Genome-wide generalized additive models

Georg Stricker[1,2,5], Alexander Engelhardt[1,5], Daniel Schulz[1], Matthias Schmid[3], Achim Tresch[4], and Julien Gagneur[1,2,*]

[1]*Gene Center Munich and Department of Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Straße 25, 81377 Munich, Germany.*

[2]*Technische Universität München, Department of Informatics, Boltzmannstr. 3, 85748 Garching, Germany.*

[3]*Institut für Medizinische Biometrie, Informatik und Epidemiologie, University Hospital Bonn, Sigmund-Freud-Straße 25, 53105 Bonn, Germany.*

[4]*Institute for Genetics, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany.*

[5]*These authors contributed equally to this work.*

**Correspondence should be addressed to: J.G. (gagneur@in.tum.de)*

ABSTRACT

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) is a widely used approach to study protein-DNA interactions. To analyze ChIP-Seq data, practitioners are required to combine tools based on different statistical assumptions and dedicated to specific applications such as calling protein occupancy peaks or testing for differential occupancies. Here, we present GenoGAM (Genome-wide Generalized Additive Model), which brings the well-established and flexible generalized additive models framework to genomic applications using a data parallelism strategy. We model ChIP-Seq read count frequencies as products of smooth functions along chromosomes. Smoothing parameters are estimated from the data eliminating ad-hoc binning and windowing needed by current approaches. We derived a peak caller based on GenoGAM with performance matching state-of-the-art methods. Moreover, GenoGAM provides significance testing for differential occupancy with

controlled type I error rate and increased sensitivity over existing methods. By analyzing a set of DNA methylation data, we further demonstrate the potential of GenoGAM as a generic analysis tool for genome-wide assays.

INTRODUCTION

Chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) is the reference method used for genome-wide quantification of protein-DNA interactions[1]. It is used to study a wide range of fundamental genome biology processes covering transcription, replication, and maintenance. ChIP-Seq consists of cross-linking DNA with chromatin, followed by DNA fragmentation and immunoprecipitation of the protein of interest along with its bound DNA fragments. The DNA fragments are then released, amplified, and sequenced. ChIP-Seq has been applied for studying DNA-bound proteins of various functions and therefore with various patterns of distribution along the genome. These include transcription factors that are bound at discrete binding sites[2,3], histone modifications[3,4] which are found at nucleosomes, or the transcription[3] and replication machinery which are even more broadly distributed. Often, the quantities of interest are the occupancies relative to technical controls, such as the input (a sample that was not subject to the immunoprecipitation step), between genetic backgrounds, treatments, or combinations thereof.

Although ChIP-Seq is a very generic methodology to study protein-DNA interactions, statistical analysis methods have been so far dedicated to specific applications. Early work has focused on transcription factors with discrete binding sites, typically DNA motifs at promoters or transcriptional enhancers[2,5]. ChIP-Seq read coverage then shows peaks localized at the binding sites. The aim of these statistical methods is to identify these peaks and their statistical significance, typically by controlling the false discovery rate. For example, MACS[5] is a widely used[6,7] peak caller that assumes a Poisson distribution for the count data and computes peak significance based on a combination of global and local rate. ZINBA[8] combines a negative binomial mixture model for background and enriched regions with a zero inflated component for regions with excessive zero counts. The specific calling of narrow and wide peaks was made possible by JAMM[9], which

makes use of replicates, and is based on a mixture model of enriched and non-enriched regions.

Testing for differential overall occupancies at regions of interest across conditions is done by testing for differences in number of reads overlapping the region [10]. Complementary to testing for overall occupancies, MMdiff[11] allows testing for differences in shapes in given regions. Lun et al.[12] provide a framework to test differential occupancies between conditions across windows in given regions while properly controlling for false discovery rate. This approach allows both testing for differences in overall occupancies and in shapes.

Hence, practitioners rely on different statistical frameworks for peak calling tasks and differential occupancies. However, flexible handling of replicates and additional control factors is not always possible. Moreover, current methods rely on binning and sliding window techniques, whose choice of the window size is not data-driven but subjective. Another limitation is that the more general task of statistical inference of a genome-wide bias-corrected occupancy track is not addressed.

Here, we introduce GenoGAM (Genome-wide Generalized Additive Model), which provides a statistical framework to simultaneously address the above issues. Our model describes genome-wide occupancy by smooth functions, which facilitate downstream applications such as peak calling or differential binding analysis. GenoGAM normalizes for sequencing depth and can handle factorial experimental designs, including biological replicates and multiple controls. The amount of smoothing is estimated in an automatic, data-driven manner and thus avoids introducing subjectivity from the analyst. When analyzing differential binding in a factorial design, we obtain well-calibrated per-base-pair p-values. Application to datasets of human and yeast shows that GenoGAM is as performant as dedicated methods for peak calling and much more sensitive than state-of-the art differential occupancy methods. By providing an approximation to a conventional generalized additive model (GAM[13]) that allows a data parallelism implementation, GenoGAM scales linearly with the number of data points and is thus computationally amenable to whole-genome applications. Our method provides a framework that is applicable not only to ChIP-Seq

data, but also to other next-generation sequencing data such as DNA methylation data (Figure 1a).

RESULTS AND DISCUSSION

A generalized additive model for ChIP-Seq data

We consider an experiment consisting of a set of ChIP-Seq samples. A data point is defined by a pair of a ChIP-Seq sample and a genomic position. We denote by x_i the genomic position of the i-th data point, by j_i its ChIP-Seq sample and by $y_i \geq 0$ the number of fragments in sample j_i centered at position x_i. For single-end libraries, the fragment center is estimated by shifting the read end position by a constant (Methods). When reducing ChIP-Seq data to fragment centers rather than full base coverage, each fragment is counted only once. This reduces artificial correlation between adjacent nucleotides. We model the counts y_i using the following generalized additive model:

$$y_i \sim \mathrm{NB}(\mu_i, \theta) \tag{1}$$

$$\log(\mu_i) = o_i + \sum_{k=1}^{K} f_k(x_i)\, z_{j_i,k} \tag{2}$$

The counts y_i are assumed to follow a negative binomial distribution with means μ_i (equation 1) and a dispersion parameter θ that relates the variance to the mean such that $\mathrm{Var}(y_i) = \mu_i + \mu_i^2/\theta$. Consequently, the model accounts for overdispersion[10]. The logarithm of the mean μ_i is the sum of an offset o_i and one or more smooth functions f_k (equation 2). The offsets o_i are predefined data-point specific constants that account for sequencing depth variations (Methods). More elaborate usage could include position- and sample-specific copy number variations, or GC-biases. The indicator variable $z_{j_i,k}$ values 1 if the smooth function f_k contributes to the mean counts of sample j_i and 0 otherwise. As demonstrated in the Methods section, this formulation allows modeling IP versus input experiments as well as factorial experimental designs.

We modeled IP versus input experiments using GenoGAM with two smooth functions: f_{input} that contributes to both input and IP samples, and f_{protein} that only contributes to IP samples. More specifically, f_{input} models local ChIP-Seq biases common to input and IP, whereas f_{protein} models the protein log-occupancy up to one genome-wide scaling factor. Figure 1b shows the application of this model to one ChIP-Seq library for the *S. cerevisiae* general transcription factor TFIIB and its input control (Methods).

