MacroH2A histone variants modulate enhancer activity to repress oncogenic programs and cellular reprogramming

Wazim Mohammed Ismail
Mayo Clinic https://orcid.org/0000-0002-1782-7296
Amelia Mazzone
Mayo Clinic
Jagneet Kaur
Mayo Clinic
Stephanie Safgren
Mayo Clinic
John Moore-Weiss
Mayo Clinic
Marina Buciuc
Mayo Clinic
Lynzie Shimp
Mayo Clinic
Luis Duarte
Icahn School of Medicine at Mount Sinai
Chandandeep Nagi
Baylor College of Medicine
Saul Carcamo
Icahn School of Medicine at Mount Sinai
Chi-Yeh Chung
Icahn School of Medicine at Mount Sinai
Dan Hasson
Icahn School of Medicine at Mount Sinai
Neda Dadgar
Mayo Clinic https://orcid.org/0000-0003-3220-7148
Alexander Revzin
Mayo Clinic
Jian Zhong
Mayo Clinic https://orcid.org/0000-0001-7157-603X
Jeong-Heon Lee

Mayo Clinic

Tamas Ordog

Mayo Clinic https://orcid.org/0000-0002-3940-7284

Emily Bernstein

Icahn School of Medicine https://orcid.org/0000-0001-6533-8326

Alexandre Gaspar-Maia (**maia.alexandre@mayo.edu**)

Mayo Clinic https://orcid.org/0000-0001-8528-5575

Article

Keywords: Active Enhancer Elements, H3 Lysine 27 Acetylation, Cis-regulatory Elements, Cell Identify Maintenance, Oncogenic Programs, Single Cell Epigenomic Profiling

Posted Date: April 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-384560/v1

MacroH2A histone variants modulate enhancer activity to repress oncogenic programs and cellular reprogramming

Wazim Mohammed Ismail[1,2,+], Amelia Mazzone[1,2,+], Jagneet Kaur[1,2], Stephanie Safgren[1,2], John Moore-Weiss[1,2], Marina Buciuc[1,2], Lynzie Shimp[1,2], Luis F. Duarte[3, 4], Chandandeep S. Nagi[5], Saul Carcamo[3,4], Chi-Yeh Chung[3,4], Dan Hasson[3], Neda Dadgar[6], Alexander Revzin[6], Jian Zhong[2], Jeong-Heon Lee[1,2], Tamas Ordog[2], Emily Bernstein[3, 4], Alexandre Gaspar-Maia[1,2*]

[1] Division of Experimental Pathology, Department of Lab Medicine and Pathology, [2] Center for Individualized Medicine, Epigenomics program, Mayo Clinic, Rochester, MN, USA

[3] Department of Oncological Sciences, [4] Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[5] Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX US

[6] Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, USA

+ These authors contributed equally to this work

*Correspondence to:

Alexandre Gaspar-Maia, PhD
Department of Laboratory Medicine and Pathology

Assistant Professor of Laboratory Medicine and Pathology
Assistant Professor of Pharmacology
Mayo Clinic College of Medicine and Science

maia.alexandre@mayo.edu
+1-507-293-2464

Mayo Clinic
200 1st Street SW, MN 55905
United States

Abstract (173), Text (4176), Figures (7), References (42), Supplementary Figures and Tables (8), Supplementary References (25)

ABSTRACT

Considerable efforts have been made to characterize active enhancer elements, which can be annotated by accessible chromatin and H3 lysine 27 acetylation (H3K27ac). However, apart from poised enhancers that are observed in early stages of development and putative silencers, the functional significance of *cis*-regulatory elements lacking H3K27ac is poorly understood. Here we show that macroH2A histone variants mark a subset of enhancers in normal and cancer cells, which we coined 'macroH2A-Bound Enhancers', that negatively modulate enhancer activity. We find macroH2A variants enriched at enhancer elements that are devoid of H3K27ac in a cell type-specific manner, indicating a role for macroH2A at inactive enhancers to maintain cell identity. In following, reactivation of macro-bound enhancers is associated with oncogenic programs in breast cancer and its repressive role is correlated with the activity of macroH2A2 as a negative regulator of BRD4 chromatin occupancy. Finally, through single cell epigenomic profiling, we show that the loss of macroH2A2 leads to increased cellular heterogeneity that may help to explain the role of macroH2A variants in defining oncogenic transcriptional dependencies.

INTRODUCTION

Enhancers are *cis*-regulatory elements found throughout the eukaryotic genome that are bound by transcription factors (TF) and coactivator complexes[1,2] playing a key modulatory role in gene expression. Chromatin landscape profiling has revealed specific patterns at enhancer regions consisting of a nucleosome-depleted region that is flanked by histones harboring specific post-translational modifications (PTMs) such as H3K4me1 and H3K27ac[3]. This combination of PTMs has been broadly utilized for epigenomic annotation of active enhancers, facilitating systematic discovery and functional understanding of this important class of *cis*-regulatory elements[4]. However, our ability to define an inactive enhancer state has been more elusive due mainly to their association with repressed transcriptional activity. In the absence of H3K27ac (and, in some instances in the presence of H3K27me3), H3K4me1, which in and of itself is largely dispensable for transcription[5], has been associated with enhancer states that are repressed or poised/primed for activation[6-9]. Recently, H3K27me3-rich genomic regions that negatively regulate gene expression via proximity or looping have been proposed as potential silencers[10]. Together, these data indicate that the regulation of repressed/poised *cis*-regulatory elements may be more complex than previously thought. By extension, inactive states may have biological relevance in the context of cellular identity and homeostasis. This is particularly true during oncogenic transformation, where plasticity and reprogramming are altered due in part to genetic or structural disruption of *cis*-regulatory regions[11] leading to re-activation or hijacking of enhancer elements[12]. Therefore, we hypothesized that dysregulation of the establishment and maintenance of repressive chromatin states in *cis*-regulatory regions could play a role in oncogenic transformation.

Histone variant incorporation into the nucleosome has distinct effects on gene expression, regulating cell specification in both development and cancer[13]. MacroH2A (mH2A) histone variants contain a 30kDa non-histone domain (macro domain) at their C-termini[14] and are associated with the inactive X chromosome[15], various forms of heterochromatin[16], and inactive genes[17-20]. MacroH2A1 and macroH2A2 isoforms are encoded by two distinct genes (*H2AFY* and *H2AFY2*, respectively), and macroH2A1 is alternatively spliced, resulting in two macroH2A1 isoforms, macroH2A1.1 and macroH2A1.2, that differ by only one exon in the macro domain[21]. The incorporation of mH2A variants into the genome occurs in large chromatin domains[22] most often marked by the Polycomb-mediated repressive histone modification H3K27me3[19,23] and in some instances by H2BK12ac[24]. While recent reports have described a dynamic process by which such mH2A domains are negatively defined by exclusion from actively transcribed regions[22], other regions of the genome are enriched for macroH2A with undefined functions. Previously, mH2A variants have been implicated in the maintenance of cell identity when challenged during somatic cell

reprogramming[23,25], acting as an epigenetic barrier in association with H3K27me3 through co-localization at pluripotency genes in differentiated cells [21]. In cancer, mH2A variants generally act as tumor suppressors, whose expression is reduced in several tumor types including melanoma, lung, bladder, and breast cancers, as compared to normal tissues and/or early cancer stages[26-30].

Accumulating evidence suggests that the genetic or structural disruption of enhancer function represents a major cause of human disease[11], particularly in cancer, with a growing body of studies pointing towards the re-activation or hijacking of enhancer elements for the activation of oncogenic pathways[12]. However, the regulation of inactive enhancer elements in normal physiology and cancer remains poorly understood. Moreover, our understanding of the role of histone variants at enhancers is limited. In this study, we demonstrate through extensive epigenomic analysis that mH2A variants regulate gene expression through enhancer modulation and identify a novel class of *cis*-regulatory elements, which we termed macro-bound enhancers (mBE). We find that mBE are inactive and play a role in preserving cell identity and limiting cellular heterogeneity with important implications for cellular reprogramming and activation of oncogenic pathways.

RESULTS

Characterization of macro-Bound Enhancers

We performed correlation analysis between macroH2A (mH2A) variants ChIP-seq signal and the ENCODE chromatin states in two different primary cell types (human mammary epithelial cells, HMEC, and normal human melanocytes, NHM), which indicated an enrichment of mH2A at enhancer elements (Fig. 1a). Interestingly, H3K27me3 and macroH2A highly correlate in the 'Repressed Polycomb' chromatin state but the presence of mH2A at enhancer elements is unique, as H3K27me3 is absent (Fig.1a, Supp. Fig. 1a). A pipeline was developed to determine the enrichment of mH2A in enhancer elements, using the ENCODE *cis*-regulatory elements (CRE) framework (Fig.1b)[31]. Following the alignment of all enhancers to the summit of cell-type specific open chromatin regions analyzed using ATAC-seq (in HMEC, NHM and the breast cancer cell line MCF-7), with the incorporation of H3K4me1 peaks and exclusion of a blacklist of ambiguous genomic regions[31], *k*-means clustering (*k*=5) was performed on the overlap with identified cell-type agnostic CREs using the ChIP-seq signal from H3K4me1, H3K4me3, H3K27ac, H3K27me3, H2A.Z, mH2A1, mH2A2 and when available, CTCF. The five clusters identified could be characterized as follows (Fig.1c, Supp. Fig. 1b,c): Active Enhancers (enriched in H3K27ac), Active Promoter-Like (APL, enriched with H3K4me3), Inactive Enhancers (low H3K27ac), ATAC-only (mostly absent of

any other mark used in the classification) and a large subset of enhancers with low H3K27ac with strong mH2A signal, a class of enhancers we coined macro-Bound Enhancers (mBEs). Genomic location distribution of the five classes revealed the expected enrichment of APL around annotated transcription start sites (TSS) and an enrichment of active enhancers and mBEs at intronic and intergenic regions (Fig. 1d, Supp. Fig. 1d). Since mH2A variants have previously been associated with H3K27me3 around TSSs and gene bodies (Supp. Fig.1a), the presence of mH2A at enhancer elements lacking H3K27me3 was unexpected, which was most pronounced in HMEC (Fig. 1d, Supp. Fig. 2a). The distribution of these five classes across the cell-type agnostic ENCODE CRE classifications also revealed a consistent enrichment of APL in the ENCODE promoter-like signatures (Supp. Fig. 1b) and a more robust distribution of mBEs at distal enhancer-like signatures. It is worth noting that the previously reported association of mH2A with H2BK12ac[24] was not found at mBEs (Supp. Fig. 2b). Also, mBEs are devoid of H2A.Z, an H2A variant associated with active TSS and enhancers[32]. Interestingly, the ATAC-only class was significantly enriched in H3K36me3, which could be explained by its predominance at intronic regions of expressed genes. Not surprisingly, all five classes show similar levels of conservation, but DNA methylation patterns are mostly low in active and mBE enhancers, which could suggest a primed state of mBEs (Supp. Fig. 2c). Moreover, super-enhancer clusters and mBEs were mostly mutually exclusive (Supp. Fig. 2d).

To validate our pipeline of enhancer mapping, we sought to compare expression at those regulatory elements through publicly available RNA-seq data, as a proxy for their activity[33]. In all three cell types, the highest expressing elements are ALP and the ATAC-only enhancers, which corroborates the idea that most of these later elements are present in intronic regions of expressed genes, since they are also enriched with H3K36me3. The class with the lowest expression detected in the non-malignant cells is mBE (Fig. 2a, Supp. Fig. 3a). The expression in normal mammary tissue[33] of the enhancer elements as identified in HMECs revealed the same pattern (Fig. 2b). This suggests that the definition of such CREs in mammary epithelial cells is also reflective of enhancer activity in human samples. Moreover, enhancer-gene association of the five classes of CRE confirms that inactive and mBE enhancers are associated with the lowest expressing genes (Fig.2c). Finally, we queried whether mBE would differ between biosamples. Interestingly, the strongest overlap from the three samples was in the APL class with mBE and the other enhancer classes having fewer common elements (Fig.2d and Supp. Fig.3b), indicating a cell specificity to regulatory elements, in general. As expected, the two samples derived from the breast (HMEC and MCF7) showed greatest overlap in APL. Gene ontology analysis (GO) performed on genes common to the two breast-derived cell lines and associated with APL or mBEs (breast-associated APL and mBE) indicated a

fundamental difference (Fig. 2e). The mBE-associated genes were highly enriched in developmental processes that are also relevant in breast cancer, including epithelial to mesenchymal transition (EMT), while APL-associated genes were mainly associated with cell cycle and gene expression (Fig.2f). These results suggest an important role of mBE as gatekeepers of cellular identity and regulation of developmental specifications.

MacroH2A is a negative modulator of enhancer activity

In order to address the functional role of mBEs, we performed cellular reprogramming in cells derived from double knockout (dKO) mice[34] lacking the genes encoding both mH2A variants (*H2AFY* and *H2AFY2*)[23]. Since these murine dermal fibroblasts (DFs) from this model were previously used to show that mH2A variants act as a barrier to reprogramming[23], we hypothesized that mBEs could be enriched in the consensus binding domains of the four iPS reprogramming factors, Oct4 (O), Sox2 (S), Klf4 (K) and Myc (M). Consistent with the results obtained in the human cell lines, CRE analysis of DFs revealed the same five classes as described above, including mBEs (Fig. 3a). We next analyzed the enrichment of mH2A variants and H3K27me3 in DFs at the OSKM binding sites[35] 48 hours after OSKM expression. Interestingly, the binding sites of the three pioneering factors (OSK) were enriched in mH2A1 and mH2A2, but not H3K27me3 (Fig. 3b). To further decipher a potential role of mBEs as an epigenetic barrier during reprogramming, ChIP-seq data for four TFs highly expressed in fibroblasts (Fra1, Cebpa, Cebpb, and Runx1), three chromatin regulators (Brg1, p300 and Hdac1), and the OSKM factors obtained at 48hr during iPS reprogramming[35] were used to calculate the enrichment of each class of enhancers (Active, Inactive, ATAC-only, and mBE) at these binding sites. Odds ratio analysis using a Fisher exact test revealed enrichment of Hdac1 and cMyc in Inactive Enhancers and enrichment of Sox2 and Oct4 in mBEs, confirming the presence of mH2A variants at the same loci bound by the pioneer factors during the early phases of reprogramming (Fig. 3c).

To functionally address whether mH2A modulates enhancer activity, DFs isolated from mH2A dKO mice were used in iPS reprogramming experiments as described[23]. These cells completely lack mH2A variants, which permitted the implementation of a strategy to assess the effect of mH2A macro domain at a single locus by means of CRISPR/Cas9-mediated epigenome editing using a dCas9 chimeric protein containing either the macro domains or the known repressor KRAB domain (Fig.3d). This approach enabled the modulation of a specific enhancer and allowed us to assess the degree of the repression. First, an embryonic stem cell (ESC) line (NG4) expressing GFP under the control of the *Nanog* promoter and regulatory region (180 Kb upstream of TSS) was used to establish cell lines with different Cas9 constructs (Supp. Fig. 4a-c). NG4 cells express green fluorescence under normal ESC

growth conditions[36], and targeting of a known regulatory element of *Nanog* (Supp. Fig. 4d,e) should decrease its expression. SpCas9 was used as a positive control for targeting of the region of interest, dCas9 alone was used as a negative control, and dCas9-KRAB was a positive control for negative modulation of the target enhancer (Supp. Fig. 4f,g). Expression of GFP was determined by FACS. The effect of dCas9-macro1.2 and dCas9-macro2 was comparable to dCas9-KRAB, especially at the enhancer locus and at the GFP gene body, indicating that the macro domains have the inhibitory effect both in transcribed regions (as expected) and at enhancer elements. After validating the enhancer modulation in ESCs, a similar experiment was then performed to examine the effect on endogenous Nanog expression (Fig.3d-f, Supp. Fig. 4h) during the process of reprogramming of mH2A dKO DFs. Expression of Nanog was also reduced upon targeting the enhancer with dCas9-macro2 after four days of iPS reprogramming (Fig. 3f). Thus, the presence of macroH2A at enhancers during reprogramming may hinder their activation, explaining in part the role of macroH2A as an epigenetic barrier for reprogramming.

