report

Nathan Lecouvreur

June 7, 2022

43628

1 Abstract

The DamID is a well-known method used to study the interactions between proteins and DNA using a Dam methylase fusion with a protein of interest. This technique evolved alongside the advances in DNA analysis, adopting high throughput sequencing to detect the methylation. But it still kept its peculiar enzymatic digestion producing fragments between 2 methylated GATC motifs next to each other. Using this particularity, we developed PseuDam, a new pipeline to analyze DamID data in a fast and innovative way. PseuDam takes advantage of these GATC fragments to use fragment-based pseudomapping approaches using GATC fragments. The use of these tools allows us to take advantage of the well-studied RNA-seq analysis to normalize and obtain fragment-wise statistical significance. To test this new approach we compared PseuDam to a reference pipeline on a published DamID dataset. To accompany this new approach we demonstrate the use of an adaptive binning strategy to fit the studied mechanism. We show that PseuDam is a faster and innovative way to look at DamID data and associated with versatile binning approaches can extract multi-scale statistical formations.

2 Introduction

The nucleus is a key element of eukaryotic cells, hosting the cell's DNA. All the mechanisms that structure, express and regulates it is extremely important for the cells to function. Most of these mechanisms are based on DNA-protein interactions. To study them, many innovative experimental methods have been developed. One of the most widely used techniques is the Chromatin Immuno Precipitation (ChIP)[9]. This technique consists in cross-linking all the DNA-proteins interactions and in selecting the proteins of interest using specific antibodies. The DNA attached to these proteins is then purified and can be studied using various methods, such as qPCR, for a very locus-focused analysis, microchips for a wider approach and now sequencing allowing to perform genome wide interaction profiles. The ChIP allows a very good characterization of the DNA-protein interactions, but, as most techniques used in biology, it is not free of bias. One of its biggest lies in the use of cross-linking agents to fix the DNA-protein interactions, which can lead to the misinterpretation of interactions. It can happen in the case of "phantom peaks" observed on several active promoters(D. Jain et al.[10]. These peaks are due to non-specific interactions that remain even after deleting the whole gene.

To bypass this limitation, other techniques have been developed, providing cross-linking free ways of studying DNA-protein interactions. One of these techniques is the DamID [18]. This technique is based on the use of the Dam methylase, a bacterial enzyme in charge of methylating adenosines in GATC motifs in the bacterial genome. This type of methylation is absent from the eukaryote genomes, which means that if we express the Dam in a eukaryotic cell, the methylated adenosines that are observed can only result from the Dam's action. The eukaryotes are also deprived of an adenosine demethylase for GATC sites, meaning that the only source of dilution of the included methylation is through DNA replication. If fused to a protein of interest, Dam will methylate around the protein's fixation site. But one issue is that the fused Dam can also randomly methylate DNA generating background methylation. This means that all the methylated GATC sites are not necessarily fixation sites for the protein of interest. Originally, to quantify GATC methylation, DpnII, a restriction enzyme that cuts unmethylated GATC sites was used. If a region was harder to cut it meant that it was methylated by Dam[18]. Then a new enzyme of the same family was used, DpnI. This enzyme also cuts GATC sites, but only those that contain a methylated adenine. If the harvested DNA is a treated with DpnI, cuts will be generated between methylated GATC motifs. This allows the study of the release of specific fragments between 2 GATC sites using PCR with specific primers [17]. To quantify the background methylation, a control containing only the Dam is also performed. This allows a separate quantification of the background methylation, and the signal from the dam fusion experiment can be compared to the Dam only experiment.

Following the advances in biology, the DamID analysis moved to the Microarray technology, allowing the study of a wider range of genomic regions by using chips coated with different cDNA and gDNA to identify the fragments cut by DpnI [17]. Jointly with the development of Microarrays techniques, the resolution of the DamID became higher and higher. It allowed the discovery of numerous interactions that were not detected using classical cross-linking methods, such as the mechanisms of Hairy transcriptional repression [3]. This microarray analysis of DamID experiments also set the bases of the analysis of Dam datasets which consists in displaying the log2 ratio between the dam-protein fusion and the dam-only control signals. This representation, while being very simple, was very well suited to show the methylation signal. Then, with the development of high throughput sequencing, new ways of measuring GATC methylation have emerged, the main one being DamIDseq [20]. In this protocol, DpnI and II are now used in combination (Fig.1.A). DpnI is first used to cut all the methylated GATC sites in the genome, generating fragments of random length between 2 GATC sites. Adapters are then introduced, containing a sequence that matches PCR primers, allowing to amplify the fragments. But before this PCR step, a second digestion is performed, this time using DpnII, to cut all the unmethylated GATC sites located in between methylated GATC sites that were previously cut. This extra digestion step allows that only the fragments corresponding to 2 adjacent methylated GATC sites will be amplified during the round of PCR. After this amplification, a second round of adapters is then added to prepare the sequencing library. Finally, after an optional round of sonication, amplified fragments are sequenced with NGS. The coverage of a fragment then gives information about the methylation of a site. This new method rapidly took over the DamID field, allowing the de-novo discovery of even more interactions, such as finding new expressed genes with a Dam-rpb6 fusion. Rpb6 being a sub-unite of the RNA polymerase II (polII), the Dam-fused protein will methylate regions that are transcribed. This fusion protein can even be put under the control of a tissue-specific promoter to study transcription in a chosen tissue [13]. These advances also opened new possibilities for innovative methods of studying the whole genome.

reference the figure

One of the techniques that emerged from this advances is the DamC [15]. It aims at observing the 3D contacts of one DNA region, also called the viewpoint, with the whole genome. These technique is very similar to the Circular Chromosome Conformation Capture (4C) technique which allows the detection of 3D contacts with the viewpoint [16]. But it requires a crosslinking step to fixate the DNA-DNA interactions with the region followed by an enzymatic digestion and ligation of the DNA fragments that were crosslinked together. Then an inverse PCR allows the amplification of the fragments of interest and their study via qPCR, microarray, or sequencing. Contrary to this method, the DamC doesn't require these two steps. However, it instead requires the insertion of a Dam fusion protein, which this time is a Dam-TetR fusion. TetR is a receptor that attaches to the DNA on a 19bp motif, TetO. The second requirement for this technique is to insert in the desired viewpoint a succession of TetO sites to allow the fixation of the Dam-TetR fusion. When the DamTetR is attached to TetO sites, it will methylate everything that comes in close proximity to the viewpoint, staining it so the interactions can be visualized. The experiment is treated very similarly to the usual DamIDseq experiments, except that the digestion with dpnII is skipped to reduce the bias toward regions with a high concentration of GATC sites. To extract the DamC signal from the usual free-roaming Dam methylation background, a new type of control is performed. It is based on the property that TetO-TetR interactions can be regulated by Doxycycline (Dox). TetR only binds to the TetO sites in presence of Dox. In Dox- conditions, TetR is not preferentially attached to TetO, therefore producing the same kind of results as a Dam-only control, but with the same construct as in the fusion condition. The resulting methylation profiles are related to contact frequency with the viewpoint and give very similar results to standard 4C [15]. This new technique to measure contact probabilities between chromatin regions without the use of crosslinking agents opens new venues on the subject.