In GenoGAM, the smooth functions are represented by cubic spline curves, which are written as linear combinations of a set of regularly spaced B-spline basis functions b_r, i.e. $f_k(x) = \sum_r \beta_r b_r(x)$. We chose second order B-splines as basis functions, which are bell-shaped cubic polynomials over a finite support[14]. To avoid overfitting, additional smoothing of the functions f_k is carried out by penalization of the second order differences of the spline coefficients, which approximately penalizes second order derivatives of f_k – an approach called P-splines (penalized B-splines [15]). The optimization criterion for P-splines is the sum of the negative binomial log-likelihood (depending on the response vector y and the vector $\boldsymbol{\beta}$ containing the coefficients of all smooth functions) plus a penalty function that is weighted by the smoothing parameter λ:

$$\hat{\boldsymbol{\beta}} = \operatorname{argmax}\ l_{\text{NB}}(\boldsymbol{\beta}; \mathbf{y}, \theta) - \lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}, \tag{3}$$

where S is a symmetric positive matrix that encodes the squared second order differences of the coefficients $\boldsymbol{\beta}$[15]. This regularization allows dense placements of the basis functions (between 20 and 50 bp), while relying on the smoothing parameter λ to protect against overfitting. Large values of λ yield smoother functions. A single smoothing parameter common to all smooth functions proved to be sufficient for our applications. For given λ and θ, model fitting was performed using penalized iteratively re-weighted least squares (Methods).

The penalized likelihood can also be interpreted in a Bayesian fashion[16], where a multivariate Gaussian prior is placed on the coefficient vector $\boldsymbol{\beta}$. Large-sample approximations then yield a

multivariate Gaussian posterior distribution for $\boldsymbol{\beta}$, and, by the linearity of $f_k(x) = \sum_r \beta_r b_r(x)$, Gaussian posteriors for the point estimates $f_k(x)$. This allows for the construction of pointwise confidence bands [16]. An example of the fitted smooth functions and their confidence bands for the yeast transcription factor TFIIB is shown in Figure 1b.

Data-driven determination of the smoothing and dispersion parameters

To determine the optimal value for λ and θ, generalized cross-validation, which is based on an analytical leave-one-out large-sample approximation[16], yielded very wiggly fits indicative of overfitting. We thus developed an empirical cross-validation scheme. To reduce computational time, cross-validation was performed on a subset of all data. To this purpose, we selected a sufficiently large set of distinct regions that are long enough to not suffer from border effects common to spline fitting. Using 40 or more distinct regions containing at least 60 basis functions gave satisfactory empirical results (Supplementary Table 1). Also, it was important to select regions relevant for the desired application. For peak calling purposes, regions were selected that had the most significant fold change of IP versus input read counts (Methods). In each region, 10-fold cross-validation was performed, where a tenth of the data points were removed, the model was fitted on the remaining data points, and the log-likelihood of the left-out data points was computed. Parameter combinations were scored for the total out-of-sample log-likelihood over all regions. Short range correlations are strong in ChIP-Seq data and are not fully controlled by replicates or input experiments. To avoid overfitting due to short range correlations, each cross-validation fold did not consist of randomly selected single base pairs but of short intervals. The length of these intervals was about a tenth the average fragment size in absence of replicates and twice the average fragment sizes when replicates were available (Methods). Investigation on grid values of θ and λ showed that the out-of-sample log-likelihood was typically unimodal. We therefore used Nelder-Mead optimization [17] to jointly fit the two parameters in a computationally faster way than grid search.

Fitting a GAM genome-wide

Since the computation time of a GAM grows polynomially with the number of basis functions, fitting one model to a whole chromosome is unfeasible. Instead, we propose to fit separate GAMs on sequential overlapping intervals (or tiles, Fig. 2a). As overlap length increases, agreement of the fit at the midpoint of the overlapping region increases. A genome-wide fit is obtained by joining together tile fits at overlap midpoints (Fig. 2a). This approximation yields computation times that are linear in the number of basis functions at no practical precision cost (Fig. 2b). Furthermore, it allows for parallelization, with speed-ups being linear in the number of cores (Fig. 2c). This approximation parallelizes the computation over the data, which will allow future implementation of GenoGAM in map-reduce frameworks such as Spark[18].

GenoGAM provides a competitive peak caller

Because analytical derivatives of P-splines are available, identifying peaks of the protein occupancy f_{protein} is straightforward by extracting local maxima where $f'_{\text{protein}}(x) = 0$ and $f''_{\text{protein}}(x) < 0$ (Methods, Supplementary Fig. S1). To assess statistical significance of the peak heights, we introduced an empirical z-score that contrasts the estimate of the log-occupancy μ at the peak to a robust estimate of background log-occupancy level μ_0, taking both background level variance σ_0^2 and uncertainty of peak height σ^2 into account (see Methods for their estimation):

$$z = \frac{\mu - \mu_0}{\sqrt{\sigma^2 + \sigma_0^2}}. \tag{4}$$

A practical approach to model the null distribution of peak scores is to assume that false positive peaks arise from symmetric fluctuations of the background and thus distribute similarly to local minima, or peaks found when inverting the role of input and IP [5]. We therefore estimated the false discovery rate using the z-score distributions of the local minima (Methods).

We first compared the performance of GenoGAM, MACS[5], JAMM[9] and ZINBA[8] in identifying binding sites of TFIIB. For about 20% of yeast promoters, recruitment of TFIIB is triggered by the well-characterized DNA element TATA-box, providing at these promoters a ground truth for a TFIIB occupancy peak[19]. We mapped 1,105 TATA-boxes genome-wide by regular expression of a consensus motif (Methods) and considered 1 kb regions centered on TATA-boxes for benchmarking. In these regions, significant peaks (FDR < 0.1) from GenoGAM were substantially closer to TATA-boxes than those of alternative methods (median absolute distance 58 bp, third quartile 144 bp for GenoGAM versus 152 and 247 bp for MACS, 82 and 174 bp for JAMM, and 155 and 237 bp for ZINBA, respectively Fig. 3a).

Moreover, the proportion of peaks within 30 bp of a TATA-box center was twice as high as for any other method independently of the number of reported peaks (Fig. 3b), showing that the improvement was robust to the score threshold. We performed a similar benchmark (Methods) on the human chromosome 22 for 6 transcription factors of the ENCODE project[6] selected to be representative of accuracies in predicting ChIP-Seq peak positions from sequence motifs [20] (CEBPB, CTCF, MAX, USF1, PAX5, and YY1). On these data, GenoGAM performance was comparable to the other methods (95% boostrap confidence intervals, Fig. 3c for significant peaks, and Supplementary Fig. S2 for distance distributions, Supplementary Fig. S3 for all cutoffs). Hence, although GenoGAM is a general framework for ChIP-Seq analysis, it nonetheless provides a peak caller that is at least as performant as dedicated tools.

We next investigated the reason for the drastic differences observed in the yeast TFIIB dataset between GenoGAM and the other methods. The TATA-box region of *IDH2* illustrates the issue (Fig. 4a). The peaks reported by GenoGAM are positions with maximal a posteriori estimate of IP over input fold-changes. In contrast, MACS and JAMM report positions with maximal statistical significance[5,9]. Because statistical significance increases with both effect size and sample size, this leads to peak calls biased toward positions with high total counts in IP and input (Fig. 4a). Across all 644 TATA-box regions at which both GenoGAM and MACS identify a peak, total counts within 50 bp of peak positions were higher for MACS, but count ratios were higher for GenoGAM (Fig.

4b), generalizing the observations made for *IDH2*. The yeast TFIIB dataset was sequenced at a much higher coverage than the ENCODE dataset (0.9 unique fragments per base in average versus less than 0.03 unique fragments per base in average), leading to stronger discrepancies between significance and robust fold-changes. As sequencing depth is expected to increase in the near future, we anticipate that robust fold-change estimates as provided by GenoGAM will be a more sustainable criterion than mere significance for calling peak positions.

Higher sensitivity in testing for differential occupancy

To assess the performance of GenoGAM for calling differential occupancy, we re-analyzed histone H3 Lysine 4 trimethylation (H3K4me3) ChIP-Seq data of a study[21] comparing wild type yeast versus a mutant with a truncated form of Set1, the H3 Lysine 4 methylase. H3K4me3 is a hallmark of promoters of actively transcribed genes. Thornton and colleagues[21] have reported genome-wide redistribution in the truncated Set1 mutant of H3K4me3, which is depleted at the promoter and enriched in the gene body. We modeled this data with GenoGAM using one smooth function f_{WT} for the wild type reference occupancy, and one further smooth function $f_{\mathrm{mutant/WT}}$ for the differential effect. The offsets were computed to control for variations in sequencing depth between replicates and overall genome-wide H3K4me3 level (Methods). This yielded base-level log-ratio estimates and their 95% confidence bands genome-wide (Methods, Figure 5a for data and fit at the gene *YNL176C* consistent with the report of reduced binding at promoter regions).

