GC content but not nucleosome positioning directly contributes to intron-splicing efficiency in *Paramecium*

Stefano Gnan, Melody Matelot, Marion Weiman, et al.

Genome Res. published online March 9, 2022
Access the most recent version at doi:10.1101/gr.276125.121

GC content but not nucleosome positioning directly contributes to intron-splicing efficiency in *Paramecium*

Stefano Gnan[1,4], Mélody Matelot[2,4], Marion Weiman[3], Olivier Arnaiz[3], Frédéric Guérin[2], Linda Sperling[3], Mireille Bétermier[3], Claude Thermes[3], Chun-Long Chen[1]* and Sandra Duharcourt[2]*

[1] Institut Curie, Université PSL, Sorbonne Université, CNRS UMR3244, Dynamics of Genetic Information, Paris, 75005, France

[2] Université de Paris, CNRS, Institut Jacques Monod, F-75013, Paris, France

[3] Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France

[4] joint First Authors

* corresponding authors

Sandra Duharcourt, Tel: [+33-(0)157278009]; Email: [sandra.duharcourt@ijm.fr]

ORCID: https://orcid.org/0000-0002-8913-8799

Chun-Long Chen, Tel: [+33-(0)156246205]; Fax: [+33-(0)156246674]; Email: [chunlong.chen@curie.fr]

ORCID: https://orcid.org/0000-0002-4795-0295

Present Address:

[Mélody Matelot], IGBMC - CNRS UMR 7104 - Inserm U 1258, 1 rue Laurent Fries, BP 10142, 67404 Illkirch CEDEX, France

[Marion Weiman], OncoDNA group, IntegraGen, Evry, 91000, France

[Frédéric Guérin], Scipio bioscience, 92120 Montrouge, France

Running Title: nucleosome position in intron splicing efficiency

Keywords: Nucleosome positioning, intron splicing, Nonsense Mediated Decay, MNase-seq

ABSTRACT

Eukaryotic genes are interrupted by introns that must be accurately spliced from mRNA precursors. With an average length of 25 nt, the >90,000 introns of *Paramecium tetraurelia* stand among the shortest introns reported in eukaryotes. The mechanisms specifying the correct recognition of these tiny introns remain poorly understood. Splicing can occur co-transcriptionally and it has been proposed that chromatin structure might influence splice site recognition. To investigate the roles of nucleosome positioning in intron recognition, we determined the nucleosome occupancy along the *P. tetraurelia* genome. We showed that *P. tetraurelia* displays a regular nucleosome array with a nucleosome repeat length of ~151 bp, amongst the smallest periodicities reported. Our analysis revealed that introns are frequently associated with inter-nucleosomal DNA, pointing to an evolutionary constraint favoring introns at the AT-rich nucleosome edge sequences. Using accurate splicing efficiency data from cells depleted for nonsense-mediated decay effectors, we showed that introns located at the edge of nucleosomes display higher splicing efficiency than those at the center. However, multiple regression analysis indicated that the low GC content of introns, rather than nucleosome positioning, is associated with high splicing efficiency. Our data reveal a complex link between GC content, nucleosome positioning and intron evolution in *Paramecium*.

INTRODUCTION

In eukaryotes, genomic DNA is compacted by histones into chromatin. The basic unit of chromatin is the nucleosome, which comprises a histone octamer made of the four core histones (H2A, H2B, H3 and H4) and 146-147 bp of DNA wrapped around it (Luger et al. 1997; Parmar and Padinhateeri 2020). Nucleosomes are not randomly located along the genome but positioned with respect to DNA sequence. The affinity of DNA for histone octamers and the energy needed to bend different DNA fragments around the histone octamer are influenced by the primary DNA sequence, which is therefore an important determinant of nucleosome positioning along the genome (Tillo and Hughes 2009; Tillo et al. 2010; Segal et al. 2006; Lorch et al. 2014; Lyer and Struhl 1995; Peckham et al. 2007; Vaillant et al. 2010). Nucleosome positioning is highly dynamic, and its regulation is crucial to control chromatin accessibility, recruitment of chromatin modifiers and transcription factors (Bartholomew 2014; Prendergast and Semple 2011; Kornberg and Lorch 2020; Jiang and Pugh 2009). In most genomes, genes display a Nucleosome Free Region (NFR) at the 5' of their Transcription Start Site (TSS) due to the formation of complexes made by transcription factors around promoter regions (Bernstein et al. 2004; Yuan et al. 2005; Jiang and Pugh 2009). A second NFR is present at Transcription Termination Sites (TTS), likely due to the adverse nucleotide composition of the poly(A) signal (Fan et al. 2010; Chereji et al. 2016). Nucleosomes are organized in regular arrays with a periodic distance called the nucleosome repeat length (NRL). Such periodicity is especially evident over gene bodies, and it is species and cell-type specific (Beshnova et al. 2014; Allan et al. 2013). Genome-wide studies have shown that nucleosomes are preferentially positioned in exons compared to introns in diverse organisms including *Schizosaccharomyces pombe*, *Drosophila*, worms and human (Andersson et al. 2009; Schwartz et al. 2009; Tilgner et al. 2009; Nahkuri et al. 2009; Spies et al. 2009; Iannone et al. 2015). Several lines of evidence indicated that a well-positioned nucleosome might slow down RNA polymerase II and favor exon inclusion and alternative splicing (Jonkers et al. 2014; Wilhelm et al. 2011), suggesting a functional role of

nucleosome arrays during mRNA maturation. This is in agreement with recent studies showing that intron splicing can occur in a co-transcriptional manner (Brody et al. 2011; Herzel et al. 2017). Some studies have suggested that GC richness at exons, and not nucleosome positioning per se, is important for intron splicing (Amit et al. 2012; Gelfman et al. 2013). Yet, the contribution of nucleosome positioning to intron splicing efficiency has not been investigated thoroughly. The nonsense-mediated decay (NMD) machinery recognizes and degrades transcripts containing premature termination codons (Lykke-Andersen and Jensen 2015; Kurosaki et al. 2019). Therefore, most of the mis-splicing or un-splicing events are removed rapidly by this powerful surveillance mechanism to avoid the production of erroneous proteins. To date, most studies estimated splicing efficiency from NMD-proficient cells, which eliminate most mis-splicing or un-splicing events, and therefore cannot provide a solid evaluation of the intrinsic efficiency of intron splicing.

The ciliate *Paramecium tetraurelia* is a unicellular eukaryotic model organism. Like all ciliates, two distinct types of nuclei co-exist within the same cytoplasm in *P. tetraurelia* (Aury et al. 2006). The diploid germline micronucleus (MIC) is transcriptionally silent during vegetative growth and transmits the germline genome to sexual progeny through meiosis, while the highly polyploid somatic macronucleus (MAC) is responsible for gene expression (Bétermier and Duharcourt 2014). The >90,000 introns annotated in the MAC genome are among the shortest reported in eukaryotes (18 to 33 nt, 25 nt on average) (Jaillon et al. 2008). How such a large number of tiny introns can be efficiently spliced is not known. In *P. tetraurelia*, no exon skipping has been reported so far (Saudemont et al. 2017; Jaillon et al. 2008). Introns are associated with weak splice signals, as shown by the very low information content of 5' and 3' splice sites, with only the first and last three bases of introns being highly constrained (Jaillon et al. 2008). A strong counter-selection for introns that cannot be detected by the NMD machinery was previously shown, suggesting that introns rely on NMD to compensate for suboptimal splicing efficiency and accuracy (Jaillon et al. 2008; Saudemont et al. 2017). Whether nucleosome positioning or other factors, such as GC

content, can regulate splicing efficiency and shape intron evolution in *Paramecium* has not been studied so far.

Here, we investigated a possible role of nucleosome positioning in the recognition of introns in *P. tetraurelia*. We mapped the nucleosome occupancy in the somatic macronucleus (MAC) through paired-end MNase-seq. We compared the positioning of nucleosomes with that of introns, whose accurate splicing efficiency data was determined from NMD-depleted cells.

RESULTS

Genome-wide nucleosome position profiling along the *Paramecium* somatic genome

Using MNase-seq, we derived a first nucleosome positioning profile of the macronuclear (MAC) genome of *P. tetraurelia* during vegetative growth. Both chromatin samples and naked MAC DNA controls were digested to mono-nucleosome size (~150 bp, Fig 1A-B and Supplemental Fig S1A). The results obtained from two biological replicates were highly reproducible (Pearson's correlation R = 0.94, Supplemental Fig S1B). We therefore combined data from both biological replicates for downstream analyses. All the data presented in the main figures were obtained with the average of two chromatin samples and two naked DNA controls, respectively. The results of each individual sample are reported in the Supplemental Figures. Using the gene annotation, together with the Transcription Start Sites (TSSs) identified by 5' CAP-seq and Transcription Termination Sites (TTSs) identified by poly(A) detection (Arnaiz et al. 2017), we investigated the nucleosome occupancy along transcription units and around their extremities. As described in other eukaryotes, *P. tetraurelia* presents an enriched nucleosome density over the transcription units compared to the flanking regions, showing regular arrays of nucleosomes over transcription units (Fig 1C-D and Supplemental Fig S1C-D). As expected, we were able to identify Nucleosome Free Regions (NFRs) upstream of the TSSs of *Paramecium* genes, followed by an array of well-positioned nucleosomes (Fig 1C-D and Supplemental Fig S1C). The analysis of TTSs shows regions with very low nucleosome occupancy downstream of the TTSs and a weakly

organized array towards the gene body (Fig 1D and Supplemental Fig S1D). We further separated gene pairs into 3 groups based on their relative orientation: tandem (n=20,233), convergent (n=8,876) and divergent (n=8,867) (Fig 1E and Supplemental Fig S1E-G). We found that nucleosome arrays are clearly visible upstream of the TTSs only when genes are positioned in tandem (Fig 1E and Supplemental Fig S1F), but not in convergent pairs (Fig 1E and Supplemental Fig S1G). This observation suggests that the nucleosome positioning at TTS observed for tandem genes might be due to the downstream TSS, as suggested for *Saccharomyces cerevisiae* (Chereji et al. 2017). Alternatively, convergent genes might be influenced by the transcription readthrough of the gene in the opposite orientation.

