signals (average C_1), from the highest to the lowest. The first and last rows in Table 1, respectively, show the most stringent parameter associations (highest values for T_1, T_3, T_4 and lowest values for T_2) and the least stringent ones (lowest values for T_1, T_3, T_4 and highest values for T_2). The number of peaks identified by the different sets of parameters varies from 24 for the most stringent to 628 for the least. Several combinations of parameters led to the same lists of peaks, with only small variations in the size of the regions detected, and which slightly changed the average C_1, C_2, C_3 and C_4 values of the lists (Table 1). This allowed us to differentiate the impact of each criterion on the final lists of selected peaks. An association between numbers of detected peaks and the values of T_3 and T_4 was clearly observed, indicating that these two parameters have, in this case, an essential influence on the stringency of the list. In contrast, T_1 and T_2 have influence which is relevant only when T_3 and T_4 are both relatively low (see the two last lists of peaks, 611 and 628 peaks, Table 1).

Figure 3A presents a comparison of the overlaps between lists of peaks associated with four different parameter sets, referred to as S1 (24 peaks), S2 (122 peaks), S3 (121 peaks) and S4 (336 peaks) (Table 1). S1 results were obtained using the most stringent combinations of parameters. They were therefore included in all other lists. In contrast, the S2 and S3 lists, although containing a similar number of peaks (122 and 121, respectively), had a partial overlap of $< 50\%$ (47 peaks; Figure 3A). This corresponds to the opposite changes of the thresholds T_3 and T_4 between the two lists. Peaks selected in S2 have lower logFC values than peaks detected in S3 ($T_3^{S2} < T_3^{S3}$) but a higher density of reads ($T_4^{S2} > T_4^{S3}$). Finally, the results of S1, S2 and S3 are all included in the S4 list, which has the lowest combinations of T_3 and T_4 values (2 and 0.7, respectively). This is coherent with the idea that T_3 and T_4 are important parameters that account for the composition of the lists of outstanding peaks. Finally, it is important to note that the parameter values have very little influence on the sizes of the detected peaks, which were around 180 base pairs for all combinations of parameters (Table 1).

Assessment of the biological significance of bPeaks lists

Once lists of peaks are generated, bPeaks provides output files that can be used to assess their biological significance (see Methods). For instance, the user can analyse the DNA sequences of peaks to predict the DNA consensus motifs recognized by the studied TF or analyse the proportion of peaks located in promoter sequences of genes (Figure 2D). This information can help the user to evaluate the specificity and sensitivity of their peak calling analyses. As an illustration, results obtained for the Pdr1p data are detailed below.

De novo cis-regulatory motif discovery

One of the important outcomes that biologists expect from ChIPseq data is the identification of the DNA binding preferences of a transcription factor. To reach this goal in the case of Pdr1p, the sequences of the peaks found in the S1, S2, S3 and S4 lists were analysed with 'peak-motifs' web tool (Thomas-Chollier *et al.*, 2012) (see Methods). DNA motifs enriched in each list were predicted using the default parameters; the results are presented in Supplementary data S1 (see supporting information) and Figure 4. With the S1 list, all detected motifs agreed with the Pdr1p-binding site, 5′-TCCGCGGA-3′ (PDRE) described in the literature. PDRE was present in eight (33%) of the peaks in the S1 list (Figure 4). Motifs detected respectively in the S2 and S3 (see supporting information) lists also agreed with the PDRE list (see supporting information, Supplementary data S1). Interestingly, the number of peaks with the PDRE motif was, this time, higher (16 peaks for the S2 list and 13 peaks for the S3 list) but the proportion of these peaks was lower (13% for the S2 list and 11% for the S3 list; Figure 4). In the larger S4 list, we observed that the PDRE was found among other unrelated motifs to Pdr1p (see supporting information, Supplementary data S1) and the global proportion of PDRE containing peaks was very low (17 peaks corresponding to 5% of the total number of predicted peaks; Figure 4). Decreasing the stringency of the

DOI: 10.1002/yea

Table 1. bPeaks results for different sets of parameter values (Pdr1p data)

T1	T2	T3	T4	Number of peaks detected	Average C1	Average C2	Average C3	Average C4	Number of peaks in promoters	Average peak size	Parameter set	Specificity	Sensitivity
6	**2**	**3**	**0.9**	**24**	**1830.4**	**140.4**	**3.41**	**8.57**	**17**	**177.0**	**S1**	**+++++**	**+**
4	2	3	0.9	24	1830.4	140.4	3.41	8.57	17	177.0			
2	2	3	0.9	24	1830.4	140.4	3.41	8.57	17	177.0			
6	4	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
4	4	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
2	4	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
6	6	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
4	6	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
2	6	3	0.9	24	1740.8	145.0	3.38	8.57	17	185.4			
6	2	2	0.9	122	876.1	140.9	2.44	8.26	49	181.1	S2		
4	2	2	0.9	122	876.1	140.9	2.44	8.26	49	181.1			
2	2	2	0.9	122	876.1	140.9	2.44	8.26	49	181.1			
6	4	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
4	4	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
2	4	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
6	6	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
4	6	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
2	6	2	0.9	122	868.3	144.6	2.43	8.26	49	184.8			
6	2	3	0.7	121	751.3	61.7	3.49	7.43	52	171.4	S3		
6	4	3	0.7	121	733.5	62.6	3.49	7.43	52	173.1			
6	6	3	0.7	121	733.5	62.6	3.49	7.43	52	173.1			
4	2	3	0.7	128	725.8	60.1	3.47	7.39	53	171.4			
2	2	3	0.7	128	725.8	60.1	3.47	7.39	53	171.4			
4	4	3	0.7	128	709.0	61.0	3.46	7.39	53	173.0			
4	6	3	0.7	128	709.0	61.0	3.46	7.39	53	173.0			
2	4	3	0.7	128	709.0	61.0	3.46	7.39	53	173.0			
2	6	3	0.7	128	709.0	61.0	3.46	7.39	53	173.0			
6	2	2	0.7	336	578.9	89.0	2.68	7.63	121	183.3	S4		
6	4	2	0.7	336	576.1	90.3	2.67	7.63	121	184.6			
6	6	2	0.7	336	576.1	90.3	2.67	7.63	121	184.6			
4	2	2	0.7	611	444.7	74.5	2.51	7.34	208	175.7			
4	4	2	0.7	611	443.1	75.3	2.50	7.34	208	176.5			
4	6	2	0.7	611	443.1	75.3	2.50	7.34	208	176.5			
2	2	2	0.7	628	438.5	73.9	2.49	7.33	215	175.3			
2	4	2	0.7	628	437.0	74.6	2.49	7.33	215	176.1			
2	6	2	0.7	628	437.0	74.6	2.49	7.33	215	176.1	S5	+	+++++

