Emergence of a new gene from an intergenic region

Tobias J.A.J. Heinen*, Fabian Staubach*, Daniela Häming and Diethard Tautz[#]
*these authors have contributed equally

Max-Planck Institute for Evolutionary Biology
August-Thienemannstrasse 2
24306 Ploen - Germany

[#]corresponding author: tautz@evolbio.mpg.de
running title: evolution of a new gene

key words: *Mus musculus*, testis expression, gene evolution

It is generally assumed that new genes would arise by gene duplication mechanisms, since the signals for regulation and transcript processing would be unlikely to evolve in parallel with a new gene function [1, 2]. We have identified here a transcript in the house mouse (*Mus musculus*) that has arisen within the past 2.5 - 3.5 million years in a large intergenic region. The region is present in many mammals, including humans, allowing us to exclude the involvement of transposable elements or other genome-rearrangements, which are typically found for other cases of newly evolved genes [3 - 8]. The gene has three exons, shows alternative splicing and is specifically expressed in postmeiotic cells of the testis. The transcript is restricted to species within the genus *Mus* and its emergence correlates with indel mutations in the 5´- regulatory region of the transcript. A recent selective sweep is associated with the transcript region in *M. m. musculus* populations. A knock-out in the laboratory strain BL6 results in reduced sperm motility and reduced testis weight. Our results show that cryptic signals for transcript regulation and processing exist in intergenic regions and can become the basis for the evolution of a new functional gene.

The role of gene duplications in generating new gene repertoires is well understood [1]. However, every genome harbours also a certain fraction of orphan genes which can not be associated with another known gene. The evolutionary origins of such genes are still rather unclear [2]. Genome comparisons have shown that de novo emergence of genes is possible, although mostly in the context of recruitment of fragments of transposable elements or other genome rearrangements [3-8]. Also, the taxa analysed in these studies have comparatively large evolutionary distances, making it difficult to infer the mechanisms of gene emergence.

We have identified a mouse-transcript (1700125F08Rik) within a 200kb region that is free of annotated transcripts or ESTs in rat and humans (suppl. Fig. 1), but spans a region that can be aligned bewteen

Mus and these species (Fig. 1A). To trace the phylogenetic origin of the transcript, we prepared a Northern blot with RNA from species with increasing evolutionary distance from the house mouse. A signal was obtained from species within the genus *Mus*, but not from the more basal *Mus* species, *M. famulus* and *M. caroli,* nor from *Apodemus* (wood mouse) and *Rattus* (rat) as outgroups (Fig. 2). This lack of hybridization is not due to sequence divergence at the locus, since it is possible to use a genomic fragment from the rat as a probe to detect the expression in *Mus* (suppl. Fig. 2). Thus, the gene originated after the split between *M. famulus* and the ingroup species approximately 2.5 -3.5 million years ago [9, 10]. Among the ingroup species expressing the transcript, *M. spicilegus* showed no signal and *M. m. musculus* and *M. spretus* yielded only a weak signal. To assess whether the weak expression is characteristic for the species or sub-species, or whether the expression level is polymorphic within the population, we used population samples from *M. m. domesticus* and *M. m. musculus* and a quantitative PCR assay to assess transcript levels. The results show that there is a high expression variance among individuals from both populations, making the difference between them non-significant in the population context (suppl. Fig. 3). Furthermore, by sequencing cDNA from both sub-species, we found that the small exon is subject to alternative splicing, with about half of the transcripts not containing it (suppl. Fig. 4). Based on these characteristics, we have named the gene *Polymorphic derived intron-containing* (*Poldi*). *In situ* hybridization to testis sections shows that the *Poldi* transcript is specifically expressed in postmeiotic round spermatids of the seminiferous tubules (Fig. 2). There is no expression in any other tissue based on the EST database ([11] - UniGene Built #178 / dbEST (06 Feb 2009)).

