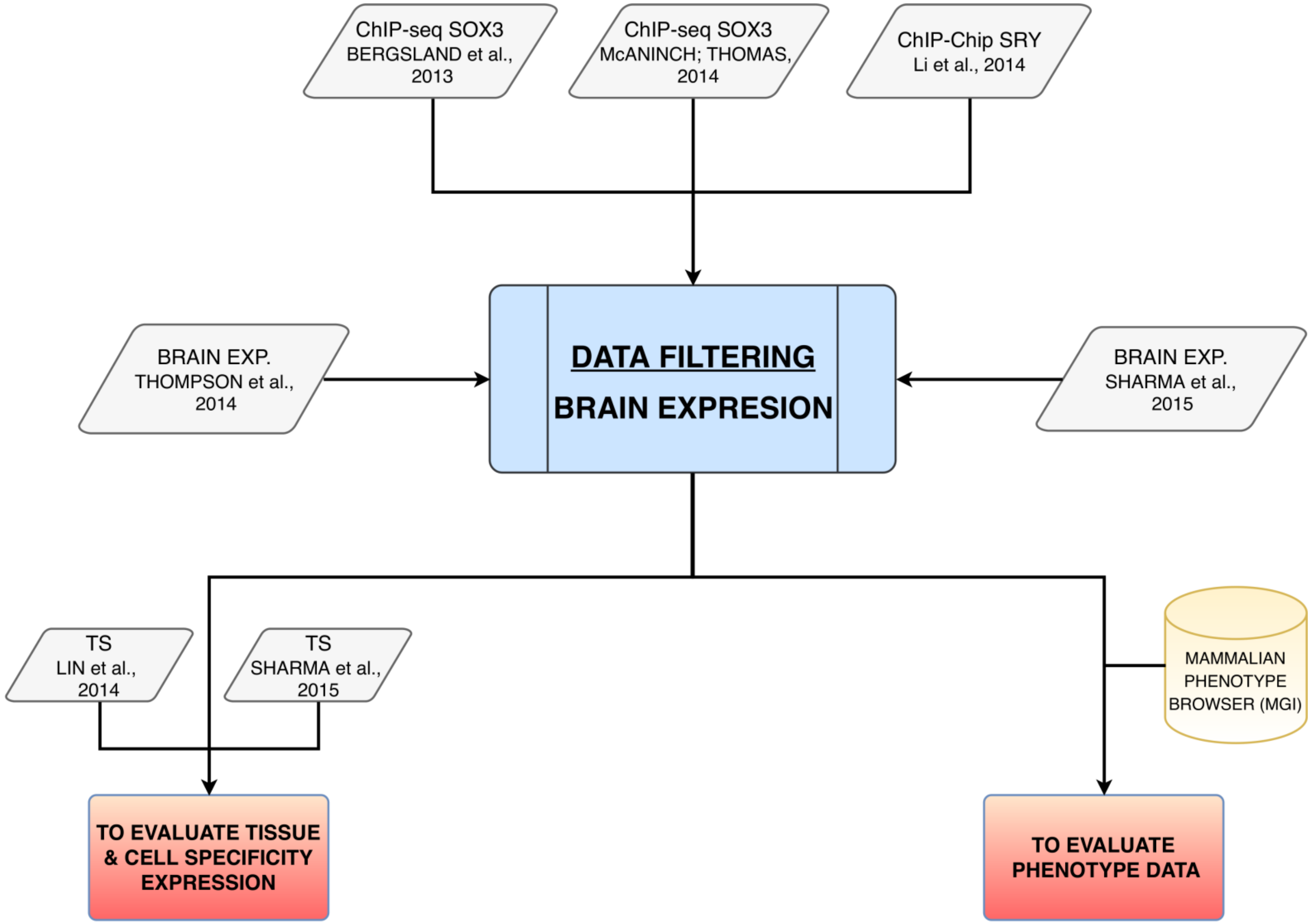
Supplementary Information

Supplementary Figures

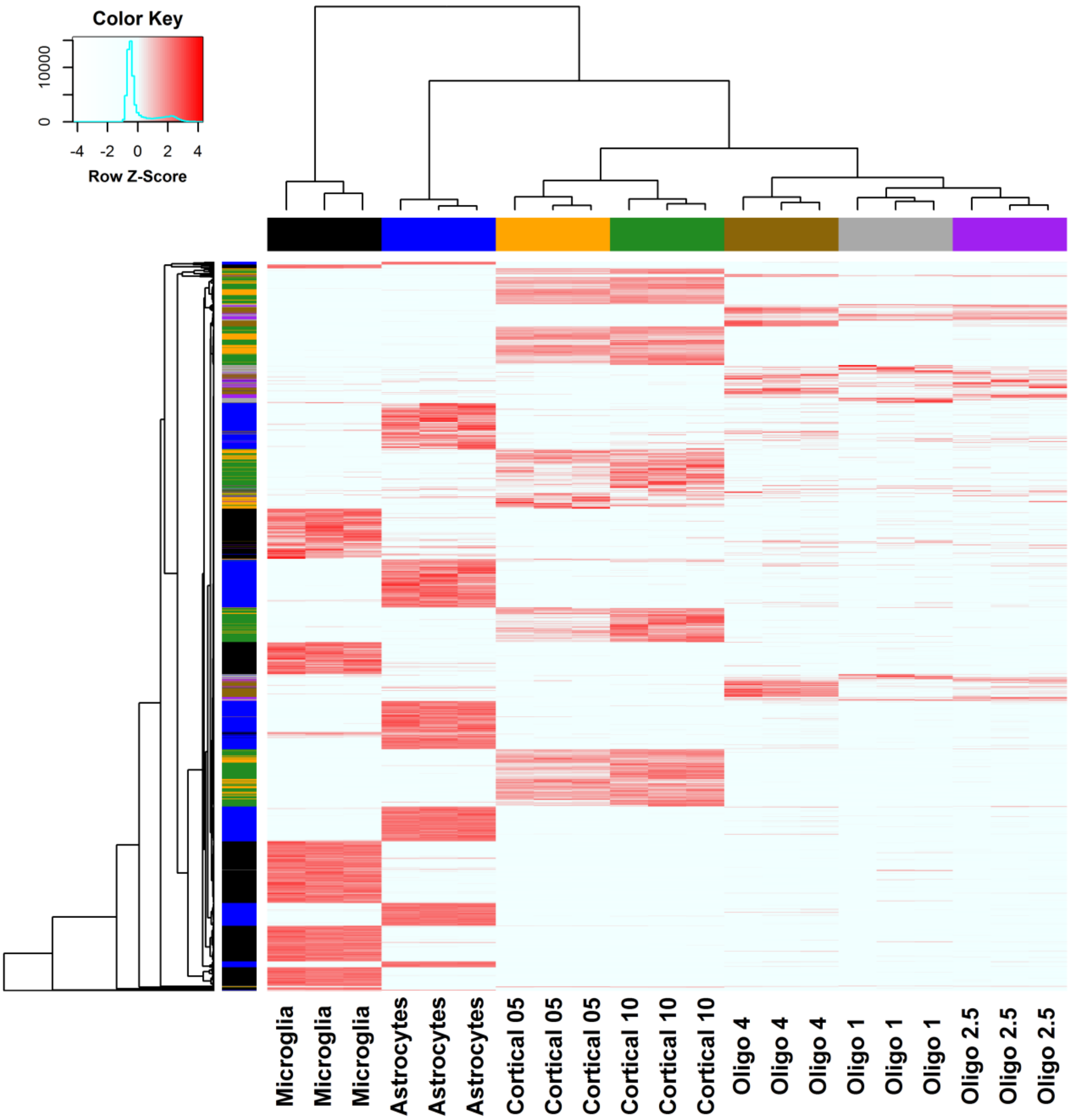
Coexpression network analysis:

