



HAL
open science

G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing

Cyril Esnault, Encar Garcia-Oliver, Amal Zine El Aabidine, Marie-Cécile Robert, Talha Magat, Kevin Gawron, Eugénia Basyuk, Magda Karpinska, Alexia Pigeot, Anne Cucchiarini, et al.

► To cite this version:

Cyril Esnault, Encar Garcia-Oliver, Amal Zine El Aabidine, Marie-Cécile Robert, Talha Magat, et al.. G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing. 2023. hal-04286714

HAL Id: hal-04286714

<https://hal.science/hal-04286714v1>

Preprint submitted on 15 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing

Cyril Esnault¹, Encar Garcia-Oliver¹, Amal Zine El Aabidine¹, Marie-Cécile Robert^{1,2,8}, Talha Magat¹, Kevin Gawron¹, Eugénia Basyuk^{1,3}, Magda Karpinska^{1,2,8}, Alexia Pigeot¹, Anne Cucchiaroni⁴, Yu Luo^{4,5}, Daniele Verga⁵, Raphael Mourad⁶, Ovidiu Radulescu⁷, Jean-Louis Mergny⁴, Edouard Bertrand^{2,8} and Jean-Christophe Andrau^{1#}

¹ Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS-UMR 5535, 1919 Route de Mende, 34293 cedex 5, Montpellier, France.

² Institut de Génétique Humaine, CNRS-UMR9002, University of Montpellier, 141 rue de la Cardonille, 34396, Montpellier, France.

³ Laboratoire de Microbiologie Fondamentale et Pathogénicité, CNRS-UMR 5234, Université de Bordeaux, Bordeaux, France.

⁴ Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris 91128 Palaiseau, France.

⁵ Université Paris Saclay, CNRS UMR9187, INSERM U1196, Institut Curie, Orsay, France.

⁶ LBCMCP, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 31062 Toulouse, France.

⁷ Laboratory of Pathogen Host Interactions, UMR CNRS 5235, University of Montpellier, Montpellier, France.

⁸ Equipe labélisée Ligue Nationale Contre le Cancer, Montpellier.

Corresponding author: jean-christophe.andrau@igmm.cnrs.fr

1 **Abstract**

2 Despite their central role in transcription, it has been difficult to define universal
3 sequences associated to eukaryotic promoters. Within chromatin context, recruitment
4 of the transcriptional machinery requires opening of the promoter but how DNA
5 elements could contribute to this process has remained elusive. Here, we show that
6 G-quadruplex (G4) secondary structures are highly enriched mammalian core
7 promoter elements. G4s are located at the deepest point of nucleosome exclusion at
8 promoters and correlate with maximum promoter activity. We found that experimental
9 G4s exclude nucleosomes both *in vivo* and *in vitro* and display a strong positioning
10 potential. At model promoters, impairing G4s affected both transcriptional activity and
11 chromatin opening. G4 destabilization also resulted in an inactive promoter state and
12 affected transition to effective RNA production in live imaging experiments. Finally, G4
13 stabilization resulted in global reduction of proximal promoter pausing. Altogether, our
14 data introduce G4s as *bona fide* promoter elements allowing nucleosome exclusion
15 and facilitating pause release by the RNA Polymerase II.

16 **Introduction**

17 Initially defined by analogy to bacterial transcription models (Jacob et al., 1964;
18 Pribnow, 1975; Rosenberg and Court, 1979) and based on *in vitro* transcription assays
19 from the pre-genomic era, eukaryotic core promoters were defined as '*the minimal
20 stretch of contiguous DNA sequence that is sufficient to direct accurate initiation of
21 transcription by the RNA polymerase II machinery*' (Butler and Kadonaga, 2002).
22 However and unlike in bacteria, eukaryotic promoters require nucleosome exclusion
23 for the transcriptional machinery to be recruited. Beyond the concept of 'core
24 promoters', eukaryotic 'promoters' can be characterized by wider sequence contexts
25 that are highly divergent depending on the species. Nevertheless, these sequences
26 carry in common to generate a well-positioned array of nucleosome (Jiang and Pugh,
27 2009; Radman-Livaja and Rando, 2010) and the ability to open chromatin. In mammals
28 this later property is in part carried over by CpG islands (CGIs) (Deaton and Bird,
29 2011), characterized by large regions of high GC and CpG content that encompass a
30 large fraction of promoters. Importantly, CGIs can open chromatin by default and in a
31 transcription-independent manner (Fenouil et al., 2012).

32 Promoters are also highly enriched in potential DNA secondary structures, suggesting
33 that these structures play a role in transcription regulation (Bansal et al., 2014). Among
34 these, G-quadruplexes (G4s) are over-represented in regulatory regions. G4s are
35 single-stranded and stable structures *in vitro* that consist in planar arrangement of
36 guanines stabilized by K⁺ at physiological concentrations. They have been involved in
37 number of nuclear processes and are present at over a million occurrence in the
38 genome and more specifically at promoters. However, and to date it was unclear to
39 what extent they could contribute as positive regulators of transcription since it was
40 described that they either inhibit or repress transcription depending on their promoter

41 context (Agarwal et al., 2014; Bochman et al., 2012; David et al., 2016; Smestad and
42 Maher, 2015). Furthermore, while it was proposed that their formation *in vivo* could be
43 dependent on high level of transcriptional activity (Hansel-Hertsch et al., 2016; Xia et
44 al., 2018), the possibility that they could represent promoter elements on their own was
45 never directly tested.

46 Here, we shed light on G4s as highly enriched mammalian core promoter elements.
47 Using both predicted and experimental G4 data, *in vitro* and *in vivo*, we find them
48 located at the deepest point of nucleosome exclusion at promoters correlating with
49 maximum promoter activity in SURE assay. Furthermore, we found that G4s exclude
50 nucleosomes both *in vivo* and *in vitro*, and display a strong nucleosome positioning
51 potential. Impairing G4 formation by specific mutations at model promoters affected
52 both transcription activity and chromatin opening. G4 destabilization also resulted in
53 increased probability of promoters to be in an inactive state (OFF times) and affected
54 transition to effective RNA production in live imaging experiments. Finally, G4
55 stabilization using ligands, resulted in global reduction of proximal promoter pausing
56 by Pol II consistent with our live imaging observations. Altogether, our data introduce
57 G4s as *bona fide* promoter elements allowing nucleosome exclusion and facilitating
58 pause release by Pol II.

59 **G-quadruplexes are highly enriched at mammalian promoters and correlate with** 60 **maximum promoter activity**

61 Based on the initial knowledge that the TFIID general transcription factor (GTF) binds
62 naked DNA *in vitro* in a window frame of 40-50 bp upstream and downstream of
63 transcription start sites (TSSs) (Buratowski et al., 1989), *in silico* sequence analyses
64 of core promoters were often restricted to this short window frame. However, these
65 searches often yielded motifs poorly enriched, lowly conserved in evolution or highly
66 degenerated (Haberle and Stark, 2018; Vo Ngoc et al., 2017). Because the most open
67 areas of chromatin extend on average up to 100 bp, we performed a motif search on
68 core promoters associated to open chromatin upstream and downstream (-100/+20)
69 of experimental TSSs in three mammalian cell types (primary T cells, K562 and Raji
70 cells - Figure 1A and S1; see also Table S1 for data sets used in this study). This
71 analysis revealed a prominent motif consistent with SP1 binding site with additional G
72 stretches in the 3 cell types. These stretches are highly compatible with the formation
73 of G4s *in vitro*, using the G4Hunter (G4H) predictive algorithm (Bedrat et al., 2016) at
74 various stringencies. To consolidate this result, we investigated the frequency of
75 predicted G4s (pG4s) or other motifs identified and found overall that stringent G-
76 quadruplex predictions (pG4s at G4H1.5 and 2.0 thresholds) (Bedrat et al., 2016) show
77 very high frequency as well as a strong enrichment above control sequences
78 (observed/expected) (Bagchi and Iyer, 2016; Fenouil et al., 2012). Enrichments of
79 pG4s (20-45%) are also higher than those of Ets and NF-Y and far above the TATA
80 box motifs (Figure 1A, S1 and Table S2). Their enrichment is similar to that of the BRE
81 and SP1 motifs, both compatible with G4 formation, which show high overlaps at

82 promoters (Figure S2A). Finally, we note that promoters containing pG4s tend to
83 exhibit less of the other motifs (Figure S2B) suggesting that they could have more
84 propensity to function autonomously.

85 Next, we analysed the enrichment of pG4s upstream of sense and antisense
86 transcription occurring at mammalian promoters. This showed that pG4s are found on
87 average at positions at -56 and -51 of the TSS, respectively (Figure 1B). We note that
88 while poorly enriched at promoters, when present, TATA boxes influence a far more
89 directional and focused transcription (Figure S2C) as described previously (Bagchi and
90 Iyer, 2016; Fenouil et al., 2012).

91 To further investigate whether G4s contribute to transcription initiation and promoter
92 activity, we took advantage of four orthogonal approaches for G4 formation
93 assessment, including G4access a technology we recently developed in our laboratory
94 (Garcia-Oliver et al., 2022). G4access is an antibody-independent technology
95 alternative and complementary to G4 ChIP. In brief, this method is based on moderate
96 nuclease digestion of chromatinized DNA and allows G4 formation assessment
97 genome-wide in the chromatin context (Figure 1C). We also used G4 ChIP (Hansel-
98 Hertsch et al., 2016; Mao et al., 2018), mapping G4s in living cells, G4seq (Chambers
99 et al., 2015) globally assessing G4s *in vitro* and ss-DNA-seq (Kouzine et al., 2013)
100 mapping single stranded DNA genome-wide. We performed G4access in Raji and
101 K562 cells and processed published data sets for the other methods whenever
102 available (Figure S2D-E). In this analysis, we ranked the promoters containing
103 predicted G4s (G4Hunter>2.0) by increasing experimental G4 signal. G4seq and
104 ssDNA-seq in Raji cells were used to further validate the selection of promoters with
105 folded G4s. G4access and G4-ChIP outputs of G4 measurement are very comparable
106 (Figure S2D-E). ssDNA-seq globally confirmed G4 formation in Raji cells while G4-
107 seq, a technique that maps G4s formed *in vitro* on naked genomic DNA (Chambers et
108 al., 2015), validated that selected sequences can form G-quadruplexes *in vitro*. Since
109 G4-seq monitors genomic G4 formation outside of nuclear context, signals of the 6
110 groups originally defined from low to high G4 formation in living cells remain largely
111 unchanged. This further indicates that G4access and ChIP do map G4s in the context
112 of chromatin in living samples.

113 We then compared G4 formation to data sets (Table S1) monitoring nascent
114 transcriptome and large-scale measurement of promoter activity assay (SURE) (van
115 Arensbergen et al., 2017) (Figure 1D-E). As illustrated for the BTG2 promoter in K562
116 cells, both predicted and experimental G4 signals correspond to the midpoint of
117 promoter divergent transcription and maximum promoter activity by SURE (Figure 1D).
118 To establish this statement globally, we analysed the correlation of G4access at pG4
119 promoter locations with that of SURE and Pol II (ENCODE) and found that both
120 increase with experimental G4 levels (Figure 1E). In addition, we observed that pG4
121 locations overlapped with the midpoints between sense and anti-sense transcription
122 initiation mapped by ChIP-exo or ChIP-seq of Pol II, TBP, TFIIB (Pugh and Venters,

123 2016) and by GRO-CAP (Core et al., 2014) (Figure 1F and S2G). Finally, we also
124 observed that promoter pG4s localise upstream of R-loops (Figure S2F) which form
125 where Pol II and GRO-cap signal start to rise, on the side of the pG4s (Figure 1F).

126 Collectively, these analyses show that pG4s represent major motifs of extended core
127 promoters, located on average at a relatively fixed position from TSSs. Moreover, G4s
128 correlate with both promoter activity and midpoints of divergent transcription (Figure
129 1G).

130 **G4 forming sequences carry an intrinsic ability for nucleosome exclusion**

131 Apparent Nucleosome Depleted Regions (NDRs) are hallmarks of core promoters in
132 eukaryotic cells allowing space for PIC recruitment (Andersson and Sandelin, 2020;
133 Haberle and Stark, 2018). To understand the link between G4s/pG4s and nucleosome
134 positioning, we performed nucleosome mapping by MNase-seq in Raji cells and re-
135 analysed published datasets, for nucleosomes and active epigenetic marks in our two
136 other mammalian models (Table S1). Strikingly, we found the center of NDRs
137 overlapping with pG4s at a very large fraction of promoters, including at the BTG2
138 promoter described above (Figure 2A). We then investigated all active promoters that
139 contain pG4s and confirmed that pG4s are found at the deepest points of NDRs
140 globally (Figure 2B, Figure S2H). By comparing increasing G4 signals to promoter
141 opening, we observed less opening in the absence of G4access signal (group1).
142 Conversely, deeper and larger NDRs were present when G4access was present
143 (group 2-5), suggesting a threshold effect in the G4access signal. Active histone marks
144 increased together with G4 strength (group 1-6), indicating overall chromatin opening
145 and modifications depending on the G4 formation (Figure 2B). At inactive promoters,
146 the presence of predicted G4s also hallmarked Polycomb-deposited H3K27me3
147 inactive chromatin mark (Figure S2I). This set of promoters carry the hallmark of CpG
148 islands, with strong GC and CpG content as expected for Polycomb signal. Since SP1
149 motifs are also highly enriched at promoters (Reed et al., 2008) and represent half of
150 a pG4 motif (Huppert et al., 2008), we questioned whether SP1 binding or motif could
151 influence the observed chromatin opening by G4s. We analysed pG4 promoters with
152 or without SP1 binding and observed similar NDR formation, reasonably allowing to
153 exclude that SP1 would be responsible for the pG4 property (Figure S3A). Similarly,
154 promoters that did not carry the canonical or non-canonical SP1 motifs did also show
155 openness at G4 motifs (Figure S3B-C). In this case, we observed residual SP1 binding
156 indicating that SP1 might directly bind G4s in accordance with published observations
157 (Lago et al., 2021; Raiber et al., 2012).

158 Next, we tested whether nucleosome exclusion at pG4s was dependent on
159 transcriptional activity. For this, we analysed nucleosome densities at inactive
160 promoters (see methods) that were separated into 2 groups, with or without pG4s.
161 Interestingly, only pG4-containing promoters showed significant nucleosome exclusion
162 (Figure 2C). We further confirmed that transcription is not required for nucleosome

163 exclusion at pG4s of active promoters by inhibiting Pol II transcription with α -amanitin
164 without substantial loss of NDRs (Figure S4A). To further infer and demonstrate the
165 direct link between pG4 sequences and chromatin opening, we made use of
166 reconstituted nucleosomes *in vitro* (Valouev et al., 2011). In this assay, chromatin was
167 reconstituted using human genomic DNA and recombinant histone before MNase
168 digestion. As for the *in vivo* analyses, promoters were split in 2 groups, with or without
169 strong G4 predictions (Figure 2D). This analysis revealed that only the pG4-containing
170 group associates with nucleosome exclusion. Thus, pG4 DNA carry the intrinsic ability
171 to exclude nucleosomes since *in vitro*, in absence of any other transcription factor or
172 proteins, we could observe this property. It also indicates that nucleosome occupancy
173 and G4 formation are mutually exclusive. This ability to exclude nucleosome is similarly
174 observed for the SP1 unbound promoters (Figure S3). Together, our data show that at
175 both active and inactive promoters, G-quadruplexes promote intrinsic nucleosome
176 exclusion *in vivo* and *in vitro*.

177 **G4 forming sequences are nucleosome organizers**

178 Next, we investigated chromatin opening and organisation around G4 predictions at
179 non-promoter locations, including intergenic regions. These locations are mainly not
180 transcribed and yet, pG4s are associated to nucleosome exclusion (Figure 3A)
181 indicating that these sequences carry this property over the whole genome and not
182 only at promoters. Further investigation of *in vitro* reconstituted chromatin (Valouev et
183 al., 2011) at all genomic locations including promoters, intergenic and intragenic
184 regions demonstrated their intrinsic abilities to evict nucleosome globally (Figure S4B-
185 C). Nucleosome exclusion was found at over 90% for promoter pG4s, and over 80%
186 for intra and intergenic pG4s (Figure S4C). Interestingly, the regions showing the
187 largest nucleosome exclusion areas also showed the highest G4 prediction densities
188 (upper part of the heatmaps), possibly because their presence in clusters increase
189 chances of G4 formation. These observations are consistent with the high stability of
190 G4 single-stranded DNA structures (Guedin et al., 2010; Sen and Gilbert, 1988) which
191 makes them incompatible with nucleosome formation.

192 To infer how nucleosome behave around G4 forming sequences, we analyzed both
193 nucleosome densities and midpoints around intergenic G4 predictions in our model
194 cell lines. Midpoints analyses allow to better assess if given sequence locations display
195 positioning properties (Figure 3A). This clearly revealed a high level of positioning
196 associated to pG4s in the model cells. The periodicity of observed nucleosome
197 positioning is highly similar to nucleosomal organization around specific pioneer
198 transcription factors (Barozzi et al., 2014) or the insulator factor CTCF (Fu et al., 2008)
199 showing an almost identical nucleosomal repeat length (Figure 3B-C), in line with
200 previous observations (Kouzine et al., 2017). Hence, our results highlight the
201 association of pG4s with open chromatin regions at promoters and at IGRs where
202 pG4s also associate with nucleosome array organization. Although experimental G4
203 signals correlate with transcription, pG4/G4 driven nucleosome depletion appears

204 independent of transcription. This observation is consistent with results described
205 recently by us and others (Garcia-Oliver et al., 2022; Shen et al., 2021). Our results
206 support the role of a novel and previously unappreciated role for G4s as global
207 chromatin organisers at transcribed and un-transcribed regions.

208 **Increased chromatin opening at CpG islands containing pG4s**

209 Since CpG islands (CGIs) are able to promote nucleosome depletion at promoters
210 (Fenouil et al., 2012), we wondered what was the contribution of pG4s in this process.
211 To address this question, we considered all human CGI annotations containing or not
212 strong G4 predictions (using G4H2.0, Figure 4). We analysed genomic features
213 associated to experimental G4s: nucleosome positioning, active chromatin marks,
214 transcription and promoter activity. As expected, G4access and ChIP-seq show
215 stronger signals at CGIs harboring pG4s. In addition, active histone marks (H3K4me3
216 and H3K27ac) and Pol II are also enhanced in this class (Figure 4A-B). Furthermore,
217 nucleosome occupancy exhibits wider and deeper chromatin opening at pG4-
218 containing CGIs. Finally, the analysis of promoter activity by SURE assay confirmed
219 that G4-containing CGIs have a ~1.5 time higher promoter activity. To further validate
220 our results, we confirmed our analyses on a set of CGIs of same lengths and CG
221 contents (Figure S5A) with a more stringent selection of non-G4 forming sequences.
222 The G4 forming sequences were considered with a G4Hunter score > 1.5 and non-
223 forming sequences for a G4Hunter score < 1.2. This analysis confirmed the association
224 of G4 forming sequences to more open chromatin and active transcription and
225 epigenetic marking within CpG islands (Figure S5B-C). Taken together, our
226 observations support the notion that features characteristic of promoters are enhanced
227 in the presence of both experimental and predicted G4s at CGIs and that G4s might
228 represent essential determinants of CGI's ability to exclude nucleosome.

229 **G4 mutations at promoters result in decreased transcription in single cells**

230 To fully demonstrate that G4s are promoter element that extend the concept of the
231 core promoter, we mutated G4s in model promoters. We inserted G4-containing
232 mouse promoters in the human genome in Hela cells using a Flip-in system as
233 described previously (Tantale et al., 2016a). The promoters were located upstream of
234 a reporter gene containing 256 MS2 repeats and allowing single cell measurement of
235 transcription by smRNA FISH (Figure 5A). We chose promoters that contained a strong
236 G4 prediction that was also verified experimentally using G4access (Garcia-Oliver et
237 al., 2022) or G4 cut&Tag (Lyu et al., 2022) in mouse ES cells (Figure S6A-E). For 3
238 out of 5 models (Taok1, Pkm and Klf6), the sequences considered did not contain any
239 TATA box, and for 3 of them no SP1/GC box site (Taok1, Pkm and Klf6). We designed
240 G4 mutations that minimally affected the promoter primary sequence while impairing
241 the G4 potential (Figure 5B). We also verified that the structure was abolished *in vitro*
242 using 3 independent biophysical assays that included circular dichroism, TDS and IDS
243 (thermal and isothermal differential spectra, Figure 5B). In the case of the Pol2ra, we

244 performed 2 independent mutations that affected differentially the G4 potential
245 assessed by the G4hunter algorithm (Figure 5B and table S3), one of which respecting
246 the GC content of the sequence. To quantify the transcriptional output of the model
247 promoters, we performed single cell measurements using smRNA FISH over hundreds
248 of cells. The data presented in Figure 5C shows G4 mutant's transcription as compared
249 to their WT promoter counterpart. The level of nascent RNA reduction ranked from 30
250 to 80% in the mutants, indicating that the G4 mutations had a significant effect on
251 transcription of the core promoter. Interestingly also, inverting the G4 orientation in the
252 promoter did not impact on its activity in the case of the Eef1a1 model (Figure S6F),
253 suggesting that the G4 functions as promoter element in an orientation independent
254 manner.

255 Together, our data show that G4 mutations impair transcription quantitatively and
256 establish that G4-forming structures most likely function as promoter elements in an
257 orientation independent manner.

258 **Modelling transcription in presence or absence of G4 and TATA box elements**

259 To decode the direct effect of G4s as promoter elements on transcription, as compared
260 to that of the TATAbox, we performed further mutational analyses in the context of the
261 Eef1a1 promoter that contains both a canonical TATA box and a stable G4. We
262 mutated the G4, the TATA, or both elements (Figure 6A). Minimal substitution of the
263 TATA box and G4s were used to affect the functions of the elements. We first analysed
264 their ability to promote transcription, to recruit the pre-initiation complex and to
265 assemble chromatin. Our smRNA FISH of MS2 reporter results show that both G4 and
266 TATA elements are required for full activity (Figure 6B). We found that pG4 mutation
267 display more pronounced effects (~60%) while less impact of the TATAbox was
268 observed and an additive or synergic effect seems at play in the double mutant.

