Supplemental Materials and Methods

Mouse Strains, Matings, and Cell/Tissue Isolation

Transgenic *Sox9-ECFP* mice (Kim et al., 2007b) were maintained on a C57BL/6J (B6) genetic background. To isolate pre-Sertoli cells, *Sox9-ECFP* homozygous transgenic males were bred to CD-1 (Charles River) females in timed matings to generate E13.5 and E15.5 embryos. Noon of the day a vaginal plug was observed was defined as E0.5. *Sox9-ECFP*$^{+/-}$ embryos were dissected and testes removed from the adjacent mesonephros. Testes from one or more litters were pooled together, incubated in 500 µl of 0.25% Trypsin-EDTA (Gibco) plus 0.25% Collagenase at 37^{o}C for 8-10 minutes. The Trypsin-EDTA was removed and the tissue rinsed with 1X PBS with 3% BSA, and then dissociated by gentle pipetting in 500 µl 1X PBS with 3% BSA. Dissociated cells were passed through a cell strainer (BD Falcon) to ensure a single cell suspension.

FACS was performed by the Duke Comprehensive Cancer Center Flow Cytometry Shared Resource facility on a BD FACStar sorter, running at 12 psi, using a water-cooled Coherent argon laser tuned to 458 nm run at 50 mW. CFP emission was collected with a 485/22 bandpass filter. Following sorting, the cells were spun down at 2 krpm for 20 min at 4^{o}C and the supernatant removed. For RNA-seq, the cell pellet was snap-frozen and stored at -80^{o}C. For DNaseI-seq, cell pellets from the CFP-positive and CFP-negative fractions were resuspended in 250 µl of Recovery-Cell Culture Freezing Media (Gibco) and slowly frozen to -80^{o}C.

Mouse tissues (kidney, liver, heart, and brain) were collected from adult B6 mice, flash frozen and then pulverized before use. The mouse fibroblast cell line was derived from adult B6 mice (Jackson Labs). ESCs, also of the B6 strain, were kindly provided by Ute Hochgeschwender (Duke University) and were grown on gelatinized plates in the absence of a feeder layer or matrigel. To harvest ESCs, plates were washed with 1X PBS and treated with 0.25% Trypsin-EDTA for 7-10 minutes at 37^{o}C. An equal

volume of medium (containing 10% FBS) was then added to the plates to stop trypsinization. Cells were collected and pipetted up/down to get a single cell suspension and were centrifuged at ~1.2 krpm for 10 minutes. All medium was removed from the cell pellet.

DNaseI-seq Assay and Data Processing

Because our experiments were severely limited by the ability to collect large numbers of FACS purified pre-Sertoli cells, DNaseI digestion optimization was first carried out on differing amounts of gonadal cells collected by FACS (CFP-negative fractions). Attempts using 1-5 million cells were performed and limited success was achieved with 3 million cells; however, consistent digestion patterns were observed when using 5 million cells. Once 5 million CFP-positive cells were collected, the DNaseI-seq assay was performed as previously described (Boyle et al., 2008a; Song et al., 2010) with few modifications. Incubation of FACS purified pre-Sertoli cells with low concentrations of exogenous DNaseI enzyme was found to severely degrade the DNA; therefore DNase digestion was performed using endogenous nucleases. Briefly, cells were lysed and incubated at 37^{o}C for various times (5 min to 1 hour) and optimal DNase digestion was confirmed by pulse field gel electrophoresis (Song et al., 2010). DNaseI-seq libraries were then prepared as previously described (Boyle et al., 2008a; Song et al., 2010) and sequenced (three lanes) on the Illumina GAII platform by the Duke Genome Sequencing and Analysis Core. DNaseI-seq data was processed as previously described (Song et al., 2011). Briefly, sequences were aligned to the mouse reference genome (UCSC mm9) using BWA (Li et al., 2009), reads were filtered to remove PCR amplification artifacts (associated with library processing), base-pair signal (Parzen score) was generated using F-seq (Boyle et al., 2008b) and discrete peaks corresponding to DHSs were called.