Briefly, a pairwise correlation matrix was computed, and adjacency was calculated by increasing the correlation matrix to a power of 7, which maximized the scale-free topology properties of the network (R^2 fit above 0.8), used for all networks. Topological overlap measures (TOMs) were used to build average linkage hierarchical clustering. Finally, modules were defined as branches of the resulting clustering tree using hybrid dynamic tree-cutting to have robustly defined modules. The minimum module size was set to 20 genes, with the deepSplit parameter set to 2. To discriminate between changes in modules comparing ASD and CON, we performed a module preservation analysis that identified which modules were preserved in both datasets using 1,000 permutations. Then, a Z-summary was calculated to exclude the possibility of randomness in preservations and to indicate whether a module was strongly (Z-summary > 10), moderately (2 pa Z-summary < 10) or not preserved (Z-summary < 2).



Supplementary Figure S1: Flowchart of analyses made in mouse model. Trapeze and cylinders represent databases and studies' datasets. Blue squares designate the addressed question. Each input is assigned according to study or database and type of data.

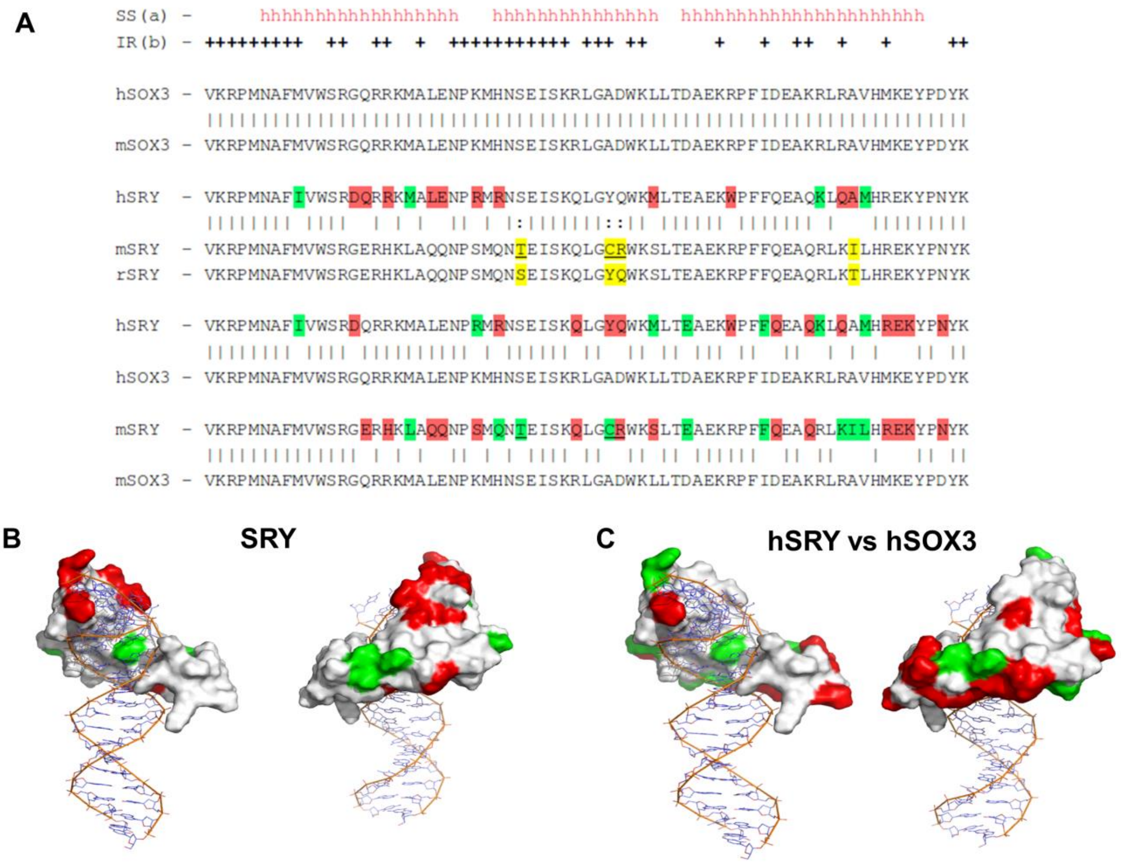
Cell type specific measure:



Supplementary Figure S2: Hierarchical clustering (Ward's method) of 21 samples (columns) based on 2,907 genes (rows) that showed tissue-specific (TS) expression according to H and Q values. Each gene was assigned to the tissue that presented the lower Q value. Colors from columns and rows assign the type of cell and TS. Black: adult microglia (821 genes), blue: astrocytes (821 genes), orange: cortical neurons division 05 (283 genes), green: cortical neurons division 10 (595 genes), brown: oligodendrocytes division 04 (180 genes), gray: oligodendrocytes division 01 (107 genes) and purple: oligodendrocytes division 2.5 (100 genes). Red and light blue colors from heatmap represent high and low expression, receptively, as summarized by z-score.

