Deimination of arginine at the C-terminal domain favors RNA polymerase II pause release

Priyanka Sharma,[1,2]; Antonios Lioutas,[1,2]; Javier Quilez,[1,2]; Roni H.G. Wright,[1,2]; Chiara Di Vona,[1,2]; Francois Le Dily,[1,2]; Roland Schüller,[3]; Dirk Eick,[3]; Miguel Beato[1,2]

[1] Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

[3] Department of Molecular Epigenetics, Helmholtz Center Munich, Center of Integrated Protein Science, Munich, Germany

Contact

E-mails: miguelbeato@crg.eu or priyanka.sharma@crg.eu

Highlights

- Peptidyl arginine deiminase 2 (PADI2) deiminates arginine (R1810) present at C-terminal domain of RNA polymerase II (RNAP2-CTD).

- PADI2 and R1810 of RNAP2-CTD regulate transcription of breast cancer cells.

- PADI2 depletion and absence of R1810 deimination at RNAP2-CTD inhibits proliferation of breast cancer cells.

- PADI2 mediated deimination of R1810 at RNAP2-CTD favors pause release.

Summary

The C-terminal domain of the large subunit of RNA polymerase II (RNAP2-CTD) coordinates transcription and associated processes by acting as a landing platform for a variety of protein complexes. In mammals, the RNAP2-CTD comprises 52 heptapeptide repeats, the first half of which (1-27) exhibit the consensus repeat sequence Y_1-S_2-P_3-T_4-S_5-P_6-S_7, whereas the second half (28-52) contains deviations from this consensus [1-3]. The residues present on CTD undergo posttranslational modifications to determine which factors will be recruited to process nascent transcripts and modify chromatin. Dynamic phosphorylation, mainly on serines 2 and 5 mediates selective recruitment of protein complexes [4-7], but recently, modifications of lysines and arginines in non-consensus repeats have expanded the functional complexity of the CTD code[8-11]. Here we show that R1810 can be deiminated by PADI2 favoring RNAP2 pause release at highly expressed genes relevant for proliferation of breast cancer cells. Depletion of PADI2 reduced expression of these genes accompanied by accumulation of RNAP2 at transcriptional start sites and resulted in inhibition of cell proliferation. As PADI2 is overexpressed in several cancers and is related to poor prognosis, selective inhibitors may help to prevent cancer progression.

Deimination of R1810 at RNAP2-CTD

Arginine deimination, also known as citrullination, is a post-translational modification of arginine residues that generates the non-coded amino acid citrulline. This reaction is catalyzed by enzymes called peptidyl arginine deiminases (PADIs)[12-14]. Arginine 1810 in repeat 31 can be asymmetrically[10] or symmetrically[11] dimethylated leading to reduced expression of small nuclear and nucleolar RNAs [10], or facilitating transcription termination[11], respectively. We explored the possibility that it could be deiminated and functionally affect the transcription. Immunoprecipitation from nuclear extracts of the breast cancer cell line T47D-MTVL[15] using an α-citrulline detects two bands migrating as the non-phosphorylated (IIA) and week phosphorylated (IIO) forms of the large subunit of RNAP2 (**Fig.1a**). The IIO form reacted with α-citrulline to much higher proportion than IIA form of RNAP2. Next, we raised a specific antibody against a 13 residues peptide centered on R1810, which was replaced by citrulline (**Extended Data**

Fig.1a top). This antibody (α-Cit1810) reacted with the citrullinated peptide, but not with the wild type, methylated (me2aR1810) or phosphorylated (S2 or S5) peptides (**Extended Data Fig.1b**). It recognized the phosphorylated form of RNAP2 in Western blots of nuclear extracts from T47D, MCF7 & HeLa cells (**Extended Data Fig.1a**). To validate that R1810 is deiminated, we transiently transfected T47D-MTVL cells with α-amanitin resistant HA-tagged wild type (WT^r) or $R1810A^r$ mutant of RNAP2, followed by α-amanitin treatment to deplete the endogenous RNAP2 (**Extended Data Fig.1c-d**). Precipitation with anti-HA antibody followed by Western blot showed that the WT^r RNAP2, but not the $R1810A^r$ mutant, reacts with α-Cit1810 (**Fig.1b**). In super-resolution immunofluorescence images of T47D cells, α-Cit1810 decorated bright clusters overlapping with RNAP2, preferentially in its S2 or S5 phosphorylated forms (**Extended Data Fig.1e-f**).

PADI2 deiminates R1810 at RNAP2-CTD

Searching for the responsible enzyme, we found that T47D cells express only *Padi2* and *Padi3* (**Fig.1c). **Depletion experiments with specific siRNAs showed that PADI2 but not PADI3 is responsible for R1810 deimination (**Fig.1d**, **Extended Data Fig.2a**). MCF7 cells express *Padi2* and *Padi4*[16,17] but only depletion of *Padi2* reduced R1810 deimination (**Extended Data Fig.2b-c**). Therefore, we incubated recombinant PADI2 with either a recombinant GST-N-CTD (repeat 1-25.5, including R1603) or with GST-C-CTD (repeat 27-52, including R1810) using GST as control (**Fig.1e, *left panel***). PADI2 deiminated R1810 in the C-CTD much more efficiently than R1603 in N-CTD (**Fig.1e, *right panel***).

Using microscale thermophoresis[18] we observed that PADI2 binds the unmodified R1810 peptide with relatively high affinity (K_d= 220±54.5 nM), whereas phosphorylated peptides were not bound (**Extended Data Fig. 2d**), suggesting that the observed S2/S5 phosphorylation in R1810 deiminated CTD must occur outside of repeats 31 and 32. The peptide with asymmetrically dimethylated R1810 was not bound by PADI2 (**Extended Data Fig.2d**). In co-immunoprecipitation experiments using T47D and MCF7 cells extracts, PADI2 but not PADI3 or PADI4 interacted with RNAP2 (**Fig.1f, Extended Data Fig.2e**). Similarly, an antibody against PADI2 precipitated Cit1810-RNAP2, along with S5P and S2P-RNAP2 (**Fig.1g, Extended Data Fig.2f).** Precipitation with monoclonal antibodies against RNAP2 phosphorylated at S2 and S5 (see methods) pulled down Cit1810 RNAP2 as well as PADI2 but not PADI3 (**Fig.1h**), suggesting an association of PADI2 with the active RNAP2. T47D nuclear extracts fractionated by size exclusion chromatography separated PADI2, along with phosphorylated RNAP2, in the high molecular weight fractions not PADI3 (**Extended Data Fig.2g**). Triple labeling immunofluorescence showed that PADI2 co-localized with Cit1810-RNAP2 and with S2P-RNAP2, consistent with PADI2 being an integral part of the engaged RNAP2 transcription machinery (**Extended Data Fig.2h**).

PADI2 is required for transcription regulation in breast cancer cells

To understand the function of PADI2, we performed replicated mRNA sequencing in wild type and PADI2-depleted T47D cells (**Extended Data Fig.3a**). PADI2 knockdown affected over 4,000 genes (**Fig.2a, Extended Data Fig.3b**), enriched in key physiological processes, including RNAP2-mediated transcription and cell proliferation (**Extended Data Fig.3c**). Reduced expression was validated by RT-qPCR for several genes including *Serpina6, c-Myc* and *Hmgn1* genes, while control genes *Gsst2* and *Lrrc39* were not affected (**Fig.2b, Extended Data Fig.3d**). Depletion of CARM1 and PRMT5, which catalyze the asymmetrical and symmetrical dimethylation of R1810, respectively[10,11], did not affect PADI2-dependent genes (**Extended Data Fig.3e-f**). In T47D cells expressing only the α-amanitin resistant R1810A^r mutant of RNAP2, expression of PADI2-dependent genes decreased significantly (**Fig.2c**). To explore the direct effect of PADI2 on transcription, we performed chromatin-RNA sequencing[19,20] (ChrRNA-seq) and found that ~2,000 genes were significantly affected by PADI2 knockdown, the majority of them (1,884) were down-regulated and 20% of them were significantly reduced at mRNA seq (**Fig.2d-f**). We substantiated expression of PADI2-dependent genes through RT-qPCR on ChrRNA (**Fig.2g**). Thus, PADI2 favors transcription at the level of chromatin associated nascent transcripts.

PADI2 directed deimination of R1810 needed for cell proliferation in breast cancer cells

Given that many PADI2 dependent genes were related to cell proliferation, we monitored T47D cell proliferation after PADI2 depletion (si*Padi2*), inhibition with Cl-amidine (pan-citrulline inhibitor), or cells expressing only R1810A^{r} mutant of RNAP2 and in all cases found a significant reduction (**Fig.2h**). PADI2 depleted cells were arrested at the G1 phase of the cell cycle (**Extended Data Fig.3g**), as expected given the downregulation of genes critical for G1 phase progression (**Extended Data Fig.3h-i, Extended Data Table 1**).

PADI2 recruitment associate with RNAP2 and correlates with genes expression level

Chromatin immunoprecipitation sequencing (ChIP-seq) of PADI2 in T47D cells showed that 60% of chromatin-bound PADI2 is located over protein-coding gene sequences, within 3kb upstream of TSS (transcription start site) and 3kb downstream of TTS (transcription termination site) (q value $\leq$0.005). PADI2 highest enrichment (2.5-fold) was found in the coding exons, followed by the 3kb downstream of the TTS (1.6-fold) (**Extended Data Fig.4a**). Overall PADI2 occupancy overlapped with RNAP2 binding measured by ChIP-seq, suggesting that PADI2 accompanies RNAP2 along the genes up to the termination region (**Fig.3a**). Quartiles of increasing levels of transcription [21] showed increased occupancy of RNAP2 and PADI2 in parallel with the expression levels (**Extended Data Fig.4b**). We verified the correlation of PADI2 and gene expression by ChIP-qPCR on highly expressed (*Serpina6,* c-*Myc*) and low expressed (*Gstt2*) genes and found that after PADI2 depletion the values decreased (**Extended Data Fig.4c**). PADI2 occupancy on PADI2 dependent genes was significantly higher on genes downregulated than non-regulated by PADI2 depletion (**Extended Data Fig.4d**). Additionally, RNAP2-ChIP followed by PADI2 re-ChIP revealed the association of RNAP2 and PADI2 at regulatory regions and gene bodies of the highly transcribed *Serpina6,* c-*Myc* genes, but not in the low expressed *Gstt2* gene, supporting the functional association of PADI2 with the transcription machinery (**Extended Data Fig.4e**).

PADI2 mediated deimination of R1810 at RNAP2-CTD is essential for pause release

Next, in RNAP2 ChIP-seq we found that T47D cells expressing the R1810A^{r} mutant exhibited accumulation of RNAP2 close to the TSS and increased pausing index particularly in highly transcribed genes (**Fig. 3b**), confirming previous findings in Raji cells[11]. Notably, this tendency was more pronounced in PADI2 dependent genes (n=2186) (**Fig. 3c-e**). Sequencing of mRNA of T47D cells expressing R1810A^r mutant RNAP2 showed 1392 down-regulated genes. Of which 939 (67.4%) were also dependent on PADI2. These 939 genes showed significant higher pausing index compared to cells expressing WT^r RNAP2 (**Extended Data Fig.4f**), confirming a role of PADI2 and R1810 in RNAP2 pause release.

Further, we validated by RNAP2 ChIP qPCR that PADI2 depletion or cells expressing R1810A^r mutant of RNAP2 showed significant accumulation of RNAP2 at the TSS of *Serpina6, Hmgn1*, c-*Myc* genes (**Extended Fig.5a-b**). Thus, PADI2 depletion or absence of R1810 lead to RNAP2 accumulation on the promoters of highly expressed genes. Also, we found that after *Padi2* knockdown S2P and S5P forms of RNAP2 are reduced on *Serpina6,* c-*Myc* genes or in cells expressing only R1810A^r mutant form of RNAP2 (**Extended Data Fig.5 c-d**), supporting a role of PADI2 mediated deimination of R1810 in RNAP2 promoter pause release. In summary, we found that PADI2 deiminates R1810 at CTD of RNAP2 and is essential for promoter pause release of highly expressed genes required for cancer cell proliferation. Depletion of PADI2 or expression of RNAP2 carrying the R1810A^r mutation leads to RNAP2 pausing along with decrease of phosphorylated RNAP2.

Discussion

The previously reported asymmetrical[10] and symmetrical[11] dimethylation of R1810 was detected only after phosphatase treatment. In contrast, R1810 deimination is preferentially found associated with the active RNAP2. Depletion of CARM1 and PRMT5 did not affect expression of PADI2 dependent genes and depletion of PADI2 did not change the expression of broad range of snRNAs, targets of arginine methylases. Thus, it seems that R1810 methylation and deiminiation are required for different sets of genes and are largely independent processes. However, we also find genes upregulated upon PADI2 depletion or R1810A RNAP2 mutant expression in T47D cells, which exhibited lower pausing index (**Extended Data Fig.5d-e**). Their upregulation could be an indirect consequence of the irreversible nature of deimination that will interfere with the role of R1810me2s on transcriptional termination[11].

Many elongation factors and kinases are implicated in the control of RNAP2 transcription pause release, a mechanism that controls expression of genes involved in cancer progression and metastasis, like MYC, JMJD6 [22,23]. Indeed, we found that PADI2 depletion or mutation of R1810 reduced cell proliferation of breast cancer cells, by modulating cell cycle progression. Among *Padi* gene family members, *Padi2* is overexpressed in breast cancers[24] and other cancers[25] (**Extended Data Fig.6 a-e**). *Padi2* overexpression correlates with poor prognosis (**Extended Data Fig. 7a-e**), suggesting that specific inhibition of deimination at R1810-RNAP2 may represent a suitable drug target for combinatorial cancer therapy.

References

1. Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36**, 541-546, doi:10.1016/j.molcel.2009.10.019 (2009).
2. Corden, J. L. RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. *Chem Rev* **113**, 8423-8455, doi:10.1021/cr400158h (2013).
3. Jeronimo, C., Collin, P., Robert, F. The RNA Polymerase II CTD: The Increasing Complexity of a Low-Complexity Protein Domain. *J Mol Biol* **428**, 2607-2622, doi:10.1016/j.jmb.2016.02.006 (2016).
4. Jonkers, I. & Lis, J.T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 167-177, doi:10.1038/nrm3953 (2015).
5. Saldi, T., Cortazar, M.A., Sheridan, R.M., Bentley, D.L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol* **428**, 2623-2635, doi:10.1016/j.jmb.2016.04.017 (2016).
6. Zaborowska, J., Egloff, S., Murphy, S. The pol II CTD: new twists in the tail. *Nat Struct Mol Biol* **23**, 771-777, doi:10.1038/nsmb.3285 (2016).
7. Harlen, K.M., Churchman, L.S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol*, doi:10.1038/nrm.2017.10 (2017).
8. Voss, K. *et al.* Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II. *Transcription* **6**, 91-101, doi:10.1080/21541264.2015.1114983 (2015).
9. Dias, J.D. *et al.* Methylation of RNA polymerase II non-consensus Lysine residues marks early transcription in mammalian cells. *Elife* **4**, doi:10.7554/eLife.11215 (2015).