RNA preparation and RNA-seq Data Processing

E15.5 *Sox9CFP*-positive cells were collected from ~165 embryos and pooled into 3 independent biological replicates, each containing ~1 million cells. RNA isolation was performed using the RNeasy Micro Kit as previously described (Jameson et al., 2012b) and according to the manufacturer's protocol (Qiagen). The RNA was DNaseI digested and eluted from the column with 14 µl of RNase-free water. QC and quantitation was performed on the Thermoscientific NanoDrop 2000 and the Agilent Bioanalyzer (NanoDrop results: ~135-160 ng/µl; Bioanalyzer results: ~90-110 ng/µl).

Library preparation and sequencing was carried out by the Duke Genome Sequencing and Analysis Core. Poly-A enriched mRNA libraries were generated from 0.8-1 µg of RNA using the standard Illumina Tru-Seq V2 protocol, then quantitated on the Agilent Bioanalyzer and adapted with index 23, 25 or 27. 7pM of the resulting library pool was run in a single Illumina HiSeq2000 lane to generate 50 bp single-end reads. Base calls were performed using CASAVA (version 1.8.2), which provides only pass-filtered reads. Output files for each biological replicate were concatenated. For each sample, 57-63 million reads were generated and over 96% of reads had a quality score equal to or greater than Q30, with a mean quality score of 38, therefore no further filtering was performed.

RNA-seq data was processed using Bowtie version 0.12.7 (Langmead et al., 2009) and RSEM version 1.2.0 (Li et al., 2011). The rsem-prepare-reference command was used to generate Bowtie-compatible index files for the UCSC mm9 transcriptome using UCSC gene transcript annotations and genome fasta files (random chromosomes were removed). Reads were aligned to the prepared reference transcript file using Bowtie with the following options: -v 3 -a -m 100 --best –strata. This resulted in ~83% of the reads having at least one reported alignment to the transcriptome.

Gene- and isoform-level abundance was estimated using the rsem-calculate-expression command. The TruSeq library fragment size distribution was estimated based on the Agilent Bioanalyzer

report; the fragment-length-mean and fragment-length-sd were set to 350 and 100, respectively. Expected counts were then quartile normalized and square-root transformed prior to further analysis.

TSS and regulatory domain assignments

To determine the overlap of DHSs with the TSS of each gene in the genome, a file was generated that assigned a single transcriptional start site to each gene. The transcriptional start and end for each gene was extracted from the refgene table downloaded from the UCSC genome browser (mm9). For genes that had multiple isoforms, the transcript's 5'-most (TxStart) and/or 3'-most (TxEnd) ends were used, resulting in one TxStart/TxEnd for each gene in the refgene table.

Gene-regulatory domains were generated by using GREAT (McLean et al., 2010) to assign each DHS to the one or two nearest genes (excluding non-coding RNAs). Briefly, for each gene a domain was extended up to 2 Mb from the TxStart and TxEnd to the next nearest gene in the 5' and 3' directions, but stopping at -5 kb from neighboring gene's TSS, leaving the proximal promoter assigned only to the nearest gene.

Genomic location assignments

Genomic locations were assigned to DHSs using previously published methods (Song et al., 2011). Briefly, DHSs were assigned to the first of the following categories that it overlapped: (1) promoter: overlaps a TSS or 2 kb upstream; (2) 5' exon/intron: overlaps the first exon/intron; (3) intragenic exon/intron: overlaps an internal exon or intron; (4) 3' exon: overlaps the last exon; or (5) intergenic: not overlapping any previous category. 5' exon, intragenic exon and 3' exon were then combined into a single category: exonic. 5' intron and intragenic intron were combined into a single category: intronic.

Cell type specificity categorization

Cell type specificity was determined by comparing DHSs from seven DNaseI-seq datasets: Sertoli, fibroblast, ESC, kidney, liver, heart and brain. DHSs were categorized into three groups: Sertoli-unique (only found in the Sertoli DNaseI-seq data), Sertoli-specific (present in Sertoli and up to 5 other cell types) and Sertoli-common (present in all seven cell types). Sequential intersections were performed for each of the DNaseI-seq datasets using the intersectBed command from the BedTools Software suite (version 2.17.0) (Quinlan et al., 2010). In each case (unique, specific or common), the boundaries for the Sertoli DHSs were maintained.

