Global identification of mammalian host and nested gene pairs reveal tissue-specific transcriptional interplay

Bertille Montibus[1,†,*], James Cain[1,†], Rocio T Martinez-Nunez[2], Rebecca J. Oakey[1,*]

[1]Department of Medical and Molecular Genetics, King's College London, London, SE1 9RT, UK

[2]Department of Infectious Diseases, King's College London, London, SE1 9RT, UK

[†] Joint Authors

*To whom correspondence should be addressed. Tel +44 (0)20 848 8147, email rebecca.oakey@kcl.ac.uk

Correspondence may also be addressed to: Tel +44 (0)20 784 84317, email bertille.montibus@kcl.ac.uk

KEYWORDS

Host gene, nested gene, isoform diversity, gene annotation, alternative polyadenylation

ABSTRACT

A large proportion of genes in mammalian genomes are embedded within another gene, sharing nucleotide sequence. Characterisation of these host/nested gene pairs is required to understand their transcriptional crosstalk. We identify all host/nested gene pairs and reveal they are found at over a sixth of genes in both human and mouse. Host genes were more highly expressed and more likely to be protein-coding and their nested gene partners mainly reside within the largest intron of their host. Individual analysis of host or nested gene expression did not reveal tissue specific profiles. Co-expression profiles did reveal tissue specific profiles, suggesting host/nested gene crosstalk plays a role during differentiation and development. To assess true co-expression or mutual expression testis scRNA-seq data were used, revealing that pairs can switch expression dynamically between these states during spermatogenesis. Host genes have a larger pool of isoforms and when co-expression occurred at some testis-specific pairs, host transcript diversity increased. Scanning experimentally validated polyadenylation sites upstream of the nested gene, shows that host polyadenylation sites are enriched, implying that alternative polyadenylation is one of the mechanisms

influenced by host/nested gene co-expression. The collection of host/nested genes and our analysis is available on an Rshiny application.

INTRODUCTION

Classically genes are represented one after each other along the chromosome's genomic sequence separated by intergenic stretches of DNA which are devoid of genes, in a strand-specific manner. However, the genomic organisation is more complex and genes can in reality occupy the same genomic space as each other, considered as overlapping genes. The first overlapping pairs of genes were identified in the genomes of viruses (1). Viruses are rapidly evolving organisms, with a small genome that must be packaged into a small capsid, and still, maintain genetic novelty (2, 3). Thus, overlapping their genes allows them to deal with these constrains while enhancing their evolutionary stability because a mutation in an overlapping pair would affect more than one gene (4). Higher order organisms, on the other hand, contain vast amount of intergenic space for genes to occupy, do not have the requirement of rapid evolution and yet, still contain genomic overlaps (5). In eukaryotes, the first of these were identified in 1986 in Drosophila and mouse (6–8).

In 2019, a study in humans estimated that about 26% of the protein coding genes are overlapping with at least one other protein coding gene (9). Most of the time, the overlap involves a non-coding sequence of the genes (5'UnTranslated Region or UTR, 3'UTR or the intronic sequences) and sometimes only a few nucleotides at the 3' or 5' end of the genes. However, larger genomic overlaps exist and can result in a gene being fully contained within a completely different gene. This organisation of genes together are known as host/nested genes and they are likely to cooperatively regulate one other because they occupy the same genomic space.

The steric hindrance observed when two RNA polymerases transcribe two different genes in the same genomic region (10) and transcriptional collision observed at convergent genes (11) suggests that host and nested genes could prevent each other's expression and that they could be mutually exclusively expressed. In addition, a repressive chromatin environment is established in gene bodies by transcription itself and prevents instances of intragenic spurious initiation of transcription (12, 13). This is likely to repress nested gene expression via the same mechanism, if the host gene is expressed. On the other hand, nested gene expression has been shown to impact host gene pre-mRNA processing mechanisms, and, as a consequence, can

generate host RNA isoforms with different stability and coding capacity (14). This was shown for example at two independent imprinted loci, where the expression of a nested gene is associated with "premature" alternative polyadenylation of the host gene upstream of the nested gene whereas its silencing is associated with polyadenylation at the canonical 3'UTR (15, 16). This extends beyond the imprinted context and has been suggested to occur at other host genes harbouring nested *LINE1* retrotransposons, nested genes or putative intragenic promoters (17, 18). Growing evidence indicates that the short isoforms generated through "premature" polyadenylation are key for cell functions (19).

Thus, even if the role and the consequences of the host/nested gene organisation are starting to be uncovered, an updated genome-wide analysis is missing. As such, characterisation and identification of all possible host/nested genes is crucial for the understanding of gene regulation both in terms of level of expression and isoform regulation.

Previous compilations of host/nested gene lists are now outdated. Two previous analyses using NCBI Refseq and microarray data, at the time, identified 373 and 128 host/nested gene pairs in human (20, 21). Not only were the number of pairs identified different between the two analyses, but conclusions regarding co-regulation of the pairs were different. Yu *et al,* demonstrated that most pairs were anti-correlated, with some showing a positive correlation, suggesting mainly a mutually exclusive expression pattern, whilst some were expressed at the same time (21). Assis *et al* did not identify any significant correlation between the pairs, suggesting that host and nested genes are not influencing each other's expression and that this organisation is "neutral" (20).

These are the most recent collations of host nested genes in human from 2005 and 2008 respectively (20, 21). Genomic annotation is now much more comprehensive and advances in measurement technologies detect transcript isoforms in a more systematic way, as such, these previous lists are likely underestimating the numbers (22). Here, we identify and characterize the host/nested gene pairs in mouse and human using contemporary genomic annotations, RNA-sequencing and scRNA-seq data, and collate this analysis into an accessible shiny app for use by the scientific community (https://hngeneviewer.sites.er.kcl.ac.uk/hn_viewer/).

MATERIAL AND METHODS

Identification of Host Nested Gene pairs in human (hg19) and mouse (mm10) genomes

To generate a list of all host/nested genes, comprehensive GENCODE annotations of the human (hg19, Release 36 (GRCh38)) and mouse (mm10, Release M25 (GRCm38.p6)) genomes were downloaded as a bed file from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables). Bedtools 2.29.2 (23) function *intersect* was used to overlap each respective genomic annotation with themselves with the option –f 1.0, meaning that the minimum overlap required was 100% of one transcript overlapping with another transcript. To keep only the overlap involving different genes, using an in-house R script we filtered out genes which contained: the same name, the same size and where the start or end were the same. When multiple transcripts from the same gene were involved in the same pair, the longest transcript was kept. Finally, transcripts annotated as to be experimentally confirmed, VDJ recombination genes, complex loci (Protocadherin's and UDP-glucuronosyltransferase) were removed from the list. A full list of genes is available in Supplementary Tables 1 and 2.

Characterisation of the host/nested genes pairs

Information about transcripts biotypes, strand, number of isoforms and gene size was retrieved from GTF files downloaded from GENCODE website (https://www.gencodegenes.org/, Release M25 (GRCm38.p6) for mouse and Release 36 (GRCh37) for human). The enrichment score for each biotype correspond to the fold change between the proportion of a specific biotype for all genes and for the specified list of genes (For example for nested genes: $\frac{\text{Prop(BiotypeX)Nested genes} - \text{Prop(biotypeX)All genes}}{\text{Prop(biotypeX)All genes}}$). The functional enrichment analysis (Gene Ontology) was performed using g:Profiler (version e107_eg54_p17_bf42210) with Bonferroni correction method applying significance threshold of 0.05 (24). Data presented on host/nested gene pairs is viewable via https://hngeneviewer.sites.er.kcl.ac.uk/hn_viewer/ which was developed using RShiny.

Annotation of the nested gene location

ChIPseeker (version 1.24.0) was used to annotate the location of the nested genes inside their host. A custom GTF file containing only the information about the host transcripts involved in host/nested genes pairs was used as a reference and a bed file containing the start and end coordinates of the nested genes was used as a query. The genomicAnnotationPriority was define as "Exon","Intron","Promoter","5UTR","3UTR","Downstream" and "Intergenic". The pairs with the nested gene annotated inside their host were kept and further round of annotation

were necessary with removal of the already annotated pairs to annotate the missing pairs (for example when a gene is nested inside two different genes). Finally, for the last pairs with missing annotation, manual annotation was performed. The table containing the annotation and exon/intron numbers was retrieved and crossed with the GENCODE GTF file for intron/exon characteristics.