Reactivation of macro-Bound Enhancers associates with oncogenic programs

Given the above, mBEs may regulate cellular homeostasis and potentially serve as gatekeepers of cell identity by limiting plasticity[37]. In turn, the loss of mH2A during cancer progression could serve as an opportunity for oncogenic gene expression programs by means of enhancer activation. Decreased mH2A expression has been described in a variety of different tumors[13,30] and has been implicated in processes such as EMT in breast cancer[37]. However, a thorough analysis of loss of mH2A variants has not been performed in mammary carcinoma. Chromatin fractionation of a panel of breast cancer lines permitted the identification of several cell lines with a prominent loss of mH2A, particularly mH2A2. The loss of mH2A2 was not limited to a particular sub-type or mutational status, even though it was more associated with aggressive tumors such as triple negative (TN) and HER2-amplified cancers (Fig. 4a). We then investigated mH2A2 levels in two cohorts of patient samples (patients from Mount Sinai Hospital and Breast Cancer Progression tissue microarrays) by immunohistochemistry. Similar to the cell lines, mH2A2 was lost in invasive tumors and in tumors with advanced grades (II and III) when compared to ductal carcinoma *in situ* or grade I tumors, respectively (Fig. 4b).

To evaluate if the loss of mH2A could also be correlated with reactivation of enhancer elements associated with oncogenic programs, activity of the mammary epithelial CREs (as defined in HMEC) in breast cancer cells lines was analyzed using the ChIP-seq signal for H3K27ac from 12 different breast cancer cell lines[38,39] including the non-tumorigenic cell line MCF10A. Principal component analysis of the H3K27ac signal at mBEs indicated a strong

correlation between cancer cell lines from the same cancer subtypes analyzed, i.e., luminal, HER2-amplified, and TN (Fig. 4c). This compelling association indicates that breast cancer sub-types can be identified based on the activity of specific enhancers that were found enriched in mH2A in normal mammary epithelial cells. To further validate the usage of such mBEs in the different breast cancer programs from different subtypes, we turned to enhancer transcription analysis from TCGA RNA-seq data[33] where patient samples were used from both the tumor and the adjacent non-malignant tissue. CRE expression of each pair (tumor vs. normal tissue) was identified for each subtype of breast cancer and showed a significant proportion of mBEs changing in the breast cancer samples for all the sub-types (Fig. 4d).

mH2A2 is a negative regulator of estrogen targets.

Next, we aimed to define the role of mBEs in a tumorigenic cell line, MCF7, which expresses both variants and in which mBEs have been characterized (Fig. 1c, 4a). Moreover, MCF7 is a known ER-responsive cell line with well-defined enhancer network[40].In light of the mBE's cell type-specificity (Fig. 2d), we first performed *in silico* analysis of TF and chromatin regulator binding to DNA in MCF7 cells to understand mBE-related regulation. Since this is a commonly used model system, several ChIP-seq datasets could be used to compare binding of different factors and the annotated enhancers (ReMAP). We applied Fisher's exact test to every binding event in MCF7 cells that exist in both the ReMAP database and our annotated enhancer sub-groups (Fig. 5a). Unsurprisingly, Active Promoters were most enriched in DNA-binding proteins. Despite mBEs being repressive and mostly depleted of significant binding events, we identified preferential mBE binding of TFs associated with ER activation, namely, GATA3 and FOXA1, in MCF7 cells. This was a surprising result given that mBEs are inactive and MCF7 cells are actively maintaining an ER transcriptional signature in order to proliferate. A potential explanation is that mBEs may help maintain enhancer stability and define TF programs in a more robust and predictable way in order to generate and preserve cellular homeostasis and prevent unwanted cellular heterogeneity. Therefore, we hypothesized that mBEs could maintain ER-responsive enhancer elements inactive and thus serve as a gatekeeper of the MCF7 enhancer network. This notion is also supported by our finding that MCF7 cells have mH2A2 levels similar to non-tumorigenic cells (Fig. 4a). It follows from this hypothesis that loss of mBEs would render the ER-dependent transcriptional program even more accentuated. To test this idea, we depleted mH2A2 in MCF7 cells by CRISPR/Cas9 genome editing using four sgRNAs (Fig. 5b,c, Supp Fig. 5a-c). After screening for efficient sgRNAs (Supp Fig. 6b), we isolated and expanded two non-targeting control clones and two mH2A2 knockout (KO) clones with similar proliferative potential (Fig. 5d). To analyze the ER response, we infected MCF7 clones with GFP-H2A

(for imaging quantification purposes) and grew them in 2D conditions with EtOH or 17β-estradiol (E2; used as an ER agonist) for 5 days before plating them as 3D spheroids for another 7 days in the presence or absence of E2. 3D spheroids were monitored using microfluidic devices and printed microwells that allow for accurate growth quantification using GFP fluorescence. After 7 days, mH2A2 KO MCF7 cells showed an increased response to E2 when compared with parental or control MCF7 cells, as measured by the tumorsphere assay (Fig. 5e). Such results suggest that the loss of the regulatory mBEs led to an overall increase of available ER regulatory regions.

mH2A2 is a negative regulator of BRD4

To gain insights into the global effect of the mH2A histone variants and specifically mH2A2 on enhancer regulation, we used a breast cancer cell line that lacks mH2A2 and expresses mH2A1 at a reduced level (MDA-MB-231L) (Fig. 4a). Over-expression of all mH2A variants and, particularly, mH2A2, lead to a decrease in the proliferative capacity of these cells (Fig. 6a). To test their oncogenic potential, we used an inducible system to over-express mH2A2 in MDA-MB-231L cells (Supp. Fig. 6a-c) and performed a tumor sphere assay. Induction of expression of mH2A2 lead to a significant decrease in the tumor sphere growth (Fig. 6b). To understand if enhancers were specifically affected by the ectopic expression of mH2A2, we performed ChIP-seq analysis of H3K4me1 and H3K27ac, the enhancer binding factor BRD4, and the histone methyltransferase p300, together with ATAC-seq (Fig. 6c). As the mH2A2 signal increased at CREs, BRD4 binding became reduced (Fig. 6c). In contrast, H3K4me1, H3K27ac, and p300 binding were not affected. This suggests a specific role for mH2A2 deposition around specific enhancers as an inhibitor of BRD4 binding (Supp. Fig. 6d). Since BRD4 function has been widely associated with the activation of transcription through its association with promoters and enhancers, we hypothesized that mBEs could negatively regulate enhancer activation through the eviction of BRD4 from chromatin. We next validated the loss of BRD4 in chromatin with over-expression of mH2A isoforms tagged with GFP in MDA-MB-231L and used GFP-H2A as a control (Fig. 6d). Loss of BRD4 followed the ectopic expression of mH2A variants as observed by Western blot of chromatin extracts, with the most robust effect seen in response to mH2A2 overexpression. Interestingly, BRD4 is also negatively associated with mBEs in MCF7 cells (Fig. 5a). Re-expression of mH2A2 in the mH2A2 KO MCF7 cell clones also showed the loss of BRD4 from the chromatin (Fig. 6e).

Loss of mH2A2 leads to increased cellular heterogeneity

Cellular heterogeneity can also be one of the mechanisms by which cancer cells increase metastatic potential, evade the immune system or resist to drug therapies, making it a central problem in medicine. Intratumoral

heterogeneity has become a major issue that can now be addressed with single-cell sequencing technologies. To gain insights into the ER-response elements involved in E2 stimulation in the presence and absence of mBEs, we mapped open chromatin regions at single-cell resolution in MCF7 cells after treatment with E2 for five days using single cell ATAC-seq (scATAC-seq). The main goal was to understand the impact of the loss of mH2A2 on ER (ESR1) motif accessibility and cellular heterogeneity. We obtained high-quality single-cell profiles derived from a MCF7 wild type and from a mH2A2 KO MCF7 clone (Supp. Fig. 7a-c). The control and mH2A2 KO scATAC-seq profiles were analyzed using UMAP projections and graph-based clustering (Fig. 7a). Notably, mH2A2 KO showed more defined separation of the clusters, indicating a higher level of heterogeneity. Similar assessment of heterogeneity can be observed by merging the two datasets and then performing the same graph-based clustering (Supp. Fig. 7d). Binding events from the ReMAP data base obtained in MCF7 cells with and without E2 exposure were then analyzed and compared to open chromatin regions (cut sites) in the two MCF7 clones (Fig. 7b). ESR1, GATA3, FOXA1 and BRD4 were among the TF and chromatin regulators that were significantly enriched in the absence of mH2A2. Notably, depletion of mH2A2 allowed for the enrichment of cut sites overlapping with mBEs, whereas the enrichment was less pronounced in Active Enhancers (Fig. 7c). This increased availability of open chromatin regions in the absence of mH2A2 was specific to enhancers, as the promoter regions showed a significantly decreased transposase accessibility in mH2A2 KO cells vs. wild-type cells ($p=0.00014$). Moreover, the co-accessibility of enhancer regions, used as a measure of interactions of regulatory elements, showed a clear increase in mH2A2 KO cells, both globally (Fig. 7d) and locally at genes associated with ER response, such as the ESR1 gene (Fig. 7e).

DISCUSSION

To understand the process by which tumors hijack regulatory elements to their benefit, we must define such regulatory elements in homeostasis and their potential role in defining cell identity and cellular heterogeneity. A better characterization of enhancer activity in cancer may reveal unknown transcriptional dependencies, novel pathways and altered enhancer states that may be valuable in designing new therapeutic approaches. In addition, the role of histone variants at enhancer elements remains poorly defined. Here we characterize a new class of regulatory elements that lack H3K27ac and are enriched with the mH2A histone variants. These mBEs are associated with transcriptional modules that reflect cell-specific functions. Further studies will be required to understand the role of mBE in cellular specification during development, however the fact that mice lacking both mH2A variants are viable suggests that mBEs function as a fine-tuning mechanism rather than indispensable regulators of normal development. It also suggests that loss of mH2A unleashes cellular plasticity in somatic cells, by contributing to age-related diseases such as neurodegenerative diseases and cancer. Studies to test this hypothesis in mouse models would help to address these important questions.

The biological parallels between reprogramming and cancer transformation led us to inquire whether mBEs could have a role in oncogenic activation. Here we demonstrate that mBEs identified in human mammary epithelial cells are associated with specific breast cancer subtypes when reactivated. In fact, the specific oncogenic programs that characterize various sub-types of breast cancer are in part encoded in such mBEs. At this point, it remains to be determined if the relationship between mBEs and tumor type-specific oncogenic programs is causal or a mere consequence of the loss of different mH2A isoforms in tumors. However, even the latter could be significant if these changes could be linked to variations in drug response or metastatic potential, and thus mH2A variants could be used as a biomarker.

To better understand how mH2A could function as a repressor of enhancer activity, we showed that the macro domain of mH2A isoforms is sufficient to promote inactivation of enhancers by making use of a chimeric dCas9-macro system to target a specific enhancer in ESCs and DFs during reprogramming. Such repression was comparable to the well characterized dCas9-KRAB system. These experiments not only provide insights into the functions of the domains that are required for mH2A-mediated repression, but also add an important new tool to the growing set of repressive systems available for experimental modulation of enhancers, alongside the repressive KRAB domain, EZH2 or DNMT domains[41], which have been shown to work in a context-dependent manner. In

addition, we demonstrated that re-expression of mH2A2 led to loss of chromatin bound BRD4, a bromodomain-containing reader of histone acetylation that binds active enhancers and promoters. Mass spectrometry studies have shown that bromodomain and extraterminal domain (BET) proteins do in fact interact with mH2A histone variants[42], consistent with a potential role for mH2A in evicting BRD4 from enhancer elements.

Finally, we showed with single-cell resolution that the loss of mH2A2 in MCF7 cells leads to increased cellular heterogeneity and more robust response to estrogen. This may occur in part due to enhancer deregulation and allows cancers to gain access to transcriptional programs that were previously silenced. For example, mH2A variants have been shown to play roles in EMT in breast cancer[37] but their role in cellular heterogeneity has been hard to define. Loss of mH2A variants in DCIS and low-grade tumors could potentially allow for metastatic programs to activate within a more heterogeneous population of cells. The same is true in the context of reprogramming, where a variety of different factors, including cellular heterogeneity, may lead to an increased efficiency of the process[35].

In closing, the role of mH2A in modulating enhancer activity reveals a novel role for mH2A variants, beyond the described associations with the histone marks H3K27me3 and H2BK12ac. Most mH2A studies have focused on deciphering their function near or at genes, including the role of the macro domains as readers of metabolic activity (mH2A1.1), in DNA repair in association with PARP, and in cancer associated features such as proliferation and invasion senescence. The presence of specific mH2A isoforms at enhancers may facilitate our understanding of how some of these processes exert transcriptional changes through enhancer regulation. It remains to be determined how different macro variants may confer unique functions at enhancers in specific cellular contexts.

REFERENCES

1. Levine, M. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**, R754-63 (2010).
2. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).
3. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-37 (2013).
4. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol Cell* **54**, 180-192 (2014).
5. Dorighi, K.M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**, 568-576 e4 (2017).
6. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-83 (2011).
7. Zentner, G.E., Tesar, P.J. & Scacheri, P.C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* **21**, 1273-83 (2011).
8. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* **44**, 148-56 (2012).
9. Koenecke, N., Johnston, J., He, Q., Meier, S. & Zeitlinger, J. Drosophila poised enhancers are generated during tissue patterning with the help of repression. *Genome Res* **27**, 64-74 (2017).
10. Cai, Y. *et al.* H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat Commun* **12**, 719 (2021).
11. Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. *Nat Struct Mol Biol* **21**, 210-9 (2014).
12. Karnuta, J.M. & Scacheri, P.C. Enhancers: bridging the gap between gene control and human disease. *Hum Mol Genet* **27**, R219-R227 (2018).
13. Vardabasso, C. *et al.* Histone variants: emerging players in cancer biology. *Cell Mol Life Sci* **71**, 379-404 (2014).
14. Pehrson, J.R. & Fried, V.A. MacroH2A, a core histone containing a large nonhistone region. *Science* **257**, 1398-400 (1992).
15. Costanzi, C., Stein, P., Worrad, D.M., Schultz, R.M. & Pehrson, J.R. Histone macroH2A1 is concentrated in the inactive X chromosome of female preimplantation mouse embryos. *Development* **127**, 2283-9 (2000).
16. Douet, J. *et al.* MacroH2A histone variants maintain nuclear organization and heterochromatin architecture. *J Cell Sci* **130**, 1570-1582 (2017).
17. Chang, E.Y. *et al.* MacroH2A allows ATP-dependent chromatin remodeling by SWI/SNF and ACF complexes but specifically reduces recruitment of SWI/SNF. *Biochemistry* **47**, 13726-32 (2008).
18. Changolkar, L.N. *et al.* Developmental changes in histone macroH2A1-mediated gene regulation. *Mol Cell Biol* **27**, 2758-64 (2007).
19. Gamble, M.J., Frizzell, K.M., Yang, C., Krishnakumar, R. & Kraus, W.L. The histone variant macroH2A1 marks repressed autosomal chromatin, but protects a subset of its target genes from silencing. *Genes Dev* **24**, 21-32 (2010).
20. Changolkar, L.N. *et al.* Genome-wide distribution of macroH2A1 histone variants in mouse liver chromatin. *Mol Cell Biol* **30**, 5473-83 (2010).
21. Chakravarthy, S. *et al.* Structural characterization of the histone variant macroH2A. *Mol Cell Biol* **25**, 7616-24 (2005).
22. Sun, Z. *et al.* Transcription-associated histone pruning demarcates macroH2A chromatin domains. *Nat Struct Mol Biol* **25**, 958-970 (2018).
23. Gaspar-Maia, A. *et al.* MacroH2A histone variants act as a barrier upon reprogramming towards pluripotency. *Nat Commun* **4**, 1565 (2013).
24. Chen, H. *et al.* MacroH2A1.1 and PARP-1 cooperate to regulate transcription by promoting CBP-mediated H2B acetylation. *Nat Struct Mol Biol* **21**, 981-9 (2014).
25. Barrero, M.J. *et al.* Macrohistone variants preserve cell identity by preventing the gain of H3K4me2 during reprogramming to pluripotency. *Cell Rep* **3**, 1005-11 (2013).
26. Sporn, J.C. *et al.* Histone macroH2A isoforms predict the risk of lung cancer recurrence. *Oncogene* **28**, 3423-8 (2009).
27. Kapoor, A. *et al.* The histone variant macroH2A suppresses melanoma progression through regulation of CDK8. *Nature* **468**, 1105-9 (2010).
28. Novikov, L. *et al.* QKI-mediated alternative splicing of the histone variant MacroH2A1 regulates cancer cell proliferation. *Mol Cell Biol* **31**, 4244-55 (2011).
29. Monteiro, F.L. *et al.* Expression and functionality of histone H2A variants in cancer. *Oncotarget* **5**, 3428-43 (2014).