Based on our nucleosome position calling and using only well-positioned nucleosomes identified in both replicates (see Methods), we calculated the nucleosome repeat length (NRL) (Methods). We found that *P. tetraurelia* displays one of the smallest NRL reported in eukaryotes (150.89±0.57 bp on average, Fig 1F-G and Supplemental Fig S1H-I), close to the 156±2 bp of *S. pombe* (Godde and Widom 1992), which is much smaller than the 167 bp of *S. cerevisiae* (Vaillant et al. 2010) (see Discussion). In human, the NRL within gene bodies is smaller than outside (Valouev et al. 2011). We performed a similar analysis sub-setting nucleosomes based on whether their centers overlap with gene bodies or not. We found a negligible difference between the NRL within gene bodies (151.00±0.94 bp, more than 80% of the analyzed sequences) and those outside of genes (150.29±1.29 bp) (Supplemental Fig S1J).

The tiny introns of *Paramecium* genes are frequently associated with inter-nucleosomal DNA

We then analyzed nucleosome positioning over gene bodies. In *P. tetraurelia*, exons range from several nucleotides to a few kilobases (Fig 2A, Supplemental Fig S2A for transcription units identified by 5' CAP-seq and poly(A) detection) and are interspersed with tiny introns, the majority spanning between 20 and 35 bp with a median size of 25 bp (Fig 2B). The

distribution of exon size shows a peak around 150 bp close to the size of nucleosomes in *P. tetraurelia*, which is smaller than the simulated exon size by assuming uniform exon sizes within each gene (Fig 2A, Supplemental Fig S2A). By visual inspection of the nucleosome occupancy profiles, we noticed a tendency of the MNase signal to be stronger over exons leaving the introns preferentially between two nucleosome peaks (Fig 2C, Supplemental Fig S2B for each MNase-digested chromatin sample). This was especially visible when we examined the nucleosome density over introns sorted by the distance of each intron center to the closest nucleosome center (Fig 2D, Supplemental Fig S2C). This distance is significantly higher than what we would expect by calculating the distance of random positions inside gene bodies to the closest nucleosome center (p value < 10^{-10} calculated with Mann-Whitney U test, one sided, alternative H_1: Intron distance from the closest nucleosome is higher than random chance. Fig 2E). Using this distance, we grouped introns into 3 categories: central, proximal and distal (as illustrated in Fig 2F and Supplemental Fig S2C). We calculated their distribution and compared it with that of exons smaller than 300 bp (roughly the same sample size) categorized in the same way (Fig 2G and Supplemental Fig S2D). Introns were found enriched at distal positions, i.e. located in the regions between two neighbor nucleosomes, compared to exons (45% vs 22%, respectively). In contrast, exons were more enriched in central positions compared to introns (46% vs 28%, respectively). These distributions are statistically significantly different: p value < 10^{-10} calculated with a χ^2 test (Fig 2G). Moreover, *P. tetraurelia* exons seem to favor mono-nucleosome length sizes with 35% of exons having sizes comprised between 100 bp and 200 bp. Such a size distribution is significantly shorter than what would be expected if we simulated exon sizes as uniformly distributed within each transcript, in which case only 24% of the exons would fall in this range (p<10^{-10}, Mann-Whitney U test, one sided, alternative H_1: simulated exons are bigger than real exons) (Fig 2A). Similar results were obtained using only exons of transcription units whose extremities are identified by both 5' CAP-seq and poly(A) detection (Supplemental Fig S2A). This distribution of exon sizes might reflect some selective

constraint keeping introns in phase in distal position, i.e. at the edge of the nucleosome.

Higher splicing efficiency for introns at the edges of nucleosomes

Previous studies have described the effect of nucleosome positioning on mRNA maturation in multiple organisms (Andersson et al. 2009; Schwartz et al. 2009; Tilgner et al. 2009; Nahkuri et al. 2009; Spies et al. 2009; Iannone et al. 2015). To address whether nucleosome positioning affects intron splicing in *P. tetraurelia*, we examined the relationship between nucleosome positioning and intron splicing efficiency, using published datasets from both wild-type (WT) and NMD-depleted cells, which provide a measurement of the splicing efficiency of *P. tetraurelia* introns (Saudemont et al. 2017). Since NMD has been shown to play an important role in removing mis-spliced transcripts and different evolutionary constraints have been observed for NMD-sensitive (presence of a premature termination codon, PTC, after retention) and NMD-insensitive (absence of a PTC after retention) introns (Saudemont et al. 2017), we further divided our 3 positional categories (central, proximal, distal) of introns into NMD-sensitive or NMD-insensitive groups (Fig 3A).

First, we observed that the proportion of distal introns is higher in NMD-insensitive introns compared to NMD-sensitive ones, independent of the introduction of a frameshift (3n versus non-3n introns) (Fig 3A). We could not observe statistically significant differences between 3n and non-3n NMD-insensitive intron distributions (p=0.38, χ^2 test), and only a minor significant increase of distal introns at the expense of central introns and proximal introns can be detected between 3n and non-3n NMD-sensitive introns ($p<10^{-3}$, χ^2 test) (Fig 3A). Since no major differences in the intron distribution between 3n and non-3n introns were observed, we decided to consider only the NMD state for subsequent analyses.

According to previous reports, a PTC is more likely to be recognized by the NMD system if it is located far away from the actual termination codon at the 3′ end of the gene (Brogna and Wen 2009; Vitali et al. 2019). We reasoned that an NMD-sensitive intron close to the TSS

has a higher probability to induce a PTC far away from the actual termination codon. Therefore, we analyzed the distribution of intron positional categories with regard to nucleosomes (distal, central, proximal) as a function of their relative position within genes and of their NMD sensitivity. We found that, for NMD-insensitive introns, the proportion of distal introns is much higher than that of central introns for all distance classes, with only a slight increase of distal intron percentage toward the gene 3' end (Fig 3B and Supplemental Fig S3A). However, for the NMD-sensitive introns, we observed a linear increase of the percentage of distal introns toward the gene 3' end (Fig 3B and Supplemental Fig S3A). This indicates that introns close to the TTS are more frequently associated with distal positions, i.e. at the edge of the nucleosome. To assess whether these introns close to the TTS are less sensitive to NMD, we monitored intron retention rates for the different intron groups. This confirmed that i) the NMD pathway is more efficient for NMD-sensitive introns close to the TSS, i.e. located at the beginning of a gene (Fig 3C), and ii) much higher retention rates in NMD-depleted cells are observed for NMD-sensitive introns located near a TSS compared to those near a TTS (Fig 3C left panel). As expected, no difference can be observed for NMD-insensitive introns (Fig 3C right panel). For NMD-sensitive introns, we observed a higher splicing efficiency (i.e. lower retention rate) for introns located in distal positions compared to those in central and proximal positions independent of their relative position within a gene (Fig 3C, left and Supplemental Fig S3B). We conclude that NMD-sensitive introns located at distal positions, i.e. at the edge of nucleosomes, are more efficiently spliced.

As shown in (Saudemont et al. 2017), the intron retention rate is inversely correlated with the gene expression level and is higher for introns that can be detected by the NMD machinery than for those that cannot. In WT cells, both NMD-sensitive and NMD-insensitive introns showed similar retention rates, with higher retention rates for genes with lower expression levels (Fig 3D). The retention rate of NMD-sensitive introns increased significantly upon NMD depletion, while it did not for NMD-insensitive introns (Fig 3D). We extended this analysis to our intron positional categories. As expected, NMD-insensitive introns showed

similar splicing efficiency for all intron classes in both WT and NMD-depleted cells (Fig 3D right panel and Supplemental Fig S3C). We found that the retention rate of NMD-sensitive introns is lower for distal introns compared to the other two categories (Fig 3D left and Supplemental Fig S3C), indicating again that NMD-sensitive introns located at the edges of nucleosomes are more efficiently spliced. This can already be observed in WT cells, while in NMD-depleted cells, where nonsense mRNAs are no longer degraded, this difference is much stronger (Fig 3D left panel and Supplemental Fig S3C). For the low-expressed genes (RPKM ≤ 1) the retention rate of central introns is 36.6% higher than that of distal introns, and it drops to 24.6% and 13.8% for the mid-expressed (1 < RPKM ≤ 10) and highly-expressed (RPKM > 10) genes, respectively (Fig 3D left panel and Supplemental Fig S3C). We further analyzed the proportion of intron positional categories within genes with different expression levels and found similar proportions for all expression classes (Supplemental Fig S3D). Similar results were also observed for genes issued from the last whole-genome duplication (Aury et al. 2006) that have different expression levels (Supplemental Fig S3E). Finally, after controlling both gene expression levels and the relative distance of the intron to the TSS, we still observed that the distal introns have a higher splicing efficiency than the central and proximal ones (Supplemental Fig S3F).