As mentioned above, the confidence bands are Bayesian credible intervals. Previous studies based on simulated data showed that these confidence bands have close to nominal coverage probabilities and can, in practice, be used in place of frequentist confidence intervals [22]. We estimated base-level p-values using the point-wise estimates and standard deviations (Methods). To empirically verify that the p-values were at least conservative, we created a negative control dataset by per-base-pair independent permutation of the counts between the four samples. The offsets were set to 0 and the smoothing and dispersion parameters were estimated again. This non-parametric permutation scheme makes less assumptions than previous simulation studies[22]. Nonetheless, per-

base-pair p-values in this negative control experiment were slightly overestimated (Figure 5b). These results show that GenoGAM can be used to identify individual positions of significant differential occupancies with controlled type I error. Here, correction for multiple testing can either be done using the Benjamini and Hochberg procedure[23] or, procedures that exploit dependencies between adjacent positions[24].

Complementary to de novo identification, predefined regions, such as genes, can be tested for differential occupancies. To test for differences at any position in a region using GenoGAM, we propose to apply Hochberg's procedure to correct the pointwise p-values for multiple testing, and to report the smallest of these corrected p-values (Methods). To validate this approach, we compared GenoGAM against the following three approaches: csaw[12], which also tests for differences at any position in the regions, DESeq[10], which tests for differences in the overall occupancies, and MMdiff[11], which tests for differences of distribution within the regions but not overall occupancy (Methods). All investigated methods empirically controlled type I error on the permuted dataset at the 5% nominal level (Supplementary Fig. S4). On the original dataset, the least number of significant genes (FDR < 0.1) were identified by DEseq (735) and MMdiff (5). The csaw algorithm gave up to 863 significant genes but the number of identified genes depended strongly on the choice of the window size (Figure 5c). Of all methods, GenoGAM was the most sensitive, reporting 4,409 significant genes. Up to 861 genes, these genes were a superset of the genes reported by csaw, indicating that GenoGAM captured the same signal but with a higher sensitivity (Figure 5d). The genes reported only by GenoGAM showed a differential occupancy pattern similar yet weaker to the genes common to csaw and GenoGAM, with depletion in the promoter and enrichment in the gene body (Figure 5d), indicating that GenoGAM captured true biological signal.

Application to DNA methylation data

Generalized additive models are based on the generalized linear modeling framework and thus allow any distribution of the exponential family for the response. Therefore, GenoGAM can be also used to model continuous responses, for instance using the Gaussian distribution, and proportions

using the Binomial distribution. For ChIP-Seq data, a log-linear predictor-response relationship of the form (2) is justified by the fact that effects on the mean are typically multiplicative. However, other monotonic link functions could also be used. Moreover, quasi-likelihood approaches are supported, allowing for the specification of flexible mean-variance relationships[25].

To test the flexibility of GenoGAM, we conducted a proof-of-principle study on modeling bisulfite sequencing of bulk embryonic mouse stem cells grown in serum[26]. Bisulfite sequencing quantifies methylation rate by converting cytosine residues to uracil, leaving 5-methylcytosine residues unaffected. At each cytosine, the data consisted of the number n_i of fragments overlapping the cytosine and the number y_i of these fragments for which the cytosine was not converted to uracil. The quantity of interest was the methylation rate, i.e. the expectation of the ratio y_i/n_i. In the original publication, single nucleotide position methylation rates were estimated using a sliding window approach with an ad-hoc choice of window size of 3 kb computed in steps of 600 bp. Figure 6 reproduces an original figure showing the fit in a 120kb section of chromosome 6. We modeled this 120 kb section with GenoGAM using a quasi-binomial model, where the response was the number of successes y_i out of n_i trials, the log-odd ratio was modeled as a smooth function of the genomic position, and the variance was equal to a dispersion parameter times the variance of the binomial distribution. Smoothing and dispersion parameters were determined by cross-validation (Methods). The GenoGAM fit was consistent with the original publication[26], but did not rely on manually set window sizes and provided confidence bands (Figure 6). As expected, wider confidence bands were obtained in regions of sparse data and tighter bands in regions with a lot of data (Figure 6).

CONCLUSION

We have introduced a generic framework based on generalized additive models to model ChIP-Seq data. We have made this possible by providing a scalable algorithm that can fit GAMs to very long longitudinal data such as whole chromosomes at base-pair resolution. Scaling was made possible by parallelization over the data and allowing approximations rather than exact computation of the

fit [27].

Smoothing and dispersion parameters were obtained by cross-validation, i.e. they were fitted for the accuracy in predicting unseen data. This criterion turned out to provide useful values of smoothing and dispersion for inference, since we obtain signal peaks close to actual binding sites of transcription factors when these are known, at least as close as dedicated tools. Moreover, this criterion also led to reasonable uncertainty estimates since confidence bands of the fits were found to be only slightly conservative.

The utilization of genome-wide GAMs comes with a number of advantages: First, we flexibly model factorial designs, as well as replicates with different sequencing depths using size factors as offsets. Second, applying GAMs yields confidence bands as a measure of local uncertainty for the estimated rates. We showed how these can be the basis to compute point-wise and region-wise p-values. Third, GAMs outputs analytically differentiable smooth functions, allowing flexible downstream analysis. We showed how peak calling can be elegantly handled by making use of the first and second derivatives. Fourth, various link functions and distributions can be used, providing the possibility to model a wide range of genomic data beyond ChIP-Seq, as we illustrated with a first application on DNA methylation. In conclusion, we foresee GenoGAM as a generic method for the analysis of genome-wide assays.

METHODS

Preprocessing

Fragments were centered, reducing each fragment to one single data point. In case of single end data, the fragment length d was estimated using the Bioconductor package `chipseq` and its *coverage* method. It is defined as the optimal shift for which the number of bases covered by any read is minimized. Thus, the center was taken as the start of the read shifted by $\frac{d}{2}$ downstream.

Sequencing depth variations

Variations for sequencing depth was controlled by using size factors computed by DESeq2 [28] (version 2_1.10.0 here and after). This method robustly estimates fold-changes in overall sequencing depth by comparing read counts of predefined regions. The selection criteria for these regions was application-specific.

For peak calling applications, the selected regions were the 1,000 tiles with smallest p-value according to DESeq2 test for enrichment of IP over input performed on total read counts per tile. This allowed to select tiles that were most likely containing peaks. For differential binding application, all tiles were considered.

Model fitting

Model fitting given λ and θ

Each chromosome was partitioned into equally-sized intervals called chunks. Tiles were defined as chunks extended on either side by equally-sized overhangs. The generalized additive model was fitted on each tile separately using the *gam* function of the R package `mgcv`. Point estimates at each base pair of the smooth functions and their standard errors were extracted with the *predict* function on the fitted object setting "type" parameter to "iterms". The tile fits were then restricted to their chunk to define the chromosome-wide fit.

Fitting of λ and θ

The parameters λ and θ were the same for all tiles and were estimated using 10-fold cross-validation on a subset of all tiles. The selection of relevant tiles for cross-validation was application-specific as outlined in the respective sections below. To avoid overfitting due to short range cor-

relation, each cross-validation fold did not consist of randomly selected single base pairs but of short intervals. When replicate samples were available (that is all except for the TFIIB dataset), intervals could be of greater length as the model can predict samples from the respective replicates. We set it to twice the estimated fragment length. In the absence of replicates (TFIIB dataset), interval length was set to 20 bp (approximately a tenth of the fragment length.). For a given pair of values for λ and θ, the score function was defined as the sum of out-of-sample log-likelihood over all cross-validation folds and all tiles, restricted to the data points within chunks to not depend on poor fitting in overhangs. The parameters λ and θ were estimated by gradient-free numerical optimization of the score function using the Nelder-Mead algorithm (R function *optim*).