30. Ghiraldini, F.G., Filipescu, D. & Bernstein, E. Solid tumours hijack the histone variant network. *Nat Rev Cancer* (2021).
31. Consortium, E.P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
32. Vardabasso, C., Hake, S.B. & Bernstein, E. Histone variant H2A.Z.2: A novel driver of melanoma progression. *Mol Cell Oncol* **3**, e1073417 (2016).
33. Chen, H. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173**, 386-399 e12 (2018).
34. Pehrson, J.R., Changolkar, L.N., Costanzi, C. & Leu, N.A. Mice without macroH2A histone variants. *Mol Cell Biol* **34**, 4523-33 (2014).
35. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442-459 e20 (2017).
36. Schaniel, C. *et al.* Smarcc1/Baf155 couples self-renewal gene repression with changes in chromatin structure in mouse embryonic stem cells. *Stem Cells* **27**, 2979-91 (2009).
37. Hodge, D.Q., Cui, J., Gamble, M.J. & Guo, W. Histone Variant MacroH2A1 Plays an Isoform-Specific Role in Suppressing Epithelial-Mesenchymal Transition. *Sci Rep* **8**, 841 (2018).
38. Xi, Y. *et al.* Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. *BMC Genomics* **19**, 150 (2018).
39. Franco, H.L. *et al.* Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. *Genome Res* **28**, 159-170 (2018).
40. Theodorou, V., Stark, R., Menon, S. & Carroll, J.S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res* **23**, 12-22 (2013).
41. O'Geen, H. *et al.* Ezh2-dCas9 and KRAB-dCas9 enable engineering of epigenetic memory in a context-dependent manner. *Epigenetics Chromatin* **12**, 26 (2019).
42. Lambert, J.P. *et al.* Interactome Rewiring Following Pharmacological Targeting of BET Bromodomains. *Mol Cell* **73**, 621-638 e17 (2019).

ACKNOWLEDGMENTS

This work was supported by the Mayo Clinic Center for Individualized Medicine and the Department of Experimental Pathology and Lab Medicine, the New York Stem Cell Foundation Druckenmiller Fellowship (NYSCF-D-F41), Postdoctoral Fellowship in Breast Cancer Research Program, United States Department of Defense (BC100975) and the Mayo Clinic National Cancer Institute-designated Comprehensive Cancer Center Ovarian SPORE grant (Career Development Award P50 CA136393) to A.G.M.; NIH R01 CA154683 and NYSTEM IIRP C029573 to E.B.. Partial support for this work was provided by the grant CA236612 awarded to A.R.. The Galaxy server used for some calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).

AUTHOR CONTRIBUTIONS

A.G.M. conceived and designed the study with the help of W.M.I., D.H. and E.B. A.M., J.K., M.B., A.G.M., L.S., C-Y.C., S.C., D.H., J.Z. and J.H.L. performed experiments. W.M.I., S.S. L.S., J.M.W. analyzed the data. L.F.D performed IHC and C.S.N. scored tissue samples. N.D and A.R. provided the microfluidic device culturing systems

and technical advice. A.G.M. wrote the manuscript with the help of W.M.I, A.M., D.H, T.O and E.B. All authors critically revised and approved the final version of the manuscript.

COMPETING INTERESTS

The authors have no conflicts of interest to declare.

METHODS

Cell culture

Normal Human Melanocytes (NHM) were cultures in Dermal Cell Basal Medium (ATCC) with the addition of 5 μg/mL Insulin, 50 μg/mL Ascorbic Acid, 6 mM L-Glutamine, 1.0 μM Epinephrine, 1.5 mM Calcium Chloride, Peptide Growth Factor and M8 Supplement. Dermal fibroblasts (DFs) were isolated from neonatal mice and iPS reprogramming was performed as described[1]. MCF-7, DFs and MDA-MB-231L cells were grown in DMEM (Gibco) with 4.5 g/L D-glucose, 110 mg/L sodium pyruvate, 10% FBS and 1% Penicillin/Streptomycin (Hyclone). HMEC cells were grown in complete Mammary Epithelial Cell Growth Media. For estradiol (E2) treatment MCF-7 cells were grown for 5 days in modified DMEM without phenol-red (Hyclone) with 4.5 g/L D-glucose, 4.0 L-glutamine, 10% charcoal-dextran–stripped FBS, 1% Penicillin/Streptomycin and 1nM of 17-β-estradiol or EtOH. For growth curves, 1000 cells stably expressing H2A-GFP were plated in each well of a 96 well plate and their growth was followed for 14 days in Incucyte (Sartorius), with acquisition every 12 hours.

Constructs

The 4 transcription factors (Oct4, Sox2, Klf4 and Myc) used for iPS reprogramming are encoded in a polycystronic lentiviral vector (Stemcca, kindly provided by Gustavo Mostoslavsky, Boston University). LentiCRISPR v2 (Addgene plasmid # 52961) and lentiCas9-Blast (Addgene Plasmid #52962) were a gift from Feng Zhang[2]. To generate CRISPR clones in MCF7 cells, sgRNAs targeting H2AFY2 were selected using CRISPR Design Tool (http://crispr.mit.edu) and cloned using BsmBI enzyme (NEB). SgRNAs targeting the H2AFY2 locus were: 1-GTTCAGCTAGGGCAGGTGTC, 2- GTTCAAGTACCGGATCAGCG, 3- GGCGGCAGTCATTGAGTACC, 4-GGATAGCCCCGAGACACATC. Human H2A and macroH2A isoforms were GFP-tagged and subcloned into pLKO.1 plasmid for lentiviral production. Tagged macroH2A2 isoform was subcloned into lentiviral vectors pLVX

(Clontech) for dox-inducible expression together with pLVX-Tet3G-Neo. pHAGE EF1α dCas9-KRAB was a gift from Rene Maehr & Scot Wolfe (Addgene plasmid # 50919). pHAGE-EF1-dCas9 plasmids were generated by cloning macro domains from mH2A1.1, mH2A1.2 and mH2A2 in replacement of the KRAB domain. SgRNAs targeting the Nanog-GFP locus were: Control (GACGGGTCTCCAGTAGTTCG), Enhancer (GACAGGAATGGGGGTTGGGGA), GFP-1(GGGCGAGGAGCTGTTCACCG), GFP-2 (GTAGGTCAGGGTGGTCACGA). SgRNAs were cloned using BbsI enzyme (NEB) into pLKO-GFP-H2A or pLKO-mCh-H2A. The packaging plasmids for the preparation of lentiviral particles were psPAX2 and pMD2G.

Lentiviral production

Transgenic cell lines with stable integration of constructs were generated by lentiviral transduction followed by selection in 2 µg/ml puromycin (Millipore) or 5 µg/ml blasticidin (InvivoGen) or 400 µg/ml neomycin (Millipore). Lentiviral particles used in this study were produced in house as previously described[3]. Briefly, lentiviral vectors containing constructs of interest were transfected into 293T cells together with packaging plasmids using calcium phosphate methods. Media containing lentiviral particles was collected at 36, 48 and 60h post-transfection, filtered and concentrated by ultracentrifugation at 25,000 rpm for 90 min.

MCF7 CRISPR/Cas9 mH2A2 knockout clones

LentiCRISPR v2 (Addgene plasmid # 52961) was used to generate CRISPR clones in MCF7 cells with sgRNAs targeting H2AFY2. After transduction, puromycin selection was performed and 1000 cells were plated in a 10cm dish. After 3 or 4 weeks, clones were identified and selected from the empty vector control or H2AFY2 targeting. Following expansion, clones were identified by western blot.

Flow Cytometry analysis

NG4 or MDA-MB-231 cells were trypsinized, washed in PBS, strained with a 100 μm filter, and resuspended at 1 X10^7 cells/ml in FACS buffer (DPBS and 2%BSA) at 4°C. GFP and mCherry fluorescence was analyzed by FACS on a LSRII machine and data was analyzed with FlowJo.

Microfluidic devices and tumorspheres

Development of spheroids in the presence or absence of E2 was achieved by 3D cultures inside microfluidic devices as previously described[4]. Briefly, microfluidic devices were fabricated using standard soft-lithography using mixture of 10:1 weight ratio of polydimethylsiloxane (PDMS) base to curing agent (Sylgard 184 Silicone Elastomer Kit, Dow Corning). 5 days after E2 treatment as described above, 4×10^5 cells were deposited in the inlet of the device and allowed to flow through the culture chamber until cells filled the bottom of the microwells. Cells were then kept in culture in the microfluidic devices for 7 days at 37 °C in the presence of E2 or EtOH, changing media every 24 h. To track the growth of the spheroids bright-field images were acquired at days 1,3, and 7 days after seeding. Spheroids sizes were assessed using ImageJ to estimate the area at each time point, and then normalized to the area at day 1 to allow growth comparisons between wells.

ChIP-seq

Approximately 1 million cells for the histone variants and 3 million cells for histone modifications and regulators, from each sample, were used for input for native chromatin immunoprecipitation (nChIP). Cells were lysed on ice for 20 minutes in lysis buffer containing 0.1% Triton X-100, 0.1% deoxycholate, and protease inhibitor. Extracted chromatin was digested with 90U of MNase enzyme (New England Biolabs) for 6 minutes at 25°C. The reaction was quenched with 250 μM of EDTA post-digestion. A mix of 1.0% Triton X-100 and 1.0% deoxycholate was added to the digested samples and incubated on ice for 20 min. Digested chromatin was pooled and pre-cleared in IP buffer (20 mM Tris-HCl [pH 7.5], 2 mM EDTA, 150 mM NaCl, 0.1% Triton X-100, and 0.1% deoxycholate) plus protease inhibitors with pre-washed Protein A/G Dynabeads (Thermo Fisher Scientific, Waltham, United States) at 4°C for 1.5 hours. Supernatants were removed from the beads and transferred to a 96-well plate containing the antibody-bead complex. Following an overnight 4°C incubation, samples were washed twice with low salt buffer (20 mM Tris-HCl [pH 8.0], 0.1% SDS, 1.0% Triton X-100, 2 mM EDTA, and 150 mM NaCl) and twice with high salt buffer (20 mM Tris-HCl [pH 8.0], 0.1% SDS, 1.0% Triton X-100, 2 mM EDTA, and 500 mM NaCl). DNA-antibody complexes were eluted in elution buffer (100 mM $NaHCO_3$, 1.0% SDS), incubated at 65°C for 90 minutes. Protein digestion was performed on the eluted DNA samples at 50°C for 30 minutes using protease mix (QIAGEN, Venlo, Netherlands). ChIP DNA was purified using Sera-Mag beads (Thermo Fisher Scientific) with 30% PEG before library construction. Size distribution and level of amplification were determined by analysis using Agilent bioanalyzer or

Fragment Analyzer. Libraries were prepared by following a modified Illumina paired-end protocol and sequenced on an Illumina HiSeq 2500 to a median depth of ~25 million (H3K4me1 and H3K4me3) or ~50 million reads (H3K27me3 and Input) or single end protocol for histone variants to a median depth of ~80 million. Reads were aligned to the hg19 reference genome using bowtie2.[5]

Immunohistochemistry

Specimens were obtained from Mount Sinai School of Medicine and considered non-human subject research. Tissue Microarray slides were provided by the NCI cancer Diagnosis program (CDP). Other investigators may have received slides from the same blocks. IHC was performed as described before[6]. In brief, 5-µm sections from formalin-fixed paraffin-embedded specimens were deparaffinized, incubated for antigen retrieval with Vector Citrate-Based Antigen Unmasking Solution (Vector Laboratories) in microwave for 10 min, and then exposed to 0.3% hydrogen peroxide to block endogenous peroxidase activity. After blocking with Vector Normal Horse Serum (2.5%) for 20 min, sections were incubated at 4 °C overnight with mH2A2 (1:350–1:500) prepared in 0.1% BSA. Slides were subsequently developed using Vector imPRESS Universal Kits anti-mouse/rabbit Ig or anti-goat Ig (Vector Laboratories), Vector DAB Peroxidase Substrate Kit as the chromagen (Vector Laboratories) and Harris Hematoxylin (Sigma) for counterstaining. Slides were then sealed and mounted with Permount (Sigma) and randomized for subsequent blinded review.

Chromatin isolation and Western blot

Chromatin fractionation was performed as described[7]. Briefly, cells were washed in PBS and resuspended in 1mL buffer A (10mM HEPES pH 7.9, 10mM KCl, 1.5mM MgCl2, 0.34M sucrose, 10% glycerol, 1mM DTT and 1 X protease inhibitor cocktail). Triton X-100 was added to 0.1% and the cells are incubated on ice for 10min. Nuclei were collected by centrifugation at 4000rpm at 4°C. The supernatant was taken as the cytosolic fraction. Nuclei were washed once with buffer A and then lysed for 30min in 'No Salt' buffer (3mM EDTA, 0.2mM EGTA, 1mM DTT and 1 X protease inhibitor cocktail) on ice. Chromatin was pelleted by centrifugation at 4000rpm at 4°C and supernatant was enriched in soluble nuclear proteins. For western blotting, equal amounts of isolated chromatin, estimated by amido black (Sigma) staining, were run on an 8%, 15% or 4-15% SDS-PAGE gel, then transferred to PVDF membranes (Millipore). After blocking with Intercept® (PBS) Blocking Buffer (LI-COR) for 1 h at room

temperature, the membrane was incubated with primary antibodies at 4 °C overnight. The membrane was then washed three times with PBST for 10 min and then incubated for 1 h at room temperature with appropriate secondary antibodies conjugated with Dylight (Invitrogen). After extensive washing, fluorescent detection was performed using the Odyssey® Fc imaging system (Li-Cor Biosciences). Alternatively, immunoblotting was performed as described[7].

