The RNA Polymerase II factor RPAP1 is critical for
Mediator-driven transcription and cell identity

Cian J Lynch\textsuperscript{1,2}, Raquel Bernad\textsuperscript{1,2}, Isabel Calvo\textsuperscript{1}, Sandrina Nóbrega-Pereira\textsuperscript{1,3}, Sergio Ruiz\textsuperscript{4}, Nuria Ibarz\textsuperscript{5}, Ana Martinez del Val\textsuperscript{5}, Osvaldo Graña\textsuperscript{6}, Gonzalo Gómez-López\textsuperscript{6}, Eduardo Andrés-León\textsuperscript{6}, Vladimir Espinosa Angarica\textsuperscript{7,8}, Antonio del Sol\textsuperscript{7}, Sagrario Ortega\textsuperscript{9}, Oscar Fernandez-Capetillo\textsuperscript{4}, Enrique Rojo\textsuperscript{10}, Javier Munoz\textsuperscript{5}, Manuel Serrano\textsuperscript{1,2,11}\textsuperscript{*}

\textsuperscript{1} Tumour Suppression Group
Spanish National Cancer Research Centre (CNIO)
Madrid 28029
Spain

\textsuperscript{2} Cellular Plasticity and Disease Group
Institute for Research in Biomedicine (IRB Barcelona)
The Barcelona Institute of Science and Technology (BIST)
Barcelona 08028
Spain

\textsuperscript{3} Instituto de Medicina Molecular
Faculdade de Medicina
Universidade de Lisboa
Lisboa 1649-028
Portugal

\textsuperscript{4} Genomic Instability Group
Spanish National Cancer Research Centre (CNIO)
Madrid 28029
Spain

\textsuperscript{5} ProteoRed-ISCIII Proteomics Unit,
Spanish National Cancer Research Centre (CNIO),
Madrid 28029, Spain

\textsuperscript{6} Bioinformatics Unit
Spanish National Cancer Research Centre (CNIO)
Madrid 28029
Spain

7 Luxembourg Centre for Systems Biomedicine
Campus Belval, Biotech II, 0.04
University of Luxembourg
L-4366 Esch-sur-Alzette
Luxembourg

8 Cancer Science Institute,
National University of Singapore
117599 Singapore

9 Transgenic Mouse Unit
Spanish National Cancer Research Centre (CNIO)
Madrid 28029
Spain

10 Department of Plant Molecular Genetics
National Center of Biotechnology (CNB-CSIC)
Madrid 280049
Spain

11 Catalan Institution for Research and Advanced Studies (ICREA)
Barcelona 08010
Spain

* Correspondence: manuel.serrano@irbbarcelona.org

Reviewer access to large datasets

- For Reviewer access to the mass spectrometry proteomics dataset, please go to the ProteomeXchange Consortium/PRIDE repository, Accession: PXD007114, where we have generated this Reviewer access token: Username: reviewer51249@ebi.ac.uk; Password: GUEzZQwc.

- For Reviewer access to RNA-seq and ChIP-seq data, please go to the GEO database, Accession:GSE78795, where we have generated this Reviewer Token: arwheeyipfcfcj.
eTOC Highlights

- RPAP1 is a novel RNA Pol II regulator, conserved from plants to mammals
- RPAP1 depletion erases cell identity gene expression, triggering de-differentiation
- Mechanistically, RPAP1 is critical for the Mediator-Pol II interaction
- RPAP1 preferentially contributes to enhancer-driven gene transcription

Blurb

Lynch et al report a novel regulator of RNA Pol II called RPAP1, displaying functional conservation from plants to mammals. RPAP1 is required to establish and maintain cell identity. Mechanistically, RPAP1 is critical for the Mediator-Pol II interaction, thereby preserving normal transcription at enhancer-driven genes.
SUMMARY

The RNA Polymerase II Associated Protein 1 (RPAP1) is conserved across metazoa and critical for stem cell differentiation in plants, however, very little is known about its mechanism of action, or its role in mammalian cells. Here, we report that RPAP1 is essential for the expression of cell identity genes and for viability. Depletion of RPAP1 triggers cell de-differentiation, facilitates reprogramming towards pluripotency, and impairs differentiation. Mechanistically, RPAP1 is essential for the interaction between Pol II and Mediator, as well as for the recruitment of important regulators, such as the Mediator-specific Pol II factor POLR2M/Gdown1 and the CTD phosphatase RPAP2. In agreement, depletion of RPAP1 diminishes the loading of Pol II and Pol II Ser5 phosphorylation levels and decreases expression of super-enhancer-driven genes. We conclude that Mediator-RPAP1-Pol II is an ancient module, conserved from plants to mammals, critical for establishing and maintaining cell identity.

Key words: transcription; RNA polymerase II; Mediator; cell identity; differentiation
INTRODUCTION

Coordinated regulation of RNA Polymerase II (Pol II) transcription is central to cell identity transitions, and reflects a common developmental principle across the plant-animal divide (Gaillochet and Lohmann, 2015; Levine, 2011; Meyerowitz, 2002). High-throughput studies have recently revealed a set of conserved RNA Pol II-Associated Proteins (RPAP1, 2, 3, 4) sharing multiple interactions among themselves (Jeronimo et al., 2004, 2007). RPAP2 is an atypical phosphatase that targets Ser5P on the Pol II carboxy-terminal domain (CTD) (Egloff et al., 2012a; Mosley et al., 2009), and RPAP2, RPAP3 and RPAP4 all have essential roles as nuclear transport chaperones for the Pol II complex (Boulon et al., 2010; Forget et al., 2010, 2013). In contrast, the function of RPAP1 remains uncharacterised in mammals.

RPAP1 is a large (153 kDa) multidomain protein with a high degree of conservation across species (Jeronimo et al., 2004, 2007; Sanmartín et al., 2011). Studies in plants, yeasts and mammals indicate that RPAP1 interacts with the RPB3 and RPB11 subunits of the Pol II complex (Giaever et al., 2002; Ito et al., 2001; Jeronimo et al., 2004, 2007; Sanmartín et al., 2011). Importantly, the heterodimer RPB3/RPB11 provides a critical interface of Pol II with the Mediator complex (Allen and Taatjes, 2015; Davis et al., 2002). Indeed, a high-throughput screen in yeast indicated that depletion of RPAP1 results in dramatic gene expression changes that were similar to depletion of the Pol II subunit RPB11, although these changes were not characterized further (Jeronimo et al., 2004, 2007).
The multiprotein Mediator complex associates with transcriptional enhancers through protein-protein interactions, being critical for enhancer-promoter looping (Allen and Taatjes, 2015). The largest accumulations of Mediator are in super-enhancers, and super-enhancer target genes are typically the most important for defining cell identity and the most heavily dependent on Mediator to drive their transcription by Pol II (Allen and Taatjes, 2015; Hnisz et al., 2013; Kagey et al., 2010; Whyte et al., 2013).

RPAP1 was recently identified in plants as a critical factor for differentiation through promoting developmental gene expression (Muñoz et al., 2017; Sanmartín et al., 2011). Specifically, in *Arabidopsis*, RPAP1 was necessary and rate limiting to initiate stem cell differentiation (Sanmartín et al., 2011, 2012). Based on this, we hypothesized that mammalian RPAP1 may also coordinate gene expression and cell identity at a global level. Here, we characterize the mammalian homolog of RPAP1 to investigate putative roles in mammalian transcription and differentiation, and reveal a mechanism involving direct RNA Pol II regulation by RPAP1 through interaction with Mediator.

**RESULTS**

**Mammalian RPAP1 expression**

The plant homolog of RPAP1 is highly expressed in stem cells compared to differentiated cells (Sanmartín et al., 2011). Based on this, we began by examining RPAP1 expression in pluripotent and differentiated mouse cells. Compared to adult tissues or mouse embryonic fibroblasts (MEFs), RPAP1
protein levels were high in embryonic stem (ES) cells, induced pluripotent stem (iPS) cells, embryo carcinoma (P19EC) cells, and embryoid bodies (EBs) (Figures 1A, 1B). Moreover, RPAP1 expression levels decreased during in vitro differentiation of ES cells by LIF removal and retinoic acid addition (Figure 1C; Figure S1A). In the case of plants, RPAP1 in stem cells is cytoplasmic and only enters into the nucleus upon differentiation, suggesting that RPAP1 functions as a differentiation switch (Sanmartín et al., 2011). Interestingly, we observed a similar behaviour in mouse cells. In particular, RPAP1 was mostly cytoplasmic at the very earliest stages of development in the morula and blastocyst (Figure 1D), as well as in ES cells undergoing self-renewal (Figure 1E and 1F). However, RPAP1 became partly nuclear upon ES cell differentiation (Figure 1E), and completely nuclear in differentiated cells and tissues (Figure 1G and 1H; S1B). Indeed, we could detect enrichment of RPAP1 at gene promoters soon after launching differentiation (Figure 1I). Moreover, treatment of ES cells with the nuclear export inhibitor leptomycin B produced rapid nuclear accumulation of RPAP1 (Figure 1F), which, similar to plants, is consistent with active nuclear export of RPAP1 during stem cell self-renewal. Therefore, mammalian RPAP1 shares similar expression and subcellular localization dynamics as observed in plants during the switch between self-renewal and differentiation.

Dependence of ES or differentiated cells on RPAP1 expression

To assess the relevance of RPAP1 in cells, we first identified shRNAs that efficiently downregulated RPAP1 (Figure 1G; S1C; see also below). RPAP1
knockdown in non-pluripotent cells, such as human 293T, monkey COS7, various human cancer cell lines, murine MEFs and immortalized primary hepatocytes, severely attenuated proliferation, induced senescence, and triggered apoptosis typically with a delay of 2-6 days (Figures 1J, 1K; and Figures S1D-G). These observations were recapitulated using a total of three different shRNAs against RPAP1 (Figure S1F). Interestingly, while 95% knockdown of RPAP1 expression had no effect on ES cell viability during self-renewal (Figures 1J, 1K; Figures S1C, S1D and S1H), we were unable to obtain viable ES clones with complete RPAP1-knockout. It is important to note that we successfully targeted RPAP1 using multiple independent CRISPR delivery systems (transient, constitutive, or inducible), guide RNAs, and several wild-type or haploid ES cell lines. In particular, we obtained many ES clones where RPAP1 suffered small deletions but never a complete loss. Also, when using an ES cell line with a LacZ reporter knocked-in within intron 8 of the RPAP1 gene, we were able to efficiently eliminate LacZ expression using guide RNAs against the first 7 exons of RPAP1, however, we never obtained clones with elimination of the remaining wild-type RPAP1 allele (Figure S1I to S1L; see Methods). Taken together, the data suggest that RPAP1 performs an essential function in all the cell types tested, including ES cells.

**Inhibition of RPAP1 affects ES cell differentiation**

While depletion of ~95% RPAP1 did not affect pluripotent cells under self-renewal conditions (see above), we next assessed differentiation by LIF removal for 24 or 72 hours (Savatier et al., 1996). We observed that RPAP1-
depleted ES cells presented a delayed differentiation based on the expression of pluripotency markers and morphological changes, followed by an increase in apoptosis (Figure S2A and S2B). Differentiation of ES cells to embryoid bodies (EBs) by hanging-drop culture constitutes a longer term and more complex in vitro differentiation assay. RPAP1 depletion in ES cells followed by EB differentiation resulted in severely reduced efficiency of cardiac center development (formation of beating cell clusters) in EBs (Figure 2A). In agreement, analyses of RNA expression also revealed a delay in the loss of pluripotency markers and delayed induction of cardiac muscle differentiation markers associated with RPAP1-depleted EBs. (Figures 2B and S2C), suggesting that a decrease in RPAP1 expression is incompatible with development. Impaired cardiac center formation by RPAP1-depleted ES cells may reflect their impaired capacity to differentiate and/or the accumulation of dying or dysfunctional cells. Consistent with a developmental defect, RPAP1(+/−) ES cells, displayed weak contribution to chimeric off-spring (10 from 254 micro-injected embryos) (Figure S2D). Furthermore, in chimera crosses to look for germline transmission, we did not obtain mice which were RPAP1(+/-) or RPAP1(-/-) (0 out of 156 pups born) (Figure S2D).

To characterize the influence of RPAP1 on early events during the pluripotency-to-differentiation transition, we performed RNA-seq analyses in ES cells, control and RPAP1-depleted, after 24 hours of differentiation (LIF removal). Of 12,827 transcripts detected, 899 (7.1%) were significantly differentially expressed in RPAP1-depleted cells (Figure S2E; Table S1). Global investigation via geneset enrichment (GSEA) and supervised network
analyses indicated that following RPAP1 depletion, differentiating ES cells maintained proliferation pathways (Myc and E2F-regulated genesets were significantly higher) and had an attenuated induction of mesenchymal identity (epithelial-mesenchymal transition (EMT)-related genesets were lower) compared to the controls, including key mesenchymal genes, such as Ctgf, Mest and Col4a2 (Figures S2F to S2I; Tables S1 and S2). This is consistent with the delayed loss of pluripotency markers and morphological changes observed upon differentiation of RPAP1 depleted ES cells (Figures 2A, 2B and S2C). Thus, RPAP1 depletion delayed ES cell differentiation, suggesting that high levels of RPAP1 endow ES cells with the ability to rapidly differentiate, whereas reduced levels of RPAP1 dramatically slow differentiation.

Inhibition of RPAP1 induces loss of differentiated identity in MEFs

Since RPAP1 depletion impaired ES cell differentiation, we investigated the role of RPAP1 in differentiated cells. Following RPAP1 depletion in MEFs, cells proliferated and still appeared morphologically normal during days 1-3, prior to the defects which subsequently emerged at days 4-6 (see above Figures 1I and 1J; S1D). Thus, RNA-seq was performed at day 3 in control or RPAP1-depleted MEFs to assess the transcriptome while avoiding death-related secondary effects. Nevertheless, transcriptomic alterations were dramatic, with >52% of the 12,249 genes detected displaying significantly altered expression (FDR q<0.05; Figure 2C; Table S3). Using multiple approaches to assess gene expression, including GSEA, gene ontology and supervised network analysis, we observed that RPAP1 triggered a rapid and pronounced loss of
multiple developmental processes and robust erasure of fibroblast identity within 3 days (Figures 2F to 2H; and S2J, S2K; Table S2 and S3), a sequence which was initiated within 24hrs as confirmed by qRT-PCR for multiple mesenchymal/ fibroblastic identity markers (Figures 2G and S2L). Notably, there was a remarkable parallel between the genesets that were downregulated in MEFs by RPAP1 loss and the genesets that failed to be upregulated in differentiating RPAP1-depleted ES cells (Figure 2H). Lastly, a defining feature of mesenchymal cell identity is a high capacity for cell migration. Consistent with the above gene expression profile, RPAP1-depletion followed by a wound healing scratch assay revealed an attenuation of MEF migration capacity (Figure 2I). In summary, after RPAP1 depletion, MEFs display rapid de-differentiation via loss of mesenchymal-fibroblastic identity.

