mathematics

MDPI

Article

Analysing the Protein-DNA Binding Sites in *Arabidopsis thaliana* from ChIP-seq Experiments

Ginés Almagro-Hernández [1,2,*], Juana-María Vivo [2,3,*], Manuel Franco [2,3,*] and Jesualdo Tomás Fernández-Breis [1,2,*]

[1] Departamento de Informática y Sistemas, Universidad de Murcia, CEIR Campus Mare Nostrum, 30100 Murcia, Spain
[2] Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), 30120 Murcia, Spain
[3] Departamento de Estadística e Investigación Operativa, Universidad de Murcia, CEIR Campus Mare Nostrum, 30100 Murcia, Spain
* Correspondence: gines.almagro@um.es (G.A.-H); jmvivomo@um.es (J.-M.V.); mfranco@um.es (M.F.); jfernand@um.es (J.T.F.-B.)

Abstract: Computational genomics aim at supporting the discovery of how the functionality of the genome of the organism under study is affected both by its own sequence and structure, and by the network of interaction between this genome and different biological or physical factors. In this work, we focus on the analysis of ChIP-seq data, for which many methods have been proposed in the recent years. However, to the best of our knowledge, those methods lack an appropriate mathematical formalism. We have developed a method based on multivariate models for the analysis of the set of peaks obtained from a ChIP-seq experiment. This method can be used to characterize an individual experiment and to compare different experiments regardless of where and when they were conducted. The method is based on a multivariate hypergeometric distribution, which fits the complexity of the biological data and is better suited to deal with the uncertainty generated in this type of experiments than the dichotomous models used by the state of the art methods. We have validated this method with *Arabidopsis thaliana* datasets obtained from the Remap2020 database, obtaining results in accordance with the original study of these samples. Our work shows a novel way for analyzing ChIP-seq data.

Keywords: bioinformatics; computational genomics; ChIP-seq experiment; protein binding functional regions; multivariate hypergeometric distribution

check for updates

Citation: Almagro-Hernández, G.; Vivo, J.-M.; Franco, M.; Fernández-Breis, J.T. Analysing the Protein-DNA Binding Sites in *Arabidopsis thaliana* from ChIP-seq Experiments. *Mathematics* **2021**, *9*, 3239. https://doi.org/10.3390/math9243239

Academic Editor: Vasile Preda

Received: 4 November 2021
Accepted: 9 December 2021
Published: 14 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computational genomics consists of the use of a wide range of mathematical tools, implemented in specific software, in order to solve challenges such as how the functionality of the genome of the organism under study is affected both by its own sequence and structure, and by the network of interaction between this genome and different biological or physical factors (proteins, metabolites, molecular complexes, electromagnetic radiation, etc.).

One of the main types of experiments included in this field is the so-called chromatin immunoprecipitation (ChIP) experiment [1], which aims to identify and localize in vivo all the binding sites of a given DNA-binding protein throughout the genome of an organism, tissue, or cell line subjected to a specific biological condition (e.g. "wild type" or "stress"). Subsequent bioinformatics analysis of the results of this experiment will be carried out to elucidate the biological implications of the binding protein on the organism under study [2]. ChIP-seq experiments [3] consist of a first ChIP phase in which the immunoprecipitated fragments of the DNA molecule (with a length of between 150 and 1000 nucleotides) to which the protein under study has been attached (hereafter referred to as target protein) are enriched over the immunoprecipitated fragments corresponding to the rest of the genome. This is followed by a phase of identification of these fragments in two steps.

The first one is their initial segmentation into subfragments with a length between 30 and 150 nucleotides and subsequent sequencing of their nucleotide chain (seq) through next generation sequencing (NGS) techniques, which, by means of "base calling" algorithms, obtains the so-called "reads" (sequenced subfragments). In a second step, the reads obtained are mapped onto the reference genome of the organism under study, i.e., the identification of the site(s) where the read is statistically presumed to belong, including both the regions to which the target protein binds and those of the rest of the genome. Once the reads obtained from the experiment have been mapped, "peak calling" algorithms apply statistical enrichment analysis to identify the binding sites of the protein in the genome, which are called "peaks", referring to the shape of the distribution of mapped reads on these regions. Thus, peaks are the enriched regions of mapped reads, and they are considered as the hosts of possible binding sites.

The following are the main challenges facing methods or algorithms used in computational genomics in general and among them those focused on ChIP-seq experiments: (i) algorithms have to process a huge amount of data; (ii) the data have a degree of uncertainty associated [4,5], which is generated and accumulated by the occurrence of technical errors, biases in the algorithms used, the nature of the data being worked with and other unknown and therefore uncontrollable factors; (iii) the inherent complexity of the genomic and biochemical information represented by an extensive network of interacting nodes to carry out complex biological processes within a cell, tissue, or organism.

One of the most important stages is elucidating the biological implications of the peaks obtained in a ChIP-seq experiment, which has the following specific challenges: (i) the noise accompanying each peak, where the length of the peak is on average about 20 times longer than the actual site where the protein under study binds; (ii) obtaining peaks that do not actually harbor any binding site for the target protein; (iii) to establish correct relationships between the peak obtained and the functional elements of the genome. The state-of-the-art methods use heuristic methods in some cases, statistical methods in others, and both at the same time. Although there are two common features of all the tools developed so far, one of them consists of modeling and treating each peak obtained [6–9] and each functional element [10] belonging to the genome under study individually, while the other consists of working with dichotomous models, where models are created containing regions of different extension, with two possible values, and which in most cases lack a mathematical formalism in the generation of the models used.

The main objective of this work consists of the design, implementation and validation of an analysis method to overcome the two main challenges seen above. Our method treats all the peaks of a ChIP-seq experiment as if they were the result of a random experiment where a multivariate hypergeometric model can be defined. This method reflects the behavior of the target protein, taking into account each and every one of the peaks generated along the entire genome where they have been mapped, as well as all the possible binding sites that this genome harbors.

Our method addresses the two challenges as follows: (i) The uncertainty that is generated at each stage throughout the whole process is handled by applying statistical techniques and applying a standardized methodology, in which each step has a precise, complete definition and setting of the parameters; (ii) the complex nature of the biological data to be interpreted is dealt with through the use of multivariate probabilistic or background models that accurately represent the entire sample space being analyzed; in this way, adjusting the dimensionality of each model (statistical variables) to this complexity. Thereby, the proposed method permits to extract precise information available from the set of peaks obtained from a ChIP-seq experiment, as well as to compare the results obtained from different experiments. The characteristics of this method are: (i) flexibility, allowing the analysis of ChIP-seq experiments for any type of organism, tissue, or cell line, as long as there is sufficient information on the configuration and structure of its genome and its component genes; (ii) extensibility, it allows for the incorporation of new information that is generated over time, thus improving the method; (iii) mathematical formalization, imple-

mented both in the generation of the probabilistic models and in the statistical techniques used on them.

Finally, we have validated the method using information from public repositories, both to obtain the datasets of peaks from ChIP-seq experiments and the information necessary for the modeling of the genome of the organism under study, which in this case is *Arabidopsis thaliana*, a well-characterized plant widely used as an experimental model. The results obtained show the capacity of our method to analyze these types of experiments from a point of view that has not been considered until now.

2. Methods

2.1. General Overview

Our basic assumptions are that the results of a ChIP-seq experiment (set of peaks) are a random vector following a multivariate hypergeometric distribution [11–13], and we can model the genome according to the expected characteristics of this type of random experiment.

Let S be a finite population formed by m elements which are classified into k mutually exclusive classes, i.e., each element belongs to one and only one of the k classes. Let S_i be the subpopulation of all the elements of the ith class, being m_i its subpopulation size $(i = 1, 2, \ldots, k)$ and $m = \sum_{i=1}^{k} m_i$. Then, the random experiment consisting in drawing without replacement n elements of S is represented by the random vector $\mathbf{X} = (X_1, \ldots, X_k)$, where each X_i denotes the number of elements of the S_i class in the sample. The random vector $\mathbf{X}$ follows a multivariate hypergeometric distribution with parameters n, $(m_1, \ldots, m_k)$ and $m = \sum_{i=1}^{k} m_i$,

$$\mathbf{X} \sim MH(n, (m_1, \ldots, m_k), m),$$

whose joint probability mass function is given by

$$P_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{1}{\binom{m}{n}} \prod_{i=1}^{k} \binom{m_i}{x_i} \tag{1}$$

where $0 \leq x_i \leq m_i$ and $\sum_{i=1}^{k} x_i = n$.

From (1), the ith component of $\mathbf{X}$ has a univariate hypergeometric distribution with parameters (n, m_i, m), $X_i \sim H(n, m_i, m)$ for $i = 1, 2, \ldots, k$. In reality there are only $k - 1$ distinct marginal random variables, since $\sum_{i=1}^{k} X_i = n$ [14], and for $j \neq i$ with $j = 1, 2, \ldots, k$, their means, variances, and covariances are

$$E(X_i) = n\frac{m_i}{m}, \quad Var(X_i) = n\frac{m_i}{m}\left(1 - \frac{m_i}{m}\right)\frac{m - n}{m - 1}, \tag{2}$$

and

$$Cov(X_i, X_j) = -n\frac{m_i m_j}{m^2}\frac{m - n}{m - 1}.$$

In our case, S is the finite population of all individual binding sites with a given length, that the genome can house. Each individual binding site must satisfy two requirements: (i) to belong only to one of the S_i classes of k functional regions (promoters, upstream, exons, ...) assigned, being also named the *annotated binding sites*; (ii) to have the same probability of being chosen. Finally, E is the set of n peaks obtained from a certain ChIP-seq experiment. To convert an element of E (peak) into an element of S (annotated binding site) with its corresponding assigned functional region S_i, its length has to be adjusted to the one given for the elements of S in a process called *peak standardization* that will be seen below. The finite population S is represented by the **background model**, which is derived from the **genome model**. Otherwise, once processed, the set E is represented by the ChIP-seq model.

The generation of the necessary models for data processing consists of two parallel and independent tracks (Figure 1). On the one hand, the **genome model** (yellow) contains the necessary information, collected from available resources regarding the structure and organization of (i) both the genome and the genes to be considered, and (ii) the tissue or organism associated with the experiment whose results are going to be analyzed.

This information is transformed from a series of parameters for subsequent use in the annotation process, generating the **background model** (blue). This model is composed of the annotated binding sites, which have been assigned a class of the functional regions that make up the **genome model**. On the other hand, a peak standardization process (Appendix A) is carried out on the set of peaks obtained from a given ChIP-seq experiment (ChIP-seq dataset). The peaks are standardized with a given length and are grouped in the **ChIP-seq model**. This length must be equal to the length configured in the **background model**. In the annotation process, each standardized peak is assigned a single class of functional region according to the annotated binding site with which it overlaps, becoming known as the annotated peaks. The set of all annotated peaks obtained from a specific **ChIP-seq model** according to a specific background model is called **ChIP-seq annotated**.

Finally, we obtain the **characteristic profile**, which includes information on the frequencies, expected values, standard deviations and Z-scores, for this annotated ChIP-seq. The **characteristic profile** depends on the count of the annotated peaks obtained according to the type of functional region class with which it has been annotated and the background model used in this annotation process. Further analyses are based on this information: (1) a goodness-of-fit test based on the multivariate hypergeometric distribution determines whether the distribution of the annotated peaks associated with the sample can be considered to belong to the population represented by the background model used in the annotation process; (ii) a test of homogeneity based on the multivariate hypergeometric model determines whether the distributions of two **ChIP-seq annotated** follow the same model.

