nature structural & molecular biology

Nucleosome positioning as a determinant of exon recognition

Hagen Tilgner[1,3], Christoforos Nikolaou[1,3], Sonja Althammer[1], Michael Sammeth[1], Miguel Beato[1], Juan Valcárcel[1,2] & Roderic Guigó[1]

Chromatin structure influences transcription, but its role in subsequent RNA processing is unclear. Here we present analyses of high-throughput data that imply a relationship between nucleosome positioning and exon definition. First, we have found stable nucleosome occupancy within human and *Caenorhabditis elegans* exons that is stronger in exons with weak splice sites. Conversely, we have found that pseudoexons—intronic sequences that are not included in mRNAs but are flanked by strong splice sites—show nucleosome depletion. Second, the ratio between nucleosome occupancy within and upstream from the exons correlates with exon-inclusion levels. Third, nucleosomes are positioned central to exons rather than proximal to splice sites. These exonic nucleosomal patterns are also observed in non-expressed genes, suggesting that nucleosome marking of exons exists in the absence of transcription. Our analysis provides a framework that contributes to the understanding of splicing on the basis of chromatin architecture.

Eukaryotic gene expression relies on the function of a battery of complex molecular machineries that execute the genetic program, from transcription and processing of primary RNAs in the nucleus to mRNA translation in the cytoplasm. These processes are highly coordinated, and their functional interplay opens a wealth of opportunities for gene regulation[1,2]. For example, there is abundant evidence for the functional coupling between transcription and pre-mRNA splicing[3,4] and for a role of this coupling in alternative splicing regulation[5–9].

Direct interactions of RNA processing factors with the largest subunit of RNA polymerase II (RNAPII)[10] through its C-terminal domain (CTD)[11] provide a mechanism for efficient co-transcriptional delivery of basal and regulatory splicing factors on nascent transcripts. In addition, some promoter-associated transcription factors and co-regulators also show splicing activities and/or recruit components of the splicing machinery[12–14]. Conversely, packaging of nascent transcripts with RNA-binding proteins precludes extended RNA-DNA hybrids, facilitates transcription elongation and prevents genomic instability[15,16]. In turn, transcription elongation factors, including splicing regulators that actively promote transcription elongation[17], can affect alternative splicing decisions by modulating the timing at which competing splice sites become available in nascent transcripts[18–21].

Eukaryotic DNA is wrapped around nucleosomes, the packaging units of chromatin, and this architecture is a key determinant of all aspects of DNA metabolism[22]. Nucleosomes contain two sets of four histone molecules. Chromatin remodeling is frequently associated with a combinatorial code of post-translational modifications of the flexible N-terminal histone tails, which can regulate chromatin compaction and the accessibility of factors responsible for DNA replication, recombination, repair and transcription[23]. The tight coupling between transcription and RNA processing opens the intriguing possibility that chromatin architecture and dynamics have a role in subsequent steps of the gene expression pathway[24]. Indeed, the Brahma subunit of the chromatin-remodeling complex SWI/SNF has been shown to interact with splicing factors, influence the accumulation of RNAPII on alternative exons and regulate alternative splicing, probably through local changes in transcription elongation[7].

A general link between nucleosomes and the gene exon-intron architecture has been proposed by Trifonov and colleagues[25–27]. These authors observed that the distance between consecutive 5′ or 3′ splice sites shows a periodicity reminiscent of the unit length of DNA wrapping around nucleosomes[25], suggesting that nucleosomes are somehow phased with the sequences that direct intron removal. On the basis of DNA sequence patterns that are characteristic of stably positioned nucleosomes, they predicted that splice sites are frequently located near the nucleosome dyad axis, a preference that they relate to the need to protect splice sites from mutation[26,27].

Genome-wide analyses of nucleosome occupancy in worms and humans have been recently published[28,29]. Using these data, we set out to investigate the relationship between gene architecture and stable nucleosome positioning on the genomes of these species' sequences. Thus, here and in the accompanying paper by Schwartz *et al.*[30], we report that stably positioned nucleosomes are more frequent in exons than in the surrounding introns, a trend that is more pronounced in exons flanked by weak splice sites and that contrasts with the decreased occupancy observed in pseudoexons. These observations are strongly suggestive of a role for chromatin organization in RNA processing. Our results also offer an explanation for the exonic enrichment of particular histone modifications recently reported[31] and suggest that

[1]Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain, [2]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia, Spain, [3]These authors contributed equally to this work. Correspondence should be addressed to R.G. (roderic.guigo@crg.cat).

Received 10 May; accepted 21 July; published online 16 August 2009; doi:10.1038/nsmb.1658

