

G-quadruplex DNA structures in human stem cells and differentiation

Katherine G. Zyner^{1,4}, Angela Simeone^{1,4}, Sean M. Flynn¹, Colm Doyle¹, Giovanni Marsico¹, Santosh Adhikari², Guillem Portella², David Tannahill¹ and Shankar Balasubramanian^{1,2,3},

Affiliations:

¹Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, United Kingdom.

²Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom.

³School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0SP, United Kingdom.

⁴These authors contributed equally to this work.

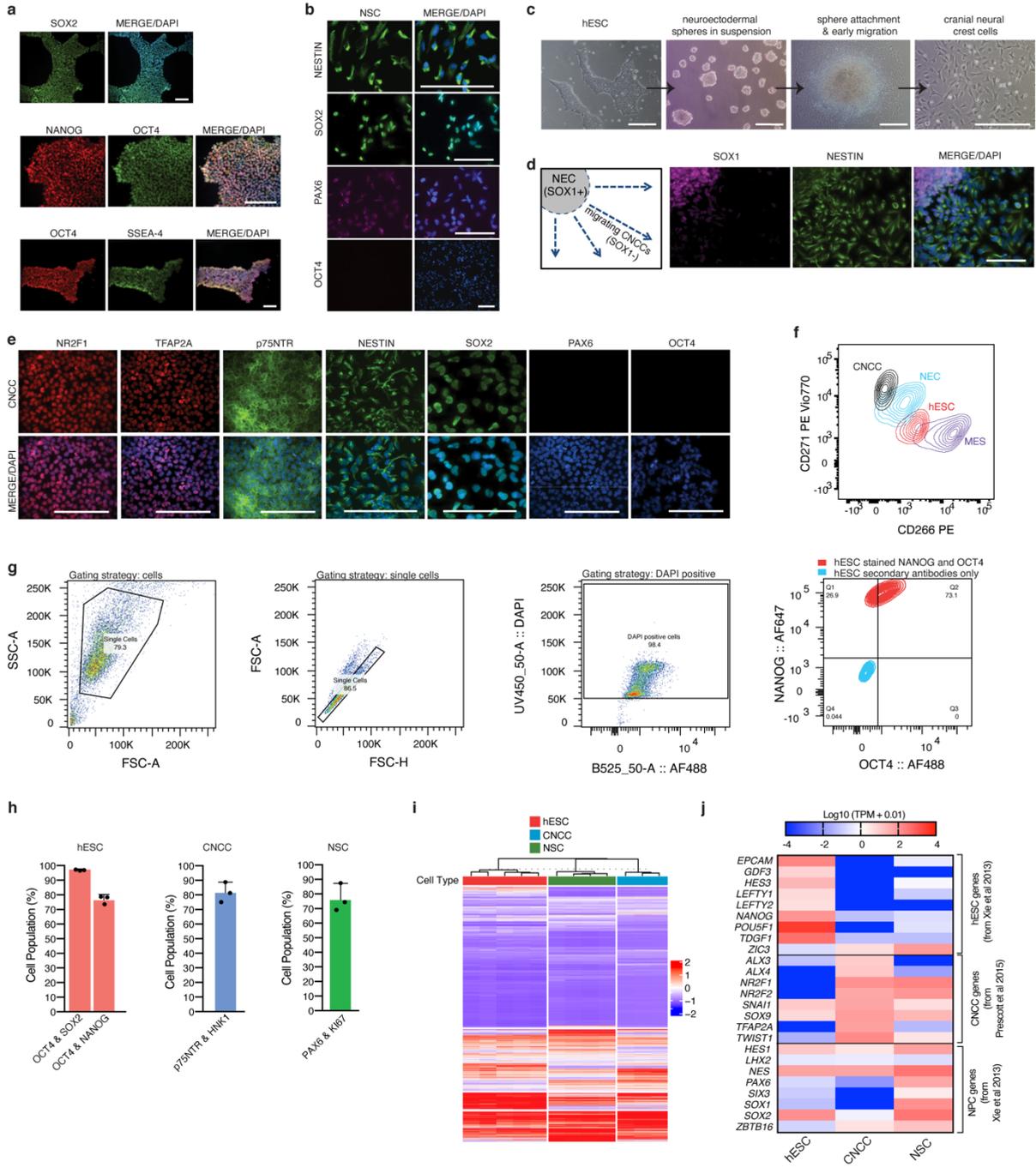
*Correspondence to: sb10031@cam.ac.uk

Inventory of Supplementary Information:

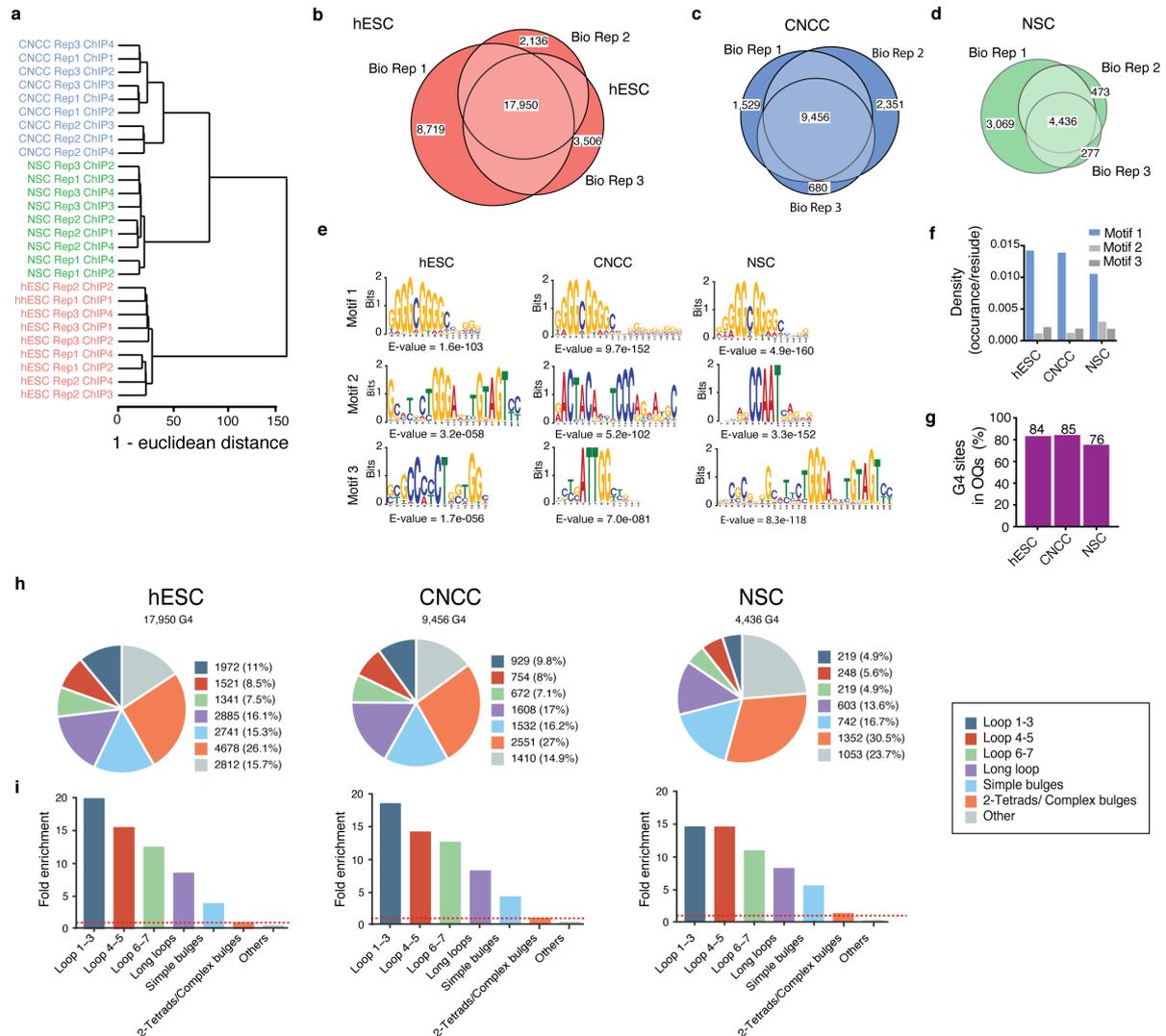
Supplementary Figures (Supplementary Fig. 1 to 19)