RPAP1-knockdown favors de-differentiation and reprogramming

Since RPAP1 is important for maintaining the mesenchymal cell identity of MEFs, we hypothesized that RPAP1 depletion may recapitulate an early stage of reprogramming to iPS. Previous investigators have found that during reprogramming there is an initial de-differentiation wave followed by a transient intermediate state, which is resolved by a second wave of transcriptional changes leading to pluripotency (Polo et al., 2012). Interestingly, the gene expression profile induced by RPAP1 depletion was significantly similar to the intermediate state of reprogramming (Figures 3A, S3A and S3B). This was supported by validation with markers of the intermediate state (Polo et al., 2012), including downregulation of Meox1 and Meox2, and upregulation of
Nup210 (Figure 3B). This suggested that RPAP1 knockdown phenocopies the de-differentiation and loss of mesenchymal identity observed in the first wave of transcriptional changes during iPS reprogramming. Consistent with this, prior knockdown of RPAP1 for 2 days in MEFs led to significantly enhanced iPS reprogramming with the four Yamanaka factors (Oct4, Sox2, Klf4 and cMyc, abbreviated as OSKM) (Figures 3C and 3D). Importantly, therefore, the lethality of RPAP1 depletion in MEFs was rescued by reprogramming to pluripotent iPS cells, suggesting that RPAP1-depleted MEFs at days 2-3 may represent de-differentiated cells without a defined identity, which subsequently collapse unless rescued re-directed by reprogramming into pluripotency.

To explore the minimal complement of the Yamanaka factors sufficient to rescue lethality of RPAP1 depletion and/or confer pluripotency, we tested all possible combinations of Oct4, Sox2, Klf4, cMyc and Nanog (32 combinations; OSKMN; Figure 3E), in combination with a panel of media supplements reported to enhance reprogramming (15 media cocktails; Figure S3C). Four interesting features emerged: (i) RPAP1 knockdown plus several of the transcription factor combinations including Klf4 or cMyc were sufficient to rescue cell survival, in particular, shRPAP1 with Klf4/cMyc together converted the majority of MEFs to putative intermediates of reprogramming, that is, rapidly proliferating colony-forming cells which were also positive for markers of the early stages of the reprogramming process, including alkaline phosphatase and SSEA1 cell surface expression, but were Sox2-eGFP-negative (Figures 3E to 3G; and Figure S3D); (ii) RPAP1 depletion increased the efficiency of all successful reprogramming combinations (Figure 3E); (iii) RPAP1 depletion can
replace Sox2 in combination with OKM or OKMN (Figures 3E to 3G); and (iv) pharmacological inhibition of TGFβ signaling, which is known to replace Sox2 (Li et al., 2010), cooperated with RPAP1 depletion in the OKM or OKMN reprogramming (Figure S3E). Taken together, phenotypic and expression data suggest that RPAP1 depletion induces a de-differentiated state that can be stabilized by Klf4/cMyc, and can be converted into full pluripotency if Oct4 is included.

RPAP1 regulates the Pol II interactome, not its expression or localization

To understand the mechanism by which RPAP1 is required for somatic cell proliferation, we first wondered if RPAP1 could affect the stability and localization of Pol II. The Pol II complex is formed by 12 subunits (RPB1-12, RPB1 being the largest and catalytic subunit) and 4 associated proteins (RPAP1-4) (Wild and Cramer, 2012). The full complex is assembled in the cytoplasm and remarkably individual depletion of the subunits RPB2 to 12 or RPAP2 to 4 prevents nuclear import of the catalytic subunit RPB1 (Boulon et al., 2010; Forget et al., 2010, 2013; Wild and Cramer, 2012), whereas the effect of RPAP1 depletion has not been reported. Therefore, we assessed the effect of RPAP1 knockdown on Pol II expression and localization in five different cell lines. In contrast to the above-mentioned subunits, RPAP1 depletion did not affect total Pol II expression levels, its global modification status by phosphorylation on serine 5 (Ser5P) or serine 2 (Ser2P) (Figure 4A and Figure S4A and S4B), or its nuclear localization (Figure 4B and S4C). This was confirmed by immunofluorescence (Figure 4C and Figure S4D). These
observations rule out Pol II destabilization and/or mislocalization as an explanation for the essential role of RPAP1 in the survival of differentiated cells.

We further investigated the mechanism by which RPAP1 might regulate Pol II. Since RPAP1 is a large protein directly associated with Pol II (see: Introduction), we compared the Pol II protein interactome of control vs. RPAP1-knockdown MEFs in order to detect changes in Pol II function through alterations in its binding partners. Immunoprecipitation of RPB1, the largest and core Pol II sub-unit, followed by mass spectrometry, revealed 294 specific interactor proteins (Figure 4D; see: Methods), with a clear enrichment for transcription-related factors, including for example all 12 subunits of the Pol II complex and almost all (28 out of 30) of Mediator subunits, illustrating the depth and specificity of this interactome analysis (Table S4). Importantly, RPAP1-knockdown did not affect the integrity of the Pol II complex itself, but it resulted in a significant reduction of 104 Pol II interactors (red circles in Figure 4D; see also Table S4), while 5 new interactors were found (green circles). Amongst Pol II interactors significantly affected by RPAP1-depletion, the Mediator Complex was ranked the highest in terms of proportion of affected subunits (Figure 4E; Table S4) suggesting an important alteration in the functions controlled by this complex. Furthermore, we observed that depletion of RPAP1 led to the loss of POLR2M (also known as Gdown1) from Pol II complexes. Gdown1 is a recently discovered protein that tightly binds approximately half of Pol II in cells, forming the so-called Pol II(G) complex (Hu et al., 2006; Jishage et al., 2012). Importantly, Pol II(G) complexes are known to contain RPAP1 (Jishage et al., 2012). We note that Gdown1 is recruited by Mediator and
associates with Pol II on Mediator-regulated target genes (Cheng et al., 2012; Hu et al., 2006; Jishage et al., 2012; Li and Price, 2012). Altogether, we conclude that RPAP1 is a critical ingredient for Mediator-competent Pol II.

RPAP1 is required for transcription of identity and developmental genes

Since Mediator has a critical role recruiting Pol II to genes controlling cell identity and development (Allen and Taatjes, 2015; D’Alessio et al., 2009; Hnisz et al., 2013; Whyte et al., 2013), we next investigated the global effect of RPAP1-depletion on Pol II binding to chromatin. For this, we performed ChIP-seq for total Pol II and for Ser5P Pol II, the latter reflecting active Pol II (Egloff et al., 2012b; Hsin and Manley, 2012). Knockdown of RPAP1 in MEFs reduced the abundance of both total and Ser5P Pol II at about 50% of detected genes, while very few genes (<0.5%) displayed an increase (Figures 5A-D; Table S5).

Interestingly, GSEA and Leading Edge analyses revealed that mesenchymal-regulators and related developmental processes were among the genesets (GSEA) and genes (Leading Edge) with the most significant loss of Pol II (Figures 5E, 5F and S5A; Table S2).

Pol II regulation at individual genes is often more complex than overall abundance, particularly in relation to two critical steps, namely, Pol II loading at promoters and transitioning into productive elongation (Chen et al., 2015; Liu et al., 2015; Rahl et al., 2010). Hence, we compared Pol II abundance at promoter versus gene body, by calculating the promoter-to-body ratio (often described as the “PI index”), as described (Chen et al., 2015; Rahl et al., 2010) (Figure S5B). Overall, in those genes with reliable Pol II signal, we observed
that 84% of the genes in MEFs had a promoter/body ratio > 2.0 (Figures S5C and S5D; Table S5), which is similar to published data in mouse ES cells (91%) (Rahl et al., 2010) or human cancer cells (90%) (Chen et al., 2015).

Following RPAP1 depletion, the promoter/body ratio was altered in many genes, in some cases it was increased and in others it was decreased (Figure S5E). Interestingly, while no significant genesets were enriched among those genes with decreased promoter/body ratios, genesets corresponding to regulators of cell identity and development were significantly present among the genes with increased ratios (Figure S5E; Table S5).

To investigate if RPAP1 depletion altered Pol II activity and abundance through altered Ser5P levels, we calculated the Ser5P/total Pol II ratio (also known as “Ser5P density”) for all genes at the promoters and gene bodies. We detected widespread changes in Ser5P density (Figure S5F), a phenomenon that has been observed before when Pol II elongation is blocked (Allepuz-Fuster et al., 2014). Notably, these changes were more pronounced at promoter regions than at gene bodies (Figure S5F). Moreover, GSEA analyses revealed that upon RPAP1 depletion, the only significantly enriched genesets were associated with increased Ser5P density at promoters, and these included genesets and GO terms such as TNFα signaling via NFκB, cell migration, locomotory behavior, and genes in which RNA levels are also downregulated (Tables S2 and S5), such as Snai2, Tgfb1i1, Tgfb3, Tgfbrap, Lox, Loxl1, Tlr2, Tlr3, Vegfa, Myo6, Smad6, Ccl7, S100a4 (Fibroblast-Specific Protein1) and S100a6 (Figure S5F; Table S5). Taken together, this suggests that RPAP1 depletion affects Pol II transcription, including the levels of Ser5P, and this
preferentially perturbs the expression of cell identity and developmental regulators.

In summary, upon RPAP1-knockdown in MEFs, the genes and genesets linked to the regulation of fibroblastic/mesenchymal identity or closely related developmental processes were the most significantly enriched in four key categories: (i) genes with the most significantly downregulated mRNA expression; (ii) genes with the greatest overall depletion of Pol II; (iii) genes with selective Pol II depletion from their gene body; (iv) genes with the most enhanced Ser5P density at their promoters.

Conservation of RPAP1 function from plants to mammals

Previously, it was shown that mutations of the RPAP1 homolog in plants inhibited cell differentiation, and microarray analyses showed a specific defect on developmental gene expression (Sanmartín et al., 2011). In order to directly compare the mouse and plant functional overlap, we converted the published plant differential gene expression data to the nearest mammalian protein homolog where possible (see Methods and Table S2). Interestingly, conversion of the plant expression data to mouse homologs also revealed significant downregulation of developmental processes (Figure 5F; Table S2). This suggests that loss of RPAP1 function in mice and plants downregulates similar developmental processes, including lineage specifiers and regulators of cell identity, such as hypoxia, cell polarity, extra-cellular matrix, and chemokine signaling.
**RPAP1 preferentially regulates Mediator-driven gene expression**

Mediator physically links enhancers with target genes, and then recruits Pol II for their transcriptional activation (Allen and Taatjes, 2015). This process is especially critical to maintain transcription of genes regulated by super-enhancers, which typically encode key markers and regulators of cell identity (Allen and Taatjes, 2015; Hnisz et al., 2013; Whyte et al., 2013). Given our observations above that RPAP1-depletion triggered both a decrease in Pol II interaction with the Mediator complex, and selective loss of cell identity gene expression, we next assessed the transcription of super-enhancer-driven genes. We found that following RPAP1-depletion in MEFs, the mRNA levels of genes proximal to super-enhancers were significantly decreased (Figure 5G), whereas, highly expressed housekeeper genes were not affected (Figure S5G).

Expression levels of enhancer RNAs (eRNAs) are proportional to their enhancer activity (Andersson et al., 2014; Li et al., 2016). In our RNA-seq, we detected eRNA expression in ~20% of super-enhancers, and we divided those enhancers into two groups, those with increased or decreased eRNA levels (Figure 5H). Interestingly, after RPAP1-depletion, enhancers with decreased eRNA levels (decreased activity) had target genes associated with Theiler Stages 20-25 (embryo day E11.5-17), while enhancers with increased eRNA levels (increased activity) had target genes associated with Theiler Stages 14-20 (E8-13) (Figures 5I and S5H). Since MEFs arise from E13.5 embryos, the data suggests that enhancers of this embryo stage are decreased in activity, while enhancers of earlier embryo stages are activated. This is consistent with the de-differentiation effects that we observed above in MEFs after RPAP1-
Taken together, this suggests that RPAP1 depletion affects Pol II transcription by disruption of the Mediator-Pol II interaction, and this preferentially reduces the expression of super-enhancer-driven cell identity and developmental regulators (Figures 5J and 5K).

**DISCUSSION**

We have characterized the function of mammalian RPAP1, and observed prominent parallels with its plant homolog, in terms of sub-cellular localization, developmental expression patterns, regulation of RNA Polymerase II transcription, and a requirement to establish and maintain differentiated cell identity. Based on this, we propose that this is an ancient mechanism to trigger the transition from pluripotency to differentiation.

**RPAP1 expression and nucleo-cytoplasmic shuttling**

We found that RPAP1 protein is largely cytoplasmic in pluripotent cells, which is consistent with the lack of effect of RPAP1 depletion on gene expression or proliferation, in self-renewing pluripotent cells. However, we were unable to isolate RPAP1-KO ES cells, suggesting that either ES cells require a small amount of RPAP1 for an essential function, or to maintain fast proliferation, under self-renewal conditions. Interestingly, we observed rapid nuclear accumulation of RPAP1 by blocking nuclear export, implying a continuous cycle of RPAP1 in/out of pluripotent cell nuclei. In contrast, the onset of
differentiation coincided with RPAP1 nuclear accumulation, observed both in vitro and in vivo, and recruitment to promoters together with Pol II. In fact, this developmental switch in nucleo-cytoplasmic shuttling is similar to the behaviour of the RPAP1 plant homolog (Muñoz et al., 2017; Sanmartín et al., 2011, 2012). This is also consistent with the existence of multiple conserved NLS/NES sequences on RPAP1 and a high homology ARM superfamily repeat region, a motif associated with nucleo-cytoplasmic shuttling, that is highly conserved in RPAP1 homologs of *Saccharomyces, Drosophila* and mammals (Jeronimo et al., 2004). Together, this suggests a conserved model for RPAP1 function in the mechanism for triggering development (Figure 5J).

**RPAP1 is required to establish and maintain cell identity**

During development, new cell identity can arise through a series of reversible epithelial-to-mesenchymal transitions (EMT) (Thiery et al., 2009). RPAP1 expression was required during ES cell differentiation, including towards cardiac muscle development, a path containing several EMT transitions (Thiery et al., 2009). Consistent with this, we failed to obtain homozygous RPAP1-null mice. Moreover, RPAP1 depletion resulted in a striking loss of the mesenchymal identity of MEFs and subsequent cell death. Similarly, all tested cell lines (a total of 8) died several days after RPAP1 depletion. Taken together, these data suggest a role for RPAP1 in both the establishment and maintenance of cell identity.

