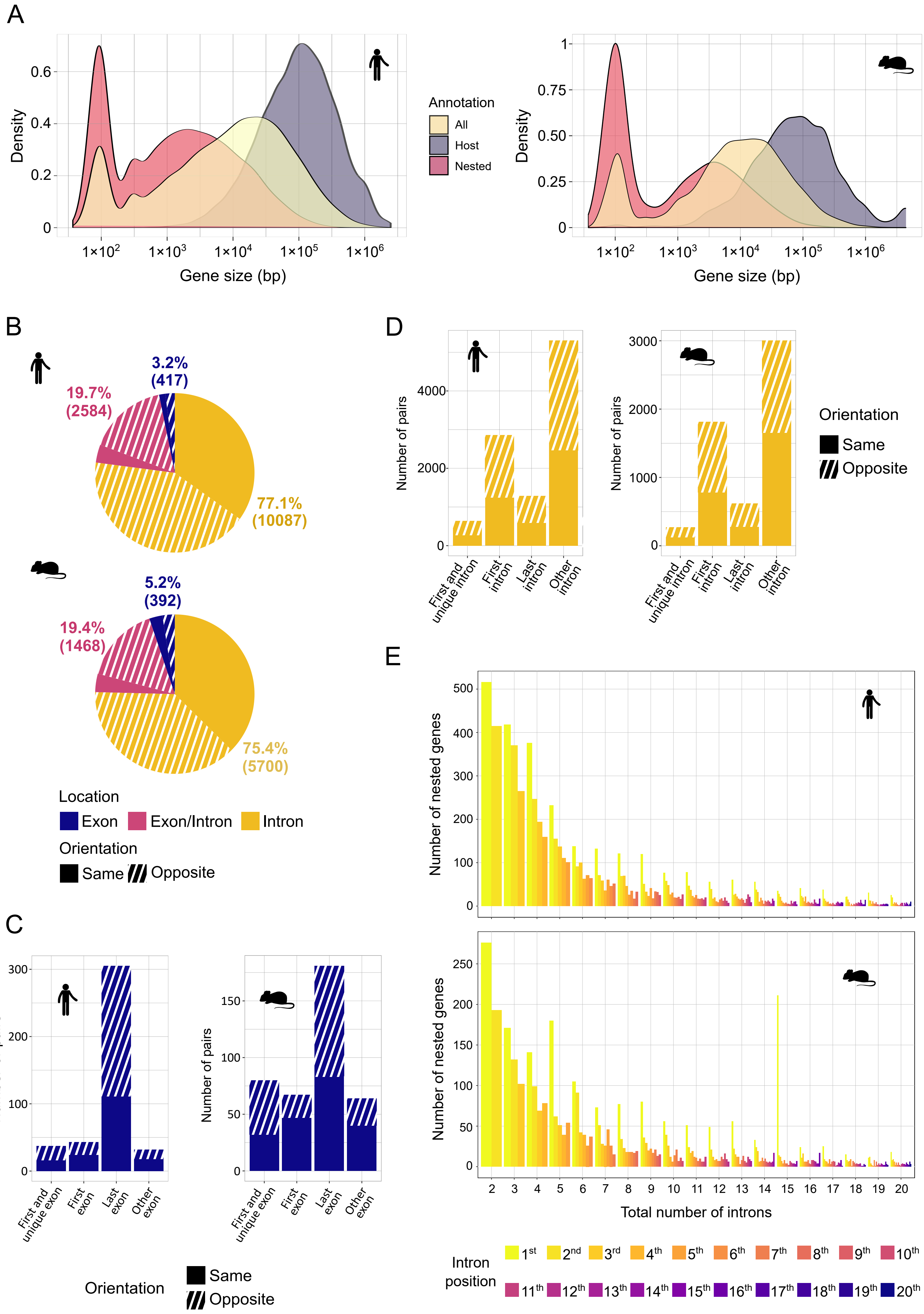
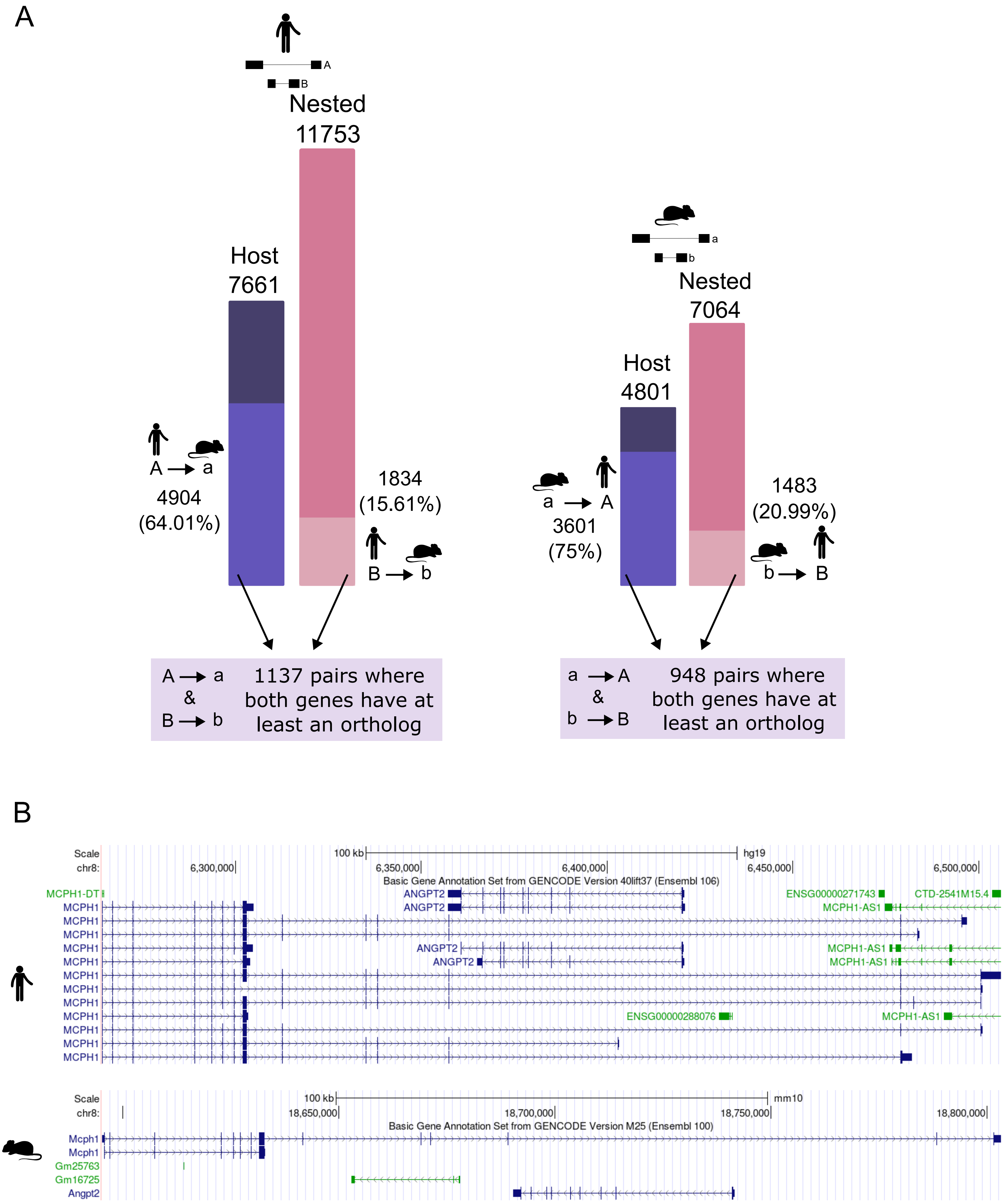
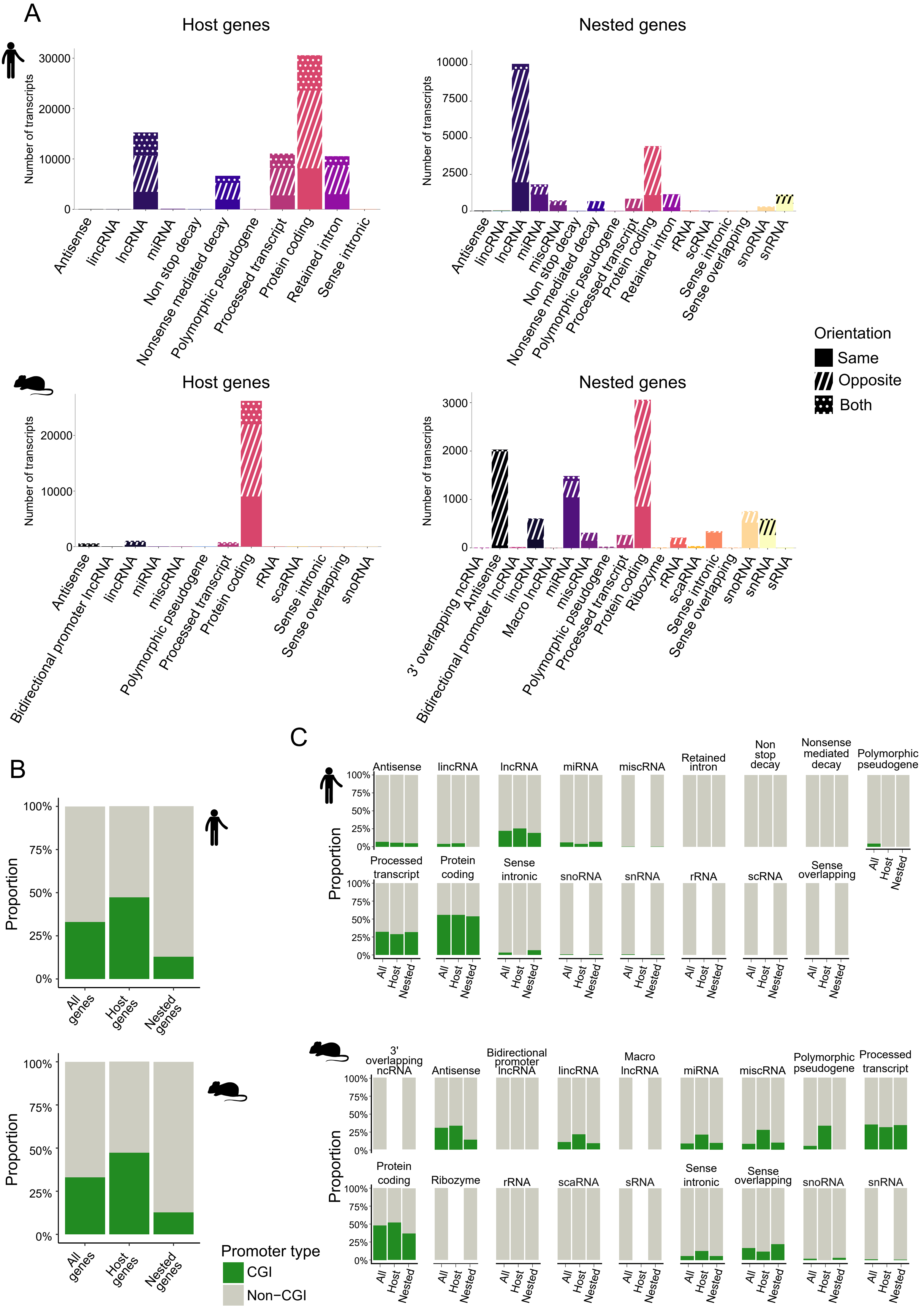

