HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2021 May 11.

Published in final edited form as:

Cell Rep. 2021 March 30; 34(13): 108927. doi:10.1016/j.celrep.2021.108927.

Epigenomic tensor predicts disease subtypes and reveals constrained tumor evolution

Jacob R. Leistico[1,2,10], **Priyanka Saini**[3,8,10], **Christopher R. Futtner**[4,8], **Miroslav Hejna**[1,2], **Yasuhiro Omura**[4,8], **Pritin N. Soni**[3,8], **Poorva Sandlesh**[3,8], **Magdy Milad**[3,8], **Jian-Jun Wei**[5,7,8], **Serdar Bulun**[3,8], **J. Brandon Parker**[3,8], **Grant D. Barish**[4,8,11], **Jun S. Song**[1,2,9,11,*], **Debabrata Chakravarti**[3,6,7,11,12,*]

[1]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[2]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[3]Department of Obstetrics and Gynecology, Northwestern University, Chicago, IL, USA

[4]Department of Medicine, Northwestern University, Chicago, IL, USA

[5]Department of Pathology, Northwestern University, Chicago, IL, USA

[6]Department of Pharmacology, Northwestern University, Chicago, IL, USA

[7]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL, USA

[8]Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

[9]Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[10]These authors contributed equally

[11]Senior author

[12]Lead contact

SUMMARY

Understanding the epigenomic evolution and specificity of disease subtypes from complex patient data remains a major biomedical problem. We here present DeCET (decomposition and classification of epigenomic tensors), an integrative computational approach for simultaneously analyzing hierarchical heterogeneous data, to identify robust epigenomic differences among tissue types, differentiation states, and disease subtypes. Applying DeCET to our own data from 21

*Correspondence: songj@illinois.edu (J.S.S.), debu@northwestern.edu (D.C.).

AUTHOR CONTRIBUTIONS

Conceptualization, D.C., J.S.S., and G.D.B.; methodology, D.C., J.S.S., and G.D.B.; investigation, P. Saini; software and formal analysis, J.R.L. and M.H.; writing – original draft, J.R.L., P. Saini, G.D.B., J.S.S., and D.C.; writing – review & editing, all authors; funding acquisition, D.C., J.S.S., and G.D.B.; supervision, D.C., J.S.S., G.D.B., and J.B.P.; resources, J.B.P., G.D.B., M.M., S.B., and J.-J.W; technical, C.R.F., Y.O., P.N.S., and P. Sandlesh.

SUPPLEMENTAL INFORMATION

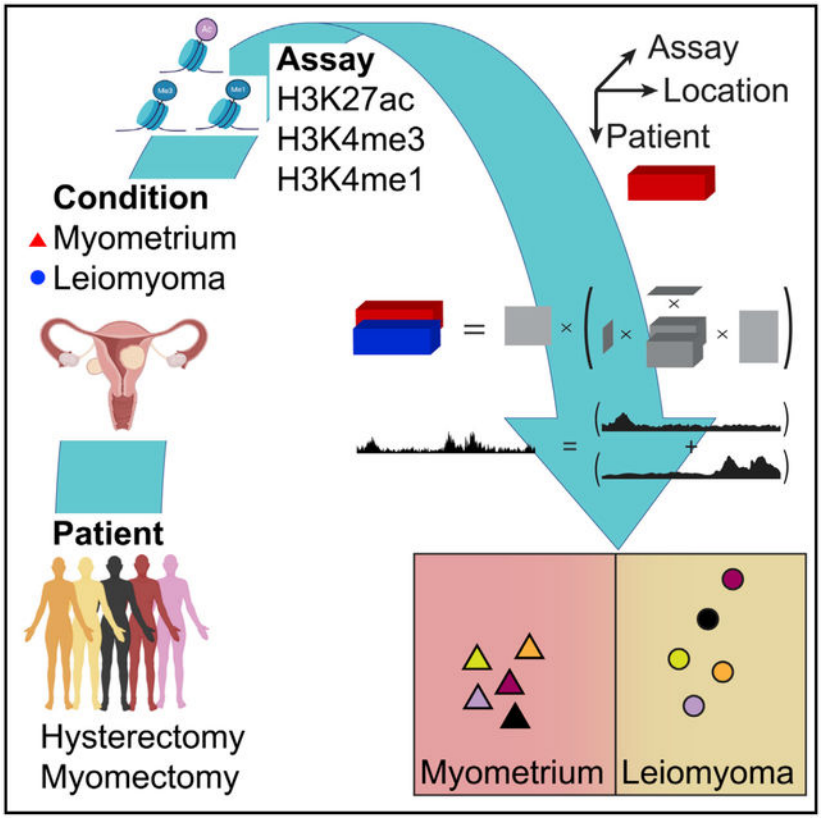
Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.108927.

DECLARATION OF INTERESTS

The authors declare no competing interests.

uterine benign tumor (leiomyoma) patients identifies distinct epigenomic features discriminating normal myometrium and leiomyoma subtypes. Leiomyomas possess preponderant alterations in distal enhancers and long-range histone modifications confined to chromatin contact domains that constrain the evolution of pathological epigenomes. Moreover, we demonstrate the power and advantage of DeCET on multiple publicly available epigenomic datasets representing different cancers and cellular states. Epigenomic features extracted by DeCET can thus help improve our understanding of disease states, cellular development, and differentiation, thereby facilitating future therapeutic, diagnostic, and prognostic strategies.

Graphical Abstract



In brief

Leistico et al. apply tensor decomposition and classification methods to integrate information from hierarchical heterogenous epigenomic datasets and identify histone modification patterns that discriminate disease conditions, tissue types, and differentiation states. Leiomyomas are shown to possess alterations in distal enhancers and large-scale regions confined to chromatin contact domains.

INTRODUCTION

Analyzing heterogeneous epigenomic data from multiple disease conditions and patients poses challenges partly attributable to technical biases in experiments and biological

variability among individuals. Most current analysis methods consider each histone modification per patient separately and then perform a meta-analysis by pooling information across epigenetic marks and patients; unfortunately, this approach suffers from loss of statistical power and requires arbitrary choices of parameters for combining individual results. We here present a powerful approach called DeCET (decomposition and classification of epigenomic tensors) that simultaneously analyzes hierarchical heterogeneous histone modification chromatin immunoprecipitation sequencing (ChIP-seq) datasets.

DeCET employs the higher-order singular value decomposition (HOSVD) of a data tensor to integrate the information of all patients, conditions (healthy versus diseased), histone modifications, and genomic locations (De Lathauwer et al., 2000) (Figure 1A; STAR Methods). By using information from all experiments, the HOSVD captures epigenomic alterations robust to experimental biases and inter-patient variability. Moreover, as the data tensor directly uses normalized ChIP-seq signals rather than peak calls, DeCET is able to identify modulation in regulatory activity beyond simple binary gain or loss events. The HOSVD simultaneously decomposes the dataset into characteristic modes in the patient, condition, assay, and genomic location spaces while capturing the interactions between these spaces in a compressed version of the tensor. In particular, the spatial decomposition yields location vectors that encode independent epigenomic patterns exhibiting spatial covariation across samples (Figure 1A). The linear HOSVD provides a direct connection between the orthogonal location vectors and the eigenmodes in the condition, patient, and assay spaces, enabling sample characterization and biological discovery.

We demonstrate the clinical applicability of DeCET by analyzing the intact epigenomes of fresh human leiomyoma and matched myometrium tissues, unperturbed by artificial *in vitro* cell culture conditions. Uterine leiomyomas (fibroids) are benign tumors of uterine smooth muscle cells characterized by deposition of excessive, disorganized extracellular matrix (ECM) (Al-Hendy et al., 2017; Commandeur et al., 2015; Doherty et al., 2014; Stewart et al., 2016). Studies have shown four major mutually exclusive clonal genetic mutations in leiomyomas: ~70% of cases harbor *MED12* mutations, either missense or small in-frame insertions or deletion in exon 2 (Je et al., 2012; Mäkinen et al., 2011; McGuire et al., 2012; Mehine et al., 2014), while high mobility group AT-hook 1 and 2 (*HMGA1* and *HMGA2*) rearrangements occur in another 15% (Ferrero, 2019; Meloni et al., 1992; Nibert and Heim, 1990; Rein et al., 1991). Biallelic inactivation of fumarate hydratase (*FH*) (Tomlinson et al., 2002) and collagen type IV alpha 5 and collagen type IV alpha 6 (*COL4A5-COL4A6*) deletions are also found in some cases (Mehine et al., 2016). Recent studies have observed characteristic changes in RNA expression (Mehine et al., 2016), DNA methylation (George et al., 2019), and H3K27ac (Moyo et al., 2020) for leiomyomas harboring distinct driver mutations, but interrogating how aberrant epigenomic structure contributes to transcriptomic alterations and leiomyoma pathology remains a difficult problem. DeCET identified informative epigenetic features that accurately distinguished myometrium and leiomyoma samples with mutational subtypes. When applied to other epigenomic datasets, DeCET also discriminated among tissue types, differentiation states, prostate cancer subtypes, and breast cancer subtypes, demonstrating its power and generalizability.

Moreover, a supervised support tensor machine (STM) classifier (Cai et al., 2006) using the epigenomic features identified by DeCET yield a robust predictor of the disease condition and subtype of unseen leiomyoma samples. The STM integrates information from all histone modifications, reducing the number of model parameters and thereby avoiding overfitting compared to other approaches that treat the modifications separately.

RESULTS

DeCET uncovers distinct epigenetic patterns specific to myometrium and leiomyoma subtypes

We profiled active histone modifications (H3K27ac, H3K4me3, and H3K4me1) (ENCODE Project Consortium, 2012; Kundaje et al., 2015) in matched myometrium and leiomyoma tissues from 21 patients. Histone modifications for 6 leiomyoma samples from 5 additional patients and a second tumor from 1 of the 21 matched patients were also profiled to be used as test data for prediction (Table S1).

Applying the HOSVD to the tensor of 21 patient-matched datasets showed that the first three genomic location vectors specified histone modification patterns highly conserved across conditions and patients, while the remaining vectors specified heterogeneities in the samples (Figure 1B; STAR Methods). To test whether the top location vectors could separate the tissues into conditions, we clustered the samples using the HOSVD projections of histone modifications onto these vectors. The clustering correctly partitioned the samples by disease condition and also revealed additional substructure within the leiomyomas (Figure 1C; STAR Methods). We investigated whether this substructure corresponded to known driver mutations (Mehine et al., 2014, 2016). We first verified the status of *MED12* exon 2 mutations, *HMGA2* overexpression, and biallelic inactivation of *FH* for 20 of the 21 patient-matched leiomyoma samples using RNA sequencing (RNA-seq), Sanger sequencing, or qRT-PCR (Figures S1 and S2; STAR Methods). Of these samples, all but three possessed one of the three driver mutations (12 *MED12* mutation, 3 *HMGA2* overexpression, and 2 biallelic loss of *FH*), and none possessed multiple alterations (Table S1). For the one patient without RNA-seq data, we confirmed by Sanger sequencing that this patient possessed a mutation in exon 2 of *MED12* but were unable to check for loss of *FH* or *HMGA2* overexpression. The clustering substructure clearly reflected the mutation-based subtypes, demonstrating that DeCET successfully uncovered clinically relevant epigenomic aberrations (Figure 1C). The three samples without any of the three driver mutations clustered together, potentially attributable to shared epigenomic alterations. We did not observe any tight clustering or consistent changes for leiomyomas treated with the gonadotrophin-releasing hormone agonist leuprorelin, suggesting its heterogeneous or weak effect on histone modifications.

The clustering of leiomyoma epigenomes by distinct driver mutations supported that leiomyoma genesis involves epigenetic reprogramming specific to each driver mutation. An analogous clustering using RNA-seq data from the 20 matched samples showed that these mutations were also associated with distinct transcriptomic states (Figure S3A), agreeing with a recent report (Mehine et al., 2016). Together, these results implied that mutation-

specific distal and local epigenomic alterations might mediate distinct transcriptional alterations in leiomyoma.

We next sought to identify the specific genomic locations separating the healthy versus disease conditions and leiomyoma subtypes. Once the location space was decomposed into singular vectors, the information distinguishing leiomyoma from myometrium was encoded in the compressed core tensor and the orthogonal decompositions in the condition, patient and assay spaces (STAR Methods). To identify the discriminatory location vectors, we thus summed over the decompositions of the condition, patient and assay spaces to represent each ChIP-seq data as a collection of projections onto the location vectors (Figure 1B; STAR Methods). For each histone modification, a one-way analysis of variance (ANOVA) of the projections onto each location vector identified vectors along which the tumor and healthy samples diverged. We also compared the within- and between-condition variances across all histone modifications to rank the separation of these projections between conditions (STAR Methods). Each histone modification individually, as well as the combination of all three together, showed clear separation between the healthy and leiomyoma conditions along the fourth location vector (Figures 1B, 1C, and S3B–S3E) (one-way ANOVA F = 183.5, 90.2, 145.8, p values = 1.1×10^{-16}, 8.4×10^{-12}, and 6.4×10^{-15} for H3K27ac, H3K4me3, and H3K4me1, respectively; STAR Methods).

Non-zero entries in this fourth vector represented genomic regions with aberrant histone modification recurrent among patients. To identify robust alterations, we used the empirical distribution of all HOSVD location vector entries to set a significance threshold (Figure S3F; STAR Methods). We thereby identified 1,818 regions (2-kb bins) with higher levels and 1,306 regions (2-kb bins) with lower levels of activating histone modifications in leiomyoma relative to matched myometrium. Plotting the heatmap of ChIP-seq signals in the significant locations clearly showed qualitative differences between healthy and disease conditions (Figures S4A–S4C), confirming the discriminatory role of these regions. Visual inspection of the identified regions further confirmed the presence of differential modifications with expected expression changes of nearby genes (Figure 2A).

A similar analysis identified the seventh location vector as strongly separating the tumors by *MED12* exon 2 mutation status (one-way ANOVA F = 67.3, 45.9, 59.1, p values = 1.1×10^{-7}, 1.8×10^{-6}, and 3.0×10^{-7} for H3K27ac, H3K4me3, and H3K4me1, respectively; STAR Methods). The variance of the projection along this vector among myometrium samples was much less than that among the leiomyoma samples (Levene's test W = 44.0, 47.9, 35.3, p values = 6.1×10^{-8}, 2.4×10^{-8}, and 5.7×10^{-7} for H3K27ac, H3K4me3, H3K4me1, respectively; STAR Methods), indicating that this vector encoded leiomyoma subtype-specific differences not present among the myometrium samples. Using the same significance threshold (Figure S3F; STAR Methods), we found 1,662 and 1,576 2-kb regions with higher activating histone modifications in *MED12* mutant (*MED12*-mut) and *MED12*-wild-type (*MED12*-WT) leiomyomas, respectively (Figures S4D–S4F and S5A). Qualitative differences between leiomyoma subtypes were again visible in the heatmaps. Most of these regions (81%) were distinct from those identified as separating leiomyomas from myometrium, indicating that they predominantly embodied new mutation-specific alterations.

DeCET yields a robust epigenomic predictor of healthy and disease conditions

The informative unsupervised clustering (Figure 1C) suggested that a supervised classifier using histone modifications may robustly predict the disease conditions. We used the HOSVD result to perform dimensionality reduction and represent each tissue sample as a 3 × 10 matrix, consisting of the projections of the three histone modification profiles onto the first 10 location vectors. Using this representation, a STM classifier (Cai et al., 2006), trained on the 21 matched samples and tested on the additional 7 leiomyoma samples (Figure 2B; STAR Methods), had 100% training and test accuracies (Figure 2C).

To evaluate the robustness of DeCET, we applied leave-one-out cross-validation (CV) (STAR Methods). For each of the 21 patients with matched leiomyoma and myometrium data, we removed the selected patient and performed the HOSVD of the reduced data tensor; a new STM classifier was then trained on the 20 patients and tested on the left-out patient and 7 additional tumor samples (STAR Methods). The training and test accuracies were 100% for each of the 21 CV steps, demonstrating that the discriminatory epigenomic features were generalizable across patients.

We next investigated whether a single histone modification was sufficient to classify healthy versus disease conditions by training a support vector machine (SVM) classifier on each row of the 3 × 10 matrix representations (STAR Methods). This approach still integrated information from the HOSVD dimensionality reduction of the full set of histone modifications but used the resulting compression of only a single histone modification to classify samples. Additionally, we performed leave-one-out CV to evaluate the robustness of each classifier. The classifier using the projections for H3K27ac correctly classified the 21 matched training samples and 7 leiomyoma test samples, suggesting it may serve as an epigenomic leiomyoma biomarker. The classifier using H3K4me1 misclassified one of the leiomyoma training samples, while the classifier using H3K4me3 misclassified two of the leiomyoma training samples and one of the additional test samples. The misclassified samples for the respective classifiers using the full training set were also consistently misclassified during CV. In addition, the classifier using H3K4me3 frequently misclassified an additional matched leiomyoma sample, while the classifier using H3K27ac misclassified the same additional test sample in two CV steps. These results showed that the disease condition of most samples could be accurately inferred from the HOSVD compression of a single histone modification profile.

Epigenomic alterations accurately classify leiomyoma subtypes

Leiomyomas harboring *MED12* mutations may exhibit distinct alterations in DNA methylation (George et al., 2019) and transcription (Mehine et al., 2016). A recent study also showed differential H3K27ac modifications between myometrium and *MED12*-mut leiomyomas (Moyo et al., 2020) but did not compare these to *MED12*-WT leiomyomas. Given the separation of *MED12*-mut from *MED12*-WT leiomyomas (Figure 1C) and the subtype differences detected by the seventh location vector (Figures S4D–S4F and S5A), we applied the DeCET prediction method to classify leiomyoma subtypes.

As before, an STM classifier was trained on the HOSVD projections of leiomyoma samples to predict the presence of *MED12* mutations. We trained the classifier on the 21 leiomyoma samples (13 *MED12*-mut) with matched myometrium using the full set of histone modifications and tested it on the 7 additional leiomyoma samples (4 *MED12*-mut). This classifier correctly classified the *MED12* mutation status of all 21 training and 7 test leiomyoma samples (Figure S5B). We also evaluated the robustness of this classifier using leave-one-out CV (STAR Methods). For 20 of the CV steps, the classifier correctly predicted the *MED12* mutation status of the 20 training, 1 left-out validation, and 7 additional test samples; for one CV step, only one test sample was misclassified.

We next examined the classification ability of a single histone modification projected onto the HOSVD location vectors obtained from the full data tensor (STAR Methods). When the full set of 21 training samples was used, the classifier for each histone modification correctly predicted the *MED12* mutation status of the 21 training and 7 test samples. The robustness of each classifier was assessed using leave-one-out CV. For the classification using H3K4me1 or H3K4me3, each CV step yielded correct classification of all training and test samples. For the classification using H3K27ac, one additional test sample was misclassified in two CV steps and a second additional test sample was misclassified in a single CV step; one of the matched leiomyoma samples was also misclassified as a test sample in the CV.

Widespread alterations in distal enhancer activity and long-range spatial correlation of epigenetic perturbations define uterine leiomyoma

To characterize the regulatory function of the regions identified by the fourth location vector, we projected the differential regions onto the condition and assay singular vectors (Figure 1A; STAR Methods). The condition space was decomposed into two orthogonal vectors representing features shared and different between the two conditions, respectively; to study aberrant alterations in leiomyoma, we fixed the condition index of the core tensor on the vector separating the conditions. Similarly, the orthogonal singular vectors of the assay space specified three independent covariation patterns of histone modifications. The second assay vector specified a tradeoff between mono- and trimethylation of H3K4, characterizing differential enhancers and promoters, respectively. We thus used the sign of the mean projection across patients onto the second assay vector to classify the regulatory function of the identified regions as enhancers or promoters (STAR Methods). This classification was supported by the roughly bimodal distribution of the mean projections at the differential regions (Figure 3A) and the association of H3K4me1 versus H3K4me3 changes with our classification (Figure 3B). The putative promoter regions were preferentially found near gene transcription start sites (TSSs), while the putative enhancer regions were more distal (Figures 3C and 3D). The majority (70%) of regions with differential histone modification were enhancers. Roughly the same fraction of regions that increased and decreased in leiomyoma were enhancers (69% and 71%, respectively); however, the enhancers with increased activity in leiomyoma tended to be more distally located than those with decreased activity in leiomyoma (Mann-Whitney $U = 784{,}770$, n1 = 1,261, n2 = 930, p value = 7.3×10^{-42}; median nearest TSS distances of 43 kb and 13 kb, respectively).