Antibodies

The following antibodies were used in this study: H3K27me3 (Millipore 07-449), H3K27ac (Abcam ab4729, Western blot), H3K27ac (Cell signaling, 8173BC, ChIP), H3K4me1 (homemade EDL), P300 (Clone NM11, Active Motif 61401), BRD4 (Bethyl A301-985A50, ChIP), BRD4 (Abcam ab128874, Western blot), Cas9 (Millipore MAC133), H2A.Z (Abcam ab150402), mH2A1 (Abcam ab37264, ChIP), mH2A1 (Millipore 07-219, Western blot), mH2A2 (Homemade, Bernstein Lab[6]), H3 (Abcam Ab1791), GFP (Roche 11814460001), Beta-Actin (Sigma, A5441), Flag (Sigma, F1804), Mouse IgG – DyLight 680 (Invitrogen SA5-10170), Rabbit IgG DyLight 800 (Invitrogen SA5-10044).

cDNA generation and RT-qPCR

Total RNA was extracted using RNeasy Mini Kit (Qiagen) according to manufacturer's protocol. cDNA was generated using First-Strand cDNA Synthesis System (ORIGENE). qPCR was performed using FastStart Universal SYBR Green Master Mix (Rox) (Roche).

ATAC-seq

DNA for ATAC-seq was prepared from 50,000 cells following the OMNI-ATAC procedures as described by *Corces et al.* with modifications using the Nextera kit (Illumina). The cells were lysed for 3 minutes on ice and transposed for 30 minutes at 37ºC following clean-up. The DNA libraries were prepared with 5-10 cycles of PCR amplification with the NEB High Fidelity Master Mix (New England Biolabs, Ipswich, United States). Clean-up was done using the Zymo DNA Clean and Concentrator kit (Zymo Research, Irvine, United States) and followed with AMPure XP (Beckman Coulter, Brea, United States) bead clean-up to remove primer dimers and under-digested chromatin.

Sequencing was performed on an Illumina HiSeq 4000 to a depth of ~30 million reads per sample. Reads were aligned to hg19 using bowtie2.

Data analysis

Aligned reads from the different immunoprecipitation and accessibility sequencing experiments were sorted and indexed using samtools[11] and peaks were called using MACS2[12] with input controls (except ATAC-seq where no input was used). The UCSC Genome Browser and deepTools were used for signal visualization.[10,13] For visualization purposes averaged, input-corrected signal tracks were created (except ATAC-seq where no input was used) using samtools merge of the corresponding aligned bam files and deepTools bamCoverage / bigwigCompare.

For classification of the cell-specific cis-regulatory elements (CRE), the following procedure was adopted for each cell-type. Firstly, the peaks that were common in both ATAC-seq and H3K4me1 were obtained using bedtools[14] intersect. Then peaks that overlap ENCODE blacklist[15] regions – regions that are known to have anomalous, unstructured or high signal[15] were removed. From this list, only peaks whose center overlapped with at least one ENCODE candidate cis-regulatory element (cCRE)[16] were chosen for downstream analysis. The signal intensity for each peak was calculated as the summation of the input-corrected signals (raw signals for ATAC-seq) over a window of 2000 base pairs around the center of each peak. The signal intensities for each peak were then normalized by the total intensity, then multiplied by a scale factor of 10,000 and then log-transformed. The z-scores of these log-normalized intensities of the seven signals – H3K4me1, H3K4me3, H3K27ac, H3K27me3, H2A.Z, macroH2A1 and macroH2A2 (and CTCF, for those cell-types where available), were fed as input to K-means clustering algorithm with $k = 5$ to classify the peaks set into 5 clusters. These clusters were then named as Active, APL, ATAC-only, Inactive and mBE, based on the signal(s) that identify each cluster.

Data are presented as median (range) unless denoted otherwise. Medians were the preferred measure of central tendency and non-parametric hypothesis tests were used for comparisons unless stated otherwise. Continuous variables were compared using the Mann-Whitney-U test, categorical variable using Fischer's exact test. Hidden Markov modeling (ChromHMM)[17] was used to discover and characterize the presence of chromatin states in promoter regions (TSS 2000bp). Genes associated with the CRE were identified using GeneHancer[18] and Genomic Regions Enrichment of Annotations Tool (GREAT). Candidate transcription factors potentially binding these regions were identified by querying the ReMap 2020 database. Genome annotation of the CRE were performed using

Hypergeometric Optimization of Motif EnRichment (HOMER)[19]. Super-Enhancer prediction was performed using LILY[20]. Expression quantification for RNA-seq data was done by read alignment using STAR[21], counting using bedtools multicov, normalization and RPKM calculations using edgeR[22].

Analysis of Odds Ratios

To analyze the enrichment of TF binding sites, we intersected the locations of the TF binding sites from REMAP (using the 2020 data in hg38 converted to hg19 with liftOver) with our enhancer sites using bedtools. From this list of all locations coded by both TF binding site and enhancer class, we performed a Fisher's Exact Test with the TFs and the enhancer classes as the categories. The result is, for each enhancer class/TF pair, a mathematical comparison of the actual occurrence of that pairing versus the statistically expected occurrence. We have displayed these results here as a heat map of the log (base 2) of odds ratio so that more positive values (redder colors) are enriched and more negative values (bluer colors) are depleted.

Single cell ATAC-seq

Approximately 100,000 viable cells per sample were subjected to transposase assays (exposing buffered nuclei to Tn5 transposase) before proceeding to single-cell partitioning into gel beads in emulsion, barcoding, library construction, and sequencing following established 10X Genomics protocols. The target cell recovery was approximately 8000 cells per sample. For details on the 10X Genomics Chromium platform including demonstrated protocols on sample preparation, library construction, instrument settings, and sequencing parameters please see the manufacturer's resources (https://support.10xgenomics.com/single-cell-atac). Genomic libraries were sequenced on an HiSeq 4000 (Illumina) before demultiplexing, alignment to the reference genome, and post-alignment quality control. The 10X Genomics Cell Ranger ATAC 1.2.0 software[23] was used for demultiplexing, alignment of the reads to the hg19 reference genome, filtering and quality control, counting of barcodes and unique molecular identifiers, identification of transposase cut sites, detection of accessible chromatin peaks, count matrix generation for peaks and transcription factors, clustering, and differential accessibility analysis.

Quality control, integration, normalization, scaling, feature selection, clustering, and dimensionality reduction of the scATAC-seq data was performed using Signac[24]. The *cis*-regulatory interactions and co-accessibility scores were predicted using Cicero[25]. The MCF7 CRE were used as the peaks set for the UMAP cluster analysis while the peaks

called by Cell Ranger ATAC were used as the peaks set for the Cicero co-accessibility analysis for need of better resolution.

For the Effect Size Bubble Plot, we computed Cohen's Effect Size comparing the means of the distributions of the number of TF binding sites per cell in the wild-type and knock-out populations. The calculation is done on TFs binding sites unique to the Estrogen line, to the control line, and to binding sites in both. *p*-values are computed using the Mann-Whitney *U* test with the null-hypothesis that the medians in the two distributions are the same. This test was used because the numbers of TF binding sites are not normally distributed so that parametric tests do not apply.

Availability of data and materials

The ChIP-seq and ATAC-seq datasets generated and analyzed in this study have been deposited into the NCBI Gene Expression Omnibus (GEO) data base (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE171599 (reviewers access token: qzshuuqwrtqlhip).

FIGURE 1

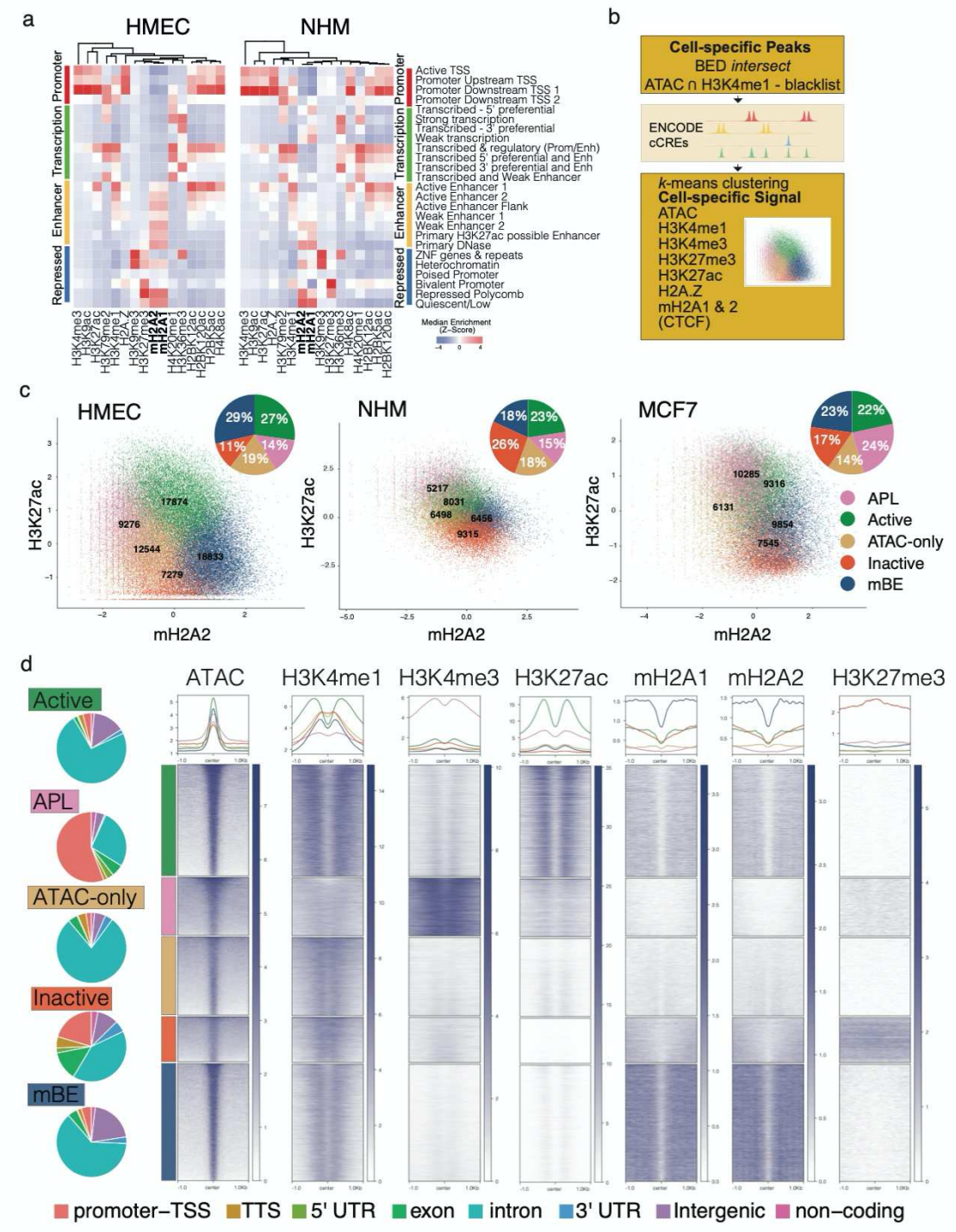

Figure 1. Characterization of macro-Bound Enhancers. a) Heatmap of the ENCODE chromatin states from HMEC and NHM with histone marks including the histone variants macroH2A1, macroH2A2. b) Algorithm to classify cell-specific *cis*-regulatory elements (CRE). c) Scatter plot showing the five classes assigned to the regulatory elements after K-means clustering in three cell types (HMEC – Human Mammary Epithelial Cells, NHM – Normal

Human Melanocytes, and MCF7 – breast cancer cell line), displayed along the Z scores for the log-normalized ChIP-seq signal of mH2A2 and H3K27ac. Pie charts show the proportions of CRE classes in each cell type. d-e) Characterization of the regulatory classes. Pie chart with genomic annotations for each regulatory class annotated using Homer (left). Signal profile (top) and heatmap (bottom) of enrichment around open chromatin regions (defined by ATAC-seq) of the five regulatory classes in HMEC (right). Enrichment of histone marks H3K4me1, H3K4me3, H3K27ac, H3K27me3 and histone variants macroH2A1 and macroH2A2 was analyzed by ChIP-seq.

FIGURE 2

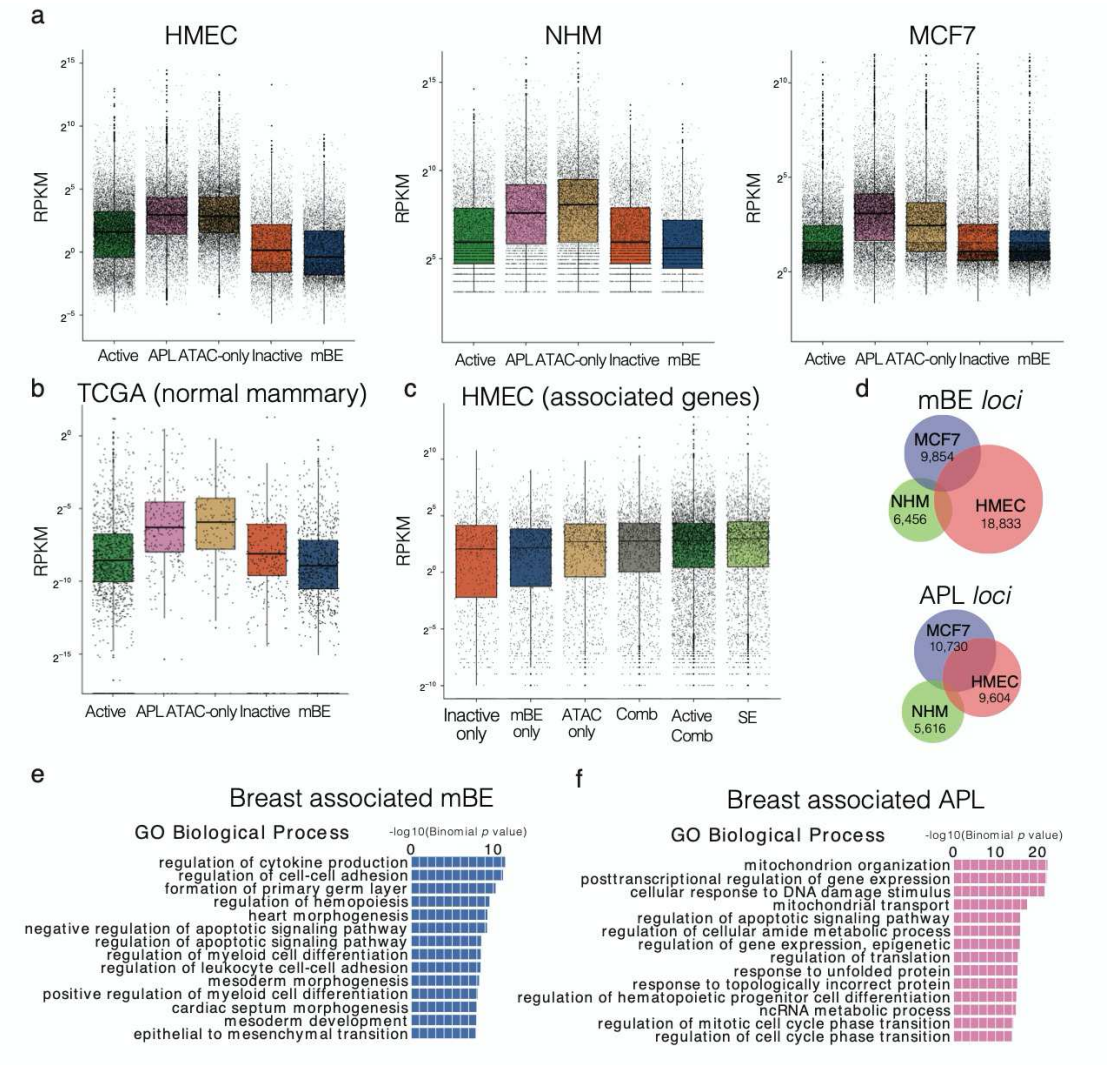

Figure 2. MacroH2A regulates enhancer activity. a) Expression levels at the CREs for HMEC, NHM and MCF7 by RNA-seq (logRPKM). b) Expression levels by RNA-seq from normal breast samples from TCGA[33] using the annotated CREs in HMEC c) Expression levels in HMEC by RNA-seq of the genes directly associated with the CREs from HMEC (associations obtained from GeneHancer v4.4) and super-enhancers (annotated using LILY). The expression data is represented as boxplots where the middle line represents the median, the lower and upper edges of the rectangle represent the first and third quartiles and the lower and upper whiskers represent the interquartile range (IQR) x 1.5. Outliers beyond the end of the whiskers are plotted individually. d) Intersection of mBE and APL (H3K4me3) CRE loci between the three cell lines. e, f) Gene Ontology Biological Process terms of genes associated with mBE and APL (H3K4me3) common in HMEC and MCF7, respectively. Enriched terms ranked by Binomial p value and presented on a -log10 scale.