10. Sims, R.J., 3rd *et al.* The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science* **332**, 99-103, doi:10.1126/science.1202663 (2011).
11. Zhao, D.Y. *et al.* SMN and symmetric arginine dimethylation of RNA polymerase IIC-terminal domain control termination. *Nature* **529**, 48-53, doi:10.1038/nature16469 (2016).
12. Witalison, E.E., Thompson,P.R., Hofseth,L.J., Protein Arginine Deiminases and Associated Citrullination: Physiological Functions and Diseases Associated with Dysregulation. *Curr Drug Targets.* **16**, 700-710 (2015).
13. Van Venrooij W.J., Pruijn G.J., Citrullination: a small change for a protein with great consequences for rheumatoid arthritis. *Arthritis Res.* **2**, 249-51, (2000).
14. Gyorgy, B., Toth, E., Tarcsa, E., Falus, A., Buzas, E.I. Citrullination: a posttranslational modification in health and disease. *Int J Biochem Cell Biol* **38**, 1662-1677, doi:10.1016/j.biocel.2006.03.008 (2006).
15. Truss, M., Bartsch, J., Schelbert, A., Haché R.J.,Beato, M. Hormone induces binding of receptors and transcription factors to a rearranged nucleosome on the MMTV promoter in vivo. *The EMBO Journal* **14**, 1737-1751 (1995).
16. Cuthbert, G. L. *et al.* Histone deimination antagonizes arginine methylation. *Cell* **118**, 545-553, doi:10.1016/j.cell.2004.08.020 (2004).
17. Sharma,P. *et al.* Citrullination of histone H3 interferes with HP1-mediated transcriptional repression. *PLoS Genet* **8**, e1002934, doi:10.1371/journal.pgen.1002934 (2012).
18. Jerabek-Willemsen, M., Wienken, C.J., Braun, D., Baaske, P., Duhr, S. Molecular interaction studies using microscale thermophoresis. *Assay Drug Dev Technol* **9**, 342-353, doi:10.1089/adt.2011.0380 (2011).
19. Werner, M.S. & Ruthenburg, A.J. Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. *Cell Rep* **12**, 1089-1098, doi:10.1016/j.celrep.2015.07.033 (2015).
20. Nojima, T., Gomes, T., Carmo-Fonseca, M., Proudfoot, N.J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc* **11**, 413-428, doi:10.1038/nprot.2016.012 (2016).
21. Baranello, L. *et al.* RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription. *Cell* **165**, 357-371, doi:10.1016/j.cell.2016.02.036 (2016).
22. Bywater, M.J., Pearson, R.B., McArthur, G.A. & Hannan, R.D. Dysregulation of the basal RNA polymerase transcription apparatus in cancer. *Nat Rev Cancer* **13**, 299-314, doi:10.1038/nrc3496 (2013).
23. Miller, T.E. *et al.* Transcription elongation factors represent in vivo cancer dependencies in glioblastoma. *Nature*, doi:10.1038/nature23000 (2017).
24. Cherrington, B.D. *et al.* Potential role for PAD2 in gene regulation in breast cancer cells. *PLoS One* **7**, e41242, doi:10.1371/journal.pone.0041242 (2012).

25. Guo, W. *et al.* Investigating the expression, effect and tumorigenic pathway of PADI2 in tumors. *Onco Targets Ther* **10**, 1475-1485, doi:10.2147/OTT.S92389 (2017).

26. Vicent, G.P. *et al.* Two chromatin remodeling activities cooperate during activation of hormone responsive promoters. *PLoS Genet* **5**, e1000567, doi:10.1371/journal.pgen.1000567 (2009).

27. Meininghaus, M., Chapman, R.D., Horndasch, M. & Eick, D. Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription. *J Biol Chem* **275**, 24375-24382, doi:10.1074/jbc.M001883200 (2000).

28. Fong, N., Bentley, D.L. Capping, splicing, and 3 processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Gene and developmenet* **12**, 1783–1179, doi:10.1101/ (2001).

29. Wright, R.H. *et al.* CDK2-dependent activation of PARP-1 is required for hormonal gene regulation in breast cancer cells. *Genes Dev* **26**, 1972-1983, doi:10.1101/gad.193193.112 (2012).

30. Christophorou, M.A. *et al.* Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature* **507**, 104-108, doi:10.1038/nature12942 (2014).

31. Wang, Y., *et al.* Human PAD4 Regulates Histone Arginine Methylation Levels via Demethylimination. *Science* **306,** 279-283 (2004).

32. Pau, G., Fuchs, F., Sklyar, O., Boutros, M., Huber, W. EBImage--an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979-981, doi:10.1093/bioinformatics/btq046 (2010).

33. Marco-Sola, S., Sammeth, M., Guigo, R., Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**, 1185-1188, doi:10.1038/nmeth.2221 (2012).

34. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185, doi:10.1093/bioinformatics/bts356 (2012).

35. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:10.1038/nature08903 (2010).

36. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

37. Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

38. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).

39. Richardson, A.L. *et al.* X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9**, 121-132, doi:10.1016/j.ccr.2006.01.013 (2006).

40. Cancer Genome Atlas Network. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

41. Adib, T.R. *et al.* Predicting biomarkers for ovarian cancer using gene-expression microarrays. *Br J Cancer* **90**, 686-692, doi:10.1038/sj.bjc.6601603 (2004).

42. Brune, V. *et al.* Origin and pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global gene expression analysis. *J Exp Med* **205**, 2251-2268, doi:10.1084/jem.20080809 (2008).

43. Cho, J.Y. *et al.* Gene expression signature-based prognostic risk score in gastric cancer. *Clin Cancer Res* **17**, 1850-1857, doi:10.1158/1078-0432.CCR-10-2180 (2011).

44. Compagno, M. *et al.* Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**, 717-721, doi:10.1038/nature07968 (2009).

45. Hou, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One* **5**, e10312, doi:10.1371/journal.pone.0010312 (2010).

46. Murat, A. *et al.* Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J Clin Oncol* **26**, 3015-3024, doi:10.1200/JCO.2007.15.7164 (2008).

47. Gyorffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* **123**, 725-731, doi:10.1007/s10549-009-0674-9 (2010).

48. Gyorffy, B., Lanczky, A., Szallasi, Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr Relat Cancer* **19**, 197-208, doi:10.1530/ERC-11-0329 (2012).

49. Gyorffy, B., Surowiak, P., Budczies, J. & Lanczky, A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One* **8**, e82241, doi:10.1371/journal.pone.0082241 (2013).

50. Szász A.M., *et al.* Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *oncotarget* **7**, 49322-49333 (2016).

51. Vathipadiekal, V. *et al.* Creation of a Human Secretome: A Novel Composite Library of Human Secreted Proteins: Validation Using Ovarian Cancer Gene Expression Data and a Virtual Secretome Array. *Clin Cancer Res* **21**, 4960-4969, doi:10.1158/1078-0432.CCR-14-3173 (2015).

52. Vicent, G.P., *et al.* Progesterone receptor interaction with chromatin. *Methods Mol Biol.* **1204**, 1-14 (2014).

53. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, r25, doi:10.1186/gb-2009-10-3-r25) (2009).

54. Xu, S., Grullon, S., Ge, K.,Peng, W., Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol* **1150**, 97-111, doi:10.1007/978-1-4939-0512-6_5 (2014).

55. Zhu, L.J., *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* (2010).

56. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

57. Iannone,C. *et al.* Relationship between nucleosome positioning and progesterone-induced alternative splicing in breast cancer cells. *RNA* **21**, 360-374, doi:10.1261/rna.048843.114 (2015).

58. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618-1619, doi:10.1093/bioinformatics/btu056 (2014).

59. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**, 1512-1516, doi:10.1038/ng.2007.26 (2007).

60. Chen, F.X et al., PAFI, a molecular regulator of promoter proximal pausing by RNA Polymerase II. *Cell* **162**,1003-15, doi: 10.1016/j.cell.2015.(2015).

Acknowledgments

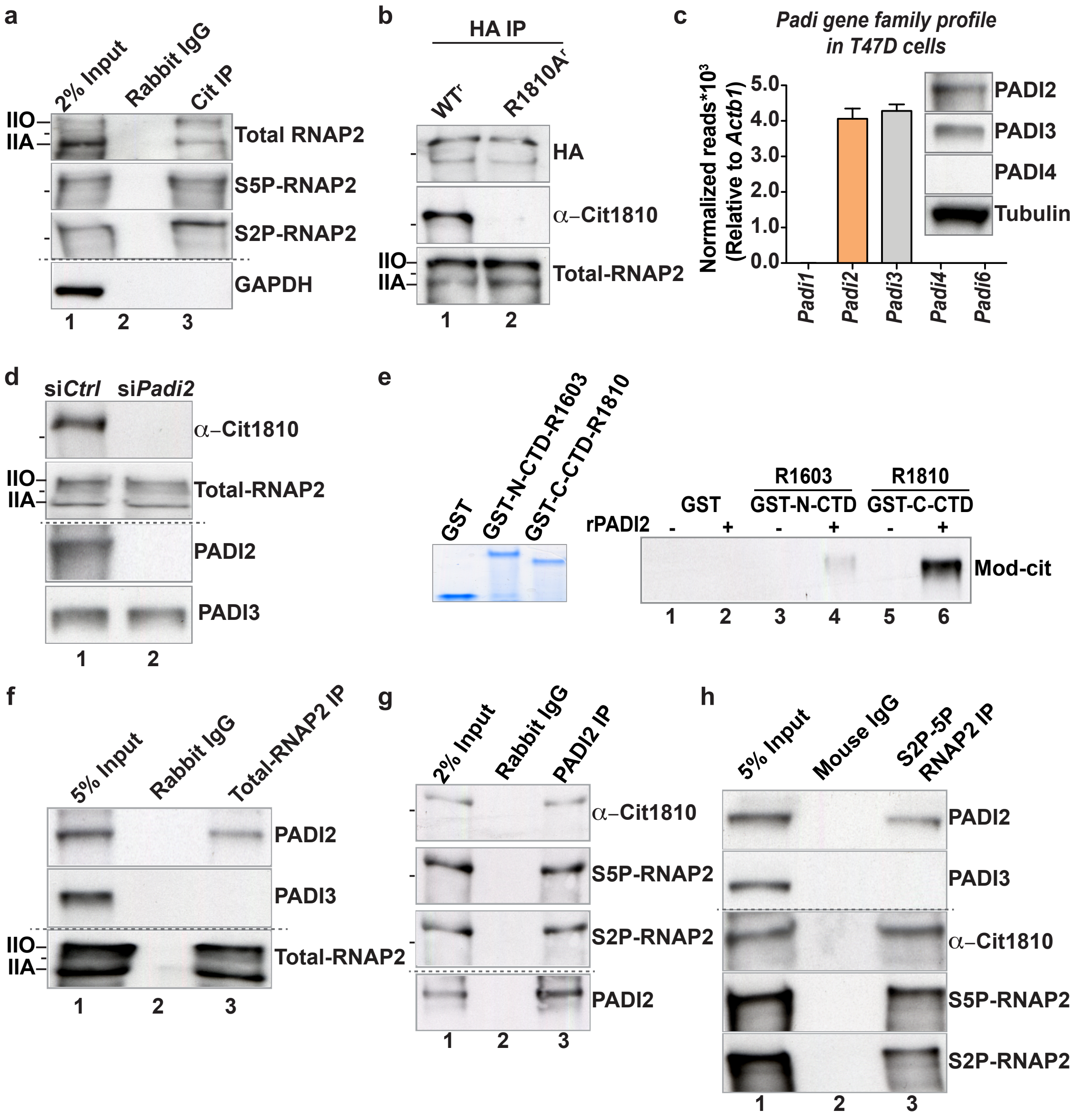
We thank David Bentley, University of Colorado, for GST-CTD plasmids; Hiroshi Kimura, Hokkaido University, for RNAP2 S2P (CMA602) /5P (CMA603) monoclonal antibodies. We thank CRG Genomics, Protein Technology and the Advanced Light Microscopy facilities for technical support. We are grateful to Enrique Vidal Ocabo, CRG, for advice on STED microscopy data quantitation. We thank Juán Valcárcel, Guillermo P. Vicent, and Gwendal Dujardin for constructive criticism and advices during the course of this work. P.S. was supported by a Novartis fellowship and Beatriu de Pinós fellowship co-founded by Marie Curie Action (2013 BP_B 00061). R.S and D.E supported by Deutsche Forschungsgemeinschaft, SFB1064, Chromatin Dynamics. This work was supported by Spanish MEC (SAF2013-42497), the Catalan Government (AGAUR 2014SGR780) and the ERC Synergy Grant “4DGenome” (609989). We acknowledge support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa” and the CERCA Programme / Generalitat de Catalunya.”

Author contributions

P.S and M.B designed the study and strategy of the project; P.S performed the large majority of the experimental work; A.L performed immunofluorescence and image analysis. J.Q provided the bioinformatics analysis of all high throughput data. R.H.G.W helped with oncomine analysis. R.S and D.E helped with the R1810A^{r} mutant experiments. C.D.V and F.L.D performed RNAP2 ChIP sequencing, P.S and M.B discussed the results and wrote the paper. All other authors contributed to editing the manuscript.