A
Density
0.6
0.4
0.2
0
1×10^2
1×10^3
1×10^4
1×10^5
1×10^6
Gene size (bp)
Annotation
All
Host
Nested
1
0.75
0.50
0.25
B
19.7%
(2584)
3.2%
(417)
77.1%
(10087)
19.4%
(1468)
5.2%
(392)
75.4%
(5700)
Location
Exon
Exon/Intron
Intron
Orientation
Same
Opposite
C
Number of pairs
300
200
100
150
50
First and unique exon
First exon
Last exon
Other exon
D
4000
2000
3000
1000
First and unique intron
First intron
Last intron
Other intron
E
Number of nested genes
500
400
300
250
Total number of introns
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Intron position
1st 2nd 3rd 4th 5th 6th 7th 8th 9th 10th
11th 12th 13th 14th 15th 16th 17th 18th 19th 20th

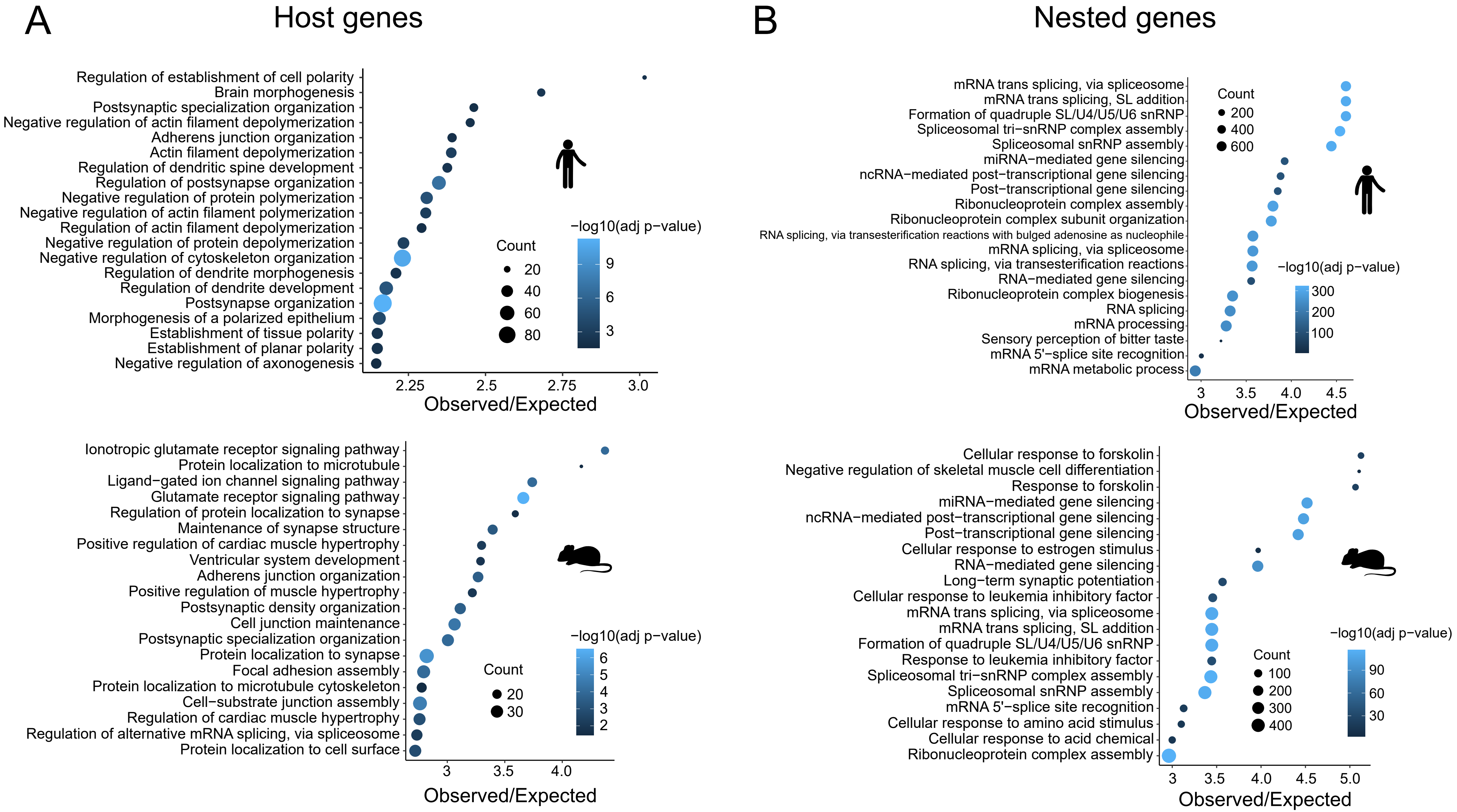
Supplementary Figure 1. Host/nested genes pairs in mouse and human shows similar characteristics. (A) Distribution of the genes' size for host genes, nested genes, and all genes. A base-10 log scale was used for the gene size. (B) Distribution of the locations of nested genes within their host genes according to their orientation to each other. (C) Detailed location of the nested genes when fully contained inside an exon. (D) Detailed location of the nested genes when fully contained inside an intron. (E) Total number of nested genes within the intron of their host gene relative to their 5′>3′ position. This was performed for host genes containing 2-20 introns.