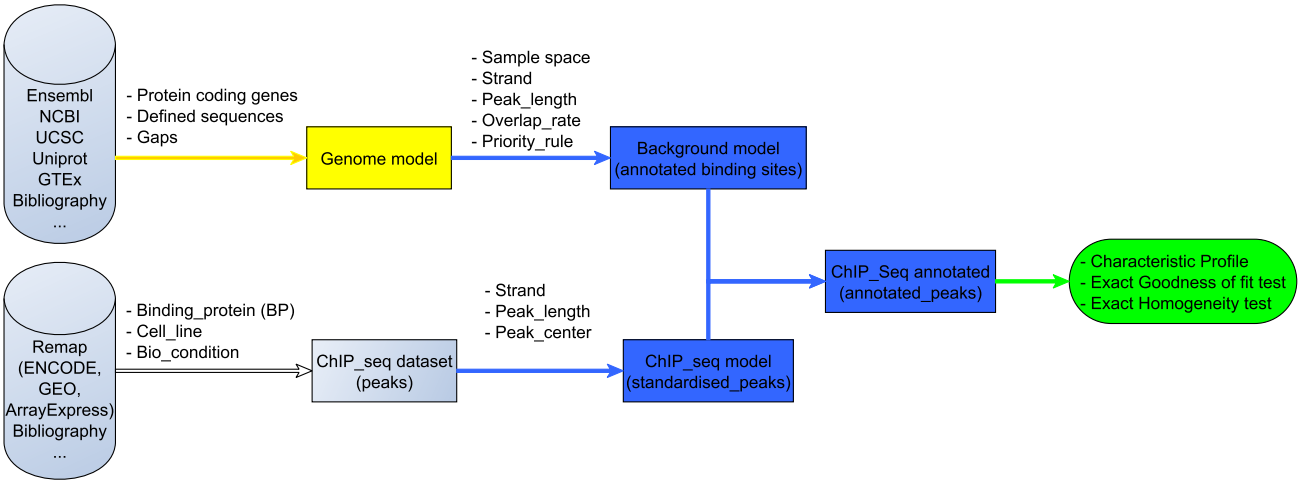

Figure 1. Overview of the different models created, in which the information collected from different databases is transformed, as well as their main components and parameters.

2.2. Representation Models

2.2.1. Model Dimensionality

The dimensionality of the model is given by the number of different classes of functional regions assumed to be part of the genome of the organism under study. These functional regions will depend mainly on the biotype (e.g., protein-coding, non-coding RNA, pseudo-gene or jumping gene) of the genes whose locus in the genome provides a functional product, which will be included in the **genome model**. Although in this work only protein-coding genes, whose functional product is a protein, have been considered, genes of different biotypes could also be included, either simultaneously or separately, depending on the objective of the research or in order to refine the **genome model** according to the target protein under study. For this work, a protein-coding gene has been considered to consist of a promoter region [15], a proximal region [16], splice regions including donor,

and acceptor regions [17], a cleavage region [18], exons, and introns. Figure 2 shows both the layout as well as the extension from the considered reference point of all functional regions mentioned above. In addition to the functional regions directly related to the different gene biotypes, there are others where their own nucleotide sequence has a characteristic function without encoding any product. These would be the enhancers, insulators, and so on. These types of functional regions are distributed throughout the genome and can also be considered in the model.

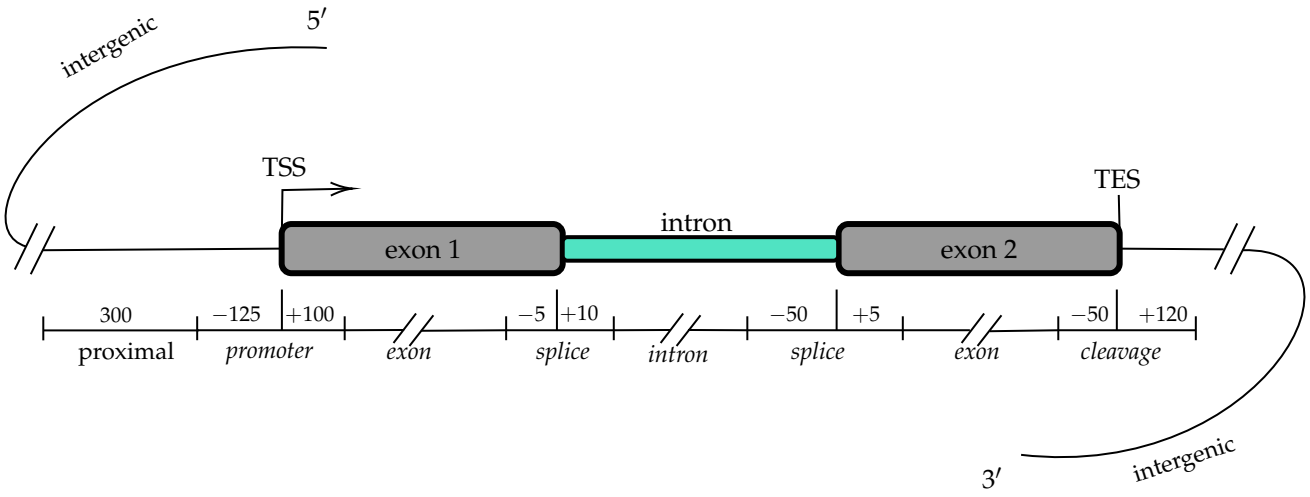

Figure 2. Depiction of the layout and extent (base pairs) of functional regions (italic) corresponding to a protein-coding gene (consisting of tow exons and an intron) lying in the forward strand of a DNA molecule (solid line). The reference points are the transcription start site (TSS), the transcription end site (TES), and the start and end coordinates of the exons of the selected transcript. Distances are in base pairs (*bp*).

It should be noted that protein-coding genes, in the case of eukaryotic organisms, such as *Arabidopsis thaliana*, can harbor different transcripts (product of the transcription process from the DNA molecule to the RNA) that include different exons, and therefore present different arrangements of the respective functional regions. This method represents for each gene a single transcript, which we call the canonical transcript, and it is usually the most representative transcript of the gene, although a tissue- or cell line-specific transcript, if available, could be chosen.

2.2.2. Genome Model

As mentioned above, this model accounts for all the existing knowledge of the genome of the cell line, tissue, or organism on which the ChIP-seq experiment was performed. This information is collected from available databases. It aims at describing the genome regions to which the experiment reads can be mapped and how they are organized (autosomal, sex, or mitochondrial chromosomes, plasmids, ...). Note that the greater the knowledge about the entity to be modeled, the more accurate the model becomes. Therefore, the regions without defined sequences, called **gaps**, are taken into account. Once the genome fragments with defined sequence have been identified, the next step is to locate the different classes of functional regions according to the genes that are considered to be part of the genome. In the genome model, the intergenic region is the default one, and it includes regions of the genome that have not been annotated with any kind of functional region and represent either regions that have no functionality (at least of interest for the analysis) or regions whose functionality is unknown. In this work, we assume a haploid genome model regardless of the organism's karyotype, because the state-of-the-art mapping and base calling algorithms do not discriminate between the alleles of homologous chromosomes [19]. However, if peaks falling on homologous chromosomes could be distinguished, they could be represented in this model.

2.2.3. Background Model

An annotated binding site is a region located in the genome with a specific length. It is defined by the chromosome to which it belongs, the start and end chromosome coordinates, the strand where it is located, and the annotation of the functional region class assigned to it. This annotation is the result of an exhaustive and mutually exclusive process through which each annotated binding site is assigned to one functional region class of the genome model under study, or the intergenic class. From the point of view of a given **genome model**, the **background model** is the set of all the annotated binding sites that it can host, according to a set of parameters. These parameters allow different **background models** to be defined for the same **genome model**. From the point of view of a target protein, a **background model** is the population of all annotated binding sites to which it could randomly bind.

The setting of a background model involves the following attributes:

- The **sample space** determines which portion of the genome model will be part of the background model: the whole genome model, one or several chromosomes, or even a portion of them;
- The **peak length** sets the length of the annotated binding sites. This must match the standardized peaks of the ChIP-seq model to be analyzed through this background model. This makes all annotated binding sites equiprobable;
- The **overlap rate** is the percentage that determines the minimum number of nucleotides of an annotated binding site that must overlap over the extent of a given functional region for its class to be assigned to it. The higher the value sets, the more restrictive the background model will be;
- The **priority rule** resolves cases where two or more classes of different functional regions overlap with each other, determining the class of functional region that should be assigned to an annotated binding site. There is mutual exclusion between the different classes;
- The **strand** determines the functional regions that are considered in the generation of the background model, according to the strand in which they are located. Its value may be *forward*, *reverse*, or *both*. The latter takes into account all functional regions regardless of the strand where they are located. It represents the flexibility of the method. The higher the value of the overlap rate attribute and the lower the value of the peak length, the more accurate the results of the analysis will be. Nevertheless, the settings of these parameters must be consistent with the accuracy and reliability of the peaks obtained in the ChIP-seq experiment under analysis.

2.2.4. ChIP-seq Model

We define a ChIP-seq model as the set of all standardized peaks derived from the peaks obtained from a certain ChIP-seq dataset through a standardization process (Appendix A). The purpose of a standardized peak is to specify both the location and the extent of the binding site that it depicts as accurately and reliably as possible. In order to achieve this, it is necessary to take into account both the discrepancy between its length and that of the real binding site, the so-called *summit* of a peak most likely being the coordinate of the peak center.

The features of a standardized peak are:

- **Location**, whose values in the genome are **chromosome**, **strand** and **start** and **end**, both expressed in chromosomal coordinates;
- **Peak center**, which is determined by the standardization process;
- **Peak length**, which may be either manually estimated by the researcher according to prior information about the target protein under study or automatically computed by motif discovery algorithms according to information gained by this analysis itself.

For each match between the set of standardized peaks of a **ChIP-seq model** and the annotated binding sites of a particular background model, an annotated peak is obtained.

An annotated peak has the additional attribute **Annotation**, which indicates the class of functional region or alternatively the class "intergenic" that has been assigned to it. Note that the value of this attribute relies on the background model used.

2.3. Analysis Models

A ChIP-seq model annotated through a certain background model can be treated as the result of a random vector following a multivariate hypergeometric distribution, and hence different analyses can be applied. On the one hand, each ChIP-seq model can be characterized by the number of annotated peaks in each class, by providing a characteristic profile of the binding sites. A goodness-of-fit test may be developed to determine the degree of similarity between the distribution of functional regions obtained from the ChIP-seq model (observed frequencies) and that of the background model used for its annotation as reference (expected frequencies). Furthermore, an homogeneity test can also be applied to analyze the similarity between the distributions of two different ChIP-seq models.

2.3.1. Characteristic Profile

The **characteristic profile** of a ChIP-seq model is obtained by counting how many standardized peaks are annotated per class, according to a random vector $\mathbf{X}$ with a multivariate hypergeometric distribution (1). For each ChIP-seq model, a profile includes a set of values, such as the absolute and relative frequencies and their confidence intervals (CI), expected values, standard deviations (sd), and Z-scores, whose experimental results have been summarized in Appendix B (Tables A1–A8).

In order to obtain a more accurate standardized profile of a protein-DNA binding sites from a cell line, the ratios of bindings over the functional regions $p_i = m_i/m$ are estimated from all ChIP-seq experiments available under the same target protein and cell line by regarding the relative frequencies $R_i = X_i/n$ as $k-1$ discrete random variables, since $\sum_{i=1}^{k} R_i = 1$ and from (2):

$$E(R_i) = p_i \text{ and } Var(R_i) = \frac{1}{n} p_i(1-p_i)\frac{m-n}{m-1}, \tag{3}$$

and hence, the method of moments yields consistent and unbiased estimators of the ratios of bindings, $\widehat{p}_i = r_i$ [20,21].

Furthermore, a non-parametric bootstrap resampling approach [22] is applied to estimate the empirical variability of the ratio of each functional region so as to calculate the 95% CIs for the ratios of the peak counts into each region by using 10,000 bootstrapped replicates. In particular, for each functional region S_i, the percentile method was applied to calculate the 95% CI for the relative frequency by selecting the 2.5th and 97.5th percentiles of its 10,000 bootstrapped replications, i.e., the $r_{(250)}$ and $r_{(9750)}$ values of R_i, where $r_{(j)}$ represents the jth value in increasing order. The distribution of the relative frequencies, along with the 95% CIs (Figures A1A–A8A in Appendix B), represents the profile of the target protein under study with respect to the sites where it is located throughout the selected portion of the genome of the given cell line under a specific biological condition.