The R code presented in Methods section allows this table to be obtained, summarizing basic information regarding the detected peaks (number of detected peaks, average values for C1, C2, C3 and C4 parameters described in Materials andethods, number of peaks in promoters and average peak sizes). Parameter associations were ordered here according to the average IP signal (C1 value). Four sets of parameters (referred to as S1, S2, S3 and S4) were selected for further analyses and are discussed in the main text. Peak calling specificity and sensitivity associated to parameter sets 1 and 4 are symbolized on the right.

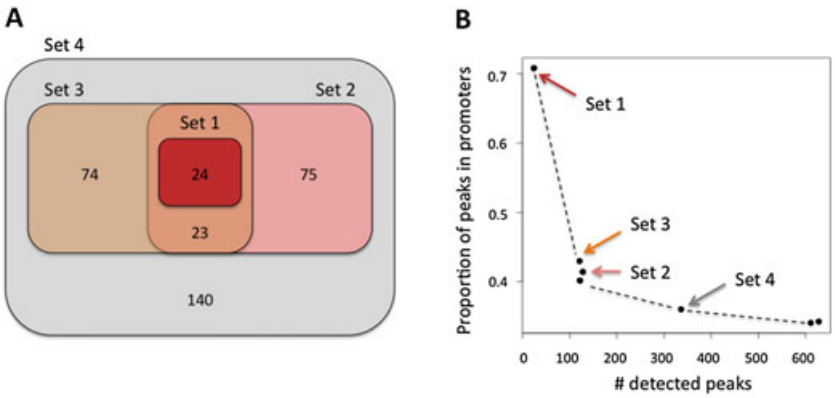

Figure 3. Comparison of lists of peaks detected using different parameter associations in bPeaks. (A) Four sets of parameters (S1, S2, S3 and S4) were selected for peak list comparison (see Table 1); overlaps between the detected peaks are shown here. Peaks detected with S1 parameters are included in peaks detected with S2 and S3 parameters, peaks detected with S2 and S3 overlap and are all included in peaks detected with S4 parameters. (B) Evolution of the proportion of peaks in promoters according to the number of detected peaks. The proportion of peaks in promoters decreases when the number of detected peaks increases; the proportions of peaks in promoters were calculated using the 'Peak-to-gene assignments' function in bPeaks (see Methods)