CTCF overlap

We used CTCF ChIP-seq data that was available on the Mouse Encode Project at Ren Lab website (http://chromosome.sdsc.edu/mouse/download.html) (Shen et al., 2012). CTCF ChIP-seq data from four adult tissues (liver, lung, spleen and testis) and four E14.5 tissues (brain, heart, liver and limb) was downloaded, and CTCF binding sites were extended 100 bp (50 bp in the 5' and 3' direction). Pairwise intersections were performed for each of the eight CTCF datasets with Sertoli-unique, Sertoli-specific and Sertoli-common DHSs using intersectBed (Quinlan et al., 2010) and at least 25 bp of overlap was required to be considered overlapping. The percentage of DHSs that overlapped a CTCF binding site was reported as the average and standard deviation of overlap across all 8 cell types (see Figure 2).

DMRT1 binding site enrichment analysis

To look for enrichment of DMRT1 binding sites in Sertoli DHSs, the overlap of DMRT1 ChIP-seq binding sites from E13.5 testes (Krentz et al., 2013) with Sertoli DHSs was analyzed using intersectBed (Quinlan et al., 2010). 25 bp of overlap was required to be considered overlapping. Sertoli DHSs were subdivided into "unique", "shared", or "common" based on their cell-type specificity as described above.

Enrichment of DHSs at Sertoli-expressed genes

To determine whether Sertoli DHSs were enriched near Sertoli-expressed genes, the number of DHSs that mapped to each gene's regulatory domain were counted (using intersectBed; (Quinlan et al., 2010) for the following categories of Sertoli DHS: common, unique+shared and active enhancers (H3K27ac-positive DHSs). Overlap was analyzed with the following categories of genes: mm9 (refers to all genes, isoforms removed as described above); GUDMAP (the subset of mm9 genes that were analyzed in our previous microarray study; (Jameson et al., 2012b); Sertoli (genes expressed >1.5 fold higher in E13.5 Sertoli cells compared to E13.5 pregranulosa cells), pregranulosa (genes expressed >1.5 fold higher in E13.5 pregranulosa cells compared to E13.5 Sertoli cells), and germ cells (genes expressed >1.5 fold higher in E13.5 male or female germ cells compared to all other gonadal cell lineages). Statistical significance was calculated using a two-sample Mann-Whitney (two-sided) test to compare the distributions for each gene set to the GUDMAP reference set.

Transient transgenics, immunocytochemistry and imaging

A putative regulatory region (UCSC mm9 coordinates chr2:104914099-104915125) upstream of *Wt1* was amplified by PCR and cloned into the NotI site of the *Hsp68 –LacZ* reporter vector (obtained from Addgene; Plasmid #33351). Cloning was carried out using In-Fusion HD (Clonetech). To prepare DNA for zygote injection, 50 µg of the *TgWt1* plasmid was linearized with NotI-HF and HindIII and gel purified by electroelution. The DNA was phenol-chloroform extracted, ethanol precipitated and resuspended in EmbryoMax Injection Buffer (Millipore, MR-095-10F). The DNA was further purified on a DNA-cleanup column (Qiagen) and eluted again in EmbryoMax Injection Buffer. Pronuclear injections into B6SJLF1/J zygotes were performed by the Duke Transgenic Core Facility to generate transient transgenics.

Embryos were dissected at E13.5 and the embryonic tail was removed for PCR genotyping to detect the *LacZ* gene (Primers (5'-3'): F-ATCCTCTGCATGGTCAGGTC and R-CGTGGCCTGATTCATTCC).

Gonads were carefully dissected from embryos and fixed in 4% paraformaldehyde for several hours or overnight at 4^{o}C. The remaining embryo bodies were fixed in 4% paraformaldehyde for 8 minutes, washed in X-gal wash buffer (2mM $MgCl_2$, 0.2% NP-40 in 1X PBS) and then incubated overnight at 37^{o}C in X-gal staining solution (5mM potassium ferrocyanide, 5mM potassium ferricyanide, 1mg/ml X-gal in X-gal wash buffer).