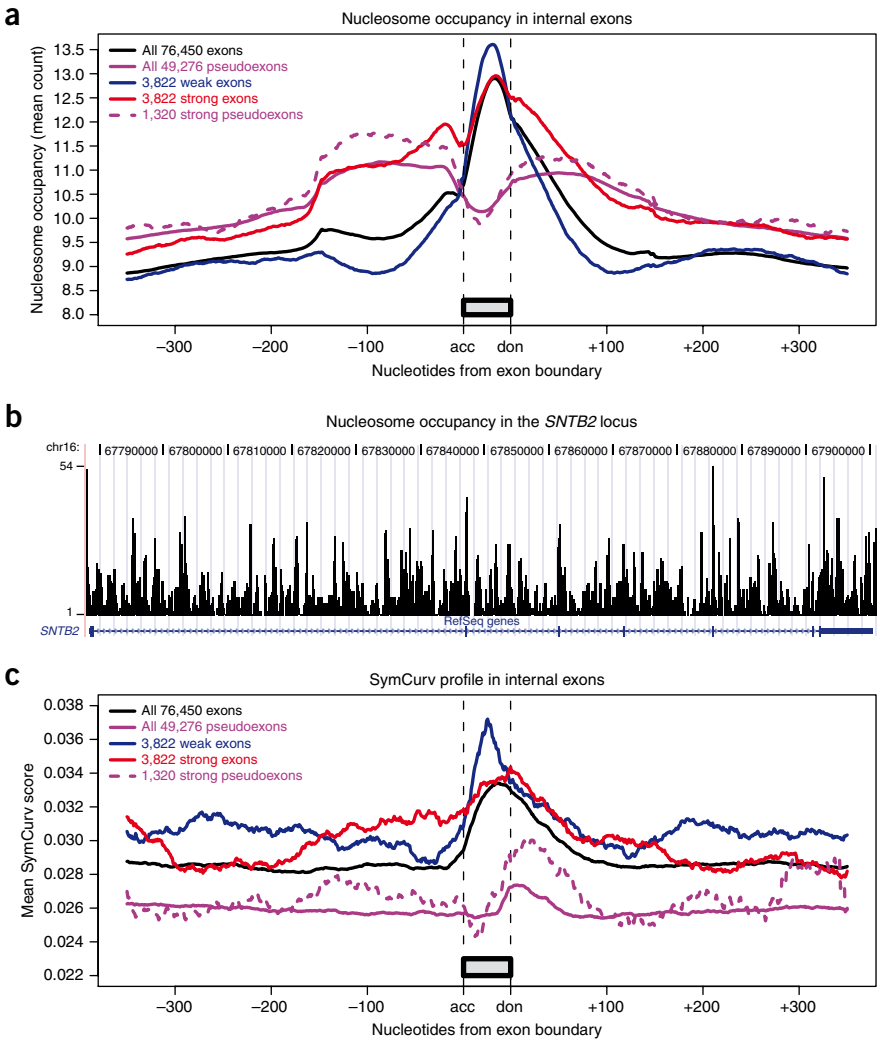

Figure 1 Observed and predicted nucleosome occupancy. (**a**) Nuclesome-occupancy profile across human internal constitutive exons in resting CD4+ T cells. We have computed the number of extended nucleosome reads overlapping each nucleotide. Upstream and downstream of an idealized internal exon, we plot the average number of nucleosome reads per nucleotide position, with negative positions relative to the acceptor (acc) site and positive positions relative the donor (don) site. Within the exon, reads have been mapped to 50 identically spaced intervals, irrespective of the length of the exon (see Online Methods). Strong exons are exons with a combined donor and acceptor score among the highest 5%; weak exons are the exons with a combined score among the lowset 5%; pseudoexons are intronic sequences bounded by splice sites; strong pseudoexons are exons with a combined score higher than the 90% percentile of real exons. (**b**) Nucleosome-occupancy profile in the human *SNTB2* locus in chromosome 16. All internal exons are clearly marked by nucleosome peaks from 28. High nucleosome peaks mark the first and fourth internal exons and the terminal exon. Notably, the first and the fourth internal exons are the exons with the weakest combined splice sites scores (first: 16.77; second: 20.60; third: 17.84; fourth: 12.26; fifth: 20.10). (**c**) Computationally predicted nucleosome-occupancy profile across human acceptor sites. The SymCurv score at each nucleotide has been averaged over all exons and pseudoexons in a way similar to that used for the nucleosome reads (**Supplementary Methods**).

the higher GC content in exons could partially result from selection to maintain sequences facilitating positioning of nucleosomes.

RESULTS

Nucleosome enrichment in human internal exons

We have analyzed data produced by Schones *et al.*[28] on genome-wide mapping of nucleosome density in resting and activated human CD4+ T cells. These data were generated by direct Solexa high-throughput sequencing of DNA purified from nucleosomes obtained by micrococcal nuclease (MNase) digestion of chromatin preparations. We extended the short sequence reads including neighboring sequences to the expected nucleosome length (147 base pairs (bp)). The number of reads mapping to a given region (nucleotide) can be assumed to be a measure of nucleosome occupancy in that region or nucleotide. When comparing the profile of nucleosome occupancy in resting CD4+ T cells with the positions of internal exons of protein-coding genes (see Online Methods) that were classified as constitutive by the AStalavista system[32] (**Supplementary Methods**), we observed strong nucleosome occupancy within the exons of human genes (**Fig. 1a**,**b**). In contrast, pseudoexons—that is, nonrepetitive intronic regions flanked by strong splice sites but showing no evidence of inclusion in mRNA, as judged by the available ESTs and cDNAs (**Supplementary Methods**)—show a weak nucleosome depletion (**Fig. 1a**). We have computed the nucleosome-occupancy ratio as the ratio between the average nucleosome occupancy per nucleotide within the exon and the average nucleosome occupancy per nucleotide in the 147-bp regions upstream and downstream of the exon. The median of the logarithm of this ratio is positive for exons (0.23) and negative for pseudoexons (−0.14), a difference that is highly significant ($P < 2.2 \times 10^{-16}$, according to both the one- and two-sided Mann-Whitney U and the Kolmogorov-Smirnov tests).

