Chromatin accessibility determines intron retention in a cell type-specific manner

Veronika Petrova[1,2], Renhua Song[3,4], DEEP Consortium, Karl J.V. Nordström[5], Jörn Walter[5], Justin J.-L. Wong[3,4], Nicola J. Armstrong[6], John E.J. Rasko[2,4,7]*†, Ulf Schmitz[1,2,4]*†

[1] Computational BioMedicine Laboratory Centenary Institute, The University of Sydney, Camperdown 2050, Australia

[2] Gene and Stem Cell Therapy Program Centenary Institute, The University of Sydney, Camperdown 2050, Australia

[3] Epigenetics and RNA Biology Program Centenary Institute, The University of Sydney, Camperdown 2050, Australia

[4] Faculty of Medicine and Health, The University of Sydney, Camperdown 2050, Australia

[5] Laboratory of EpiGenetics, Saarland University, Campus A2 4, 66123 Saarbrücken, Germany

[6] Mathematics and Statistics, Curtin University, Bentley, WA 6102, Australia

[7] Cell and Molecular Therapies, Royal Prince Alfred Hospital, Camperdown 2050, Australia

*To whom correspondence should be addressed:

Dr. Ulf Schmitz; email: u.schmitz@centenary.org.au or Prof. John Rasko, email: j.rasko@centenary.org.au

†These authors should be regarded as joint last authors.

Summary

Dynamic intron retention (IR) in vertebrate cells is of widespread biological importance. Aberrant IR is associated with numerous human diseases including cancer. Despite consistent reports demonstrating intrinsic sequence features that predispose introns to become retained, conflicting findings about cell type-specific IR regulation demand a systematic analysis in a controlled experimental setting. We integrated matched transcriptomics and epigenetics data (including DNA methylation, nucleosome occupancy, histone modifications) from primary human myeloid and lymphoid cells. Using machine learning we trained two complementary models to determine the role of epigenetic factors in the regulation of IR. We show that increased chromatin accessibility contributes substantially to the retention of introns in a cell-specific manner. We also confirm that intrinsic characteristics of introns are key for them to evade splicing. With mounting reports linking pathogenic alterations to RNA processing, our findings may have profound implications for the design of therapeutic approaches targeting aberrant splicing.

Keywords: chromatin accessibility, intron retention, epigenetics, alternative splicing, histone marks, CpG methylation, nucleosome occupancy

Introduction

The role of introns in mammalian genomes remains largely unexplained. Given the time and energy required for the transcription and subsequent excision of introns from pre-mRNA, it was important to recognise in recent years that introns can be selectively retained in mature mRNA transcripts and thereby contribute significantly to transcriptomic complexity (Schmitz et al., 2017; Wong et al., 2013). Intron retention (IR) is a form of alternative splicing that was assumed to occur due to the failure of the spliceosome to excise an intron from a pre-mRNA transcript. However, growing evidence suggests that IR is highly regulated by multiple complementary factors (Monteuuis et al., 2019).

IR is widespread across all human tissues and affects more than 80% of protein-coding genes (Middleton et al., 2017). For example, dynamic IR profiles have been identified in key genes involved in hematopoietic cell differentiation and activation (Edwards et al., 2016; Green et al., 2020; Ni et al., 2016; Ullrich and Guigo, 2020; Wong et al., 2013). Fates of intron-retaining transcripts can be diverse and include (i) nonsense-mediated decay triggered by intronic premature termination codons, (ii) detention in the nucleus or nuclear degradation, and (iii) translation into alternative protein isoforms or creation of neoepitopes (Monteuuis et al., 2019; Smart et al., 2018; Wong et al., 2016). A better understanding of how IR is regulated is crucial to determine factors leading to aberrant IR, which has

been associated with multiple diseases including cancer (Dvinge et al., 2019; Hershberger et al., 2020; Monteuuis et al., 2020)

Despite numerous studies that describe the role of retained introns in key biological functions in animals and in human diseases (Monteuuis et al., 2020; Monteuuis et al., 2019; Wong et al., 2016), a comprehensive understanding of their regulation is still lacking. Retained introns have conserved intrinsic characteristics such as a higher GC content, shorter lengths, and weaker splice sites in comparison to their non-retained counterparts (Braunschweig et al., 2014; Monteuuis et al., 2019; Schmitz et al., 2017). These features predispose introns to retention but cannot explain the dynamic IR profiles observed in numerous biological processes.

The regulation of alternative splicing has been the focus of many studies. Evidence suggests that alternative splicing is regulated at least at two levels: (i) locally, where *trans*-acting splicing regulators interact with *cis*-acting regulatory elements, and (ii) globally, through the structure of chromatin, which is largely governed by epigenetic factors, including nucleosome assembly, histone modifications and CpG methylation (Zhou et al., 2014).

Previous reports have shown that, apart from intrinsic sequence-based features, intron expression can be regulated through (i) *cis*-regulatory elements, such as sequence motifs attracting *trans*-acting splicing-regulatory RNA binding proteins (Middleton et al., 2017), (ii) core components of the splicing machinery (Wong et al., 2013), and (iii) change in the RNA Pol II elongation rate (Fong et al., 2014). Moreover, an increasing number of studies have found links between epigenetic profiles and IR; reporting that IR is associated with reduced CpG methylation (Gao et al., 2019; Green et al., 2020; Kim et al., 2018; Wong et al., 2017a) and various histone modifications (Guo et al., 2014; Wei et al., 2018). However, these reports have typically established the association of IR with only one epigenetic factor at a time. In general, the question of whether there are dominant epigenetic factors that underpin IR regulation remain unanswered.

In the quest to find a splicing regulatory 'code', several studies have used machine learning methods to train models that predict exon usage with increasing precision (Barash et al., 2010; Leung et al., 2014). Moreover, some models were developed to predict cryptic splicing events caused by genetic variations and to link these to human diseases (Baeza-Centurion et al., 2019; Jaganathan et al., 2019; Xiong et al., 2015). However, the computational prediction of IR events has not been attempted to date and the role of epigenetic marks has rarely been considered in computational models of splicing regulation (Monteuuis et al., 2019; Pacini and Koziol, 2018).

In this study, to sought to systematically elucidate the role of epigenetic marks in the regulation of IR. We analysed genome-wide profiles of 6 histone modifications, CpG methylation and nucleosome occupancy at single-base resolution in primary lymphoid and myeloid cells. Using machine learning, we developed two models that predict IR in primary human immune cells. More specifically, we trained

89 a logistic regression with elastic net (EN) classifier and a conditional Random Forest (RF) classifier
90 with matched transcriptomics and epigenomics data from monocytes, macrophages, naïve T-cells, T-
91 central memory, and T-effector memory cells (Figure 1).

92 Our results show that intrinsic characteristics are key for introns to evade splicing and that epigenetic
93 marks may modulate IR levels in a cell type-specific manner, where the dominant factor for dynamic
94 IR regulation is chromatin organisation.

95 Results

96 Intrinsic features of retained introns are consistent across cell types

97 To investigate how IR is regulated in primary immune cells (CD4+ T-cells, monocytes, and
98 macrophages), we integrated transcriptomics (mRNA-Seq) data with epigenomics data including
99 genome-wide CpG methylation (WGBS), histone modifications (ChIP-Seq), and nucleosome
100 occupancy (NOMe-Seq) (Table S1). The cells were isolated from peripheral blood of 2 healthy donors,
101 except for the monocyte-derived macrophages. Using the IR identification software IRFinder
102 (Middleton et al., 2017), we quantified IR events of expressed genes (FPKM>1) in five cell types across
103 myeloid and lymphoid cells, representing two modes of differentiation: monocyte-to-macrophage
104 differentiation and naïve T-cell differentiation into central memory (CM) and effector memory (EM)
105 T-cells. Introns that were present in at least 10% of a gene's mature mRNA transcripts ($IR_{ratio} \geq 0.1$)
106 with an overall intron depth ≥ 10 were considered retained. Non-retained introns were defined as those
107 with an $IR_{ratio} \leq 0.01$ and intron depth < 10.

108 We identified a total of 26,147 retained introns in 12,379 genes, some of which were retained in both
109 myeloid and lymphoid cells while others were cell type-specific (Figure S1A). Consistent with previous
110 reports, retained introns in our dataset are shorter in length, exhibit a higher GC content and weaker
111 splice site strengths compared to non-retained introns (Figure S1B-E). Our analysis revealed diverse
112 splicing patterns in myeloid and lymphoid cells. While 40% of the retained introns in myeloid cells
113 were significantly differentially retained ($\Delta IR \geq 0.1$; $p < 0.05$ Audic-Claverie test) between monocytes
114 and macrophages (571/1425), T cells displayed greater stability in regard to IR with only 8% of introns
115 classified as differentially retained (146/1812 in naïve T vs CM, and 80/969 in CM vs EM). In contrast
116 to the monocyte-to-macrophage differentiation, where we observed a reduction in IR events (Figure
117 2A), the overall number of retained introns remained consistent in all CD4+ T cells. These patterns
118 coincide with fewer changes in gene expression during T cell differentiation in contrast to major gene
119 expression changes in monocyte-to-macrophage differentiation (Figure S1F).

120 Most retained introns in our analysis overlapped with histone marks (HM) or with a nucleosome free
121 region (NFR, predicted from NOMe-seq data) around their 5′ and 3′ splice sites (+/100 bp) as well as

122 the middle of an intron (Figure S2A). Interestingly, many non-retained introns (~50%) lacked such
123 epigenetic marks in lymphoid cells (as opposed to only 20-30% of retained introns). H3K36me3 was
124 the most frequently observed histone modification followed by NFR peaks. In retained introns, between
125 30% and 60% of H3K36me3 signals were classified as strong (see Methods), whilst in non-retained
126 introns the proportion of overlap with the regions of strong signal ranged between 2% and 18%. Again,
127 the patterns of signal strength varied between the cell types (Table S3).

128 CpG methylation profiles (extracted from WGBS data) for retained and non-retained introns displayed
129 a characteristic bimodal distribution with two distinct peaks at 0% and 100%. Differential methylation
130 was predominantly found at the splice sites when we compared regions of genomic DNA associated
131 with IR and no IR. At the 5′ splice sites, we observed higher methylation levels in retained compared
132 to non-retained introns in all five cell types. However, this trend was reversed in the lymphoid cells at
133 the 3′ splice sites and in the middle of introns (Figure S2B).