For immunostaining, fixed gonads were washed three times in 1X PBS and incubated in blocking solution (10% FBS, 3% BSA and 0.1% Triton-X-100 in 1X PBS) for 1 hr at room temperature. Blocking solution was replaced with primary antibodies diluted in blocking solution and incubated overnight at 4^{o}C. The next morning, samples were washed three times in washing solution (1% FBS, 3% BSA and 0.1% Triton-X-100 in 1X PBS) followed by one-hour incubation with blocking solution. Samples were then incubated with secondary antibodies, diluted in blocking solution, overnight at 4^{o}C. Following three washes, samples were mounted in DABCO (2.5% 1,4, diazagicyclo octane, 90% glycerol in 1X PBS). Images were captured on a Leica SP2 confocal microscope.

Primary and secondary antibodies were used at the following dilutions: rat-anti-CDH1, 1:250 (Zymed, 13-1900); rabbit-anti-β-galactosidase, 1:10,000 (MP Biomedicals, 55976); goat-anti-MIS/AMH, 1:250 (Santa Cruz, sc-6886); Alexa Fluor 488-anti-rat, 1:500 (Molecular Probes, A21208); Cy3-anti-rabbit, 1:500 (Jackson ImmunoResearch Laboratories, 711-165-15); Alexa Fluor 647-anti-goat, 1:500 (Molecular Probes, A21447).

ChIP-seq assay and data processing

For ChIP-seq analysis, FACS purified Sertoli cells were pelleted, resuspended in 360 µl of PBS and cross-linked with 10 µl of 37% formaldehyde at room temperature for 10 minutes. Cross-linking was stopped by addition of 46.3 µl of 1M glycine for 5 minutes at room temperature. Cells were then pelleted, supernatant was removed and stored at -80^{o}C. 1 million cells were pooled from multiple sorts

washed twice in 500 µl of PBS with protease inhibitors. Cells were resuspended in 500 µl of lysis buffer (50mM Tris-HCL, 10mM EDTA, 1% SDS and protease inhibitors) and sonicated with a Branson 450 Sonicator (output power of 3, duty cycle of 30% for 16 cycles of 30 seconds with 1 minute rest time between sonications). Bead-antibody complexes were prepared by incubating 30 µl of dynabeads (Protein A; Life Technologies 10002D) with 2.5 µg antibody (Rabbit-anti-H3K27ac; Abcam ab4729).

The sonicated lysate was spun down at 4^{o}C for 10 minutes at 10 krpm with 40 µl of the supernatant set aside as input and 200 µl transferred to tubes containing pre-incubated bead-antibody complexes. We added 700 µl of ChIP Dilution Buffer (CDB) (1 % Triton X-100 (Sigma T8787), 2mM EDTA, 150mM NaCl, 20mM Tris (pH=8.0)) with protease inhibitors to IP tubes, and incubated overnight at 4^{o}C. 160 µl of CDB and 8 µl of 5M NaCl were added to the input tube which was incubated at 65^{o}C overnight. The following day, IP tubes were washed as follows: Once with Wash Buffer 1 (50mM Tris HCl, 1mM EDTA, 150 mM NaCl, 0.1% SDS, 0.1% Triton X-100, 0.1% Sodium deoxycholate), twice with Wash Buffer 2 (50mM Tris HCl, 1mM EDTA, 500 mM NaCl, 0.1% SDS, 0.1% Triton X-100, 0.1% Sodium deoxycholate), once with Wash Buffer 3 (10mM Tris HCl, 1mM EDTA, 1% NP-40, 1% Sodium deoxycholate, 250mM LiCl), twice with Wash Buffer 4 (50mM Tris HCl, 1mM EDTA, 500 mM LiCl, 1% NP-40, 0.7% Sodium deoxycholate), twice with TE buffer (pH=8.0). All washes were done in 1 ml, with added protease inhibitors, at 4^{o}C for 5 minutes. Solutions for wash buffers were modified from the protocols posted on the Epigenomics Roadmap website.

