Tissue-specific 5-hydroxymethylcytosine landscape of the human genome

Bo He
Peking University
Chao Zhang
Peking University
Xiaoxue Zhang
Peking University
Yu Fan
Peking University First Hospital
Hu Zeng
Peking University
Jun'e Liu
Peking University
Haowei Meng
Peking University, Beijing 100871 https://orcid.org/0000-0001-9695-5060
Dongsheng Bai
Peking University
Jinying Peng
Peking University
Qian Zhang
Peking University First Hospital
Wei Tao
Peking University
Chengqi Yi (chengqi.yi@pku.edu.cn)
Peking University https://orcid.org/0000-0003-2540-9729

Article

Keywords: Epigenetics, human tissue, 5hmC, tsDhMRs

Posted Date: July 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-39144/v1

Version of Record: A version of this preprint was published at Nature Communications on July 12th, 2021. See the published version at https://doi.org/10.1038/s41467-021-24425-w.

Tissue-specific 5-hydroxymethylcytosine landscape of the human genome

Bo He,[1,2,10] Chao Zhang,[3,5,10] Xiaoxue Zhang,[1,2,10] Yu Fan,[6,7,8,10] Hu Zeng,[1] Jun'e Liu,[1] Haowei Meng,[1] Dongsheng Bai,[1] Jinying Peng,[1] Qian Zhang,[6,7,8,9,*] Wei Tao,[3,*] and Chengqi Yi[1,2,4*]

1 State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China

2 Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

3 Key Laboratory of Cell Proliferation and Differentiation, School of Life Sciences, Peking University, Beijing, 100871, China

4 Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

5 Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program (PTN), Peking University, Beijing 100871, China

6 Department of Urology, Peking University First Hospital, Beijing 100034, China

7 Institute of Urology, Peking University, Beijing 100034, China

8 National Urological Cancer Center, Beijing 100034, China

9 Peking University Binhai Hospital, Tianjian 300450, China

10 These authors contributed equally

ABSTRACT

5-Hydroxymethylcytosine (5hmC) is an important epigenetic mark that regulates gene expression. Charting the landscape of 5hmC in human tissues is fundamental to understand its regulatory functions. Here, we systematically profiled the whole-genome 5hmC landscape at single-base resolution for 19 types of human tissues. We found that 5hmC preferentially decorates gene bodies and outperforms gene body 5mC in reflecting gene expression. Approximately one-third of 5hmC peaks are tissue-specific differentially hydroxymethylated regions (tsDhMRs), which are deposited in regulatory elements that regulate the expression of nearby tissue-specific functional genes. In addition, tsDhMRs are enriched with tissue-specific transcription-factor-binding sites and may rewire tissue-specific gene expression networks. Moreover, tsDhMRs are associated with SNPs identified by genome-wide association study (GWAS), linked to tissue-specific phenotypes and diseases. Collectively, our results show the tissue-specific 5hmC landscape of the human genome and demonstrate that 5hmC serves as a fundamental regulatory element affecting tissue-specific development and diseases.

Introduction

DNA methylation at the fifth position of cytosine (5mC), which is established and maintained by DNA methyltransferases (DNMTs), is a predominant epigenetic modification that is critical for various biological and pathological processes, including silencing of transposable elements, regulation of gene expression, genomic imprinting and X-chromosome inactivation[1,2]. 5-Hydroxymethylcytosine (5hmC), also known as the "sixth base" of DNA, was discovered as another relatively abundant form of cytosine modification in Purkinje neurons and mouse embryonic stem cells (mESCs)[3,4]. Further studies found that the ten-eleven translocation (TET) family of Fe(II)- and α-ketoglutarate (α-KG)-dependent DNA dioxygenases (including TET1, TET2, and TET3) catalyzes the sequential oxidation of 5mC to 5hmC, 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[4-6]. Subsequently, the DNA repair enzyme thymine-DNA glycosylase (TDG) can excise 5fC and 5caC to generate abasic sites, which eventually results in the regeneration of unmodified cytosines by base excision repair (BER) pathway[7-9]. This TET-TDG pathway is known as the active DNA demethylation pathway.

In addition to being an intermediate of 5mC oxidation, evidence shows that 5hmC is a stable epigenetic mark with regulatory functions[10]. The specific genomic distribution pattern of 5hmC, such as its high enrichment in the gene bodies of transcriptionally active genes, promoters and enhancers, hints at specific biological roles of 5hmC[11-13]. In addition, the 5hmC level undergoes highly dynamic changes during development, differentiation and cancer[4,14,15]. For instance, the global 5hmC

content is dramatically reduced in multiple human cancers compared with that in the normal tissues adjacent to the cancer[16-18], suggesting that dysregulation of genomic 5hmC may be involved in tumorigenesis. Mechanistically, abnormal hydroxymethylation status impacts chromatin structure by interrupting the interaction of 5hmC-specific binding proteins with 5hmC[19]. Together, existing data have suggested a critical role of 5hmC in developmental processes and an association of dysregulation of 5hmC with human diseases.

Different methods have been developed to map the genomic distribution of 5hmC, including affinity-based methods (5hmC-DIP-Seq, CMS-Seq, GLIB and hMe-Seal) and high-resolution methods (oxBS-Seq, TAB-Seq, hmC-CATCH, TAPS and ACE-Seq)[20,21]. For instance, we previously developed hmC-CATCH[22], which is a bisulfite-free method for genome-wide detection of 5hmC. hmC-CATCH couples selective chemical labeling and biotin pulldown[22-25], thus allowing 5hmC enrichment and detection at single-base resolution. While hydroxymethylome maps have been obtained for mammalian cell lines and mouse tissues, 5hmC profiles in human tissues are poorly characterized thus far.

In this study, we generated genome-wide, base-resolution 5hmC data via hmC-CATCH across 19 human tissue types, represented by 60 tissue samples derived from 6 Chinese donors. We found that 5hmC is enriched in gene bodies and that gene body 5hmC exhibits a better positive correlation with gene expression than gene body 5mC. Tissue-specific differentially hydroxymethylated regions (tsDhMRs) may serve as regulatory elements and colocalize with transcription-factor-binding sites to regulate

the tissue-specific gene expression program. Furthermore, tsDhMRs were shown to enrich disease-related GWAS SNPs, linking 5hmC to human phenotypes and pathologies. Collectively, our results provide a high-resolution, high-quality atlas of DNA hydroxymethylation across diverse human tissues and provide a resource for exploring the role of this modification in development and human diseases.

Results

5-Hydroxymethylcytosine landscape of diverse human tissues

To systematically investigate the dynamics of DNA hydroxymethylation across different human tissues, we collected samples of 19 tissue types from 6 Chinese donors, including 3 males and 3 females (Fig. 1a, Supplementary Table 1). We profiled the 5hmC signals of various human tissues via hmC-CATCH (60 hmC-CATCH samples; 83-183 million paired-end reads per sample; average, 128 million paired-end reads) (Supplementary Table 2). We found that the 5hmC-containing spike-in sequence was specifically and efficiently enriched (Extended Data Fig. 1a) and displayed a high detection rate for 5hmC (~91% and ~98% before and after pull down) (Extended Data Fig. 1b, c, Supplementary Table 3), both of which showed the high quality of the hydroxymethylome generated by hmC-CATCH. In addition, tissues derived from the same organ system from different donors were clearly clustered together (Fig 1b, Extended Data Fig. 1d), further demonstrating the confidence of the 5hmC data.

We next analyzed the genomic features of 5hmC. We identified 721,404 reproducible 5hmC peaks, most of which were located in intron and intergenic regions (Fig. 1c). The peak number of each tissue type ranged from 272,391 to 525,080 (Extended Data Fig. 1e). We found that 5hmC is highly enriched in gene bodies, especially in exon regions, while it is depleted in intergenic regions (Extended Data Fig. 1f). Taking the HOX gene clusters as examples, we observed 5hmC signals in distal regulatory elements, promoters and gene bodies; such signals can also be dynamic among tissues (Fig. 1d, Extended Data Fig. 2). Hence, via hmC-CATCH, we were able to generate the whole-genome 5hmC landscape of different human tissues.

Hydroxymethylome at single-base resolution

To obtain a more detailed and clearer picture of the hydroxymethylome, we analyzed 5hmC sites at single-base resolution. We identified 9,416,937 reproducible 5hmC sites, with the site number of each tissue type ranging from 1,217,850 to 2,429,878 (approximately 3%-10% CpG sites were hydroxymethylated) (Extended Data Fig. 3a). More than half of the 5hmC sites were found in at least two tissues, suggesting that a substantial number of 5hmC sites could be conversed (Extended Data Fig. 3b). Most of the 5hmC sites were located in intron and intergenic regions (Fig. 2a) and were highly enriched in gene bodies (Extended Data Fig. 3c). The number of 5hmC sites in the brain was significantly higher than that in other tissues (p value: 6.57×10^{-7}), consistent with the higher 5hmC content in the brain[21].

5mC in the CG context (5mCG) is almost symmetric due to the maintenance of DNA methyltransferase 1 (DNMT1). A total of 98.46% of the 5hmC sites identified are located in the CG context (Fig. 2b), but only approximately 13% of the 5hmCG sites are symmetric (Extended Data Fig. 3d). In addition, the proportion of 5hmC in the non-CG context (5hmCH) is variable in different tissues, ranging from 0.96% to 2.79% (Fig. 2c, Extended Data Fig. 3e). Previous research suggests that DNA methylation is rapidly accumulated in 5mCH sites during synaptogenesis[26], and we also found a higher 5hmCH ratio in the brain than in other tissues (Fig. 2c, Extended Data Fig. 3e), indicating that 5hmCH may play a role during brain development.

We next analyzed the 5hmC context in tissues and found a "CA^{hm}CGT" motif for 5hmCG (Fig. 2d, Extended Data Fig. 3f, g), which coincides with the binding site of ARNT (Fig. 2e), a housekeeping gene that participates in important metabolic processes. Within ARNT ChIP-seq signals, cytosines are hydroxymethylated in most tissues (Fig. 2f), further indicating that 5hmC may positively impact the genomic occupancy of ARNT. Consistent with this, it has been reported that the ARNT binding motif loses 5hmC signals in esophageal cancer patients compared to healthy individuals[27]. With regard to the 5hmCH modification, we found that the most frequent base following 5hmCH is adenine (Fig. 2g, h, Extended Data Fig. 3h, i), consistent with the 5mCH sequence preference[28].

Taken together, the base-resolution hydroxymethylome analysis results presented above revealed the varied 5hmCH ratio among tissues, the asymmetric distribution of

5hmCG and the sequence preference of 5hmC, providing a potential mechanism for how 5hmC can be modified, recognized and dynamically regulated.

Gene body 5hmC excels 5mC in reflecting gene expression

Because the 5hmC level in the gene body shows a positive correlation with gene expression in mECSs and the mouse brain[19], we analyzed whether the gene body 5hmC level in human tissues also corresponds to gene expression. Indeed, we found that stronger gene body 5hmC signals were correlated with higher gene expression levels (Fig. 3a, Extended Data Fig. 4a). For instance, the genes that escape X chromosome inactivation display higher gene body 5hmC levels than the inactive genes (Extended Data Fig. 4b, c). Because the gene body 5mC level is reported to be positively correlated with transcription (Fig. 3b, c)[1], we also compared the gene body 5mC and 5hmC levels. Although 5mC colocalizes with 5hmC in all gene bodies (Fig. 3d), we found that the gene body 5hmC level exhibits a stronger quantitative correlation with gene expression levels than the gene body 5mC level.