134 M.CviPI enzyme, used in NOMe-seq experiment, methylates cytosine dyads in GC sequence and GCH
135 methylation levels (where H is any nucleobase except guanine) provide information about chromatin
136 accessibility. Unlike endogenous CpG methylation, GC dinucleotides are rarely fully methylated,
137 therefore the mid-range levels (anywhere between 20 to 50%) are usually sufficient to indicate open
138 chromatin regions. In our data, chromatin accessibility (i.e. GCH methylation) increased from
139 monocytes to macrophages with slightly higher levels in retained introns, while lymphoid cells had
140 increased chromatin accessibility (GCH methylation levels 15-35%) but with lower levels in retained
141 introns compared to non-retained introns (Figure S2C).

142 To determine important factors of IR regulation, we compiled sequence-based and epigenetic features:
143 (i) sequence-based features: intron length, GC content, splice site strength, CpG density (also referred
144 to as intrinsic features), (ii) transcriptomics features: percent spliced-in (PSI) values of the flanking
145 exons, and (iii) epigenomics features extracted from the WGBS, ChIP-Seq (H3K9me3, H3K27me3,
146 H3K27ac, H3K36me3, H3K4me1, H3K4me3), and NOMe-Seq data (Table S2). We then used these
147 features (n=46) to train EN models for each cell type and predict whether introns are either retained or
148 non-retained. The performance of our models was assessed based on the area under the receiver
149 operating characteristic curve (AUC) values, which ranged between 0.87 and 0.95 (Figure 2B) and
150 values for the area under the prediction-recall curve (accuracy) ranging between 0.85 and 0.95 (Figure
151 2C). The consistently high values suggest that the model choice was appropriate for the task.

152 Next, in order to evaluate whether the learned relationship between the model features and IR was
153 generalizable across cell types we trained our model with data from one cell type and tested it with data
154 from another cell type. For all training/test data pairs, the AUC and accuracy metrics were comparable
155 to those models that were trained and tested on the same cell type (Table S4).

156 The EN model assumes a monotonic linear relationship between the class variable and the model
157 features. To determine whether this assumption is adequate for IR classification, we also trained
158 conditional random forest (cRF) models, which do not make any prior assumption about the relationship
159 between the outcome of interest and the model features. Comparing the results from both types of
160 models, we found that cRF performed slightly better than EN with AUC values ranging between 0.91
161 and 0.98 (Figures 2D, S3A) and PR values between 0.87 and 0.95 (Figure S3B).

162 To assess which features contribute most to the model performance (and thus, the relevance of a feature
163 to IR), we used variable-importance measures (VIM). For EN, these are the regression coefficients
164 ordered from lowest to highest, where parameters with larger values have a greater effect. For cRF,
165 variable importance was calculated as the mean decrease in accuracy after permutation of each model
166 feature (Figure 2E). Given the known properties of retained introns it was no surprise that intrinsic
167 features, such as length, GC content and CpG density were ranked as the top predictors with a high
168 level of agreement across all cell types analysed. Again, we observed consistency between the EN and
169 cRF models, except for minor variations in the order that important features were ranked in.

170 Epigenetic features were also ranked among the top 5 predictors across all models and cell types,
171 however their nature and relative importance varied between cell types (Figure 2E). Overall EN models
172 ranked epigenetic features as moderately to very important (VIM between 0.4 and 0.8), which is
173 comparable to the intrinsic features (ranging between 0.3 and 1). In contrast, cRF identified epigenetic
174 features as somewhat important with VIM mostly below 0.50 (Figures S3C, S3D). Nevertheless,
175 intrinsic features were consistently identified as most relevant for correctly classifying IR, suggesting
176 that these features predispose introns to being retained irrespective of cell or tissue type.

177

178 Chromatin accessibility is predicted to be the strongest regulator of IR

179 In the previous section we classified IR on a cell type-specific basis and determined the intrinsic features
180 as having the strongest association with IR outcomes. However, we often find that an intron is retained
181 in one cell type but not in another. In those cases, factors beyond intrinsic features are the likely drivers
182 of this transition.

183 To find these IR determinants, we modified our initial modelling approach by focusing only on the
184 dynamic introns - those that changed their retention status between cell types (Figure 3A). In total,
185 1,540 introns matched this criterion with various IR patterns (Figure 3B). We used these introns to train
186 EN and cRF models with both epigenetic and intrinsic features. The cRF model was performed superior
187 to the EN model achieving AUCs of 0.85 and 0.76, respectively (Figure 3C). cRF also achieved a higher
188 area under the precision-recall curve value (0.83) than EN (0.73) (Figure 3D). The poorer performance
189 of EN might be a reflection of the model's inability to fully utilise complex structures within the omics

190 data, thus supporting the notion that a relationship between chromatin modifiers and IR is indeed non-
191 linear, as previously suggested (Singer et al., 2015).

192 Evaluation of feature rankings revealed that, despite varying model performances, both EN and cRF
193 models identified features related to chromatin accessibility as most important for correct IR
194 classification (Figure 3E). These features include GCH methylation and GCH (i.e. nucleosome)
195 occupancy and the presence of nucleosome free regions (NFRs). GCH methylation at the 5′ and 3′ splice
196 sites were determined as most important features discriminating retained form non-retained introns in
197 both models. The cRF classifier also identified CpG methylation as somewhat important for IR
198 classification, which has a known relationship with chromatin accessibility (Farlik et al., 2016; Lay et
199 al., 2015; Taberlay et al., 2014). Interestingly, the cRF model also identified GC content as a moderately
200 important contributor to IR outcomes, whilst the EN model included histone marks (H3K27ac and
201 H3K36me3) in their top 10 predictors (Figure S4A).

202

203 Epigenetic IR regulation is independent of gene expression regulation

204 It is reasonable to assume that changes in the epigenomic landscape might not directly affect IR but
205 rather gene expression (Jaenisch and Bird, 2003). To confirm that the features identified as relevant to
206 IR are independent from gene expression regulation, we split dynamically retained introns into three
207 groups: (i) host gene expression is reduced along with the change in IR status, (ii) host gene expression
208 remained stable ($\log_2$ FC FPKM $\leq$ 2), and (iii) host gene expression increased (Figure 4A). For most
209 of the dynamic introns the host gene expression remained unchanged (N = 1,220), whilst down- and
210 upregulated host genes were associated with 73 and 247 alternately retained introns, respectively. We
211 repeated the classification analysis on the group of introns where the IR changes were not accompanied
212 by host gene expression changes. Since the relationship between IR and epigenetic model features is
213 not linear, as was established in the previous section, we only used the cRF algorithm.

214 The model fitted on this data subset achieved an AUC of 0.83 (Figure 4B) and an area under the
215 precision-recall curve value of 0.78 (Figure S4B). The features that were selected as important were
216 GCH methylation at the 5′ and 3′ splice sites and GC content in the same order as in the model trained
217 on all dynamically retained introns (Figure 4C). This observation held true for both highly and lowly
218 expressed host genes (Figures S4C). We therefore concluded that the observed epigenetic changes
219 associated with IR modulation are independent from gene expression regulation. In Figure 4D, we show
220 two exemplary introns where greater chromatin accessibility was associated with an increase in IR:
221 Phosphatidylinositol Glycan Anchor Biosynthesis Class T (PIGT) helps building the
222 glycosylphosphatidylinositol-anchor which is found on the surface of various blood cells (Figure 4D,
223 left). PIGT is known to express many isoforms through alternative splicing including IR. The nucleotide

224 binding protein SEPTIN8 is a regulator of cytoskeletal organization, which has multiple alternatively
225 spliced transcript variants as well (Figure 4D, right).

226

227 Dynamic changes in chromatin structure are responsible for cell type-
228 specific IR

229 As chromatin accessibility was identified as the strongest predictive factor for differential IR, we closely
230 examined its relationship with retained and non-retained introns. We identified 5 distinct GCH
231 methylation profiles in the +/- 200 bp region around the 5′ splice site of retained introns (Figure 5A,
232 left). Similar clustering profiles were identified in the region around 3′ splice sites and the middle of
233 introns (Figure S5). To understand changes in chromatin status in the context of differential IR, we
234 plotted the GCH methylation values of the same introns when they were not retained (Figure 5A). The
235 associated heatmap shows that GCH methylation is widely depleted in non-retained introns, with no
236 distinct clustering. In retained intron, however, we observed a clear increase in GCH methylation
237 immediately upstream or downstream from the 5′ splice site (Figure 5B, clusters 1, 3 and 4). We also
238 identified a group of retained introns with relatively low levels of GCH methylation (cluster 2) and
239 another with particularly strong GCH methylation (cluster 5).

240 Upon visualising the intronic regions that changed their IR status between cell types, we observed
241 greater chromatin accessibility levels in retained introns (Figure 6A). Moreover, for the majority of
242 introns, we found that IR gain was accompanied with a reduction in H3K36me3 signal (Figure 6A).

243 Based on the observed patterns, we hypothesise that there is an association between chromatin dynamics
244 and IR: chromatin is more likely to be in a permissive state (high GCH methylation) in the vicinity of
245 retained introns and more compact (low GCH methylation) around constitutively spliced introns.
246 Indeed, we observed that chromatin becomes more accessible as introns become retained (65% of
247 observations). In other cases, the IR status changes without any change to the chromatin state (35% of
248 observations).

249 Based on the observations concerning chromatin accessibility, we sought to assess the relationship
250 between IR and epigenetic factors in the context of changing chromatin states, i.e. differential GCH
251 methylation (Figure 6B), and stable chromatin status, i.e. non-differential GCH methylation (Figure
252 6C). In our analysis, we separated first introns from other introns to detach epigenetic signals associated
253 with gene promoters.

254 The patterns of CpG methylation, H3K27ac, H3K4me3 and H3K4me1 levels in retained and non-
255 retained introns were similar in both chromatin modes (dynamic and stable). First non-retained introns
256 displayed enrichment for histone marks and reduced CpG methylation levels, while first retained introns

257 had negligible levels of histone marks and were marked by the absence of CpG methylation (Figure 6B
258 and 6C, top rows). In contrast, the above-mentioned histone marks were silenced in the internal introns
259 irrespective of the IR status, while the H3K36me3 signal increased. Interestingly, H3K36me3 levels
260 were reduced in retained introns associated with dynamic chromatin (Figure 6B, 2nd row, far right),
261 while they remained similar in retained- and non-retained introns associated with stable chromatin
262 (Figure 6C, 2nd row, far right).

263 A most interesting result of this analysis was that there are no differences in epigenetic marks between
264 internal retained and non-retained introns when a stable chromatin state is maintained (Figure 6C,
265 bottom row). This suggests that there must be unknown factors that are independent of chromatin
266 accessibility responsible for modulating IR. Thus, further investigations are required to identify
267 additional factors that impact on IR in haematopoietic cells.

