Common tissue-specific expressions and regulatory mechanisms of c-KIT isoforms with and without GNNK and GNSK sequences across five mammals

Running title: c-KIT isoforms with and without GNNK and GNSK

Rikuto Goto[1], Naoaki Sakamoto[1,2], Akinori Awazu[1,2]

[1]Graduate School of Integrated Sciences for Life, Hiroshima University, Higashihiroshima, Hiroshima, Japan

[2]Research Center for the Mathematics on Chromatin Live Dynamics, Hiroshima University, Higashihiroshima, Hiroshima, Japan

Corresponding author: Akinori Awazu (awa@hiroshima-u.ac.jp)

Abstract

c-KIT is a transmembrane receptor tyrosine kinase involved in various signaling pathways. Alternative pre-mRNA splicing of *KIT* results in isoforms that differ in the presence or absence of four amino acid sequences in the extracellular juxtamembrane region, such as, isoforms with and without the GNNK sequence (GNNK+ and GNNK-, respectively) in humans and mice, and those with and without GNSK (GNSK+ and GNSK-, respectively) in domestic dogs, cats, and sheep. These isoforms have been extensively studied as disease-associated splice variants with differing kinase activities. However, the expression patterns and regulatory mechanisms of each isoform in various animal species remain poorly understood. In the present study, a comprehensive expression analysis was conducted using tissue-wide transcriptome data from humans, mice, domestic dogs, domestic cats, and domestic sheep. We found that the expression ratio of c-KIT isoforms differs across tissues, and such features are conserved across animal species: GNNK+ and GNSK+ isoforms have high expression ratios in the central nervous system, while GNNK- and GNSK- predominate in other tissues. Furthermore, the RNA-binding proteins NOVA1/2 were involved in regulating such tissue-specific isoform expressions through alternative 5' splice donor site selections.

Keywords: c-KIT, RNA-binding proteins, isoforms, transcriptome data, central nervous system

Introduction

c-KIT is the transmembrane receptor tyrosine kinase encoded by the *KIT* gene, which is activated by its ligand, stem cell factor (SCF), to trigger multiple downstream signaling cascades, such as PI3K, Src, JAK/STAT, PLC-γ, and MAPK pathways, essential for hematopoiesis, stem cell maintenance, gametogenesis, melanogenesis, mast cell development, and cell migration.[1,2] c-KIT comprises five N-terminal immunoglobulin-like domains, a single transmembrane domain, and a C-terminal intracellular tyrosine kinase domain. In humans and mice, the extracellular juxtamembrane region, which is located adjacent to the fifth immunoglobulin□like domain (D5) that mediates dimer contacts,[3] contains a four-amino-acid sequence Gly□Asn□Asn□Lys (GNNK) (Figure 1A and B). Whereas in domestic dogs, cats, and sheep, it contains another four-amino acid sequence, Gly-Asn-Ser-Lys (GNSK).[4-8]

c-KIT synthesizes isoforms that differ in the presence or absence of the GNNK/GNSK sequence. Nevertheless, the current knowledge of these isoforms differs among species. In the Reference Sequence Database (RefSeq) (https://www.ncbi.nlm.nih.gov/refseq/),[9] the isoforms with and without GNNK (GNNK+ and GNNK-, respectively) in human and mouse have been curated and registered as Known RefSeq entries, with accession numbers beginning with "NP_" (Figure 1B). In contrast, in domestic dog and sheep, only GNSK- is included in the Known RefSeq, while the GNSK+ is represented as a predicted transcript in the Model RefSeq category and indicated by accession numbers beginning with "XP_". Nonetheless, GNSK+ and GNSK- expressions in various canine tissues,[5,6,10] have been reported. Moreover, the co-expression of the two isoforms in the skin tissue of Merino sheep has been observed.[8] Additionally, in domestic cats, only GNSK+ has been annotated in the Known RefSeq. However, GNSK- isoforms have also been detected in feline samples.[7]

Clinically, the mutations of *KIT*, a proto-oncogene, have been frequently reported in various cancers, including gastrointestinal stromal tumors,[11,12] mast cell tumors, [13,14] mastocytosis, [15,16] and acute myeloid leukemia (AML).[17] Consequently, studies on c-KIT isoforms have mainly focused on their involvement in various diseases. For instance, in humans, the GNNK-/+ balance correlates with prognosis in AML. [18] The expression ratio of GNSK+ to GNSK- differs between healthy and tumorous canine mammary tissues.[10] Functionally, SCF-induced phosphorylation occurs via isoform-specific kinetics. GNNK- and GNSK- variants exhibit rapid and robust phosphorylation.[7,19-23] These differences in activity between the isoforms may be explained by differences in the orientation and flexibility of each protomer within the dimer, based on differences in the length of the juxtamembrane domain.[3,23,24] These findings suggest that isoform-dependent differences in phosphorylation and downstream signaling play critical roles in tumor initiation and progression.

Existing studies have predominantly focused on disease-specific contexts, thereby limiting the range of tissues examined. Thus, studies on the comprehensive tissue-dependent characterization of c-KIT isoform expression and its regulatory mechanisms, which should provide basic knowledge for clinical research, are lacking, even though *KIT* is expressed in the entire animal body.

Materials and Methods

Processing of transcriptome data

For the analysis of human and mouse transcriptome data, we first downloaded the aligned RNA sequencing (RNA-seq) data files (BAM format files, aligned to GRCh38.p12 and GRCm38.p6, respectively) with their corresponding transcriptome-level gene expression quantification files (TSV format files) from the ENCODE Project (https://www.encodeproject.org) [25,26] (Table S1). Reference genome files and gene transfer format (GTF) files for humans and mice were obtained from GENCODE, matching the versions used by ENCODE: *Homo sapiens* GRCh38.p12 (GENCODE Release 29; https://www.gencodegenes.org/human/release_29.html) and *Mus musculus* GRCm38.p6 (GENCODE Release M21; https://www.gencodegenes.org/mouse/release_M21.html).

To obtain splice junction data (SJ.out.tab format file) for target genes, we extracted reads for *KIT* (human: chr4:54652918-54745715, mouse: chr5:75569987-75661722), *Ccdc9* (mouse: chr7:16274042-16286795), and *Ncdn* (chr4:126743750-126753438) from the downloaded BAM files using the samtools view command (options -h -b) of samtools (version 1.21), and converted them to FASTQ files using the samtools fastq command (no options for single-end sequence, and options -1, -2 to split FASTQ for pair-end sequence). Subsequently, we employed repair.sh from BBMap (version 39.01) with default parameters to make both paired-end FASTQ files (R1 and R2) contain the same number of reads, and that read pairs were synchronized in the same order. The FASTQ files were re-mapped to each reference genome using STAR (version 2.5.2b, options -outFilterMultimapNmax 1000, -outSAMattributes All, and -outSAMtype BAM SortedByCoordinate).[29]

Reference genomes for domestic dogs (*Canis lupus familiaris;* GCF_011100685.1), cats (*Felis catus;* GCF_018350175.1, AnAmus 1.0 [27]), and sheep (*Ovis aries;* GCF_016772045.2) were downloaded from the National Center for Biotechnology Information (NCBI) Datasets portal (https://www.ncbi.nlm.nih.gov/datasets/genome/) for transcriptome analyses. RNA-Seq data were obtained from the NCBI Sequence Read Archive (NCBI SRA). Quality control of each raw read of the RNA-seq data was performed using fastp (version 0.22.0) [28] with default parameters. Filtered reads were aligned to the reference genome using STAR (version 2.7.4a, options -outSAMtype BAM

SortedByCoordinate, -quantMode TranscriptomeSAM, -outSAMattributes All, -outFilterMultimapNmax 1000, -outSAMunmapped Within KeepPairs, -outSAMstrandField intronMotif). Transcript quantification was performed using RSEM (version 1.3.1, options: -estimate-rspd, -strandedness reverse, -no-bam-output).[30]

Tissue categories were assigned based on the ENCODE "human and mouse body maps". Samples that did not conform to any predefined category were excluded. Additionally, we introduced "adipose tissue" and "embryo" as separate categories.

Quantification of splicing variants

To distinguish between the long and the short exon of *KIT*, *Ccdc9*, and *Ncdn*, we focused on RNA-seq reads spanning splice junctions generated by the selective use of alternative donor sites (5′ splice sites). Based on the genomic coordinates of splice sites, reads were classified as either longer or shorter exons (for c-KIT, inclusion or exclusion of GNNK or GNSK), and the number of reads supporting each splice variant was counted (Figure S1). Samples were filtered to include only those with *KIT* expression levels of ≥1 transcript per million (TPM) and a combined total of at least three junction-spanning reads, supporting either the inclusion or exclusion of the relevant domains. Using this filtering method, we obtained 334 human samples across 14 tissue types, 184 mouse samples across 13 tissue types, 34 dog samples across seven tissue types, 76 cat samples across 12 tissue types, and 197 sheep samples across 11 tissue types for subsequent analysis. A similar approach was employed, yielding 217 and 209 mouse samples after *Ccdc9* and *Ncdn* filtering, respectively (Table S1).

For data from these samples, the inclusion-exclusion ratio (in-ex ratio) for the spliced domains was calculated using the following formula:

$$\text{in-ex ratio} = ([\text{SDS2\&SAS}] - [\text{SDS1\&SAS}])/([\text{SDS2\&SAS}] + [\text{SDS1\&SAS}])$$

Where [SDS1&SAS] indicates the number of reads aligned to the genome region containing only splice donor site 1 (SDS1) and the splice acceptor site (SAS), and [SDS2&SAS] indicates those aligned to the genome region with SDS1, SDS2, and SAS (Figure S1). For c-KIT, [SDS2&SAS] corresponds to the count of GNNK+/GNSK+ and [SDS1&SAS] corresponds to that of GNNK-/GNSK-.

Statistical analysis

Statistical analyses were performed using R software (version 4.4.1). Wasserstein distance was computed using the wasserstein1d function (parameter p = 1) from the "transport" package. Hierarchical

clustering was computed using pheatmap function (parameter clustering_method = “complete,” scale = “none,” cluster_rows = TRUE, cluster_cols = TRUE, treeheight_row = 0) from the “pheatmap” package. Wilcoxon rank-sum test was computed using the wilcox.exact function (parameter alternative=‘two.sided’, exact=TRUE, paired=FALSE) from the “exactRankTests” package.

Gene set enrichment analysis (GSEA)

The gene set files (GAF format) for humans and mice were downloaded from the Gene Ontology Consortium (https://current.geneontology.org/products/pages/downloads.html),[31,32] and those for dogs, cats, and sheep were downloaded from the NCBI portal and then converted to the GMT format. The Pearson correlation coefficient between the in-ex ratio and the expression level of all genes (TPM) was calculated using the cor function from the “StatsBase” package in Julia software. GSEA was performed by setting the seed value with set.seed(1111) and using the GSEA function (parameter TERM2GENE={gmt files for each species}, minGSSize = 20, maxGSSize = 500, pvalueCutoff = 0.01, seed = TRUE, eps = 0) from the “clusterProfiler” package in R. Top and bottom ranked genes by their correlation coefficient were used as input.

Analysis of high-throughput sequencing-crosslinking immunoprecipitation (HITS-CLIP) data

HITS-CLIP data for NOVA1/2 collected by Saito et al.[33] were downloaded from NCBI (Accession PRJNA286234; GEO: GSE69710). The NOVA1 sequence tag dataset (FASTQ format file, replicate1: SRR3532943, replicate2: SRR3532944, replicate3: SRR3532945) and NOVA2 sequence tag dataset (FASTQ format file, replicate1: SRR3532940, replicate2: SRR3532941, replicate3: SRR3532942) were quality-controlled using fastp with default parameters and aligned to GRCm38.p6 using STAR (version 2.5.2b, options -outFilterMultimapNmax 1000, -outSAMattributes All, -outSAMtype BAM SortedByCoordinate). Subsequently, the triplicate BAM files for NOVA1 and NOVA2 were merged into a single BAM file using the samtools merge command with the default parameters. Then, by referencing the CIGAR column of the BAM file, reads containing N were removed using the samtools view and awk commands. The resulting BAM file was converted to bed format using the bedtools bamtobed command with default parameters (version v2.31.1).

Result

Two selective splicing 5’ donor sites were closely positioned in *KIT*

For human, mouse, dog, cat, and sheep, we confirmed both GNNK-/+ and GNSK-/+ isoforms of c-KIT are generated via alternative 5′ donor site splicing that follows the canonical GU-AG rule [34,35] (Figure

1C). Within the pre-mRNA, two closely spaced GU donor sequences are present at the 5′ end of the intron immediately upstream of the exon encoding the transmembrane domain. Selection of the upstream donor site resulted in the exclusion of GNNK/GNSK, whereas selection of the downstream donor site, located 12 nucleotides downstream, led to GNNK/GNSK inclusion.