269 To infer how the G4 and TATAbox influence promoter activity and transcription, we
270 then analysed nucleosome organization and PIC assembly at these promoters in bulk
271 assays. MNase assay coupled to PCR analyses revealed that mutating the G4
272 sequence led to increased nucleosome density at the level of the NDR location, that
273 also corresponds to the G4 structure coordinate (Figure 6C). Remarkably, this effect
274 was not observed in the TATA box mutant. The double mutant showed synergic effects
275 of the elements, consistent with nascent transcription data. These results suggest that
276 the pG4 is the main element controlling Eef1a1 promoter opening and that the
277 TATAbox influence could only be seen in the TATAmutG4mut context. We next
278 monitored PIC recruitment using Pol II and TBP ChIP qPCR assays. Interestingly,
279 while Pol II recruitment was reduced in all mutants, TBP binding was only impaired
280 when the TATAbox was mutated but not in the G4mut (Figure 6D). We conclude that
281 the primary effect of the pG4 mutation is to restrict chromatin and Pol II accessibility
282 but not TBP recruitment, while the TATAbox mutation affects TBP and ultimately Pol
283 II recruitment (Figure 6A-D).

284 Next, we analysed the Eef1a1 promoter dynamics. We examined transcription in living
285 single cell by imaging an MCP-GFP fusion protein. We recorded long movies on the
286 WT, TATAmut and G4mut Eef1a1 cell lines, by taking 3D image stacks every 3 min
287 during 8 h (see representative examples in Figure S6G). Since a single polymerase
288 remains more than 3 min at the reporter transcription sites (Tantale et al., 2016a), and
289 since the sensitivity of this assay allows single polymerase tracking, every single
290 transcription event can be detected. We quantified the brightness of transcription sites
291 in hundreds of cells (see methods) and quantified permissive (ON) periods, from which
292 Pol II regularly initiate transcription, from inactive (OFF) periods (Figure 6E). These
293 analyses clearly show that both TATA and G4 mutants displayed shorter ON and
294 longer OFF periods. Moreover, OFF periods were slightly longer for the G4 mutant as
295 compared to TATAmut (Figure 6E-F). To complete this picture, we also recorded short
296 movies at high temporal resolution (one stack every 3 seconds), to model the entire
297 dynamics of the Eef1a1 promoter (Figure 6G). By performing mathematical modelling
298 of the distribution of initiation events and interpolating the signal intensities (Figure
299 S6H) (Tantale et al., 2021), we were able to show that the WT promoter is described
300 by a two states promoter model (ON and OFF). In contrast, both G4 and TATA mutants
301 require three states model to describe the experimental data, with the additional
302 promoter state being a long-lived inactive state (~2h lifetime; Figure S6H-I). Given the
303 results of the biochemical analysis of the mutant promoters, the additional state likely
304 corresponds to a TBP unbound state for the TATAmut, and a closed chromatin state
305 for G4mut. Moreover, examination of transcription initiation rates derived from the
306 models (k_3), indicates a slower transition into processive elongation for the G4 mutant
307 (0.3 vs 0.15 seconds). These live cell kinetic data are consistent with the idea that the
308 G4 mutations severely limit chromatin opening. In addition, changes in the k_3 constant
309 (Figure 6G, S6I) suggest that G4 mutations also slow down promoter escape and/or
310 Pol II pause release (Figure 6H).

311 Our observations on the model Eef1a1 promoter (summarized in Figure 6H) lead us to
312 propose that the role of G4s at promoters is to promote and maintain nucleosome
313 exclusion as a prerequisite for stable PIC recruitment. In contrast, when the TATAbox
314 is mutated, formation of an active promoter state is also impacted due to defects in
315 TBP and PIC recruitment. Collectively, our reporter experiments in bulk and single cells
316 further support the role of G4s/pG4s as promoter elements, conditioning nucleosome
317 exclusion and the rates of promoter transition toward an active state competent for
318 transcription.

319 **G4 stabilization results in Pol II promoter proximal pause release**

320 To investigate the influence of G4 stabilization on transcription at the global, we treated
321 human Raji cells with Pyridostatin (PDS), a well-known G4 ligand (Rodriguez et al.,
322 2008). This ligand stabilizes G4's structure by limiting the transition from G4-structured
323 to unstructured ssDNA or ds B-DNA (Rodriguez et al., 2008). Therefore, and since
324 promoters are highly enriched in pG4s, their stabilization could either positively or

325 negatively impact transcription. To address this question, we monitored PDS effects
326 on Pol II densities and nascent transcripts, using ChIP-seq and chrRNA-seq,
327 respectively. We used short time points of treatment (10 to 60 min) to avoid indirect
328 effects (Olivieri et al., 2020; Rodriguez et al., 2008) resulting from the appearance of
329 double strand breaks at later time points (Figure S7A). In these assays, we observed
330 that PDS led to changes in Pol II profiles after only 10 minutes of treatment, with
331 increased signal in gene bodies while signal at promoters was decreased. Examples
332 in Figure 7A illustrate this effect for model genes in which Pol II release is illustrated
333 by either promoter density decrease, gene body increase or both (the sequence of the
334 G4 upstream of TSS is indicated). These effects most likely reflect a general promoter
335 proximal pause release of Pol II. We confirmed a global decrease of Pol II pausing by
336 computing apparent pausing scores at the genome-wide level (Figure 7B and Figure
337 S7B). The effect was found more pronounced for a subset of 556 genes (see methods).
338 Interestingly these genes tend to display slightly less stable G4s suggesting that the
339 ligand preferentially act on weaker structure (Figure 7C) in agreement with our recent
340 observation using the G4access procedure (Garcia-Oliver et al., 2022). This selection
341 also appeared as enriched in mRNA splicing functions (Figure S7C). Consistent with
342 pause release, Pol II average profiles at this subset show a decrease around TSSs
343 and an increase of Pol II density over gene bodies (Figure 7D-E). This was further
344 confirmed by nascent chrRNA-seq analysis (Figure 7F-G). We also analysed later time
345 points, showing that, although reduced pausing is still visible at 30 min, the impact of
346 PDS is partially reversed after 60 min over gene bodies (Figure S7B, middle panel).
347 Altogether, our observations suggest that G4 stabilization results in increased ability
348 for Pol II to escape from pause states. This is also consistent with our mathematical
349 modelling of the *eef1a1* model promoter where G4 mutations affected efficiency of Pol
350 II released from the promoter.

351 Discussion

352 Our study has shown that G4 forming sequences are highly enriched in extended
353 mammalian core promoters. We propose a novel function for these elements that is to
354 intrinsically exclude nucleosomes, possibly defining one essential property of
355 promoters *in vivo*. The mechanism of nucleosome exclusion by G4s could be simply
356 explained by the incompatibility of stable single-strand DNA (ssDNA) formation and its
357 incorporation into stable nucleosomes. Since G4s are not significantly present in all
358 eukaryotic promoters (Marsico et al., 2019), other secondary structures or sequence
359 context could have the same role in other organisms, for example AT stretches in yeast
360 (Kaplan et al., 2010; Segal and Widom, 2009). This G4 property is however in contrast
361 to previous observation proposing that it is transcription that stimulates G4 formation
362 based on *in vitro* transcription (Xia et al., 2018) assays or transcription activation *in*
363 *vivo* (Hansel-Hertsch et al., 2016) but in line with more recent observations (Shen et
364 al., 2021), including work from our laboratory (Garcia-Oliver et al., 2022). In this work
365 (Garcia-Oliver et al., 2022), we show that transcription inhibition results in
366 maintenance, but reduction of G4 signal indicating that transcription does not precede

367 G4 formation at promoter but that it does increase or stabilize its structure. Taken
368 altogether, our data plead for a model in which G4s are formed prior from/or in the
369 absence of transcription since pG4s/G4s are detected at NDRs *in vitro* or at silent
370 promoters and since transcription inhibition does not significantly alter NDRs at pG4
371 locations. The presence of frequent transcription factor binding site (TFBS) motifs in
372 addition to pG4 elements also opens the possibility that PIC recruitment *in vivo* does
373 not directly rely on motifs that recruit the GTFs but rather on various already bound
374 TFs, possibly in combination to co-activator, to allow further recruitment of the PIC.
375 Arguing also for this model is the observation that TFIID does not contact DNA on its
376 own upstream of TSSs on TATA-less promoters (Burke and Kadonaga, 1996; Parry et
377 al., 2010), which represent the vast majority of promoters. Furthermore, TFIID
378 recruitment was previously reported to occur through direct chromatin contacts
379 (Bhuiyan and Timmers, 2019; Muller and Tora, 2014; Vermeulen et al., 2007) or via
380 interaction of TF such as NF-Y (Frontini et al., 2002) that we also find to be a major
381 TFBS enriched at promoters, consistent with previous observations (Oldfield et al.,
382 2019).

383 Our data are consistent with a role of G4s in favouring pause release by Pol II. First,
384 modelling transcription at the Eef1a1 G4 mutant promoter indicates one additional
385 limiting step for transcription (OFF2) likely corresponding to chromatin opening in the
386 absence of G4. Another limiting step includes transition to productive elongation, which
387 comprises pausing, also show a significant time increase (k3, Figure 6H). Second,
388 stabilizing G4s globally with PDS *in vivo* resulted in pause release of genes containing
389 weaker G4s in their promoter (Figure 7C). These results are in apparent contrast with
390 recent work indicating that treatment with another G4 ligand reduces transcription
391 initiation (Li et al., 2021). However, the time of treatments performed in this study were
392 much longer, opening the possibility that more indirect effects might have come into
393 play and also allowing accumulation of DNA breaks. Overall, our data are in agreement
394 with previous work having shown that G4-containing promoters tend to show less
395 poised Pol II (Dao et al., 2016). How could thus G4s facilitate pause release? Because
396 G4s at promoter exclude nucleosome, the presence of one or multiple G4s would
397 favour not only open chromatin but also a pre-melted template for Pol II. Such
398 structures would potentially facilitate the formation and the extension of the
399 transcription bubble. As a consequence, crossing the +1 nucleosomal barrier would
400 become easier for the Pol II complex, resulting in pause release. The comparison of
401 experimental G4 signal (G4access) to ssDNA scored by KMnO4 footprinting supports
402 this hypothesis since it indicates a correlation between G4 formation and open complex
403 at active promoters (Figure S2E).

404 Another striking property of G4 forming sequences, also visible at experimental G4s
405 determined by G4access, is their ability to position nucleosomes. The nucleosome
406 repeat length observed is in the range of that described for CTCF but also of strongly
407 positioning nucleosome *in vitro* (Valouev et al., 2011). Because G4s tend to be present
408 as clusters in promoters, at these locations, nucleosome positioning is less visible

409 probably due to the presence of multiple G4 introducing fuzziness in the adjacent
410 nucleosomes and to their additive effects. We do not know at this stage if the
411 positioning property in IGRs is inherent to G4 structure or could be explained by the
412 recruitment of G4 binders. In any of these scenarios, the barrier constituted by the G4s
413 would dock the surrounding nucleosomes. It will thus be interesting in future studies to
414 investigate the precise interplay between G4 and CTCF since recent work suggest
415 they could be locally associated (Tikhonova et al., 2021).

416 Our data point to a role of G4 as driver of CGI's properties, possibly because they yield
417 a more robust and/or constitutive NDR. Over 70% of G4access or G4 ChIP (Mao et
418 al., 2018) enriched areas are actually located in CGIs and consistently, CGIs also
419 contain large G4 clusters, increasing the likelihood of their formation locally. Our
420 analyses suggest that pG4s could be one essential determinant of the ability of CGIs
421 to exclude nucleosomes, thus adding a novel determinant, besides GC and CpG
422 content (Deaton and Bird, 2011), of these essential areas of the genome. Importantly,
423 while our study shows that G4 forming sequences behave as promoter elements by
424 excluding nucleosome, it does not demonstrate per se that the G4 structures are
425 formed in situ, in the context of chromatin. Nevertheless, the use of various orthogonal
426 techniques to score for pG4s based on different principles and the pG4 ability to
427 exclude nucleosome intrinsically pleads for their direct structural involvement as
428 promoter element rather than protein docking sites on DNA. All in all, our work opens
429 a new gate in our understanding and definition of a promoter *in vivo* and readjusts the
430 existing paradigms. It will also support future work on targeting secondary structures
431 to control their activity using specific ligands in cancer therapy.

432 **Methods**

433 **Cell lines and culture**

434 Data presented in this article were issued from the analysis of human cell lines (K562,
435 Raji, HeLa) or mouse primary thymocytes (CD⁺ CD8⁺ (DP)). Original data presented
436 concern essentially Raji and HeLa cells but all cellular models are described in this
437 section.

438 K562 is a pseudotriploid ENCODE Tier I erythroleukemia cell line derived from a
439 female (age 53) with chronic myelogenous leukemia. The Raji cells are lymphoblast-
440 like cells from a male (age 11) with Burkitt's lymphoma. HeLa Flp-in H9 cells (a kind
441 gift of S. Emiliani) is a cell line derived from the parent HeLa line. The HeLa line is
442 derived from a female (age 31) with adenocarcinoma. Mouse CD4⁺ CD8⁺ DP cells
443 were sorted from thymuses of 5 to 6 weeks old mice as described (Fenouil et al., 2012;
444 Koch et al., 2011).

445 Raji cells were grown in RPMI 1640 medium supplemented with 10% fetal calf serum,
446 penicillin/streptomycin (100 units/L) and glutamin (2 mg/L) at 37°C and 5% CO₂. For
447 the α -amanitin experiments (ED Figure6A), cells were treated with 2.5 μ g/L at the

448 indicated times as described(Fenouil et al., 2012). For the PDS experiments (Figure 7
449 and Figure S7), cells were treated with 10 μ M PDS at the indicated times.

450 HeLa Flp-in H9 cells used for reporter assays (Figure 5-6, Figure S6) were maintained
451 in DMEM supplemented with 10% fetal calf serum, penicillin/streptomycin (100 units/L)
452 and glutamin (2.9 mg/L) at 37°C and 5% CO₂. HeLa cells with integrated constructs
453 were transfected with plasmids using JetPrime (Polyplus), following manufacturer
454 recommendations.

455 **Genome-wide data sets**

456 All data sets used in this study including published and original experiments are
457 described in Table S1. All GEO accession numbers are included. The GEO accessions
458 for specific experiments related to this study are recorded under GSE52914.

459 **MNase and MNase-seq**

460 For sequencing of nucleosomal DNA in Raji cells, 3.5×10^7 cells were resuspended in
461 350 μ l Solution I (150 mM sucrose, 80 mM KCl, 5 mM K₂HPO₄, 5 mM MgCl₂, 0.5 mM
462 CaCl₂, 35 mM HEPES pH 7.4) and NP40 was added to a final concentration of 0.2%.
463 Cell membranes were permeabilized for 5 min at 37°C. MNase was prepared at 50,
464 25, 12, 6 or 3 units in 0.5 mL of Solution II (150 mM sucrose, 50 mM Tris pH 8, 50 mM
465 NaCl, 2 mM CaCl₂) and incubated with 50 μ L of cellular preparation, corresponding to
466 5×10^6 cells, for exactly 10 min at 37°C. The reactions were stopped by adding EDTA
467 to a final concentration of 10 mM. The cells were lysed using 1.45 mL of SDS Lysis
468 Buffer (1% SDS, 10 mM EDTA pH 8, 50 mM Tris pH 8), with 10 min incubation at 4°C.
469 200 μ l aliquots were taken for purification and the remaining extracts were stored at -
470 80°C. An equal volume of TE (200 μ l) was added to the aliquots, followed by
471 subsequent 2h treatments with 0.2 μ g/mL of RNase A and Proteinase K at 37°C and
472 55°C, respectively. DNA was extracted by two subsequent
473 phenol:chloroform:isoamylalcohol (25:24:1) extractions, further purified using
474 QIAquick PCR purifications columns (Qiagen, Germany). Nucleosomal digestion was
475 verified by running 500ng of DNA on a 1.5% agarose gel as well as on DNA high-
476 sensitivity 2100 Bioanalyzer chips (Agilent, USA). Digestions showing 75% of
477 mononucleomes (running at 150bp) were selected for library preparations. Fragments
478 below 250 bp were purified with Ampure XP Beads (Beckman Coulter, USA) following
479 manufacturer instructions. Libraries were prepared with TruSeq ChIP Library
480 Preparation Kit (illumina) and sequenced on Hiseq 2000 or 4000 sequencers
481 (Illumina).

482 Chromatin analysis by MNase treatment on HeLa cells (Figure 6C) were performed as
483 follows, since adherent cells harvested with trypsin tend to clamp using the method
484 described above. HeLa cells were harvested using trypsin and washed twice with ice-
485 cold PBS. Cells were resuspended in 250 μ L of ice-cold Nuclei buffer I (15 mM Tris-
486 HCl pH7.5, 300 mM sucrose, 60 mM KCl, 15 mM NaCl, 5 mM MgCl₂, 0.1 mM EGTA,

487 0.5 mM DTT, 0.1 mM PMSF, 3.6 µg/mL aprotinin) before addition of 250 µL of ice-cold
488 Nuclei buffer II (15 mM Tris-HCl pH7.5, 300 mM sucrose, 60 mM KCl, 15 mM NaCl, 5
489 mM MgCl₂, 0.1 mM EGTA, 0.5 mM DTT, 0.1 mM PMSF, 3.6 µg/mL aprotinin, 0.4%
490 IGEPAL CA-630). Extracts were incubated 10 min on ice and layered on 1 mL Nuclei
491 buffer III - sucrose cushion (15 mM Tris-HCl pH7.5, 1.2 M sucrose, 60 mM KCl, 15 mM
492 NaCl, 5 mM MgCl₂, 0.1 mM EGTA, 0.5 mM DTT, 0.1 mM PMSF, 3.6 µg/mL aprotinin).
493 Nuclei were isolated by centrifugation at 10,000 g for 20 min at 4°C, and were
494 resuspended in 600 µL of MNase digestion buffer (50 mM Tris-HCl pH7.5, 320 mM
495 sucrose, 4 mM MgCl₂, 1 mM CaCl₂, 0.1 mM PMSF) and incubated on ice for 3 min.
496 MNase was added for exactly 10 min at 37°C, using 50, 25, 12, 6 or 3 units of the
497 enzyme. The reactions were then stopped by adding EDTA to a final concentration of
498 10 mM. 100 µL of SDS Lysis Buffer were added (1% SDS, 10 mM EDTA pH 8, 50 mM
499 Tris pH 8) and after 10 min of incubation at 4°C, samples were processed as previously
500 described for DNA purification. qPCR quantifications were performed by using the
501 primers described in Table S4.

502 **G4access**

503 The complete G4access procedure is described in(Garcia-Oliver et al., 2022) and the
504 principle of the method is summarized in Figure S3a. In short, K562 cells were pelleted
505 and rinsed twice in phosphate-buffered saline buffer (PBS). For each experiment,
506 5x10⁶ cells per titration point were re-suspended in 50 µL of prewarmed
507 permeabilization buffer (150 mM of sucrose, 80 mM KCl, 5 mM KH₂PO₄, 5 mM MgCl₂,
508 0.5 mM CaCl₂ and 35 mM HEPES pH 7.4) supplemented with 0.2% (v/v) NP40 and
509 incubated for 5 minutes at 37°C prior digestion. MNase digestions, were then
510 performed by adding a volume of 500 µL of prewarmed MNase reaction buffer (150
511 mM sucrose, 50 mM Tris-HCl pH 8, 50 mM NaCl and 2 mM CaCl₂) supplemented with
512 either 3, 6, 12, 25 or 50U of MNase (Merck, 10107921001). Digestions were incubated
513 at 37°C for 10 min and stopped on ice and by adding 11 µL of 500 mM EDTA to each
514 reaction. Samples were then incubated 10 minutes on ice with 550 µL of SDS lysis
515 buffer (1% (v/v) SDS, 10 mM EDTA and 50 mM Tris.HCl pH 8). Before DNA
516 purification, 1 mL of water was added to dilute the SDS and the samples were
517 incubated with 5 µL of RNase A (ThermoFisher, EN0531) at 37 °C for 2 hours and with
518 8 µL of proteinase K (Euromedex, 09-0911) at 56 °C for 2 hours to complete the lysis.
519 To then quality control the MNase digestions: 125 µL of each sample were cleaned-up
520 using QIAquick PCR Purification Kit (QIAGEN, 28106) and assessed by agarose gel
521 and Bioanalyzer. At this step, for efficient G4access, samples should present ~30%
522 (+/-5%) of mono-nucleosomes. Importantly, this assessment should be performed on
523 purified DNA that does not contain the subnucleosomal fraction, using a bioanalyzer
524 equipment. The remaining of the samples was then purified by phenol-chloroform and
525 ethanol precipitation for subsequent steps. We recommend that, when implementing
526 this method, a wide range of MNase concentrations shall be tested in a first round of
527 preparative experiments to narrow down the condition in which the critical fraction of

528 30% of mononucleosome shall be obtained. Our experiences showed this fraction is
529 on average optimal for best G4 sequence recovery.

530 The 0-100 bp size-selected fragments from MNase digestions that have ~30% of
531 mono-nucleosomes were subjected to DNA library preparation. In parallel, genomic
532 DNA libraries were sonicated by Bioruptor[®] Pico sonicator (Diagenode) to obtain DNA
533 fragments of ~150 bp to be used later as reference data sets for bioinformatic analyses.
534 Paired-end libraries were constructed using NEBNext[®] Ultra[™] II DNA Library Prep Kit
535 for Illumina (New England Biolabs, E7645S) using a starting material of 50 ng. DNA
536 fragments were treated with end-repair, A-tailing and ligation of Illumina-compatible
537 adapters. Clean-up of adaptor-ligated DNA was performed by using CleanNGS beads
538 (CNGS-0050) with a bead:DNA ratio of 2:1. The purified products were amplified with
539 8 cycles of PCR. Finally, samples were cleaned up with a bead:DNA ratio of 0.8:1 to
540 remove the free sequencing adapters. Libraries were sequenced on the Illumina
541 NextSeq-500 Sequencer using paired 50-30 bp reads. The G4access data is deposited
542 to GEO database under GSE31755.