Conservation analysis

The list of all the human-mouse orthologs was downloaded using the Ensembl BioMart software suite (25) with the following filter "Homolog filters" option "Orthologous Mouse Genes" from the human gene list. The table was next downloaded with the following attributes: Gene stable ID, Transcript stable ID, Mouse gene stable ID, Mouse gene name, Mouse protein or transcript stable ID. The orthologs information was next overlap with the lists of host/nested genes using an in-house R script to identify the 474 combinations of orthologous pairs (Supp Tables 1 and 2).

ENCODE RNA-Sequencing analysis

Paired end RNA sequencing data was taken from multiple tissues in both human and mouse (for accession numbers, see Supplementary Tables 3,4). Reads were aligned using kallisto (26) and a read count table per sample was generated using tximport (27); reads were then normalised using DESeq2 (28). Pairs of host/nested genes were processed with a pair_id which was used to sort normalised read matrix data based on the corresponding host and nested gene ENST_ID using a custom R script. Correlations of host/nested gene expression across all samples was performed using the Pearson correlation coefficient. Highly correlated (coefficient >0.7) host/nested gene pairs were hierarchically clustered using the complete-linkage clustering method through the pheatmap package. Tissue-specificity of host, nested and all genes were evaluated using standard deviation of expression across all tissues, and, calculation of Tau index (29) using the package tspex (30) on the normalised count matrix.

scRNA-sequencing analysis

Integrated, normalized counts per cell table from scRNA-seq dataset during spermatogenesis was retrieved from Gene Expression Omnibus (GEO) database (accession number GSE153947) (31). The dataset was filtered to only include host and nested genes which were highly

correlated from the ENCODE-RNA sequencing analysis in testis. The expression of a gene was reduced to a binary signal whereby if expression reached above the threshold (a quarter of the maximum expression), a value of 1 was given, and a value of 0 if below this threshold. Host/Nested gene pairs in single cells were co-expressed if the sum of this value was 2. If the sum equal to 1, then single cells were only expressing either the Host or Nested gene. If the sum equal to 0, the single cells were not expressing either gene. A schematic illustrating this method is in Supplementary Figure 3B. For further analysis, pairs where over 5% of cells were expressing both transcripts were considered. The proportion of cells co-expressing each Host/Nested gene pairs was clustering via k-means clustering using the factoextra package (Version 1.0.7.999).

Analysis of polyadenylation sites at Host/Nested Gene pairs

Experimentally validated human (hg38) and mouse (mm10) polyadenylation site coordinates were downloaded from PolyASite 2.0 (32), with the hg38 coordinates being converted to hg19 using UCSC LiftOver. Bedtools 2.29.2 (23) *intersect* was used to overlap polyadenylation sites with regions and upstream of any nested gene coordinate by 10% of the size of the host gene. A control dataset was generated of random regions across host genes with a matched length (end-start site) distribution as the 10% upstream regions using bedtools *shuffle*.

RESULTS

About a sixth of human (23%) and mouse (14%) genes contains at least one nested gene

To obtain the most accurate and exhaustive list of host and nested genes in mouse and human we decided to use the GENCODE consortium genomic annotation. We used the "comprehensive" set which includes the highest number of transcripts associated with protein-coding genes, pseudogenes, long non-coding RNAs and small non-coding RNAs genes (22). Transcripts with "To be Experimentally Confirmed" biotypes and complex loci where alternative promoter or recombined segments are annotated with different genes' names but cannot be considered as independent genes and thus were removed (33–35) (see methods and Figure 1A) to ensure that a gene was defined as a validated transcriptional unit. Coordinate information for all the remaining genes were overlapped and the result filtered according to

the extent of the overlap. Only pairs where one transcript was fully contained inside another and involving transcripts from genes with different names were retained. When different transcripts of the same genes were involved in a pair, the longest was selected. After filtering 8659 and 17992 host/nested genes pairs in mouse and human remained respectively (Supplementary Tables 1 and 2). This represented 5511 (~14% of the genes) and 10061 (~23% of the genes) genes with at least one nested gene in mouse and human respectively (Figure 1A, Supplementary Tables 1 and 2)

Hosts harbour nested genes preferentially inside a long intron and show no strong orientation bias

To identify any preferences of host/nested gene structure and location, we performed a suite of analyses on their basic characteristics. Unsurprisingly, host genes were longer than their nested gene partner and when compared to the average size of the corresponding transcriptome (Supplementary Figure 1A). However, there was no correlation between the size of the host and their nested gene (Figure 1E). To minimize impact to the transcriptome, we hypothesised that host genes would not contain a large number of nested genes and that they would preferentially locate within non-coding intronic host gene regions. Most of the host genes, indeed, harboured less than 4 nested genes (95th percentile, 3 in mouse and 4 in human) but some contained up to 80 nested genes in human and 187 in mouse (Figure 1B). In addition, ~70% of the nested genes are fully contained inside one intron of their host gene. Surprisingly, the remaining overlap with exonic sequences and 3-5% of nested genes are completely embedded within exons (Supplementary Figure 1A). Within introns nested genes preferentially locate to the first intron (Supplementary Figure 1B) or the largest host intron (Figure 1D), as observed before (21). Assessment of pair orientation did reveal a slight bias towards being in the opposite orientation between pairs, but the distribution between opposing and same orientation were broadly similar (Figure 1C).

Given host/nested gene pairs surprising coverage (over a sixth) across the genome, we asked if these pairs were conserved between mouse and human. Only pairs where both genes had an identified orthologue in the other species were considered. Using these filtered pairs, it was found that 21% of the human pairs and 35% of the mouse pairs were conserved in the other species (Figure 1F) (for example *Mcph1/Angpt2* and *MCPH1/ANGPT2* (Supplementary Figure 1C)).

Host and Nested genes are enriched for different functional classes

To better characterise the genes involved in host/nested gene pairs, we took advantage of the broad functional classes of biotypes available in GENCODE (22). By comparison to the distribution of biotypes among all the genes, we found that host genes were enriched for protein coding genes both in human and mouse. The nested genes were enriched for small RNAs in general whereas host genes were depleted for these transcripts in both species (Figure 2A). When looking at the number of instances for each biotype, we observed that most of the host genes, but also a high proportion of the nested genes were protein coding genes. As expected from the enrichments, we also detected a high number of non-coding RNA in the nested genes list (lncRNAs, antisense RNAs and miRNAs) (Supplementary Figure 2A). Finally, when considering the correspondence of biotype between the host and nested genes per pair, no trend was observed for host and nested genes biotype association, other than the expected higher proportion of pairs of genes being in the opposite direction when one of the partners was annotated as an antisense transcript (Figure 2B).

Host or Nested gene expression is not linked to tissue-specificity.

Host/Nested genes show enrichment for distinct functional classes and so we next sought to annotate their biological function via gene ontology and expression analysis. Ontology analysis shows that host genes are enriched for terms associated with neural differentiation and synapse development (Supplementary Figure 2B) and nested genes for terms associated with negative regulation of translation and gene silencing as expected given that a large degree of small RNAs are nested genes (Supplementary Figure 2C).

The gene ontology analysis of host genes suggested that they could be enriched for genes important for tissue specification (Supplementary Figure 2B). Therefore, we decided to investigate if the host genes were expressed in a tissue-specific way using bulk tissue ENCODE RNA-sequencing data (36) (Supplementary Table 3,4). As a first approach, we calculated the standard deviation of host and nested gene expression, showing that host genes had a greater variance in expression level (Supplementary Figure 2D), suggesting these genes are expressed in a limited set of tissues. As the standard deviation is influenced by the level of expression we investigated the global level of expression across each tissue. Host genes were globally expressed at a higher level in all tissues tested whereas nested genes had lower expression

levels (Figure 2C), suggesting that the differences between standard deviations could be due to differences in basal expression levels. To explore this, we next computed the τ (Tau) index which is a tissue-specificity metric (29). The values of τ varies between 0 for housekeeping genes (or broadly expressed genes) and 1 for tissue-specific genes. In neither mouse nor human, was there a major difference in the value of τ indicating that host and nested genes are not enriched for tissue-specific genes when their expression is considered separately (Figure 2D).