Similarities of HMG-box SOX3 and SRY among species.

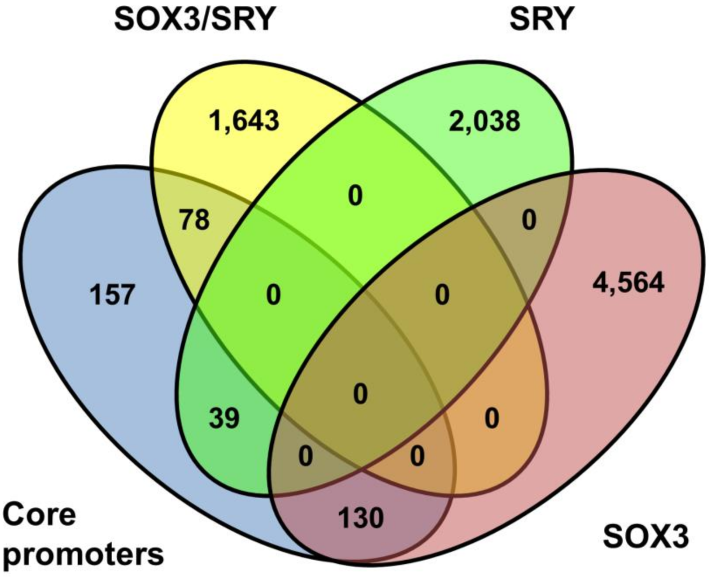
Comparisons between human and mouse SOX3 shows that HMG-box are identical , while SRY shows high conservation with 84% of residues conserved among human, mouse and rat. The alignment showed that 54 residues are identical (78%), four residues are similar (6%) and only 11 positions have amino acids with different properties (16%). From the 11 residues, four are in the protein-DNA interface. Comparing hSOX3 and hSRY, 49 residues are identical residues (71%), seven are similar residues (10%) and 13 are different residues (19%), from which five are located in the interface between DNA and protein-domain. The differences observed between the SRY HMG-box between species do not allow us to state that the proteins bind to identical DNA sequences with the same affinity; likewise, the high similarity of the domains does not allow us to conclude that there is a difference of specificity. Protein sequences were retrieved from UniProt database (EMBL, SIB Swiss Institute of Bioinformatics, & Protein Information Resource (PIR), 2013), alignments were performed in T-coffee (Notredame, Higgins, & Heringa, 2000) and 3D images in Pymol (DeLano, 2002).



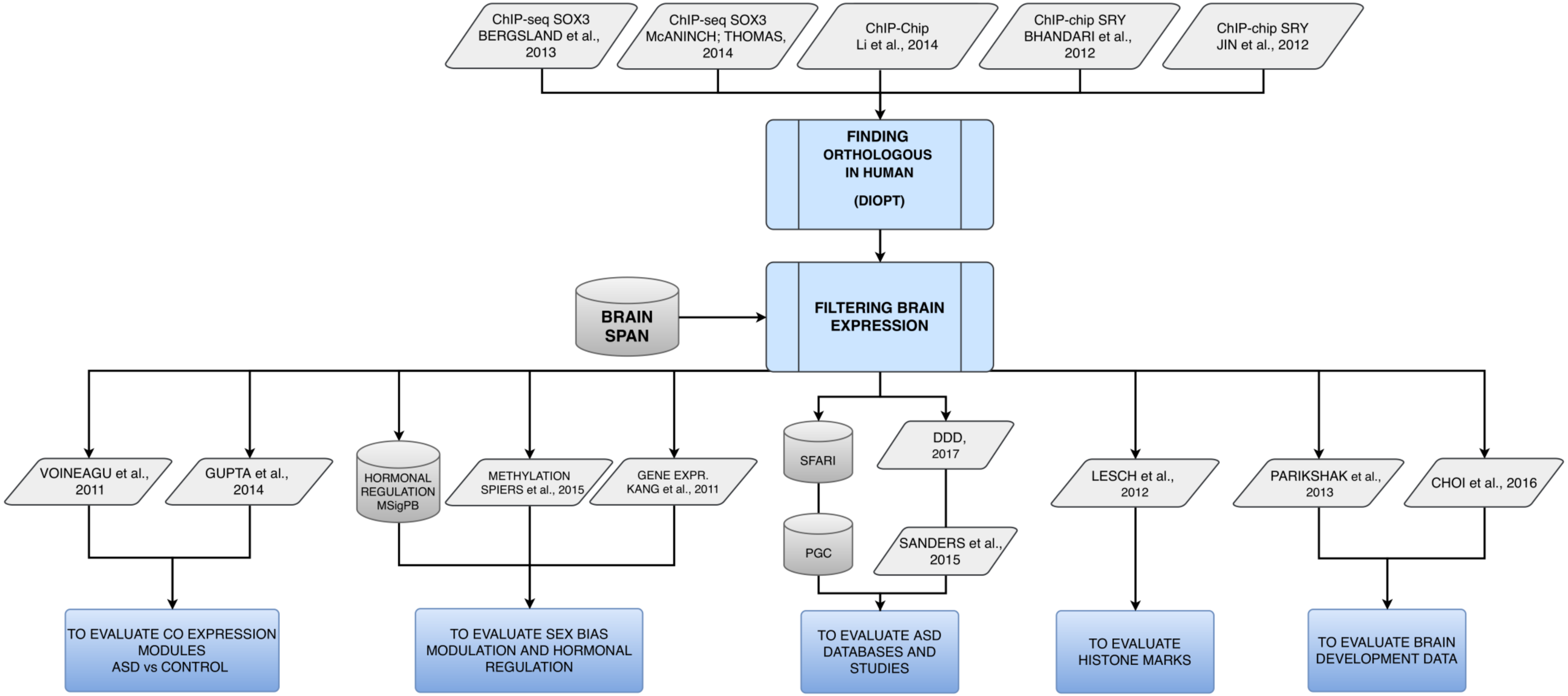
```
SS(a) -      hhhhhhhhhhhhhhhhhh   hhhhhhhhhhhhhhh  hhhhhhhhhhhhhhhhhhhhhh
IR(b) - +++++++++  ++  ++  +  +++++++++++ +++ ++      +   +  ++  +   +     ++

hSOX3 - VKRPMNAFMVWSRGQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFIDEAKRLRAVHMKEYPDYK
        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
mSOX3 - VKRPMNAFMVWSRGQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFIDEAKRLRAVHMKEYPDYK

hSRY  - VKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYK
        |||||||| ||||  | | |  || | |:|||||||::|| |||||| ||||||| |   |||||||||
mSRY  - VKRPMNAFMVWSRGERHKLAQQNPSMQNTEISKQLGCRWKSLTEAEKRPFFQEAQRLKILHREKYPNYK
rSRY  - VKRPMNAFMVWSRGERHKLAQQNPSMQNSEISKQLGYQWKSLTEAEKRPFFQEAQRLKTLHREKYPNYK

hSRY  - VKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYK
        |||||||| |||| |||||||||| | |||||| ||  || || ||| ||  ||  | | |   || ||
hSOX3 - VKRPMNAFMVWSRGQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFIDEAKRLRAVHMKEYPDYK

mSRY  - VKRPMNAFMVWSRGERHKLAQQNPSMQNTEISKQLGCRWKSLTEAEKRPFFQEAQRLKILHREKYPNYK
        |||||||||||||| | | |  || | | |||| ||  || || ||||||  || ||   |   || ||
mSOX3 - VKRPMNAFMVWSRGQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFIDEAKRLRAVHMKEYPDYK
```