Recently, the groups of Daniel Jost and Gaël Yvert at LBMC aim at developing a Dam-C protocol to investigate its applicability and evaluate its interest in future studies regarding nuclear organization and epigenetic regulation (see Discussion). As a proof of concept, they want to transpose the DamC method into S. cerevisiae and develop a quantitative framework to rationalize the experiments and extract quantitative information on the 3D contact frequencies. Prior to my internship, a first set of calibration experiments was performed in Yvert's group and a theoretical framework was developed in the Jost's group. The goal of my internship was thus to develop (1) bioinformatic tools to process the generated raw data ; (2) methods to test the applicability of the theoretical framework. In particular, since the bioinformatic and statistical analysis of DamID data

Say that you only have 1 replicate here

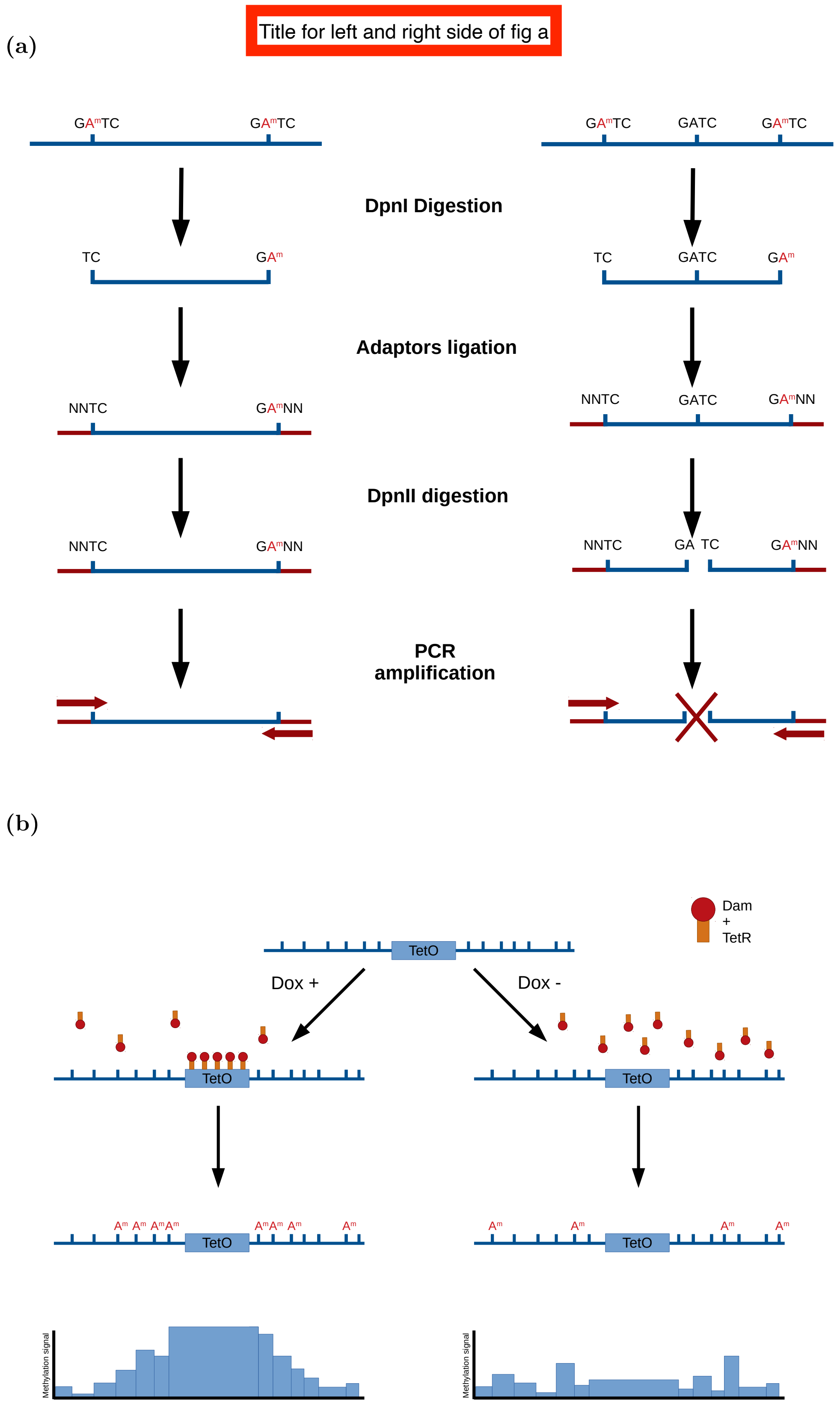


Figure 1: ***The Dam techniques*** *(**A.**) Schema of the processing of the gDNA for DamIDseq experiments. After harvesting and purifying the DNA, a DpnI digestion is performed to cut the methylated GATC sites followed by the ligation of adapters. The fragments are then digested using DpnII, cutting only the fragments that contain a GATC site that remain in the middle of some fragments. A PCR is then performed using primers that hybridize to the adapters. Only the fragments that did not get digested by DpnII will be amplified, resulting in fragments located between 2 nearby methylated GATC sites. (**B.**) Schema describing the design of the DamC experiments on S.cerevisae. TetO sites have been inserted in the Leu2 motif. This will allow, in presence of Dox, the Dam-TetR to attach to the TetO motifs. In the control condition, Dox is not added to the medium and the fusion won't be able to attach to the TetO motifs. The Dam will then methylate radnomely the genome creating background methylation. Whereas in the test condition, Dox is added and the Dam fusion will attach to the TetO sites and methylate the GATC sites around the Leu2 locus in addition to the background methylation.*

did not really evolve since the time of the Micro-array DamID, except for new ways to normalize the data [14], we first decided to take a new look at the processing of the DamID data in general and to develop a new DamIDseq pipeline allowing to efficiently process and analyze them. In the first part of my report, I will explain the concept of the new analysis that we implemented and demonstrate it on published data. In the second part, I will describe the results obtained in the DamC experiments. And in the third part, I will describe the model that was developed to describe these DamC experiments and how we tested it. Then I will conclude and discuss the future perspectives of my work.

3 Results

3.1 DamID pipeline (PseuDam ?)

The enzymatic treatment that the DamID samples provide particular properties to the fragments that are generated. The first DpnI digestion will create fragments that are located between 2 methylated GATC motifs and separated by varying distances. After the addition of the adapters, the digestion by DpnII will eliminate all the fragments that had unmethylated GATC motifs in between the methylated ones. This means that the fragments that we obtain after a PCR amplification using the adapters will all be located between 2 adjacent methylated GATC sites. This unique property implies that to be relevant, our reads have to be located in defined genomic intervals. Previously, this was only used in later steps of the analysis to define the resolution of the methylation signal. Moreover, despite this unique property the statistical analysis that was performed after the mapping short definition of mapping was very similar to what is used in the analysis of ChIP data. The whole statistical analysis was done by binning the genome with a defined length, ignoring the attribution of the reads to specific GATC fragments. This is why we decided to use a different approach to take full advantage of the properties given by the enzymatic treatment. Since we have intervals where our reads are supposed to map, we do not have to locate exactly where they are located in the genomic intervals defined by adjacent GATC sites, as long as they are fully in these regions. A parallel can then be done with RNAseq data since the reads from these experiments are also supposed to be located in defined regions (genes) to carry information. This property is at the center of the modern analysis of RNAseq data. This parallel with RNAseq allows us to use all the tools that were developed over the years to map and analyze RNAseq data.

3.1.1 Mapping the DamID reads

First, to map DamID sequencing reads, we use Kallisto (article Kallisto)[4]. It is based on the pseudo-mapping of reads to the given fragments of a genome. It is a new approach to RNA sequencing data, which speeds up the processing time. This increase in speed allows the addition of The computation of the uncertainty around the count estimate a Bootstrap re-sampling step, allowing the computation of the uncertainty of the counts by resampling the dataset with replacement and performing the analysis on the reconstructed datasets. To gain that much time, the pseudomapping ~~removes~~ skip the exact mapping of the reads ~~to~~ on the genome. Briefly, ~~it instead bases its approach on computing~~ pseudo-mapping computes a compatibility score for each read on a simplified version of the genome obtained by transforming each transcript into a succession of nodes for which the likelihood of the read to belong to it will be computed. This allows a fast and accurate estimation of the transcripts the read could have originated from. This method also removes the need for the extra counting step that was needed ~~to attribute each read to a transcript.~~ estimate the abondance of a given transcript

small definition of nextflow

We built a Nextflow pipeline to analyse DamID datas with this new approach, allowing us to gain in ease of use and reproducibility (ref nextflow)[8]. It follows ~~a very simple~~ mai non c'est super ce que tu as fait workflow : first a quality check and removal of the adapters is performed using fastp (ref fastp). In parallel, a ~~map~~ reference of all the possible fragments between 2 adjacent GATC sites is generated. This ~~map~~ reference is then used to create a pseudo-transcriptome of these fragments. This pseudo-transcriptome can then be indexed and the reads pseudo-mapped on it using Kallisto [4]. This produces abundance files containing the pseudo count of each GATC fragment. It also produces another abundance file, which contains the results of the bootstraps that were performed during the pseudo-mapping.