Supplementary Discussion

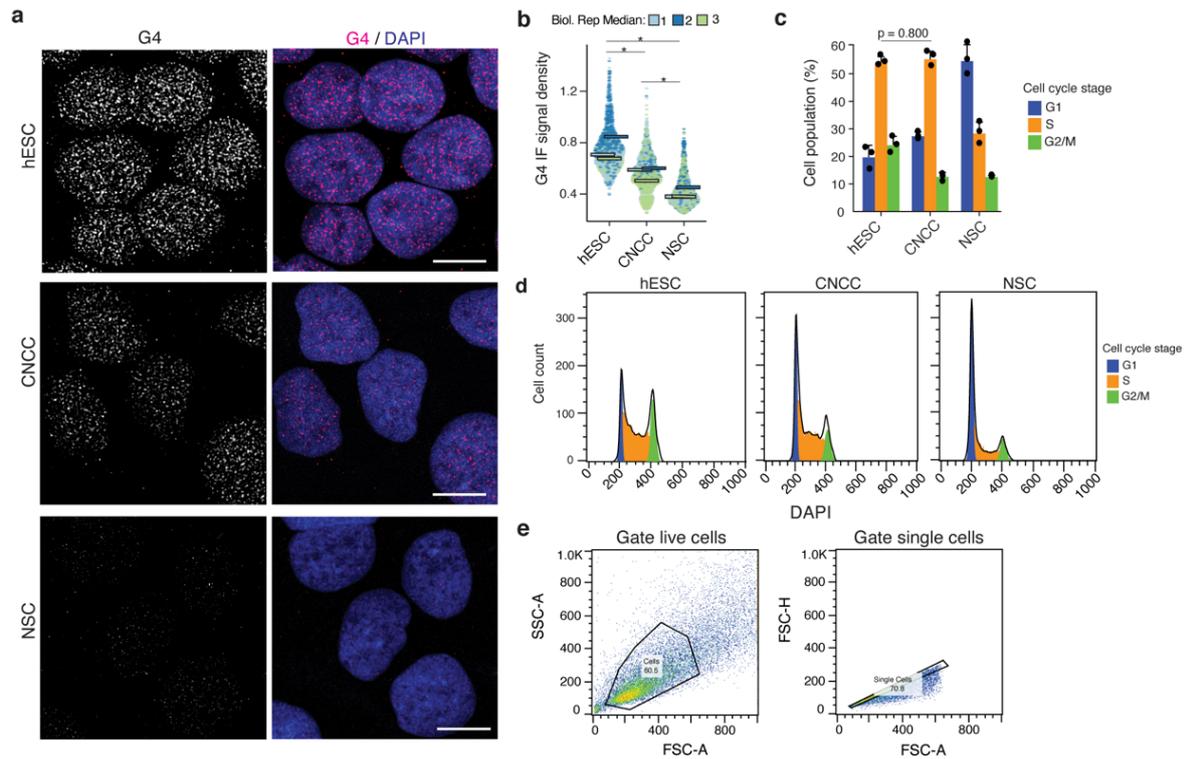
Supplementary References



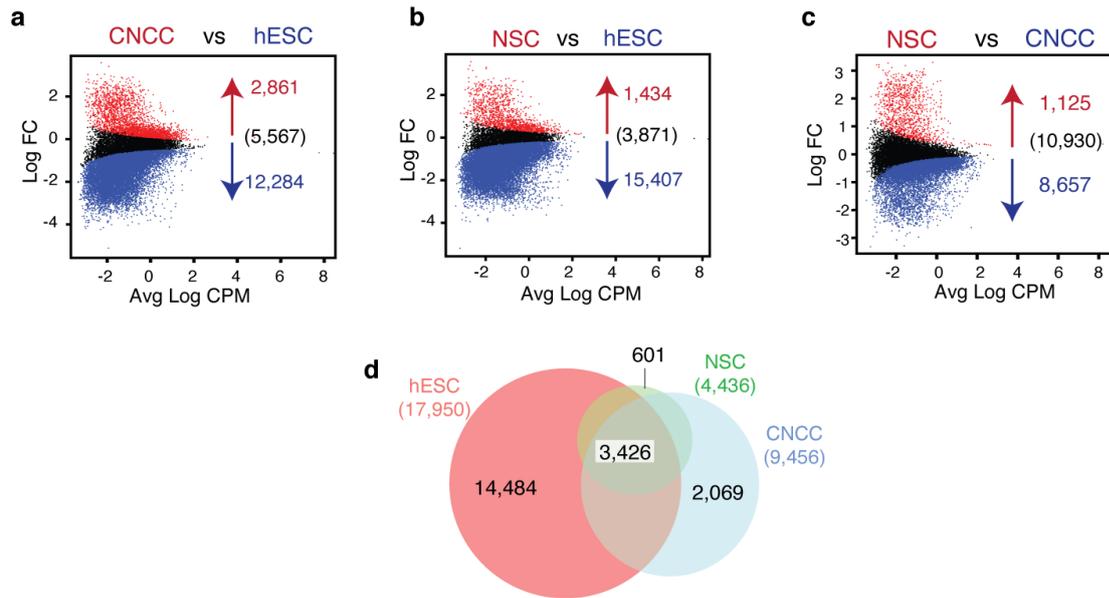
Supplementary Figure 1: Characterisation of human stem cell types. a-b, Representative immunofluorescent microscopy (IF) images of (a) hESCs and (b) NSCs stained for hESC markers: OCT4, NANOG, SOX2 and SSEA-4 (OCT4, NANOG IF data are also shown in Fig 1.c) and known NSC markers: SOX2, NESTIN and PAX6 (NESTIN and PAX6 IF data are also shown in Fig 1c). Scale bar = 100 μ m. **c,** Brightfield images of steps in derivation of CNCCs: hESCs are first differentiated into neuroectodermal spheres (NEC) which are then allowed to attach for the generation of migratory CNCCs. Scale bar = 250 μ m. **d,** Representative IF images showing NESTIN+/SOX1- (neural crest) cells migrating out from NESTIN+/SOX1+ neural rosettes. Scale bar = 100 μ m. **e,** Representative IF images of CNCCs stained with antibodies against neural crest markers NR2F1, TFAP2A, p75NTR, NESTIN, SOX2; hESC marker OCT4 and NSC marker PAX6 (TFAP2A and p75NTR IF data are also shown in Fig 1c). Scale bar = 100 μ m. **f,** Flow cytometry plots for cell surface markers CD271 and CD266 during derivation of CNCCs from hESCs, as well as further differentiation to mesenchymal lineages (in agreement with Prescott 2015¹). **g,** Example of gating strategy used for determining percentage stem cells positive for tested cell lineage markers (e.g., hESCs positive for NANOG and OCT4) via flow cytometry. **h,** Percentage of hESCs positive for OCT4 and NANOG or OCT4 and SOX2; NSCs positive for PAX6 and Ki67 and CNCCs positive for neural crest markers p75NTR and HNK1 via flow cytometry. Mean \pm SD from 3 independent biological replicates. **i,** Hierarchical clustering of genes based on z-score of expression values (median TPM). Row: genes, columns: individual stem cell libraries. **j,** Heatmap showing median gene expression (Log₁₀(TPM + 0.01)) for the indicated genes for hESCs², CNCCs¹ and neural progenitor cells (NPC) developmentally analogous to NSCs² for each stem cell line.



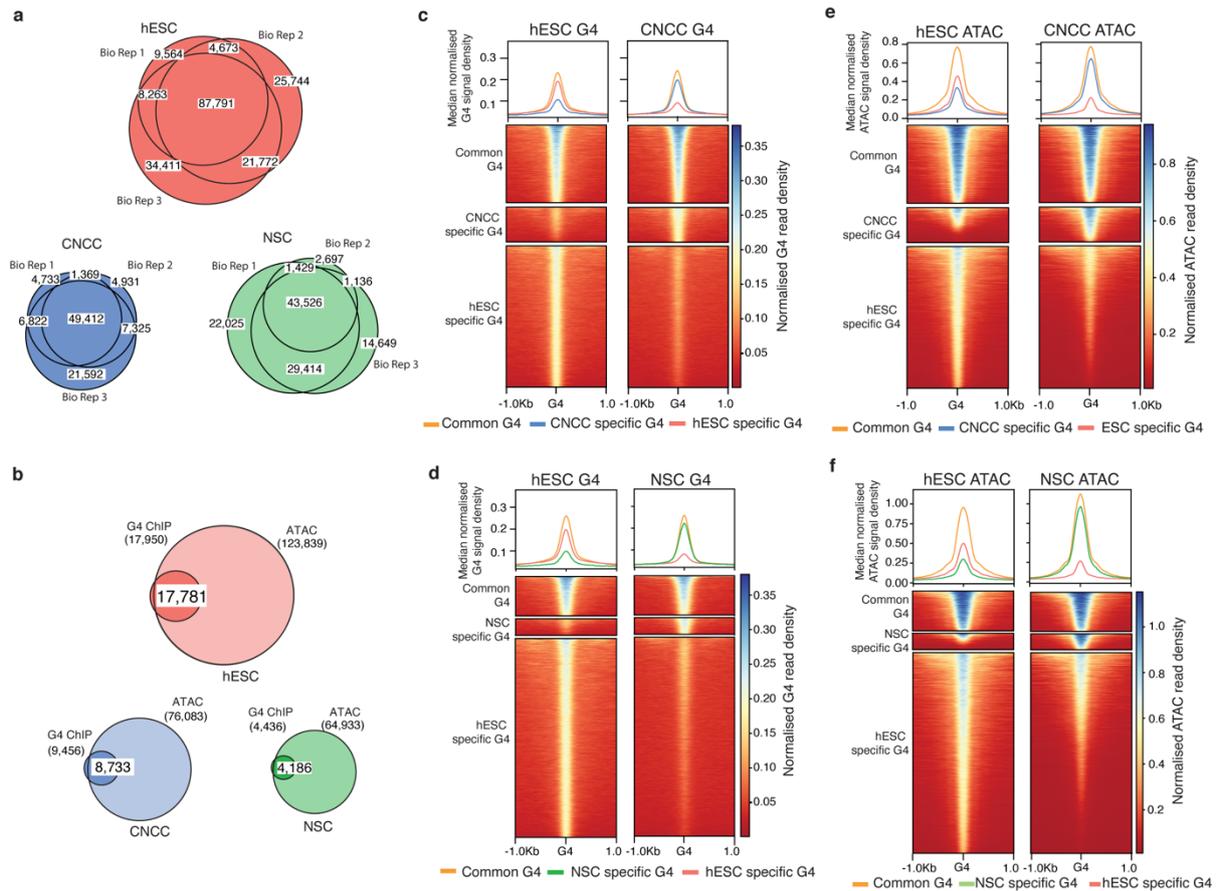
Supplementary Figure 2: Quality control for G4 ChIP-seq. **a**, Hierarchical clustering of reads per million (RPM) for G4s (G4 ChIP-seq libraries) generated for 3 biological replicates (i.e., Rep) with 3 technical replicates (i.e., ChIP) each for stem cell type. **b-d**, Area proportional Venn diagrams of G4 ChIP-seq peaks called by macs2 for (b) hESCs, (c) CNCCs and (d) NSCs for 3 biological replicates. Values given in the center correspond to the number of G4 ChIP-seq peaks in common for at least 2 out of 3 biological replicates, hereinafter called G4s. **e**, Top 3 sequence motifs identified by Multiple Em for motif elicitation (MEME)-ChIP³ in G4s for each stem cell type. **f**, Motif density (motif occurrence per residue) in G4s calculated via Find individual motif occurrences (FIMO)⁴. Only the G-rich motif showed an increase in density for G4s relative to other motifs. **g**, Percentage of G4s identified in sequences with the ability to fold into a G4 *in vitro* (observed quadruplex sequences; OQs)⁵. **h**, Enrichment analysis for different G4 structural classes. The total number of G4 regions in each structural class, per stem cell line are shown. Loop size 1–3, 4–5 and 6–7, indicates that at least one loop of this length is present in the G4; long loop indicates a G4 with any loop of length >7 (up to 12 for any loop and 21 for the middle loop); simple bulge indicates a G4 with a bulge of 1–7 bases in one G-run or multiple 1-base bulges; 2-tetrads / complex bulge indicates G4s with two G-bases per G-run or several bulges of 1–5 bases; and other indicates other G4-types that do not fall into the former categories. **i**, Fold enrichment for each structural class in (g) compared to random (average of 3 randomisations). Higher enrichment values equate to the higher likelihood of being present among the G4 regions.



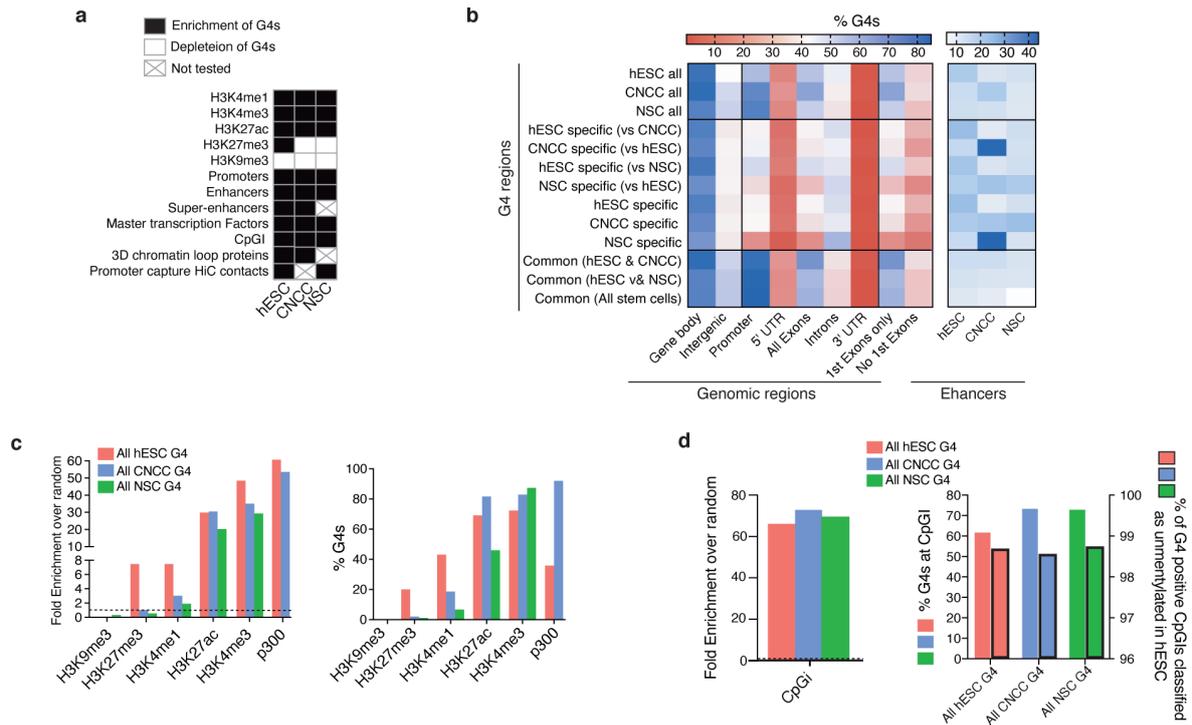
Supplementary Figure 3: G4 abundance decreases with loss of stem cell potency. a, Representative immunofluorescent microscopy images for hESCs, CNCCs and NSCs stained with the G4 structure-specific antibody BG4⁶ after pre-treatment with RNase A. Central slice of 13 slice confocal Z-stack (captured 0.3 μm apart) is shown. Left panels show G4 foci and right panels show G4 foci together with the nuclear counterstain DAPI (blue). Scale bar is 10 μm . **b,** Quantification of the total G4 signal density (sum G4 signal /sum DAPI signal) per nucleus calculated on the sum projection of the Z-stack. Each point represents a single nucleus, and the coloured horizontal bars represent the median from each biological replicate. $N = 3$ biologically independent experiments. Exact number of cells can be found in Source Data. *: $p < 2E-16$, one-sided Kolmogorov-Simonov test, per biological replicate. **c-d,** Cell cycle analysis of stem cell populations as determined by DAPI staining and flow cytometry. Quantification of the proportion of the cells in each phase of the cell cycle presented in (c) and representative cell cycle profiles are shown in (d). Bar graph shows Mean \pm SD. $N = 3$ independent biological replicates of 10,000 cells each. Statistical test: one-sided Mann-Whitney U test. **e,** Example of flow cytometry gating strategy used to generate histograms in (d).