To characterize the length scale of the epigenomic alterations, we performed a multiresolution analysis of the differential location vector. The fourth location vector was first split into positive and negative components representing the decreased and increased histone modification changes in leiomyoma, respectively. A discrete wavelet transform (DWT) of each vector, binarized using the threshold for identifying significant alterations, yielded the magnitude of epigenetic fluctuations at a given length scale from 4 kb to 2,048 kb (STAR Methods). While we found examples of both broad increases and decreases in histone modifications in leiomyoma, there was a greater occurrence of large-scale (~1 Mb) increases in activating histone marks, suggesting a long-range spatial correlation of chromatin alterations conducive to coordinated gene activation (Figure 4A). We further performed the same DWT analysis on the histone modification ChIP-seq signal at the identified differential regions for each condition, patient, and assay (STAR Methods). For each condition and assay, the genome-wide wavelet coefficients at a fixed length scale were averaged across patients, and the resulting values were compared between conditions. The comparison for H3K27ac again showed an increase in the large-scale (~1 Mb) coefficients for leiomyomas, demonstrating preponderant acquisition of long-range blocks of H3K27ac modification in leiomyoma compared to myometrium (Figure 4A). No clear trend was seen for H3K4me3 and H3K4me1.

Characterizing the regions with epigenetic alterations specific to leiomyomas with and without *MED12* mutations (STAR Methods) showed that these alterations also fell into two classes representing changes in promoters and enhancers (Figure S6A). Alterations in enhancers comprised most of the differences between leiomyomas with and without *MED12* mutations, accounting for 70% of the identified differences. Enhancers comprised a similar percentage of the regions with higher activating histone marks in *MED12*-mut versus *MED12*-WT leiomyomas (71% and 69%, respectively). There was no significant difference in the spatial distribution of these elements around the nearest gene TSS (Mann-Whitney U = 627,845.5, n1 = 1,173, n2 = 1,083, p value = 0.64; median nearest TSS distances of 30 kb for enhancers higher in *MED12*-mut and *MED12*-WT leiomyomas, respectively) (Figures S6B and S6C). Quantification of the length scale of the mutation-specific alterations using DWT of the seventh location vector showed preferential large-scale changes in *MED12*-mut leiomyomas (Figure S6D; STAR Methods).

Chromatin contact domains confine epigenomic alterations during tumor evolution

Since the human genome is thought to be partitioned into distinct domains of contact interactions (Dixon et al., 2012; Lieberman-Aiden et al., 2009; Rao et al., 2014), we tested whether the blocks of long-range epigenomic alterations agreed with this domain structure. Hi-C data from nine different cell types revealed several histone modifications (including H3K4me3 and H3K4me1) to be much more correlated for loci within these contact domains than for loci located at a similar distance but separated by a boundary, suggesting the contact domains may compartmentalize the epigenome (Rao et al., 2014). Comparing the epigenomic alterations in leiomyoma with the annotated contact domains in HeLa cells (Rao et al., 2014) showed that transitions in the differential HOSVD location vector often occurred at domain boundaries (Figure 4B; STAR Methods). To quantify this phenomenon, for each contact domain, we extracted the differential vector signal at loci within the domain

and at loci flanking the domain within the domain length in each direction; we then binned each of the three segments into five bins. Computing the pairwise correlation of differential vector signals at these bins across all contact domains showed a pronounced within-domain coupling of epigenomic alterations and minimal leakage across domain boundaries (Figure 4C). This trend was lost when either the positions of the contact domains (Figure 4D) or the differential vector components (Figure 4E; STAR Methods) were randomized. The *MED12*-mutation-specific epigenetic alterations were also predominantly confined within individual contact domains (Figures S6E–S6G). These results thus demonstrated that contact domains may confine epigenomic alterations during tumor evolution.

To identify contact domains with significant changes in chromatin state, we summed the significant differential vector signal across contact domains (STAR Methods) and observed a rapid decline and increase in the sorted net change, corresponding to 33 and 19 most altered contact domains with broadly increased and decreased activating histone marks in leiomyoma, respectively (Figure 4F; STAR Methods). The majority of these domains contained at least one differentially expressed gene changing in the expected direction (23/33 and 17/19).

Epigenomic alterations elevate collagenous ECM production in leiomyoma and dysregulate homeotic genes

To study the functions of target genes dysregulated by the identified epigenomic alterations, we analyzed matched RNA-seq data from 20 of the patients used in the HOSVD (Table S2; STAR Methods). Putative target genes were obtained using a distance criterion of 10 kb from the altered regions (Table S3; STAR Methods). To reduce false positives, the resulting list was limited to the genes that were also differentially expressed in the corresponding direction (Table S3; STAR Methods). The genes both elevated in leiomyoma and nearby a region with higher activating histone modifications in leiomyoma were significantly enriched for Gene Ontology (GO) terms (Huang et al., 2009a, 2009b) related to glycoproteins, collagen, and ECM (Table S3; STAR Methods), in line with the fact that excessive deposition of ECM composed primarily of collagen is a hallmark of uterine leiomyoma (Bulun, 2013; Moyo et al., 2020). Genes that were suppressed and nearby a region with lower activating histone modifications in leiomyoma were significantly enriched for terms related to homeobox genes, specifically the HOX genes (Table S3). The HOX genes are transcription factors (TFs) with diverse roles in development and cellular differentiation, and are dysregulated in several tumors (Cillo et al., 2001). Our data thus suggested that epigenetic alterations might dysregulate HOX genes in leiomyoma.

A similar approach identified the biological functions of target genes dysregulated by the mutation-specific alterations (STAR Methods). We performed differential expression analysis separately for leiomyomas with or without *MED12* mutations relative to corresponding myometrium (Table S2). Two gene lists were obtained for GO analysis (Huang et al., 2009a, 2009b): genes nearby regions with increased activating histone marks and elevated expression in *MED12*-mut leiomyomas, and genes near regions with increased activating histone marks and elevated expression in *MED12*-WT leiomyomas (Table S3). The genes elevated in *MED12*-mut leiomyomas were strongly enriched for terms related to

ECM, collagen, and focal adhesion (Table S3), known hallmarks of these tumors (Bulun, 2013; Moyo et al., 2020). The genes elevated in *MED12*-WT leiomyomas were not strongly enriched for any terms (Table S3), perhaps as a result of the heterogeneity of *MED12*-WT tumors and the smaller sample size.

Altered regions are enriched for TF binding motifs

To identify TFs associated with altered regulatory regions, we performed assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) in eight of the matched leiomyoma and myometrium samples (Table S1; STAR Methods). We intersected the full set of ATAC-seq peaks with the regulatory regions having increased or decreased histone marks in leiomyoma, yielding two refined sets of epigenetically altered open chromatin sites. Changes in ATAC-seq signal at these regions were consistent with the expected changes in accessibility based on histone modifications (Figures S7A and S7B). A discriminatory motif analysis using known human TF DNA binding motifs (Kulakovskiy et al., 2018) on the two sets of altered open chromatin regions identified differentially enriched TFs (STAR Methods). For the regions with increased activating histone marks in leiomyoma, serum response factor (SRF), HOXA9, and HOXA10 were some of the most enriched motifs (Table S4). Interestingly, SRF is known to confer phenotypic plasticity to vascular smooth muscle cells through competitive cofactor binding (Pipes et al., 2006; Wang et al., 2004), and changes in the post-translational modification of SRF have been found in FH-deficient uterine leiomyomas (Raimundo et al., 2009). The enrichment of HOXA9 and HOXA10 motifs was also notable, considering the epigenetic and transcriptomic dysregulation detected across the *HOXA* cluster (Figure 2A) and the essential role of these TFs in female reproductive tract development and function (Du and Taylor, 2015). The motifs enriched in the regions with reduced activating histone marks in leiomyoma included several ETS family TFs as well as glucocorticoid receptor (NR3C1) (Table S5). Consistent with this observation, NR3C1 (Yin et al., 2013) and the ETS family members ETS1, ETS2, ERG, FLI1, ELF1, and ELF3 showed reduced expression in leiomyoma.

To assess the function of these motifs, we built a logistic regression classifier with lasso regularization to predict whether an epigenetically altered region would have increased or decreased activating histone modifications in leiomyoma given the motif content of the region (STAR Methods). The classifier obtained 61% test and 67% training mean accuracy over 500 iterations of Monte Carlo CV with 20% of the data left out for testing at each iteration. As we were not considering differential accessibility but rather accessible sites (as measured by ATAC-seq peaks) with altered histone modifications, it is plausible that some accessible sites might have been unrelated to the observed alterations. While the prediction accuracy was moderate, it did show that the presence or absence of certain TF motifs was informative of the observed epigenetic alterations. We found SRF and HOX motifs to be the most informative predictors of increased histone modifications in leiomyoma (Table S4), while progesterone receptor (PGR) and ETS family sites were some of the most informative predictors of decreased histone modifications (Table S5).

HOXA13 is dysregulated and modifies the expression of genes related to leiomyoma pathogenesis

The epigenetic and transcriptional dysregulation at the *HOXA* cluster and the enrichment of posterior HOXA DNA binding motifs in epigenetically altered regions suggested a potential tumorigenic role of posterior *HOXA* genes in leiomyomas. *HOXA13* was the posterior *HOXA* gene exhibiting the greatest transcriptional dysregulation ($log_2FC = 2.61$ higher in leiomyoma) (Figure 2A). A recent study also implicated dysregulation of *HOXA13* and the *HOXA* cluster in uterine leiomyoma (George et al., 2019). The aberrant protein expression of HOXA13 in leiomyoma was confirmed by immunohistochemical staining (Figure 5A, left), with mean QH score of 120.9 compared to 31.3 in myometrial tissue (Figure 5A, right). To determine the biological significance of overexpressed *HOXA13*, we knocked down *HOXA13* using small hairpin RNA (shRNA) in primary leiomyoma cells. The shRNA targeting *HOXA13*, but not the control shRNA, decreased HOXA13 protein and mRNA levels similar to those of primary myometrial cells (Figures S7C and S7D). We performed RNA-seq on matched HOXA13 knockdown and control leiomyoma primary cells from two patients (pt20 and ptC; STAR Methods). HOXA13 knockdown affected the expression of several genes dysregulated in leiomyoma (Tables S2 and S6): genes with higher expression in fibroids and downregulated upon knockdown of HOXA13 included *COL6A3*, *THSD4*, and *ADAMTS2*, which either code for components of the ECM or the proteins that modulate its organization (Table S6); genes that were upregulated after HOXA13 knockdown and had lower expression in leiomyoma were mostly enriched for transcription regulation and response to hypoxia (Table S7).

To validate the RNA-seq results, we selected a subset of genes differentially expressed both in leiomyoma compared to myometrium and upon HOXA13 knockdown and performed qRT-PCR in primary leiomyoma cells (STAR Methods). MEDAG and PDK4 are involved in cell differentiation and glucose metabolism, respectively, showing highly reduced expression in leiomyoma; LIMK1 and SDC1 play important roles in cytoskeletal organization and are elevated in leiomyoma. We observed consistent mRNA changes upon HOXA13 knockdown, with *MEDAG* and *PDK4* showing a significant increase and *LIMK1* and *SDC1* showing a significant decrease (Figure 5B).

To further investigate the role of HOXA13, we overexpressed *HOXA13* and performed RNA-seq in primary myometrium cells from three patients (Figure S7E; STAR Methods). Increased protein levels of HOXA13 in myometrial cells led to dysregulation of 1,092 genes (Figure 5C; Table S6; STAR Methods) enriched for GO terms related to ECM (Figure 5D; Table S6). These results corroborated the findings from the knockdown analysis and further indicated a function of HOXA13 in leiomyoma via regulation of ECM-related genes. Furthermore, genes involved in the transforming growth factor β (TGFβ) signaling pathway (*ID1*, *BMP2*, and *SMAD7*) were upregulated in leiomyoma primary cells when *HOXA13* was knocked down and repressed when *HOXA13* was overexpressed in myometrial primary cells (Tables S2 and S6). These genes were also downregulated in leiomyoma tissues. Aberrant concentration of TGFβ family proteins or expression of genes involved in the TGFβ pathway has been observed in leiomyomas (Ciebiera et al., 2017; McWilliams and Chennathukuzhi, 2017). The consistent alterations in the expression of *ID1, BMP2*, and

SMAD7 in the *HOXA13* knockdown and overexpression analyses suggested HOXA13 functions to repress TGFβ pathway genes in leiomyoma.

DeCET identifies epigenomic signatures discriminating distinct tissues types, differentiation states, and disease subtypes

We further demonstrated DeCET's power and generalizability on Roadmap Epigenomics Mapping Consortium (REMC) histone modification data from 34 adult tissues representing a range of phenotypically and functionally distinct subgroups (Kundaje et al., 2015). Hierarchical clustering using the projections onto the first 13 DeCET location vectors grouped biologically related samples together even when related samples were from different labs (Figure 6A). Moreover, the location vectors encoded interpretable biological features associated with each cluster, and the pattern of projections across tissues clearly reflected the corresponding tissue-specific functional annotation of the vectors. For example, the regions with the most positive components of the fourth vector were enriched for genes related to lymphocyte activation and immune response, while those with the most negative components of the seventh vector were enriched for glial cell differentiation and myelination (Table S8; STAR Methods). Similarly, the regions with the most negative components of the twelfth vector were enriched for striated muscle cell development genes (Table S8).

As a more targeted application, we applied DeCET to the 10 REMC adult muscle tissue samples. Clustering using the projections onto the first 10 location vectors stratified the samples by muscle type, with smooth muscle clustering separately from striated muscle, and striated muscle being further stratified into cardiac and skeletal subtypes (Figure 6B). Notably, aorta (E065) clustered with smooth and not cardiac muscle, consistent with its smooth muscle composition. The projections onto the sixth location vector differed between the smooth and striated muscles samples, while those onto the seventh vector distinguished skeletal and cardiac muscles. In line with these discriminatory roles, genomic regions with the most positive components of the sixth vector were enriched for blood vessel and muscle structure development genes, while regions with the most positive and negative components of the seventh vector were enriched for skeletal muscle and heart development, respectively (Table S8).

We next tested DeCET's ability to extract epigenomic signatures of immune cell activation and differentiation from eight REMC primary T cell samples (Figure 6C). This dataset represented three differentiation states, with three samples selected for naive markers, three for memory markers, and two activated by phorbol myristate acetate (PMA)-ionomycin (PMA-I) treatment. Hierarchical clustering grouped the samples by differentiation state. The PMA-I-stimulated T cells had the most negative projections of active histone modifications onto the seventh location vector, the most negative components of which encoded regions enriched for lymphocyte activation and immune response genes (Table S8). Similarly, naive cells had the most positive projections onto the eighth location vector, the most positive values of which signified regions enriched for genes related to lymphocyte differentiation (Table S8) and near genes downregulated during T cell differentiation from naive to effector/memory states, including LEF1 and TCF7 (Danilo et al., 2018; Willinger et al., 2006).

We further applied DeCET to data from 11 breast cancer and 2 immortalized mammary tissue cell lines (Xi et al., 2018). For breast cancer samples, two replicates showed high reproducibility in the DeCET projections for each assay, and we used a tensor-based consensus profile for each pair (STAR Methods). Hierarchical clustering of the consensus profiles stratified the samples into luminal and basal subtypes, with further sub-stratification capturing estrogen receptor (ER) status in the luminal subtype and claudin-low and immortalized subgroups in the basal subtype (Figure 7A). The clustering reflected the corresponding expression pattern of known marker genes for the subtypes (Dai et al., 2017) (Figure 7B). The third DeCET location vector distinguished the luminal from basal subtypes, with the most positive components (corresponding to higher histone modifications in the basal subtypes) encoding regions enriched for genes related to ECM organization, positive regulation of locomotion, and negative regulation of cell death (Table S8). By contrast, the regions with the most negative components (corresponding to higher modifications in the luminal subtypes) were enriched for genes related to mammary gland development and epithelial cell differentiation (Table S8). These annotations were consistent with the more aggressive and invasive phenotype of basal breast cancers compared to the more differentiated luminal breast cancers (Dai et al., 2017). The fifth location vector further separated the luminal cell lines by ER status; the regions with the most positive components of this vector (corresponding to higher modifications in the ER- $HER2^+$ luminal subtype) were enriched for placenta development and vasculo-genesis genes (Table S8), while those with the most negative components (corresponding to higher modifications in the ER^+ luminal subtype) were enriched for negative regulation of cell cycle and mammary gland alveolus development (Table S8). These annotations were consistent with the more aggressive nature of HER2-enriched cell lines between ER^+ luminal and basal subtypes (Dai et al., 2017).

Interestingly, MCF7, a commonly used model of luminal A subtype supposed to express ER and progesterone receptor (PR) (Dai et al., 2017; Xi et al., 2018), clustered with the claudin-low and immortalized subtypes. This MCF7 cell line also exhibited a more basal-like expression profile with lost expression of many luminal markers and gained expression of several basal markers (Figure 7B). In particular, it had low expression of PR and no detectable histone marks near the *PGR* promoter, contrary to the other PR^+ luminal A and B cell lines. *PGR* loss in MCF7 cell lines through copy number deletion has been previously reported and may lead to the acquisition of a more aggressive phenotype (Mohammed et al., 2015).

We also applied DeCET to identify prostate cancer subtypes in a patient cohort with H3K27ac, H3K27me3, and androgen receptor (AR) ChIP-seq data (Stelloo et al., 2018). The samples were taken from primary prostate cancer tissues and labeled according to the status of a biochemical recurrence (case) within 5 years of diagnosis or no biochemical recurrence (control) within 10 years of diagnosis. Hierarchical clustering revealed three clusters of patients (Figure 7C; STAR Methods). Consistent with the original finding (Stelloo et al., 2018), we did not observe preponderant epigenetic differences between case and control groups. Two of the clusters were characterized by low and high ERG expression (STAR Methods), consistent with the previous report (Stelloo et al., 2018). The third cluster (low-metabolically active (low-MA]) showed widespread downregulation of genes related to

mitochondria and metabolism. A dormant-like, metabolically inactive state of prostate cancer cells may arise following induced epithelial-to-mesenchymal transition (Stylianou et al., 2019). In our analysis, this subgroup showed a trend of increased biochemical recurrence (one-sided Fisher's exact test odds ratio [OR] = 3.9, p value = 0.05). DeCET can thus successfully analyze diseases showing high patient variability in mutational landscape within and between subtypes.

DeCET outperforms peak-based clustering in stratifying distinct tissues types, differentiation states, and disease subtypes

To demonstrate the relative advantages of DeCET, we compared the above clustering results to those based on the Jaccard index of peak calls (STAR Methods). For comparing the distinct tissue groups from REMC, Jaccard index clustering was sufficient to stratify tissues into functionally similar types, although the clustering was not as clean as the DeCET result (Figure S8A; cf. Figure 6A). The Jaccard index approach, however, failed to stratify more functionally related tissue subtypes such as distinct muscle types (Figure S8B; cf. Figure 6B) and different immune cell subsets (Figure S8C; cf. Figure 6C). It also could not stratify breast cancer cell lines into relevant subgroups (Figure S8D; cf. Figures 7A and 7B) or leiomyoma and myometrium by disease condition and subtype (Figure S8E; cf. Figure 1C).

