Article

Deficient H2A.Z deposition is associated with genesis of uterine leiomyoma

https://doi.org/10.1038/s41586-021-03747-1

Received: 26 June 2020

Accepted: 18 June 2021

Published online: 4 August 2021

Check for updates

Davide G. Berta[1,2,7], Heli Kuisma[1,2,7], Niko Välimäki[1,2], Maritta Räisänen[1,2], Maija Jäntti[1,2], Annukka Pasanen[3], Auli Karhu[1,2], Jaana Kaukomaa[1,2], Aurora Taira[1,2], Tatiana Cajuso[1,2], Sanna Nieminen[1,2], Rosa-Maria Penttinen[1,2], Saija Ahonen[1,2], Rainer Lehtonen[1,2], Miika Mehine[1,2], Pia Vahteristo[1,2], Jyrki Jalkanen[4], Biswajyoti Sahu[2], Janne Ravantti[1,2], Netta Mäkinen[1,2], Kristiina Rajamäki[1,2], Kimmo Palin[1,2,5], Jussi Taipale[2], Oskari Heikinheimo[6], Ralf Bützow[2,3], Eevi Kaasinen[1,2]✉ & Lauri A. Aaltonen[1,2,5]✉

One in four women suffers from uterine leiomyomas (ULs)—benign tumours of the uterine wall, also known as uterine fibroids—at some point in premenopausal life. ULs can cause excessive bleeding, pain and infertility[1], and are a common cause of hysterectomy[2]. They emerge through at least three distinct genetic drivers: mutations in *MED12* or *FH*, or genomic rearrangement of *HMGA2*[3]. Here we created genome-wide datasets, using DNA, RNA, assay for transposase-accessible chromatin (ATAC), chromatin immunoprecipitation (ChIP) and HiC chromatin immunoprecipitation (HiChIP) sequencing of primary tissues to profoundly understand the genesis of UL. We identified somatic mutations in genes encoding six members of the SRCAP histone-loading complex[4], and found that germline mutations in the SRCAP members *YEATS4* and *ZNHIT1* predispose women to UL. Tumours bearing these mutations showed defective deposition of the histone variant H2A.Z. In ULs, H2A.Z occupancy correlated positively with chromatin accessibility and gene expression, and negatively with DNA methylation, but these correlations were weak in tumours bearing SRCAP complex mutations. In these tumours, open chromatin emerged at transcription start sites where H2A.Z was lost, which was associated with upregulation of genes. Furthermore, *YEATS4* defects were associated with abnormal upregulation of bivalent embryonic stem cell genes, as previously shown in mice[5]. Our work describes a potential mechanism of tumorigenesis—epigenetic instability caused by deficient H2A.Z deposition—and suggests that ULs arise through an aberrant differentiation program driven by deranged chromatin, emanating from a small number of mutually exclusive driver mutations.

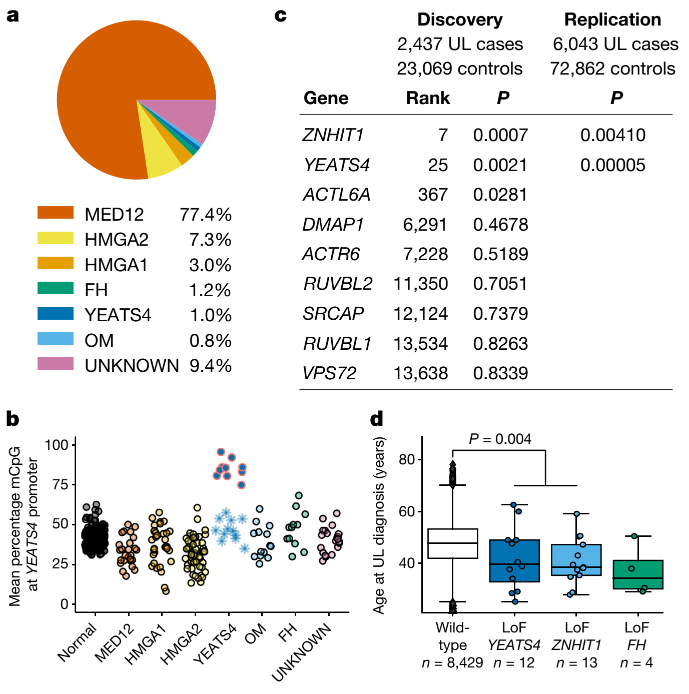
Uterine leiomyoma is the most common noncutaneous tumour. The cumulative incidence is more than 70% by the age of 50 years. Between 25 and 30% of women manifest symptoms[6,7] including pain, heavy menstrual bleeding, ascites and reduced fertility[1]. Three UL driver mutations have been identified: 70% of ULs carry a mutation in the *MED12* gene, 10–15% have chromosomal aberrations that activate *HMGA2* expression[3,8], and 1% have biallelic loss of *FH*[9].

To characterize the molecular landscape of UL, we set up the Finland Myoma Study, which currently comprises 728 patients and 2,263 fresh-frozen tumours with paired normal myometrium (Supplementary Tables 1, 2, Supplementary Fig. 1). Each tumour was subjected to single nucleotide polymorphism (SNP)-chip analysis of chromosomal gains and losses, followed by identification of driver mutations (*MED12*, *HMGA2*, *FH*, or unknown, hereafter referred to as MED12, HMGA2, FH and UNKNOWN subclasses, respectively) (Extended Data Fig. 1a). These efforts were amended by genome-wide analyses in normal myometria and tumours representing the four subclasses (Supplementary Table 2) through RNA sequencing (RNA-seq) (276 UL and 162 normal myometrium samples), nanopore sequencing for DNA methylation (106 UL, 96 normal), ATAC with high-throughput sequencing (ATAC–seq) (16 UL, 15 normal), ChIP with sequencing (ChIP–seq) (for H2A.Z from 24 UL, 15 normal; H3K27ac from 5 UL, 5 normal; H3K27me3 from 6 normal; H3K4me3 from 7 normal) and HiChIP (5 UL, 5 normal).

SRCAP complex gene mutations in ULs

As the known driver mutations seem to be mutually exclusive[3,8], we chose a staged mutation detection approach. MED12, HMGA2 and FH tumours were encountered as expected[3] (Fig. 1a). Samples representing each subclass (42 MED12, 44 HMGA2, 15 FH) and all tumours for

[1]Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland. [2]Applied Tumor Genomics Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland. [3]Department of Pathology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [4]Department of Obstetrics and Gynecology, Central Finland Central Hospital, Jyväskylä, Finland. [5]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, Helsinki, Finland. [6]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [7]These authors contributed equally: Davide G. Berta, Heli Kuisma. ✉e-mail: eevi.kaasinen@helsinki.fi; lauri.aaltonen@helsinki.fi



	Discovery 2,437 UL cases 23,069 controls		Replication 6,043 UL cases 72,862 controls
Gene	Rank	*P*	*P*
ZNHIT1	7	0.0007	0.00410
YEATS4	25	0.0021	0.00005
ACTL6A	367	0.0281	
DMAP1	6,291	0.4678	
ACTR6	7,228	0.5189	
RUVBL2	11,350	0.7051	
SRCAP	12,124	0.7379	
RUVBL1	13,534	0.8263	
VPS72	13,638	0.8339	

Fig. 1 | Mutations in SRCAP complex genes identified as a driver of UL. **a**, Driver change spectrum in 2,263 ULs. **b**, Allele-specific methylation of *YEATS4* promoter (1,000 bp upstream from TSS) for 84 normal and 93 UL samples. Each dot depicts average methylation for one allele at the *YEATS4* promoter. For YEATS4 tumours, dots with a red outline represent wild-type alleles and asterisks represent *YEATS4* mutant alleles (one YEATS4 tumour had both alleles mutated). The methylation difference between the alleles in YEATS4 tumours was tested with paired two-sided *t*-test ($P = 4.47 \times 10^{-5}$, 95% confidence interval (CI) of difference 28–48 percentage points). **c**, **d**, Gene-based UL associations of LoF variants in UK Biobank WES data. **c**, *YEATS4* and *ZNHIT1* were among the top 25 associations, out of 15,341 genes examined, and were validated in an independent batch of UK Biobank WES data (SKAT-O test). **d**, Individuals with germline LoF variants were diagnosed with UL at an earlier age than wild-type individuals ($P = 0.004$; Welch's two-sided *t*-test on *YEATS4* and *ZNHIT1* LoF variant carriers). Age at diagnosis was available for 8,458 individuals with WES data. *FH* LoF carriers, who are known to be predisposed to developing UL[17], are shown for reference. Centre line, median; box limits, first and third quartiles; whiskers, 1.5 × interquartile range (IQR) past the quartiles; points outside this range are shown as a dot plot.

which no driver mutation was identified and RNA was available (175 UNKNOWN samples) entered RNA-seq. Sixty-six of the UNKNOWN samples were found to overexpress *HMGA1* here (Supplementary Fig. 2), and one in a previous analysis. A search of the RNA-seq data identified a subclass of ULs with heterozygous somatic mutations in SRCAP complex subunits (hereafter referred to as SRCAP tumours). Using all our data sources, we found 39 such ULs, and identified another four in which SRCAP complex gene expression was notably reduced (Supplementary Table 3, Extended Data Fig. 1, Supplementary Figs. 3, 4; Methods). Twenty-four tumours bore *YEATS4* alterations (hereafter referred to as YEATS4 tumours) and 19 had alterations in other SRCAP member genes (hereafter other member (OM) tumours). All driver changes were typically mutually exclusive (Supplementary Discussion, Supplementary Table 4).

The SRCAP complex deposits the histone variant H2A.Z onto chromatin[4,5]. P400–TIP60 is another complex that is involved in H2A.Z loading[4]. H2A.Z is involved in processes such as transcriptional regulation and DNA repair[4]. Overexpression of H2A.Z has been linked to cancer[10–14], and its acetylation and ubiquitination to positive and negative regulation of gene expression, respectively[4,15,16]

Six of the nine SRCAP complex members (HGNC gene group 1329) (Extended Data Fig. 1b, Supplementary Fig. 3) contained mutations. The second allele was commonly inactivated by allelic loss (Supplementary Discussion, Supplementary Tables 3, 5, 6) or, in *YEATS4*, hypermethylation (Fig. 1b, Supplementary Figs. 5, 6). Six patients had at least two SRCAP tumours, ($P = 1 \times 10^{-6}$, permutation test), which suggests a predisposition due to hereditary or environmental factors.

We examined the effects of hereditary loss-of-function (LoF) variants in SRCAP complex genes on UL predisposition in 50,000 individuals for whom whole-exome sequencing (WES) data were available at the UK Biobank (UKB) (Fig. 1c). This discovery cohort revealed two strong candidate associations: *ZNHIT1* and *YEATS4*. The UKB October 2020 data release enabled us to independently validate these findings (*ZNHIT1* $P = 4.1 \times 10^{-3}$, SKAT-O test and odds ratio (OR) = 3.8; *YEATS4* $P = 5.1 \times 10^{-5}$, OR = 5.7) (Fig. 1c, d, Extended Data Fig. 1b, Supplementary Tables 7–10). *ZNHIT1* and *YEATS4* LoF germline mutations were more frequent ($n = 25$) in individuals with ULs than were similar *FH* mutations ($n = 4$; Fig. 1d) that are known to predispose women to UL[17].

H2A.Z immunohistochemistry in ULs

To study whether SRCAP complex mutations affect cellular levels of H2A.Z, we used immunohistochemical (IHC) staining of 265 tumours (representing the different subclasses) for H2A.Z and acetylated H2A.Z (H2A.Zac). Normal myometrium and MED12 and FH tumour samples showed strong nuclear H2A.Z staining (Extended Data Figs. 2a, b, 3). YEATS4 and OM tumours showed weak or absent staining (Extended Data Fig. 2b, c, Supplementary Fig. 7). Three tumours containing mutations in the SRCAP complex-specific *ZNHIT1* gene showed similar H2A.Z loss to tumours with mutations in *YEATS4* (a member of both SRCAP and P400–TIP60), which suggests that P400–TIP60 did not rescue this defect. HMGA2 and HMGA1 tumours combined displayed an intermediate pattern with significantly reduced staining (Extended Data Fig. 2b, c, Supplementary Fig. 7). These differences were replicated for H2A.Zac (Extended Data Fig. 2b, c, 3, Supplementary Fig. 7). Western blots of chromatin extracts from one MED12 and one YEATS4 tumour, and related myometrium, showed similar results (Extended Data Fig. 2d). These results could arise through degradation of free H2A.Z[18].

Deposition of H2A.Z in chromatin

To examine genome-wide H2A.Z deposition in UL subclasses and myometrium, we performed ChIP–seq on tissue specimens. As proof of ChIP quality, peaks from myometrium were enriched at H2A.Z peak regions derived from fibroblasts by the Roadmap epigenomics project[19] (OR = 223, one-sided Fisher's exact test $P < 1 \times 10^{-300}$) (Supplementary Fig. 8a). The peaks were also enriched at active and bivalent transcription start site (TSS) regions (TssA and TssBiv, respectively) (Supplementary Fig. 9a), derived from our five-state genome segmentation on pooled myometrium data using H3K27ac, H3K27me3 and H3K4me3 ChIP–seq. The same enrichment was observed using published segmentations of 21 mesenchymal or stem cell types[19] (Supplementary Fig. 8b, Supplementary Table 11).

We then examined differential H2A.Z binding by comparing multiple tumours per subclass against multiple myometrium samples. The correlation of read counts on the peaks showed that H2A.Z binding patterns differed between tumours and normal samples, and between tumour subclasses, with YEATS4 and OM tumours having the highest correlation (Fig. 2a, Supplementary Fig. 9b, c). The differential binding analysis was validated using a spike-in strategy in the ChIP against H2A.Z. MED12 tumours showed a reduction in H2A.Z, especially at TssA (Extended Data Figs. 4a, 5, Supplementary Figs. 10–13). In the HMGA2 subclass, the overall effect was a redistribution of H2A.Z peaks (Extended Data Figs. 4b, 5, Supplementary Figs. 10–13). None of the HMGA2 tumours that were randomly selected for H2A.Z ChIP showed

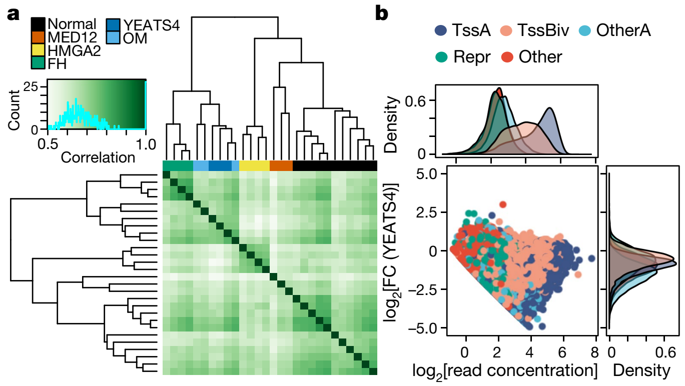


Fig. 2 | H2A.Z binding on chromatin is changed in UL. a, A correlation heat map shows distinct H2A.Z binding on chromatin in tumours ($n = 17$) as compared to normal myometria ($n = 11$). YEATS4 and OM tumours are intermixed. Correlations were calculated using normalized read counts on 59,647 H2A.Z peaks. **b**, H2A.Z binding difference in YEATS4 tumours ($n = 3$) as compared to normal samples ($n = 11$). $\log_2$FC and average binding strength (normalized read concentration) were calculated on each H2A.Z peak region, represented by a dot. H2A.Z peak regions were stratified by five-state genome annotations from myometrium. States were annotated as bivalent TSS regions (TssBiv marked by both H3K4me3 and H3K27me3), active TSS regions (TssA marked by both H3K4me3 and H3K27ac), active chromatin outside TSS regions (OtherA marked by H3K27ac), repressed chromatin (Repr marked by H3K27me3) and other chromatin (Other).

weak H2A.Z IHC staining, and a global reduction in H2A.Z binding was not seen. Analysis of HMGA2 tumours with moderate H2A.Z staining showed the expected reduction in ChIP–seq (Supplementary Fig. 14; data not included in other analyses). H2A.Z binding changes in FH tumours did not accumulate at any particular chromatin segmentation (Extended Data Figs. 4c, 5, Supplementary Figs. 10–13). SRCAP tumours typically displayed a global reduction in H2A.Z peaks (Fig. 2b, Extended Data Figs. 4d, 5, Supplementary Figs. 10–13), similar to the results of *Yeats4* knockdown in mouse embryonic stem cells (mES cells)[5]. This phenomenon was particularly prominent at active chromatin outside TSSs (OtherA) (Fig. 2b, Supplementary Figs. 10–13). The results demonstrate a marked defect in H2A.Z deposition in SRCAP tumours that is compatible with model systems[20–22].

H2A.Z sites and chromatin accessibility

We generated ATAC–seq data to study the effects of H2A.Z loss and differences in chromatin organization in ULs of different subclasses. Myometria had different ATAC–seq profiles from tumours (Supplementary Fig. 15a, b). Compatible with the literature[23,24], changes in H2A.Z occupancy in ULs correlated positively with chromatin accessibility (Extended Data Fig. 6a, Supplementary Fig. 16). In YEATS4 and MED12 tumours that displayed a global reduction in H2A.Z by ChIP–seq, this correlation was weak (Extended Data Fig. 6a, Supplementary Fig. 16).

We performed a genome-wide analysis to measure whether myometrial H2A.Z peak regions were overall more open or closed in tumours, and whether these differentially accessible H2A.Z regions were enriched on particular chromatin state regions from myometria. More open H2A.Z sites in YEATS4 tumours (4.1% of all H2A.Z peaks) were enriched on TssBiv regions (OR = 2.6, one-sided Fisher's exact test $P = 1 \times 10^{-98}$), whereas more closed H2A.Z sites (3.5% of all H2A.Z peaks) were enriched in OtherA regions (OR = 4.1, one-sided Fisher's exact test $P = 1 \times 10^{-273}$) (Supplementary Fig. 17).

We next identified regions that were differentially more accessible in each UL subclass ($\log_2$ fold change ($\log_2$FC) > 0 and FDR < 0.05; hereafter referred to as differentially accessible regions (DARs)). FH and HMGA2 tumours had the largest number of more open regions (7,837 and 5,716 DARs, respectively) (Supplementary Fig. 15c). One hundred and fifty-six DARs were shared by all UL subclasses, including eighteen in promoter-proximal regions (Supplementary Table 12). DARs in YEATS4 tumours (2,394) were strongly enriched on TssA and TssBiv, and the presence of H2A.Z in myometrium was associated with the emergence of DARs in the tumours (Supplementary Fig. 18, Supplementary Table 13). DARs in YEATS4 ULs displayed strong H2A.Z binding in pooled myometrium data, and were mostly unique to the YEATS4 subclass (Extended Data Fig. 6b, c). We measured differences in H2A.Z binding at DARs using spike-in ChIP–seq and found that H2A.Z was significantly reduced in MED12 and YEATS4 tumours (Extended Data Fig. 6d). These results suggest that chromatin opening in YEATS4 tumours displays a preference for active and bivalent TSSs that are bound by H2A.Z in normal myometrium, and that H2A.Z is reduced at these new open chromatin regions.

H2A.Z sites and DNA methylation

In a hierarchical clustering analysis of DNA methylation, normal myometrium and MED12, FH and YEATS4 tumours clustered according to mutation status (Supplementary Figs. 19, 20). We measured methylation at sites occupied by H2A.Z in pooled myometrium data and found that these regions were significantly more methylated in FH, HMGA2, YEATS4 and OM tumours than in normal myometrium (Extended Data Fig. 7h). MED12 tumours did not show an increase in methylation at these loci (Extended Data Fig. 7h) despite having enrichment of hypermethylation at bivalent chromatin similar to HMGA2 tumours (Supplementary Discussion, Extended Data Fig. 7b, d). We then measured DNA methylation at H2A.Z sites that showed differential binding in UL subclasses (FDR < 0.05 and $|\log_2\text{FC}| > 1$; from the spike-in data) and found an inverse correlation (Extended Data Fig. 6e), consistent with data from model systems[4].

H2A.Z loss and gene expression

YEATS4 ULs showed aberrant upregulation of differentiation genes regulated by bivalent promoters, similar to the results of *Yeats4* knockdown in mES cells[5] (Fig. 3a, Supplementary Fig. 21). To better understand how loss of H2A.Z exerts its tumorigenic action, we examined the effects of changes in H2A.Z occupancy on gene expression. We measured the expression of genes with a differential H2A.Z binding site (FDR < 0.05, $|\log_2\text{FC}| > 1$; from the spike-in data) at up to 250 bp from the TSS (Supplementary Fig. 22a). In the HMGA2, HMGA1 and FH subclasses, decreased H2A.Z binding led to downregulation and increased H2A.Z binding to upregulation of genes (Extended Data Fig. 8a, Supplementary Fig. 23). In YEATS4, OM and MED12 tumours, the positive correlation was less clear (Extended Data Fig. 8a). In YEATS4 tumours, 204 genes with reduced H2A.Z at TSS were overexpressed and 296 underexpressed (FDR < 0.05) (Fig. 3b, Supplementary Table 14). The most significantly enriched ontologies for overexpressed genes were related to morphogenesis, whereas no such enrichment was detected for underexpressed genes. Almost all genes with DARs in proximity to reduced H2A.Z binding were overexpressed (Fig. 3b).