It has been shown that the splicing efficiency of *P. tetraurelia* introns depends on the sequences at the donor and acceptor sites (Jaillon et al. 2008). We thus assessed whether this difference in splicing efficiency between our nucleosome-positional classes could be explained by a different distribution of stronger donor (5’ GTA) and/or stronger acceptor (3’ TAG) sites (Supplemental Fig S3G) within different intron groups. As expected, NMD-insensitive introns were more frequently associated with both stronger donors and acceptors whatever the distance of the intron to the closest nucleosome center (Fig 3E). In contrast, we found a minor increase, for the NMD-sensitive introns, in the association of distal introns with “weaker donor and acceptor” (0.82%) and “stronger acceptor only” (4.46%) intron groups compared to central introns (0.50% and 3.80%, respectively) with respectively 64%

and 17% increase (Fig 3E). This slight increase was not associated with a higher retention rate for distal introns compared to central introns. Instead, we did observe a reduced retention rate in distal introns (Fig 3C-D). We conclude that the reduced retention rate in distal introns is not due to a difference of donor/acceptor signals in this class.

GC content related to nucleosome positioning contributes to intron-splicing efficiency at the edges of nucleosomes

It is well known that nucleosome positioning is highly associated with GC content: nucleosome centers show higher GC than distal regions (Tillo and Hughes 2009; Lorch et al. 2014; Lyer and Struhl 1995; Peckham et al. 2007; Vaillant et al. 2010). In *P. tetraurelia*, we observed that NMD-sensitive introns have a higher GC content (18.9%, 18.4% and 15.7% for central, proximal and distal introns, respectively) than their NMD-insensitive counterparts (16.3%, 16.1% and 13.2% for central, proximal and distal introns, respectively) (Fig 4A). Moreover, as we would expect, the central introns have the highest GC content followed by proximal and distal introns (Fig 4A). We therefore analyzed the impact of GC content on intron retention rates. We found a direct correlation between GC percentage and retention rate in NMD-sensitive introns, yet no statistically significant difference could be observed between different intron groups (Fig 4B and Supplemental Fig S4A). This suggests that GC-content anti-correlates with intron splicing efficiency. To further evaluate how different parameters, such as GC content, gene expression level, intron relative position within genes, nucleosome positioning and RNA secondary structure prediction (Supplemental Table S1) affect intron splicing efficiency, we first filtered the parameters by trying to lower the variance inflation factor below 5, and then used the resulting parameters to train a multivariate regression model as previously described (Chen et al. 2010). Only the parameters with a statistically significant contribution were retained (Methods). The final fitted model has an R=0.62, which explains 39% of the variation in intron splicing efficiency measured in NMD-

depleted cells. The model allowed us to estimate the contribution of each parameter (full list of parameters in Supplemental Table S1).

The highest contribution came from the level of gene expression that accounts for about 46% of the model (Fig 4C). The GC content of the intron accounts for 15%, and together with other parameters associated with the base composition of the introns (e.g. TC% accounts for 6.2%), the total contribution of base composition reaches ~22%. Although the difference between GC content of an intron and the flanking exons (ΔGC) has been previously reported to be linked to intron splicing efficiency (Amit et al. 2012), ΔGC was not retained in our final model (Supplementary Table S1). As expected, GC content and ΔGC (introns – flanking exons) are highly correlated (Supplementary Fig S4B) and forcing the usage of the latter does not improve the model. Splicing signals account for 9.5% of the model and the parameters associated with the size and base composition of the transcript account for about 8.7%. Intron size and position in the transcript account for 7.1% of the model, followed by intron sensitivity to the NMD pathway (2.5%), the size and base composition of the flanking exons (2.4%), whether an intron is 3n or not (0.66%) and the parameters associated with the formation of secondary structures (0.35%). All the parameters relative to nucleosome positioning account for only 0.43% of the model (Fig 4C-D and Supplemental Table S1). Moreover, if we divide the introns based on their NMD sensitivity, our model can explain 40% of the variation in intron splicing efficiency for the NMD-sensitive introns, while only 27% for the NMD-insensitive ones (Supplemental Fig S4C). We therefore conclude that the GC content, which is tightly linked to nucleosome positioning, contributes to intron-splicing efficiency: a high GC content, which is correlated with high nucleosome occupancy, is associated with low splicing efficiency.

DISCUSSION

We have performed the first nucleosome position profiling in the *P. tetraurelia* MAC genome during vegetative growth. Despite its high AT richness (72% AT), the *P. tetraurelia* MAC

genome displays a very regular nucleosome positioning pattern as observed in other eukaryotes: NFRs at the TSSs and TTSs, and a regular nucleosome array along genes. An independent study reached the same conclusions (Drews et al. 2021). Unlike *Tetrahymena thermophila,* another AT-rich ciliate (78% AT) (with an NRL of 199 bp) (Beh et al. 2015), the NRL in the *P. tetraurelia* MAC genome presents a smaller periodicity (151±1 bp), very similar to that of *S. pombe* (156 bp) (Godde and Widom 1992) and of *Plasmodium falciparum* (155 bp) (>80%AT) (Kensche et al. 2015; Silberhorn et al. 2016). This short NRL means that the naked "linker" DNA between nucleosomes in *Paramecium* is extremely small (only a few bp) compared to that of most other eukaryotic genomes, at least tens of bp or even larger (Arceci and Gross 1980). A higher H1/core-histone ratio has been previously reported being associated with a longer NRL (Fan et al. 2003, 2005; Woodcock et al. 2006). For the three eukaryotes with the smallest NRL, *P. tetraurelia, P. falciparum* and *S. pombe*, no orthologue of histone H1 has been identified so far. This strongly suggests that the absence of H1 might contribute to the extremely short NRL observed in *Paramecium* chromatin organization in the somatic MAC genome.

In yeast and human, actively transcribed genes tend to have shorter NRL than transcriptionally inactive genes, partially due to the binding of H1 generating inaccessible chromatin at inactive genes (Correll et al. 2012; Barbier et al. 2021; Valouev et al. 2011). With the separation of the germline MIC and the somatic MAC genomes in two distinct nuclei, the *Paramecium* MAC genome is characterized by very high coding density. Indeed, >80% of the MAC is covered by annotated genes and 65% of the coding genes are expressed (RNA-seq coverage of at least 1 RPKM) during vegetative growth (Aury et al. 2006; Arnaiz et al. 2017), which might explain the extremely short length and narrow distribution of NRL. A significant difference in the nucleosome organization between MAC and MIC genomes has been reported for *T. thermophila* (Xiong et al. 2016). How nucleosomes are organized in the *Paramecium* MIC genome is unknown. At each sexual cycle of *Paramecium*, the parental MAC is destroyed and the new MIC and MAC are generated from the parental germline MIC

(Bétermier and Duharcourt 2014). During new MAC development, at least 30% of the germline DNA is eliminated during massive genome rearrangements (Guérin et al. 2017; Sellis et al. 2021). A large amount of extremely short (26 to ~1000 bp) non-coding germline sequences, called IESs (Internal Eliminated Sequences), need to be precisely excised to correctly assemble functional genes in the new MAC genome of *Paramecium* species (Sellis et al. 2021). How nucleosome positioning is organized in the germline MIC genome relative to IESs and whether nucleosome positioning and/or GC content might play a role in IES excision are open questions (Coyne et al. 2012; Lhuillier-Akakpo et al. 2014).

In multicellular eukaryotes, long introns are recognized through exon definition and nucleosomes positioned along exons might contribute to the exon-intron architecture, possibly pointing to a function in exon definition (Andersson et al. 2009; Schwartz et al. 2009; Tilgner et al. 2009; Nahkuri et al. 2009; Spies et al. 2009; Iannone et al. 2015). By contrast, short introns are recognized through intron definition. With an average length of 25 nt, introns of *P. tetraurelia* are among the shortest reported in eukaryotes (Jaillon et al. 2008). The large number of introns (>90,000) are associated with weak splicing signals. In the current study, we examined the role of nucleosome positioning in intron splicing. We found a regular nucleosome array associated with intron positions within genes, with exons wrapped around nucleosomes and introns frequently located at the edge of nucleosomes. By using the accurate splicing efficiency data obtained from NMD-depleted cells (Saudemont et al. 2017), we performed a thorough investigation on the effect of nucleosome positioning on splicing efficiency. We showed that the NMD-sensitive introns located at the edge of nucleosomes display higher splicing efficiency than those at the nucleosome centers. However, we found that this higher splicing efficiency is due to the fact that the introns located at the edges of nucleosomes display lower GC content. Our multiple regression analysis indicated that the nucleosome positioning has a minimal contribution (0.43%) to the intron splicing efficiency (Supplementary Fig S4C and Supplemental Table S1). Our results strongly indicate that GC content, and more broadly intron base composition, rather than

nucleosome positioning, directly influences intron splicing efficiency in *Paramecium.* This conclusion may pave the way for future mechanistic studies to decipher how GC content impinges on intron splicing efficiency. Whether the effect of GC content and nucleosome positioning on intron splicing efficiency observed in *Paramecium* can be extended to other eukaryotes remains an open question.