3.1.2 Normalizing and analyzing the data

The bootstraps information can then be used by another algorithm, Sleuth [4], to normalize the samples ~~altogether~~ by computing a size factor on all the pseudo-counts of each experiment. For example, to visualize the effect of such normalization, we treated a DamIDseq dataset constituted of 2 Dam-fusion samples as test samples and 2 Dam-only control samples. This experimental structure is the bare minimum to efficiently use sleuth. Sleuth Most RNAseq count processing packages such as DESeq2 or sleuth Sleuth require at least 2 replicates for each condition. 2 replicates might be enough to run DESeq2 but normalization is very poor and barely visible for the datasets that we tested. Sleuth, by using the bootstraps from Kallisto manages to provide a very good normalization with only 2 replicates for each condition and still provide a similar downstream statistical analysis (**Fig.2.B-C**). When we compare the raw counts of each GATC fragment between the dam fusion experiment and the raw counts of a Dam only experiment, we notice a distinct shift of the densest part of the fragments from the identity line (**Fig.2.B**). But when we compare the normalized counts of the same experiments, we notice that the size factor heavily reduced the shift and that now the dense part is on the identity line (**Fig.2.C**) which is what is expected since XXX. Moreover, we can still see the outliers that differ from the control and they are now even further from the identity line. Using this normalization we seem to be able to accurately account for library size.

Tu n'as pas défini size factor et library size

After this normalization step, still using Sleuth, data are further transformed into log2 to compare the Log Fold Change (LFC) of the test and control conditions, a pseudocount of 0.5 is also added to escape computing logarithms of 0. These transformed data are then fitted with an error estimation model and a Wald test is performed on the fitted data. This tests the statistical relevance of the difference between the control and test samples by testing the difference to 0 of their LFC. The output of this step is a table presenting the b value, an estimator of the LFC between the Dam-fusion and Dam-only conditions computed during the Wald test, the pvalues associated with the LFC of each fragment, and the qvalue which is the adjusted pvalue using the Benjamini-Hochberg procedure. This gives to PseuDam new statistical information at the highest resolution achievable by Dam, giving a way to estimate the significance of the observed LFC.

3.1.3 Binning the data

DamID data can be very noisy and wield a bias toward the small GATC fragments that were eliminated during the library preparation step or not sequenced due to their small size. In PseuDam this induces small GATC fragments to often contain 0 reads and to be filtered out during Sleuth's filtration step that removes from the analysis all the fragments in which 53% of the replicates don't have at least 5 reads. This cause many holes in track representation (**Fig3.C**). To smooth the data and obtain broader information, a binning can be performed on the data. Here we decided to use an approach inspired by Serpentine (L Baudry et al. 2020)[2]. This binning revolves around the processing of both the Dam-only and Dam-fusion conditions in parallel to establish the bins. The binning algorithm starts at the beginning of each chromosome and iterates over the GATC fragments adding up the pseudocounts in each condition. The bins are terminated using 3 parameters, θ, the highest number of pseudocounts one condition can have before the end of the bin is set. ϵ, the highest number of pseudocounts both conditions can have at the same time before finishing the bin. The third is the maximum length of the bin, to stop the bins from carrying counts over too many empty GATC fragments. The different parameters are to be set depending on what is observed. For example, very punctual signals such as transcription factors require very small values of parameters to obtain very small bins and keep a good precision of the signal. Whereas wider signal like LaminB requires larger values of the parameters to obtain wider bins that will give a better estimation of the global trend in the region. To process the results of the binning, due to the high speed of PseuDam, we re-process the reads by replacing the DpnII pseudo-transcriptome with the newly define bins in a bed file. This allows the processing of qvalues and LFC for the newly created bins. This binning approach is just one of many binning strategies that can be applied with PseuDam and more specific binning strategies can be used, such as using an annotation file containing the genes of the model organism with an offset on the nearest GATC sites to analyze TaDa approaches, giving a per-gene LFC and qvalue, similarly to an RNA-seq approach.

3.2 Testing the pipeline on existing data

To test this new approach on DamID data, we decided to use a very recent dataset using the Tada technique, a variation of the DamIDseq technique that aims at identifying new expressed genes in a tissue-specific manner by using a Dam-polII fusion protein under the control of the promoter of an active gene in the tissue [11]. The study focuses on 2 tissues: the seam, using the srf-3i1 promoter to express the Dam-rpb6 fusion; and the hypodermis, which is targeted by adding using the dpy-7syn1 promoter. Then worms are harvested at 2 developmental stages, L2 and L4 to also study the temporal difference in gene expression. The DamIDseq data from their experiments were originally processed using the reference DamID pipeline for processing DamIDseq [14]. This pipeline follows a classical approach to sequencing high throughput sequencing data, aligning the genes to a reference genome, then counting the reads in each GATC fragment and performing a normalization. The LFCs are then computed, comparing one Dam-only control replicate to all the Dam-fusion replicates. This forced us to launch the pipeline 2 times to perform the LFC over all the replicates. For each fragment, the mean of the LFC was computed between the LFC of the replicates. To compare our new approach to the reference one, we also processed all the experiments using PseuDam. The processing time of the data was faster for PseuDam (9 minutes vs

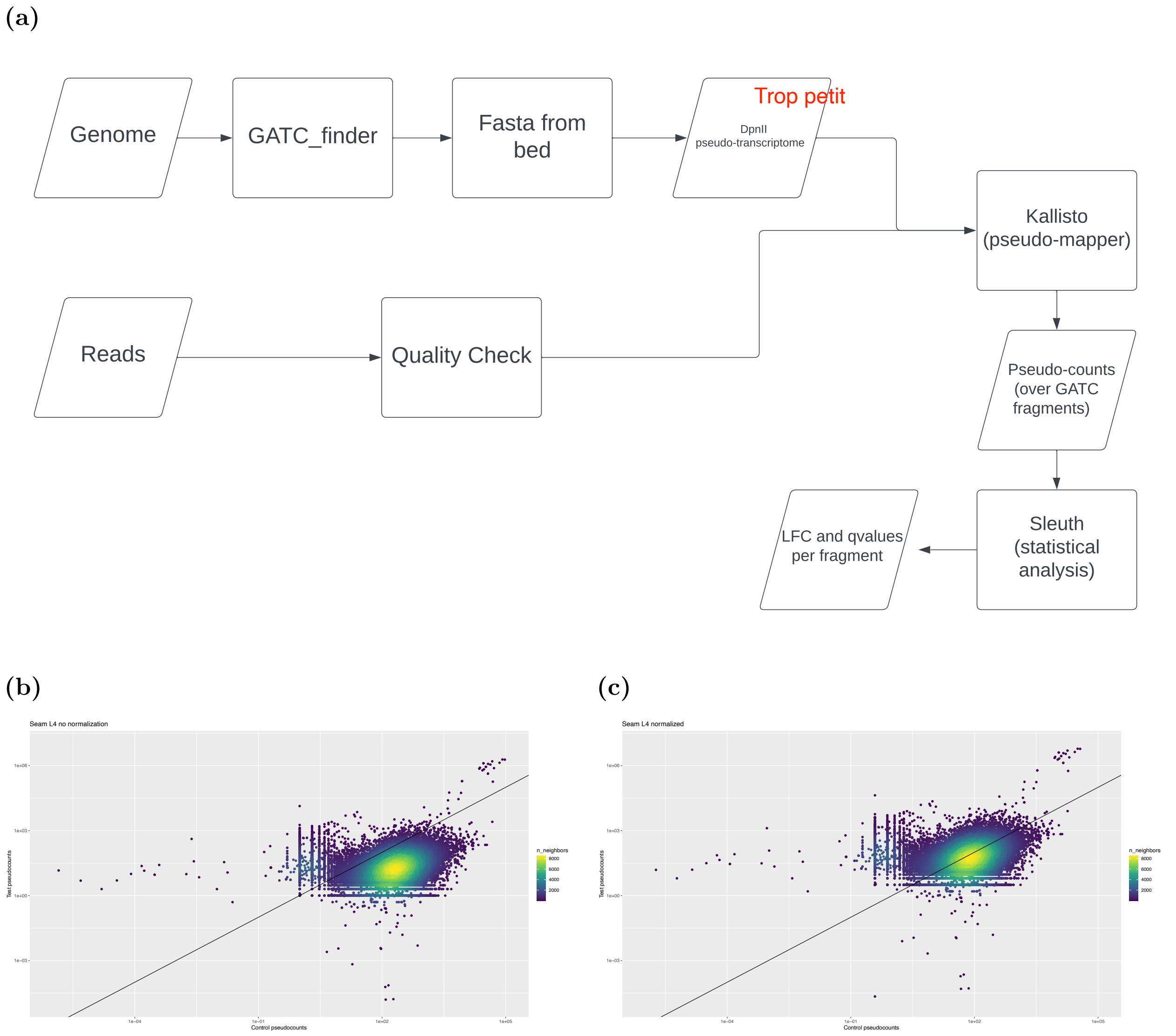


Figure 2: ***Pseudomapping pipeline,*** *(**A.**) Workflow of the pipeline. The genome is processed to create a map of the DpnII cutting sites which are then transformed in a pseudo transcriptome of the DpnII fragments that is then indexed. At the same time the reads' quality is assessed and the remaining adaptors are trimmed. The reads are then pseudo-aligned to the pseudo-transcriptome using Kallisto which outputs pseudo-counts for each DpnII fragment. This output is then analysed using Sleuth to produce the Log2 Fold Change and associated qvalue of the Dam fusion and Dam only conditions for each fragment. (**B.-C.**)Scatter plot of Dam-only vs Dam-fusion conditions (**B.**) raw counts and (**C.**) normalized counts. The color represents the ~~k mean~~ points density*