DNA-protein complexes were eluted twice from the beads with 100 µl elution buffer (100mM sodium bicarbonate, 1% SDS, 8mM NaOH). 8 µl of 5M NaCl was added to the eluates, as well as the input tube, and incubated at 65^{o}C overnight. Samples were treated with 1 µl RNase-cocktail (Life Technologies AM2286) for 30 minutes at 37^{o}C, then 4 µl of 0.5M EDTA, 8 µl of 1M Tris and 1 µl of 10 mg/ml Proteinase K was added for 60 minutes at 45^{o}C. Finally, DNA was purified using PCR purification columns (Qiagen; 28104).

For library preparation for sequencing, DNA was concentrated using a vacuum centrifuge to ~10 µl. 10 µl of IP DNA and 1 µl of input DNA was used in the library preparation using the Rubicon ThruPLEX FD kit according to the manufacturer's protocol. Size selection of smaller size amplified DNA was done with SPRI beads (Agencourt AMPure XP A63880) at 0.6x concentration. Sequencing was performed at Duke's Genome Sequencing and Analysis core facility on the Illumina HiSeq2000/2500

ChIP-seq reads were aligned with Bowtie (Langmead et al., 2009) with only uniquely aligning reads used for future processing. Peaks were called with SICER (Zang et al., 2009) with enrichment called for the histone modifications using the input track as the control. The species variable was set to mm9, redundancy threshold to 2, window size to 200, fragment size to 150, effective genome fraction to 0.7, gap size to 600 and FDR to 0.01. Peaks found in both replicates were used to identify active enhancers and inactive DHSs.

Sequence analysis to identify predictive 6-mers and matching motifs

To identify enriched motifs, a discriminative classification approach using sequence features was performed incorporating previous methods (Lee et al., 2011; Natarajan et al., 2012). 6-mers, with reverse complements counted as the same 6-mer, were counted in each region, normalized by the length of the region, and used as features for an L1-norm sparse logistic regression classifier (Koh et al., 2007). Ten randomized iterations of 4-fold cross-validation were performed to generate 40 different partitions of the data. 6-mers that showed non-zero regression coefficients in over 75% of the partitions were deemed to be significant and shown in Table S1A-D. We evaluated the 6-mers identified as consistently important in our classifiers by two metrics. First, we used the regression coefficient averaged over the cross-validation iterations. The regression coefficient in a logistic regression classifier is the log odds ratio as a result of a unit increase in the variable. In addition, we calculated the prevalence ratio of the 6-mers. Prevalence ratio is defined as the ratio of the average length normalized 6-mer frequency between Sertoli specific DHS (active enhancer) vs. flanking regions (inactive DHS). 6-

mers were matched to known TF binding sites with TOMTOM (Gupta et al., 2007) using the following options: -no- -min-overlap 5 -mi 1 -dist pearson -evalue -thresh 0.5 -query-pseudo 0.01. The motif database included motifs from the JASPAR core vertebrate motifs, the UNIPROBE database and those generated by Jolma et al. (Bryne et al., 2008; Jolma et al., 2013; Hume et al., 2015). Additionally, the SF1 (Baba et al., 2014) and DMRT1 (Krentz et al., 2013) motifs were included.

Analysis of significance of motif matches

To assess significance of motif matches in different regions, we used FIMO from the MEME suite (Bioinformatics 27:1017) with default settings to scan different regions of the genome with three motifs SOX9, SF1, and GATA4. The score of the motif matches were then summed up across each region and length normalized. We used a one-sided Mann-Whitney U test to test for significance between the length normalized motif matches in each region.