**RPAP1 depletion permits de-differentiation and reprogramming**
RPAP1 depletion induced a loss of MEF mesenchymal/fibroblastic identity. Strikingly however, such de-differentiation complemented the early stages of reprogramming to pluripotent iPS cells, and thus, RPAP1 depletion enhanced the efficiency of recapturing pluripotency. Therefore, reprogramming with OSKM rescued the lethality of RPAP1 depletion, a phenomenon we found could be attributed to the overexpression of Klf4 plus cMyc in particular. We hypothesize that Klf4/cMyc dual overexpression may revert or compensate the lethal effects of RPAP1 depletion since cMYC amplifies active Pol II transcription (Lin et al., 2012; van Riggelen et al., 2010), while the ectodermal lineage specifier KLF4 may help to specify a new epithelial identity. In this way, RPAP1 depletion plus Klf4/cMyc overexpression may stabilize a highly proliferative reprogramming intermediate.

**RPAP1 acts at the interface between RNA Pol II and Mediator**

RPAP1 is a large (153 kDa) multidomain protein that has been reported to bind a number of interesting Pol II regulators, most notably the RPB3/11 heterodimer and this is well substantiated in plants, yeasts and mammals (Giaever et al., 2002; Hazbun et al., 2003; Ito et al., 2001; Jeronimo et al., 2004, 2007; Sanmartín et al., 2011). Indeed, loss of RPAP1 in yeast produces global changes in gene expression that resemble those produced by loss of RPB11 (Jeronimo et al., 2004). The RPB3/RPB11 heterodimer provides the interface between Pol II and the Mediator complex (Davis et al., 2002). Importantly, Mediator plays a critical role in establishing cell identity (Allen and Taatjes, 2015; Hnisz et al., 2013; Whyte et al., 2013), and RPB3 is reported to specify
muscle identity (Corbi et al., 2002). Here, we detected a major disruption of the 
Pol II interactome following RPAP1-depletion, and most notably, out of 3,000 
known protein complexes in the Corum database, the complex most heavily 
affected was the Mediator complex. Therefore, our current findings suggest a 
model whereby RPAP1 operates at the interface between Pol II and Mediator to 
direct the transcription of cell identity genes.

RPAP1 is required for Pol II transcription at cell identity genes

Consistent with the pivotal role of RPAP1 in the Mediator/Pol II axis, we 
observed widespread transcriptional changes in RPAP1-depleted MEFs, with 
significantly altered gene expression in 52% of all detectable mRNAs, and 
decreased Pol II loading in 50-60% of all genes. However, we also observed 
that about 40% of genes displayed minimal changes in Pol II abundance 
(Figure 5C), and many highly-expressed mRNAs remained unaffected (Figures 
2C and S5G), arguing against a non-specific defect in Pol II transcription. 
Furthermore, upon RPAP1-knockdown in MEFs, genes regulating 
developmental processes and fibroblastic/mesenchymal identity were the most 
significantly affected according to four criteria: (i) downregulated mRNA 
expression; (ii) greatest overall depletion of Pol II; (iii) increased Ser5P Pol II 
density at promoters; and, (iv) depletion of Pol II within gene bodies relative to 
promoters. These features are consistent with RPAP1 deletion affecting Pol II 
loading on promoters and promoter escape into gene bodies. Remarkably, 
these aspects mirror Mediator’s best known functions (Allen and Taatjes, 2015).
Our proteomic data provide mechanistic explanations for the relative increase in Ser5P Pol II at promoters and for the relative reduction of Pol II from gene bodies. In particular, RPAP1 has conserved interactions with the Ser5P phosphatase RPAP2 in plants and mammals (Egloff et al., 2012a; Jeronimo et al., 2007; Mosley et al., 2009; Muñoz et al., 2017). We observed that RPAP2 phosphatase was depleted from the Pol II interactome upon knock-down of RPAP1 and this may explain the relative accumulation of Ser5P Pol II at promoters. Meanwhile, POLR2M (also known as Gdown1) is a recently discovered protein, often referred as “the 13th subunit”, that tightly binds approximately half of Pol II complexes in cells, forming Pol II(G) (Hu et al., 2006; Jishage et al., 2012). Specifically, Gdown1 is recruited by Mediator and associates with Pol II on Mediator-regulated target genes (Cheng et al., 2012; Hu et al., 2006; Jishage et al., 2012; Li and Price, 2012). It has been reported that Pol II(G) contains RPAP1 (Jishage et al., 2012), and, here, we show that depletion of RPAP1 leads to the loss of Gdown1 from Pol II complexes. Therefore, RPAP1 behaves as a critical ingredient for Mediator-competent Pol II.

Mediator is most abundant in super-enhancers, and super-enhancer target genes are typically the most important for defining cell identity and the most heavily dependent on Mediator to drive their transcription by Pol II (Allen and Taatjes, 2015; Hnisz et al., 2013; Whyte et al., 2013). In agreement, the gene expression of super-enhancer target genes was preferentially decreased following RPAP1-depletion in MEFs, and this pattern of gene expression correlates closely with the first 3 days of iPS reprogramming, constituting a de-
differentiation effect. Consistent with a de-differentiation effect, we observed that the activity of enhancers, measured by their eRNA levels, shifted from the developmental stage of MEFs towards an earlier developmental stage. This is consistent with recent evidence that during cell identity transitions, coordinated changes in enhancer activity lead the re-organization of transcriptional networks (Arner et al., 2015; Factor et al., 2014). Taken together, the data point toward a primary role for RPAP1 in maintaining the expression of identity regulators, through the Mediator/Pol II axis.

Concluding remarks

Collectively, our data points toward a developmental requirement for mammalian RPAP1, both in establishing and maintaining cell identity, through direct regulation of RNA Polymerase II transcription. Mechanistically, we present evidence suggesting a unified model whereby RPAP1 operates by coordinating the communication between Mediator and RNA Pol II, particularly on super-enhancer-driven genes.
EXPERIMENTAL PROCEDURES – STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

KEY RESOURCES TABLE

CONTACT FOR REAGENT AND RESOURCE SHARING

EXPERIMENTAL MODELS AND SUBJECT DETAILS

Mice

Cells and Culture Conditions

METHOD DETAILS

CRISPR/Cas9-based gene editing

Production of Retrovirus and Lentivirus, and infection of recipient cells

Generation of iPS cells from primary MEFs or i4F-MEFs

Growth factors and small molecules to improve iPS reprogramming

Differentiation with retinoic acid

EB Hanging-Drop Differentiation

Wound healing scratch assay

Cytometry

Cell lysis and Western blot

Histopathology and Immunohistochemistry

Immunofluorescence

RNA Pol II interactome analysis and LC/LC Mass Spectrometry

Immunoprecipitate sample preparation for Mass Spectrometry.

LC–MS/MS Analysis
Protein Pol II-interactome Data Collection and Analysis
Protein Pol II-interactome Functional analysis
RNA isolation and Quantitative real-time PCR (qPCR)
RNA-seq transcriptomic analyses
Functional analyses of differential gene expression
Supervised Network Analysis
Comparison of differential gene expression with the iPS roadmap
Conversion of Plant gene expression data to Mammalian homologs
Chromatin Immunoprecipitation (ChIP) and deep-sequencing
Pol II ChIP-seq data analyses
Definition of MEF super-enhancers, their target genes, and eRNA levels

QUANTIFICATION AND STATISTICAL ANALYSIS
DATA AND SOFTWARE AVAILABILITY
Data Resources: Accession numbers.

SUPPLEMENTAL INFORMATION
Supplemental Information includes 5 Supplemental Figures, 6 Supplemental tables, and Supplemental Experimental Procedures in STAR METHODS.
AUTHOR CONTRIBUTIONS

C.J.L. performed most of the experiments, contributed to experimental design, data analysis, and co-wrote the manuscript; R.B., I.C., S.N.P., S.R., and N.I., contributed to experimental work; C.J.L., A.M.D.V., O.G., G.G. and E.A.L. contributed to bioinformatic analyses; V.E.A. and A.D.S performed supervised network analyses; S.O. performed blastocyst injections; E.R., O.F.C., and J.M., provided reagents, discussion, and revisions. M.S. designed and supervised the study, secured funding, analyzed the data, and co-wrote the manuscript. All authors discussed the results and commented on the manuscript.

ACKNOWLEDGMENTS

We are grateful to Elisa Varela for assistance with morula and blastocyst fixation. Work in the laboratory of M.S. is funded by the CNIO and the IRB and by grants from the Spanish Ministry of Economy co-funded by the European Regional Development Fund (SAF project), the European Research Council (ERC Advanced Grant), the Regional Government of Madrid co-funded by the European Social Fund (ReCaRe project), the European Union (RISK-IR project), the Botin Foundation and Banco Santander (Santander Universities Global Division), the Ramon Areces Foundation, and the AXA Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
REFERENCES


RNA Pol II regulator RPAP1 and cell identity


FIGURE LEGENDS

Figure 1. RPAP1 expression, localisation, and requirement for survival in stem cells versus differentiated cells. See also Figure S1.

(A,B) Western blot of RPAP1 expression in a range of pluripotent cell types versus adult tissues (A), or MEFs (B).

(C) Western blot of RPAP1 expression and the ES marker NANOG during a timecourse of ES cell differentiation by LIF-removal and retinoic acid addition.

(D) Immunohistochemical and Immunofluorescence staining for RPAP1 in mouse E3.0 morula (upper panel) or E4.0 blastocyst (lower panel). Scale bars are 20 µm.

(E-G) Immunofluorescence for RPAP1. (E) ES cells undergoing self-renewal versus 24 hrs differentiation by LIF-removal. (F) ES cells exposed to Leptomycin B for 3 hrs. (G) MEFs at day 3 after lentiviral shRNA with non-targeting control (shSCR) or for RPAP1 depletion (shRPAP1).

(H) Immunohistochemical staining for RPAP1 in mouse adult testis. Scale bar 30 µm.

(I) ChIP-qPCR for Pol II or RPAP1 enrichment at promoters of the indicated genes. Analysis was performed on ES cells maintained in self-renewal conditions, or after 24 hrs of differentiation by LIF withdrawal and addition of Retinoic Acid (Diff). See Table S6 for ChIP-qPCR primers.

(J) Quantification of apoptosis by AnnexinV/propidium iodide co-staining and FACS of the indicated cell lines at day 6 after lentiviral control (shSCR) or
RPAP1 depletion (shRPAP1). Mean +/- SEM, n=3 replicates; *p<0.05, **p<0.01.

(K) Proliferation curves (shown by cumulative population doubling) treated by control (shSCR) or lentiviral shRNA against RPAP1 in the indicated cell lines.

Figure 2. RPAP1 is required for the establishment and maintenance of cell identity. See also Figure S2.

(A) Effect of RPAP1 depletion on Embryoid Body (EB) cardiac centre development. EBs were scored daily by microscopy for the appearance of clusters of actively-beating cells indicative of cardiac muscle development. The graph shows the kinetics of this development over several days. Representative pictures of EBs are shown.

(B) qPCR analyses of pluripotency or cardiac development markers at the indicated time points from the EB differentiation assay in (A). Mean +/- SD, n=3 replicates; *p<0.05.

(C) Overview of RNA-seq transcriptome analyses summarizing differential gene expression (FDR q<0.05) in MEFs at day 3 after RPAP1 depletion. Upper panel: Proportional representation pie-chart of significantly differentially expressed genes. Lower panel: dot plot of FPKM values for all genes, showing that many genes of high and low expression level remain unchanged.

(D) Ingenuity Pathway Analysis showing the top 25 most significantly enriched GO terms amongst those genes which were significantly downregulated at
day 3 after RPAP1 depletion in MEFs (FDR q<0.01). Terms highlighted in
red contain “development” or “morphogenesis”. Dotted line indicates the
basal threshold of significance.

(E) Examples of the most significantly up- or downregulated genesets
identified by GSEA analysis in RNA-seq data at day 3 after RPAP1
depletion in MEFs (FDR q<0.01). See also Figures S3J, S3K; and Table
S2.

(F) qPCR validation of RNA-Seq data. Mesenchymal, fibroblast, and
epithelial marker mRNA expression levels were assessed by RNA-Seq
(left) or qPCR (right) at day 3 after RPAP1-knockdown in MEFs. Data
indicates fold change relative to control shSCR, Mean +/-SD, n= 3
independent MEF lines; *p<0.05.

(G) qPCR measurement of mesenchymal and fibroblast marker mRNA levels
during days 1-3 after lentiviral control (shSCR) or RPAP1 depletion
(shRPAP1) in MEFs. Data indicates fold change relative to control, Mean
+/− SD, n=3 independent MEF lines; *p<0.05. See also Figure S2L.

(H) Heatmap summarizing the most significantly up- or downregulated
hallmark genesets identified by GSEA analysis amongst all gene
expression at day 3 after RPAP1 depletion in MEFs (FDR q<0.01; left
column; see Tables S2 and S3), or in ES cells 24hrs after triggering
differentiation (FDR q<0.05; right column see Tables S1 and S2).
Hallmark genesets with FDR q<0.25 are significant. Also highlighted in
the heatmap are borderline genesets (where FDR q = 0.35-0.25).
Wound assay scratch test recovery. Graph shows the percent damaged area remaining at +24hrs. Mean +/- SD, n=3 independent MEF lines with 12 replicates each; *p<0.05.

Figure 3. RPAP1-knockdown favors de-differentiation and reprogramming

See also Figure S3.

(A) Comparison of gene expression at day 3 after RPAP1 depletion in MEFs, versus, a published iPS roadmap gene expression profile (Polo et al., 2012). Panels show GSEA comparison of the published top 500 genes up-or downregulated at day 3 of the iPS roadmap, versus, a ranked list of the gene expression profile in the current study at day 3 after RPAP1 depletion in MEFs (x-axis). See Methods for assessment of the iPS Roadmap data from parental MEFs verses Thy1-negative cells at day 3 of iPS reprogramming. FDR q<0.25 are significant. See also Figures S3A and S3B.

(B) qPCR measurement of selected genes at day 3 after RPAP1 depletion in MEFs. Downregulation of Meox1 and Meox2, and, upregulation of Nup210, were reported to correlate with cell gene expression during the intermediate stages of iPS reprogramming (Hansson et al., 2012; Polo et al., 2012). Mean +/-SD, n= 3 independent MEF lines; *p<0.05; **p<0.01.