FIGURE 3

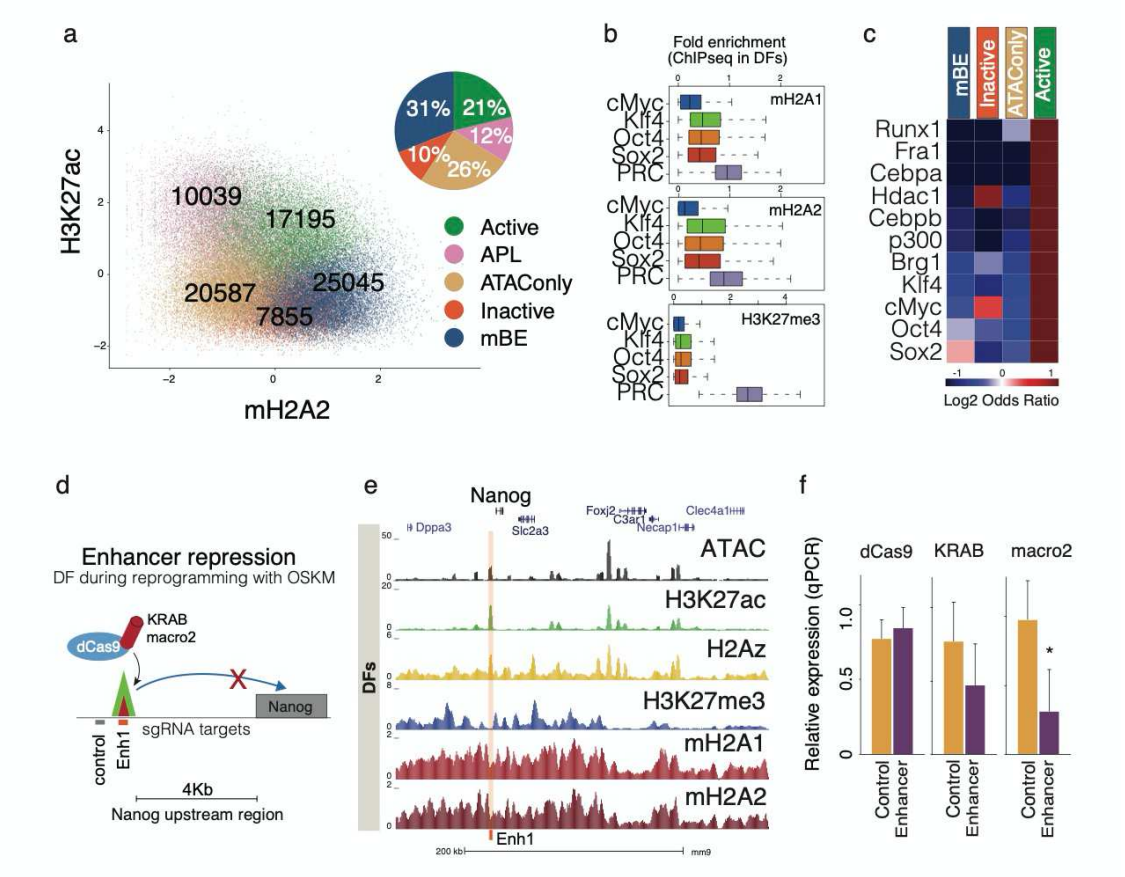

Figure 3. MacroH2A is a negative modulator of enhancer activity. a) Scatter plot upon K-means clustering of regulatory elements in mouse dermal fibroblasts (DFs), displayed along the Z scores for log-normalized ChIP-seq signals of mH2A2 and H3K27ac. Pie charts show proportions of CRE classes. b) Boxplot graph of ChIP-seq signal (fold enrichment) from mH2A1, mH2A2 and H3K27me3 in DFs at OSKM binding sites and PRC2 sites for comparison (48hrs after OSKM induction, binding sites data collected from[35]). The middle line represents the median, the lower and upper edges of the rectangle represent the first and third quartiles and the lower and upper whiskers represent the interquartile range (IQR) x 1.5. c) Odds ratio of enrichment of TF and chromatin binding factors after 48h of OSKM expression[35] at enhancer elements of DF cells, ranked by enrichment in mBE (Fisher Exact test). d) Schematic of enhancer targeting in DF during reprogramming with OSKM using dCas9-KRAB and dCas9-macro2 and sgRNAs complementary to regions around the enhancer site. e) UCSC genome browser snapshot of the Klf4 binding site upstream of the Nanog TSS in DF with open chromatin (ATAC-seq), H3K27me3, H3K27ac, H2A.Z, mH2A1 and mH2A2 data. f) *Nanog* relative expression after 96h of OSKM infection in DFs with dCas9, dCas9-KRAB or dCas9-macro2 with sgRNAs targeting the enhancer site upstream of Nanog or control. Unpaired (two tailed) student's t-test: * p value= 0.02.

FIGURE 4

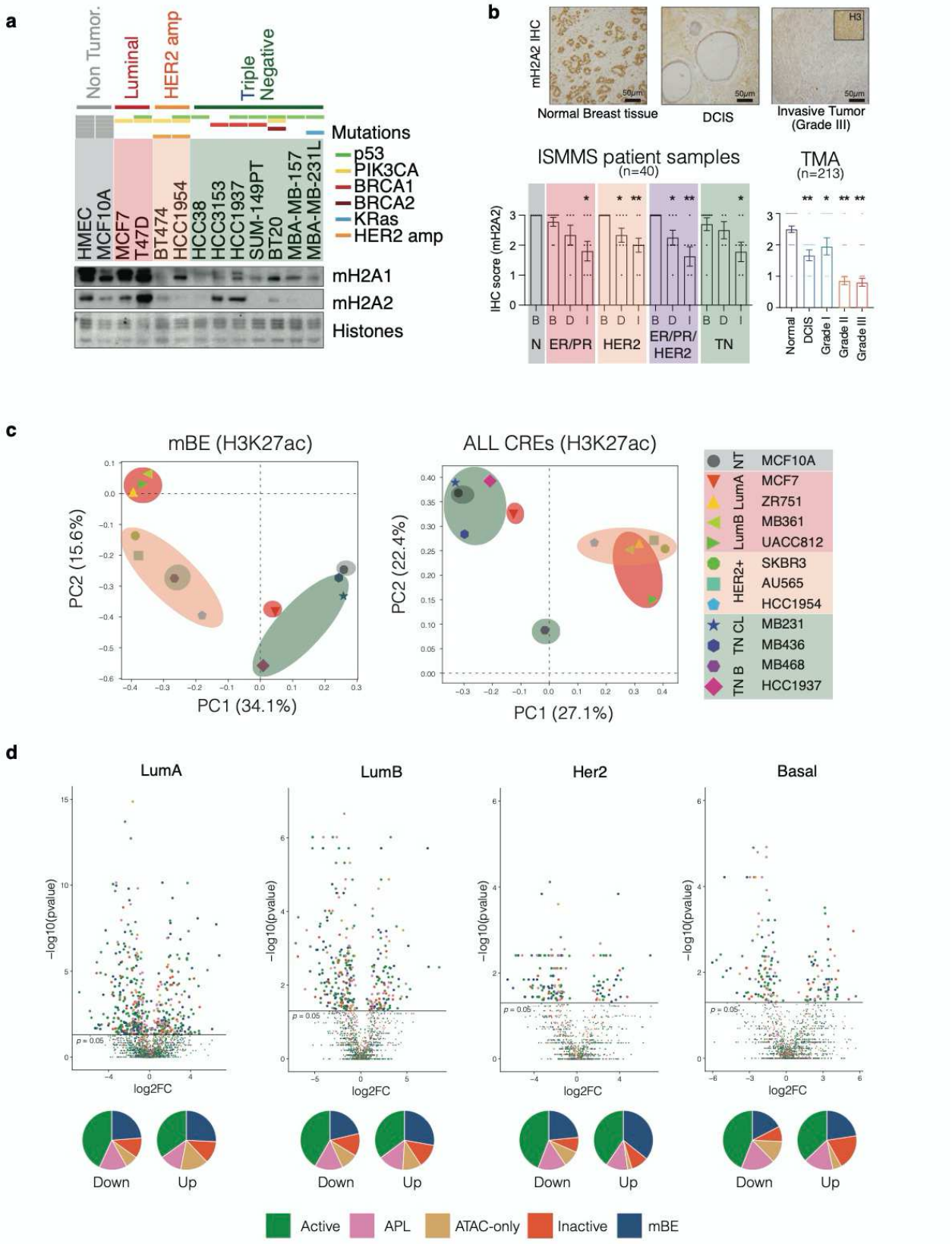
a
Non Tumor.
Luminal
HER2 amp
Triple Negative
Mutations
p53
PIK3CA
BRCA1
BRCA2
KRas
HER2 amp
HMEC
MCF10A
MCF7
T47D
BT474
HCC1954
HCC38
HCC3153
HCC1937
SUM-149PT
BT20
MBA-MB-157
MBA-MB-231L
mH2A1
mH2A2
Histones
b
mH2A2 IHC
Normal Breast tissue
DCIS
Invasive Tumor (Grade III)
ISMMS patient samples (n=40)
TMA (n=213)
IHC socre (mH2A2)
N
ER/PR
HER2
ER/PR/ HER2
TN
Normal
DCIS
Grade I
Grade II
Grade III
c
mBE (H3K27ac)
ALL CREs (H3K27ac)
PC1 (34.1%)
PC2 (15.6%)
PC1 (27.1%)
PC2 (22.4%)
NT
LumA
LumB
HER2+
TN CL
TN B
MCF10A
MCF7
ZR751
MB361
UACC812
SKBR3
AU565
HCC1954
MB231
MB436
MB468
HCC1937
d
LumA
LumB
Her2
Basal
−log10(pvalue)
log2FC
p = 0.05
Down
Up
Active
APL
ATAC-only
Inactive
mBE

Figure 4. Reactivation of macro-Bound Enhancers associates with oncogenic programs. a) Immunoblot of chromatin extracts were probed for mH2A1 and mH2A2 across a panel of breast cancer cell lines including the three different major sub-types (estrogen and progesterone positive, HER2 positive and triple negative), and non-tumorigenic cells (human mammary epithelial cells, HMEC; and immortalized mammary cells MCF10A). Mutational status defined above. Amido Black of core histones used as loading control. b) Immunohistochemistry (IHC) from normal breast tissue, ductal carcinoma in situ (DCIS) and grade III invasive tumors for mH2A2. Histone H3 IHC was used as a control (top right). Quantification of mH2A2 scoring for TMA and ISMMS patient samples according to the tumor grade or sub-class (below). B – Benign tissue, D – DCIS, I – Invasive. Column bar represents mean and SE. Unpaired (two tailed) student's t-test *p<0,05 ** p<0,005. c) Principal component analysis of H3K27ac ChIP-seq signal in HMEC macro-Bound enhancers (left) and all CREs (right) in 12 breast cancer cell lines from[38]. d) Volcano plot of differential enhancer expression in paired normal vs. tumor patient samples in breast cancer sub-types from RNA-seq (data collected from TCGA as described in ref.[33]), including the proportion of enhancers in each enhancer class that are significantly differentially expressed, down-regulated (down) or up-regulated (up), p-value < 0.05.

FIGURE 5

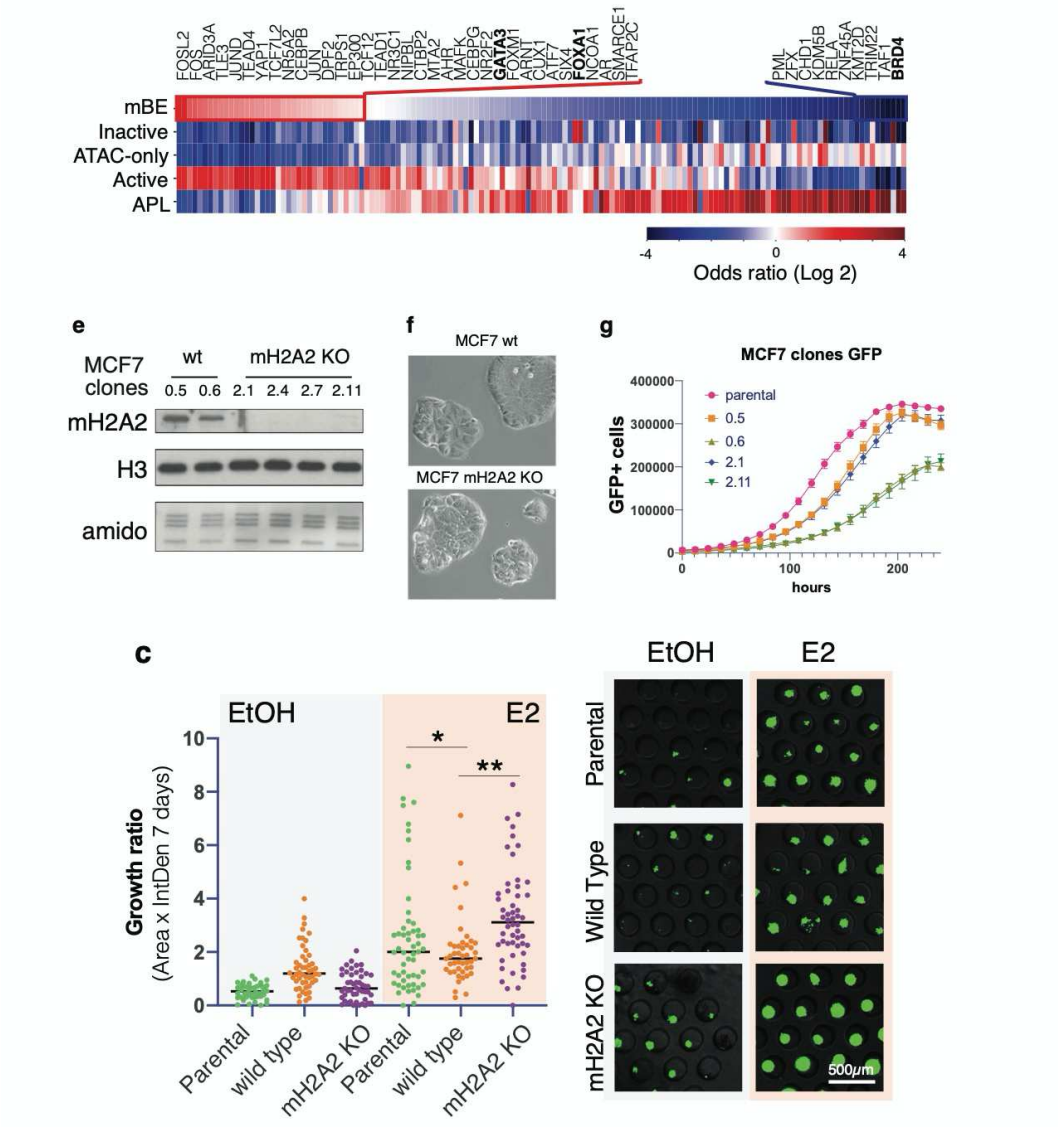

Figure 5. mH2A2 is a negative regulator of estrogen targets. a) Odds ratio of enrichment of TF and chromatin binding factors (ReMAP data for MCF7 cells) at enhancer elements of MCF7 cells, ranked by enrichment in mBEs (Fisher Exact test). b) Immunoblots for mH2A2 from chromatin extracts in MCF7 clones. H3 and histones (amido black) used as loading controls. c) Brightfield images of wild type and mH2A2 KO clones. d) Proliferation of MCF7 clones (wild type and mH2A2 *KO*) transduced with H2A-GFP and analyzed by number of GFP cells using Incucyte. Data represented are mean with SE (n=3). e) Growth ratio of MCF7 3D spheroids after treatment with EtOH and E2 (Estradiol) in microwells after 7 days. Scatter plot of area factored with GFP Intensity Density in individual spheroids. Horizontal bars signify mean values (left). Representative images of MCF7 spheroids (right). Scale bar, 500µm. Unpaired (two tailed) student's t-test *p<0,05 ** p<0,005.