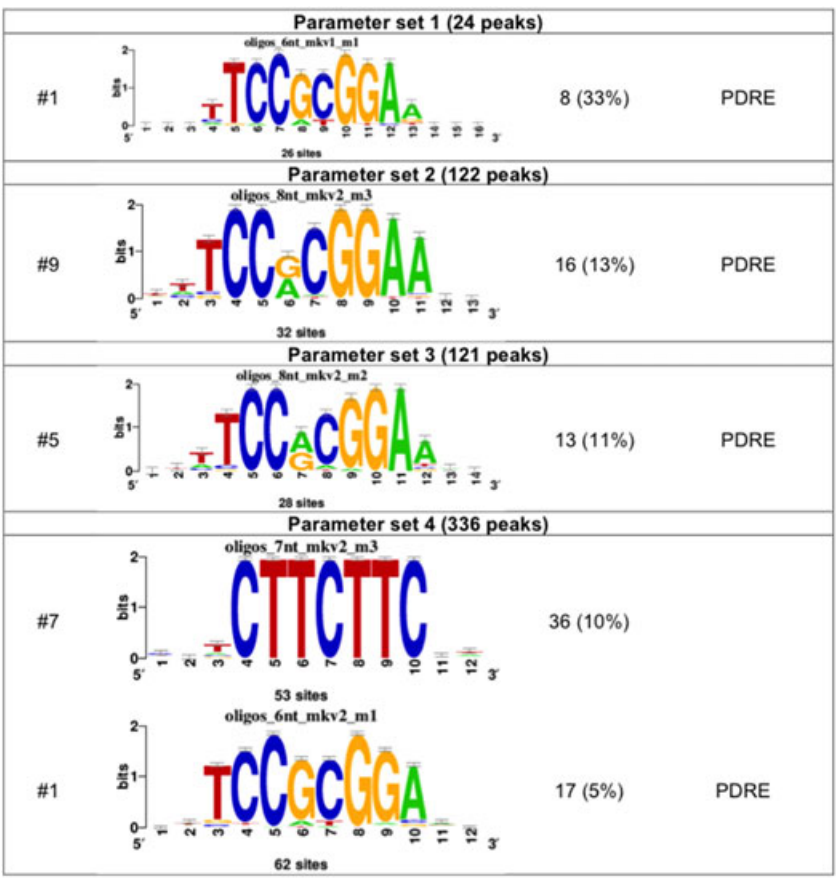

Figure 4. *De novo cis*-regulatory motif discovery. DNA sequences associated with the genomic positions of peaks detected with parameter sets S1, S2, S3 and S4 (see Table 1 and main text) were analysed using the 'peak-motif' program with default parameters. Complete results are presented in Supplementary data S1 (see supporting information). Motifs that exhibited the highest percentages in initial lists of peaks are shown here (motif 1 for the S1 list, motif 9 for the S2 list, motif 5 for the S3 list and motifs 7 and 1 for the S4 list); motifs that match the PDRE are indicated by 'PDRE'

DOI: 10.1002/yea

parameters in S4 therefore increased the sensitivity of the motif discovery (number the peaks with PDRE motif is higher) but also strongly decreased the associated specificity (the proportion of peaks with PDRE motif is lower). Considering these criteria, the S1 list was the most specific list, with one-third of the peaks containing a PDRE, but the S2 list presented a better compromise between sensitivity and specificity, as it allowed identifying twice as many peaks containing PDRE motifs. Importantly, the peaks containing PDRE motifs were ranked as the best positions in all lists (see supporting information, Supplementary data S2), emphasizing the relevance of the peak order provided by bPeaks. Therefore, bPeaks was efficient in providing high quality data that allowed the identification, without any *a priori*, of the correct Pdr1p binding site. This is connected with the fact that it efficiently ranked as best positions peaks with Pdr1p DNA consensus sequences. We also observed that the PDRE motifs were located very close to the centres of the peaks, as illustrated in Figure 5 (green boxes). This indicates that bPeaks was very precise in the prediction of the actual DNA binding site of Pdr1p from the ChIPseq data.

Proportions of peaks in promoters

Using the 'peak-to-gene' function available in bPeaks (see Methods), we identified peaks located in the 800 nucleotides upstream of the ATG of a protein-encoding gene. They were defined as 'peaks in promoters'. In yeasts, most of the regulatory binding sites for specific TFs have been found in these regions (Harbison *et al.*, 2004). We can therefore expect that the majority of the peaks, which are meaningful in terms of transcriptional regulation, stand in promoters. The proportion of 'peaks in promoters' in a list may thus reflect the specificity and the relevance of a particular combination of parameters. This information is presented in Table 1. We observed that the proportion of 'peaks in promoters' decreased together with the stringency of the bPeaks parameters (Figure 3B). It is very high in the S1 list (>70%) and the proportion of peaks in promoters rapidly dropped to 45% in the longer lists (S2, S3 and S4). This

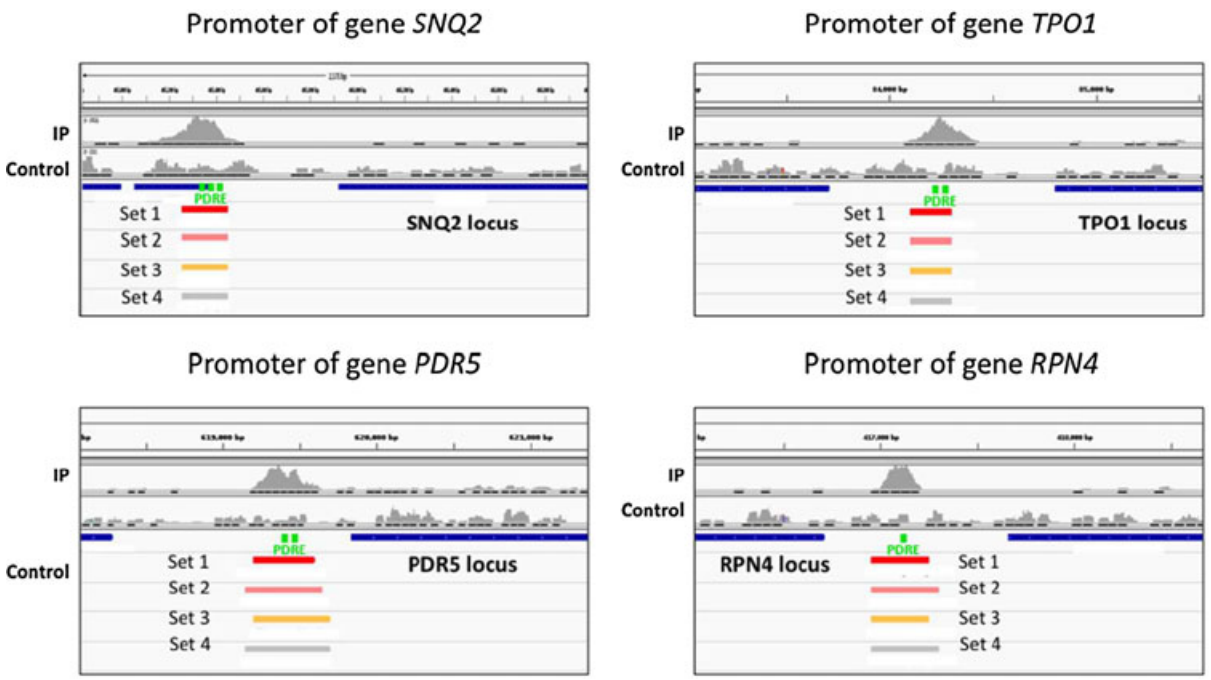

Figure 5. Illustration of peaks detected in promoter sequences of Pdr1p targets. Promoters of genes *SNQ2*, *TPO1*, *PDR5* and *RPN4* are shown here. IP and control signals are shown in grey, gene locations in blue, and genomic positions (peaks) detected with bPeaks are represented in red (parameter set S1), pink (parameter set S2), yellow (parameter set S3) and grey (parameter set S4); PDRE motifs are shown in green. These images are screenshots of the IGV software

DOI: 10.1002/yea

parameter can also be used to assess the interest of choosing one list or another for further analyses.