We investigated 3D chromatin interactions in tumours and paired myometrium with HiChIP and ChIP–seq against H3K27ac, a mark of active chromatin that is associated with promoter–enhancer loops. We identified 33,014; 20,523; 26,719; 20,275; and 11,137 differential 3D interactions (FDR < 0.05) at 40-kb resolution in MED12; HMGA2; FH; YEATS4; and OM tumours, respectively. We then examined the overlap between DARs identified in UL subclasses and bins involved in the increased interactions. Most of the genes with increased 3D interaction and a DAR were significantly overexpressed in MED12, HMGA2 and YEATS4 tumours (Supplementary Tables 15–18). We then studied differential 3D interactions connected to H2A.Z sites outside TSS regions and their effects on gene expression (Supplementary Fig. 22b). We resolved

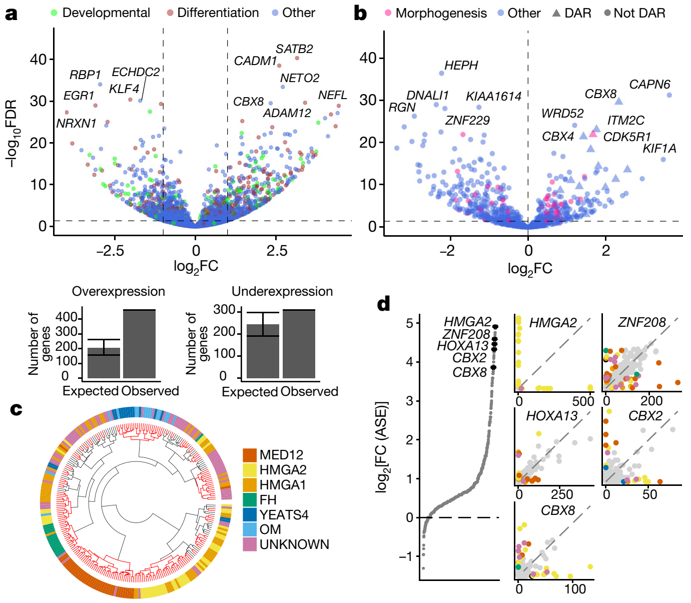


Fig. 3 | Clues to the pathogenesis of UL derived from expression data. a, Top, expression of differentiation genes with bivalent promoter in YEATS4 tumours ($n = 16$) against normal samples ($n = 162$) is similar to one observed in *Yeats4* knockdown mouse ES cells[5]. To replicate the setting, genes with bivalent chromatin annotation in the human H9 ES cell line within 5 kb of the TSS were selected (3,356 genes). Differentiation genes (gene ontology (GO):0030154) and developmental genes (GO:0032502) are highlighted. Bottom, expected values were calculated by permutation with 10,000 iterations. Mean and maximum and minimum values for number of over- and under-expressed genes. **b**, Expression differences for genes with reduced H2A.Z binding site at TSS in YEATS4 tumours against normal samples. Morphogenesis genes (GO:0032989) are highlighted. In particular, genes where the decreased H2A.Z binding site is within 1 kb of the YEATS4 tumour DAR (triangles) are overexpressed. **c**, Unsupervised hierarchical clustering of ULs based on global expression patterns. Red branches denote clusters that are strongly supported with bootstrapping at a significance level of 10%. **d**, Genes with highest proportions of ASE. Left, genes ranked by enrichment of ASE among tumours ($\log_2$FC for tumours (one tumour per patient, $n = 240$) compared to normal samples ($n = 160$)). Right, ASE in five high-ranking genes (grey dots, samples with balanced allelic expression; coloured dots, samples with ASE). The high rank of *HMGA2*, which is typically activated by allele-specific translocation, serves as validation for the approach.

significantly more H2A.Z-associated differential 3D interactions to differentially expressed genes in MED12 ($P < 0.0346$), FH ($P < 0.0008$), YEATS4 ($P < 0.0019$) and OM ($P < 0.0321$) tumours genome-wide than expected by chance (Supplementary Table 19). Interactions from decreased H2A.Z sites tended to be weakened and expression of the connected genes decreased. This suggests that in SRCAP tumours, the decrease in H2A.Z binding at active regions outside TSSs typically leads to decreased expression of the connected genes.

Gene expression patterns in ULs

Next, we studied global expression profiles in ULs. Normal myometrium samples and MED12, HMGA, SRCAP and FH tumours tended to form separate clusters (Fig. 3c, Extended Data Fig. 8b, c). Linear discriminant analysis shortlisted genes with distinct expression patterns between subclasses (Extended Data Fig. 8b, Supplementary Table 20). Analyses for differentially expressed genes revealed altered pathways shared by all or subsets of UL subclasses (Supplementary Tables 21, 22, Extended Data Fig. 9).

Three adjacent CBX family member genes, *CBX2*, *CBX4* and *CBX8* (hereafter referred to as *CBX2/4/8*) were highly overexpressed across all UL subclasses, whereas *CBX7* was downregulated (Supplementary Table 21). In allele-specific expression (ASE) analysis, *CBX2*, *CBX8*, *ZNF208* and *HOXA13* were the most intriguing high-ranking genes (Fig. 3d, Supplementary Fig. 24, Supplementary Tables 23, 24). The *CBX2/4/8* locus was also highlighted through differential H2A.Z binding (Extended Data Figs. 5, 10), DARs (Supplementary Fig. 25), differential 3D interactions (Supplementary Fig. 26, Supplementary Table 15) and methylation-expression correlations (Supplementary Discussion, Fig. 4a, Supplementary Fig. 27a). In addition to *CBX2/4/8*, the strongest methylation-expression correlations included the developmental transcription factor genes *SATB2* and *HOXA13* (Supplementary Fig. 27b, c), which were also highlighted through overexpression and promoter-proximal DARs shared by all subclasses (Supplementary Tables 12, 21) and differential 3D interactions (Supplementary Table 15).

We also detected overexpression of *SRD5A2* and *HSD17B6* across all UL subclasses (Supplementary Table 21, Supplementary Fig. 28). These are key enzymes in synthesis of dihydrotestosterone, which can also be transformed into metabolites with oestrogen receptor affinity[25–27]. One of only eighteen promoter-proximal DARs that were shared between all UL subclasses genome-wide was at *HSD17B6* (Supplementary Table 12, Supplementary Fig. 28a) and multiple differentially methylated loci correlated with *HSD17B6* expression (Supplementary Fig. 28b). A genome-wide association study combining three independent cohorts detected a UL predisposition locus at *rs2277339*, which is only 11,039 base pairs from *HSD17B6* ($P = 1.5 \times 10^{-8}$) (Supplementary Fig. 29).

Thus, although subclasses of ULs had distinct global expression profiles, CBX family members and dihydrotestosterone-synthesizing enzymes were found to be dysregulated across all UL subclasses.

Discussion

We have carried out an extensive multi-omics characterization of ULs and identified inactivating mutations in SRCAP complex genes and a consequent H2A.Z loading defect as a potential driver of neoplasia (Fig. 4b). Immunohistochemistry and ChIP–seq demonstrated loss of H2A.Z in SRCAP ULs. Many, though not all, *HMGA*-overexpressing tumours displayed reduced H2A.Z. Possible mechanisms for this include competing binding of overexpressed HMGA proteins with SRCAP complex to AT-rich sequences[28]. ChIP–seq showed that H2A.Z in chromatin was reduced in MED12 tumours, especially at active TSS regions, but a strong H2A.Z signal was detected by IHC and western blotting. In IHC, an antigen-retrieval step enhances epitope presentation[29], whereas in ChIP–seq, epitope masking by surrounding molecules[30] could contribute to the reduction in H2A.Z. The possible relevance of H2A.Z to UL genesis in HMGA and MED12 tumours needs further investigation. Immunohistochemistry also displayed a reduction in H2A.Zac in SRCAP and HMGA ULs, and H2A.Z acetylation and ubiquitination in ULs is another area of interest.

Increased expression of *CBX2/4/8* and decreased expression of *CBX7* may be relevant across UL subclasses. These proteins serve as mutually exclusive components of the canonical polycomb repressor complex 1 (PRC1), a key development regulator[31]. CBX7 maintains pluripotency, whereas the introduction of subunits CBX2, 4 or 8 pushes cells from stemness towards differentiation[32]. The introduction of CBX2/4/8 to PRC1 at the expense of CBX7 could conceivably contribute to the activation of a deranged differentiation programme facilitated by increased expression of developmental transcription factors, such as *HOXA13* and *SATB2*, that have also previously been implicated in UL genesis[33,34] (Fig. 4b). Notably, MED12 has been proposed to work with CBX7–PRC1 to repress differentiation genes[35].

Compounds that enhance H2A.Z occupancy, such as metformin[36], may have an effect against some ULs. The use of metformin reduces UL risk, and it has been proposed that it works by inhibiting mTOR[37].

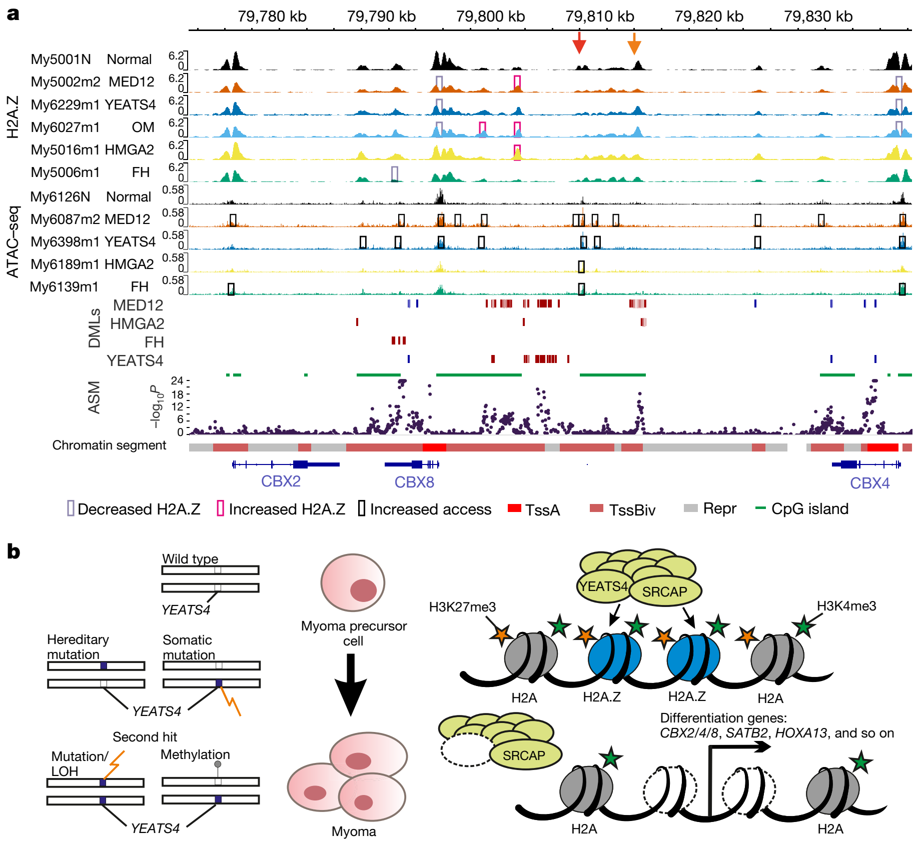


Fig. 4 | Chromatin changes emerge at the locus containing *CBX2*, *CBX4* and *CBX8* in ULs. a, H2A.Z ChIP fragment pileup from one representative normal and tumour from each subclass, ATAC-seq Tn5 insertion counts from one representative normal sample and tumour from each subclass, differentially methylated loci (DMLs) (red hypermethylation, blue hypomethylation) and allele-specific methylation (ASM) association to *CBX8* allelic expression at *CBX2*/*4*/*8* locus. Pink and light grey squares depict increased and decreased H2A.Z binding sites (FDR < 0.05), respectively, detected in UL subclasses from the spike-in ChIP–seq experiments (Extended Data Figs. 5, 10). Dark grey squares depict DARs (FDR < 0.05) detected in UL subclasses from the ATAC–seq Tn5 insertion counts (Supplementary Fig. 25). Both *CBX8* and *CBX4* show reduced H2A.Z binding and DARs at TSSs in YEATS4 and MED12 tumours (Extended Data Fig. 6d, peaks S0 and S41). Red arrow, region with DAR in all tumour subclasses (one of only 156 such sites in the whole genome). Orange arrow, region where increased methylation correlated notably with increased expression of *CBX2*, *CBX4* and *CBX8*. **b**, Model for genesis of *YEATS4*-mutated UL. Myoma precursor cells maintain their stem cell-like status through appropriate regulation (repression) of bivalent developmental genes. In the hereditary setting (left), each cell contains one mutant *YEATS4* allele, facilitating UL genesis at a younger age. In the sporadic setting, loss of both *YEATS4* alleles leads to epigenetic instability through defective H2A.Z deposition and initiation of a tumorigenic differentiation program including overexpression of *CBX2*/*4*/*8* and developmental transcription factors such as *SATB2* and *HOXA13*. Grey, canonical nucleosomes; blue, variant nucleosomes. LOH, loss of heterozygosity; H3K27me3 and H3K4me3, trimethylation of lysines 27 and 4, respectively, of histone H3.

Whether enhanced H2A.Z occupancy is involved remains to be examined. H2A.Z is involved in transferring cues from oestrogen and androgens[4,5]. *HSD17B6* and *SRD5A2* were strongly upregulated across UL subclasses, and inhibition of these sex hormone-synthesizing enzymes might offer another treatment option if appropriately validated, similar to the treatment of prostate hyperplasia in men[38].

Although our study firmly associates SRCAP complex gene mutations with the genesis of UL, further research is required to unravel the mechanisms of this effect. Loss of H2A.Z could provide the UL precursor cells with increased, albeit confined, plasticity; this is capable of promoting a benign lesion, but perhaps provides little potential for malignancy, as SRCAP complex genes are not listed in the Cancer Gene Census (COSMIC v91)[39]. The roles of H2A.Z and bivalent chromatin[40–42] might be particularly important in myometrium, which responds rapidly to external cues such as those that govern the menstrual cycle and pregnancy. Thus, when the epigenetic code at bivalent regions becomes aberrant, myometrium might be more likely than most other tissues to become prone to neoplastic degeneration. This dynamic character of myometrium could contribute to the exceptionally high prevalence of UL.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03747-1.

1. Wallach, E. E., Buttram, V. C. Jr & Reiter, R. C. Uterine leiomyomata: etiology, symptomatology, and management. *Fertil. Steril.* **36**, 433–445 (1981).
2. Gurusamy, K. S., Vaughan, J., Fraser, I. S., Best, L. M. J. & Richards, T. Medical therapies for uterine fibroids — a systematic review and network meta-analysis of randomised controlled trials. *PLoS ONE* **11**, e0149631 (2016).
3. Mehine, M., Mäkinen, N., Heinonen, H.-R., Aaltonen, L. A. & Vahteristo, P. Genomics of uterine leiomyomas: insights from high-throughput sequencing. *Fertil. Steril.* **102**, 621–629 (2014).
4. Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B. & Borggrefe, T. The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin* **12**, 37 (2019).
5. Hsu, C.-C. et al. Gas41 links histone acetylation to H2A.Z deposition and maintenance of embryonic stem cell identity. *Cell Discov.* **4**, 28 (2018).
6. Baird, D. D., Dunson, D. B., Hill, M. C., Cousins, D. & Schectman, J. M. High cumulative incidence of uterine leiomyoma in black and white women: ultrasound evidence. *Am. J. Obstet. Gynecol.* **188**, 100–107 (2003).

7. Okolo, S. Incidence, aetiology and epidemiology of uterine fibroids. *Best Pract. Res. Clin. Obstet. Gynaecol*. **22**, 571–588 (2008).
8. Bertsch, E. et al. *MED12* and *HMGA2* mutations: two independent genetic events in uterine leiomyoma and leiomyosarcoma. *Mod. Pathol*. **27**, 1144–1153 (2014).
9. Lehtonen, R. et al. Biallelic inactivation of fumarate hydratase (FH) occurs in nonsyndromic uterine leiomyomas but is rare in other tumors. *Am. J. Pathol*. **164**, 17–22 (2004).
10. Hua, S. et al. Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression. *Mol. Syst. Biol*. **4**, 188 (2008).
11. Dunican, D. S., McWilliam, P., Tighe, O., Parle-McDermott, A. & Croke, D. T. Gene expression differences between the microsatellite instability (MIN) and chromosomal instability (CIN) phenotypes in colorectal cancer revealed by high-density cDNA array hybridization. *Oncogene* **21**, 3253–3257 (2002).
12. Yang, B. et al. H2A.Z regulates tumorigenesis, metastasis and sensitivity to cisplatin in intrahepatic cholangiocarcinoma. *Int. J. Oncol*. **52**, 1235–1245 (2018).
13. Vardabasso, C. et al. Histone variant H2A.Z.2 mediates proliferation and drug sensitivity of malignant melanoma. *Mol. Cell* **59**, 75–88 (2015).
14. Hsu, C.-C. et al. Recognition of histone acetylation by the GAS41 YEATS domain promotes H2A.Z deposition in non-small cell lung cancer. *Genes Dev*. **32**, 58–69 (2018).
15. Bellucci, L., Dalvai, M., Kocanova, S., Moutahir, F. & Bystricky, K. Activation of p21 by HDAC inhibitors requires acetylation of H2A.Z. *PLoS ONE* **8**, e54102 (2013).
16. Ku, M. et al. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biol*. **13**, R85 (2012).
17. Tomlinson, I. P. M. et al. Germline mutations in *FH* predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nat. Genet*. **30**, 406–410 (2002).
18. Takahashi, D. et al. Quantitative regulation of histone variant H2A.Z during cell cycle by ubiquitin proteasome system and SUMO-targeted ubiquitin ligases. *Biosci. Biotechnol. Biochem*. **81**, 1557–1560 (2017).
19. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
20. Ye, B. et al. Suppression of SRCAP chromatin remodelling complex and restriction of lymphoid lineage commitment by Pcid2. *Nat. Commun*. **8**, 1518 (2017).
21. Bowman, T. A., Wong, M. M., Cox, L. K., Baldassare, J. J. & Chrivia, J. C. Loss of H2A.Z is not sufficient to determine transcriptional activity of Snf2-related CBP activator protein or p400 complexes. *Int. J. Cell Biol*. **2011**, 715642 (2011).
22. Slupianek, A., Yerrum, S., Safadi, F. F. & Monroy, M. A. The chromatin remodeling factor SRCAP modulates expression of prostate specific antigen and cellular proliferation in prostate cancer cells. *J. Cell. Physiol*. **224**, 369–375 (2010).
23. Murphy, K. E., Meng, F. W., Makowski, C. E. & Murphy, P. J. Genome-wide chromatin accessibility is restricted by ANP32E. *Nat. Commun*. **11**, 5063 (2020).
24. Brunelle, M. et al. The histone variant H2A.Z is an important regulator of enhancer activity. *Nucleic Acids Res*. **43**, 9742–9756 (2015).
25. Andersson, S., Berman, D. M., Jenkins, E. P. & Russell, D. W. Deletion of steroid 5 α-reductase 2 gene in male pseudohermaphroditism. *Nature* **354**, 159–161 (1991).
26. Bauman, D. R., Steckelbroeck, S., Williams, M. V., Peehl, D. M. & Penning, T. M. Identification of the major oxidative 3α-hydroxysteroid dehydrogenase in human prostate that converts 5α-androstane-3α,17β-diol to 5α-dihydrotestosterone: a potential therapeutic target for androgen-dependent disease. *Mol. Endocrinol*. **20**, 444–458 (2006).
27. Weihua, Z., Lathe, R., Warner, M. & Gustafsson, J.-A. An endocrine pathway in the prostate, ERβ, AR, 5α-androstane-3β,17β-diol, and CYP7B1, regulates prostate growth. *Proc. Natl Acad. Sci. USA* **99**, 13589–13594 (2002).
28. Solomon, M. J., Strauss, F. & Varshavsky, A. A mammalian high mobility group protein recognizes any stretch of six A.T base pairs in duplex DNA. *Proc. Natl Acad. Sci. USA* **83**, 1276–1280 (1986).
29. Fowler, C. B., Evers, D. L., O'Leary, T. J. & Mason, J. T. Antigen retrieval causes protein unfolding: evidence for a linear epitope model of recovered immunoreactivity. *J. Histochem. Cytochem*. **59**, 366–381 (2011).
30. Kidder, B. L., Hu, G. & Zhao, K. ChIP-seq: technical considerations for obtaining high-quality data. *Nat. Immunol*. **12**, 918–922 (2011).
31. Kim, J. & Kingston, R. E. The CBX family of proteins in transcriptional repression and memory. *J. Biosci*. **45**, 16 (2020).
32. Klauke, K. et al. Polycomb Cbx family members mediate the balance between haematopoietic stem cell self-renewal and differentiation. *Nat. Cell Biol*. **15**, 353–362 (2013).
33. George, J. W. et al. Integrated epigenome, exome, and transcriptome analyses reveal molecular subtypes and homeotic transformation in uterine fibroids. *Cell Rep*. **29**, 4069–4085.e6 (2019).
34. Sato, S. et al. SATB2 and NGR1: potential upstream regulatory factors in uterine leiomyomas. *J. Assist. Reprod. Genet*. **36**, 2385–2397 (2019).
35. Papadopoulou, T., Kaymak, A., Sayols, S. & Richly, H. Dual role of Med12 in PRC1-dependent gene repression and ncRNA-mediated transcriptional activation. *Cell Cycle* **15**, 1479–1493 (2016).
36. Tyagi, M., Cheema, M. S., Dryhurst, D., Eskiw, C. H. & Ausió, J. Metformin alters H2A.Z dynamics and regulates androgen dependent prostate cancer progression. *Oncotarget* **9**, 37054–37068 (2018).
37. Tseng, C.-H. Metformin use is associated with a lower risk of uterine leiomyoma in female type 2 diabetes patients. *Ther. Adv. Endocrinol. Metab*. https://doi.org/10.1177/2042018819895159 (2019).
38. Gormley, G. J. et al. The effect of finasteride in men with benign prostatic hyperplasia. *N. Engl. J. Med*. **327**, 1185–1191 (1992).
39. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. **47**, D941–D947 (2019).
40. Ikeda, H., Sone, M., Yamanaka, S. & Yamamoto, T. Structural and spatial chromatin features at developmental gene loci in human pluripotent stem cells. *Nat. Commun*. **8**, 1616 (2017).
41. Ku, M. et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet*. **4**, e1000242 (2008).
42. Surface, L. E. et al. H2A.Z.1 monoubiquitylation antagonizes BRD2 to maintain poised chromatin in ESCs. *Cell Rep*. **14**, 1142–1155 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Experimental models and subject details

The study was conducted in accordance with the Declaration of Helsinki and approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00/2017, THL/723/5.05.00/2018), and the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/2509/2016). The sample set consisted of six prospectively collected sample series (M, My, My1000, My5000, My6000 and My8000). The anonymous M-sample series was collected according to Finnish laws and regulations after authorization from the director of the health care unit, between the years 2001 and 2002. For all subsequent samples, written informed consent was obtained. Participants were not compensated. Samples were collected after hysterectomy at the department of pathology, and tumours down to 4 mm in diameter were derived as systematically as possible. Leiomyomatous tumours containing an epithelioid and endometrioid stromal component (adenomyomas) were excluded from the study. To minimize risk of contamination by normal myometrium and to ascertain the leiomyomatous nature of the tumours, all samples were collected and reviewed by a pathologist with expertise in gynaecological pathology (R.B., A.P.). By both histological and immunohistochemical estimation (desmin and caldesmon negativity), the percentage of non-smooth cell compartment did not exceed 20%. Tumours with ischaemic necrosis and/or pronounced inflammation were excluded from the study. Surrounding myometrium, peritoneal and subperitoneal tissue and endometrium were carefully removed from the samples.

Altogether, 2,263 uterine leiomyomas and 728 corresponding myometrium samples were successfully used in this study.

The biological materials are limited-sized primary tumour samples. Any request related to obtaining samples needs to be evaluated individually in light of the relevant local rules and regulations, and unrestricted access cannot be guaranteed.