543 **ChIP-seq and ChIP qPCR**

544 Fifty million cells were used to perform each Pol II ChIP-seq experiment. Cells were
545 crosslinked for 10 min at 20°C with the crosslinking solution (10 mM NaCl, 0.1 mM
546 EDTA pH 8, 0.05 mM EGTA pH 8, 5 mM HEPES pH 7.8 and 1% formaldehyde). The
547 reaction was stopped by adding glycine to reach a final concentration of 250 mM. After
548 5 min of formaldehyde quenching, cells were washed twice with cold PBS and
549 resuspended in cold 2.5mL LB1 (50 mM HEPES pH 7.5, 140 mM NaCl, 1 mM EDTA
550 pH 8, 10% glycerol, 0.75% NP-40, 0.25% Triton X-100) at 4°C for 20 min on a rotating
551 wheel. Nuclei were pelleted down by spinning at 1350 rcf in a refrigerated centrifuge
552 and washed in 2.5mL LB2 (200 mM NaCl, 1 mM EDTA pH 8, 0.5 mM EGTA pH 8, 10
553 mM Tris pH 8) for 10 min at 4°C on a rotating wheel followed by centrifugation to collect
554 nuclei. Nuclei were then resuspended in 1mL LB3 (1 mM EDTA pH 8, 0.5 mM EGTA
555 pH 8, 10 mM Tris pH 8, 100 mM NaCl, 0.1% Na-Deoxycholate, 0.5% N-
556 lauroylsarcosine) and sonicated using Bioruptor Pico (Diagenode) in 15mL tubes for
557 20 cycles of 30 s ON and 30 s OFF pulses in 4°C bath. All buffers (LB1, LB2 and LB3)
558 were complemented with EDTA free Protease inhibitor cocktail (Roche), 0.2 mM PMSF
559 and 1 µg/mL Pepstatin just before use. After sonication, Triton X-100 was added to a
560 final concentration of 1% followed by centrifugation at 20000 rcf and 4°C for 10 min to
561 remove particulate matter. After taking aside a 50 µl aliquot to serve as input and to
562 analyze fragmentation, chromatin was aliquoted and snap-frozen in liquid nitrogen and
563 stored at - 80°C until use in ChIP assays. Input aliquots were mixed with an equal
564 volume of 2X elution buffer (100 mM Tris pH 8.0, 20 mM EDTA, 2% SDS) and
565 incubated at 65°C for 12 hours for reverse-crosslinking. An equal volume of TE buffer
566 (10 mM Tris pH 8 and 1 mM EDTA pH 8) was added to dilute the SDS to 0.5% followed
567 by treatment with RNase A (0.2µg/mL) at 37°C for one hour and Proteinase K (0.2
568 µg/L) for two hours at 55°C. DNA was isolated by phenol:chloroform:isoamylalcohol

569 (25:24:1 pH 8) extraction followed by Qiaquick PCR Purification (QIAGEN, Germany).
570 Purified DNA was then analyzed on a 1.5% agarose gel and on Bioanalyzer (Agilent,
571 USA) using a High Sensitivity DNA Assay.

572 For Pol II ChIP, Protein-G coated Dynabeads were incubated at 4°C in blocking
573 solution (0.5% BSA in PBS) carrying Pol II N20 (Santa-Cruz sc-899x, lot H3115) and
574 TBP N12 (Santa-Cruz sc-204, lot LO214) specific antibodies. Sonicated chromatin
575 (1mL) was added to pre-coated beads (250µL) and the mix was incubated overnight
576 at 4°C on a rotating wheel. After incubation with chromatin, beads were washed 7 times
577 with Wash buffer (50 mM HEPES pH 7.6, 500 mM LiCl, 1 mM EDTA pH 8, 1% NP-40,
578 0.7% Na-Deoxycholate, 1X protease inhibitor cocktail) followed by one wash with TE-
579 NaCl buffer (10 mM Tris pH 8 and 1 mM EDTA pH 8, 50 mM NaCl) and a final wash
580 with TE buffer (10 mM Tris pH 8 and 1 mM EDTA pH 8). Immunoprecipitated chromatin
581 was eluted by two sequential incubations with 50 µL Elution buffer (50 mM Tris pH 8,
582 10 mM EDTA pH 8, 1% SDS) at 65°C for 15 min. The two eluates were pooled and
583 incubated at 65°C for 12 hours to reverse-crosslink the chromatin followed by
584 treatment with RNase A and Proteinase K and purification of DNA, as described above
585 for input samples. Both input and ChIP samples were subjected to Bioanalyzer
586 analysis to check that the major bulk of isolated DNA was in the 250 bp size range.

587 Samples were analyzed by qPCR (Stratagene) in HeLa cells following the
588 manufacturer recommendations. Oligonucleotides pairs used for qPCR in this study
589 are presented in the Table S4. For ChIP-seq experiments in Raji cells, purified DNA
590 was quantified with Qubit DS DNA HS Assay (ThermoFisher Scientific, USA). Five ng
591 of ChIP DNA were used to prepare sequencing libraries with Illumina ChIP Sample
592 Library Prep Kit (Illumina, USA). After end-repair and adaptor ligation, library fragments
593 were amplified by 12 cycles of PCR. Barcoded libraries from different samples were
594 pooled together and sequenced on Illumina HiSeq2000 platform in paired-end
595 sequencing runs.

596 **Chromatin RNA sequencing (chrRNA-seq)**

597 Chromatin associated RNAs (ChrRNAs) were isolated from 2×10^7 Raji cells before and
598 after 10 min of PDS treatment (Figure 7F-G) as described previously (Nojima et al.,
599 2015) followed by TurboDNase treatment. Purified RNAs were quantified by Qubit and
600 quality was assessed using RNA Pico Assay kit with Bioanalyzer (Agilent
601 Technologies, USA). chrRNA were then subjected to library preparation using True-
602 seq stranded total RNA library prep gold kit (Ref#220599) from Illumina using 1 µg of
603 chrRNA, 15 cycles of amplification and following manufacturer instructions (including
604 ribo-depletion). The data is submitted to GEO database together with the
605 manuscript(Garcia-Oliver et al., 2022) (Table S1).

606 **Bioinformatics**

607 *Motif analysis*

608 Canonical promoter elements, *de novo* and known motifs (TRANSFAC) were analyzed
609 across all expressed genes in K562, Raji and DP cells (Figure1 and FigureS1). *De*
610 *novo* motif discovery and known motif identification were performed using MEME and
611 DREME (Bailey, 2011; Bailey et al., 2009) using fragments from -100 to +20 bp of
612 experimental TSSs since this area encompasses not only the promoter but also the
613 majority of the NDR. Enrichment of canonical promoter elements were tested using
614 bedtools 'intersect' against all promoters of expressed genes (-100 to +20 bp of
615 experimental TSSs) and against 10,000 permutations of random genomic areas of
616 121bp. Random controls were generated using bedtools 'random' using or not GC
617 constraints. GC thresholds were determined to fit exactly the GC biased observed at
618 promoters of expressed genes. Motifs used for this analysis (Table S2) were the BRE,
619 the canonical or non-canonical TATAboxes as indicated in ED Figure1. Additionally,
620 we also used the quadparser QP1-7 ((G_{n>2}N1-7)x4) and bed files generated by the
621 G4Hunter algorithm G4H2.0 or G41.5 using a window of 25 bp as described (Bedrat
622 et al., 2016).

623 *G-quadruplex predictions*

624 G-quadruplex predictions were performed using the G4Hunter(Bedrat et al., 2016)
625 algorithm. Predicted G-quadruplexes (pG4s) at stringencies 1.52 and 2.0 were used
626 throughout this study. Previous experiments have shown that these G4Hunter
627 thresholds allow to experimentally confirm 92 and 100% of the predicted G-quadruplex
628 structures(Bedrat et al., 2016).

629 *ChIP-seq, ChIP-exo and MNase-seq analyses*

630 All genomic experiments from this study or re-analyzed from available datasets were
631 processed using our pipeline. Sequencing files were analysed using
632 Bowtie2(Langmead and Salzberg, 2012) and PASHA(Fenouil et al., 2016). Raw
633 sequencing reads were aligned to human Hg19 or mouse genome (mm9) using
634 Bowtie2. Duplicate reads with identical coordinates (sequencing depth taken into
635 account) to remove potential sequencing and alignment artifacts. For ChIP-seq and
636 MNase-seq (nucleosome density) signal analyses, aligned reads were elongated *in*
637 *silico* using the DNA fragment size inferred from paired-reads or an estimated optimal
638 fragment size for orphan reads using Pasha R package. These elongated reads were
639 then used to calculate the number of fragments that overlapped at a given nucleotide
640 thus representing an enrichment score for each bin in the genome. For nucleosome
641 positioning analyses (midpoints) presented in Figure 3, to determine the average
642 nucleosome positions, wiggle files representing the central nucleotides of DNA
643 fragments were also generated. For ChIP-exo, the nucleotide located at the 5'
644 extremity of the DNA fragments was considered to generate wiggle files, since it

645 represents the exact points where the nucleases have stopped. Wiggle files
646 representing average enrichment score every 50bp or 10bp were generated.
647 Sequencing data from Input samples were treated in the same way to generate Input
648 wiggle files. All wiggle files were then rescaled to normalize the enrichment scores to
649 reads per million. For ChIP-seq datasets, enrichment scores from input sample wiggle
650 files were subtracted from ChIP sample wiggle files. This allows removing/reducing the
651 over-representation of certain genomic regions due to biased sonication, local
652 duplications, and DNA sequencing. Finally, for MNase-seq, we smoothed the signal by
653 replacing each 10bp bin by the average of the 5 surrounding bins on each side.

654 *RNA-seq analysis*

655 All RNA-seq datasets re-analyzed in this study were processed using our in house
656 pipeline. Raw sequencing reads were aligned to mouse genome (mm9) or human
657 genome (hg19) using TopHat2(Kim et al., 2013). Alignment files were then treated
658 using PASHA(Fenouil et al., 2016) to generate wiggle files. In Raji and DP cells,
659 experimental TSSs were determined as the summit in short-RNA-seq signals in a
660 window of 300 bp of annotated TSSs.

661 *Average binding profiles and heatmaps*

662 To generate average binding profiles (Figure 1-4, and Figure S2-5), R scripts were
663 developed and used for retrieving bin scores in defined regions from 10 or 50 bp bin
664 sized wiggle files(Fenouil et al., 2016). Heatmaps were generated, viewed and color-
665 scaled according to sample read depth using Java TreeView(Saldanha, 2004).
666 Regions were defined either as centered on experimental TSSs (see above), on the
667 center of predicted G4 from G4Hunter or the center of the area if no G4 was predicted.
668 In addition, pG4s that were not located in annotated gene features or further than 200
669 bp from annotated TSSs were considered as intergenic.

670 To generate average binding profiles of Pol II and of chrRNA (Figure 7D-G), hg19
671 Refseq genes annotations were used to extract values from wiggle files associated
672 with the selected genes. Bin scores inside these annotations and in a region of 5kb
673 before the TSSs and after 5kb of annotated termination sites were determined. Based
674 on the gene list selections, bin scores from wiggle files were used to re-scale values
675 between TSSs and transcription termination sites (gene body) of all genes using linear
676 interpolation. In total, 1000 points were interpolated for the gene body of each selected
677 gene in all average profiles presented.

678 *Identification of inactive promoters*

679 To identify inactive promoters, we selected the bottom 30% of genes of Pol II signal
680 over the defined areas (Figure 2C, Figure S2I). Hg19 Refseq genes annotations were
681 used to extract values from wiggle files associated with the selected genes and bin
682 scores in a region of 2kb before and after the TSSs were determined.

683 *Nucleosome arrays:*

684 We have performed a precise assessment of nucleosome repeat length (NRL) and
685 phasing, comparing pG4s and CTCF(Maurano et al., 2015) in K562 (Figure 3C), which
686 yielded an average NRL of 194 and 193 nt calculated over 5 nucleosomes from the
687 docking site and 215 and 216 nt over 10 nucleosomes, respectively.

688 *Pausing scores*

689 To analyze how the G-quadruplex ligand pyridostatin (PDS) impacts Pol II pausing
690 (Figure 7B and Figure S7B), we have determined pausing scores based on the ratio
691 of Pol II signals at promoters and in gene bodies(Adelman and Lis, 2012). Our
692 approach for pausing scores determination is comparable to the one previously
693 described(Fenouil et al., 2012) with modifications. It takes into account Pol II density
694 on either promoter regions (TSS) or gene bodies (GB). Promoter regions were
695 considered between -300 and +100 of TSS to define paused Pol II density for
696 calculations. Densities at genes bodies were analyzed in the intervals of 50-100% of
697 the length. The use of these intervals avoids detecting signal originating from the
698 promoters for short genes or genes with exceptionally large initiation areas and allows
699 detecting more significant signal of elongating Pol II. To avoid interferences between
700 promoter and gene body read counts, only genes larger than 3kb were considered.
701 Read count was performed using HTseq(Anders et al., 2015), normalized to the length
702 of the genomic regions and expressed as RPKM (reads per kb per millions). Only
703 genes with sufficient read coverage were considered (>75 RPKM at promoters and
704 >25 RPKM at gene bodies n = 7617). Pausing scores were expressed as the ratio
705 TSS/GB. To define a high confidence set of genes with pause release effect and since
706 in our datasets PDS globally affected Pol II pausing (ED Figure9b, linear regression
707 slope T0 versus T10 minutes = 0.71, Wilcoxon test <0.00001, n= 7617), we further
708 selected genes with significant Pol II signal increase in their gene bodies using DESEQ
709 (Pvalue< 0.05, n= 556; Pvalue ≥ 0.05, n=7061).

710 **Plasmids and cloning**

711 The repeats of the 256xMS2 binding sites were cloned from chemically-synthesized
712 oligonucleotides into pMK123(Alexander et al., 2010). The MS2 stem loops are
713 separated by a linker of only three nucleotides and cloned in a pIntro-MS2x256
714 plasmid, which also contained an FRT-Hygro cassette for Flp-in
715 recombination(Boireau et al., 2007; Tantale et al., 2016a). pUC57 containing the
716 Eef1a1 (1513bp) or Polr2a (711bp) WT mouse promoters were purchased at
717 genescript; pGL4.17 containing the Pkm (200bp), Klf6 (400bp) or Taok1 (300bp) or
718 mutagenized mouse promoters were purchased at genecust (Table S5). All construct
719 were then subcloned into pIntro-MS2x256 between SnaBI and MluI sites. To introduce
720 mutations into the largest promoters Polr2a and Eef1a1, smaller fragments (206 and
721 219 bp respectively) were purchased and subcloned to the full-length promoter

722 between NotI and NheI sites for Eef1a1 and between NotI and MluI for Polr2a.
723 Additionally, for a second version of Eef1a1 and Polr2a WT promoters and for
724 Eef1a1*inv*, Polr2aG4*mut2* and for Polr2aG4*mut3* promoters an additional luciferase
725 reporter was added downstream of the MS2 reporter using NEB assembly builder kit
726 and the two following oligonucleotides Gibintro-BsrG1-ires-fwd:
727 GGTTCCTCCAGTCACACCTCATGTACAGGCCCTCTCCCTCCCCCCC and Gibluc-
728 BsiW1-intro-Rev: :
729 TGTAAGTCATTGGTCTTAAACGTACGTCTAGAATTACACGGCGATC.

730 Stable expression of MCP-GFP was achieved by retroviral-mediated integration of a
731 self-inactivating vector containing an internal ubiquitin promoter. The MCP used
732 dimerizes in solution and contained the deltaFG deletion, the V29I mutation, and an
733 SV40 NLS24(Tantale et al., 2016a). MCP-GFP expressing cells were grown as pool
734 of clones and FACS-sorted to select cells expressing low levels of fluorescence.
735 Isogenic stable cell lines expressing the reporter genes were created using the Flp-In
736 system and a HeLa Flp-in H9 integrants were selected on hygromycin (150 µg/L). For
737 each construct, several individual clones were picked and analysed by *in situ*
738 hybridization. Clones usually looked similar, and two of them were further selected for
739 the experiments after PCR and sequencing check.

740 **Circular dichroism (CD) spectroscopy**

741 CD spectra were recorded on a Jasco J-815 spectropolarimeter equipped with a Peltier
742 temperature control accessory (JASCO Co., Ltd., Hachioji, Japan). Each spectrum was
743 obtained by averaging three scans at a speed of 100 nm/min. A background CD
744 spectrum of corresponding buffer solution was subtracted from the average scan for
745 each sample. The CD profile was monitored between 220 nm and 300 nm using quartz
746 cells of 5 mm path-length and a volume of 1000 µl.

747 **Absorbance spectroscopy**

748 All spectra were recorded on a Cary-300 (Agilent Technologies) spectrophotometer in
749 10 mM lithium cacodylate buffer (pH 7.2) at 3 or 4 µM oligonucleotide strand
750 concentration, in the presence or absence of 100 mM KCl.

751 Thermal difference spectra (TDS) were obtained by taking the difference between the
752 absorbance spectra of unfolded and folded oligonucleotides that were recorded at high
753 (95°C) and low (25°C) temperatures, respectively, in a buffer containing 100 mM KCl.
754 TDS provides specific signatures of different DNA structural conformations, provided
755 that the structure is not too heat-stable (a number of G4 structures do not melt at high
756 temperatures).

757 Isothermal difference spectra (IDS) were obtained as described previously(Renaud de
758 la Faverie et al., 2014) by taking the difference between the absorbance spectra from
759 unfolded and folded oligonucleotides. These spectra were recorded at 25°C before

760 and after potassium cation addition (100 mM KCl), respectively. IDS provide specific
761 signatures of different DNA structural conformations.

762 **Acquisition and analysis of smFISH images**

763 SmFISH was performed as previously described (Tantale et al., 2016a), with a mix of
764 10 fluorescent oligos hybridizing against the MS2x32 repeat, each oligo containing four
765 molecules of Cy3. Since each oligo bound eight times across the MS2x256 repeats,
766 each molecule of pre-mRNA hybridized with 80 oligos, thereby providing excellent
767 single molecule detection and signal-to-noise ratios.

768 To obtain the number of released, nucleoplasmic and nascent mRNA per cell, smFISH
769 images were recorded with an upright widefield Leica microscope as 3D image stacks
770 with a z-spacing of 0.3 μ M, with a x100 objective, and an Evolve 512x512 EMCCD
771 camera (Photometrics). The images were analyzed with FISH-quant (Mueller et al.,
772 2013) to count the number of pre-mRNA per nuclei, using populations of 400–500 cells
773 per experiment. To obtain the number of nascent pre-mRNA per cell, the transcription
774 sites (TS) were identified manually and isolated pre-mRNA molecules located in the
775 nucleoplasm were used to define the point spread function (PSF) and the total light
776 intensity of single molecules, which finally allowed determining the intensity of TS
777 expressed in number of full-length transcripts.

778 **Live-cell image acquisition**

779 Cells were plated on 25-mm diameter coverslips (0.17- mm thick). After 24-48 hours
780 the coverslips were mounted in the GFP-imaging medium (DMEM-GFP-2, Evrogen)
781 with rutin in a temperature-controlled chamber with CO₂ and imaged on an inverted
782 OMX Deltavision microscope in time-lapse mode. A x100, NA 1.4 objective was used,
783 with an intermediate x2 lens and an Evolve 512x512 EMCCD camera (Photometrics).
784 Stacks of 11 planes with a z-spacing of 0.6 μ m were acquired, with one stack collected
785 every 3 min for 8h.

786 **Quantification of short movies**

787 Short movies were analysed as previously described (Tantale et al., 2021; Tantale et
788 al., 2016a) (Figure 6G). In short, we manually defined the nuclear outline and the
789 region within which the transcription site (TS) is visible and stacks were corrected for
790 photobleaching using a fitted curve with a sum of three exponentials. This curve was
791 used to normalize each time-point such as nuclear intensities were equal to the
792 intensity of the first time-point. We then filtered the image with a 2-state Gaussian filter.
793 First, the image was convolved with a larger kernel to obtain a background image,
794 which was then subtracted from the original image before the quantification is
795 performed. Second, the background-subtracted image was smoothed with a smaller
796 Kernel, which enhances the SNR of single particles to facilitate spot pre-detection. TS
797 positions in each frame of the filtered images were determined as the brightest pixel

798 above a user-defined threshold in the pre-detected region of the TS. When no pixel
799 was above the threshold, the last known TS position was used. Then the TS signal
800 was fitted with a 3D Gaussian estimating its standard deviation σ_{xz} and σ_z , amplitude,
801 background, and position. We performed two rounds of fitting: in the first round all fitting
802 parameters were unconstrained. In the second round, the allowed range was restricted
803 for some parameters, to reduce large fluctuations in the estimates especially for the
804 frames with a dim or no detectable TS. More specifically, the σ_{xz} and σ_z were restricted
805 to the estimated median value \pm standard deviation from the frames where the TS
806 could be pre-detected, and the background was restricted to the median value. The
807 TS intensity was finally quantified by estimating the integrated intensity above
808 background expressed in arbitrary intensity units. With the live cell acquisition settings,
809 the illumination power was low and we could not reliably detect all individual molecules.
810 We therefore collected right after the end of the movies one 3D stack with increased
811 laser intensity (50% of max intensity, compared to 1% for the movie), which allowed
812 reliable detection of individual RNA molecules. We also collected slices with a smaller
813 z-spacing for a better quantification accuracy (21 slices every 300 nm). Quantification
814 of TS site intensity in the calibration stack was done with *FISH-quant* as follows: (a)
815 when calculating the averaged image of single RNA molecules, we subtracted the
816 estimated background from each cell to minimize the impact of the different
817 backgrounds; (b) when quantifying the TS in a given cell, we rescaled the average
818 image of single RNA molecules such that it had the same integrated intensity as the
819 molecules detected in the analyzed cell. To calibrate the TS intensities in the entire
820 movie, i.e. to express the TS intensity as a number of equivalent full-length transcripts,
821 we used the fact that the last movie frame was acquired at the same time as the
822 calibration stack. We then normalized the extracted TS intensity in the movies, I_{MS2} , to
823 get the nascent counts $N_{\text{nascent;calib}}$: $N_{\text{nascent;calib}}(t) = I_{MS2}(t) \times (N_{\text{nascent;final}} / I_{\text{final}})$, where
824 $N_{\text{nascent;final}}$ stands for the estimated number of nascent transcripts in the calibration stack
825 and I_{final} for the averaged intensity of the last four frames.

826 **Analysis of long movies**

827 To quantify long movies acquired at low frames rate (one 3D image stack every 3min),
828 we used ON-quant, a rapid analysis tool that identifies transcription sites, measures
829 their intensities, attributes the ON or OFF states of transcription, based on the defined
830 intensity threshold under which a TS is considered to be silent, and above which a TS
831 is considered to be active. The intensity threshold was defined based on the mean
832 intensity of single molecules (Tantale et al., 2016a).

833 **Mathematical modelling, short and long movies analysis**

834 Intermittent transcriptional activity of the promoters is modelled using a Markov
835 process with one active and multiple inactive epigenetic states. The number of states
836 and the transition rate parameters are obtained using the algorithms and pipeline first
837 described in (Tantale et al., 2021).

838 For the sake of consistency, we provide a short description of the pipeline. The cells
839 were imaged live for 30 minutes every 3 seconds (short movies), or for 8-9h every 3
840 minutes (long movies).

