Supplementary Information for Comparative Analysis of Mammalian Y Chromosomes Illuminates Ancestral Structure and Lineage-Specific Evolution

Gang Li, Brian W. Davis, Terje Raudsepp, Alison J. Pearks Wilkerson, Victor C. Mason, Malcolm Ferguson-Smith, Patricia C. O'Brien, Paul Waters and William J. Murphy.

Figures 1-12 Pgs. 2-16

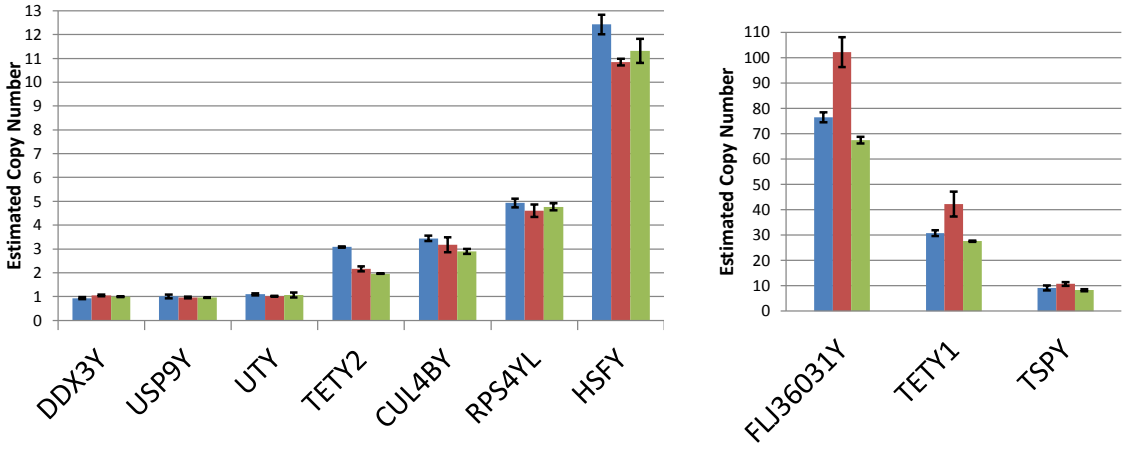
Tables 1-3 * Pgs. 17-19

Supplemental Methods Pgs. 20-24

*Supplementary Table 4 is attached as a separate Excel file

Figure 1. Copy number analysis of felid Y chromosome genes using qPCR and FISH

A. Quantitative PCR analysis in three domestic cats



B. FISH mapping of domestic cat multicopy cDNA probes on snow leopard (*Panthera uncia*) Y chromosomes

Figure 2. Summary of copy number and expression information for domestic cat and dog Y chromosome genes.

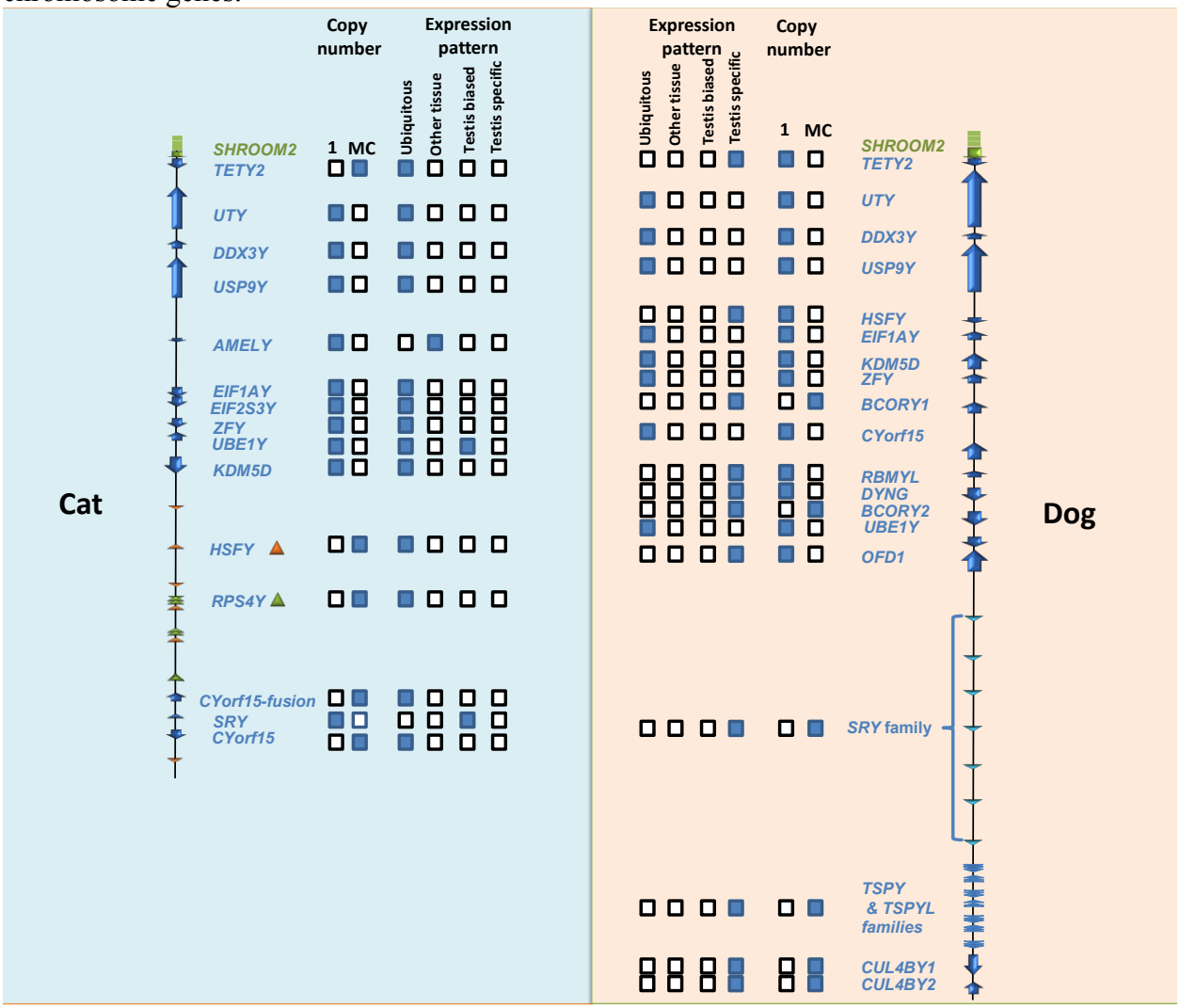
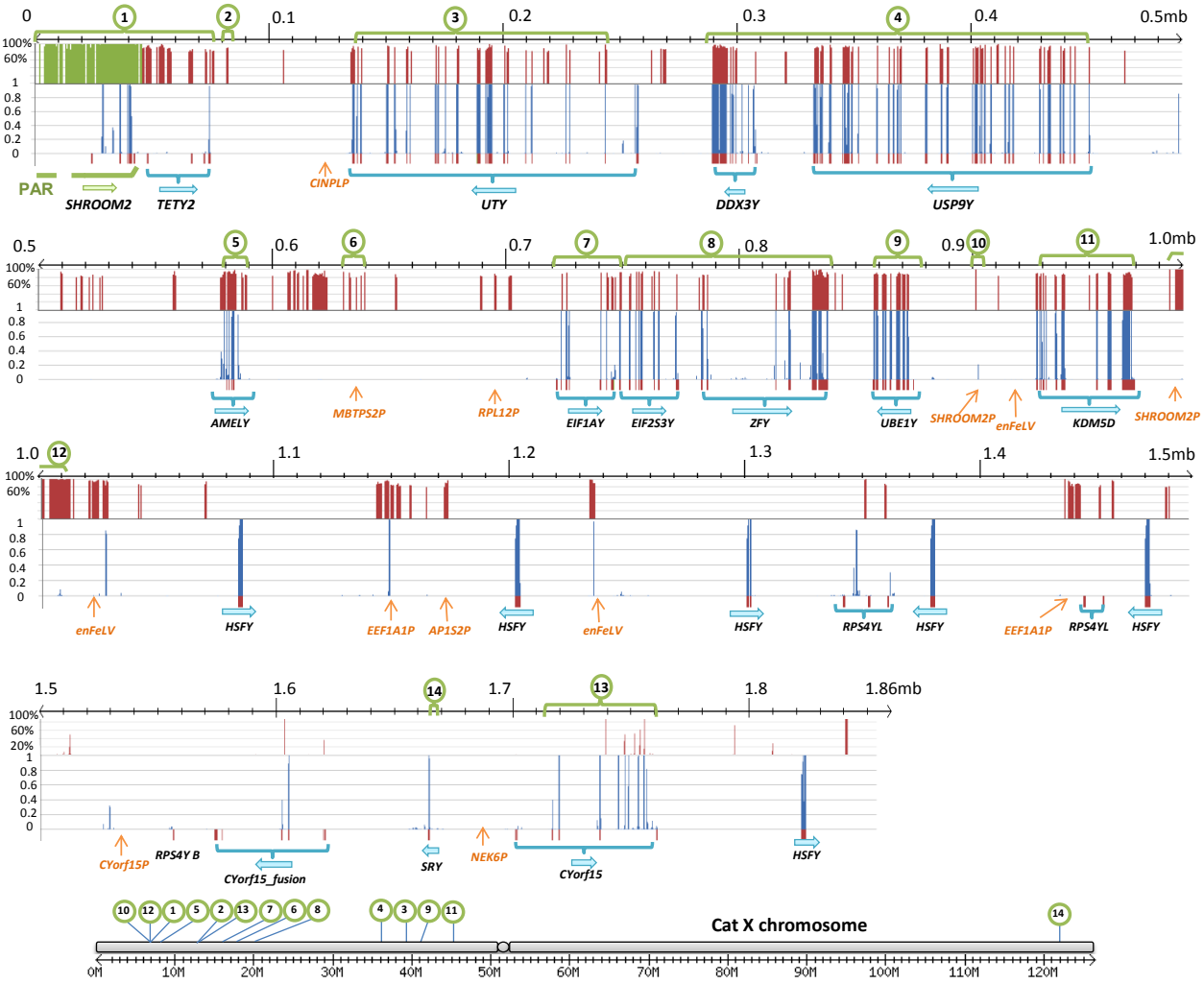


Figure 3: Comparison of sequence conservation among 5 Y mammalian chromosomes (human, chimpanzee, rhesus, dog, and cat) using the sequenced cat **(A)** and dog (**B**, next page) Y chromosome sequence as a reference. Multispecies sequence conservation scores are shown by blue columns (bottom panel), and pairwise sequence identity between X and Y chromosome sequence is shown by red columns (top panel). Green circles with numbers indicate X-Y orthologous regions, the order of which is shown on a schematic of the X chromosome at the bottom. Gene and transcript information is listed below each panel. Functional genes are shown in black, while pseudogenes are shown in orange. Red bars at the bottom of each panel indicate known gene exons.