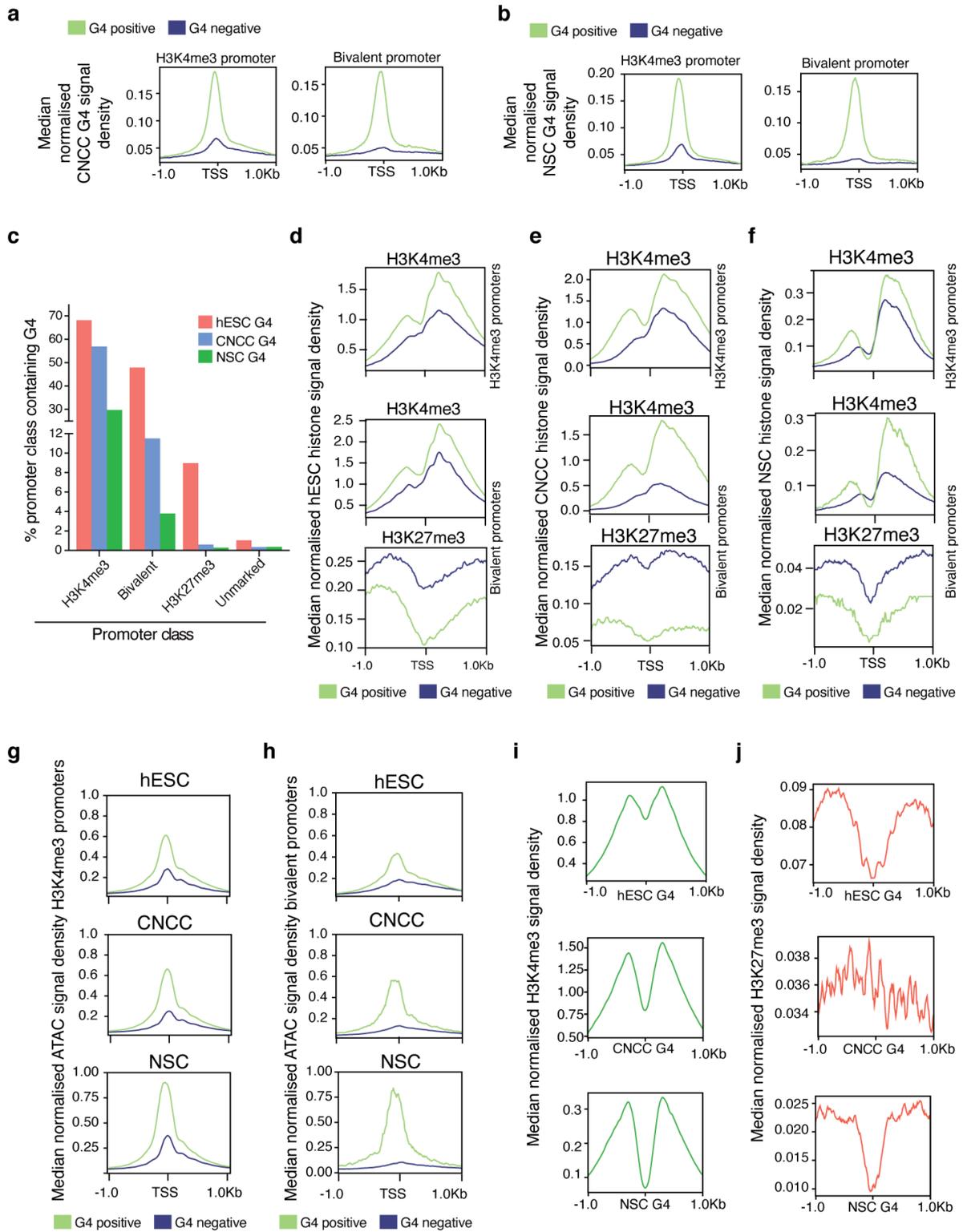
Supplementary Figure 4: Most CNCC and NSC G4s are found in hESC. **a-c**, MA plot showing differential binding analysis across 20,712 sites showing significant ($FDR < 0.05$) differences in G4 signal intensities between (a) hESCs and CNCCs, (b) hESCs and NSCs or (c) CNCCs and NSCs (\log_2 fold change vs \log_2 counts per million). 66% (5,567/8,428) of CNCC and 73% (3,871/ 5,305) of NSC G4s are in common with hESCs. 91% (10,930/12,055) of NSC G4s are in common with CNCCs. **d**, Intersection of G4s between all three stem cell types. Only G4s common to all three stem cells or unique to each stem cell type are shown.



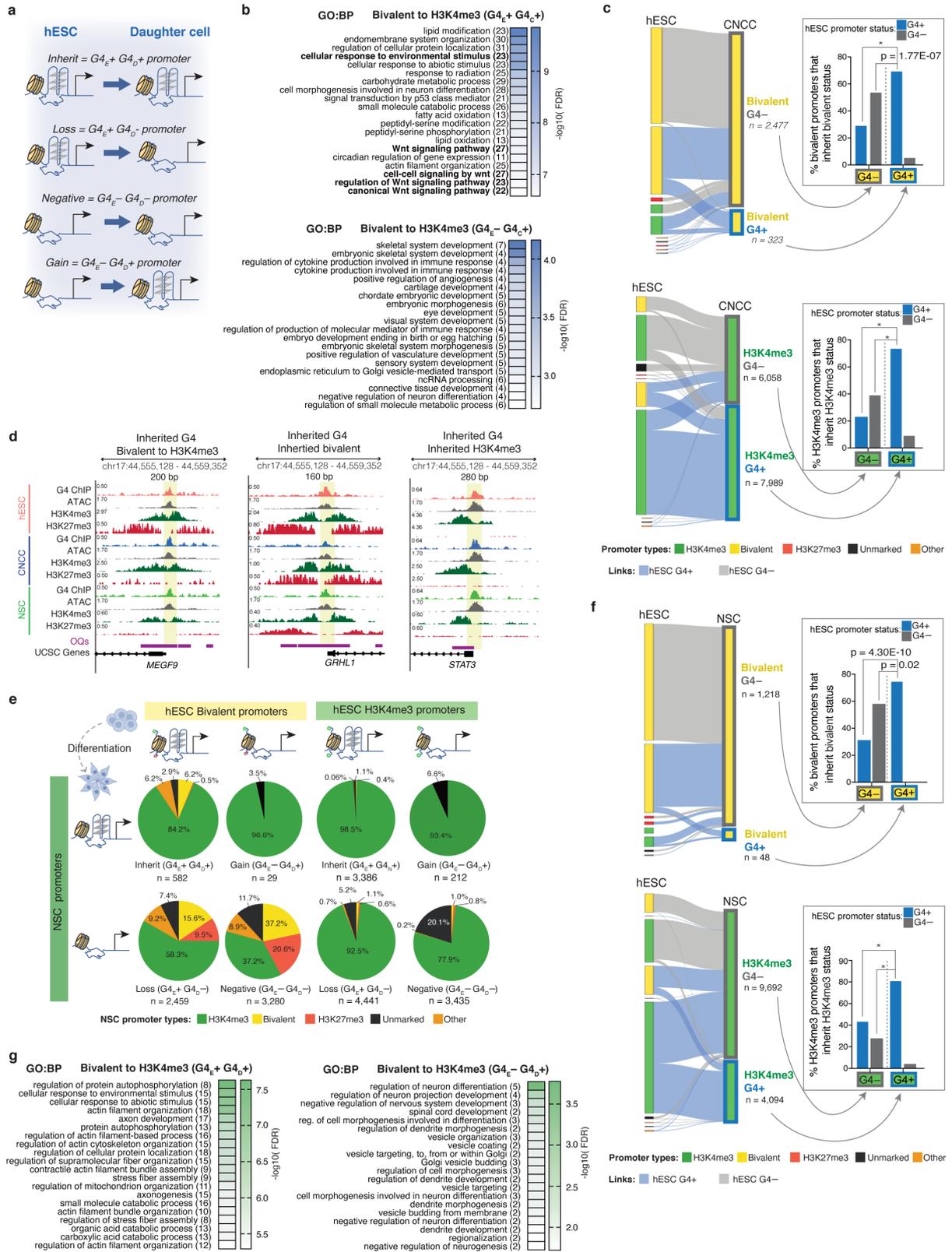
Supplementary Figure 5: G4 loss is linked to reduced chromatin accessibility during differentiation. **a-b**, Area proportional Venn diagrams of (a) ATAC-seq peaks (ATAC) called by macs2 for hESCs, CNCCs and NSCs for 3 biological replicates and (b) G4s overlapping ATAC sites in each stem cell type. **c-d**, Heatmaps showing normalised G4 read density across common or cell type-specific G4 sites derived for pairwise comparisons between (c) hESCs and CNCCs and (d) hESCs and NSCs. **e-f**, Heatmaps showing normalised ATAC-seq read density across common or cell type specific G4s derived from pairwise comparisons between (e) hESCs and CNCCs and (f) hESCs and NSCs.



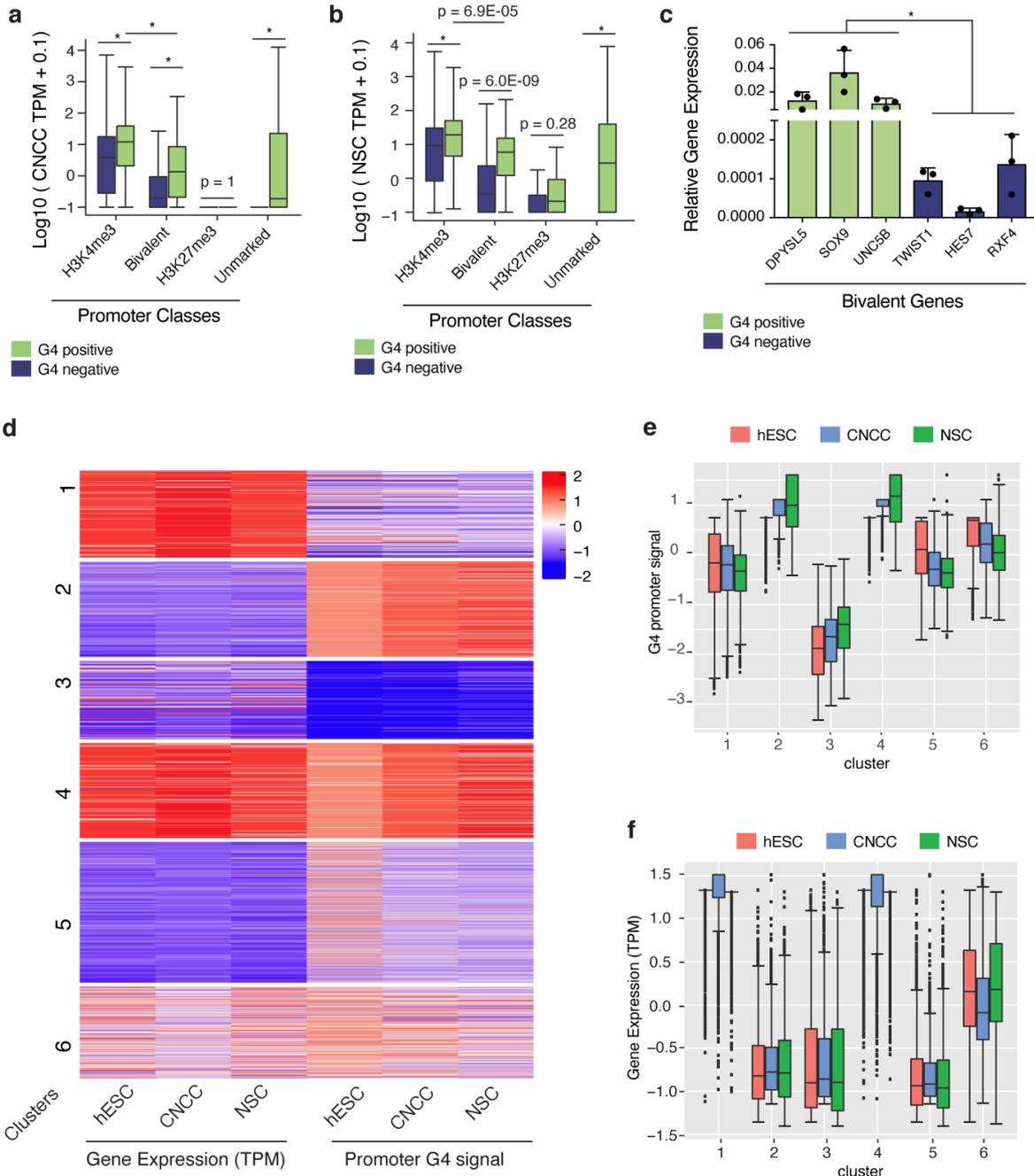
Supplementary Figure 6: G4s are found in stem cell regulatory elements. a, Summary table of genomic and epigenomic features compared to G4 presence. **b**, Proportion of G4s found at genomic features, as defined in Fig.1 (TSS, transcription start site), and enhancers. Rows; G4s and columns; genomic and epigenomic features. **c-d**, Fold-enrichments over random ($n = 1,000$ permutations) and proportion of G4s per stem cell type at (c) the histone marks and histone acetyltransferase (p300); and (d) CpG islands (CpGI, tracks from USCS genome browser). hESC hypomethylated CpG islands are defined as in Lee *et al* 2017⁷.