Animal species-independent tissue-specific expressions of c-KIT isoforms

To clarify the whole-body expressions of the isoforms in animals, we obtained RNA sequence read data from various tissues of five mammals and aligned transcriptome reads to their reference genomes to quantify *KIT* expression levels in TPM (Table S1). The data analyzed in this study did not include samples from patients with tumors or cancer. Next, referring to the splice junction coordinates, we respectively estimated expression levels of GNNK+ and GNNK- for human and mouse, and those of GNSK+ and GNSK- for domestic dog, cat, and sheep. Subsequently, the in-ex ratio of GNNK and GNSK sequences was calculated for each dataset, where $-1 \leq$ 'in-ex ratio' ≤ 1 was satisfied, and 'in-ex ratio' = 1 when all c-KIT proteins contained GNNK or GNSK, and 'in-ex ratio' = –1 vice versa (see Materials and Methods for details).

First, we identified the whole-body expressions of both GNNK+ and GNNK- in human and mouse, as well as those of both GNSK+ and GNSK- in dog, cat, and sheep (Figure 2). Second, we observed the following animal species-independent features of in-ex ratios across tissue types: GNNK- and GNSK- were predominantly expressed in most tissues, whereas the central nervous system (CNS) exhibited a higher proportion of GNNK+ and GNSK+ than other tissues (Figure 3 and Figure S2).

To quantify the differences in in-ex ratio distributions across tissues, we calculated Wasserstein distances and performed hierarchical clustering (Figure S3). In all species analyzed, CNS samples were consistently distinct from other tissue types in the first branch of the dendrogram, indicating that the in-ex ratio distribution in the CNS differed specifically from that in other tissues. Additionally, we performed a Wilcoxon rank-sum test to determine significant differences in the in-ex ratios between CNS and non-CNS tissue samples. The results were as follows: human ($p = 4.197 \times 10^{-41}$), mouse ($p = 1.257 \times 10^{-17}$), dog ($p = 1.943 \times 10^{-12}$), cat ($p = 2.953 \times 10^{-16}$), and sheep ($p = 6.323 \times 10^{-12}$).

Enrichment analysis of genes correlated with in-ex ratio

To investigate the roles of GNNK+/GNSK+ and GNNK-/GNSK- in biological processes, the functions of genes whose expression correlated with the in-ex ratio were examined. GSEA was performed on a ranked list of genes based on the correlation coefficient between the in-ex ratio and gene expression

(TPM). Gene ontology (GO) terms were used to categorize the roles of each gene. In all five species, the top-ranked gene groups whose expression correlated with GNNK+/GNSK+ expression were enriched in categories related to neural development, neural function, ion transport, and cell adhesion. In contrast, the lower-ranked gene groups whose expression correlated with GNNK-/GNSK- expression were enriched in categories related to DNA repair and replication, immune response, and cell matrix (Figure 4).

Here, the GO terms enriched for genes whose expression positively correlated with GNNK+ in humans and mice were designated as G_H and G_M, respectively, and the GO terms enriched for genes whose expression positively correlated with GNSK+ in dogs, cats, and sheep were designated as G_D, G_C, and G_S. We found 53 GO terms that were commonly contained in G_H and G_M but not in G_D, G_C, and G_S, that is, GO terms associated only with GNNK+ correlated genes; for example, neuron migration (GO:0001764), regulation of presynapse assembly (GO:1905606), and social behavior (GO:0035176). Additionally, four GO terms, monoatomic ion transport (GO:0006811), nervous system process (GO:0050877), proton transmembrane transport (GO:1902600), and ubiquitin-dependent protein catabolic process (GO:0006511) were associated only with the GNSK+ correlated genes (Table S2). This result suggests that the disparity between the third amino acid residues, asparagine (N) and serine (S), in c-KIT is animal species-dependent.

Splice site switching is regulated by splicing factors

Splicing for GNNK+/GNSK+ production occurred significantly more frequently in the CNS than in other tissues. Based on this observation, we focused on *NOVA1/2* genes, splicing regulators with CNS-specific expression. NOVA1/2 proteins are expressed at high levels in the CNS and regulate tissue-specific alternative splicing by binding to YCAY (Y indicates pyrimidine C or U) clusters in target pre-mRNAs.[36-41]

Similar to previous studies, *NOVA1/2* was highly expressed in the CNS compared with other tissue types in the sample group used in this study (Figure S4). To determine whether NOVA1/2 proteins modulate the alternative pre-mRNA splicing of *KIT*, we calculated the Pearson correlation coefficients and p-values to test for a lack of correlation between the in-ex ratio values and the expression levels of *NOVA1* or *NOVA2* in the five mammals. In human, *NOVA1* expression correlated positively with the in-ex ratio ($r = 0.5948$, $p = 2.469 \times 10^{-33}$), in mouse ($r = 0.2488$, $p = 0.03259$), dog ($r = 0.8223$, $p = 2.46 \times 10$□□), cat ($r = 0.6316$, $p = 9.480 \times 10$□¹□), and sheep ($r = 0.3756$, $p = 5.38 \times 10$□□). *NOVA2* demonstrated comparable positive correlations: human ($r = 0.6247$, $p = 1.516 \times 10$□³□), mouse ($r = 0.3443$, p =

0.002665), dog ($r = 0.7783$, $p = 6.01 \times 10^{\square\square}$), cat ($r = 0.6672$, $p = 4.65 \times 10^{\square 11}$), and sheep ($r = 0.4139$, $p = 1.50 \times 10^{\square\square}$).

Subsequently, we assessed YCAY cluster density in the region surrounding the sequences encoding GNSK and GNNK. This assessment was performed using the methodology described by Ule et al.[38] In all animals, enrichment of the YCAY motifs was observed in the region surrounding the alternatively spliced donor sites (Figure 5). To confirm the direct binding of NOVA1/2 proteins to *KIT* pre-mRNA, we analyzed HITS-CLIP data obtained from wild-type mouse cortex samples by Saito et al.[33] The analysis revealed multiple NOVA1 and NOVA2 CLIP tags in the intron immediately downstream of alternative donor sites (Figure 6). To assess the conservation of the RNA-binding specificity of NOVA1/2 proteins among mice and four other mammals (humans, dogs, cats, and sheep), their amino acid sequences were compared. This analysis revealed isoleucine-valine substitutions at positions 106 and 197 of NOVA1. However, considering the hydrophobic nature of these residues, these substitutions seemed to cause little substantial alteration in the protein structure. The V197I substitution observed in humans has not been demonstrated to affect the sequence specificity or binding affinity to RNA.[42] In addition, the three KH domains that constitute the RNA-binding domain of NOVA2 display identical amino acid sequences across these species (Figure S5). These observations provide a rational framework for explaining the specificity of NOVA1/2-mediated regulation of splicing in mammals.

Regulation of alternative 5′ splice donor site selection by *NOVA1/2*

We ascertained whether NOVA1/2, a known regulatory factor for cassette exon inclusion/skipping, plays a role in the selection of the alternative 5’ donor site. We compared the previous results for genes undergoing alternative 5’ donor site selection, *Ccdc9* and *Ncdn*. *Ccdc9* is known to prefer to insert longer exon in the presence of NOVA1/2 proteins,[38] and we confirmed that CNS had a higher in-ex ratio than other tissues (Wilcoxon rank-sum test: $p = 9.141 \times 10^{\square 37}$) (Figure 7 and Figure S6) and the NOVA1/2 CLIP tags were densely packed within the downstream introns (Figure 8). In contrast, *Ncdn* tends to express shorter exons specifically in the presence of NOVA,[37] and we confirmed that the CNS has a lower in-ex ratio than other tissues (Wilcoxon rank-sum test: $p = 1.327 \times 10^{\square 24}$), and NOVA1/2 CLIP tags were enriched within the exon. The aforementioned scenarios in *KIT* and *Ccdc9* are analogous. Therefore, the mechanism of present-day specific splicing in *KIT* mRNA can be explained by the analogy of the case of *Ccdc9*.

Discussion

In this study, we performed a comprehensive tissue-wide analysis of the expression features of the c-KIT isoform, GNNK+/GNSK+ and GNNK-/GNSK- in five mammals (humans, mice, domestic dogs, cats, and sheep). Recent studies on human and mouse c-KITs have reported the co-expression of GNNK+ and GNNK- in multiple tissues and cell lines,[43-46] and both isoforms have already been registered as Known RefSeq entries. The expression of GNNK+ and GNNK- in various diseases, particularly tumorigenesis, has been extensively investigated.[18,45,47-49] In contrast, in dogs and sheep, only GNSK- was registered as a Known RefSeq entries, whereas GNSK+ was registered as a predicted transcript. However, elevated GNSK+ expression in the cerebellum and higher GNSK- levels in C2 mast cell tumors and angiosarcoma samples from the spleen, skin, and mammary tissues have been reported in domestic dogs.[4,6,10] In cats, only GNSK+ has been registered in RefSeq. GNSK- isoforms have also been isolated from feline spleen.[7] Consequently, the present study suggests that both GNNK+/GNSK+ and GNNK-/GNSK- are synthesized throughout the body without restriction to specific tissues or tumors and that their presence is conserved across animal species. However, the present results also suggest that significant tissue-dependent changes in the expression of these isoforms occur independently of tumorigenesis; a higher proportion of GNNK+/GNSK+ exists in the CNS, whereas GNNK-/GNSK- isoforms predominate in other tissues.

Enrichment analysis revealed that genes whose expression correlated with GNNK+/GNSK+ contained significantly more GO terms related to CNS functions. Moreover, some of these terms were specific to GNNK+, obtained only from genes whose expression correlated with GNNK+ in humans and mice, but not from genes whose expression correlated with GNSK+ in dogs, cats, and sheep, and vice versa. In other words, the GNNK+ c-KIT isoform in humans and mice and the GNSK+ isoform in dogs, cats, and sheep function in the CNS, but may have slightly different roles. However, the lack of GO terms specific to GNSK+ may be due to the lack of progress in gene annotation by GO terms for dog, cat, and sheep, compared with humans and mice. Future experiments are required to identify these terms for the development of veterinary medicine. These differences may also be due to differences in molecular activity, depending on whether the third amino acid is asparagine or serine. This is also an important issue from both the biological and clinical perspectives and should be verified in future experiments.

In this study, we also revealed the presence of a binding motif for the CNS-specific splicing regulator NOVA1/2 proteins and its binding to the downstream intron in the vicinity of the alternative 5' donor site in the *KIT* pre-mRNA. Previous studies on NOVA1/2 have primarily focused on NOVA-mediated "cassette exon" regulation, proposing RNA regulatory map in which binding downstream of the target

exon promotes exon inclusion, while binding on or upstream of the target exon induces exon skipping.[38] However, minor alternative splicing, including alternative 5′ donor sites, remains underexplored. Our comparative analysis of NOVA1/2 binding distributions and splicing patterns in *KIT*, *Ccdc9*, and *Ncdn* pre-mRNAs suggests that NOVA1/2 binding and the frequency of proximal upstream donor site usage are highly correlated. These results may provide insights into the potential of multiple RNA maturation pathways involving NOVA1/2 proteins.

Although c-KIT is expressed at high levels in the brain during development and adulthood, no overt phenotypic changes have been observed in *KIT* mutant mice, suggesting that c-KIT expression in the brain is functionally redundant.[50] However, subsequent studies have demonstrated that brain-specific *KIT* mutations induced during development suppress neural crest cell proliferation.[51,52] The present study showed that the GNNK+/GNSK+ isoform of c-KIT was excessively synthesized in the CNS compared with other tissues. In addition, the intensity, modification rate, and lifetime of phosphorylation of each tyrosine residue in c-KIT after SCF stimulation differed between GNNK+ and GNNK-, and signal transduction activity was known differently depending on cell types.[20,22] Thus, one of the roles of the GNNK+/GNSK+ isoform of c-KIT in the brain is the fine regulation of the pathway and the strength of signal transduction by adjusting the relative abundance of GNNK+/GNSK+ and GNNK-/GNSK- to contribute to brain development and morphogenesis.

The findings of this study indicate that, in both human clinical and veterinary research, it is crucial to consider tissue-specific variations in c-KIT isoform ratios when evaluating the efficacy of c-KIT-targeting agents (such as imatinib). In addition to the isoforms considered in the present study, c-KIT in sheep and cats was predicted to synthesize another type of isoform lacking the domain corresponding to exon 1 (called d1KIT in domestic cats), which might contribute to the development of the neural crest and regulation of skin coat color.[8,53] Therefore, identifying potential tissue-specific, particularly CNS-specific, functions and regulatory mechanisms of various c-KIT isoforms using in vivo and in vitro studies, as well as in silico predictions, should be a crucial focus of future research.

Acknowledgements

This work was supported by the Japan Science and Technology Agency (JST) Moonshot R&D (Grant Number JPMJMS2021), https://www.jst.go.jp/ (A. A.), and Japan Society for the Promotion of Science (JSPS) KAKENHI grants (award numbers 21K06124 and 24K09248 (A. A.)).

Author contributions

Conceptualization: R.G. and A.A., Data curation: R.G., Formal analysis: R.G., Funding acquisition: A.A., Investigation: R.G., N.S., and A.A., Methodology: R.G. and A.A., Project administration: R.G. and A.A., Supervision: N.S. and A.A., Validation: N.S. and A.A., Visualization: R.G., Writing - original draft: R.G., Writing - review & editing: N.S. and A.A.