Postmeiotic gene expression and regulation in testis follows different rules than somatic gene regulation. There is generally a broad expression of many genes during this stage, although most of them are not translated [12]. Regulatory regions of genes that are specifically activated during this stage are usually short (less than 400bp - [13]). To assess which mutations in the lineage *Mus* could have led to the emergence of the transcript, we sequenced an approximately 1kb upstream region from several species of the genus *Mus*. Within the group of species expressing *Poldi*, we find fixed derived changes at -390 (an 11bp indel), at -160 (a 2bp substitution) and at + 11 (a 3bp indel) (Fig. 3). Sequencing of the exon junctions shows that all splice donor and acceptor sites, as well as the poly-adenylation signal are identical between all mouse species sequenced, including those that do not express the transcript (Fig. 3). Thus, these sites existed as cryptic sites before the emergence of the transcript. In *M. spicilegus*, which belongs to the ingroup species but lacks a transcript (Fig. 2) we find a derived mutation that affects the first splice donor site (Fig. 3).

Analysis of polymorphisms in the cDNA sequence samples from the *M. m. musculus* population showed a marked reduction compared to the *M. m. domesticus* samples. To assess whether this was caused by a selective sweep in the *M. m. musculus* lineage, we have obtained polymorphism data from 14 fragments within a 500kb region surrounding the transcript and find that the reduction is indeed specific to the *Poldi* transcript region in *M. m. musculus* (Fig. 4). Small ZnS values (Kaz: 0.1, Cze:

0.04) [14] are compatible with a star phylogeny of the *M. m. musculus* alleles, indicative of a selective sweep in the recent past.

To study the function of *Poldi*, we designed a conditional knock out of the whole gene region (see suppl. Fig. 5). We find that mice lacking the *Poldi* transcript are viable and fertile and testis morphology is not changed (suppl. Fig. 6). However, they show a significantly reduced testis weight and reduced sperm motility (Figure 5A). Testis transcriptome comparisons between mice lacking *Poldi* and the strain from which they were derived shows that the expression level of several genes are changed in the knock out mice (Figure 5B; suppl. Table 1), indicating that *Poldi* is part of a regulatory network. The GO terms of the list of genes that are affected do not provide a specific clue towards a particular function (suppl. Table 2). However, one of the significantly downregulated genes (*Hmgb2*) is a chromosomal protein known to have a specialised role in germ cell differentiation and a knockout of this gene results in a similar phenotype as the *Poldi* phenotype, namely reduced fertility and immobile spermatozoa [15]. The two other downregulated genes are involved in altering epigenetic modifications (*Arid4b* - [16]) and in transcription coupled repair processes (*Xab2* - [17]). Among the upregulated genes we find five further genes involved in chromatin functions (*Rpa1*, *Thumpd3*, *Prkrip1*, *Larp2*, *H2afz*). This could suggest that *Poldi* has a role in chromatin modification pathways, but this will require further experimental verification.

Although the transcript has two potential open reading frames longer than 100 aminoacids, it is unlikely that either of them is translated. First, their AUG codons are embedded in a suboptimal context for translation initiation and three AUG codons with much shorter reading frames precede these long ORFs. Second, dn/ds ratios calculated between the different in-group species are not significantly different from 1, i.e. there is no evidence for positive or negative selection on these ORFs. Finally, preliminary experiments with antibodies produced against both ORFs did not yield a specific signal on Western blots. Hence, it appears that the RNA exerts its function as a non-coding RNA. The fact that we find a selective sweep in *M. m. musculus* indicates that the gene is subject to ongoing positive selection. Interestingly, analysis of the fixed derived sites in the *M. m. musculus* transcript suggest that two such sites in exon1 would change the folding structure of the RNA (suppl. Fig. 7). A genome-wide search for possible antisense interacting regions in other genes [18] did not yield any candidates.

A second transcript (AK158810) is annotated in the data base on the opposite strand of the *Poldi* transcript . This annotation is based on EST fragments derived from brain RNA, but there are no testis ESTs for this transcript. Northern blots of testis RNA show indeed no signal for *M. m. domesticus*, *M. m. musculus*, *M. m. castaneus* and *M. spretus*, and only a smear for *M. macedonicus*, *M. cypriacus* and *M. spicilegus* (suppl. Fig. 8). It appears therefore that this is one of the spurious transcripts that can be found across the genome [19].