Supplementary Figure S3: A-) Protein alignment of HMG-box domain. SS and IR represent the secondary structure and interface residue between DNA and protein. Green and red colors represent similar and non similar substitutions and white represents identical residues. First alignment shows that human (hSOX3) and mouse (mSOX3) SOX3 are identical concerning the HMG-box domain, with all 69 aa conserved. The second alignment shows the similarities and differences of SRY protein among three species (mouse, rat and human). Yellow represent positions with different residues between mouse and rat and "**:**" represents partially conserved residues. Third and fourth alignments shows human and mouse SRY and SOX3, respectively, which present high similarities for both species. **B -)** HMG-box and DNA 3D images from SRY in all three species. **C-)** HMG-box and DNA 3D image of hSRY and hSOX3.

<u>Histone core promoter's analysis:</u>

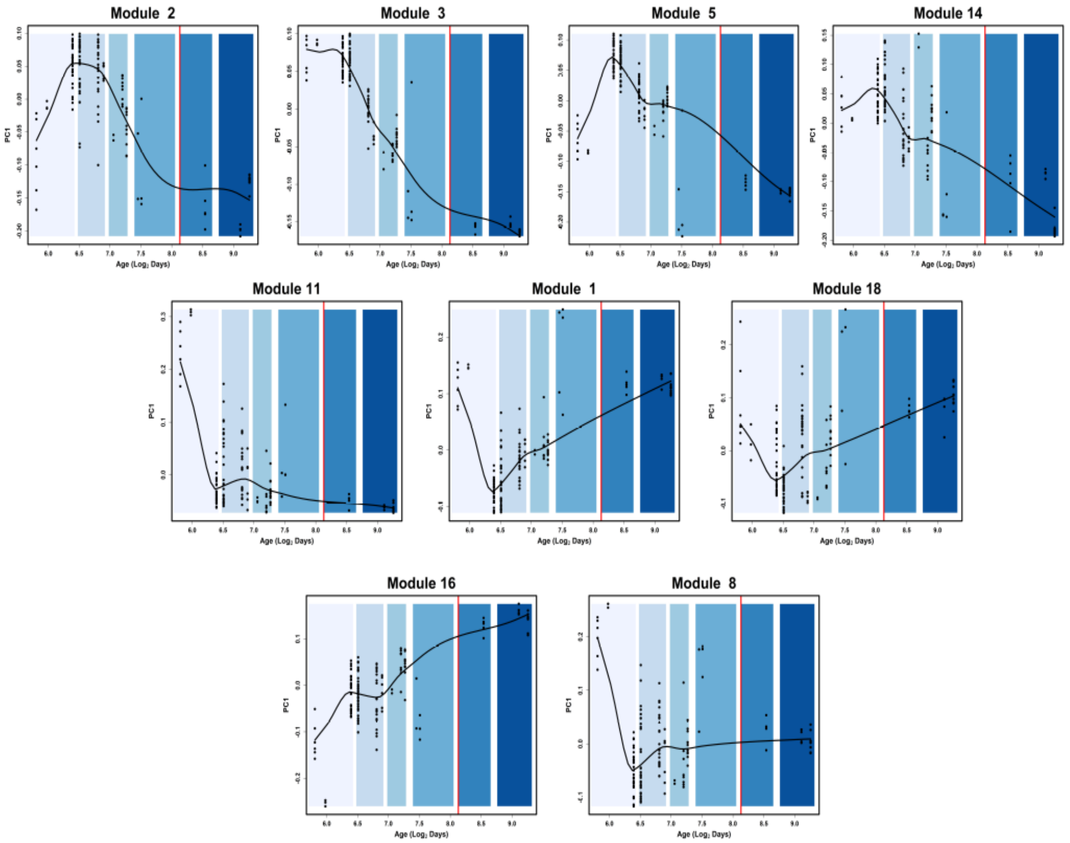


Supplementary Figure S4: Venn diagram showing the overlap of 404 core promoters (Lesch, Silber, McCarrey, & Page, 2016) between five species (human, mouse, rhesus, bull and opossum) in genes presenting a role during development of somatic cells. The overlap shows 247 (61%) genes with core promoter constrained during evolution, which were in one of our three gene datasets (SOX3/SRY, SOX3 or SRY). These genes seem to be under the SOX3 and/or SRY sex chromosome regulation proteins.