SNP array

DNA was extracted using a QIAamp FAST DNA Tissue Kit (Qiagen). A total of 2,263 individual tumours and their respective normal myometria were genotyped using either Illumina Infinium HumanCore-24 or HumanOmni 2.5-8 chips. B-allele frequencies (BAFs) and log-R ratios (LRRs) were extracted with Illumina GenomeStudio, GC wave adjustments with PennCNV (v. 1.0.4)[43], and allelic imbalance (AI) segments with BAF segmentation (v. 1.2.0). Segments with at least eight heterozygous SNPs and a proportion of ≥3% heterozygous markers were kept: segments with LRR ≥ 0.05 and mirrored-BAF (mBAF) ≥ 0.56 denoted regions of allelic gain, and LRR < 0.05 and mBAF ≥ 0.56 denoted loss, including copy-neutral loss of heterozygosity. Subclass-specific enrichment of loss was assessed with one-sided Wald tests after fitting a generalized estimating equations (GEE) model (R package geepack v. 1.3-2), including the sum of AI segment length over other chromosomes as a covariate. The model accounted for unknown dependencies between observations (multiple tumours per patient): exchangeable correlation structure, binomial random component and logit link function were assumed. Numbers of double-strand breaks (DSBs) were estimated by merging together any adjacent (<1-Mb window) AI segments with an absolute LRR difference of <0.1 and counting the remaining segment ends (telomere ends excluded).

Detection of the three known UL subclasses

All 2,263 tumours entered analysis for known UL drivers in the following sequential stages. *MED12* mutation screening of exons 1 and 2, the loci exclusively containing the MED12 mutations detected in UL[44,45], was performed by Sanger sequencing. *HMGA2* overexpression was assessed in all *MED12* mutation-negative and 103 *MED12*-mutated tumours using *HMGA2* (Hs04397751) and *RNA18S5* (Hs03928990_g1) TaqMan probes and the 7500 Fast Real-Time PCR System (Thermo Fisher Scientific). The qPCR expression analysis was carried out according to the manufacturer's instructions. Samples were analysed in triplicate and only tumours showing expression of the *RNA18S5* reference gene were included in the analysis. All runs contained a previously assessed UL sample with weak *HMGA2* expression (negative control) and a tumour showing a relative *HMGA2* expression >100 (positive control) when compared to the negative control[46]. *HMGA2* overexpression was defined if the *HMGA2* expression was >100 when compared to the negative control. If a UL showed relative expression of 2–100, the HMGA2 expression status was further examined by immunohistochemistry (IHC). The HMGA2 IHC was performed as previously described[47]. Samples showing deletions at the *FH* locus in 1q on an SNP chip array entered *FH* Sanger screening covering all exons.

RNA-seq and driver mutation identification

For RNA-seq, we chose representative sets of samples belonging to the MED12, FH and HMGA2 groups, and all tumours for which no driver mutation was identified and RNA was available (the UNKNOWN group). The selection of tumours was mostly random, although a few samples were included because of their interesting patterns of allelic imbalance. RNA was extracted using TRIzol Reagent (Invitrogen) and DNase treated with RNase-free DNase set (Qiagen), and purified with RNeasy MiniElute clean-up kit (Qiagen). RNA-seq libraries were prepared according to the Illumina TrueSeq stranded total RNA with Ribo-Zero human library preparation according to the manufacturer's protocol (Illumina). We sequenced 130 samples with Illumina HiSeq 2500 in SciLife and 308 samples with NovaSeq 6000 in Macrogen. Altogether, 276 myoma and 162 myometrium samples were sequenced. One patient contributed 1–3 tumours to the analysis and altogether 187 tumours had their corresponding myometrium sample sequenced. First, the quality and adaptor trimming were done with Trim Galore (v. 0.5.0). Then, the reads were aligned to the human reference genome (GRCh37) with HISAT2 (v.2.1.0)[48] with the parameter -dta and RNA-strandness set to RF. The aligned reads were assembled to transcripts with StringTie (v.1.3.4d)[49] with parameters -e and RNA-strandness -rf using the ensembl GRCh37.75 reference annotation (source: ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/) to guide the assembly. Finally, the read counts for each gene were calculated with prepDe.py. To allow the methylation expression correlation analysis, the reads were also aligned to the GRCh38 reference genome and refseq reference annotation (accession: GCA_000001405.15, release: 109.20190125). hisat2_extract_splice_sites.py and hisat2_extract_exons.py were used to preprocess the annotation.

RNA-seq data were analysed for the presence of somatic variants, to search for novel driver mutations. For the variant calling, samples with corresponding normal samples were preprocessed with picard (v.2.8.16) AddOrReplaceReadGroups and MarkDuplicates. The GATK[50] (v. 3.5) SplitNCigarReads (with the option ALLOW_N_CIGAR_READS), BaseRecalibrator and PrintReads were used. DBSNP v.137 and 1000G phase1 and gold standard indels were marked as known sites. The variants were called with MuTect2 in tumour-normal mode. The variants were further filtered to have a minimum coverage of 4 and minimum allele fraction of 10%. Samples with no corresponding normal samples were analysed with picard (v.2.8.16) and GATK (v. 4.1.4.1). Gnomad (v.3.1.1) was used as a germline variant database. To account for the lack of corresponding normal samples, stricter filtering was used: variants were required to have a minimum coverage of 6 and minimum allele fraction of 25%. Observed mutations with the highest potential biological relevance, and their somatic nature, were confirmed by Sanger sequencing. We also screened *YEATS Domain Containing 4* (*YEATS4*, also called *GAS41*), the most frequently mutated gene in the RNA-seq data, by Sanger sequencing in all UNKNOWN samples as well as a set of MED12 and HMGA2 samples. These analyses yielded a fourth driver mutation group, tumours with a mutation in genes belonging to the SRCAP histone H2A.Z loading complex. These tumours are referred to as SRCAP tumours. SRCAP tumours were further divided into YEATS4 and OM (other SRCAP member) subclasses, as the high number of YEATS4 tumours allowed stratified analyses. In addition to the observed somatic mutations, three tumours were classified as YEATS4 and one tumour as ZNHIT1 based on particularly low YEATS4 and ZNHIT1 expression, respectively (Supplementary Fig. 4c). These four tumours were found to cluster with somatically mutated YEATS4 and OM tumours in the RNA-seq analysis. One of the tumours (M23m8) had entered nanopore sequencing and was shown to have hypermethylation of one allele at the *YEATS4* promoter, compatible with an undetected DNA-level defect in the other allele. Two of the tumours had entered IHC analysis (My6136m3 and My6492m1), and both displayed negative H2A.Z and acH2A.Z IHC staining. A permutation test[44] with 1,000,000 permutations was used to test whether the SRCAP tumours aggregated to a subset of patients instead of being randomly distributed among the patients.

HMGA1 overexpression was occasionally seen with other driver changes. For the purpose of subsequent analyses, HMGA1-labelled tumours were identified using the following criteria: the tumour did not have any other known driver changes (*MED12* mutation, *FH* mutation, *HMGA2* overexpression, SRCAP complex mutation) and *HMGA1* expression was higher than the 75th percentile of the myometrium samples (Supplementary Fig. 2). In addition, one tumour with no RNA-seq data was categorized as *HMGA1*-overexpressing on the basis of previous analysis[46].

For the differential expression (DE) analysis, only genes with coverage of at least ten reads in at least ten samples were considered (28,042 genes). The DE analysis was performed with DESeq2 (v. 1.22.2)[51] using the option betaPrior = T to shrink the $\log_2$FC and including the sequencing batch as a confounder. DE was calculated for each subclass (*MED12* $n = 38$; *HMGA2* $n = 44$; *HMGA1* $n = 62$; *FH* $n = 15$: *YEATS4* $n = 16$; OM $n = 15$) against all myometrium samples so that only one tumour per patient per subclass entered the analysis (third column in Supplementary Table 2). The DE between all tumour subclasses combined against myometrium was calculated with the option listValues = c(1,-1/6) to weight each subclass equally; only one tumour per patient ($n = 240$) entered this analysis. For unsupervised hierarchical clustering, the effect of the sequencing batch was removed with limma (v. 3.42.0)[52] removeBatchEffect and 5% ($n = 1,355$) of the most variable genes processed through varianceStabilizingTransformation (blind = F) were used as follows. The data were scaled by subtracting the genes' mean for each gene over all tumours ($n = 257$) and ward. D2 linkage and 1 − correlation distance were used. Approximately unbiased P values for this unsupervised hierarchical clustering were obtained with pvclust (v.2.2-0) using 100,000 bootstrap iterations. Clusters that rejected the null hypothesis ('the cluster does not exist') at a significance level of 10% are reported in Fig. 3c. The most important genes driving the differences between the subclasses were selected using penalizedLDA[53] (v. 1.1) with lambda = 0.0535 and $K = 5$. YEATS4 and OM were considered as one group. The resulting 426 genes are presented in the heat map (Extended Data Fig. 8b). As an alternative unsupervised clustering method, 10% of the most variable genes in one tumour per patient per subtype ($n = 257$) and all myometrium samples ($n = 162$) were consensus clustered with ConsensusClusterPlus (v. 1.50.0)[54,55] using 1,000 resamplings, the maximum cluster count 26, and average linkage.

Analysis of mutual exclusivity of driver changes

Extended Data Figure 1a summarizes the numbers of tumours screened per subclass: *MED12* mutations were screened from all tumours (exon 1 and 2 Sanger sequencing). *HMGA2* overexpression was assessed from a total of 655 tumours: 510 *MED12* mutation-negative tumours and 145 MED12 tumours (Supplementary Table 2). *YEATS4* mutations were screened from a total of 638 tumours (all Sanger, whole genome sequencing (WGS) and RNA-seq data combined; Supplementary Table 2). The other SRCAP complex genes were inspected for mutations from all WGS and RNA-seq samples ($n = 332$ tumours), and *HMGA1* overexpression was assessed among the RNA-seq samples ($n = 276$). The *FH* locus was inspected for any allelic losses from all SNP array samples ($n = 2,186$).

A cross-comparison of mutation statuses was performed to determine whether the known driver mutations are mutually exclusive. All HMGA2 tumours were screened for *MED12* mutations, and *HMGA2* overexpression was assessed from a random sample of MED12 tumours. FH was screened in additional MED12 and HMGA2 tumours. To show that *MED12* mutations, *HMGA2* overexpression, FH deficiency and *YEATS4* mutations are mutually exclusive, *YEATS4* was screened in a random sample covering all these subclasses. Here, at most one tumour per patient was counted in each comparison to avoid patient-specific confounding. See Supplementary Discussion for details of the mutual exclusivity of the UL driver changes.

Analysis of germline loss-of-function variants

UK Biobank (UKB) WES data were processed in two separate batches. Data for the first 50,000 WES individuals became available in March 2019 (referred to as the discovery set) and for an additional 150,000 WES individuals later in October 2020 (referred to as the replication set). Similar to a previous publication[56], we extracted European ancestry samples based on the first two genetic principal components (UKB data-field 22009), resulting in a total of 46,969 and 142,732 individuals from the discovery and replication sets, respectively. These subsets covered 94.6% of all the WES samples and were predominantly self-reported 'white British' ($n = 175,740$), 'any other white background' ($n = 6,478$) or 'white Irish' ($n = 5,407$). Our association analysis was restricted to females only (data-fields 22001 and 31), thus, a total of 25,506 and 78,905 females of European ancestry remained in the discovery and replication sets, respectively. UL phenotype was defined on the basis of ICD10 diagnoses (D25), ICD9 diagnoses (2189), self-reported ULs (uterine fibroids) and self-reported operations (myomectomy/fibroids removed): 2,437 and 6,043 UL cases were identified in the discovery and replication sets, respectively (UKB data accessed on December 2019). Genetic relationship matrices were precomputed as follows. High-quality genotypes that passed UKB data QC (UKB resource 1955; 'input for phasing' $n = 688,581$ SNPs) were pruned with PLINK (window size of 50 SNPs, shift of 5 SNPs at each step and variance inflation factor threshold of 2). The remaining 372,671 pruned, autosomal SNPs were then used to initialize the genetic relationship matrices and to fit the null logistic mixed model (SAIGE-GENE v.0.42.1)[57] with default settings and covariates for age at first assessment centre visit (data-field 21003) and first four principal components (data-field 22009).

Owing to known issues in the early 50,000 UKB WES variant calls (see 'UK Biobank – Exome Data Release FAQs', accessed on June 2020 (see https://www.ukbiobank.ac.uk/media/cfulxh52/uk-biobank-exome-release-faq_v9-december-2020.pdf (accessed December 2020)), all analyses presented here were implemented using the improved and unified pipeline (OQFE) variant call material (data-field 23155; accessed October 2020). Variant effect annotation was done with SnpEff (v.4.3t) using the database version GRCh38.86. Variants annotated as stop gained, frameshift, splice acceptor variant, splice donor variant, start lost or stop lost were included in the analysis and are referred to as loss-of-function (LoF) variants. A total of 133,458 LoF variants were annotated among the discovery set females. Gene-based association tests (SAIGE-GENE v0.42.1; SKAT-O test) were computed by grouping the LoF variants by gene name (HGNC; LoF variants affecting multiple genes were grouped according to their highest impact SnpEff annotation). Following ref. [57], SKAT-O tests were applied to MAF $\leq 1\%$ variants, which resulted in 128,173 LoF variants (MAF $\leq 1\%$) from 15,341 genes that passed minimum sum(MAC) ≥ 3 per gene. SRCAP complex genes were annotated to contain in total 44 LoF variants, and all but one *DMAP1* LoF variant had MAF $\leq 1\%$. A full summary of discovery stage, gene-based association tests is given in Supplementary Table 8.

The subsequent replication set analysis was restricted to any candidate gene associations (SKAT-O $P < 0.05/9$) identified in the discovery stage. Odds ratio (OR) was calculated based on the numbers of [(LoF UL cases)/(wild-type UL cases)]/[(LoF controls)/(wild-type controls)].

Detection of clonally related tumours

We examined the occurrence of clonally related tumours, as inclusion of these could bias subsequent analyses. The following criteria were used to detect possibly clonally related tumours within a uterus. For MED12 tumours, identical *MED12* mutations and at least one shared AI segment were required; for HMGA2 tumours, *HMGA2* overexpression and at least one shared AI segment; for UNKNOWN tumours, at least one shared AI segment; and for YEATS4 and OM tumours, the same complex gene mutation was deemed sufficient. In the FH subclass no evidence for clonal relationship emerged. Shared AI segments were determined as segments for which both start and end positions matched within 250 kb tolerance. From the clonally related tumour sets, one tumour was arbitrarily chosen for subsequent analyses.

Allele-specific gene expression analysis

ASE was analysed using phASER (v.1.1.1)[58] over pre-phased, imputed genotypes. The imputed genotypes were produced with the EAGLE2-PBWT pipeline and Haplotype Reference Consortium (r1.1) panel (Sanger imputation services; accessed on 3 December 2019). phASER analysis was run after MarkDuplicates (Picard v.2.18.16) using uniquely mapping read-pairs and requiring minimum RNA-seq base-quality of 10 and coverage of 16× at the heterozygous imputed genotypes. The phASER Gene AE tool was run to produce gene-level estimates of maternal and paternal haplotype counts: a binomial test of haplotype counts—at thresholds of $P < 10^{-4}$ and $\log_2$(fold-change between haplotypes) > 1—was used to determine significant allele-specificity. Haplotype counts with overlapping somatic AI were excluded from the analysis of allele-specific gene expression. The tumour-enrichment analysis included one tumour per patient to avoid patient-specific confounding, giving a total of 240 tumours and 160 normal myometria (2 unpaired normal samples excluded). Clonality of X-inactivation was analysed, using all 276 tumours and 160 normal myometria, by estimating the proportion of ASE genes over the X chromosome: bulk RNA-seq from tumour tissue was expected to display a clonal pattern of X-inactivation (Supplementary Fig. 24).

Pathway analysis

The pathway data were generated with Ingenuity Pathways Analysis (IPA) software (QIAGEN IPA Winter 2021 Release, Version 60467501). A dataset derived from the differential expression analysis between each UL subclass against normal myometrium samples ($n = 162$) was used in the pathway enrichment analysis. The subclasses were MED12 ($n = 38$), HMGA2 ($n = 44$), HMGA1 ($n = 62$), YEATS4 ($n = 16$), OM ($n = 15$) and FH ($n = 15$). For each comparison, the absolute $\log_2$FC threshold was set to >1 and the 500 most significantly differentially expressed genes in each subclass were selected based on Benjamini–Hochberg adjusted *P* value. The data were mapped into relevant pathways based on their functional annotation and known molecular interactions in Ingenuity's Knowledge Base (IPKB). The −log of *P* values were calculated by Fisher's exact test (−log cutoff of 1.3). The overall activation/inhibition states of canonical pathways were predicted based on a *z*-score algorithm. A positive *z*-score predicted activation and a negative *z*-score predicted inactivation of the enriched pathway[59].

The comparison analyses of canonical pathways relevant to multiple UL subclasses was performed using right-tailed Fisher's exact test *P* values (scores). The pathways with the highest total score across the set of subclasses were sorted to the top.

Illumina and Complete Genomics whole-genome sequencing

Whole-genome sequencing data from 88 leiomyomas, excluding clonally related tumours, and their corresponding myometrium samples were used in driver mutation screening and in mutual exclusivity analysis. The first WGS set was prepared and sequenced by Illumina and Complete Genomics paired-end sequencing services as previously described[60]. An additional 54 tumour–normal pairs entered WGS in later sequencing batches using Illumina paired-end sequencing protocol. All samples were prepared according to the manufacturer's instructions and fulfilling their sample quality and quantity criteria. Samples using the Illumina platform were sequenced on the HiSeq 2000/4000/X platform (details in Supplementary Table 2; sequencing instrument column) and prepared with TruSeq DNA PCR-free (Illumina) using 1 µg genomic DNA. Samples sequenced on the Complete Genomics platform were prepared following DNA nanoarray technology as previously described[61]. Supplementary Table 2 gives detailed batch assignment information, number of reads sequenced and a summary of sequencing coverage over all SRCAP complex genes (median and IQR of coverage; proportion of nucleotides with minimum 10× coverage). Illumina platform data were processed with two different pipeline versions as detailed in Supplementary Table 2: short-read paired-end alignment (bwa v.0.6.2 aln with parameters -n0.06 and -q5; or bwa v.0.7.12 mem with default parameters) onto the 1000 Genomes Project Phase 2 reference assembly hs37d5 was followed by PCR duplicate removal (SAMtools v.0.1.18; or Picard MarkDuplicates v.1.79), local indel realignment and base score quality recalibration using Genome Analysis Toolkit (GATK IndelRealigner and BaseRecalibrator v.2.3-9 or v.3.5-0). Somatic variant calling was performed using MuTect (v.2.2-25-g2a68eab; default parameters) and somatic insertion-deletion calling using either GATK SomaticIndelDetector (v.2.3-9-ge5ebf34) or VarScan (v.2.3). For Complete Genomics data, somatic variant calling was done as a service (Complete Genomics CGApipeline version 2.0.2.22–2.0.3.2; details in Supplementary Table 2) and summary of sequencing coverage was computed at a 100-kb resolution (CGApipeline, depth of coverage report). One of the tumours (My6153m1) had exceptionally low coverage at SRCAP complex genes and was discarded from the analysis. All SRCAP complex mutations identified in the WGS data were subsequently validated by Sanger sequencing of both normal myometrium and tumour DNA.

Immunohistochemistry

HMGA2 immunostaining was performed as described earlier using anti-HMGA2 antibody (dilution 1:2,000; Biocheck; 59170AP)[47]. The immunoreaction was classified into four groups: 0 = fully negative, (1) = single-cell positivity, 1 = low positivity, 2 = strong positivity. Only samples that showed strong HMGA2 positivity were considered overexpressed.

To examine whether H2A.Z loading was compromised in different UL subclasses, a set of 265 ULs containing 96 MED12, 68 HMGA2, 58 HMGA1, 19 YEATS4, 14 OM, and 10 FH tumours and six normal tissue sections were examined using antibodies for non-acetylated- and acetylated forms of H2A.Z, hypothesizing that unbound H2A.Z is degraded. Formalin fixed paraffin-embedded (FFPE) sections (5 µm) were immunostained with recombinant anti-histone H2A.Z (dilution 1:2,500; Abcam, ab150402) and rabbit polyclonal histone H2A.Z (acetyl Lys5/Lys7/Lys11) (1:500; GeneTex, GTX60813). The Orion two-component detection system (peroxidase, goat anti-rabbit/mouse IgG HRP (ready-to-use); WellMed BV, Cat. No. T100-HRP) was used for detection. The intensity of immunoreaction was classified into three groups: 0 = negative or weak, 1 = moderate, 2 = strong. Normal myometrium sections—typically showing strong H2A.Z intensity—were used as internal controls in each separate staining batch. The association between the immunohistochemistry staining intensity and tumour subclass was calculated by fitting a generalized estimating equations (GEE) model (R package geepack v. 1.3-1). The model accounted for dependent observations (multiple tumours per patient): exchangeable correlation structure, binomial random component and logit link function were assumed. Tumours with low staining intensity, 0 or 1, were compared to tumours with score 2 staining. Staining batch was used as a covariate. See full details in Supplementary Fig. 7.

Western blot

We cut 200 mg of frozen tissue into small pieces and processed them with the Subcellular Protein Fractionation Kit for Tissues (Thermo Scientific, Cat. No. 87790). Proteins from chromatin extract (18 µg) were separated by SDS–PAGE on a mini-Protean TGX precast gel 4–20% (Bio-Rad, Cat. No. 4561096). After transfer onto PVDF membrane, proteins of interest were detected by immunoblotting with the following antibodies: H2A.Z 1:500 (Merck, Cat. No. 07-594), H2A.Zac 1:5,000 (Genetex, Cat. No. GTX60813) and TBP 1:1,000 (Abcam, Cat. No. 51841). Blots were washed three times with PBS-T (0.1% Tween-20), then incubated with a secondary antibody: goat anti-rabbit 1:12,000 (Sigma, Cat. No. A6154) for H2A.Z and H2A.Zac, goat anti-mouse 1:10,000 (Sigma, Cat. No. A4416) for TBP. After three washes with PBS-T (0.1% Tween-20), blots were developed using Amersham ECL Prime Western Blotting Detection kit (GE Healthcare, Cat. No. RPN2232). MagicMark XP Western Protein Standard (ThermoFisher, Cat. No. LC5603) was used as a ladder for the signal detection and Precision Plus Protein Kaleidoscope Prestained Protein Standard (Bio-Rad, Cat. No 1610375) as a ladder for electrophoresis.

ChIP–seq

We cut 200 mg of frozen tissue into small pieces and subsequently pulverized them using a Covaris CP02 (5 strokes with power level 5). Tissue powder was crosslinked with formaldehyde (1% for H3K27ac, H3K4me3 and H3K27me3; 0.15% for H2A.Z) for 10 min at room temperature (RT) and the reaction stopped by adding 0.125 M glycine, in the presence of protease inhibitor cOmplete (Roche, Cat. No. 11873580001). The pellet was washed in ice-cold PBS (for H3K27ac, H3K4me3 and H3K27me3) or 20 mM NaCl PBS (for H2A.Z) twice and then processed as follows.