Author Information

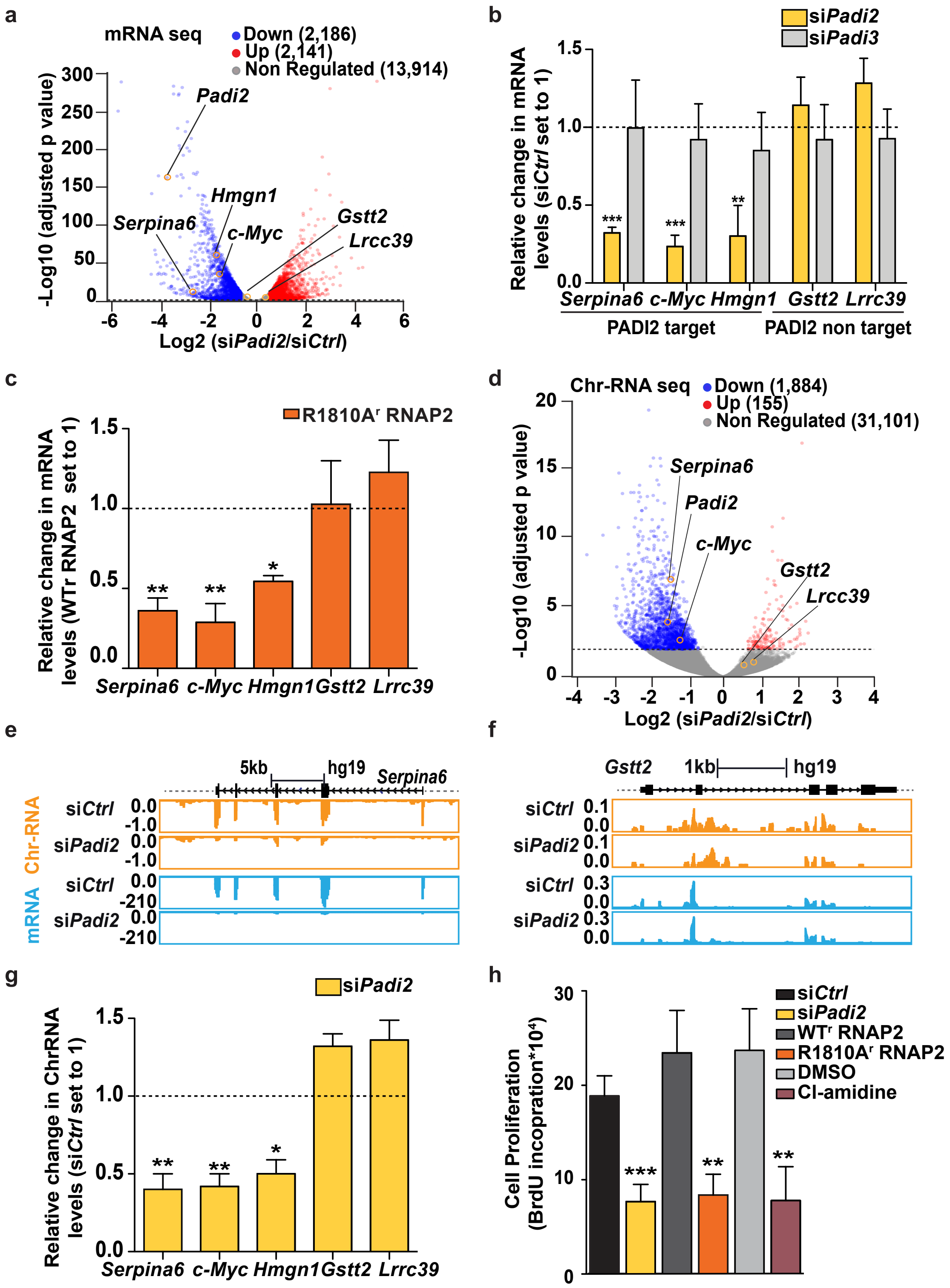
Global data sets of ChIP-Seq and RNA Seq analysis had been deposited at GEO with accession code (GEO Submission (GSE105795, GSE105792, GSE105793)). The authors declare no competing interest. Correspondence and request for material should be correspondence to miguel.beato@crg.eu and priyanka.sharma@crg.eu.



Sharma *et al.*, Figure 1

Figure 1: PADI2 deiminates R1810 at CTD of RNAP2. (a) T47D cells extracts immunoprecipitated with pan-citrulline antibody and analyzed by Western blot using antibodies against total-RNAP2, S5 and S2 phosphorylated forms of RNAP2. The dotted line indicates a separate gel (see methods) and lines on the left mark the migration of the 250 kDa size marker. (**b**) Extracts from T47D cells expressing α-amanitin resistant WT^r or R1810A^r mutant of RNAP2 were precipitated with α-HA antibody and probed with α-

Cit1810 or α-RNAP2. (**c**) Bar plot showing the normalized reads of *Padi* gene family members from two RNA-sequencing experiments performed in T47D cells. The normalized reads are represented relative to *Actb1* gene. Values are means ± SEM. Inset-Western blots performed on T47D nuclear extract with antibodies to PADI2, 3 and 4. (**d**) Nuclear extracts from T47D cells transfected with siRNA control (si*Ctrl*) or siRNA against PADI2 (si*Padi2*) are probed with α-Cit1810, α-total-RNAP2, α-PADI2 and α-PADI3. (**e**) *Left*: Coomassie blue staining of SDS-PAGE with recombinant GST-tagged, GST-N-CTD and GST-C-CTD proteins used for the deimination assay. *Right*: *In vitro* deimination immunoblot with or without recombinant PADI2 (rPADI2) using as substrate the N-terminal half of the CTD containing R1603 (*lanes 3 & 4*) or the C-terminal half containing R1810 (*lanes 5 & 6*), both linked to GST. As a control, GST was also tested (*lanes 1 & 2*). (**f-h**) Immunoprecipitation with (**f**) α-total RNAP2 (**g**) α-PADI2 (**h**) α-S2P/S5P RNAP2 or non-immune mouse or rabbit IgG of T47D extracts followed by Western blot with the indicated antibodies.



Sharma *et al.* Figure 2

Figure 2. PADI2 mediated deimination of R1810 regulates transcription and cell proliferation. (a) Volcano plot showing genome-wide mRNA changes after PADI2 depletion from biological replicates. The X-axis represents log2 expression fold changes (FC) and the Y-axis represents the adjusted p-values (as -log10). Differentially expressed genes (FC > 1.5 or <1/1.5 and p value < 0.01) are shown, the positions of *Padi2* and genes used for validation are also indicated. (**b**) Quantitative RT-qPCR validation in T47D cells transfected with si*Ctr,* si*Padi2* and si*Padi3*. Changes in mRNA levels were normalized to *Gapdh* mRNA. Data represent mean ± SEM of at least three biological experiments as in other plots in the figure. **(c)** T47D cells expressing only α-Amanitin resistant WT^{r} or R1810A^{r} mutant form of RNAP2 were used for quantitative RT-qPCR of mRNA from PADI2 dependent genes (*Serpina6, c-Myc* and *Hmgn1*), and for control genes (*Gstt2* and *Lrrrc39*). **(d)** Volcano plot showing genome-wide chromatin associated RNAs changes before and after PADI2 depletion from independent replicates. The X-axis represents the log2 fold changes (FC) and the Y-axis represents the adjusted p values (-log10). The dotted line indicates the cutoff p value < 0.01. Differential expressed genes (FC > 1.5 or 1/1.5 and the p value < 0.01) are shown. (**e**, **f**) Browser snapshots showing chromatin associated RNAs sequencing profile (orange) and mRNA sequencing profile (blue) of (e) *Serpina6* and (f) control gene *Gstt2* after *PADI2* knockdown performed in T47D cells. Scale is indicated on the top of each gene. (**g**) Quantitative RT-qPCR on chromatin associated RNA (ChrRNA) in T47D cells transfected with si*Ctr,* si*Padi* RNA. Data normalized to *Gapdh* ChrRNA expression level. (**h**) Cell proliferation of T47D cells in the absence (si*Ctrl* or DMSO) or presence of PADI2 depletion (si*Padi2*), PADI2 inhibitor (Cl-amidine in DMSO) and of cells expressing α-Amanitin resistant WT^{r} and R1810A^{r} mutant form of RNAP2.

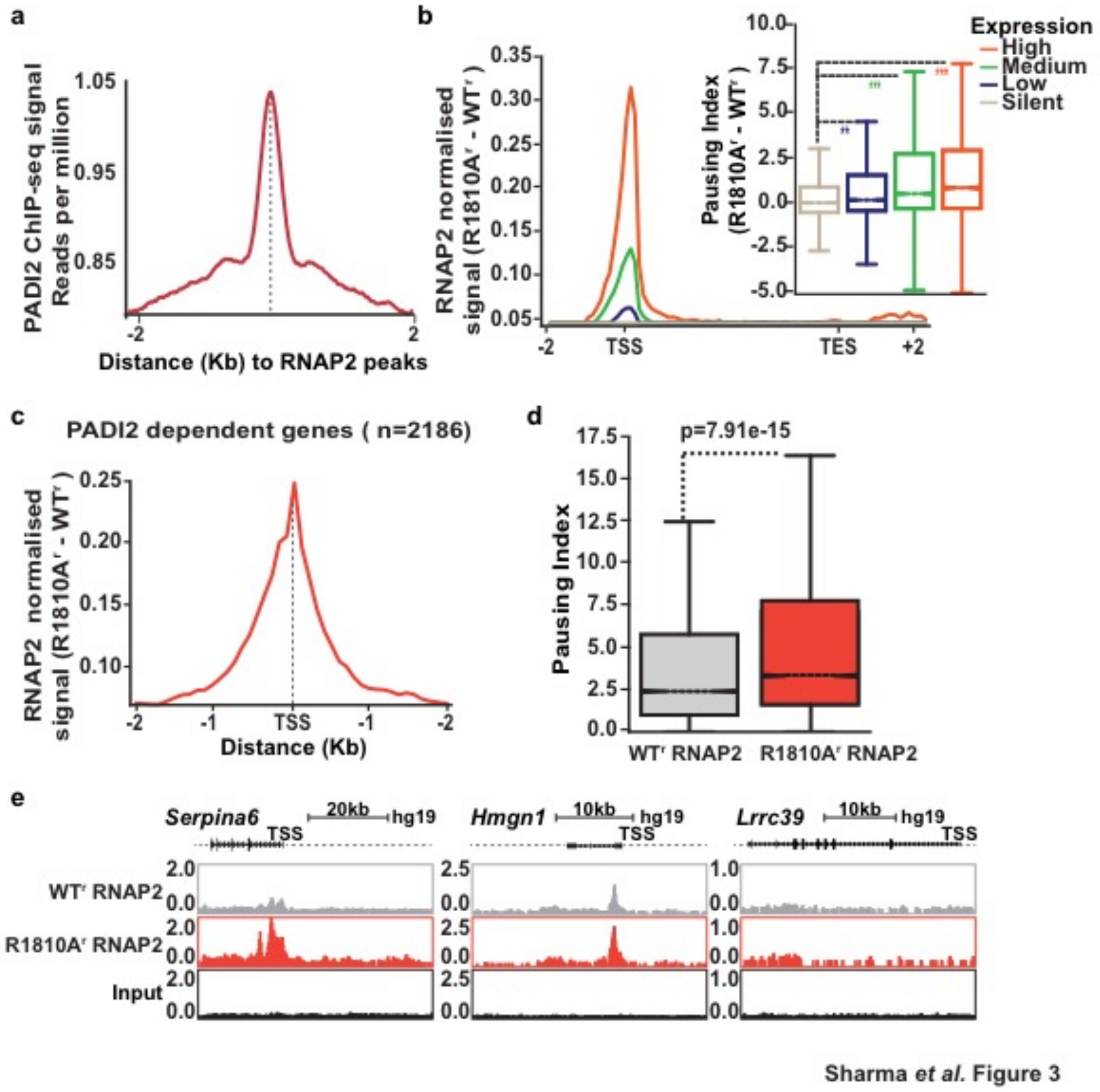


Figure 3. **PADI2 mediated deimination of R1810 CTD-RNAP2 regulates pausing.** **(a)** Distribution of normalized PADI2 reads around the center of RNAP2 peaks in T47D cells. p value was calculated by Wilcoxon-Mann-Whitney test in comparison to silent genes as indicated (** p value < 10^{-3}; *** p value < 10^{-5}). **(b)** Difference in RNAP2 occupancy in cells expressing HA-tagged R1810A^{r} mutant versus WT^{r} RNAP2 across genes classified by expression. ***Inset:*** Pausing index of RNAP2 as indicated. **(c-d)** PADI2 dependent genes showing (**c**) Difference in RNAP2 normalized ChIP-seq signal (R1810A^{r} -WT^{r}) (**d**) Higher pausing index in cell expressing R1810A^{r} mutant as compared to WT^{r} form of RNAP2. (**e**) Browser snapshots showing RNAP2 occupancy for *Sepina6*, *Hmgn1* & control gene *Lrrc39* in cells expressing HA-tagged WT^{r} or R1810A^{r} form of RNAP2.

METHODS

Cell culture: The breast cancer cell line T47D was used either as wild type or carrying a single copy of luciferase reporter gene driven by MMTV promoter (T47D-MTVL[15]). Cells were grown in RPMI-1640 medium without phenol red supplemented with 10% dextran coated charcoal treated FBS (DCC/FBS), 2mM L-glutamine, 100U/ml penicillin-streptomycin as reported previously[26,29]. MCF7 cells were grown in DMEM (without phenol red) with 10% DCC/FBS and 100U/ml penicillin-streptomycin. HeLa and HEK293 cells were cultured in DMEM with 10% FBS (fetal bovine serum) with 100U/ml penicillin-streptomycin. All cell transfections were carried out using Lipofectamine 3000 (Invitrogen) according to manufacturer's instructions. Cells were treated with 200µM citrulline inhibitor Cl-amidine (506282, Calbiochem) or vehicle (DMSO) for 2 hours. Cells expressing recombinant RNAP2 were treated with α-amanitin (A2263, Sigma) for 12 hours.

Antibodies

Anti-CTD cit1810 Antibody

The citrulline specific antibody was raised in rabbits by Eurogentec using a KLH coupled CTD peptide sequence (`YSPSSP-cit-YTPQSP`). Affinity purification was performed first on a column containing the citrullinated peptide, followed by removal of non-citrullinated specific antibodies on a column containing the non-citrullinated peptide.