References

1. Sheikh, E., Tran, T., Vranic, S., Levy, A., and Bonfil, R. D. 2022, Role and significance of c-KIT receptor tyrosine kinase in cancer: A review, Bosn. J. Basic. Med. Sci., 22, 683-698.

2. Roskoski, R. 2005, Structure and regulation of Kit protein-tyrosine kinase-The stem cell factor receptor, Biochem. Biophys. Res. Commun., 338, 1307-1315.

3. Krimmer, S. G., Bertoletti, N., Suzuki, Y., et al. 2023, Cryo-EM analyses of KIT and oncogenic mutants reveal structural oncogenic plasticity and a target for therapeutic intervention, Proc. Natl. Acad. Sci. U S A., **120**, e2300054120.

4. London, C. A., Galli, S. J., Yuuki, T., Hu, Z. Q., Helfand, S. C., and Geissler, E. N. 1999, Spontaneous canine mast cell tumors express tandem duplications in the proto-oncogene c-KIT, Exp. Hematol., 27, 689-697.

5. Gramer, I., Kessler, M., and Geyer, J. 2016, Detection of novel polymorphisms in the c-KIT gene of canine patients with lymphoma, melanoma, haemangiosarcoma, and osteosarcoma, Vet. Res. Commun., 40, 89-95.

6. Chen, Y.C., Liao, J. W., Hsu, W. L., and Chang, S. C. 2016, Identification of the two KIT isoforms and their expression status in canine hemangiosarcomas, BMC Vet. Res., 12, 142.

7. Hasegawa, Y., Shosu, K., Tsuji, K., et al. 2022, Intratumoral heterogeneity of c-KIT mutations in a feline splenic mast cell tumor and their functional effects on cell proliferation, Sci. Rep., 12, 15791.

8. Saravanaperumal, S. A., Pallotti, S., Pediconi, D., Renieri, C., and La Terza, A. 2021, Exon-1 skipping and intron-1 retaining by alternative splicing of the c-KIT gene encodes a novel splice variant in the skin of Merino sheep (Ovis aries), Mol. Biol. Rep., 48, 4987-4994.

9. Goldfarb, T., Kodali, V. K., Pujar, S., et al. 2025, NCBI RefSeq: Reference sequence standards through 25 years of curation and annotation, Nucleic Acids Res., 53, D243-D257.

10. Chen, Y. C., Liao, J. W., Chang, S. C., and Hsu, W. L. 2020, Expression frequency of c-KIT isoforms and its prognostic potential in canine mammary tumours, Vet. Comp. Oncol., 18, 303-314.

11. Hirota, S., Isozaki, K., Moriyama, Y., et al. 1998, Gain-of function mutations of c-KIT in human gastrointestinal stromal tumors, Science., 279, 577-580.

12. Gregory-Bryson, E., Bartlett, E., Kiupel, M., Hayes, S., and Yuzbasiyan-Gurkan, V. 2010, Canine and human gastrointestinal stromal tumors display similar mutations in c-KIT exon 11, BMC Cancer, 10, 559.

13. Isotani, M., Yamada, O., Lachowicz, J. L., et al. 2010, Mutations in the fifth immunoglobulin-like domain of kit are common and potentially sensitive to imatinib mesylate in feline mast cell tumours, Br. J. Haematol., 148, 144-153.

14. Webster, J. D., Yuzbasiyan-Gurkan, V., Kaneene, J. B., Miller, R., Resau, J. H., and Kiupel, M. 2006, The role of c-KIT in tumorigenesis: Evaluation in canine cutaneous mast cell tumors, Neoplasia, 8, 104-111.

15. Nagata, H., Worobec, A. S., Oh, C. K., et al. 1995, Identification of a point mutation in the catalytic domain of the protooncogene c-KIT in peripheral blood mononuclear cells of patients who have mastocytosis with an associated hematologic disorder, Proc. Natl. Acad. Sci. U S A., 92, 10560-10564.

16. Garcia-Montero, A. C., Jara-Acevedo, M., Teodosio, C., et al. 2006, KIT mutation in mast cells and other bone marrow hematopoietic cell lineages in systemic mast cell disorders: A prospective study of the Spanish Network on Mastocytosis (REMA) in a series of 113 patients, Blood, 108, 2366-2372.

17. Beghini, A., Ripamonti, C.B., Cairoli, R. et al. 2004, KIT activating mutations: Incidence in adult and pediatric acute myeloid leukemia, and identification of an internal tandem duplication, Haematologica, 89, 920-925.

18. Guerrini, F., Galimberti, S., Ciabatti, E., et al. 2007, Molecular detection of GNNK− and GNNK+ c-KIT isoforms: A new tool for risk stratification in adult acute myeloid leukaemia, Leukemia, 21, 2056-2058.

19. Caruana, G., Cambareri, A. C., and Ashman, L. K. 1999, Isoforms of c-KIT differ in activation of signalling pathways and transformation of NIH3T3 fibroblasts, Oncogene, 18, 5573-5581.

20. Voytyuk, O., Lennartsson, J., Mogi, A., et al. 2003, Src family kinases are involved in the differential signaling from two splice forms of c-KIT, J. Biol. Chem., 278, 9159-9166.

21. Young, S. M., Cambareri, A. C., Odell, A., Geary, S. M., and Ashman, L. K. 2007, Early myeloid cells expressing c-KIT isoforms differ in signal transduction, survival, and chemotactic responses to stem cell factor, Cell Signal., 19, 2572-2581.

22. Sun, J., Pedersen, M., and Rönnstrand, L. 2008, Gab2 is involved in differential phosphoinositide 3-kinase signaling by two splice forms of c-KIT, J. Biol. Chem., 283, 27444-27451.

23. Phung, B., Steingrímsson, E., and Rönnstrand, L. 2013, Differential activity of c-KIT splice forms is controlled by extracellular peptide insert length, Cell Signal., 25, 2231-2238.

24. Bell, C. A., Tynan, J. A., Hart, K. C., Meyer, A. N., Robertson, S. C., and Donoghue, D. J. 2000, Rotational coupling of the transmembrane and kinase domains of the neu receptor tyrosine kinase, Mol. Biol. Cell., 11, 3589-3599.

25. ENCODE Project Consortium. 2012, An integrated encyclopedia of DNA elements in the human genome, Nature, 489, 57-74.

26. Luo, Y., Hitz, B. C., Gabdank, I., et al. 2020, New developments on the Encyclopedia of DNA Elements (ENCODE) data portal, Nucleic Acids Res., 48, D882-D889.

27. Matsumoto, Y., Yik-Lok Chung, C., Isobe, S., et al. 2024, Chromosome-scale assembly with improved annotation provides insights into breed-wide genomic structure and diversity in domestic cats, J. Adv. Res., S2090-1232, 00478-8..

28. Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018, Fastp: An ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, 34, i884-1890.

29. Dobin, A., Davis, C. A., Schlesinger, F., et al. 2013, STAR: Ultrafast universal RNA-seq aligner, Bioinformatics, 29, 15-21.

30. Li, B., and Dewey, C. N. 2011, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics, 12, 323.

31. Ashburner, M., Ball, C. A., Blake, J. A., et al. 2000, Gene ontology: Tool for the unification of biology, Nat. Genet., 25, 25-29.

32. The Gene Ontology Consortium; Aleksander, S. A., Balhoff, J., et al. 2023, The gene ontology knowledgebase in 2023, Genetics, 224, iyad031.

33. Saito, Y., Miranda-Rottmann, S., Ruggiu, M., et al. 2016, NOVA2-mediated RNA regulation is required for axonal pathfinding during development, Elife, 5, e14371.

34. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. 1978, Ovalbumin gene: Evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries, Proc. Natl. Acad. Sci U S A., 75, 4853-4857.

35. McManus, C. J., and Graveley, B. R. 2011, RNA structure and the mechanisms of alternative splicing, Curr. Opin. Genet. Dev, 21, 373-379.

36. Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. 2003, CLIP identifies nova-regulated RNA networks in the brain, Science, 302, 1212-1215.

37. Ule, J., Ule, A., Spencer, J., et al. 2005, Nova regulates brain-specific splicing to shape the synapse, Nat. Genet., 37, 844-52.

38. Ule, J., Stefani, G., Mele, A., et al. 2006, An RNA map predicting Nova-dependent splicing regulation, Nature, 444, 580-586.

39. Licatalosi, D. D., Mele, A., Fak, J. J., et al. 2008, HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature, 456, 464-9.

40. Zhang, C., Frias, M. A., Mele, A., et al. 2010, Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls, Science, 329, 439-43.

41. Fisher, E., and Feng, J. 2022, RNA splicing regulators play critical roles in neurogenesis, Wiley Interdiscip. Rev. RNA, 13, e1728.

42. Tajima, Y., Vargas, C. D. M., Ito, K., et al. 2025, A humanized NOVA1 splicing factor alters mouse vocal communications, Nat. Commun., 16, 1542.

43. Crosier, P., Ricciardi, S., Hall, L., Vitas, M., Clark, S., and Crosier, K. 1993, Expression of isoforms of the human receptor tyrosine kinase c-KIT in leukemic cell lines and acute myeloid leukemia, Blood, 82, 1151-1158.

44. Zhu, W. M., Dong, W. F., and Minden, M. 1994, Alternate splicing creates two forms of the human kit protein, Leuk. Lymphoma, 12, 441-447.

45. Sakuma, Y., Sakurai, S., Oguni, S., Hironaka, M., and Salto, K. 2003, Alterations of the c-KIT gene in testicular germ cell tumors, Cancer Sci., 94, 486-91.

46. Tuck, A. R., Robker, R. L., Norman, R. J., Tilley, W. D., and Hickey, T. E. 2015, Expression and localisation of c-KIT and KITL in the adult human ovary, J. Ovarian Res., 8, 31.

47. Théou, N., Tabone, S., Saffroy, R., et al. 2004, High expression of both mutant and wild-type alleles of c-KIT in gastrointestinal stromal tumors, Biochim. Biophys. Acta., 1688, 250-256.

48. Chan, E. C., Bai, Y., Bandara, G., et al. 2013, KIT GNNK splice variants: Expression in systemic mastocytosis and influence on the activating potential of the D816V mutation in mast cells, Exp. Hematol., 41, 870-881.e2.

49. Lebedev, T. D., Vagapova, E. R., Popenko, V. I., Leonova, O. G., Spirin, P. V., and Prassolov, V. S. 2019, Two receptors, two isoforms, two cancers: Comprehensive analysis of KIT and TrkA expression in neuroblastoma and acute myeloid leukemia, Front. Oncol., 9, 1046.

50. Motro, B., Kooy, D. V. D., Rossant, J., Reith, A., and Bernstein, A. 1991, Contiguous patterns of c-KIT and steel expression: Analysis of mutations at the W and Sl loci, Development, 113, 1207-1221.

51. Aoki, H., Hara, A., and Kunisada, T. 2017, Induced haploinsufficiency of Kit receptor tyrosine kinase impairs brain development, JCI Insight, 2**,** e94385.

52. Minei, R., Aoki, H., Ogura, A., and Kunisada, T. 2023, Compensatory gene expression potentially rescues impaired brain development in Kit mutant mice, Sci. Rep., 13, 4166.

53. Awazu, A., Takemoto, D., Watanabe, K., and Sakamoto, N. 2023, Possibilities of skin coat color-dependent risks and risk factors of squamous cell carcinoma and deafness of domestic cats inferred via RNA-seq data, Genes. Cells, 28, 893-905.