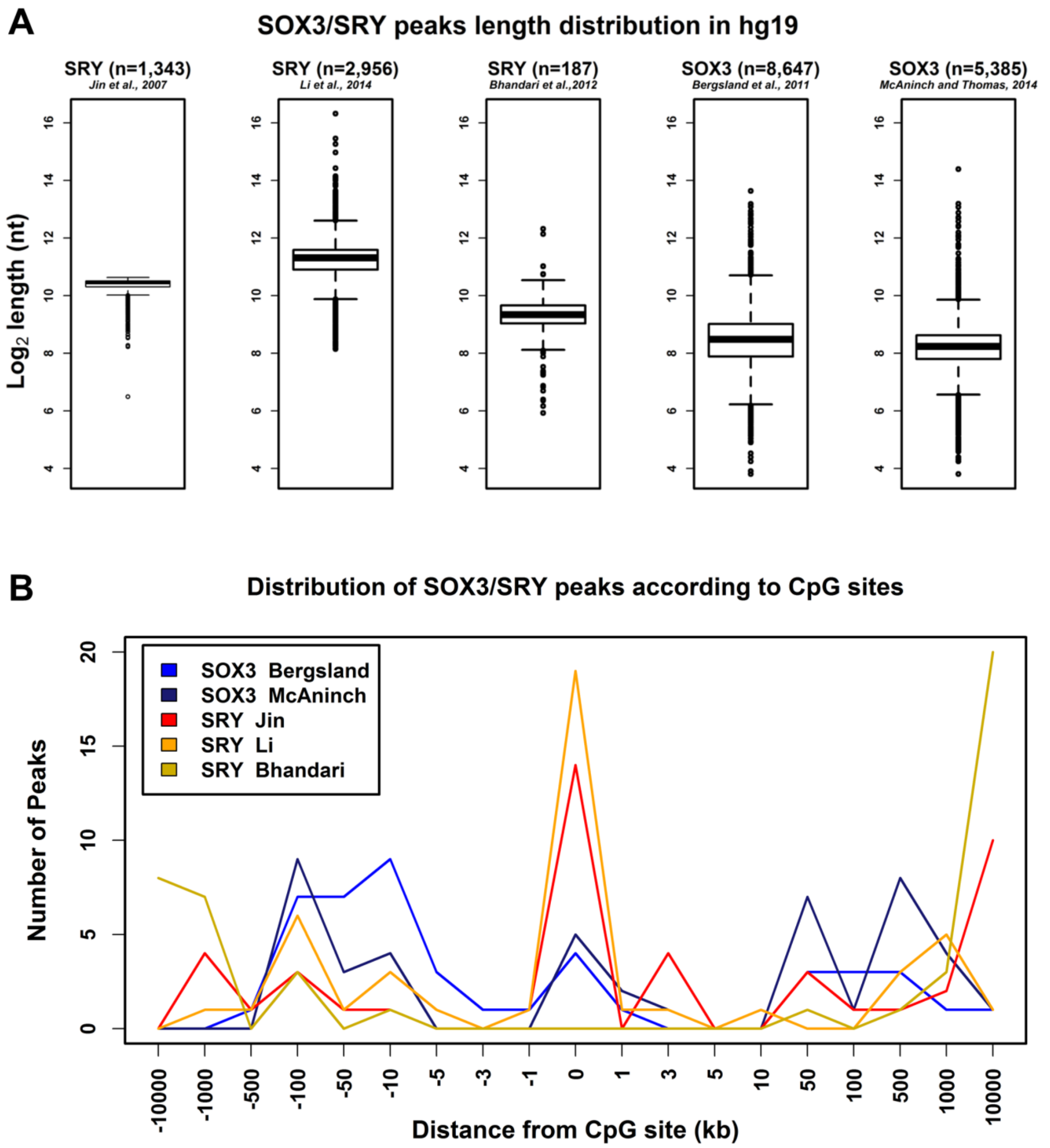

ChIP-seq SOX3 BERGSLAND et al., 2013
ChIP-seq SOX3 McANINCH; THOMAS, 2014
ChIP-Chip Li et al., 2014
ChIP-chip SRY BHANDARI et al., 2012
ChIP-chip SRY JIN et al., 2012
FINDING ORTHOLOGOUS IN HUMAN (DIOPT)
BRAIN SPAN
FILTERING BRAIN EXPRESSION
VOINEAGU et al., 2011
GUPTA et al., 2014
HORMONAL REGULATION MSigPB
METHYLATION SPIERS et al., 2015
GENE EXPR. KANG et al., 2011
SFARI
PGC
DDD, 2017
SANDERS et al., 2015
LESCH et al., 2012
PARIKSHAK et al., 2013
CHOI et al., 2016
TO EVALUATE CO EXPRESSION MODULES ASD vs CONTROL
TO EVALUATE SEX BIAS MODULATION AND HORMONAL REGULATION
TO EVALUATE ASD DATABASES AND STUDIES
TO EVALUATE HISTONE MARKS
TO EVALUATE BRAIN DEVELOPMENT DATA

Supplementary Figure S5: Flowchart of analyses performed in humans datasets. Cylinders represent database entries and trapezes represent datasets used in the analysis. Each input is assigned according to study or database and type of data. Blue squares describe the type of analysis.