Commercial antibodies used in this study were as follows: anti-PADI2 for ChIP: (sc-133877 lot no # E1214 and H0715, Santa Cruz); anti-PADI2 for Western blots (12110-1-AP from Proteintech and WH0011240M1 from Sigma); anti-citrulline (AB5612, Millipore), anti-citrulline detection kit (17-347, Millipore,), anti-CARM1 (09-818, Millipore), anti-PRMT5 (07-405, Millipore), anti-TUBULIN (T9026, Sigma), anti-GAPDH (sc-32233, Santa Cruz), anti-PADI4 (ab50247, Abcam), anti-PADI3 (sc-393622, Santa Cruz), anti-HA (ab9110, Abcam), anti-Total RNAP2 for ChIP (CTD4H8, Millipore), for Western blot (N-20, sc-899, Santa Cruz), anti-phospho-S2 RNAP2 CTD

(3E10, 04-1571 from Millipore). IgG negative control for ChIP and immunoprecipitation assays (12-371, Millipore; 2729S Cell Signalling)

Anti-S2P-RNAP2 (CMA602, MBL Life science) and anti-S5P-RNAP2 (CMA603, MBL Life science) were kindly provided by Hiroshi Kimura´s laboratory.

Peptides

All peptides were synthesized and purified by Eurogentec. The CTD peptide PSY**S**$_2$PS**S**$_5$P**R**YTPQ**S**$_5$PTYTP was used for dot blots, and microscale thermophoresis (MST) experiments, either unmodified (WT) or with Cit1810, R1810me^{2a}, S$_2$-P (1805) or S$_5$-P (1808, 1815) modifications. Peptides were quantified by amino acid analysis, and the presence of the modifications was confirmed by mass spectrometry.

Plasmids

The α-amanitin resistant HA-tagged wild type (WT^r) or R1810A^r mutant RNAP2 plasmids were previously published[10,27]. The GST-N-CTD (repeats 1-25.5), the GST-C-CTD (repeats 27-52) of RNAP2 were kindly provided by David Bentley[31].

RNA interference experiments

For siRNAs inhibition experiments T47D-MTVL or MCF7 cells were transfected with 100µM siRNA using Lipofectamine 3000 (Invitrogen) for 72 hours according to manufacturer´s instructions. SMARTpool On-target plus siRNAs for PADI2 (M-019485-01) and PADI3 (M-021051-01) were purchased from Dharmacon (Thermo Scientific). siRNAs for CARM1 (sc-44875), PRMT5 (sc-41073) and PADI4 (sc-61283) were purchased from Santa Cruz.

RNA extraction and RT-qPCR

Total RNA from T47D-MTVL cells was extracted with RNeasy (Qiagen) and quantified with a Qubit 3.0 Fluorometer (Life technologies). After DNase treatment (Ambion), reverse transcription was performed using SuperScript III (Invitrogen) according to manufacturer´s instructions. Complementary DNA was quantified by qPCR using Roche Lightcycler (Roche), as previously described[26]. For each gene product, relative RNA abundance was calculated using the standard curve method and expressed as relative RNA abundance after normalizing against the human *Gapdh* gene level. All the gene expression data generated by RT-qPCR represent the average and SEM of at least 3 biological replicates. Primers used for RT-qPCR were as listed here:

Padi2 f CATGTCCCAGATGATCCTGC,
Padi2 r CATGGTAGAGCTTCCGCC,
Padi3 f TCTACATCTCTCCCAACATGG,
Padi*3 r* GTCAACACAGGTGAGGTAGA,
Hmox1 f GCAGAGAATGCTGAGTTCAT,
Hmox1 r ACATAGATGTGGTACAGGGA,
Gapdh f GAGTCAACGGATTTGGTCGT,
Gapdh r TTGATTTTGGAGGGATCTCG,
Serpina6 f TGGCTATGAATTTCCAGGAC,
Serpina6 r CAGTTGTCTCGTCCACATAG,
c-*Myc f* AGAGTCTGGATCACCTTCTG,
c-*Myc r* GGTTGTTGCTGATCTGTCTC,
Hmgn1 f CAGCAGCGAAGGATAAATCT,
Hmgn1 r GACTTGGCTTCTTTCTCTCC,
Gstt2 f ATCTTCGCCAAGAAGAATGG,
Gstt2 r GGTACTTACAGCTCAGGTAAATC,
Lrrc39 f CCCTTCTTCTCTGCTGAAAC,
Lrrc39 r CTGAGAATCAGTTCCTGAAGTC,
Carm1 f GCAAGCAGTCCTTCATCAT,
Carm1 r GGATGTTGTAGAAGGAACAGAA,
Prmt5 f TTCATTCAGGAACCTGCTAAG,
Prmt5 r GGACGAATCCATGGAGAAAG,
Ccnd1 f CCCTCGGTGTCCTACTTCAA,
Ccnd1 r AGGAAGCGGTCCAGGTAGTT,
Thbs1 f CAGAATGTGAGGTTTGTCTTTG
Thbs1 r TCTTGTGGCCAATGTAGTTAG
Pgr f AGGTCTACCCGCCTATCTC
Pgr r GATGCTTCATCCCCACAGAT
Gnmt f CCATTATGCTGGTGGAAGAG
Gnmt r CATCTTTGTCCAGAGTCATCC
Nfkbia f TGAGGATGAGGAGAGCTATG
Nfkbia r CCTCCAAACACACAGTCATC
Cald1 f AGTCCTTCCTTTCCGACTTA
Cald1 r CCCTGTGGAATTTGATTTGATG

Irx2 f CTACGAGCCCAAGAAAGATG
Irx2 r CACTTACTTGCATTGCTGTG
Ehbp1 f CGAAGAAATGGCAACCAGATA
Ehbp1 r CACAACAACACCACGATAGG.

Chromatin RNA was prepared as described previously[20]. Briefly, T47D-MTVL cells transfected with si*Ctrl* or si*Padi2* were lysed for 5 minutes in 4ml of ice-cold HLB+N [10mM Tris-HCl pH 7.5, 10mM NaCl, 2.5mM $MgCl_2$ and 0.5% (vol/vol) NP-40], followed by addition of 1ml of ice cold HB+NS buffer [10 mM Tris-HCl pH 7.5, 10mM NaCl, 2.5mM MgCl2, 0.5% vol/vol, NP-40 and 10% wt/vol sucrose]. Cell nuclei were collected by centrifugation at 1,400rpm for 5 minutes at 4°C, resuspended in 125µl of NUN1 buffer [20mM Tris-HCl pH 7.9, 75mM NaCl, 0.5mM EDTA and 50% vol/vol glycerol]. After addition of 1.2 ml of ice cold NUN2 buffer [20mM HEPES-KOH pH 7.6, 300mM NaCl, 0.2mM EDTA, 7.5mM MgCl2, 1% vol/vol NP-40 and 1M urea], samples were incubated on ice for 15 minute, mixing by vortexing every 3 minutes. The chromatin pellet was resuspended in HSB buffer [10mM Tris pH 7.5, 500mM NaCl and 10mM MgCl2] and treated with TURBO™ DNase (AM2239, Ambion) at 37˚C for 15 minutes, followed by Proteinase K (AM2546, Ambion) for 10 minutes at 37˚C. Chromatin-RNA was purified by Trizol (ThermoFisher Scientific) and quantified with a Qubit 3.0 Fluorometer (Life Technologies). Reverse transcription was performed using random hexamers and SuperScript III (Invitrogen) according to manufacturer´s instructions. Complementary DNA was quantified by qPCR. Before preparing chromatin RNA libraries, 1µg of chromatin RNA was used to deplete rRNAs followed by libraries preparation using TruSeq Stranded total RNA Library Prep Kit (Illumina) and sequenced using Illumina HiSeq 2500.

Protein extract preparation, Co-immunoprecipitation (IP), and Western blots

Cells were prepared as described previously[29]. Briefly, 5 x 10^6 to 10^7 cells were lysed on ice for 30 minutes in lysis buffer (1% Triton X-100 in 50mM Tris pH 7.4-7.6, 130mM NaCl) containing proteases inhibitors (11836170001, Roche) with rotation, followed by mild sonication pulses each 10 seconds. After centrifugation at 4ºC and 13,000rpm for 10 minutes, extracts were used for protein quantitation. For IP 2mg of extract proteins were incubated for 12 hours with protein G/A agarose beads (for rabbit antibodies, IP05, Millipore) or Dynabeads (for mouse antibodies, 11201D, ThermoFisher Scientific), previously coupled with 5-7µg of the corresponding antibodies or a control IgGs. For RNAP2-S2P and -S5P 7µg of each mouse monoclonal antibodies (CMA602 or CMA603, respectively) were coupled with Dynabeads, followed by 12 hours incubation with extract at 4ºC. The samples were washed 6 times with lysis buffer and boiled for 5 minutes in SDS gel sample buffer. For detection of mentioned proteins of molecular weight (<200 kDa) 4-12% SDS-PAGE gels were used; while for the RNAP2 large subunits (> 200kDa) we used 3-8% SDS-PAGE. For Western blots primary antibodies were used at 1:250 to 1:1000 dilution and incubated overnight at 4ºC, followed by an hour incubation with horseradish peroxidase conjugated anti-mouse (NA931V) or anti-rabbit (NA934V, Amersham) and blots were developed using ECL prime Western blotting detection reagent (RPN2232, GE Healthcare) according to the manufacture instructions.

Size exclusion chromatography

The size exclusion chromatography of T47D-MTVL cell extracts were carried out using Superdex 200 10/300mm columns (17517501, GE healthcare). As per manufacturer´s instructions, for high molecular weight (Ferritin ,440 KDa) and for low molecular weight (Conalbumin,75 KDa & Carbonic anhydrase 29KDa) were run along with cell extracts. Samples were chosen according to chromatography profile and used for Western blots.

BrdU (5´-bromo-2´-deoxyuridine) cell proliferation assay

T47D-MTVL cells (1×10^4) were plated in a 96-well plate followed by transfection with control/ PADI2 siRNAs or treated with the PADI inhibitor Cl-amidine at concentration of 200µM or DMSO or expressing α-amanitin resistant HA-tagged wild type (WT^r) or R1810A^r mutant form of RNAP2.The cell proliferation ELISA BrdU Colorimetric assay (Roche,11647229001) was performed as per manufacturer's instructions. The experiments were performed at least four biological replicates.

Fluorescence-activated cell sorting (FACS) experiments

FACS assay was performed in T47D-MTVL cells transfected with control or PADI2 siRNAs from three biological replicates. Briefly, cells were trypsinized, washed three times with 1x PBS and fixed with cold absolute ethanol in suspension at 70% final concentration. Cells were stained with propidium iodide (P-1304, Molecular Probe) followed by DNase free RNase A (AM2222, Ambion) treatment and stored for 24 hours at 4ºC and DNA contents of cell cycle phases were analyzed using a BD™ LSR II flow cytometer.

In vitro deimination assay with recombinant PADI2 and fragments of the RNAP2-CTD

The PADI2 open reading frame (ORF) was cloned into the HIS-tagged expression vector pCOOFY40 and the plasmid sequence were verified. Recombinant proteins were expressed in HEK-293 cells and purified following the standard method of histidine-tagged recombinant protein. Briefly, cells were lysed in Buffer A (50mM Tris HCl pH7.4, 500mM NaCl, 10% glycerol, 2mM DTT, 20mM Imidazole, 1% and triton X-100) complemented with proteinase inhibitors (11836170001, Roche). Purification was performed using the HiTrap TALON crude (28953766, GE Healthcare) according to manufacturer´s instruction. Proteins eluted in buffer containing 50mM Tris-HCl pH 7.4, 300mM NaCl and 10% glycerol, were stored at -80ºC until required. The GST-N-CTD (repeats 1-25.5), GST-C-CTD (repeats 27-52) of RNAP2 were expressed and purified following the standard glutathione bead purification[28]. *In vitro* citrullination was carried with recombinant His-PADI2 in deimination buffer (50mM HEPES pH 7.5, 10mM $CaCl_2$, 4mM DTT) at 37 ºC for 1hour. Samples were dissolved in sample Laemmli buffer for immunoblot analysis using an anti-citrulline antibody[30,31] (Millipore,17-347).

Microscale Thermophoresis (MST[18]) of recombinant PADI2 with RNAP2-CTD peptides

Wild type peptide or peptides carrying different modifications at R1810, S1805, S1808 and S1815 (20nM to 500µM) were titrated against a fixed concentration of fluorescent recombinant His-PADI2 (50nM). MST data were acquired at 20°C using the red LED at 20% and IR- Laser at 40% using a (Monolith NT.115, Nano Temper Technologies) according to manufacturer's instructions. The results are plotted as normalized fluorescence (Fnorm, representing binding affinity) against the concentration of the unlabeled ligand and fitted according to the law of mass action.

Immunofluorescence, image acquisition and analysis

T47D-MTVL cells were grown on round 10mm glass coverslips transfected with sicontrol or si*Padi2* prior to fixation with 4% paraformaldehyde in PBS for 5 minutes and permeabilized with PBS 0.1% Triton X-100 (PBST) at room temperature for 5 minutes. Coverslips were blocked with IF buffer (5% BSA, 0.1%Triton X-100 in PBS) for 20 minutes at room temperature and incubated overnight with primary antibodies diluted in IF buffer at 1:50 of α-Cit1810 (Rabbit), 1:50 of α-CTD4H8 (mouse, Santa Cruz) to detect the total RNAP2, 1:250 of α-S2P-RNAP2 (mouse, CMA602) and 1:250 of α-S5P-RNAP2 (mouse, CMA603). For triple staining 1:500 of α-PADI2 (Mouse, WH0011240M1, Sigma) was used with 1:50 of α-Cit1810 along with α-S2P-RNAP2 (3E10; Rat, Millipore; Extended Data Fig.2h). After 3x washes with PBST (1X PBS with Triton X-100 0.1 %) samples were incubated with secondary antibodies at a dilution 1:500 (AlexaFluor 594 anti-rabbit, AlexaFluor 488 anti-mouse or AlexaFluor 680 anti-mouse and AlexaFluor 488 anti-rat, Invitrogen-Molecular Probes) for 1h at room temperature followed by three washes with PBST. Samples were mounted with Mowiol mounting medium.