268 Discussion

269 In this study, we have employed a machine-learning approach to determine regulators of IR in primary
270 hematopoietic cells. For the first time we provide integrated matched transcriptomic, nucleosome
271 occupancy, CpG methylation, and 6 histone modification profiles from 5 primary human cell types
272 representing 2 independent systems of haematopoietic cell differentiation. Previous studies have
273 described features that are associated with retained introns, including a higher intronic GC content,
274 shorter intron lengths, weaker 5′ and 3′ splice site strengths, and some epigenetic marks (Braunschweig
275 et al., 2014; Schmitz et al., 2017; Wong et al., 2017a). However, these studies have focused on single
276 or paired omics layers only and often used individual cell lines for their analyses.

277 We applied supervised machine learning using EN and conditional RF algorithms. Unlike deep learning
278 methods, that are very capable of identifying complex relationship patterns but do not provide tools to
279 determine how exactly an outcome was determined (Rauschert et al., 2020), these multivariate models
280 allows the identification of features that contribute most to the outcome of interest (IR). Such modelling
281 strategy is "data-independent" and can be applied to other forms of alternative splicing as well. For
282 example, RF has been used to study the importance of chromatin modifications in the interaction
283 between topologically associated domains (Dixon et al., 2015) and EN was used to model prognostic
284 alternative splicing signatures in breast cancer (Wang et al., 2020).

285 Previous studies have mostly focussed on investigating the functional links between chromatin
286 organisation and gene expression regulation and found that nucleosome free regions at a transcription
287 start site are strongly associated with transcription initiation (Radman-Livaja and Rando, 2010).
288 Nucleosomes were also reported to be preferentially positioned in exons to facilitate their identification
289 among flanking introns by the splicing machinery (Schwartz et al., 2009; Tilgner et al., 2009). However,
290 it is important to note that these findings were made using the micrococcal nuclease digestion with deep

291 sequencing (MNase-seq) protocol, which is more susceptible to GC content bias. Kelly et al. (Kelly et
292 al., 2012) showed that nucleosome enrichment in exons vs. introns was not observed in NOMe-seq data,
293 which they attributed to the technical differences between the two experimental approaches. NOMe-
294 seq data includes the percentage of methylated reads at a given position as opposed to the count of
295 mapped reads in MNase-seq data. Similarly, our NOMe-seq based analysis of chromatin accessibility,
296 quantified by GCH methylation, did not reveal a specific preference for nucleosomes to be positioned
297 in exons rather than introns.

298 Our study did reveal the regions of clear GCH enrichment clusters either upstream, downstream or
299 directly at the splice sites of retained introns in contrast to non-retained introns. High GCH methylation
300 levels, like those observed in retained introns, are indicative of nucleosome free regions or NFRs,
301 regions of possible nucleosome eviction that are characterised by a high density of methylated GCH
302 sites and unmethylated CpG dinucleotides (Nordström et al., 2019). Interestingly, You et al. showed
303 that a loss of nucleosome depleted regions accompanied by nucleosome occupancy precedes changes
304 in endogenous CpG methylation in OCT4 and NANOG genes in embryonic carcinoma cell line NCCIT
305 (You et al., 2011). Formation of an NFR upstream from the 5′ exon/intron boundary led to DNA
306 hypomethylation and the depletion of H3K36me3 in SETD2 deficient tumours (Simon et al., 2014). It
307 is therefore reasonable to conclude that alteration of the epigenetic landscape attributed to IR initially
308 starts with changes in nucleosome architecture and subsequent transcriptome rewiring.

309 Apart from signalling a nucleosome eviction, high levels of GCH methylation potentially mark regions
310 with longer internucleosomal spacing, also known as DNA linker regions. A study in estimating
311 nucleosome phasing in single cell found great agreement between average linker length measured with
312 scNOMe-seq data and the phase estimates derived from MNase-seq (Pott, 2017). Linker length ranges
313 between ~20–90 bp and varies among different species, tissues, and even fluctuates within a single
314 cellular genome (Szerlong and Hansen, 2011). Nucleosome phasing has been linked to alternative
315 splicing before, where RNA Pol II elongation rates increase upon histone depletion and pre-mRNA
316 splicing is delayed (Jimeno-González et al., 2015). Previous studies identified nucleosomes as physical
317 barriers to efficient transcription elongation *in vitro*, however *in vivo* they are efficiently removed from
318 transcribed chromatin (Saldi et al., 2016). Pol II were also found to be involved in maintaining
319 nucleosome phasing in the transcribed region, where longer Pol II dwell times, associated with slow
320 transcription, allowed for remodelling of H3K36me3 profiles (Fong et al., 2017).

321 In regions further downstream of transcription start sites, nucleosome positioning becomes less stable
322 (Radman-Livaja and Rando, 2010) and linker region lengths become nonuniform. We therefore propose
323 that the differences in DNA methylation and H3K36me3 signal observed over internal introns reflect
324 the underlying changes in nucleosome organisation, that in turn propagates IR (Figure 7). In the
325 presence of IR, transcription rates are faster over more spaced out nucleosomes that does not allow

326 sufficient time for a "writer" to deposit H3K36me3 in the splicing region (Fong et al., 2017). CpG sites
327 in the DNA linker regions are usually unmethylated (Pott, 2017) and therefore may explain the reduced
328 DNA methylated levels associated with IR (Wong et al., 2017b).

329 In the proximity of transcription start sites, strong histone modification levels (like we observed for
330 H3K4me3 and H3K27ac) indicate a well-positioned nucleosome (Andersson et al., 2009), while
331 reduced histone modification levels, particularly reduced H3K4me3, are associated with transcription
332 factor (TF) binding (Wu et al., 2015). TF binding sites can undergo nucleosome remodelling (Ballaré
333 et al., 2013) in the form of nucleosome shifts or nucleosome eviction and the formation of an NFR with
334 associated changes to RNA polymerase II elongation rates. We propose that IR in first introns might be
335 a biproduct of functional histone modifications and nucleosome remodelling for the purpose of TF
336 recruitment in the regions proximal to transcription start sites.

337 In summary, our results provide a major conceptual advance in our understanding of alternative splicing
338 regulation. We found an unanticipated strong contribution of chromatin organisation in IR modulation
339 where nucleosomes position upstream or downstream of retained introns (determined by the length of
340 linker regions and NFRs) to facilitate an acceleration of RNA Pol II elongation and increased IR.
341 Furthermore, the models generated in this study can be adapted to study epigenetic gene expression and
342 alternative splicing regulation in other cell systems, other species, in health or disease, and further our
343 understanding of these essential biological mechanisms.

344

345 Acknowledgements

346 We thank Benedikt Brors and Roland Eils from DKFZ Heidelberg and Alf Hamann from DRFZ Berlin,
347 Wie Chen, Nikolaus Rajewsky and Sascha Sauer from MDC Berlin, Ho-Ryun Chung and Martin
348 Vingron from MPI-MG Berlin, Thomas Jenuwein, Thomas Manke and Andrew Pospisilik from MPI-
349 IE Freiburg, Philip Rosenstiel and Stefan Schreiber from CAU Kiel, Jan G. Hengstler from IfADo
350 Dortmund, Thomas Lengauer from MPI-INF Saarbrücken, Bernhard Horsthemke from Universität
351 Duisburg-Essen, Alexandra Kiemer from Universität des Saarlandes Saarbrücken, Thomas Pap from
352 WWU Münster and Gerd Schmitz from Universität Regensburg who were involved in the work with
353 biological samples, sequencing and generation of WGBS, NOMe-Seq, ChIP-Seq and RNA-Seq data
354 for the DEEP Consortium.

355 This work was supported by the National Health and Medical Research Council (Investigator Grant
356 #1177305 to J.E.J.R., Project #1080530 to J.E.J.R., Project #1128175 and #1129901 to J.E.J.R. and J.J.-
357 L.W., #1126306 to J.J.-L.W.; the NSW Genomics Collaborative Grant (J.E.J.R. and J.J.-L.W.); Cure
358 the Future (J.E.J.R.), and an anonymous foundation (J.E.J.R.). U.S. and J.J.-L.W. hold Fellowships

from the Cancer Institute of New South Wales. U.S. also received support from the Australian Academy of Science in form of an Australia-India Early and Mid-Career Fellowship. This research was funded by the Cancer Council NSW Project Grants (RG11-11 and RG20-12) to J.E.J.R. and U.S. K.V.J.N. and J.W. were supported by the German Epigenome Program (DEEP) funded by the Ministry of Education and Research in Germany (BMBF 01KU1216).

The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney.

Author Contributions

J.E.J.R., J.J.-L.W. and U.S. designed the study and supervised the project, V.P. and R.S. performed bioinformatic analyses, V.P. performed statistical analysis and data modelling, N.J.A. advised on statistical methodology, DEEP Consortium provided sequencing data, J.W. designed and coordinated sequencing experiments, K.J.V.N. data management, V.P. and U.S. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Declaration of interests

J.E.J.R. has received honoraria or speakers' fees (GSK, Miltenyi, Takeda, Gilead, Pfizer, Spark, Novartis, Celgene, bluebird bio); Director of Pathology (Genea); equity ownership (Genea, Rarecyte); consultant (Rarecyte, Imago); chair, Gene Technology Technical Advisory, OGTR, Australian Government. K.J.V.N. is currently employed by AstraZeneca. The remaining authors declare no competing financial interests.

Figure legends

Figure 1 Experimental design and workflow to determine regulators of IR. Raw high-throughput data were processed for each biological replicate and amalgamated by cell type from the indicated number of samples (n). The output was used for feature extraction: IR events were treated as a binary outcome and we trained an Elastic Net (EN) regression model and a conditional Random Forest model with a total of 46 sequence-based and epigenetic features. Using feature ranking, we identified the factors that were most strongly associated with IR outcomes and compared the performances of both modelling strategies. These steps were repeated for each cell type.

Figure 2 IR prediction and model feature association analyses. (A) Scatter plot of differential IR events (Sig blue – significant; Not Sig yellow – not significant) between monocytes (Mo) vs macrophages (Ma) (left), Naïve (TN) vs Central Memory (CM) T cells (middle), and Central Memory vs Effector Memory (EM) T cells (right). **(B)** Receiver operating characteristic (ROC) curves and **(C)** precision recall (PR) curves comparing the performance of the EN classifier in five cell types. **(D)** Comparison of AUC values between EN and cRF algorithms, error bars show 95% confidence interval. **(E)** Variable importance scores for the top 10 features identified by EN and conditional RF algorithms. The scores were scaled to values that add up to 1.0 and the size of a bar corresponds to the effect size.