Co-expression of host and nested gene is linked to tissue-specificity

As intragenic elements have been previously associated with tissue-specificity (17), we hypothesised that co-expression is likely to influence tissue-specific profiles and tested this by focusing on pair co-expression using the Pearson correlation coefficient method. This also tested the model whereby there is a steric hindrance of two RNA polymerases transcribing in the same genomic region (10) and host gene expression in turn represses the nested gene expression (13). In this case, host and nested genes should be expressed in a mutually exclusive way. A Pearson correlation coefficient was calculated for each pair between the expression profile of the host and the nested gene across tissues (as exemplified for the conserved pair *Mcph1/Angpt2* in Figure 3A). The coefficient distribution showed a peak between 0 and 0.5, indicating that for most pairs there is no relationship between the expression level of the host and the nested gene in both species (Figure 3B). In addition, anti-correlated expression was rare and there was a positive correlation (coefficient distribution > 0.7) (Figure 3B). To determine whether this correlation was linked with of the host and nested genes expression profiles, the relationship between the correlation and the tissue-specificity was evaluated. The higher the correlation, the more the host and nested genes were expressed in a tissue specific manner (Figure 3C). Finally, when looking at the expression of these highly correlated pairs, an important group of host/nested gene pairs emerged that exhibited co-expression in testis both in mouse and human (Figure 3D) and provided a model to study host/nested genes crosstalk.

Single-cell RNA sequencing reveals patterns of expression of host/nested genes pairs

Because, testis is a complex tissue, consisting of both somatic and germline cells at different stages of differentiation, bulk RNA-sequencing does not offer the resolution to determine if pairs of genes are expressed in the same cells or in different cells and are mutually exclusive. To assess this, a human testis single cell RNA-sequencing (scRNA-seq) dataset was leveraged (31). A Pearson correlation analysis was performed between host and nested gene expression per cell on the subset of testis-specific pairs identified by the ENCODE analysis (Figure 3D). Most of the 135 pairs (pairs with detected expression in the scRNA-seq dataset) demonstrated a low absolute value of the Pearson correlation coefficient (Supplementary Figure 3A). However, as Pearson correlation is skewed by the low and zero counts of scRNA-seq data, it was not suitable to assess co-expression, nor, mutually exclusive expression of genes (37–39). Indeed, host/nested gene pairs demonstrate different profiles of co-expression despite similar Pearson correlation coefficients. (*ABHD2/AC013565*.3, R=0.31 and *MYBPC1/AC117505.1*, R=0.32, Supplementary Figure 3B). To overcome this, we classified the single cells according to the host and nested gene expression based on a quarter of the maximum expression value for each gene (Supplementary Figure 3C). Expression above the threshold was indicative of robust detection and allows cell classification. Because we wanted to investigate the relationship between host and nested genes, both genes had to be detected in a reasonable number of cells by the scRNA-seq. We applied a filter of a minimum of 5% of the cells of the dataset had to have robust expression for both the host and the nested gene which identified 45 pairs. Finally, expression above the threshold for both the host and the nested gene allows identification of cells which co-express the pair. Expression above one threshold but below the other indicates expression of only one member of the pair whereas expression below both thresholds classified the cells as non-expressing the pair (Supplementary Figure 3C). This generated an accurate representation of co-expression and mutual exclusive expression, e.g., for *AL121974.1/PGK2* (36.98% of cells where both genes are expressed) and *AL135878.1/LINC02326* (0.45% of cells where both genes are expressed but 4.6% and 5% of cells with exclusive expression of the host and the nested gene, respectively) respectively (Figure 4A).

Transcriptional interplay between host and nested genes during spermatogenesis

Co-expression or exclusive expression were limited to some of the cell types, as such we hypothesised they are tightly regulated during development. For example,

ABHD2/AC013565.3 was co-expressed mainly during the late stages of spermatogenesis (diplotene, meiotic division, spermatids) and expression of *ABHD2* was only at earlier stages (spermatogonia, leptotene, zygotene, pachytene) and in some of the somatic cells (Supplementary Figure 3D). Globally, the interplay between host and nested genes expression was defined for each pair based on the percentage of cells defined as co-expressing or mutually expressing in every cell type (Supplementary Figure 3E). We performed K-mean clustering (Figure 4B and Supplementary Figure 4A) and identified distinct profiles of co-expression or exclusive expression, indicating that host/nested gene expression and interplay is regulated according to cell types and during the process of spermatogenesis. However, a profile with a simple relationship (*i.e.* mostly co-expressed or host expressed in some cell types and nested gene in others) was rare. For example, *GAGE12E/GAGE12F* were mainly co-expressed and *AC021242.2/AC022784.1* exhibited mutually exclusive expression (Figure 4C). Most of the pairs showed a complex profile sometimes with co-expression and sometimes exclusive expression, as exemplified by *PWRN1/PWRN2, IGSF11/IGSF11-AS1 and ABHD2/AC013565.2* (Figure 4C and Supplementary Figure 4B). For some pairs presenting this more complex expression profile, we observed that while the host gene was broadly expressed, the nested gene expression was limited to testis (Figure4D). Using the GTEx portal data on isoforms expression (40), we also observed that the nested gene expression was associated with a greater diversity of host isoforms expressed in testis (Figure 4D and Supplementary Figure 4B).

Host genes show a greater number of isoforms which seem to be linked with alternative polyadenylation

Given the correlation between co-expression with enhanced transcript diversity, at *IGSF11/IGSF11-AS1 and ABHD2/AC013565.2,* at regions with intragenic promoters (17) and at other host/nested pairs (15, 16), we assessed transcript diversity at host/nested genes. Evaluation of isoform enrichment identified that in comparison to the transcriptome, host genes are isoform rich, in the GENCODE annotation ($p<0.05$, two-sample Kolmogorov-Smirnov test) (Figure 5A). We observed that host genes were enriched for longer genes (Supplementary Figure 1A), and because of the slight positive correlation between gene size and number of isoforms (Supplementary Figure 5A), we asked if the enrichment for higher number of isoforms was associated with host genes being longer. However, when we classified

genes into size categories, we still observed that host genes were exhibiting a higher number of isoforms, suggesting that this was independent of their length (Supplementary Figure 5B). At similar regions, increased isoform diversity happens through alternative polyadenylation e.g., at *H13/Mcts2 or Herc3/Nap1L5,* suggesting that this is one method of transcriptional interplay between host and nested genes. We sought to assess whether alternative host transcripts were generated through alternative polyadenylation by leveraging a collection of experimentally validated polyadenylation sites from the PolyASite 2.0 (32). All polyadenylation sites were mapped to an upstream window of the nested gene which was proportional depending on the size of the host gene partner, with randomly placed windows of the same size across the host gene as a control (Supplementary Figure 5C). Polyadenylation sites were enriched in the window upstream of the nested gene compared to the expected number by chance in both human and mouse genomes (Figure 5B). The majority of these sites were attributed to intronic or terminal exonic termination regions (Figure 5C). Those attributed to intronic regions, indicate potential novel transcripts whilst those at exonic regions are already known transcripts of the respective host gene.

To determine if host gene poly(A) site enrichment was due to more sites per window or a higher probability that they contain a polyadenylation site, the number of sites per window was computed (Figure 5D). There was a higher probability of encountering at least a polyadenylation site upstream of a nested gene (Figure 5D, bar chart) and most of the windows contained one polyadenylation site (Figure 5D, line chart). Given that host polyadenylation sites are enriched close to nested genes, host genes are likely to be prone to being prematurely terminated when the host and nested gene pair is co-expressed.

DISCUSSION

Global compilation of host/nested genes pairs in mammals illustrates their transcriptional profile during tissue-specification and interplay with 3'end processing mechanisms. Compared to the most exhaustive previous known list (21), we identified ~50 times more pairs likely due to the continuous improvement of sequencing technologies allowing the annotation of more genes and transcripts. For example, between 2003 and 2013, the number of transcripts annotated in the RefSeq database increased by 23% (41). In addition, the comprehensive GENCODE annotation provides an exhaustive annotation for transcripts (22, 42). The resulting

collection is the most comprehensive list of host/nested gene pairs in mouse and human which can be explored through the web application developed for this study (https://hngeneviewer.sites.er.kcl.ac.uk/hn_viewer/). Ultimately, the accuracy of the pair identification is dependent the genome annotation itself. The development of new technologies, such as long-read sequencing which will iteratively advance the detection of all possible genomic transcripts (43).