841 *Deconvolution and position of Pol II initiation events in short movies*

842 The Pol II positions were found by combining a genetic algorithm with a local
843 optimisation procedure. Before initiation of the analysis algorithm, several key
844 parameters were established. The Pol II elongation speed was fixed at 67 bp/s (Tantale
845 et al., 2021; Tantale et al., 2016b). The reporter construct transcript was divided into
846 three sections consisting of the pre-MS2 fragment (PRE=700 bp), 256xMS2 loops
847 (SEQ=5800 bp), and post-MS2 fragment until the polyA site (POST=1600 bp). An extra
848 time $P_{poly}=100s$ was added to POST, corresponding to the polyadenylation signal
849 (during this time the polymerase is past the polyA site and remains on the DNA,
850 see (Tantale et al., 2016b)). The frame rate of short movies is sufficient to detect
851 processes that occur on the order of seconds.

852 In order to find the positions of initiation events via the deconvolution pipeline, all
853 the possible initiation times were discretized using a step size of 0.45 seconds (or 30
854 bp at 67 bp/s). This step was chosen as it is smaller than the minimum polymerase
855 spacing and large enough to still accommodate a reasonable computation time. For a
856 movie of 30 min length this choice corresponds to a maximum number of 4020
857 positions. The deconvolution algorithm was implemented in Matlab R2020a using
858 Global Optimization and Parallel Computing Toolboxes for optimizing Pol II positions
859 in parallel for all nuclei in a collection of movies. Waiting times were then computed
860 from the position of each initiation event (Figure 6G).

861 *Long movies waiting time distribution*

862 For long movies, the low resolution (3 min) does not allow a precise positioning of
863 initiation events. In this case we binarize the signal by considering that the transcription
864 site is active or inactive if the measured intensity is above or below a threshold level,
865 respectively, which is set to be slightly higher than the intensity similar of a single
866 polymerase. The inactive intervals then indicate long waiting times between successive
867 polymerases. The active intervals are used to estimate the probability that waiting
868 times are larger than the movie frame rate (3 min), which is one of the parameters
869 needed for connecting long and short time distributions and obtain a multiscale
870 distribution (see (Tantale et al., 2021)).

871 *Multi-exponential regression fitting of the survival function and model reverse 872 engineering using the survival function*

873 Waiting times were extracted as differences between successive Pol II initiation events
874 from all the resulting traces and the corresponding data was used to estimate the
875 nonparametric cumulative short movie distribution function by the Meyer-Kaplan

876 method. Data from long movies is used to generate the nonparametric cumulative long
877 movie distribution function. The two distribution functions are fitted together into a
878 multiscale cumulative distribution function using the total probability theorem and
879 estimates of two parameters p_l and p_s , representing the probabilities that waiting times
880 are longer than the long movie frame rate, and longer than the length of the short
881 movie, respectively (see ref(Tantale et al., 2021) for details).

882 Then, a multi-exponential regression fitting of the multiscale distribution function
883 produces a set of $2N-1$ distribution parameters, where N is the number of exponentials
884 in the regression procedure (3 for $N=2$ and 5 for $N=3$). The regression procedure was
885 initiated with multiple log-uniformly distributed initial guesses and followed by local
886 gradient optimisation. It resulted in a best-fit solution with additional suboptimal
887 solutions (local optima with objective function value larger than the best fit).

888 The $2N-1$ distribution parameters can be computed from the $2N-1$ kinetic
889 parameters of a N state transcriptional bursting model. Conversely, a symbolic solution
890 for the inverse problem was obtained, allowing computation of the kinetic parameters
891 from the distribution parameters and reverse engineering of the transcriptional bursting
892 model. In particular, it is possible to know exactly when the inverse problem is well-
893 posed, i.e. when there is a unique solution in terms of kinetic parameters for any given
894 distribution parameters in a domain (Figure 6H and Figure S6I).

895 *Transcriptional bursting models*

896 The transcriptional bursting models used in this paper are as following:

897 For a promoter two-state model ($N=2$), the model corresponds to the well-known ON-
898 OFF telegraph model. In this case there are 3 distribution parameters and 3 transition
899 rates parameters.

900 The distribution parameters are $A_1, \lambda_1, \lambda_2$, defining the survival function

$$901 \quad S(t) = A_1 e^{\lambda_1 t} + (1 - A_1) e^{\lambda_2 t}.$$

902 These parameters are obtained by bi-exponential fit of the empirical survival function.

903 The transition rates parameters of the ON-OFF telegraph model can be obtained from
904 the distribution parameters using the formulas

$$905 \quad k_3 = -S_1, k'_2 = S_1 - \frac{S_2}{S_1}, k_2 = \frac{S_3 S_1 - S_2^2}{S_1 (S_1^2 - S_2)},$$

$$906 \quad S_1 = A_1 \lambda_1 + A_2 \lambda_2, S_2 = A_1 \lambda_1^2 + A_2 \lambda_2^2, S_3 = A_1 \lambda_1^3 + A_2 \lambda_2^3, A_2 = 1 - A_1,$$

907 where k_3, k_2, k_2' are the initiation rate, the OFF to ON and ON to OFF transition rates,
908 respectively.

909 For a promoter three-state model (N=3), there are 5 distribution parameters and 5
910 kinetic parameters.

911 The distribution parameters are $A_1, A_2, \lambda_1, \lambda_2, \lambda_3$, defining the survival function

$$912 \quad S(t) = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} + (1 - A_1 - A_2) e^{\lambda_3 t}.$$

913 The corresponding model represented in the Figure has two OFF and one ON state.

914 The five transition rate parameters can be obtained from the distribution parameters:

$$915 \quad k_3 = -S_1, k_2 = \frac{S_2^2 - S_1 S_3}{S_1(-S_1^2 + S_2)}, k_2' = S_1 - \frac{S_2}{S_1},$$

$$916 \quad k_1 = \frac{L_3(-S_1^2 + S_2)}{S_2^2 - S_1 S_3}, k_1' = \frac{A_1 A_2 A_3 S_1 (\lambda_1 - \lambda_2)^2 (\lambda_1 - \lambda_3)^2 (\lambda_2 - \lambda_3)^2}{(-S_1^2 + S_2)(S_2^2 - S_1 S_3)},$$

917 where

$$918 \quad S_1 = A_1 \lambda_1 + A_2 \lambda_2 + A_3 \lambda_3, S_2 = A_1 \lambda_1^2 + A_2 \lambda_2^2 + A_3 \lambda_3^2, S_3 = A_1 \lambda_1^3 + A_2 \lambda_2^3 + A_3 \lambda_3^3, A_3 =$$

919 $1 - A_1 - A_2,$

$$920 \quad L_1 = \lambda_1 + \lambda_2 + \lambda_3, L_2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2, L_3 = \lambda_1^3 + \lambda_2^3 + \lambda_3^3,$$

921 and $k_3, k_2, k_2', k_1, k_1'$ are the transcription initiation, OFF2 to ON, ON to OFF2, OFF1 to
922 OFF2, and OFF2 to OFF1 rates, respectively.

923 *Error intervals*

924 Distribution parameters result from multi-exponential regression fitting using gradient
925 methods with multiple initial data. These optimization methods provide a best fit (global
926 optimum) but also suboptimal parameter values. Using an overflow ratio (a number
927 larger than one, in our case 2) to restrict the number of suboptimal solutions, we define
928 boundaries of the error interval as the minimum and maximum parameter values
929 compatible with an objective function less than the best fit times the overflow.

930 *Choice of the number of exponentials*

931 The number of exponentials was determined by a parsimony principle: we have chosen
932 the smallest N that fits well. More precisely, starting with N=2, we have increased N
933 as long as the goodness of fit reduced without increase of overfitting. We have used
934 parametric uncertainty (error intervals) as a proxy for overfitting (Figure S6H).

935 **Code availability**

936 Softwares and codes are all publicly available and have been previously described
937 (Descostes et al., 2014; Fenouil et al., 2012; Fenouil et al., 2016).

938 **Data availability**

939 The GEO accessions for specific experiments related to this study are recorded under
940 GSE52914.

941 **Acknowledgements**

942 In the JCA lab, this work was supported by institutional grants from the CNRS including
943 an 80prime2021 ‘Deciph G4’, and a grant from “Agence Nationale de la Recherche”
944 (G4access, ANR-20-CE12-0023), ‘amorçage jeunes équipes’ Fondation pour la
945 Recherche Medicale FRM AJE20130728183 and INCA PLBIO20-225. CE was
946 supported by a grant from ARC (retour postdoc), EGO by a grant from EpiGenMed
947 labex of excellence. JLM was supported by Inserm, CNRS and Ecole Polytechnique.
948 We are grateful to Salvatore Spicuglia, Dan Fisher, Robert Feil and Mounia Lagha for
949 critical reading of the manuscript and to Florian Mueller for help with quantification of
950 live cells imaging data.

951 **Author contributions**

952 JCA and CE designed experiments; CE, EGO, TM, KG and AP performed the
953 experiments; CE and AZEA analyzed genomic data; CE, EB, MK, MCR and EB
954 performed and interpreted the microscopy experiments, YL, AC, DV and JLM
955 characterized and interpreted the G-quadruplexes *in vitro*, RM performed the eSNP
956 analysis. OR performed the mathematical modelling. JCA and CE conceived the
957 project, suggested and interpreted experiments, and wrote the article. All authors
958 reviewed the manuscript.

959 **Competing interests**

960 Authors declare no competing interests.

961

962 Figure legends

963 **Figure 1: Predicted G-quadruplexes (pG4s) are highly enriched at promoters and** 964 **correlate with maximum of mammalian promoter activity.**

- 965 (a) Promoter motif search around experimental TSSs highlights pG4 motifs in
966 various cells. (Top) experimental strategy to determine promoter elements
967 associated to experimental TSSs of expressed and annotated Refseq coding
968 genes in our 3 model cell types (K562, Raji cells and mouse primary T cells);
969 (bottom) motif distances to experimental TSSs, overlay of motif densities,
970 relative representation at promoters and enrichment over control sequences
971 are shown (INR= initiator). See also Figure S1 and Table S2 for detail of motif
972 discovery analyses in 3 independent human and mouse cell types.
- 973 (b) pG4s mark genomic areas comprised between sense and antisense
974 promoters. Metaprofiles at promoters show enrichment of pG4s (G4H2.0) at
975 56 bp upstream of sense TSSs (n =8346) and at 51 bp of antisense TSSs for
976 genes where divergent transcription initiation is detected (n=5689). See the
977 heatmaps of Figure S1 for Pol II and short chromatin RNA profiles. A green
978 arrow indicates the sense of transcription in each graph.
- 979 (c) G4access principle. Chromatin is digested at moderate MNase level at which
980 G4s show apparent resistance and are freed in subnucleosomal fractions.
981 These are purified and subsequently subjected to library preparation and
982 high-throughput sequencing. The details of the method are described in ref
983 (Garcia-Oliver et al., 2022).
- 984 (d) Example of a pG4 sequence fitting midpoint of upstream/downstream Pol II
985 and nascent transcription (chrRNA and GROcap) at the *BTG2* promoter,
986 experimental G4s (G4access and G4-ChIP), and maximum of promoter
987 activity (SURE assay) in K562 cells are displayed. Sequence of the pG4
988 (G4H2.0) is indicated below the gene.
- 989 (e) Experimental G4 signals (G4access) correlate with levels of transcription (Pol
990 II ChIP) and SURE genome-wide promoter activity assay. Groups 1-6
991 correspond to increasing level of G4access signals. Green arrows represent
992 the sense of gene transcription. Metaprofiles were centered on the G4 motif
993 upstream of TSS as shown below the tracks.
- 994 (f) pG4s are located at the midpoints of GROcap, Pol II, TBP and TFIIB GTFs
995 using ChIP-exo datasets. See also Figure S3d-e for analyses of ChIP-seq in
996 Raji and Mouse T cells.
- 997 (g) Model of average pG4 locations as determined by G4Hunter algorithm
998 upstream of TSSs and as a midpoint of upstream and downstream Pol II
999 peaks.

1000 See Table S1 for data sets used, references and GEO accession numbers.

1001 **Figure 2: G4s promote nucleosome exclusion at active and inactive promoters,** 1002 ***in vivo* and *in vitro* (see also Figure S2 and S4).**

- 1003 (a) Example of a pG4 sequence fitting the maximum of the nucleosome depleted
1004 regions (NDRs) at the *Btg2* promoter in K562 cells using MNase-seq, and
1005 ChIP-seq of active chromatin marks (H3K4me3, H3K27ac).

- 1006 (b) Experimental G4 signals correlate with levels of nucleosome depletion and
1007 with active histone modification marks. Promoters that harbour strong G4
1008 predictions using G4Hunter algorithm G4H2.0 were ranked by G4access
1009 signal as depicted in the heatmap, corresponding heatmaps of G4-ChIP and
1010 G4H2.0 predictions are also shown (left). Promoters were split in 6 groups, as
1011 in Figure1E. Metaprofiles of nucleosome densities (MNase-seq) and of
1012 H3K4me3 and H3K27ac of all groups are displayed (right). See also Figure
1013 S2H for other cell types. Green arrows represent the sense of gene
1014 transcription.
- 1015 (c) pG4 promote nucleosome exclusion at inactive promoters in K562 cells
1016 (ENCODE). Transcriptionally inactive promoters were split in groups with or
1017 without G4H2.0 prediction. See also Figure 3 for pG4 influence on
1018 nucleosome positioning at inactive intergenic regions and Figure S4A after
1019 transcription inhibition with α -amanitin.
- 1020 (d) pG4-containing promoters have intrinsic nucleosome exclusion properties on
1021 *in vitro* reconstituted chromatin (analysed from (Valouev et al., 2011)). The
1022 promoter selections are based on K562 active promoters shown in Figure 1.
1023 See also Figure S4B-C for pG4s effect on nucleosome exclusion properties
1024 on *in vitro* at inter and intragenic regions.

1025 **Figure 3: pG4s organise nucleosome at intergenic regions (IGRs).**

- 1026 (a) pG4s mark the center of organised nucleosome arrays at IGRs in K562
1027 (ENCODE) and Raji (this study). All G4H2.0 predictions from IGRs are shown.
1028 Heatmaps of nucleosome organisation mapped by MNase-seq is displayed.
1029 Nucleosome positioning (see methods, left), MNase-seq density (right).
- 1030 (b) Nucleosomal organisation at CTCF sites in K562 cells. Metaprofiles of
1031 MNase-seq signal (Top), positioning (Bottom) and CTCF ChIP-seq
1032 (GSE30263) at the 13775 identified binding sites in K562 cells.
- 1033 (c) Nucleosomal organisation at CTCF and pG4 show similar nucleosome
1034 phasing at IGRs. (Top) Overlay of metaprofiles of MNase-seq at IGRs centred
1035 either on pG4s or CTCF ChIP-seq sites in K562 cells; (Bottom) the two
1036 nucleosomes surrounding CTCF or pG4 sites (-1/+1) were aligned to compare
1037 similarity of nucleosome phasing (observed nucleosome repeat lengths of 193
1038 and 194 nt, respectively).

1039 **Figure 4: pG4s and experimental G4s hallmark nucleosome exclusion at CpG
1040 islands (CGIs) (see also Figure S5).**

- 1041 (a) CGIs with pG4s have enhanced chromatin opening, histone modifications and
1042 transcription activity. Heatmaps of the 27702 human CGIs were split in two
1043 groups with (n=21520) or without (n=6182) pG4 (G4H1.5) annotations.
1044 Corresponding signals for pG4, G4access, G4 ChIP, GC and CpG contents,
1045 Pol II, nucleosomes (MNase-seq), chromatin marks and SURE promoter
1046 activity are shown as indicated.
- 1047 (b) CGIs with pG4s have deeper chromatin opening, increased active histone
1048 modifications and promoter activity. Metaprofiles of all marks at CGIs with or
1049 without pG4 displayed in the heatmaps from (a). Further selection and
1050 controls for this analysis are presented in Figure S5.

1051 **Figure 5: G4 mutations impair promoter activity in single cells (see also Figure**
1052 **S6 A-F and Table S3).**

- 1053 (a) Scheme of integrated reporter constructs for the indicated mouse G4-
1054 containing promoters used for smRNA FISH. These constructs were
1055 integrated in human genome (Hela) using a Flp-In strategy. Experimental G4
1056 signals for the mouse promoters in ES cells are shown in Figure S6A-E.
- 1057 (b) Sequence features and G4 structural assessment in the model promoters
1058 indicated in (a). The G4 scores determined by the G4Hunter algorithm,
1059 reflecting stability and likelihood of formation are indicated for WT and mutant
1060 sequences (see Table S3 for the individual sequences). Three independent
1061 assays were performed to conclude for G4 formation *in vitro* on the
1062 oligonucleotide (last column).
- 1063 (c) Quantification of smFISH images of MS2 reporter activity of Eef1a1, Pkm,
1064 Klk6, Taok1 and Polr2a WT and mutant promoters. Mann-Whitney tests were
1065 used (ns $P > 0.05$; **** $P \leq 0.0001$). For Polr2a the two mutants have
1066 moderate (mut2) or strong mutations (mut3) (see panel b). Sequences of all
1067 promoters are provided in Table S3.

1068 **Figure 6: G4 mutations at Eef1a1 model promoters decrease transcription and**
1069 **increase nucleosome density at NDRs, while increasing promoter OFF times in**
1070 **single cells (see also Figure S6 G-I).**

- 1071 (a) Scheme of Eef1a1 mouse model promoters inserted in human HeLa cells with
1072 indicated mutations (see also Table S3 for the sequences) (see also Table S3
1073 for the sequences).
- 1074 (b) G4 forming sequences and the TATAbox regulate gene expression from the
1075 Eef1a1 model promoter in single cells. Quantification of smFISH images of
1076 MS2 reporter activity of WT and mutant Eef1a1 promoters. Representative
1077 smFISH images are shown in Figure S6G. Mann-Whitney tests were used.
- 1078 (c) Eef1a1 G4 mutations result in increased nucleosome density at the apparent
1079 nucleosome depleted region (NDR) location. MNase signal of WT and mutant
1080 promoters was quantified by qPCR ($n=3$; means \pm s.e.m), the NDR location is
1081 highlighted in grey. Quantifications of the nucleosome density signal
1082 variations at NDR are shown on the right of the graph, together with a GAPDH
1083 control. (qPCR oligonucleotides used in are presented Table S4)
- 1084 (d) Pol II and TBP recruitment at Eef1a1 model promoter and mutants. ChIP were
1085 assayed and quantified by qPCR ($n=3$; means \pm s.e.m). Pol II is affected in all
1086 mutant contexts while TBP is impaired specifically in the TATAmut. (qPCR
1087 oligonucleotides used in are presented Table S4)
- 1088 (e) Heatmaps of permissive (red) and non-permissive (blue) transcription periods
1089 for WT and mutant Eef1a1 promoters in single cell live imaging (Long movies,
1090 8h-9h, with stacks every 3 min). Each line represents an individual cell
1091 assessment. Representative movie images and promoter threshold are
1092 shown in Figure 6G.
- 1093 (f) Violin plots of ON and OFF period average duration in WT or mutated Eef1a1
1094 promoters measured in long movies). Computed Pvalues (Mann-Withney
1095 test) are as follows: **** $<1e^{-4}$, *** $<1e^{-3}$, ** $<1e^{-2}$, * $<5e^{-2}$

1096 (g) Heatmaps of the density of polymerases (number of polymerases every 30s)
1097 for WT and mutant Eef1a1 promoters in single cell live imaging (short movies,
1098 30 min, with stacks every 30s). Each line represents an individual cell
1099 assessment.

1100 (h) Interpretative scheme of the signal deconvolution and time constant analysis
1101 derived from the long and short movies. While WT transcription can be
1102 described in 2 main steps both TATA and G4 mutants require at least 3 steps.
1103 The G4 mutant has limiting OFF2 and ON state (see also Figure S6H-I and
1104 methods for mathematical modelling).

1105

1106 **Figure 7: Stabilization of G4 by ligand results in global pause release by Pol II.**
1107 **(see also figure S7)**

1108 (a) Pyridostatin (PDS) treatment (10 min) results in Pol II release from the promoter
1109 area and reduction of the +1 nucleosome at the ACTB, MAT2a and KHLH9 loci.
1110 Tracks of Pol II ChIP-seq and MNase-seq (zoomed around TSS) before and
1111 after PDS treatment are shown. G4 predictions and the sequence of the pG4
1112 upstream of the TSSs associated with the NDRs are indicated and are shown
1113 below the tracks (G4H2.0).

1114 (b) Pausing scores are globally reduced at the genome-scale following PDS
1115 treatment. Scatter plots comparing pausing scores before and after PDS
1116 treatment (10 min) is shown. Genes with enhanced Pol II signal at gene bodies
1117 (Pvalue <0.05) are highlighted in dark blue (n=556). The light blue slope
1118 represents the linear regression curve of the 7061 other points. The whole
1119 kinetic analysis is presented in Figure S7B.

1120 (c) G4Hunter scores distribution of pause release genes and others.

1121 (d) Metaprofiles of Pol II ChIP-seq signal at selected genes with decreased pausing
1122 scores (n=556) depict reduced promoter and increased gene body
1123 occupancies.

1124 (e) Metaprofiles of Pol II ChIP-seq signal at control genes following PDS treatment.
1125 (n=7061).

1126 (f) Metaprofiles nascent ChrRNA-seq genes at pause release genes. (n=556).

1127 (g) Metaprofiles nascent ChrRNA-seq genes at other genes. (n=7061).

1128

1129

1130

1131 **Supplementary data**

1132 **Figure S1: Analysis of promoter elements and motifs in 3 independent cell lines.**

1133 (a) Definition of major TSS (mTSS), MEME and DREME promoter motif analyses
1134 in K562 cells. CAGE datasets (FANTOM consortium) were used to define mTSSs
1135 at the nucleotide resolution (n=8346). Heatmaps show short chromatin (nascent)
1136 RNA-seq (analysed from GSE52914) and Pol II ChIP-seq (ENCODE) docked on
1137 the main sense mTSS and ranked by increasing distance between sense and
1138 antisense short RNAs (top left panel). Sequence motif analysis of the
1139 transcription initiating nucleotides (INR) at the sense and antisense mTSSs are
1140 shown (bottom left). On the right panels are shown TFBS analyses using MEME
1141 or DREME and frequency analyses for all sequences features including
1142 Quadparser (QP1-7) and BRE motifs. The random control column depicts a
1143 search for the motif in 8346 random genomic sequences (with 10000
1144 permutations). See also Table S2 for detailed frequencies of all motifs.

1145 (b) Definition of mTSSs and MEME and DREME motif analyses in Raji B cells.
1146 The analysis was performed as in (a) over 8356 human promoters and using
1147 chrRNA-seq in Raji for mTSS determination (analysed from GSE52914).