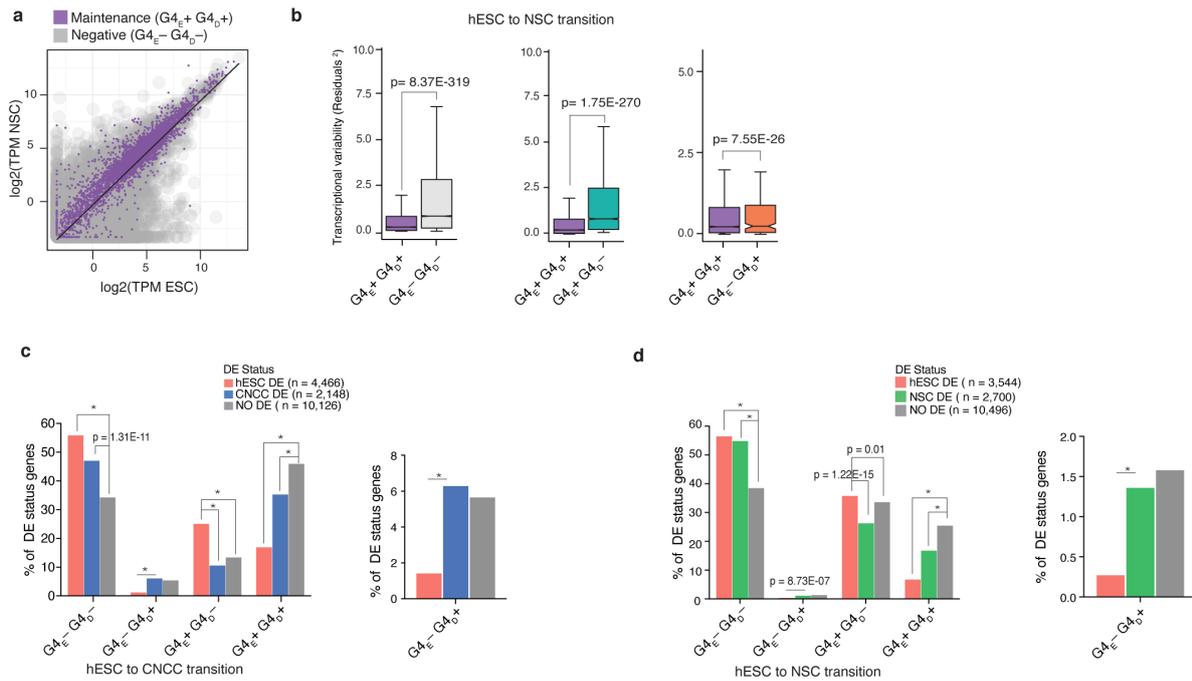
Supplementary Figure 7: G4s are a structural feature of bivalent and H3K4me3 promoters. **a-b**, Median normalised (a) CNCC or (b) NSC G4 ChIP-seq signal density at H3K4me3 and bivalent promoter regions. TSS, transcription start site. **c**, Proportion of all hESC, CNCC or NSC G4s overlapping histone modification promoter classes. **d-f**, Median normalised H3K4me3 read density across (d) hESC (also presented in Fig. 3), (e) CNCC and (f) NSC H3K4me3 promoters (top) or H3K4me3 and H3K4me27 read density across bivalent promoter regions (bottom) segregated by presence (green) or absence (dark blue) of G4. **g-h**, Median normalised ATAC-seq signal at (g) H3K4me3 and (h) bivalent promoters in hESCs (top, also presented in Fig. 3), CNCCs (middle) or NSCs (bottom). **i-j**, Median normalised ChIP-seq signal density of H3K4me3 (i) and H3K27me3 (j) at hESCs (top), CNCCs (middle) and NSCs (bottom) promoters centered around G4 sites.



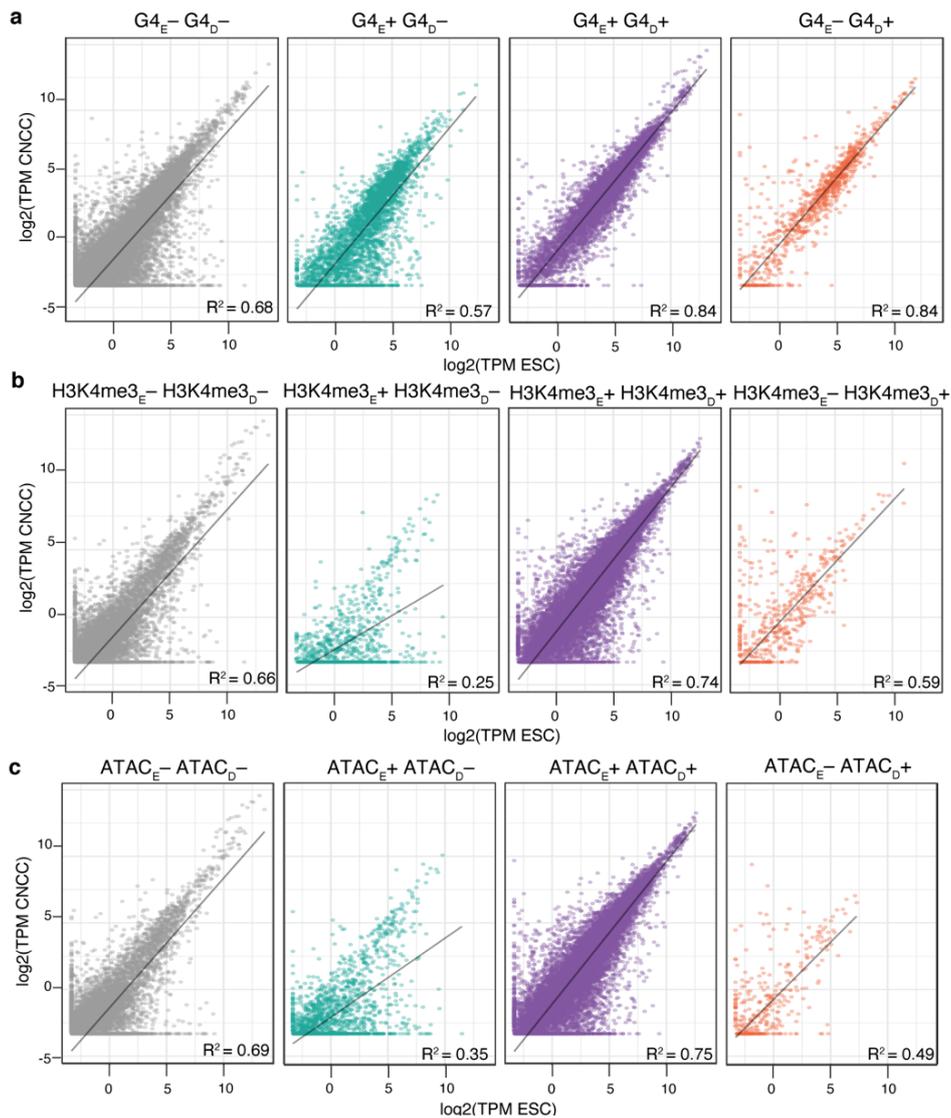
Supplementary Figure 8: G4s are associated with hESC H3K4me3 and bivalent promoters during differentiation. **a**, Schematic of G4 promoter signature: $G4_{E+}G4_{D+}$: G4 inheritance from hESCs to daughter cells; $G4_{E+}G4_{D-}$: G4 lost in daughter cells; $G4_{E-}G4_{D+}$: G4 gained in the daughter cells and $G4_{E-}G4_{D-}$: promoters lacking a G4s in both stem cells. **b**, Top 20 significant (FDR <0.05) Gene Ontology biological processes (GO: BP) obtained using g:Profiler, for genes with promoters that transition from bivalent to H3K4me3 and maintain or gain a G4 after differentiation of hESCs into CNCCs. The number of genes intersected is shown in brackets. Terms in bold are associated with Wnt signalling pathways. **c**, Alluvial plots showing origin of all CNCC bivalent (top) and H3K4me3 (bottom) promoters, classified by G4 presence or absence. Top inset, proportion of bivalent CNCC promoters that originate from hESC bivalent promoters, classified by G4 presence or absence. Bottom inset, similar quantification for CNCC H3K4me3 promoters. See Source Data for further details. *: $p < 2E-16$, one-sided Pearson's χ^2 test for proportions. **d**, Genome browser view for representative gene promoters demonstrating maintenance of G4 leading to transition from bivalent to H3K4me3 status (*MEGF*), maintenance of bivalent status (*GRHL1*) or maintenance of H3K4me3 status (*STAT3*) after hESC differentiation into CNCCs or NSCs. Yellow box highlights overlap of G4s, open chromatin sites (ATAC) and observed quadruplex sequences, (OQs: G4s confirmed to fold *in vitro*)⁵. **e**, Pie charts showing histone modification transitions for bivalent and H3K4me3 hESC promoters after differentiation to NSCs, segregated by G4 promoter signature. **f**, Same analysis as described in (b) for hESC to NSC transition. **g**, Same analysis as described in C for hESC to NSC transitions. See Source Data for further details. *: $p < 2E-16$, one-sided Pearson's χ^2 test for proportions.