A:



B:

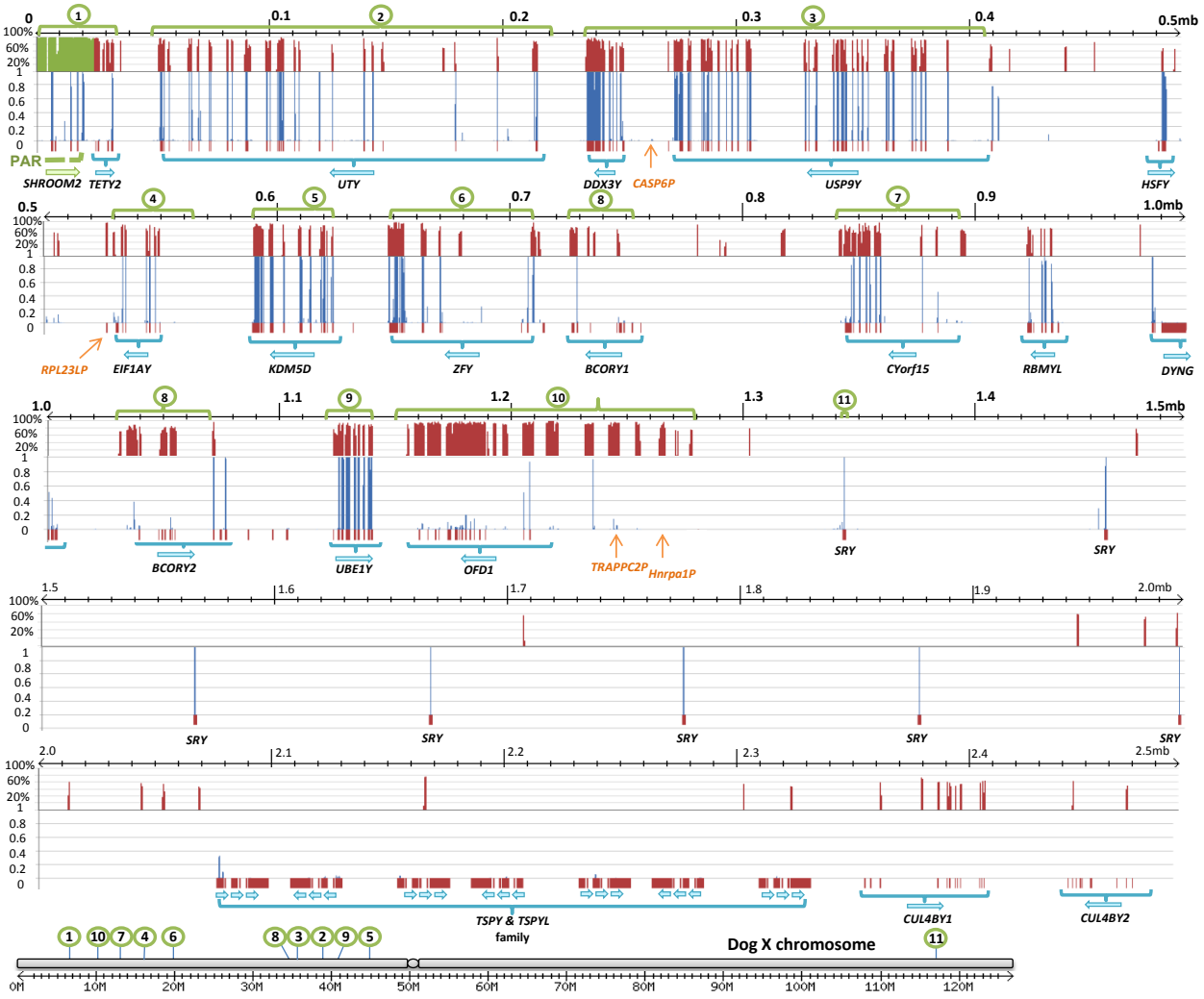

0
0.1
0.2
0.3
0.4
0.5mb
100%
60%
20%
PAR
SHROOM2
TETY2
UTY
DDX3Y
CASP6P
USP9Y
HSFY
0.5
0.6
0.7
0.8
0.9
1.0mb
RPL23LP
EIF1AY
KDM5D
ZFY
BCORY1
CYorf15
RBMYL
DYNG
1.0
1.1
1.2
1.3
1.4
1.5mb
BCORY2
UBE1Y
OFD1
TRAPPC2P
Hnrpa1P
SRY
1.5
1.6
1.7
1.8
1.9
2.0mb
2.0
2.1
2.2
2.3
2.4
2.5mb
TSPY & TSPYL
family
CUL4BY1
CUL4BY2
Dog X chromosome
0M
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M

Figure 4. Triangular self dot plots of self-self DNA sequence identities within the dog and cat MSY sequences. The horizontal panel below each figure displays the average % self-self identity, based on a sliding window analysis (window size=50 nucleotides, distance between each window step=5 nucleotides). Sequences with 100 percent identity present a dot in the triangular dot plot. Horizontal and vertical lines represent direct and inverted repeats. The MSY schematic with scale is shown below the triangular plot with colored background representing the different gene regions. Gene bodies are represented within the colored blocks as directional arrows, drawn to scale.

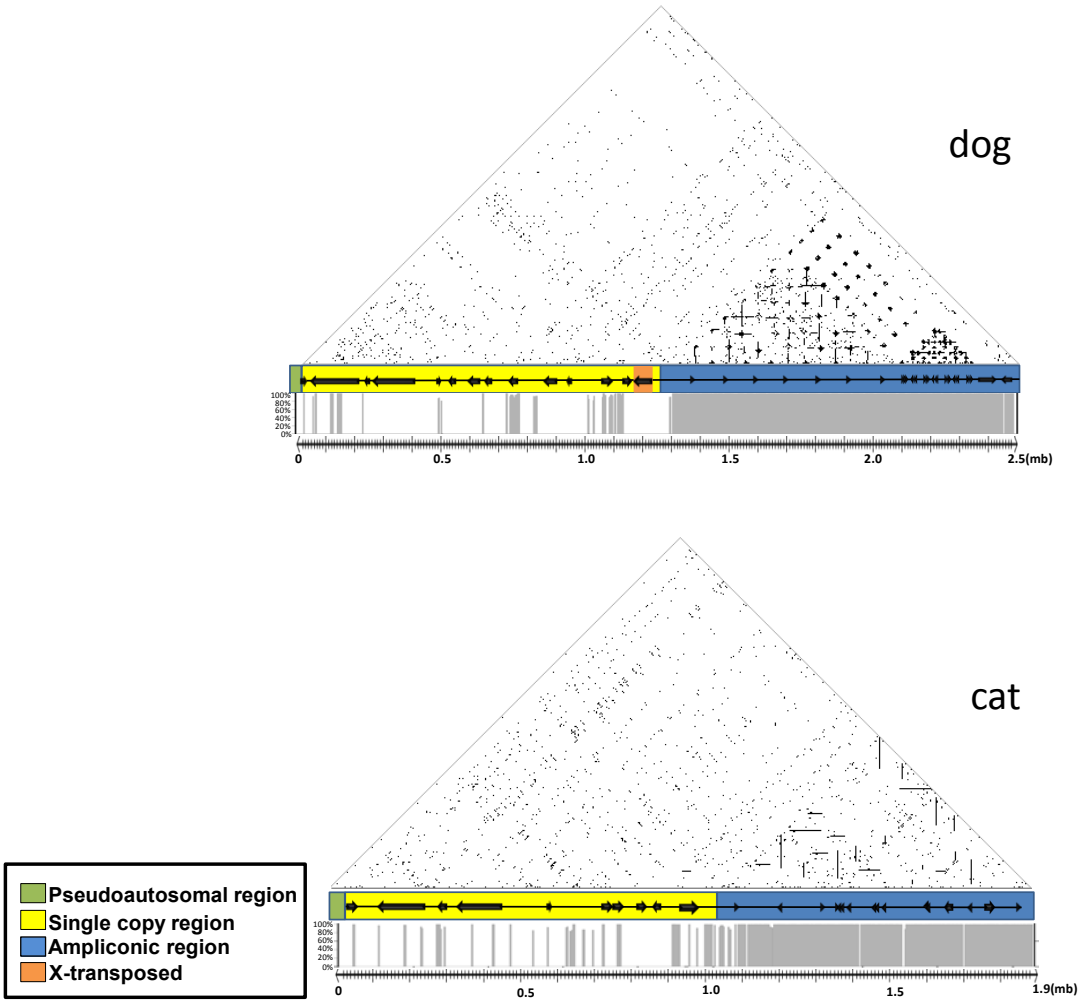
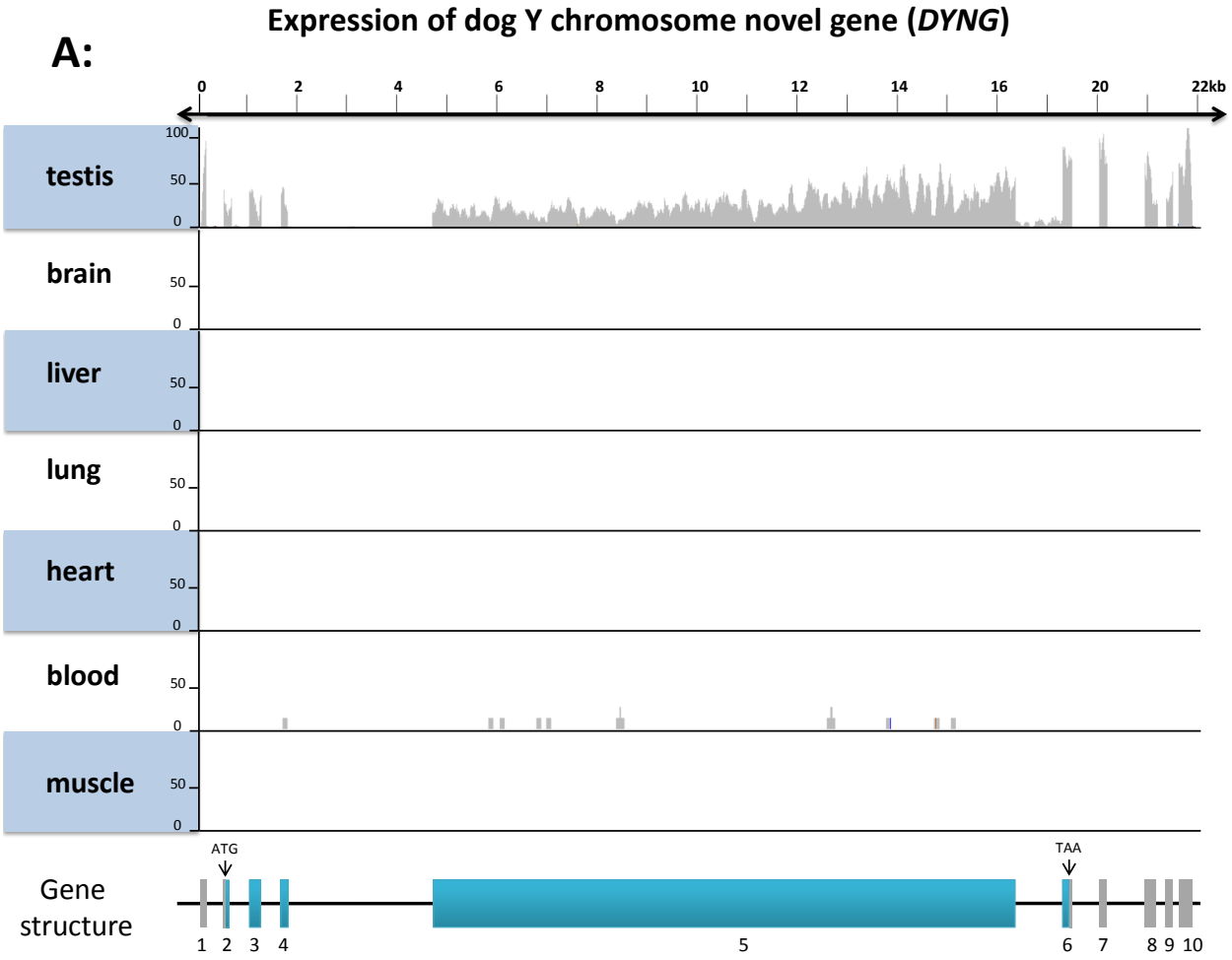