Validation of bPeaks results by comparison with Pdr1p targets described in the literature

A last outcome expected from ChIPseq data is to identify target genes of the studied transcription factor. In that respect, we searched for all documented DNA binding evidences for Pdr1p in the YEASTRACT database (Teixeira *et al.*, 2014) and compared this list with the lists of peaks found by bPeaks in the promoters of protein-encoding genes (see supporting information, Supplementary data S2). We observed that the DNA binding evidences found in the literature largely overlapped the PDRE-containing peaks that we identified above. Hence, the proportion of Pdr1p targets was 45% in the S1 list, 29% in the S2 list, 19% in the S3 list and 14% in the S4 list. As the proportion of yeast promoters annotated in the database as Pdr1p targets is only 6% (see supporting information, Supplementary data S2), bPeaks appeared to be very efficient in identifying, as 'best' peaks, promoter regions of genes that were already described as Pdr1p targets. Moreover, as was observed for PDRE-containing peaks, the validated targets of Pdr1p were essentially found in the first positions of lists retrieved by the bPeaks program (see supporting information, Supplementary data S2).

Generalization to other datasets

With the case study of the Pdr1p transcription factor, we presented a general protocol for applying bPeaks to ChIPseq data that consists in: (a) testing the influence of bPeaks parameter values on detected peaks; and (b) assessing the biological significance of the retrieved lists of peaks using additional information (e.g. detection of regulatory motifs, proportion of peaks in promoters or a reference list of target genes described in the literature). This protocol can be easily used for other ChIPseq contexts, i.e. different transcription factors and different species. To demonstrate this point, we used ChIPseq datasets obtained in the species *C. albicans* and *C. glabrata*. All results are available in Supplementary data S3 (see supporting information).

Sfl1p transcription factor in C. albicans

In *Candida albicans*, Sfl1p is a transcription factor involved in morphogenesis and virulence. Its transcriptional targets were identified recently using ChIPseq experiments and it has been demonstrated that Sfl1p exerts its regulatory activity by binding two transcriptional co-factors, Ndt80p and Efg1p (Znaidi *et al.*, 2013). In their original publication, the authors used the MACS program (Zhang *et al.*, 2008) to perform peak calling. To test the relevance of the bPeaks approach, ChIPseq data on Sfl1p were downloaded (see Methods) and used for peak calling with bPeaks. As described for Pdr1p datasets, we first evaluated the influence of parameters using the 36 combinations of values (identical to these used Pdr1p data), and seven different lists of peaks were obtained (see supporting information, Supplementary data S3). To assess the biological relevance of the retrieved lists, we searched for regulatory motifs using the 'peak-motifs' program (Thomas-Chollier *et al.*, 2012). For all lists of peaks, we were able to detect motifs similar to the consensus-binding sites of the Sfl1p co-factors (Ndt80p and Efg1p; an illustration is shown Figure 6A), as described in the original publication (Znaidi *et al.*, 2013).

CgAp1p transcription factor in Candida glabrata

CgAp1p is the orthologous TF of Yap1p, which controls the transcriptional response to oxidative stress in *S. cerevisiae* (Lucau-Danila *et al.*, 2003; Lelandais *et al.*, 2008; Goudot *et al.*, 2011). These TFs have been shown to enter the nucleus and bind their DNA consensus motifs, called Yap Response Element (YRE), only in stress conditions. Comparing the DNA binding pattern of CgAp1p TF in stressful and normal growth conditions represents a good opportunity to test the specificity of the bPeaks program. ChIPseq experiments were therefore performed in oxidative stress and in optimal growth conditions (see Methods). Each stress and non-stress condition was independently compared to the associated control samples (INPUT), applying the bPeaks general protocol (parameter evaluation followed by detections of regulatory motifs). As expected, the 36 combinations of parameters all retrieved numbers of detected peaks much higher in stress condition than in the absence of stress [see Figure 6B for S1 and S4 results, and Supplementary data S3 (see supporting information) for all other combinations]. Notably, only five peaks were found in normal conditions with the most stringent sets of S1 parameters