Figure 3. Analysis of dynamics intron retention. (A) Modified modelling strategy from Figure 1. Only introns that were found to be in retained and non-retained states in different cell types were included in the analysis. **(B)** Alluvial plot illustrating the dynamics of IR states among the five cell types. **(C)** ROC and **(D)** PR curves comparing the performance of cRF (brown) and EN (black). **(E)** Variable importance scores for the top 5 features identified by EN and conditional RF algorithms, scaled between 0 and 1.

Figure 4 Analysis of introns from genes with non-differential expression levels. (A) Scatter plot of host gene expression for introns that change their IR status. **(B)** ROC curve indicating the performance of a conditional RF model fitted on the data from non-differentially expressed genes (GE, gene expression). **(C)** Ranking of the features based on the scaled variable importance scores. **(D)** Integrative Genomics Viewer (IGV) plots revealing higher density and hypermethylation levels of GCH sites in the splice site regions of differentially retained introns in both highly- and lowly- expressed gene examples (NFR – Nucleosome Free Region, GCH Methylation – methylation levels of GC dinucleotides followed by any nucleobase except guanine).

Figure 5 GCH methylation clustering in differentially retained introns. **(A)** Clustering of GCH methylation in the +/- 200 bp region around the 5′ splice site (ss). Each line corresponds to one intron that is either in a retained (left) or non-retained state (right). **(B)** Line plots showing average GCH methylation values (i.e. chromatin accessibility) in retained vs non-retained introns across 5 clusters.

415 **Figure 6 Interplay between chromatin accessibility, CpG methylation and histone modifications.**
416 **(A)** IGV plots of mRNA-seq, H3K36me3 ChIP-seq, NOMe-seq, and WGBS-seq data indicating
417 different levels of GCH methylation between retained and non-retained introns and higher prevalence
418 of NFRs in the regions proximal to IR. **(B)** Line graphs show the average levels of GCH methylation,
419 CpG methylation, and the difference between ChIP-seq H3K4me3, H3K27ac, H3K4me1, and
420 H3K36me3 signals and ChIP-Seq Input, normalised to the Bins Per Million (BPM), in retained (red)
421 and non-retained (blue) introns associated with chromatin status. The first row shows epigenetic signals
422 at the 5′ splice site of first introns (close to the promoter region) and the second row represents all other
423 introns. **(C)** The same analysis performed in (B) is repeated for introns where the chromatin status
424 remains the same, i.e. non-differential GCH methylation.

425 **Figure 7 Proposed role of chromatin accessibility in IR regulation.** More dense positioning of
426 nucleosomes slows down RNA Pol II elongation rate, allowing sufficient time for a histone
427 modification (in this case, H3K36me3). Methylated CpG dinucleotides and unmethylated GCH sites
428 over the nucleosome core explain higher CpG methylation levels and lower GCH methylation levels in
429 constitutively spliced introns.

430

431

432

STAR Methods

Quantification and statistical analysis

To investigate how IR is regulated in primary immune cells, we integrated epigenomics and transcriptomics data from the German Epigenome Program (DEEP). Primary monocytes, monocyte-derived macrophages, and primary T-cells (naïve, central memory, effector memory) were retrieved from 2 healthy donors. Cell isolation, differentiation, DNA/RNA extraction and library preparation for mRNA-Seq, WGBS, NOMe- and ChIP-Seq experiments are described in detail in these articles (Durek et al., 2016; Wallner et al., 2016).

mRNA-Seq data processing and identification of IR events

RNA-Seq reads (FASTQ format) of each technical replicate were tested for quality using FastQC v.0.11.5 (github.com/s-andrews/FastQC). Further processing, including adaptor trimming, was performed within the IRFinder algorithm for IR quantification (Middleton et al., 2017). Sequencing reads were mapped to the human reference genome (GRCh38) using STAR v2.7 with default parameters (Dobin et al., 2013). IR-ratios, a quantitative measure of IR levels, were determined as:

$$IRratio = \frac{Intronic\ Abundance}{Intronic\ Abundance + Exonic\ Abundance},$$

where the Intronic Abundance is defined as the trimmed mean of the reads that map to an intron, after having excluded features that overlap the intron, with the highest and lowest 30% of values being excluded. Exonic Abundance is defined as the number of reads that map across an exon-exon junction. Library size normalisation (between-sample normalisation) was not required as the ratio between intronic and exonic abundance is determined from within the same transcriptome (Middleton et al., 2017).

Introns that were present in at least 10% of a gene's mature mRNA transcripts ($IR_{ratio} \geq 0.1$) with an overall intron depth ≥ 10 were considered retained. Non-retained introns were defined as those with an $IR_{ratio} \leq 0.01$ and intron depth < 10.

We used Cufflinks v2.1.1 (Trapnell et al., 2010) to estimate gene abundance in fragments per kilobase per million (FPKM). Only introns from host genes with FPKM ≥ 1 were selected for the downstream analyses.

WGBS data processing

Raw WGBS FASTQ files were assessed for quality using FastQC v.0.11.5 (github.com/s-andrews/FastQC). Standard Illumina adaptors used for the library preparation were trimmed using cutadapt v.1.10 (Martin, 2011) with a quality cutoff of 20 base pairs (bp) and minimum read length of

464 30 bp. Trimmed reads were mapped to the GRCh38 reference genome, duplicate reads removed, and
465 methylation calling performed using Bismark v.0.19.0 (Krueger and Andrews, 2011). Only CpG sites
466 with a coverage of more than 5 reads were retained for further analysis.

467 **ChIP-Seq data processing**

468 ChIP-Seq data for six histone modifications (H2K27ac, H3K27me3, H3K36me3, H3K4me1,
469 H3K4me3, H3K9me3) were aligned to the human GRCh38 reference genome using STAR v2.7 (Dobin
470 et al., 2013). Duplicate reads were removed using Picard v.2.18.4 (broadinstitute.github.io/picard/) and
471 further processed using MACS2 v.2.2.6 (Zhang et al., 2008) to identify histone modification peaks,
472 with default parameters and q-value cut-off of 0.01. All histone modifications were processed in the
473 "narrow peak" mode in order to extract peak summit coordinates. For visualisation in IGV (Robinson
474 et al., 2012), we generated coverage tracks using `bamCoverage` from deepTools2 (Ramirez et al.,
475 2016) with the following parameters `--binSize 1 --normalizeUsing BPM --`
476 `effectiveGenomeSize 2913022398 --extendReads 200`. For HM line plots, we
477 substracted ChiP-Seq Input from a respective HM ChiP-seq read counts and normalised based on Bins
478 Per Million (BPM) mapped reads using bamCompare and parameters `--binSize 1 --`
479 `scaleFactorsMethod readCount --effectiveGenomeSize 2913022398 --`
480 `operation subtract --normalizeUsing BPM`.

481 **NOMe-Seq data processing**

482 Raw FASTQ files were assessed for quality using FastQC v.0.11.5 (github.com/s-andrews/FastQC).
483 Reads were mapped to the GRCh38 reference genome, duplicate reads removed, and methylation
484 calling performed using Bismark v.0.19.0 (Krueger and Andrews, 2011). GCH methylation information
485 was extracted with the `coverage2cytosine` utility with `--nome` parameter.

486 NFRs were predicted using gNOMePeaks tool (Nordström et al., 2019) with default parameters, which
487 include 4,000 bp up- and downstream from each peak for background signal calculation and the
488 maximum distance between GpC sites of 150 bp. We used the same algorithms to predict nucleosome
489 positioning by substituting GCH methylation, as required input, with GCH occupancy ($1 -$
490 $GCH\ methylation$) and reducing the background region to 1,000 bp up- and downstream from each
491 peak and the distance between GCH sites to 20bp.

492 **Feature selection**

493 Model features were associated with three genomic regions around retained and non-retained introns:
494 (i) +/- 100 bp from the 5´splice site, (ii) +/- 100 bp from the 3´splice site, and (iii) +/- 100bp from the
495 middle of an intron, each region being 200 bp long. GC content was extracted using bedtools v.2.26.0
496 (Quinlan and Hall, 2010) `nuc` command. For splice site strength calculations, we used MaxEntScan

497 (Yeo and Burge, 2004). CpG density values was obtained using Repitools (Statham et al., 2010). The
498 percent spliced in (PSI) index of flanking exons was calculated as described in (Schafer et al., 2015).
499 Exons with PSI $\geq$ 0.9 were considered as included.

500 To generate epigenetic features, we overlapped three regions of interest with the pre-processed
501 epigenetic data. NFR regions were defined as regions greater than 40bp in length with p-value $\leq$ 0.05
502 (Fisher test comparing CpG methylation in the NFR to the surrounding background). Presence or
503 absence of an NFR was dichotomised as "yes" – 1 and "no" – 0. Information about nucleosome location
504 was included into the model in the similar manner (nucleosomes were defined as regions greater than
505 140bp in length with p-value $\leq$ 0.05).

506 The relationship between histone modification and IR was included into the model through the presence
507 or absence of an overlap with a histone signal region. It was categorised as 0 – no overlap, 1 – overlap
508 with a region of HM signal, 2 – overlap with a region of strong signal (strong signal = mean (HM pile-
509 up) + sd (HM pile-up)). The full list of features is presented in Table S1.

510 Elastic Net and Conditional Random Forest Modelling

511 To identify features important for IR, we constructed a binary classification model using the EN
512 algorithm. We approached the problem in a naïve manner, i.e. we did not impose any prior assumptions
513 about the factors that might potentially play a role and therefore an equal penalty factor was applied to
514 all features. EN classification was performed in the *caret* R package (Kuhn, 2008) using *glmnet* method
515 (Friedman J, 2010) for a binary outcome. The group imbalance, due to the different number of retained
516 and non-retained introns identified suitable for modelling, was handled by down-sampling, using
517 *downSample* command. Parameter λ, determining the overall size of the regularization penalty, was
518 optimised by 10-fold cross validation procedure. Features were ranked based on the absolute values of
519 the model coefficients.

520 We repeated this *in-silico* analysis to validate our results using an independent machine learning
521 algorithm, cRF. In cRF, unlike standard RF where the first split variable is randomly selected, an
522 association test between the outcome and the model predictors is performed first. The ranked p-values
523 are then used to identify the covariate with the strongest association to the outcome, which is later used
524 for the first binary split at cutpoint c for a continuous covariate or at category C for a categorical
525 covariate. cRF classification was also performed in *caret* using *cforest* method as implemented in the
526 *party* R package (Strobl C, 2008). The cRF model provides an unbiased measure of variable importance,
527 which we used to rank the most important features for IR prediction.