For H2A.Z and H3K27ac: resuspended in lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40, 1× protease inhibitor cOmplete). The lysate was transferred to a Dounce tissue grinder tube and thoroughly homogenized, then resuspended in 800 µl RIPA buffer (1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 1 mM EDTA, 140 mM NaCl, 10 mM Tris-HCl pH 8, 1× protease inhibitor cOmplete). The chromatin was sonicated to an average fragment length of 200–500 bp using Sonicator Misonix S-4000 with the following parameters: amplitude 30%, pulse-on time 30 s, pulse-off time 60 s, for a total sonication time of 10 min.

For H3K4me3 and H3K27me3: resuspended in micrococcal lysis buffer (10 mM NaCl, 0.4% NP40, 3 mM $MgCl_2$, 10 mM Tris pH 7.8, 1× protease inhibitor cOmplete), transferred to a Dounce tissue grinder tube and thoroughly homogenized. Lysate was centrifuged for 5 min at 500*g* at 4 °C and pellets resuspended in micrococcal lysis buffer supplemented with 5 mM CaCl and 30 U micrococcal nuclease (Thermo Fisher Scientific, 88216). The reaction was incubated for 10 min at 37 °C on a thermomixer with 500 rpm shaking, then stopped with the addition of 20 mM EGTA. The nuclei were pelleted by centrifugation for 5 min at 500*g* at 4 °C and resuspended in 800 µl RIPA buffer. Samples were sonicated for 30 s with 20% power, using Sonicator Misonix S-4000.

All samples were then centrifuged at 18,000*g* for 15 min at 4 °C and supernatant collected. Dynabeads protein-A (ThermoFisher Scientific, Cat. No 10002D) were washed with 0.05% Tween-20 in PBS. Seven micrograms of antibodies against H3K27ac (Abcam, Cat. No. ab4729), H2A.Z (Merck, Cat. No. ABE1348 and Abcam, Cat. No. ab150402), H3K4me3 (Abcam, Cat. No. ab8580) and H3K27me3 (Cell Signaling, Cat. No. 9733S) were incubated with the beads for 15 min on a rotator at RT. For spike-in experiments, 2 µg of spike-in antibody (Active Motif, Cat. No. 61686) was incubated together with the antibody of interest. For each sample, 50 µl of sonicated chromatin was taken as input fraction and 50 µl as control for fragmentation efficiency and quantification. Twenty-five micrograms of chromatin was incubated with the antibody-coupled magnetic beads on a rotator overnight at 4 °C. For spike-in experiments, 20 ng spike-in chromatin (Active Motif, Cat. No. 53083) was mixed with 25 µg human chromatin before antibody incubation. After incubation, the beads were washed twice with low-salt washing buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 150 mM NaCl), twice with high-salt washing buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM NaCl), once with lithium chloride washing buffer (1% NP-40, 1% sodium deoxycholate, 1 mM EDTA, 20 mM Tris-HCl pH 8.0, 500 mM LiCl) and twice with 1× TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA). Antibody-bound chromatin samples were eluted from the beads by incubating for 1 h at 65 °C with shaking at 1,300 rpm in 200 µl IP elution buffer (1% SDS, 0.1 M NaHCO3, 10 mM Tris-HCl pH 7.5) and crosslinking was reversed by overnight incubation at 65 °C. The eluted DNA was purified using the phenol–chloroform method, followed by library preparation for 100-bp single-end Illumina sequencing.

Raw sequencing reads were quality- and adaptor-trimmed with cutadapt version 1.16 in Trim Galore version 0.3.7 using default parameters. Trimmed reads were aligned to the hs37d5 reference genome using Bowtie 2 (version 2.1.0) and reads with mapping quality <20 were filtered out with samtools (version 1.7). Peak calling for sample- and antibody-specific reads with mapping quality >20 was performed with MACS2 (version 2.1.2) with default parameters, except that FDR cutoff for narrowPeak regions was set to 0.01. Each sample included in the study was required to have at least 5% of reads in broadPeaks. The ENCODE blacklist genomic regions[62] were filtered out from the final narrowPeaks located at autosomes and the X chromosome.

ChIP signals for H2A.Z, H3K27ac, H3K27me3 and H3K4me3 in myometrium were measured by pooling data from normal specimens to get a high signal-to-noise ratio devoid of sample-to-sample variability. Peak calling for pooled normal samples was performed using merged reads after duplicate removal (samtools version 1.7) and running MACS2 (version 2.1.2) with default parameters except for --keep-dup parameter, which was set to the value that corresponded to the number of pooled normals. The ENCODE blacklist genomic regions[62] were filtered out from the final narrowPeaks located at autosomes and the X chromosome.

Enrichment of H2A.Z narrowPeak regions from pooled normal samples on specific chromatin regions was analysed with Locus Overlap Analysis (LOLA) R package (version 1.8.0)[63]. All regions that had read coverage of at least two in the aligned ChIP–seq data and did not overlap with the ENCODE blacklist genomic regions were considered as a background set. Region set databases used in this study included Roadmap Epigenomics ChIP– and DNase–seq data provided in the LOLA extended and core databases, as well as chromatin states provided by the Roadmap Epigenomics project and derived from myometrium specimens. For this purpose, Roadmap epigenomics data from 21 cell types representing smooth muscle cells, fibroblasts, mesenchymal stem cells, ES cells, ES cell-derived mesoderm cells, breast myoepithelial cells and placenta cells were used. The chromatin states provided by Roadmap Epigenomics were learned by ChromHMM v.1.10[64] using data corresponding to five chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3).

Additional, in-house segmentation was run using ChromHMM (v.1.21) on pooled myometrium data from H3K27ac ($n = 4$), H3K27me3 ($n = 6$) and H3K4me3 ($n = 7$) ChIP–seq samples. The binarization step was run with a shift value of 150 to account for the fragment size, and the rest of the parameters were kept at default. These three markers enabled us to train a five-state model to estimate the following chromatin states: (1) repressed chromatin (H3K27me3 only), (2) bivalent TSS (H3K4me3 and H3K27me3), (3) active TSS (H3K4me3 and H3K27ac), (4) other active chromatin (H3K27ac only) and (5) other chromatin (none of the three).

Clustering and differential binding analysis of H2A.Z ChIP–seq samples were performed with DiffBind[65] version 2.14.0 in R 3.6.3 (data without spike-in) and version 3.0.6 in R 4.0.3 (spike-in data). Duplicate reads were removed with samtools (version 1.7) rmdup from the ChIP alignment files that were used in the DiffBind analyses. For principal components analysis (PCA), read counts in peaks for each sample were counted with dba.count-function with parameters minOverlap = 1 and summits = 250, which includes in the peak set binding sites defined as 250 bp up- and downstream from the summit detected in all samples. Normalized read counts for the binding sites in the peak set were visualized with dba.plotPCA-function. For correlation clustering and differential binding analysis, read counts in peaks for each sample were counted using dba.count-function with parameter minOverlap = 2, which included in the peakset all narrowPeaks that overlapped in at least two samples, and these consensus peaks were re-centred around a consensus summit including 250 bp up- and downstream of the summit. The resulting peaksets without and with spike-in consisted of 59,647 and 84,663 peaks, respectively. Reads in the peaks were counted by extending reads to average fragment length that were derived from MACS. Then the number of reads in the respective input DNA was subtracted from the number of reads in ChIP at each peak region from each sample, and normalized using the total number of reads to make all the tumour and normal samples in the analysis comparable. In the normalization of the spike-in data, total numbers of reads were derived from reads aligned to the *Drosophila* genome (BDGP6). Fold change and FDR for differential H2A.Z binding were derived from DESeq2 analysis as implemented in the DiffBind[65] R package. Normalized read counts for the binding sites in the peakset were visualized with dba.plotHeatmap -function. Correlation clustering using default scoring was performed with DiffBind version 2.14.0 while binding affinity heatmaps using score = DBA_SCORE_NORMALIZED were created with DiffBind version 3.0.6.

ATAC–seq

ATAC–seq was performed on 4 myomas for each of the MED12, HMGA2, FH and YEATS4 subclasses, as well as 15 normal myometria. Following the omni-ATAC–seq protocol[66], 50,000 nuclei were taken for each transposition. ATAC was performed simultaneously in triplicate for each sample (150,000 nuclei in total), and DNA fragments were pooled after transposition. DNA fragments were purified with Zymo DNA clean and Concentrator-5 kit (Nordic BioSite, Cat. No. D4014) and amplified by PCR. Libraries were quantified by qPCR (KAPA library quant. kit, Illumina, Cat. No. KK4824) and quality measured by TapeStation 2200 (Agilent Technologies). ATAC–seq libraries were sent for 100-bp paired-end Illumina sequencing, aiming for 80 million paired-end reads.

Raw sequencing reads were quality and adaptor trimmed with cutadapt version 1.16 in Trim Galore version 0.3.7 using default parameters. Trimmed reads were aligned to hs37d5 reference genome using Bowtie 2 (version 2.1.0). Samtools (version 1.8) was used to filter out reads with mapping quality <20, to count reads that mapped to the mitochondrial genome and to remove PCR duplicates. Peak calling for deduplicated reads aligned on autosomes and the X chromosome with mapping quality ≥ 20 was performed with MACS2 (version 2.1.2). The quality of the ATAC–seq data, including TSS enrichment and fraction of reads in peaks (FRiP), was checked using ataqv[67]. We included only samples that passed ATAC–seq quality criteria (FRiP > 5% and TSS enrichment > 3). Fixed-width peaks were generated using MACS2 callpeak command with parameters '--shift 75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -p 0.01'. The resulting peak summits were extended to both directions by 250 bp and blacklist genomic regions were filtered out. Furthermore, MACS2 peak scores ($-\log_{10}P$) for each sample were converted to a 'score per million' in a similar manner to that previously described[68]. In short, this score normalization was performed by dividing each individual peak score by the sum of all of the peak scores in the given sample divided by 1 million.

Clustering of ATAC–seq samples was performed with DiffBind[65] (version 2.14.0) in R 3.6.3. Peak sets were read in by creating a dba-object with default parameters except filter = 5, to exclude peaks with normalized peak score below five from analysis. Read counts in peaks for each sample were counted with dba.count-function with parameter minOverlap = 2, which includes only peaks present in at least 2 samples in the peak set. Normalized read counts for the binding sites in the peak set were visualized with dba.plotPCA -function.

Differential accessibility was calculated separately for four tumours from each subclass, against 15 normal samples. For each tumour and normal sample, a single-sample peak set was first generated by removing overlapping fixed-width peaks using an iterative removal procedure that keeps the highest scoring peak, as previously described[68]. Reproducible peaks for each subclass were then generated by selecting peaks that were shared by at least two out of four tumours with a score per million value ≥ 5. Counts matrices representing the number of Tn5 insertions in each of the 19 samples (four tumours and 15 normal) were calculated for the subclass-specific reproducible peak sets. To obtain the number of Tn5 insertions, first the properly paired paired-end read alignments were corrected for the Tn5 offset ('+' stranded +4 bp, '−' stranded −5 bp)[69]. Then the corrected insertion sites (ends of fragments) were counted per peak. Finally, differential accessibility using the counts matrix was calculated with DESeq2 v.1.14.1[51] in R v.3.3.3 with default settings and Wald statistics. From the subclass-specific reproducible peak set analysis, regions with BH-adjusted *P* value (FDR) < 0.05 and FC > 0 were considered as differentially more accessible regions (DARs) in the subsequent analyses. Overlap analysis was performed with HOMER mergePeaks requiring a maximum 200-bp distance between centre positions of DARs (-d 200 parameter). Annotation of DARs was performed with HOMER annotatePeaks.pl against hg19 v.6.4 RefSeq database. Enrichment of DARs on specific chromatin regions was analysed with Locus Overlap Analysis (LOLA) R package (version 1.8.0)[63]. The same region set databases were used as described for ChIP–seq. All subclass-specific reproducible peaks were considered as a background set.

Differential accessibility was also calculated for H2A.Z narrow peak regions from pooled normal samples as well as for the H2A.Z peaks from the spike-in data by comparing corrected insertion sites (ends of fragments) from tumours of each subclass ($n = 4$) against 15 normal samples at these regions. For this, DESeq2 v.1.26.0[51] in R v.3.6.3 was used with default settings and Wald statistics. Differentially accessible H2A.Z regions from the analysis of YEATS4 tumours versus normals were selected with FDR < 0.05 cutoff for LOLA enrichment analysis and for heatmap visualization.

HiChIP

The protocol was adapted from ref. [70]. Two-hundred milligrams of frozen tissue was pulverized and crosslinked in the same way as for ChIP (1% formaldehyde). After two washes with PBS, the pellet was resuspended in 1.5 ml ice-cold Hi-C lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.2% Igepal CA630 or NP40, 1× Roche cOmplete protease inhibitor), transferred to a Dounce tissue grinder tube and thoroughly homogenized. Lysate was aliquoted in 3 tubes and centrifuged at 500*g* for 5 min at 4 °C. Each pellet was resuspended in 333 µl 1.27× NEBuffer 3.1 containing 9 µl of 50× cOmplete protease inhibitor. Thirty-eight microlitres of 1% SDS (final concentration 0.1%) was added and samples incubated in a thermomixer at 65 °C for 10 min (800 rpm shaking). SDS was quenched by adding 43 µl of 10% Triton X-100 (final concentration 1%) and incubating in a thermomixer at 37 °C for 15 min (800 rpm shaking). Ten microlitres of lysate was taken as lysis control and replaced by 10 µl of 1× NEBuffer 3.1, then 3.7 µl of 10× NEBuffer, 13.3µl of water and 200 U of DpnII (NEB, Cat. No. R0543L) were added to the reaction. Digestion was performed overnight in a thermomixer at 37 °C with 500 rpm shaking. After enzyme inactivation, overhangs were filled in and marked with biotin-14-dATP (Life Technologies, Cat. No. 19524-016) for 4 h at 23 °C with slow rotation (each tube received 53 µl of mastermix containing 37.5 µl of 0.4 mM biotin-14-dATP (Life Technologies, Cat. No. 19524-016), 1.5 µl each of 10 mM dCTP, dGTP, dTTP, 1 µl of 50× cOmplete protease inhibitor and 10 µl of 5 U/µl DNA Polymerase I, Large Klenow (NEB, Cat. No. M0210)). Ten microlitres was taken as digestion control, then 2.5 ml of ligation master mix was added (1× NEB T4 DNA ligase buffer (NEB, Cat. No. B0202), 1% Triton X-100, 0.1 mg/ml BSA, 2,000 U/µl T4 DNA Ligase (NEB, Cat. No. M0202), 1× cOmplete protease inhibitor) and samples incubated overnight at 4 °C with slow rotation. Ten microlitres was taken as ligation control, then pellets combined and resuspended in 800 µl RIPA buffer containing Roche protease inhibitor cOmplete. Sonication, antibody incubation (H3K27ac, Abcam, Cat. No. ab4729) and washes were performed as for ChIP. After elution, crosslink reversion and proteinase K treatment, samples were purified with Zymo DNA clean and Concentrator-5 kit (Nordic BioSite, Cat. No. D4014). One hundred and twenty-five nanograms of DNA underwent pulldown with 50 µg beads conjugated with Streptavidin C-1 (Invitrogen, Cat. No. 65001) and subsequent tagmentation with 4 µl transposase Tn5 (Illumina, Cat. No. 20034198). Bead-conjugated DNA was amplified by PCR and libraries quantified by qPCR (KAPA library quant. kit, Illumina, Cat. No. KK4824) and quality measured by TapeStation 2200 (Agilent Technologies). Libraries were sent for 100-bp paired-end Illumina sequencing, aiming for 250 million paired end reads.

HiC-Pro v. 2.11.1[71] was used to identify valid interaction pairs from raw reads. Bowtie2 v.2.3.4.1 was used to map reads to reference genome hs37d5 with HiC-Pro default parameters and minimum mapping quality 10. After alignment, reads were paired and assigned to restriction fragment GATCGATC. Self-circle and dangling end fragments were discarded, and valid interaction pairs were detected. FitHiChIP[71,72] was used to call significant interactions from all valid pairs identified by HiC-Pro. Interactions were called with bin size 40 kb and FDR threshold 0.01 for significant loops. Coverage bias correction and peak to all (L) background were used. H3K27ac ChIP–seq narrow peaks were used to identify interactions related to binding of H3K27ac. Differential analysis of FitHiChIP significant loops between tumour–normal pairs representing each subclass was run with FitHiChIP using EdgeR v.3.26.5 and default parameters including --DiffFDRThr 0.05. FitHiChIP interactions and differential loops were visualized together with H3K27ac ChIP–seq profiles using HiCPlotter[73]. Bedtools v.2.29.1 intersect was used to identify DARs from the respective subclass that fall into bins containing differential interactions in the tumour identified with FitHiChIP. Annotation of DARs located in HiChIP bins containing differential interactions was performed with HOMER (v.4.10.4) annotatePeaks.pl against hg19 v.6.4 RefSeq database. Promoter-proximal regions were defined as TSS + 1 kb flank.

For significance analysis we considered bridging links (Supplementary Fig. 22b) as (1) significant in normal myometrium or in the tumour, (2) significantly differentially interacting (FDR < 0.05), (3) having a gene in one end of the interaction and (4) having a H2Az binding site in the other end of the interaction. Finally, we computed chi-squared statistics for independence of the count of genes with or without differential expression (FDR < 0.05) compared to having or not being involved with a bridging link. Only genes with observed expression in any of the tumour types were used. The *P* values were estimated by 10,000 permutations of the differential expression status among the genes.

Nanopore long-read sequencing

Long-read libraries were prepared for 106 ULs (13 MED12, 28 HMGA2, 22 HMGA1, 6 FH, 14 YEATS4, 9 OM, 14 UNK) and 96 myometria following the Genomic DNA by Ligation (SQK-LSK109) protocol as per the manufacturer's instructions (Oxford Nanopore Technologies). Sequencing and base calling were performed on the PromethION platform using MinKnow-Live-Basecalling (version 3.4.6). Subsequently, reads were aligned with minimap2 (v.2.16; preset: map-ont)[74] against the GRCh38 reference genome (GCA_000001405.15, excluding alt contigs). Data quality was inspected with NanoStat (v.1.1.2) and NanoPlot (v.1.20.0)[75]. Reads were phased to parental chromosomes using WhatsHap (v.0.18)[76] and genotype information from SNP arrays.

Genome-wide DNA methylation profiles were derived from the nanopore data. Methylation status for each read at each CpG site aligned to the reference genome was called with Nanopolish[77]. Nanopolish considers methylation signal 5 bp before and after each C in CpG context. If two such regions overlap, the methylation call is determined for the combined region. Ninety-nine per cent of CpG sites in the genome are in regions consisting of maximum 29 bp or 11 CpG sites. Hierarchical clustering was performed including one myoma per patient per subtype using tumours and normal samples and alternatively using only tumour samples. The Ward method ('ward.D2') was used for Euclidean distances of methylation values on 99,622 or 180,697 CpG-containing regions within CpG islands on which all samples or tumour samples only, respectively, had sequencing coverage of at least six reads. The locations for CpG islands were extracted from UCSC Genome Browser[78] using the Table Browser tool[78,79].

Mean genome-wide methylation was computed on CpG sites with 2–60× coverage as the sum of methylated calls out of all calls. The average methylation for each sample over an annotation set and normal myometria H2A.Z ChIP–seq peak summits was calculated as mean (over all sites ±250 bp distance from region centre) of mean (for each distance from region centre over all regions in the annotation set) of mean (for a specific CpG over the reads covering that site) methylation. The statistical significance for the overall methylation difference over an annotation set was assessed with ordinary least squares regression with sample mean methylation as a dependent variable, and mean genome-wide methylation and sample subclass as independent variables. The sample subclass was encoded as treatment with respect to normal samples.

Differentially methylated loci (DMLs) were determined from 102 nanopore-sequenced samples with the DSS[80] R package (version 2.28.0) using bsseq R package (version 1.16.1) in R 3.5.1. CpG-containing regions with minimal coverage of six were included and sites that coincided with somatically deleted regions based on BAF segmentation of SNP array data were excluded. One myoma per patient per subtype was included in the analysis. DMLs were detected by Wald test requiring a methylation difference greater than 0.2 and posterior probability greater than 0.99 in comparison of tumours belonging to the MED12, HMGA2, FH or YEATS4 subclasses against all normal samples. Enrichment of hyper- and hypomethylated Cs was analysed with Locus Overlap Analysis (LOLA) R package (version 1.8.0)[63]. The same region set databases were used as described for ChIP–seq. All CpG-containing regions that were tested for DMLs were included in the background set.

The *YEATS4* mutations on nanopore data were visualized using Integrative Genomics Viewer (IGV). Using genetic variation in the long reads provided by this method we could confidently identify the haplotype at the mutated allele in 11 out of 14 of the mutant samples. Reads belonging to the same phase set as the reads from the mutation site were collected and sorted according to their allelic origin, providing a tool to examine DNA methylation in the *YEATS4* mutant versus normal allele. The average allele-specific methylation value from 1,000 bp upstream of the *YEATS4* TSS was calculated for all samples with phased data at the *YEATS4* promoter (84 myomas, 93 normal). For *YEATS4*-mutated myomas, the methylation difference between the mutated and the wild-type alleles was tested using a paired two-sided *t*-test with one myoma per patient included. The sample with both alleles mutated was excluded from the testing. Methylation values around *YEATS4* from both alleles for 11 *YEATS4* mutant myomas and 9 corresponding normal myometria (Supplementary Figs. 5, 6) were visualized using R 3.5.1 and packages data.table (version 1.13.0), ggplot2 (version 3.3.2), ggpubr (version 0.4.0) and tidyverse (version 1.3.0). Smoothing curves were produced with the geom_smooth() -function from package ggplot2.

Allele-specific methylation for the locus that contains *CBX2*, *CBX4* and *CBX8* was analysed with the DSS[80] R package (version 2.34.0) using

100-bp smoothing windows and applied to methylation levels on the reads phased to high- and low-expressing haplotypes. The phasing was determined with WhatsHap (version 0.18) and haplotype expression with ASE analysis as described above.

Methylation–expression correlation

Methylation of each sample at DML was correlated to RNA-seq expression values of genes in *cis* within 5 Mb distance using a linear model in Matrix eQTL[81] R package v.2.3. 'Individual' was used as a covariate in the model. Before correlation, RNA-seq expression data were aligned to the GRCh38 reference genome (version GCA_000001405.15), and processed through variance stabilization in DESeq2 (v. 1.22.2)[51] and limma (v. 3.42.0)[52] removeBatchEffect. Correlations with $P < 1 \times 10^{-10}$ were considered significant.