To further explore the role of the 5hmC gene body in controlling the expression of tissue-specific genes, we defined 4,031 such genes using RNA expression data generated by the Genotype-Tissue Expression project (GTEx)[29] (Extended Data Fig. 4d). The expression of tissue-specific genes and gene body 5hmC levels also demonstrated a positive correlation, which was absent for 5mC (Extended Data Fig.

4d-g). Correlations of representative tissue-specific marker genes are shown in Fig. 3e-g. For instance, *PGA4* is specifically expressed in the stomach and encodes the precursor of pepsin; we found the highest level of gene body 5hmC signals in the stomach among all tissues (Fig. 3e). In contrast, gene body 5mC failed to display a positive correlation with tissue-specific gene expression (Fig. 3f). Using *CYP4A11* as another example showed the following: this gene is specifically expressed in the liver and is involved in drug metabolism and the synthesis of cholesterol; consistent with this function, we also found higher gene body 5hmC levels in the liver than in other tissues (Fig. 3h). Again, this pattern was not observed for 5mC (Extended Data Fig. 4h). Collectively, our data show that the gene body 5hmC level correlates well with gene expression and could play an important role in maintaining tissue-specific functions.

tsDhMRs function as tissue-specific regulatory elements

5hmC is dynamic during development and differentiation[4,14,15,30]; nevertheless, the dynamics of 5hmC in different tissues have not been reported. Based on our hydroxymethylome of human tissues, we identified 33.31% (240,269 out of 721,404 peaks tested) of the 5hmC peaks as being differentially hyperhydroxymethylated in different tissues, which we term tissue-specific differentially hydroxymethylated regions (tsDhMRs) (Fig. 4a, See Methods). A majority of the tsDhMRs were located in intron and intergenic regions (56.0% and 31.5%, respectively, Fig. 4b, Extended Data Fig. 5a). We further analyzed the histone marks using data from the ENCODE project[31]

and found that tsDhMRs are showed high enrichment of histone modification signals, including H3K4me1 and H3K27ac, in the corresponding tissues (Fig. 4c). Thus, tsDhMRs may predominantly function as regulatory elements. In addition, tsDhMRs show higher evolutionary conservation than random regions (Extended Data Fig. 5b), suggesting the importance of such functional elements.

We then investigated how these tsDhMRs may regulate the gene expression programs of specific tissues. We adopted an existing method to identify putative regulatory element-gene linkages occurring within a 500-kb window[32]. In total, we identified 4,154 genes whose expression showed significant correlations with 5hmC signals in tsDhMRs (Fig. 4d). These are genes with tissue-specific expression, such as *CYP11B2* in adrenal glands and *NPPB* in heart. Gene ontology (GO) analysis demonstrated that these genes are enriched in tissue-specific functions, such as learning or memory (brain), kidney epithelium development (kidney), female gonad development (ovary) and sex differentiation (uterus) (Fig. 4e). Similarly, KEGG pathway analysis revealed that tsDhMR-associated genes participate in tissue-specific functional pathways (Extended Data Fig. 5c). *CYP2C8*, which is involved in the metabolism of xenobiotics, is highly expressed in liver, with sharp and specific 5hmC peaks present upstream of *CYP2C8* in liver (Fig. 4f). Meanwhile, *PTF1A* plays a vital role in mammalian pancreatic development and displays increased 5hmC levels around the gene (Extended Data Fig. 5d). In summary, tsDhMRs act as regulatory elements specifically affecting the expression of nearby functional genes.

Tissue-specific transcription factors are enriched in tsDhMRs

We next analyzed, within tsDhMRs, potential transcription-factor-binding site (TFBS) clusters profiled by the ENCODE project. We found that more than half of the tsDhMRs overlapped with at least one TFBS, and 20.3% of the tsDhMRs even overlapped with no less than four TFBSs (Fig. 5a). We further investigated different transcription factors (TFs) enriched in tsDhMRs derived from each tissue type and found that tissue-specific TFs were significantly enriched in the corresponding tsDhMRs (Fig. 5b, Extended Data Fig. 6a). Taking NEUROD1 and HNF4A as examples, NEUROD1 plays an important role in the regulation of the differentiation process of various nervous system cells during development, while HNF4A regulates the expression of several hepatic genes. By comparing to ENCODE ChIP-seq data, we found that NEUROD1 specifically binds to brain-specific DhMRs, while HNF4A specifically binds to liver-specific DhMRs (Fig. 5c, Extended Data Fig. 6b). We further used *LINGO1* and *CYP2C18*, which are two downstream targets of NEUROD1 and HNF4A in the brain and liver, respectively, to demonstrate the tissue-specific coexistence of 5hmC signals (Fig. 5d). Thus, key tissue-specific TFs may rewire their tissue-specific network through binding to tsDhMRs.

To further understand the regulatory functions, we constructed regulatory networks mediated by HNF4A and NEUROD1 with their corresponding tsDhMRs. We found that HNF4A can regulate genes with tissue-specific expression, such as *CYP4F2*,

CYP8B1, and *UGT1A9*, via a process potentially mediated by liver-specific DhMRs (Fig. 5e). Within the NEUROD1 network, brain-specific DhMRs regulate *GRIK3, CSMD2* and other genes to perform their brain-specific functions (Fig. 5f). Through network analysis, we illustrated that the key TFs interact with tsDhMRs, which may further affect the expression of tissue-specific functional genes.

GWAS SNPs prefer to locate within tsDhMRs

We next analyzed the potential relationship of tsDhMRs with functional single-nucleotide polymorphisms (SNPs). We found that the tsDhMRs derived from each tissue highly overlapped with the GTEx single-tissue eQTL SNPs (Fig. 6a). In fact, tsDhMRs contain SNPs that are functional in the corresponding tissues. Moreover, tsDhMRs are also enriched for GWAS SNPs[33] with phenotypes related to the corresponding tissue functions (Extended Data Fig. 6c), indicating that tsDhMRs contribute to tissue-related diseases (Fig. 6b). We used several examples to elaborate the findings: (1) SNPs related to electrocardiographic traits and QRS duration are highly enriched in heart-specific DhMRs; (2) SNPs related to metabolite levels and LDL cholesterol are enriched in liver-specific DhMRs; and (3) SNPs related to type 2 diabetes are enriched in pancreas- and adrenal gland-specific DhMRs.

More specifically, we used an example to illustrate the potential mechanism by which distal GWAS SNPs may impact diseases. HCN4, which is necessary for the

cardiac pacemaking process, is specifically expressed in the heart (Fig. 6c). We found that several heart-related GWAS SNPs were localized in the heart-specific DhMR (Fig. 6c), indicating that the DhMR is associated with heart diseases. Moreover, several enhancers of the heart identified by ENCODE candidate cis-regulatory elements (cCREs) were located in this DhMR (Fig. 6c). We further integrated the VISTA enhancer data[34], which were validated by transgenic mouse assays, and found an enhancer (enhancer element 2161 from VISTA) within the DhMR. This enhancer is specifically expressed in the mouse heart (images from VISTA database) (Fig. 6d). These data confirm that DhMR, as an enhancer, is functional in the heart and is related to heart diseases. These results indicate that dysregulation of tsDhMRs may be involved in human disease pathologies. Collectively, our data show that tsDhMRs could help us understand the function of distal GWAS SNPs in the corresponding tissues.

Discussion

In this study, we present a base-resolution atlas of 5hmC in human tissues. Hundreds of thousands of 5hmC peaks and millions of 5hmC sites were identified in this dataset, expanding the epigenomic landscape determined by previous large-scale efforts, for example, the ENCODE project.

Gene body 5hmC levels are positively correlated with gene expression, especially of genes with tissue-specific expression. Thus, the gene body 5hmC levels may be used to infer gene expression in tissues. This can be particularly useful in some precious clinical samples, where RNA degradation is severe (frozen samples, formalin-fixed

paraffin-embedded samples, body fluids and so on). For instance, 5hmC levels in cell-free DNA (cfDNA) have been utilized as biomarkers for cancer diagnosis[22,35-37]. While it is anticipated that healthy and cancerous tissues may have different hydroxymethylomes, cancer-specific 5hmC signatures, which may reflect the gene expression program in different cancers, could be identified for noninvasive cancer diagnosis. Thus, the relative stability of epigenetic modifications makes 5hmC a promising candidate for prediction of gene expression in clinical samples.

Using the hydroxymethylome of multiple tissues, we discovered that approximately one-third of all 5hmC peaks are tissue-specifically hyperhydroxymethylated. We found that tsDhMRs, as cis-regulatory elements, are positively correlated with gene expression, which is in contrast to the fact that differentially methylated regions (DMRs) show a negative correlation[38]. Moreover, tissue-specific TFs are enriched in tsDhMRs, providing a mechanism by which key TFs may regulate tissue-specific gene expression via tsDhMRs. Based on our identified tsDhMRs, future studies could be designed to illustrate the specific mechanisms by which 5hmC can regulate tissue development and differentiation.

Through integration of GWAS SNP and 5hmC data, we discovered that tsDhMRs were significantly enriched with GWAS SNPs. Our data indicate that tsDhMRs, as cis-regulatory elements, contribute to tissue-related diseases. Although GWAS SNPs are consistent among somatic cells, they may affect tsDhMR 5hmC levels and ultimately lead to dysfunctions of the corresponding tissues. Our analysis provides new insights into the understanding of GWAS data, where distal GWAS SNPs interact with tsDhMRs

to regulate target genes. Disruption of tsDhMR 5hmC levels may result in tissue-related disease phenotypes.

Collectively, our data provide a rich resource for understanding the 5hmC landscape in human tissues. The reported human tissue hydroxymethylome adds to the knowledge of how this epigenetic mark may affect tissue-specific differentiation and diseases.

ACKNOWLEDGMENT

The authors would like to thank Jiabin Cai and Jia Fan for proving tissue samples; Xushen Xiong for bioinformatics assistance; and National Center for Protein Sciences at Peking University in Beijing, China, for assistance with mass spectrometry analysis. Part of the analysis was performed on the High Performance Computing Platform of the Center for Life Science. This work was supported by National Key R&D Program (no. 2019YFA0110900 to C.Y. and no. 2020YFC2002900 to W.T.), the National Natural Science Foundation of China (nos. 21825701 and 91953201 to C.Y.), and Peking University Ge Li and Ning Zhao Education Fund.

Author Contributions

C.Y., W.T. and Q.Z. designed the study. 5fC labeling chemical compounds are synthesized by B.D. X.Z., B.H., Y.F. and H.Z. performed experiments. C.Z. and B.H. analyze the high-throughput sequencing data with help from H.M. B.H., C.Z., X.Z., W.T. and C.Y. wrote the manuscript.

Competing Interests Statement

B.X., A.Z., C.Z. and C.Y. are co-inventors on filed patents (PCT/CN2014/087479 and 201710111600.9) for the labeling strategies of 5fC and 5hmC.

References

1 Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590-607,(2019).

2 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204-220,(2013).

3 Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science* **324**,(2009).

4 Tahiliani, M. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**,(2009).

5 Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303,(2011).

6 He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307,(2011).

7 Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**,(2011).

8 Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706,(2013).

9 He, Y. F. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**,(2011).

10 Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable

DNA modification. *Nat Chem* **6**, 1049-1055,(2014).

11 Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402,(2011).

12 Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68-72,(2011).

13 Schutsky, E. K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat Biotechnol*,(2018).

14 Song, C. X. *et al.* Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Methods* **9**, 75-77,(2011).

15 Haffner, M. C. *et al.* Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. *Oncotarget* **2**, 627-637,(2011).

16 Pfeifer, G. P., Xiong, W., Hahn, M. A. & Jin, S. G. The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res* **356**, 631-641,(2014).