Images were acquired with a Leica SP8 STED 3X microscope using a HC PL APO CS2 100x/1.4 Oil immersion lens, a pulsed supercontinuum light source (white light laser) and HyD detectors, using the Leica acquisition software. Laser and spectral detection bands were chosen for the optimal imaging of nuclear optical sections with a z-distance of 0.16µm. Deconvolution was performed using the Huygens deconvolution software (Scientific Volume Imaging, SVI) for STED modes using shift correction to account for drift during stack acquisition. Adjustments of individual channels were applied to the whole image, pseudo-coloring, cropping and different channel composite images were done with FIJI (https://www.fiji.sc). Raw cropped images were used to calculate the correlation of the fluorescent signal with R (https://www.R-project.org/). All images were imported to R with the package EBImage[32].

mRNA and Chromatin RNA (ChrRNA) sequencing

RNA-sequencing experiments were performed in T47D-MTVL cells transfected with control siRNAs (si*Ctrl*) or si*Padi2* and cells expressing HA-tagged wild type (WT^r) or $\mathrm{R1810A}^r$ mutant form of RNAP2 in two different biological replicates, PolyA plus RNA mRNA and ChrRNA were sequenced using Illumina HiSeq 2500. Paired end reads were mapped using GEM RNA Mapping Pipeline[33] (v1.7) using GENCODE annotation version 19. BAM alignment files were used to generate genome-wide normalized profiles using RseQC tools[34]. Exon quantifications (summarized per-genes) were used for expression level determination, either as raw read counts or as reads per kilobase per million mapped reads (RPKM) using Flux-Capacitor[35]. Spearman pairwise correlation

(R2) between the two biological replicates (E1 and E2) were calculated, for the mRNA-sequencing si*Ctrl* (R^2=0.96), si*Padi2* (R^2=0.97), WT^r (R^2=0.98) and $R1810A^r$ (R^2=0.97) and for the ChrRNA-sequencing experiments si*Ctrl* (R^2=0.95) and si*Padi2* (R^2=0.93). Differential expression analysis was performed using DESeq2 Bioconductor package[36]. Analysis was performed by using 196,520 number of annotated transcripts of hg19 (correspond to 57,280 number of genes). Out of this, we quantitated data for total 18,241 (mRNA seq) and 33,140 (ChrRNA seq) genes. Genes with FC < 1/1.50 and adjusted p-value < 0.01 were considered as down-regulated and genes with FC > 1.50 and adjusted p-value < 0.01 as up-regulated (Fig. 2a and 2d).

Gene Ontology (GO) analysis

Go Annotation were performed using the online tool GSEA[37] (**G**ene **S**et **E**nrichment **A**nalysis, http://software.broadinstitute.org/gsea/index.jsp) collection database v5 . The significant cut off p value and FDR q-value <0.05.

All 315 genes documented with the parent cell cycle GO:0007049 were considered (http://software.broadinstitute.org/gsea/msigdb/cards/CELL_CYCLE_GO_0007049) for analysis. Among them, 282 genes expressed in T47D cells; 101 genes classified as PADI2 dependent (si*Ctrl*/si*Padi2* FC < 1/1.5, p value < 0.01) and rest 181 genes as PADI2 independent (Extended Data Fig. 3g, Extended Data Table1).

Oncomine Gene Expression and mutation analysis in Cancer cohorts

PADI gene family expression was analyzed by using the using Oncomine 4.5 database (Compendia Bioscience) in breast [38-40] and other cancer cohorts[41-46] (Extended Data Table 2). The p-value for a gene is its p-value for the median-ranked analysis (extended data figure 6a & 6e). *Padi2* and *Padi3* genes expression level were analyzed in normal breast versus ductal or invasive ductal breast carcinoma cohorts (extended data figure 6b, 6d) and also in normal breast vs invasive lobular breast carcinoma cohorts (Extended Data Fig. 6c). Data is represented as log2 median expression in cancer versus normal.

Kaplan Meier Survival Analysis

Kaplain Meier analysis was performed using the publically available tool Kaplain Meier plotter[47-51] (http://kmplot.com/analysis/index.php?p=service). *Padi2* over-expression was considered significant with a corrected p value of less than 0.05, the number of patients considered in each cohort tested is given on the graphs (Extended Data Table 2).

Chromatin immunoprecipitation (ChIP) and sequencing analysis

ChIP assays were performed as described previously[52].Briefly, 10 x 10^6 of T47D-MTVL cells, transfected with either si*Ctrl* or si*PADI2* and expressing α-amanitin resistant HA-tagged wild type (WT^r) or $R1810A^r$ mutant RNAP2 were cross-linked for 10 minutes with 1% formaldehyde. Lysate were sonicated to a DNA fragment size range of 200-300bp using a Biorupter sonicator (Diagenode). PADI2 was immunoprecipitated with 6µg of PADI2 antibody (z-22):sc-133877 lot number E1214 and H0715 by 50 µg of chromatin and 42µl of Protein A-Agarose Beads (Diagenode). For RNAP2, 150 µg chromatin was incubated with 20µg of antibody (Rpb1 NTD (D8L4Y), 14958, Cell Signaling; CTD4H8, Millipore), anti-HA (ab9110) for HA-tagged wild type (WT^r) or $R1810A^r$ mutant form of RNAP2, S2P; CMA602 and S5P; CMA603) or control IgG in IP Buffer with 2X SDS buffer (100mM NaCl, 50mM TrisHCl, pH8, 5mM EDTA and 0.5% SDS) and 1X Triton buffer (100mM Tris-HCl, pH8.8, 100mM NaCl, 5mM EDTA and 5% Triton-X) with protease inhibitors (11836170001, Roche) for 16 hours at 4° C. Followed by incubating with Protein A Sepharose beads CL-4B (17-0780-01, GE Healthcare) or 50µl of Dynabeads® M-280 sheep anti-mouse IgG (11201D, ThermoFisher Scientific) for 3 hours. Beads were washed with 3 times with low salt buffer (140mM NaCl, 50mM HEPES, pH 7.4, 1% Triton-X 100), 2 times with high salt buffer (500 mM NaCl, 50mM HEPES, pH 7.4, 1% Triton-X 100) followed by single wash of LiCl Buffer (10mM Tris HCl pH 8.0, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholic acid and 1mM EDTA) and 1X TE buffer in cold room. Subsequently, crosslinks were reversed at 65° C overnight and bound DNA was purified by Phenol-Chloroform. The resultant eluted DNA was quantified by Qubit 3.0 Fluorometer (Life technologies) and followed by real time qPCR and data represented as fold change over input fraction from at least 3 biological replicate experiments. Primers used for ChIP-qPCR were listed here

Serpina6-A *f* CTATACTGGACAATGCCACTC,
Serpina6-A *r* CGGTGATGGTTACTCATGTT,
Serpina6-B *f* CTCACTCCATCTTGTCCTTTG,
Serpina6-B *r* GAGGAGATCCCAGATAGGTATG,
Serpina6-C *f* GGGTAAGCAGACAGAGTTAGA,
Serpina6-C *r* GGAGCGCATGAGGAAATAAG,
Serpina6-D *f* GAGAAGAACTTGGAGGAGATTG,
Serpina6-D *r* GGGTTATGAACCCAGTGTAAG,
Serpina6-E *f* TCCCAAAGTTGATGGTGTTAG,
Serpina6-E *r* GGCAGAAAGAGCAGAGAAAG,
c-Myc-A *f* CTGCTTAGACGCTGGATTT,
c-Myc-A *r* TTTAGGCATTCGACTCATCTC,
c-Myc-B *f* AGGAACTATGACCTCGACTAC,
c-Myc-B *r* CAGCTCGAATTTCTTCCAGATA,
c-Myc-C *f* GAGAGGTAGGCAAAGGAGATA,
c-Myc-C *r* GTAACAGGAGTTTCTTCCTCAC,
Gstt2-A *f* GCCTAGAGCTGTTTCTTGAC,
Gstt2-A *r* CCACCTTTGACCAAATCCA,
Gstt2-B *f* GAACCTTAGCTTGCCTTCT,
Gstt2-B *r* CTTCTGGGCTTTGTGAATTTAT,
Gstt2-C *f* CTCCCAGACTGAAGGGTATTA,
Gstt2-C *r* TGGGTAGATTGGGTAGTCTG,
Gstt2-D *f* CAACGGTCTTTGAAGTCTAGG,
Gstt2-D *r* GAGGTCTGTGTGACAATGTATAA,
Lrrc39 Promoter *f* GGTCCTTTCTGAAATGGTATC,
Lrrc39 Promoter *r* TCAGGTCTTCATTGAGTTTCTT,
Hmgn1 Promoter *f* CTTATCCTTCTCCCTTCTTCAC,
Hmgn1 Promoter *r* GTAACTTCCTTCCTTCCTTCC.

PADI2 and RNAP2 ChIP purified DNA was used to prepare ChIP-seq libraries (biological replicates) and sequenced using Illumina HiSeq 2000 to obtain ~80–100 million reads per replicate. Single-ended sequences were trimmed to 50 bp and mapped to the human genome assembly hg19 using Bowtie[53], keeping only tags that mapped uniquely and with no more than two mismatches. In both biological replicate significant enrichments compared to Input DNA (hereafter referred to as peaks) were identified using SICER[54] with the following parameters: window size 200; fragment size 233 bp ; gap size 600; and FDR 0.01. Genome-wide 10kb Spearman pairwise correlation of PADI2 ChIP-sequencing signal was R^2=0.98 .We then applied ChIPpeakAnno R library[55] to create a merged list of peaks present in both replicates. Fragment sizes were estimated using HOMER tools[56], and used to extend reads towards 3' direction in order to compute genome-wide reads per million (RPM) normalized profiles. RNAP2 ChIP-seq sequences were analyzed as mentioned previously [57]. ChIP-seq RPM normalized profiles were used to generate average profiles over different genomic features using bwtool[58] .

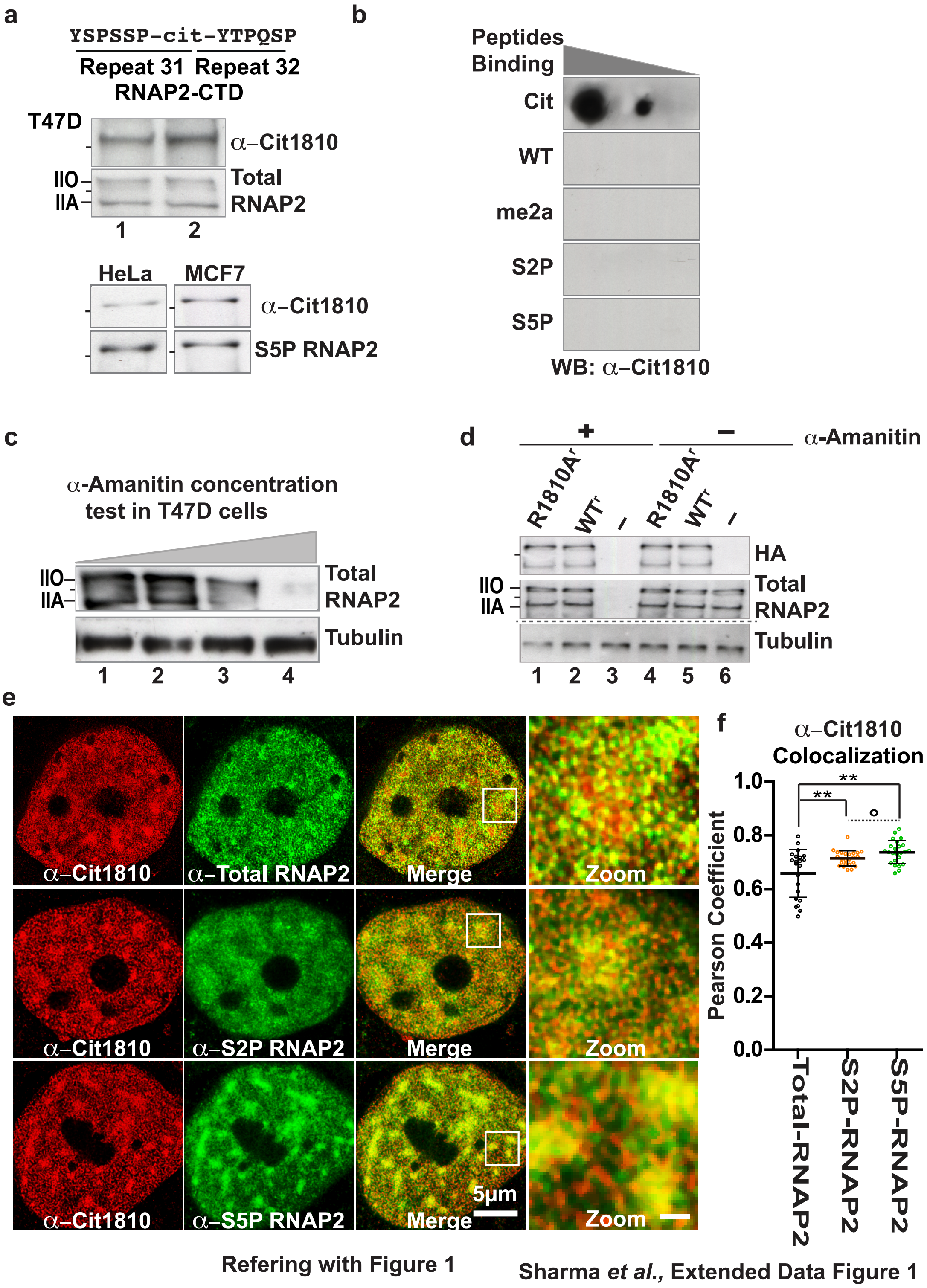
Pausing Index (PI) of RNAP2 Analysis

RNAP2 pausing index was calculated as mentioned previously[21,59-60]. Briefly, RNAP2 pausing index represents the dynamics of RNAP2 assembly and promoter release and hence not only presence or absence of transcription. We calculated pausing index (PI) as the ratio of the RNAP2 read count 1kb (kilo base) flanking to the TSS divided by size and read count in the same gene body divided by size of the gene body.