In yeast, a newly evolved gene was shown to have assumed a functional role in DNA repair processes [20]. Interestingly, the transcript of this gene existed already as a non-coding transcript in other species, before its reading frame acquired a function. *Poldi* is thus the first transcript for which a recent direct emergence from intergenic DNA can be unequivocally shown. Still, it is likely that more such genes will be found once more comparative data are available. A recent search for orphan genes in primates, where a sufficiently dense genomic taxon sampling is available, has yielded 15 candidates for genes that appear to have evolved from intergenic regions [7]. This search was focussed on protein coding genes, i.e. even more are likely to be detected if non-coding genes are included. Hence, de novo emergence of genes from intergenic DNA may not be a rare phenomenon. In our case, we find that the relevant splicing signals for the three exons, as well as the polyadenylation signal were already preexisting in the intergenic DNA, suggesting that such cryptic signals are a suitable basis for functionalisation.

The observation that newly emerged genes are preferentially associated with a specific expression in testis [4-6, 21, 22] may be related to the fact that postmeiotic expression requires only relatively simple promotors [12, 13]. Furthermore, sexual conflict mechanisms might be involved to drive the emergence of new gene functions in the testis [23]. By acquiring more complex regulatory elements over time, such genes might eventually become also relevant for somatic tissues and thus to other phenotypic adaptations. De novo functionalization via an initial testis expression might thus be an important additional process leading to the emergence of orphan genes [2, 7].

Material and Methods

M. m. domesticus samples were obtained from locations in France and Germany and *M. m. musculus* from the Czech Republic and Kazakhstan as described [24]. *M. spretus* and *M. m. castaneus* were provided by T. Harr (Plön). *M. famulus*, *M. macedonicus*, *M. cypriacus*, *M. spicilegus*, and *M. caroli* were provided by F. Bonhomme (Montpellier). Animals used in qRT-PCR experiments were caught in live traps and kept in the laboratory for 3-5 days under controlled conditions.

Conditional knock-out mice were generated via gene targeting in Bl/6 mice by Artemis Pharmaceutical (Cologne). Total knock-out mice were obtained from crossing mice carrying the conditional allele with a total Cre-deleter strain [25]. For the sperm analyis 9 week old mice were housed together with a female for one week and separated thereafter into single cages. At the age of 12 weeks the males were sacrificed and the left cauda epididymis with 1cm of vas deferens were dissected. Approx. 10^6 sperm cells were measured and classified with the Ceros sperm analysis system version 12 (Hamilton Thorne Biosciences, Beverly, MA, USA) following the procedure described in [26].

In situ hybridizations were done on paraffin sectioned material using a DIG based labeling system (Roche). Northernblots were done with 10µg total RNA via denaturing agarose gel electrophoresis. Microarray anaylsis was done on the Mouse Genome 430 2.0 array (Affymetrix, Santa Clara). Probe data from Affymetrix CEL-files were normalized with the MAS5 method [27] using the R-based bioconductor software (http://www.bioconductor.org/). The normalized probe set data was searched for differentially expressed genes with the significance analysis of microarrays (SAM) [28].

Acknowledgements

We thank S. Ihle, T. Harr, C. Voolstra, C. Pfeifle and F. Bonhomme for providing DNA, RNA samples and mice, P. Scholl for histology work and S. Carstensen for fragment sequencing. This work was supported by funds from the SFB 680 and funds of the Max-Planck Society.

References

1. Ohno, S. *Evolution by Gene Duplication*. Springer-Verlag, Berlin (1970).
2. Domazet-Loso, T. & Tautz, D. An evolutionary analysis of orphan genes in Drosophila. Genome Res 13, 2213-2219 (2003).
3. Volff J. N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28, 913-922 (2006).
4. Levine, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors & D. J. Begun Novel genes derived from non- coding DNA in Drosophila melanogaster are frequently X-linked and show testis-biased expression. Proc. Natl. Acad. Sci USA 103, 9935– 9939 (2006).
5. Chen S.T., Cheng H.C., Barbash D.A. & Yang H.P. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster. PLoS Genet 3, e107 (2007).
6. Begun D.J., Lindfors H.A., Kern A.D. &Jones C.D. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics 176, 1131-1137 (2007).
7. Toll-Riera M., Bosch N., Bellora N., Castelo R., Armengol L., Estivill X.& Albà M. M. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26, 603-612 (2009).
8. Xiao W, Liu H, Li Y, Li X, Xu C, et al. A Rice Gene of De Novo Origin Negatively Regulates Pathogen-Induced Defense Response. PLoS ONE 4: e4603 (2009).
9. Guenet J.L. & Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet 19, 24-31(2003).
10. Chevret, P., Veyrunes, F., and Britton-Davidian, J. Molecular phylogeny of the genus Mus (Rodentia: Murinae) based on mitochondrial and nuclear data. Biological Journal of the Linnean Society 84, 417-427 (2005).
11. Boguski M.S., Lowe T.M. &Tolstoshev C.M. dbEST--database for "expressed sequence tags". Nature Genetics 4, 332-333 (1993).
12. Kleene, K. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mechanisms of Development 106, 3-23(2001).
13. Acharya K.K., Govind C.K., Shore A.N., Stoler M.H. & Reddi P.P. cis-requirement for the maintenance of round spermatid-specific transcription. Dev Biol 295, 781-790 (2006).
14. Kelly, J.K. A test of neutrality based on interlocus associations. Genetics 146: 1197-1206 (1997).