Notably, the intensity of nucleosome occupancy is inversely related to exon splice site strength. Indeed, we have computed the scores of the splice sites and ranked the exons according to the sum of the acceptor and donor scores (**Supplementary Methods**). We have considered the lowest-scoring 5% of exons as weak exons and the highest-scoring 5% of exons as strong exons—with these terms referring only to the strength of the exons splice sites and not to their inclusion level. We have found that nucleosome occupancy is stronger in weak than in strong exons (**Fig. 1a**). The median of the logarithm of the nucleosome occupancy ratio is 0.38 for weak exons and 0.11 for strong exons (a statistically significant difference; $P < 2.2 \times 10^{-16}$). In exons with strong splice sites, in contrast, an extended region of nucleosome occupancy occurs upstream of the 3′ splice site. This region is also observed in pseudoexons and is more accentuated in pseudoexons flanked by strong splicing signals. As a result, pseudoexons with strong splice sites—in which splicing does not occur despite the strength of the sites—show a pattern of nucleosome occupancy that is the mirror image of that observed on exons with weak sites—in which splicing occurs despite the weakness of the sites (**Fig. 1a**). A similar pattern is observed in activated cells, albeit less sharp (**Supplementary Fig. 1**). These observations strongly suggest a relationship between

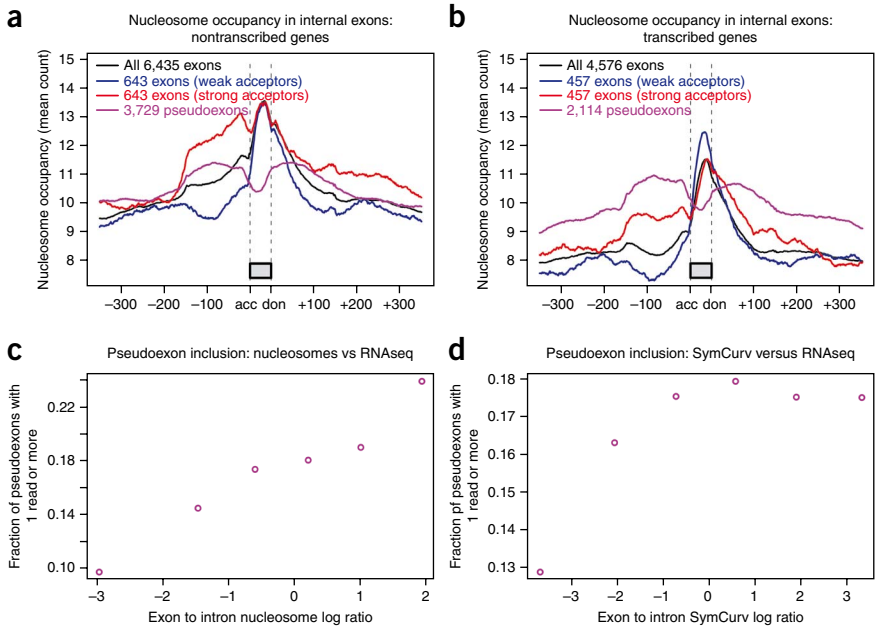

Figure 2 Nuclesome occupancy and expression of genes and exons. (**a**) Nuclesome-occupancy profile across internal acceptor sites from genes that are not expressed in resting CD4+ T cells. Gene expression has been determined using the Affymetrix platform[28]. We plot the average number of nucleosome reads per position in all exons considered together (black), only in exons with strong (red) and weak (blue) acceptor sites, and in intronic pseudoexons. (**b**) Nucleosome-occupancy profile across internal acceptor sites from genes expressed in resting CD4+ T cells, shown as in **a**. (**c**,**d**) Nucleosome-occupancy ratio versus inclusion of pseudoexons. Large values of the nucleosome-occupancy ratio are typical of bona fide exons. Nucleosome occupancy has been measured using both the nucleosome sequence reads obtained experimentally by Schones *et al.*[28] (**c**) and our SymCurv theoretical predictions (**d**). In each case, we have binned the set of pseudoexons into six classes depending on this ratio (**Supplementary Methods**). Within each class, we have computed the proportion of pseudoexons that have evidence of inclusion (at least one read from one tissue mapping entirely within the pseudoexon) according to a published RNAseq data set[33].

nucleosome occupancy and exon recognition during pre-mRNA splicing. Specifically, nucleosome occupancy within the exon may promote exon inclusion—an effect that is particularly relevant in exons with weak splice sites, whereas nucleosome depletion within the exon and stable nucleosome occupancy upstream of the acceptor site—as observed in pseudoexons—could have a repressing effect. As an example, **Figure 1b** shows the relationship between peaks of nucleosome occupancy and internal or 3′ exons in the human syntrophin β2 (*SNTB2*) gene.

Further supporting this relationship is the pattern of nucleosome occupancy computed theoretically using the SymCurv algorithm (http://genome.crg.cat/software/#SymCurv), which predicts nucleosomal sequences on the basis of the symmetry of the DNA curvature (**Supplementary Methods**). The SymCurv computational predictions on human exons closely reproduced the pattern of exon nucleosome occupancy that we observed experimentally in CD4+ T cells, including the differential behavior between weak and strong exons (**Fig. 1c**).

Exonic nucleosome enrichment is not transcription-dependent

The SymCurv nucleosome-occupancy profile is based exclusively on the structural properties of the DNA sequence, suggesting that nucleosomes mark exons in chromatin in the absence of transcription. Indeed, after analyzing expression data from resting CD4+ T cells[28] (see Online Methods), we found that non-expressed genes show an exonic nucleosome-occupancy pattern that is similar, albeit less sharp, than that observed in expressed genes (**Fig. 2a**,**b**). Notably, although expressed genes show reduced frequencies of stably positioned nucleosomes overall (**Fig. 2a**,**b**), this reduction is less prominent within weak exons and upstream of pseudoexons, consistent with the hypothesis that nucleosome positioning has a particularly relevant role in regulating the splicing of these elements.