We also observed that during evolution, nucleosome positioning has been displaced relative to introns, frequently locating the AT-rich intron sequences at the edge of nucleosomes (Fig 4A). Although both NMD-sensitive and NMD-insensitive introns present a higher proportion of distal positions, NMD-insensitive introns show a significantly higher proportion (50% for 3n and 48% for non-3n introns) than NMD-sensitive introns (40% for 3n and 44% for non-3n introns) (Fig 3A). This strongly suggests that the NMD-insensitive introns not located at the AT-rich nucleosome edges, whose retention in transcripts cannot be cleaned up by the NMD pathway, are counter-selected during evolution. Whether introns in *Paramecium* might play a functional role is still unclear. These introns do not seem to contribute to alternative splicing to generate protein diversity or to encode ncRNAs as in large other genomes with long and abundant introns (Lee and Rio 2015; Chen et al. 2003; Ruby et al. 2007). Due to their extremely small size, it seems unlikely that these introns play a role in regulating transcription rate as suggested in recent publications (Fong et al. 2014; Aslanzadeh et al. 2018; Alexander et al. 2010). As the parameters analyzed in this study only explain ~40% of the variation in intron splicing efficiency, other parameters remain to be identified and perhaps other models would be necessary to fully understand what intron properties determine splicing efficiency. How such a large number of tiny introns in *Paramecium* is maintained during evolution and how these introns can be efficiently spliced need to be further investigated.

METHODS

402 ***Paramecium* strains, cultivation and autogamy**

403 All experiments were carried out with the entirely homozygous wild type strain 51 of *P.*
404 *tetraurelia*. Cells were grown at 27°C in wheat grass powder (WGP) infusion medium
405 bacterized the day before use with *Klebsiella pneumoniae* and supplemented with 0.8
406 mg/mL β-sitosterol (Beisson et al. 2010a, 2010b).
407
408 **Macronuclei preparation**
409 Cells were exponentially grown for 12 divisions then cultures at 1,000 cells/mL were filtered
410 through eight layers of sterile gauze. Cells were collected by low-speed centrifugation (550 g
411 for 1 min) and washed once with 10 mM Tris-HCl pH 7.4. The pellet was diluted 3-fold by
412 addition of lysis buffer (0.25 M sucrose, 10 mM $MgCl_2$, 10 mM Tris pH 6.8, 0.2% Nonidet P-
413 40) and processed at 4°C as described in (Arnaiz et al. 2012) with some modifications.
414 Briefly, cells were lysed with 10 strokes of a Dounce homogenizer. Particular care was taken
415 to make sure that macronuclei were still intact under the microscope. Washing buffer (0.25
416 M sucrose, 10 mM $MgCl_2$,10 mM Tris-HCl pH 7.4) was added to a final volume of 10 times
417 the initial pellet. Macronuclei were collected by centrifugation at 2,000 g for 1 min and
418 washed once in washing buffer. The pellet was diluted 2-fold in 2.1 M sucrose, 10 mM $MgCl_2$,
419 10 mM Tris pH 7.4 and loaded on top of a 3-mL sucrose layer (2.1 M sucrose, 10 mM $MgCl_2$,
420 10 mM Tris-HCl pH 7.4) and centrifuged in a swinging rotor for 1 hr at 210,000 g. The
421 macronuclear pellet was washed once, centrifuged at 2,000 g for 1 min and resuspended in
422 washing buffer at 10^7 nuclei/mL. The macronuclei recovery is quite low, of the order of 10-
423 20%.
424
425 **MNase digestion on chromatin isolated from macronuclei**
426 Samples containing 10^5 macronuclei were incubated in the digestion buffer (0.25 M sucrose,
427 10 mM $MgCl_2$,10 mM Tris pH 7.4, 1 mM $CaCl_2$) with increasing amounts (0, 0.5, 1, 2, 5, 7.5,
428 10 U) of MNase (Sigma-Aldrich) at 30°C for 10 min. Reactions were stopped by the addition
429 of 3 volumes of 0.5 M EDTA pH 9.0, 1% N-laurylsarcosine (Sigma-Aldrich), 1% SDS, 1

mg/mL Proteinase K (Merck) and incubated at 55°C overnight. DNA from each sample was gently extracted once with phenol, and dialyzed twice against TE (10 mM Tris-HC1, 1 mM EDTA at pH 8.0) containing 25% ethanol, and once against TE. Samples were then treated with RNase A and DNA was quantified with a NanoDrop spectrophotometer (Thermo Fisher Scientific) and separated on a 1.2% agarose gel. The reactions containing mostly mono-nucleosomal DNA fragments (see Fig 1) were selected and mono-nucleosomal DNA fragments were purified from 3% low melting-temperature agarose gels and treated with β-agarase (Sigma-Aldrich) for sequencing.

MNase digestion on naked DNA

Following purification on a sucrose layer, the macronuclear pellet was washed once, centrifuged at 2,000 g for 1 min, and was resuspended in three volumes of lysis solution (0.5 M EDTA at pH 9.0, 1% SDS, 1% N-laurylsarcosine (Sigma-Aldrich), 1 mg/mL of Proteinase K (Merck) then incubated at 55°C overnight. DNA was gently extracted with phenol, and dialyzed twice against TE (10 mM Tris-HC1, 1 mM EDTA at pH 8.0) containing 20% ethanol, and once against Tris 10mM pH 8.0. 1.6 µg of DNA was digested with increasing amounts of MNase (0 to 1×10^{-3} U) in the digestion buffer at 30°C for 10 min. The reactions were stopped with 250 mM EDTA. The samples were analyzed on a 1.2 % agarose gel and reactions containing fragments of 100-200 bp were gel-purified for DNA sequencing (see Supplemental Fig S1).

MNase library preparation and sequencing

Sequencing libraries were generated using the sequencing kit: TruSeq SBS Kit v5 – GA (36 Cycle) (FC-104-5001, Illumina). Samples were then sequenced on an Illumina GA-IIx sequencer using paired-end (PE) 74 bp setting. The MNase-seq datasets used in this study are from (Hardy et al. 2021) and are available under accession number PRJEB39679 at the European Nucleotide Archive (ENA: https://www.ebi.ac.uk/ena)

Alignment was performed using Bowtie 2 (v2.3.3 --local and other default parameters) (Langmead and Salzberg 2012) and mapping to the MAC genome of strain 51 v1.0 (ptetraurelia_mac_51.fa), available at ParameciumDB (https://paramecium.i2bc.paris-saclay.fr/) (Arnaiz et al. 2019).

Nucleosome positioning calling

After aligning reads to the reference MAC genome, PCR duplicates with the same start and end positions were removed. Only reads mapped in proper pair with a mapping quality score equal or higher than 30 were kept. Filtering, sorting and filling of mate related flags were performed using Samtools, version 1.9 (Danecek et al. 2021). Bamfiles were converted into bed using BEDTools, version 2.29.2 (Quinlan and Hall 2010) and a customized script. We aimed to use only reads deriving from mono-nucleosomes, therefore, read pairs longer than 150 bp and shorter than 75 bp were excluded. We used only the data within the scaffolds larger than 200 kb. A nucleosome score was calculated using the central 75 bp of each read pair. Signal was then smoothed with a gaussian filter and a sigma of 30 over 90 bp for visual assessment of nucleosome position calling. Local maxima and local minima were identified by convoluting the nucleosome score with a first derivative of a gaussian (sigma 30 over ±90 bp). The points of inflection were identified by convoluting the nucleosome score with a kernel containing the second derivative of a gaussian (sigma 30 over ± 90 bp). Peaks were called as a local maximum between two inflection points with opposite inclination. Peaks were called independently in the two chromatin samples, and then a list of well-positioned nucleosomes was compiled using those nucleosomes whose dyad (i.e. center) differs by less than 10 bp between the two biological replicates (about 75% of all nucleosomes). These well-positioned nucleosomes were used for downstream analyses.