2*45 minutes to process 4 samples) but this simple time estimation is biased due to Nextflow's high parallellization capability. To cope with this bias we computed the number of CPU hours used for each computation. The reference pipeline used CPU hours while our pipeline only used 1.303 CPU hours to compute the same files. The alignment rates were similar between the pseudomapping and classical mapping approachs, being around 2% difference between them.

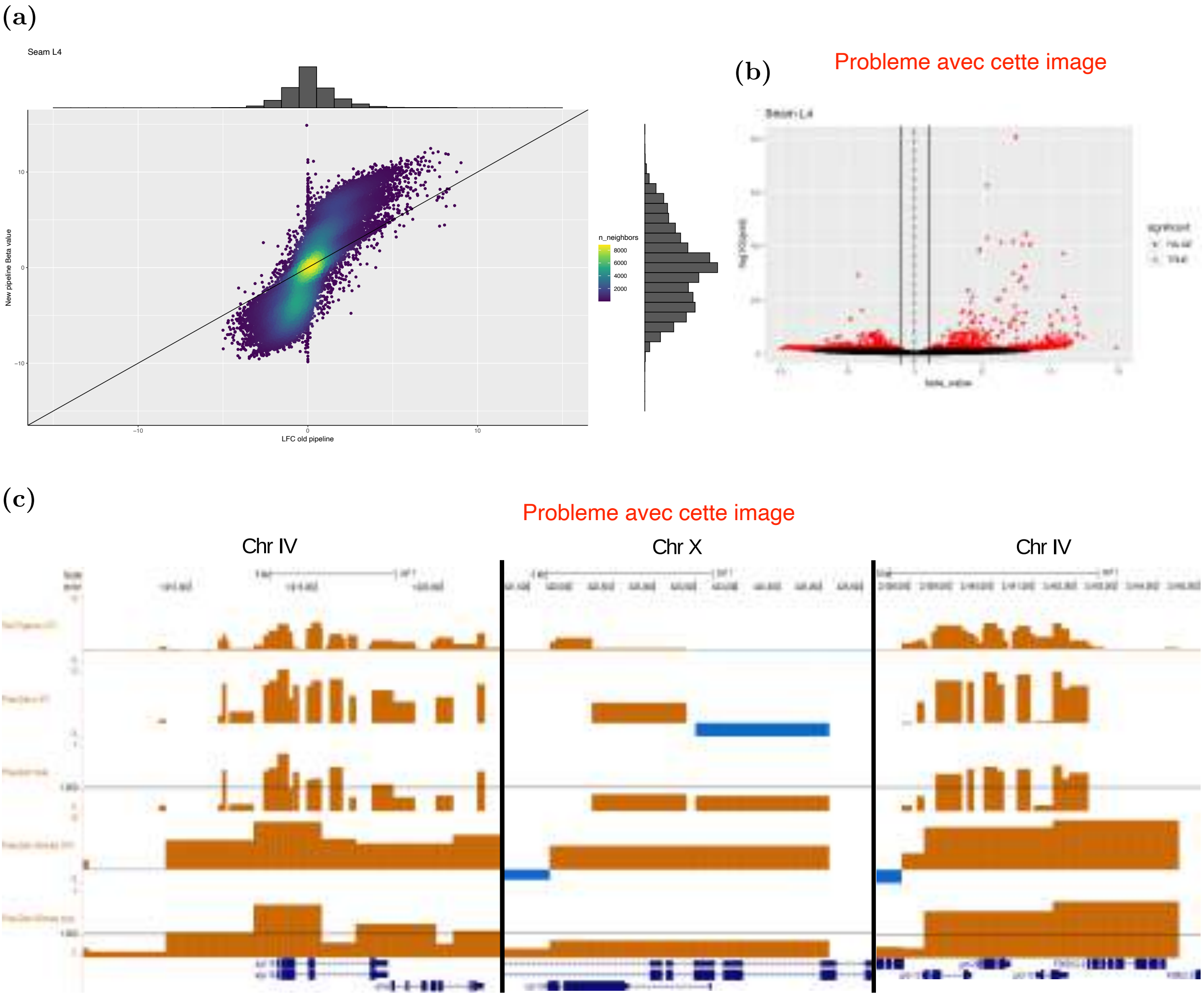
When looking at the trend of the LFC values of the experiments analysed with the 2 pipelines, we can see that it is globally similar between the two pipelines, the highest density beeing situated on the equality line **(Fig.3A)**. But we can see that the LFC computed with PseuDam are distributed in a wide range, with LFC ranging from 13 to -10. While the reference pipeline produced narrower distribution only ranging from 10 to -5. But while looking at the actual trend of the LFC on genomic regions, we can see that they are very similar between the reference pipeline and PseuDam(two first lines in Fig.3C). Egl-18 is known to be expressed in the seam at stage S4 and we can observe this expression through both PseuDam and the reference pipeline(**Fig3.C** left panel) (L. Gorrepati et al. 2013). The same goes for grd-3, expressed in the same way in the seam at L4 [1]. For genes that are not expressed in the seam at L4, such as col-19, we also observe LFCs closer to 1 but also negative values (middle panel of **Fig3.C**)[19].

Our new pipeline also provides new ways of analyzing DamID data. For example, we can visualize the track of the qvalue corresponding to each fragment's LFC (third track in **Fig.3C**). This gives us a very useful information on the significativity of the changes that we observe. More generally, we can use our output information to generate a volcano plot for each experiments **(Fig 3.B)** where we can easily locate outliers (high LFC and low q-value).

The signal that we obtain with the new pipeline seems to have many undefined values due to Sleuth's filtering out the fragments for which 47% have less than 5 reads. To better identify the trend over the genes we chose to perform a binning of the data, we applied our Serpentine-inspired binning on the data, creating an alternative pseudo-transcriptome that we used to re-process the data. We test multiple parameters for the binning and noticed that higher the values of θ and ϵ were, higher was the ratio $\frac{qvalue<\alpha}{qvalue_{tot}}$. For the Tada dataset, we chose to use $\epsilon = 10$ and $\theta = 100$ and a maximum fragment length of 3kb, keeping the $10x$ ratio between θ and ϵ advised by L. Baudry et al. . We found this values to suit best the polII signal, giving a gene resolution (5th track of **Fig3.C**). But in the case of some genomic regions with close but very differently expressed genes, this binning can overflow from highly expressed genes to nearby non-expressed genes (last 2 tracks of the right panel of **Fig3.C**). This overflow can even reach the $\alpha = 0.05$ threshold and induce false-positive expressed genes in the analysis (5th track of the left panel of **Fig3.C**). It also allows to strengthen the qvalues that we observe in some regions and make some regions reach the threshold $\alpha = 0.05$, which they were not able to do before.