1148 (c) Definition of mTSSs, MEME and DREME motif analyses in mouse primary T
1149 cells. The analysis was performed as in (a) over 7947 mouse promoters and
1150 using short-RNA datasets (size-selected below 50 bp, analysed from
1151 GSE38577) to define mTSSs at the nucleotide resolution. Short RNA-seq and
1152 Pol II ChIP-seq data sets are shown.

1153 **Figure S2: Association of the TATAbox and G4 motifs with transcription** 1154 **initiation**

1155 (a) pG4, BRE and SP1 motifs largely overlap at active promoters. Venn diagram
1156 of active promoters containing SP1, BRE motifs or pG4s. G4Hunter is
1157 displayed at two stringencies (1.5, red circle or 2.0, dotted inner circle). The
1158 principle of the G4Hunter algorithm is to score positively Gs and G stretches,
1159 while scoring negatively Cs and C stretches within a defined window (typically
1160 25nt). Overlapping G4s, above a defined threshold, are concatenated. A and
1161 T nucleotides score are fixed as null. G4Hunter scores >1.5 and 2.0
1162 correspond to likelihood of G4 formation *in vitro* of >95 and 99% (Bedrat et
1163 al., 2016; Garcia-Oliver et al., 2022).

1164 (b) Promoters containing G4 predictions (G4H1.5 or 2.0) tend to harbour less
1165 other TFBS or promoter elements as compared to all promoters. This analysis
1166 was performed with the selection described in Figure 1 for K562 cells.

1167 (c) Promoters with TATA boxes show more focused and directional transcription.
1168 Pol II ChIP-seq profiles in K562 cells (ENCODE) at promoters of expressed
1169 genes that contain either no TATA, a non-canonical (TATAW) or a canonical
1170 (TATAWAAG) TATAbox.

1171 (d) Experimental G4 signals definition. Groups 1-6 correspond to increasing level
1172 of G4access signals. Heatmaps in K562 of predicted G4 (H4hunter 2.0) and
1173 G4 signals using G4access (Garcia-Oliver et al., 2022), G4-ChIP (Hansel-
1174 Hertsch et al., 2016; Mao et al., 2018) and G4seq (Chambers et al., 2015); in
1175 Raji of predicted G4 (H4hunter 2.0) and G4 signals using G4access (Garcia-

1176 Oliver et al., 2022), ssDNA-seq (Kouzine et al., 2013) and G4seq (Chambers
1177 et al., 2015). See also Figure 1E and 2B. The heatmaps were centered on the
1178 G4 motif upstream of TSS as shown below the tracks.

1179 (e) Experimental G4 signal metaprofiles. Metaprofiles of the heatmaps shown in
1180 d. The signals were divided in 6 groups of ascending G4 access signals.
1181 Metaprofiles were centered on the G4 motif upstream of TSS as shown below
1182 the tracks.

1183 (f) Metaprofiles of R-loops at active promoters and docked on pG4s (G4H 2.0)
1184 upstream of experimental TSSs in K562 cells. The signals were divided in 6
1185 groups of ascending G4 access signals

1186 (g) Metaprofiles of ChIP-seq of Pol II centered on pG4s (G4H2.0) on 1444 pG4-
1187 containing promoters of active genes (-100,+20 bp) in Raji cell.

1188 (h) Metaprofiles of ChIP-seq of Pol II, TBP, TFIIB, centered on pG4s (G4H2.0)
1189 on 1291 pG4-containing promoters of active genes (-100,+20 bp) in mouse T
1190 cells (GSE38577).

1191 (i) Metaprofiles of Polycomb-deposited H3K27me3 mark, GC and CpG-content
1192 in groups 1 and 2 defined in Figure2C.

1193 A green arrow indicates the sense of transcription in each graph,

1194 All accession numbers are presented in the table S1.

1195

1196 **Figure S3: pG4-dependent nucleosome exclusion does not depend on SP1**
1197 **binding in K562 cells.**

1198 (a) Heatmaps ranked by increasing signals of SP1 binding and centred on a pG4
1199 (ChIP-seq, ENCODE). All promoters of expressed genes that contain a pG4
1200 in K562 cells are presented (n=1766). Group 1 and 2 are depleted or enriched
1201 for SP1, respectively. Metaprofiles derived from the heatmaps (groups 1 and
1202 2) of SP1 density, GC content nucleosome densities and G4access are
1203 shown on the right.

1204 (b) Metaprofiles of promoters not containing a canonical GC box/SP1 binding
1205 site.

1206 (c) Metaprofiles of promoters not containing a non-canonical GC box/SP1 binding
1207 site.

1208 Green arrows indicate the sense of transcription in each graph.

1209

1210 **Figure S4: pG4s show intrinsic nucleosome eviction property**

1211 (a) Persistence of NDRs at pG4 sites following transcription inhibition by α -
1212 amanitin. Raji cells were treated for 0, 12, 18 or 36 h with 2.5 μ g/mL α -
1213 amanitin. Pol II clearance and nucleosome depletion mapped by MNase-seq
1214 from active genes containing pG4s (G4H2.0) are shown at the different time
1215 points (GSE38577 and this study). A green arrow indicates the sense of
1216 transcription in each graph

- 1217 (b) Metaprofiles of MNase-seq from *in vitro* reconstituted chromatin on T cell
1218 genomic DNA (GSE25133) centred on pG4 at the active pG4-containing
1219 promoters (-100, +20 bp), intergenic or intragenic regions defined in K562
1220 cells.
- 1221 (c) Heatmaps ranked by increasing MNase signals showing nucleosome or
1222 G4H1.5 (docked on G4H2.0) signals. Six manually defined groups based on
1223 relative nucleosome densities are further plotted as graphs below the
1224 heatmaps.

1225

1226 **Figure S5: G4s contribute to CpG islands openness and activity.**

- 1227 (a) Selection of G4-containing and G4-depleted CpG islands. The selections were
1228 performed on the same number of sequences (2 x 1112) with similar length and
1229 GC content, with the indicated G4Hunter thresholds. For the groups to be of
1230 equal size, equivalent length and GC content the initial populations of CGIs with
1231 G4H>1.5 (21536) or G4H<1.2 (2191) were randomized to end up with 2 groups
1232 of 1112 sequences.
- 1233 (b) Heatmaps as in Figure S8a for the selections presented in a. The DNase data
1234 used here indicate more chromatin opening in the G4-containing group
1235 (ENCODE data for K562 cells, GSE32970).
- 1236 (c) Average profiles of the group 1 and 2 shown in a and b. For Nucleosome
1237 density, a zoom over 4kb is indicated to best show the differences in the 2
1238 groups. The differences are observed essentially for the width of the NDRs.

1239

1240 **Figure S6: G4s are required for full activities of model promoters**

- 1241 (a)–(e) G4access (Garcia-Oliver et al., 2022) and CUT&Tag (Lyu et al., 2022)
1242 at model promoters. Signals were extracted from data obtained in mouse ES
1243 cell lines at indicated model promoters. Predicted G4s scored by G4Hunter
1244 are indicated. Their sequences, used for promoter assays in single cells, are
1245 shown below the tracks in the zoomed areas. Each of the G tracks ($n>1$) are
1246 indicated in light blue. Complete sequences of the model core promoters are
1247 indicated in Table S3.
- 1248 (f) Quantification of smFISH images of MS2 reporter activity of Eef1a1, WT and
1249 G4inv mutant where the strand of the canonical G4 has been inverted.
1250 Mann-Whitney test was used (ns: $P > 0.05$). Scheme of the promoters and
1251 mutations are indicated below the charts. pG4s are represented in red
1252 rectangles (changes in font orientation indicates swap of strand) and
1253 TATAboxes in green rectangles.
- 1254 (g) Representative images of long movie analyses. Long movies of MS2
1255 reporter activity in WT and mutant cell lines as indicated. (Left) maximum image
1256 projections of selected 3-D image stacks from the 8h movies. The arrows
1257 indicate transcription sites. (Right) graphs display the corresponding
1258 quantifications of the movies, with the bars representing ON (green) and OFF
1259 (red) states.

1260 (h) Survival functions describing transcriptional bursting at the model
1261 promoters and deconvolution indicating 2 fits to describe optimal function in the
1262 WT situation and 3 fits for both TATA and G4 mutants.

1263 (i) WT promoters require 2 main steps for transcription, and mutant
1264 promoters 3 main steps defined by 3 or 5 time constants, respectively. The
1265 values of the derived time constants (described in methods) are indicated in the
1266 right panel.

1267 **Figure S7: Time course analysis of G4s stabilization by PDS effects on**
1268 **nucleosome exclusion and positioning, and Pol II pausing.**

1269 (a) FACS analysis of gH2AX levels following PDS treatment at the indicated time
1270 (n=3).

1271 (b) PDS influences Pol II pausing. Changes of Pol II pausing scores (see
1272 methods) in response to 10 μ M PDS of all genes with detectable Pol II at
1273 promoters and gene bodies for 0, 10, 30 and 60 min are displayed. Genes that
1274 have increased gene body signals with a $P_{val} < 0.05$ at $t=10$ min (n=556) are
1275 highlighted in dark blue (t=10 min), dark green (t=30 min) or dark orange (t=60
1276 min), showing that pause release is relatively transient following treatment.
1277 Changes in pausing scores, promoter and gene body signals are shown.

1278 (c) Gene ontology analysis of the 556 selected pause release genes using
1279 DAVID webtool.

1280 **Table S1: Genomic files**

1281 Table summarizing all resources of the genomic analyses presented in this study.
1282 The table depicts the cell type, the experiment type and the source of the files.

1283 **Table S2: Motif search analysis**

1284 Frequency analyses of sequence features in a 120 bp window around experimental
1285 TSSs (-100, +20). 14 motifs or features were analysed (BREd, BREu, ETS,
1286 G4H1.5, G4H2.0, QP1-7, INR, NF-Y, Half G-quadruplex, SP1 motifs, two TATAbox
1287 consensus and two negative control motifs).

1288 **Table S3: Sequence of all mouse model promoters used in this study.**

1289 These sequences were inserted upstream of the MS2 reporter and inserted in Hela
1290 genome using FRT system.

1291 **Table S4: Table of qPCR oligonucleotides used in this study**

1292 All oligonucleotide pairs used for qPCR are presented; sequences, complementary
1293 genomes and genomic location of amplicons are described.

1294

1295 **References**

1296

- 1297 Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II:
1298 emerging roles in metazoans. *Nat Rev Genet* 13, 720-731.
- 1299 Agarwal, T., Roy, S., Kumar, S., Chakraborty, T.K., and Maiti, S. (2014). In the sense
1300 of transcription regulation by G-quadruplexes: asymmetric effects in sense and
1301 antisense strands. *Biochemistry* 53, 3711-3718.
- 1302 Alexander, R.D., Barrass, J.D., Dichtl, B., Kos, M., Obtulowicz, T., Robert, M.C., Koper,
1303 M., Karkusiewicz, I., Mariconti, L., Tollervey, D., *et al.* (2010). RiboSys, a high-
1304 resolution, quantitative approach to measure the in vivo kinetics of pre-mRNA splicing
1305 and 3'-end processing in *Saccharomyces cerevisiae*. *RNA* 16, 2570-2580.
- 1306 Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with
1307 high-throughput sequencing data. *Bioinformatics* 31, 166-169.
- 1308 Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter
1309 activities of regulatory elements. *Nat Rev Genet* 21, 71-87.
- 1310 Bagchi, D.N., and Iyer, V.R. (2016). The Determinants of Directionality in
1311 Transcriptional Initiation. *Trends Genet* 32, 322-333.
- 1312 Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data.
1313 *Bioinformatics* 27, 1653-1659.
- 1314 Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li,
1315 W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching.
1316 *Nucleic acids research* 37, W202-208.
- 1317 Bansal, M., Kumar, A., and Yella, V.R. (2014). Role of DNA sequence based structural
1318 features of promoters in transcription initiation and gene expression. *Curr Opin Struct*
1319 *Biol* 25, 77-85.
- 1320 Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli,
1321 G. (2014). Coregulation of transcription factor binding and nucleosome occupancy
1322 through DNA features of mammalian enhancers. *Mol Cell* 54, 844-857.
- 1323 Bedrat, A., Lacroix, L., and Mergny, J.L. (2016). Re-evaluation of G-quadruplex
1324 propensity with G4Hunter. *Nucleic acids research* 44, 1746-1759.
- 1325 Bhuiyan, T., and Timmers, H.T.M. (2019). Promoter Recognition: Putting TFIID on the
1326 Spot. *Trends Cell Biol* 29, 752-763.
- 1327 Bochman, M.L., Paeschke, K., and Zakian, V.A. (2012). DNA secondary structures:
1328 stability and function of G-quadruplex structures. *Nat Rev Genet* 13, 770-780.
- 1329 Boireau, S., Maiuri, P., Basyuk, E., de la Mata, M., Knezevich, A., Pradet-Balade, B.,
1330 Backer, V., Kornblihtt, A., Marcello, A., and Bertrand, E. (2007). The transcriptional
1331 cycle of HIV-1 in real-time and live cells. *J Cell Biol* 179, 291-304.

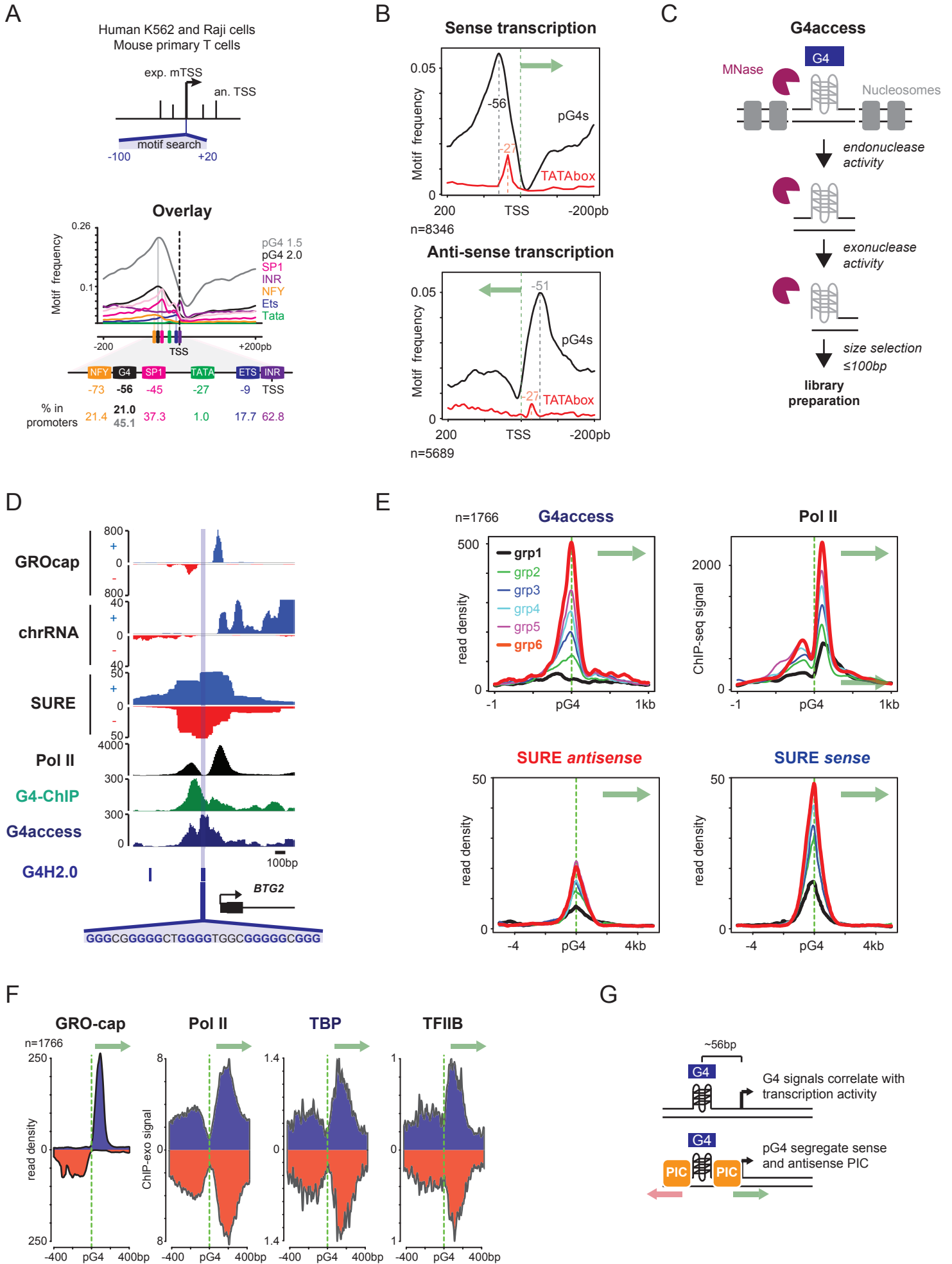
- 1332 Buratowski, S., Hahn, S., Guarente, L., and Sharp, P.A. (1989). Five intermediate
1333 complexes in transcription initiation by RNA polymerase II. *Cell* 56, 549-561.
- 1334 Burke, T.W., and Kadonaga, J.T. (1996). *Drosophila* TFIID binds to a conserved
1335 downstream basal promoter element that is present in many TATA-box-deficient
1336 promoters. *Genes Dev* 10, 711-724.
- 1337 Butler, J.E., and Kadonaga, J.T. (2002). The RNA polymerase II core promoter: a key
1338 component in the regulation of gene expression. *Genes Dev* 16, 2583-2592.
- 1339 Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P., and
1340 Balasubramanian, S. (2015). High-throughput sequencing of DNA G-quadruplex
1341 structures in the human genome. *Nature biotechnology* 33, 877-881.
- 1342 Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014).
1343 Analysis of nascent RNA identifies a unified architecture of initiation regions at
1344 mammalian promoters and enhancers. *Nat Genet* 46, 1311-1320.
- 1345 Dao, P., Wojtowicz, D., Nelson, S., Levens, D., and Przytycka, T.M. (2016). Ups and
1346 Downs of Poised RNA Polymerase II in B-Cells. *PLoS Comput Biol* 12, e1004821.
- 1347 David, A.P., Margarit, E., Domizi, P., Banchio, C., Armas, P., and Calcaterra, N.B.
1348 (2016). G-quadruplexes as novel cis-elements controlling transcription during
1349 embryonic development. *Nucleic acids research* 44, 4163-4173.
- 1350 Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription.
1351 *Genes Dev* 25, 1010-1022.
- 1352 Descostes, N., Heidemann, M., Spinelli, L., Schuller, R., Maqbool, M.A., Fenouil, R.,
1353 Koch, F., Innocenti, C., Gut, M., Gut, I., *et al.* (2014). Tyrosine phosphorylation of RNA
1354 polymerase II CTD is associated with antisense promoter transcription and active
1355 enhancers in mammalian cells. *Elife* 3, e02105.
- 1356 Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier,
1357 P., Spicuglia, S., Gut, M., Gut, I., *et al.* (2012). CpG islands and GC content dictate
1358 nucleosome depletion in a transcription-independent manner at mammalian
1359 promoters. *Genome Res* 22, 2399-2408.
- 1360 Fenouil, R., Descostes, N., Spinelli, L., Koch, F., Maqbool, M.A., Benoukraf, T.,
1361 Cauchy, P., Innocenti, C., Ferrier, P., and Andrau, J.C. (2016). Pasha: a versatile R
1362 package for piling chromatin HTS data. *Bioinformatics* 32, 2528-2530.
- 1363 Frontini, M., Imbriano, C., diSilvio, A., Bell, B., Bogni, A., Romier, C., Moras, D., Tora,
1364 L., Davidson, I., and Mantovani, R. (2002). NF-Y recruitment of TFIID, multiple
1365 interactions with histone fold TAF(II)s. *J Biol Chem* 277, 5841-5848.
- 1366 Fu, Y., Sinha, M., Peterson, C.L., and Weng, Z. (2008). The insulator binding protein
1367 CTCF positions 20 nucleosomes around its binding sites across the human genome.
1368 *PLoS Genet* 4, e1000138.