528 To avoid overfitting, we ranked the features' importance using both EN and cRF techniques (Ding et
529 al., 2018). Moreover, our findings were validated across different blood cell lineages from different
530 humans.

Statistical Analysis

All statistical analyses were performed in R v.4.0. For the identification of differentially retained introns we used the Audic and Claverie Test (Audic and Claverie, 1997). P-values ≤ 0.05 were considered significant. Clustering was performed using unsupervised hierarchical clustering with complete linkage.

Data and Software Availability

Sequencing data are deposited at the European Genome-Phenome Archive under the accession numbers EGAS00001001595 and EGAS00001001624. Access is subject to an application process as per the EGA requirements. R scripts developed for this study are available at https://github.com/combiomed/IR_code. Processed sequencing data used to train the models was deposited at Mendeley Data: http://dx.doi.org/10.17632/b6crxbxbk2.1.

References

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome research *19*, 1732-1741.

Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome Res *7*, 986-995.

Baeza-Centurion, P., Minana, B., Schmiedel, J.M., Valcarcel, J., and Lehner, B. (2019). Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell *176*, 549-563 e523.

Ballaré, C., Castellano, G., Gaveglia, L., Althammer, S., González-Vallinas, J., Eyras, E., Le Dily, F., Zaurin, R., Soronellas, D., Vicent, Guillermo P.*, et al.* (2013). Nucleosome-Driven Transcription Factor Binding and Gene Regulation. Molecular Cell *49*, 67-79.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53-59.

Braunschweig, U., Barbosa-Morais, N.L., Pan, Q., Nachman, E.N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., and Blencowe, B.J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res *24*, 1774-1786.

Ding, M.Q., Chen, L., Cooper, G.F., Young, J.D., and Lu, X. (2018). Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. Mol Cancer Res *16*, 269-278.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W.*, et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Durek, P., Nordström, K., Gasparoni, G., Salhab, A., Kressler, C., de Almeida, M., Bassler, K., Ulas, T., Schmidt, F., Xiong, J.*, et al.* (2016). Epigenomic Profiling of Human CD4(+) T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. Immunity *45*, 1148-1161.

Dvinge, H., Guenthoer, J., Porter, P.L., and Bradley, R.K. (2019). RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. Genome Res *29*, 1591-1604.

Edwards, C.R., Ritchie, W., Wong, J.J., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J.E., and Blobel, G.A. (2016). A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood *127*, e24-e34.

575 Farlik, M., Halbritter, F., Muller, F., Choudry, F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A.,
576 Ciaurro, V., Mathur, A., *et al.* (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell
577 Differentiation. Cell Stem Cell *19*, 808-822.
578 Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.D., and Bentley, D.L.
579 (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev
580 *28*, 2663-2676.
581 Fong, N., Saldi, T., Sheridan, R.M., Cortazar, M.A., and Bentley, D.L. (2017). RNA Pol II Dynamics
582 Modulate Co-transcriptional Chromatin Modification, CTD Phosphorylation, and Transcriptional
583 Direction. Molecular cell *66*, 546-557.e543.
584 Friedman J, H.T., Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via
585 Coordinate Descent. Journal of Statistical Software *33*, 1-22.
586 Gao, D., Pinello, N., Nguyen, T.V., Thoeng, A., Nagarajah, R., Holst, J., Rasko, J.E., and Wong, J.J.
587 (2019). DNA methylation/hydroxymethylation regulate gene expression and alternative splicing during
588 terminal granulopoiesis. Epigenomics *11*, 95-109.
589 Green, I.D., Pinello, N., Song, R., Lee, Q., Halstead, J.M., Kwok, C.T., Wong, A.C.H., Nair, S.S., Clark,
590 S.J., Roediger, B., *et al.* (2020). Macrophage development and activation involve coordinated intron
591 retention in key inflammatory regulators. Nucleic Acids Res *48*, 6513-6529.
592 Guo, R., Zheng, L., Park, J.W., Lv, R., Chen, H., Jiao, F., Xu, W., Mu, S., Wen, H., Qiu, J., *et al.* (2014).
593 BS69/ZMYND11 reads and connects histone H3.3 lysine 36 trimethylation-decorated chromatin to
594 regulated pre-mRNA processing. Mol Cell *56*, 298-310.
595 Hershberger, C.E., Moyer, D.C., Adema, V., Kerr, C.M., Walter, W., Hutter, S., Meggendorfer, M.,
596 Baer, C., Kern, W., Nadarajah, N., *et al.* (2020). Complex landscape of alternative splicing in myeloid
597 neoplasms. Leukemia.
598 Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates
599 intrinsic and environmental signals. Nat Genet *33 Suppl*, 245-254.
600 Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I.,
601 Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., *et al.* (2019). Predicting Splicing from Primary
602 Sequence with Deep Learning. Cell *176*, 535-548 e524.
603 Jimeno-González, S., Payán-Bravo, L., Muñoz-Cabello, A.M., Guijo, M., Gutierrez, G., Prado, F., and
604 Reyes, J.C. (2015). Defective histone supply causes changes in RNA polymerase II elongation rate and
605 cotranscriptional pre-mRNA splicing. Proceedings of the National Academy of Sciences *112*, 14840-
606 14845.
607 Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., and Jones, P.A. (2012). Genome-wide
608 mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome
609 research *22*, 2497-2506.
610 Kim, D., Shivakumar, M., Han, S., Sinclair, M.S., Lee, Y.J., Zheng, Y., Olopade, O.I., Kim, D., and
611 Lee, Y. (2018). Population-dependent Intron Retention and DNA Methylation in Breast Cancer. Mol
612 Cancer Res *16*, 461-469.
613 Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-
614 Seq applications. Bioinformatics *27*, 1571-1572.
615 Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical
616 Software *28*, 1-26.
617 Lay, F.D., Liu, Y., Kelly, T.K., Witt, H., Farnham, P.J., Jones, P.A., and Berman, B.P. (2015). The role
618 of DNA methylation in directing the functional organization of the cancer epigenome. Genome Res *25*,
619 467-477.
620 Leung, M.K., Xiong, H.Y., Lee, L.J., and Frey, B.J. (2014). Deep learning of the tissue-regulated
621 splicing code. Bioinformatics *30*, i121-129.
622 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
623 EMBnet.journal *17*, 10-12.
624 Middleton, R., Gao, D., Thomas, A., Singh, B., Au, A., Wong, J.J., Bomane, A., Cosson, B., Eyras, E.,
625 Rasko, J.E., *et al.* (2017). IRFinder: assessing the impact of intron retention on mammalian gene
626 expression. Genome Biol *18*, 51.
627 Monteuuis, G., Schmitz, U., Petrova, V., Kearney, P.S., and Rasko, J.E.J. (2020). Holding on to junk
628 bonds: intron retention in cancer and therapy. Cancer Res.

629 Monteuuis, G., Wong, J.J.L., Bailey, C.G., Schmitz, U., and Rasko, J.E.J. (2019). The changing
630 paradigm of intron retention: regulation, ramifications and recipes. Nucleic Acids Res *47*, 11497-11513.
631 Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., Zhou, Z., Dai, Y., Yang, Y., Liu, P*., et al.*
632 (2016). Global intron retention mediated gene regulation during CD4+ T cell activation. Nucleic Acids
633 Res *44*, 6817-6829.
634 Nordström, K.J.V., Schmidt, F., Gasparoni, N., Salhab, A., Gasparoni, G., Kattler, K., Müller, F., Ebert,
635 P., Costa, I.G., consortium, D*., et al.* (2019). Unique and assay specific features of NOMe-, ATAC- and
636 DNase I-seq data. Nucleic Acids Research *47*, 10580-10596.
637 Pacini, C., and Koziol, M.J. (2018). Bioinformatics challenges and perspectives when studying the
638 effect of epigenetic modifications on alternative splicing. Philos Trans R Soc Lond B Biol Sci *373*.
639 Pott, S. (2017). Simultaneous measurement of chromatin accessibility, DNA methylation, and
640 nucleosome phasing in single cells. Elife *6*.
641 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
642 features. Bioinformatics *26*, 841-842.
643 Radman-Livaja, M., and Rando, O.J. (2010). Nucleosome positioning: How is it established, and why
644 does it matter? Developmental Biology *339*, 258-266.
645 Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F.,
646 and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis.
647 Nucleic Acids Res *44*, W160-165.
648 Rauschert, S., Raubenheimer, K., Melton, P.E., and Huang, R.C. (2020). Machine learning and clinical
649 epigenetics: a review of challenges for diagnosis and classification. Clinical Epigenetics *12*, 51.
650 Robinson, J.T., Thorvaldsdottir, H., and Mesirov, J. (2012). Exploring cancer datasets in the integrative
651 genomics viewer (IGV). Cancer Research *72*.
652 Saldi, T., Cortazar, M.A., Sheridan, R.M., and Bentley, D.L. (2016). Coupling of RNA Polymerase II
653 Transcription Elongation with Pre-mRNA Splicing. J Mol Biol *428*, 2623-2635.
654 Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. (2015). Alternative
655 Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI). Curr Protoc Hum Genet *87*, 11161-
656 111614.
657 Schmitz, U., Pinello, N., Jia, F., Alasmari, S., Ritchie, W., Keightley, M.C., Shini, S., Lieschke, G.J.,
658 Wong, J.J., and Rasko, J.E.J. (2017). Intron retention enhances gene regulatory complexity in
659 vertebrates. Genome Biol *18*, 216.
660 Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure.
661 Nature Structural & Molecular Biology *16*, 990-995.
662 Simon, J.M., Hacker, K.E., Singh, D., Brannon, A.R., Parker, J.S., Weiser, M., Ho, T.H., Kuan, P.-F.,
663 Jonasch, E., Furey, T.S*., et al.* (2014). Variation in chromatin accessibility in human kidney cancer links
664 H3K36 methyltransferase loss with widespread RNA processing defects. Genome research *24*, 241-
665 250.
666 Singer, M., Kosti, I., Pachter, L., and Mandel-Gutfreund, Y. (2015). A diverse epigenetic landscape at
667 human exons with implication for expression. Nucleic Acids Research *43*, 3498-3508.
668 Smart, A.C., Margolis, C.A., Pimentel, H., He, M.X., Miao, D., Adeegbe, D., Fugmann, T., Wong,
669 K.K., and Van Allen, E.M. (2018). Intron retention is a source of neoepitopes in cancer. Nat Biotechnol
670 *36*, 1056-1058.
671 Statham, A.L., Strbenac, D., Coolen, M.W., Stirzaker, C., Clark, S.J., and Robinson, M.D. (2010).
672 Repitools: an R package for the analysis of enrichment-based epigenomic data. Bioinformatics *26*,
673 1662-1663.
674 Strobl C, B.A., Kneib T, Augustin T, Zeileis A (2008). Conditional Variable Importance for Random
675 Forests. BMC Bioinformatics *9*.
676 Szerlong, H.J., and Hansen, J.C. (2011). Nucleosome distribution and linker DNA: connecting nuclear
677 function to dynamic chromatin structure. Biochem Cell Biol *89*, 24-34.
678 Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J., and Jones, P.A. (2014). Reconfiguration of
679 nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers
680 and insulators in cancer. Genome Research *24*, 1421-1432.
681 Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009).
682 Nucleosome positioning as a determinant of exon recognition. Nature Structural & Molecular Biology
683 *16*, 996-1001.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Ullrich, S., and Guigo, R. (2020). Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. Nucleic Acids Res *48*, 1327-1340.