Nucleosome occupancy predicts inclusion of exons

As our definition of pseudoexons is relative to the available transcript evidence, we cannot rule out that some of the pseudoexons in our data set may actually be included in some cell type or condition that has not yet been surveyed. In this regard, as a corollary of our observations, we hypothesized that pseudoexons with a nucleosome-occupancy pattern similar to that of 'bona fide' exons—that is, nucleosome enrichment within the exons and depletion in the flanking regions—may actually be included as real (alternative) exons in some cell type. For each pseudoexon, we have thus computed the log ratio of nucleosome occupancy within the pseudoexon over the flanking intronic regions (**Supplementary Methods**), using both SymCurv scores and sequence reads. We have binned the set of pseudoexons in six classes depending on these ratios. Within each bin, we assessed inclusion rate using RNAseq data recently obtained using the Solexa platform in nine human tissues[33]. Consistent with our hypothesis, the larger the nucleosome-occupancy ratio, the larger the proportion of pseudoexons with evidence of inclusion (that is, with at least one read from one tissue mapping entirely within it; **Fig. 2c**,**d**).

Nucleosome positioning contributes to global exon definition

We binned the set of exons with the weakest 10% of acceptor sites according to their length. Short exons (less than 90 bp long) show a weak nucleosome-occupancy peak downstream of the acceptor sites. The peak grows with exon length and 'moves' toward the center of the exon, a pattern that closely reproduces the partitioning of exons in length classes (**Fig. 3a**). A less sharp pattern is observed in exons with the strongest 10% of acceptor sites (**Fig. 3b**). These observations are more compatible with nucleosome positioning defining the exon globally rather than specifically affecting the acceptor or the donor site. However, our analysis also suggests that nucleosome occupancy may have an additional role in the definition of acceptor 3′ sites. Indeed, the peak nucleosome pattern is obvious downstream of weak acceptor sites of terminal exons (**Fig. 3c**). In contrast, the peak nucleosome pattern is weak upstream of the donor sites of initial exons, and no differences can be observed here between weak and strong donor sites (**Fig. 3d**).

Exon marking by histone modifications and nucleosomes

Trimethylation of Lys36 in Histone 3 (H3K36me3) has been recently described as a marker of exons in expressed genes from *C. elegans*[31]. We have analyzed ChIP-Seq data on this modification recently obtained in CD4+ T cells[34] (see Online Methods). Consistent with

the previous observation[31], we also observed a peak of H3K36me3 within human internal exons from expressed genes (**Fig. 4a**). The peak showed the differences between weak and strong exons, albeit less sharp, that we had also observed for nucleosomes. We therefore performed a crude and qualitative normalization of the H3K36me3 profile against the nucleosome-occupancy profile. At each nucleotide position, we simply divided the number of H3K36me3 reads by the number of nucleosome reads overlapping the position (see Online Methods). After processing the data in this way, the H3K36me3 peak within exons essentially vanished (**Fig. 4b**). This indicates that the H3K36me3 peak mostly reflects underlying nucleosome occupancy. For another epigenetic mark (H4K20me1), however, nucleosome normalization uncovers a potential anticorrelation with exon positions that is not apparent from the raw data (**Fig. 4c**,**d**).

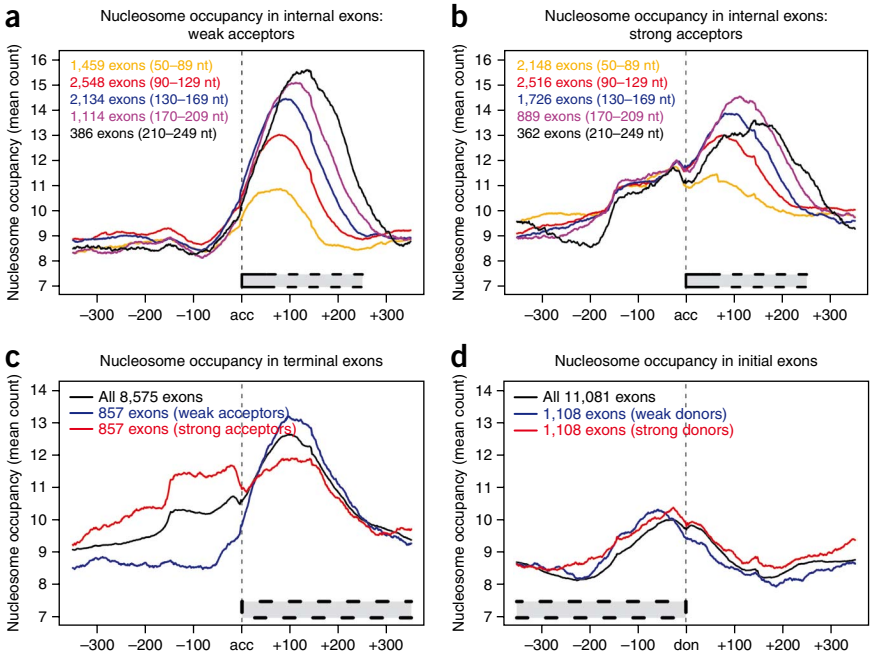

Figure 3 Nucleosome occupancy in internal exons of different lengths, initial exons and terminal exons. (**a**,**b**) Nucleosome-occupancy profiles across internal acceptors for different exon length classes. We plot the average number of nucleosome reads per position. Positions are aligned at the acceptor (acc) site. Different profiles are plotted (using different colors) for different exon length classes. Nucleosome occupancy is shown for weak acceptors and strong acceptors (**b**). (**c**) Nucleosome-occupancy profile in terminal exons. Positions have been aligned at the acceptor site. (**d**) Nucleosome-occupancy profile in initial exons. Positions have been aligned at the donor (don) site.