Genome-wide association study

Meta-analysis of germline UL predisposition was done in three stages. First, genome-wide summary statistics for 'uterine fibroids' GWAS were publicly available from Biobank Japan (BBJ; http://jenger.riken.jp/en/, accessed on 2 September 2020) and FINNGEN (https://finngen.gitbook.io/documentation/v/r3/, accessed on 2 September 2020). The BBJ cohort had a total of 5,954 UL cases and 95,010 female controls[82], and FINNGEN a total of 11,490 UL cases and 64,898 female controls. Both these cohort statistics were precomputed for mixed model logistic regression (SAIGE); details are available at the sources listed above. FINNGEN data were lifted over to GRCh37/hg19 coordinates (UCSC liftOver). Second, case-control statuses within the UK Biobank cohort (white British women; accessed in March 2019) were determined using both self-reported ULs and ICD10/9 codes (see ref. [83] for phenotype definition, population stratification and genotype quality-control steps), resulting in 16,076 UL cases and 204,495 female controls. SAIGE (v. 0.35.8) was run to determine the genome-wide summary statistics for UK Biobank. Finally, an inverse-variance weighted fixed effects meta-analysis was applied to 6.0 million imputed SNPs that were available from all three cohorts (R package 'meta' v4.8-4). Supplementary Table 25 gives pheweb (v1.1.14) annotated SNPs that passed meta-analysis with $P < 1 \times 10^{-6}$ and their summary statistics for each of the three cohorts. SNP annotations also include the gene symbols from ref. [83] and any GWAS catalogue 'uterine fibroid' associations found within 3 Mb (accessed on 1 October 2020).

Statistical information

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to groups, except when noted otherwise. The scoring of the immunohistochemistry staining was done blind to the tumour subclass.

Each statistical test included only one tumour per patient per subclass, unless noted otherwise. The tumour selection was prioritized by data availability and was otherwise random: the third column of Supplementary Table 2 denotes which of the tumours were selected. Further details can be found in the respective Methods sections.

Subclass-specific enrichment of allelic loss was calculated as described in 'SNP array'. Mutual exclusivity of the UL subclasses was assessed with a one-sided Fisher's exact test. Gene-based UL association of germline LoF variants was assessed with SKAT-O tests as described in 'Analysis of germline loss-of-function variants'. Age at UL diagnosis was compared using a two-sided Welch's *t*-test. FDR and fold change (FC) for differential expression were derived from DESeq2 analysis as described in 'RNA-seq and driver mutation identification'. The accumulation of SRCAP tumours to a subset of patients was studied with a permutation test as described in 'RNA-seq and driver mutation identification'. Allele-specific expression was calculated with a binomial test as described in 'Allele-specific gene expression analysis'. Fisher's exact test was used to calculate the *P* value in pathway analysis as described in 'Pathway analysis'. Odds ratios and *P* values between the immunohistochemistry score and the tumour subclass were calculated with general estimating equations implemented in the R package geeglm. The full model is described in 'Immunohistochemistry'. All enrichment odds ratios and *P* values were calculated with one-sided Fisher's exact test as implemented in the LOLA[63] R package. FC and FDR for differential H2A.Z binding and chromatin accessibility were derived from DESeq2 analysis as described in 'ChIP–seq' and 'ATAC–seq'. Significance analysis for bridging links is described in 'HiChIP'. Differentially methylated loci, allele-specific methylation and overall methylation difference over an annotation set were calculated using the DSS[80] R package, paired *t*-test or ordinary least squares regression as described in 'Nanopore long-read sequencing'. Correlations between methylation and expression were calculated with the Matrix eQTL[81] R package as described in 'Methylation–expression correlation'. GWAS methods are described under 'Genome-wide association study'.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The peak level data used in the study are available for research use through Zenodo (https://doi.org/10.5281/zenodo.4745433). Genetic data presented in this manuscript have been deposited at the European Genome–phenome Archive (EGA; https://www.ebi.ac.uk/ega/) under accession number EGAS00001004499. A data access committee (DAC) has been established from two University of Helsinki representatives that are independent of the authors of the current study. See Supplementary Table 26 for EGA dataset accession numbers. Requests for the data should be sent to the DAC via email (dac-finlandmyomastudy@helsinki.fi). The DAC ensures that the intended use of data as detailed in the request is compatible with the requirements of the European General Data Protection Regulation (GDPR), consistent with the consents given and otherwise ensures the protection of data subjects' rights as required by the GDPR. The DAC will always grant access to the data if the University is legally allowed to do so without infringing the rights and freedoms of data subjects. Subject to the requirements of the GDPR, the DAC grants access to the genetic data to non-commercial academic research on neoplasia and chromatin. Roadmap Epigenomics ChIP–seq and DNase–seq data (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) provided in the LOLA extended and core databases were downloaded from http://cloud.databio.org/regiondb/. Chromatin states provided in mnemonics bed files by the Roadmap Epigenomics project were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/. GWAS cohort material was downloaded from http://jenger.riken.jp/en/ (accessed on 2 September 2020) and https://finngen.gitbook.io/documentation/v/r3/ (accessed on 2 September 2020). UK Biobank material access can be applied for at https://www.ukbiobank.ac.uk/. For RNA-seq, the reference annotation for GRCh37 was downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/) and for GRCh38 from GenBank (accession: GCA_000001405.15). For RNA-seq variant calling, the SNP and indel resources were downloaded from Broad institute (https://console.cloud.google.com/storage/browser/gcp-public-data–broad-references/hg19/v0). DAR annotation data are available from http://homer.ucsd.edu/homer/data/genomes/hg19.v6.4.zip. Source data are provided with this paper.

Code availability

Code for performing the analyses are available from Zenodo (https://doi.org/10.5281/zenodo.4745433).

43. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. **17**, 1665–1674 (2007).
44. Mäkinen, N. et al. *MED12*, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. *Science* **334**, 252–255 (2011).
45. McGuire, M. M. et al. Whole exome sequencing in a random sample of North American women with leiomyomas identifies *MED12* mutations in majority of uterine leiomyomas. *PLoS ONE* **7**, e33251 (2012).
46. Mehine, M. et al. Integrated data analysis reveals uterine leiomyoma subtypes with distinct driver pathways and biomarkers. *Proc. Natl Acad. Sci. USA* **113**, 1315–1320 (2016).
47. Mäkinen, N., Kämpjärvi, K., Frizzell, N., Bützow, R. & Vahteristo, P. Characterization of *MED12*, *HMGA2*, and *FH* alterations reveals molecular variability in uterine smooth muscle tumors. *Mol. Cancer* **16**, 101 (2017).
48. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
49. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol*. **33**, 290–295 (2015).
50. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
51. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
52. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. **43**, e47 (2015).
53. Witten, D. M. & Tibshirani, R. Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Series B Stat. Methodol*. **73**, 753–772 (2011).
54. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
55. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn*. **52**, 91–118 (2003).
56. Van Hout, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at https://doi.org/10.1101/572347 (2019).
57. Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet*. **52**, 634–639 (2020).
58. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun*. **7**, 12817 (2016).
59. Krämer, A., Green, J., Pollard, J., Jr & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530 (2014).
60. Mehine, M. et al. Characterization of uterine leiomyomas by whole-genome sequencing. *N. Engl. J. Med*. **369**, 43–53 (2013).
61. Drmanac, R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
62. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep*. **9**, 9354 (2019).
63. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).
64. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
65. Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
66. Corces, M. R. et al. An improved ATAC–seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
67. Orchard, P., Kyono, Y., Hensley, J., Kitzman, J. O. & Parker, S. C. J. Quantification, dynamic visualization, and validation of bias in ATAC-seq data with ataqv. *Cell Syst*. **10**, 298–306.e4 (2020).
68. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
69. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
70. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
71. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. **16**, 259 (2015).
72. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun*. **10**, 4221 (2019).
73. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol*. **16**, 198 (2015).
74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
75. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
76. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. Preprint at https://doi.org/10.1101/085050 (2016).
77. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
78. Haeussler, M. et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*. **47**, D853–D858 (2019).
79. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. **32**, D493–D496 (2004).
80. Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**, 1446–1453 (2016).
81. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
82. Ishigaki, K. et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet*. **52**, 669–679 (2020).
83. Välimäki, N. et al. Genetic predisposition to uterine leiomyoma is determined by loci for genitourinary development and genome stability. *eLife* **7**, e37110 (2018).

Acknowledgements The Finland Myoma Study has been made possible by professionals from operating theatre nurses to data scientists, and in particular the study participants, and we are deeply grateful for their vital support during the years. We thank S. Marttinen, S. Soisalo, M. Rajalaakso, I.-L. Åberg, I. Vuoristo, A. London, J. Kolakowska, H. Metsola and K. Pylvänäinen for technical support; and P. Ikonen and the rest of the staff of the Kätilöopisto Maternity Hospital, and the staff of the Department of Pathology, University of Helsinki for technical assistance. Capillary sequencing was performed by the sequencing unit of FIMM Technology Centre, University of Helsinki supported by Biocenter Finland. This research was conducted using the UK Biobank Resource under Application Number 32506. We acknowledge the participants and investigators of the FinnGen study; the Biobank Japan Project for the 'uterine fibroids' GWAS summary statistics; and the computational resources provided by the ELIXIR node, hosted at the CSC–IT Center for Science, Finland. The study was supported by grants from Academy of Finland (Finnish Center of Excellence Program 2012–2017, 2018–2025, No. 250345 and 312041, Academy Professor grants No. 319083 and 320149), European Research Council (ERC, 695727), iCAN Digital Precision Cancer Medicine Flagship (320185), Cancer Society of Finland, Sigrid Juselius Foundation and Jane and Aatos Erkko Foundation.

Author contributions D.G.B., H.K., N.V., A.K., K.P., E.K. and L.A.A. conceived the ideas, planned the experiments and wrote the manuscript. D.G.B., M.R. and M.J. performed the ATAC and ChIP–seq experiments. D.G.B. performed HiChIP and western blot experiments. M.R., E.K., K.P and H.K. analysed ATAC, ChIP and HiChIP data. H.K., N.V. and T.C. performed the mutation analysis. H.K. and N.V. performed the gene expression analysis. R.-M.P. and N.V. performed the allele-specific expression analysis. A.T., K.P. and E.K. performed the methylation analysis. A.K., S.N. and T.C. performed the IHC experiments. R.B., A.K. and H.K. analysed the IHC data. N.V. analysed the germline predisposition data. J.K., S.A., P.V., M.M. and N.M. performed the mutation screening and managed clinical data. A.P., J.J., O.H. and R.B. collected the tissue samples. J.R. and R.L. provided support with analysis and management of data. J.T., K.R. and B.S. provided support with the experimental design and interpretation.

Competing interests The authors declare no competing interests.

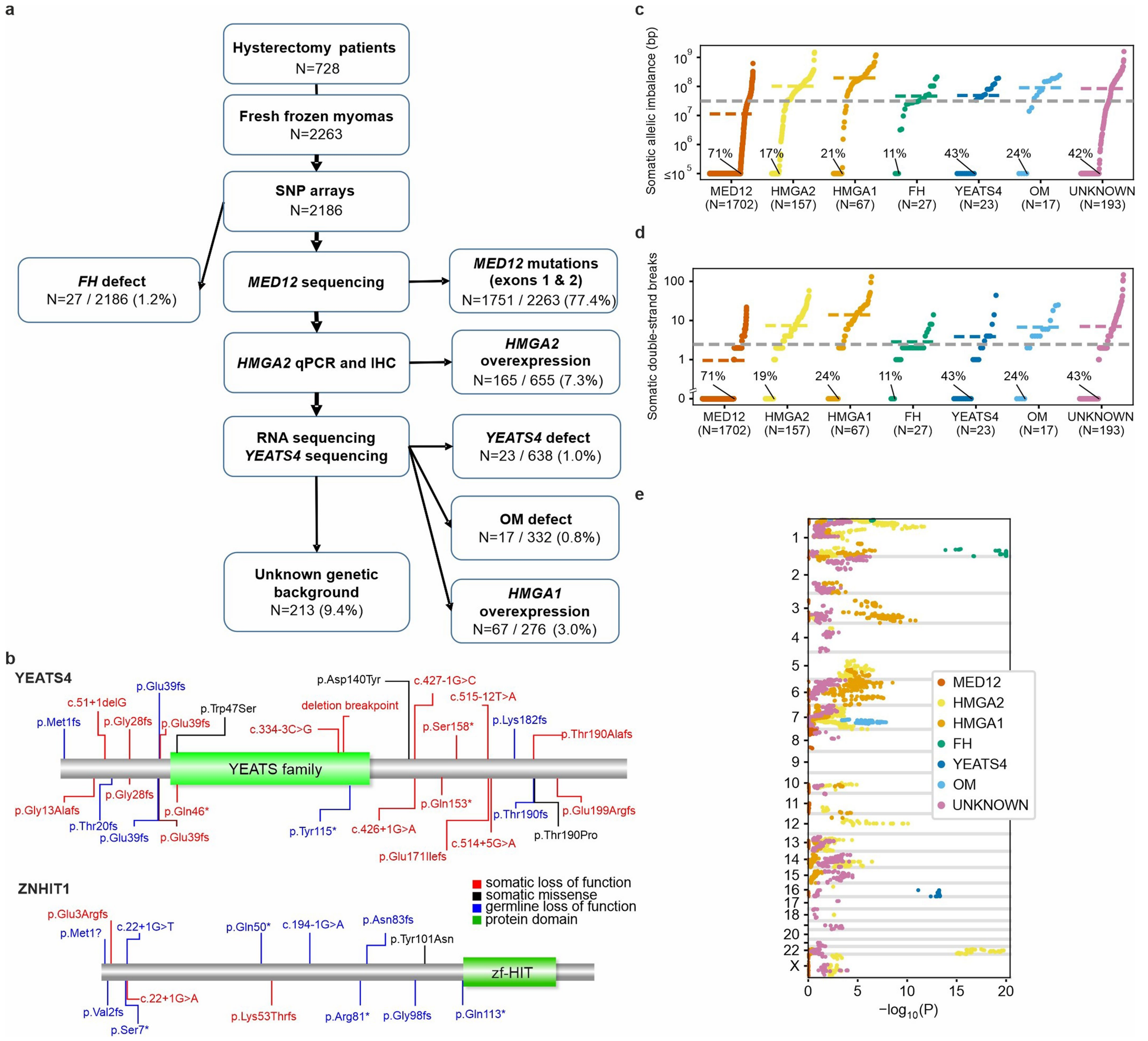
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03747-1.

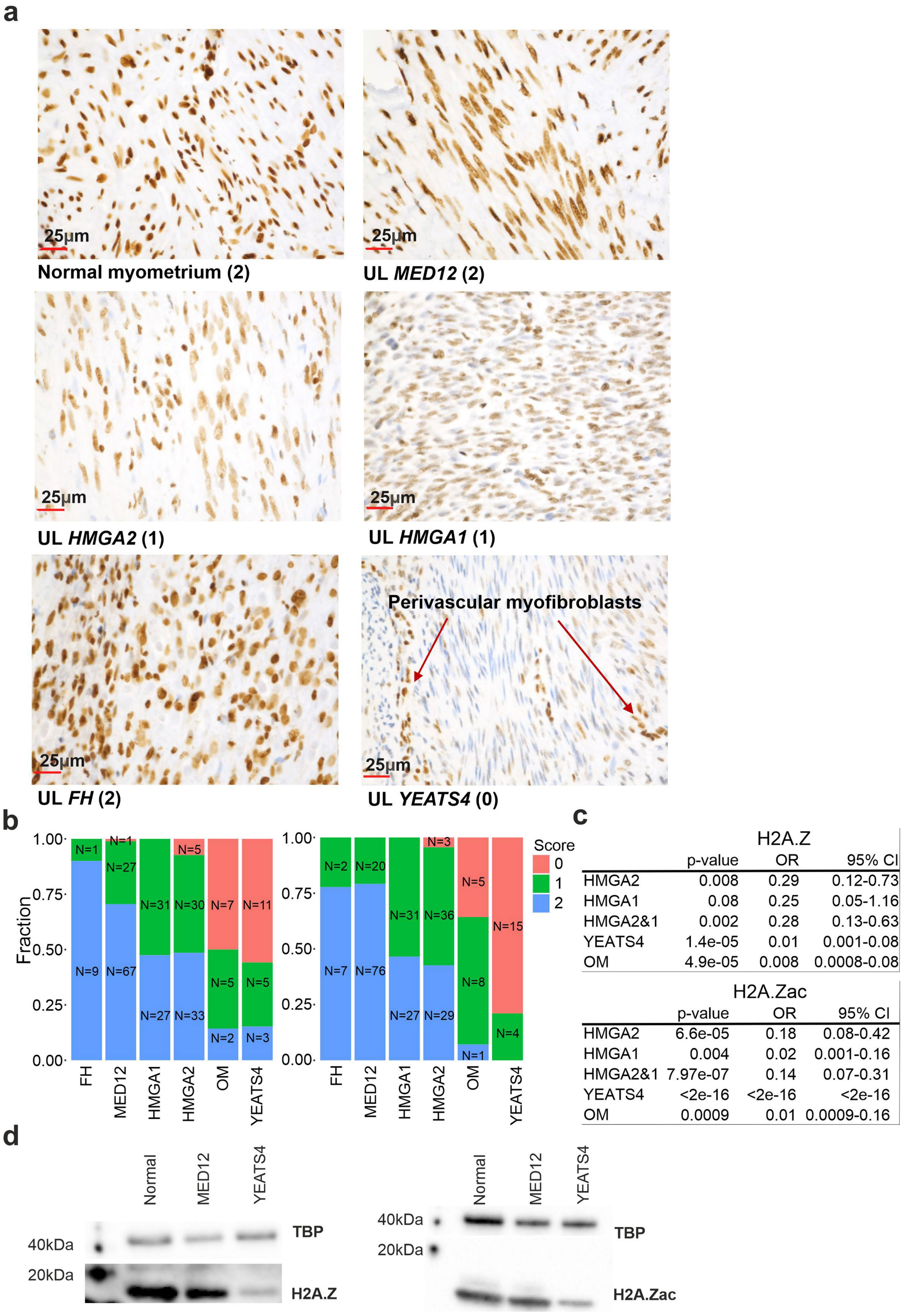
Correspondence and requests for materials should be addressed to E.K. or L.A.A.

Peer review information *Nature* thanks Zehra Ordulu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Overview of study material and somatic allelic imbalance. a, The sequential stages of UL subclass detection. Denominators are numbers of tumours screened at each stage; see Methods for a detailed description of data used at each stage. Percentages refer to proportions among all 2,263 tumours. HMGA1 subclass comprised 67 tumours without any other known driver change. Three tumours with an SRCAP complex gene mutation are shown here as one MED12 and two HMGA2 tumours as they also showed these driver changes. **b**, Germline LoF variants (blue; UK Biobank WES) compiled with somatic LoF (red) and missense (black) variants. Green boxes represent protein domains from Pfam. **c**–**e**, Overview of genome-wide somatic allelic imbalance in 2,186 SNP-arrayed ULs. **c**, *y*-axis gives the total length of the genome affected by somatic loss and gain aberrations per tumour, stratified by subclass (logarithmic and truncated to 10^5 bp). Dashed lines show the overall and subclass-specific mean values. Percentage units refer to the proportion of chromosomally stable tumours within each subclass. **d**, Estimated numbers of somatic DSBs. **e**, Subclass-specific enrichment of allelic loss: *x*-axis gives log-transformed, one-sided tests of loss-event enrichment in each subclass compared to the rest of the tumours (truncated to 1.0×10^{-20}). *y*-axis indicates the genomic position (autosomes and X). See Supplementary Table 6 for detailed statistics.



H2A.Z

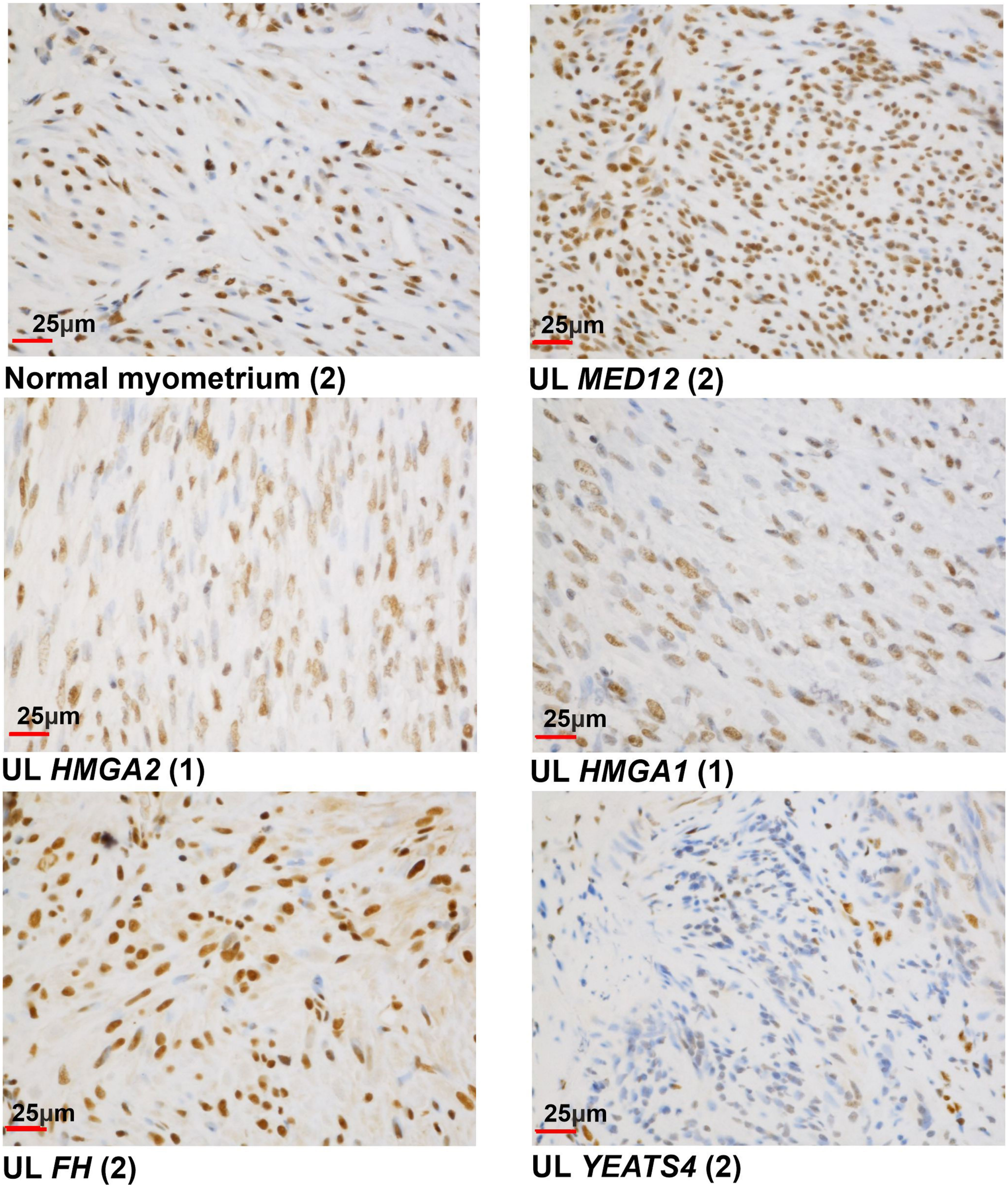
	p-value	OR	95% CI
HMGA2	0.008	0.29	0.12-0.73
HMGA1	0.08	0.25	0.05-1.16
HMGA2&1	0.002	0.28	0.13-0.63
YEATS4	1.4e-05	0.01	0.001-0.08
OM	4.9e-05	0.008	0.0008-0.08