Computation of Nucleosome Repeat Length

To compute the Nucleosome Repeat Length (NRL), we first calculated the distance of each nucleosome to all the other nucleosomes on the same scaffold, then used the distances

485 obtained to generate the density distribution. This density distribution was then smoothed
486 using a gaussian filter (sigma=10 over ±30bp) and local maxima identified convoluting the
487 density distribution with the first derivative of a gaussian (sigma=10 over ±30bp). The first n
488 local maxima were then ordered by increasing distances and fitted using a linear model. The
489 slope of the fitted model corresponds to the NRL.
490
491 **Nucleosome distribution calculation**
492 Gene annotation v2.0 of MAC was from (Arnaiz et al. 2019), and the transcription start sites
493 and transcription termination sites were from (Arnaiz et al. 2017). The gene annotations and
494 RNA-seq data are available at ParameciumDB (https://paramecium.i2bc.paris-saclay.fr/). To
495 compare with the distribution of real exon sizes, a set of simulated exons was created
496 assuming uniform exon sizes within each gene, i.e. for a given gene with *n* exons, we
497 divided its total exon length by *n* to get the length of *n* simulated exons of the corresponding
498 gene. The NMD data were obtained from (Saudemont et al. 2017), splicing efficiency of each
499 intron was calculated as the Splicing events / Total number of observations (i.e. spliced +
500 unspliced reads). The mean profiles and heatmaps were drawn using a customized script
501 and plotting using Matplotlib (version 3.1.0) (Hunter 2007). All statistical analyses were
502 performed with Python (version 3.7.4, https://docs.python.org/release/3.7.4/) using
503 statsmodels (version 0.10.1)(Seabold and Perktold 2010) and SciPy (version 1.3.1)(Virtanen
504 et al. 2020) modules.
505
506 **Multilinear regression**
507 The starting parameters used for the multiple linear regression can be found in
508 Supplemental Table S1. Parameters were transformed using appropriate functions in order
509 to maximize their linearity with intron splicing efficiency, e.g. log transformation of expression
510 levels. Values were then standardized. Variance inflation factors (VIF) were calculated for
511 the whole pool of parameters. If the parameter with highest VIF exceeded the threshold of 5
512 it was excluded from the pool. Parameters VIF was then recalculated and the process

repeated until VIF was greater than 5. Parameters from this first selection were used to fit our linear regression model. A randomly selected set of introns (10% of all introns) was kept from the multilinear regression model fitting, and used as a test dataset to evaluate the model performance. We performed a two-sided Z-test for each coefficient with H0: C=0 and H1: C≠0. Statistically significant coefficients were then retained and the linear model was trained again with the associated parameters. This step was repeated until the number of variables is stabilized. Estimation of the contribution of each parameter is calculated as in (Chen et al. 2010), which is based on the absolute value of the product of each coefficient and the Pearson's correlation value of its parameter with the splicing efficiency. Contributions were then converted to percentages. Using the intron test dataset, we calculated the Pearson's correlation between real and predicted data. To calculate the Pearson's correlation between prediction and real data divided by NMD-sensitive and NMD-insensitive, all the introns belonging to either group were used. For this part, Python (version 3.7.4, https://docs.python.org/release/3.7.4/) was used with scikit-learn (version 0.21.3) (Pedregosa et al. 2011), statsmodels (version 0.10.1) (Seabold and Perktold 2010) and SciPy (version 1.3.1) (Virtanen et al. 2020) modules. The full list of parameters can be found in Supplemental Table S1.

DATA ACCESS

The customized script and Jupyter notebooks used for this study are available as a Supplemental Code File and on our GitHub page (https://github.com/CL-CHEN-Lab/Nucleosome).

COMPETING INTEREST STATEMENT

The authors declare no competing interest.

ACKNOWLEDGMENTS

The authors would like to thank Laurent Duret for useful suggestions and discussion, Laurent Duret and Eric Meyer for sharing with us the NMD data, and to acknowledge the high-throughput sequencing facility of I2BC for its sequencing and bioinformatics expertise.

LS, MB, CT, CLC and SD conceived and planned the study. MM, FG and SD conducted the experiments. SG, MW, OA and CLC performed the bioinformatics analyses. CT and CC supervised the bioinformatics analyses. SG, CLC and SD wrote the manuscript, and all the authors reviewed it.

REFERENCES

Alexander RD, Innocente SA, Barrass JD, Beggs JD. 2010. Splicing-Dependent RNA polymerase pausing in yeast. *Mol Cell* **40**: 582–593.

Allan J, Fraser RM, Owen-Hughes T, Docherty K, Singh V. 2013. A comparison of in vitro nucleosome positioning mapped with chicken, frog and a variety of yeast core histones. *J Mol Biol* **425**: 4206–4222.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, et al. 2012. Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Rep* **1**: 543–556.

Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19**: 1732–41.

Arceci RJ, Gross PR. 1980. Sea urchin sperm chromatin structure as probed by pancreatic DNase I: Evidence for a novel cutting periodicity. *Dev Biol* **80**: 210–224.

Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Denby Wilkes C, Garnier O, Labadie K, Lauderdale BE, Le Mouël A, et al. 2012. The Paramecium germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet* **8**: e1002984.

Arnaiz O, Meyer E, Sperling L. 2019. ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. *Nucleic Acids Res* **48**: D599–D605.

568 Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt S, Sallet
569 E, Gouzy J, Sperling L. 2017. Improved methods and resources for paramecium
570 genomics: Transcription units, gene annotation and gene expression. *BMC Genomics*
571 **18**: 1–12.
572 Aslanzadeh V, Huang Y, Sanguinetti G, Beggs JD. 2018. Transcription rate strongly affects
573 splicing fidelity and cotranscriptionality in budding yeast. *Genome Res* **28**: 203–213.
574 Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard
575 V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the
576 ciliate Paramecium tetraurelia. *Nature* **444**: 171–178.
577 Barbier J, Vaillant C, Volff J-NN, Brune FG, Audit B, Brunet F, Audit B. 2021. Coupling
578 Between Sequence-Mediated Nucleosome Organization and Genome Evolution.
579 *Genes (Basel)* **12**: 1–23.
580 Bartholomew B. 2014. Regulating the Chromatin Landscape: Structural and Mechanistic
581 Perspectives. *Annu Rev Biochem* **83**: 671–696.
582 Beh LY, Müller MM, Muir TW, Kaplan N, Landweber LF. 2015. DNA-guided establishment of
583 nucleosome patterns within coding regions of a eukaryotic genome. *Genome Res* **25**:
584 1727–38.
585 Beisson J, Bétermier M, Bré MH, Cohen J, Duharcourt S, Duret L, Kung C, Malinsky S,
586 Meyer E, Preer JR, et al. 2010a. Maintaining clonal paramecium tetraurelia cell lines of
587 controlled age through daily reisolation. *Cold Spring Harb Protoc* **5**: pdb.prot5361.
588 Beisson J, Bétermier M, Bré MH, Cohen J, Duharcourt S, Duret L, Kung C, Malinsky S,
589 Meyer E, Preer JR, et al. 2010b. Mass culture of paramecium tetraurelia. *Cold Spring*
590 *Harb Protoc* **5**: pdb.prot5362.
591 Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. 2004. Global nucleosome
592 occupancy in yeast. *Genome Biol* **5**: R62.
593 Beshnova DA, Cherstvy AG, Vainshtein Y, Teif VB. 2014. Regulation of the Nucleosome
594 Repeat Length In Vivo by the DNA Sequence, Protein Concentrations and Long-Range
595 Interactions ed. Ioshikhes. *PLoS Comput Biol* **10**: e1003698.

Bétermier M, Duharcourt S. 2014. Programmed Rearrangement in Ciliates: Paramecium. *Microbiol Spectr* **2**: MDNA3-0035–2014.

Brody Y, Neufeld N, Bieberstein N, Causse SZ, Böhnlein EM, Neugebauer KM, Darzacq X, Shav-Tal Y. 2011. The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol* **9**: e1000573.

Brogna S, Wen J. 2009. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol* **16**: 107–113.

Chen C-L, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, D'Aubenton-Carafa Y, Arneodo A, Hyrien O, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**: 447–57.

Chen CL, Liang D, Zhou H, Zhuo M, Chen YQ, Qu LH. 2003. The high diversity of snoRNAs in plants: Identification and comparative study of 120 snoRNA genes from Oryza sativa. *Nucleic Acids Res* **31**: 2601–2613.

Chereji Ř V., Kan TW, Grudniewska MK, Romashchenko A V., Berezikov E, Zhimulev IF, Guryev V, Morozov A V., Moshkin YM. 2016. Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Res* **44**: 1036–1051.

Chereji R V., Ocampo J, Clark DJ. 2017. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Mol Cell* **65**: 565-577.e3.

Correll SJ, Schubert MH, Grigoryev SA. 2012. Short nucleosome repeats impose rotational modulations on chromatin fibre folding. *EMBO J* **31**: 2416–2426.

Coyne RS, Lhuillier-Akakpo M, Duharcourt S. 2012. RNA-guided DNA rearrangements in ciliates: is the best genome defence a good offence? *Biol cell* **104**: 309–25.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: 1–4.

Drews F, Salhab A, Karunanithi S, Cheaib M, Jung M, Schulz MH, Simon M. 2021. Broad domains of histone marks in the highly compact Paramecium macronuclear genome.

bioRxiv 2021.08.05.454756.

Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, Struhl K. 2010. Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3′-end formation. *Proc Natl Acad Sci U S A* **107**: 17945–17950.