Yeast TFIIB dataset

ChIP-Seq library preparation, sequencing and read alignment

ChIP-Seq for TFIIB was performed essentially as described previously [29] with a few modifications. Briefly, 600 ml BY 4741 S. cerevisiae culture with C-terminally TAP-tagged TFIIB (Open Biosystems) was used. Immunoprecipitation was performed with 75 μl of IgG SepharoseTM 6 Fast Flow beads (GE Healthcare) for 3 hours at 4°C on a turning wheel. 30 μl of Input sample was taken before immunoprecipitation and stored at 4°C. IP and Input samples were reverse cross-linking for 2 hours with Proteinase K at 65°C and purified using Quiagen MinElute Kit. Samples were digested with 2.5 μl RNase A/T1 Mix (2 mg/ml RNase A, 5000 U/ml RNase T1; Fermentas) at 37°C for 1 h, purified and eluted in 50 μl H2O. ChIP-Seq libraries were prepared using NEB Next library preparation kit following manufacturer's instructions using the complete 50 μl as input. 2 μl of 1.7 μM adapters containing a GGAT barcode and 2 μl of a 0.25 μM adapter containing a CACT barcode were used for ligation with Input and IP samples, respectively. The final library was amplified for 22 cycles using Phusion Polymerase and purified using Agencourt Magnetic beads. 36 bp single end sequencing was performed on an Illumina GAIIX sequencer at the LAFUGA core facility of the Gene Center, Munich. Single-end 36 base reads and 4 base reads of barcodes were

obtained and processed using the Galaxy platform[30]. Reads were demultiplexed, quality-trimmed (Fastq Quality Filter), and mapped with Bowtie 0.12.7 [31] to the SacCer2 genome assembly (Bowtie options: -q -p 4 -S –sam-nohead -phred33-quals).

GenoGAM model

This dataset consisted of two samples: one input and one IP without replicates. Hence there was no need for an offset. We used the following GenoGAM model:

$$y_i \sim \text{NB}(\mu_i, \theta)$$
$$\log(\mu_i) = f_{\text{input}}(x_i) + f_{\text{protein}}(x_i) z_{j_i,\text{protein}},$$

where $z_{j_i,\text{protein}} = 1$ whenever j_i is the index of an IP sample and $z_{j_i,\text{protein}} = 0$ whenever j_i is the index of an input sample. Further parameter details are given in Supplementary Table S1.

TATA box mapping

Promoter TATA boxes where defined as instances of the motif TATAWAWR[19] at most 200 bp 5' and 50 bp 3' of one of the 7,272 transcript 5'-ends reported by Xu et al.[32].

ENCODE transcription factors

Data processing

Alignment files (BAM files, aligned for the human genome assembly hg19) for ChIP-Seq data for the transcription factors CEBPB, CTCF, MAX, USF1, PAX5, and YY1 were obtained from the ENCODE website `www.encodeproject.org`. All these datasets contained two biological replicates for the protein samples and at least one input sample. However, the library sizes of the

input samples were so small that including them resulted in higher uncertainty about the peaks, for our approach and for alternative approaches. We therefore conducted the analyses without correction for input.

GenoGAM model

For each transcription factor the dataset was modeled separately. Each one consisted of IPs with replicates. The following GenoGAM model was used:

$$y_i \sim \text{NB}(\mu_i, \theta)$$
$$\log(\mu_i) = \log(s_{j_i}) + f_{\text{protein}}(x_i),$$

where the offsets $\log(s_{j_i})$ are log-size factors computed to control for sequencing depth variation between the replicates (see section). Further parameter details are given in Supplementary Table S1.

Transcription factor motif mapping

Motif occurrences in the genome were determined by FIMO[33] using default threshold 10^{-4} with position weight matrices (PWMs) from the JASPAR 2014 database[34] with the following IDs: CEBPB: MA0466.1, CTCF: MA0139.1, MAX: MA0058.1, PAX5: MA0014.2, USF1: MA0093.2, YY1: MA0095.2

Peak calling

GenoGAM-based peak calling and z-score

Values of first and second derivatives of fitted smooth functions were obtained by multiplying the estimated coefficients with the corresponding derivatives of the B-splines as obtained from the *spline.des* function of the R package `splineDesign`. Local extrema (at base pair resolution) were identified as positions at which the sign of the first derivative differed from the one of the preceding position. For the z-score (equation 4), μ_0 is the global background mean and σ_0^2 is the global background variance of $f(x)$. In order to account only for the background without potential peaks, μ_0 was estimated as the *shorth* from the Bioconductor `genefilter` package for all $f(x_i)$, $i = 1, \ldots, n$ (midpoint of the shortest interval containing half of the data) of all fitted values. The fitted values smaller than the shorth were mirrored on it, such that a symmetric density was created that excludes the values larger than the shorth, in particular those high values representing peaks. The variance of this newly created distribution was then estimated in a robust fashion by the *median absolute deviation* (MAD) giving σ_0^2 (Supplementary Figure S1).

False Discovery Rate for GenoGAM peaks

To estimate false discovery rates (FDR), peaks were called on $-f_{\text{protein}}$. Their z-scores were obtained by recomputing μ_0 and σ_0 and applying the same formula. The FDR for a given minimum z-score z was estimated by $\frac{|V_z|}{|P_z|+|V_z|}$, where P_z and V_z are the sets of peaks and valleys, respectively, with a z-score greater than or equal to z.

MACS, JAMM, and ZINBA

The version 2 of the MACS software, MACS2, was run with the default parameters and the additional flag *call-summits*. In case of TFIIB, the *nomodel* parameter was used to avoid building the shifting model. This was necessary since the default values for *mfold* were too high and resulted in worse performance if reduced, compared to absence of a model.

JAMM was run with default values and peak calling mode (*-m*) set to narrow assuming a three component mixture model for background, enriched regions and tails of enriched regions. Although JAMM computes a score to rank peaks it does not provide a method to define a threshold for a given FDR or significance. Nevertheless, JAMM applies some filtering on the complete list of peaks to output a filtered list. Instead of using this filtered output directly, we used the complete, sorted (by score) peak list and took the top N results where N is the number of peaks in the filtered output. This improved the performance of JAMM in some cases (and left unchanged in others). For analysis, where a cutoff for JAMM was still needed we used the same number of peaks that MACS reported.

For ZINBA, the mappability score was generated (*generateAlignability*) with the mappability files for 36 bp reads, taken from the ZINBA website `https://code.google.com/p/zinba/`. The average fragment length (*extension*) was specified at 190 bp, window size (*winSize*) at 250 and offset (*offset*) at 125. The FDR threshold was set to 0.1 and window gap to 0. Peaks were refined (default) and model selection was activated. The complete model was used (*selecttype = "complete"*), input was included as a covariate (*selectcovs = "input_count"*) and interactions were allowed. The chromosome used to build the model was selected randomly to be "chrXVI" (*selectchr*). The parameter "method" was set *method = "mixture"*.

Differential binding

Data processing

Raw sequencing files (H3K4ME3_Full_length_Set1_Rep_1.fastq, H3K4ME3_Full_length_Set1_Rep_2.fastq, H3K4ME3_aa762-1080_Set1_Rep_1.fastq, and H3K4ME3_aa762-1080_Set1_Rep_2.fastq) were obtained from the Sequence Read Archive (SRA) repository (`http://www.ncbi.nlm.nih.gov/sra`). These were paired-end reads. Reads were aligned to the SacCer3 build of the *S. cerevisiae* genome with the STAR aligner[35] (version 2.4.0, default parameters). Reads with ambiguous mapping were removed using samtools[36] (version 1.2 option *-q 255*). Gene boundaries were obtained from the *S. cerevisiae* genome annotation R64.1.1, restricting gff file entries of type "gene".