In addition to the improved clustering, the DeCET projections provided a low-dimensional representation of the heterogenous epigenetic data that could be used for biological interpretation, supervised learning, and classification tasks. By contrast, the Jaccard index approach only provided a notion of similarity between samples and required a separate meta-analysis to build a classifier or to identify differential regions. Plotting the difference in the number of myometrium and leiomyoma samples with a peak across the *HOXA* cluster (Figure S8F) demonstrated some difficulties in performing a meta-analysis of peak calls to identify differential regions. While the difference in the number of samples with a called peak qualitatively reflected the differences in histone modifications (Figure 2A), there was significant variability in the peaks called for the different samples and histone modifications within each biological condition. In addition, comparing peak calls could not directly account for modulations, which might not affect peak calling, but nevertheless reflect a change in mRNA regulatory activity. This phenomenon was observed downstream of *HOXA13* (Figure S8F).

DISCUSSION

DeCET overcomes drawbacks of current analysis methods, such as arbitrary choices of meta-analysis parameters and inter-sample variability in peak calls. As the epigenome carries essential instructions for specifying cellular identity, epigenomic assays may offer a robust diagnostic marker for a wide range of diseases. For example, the pattern of DNA methylation can classify primary and secondary central nervous system tumors, as well as identify the cell type of cancer origin (Capper et al., 2018; Moran et al., 2016; Orozco et al., 2018). Histone modifications provide complementary information about regulatory function and chromatin state (ENCODE Project Consortium, 2012; Kundaje et al., 2015). The challenges presented by heterogeneous datasets and the additional resources required for

multiple profiles, however, have made histone modifications less attractive for diagnostic applications to date. By contrast, we have shown that after an initial tensor decomposition of heterogeneous data even from a small patient cohort, new patient samples can be accurately classified by measuring only the most predictive histone modification and compressing the data along the precomputed HOSVD location features. The DeCET framework, combined with technologies requiring few input cells, such as cleavage under targets and release using nuclease (CUT&RUN) (Skene and Henikoff, 2017) or chromatin integration labeling followed by sequencing (ChIL-seq) (Harada et al., 2019), could offer an efficient diagnostic or prognostic tool for diseases, including cancers.

DeCET has identified epigenomic features capable of accurately distinguishing tissue types and disease conditions. Alterations in metabolic activity may have profound influence on chromatin structure in cancers (Wallace, 2012). Our analysis of prostate cancer epigenomes provides evidence for a metabolically inactive state with distinct epigenomic signatures and AR binding. These results together demonstrate the broad utility and power of DeCET in identifying clinically relevant disease subtypes and consolidating information from complex hierarchical datasets.

Interestingly, the majority of epigenomic alterations in leiomyoma occur in distal enhancers, with several alterations being large scale (0.1–1Mb). This length scale reminds of super-enhancers that may control cell identity and are highly sensitive to changes in TF concentration and bromodomain inhibitors (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). Calling super-enhancers in our H3K27ac data (Lovén et al., 2013; Whyte et al., 2013) shows that the regions with increased activating histone marks in leiomyoma are enriched for putative super-enhancers and that putative super-enhancers are more likely to be altered than typical enhancers (permutation test p values $< 10^{-3}$) (STAR Methods). We have also shown that epigenomic alterations in leiomyomas are largely confined within contact domains (Figure 4C), suggesting that the aberrant changes predominantly result from activation or silencing of chromatin compartments already established in the uterus.

Our analysis has identified HOXA13 as a potential tumorigenic factor (Gu et al., 2009; Li et al., 2015; Quagliata et al., 2018) affecting ECM-related genes in leiomyoma. Previous studies have shown that the nearby long non-coding RNA HOTTIP may drive the expression of 5′ HOXA genes through chromatin domain organization and recruiting the WDR5/mixed-lineage leukemia (MLL) complex to establish an active chromatin state (Luo et al., 2019; Wang et al., 2011). Our study shows that leiomyomas exhibit recurrent epigenetic alterations consistent with this activation mechanism (Figure 2A) (H3K4me3 not shown), with a ~1.8-fold increase in HOTTIP expression.

In summary, we have presented a powerful computational framework for integrating heterogeneous epigenomic data and demonstrated how this method can facilitate the discovery of rich biological knowledge. In particular, we envision that our approach can be extended to cells isolated from bodily fluids or biopsies to identify disease states and subtypes, thereby improving diagnosis and treatment.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Debabrata Chakravarti (debu@northwestern.edu).

Materials availability—The plasmid created in this study will be submitted to Addgene once manuscript is published.

Data and code availability—The datasets generated in this study are available at the Gene Expression Omnibus (GEO) database (GEO: GSE142332). The code used for this study can be found at https://github.com/jssong-lab/DeCET (https://doi.org/10.5281/zenodo.4540815).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human samples—Human tissues were collected upon approval by Institutional Review Board of Northwestern University (Study number STU00018080, renewal approved on Dec17, 2018). Premenopausal women (age ranging from 41–52 years with average age of 48 years) undergoing hysterectomy or myomectomy at Northwestern University Prentice Women's Hospital were consented and leiomyoma and/or matched myometrium tissues were obtained from them (Table S1).

Primary cells—Leiomyoma and myometrial tissues obtained from the patients were digested and primary cells were plated and grown at 37°C in a humidified cell culture incubator containing 5% CO_2.These cells were cultured in Smooth Muscle Cell complete growth media SmGM™- 2 (Lonza, CC-3182).

Cell lines—HEK293T/17 cells (ATCC, CRL-11268) were used to produce virus particles containing shRNA sequence. The HEK293T/17 cells were grown in DMEM media (ThermoFisher, 11965092) with 10% FBS (Fisher Scientific, 10437028) at 37°C in a humidified cell culture incubator containing 5% CO_2.

METHOD DETAILS

Tissue collection—Matched tissues from 21 patients and leiomyoma from 5 patients with diverse ethnicity (comprising 8 African-Americans, 10 Caucasian, 1 Hispanic/Latino and 8 with unknown status) were dissected, snap-frozen in liquid nitrogen and immediately stored at −80°C in small aliquots (about 1g) until further processing for RNA and ChIP-sequencing. 7 of these 26 patients were on hormonal treatment before surgery. RNA from leiomyoma of one patient (pt16) was of poor quality and was not included in the RNA-seq study (20 matched, 5 leiomyoma-only tissues). ATAC-sequencing was performed on matched tissues from eight patients (Table S1).

Tissue preparation for ChIP—About 1g of frozen tissues for each case were pulverized in a Covaris CP02 cryoPREP dry pulverizer (setting 5, 4 strikes) making sure to rechill the

tissue in liquid nitrogen between strikes. Further, tissue was finely powdered with mortar-pestle in liquid nitrogen, followed by fixation with 1% paraformaldehyde for 15 minutes at room temperature. To stop cross-linking 1X glycine (Cell Signaling Technology, 9003) was added for 5 minutes. Cross-linked tissues were either stored at −80°C or used immediately for ChIP.

Tissue digestion and primary cell culture—Tissues were rinsed twice with HBSS and then cut into ~2mm^3 pieces in a sterile cell culture plate, discarding tissue with blood traces. Afterward, tissue pieces were transferred to 50ml tubes and freshly prepared digestion buffer [1.5mg/ml Collagenase type I (Sigma-Aldrich, C0130), 3mM $CaCl_2$, 20ug/ml DNase I (Sigma-Aldrich, D5025) in Hank's Balanced Salt Solution (ThermoFisher, 14025)] was added at a ratio of 1 part (w/v) of tissue to 5 parts of buffer. Tissue was digested in 37°C shaker at 100rpm for 4–5hrs. Digested tissue was filtered through 12-ply, 4×4 in. sterile gauze sponges, spun at 0.4 × g for 5 minutes to harvest primary cells, followed by resuspension of cells in PBS and filtration through 70μm cell strainer. Cells were centrifuged at 0.4 × g for 5 minutes and subsequently resuspended in Smooth Muscle Cell complete growth media SmGM™- 2 (Lonza, CC-3182). Primary cells were plated and grown at 37°C in a humidified cell culture incubator containing 5% CO_2. Cells were used in experiment after two passages as leiomyoma cells become senescent if cultured longer. Leiomyoma tissues were processed from four patients (pt20, ptA, ptB and ptC) for silencing *HOXA13*. After *HOXA13* knockdown, primary cells from pt20 and ptC were used for RNA-seq, while those from ptA, ptB and ptC were used for qRT-PCR validation of target genes. Primary myometrial cells were processed from three patients for overexpression of *HOXA13*.

ChIP-seq—The cross-linked tissues were processed for ChIP using the SimpleChIP Kit (Cell Signaling Technology, 9003). Briefly, tissue pellet was lysed and homogenized in 10 mL of 1X Buffer A by 15–20 strokes in Dounce homogenizer, then rotated for 20 minutes at 4°C for end-to-end mixing. For chromatin fragmentation, lysed tissue was resuspended in 4 mL 1X Buffer B, transferred to 1.5ml Eppendorf tubes and incubated with 2000–4000 gel units of MNase/ml. Tubes were kept in thermomixer at 37°C for about 20 minutes or until approximately 75% mono-nucleosomal profile of purified digested DNA was observable on agarose gel. Digested chromatin was retrieved by 2–3 sonication pulses of 15 s on and 45 s off (Misonix, setting 5) in 500μl of 1X ChIP buffer per tube followed by centrifugation at 16000 × g for 20 minutes to remove cell debris and ECM proteins. For ChIP, 5μg (DNA) of the solubilized chromatin was incubated overnight at 4°C while rotating with 5μg antibodies for the histone marks. Antibodies used for ChIP: anti-H3K27ac (Active motif, 39133), anti-H3K4me3 (Diagenode, C15410003), anti-H3K4me1 (Diagenode, C15410194). Afterward, protein G Dynabeads were added and incubation continued for another 2 hours. To obtain ChIP products, beads were washed and DNA eluted. Chromatin supernatant was reverse cross-linked overnight, and DNA was purified using PCR purification kit (QIAGEN, 28006). Library preparation of the purified DNA from ChIP was performed by following the Kapa hyper prep protocol (Kapa Biosystems, KK8502). 1–5ng of the DNA was end-repaired and A-tailed according to the kit instructions. Adapters were ligated at 20°C for 15 minutes with post-ligation cleanup and libraries were size selected for 250–450 bp fragments using 0.6X–0.8X ratio of AMPure beads (Beckman Coulter, A63881). The size selected libraries were

amplified and checked for size by Agilent Bioanalyzer for high sensitivity DNA (Agilent technologies, 5067). Libraries were diluted to 2nM, pooled and quantified with Kapa quantification kit (Kapa biosystems, KK4835) before single end sequencing with Nextseq 500/550 high output kit (v2, 75 cycles).

RNA isolation and next generation sequencing—RNA was isolated from approximately 30mg of the frozen tissue using RNeasy Fibrous tissue kit (QIAGEN, 74704). Frozen tissue was finely grounded in mortar-pestle in liquid nitrogen. Tissue lysate prepared by addition of Buffer RLT (along with 10μl β-mercaptoethanol/ml of RLT buffer) was homogenized by spinning down through QIAshredder (QIAGEN, 79656). Following the manufacturer's instructions, RNA was isolated from the homogenized tissue lysates. RNA obtained was analyzed on Bioanalyzer eukaryote total RNA 6000 nano assay (Agilent technologies, 5067) for RNA Integrity. Samples with RNA Integrity Number (RIN) values 8 or above were processed further for next generation sequencing. TruSeq stranded mRNA kit (Illumina, 20020594) was used for the preparation of libraries for RNA-seq. From 1ug of the total RNA, mRNA was purified and fragmented using the reagents provided with the kit. Blunt ends of cDNA were adenylated at 3′ end, adapters ligated, and library amplified. 2nM libraries were pooled, quantified and sequenced as paired-end with 42 cycles using Nextseq 500/550 high output kit (v2, 75 cycles).

ATAC-seq—Tissues were pulverized in a Covaris CP02 cryoPREP dry pulverizer. About 30mg of powdered frozen tissue was transferred to 1.5ml tubes and resuspended in ice cold Nuclei Lysis Buffer (NIB) (20mM Tris-HCl, 50mM EDTA, 5mM spermidine, 0.15mM spermine, 0.1% mercaptoethanol, 40% glycerol, pH 7.5). Samples were agitated for 5 minutes on ice, filtered through Miracloth (EMD Millipore Corp), and then centrifuged at 1,100 × g for 10 minutes at 4°C. The resulting nuclear pellet was then subjected to transposition as per (Buenrostro et al., 2015), with no modifications. Briefly, the pellet was resuspended in the transposase reaction mix (25 μL 2x TD buffer (Illumina, 15027865), 2.5 μL Transposase (Illumina, 15027865) and 22.5 μL of nuclease free water). The transposition reaction was performed at 37°C for 30 minutes. Following transposition, the samples were purified using a QIAGEN MinElute kit (QIAGEN, 28006). After transposition, the library fragments were amplified for 10–12 cycles. Traces were then analyzed on Bioanalyzer and 42 cycles of paired-end sequencing were performed on an Illumina Nextseq 500/550.

Mutation analysis—1ug of the RNA isolated from normal myometrium and leiomyoma tissues was reverse transcribed with qScript™ cDNA super mix (QuantaBio, VWR, 101414–102) as per the manual. To check mutation status, sequences from RNA-seq were visualized with Integrated genome viewer (Robinson et al., 2011) and noted for the *MED12* exon 2 mutations, *HMGA2* overexpression, and loss of *FH*. For further validation, *MED12* primers were used to amplify exon 2 and sequenced by Sanger sequencing. Sequences were analyzed manually using BioEdit software (Hall, 1999). For *HMGA2* overexpression analysis, qRT-PCR was performed using primers mentioned below and chromatin prepared as above was used for HMGA2 (Genetex, GTX629478) protein detection by Western Blot. Anti-Histone 3 (Abcam, ab1791) antibody was used as a loading control.

Genomic DNA isolation—For isolation of genomic DNA from tissues, DNeasy blood and tissue kit (QIAGEN, 69504) was used. About 25mg of frozen tissue was finely grounded in mortar-pestle in liquid nitrogen, followed by addition of buffer ATL and proteinase K (to a final concentration of 2mg/ml). Powdered tissue was incubated at 56°C for about 20–30 minutes so the tissue is completely lysed. Buffer AL and 33% ethanol were added to the samples with thorough mixing by vortexing and DNA was purified using spin columns provided with the kit. For validation of *MED12* mutation, *MED12* primers for genomic DNA were used to amplify exon 2 and sequenced by Sanger sequencing. Sequences were analyzed using Indigo (https://www.gear-genomics.com/).

shRNA-mediated knockdown in primary cells—HEK293T/17 cells (ATCC, CRL-11268) were used to produce virus particles containing shRNA sequence. The HEK293T/17 cells were grown in DMEM media (Thermo Fisher, 11965092) with 10% FBS (Fisher Scientific, 10437028) were transfected with lentiviral pLKO.1 plasmid construct with shRNA against human *HOXA13* (Sigma, TRCN0000015406) along with pMD2.G and psPAX2 plasmids. Transfection was performed using Lipofectamine 2000 reagent (ThermoFisher, 11668019) according to manufacturer's instructions. Media was changed after overnight incubation. Virus particles were collected 24 hours later by spinning down the supernatant and stored at −80°C in small aliquots when not used immediately. 2×10^5 primary leiomyoma cells (Passage 2 or 3) were transduced in a 6-well plate with either control or *HOXA13* gene silencing lentiviral particles in the presence of 6ug/ml polybrene for 18 hours. Fresh SmGM complete media was added and then after 24 hours, cells were selected in 2ug/ml puromycin for 3 days. Cells were then harvested for RNA using RNeasy mini kit (QIAGEN, 74104) and for protein with modified RIPA lysis buffer (20 mM Tris-HCl [pH 7.6], 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% IGEPAL CA-630, 1% sodium deoxycholate, 0.25% SDS). Whole-cell extracts prepared using modified radioimmunoprecipitation assay (RIPA) buffer were processed as described previously (Parker et al., 2012). Extracts were clarified by centrifugation at 20,000 × g for 15 min at 4°C, and protein concentrations determined by bicinchoninic acid (BCA) assay (Thermo Fisher Scientific, 23225). About 30μg of protein lysate was loaded on precast 8 to 16% polyacrylamide gels (Thermo Fisher Scientific, XP08162BOX). Following transfer to nitrocellulose membrane, western blot was performed to detect HOXA13 (Abcam, ab106503). Anti- GAPDH (Sigma-Aldrich, G9545) antibody was used as a loading control. About 100ng of RNA was used for library preparation using TruSeq stranded mRNA kit (Illumina, 20020594) and paired-end sequenced on Illumina Nextseq 500/550 as described above.

In Figure S7C and S7D, protein and RNA were extracted from the primary leiomyoma cells 5 days after being transduced with lentivirus containing negative control (shControl) or HOXA13 (shHOXA13) shRNA construct.

HOXA13 overexpression in primary cells—*HOXA13* cDNA was amplified from pLV expression plasmid (Vector builder, VB180306–1076naw) with Platinum Superfi DNA polymerase (Thermo Fisher Scientific, 12359010) using primers with attB sequence. Amplified product was cloned into pLEX_306 plasmid (Addgene, 41391) containing V5-

epitope tag at 3′end of insertion site using a Gateway reaction (Thermo Fisher Scientific, 11-789-020 and 11-791-020). Empty and HOXA13 inserted pLEX_306 plasmids were packaged into lentiviral particles following transfection in HEK293T cells as described above for shRNA constructs. Sanger sequencing was performed to ascertain integrity of the insert. Similar to knockdown studies, primary myometrial cells were transduced with lentivirus and selected in 2ug/ml puromycin for five days. Cells were then harvested for RNA using RNeasy mini kit (QIAGEN, 74104) and for protein with modified RIPA lysis buffer (20 mM Tris-HCl [pH 7.6], 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% IGEPAL CA-630, 1% sodium deoxycholate, 0.25% SDS). To confirm overexpression of V5-tagged HOXA13, western blot was performed using anti- HOXA13 (Abcam, ab106503), or anti V5-epitope specific antibodies (Thermo Fisher Scientific, R96025). Anti- GAPDH (Sigma-Aldrich, G9545) antibody was used to detect a loading control protein (GAPDH) (Figure S7E). About 100ng of RNA was used for library preparation using TruSeq stranded mRNA kit (Illumina, 20020594) and paired-end sequenced on Illumina Nextseq 500/550 as described above.

Immunohistochemistry (IHC) staining of HOXA13—Tissue microarray (TMA) was prepared from randomly selected leiomyoma (57 tumors) and matched myometrium (57 cases) and additional myometrium without leiomyoma (13 cases). TMA block was sectioned in 4μm and IHC staining was performed at Pathology core facility, Northwestern University. Briefly, tissue sections were de-paraffinized, rehydrated and processed for antigen retrieval at pH 9.0. Afterward, sections were subjected to immunohistochemical staining at Bond™ Polymer Refine detection system (Leica, DS9800). Primary antibody against HOXA13 (Abcam, ab106503) at a dilution of 1:400 was used for the staining. Sections were mounted with aqueous media and imaged with Nanozoomer 2.0 HT (Olympus). Immunoreactivity of the cores was graded semiquantitatively with visual inspection based on intensity of 0 (negative), 0.5 (faint), 1 (weak), 2 (moderate) and 3 (strong). Quantitative H score (QH) was determined for nuclear HOXA13 as follows: $QH = Pi$ (where, i is intensity 0–4, P is the percentage of positive cells for each given i) (Wei et al., 2005), and p value was calculated using the two-tailed t test.