Nucleosome enrichment in *C. elegans* exons

We have mapped high-throughput sequence data on nucleosome positioning obtained using the SOLiD platform in *C. elegans*[29] across internal constitutive exons (see Online Methods). The mapping reveals a well-defined peak of nucleosome occupancy within internal exons from *C. elegans* genes. (**Supplementary Fig. 2**). Such a peak is not observed in the DNA used as a control in these experiments, demonstrating that GC sequencing bias is not confounding our observations.

GC content and nucleosome occupancy in human exons

Human exons tend to be GC-rich when compared to the surrounding intronic regions. Because nucleosome sequences have also been postulated to prefer GC-rich regions[35,36], it is a possibility that nucleosome positioning within exons could be mediated by increased GC content. We therefore computed the profile of GC content in human internal exons (**Supplementary Fig. 3** and **Supplementary Methods**). The profile (**Supplementary Fig. 3**) is indeed markedly similar to that of nucleosome occupancy, including the higher GC content in weak than in strong exons and the reduced GC content within pseudoexons. GC content by itself, however, cannot fully explain the pattern of

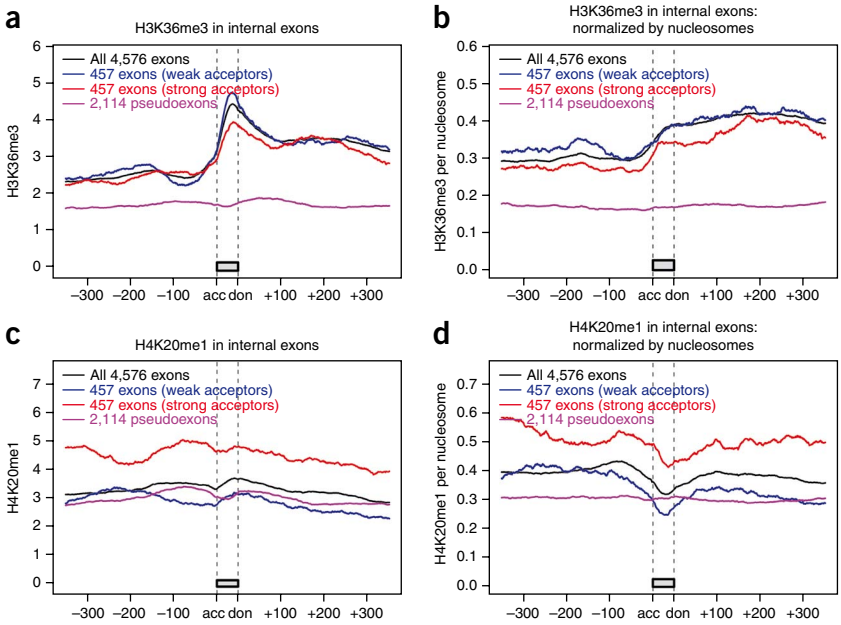

Figure 4 Profile of histone modifications in expressed genes in resting CD4$^+$ T cells. (**a**) H3K36me3. Note that the plateau downstream from the exonic peak of H3K36me3 is higher than the plateau upstream. This is in agreement with previous work[34] that showed that the levels of H3K36me3 increase 3′ to 5′ along the transcript. Our results indicate that the increase is not entirely linear but that it occurs, at least partially, in a stepwise fashion with the exons. Acc, acceptor; don, donor. (**b**) H3K36me3 normalized by nucleosome occupancy. (**c**) H4K20me1 without normalization. (**d**) H4K20me1 normalized by nucleosome occupancy. The plots of the raw data (**a**,**c**) were generated in a similar way to that for nucleosomes (**Fig. 1a** and Online Methods), but relying on ChIP-Seq data for histone modifications[34]. Normalized plots (**b**,**d**) were obtained after dividing the values corresponding to histone modifications by those corresponding to nucleosome occupancy (see Online Methods for details).

nucleosome occupancy that we have observed in human exons. First, the region upstream of the donor sites of initial exons is more GC-rich than the region downstream from acceptors of terminal exons (63% on average versus 51%), but nucleosome occupancy is higher in the latter (9.8 extended read counts per nucleotide on average versus 12.0; **Fig. 3**). Second, we have computed the correlation coefficient between the raw nucleosome occupancy and the GC content in exons and pseudoexons. Although the correlation is positive and significant in both cases, it is much higher for pseudoexons (0.422) than for exons (0.182), suggesting that factors other than GC content (for instance, factors involved in splicing) have a stronger influence on nucleosome occupancy in exons than in pseudoexons. Finally, we have selected subsets of exons and pseudoexons that are almost identical in terms of their log ratio of GC content between the exon (pseudoexon) and the flanking intronic regions (**Supplementary Methods**). Even in these subsets, exons have a significantly higher nucleosome occupancy log ratio than pseudoexons do (0.084 versus −0.061, $P < 2.2 \times 10^{-16}$; **Supplementary Fig. 4**). An equivalent conclusion can be reached when comparing weak and strong exons (**Supplementary Fig. 5**).

Nucleosome enrichment in noncoding exons

To further investigate the relationship between coding function, GC content and nucleosome occupancy, we investigated exons from noncoding transcripts. We considered the noncoding genes from the GENCODE annotation[37] (see Online Methods). Noncoding exons also have a higher GC content than the surrounding intronic areas (**Supplementary Fig. 6a**). Notably, they also show strong nucleosome occupancy (**Supplementary Fig. 6b**).

DISCUSSION

Taken together, we believe that our analyses suggest a role for chromatin structure in splicing. More specifically, the interplay between nucleosome positioning within exons and upstream from the acceptor sites seems to contribute to exon recognition: nucleosome positioning within the exon coupled with nucleosome depletion upstream from the acceptor site would promote inclusion of exons with weak splice sites, whereas nucleosome depletion within the exon coupled with stable nucleosome occupancy upstream of the acceptor site would have a repressing effect.