H2A.Zac

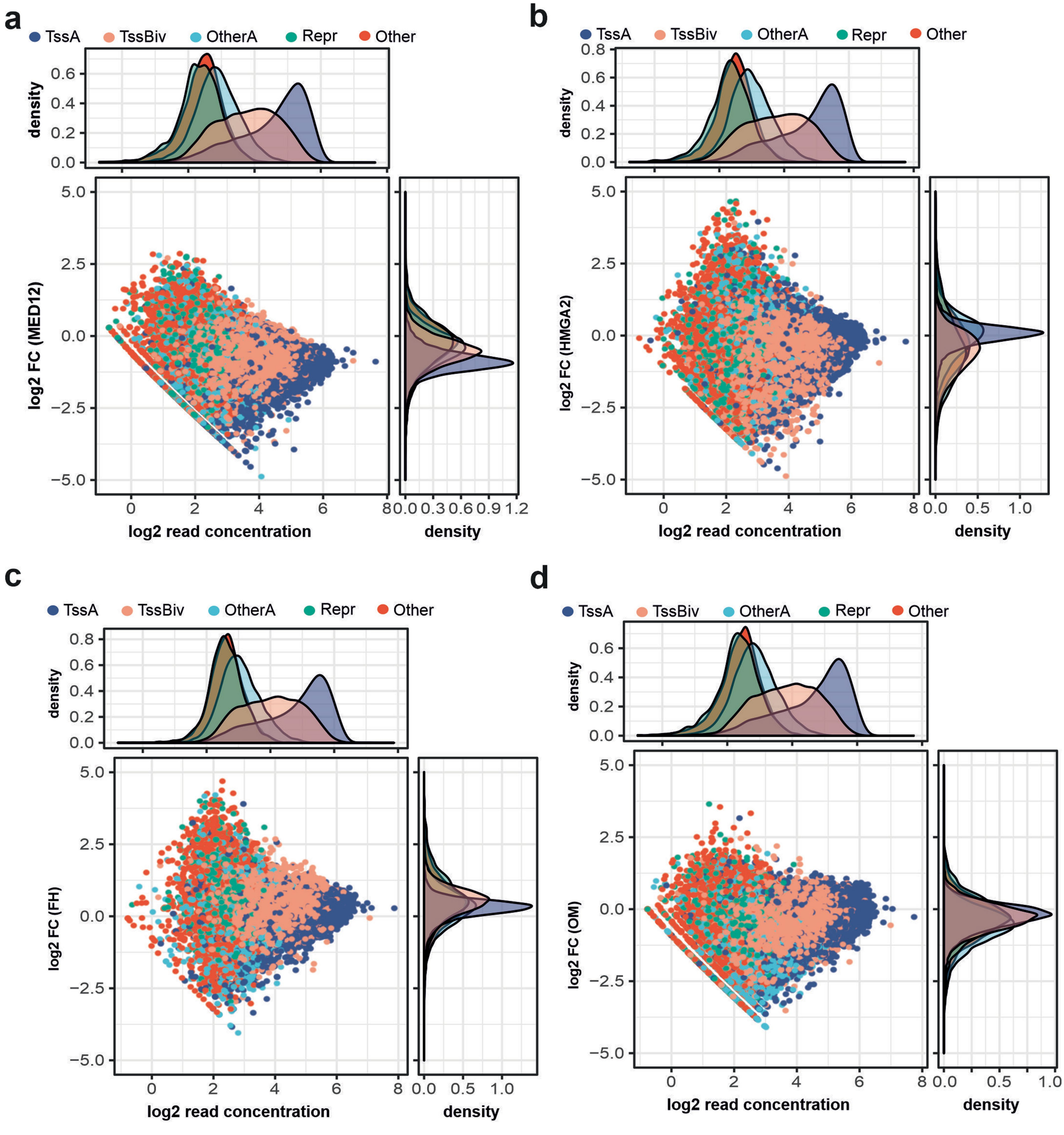
	p-value	OR	95% CI
HMGA2	6.6e-05	0.18	0.08-0.42
HMGA1	0.004	0.02	0.001-0.16
HMGA2&1	7.97e-07	0.14	0.07-0.31
YEATS4	<2e-16	<2e-16	<2e-16
OM	0.0009	0.01	0.0009-0.16

Extended Data Fig. 2 | See next page for caption.

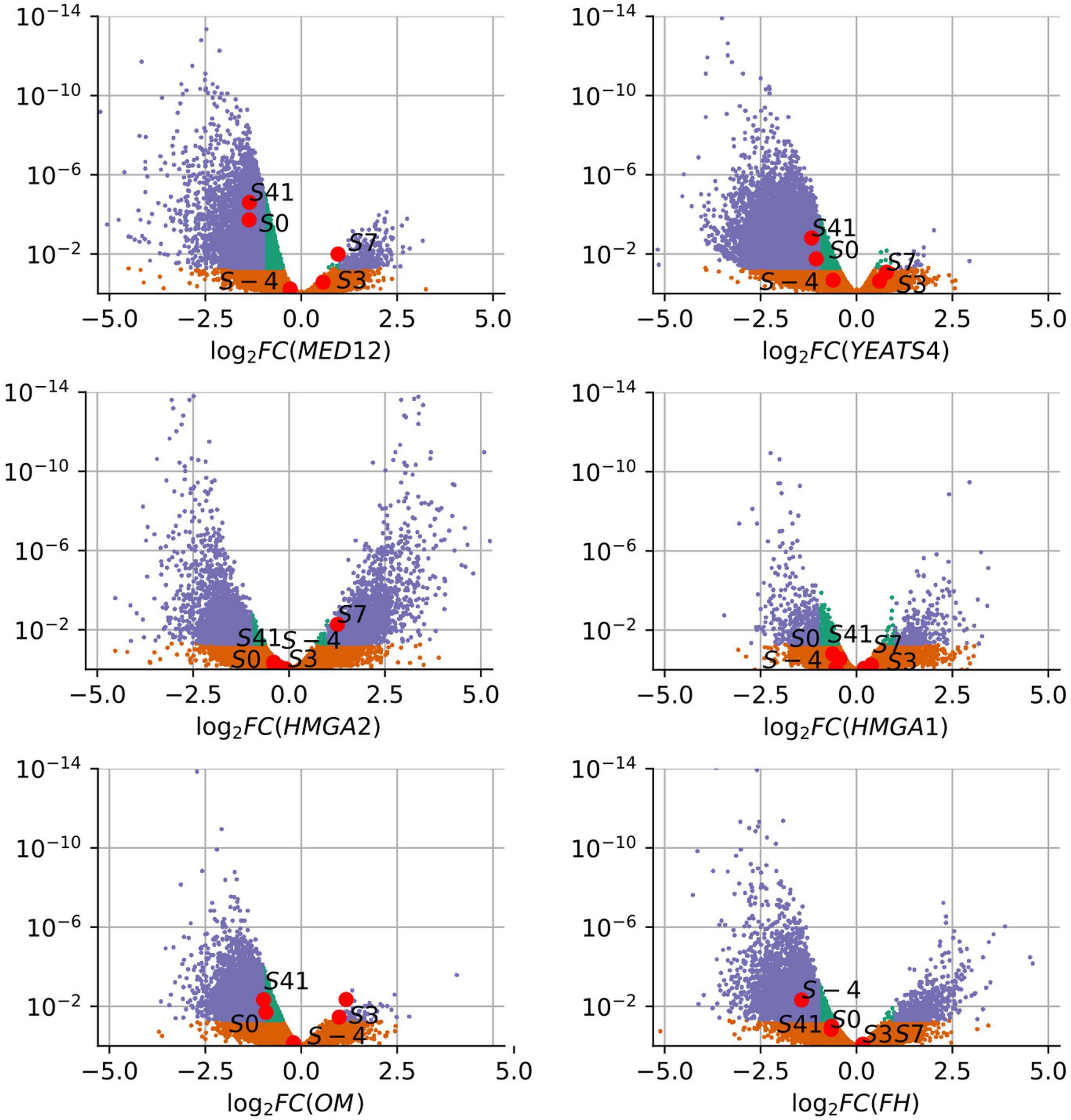
Extended Data Fig. 2 | H2AZ staining in myometrium, MED12, HMGA, FH, OM and YEATS4 tumours. a, Representative immunostaining of non-acetylated H2A.Z in normal myometrium and MED12, HMGA2, HMGA1, FH and YEATS4 ULs. The intensity of immunoreaction is given in parentheses: 0 = negative or weak, 1 = moderate, 2 = strong. YEATS4 mutated myoma displays negative or weak H2A.Z staining in neoplastic cells, but preserved staining in endothelial, perivascular and scattered inflammatory cells. 40× magnification. The scores for all the stained samples are given in **b** (left). MED12 (*n* = 95), HMGA2 (*n* = 68), HMGA1 (*n* = 58), FH (*n* = 10), OM (*n* = 14), YEATS4 (*n* = 19). **b**, Distribution of staining scores on non-acetylated H2A.Z antibody (left) and acetylated H2A.Z (H2A.Zac) antibody (right). **c**, H2A.Z (top) and H2A.Zac (bottom) staining intensity in HMGA2, HMGA1, HMGA2 and HMGA1 combined (HMGA2&1), YEATS4 and OM tumours compared to MED12 tumours (GEE-model). Nominal two-sided *P* values are shown. **d**, Western blot analysis of H2A.Z (left), H2A.Zac (right) and TBP (for both, as loading control) in chromatin fraction from MED12 and YEATS4 tumours and related normal myometrium, all from the same individual. Molecular weights of the protein ladder are indicated. Owing to the limited amount of the respective myometrium tissue as control, it was possible to perform the western blots only once. Uncropped blot available in Supplementary Fig. 30.



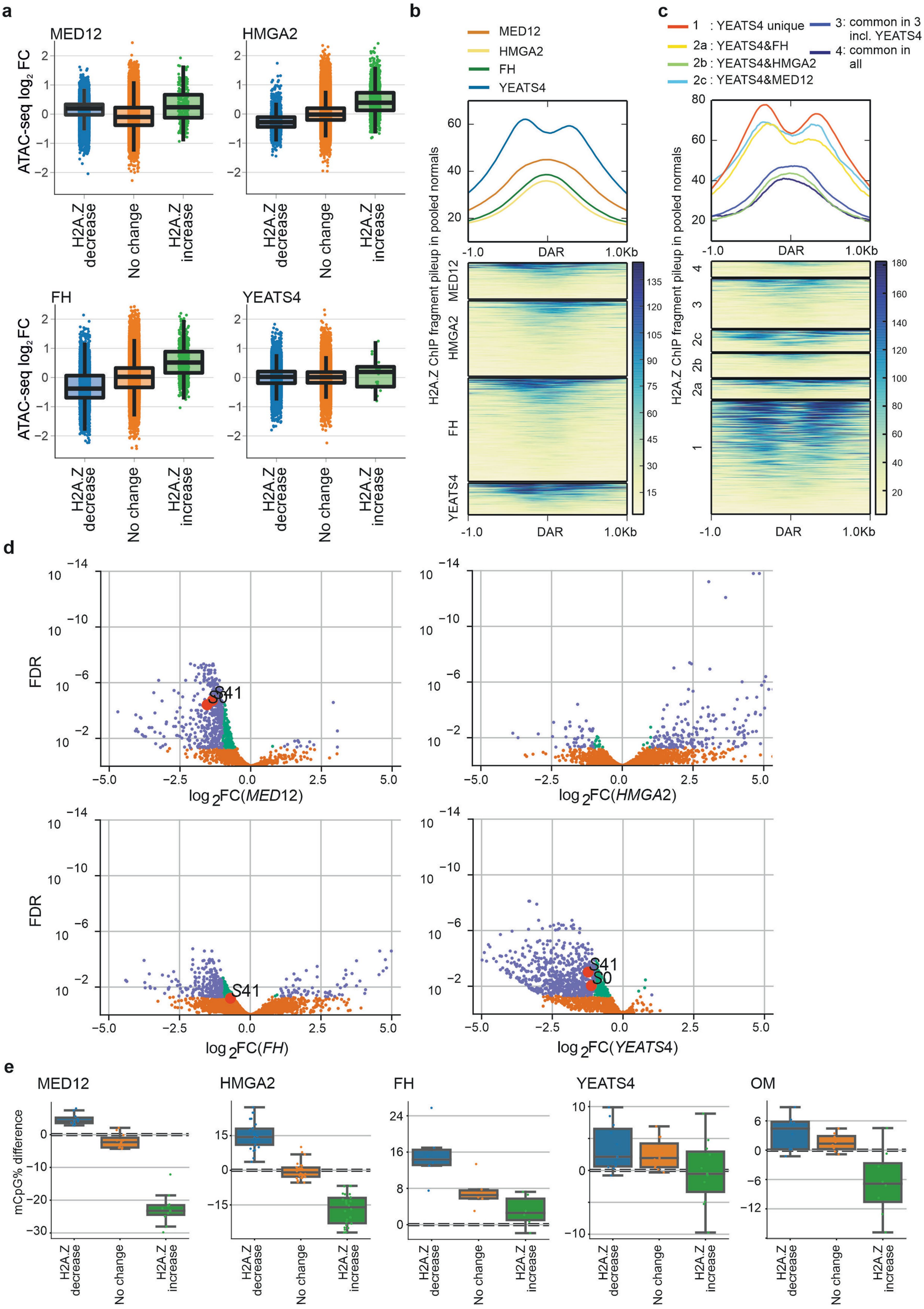
Extended Data Fig. 3 | Reduction in acetylated H2A.Z in MED12, HMGA, FH and YEATS4 tumours. Representative immunostaining of H2A.Zac in normal myometrium and MED12, HMGA, FH and YEATS4 ULs. The intensity of immunoreaction is shown in parentheses: 0 = negative or weak, 1 = moderate, 2 = strong. 40× magnification. The scores for all the stained samples are presented in Extended Data Fig. 2b (right). MED12 (*n* = 96), HMGA2 (*n* = 68), HMGA1 (*n* = 58), FH (*n* = 9), OM (*n* = 14), YEATS4 (*n* = 19).



Extended Data Fig. 4 | Differences in H2A.Z binding between UL subclasses and normal samples. $\log_2$ FC and average binding strength (normalized read concentration) were calculated on each H2A.Z peak region and are represented by a dot. **a**–**d**, MED12 (*n* = 3) (**a**), HMGA2 (*n* = 4) (**b**), FH (*n* = 4) (**c**) and OM (*n* = 3) (**d**) tumours were separately compared to all normal samples (*n* = 11). H2A.Z peak regions were stratified by five-state genome annotations from myometrium. States were annotated as bivalent TSS regions (TssBiv marked by both H3K4me3 and H3K27me3), active TSS regions (TssA marked by both H3K4me3 and H3K27ac), active chromatin outside TSS regions (OtherA marked by H3K27ac), repressed chromatin (Repr marked by H3K27me3) and other chromatin (Other).

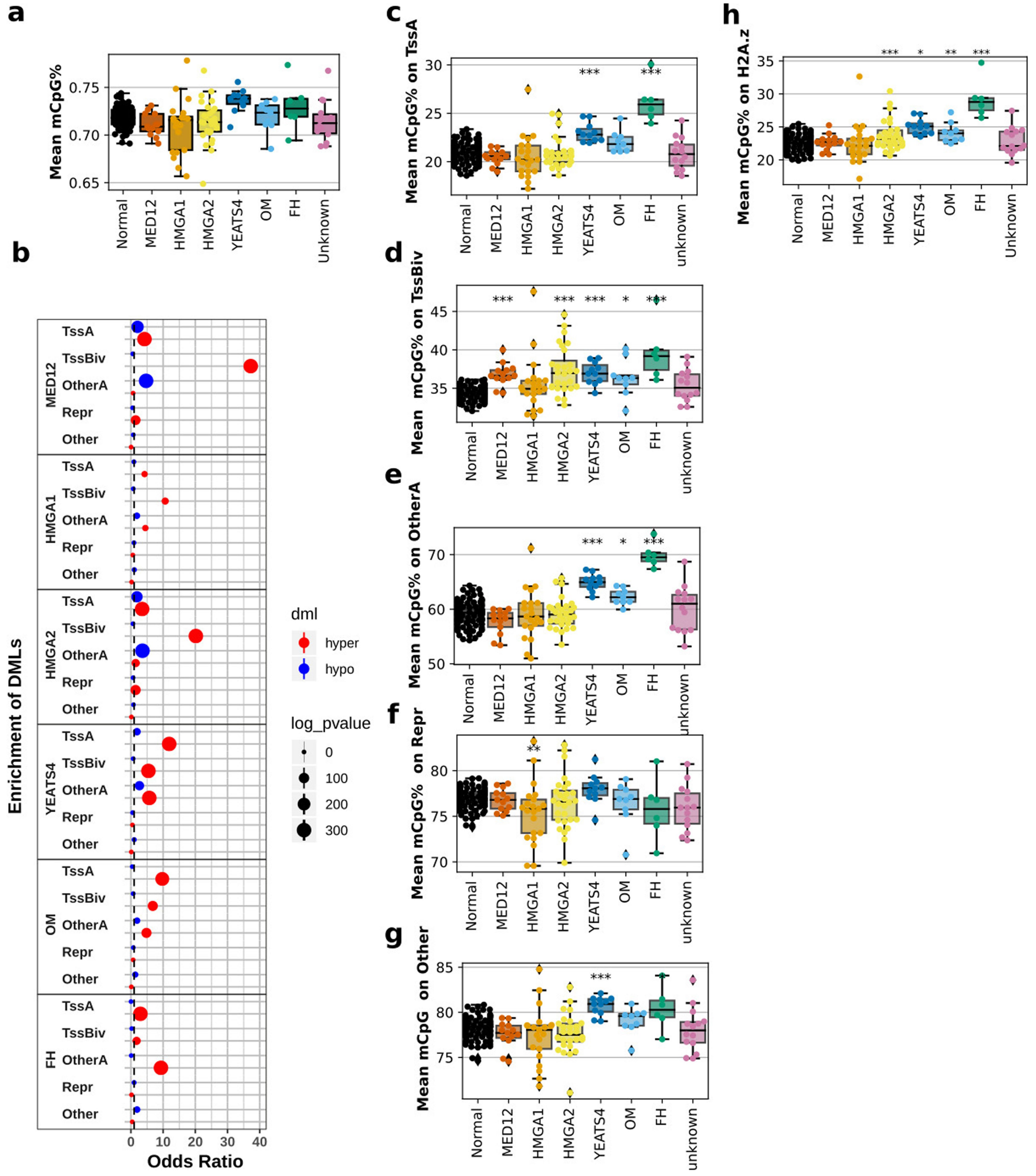


Extended Data Fig. 5 | Differential H2A.Z binding between tumours and myometria. Volcano plots displaying differences in H2A.Z binding for MED12 ($n = 2$), YEATS4 ($n = 2$), HMGA2 ($n = 2$), HMGA1 ($n = 4$), OM ($n = 2$) and FH ($n = 2$) tumours against normal samples ($n = 4$) from the spike-in ChIP–seq experiments. FDR (*y*-axis) and $\log_2$FC (*x*-axis) from DESeq2 analysis as implemented in the DiffBind R package. Violet dots represent differential H2A.Z binding sites (FDR < 0.05, $|\log_2 FC| > 1$). The highlighted peaks (pink dots) are located close to *CBX8* and named as *S* with distance in kilobases from the *CBX8* TSS.

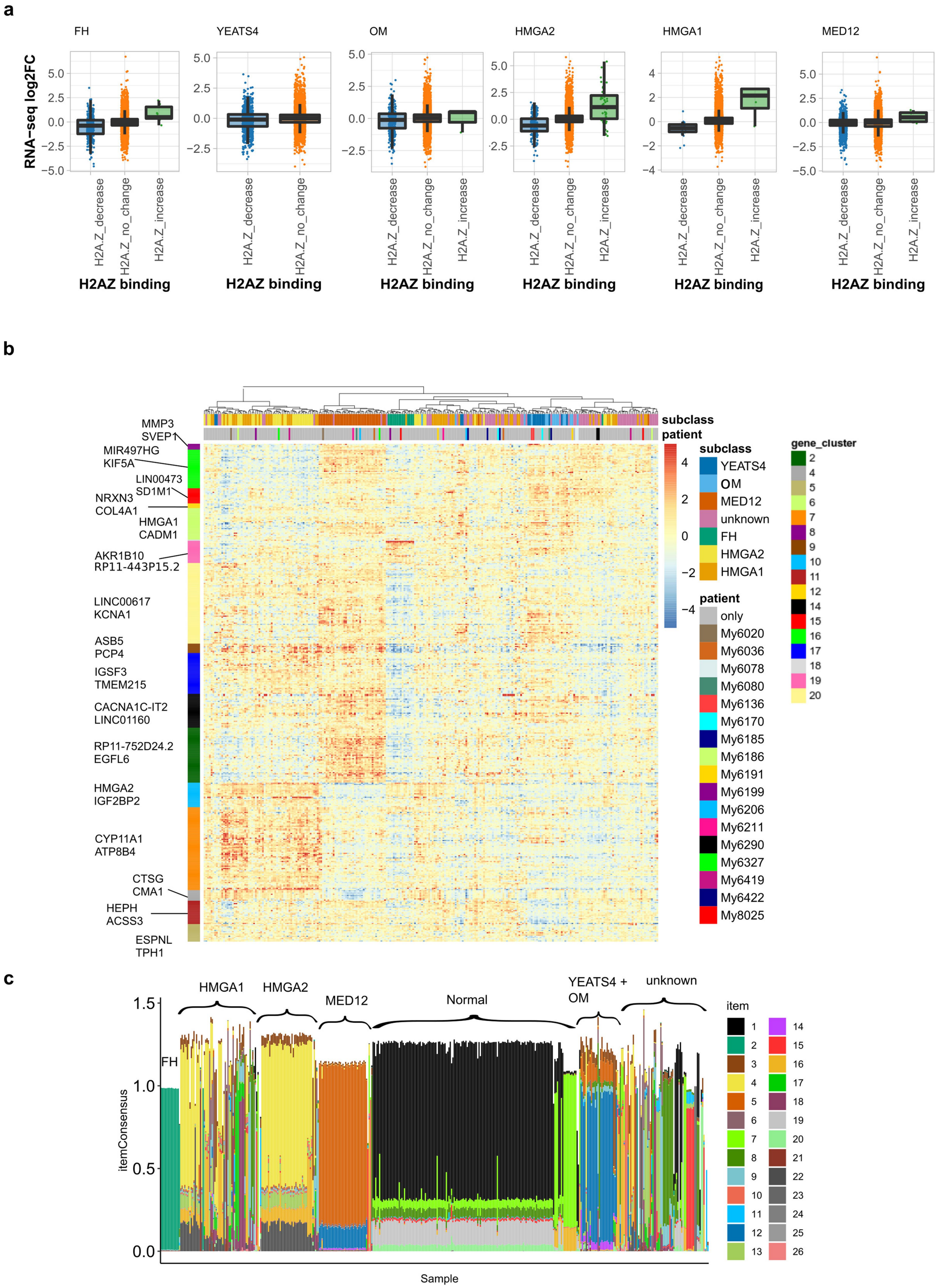


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | H2A.Z occupancy associates with chromatin opening and DNA methylation. **a**, Chromatin accessibility as a function of H2A.Z occupancy. Chromatin accessibility $\log_2$FC (*y*-axis) was measured by DESeq2-normalized ATAC–seq Tn5 insertion counts in MED12 ($n = 4$), HMGA2 ($n = 4$), FH ($n = 4$) and YEATS4 ($n = 4$) tumours compared with normal samples ($n = 15$) at H2A.Z peaks stratified into increased, no change and decreased binding. **b**, **c**, Pileup of H2A.Z ChIP fragments from pooled myometrium data at DARs shown by composite plots and heatmaps. DARs in each UL subclass (**b**) or in YEATS4 tumours stratified by overlap with other UL subclasses (**c**) are represented by heatmap rows over which mean H2A.Z fragment coverage is calculated. **d**, Differences in H2A.Z binding at DARs in UL subclasses as compared to normal samples. Violet dots represent differential H2A.Z binding sites (FDR < 0.05, $|\log_2\mathrm{FC}| > 1$). The highlighted peaks (pink dots) are located close to *CBX8* and named as *S* with distance in kilobases from the *CBX8* TSS. **e**, Mean sample-wise DNA methylation differences in MED12 ($n = 11$), HMGA2 ($n = 26$), FH ($n = 6$), YEATS4 ($n = 14$) and OM ($n = 8$) tumours compared with respective normal samples at H2A.Z peaks stratified into increased, no change and decreased binding. H2A.Z binding differences in **a**, **d**, **e** are from the spike-in ChIP–seq experiments comparing two tumours from each subclass to four normal samples. Increased, no change and decreased binding were defined using FDR < 0.05 and $|\log_2\mathrm{FC}| > 1$ cutoffs. Boxplots show the median and the first and third quartiles. Error bars extend up to 1.5 IQR beyond the quartiles.