Figure 5. **A**) Gene structure and expression profile (RNASeq) of the novel coding gene, *DYNG*, found on the domestic dog Y chromosome. **B**) Self-self blast result of the transcript identifies the internal repeats within the open reading frame.



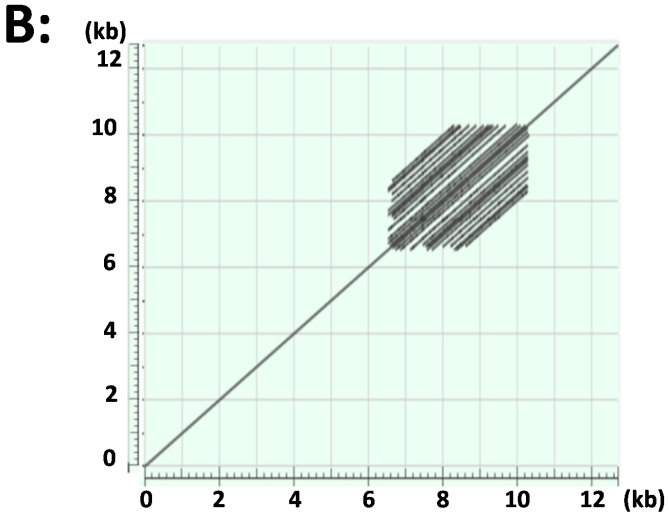


Figure 6. Evolution of the *CYorf15* gene family in select mammals. **A)** Structural comparison of human *CYorf15* gene family, with **B)** dog *CYorf15* gene, cat *CYorf15* fusion gene and the X chromosome gene Gamma taxilin, (*TXLNG*). Dashed arrows and boxes indicate regions of sequence homology between the different gene family members, based on pairwise sequence comparisons.

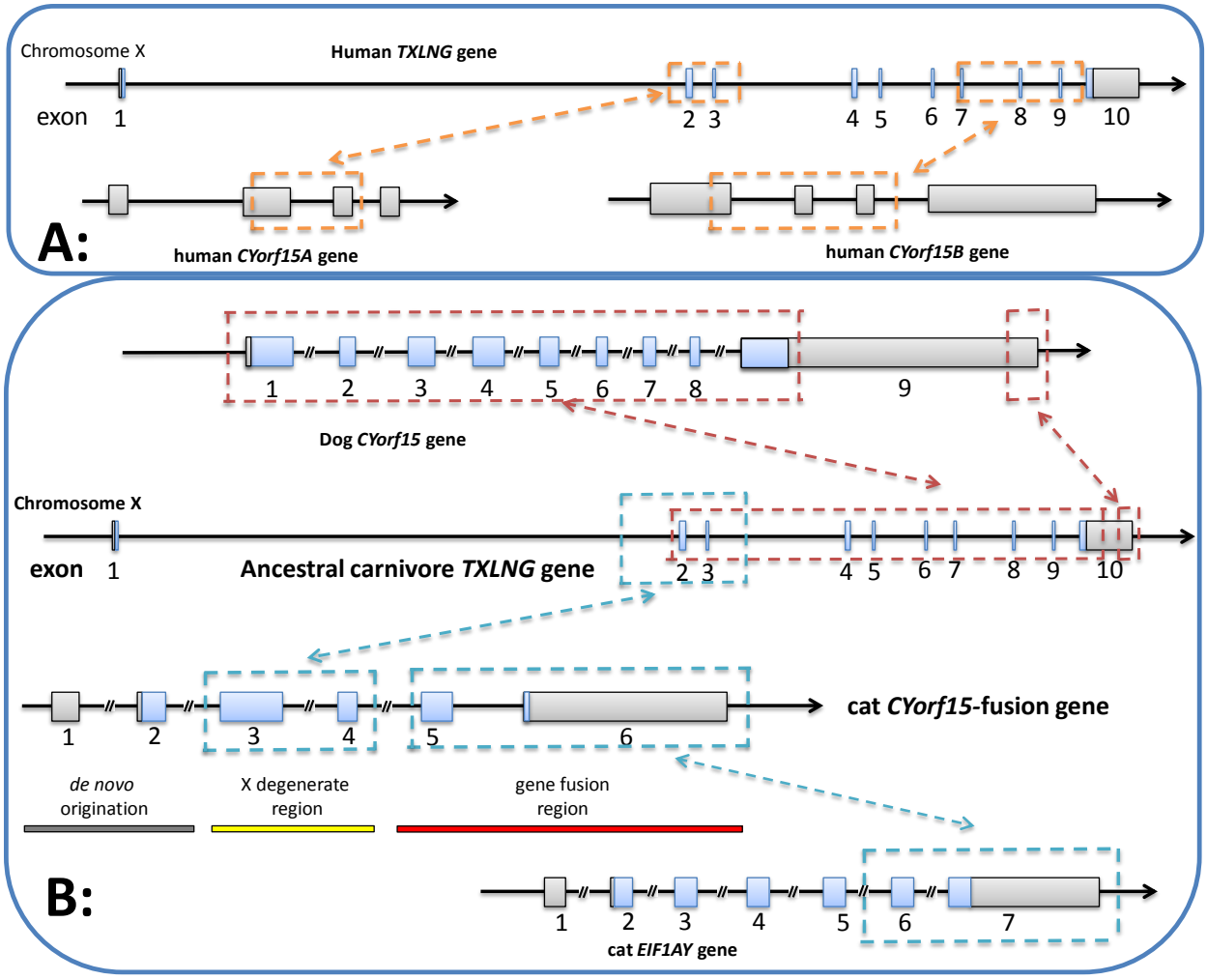


Figure 7. *OFD1* gene tree estimated using maximum likelihood method under a GTR+G+I model. Numbers near each node on the tree are bootstrap values (500 pseudoreplicates: only values higher than 80 have been shown). Taxon names are denoted by (species common name)/(ENSEMBL transcript name or NCBI accession number)/(available chromosome information). The chimpanzee Y chromosome *OFD1* pseudogene is denoted by the start and end locations within the published Y chromosome sequence.

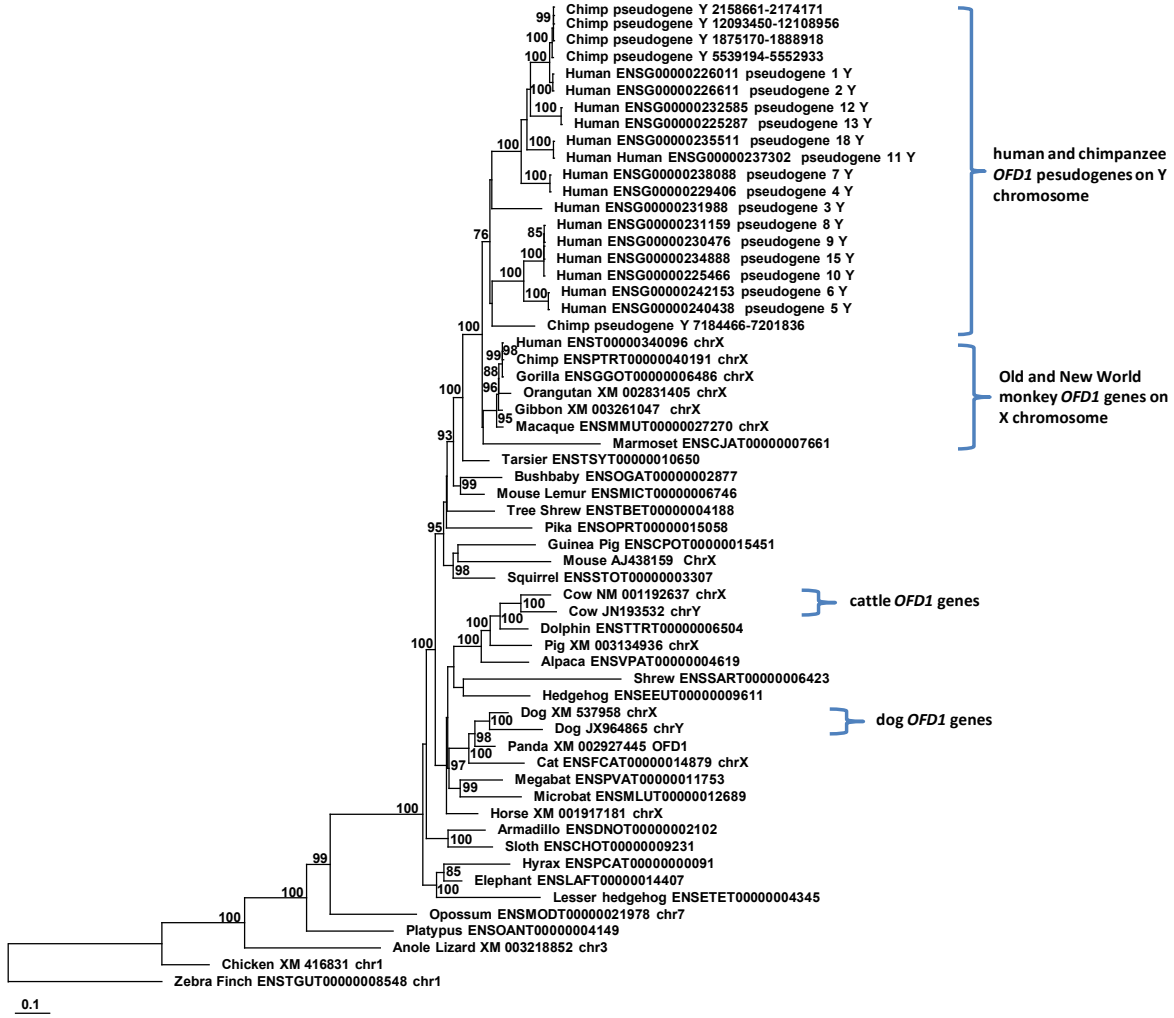
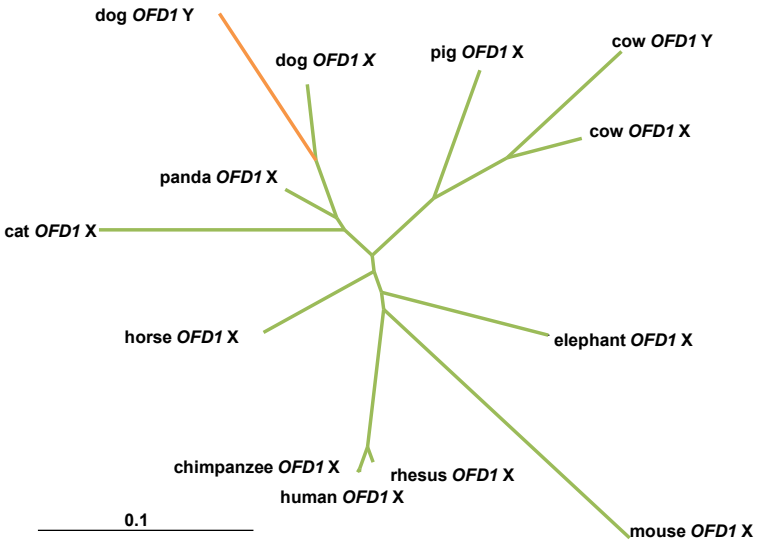