Wallner, S., Schroder, C., Leitao, E., Berulava, T., Haak, C., Beisser, D., Rahmann, S., Richter, A.S., Manke, T., Bonisch, U.*, et al.* (2016). Epigenetic dynamics of monocyte-to-macrophage differentiation. Epigenetics Chromatin *9*, 33.

Wang, L., Wang, Y., Su, B., Yu, P., He, J., Meng, L., Xiao, Q., Sun, J., Zhou, K., Xue, Y.*, et al.* (2020). Transcriptome-wide analysis and modelling of prognostic alternative splicing signatures in invasive breast cancer: a prospective clinical study. Scientific Reports *10*, 16504.

Wei, G., Liu, K., Shen, T., Shi, J., Liu, B., Han, M., Peng, M., Fu, H., Song, Y., Zhu, J.*, et al.* (2018). Position-specific intron retention is mediated by the histone methyltransferase SDG725. BMC Biol *16*, 44.

Wong, J.J., Au, A.Y., Ritchie, W., and Rasko, J.E. (2016). Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. Bioessays *38*, 41-49.

Wong, J.J., Gao, D., Nguyen, T.V., Kwok, C.T., van Geldermalsen, M., Middleton, R., Pinello, N., Thoeng, A., Nagarajah, R., Holst, J.*, et al.* (2017a). Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. Nat Commun *8*, 15134.

Wong, J.J., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K.*, et al.* (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583-595.

Wong, J.J.L., Gao, D.D., Nguyen, T.V., Kwok, C.T., van Geldermalsen, M., Middleton, R., Pinello, N., Thoeng, A., Nagarajah, R., Holst, J.*, et al.* (2017b). Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. Nature Communications *8*.

Wu, J.N., Pinello, L., Yissachar, E., Wischhusen, J.W., Yuan, G.-C., and Roberts, C.W.M. (2015). Functionally distinct patterns of nucleosome remodeling at enhancers in glucocorticoid-treated acute lymphoblastic leukemia. Epigenetics & Chromatin *8*, 53.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R.*, et al.* (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394.

You, J.S., Kelly, T.K., De Carvalho, D.D., Taberlay, P.C., Liang, G., and Jones, P.A. (2011). OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. Proceedings of the National Academy of Sciences *108*, 14497-14502.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhou, H.L., Luo, G., Wise, J.A., and Lou, H. (2014). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. Nucleic Acids Res *42*, 701-713.

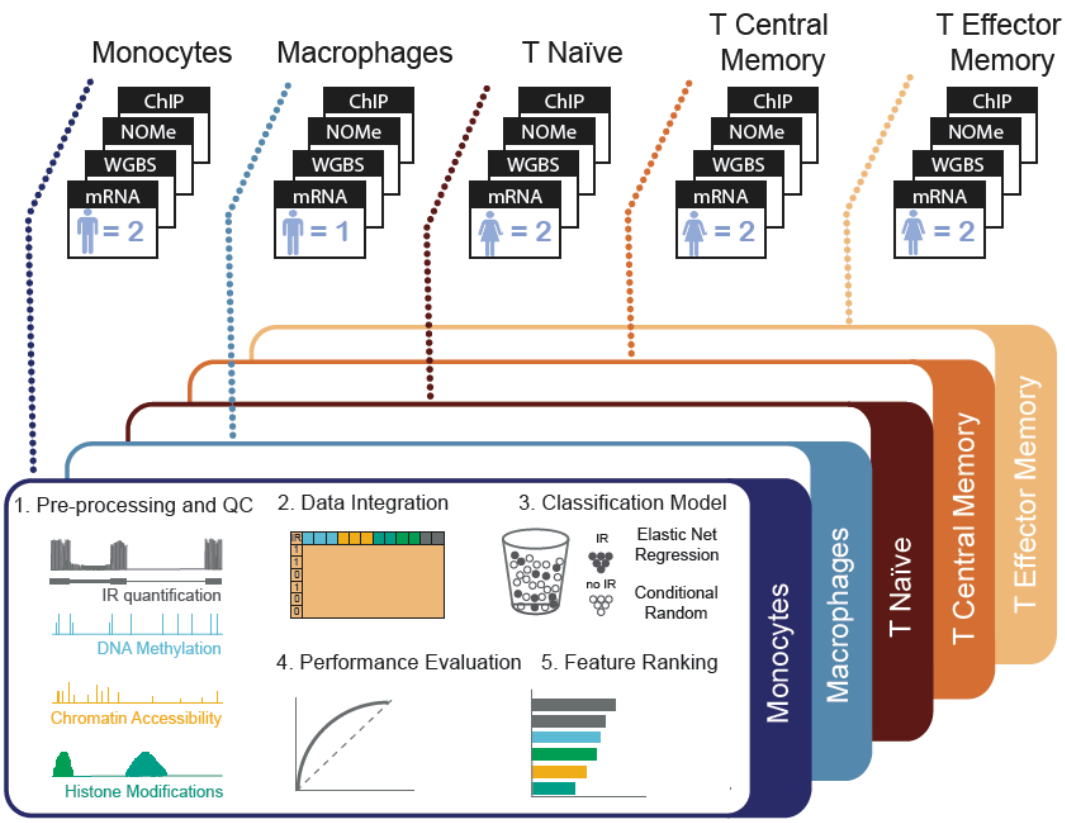


Figure 6 Experimental design and workflow to determine regulators of IR. Raw high-throughput data were processed for each biological replicate and amalgamated by cell type from the indicated number of samples (n). The output was used for feature extraction: IR events were treated as a binary outcome and we trained an Elastic Net (EN) regression model and a conditional Random Forest model with a total of 46 sequence-based and epigenetic features. Using feature ranking, we identified the factors that were most strongly associated with IR outcomes and compared the performances of both modelling strategies. These steps were repeated for each cell type.

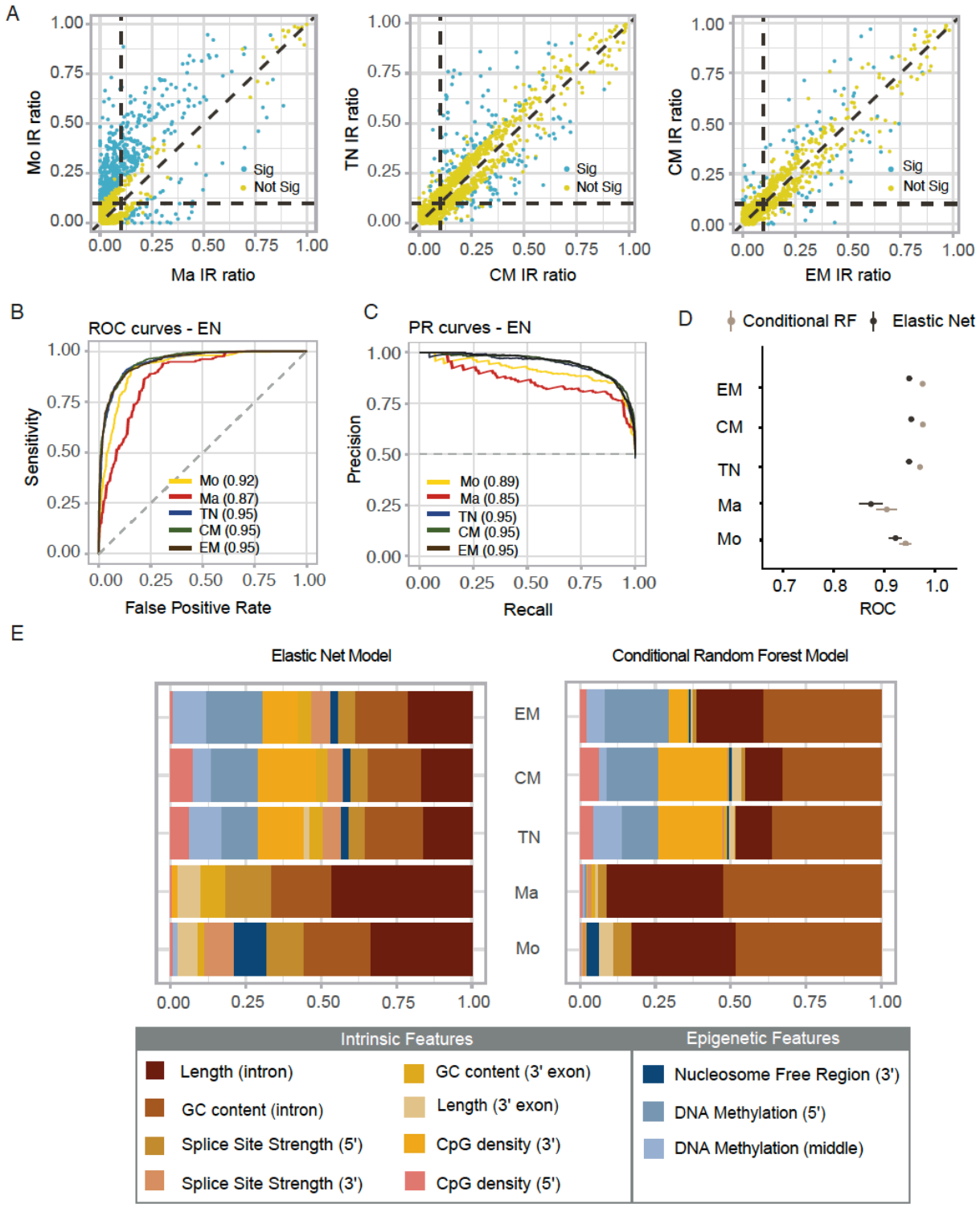


Figure 7 IR prediction and model feature association analyses. (A) Scatter plot of differential IR events (Sig blue – significant; Not Sig yellow – not significant) between monocytes (Mo) vs macrophages (Ma) (left), Naïve (TN) vs Central Memory (CM) T cells (middle), and Central Memory vs Effector Memory (EM) T cells (right). **(B)** Receiver operating characteristic (ROC) curves and **(C)** precision recall (PR) curves comparing the performance of the EN classifier in five cell types. **(D)** Comparison of AUC values between EN and cRF algorithms, error bars show 95% confidence interval. **(E)** Variable importance scores for the top 10 features identified by EN and conditional RF algorithms. The scores were scaled to values that add up to 1.0 and the size of a bar corresponds to the effect size.