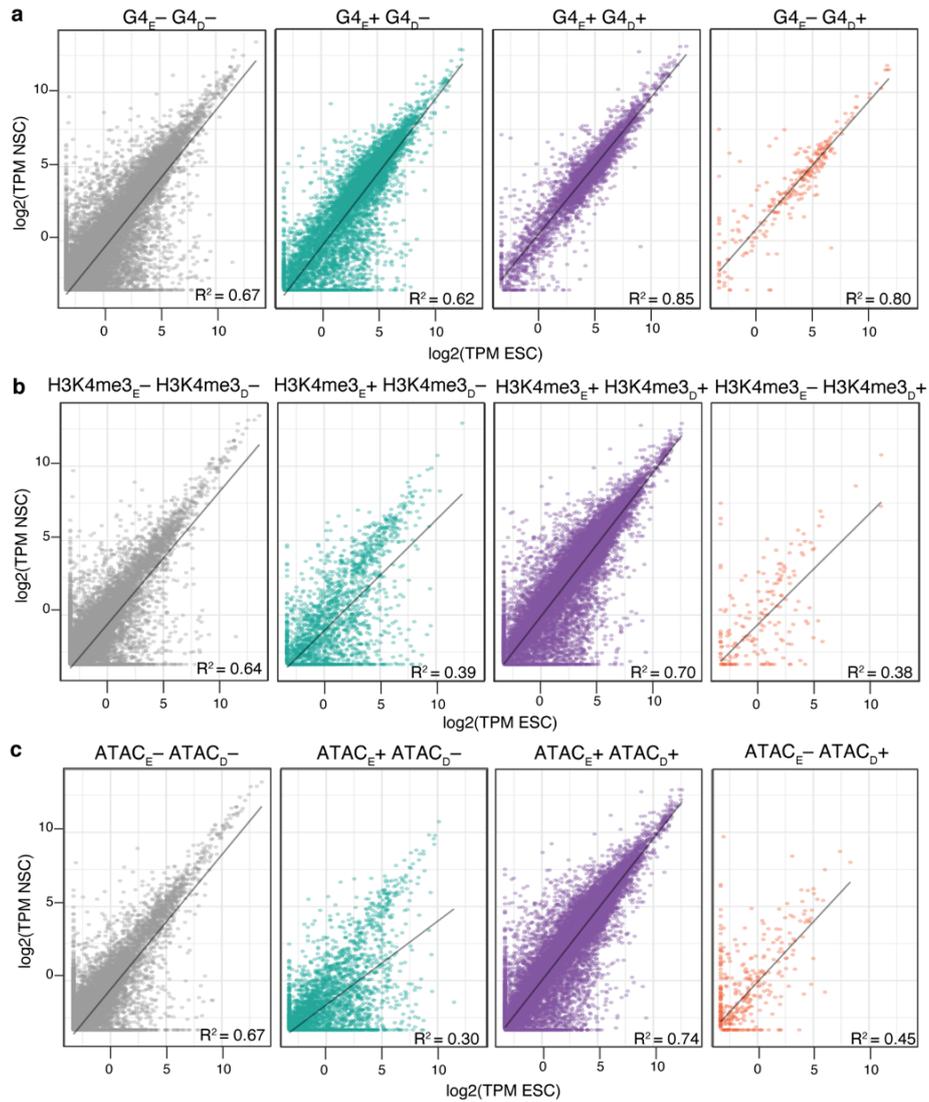
Supplementary Figure 9: Promoter G4s and gene expression. **a-b**, Gene expression levels ($\log_{10}[\text{average TPM} + 0.1]$, transcript per million averaged across biological replicates) for (a) CNCCs or (b) NSCs. Histone modification promoter classes as defined in Fig. 2B. G4-positive promoters (green) and G4-negative (dark blue). Data are presented as median (centre) and interquartile range (box; the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively). Whiskers represent $\pm 1.5x$ interquartile range. $N = 3$ and 4 biologically independent samples for CNCCs and NSCs, respectively. Number of genes per group: CNCC G4 positive (H3K4me3: 7989, bivalent: 323, H3K27me3: 10 and unmarked: 135), CNCC G4 negative (H3K4me3: 6058, bivalent: 2477, H3K27me3: 1657 and unmarked: 39347), NSC G4 positive (H3K4me3: 4094, bivalent: 48, H3K27me3: 4 and unmarked: 150), NSC G4 negative (H3K4me3: 9692, bivalent: 1218, H3K27me3: 1420 and unmarked: 41147). $*p < 2E-16$, one-sided Kolmogorov-Smirnov test. **c**, Relative expression levels (normalised to *GAPDH*) of hESC genes with bivalent promoters that are G4 positive (green) or G4 negative (dark blue). Mean \pm SD of $N = 3$ independent biological samples. $*$: $p < 0.05$ (unpaired two-sided t-test, see Source Data). **d**, K-medoid clustering ($K=6$) of genes based on z-score of expression level (median TPM) and G4 promoter signal; rows: promoters ($n = 11,876$), columns: individual stem cell types. $N = 5, 3$ and 4 biologically independent samples for hESCs, CNCCs and NSCs, respectively. Number of genes per cluster: cluster 1 (1754), cluster 2 (1929), cluster 3 (1577), cluster 4 (1920), cluster 5 (2846) and cluster 6 (1850). **e-f**, Box and whisker plots showing (e) z-score of G4 promoter signal and (f) z-score of gene transcription level for genes in each of the 6 clusters identified in (d). Data are presented as median (centre) and interquartile range (box; the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively). Whiskers represent $\pm 1.5x$ interquartile range. See Source Data.



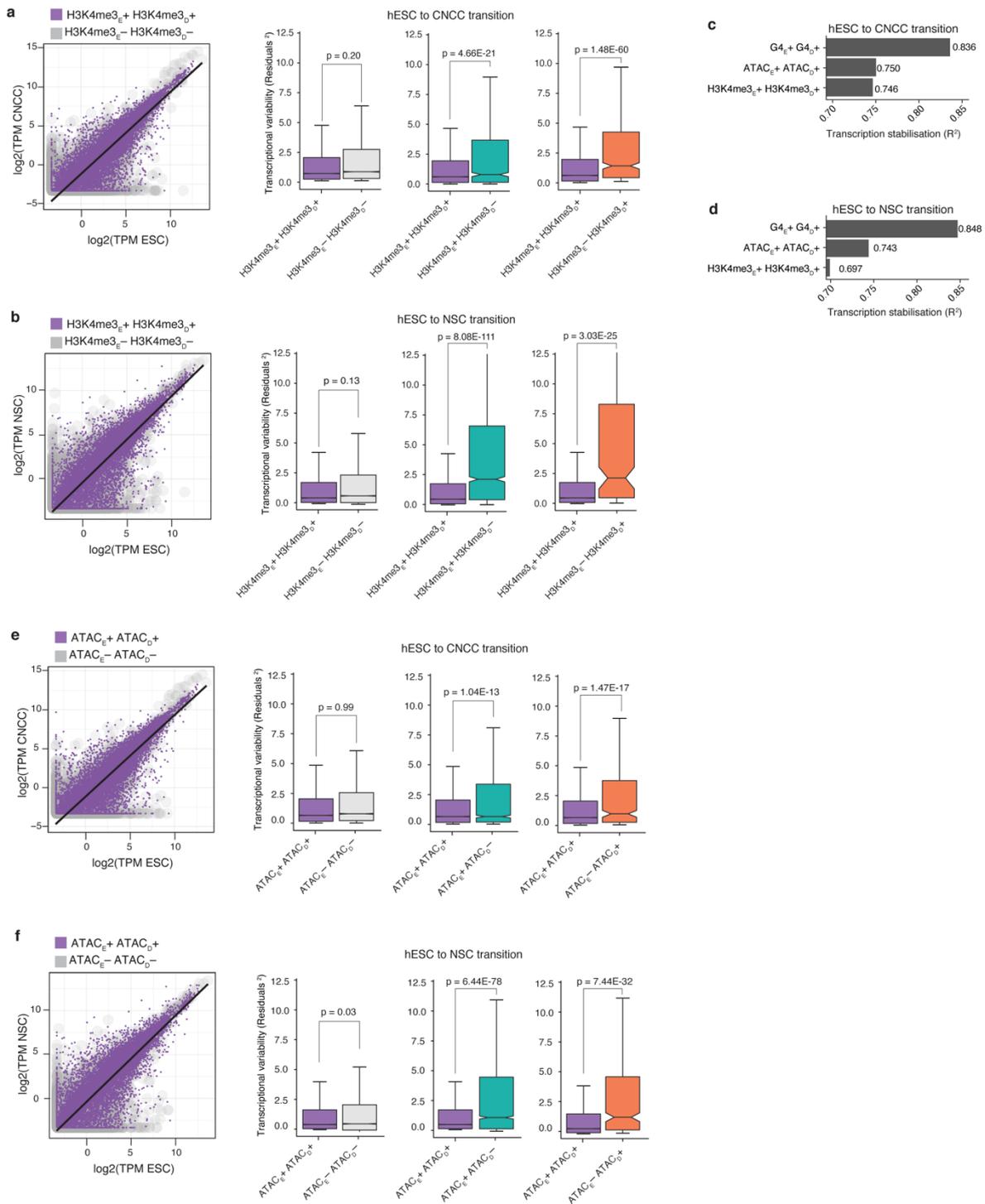
Supplementary Figure 10: G4s and differential gene expression changes during differentiation. **a**, Scatter plot of the log₂ median transcripts per million (TPM) for hESCs (x-axis) vs NSCs (y-axis) with fitted weighted linear regression model (black line). $G4_E$; $G4_D$ promoter status in hESCs and $G4_D$; $G4$ promoter status in daughter cell (NSCs). **b**, Box plots of the calculated residuals of $G4_E^+ G4_D^+$ and $G4_E^- G4_D^-$, $G4_E^+ G4_D^-$ or $G4_E^- G4_D^+$. Data are presented as median (centre) and interquartile range (box; the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively). Whiskers represent ± 1.5 x interquartile range. N = 5 and 4 biologically independent samples for hESCs and NSCs, respectively. Number of genes per group: $G4_E^+ G4_D^+$ (3849) and $G4_E^- G4_D^-$ (18698), $G4_E^+ G4_D^-$ (6838) or $G4_E^- G4_D^+$ (254). One-tailed F- test for variances. **c-d**, Proportion of differentially expressed (DE) genes categorised by G4 promoter signature for the comparison between (c) hESCs vs CNCCs and (d) hESCs vs NSCs. For each cell pair, genes were stratified into three groups: DE in hESCs ($\text{Log}_2\text{FC} < -1$, $\text{FDR} < 0.05$), DE in the daughter cell ($\text{Log}_2\text{FC} > 1$, $\text{FDR} < 0.05$) and not DE (NO DE). N = 5, 3 and 4 biologically independent samples for hESCs, CNCCs and NSCs, respectively. See Source Data for number of genes per group. *: $p < 2E-16$; one-sided Pearson's χ^2 test for proportions. Right graph in (c) and (d) is a zoomed in view for $G4_E^- G4_D^+$.



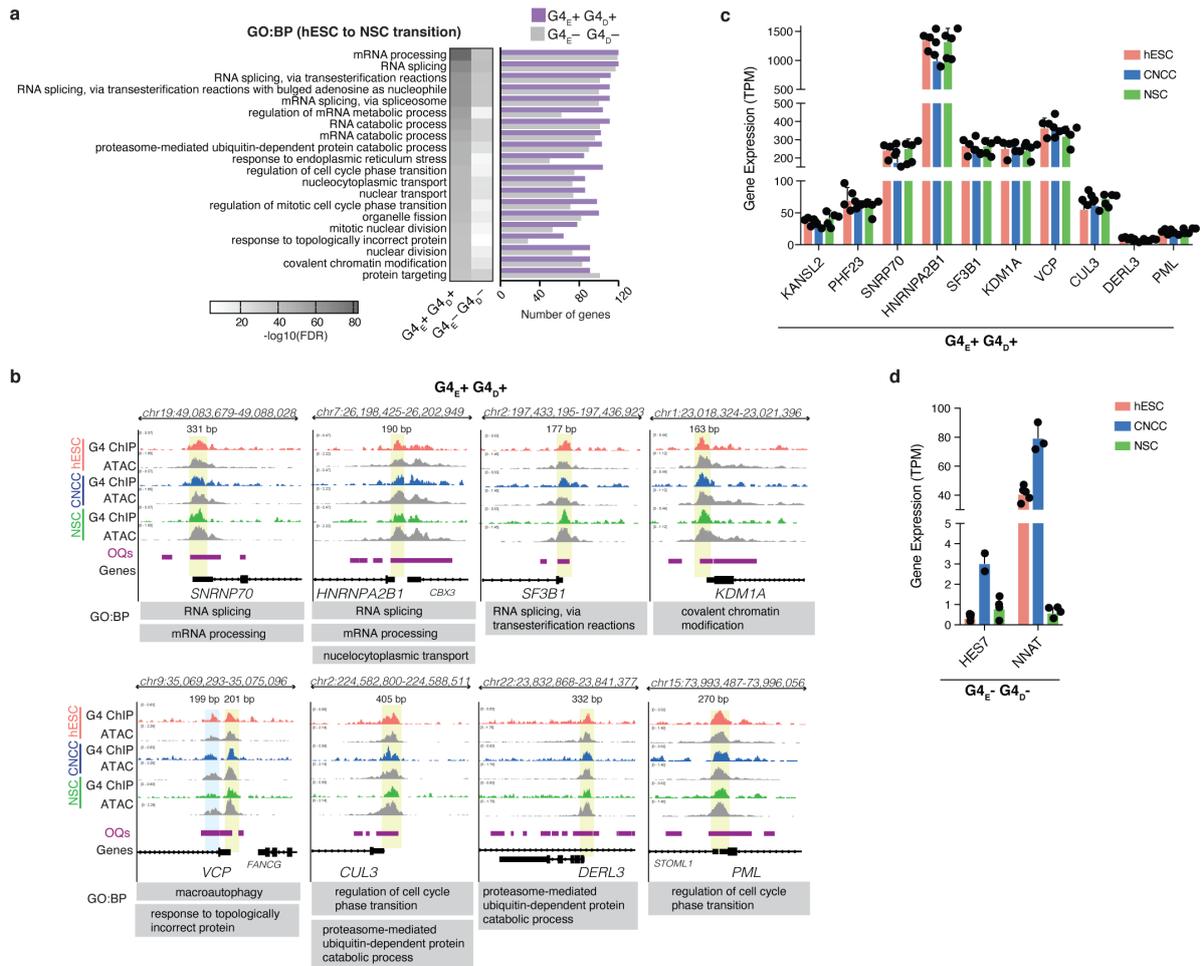
Supplementary Figure 11: Maintenance of G4s, ATAC or H3K4me3 promoter status is linked to transcriptional stabilisation during hESC to CNCC differentiation. a-c, Scatter plots of median transcripts per million (TPM) for hESCs vs CNCCs for genes categorised by (a) G4s, (b) H3K4me3 or (c) open chromatin (defined by ATAC-seq, ATAC) promoter classification. For each pairwise analysis, a weighted linear regression model was fitted to gene expression values (grey line) and the residuals computed to quantify transcriptional variation (see Methods). The coefficient of determination and goodness of fit, R^2 are displayed in the bottom right-hand corner.