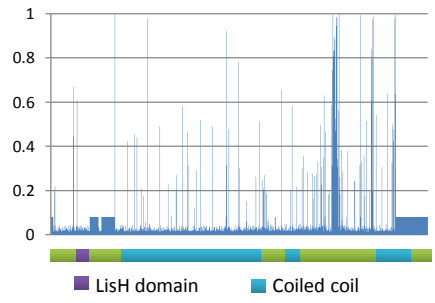
Figure 8. Results of positive selection test for the branch leading to domestic dog *OFD1Y* gene. **A)** *OFD1* gene tree, with branch lengths based on non-synonymous substitution rates. **B)** Bayesian posterior probabilities for protein sites under positive natural selection (branch-site models comparison). Colored bars underneath indicate different structural domains. **C)** Results of likelihood ratio tests for two pairs of model comparisons.

A:



B:

Bayesian posterior probabilities of positive selection



C:

Results of likelihood ratio tests for dog *OFD1Y* branch

Models	Significance
Two ratio/One ratio	P<0.001
Branch-site models	P<0.001

Figure 9. Location of the dog (top) and cat (bottom) pseudoautosomal boundaries (dashed vertical line), inferred from X-Y sequence similarity plots. Gene annotations are shown at the bottom of each figure.

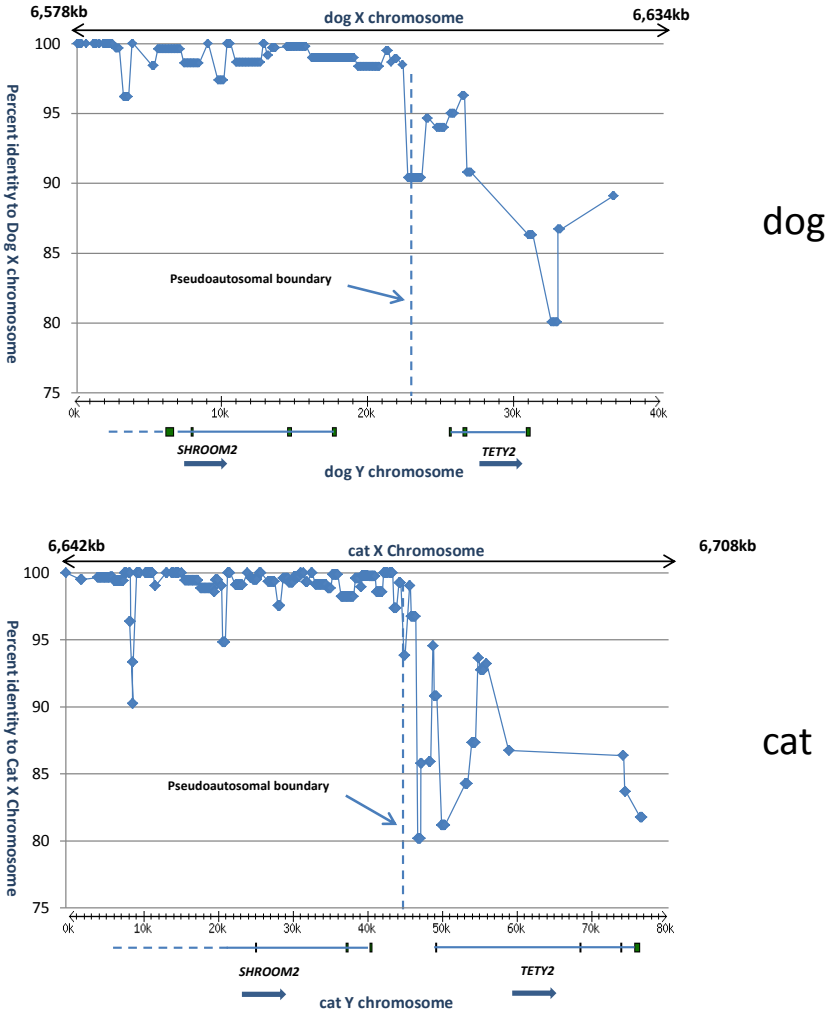


Figure 10. **A)** Maximum likelihood tree of *TETX2* orthologues sampled from different eutherian X chromosomes, estimated under a GTR+G+I model. **B)** Schematic showing the origination and divergence of the Y-linked *TETY2* gene on dog and cat Y chromosomes from the X chromosome progenitor gene, resulting from several mechanisms, including PAB movement and exon loss and *de novo* exon gain.

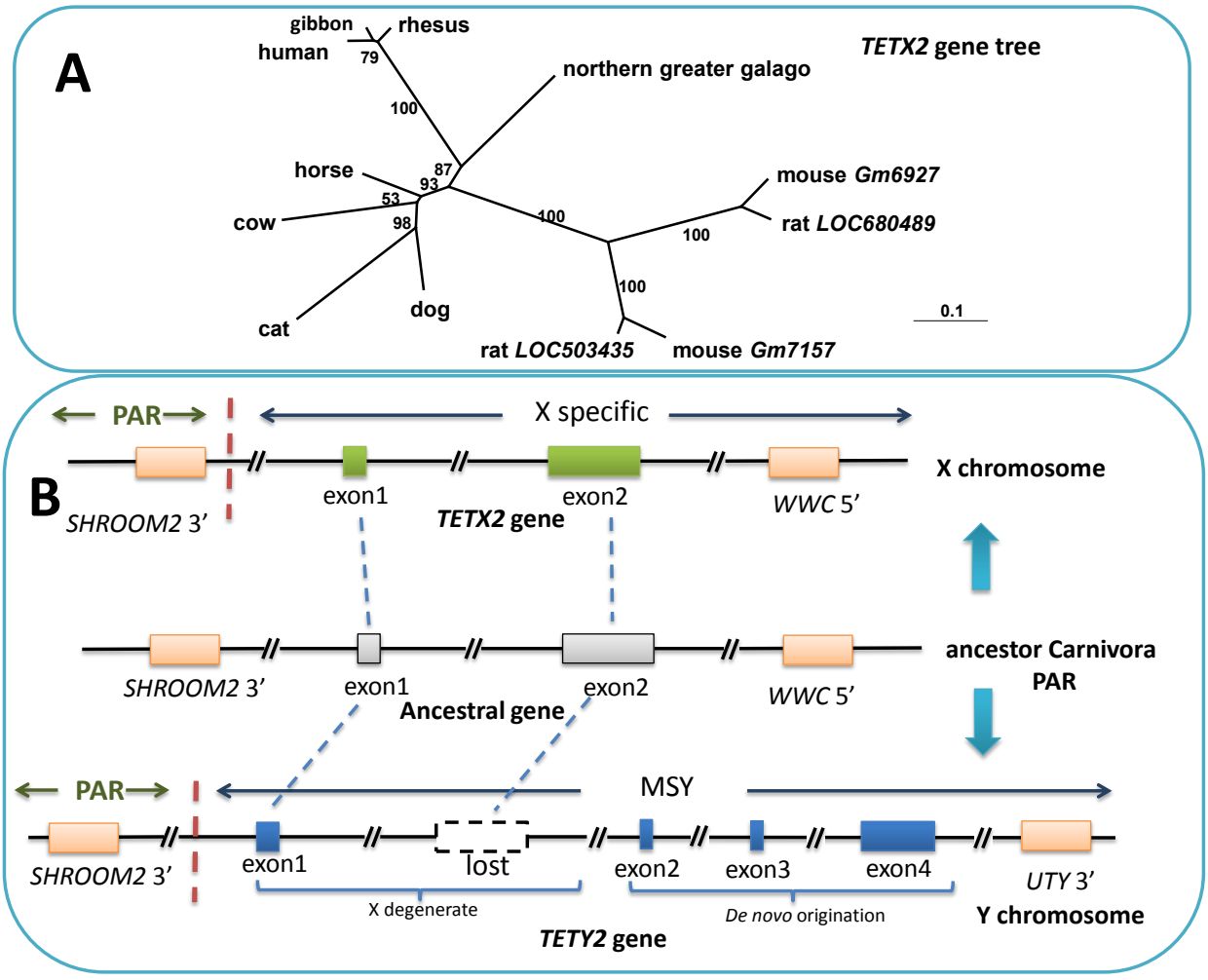


Figure 11. X-Y dot plots show regions of sequence conservation (>97%). The blue box highlights additional remnants of *SHROOM2* X-Y similarity, well inside the MSY and distant from the PAB, reflecting ancient chromosomal inversions that likely occurred on the ancestral felid branch.