Oligonucleotides—*MED12* exon 2 cDNA sequencing primers:

Forward Primer: 5′- CTTCGGGATCTTGAGCTACG- 3′

Reverse Primer: 5′- GTTGGAACTGATCTTGGCAGG- 3′

MED12 exon 2 genomic DNA sequencing primers:

Forward Primer: 5′- GCC CTT TCA CCT TGT TCC TT- 3′

Reverse Primer: 5′- TGTCCCTATAAGTCTTCCCAACC- 3′

HOXA13 cDNA PCR primers:

attB1_HOXA13_Forward Primer:

5′-
GGGGACAAGTTTGTACAAAAAAGCAGGCTTCACCATGACAGCCTCCGTGCTCCTC
C- 3′

attB2_HOXA13_Reverse Primer:

5′- GGGGACCACTTTGTACAAGAAAGCTGGGTGACTAGTGGTTTTCAGTTTGTTG-
3′

qRT-PCR primers:

HMGA2:

Forward Primer: 5′- AGCAGCAGCAAGAACCAACC- 3′

Reverse Primer: 5′- CTTGGCCGTTTTTCTCCAGTG- 3′

HOXA13:

Forward Primer: 5′- ACTCTGCCCGACGTGGT- 3′

Reverse Primer: 5′- CCGCTCAGAGAGATTCGTCG- 3′

MEDAG:

Forward Primer: 5′- TCAAGAGGTATGTGGAACTGACC- 3′

Reverse Primer: 5′- TGACCATGTCCATCCCTTGC- 3′

PDK4:

Forward Primer: 5′- CAGACAGGAAACCCAAGCCA- 3′

Reverse Primer: 5′- TTGCCCGCATTGCATTCTTA- 3′

LIMK1:

Forward Primer: 5′- ATCAGGGATGGCCTACCTCC- 3′

Reverse Primer: 5′- CAGGCTGAGTCTTCTCGTCC- 3′

SDC1:

Forward Primer: 5′- GGAAGGGCCTGTGGGTTTA- 3′

Reverse Primer: 5′- CGCTCTCTACTGCCGGATTC- 3′

GAPDH:

Forward Primer: 5′- TGCACCACCAACTGCTTAGC- 3′

Reverse Primer: 5′- GGCATGGACTGTGGTCATGAG- 3′

shRNA sequences:

HOXA13 (TRCN0000015406):

5′-CCGGGTTCCAGAACAGGAGGGTTAACTCGAGTTAACCCTCCTGTTCTGGAACTTTTT-3′

Constructing the data tensor

The histone modification ChIP-seq dataset generated in this study fits naturally into an order four tensor (Figure 1A). The four indices of the tensor are the condition (healthy or tumor), patient ID, ChIP-seq assay type (H3K27ac, H3K4me3, H3K4me1) and genomic location. The values stored in the tensor correspond to a measure of the corresponding ChIP-seq signal:

$$\mathcal{X}^{i}{}_{condition}{}^{i}{}_{patient}{}^{i}{}_{assay}{}^{i}{}_{location} = (Strength\ of\ ChIP\text{-}seq\ signal\ for\ condition, patient\ and\ assay\ at\ genomic\ location).$$

For the location index, we binned the genome into non-overlapping 2000bp intervals. The width was chosen to ensure sufficient read coverage and reduce the effect of statistical fluctuations. To obtain the strength of the ChIP-seq signal in each bin, we first estimated the center of each read by shifting each aligned read by 100bp in the 3′ direction of the strand to which the read aligned. The strength of the ChIP-seq signal was then measured as the number of read centers assigned to each bin.

Data tensor processing and normalization

ChIP-seq assays are subjected to variability in library preparation, sequencing depth, immunoprecipitation (IP) enrichment, and antibody qualities; thus, normalizing the heterogenous data across experiments and patients is crucial for downstream analysis. We applied the following three data processing and normalization steps:

The first step in the data processing procedure accounted for potential biases in library preparation or other artifacts not corresponding to true IP enrichment signal. This step used a control file generated for each condition i for each patient j, $control_{i,j}$, vectorized in the same way as the signal datasets. For each of the three histone modification ChIP-seq datasets, $signal_{i,j,k}$ where $k \in \{H3K27ac, H3K4me3, H3K4me1\}$, we used the Signal Extraction Scaling method described in Diaz et al. (2012) to scale the corresponding control dataset to each signal dataset. The scaling factor, $\alpha_{i,j,k}$, was computed using the binned signal and control vectors. The control vector was then scaled using the obtained scaling factor and subtracted from the signal dataset with any resulting negative entries set to zero:

$$\mathcal{X}^{ijkl}_{s1} = \max\left(0, signal_{i,j,k}(l) - \left(\alpha_{i,j,k}\right) control_{i,j}(l)\right).$$

The second data processing step scaled the data to account for differences in sequencing depth or differences in IP enrichment between samples for a given ChIP-seq assay. To

account for actual differences in the strength of the signal across patients and condition, the median of ratios scaling method described in Anders and Huber (2010) was used to scale the genomic location vectors across samples for a given ChIP-seq assay. Specifically, for each location index l, the geometric mean across conditions and patients

$$g\left(\mathcal{X}_{s1}^{:,\ :,\ i_a=a,l}\right)=\sqrt[n_c n_{pt}]{\prod_{i_c=1}^{n_c}\prod_{i_{pt}=1}^{n_{pt}}\mathcal{X}_{s1}^{i_c,i_{pt},i_a=a,l}}$$

was computed for each assay, with n_c equal to the number of conditions and n_{pt} the number of patients. For each sample, the ratio of the sample data to the geometric mean was taken at all locations where the geometric mean was non-zero. The inverse scaling factor is taken to be the median of these ratios

$$s\left(\mathcal{X}_{s1}^{i_c=c,i_{pt}=j,i_a=a,:}\right)=s_{c,j,a}=\underset{l}{median}\left\{\left|\frac{\mathcal{X}_{s1}^{i_c=c,i_{pt}=j,i_a=a,l}}{g\left(\mathcal{X}_{s1}^{:,\ :,\ i_a=a,l}\right)}\right|g\left(\mathcal{X}_{s1}^{:,\ :,,\ i_a=a,l}\right)>0\right\}.$$

The scaled subtensor was then obtained by dividing each entry in $\mathcal{X}^{i_c=c,i_{pt}=j,i_a=a:}$ by $s_{c,j,a}$

$$\mathcal{X}_{s2}^{i_c=c,i_{pt}=j,i_a=a,l}=\frac{1}{s_{c,j,a}}\mathcal{X}_{s1}^{i_C=c,i_{pt}=j,i_a=a,l}.$$

The final data processing step was to normalize across the different ChIP-seq assays. This step was necessary because of variability in antibodies for different histone modifications, which could skew certain component values in the tensor. We thus scaled the subtensor obtained by fixing the assay index to have total value 1:

$$\mathcal{X}_{s3}^{i_c,i_{pt},i_a=a,l}=\left(\sum_{c=1,\ pt=1,l=1}^{n_c,n_{pt},L}\mathcal{X}_{s2}^{c,pt,i_a=a,l}\right)^{-1}\mathcal{X}_{s2}^{i_c,i_{pt},i_a=a,l}.$$

This final processed data tensor was used for all further analysis.

The DeCET method—We designed the DeCET (Decomposition and Classification of Epigenomic Tensors) method to identify differential epigenetic signals in heterogeneous histone modification ChIP-seq datasets and to classify disease conditions using the identified features. The power of a tensor method comes from fully leveraging the structure of a complex dataset. For the dataset generated in this study, it involved integrating information from multiple histone modifications from matched healthy and tumor samples across a set of patients into a single tensor. The DeCET method for identifying differential epigenetic features consisted of three steps that will be described in detail below:

- The data tensor was decomposed to obtain a set of vectors in genomic space and corresponding sets of projections onto these vectors for each sample.

- The projections were used to identify genomic location vectors along which groups of samples diverged.
- The components of a discriminatory genomic location vector were used to identify regions in the genome with differential epigenetic signal.

Support Tensor Machine (STM) and Support Vector Machine (SVM) classifications of tissue samples were performed using the projections obtained in the first step, as described in detail below.

The DeCET framework is flexible, and our method can easily be generalized to accommodate additional structure in the data, such as time points at which samples were taken, by simply adding additional indices to the data tensor. Similarly, the DeCET method can also be applied to datasets with less structure, such as when matched tissue samples are not available.

Decomposing the data tensor—Higher order singular value decomposition (HOSVD) (De Lathauwer et al., 2000) of the normalized data tensor was used to identify correlated features. The HOSVD representation of the data tensor took the following n-mode product form:

$$\mathcal{X} = \mathcal{S} \times_1 U^{(condition)} \times_2 U^{(patient)} \times_3 U^{(assay)} \times_4 U^{(location)},$$

where $\mathcal{S}$ is the core tensor, and $U^{(condition)}$; $U^{(patient)}$; $U^{(assay)}$; $U^{(location)}$ are orthogonal matrices with size equal to the dimension of the corresponding unfolded subspace (the number of conditions $n_c = 2$, the number of patients $n_{pt} = 21$, the number of assays $n_a = 3$, and the number of genomic location bins L).

The HOSVD result was used to express the vectorized data for a given condition i_c, patient i_{pt}, and assay i_a as a weighted sum over characteristic, orthonormal vectors in location space

$$\mathcal{X}^{i_c i_{pt} i_a :} = \sum_l \left(\sum_{\alpha_1 \alpha_2 \alpha_3} \mathcal{S}^{\alpha_1 \alpha_2 \alpha_3 l} U^{(condition)}_{i_c, \alpha_1} U^{(patient)}_{i_{pt}, \alpha_2} U^{(assay)}_{i_a, \alpha_3} \right) U^{(location)}_{:, l} = \sum_l \mathcal{B}^{i_c i_{pt} i_a l} U^{(location)}_{:, l}.$$

The number of location vectors needed to obtain a complete representation of the data was limited by the rank of the corresponding unfolded matrix, with the upper limit being the product of the dimensions of the other three indices ($2 \times 21 \times 3 = 126$). We used the projection weights $\mathcal{B}^{i_c i_{pt} i_a l}$ obtained from this representation to compare disease condition subtypes.

We developed our own implementation of the HOSVD on GPU. We used the TensorLy package (Kossaifi et al., 2019) to perform all n-mode products and matrix unfolding of tensors, with a PyTorch (A. Paszke et al., 2017, NIPS Autodiff Workshop, conference) backend for matrix operations and eigenvalue decomposition of symmetric matrices on GPU processors. Our implementation of the HOSVD takes advantage of the rank restrictions imposed by matrix SVD to avoid eigen decomposition of a large matrix. This is done by

observing that the informative location vectors (those with non-zero projections) can be obtained by first obtaining the right singular vectors of the genomic-location mode unfolding of the data tensor and then obtaining the corresponding left singular vectors through a matrix multiplication. For most genomic datasets, the implementation of the HOSVD used in DeCET will be very efficient as the product of the number of samples and assays will be much less than the dimension of the genomic location space. For the datasets used in this study, the HOSVD of the normalized tensor took only seconds, with the greatest computation time coming from binning and normalizing the ChIP-seq data.

Identifying epigenetic alterations—The difference in histone modification ChIP-seq signal between conditions for a given patient j and assay k can be expressed using the HOSVD results as

$$\sum_{l=1}^{126} \left(\mathcal{B}^{L,j,k,l} - \mathcal{B}^{M,j,k,l} \right) U^{(location)}_{:,l}.$$

For each assay, we used a one-way analysis of variance (ANOVA) of the projection weights to identify the vectors in the genomic location space decomposition along which the projections separated the two conditions and, hence, the mean difference between conditions was significantly non-zero. To prioritize epigenetic alterations showing a high effect size, we considered both the p value and the relative values of the ANOVA test statistics when identifying location vectors with significant between-group separation. For the comparison of leiomyoma and. myometrium samples, 1 and 40 were used for the numerator and denominator degrees of freedom in the ANOVA test. For the comparison of *MED12*-mut and *MED12*-wt leiomyomas, 1 and 19 were used for the numerator and denominator degrees of freedom. The location vector with the greatest ANOVA test statistic was selected, while the ANOVA p value was used to test the statistical significance of the separation under the assumptions of ANOVA.

In addition to the ANOVA analysis for each assay, we sought to quantify the overall separation between the tumor and healthy samples across all assays along a given location vector. To quantify the overall separation into a single test statistic, we first computed the within- and between-condition separation (the denominator and numerator, respectively, of the ANOVA test statistic) for each assay. These quantities were then summed over assays, and the ratio of the summed between-condition separation to summed within-condition separation was taken as the test statistic for each location vector. Because the projections onto a location vector for different assays were not independent, the test statistic was not F-distributed as in ANOVA. Instead of computing *p-value*s, we thus compared the relative values of the test statistic to identify the location vector with the greatest overall separation between the groups. We found the separation of the condition projection means for the fourth location vector to be far greater than all the others (Figures S3B–S3E). Therefore, considering the equation above describing the condition difference, this location vector specified the differences in histone modifications between the two conditions. As the condition projection means were fixed for this location vector, the sign of the difference in projection means and the sign of the location vector component specified the direction of the

alterations (either increasing or decreasing in the tumors relative to healthy tissue). The magnitude of the vector entries specified the effect size of the alteration. We used the empirical distribution of the absolute value of all location vector entries to set a threshold for a location vector entry to be considered significantly non-zero. We set the threshold at the 99.9^{th} percentile of this distribution (Figure S3F). Genomic regions with differential epigenetic states were identified as those for which the absolute value of the corresponding vector entry exceeded the threshold value.

The same procedure was used to identify regions different between leiomyomas with and without *MED12* mutations. Here, the mean projections for leiomyomas with and without *MED12* mutations were compared to identify the seventh location vector to be specifying the mutation-specific alterations.

Projection of additional samples—Given a histone modification ChIP-seq dataset for a new patient, projections onto the location vectors were obtained using the location vector matrix $U^{(location)}$ from the HOSVD. This was accomplished by first vectorizing the new dataset and subtracting a corresponding control, as was done for the samples used in the tensor. Denoting this new data vector by v_k where $k \in$ {H3K27ac, H3K4me3, H3K4me1}, the projection onto the I^{th} location vector was obtained by taking the dot product of v_k with the I^{th} column of $U^{(location)}$. The new vector of projections is then given by

$$projections_k = v_k^t U^{(location)}.$$

Tissue sample classification—We applied two different techniques for classifying the condition of tissue samples. Each technique has advantages in certain conditions or applications. Both techniques use supervised machine learning to train a classifier to predict a label for a given tissue sample based on the projections obtained by the HOSVD used in the DeCET method.

To enable direct cross-sample comparison, we applied an additional standardization step to each tissue sample. For each condition, patient, and assay, the vector consisting of the projections onto the first 10 location vectors from the HOSVD was scaled to unit ℓ_2-norm. This same standardization was applied to both the tissue samples used in the tensor for the HOSVD and the additional test samples that were projected onto the obtained location vectors. Each tissue sample (consisting of a fixed condition and patient index) was represented as a 3×10 matrix of scaled projections, with the row index corresponding to the assay and the column index to the HOSVD location vector.

Classification based on a single histone modification was performed using a support vector machine (SVM) classifier with a linear kernel. The SVM was implemented using SciKit-learn (Pedregosa et al., 2011) linearSVC with an ℓ_2-penalty regularization and a squared hinge loss function (options C = 1, loss = ‘squared_hinge’, penalty = ‘l2′, max_iter = 10000, tol = 1e-5, class_weight = ‘balanced’). For each histone modification, the SVM classifier was trained to classify tissue samples based on the scaled 10 dimensional vector of HOSVD projections for that assay. We also applied this method to classify samples based on the joint

profile of all three histone modifications. To do this, each tissue sample was encoded as a 30-dimensional vector containing all histone modifications by concatenating the three 10-dimensional vectors representing the three histone modifications.

We next built a support tensor machine (STM) classifier (Cai et al., 2006; Tao et al., 2007) to improve the classification based on the full set of histone modifications. The STM classifier utilizes the tensor form of the predictor variable (in this case the 3× 10 matrix of scaled projections) to obtain a robust classifier with a smaller set of parameters than the full SVM. As fewer parameters are used, the STM classifier is less prone to overfitting and hence ideal for small training sets. The STM classifier fits vectors $u \in \mathbb{R}^3$ and $v \in \mathbb{R}^{10}$ that form the rank-1 CP-decomposition of the weight matrix $W = uv^T$. Optimal parameters of the STM are found by minimizing the following loss function

$$Loss\ Function = \frac{1}{2}\lambda u^2 v^2 + \sum_n max\left(0, 1 - y^{(n)}\left[\sum_{ij} x_{ij}^{(n)} u_i v_j + b\right]\right),$$

where for each of the tissue samples indexed by n, $x^{(n)}$ denotes the feature tensor and $y^{(n)} \in \{-1, 1\}$ is the indicator distinguishing the condition (either leiomyoma/myometrium, or *MED12* mutation status), b is the bias constant, and $\lambda = 0{:}0005$ is the regularization constant. We minimized the Loss Function by optimizing each of the u, v vectors in turn while holding the other vector fixed. We have run 400 iterations of successive u, v optimization, with each such optimization done using gradient descent method with 250 iterations and learning rate $\alpha = 0{:}0001$.

QUANTIFICATION AND STATISTICAL ANALYSIS

ChIP-seq data processing—Illumina adaptor sequences were removed from sequenced reads using Trim Galore (options–illumina -stringency 13) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Quality control was performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After adaptor trimming, reads were aligned to the hg19 genome using Bowtie2 (options -end-to-end–sensitive–score_min L,-1.5,-0.3) (Langmead and Salzberg, 2012). Aligned reads were sorted using Picard (https://broadinstitute.github.io/picard), and supplementary and low quality alignments were removed using SAMtools view (options -F 3588 -q 13) (Li et al., 2009). Duplicate reads were removed using Picard MarkDuplicates. Coverage files were generated from bam files using bedtools genomecov (options -bg -split -fs 200) (Quinlan and Hall, 2010) and then converted to bigWig format using kentUtils bedGraphToBigWig (Kent et al., 2010). The bigWig files were used for visualization in the UCSC genome browser (Kent et al., 2002; Raney et al., 2014). For visualization, each track was scaled by the inverse of the Signal Extraction Scaling method (Diaz et al., 2012) scale factor (described above) and the corresponding control sequencing depth. All tracks were then scaled evenly to fit in the viewing range. To remove potential alignment bias, we removed all aligned reads overlapping repeat regions annotated by RepeatMasker (Smit et al., 1996) and segmental duplication regions annotated by Variant Annotation Tools (San Lucas et al., 2012) for the hg19 genome using bedtools subtract (options -A) (Quinlan and Hall, 2010); annotated regions were downloaded from the UCSC table browser (Karolchik et al., 2004). Peak

calling for histone modification ChIP-seq was performed relative to a corresponding control using MACS2 (options–broad -g hs–broad-cutoff 0.05) after removing reads overlapping repeat regions.

RNA-seq data processing—Adaptor sequences were removed from paired-end RNA-seq reads using Trim Galore (options–illumina–stringency 13–paired) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Quality control was performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed paired-end reads were aligned to the hg19 genome using STAR (Dobin et al., 2013). Gene counts for the UCSC Known Genes (Hsu et al., 2006) annotation were obtained using STAR option–quantMode. Differential expression analysis was performed using DESeq2 (Love et al., 2014) with a design matrix including variables for patient and condition (design = ~patient + condition). P values were calculated using the Wald test, and a significance threshold was set at Benjamini-Hochberg adjusted p value of 0.01 and a minimum magnitude $\log_2$ fold change of 0.5. The same differential expression analysis was performed for comparing primary leiomyoma cells treated with control shRNA or HOXA13 shRNA. Lowly expressed genes (< 1 FPKM in all patient tissue samples, see below) were removed for gene ontology analyses. For comparing cells with control or HOXA13 construct, gene counts with row-sum below 10 were excluded from the DESeq2 analysis. Gene counts were normalized using cpm function of edgeR (v 3.28.1) (McCarthy et al., 2012; Robinson et al., 2010) in R (v 3.6.3) (options prior.count = 2, log = TRUE) (https://www.r-project.org) and z-scores calculated by scale function in R. The matrix obtained for differential genes with adjusted p value < 0.05 and FPKM of at least 1 was used for plotting hierarchical clustered heatmap using pheatmap (v 1.0.12) (Gu et al., 2016) (Figure 5C).