FIGURE 6

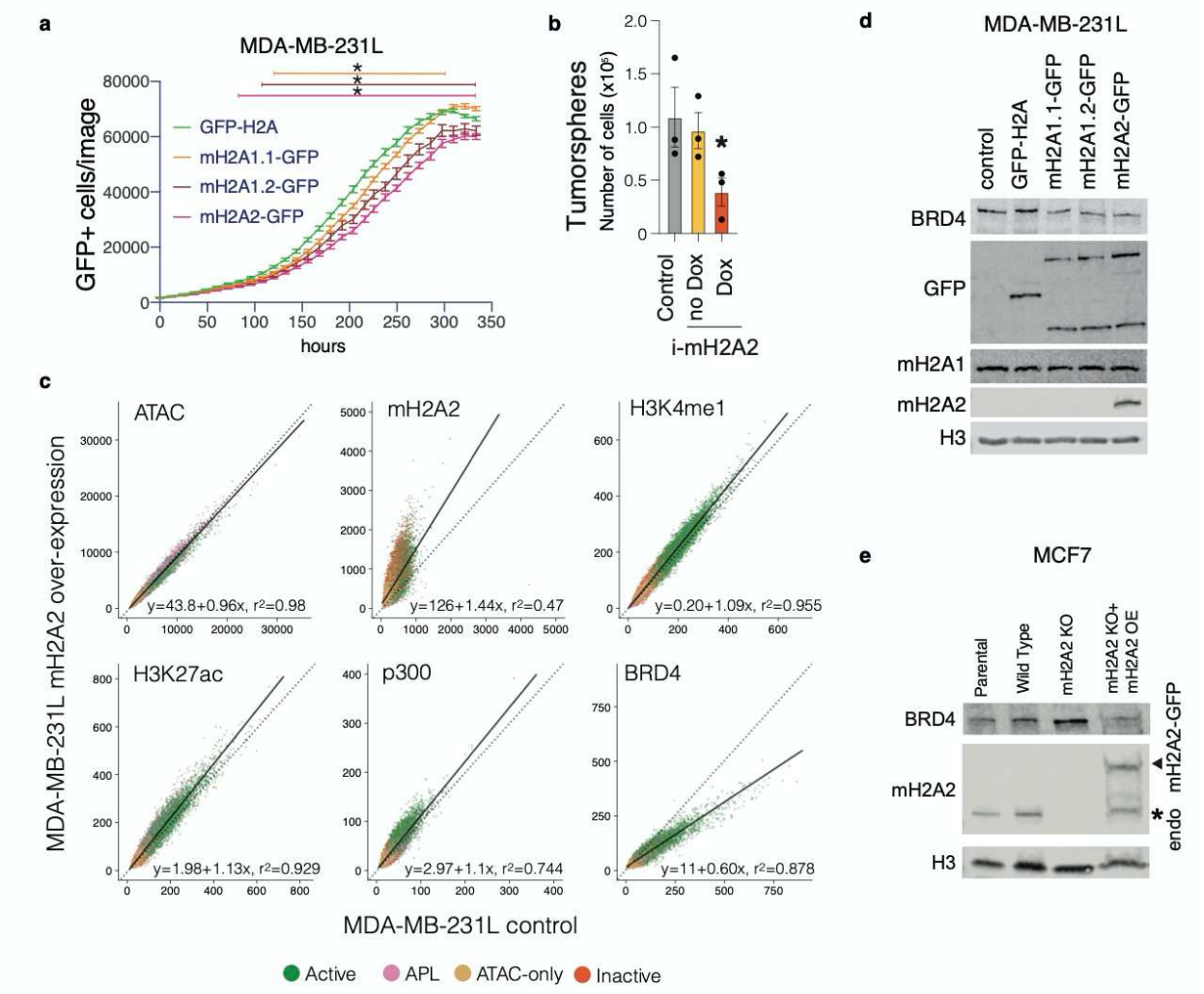

Figure 6. mH2A2 is a negative regulator of BRD4. a) Proliferation of MDA-MB-231L cells with over-expression of mH2A-GFP constructs (and H2A-GFP as control) determined by the number of GFP positive cells on Incucyte. Data are mean with SE (n=9). Two-way ANOVA (Dunnett's multiple comparison) *p<0,05. b) Tumorsphere formation assessed by number of cells upon mH2A2-GFP induction. Data represented are mean with SEM (n=3). Unpaired two tailed student's t-test *p<0,05. c) Scatter plot of ChIP-seq signals for mH2A2, H3K4me1, H3K27ac, p300 and BRD4, and ATAC-seq signal enrichment after over-expression of mH2A2 against control over-expression (GFP-H2A) at CREs in MDA-MB-231L cells. The regression line (solid line) is shown along with the line y = x for reference. d) Immunoblots from chromatin extracts in MDA-MB-231L cells with over-expression of mH2A-GFP constructs (and H2A-GFP as control) probed for BRD4, GFP and histone H3 (loading control). e) Immunoblots from chromatin extracts in MCF7 clones with over-expression of mH2A-GFP constructs (and H2A-GFP as control) probed for BRD4, mH2A2 and histone H3 (loading control).

FIGURE 7

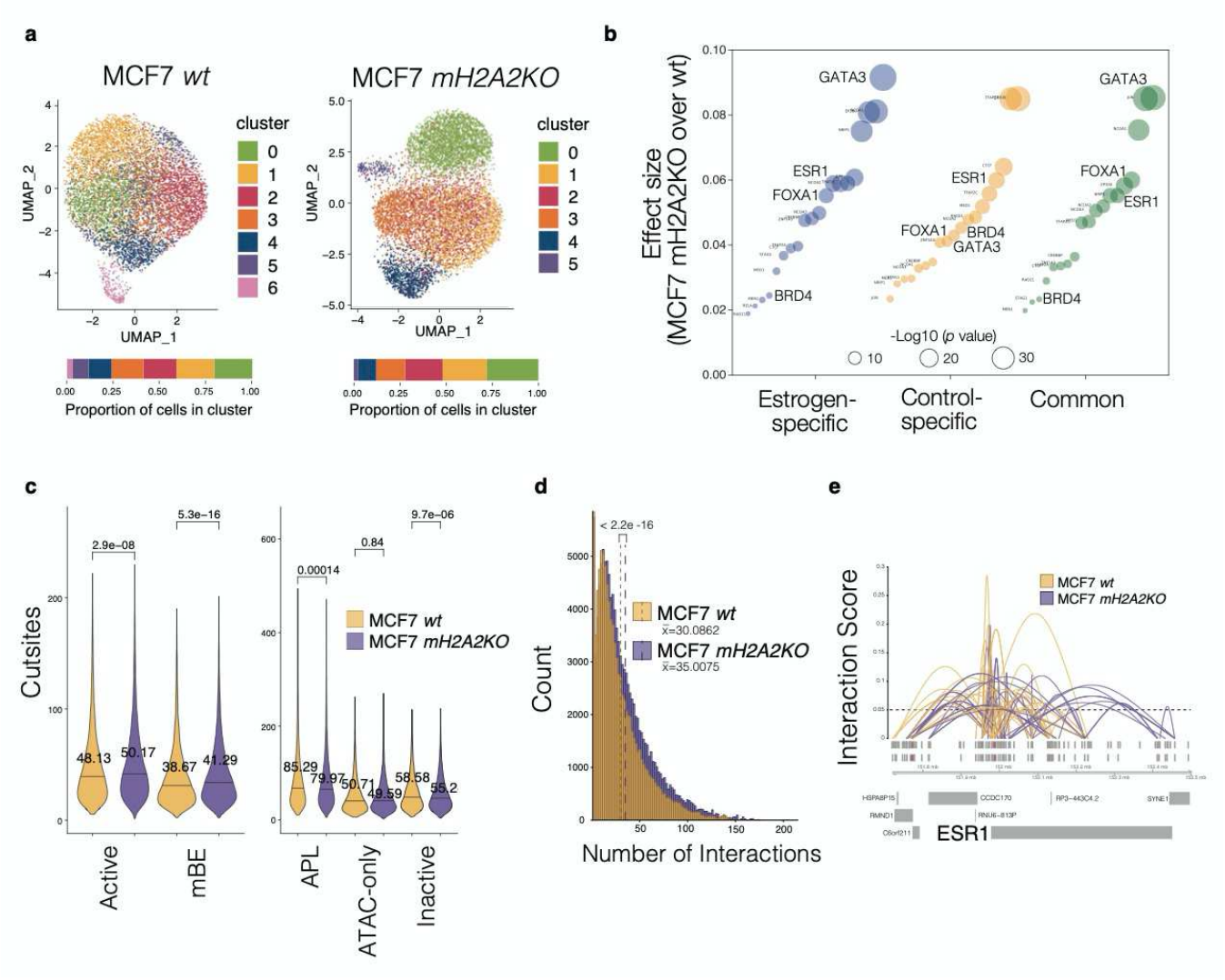

Figure 7. Loss of mH2A2 leads to increased cellular heterogeneity. a) UMAP plot of cells showing clusters identified by graph-based clustering of scATAC-seq signals from MCF7 wild type and MCF7 mH2A2 KO after 5 days of treatment with E2. Proportion of cells in each cluster represented as a bar graph (bottom). b) Cohen's effect size distributions of TF binding sites in mH2A2KO cells compared to wild-type MCF7 cells, grouped by binding sites found in estrogen- or control-specific cell lines (or both). c) Violin plot showing the number of cut sites overlapping different classes of enhancers in MCF7 control and mH2A2 KO clone. Mean and *p*-values shown were obtained by comparing means using unpaired Wilcoxon test. d) Histogram showing the distribution of the number of interactions between open chromatin regions predicted from the scATAC-seq data using Cicero. Mean is represented in dotted lines and wilcoxon rank sum test with continuity correction was used. e) Interaction plot around ESR1 locus for MCF7 wild type and mH2A2 KO clone with co-accessibility score given by Cicero.

Figures

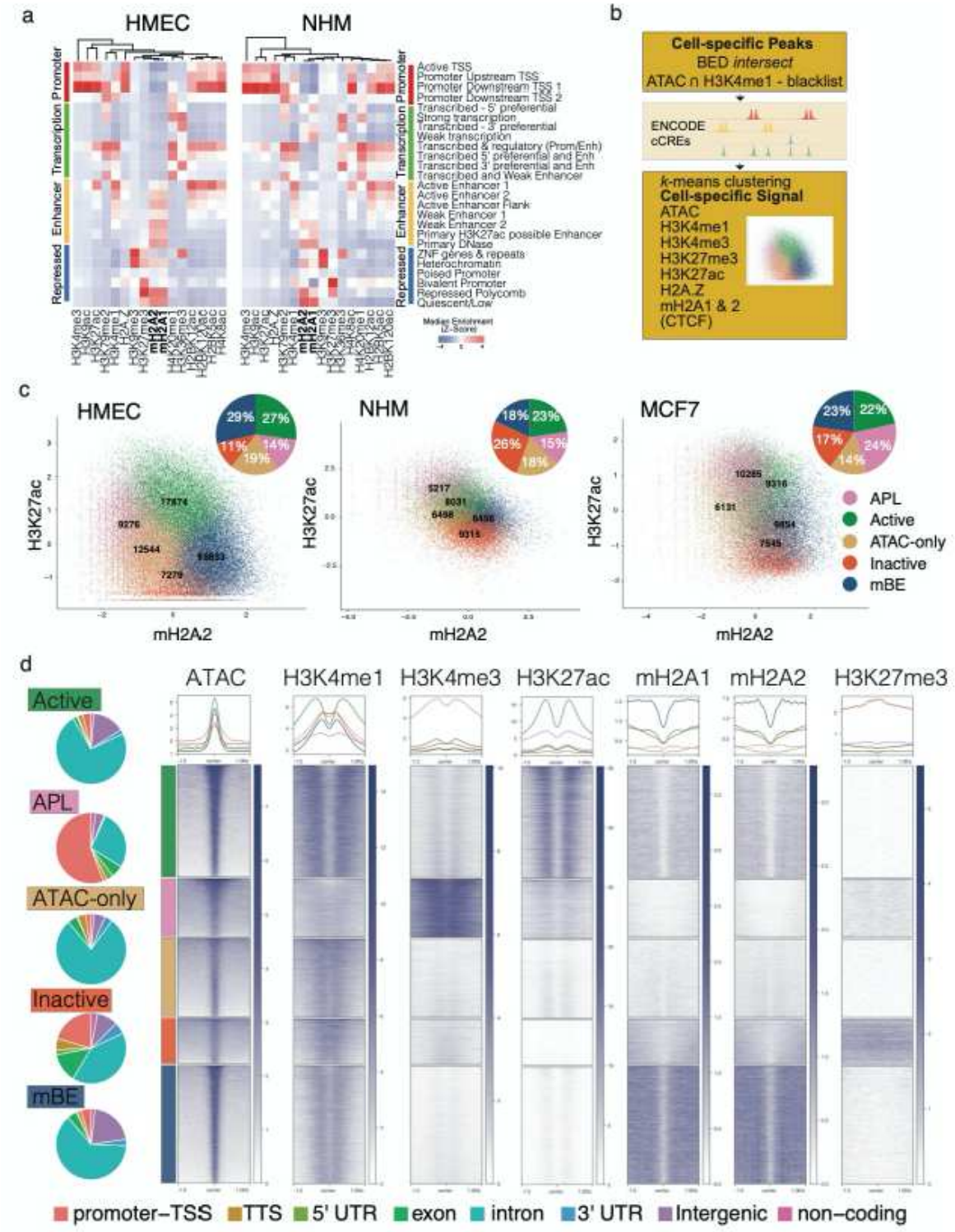

Figure 1

Characterization of macro-Bound Enhancers. a) Heatmap of the ENCODE chromatin states from HMEC and NHM with histone marks including the histone variants macroH2A1, macroH2A2. b) Algorithm to classify cell-specific cis-regulatory elements (CRE). c) Scatter plot showing the five classes assigned to

the regulatory elements after K-means clustering in three cell types (HMEC – Human Mammary Epithelial Cells, NHM – Normal Human Melanocytes, and MCF7 – breast cancer cell line), displayed along the Z scores for the log-normalized ChIP-seq signal of mH2A2 and H3K27ac. Pie charts show the proportions of CRE classes in each cell type. d-e) Characterization of the regulatory classes. Pie chart with genomic annotations for each regulatory class annotated using Homer (left). Signal profile (top) and heatmap (bottom) of enrichment around open chromatin regions (defined by ATAC-seq) of the five regulatory classes in HMEC (right). Enrichment of histone marks H3K4me1, H3K4me3, H3K27ac, H3K27me3 and histone variants macroH2A1 and macroH2A2 was analyzed by ChIP-seq.