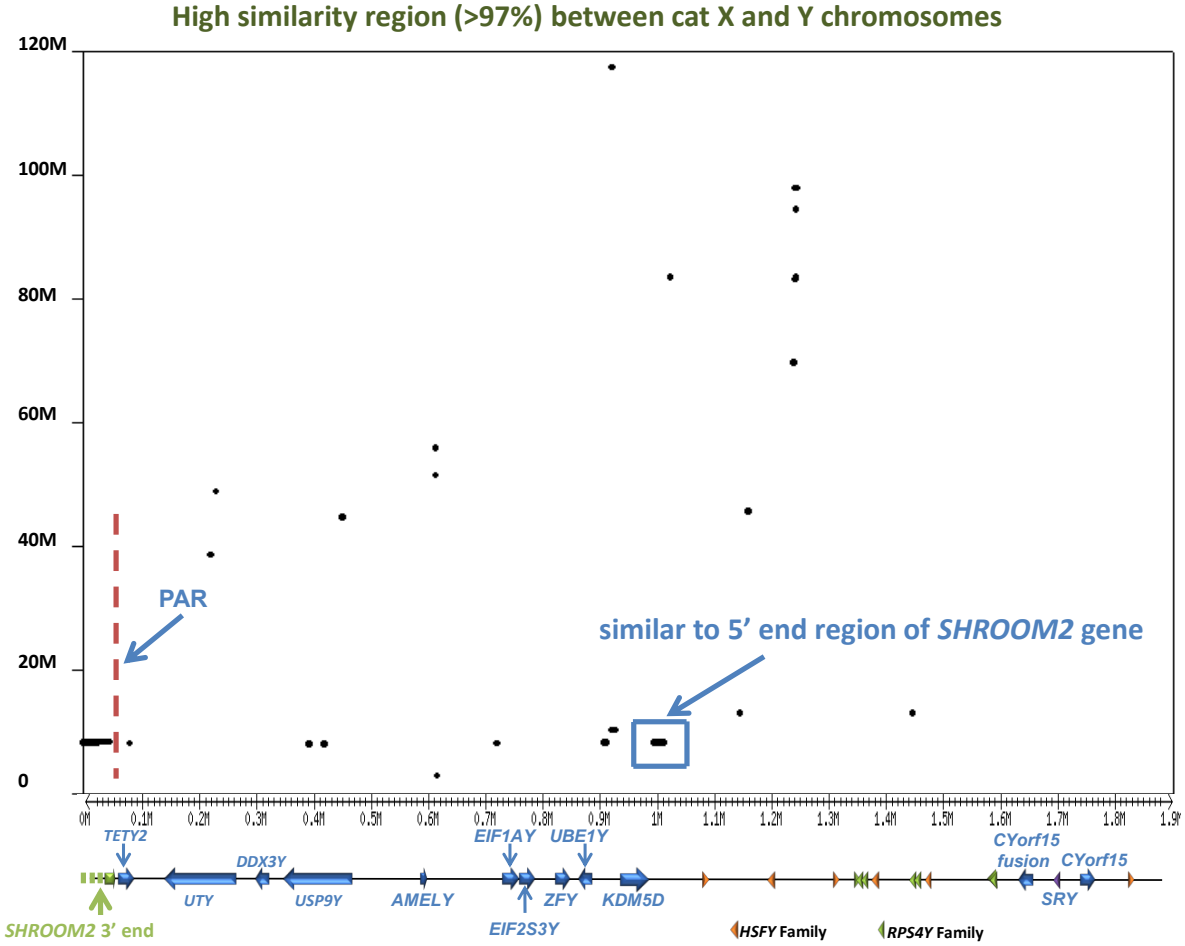


Figure 12. Pairwise sequence comparisons of four Y chromosomes (dog versus cat, rhesus, and human), using a BLAST-based method on repeatmasked sequences. Each dot represents a 50 nucleotide window with at least 50 percent identity between two species. Green boxes indicate the *UTY/DDX3Y/USP9Y* conserved gene region. The schematic on the x and y axes of each figure indicates the size and location of each MSY gene (green bars).

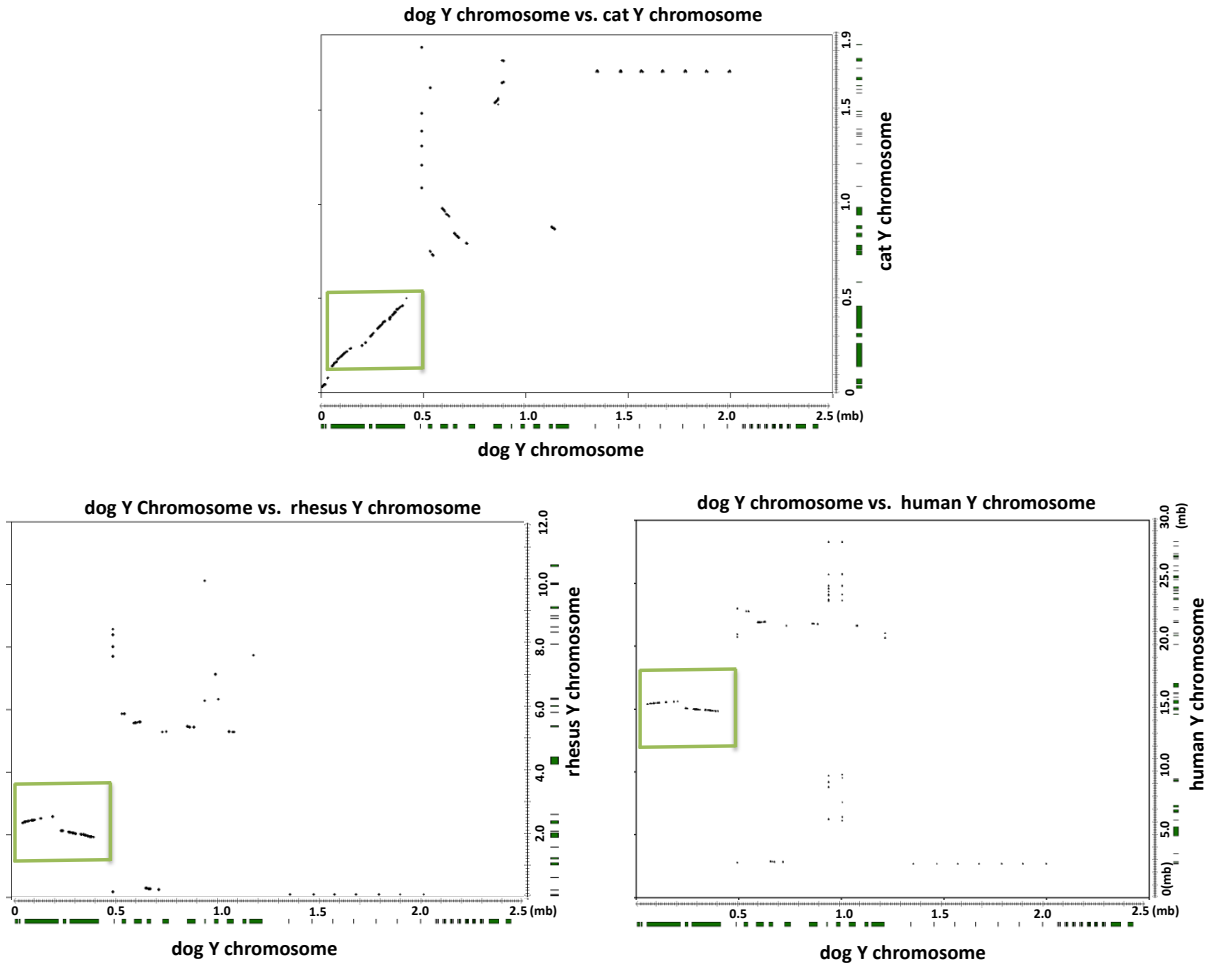


Figure 13. Plots showing rates of MSY gene loss during the evolution of mammals and continuing in the carnivore lineage, assuming an exponential model of decay with baseline (Hughes et al. 2012). Plots are divided into two groups of X-degenerate gene classes: Stratum I (Murphy et al. 2006; Hughes et al. 2012), and stratum II/III (following Pearks Wilkerson et al. 2008). Estimated ancestral gene numbers (Hughes et al. 2012), and active genes inferred for different eutherian ancestors (with divergence times drawn from Meredith et al. 2011) are shown as blue diamonds. Age of the ancestral Stratum I genes (a) is based upon the inferred ~218 Mya prototherian/therian divergence time (Meredith et al. 2011), whereas the age of the origin of Stratum II/III genes (a) is derived from Pearks Wilkerson et al. 2008. Other nodes are as follows: (b) primate/rodent/carnivore ancestor, (c) primate/rodent ancestor, (d) carnivore ancestor, (e) catarrhine primate ancestor.