Although the evidence for the relationship is convincing, the molecular mechanisms by which nucleosome positioning influence splice site recognition remain to be elucidated. A possible mechanism would be mediated by changes in transcription elongation rates caused by the presence of a positioned nucleosome near the splice sites. Such changes are indeed known to influence splice site selection[5–9,20,21]. It is conceivable that the presence of a stably positioned nucleosome reduces the elongation rate of the polymerase complex, and this, in turn, provides a window of opportunity for RNAPII CTD–associated splicing factors to interact with splice sites. Alternatively, nucleosomes could contribute to specifically recruit some splicing factors during transcription, or to co-transcriptionally enhance the molecular interactions by which splicing factors bound at the flanking splice sites stabilize each other. This phenomenon, known as exon definition, has been linked to the optimal length of internal exons, which may provide an optimal distance to accommodate direct or indirect interactions between factors involved in early recognition of the flanking 3′ and 5′ splice sites[38]. In this regard, it is notable that the average length of human internal exons (151 bp, for all exons, not only those in our size-selected data set) is similar to that of the nucleosome sequences (approximately 147 bp) and that this similarity is greater and more constrained for exons with weak splice sites (mean 153 bp, s.d. 177 bp), where nucleosomes are positioned more stably, than for exons with strong splice sites (mean 164 bp, s.d. 313 bp). Thus, nucleosome positioning in internal exons can contribute to the proper positioning of molecular interactions across the exon that characterize the process of exon definition. The fact that the splicing process can occur *in vitro* on exogenously added RNA molecules clearly demonstrates that nucleosome positioning is not a prerequisite for splicing to occur. Splicing is, however, substantially more efficient when coupled to transcription[39], and nucleosome positioning may further increase this efficacy.

Our analyses indicate that positioning of nucleosomes within exons is not dependent on transcription. Nucleosome organization along the genome would therefore partially reflect the underlying exonic structure of genes and, thus, constitute a code for splicing present in the DNA but not in the sequence of the primary transcript. Our analyses also suggest that the enrichment of certain histone modifications, notably H3K36me3, in exons[31] is, at least partially, the reflection of the enrichment of stably positioned nucleosomes within exons of active genes. Indeed, in human $CD4^+$ T cells, when normalized with respect to nucleosome occupancy, the H3K36me3 peak within exons essentially disappears. Notably, however, nucleosome enrichment upstream of the acceptor sites that we observe in strong exons does not seem to correlate with depletion of H3K36me3. Also, even after normalizing for nucleosome occupancy, some histone modifications (H4K20me1, for instance) show a characteristic exonic pattern. These observations argue that, as suggested[31], histone modifications may indeed have a role in splicing. In our opinion, however, this role can be fully understood only when the underlying pattern of nucleosome occupancy is taken into account.

We have found that the pattern of GC content on human exons is markedly similar to the pattern of nucleosome occupancy but that GC content alone can not fully explain the pattern of nucleosome occupancy observed within exons. One is tempted to speculate that the elevated GC content of exons may partially result from the need to accommodate nucleosome sequences, which have been postulated to also prefer GC-rich regions[35,36]. Indeed, the increased GC content in exons has often been attributed to the codon biases that resulting from protein-coding functionality. Among other hypotheses, GC-rich codons would be preferred because of the higher relative abundance of the cognate tRNAs (see, for example, refs. 40,41), which would lead to increased efficiency of translation. However, we have also found elevated GC content within exons from noncoding RNAs (**Supplementary Fig. 6a**), where a codon bias cannot be invoked. Notably, the fact that we have also found nucleosome enrichment within noncoding exons (**Supplementary Fig. 6b**) would support the hypothesis that GC content in noncoding exons (as well as in coding exons) could at least partially be the consequence of selection to favor splicing-related nucleosome occupancy. Further supporting this hypothesis is the striking observation of reduced GC content within pseudoexons (**Supplementary Fig. 3**). Low GC content would explain nucleosome depletion in pseudoexons (**Fig. 1a**), which in turn would have a repressive effect on their inclusion in mRNA sequences. Although other selective pressures that are not related to translation efficiency could explain GC enrichment within noncoding exons, such as DNA stability[42], RNA structure[43] or RNA processsing[44,45], such selective pressures cannot easily explain GC reduction within pseudoexons.

Several interesting examples of connections between chromatin structure and RNA processing have been reported[6–9,20,21], but our findings provide a general concept for how the architecture of genome packaging can influence pre-mRNA splicing. Despite great progress, the determinants of splice site identification are not totally

understood, and it is not possible to predict from the analysis of the primary RNA sequence alone the resulting pattern of splicing products. Our results indicate that some of these determinants may actually reside outside the primary transcript, in the chromatin structure itself, and represent instructions for splicing encoded in the DNA sequence but not in the sequence of the primary transcript. Taking this concept into account should provide a new framework to understand features of splice site recognition, exon definition and alternative splicing on the basis of chromatin architecture.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank D.E. Schones for help with the data and its interpretation and members of the Guigó laboratory, especially D. Gonzalez, for help with data analysis. This work was supported by the Spanish Ministry of Science with fellowships to M.S. and S.A., and with grant number BIO2006-03380 to R.G.