Contrary to the previous studies, a lower conservation of the host/nested genes was observed between mouse and human (between 21 and 34% of the pairs with known orthologues). This suggests that these events are mainly species-specific and the difference to previous studies can be explained by the updated annotation and inclusion of more diverse transcripts type (multiple types *versus* protein coding only). Furthermore, the conservation analysis is highly dependent on the annotation of the orthologue gene list and it cannot be excluded that some conserved pairs were missed or excluded due to annotation defaults. A more detailed conservation study, including other species could improve the conservation level of host nested genes pairs. This would also be useful for investigating the origins of host/nested gene pairs and identify mechanisms involved such as *de novo* promoter generation inside a gene, transfer of a gene to inside another through genomic rearrangement or transposition, or extension of the host gene by acquisition of new exons which internalise an adjacent gene (44). These data indicate that most of the nested genes are fully contained within big introns of their hosts which suggests that there is a bias toward the acquisition of nested genes in these regions or a selection against events that interfere with the coding sequences of the host gene. This is also an indication that the impact of such a configuration would mainly be at the transcriptional level.

This was tested by conducting a correlation analysis between host and nested gene expression across multiple tissues. Previous analysis suggested pairs were anti-correlated and the other that there was no relationship regarding expression (20, 21). Our results agree with the latter, as they show there is not a straight and direct correlation, or, anti-correlation between host and nested genes expression for most of the pairs. We identify that, only when co-expression is considered, the pairs show patterns related to tissue-specificity and not when partners, either host or nested, are examined on their own. This suggests that the organisation, pattern and potential function of the host and nested gene pairs are during developmental processes during tissue specification. Given that testis showed the largest proportion of bulk

co-expressed pairs, we used single-cell RNA sequencing data from this tissue with a novel analysis method to assess true co-expression in an individual cell-dependent manner. Thus, we investigated co-expression and mutual exclusive behaviour of host nested genes in the same cells and showed that host and nested genes can be co-expressed in the same cells. For most of the pairs the profile was complex and for some of these pairs such as *IGSF11/IGSF11-AS1* and *ABHD2/AC013565.2,* increased isoform diversity at the host is associated with expression of the nested gene only in testis. Interestingly, *ABDH2* as an important role in sperm activation in mammals (45) and *IGSF11* is a regulator of meiosis during spermatogenesis (46) suggesting that the co-expression of host and nested genes and the increased number of host isoforms could be key for their function in testis.

In line with this, host genes were found to globally exhibit more transcript isoforms than other genes and the global analysis of polyadenylation signals suggests that isoform diversity is expanded at least through alternative polyadenylation with sites upstream of the nested genes, in line with premature polyadenylation events at two imprinted host/nested loci and the previous reports nine other host/nested genes pairs (15, 16, 18). In addition, the other mechanisms leading to transcript diversity, such as the use of alternative promoters or alternative splicing could be regulated by the interplay between host and nested genes. The dissection of mechanisms involved would require the improvement of technical approaches, such as long read sequencing, capturing precisely the transcription start site, end site and splicing profile of individual RNA molecules at a sufficient depth.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

BM and JC concieved the idea, performed the analysis, and wrote the manuscript. JC developed the shiny app with support and guidance from RTM. RJO provided supervision, discussed the work with BM and JC and critically reviewed and contributed to the final manuscript.

ACKNOWLEDGEMENTS

We thank Hannah Mischo for advice and useful discussions. We thank Lukasz Zalewski & Stuart Morrison for their assistance with the R Shiny app. We thank Hallgerdur Kolbeinsdottir for critical reading of the manuscript. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 04/05/23.

FUNDING

JC is supported by the UK Medical Research Council [MR/ N013700/1], and King's College London and is a member of the MRC Doctoral Training Partnership in Biomedical Sciences. BM is supported by funds from King's College London (to RJO) and by the Faculty of Life Sciences & Medicine Innovation Fund (to RJO & H. Mischo).

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

1. Weisbeek,P.J., Borrias,W.E., Langeveld,S.A., Baas,P.D. and Arkel,G.A.V. (1977) Bacteriophage phiX174: gene A overlaps gene B. *Proc. Natl. Acad. Sci.*, **74**, 2504–2508.

2. Feiss,M., Fisher,R.A., Crayton,M.A. and Egner,C. (1977) Packaging of the bacteriophage λ chromosome: Effect of chromosome length. *Virology*, **77**, 281–293.

3. Wu,Z., Yang,H. and Colosi,P. (2010) Effect of Genome Size on AAV Vector Packaging. *Mol. Ther.*, **18**, 80–86.

4. Simon-Loriere,E., Holmes,E.C. and Pagán,I. (2013) The Effect of Gene Overlapping on the Rate of RNA Virus Evolution. *Mol. Biol. Evol.*, **30**, 1916–1928.

5. Veeramachaneni,V., Makalowski,W., Galdzicki,M., Sood,R. and Makalowska,I. (2004) Mammalian Overlapping Genes: The Comparative Perspective. *Genome Res.*, **14**, 280–286.

6. Henikoff,S., Keene,M.A., Fechtel,K. and Fristrom,J.W. (1986) Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell*, **44**, 33–42.

7. Spencer,C.A., Gietz,R.D. and Hodgetts,R.B. (1986) Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature*, **322**, 279–281.

8. Williams,T. and Fried,M. (1986) A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. *Nature*, **322**, 275–279.

9. Chen,C.-H., Pan,C.-Y. and Lin,W. (2019) Overlapping protein-coding genes in human genome and their coincidental expression in tissues. *Sci. Rep.*, **9**, 13377.

10. Billingsley,D.J., Bonass,W.A., Crampton,N., Kirkham,J. and Thomson,N.H. (2012) Single-molecule studies of DNA transcription using atomic force microscopy. *Phys. Biol.*, **9**, 021001.

11. Prescott,E.M. and Proudfoot,N.J. (2002) Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci.*, **99**, 8796–8801.

12. Latos,P.A., Pauler,F.M., Koerner,M.V., Şenergin,H.B., Hudson,Q.J., Stocsits,R.R., Allhoff,W., Stricker,S.H., Klement,R.M., Warczok,K.E., *et al.* (2012) Airn Transcriptional Overlap, But Not Its lncRNA Products, Induces Imprinted Igf2r Silencing. *Science*, **338**, 1469–1472.

13. Neri,F., Rapelli,S., Krepelova,A., Incarnato,D., Parlato,C., Basile,G., Maldotti,M., Anselmi,F. and Oliviero,S. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, **543**, 72–77.

14. Licatalosi,D.D. and Darnell,R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.

15. Cowley,M., Wood,A.J., Böhm,S., Schulz,R. and Oakey,R.J. (2012) Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. *Nucleic Acids Res.*, **40**, 8917–8926.

16. Wood,A.J., Schulz,R., Woodfine,K., Koltowska,K., Beechey,C.V., Peters,J., Bourc'his,D. and Oakey,R.J. (2008) Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev.*, **22**, 1141–1146.

17. Amante,S.M., Montibus,B., Cowley,M., Barkas,N., Setiadi,J., Saadeh,H., Giemza,J., Contreras-Castillo,S., Fleischanderl,K., Schulz,R., *et al.* (2020) Transcription of intragenic CpG islands influences spatiotemporal host gene pre-mRNA processing. *Nucleic Acids Res.*, **48**, 8349–8359.

18. Kaer,K., Branovets,J., Hallikma,A., Nigumann,P. and Speek,M. (2011) Intronic L1 Retrotransposons and Nested Genes Cause Transcriptional Interference by Inducing Intron Retention, Exonization and Cryptic Polyadenylation. *PLOS ONE*, **6**, e26099.

19. Singh,I., Lee,S.-H., Sperling,A.S., Samur,M.K., Tai,Y.-T., Fulciniti,M., Munshi,N.C., Mayr,C. and Leslie,C.S. (2018) Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.*, **9**, 1716.

20. Assis,R., Kondrashov,A.S., Koonin,E.V. and Kondrashov,F.A. (2008) Nested genes and increasing organizational complexity of metazoan genomes. *Trends Genet.*, **24**, 475–478.

21. Yu,P., Ma,D. and Xu,M. (2005) Nested genes in the human genome. *Genomics*, **86**, 414–422.

22. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J., *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.

23. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

24. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.

25. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R., *et al.* (2022) Ensembl 2022. *Nucleic Acids Res.*, **50**, D988–D995.

26. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

27. Soneson,C., Love,M.I. and Robinson,M.D. (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. 10.12688/f1000research.7563.1.

28. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

29. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E., *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.

30. Camargo,A.P., Vasconcelos,A.A., Fiamenghi,M.B., Pereira,G.A.G. and Carazzolle,M.F. (2020) tspex: a tissue-specificity calculator for gene expression data. 10.21203/rs.3.rs-51998/v1.

31. Di Persio,S., Tekath,T., Siebert-Kuss,L.M., Cremers,J.-F., Wistuba,J., Li,X., Meyer zu Hörste,G., Drexler,H.C.A., Wyrwoll,M.J., Tüttelmann,F., *et al.* (2021) Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis. *Cell Rep. Med.*, **2**, 100395.

32. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res.*, **48**, D174–D179.

33. Jia,Z. and Wu,Q. (2020) Clustered Protocadherins Emerge as Novel Susceptibility Loci for Mental Disorders. *Front. Neurosci.*, **14**.

34. Jung,D., Giallourakis,C., Mostoslavsky,R. and Alt,F.W. (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu. Rev. Immunol.*, **24**, 541–570.

35. Strassburg,C.P., Kalthoff,S. and Ehmer,U. (2008) Variability and Function of Family 1 Uridine-5′-Diphosphate Glucuronosyltransferases (UGT1A). *Crit. Rev. Clin. Lab. Sci.*, **45**, 485–530.

36. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K., *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

37. Li,W.V. and Li,Y. (2021) scLink: Inferring Sparse Gene Co-expression Networks from Single-cell Expression Data. *Genomics Proteomics Bioinformatics*, **19**, 475–492.

38. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P., *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.

39. Sanchez-Taltavull,D., Perkins,T.J., Dommann,N., Melin,N., Keogh,A., Candinas,D., Stroka,D. and Beldi,G. (2020) Bayesian correlation is a robust gene similarity measure for single-cell RNA-seq data. *NAR Genomics Bioinforma.*, **2**, lqaa002.

40. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

41. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M., *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.

42. Frankish,A., Uszczynska,B., Ritchie,G.R., Gonzalez,J.M., Pervouchine,D., Petryszak,R., Mudge,J.M., Fonseca,N., Brazma,A., Guigo,R., *et al.* (2015) Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics*, **16**, S2.

43. Leung,S.K., Jeffries,A.R., Castanho,I., Jordan,B.T., Moore,K., Davies,J.P., Dempster,E.L., Bray,N.J., O'Neill,P., Tseng,E., *et al.* (2021) Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep.*, **37**.

44. Wright,B.W., Molloy,M.P. and Jaschke,P.R. (2021) Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.*, 10.1038/s41576-021-00417-w.

45. Miller,M.R., Mannowetz,N., Iavarone,A.T., Safavi,R., Gracheva,E.O., Smith,J.F., Hill,R.Z., Bautista,D.M., Kirichok,Y. and Lishko,P.V. (2016) Unconventional endocannabinoid signaling governs sperm activation via the sex hormone progesterone. *Science*, **352**, 555–559.

46. Chen,B., Zhu,G., Yan,A., He,J., Liu,Y., Li,L., Yang,X., Dong,C. and Kee,K. (2021) IGSF11 is required for pericentric heterochromatin dissociation during meiotic diplotene. *PLOS Genet.*, **17**, e1009778.

47. Dowd,C. (2020) A New ECDF Two-Sample Test Statistic.

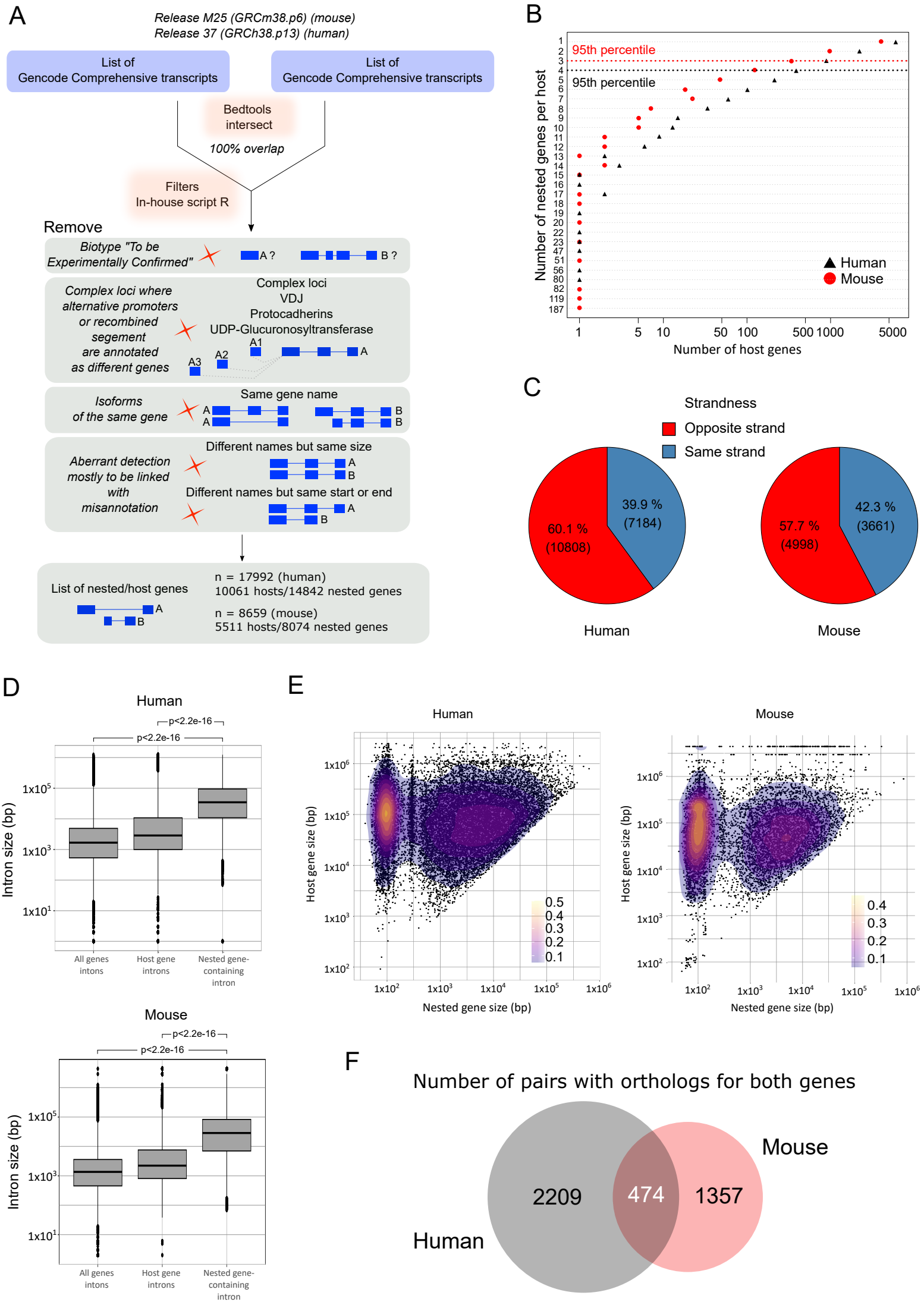
A
Release M25 (GRCm38.p6) (mouse)
Release 37 (GRCh38.p13) (human)
List of Gencode Comprehensive transcripts
List of Gencode Comprehensive transcripts
Bedtools intersect
100% overlap
Filters In-house script R
Remove
Biotype "To be Experimentally Confirmed"
A ?
B ?
Complex loci where alternative promoters or recombined segement are annotated as different genes
Complex loci
VDJ
Protocadherins
UDP-Glucuronosyltransferase
A1
A2
A3
A
Isoforms of the same gene
Same gene name
Aberrant detection mostly to be linked with misannotation
Different names but same size
Different names but same start or end
List of nested/host genes
n = 17992 (human)
10061 hosts/14842 nested genes
n = 8659 (mouse)
5511 hosts/8074 nested genes
B
95th percentile
95th percentile
Number of nested genes per host
Number of host genes
Human
Mouse
C
Strandness
Opposite strand
Same strand
39.9 % (7184)
60.1 % (10808)
42.3 % (3661)
57.7 % (4998)
Human
Mouse
D
Human
p<2.2e-16
p<2.2e-16
Intron size (bp)
All genes intons
Host gene introns
Nested gene-containing intron
Mouse
p<2.2e-16
p<2.2e-16
Intron size (bp)
All genes intons
Host gene introns
Nested gene-containing intron
E
Human
Host gene size (bp)
Nested gene size (bp)
Mouse
Host gene size (bp)
Nested gene size (bp)
F
Number of pairs with orthologs for both genes
Mouse
Human
2209
474
1357

Figure 1. Host/nested genes pairs in mouse and human shows similar characteristics. (A) Schematic of the pipeline used to generate the list of host/nested genes pairs. (B) Scatter plots showing the number of nested genes per host. The dotted line represents the 95th percentile of the distribution. (C) Proportion of host/nested genes pairs with both genes in the same orientation or in the opposition direction. (D) Distribution of the introns size for all introns in all the genes, all introns in host genes only or the intron of the host containing a nested gene. A base-10 log scale was used for the intron size. The significance of the difference was tested using a Welch Two Sample t-test. The resulting p-value is mentioned on the graph. (E) Scatter plot showing the correspondence between the sizes of host and nested genes. The colour gradient shows the density. A base-10 log scale was used for the gene size. (F) Venn diagram showing the overlap and conservation between the pairs having an ortholog for both host and nested in the other species.