Statistical analysis

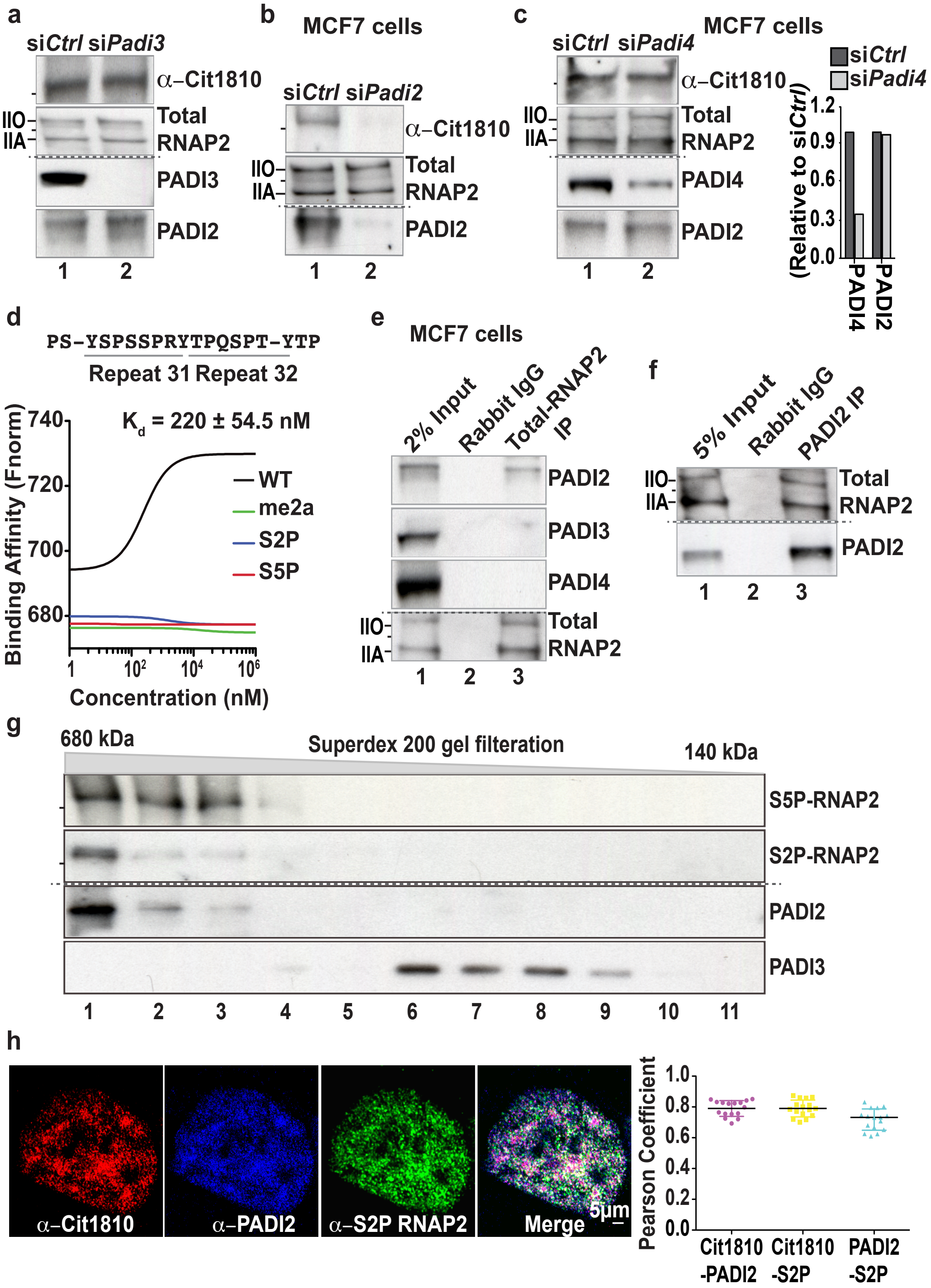
For super resolution image analysis in Extended Data Fig. 1e and 2h, Pearson correlation was calculated for the fluorescent signal of the two channels of interest for each individual stack. Statistical analysis (two tailed Student's t-test) for the average z-stack Pearson´s correlation of each individual cell. Plots and indicated statistical analysis were done with the use of Prism (GraphPad Prism 6.0 for MacOS), unless otherwise stated. For all experiments of RT-qPCR, ChIP-qPCR and cell proliferation, a Two-tailed unpaired Student's t-test was used to determine statistical significance between the groups. Correction between biological replicated of RNA and ChIP sequencing was calculated by Spearman´s correlation (R2). For all other experiments, significance between groups calculated by Wilcoxon-Mann-Whitney test. If exact p-value are not shown or indicated in legend then p-values are represented in all figures as follows: *, p-value ≤ 0.05; **, p-value ≤ 0.01; ***, p-value ≤ 0.001; ^{o}, p value > 0.05.

Data availability. The authors declare that all the data supporting this study are available within paper or as supplementary files. The investigators were not blinded during experiment and outcome assessments.



Extended Data Figure 1. Specificity of the anti-Cit1810 RNAP2 antibody. (a) *Top*: the epitope within repeats 31/32 of the CTD domain of RNAP2 used to generate α-Cit1810. *Center*: duplicated Western blot of T47D nuclear extract with α-Cit1810 and α-total-RNAP2. *Bottom*: HeLa and MCF7 nuclear extracts probed with α-Cit1810 and α-

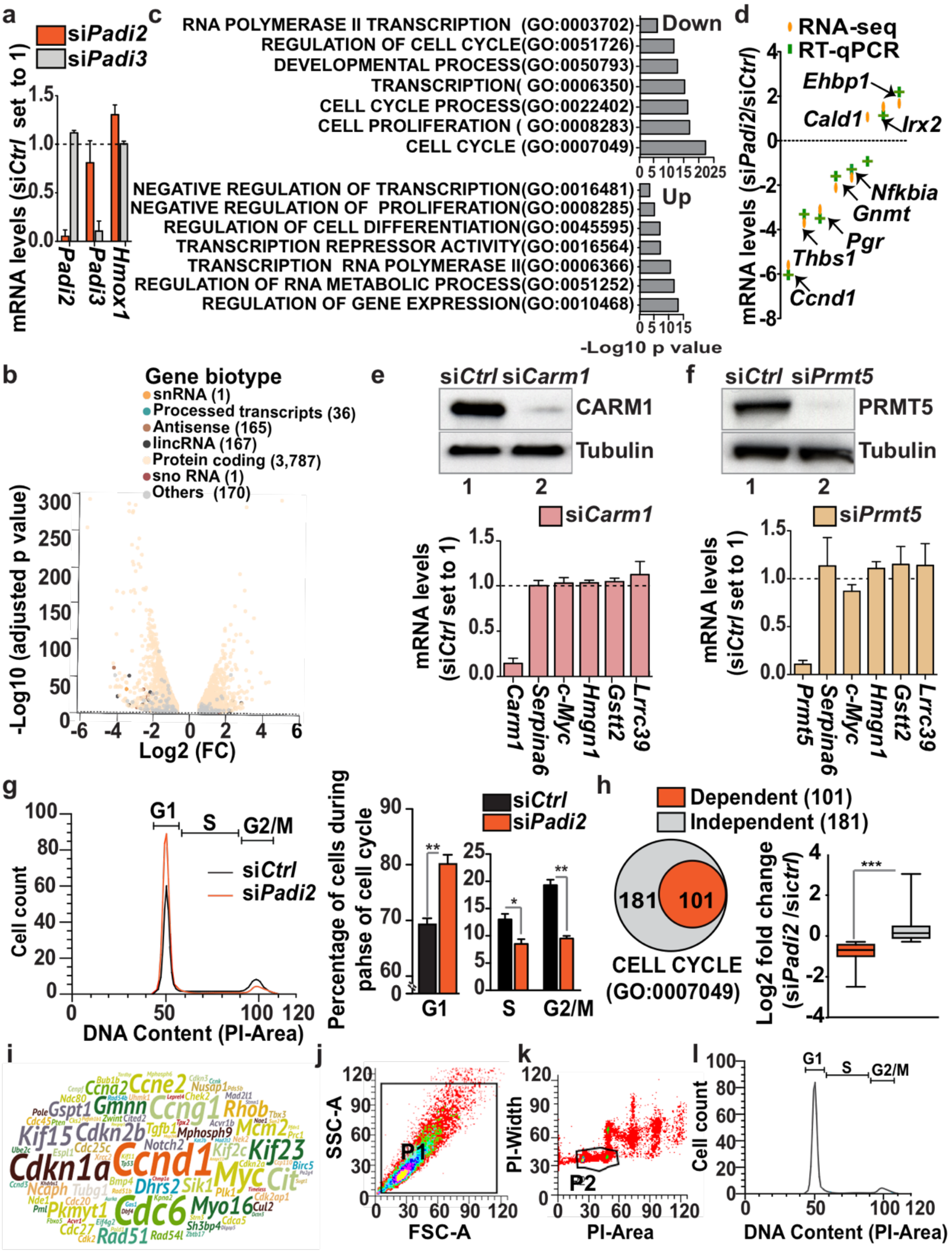
S5P-RNAP2. (**b**) Dot blot showing the specificity of α−Cit1810, performed with 1, 0.2 and 0.04 μg of mentioned RNAP2-CTD peptides encompassing R1810 with specific post-translational modifications. Only, peptide bearing the Cit1810 modification were specifically recognized by α-Cit1810. **(c)** Titration of the concentration of α−amanitin needed to deplete endogenous RNAP2 in T47D cells. Cells were incubated with increasing concentrations of α−amanitin (0, 2, 4 and 6μg/ml; *lanes 1 to 4*) for 12h before preparing nuclear extracts. The extracts were probed in Western blot with an antibody against total RNAP2; Tubulin was used as loading control. (**d**) Western blot of extracts from cells depleted of endogenous RNAP2 by preincubation with 6μg/ml α−amanitin, and transfected with empty vector (-) or with expression vectors for WT^r or the R1810A^r mutant HA-tagged recombinant RNAP2 carrying an additional mutation that makes the enzyme resistant to α−amanitin (see Method). The Western blots were probed with antibodies against the HA tag or total-RNAP2; Tubulin was used as loading control. (**e**) Representative super resolution images of T47D cells immunostained with α-Cit1810 (red) in combination to α-total-RNAP2 (green), α-S2P-RNAP2 (green) and α-S5P-RNAP2 (green). (**f**) Plot representing the mean Pearson correlation coefficient of individual cells for α-Cit1810-RNAP2 with α-total-RNAP2 (n=22), α-S2P-RNAP2 (n=24) and α-S5P-RNAP2 (n=24); values presented as the mean ± SEM. ** p value < 0.005; ^{o} p value > 0.05.



Refering with Figure 1

Sharma *et al.*, Extended Data Figure 2

Extended Data Figure 2. PADI2 interacts with CTD-R1810 in *vivo* and *in vitro* (**a**) Western blot of extracts from T47D cells transfected with siRNA control (si*Ctrl)* or siRNA against *Padi3* (si*Padi3)* and probed with the indicated antibodies. (**b**) Western blot of extracts from MCF7 cells transfected with si*Ctrl* or si*Padi2* and probed with the indicated antibodies. (**c**) Western blot of extracts from MCF7 cells transfected with siRNA *Ctrl* or si*Padi4* and probed with the indicated antibodies. The bar plot represents the total intensity of PADI4 and PADI2 band quantitated by Image J after PADI4 depletion. (**d**) Microscale thermophoresis assay showing the affinity of recombinant PADI2 for the indicated wild type and modified CTD-RNAP2 peptides encompassing the R1810. Y-axis represents the binding affinity as normalized fluorescence (Fnorm, see methods) (**e**) Immunoprecipitation of MCF7 extracts with an α-total-RNAP2 or with non-immune rabbit IgG followed by Western blot with the indicated antibodies. (**f**) Immunoprecipitation of T47D extracts with α-PADI2 or non-immune rabbit IgG probed with indicated antibodies. **(g)** Fractionation of T47D cells extract using a Superdex 200 gel filtration column and analysis of the eluted fractions by Western blotted with the indicated antibodies. (**h**) *Left*: Representative super-resolution immunofluorescence images of T47D cells using triple labelling with α-Cit1810 (red) in combination with α-S2P-RNAP2 (green) and α-PADI2 (blue). Plot representing the mean Pearson correlation coefficient of several individual cells (n=16), for PADI2 with Cit1810 (purple), for S2P-RNAP2 with Cit1810 (yellow) and for PADI2 with S2P-RNAP2 (cyan). *Right*: Person correlation coefficient presented as value of mean ± SEM. Note Cit1810-RNAP2 showed co-localization with PADI2 as well as with S2P-RNAP2 (shown as white in the merged image).