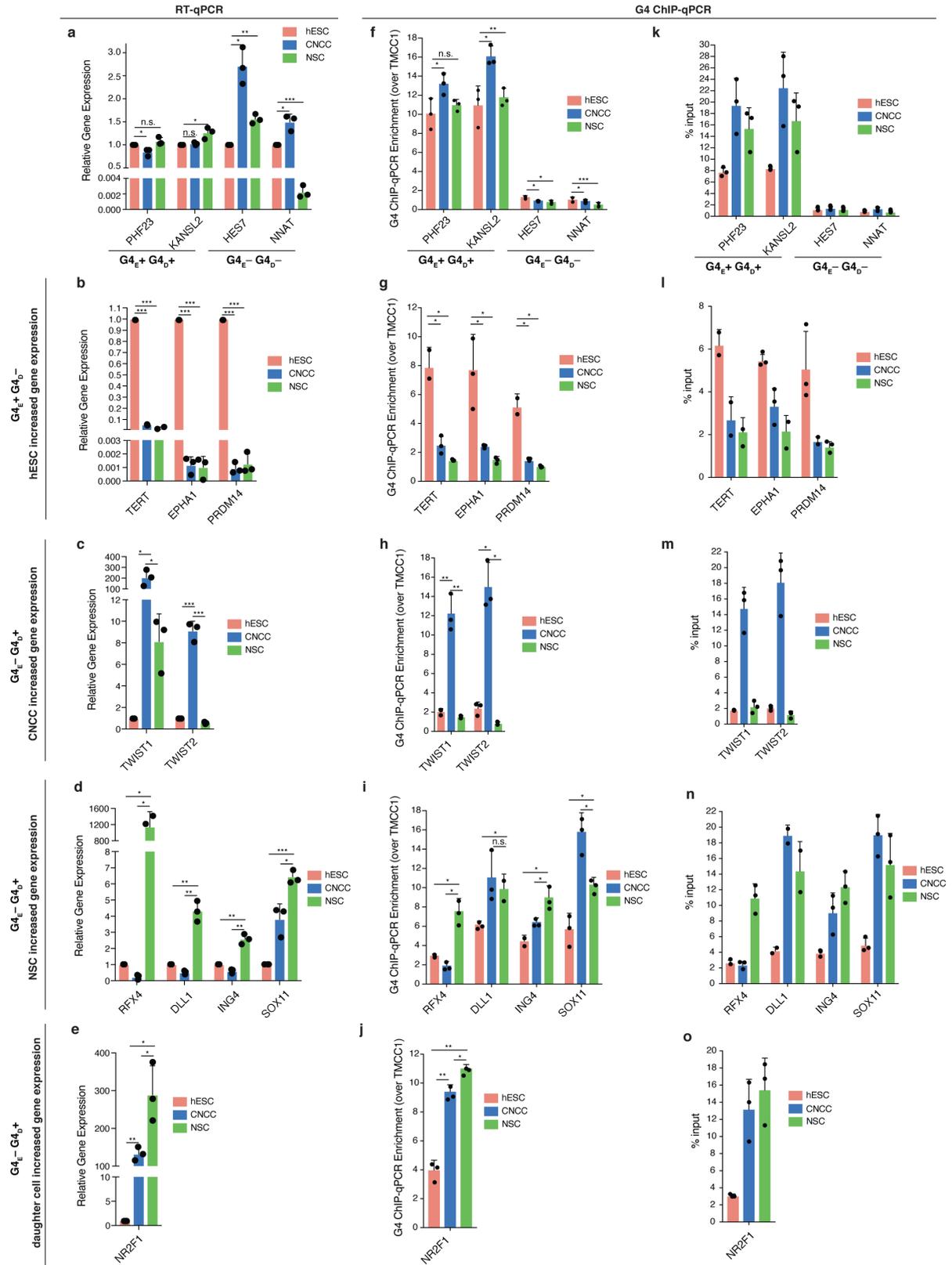
Supplementary Figure 12: Maintenance of G4s, ATAC or H3K4me3 promoter status is linked to transcriptional stabilisation during hESC to NSC differentiation. a-c, Scatter plots of median transcripts per million (TPM) for hESCs vs NSCs for genes categorised by (a) G4s, (b) H3K4me3 or (c) open chromatin (defined by ATAC-seq, ATAC) promoter classification. For each pairwise analysis, a weighted linear regression model was fitted to gene expression values (grey line) and the residuals computed to quantify transcriptional variation (see Methods). The coefficient of determination and goodness of fit, R^2 , are displayed in the bottom right-hand corner.



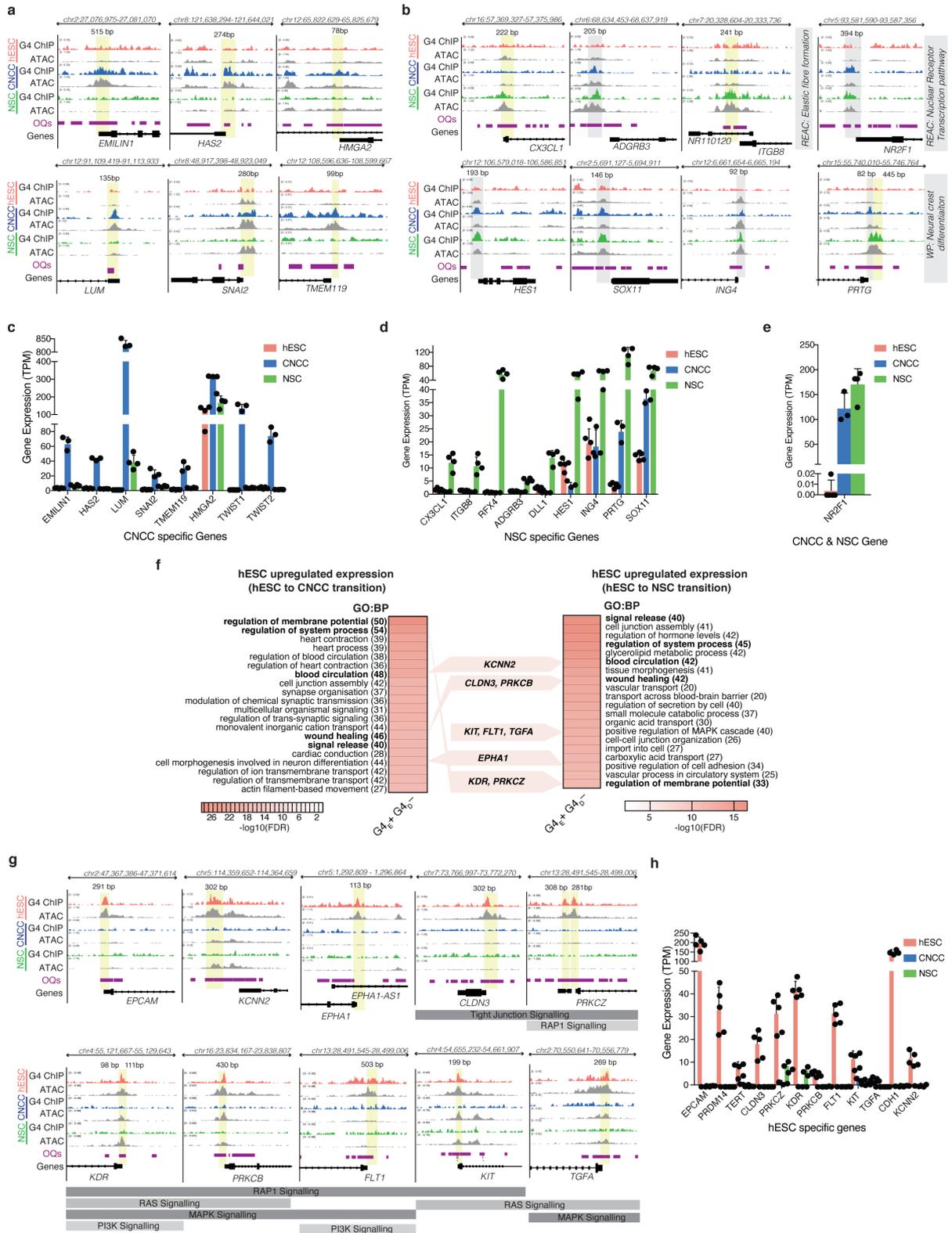
Supplementary Figure 13: Maintenance of G4s during differentiation results in greater transcriptional stabilisation compared to H3K4me3 or open chromatin. **a**, Scatter plot (left) of the log₂ median transcripts per million (TPM) for hESCs vs CNCCs with fitted weighted linear regression model (black line). H3K4me_{3E}; H3K4me₃ promoter status in hESC and H3K4me_{3D}; H3K4me₃ promoter status in daughter cells. Box plots (right) of the calculated residuals of H3K4me_{3E+} H3K4me_{3D+} and H3K4me_{3E+} H3K4me_{3D-}, H3K4me_{3E-} H3K4me_{3D+} or H3K4me_{3E-} H3K4me_{3D-}. Data are presented as median (centre) and interquartile range (box; the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively). Whiskers represent $\pm 1.5x$ interquartile range. N = 5 and 3 biologically independent samples for hESCs and CNCCs, respectively. Number of genes per group: H3K4me_{3E+} H3K4me_{3D+} (15100) and H3K4me_{3E-} H3K4me_{3D-} (11619), H3K4me_{3E+} H3K4me_{3D-} (1059) or H3K4me_{3E-} H3K4me_{3D+} (506). One-tailed F- test for variances. **b**, Same analysis as described in (a) for hESC vs NSCs. N = 5 and 4 biologically independent samples for hESCs and NSCs, respectively. Number of genes per group: H3K4me_{3E+} H3K4me_{3D+} (14459) and H3K4me_{3E-} H3K4me_{3D-} (13036), H3K4me_{3E+} H3K4me_{3D-} (1962) or H3K4me_{3E-} H3K4me_{3D+} (182). **c-d**, Bar chart comparing transcriptional stabilisation, as determined by the coefficient of determination and goodness of fit, R², for genes which either maintain G4 (G4_{E+} G4_{D+}), chromatin accessibility (defined by ATAC-seq, ATAC_{E+} ATAC_{D+}) or H3K4me₃ status (H3K4me_{3E+} H3K4me_{3D+}) of their promoter upon (c) hESC to NSC or (d) hESC to CNCC differentiation. **e**, Scatter plot (left) of the log₂ median transcripts per million (TPM) for hESCs vs CNCCs with fitted weighted linear regression model (black line). ATAC_E; ATAC promoter status in hESC and ATAC_D; ATAC promoter status in daughter cells. Box and whisker plots (right) of the calculated residuals for ATAC promoter classifications. Data are presented as median (centre) and interquartile range (box; the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively). Whiskers represent $\pm 1.5x$ interquartile range. N = 5 and 3 biologically independent samples for hESCs and CNCCs, respectively. Number of genes per group: ATAC_{E+} ATAC_{D+} (16010) and ATAC_{E-} ATAC_{D-} (9869), ATAC_{E+} ATAC_{D-} (2034) or ATAC_{E-} ATAC_{D+} (371). One-tailed F- test for variances. **f**, Same analysis as described in (e) for hESC vs NSCs. N = 5 and 4 biologically independent samples for hESCs and NSCs, respectively. Number of genes per group: ATAC_{E+} ATAC_{D+} (15407) and ATAC_{E-} ATAC_{D-} (10856), ATAC_{E+} ATAC_{D-} (3015) or ATAC_{E-} ATAC_{D+} (361). One-tailed F- test for variances. See Supplementary Fig. 11-12 for full set of scatter plots used to calculate residuals.