GenoGAM model

This dataset consisted of four samples: two biological replicate IPs for the wild type strain, and two biological replicate IPs for the mutant strain. We used the following GenoGAM model:

$$y_i \sim \text{NB}(\mu_i, \theta)$$
$$\log(\mu_i) = \log(s_{j_i}) + f_{\text{WT}}(x_i) + f_{\text{mutant/WT}}(x_i) z_{j_i,\text{mutant}}$$

where $z_{j_i,\text{mutant}} = 1$ for j index of one mutant sample and 0 for wild-type samples. The offsets $\log(s_{j_i})$ are log-size factors computed to control for sequencing depth variation and overall H3K4me3 across all four samples (see section). Further parameter details are given in Supplementary Table S1.

Position-level significance testing

Null hypotheses of the form $H_0 : f_k(x) = 0$ for a smooth function $f_k()$ at a given position x of interest were tested assuming approximate normal distribution of the corresponding z-score, i.e.:

$$T_k(x) \sim N(0,1)$$

where

$$T_k(x) = \frac{\hat{f}_k(x)}{\hat{\sigma}^2_{f_k(x)}}$$

where $\hat{f}_k(x)$ and $\hat{\sigma}^2_{f_k(x)}$ denote point estimate and standard error of the smoothed value [16] as returned by the function *predict(..., type="iterms", se.fit=TRUE)* of the R package `mgcv`.

False discovery rate for predefined regions

Let $R_1, ..., R_p$ be p regions of interest, where a region is defined as a set of genomic positions. Regions are typically but not necessarily, intervals (e.g. genes or promoters). For instance, all exons of a gene could make up a single region. Regions can be a priori defined or defined on the data using independent filtering [37]. For instance, when testing for significant differences between two conditions, regions can be selected for having a large total number of reads over the two conditions [12].

For j in $1, .., p$, let H_0^j be the composite null hypothesis that the smooth function f_k values 0 at every position of the region R_j:

$$H_0^j = \bigwedge_{x_i \in R_j} (f_k(x_i) = 0)$$

The False Discovery Rate was controlled using the following procedure [12]:

1. Position-level p-values at all region positions were computed using position-level significant testing as described above.

2. Within each region R_j, position-level p-values were corrected for multiple testing using Hochberg family-wise error rate correction [38]. The Hochberg correction was applied because position-level p-values of one smooth function are positively associated. The p-value for the null hypothesis H_0^j was then computed as the minimal family-wise error rate corrected position-level p-value. This step gives one p-value per region.

3. FDRs were computed using Benjamini-Hochberg procedure [23] applied to the region-level p-values.

Benchmarking

The R/Bioconductor packages `DESeq2`[28], `MMDiff`[11], and `csaw`[12] were applied on original count data and on the base-level permuted dataset, for all genes. The log-size factors were set 0 for all methods when applied to the permuted datasets. DESeq2 was applied with default parameters. MMDiff was applied with a bin length of 50 bp, the DESeq method for the normalization factor, and the Maximum Mean Discrepancy (MMD) histogram distance. The csaw method was applied with window size of 150 bp and otherwise default parameters. The window size was determined through a grid search (see Figure 5c), choosing the window size with the most significant genes. In particular, csaw uses a different procedure to estimate normalization factors than DESeq and MMDiff. We used the default one as it was in favor of csaw for returning more significant genes.

Methylation data

Data processing

We obtained the data in text table format from Smallwood et al.[26] from the Gene Expression Omnibus (GEO) repository `http://www.ncbi.nlm.nih.gov/gds`. The data provided, was one record per CpG site, with the number of methylated and unmethylated fragments at the respective site. We used a Python script to bin this data into bins of 3,000 bp width every 600 bp, as was done in the original paper.

GenoGAM model

To model y_i, the number of reads of methylated state, out of n_i, the total number of reads, we used the quasi-binomial model defined by:

$$\begin{aligned} \mathrm{E}(y_i/n_i) &= \mu_i \\ \log(\frac{\mu_i}{1-\mu_i}) &= f_{\text{methylation}}(x_i) \\ \mathrm{Var}(y_i/n_i) &= \theta \cdot \frac{\mu_i(1-\mu_i)}{n_i}, \end{aligned}$$

where the scale parameter $\theta > 0$ models dispersion. The model was applied on only one tile with a width of 120 kb, reproducing Figure 2a of Smallwood et al.[26]. Further parameter details are given in Supplementary Table S1.

Accession code

ChIP-Seq data are available at Array Express under the accession number E-MTAB-4175. For review, user: Reviewer_E-MTAB-4175, password: NF3Qvgio

Code availability

Scripts used for this study are provided in Additional data file 2. A R package called GenoGAM has been submitted to Bioconductor[39]. Refer to the Bioconductor web page at `http://www.bioconductor.org` for installation procedures.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceived the project and supervised the work: JG AT. Developed the software and carried out the analysis: GS AE JG Carried out the ChIP-Seq experiments for TFIIB on yeast: DS. Gave advice on statistics: MS. Wrote the manuscript: JG AE GS MS AT

Acknowledgements

We thank Ulrich Unnerstall and Michael Lidschreiber for fruitful discussions on data analysis, Martin Morgan and Hervé Pagès for support during the implementation of the GenoGAM package, Stefan Krebs for sequencing and raw data preprocessing, and Ulrich Mansmann and Patrick Cramer for institutional support. JG was supported by the Bavarian Research Center for Molecular Biosystems and by the Bundesministerium für Bildung und Forschung through the Juniorverbund in der Systemmedizin 'mitOmics' (FKZ 01ZX1405A). We acknowledge support from the European Commission through the Horizon 2020 project SOUND (GS, JG) and from the Graduate School for Quantitative Biosciences Munich (AE).

REFERENCES

1. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**, 651–657 (2007).

2. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)* **316**, 1497–502 (2007). URL `http://www.sciencemag.org/content/316/5830/1497.long`.

3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–37 (2007). URL `http://www.sciencedirect.com/science/article/pii/S0092867407006009`.

4. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572–6 (2007). URL `http://dx.doi.org/10.1038/nature05632`.

5. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

6. Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

7. Yan, H. *et al.* HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. *BMC bioinformatics* **15**, 280 (2014).

8. Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W. & Lieb, J. D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology* **12**, R67 (2011).

9. Ibrahim, M. M., Lacadie, S. A. & Ohler, U. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics (Oxford, England)* **31**, 48–55 (2015).

10. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106–R106 (2010). URL `http://genomebiology.com/2010/11/10/R106papers2://publication/doi/10.1186/gb-2010-11-10-r106`.

11. Schweikert, G., Cseke, B., Clouaire, T., Bird, A. & Sanguinetti, G. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC genomics* **14**, 826 (2013).

12. Lun, A. T. & Smyth, G. K. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic acids research* **42**, e95–e95 (2014).

13. Hastie, T., Tibshirani, R. *et al.* Generalized additive models. *Statistical science* **1**, 297–310 (1986).

14. De Boor, C. *A practical guide to splines*, vol. 27 (Springer-Verlag New York, 1978).

15. Eilers, P. H. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Statistical science* 89–102 (1996).

16. Wood, S. *Generalized additive models: an introduction with R* (CRC press, 2006).

17. Nelder, J. A. & Mead, R. A simplex method for function minimization. *The computer journal* **7**, 308–313 (1965).

18. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S. & Stoica, I. Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, vol. 10, 10 (2010).

19. Basehoar, A. D., Zanton, S. J. & Pugh, B. F. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**, 699–709 (2004).

20. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* (2015).

21. Thornton, J. *et al.* Context dependency of Set1/ COMPASS-mediated histone H3 Lys4 trimethylation. *Genes & Development* **28**, 115–120 (2014).

22. Marra, G. & Wood, S. N. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* **39**, 53–74 (2012).

23. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300 (1995).