Supplemental Figures

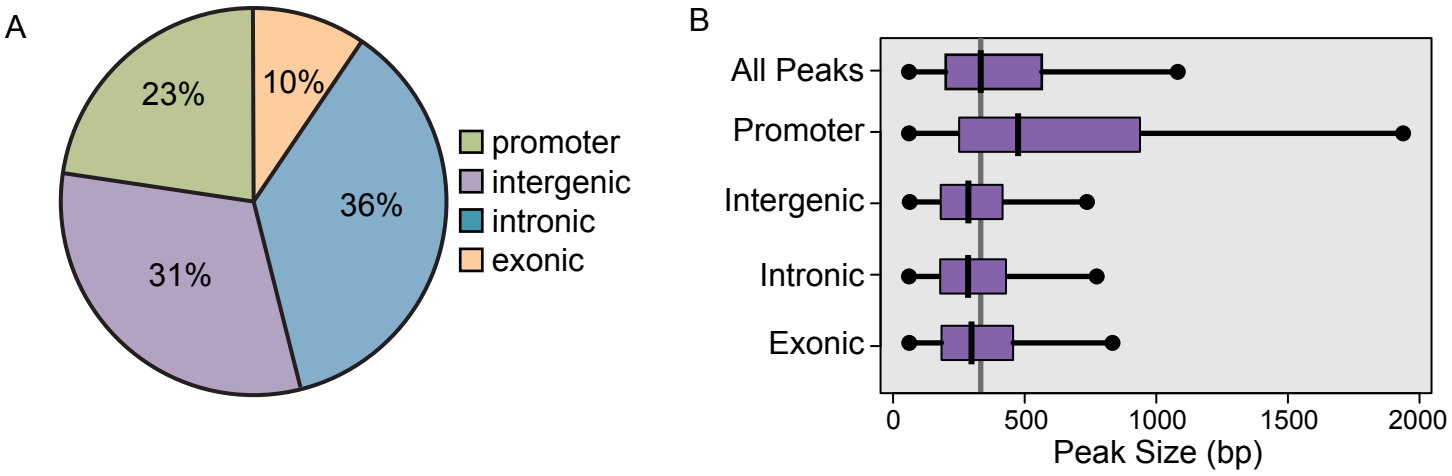
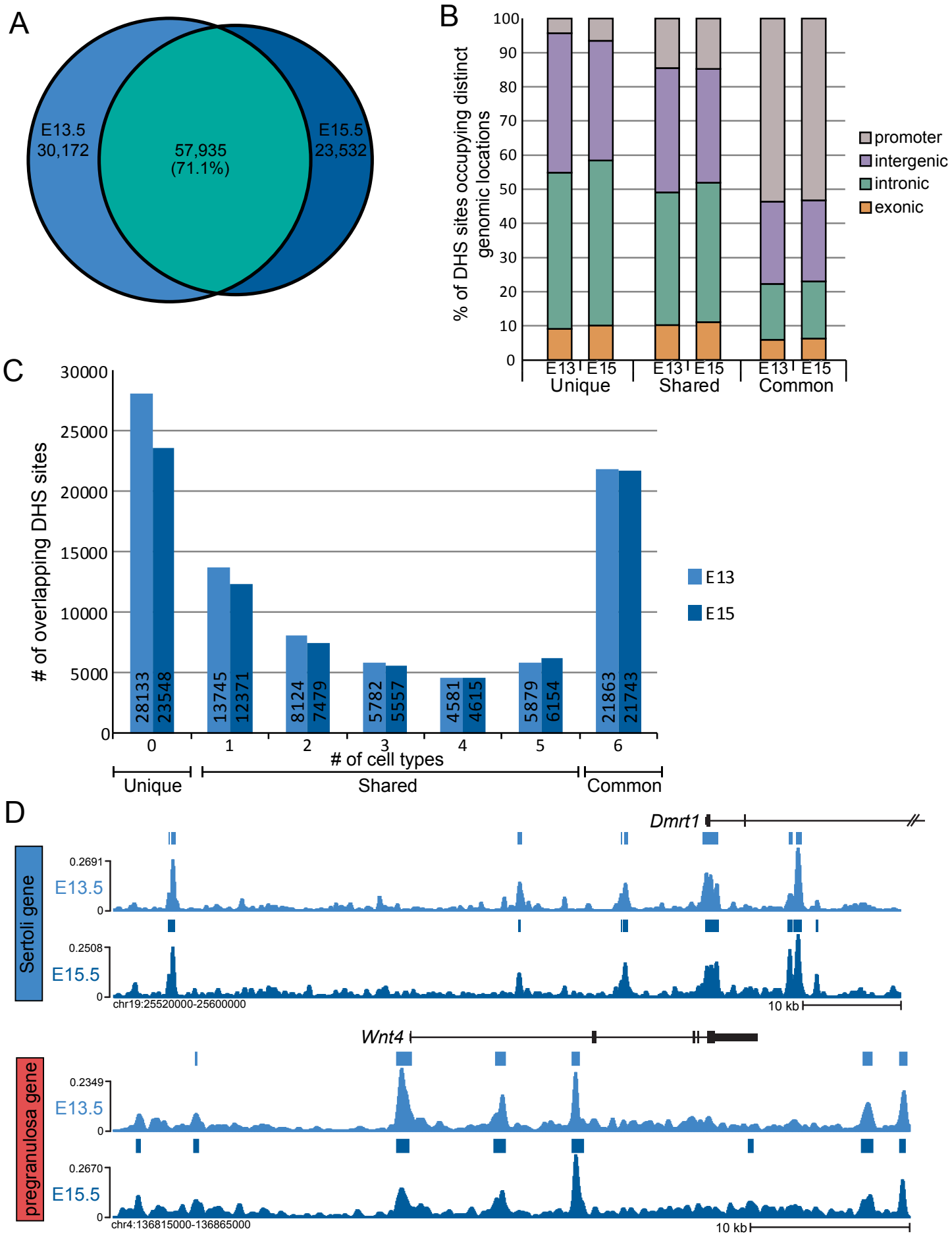