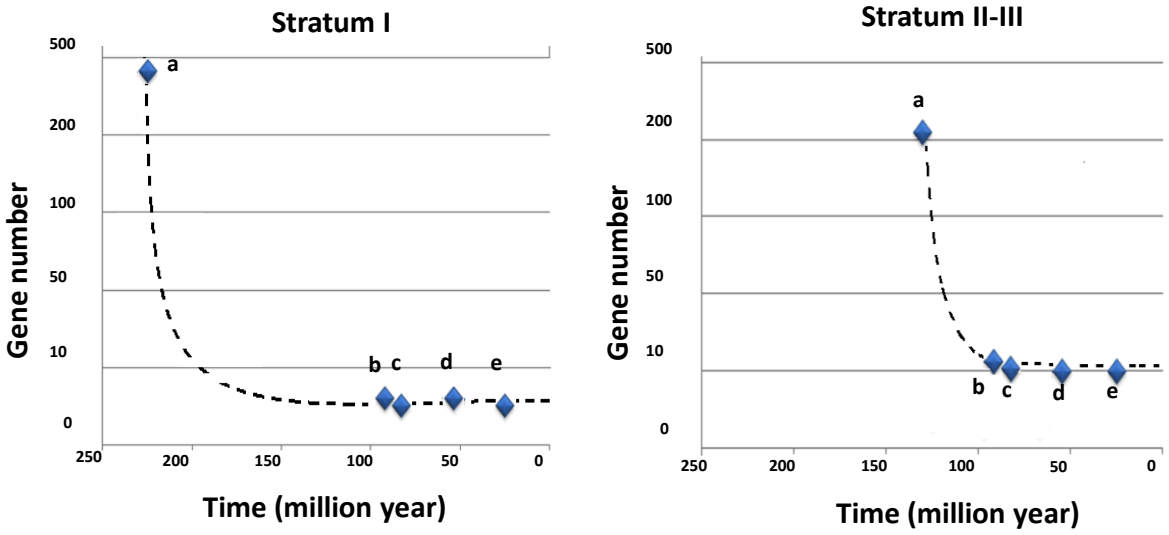
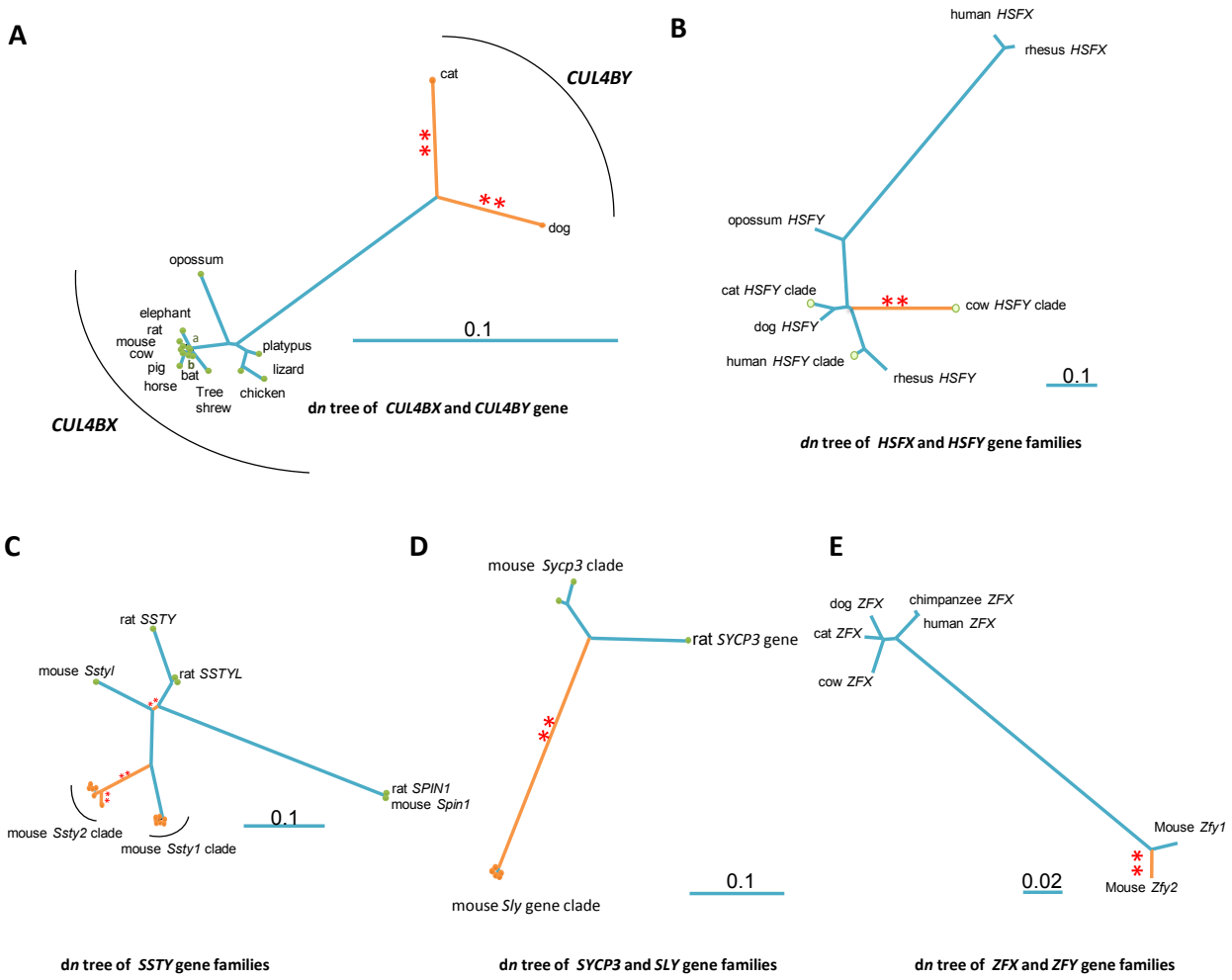


Figure 14. Results of positive selection tests for several multicopy MSY gene families, performed using the branch-site model. Asterisks indicate significant (p<0.01) positive selection detected with likelihood ratio tests.



Supplementary Table 1. Structural domains of the domestic cat and dog Y chromosomes, and relative sequence coverage. Estimated sizes are based on cytogenetic and physical mapping data (Murphy et al. 2006, Pearks-Wilkerson et al. 2008).

region	PAR		Y single copy		Multicopy/ampliconic		NOR	
	estimated size	sequenced	estimated size	sequenced	estimated size	sequenced	estimated size	sequenced
cat	6.6 Mb Xp/Yp	23 Kb[a] Yp	1-2 Mb Yp	0.7 Mb Yp	~35 Mb Yp+Yq	1 Mb Yp	NA*	NA
dog	6.6 Mb Xp/Yq	45 Kb[a] Yq	1-2 Mb Yq	1.1 Mb Yq	1-2 Mb Yq	1.5 Mb Yq	8-10 Mb Yp	0 bp Yp

[a]length of corresponding Y-PAR.

[b]NA=not applicable

Supplementary Table 2. Comparison of synonymous substitution rates between canine X-Y paralogs.

Orthologous genes		Synonymous substitution rate	average sequence divergence within intron region
X copy	Y copy		
OFD1X	*OFD1Y*	0.155	0.149
TRAPPC2	*TRAPPC2P*	NA	0.146
DDX3X	*DDX3Y*	0.527	not alignable
UTX	*UTY*	0.334	not alignable
KDM5C	*KDM5D*	0.751	not alignable
UBA1	*UBE1Y*	0.560	not alignable
USP9X	*USP9Y*	0.561	not alignable
ZFX	*ZFY*	0.266	not alignable
TXLNG	*CYorf15*	0.498	not alignable
BCOR	*BCORY1/2*	0.774/0.687	not alignable
EIF1AX	*EIF1AY*	0.532	not alignable

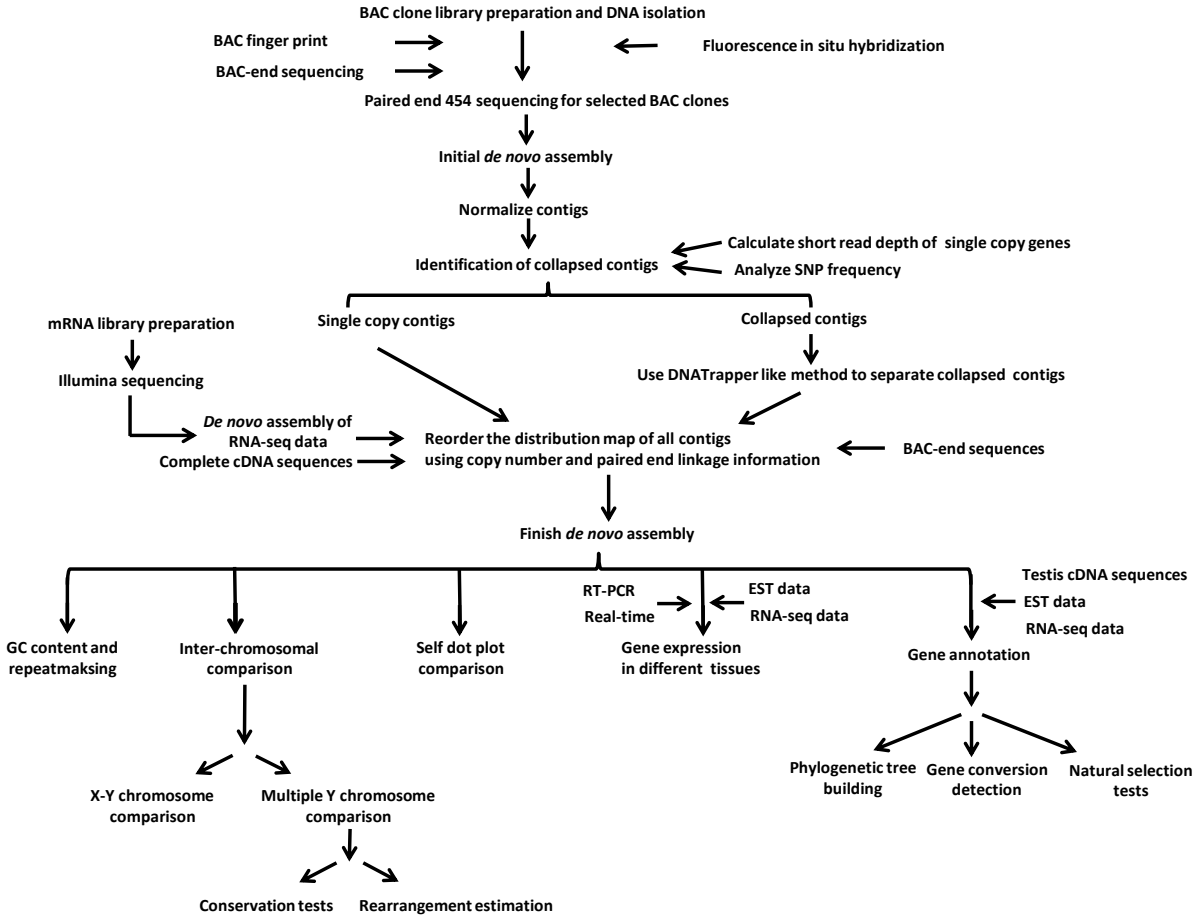
Supplementary Table 3. Sequences involved in selection tests.