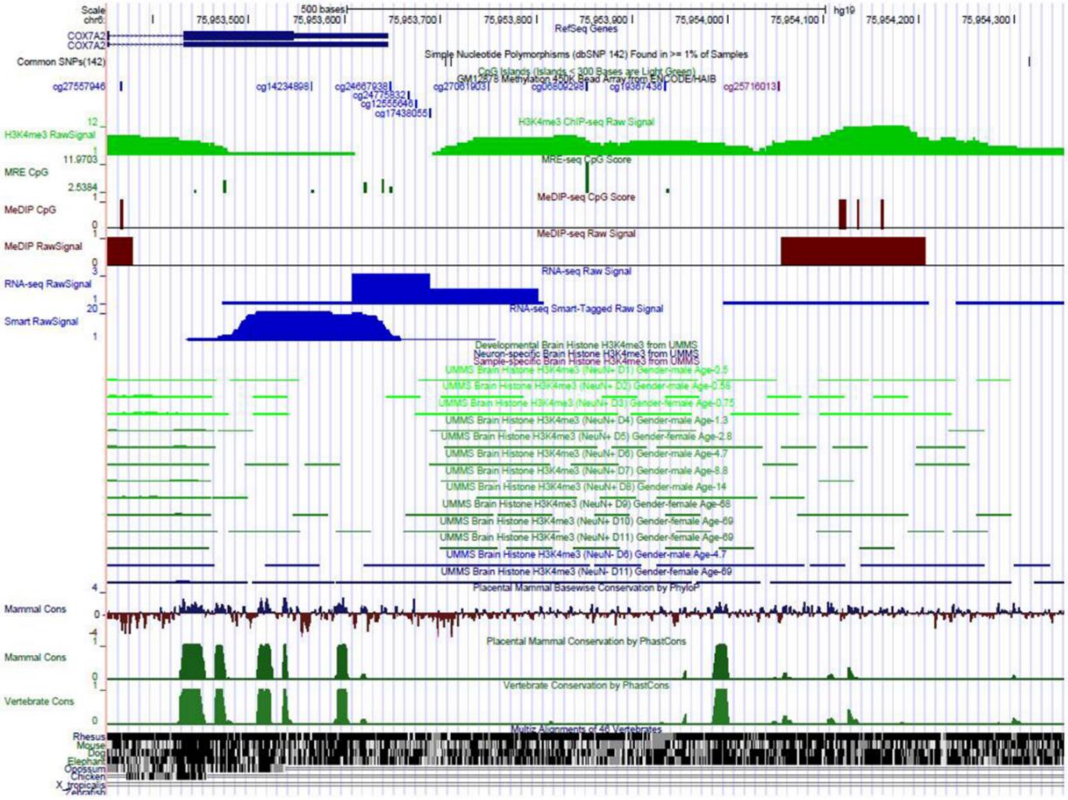
Brain Development Modules:



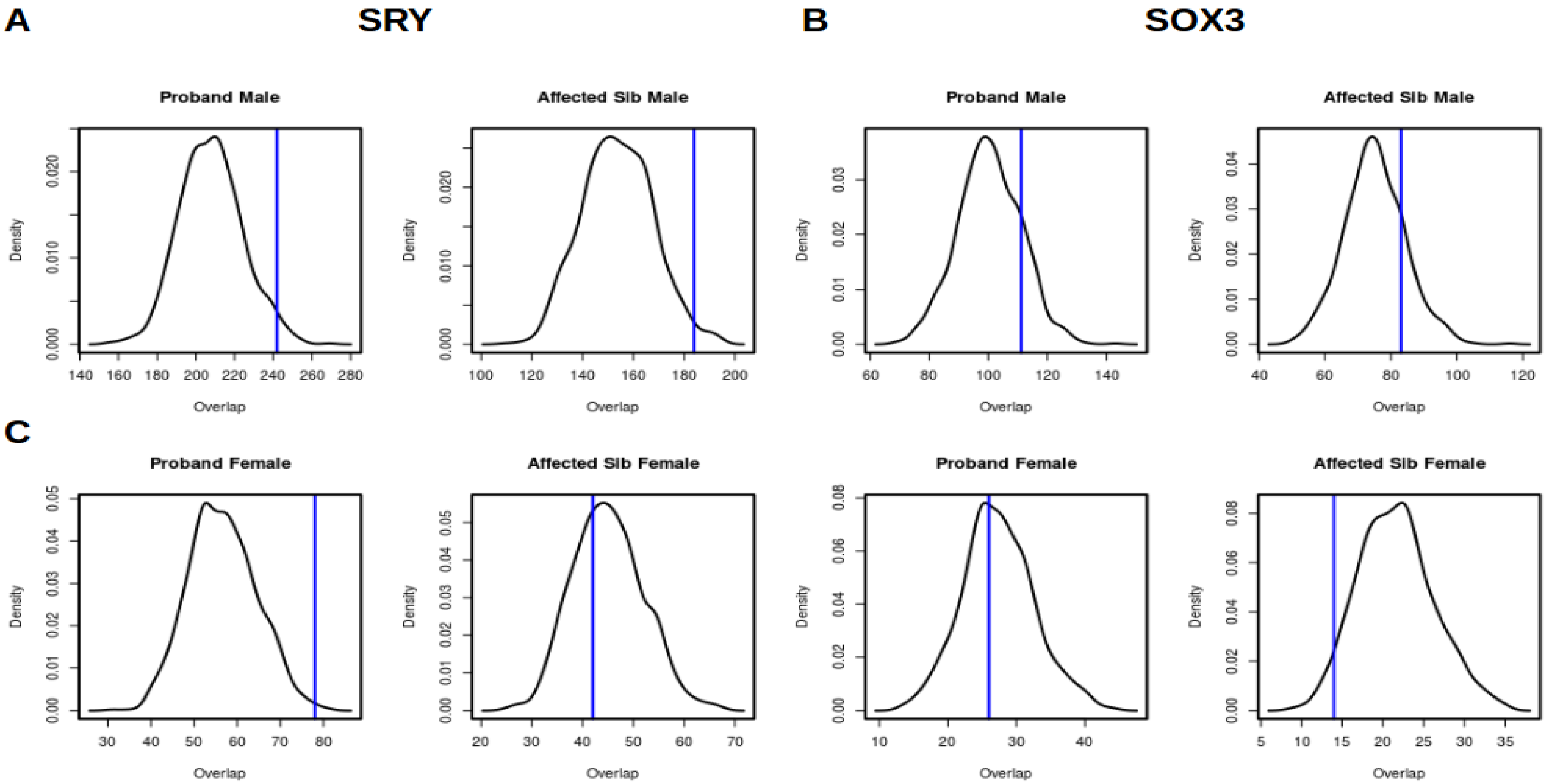
Supplementary Figure S6: Scatter plot of module eigengene values during neocortical development (Parikshak et al., 2013) . The module eigengene values are defined as the first principal component (PC) of the expression matrix. Module eigengene values are represented in the y-axis, and age is provided in log$_2$days. Different life periods were scaled from light to dark blue colors: early fetal (8-12 PCW), early-mid fetal (13-17 PCW), late midfetal (18-22 PCW), late fetal (24-35 PCW), neonatal/early infancy (up to 7 months) and late infancy (up to 1 yr.). The red line indicates birth. The upper four graphics show the modules that were enriched with SOX3/SRY target genes dataset only (M2, M3, M5 and M14), which consists of genes with higher expression during the early- and mid-fetal periods. The middle panel shows three gene modules enriched for (M11, M1, M18), consisting of genes with higher expression during the early fetal period and M1 and M18 after birth. The left lower panel shows M16 is comprised by genes with higher expression after birth. The right lower panel shows M8, which is enriched in SOX3/SRY and ER with higher expression in early fetal period.