In addition, for the functional region S_i according to a ChIP-seq annotated through a given background model, a Z_i-score can be obtained by standardising the ith component of $\mathbf{X}$ from (2), or equivalently, by standardising the ith relative frequency R_i from (3), as follows

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{Var(X_i)}} = \frac{R_i - E(R_i)}{\sqrt{Var(R_i)}} \tag{4}$$

being $E(Z_i) = 0$, $Var(Z_i) = 1$, and the Z-scores' covariances are given by

$$Cov(Z_i, Z_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}, \text{ for } j \neq i.$$

The Z-scores obtained for the functional regions quantify the preferences that the target protein has for each one of these classes in such a particular ChIP-seq model (Figures A1B–A8B in Appendix B).

2.3.2. Goodness-of-Fit Test

Considering the multivariate hypergeometric distribution of the peak counts into the functional regions on the genome as the reference model (**background model**), i.e., the null hypothesis

$$H_0 : (X_1, \ldots, X_k) \sim MH(n, (m_1, \ldots, m_k), m),$$

this test measures the fit of the observed peak counts (**ChIP-seq model**) to the expected peak counts (**background model**), which analyzes whether their differences were by chance.

For population size m large relative to n, sampling without replacement is closely approximated by sampling with replacement. Thus, the joint modeling of functional regions will be analyzed by assuming independence among the Z_i-scores (4) due to large m, and in addition, each Z_i may be approximated by a standard normal distribution [23]. Other approximations related to the hypergeometric distribution or transformations of it have been studied in [20,24–27].

In this first approach, the statistic T defined by the sum of squares of Z_i-scores can be approximated by a chi-squared distribution with $k-1$ degree of freedom:

$$T = \sum_{i=1}^{k} Z_i^2 \approx \chi^2_{k-1}, \tag{5}$$

which is similar to the multinomial test of Pearson, by using the binomial approximation to the hypergeometric given by Sandiford [28] for each Z_i-score. Some alternative statistics and corrections for continuity have been discussed in the literature, e.g., see [14,23,24,26–30], although the difference is negligible for large m [24]. Moreover, the maximum, range and rate of the extreme order statistics can also be used for detecting outliers and extreme values [27,31].

From the result of this test (p-value), the degree of similarity between the corresponding ChIP-seq model and its associated background model can be quantified. Thus, the decision can be made on the randomness (or not) of the binding sites where the target protein has been located, i.e., whether the distribution obtained from the classes of the annotated peaks is what would be expected by chance according to the background model used, or otherwise whether there is a specific functional region whose frequency is significantly different from the expected.

2.3.3. Test of Homogeneity

From two ChIP-seq experiments, this test measures the difference between the observed peak counts into the same set of k functional regions under an underlying multivariate hypergeometric distribution, i.e., the null hypothesis

$$H_0 : \mathbf{Y}_1 \sim MH(n_1, (m_{11}, \ldots, m_{1k}), m) \equiv \mathbf{Y}_2 \sim MH(n_2, (m_{21}, \ldots, m_{2k}), m).$$

Hence, each Z_{ji}-score associated with Y_{ji} given by (4) depends on the parameters (n_j, m_{ji}, m) for $j = 1, 2$, which can be reduced to (n_j, m_i, m), since $m_{1i} = m_{2i}$ under the null hypothesis. Thereby, the mean and variance of Y_{ji} can be written as

$$E(Y_{ji}) = n_j \frac{m_i}{m} \quad \text{and} \quad Var(Y_{ji}) = n_j \frac{m_i}{m}\left(1 - \frac{m_i}{m}\right)\frac{m - n_j}{m - 1},$$

where the subpopulation size m_i is estimated from the observed values $y_{1i} + y_{2i}$ of $Y_{1i} + Y_{2i} \sim H(n_1 + n_2, m_i, m)$. Thus, the maximum likelihood estimate of m_i is an integer value in the interval

$$[l_i, u_i] = \left[\frac{(m+1)(y_{1i}+y_{2i})}{n_1+n_2} - 1, \frac{(m+k-1)(y_{1i}+y_{2i})}{n_1+n_2} \right],$$

i.e., $\widehat{m}_i$ is between the greatest integer less than or equal to u_i and the lowest integer greater than or equal to l_i, which can be found by using the algorithm of Oberhofer and Kaufmann [32]. In particular, if $m/(n_1+n_2)$ is an integer, the maximum likelihood estimates $\widehat{m}_i$ can be expressed as (see [20,32,33])

$$\widehat{m}_i = (y_{1i}+y_{2i})\frac{m}{n_1+n_2},$$

which are as the ones obtained by the method of moments [20,21].

Therefore, under the same assumptions of Section 2.3.2, the statistic T defined by the sum of squares of Z_{ji}-scores can be approximated by a chi-squared distribution with $k-1$ degree of freedom:

$$T = \sum_{j=1}^{2}\sum_{i=1}^{k} Z_{ij}^2 \approx \chi_{k-1}^2. \tag{6}$$

From the result of this test (p-value), the degree of similarity between the two ChIP-seq models annotated through the same background model can be quantified. Thus, we can identify similar characteristic profiles and the functional region class whose frequencies are most significantly different between both profiles, and therefore also the biological conditions where the target protein location is differentially altered.

3. Results

3.1. The Use Case

The validation of our method was carried out using ChIP-seq experiments collected from the ReMap2020 database (https://remap2020.univ-amu.fr, accessed on 30 November 2021) [34]. The peaks of the ReMap2020 ChIP-seq datasets were generated by applying its own pipeline from reads generated by ChIP-seq, ChIP-exo, and DAP-seq experiments collected from public resources such as GEO (https://www.ncbi.nlm.nih.gov/geo, accessed on 30 November 2021), ENCODE (https://www.encodeproject.org, accessed on 30 November 2021) or ENA (https://www.ebi.ac.uk/ena/browser/home, accessed on 30 November 2021). ReMap2020 contains 5798 ChIP-seq datasets performed on cell lines belonging to *Homo sapiens*, and 795 performed on *Arabidopsis thaliana*.

In particular, four *Arabidopsis thaliana* ChIP-seq datasets correspond to the GSE112951 experiment carried out by Nassrallah et al. [35], which analyzed the influence of the light-mediated development protein (DET1) on the pattern of monoubiquitination (chemical modification where a ubiquitin molecule is added to the target molecule) of histone H2B (H2Bub). The DET1 is a component of light signal transduction machinery, involved in the repression of photomorphogenesis in darkness through regulation of the activity of ubiquitin conjugating enzymes, involved in the repression of de-etiolation in developing seedling, and involved in the repression of the blue light responsive promoter in chloroplasts (UniProt Consortium https://www.uniprot.org/uniprot/P48732, accessed on 30 November 2021). In such a study, two replicates of a ChIP-seq experiment using an antibody against H2Bub were performed in each condition, where each peak obtained indicates that histone H2B is monoubiquitinated. For both, a gene had been tagged if a peak fell within the region spanning from -1 kb from the TSS to $+1$ kb from the TES. The results of Nassrallah et al. [35] revealed that DET1 is directly linked to the histone H2B monoubiquitination pathway. They found that over 6900 genes supported a peak in all four samples, while the number of genes decayed over 20% in the samples with the mutated gene relative to the wild type samples, regardless of the light or dark condition. Table 1 shows the peaks obtained in the experiments performed on the Col-0-seedling cell line for the wild type and the knockout mutant for the DET1 gene in both light and darkness conditions for 5 days.

These four *Arabidopsis thaliana* ChIP-seq datasets synthesised in Table 1 have been chosen to illustrate the usefulness of the proposed method.

Table 1. Number of peaks obtained in the four ChiP-seq experiments carried out in the GSE112951 study on the Col-0-seedling cell line of the organism *Arabidopsis thaliana*, for samples with the wild-type and knockout mutant DET1 gene subjected to light and darkness conditions for 5 days.

Study	Cell Line	ChIP-seq Dataset	Bio-Condition		Peaks
			DET1	Light	
GSE112951	Col-0-seedling	5d-L	Wild type	Light	12,782
		5d-D		Darkness	12,305
		det1_5d-L	Knockout mutant	Light	12,165
		det1_5d-D		Darkness	12,173

3.1.1. The *Arabidopsis thaliana* Genome Model

The *Arabidopsis thaliana* assembly TAIR10 (https://plants.ensembl.org/Arabidopsis_thaliana/Info/Index accessed on 30 November 2021) was used for generating the genome model. A total of 27,420 protein-coding genes (Table A9) were collected from Ensembl Plants release 51 [36] using the BioMart tool. The components of the BioMart query used are specified in Table A10. We only used the 5 genomic chromosomes, with a total length of 119,146,348 bp, which were the ones used in the original experiment. The *gaps* were collected from https://genome.ucsc.edu/goldenPath/help/examples/hubExamples/hubAssembly/plantAraTha1/araTha1/gap.html (accessed on 30 November 2021), which represent 0.156% of the total genome. Note that gaps smaller than 10 nucleotides (*nt*) were not considered.

We used the regions indicated in Section 2.2.1 as functional regions of this model, and their description was shown in Figure 2, including the intergenic region. Therefore, we used a 7-dimensional genome model. On the other hand, an 8-dimensional genome model was created by including the enhancer as the functional region. The enhancers were obtained from Zhu et al. [37].

3.1.2. The *Arabidopsis thaliana* Background Model

The values of the parameters set for this background model were the following:

- *Peak length* = *31 nt*, which is an average length to deal with inaccuracy in the peaks;
- *Overlap rate* = 20%; an overlap of at least 7 *nt* of a standardized peak over a certain class of functional region is necessary for that class to be assigned to that peak;
- *Sample space* = $\{1, 2, 3, 4, 5\}$, that is, the 5 genomic chromosomes;
- *Strand* = *both*. All functional regions belonging to the two strands of the DNA molecule, forward and reverse, are considered;
- *Prior rule* = *promoter* > *proximal* > *enhancer* > *cleavage* > *splice* > *exon* > *intron* > *intergenic*. A functional region class had higher priority than those to its right and lower priority than those to its left. Thus, the promoter class had the highest priority over the other classes.

Once all the above parameters were applied to the information stored in the genome model, the two background models with 7 and 8 dimensions (dm) were generated (Table 2). It should be noted that the highest percentage of annotated binding sites corresponds to the intergenic class, followed by exon and splice. The comparison of both background models reveals that 2% out of the annotated binding sites of the enhancer class in the 8 dm are found in regions of the intergenic class in the 7 dm.

Table 2. Percentage of annotated binding sites of each functional region class considering the 5 genomic chromosomes of *Arabidopsis thaliana*, both for the 8 dm background model that includes the enhancer class, and for the 7 dm background model that does not include the enhancer class.

Functional Regions	**Arabidopsis Background Model**	
	8 dm (%)	**7 dm** (%)
Promoter	5.49	5.49
Proximal	6.36	6.36
Enhancer	2.04	-
Cleavage	3.02	3.22
Splice	10.42	10.43
Exon	28.89	29.07
Intron	8.21	8.24
Intergenic	35.57	37.19
Total	**100**	**100**

3.1.3. ChIP-seq Models for GSE112951

Although in the original study two replicates were performed for each of the four biological conditions, the Remap2020 database unified the reads from both replicates as input data to its pipeline to generate a single ChIP-seq dataset for each biological condition. Thus the ChIP-seq datasets collected from Remap2020 consisted of 12,782 and 12,305 peaks for the wild type condition in light (5d-L) and dark (5d-D) respectively, and 12,165 and 12,173 peaks for the mutant in light (det1_5d-L) and dark (det1_5d-D), respectively (Table 1).

The peaks of all four ChIP-seq datasets were standardized (Appendix A). The *peak center* parameter was the summit of each peak and the *peak length* was 31 nt. The standardized peaks of each sample were annotated across the two background models (7 dm and 8 dm), so two annotated ChIP-seqs were obtained for each ChIP-seq model.