Principal component analysis (PCA) of RNA-seq data was performed using DESeq2. Prior to PCA analysis we applied a regularized log transform to the RNA-seq gene count data using the DESeq2 rlog function (option blind = False). To obtain the higher principal components used in (Figure S3A), we modified the DESeq2 plotPCA function to return the projections onto additional principal components. The projections onto the first 10 principal components were used to generate (Figure S3A). The clustering was obtained using the adjusted cosine distance metric and an average linkage.

ATAC-seq data processing—Adaptor trimming for paired-end ATAC-seq reads followed the same approach as for RNA-seq reads. Paired-end reads were aligned using Bowtie2 (Langmead and Salzberg, 2012) with all other options the same as for histone modification ChIP-seq. Quality control was performed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and ATACseqQC (Ou et al., 2018). After alignment, SAMtools was used to remove read pairs with PCR duplicates, supplementary alignments, poor quality alignments, or pairs for which only a single read aligned (options -b -f 2 -F 2828 -F 1024 -q13). Peaks were called using MACS2 (Zhang et al., 2008) callpeak command (options -f BAMPE -g hs -B–keep-dup 1) with a significance q-value threshold of q < 0.05. To define the set of relevant open chromatin regions, we merged the set of called ATAC-seq peaks from matched samples from all eight patients and took the union of these regions. This final set was used for all further ATAC-seq analysis.

REMC data processing—Consolidated alignment files were downloaded from the NIH Roadmap Epigenomics web portal (https://egg2.wustl.edu/roadmap/) (Kundaje et al., 2015). Reads aligning to repeat regions or segmental duplications were removed, as was done for the leiomyoma data. Reads aligning to the sex chromosomes were also removed to prevent gene-specific effects from confounding the tensor decomposition. The order-3 data tensor – with indices for sample, assay, and location – was constructed following the same approach as for the leiomyoma data. Consolidated broakPeak files were downloaded from the Roadmap Epigenomics web portal.

Breast cancer cell line data processing—Raw ChIP-seq data for histone modifications in 13 breast cancer cell lines were obtained from the Gene Expression Omnibus (GEO) (GEO: GSE85158). RNA-seq data were obtained from Xi et al. (2018). ChIP-seq data were processed through the same pipeline as for the leiomyoma data. The order-4 data tensor – with indices for cell line, assay, replicate and location – was constructed using the same approach as for the leiomyoma data. Peak calling for histone modification ChIP-seq was performed relative to a corresponding control using MACS2 (options–broad -g hs–broad-cutoff 0.05) after removing reads overlapping repeat regions.

Prostate cancer data processing—Processed data for RNA-seq and raw data for histone modifications and AR ChIP-seq in a prostate cancer cohort were downloaded from the GEO (GEO: GSE120741, GEO: GSE120738). The ChIP-seq data were processed through the same pipeline as for the leiomyoma data, but reads aligning to the Y chromosome were not removed for this dataset, as all patients were male. We included only those samples that had data for all four ChIP-seq assays included in the original study, and that matched the data summary provided in the published supplementary information (Stelloo et al., 2018). We observed significant variability in the enrichment of H3K4me3 data across the samples. As this mark was found to contribute least to the clustering of samples obtained in the original study, we did not include H3K4me3 data in our DeCET analysis. The order-3 data tensor – with indices for patient, assay, and location – was constructed following the same approach as for the leiomyoma data. For differential expression analysis, the normalized $\log_2$-transformed and ComBat-corrected (Leek et al., 2012) read counts of RNA-seq data were obtained from Stelloo et al. (2018). Differential expression between groups was evaluated using a t test with unequal variance (SciPy ttest_ind option equal_var = False). A p value threshold of 10^{-3} was used to call significantly differentially expressed genes. Gene ontology analysis was performed with DAVID (Huang et al., 2009a, 2009b) using default parameters and the default *Homo sapiens* background.

Gene ontology analysis for leiomyoma samples—For gene ontology (Tables S3 and S7), we used a background gene set consisting of all genes with at least 1 FPKM in RNA-seq data for at least one sample used in the tensor decomposition analysis. Gene lengths for FPKM were defined as the length of the union of all exons for a given gene. The obtained read counts for each gene (number of reads mapping to the gene obtained from STAR option–quantMode) were used to calculate FPKM. Small non-coding RNAs were removed by excluding genes with a total length less than 200nt.

To assign genes to genomic regions, each gene was assigned a 20kb regulatory region consisting of 10kb in each direction of the gene TSS. We used the TSS for the UCSC knownCanonical (Hsu et al., 2006) transcript for each gene. Genes were assigned to identified regions if there was any overlap with the corresponding gene regulatory region. For gene ontology, only those genes that were also differentially expressed in the corresponding direction in the differential expression analysis were included. Gene ontology was performed using DAVID (Huang et al., 2009a, 2009b) with default settings.

Gene ontology analysis for HOXA13 overexpressing cells—Differentially expressed genes obtained above (with adjusted p value < 0.05 and FPKM of 1 or more in at least one of the 21 matched tissue samples) were listed in Metascape (Zhou et al., 2019) and gene ontologies enriched for those genes in *Homo sapiens* were identified (Figure 5D).

DeCET statistical analysis—The histone modification ChIP-seq data for patients 1–21 (Table S1) was used to build the data tensor used in this study. For pt10, data from one tumor (pt10_1) was included in the data tensor, while data from a second tumor (pt10_2) was only used as a test set for the SVM and STM classification. The additional leiomyoma samples (pt22, pt23, pt24_1, pt24_2, pt25 and pt26) were used only as test sets for the SVM and STM classification.

To identify HOSVD location vectors specifying differences between two groups, a one-way ANOVA analysis of the HOSVD projections for each assay was used. The ANOVA test statistics were computed in Python and p values were calculated by the F-distribution using SciPy 1 - f.cdf (options loc = 0, scale = 1, numerator and denominator degrees of freedom provided in main text) (Jones et al., 2001). Levene's test was used to test for the equivalence of the variance between myometrium and leiomyoma samples of the projections onto an HOSVD location vector. Levene's test was implemented using SciPy levene (option center = 'mean'). Mann-Whitney U test was implemented using SciPy mannwhitneyu (option alternative = 'two-sided'). The construction of the data tensor from bed files, the HOSVD and the classification cross-validation were performed using Python v3.6 (https://www.python.org), while the downstream analysis was performed using Python v3.7. Additional Python libraries used were NumPy (Harris et al., 2020) and pandas (W. McKinney, 2010, Proc. Python Sci. Conf., conference). Seaborn (Waskom et al., 2018) and Matplotlib (Hunter, 2007) were used for generating Figures 1B, 1C, 2B, 2C, 3A–3D, 4A, 4C–4F, 6A–6C, 7A–7C, S3A–S3F, S4A–S4F, S5B, S6A–S6G, and S8A–S8E; the code for generating these figures can be found at https://github.com/jssong-lab/DeCET.

Consensus profiles of the replicate data for breast cancer cell lines—The consensus projections onto DeCET location vectors for each breast cancer cell line were obtained by decomposing the replicate space in the HOSVD. That is, the histone modification ChIP-seq data were projected onto the singular vectors for the replicate and location spaces, as

$$\mathcal{X}^{i_{\mathrm{r}}\, i_{\mathrm{cl}}\, i_{\mathrm{a}} i_{l}} = \sum_{\alpha_1 \alpha_4} \left(\sum_{\alpha_2 \alpha_3} \mathcal{S}^{\alpha_1 \alpha_2 \alpha_3 \alpha_4} U^{(cell\ line)}_{i_{\mathrm{cl}},\, \alpha_2} U^{(\mathrm{assay})}_{i_{\mathrm{a}},\, \alpha_3} \right) U^{(replicate)}_{i_{\mathrm{r}},\, \alpha_1} U^{(location)}_{i_{l},\, \alpha_4}.$$

The two columns of the matrix $U^{(\text{replicate})}$ represented roughly the similarities and differences between the replicates, respectively; the consensus projections were obtained by taking the projections onto the first “similarity” vector. These consensus profiles captured the epigenetic features common to both replicates, while correcting for potential batch effects resulting from differences in sequencing depth or antibody efficiency.

Function annotation of location vector regions—Regions with a significant histone modification signal along a given location vector were identified using the thresholding method based on the distribution of all location vector components described above. Functional annotation of the regions with significant histone modifications along a location vector was performed using GREAT (McLean et al., 2010) with default parameters. The summaries of the GREAT annotation in Table S8 were obtained by sorting the annotation results for GO Biological Process by the hypergeometric test rank and taking the first 20 terms.

Unsupervised hierarchical clustering—We used the projections onto the location vectors obtained by the HOSVD to perform unsupervised hierarchical clustering (Figure 1C). Prior to clustering, we observed extreme outlier projections onto the fifth and sixth location vectors. The outliers corresponded to two leiomyoma samples, pt19 and pt21, for which we observed much lower IP enrichment in the H3K4me3 ChIP-seq data compared to other samples. The H3K27ac and H3K4me1 ChIP-seq data did not exhibit such outlier behavior for these samples, and the extreme outlier behavior for H3K4me3 was confined to the projections onto only these two location vectors. To remove bias from these outliers, we only used the projections onto the other 8 of the first 10 HOSVD projections for the unsupervised hierarchical clustering. The distance between two tissue samples was obtained by first computing the adjusted cosine distance between the 8 dimensional vector of projections for each histone modification separately. These distances were then summed over all histone modifications to obtain the total pairwise distance between two samples. Hierarchical clustering using average linkage was performed on the obtained pairwise distance matrix. We used the Python package SciPy (Jones et al., 2001) to calculate the linkage and seaborn (Hunter, 2007; Waskom et al., 2018) to plot the dendrogram and heatmap shown in (Figure 1C). The heatmap in (Figure 1C) shows the HOSVD projections, mean-centered and scaled to unit variance, of all assays for each tissue sample in the dataset; the color bar ranges from the 5^{th} to the 95^{th} percentile.

For the unsupervised clustering of REMC data, the first few location vectors with top singular values were used to cluster the samples. This was compared to the clustering using the full set of location vectors to ensure that the clustering result was robust. For the breast and prostate cancer datasets, the location vectors used for clustering were selected based on the variance of the projections onto these vectors. To select the location vectors, the variance in the projections onto each location vector was computed for each histone modification. The top location vectors with greatest variance (4 for the breast cancer data and 5 for the prostate cancer data) were taken for each histone modification and then pooled into a common list. The vector of projections of each histone modification onto the pooled list was used to perform hierarchical clustering as described above for the leiomyoma data. The color

bar ranges from the 10th to the 90th percentile for the REMC, breast cancer, and prostate cancer datasets (Figures 6, 7A, and 7C).

Visualization of tissue classification—Figures 2C and S5B were generated after training the respective classifier for the corresponding set of training samples. The figures were generated in the same way for both the STM and SVM classifiers. The fitted model for each classifier consisted of a hyperplane parameterized by an un-normalized direction vector $\vec{w}$ orthogonal to the fitted hyperplane and an intercept value b specifying the offset of the hyperplane from the origin. For the SVM the vector $\vec{w}$ was obtained directly, while for the STM it was obtained by vectorizing the matrix whose rank-1 CP decomposition gives the vectors u and v

$$\vec{w} = vec\left(uv^T\right).$$

We computed the signed distance of a given sample vector $\vec{x}$ from the hyperplane as

$$d = \frac{\vec{w} \cdot \vec{x} + b}{\sqrt{\vec{w} \cdot \vec{w}}}.$$

The absolute value of this quantity is the distance from the hyperplane, while the sign specifies the side of the hyperplane the sample is on. Principal component analysis (PCA) was then applied to the vectors after removing the component perpendicular to the hyperplane to identify the direction with the greatest sample variance within the hyperplane.

Classification cross-validation—Leave-one-out cross-validation was used to verify the robustness of the classification. For each of the 21 patients in the full tensor dataset, the tensor scaling and HOSVD were applied to the subtensor with that patient's data removed. The SVM or STM classifier was then trained using the projections from the HOSVD with the 20 remaining patients. After training, the classifier was tested on the patient that was removed and the additional leiomyoma datasets not included in the full training tensor. For classifying subpopulations of leiomyoma samples, the full set of patient data (including matched healthy tissue) was included in the HOSVD. The classifier was then trained and tested using only the projections for leiomyoma samples.

Functional characterization of altered regions—We used the results of the HOSVD to characterize the functional role of the identified regions. The histone modification ChIP-seq data samples were projected onto the basis vectors found for the assay and condition space:

$$\mathcal{X}^{i_c\, i_{pt}\, i_a i_l} = \sum_{\alpha_1\alpha_3} \left(\sum_{\alpha_2\alpha_4} \mathcal{S}^{\alpha_1\alpha_2\alpha_3\alpha_4} U^{(patient)}_{i_{pt},\alpha_2} U^{(location)}_{i_l,\alpha_4} \right) U^{(condition)}_{i_c,\alpha_1} U^{(assay)}_{i_a,\alpha_3}.$$

The two columns of the condition matrix $U^{(\text{condition})}$ represented roughly the similarities and differences between leiomyoma and myometrium samples, respectively. The three columns

of the assay matrix $U^{(assay)}$ described the combinatorial histone modification patterns observed in the data, with the first column corresponding to a weighted combination of all three histone modifications moving together, the second to a tradeoff between H3K4me3 and H3K4me1, and the third to a tradeoff between H3K4me1 and H3K27ac.

For the regions identified as discriminating leiomyoma from myometrium, we fixed the condition basis index α_1 to that specifying the differences between leiomyoma and myometrium. The average projection across patients onto each assay basis vector was then taken for the binned genomic locations with differential signal. The mean projection onto the second assay basis vector specified a tradeoff between H3K4me3 and H3K4me1, and hence the functional role of the region as a promoter versus enhancer (Figures 3A and 3B). The sign of the mean projection was used to classify a region as an enhancer or a promoter. Because the projections are taken onto the condition space basis vector discriminating the conditions, the interpretation of the sign was opposite for the regions that were lower in leiomyoma compared to the regions that were higher. In Figure 3A, we flipped the sign of the projections for regions that were lower in myometrium to make the interpretation of the x axis consistent between the top and bottom panels.

For the regions identified as separating leiomyomas with and without *MED12* mutations, we again fixed the condition basis index α_1 to that specifying the differences between leiomyoma and myometrium. For these regions, *MED12*-mut leiomyomas separated from their corresponding myometrium samples in the opposite direction to *MED12*-wt leiomyomas. This resulted in a change in the sign of the projections onto the condition basis vector for patients with and without *MED12* mutations. We used the sign of the projection onto the first assay basis vector to account for this in an unsupervised way. Since we observed the first assay basis vector specified a weighted change in the profile of all three histone modifications, the sign specified the overall direction of the change in activating histone marks. Therefore, for each region and patient, we first multiplied the projection onto the second assay basis vector by the sign of the projection onto the first assay basis vector before averaging over patients. The sign-corrected patient mean projection onto the second assay basis vector was then used to classify the regions as promoters or enhancers following the same procedure (Figure S6A).

For the distribution around the nearest gene TSS, the minimum distance between the center of each bin and the nearest gene TSS from the UCSC Known Gene (Hsu et al., 2006) annotation was obtained. Distances up to 200kb were binned into 10 equal sized bins, and an eleventh overflow bin was used for all values greater than 200kb. The distribution shows the fraction of the 2kb differential genomic bins that have distance to the nearest TSS within the respected range.

Length scale characterization of epigenetic alterations—To characterize the length scale of the identified epigenetic changes, we performed a discrete wavelet transform (DWT) multiresolution analysis of the identified location vector. Specifically, we performed the DWT of the signal vector v as

$$v[n] = \sum_{k=1}^{L(J)} a_k \phi_{-Jk}[n] + \sum_{j=1}^{J} \sum_{k=1}^{L(j)} b_{jk} \psi_{-jk}[n]$$

where ϕ_{-Jk} are the scaling functions, ψ_{-jk} the corresponding wavelet functions, and k and j the translation and scaling indices, respectively. For the genomic location vectors used in the tensor model, k specifies the genomic location of a 2000bp bin, while j specifies the scale of the wavelet. The scale index j can be interpreted as specifying a wavelet window of width 2^j * 2000bp. For our analysis we used the Coiflet 5 wavelets with a maximum level $J = 10$, which corresponds to a scale of 2048kb. The discrete wavelet transforms were performed using the PyWavelets Python package (Lee et al., 2019).

Because we were interested in comparing the scale of the differential epigenetic signal between two conditions, we first identified the location vector along which the two conditions diverged. This location vector was then split into positive and negative components

$$v = v^+ - v^-$$

where $v_l^+ = \max(v_l, 0)$; $v_l^- = -\min(v_l, 0)$. These vectors were then binarized to restrict the transform to the regions with significant changes and to focus on the length scale rather than the effect size. The binarization set the insignificant regions to 0 and the significant regions to 1, and then scaled each vector to unit ℓ_2-norm. The DWTs of the binarized positive and negative component vectors were then computed separately. Denoting the obtained wavelet coefficients for scale index j and translation index k by b_{jk}^+ and b_{jk}^- for the binarized positive and negative component vectors, respectively, we computed the squared sum of the wavelet coefficients across the translation index

$$b_j^{\pm} = \sqrt{\sum_{k=1}^{L(j)} \left(b_{jk}^{\pm}\right)^2}.$$

These values were used to compare the scale of the epigenetic signal along the positive and negative components of the vector (Figures 4A and S6D).

As a separate characterization of the length scale, we extracted normalized ChIP-seq signals at the differential regions and performed a DWT. For each patient, condition, and assay a vector was obtained consisting of the corresponding normalized ChIP-seq signals at the identified differential regions (both increased and decreased histone modification regions). Regions without differential histone modifications were set to zero. These vectors were first scaled to unit ℓ_2-norm and then the DWT was performed on each. The squared wavelet coefficients were summed across the translation index, and the resulting values were then averaged over patients. The square roots of the group mean coefficients were used to compare the scale usage between the two groups.

Chromatin contact domain confinement—Chromatin contact domains identified from Hi-C data in HeLa cells (Rao et al., 2014) were accessed from Gene Expression Omnibus (GEO: GSE63525). Overlapping contact domains, but not adjacent domains, were merged into a single domain. To correlate the histone modification alterations around the domains, each domain was first split into 5 consecutive windows of equal size extending the full length of the domain. A region of the same length on each side of the domain was also split into 5 consecutive windows. Each window was assigned the mean value of the differential vector entries corresponding to the bins within that window. This resulted in a 15 dimensional vector for each of the merged contact domains; the vector entries were sorted by the corresponding genomic location order. The Pearson correlation coefficient taken across the domains was obtained for each pair of windows and plotted as a heatmap (Figures 4C and S6E). The Pearson correlation was calculated using SciPy pearsonr. We also repeated this analysis using an extension of 50kb or 100kb on each side of the domain, which qualitatively showed the same results. As a negative control, we shuffled the contact domains by randomly repositioning each domain along the corresponding chromosome; this resulted in distance-dependent correlations, but removed the domain block structure seen with the true domain locations (Figures 4D and S6F). As an additional negative control, we shuffled the differential location vector components with the contact domains fixed at the true locations, which removed all significant correlations (Figures 4E and S6G).

To identify contact domains with changes in chromatin state, the net change in histone modification profiles across a domain was quantified by summing the significant entries of the differential location vector across the domain. Only binned genomic regions completely contained within the contact domain were included in the sum, and entries not passing the threshold for significant changes were set to zero to remove bias from insignificant changes. When the contact domains were sorted by the net change, we observed a very rapid increase in the net change at the edges of this distribution (Figure 4F). We applied a method similar to that used by the ROSE algorithm for calling super enhancers (Whyte et al., 2013) to identify the point in the distribution where the net change increased rapidly. Contact domains with non-zero change were sorted by the absolute value of the net change, and scaled so that the x and y axis ranged from 0 to 1. The point of rapid increase was then found as the point for which a line with slope 1 was tangent to the resulting curve.