(C, D) MEF to iPS reprogramming after RPAP1 depletion. Expression of the OSKM reprogramming factors was initiated at day 2 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). In (C), top panel: kinetics of iPS colony appearance during doxycyclin-induced reprogramming of i4F MEFs
which express the four Yamanaka factors (see: Experimental Procedures). A profile representative of three independent i4F MEF lines is shown (Mean +/-SD, 3 technical replicates). In (C), bottom panel: quantification of iPS colony yield at day 14 of doxycyclin-induced 4F-reprogramming. Mean +/-SD, n= 3 independent MEF lines; **p<0.01. In (D), examples of Alkaline Phosphatase staining to indicate iPS colonies formed at day 12 of i4F-MEF doxycyclin-induced-OSKM iPS reprogramming (top panel), or retroviral delivery of the OSKM factors (bottom panel). FGF2 was added to stimulate reprogramming efficiency.

(E) Summary of outcomes from 32 combinations of OSKMN Yamanaka transcription factors. Sox2-eGFP MEFs at day 2 after control or RPAP1 depletion received the indicated factors by retroviral delivery, followed by culture in standard iPS reprogramming media. Progress of iPS reprogramming was assessed by cell proliferation rate, morphology changes, colony formation, staining for alkaline phosphatase, SSEA1 expression, and Sox2-eGFP levels. Sox2-eGFP-positive cells forming typical iPS colonies were scored as successfully reprogrammed iPS cells. Rapidly proliferating cells which initiated colony formation, and which were positive for alkaline phosphatase and SSEA1, but negative for Sox2-eGFP, were scored as putative intermediate-stages of reprogramming.

(F) Examples of Alkaline Phosphatase staining to indicate formation rates of iPS colonies and putative intermediate cell types at day 14 of MEF reprogramming with the indicated combinations of Yamanaka factors +/-
RPAP1 depletion. Green dot indicates those combinations which produced Sox2-eGFP-positive full reprogrammed iPS colonies.

(G) qPCR measurement of mesenchymal, epithelial and pluripotency marker mRNA expression levels. Data was converted to heatmap format to highlight the intermediate nature of marker expression displayed by the cells which were generated by shRPAP1+Klf4/cMyc.

**Figure 4. RPAP1 regulates the Pol II interactome, not its expression or localization. See also Figure S4.**

(A) Western blots of Pol II Total (RPB1), Ser5P or Ser2P expression in whole cell lysates from a range of cell lines at day 3 after RPAP1 depletion. GAPDH, β-ACTIN, and LAMIN A/C used as internal controls.

(B) Western blots of Pol II Total (RPB1), Ser5P, or Ser2P expression in Nuclear/Cytoplasmic fractions from a range of cell lines at day 3 after RPAP1 depletion. GAPDH and LAMIN A/C used as indicators of fraction separation. N, Nuclear fraction. C, Cytoplasmic fraction.

(C) Immunofluorescence of Pol II Total (RPB1), Ser5P or Ser2P in MEFs at day 3 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). Nuclei stained with DAPI.

(D) Schematic of the 294 specific interactors of Polr2a/RPB1 detected in primary MEFs in this study by Pol II immunoprecipitation and mass spectrometry analysis (see also: Table S4). Interactors were displayed as a network using Cytoscape, and grouped manually by their known physical interactions and general primary function, wherein the thickness
and intensity of the connecting edges indicates the strength of their known interactions in the STRING database. Following RPAP1-depletion, the Pol II-interactors reduced (circled in red) and Pol II-interactors gained (circled in green) are indicated. The Mediator complex is depicted centrally and in full colour based on the data in Figure 4E, below.

(E) Pol II interactors which were decreased following RPAP1 depletion were assigned to all 3,000 known protein complexes in the Corum database (see Methods). On left, complexes are ranked according to the highest percentage of proteins whose interaction were decreased upon RPAP1-depletion from the cells. On right, the total number of subunits per complex is indicated, together with the number of subunits detected in this study and the number of subunits decreased following RPAP1 depletion.

Figure 5. RPAP1 is required for Pol II transcription in MEFs, particularly on developmental and mesenchymal genes. See also Figure S5.

(A,B) ChIP-Seq enrichment data plotted as heatmaps of Pol II total (A), or Pol II Ser5P (B), occupancy around the TSS region +/- 5Kb. Rows are sorted by decreasing Pol II occupancy at the promoter (−100 to +300bp) in the shSCR control. Color-scaled intensities are in units of reads per million mapped reads (rpm; see Methods – ChIP-seq analysis).

(C) Proportional representation of ChIP data, classifying genes according to the changes in abundance of Pol II total (upper panel) or Ser5P (lower panel) at the promoter (−100 to +300bp), following RPAP1 depletion in MEFs for 3 days.
(D) Schematics of Pol II total and Ser5P abundance on selected genes, showing examples of Pol II depletion (S100a4, Snai1, Snai2) or minimal effects (Asap3, Tulp3).

(E) Table summarizing the most significantly up- or downregulated GO-term genesets identified by GSEA among the genes with >2xfold decrease in Pol II following RPAP1 depletion in MEFs, (see also: Tables S2 and S5). Genesets with FDR q-value <0.25 are significant.

(F) Summary heatmap displaying the overlay of significant GSEA hallmark genesets across 4 experiments (Columns 1 and 2: see also Figure 2H). Column 1: GSEA on the ranked list of differential mRNA expression in MEFs at day 3 +/- RPAP1-depletion. Column 2: GSEA on the ranked list of differential mRNA expression in ES cells at +24hrs after inducing differentiation, +/- RPAP1-depletion. Column3: GSEA on the ranked list of differential Pol II abundance at all promoters in MEFs at day 3 +/- RPAP1 depletion. Column4: GSEA on the ranked list of differential gene expression in Arabidopsis thaliana plant tissues +/- RPAP1-mutation (Sanmartín et al., 2011), following conversion to the nearest mouse homolog based on protein sequence conservation (see: Methods – Conversion of Plant to Mouse homologs, and Table S2).

(G) GSEA to assess mRNA expression levels of MEF super-enhancer target genes (n= 661, defined by GREAT analysis as described, see Methods) within the transcriptome of primary MEFs at day3 after RPAP1 knockdown. Compare with Housekeeper gene expression in Figure S5G.
(H) Plots show the average eRNA levels within two groups of MEF super-enhancers regions: those which were increased (n=63 enhancers, top panel) or decreased (n=64 enhancers, bottom panel) in MEFs at day 3 after RPAP1 knockdown.

(I) GREAT analysis was used to identify a set of target genes for each of the two super-enhancer groups identified in (H) (see: Methods). Next, the developmental stages significantly associated with the enhancer-target genes in each group were identified and plotted according to their group, that is whether the associated enhancer had increased or decreased eRNA abundance, in MEFs at day 3 after RPAP1 knockdown. See also, Figures S5H to S5I for tables of functions and developmental stages enriched within the target genes of enhancers identified in (H) above.

(J,K) Model for RPAP1 function in the mechanism for triggering development.

(J) In self-renewing ES cells, RPAP1 is expressed but largely inhibited from interacting with Pol II, being continuously imported and exported from the nucleus such that it is predominantly cytoplasmic in its overall abundance. Upon differentiation, nuclear accumulation of RPAP1 permits increased transcriptional regulation. Hence, depletion of RPAP1 in self-renewing ES cells does not affect proliferation or pluripotency, however ES cell differentiation is aberrant. Depletion of RPAP1 in MEFs, where it is nuclear, results in Pol II transcriptional dysfunction preferentially on cell identity and developmental regulator genes, which is associated with de-differentiation and enhanced susceptibility for reprogramming towards pluripotency. (K): Taken together, our data suggest a model where
RPAP1 exists in complex with RNA Pol II, and plays an essential role in the Mediator-Pol II regulatory axis. Thus, loss of RPAP1 triggers a decrease in the association between Mediator and Pol II (including the key regulators Gdown1 (G) and the Ser5P phosphatase RPAP2) preferentially affecting the ability of enhancers to activate Mediator target genes, which are known to include the key markers and regulators of cell identity. In somatic cells such as MEFs, this leads to de-differentiation, as expression of fibroblastic, mesenchymal and developmental markers is erased.
Figure 1

A. Pluripotent vs. Adult Tissues

B. Pluripotent vs. MEFs

C. Self-Renewal vs. Differentiation

D. RPAP1 vs. RPAP1/DAPI

E. ES vs. Self-Renewal vs. Differentiation

F. ES vs. Self-Renewal vs. Leptomycin B

G. MEFs

H. Testis

J. ESC vs. 293T vs. COS7 vs. MEF

K. Cumulative PD

Figure 1

**Figure 1**

A. Pluripotent vs. Adult Tissues

B. Pluripotent vs. MEFs

C. Self-Renewal vs. Differentiation

D. RPAP1 vs. RPAP1/DAPI

E. ES vs. Self-Renewal vs. Differentiation

F. ES vs. Self-Renewal vs. Leptomycin B

G. MEFs

H. Testis

J. ESC vs. 293T vs. COS7 vs. MEF

K. Cumulative PD
Figure 2

A) Cardiac Development

B) Nanog, Cardiac Alpha-Actinin, and Atrial Natriuretic Peptide

C) Differentially Expressed Genes (FDR<0.05)

D) -Log(p-value)

E) Epithelial-Mesenchymal Transition and Skeletal Development

F) DNA Replication

G) RNA-Seq and qPCR

H) FDR q-value

I) Wound Assay

Legend:
- shSCR
- shRPAP1

Key Genes:
- Rpap1
- Snai1
- Zeb1
- Fbn1
- BMP4
- Grem1
- Snai2
- Cdkn1a
- Thy1

Significant: *
Not significant: **
Figure 3

A. Geneset: iPS Roadmap Day 3, Top500

B. Meox1

C. Meox2

D. Nup210

E. Ranked list: MEF+/- RPAP1 depletion

F. SSEA1+ AP+ Sox2-eGFP-

Putative Intermediate colonies

Sox2 eGFP+ iPS colonies

Death

G. Target/GAPDHx10

-5

0

5

10

15

20

+ FGFR2

GFP

AP+ colonies/well

shSCR

shRPAP1

AP+ colonies/well

shSCR

shRPAP1

iPS colonies/well

shSCR

shRPAP1

Dox-OsKm

Control + FGF2

Retroviral OsKm

GFP

A: Oct4

M: Myc

K: Klf4

S: Sox2

N: Nanog

O: Nanog

No proliferation

Proliferation

Death

FDR q-value <0.001

Roadmap Day3: Top 500 Genes UP

500 genes FDR q-value <0.001

Roadmap Day3: Top 500 Genes DOWN

500 genes FDR q-value <0.001

100

50

0

150

Days

12

17

10

5

0
### A

<table>
<thead>
<tr>
<th></th>
<th>293T</th>
<th>MEF</th>
<th>HCT</th>
<th>SCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>shRPAP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPAP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pol II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser5P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser2P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAPDH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b-ACTIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMIN A/C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMIN A/C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### B

<table>
<thead>
<tr>
<th></th>
<th>293T</th>
<th>MEF</th>
<th>HCT116</th>
<th>SCC42B</th>
</tr>
</thead>
<tbody>
<tr>
<td>shRPAP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pol II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser5P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser2P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAPDH</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMIN A/C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### C

- **Total Pol II**
- **Pol II Ser5P**
- **Pol II Ser2P**

### E

- All protein sub-units in each complex (Corum Database)
- Sub-units identified here in RNA Pol II interactome
- Sub-units decreased in RNA Pol II interactome after RPAP1-knockdown

**Figure 4**

- **E**: % of proteins in complex vs Total number of proteins in complex
- **D**: RNA Pol II - associated with various functional categories
- **C**: Western blots showing protein expression
- **A**: Summary of Western blots for different cell lines

**Note**: The images depict the results of Western blot analyses for different cell lines (293T, MEF, HCT, SCC) under control (C) and knockdown (N) conditions for shRPAP1 and shSCR. The blots show protein expression levels for Total Pol II, Pol II Ser5P, and Pol II Ser2P. The diagram (D) illustrates the interaction network among various proteins, including transcription factors, kinases, ligases, RNA Pol II, and other functional categories.
Supplemental Information

- Supplemental Figures S1-S5
- Supplemental Figure Legends
- Supplemental Experimental Procedures – STAR METHODS
- Supplemental References
**Figure S1**

**A**

P19 EC cells

<table>
<thead>
<tr>
<th>Control</th>
<th>Differentiation + Retinoic Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**B**

E13.5 Liver

Ovary

**C**

shSCR

#1 #2 #3 #4 #5 #1-5

shRPAP1

**D**

shSCR

shRPAP1

**E**

SA-ji-gal+ MEF cells

**F**

Cumulative PD

Day: 0 3 6 9 12

293T

shSCR

shRPAP1#2

shRPAP1#4

shRPAP1#5

**G**

HCT116

SCC42B

HEP

H26

<table>
<thead>
<tr>
<th>Day: 0 3 6 9 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>30</td>
</tr>
</tbody>
</table>

**H**

Nanog-GFP

Intensity/Cell

Actin

<table>
<thead>
<tr>
<th>Control</th>
<th>No-CG</th>
<th>sU6L</th>
<th>sG6L</th>
<th>Sarc/4L</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**I**

CRISPR gRNAs

<table>
<thead>
<tr>
<th>5' ATG</th>
<th>Method</th>
<th>Cells</th>
<th>Expression</th>
<th>Gene</th>
<th>RPAP1 KO clones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>E14</td>
<td>Transient</td>
<td>E14</td>
<td>0/126</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R1</td>
<td>Constitutive</td>
<td>R1</td>
<td>5/6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G4</td>
<td>Constitutive</td>
<td>G4</td>
<td>5/6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HAP1</td>
<td>Constitutive</td>
<td>HAP1</td>
<td>4/6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EESC9</td>
<td>Inducible</td>
<td>EESC9</td>
<td>6/6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CS6919</td>
<td>Transient</td>
<td>CS6919</td>
<td>7/6</td>
</tr>
</tbody>
</table>

**J**

CRISPR gRNAs A-D

Pool

<table>
<thead>
<tr>
<th>Pool</th>
<th>CRISPR gRNAs A-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

**K**

CRISPR gRNAs A-D

Pool

<table>
<thead>
<tr>
<th>Pool</th>
<th>CRISPR gRNAs A-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

**L**

LacZ+ • LacZ-
Supplemental Figure S1. Related to Figure 1.

RPAP1 expression, localisation, and requirement for survival in stem cells versus differentiated cells.

(A) Western blot of RPAP1 expression during a timecourse of P19EC cell differentiation by Retinoic Acid addition.

(B) Immunohistochemical staining for RPAP1 in mouse E13.5 liver, adult ovary, and adult skin. Scale bars represent 30 µm.