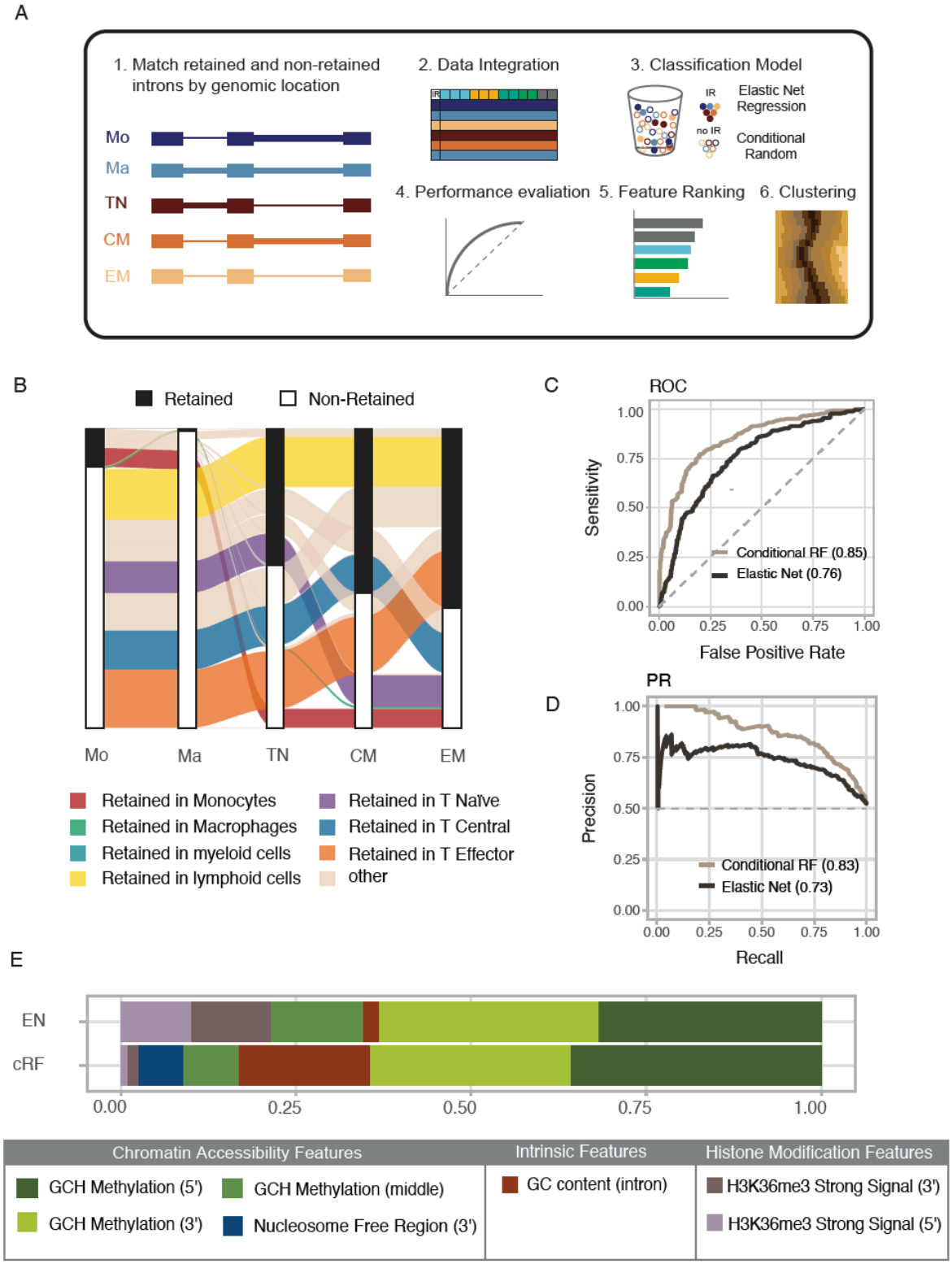


Figure 8. Analysis of dynamics intron retention. (A) Modified modelling strategy from Figure 1. Only introns that were found to be in retained and non-retained states in different cell types were included in the analysis. **(B)** Alluvial plot illustrating the dynamics of IR states among the five cell types. **(C)** ROC and **(D)** PR curves comparing the performance of cRF (brown) and EN (black). **(E)** Variable importance scores for the top 5 features identified by EN and conditional RF algorithms, scaled between 0 and 1.

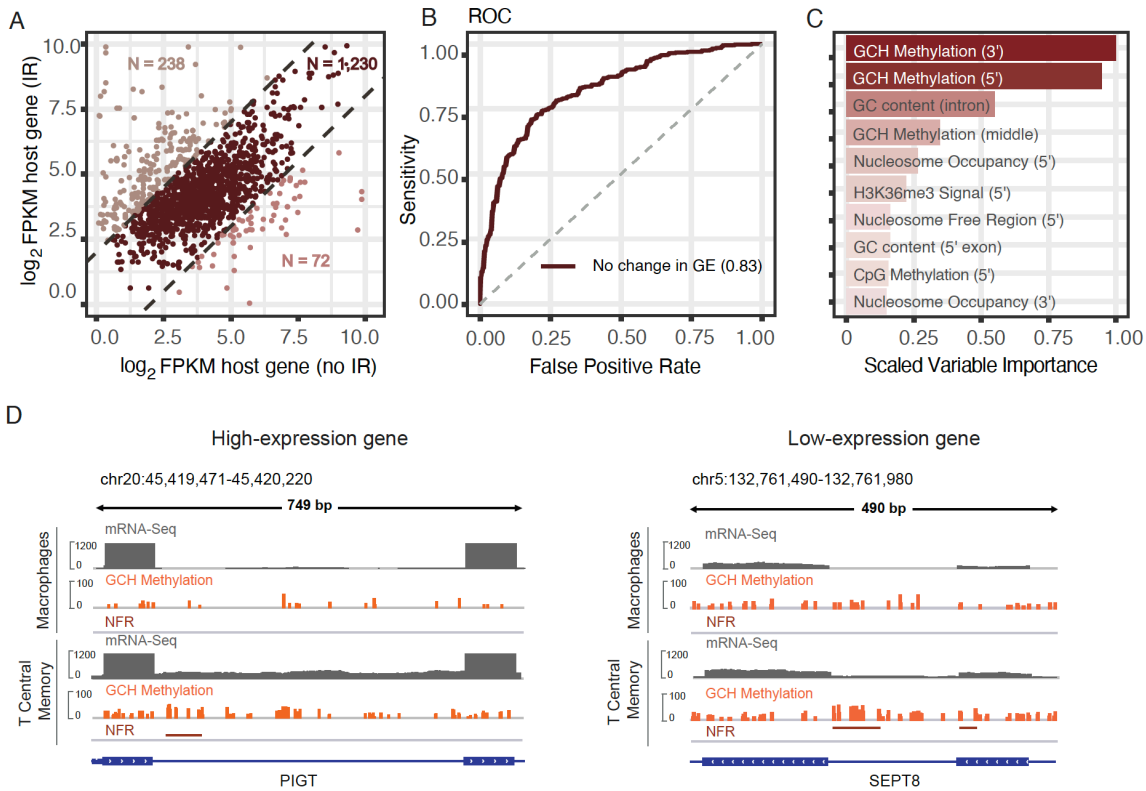
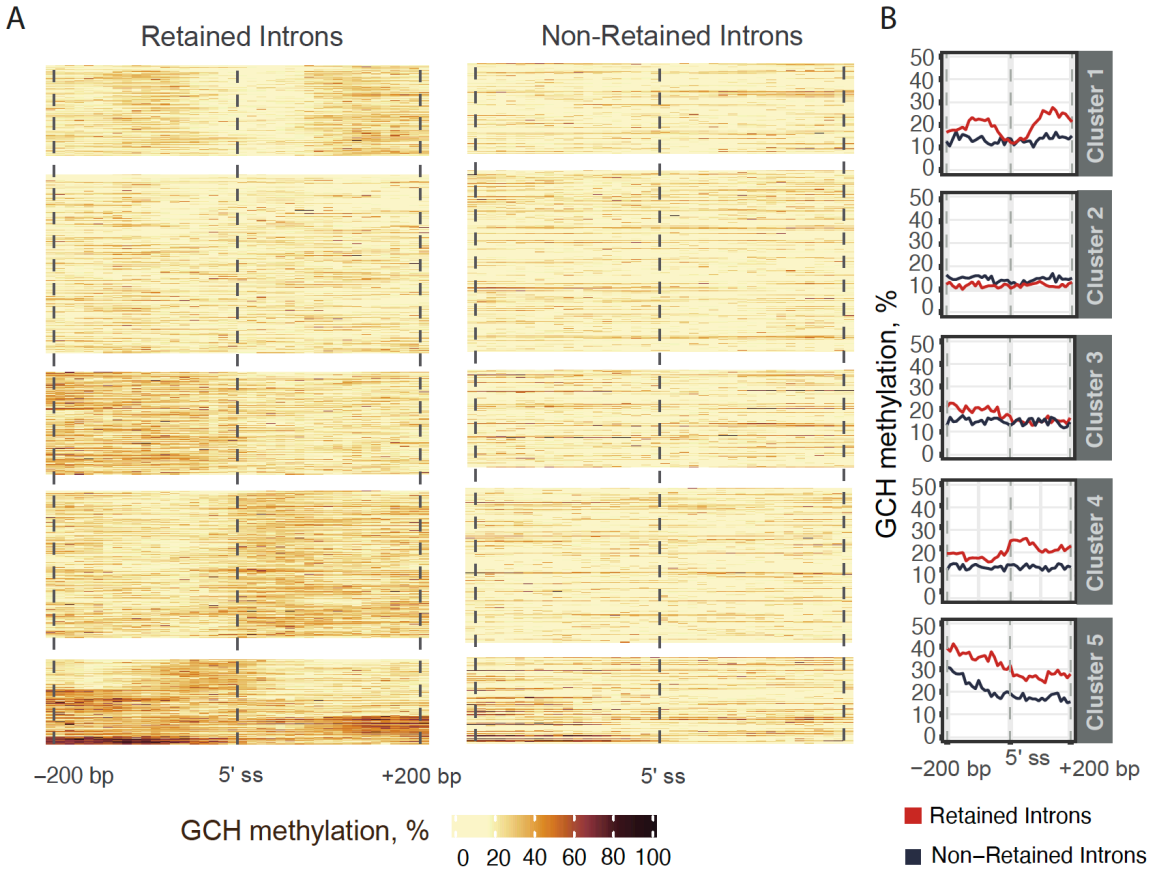


Figure 9 Analysis of introns from genes with non-differential expression levels. (A) Scatter plot of host gene expression for introns that change their IR status. **(B)** ROC curve indicating the performance of a conditional RF model fitted on the data from non-differentially expressed genes (GE, gene expression). **(C)** Ranking of the features based on the scaled variable importance scores. **(D)** Integrative Genomics Viewer (IGV) plots revealing higher density and hypermethylation levels of GCH sites in the splice site regions of differentially retained introns in both highly- and lowly- expressed gene examples (NFR – Nucleosome Free Region, GCH Methylation – methylation levels of GC dinucleotides followed by any nucleobase except guanine).