15. Ronfani L, Ferraguti M, Croci L, Ovitt CE, Schöler HR, Consalez GG, Bianchi ME. Reduced fertility and spermatogenesis defects in mice lacking chromosomal protein Hmgb2. Development. 128: 1265-1273 (2001).
16. Wu MY, Tsai TF, Beaudet AL. Deficiency of Rbbp1/Arid4a and Rbbp1l1/Arid4b alters epigenetic modifications and suppresses an imprinting defect in the PWS/AS domain. Genes Dev. 20: 2859-2870.
17. Yonemasu R, Minami M, Nakatsu Y, Takeuchi M, Kuraoka I, Matsuda Y, Higashi Y, Kondoh H, Tanaka K. Disruption of mouse XAB2 gene involved in pre-mRNA splicing, transcription and transcription-coupled DNA repair results in preimplantation lethality. DNA Repair 4: 479-491. Epub 2005 Jan 19.
18. Li, J.T., Zhang, Y., Kong, L., Liu, Q.R. & Wei, L. Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation. Nucleic Acids Research 36: 4833-4844 (2008).
19. Birney, E.J. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816 (2007).
20. Cai, J., R. Zhao, H. Jiang und W. Wang. De novo origination of a new proteincoding gene in Saccharomyces cerevisiae. Genetics 179, 487-496 (2008).
21. Betrán, E, Thornton, K & Long, M. Retroposed new genes out of the X in Drosophila. Genome Res. 12: 1854-1859 (2002).
22. Emerson, J.J., Kaessmann, H., Betrán, E. & Long, M. Extensive Gene Traffic on the Mammalian X Chromosome. Science 303, 537-540 (2004).
23. Kleene, K. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Developmental Biology 277, 16-26 (2005).
24. Ihle, S., I. Ravaoarimanana, M. Thomas & Tautz, D. An analysis of signatures of selective sweeps in natural populations of the house mouse. Mol Biol Evol 23, 790-797 (2006).
25. Schwenk, F., Baron, U. & Rajewsky, K. A cre-transgenic mouse strain for the ubiquitous deletion of loxP-flanked gene segments including deletion in germ cells. Nucleic Acids Res 23, 5080-5081 (1995).
26. Goossens E., De Block G. & Tournaye H. Computer-assisted motility analysis of spermatozoa obtained after spermatogonial stem cell transplantation in the mouse. Fertility and Sterility 90, 1411-1416 (2008).
27. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307-315 (2004).
28. Tusher, V.G., Tibshirani, R. &Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98, 5116-5121 (2001).
29. Kent, W.J. et al. The human genome browser at UCSC. Genome Res 12, 996-1006 (2002).
30. Kent, W.J. BLAT - the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).

Figures

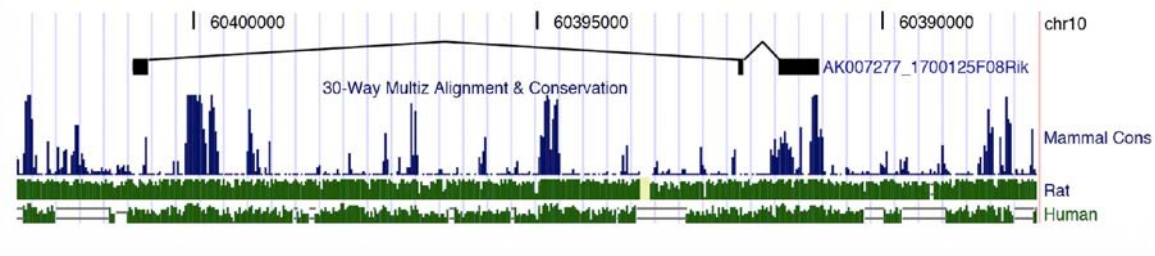


Figure 1

Transcript structure and mammalian conservation pattern of the genomic region on chromosome 10 with the newly evolved transcript. The picture is taken from the UCSC browser [29] (http://genome.ucsc.edu/). The black boxes depict the three exons. The blue track represents the conservation pattern across various mammalian genomes [30]. The green tracks reflect specifically the conservation in Rat and humans respectively.