A
A
B
Nested
11753
Host
7661
A → a
4904
(64.01%)
1834
(15.61%)
B → b
a
b
Nested
7064
Host
4801
a → A
3601
(75%)
1483
(20.99%)
b → B
A → a
&
B → b
1137 pairs where both genes have at least an ortholog
a → A
&
b → B
948 pairs where both genes have at least an ortholog
B
Scale
chr8:
100 kb
hg19
6,300,000
6,350,000
6,400,000
6,450,000
6,500,000
Basic Gene Annotation Set from GENCODE Version 40lift37 (Ensembl 106)
MCPH1-DT
MCPH1
ANGPT2
ENSG00000271743
CTD-2541M15.4
MCPH1-AS1
ENSG00000288076
Scale
chr8:
100 kb
mm10
18,650,000
18,700,000
18,750,000
18,800,000
Basic Gene Annotation Set from GENCODE Version M25 (Ensembl 100)
Mcph1
Gm25763
Gm16725
Angpt2

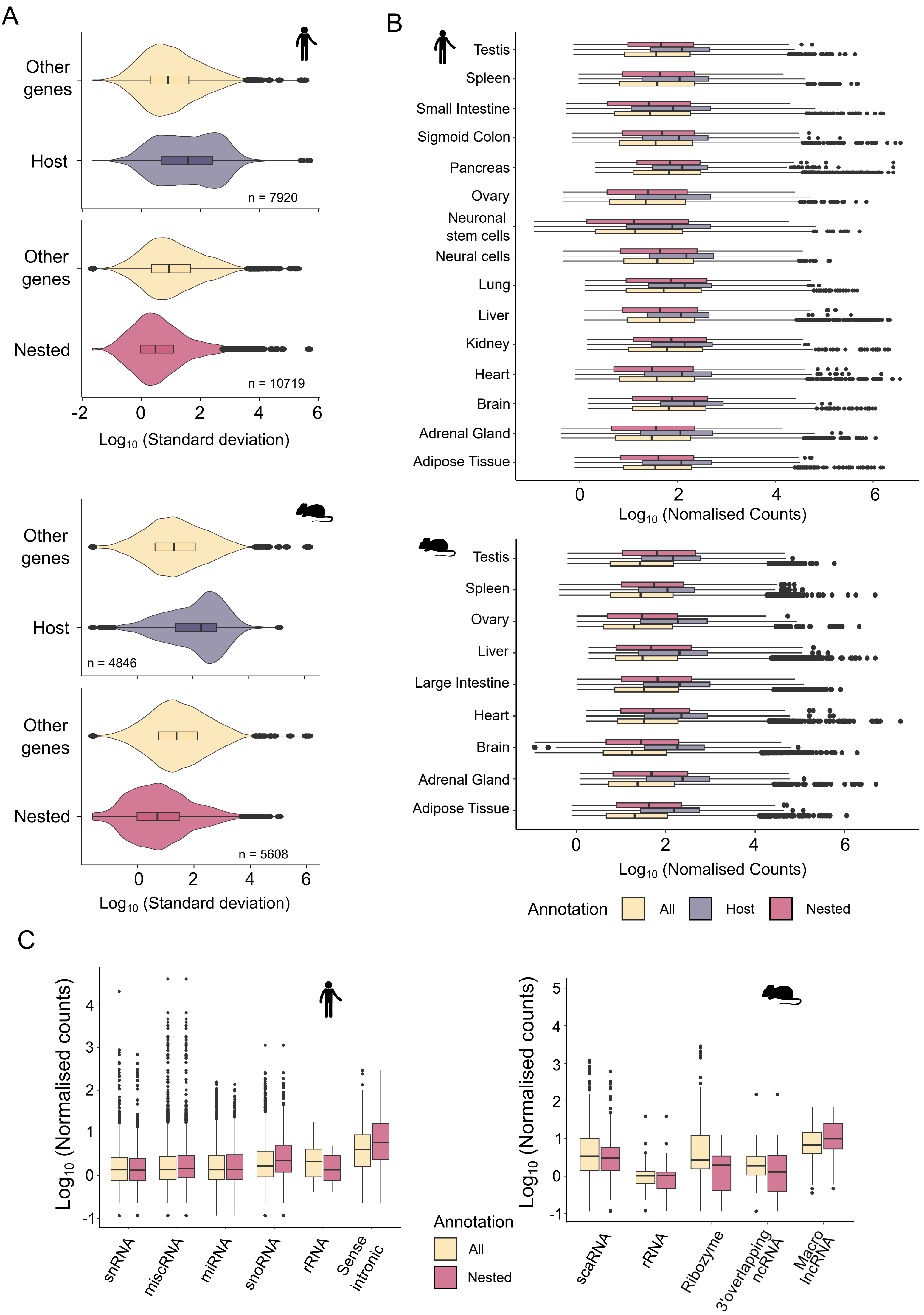
Supplementary Figure 2. Conservation of host and nested gene pairs between human and mouse genomes. (A) Host and nested gene conservation between mouse and human genomes. (B) Screenshots of the UCSC genome browser for an example of a host/nested genes pair conserved between mouse and human.