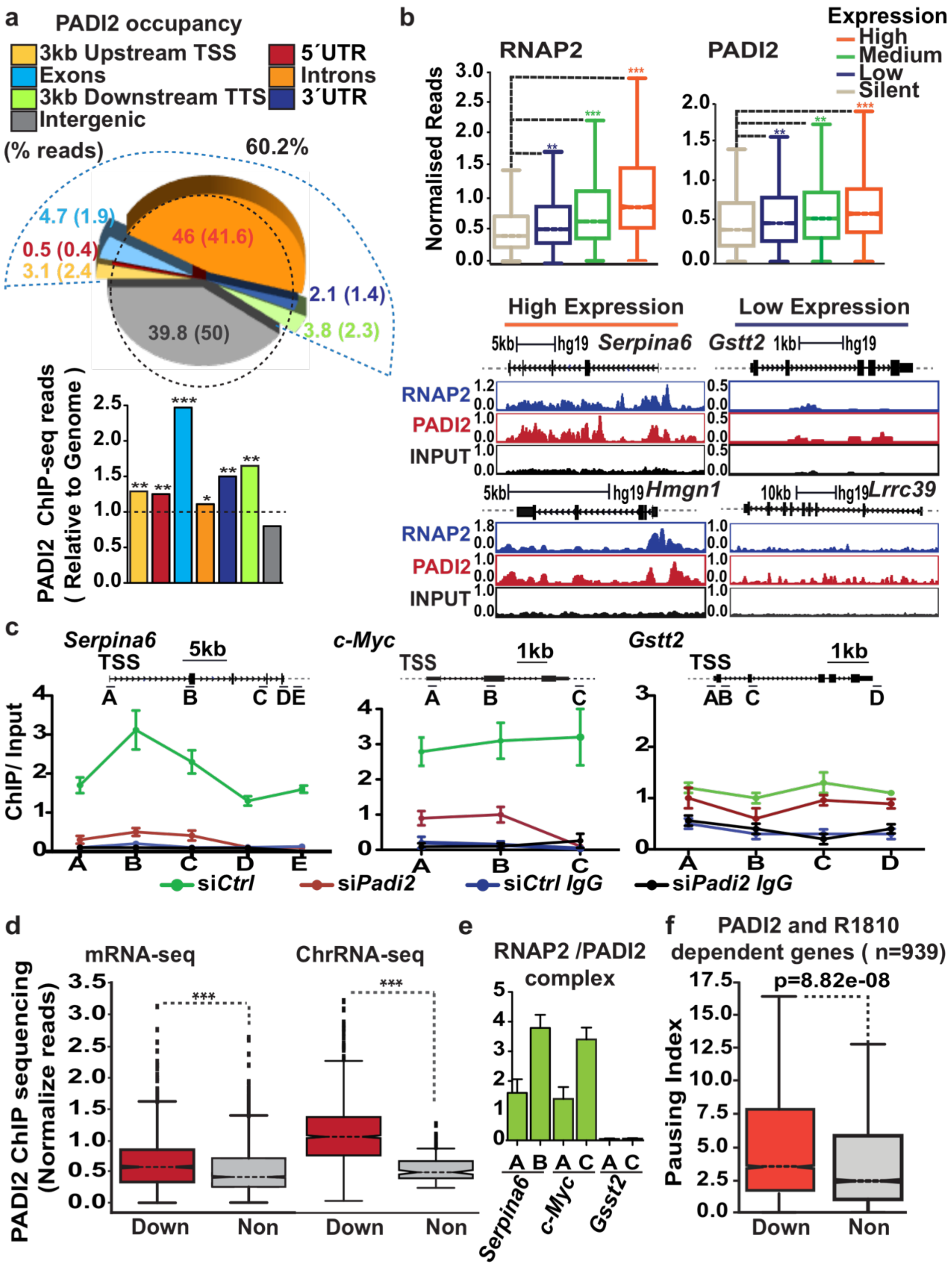


Refering with Figure 2

Sharma *et al.*, Extended Data Figure 3

Extended Data Figure 3. PADI2 depletion affects genes involved in cell cycle progression. (a) Changes in *Padi2*, *Padi3* and *Hmox1* mRNA quantified by RT-qPCR in T47D cells transfected with siRNA *Padi2* (orange bars) or siRNA *Padi3* (grey bars). mRNA levels were normalized to a *Gapdh* mRNA expression level, which was not affected in these experimental conditions, and are presented as ratio to levels in cells transfected siControl RNA (si*Ctr*). Values are the mean ± SEM of at least three biological replicates as in other plots in the figure. **(b)** Volcano plot showing indicated gene biotypes of differentially expressed genes determined from polyA-RNA sequencing as shown in Fig.2a. The X-axis represents log2 expression fold changes (FC) and the Y-axis represents the adjusted p-values (as -log10). **(c)** Gene set enrichment analysis (GSEA) for biological processes and molecular functions. Seven representative processes are presented. The X-axis shows the -log10 transformed p values. GO, Gene Ontology. **(d)** The mRNA-seq data validated by RT-qPCR for indicated genes. Y-axis shows the fold change over si*ctrl*, orange color indicates data from duplicate RNA-seq and green color from RT-qPCR (mean ± SEM from at least three experiments). **(e)** Knock down of *Carm1* mRNA (>80% depletion relative to si*Ctrl RNA*) did not significantly affect PADI-dependent gene transcripts. Data are shown as fold change over si*Ctrl* and represent mean ± SEM. Extracts from si*Ctrl* and si*Carm1* transfected cells were probed with indicated antibodies to estimate CARM1 depletion (>80%). (**f**) Knock down of *Prmt5* mRNA (>80% relative to *siCtrl* mRNA) did not significantly affect PADI-dependent gene transcripts. Extracts from si*Ctrl* and *siPrmt5* transfected cells were probed with indicated antibodies to estimate PRMT5 depletion (>90%). (**g**) *Left*: Cell cycle profile of T47D cells transfected with si*Ctrl (*black) and si*Padi2* (orange) obtained by propidium iodide labelling followed by fluorescence-activated cell sorting (FACS) analysis. *Right*: Histogram showing the percentage of cells during cell cycle phases. Data presented as mean ± SEM from three biological replicates. p value was calculated by student's t test, * p value < 0.05, ** p value < 0.001. (**h**) *Left*: Venn diagram showing the set of genes related to cell cycle (GO:0007049) downregulated in T47D cells after PADI2 depletion (PADI2 dependent) versus PADI2 independent gene (see Extended Data Table 1). *Right*: Box plot showing log2 fold change (si*Ctrl*/si*Padi2*) for PADI2 dependent and independent cell cycle genes. Each box in the panel represents the interquartile range; Whisker extend the box to the highest and lowest values. Horizontal lines across each box indicate the medium value. Dependent genes showed significant lower mRNA levels than independent genes (***p value < 0.0001, calculated by Wilcoxon-Mann-Whitney test).

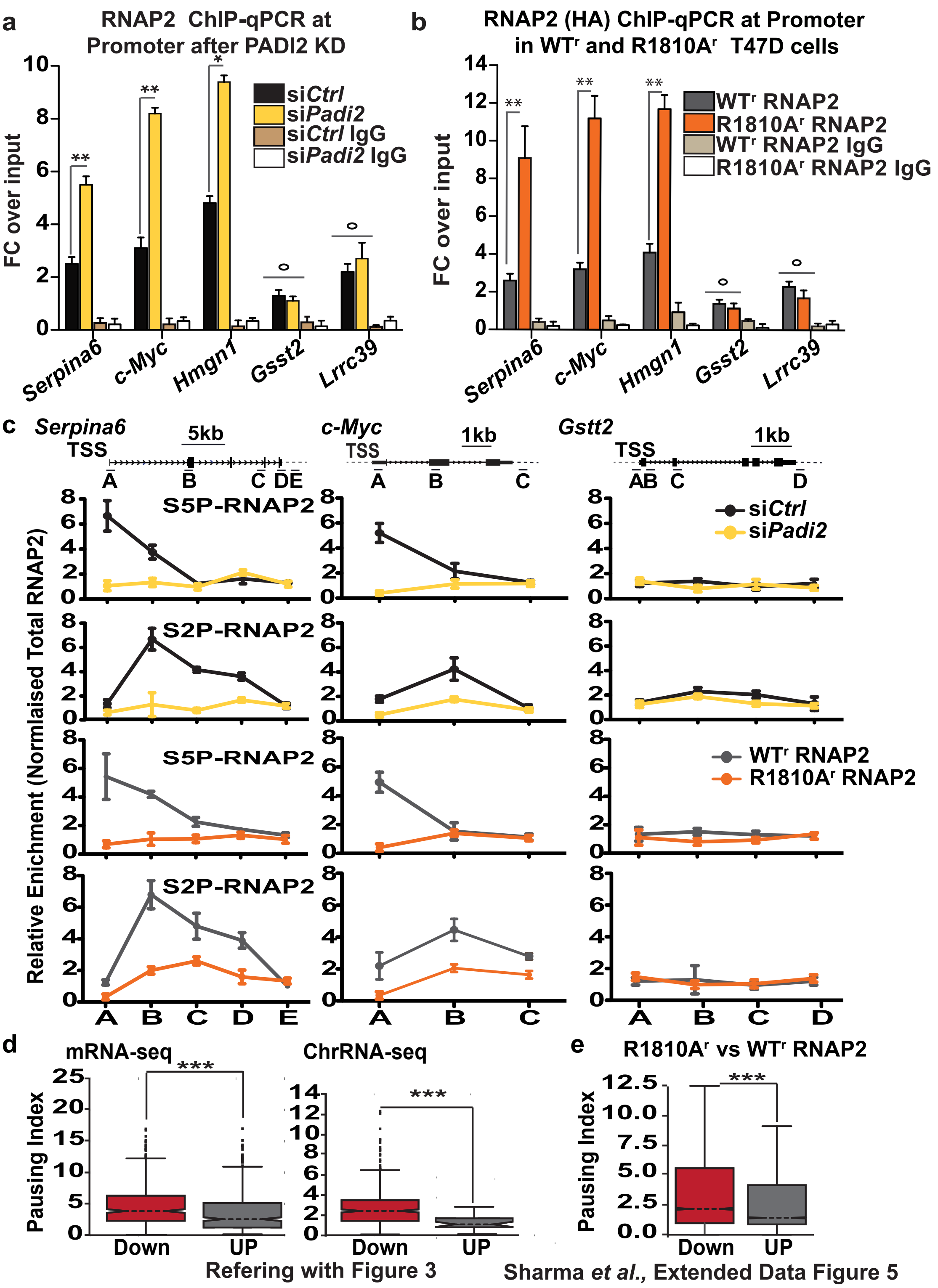
(i) Word cloud of all the PADI2 dependent genes represented in (**h**); size of genes indicates extent of down-regulation after PADI2 depletion as measure by the log2 fold change si*Ctrl*/si*Padi2***. (j-l)** Representated gating strategy used for above mentioned FACS analysis (**j**) Reference dot plot showing morphological related parameters, main selected cell population area marked as ¨P1¨, x-axis: FSC-A (forward scatter area), y-axis: SSC-A (side scatter area). **(k)** Single cell analysis dot plot from selected ¨P1¨ and marked as ¨P2¨ to exclude cell aggregates, x-axis: propidium iodide (PI) area and y-axis: PI width. **(l)** Cell cycle profile indicated x-axis as DNA content (PI-area) and y-axis as cell counts.



Refering with Figure 3

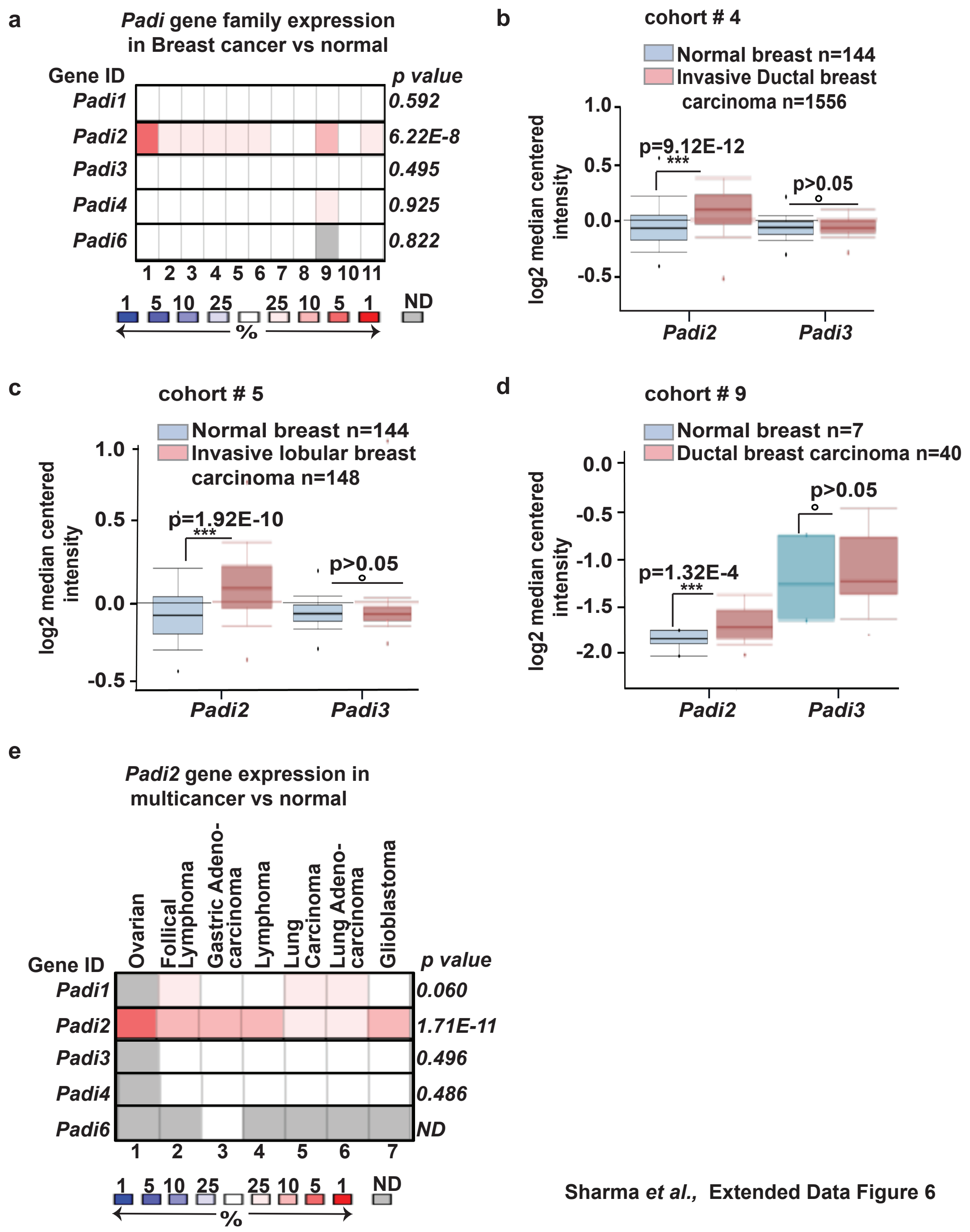
Sharma *et al.*, Extended Data Figure 4

Extended Data Figure 4. RNAP2 and PADI2 co-localize on highly transcribed genes. **(a)** ***Top:*** Spie chart showing the distribution of PADI2 ChIP-seq peaks over various genomic regions. Dashed curved line indicates the region from 3kb upstream of TSS to 3kb downstream of TTS; the numbers in parenthesis show the proportion occupied by each region in the genome. ***Bottom****:* Enrichment of normalized PADI2 reads in various genome regions relative to random distribution (* p-value $<10^{-2}$; ** p-value $<10^{-3}$; *** p value $<10^{-4}$). (**b**) ***Top***: RNAP2 and PADI2 occupancy across genes classified with increasing levels of expression. p value was calculated by Wilcoxon-Mann-Whitney test in comparison to silent genes as indicated (** p value $< 10^{-3}$; *** p value $< 10^{-5}$). ***Bottom****:* Browser snapshots representing the PADI2 and RNAP2 occupancy on highly transcribed genes *Serpina6* and *Hmgn1* versus lowly expressed genes (*Gstt2* and *Lrrc39*. Scale is shown on the top for each gene. Y axis: reads per million (RPM). **(c)** PADI2 occupancy monitored along the gene bodies by PADI2 ChIP assay in T47D cells transfected with si*Padi2* or si*Ctrl* followed by qPCR along the gene bodies of two highly transcribed genes (*Serpina6*, *c-Myc*) and a low transcribed gene (*Gstt2*). Non-immune IgG was used as negative control. Y-axis: PADI2 enrichment over input samples. Data represented as mean ± SEM from at least three biological replicates as in other plots of the figure. Top: basic gene structure and scale with the positions of the amplicons. **(d)** PADI2 ChIP seq normalized reads in a window from 3kb upstream of TSS to 3kb downstream of TTS on genes downregulated and non-regulated genes after PADI2 depletion in mRNA-sequencing (left) and ChrRNA-sequencing (right). Downregulated genes showed significantly higher PADI2 recruitment both from mRNA-seq (p value=2.06e^{-28}) and ChrRNA-seq (p value =1.53e^{-72}); p value was calculated by Wilcoxon-Mann-Whitney test. **(e)** RNAP2 ChIP followed by PADI2 re-ChIP assay performed in T47D cells to examine co-localization on promoter regions (A primer) and exons (B or C primers) of two highly transcribed genes (*Serpina6*, *c-Myc*) and a low transcribed gene (*Gstt2*). (**e**) Higher pausing index in cell expressing R1810A^{r} mutant as compared to WT^{r} form of RNAP2 calculated for shared 939 genes downregulated after PADI2 depletion (2,186) and R1810 dependent (1392 genes found significantly downregulated by considering FC < 1/1.5 and p value < 0.01, after expressing R1810A^{r} mutant as compared to WT^{r} RNAP2) in T47D.