Supplementary Figure S7: Distribution of SOX3 and SRY peaks' size in the hg19 assembly. First, all SRY/SOX3 peaks from a different animal model were converted to the hg19 assembly. **A-)** Boxplot showing the distributions of sizes ($\log_2$) for data from five studies (3 from SRY ChIP-chip and 2 from SOX3 ChIP-seq). SRY peaks had a median size of 1,400 nt (10.45 in $\log_2$) (Jin, O'Geen, Iyengar, Green, & Farnham, 2007), 2557 nt (11.32 in $\log_2$) (Li, Zheng, & Lau, 2014) and 650 nt (9.34 in $\log_2$) (Bhandari, Haque, & Skinner, 2012), SOX3 peaks had median sizes of 358 nt (8.48 in $\log_2$) (Bergsland et al., 2011) and 302 nt (8.24 in $\log_2$) (McAninch & Thomas, 2014). The number of peaks for each study is displayed in the top of the boxplots. Note that the median size of the SRY peaks (median joining all studies 1859 nt) is larger than that of the SOX3 peaks. **B-)** Distribution of the distance according to 45 CpG sites. The x-axis represents the distance according to the hg19 coordinate of each CpG site from the peak, and the y-axis is the number of the closest peaks (SRY or SOX3). SOX3 distribution is represented by blue (Bergsland et al., 2011) and midnight blue (McAninch & Thomas, 2014) lines, and SRY distribution is represented by red (Jin et al., 2007), orange (Li et al., 2014) and yellow (Bhandari et al., 2012) lines.


Scale
chr6:
500 bases
hg19
75,953,500
75,953,600
75,953,700
75,953,800
75,953,900
75,954,000
75,954,100
75,954,200
75,954,300
RefSeq Genes
COX7A2
COX7A2
Common SNPs(142)
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
CpG Islands (Islands < 300 Bases are Light Green)
GM12878 Methylation 450K Bead Array from ENCODE/HAIB
cg27557946
cg14234898
cg24667938
cg24775832
cg12555646
cg17438055
cg27061903
cg06809298
cg19387436
cg25716013
12
H3K4me3 RawSignal
1
H3K4me3 ChIP-seq Raw Signal
11.9703
MRE CpG
2.5384
MRE-seq CpG Score
1
MeDIP CpG
0
MeDIP-seq CpG Score
1
MeDIP RawSignal
0
MeDIP-seq Raw Signal
3
RNA-seq RawSignal
1
RNA-seq Raw Signal
20
Smart RawSignal
1
RNA-seq Smart-Tagged Raw Signal
Developmental Brain Histone H3K4me3 from UMMS
Neuron-specific Brain Histone H3K4me3 from UMMS
Sample-specific Brain Histone H3K4me3 from UMMS
UMMS Brain Histone H3K4me3 (NeuN+ D1) Gender-male Age-0.5
UMMS Brain Histone H3K4me3 (NeuN+ D2) Gender-male Age-0.58
UMMS Brain Histone H3K4me3 (NeuN+ D3) Gender-female Age-0.75
UMMS Brain Histone H3K4me3 (NeuN+ D4) Gender-male Age-1.3
UMMS Brain Histone H3K4me3 (NeuN+ D5) Gender-female Age-2.8
UMMS Brain Histone H3K4me3 (NeuN+ D6) Gender-male Age-4.7
UMMS Brain Histone H3K4me3 (NeuN+ D7) Gender-male Age-8.8
UMMS Brain Histone H3K4me3 (NeuN+ D8) Gender-male Age-14
UMMS Brain Histone H3K4me3 (NeuN+ D9) Gender-female Age-68
UMMS Brain Histone H3K4me3 (NeuN+ D10) Gender-female Age-69
UMMS Brain Histone H3K4me3 (NeuN+ D11) Gender-female Age-69
UMMS Brain Histone H3K4me3 (NeuN- D6) Gender-male Age-4.7
UMMS Brain Histone H3K4me3 (NeuN- D11) Gender-female Age-69
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
1
Mammal Cons
0
Placental Mammal Conservation by PhastCons
1
Vertebrate Cons
0
Vertebrate Conservation by PhastCons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish

Supplementary Figure S8: UCSC genome browser of cg06809298 and *COX7A2* gene. The two isoforms of the RefSeq gene are shown in dark blue at the top of the figure. The CpG site of interest is represented by the red square. The *USCS Brain DNA Methylation* track consists of H3K4me3 ChIP-seq (green), MRE-seq (dark green), MeDIP-seq (brown) and RNA-seq (blue) of brain samples from a 57-year-old male individual. The histone mark is evidence of an active regulatory promoter region. The MRE-seq shows that this CpG site (cg06809298) is hypomethylated. The MeDIP-seq assay confirms this finding because there is no raw signal of the immunoprecipitation of methylated CpG. In addition, there is evidence of transcription in both RNA-seq experiments. The other histone marks (*UMMS Brain Histone H3K4me3 ChIP-seq*) from 11 prefrontal brain samples (6 male samples and 5 female samples aged 0.5-69 yrs), and the signal is proportional to the remapping of reads to hg19. The scale of these signal ranges from 0-50. Thus, little signal of H3K4me3 is shown in these graphics. At the position of the CpG site (cg06809298), little conservation is observed in the PhyloP or PhastCons Scores (mammal or placental).