A
Host genes
Number of transcripts
30000
20000
10000
0
Antisense
lincRNA
lncRNA
miRNA
Non stop decay
Nonsense mediated decay
Polymorphic pseudogene
Processed transcript
Protein coding
Retained intron
Sense intronic
Nested genes
10000
7500
5000
2500
0
Antisense
lincRNA
lncRNA
miRNA
miscRNA
Non stop decay
Nonsense mediated decay
Polymorphic pseudogene
Processed transcript
Protein coding
Retained intron
rRNA
scRNA
Sense intronic
Sense overlapping
snoRNA
snRNA
Orientation
Same
Opposite
Both
Host genes
Number of transcripts
20000
10000
0
Antisense
Bidirectional promoter lncRNA
lincRNA
miRNA
miscRNA
Polymorphic pseudogene
Processed transcript
Protein coding
rRNA
scaRNA
Sense intronic
Sense overlapping
snoRNA
Nested genes
Number of transcripts
3000
2000
1000
0
3' overlapping ncRNA
Antisense
Bidirectional promoter lncRNA
lincRNA
Macro lncRNA
miRNA
miscRNA
Polymorphic pseudogene
Processed transcript
Protein coding
Ribozyme
rRNA
scaRNA
Sense intronic
Sense overlapping
snoRNA
snRNA
sRNA
B
Proportion
100%
75%
50%
25%
0%
All genes
Host genes
Nested genes
Proportion
100%
75%
50%
25%
0%
All genes
Host genes
Nested genes
Promoter type
CGI
Non−CGI
C
Proportion
Antisense
lincRNA
lncRNA
miRNA
miscRNA
Retained intron
Non stop decay
Nonsense mediated decay
Polymorphic pseudogene
Processed transcript
Protein coding
Sense intronic
snoRNA
snRNA
rRNA
scRNA
Sense overlapping
100%
75%
50%
25%
0%
All
Host
Nested
Proportion
3' overlapping ncRNA
Antisense
Bidirectional promoter lncRNA
lincRNA
Macro lncRNA
miRNA
miscRNA
Polymorphic pseudogene
Processed transcript
Protein coding
Ribozyme
rRNA
scaRNA
sRNA
Sense intronic
Sense overlapping
snoRNA
snRNA
100%
75%
50%
25%
0%
All
Host
Nested

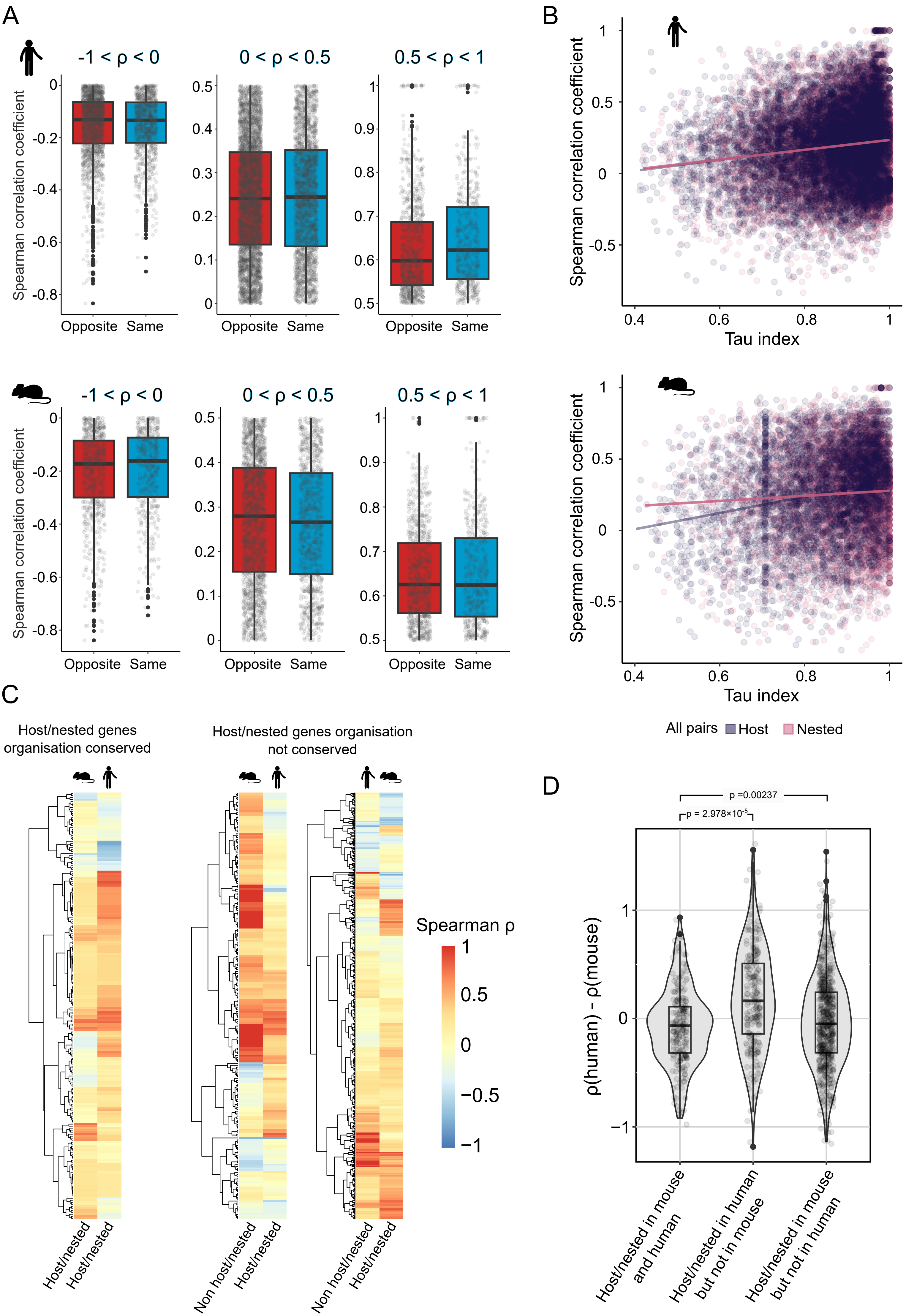
Supplementary Figure 3. Host/nested genes pair’s biotypes, functional enrichment analysis and expression dispersion. (A) Number of host and nested genes associated with each biotype according to their orientation in pairs. Both refers to genes involved in multiple pairs with different relative orientation. (B) Proportion of host, nested and all genes that are associated with a CpG island (CGI) promoter. (C) Proportion of host, nested and all gene transcripts that are associated with a CGI promoter across all biotypes examined.