Gene	species name	Accession number	Gene	species name	Accession number
AMELY	pig	AB091792	*UTY*	chimpanzee	ENSPTRT00000042001
AMELY	horse	AB032194	*UTY*	mouse	NM_009484
AMELY	cow	M63500	*ZFY*	mouse	AK076618
AMELY	human	ENST00000215479	*ZFY*	mouse	M24401
AMELY	chimpanzee	ENSPTRT00000041965	*ZFY*	cow	AF032867
DDX3Y	human	ENST00000360160	*ZFY*	human	ENST00000383052
DDX3Y	chimpanzee	ENSPTRT00000041996	*ZFY*	chimpanzee	ENSPTRT00000041960
DDX3Y	mouse	ENSMUST00000091190	*HSFY*	human	AF332227
EIF1AY	human	ENST00000361365	*HSFY*	human	AK058182
EIF1AY	chimpanzee	ENSPTRT00000042030	*HSFY*	rhesus	FJ527015
EIF1AY	rhesus	FJ527014	*HSFY*	opossum	GQ253469
EIF1AY	cow	FJ195366	*HSFY*	cow	NM_001077006
EIF1AY	pig	AY609402	*HSFY*	cow	XM_001250229
EIF2S3Y	rat	FJ775732	*HSFY*	cow	XM_003585244
EIF2S3Y	mouse	AJ006584	*TSPY*	human	ENST00000451548
KDM5D	human	ENST00000317961	*TSPY*	human	ENST00000457222
KDM5D	chimpanzee	ENSPTRT00000042025	*TSPY*	human	ENST00000428845
KDM5D	mouse	NM_011419	*TSPY*	human	ENST00000426950
SRY	human	AM884750	*TSPY*	human	ENST00000426950
SRY	chimpanzee	DQ977342	*TSPY*	human	ENST00000457222
SRY	rhesus	FJ527022	*TSPY*	human	ENST00000287721
SRY	rabbit	AY785433	*TSPY*	chimpanzee	ENSPTRT00000055849
SRY	mouse	U03645	*TSPY*	chimpanzee	ENSPTRT00000041983
SRY	rat	AF274872	*TSPY*	chimpanzee	ENSPTRT00000055827
SRY	cow	AB039748	*TSPY*	chimpanzee	ENSPTRT00000055807
SRY	pig	U49860	*TSPY*	pig	ENSSSCT00000022749
SRY	horse	AB004572	*TSPY*	cow	ENSBTAT00000049474
SRY	panda	AB292070	*CDY*	human	ENST00000426790
UBE1Y	mouse	AF150963	*CDY*	human	ENST00000306882
UBE1Y	rat	EF690356	*CDY*	human	ENST00000544303
UBE1Y	opossum	GQ253467	*CDY*	human	ENST00000306609
USP9Y	human	ENST00000338981	*CDY*	chimpanzee	ENSPTRT00000076216
USP9Y	chimpanzee	ENSPTRT00000041991	*CDY*	chimpanzee	ENSPTRT00000073911
USP9Y	mouse	NM_148943	*CDY*	chimpanzee	ENSPTRT00000076243
USP9Y	cow	NM_001145509	*CDY*	chimpanzee	ENSPTRT00000072418
UTY	human	ENST00000331397	*CDY*	chimpanzee	ENSPTRT00000072527

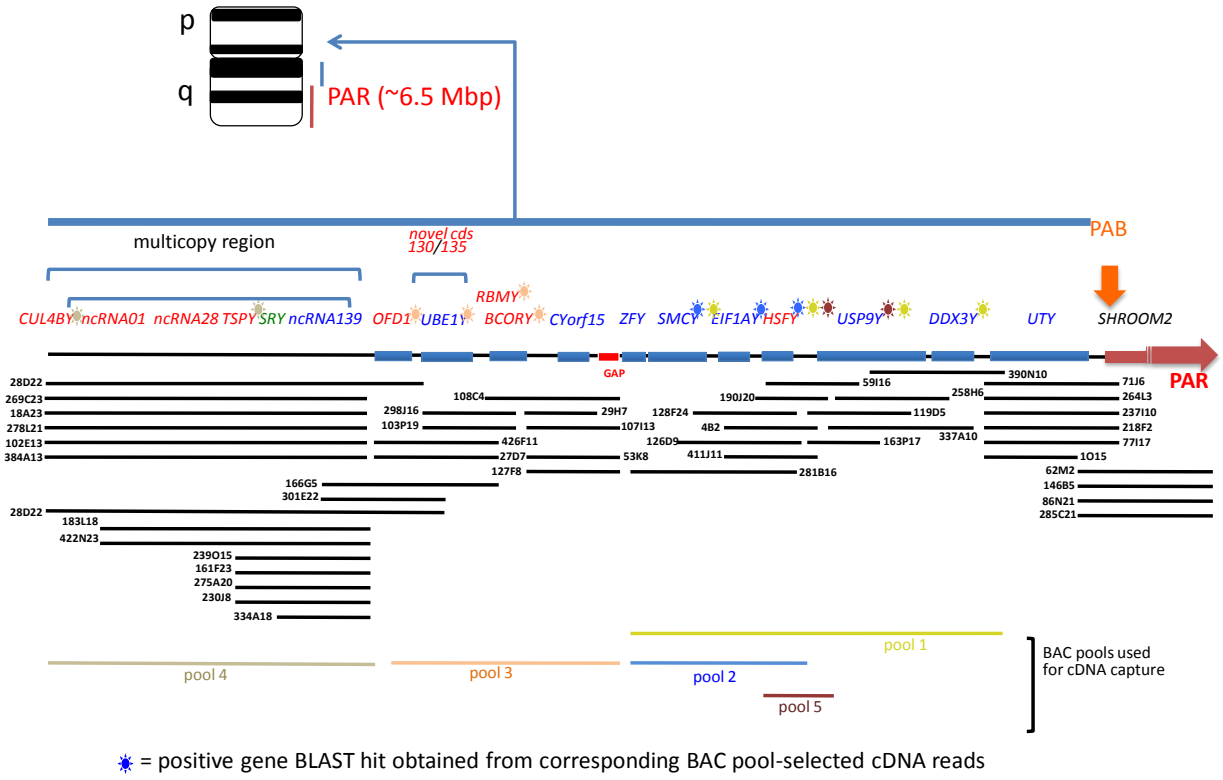
Gene	species name	Accession number	Gene	species name	Accession number
CDY	rhesus	FJ527011	*SSTY1L*	rat	XM_222167
DAZ	chimpanzee	ENSPTRT00000062513	*SSTY1L*	rat	XM_001078732
DAZ	chimpanzee	ENSPTRT00000062535	*SstyL*	mouse	XM_144574
DAZ	chimpanzee	ENSPTRT00000062545	*Ssty1*	mouse	X05260
DAZ	chimpanzee	ENSPTRT00000055821	*Ssty1*	mouse	Mm#S38822574 (EST)
DAZ	human	ENST00000382440	*Ssty1*	mouse	Mm#S38822973
DAZ	human	ENST00000405239	*Ssty1*	mouse	Mm#S38822990
DAZ	human	ENST00000382365	*Ssty1*	mouse	Mm#S8039565
DAZ	human	ENST00000440066	*Ssty1*	mouse	Mm#S7116079
DAZ	rhesus	FJ648738	*Ssty1*	mouse	Mm#S38821906
Cyorf15A	human	ENST00000407724	*Ssty1*	mouse	Mm#S32784017
Cyorf15A	chimpanzee	NM_001009080	*Ssty1*	mouse	Mm#S26680069
Cyorf15A	rhesus	FJ527012	*Ssty1*	mouse	Mm#S26702624
Cyorf15A	gorilla	FJ532256	*Ssty1*	mouse	Mm#S40811957
Cyorf15B	human	ENST00000382832	*Ssty2*	mouse	Mm#S7227131
Cyorf15B	rhesus	FJ648737	*Ssty2*	mouse	Mm#S52502287
Cyorf15B	gorilla	FJ532257	*Ssty2*	mouse	Mm#S7227130
CUL4BX	microbat	ENSMLUT00000013243	*Ssty2*	mouse	Mm#S38822045
CUL4BX	dog	ENSCAFT00000029435	*Ssty2*	mouse	Mm#S19760504
CUL4BX	platypus	ENSOANT00000004088	*Ssty2*	mouse	Mm#S14960481
CUL4BX	pig	ENSSSCT00000013782	*Sly*	mouse	BC049626
CUL4BX	lizard	ENSACAT00000010801	*Sly*	mouse	Mm#S38821277 (EST)
CUL4BX	chicken	ENSGALT00000013942	*Sly*	mouse	Mm#S38821362
CUL4BX	marmoset	ENSCJAT00000005037	*Sly*	mouse	Mm#S38822154
CUL4BX	human	ENST00000404115	*Sly*	mouse	Mm#S38822342
CUL4BX	tree shrew	ENSTBET00000007462	*Sly*	mouse	Mm#S60235098
CUL4BX	mouse	NM_001110142	*Sly*	mouse	Mm#S38822478
CUL4BX	cat	ENSFCAT00000010442	*Sly*	mouse	Mm#S21508819
CUL4BX	elephant	ENSLAFT00000034221	*Sly*	mouse	Mm#S14960473
CUL4BX	cow	ENSBTAT00000024713	*Sycp3*	mouse	ENSMUST00000033458
CUL4BX	horse	ENSECAT00000025736	*Sycp3*	mouse	ENSMUST00000067763
CUL4BX	rat	ENSRNOT00000065023	*Sycp3*	rat	ENSRNOT00000056700
CUL4BX	rabbit	ENSOCUT00000015646	*FLJ36031Y*	cat (clone a260)	DQ329513
CUL4BX	opossum	XM_001372861	*FLJ36031Y*	cat (b321)	DQ329514
Spin1	mouse	AK168544	*FLJ36031Y*	cat (c254)	DQ329515
SPIN1	rat	BC097359	*FLJ36031Y*	cat (e268)	DQ329517
SSTY	rat	AABR06008269	*FLJ36031Y*	cat (d340)	DQ329516

Supplementary Methods

Pipeline of data generation and analyses.



Domestic dog Y chromosome BAC clone map.

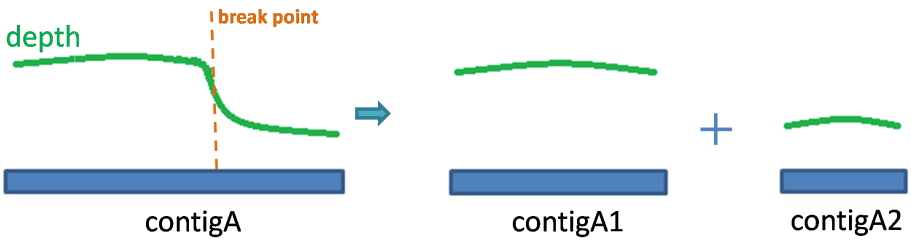


✸ = positive gene BLAST hit obtained from corresponding BAC pool-selected cDNA reads

***De novo* assembly of cat and dog Y chromosome sequences**

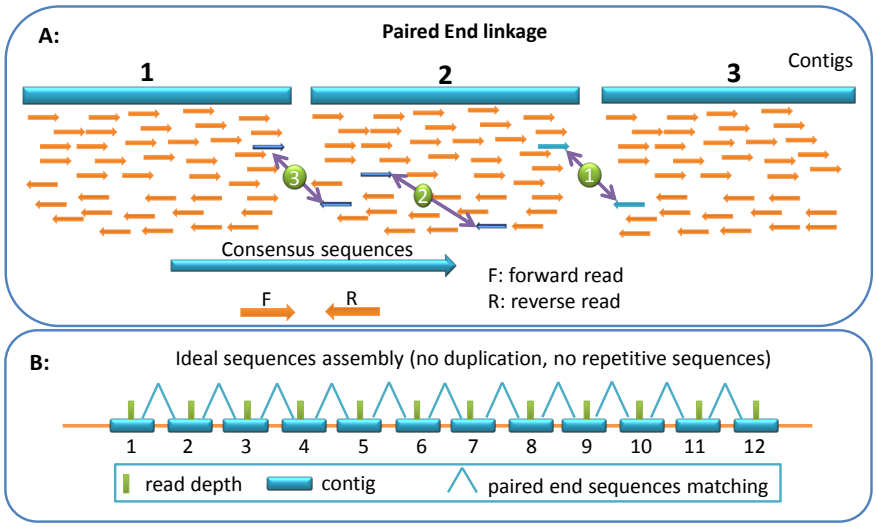
The initial sequence assembly was subjected to custom manual annotation, as follows:

a). We normalized the initial *de novo* contigs containing significantly uneven depth of coverage by identifying regions with significant shifts in coverage, and cutting the contigs into new sub-contigs with even read depth coverage (see below).