3.2. Characteristic Profiles for GSE112951

The characteristic profiles of the four ChIP-seq models of the GSE112951 study annotated through each of the two background models, 7 dm and 8 dm, were obtained (see Table 3). Taking into account the high percentage of annotated peaks (above 98%) that were assigned a class of non-intergenic functional regions in each of the ChIP-seq models for both background models, it can be stated that these background models fit well the characteristics of the protein under study. On the other hand, the difference in the percentages between both background models was in the range $[0.016, 0.033]$. This, along with the high percentage value for the background model 7 dm, indicated that the incorporation of the enhancer class in the background model 8 dm is not relevant for the protein under study, at least considering the sample space formed by the 5 genomic chromosomes.

Table 3. Characteristic profiles and percentage of annotated peaks for the four ChIP-seq models belonging to study GSE112951, according to both background models 7 dm and 8 dm.

ChIP-seq	**Background**	**Characteristic Profile**			**% Annotated Peaks**	
Model	**Model**	**Table**	**Relative freq.**	***Z*-Score**	**Functional Regions**	**Intergenic**
5d-D	**7 dm**	Table A1	Figure A1A	Figure A1B	98.635	1.365
	8 dm	Table A2	Figure A2A	Figure A2B	98.651	1.349
5d-L	**7 dm**	Table A3	Figure A3A	Figure A3B	98.537	1.463
	8 dm	Table A4	Figure A4A	Figure A4B	98.560	1.440
det1_5d-D	**7 dm**	Table A5	Figure A5A	Figure A5B	98.850	1.150
	8 dm	Table A6	Figure A6A	Figure A6B	98.883	1.117
det1_5d-L	**7 dm**	Table A7	Figure A7A	Figure A7B	98.775	1.225
	8 dm	Table A8	Figure A8A	Figure A8B	98.792	1.208

If we visually examine both the relative frequencies and the Z-score values obtained in the four samples with respect to both background models (Figures A1–A8), two patterns could be distinguished depending on the number of monoubiquitinated H2ub sites in the observed values and the expected ones according to the background model used. One pattern was shared between wild type samples and another between those with the mutated DET1 gene, regardless of light or dark conditions. For the background model 7 dm, in the two wild type 5d-D and 5d-L samples, the intron class showed the greatest difference with positive values of 72.5 and 74.3 respectively, followed by the exon class with values of 46.3 and 47.8 respectively, and the splice with values of 44.6 and 44.3 respectively. In the two samples with the mutated DET1 gene det1_5d-D and det1_5d-L, the intron class showed the greatest difference with positive values of 95.2 and 96.7, respectively, followed in this case by the splice class with values of 46.9 and 47.2 respectively and the exon with values of 31.0 and 29.9, respectively. In all four samples the intergenic class had the lowest observed number of monoubiquitinated H2ub sites compared to the expected ones. These same patterns could be observed for the background model 8 dm, with very similar Z-scores, with those corresponding to the enhancer class being found in the four samples within the range $[-16.0, -15.4]$, demonstrating that this class was one of the least relevant for the study of the target protein.

3.3. Goodness-of-Fit Test for GSE112951

The results of the goodness-of-fit test (5) for each of the four samples with respect to both background models are shown in Table 4. The p-value was 0 for each of them, which means that the observed H2Bub sites do not fit to the pattern of the background model, but showed a bias towards certain classes, as described in the previous section.

Table 4. Results of the goodness-of-fit test for the four ChIP-seq models belonging to the GSE112951 study on the DET1 target protein according to the 7 dm and 8 dm background models.

ChIP-seq Model	7 dm			8 dm		
	Statistic	**df**	***p*-Value**	**Statistic**	**df**	***p*-Value**
5d-D	17,712	6	0	17,553	7	0
5d-L	18,402	6	0	18,238	7	0
det1_5d-D	20,556	6	0	20,394	7	0
det1_5d-L	20,799	6	0	20,639	7	0

3.4. Test of Homogeneity for GSE112951

Based on the profile drawn by the four relative frequencies, two patterns could be distinguished in Figures 3A and A9A. One pattern was exhibited by the two ChIP-seq models with the wild type condition and the other one by the two ChIP-seq models with the mutated DET1 gene, regardless of the light or dark condition. The results of the homogeneity tests (6) carried out on the six pairs formed by the four ChIP-seq models under study are shown in Table 5.

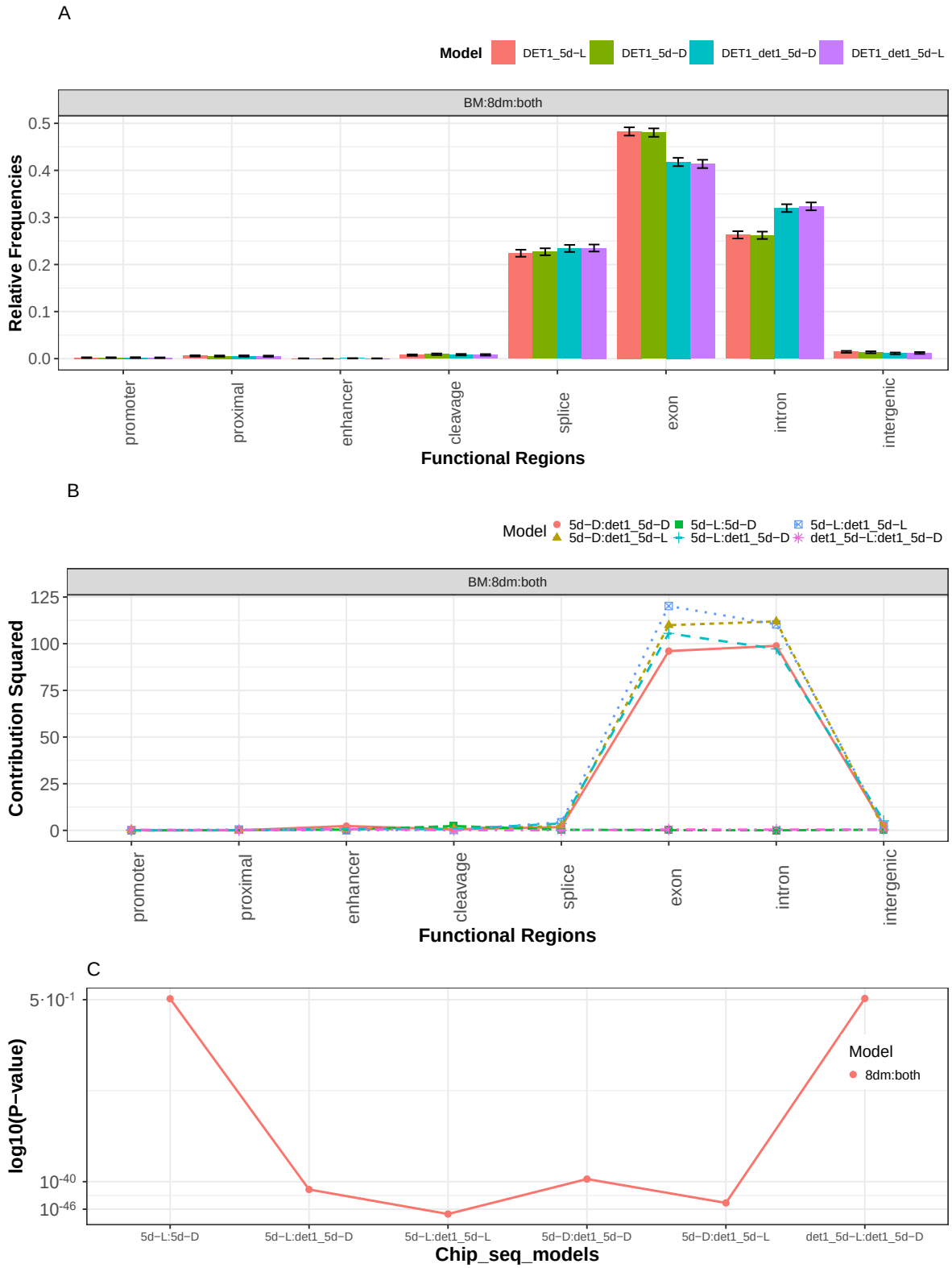

Figure 3. (**A**) Multiple bar diagram of relative frequencies with their corresponding 95% confidence interval by functional region for the four ChIP-seq models **5d-L**, **5d-D**, **det1_5d-L**, and **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedlng, according to the 8 dm background model. (**B**) Depiction of the contribution squared by functional region for each ChIP-seq model pair **5d-L:5d-D**, **5d-L:det1_5d-D**, **5d-L:det1_5d-L**, **5d-D:det1_5d-L**, **5d-D:det1_5d-D**, and **det1_5d-L:det1_5d-D**, according to the 8 dm background model, calculated on the basis of the test of homogeneity (6). (**C**) *p*-values result of such a test applied to each ChIP-seq model pair, according to the 8 dm background model. The values have been transformed to a logarithmic scale in base 10 for the sake of clarity.

Table 5. Results of the homogeneity test (6) between the six pairs formed by the four ChIP-sep models (5d-L, 5d-D, det1_5d-L, det1_5d-D), belonging to the GSE112951 study on the DET1 target protein, according to both 7 dm and 8 dm background models.

ChIP-seq Model	7 dm			8 dm		
	Statistic	**df**	***p*-Value**	**Statistic**	**df**	***p*-Value**
5d-L:5d-D	3	6	0.788	4	7	0.805
5d-L:det1_5d-D	211	6	$8.72 \cdot 10^{-43}$	213	7	$1.94 \cdot 10^{-42}$
5d-L:det1_5d-L	238	6	$1.45 \cdot 10^{-48}$	238	7	$8.79 \cdot 10^{-48}$
5d-D:det1_5d-D	199	6	$3.18 \cdot 10^{-40}$	202	7	$3.73 \cdot 10^{-40}$
5d-D:det1_5d-L	226	6	$4.94 \cdot 10^{-46}$	227	7	$2.32 \cdot 10^{-45}$
det1_5d-L:det1_5d-D	2	6	0.935	3	7	0.918

According to the corresponding *p*-values (see Figures 3C and A9C), significant differences were found between all pairs formed by one wild type sample and one sample with the mutated gene (*p*-value with magnitudes below 10^{-40}), while *p*-values above 0.5 were obtained between pairs formed by the two wild type samples or the two samples with the mutated gene. This indicates that the pattern of monoubiquitination is more dependent on the wild type or mutated gene condition than on the light or dark condition.

Further, Figures 3B and A9B display which functional regions contribute the most to the differences between samples. The exon and intron classes were practically the only ones responsible for the difference between the wild type condition and that of the mutated gene. The results presented in this section held for both 7 dm and 8 dm background models.

4. Discussion

In this work we have presented a statistical method that allows one to generate more accurate and reliable information from ChIP-seq experiments by modeling the peaks through a random vector with a multivariate hypergeometric distribution. We have showed how multidimensional models can improve the analysis of this type of experiments and how these generated models might facilitate the comparison between different experiments, regardless of when and where they have been carried out.

The results obtained are in line with those reported in the earlier study [35], the difference between samples mainly being due to the wild type/mutated gene conditions rather than by the light/darkness conditions. While the earlier study was based on the number of genes, our method was able to determine the most affected classes of functional regions by using multivariate distribution models. Furthermore, the models can be updated, which has been illustrated by the background models with different dimensions.

The sample space of the background models is related to the scalability of the method. For the sake of simplicity, we have not demonstrated such scalability in this work. As future work, we will study how the scalability is achieved by exploiting the grouping property of the multivariate hypergeometric distribution. Both individual genes and chromosomes can be defined as background models, thus allowing one to focus on the extent of the analysis, but always considering the whole genome and from a biologically reasonable point of view. Further work will also show how the use of more functional regions can also have a positive impact on the level of detail of the study.

5. Conclusions

The design, implementation, and validation of an analytical method have been introduced to deal with some challenging issues assisting in elucidating the biological implications of the peaks obtained in a ChIP-seq experiment. The coherence of the method has also been shown in this work, since the inclusion of an irrelevant functional region for the behavior of the protein under study did not affect the final results.