A
Host genes
Regulation of establishment of cell polarity
Brain morphogenesis
Postsynaptic specialization organization
Negative regulation of actin filament depolymerization
Adherens junction organization
Actin filament depolymerization
Regulation of dendritic spine development
Regulation of postsynapse organization
Negative regulation of protein polymerization
Negative regulation of actin filament polymerization
Regulation of actin filament depolymerization
Negative regulation of protein depolymerization
Negative regulation of cytoskeleton organization
Regulation of dendrite morphogenesis
Regulation of dendrite development
Postsynapse organization
Morphogenesis of a polarized epithelium
Establishment of tissue polarity
Establishment of planar polarity
Negative regulation of axonogenesis
Count
20
40
60
80
−log10(adj p−value)
9
6
3
2.25
2.5
2.75
3.0
Observed/Expected
Ionotropic glutamate receptor signaling pathway
Protein localization to microtubule
Ligand−gated ion channel signaling pathway
Glutamate receptor signaling pathway
Regulation of protein localization to synapse
Maintenance of synapse structure
Positive regulation of cardiac muscle hypertrophy
Ventricular system development
Adherens junction organization
Positive regulation of muscle hypertrophy
Postsynaptic density organization
Cell junction maintenance
Postsynaptic specialization organization
Protein localization to synapse
Focal adhesion assembly
Protein localization to microtubule cytoskeleton
Cell−substrate junction assembly
Regulation of cardiac muscle hypertrophy
Regulation of alternative mRNA splicing, via spliceosome
Protein localization to cell surface
Count
20
30
−log10(adj p−value)
6
5
4
3
2
3
3.5
4.0
Observed/Expected
B
Nested genes
mRNA trans splicing, via spliceosome
mRNA trans splicing, SL addition
Formation of quadruple SL/U4/U5/U6 snRNP
Spliceosomal tri−snRNP complex assembly
Spliceosomal snRNP assembly
miRNA−mediated gene silencing
ncRNA−mediated post−transcriptional gene silencing
Post−transcriptional gene silencing
Ribonucleoprotein complex assembly
Ribonucleoprotein complex subunit organization
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
RNA splicing, via transesterification reactions
RNA−mediated gene silencing
Ribonucleoprotein complex biogenesis
RNA splicing
mRNA processing
Sensory perception of bitter taste
mRNA 5'−splice site recognition
mRNA metabolic process
Count
200
400
600
−log10(adj p−value)
300
200
100
3
3.5
4.0
4.5
Observed/Expected
Cellular response to forskolin
Negative regulation of skeletal muscle cell differentiation
Response to forskolin
miRNA−mediated gene silencing
ncRNA−mediated post−transcriptional gene silencing
Post−transcriptional gene silencing
Cellular response to estrogen stimulus
RNA−mediated gene silencing
Long−term synaptic potentiation
Cellular response to leukemia inhibitory factor
mRNA trans splicing, via spliceosome
mRNA trans splicing, SL addition
Formation of quadruple SL/U4/U5/U6 snRNP
Response to leukemia inhibitory factor
Spliceosomal tri−snRNP complex assembly
Spliceosomal snRNP assembly
mRNA 5'−splice site recognition
Cellular response to amino acid stimulus
Cellular response to acid chemical
Ribonucleoprotein complex assembly
Count
100
200
300
400
−log10(adj p−value)
90
60
30
3
3.5
4.0
4.5
5.0
Observed/Expected

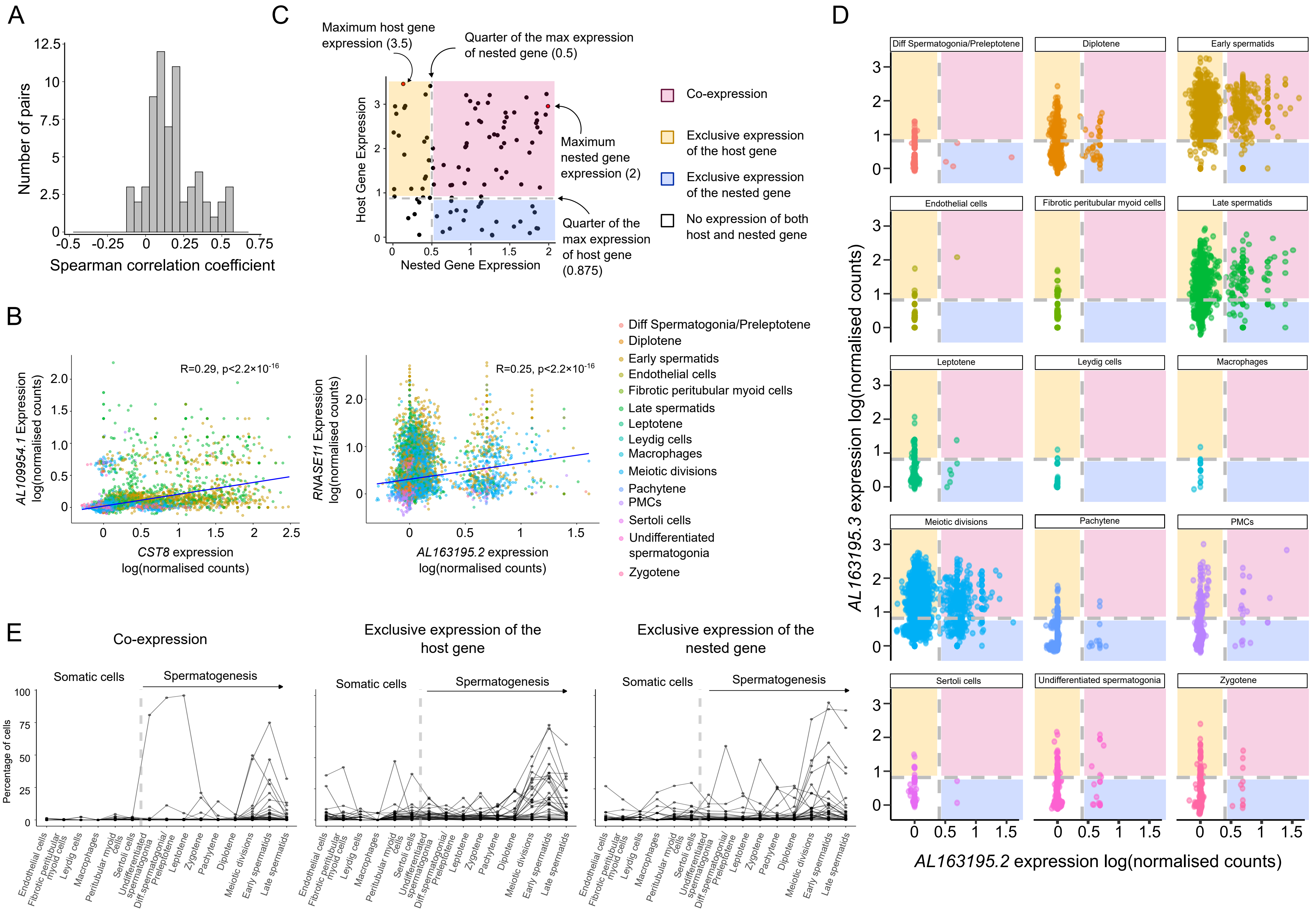
Supplementary Figure 4. Host/nested gene pairs functional enrichment analysis. (A) Gene sets enrichment analysis for host gene lists. Only the top 20 (according to the observed/expected ratio) significant biological processes are represented. (B) Same for nested genes.


A
Other genes
Host
n = 7920
Other genes
Nested
n = 10719
-2
0
2
4
6
Log10 (Standard deviation)
Other genes
Host
n = 4846
Other genes
Nested
n = 5608
0
2
4
6
Log10 (Standard deviation)
B
Testis
Spleen
Small Intestine
Sigmoid Colon
Pancreas
Ovary
Neuronal stem cells
Neural cells
Lung
Liver
Kidney
Heart
Brain
Adrenal Gland
Adipose Tissue
0
2
4
6
Log10 (Nomalised Counts)
Testis
Spleen
Ovary
Liver
Large Intestine
Heart
Brain
Adrenal Gland
Adipose Tissue
0
2
4
6
Log10 (Nomalised Counts)
Annotation
All
Host
Nested
C
Log10 (Normalised counts)
-1
0
1
2
3
4
snRNA
miscRNA
miRNA
snoRNA
rRNA
Sense intronic
Annotation
All
Nested
Log10 (Normalised counts)
-1
0
1
2
3
4
5
scaRNA
rRNA
Ribozyme
3'overlapping ncRNA
Macro lncRNA