DOI: 10.1002/yea

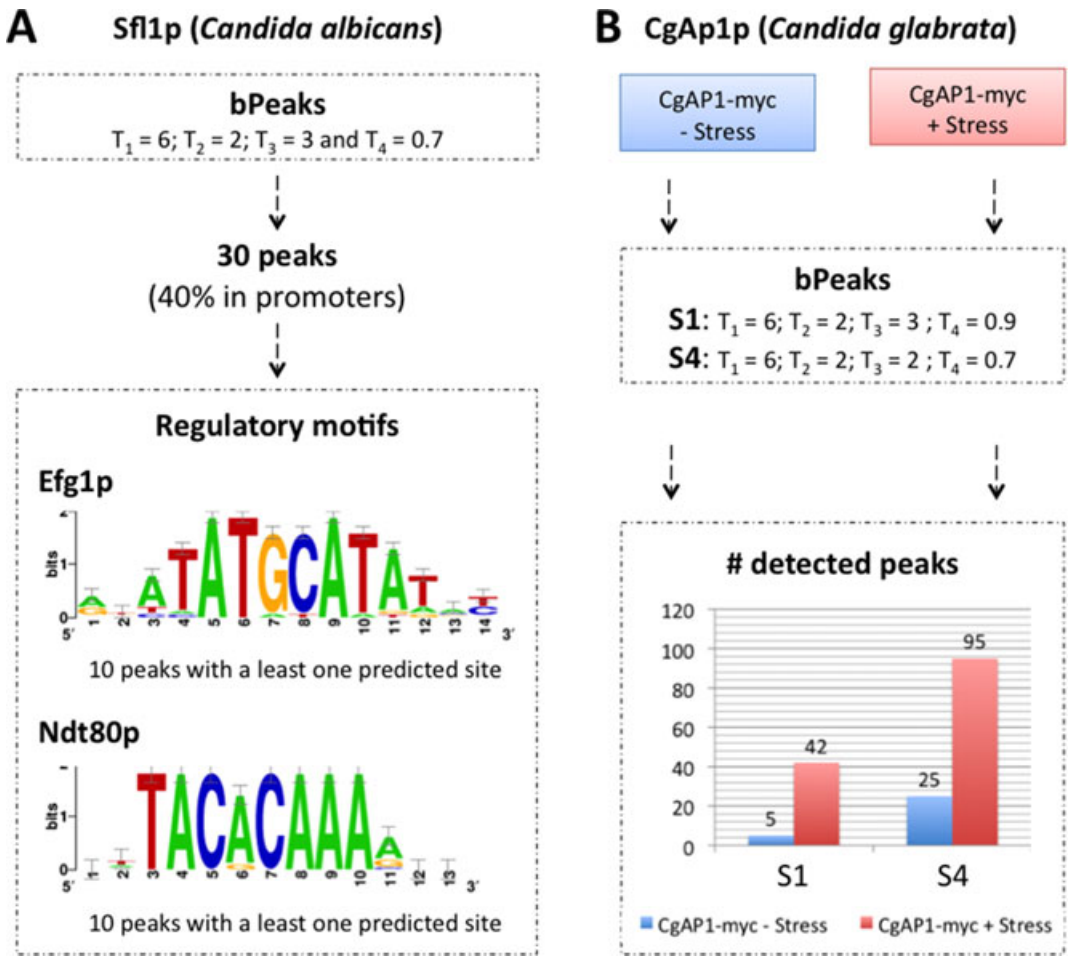

Figure 6. Results of the bPeaks program analysing ChIPseq data in *Candida albicans* (Sfl1p transcription factor) and *Candida glabrata* (Cgap1p transcription factor). An identical procedure to that presented in the main text for Pdr1p data was applied. Parameter evaluation was first performed and lists of detected peaks were retrieved. Associated DNA sequences were analysed for regulatory motif discovery; all the results are provided in Supplementary data S3 (see supporting information). Illustrations of the bPeaks results obtained in *C. albicans* and *C. glabrata* are shown in (A) and (B), respectively

(Figure 6B). Visual inspection of these peaks suggested experimental bias associated to ChIPseq experiments; this will be discussed elsewhere (Merhej *et al.*, manuscript in preparation). Also, the YRE motif was found as the best motif from all the lists of peaks defined by bPeaks in stress conditions, whereas it was not found with the lists of peaks obtained from the control experiment (see supporting information, Supplementary data S3).

In conclusion, the Sfl1p and Cgap1p analyses presented here confirmed that bPeaks is efficient in sorting out biologically meaningful lists of peaks from ChIPseq data conducted on different types of TF, in different yeast species and with different sequencing coverages.

Evaluation of bPeaks performances in the light of other peak-calling methods

Numerous tools are available for peak-calling analyses. To evaluate bPeaks performances, we compared the results obtained by bPeaks with those obtained with three popular peak-calling programs – MACS (Zhang *et al.*, 2008), SPP (Kharchenko *et al.*, 2008) and BayesPeak (Spyrou *et al.*, 2009; Cairns *et al.*, 2011). Results obtained in the yeasts *S. cerevisiae*, *C. albicans* and *C. glabrata* using the Pdr1p, Sfl1p and CgAp1p ChIPseq data are shown Table 2. We observed that the three programs produced very different results in terms of detected peak numbers and peak sizes.

DOI: 10.1002/yea

Table 2. Summary table for peaks identified by different peak calling methods

Peak calling program	Computational time	Number of peaks detected	Average peak size (nt)
Pdr1p – *S. cerevisiae*			
MACS	20 min 04 s	248	932
SPP	69 min 46 s	67	2878
BayesPeak	28 min 28 s	3500	183
bPeaks (S2 parameters)	04 min 07 s	122	181
Sfl1p – *C. albicans*			
MACS	025 min 00 s	170	2678
SPP	138 min 55 s	0	0
BayesPeak	023 min 47 s	338	207
bPeaks (S2 parameters)	006 min 48 s	35	278
CgAp1p – *C. glabrata* (+stress)			
MACS	019 min 30 s	615	1737
SPP	415 min 17 s	3068	3765
BayesPeak	024 min 51 s	1591	229
bPeaks (S2 parameters)	005 min 07 s	95	298
CgAp1p – *C. glabrata* (– stress)			
MACS	017 min 58 s	340	1149
SPP	364 min 08 s	2733	2644
BayesPeak	024 min 03 s	2560	210
bPeaks (S2 parameters)	005 min 02 s	25	252

Default parameter values were used for MACS, SPP and BayesPeak and parameter set S2 was used for bPeaks (T1 = 6, T2 = 2, T3 = 2 and T4 = 0.9; see Table 1); the Pdr1p, CgAp1p and Sfl1p datasets are detailed in the main text. Number of detected peaks and average peak size are indicated for each method applied to each dataset. Computational times were obtained on an HP Z820 Workstation [Intel Xeon E5-2609 2.4 Ghz CPU and 16 GB DDR3-1600 (8 × 2 GB) RAM].