24. Wei, Z., Sun, W., Wang, K. & Hakonarson, H. Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics (Oxford, England)* **25**, 2802–2808 (2009). URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp476.

25. Wedderburn, R. W. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika* **61**, 439–447 (1974).

26. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods* **11**, 817–820 (2014).

27. Heinis, T. Data analysis: approximation aids handling of big data. *Nature* **515**, 198 (2014). URL http://dx.doi.org/10.1038/515198d.

28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

29. Schulz, D. *et al.* Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell* **155**, 1075–1087 (2013).

30. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**, R86 (2010). URL http://genomebiology.com/2010/11/8/R86.

31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009). URL http://genomebiology.com/2009/10/3/R25.

32. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* (2009). URL http://www.nature.com/nature/journal/vaop/

565 ncurrent/full/nature07728.htmlpapers2://publication/doi/doi:
566 10.1038/nature07728.

567 33. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.
568 *Bioinformatics (Oxford, England)* **27**, 1017–8 (2011). URL http://bioinformatics.
569 oxfordjournals.org/content/27/7/1017.

570 34. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access
571 database of transcription factor binding profiles. *Nucleic acids research* **42**, D142–7
572 (2014). URL http://nar.oxfordjournals.org/content/early/2013/11/
573 04/nar.gkt997.full.

574 35. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, Eng-*
575 *land)* **29**, 15–21 (2013). URL http://bioinformatics.oxfordjournals.org/
576 content/29/1/15.long.

577 36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford,*
578 *England)* **25**, 2078–9 (2009). URL http://bioinformatics.oxfordjournals.
579 org/content/25/16/2078.long.

580 37. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for
581 high-throughput experiments. *Proceedings of the National Academy of Sciences* **107**, 9546–
582 9551 (2010).

583 38. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*
584 **75**, 800–802 (1988).

585 39. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology
586 and bioinformatics. *Genome biology* **5**, R80 (2004). URL http://genomebiology.
587 biomedcentral.com/articles/10.1186/gb-2004-5-10-r80.

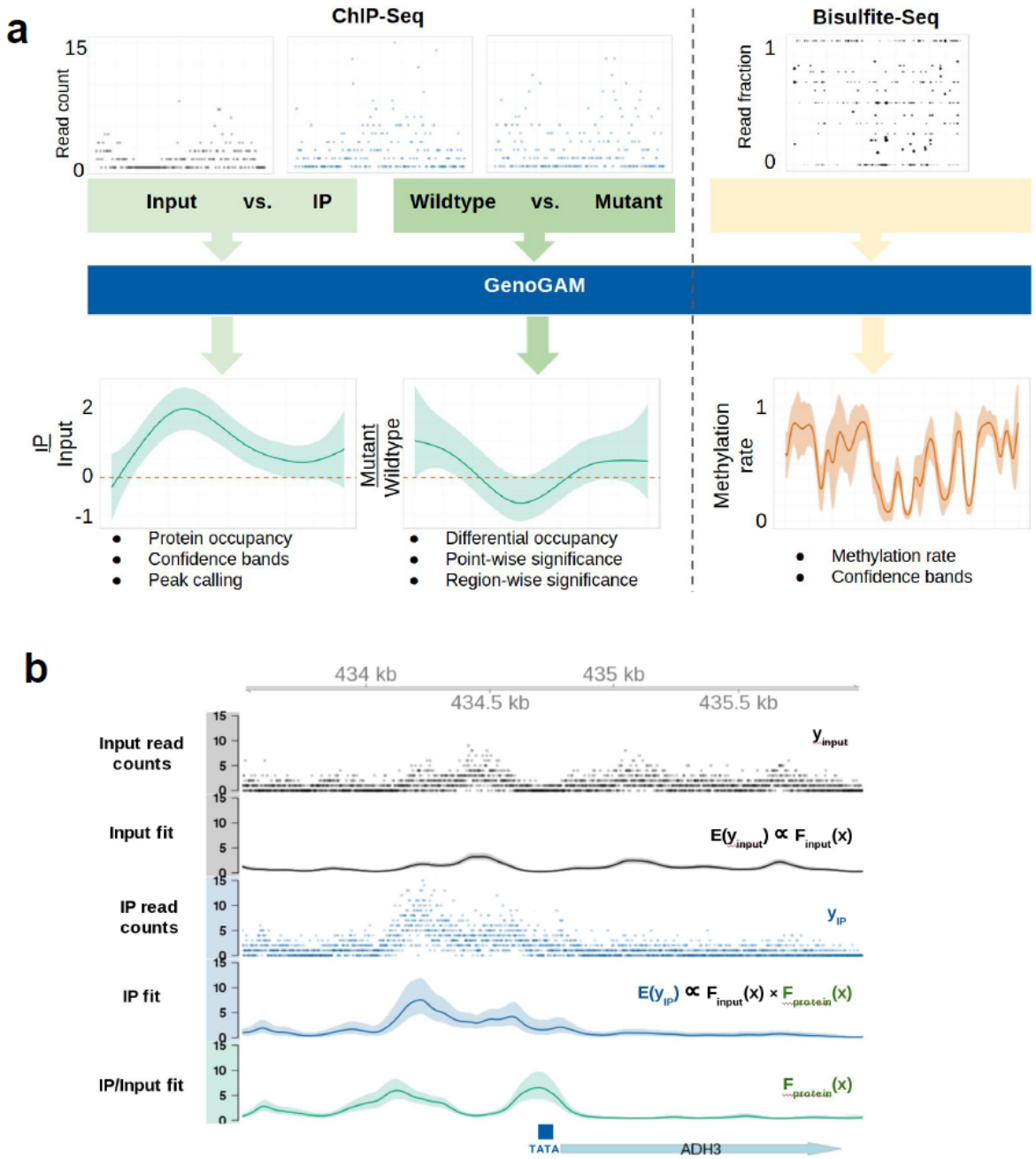

Figure 1: **GenoGAM applications and concept** (**a**) GenoGAM provides a general framework to analyze ChIP-Seq data for both absolute (left arrow) and differential protein (center arrow) occupancy. It can also be applied to infer DNA methylation rate from bisulfite sequencing data (right arrow). (**b**) ChIP-Seq analysis with GenoGAM yields base-pair resolution occupancy profiles with confidence bands. Input (black) and IP (blue) centered read counts (dots) and fitted smooth (solid line) with 95% confidence intervals (ribbons) for the transcription factor TFIIB for a section of the chromosome XIII of *S. cerevisiae*. Additionally, the extracted fold change of IP over Input (green) and gene annotation at the very bottom. Simplified equations depict model constituents.