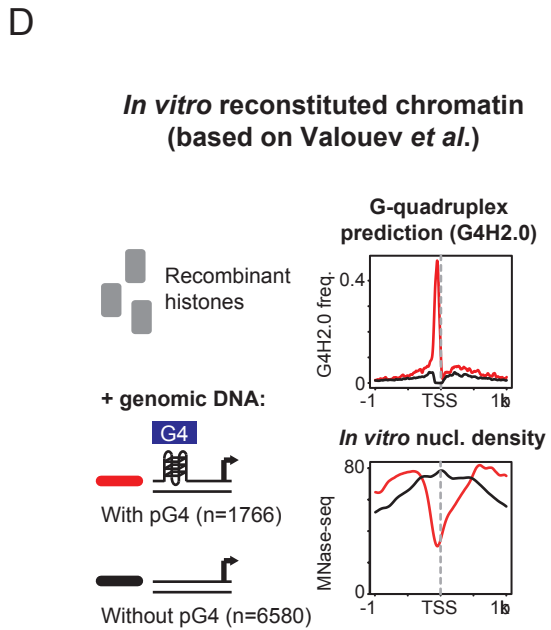
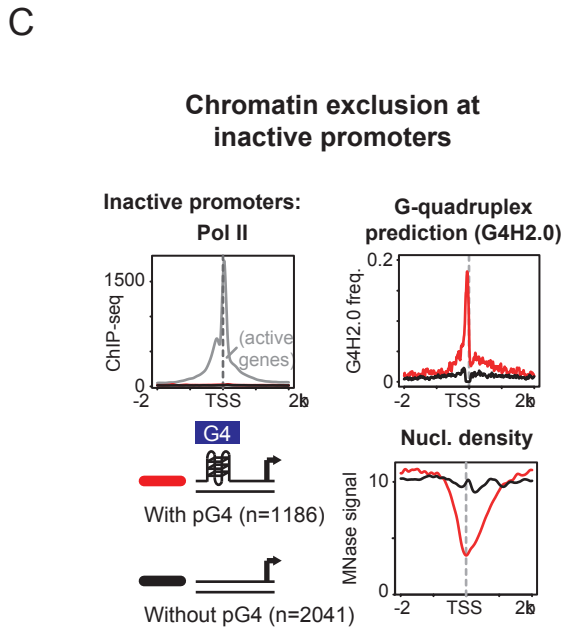
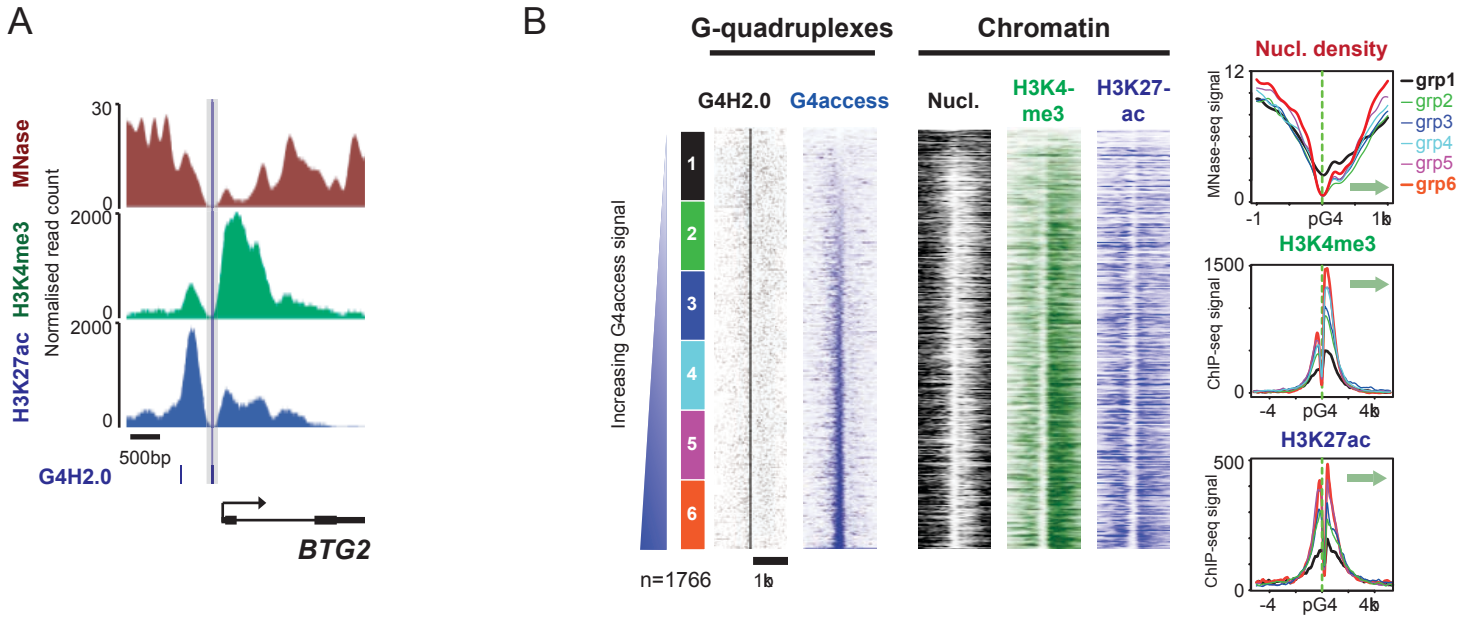
- 1369 Garcia-Oliver, E., Esnault, C., Zine El Aabidine, A., Magat, T., Cucchiarini, A., Lleres,
1370 D., Goerke, L., Verga, D., Lacroix, L., Feil, R., *et al.* (2022). G4access reveals G-
1371 quadruplexes association to open chromatin and to imprinting regions control. *Nature*
1372 *Genetics In revision.*
- 1373 Guedin, A., Gros, J., Alberti, P., and Mergny, J.L. (2010). How long is too long? Effects
1374 of loop size on G-quadruplex stability. *Nucleic acids research* 38, 7858-7868.
- 1375 Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis
1376 of transcription initiation. *Nat Rev Mol Cell Biol* 19, 621-637.
- 1377 Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di
1378 Antonio, M., Pike, J., Kimura, H., Narita, M., *et al.* (2016). G-quadruplex structures
1379 mark human regulatory chromatin. *Nat Genet* 48, 1267-1272.
- 1380 Huppert, J.L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008). G-
1381 quadruplexes: the beginning and end of UTRs. *Nucleic acids research* 36, 6260-6268.
- 1382 Jacob, F., Ullman, A., and Monod, J. (1964). [the Promotor, a Genetic Element
1383 Necessary to the Expression of an Operon]. *C R Hebd Seances Acad Sci* 258, 3125-
1384 3128.
- 1385 Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation:
1386 advances through genomics. *Nat Rev Genet* 10, 161-172.
- 1387 Kaplan, N., Moore, I., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., Hughes,
1388 T.R., Lieb, J.D., Widom, J., and Segal, E. (2010). Nucleosome sequence preferences
1389 influence in vivo nucleosome organization. *Nat Struct Mol Biol* 17, 918-920.
- 1390 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013).
1391 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions
1392 and gene fusions. *Genome Biol* 14, R36.
- 1393 Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia,
1394 S., de la Chapelle, A.L., Heidemann, M., Hintermair, C., *et al.* (2011). Transcription
1395 initiation platforms and GTF recruitment at tissue-specific enhancers and promoters.
1396 *Nat Struct Mol Biol* 18, 956-963.
- 1397 Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-
1398 Kwon, K.R., Benham, C.J., Casellas, R., Przytycka, T.M., *et al.* (2017).
1399 Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with
1400 Regulatory Potential across a Mammalian Genome. *Cell Syst* 4, 344-356 e347.
- 1401 Kouzine, F., Wojtowicz, D., Yamane, A., Resch, W., Kieffer-Kwon, K.R., Bandle, R.,
1402 Nelson, S., Nakahashi, H., Awasthi, P., Feigenbaum, L., *et al.* (2013). Global regulation
1403 of promoter melting in naive lymphocytes. *Cell* 153, 988-999.
- 1404 Lago, S., Nadai, M., Cernilogar, F.M., Kazerani, M., Domínguez Moreno, H., Schotta,
1405 G., and Richter, S.N. (2021). Promoter G-quadruplexes and transcription factors

- 1406 cooperate to shape the cell type-specific transcriptome. *Nature communications* 12,
1407 3885.
- 1408 Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2.
1409 *Nat Methods* 9, 357-359.
- 1410 Li, C., Wang, H., Yin, Z., Fang, P., Xiao, R., Xiang, Y., Wang, W., Li, Q., Huang, B.,
1411 Huang, J., *et al.* (2021). Ligand-induced native G-quadruplex stabilization impairs
1412 transcription initiation. *Genome Res* 31, 1546-1560.
- 1413 Lyu, J., Shao, R., Kwong Yung, P.Y., and Elsässer, S.J. (2022). Genome-wide
1414 mapping of G-quadruplex structures with CUT&Tag. *Nucleic acids research* 50, e13.
- 1415 Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio,
1416 M., Marsico, G., Hansel-Hertsch, R., Tannahill, D., and Balasubramanian, S. (2018).
1417 DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol Biol* 25, 951-
1418 957.
- 1419 Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Antonio,
1420 M.D., and Balasubramanian, S. (2019). Whole genome experimental maps of DNA G-
1421 quadruplexes in multiple species. *Nucleic acids research* 47, 3862-3874.
- 1422 Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and
1423 Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants
1424 influencing human transcription factor occupancy in vivo. *Nat Genet* 47, 1393-1401.
- 1425 Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E.,
1426 Bertrand, E., Darzacq, X., and Zimmer, C. (2013). FISH-quant: automatic counting of
1427 transcripts in 3D FISH images. *Nat Methods* 10, 277-278.
- 1428 Muller, F., and Tora, L. (2014). Chromatin and DNA sequences in defining promoters
1429 for transcription initiation. *Biochim Biophys Acta* 1839, 118-128.
- 1430 Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca,
1431 M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent
1432 Transcription Coupled to RNA Processing. *Cell* 161, 526-540.
- 1433 Oldfield, A.J., Henriques, T., Kumar, D., Burkholder, A.B., Cinghu, S., Paulet, D.,
1434 Bennett, B.D., Yang, P., Scruggs, B.S., Lavender, C.A., *et al.* (2019). NF-Y controls
1435 fidelity of transcription initiation at gene promoters through maintenance of the
1436 nucleosome-depleted region. *Nature communications* 10, 3072.
- 1437 Olivieri, M., Cho, T., Álvarez-Quilón, A., Li, K., Schellenberg, M.J., Zimmermann, M.,
1438 Hustedt, N., Rossi, S.E., Adam, S., Melo, H., *et al.* (2020). A Genetic Map of the
1439 Response to DNA Damage in Human Cells. *Cell* 182, 481-496.e421.
- 1440 Parry, T.J., Theisen, J.W., Hsu, J.Y., Wang, Y.L., Corcoran, D.L., Eustice, M., Ohler,
1441 U., and Kadonaga, J.T. (2010). The TCT motif, a key component of an RNA
1442 polymerase II transcription system for the translational machinery. *Genes Dev* 24,
1443 2013-2018.

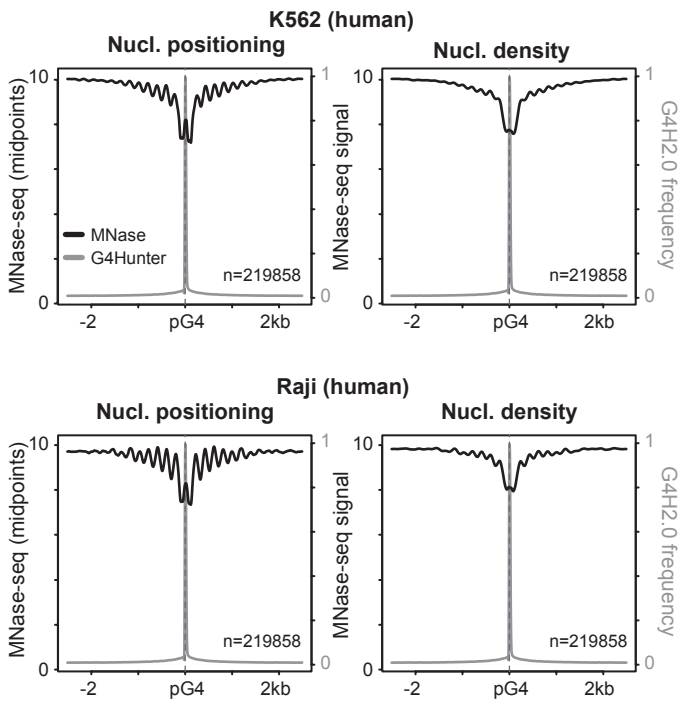
- 1444 Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an
1445 early T7 promoter. *Proc Natl Acad Sci U S A* 72, 784-788.
- 1446 Pugh, B., and Venters, B. (2016). Genomic Organization of Human Transcription
1447 Initiation Complexes. *PLoS One Feb 11;11(2):e0149339*.
- 1448 Radman-Livaja, M., and Rando, O.J. (2010). Nucleosome positioning: how is it
1449 established, and why does it matter? *Dev Biol* 339, 258-266.
- 1450 Raiber, E.A., Kranaster, R., Lam, E., Nikan, M., and Balasubramanian, S. (2012). A
1451 non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro.
1452 *Nucleic acids research* 40, 1499-1508.
- 1453 Reed, B.D., Charos, A.E., Szekely, A.M., Weissman, S.M., and Snyder, M. (2008).
1454 Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel
1455 functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet*
1456 4, e1000133.
- 1457 Renaud de la Faverie, A., Guedin, A., Bedrat, A., Yatsunyk, L.A., and Mergny, J.L.
1458 (2014). Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic acids*
1459 *research* 42, e65.
- 1460 Rodriguez, R., Muller, S., Yeoman, J.A., Trentesaux, C., Riou, J.F., and
1461 Balasubramanian, S. (2008). A novel small molecule that alters shelterin integrity and
1462 triggers a DNA-damage response at telomeres. *J Am Chem Soc* 130, 15758-15759.
- 1463 Rosenberg, M., and Court, D. (1979). Regulatory sequences involved in the promotion
1464 and termination of RNA transcription. *Annu Rev Genet* 13, 319-353.
- 1465 Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data.
1466 *Bioinformatics* 20, 3246-3248.
- 1467 Segal, E., and Widom, J. (2009). What controls nucleosome positions? *Trends Genet*
1468 25, 335-343.
- 1469 Sen, D., and Gilbert, W. (1988). Formation of parallel four-stranded complexes by
1470 guanine-rich motifs in DNA and its implications for meiosis. *Nature* 334, 364-366.
- 1471 Shen, J., Varshney, D., Simeone, A., Zhang, X., Adhikari, S., Tannahill, D., and
1472 Balasubramanian, S. (2021). Promoter G-quadruplex folding precedes transcription
1473 and is controlled by chromatin. *Genome Biol* 22, 143.
- 1474 Smestad, J.A., and Maher, L.J., 3rd (2015). Relationships between putative G-
1475 quadruplex-forming sequences, RecQ helicases, and transcription. *BMC Med Genet*
1476 16, 91.
- 1477 Tantale, K., Garcia-Oliver, E., Robert, M.C., L'Hostis, A., Yang, Y., Tsanov, N., Topno,
1478 R., Gostan, T., Kozulic-Pirher, A., Basu-Shrivastava, M., *et al.* (2021). Stochastic
1479 pausing at latent HIV-1 promoters generates transcriptional bursting. *Nature*
1480 *communications* 12, 4503.

- 1481 Tantale, K., Mueller, F., Kozulic-Pirher, A., Lesne, A., Victor, J.M., Robert, M.C.,
1482 Capozzi, S., Chouaib, R., Backer, V., Mateos-Langerak, J., *et al.* (2016a). A single-
1483 molecule view of transcription reveals convoys of RNA polymerases and multi-scale
1484 bursting. *Nature communications* 7, 12248.
- 1485 Tantale, K., Mueller, F., Kozulic-Pirher, A., Lesne, A., Victor, J.M., Robert, M.C.,
1486 Capozzi, S., Chouaib, R., Backer, V., Mateos-Langerak, J., *et al.* (2016b). A single-
1487 molecule view of transcription reveals convoys of RNA polymerases and multi-scale
1488 bursting. *Nature communications* 7, 12248.
- 1489 Tikhonova, P., Pavlova, I., Isaakova, E., Tsvetkov, V., Bogomazova, A., Vedekhina,
1490 T., Luzhin, A.V., Sultanov, R., Severov, V., Klimina, K., *et al.* (2021). DNA G-
1491 Quadruplexes Contribute to CTCF Recruitment. *Int J Mol Sci* 22.
- 1492 Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011).
1493 Determinants of nucleosome organization in primary human cells. *Nature* 474, 516-
1494 520.
- 1495 van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J.,
1496 Bussemaker, H.J., and van Steensel, B. (2017). Genome-wide mapping of
1497 autonomous promoter activity in human cells. *Nature biotechnology* 35, 145-153.
- 1498 Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier,
1499 R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T. (2007).
1500 Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.
1501 *Cell* 131, 58-69.
- 1502 Vo Ngoc, L., Wang, Y.L., Kassavetis, G.A., and Kadonaga, J.T. (2017). The punctilious
1503 RNA polymerase II core promoter. *Genes Dev* 31, 1289-1301.
- 1504 Xia, Y., Zheng, K.W., He, Y.D., Liu, H.H., Wen, C.J., Hao, Y.H., and Tan, Z. (2018).
1505 Transmission of dynamic supercoiling in linear and multi-way branched DNAs and its
1506 regulation revealed by a fluorescent G-quadruplex torsion sensor. *Nucleic acids*
1507 *research* 46, 7418-7424.
- 1508

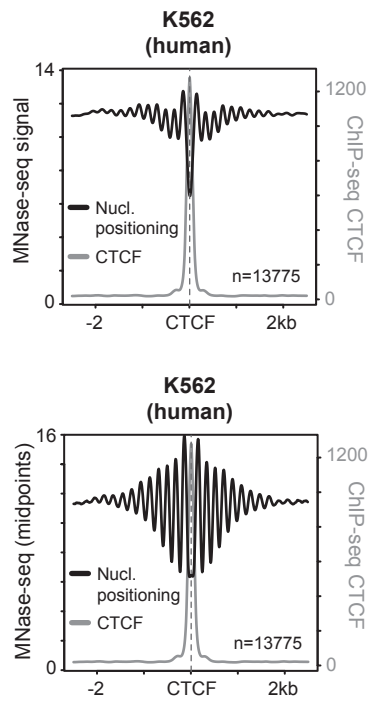




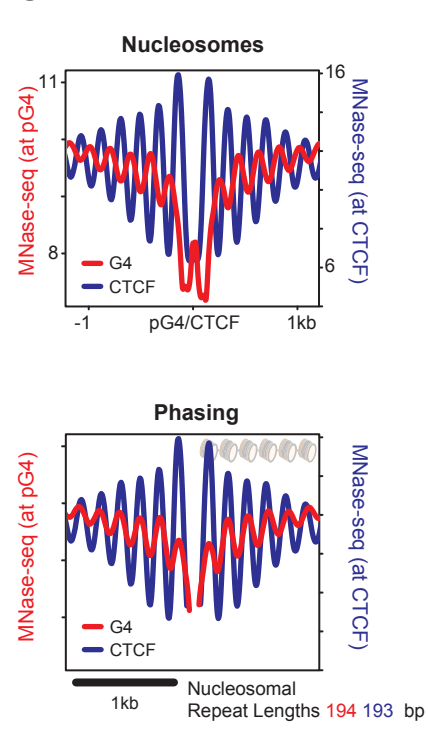
A

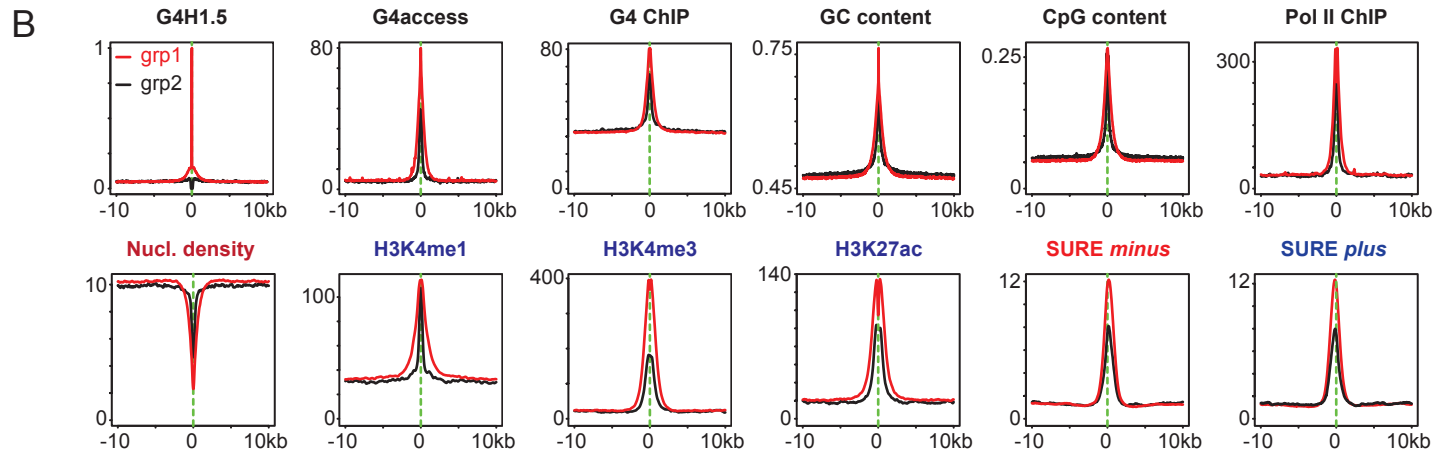
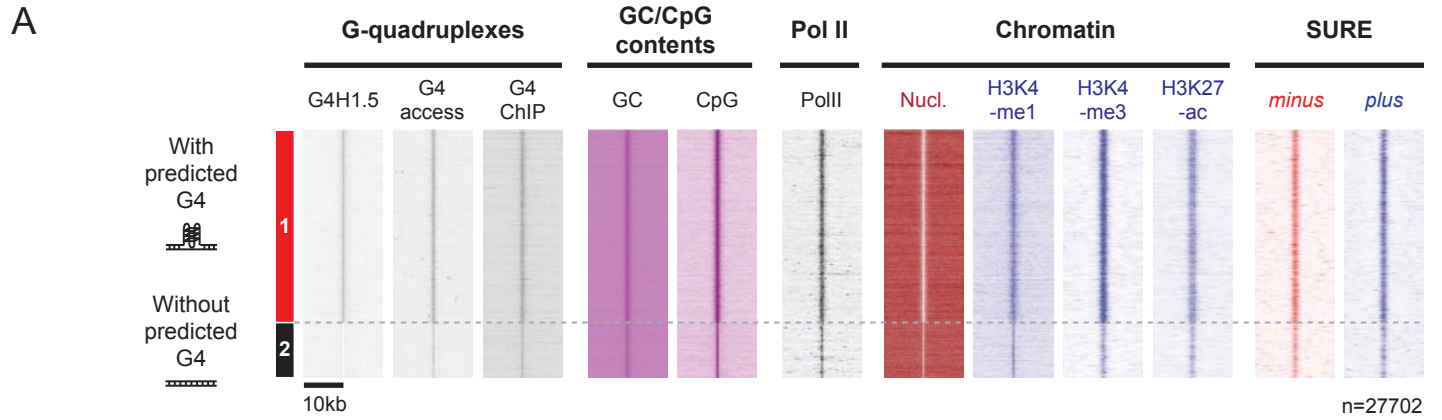


B



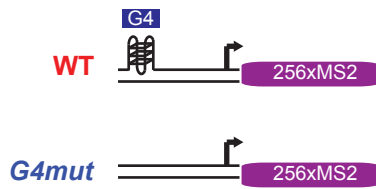
C





A

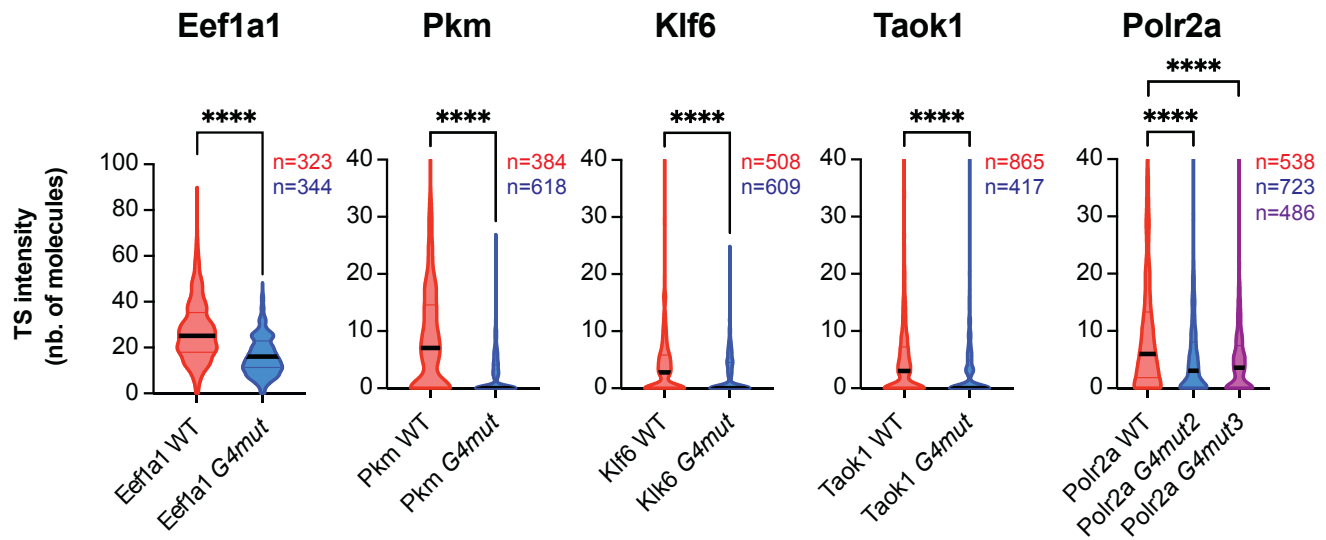
Pkm, Klf6, Taok1, Pol2ra, Eef1a1 promoters