3.3 Using the pipeline on our data

Noémie Quesnel performed DamC-type experiments as a proof of concept in yeast. As in the original paper, we used a Dam-TetR fusion that will attach to a TetO region constituted of 199 repetitions of the motif inserted in the Leu2 region. The Dam-TetR gene is under the regulation



Probleme avec cette image

Probleme avec cette image

Figure 3: ***Comparison of the pipelines,*** *Analysis of experiments using a Dam-rpb6 construct expressed in the seam at stage L4 in C.elegans. (**A.**) Volcano Plot of the Beta values and -log10 of the qvalues for each GATC fragment in C.elegans genome. Red points have a qvalue higher than the threshold alpha = 0.05.(**B.**) Scatter plot of the LFC obtained with the reference pipeline vs the beta values of the new pipeline for each GATC fragment. The colors represent the k-mean density of points and the histograms represent the distributions on the y and x axis.(**C**) Genome tracks visualisation of the genomic region surrounding multiple genes. Left panel shows the egl-18 gene region, middle panel displays the col-19 region and right panel displays the grd-3, grd-10, grd-13 region. The first track displays the LFCs obtained with the reference pipeline, the second track shows the LFCs obtained with PseuDam, third track displays, the qvalues associated to the PseuDam LFCs. The Fourth track shows LFCs for a binned analysis with the serpentine approach and the fifth track displays the corresponding qvalues.*

of a Gal1 promoter to control its expression playing with galactose concentration. This type of DamC experiments not being described in Saccharomyces Cerevisiae, we had to test multiple concentrations of Galactose (Gal). We also varied the concentrations of Doxycycline (Dox), a drug that enables the fixation of TetR to the TetO motifs. 2 control conditions were also performed to extract the signal, one Dam-only control and one experiment without Dox so TetR is not recruited at TetO. The different Gal conditions aim at estimating the required level of expression of the Dam fusion. Indeed, the gal concentration (and subsequent Dam fusion) should be high enough so the methylation signal is measurable, but not too high such that random methylation events are not too frequent. We used the classical DamIDseq protocol to treat the harvested DNA using the DpnII and DpnI digestions. In addition, GATC fragments were sonicated to fit the paired-end library generation. The paired-end sequencing is particularly useful for DamIDseq experiments since if a paired-end fragment would overlap a GATC fragment the paired-end reads couple will be eliminated. These overlapping paired-end fragments have to be eliminated since they most likely result from undigested fragments and do not carry valid information.

To process our data, we have to add an extra step to our data processing. Since our experiments were only a proof of concept, only one replicate was performed for each condition. This means that is not possible to pass this datas throught sleuth after the pseudomapping by kalisto because it requires at least 2 replicates for each condition. To still be able to analyse our data and perform a normalisation, we splitted the fastq files containing the reads for both experiments and the controls. This did not create real replicates, but allowed us to process the data through sleuth to normalize our data.

When looking at the log fold change at the fragment level, we can observe that the signal is very noisy on all the Gal and Dox conditions **(Fig.4.A)**. The LFC changes widely and the TetO region methylation is almost indistinguishable without zooming in. When we do so, we observe a distinctly LFC peak where the TetO sites have been inserted but there exist nearby peaks also with high LFC. To reduce the noise of the data, a binning was performed using an ϵ of 10 and a θ of 100. With the binning the methylation peak in the TetO region is now noticeable while looking at the whole chromosome III. When we focus around the region where the TetO motifs are inserted, we can further oberve the effects of the binning, assembling mutliple peaks in the TetO region into higher LFC bins compared to the neighboring region.

When observing the other chromosomes, the LFC peaks that we observe appear to be random (**4.E**). To estimate if such peaks may be related to actual long range interactions with the Leu2 region where the TetO array was inserted, we look at 4C-like data generated from micro-C data of S. Cerevisiae [6] (see Materials and Methods). This pseudo-4C data reflects the interactions of the 20kb region around the Leu2 gene with all the yeast's genome.

Version avec circos : The mean of the interactions with the region is then computed with a bin size of 400 bp. The interactions that passed the threshold can then be represented as links on the circos plot (**4.E**). We can see that the 4C interactions don't seem to follow the peaks of LFC. This confirms the observation that our data are very noisy and need to be reproduced with more replicates and a finer control of the Dam level.

Version sans circos : When we look at the 4C signal observed between the region and the whole genome, we do not see an association with the LFC peaks that we observe on the DamC data, binned or not. This confirms the observation that our data are very noisy and need to be reproduced with more replicates and a finer control of the Dam level.

3.4 Modeling Dam-C experiments

3.4.1 The model

To better understand the relation between methylation profile (DamID data) and 3D contact frequencies (4C-data), and potentially extract directly 4C-like information from the DamID signal, a mathematical model was designed in the team. Indeed, while, in the original DamC paper, Redolphi et al. already proposed a quantitative framework,the underlying hypotheses used to derive it might not be justified and valid in the context of our experiments[15]. In our model, we added more details such as we included the notion of hemimethylation of the GATC cites since the hemimethylation may affect the DpnII cutting step (Melissa A. Calmann et al. 2003)[5] by reducing its cutting affinity. The model also accounts for structural transition between a looped state where the TetO region is in close proximity with a GATC site and an complementary, unlooped state (**5.A**). In the looped state, methylation occurs in multiple steps: one Dam bound at TetO interacts with the

GATC site and then (hemi)methylate it (**5.B**). In the unlooped state, methylation of the GATC site is possible by the freely diffusing Dam-TetR fusion in the nucleoplasm (**5.C**). This model can be analytically solved and led to different equations linking methylation rates in presence or absence of Dox, the contact probability P derived from the model's state switching rates and a factor Y that describes the relationship between number of TetO motifs and dissociation constants with the TetO and GATC sites. (**5.D**). From these formulas, it is possible to extract a relationship between the methylation states in presence of Dox, S and absence of Dox B and the contact probability between TetO and the locus. We can use these values to estimate P, the contact probability with the TetO site in presence of Dox. To do so we use 2 different formulas depending on the regime of αT, T being the time of the experiment after the introduction of Dox. in the case where $\alpha T << 1$ then we use $YP = \sqrt{\frac{S}{B}} - 1$ and in the case of $\alpha T \equiv 1$ we use $YP = \frac{S-B}{BY}$. This allows us to model most of the experimental conditions.

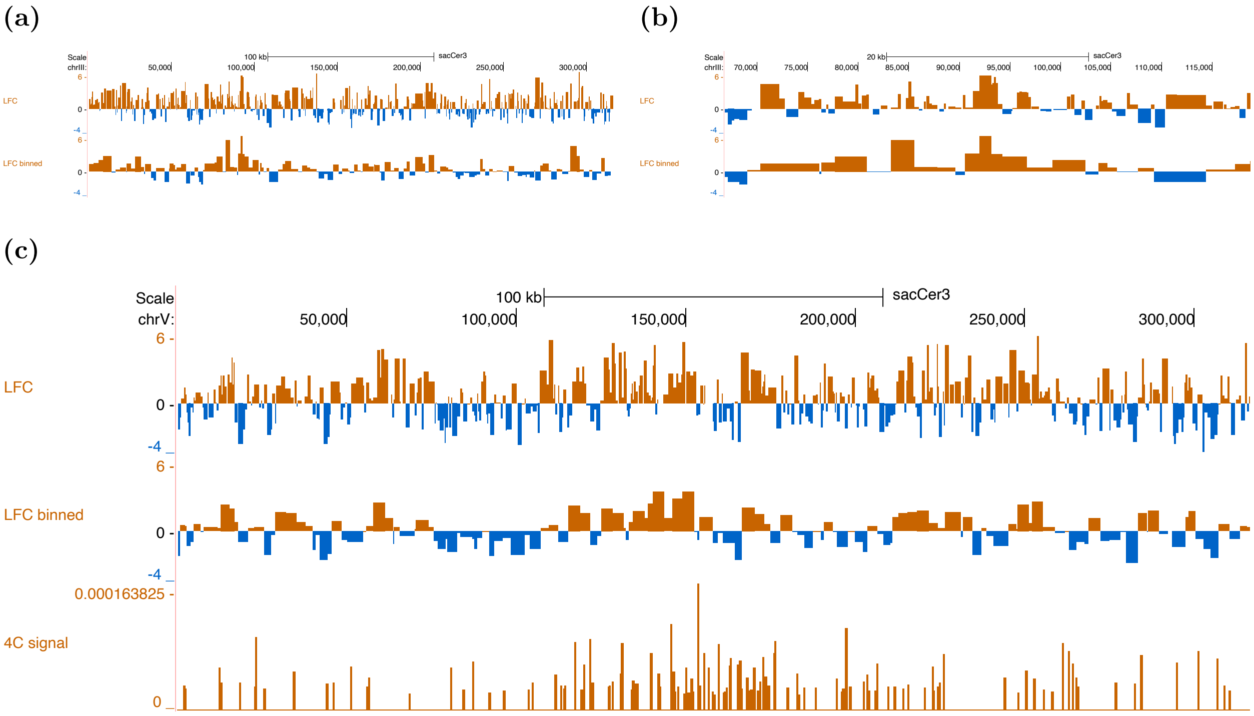


Figure 4: ***Analysis of DamC datas,*** *(**A-D**) Genomic tracks on the chromosome III of S.cerevisae for 0.005% of Gal and 1.58 g/mL of Dox. (**A.**) LFC on the whole chromosome 3 and (**B.**) zoom on the Leu2 region. (**C.**) LFC of the binning 10 100 and (**D.**) zoom on the Leu2 region. (**E.**) Circos plot of the chromosome V and Leu2 locus of the chromosome III. Outer track represents the LFC, yellow beeing positive LFC and blue negative LFC. Inner circle represents the -log10 of the qvalues of the LFCs. Links represent the pseudo 4C interactions between the Leu2 region and the chromosome V.*