17 Ficz, G. & Gribben, J. G. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? *Genomics* **104**, 352-357,(2014).

18 Lian, C. G. *et al.* Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* **150**, 1135-1146,(2012).

19 Mellen, M. *et al.* MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417-1430,(2012).

20 Zeng, H., He, B. & Yi, C. Compilation of Modern Technologies To Map

Genome-Wide Cytosine Modifications in DNA. *Chembiochem* **20**, 1898-1905,(2019).

21 Wu, H. & Zhang, Y. Charting oxidized methylcytosines at base resolution. *Nature structural & molecular biology* **22**, 656-661,(2015).

22 Zeng, H. *et al.* Bisulfite-Free, Nanoscale Analysis of 5-Hydroxymethylcytosine at Single Base Resolution. *J Am Chem Soc* **140**, 13190-13194,(2018).

23 Zeng, H. *et al.* Unnatural Cytosine Bases Recognized as Thymines by DNA Polymerases by the Formation of the Watson-Crick Geometry. *Angew Chem Int Ed Engl* **58**, 130-133,(2019).

24 Zhu, C. *et al.* Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. *Cell Stem Cell* **20**, 720-731.e725,(2017).

25 Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nature Methods* **12**, 1047-1050,(2015).

26 Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905,(2013).

27 Tian, X. *et al.* Circulating tumor DNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer. *Cell Res* **28**, 597-600,(2018).

28 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322,(2009).

29 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat*

Genet **45**, 580-585,(2013).

30 Wang, L. *et al.* Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell* **157**, 979-991,(2014).

31 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74,(2012).

32 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**,(2018).

33 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012,(2019).

34 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92,(2007).

35 Song, C. X. *et al.* 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res* **27**, 1231-1242,(2017).

36 Mellen, M., Ayata, P. & Heintz, N. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc Natl Acad Sci U S A* **114**, E7812-E7821,(2017).

37 Li, W. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res* **27**, 1243-1257,(2017).

38 Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212-216,(2015).

39 Peng, X. *et al.* TELP, a sensitive and versatile library construction method for next-generation sequencing. *Nucleic Acids Res* **43**, e35,(2015).

40 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606,(2014).

Methods

Biospecimen Collection

We collected a total of 60 tissue samples at Zhongshan Hospital (Shanghai, China) after obtaining research consent from the families; we targeted 19 distinct tissues from 6 postmortem Chinese individuals, including 3 males and 3 females. The collected samples were transferred to cryovials for long-term storage at -80℃.

Genomic DNA library construction and sequencing

Genomic DNA was isolated from tissues using the Blood/Cell/Tissue Genomic DNA Extraction Kit (TIANGEN, DP304) following the manufacturer's specifications. The 5hmC sequencing libraries were constructed with 200 ng of genomic DNA through the hmC-CATCH method reported previously[22], which is a bisulfite-free, base-resolution method for genome-wide detection of 5hmC. Briefly, the DNA was fragmented with a Covaris device (Gene Company Limited, ME220) to 300-500 bp, followed by end repair and addition of the 3'dA nucleotide. Endogenous 5fC was blocked by hydroxylamine, and endogenous 5hmC was oxidized to 5fC using K_2RuO_4. The newly generated 5fC was labelled with 5-(2-azidoethyl)-1,3-indandione (AI) (J&K, 2793948), and the AI-labeled ssDNA was used directly for library preparation with the TELP protocol[39]. It is worth mentioning that we designed a homemade adapter with a unique molecular identifier (UMI) for deduplication[40]. Subsequently, click chemistry was performed by adding DBCO-S-S-PEG3-Biotin (Click Chemistry Tools, A112-10), followed by pull-down for enrichment of 5hmC-containing DNA. Finally, we obtained

libraries by performing PCRs on subsets of 5hmC-containing DNA.

Data processing

Illumina sequencing adapters and low-quality reads were removed from raw sequencing data to obtain clean data. We added a sequencing barcode through the hmC-CATCH protocol to mark PCR duplicated reads. Then, the PCR-duplicated reads were filtered out, and only one read was retained using an in-house script. The final cleaned reads were mapped to hg38 by Bismark (Version: v0.15.0). To enhance the signal-to-noise ratio, we used only read pairs with more than one C-to-T conversion for further analysis.

Assessing the C-to-T conversion rate of 5hmC sites

We added a model sequence as a spike-in before constructing the hmC-CATCH library, of which one C site was 100% hydroxymethylated. After treatment through the hmC-CATCH protocol, we observed a C-to-T conversion signal in this 5hmC site. We used the T/(C+T) of the sequencing data at this site to estimate the C-to-T conversion rate.

Identification of 5hmC sites

We used Bismark (Version: v0.15.0) to extract single-base-resolution information. Sites with less than five total bases (NT+NC) or three NTs were discarded for 5hmC calling. Then, we used the binomial distribution with N as the sequencing depth (NC+NT) and p as the normal cytosine conversion rate to assess the probability of observing NT by

chance. We considered 5hmC sites with the Holm-Bonferroni method-adjusted P < 0.001 and located within a type of tissue 5hmC-enriched region as high-confidence 5hmC sites.

Identification of 5hmC peaks

Our hmC-CATCH approach could enrich DNA fragments with 5hmC. Here, we used a peak calling method to identify these regions with 5hmC. MASC2 was applied to call peaks in each sample with the following command:

"macs2 callpeak -t <5hmC bam> -c <input bam> -g hs -f BAMPE --keep-dup all --outdir <outdir> -n <sample name>"

Annotation of 5hmC sites and peaks

The 5hmC sites or peaks were annotated by annotatePeaks.pl (Homer, Version: v4.5), and the "-annStats" parameter was added to quantify the enrichment of genomic elements compared to the background. Then, the log2 ratios of observation to expectation in all tissues were plotted as a bar chart by ggplot2 (R package).

tSNE cluster of the global 5hmC signals

The whole hg38 genome was first cut into 10-kb nonoverlapping bins. Then, the read counts in each 10-kb bin of all samples were calculated by "Bedtools multicov" (Version: v2.27.1). After normalizing the sequencing depth (DESeq2, R package) and batch effect (limma, R package), we performed tSNE clustering (Rtsne, R package) to reduce the high-dimensional data to two dimensions. ggplot2 (R package) was used to

visualize the data.

Analysis of the correlation of gene body 5hmC and 5mC

The 5mC data of human tissues were downloaded from the ENCODE project. Then, the RPKM values of all protein-coding genes were calculated and normalized as mentioned above. Spearman's correlation coefficients of the RPKM values between 5hmC and 5mC in matched tissues were calculated by R (cor, method="spearman").

Identification of genes with tissue-specific expression

We used GTEx gene expression data (RNASeQCv1.1.9_gene_median_tpm) to identify the genes with tissue-specific expression that were defined as being highly expressed in one tissue. We first filtered out the genes with a mean TPM less than 1, which was regarded as low expression in all tissues. Then, we calculated the fold changes in the expression of protein-coding genes in one tissue over the mean values for other tissues. We ordered the genes according to fold change levels, and the top 300 in each tissue were regarded as genes with tissue-specific expression. To ensure that the top 300 genes with tissue-specific expression in each tissue were significantly more highly expressed than others, we filtered out the genes with fold changes lower than 2.

Identification of tissue-specific differential 5hmC regions

We first merged all 5hmC peaks from the 60 samples to obtain the total peaks using "Bedtools merge". Then, the read counts in each merged peak of all samples were

calculated by “Bedtools multicov” and normalized as mentioned above. We merged all biological replicates from the same tissue to enhance the signals. We used the Poisson distribution to estimate the p value of each peak in a tissue. The probability of read counts in each peak of one tissue was estimated by the one-tail Poisson distribution with the parameter λ as the mean for other tissues. The p values were further adjusted by the Bonferroni method. The peaks with adjusted p values less than 0.05 and fold changes more than 2 were regarded as significant tissue-specific differential 5hmC regions.

Identification of tsDhMR-associated genes

We adapted a method to link tsDhMRs to putative genes[32]. We first identified all possible genes linked to tsDhMRs by searching any TSS of GENCODE protein-coding genes within 500 kb of the tsDhMRs. Then, we calculated the Pearson correlation coefficients between normalized 5hmC signals and gene expression (TPM). To avoid spurious associations, we used cor.test (R) to assess the stochastic links. Finally, we required that the confident links had a Pearson correlation over 0.8 and p value lower than 0.05.

Motif enrichment analysis

For each tissue, we used findMotifsGenome.pl (Homer) to find the motifs enriched in tsDhMRs with the following command:

“findMotifsGenome.pl <tsDhMRs bed> hg38 <output> -p 5”

Analysis of tsDhMRs associated with GWAS SNPs

We downloaded previously published GWAS datasets from the GWAS catalog. Then, we calculated the associations of GWAS phenotypes and tsDhMRs. A GWAS phenotype always has multiple associated SNPs. For each phenotype, we used Fisher's test to compute the significance and odds ratios between the phenotype-associated SNPs and tsDhMRs.

Data availability

Sequencing data have been deposited into the Gene Expression Omnibus (GEO) under the accession number GSE134078.

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134078).

Figure legends

Figure 1. Hydroxymethylation landscape of human tissues. a. Human tissues analyzed in this study. Samples are denoted by the two-letter code in parentheses. Colors indicate each tissue type. **b.** tSNE cluster of all tissues using the global 5hmC signals under 10-kb bins. **c.** Pie chart showing the percentages of 5hmC peaks in each class of genomic features. The promoter regions are defined as 2000 bp around the TSS. **d.** IGV visualization of the 5hmC signals surrounding the HOXD gene cluster on chromosome 2. 5hmC signals in promoter, gene body or intergenic region are highlighted.

Figure 2. Single-base resolution profiles of 5hmC. a. Pie chart showing the percentages of 5hmC sites in each class of genomic features. **b.** Distribution of 5hmCG, 5hmCHH and 5hmCHG sites. **c.** Proportions of 5hmCH in different tissues. **d.** Sequence context ± 10 bp around 5hmCG sites. **e.** The most significant 5hmC motif (the upper panel) demonstrates partial sequence overlap with the known transcription factor motif of ARNT (the lower panel). **f.** IGV visualization of the 5hmC site signals. 5hmC is asymmetric at the 5hmCG site. **g.** Sequence context ± 10 bp around 5hmCHG sites. **h.** Sequence context ± 10 bp around 5hmCHH sites.

Figure 3. Gene body 5hmC correlates well with gene expression in human tissues. **a.** 5hmC profiles of genes expressed at high (yellow), low (cyan) and silenced (blue) levels in heart tissue. **b.** Spearman's correlation of gene body 5hmC (5mC) signals and

gene expression levels. Random regions were selected as controls. *** represents *P* values < 0.0001 (Wilcox test was performed to compare the difference). **c.** 5mC profiles of genes expressed at high (yellow), low (cyan) and silenced (blue) levels in heart tissue. **d.** Scatter plot showing the correlation of 5mC and 5hmC signals at gene bodies. **e.** Heatmap displaying the expression levels of tissue-specific marker genes. Gene expression of matched tissue samples was downloaded from the GTEx project. **f.** Heatmap displaying the normalized gene body 5hmC signals of tissue-specific marker genes. The order of the row is identical to (e). **g.** Heatmap displaying the normalized gene body 5mC signals of tissue-specific marker genes. The order of the row is identical to (e). The 5mC data were downloaded from ENCODE. **h.** IGV visualization of the 5hmC signals at the gene body of *CYP4A11* and surrounding regions. Gene expression levels of the gene in different tissues are shown on the right.

Figure 4. tsDhMRs are tissue-specific regulatory elements. a. Heatmap showing the normalized 5hmC signals in tsDhMRs. **b.** Pie chart showing the percentages of tsDhMRs in each class of genomic features. **c.** Profiles of H3K27ac and H3K4me1 modifications around tsDhMRs in the liver (LI) and spleen (SX). **d.** Heatmap showing the expression of genes associated with tsDhMRs within 500 kb. **e.** GO enrichment and representative genes of tsDhMR-associated genes. **f.** IGV visualization of the 5hmC signals near *CYP2C8* on chromosome 10. The highlighted regions are liver-specific DhMRs.