Figure legends

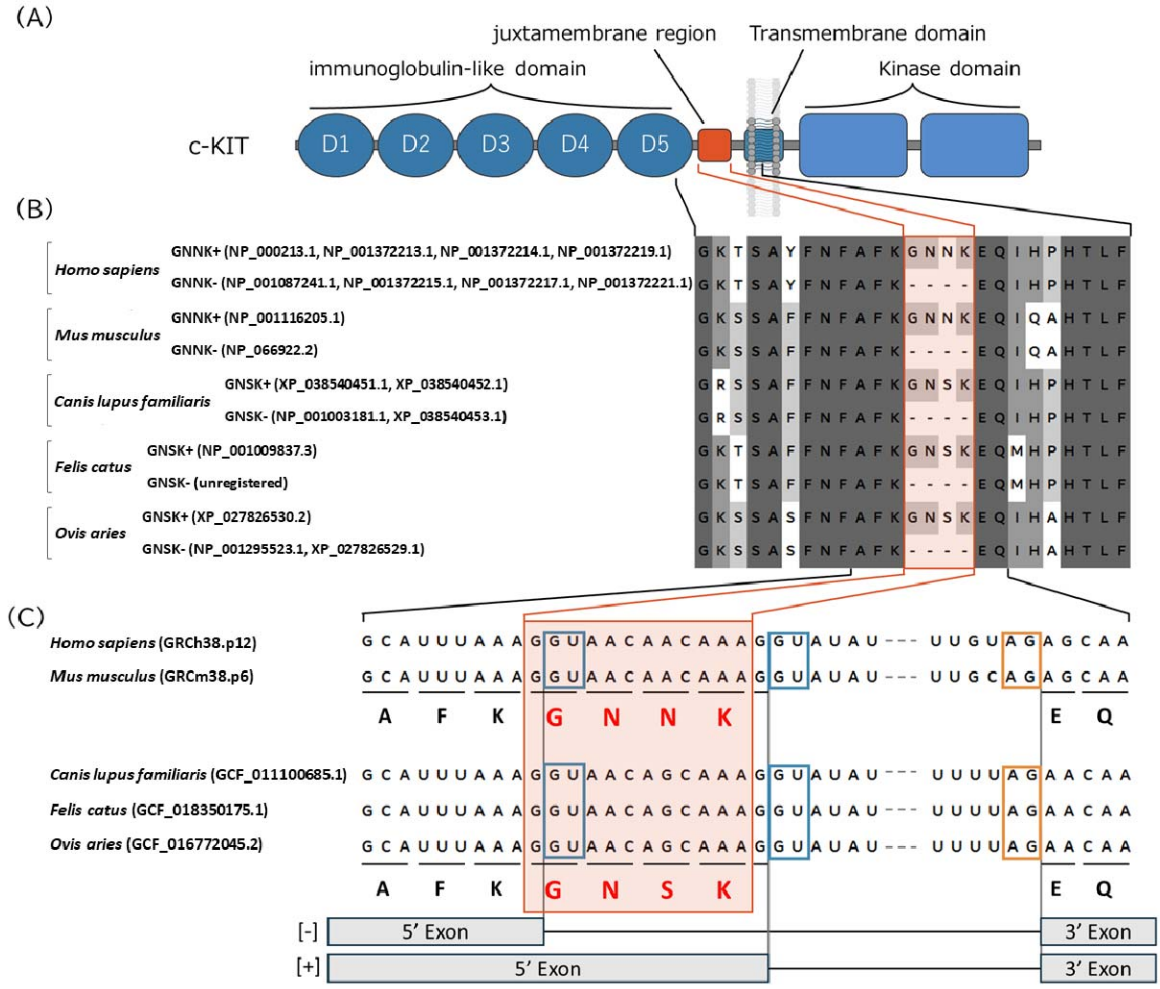
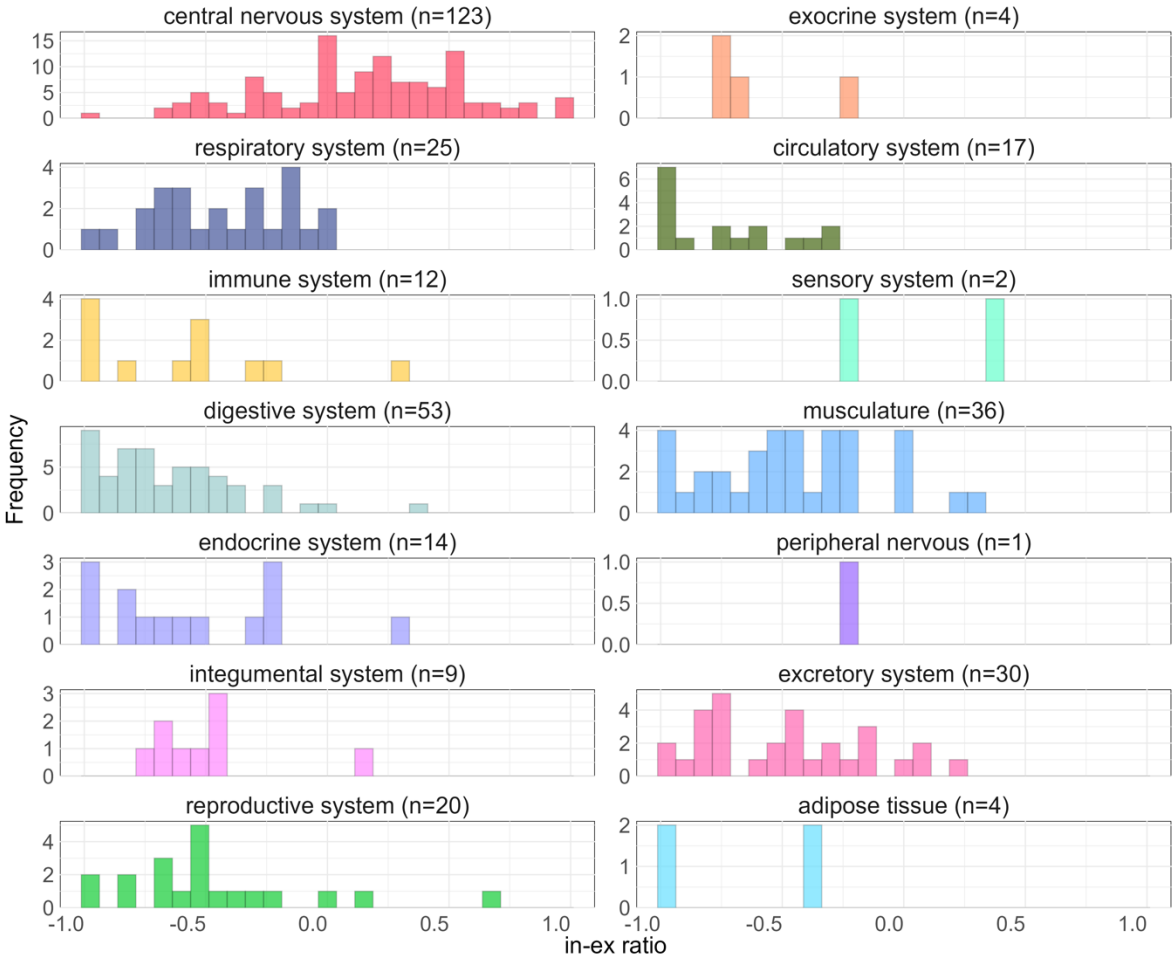


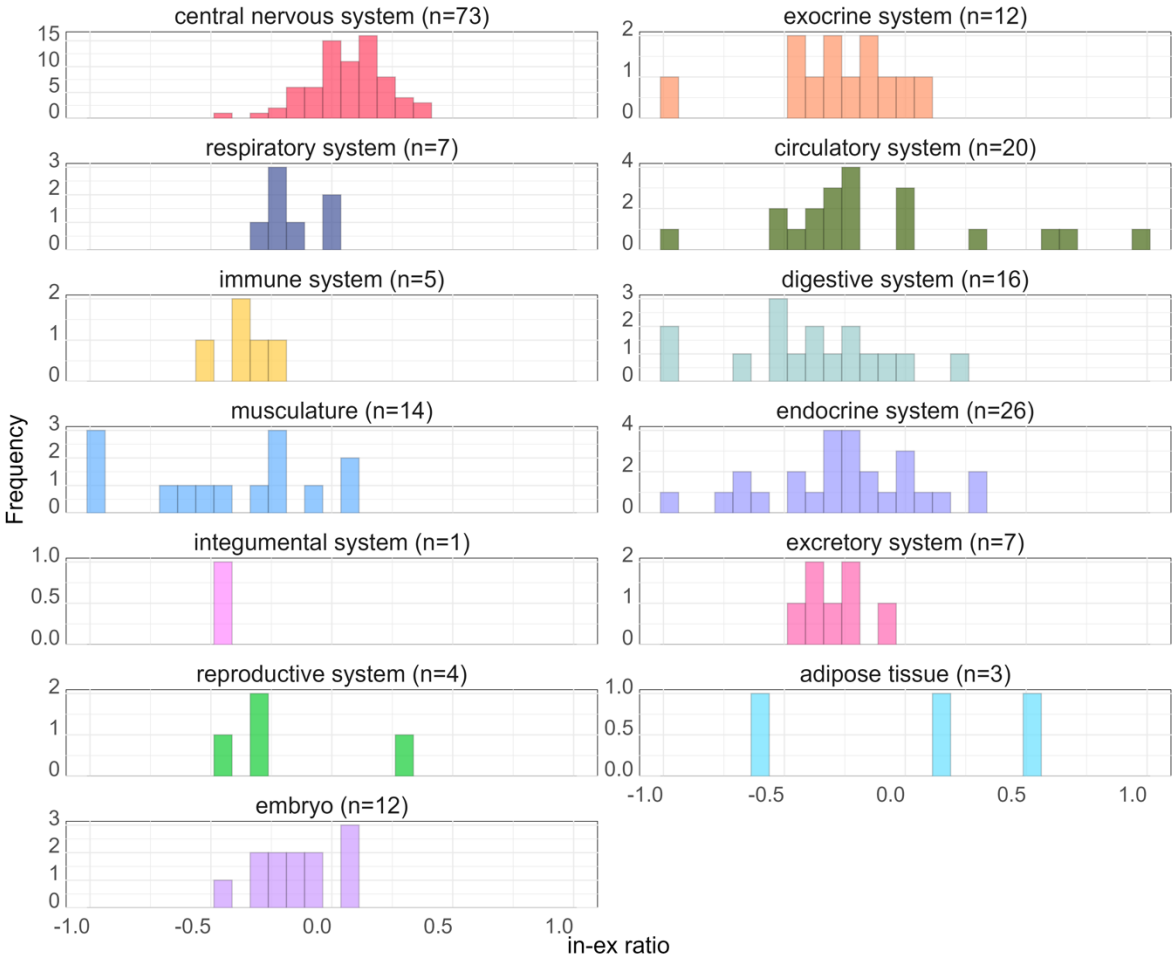
Figure 1. Structural features and sequences of c-KIT across five mammals. (A) Schematic representation of the structure of c-KIT protein. Blue indicates the functional domain of c-KIT, and orange indicates the location of the four amino acid sequence, GNNK/GNSK. (B) Amino acid sequences around the juxtamembrane region of each species. The numbers in parentheses indicate the accession number in RefSeq, and the prefix "NP_" represents curated Known RefSeq, while "XP_" represents computationally predicted Model RefSeq. (C) Nucleotide sequence of exon-intron boundaries around the region encoding the juxtamembrane region, arranged based on the genome assembly of each species. The sequences enclosed in squares indicate splice sites.