(C) Western blot of RPAP1 expression in mouse ES cells at day 6 following five separate lentiviral shRNA against RPAP1 (#1-#5).

(D) Photographs of the indicated cell lines at days 6-9 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1).

(E) Quantification of senescence-associated β-galactosidase staining in MEFs at day 9 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). Mean+/−SD, n=3 replicates; **p<0.01.

(F) Proliferation curves (shown by cumulative population doubling) following lentiviral control (shSCR) or RPAP1 depletion (shRPAP1) in 293T cells using 3 different shRNAs.

(G) Proliferation curves (shown by cumulative population doubling) following lentiviral control (shSCR) or RPAP1 depletion (shRPAP1) in the indicated cell lines.

(H) FACS analysis of a Nanog-GFP reporter ES cell line (TNGA) cultured in three different media cocktails, at day 6 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). Wild type non-GFP ES cells were used as negative control.
Schematic of the 26-Exon mouse RPAP1 gene (Gene ID: 68925; NM_177294.5). Indicated are: the open reading frame ATG start (green, exon3) and TGA stop (red, exon26); the location of the gene trap in CSI619 ES cells (grey, intron8); and the location of CRISPR guide RNAs used in this study (blue, exons 3-7; see also: Table S6). Table (below) summarizes the effect of multiple CRISPR approaches on the expression of RPAP1. While RPAP1 protein levels were decreased in cell pools and in clonal lines, no RPAP1-null clones could be derived.

Example of Western blot analyses of whole population from haploid HAP1 cells following CRISPR against RPAP1 using lentiviral constitutive CRISPR/Cas9 expression.

Examples of Western blot analyses of ES clones following CRISPR against RPAP1 using CRISPR/Cas9 expression systems which were transient in wild-type E14 ES cells (pX330; above), or constitutive in G4 ES cells (lentiviral; below).

CSI619 (RPAP1+/Trap) reporter ES cells stained for LacZ 3 days after CRISPR against RPAP1. Inset shows examples of ES colonies expanded from single cells and stained for LacZ (blue) expressed from the RPAP1 β-geo reporter allele. Arrow: a white/non-stained colony, indicating that CRISPR has successfully mutated at least one of the RPAP1 alleles, by knocking out the RPAP1 β-geo reporter allele. Bar chart below, shows the percentage of non-staining colonies at Days +3 or +10 after CRISPR in the whole population. Significantly fewer non-staining colonies are observed at Day+10 (after a passage). This suggests that where CRISPR
is active, and also knocks out the WT RPAP1 allele, the ES cells display a growth or survival phenotype within a few days.
Figure S2

GSEA Hallmark Pathways

FDR

<table>
<thead>
<tr>
<th>Pathway Description</th>
<th>FDR</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interferon gamma response</td>
<td>0.049</td>
<td>0.117</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>0.049</td>
<td>0.117</td>
</tr>
<tr>
<td>Colonic epithelial cell differentiation</td>
<td>0.049</td>
<td>0.117</td>
</tr>
</tbody>
</table>

GSEA Gene Ontology Terms

FDR

<table>
<thead>
<tr>
<th>Term Description</th>
<th>FDR</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid metabolic process</td>
<td>0.011</td>
<td>0.049</td>
</tr>
<tr>
<td>Nucleotide metabolic process</td>
<td>0.011</td>
<td>0.049</td>
</tr>
<tr>
<td>Tyrosine kinase activity</td>
<td>0.011</td>
<td>0.049</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>0.011</td>
<td>0.049</td>
</tr>
</tbody>
</table>

Results from RNA-seq analysis of MEFs at day 3, showing significant enrichment of pathways related to interferon-gamma response and apoptosis. FDR values are below 0.001 for several pathways, indicating strong statistical significance.
Supplemental Figure S2. Related to Figure 2.

RPAP1 is required for the establishment and maintenance of cell identity.

(A) Following shRNA-knockdown of RPAP1, ES cells were differentiated for 24 hours by LIF-removal, then fixed and scored per colony for morphology and Alkaline Phosphatase staining intensity. Photographs show examples of the delay in colony morphology changes and delay in attenuation of AP-staining intensity associated with RPAP1-depletion at +24 hours after LIF-removal. Mean +/- SEM, n=3 replicates; *p<0.05.

(B) FACS analyses for apoptosis levels by AnnexinV/Propidium Iodide double-staining in ES cells following induction of differentiation by LIF-removal for 24 hrs and then addition of retinoic-acid for 48hrs.

(C) qPCR analyses of pluripotency or cardiac development markers at the indicated time points from the EB differentiation assay in Figure 2A. Mean +/- SD, n=3 replicates; *p<0.05. Panel on right: Western blot confirming RPAP1 knockdown in ES cells during self renewal.

(D) On left, table summarizing the generation of chimeric mice using CSI619, RPAP1(+/Trap) ES cells (1 wild-type and 1 null allele). A low percentage of chimeric pups survived to birth (21/254 micro-injected embryos), of which 10/21 pups displayed moderate chimerism based on coat colour 20-60% Agouti coat colour). On right, table summarizing the offspring generated by mating chimeric mice with eachother, or with wild-type mice, to look for germline transmission. Of 156 pups born from these matings, 7 pups had Agouti coat colour, indicating that the parental RPAP1(+/Trap) ES cells were viable. However, no pups carried the RPAP1(+/Trap)
geneotype, suggesting that a single RPAP1 allele was insufficient for germline transmission.

**(E)** Overview of differential gene expression in RNA-Seq transcriptome analysis of ES cells following RPAP1-knockdown then differentiation for 24 hours, as above, in Figure S2A. Proportional representation pie-chart indicates the proportion of mRNAs significantly up- or down-regulated with FDR q<0.05. See also Table S1.

**(F,G)** Table summarizing the most significantly up- or downregulated Hallmark genesets (F), or GO-term genesets (G), identified by GSEA analysis in RPAP1-depleted ES cells after 24 of differentiation, as above, in Figure S2A (FDR q<0.05; see also Tables S1 and S2). Genesets with FDR q<0.25 are significant.

**(H)** GSEA Leading Edge analysis of the most prevalent genes among those GO terms database genesets which were significantly downregulated in RPAP1-depleted ES cells after 24 hrs of differentiation, as above, in Figure S2A (FDR q<0.05; see also Tables S1 and S2).

**(I)** Normalized RNA-seq expression levels of mesenchymal, fibroblastic and development markers in RPAP1-depleted ES cells after 24 of differentiation, as above, in Figure S2A. Data based on Mean FPKM values, n=3 replicates; * FDR q-value <0.05. See also Table S1.

**(J)** Table summarizing the most significantly up- or down-regulated hallmark genesets identified by GSEA analysis of RNA-seq data at day 3 after RPAP1 depletion in MEFs (FDR q<0.05; see also Tables S2 and S4). Hallmark genesets with FDR q<0.25 are significant.
Table summarizing GSEA Leading Edge analysis of the most prevalent genes among the genesets which were significantly downregulated at day 3 after RPAP1 depletion in MEFs (FDR q<0.05; see also Table S6) in a comparison versus the GSEA C5 GO terms database.

qPCR analyses of fibroblastic, mesenchymal and development markers at the indicated early time points after RPAP1 shRNA depletion in MEFs. Mean+/−SD, n=3 replicates. Raw data from Figure 2G is displayed here relative to the housekeeper internal control Gapdh (whereas, in Figure 2G, the data is shown as fold-change, normalized to the shSCR non-targeting control).
**A** IPS Roadmap Significant Genes (FDR<0.05) TOP 500 genes vs MEFs+shRPAP1 Day3

- Geneset = IPS Roadmap Genes Up Day 3 (Top500)
  - 500 genes FDR q-value <0.001
- Geneset = IPS Roadmap Genes DOWN Day 3 Top500
  - 500 genes FDR q-value <0.001

Day3

- Geneset = IPS Roadmap Genes Up Day 9 Top500
  - 500 genes FDR q-value <0.001
- Geneset = IPS Roadmap Genes DOWN Day 9 Top500
  - 500 genes FDR q-value <0.001

Day9

**B** MEFs+shRPAP1 Day3 Significant Genes (FDR<0.01) vs IPS Roadmap

- Geneset = Genes Up Day 3 shRPAP1 (FDR<0.01)
  - 2535 genes FDR q-value <0.001
- Geneset = Genes DOWN Day 3 shRPAP1 (FDR<0.01)
  - 1288 genes FDR q-value <0.001

**C**

Control = KSR/LIF standard IPS media
-plus 12 growth-factor/inhibitor combinations

- Growth Promoters +Inhibitors:

  - Growth Factor Combinations
  - FGF2 @10ng/mL
  - EGF @20ng/mL
  - Alk4 (ALK4/5/7, SB431542) @2/4uM
  - DLPC @1-5uM
  - SCF @10ng/mL
  - Forskolin @5uM
  - 5-AzaC @0.5-2uM
  - VPA @0.5mM
  - TSA @20nM
  - BIX @100uM
  - Kenpaullone @5uM
  - Flavopiridol @50nM
  - FGF2/KSR
  - FGF2/LIF
  - FGF2/Alk4
  - FGF2/DLPC
  - FGF2/SCF
  - FGF2/Forskolin
  - FGF2/VPA
  - FGF2/TSA
  - FGF2/Kenpaullone
  - FGF2/Flavopiridol

**D**

![Cell Count vs SSEA1 intensity](image)

- MEF shRP+KM ES
- SSEA1 intensity

**E**

![Graph](image)

- Sox2-GFP+/iPS Colonies well
- OKM+

**Figure S3**
Supplemental Figure S3. Related to Figure 3.

**RPAP1-knockdown favors de-differentiation and reprogramming.**

**(A)** GSEA comparison of gene expression at day 3 after RPAP1 depletion in MEFs, versus, a published iPS roadmap gene expression profile (Polo et al., 2012). See Methods for assessment of the iPS Roadmap data from control MEFs versus Thy1-negative cells at day 3, or day 9, of reprogramming. The data here can be compared with Figure 3A. GSEA comparison of the published top 500 genes (on left), or Top 100 genes (on right (here, the reduced geneset size was used), up- or down-regulated at day 3, or day 9, of the iPS roadmap, versus, a ranked list of the gene expression profile at day 3 after RPAP1 depletion in MEFs in the current study (x-axis). FDR q<0.25 are significant.

**(B)** GSEA comparison of the significantly up- or down-regulated genes (FDR q<0.01) at day 3 after RPAP1 depletion in MEFs, versus, a ranked list of the published gene expression profile of the iPS roadmap at day 3, or day 9 (x-axis) as indicated. FDR q<0.25 are significant.

**(C)** Left: a list of growth factors and small molecule inhibitors, and the concentrations used, to test culture media supplementation in relation to the iPS reprogramming screen in Figure 3E. Right: the list of 12 combinations of the media supplements that were tested.

**(D)** Cells resembling putative reprogramming intermediates, which were generated by shRPAP1+Klf4/cMyc over-expression (see Figure 3E), were expanded to passage 4, then independent clones were analyzed for
surface expression of SSEA1 by FACS. MEFS and ES cells were included as negative and positive controls for SSEA1 staining respectively.

(E) Sox2-eGFP-positive iPS colonies were counted per well at day14 following retroviral expression of the indicated combinations of Yamanaka factors, RPAP1 depletion, and/or TGFβ-signaling inhibition (see: Experimental Procedures). Mean +/-SD, 3 replicates; ***p<0.001, versus the control (lane 1).
**Figure S4**

A: Western blots showing the expression of shRPAP1 and shSCR in MEF#1 and MEF#2. The blots are for RPAP1, Total Pol II, Ser5P, Ser2P, and β-ACTIN.

B: Western blots for shRPAP1 and shSCR in HEP and H226. The blots are for RPAP1, Total Pol II, Ser5P, Ser2P, GAPDH, and LAMIN A/C.

C: Western blots for shRPAP1 and shSCR in HEP and H226. The blots are for Total Pol II, Ser5P, LAMIN A/C, and GAPDH.

D: Immunofluorescence images showing the localization of Total Pol II, Pol II Ser5P, and Pol II Ser2P in SCC42B, 293T, and HCT116 cells with shRPAP1 and shSCR.

E: Gene ontology analysis showing the biological processes regulated by shRPAP1 and shSCR in HEP and H226 cells. The processes include splicing, transcription, RNA processing, and metabolic processes. The log(p-value) is plotted against -log10(p-value).
Supplemental Figure S4. Related to Figure 4.

RPAP1 regulates the Pol II interactome, not its expression or localization.

(A,B) Western blots of RPAP1, Pol II Total (RPB1), Ser5P, or Ser2P expression in whole cell lysates from two independent MEF lines at day 3 (A), or from HEP and H226 cell lines (B), at day 3 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). GAPDH, β-ACTIN, and LAMIN A/C used as internal loading controls.

(C) Western blots of Pol II Total (RPB1), Ser5P, or Ser2P expression in Nuclear/Cytoplasmic fractions from HEP and H226 cell lines at day 3 after RPAP1 depletion. GAPDH and LAMIN A/C used as indicators of fraction separation. N, Nuclear fraction. C, Cytoplasmic fraction.

(D) Immunofluorescence of Pol II Total (RPB1), Ser5P or Ser2P in a range of human cell lines at day 3 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1). Nuclei stained with DAPI.

(E) Gene ontology analysis for the enrichment of biological processes among the Pol II interactors lost following RPAP1-depletion, with p-value corrected for multiple testing (Bonferroni).
Figure S5

A. Number of Genesets

B. Pausing Index

C. All genes

D. Change in Pausing Index

E. Change in Ser5P Density

F. Change in Ser5P Density

G. Housekeeper Gene Expression

H. Expression patterns associated with the Target genes of MEF SEs that have increased eRNAs, n=64.

I. Expression patterns associated with the Target genes of MEF SEs that have decreased eRNAs, n=63.

J. Expression patterns associated with the Target genes of MEF SEs that have decreased eRNAs, n=63.
Supplemental Figure S5. Related to Figure 5.

RPAP1 is required for Pol II transcription in MEFs, particularly on developmental and mesenchymal genes.

(A) GSEA Leading Edge analysis. The genesets which were significantly depleted in Pol II abundance were identified in MEFs at day 3 after RPAP1 depletion (see Tables S2 and S6; and Figure 5E). The table lists the most prevalent genes among the GO term genesets.

(B) Definition and analysis of Pol II loading ratio on Promoter-Body (or Pausing Index, PI). Schematics outline the parameters used to define the whole gene, promoter, gene body, and Pausing Index (PI) in this study, (see also Methods). An example of preferential depletion of Pol II from the gene body is shown, in lower panel.