Super enhancer analysis—To identify putative super enhancers in leiomyoma, we first defined a common set of constituent enhancers by merging the H3K27ac peaks called with MACS2 from the 21 leiomyoma samples with patient-matched myometrium. For each of the 21 leiomyoma samples, the ROSE algorithm (Lovén et al., 2013; Whyte et al., 2013) was used to compute the H3K27ac ChIP-seq read density (using the bam files prior to repeat filtering) at the common enhancers, and to call super enhancers from the common enhancer set using this read density. Putative super enhancers were identified as stitched enhancers from the common enhancer set that were called as a super enhancer based on the read density of at least one of the patients. We performed a permutation test to statistically test for the enrichment of the differential genomic bins in the putative super enhancers. The enhancer centers were permuted for all enhancers on the same chromosome with the enhancer lengths fixed. The number of altered genomic bins overlapping the permuted super

enhancers was then computed. The same permutation test was used to assess whether super enhancers were preferentially altered. For this test, the number of permuted super enhancers overlapping an altered region was computed. Both tests were performed separately for the regions with increased and decreased histone modifications in leiomyoma. In 1000 iterations of both tests, the true number of differential regions overlapping super enhancers and the number of super enhancers containing alterations were always greater than the values from permutations.

Changes in ATAC-seq signal—ATAC-seq peak summits were identified with the MACS2 peak caller (Zhang et al., 2008) for matched leiomyoma and myometrium from 8 patients. For each peak summit that overlapped a region with altered histone modifications in leiomyoma, a peak region was defined by extending 500bp on each side of the peak summit. For all eight matched leiomyoma and myometrium ATAC-seq datasets, the pileup of reads across each of these peak regions was obtained using samtools mpileup (Li et al., 2009). The pileup for each sample was scaled by 10^7 divided by the total number of aligned reads for the corresponding sample. For each peak region, a consensus ATAC-seq signal was obtained separately for leiomyoma and myometrium by taking the median scaled signal at each base across the eight patients. The mean of these consensus profiles across peak regions was then taken separately for the summits overlapping regions with increased or decreased active histone modifications in leiomyoma. Finally, the ratio of the mean of the signal at the ± 500bp locations was used to scale the myometrium to the leiomyoma signal in Figure S7A and the leiomyoma to the myometrium signal in Figure S7B.

Motif scanning—Motif scanning was performed using the method described in Hejna et al. (2019), using the position specific scoring matrices (PSSM) from the HOCOMOCO core collection of human transcription factors (Kulakovskiy et al., 2018). We only used motifs for transcription factors having FPKM greater than 1 in at least one sample. Motifs were called by computing the log-likelihood ratio between the PSSM matrix and a second-order Markov background distribution fit to the nucleotide content of the hg19 human genome. We set a cutoff threshold for calling a motif hit at the relative entropy of the distribution specified by the PSSM matrix with respect to the background distribution.

Motif enrichment analysis—After identifying sets of regions S_1 and S_2 associated with two conditions, we performed motif scanning to identify transcription factor motifs that were differentially enriched between the two sets of regions. The union of the regions within a set was taken to remove any arbitrary separation of neighboring regions that resulted from the genome binning. The resulting sets of regions were then intersected with the combined set of ATAC-seq peaks to limit the analysis to chromatin regions open in at least one condition for some patient. Intersected regions smaller than 25bp in length were removed prior to motif calling. For each region, a binary vector was obtained indicating the presence or absence of TF motifs. Fisher's exact test was used to test for the enrichment of a given motif between the two sets of regions. A threshold for significance was set using the Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedure with a false discovery rate of 0.001. The Fisher's exact test p values were calculated with SciPy fisher_exact (options alternative = 'two-sided').

Sequence-based region classification—Sequence-based classification was performed using the binary motif presence vectors described in the preceding section. An ℓ_1 regularized logistic regression classifier was trained on the motif presence vectors to classify a region as having come from set S_1 or S_2. The logistic regression was implemented using sklearn.linear_model.LogisticRegression (Pedregosa et al., 2011). A weight was included for each region in the loss function to account for imbalance in the number of regions from each class (option class_weight = balanced). The regularization parameter was chosen by a grid search to minimize the 5-fold cross-validation loss. The regions from both sets S_1 and S_2 were split into 5 approximately equal-sized groups. The 5-fold cross-validation loss on the test sets was obtained for each value of the regularization parameter between 0.01 and 1.0 using a step size of 0.01. After the regularization parameter was chosen, we performed 500 iterations of Monte Carlo cross-validation. At each iteration, 20% of the regions were removed for testing, and the classifier was trained on the remaining 80% of regions. The mean accuracies after the 500 iterations were reported as the training and validation accuracies. To identify the most informative motifs, the classifier was trained on the entire set of regions and the fitted coefficients were sorted by absolute value. The sign of the fitted coefficient was used to determine the associated condition.

Relative mRNA expression—For quantifying relative mRNA expression, Ct (cycle threshold) values of genes in each sample were first normalized with *GAPDH* and then expression relative to either myometrium or shControl was determined by $2^{-\Delta\Delta Ct}$ method. Data was plotted as mean of the technical replicates for each patient and error bar represents the standard deviation. Significance was determined by performing student's two-tailed t test in excel (*** p value < 0.001, ** p value < 0.01).

Jaccard index clustering—For each histone modification the pairwise Jaccard index between two samples was computed using bedtools jaccard. To integrate information from multiple histone modifications in clustering a set of samples, we first quantile normalized the pairwise similarity matrix to account for differences in the distribution of Jaccard indices for different histone modifications. Quantile normalization was performed by flattening the pairwise similarity matrix into a vector and then quantile normalizing the vector for each histone to a reference distribution obtained as the mean of the sorted pairwise similarity vectors. A consensus similarity matrix was obtained by unfolding the quantile normalized vectors and taking the mean across histones. For clustering a set of samples, a distance matrix was obtained by subtracting this consensus similarity matrix from a matrix of ones and performing complete linkage hierarchical clustering.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank all tissue donors who consented. Stacy A. Kujawa helped with informed patient consent and collection of tissues from the Prentice hospital. Northwestern University NUSeq core and Pathology Core performed Agilent Bioanalyzer assay and IHC staining on TMA slides, respectively. This research is supported by NIH grants R01HD089552, P01HD057877, P50HD098580, R01CA163336, and R01CA196270.

REFERENCES

Al-Hendy A, Myers ER, and Stewart E (2017). Uterine Fibroids: Burden and Unmet Medical Need. Semin. Reprod. Med 35, 473–480. [PubMed: 29100234]

Anders S, and Huber W (2010). Differential expression analysis for sequence count data. Genome Biol. 11, R106. [PubMed: 20979621]

Benjamini Y, and Hochberg Y (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Stat. Methodol 57, 289–300.

Buenrostro JD, Wu B, Chang HY, and Greenleaf WJ (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. 109, 21.29.21–21.29.29.

Bulun SE (2013). Uterine fibroids. N. Engl. J. Med 369, 1344–1355. [PubMed: 24088094]

Cai D, He X, Wen J. r., Han J, Ma W. y., Cai D, He X, Wen J. r., Han J, and Ma W. y. (2006). Support tensor machines for text categorization (Department of Computer Science Technical Report No. 2714) (University of Illinois at Urbana-Champaign).

Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE, et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature 555, 469–474. [PubMed: 29539639]

Ciebiera M, Włodarczyk M, Wrzosek M, Męczekalski B, Nowicka G, Łukaszuk K, Ciebiera M, Słabuszewska-Jóźwiak A, and Jakiel G (2017). Role of Transforming Growth Factor β in Uterine Fibroid Biology. Int. J. Mol. Sci 18, 2435.

Cillo C, Cantile M, Faiella A, and Boncinelli E (2001). Homeobox genes in normal and malignant cells. J. Cell. Physiol 188, 161–169. [PubMed: 11424082]

Commandeur AE, Styer AK, and Teixeira JM (2015). Epidemiological and genetic clues for molecular mechanisms involved in uterine leiomyoma development and growth. Hum. Reprod. Update 21, 593–615. [PubMed: 26141720]

Dai X, Cheng H, Bai Z, and Li J (2017). Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. J. Cancer 8, 3131–3141. [PubMed: 29158785]

Danilo M, Chennupati V, Silva JG, Siegert S, and Held W (2018). Suppression of Tcf1 by Inflammatory Cytokines Facilitates Effector CD8 T Cell Differentiation. Cell Rep 22, 2107–2117. [PubMed: 29466737]

De Lathauwer L, De Moor B, and Vandewalle J (2000). A multilinear singular value decomposition. Siam J. Matrix Anal. A 21, 1253–1278.

Diaz A, Park K, Lim DA, and Song JS (2012). Normalization, bias correction, and peak calling for ChIP-seq. Stat. Appl. Genet. Mol. Biol 11, 9.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. [PubMed: 22495300]

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. [PubMed: 23104886]

Doherty L, Mutlu L, Sinclair D, and Taylor H (2014). Uterine fibroids: clinical manifestations and contemporary management. Reprod. Sci 21, 1067–1092. [PubMed: 24819877]

Du H, and Taylor HS (2015). The Role of Hox Genes in Female Reproductive Tract Development, Adult Function, and Fertility. Cold Spring Harb. Perspect. Med 6, a023002. [PubMed: 26552702]

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Ferrero H (2019). Growth disparities in uterine leiomyomas associated with MED12 mutation. Fertil. Steril 111, 58–59. [PubMed: 30527838]

George JW, Fan H, Johnson B, Carpenter TJ, Foy KK, Chatterjee A, Patterson AL, Koeman J, Adams M, Madaj ZB, et al. (2019). Integrated Epigenome, Exome, and Transcriptome Analyses Reveal Molecular Subtypes and Homeotic Transformation in Uterine Fibroids. Cell Rep 29, 4069–4085.e6. [PubMed: 31851934]

Gu ZD, Shen LY, Wang H, Chen XM, Li Y, Ning T, and Chen KN (2009). HOXA13 promotes cancer cell growth and predicts poor survival of patients with esophageal squamous cell carcinoma. Cancer Res. 69, 4969–4973. [PubMed: 19491265]

Gu Z, Eils R, and Schlesner M (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849. [PubMed: 27207943]

Hall TA (1999). BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT. Nucleic Acids Symp. Ser 41, 95–98.

Harada A, Maehara K, Handa T, Arimura Y, Nogami J, Hayashi-Takanaka Y, Shirahige K, Kurumizaka H, Kimura H, and Ohkawa Y (2019). A chromatin integration labelling method enables epigenomic profiling with lower input. Nat. Cell Biol 21, 287–296. [PubMed: 30532068]

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, et al. (2020). Array programming with NumPy. Nature 585, 357–362. [PubMed: 32939066]

Hejna M, Moon WM, Cheng J, Kawakami A, Fisher DE, and Song JS (2019). Local genomic features predict the distinct and overlapping binding patterns of the bHLH-Zip family oncoproteins MITF and MYC-MAX. Pigment Cell Melanoma Res. 32, 500–509. [PubMed: 30548162]

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, and Young RA (2013). Super-enhancers in the control of cell identity and disease. Cell 155, 934–947. [PubMed: 24119843]

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, and Haussler D (2006). The UCSC Known Genes. Bioinformatics 22, 1036–1046. [PubMed: 16500937]

Huang W, Sherman BT, and Lempicki RA (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13. [PubMed: 19033363]

Huang W, Sherman BT, and Lempicki RA (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc 4, 44–57. [PubMed: 19131956]

Hunter JD (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng 9, 90–95.

Je EM, Kim MR, Min KO, Yoo NJ, and Lee SH (2012). Mutational analysis of MED12 exon 2 in uterine leiomyoma and other common tumors. Int. J. Cancer 131, E1044–E1047. [PubMed: 22532225]

Jones E, Oliphant T, Peterson P, et al. (2001). SciPy: Open source scientific tools for Python..

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–D496. [PubMed: 14681465]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D (2002). The human genome browser at UCSC. Genome Res. 12, 996–1006. [PubMed: 12045153]

Kent WJ, Zweig AS, Barber G, Hinrichs AS, and Karolchik D (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics 26, 2204–2207. [PubMed: 20639541]

Kossaifi J, Panagakis Y, Anandkumar A, and Pantic M (2019). TensorLy: Tensor Learning in Python. J. Mach. Learn. Res 20, 1–6.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 46 (D1), D252–D259. [PubMed: 29140464]

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330. [PubMed: 25693563]

Langmead B, and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. [PubMed: 22388286]

Lee GR, Gommers R, Waselewski F, Wohlfahrt K, and O'Leary A (2019). PyWavelets: A Python package for wavelet analysis. J. Open Source Softw 4, 1237.

Leek JT, Johnson WE, Parker HS, Jaffe AE, and Storey JD (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883. [PubMed: 22257669]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. [PubMed: 19505943]

Li Z, Zhao X, Zhou Y, Liu Y, Zhou Q, Ye H, Wang Y, Zeng J, Song Y, Gao W, et al. (2015). The long non-coding RNA HOTTIP promotes progression and gemcitabine resistance by regulating HOXA13 in pancreatic cancer. J. Transl. Med 13, 84. [PubMed: 25889214]

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293. [PubMed: 19815776]

Love MI, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. [PubMed: 25516281]

Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, and Young RA (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 153, 320–334. [PubMed: 23582323]

Luo H, Zhu G, Xu J, Lai Q, Yan B, Guo Y, Fung TK, Zeisig BB, Cui Y, Zha J, et al. (2019). HOTTIP lncRNA Promotes Hematopoietic Stem Cell Self-Renewal Leading to AML-like Disease in Mice. Cancer Cell 36, 645–659.e8. [PubMed: 31786140]

Mäkinen N, Mehine M, Tolvanen J, Kaasinen E, Li Y, Lehtonen HJ, Gentile M, Yan J, Enge M, Taipale M, et al. (2011). MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. Science 334, 252–255. [PubMed: 21868628]

McCarthy DJ, Chen Y, and Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40, 4288–4297. [PubMed: 22287627]

McGuire MM, Yatsenko A, Hoffner L, Jones M, Surti U, and Rajkovic A (2012). 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. [PubMed: 22428002]

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, and Bejerano G (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol 28, 495–501. [PubMed: 20436461]

McWilliams MM, and Chennathukuzhi VM (2017). Recent Advances in Uterine Fibroid Etiology. Semin. Reprod. Med 35, 181–189. [PubMed: 28278535]

Mehine M, Mäkinen N, Heinonen HR, Aaltonen LA, and Vahteristo P (2014). Genomics of uterine leiomyomas: insights from high-throughput sequencing. Fertil. Steril 102, 621–629. [PubMed: 25106763]

Mehine M, Kaasinen E, Heinonen HR, Mäkinen N, Kämpjärvi K, Sarvilinna N, Aavikko M, Vähärautio A, Pasanen A, Bützow R, et al. (2016). Integrated data analysis reveals uterine leiomyoma subtypes with distinct driver pathways and biomarkers. Proc. Natl. Acad. Sci. USA 113, 1315–1320. [PubMed: 26787895]

Meloni AM, Surti U, Contento AM, Davare J, and Sandberg AA (1992). Uterine leiomyomas: cytogenetic and histologic profile. Obstet. Gynecol 80, 209–217. [PubMed: 1635734]

Mohammed H, Russell IA, Stark R, Rueda OM, Hickey TE, Tarulli GA, Serandour AA, Birrell SN, Bruna A, Saadi A, et al. (2015). Progesterone receptor modulates ERα action in breast cancer. Nature 523, 313–317. [PubMed: 26153859]

Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, Moutinho C, Heyn H, Diaz-Lagares A, de Moura MC, et al. (2016). Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 17, 1386–1395. [PubMed: 27575023]

Moyo MB, Parker JB, and Chakravarti D (2020). Altered chromatin landscape and enhancer engagement underlie transcriptional dysregulation in MED12 mutant uterine leiomyomas. Nat. Commun 11, 1019. [PubMed: 32094355]

Nibert M, and Heim S (1990). Uterine leiomyoma cytogenetics. Genes Chromosomes Cancer 2, 3–13. [PubMed: 2278965]

Orozco JIJ, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jalas JR, Wilmott JS, Hothi P, Wang X, Takasumi Y, et al. (2018). Epigenetic profiling for the molecular classification of metastatic brain tumors. Nat. Commun 9, 4627. [PubMed: 30401823]

Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, and Zhu LJ (2018). ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genomics 19, 169. [PubMed: 29490630]

Parker JB, Palchaudhuri S, Yin H, Wei J, and Chakravarti D (2012). A transcriptional regulatory role of the THAP11-HCF-1 complex in colon cancer cell function. Mol. Cell. Biol 32, 1654–1670. [PubMed: 22371484]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res 12, 2825–2830.

Pipes GC, Creemers EE, and Olson EN (2006). The myocardin family of transcriptional coactivators: versatile regulators of cell growth, migration, and myogenesis. Genes Dev. 20, 1545–1556. [PubMed: 16778073]

Quagliata L, Quintavalle C, Lanzafame M, Matter MS, Novello C, di Tommaso L, Pressiani T, Rimassa L, Tornillo L, Roncalli M, et al. (2018). High expression of HOXA13 correlates with poorly differentiated hepatocellular carcinomas and modulates sorafenib response in in vitro models. Lab. Invest 98, 95–105. [PubMed: 29035381]

Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. [PubMed: 20110278]

Raimundo N, Vanharanta S, Aaltonen LA, Hovatta I, and Suomalainen A (2009). Downregulation of SRF-FOS-JUNB pathway in fumarate hydratase deficiency and in uterine leiomyomas. Oncogene 28, 1261–1273. [PubMed: 19151755]

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, and Kent WJ (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30, 1003–1005. [PubMed: 24227676]

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, and Aiden EL (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680. [PubMed: 25497547]

Rein MS, Friedman AJ, Barbieri RL, Pavelka K, Fletcher JA, and Morton CC (1991). Cytogenetic abnormalities in uterine leiomyomata. Obstet. Gynecol 77, 923–926. [PubMed: 2030869]

Robinson MD, McCarthy DJ, and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. [PubMed: 19910308]

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP (2011). Integrative genomics viewer. Nat. Biotechnol 29, 24–26. [PubMed: 21221095]

San Lucas FA, Wang G, Scheet P, and Peng B (2012). Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics 28, 421–422. [PubMed: 22138362]

Skene PJ, and Henikoff S (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 6, e21856. [PubMed: 28079019]

Smit AFA, Hubley R, and Green P (1996–2010). RepeatMasker Open-3.0. http://www.repeatmasker.org.

Stelloo S, Nevedomskaya E, Kim Y, Schuurman K, Valle-Encinas E, Lobo J, Krijgsman O, Peeper DS, Chang SL, Feng FY, et al. (2018). Integrative epigenetic taxonomy of primary prostate cancer. Nat. Commun 9, 4900. [PubMed: 30464211]

Stewart EA, Laughlin-Tommaso SK, Catherino WH, Lalitkumar S, Gupta D, and Vollenhoven B (2016). Uterine fibroids. Nat. Rev. Dis. Primers 2, 16043. [PubMed: 27335259]

Stylianou N, Lehman ML, Wang C, Fard AT, Rockstroh A, Fazli L, Jovanovic L, Ward M, Sadowski MC, Kashyap AS, et al. (2019). A molecular portrait of epithelial-mesenchymal plasticity in prostate cancer associated with clinical outcome. Oncogene 38, 913–934. [PubMed: 30194451]

Tao D, Li X, Wu X, Hu W, and Maybank SJ (2007). Supervised tensor learning. Knowl. Inf. Syst 13, 1–42.

Tomlinson IP, Alam NA, Rowan AJ, Barclay E, Jaeger EE, Kelsell D, Leigh I, Gorman P, Lamlum H, Rahman S, et al.; Multiple Leiomyoma Consortium (2002). Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. Nat. Genet 30, 406–410. [PubMed: 11865300]

Wallace DC (2012). Mitochondria and cancer. Nat. Rev. Cancer 12, 685–698. [PubMed: 23001348]

Wang Z, Wang DZ, Hockemeyer D, McAnally J, Nordheim A, and Olson EN (2004). Myocardin and ternary complex factors compete for SRF to control smooth muscle gene expression. Nature 428, 185–189. [PubMed: 15014501]

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472, 120–124. [PubMed: 21423168]

Waskom M, Botvinnik O, O'Kane D, Hobson P, Ostblom J, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Cole JB, et al. (2018–). v0.9.0 (7 2018). https://zenodo.org/record/1313201#.YFjmemhKjIU.