(A) *Homo sapiens*



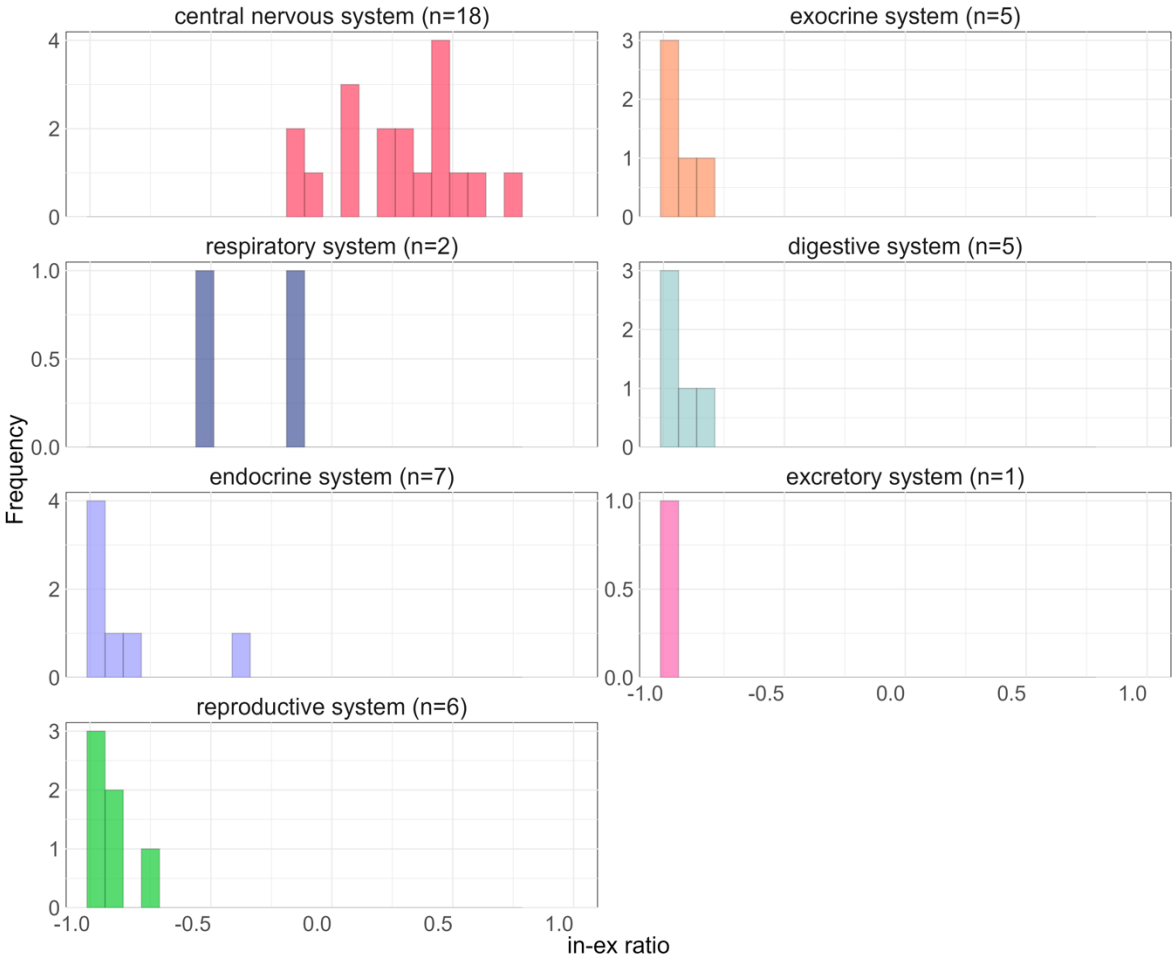
468

(B) *Mus musculus*



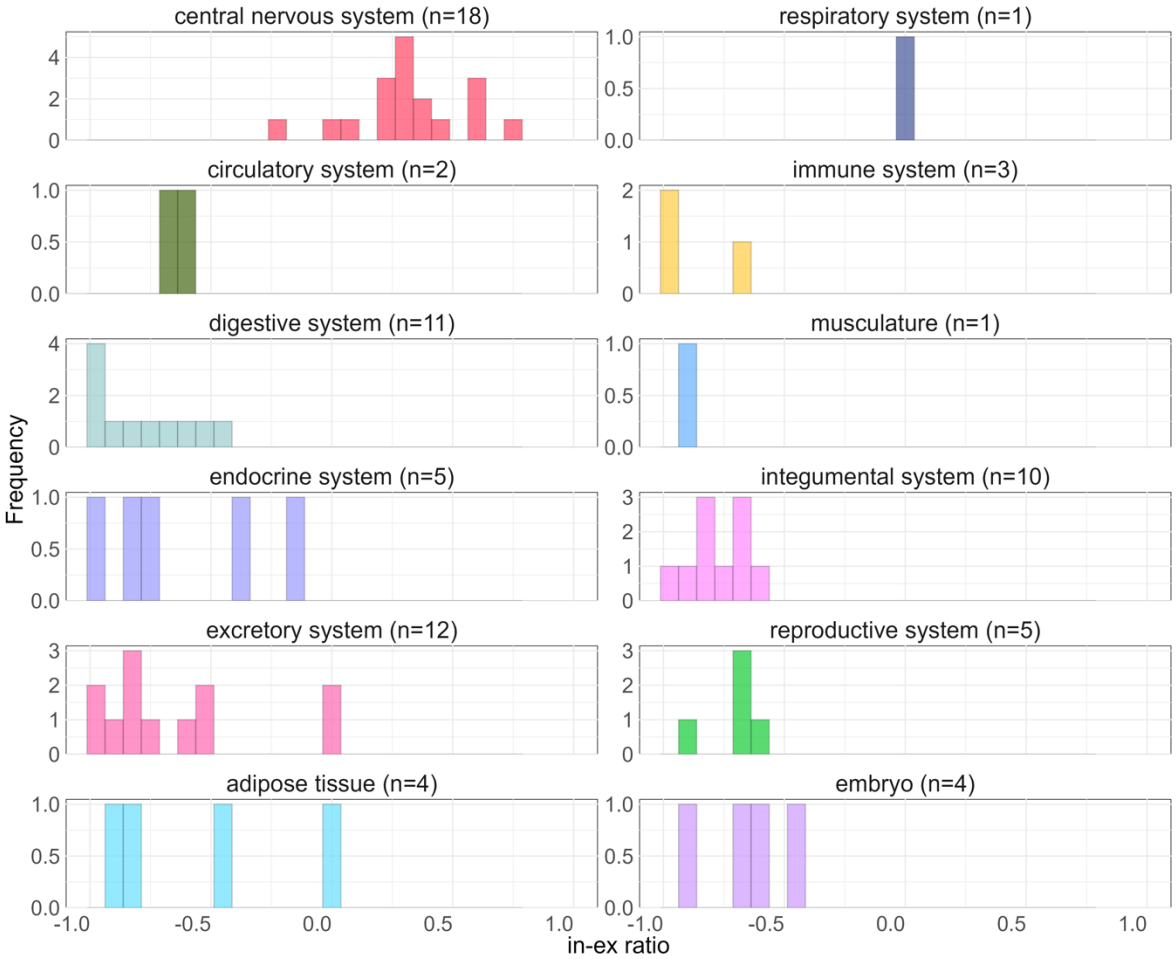
469

(C) *Canis lupus familiaris*



470

(D) *Felis catus*



471

(E) *Ovis aries*

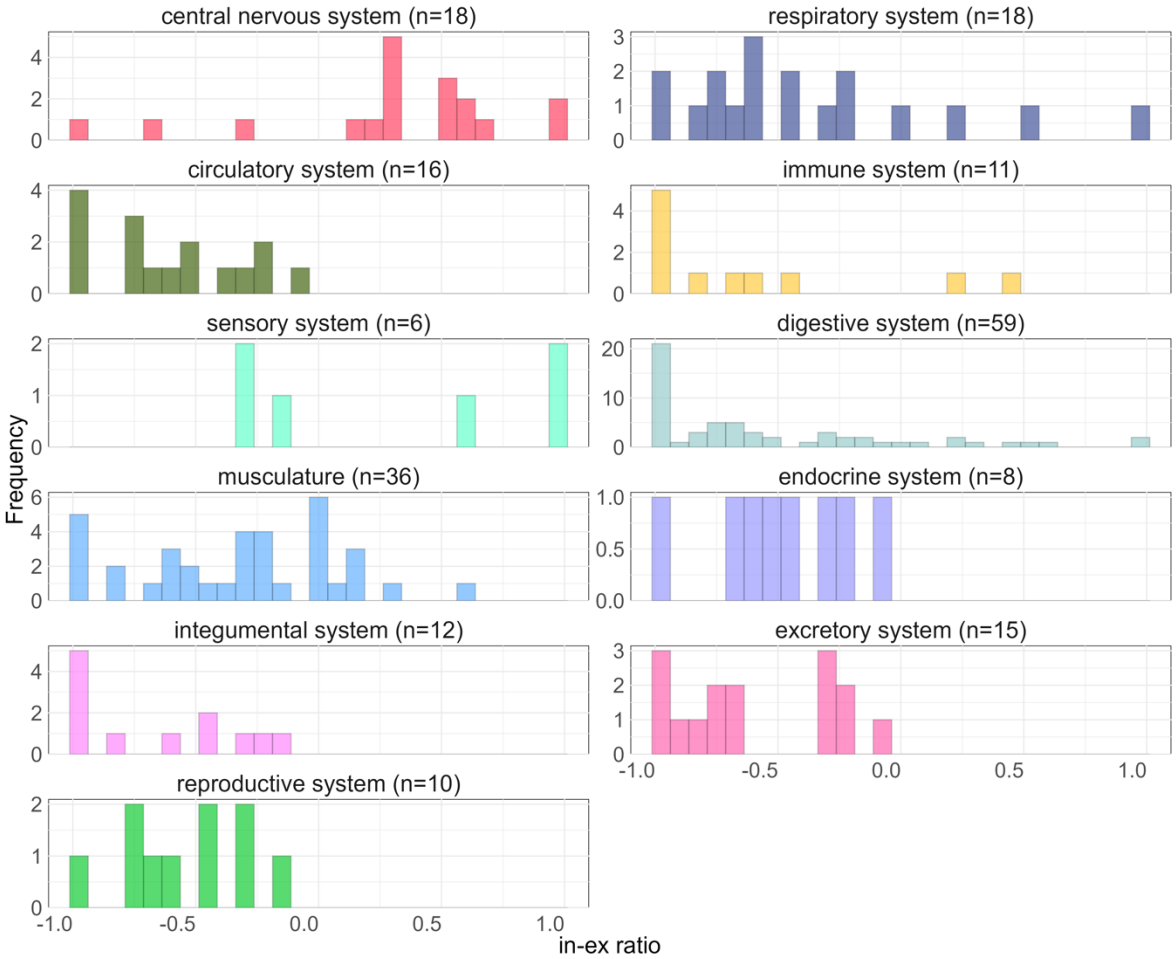
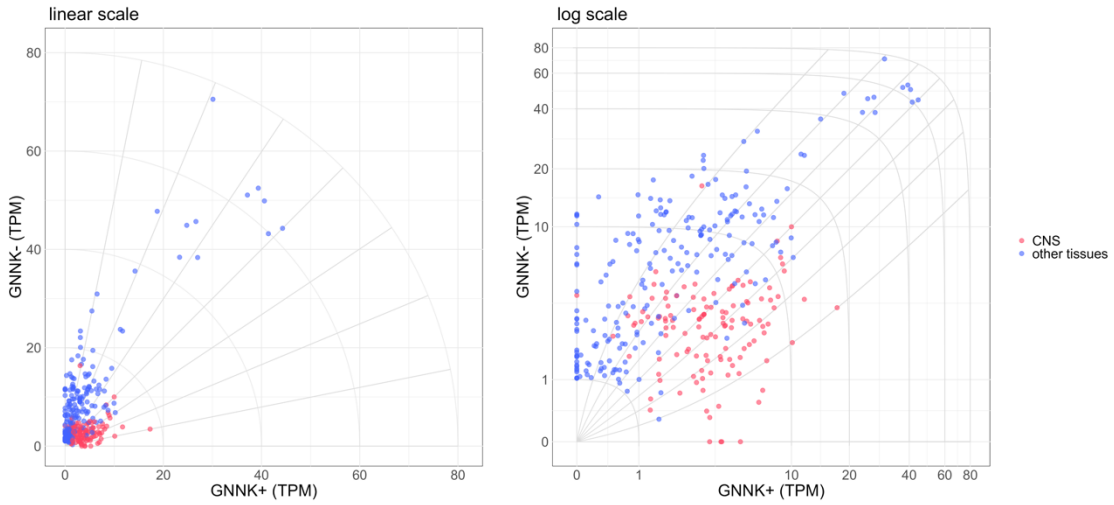


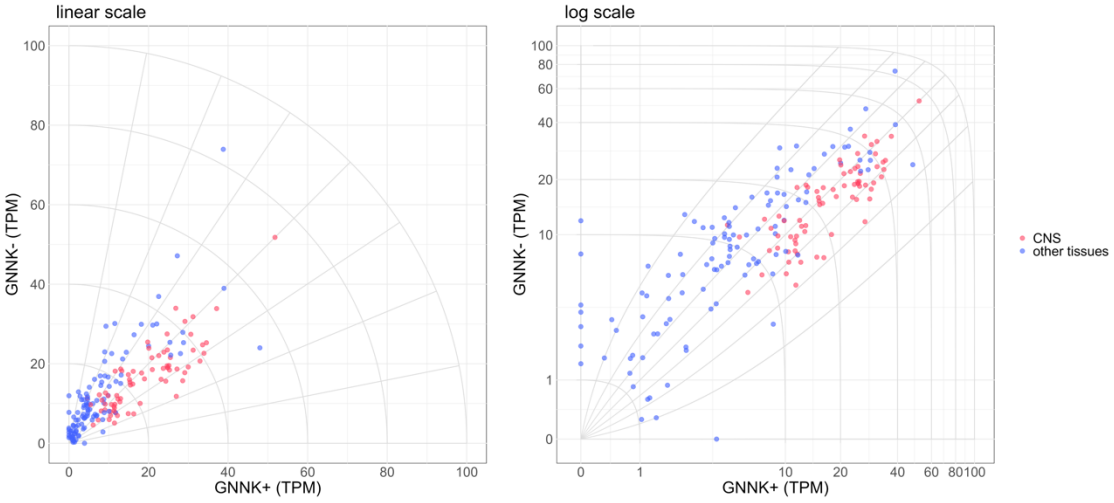
Figure 2. Histograms of the in-ex ratio of each tissue. Histograms of in-ex ratios for human (*Homo sapiens*) (A), mouse (*Mus musculus*) (B), dog (*Canis lupus familiaris*), cat (*Felis catus*) (D), and sheep (*Ovis aries*) (E). "n" indicates the number of samples for each histogram.