Fan Y, Nikitina T, Morin-Kensicki EM, Zhao J, Magnuson TR, Woodcock CL, Skoultchi AI. 2003. H1 Linker Histones Are Essential for Mouse Development and Affect Nucleosome Spacing In Vivo. *Mol Cell Biol* **23**: 4559–4572.

Fan Y, Nikitina T, Zhao J, Fleury TJ, Bhattacharyya R, Bouhassira EE, Stein A, Woodcock CL, Skoultchi AI. 2005. Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**: 1199–1212.

Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, Diener K, Jones K, Fu XD, Bentley DL. 2014. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* **28**: 2663–2676.

Gelfman S, Cohen N, Yearim A, Ast G. 2013. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res* **23**: 789–799.

Godde JS, Widom J. 1992. Chromatin structure of Schizosaccharomyces pombe. A nucleosome repeat length that is shorter than the chromatosomal DNA length. *J Mol Biol* **226**: 1009–1025.

Guérin F, Arnaiz O, Boggetto N, Denby Wilkes C, Meyer E, Sperling L, Duharcourt S. 2017. Flow cytometry sorting of nuclei enables the first global characterization of Paramecium germline DNA and transposable elements. *BMC Genomics* **18**: 327.

Hardy A, Matelot M, Touzeau A, Klopp C, Lopez-Roques C, Duharcourt S, Defrance M. 2021. DNAModAnnot: A R toolbox for DNA modification filtering and annotation ed. P. Robinson. *Bioinformatics* **37**: 2738–2740.

Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. 2017. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18**: 637–650.

652 Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**: 90–95.
653 Iannone C, Pohl A, Papasaikas P, Soronellas D, Vicent GP, Beato M, ValcáRcel J. 2015.
654 Relationship between nucleosome positioning and progesterone-induced alternative
655 splicing in breast cancer cells. *RNA* **21**: 360–374.
656 Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V,
657 Porcel BM, Ségurens B, et al. 2008. Translational control of intron splicing in
658 eukaryotes. *Nature* **451**: 359–362.
659 Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: Advances through
660 genomics. *Nat Rev Genet* **10**: 161–172.
661 Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its
662 interplay with promoter proximal pausing, chromatin, and exons. *Elife* **2014**: e02407.
663 Kensche PR, Hoeijmakers WAM, Toenhake CG, Bras M, Chappell L, Berriman M, Bártfai R.
664 2015. The nucleosome landscape of Plasmodium falciparum reveals chromatin
665 architecture and dynamics of regulatory sequences. *Nucleic Acids Res* **44**: 2110–24.
666 Kornberg RD, Lorch Y. 2020. Primary Role of the Nucleosome. *Mol Cell* **79**: 371–375.
667 Kurosaki T, Popp MW, Maquat LE. 2019. Quality and quantity control of gene expression by
668 nonsense-mediated mRNA decay. *Nat Rev Mol Cell Biol* **20**: 406–420.
669 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*
670 *2012 94* **9**: 357–359.
671 Lee Y, Rio DC. 2015. Mechanisms and regulation of alternative Pre-mRNA splicing. *Annu*
672 *Rev Biochem* **84**: 291–323.
673 Lhuillier-Akakpo M, Frapporti A, Denby Wilkes C, Matelot M, Vervoort M, Sperling L,
674 Duharcourt S. 2014. Local effect of enhancer of zeste-like reveals cooperation of
675 epigenetic and cis-acting determinants for zygotic genome rearrangements. *PLoS*
676 *Genet* **10**: e1004665.
677 Lorch Y, Maier-Davis B, Kornberg RD. 2014. Role of DNA sequence in chromatin
678 remodeling and the formation of nucleosome-free regions. *Genes Dev* **28**: 2492–2497.
679 Luger K, Mä Der AW, Richmond RK, Sargent DF, Richmond TJ, Mäder AW, Richmond RK,

Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**: 251–260.

Lyer V, Struhl K. 1995. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J* **14**: 2570–2579.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: An intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**: 665–677.

Nahkuri S, Taft RJ, Mattick JS. 2009. Nucleosomes are preferentially positioned at exons in somatic and sperm cells. *Cell Cycle* **8**: 3420–3424.

Parmar JJ, Padinhateeri R. 2020. Nucleosome positioning and chromatin organization. *Curr Opin Struct Biol* **64**: 111–118.

Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z. 2007. Nucleosome positioning signals in genomic DNA. *Genome Res* **17**: 1170–1177.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.

Prendergast JGD, Semple CAM. 2011. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res* **21**: 1777–1787.

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

Ruby JG, Jan CH, Bartel DP. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**: 83–86.

Saudemont B, Popa A, Parmley JL, Rocher V, Blugeon C, Necsulea A, Meyer E, Duret L. 2017. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biol* **18**: 1–15.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16**: 990–5.

Seabold S, Perktold J. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference* (eds. S. Van der Walt and J.

Millman), Vol. SCIPY 2010 of, pp. 92–96.

Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772–778.

Sellis D, Guérin F, Arnaiz O, Pett W, Lerat E, Boggetto N, Krenek S, Berendonk T, Couloux A, Aury J-M, et al. 2021. Massive colonization of protein-coding exons by selfish genetic elements in Paramecium germline genomes ed. H.S. Malik. *PLOS Biol* **19**: e3001309.

Silberhorn E, Schwartz U, Löffler P, Schmitz S, Symelka A, de Koning-Ward T, Merkl R, Längst G. 2016. Plasmodium falciparum Nucleosomes Exhibit Reduced Stability and Lost Sequence Dependent Nucleosome Positioning ed. K.W. Deitsch. *PLOS Pathog* **12**: e1006080.

Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Mol Cell* **36**: 245–254.

Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, Guigó R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16**: 996–1001.

Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**: 442.

Tillo D, Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ. 2010. High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. *PLoS One* **5**: 9129.

Vaillant C, Palmeira L, Chevereau G, Audit B, D'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res* **20**: 59–67.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474**: 516–20.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**: 261–272.

Vitali V, Hagen R, Catania F. 2019. Environmentally induced plasticity of programmed DNA elimination boosts somatic variability in Paramecium tetraurelia. *Genome Res* **29**: 1693–1704.

Wilhelm BT, Marguerat S, Aligianni S, Codlin S, Watt S, Bähler J. 2011. Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. *Genome Biol* **12**: R82.

Woodcock CL, Skoultchi AI, Fan Y. 2006. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosom Res* **14**: 17–25.

Xiong J, Gao S, Dui W, Yang W, Chen X, Taverna SD, Pearlman RE, Ashlock W, Miao W, Liu Y. 2016. Dissecting relative contributions of cis-and trans-determinants to nucleosome distribution by comparing Tetrahymena macronuclear and micronuclear chromatin. *Nucleic Acids Res* **44**: 10091–10105.

Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-Scale Identification of Nucleosome Positions in S. cerevisiae. *Science (80-)* **309**: 626–630.

FIGURE LEGENDS

Figure 1. Nucleosome occupancy along the *Paramecium* MAC genome. (A) Schematic representation of the MNase-seq experiment. **(B)** MNase digestion of MAC chromatin with increasing MNase enzyme concentration. **(C)** Heatmap showing nucleosome occupancy ± 1 kb around the center of each gene ordered by gene size (small genes on top and large genes at the bottom) for 38,143 genes located on scaffolds that are at least 200 kb long. Left panel: average of two chromatin treated samples (Chromatin). Right panel: average of two naked DNA control samples (Naked DNA). **(D)** Average nucleosome occupancy around Transcription Start Sites (TSSs) identified by 5' CAP-seq on the left, and Transcription Termination Sites (TTSs) identified by poly(A) detection on the right: in green, the average profile of chromatin treated sample (Chromatin); in blue, average profile of naked DNA treated sample (Naked DNA); and in magenta the Chromatin/Naked DNA ratio, enrichment of which is shown on the second axis on the right (red axis). **(E)** Average nucleosome occupancy ±1 kb around the center of intergenic regions: same color code as in panel D.

Intragenic regions have been divided into three groups based on the relative positions of gene pairs: tandem (left), convergent (middle) or divergent (right). **(F)** Inter-center distance between well-positioned nucleosomes (Methods) on the same scaffold. In blue, distance distributions from actual data (from 1 bp to 2 kb, binning=1); and in orange, the gaussian smoothed signal. Black dashed lines indicate the local maxima (peak centers) of the smoothed data (Methods). (**G**) In orange, the first 8 local maxima from panel F ordered by increasing distance, and in blue the linear fitted model. At the bottom right, information about linear fitting and estimated NRL (Mean±SD) is given. P-value is calculated using a two-sided Z-test.