Figure 5. tsDhMRs are enriched for tissue-specific TFBSs. a. Overlap of tsDhMRs with ENCODE TFBSs. **b.** The TF motifs enriched in tsDhMRs of each tissue. The color scale represents the –log10(p value). **c.** ChIP-seq signals of NEUROD1 and HNF4A around BR-specific DhMRs and LI-specific DhMRs. **d.** IGV visualization of 5hmC signals and ChIP-seq signals of NEUROD1 and HNF4A in the brain and liver. **e.** Key TF regulatory networks in the liver and brain. Hexagon represents Key TF; rhombus represents tsDhMRs; circle represents regulated genes. The line linking the hexagon and rhombus indicate that the TF binds to the tsDhMRs, which is supported by TF ChIP-seq data.

Figure 6. tsDhMRs enrich tissue-specific, phenotypic GWAS SNPs. a. Overlap of tsDhMRs with single-tissue eQTL SNPs. eQTL SNP data are from GTEx. **b.** Representative GWAS phenotypes enriched in tsDhMRs. **c.** IGV visualization of the 5hmC signals around *HCN4* on chromosome 15. The locations of GWAS SNPs and VISTA enhancers are also shown. The highlighted region shows the LI-specific DhMRs. **d.** In vivo reporter assay of enhancer activity for VISTA enhancer element 2162, as obtained from the VISTA enhancer browser[34].

Figure 1

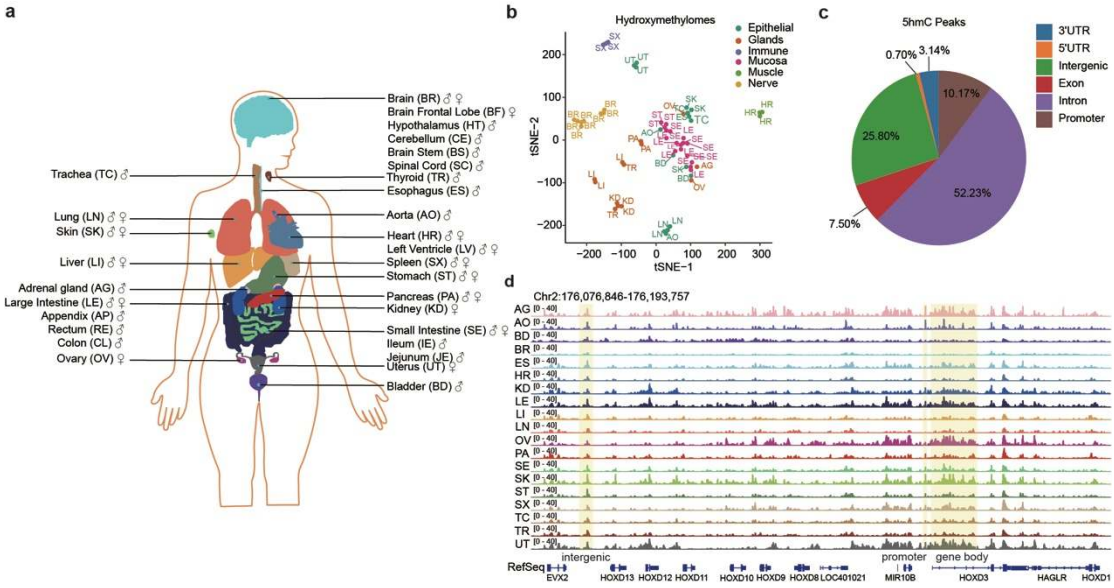

a
Brain (BR) ♂ ♀
Brain Frontal Lobe (BF) ♀
Hypothalamus (HT) ♂
Cerebellum (CE) ♂
Brain Stem (BS) ♂
Spinal Cord (SC) ♂
Trachea (TC) ♂
Thyroid (TR) ♂
Esophagus (ES) ♂
Lung (LN) ♂ ♀
Aorta (AO) ♂
Skin (SK) ♂ ♀
Heart (HR) ♂ ♀
Left Ventricle (LV) ♂ ♀
Liver (LI) ♂ ♀
Spleen (SX) ♂ ♀
Stomach (ST) ♂ ♀
Adrenal gland (AG) ♂
Pancreas (PA) ♂ ♀
Large Intestine (LE) ♂ ♀
Kidney (KD) ♀
Appendix (AP) ♂
Small Intestine (SE) ♂ ♀
Rectum (RE) ♂
Ileum (IE) ♂
Colon (CL) ♂
Jejunum (JE) ♂
Ovary (OV) ♀
Uterus (UT) ♀
Bladder (BD) ♂
b
Hydroxymethylomes
Epithelial
Glands
Immune
Mucosa
Muscle
Nerve
tSNE−2
tSNE−1
c
5hmC Peaks
3'UTR
5'UTR
Intergenic
Exon
Intron
Promoter
0.70%
3.14%
10.17%
25.80%
52.23%
7.50%
d
Chr2:176,076,846-176,193,757
AG
AO
BD
BR
ES
HR
KD
LE
LI
LN
OV
PA
SE
SK
ST
SX
TC
TR
UT
intergenic
promoter
gene body
RefSeq
EVX2
HOXD13
HOXD12
HOXD11
HOXD10
HOXD9
HOXD8
LOC401021
MIR10B
HOXD3
HAGLR
HOXD1

Figure 2

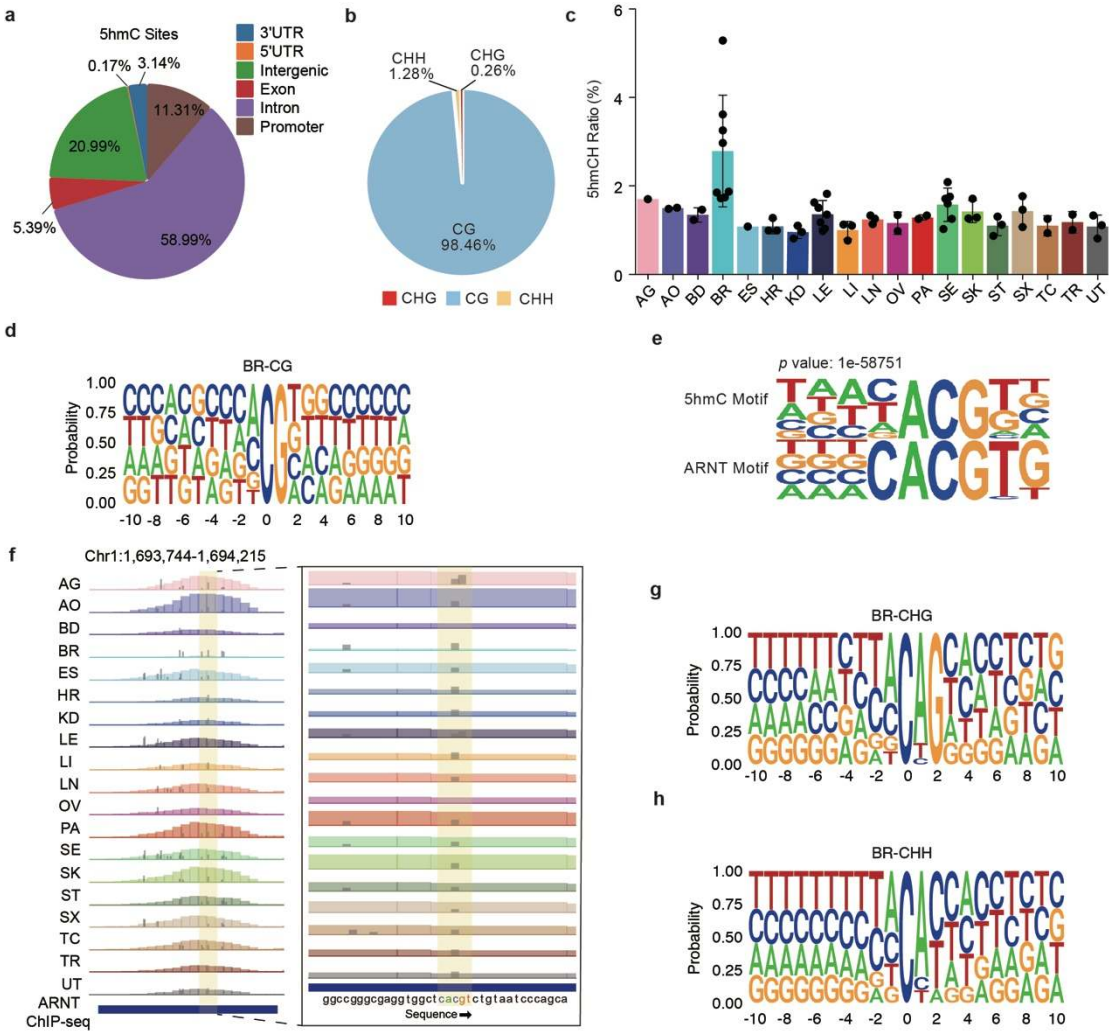

a
5hmC Sites
3'UTR
5'UTR
Intergenic
Exon
Intron
Promoter
0.17%
3.14%
11.31%
20.99%
5.39%
58.99%
b
CHH
1.28%
CHG
0.26%
CG
98.46%
CHG CG CHH
c
5hmCH Ratio (%)
AG AO BD BR ES HR KD LE LI LN OV PA SE SK ST SX TC TR UT
d
BR-CG
Probability
e
p value: 1e-58751
5hmC Motif
ARNT Motif
f
Chr1:1,693,744-1,694,215
ARNT
ChIP-seq
Sequence
g
BR-CHG
h
BR-CHH