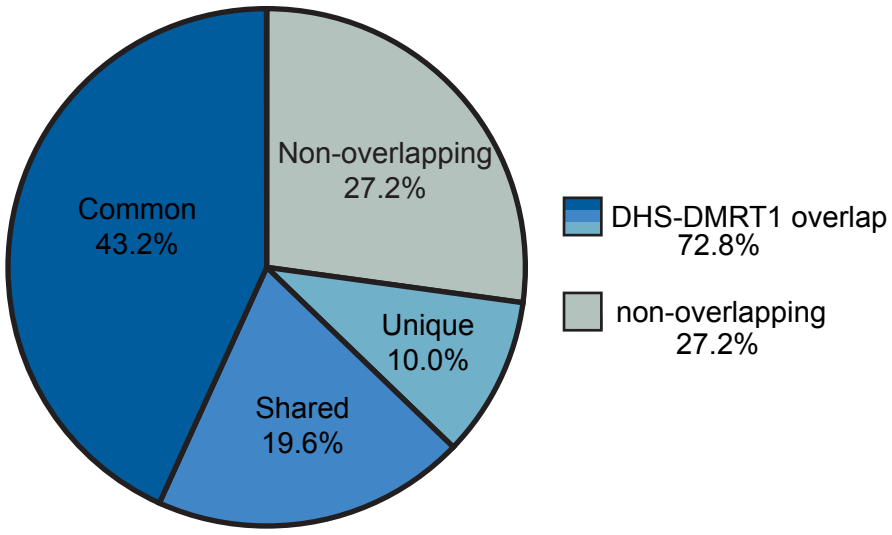
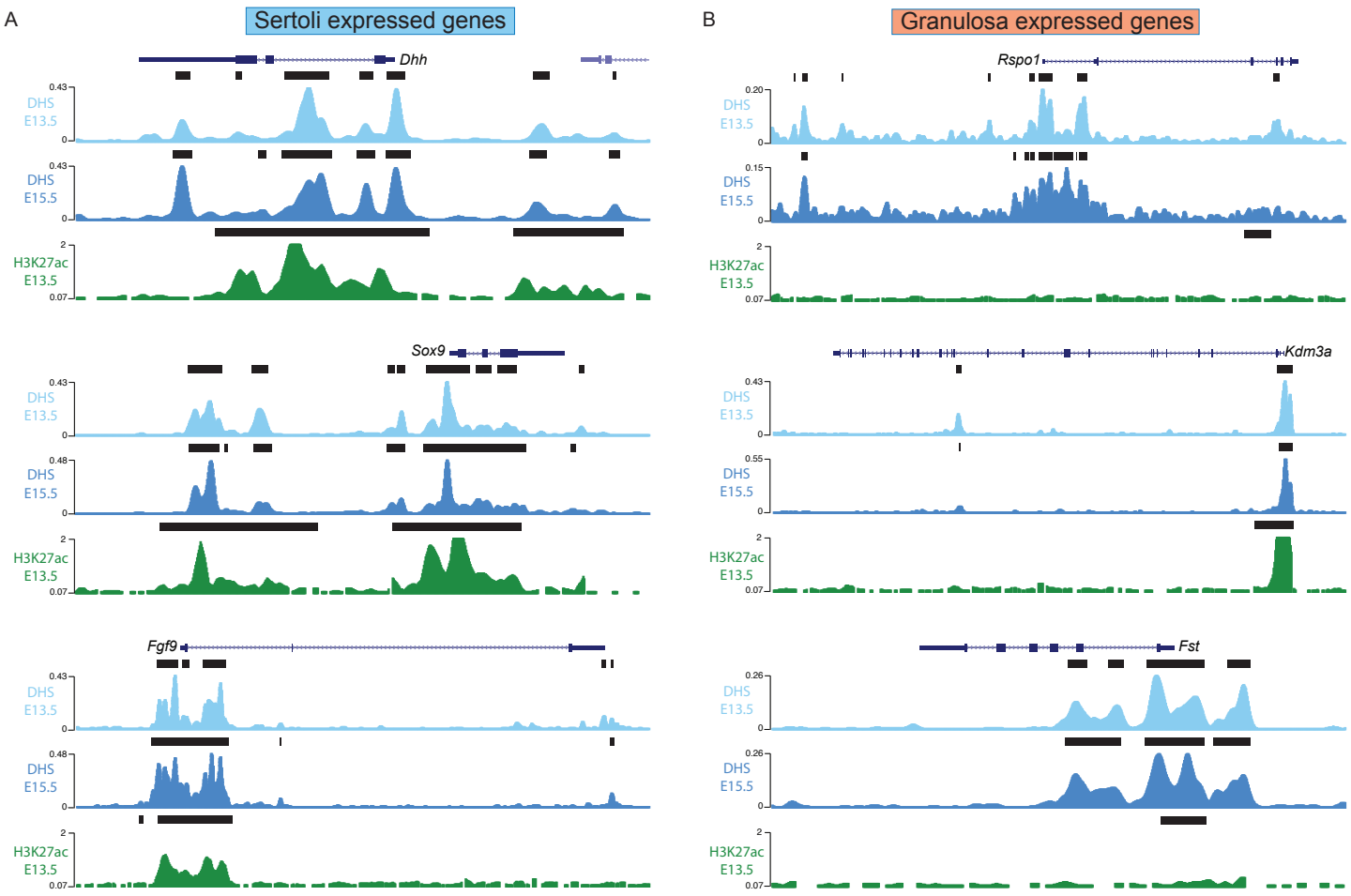