(A) *Homo sapiens* [*KIT*]



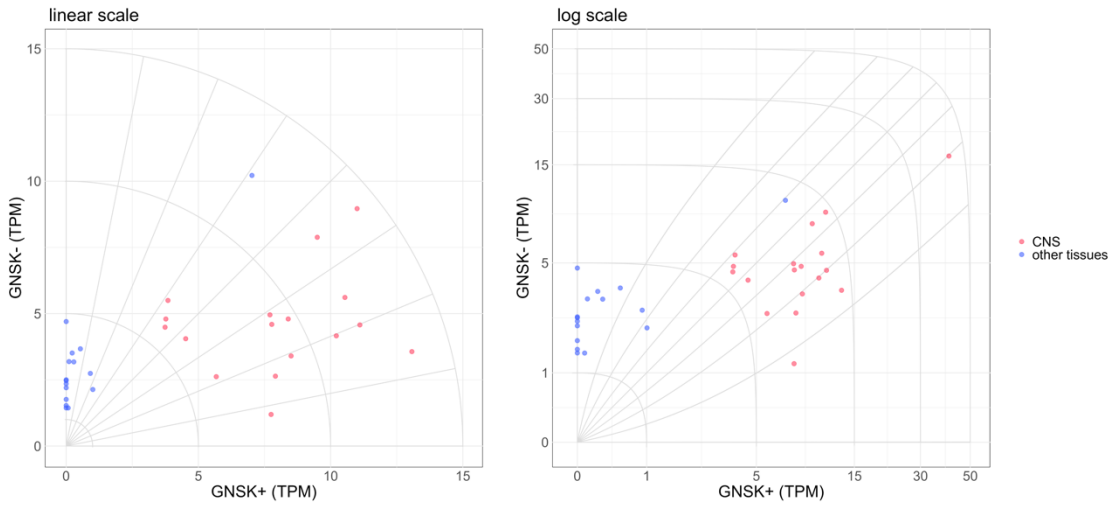
476

(B) *Mus musculus* [*KIT*]



477

(C) *Canis lupus familiaris* [*KIT*]



478

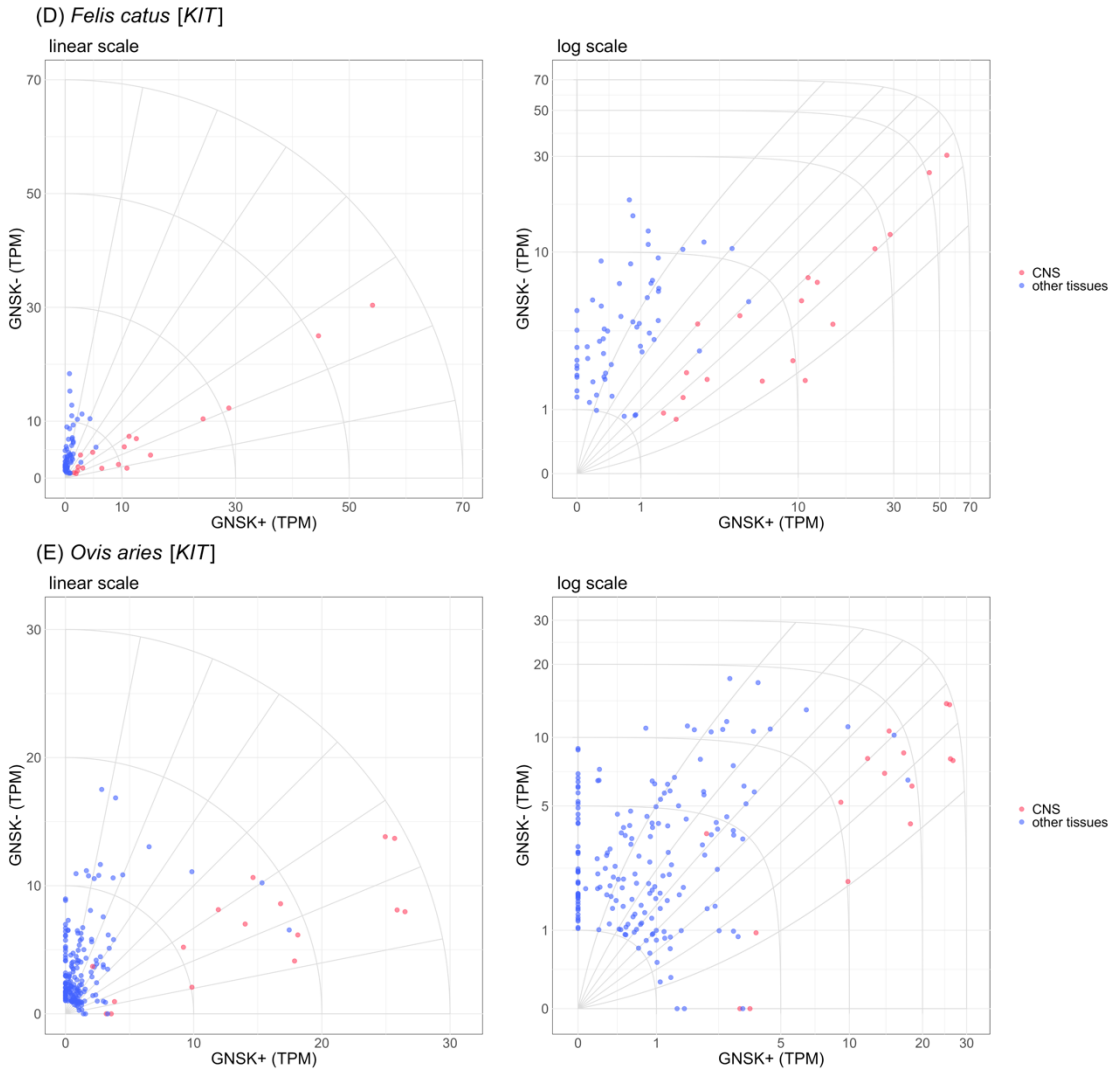
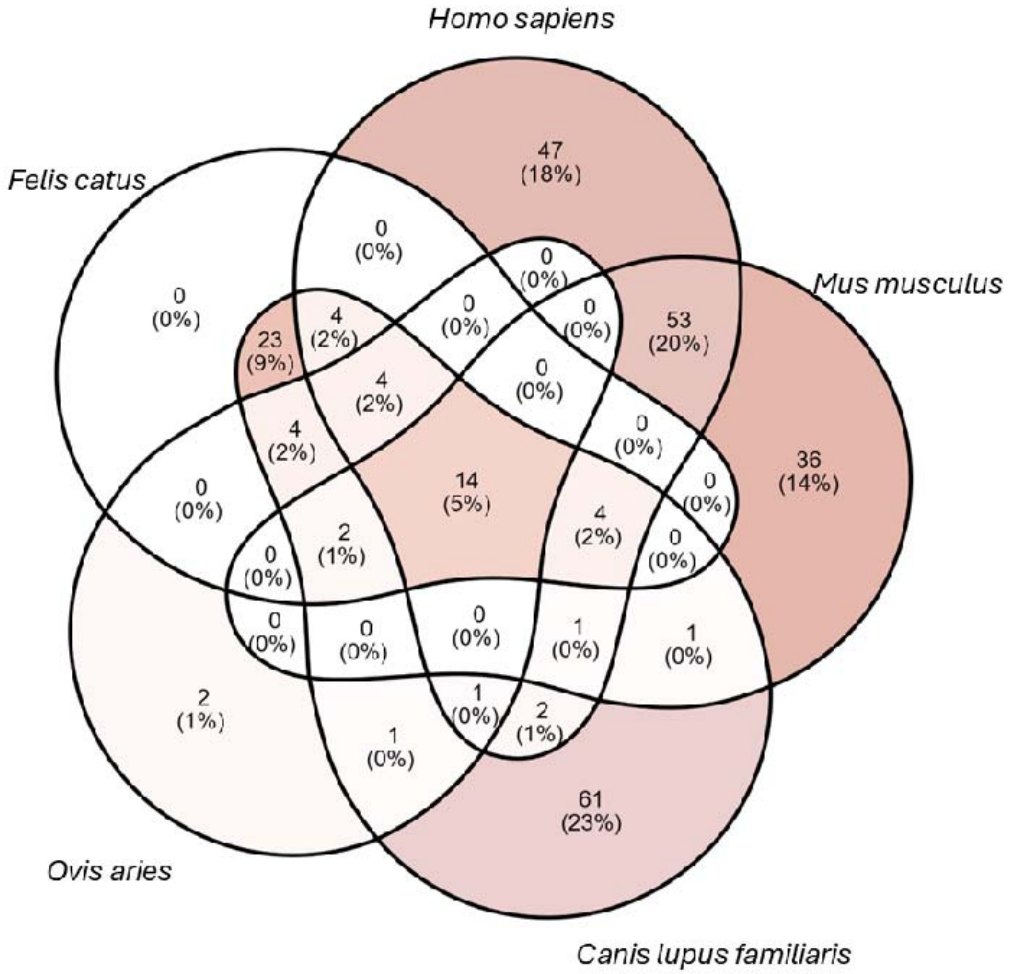


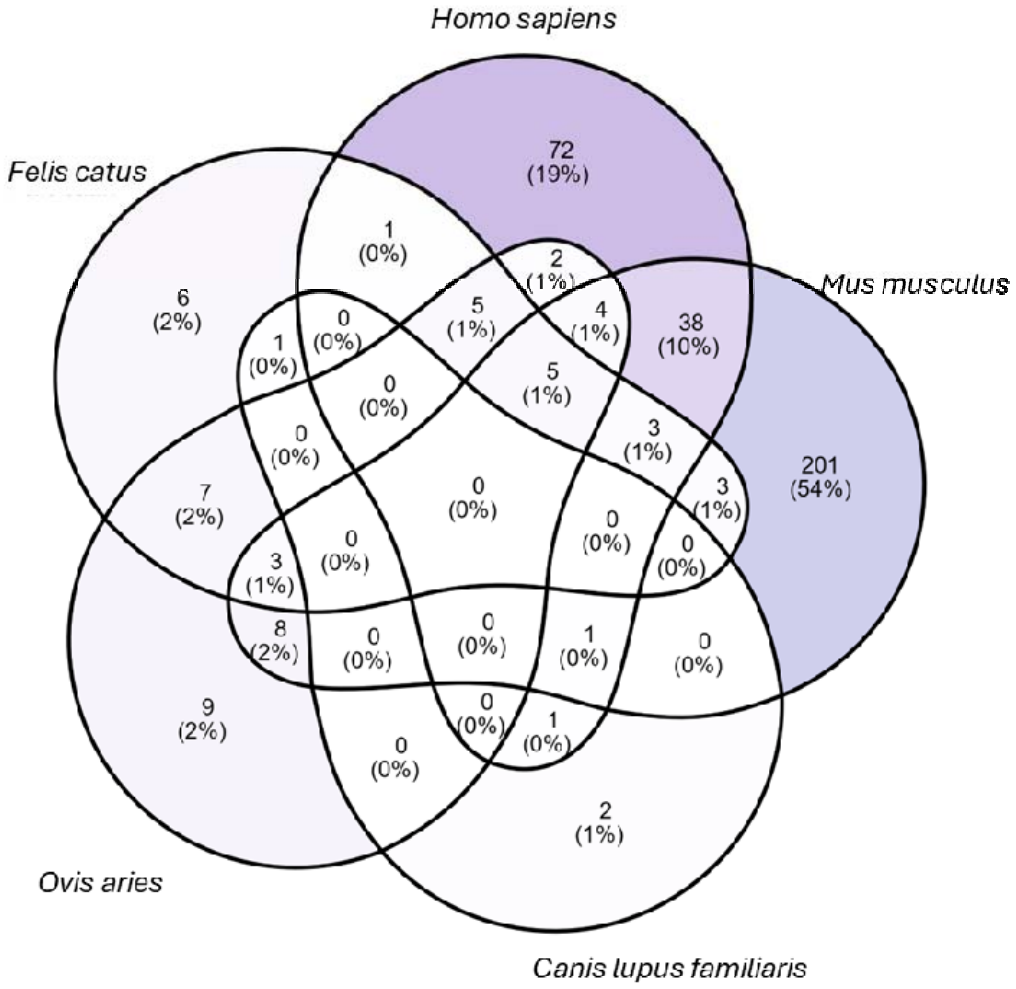
Figure 3. Expression levels of GNNK+/GNSK+ and GNNK/GNSK- in CNS and other tissues for five mammals. Scatter plots of TPM values (lefts) and ln(TPM+1) values (rights) of GNNK+/GNSK+ and GNNK-/GNSK- in humans (A), mice (B), dog (C), cat (D), and sheep (E). Red and blue points indicate data obtained from CNS samples and other tissue samples, respectively. For dogs, one more data point with TPMs of (GNSK+ and GNSK- = 40.21 and 16.46, respectively) was obtained from the CNS (SRR10915306) but excluded from the plot in linear scale on the left side for convenience. CNS, central nervous system; TPM, transcripts per million.