Remarkably, bPeaks exhibited the smaller computational time (around 5 min). This is an advantage of its specific development to study small genomes. Interestingly, we could observe that the numbers of peaks found were, in many occurrences, in contradiction with what is known of the biology of the system studied. For instance, BayesPeak systematically detected a very high number of peaks and SPP proposed similar numbers of peaks for CgAp1p with and without stress, whereas this factor is known to be enriched in the nucleus only in stress conditions. Also, no peaks were found by SPP in the case of Sfl1p. This result is in contradiction with what has been published on these data (Znaidi *et al.*, 2013).

Lists of peaks detected with MACS, SPP and bPeaks in the Pdr1p dataset (*S. cerevisiae*) were next retrieved for further comparisons. BayesPeak results were not used because this program provides lists of peaks sorted by genomic positions, and not by confidence levels (as the MACS, SPP and bPeaks programs do). Since the numbers of peaks were very different from one method to another, we selected for comparison the top 50 'best peaks', i.e. peaks with the highest confidence, retrieved by MACS, SPP and bPeaks, respectively. We first searched for peaks in promoters of well-characterized Pdr1p target genes (see supporting information, list in Supplementary data S4). We found nine peaks in promoters of Pdr1p targets in the MACS list, 14 peaks in the SPP list and 12 peaks in the bPeaks list (see supporting information, Supplementary data S4, for all lists of peaks with Pdr1p target annotations). We next analysed the ranks that exhibited peaks in promoters of Pdr1p targets, in the top 50 'best peaks'; the results are presented in Figure 7. In terms of detection and ranking of Pdr1p targets, the performances obtained by bPeaks were similar to those of SPP or MACS. We finally compared the sizes of the peaks predicted by the different methods. Our rationale was that the smaller the peak sizes are, the better the accuracy achieved for DNA binding site predictions. The average size of peaks is remarkably small for bPeaks (181 bp) compared to MACS (2310 pb) and SPP (2878 bp) (Table 2). Our bioinformatics tool is more precise than MACS and SPP in localizing the actual binding site of the transcription factor. This is well illustrated in Supplementary data S5 (see supporting information). Altogether, these results conform to the idea that whereas bPeaks does not use a sophisticated statistical model, it is at least as efficient as existing tools in proposing lists of peaks that are enriched in potential targets, and is more precise in defining the peak loci.

Discussion

With the explosion of high-throughput sequencing in general and ChIPseq in particular, dozens of tools for peak calling were developed (for a detailed list, see Bailey *et al.*, 2013). In this context, the

DOI: 10.1002/yea

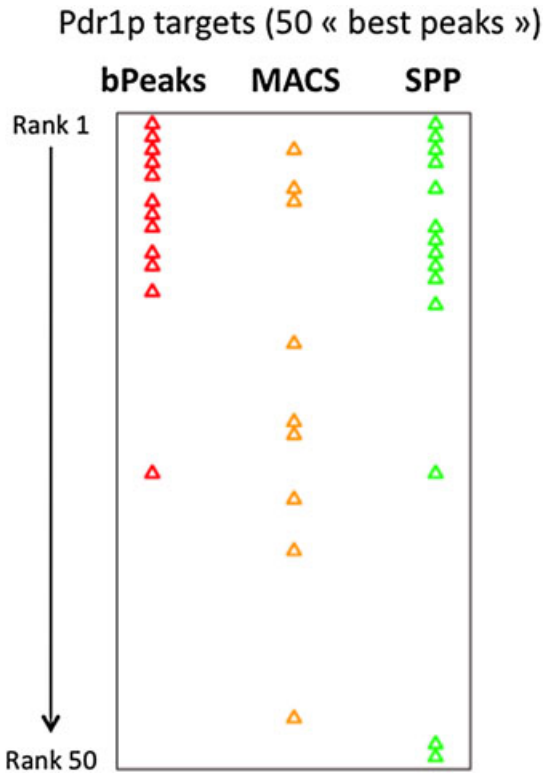

Figure 7. Positions of Pdr1p target genes in the list of 50 best peaks detected with the bPeaks, MACS and SPP programs. A complete list of Pdr1p targets, together with detailed annotation of peaks, are available in Supplementary data S4 (see supporting information)

originality of our bioinformatics tool rely on its dedicated design to study ChIPseq data related to specific transcription factors in small eukaryotic genomes, such as yeasts. In Table 2 and Supplementary data S4 and S5 (see supporting information) we compared bPeaks performances with other commonly used peak-calling algorithms (MACS, SPP, BayesPeak). As a result, bPeaks provided lists of peaks which were better in terms of biological relevance (i.e. better enrichment of potential targets and more precise definition of the transcription factor binding sites). Importantly, this does not mean that bPeaks is intrinsically better than the other methods. This is due to the fact that bPeaks was optimized for yeast ChIPseq data, when the other methods have been designed to have a wider range of applications. Accurate optimization of parameters would certainly lead to increasing the performance of these programs. Still, this comparison demonstrates that we succeeded in making an appropriate and efficient tool to study ChIPseq data in yeasts. We believe that this specialization of bPeaks has three main advantages compared to more general methods.