3.4.2 Simulating the error for the models

To investigate the applicability of such model to real, noisy data, we studied the sensitivity of the model to an experimental error in the methylation profiles. To do so we produced simulations of the model over ranges of chromatin while modifying different parameters. We estimated the value of YP using different formulas. $YP = \frac{S-B}{BY}$ with S and B the methylation probability in presence and absence of Dox when A(x) is close to 1). We used a second formula, $YP = \sqrt{\frac{S}{B}} - 1$ to model values when A(x) is small compared to 1. These 2 formulas were used to compare P to $\frac{1}{x}$, a theoretical contact probability. To observe the behavior of this comparison, we computed for each formula $\frac{P_{calculated}}{Y\frac{1}{x}}$ over a genomics distance by setting parameters. We observe that the ”square” formula tends toward 1 even at short distances with $A(x) << 1$ but needs more distance to get close to 1 with $A(x) = 1$ (**Fig.5.B**). For the ”ratio” formula, we observe 2 distinct aspect of the curves depending on the A(x) value (**Fig.5.A**). With $A(x) << 1$ we can see that the curves do not tend toward 1 but toward 2 (**Fig.5.A** plain curves). With $A(x) = 1$ and $[Dam]$ close to K_d^{GATC} the curve tends to 1 (**Fig.5.A** blue dotted curve). We also modeled the addition of an additive or multiplicative error to the model calculated with random function. We added this error to both S and B and used the ”ratio” formula to calculate. Then over 1000 simulations, we calculated the % of cells that met the thresholds and plotted it as a heatmap over the distance from TetO (**Fig.6.C-D**). For the additive error, as expected it is constant over the distance from TetO and its value changes makes the data noiser, since less simulations make it past the thresholds (**Fig.6.C**). For the multiplicative error, we can see this time and effect of the distance from TetO on the multiplicative error, the multiplicative error beeing bigger if S and B are larger **Fig.6.D**.

4 Discussion

4.1 advantages and limits of Pseudam

We showed that Pseudam offers a new fast and information rich way of analysing DamID type experiments. The pseudomapping speeds up multiple times the processing speed a the cost of the exact positioning of the reads. But due to the nature of DamID library preparation the reads position doesn't matter and only the fragment in which they are located in gives information on the 2 neighboring GATC sites methylation state. Pseudam also provides a new statistical information for each GATC fragment in the genome, allowing to easily identify fragments are region with increased methylation. With Pseudam we are now able to identify the methylated region much faster using the computed qvalues. With the addition of the our serpentine inspired binning approach, the resolution of the Dam experiments can now be adapted to the biological phenomenon studied. Indeed by varying the different parameters, the GATC sites can be pooled to smooth the signal at the desired scale. But the binning of DamID data can be used to bin GATC to span any regions, such as a genes by offsetting them to the nearest GACT sites. This approach would be particularly useful to analyse TaDa experiments, producing the same results as RNAseq data, even using the same statistical analysis.

4.2 Discussing our DamID datas

Using this new and innovative method, we aimed at producing a synthetic system in Saccharomyces Cerevisae to study important functions of regulatory proteins in the nucleus. We demonstrate here a first proof of concept by transposing the DamC method into S. cerevisae. The introduction of TetR and TetO reproduces a very simple reader-writer interactions, TetR reading the TetO motif and bringing Dam close enough to methylate. This first proof of concept will allow us to then fine tune the conditions of the experiment and produce more complex interaction networks by adding other readers, writers or bridgers. The results that we obtained are very noisy, making it very hard to extract the peaks surrounding the TetO instertion, and even harder to observe interactions with other regions. But the binning that we applied was very usefull to extract the peaks, but it was still hard to attribute the other peaks to either long range interactions or background noise. These DamC experiments still have some potential since we can see the main peak with the binning, meaning that with the right tunning of the parameters, it will be possible to engineer these DamC and observe the 3D contacts with the region.

4.3 model

5 Materials and Methods

5.1 DamID exepriments (short)

The DamID experiements were performed on Saccharomyces Cerevisae using culture and DNA purification protocols form the A. Piazza's team. To activate the transcription of the Dam-TetR construction, Galactose was added to obtain 0.05%, 0.0158% or 0.005% of galactose depending on the condition. At the same time, Doxycycline was added to activate the TetR binding to TetO with concentrations of 5 µg/mL, 1.58 µg/mL or 0.5 µg/mL. The cells were then harvested and the DamID protocol from Georgina Gómez-Saldivar et al., 2016 [] was used to treat the different conditions. The sequencing libraries were then prepared flowing the NEBNext's kit instruction. The libraries were then sequenced by NOVAGEN using the Illumina technology to generate 150bp paired end reads

5.2 pseudo-alignment part of the pipeline

The fastq files first undergo a quality control and adaptor removal step using fastp. Then they were pseudoaligned with Kallisto (Bray, N., Pimentel, H., Melsted, P. et al. Nat Biotechnol. 2016) to a fragmented fasta file of the yeast genome, generated with bedtools (Quinlan AR, Hall IM. Bioinformatics. 2010) from a bed file of all DpnII framgents in the yeast genomes. The bedfile is generated with GATC finder a homemade script which finds all the GATC sites in a given genome file and generates intervals between the GATC sites. In case of single end reads, the mean length of the reads is used as the fragment size in case the average length / standard deviation of the fragment is not accessible. Kallisto is run with the bias option and 100 bootstraps. For single end data, the mean length of the reads was used (150bp) instead of the mean length of the fragments and for the standard deviation of the length of the fragments 10 was used[4]. This was done since this information wasn't accessible. All the above steps are embed in a Nextflow pipeline to ensure the reproducibility of the analysis[8]. To perform the normalisation with sleuth on experiments that do not contain replicates, the fastq files were splitted using fastq splitter.

5.3 Reference pipeline

To compare our piepline we computed the datas from D. Kastanos et al. (2020) from the GEO accession number gse164775 using the perl pipeline from Owen J. Marshall et al. Bioinformatics 2015[11, 14]. To obtain the raw coverage of the GATC fragments from the BAM files generated by the reference pipeline, bedtools coverage was used with the bedfile containing the GATC fragments from C. Elegans' genome, generated by GATC finder.

5.4 statistical analysis

The abundance files generated by kallisto for each couple of Test/Dam-only conditions using sleuth [4]. A regression fit was performed on the control and tests of the different conditions. A 0 centered bilateral wald test was then performed on the fitted data. The pvalues are adujsted using Benjamini-Hochberg procedure and the beta value for each fragment is used as an estimator of the LFC of each GATC fragment.

5.5 4C maps generation

Due to the absence of existing 4C datas on the LEU2 region of S. Cerevisae, the 4C data used here were generated from microC experiments obtained from Lorenzo Costantino et. al., elife (2020) from the GEO accession number GSE151553[7]. The pseudo-4C was generated by computing the whole HiC matrix with 400bp resolution from the mcool file with a script from the team.The rows that corresponded to a 20kb region around the LEU2 locus were selected to obtain an interaction matrix of the region versus the whole genome. The mean contact probability in the studied region was then computed for each bin of the genome.

5.6 Plotting

The volcano plots were computed using sleuth and the bedgraphs of the LFC and qvalue for each GATC fragment were generated using a homemade R script and plotted using the UCSC genome browser (Kent WJ, et al. Genome Res. 2002)[12].