Author Contributions: Conceptualization and methodology, G.A.-H., M.F., J.-M.V., and J.T.F.-B.; software, G.A.-H. and M.F.; validation, G.A.-H., M.F., and J.-M.V.; writing—original draft preparation, G.A.-H.; writing—review and editing, G.A.-H., M.F., J.-M.V., and J.T.F.-B.; supervision, M.F. and J.T.F.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research is part of the grants TIN2017-85949-C2-1-R and PID2020-113723RB-C22 funded by MCIN/AEI/10.13039/501100011033/.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data generated in this work is available at https://github.com/gines-almagro/AnalysingProtein-DNAbindingSitesArabidopsisThalianaChip-seqExperiments (accessed on 1 November 2021).

Acknowledgments: The authors would like to thank the three anonymous reviewers for their comments and suggestions, which have improved our manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ChIP-seq	Chromatin immunoprecipitation sequencing
DNA	Deoxyribonucleic acid
NGS	Next generation sequencing
RNA	Ribonucleic acid
TSS	Transcription start site
TES	Transcription end site

Appendix A. Peak Standardization

The standardization process consists of combining the peaks of selected ChIP-seq datasets corresponding to the same target protein and cell line, regardless of the biological condition to which the cell line is subjected.

(**i**) **Peak selection** consists of a progressive selection process of peaks from different ChIP-seq datasets.
From the first peak selected, the peak with the start coordinate immediately above the first one, must meet two conditions to be selected. The first condition is that it must belong to a different ChIP-seq dataset than the raw peaks already selected for the same standardize peak. The second one is that either the summit of each raw peak falls within the extension of the other raw peak, or that they overlap the intervals formed by symmetrically extending the summit of each raw peak until the value of the *peak length* parameter is reached. In case of a positive selection, both raw peaks are merged, with the start coordinate value being the smaller of the two, and the end coordinate being the larger of the two, keeping the summits of each raw peak for future comparisons. This new merged peak is the one to be compared with the next raw peak. If this third party is also selected, the merged peak is updated, with its start being the lowest of the three and its end being the highest of the three, thereby maintaining their respective summits, and so on. If the selection is negative, the center of the standardized peak representing each of the selected raw peaks in their respective ChIP-seq model would be the position that determines the arithmetic mean of the summits of each of the selected raw peaks;

(**ii**) **Peak center** determines the center of the standardized peak, which is a position obtained as the arithmetic mean of the summits of the selected peaks from which it is derived;

(**iii**) **Peak length** is the length of the standardized peak as it extends symmetrically on both sides of its center. The higher the value of this parameter, the more ambiguous the standardized peak obtained. Notice that this value must match the value of the background model with which the ChIP-seq model will be analyzed.

Appendix B. Tables and Figures

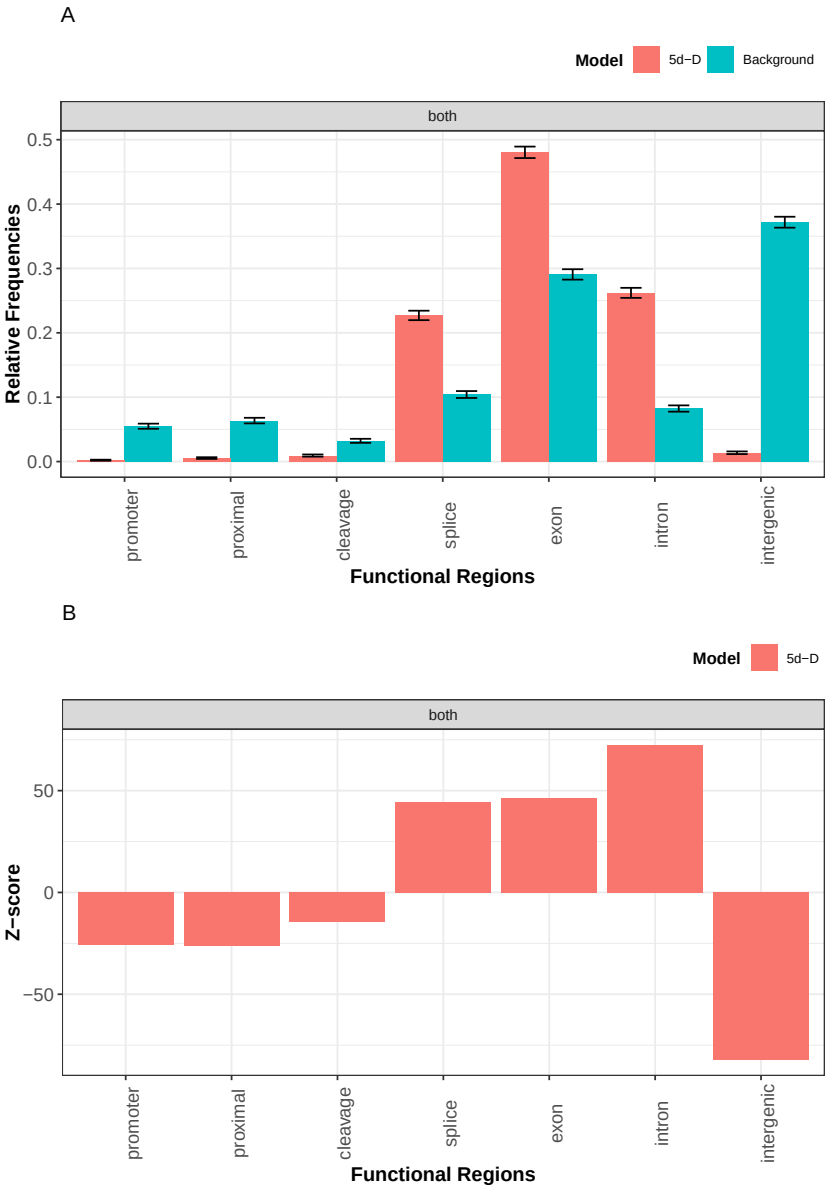

Figure A1. Graphics of the ChIP-seq model **5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model. (**A**) Bar diagram of **relative frequencies** with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A1. Characteristic profile for the ChIP-seq model **5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	26	0.00211	(0.00138, 0.00293)	675.5	25.3	−25.7
Proximal	67	0.00545	(0.00414, 0.00675)	783.1	27.1	−26.4
Cleavage	114	0.00927	(0.00764, 0.01105)	395.8	19.6	−14.4
Splice	2794	0.22706	(0.21967, 0.23438)	1283.1	33.9	44.6
Exon	5911	0.48037	(0.47152, 0.48931)	3577.0	50.4	46.3
Intron	3225	0.26209	(0.25429, 0.26989)	1014.5	30.5	72.5
Intergenic	168	0.01365	(0.01162, 0.01577)	4576.0	53.6	−82.2
Total	12,305	1		12,305		

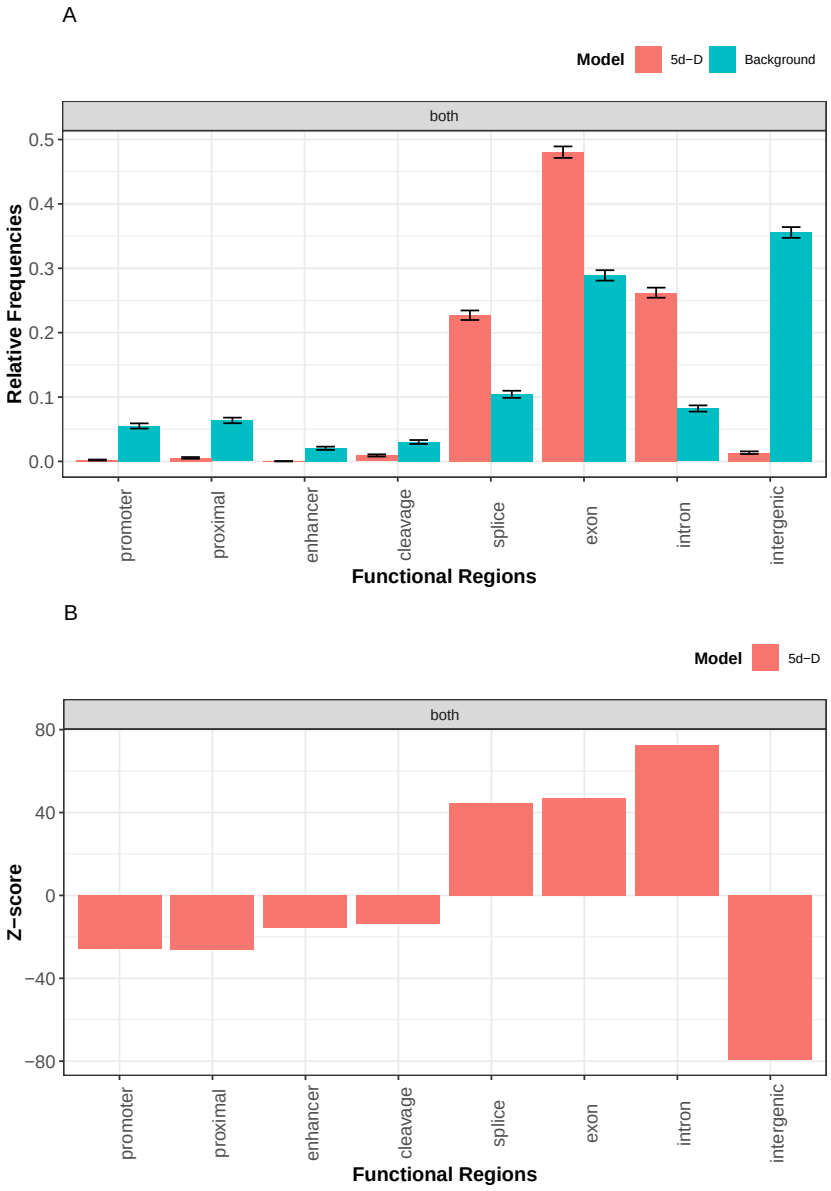

Figure A2. Graphics of the ChIP-seq model **5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A2. Characteristic profile for the ChIP-seq model **5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	26	0.00211	(0.00138, 0.00293)	675.5	25.3	−25.7
Proximal	67	0.00545	(0.00414, 0.00675)	783.1	27.1	−26.4
Enhancer	3	0.00024	(0.00000, 0.00057)	251.1	15.7	−15.8
Cleavage	114	0.00927	(0.00764, 0.01105)	370.9	19.0	−13.5
Splice	2794	0.22706	(0.21967, 0.23438)	1281.9	33.9	44.6
Exon	5910	0.48029	(0.47144, 0.48923)	3555.3	50.3	46.8
Intron	3225	0.26209	(0.25429, 0.26989)	1010.6	30.5	72.7
Intergenic	166	0.01349	(0.01146, 0.01560)	4376.6	53.1	−79.3
Total	12,305	1		12,305		