Figure 3

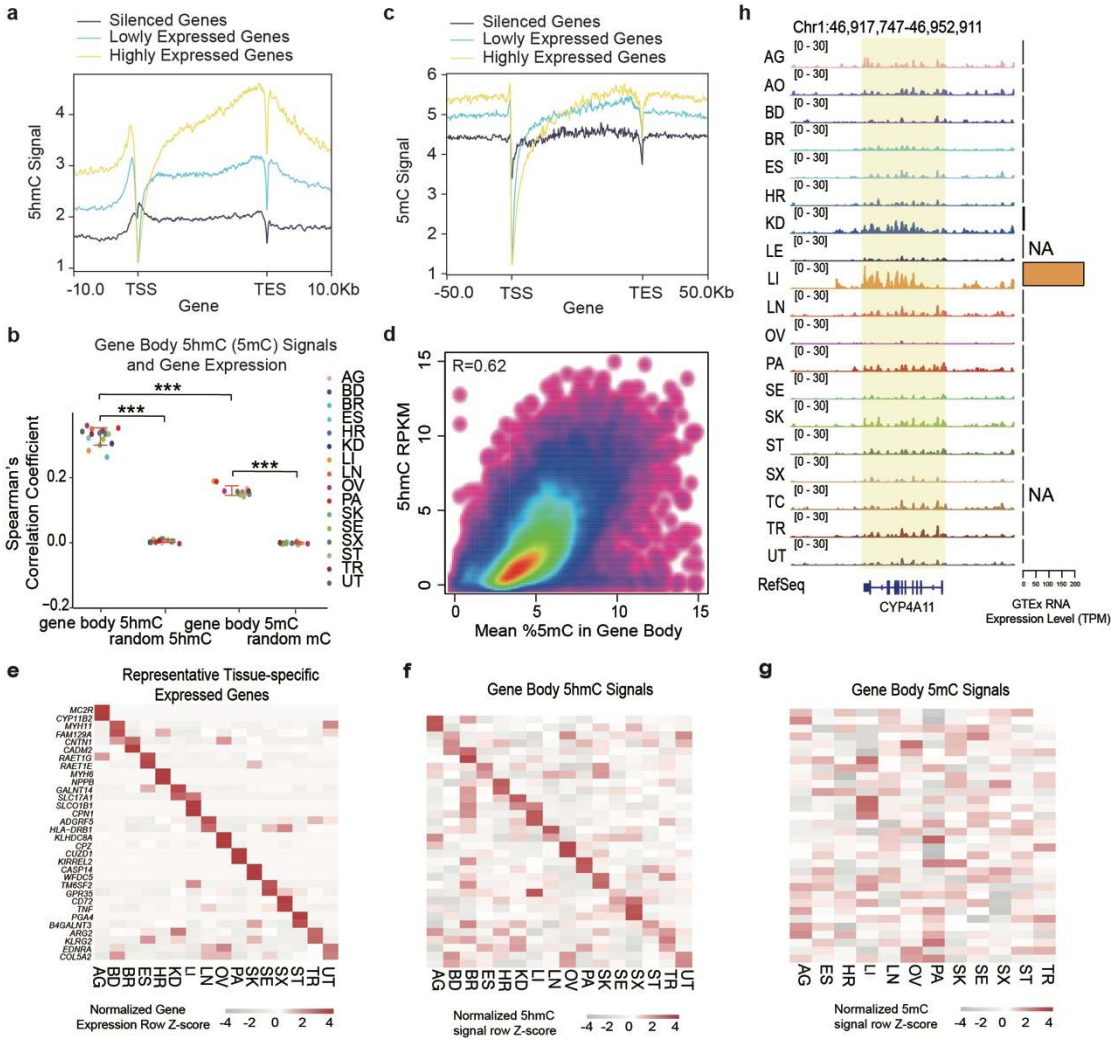

a
Silenced Genes
Lowly Expressed Genes
Highly Expressed Genes
5hmC Signal
-10.0 TSS TES 10.0Kb
Gene
c
Silenced Genes
Lowly Expressed Genes
Highly Expressed Genes
5mC Signal
-50.0 TSS TES 50.0Kb
Gene
h
Chr1:46,917,747-46,952,911
AG AO BD BR ES HR KD LE LI LN OV PA SE SK ST SX TC TR UT
[0 - 30]
NA
NA
RefSeq
CYP4A11
GTEx RNA Expression Level (TPM)
b
Gene Body 5hmC (5mC) Signals and Gene Expression
Spearman's Correlation Coefficient
***
gene body 5hmC
random 5hmC
gene body 5mC
random mC
d
R=0.62
5hmC RPKM
Mean %5mC in Gene Body
e
Representative Tissue-specific Expressed Genes
Normalized Gene Expression Row Z-score
f
Gene Body 5hmC Signals
Normalized 5hmC signal row Z-score
g
Gene Body 5mC Signals
Normalized 5mC signal row Z-score

Figure 4

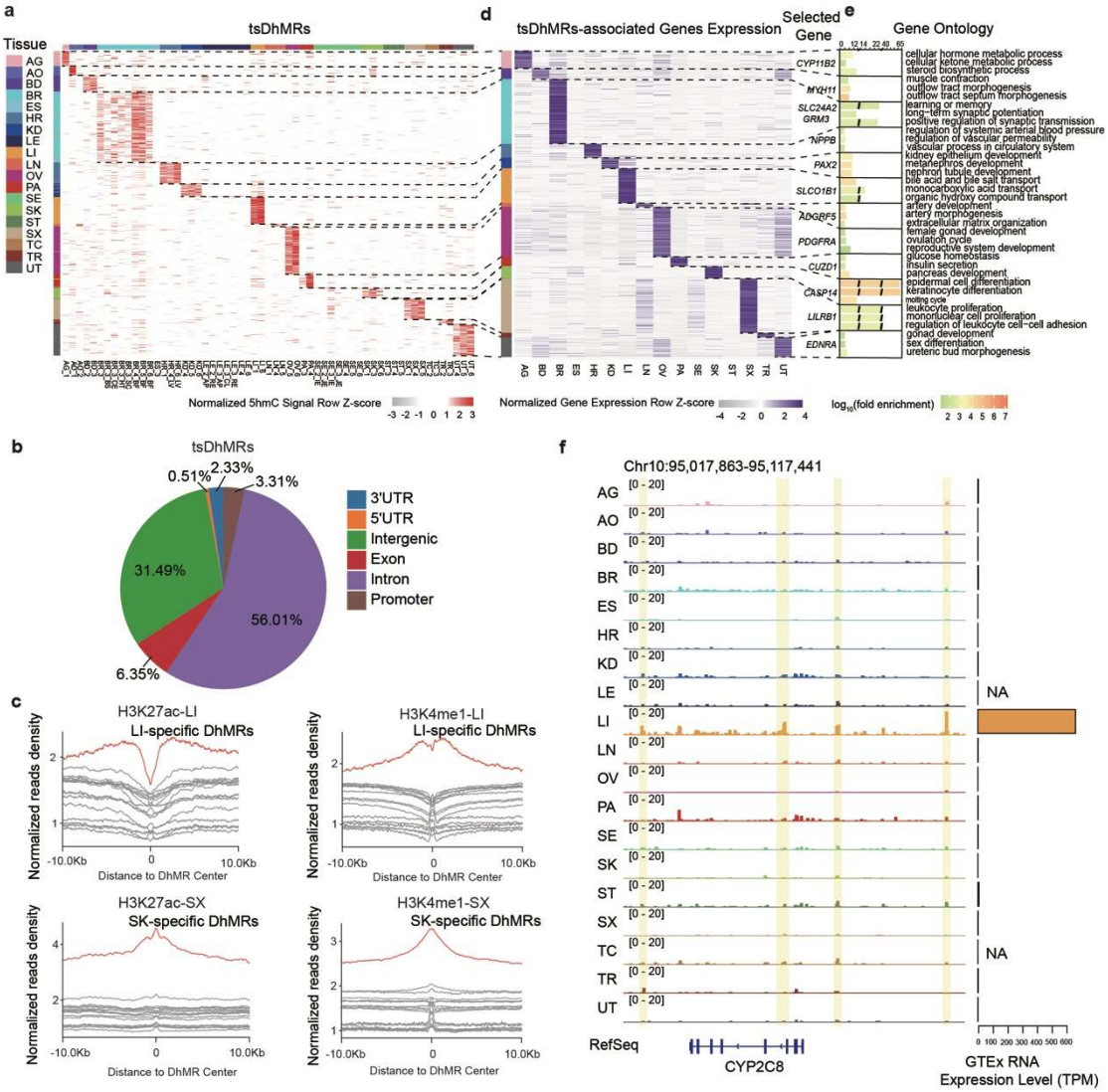

a
tsDhMRs
Tissue
AG
AO
BD
BR
ES
HR
KD
LE
LI
LN
OV
PA
SE
SK
ST
SX
TC
TR
UT
Normalized 5hmC Signal Row Z-score
-3 -2 -1 0 1 2 3
d
tsDhMRs-associated Genes Expression
Selected Gene
CYP11B2
MYH11
SLC24A2
GRM3
NPPB
PAX2
SLCO1B1
ADGRF5
PDGFRA
CUZD1
CASP14
LILRB1
EDNRA
Normalized Gene Expression Row Z-score
-4 -2 0 2 4
e
Gene Ontology
0 12.14 22.40 65
cellular hormone metabolic process
cellular ketone metabolic process
steroid biosynthetic process
muscle contraction
outflow tract morphogenesis
outflow tract septum morphogenesis
learning or memory
long-term synaptic potentiation
positive regulation of synaptic transmission
regulation of systemic arterial blood pressure
regulation of vascular permeability
vascular process in circulatory system
kidney epithelium development
metanephros development
nephron tubule development
bile acid and bile salt transport
monocarboxylic acid transport
organic hydroxy compound transport
artery development
artery morphogenesis
extracellular matrix organization
female gonad development
ovulation cycle
reproductive system development
glucose homeostasis
insulin secretion
pancreas development
epidermal cell differentiation
keratinocyte differentiation
molting cycle
leukocyte proliferation
mononuclear cell proliferation
regulation of leukocyte cell–cell adhesion
gonad development
sex differentiation
ureteric bud morphogenesis
log10(fold enrichment)
2 3 4 5 6 7
b
tsDhMRs
0.51%
2.33%
3.31%
31.49%
56.01%
6.35%
3'UTR
5'UTR
Intergenic
Exon
Intron
Promoter
c
H3K27ac-LI
LI-specific DhMRs
H3K4me1-LI
LI-specific DhMRs
H3K27ac-SX
SK-specific DhMRs
H3K4me1-SX
SK-specific DhMRs
Normalized reads density
-10.0Kb
0
10.0Kb
Distance to DhMR Center
f
Chr10:95,017,863-95,117,441
[0 - 20]
NA
NA
RefSeq
CYP2C8
0 100 200 300 400 500 600
GTEx RNA
Expression Level (TPM)