Figure S1. Location and size distribution of E15.5 DHSs. (A) Genomic location distribution analysis of all DHSs. Peaks were categorized by overlap with specific genomic features (promoter, exonic, intronic or intergenic) as described in the Online Methods. (B) A boxplot of the peak size distributions. Purple boxes indicate the middle 50% of peaks, lines mark the lower (left) and upper (right) 25%. Outliers are not shown. The vertical gray line indicates the median length for all DHS peaks (330bp).


A
E13.5
30,172
57,935
(71.1%)
E15.5
23,532
B
% of DHS sites occupying distinct genomic locations
100
90
80
70
60
50
40
30
20
10
0
E13
E15
Unique
Shared
Common
promoter
intergenic
intronic
exonic
C
# of overlapping DHS sites
30000
25000
20000
15000
10000
5000
0
28133
23548
13745
12371
8124
7479
5782
5557
4581
4615
5879
6154
21863
21743
0
1
2
3
4
5
6
# of cell types
Unique
Shared
Common
E13
E15
D
Dmrt1
Sertoli gene
0.2691
E13.5
0
0.2508
E15.5
0
chr19:25520000-25600000
10 kb
Wnt4
pregranulosa gene
0.2349
E13.5
0
0.2670
E15.5
0
chr4:136815000-136865000
10 kb


Non-overlapping
27.2%
Common
43.2%
Unique
10.0%
Shared
19.6%
DHS-DMRT1 overlap
72.8%
non-overlapping
27.2%


A
Sertoli expressed genes
B
Granulosa expressed genes
Dhh
Rspo1
Sox9
Kdm3a
Fgf9
Fst
DHS E13.5
DHS E15.5
H3K27ac E13.5