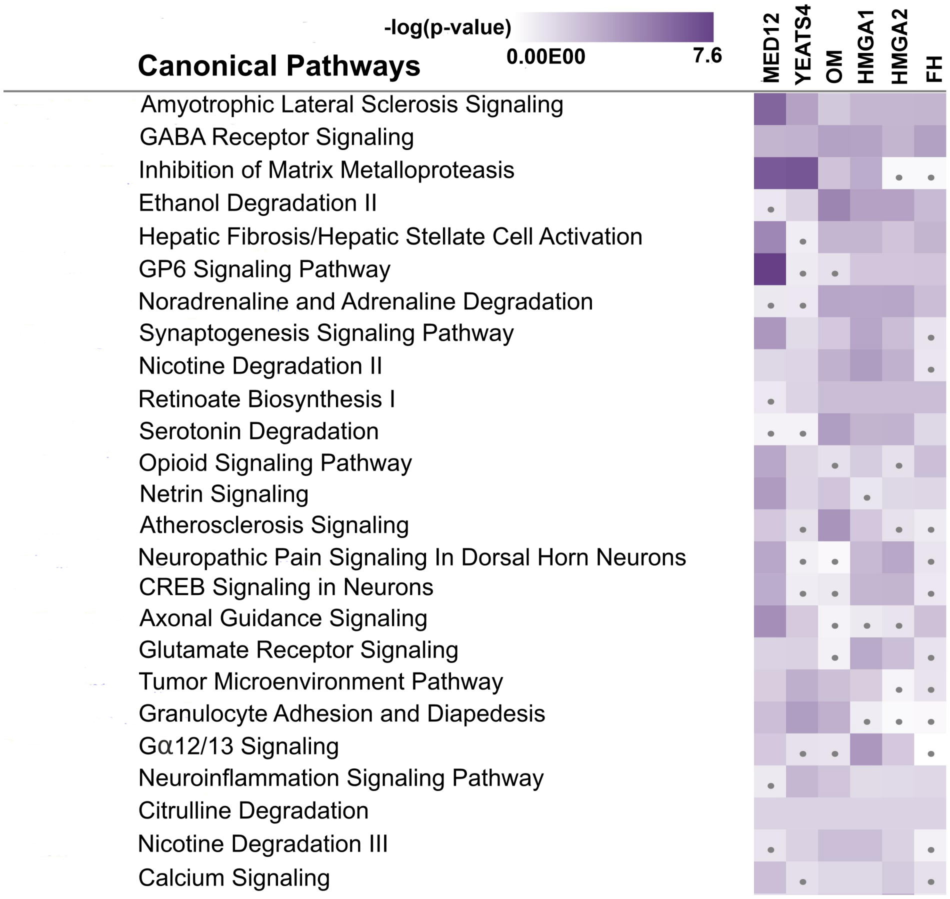


Extended Data Fig. 7 | DNA methylation displays distinct patterns in UL subclasses. a, Overall genome-wide DNA methylation in ULs and normal myometrium. Each dot represents a sample. **b**, Enrichment of hyper- and hypomethylated loci in different tumour subclasses on five chromatin states from normal myometrium. ORs and *P* values (log_pvalue stands for −$\log_{10}$ (*P* value)) from one-sided Fisher's exact test implemented in the LOLA R package. **c**–**g**, Overall DNA methylation on active (**c**) and bivalent (**d**) TSSs, other active chromatin (**e**), repressed/poised chromatin (**f**) and other, quiescent, chromatin regions (**g**). Significance of methylation difference against normals evaluated by ordinary least squares regression: ***$P < 0.0001$, **$P < 0.001$, *$P < 0.01$. Test controlled by global methylation levels. Sample sizes: normal 96, MED12 13, HMGA1 21, HMGA2 28, YEATS4 11, OM 9, FH 6, Unknown 14. Box covers the central two quartiles of the distribution. Median is highlighted. Whiskers extend to the minimum and maximum of the distribution or at most 1.5× IQR from the edge of the box, whichever is closer. Multiple testing correction was not performed. **h**, Mean sample-wise DNA methylation in ULs and normal myometrium at H2A.Z binding sites derived from pooled normal myometrium tissue samples. Asterisks, box-and-whiskers plots and sample sizes are as in **c**–**g**.

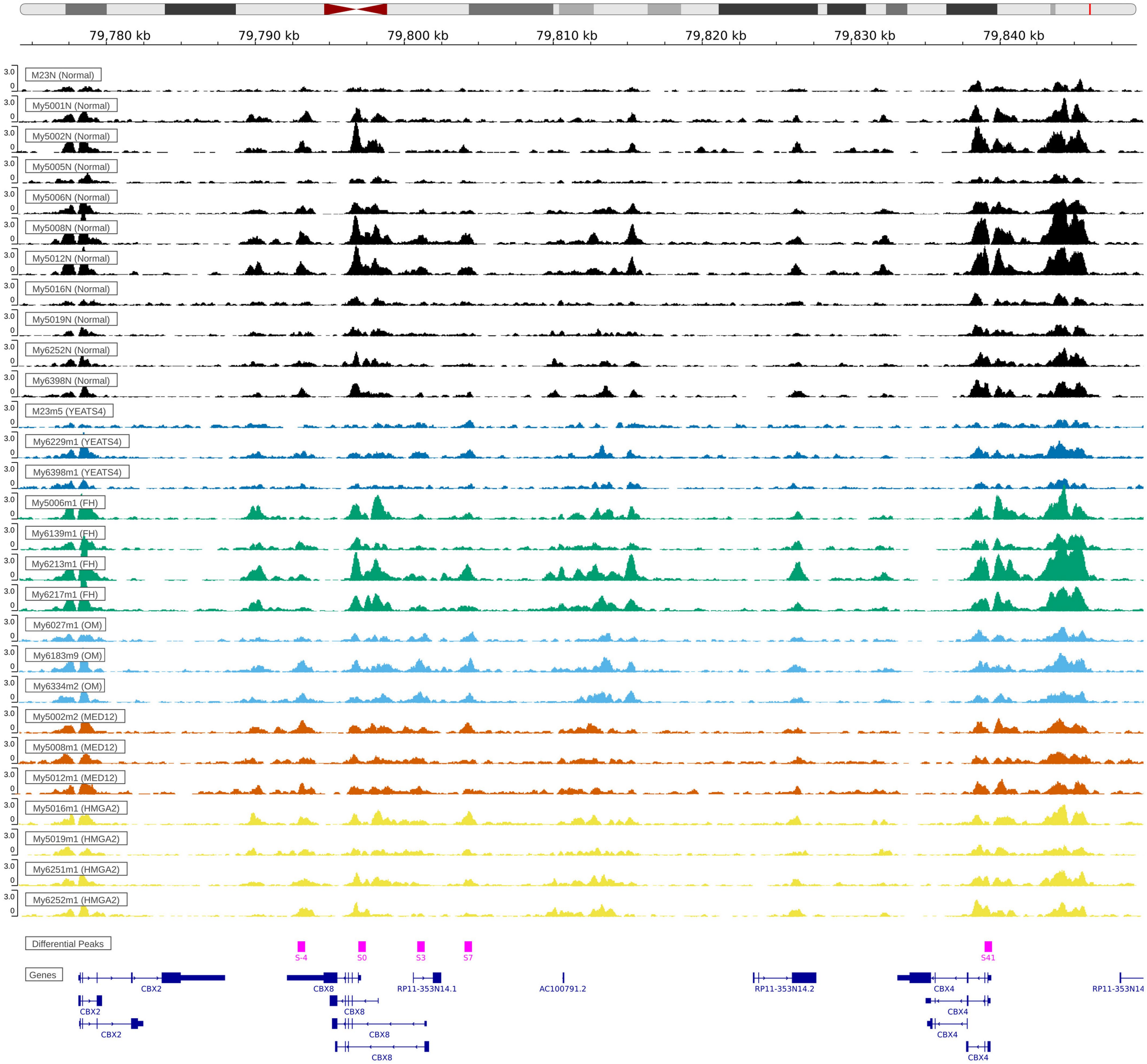


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Clustering of RNA-seq samples. a, $\log_2$FC of genes for which decreased (FDR < 0.05, FC < −1), increased (FDR < 0.05, FC > 1) or no-change H2A.Z peaks are located at the TSS. $\log_2$FC measured by differential expression analysis of tumours (MED12 ($n = 38$), HMGA2 ($n = 44$), HMGA1 ($n = 62$), FH ($n = 15$), OM ($n = 15$) and YEATS4 ($n = 16$)) against normal myometrium ($n = 162$). H2AZ binding differences are from the spike-in ChIP experiments comparing MED12 ($n = 2$), YEATS4 ($n = 2$), HMGA2 ($n = 2$), HMGA1 ($n = 4$), OM ($n = 2$) and FH ($n = 2$) tumours against normal samples ($n = 4$). Boxplots show the median and the first and third quartiles. Error bars extend up to 1.5× IQR beyond the quartiles. **b**, Heatmap presentation of 426 genes that separate myoma subclasses, selected on the basis of linear discriminant analysis. The ordering of samples and genes is based on an unsupervised hierarchical clustering of the 5% ($n = 1{,}355$) most variable genes. Two genes per gene cluster are highlighted on the basis of the highest absolute value in discriminant vectors. Patients from whom more than one tumour entered the analysis are highlighted in separate colours. All 426 genes are presented in Supplementary Table 20. **c**, Consensus clustering of RNA-seq samples. *x*-axis is sorted by subclass (from left to right: FH, HMGA1, HMGA2, MED12, normal myometrium, OM, YEATS4, and unknown; subclass labels are shown for reference). Item consensus is the mean consensus of an item with all the other items in the same cluster. For each sample, the item consensus value corresponding to each cluster ($k = 26$) is represented by a colour. For example, all FH samples have the largest item consensus on cluster 2, represented by dark green. Both YEATS4 and OM samples cluster predominantly to cluster 12, represented by blue. Unknown samples form several small clusters. The item consensus value (e_i) of each cluster (k) is presented on the *y*-axis. It is defined as: $m_i(k) \;=\; \frac{1}{N_k - 1\{e_i \in I_k\}} \sum_{j \in I_k, j \neq i} M(i\,.j)$, where M is distance matrix and N_k is the number of items in the cluster. See Monti et al.[55] for further details.



Extended Data Fig. 9 | Canonical IPA pathway comparison analyses. Pathways are relevant to multiple UL subclasses. Pathways with the highest total score (*P* values; right-tailed Fisher's exact test) across the set of subclasses are sorted to the top. Heat map cells with insignificant *P* values (−log$_{10}$1.3) are marked with a dot.



Extended Data Fig. 10 | H2A.Z ChIP–seq fragment pileup for individual samples at the locus harbouring *CBX2*, *CBX4* and *CBX8*. Pink squares depict differential H2A.Z binding sites close to *CBX8* and the name of each of these sites refers to distance from the *CBX8* TSS in kilobases. UL subclasses are colour-coded as in the main Figures. Coordinates in GRCh38.

nature research

Corresponding author(s): Eevi Kaasinen and Lauri Aaltonen

Last updated by author(s): Jun 11, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No software was used for the data collection.

Data analysis:

SNP-array processing
B-allele frequencies (BAF) and log-R-ratios (LRR) were extracted with Illumina GenomeStudio, GC wave adjustments with PennCNV (v. 1.0.4), and allelic imbalance (AI) segments with BAF segmentation (v. 1.2.0). Subclass-specific enrichment of loss was calculated utilizing GEE-model (geepack v. 1.3-2).

RNA sequencing
The quality and adapter trimming was done with Trim Galore (v. 0.5.0). Then, the reads were aligned to the human reference genome with HISAT2 (v2.1.0). The aligned reads were assembled to transcripts with StringTie (v1.3.4d). The differential expression analysis was performed with DESeq2 (v. 1.22.2). For clustering, the effect of the sequencing batch was removed with limma (v. 3.42.0). P-values for the clustering were obtained with pvclust (v2.2-0). The most variable genes were also consensus clustered with ConsensusClusterPlus (v. 1.50.0). Linear discriminant analysis was performed with penalizedLDA (v.1.1). Preprocessing for variant calling was done with picard (v. 2.8.16) and variant calling with GATK (v.3.5 or 4.1.4.1; see Methods for details).

Allele specific expression
Allele-specific gene expression (ASE) was analysed using phASER (v1.1.1). phASER analysis was run after MarkDuplicates (Picard v2.18.16)

Pathway analysis
The pathway data was generated with Ingenuity Pathways Analysis (IPA) software (QIAGEN IPA Spring 2020 Release, Version 51963813) utilizing the z-score algorithm.

ATAC sequencing

Reads were quality and adapter trimmed with cutadapt version 1.16 in Trim Galore version 0.3.7. Trimmed reads were aligned to reference genome using Bowtie 2 (version 2.1.0). Samtools (version 1.8) was used to filter out reads with mapping quality<20, count reads mapping to the mitochondrial genome and to remove PCR duplicates. Peak calling was performed with MACS2 (version 2.1.2). Quality was checked with ataqv. Fixed-width peaks were generated using MACS2 callpeak command. Clustering of ATAC-seq samples was performed with DiffBind (version 2.14.0). Differentially accessible regions were calculated with DESeq2 v1.14.1.

ChIP sequencing
Raw sequencing reads were quality and adapter trimmed with cutadapt version 1.16 in Trim Galore version 0.3.7. Trimmed reads were aligned to the hs37d5 reference genome using Bowtie 2 (version 2.1.0) and reads with mapping quality<20 were filtered out with samtools (version 1.7). Peak calling was performed with MACS2 (version 2.1.2). Clustering and differential binding analysis of H2A.Z ChIP-seq samples was performed with DiffBind version 2.14.0 in R 3.6.3 (data without spike-in) and version 3.0.6 in R 4.0.3 (spike-in data). Duplicate reads were removed with samtools (version 1.7) rmdup from the ChIP alignment files that were utilized in the DiffBind analyses.

HiChip
HiC-Pro v. 2.11.1 was used to identify valid interaction pairs. Bowtie2 v2.3.4.1 was used to map reads to reference genome. FitHiChIP was used to call significant interactions and differential links. Differential analysis utilized EdgeR v3.26.5.

Nanopore long-read sequencing
Sequencing and base calling were performed on PromethION platform using MinKnow-Live-Basecalling (version 3.4.6). Reads were aligned with minimap2 (v2.16; preset: map-ont). Data quality was inspected with NanoStat (v1.1.2) and NanoPlot (v1.20.0). Reads were phased to parental chromosomes using Whatshap (v0.18) and genotype information from SNP arrays. Methylation status for each read at each CpG site aligned to the reference genome was called with Nanopolish. Differentially methylated loci (DML) were determined from sequenced samples with the DSS R package (version 2.28.0) utilizing bsseq R package (version 1.16.1). Allele Specific Methylation was analysed with the DSS R package (version 2.34.0). The YEATS4 mutations on nanopore data were visualized using IGV. Methylation values around YEATS4 gene were visualized using R 3.5.1 and packages data.table (version 1.13.0), ggplot2 (version 3.3.2), ggpubr (version 0.4.0) and tidyverse (version 1.3.0). Smoothing curves were produced with geom_smooth() -function from package ggplot2

Methylation expression correlation
Methylation of each sample at DMLs were correlated to RNA-seq expression values using Matrix eQTL R package v2.3

Genome-wide association study
Inverse-variance weighted fixed effects meta-analysis (R package "meta" v4.8-4). Mixed model logistic regression was computed with SAIGE (v0.35.8).

Germline loss-of-function associations to UL
Variant effect annotation was done with SnpEff (v4.3t) using the database version GRCh38.86. Gene-based association tests were computed with SAIGE-GENE (v0.42.1; SKAT-O test).

Illumina and Complete Genomics whole genome sequencing
Illumina platform data was processed with bwa (v0.6.2 aln; or v0.7.12 bwa-mem), PCR duplicate removal (SAMtools v0.1.18; or Picard MarkDuplicates v1.79), Genome Analysis Toolkit (GATK IndelRealigner and BaseRecalibrator v2.3-9 or v3.5-0), MuTect (v2.2-25-g2a68eab) and GATK SomaticIndelDetector (v2.3-9-ge5ebf34; or VarScan v2.3). For Complete Genomics data, somatic variant calling was done as a service (Complete Genomics' CGApipeline version 2.0.2.22–2.0.3.2; details in Supplementary Table 2).

Immunohistochemistry
The association between staining score and tumor subclass was calculated utilizing GEE-model (geepack v1.3-1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The peak level data utilized in the study are available for research use through Zenodo (DOI:10.5281/zenodo.4745433). Source data for Figures 1a,b,d and 3a,b,d and Extended Data Figures 1a,c,d,e, 2b,d and 8a,c and Supplementary Figure 1a are included.

Genetic data presented in this manuscript have been deposited at the European Genome-phenome Archive (EGA) under accession number EGAS00001004499. A data access committee (DAC) has been established from two University of Helsinki representatives that are independent of the authors of the current study. See Supplementary Table 26 for EGA dataset accession numbers. Requests for the data should be sent to the DAC via email (dac-finlandmyomastudy@helsinki.fi).

The DAC ensures that the intended use of data as detailed in the request is compatible with the requirements of the European General Data Protection Regulation (GDPR), consistent with the consents given and otherwise ensures the protection of data subjects' rights as required by the GDPR. The DAC will always grant access to the data if the University is legally allowed to do so without infringing the rights and freedoms of data subjects. Subject to the requirements of the GDPR, the DAC grants access to the genetic data to non-commercial academic research on neoplasia and chromatin. University of Helsinki aims at initiating processing data access requests within five business days from receipt.

Roadmap Epigenomics ChIP- and DNase-seq data (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) provided in the LOLA extended and core databases were downloaded from http://cloud.databio.org/regiondb/. Chromatin states provided in mnemonics bed files by the Roadmap Epigenomics project were downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/. GWAS cohort material was downloaded from http://jenger.riken.jp/en/ (accessed on Sept 2, 2020) and http://r3.finngen.fi (accessed on Sept 2, 2020). UK

Biobank material access can be applied from https://www.ukbiobank.ac.uk/. For RNA-seq, the reference annotation for GRCh37 was downloaded from ensembl (ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/) and for GRCh38 from Genbank (Accession: GCA_000001405.15). For RNA-seq variant calling, the SNP and indel resources were downloaded from Broad institute (https://console.cloud.google.com/storage/browser/gcp-public-data--broad-references/hg19/v0). DAR annotation data is available from http://homer.ucsd.edu/homer/data/genomes/hg19.v6.4.zip

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine the sample size. We aimed at a large sample size, in thousands rather than hundreds of ULs, to ensure appropriate representation of previously known and possible new UL subclasses. Prospectively collected hysterectomy samples from six sample collections were utilized in the study; altogether 2263 uterine leiomyomas (ULs) and 728 corresponding normal myometrium tissue samples. All the samples collected during the study period were utilized and the sample size was sufficient to reach the aims and make the presented conclusions.
Data exclusions	Tumors with a likely common clonal origin were excluded from the analysis using criteria that were established during the study. For MED12 tumors, identical MED12 mutations and at least one shared AI segment was required. For HMGA2 tumors, HMGA2 overexpression and at least one shared AI segment. For UNKNOWN tumors, at least one shared AI segment. For YEATS4 and OM tumors, the same complex gene mutation was deemed sufficient. In the FH subclass no evidence for clonal relationship emerged. Shared AI-segments were determined as segments where both start and end position matched within 250kb tolerance. From the clonally related tumor sets, one tumor was arbitrarily chosen for subsequent analyses. Only one tumor per patient entered in a statistical analysis if the statistics assumed independence between the tumors. When analyzing separately the characteristics of the different UL subclasses, we have allowed inclusion of more than one tumor per patient, if the respective tumors have entered analysis within separate subclasses. E.g. a patient with one MED12 tumor and one HMGA2 tumor was allowed to contribute to analysis of characters of MED12 tumors and HMGA2 tumors, but per analysis, only one tumor per patient was included. ChIP-seq and ATAC-seq data that did not meet our quality criteria (FRiP>=5 % for ChIP-seq, and FRiP>=5 % and TSS enrichment>=3 for ATAC-seq) were excluded from the study.
Replication	Multiple biological replicates from each UL subclass and normal tissue specimens were included in all analyses except Western blots. The number of replicates are indicated in each figure legend and/or respective results section. Due to the limited amount of the respective myometrium tissue as control, it was possible to perform the Western blots shown in Extended Data Fig. 2d only once. Here the similar results derived with H2A.Z and H2A.Zac antibodies provide confidence, and the Western blot results were compatible with immunohistochemistry findings.
Randomization	Experimental groups were not used in the study. Relevant background variables (e.g. sequencing batch, immunohistochemistry staining batch) were used as confounders in the statistical analyses of RNA-seq and immunohistochemistry stainings.
Blinding	Laboratory staff performing DNA and RNA extractions, and whole-genome and RNA sequencing library preparations were blinded to tumor subclasses. In ChIP-seq, ATAC-seq and HiChIP experiments, sample sizes were small and each tumor subclass had to be represented. Thus, blinding was not possible during DNA processing and library preparations for these three data types, except that library preparations for ChIP-seq were done blinded. Some data analyses such as clusterings were always done blinded. Blinding was not relevant in analyses comparing subclasses. Immunohistochemistry stainings and scoring were done blinded to tumor subclasses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
□	☒	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

Immunohistochemistry
HMGA2 (dilution 1:2000; Biocheck Inc., Cat. No. 59170AP, polyclonal, lot RN-60332)
Histone H2A.Z (dilution 1:2500; Abcam, Cat. No. ab150402, IgG clone EPR6171(2)(B), lot GR3254223-2)
Histone H2A.Z acetyl Lys5/Lys7/Lys11 (dilution 1:500; GeneTex, Cat. No. GTX60813, polyclonal, lot 821904190).
Orion, two components detection system, peroxidase, goat anti-rabbit/mouse IgG HRP (ready-to-use) (WellMed BV, Cat.No. T100-HRP, lot. 030519) was used for detection.