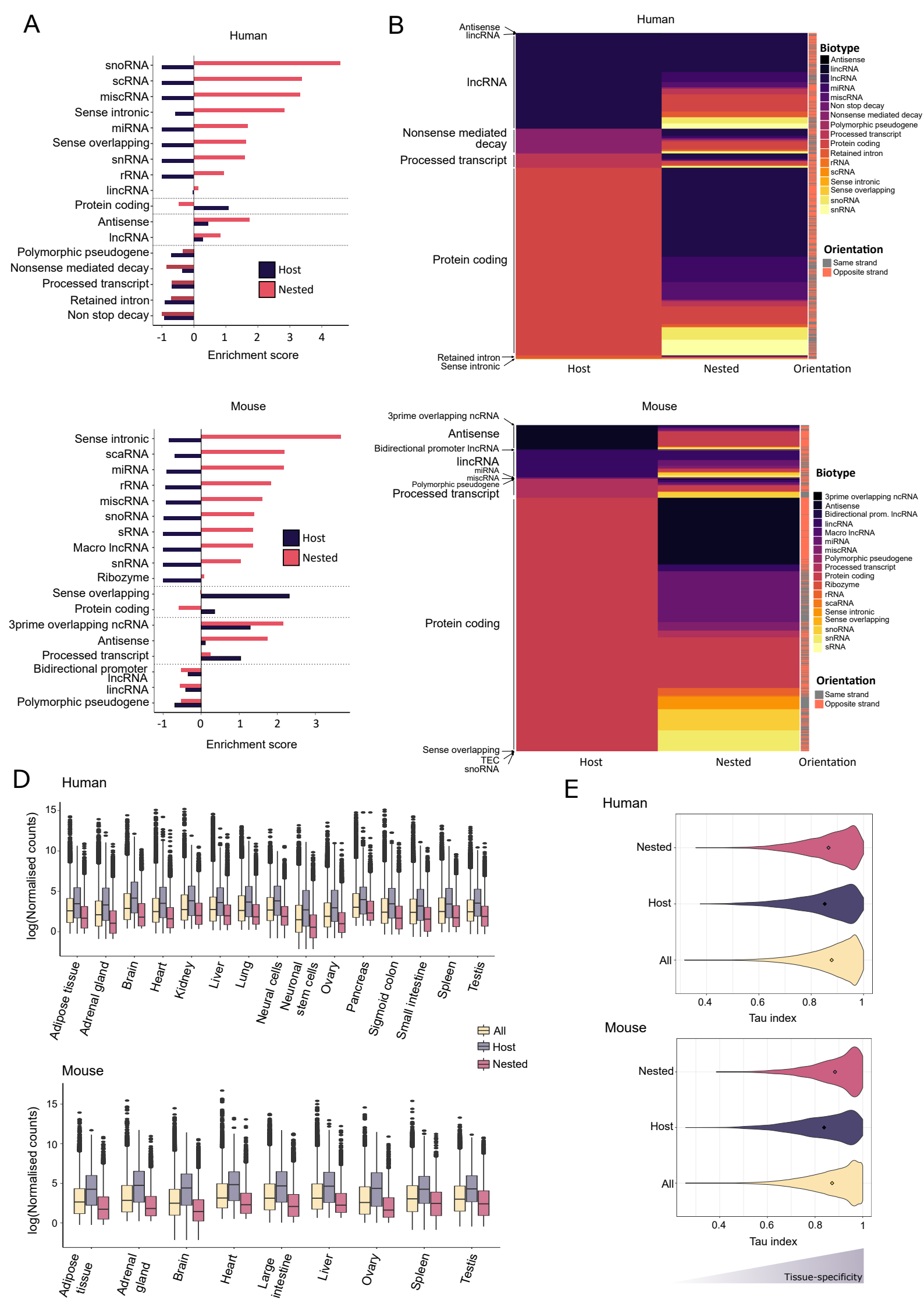
A
Human
snoRNA
scRNA
miscRNA
Sense intronic
miRNA
Sense overlapping
snRNA
rRNA
lincRNA
Protein coding
Antisense
lncRNA
Polymorphic pseudogene
Nonsense mediated decay
Processed transcript
Retained intron
Non stop decay
Host
Nested
-1 0 1 2 3 4
Enrichment score
B
Human
Antisense
lincRNA
lncRNA
Nonsense mediated decay
Processed transcript
Protein coding
Retained intron
Sense intronic
Host
Nested
Orientation
Biotype
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 strand
Opposite strand
Mouse
Sense intronic
scaRNA
miRNA
rRNA
miscRNA
snoRNA
sRNA
Macro lncRNA
snRNA
Ribozyme
Sense overlapping
Protein coding
3prime overlapping ncRNA
Antisense
Processed transcript
Bidirectional promoter lncRNA
lincRNA
Polymorphic pseudogene
Host
Nested
-1 0 1 2 3
Enrichment score
Mouse
3prime overlapping ncRNA
Antisense
Bidirectional promoter lncRNA
lincRNA
miRNA
miscRNA
Polymorphic pseudogene
Processed transcript
Protein coding
Sense overlapping
TEC
snoRNA
Host
Nested
Orientation
Biotype
3prime overlapping ncRNA
Antisense
Bidirectional prom. lncRNA
lincRNA
Macro lncRNA
miRNA
miscRNA
Polymorphic pseudogene
Processed transcript
Protein coding
Ribozyme
rRNA
scaRNA
Sense intronic
Sense overlapping
snoRNA
snRNA
sRNA
Orientation
Same strand
Opposite strand
D
Human
log(Normalised counts)
0 5 10 15
Adipose tissue
Adrenal gland
Brain
Heart
Kidney
Liver
Lung
Neural cells
Neuronal stem cells
Ovary
Pancreas
Sigmoid colon
Small intestine
Spleen
Testis
All
Host
Nested
Mouse
log(Normalised counts)
0 5 10 15
Adipose tissue
Adrenal gland
Brain
Heart
Large intestine
Liver
Ovary
Spleen
Testis
E
Human
Nested
Host
All
0.4 0.6 0.8 1
Tau index
Mouse
Nested
Host
All
0.4 0.6 0.8 1
Tau index
Tissue-specificity

Figure 2. Host/nested genes pair's biotypes and expression profile. (A) Bar chart representing the enrichment score for each of the biotype represented in the lists of host/nested genes in mouse and human. Positives values are indicatives of an enrichment whereas negative values correspond to a depletion in the list comparing to the proportions for all genes. (B) Heatmap showing the correspondence between host and nested genes biotypes per pair. For clarity reasons the pairs were grouped according to the biotype of the host. (D) Distribution of the normalised expression level detected in RNA-sequencing datasets from the ENCODE project (see material and methods) in different tissues, for the host genes, nested genes and all the genes. (E) Distribution of the Tau (τ) index calculated using the normalised expression level detected in RNA-sequencing datasets from the ENCODE project. The highest the value of τ index is the more the gene is showing tissue specific expression.