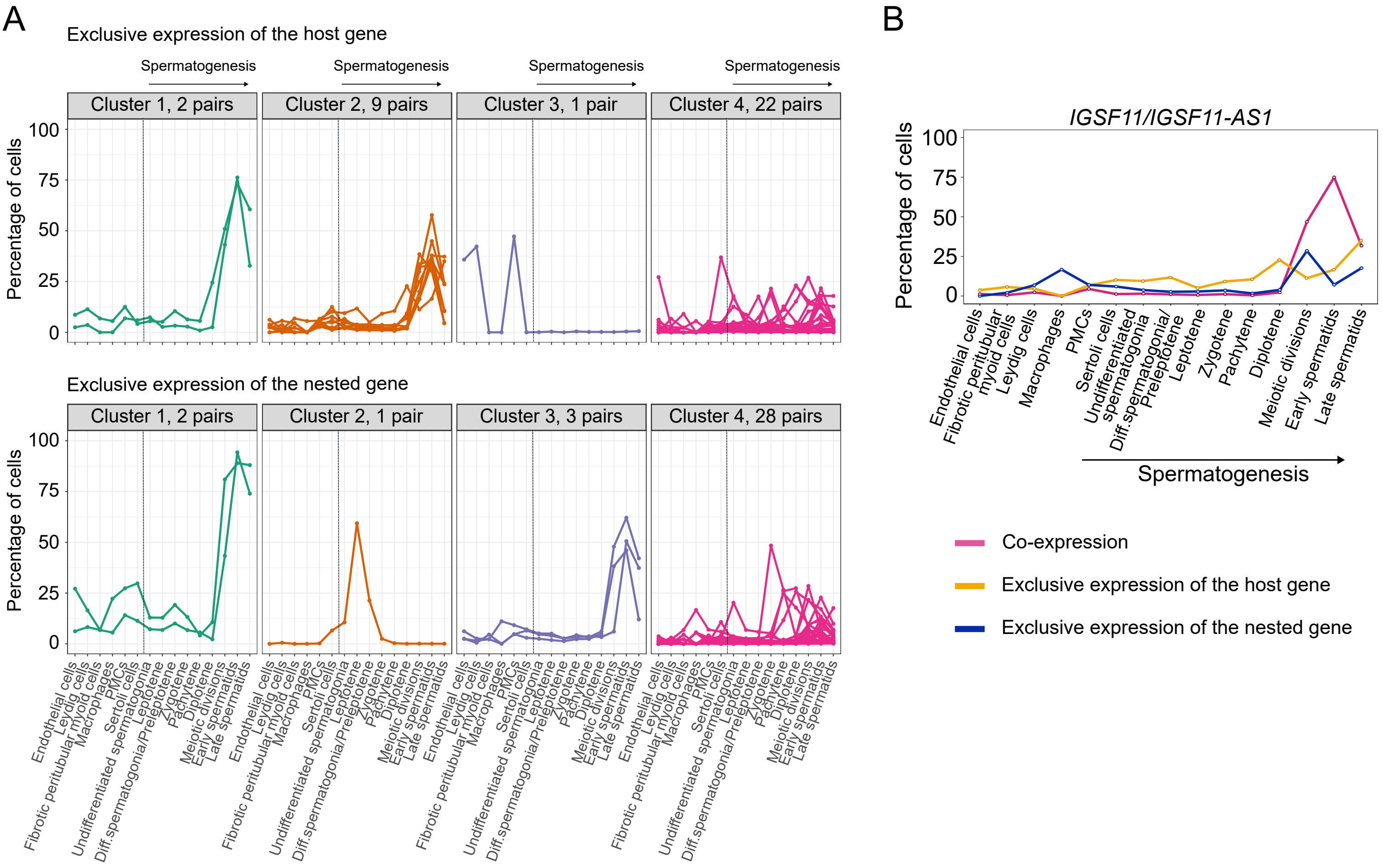
Supplementary Figure 5. Distribution of the expression of host/nested genes in ENCODE RNA sequencing datasets. (A) Distribution of standard deviation between the normalized counts across tissues Across 1000 matched size random samples, application of a Kolmogorov–Smirnov test to these distributions always resulted in a $p<0.05$. The figure shows one representative matched sample comparison. (B) Distribution of the expression of all, host and nested genes for protein coding genes across tissues, in human and mouse transcriptomes (C) Distribution of the expression of all and nested genes for biotypes unique to nested genes, in human and mouse transcriptomes.


A
-1 < ρ < 0
0 < ρ < 0.5
0.5 < ρ < 1
Spearman correlation coefficient
Opposite
Same
B
Spearman correlation coefficient
Tau index
All pairs
Host
Nested
C
Host/nested genes organisation conserved
Host/nested genes organisation not conserved
Host/nested
Non host/nested
Spearman ρ
1
0.5
0
−0.5
−1
D
p =0.00237
p = 2.978×10-5
ρ(human) - ρ(mouse)
Host/nested in mouse and human
Host/nested in human but not in mouse
Host/nested in mouse but not in human

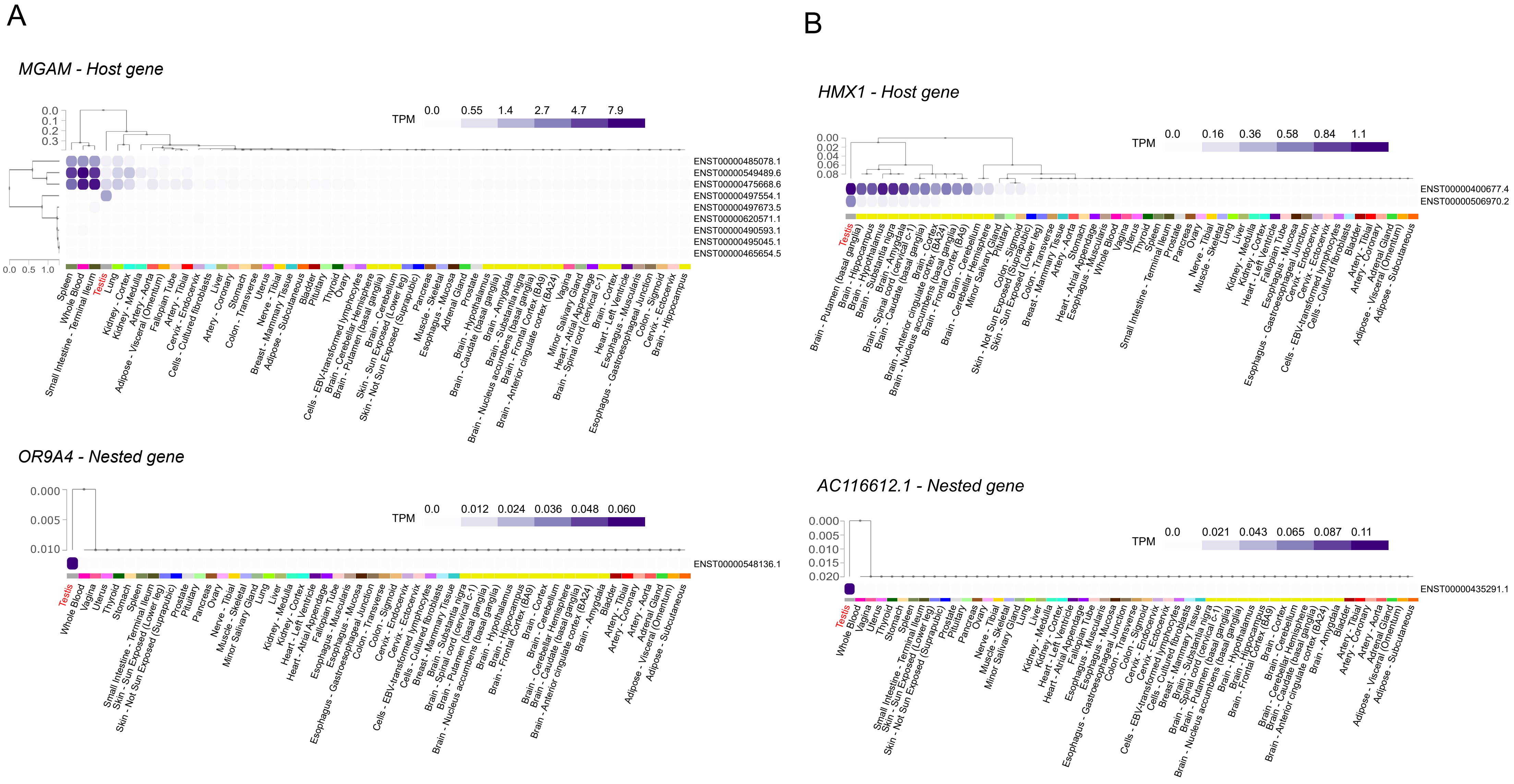
Supplementary Figure 6. Expression correlation between host and nested genes and its association with tissue specificity and conservation. (A) Distribution of Spearman's rank correlation coefficient between pairs according to their orientation and depending on the range of ρ. (B) Correlation between the τ index and Spearman's rank correlation coefficient across all host nested gene pairs. (C) Heatmap showing ρ between conserved host/nested gene pairs and then, between pairs of genes that were conserved but were only host/nested in one species. (D) Difference in ρ between human and mouse, for conserved host/nested gene pairs and then, between pairs of genes that were conserved but were only host/nested in one species. Levene Test was used to determine differences in variance across the values, p-values indicated above the graph.