Wei J, Chiriboga L, Mizuguchi M, Yee H, and Mittal K (2005). Expression profile of tuberin and some potential tumorigenic factors in 60 patients with uterine leiomyomata. Mod. Pathol 18, 179–188. [PubMed: 15467714]

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, and Young RA (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319. [PubMed: 23582322]

Willinger T, Freeman T, Herbert M, Hasegawa H, McMichael AJ, and Callan MF (2006). Human naive CD8 T cells down-regulate expression of the WNT pathway transcription factors lymphoid enhancer binding factor 1 and transcription factor 7 (T cell factor-1) following antigen encounter in vitro and in vivo. J. Immunol 176, 1439–1446. [PubMed: 16424171]

Xi Y, Shi J, Li W, Tanaka K, Allton KL, Richardson D, Li J, Franco HL, Nagari A, Malladi VS, et al. (2018). Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. BMC Genomics 19, 150. [PubMed: 29458327]

Yin H, Lo JH, Kim JY, Marsh EE, Kim JJ, Ghosh AK, Bulun S, and Chakravarti D (2013). Expression profiling of nuclear receptors identifies key roles of NR4A subfamily in uterine fibroids. Mol. Endocrinol 27, 726–740. [PubMed: 23550059]

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, and Liu XS (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. [PubMed: 18798982]

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, and Chanda SK (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun 10, 1523. [PubMed: 30944313]

Highlights

- Tensor decomposition provides integrative analysis of epigenomic data
- Leiomyomas exhibit recurrent subtype-specific alterations in histone modifications
- Chromatin contact domains constrain histone modification alterations in leiomyoma
- HOXA13 is a potential tumorigenic factor in leiomyoma

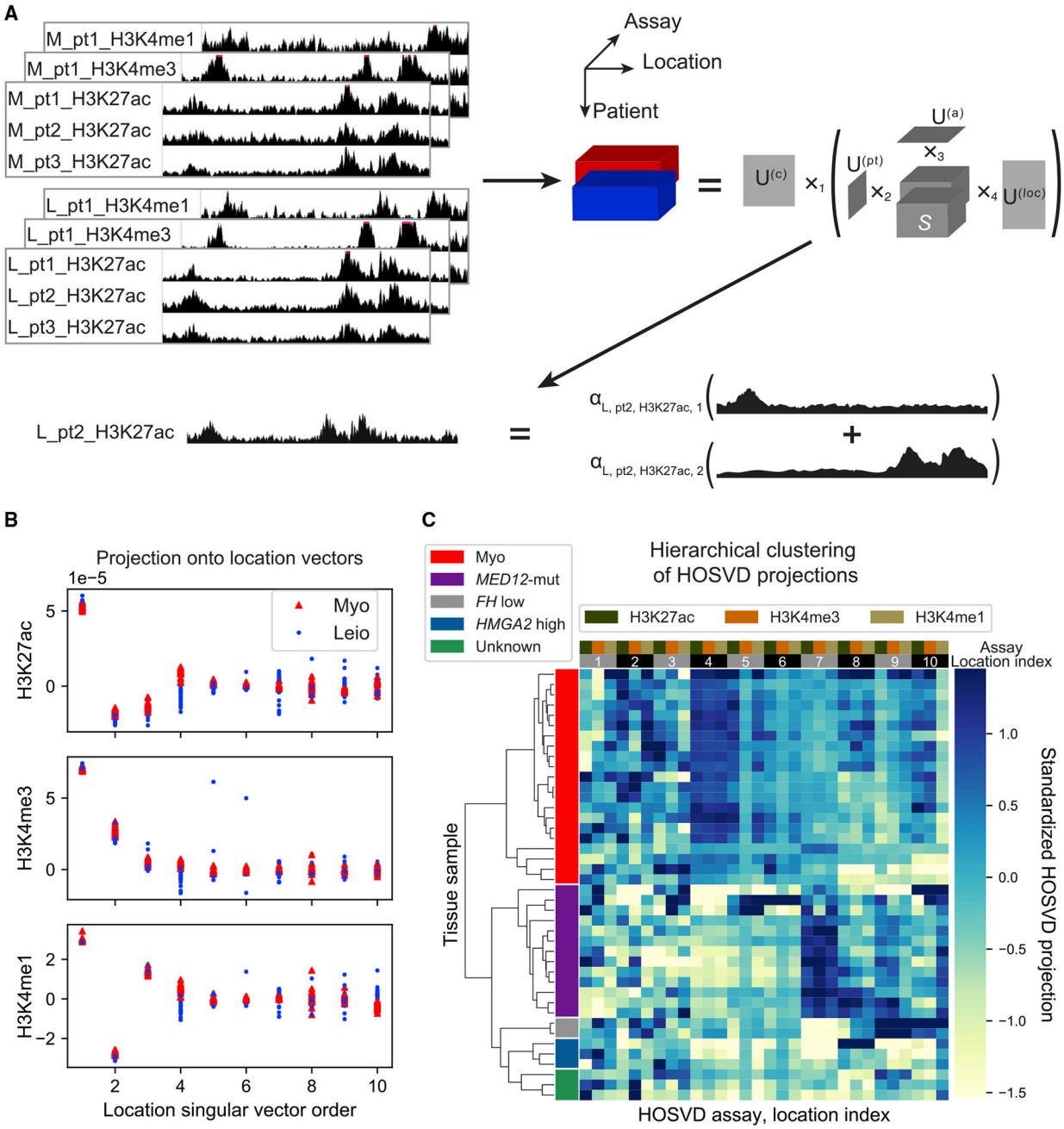


Figure 1. DeCET uncovers epigenetic patterns specific to myometrium and uterine leiomyoma subtypes

(A) Schematic illustration of how the data tensor is decomposed into characteristic modes in each index space (condition (c), patient (pt), assay (a), and genomic location (loc). The bottom portion shows how the decomposition represents each ChIP-seq profile as a projection onto independent spatial patterns of histone modifications.

(B) The projections of ChIP-seq datasets onto the first 10 HOSVD location vectors.

(C) Unsupervised hierarchical clustering of the 21 patient-matched samples using 8 of the first 10 HOSVD projections for each assay (Figure 1B; STAR Methods). The columns correspond to an assay and location vector index pair. Leiomyoma tissues are labeled by

observed mutations (Table S1): *MED12* exon 2 mutations (*MED12*-mut), *HMGA2* overexpression (*HMGA2* high), biallelic loss of FH (*FH* low), or unknown if none of the above three were observed.

See also Figures S1–S4 and Tables S1, S2, S3, S4, and S5.

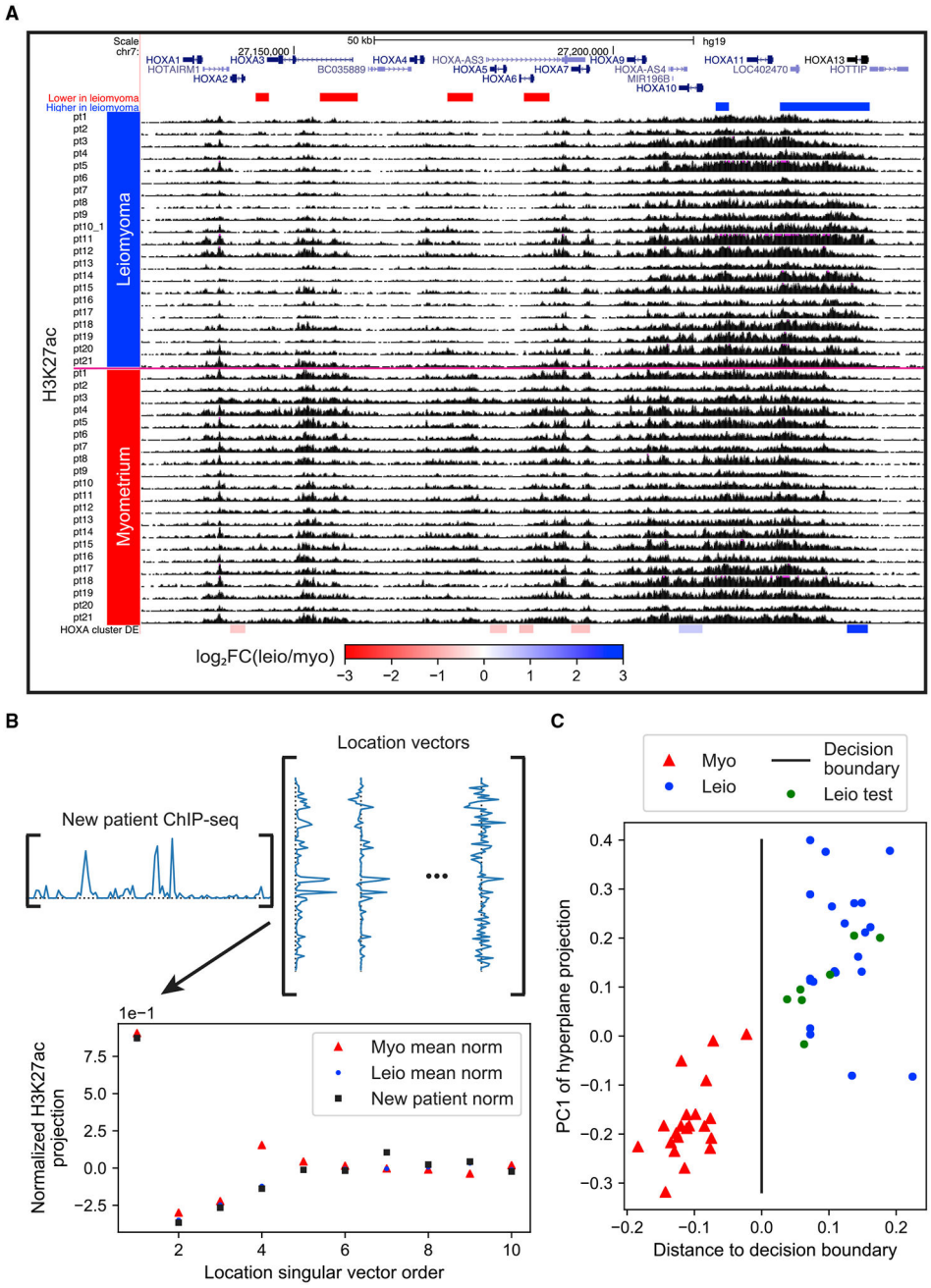


Figure 2. DeCET yields a robust epigenomic classifier of uterine leiomyoma disease status

(A) UCSC genome browser tracks (http://genome.ucsc.edu) of the H3K27ac ChIP-seq data and the identified differential regions at the *HOXA* cluster. Significantly differentially expressed genes (STAR Methods) are shown below the ChIP-seq tracks (FC = fold change).

(B) Illustration of projecting the data for an additional patient not used in the HOSVD onto the location vectors (STAR Methods). The projections were ℓ_2-normalized across the first 10 location vectors; this normalization was used for all classifiers.

(C) STM classification of the 21 patient-matched samples and 7 additional leiomyoma samples. The x axis shows the signed distance from the decision boundary hyperplane (black

line). The y axis shows the first principal component of the data projected onto the decision boundary.

See also Figure S5.

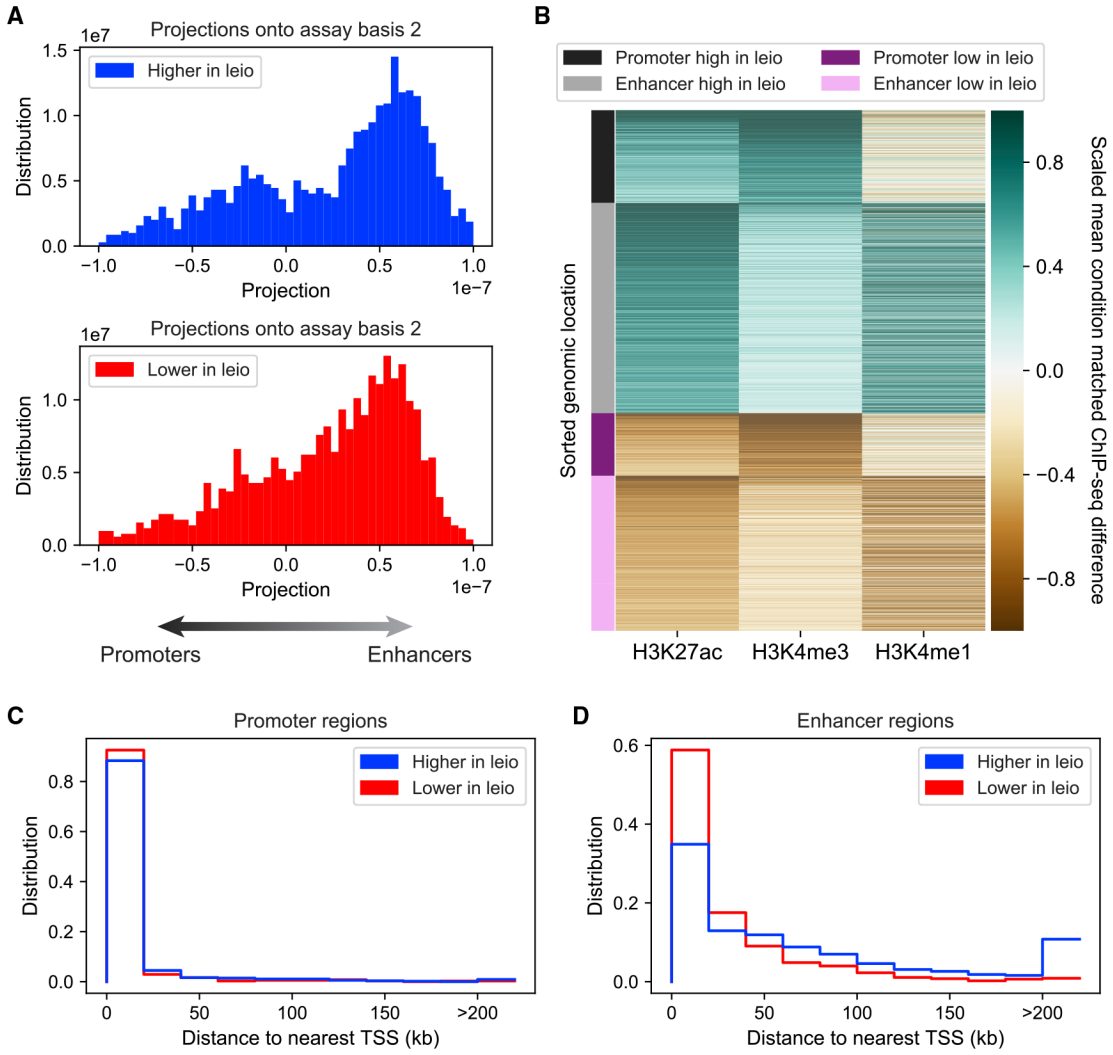


Figure 3. Genomic annotation and distribution of epigenetic alterations in leiomyomas

(A) Distribution of the patient mean projection onto the second basis vector of the assay space for the differential genomic bins higher (top) or lower (bottom) in leiomyoma (STAR Methods). The sign of the x axis was flipped in the bottom plot to make the interpretation of the direction consistent (STAR Methods).

(B) Heatmap showing the mean difference in the normalized ChIP-seq signal between leiomyoma and myometrium at the differential regions identified from the fourth HOSVD vector. Within each functional class, the bins (rows) are sorted by the absolute value of the fourth location vector component (the most differential bins being at the top). Each column is scaled by the 90th percentile of its absolute entry values.

(C and D) The spatial distributions of the center of the 2-kb genomic bins identified in (B) as being differential promoters and enhancers, relative to the nearest gene TSS.

See also Figure S6.

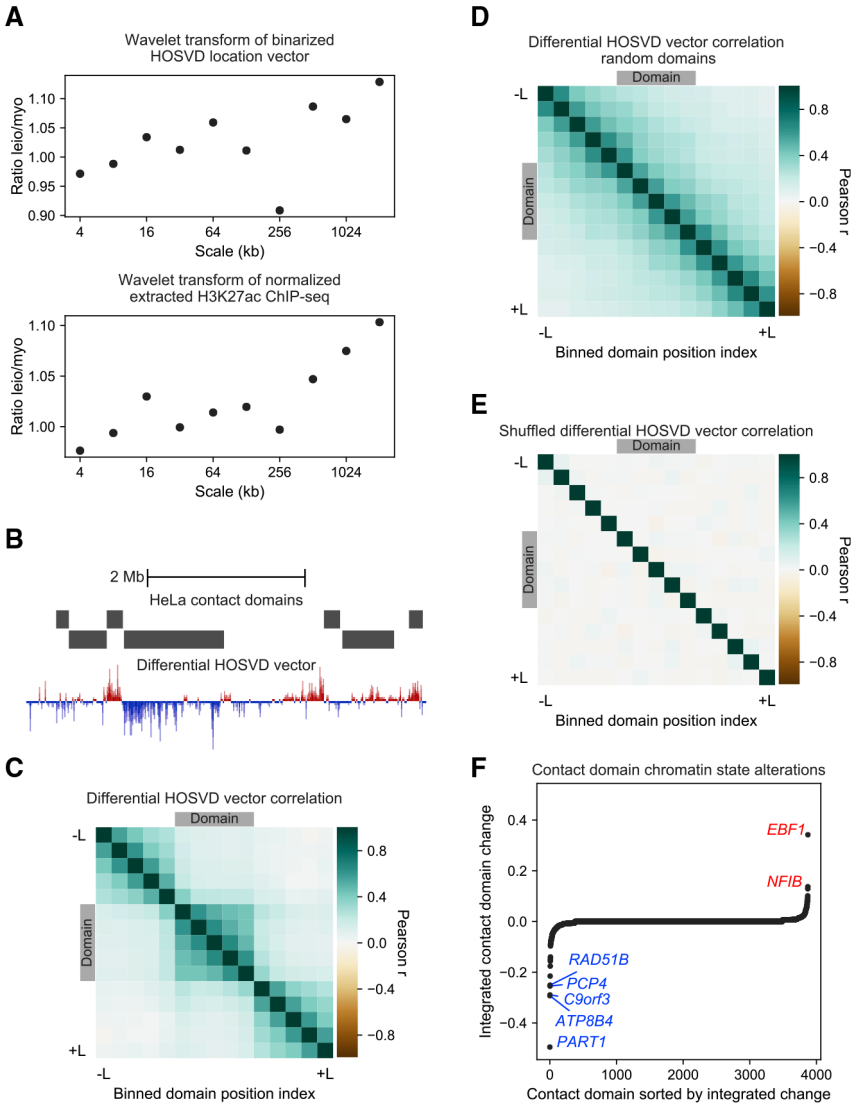


Figure 4. Contact domains confine epigenetic alterations in uterine leiomyomas

(A) Ratio of the discrete wavelet transform coefficients of the binarized fourth location vector (top) and patient mean normalized H3K27ac ChIP-seq data extracted at the differential regions identified from this vector (bottom).

(B) UCSC genome browser track of HeLa contact domains (Rao et al., 2014) and the fourth location vector signal in the region chromosome 6 (chr6): 125,088,000–130,138,000.

(C) Pairwise correlation of the fourth location vector signal at binned regions within and flanking contact domains (STAR Methods). The 15 bins are of the same size and ordered by their genomic position.

(D) Same as (C), but for random domain locations obtained by moving each contact domain to a random location along the same chromosome.

(E) Same as (C), but after shuffling the vector components, while fixing the contact domains at the true locations.

(F) Contact domains sorted by the summed fourth location vector signal at significantly altered regions within each domain (STAR Methods). Some top domains are labeled by the gene showing the greatest differential expression within the corresponding domain.
See also Figure S6.