First, bPeaks uses simple parameters to identify peaks, which actually correspond to the parameters that a user is considering when manually checking peaks on a genome browser. Because of this simplicity, and because it is meant to be applied to small genomes, bPeaks can be systematically used to explore its parameter space in a very reasonable computing time. Running the 36 different combinations of parameters on a yeast genome takes 1 h on a standard working desk station. This approach has the advantage of helping the user to get a comprehensive view of the influence of the different parameter changes on the final output of the program. There is a very direct relationship between the specified parameter values and the basic features of the detected peaks. This allows precise control of the balance between stringency and sensitivity in peak-calling analyses.

Second, while most of the existing methods use density estimators in fixed windows to limit the calculation time required when analysing large genomes, bPeaks analyses the ChIPseq signal at the nucleotide level, which provides a higher spatial resolution of the detected peaks. This is well illustrated by the sharpness of the peaks found by bPeaks, which were perfectly centred on the putative consensus-binding site of the studied TF (Figure 5 and Supplementary data S5, see supporting information). Together with the relevance of the parameters used to define peaks, this characteristic may explain why bPeaks was so efficient in predicting the consensus sequence of TF binding sites on the three examples we studied.

Third, bPeaks output files were designed to be used to estimate the relevance of the detected peaks, using complementary tools to, for instance: (a) visualize results in a genome browser; (b) search for regulatory motifs; or (c) automatically assign peaks to particular annotated regions in a genome. In the specific case of yeast transcription factors, the number of peaks that contain a predicted TF binding site, together with the proportion of peaks in promoter regions, appeared to be very biologically meaningful information. Still it is important to consider that ChIPseq technique generates numerous false positives that are due to sequencing biases, libraries complexity, the chromatin state of highly expressed genes, etc. (for a detailed

DOI: 10.1002/yea

description, see Park *et al.*, 2013). This could explain why detected peaks are not systematically associated to regulatory motifs and why peaks can be detected in conditions where a TF is not expected to be functional. Of course, the intrinsic complexity of transcriptional regulations that occur in a cell is another explanation, and distinguishing 'real' peaks (interaction between TF and DNA) from 'artefact' peaks (other peaks) is the main challenge faced by peak-calling programs. In this context, the use of a control sample, as in bPeaks, is a prerequisite to improving peak-calling results. Also, we believe that it is important not to trust only final statistical parameters (e.g. *p* values). In this article we used an original strategy in which different parameter values for peak detection were systematically tested and the lists of peaks obtained were used to predict the DNA-binding motif(s) of the TF and calculate the proportion of peaks in promoters. These are two other types of reference information, which in turn were used to address the sensitivity and specificity of the initial parameter choices. The originality of the bPeaks program is therefore to give the opportunity to the user to define, based on the user's own criteria, the list of peaks that present a good compromise between sensitivity and specificity.

We used bPeaks to analyse three different transcription factors, from three different protein families with different DNA binding properties, in three different yeast species. In all three cases, bPeaks provided lists of peaks that allowed accurate predictions of the DNA consensus sequences. Remarkably, the peaks containing these motifs were ranked among the top of the best peaks in the lists (see supporting information, Supplementary data S2 and S3). Notably, when we applied bPeaks to the CgAp1p ChIPseq conducted in the absence of stress, a condition in which no binding events was expected, we obtained very short lists of peaks (five at minimum, 30 with the less stringent combinations of parameters). These results fully validate the efficiency and relevance of the bPeaks program in identifying TF binding sites from ChIPseq experiments conducted in simple eukaryotes with a genome size of ca. 10–20 Mb.

Acknowledgements

This study was funded by the STRUDYEV project of the French National Research Agency (ANR; Grant No. ANR-JCJC-2010). Jawad Merhej is an ANR post-doctoral fellow hired in the frame of the STRUDYEV project. High-throughput sequencing was performed on the Genomic Paris Centre IBENS platform, member of 'France Génomique' (Grant No. ANR10-INBS-09-08). This platform has received support under the program 'Investissements d'Avenir' launched by the French Government and implemented by the ANR (references ANR-10-LABX-54 MEMO LIFE and ANR-11-IDEX-0001-02 PSL* Research University).

References

Bailey T, Krajewski P, Ladunga I, *et al*. 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol* **9**(11): e1003326.

Boeva V, Lermine A, Barette C, *et al*. 2012. Nebula – a web-server for advanced ChIP-seq data analysis. *Bioinformatics* **28**(19): 2517–2519.

Cairns J, Spyrou C, Stark R, *et al*. 2011. BayesPeak – an R package for analysing ChIP-seq data. *Bioinformatics* **27**(5): 713–714.

Cheng C, Min R, Gerstein M. 2011. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27**(23): 3221–3227.

Cherry JM, Hong EL, Amundsen C, *et al*. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**(database issue): D700–705.

DeRisi J, van den Hazel B, Marc P, *et al*. 2000. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Lett* **470**(2): 156–160.

Devaux F, Marc P, Bouchoux C, *et al*. 2001. An artificial transcription activator mimics the genome-wide properties of the yeast Pdr1 transcription factor. *EMBO Rep* **2**(6): 493–498.

Diaz A, Park K, Lim DA, *et al*. 2012. Normalization, bias correction, and peak calling for ChIP-seq. *Stat Appl Genet Mol Biol* **11**(3): 1515–1544.