AUTHOR CONTRIBUTION

H.T., C.N., S.A. and M.S. performed computational analysis; R.G., J.V. and M.B. desgined the analysis and wrote the paper; all authors discussed the data.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506 (2002).
2. Moore, M.J. & Proudfoot, N.J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688–700 (2009).
3. Bentley, D.L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr. Opin. Cell Biol.* **17**, 251–256 (2005).
4. Pandit, S., Wang, D. & Fu, X.D. Functional integration of transcriptional and RNA processing machineries. *Curr. Opin. Cell Biol.* **20**, 260–265 (2008).
5. Kornblihtt, A.R. Coupling transcription and alternative splicing. *Adv. Exp. Med. Biol.* **623**, 175–189 (2007).
6. Kadener, S. *et al.* Antagonistic effects of T-Ag and VP16 reveal a role for RNA Pol II elongation on alternative splicing. *EMBO J.* **20**, 5759–5768 (2001).
7. Batsché, E., Yaniv, M. & Muchardt, C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat. Struct. Mol. Biol.* **13**, 22–29 (2006).
8. Sims, R.J., III *et al.* Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol. Cell* **28**, 665–676 (2007).
9. Schor, I.E., Rascovan, N., Pelisch, F., Allo, M. & Kornblihtt, A.R. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc. Natl. Acad. Sci. USA* **106**, 4325–4330 (2009).
10. Das, R. *et al.* SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Mol. Cell* **26**, 867–881 (2007).
11. Phatnani, H.P. & Greenleaf, A.L. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.* **20**, 2922–2936 (2006).
12. Nogues, G., Kadener, S., Cramer, P., Bentley, D. & Kornblihtt, A.R. Transcriptional activators differ in their abilities to control alternative splicing. *J. Biol. Chem.* **277**, 43110–43114 (2002).
13. Auboeuf, D., Honig, A., Berget, S.M. & O'Malley, B.W. Coordinate regulation of transcription and splicing by steroid receptor coregulators. *Science* **298**, 416–419 (2002).
14. Monsalve, M. *et al.* Direct coupling of transcription and mRNA processing through the thermogenic coactivator PGC-1. *Mol. Cell* **6**, 307–316 (2000).
15. Li, X. & Manley, J.L. Cotranscriptional processes and their influence on genome stability. *Genes Dev.* **20**, 1838–1847 (2006).
16. Luna, R., Gaillard, H., Gonzalez-Aguilera, C. & Aguilera, A. Biogenesis of mRNPs: integrating different processes in the eukaryotic nucleus. *Chromosoma* **117**, 319–331 (2008).
17. Lin, S., Coutinho-Mansfield, G., Wang, D., Pandit, S. & Fu, X.D. The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.* **15**, 819–826 (2008).
18. de la Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing *in vivo*. *Mol. Cell* **12**, 525–532 (2003).
19. Howe, K.J., Kane, C.M. & Ares, M. Jr. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA* **9**, 993–1006 (2003).
20. Muñoz, M.J. *et al.* DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* **137**, 708–720 (2009).
21. Allo, M. *et al.* Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat. Struct. Mol. Biol.* **16**, 717–724 (2009).
22. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413–417 (2007).
23. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
24. Allemand, E., Batsche, E. & Muchardt, C. Splicing, transcription, and chromatin: a *ménage à trois*. *Curr. Opin. Genet. Dev.* **18**, 145–151 (2008).
25. Beckmann, J.S. & Trifonov, E.N. Splice junctions follow a 205-base ladder. *Proc. Natl. Acad. Sci. USA* **88**, 2380–2383 (1991).
26. Denisov, D.A., Shpigelman, E.S. & Trifonov, E.N. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene* **205**, 145–149 (1997).
27. Kogan, S. & Trifonov, E.N. Gene splice sites correlate with nucleosome positions. *Gene* **352**, 57–62 (2005).
28. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
29. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
30. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron architecture. *Nat. Struct. Mol. Biol.* advance online publication, doi:10.1038/nsmb.1659 (16 August 2009).
31. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
32. Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. *PLOS Comput. Biol.* **4**, e1000147 (2008).
33. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
34. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
35. Kharchenko, P.V., Woo, C.J., Tolstorukov, M.Y., Kingston, R.E. & Park, P.J. Nucleosome positioning in human HOX gene clusters. *Genome Res.* **18**, 1554–1561 (2008).
36. Peckham, H.E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17**, 1170–1177 (2007).
37. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (Suppl 1), S4 (2006).
38. Berget, S.M. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414 (1995).
39. Das, R. *et al.* Functional coupling of RNAP II transcription to spliceosome assembly. *Genes Dev.* **20**, 1100–1109 (2006).
40. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**, 573–597 (1982).
41. Kotlar, D. & Lavner, Y. The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics* **7**, 67 (2006).
42. Jabbari, K., Clay, O. & Bernardi, G. GC3 heterogeneity and body temperature in vertebrates. *Gene* **317**, 161–163 (2003).
43. Katz, L. & Burge, C.B. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **13**, 2042–2051 (2003).
44. Duret, L. Detecting genomic features under weak selective pressure: the example of codon usage in animals and plants. *Bioinformatics* **18** (Suppl 2), S91 (2002).
45. Willie, E. & Majewski, J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends Genet.* **20**, 534–538 (2004).

ONLINE METHODS

Human exons and nucleosome data. We downloaded the human RefSeq transcripts[46] and GenBank mRNAs[47] (both relative to the hg18 version of the human genome) from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables)[48] on 12 August 2008. From a set of 25,818 RefSeq transcripts aligning to the human chromosomes 1 to 22 and X (including only one genomic location if the transcript mapped to multiple locations), we chose a set of 76,450 nonredundant internal exons that (i) had appropriate length for exon definition (between 50 nt and 250 nt, a little more conservative than the 50–300 nt mentioned before[38]), (ii) were classified as constitutive using the Astalavista framework[32] (**Supplementary Methods**) with RefSeq and mRNA exons (an internal exon was classified as constitutive, if and only if all annotated RefSeq transcripts and GenBank mRNAs whose transcript sequences overlapped the complete exon show the exon as part of their annotated gene structure), (iii) did not have any adjacent intron of U12 type (using geneid[49,50]) or of less than 70 nt and (iv) had AG acceptors and GT donors.