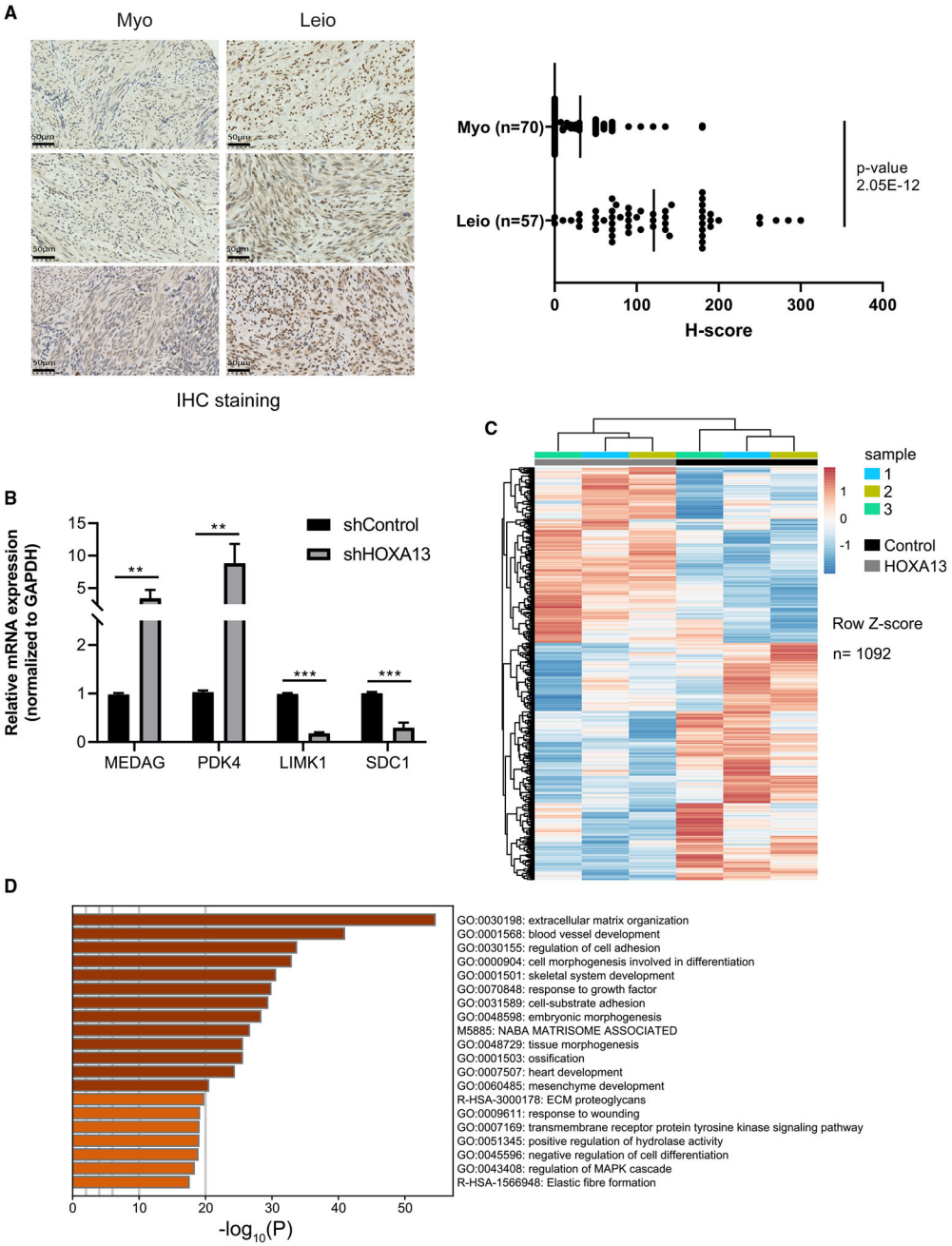


Figure 5. HOXA13 is elevated in and regulates leiomyoma pathogenesis

(A) (Left) Representative images showing HOXA13 IHC staining in the normal myometrium (Myo) and leiomyoma (Leio) (scale = 50 μm). (Right) Scatter dot plot of H-score measured for nuclear HOXA13 in normal myometrium (Myo; n = 70) and leiomyoma (Leio; n = 57) from HOXA13-stained tissue microarrays. Vertical dashed line shows mean H-score for each condition. p value was from the two-tailed t test.

(B) Relative mRNA levels of MEDAG, PDK4, LIMK1, SDC1 measured by qRT-PCR in primary leiomyoma cells treated with shControl or shHOXA13. Mean fold-change in shHOXA13 relative to shControl is shown, with error bars representing standard deviation of

biological replicates (n = 3). Significance was from the two-tailed t test (***p < 0.001, **p < 0.01).

(C) Hierarchically clustered heatmap of differentially expressed genes (adjusted p value < 0.05) in *HOXA13*-overexpressing primary myometrial cells from three patients (samples 1, 2, and 3). Gene expression relative to the mean expression in control and HOXA13 construct-containing cells are shown as row *Z* scores.

(D) Bar plot of GO enrichment scores (Zhou et al., 2019) for differentially expressed genes in *HOXA13*-overexpressing cells.

See also Figure S7 and Tables S2, S6, and S7.

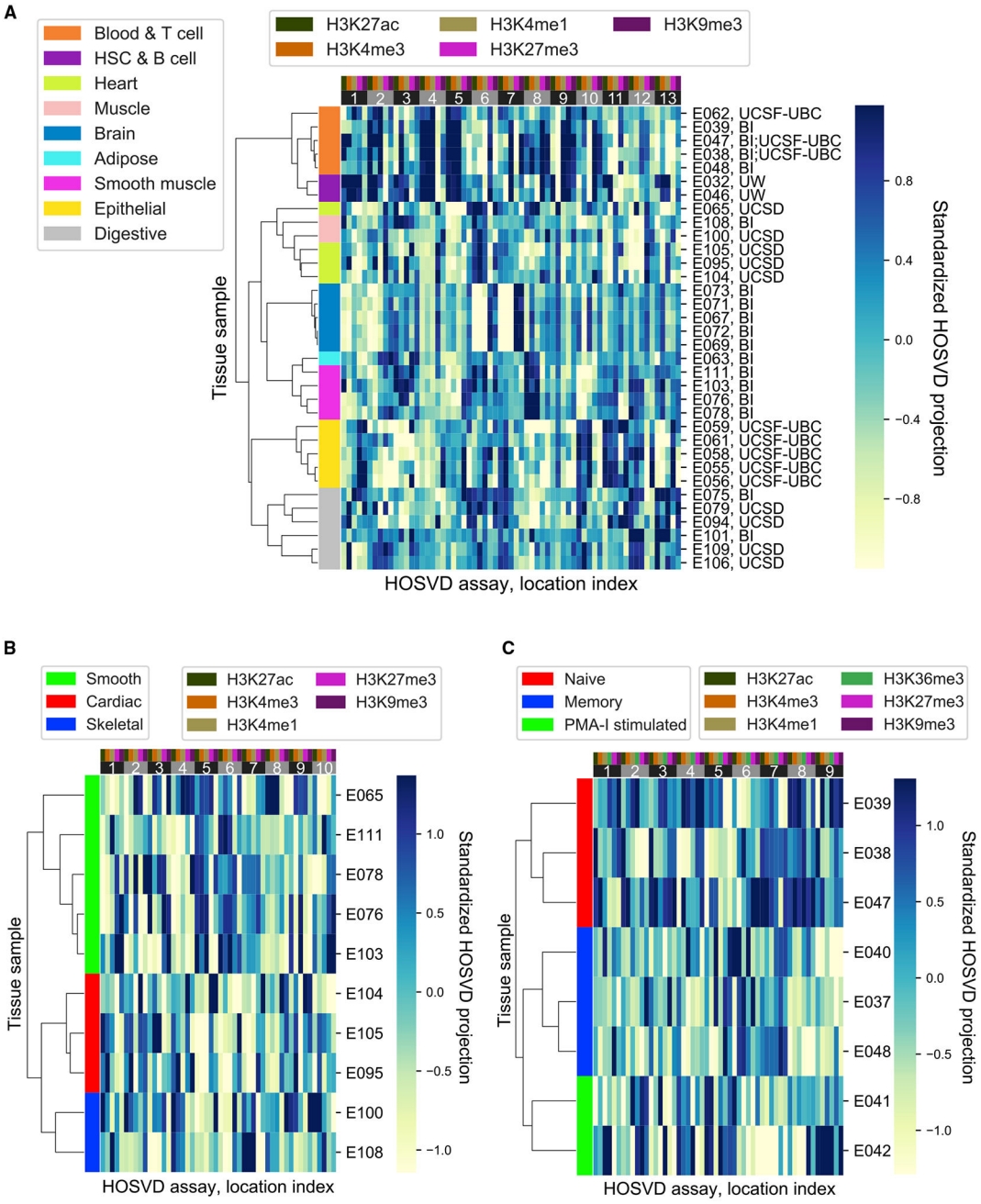


Figure 6. DeCET reveals epigenetic organization of tissue types and differentiation states in REMC data

(A) Hierarchical clustering of 34 adult human tissues using the projections onto the first 13 HOSVD location vectors. The REMC sample identifiers and associated laboratory are shown on the right.

(B) Hierarchical clustering of 10 adult human muscle tissues using the projections onto the first 10 location vectors.

(C) Hierarchical clustering of 8 T cell samples representing three differentiation states using the projections onto the first 9 location vectors.

See also Figure S8 and Table S8.

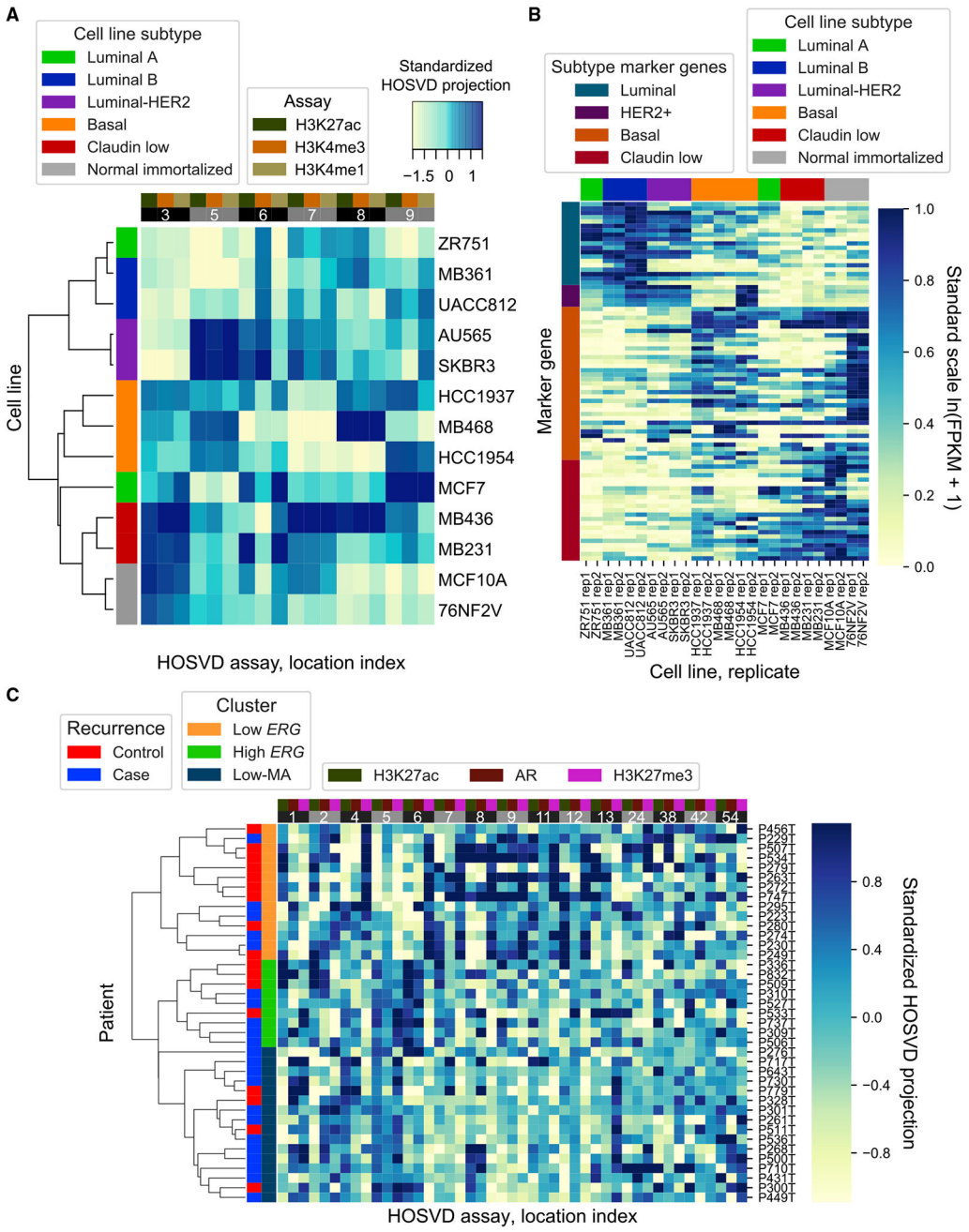


Figure 7. DeCET provides epigenetic stratification of breast and prostate cancer subtypes

(A) Hierarchical clustering of 13 breast cancer cell lines based on three histone modifications (STAR Methods). The 6 location vectors with greatest variance in the projections across samples were used for clustering (STAR Methods).

(B) The expression pattern of known marker genes for breast cancer subtypes (Dai et al., 2017). The cell lines are ordered according to the clustering in (A). Each row is scaled by its largest value for visualization.

(C) Hierarchical clustering of prostate cancer samples using the 15 location vectors with greatest variation in the projections (STAR Methods). Samples are colored according to the

status of a biochemical recurrence (case) or no recurrence (control). The cluster labels, named based on expression signatures, represent the DeCET clustering.
See also Figure S8 and Table S8.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-GAPDH	Sigma-Aldrich	G9545; RRID:AB_796208
Rabbit polyclonal anti-HOXA13	Abcam	ab106503; RRID:AB_11128701
Rabbit polyclonal anti-H3K27ac	Active Motif	39133; RRID:AB_2561016
Rabbit polyclonal anti-H3K4me3	Diagenode	C15410003; RRID:AB_2616052
Rabbit polyclonal anti-H3K4me1	Diagenode	C15410194; RRID:AB_2637078
Rabbit polyclonal anti-Histone3	Abcam	ab1791; RRID:AB_302613
Mouse monoclonal anti- V5	Thermo Fisher Scientific	R960-25; RRID:AB_2556564
Mouse monoclonal anti-HMGA2	Genetex	GTX629478
Biological samples		
Fresh human uterine leiomyoma and matched myometrium tissues	Northwestern University Prentice Women's Hospital	N/A
Chemicals, peptides, and recombinant proteins		
HOXA13 inserted pLEX_306 plasmid	This paper	N/A
Critical commercial assays		
SimpleChIP kit	Cell Signaling Technology	9003
Kapa hyper prep kit	Kapa Biosystems	KK8502
Kapa quantification kit	Kapa Biosystems	KK4835
RNeasy Fibrous tissue kit	QIAGEN	74704
TruSeq stranded mRNA kit	Illumina	20020594
DNeasy blood and tissue kit	QIAGEN	69504
CellTiter-Glo® 2.0 Cell Viability Assay	Promega	G9241
Deposited Data		
ChIP-seq, RNA-seq, ATAC-seq	This paper, Gene Expression Omnibus	GEO: GSE142332
Experimental models: Cell lines		
Primary uterine leiomyoma and myometrial cells	Fresh tissues	N/A
Oligonucleotides		
Primers for *MED12* exon2 cDNA sequencing Forward: CTTCGGGATCTTGAGCTACG Reverse: GTTGGAACTGATCTTGGCAGG	This paper	N/A
Primers for *MED12* exon2 genomic DNA sequencing Forward: GCCCTTTCACCTTGTTCCTT Reverse: TGTCCCTATAAGTCTTCCCAACC	This paper	N/A
Primers for *HMGA2* qRT-PCR Forward: AGCAGCAGCAAGAACCAACC Reverse: CTTGGCCGTTTTTCTCCAGTG	This paper	N/A
Primers for *HOXA13* qRT-PCR Forward: ACTCTGCCCGACGTGGT Reverse: CCGCTCAGAGAGATTCGTCG	This paper	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Primers for *MEDAG* qRT-PCR Forward: TCAAGAGGTATGTGGAACTGACC Reverse: TGACCATGTCCATCCCTTGC	This paper	N/A
Primers for *PDK4* qRT-PCR Forward: CAGACAGGAAACCCAAGCCA Reverse: TTGCCCGCATTGCATTCTTA	This paper	N/A
Primers for *LIMK1* qRT-PCR Forward: ATCAGGGATGGCCTACCTCC Reverse: CAGGCTGAGTCTTCTCGTCC	This paper	N/A
Primers for *SDC1* qRT-PCR Forward: GGAAGGGCCTGTGGGTTTA Reverse: CGCTCTCTACTGCCGGATTC	This paper	N/A
Primers for *GAPDH* qRT-PCR Forward: TGCACCACCAACTGCTTAGC Reverse: GGCATGGACTGTGGTCATGAG	This paper	N/A
shRNA (sh*HOXA13*): CCGGGTTCCAGAACAGGAGGGTTAACTCGAGTTAACCCTCCTGTTCTGGAACTTTTT	Sigma	TRCN0000015406
Software and algorithms		
Indigo	GitHub	https://www.gear-genomics.com
FastQC v0.11.5	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Bowtie2 v2.3.2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
trim galore v0.4.4	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Picard v2.10.1	Broad Institute	https://broadinstitute.github.io/picard/
MACS2 v2.1.1	Zhang et al., 2008	https://github.com/macs3-project/MACS
STAR v2.5.3a	Dobin et al., 2013	https://github.com/alexdobin/STAR
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ATACseqQC	Ou et al., 2018	https://bioconductor.org/packages/release/bioc/html/ATACseqQC.html
SAMtools v1.7	Li et al., 2009	http://samtools.sourceforge.net
Bedtools v2.26.0	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/
DAVID	Huang et al., 2009a, 2009b	https://david.ncifcrf.gov
Python v3.6.1 and v3.7.3	Python	https://www.python.org
NumPy v1.18.5 and v1.16.4	Harris et al., 2020	https://numpy.org
PyTorch v0.4.0	A. Paszke et al., 2017, NIPS Autodiff Workshop, conference	https://pytorch.org
Tensorly	Kossaifi et al., 2019	http://tensorly.org/stable/index.html
SciPy	Jones et al., 2001	https://www.scipy.org
Scikit-learn v0.21.2	Pedregosa et al., 2011	https://scikit-learn.org/stable/
pandas v0.24.2	W. McKinney, 2010, Proc. Python Sci.	https://pandas.pydata.org

REAGENT or RESOURCE	SOURCE	IDENTIFIER
	Conf., conference	
Seaborn v0.9.0	Waskom et al., 2018	https://seaborn.pydata.org
PyWavelets v1.0.3	Lee et al., 2019	https://pywavelets.readthedocs.io/en/latest/#
Ranking of Super Enhancer (ROSE)	Lovén et al., 2013; Whyte et al., 2013	https://bitbucket.org/young_computation/rose/src/master/
The Human Genome Browser at UCSC	Kent et al., 2002	https://genome.ucsc.edu
Integrative Genomics Viewer	Robinson et al., 2011	https://software.broadinstitute.org/software/igv/
BioEdit	Hall, 1999	http://en.bio-soft.net/format/BioEdit.html
GREAT v4.0.4	McLean et al., 2010	http://great.stanford.edu/public/html/index.php
DeCET	This paper	https://github.com/jssong-lab/DeCET (https://doi.org/10.5281/zenodo.4540815)
Metascape	Zhou et al., 2019	https://metascape.org
R v3.6.3	R	https://www.r-project.org
edgeR v3.28.1	McCarthy et al., 2012; Robinson et al., 2010	http://bioinf.wehi.edu.au/edgeR
pheatmap v1.0.12	Gu et al., 2016	http://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html