A
Number of pairs
Spearman correlation coefficient
C
Maximum host gene expression (3.5)
Quarter of the max expression of nested gene (0.5)
Maximum nested gene expression (2)
Quarter of the max expression of host gene (0.875)
Host Gene Expression
Nested Gene Expression
Co-expression
Exclusive expression of the host gene
Exclusive expression of the nested gene
No expression of both host and nested gene
B
R=0.29, p<2.2×10^-16
AL109954.1 Expression log(normalised counts)
CST8 expression log(normalised counts)
R=0.25, p<2.2×10^-16
RNASE11 Expression log(normalised counts)
AL163195.2 expression log(normalised counts)
Diff Spermatogonia/Preleptotene
Diplotene
Early spermatids
Endothelial cells
Fibrotic peritubular myoid cells
Late spermatids
Leptotene
Leydig cells
Macrophages
Meiotic divisions
Pachytene
PMCs
Sertoli cells
Undifferentiated spermatogonia
Zygotene
D
AL163195.3 expression log(normalised counts)
AL163195.2 expression log(normalised counts)
E
Co-expression
Exclusive expression of the host gene
Exclusive expression of the nested gene
Somatic cells
Spermatogenesis
Percentage of cells
Endothelial cells
Fibrotic peritubular myoid cells
Leydig cells
Macrophages
Peritubular myoid cells
Sertoli cells
Undifferentiated spermatogonia
Diff.spermatogonia/ Preleptotene
Leptotene
Zygotene
Pachytene
Diplotene
Meiotic divisions
Early spermatids
Late spermatids

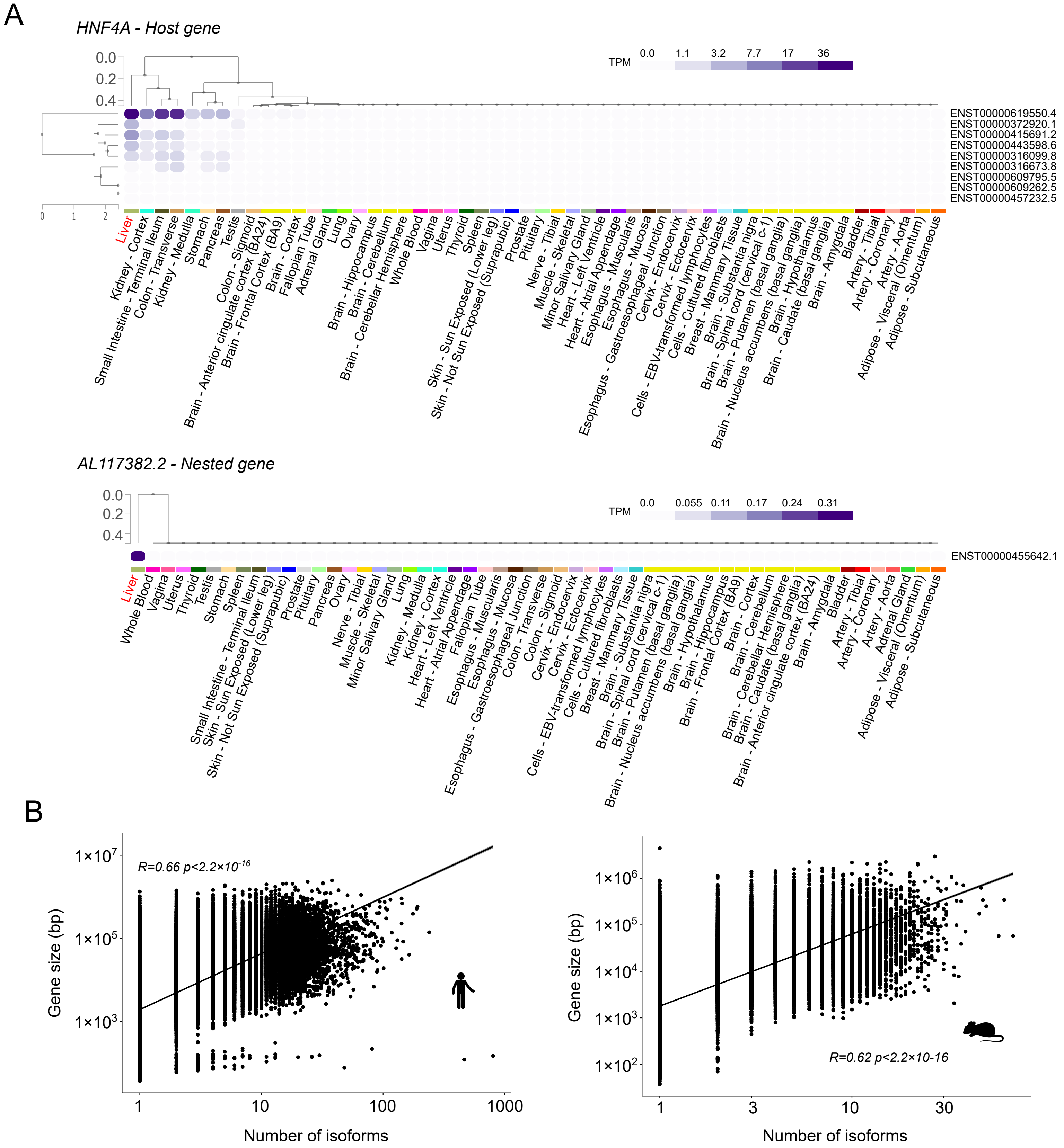
Supplementary Figure 7. Host and nested genes with high coexpression in testis can exhibit dynamic coexpression patterns during spermatogenesis. (A) Distribution of the Spearman's rank correlation coefficient values between expression of the host and nested genes across testis single cells (B) Scatter plots showing examples of host and nested gene expression and Spearman's rank correlation across single cells in testis for *AL109954.1/CST8*, R=0.29 and *RNASE11/AL163195.2* R= 0.25 (C) Schematic representation of the filtering methodology applied to single cells to determine proportion of cells which co-express a host and nested gene pair, or only express one of the partner. In this randomised mock example, host (0.875) and nested (0.5) gene expression thresholds were used to calculate that 46% of single cells co-express the gene pair. (D) Scatter plots showing examples of host and nested genes coexpression in the different testis cell types for the pair *AL163195.3/AL163195.2* (E) Percentage of cells co-expressing host and nested gene pairs or expressing only one of the partners across spermatogenesis for the 34 pairs with tissue specific expression in testis and robust expression data in the scRNA-seq.