Consequently, we used initial and terminal exons in the analysis if all annotated RefSeq transcripts and GenBank mRNAs whose transcript sequences overlapped the splice site (donor for an initial exon; acceptor for a terminal exon) had this splice site as part of their annotated gene structure. Furthermore we used no length criterion for initial and terminal exons.

Exons from noncoding RNA. We extracted noncoding RNAs from the GENCODE annotation (the reference annotation being built within the framework of the ENCODE project[37]). We considered only noncoding RNAs that were not more than 1 kb away from the boundaries of the closest annotated protein-coding loci. This resulted in 3,019 transcripts (corresponding to 2,258 loci), from which 1,539 had at least one internal exon. We extracted internal exons from these transcripts as for coding exons. We retained exons only from transcripts that were classified as 'processed_transcript' or 'noncoding'. This resulted in a set of 1,403 internal exons from noncoding transcripts.

Nucleosome and histone modification data. To address nucleosome occupancy, we extended nucleosomal and histone modification reads that mapped uniquely to the genome[28,34] to the length of a full nucleosome (147 bp). The number of extended reads ('nucleosome occupancy') overlapping each genomic position was calculated at single-nucleotide resolution. To calculate the aggregate values for pseudo(exon) representation, we represented each category (exons, pseudoexons, weak and strong exons) by a series of 750 values. We calculated 350 intronic values for the upstream intron at each point as the average nucleosome occupancy of all (pseudo)exons in a category. We calculated 350 values for the downstream intron analogously. To represent an idealized (pseudo)exon with a fixed size despite the varying length of (pseudo)exons, we calculated 50 values as follows. The first value is the average nucleosome occupancy over all (pseudo)exons for a given category in a 7-bp window centered at nucleotide 1 of the (pseudo)exon. To the second value, every exon e with $l(e)$ nucleotides contributes with the average nucleosome occupancy in a 7-bp window centered around nucleotide $1 + 1*l(e)/50$, rounded to the next integer. These contributions are averaged to produce one single value. In the same way, for $i = 3,\ldots,50$ windows of 7 bp each centered around nucleotide $1+ (i-1)*l(e)/50$ (rounded to the next integer) are used. We used the 7-bp window approach to guarantee that all bases in all exons (up to 250 bp) would be taken into account. Averaging within windows made sure that no artificial overcounts were produced when projecting longer exons (up to 250 bp) to the idealized length of 50 bp.

For histone modification normalization against nucleosomes, we treated both nucleosomes and histone-modification data as described in the previous section, so that for nucleosomes and for each kind of histone modification 750 values (representing the upstream intron, the exon and the downstream intron) were obtained. Normalization for for example, H3K36me3 (H3K27me1, H4K20me1 and so on) was performed by dividing the 750 values for H3K36me3 by the corresponding values for nucleosomes. We chose deliberately to perform this at the average level of all (pseudo)exons and not for each (pseudo)exon separately to avoid biases due to excessive pseudocounts.

Transcript sets of transcribed and nontranscribed RefSeq transcripts. A list of nontranscribed ('absent' in their terminology) and transcribed ('present' in their terminology) UCSC transcript sets derived from microarray analysis in resting T cells was provided by D. Schones[28]. Using the correspondence tables provided by the UCSC genome browser[48], we labeled a RefSeq gene 'transcribed' (or nontranscribed, respectively) if and only if all corresponding UCSC transcripts were in the list of transcribed (or nontranscribed) UCSC transcripts. When comparing exons of transcribed versus those of nontranscribed genes, we defined the strength of exons and pseudoexons using only the strength of the acceptor. To obtain larger exon sets, 10% (instead of the previously used 5%) were chosen for the definition of 'weak' and 'strong'.

Human internal exons for length analysis. To investigate length constraints of exons with weak and strong splice sites without the artificially imposed minimal and maximal exon length, we chose a set of 158,725 internal RefSeq exons bounded by AG acceptors and GT donors, regardless of length, splicing type (constitutive or alternative) or the type of their surrounding introns (U2 or U12). For cases where an exon can be used by more than one transcript, the exon was counted only once. Again we scored all splice sites and determined weak and strong exons according to the sum of their acceptor and donor scores, as described before. Means and s.d. of the length distributions of weak and strong exons were calculated.

***Caenorhabditis elegans* exons and nucleosome data.** We downloaded *C. elegans* RefSeq transcripts and mRNAs (both relative to the ce4 version of the *C. elegans* genome) from the UCSC table browser on 24 November 2008. We chosen exons from these transcripts in a similar way as the human exons were chosen with the following differences. (i) Exons had to be surrounded by introns of a minimal length of 150 nt, and all introns were assumed to be of U2 type (because *C. elegans* is not known to contain any U12 introns[51,52]). (ii) No pseudoexons were defined, as most *C. elegans* introns are too short to harbor pseudoexons. *C. elegans* nucleosomal and control reads[29] were downloaded from the UCSC table browser. No extension was necessary as they were already of nucleosome size. We obtained nucleosome occupancies and control occupancies as well as aggregate plots as described for the human nucleosome data for both nucleosomes and control.

46. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
47. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic Acids Res.* **36**, D25–D30 (2008).
48. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
49. Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics*, Chapter 4: Unit 4.3 (2007).
50. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10**, 511–515 (2000).
51. Sheth, N. *et al.* Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* **34**, 3955–3967 (2006).
52. Alioto, T.S. U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.* **35**, D110–D115 (2007).

doi:10.1038/nsmb.1658