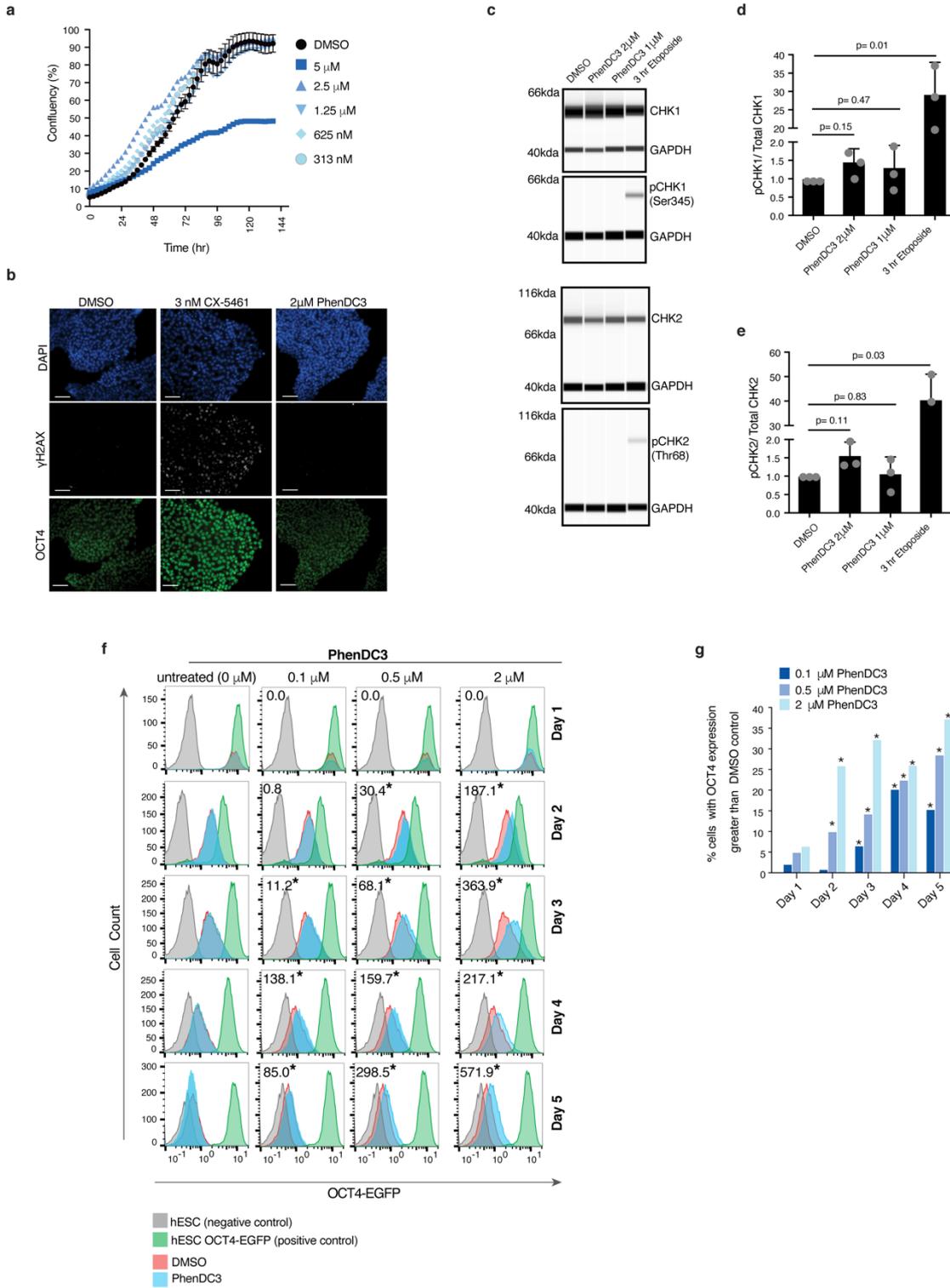
Supplementary Figure 14: Gene enrichment analysis for genes with a promoter $G4$ in both hESCs and daughter cells. **a**, Top 20 significant (FDR < 0.05) biological processes obtained from Gene Ontology enrichment analysis (GO:BP) performed with g:Profiler⁸ for genes that are not differentially expressed between hESCs and NSCs and maintain ($G4_{E+}G4_{D+}$) a $G4$ ($n = 2,696$ genes) or never ($G4_{E-}G4_{D-}$) had a $G4$ ($n = 4,073$ genes) in their promoter. Number of genes intersecting GO term shown in adjacent bar plot. **b**, Genome browser view for representative gene promoters demonstrating $G4_{E+}G4_{D+}$ promoter status after hESC differentiation into CNCCs or NSCs. Yellow shading highlights overlap of $G4$ s, open chromatin sites (ATAC) and $G4$ s that form *in vitro* (observed quadruplex sequences, OQs)⁵. **c-d**, Gene expression levels (transcript per million, TPM) for genes (as shown in **b** and Fig 4.e) which (c) maintain a $G4$ ($G4_{E+}G4_{D+}$) or (d) never had a $G4$ in their promoter. Mean \pm SD, $N = 5, 3$ and 4 biologically independent samples for hESCs, CNCCs and NSCs, respectively.



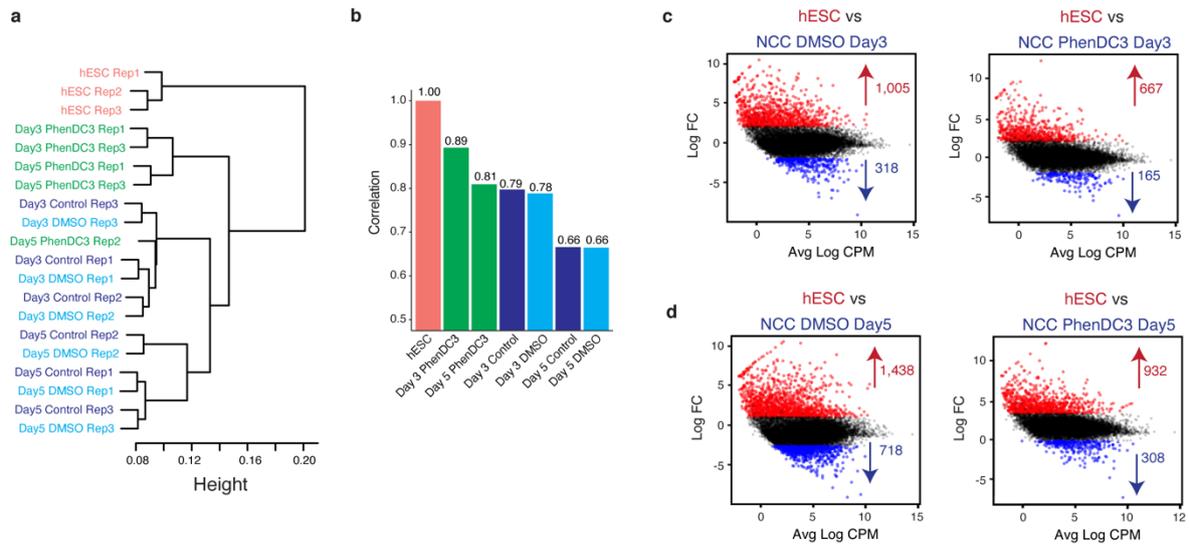
Supplementary Figure 15: RT-qPCR and G4-ChIP qPCR validation of selected genes uncovered by gene enrichment analyses. a-e, Relative gene expression levels (normalised to GAPDH and relative to hESCs) and **f-o**, G4 ChIP-qPCR of a selection of genes shown in Fig 4e; Fig 5 b,d,f; Supplementary Fig. 15. (f-j) The G4-negative *TMCC1* regulatory region⁹ was used as internal reference to normalise G4-ChIP qPCR promoter enrichments. Mean \pm SD, N = 3 biologically independent samples (see methods). Unpaired t-test, two-sided: *:p <0.05; **:p<0.001, ***: p <0.0001. See Source Data for exact p-values and Supplementary Data 4 for PCR primer sequences.



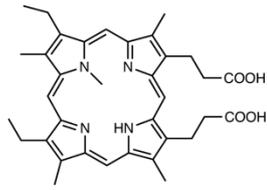
Supplementary Figure 16: Genome browser views for selected genes uncovered by gene enrichment analysis that maintain a G4 promoter during differentiation of hESCs to daughters. **a-b**, Genome browser view for representative gene promoters demonstrating G4_{E-} G4_{D+} promoter status after hESC differentiation into (a) CNCCs or (b) NSCs from enriched GO: BP terms (see Fig. 5 a and c). Yellow shading highlights location of cell specific G4 across open chromatin sites (ATAC) and G4s that can form *in vitro* (observed quadruplex sequences, OQs)⁵. Grey shading highlights G4 common to both CNCCs and NSCs but absent in hESCs. **c-e**, Gene expression levels (transcript per million, TPM) for genes shown in a, b and Fig 5 a and d. Mean \pm SD, N = 5, 3 and 4 biologically independent samples for hESCs, CNCCs and NSCs, respectively (see methods). **f**, Top 20 significant (FDR < 0.05) GO: BP terms for genes which have increased expression in hESC compared to CNCC (n = 1,128 genes) or NSC (n = 1,175 genes) which lose a promoter G4 (G4_{E+} G4_{D-}). Number of genes in intersection for each term shown in brackets. See Supplementary Data 2 and 3 for full list of terms. **g**, Genome browser view for representative gene promoters demonstrating G4_{E+} G4_{D+} promoter status after hESC differentiation into CNCCs or NSCs. Yellow box highlights hESC specific G4s and their overlap with open chromatin sites (ATAC) and G4s that can form *in vitro* (OQs)⁵. **h**, Gene expression levels (transcript per million, TPM) for genes shown in g and Fig 5f which lose a G4 (G4_{E+} G4_{D-}) G4 in their promoter. Mean \pm SD, N = 5, 3 and 4 biologically independent samples for hESCs, CNCCs and NSCs, respectively (see methods).