(C,D) Plots showing the PI ratio for all genes, or all regulators of the Epithelial-Mesenchymal Transition (defined by the GSEA Hallmark geneset #M5930, MySigDB, Broad Institute). In (D), data are Mean +/- SEM of “n” genes as indicated; **p<0.01. In (C) and (D), at day 3 after RPAP1 depletion in MEFs, the Mean PI significantly decreases for many genes (ΔPI <1.0), however, the Mean PI increases for EMT-regulatory genes (ΔPI >1.0). This is consistent with preferential depletion of Pol II from the gene body, as depicted in lower panel of (B), above.

(E) Ratio of shRPAP1/shSCR for the change in PI for each gene (ΔPI), at day 3 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1) in MEFs. Arrow highlights the region containing genes with increased PI at their promoters. Table shows the top three GSEA results which identify that
genesets and genes with increased pausing index (region highlighted in plot) are enriched for MEF cell identity and developmental regulators (FDR q<0.25 is significant). Below: examples of GSEA plots for the most significantly enriched genesets with increased PI (see also Table S5 for PI calculations per gene, and Table S2 for full GSEA results).

(F) Graph of the change in Ser5P density comparing shRPAP1/shSCR, at the promoter (red), or in the gene body, (black). Data from day 3 after lentiviral control (shSCR) or RPAP1 depletion (shRPAP1) in MEFs. Arrow highlights the region containing genes with increased Pol II Ser5P density at their promoters. GSEA analyses were performed on the entire ranked lists for promoters (red line) or the gene bodies (black line), however, significant enrichment of genesets was only observed for those genes with increased Pol II Ser5P density at their promoters (Red plot line, region as indicated by arrow). Table below shows the top three GSEA results which identify that genes with increased promoter Ser5P density (region highlighted in red plot line by arrow, above) are enriched for cell identity and developmental regulators (FDR q<0.25 is significant). Below: examples of GSEA plots for the most significantly enriched genesets with increased Ser5P density at their promoter (see also Table S5 for Ser5P density calculations per gene, and Table S2 for full GSEA results).

(G) GSEA to assess mRNA expression levels of housekeeper genes, (as defined: see Methods) in primary MEFs at day3 after RPAP1 knockdown. No significant change in housekeeper geneset expression was observed.
(H) Developmental stages significantly associated with the super-enhancer target genes ($P < 10^{-4}$) where the enhancers display increased eRNA levels in MEFs at day 3 after RPAP1 knockdown.

(I,J) Developmental stages (I) and GO Biological Processes (J) significantly associated with the super-enhancer target genes ($P < 10^{-4}$) where the enhancers display decreased eRNA levels in MEFs at day 3 after RPAP1 knockdown.
SUPPLEMENTAL EXPERIMENTAL PROCEDURES –STAR METHODS

KEY RESOURCES TABLE

CRISPR gRNAs, Primers, Antibodies and shRNAs used in this study. See Table S6.

CONTACT FOR REAGENT AND RESOURCE SHARING

Please contact Manuel Serrano. Manuel.serrano@irbbarcelona.org

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

Animal experimentation at the CNIO, Madrid, was performed according to protocols approved by the CNIO-ISClII Ethics Committee for Research and Animal Welfare (CEIyBA).

Cells and Culture conditions

Primary mouse embryo fibroblasts (wild-type, MEFs, passage 2) were obtained at E13.5 from pure inbred C57BL6 background mice, as described previously (Palmero et al. 2001). Mouse P19EC cells and HEP cells (immortalized mouse hepatocytes), monkey COS7 cells, and the human cell lines 293T, HCT116, SCC42B and H226, were from ATCC and were maintained in DMEM medium with 10% FBS (Gibco) with antibiotics (penicillin/streptomycin 100 U/ml). The mouse ES cells E14Tg2a.4 (wild-type parental) and CSI619 RPAP1 (+/Trap) mouse ES cells containing a pGT0Lxf genetrap in Intron8 were from BayGenomics/MMRRC genetrap resource, University of California. Nanog-GFP knockin mouse ES cells (TNGA) were previously described (Chambers et al., 2007) and were shared by the laboratory of Austin Smith. The mouse ES cells R1, G4, doxy-inducible ESCas9 as described (Ruiz et al., 2016). HAP1 cells (a kind gift from T Brummelkamp) were grown in IMDM (Invitrogen) and 15%FBS. The Sox2-eGFP MEFs (Sox2-Promoter/GFP transgenic) were as described (D’Amour and Gage, 2003). MEFs and P19EC cells were cultured in standard DMEM medium with 10% FBS (Gibco) with antibiotics (penicillin/streptomycin 100 U/ml). Mouse ES cells and iPSCs, were routinely cultured on gelatin-coated plates in either “Serum/LIF” (15% FBS), or Knockout Serum Replacement (KSR, Invitrogen) “KSR/LIF” (15% KSR), in DMEM (high glucose) basal media, with LIF (1000U/nits/mL), non-essential amino acids, glutamax and beta-mercaptoethanol plus antibiotics. Where used, the “2i” drug cocktail comprised 1 uM Mek-inhibitor (PD0325901, Axon Medchem, #1408) plus 3 uM GSK3b-inhibitor (CHIR 99021, Axon Medchem #1386) as described (Ying et al., 2008). Reprogrammed iPSC cells were initially derived and expanded on mitomycin-C inactivated feeder cells on gelatin-coated plates, before transfer to gelatin-only. Cultures were routinely tested for mycoplasma and were always negative. C57BL/6 ES cells were derived at the Transgenic Mice Unit of the Spanish National Cancer Research Center (commonly abbreviated as CNIO, from the name in Spanish: Centro Nacional de Investigaciones Oncológicas) from E4.5 C57BL6 blastocysts, or mixed background C57BL6/129 blastocysts. ES cell self-renewal and pluripotency was scored, by cytometry (Nanog-GFP heterogeneity and overall intensity), by immunofluorescence (see below), by colony morphology (see Figures SID and S2A), by alkaline phosphatase staining of fixed cells (Promega #S3771), and by qPCR for stemness markers Nanog, Oct4 and Sox2 (See: Figures 2B and S2C), in addition to their differentiation capacity in retinoic acid or embryoid body cardiac centre development (see below). To inhibit CRM1-dependent nuclear export, cells were treated for 3hrs with 10 nM Leptomycin B (Sigma #L2913). For proliferation curves, cells were counted and serially passaged every 3 days to monitor the cumulative doubling rate. Senescence-associated β-gal staining was performed as described (Munoz et al., 2013). Staining for LacZ expression in the CSI619 ES cells RPAP1(+/Trap), where the genetrap contains a β-geo reporter, was performed as described (Munoz et al., 2013).
METHOD DETAILS

CRISPR/Cas9-based gene editing
To target human or mouse RPAP1 sequences, we used the MIT CRISPR design tool (http://tools.genome-engineering.org/) to design the sgRNAs as described (Ran et al., 2013). Six mouse sgRNAs were used targeting mouse/human RPAP1 Exons 4-7 (see Figure S11) or 4 sgRNAs targeting human RPAP1 Exons 2-4, either individually to generate indels, or in combinations to generate deleted regions (see: Table S6, Sheet#4, for sgRNA sequences, plasmid details and gRNA combinations used). RPAP1-knockout was assessed by Western blot of entire cellular pools, or derivation and expansion of individual clones.

Briefly, three CRISPR strategies were pursued. Transient CRISPR/Cas9 expression was by electroporation of mouse ES cells (Neon Transfection System; 1200V, 20 msec, 2 pulses) using the pX330 plasmid (Addgene #42230). Constitutive CRISPR/Cas9 expression was by pLentiCRISPRv2 (Addgene: #52961) as described (Ruiz et al., 2010). For the human HAP1 cell line, human specific CRISPR-sgRNAs oligos (Table S6, Sheet#4) were cloned into the pLenti-CRISPRV2 (Addgene plasmid #52961). For doxycyclin-inducible CRISPR/Cas9: CRISPR-sgRNAs oligos cloned into the pKLV-U6-grRNA (BbsI)-PGKpuro2ABFP (Addgene #50946) to generate doxy-inducible ESCas9 cells as described (Ruiz et al., 2016). Individual lentiviral vectors pKLV-U6gRNA-PGTKpuro2ABFP (Addgene #50946) or pLentiCRISPR v2 (Addgene plasmid #52961) were co-transfected with 3rd generation packaging vectors in 293T cells using Lipofectamine 2000 (Invitrogen) in order to generate viral supernatants as described (Ruiz et al., 2010). A total of 10^5 ES cells were infected in suspension with 500 µl of viral supernatant for 1 hour at 37°C and plated on a layer of fresh feeder cells. Two days after infection, G4 and R1 ES cells were selected with Puromycin 1µg/ml and maintained for a week in culture to allow efficient gene editing. For the doxycycline-inducible ESCas9 cell line (Ruiz et al, 2016), two days after infection, cells were split into media with or without 1 µg/ml doxycycline and maintained for an additional week in culture to allow efficient gene editing. In the case of Hap1 cells, spinfection was used to infect as follows: a total of 10^5 HAP1 cells in one 6-well plate were incubated with 1.5 mls of viral supernatant and centrifuged at 1850rpm for 1 hour. Two days after infection, cells were selected with Puromycin 1µg/ml and maintained for a week in culture to allow efficient gene editing.

Production of Retrovirus and Lentivirus, and infection of recipient cells
Briefly, retroviral and lentiviral supernatants were produced in HEK-293T cells (5x10^6 cells per 100-mm-diameter dish). Vector transfections were performed using Fugene-6 transfection reagent (Roche) according to the manufacturer’s protocol. Two days later, viral supernatants (10 ml) were collected serially during the subsequent 48 hours, at 12-hour intervals, each time adding fresh medium to the cells (10 ml). The recipient cells were seeded the previous day (1.5x10^6 cells per well in a 6-well plate) and each well received 1.0 ml of the corresponding retroviral and/or lentiviral supernatants as indicated in each Figure. This procedure was repeated every 12 hours for 2 days (a total of 4 additions).

For lentiviral shRNA production, per dish, 293T cells were transfected with 3 plasmids: (i) the ecotropic lentiviral envelope packaging plasmid pMD2.G (0.3 µg; Addgene, plasmid #12259; containing the VsVg gene); (ii) the lentiviral packaging plasmid pCMV-dR8.91 (3.0 µg); (from: Harvard Medical School, plasmid #516); (iii) plus either one of the following 6 lentiviral shRNA constructs (3.0 µg) expressing mouse shRNAs against RPAP1 (shRPAP1#1-5, respectively), or the corresponding non-targeting control (Scramble, shSCR) vector. After lentiviral infection was completed, lentiviral RPAP1-knockdown shRNA recipient cells were selected with puromycin (1ug/mL). A panel of five lentiviral shRNA against RPAP1 were from Open Biosystems (#RMM4534-NM_177294; TRC Mission Library) with a pLKO.1 lentiviral backbone. From these 5 clones we identified that the best knockdown of RPAP1 expression was achieved using clone TRCN0000173186, hereafter “shRPAP1#5”. See shRNA clone details in Table S6, sheet#3.

For retrovirus, per dish, 293T cells were transfected with the ecotropic packaging plasmid pCL-Eco (4 µg) together with one of the following retroviral constructs (4 µg): pMXs-Oct4, pMXs-Sox2, pMXs-Klf4, pMXs-cMyc, or pMXs-Nanog (obtained from Addgene and previously described (Takahashi and Yamanaka, 2006) -the backbone is pMXs plasmid in all cases and the expression of the coding sequences of the reprogramming factors are driven by the MMLV LTR promoter.

Generation of iPS cells from primary MEFs or i4F-MEFs
For retroviral-mediated iPS reprogramming of primary (passage 2-4) mouse embryo fibroblasts was performed by a previous protocol (Li et al., 2009a). Briefly, after infection of primary MEFs with retrovirus expressing the four Yamanaka transcription factors (OSKM), as outlined above, MEF media
was replaced by KSR/LIF medium (see above). Cultures were maintained in the absence of drug selection with medium changes every 48 hrs (Li et al., 2009a).

For reprogramming of the secondary-system doxycyclin-inducible 4-Factor (i4F) MEFs which inducibly-express the four Yamanaka factors Oct4, Sox2, Klf4, and cMyc (OSKM) was performed as previously described (Abad et al., 2013). Briefly, i4F-MEFs were treated with doxycyclin (1µg/mL) continuously to induce expression of the OSKM transcription factors in the presence of the KSR/LIF iPSC medium described above, which was replaced every 48hrs.

After 7-10 days, iPSC colonies with ES-like morphology were counted as they became visible and were subsequently scored by Alkaline Phosphatase staining according to manufacturer’s protocol (AP detection kit, Chemicon International, or, Promega #S3771). Colonies of iPSCs were picked after 2 weeks and expanded on feeder fibroblasts using standard procedures. Sox2-eGFP MEFs (D’Amour and Gage, 2003) were used in iPSC reprogramming experiments since they become Sox2-GFP-positive (reflecting activation of the endogenous pluripotency network) only in the final stages of iPSC reprogramming (see: Figures 3E, 3F and S3E).

Growth factors and small molecules to improve iPSC reprogramming

The media supplements to improve iPSC reprogramming, at the indicated concentrations shown in Figure S4C, are as follows: FG2 (R+D Systems #233-FB/CF); EGF (Sigma # E9644); Alki (SB431542; ALK4/5/7 inhibitor; Sigma# #S4317); Forskolin (Sigma # F6886); SCF (R+D Systems #455-MC/CF); DLPC (Lrh1 agonist; Stratech #850335P); 5-Aza-Deoxycytidine (Sigma # A3656-5MG); VPA (Calbiochem # 676380); TSA (Trichostatin A; Sigma; T8552); BIX (BIX 01294; Tocris #3364); Kenpaullone (Tocris #1398); Flavopiridol (Santa Cruz # CAS 146426-40-6).

Differentiation with retinoic acid

Differentiation of ES cells with retinoic acid (RA) was performed essentially as described (Savatier, 1996). LIF was first removed for 24 hrs by culture in LIF-free Differentiation medium (that is DMEM (high glucose) supplemented with serum 15%, non-essential amino acids, glutamax and beta-mercaptoethanol; hereinafter referred as “differentiation medium”). Next, LIF-free differentiation media was supplemented with Retinoic Acid at 10⁻⁶ M from +24 to +72 hrs, followed by LIF-free differentiation medium alone from +72 to +96 hrs. P19EC cell differentiation was by Retinoic Acid addition at 10⁻⁶ M.