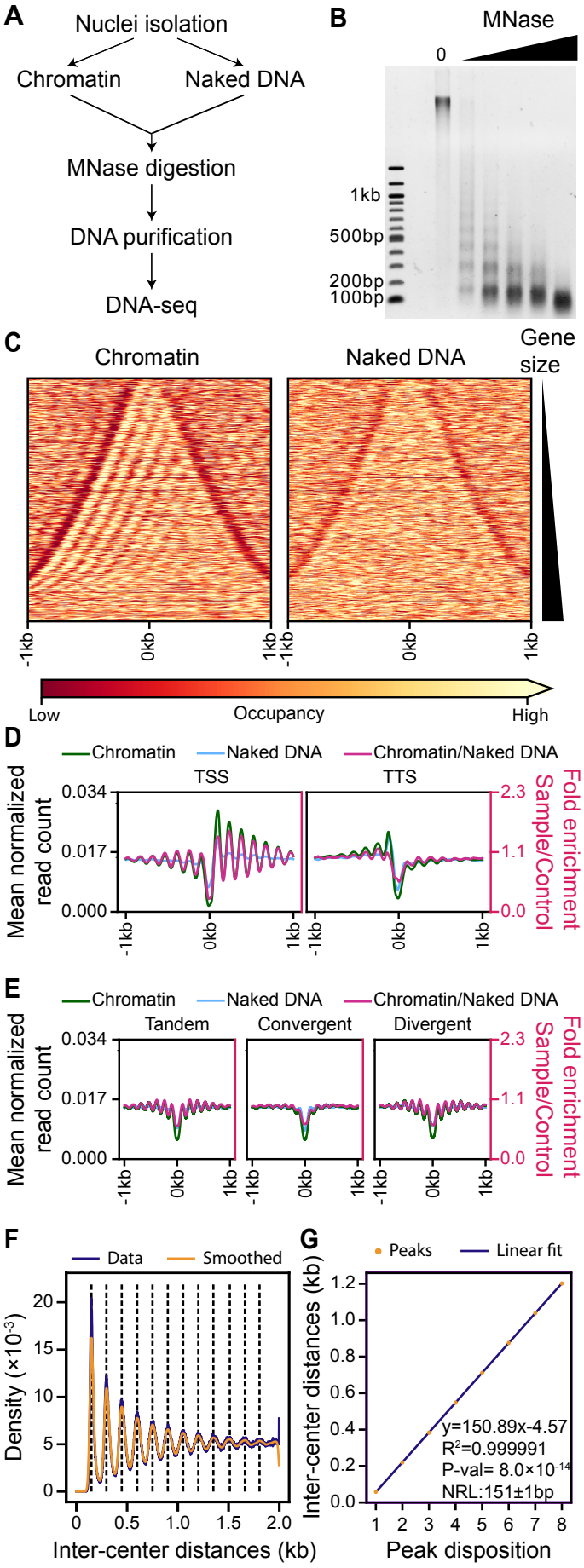
Figure 2. Inter-nucleosomal DNA is frequently associated with intron position. (A) Histogram showing exon size distribution (bin size = 25 bp): in blue, real exons; and in orange, simulated exons created assuming uniform exon sizes within each gene. **(B)** Histogram showing intron size distribution (bin size = 1 bp). **(C)** Example track reporting nucleosome occupancy over genes with intron locations indicated by vertical dashed lines. We can observe nucleosome free regions (NFRs) around the gene promoters and introns frequently associated with inter-nucleosomal DNA. **(D)** Heatmap showing nucleosome occupancy ± 200 bp around intron centers. Introns are ordered based on increasing distances from their center to the closest nucleosome center, from top to bottom. The average of the chromatin samples is shown on the left and the average of the naked DNA samples on the right, with the same color code as in Fig 1C. Vertical black dashed lines delineate the average size of an intron (25 bp). Individual samples are displayed in Supplemental Fig S2C. (**E**) Histogram reporting the distance of an intron center to the closest nucleosome center (red). For each intron, a random position inside the corresponding gene body was selected and the distance to its closest nucleosome center is reported (green). Bin size= 5 bp. **(F)** Schematic representation of the criteria to assign features for each intron (or exon) into one of the 3 classes, based on the distance (*d*) between its center and the closest nucleosome center position: central, $d \leq 25$ bp; proximal, 25 bp $< d <$ 50 bp; distal 50 bp $\leq d \leq$ 75 bp. **(G)** Relative distribution of introns, exons and both features over categories defined in panel F for the introns overlapping with a fixed nucleosome (about 70% of all introns, see Methods) and exons with a size below 300 bp overlapping with fixed peaks. See Supplemental Fig S2D including the features with $d >$ 75 bp.

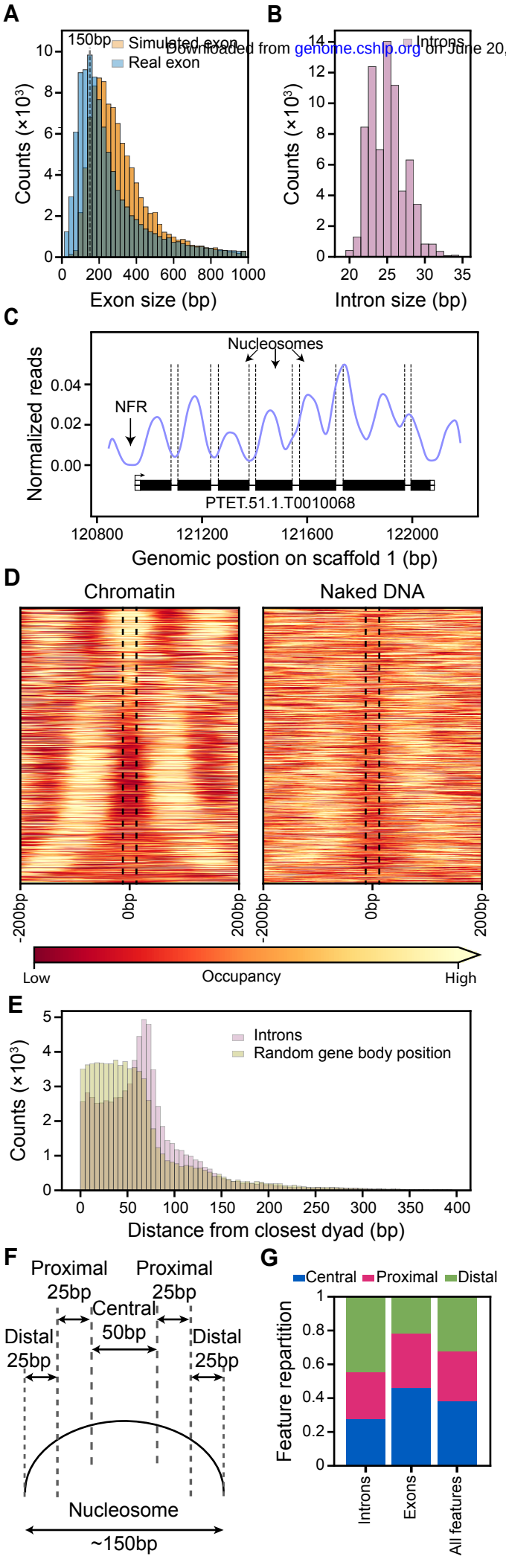
Figure 3. Nucleosome positioning is associated with intron-splicing efficiency. (A) Relative distribution of different classes of introns. Introns are grouped based on their length (3n or non-3n) and whether their retention causes a premature termination codon making