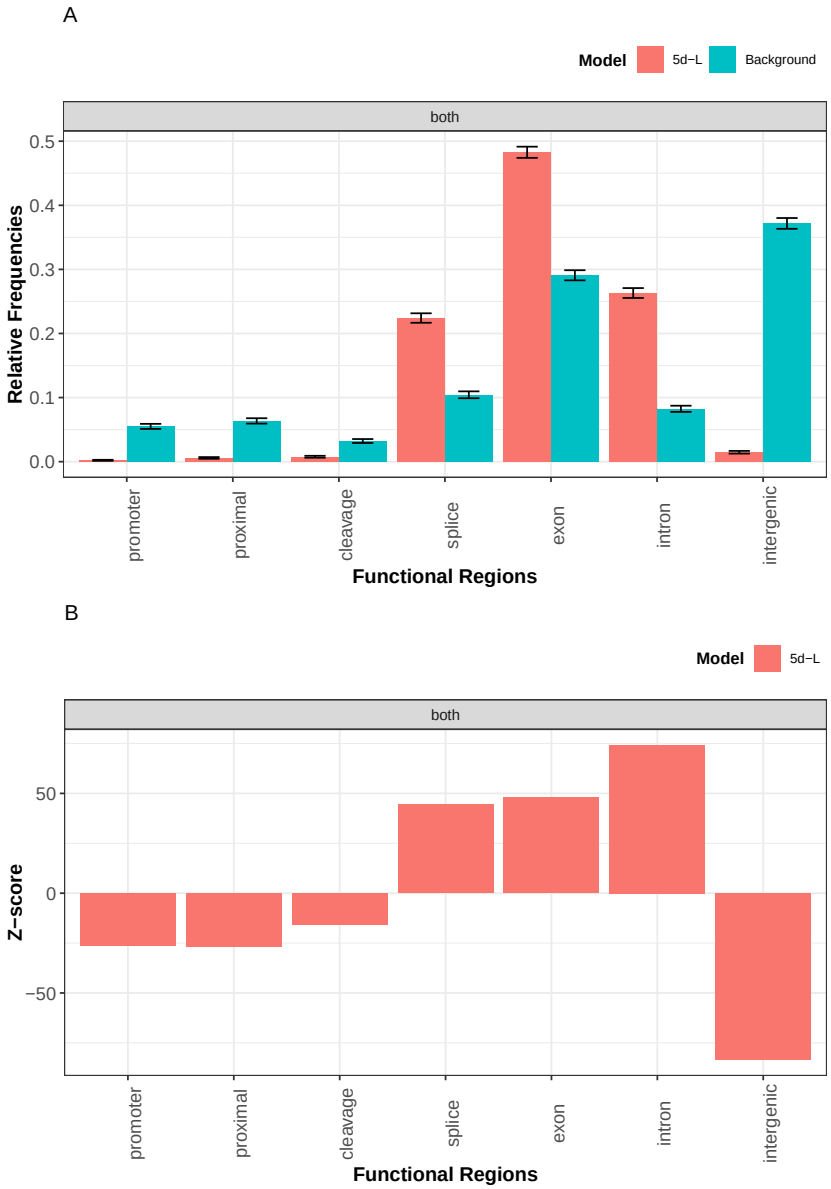

Figure A3. Graphics of the ChIP-seq model **5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A3. Characteristic profile for the ChIP-seq model **5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	27	0.00211	(0.00133, 0.00289)	701.7	25.8	−26.2
Proximal	74	0.00579	(0.00454, 0.00712)	813.4	27.6	−26.8
Cleavage	97	0.00759	(0.00618, 0.00915)	411.1	19.9	−15.7
Splice	2864	0.22407	(0.21671, 0.23150)	1332.9	34.6	44.3
Exon	6170	0.48271	(0.47403, 0.49147)	3715.7	51.3	47.8
Intron	3363	0.26310	(0.25544, 0.27085)	1053.8	31.1	74.3
Intergenic	187	0.01463	(0.01260, 0.01674)	4753.4	54.6	−83.6
Total	12,782	1		12,782		

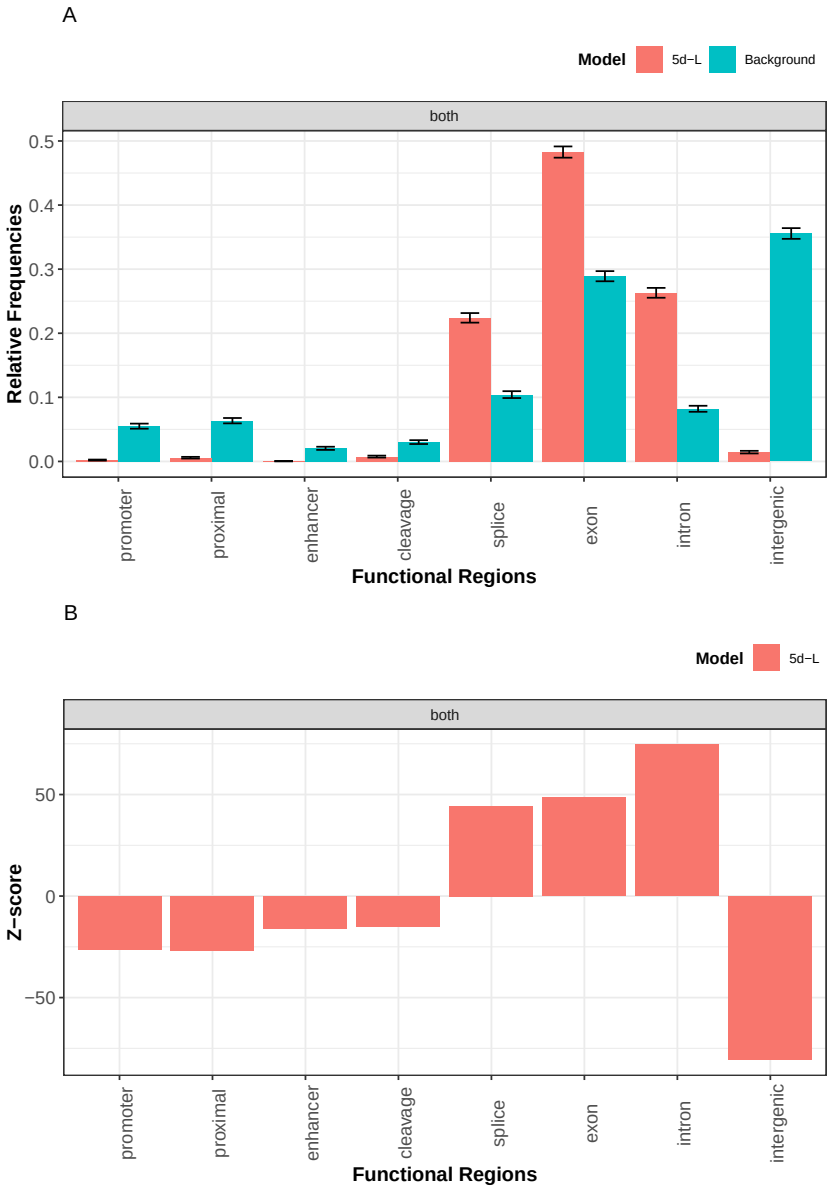

Figure A4. Graphics of the ChIP-seq model **5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A4. Characteristic profile for the ChIP-seq model **5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	27	0.00211	(0.00133, 0.00289)	701.7	25.8	−26.2
Proximal	74	0.00579	(0.00454, 0.00712)	813.4	27.6	−26.8
Enhancer	5	0.00039	(0.00008, 0.00078)	260.8	16.0	−16.0
Cleavage	96	0.00751	(0.00603, 0.00908)	385.2	19.3	−15.0
Splice	2863	0.22399	(0.21655, 0.23150)	1331.6	34.5	44.3
Exon	6170	0.48271	(0.47395, 0.49147)	3693.2	51.2	48.3
Intron	3363	0.26310	(0.25536, 0.27077)	1049.8	31.0	74.5
Intergenic	184	0.01440	(0.01244, 0.01651)	4546.3	54.1	−80.6
Total	12,782	1		12,782		

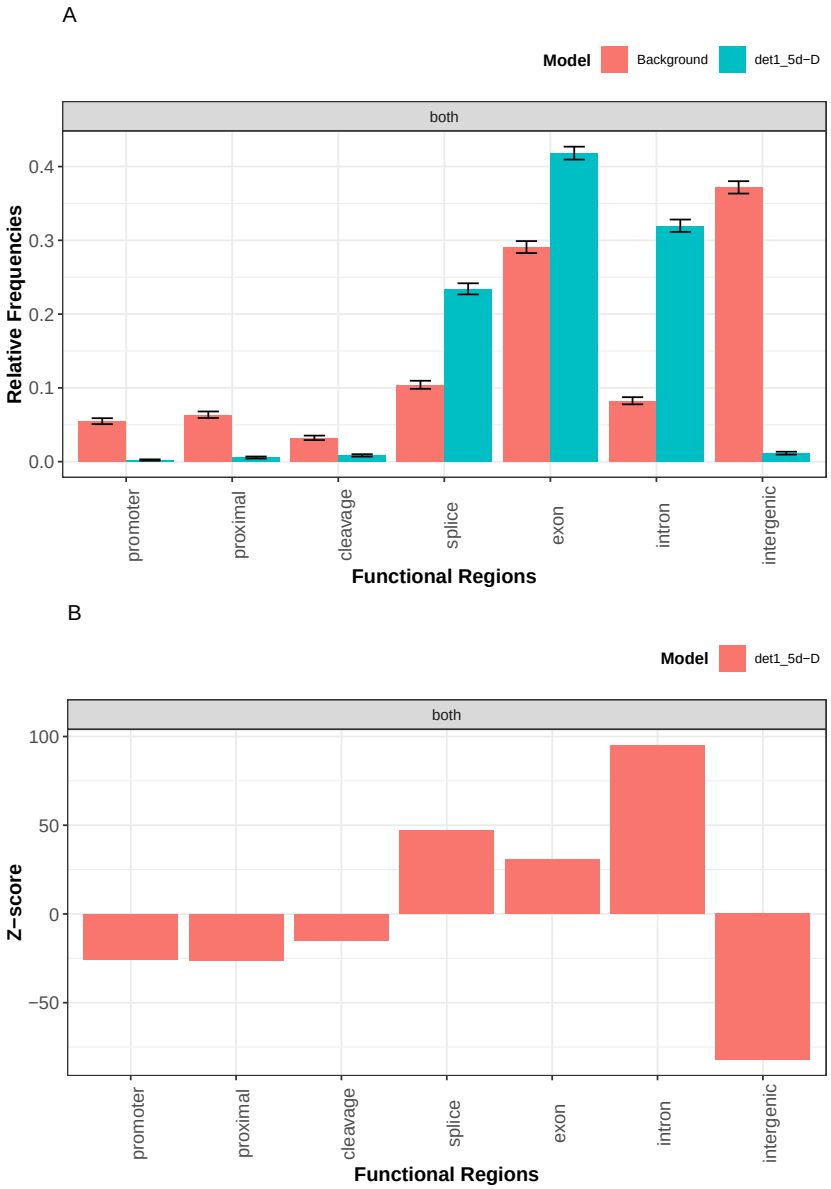

Figure A5. Graphics of the ChIP-seq model **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A5. Characteristic profile for the ChIP-seq model **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	27	0.00221	(0.00140, 0.00312)	668.3	25.1	−25.5
Proximal	68	0.00559	(0.00427, 0.00698)	774.7	26.9	−26.2
Cleavage	103	0.00846	(0.00690, 0.01010)	391.5	19.5	−14.8
Splice	2851	0.23421	(0.22673, 0.24176)	1269.4	33.7	46.9
Exon	5091	0.41822	(0.40943, 0.42693)	3538.6	50.1	31.0
Intron	3893	0.31981	(0.31143, 0.32819)	1003.6	30.3	95.2
Intergenic	140	0.01150	(0.00961, 0.01347)	4526.9	53.3	−82.3
Total	12,173	1		12,173		