EB Hanging-Drop Differentiation

This was performed essentially according to Marikawa et al., 2009. ES cells were transferred to Differentiation medium (that is DMEM (high glucose) supplemented with serum 15%, non-essential amino acids, glutamax and beta-mercaptoethanol; hereinafter referred as “differentiation medium”), and suspended in hanging drop culture at a cell density of 5000 cells/20 µL. ES cells were allowed to form spherical aggregates known as Embryoid Bodies (EBs) for 48 hours in the hanging drops before transfer to suspension culture in low-adherence petri-dishes. In suspension culture, fresh Differentiation medium was added every 3 days, and the percent of EBs was scored daily for the development of beating cells in cardiac centres.

Wound healing scratch assay

Three MEF clones were assessed for their ability to migrate and close a scratched region at day 3 +/- RPAP1 depletion. Scratch wounds (12 per experimental condition) were made in shSCR and shRPAP1 cultures and photographed at both +0 and +24 hrs in order to quantify the percent area of the original damage which remained at +24 hrs, using ImageJ software analyses of the photographs.

Cytometry

FACS was performed as described (Li et al 2009a). Briefly, for SSEA1 analysis, cells were collected by scraping and pipetting to unicellularize, before resuspension in 500 µLs 1xPBS and incubation with anti-SSEA1 antibody conjugated to allophycocyanin (R+D Systems, #FAB2155A) for 15 mins at room temperature. For AnnexinV analysis of apoptosis, the cells were collected by trypsinization before resuspension in 1xbinding buffer and incubation with anti-AnnexinV antibody conjugated to FITC (BD Pharmingen, # 556570). Data were analyzed with FlowJo 9.6.2 software. The percent of cells in S-phase was quantified using the Click-it EdU staining kit (Invitrogen #C35002). Briefly, cells were exposed to EdU in culture for 45 minutes followed by fixation and staining according to the manufacturer’s protocol.

Cell lysis and Western blot

Whole cell extracts were prepared using 50 mM TrisHCl pH8; 1 mM EDTA; 150 mM NaCl; 1% NP40; 0.5% Triton X-100; 1% SDS, with freshly added protease inhibitors (Roche #11873580001). A total
protein of 10 µg was loaded per lane and resolved on NuPAGE 4-12% gradient Bis-Tris gels, transferred to nitrocellulose and hybridized using antibodies as described in Table S6, sheet#2. Nuclear/Cytosolic Fractionation was performed by using the NE-PER Nuclear and Cytoplasmic Extraction Kit by Thermo Scientific, following the manufacturer’s instructions.

Histopathology and Immunohistochemistry
Mouse tissues were fixed in formalin at 4°C, embedded in paraffin block, and sectioned at a thickness of 5 µm. Sections were stained with hematoxylin and eosin for pathological examination or processed for immunohistochemical analysis with antibodies against mouse RPAP1 (for a list of the antibodies used, see Table S6, sheet#2). E3.0 morulae and E4.0 blastocyst embryos were collected in KSOM media (Chemicon #3699) and gently resuspended in 10% Formalin at 4C overnight to fix. Next day, embryos were resuspended in 100-200 µl of sterile 5% gelatin/dH2O pre-warmed at 37°C, then placed at 4C to allow gelatin solidification, followed by equilibration of the solid gelatin pellet in cold 10% formalin before embedding in paraffin block, and sectioning as above.

Immunofluorescence
Cells were grown on chamber slides using the same protocols as for the rest of the experiments. Briefly, at day 3 after RPAP1 depletion, cells were fixed with 4% paraformaldehyde for 2 minutes at room temperature, washed with PBS and permeabilized with PBS containing 0.02% Tween-20 for 20 minutes. Cells were blocked in PBS with 50% Australian FBS for 1 h and incubated with antibodies against RPAP1 or Pol II (for a list of the antibodies used, see Table S6, sheet#2) at 1:200 to 1:1000 in PBS-4%BSA, for 3 h, washed with PBS and further incubated with secondary anti-rabbit antibodies conjugated with Alexa-488, Alex-555 and/or Alexa-647 (1:500 in PBS-4%BSA). Nuclei were counterstained with DAPI. Confocal immunofluorescence cell images were captured using a Leica SP5, equipped with white light laser and hybrid detection.

RNA Pol II interactome analysis and LC/LC Mass Spectrometry
RNA Pol II immunoprecipitation was performed on Day+2 after lentiviral shRNA knockdown of RPAP1 in primary MEFs. Cells were washed x2 with ice-cold 1xPBS, then scrape-harvested in ice-cold 1xPBS. Lysates were prepared from two replicate experiments, sonicated, and clarified by centrifugation at 10C, at 10,000g, for 10 minutes. The supernatants were pre-cleared by exposure to Protein A/G beads (Santa Cruz #sc-2003). The Pol II complex was immunoprecipitated using a cocktail of three antibodies against RPBP1/Polr2a, the largest and core catalytic subunit of Pol II, in order to immunoprecipitate Pol II throughout all the stages of transcription. The antibodies targeted the N-terminus of Pol II (Santa Cruz, sc-899x), the Serine5-phosphorylated C-terminal domain (Abcam #5131), and the Serine-2-phosphorylated C-terminal domain (Abcam #5095). The immunoprecipitate fraction was eluted, specific Pol II protein interactors were determined by Mass Spectrometry, and the Pol II-interactome was analysed, as described below.

Immunoprecipitate sample preparation for Mass Spectrometry
Proteins were eluted from the agarose beads in two consecutive steps by shaking for 10 min at 1250 rpm in an Eppendorf Thermomixer in 2 bead volumes (~100ul) of elution buffer (UT: 8M Urea, 100mM Tris-HCl pH=8.0). The supernatant obtained was digested by means of standard FASP (Filter Aided Sample Preparation) protocol (Wiśniewski et al., 2009). Proteins were then reduced with 10 mM DTT, alkylated using 50 mM IAA for 20 min in the dark. Proteins were digested with Lys-C (Wako, Neuss, Germany) for 6 hours (1:50). Finally, samples were diluted in 50 mM ammonium bicarbonate to reduce the urea concentration to less than 1M, and were subsequently digested with Trypsin (Promega, Madison, WI; 1:100 sample concentration, overnight at 37 °C). Resulting peptides were desalted using a Sep-Pak C18 cartridge for SPE (Waters Corp., Milford, MA). Eluted peptides were vacuum-dried. To comprehensively identify the Pol II interactome, peptides were further pre-fractionated into five fractions using high pH reverse phase micro-columns (Batth et al., 2014), packing three discs (16 g diameter) of 3M Empore C18 at the bottom of a conventional 200 µL micropipette tip. After conditioning the tip, peptides were dissolved in 50 µl of Buffer A (20mM NH4, pH ≥ 10). Using an adapter, the tip was mounted on a 1.5 mL tube and fit in a benchtop centrifuge. During each fractionation step, centrifugation was operated at 1500 g for 2 min until all the volume passed through the C18 membrane. Peptides were subsequently eluted increasing the percentage of Buffer B (20mM NH3 in CH3CN) (i.e. 4, 8, 12, 80%) of Buffer B. All the five fractions and the flow through were dried in the speed-vac and resuspended in 22 µl 0.5% FA.

LC–MS/MS Analysis
The five fractions of the eight different samples were analyzed by RP chromatography using a nanoLC Ultra system (Eksigent, Dublin, CA), directly coupled with a LTQ-Orbitrap Velos instrument (Thermo) via nanoESI (Proxeon Biosystems, Waltham, MA). Peptides were loaded onto a Reprosil-Pur C18 column (3 μm, 400x0.075 mm; Dr. Maisch, Ammerbuch-Entringen Germany), with a trapping column (Prot Trap Column 0.3 x 10 mm, ReproSil C18-AQ, 5 μm), for 10 min with a flow rate of 2.5 L/min of loading buffer (0.1% FA). Elution was performed with a 120 min linear gradient (buffer A: 2% ACN, 0.1%FA; buffer B: 100% ACN, 0.1%FA) at 300 nL/min. Peptides were directly electrosprayed into the mass spectrometer using a PicoTip emitter (360/20 OD/μm tip id 10 μm, New Objective) at 1.4 kV spray voltage with a heated capillary temperature of 325°C and S-Lens of 60%. Mass spectra were acquired in a data-dependent manner, with an automatic switch between MS and MS/MS scans using a top 10 method. MS spectra were acquired with a resolution of 60,000 (FWHM) at 400 m/z in the Orbitrap, scanning a mass range between 350 and 1500 m/z (AGC = 1e6, Max IT = 500 ms). Peptide fragmentation was performed using collision-induced dissociation (CID) with read out in the ion trap (AGC = 5e3, Max IT = 100 ms) and a normalized collision energy of 35%.

Protein Pol II-interactome Data Collection and Analysis
Forty raw files (i.e. two experiments “SCR-Pol II vs SCR-IgG” and “shRPAP1-Pol II vs shRPAP1-IgG” with two biological replicates each and fractionated into five fractions), were analyzed using MaxQuant 1.5.3.30 (Cox and Mann, 2008) with Andromeda (Cox et al., 2011) as the search engine against a Mus musculus database (UniProtKB/Swiss-Prot, 43,539 sequences). Carbamidomethylation of cysteine was included as fixed modification. Oxidation of methionine, acetylation of protein N-terminal were included as variable modifications. Precursor mass tolerance was 20 ppm for the first search, and 4.5 ppm for the main search. Fragment mass tolerance was set to 0.5 Da. Minimal peptide length was set to 6 amino acids and a maximum of two missed-cleavages were allowed. Peptides were filtered at 1% FDR. For protein assessment (FDR <1%) in MaxQuant, at least one unique peptide was required for both identification and quantification. Other parameters were set as default. A total of 4,384 proteins were identified. Afterwards, the “protein-group” file was loaded in Perseus (v1.5.1.6) (Tyanova et al., 2016). After removing proteins annotated as contaminants, only identified by site and/or reversed a total of 3,944 proteins were quantified. Missing values in the IgG runs were replaced by the minimum LFQ value (i.e. 10) detected in the whole experiment. Using the LFQ values, all four possible pairwise comparisons between the two biological replicates of “SCR-Pol II vs SCR-IgG” were calculated. The same four comparisons were calculated for the “shRPAP1-Pol II vs shRPAP1-IgG” experiments. A protein was declared as specific interactor when the log2 enrichment ratio against its IgG was larger than 2.5 in three out of the four comparisons at least one of the two IP experiments. In total, 294 proteins were found as specific interactors (see Table S4). Among them, we identified all the subunits of the RNA pol II complex (12 proteins) and 28 out of 30 subunits of the Mediator complex. To identify interactors affected upon RPAP1 depletion, the data was normalized using the RPB1/Polr2a bait protein levels. Then, all four possible pairwise comparisons between “shRPAP-Pol II vs SCR-Pol II” experiments were calculated, and proteins were declared to be decreasing in the shRPAP1 if the log2 ratio was smaller than -1.5 in three out of the four comparisons. Proteins were declared to be increasing in the shRPAP-pol II if the log2 ratio was larger than 1.5 in three out of the four comparisons. The RPB1/Polr2a interactome in cells treated with the shRPAP showed alterations, specifically 104 interactors were absent or significantly reduced, while 5 new interactors were found (see Table S4). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino et al., 2016) partner repository with the dataset identifier PXD007114.

Protein Pol II-interactome Functional analysis
The interactors found to be affected in the RPAP1-depleted cells were functionally categorized using Panther database (http://pantherdb.org) by GO molecular function, GO biological process and GO cellular component. Statistical over-representation of GO terms (mouse genome was used as the background data set) was determined with a Binomial test and used the Bonferroni correction for multiple testing. P-values were then –log10 transformed for better graphical representation. These analyses revealed that the affected interactors in shRPAP1 were enriched in processes related to transcription and splicing (p<0.00001) (Figure S4E).

To find out whether these affected pol II-interactor proteins belong to specific complexes, we mapped our interactome data to the Corum database (Comprehensive resource of mammalian protein complexes) (http://mips.helmholtz-muenchen.de/corum/) that contains more than 3000 manually curated mammalian protein complexes. The number of subunits identified in the interactome data (specific interactors) for each known complex was retrieved. The same mapping was done with the list of
interactors found to be affected in the RPAP1-depleted cells. Corum complexes with less than 6 subunits were not considered and redundant complexes (those sharing identical subsets of proteins) were also removed. Several well-known complexes were represented in our dataset of Pol II-interactors which were affected by RPAP1 depletion (Table S4). Among them, the Mediator complex (which is formed by 30 subunits) was ranked the highest (Figure 4E) with eleven subunits affected following RPAP1 depletion (MED27, MED28, MED9, MED13, MED25, MED22, MED29, MED10, MED31, CDK8, MED14) indicating an important alteration in the functions controlled by this complex.

RNA isolation and Quantitative real-time PCR (qPCR)
Total RNA was extracted from cells on column by RNeasy kit with DNA digestion following provider’s recommendations (Qiagen # 74104) and retrotranscribed into cDNA following manufacturer’s protocol with Superscript Reverse Transcriptase (Life Technologies). Quantitative real-time-PCR was performed using Syber Green Power PCR Master Mix (Applied Systems) in an ABI PRISM 7700 thermocycler (Applied Biosystem). Input normalization of all the qRT-PCR data was by the 2–ΔΔCt method (Yuan et al., 2006) using the housekeeping genes β-Actin or Gapdh as indicated in each Figure, and as described (Ortega-Molina et al., 2015). Primers used are in Table S6, sheet#1.

RNA-seq transcriptomic analyses
For RNA-seq, samples of 1 μg of total RNA, with RIN numbers in the range 9.8 to 10 (Agilent 2100 Bioanalyzer), was used. PolyA+ fractions were processed using TruSeq Stranded mRNA Sample Preparation Kit (Agilent). Adapter-ligated library was completed by PCR with Illumina PE primers (8 cycles). The resulting directional cDNA libraries were sequenced for 40 bases in a single-read format (Genome Analyzer IIx, Illumina). The complete set of reads has been deposited in GEO (GSE78795). Sequencing quality for RNA-seq samples was analyzed with FastQC. Reads were aligned to the mouse genome (GRCm38/mm10) with TopHat-2.0.4 (Trapnell et al., 2012) (using Bowtie 0.12.7 (Langmead et al., 2009) and Samtools 0.1.16 (Li et al., 2009c), allowing two mismatches and five multihits. Transcripts assembly, estimation of their abundance, and differential expression, were calculated with Cufflinks 1.3.0 (Trapnell et al., 2012), using the mouse genome annotation data set GRCm38/mm10 from the UCSC Genome Browser (Rosenbloom et al., 2015).