A
Exclusive expression of the host gene
Spermatogenesis
Cluster 1, 2 pairs
Cluster 2, 9 pairs
Cluster 3, 1 pair
Cluster 4, 22 pairs
Percentage of cells
100
75
50
25
0
Exclusive expression of the nested gene
Cluster 1, 2 pairs
Cluster 2, 1 pair
Cluster 3, 3 pairs
Cluster 4, 28 pairs
Endothelial cells
Leydig cells
Fibrotic peritubular myoid cells
Macrophages
PMCs
Sertoli cells
Undifferentiated spermatogonia
Diff.spermatogonia/Preleptotene
Leptotene
Zygotene
Pachytene
Diplotene
Meiotic divisions
Early spermatids
Late spermatids
B
IGSF11/IGSF11-AS1
Percentage of cells
Endothelial cells
Fibrotic peritubular myoid cells
Leydig cells
Macrophages
PMCs
Sertoli cells
Undifferentiated spermatogonia
Diff.spermatogonial Preleptotene
Leptotene
Zygotene
Pachytene
Diplotene
Meiotic divisions
Early spermatids
Late spermatids
Spermatogenesis
Co-expression
Exclusive expression of the host gene
Exclusive expression of the nested gene

Supplementary Figure 8. Profiles of host and nested genes expression during spermatogenesis. (A) Proportion of cells which only express host or nested gene across spermatogenesis. Four distinct coexpression profiles were determined by *k*-means clustering. (B) Line plot showing the profile of expression for the IGSF11/IGSF11-AS1 gene pair calculated using the method detailed in Supplemental_Fig_7C.


A
MGAM - Host gene
TPM
0.0
0.55
1.4
2.7
4.7
7.9
ENST00000485078.1
ENST00000549489.6
ENST00000475668.6
ENST00000497554.1
ENST00000497673.5
ENST00000620571.1
ENST00000490593.1
ENST00000495045.1
ENST00000465654.5
Testis
OR9A4 - Nested gene
TPM
0.0
0.012
0.024
0.036
0.048
0.060
ENST00000548136.1
Testis
B
HMX1 - Host gene
TPM
0.0
0.16
0.36
0.58
0.84
1.1
ENST00000400677.4
ENST00000506970.2
Testis
AC116612.1 - Nested gene
TPM
0.0
0.021
0.043
0.065
0.087
0.11
ENST00000435291.1
Testis

Supplementary Figure 9. Host and nested genes coexpression is correlated with regulation of isoform diversity. (A) Profile of expression of the different isoforms (rows) of the *MGAM/OR9A4* pair in different tissues (columns) from the GTEx portal data on isoforms expression (https://gtexportal.org/home/). Isoform and tissue were ordered by hierarchal clustering using the Euclidean distance and average linkages. Tissues where coexpression is happening are in red. (B) Profile of expression of the different isoforms (rows) of the *HMX1/AC116612.1* pair in different tissues (columns) from the GTEx portal data on isoforms expression (https://gtexportal.org/home/). Isoform and tissue were ordered by hierarchal clustering using the Euclidean distance and average linkages. Tissues where coexpression is happening are in red.


A
HNF4A - Host gene
0.0
0.2
0.4
TPM 0.0 1.1 3.2 7.7 17 36
ENST00000619550.4
ENST00000372920.1
ENST00000415691.2
ENST00000443598.6
ENST00000316099.8
ENST00000316673.8
ENST00000609795.5
ENST00000609262.5
ENST00000457232.5
0 1 2
Liver
Kidney - Cortex
Small Intestine - Terminal Ileum
Colon - Transverse
Kidney - Medulla
Stomach
Pancreas
Testis
Colon - Sigmoid
Brain - Anterior cingulate cortex (BA24)
Brain - Frontal Cortex (BA9)
Brain - Cortex
Fallopian Tube
Adrenal Gland
Lung
Ovary
Brain - Hippocampus
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Whole Blood
Vagina
Uterus
Thyroid
Spleen
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Pituitary
Nerve - Tibial
Muscle - Skeletal
Minor Salivary Gland
Heart - Left Ventricle
Heart - Atrial Appendage
Esophagus - Muscularis
Esophagus - Mucosa
Esophagus - Gastroesophageal Junction
Cervix - Endocervix
Cervix - Ectocervix
Cells - EBV-transformed lymphocytes
Cells - Cultured fibroblasts
Breast - Mammary Tissue
Brain - Substantia nigra
Brain - Spinal cord (cervical c-1)
Brain - Putamen (basal ganglia)
Brain - Nucleus accumbens (basal ganglia)
Brain - Hypothalamus
Brain - Caudate (basal ganglia)
Brain - Amygdala
Bladder
Artery - Tibial
Artery - Coronary
Artery - Aorta
Adipose - Visceral (Omentum)
Adipose - Subcutaneous
AL117382.2 - Nested gene
0.0
0.2
0.4
TPM 0.0 0.055 0.11 0.17 0.24 0.31
ENST00000455642.1
Liver
Whole Blood
Vagina
Uterus
Thyroid
Testis
Stomach
Spleen
Small Intestine - Terminal Ileum
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Pituitary
Pancreas
Ovary
Nerve - Tibial
Muscle - Skeletal
Minor Salivary Gland
Lung
Kidney - Medulla
Kidney - Cortex
Heart - Left Ventricle
Heart - Atrial Appendage
Fallopian Tube
Esophagus - Muscularis
Esophagus - Mucosa
Esophagus - Gastroesophageal Junction
Colon - Transverse
Colon - Sigmoid
Cervix - Endocervix
Cervix - Ectocervix
Cells - EBV-transformed lymphocytes
Cells - Cultured fibroblasts
Breast - Mammary Tissue
Brain - Substantia nigra
Brain - Spinal cord (cervical c-1)
Brain - Putamen (basal ganglia)
Brain - Nucleus accumbens (basal ganglia)
Brain - Hypothalamus
Brain - Hippocampus
Brain - Frontal Cortex (BA9)
Brain - Cortex
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Brain - Caudate (basal ganglia)
Brain - Anterior cingulate cortex (BA24)
Brain - Amygdala
Bladder
Artery - Tibial
Artery - Coronary
Artery - Aorta
Adrenal Gland
Adipose - Visceral (Omentum)
Adipose - Subcutaneous
B
R=0.66 p<2.2×10-16
Gene size (bp)
1×10^7
1×10^5
1×10^3
1 10 100 1000
Number of isoforms
R=0.62 p<2.2×10-16
Gene size (bp)
1×10^6
1×10^5
1×10^4
1×10^3
1×10^2
1 3 10 30
Number of isoforms

Supplementary Figure 10. **Host genes display a higher number of isoforms.** (A) Profile of expression of the different isoforms (rows) of the *HNF4A/AL117382.2* pair in different tissues (columns) from the GTEx portal data on isoforms expression (https://gtexportal.org/home/). Isoform and tissue were ordered by hierarchal clustering using the Euclidean distance and average linkages. Tissues where coexpression is happening are in red. (B) Scatter plot showing the correspondence between the gene size and number of isoforms. A base-10 log scale was used for the number of isoforms and the gene size. The correlation was determined by Spearman's rank correlation.