Figure 5

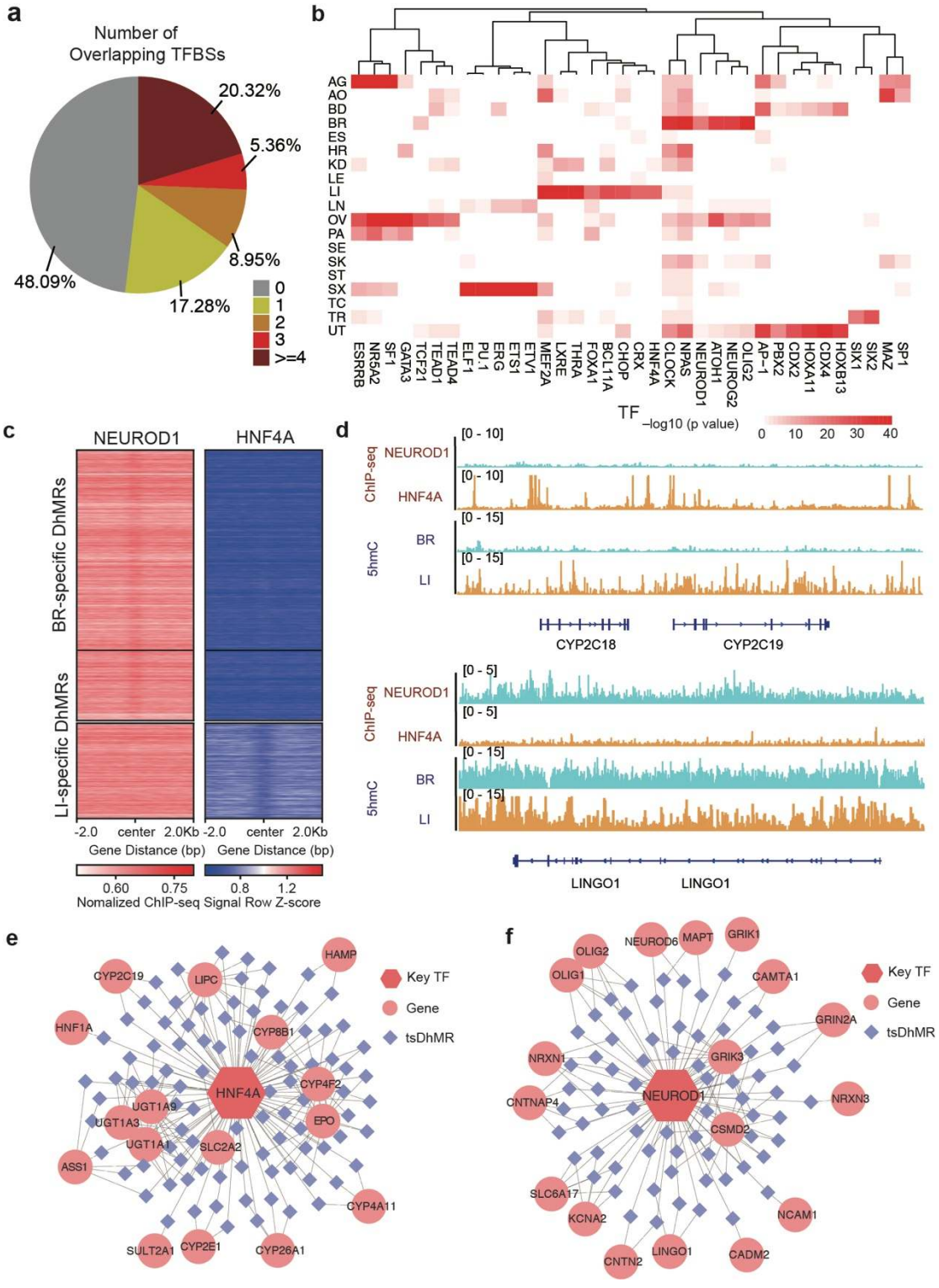
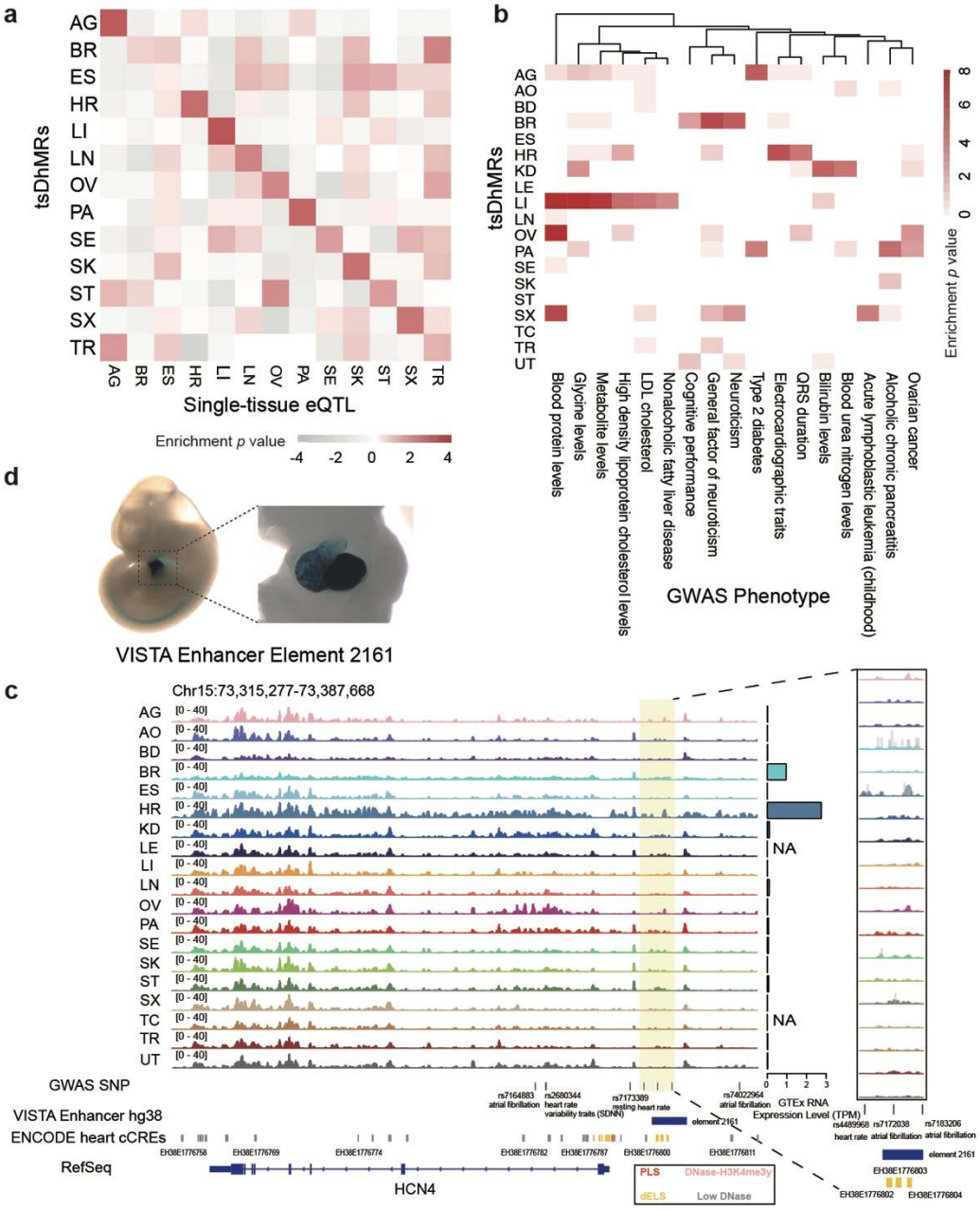

a
Number of Overlapping TFBSs
20.32%
5.36%
8.95%
17.28%
48.09%
0
1
2
3
>=4
b
AG
AO
BD
BR
ES
HR
KD
LE
LI
LN
OV
PA
SE
SK
ST
SX
TC
TR
UT
ESRRB NR5A2 SF1 GATA3 TCF21 TEAD1 TEAD4 ELF1 PU.1 ERG ETS1 ETV1 MEF2A LXRE THRA FOXA1 BCL11A CHOP CRX HNF4A CLOCK NPAS NEUROD1 ATOH1 NEUROG2 OLIG2 AP-1 PBX2 CDX2 HOXA11 CDX4 HOXB13 SIX1 SIX2 MAZ SP1
TF
−log10 (p value)
0 10 20 30 40
c
NEUROD1
HNF4A
BR-specific DhMRs
LI-specific DhMRs
-2.0 center 2.0Kb
Gene Distance (bp)
0.60 0.75
0.8 1.2
Nomalized ChIP-seq Signal Row Z-score
d
ChIP-seq
NEUROD1
HNF4A
5hmC
BR
LI
[0 - 10]
[0 - 15]
[0 - 5]
CYP2C18
CYP2C19
LINGO1
LINGO1
e
Key TF
Gene
tsDhMR
HAMP
CYP2C19
LIPC
HNF1A
CYP8B1
HNF4A
CYP4F2
UGT1A9
UGT1A3
UGT1A1
EPO
SLC2A2
ASS1
CYP4A11
SULT2A1
CYP2E1
CYP26A1
f
Key TF
Gene
tsDhMR
OLIG2
NEUROD6
MAPT
GRIK1
OLIG1
CAMTA1
GRIN2A
NRXN1
GRIK3
CNTNAP4
NEUROD1
NRXN3
CSMD2
SLC6A17
KCNA2
NCAM1
CNTN2
LINGO1
CADM2

Figure 6


a
tsDhMRs
AG
BR
ES
HR
LI
LN
OV
PA
SE
SK
ST
SX
TR
Single-tissue eQTL
Enrichment p value
-4 -2 0 2 4
b
tsDhMRs
AG AO BD BR ES HR KD LE LI LN OV PA SE SK ST SX TC TR UT
Enrichment p value
0 2 4 6 8
Blood protein levels
Glycine levels
Metabolite levels
High density lipoprotein cholesterol levels
LDL cholesterol
Nonalcoholic fatty liver disease
Cognitive performance
General factor of neuroticism
Neuroticism
Type 2 diabetes
Electrocardiographic traits
QRS duration
Bilirubin levels
Blood urea nitrogen levels
Acute lymphoblastic leukemia (childhood)
Alcoholic chronic pancreatitis
Ovarian cancer
GWAS Phenotype
d
VISTA Enhancer Element 2161
c
Chr15:73,315,277-73,387,668
AG AO BD BR ES HR KD LE LI LN OV PA SE SK ST SX TC TR UT
[0 - 40]
NA
NA
GWAS SNP
VISTA Enhancer hg38
ENCODE heart cCREs
RefSeq
HCN4
rs7164883 atrial fibrillation
rs2680344 heart rate variability traits (SDNN)
rs7173389 resting heart rate
rs74022964 atrial fibrillation
element 2161
GTEx RNA Expression Level (TPM)
rs4489968 heart rate
rs7172038 atrial fibrillation
rs7183206 atrial fibrillation
EH38E1776758
EH38E1776769
EH38E1776774
EH38E1776782
EH38E1776787
EH38E1776800
EH38E1776811
EH38E1776802
EH38E1776803
EH38E1776804
PLS
DNase-H3K4me3y
dELS
Low DNase

Figures

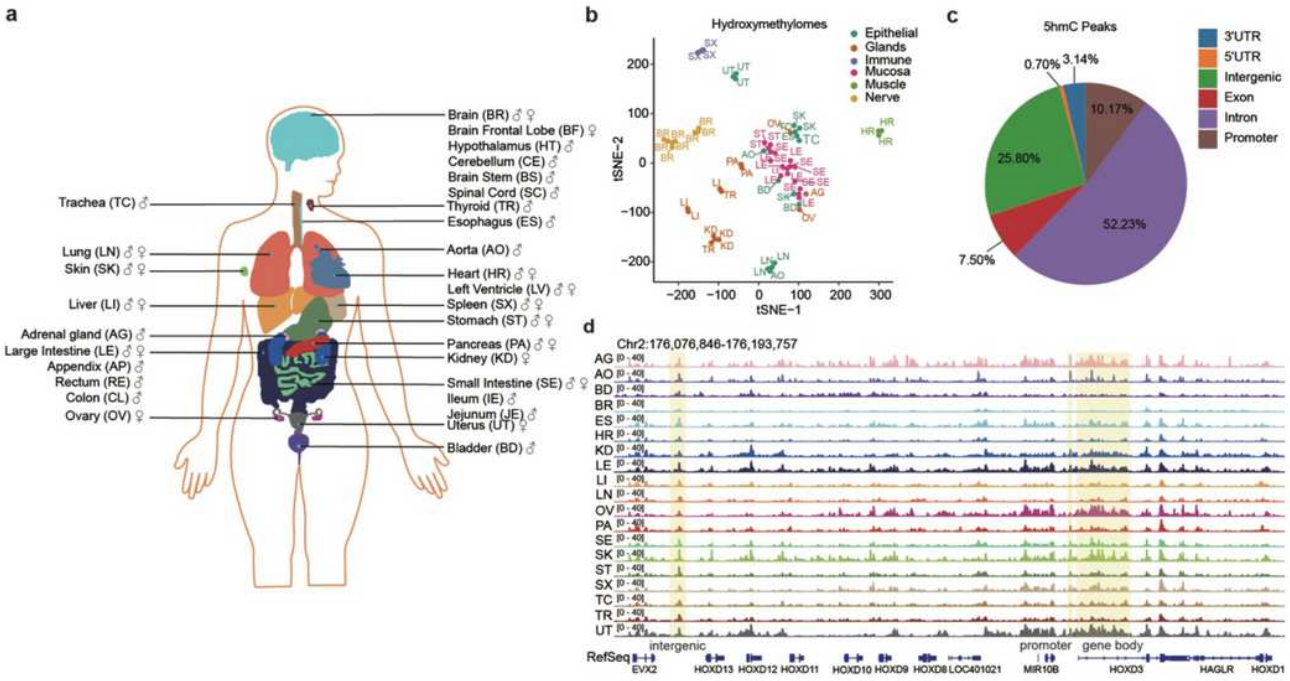


Figure 1

Hydroxymethylation landscape of human tissues. a. Human tissues analyzed in this study. Samples are denoted by the two-letter code in parentheses. Colors indicate each tissue type. b. tSNE cluster of all tissues using the global 5hmC signals under 10-kb bins. c. Pie chart showing the percentages of 5hmC peaks in each class of genomic features. The promoter regions are defined as 2000 bp around the TSS. d. IGV visualization of the 5hmC signals surrounding the HOXD gene cluster on chromosome 2. 5hmC signals in promoter, gene body or intergenic region are highlighted.