Fardeau V, Lelandais G, Oldfield A, *et al*. 2007. The central role of PDR1 in the foundation of yeast drug resistance. *J Biol Chem* **282**(7): 5063–5074.

Fejes AP, Robertson G, Bilenky M, *et al*. 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**(15): 1729–1730.

Goudot C, Etchebest C, Devaux F, *et al*. 2011. The reconstruction of condition-specific transcriptional modules provides new insights in the evolution of yeast AP-1 proteins. *PLoS One* **6**(6): e20924.

Harbison CT, Gordon DB, Lee TI, *et al*. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99–104.

Inglis DO, Arnaud MB, Binkley J, *et al*. 2012. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res* **40**(database issue): D667–674.

Johnson DS, Mortazavi A, Myers RM, *et al*. 2007. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**(5830): 1497–1502.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**(12): 1351–1359.

Kidder BL, Hu G, Zhao K. 2011. ChIP-Seq: technical considerations for obtaining high-quality data. *Nat Immunol* **12**(10): 918–922.

Kim TH, Ren B. 2006. Genome-wide analysis of protein–DNA interactions. *Annu Rev Genomics Hum Genet* **7**: 81–102.

Kolaczkowska A, Goffeau A. 1999. Regulation of pleiotropic drug resistance in yeast. *Drug Resist Update* **2**(6): 403–414.

Landt SG, Marinov GK, Kundaje A, *et al*. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9): 1813–1831.

Langmead B, Trapnell C, Pop M, *et al*. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Lefrancois P, Euskirchen GM, Auerbach RK, *et al*. 2009. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genom* **10**: 37.

Lelandais G, Tanty V, Geneix C, *et al*. 2008. Genome adaptation to chemical stress: clues from comparative transcriptomics in *Saccharomyces cerevisiae* and *Candida glabrata*. *Genome Biol* **9**(11): R164.

Li H, Handsaker B, Wysoker A, *et al*. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078–2079.

Liang K, Keles S. 2012. Normalization of ChIP-seq data with control. *BMC Bioinformat* **13**: 199.

Lucau-Danila A, Delaveau T, Lelandais G, *et al*. 2003. Competitive promoter occupancy by two yeast paralogous transcription factors controlling the multidrug resistance phenomenon. *J Biol Chem* **278**(52): 52641–52650.

Malone BM, Tan F, Bridges SM, *et al*. 2011. Comparison of four ChIP-Seq analytical algorithms using rice endosperm H3K27 trimethylation profiling data. *PLoS One* **6**(9): e25260.

Mamnun YM, Pandjaitan R, Mahe Y, *et al*. 2002. The yeast zinc finger regulators Pdr1p and Pdr3p control pleiotropic drug resistance (PDR) as homo- and heterodimers *in vivo*. *Mol Microbiol* **46**(5): 1429–1440.

Nagasaki H, Mochizuki T, Kodama Y, *et al*. 2013. DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res* **20**(4): 383–390.

Park D, Lee Y, Bhupindersingh G, *et al*. 2013. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS One* **8**(12): e83506.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**(10): 669–680.

Pepke S, Wold B, Mortazavi A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**(11, suppl): S22–32.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841–842.

Robertson G, Hirst M, Bainbridge M, *et al*. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**(8): 651–657.

Rozowsky J, Euskirchen G, Auerbach RK, *et al*. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**(1): 66–75.

Schjerling P, Holmberg S. 1996. Comparative amino acid sequence analysis of the C6 zinc cluster family of transcriptional regulators. *Nucleic Acids Res* **24**(23): 4599–4607.

Sherman DJ, Martin T, Nikolski M, *et al*. 2009. Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res* **37**(database issue): D550–554.

Spyrou C, Stark R, Lynch AG, *et al*. 2009. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformat* **10**: 299.

Teixeira MC, Monteiro PT, Guerreiro JF, *et al*. 2014. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **42**(database issue): D161–166.

Thomas-Chollier M, Herrmann C, Defrance M, *et al*. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* **40**(4): e31.

Wang J, Lunyak VV, Jordan IK. 2013. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics* **29**(4): 492–493.

Wilbanks EG, Facciotti MT. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One* **5**(7): e11471.

Xu H, Handoko L, Wei X, *et al*. 2010. A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **26**(9): 1199–1204.

Zhang Y, Liu T, Meyer CA, *et al*. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Znaidi S, Nesseir A, Chauvel M, *et al*. 2013. A comprehensive functional portrait of two heat shock factor-type transcriptional regulators involved in *Candida albicans* morphogenesis and virulence. *PLoS Pathog* **9**(8): e1003519.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Supplementary data S1. All motifs detected in S1, S2, S3 and S4 lists of peaks. The 'peak-motif' tool was used, with default parameters

Supplementary data S2. Detailed lists of peaks associated with parameters in S1, S2, S3 and S4, located in the promoters of genes (800 bp). The names of the genes are indicated, together with information related to the presence of a PDRE motif in promoters and DNA-binding evidence in the YEASTRACT database. All Pdr1p targets stored in YEASTRACT are also noted

Supplementary data S3. Detailed results of bPeaks analyses in *Candida* yeast species

Supplementary data S4. Peak-calling results obtained with MACS, SPP, BayesPeak and bPeaks programs on Pdr1p data

Supplementary data S5. Illustrations of peaks detected in promoter sequences of Pdr1p targets. bPeaks results are shown in red, MACS in orange, SPP in green and BayesPeak in purple