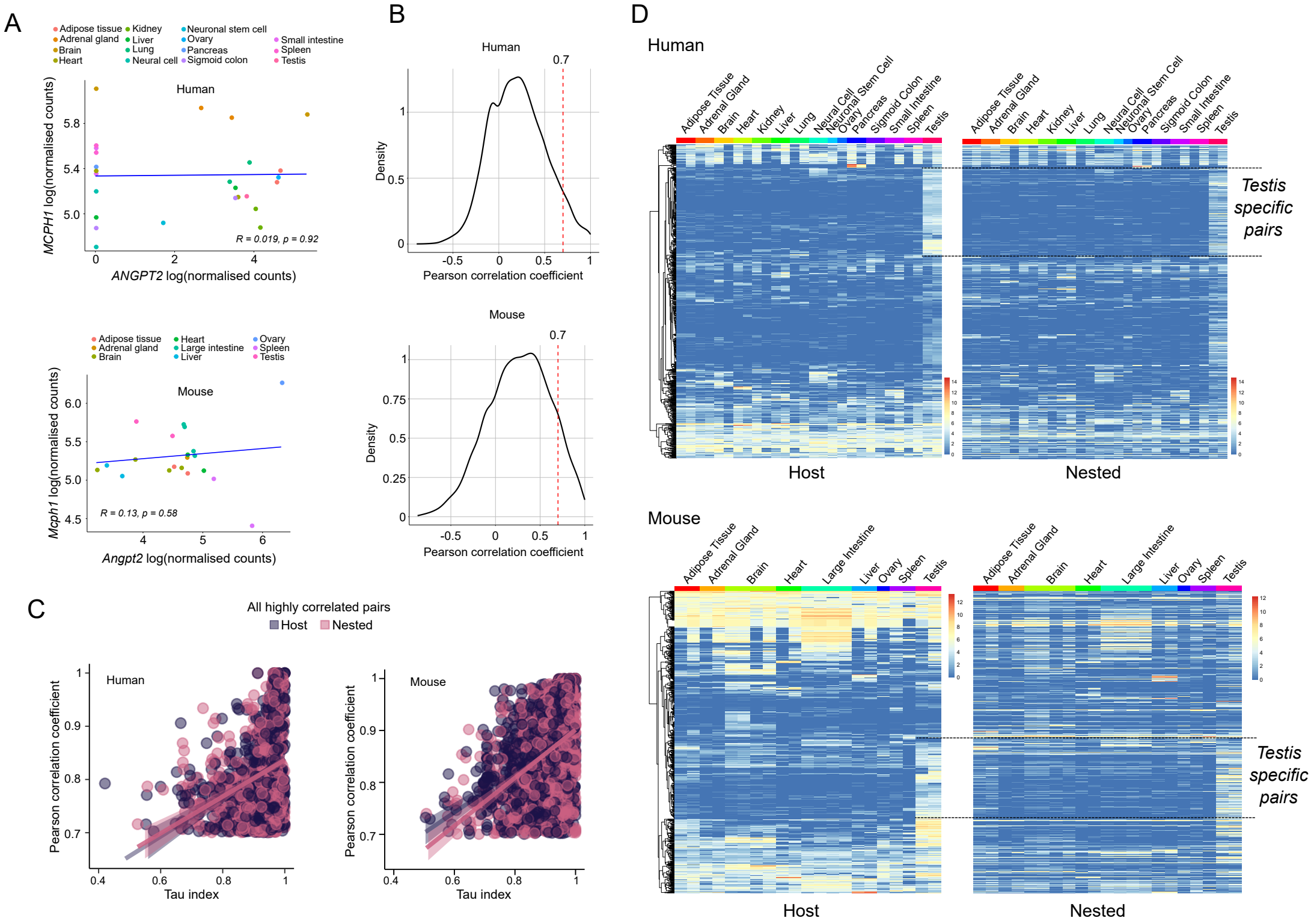
A
Adipose tissue
Adrenal gland
Brain
Heart
Kidney
Liver
Lung
Neural cell
Neuronal stem cell
Ovary
Pancreas
Sigmoid colon
Small intestine
Spleen
Testis
Human
MCPH1 log(normalised counts)
ANGPT2 log(normalised counts)
R = 0.019, p = 0.92
Adipose tissue
Adrenal gland
Brain
Heart
Large intestine
Liver
Ovary
Spleen
Testis
Mouse
Mcph1 log(normalised counts)
Angpt2 log(normalised counts)
R = 0.13, p = 0.58
B
Human
0.7
Density
Pearson correlation coefficient
Mouse
0.7
Density
Pearson correlation coefficient
C
All highly correlated pairs
Host
Nested
Human
Pearson correlation coefficient
Tau index
Mouse
Pearson correlation coefficient
Tau index
D
Human
Adipose Tissue
Adrenal Gland
Brain
Heart
Kidney
Liver
Lung
Neural Cell
Neuronal Stem Cell
Ovary
Pancreas
Sigmoid Colon
Small Intestine
Spleen
Testis
Testis specific pairs
Host
Nested
Mouse
Adipose Tissue
Adrenal Gland
Brain
Heart
Large Intestine
Liver
Ovary
Spleen
Testis
Testis specific pairs
Host
Nested

Figure 3. Most of the host/nested genes pairs do not exhibit expression correlation. (A) Examples of correlation plots used to calculate the Pearson's correlation coefficient for the conserved pair *Mcph1/Angpt2*. (B) Distribution of the Pearson correlation coefficient for all host/nested genes pairs. The vertical red dotted line indicates a correlation of 0.7 which was used as a threshold for high correlation (C) Correlation between the Pearson correlation coefficient value and the τ index for all the pairs exhibiting a Pearson correlation coefficient higher than 0.7. (D) Heatmap showing the normalised expression level detected in RNA-sequencing datasets from the ENCODE project for host and nested genes from pairs with a Pearson correlation coefficient higher than 0.7. Host genes were clustered using hierarchical clustering according to their expression profile and their nested counterpart were maintained in the same order. The dotted lines are flanking the groups of pairs specifically co-expressed in testis.

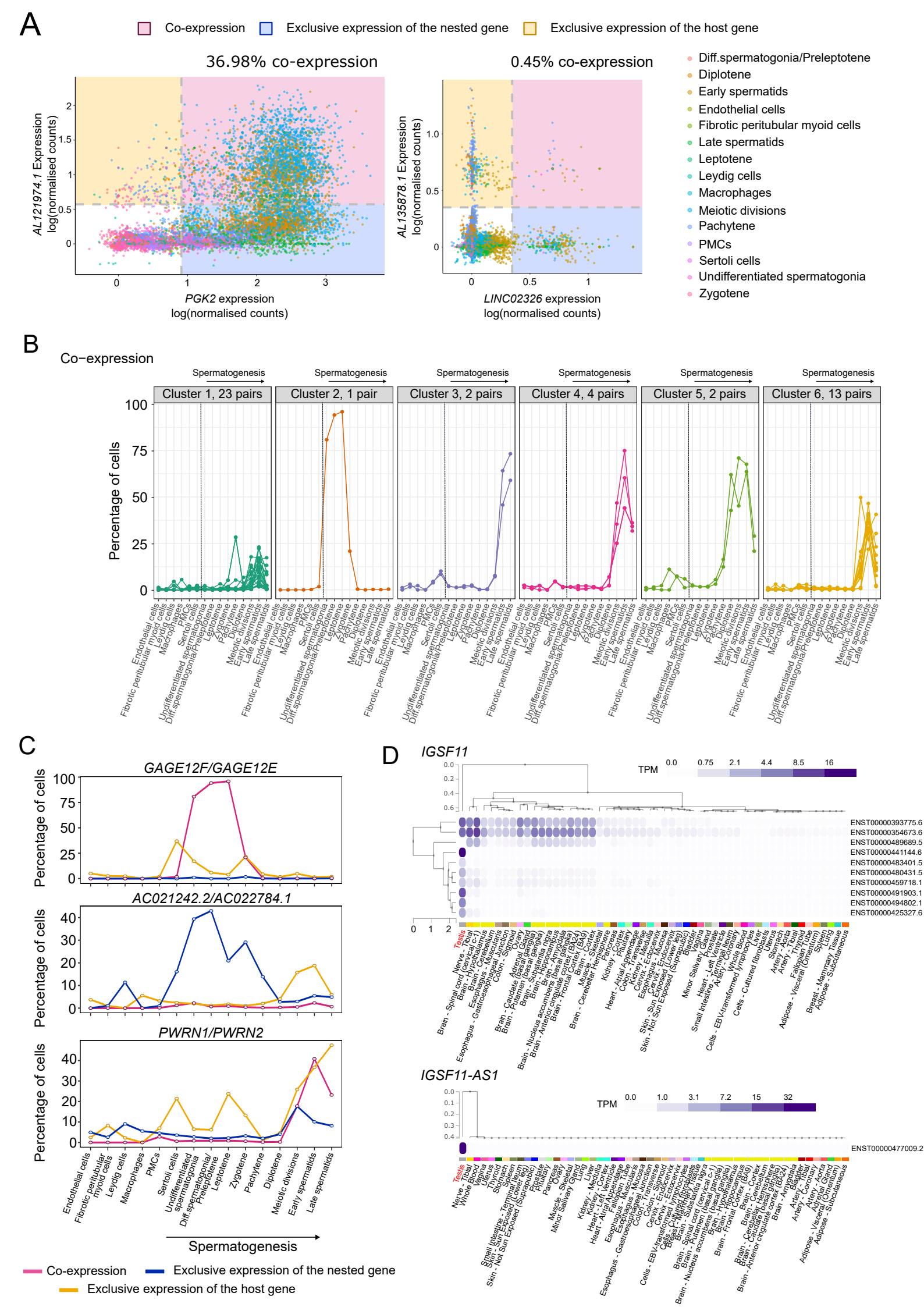
A
Co-expression
Exclusive expression of the nested gene
Exclusive expression of the host gene
36.98% co-expression
0.45% co-expression
AL121974.1 Expression log(normalised counts)
PGK2 expression log(normalised counts)
AL135878.1 Expression log(normalised counts)
LINC02326 expression log(normalised counts)
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
B
Co−expression
Spermatogenesis
Cluster 1, 23 pairs
Cluster 2, 1 pair
Cluster 3, 2 pairs
Cluster 4, 4 pairs
Cluster 5, 2 pairs
Cluster 6, 13 pairs
Percentage of cells
C
GAGE12F/GAGE12E
AC021242.2/AC022784.1
PWRN1/PWRN2
Percentage of cells
Spermatogenesis
Co-expression
Exclusive expression of the nested gene
Exclusive expression of the host gene
D
IGSF11
TPM
0.0
0.75
2.1
4.4
8.5
16
ENST00000393775.6
ENST00000354673.6
ENST00000489689.5
ENST00000441144.6
ENST00000483401.5
ENST00000480431.5
ENST00000459718.1
ENST00000491903.1
ENST00000494802.1
ENST00000425327.6
IGSF11-AS1
TPM
0.0
1.0
3.1
7.2
15
32
ENST00000477009.2