B

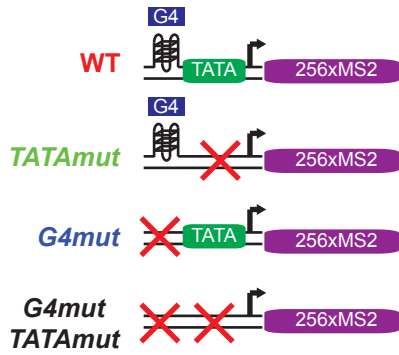
		Score	length	IDS	TDS	CD	Conclusion
Eef1a1	WT	2.14	59	Yes	Yes	Yes	G4
	G4mut	1.05	59	No	No	n.c.	No G4
	G4inv	2.14	59	Yes	Yes	Yes	G4
Polr2a	WT	1.62	37	Yes	Yes	Yes	G4
	G4mut2	0.81	37	No	No	No	no G4
	G4mut3	0.43	37	No	No	No	no G4
Taok1	WT	2.22	37	Yes	Yes	Yes	G4
	G4mut	0.65	37	No	No	?	no G4
Pkm	WT	1.93	27	Yes	Yes	Yes	G4
	G4mut	0.44	27	No	No	?	no G4
Klf6	WT	2.29	41	Yes	Yes	Yes	G4
	G4mut	0.56	41	No	No	?	no G4

C

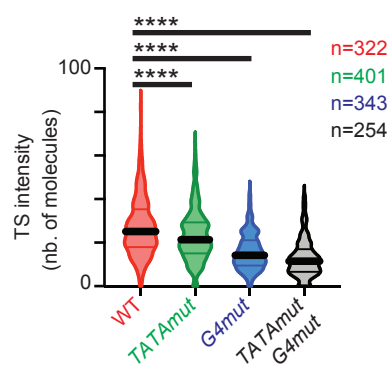


A

Eef1a1 promoters

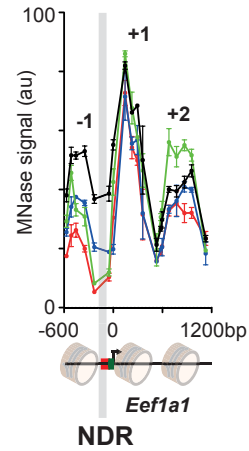


B

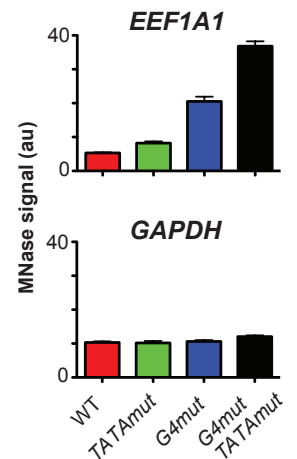


C

Nucleosomes

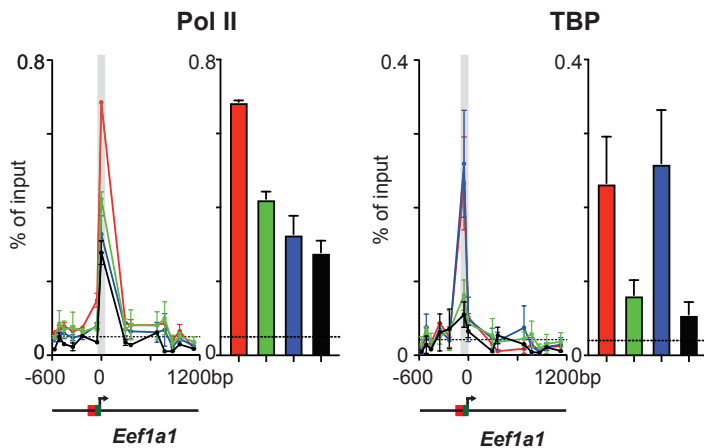


Nucl. density at NDR



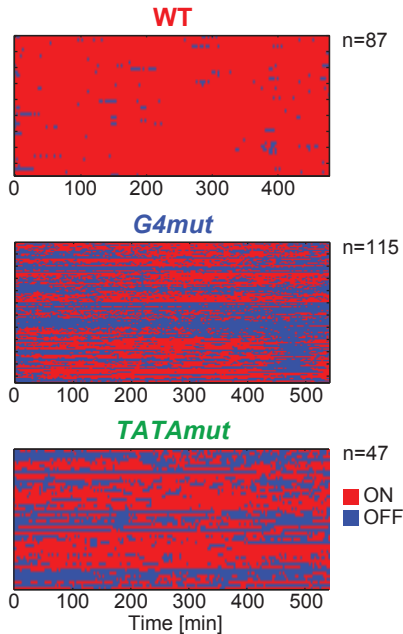
D

— WT
 — TATAmut
 — G4mut
 — G4mutTATAmut

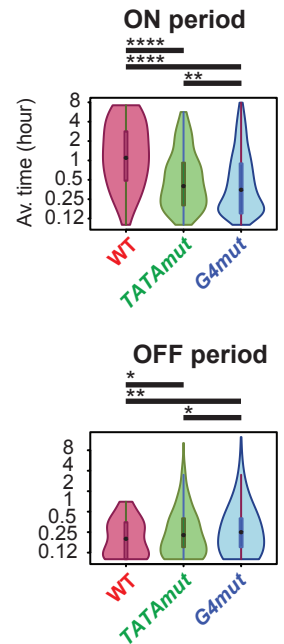


E

Long movies

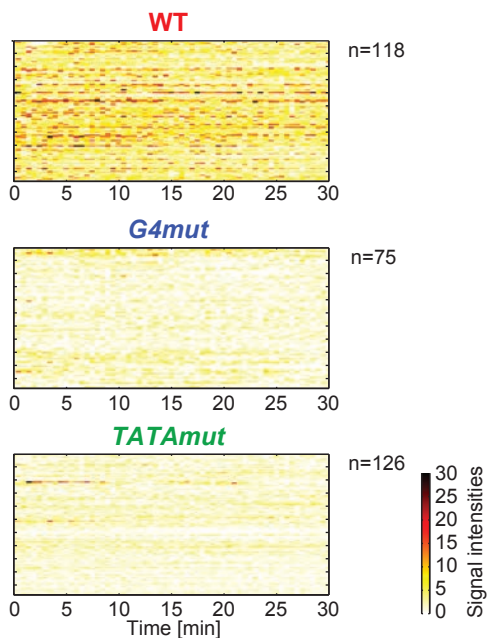


F

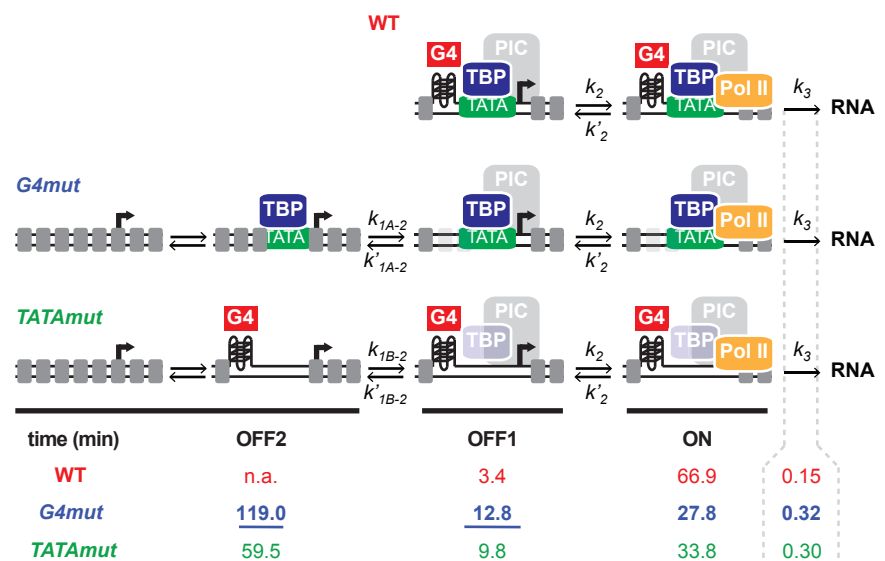


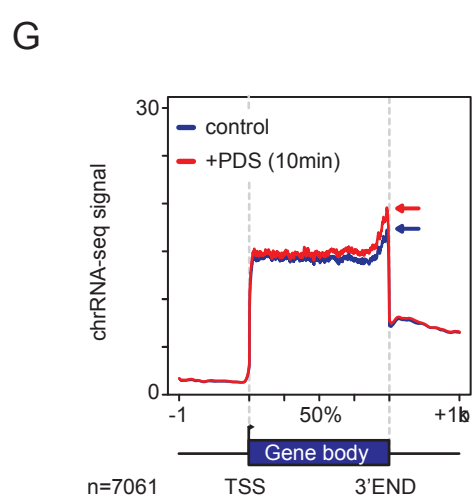
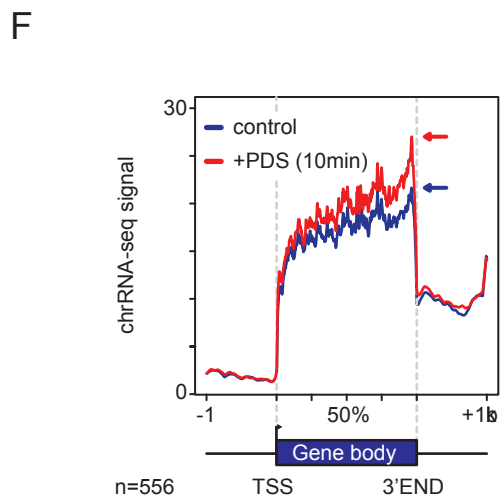
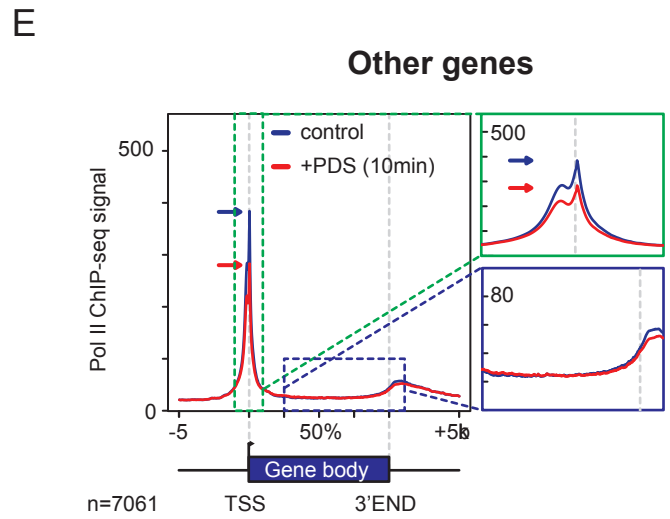
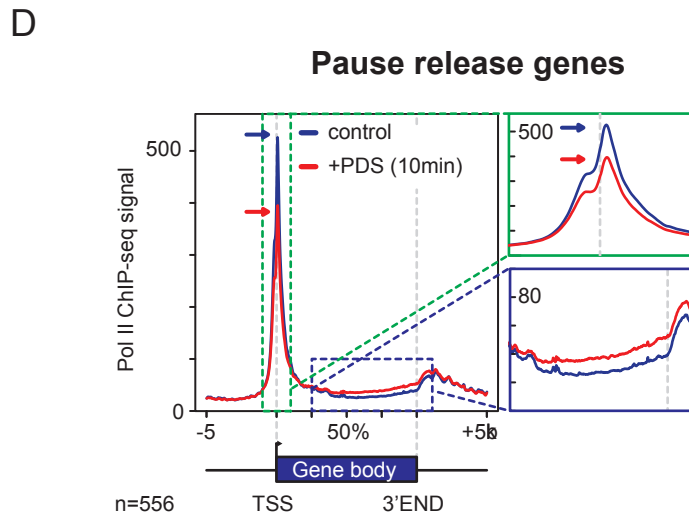
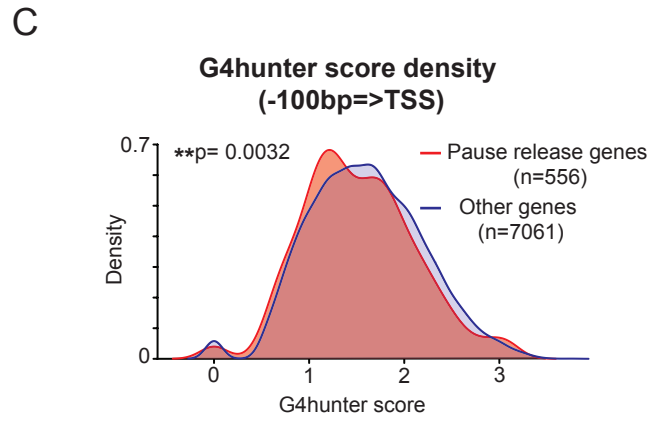
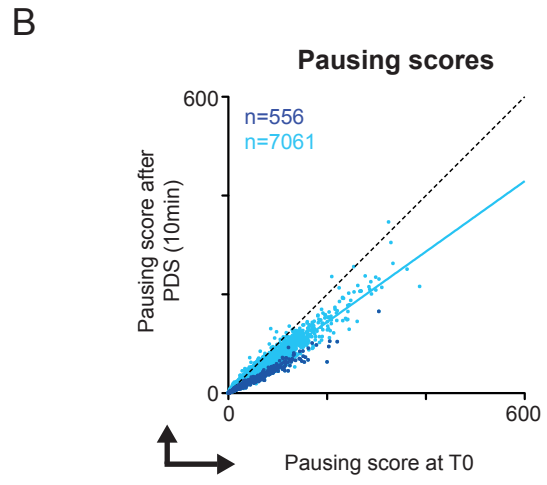
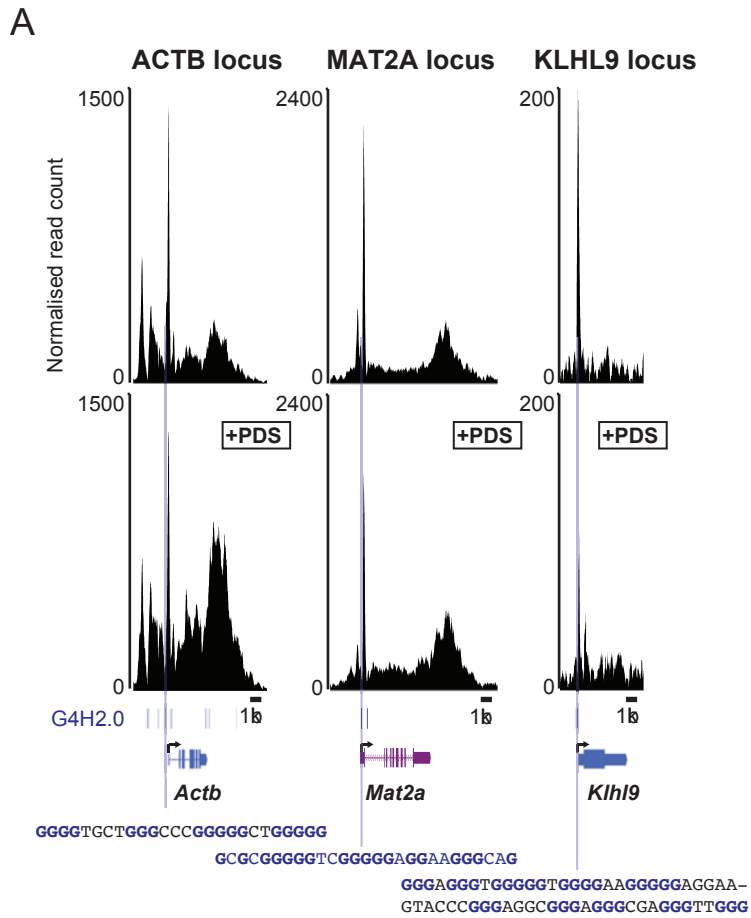
G

Short movies



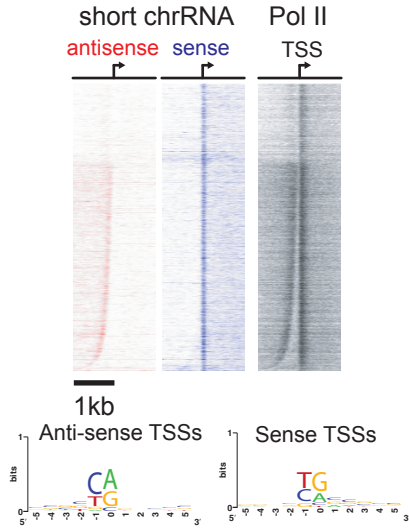
H





A

K562 EL cells (n=8346)

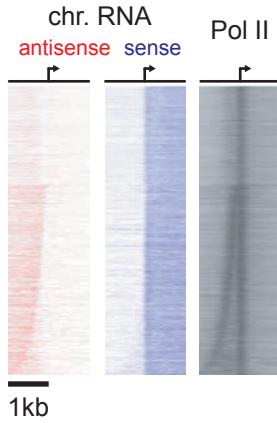


Motif	MEME				DREME		
	Sites	E-value	Rank	Seq. Logo	E-value	Rank	Seq. Logo
G4/SP1	2846	2.2e ⁻¹⁵⁵	1		1.4e ⁻²⁹⁸	1	
Ets	1308	1.2e ⁻³²	2		3.7e ⁻¹⁹⁴	2	
NFY	1071	2.2e ⁻²⁰	3		1.2e ⁻¹⁰⁴	3	
TATA			n.a		6.6e ⁻⁵⁷	9	

	% in promoters	% (cor. for GC)	obs./exp.	Pvalue	Random cont.
BRE (SSRCGCC)	45.2% (3772)	6.3% (528)	7.1	<1e ⁻⁹⁹	2.9% (242)
G4 (G4H2.0)	21.1% (1766)	3.1% (262)	6.7	<1e ⁻⁹⁹	1.5% (121)
G4 (QP1-7)	21.0% (1757)	3.2% (274)	6.4	<1e ⁻⁹⁹	1.5% (125)
G4 (G4H1.5)	45.1% (3763)	12.0% (1005)	3.7	<1e ⁻⁹⁹	5.9% (491)
TATA (TATAWAAG)	1.0% (80)	0.6% (46)	1.7	<1e ⁻⁹⁹	1.5% (127)
TATA (TATAW)	8.6% (721)	16.5% (1373)	0.5	ns	33.7% (2811)

B

Raji B cells (n=8356)

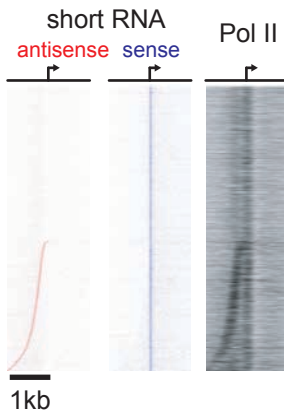


Motif	MEME				DREME		
	Sites	E-value	Rank	Seq. Logo	E-value	Rank	Seq. Logo
G4/SP1	3230	3.9e ⁻²³⁸	1		7.4e ⁻²³⁸	2	
Ets	2187	1.0e ⁻⁹⁶	2		7.3e ⁻²⁶⁰	1	
NFY	934	6.9e ⁻⁵⁷	3		8.0e ⁻¹⁵⁶	3	
TATA			n.a		3.5e ⁻²⁸	10	

	% in promoters	% (cor. for GC)	obs./exp.	Pvalue
BRE (SSRCGCC)	44.0% (3680)	6.3% (529)	7.0	<1e ⁻⁹⁹
G4 (G4H2.0)	17.3% (1444)	3.1% (262)	5.5	<1e ⁻⁹⁹
G4 (QP1-7)	17.6% (1467)	3.2% (275)	5.3	<1e ⁻⁹⁹
G4 (G4H1.5)	40.7% (3398)	12.0% (1007)	3.4	<1e ⁻⁹⁹
TATA (TATAWAAG)	0.9% (73)	0.6% (47)	1.6	<1e ⁻⁹⁹
TATA (TATAW)	6.3% (523)	16.5% (1375)	0.4	ns

C

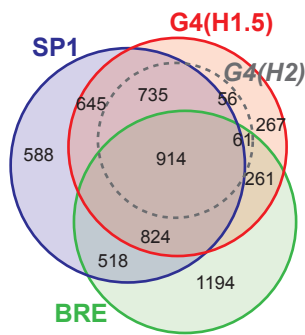
Mouse primary T cells (n=7947)



Motif	MEME				DREME			
	Sites	E-value	Rank	Seq. Logo	Motif	E-value	Rank	Seq. Logo
G4/SP1	2249	3.9e ⁻²³⁸	1		G4/SP1	1.5e ⁻²²⁷	1	
A-stretch	210	2.1e ⁻⁴⁷	2		Ets	1.3e ⁻¹³⁵	2	
Ets or NFY			n.a		NFY	1.8e ⁻⁹⁶	3	
TATA			n.a		TATA	6.3e ⁻⁴⁶	7	

	% in promoters	% (cor. for GC)	obs./exp.	Pvalue
BRE (SSRCGCC)	42.3% (3364)	6.5% (512)	6.6	<1e ⁻⁹⁹
G4 (G4H2.0)	16.2% (1291)	3.0% (235)	5.5	<1e ⁻⁹⁹
G4 (QP1-7)	16.3% (1294)	3.0% (238)	5.4	<1e ⁻⁹⁹
G4 (G4H1.5)	40.1% (3184)	11.0% (944)	3.4	<1e ⁻⁹⁹
TATA (TATAWAAG)	1.0% (83)	0.6% (46)	1.8	<1e ⁻⁹⁹
TATA (TATAW)	6.4% (509)	16.9% (1345)	0.4	ns