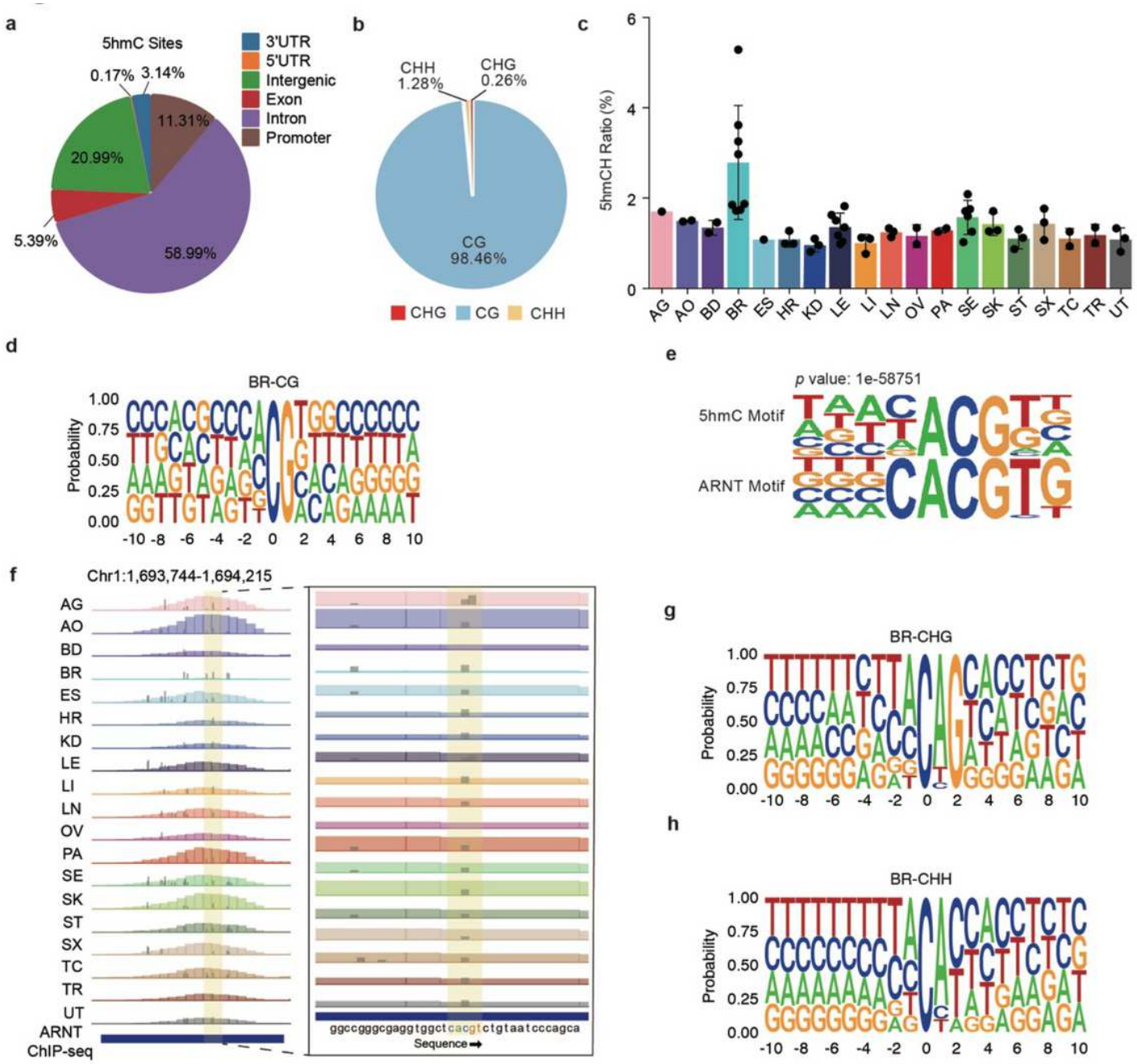


Figure 2

Single-base resolution profiles of 5hmC. a. Pie chart showing the percentages of 5hmC sites in each class of genomic features. b. Distribution of 5hmCG, 5hmCHH and 5hmCHG sites. c. Proportions of 5hmCH in different tissues. d. Sequence context ± 10 bp around 5hmCG sites. e. The most significant 5hmC motif (the upper panel) demonstrates partial sequence overlap with the known transcription factor motif of ARNT (the lower panel). f. IGV visualization of the 5hmC site signals. 5hmC is asymmetric at the 5hmCG site. g. Sequence context ± 10 bp around 5hmCHG sites. h. Sequence context ± 10 bp around 5hmCHH sites.

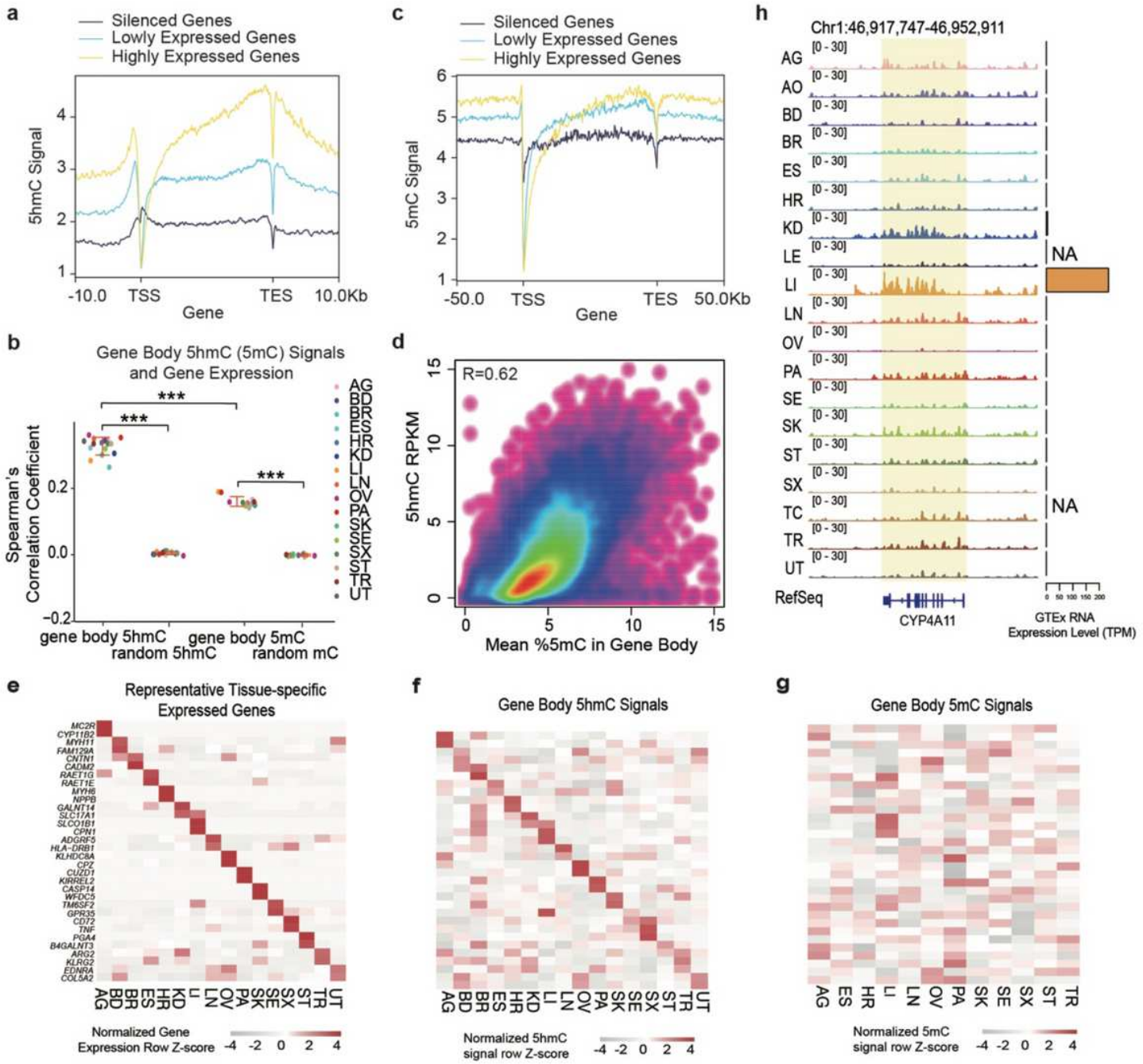


Figure 3

Gene body 5hmC correlates well with gene expression in human tissues. a. 5hmC profiles of genes expressed at high (yellow), low (cyan) and silenced (blue) levels in heart tissue. b. Spearman's correlation of gene body 5hmC (5mC) signals and gene expression levels. Random regions were selected as controls. *** represents P values < 0.0001 (Wilcox test was performed to compare the difference). c. 5mC profiles of genes expressed at high (yellow), low (cyan) and silenced (blue) levels in heart tissue. d. Scatter plot showing the correlation of 5mC and 5hmC signals at gene bodies. e. Heatmap displaying the expression levels of tissue-specific marker genes. Gene expression of matched tissue samples was downloaded from the GTEx project. f. Heatmap displaying the normalized gene body 5hmC signals of tissue-specific marker genes. The order of the row is identical to (e). g. Heatmap displaying the

normalized gene body 5mC signals of tissue-specific marker genes. The order of the row is identical to (e). The 5mC data were downloaded from ENCODE. h. IGV visualization of the 5hmC signals at the gene body of CYP4A11 and surrounding regions. Gene expression levels of the gene in different tissues are shown on the right.