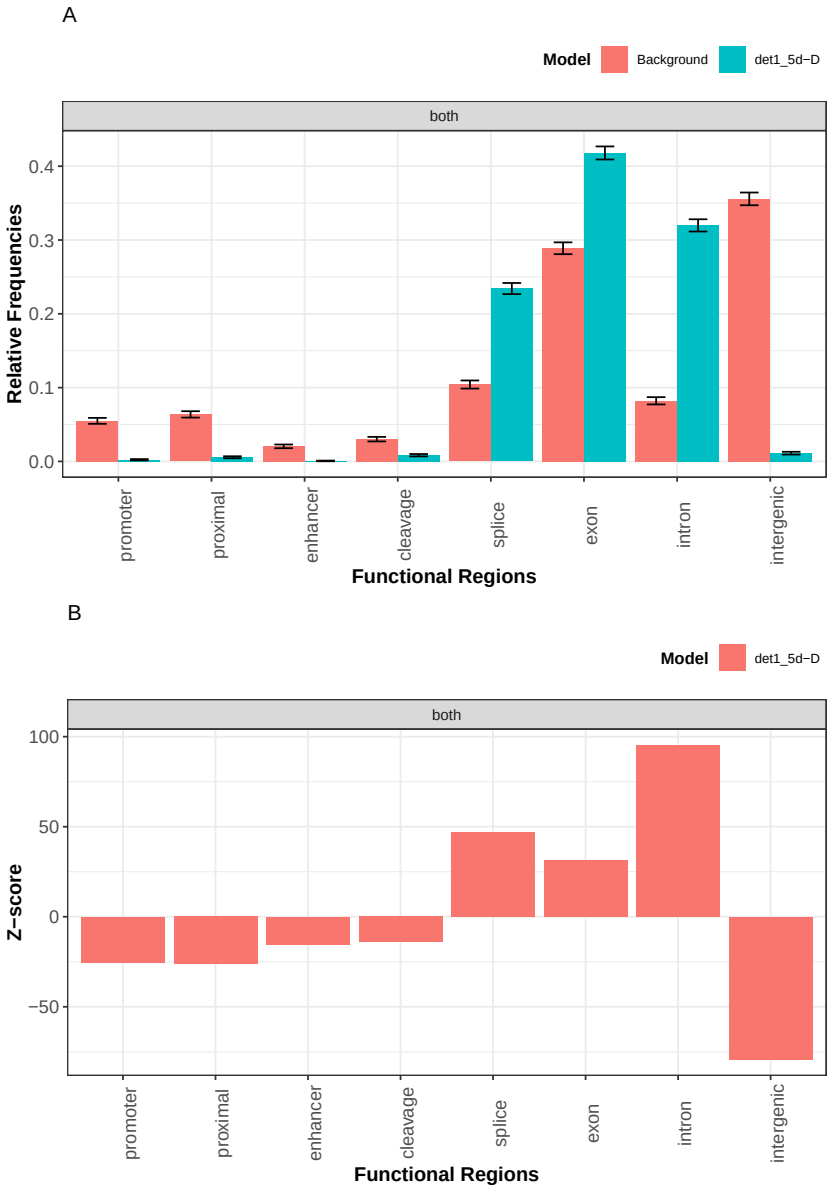

Figure A6. Graphics of the ChIP-seq model **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A6. Characteristic profile for the ChIP-seq model **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	27	0.00221	(0.00140, 0.00312)	668.2	25.1	−25.5
Proximal	68	0.00559	(0.00427, 0.00698)	774.7	26.9	−26.2
Enhancer	8	0.00066	(0.00025, 0.00115)	248.4	15.6	−15.4
Cleavage	102	0.00838	(0.00682, 0.01002)	366.9	18.9	−14.0
Splice	2851	0.23421	(0.22665, 0.24176)	1268.1	33.7	47.0
Exon	5088	0.41797	(0.40910, 0.42676)	3517.2	50.0	31.4
Intron	3893	0.31981	(0.31151, 0.32810)	999.8	30.3	95.5
Intergenic	136	0.01117	(0.00928, 0.01306)	4329.7	52.8	−79.4
Total	12,173	1		12,173		

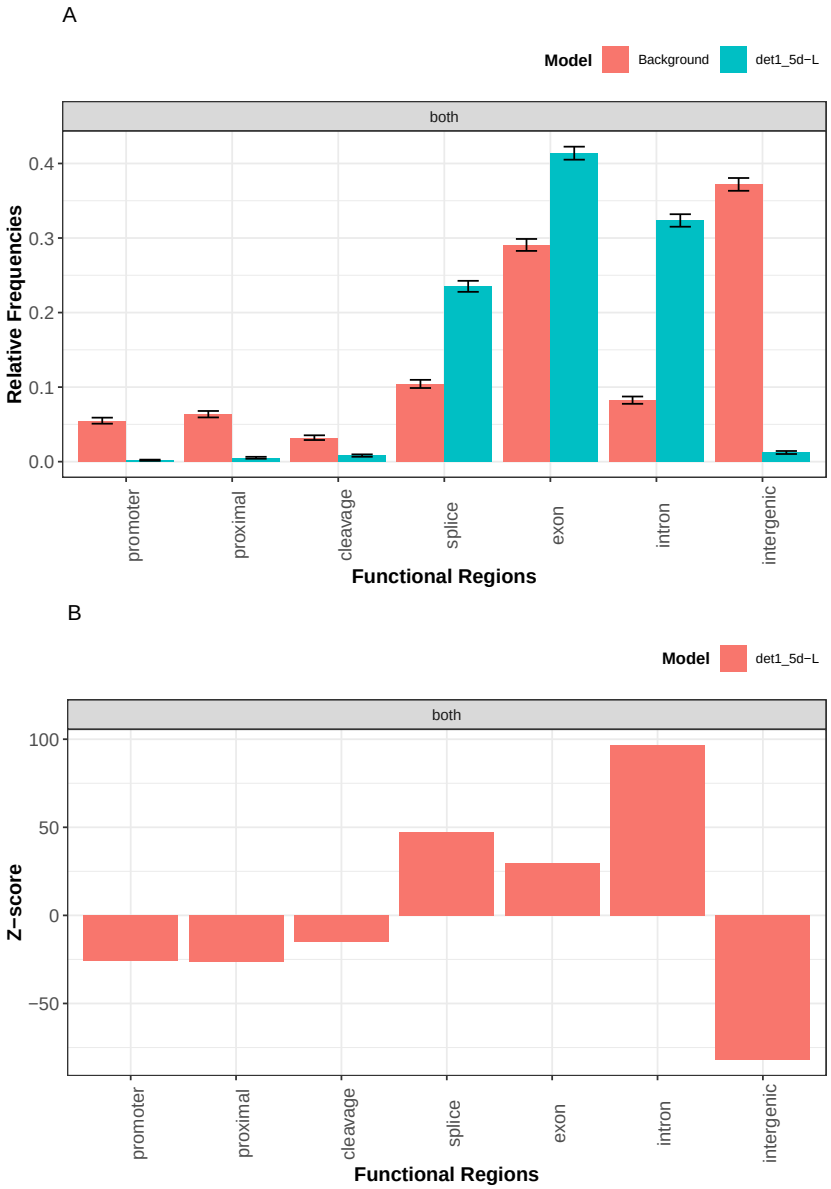

Figure A7. Graphics of the ChIP-sep model **det1_5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

Table A7. Characteristic profile for the ChIP-seq model **det1_5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **7 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	*Z*-Score
Promoter	23	0.00189	(0.00115, 0.00271)	667.8	25.1	−25.7
Proximal	63	0.00518	(0.00395, 0.00649)	774.2	26.9	−26.4
Cleavage	99	0.00814	(0.00658, 0.00978)	391.3	19.5	−15.0
Splice	2861	0.23518	(0.22778, 0.24258)	1268.6	33.7	47.2
Exon	5033	0.41373	(0.40510, 0.42252)	3536.3	50.1	29.9
Intron	3937	0.32363	(0.31517, 0.33194)	1002.9	30.3	96.7
Intergenic	149	0.01225	(0.01028, 0.01430)	4523.9	53.3	−82.1
Total	12,165	1		12,165		

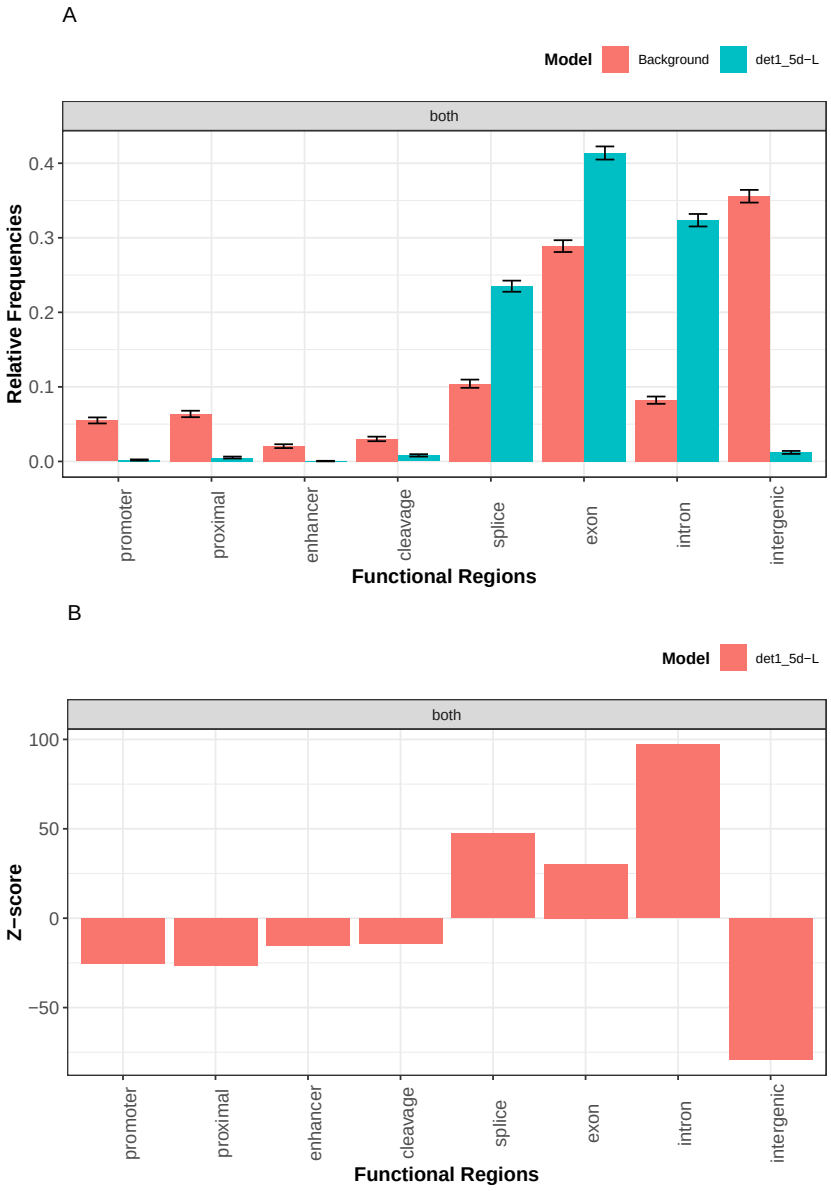

Figure A8. Graphics of the ChIP-sep model **det1_5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model. (**A**) Bar diagram of relative frequencies with their 95% confidence interval for each functional region. (**B**) Bar chart for the values of Z_{score} for each functional region.