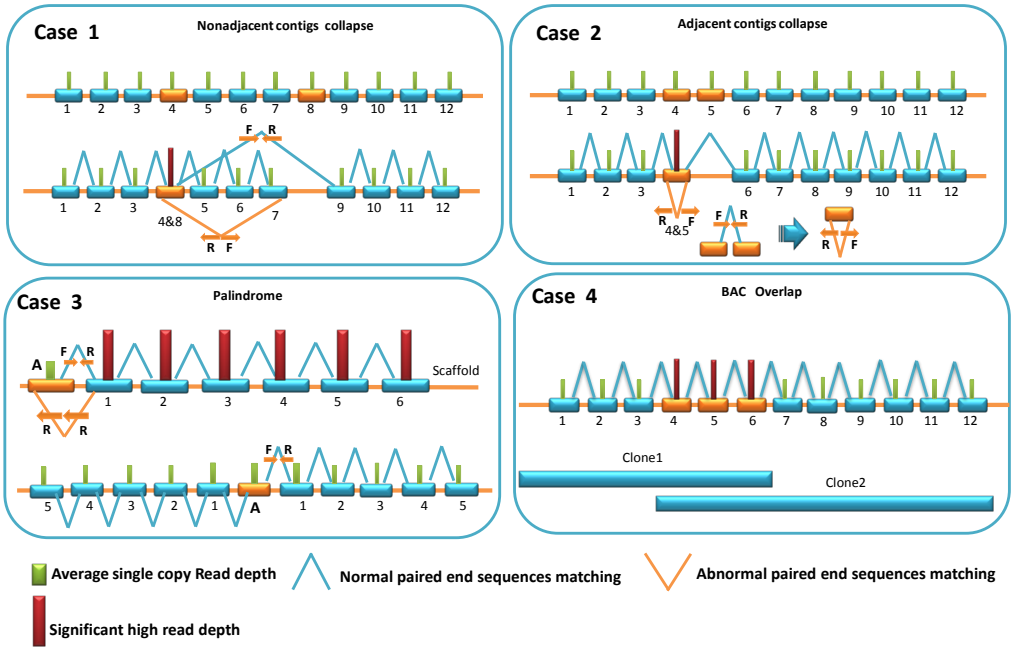


b). Paired end sequences are favorable for sequence assembly by utilizing upstream and

downstream read recognition (see Fig. A below). An ideal *de novo* sequence assembly based on NGS data will exhibit relatively even read-depth within each contig and a contiguous distribution of matching paired-end reads (Fig. B below).



Using this information, we identified collapsed contigs (caused by gene/sequence duplication) in the initial *de novo* assembly. Several examples of different assembly artifacts are shown below. For example, in a case where two regions with highly similar, paralogous sequences collapse into one contig with greater than average read coverage, this results in two distinct and abnormal patterns of paired end read distribution (Case 1 and Case 2 below).

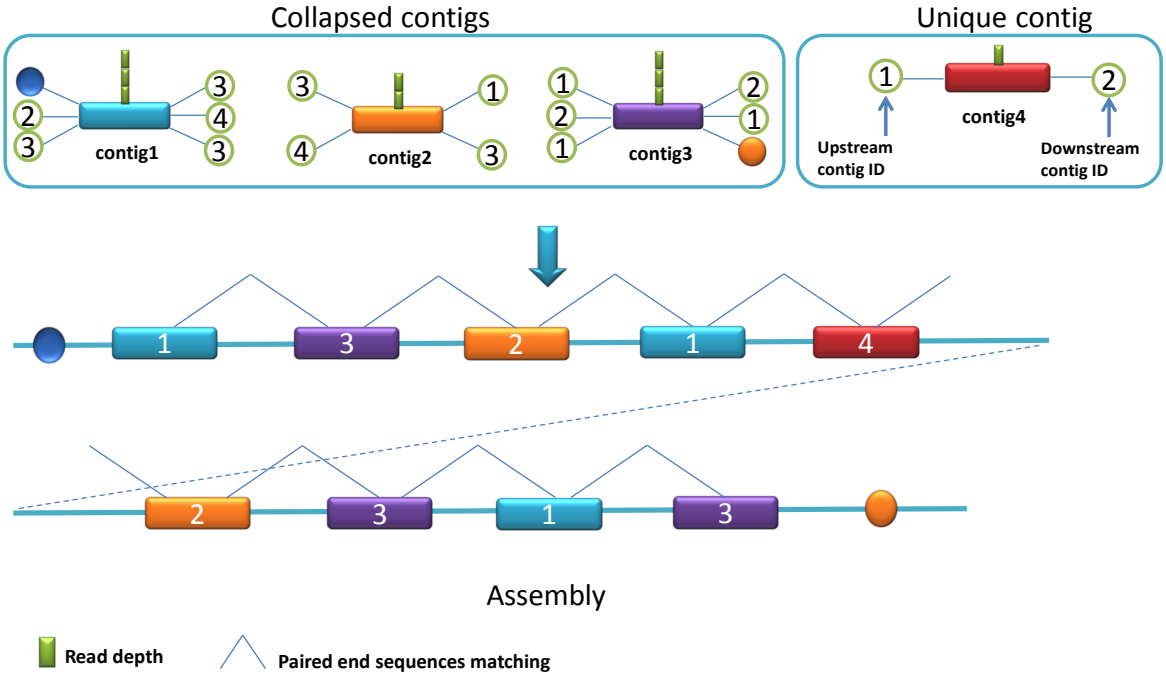


Case 3 indicates a palindrome structure of highly similar sequences with read depth and paired end reads matching. Regions of BAC clone overlap also can cause significantly higher read

depth (Case 4), but paired end sequences patterns that are distinct from collapsed contigs caused by gene duplication.

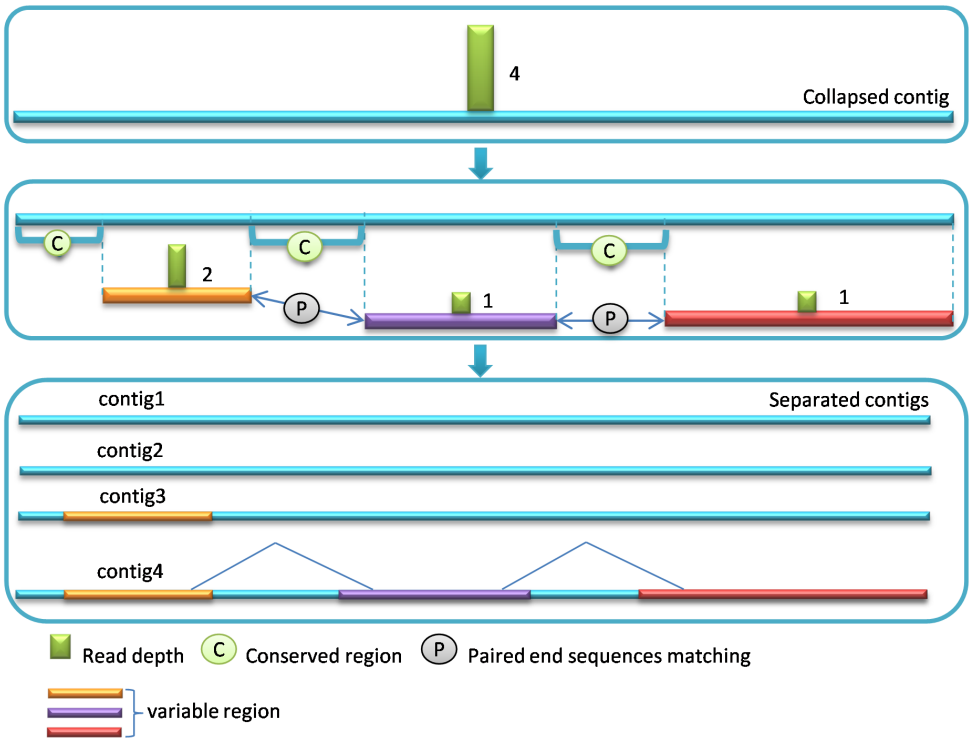
Using this information we reanalyzed the initial *de novo* sequence assembly. First we evaluated the copy number of each contig that was identified as potential collapsed contigs containing sequences from highly similar, duplicated Y chromosome regions. In this step, two criteria were used. We estimated the average expected depth of sequence coverage in contigs which contain known single copy coding genes (e.g. *UTY*/*DDX3Y*/*USP9Y* gene coding region). All normalized contigs were then compared to this value to provide an initial estimate of the potential collapsed copy number. Second, because the sequenced MSY region is haploid, only sequencing error or collapsed sequences could account for polymorphisms found in the initial *de novo* assembled contigs. We analyzed the SNP frequency of polymorphic sites found in contigs exhibiting a significant excess of coverage depth to adjust our initial copy number estimate.

c). Using paired-end sequence information and the estimated copy number of collapsed contigs, we calculated the maximum parsimony distribution map of contigs to rebuild scaffolds (below).



d). We used a method similar to that applied in the software DNPtrapper(Arner et al. 2006) to separate collapsed contigs into unique contigs. Collapsed contigs include highly conserved regions and polymorphic region that result from collapse of nearly identical sequences distinguished by rare sequence variants. Variable region sequences are separated from the original contig and then sequence coverage is recalculated. Using paired end information the new

separated contigs are rebuilt.



The *de novo* assembly is completed by repositioning the formerly collapsed contigs into unique positions in the existing assembly. Copy number information inferred from read depth data was validated with quantitative PCR.

References

Arner E, Tammi MT, Tran AN, Kindlund E, Andersson B. 2006. DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. *BMC Bioinformatics* **7**:155.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**(7387): 82-86.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TLL, Stadler T et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* **334**(6055): 521-524.

Murphy WJ, Wilkerson AJP, Raudsepp T, Agarwala R, Schaffer AA, Stanyon R, Chowdhary BP. 2006. Novel gene acquisition on carnivore Y chromosomes. *Plos Genet* **2**(3): 353-363.

Pearks Wilkerson AJ, Raudsepp T, Graves T, Albracht D, Warren W, Chowdhary BP, Skow LC, Murphy WJ. 2008. Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics* **92**(5): 329-338.