5.7 Test of the model

The effect of adding an error to the model was done by adding a multiplicative error proportial to the value of either the methylation rate with or without Dox. Or adding a fixed additive error to the methylation rates. The effect of different calculation formulas for the contact probalility using :
The contact probability

5.8 Data availability

The tracks used to generate the figures from ucsc are available as follow :
The following datasets were used :

C.Elegans datas hypodermis and seam data : (lien)
DamC experiments on the yeast :
(lien) MicroC data :

5.9 Code availability

The DamIDseq pipeline is obtainable at ......
main main

References

[1] Gudrun Aspöck et al. "*Caenorhabditis elegans* Has Scores of *hedgehog* Related Genes: Sequence and Expression Analysis". en. In: *Genome Research* 9.10 (Oct. 1999), pp. 909–923. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.9.10.909. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.9.10.909 (visited on 06/06/2022).

[2] Lyam Baudry et al. "Serpentine: a flexible 2D binning method for differential Hi-C analysis". en. In: *Bioinformatics* 36.12 (June 2020). Ed. by Alfonso Valencia, pp. 3645–3651. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btaa249. URL: https://academic.oup.com/bioinformatics/article/36/12/3645/5822880 (visited on 06/06/2022).

[3] Daniella Bianchi-Frias et al. "Hairy Transcriptional Repression Targets and Cofactor Recruitment in Drosophila". en. In: *PLoS Biology* 2.7 (July 2004). Ed. by Michael Levine, e178. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020178. URL: https://dx.plos.org/10.1371/journal.pbio.0020178 (visited on 06/06/2022).

[4] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (May 2016), pp. 525–527. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3519. URL: http://www.nature.com/articles/nbt.3519 (visited on 03/23/2022).

[5] Melissa A. Calmann and M. G. Marinus. "Regulated Expression of the *Escherichia coli dam* Gene". en. In: *Journal of Bacteriology* 185.16 (Aug. 2003), pp. 5012–5014. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.185.16.5012-5014.2003. URL: https://journals.asm.org/doi/10.1128/JB.185.16.5012-5014.2003 (visited on 06/06/2022).

[6] Lorenzo Costantino et al. "Cohesin residency determines chromatin loop patterns". In: *eLife* 9 (Nov. 2020). Ed. by Adèle L Marston, Jessica K Tyler, and Adèle L Marston. Publisher: eLife Sciences Publications, Ltd, e59889. ISSN: 2050-084X. DOI: 10.7554/eLife.59889. URL: https://doi.org/10.7554/eLife.59889 (visited on 11/04/2021).

[7] Lorenzo Costantino et al. "Cohesin residency determines chromatin loop patterns". In: *eLife* 9 (Nov. 2020), e59889. ISSN: 2050-084X. DOI: 10.7554/eLife.59889. URL: https://elifesciences.org/articles/59889 (visited on 05/24/2022).

[8] Paolo Di Tommaso et al. "Nextflow enables reproducible computational workflows". en. In: *Nature Biotechnology* 35.4 (Apr. 2017), pp. 316–319. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3820. URL: http://www.nature.com/articles/nbt.3820 (visited on 06/06/2022).

[9] D S Gilmour and J T Lis. "Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes." In: *Proceedings of the National Academy of Sciences* 81.14 (July 1984), pp. 4275–4279. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.81.14.4275. URL: https://pnas.org/doi/full/10.1073/pnas.81.14.4275 (visited on 06/06/2022).

[10] Dhawal Jain et al. “Active promoters give rise to false positive ‘Phantom Peaks’ in ChIP-seq experiments”. In: *Nucleic Acids Research* 43.14 (Aug. 2015), pp. 6959–6968. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv637. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv637 (visited on 05/30/2022).

[11] Dimitris Katsanos et al. “Gene expression profiling of epidermal cell types in ¡span class=”nocase”¿*C. elegans*¡/span¿ using Targeted DamID”. In: *Development (Cambridge, England)* 148.17 (Sept. 2021), dev199452. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.199452. URL: https://journals.biologists.com/dev/article/148/17/dev199452/272042/Gene-expression-profiling-of-epidermal-cell-types (visited on 04/14/2022).

[12] W James Kent et al. “The Human Genome Browser at UCSC”. en. In: (), p. 11.

[13] Owen J Marshall et al. “Cell-type-specific profiling of protein–DNA interactions without cell isolation using targeted DamID with next-generation sequencing”. In: *Nature Protocols* 11.9 (Sept. 2016), pp. 1586–1598. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2016.084. URL: http://www.nature.com/articles/nprot.2016.084 (visited on 05/27/2022).

[14] Owen J. Marshall and Andrea H. Brand. “damidseq_pipeline: an automated pipeline for processing DamID sequencing datasets: Fig. 1.” In: *Bioinformatics (Oxford, England)* 31.20 (Oct. 2015), pp. 3371–3373. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv386. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv386 (visited on 05/24/2022).

[15] Josef Redolfi et al. “DamC reveals principles of chromatin folding in vivo without crosslinking and ligation”. In: *Nature Structural & Molecular Biology* 26.6 (June 2019), pp. 471–480. ISSN: 1545-9993, 1545-9985. DOI: 10.1038/s41594-019-0231-0. URL: http://www.nature.com/articles/s41594-019-0231-0 (visited on 01/10/2022).

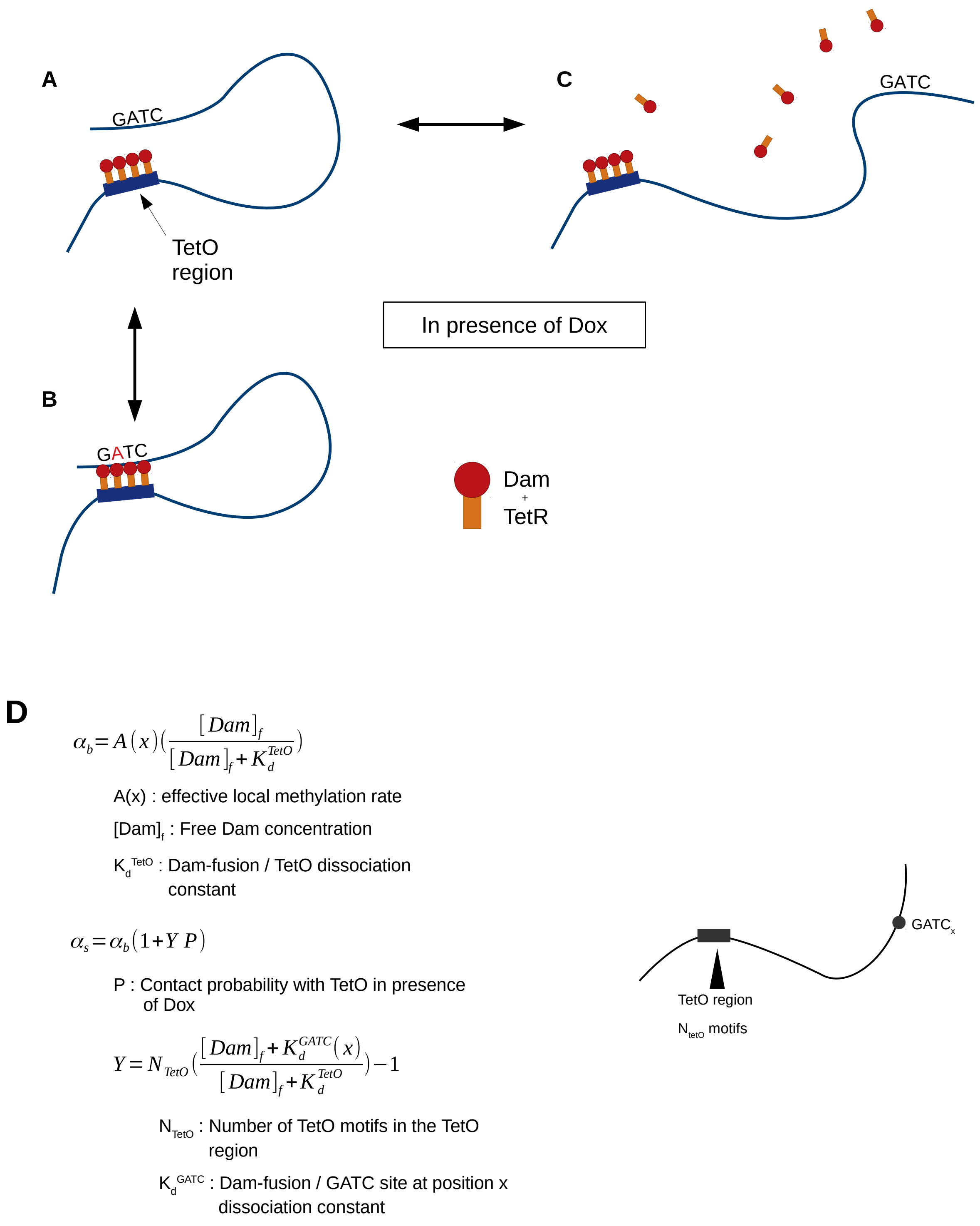
[16] Marieke Simonis et al. “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C)”. In: *Nature Genetics* 38.11 (Nov. 2006), pp. 1348–1354. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng1896. URL: http://www.nature.com/articles/ng1896 (visited on 05/31/2022).