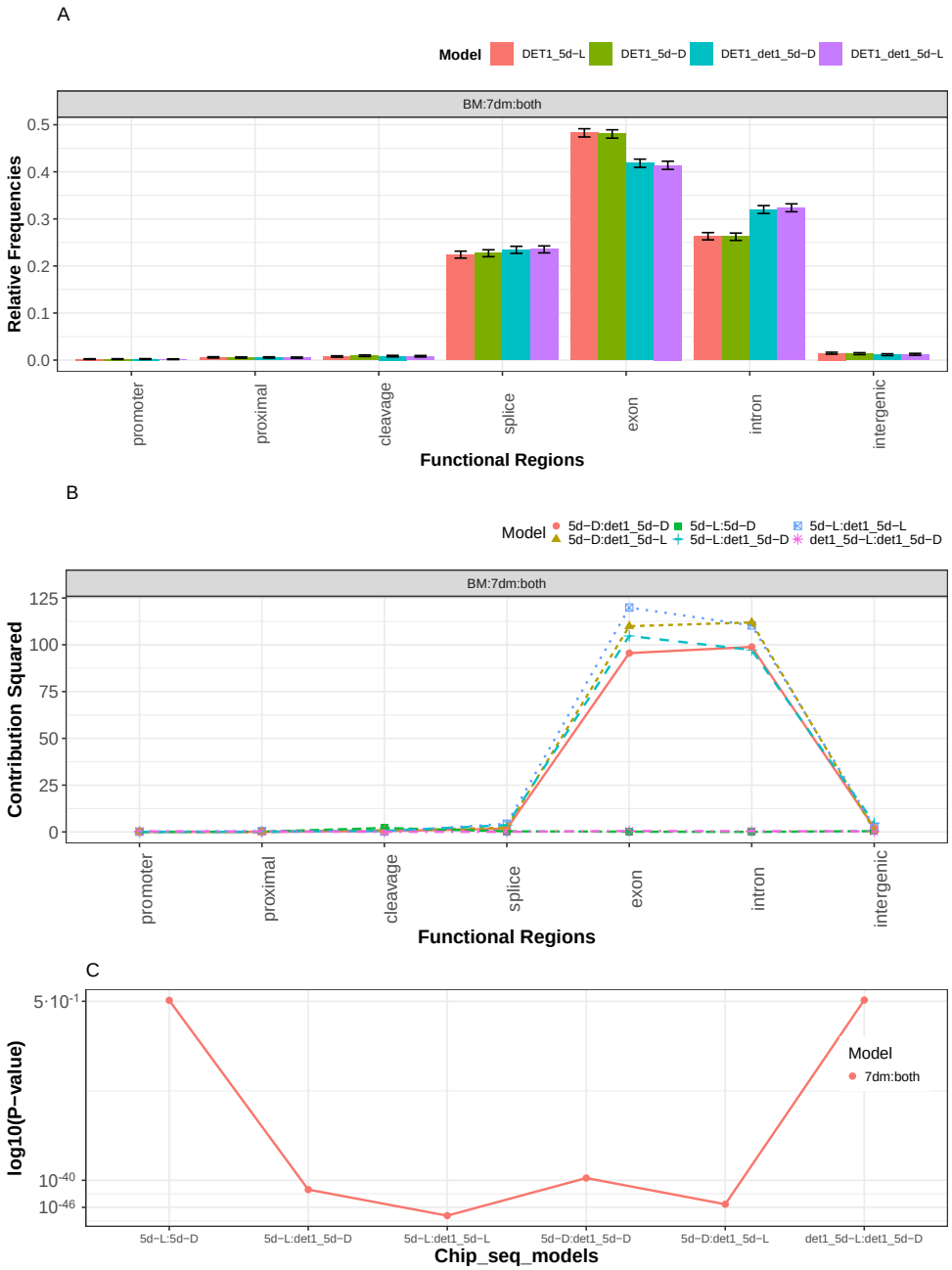

Figure A9. (**A**) Multiple bar diagram of relative frequencies with their corresponding 95% confidence interval by functional region for the four ChIP-seq models **5d-L**, **5d-D**, **det1_5d-L**, and **det1_5d-D**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedlng, according to the 7 dm background model. (**B**) Depiction of the contribution squared by functional region for each ChIP-seq model pair **5d-L:5d-D**, **5d-L:det1_5d-D**, **5d-L:det1_5d-L**, **5d-D:det1_5d-L**, **5d-D:det1_5d-D**, and **det1_5d-L:det1_5d-D**, according to the 7 dm background model, calculated on the basis of the test of homogeneity (6). (**C**) *p*-values result of such a test applied to each ChIP-seq model pair, according to the 7 dm background model. The values have been transformed to a logarithmic scale in base 10 for the sake of clarity.

Table A8. Characteristic profile for the ChIP-seq model **det1_5d-L**, belonging to study **GSE112951** on target protein DET1 in the cell line Col-0-seedling, according to the **8 dm** background model.

Functional Regions	Absolute Frequencies	Relative Frequencies	95% CI	Expected Values	sd	Z-Score
Promoter	23	0.00189	(0.00115, 0.00271)	667.8	25.1	−25.7
Proximal	63	0.00518	(0.00395, 0.00649)	774.2	26.9	−26.4
Enhancer	5	0.00041	(0.00008, 0.00082)	248.2	15.6	−15.6
Cleavage	98	0.00806	(0.00649, 0.00970)	366.6	18.9	−14.2
Splice	2860	0.23510	(0.22762, 0.24250)	1267.3	33.7	47.3
Exon	5032	0.41365	(0.40493, 0.42252)	3514.9	50.0	30.3
Intron	3937	0.32363	(0.31517, 0.33194)	999.1	30.3	97.0
Intergenic	147	0.01208	(0.01011, 0.01406)	4326.9	52.8	−79.2
Total	12,165	1		12,165		

Table A9. General description of the number of genes, length of chromosomes, percentage of gap sequence by chromosome, and number of enhancers for the genome of the *Arabidopsis thaliana* TAIR10.

	Chr1	Chr2	Chr3	Chr4	Chr5	Total
# genes protein-coding	7149	4315	5455	4174	6327	**27,420**
Total length (*bp*)	30,427,671	19,698,289	23,459,830	18,585,056	26,975,502	**119,146,348**
% undefined sequence	0.539	0.013	0.025	0.016	0.038	**0.156**
# enhancers	1574	942	1133	882	1340	**5871**

Table A10. Search setting for Ensembl Plant's Biomart tool to get a list with the genes that make up the Arabidopsis genome model.

Dataset	Filters	Attributes
Arabidopsis thaliana Genes (TAIR10)	Gene type: protein_coding Transcript type: protein_coding	Gene stable ID Gene name

References

1. Ji, H.; Jiang, H.; Ma, W.; Johnson, D.S.; Myers, R.M.; Wong, W.H. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **2008**, *26*, 1293–1300. [CrossRef]
2. Nakato, R.; Shirahige, K. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.* **2017**, *18*, 279–290. [CrossRef] [PubMed]
3. Mundade, R.; Ozer, H.G.; Wei, H.; Prabhu, L.; Lu, T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* **2014**, *13*, 2847–2852. [CrossRef] [PubMed]
4. Park, P.J. ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **2009**, *10*, 669–680. [CrossRef] [PubMed]
5. Furey, T.S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **2012**, *13*, 840–852. [CrossRef] [PubMed]
6. Salmon-Divon, M.; Dvinge, H.; Tammoja, K.; Bertone, P. PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinform.* **2010**, *11*, 415. [CrossRef] [PubMed]
7. Zhu, L.J.; Gazin, C.; Lawson, N.D.; Pagès, H.; Lin, S.M.; Lapointe, D.S.; Green, M.R. ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinform.* **2010**, *11*, 237. [CrossRef] [PubMed]
8. McLean, C.Y.; Bristor, D.; Hiller, M.; Clarke, S.L.; Schaar, B.T.; Lowe, C.B.; Wenger, A.M.; Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **2010**, *28*, 495–501. [CrossRef]
9. Kondili, M.; Fust, A.; Preussner, J.; Kuenne, C.; Braun, T.; Looso, M. UROPA: A tool for Universal RObust Peak Annotation. *Sci. Rep.* **2017**, *7*, 2593. [CrossRef] [PubMed]
10. Huang, W.; Loganantharaj, R.; Schroeder, B.; Fargo, D.; Li, L. PAVIS: A tool for Peak Annotation and Visualization. *Bioinformatics* **2013**, *29*, 3097–3099. [CrossRef] [PubMed]
11. Steyn, H. On discrete multivariate probability functions of Hypergeometric type. *Proc. K. Ned. Akad. Wetensh. Ser. A* **1955**, *58*, 588–595. [CrossRef]
12. Janardan, K.G.; Patil, G.P. A unified approach for a class of multivariate hypergeometric models. *Sankhyā Ser. A* **1972**, *34*, 363–376.
13. Nevill, A.; Kemp, C. On characterizing the hypergeometric and multivariate hypergeometric distributions. In *Statistical Distributions in Scientific Work*; Patil, G., Kotz, S., Ord, J., Eds.; Reidel: Dordrecht, The Netherlands, 1975; Volume 3, pp. 353–357.
14. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Discrete Multivariate Distributions*; Wiley: New York, NY, USA, 1997.

15. Thieffry, A.; Bornholdt, J.; Ivanov, M.; Brodersen, P.; Sandelin, A. Characterization of Arabidopsis Thaliana Promoter Bidirectionality and Antisense RNAs by Depletion of Nuclear RNA Decay Enzymes. Unpublished. Available online: https://www.biorxiv.org/content/early/2019/10/23/809194.full.pdf (accesed on 1 February 2021).
16. Korkuć, P.; Schippers, J.H.; Walther, D. Characterization and identification of cis-regulatory elements in arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.* **2014**, *164*, 181–200. [CrossRef]
17. Hebsgaard, S.M.; Korning, P.G.; Tolstrup, N.; Engelbrecht, J.; Rouzé, P.; Brunak, S. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **1996**, *24*, 3439–3452. [CrossRef] [PubMed]
18. Wu, X.; Liu, M.; Downie, B.; Liang, C.; Ji, G.; Li, Q.Q.; Hunt, A.G. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12533–12538. [CrossRef] [PubMed]
19. Feng, J.; Liu, T.; Qin, B.; Zhang, Y.; Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **2012**, *7*, 1728–1740. [CrossRef]
20. Hartley, H.; Rao, J. A new estimation theory for sample surveys. *Biometrika* **1968**, *55*, 547–557. [CrossRef]
21. Guenther, W. Hypergeometric distributions: Overview. In *Encyclopedia of Statistical Sciences*; Wiley: Hoboken, NJ, USA, 2006; Volume 5, pp. 3283–3288.
22. Davison, A.; Hinkley, D. *Bootstrap Methods and Their Applications*; Cambridge University Press: Cambridge, UK, 1997.
23. Nicholson, W.L. On the normal approximation to the hypergeometric distribution. *Ann. Math. Stat.* **1956**, *27*, 471–483. [CrossRef]
24. Hemelrijk, J. The hypergeometric, the normal and chi-squared. *Stat. Neerl.* **1967**, *21*, 225–228. [CrossRef]
25. Molenaar, W. Simple approximations to the Poisson, bimomial and hypergeometric distributions. *Biometrics* **1973**, *29*, 403–407. [CrossRef]
26. Patel, J.; Read, C.B. *Handbook of the Normal Distribution*; Marcel Dekker: New York, NY, USA, 1982.
27. Childs, A.; Balakrishnan, N. Some approximations to the multivariate hypergeometric distribution with applications to hypothesis testing. *Comput. Stat. Data Anal.* **2000**, *35*, 137–154. [CrossRef]
28. Sandiford, P.J. A new binomial approximation for use in sampling from finite populations. *J. Am. Stat. Assoc.* **1960**, *55*, 718–722. [CrossRef]
29. Miller, R. *Simultaneous Statistical Inference*; Springer: New York, NY, USA, 1981.
30. Johnson, N.; Kotz, S.; Kemp, A. *Univariate Discrete Distributions*; Wiley: New York, NY, USA, 1992.
31. Corrado, C. The exact distribution of the maximum, minimum and the range of multinomial/Dirichlet and multivariate hypergeometric frequencies. *Stat. Comput.* **2011**, *21*, 349–359. [CrossRef]
32. Oberhofer, W.; Kaufmann, H. Maximum likelihood estimation of a multivariate hypergeometric distribution. *Sankhyā Ser. B* **1987**, *49*, 188–191.
33. Boland, P.; Proschan, F. Schur convexity of the maximum likelihood function for the multivariate hypergeometric and multinomial distributions. *Stat. Probab. Lett.* **1987**, *5*, 317–322. [CrossRef]
34. Chèneby, J.; Ménétrier, Z.; Mestdagh, M.; Rosnet, T.; Douida, A.; Rhalloussi, W.; Bergon, A.; Lopez, F.; Ballester, B. ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* **2019**, *48*, D180–D188. [CrossRef]
35. Nassrallah, A.; Rougée, M.; Bourbousse, C.; Drevensek, S.; Fonseca, S.; Iniesto, E.; Ait-Mohamed, O.; Deton-Cabanillas, A.F.; Zabulon, G.; Ahmed, I.; et al. DET1-mediated degradation of a SAGA-like deubiquitination module controls H2Bub homeostasis. *eLife* **2018**, *7*, e37892. [CrossRef] [PubMed]
36. Howe, K.L.; Contreras-Moreira, B.; De Silva, N.; Maslen, G.; Akanni, W.; Allen, J.; Alvarez-Jarreta, J.; Barba, M.; Bolser, D.M.; Cambell, L.; et al. Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* **2020**, *48*, D689–D695. [CrossRef] [PubMed]
37. Zhu, B.; Zhang, W.; Zhang, T.; Liu, B.; Jiang, J. Genome-wide prediction and validation of intergenic enhancers in arabidopsis using open chromatin signatures. *Plant Cell* **2015**, *27*, 2415–2426. [CrossRef] [PubMed]