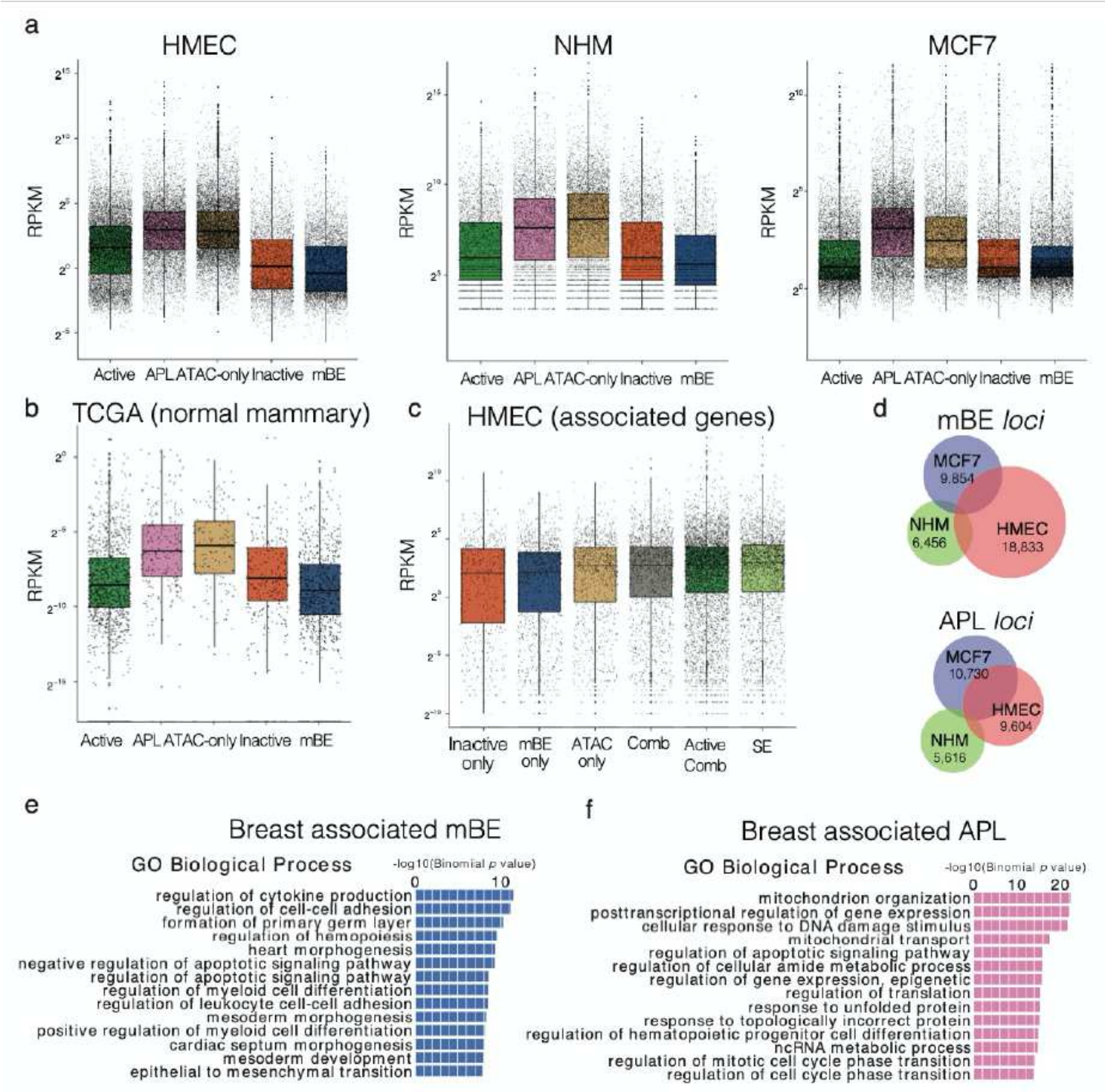

Figure 2

MacroH2A regulates enhancer activity. a) Expression levels at the CREs for HMEC, NHM and MCF7 by RNA-seq (logRPKM). b) Expression levels by RNA-seq from normal breast samples from TCGA33 using the annotated CREs in HMEC c) Expression levels in HMEC by RNA-seq of the genes directly associated with the CREs from HMEC (associations obtained from GeneHancer v4.4) and super-enhancers (annotated using LILY). The expression data is represented as boxplots where the middle line represents the median, the lower and upper edges of the rectangle represent the first and third quartiles and the lower and upper whiskers represent the interquartile range (IQR) x 1.5. Outliers beyond the end of the whiskers are plotted individually. d) Intersection of mBE and APL (H3K4me3) CRE loci between the three cell lines. e, f) Gene Ontology Biological Process terms of genes associated with mBE and APL (H3K4me3) common in HMEC and MCF7, respectively. Enriched terms ranked by Binomial p value and presented on a -log10 scale.

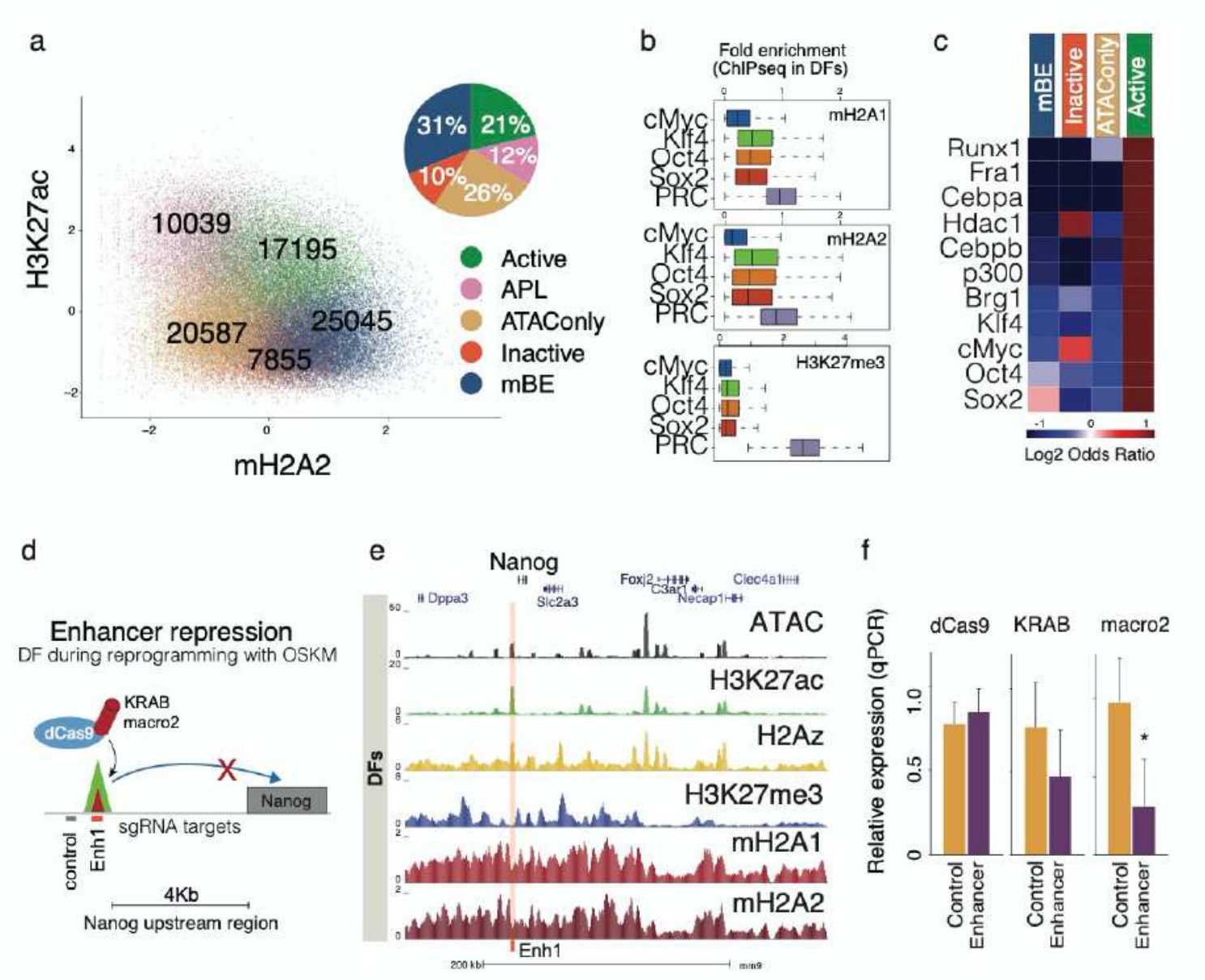

Figure 3

MacroH2A is a negative modulator of enhancer activity. a) Scatter plot upon K-means clustering of regulatory elements in mouse dermal fibroblasts (DFs), displayed along the Z scores for log-normalized ChIP-seq signals of mH2A2 and H3K27ac. Pie charts show proportions of CRE classes. b) Boxplot graph of ChIP-seq signal (fold enrichment) from mH2A1, mH2A2 and H3K27me3 in DFs at OSKM binding sites and PRC2 sites for comparison (48hrs after OSKM induction, binding sites data collected from35). The middle line represents the median, the lower and upper edges of the rectangle represent the first and third quartiles and the lower and upper whiskers represent the interquartile range (IQR) x 1.5. c) Odds ratio of enrichment of TF and chromatin binding factors after 48h of OSKM expression35 at enhancer elements of DF cells, ranked by enrichment in mBE (Fisher Exact test). d) Schematic of enhancer targeting in DF during reprogramming with OSKM using dCas9-KRAB and dCas9-macro2 and sgRNAs complementary to regions around the enhancer site. e) UCSC genome browser snapshot of the Klf4 binding site upstream of the Nanog TSS in DF with open chromatin (ATAC-seq), H3K27me3, H3K27ac, H2A.Z, mH2A1 and mH2A2 data. f) Nanog relative expression after 96h of OSKM infection in DFs with dCas9, dCas9-KRAB or dCas9-macro2 with sgRNAs targeting the enhancer site upstream of Nanog or control. Unpaired (two tailed) student’s t-test: * p value= 0.02.

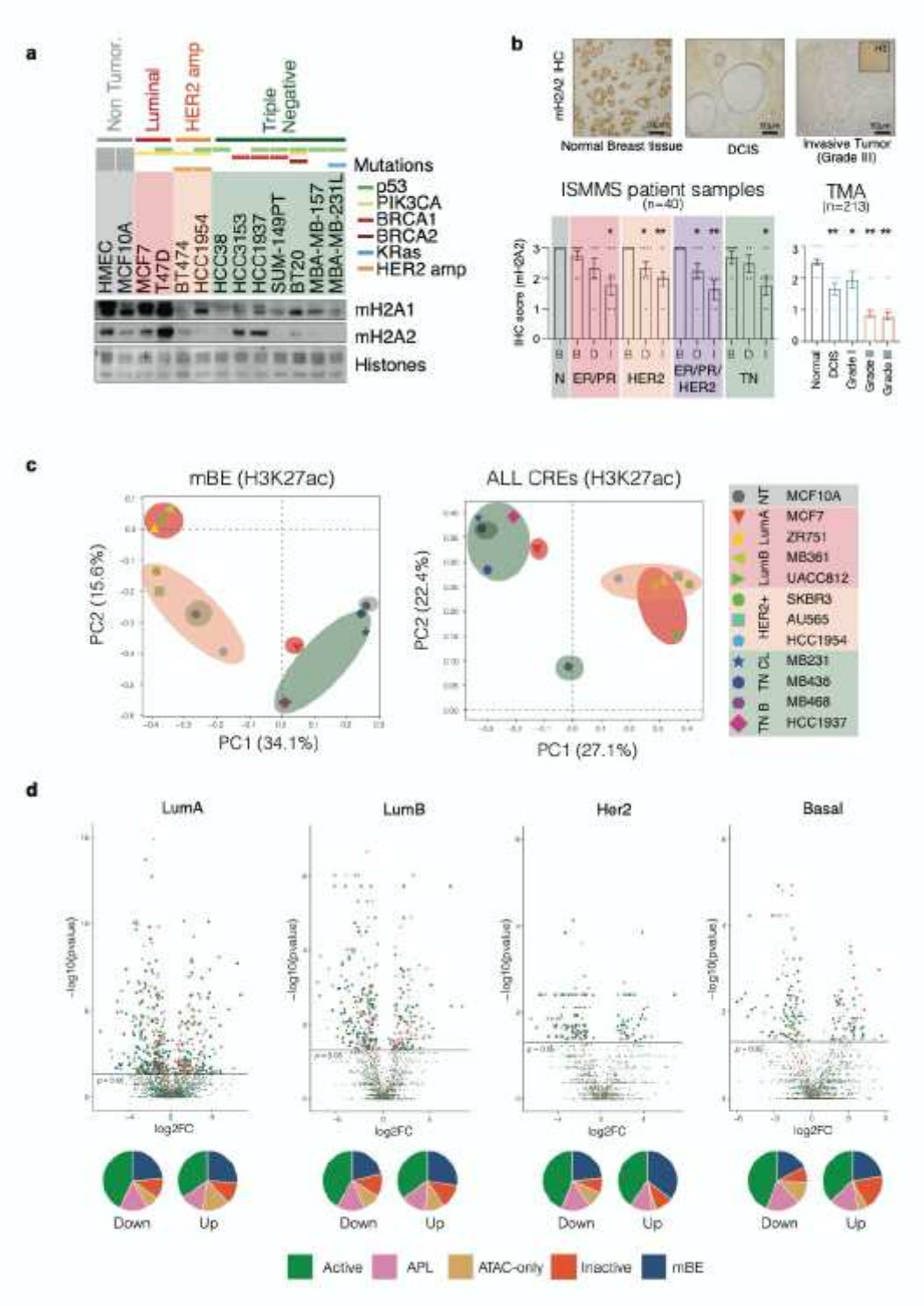

Figure 4

Reactivation of macro-Bound Enhancers associates with oncogenic programs. a) Immunoblot of chromatin extracts were probed for mH2A1 and mH2A2 across a panel of breast cancer cell lines including the three different major sub-types (estrogen and progesterone positive, HER2 positive and triple negative), and non-tumorigenic cells (human mammary epithelial cells, HMEC; and immortalized mammary cells MCF10A). Mutational status defined above. Amido Black of core histones used as loading control. b) Immunohistochemistry (IHC) from normal breast tissue, ductal carcinoma in situ

(DCIS) and grade III invasive tumors for mH2A2. Histone H3 IHC was used as a control (top right). Quantification of mH2A2 scoring for TMA and ISMMS patient samples according to the tumor grade or sub-class (below). B – Benign tissue, D – DCIS, I – Invasive. Column bar represents mean and SE. Unpaired (two tailed) student's t-test *p<0,05 ** p<0,005. c) Principal component analysis of H3K27ac ChIP-seq signal in HMEC macro-Bound enhancers (left) and all CREs (right) in 12 breast cancer cell lines from38. d) Volcano plot of differential enhancer expression in paired normal vs. tumor patient samples in breast cancer sub-types from RNA-seq (data collected from TCGA as described in ref.33), including the proportion of enhancers in each enhancer class that are significantly differentially expressed, down-regulated (down) or up-regulated (up), p-value < 0.05.