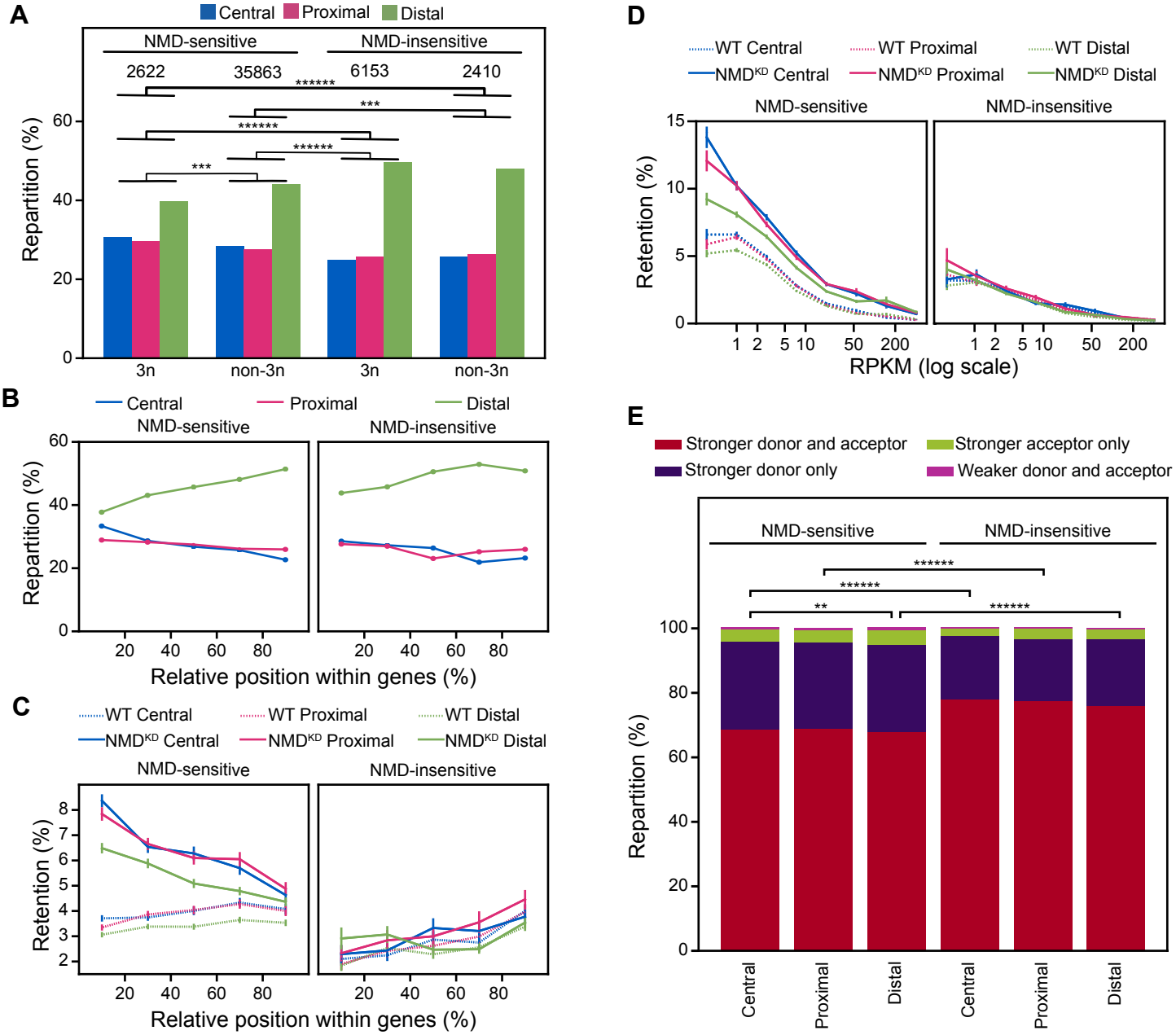
them sensitive to the nonsense-mediated decay mechanism (NMD-sensitive) or not (NMD-insensitive). Within each group, introns are classified based on the distance to the closest nucleosome center as in Fig 2G. P-values are calculated using the χ^2 test and only the significant ones are indicated. **(B)** Intron repartition according to the categories defined in Fig 2F as a function of their relative position within a gene. Introns are grouped based on their NMD sensitivity. Bin size= 20%. A barplot representation with relative p-values is displayed in Supplemental Fig S3A. **(C)** The retention rate of introns in WT (dashed lines) and in NMD-depleted (NMD^{KD}, solid lines) cells as a function of their relative position within a gene. Introns are grouped as in panel B. Error bars represent the standard error of the mean. P-values calculated using Mann-Whitley U test, and adjusted with false discovery rate (5%), are displayed in Supplemental Fig S3B. Bin size= 20%. **(D)** The retention rate of introns in WT (dashed lines) and in NMD-depleted (NMD^{KD}, solid lines) cells as a function of gene expression levels. Error bars represent the standard error of the mean. Colors and groups are as in panel B. P-values calculated using Mann-Whitley U test, and adjusted with false discovery rate (5%), are displayed in Supplemental Fig S3C. **(E)** Relative characterization of introns, within the same categories as in panel B, based on the strength of splicing acceptor and donor sites. P-values are calculated using the χ^2 test and adjusted with false discovery rate (5%). Tests were run between introns belonging to the same positional group or between introns belonging to the same NMD group. (P-value in all the plots *<0.05, **< 10^{-2}, ***< 10^{-3}, ****< 10^{-4}, *****< 10^{-5}, ******< 10^{-6}).

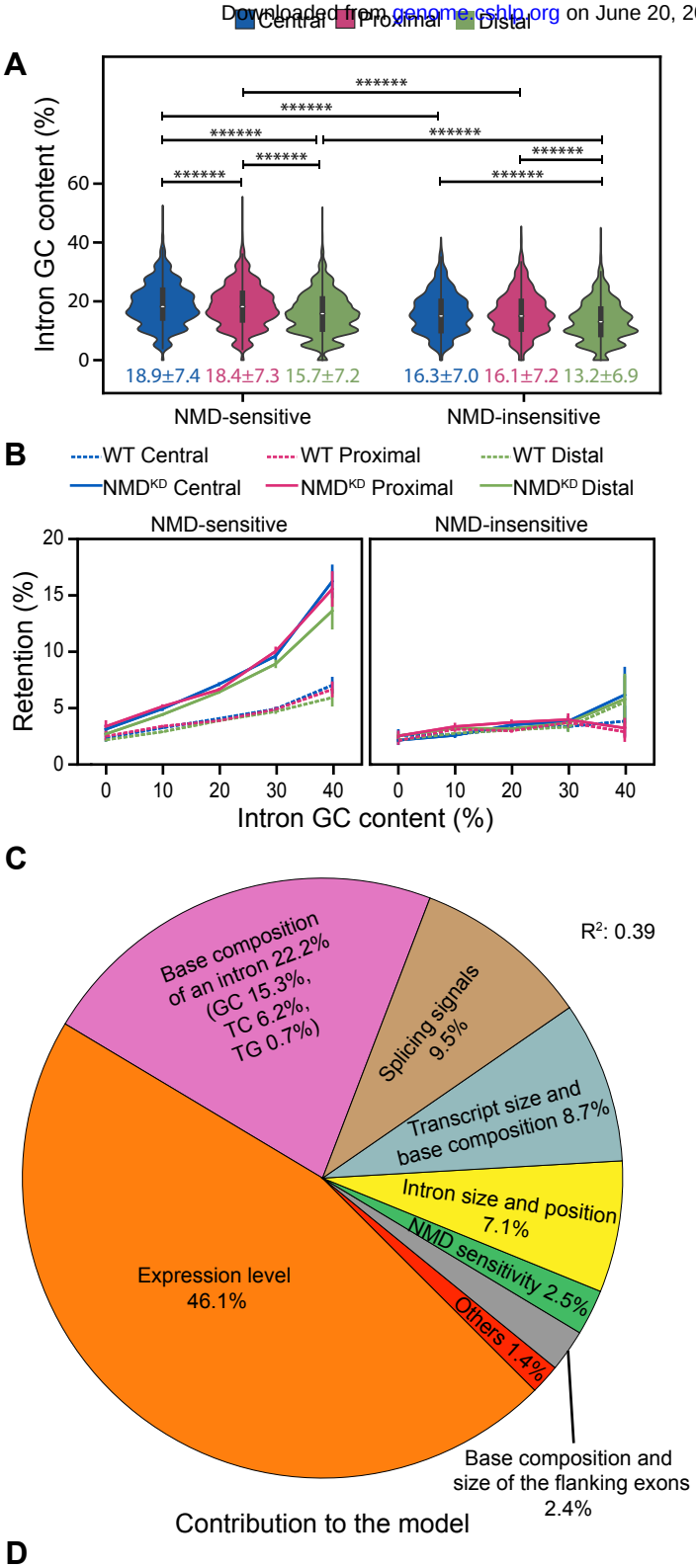
Figure 4. GC content related to nucleosome positioning contributes to intron-splicing efficiency. (A) GC content (%) distribution of introns based on the distance to the closest nucleosome center and NMD sensitivity. Mean and standard deviation for each group is reported at the bottom. P-values were calculated using the Mann–Whitney U test and adjusted using the false discovery rate (5%). Tests were run between introns belonging to the same positional group or between introns belonging to the same NMD group (P-value *<0.05, **< 10^{-2}, ***< 10^{-3}, ****< 10^{-4}, *****< 10^{-5}, ******< 10^{-6}). **(B)** The retention rate of introns in WT and NMD-depleted (NMD^{KD}) cells as a function of their GC content (excluded GT and AG dinucleotides at both extremities). Introns are classified based on their distance to the closest nucleosome center and on whether they are NMD sensitive or not. Binning = 10%. Error bars represent the standard error of the mean. P-values calculated using the Mann–Whitney U test, and adjusted using the false discovery rate (5%), are displayed in Supplemental Fig S4A. **(C)** Modelling Splicing Efficiency (SE) in NMD-depleted cells: the pie chart reports the contribution of each parameter or group of parameters used in the final model. The full list of retained parameters, reporting their contribution and their statistical significance, is displayed in Supplemental Table S1 as well as in Supplemental Fig S4C. **(D)**

841 The full fitted model in explaining intron splicing efficiency, indicating whether each
842 parameter is positively or negatively correlated with splicing efficiency. The parameter
843 abbreviations are explained in Supplemental Table S1.
844









D

$$e^{SE} = 0.295 \times \log_{10}(EL) - 0.170 \times GC + 0.098 \times SA + 0.109 \times TC + 0.09 \times N + 0.054 \times D_{TSS} + 0.052 \times L_I - 0.037 \times NMD + 0.067 \times SD + 0.074 \times L_T - 0.03 \times GC_T - 0.023 \times L_{FE} - 0.024 \times TG_I + 0.017 \times 3N + 0.032 \times TG_T - 0.032 \times TC_T + 0.026 \times GC_{FE} + 0.013 \times GC_{PE} + 0.023 \times TG_{FE} - 0.019 \times TC_{FE} + 0.011 \times D_N - 0.021 \times \Delta G_{p50} + 0.01 \times M_I + 0.018 \times \Delta G_{p50\text{-}I} + 0.022 \times \Delta G_{pf50\text{-}I} - 0.023 \times TG_{PE} + 0.009 \times M_{FE}$$