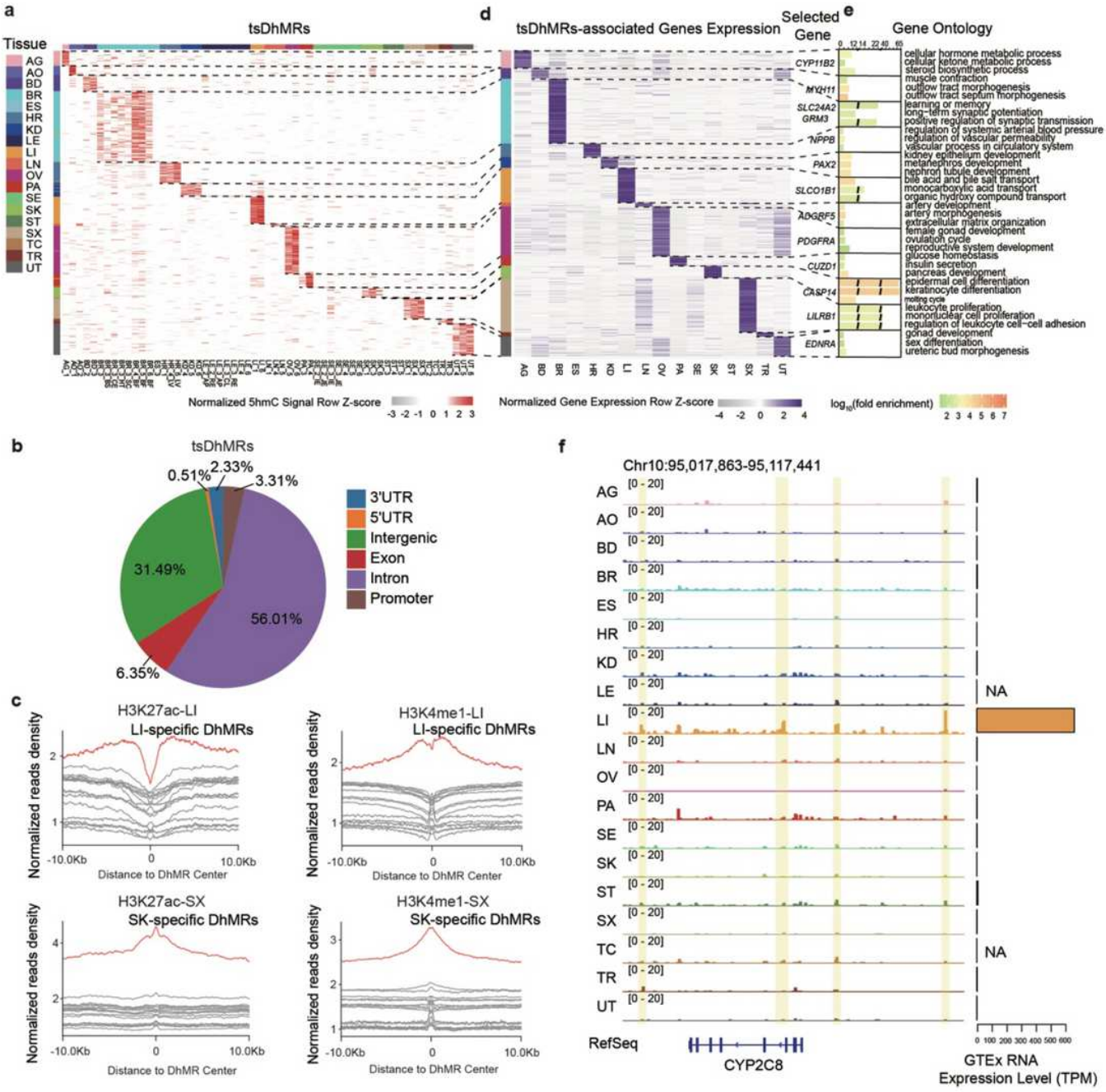


Figure 4

tsDhMRs are tissue-specific regulatory elements. a. Heatmap showing the normalized 5hmC signals in tsDhMRs. b. Pie chart showing the percentages of tsDhMRs in each class of genomic features. c. Profiles of H3K27ac and H3K4me1 modifications around tsDhMRs in the liver (LI) and spleen (SX). d. Heatmap showing the expression of genes associated with tsDhMRs within 500 kb. e. GO enrichment and

representative genes of tsDhMR-associated genes. f. IGV visualization of the 5hmC signals near CYP2C8 on chromosome 10. The highlighted regions are liver-specific DhMRs.

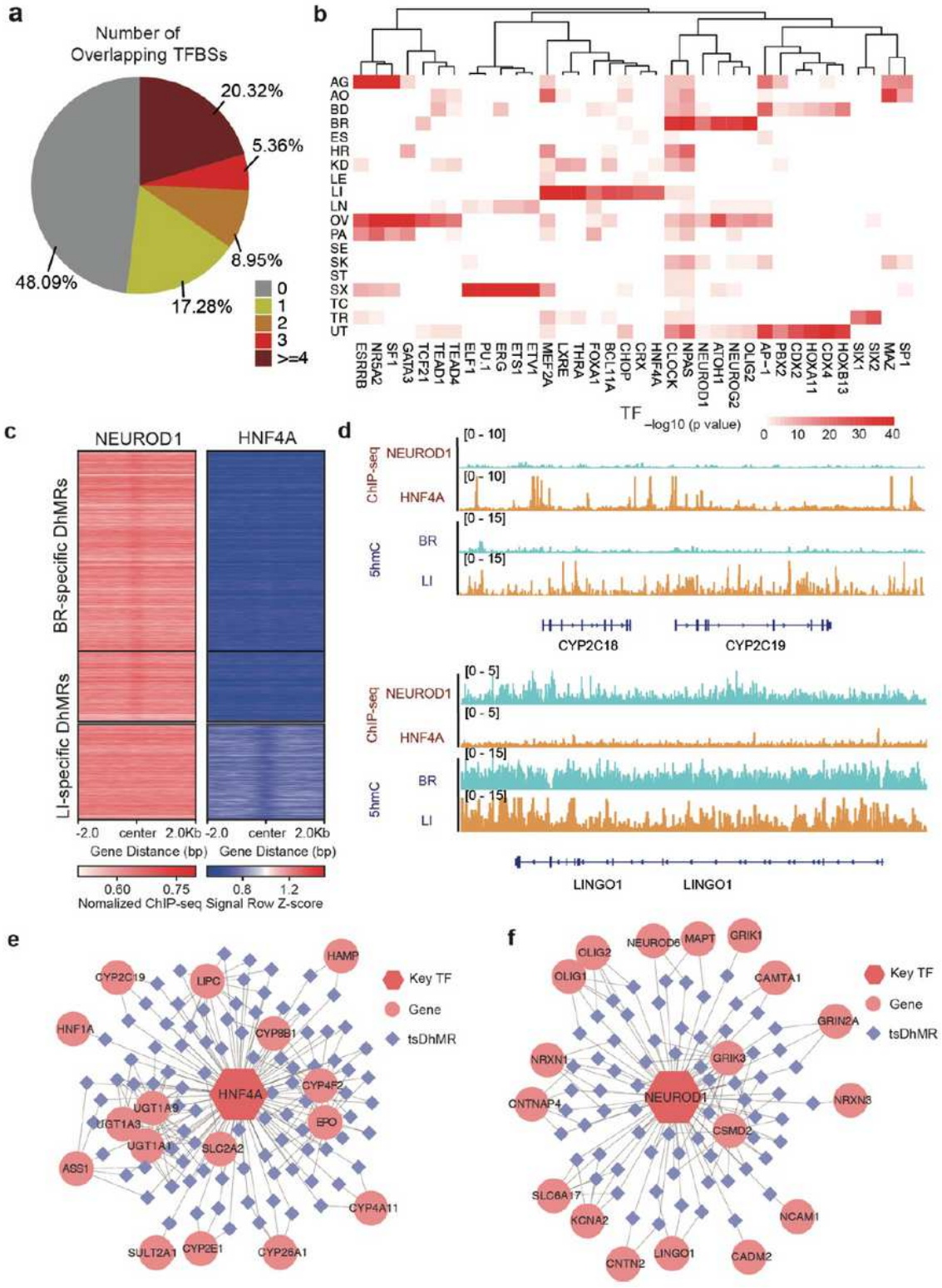


Figure 5

tsDhMRs are enriched for tissue-specific TFBSs. a. Overlap of tsDhMRs with ENCODE TFBSs. b. The TF motifs enriched in tsDhMRs of each tissue. The color scale represents the −log10(p value). c. ChIP-seq signals of NEUROD1 and HNF4A around BR-specific DhMRs and LI-specific DhMRs. d. IGV visualization

of 5hmC signals and ChIP-seq signals of NEUROD1 and HNF4A in the brain and liver. e. Key TF regulatory networks in the liver and brain. Hexagon represents Key TF; rhombus represents tsDhMRs; circle represents regulated genes. The line linking the hexagon and rhombus indicate that the TF binds to the tsDhMRs, which is supported by TF ChIP-seq data.

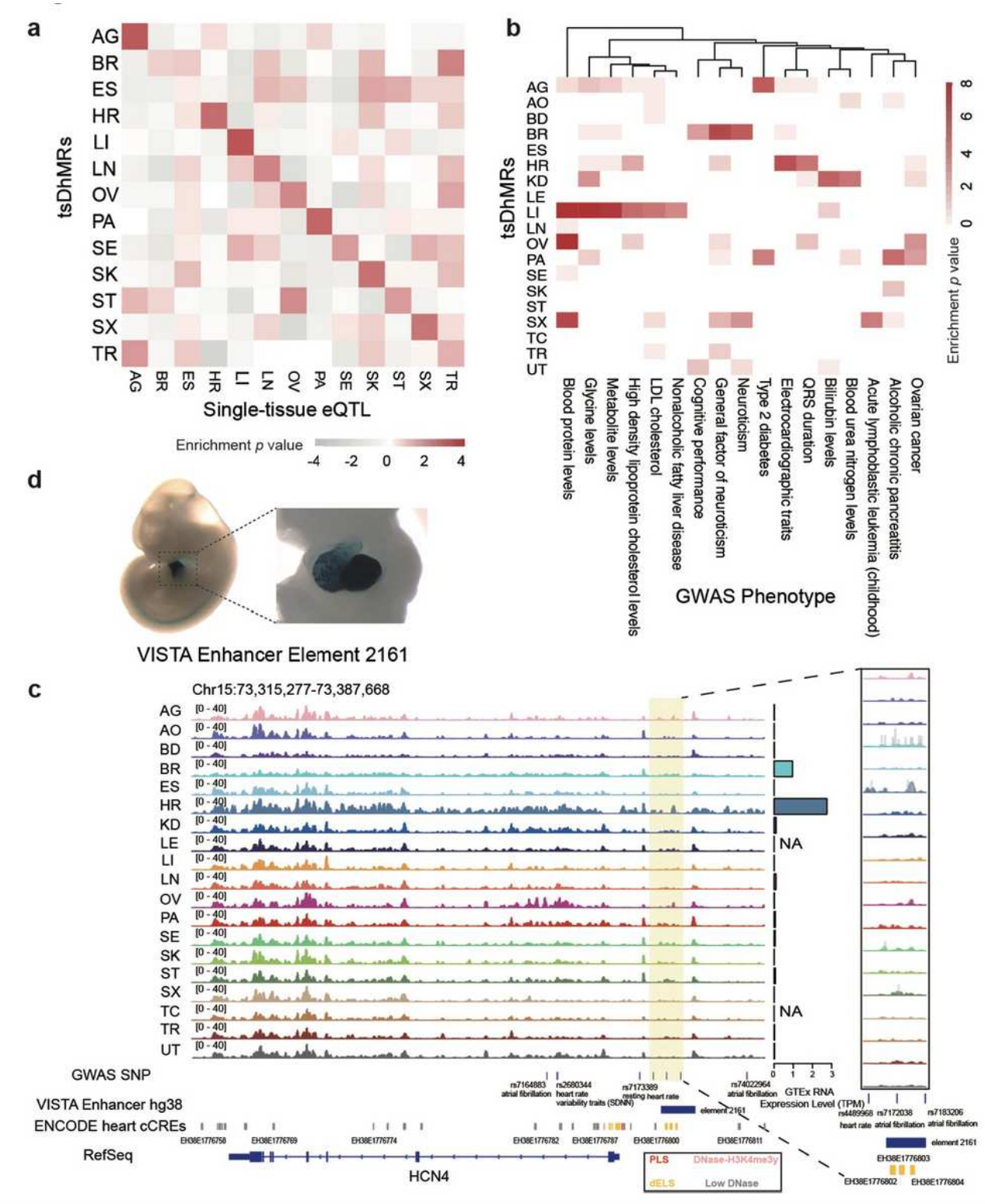


Figure 6

tsDhMRs enrich tissue-specific, phenotypic GWAS SNPs. a. Overlap of tsDhMRs with single-tissue eQTL SNPs. eQTL SNP data are from GTEx. b. Representative GWAS phenotypes enriched in tsDhMRs. c. IGV visualization of the 5hmC signals around HCN4 on chromosome 15. The locations of GWAS SNPs and VISTA enhancers are also shown. The highlighted region shows the LI-specific DhMRs. d. In vivo reporter assay of enhancer activity for VISTA enhancer element 2162, as obtained from the VISTA enhancer browser34.

Supplementary Files

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

- Supplementaryinformation0623.docx