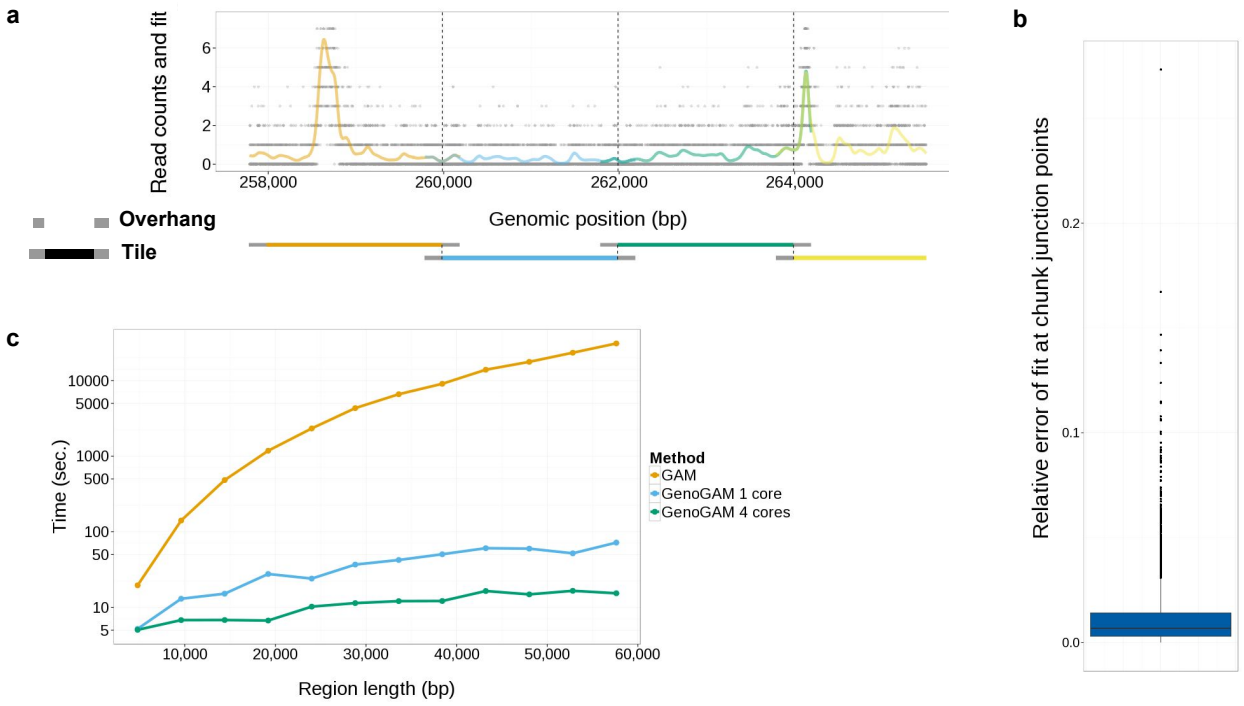

Figure 2: **Per tile-parallelization allows map-reduce implementation of GAM.** (**a**) Read count (black dots, capped at $>= 7$) and predicted rates (orange, blue, green, and yellow transparent lines) for four successive tiles (lower track). Vertical dashed lines denote the junction points (**b**) Distribution of the relative error (difference over mean) at the junction point of two neighbor tiles, for an overhang of 8 basis functions. (**c**) Computing time in seconds (y-axis in log scale) versus region length in bp (x-axis) for a standard GAM (orange), GenoGAM on a single core (blue), and GenoGAM on four cores (green). Tiles were 2,400 bp long and contained 100 basis functions each.

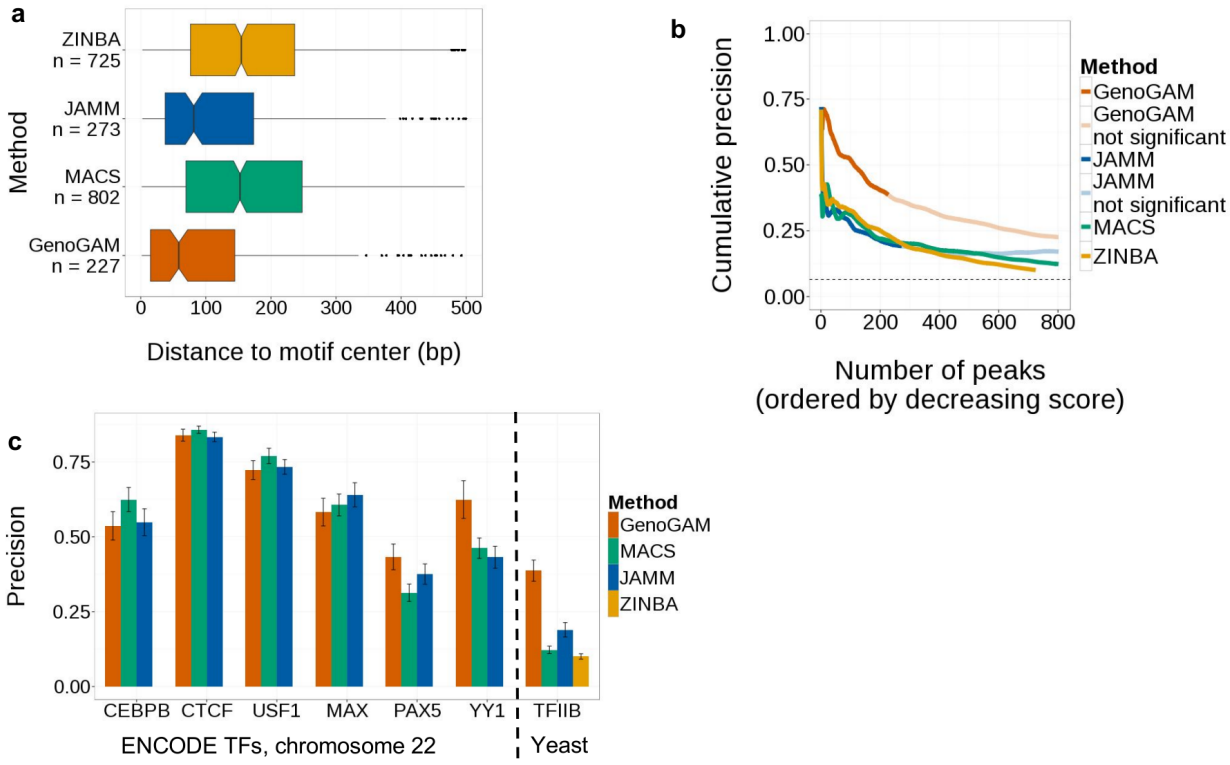

Figure 3: **GenoGAM identifies protein binding sites with similar accuracy to state-of-the-art peak callers.** (**a**) Boxplot of distances between significant peaks (FDR< 0.1, Methods) and TATA box for the yeast TFIIB dataset (Methods) for GenoGAM (orange), MACS (green) and JAMM (blue) and ZINBA (yellow). (**b**) Proportion of TFIIB peaks (y-axis) within 30 bp of a TATA box for GenoGAM (orange), MACS (green), JAMM (blue) and ZINBA (yellow) versus number of selected peaks when ordered by decreasing score (x-axis). For each method transparent colors indicate peaks that the method considers not significant (FDR> 0.1, Methods). (**c**) Proportion of significant peaks within 30 bp of motif center and 95% bootstrap confidence interval (error bars, Methods) for all six ENCODE transcription factors (CEBPB, CTCF, USF1, MAX, PAX5, YY1) on chromosome 22 and for the yeast TFIIB dataset.

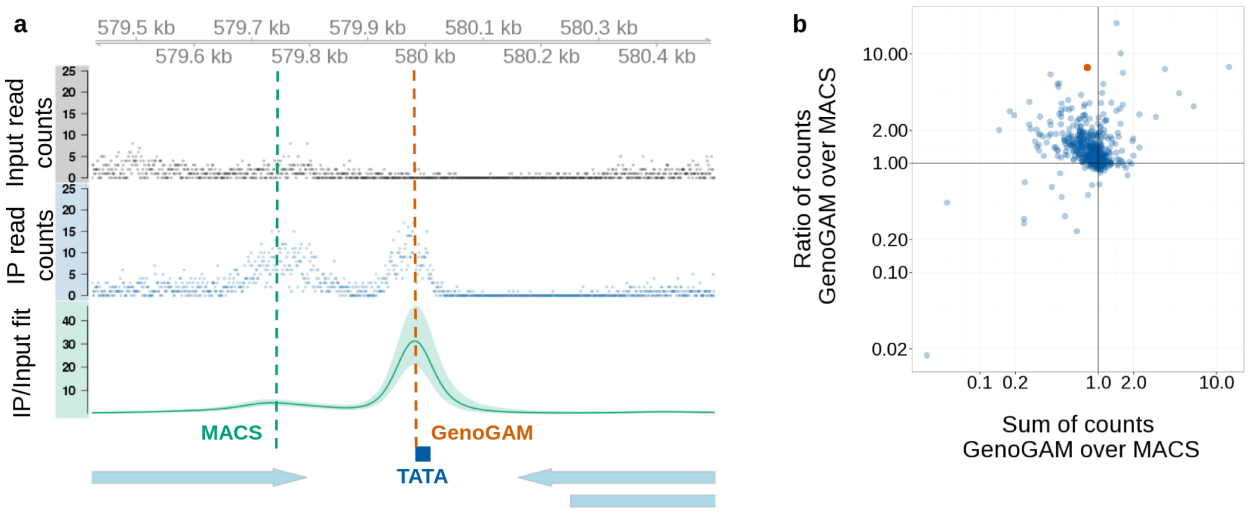

Figure 4: **Peak position represent maximal fold change rather then maximal significance.** (**a**) Example region in yeast with input (black dots), IP (blue dots) and the smooth function IP over input with 95% confidence interval (green line) showing a correctly identified peak by GenoGAM (orange vertical dashed line) and an incorrect identified by MACS (green vertical dashed line), due to enrichment in input. (**b**) Scatterplot of the sum of counts (input + IP) vs ratio of counts (input/IP) for GenoGAM divided by MACS on all mutually called TATA box positions. The red dot denotes the example region shown in (a)