(A)


Homo sapiens
Felis catus
Mus musculus
Ovis aries
Canis lupus familiaris
47 (18%)
53 (20%)
36 (14%)
61 (23%)
14 (5%)
23 (9%)
0 (0%)
4 (2%)
2 (1%)
1 (0%)

488

(B)


Homo sapiens
Felis catus
Mus musculus
Ovis aries
Canis lupus familiaris
72 (19%)
1 (0%)
2 (1%)
6 (2%)
0 (0%)
5 (1%)
4 (1%)
38 (10%)
1 (0%)
0 (0%)
5 (1%)
0 (0%)
3 (1%)
201 (54%)
7 (2%)
0 (0%)
3 (1%)
0 (0%)
0 (0%)
0 (0%)
3 (1%)
8 (2%)
0 (0%)
0 (0%)
1 (0%)
0 (0%)
9 (2%)
0 (0%)
1 (0%)
0 (0%)
2 (1%)

489

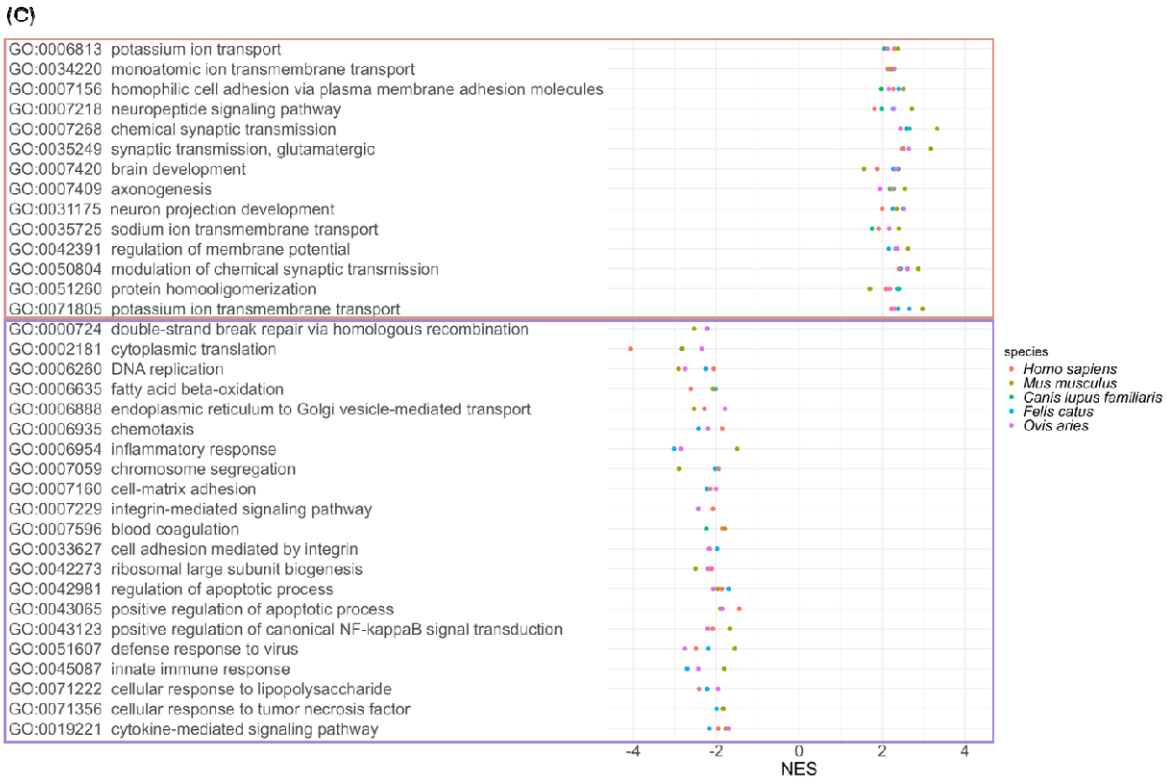
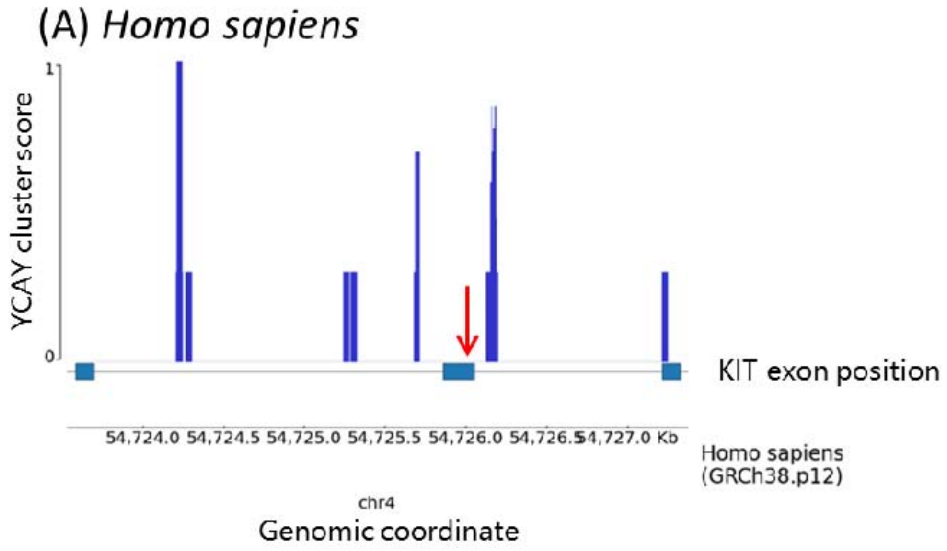
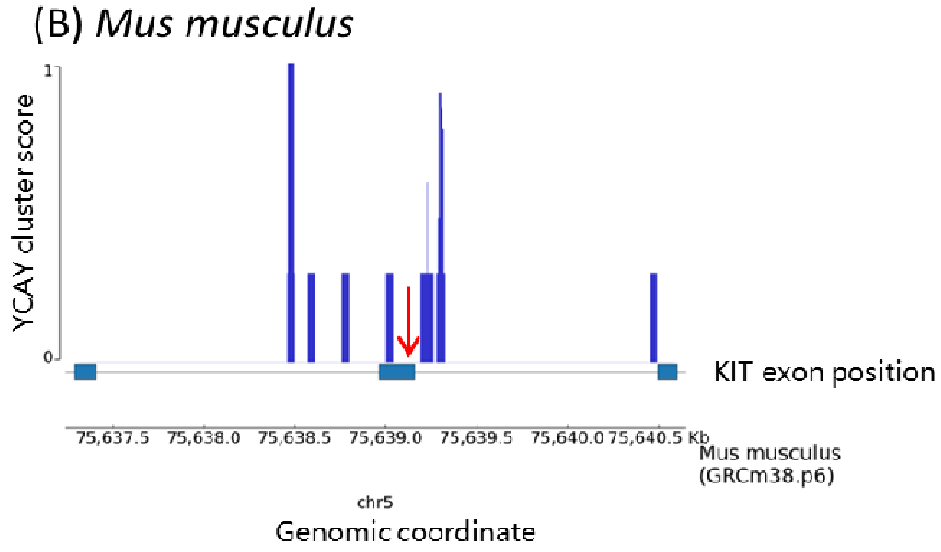


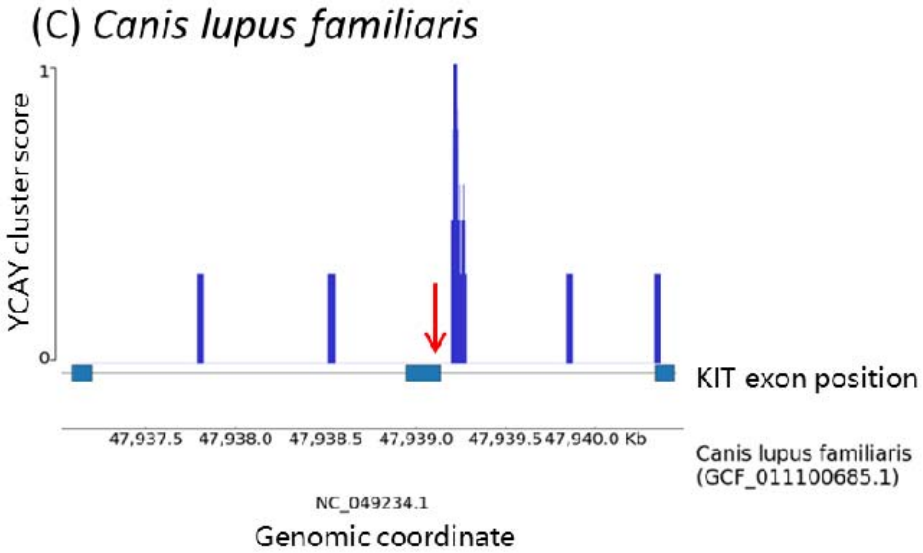
Figure 4. Cross-species comparison of enriched GO terms. Enrichment analysis was performed based on a list of genes sorted by the correlation coefficient between the in-ex ratio and gene expression level (TPM). Venn diagram shows the overlap between species for GO terms that meet the threshold (p-value < 0.01). Count of GO terms with positive normalized enrichment score (NES) (A), Count of GO terms with negative NES (B). Distribution of NES for GO terms and their descriptions that are shared across all five species (as identified in panel A) (C, top). Distribution of NES for GO terms and descriptions that are common to at least three species (as identified in panel B) (C, bottom). GO, gene ontology; TPM, transcripts per million.


(A) Homo sapiens
YCAY cluster score
1
0
KIT exon position
54,724.0 54,724.5 54,725.0 54,725.5 54,726.0 54,726.5 54,727.0 Kb
Homo sapiens
(GRCh38.p12)
chr4
Genomic coordinate

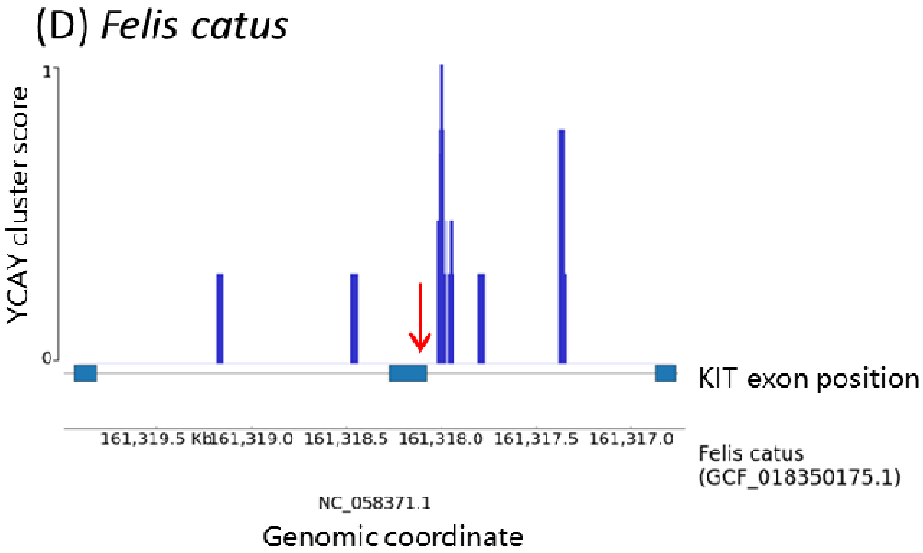
499


(B) Mus musculus
YCAY cluster score
1
0
KIT exon position
75,637.5 75,638.0 75,638.5 75,639.0 75,639.5 75,640.0 75,640.5 Kb
Mus musculus
(GRCm38.p6)
chr5
Genomic coordinate

500


(C) Canis lupus familiaris
YCAY cluster score
1
0
KIT exon position
47,937.5 47,938.0 47,938.5 47,939.0 47,939.5 47,940.0 Kb
Canis lupus familiaris
(GCF_011100685.1)
NC_049234.1
Genomic coordinate

501


(D) Felis catus
YCAY cluster score
1
0
KIT exon position
161,319.5 Kb 161,319.0 161,318.5 161,318.0 161,317.5 161,317.0
Felis catus
(GCF_018350175.1)
NC_058371.1
Genomic coordinate

502

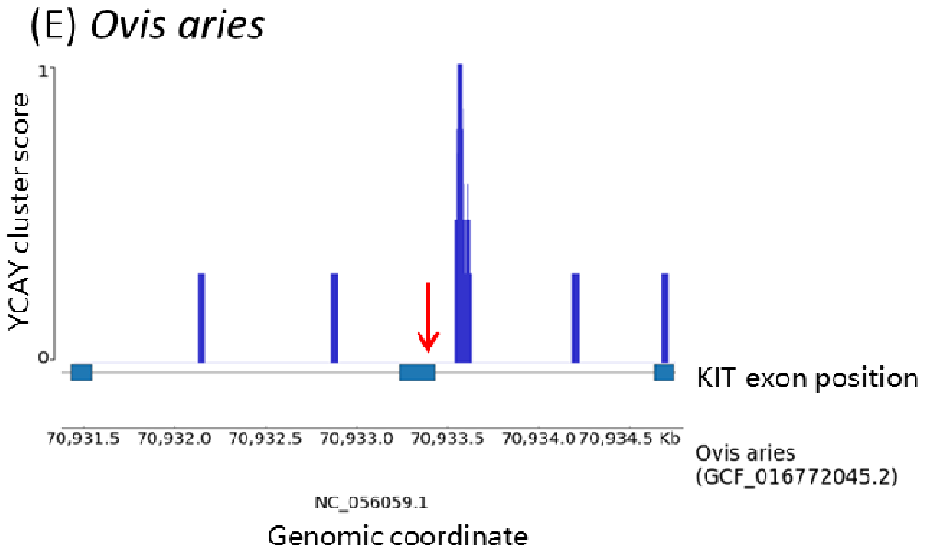


Figure 5. Distribution of YCAY cluster scores around genomic regions encoding GNNK+/GNSK+ and GNNK-/GNSK-. Distribution of YCAY cluster scores around the sequences encoding GNSK and GNNK in reference genomes of human (A), mouse (B), dog (C), cat (D), and sheep (E). The red arrow indicates the position of the alternative donor site.