[17] Bas van Steensel, Jeffrey Delrow, and Steven Henikoff. “Chromatin profiling using targeted DNA adenine methyltransferase”. In: *Nature Genetics* 27.3 (Mar. 2001), pp. 304–308. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/85871. URL: http://www.nature.com/articles/ng0301_304 (visited on 05/11/2022).

[18] Bas van Steensel and Steven Henikoff. “Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase”. In: *Nature Biotechnology* 18.4 (Apr. 2000), pp. 424–428. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/74487. URL: http://www.nature.com/articles/nbt0400_424 (visited on 05/11/2022).

[19] Melanie C. Thein et al. “*Caenorhabditis elegans* exoskeleton collagen COL-19: An adult-specific marker for collagen modification and assembly, and the analysis of organismal morphology”. en. In: *Developmental Dynamics* 226.3 (Mar. 2003), pp. 523–539. ISSN: 10588388, 10970177. DOI: 10.1002/dvdy.10259. URL: https://onlinelibrary.wiley.com/doi/10.1002/dvdy.10259 (visited on 06/06/2022).

[20] Feinan Wu, Brennan G. Olson, and Jie Yao. “DamID-seq: Genome-wide Mapping of Protein-DNA Interactions by High Throughput Sequencing of Adenine-methylated DNA Fragments”. In: *Journal of Visualized Experiments* 107 (Jan. 2016), p. 53620. ISSN: 1940-087X. DOI: 10.3791/53620. URL: http://www.jove.com/video/53620/damid-seq-genome-wide-mapping-protein-dna-interactions-high (visited on 05/27/2022).



$$\alpha_b = A(x)\left(\frac{[Dam]_f}{[Dam]_f + K_d^{TetO}}\right)$$

A(x) : effective local methylation rate

$[Dam]_f$: Free Dam concentration

K_d^{TetO} : Dam-fusion / TetO dissociation constant

$$\alpha_s = \alpha_b(1 + Y P)$$

P : Contact probability with TetO in presence of Dox

$$Y = N_{TetO}\left(\frac{[Dam]_f + K_d^{GATC}(x)}{[Dam]_f + K_d^{TetO}}\right) - 1$$

N_{TetO} : Number of TetO motifs in the TetO region

K_d^{GATC} : Dam-fusion / GATC site at position x dissociation constant

Figure 5: ***Modeling DamC experiments*** *Schema explaining the different states of the model, (**A.**) the chromatine is in a looped state creating a proximity between the TetO sites and a GATC site. (**B.**) In the same state, the Dam-TetR fusion attaches to the GATC and methylates. (**C.**) The chromatin is not in a looped state so there is no proximity created so the methylation happens through the freely defusing Dam-TetR. (**D.**)*

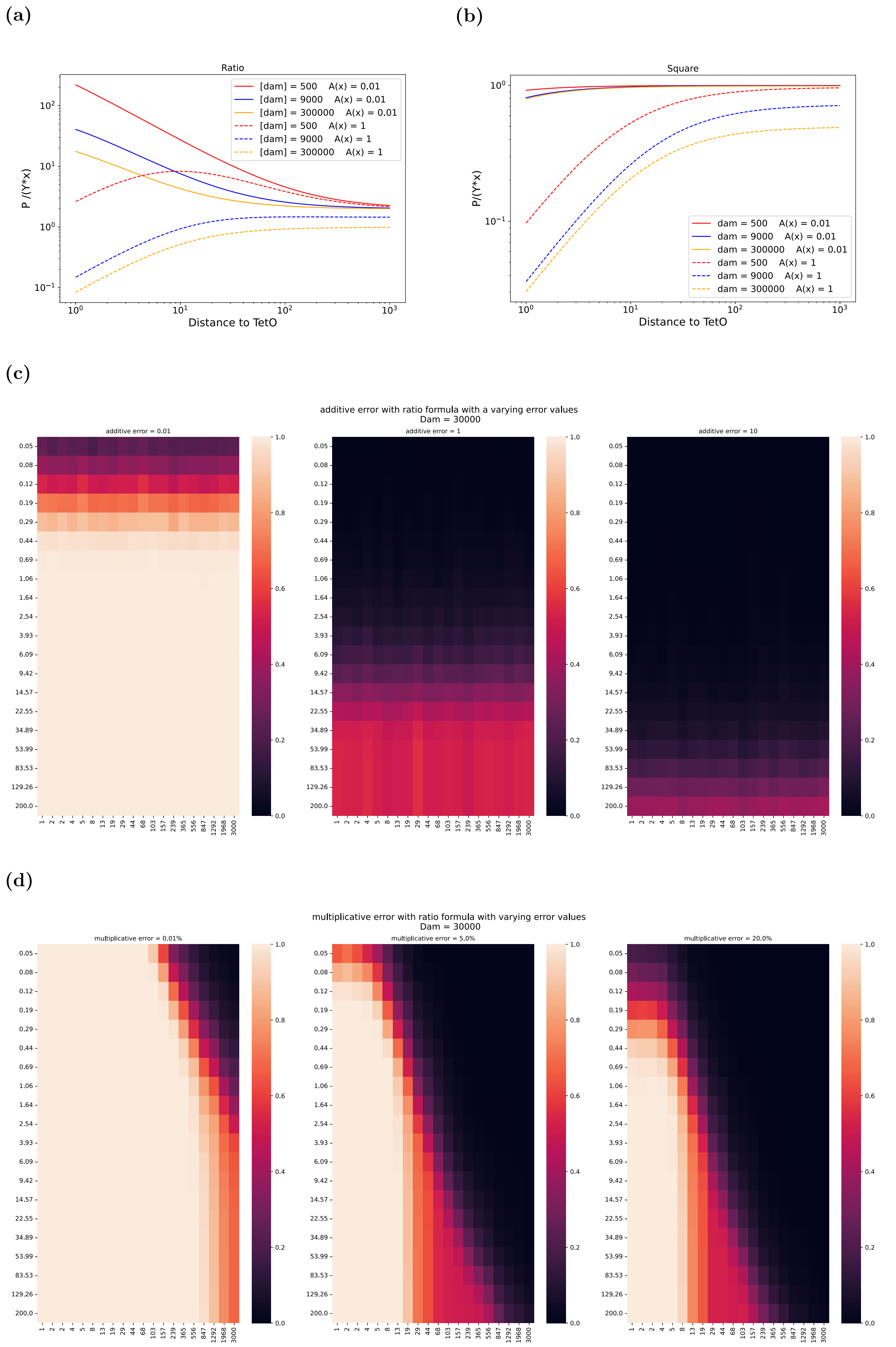


Figure 6: ***Study of the model***

, (**A-B**) lineplot of the $\frac{P_{calculated}}{Y\frac{1}{x}}$ for 3 ratios of $\frac{[Dam]}{k_d^{GATC}}$. The redline represents $[Dam] << k_d^{GATC}$, the blue line $[Dam] = k_d^{GATC}$ and the yellow line $[Dam] >> k_d^{GATC}$. The plain line represents a small value of $A(x)$ while a dotted line represents a value close to 1. (**A.**) Is computed using $YP = \frac{S-B}{BY}$. (**B.**) Is computed using $YP = \sqrt{\frac{S}{B}} - 1$. (**C-D**) Heatmaps representation of $\frac{P_{calculated}}{Y\frac{1}{x}}$ while adding an (**C.**) additive error and (**D.**) multiplicative error. The y axis represents thresholds of difference from 1, and x distance from the TetO site. The color-scale represents percentage of simulations meeting the threshold. (**C.**) Multiple values of error were used to do the heatmaps, 0.01 (**left**), 1 (**middle**), 10 (**right**). (**D.**) Mutliple values of multiplicative errors were also used, 0.01% (**left**), 5% (**middle**), 20% (**right**).