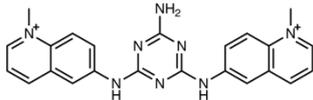
Supplementary Figure 17: Differentiation delay with PhenDC3 treatment is not due to growth arrest or induction of DNA damage. **a**, Growth curves for OCT4-EGFP hESCs grown in mTESR1 media on Matrigel over 144 hours in the presence of 0.02% DMSO (black) or the G4-specific ligand PhenDC3 at indicated concentrations. Error bars represent the Mean \pm SD of measurements from 3 independent wells. **b**, Representative IF images of OCT4-EGFP hESCs stained with antibodies against γ H2AX and OCT4 after 5 day treatment with 0.02% DMSO as negative control, the G4 ligand CX-5461 known to induce DNA damage¹⁰ as positive control or PhenDC3. Scale Bar = 100 μ m. **c**, hESCs were differentiated to CNCCs in the presence of PhenDC3 (1 μ M and 2 μ M) or 0.2% DMSO. On day 5, cell lysates were probed with antibodies against CHK1, phosphorylated CHK1(Ser354), CHK2, phosphorylated CHK2 (Thr68) and GAPDH control by western blotting. **d-e**, Normalised protein levels for d) pCHK1 and e) pCHK2. N = 3 independent biological replicates (Mean \pm SD), Welch's t-test, two-tailed. hESCs treated with 100 nM etoposide for 3 hours was used as a positive control. **f-g**, OCT4-EGFP expression levels over 5-day CNCC differentiation obtained from live cell flow cytometry. (f) FACs histogram plots and (g) calculated proportion of cells with higher levels of OCT4-EGFP expression compared to DMSO controls. χ^2 values, calculated using the Chi squared T(x) test in 'compare populations' using FlowJo software, for each PhenDC3 vs DMSO comparison are shown in top-left of each FACs histogram. *: comparisons are statistically different ($p < 0.01$, T(x) metric) compared to untreated (0 μ M PhenDC3 and no DMSO) vs 0.02% DMSO (see Methods). N = 10,000 cells analysed per day (apart from day 1 where the whole sample was analysed (5,000 cells)). N = one independent biological replicate. See Source Data.



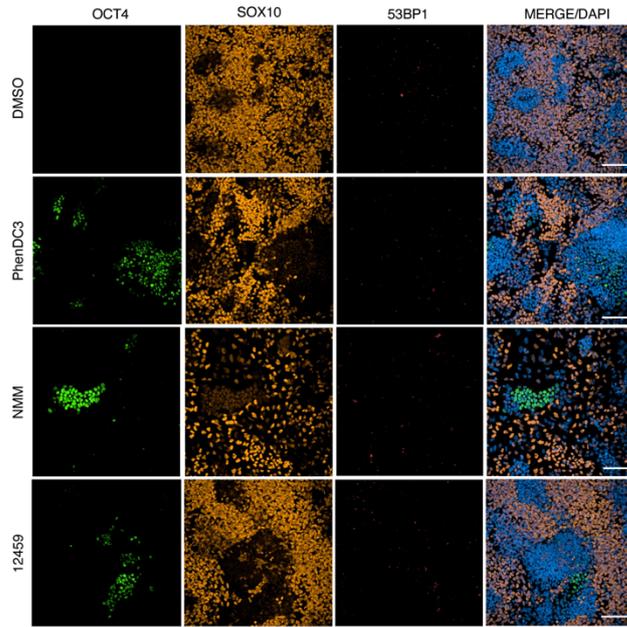
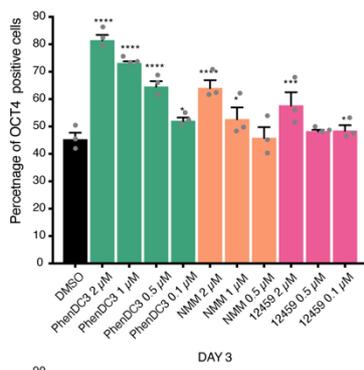
Supplementary Figure 18: PhenDC3 treatment results in differentiation delay. RNA-seq was performed at day 3 and day 5 after induction of the 5-day CNCC differentiation protocol. hESCs grown in mTESR1 were used as the pluripotency control. **a**, Hierarchical clustering of counts per million (CPM) generated from 3 biological replicates (i.e. Rep) at day 3 and day 5 for hESCs treated with NCC induction media only (Control) or with NCC induction media and DMSO or 2 μ M PhenDC3. Height calculated as 1-euclidean distance. **b**, Spearman correlation of expression levels (average counts per million) for stem cell maintenance genes (GO:0019827). **c-d** MA plots showing differential expressed genes (Log2FC >2, FDR <0.05) for the comparison between hESC and DMSO treated (left) and hESC and 2 μ M PhenDC3 treated after NCC induction (right) at (c) day 3 and (d) day 5.

a

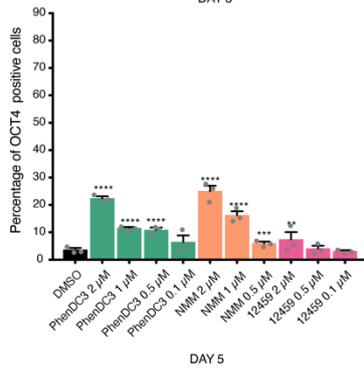
NMM



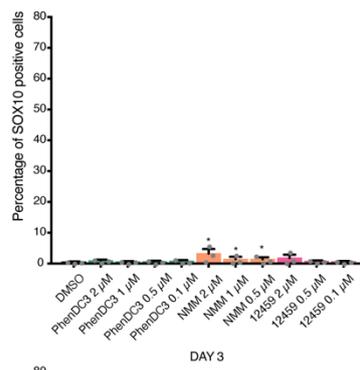
12459

b**c**

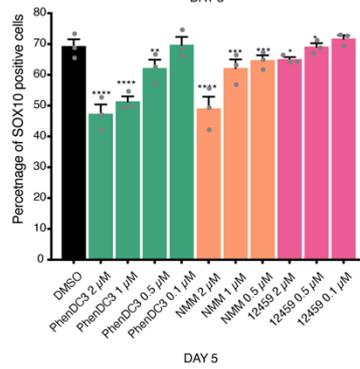
DAY 3



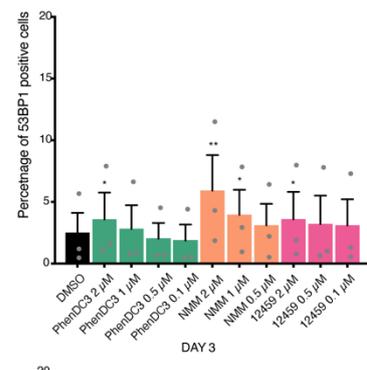
DAY 5

d

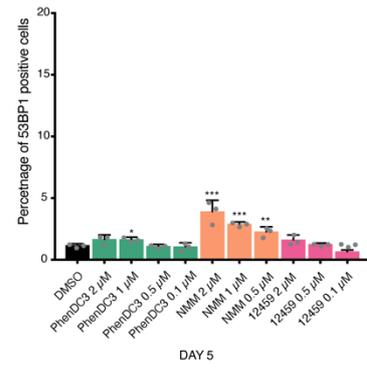
DAY 3



DAY 5

e

DAY 3



DAY 5

Supplementary Figure 19: Treatment with further G4-specific small molecules results in hESC differentiation delay. hESCs were differentiated into CNCCs as described in Figure 6. Samples were taken at the day 3 and day 5 to determine the percentage of cells positive for the pluripotency marker OCT4 or SOX10 (lineage marker for CNCC) by immunofluorescence (IF). **a**, Chemical structures of G4-specific ligands *N*-methyl mesoporphyrin IX (NMM)^{11,12} and 12459¹³. **b**, Representative confocal IF images of differentiating hESCs treated with either 0.02% DMSO, 2 μ M PhenDC3, 2 μ M NMM or 2 μ M 12459. SOX10 = red, OCT4 = green and DAPI (nuclear stain) = blue. Scale bar = 100 μ m. **c-d**, Proportion of (c) OCT4, (d) SOX10 and (e) 53BP1 (DNA damage marker¹⁴) positive cells determined in IF studies. $n = 7,000$ to $75,000$ cells obtained from 52 fields of view per sample. Mean \pm SEM shown for 3 independent biological replicates *: $p < 0.05$, **: $p < 1E-10$, ***: $p < 1E-50$, ****: $p < 1E-100$; one-sided Pearson's χ^2 test for proportions (see Source Data for exact p-values).

Supplementary Discussion

hESC cell cycle changes do not account for high G4 abundance

We considered whether hESCs have more G4s due to a higher proportion of asynchronously dividing cells in S phase¹⁵ as the increased transient occurrence of single-strand DNA during replication could favour G4 formation¹⁶. To rule this out, cell cycle analysis of asynchronous hESCs, CNCCs and NSCs was performed using flow cytometry. As expected, there is an increased percentage of cells in G1 in CNCCs and NSCs compared to hESC, due to lengthening of the G1 phase which accompanies differentiation^{17,18} (Supplementary Fig. 3c-e). Notably, the fraction of CNCCs in S phase ($55.4\% \pm 2.7\%$) was not significantly lower than hESCs ($54.3\% \pm 1.8\%$) ($p = 0.800$) thus the number of cells in S-phase alone does not account for the high number of G4s in hESCs.

Expression level differences across promoter classes

Compared to genes carrying promoters marked by H3K4me3, promoters positive for both active H3K4me3 and repressive H3K27me3 marks (bivalent) are generally either transcriptional silent or lowly expressed¹⁹. To determine whether G4 presence in a promoter is associated with their transcriptional output, the 4 histone promoter signatures (H3K4me3, bivalent, H3K27me3 and unmarked) were stratified by the presence or absence of a G4 (G4+, G4-, respectively). In all 3 stem cell types, genes with H3K4me3 or bivalent promoters that contained a G4 showed significantly ($p \leq 6.0E-09$) increased transcription compared to those without at G4. However bivalent G4-positive promoters had significantly ($p \leq 6.9E-5$) lower expression than H3K4me3 genes (Fig. 3h, Supplementary Fig. 9a-b). This demonstrates that the presence of a G4 in a promoter leads to on average higher expression levels, supporting our previous observations in cancer²⁰. However, for bivalent promoters, promoter histone classification appears to determine transcriptional levels relative to G4 presence.

To assess if the G4 landscape alone directly recapitulates transcriptional levels, we performed a k-medoid clustering of the 11,876 genes (total 58,381 gene set filtered for genes with median TPM >1) considering normalised expression (median TPM) levels and promoter G4 signal (median from replicates) for each gene across the three stem cell types (Supplementary Fig. 9d-f). This clustering analysis revealed that G4 levels do not directly account for expression levels, thus G4s are not sufficient to recapitulate overall expression patterns. Furthermore, this integrative analysis revealed 6 distinct trends and that G4 promoters can be associated with all ranges of expression levels.

Supplementary References

1. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**, 68–83 (2015).
2. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
3. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* (2011) doi:10.1093/bioinformatics/btr189.
4. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* (2011) doi:10.1093/bioinformatics/btr064.
5. Marsico, G. *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.* **47**, 3862–3874 (2019).
6. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–186

- (2013).
7. Lee, S. M. *et al.* Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1885–E1894 (2017).
 8. Raudvere, U. *et al.* G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz369.
 9. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **13**, 551–564 (2018).
 10. Xu, H. *et al.* CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat. Commun.* **8**, 1–18 (2017).
 11. Yingfu Li, C. Ronald Geyer, and & Sen*, D. Recognition of Anionic Porphyrins by DNA Aptamers†. *Biochemistry* **35**, 6911–6922 (1996).
 12. Nicoludis, J. M., Barrett, S. P., Mergny, J.-L. & Yatsunyk, L. A. Interaction of human telomeric DNA with N- methyl mesoporphyrin IX. *Nucleic Acids Res.* **40**, 5432–5447 (2012).
 13. Riou, J. F. *et al.* Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. *Proc. Natl. Acad. Sci.* **99**, 2672–2677 (2002).
 14. Rappold, I., Iwabuchi, K., Date, T. & Chen, J. Tumor Suppressor P53 Binding Protein 1 (53bp1) Is Involved in DNA Damage–Signaling Pathways. *J. Cell Biol.* **153**, 613 (2001).
 15. Becker, K. A. *et al.* Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. *J. Cell. Physiol.* **209**, 883–893 (2006).
 16. Técher, H., Koundrioukoff, S., Nicolas, A. & Debatisse, M. The impact of replication stress on replication dynamics and DNA damage in vertebrate cells. *Nature Reviews Genetics* vol. 18 535–550 (2017).
 17. Calegari, F. & Huttner, W. B. An inhibition of cyclin-dependent kinases that lengthens, but does not arrest, neuroepithelial cell cycle induces premature neurogenesis. *J. Cell Sci.* **116**, 4947–4955 (2003).
 18. Borghese, L. *et al.* Inhibition of Notch Signaling in Human Embryonic Stem Cell-Derived Neural Stem Cells Delays G1/S Phase Transition and Accelerates Neuronal Differentiation In Vitro and In Vivo. *Stem Cells* **28**, 955–964 (2010).
 19. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).
 20. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).