Western Blot:
Histone H2A.Z (dilution 1:500; Merck, Cat. No. 07-594, polyclonal, lot 3046749)
Histone H2A.Z acetyl Lys5/Lys7/Lys11 (dilution 1:5000; GeneTex, Cat. No. GTX60813, polyclonal, lot 821904190).
TBP (dilution 1:1000; Abcam, Cat. No. 51841, IgG clone mAbcam 51841, lot GR297600-5)
goat anti-rabbit (dilution 1:12000; Sigma, Cat. No. A6154, lot. 101M6251) for H2A.Z and H2A.Zac
goat anti-mouse (dilution 1:10000; Sigma, Cat. No. A4416, lot. SLBS1010V) for TBP

ChIP-seq:
H3K27ac (Abcam, Cat. No. ab4729, IgG polyclonal, lot GR3216173-1)
Histone H2A.Z (Merck, Cat. No. ABE1348, IgG polyclonal, lot 3275084, 3310689 and 3439804)
Histone H2A.Z (Abcam, Cat. No. ab150402, IgG clone EPR6171(2)(B), lot GR3254223-2)
H3K4me3 (Abcam, Cat. No. ab8580, polyclonal, lot. GR3275503-1)
H3K27me3 (Cell Signaling, Cat. No. 9733S, IgG clone C36B11, lot. 16)
Spike-in antibody (Active Motif, Cat. No. 61686, lot. 00419007)

Validation

Antibodies were validated by the suppliers.

HMGA2 (Biocheck Inc., Cat. No. 59170AP): Antigen source: HMGA2-P1-KLH (Synthetic peptide KLH conjugate). Specificity: HMGA2-P1. Suggested Use: for Western Blot or Immunohistochemistry (IHC) use. Applications: the antibody has been used for IHC staining on human and mouse tissues (https://doi.org/10.1038/modpathol.2010.174, https://doi.org/10.1186/s13000-017-0603-x, https://doi.org/10.1038/s41598-018-32159-x)

Histone H2A.Z (Abcam, Cat. No. ab150402): Immunogen: synthetic peptide within Human Histone H2A.Z aa 1-100. Tested applications: suitable for WB, IHC-P, ICC/IF, Flow Cyt, ChIP. Species reactivity: Mouse, Rat, Human.

Histone H2A.Z acetyl Lys5/Lys7/Lys11 (GeneTex, Cat. No. GTX60813): Immunogen: the region of histone H2A.Z containing the acetylated lysines 4, 7 and 11, using a KLH-conjugated synthetic peptide. Application: WB, ICC/IF, Dot, ELISA, ChIP assay. Reactivity: Human, Mouse.

Histone H2A.Z (Merck, Cat. No. 07-594): Immunogen: peptide containing the sequence SLIGKKGQQ, corres-ponding to the C-terminus of human histone H2A.Z. Species reactivity: Human. Key applications: Western blotting. “Detect Histone H2A.Z also known as H2AZ histone with Anti-Histone H2A.Z Antibody (Rabbit Polyclonal Antibody) that has been demonstrated to work in WB.”
TBP (Abcam, Cat. No. 51841): Immunogen: Synthetic peptide corresponding to Human TATA binding protein TBP aa 1-100 conjugated to keyhole limpet haemocyanin. Species reactivity: Reacts with: Human; predicted to work with: Mouse, Rat, Chicken, Cow, Xenopus laevis, Chimpanzee, Zebrafish. Tested applications: Suitable for WB, ChIP, ICC/IF, Flow Cyt, IP, IHC-P.

H3K27ac (Abcam, Cat. No. ab4729): Immunogen: Synthetic peptide corresponding to Human Histone H3 aa 1-100 (acetyl K27) conjugated to keyhole limpet haemocyanin. Tested applications: suitable for ICC/IF, WB, IHC-P, ChIP, PepArr. Species reactivity: Reacts with Mouse, Cow, Human; predicted to work with Rat, Chicken, Xenopus laevis, Arabidopsis thaliana, Drosophila melanogaster, Monkey, Zebrafish, Plasmodium falciparum, Rice, Cyanidioschyzon merolae.
Histone H2A.Z (Merck, Cat. No. ABE1348): Immunogen: KLH-conjugated linear peptide corresponding to region the near C-terminus of Human Histone H2A.Z. Tested applications: Western Blotting, Chromatin Immunoprecipitation (ChIP), ChIP-seq, Immunocytochemistry, ELISA. Species reactivity: Human, wide range expected.

H3K4me3 (Abcam, Cat. No. ab8580): Immunogen: Synthetic peptide within Human Histone H3 aa 1-100 (tri methyl K4) conjugated to keyhole limpet haemocyanin. The exact sequence is proprietary. Tested applications: suitable for PepArr, ChIP, WB, IHC-P, ICC/IF. Species reactivity: reacts with Cow, Human; predicted to work with Mouse, Rat, Rabbit, Pig, Saccharomyces cerevisiae, Tetrahymena, Xenopus laevis, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Indian muntjac, Oikopleura, Plants, Zebrafish, Mammals, Trypanosoma cruzi, Common marmoset, Rice, Xenopus tropicalis.

H3K27me3 (Cell Signaling, Cat. No. 9733S): “Tri-Methyl-Histone H3 (Lys27) (C36B11) Rabbit mAb detects endogenous levels of histone H3 only when tri-methylated on Lys27. The antibody does not cross-react with non-methylated, mono-methylated or di-methylated Lys27. In addition, the antibody does not cross-react with mono-methylated, di-methylated or tri-methylated histone H3 at Lys4, Lys9, Lys36 or Histone H4 at Lys20.” Species Reactivity: Human, Mouse, Rat, Monkey. Species predicted to react based on 100% sequence homology: Xenopus, Zebrafish.

Spike-in antibody (Active Motif, Cat. No. 61686): “The Spike-in Antibody recognizes a histone variant (H2Av) that is specific to the species of the Spike-in Chromatin (Drosophila). Each lot of Spike-in Chromatin is quantified and tested with the Spike-in Antibody. This enables specific detection of the Spike-in Chromatin without any significant increase in background signal.”

Human research participants

Policy information about studies involving human research participants

Population characteristics

All participants were females who underwent hysterectomy and had at least one uterine leiomyoma. Patient chartacteristiscs are reported in Supplementary Table 1 and Supplementary Figure 1.

Recruitment

The sample set consists of six prospectively collected sample series (M, My, My1000, My5000, My6000 and My8000). The anonymous M-sample series was collected according to Finnish laws and regulations after authorization from the director of the health care unit, between the years 2001 and 2002. For all subsequent samples, a written informed consent was obtained. Participants were not compensated. See full details of sample collection in the Methods section. Self-selection bias did not affect the study.

Ethics oversight

The study was conducted in accordance with the Declaration of Helsinki and approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00/2017, THL/723/5.05.00/2018), and the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/2509/2016).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

https://www.ebi.ac.uk/ega/studies/EGAS00001004499

Files in database submission

Raw data in fastq format, aligned data in bam files and peak calls in BED files are deposited to EGA under accession number EGAS00001004499. Peaks calls in BED files are also available through the Zenodo platform under DOI:10.5281/zenodo.4745433

Genome browser session
(e.g. UCSC)

no longer applicable

Methodology

Replicates

Technical replicates with two different antibodies (denoted with "H2A.Z-ABE1348" and "H2A.Z-ab150402") from one normal sample (M23N) were done and they cluster together in principal component analysis (PCA) of normalized ChIP-seq read counts in peaks. ChIP-seq with H2A.Z-ABE1348 antibody were performed without spike-in and with spike-in chromatin (spike-in samples denoted with "H2A.Z-ABE1348-s"). Biological replicates were used to study H2A.Z deposition changes in MED12, HMGA2, HMGA1, FH, YEATS4 and OM UL subclasses as compared to normal myometrium samples.

Sequencing depth

All ChIP-seq data were sequenced with single-end 100bp reads.
My5001N1_H2A.Z-ABE1348 (normal): total reads 25632259; total reads mapped (q20) 17277166; total nonredundant reads mapped (q20) 16333776
My5005N1_H2A.Z-ABE1348 (normal): total reads 27912055; total reads mapped (q20) 24387081; total nonredundant reads mapped (q20) 22441087
My5006m1_H2A.Z-ABE1348 (FH): total reads 23859084; total reads mapped (q20) 21077443; total nonredundant reads mapped (q20) 20382967
My5006N1_H2A.Z-ABE1348 (normal): total reads 33275331; total reads mapped (q20) 29858161; total nonredundant reads mapped (q20) 28467025
My5016m1_H2A.Z-ABE1348 (HMGA2): total reads 24412808; total reads mapped (q20) 20787099; total nonredundant reads mapped (q20) 19091840
My5016N3_H2A.Z-ABE1348 (normal): total reads 26097871; total reads mapped (q20) 20580230; total nonredundant reads mapped (q20) 19855593
My5019m1_H2A.Z-ABE1348 (HMGA2): total reads 25516910; total reads mapped (q20) 21032343; total nonredundant reads mapped (q20) 20705395
My5019N1_H2A.Z-ABE1348 (normal): total reads 24188544; total reads mapped (q20) 21023340; total nonredundant reads mapped (q20) 20711622
My6027m1_H2A.Z-ABE1348 (OM): total reads 20522809; total reads mapped (q20) 17951357; total nonredundant reads mapped (q20) 17482469
My6183m9_H2A.Z-ABE1348 (OM): total reads 22712851; total reads mapped (q20) 20072266; total nonredundant reads mapped (q20) 19246439
My6229m1_H2A.Z-ABE1348 (YEATS4): total reads 22191564; total reads mapped (q20) 19210685; total nonredundant reads mapped (q20) 18771014
My6251m1_H2A.Z-ABE1348 (HMGA2): total reads 25246013; total reads mapped (q20) 22812776; total nonredundant reads mapped (q20) 21820445
My6252m1_H2A.Z-ABE1348 (HMGA2): total reads 21103033; total reads mapped (q20) 18979781; total nonredundant reads mapped (q20) 17979806
My6252N1_H2A.Z-ABE1348 (normal): total reads 25185591; total reads mapped (q20) 22575174; total nonredundant reads mapped (q20) 21440831

My6334m2_H2A.Z-ABE1348 (OM): total reads 22760804; total reads mapped (q20) 19910684; total nonredundant reads mapped (q20) 19091878

My6398m1_H2A.Z-ABE1348 (YEATS4): total reads 21679504; total reads mapped (q20) 18730494; total nonredundant reads mapped (q20) 17531206

My6398N1_H2A.Z-ABE1348 (normal): total reads 21202381; total reads mapped (q20) 18924368; total nonredundant reads mapped (q20) 18340549

M23m1_H2A.Z-ab150402 (MED12): total reads 27615111; total reads mapped (q20) 22269626; total nonredundant reads mapped (q20) 18487992

M23m5_H2A.Z-ABE1348 (YEATS4): total reads 27874129; total reads mapped (q20) 16659850; total nonredundant reads mapped (q20) 15571358

M23N_H2A.Z-ab150402 (normal): total reads 28346970; total reads mapped (q20) 18138653; total nonredundant reads mapped (q20) 17217479

M23N_H2A.Z-ABE1348 (normal): total reads 28973014; total reads mapped (q20) 23735628; total nonredundant reads mapped (q20) 21173219

M22m1_H3K27ac (MED12): total reads 23975593; total reads mapped (q20) 21546379; total nonredundant reads mapped (q20) 20201560

M22N_H3K27ac (normal): total reads 27232589; total reads mapped (q20) 24867957; total nonredundant reads mapped (q20) 24276232

My5002N1_H2A.Z-ABE1348 (normal): total reads 20913722; total reads mapped (q20) 18730273; total nonredundant reads mapped (q20) 10981249

My5002m2_H2A.Z-ABE1348 (MED12): total reads 21108087; total reads mapped (q20) 18616767; total nonredundant reads mapped (q20) 16728964

My5008N1_H2A.Z-ABE1348 (normal): total reads 21190911; total reads mapped (q20) 19008930; total nonredundant reads mapped (q20) 17774208

My5008m1_H2A.Z-ABE1348 (MED12): total reads 21104257; total reads mapped (q20) 18482428; total nonredundant reads mapped (q20) 17415820

My5012N1_H2A.Z-ABE1348 (normal): total reads 21341773; total reads mapped (q20) 19242613; total nonredundant reads mapped (q20) 17827786

My5012m1_H2A.Z-ABE1348 (MED12): total reads 20665854; total reads mapped (q20) 18275317; total nonredundant reads mapped (q20) 12523974

My6139m1_H2A.Z-ABE1348 (FH): total reads 21047451; total reads mapped (q20) 18494069; total nonredundant reads mapped (q20) 17070175

My6213m1_H2A.Z-ABE1348 (FH): total reads 20960661; total reads mapped (q20) 18828416; total nonredundant reads mapped (q20) 17090231

My6217m1_H2A.Z-ABE1348 (FH): total reads 29751548; total reads mapped (q20) 26119600; total nonredundant reads mapped (q20) 25632077

My6011N1_H2A.Z-ABE1348-s (normal): total reads 24556084; total reads mapped (q20) 22154691; total nonredundant reads mapped (q20) 19949680

My6018m1_H2A.Z-ABE1348-s (HMGA1): total reads 20235398; total reads mapped (q20) 18170790; total nonredundant reads mapped (q20) 16960888

My6023m1_H2A.Z-ABE1348-s (HMGA1): total reads 21838268; total reads mapped (q20) 19803470; total nonredundant reads mapped (q20) 18295016

My6047m6_H2A.Z-ABE1348-s (HMGA1): total reads 22730982; total reads mapped (q20) 20456750; total nonredundant reads mapped (q20) 19036888

My6155N1_H2A.Z-ABE1348-s (normal): total reads 24076826; total reads mapped (q20) 21966087; total nonredundant reads mapped (q20) 19585513

My6267m1_H2A.Z-ABE1348-s (HMGA1): total reads 20676457; total reads mapped (q20) 18514917; total nonredundant reads mapped (q20) 16966894

My5001N1_H2A.Z-ABE1348-s (normal): total reads 23890056; total reads mapped (q20) 21898988; total nonredundant reads mapped (q20) 20632330

My5002m2_H2A.Z-ABE1348-s (MED12): total reads 25612966; total reads mapped (q20) 22863666; total nonredundant reads mapped (q20) 22284503

My5005N1_H2A.Z-ABE1348-s (normal): total reads 28529819; total reads mapped (q20) 26252442; total nonredundant reads mapped (q20) 24619155

My5006m1_H2A.Z-ABE1348-s (FH): total reads 26275664; total reads mapped (q20) 23574702; total nonredundant reads mapped (q20) 22806143

My5008m1_H2A.Z-ABE1348-s (MED12): total reads 21902184; total reads mapped (q20) 19554272; total nonredundant reads mapped (q20) 19169881

My5016m1_H2A.Z-ABE1348-s (HMGA2): total reads 24560464; total reads mapped (q20) 22564698; total nonredundant reads mapped (q20) 21419925

My5019m1_H2A.Z-ABE1348-s (HMGA2): total reads 25434726; total reads mapped (q20) 23080645; total nonredundant reads mapped (q20) 22238952

My6027m1_H2A.Z-ABE1348-s (OM): total reads 24490610; total reads mapped (q20) 22022753; total nonredundant reads mapped (q20) 21262269

My6217m1_H2A.Z-ABE1348-s (FH): total reads 22704681; total reads mapped (q20) 20602294; total nonredundant reads mapped (q20) 19828203

My6229m1_H2A.Z-ABE1348-s (YEATS4): total reads 25208604; total reads mapped (q20) 22440871; total nonredundant reads mapped (q20) 21640442

My6334m2_H2A.Z-ABE1348-s (OM): total reads 24254680; total reads mapped (q20) 21807497; total nonredundant reads mapped (q20) 21058179

My6398m1_H2A.Z-ABE1348-s (YEATS4): total reads 25080752; total reads mapped (q20) 22251109; total nonredundant reads mapped (q20) 12752583

My6024m1_H2A.Z-ABE1348-s (HMGA2): total reads 24782337; total reads mapped (q20) 21651208; total nonredundant reads mapped (q20) 19467977

My6026N1_H2A.Z-ABE1348-s (normal): total reads 22178143; total reads mapped (q20) 20076198; total nonredundant reads mapped (q20) 17975294

My6251N1_H2A.Z-ABE1348-s (normal): total reads 26069412; total reads mapped (q20) 23119338; total nonredundant reads mapped (q20) 22257309
My6252N1_H2A.Z-ABE1348-s (normal): total reads 21671208; total reads mapped (q20) 19216820; total nonredundant reads mapped (q20) 16719997
My6254m1_H2A.Z-ABE1348-s (HMGA2): total reads 23319886; total reads mapped (q20) 20566506; total nonredundant reads mapped (q20) 18858601
M23N_H3K4me3-ab8580-Mnase (normal): total reads 32027413; total reads mapped (q20) 27096974; total nonredundant reads mapped (q20) 26002570
My5001N1_H3K4me3-ab8580-Mnase (normal): total reads 36465966; total reads mapped (q20) 31057140; total nonredundant reads mapped (q20) 30108262
My5005N1_H3K4me3-ab8580-Mnase (normal): total reads 29315529; total reads mapped (q20) 25825087; total nonredundant reads mapped (q20) 24503410
My5016N3_H3K4me3-ab8580-Mnase (normal): total reads 22519916; total reads mapped (q20) 19967245; total nonredundant reads mapped (q20) 15782432
My5006N1_H3K4me3-ab8580-Mnase (normal): total reads 20881237; total reads mapped (q20) 17972548; total nonredundant reads mapped (q20) 13878256
My6027N1_H3K4me3-ab8580-Mnase (normal): total reads 20960360; total reads mapped (q20) 17760403; total nonredundant reads mapped (q20) 14613543
My6183N1_H3K4me3-ab8580-Mnase (normal): total reads 20000881; total reads mapped (q20) 16275455; total nonredundant reads mapped (q20) 14820688
M23N_H3K27me3-9733S-Mnase (normal): total reads 30223145; total reads mapped (q20) 26707721; total nonredundant reads mapped (q20) 25798371
My5001N1_H3K27me3-9733S-Mnase (normal): total reads 33686418; total reads mapped (q20) 29601383; total nonredundant reads mapped (q20) 28781593
My5016N3_H3K27me3-9733S-Mnase (normal): total reads 27162893; total reads mapped (q20) 23949654; total nonredundant reads mapped (q20) 22793137
My5019N1_H3K27me3-9733S-Mnase (normal): total reads 24311326; total reads mapped (q20) 21548096; total nonredundant reads mapped (q20) 20024684
My6027N1_H3K27me3-9733S-Mnase (normal): total reads 20811179; total reads mapped (q20) 18319871; total nonredundant reads mapped (q20) 16776348
My6183N1_H3K27me3-9733S-Mnase (normal): total reads 20496847; total reads mapped (q20) 18000518; total nonredundant reads mapped (q20) 16631096
M23N_H3K27ac-ab4729 (normal): total reads 21053090; total reads mapped (q20) 19471805; total nonredundant reads mapped (q20) 18433747
My5006N1_H3K27ac-ab4729 (normal): total reads 21119008; total reads mapped (q20) 19825102; total nonredundant reads mapped (q20) 18569609
My5019N1_H3K27ac-ab4729 (normal): total reads 23013969; total reads mapped (q20) 21419565; total nonredundant reads mapped (q20) 20492501
My6027N_H3K27ac-ab4729 (normal): total reads 24232162; total reads mapped (q20) 22541194; total nonredundant reads mapped (q20) 21650246
M23m5_H3K27ac-ab4729 (YEATS4): total reads 23960151; total reads mapped (q20) 22388970; total nonredundant reads mapped (q20) 21430165
My5006m1_H3K27ac-ab4729 (FH): total reads 20538883; total reads mapped (q20) 19210578; total nonredundant reads mapped (q20) 18533643
My5019m1_H3K27ac-ab4729 (HMGA2): total reads 20185481; total reads mapped (q20) 18900937; total nonredundant reads mapped (q20) 18197526
My6027m1_H3K27ac-ab4729 (OM): total reads 20812459; total reads mapped (q20) 19472371; total nonredundant reads mapped (q20) 18658261

Antibodies

H3K27ac (Abcam, Cat. No. ab4729, lot GR3216173-1)
H2A.Z ("H2A.Z-ABE1348"; Merck, Cat. No. ABE1348, lot 3275084, 3310689 and 3439804)
H2A.Z ("H2A.Z-ab150402"; Abcam, Cat. No. ab150402, clone EPR6171(2)(B), lot GR3254223-2)
H3K4me3 (Abcam, Cat. No. ab8580, lot. GR3275503-1)
H3K27me3 (Cell Signaling, Cat. No. 9733S, lot. 16)
Spike-in antibody (Active Motif, Cat. No. 61686, lot. 00419007)
Spike-in chromatin (Active Motif, Cat. No. 53083, lot. 06420010)

Peak calling parameters

Trimmed reads were aligned to hs37d5 reference genome using Bowtie 2 and reads with mapping quality<20 were filtered out with samtools (version 1.7) view -q 20. Peaks were called with MACS2 callpeak function using parameters --broad -q 0.01 and input for each biological replicate as a control. Peak calling for pooled normals was performed utilizing aligned reads from individual samples after duplicate removal (samtools version 1.7), merging (samtools version 1.7) and running MACS2 (version 2.1.2) with default parameters except --keep-dup that was set to the value that corresponded to the number of pooled normals.

Data quality

Each ChIP included in the study was required to have at least 5% of reads in broadPeaks. Using this cutoff to include samples in the study, median and mean number of broad peaks at FDR 5% and 5-fold increase were 27911 and 30827, respectively.

Software

Trim Galore version 0.3.7, Bowtie 2 version 2.1.0, samtools version 1.7, MACS version 2.1.2, DiffBind version 2.14.0 and version 3.0.6