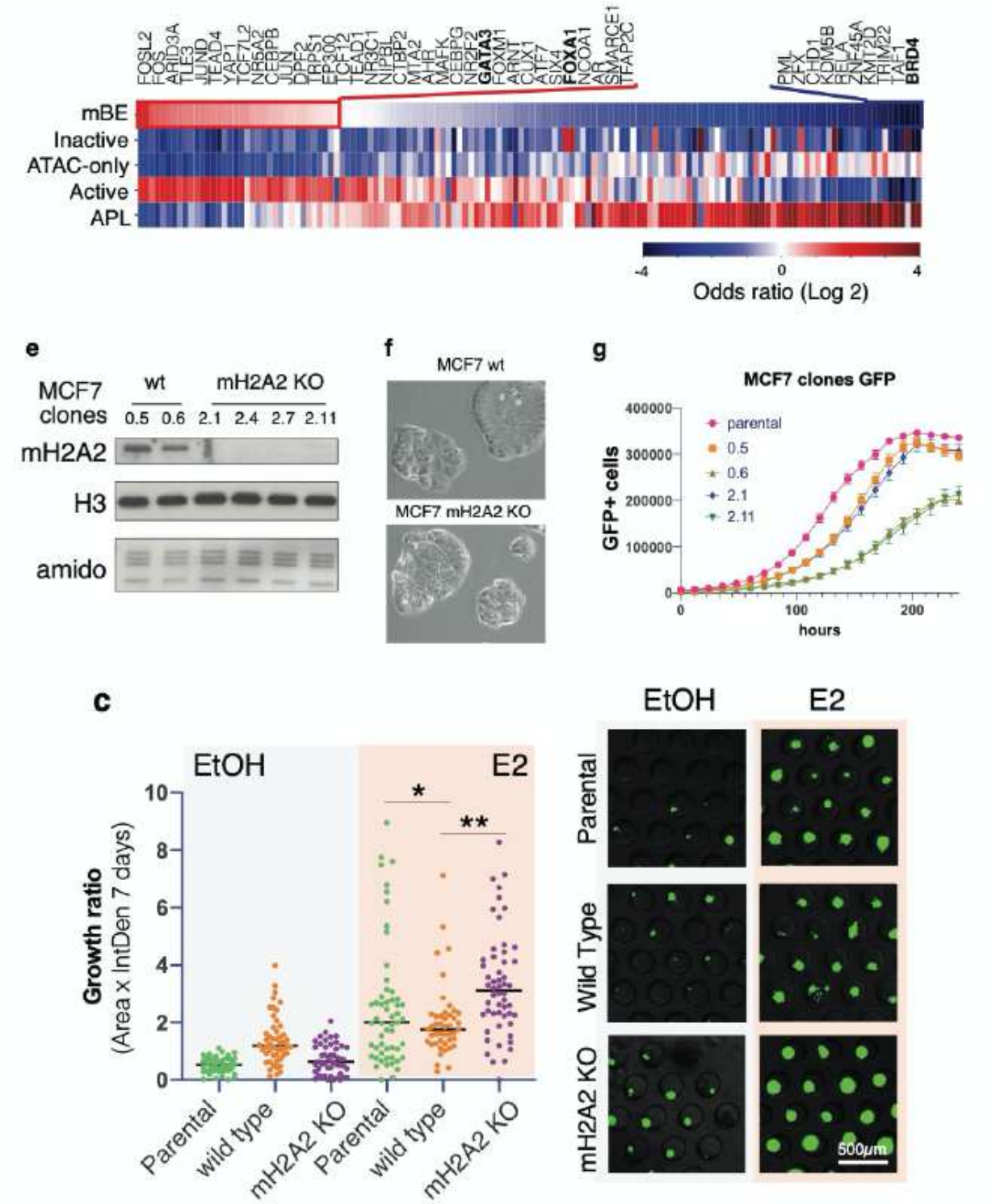

Figure 5

mH2A2 is a negative regulator of estrogen targets. a) Odds ratio of enrichment of TF and chromatin binding factors (ReMAP data for MCF7 cells) at enhancer elements of MCF7 cells, ranked by enrichment in mBEs (Fisher Exact test). b) Immunoblots for mH2A2 from chromatin extracts in MCF7 clones. H3 and histones (amido black) used as loading controls. c) Brightfield images of wild type and mH2A2 KO clones. d) Proliferation of MCF7 clones (wild type and mH2A2 KO) transduced with H2A-GFP and

analyzed by number of GFP cells using Incucyte. Data represented are mean with SE (n=3). e) Growth ratio of MCF7 3D spheroids after treatment with EtOH and E2 (Estradiol) in microwells after 7 days. Scatter plot of area factored with GFP Intensity Density in individual spheroids. Horizontal bars signify mean values (left). Representative images of MCF7 spheroids (right). Scale bar, 500µm. Unpaired (two tailed) student's t-test *p<0,05 ** p<0,005.

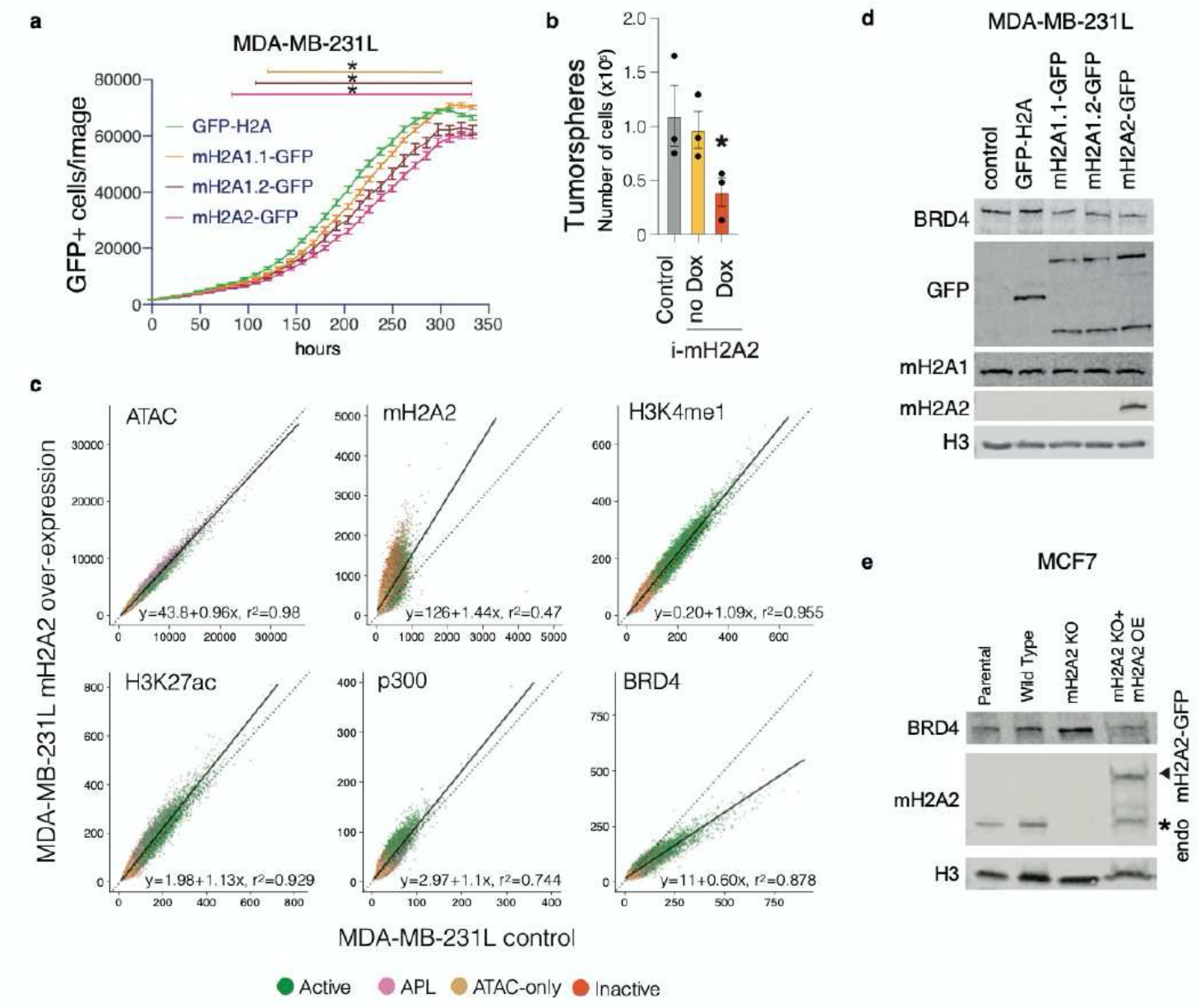

Figure 6

mH2A2 is a negative regulator of BRD4. a) Proliferation of MDA-MB-231L cells with over-expression of mH2A-GFP constructs (and H2A-GFP as control) determined by the number of GFP positive cells on Incucyte. Data are mean with SE (n=9). Two-way ANOVA (Dunnett's multiple comparison) *p<0,05. b) Tumorsphere formation assessed by number of cells upon mH2A2-GFP induction. Data represented are mean with SEM (n=3). Unpaired two tailed student's t-test *p<0,05. c) Scatter plot of ChIP-seq signals for mH2A2, H3K4me1, H3K27ac, p300 and BRD4, and ATAC-seq signal enrichment after over-expression of

mH2A2 against control over-expression (GFP-H2A) at CREs in MDA-MB-231L cells. The regression line (solid line) is shown along with the line y = x for reference. d) Immunoblots from chromatin extracts in MDA-MB-231L cells with over-expression of mH2A-GFP constructs (and H2A-GFP as control) probed for BRD4, GFP and histone H3 (loading control). e) Immunoblots from chromatin extracts in MCF7 clones with over-expression of mH2A-GFP constructs (and H2A-GFP as control) probed for BRD4, mH2A2 and histone H3 (loading control).

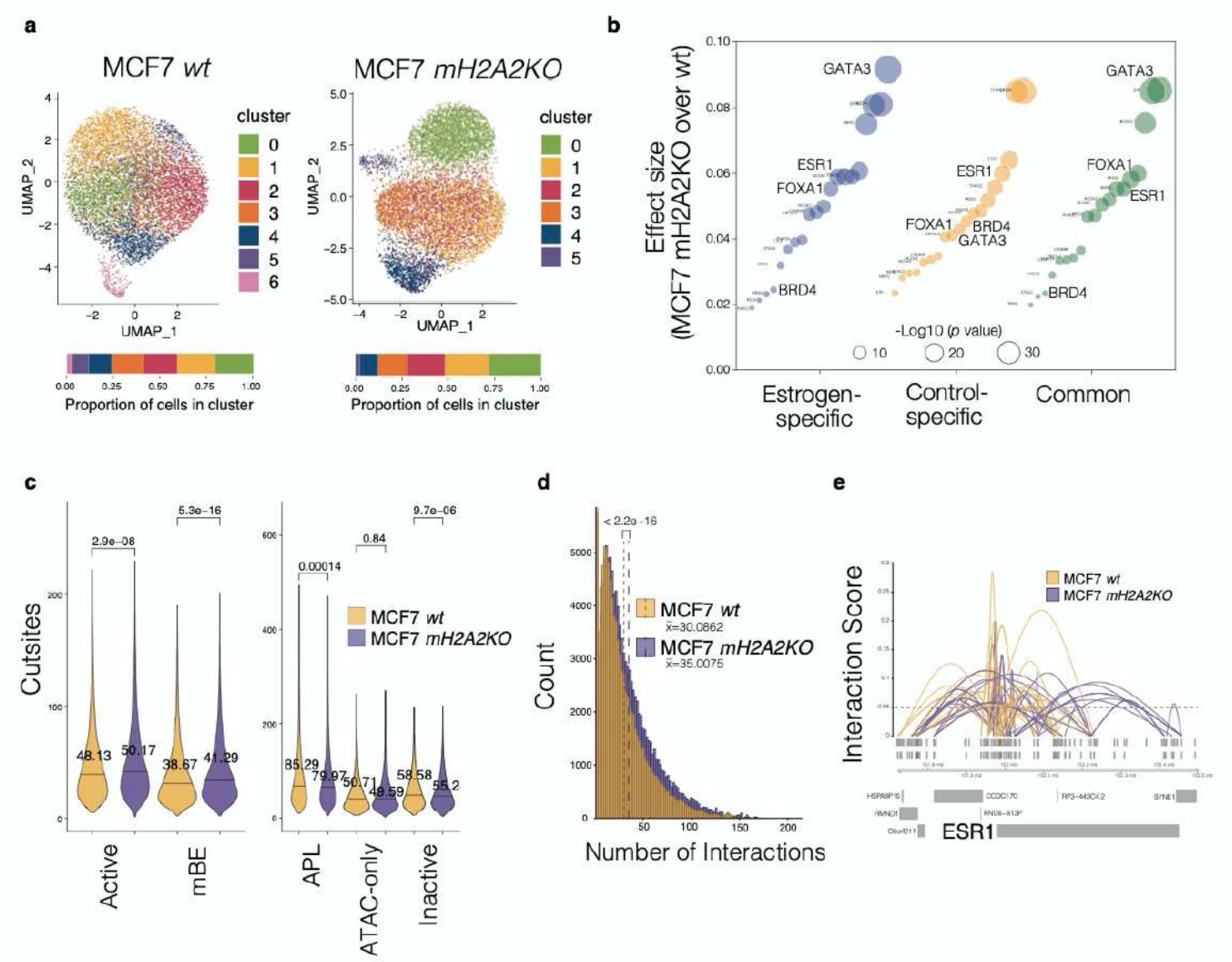

Figure 7

Loss of mH2A2 leads to increased cellular heterogeneity. a) UMAP plot of cells showing clusters identified by graph-based clustering of scATAC-seq signals from MCF7 wild type and MCF7 mH2A2 KO after 5 days of treatment with E2. Proportion of cells in each cluster represented as a bar graph (bottom). b) Cohen's effect size distributions of TF binding sites in mH2A2KO cells compared to wild-type MCF7 cells, grouped by binding sites found in estrogen- or control-specific cell lines (or both). c) Violin plot showing the number of cut sites overlapping different classes of enhancers in MCF7 control and mH2A2 KO clone. Mean and p-values shown were obtained by comparing means using unpaired Wilcoxon test.

d) Histogram showing the distribution of the number of interactions between open chromatin regions predicted from the scATAC-seq data using Cicero. Mean is represented in dotted lines and wilcoxon rank sum test with continuity correction was used. e) Interaction plot around ESR1 locus for MCF7 wild type and mH2A2 KO clone with co-accessibility score given by Cicero.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- macroXEnh2021v5.0Supp.pdf