759

760 **Figure 10 GCH methylation clustering in differentially retained introns. (A)** Clustering of GCH methylation in the +/-
761 200 bp region around the 5′ splice site (ss). Each line corresponds to one intron that is either in a retained (left) or non-retained
762 state (right). **(B)** Line plots showing average GCH methylation values (i.e. chromatin accessibility) in retained vs non-retained
763 introns across 5 clusters.

764

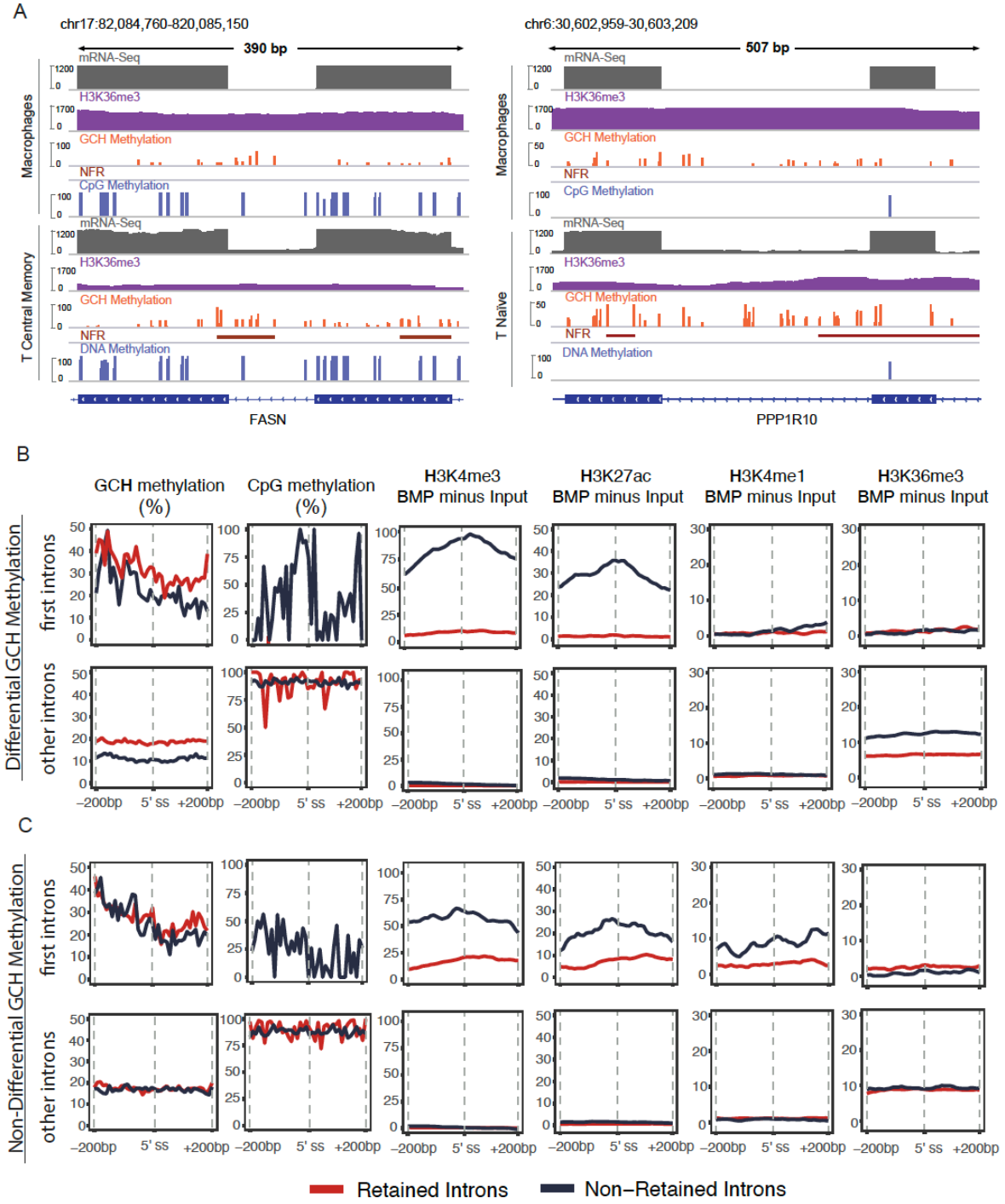


Figure 6 Interplay between chromatin accessibility, CpG methylation and histone modifications. (A) IGV plots of mRNA-seq, H3K36me3 ChIP-seq, NOMe-seq, and WGBS-seq data indicating different levels of GCH methylation between retained and non-retained introns and higher prevalence of NFRs in the regions proximal to IR. **(B)** Line graphs show the average levels of GCH methylation, CpG methylation, and the difference between ChIP-seq H3K4me3, H3K27ac, H3K4me1, and H3K36me3 signals and ChIP-Seq Input, normalised to the Bins Per Million (BPM), in retained (red) and non-retained (blue) introns associated with chromatin status. The first row shows epigenetic signals at the 5′ splice site of first introns (close to the promoter region) and the second row represents all other introns. **(C)** The same analysis performed in **(B)** is repeated for introns where the chromatin status remains the same, i.e. non-differential GCH methylation.

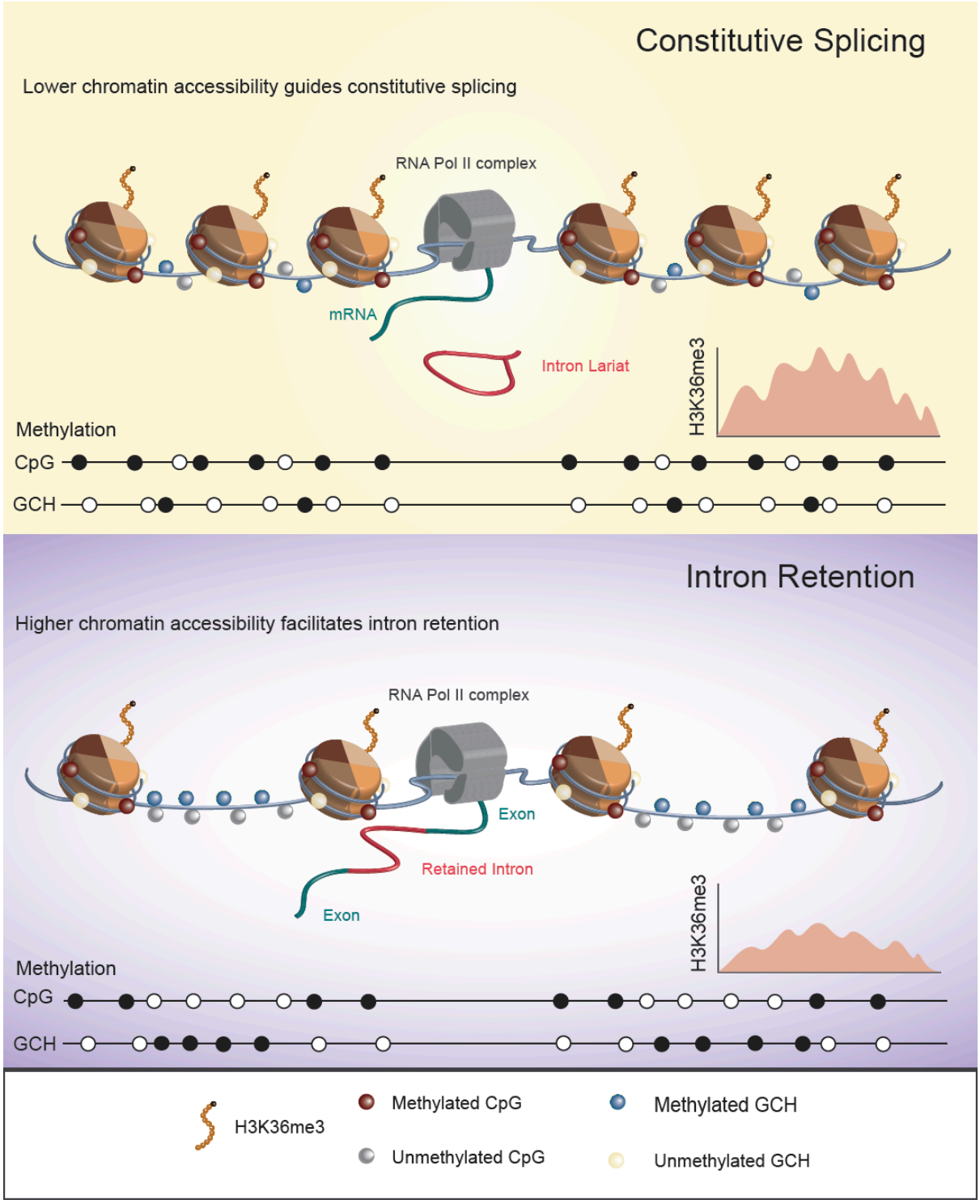


Figure 7 Proposed role of chromatin accessibility in IR regulation. More dense positioning of nucleosomes slows down RNA Pol II elongation rate, allowing sufficient time for a histone modification (in this case, H3K36me3). Methylated CpG dinucleotides and unmethylated GCH sites over the nucleosome core explain higher CpG methylation levels and lower GCH methylation levels in constitutively spliced introns.