Supplementary Figure S9: Enrichment analyses of DNVs under SRY and SOX3 peaks. Each graphic represents the density distribution of overlapping number of DNVs across 1,000 random groups. The x axis shows the number of overlapping and y axis the density. Blue line shows the number of overlapping genes of the corresponded experiment.

SOX3 and SRY peaks distribution according to *De novo* variantions in whole genome sequencing in ASD samples.

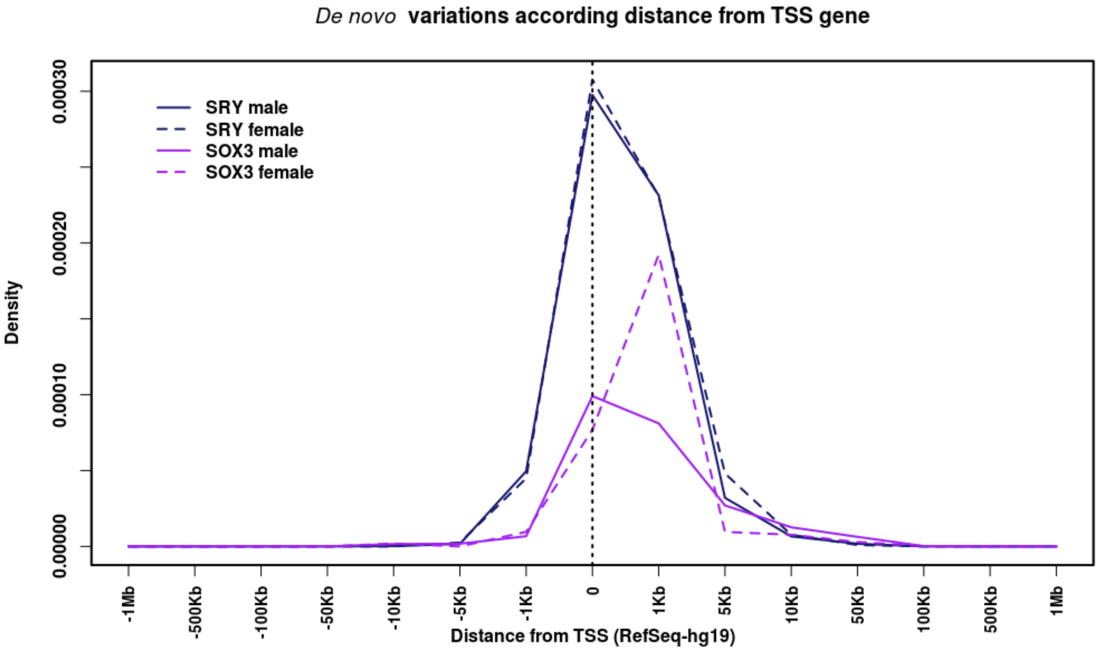


Figure S10: *De novo* variations according to TSS gene. Distribution of the distance of 242 DNVs over SRY peaks and 78 DNVs over SOX3 peaks according to TSS genes. The x-axis represents the distance according to the hg19 coordinate of each DNV from the TSS gene, and the y-axis is the number of overlapping peaks (SRY or SOX3). SOX3 frequency is represented by purple lines and SRY by blue lines, solid and dashed lines represent male and female ASD samples. The DNVs were obtained from whole genome sequencing (Yuen et al., 2017).

REFERENCES:

Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., & Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes and Development*, *25*(23), 2453–2464. https://doi.org/10.1101/gad.176008.111

Bhandari, R. K., Haque, M. M., & Skinner, M. K. (2012). Global Genome Analysis of the Downstream Binding Targets of Testis Determining Factor SRY and SOX9. *PLoS ONE, 7*. https://doi.org/10.1371/journal.pone.0043380

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. *Schrödinger LLC Wwwpymolorg, Version 1.*, http://www.pymol.org. https://doi.org/citeulike-article-id:240061

EMBL, SIB Swiss Institute of Bioinformatics, & Protein Information Resource (PIR). (2013). UniProt. In *Nucleic acids research* (p. 41: D43-D47).

Jin, V. X., O'Geen, H., Iyengar, S., Green, R., & Farnham, P. J. (2007). Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Research*, *17*, 807–817. https://doi.org/10.1101/gr.6006107

Lesch, B. J., Silber, S. J., McCarrey, J. R., & Page, D. C. (2016). Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nature Genetics*, *48*(8), 888–894. https://doi.org/10.1038/ng.3591

Li, Y., Zheng, M., & Lau, Y.-F. (2014). The Sex-Determining Factors SRY and SOX9 Regulate Similar Target Genes and Promote Testis Cord Formation during Testicular Differentiation. *Cell Reports*, *8*(3), 723–733. https://doi.org/10.1016/j.celrep.2014.06.055

McAninch, D., & Thomas, P. (2014). Identification of highly conserved putative developmental enhancers bound by SOX3 in neural progenitors using ChIP-Seq. *PloS One*, *9*(11), e113361. https://doi.org/10.1371/journal.pone.0113361

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–17. https://doi.org/10.1006/jmbi.2000.4042

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., … Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, *155*(5), 1008–21. https://doi.org/10.1016/j.cell.2013.10.031

Yuen, R. K. C., Merico, D., Bookman, M., Howe, J. L., Thiruvahindrapuram, B., Patel, R. V., … Scherer, S. W. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience*, *20*(4), 602–611. https://doi.org/10.1038/nn.4524