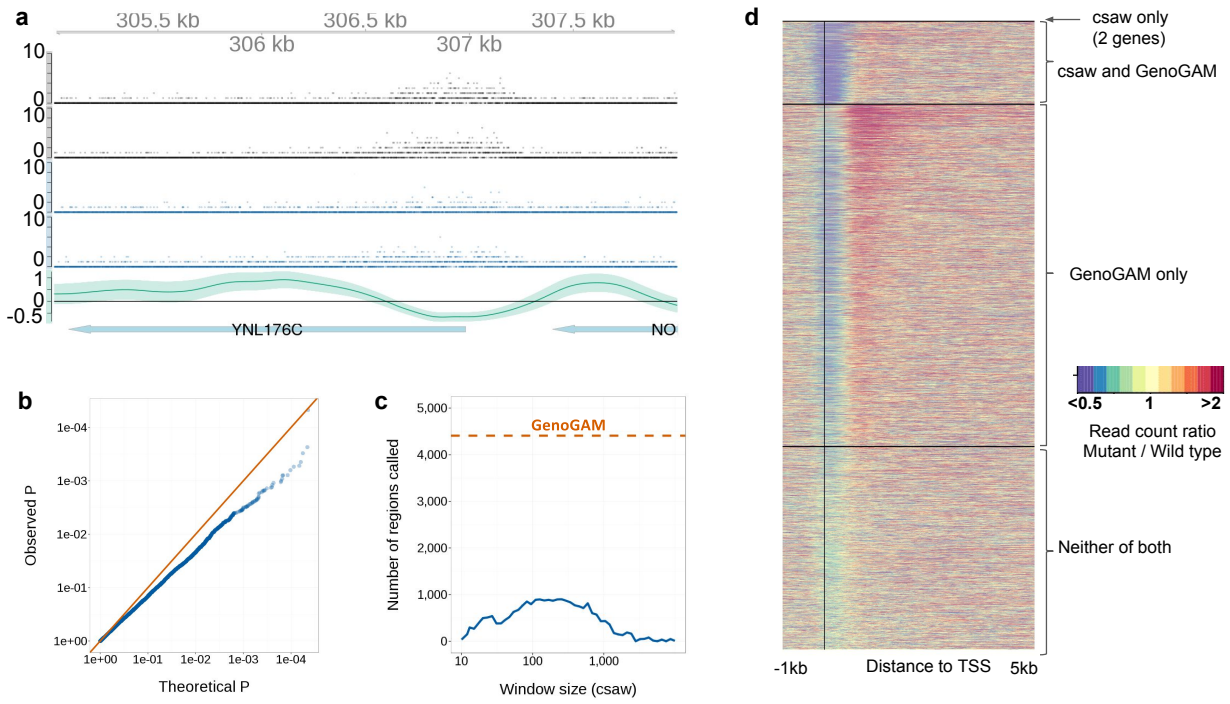

Figure 5: **Statistical testing for factorial designs.** (**a**) Read counts (dots) and fitted rates with 95% confidence bands for wild-type (black) and mutant (blue) and the log-ratio of mutant over wild-type with confidence band (bottom row, green) around *YNL176C*. (**b**) Empirical (y-axis) versus theoretical (x-axis) p-values in base-level permuted count data (Methods). P-values at every 200 bp positions are shown. (**c**) Number of genes with significant differential occupancies in mutant over wild type (FDR<0.1) reported by GenoGAM (orange) and by csaw (blue) as function of window size (x-axis). (**d**) Fold-change of counts in mutant over wild-type in 150 bp windows for all 6607 yeast genes in the -1 to 5 kb region centered on TSS (vertical black line). The genes are sorted into four groups (separated by the black horizontal lines) according to which method reports them significant. From top to bottom: csaw only (2 genes), csaw and GenoGAM (861 genes), GenoGAM only (3,548 genes) and none (2,196 genes). Within each group genes are ordered by p-values (lowest to highest from top to bottom). The "csaw and GenoGAM" group is sorted by GenoGAM p-values.

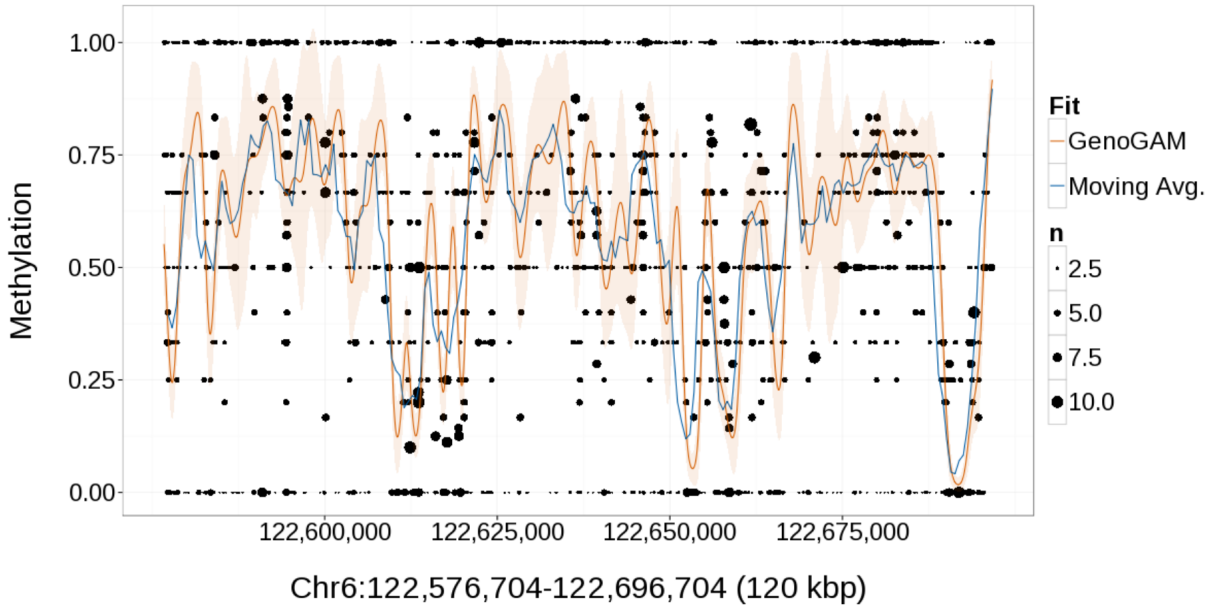

Figure 6: **Application to DNA methylation data.** Estimated DNA methylation rates in a 120 kb region of chromosome 6 of the mouse (cf. Smallwood et al.[26]). Shown are the data for bulk embryonic mouse stem cells grown in serum; ratios of methylated counts for each CpG position (black dots), with point size proportional to the number of reads. The estimated rates are shown for the moving average approach[26] of 3,000 bp bins in 600 bp steps (blue line) and for the GenoGAM (orange line) with 95% confidence band (ribbon).