Functional analyses of differential gene expression
For differential gene expression lists (see data in Table S1: ES cells +24hr differentiation; or Table S3: MEFs at day 3 after RPAP1 depletion). Genes were ranked using the FDR q-value statistic to identify significant genes (FDR<0.05 or FDR<0.01, as indicated in the Figures), then by fold change in expression. Selected differentially-expressed genes identified in the RNA-seq were validated by qPCR. Venn diagrams were generated by JVenn (Bardou et al., 2014) and hypergeometric testing was performed to assess any significant overlaps. Pathway analyses were by Ingenuity Pathway Analysis software (www.ingenuity.com).Gene Set Enrichment Analysis (GSEA; Subramanian et al., 2005). GSEAPre-ranked was used to perform a gene set enrichment analysis of annotations from the MsigDB Hallmarks, C5-Gene Ontoly (GO) terms, C2-Curated, KEGG, Reactome and NC1 databases, with standard GSEA and Leading Edge analysis settings. We used the RNA-seq gene list ranked by statistic, setting ‘gene set’ as the permutation method and ran it with 1000 permutations for Kolmogorov-Smirnov correction for multiple testing. We considered only those gene sets with significant enrichment levels (FDR q-value < 0.25) (Subramanian et al., 2005) (see: Table S2). GSEA Enrichment data were obtained and ranked according to their FDR q-value (see: Table S2). Heatmaps of GSEA data (Figures 2H and 5F) or qPCR data (Figure 3G) were generated using Gene Pattern (Reich et al., 2006).

Supervised Network Analysis
Investigation of differential gene expression for dominant gene-ontologies or functions was performed by supervised network analyses. Briefly, network analyses were performed starting from the list of differentially expressed genes induced by RPAP1 depletion followed by 24hrs of ES differentiation, or separately, RPAP1 depletion for 3 days in MEFs. Next these lists were used to find gene interaction information in the MetaCore™ database, including manually curated experimentally validated interaction data. The interaction datasets generated (including information of the interaction direction –i.e. source and target genes– and interaction effect –i.e. inhibition or activation–) were contextualized for obtaining the gene regulatory networks of the RPAP1-depletion and control phenotypes, using an algorithm developed in-house (Crespo et al., 2013; Zickenrott S et al., 2016). Finally, the phenotype-specific networks were compared to identify the pathway enrichment in genes in the “up-regulated” or “down-regulated” lists. In this comparison we estimate the statistical significance (i.e. enrichment) of the
interactions among genes in each category, which constitute an indication of the differences in the regulatory mechanisms underlying the phenotypical changes caused by RPAP1 depletion. See Table S1, sheets#6-11; and Table S3, sheets#6-10.

Comparison of differential gene expression with the iPS roadmap

Gene expression changes have been comprehensively characterized in the subset of successfully-progressing cells during iPS reprogramming by overexpression of the OSKM Yamanaka factors (Polø et al., 2012; Hansson et al., 2012). We first identified the gene expression changes which occur between day 0 and day +3, or between day 0 and day +9, of successful iPS reprogramming by comparing RNA-seq data in the parental MEFs (day 0) versus day +3 (or day +9) Thy1-negative cells in the published datasets. Next, we used GSEA to compare these iPS roadmap genesets of top 100 or top 500 up- or down-regulated mRNAs versus the complete ranked list of differential gene expression in MEFs at day 3 after RPAP1 depletion. We also performed the analysis in reverse, comparing the genomes of significantly differentially expressed mRNAs up- or down-regulated in MEFs at day 3 after RPAP1 depletion, versus, the complete ranked list of differential gene expression at day 3 (or day 9) of the iPS roadmap. GSEA results are shown in Figure 3A and S3A and S3B. Data with P<0.05 and FDR<0.25 are considered significant.

Conversion of Plant gene expression data to Mammalian homologs

The effect of RPAP1-mutation on mRNA expression levels was previously published in Arabidopsis (Sanmartín et al., 2011). We converted the published data from plant (31,200 genes; see: Table S2, sheet#4) to mammal (mouse) via protein sequence similarity (Table S2, sheets #5 and #6), filtering the data by three thresholds: (i) “100% coverage”, that is, the whole plant protein is included in the alignment against whole mouse proteins; (ii) the best “%amino acid Identity” possible, always greater than 20% (% of amino acids than are totally conserved in both sequences); (iii) the best “%Positive amino acids” as possible (this takes into account synonymous amino acids (that is based on similarity in terms of size and charge). We filtered out: 250 genes that did not map (neither in Arabidopsis TAIR 10 database nor in EnsEMBL), and a further 1933 Arabidopsis genes were without homology/orthology in mouse, however, the majority of these were transposons (Table S2, sheet#7). We ran GSEA using the MSigDB Hallmark, C5-GO terms, and C2-Curated databases against the entire remaining ranked list of homologous plant proteins/genes (for the ranked list of genes converted to mouse, see Table S2, sheets #5 and #6) to identify significant genesets up or down-regulated by RPAP1-mutation in plants which have a homolog in mouse (see summary of results in Table S2, sheet #7). In Figure 5F, the heatmap compares the GSEA hallmark database analyses results from the plant-mammal conversion above (Table S2, sheet #7), versus, the GSEA hallmark database analyses results for three other experiments: (i) GSEA on the ranked list of differential mRNA expression in MEFs at day 3 +/- RPAP1-depletion; (ii) GSEA on the ranked list of differential mRNA expression in ES cells at +24hrs after inducing differentiation, +/- RPAP1-depletion; (iii) GSEA on the ranked list of differential RNA Pol II abundance at the promoter at day 3 +/- RPAP1 depletion.

Chromatin Immunoprecipitation (ChIP) and deep-sequencing

ChIP-qPCR was performed as described (Li et al., 2012) with primers listed in Table S6 and antibodies for Total Pol II (Santa Cruz N20, sc-899x) and RPAP1 (Cosmo Bio MK14030910). ChIP-seq for Pol II was performed as described (Rahl et al., 2010). Briefly, cells were fixed using 1% formaldehyde, scrape-harvested, resuspended in ChIP lysis buffer (1% SDS, 10mM EDTA, 50mM Tris-HCl, pH 8.1) and sonicated using Covaris water bath sonicator to generate fragments of 150 to 500 bp. Soluble chromatin was diluted 10 fold in ChIP Dilution buffer (1% Triton X-100, 2 mM EDTA pH 8.0, 150 mM NaCl) precleared with Agarose Protein A/G beads (Santa Cruz), and then incubated with antibody specific for total RNA Pol II (N-20, sc-899x, Santa Cruz) or specific for the RNA Pol II Ser5P-phosphorylated form (Abcam #ab5131). After incubation, immunocomplexes were collected with Agarose Protein A/G beads (Santa Cruz). Next, the immunocomplexes were washed sequentially with Low Salt Wash Buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl, pH 8.1, 150mM NaCl), High Salt Wash Buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl, pH 8.1, 500mM NaCl), LiCl Wash Buffer (0.25M LiCl, 1% NP40, 1% deoxycholate-Na, 1mM EDTA, 10mM Tris-HCl, pH 8.1) and washed twice with TE (10 mM Tris-HCl pH7.5, 1mM EDTA). Immunocomplexes were eluted in ChIP elution buffer (1%SDS, 0.1M NaHCO3) and the crosslinking was reverted by incubation at 65 °C for 8 hrs with 200 mM NaCl. Samples were treated with Proteinase K and RNase A ,and DNA was extracted using Phenol-Chloroform. DNA precipitation was in 100% ethanol with 0.1 M NaAcetate pH5.2 and 2 uLs glycoen (Roche). The DNA pellet was washed with 70% ethanol, and resuspended in ddH2O. Purified chromatin was used for library construction.
For ChIP-seq the amount of DNA used was ~5 ng from each sample (as quantitated by fluorometry). Samples were were processed through subsequent enzymatic treatments of end-repair, dA-tailing, and ligation to adapters as in Illumina's "TruSeq DNA Sample Preparation Guide" (part # 15005180 Rev. C). Adapter-ligated libraries were completed by limited-cycle PCR with Q5 High-Fidelity DNA Polymerase (NEB) and Illumina PE primers (15 cycles), and further purified with a double-sided SPRI size selection to obtain a size distribution in the range of 230-500bp. Libraries were applied to an Illumina flow cell for cluster generation (TruSeq cluster generation kit v5) and sequenced on the Genome Analyzer IIx with SBS TruSeq v5 reagents by following manufacturer's protocols, to 20-25 million reads per sample.

**Pol II ChIP-seq data analyses**

Definition of promoter and gene body regions (See: Figure S5B) and the calculation of Pol II total and Ser5P abundance along genes was based on methods of Young and colleagues (Rahl et al., 2010) (see Table S5). Sequencing quality for ChIP-seq samples was analyzed with FastQC (Andrews, 2011). Reads were aligned with Bwa 0.7.5a (Li and Durbin, 2009) to the mouse reference genome (GRCm38/mm10) using the default seed length (32) and allowing 1 mismatch in the seed. SAMtools 0.1.16 (et al., 2009b) was used to convert the output alignment SAM files to the BAM file format, sort the alignments and eliminate duplicated reads. BEDTools 2.23.0 (Quinlan, 2014) was used to convert the resulting files to the BED format. All ChIP and input samples were randomly normalized to the same number of reads. Peak calling was performed with MACS 2.0.10.20130712 (Feng et al., 2012) using the input sample as control for each one of the ChIP samples. BigWig files were obtained with bedGraphToBigWig (Kent et al., 2010) from the BedGraph files generated with MACS. Resulting peaks were annotated with PeakAnalyzer 1.4 (Salmon-Divon et al., 2010), and the distribution of peaks was plotted with SeqMiner 1.3.3e (Ye et al., 2014) with color-scaled intensities are in units of reads per million mapped reads (rpm). Transcription Start Sites (TSS) and Transcription Termination Sites (TTS) were identified using the Database of Transcriptional Start Sites (http://dbtss.hgc.jp). Metagenes were aligned +/- 5 Kb around the TSS. The Pausing Index (PI) for gene promoters versus gene bodies was calculated as described (Rahl et al., 2010; see Figure S5B). First, the number of reads per nucleotide was computed with BEDTools 'genomecov'; second, to extend this number to the number of reads per gene promoter or gene body, BEDTools 'map' was used; and third, the Pausing Index was calculated for each gene promoter or gene body as PI = ((number of reads in region / region size)*scaling factor)*10^5. Scaling factor = (total number of reads in sample/genome length).

**Definition of MEF super-enhancers, their target genes, and their eRNA levels**

For Figure 5G: MEF super-enhancers were defined by H3K27Ac ChIP-seq signal and ranking by ROSE, as previously described (Whyte et al., 2013; Shen et al., 2012; Khan and Zhang, 2016; dbSUPER, http://bioinfo.au.tsinghua.edu.cn/dbsuper/). To identify the single-nearest target gene to each MEF super-enhancer, GREAT analysis was performed as described (GREAT v3.0.0; McLean et al., 2010). In Figure 5G, this genset of MEF super-enhancer target genes was used in GSEA analysis of the mRNA expression levels of these genes at day 3 after RPAP1 knockdown in primary MEFs in our data.

For Figure 5H: the same MEF super-enhancer regions were assessed for enhancer-RNA (eRNA) abundance which has been reported to be increased (~10%, n=64), decreased (~10%, n=63), no-change/not-detected (~80%). In Figure 5H, average RNA abundance on the super-enhancers with increased or decreased eRNA levels was visualized in 50bp bins from start to end of feature using SeqMINER (Ye et al., 2014).

For Figure S5G, GSEA was performed as described above for Figure 5G, except here, in order to assess any changes in housekeeper mRNA expression levels. No significant change in housekeeper gene expression was detected, despite performing GSEA using the following housekeeper genestsets: (i) a full set of 3,384 housekeeper genes (defined in Eisenberg and Levanon, 2013); (ii) 10 sets of 500 genes, each randomly selected from the full 3,384 housekeepers, performed so that the individual genes could be visualized in the GSEA enrichment plot. The data shown in Figure S5G is representative of one of the random selections of 500 housekeeper genes from the above list of 3,384 genes where no significant change in housekeeper mRNA was detected.

For Figures S5H, S5I, and S5J: GREAT analysis (GREAT v3.0.0; McLean et al., 2010) was performed on the following two groups of super-enhancers, defined above, to identify the single-nearest target gene of each super-enhancer: (i) super-enhancers with increased eRNA levels (Figure S5H); (ii) super-enhancers with decreased eRNA levels (Figures S5I and S5J). Next, these two enhancer-target-gene groups were assessed separately for any enrichments in their functions (Gene Ontology Biological process) or the developmental stage associated with their expression (MGI Expression-Detected; Theiler
Stage of embryo development). The data is presented in Figures S5H to S5J. In Figure 5I, the Theiler Stage of embryo development associated with these two enhancer-target-gene groups (the enhancers with increased or decreased eRNA/activity levels) is shown, together with the approximate embryo day-post-coitus (dpc) (emouseatlas.org; Bard et al., 1998). Super-enhancers with decreased eRNA levels (and thus putatively decreased activity) were associated with target-genes expressed during the period embryo dpc E11.5-E17. Conversely, super-enhancers with increased eRNA levels (and thus putatively increased activity) were associated with target-genes expressed during the period embryo dpc E8-E13. Since primary MEFs derive from E13.5, the decrease in activity of enhancers associated with E11.5-E17, coupled with the increase in activity of enhancers associated with E8-E13 mirrors the gene expression analysis in Figure 2, where MEFs at Day3 after RPAP1 knockdown appear to have de-differentiated, and in Figure 3, where this pattern of de-differentiation correlates significantly with the first 3 days of iPS reprogramming when MEF cell identity is erased (Polo et al., 2012).

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise specified quantitative data are presented as mean +/- SD and significance was assessed by the two-tailed Student’s t test; *p<0.05, **p<0.01, ***p<0.001.

DATA AND SOFTWARE AVAILABILITY

Data Resources. Accession Numbers: Three datasets (two RNA-seq and one ChIP-seq experiment) are available from the GEO database: GSE78795. The mass spectrometry proteomics data are available from the ProteomeXchange Consortium/PRIDE repository with the dataset identifier PXD007114.

SUPPLEMENTAL REFERENCES


Click here to access/download
**Supplemental Movies and Spreadsheets**
Suppl Table S1.xlsx
Click here to access/download
Supplemental Movies and Spreadsheets
Suppl Table S2.xlsx
Click here to access/download

Supplemental Movies and Spreadsheets
Suppl Table S3.xlsx
Click here to access/download
Supplemental Movies and Spreadsheets
Suppl Table S4.xlsx
Click here to access/download
Supplemental Movies and Spreadsheets
Suppl Table S5.xlsx
Click here to access/download

**Supplemental Movies and Spreadsheets**

Supp Table S6.xlsx