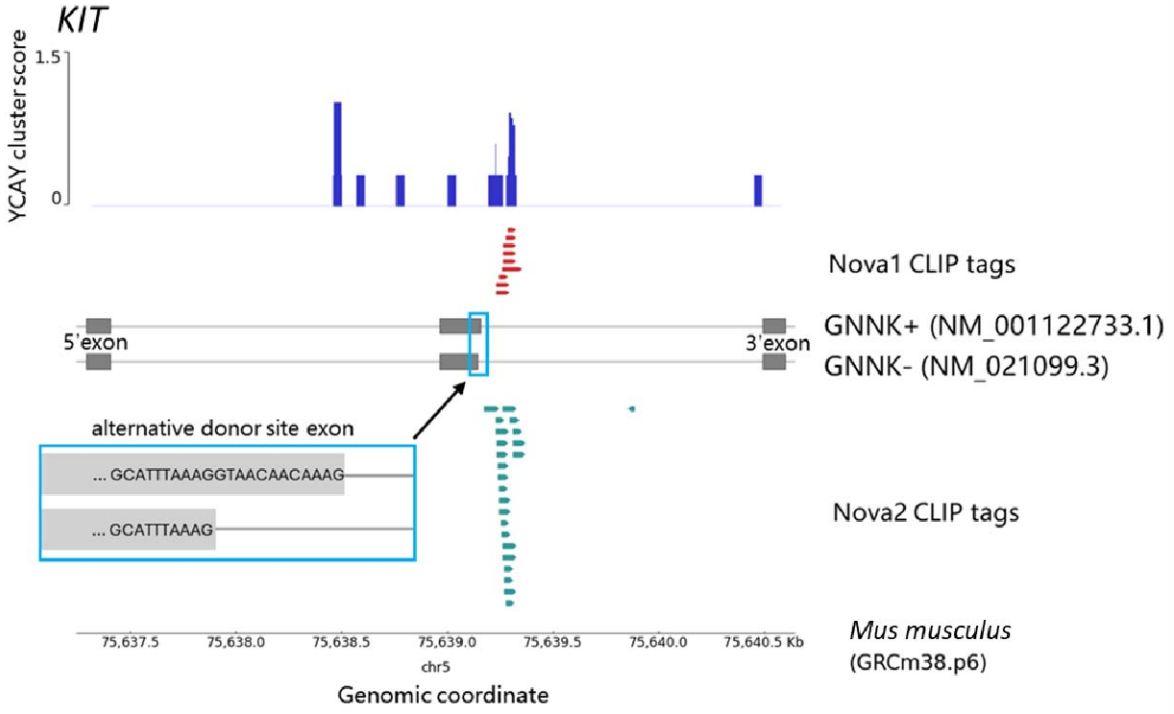


Figure 6. Distribution of binding tags of NOVA1/2 proteins obtained from CNS tissues in the mouse. Mapped position of each read of the binding tag of NOVA1 (red bar) and NOVA2 (green bar) around the *KIT* pre-mRNA GNNK coding regions in mouse, which were aligned with the distribution of YCAY cluster scores (blue bars). CNS, central nervous system.

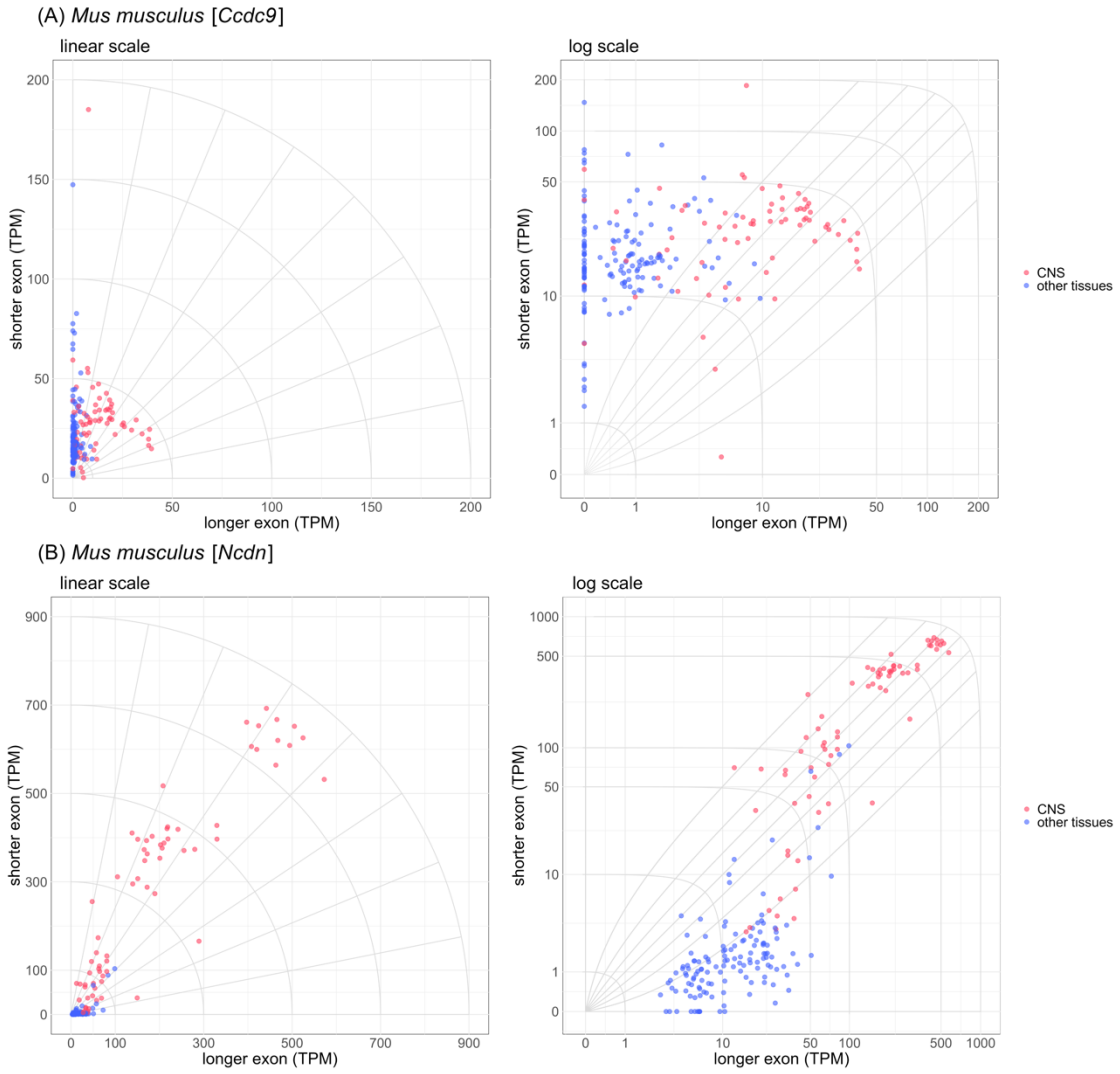
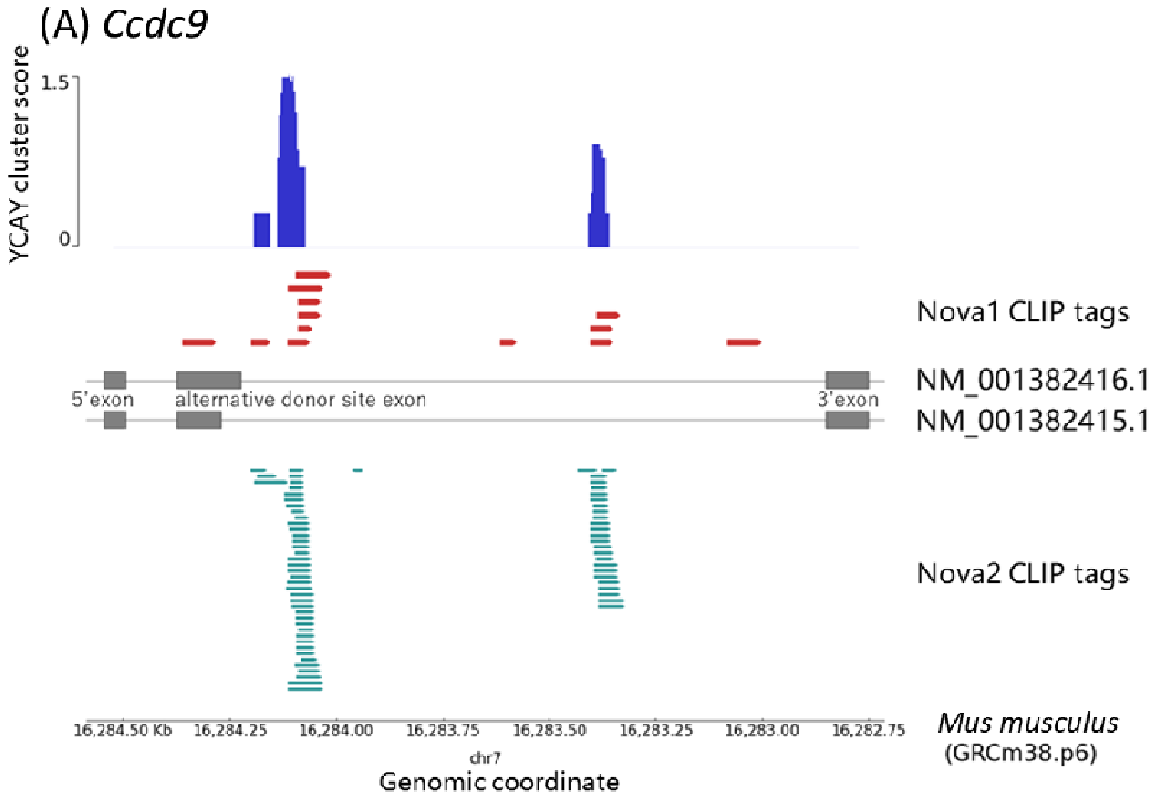


Figure 7. Expression levels of longer and shorter exons of *Ccdc9* and *Ncdn* in CNS and other tissues. Scatter plots of TPM values (lefts) and ln(TPM+1) values (rights) of longer and shorter exons of *Ccdc9* (A) and *Ncdn* (B) in the mouse. Red and blue points indicate data obtained from CNS samples and other tissue samples, respectively. CNS, central nervous system.


(A) Ccdc9
YCAY cluster score
1.5
0
Nova1 CLIP tags
NM_001382416.1
NM_001382415.1
5'exon
alternative donor site exon
3'exon
Nova2 CLIP tags
16,284.50 Kb
16,284.25
16,284.00
16,283.75
16,283.50
16,283.25
16,283.00
16,282.75
chr7
Genomic coordinate
Mus musculus
(GRCm38.p6)

520

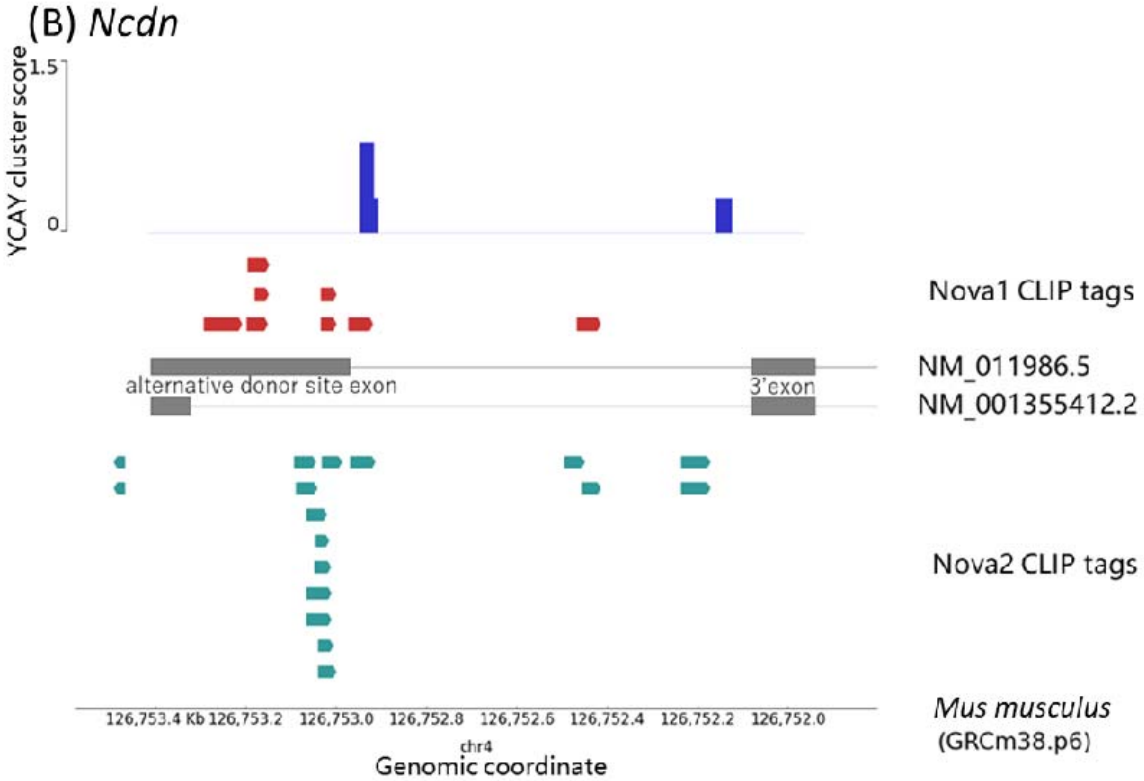


Figure 8. Distribution of binding tags of NOVA1/2 proteins within *Ccdc9* and *Ncdn* obtained from CNS tissues in mice. Mapped position of each read of binding tag of NOVA1 (red bar) and NOVA2 (green bar) around the alternative 5' donor exon *Ccdc9* (A) and *Ncdn* (B) in mouse, which were aligned with the distribution of YCAY cluster scores (blue bars). CNS, central nervous system.