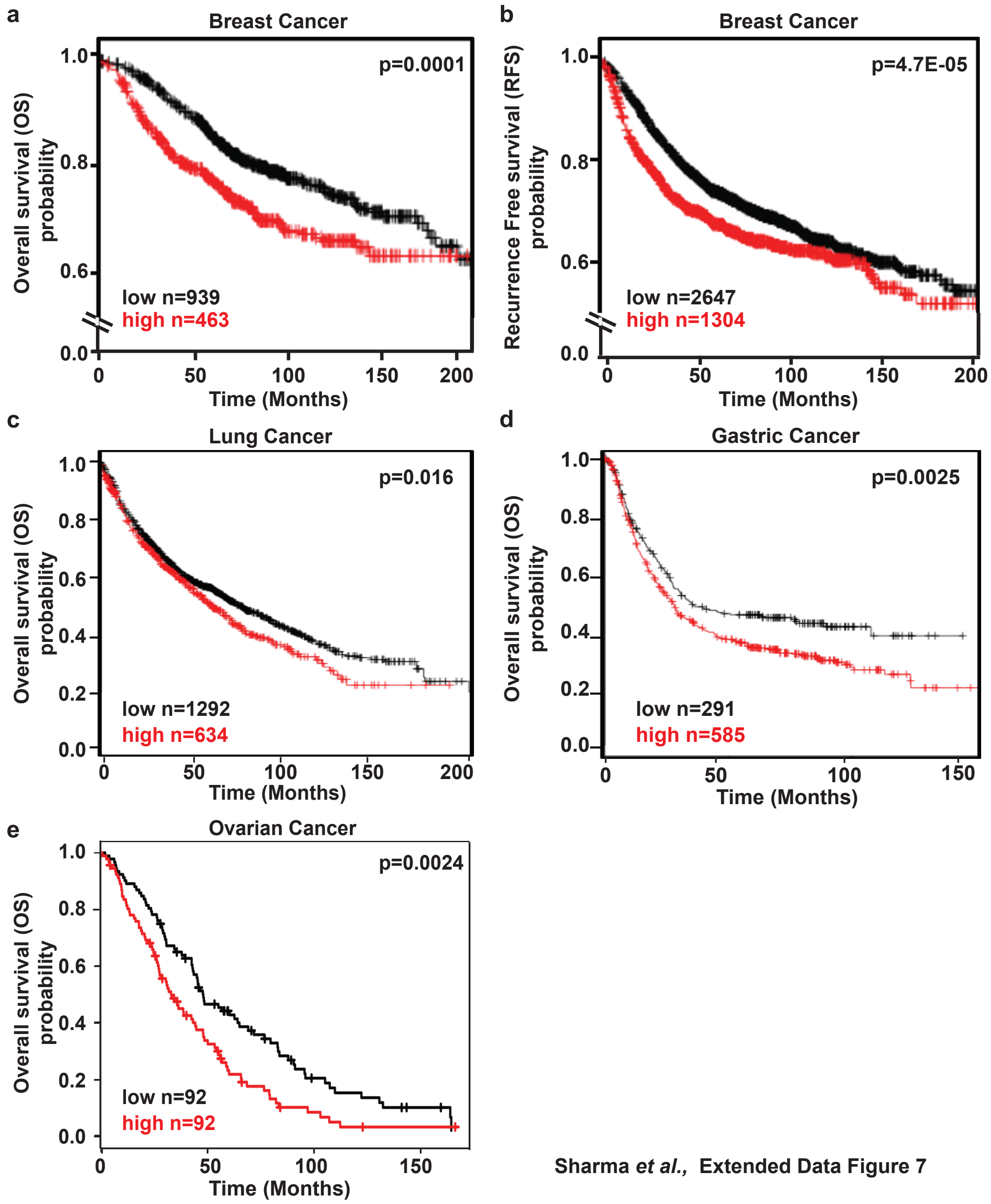
Extended Data Figure 5. PADI2 depletion leads to RNAP2 pausing of highly transcribed genes. **(a-b)** RNAP2 ChIP qPCR assay performed in T47D cells after PADI2 depletion with an antibody to RNAP2 (**a**) or in cells expressing only the HA-tagged wild type (WT^r) or R1810A^r mutant of RNAP2 with anti-HA antibody (**b**). Non-immune IgG was used as negative control. Y-axis: fold change over the input samples. **(c)** S5P and

S2P RNAP2 occupancy was monitored along the gene bodies by performing ChIP assay in T47D cells transfected with si*Ctrl or* si*Padi2 (black or yellow),* or T47D cells expressing only WT^r or the R1810A^r mutant of RNAP2 (grey or orange) followed by quantitative PCRs along the gene bodies. Y-axis: relative enrichment normalized to total RNAP2; values are means ± SEM. *Top*: For each gene, the basic gene structure and the position of the amplicons are indicated. (**d-e**) Box plot of the pausing index of (**d**) top 25% down-regulated and up-regulated genes after PADI2 depletion from mRNA-seq (*left*, p value =3.43e^{-09}) and ChrRNA-seq (*right*, p value =2.09e^{-02}). (**e**) Downregulated (n=1392) and upregulated (n=1372) genes obtained from replicate RNA-seq performed in T47D cells expressing R1810A^r mutant as compared to WT^r form of RNAP2 (p value =7.36e$^{-48)}$. Each box represents the interquartile range; Whisker extend the box to the highest and lowest values. Line across each box indicate the medium value. p value was calculated by Wilcoxon-Mann-Whitney test).



Sharma *et al.*, Extended Data Figure 6

Extended Data Figure 6. ***Padi2* overexpressed in breast and other cancers. (a)** *Padi* gene family expression in breast cancer cohorts (n=2155 patients). Each column represents a breast cancer cohort (detailed information of all the 11 data sets analyzed are given in extended method section Table 3). Scale on the bottom is the best gene rank percentile; color indicated overexpression (red) and underexpression (blue) compared to normal tissue from same study. The p-value for a gene is its p-value for the median-ranked analysis. Note that only over expression of *Padi2* but no other *Padi* family member is significant**. (b-d)** *Padi2* and *Padi3* expression level were analyzed in various subtypes of breast cancer compared to normal breast: (**b**) Cohort-4, normal breast vs invasive ductal breast carcinoma; (**c**) Cohort-5, normal breast vs invasive lobular breast carcinoma; (**d**) Cohort-9, normal vs ductal breast carcinoma. Y-axis represents log2 median intensity. The p value indicates the significance of difference. (**e**) *Padi* gene family expression in 7 different cancer cohorts, represented by each column (details are given in Extended Data Table 2). The p-value for a gene is its p-value for the median-ranked analysis. Scale at the bottom is as (6a).



Sharma *et al.*, Extended Data Figure 7

Extended Data Figure 7. Elevated levels of PADI2 in patient samples associates with poor prognosis. (**a-b**) Kaplan-Meier survival graph segregated according to PADI2 expression, high (> 75% percentile) or low (<25% percentile): (**a**) overall survival; (**b**) recurrent free survival for breast cancer patients. (**c-e**) Kaplan-Meier survival graph showing overall survival for (**c**) lung cancer (**d**) gastric cancer and (**e**) ovarian cancer (GSE26712). The number of patients and the p value of the comparison are indicated.

Extended Data Table 1- PADI2 cell cycle dependent genes

***Ccnd1,Ccng1,Cdc6,Cdkn1a,Cdkn2b,Cdc27,Cdc25c,Cdk2ap1,Cdc45,Cdca5,Stmn1,
Cdc20,Ccne2,Cdkn2a,Plk1,Myc,Gspt1,Rhob,Dhrs2,Ncaph,Mphosph9,Cks2,Pole,
Mphosph6,Cit,Nusap1,Kif2c,Sik1,Sh3BP4,Notch2,Tgfb1,Mcm2,Kif23,Ccnk,Mad2l2,
Rad54l,Myo16,Tubg1,Gmnn, Mad2l1,Pkmyt1,Acvr1b,Bmp4,Ndc80,Cited2,Pa2g4,
Chek2,Espl1,Kif15,Rad51,Birc5,Cul2,Bub1b,Nde1,Zwint,Kpna2,Gas1,Dbf4,Tpx2,
Ccna2,Cdkn3,Uhmk1,Eif4g2,Ube2c,Cdk2,Tbx3,Prc1,Nek2,Cenpf,Xrcc2,Sun2,
Ccp110,Rad51b,Ccnd3,Leprel4,Zbtb17,Nae1,Strn3,Pml,Rad54b,Chmp1a,
Pafah1b1,Pten,Fbxo5,Timeless,Pds5b,Dlgap5,Kif11,Khdrbs1,Tp53,Acvr1,
Tardbp,Sugt1,Map2k6,Racgap1,Pfdn1,Aurka,Anapc11,Dctn3,Pold1,Kat2b,***

PADI2 cell cycle independent genes

***E2F1,Rassf1,Cks1b,Hus1,Top3a,Herc5,Cdk10,Brac2,Rbbp8,kiF22,Anln,Pcbp4,Mapk12,
Bub1,Smc4,Ercc2,Dtymk,Pim2,Cdc37,Triap1,Ppm1g,Anapc2,Ppp5c,Cdc7,Anapc5,Rcc1,
Sertad1,Cdc16,Gps1,Cdc123,Inhba,Hspa2,Chek1,Ran,Upf1,Bccip,Ing4,Fancg,Ckap5,
Sac3d1,Ddb1,Cdk4,Dctn2,Kntc1,Rad21,Prmt5,Smc1a,Usp16,Ubb,Egf,Dlg1,Nek6,Ercc3,
Cdc25a,Tbrg4,Cul4a,Papd7, npm1,Cetn3,Akap8,Skp2,Ddx11,Cdc25b,Hcfc1,Anapc10,
Jmy,Pin1,Nasp,Katna1,Map3k11,Cdkn1c,Asns,Ppp6c,Sssca1,Lats2,Tmem8b,Rad50,
Hexim1,Rb1,Atr,Mnat1,Tpd52l1,Arap1,Gas7,Spdya,Numa1,Chfr,Mdm4,Sesn1,Cul5,Ahr,
Anapc4,Smc3,Hexim2,Mlf1,Pbrm1,Cdc23,Cdkn2c,Ccnt2,Rad52,Gtpbp4,Rprm,Cdk7,
Gadd45a,Cdkn1b,Msh5,Mre11a,Cdk5rap3,Inha,Ppp2r3b,Rabgap1,Ppp1r15a,Lats1,Cdk13,
Tp53bp2,Ccnt1,Myh10,Rad17,Tbrg1,Tgfb2,Cep250,Dmc1,Cdkn2d,Stk11,Nek11,Sphk1,
Madd,Foxn3,Sass6,Tgfa,Rad51d,Zw10,Clip1,Nolc1, Npm2,Apbb1,Xpc,Foxo4,Hpgd,Foxc1,
Diras3,Bmp7,Smad3,Rad9a,PPP1r13b,Mfn2,Epgn,Gps2,Ccng2,Atrip,Pola1,Cul3,Cntrob,
Rad1,Rint1,Lig3,Nbn,Cdk5rap1,Atm,Taf1,Ttk,Cul1,Tipon,Cenpe,Abl1,APC,,Cdk5r1,Rec8,
Dusp13,Brsk1,Cdk6,Plagl1,Krt7,Cdt1,Uhrf2,Ppp1r9b,Apbb2,Tube1, Pam,Gtf2h1,Ern1,***

Extended Data Table 2- Cancer cohorts used in Cancer versus Normal analysis

Figures	Column	Study	Samples (Normal/Cancer)
Extended Data Figure 6a	1	Breast Neoplasm vs Normal[38]	144/10
Extended Data Figure 6a	2	Ductal Breast Carcinoma vs Normal [38]	144/148
Extended Data Figure 6a	3	Invasive Ductal / Lobular Carcinoma vs Normal[38]	144/90
Extended Data Figure 6a-b	4	Invasive Ductal Breast Carcinoma vs Normal[38]	144/1556
Extended Data Figure 6a-c	5	Invasive Lobular Breast Carcinoma vs Normal [38]	144/148
Extended Data Figure 6a	6	Medullary Breast Carcinoma vs Normal [38]	144/32
Extended Data Figure 6a	7	Mucinous Breast Carcinoma vs Normal [38]	144/46
Extended Data Figure 6a	8	Tubular Breast Carcinoma vs Normal[38]	144/67
Extended Data Figure 6a-d	9	Ductal Breast Carcinoma vs Normal [39]	7/40
Extended Data Figure 6a	10	Invasive Ductal /Lobular Carcinoma vs Normal [40]	111/14
Extended Data Figure 6a	11	Medullary Breast Carcinoma vs Normal[40]	111/4
Extended Data Figure 6e	1	Ovarian Serous Adenocarcinoma vs Normal [41]	4/6
Extended Data Figure 6e	2	Follicular Lymphoma vs Normal [42]	5/5
Extended Data Figure 6e	3	Gastric Mixed Adenocarcinoma vs Normal[43]	19/10
Extended Data Figure 6e	4	Diffuse Large B Cell Lymphoma vs Normal[44]	20/44
Extended Data Figure 6e	5	Large Cell Lung Carcinoma vs Normal [45]	65/19
Extended Data Figure 6e	6	Lung Adenocarcinoma vs Normal[45]	65/45
Extended Data Figure 6e	7	Glioblastoma vs Normal [46]	4/80
Extended Data Figure 7a		Breast Cancer[47]	939 low versus 463 high PADI2 level
Extended Data Figure 7b		Breast Cancer[47]	2647 low versus 1304 high PADI2 level
Extended Data Figure 7c		Lung Cancer[49]	1292 low versus 634 high PADI2 level
Extended Data Figure 7d		Gatric Cancer[50]	291 low versus 585 high PADI2 level
Extended Data Figure 7e		Ovarian Cancer[51]	92 low versus 92 high PADI2 level