Figure 4. Host and nested genes with high co-expression in testis can exhibit dynamic co-expression patterns during spermatogenesis at the single cell level. (A) Scatter plots showing host and nested gene normalised expression in testis single cells for two examples pairs *AL121974.1/PGK2* and *AL135878.1/LINC02326*. (B) Proportion of cells which co-express host and nested gene pairs across spermatogenesis. Six distinct co-expression profiles were determined by K-means clustering. (C) Scatter plot showing the profiles of expression determined as in SuppFig3C for three different pairs of host and nested genes. *CAGE12F/CAGE12E* is an example of pair where host and nested genes are mainly co-expressed, *AC021242.2/AC022784.1* an example of pair with mutually exclusive expression and *PWRN1/PWRN2* a pair with a complex expression profile in testis. (D) Profile of expression of the different isoforms (rows) of the *IGSF11/IGSF11-AS1* 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.

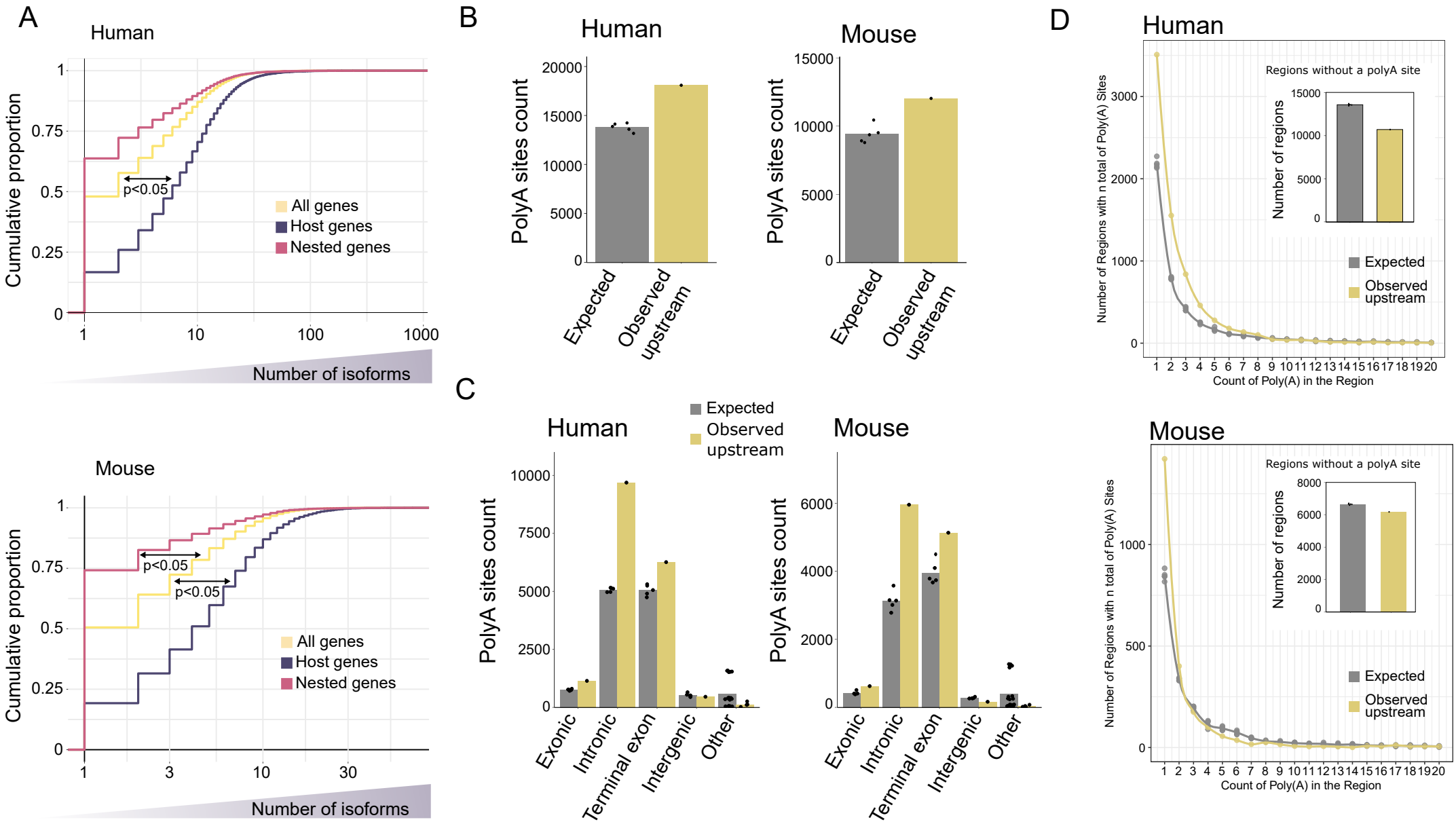
A
Human
Cumulative proportion
p<0.05
All genes
Host genes
Nested genes
Number of isoforms
Mouse
p<0.05
p<0.05
B
Human
Mouse
PolyA sites count
Expected
Observed upstream
C
Expected
Observed upstream
Exonic
Intronic
Terminal exon
Intergenic
Other
D
Regions without a polyA site
Number of regions
Number of Regions with n total of Poly(A) Sites
Count of Poly(A) in the Region

Figure 5. Nested genes are likely to impact host gene alternative polyadenylation. (A) Cumulative proportion of the number of isoforms per host genes, nested genes and for all genes. The significance of the difference was tested using a DTS statistic test which compare empirical cumulative distribution functions (ECDF) (47). The significant resulting p-values are mentioned on the graph. (B) Bar chart showing the total number of polyadenylation sites counted per window placed upstream of the nested gene (yellow) or randomly inside the host (grey) (See Supplementary Figure 4) (C) Genomic annotation of the polyadenylation sites identified in B. (D) Number of polyadenylation sites per window (line chart) and number of windows without any polyadenylation site (bar chart).