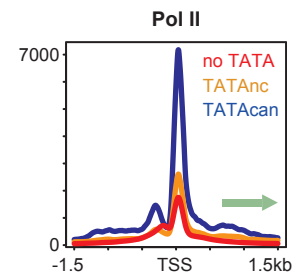
A



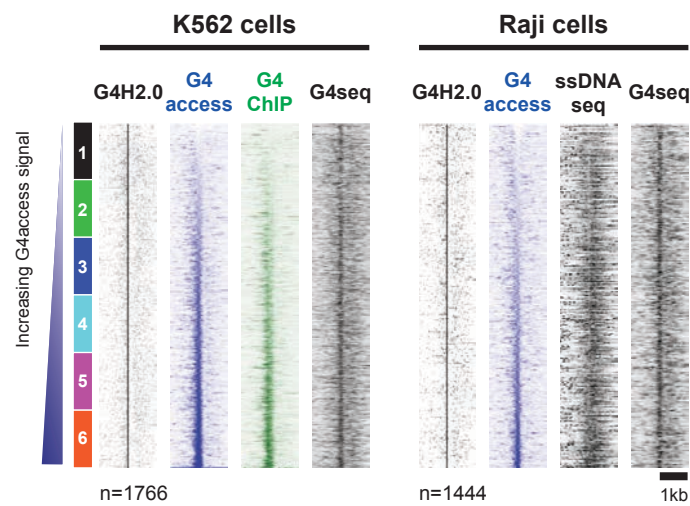
B

	All prom	G4H1.5	G4H2.0	Ets	NFY	TATAcan	TATA
All prom	100						
G4H1.5	45	100					
G4H2.0	21	47	100				
Ets	18	13	10	100			
NFY	21	17	12	17	100		
TATAcan	1.0	0.6	0.3	0.5	1.5	100	
TATA	9	4.6	3.9	5	12	100	100

C

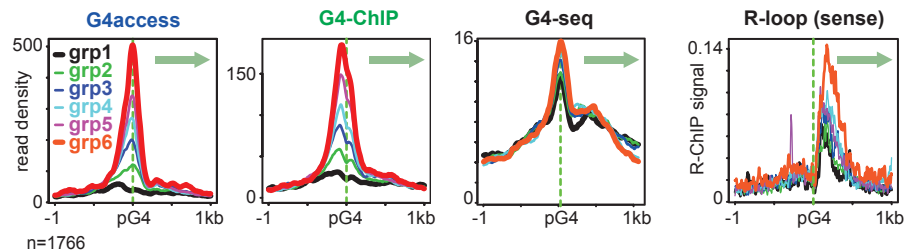


D



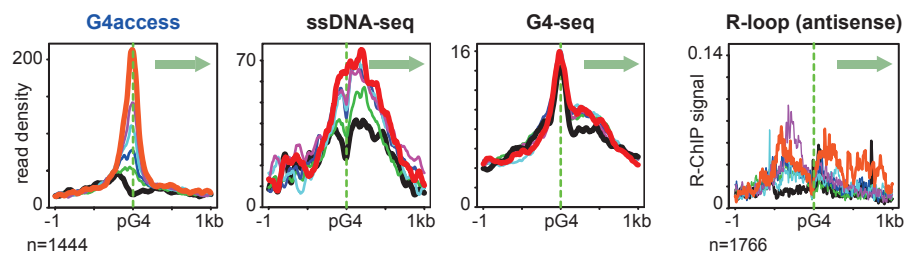
E

K562 cells



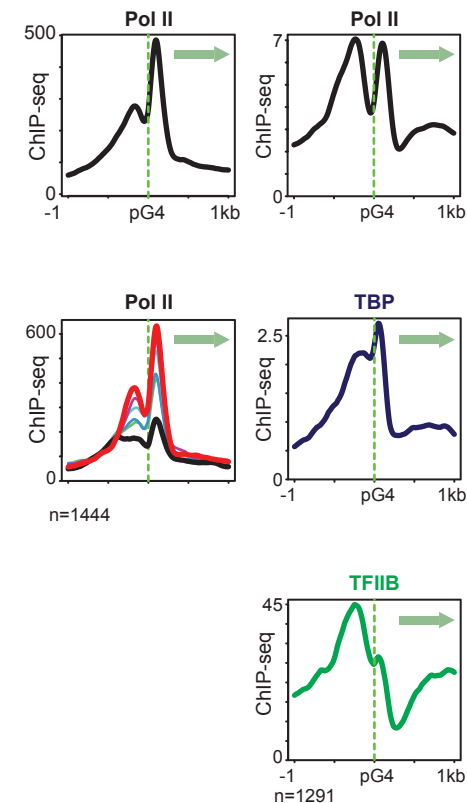
F

Raji cells



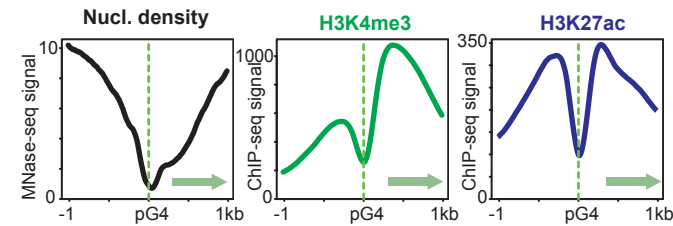
G

Raji cells Mouse primary T cells

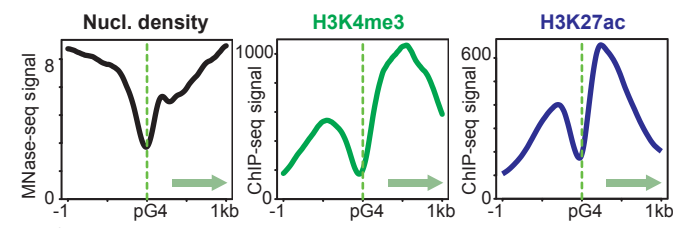


H

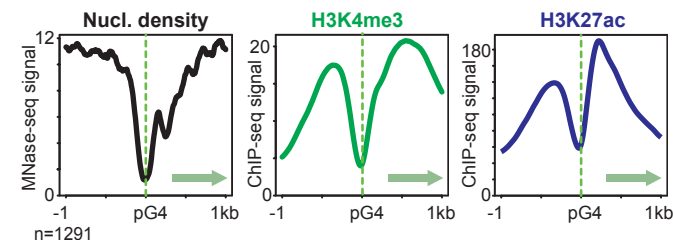
K562 cells



Raji cells

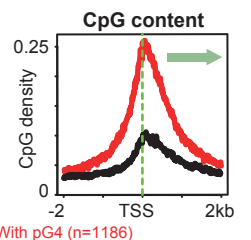
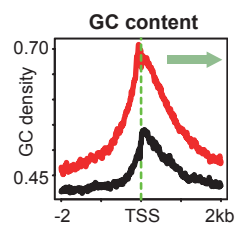
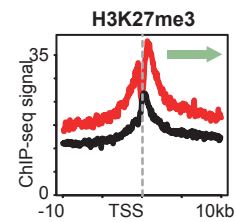


Mouse primary T cells



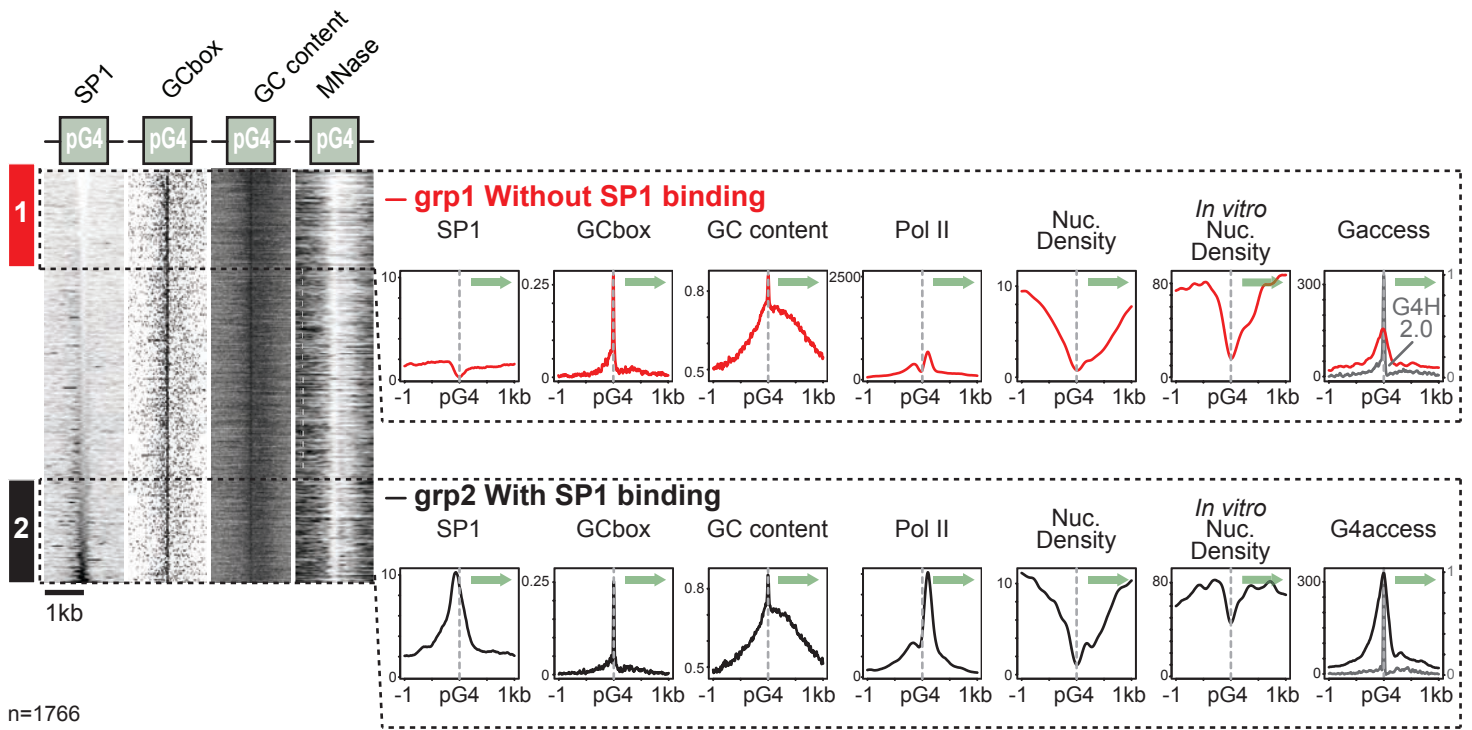
I

K562 inactive promoters



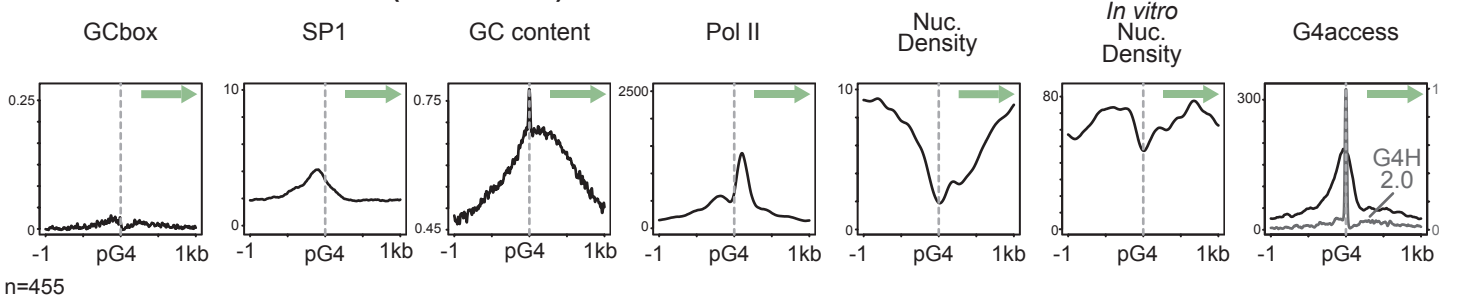
With pG4 (n=1186)
Without pG4 (n=2041)

A



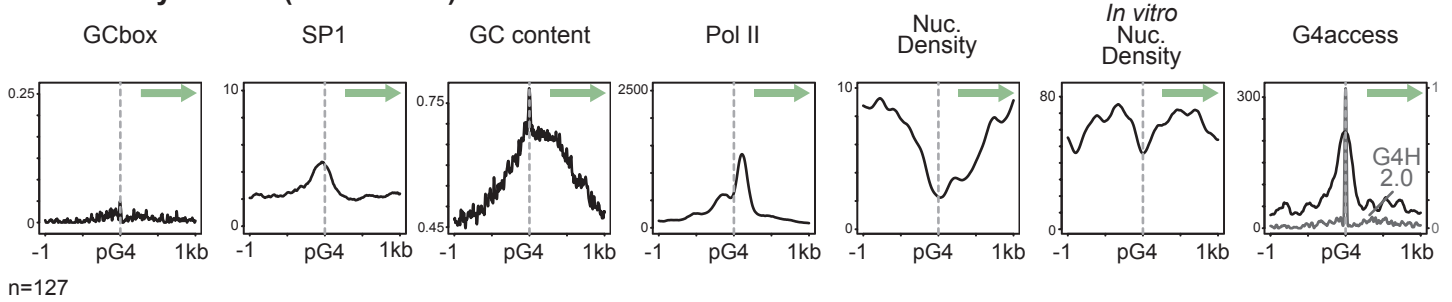
B

Without a canonical GCbox (GGGCGGG)

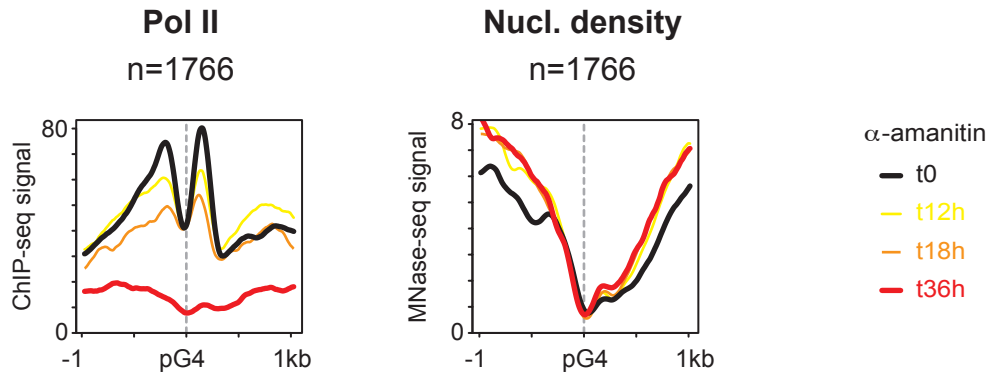


C

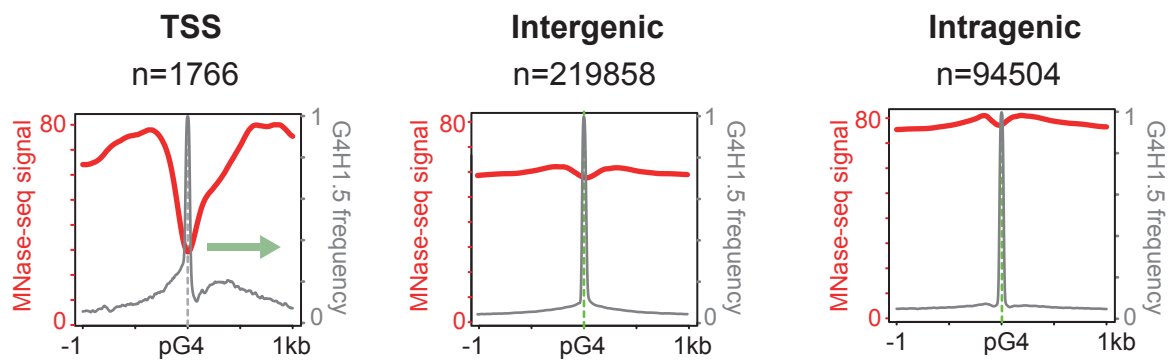
Without any GCbox (GGGNGGG)



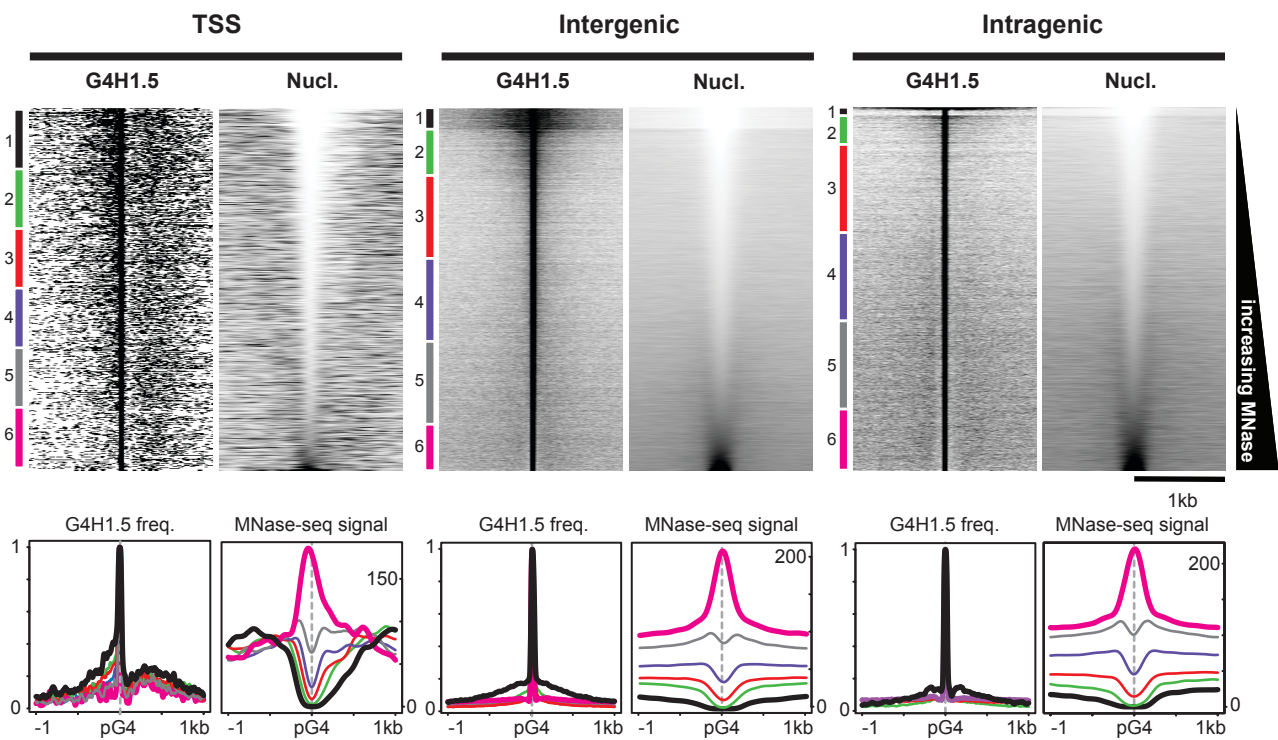
A



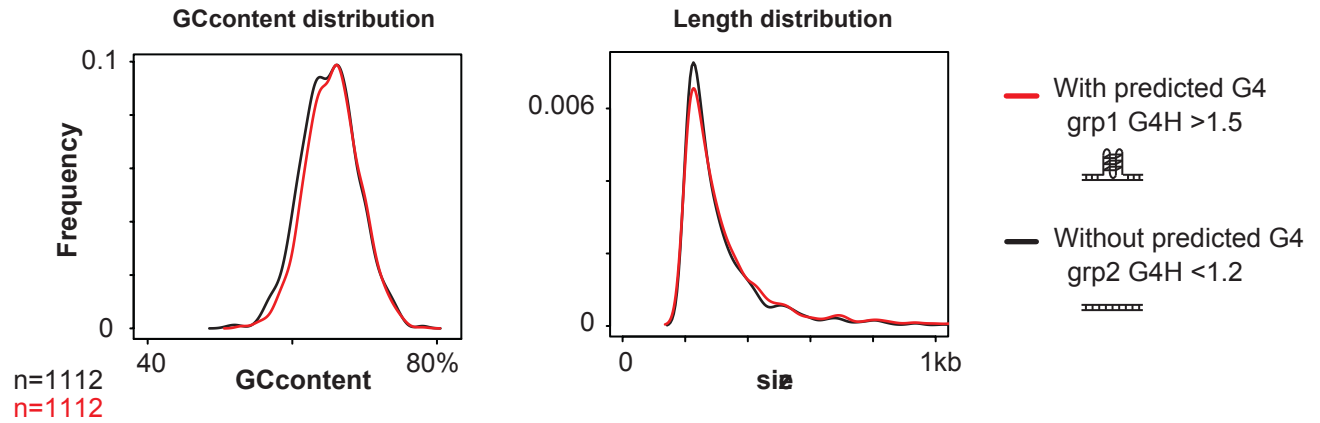
B



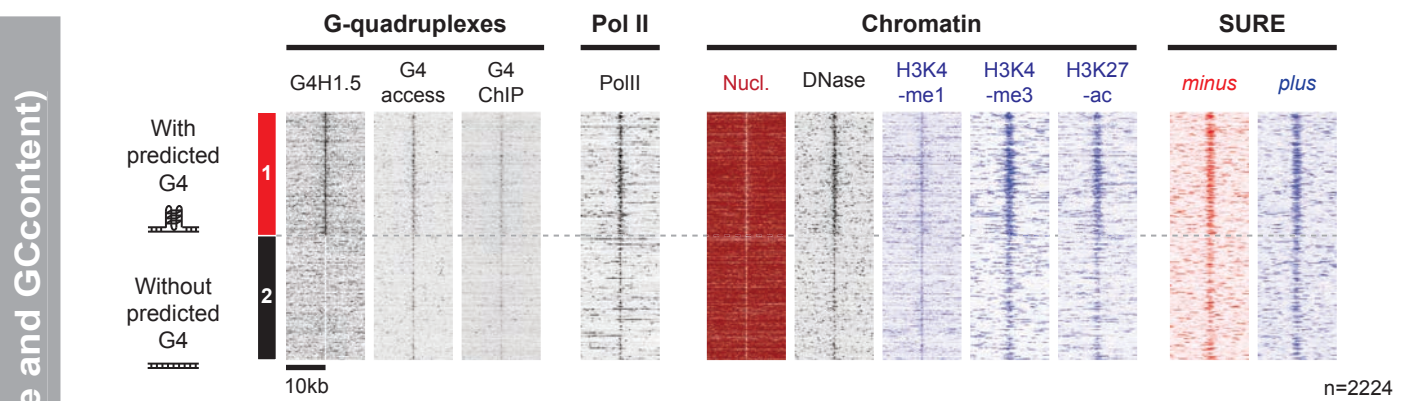
C



A



B



C