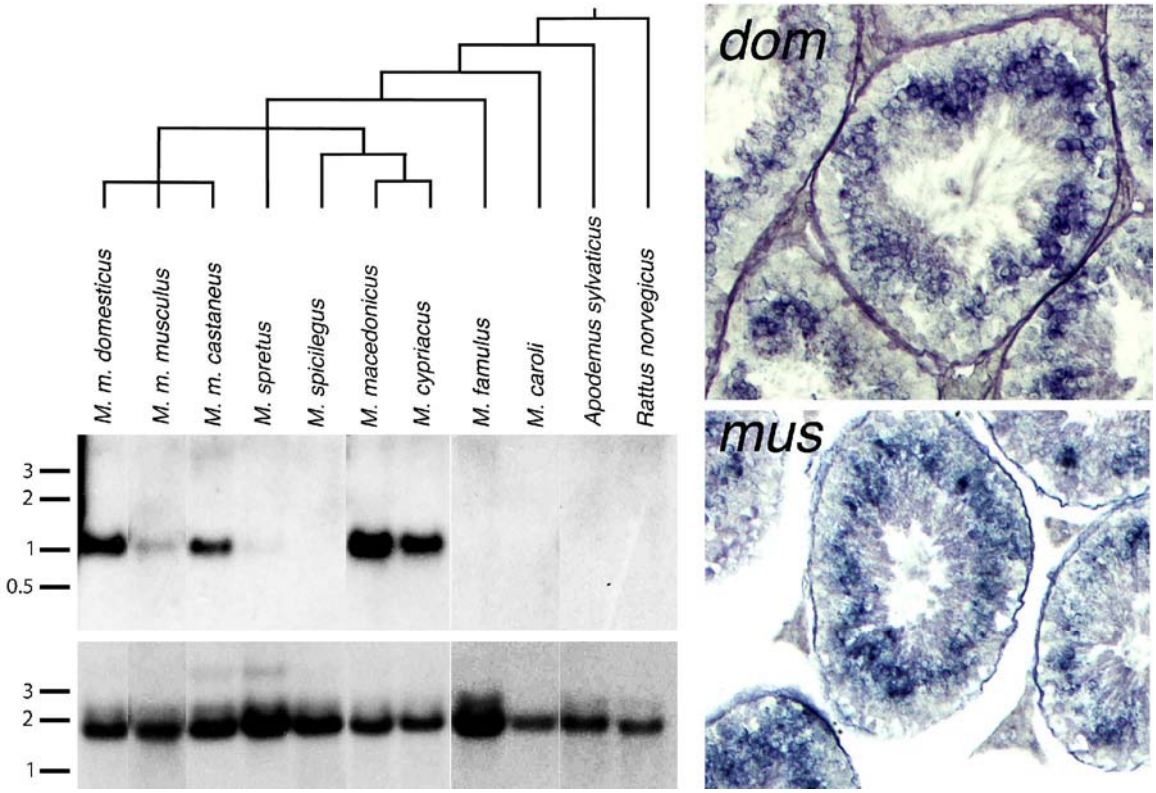


Figure 2

Expression analysis of *Poldi*, based on Northern blots (left) and in situ hybridization (right). Testis RNA from a number of species of the genus *Mus*, as well as the the woodmouse (*Apodemus sylvaticus*) and the rat (*Rattus norvegicus*) was hybridized with a cDNA probe derived from *M. m. domesticus* (upper part). After signal detection, the same blot was rehybridised with a tubulin probe as a loading control (lower part). The phylogenetic relationships of the species used are depicted at the top. Tissue sections (right) of testis from *M. m. domesticus* (*dom*) and *M. m. musculus* (*mus*) were hybridized with a *Poldi* cDNA probe. The signal is restricted to the round spermatid cells of the seminiferous tubules in both sub-species.

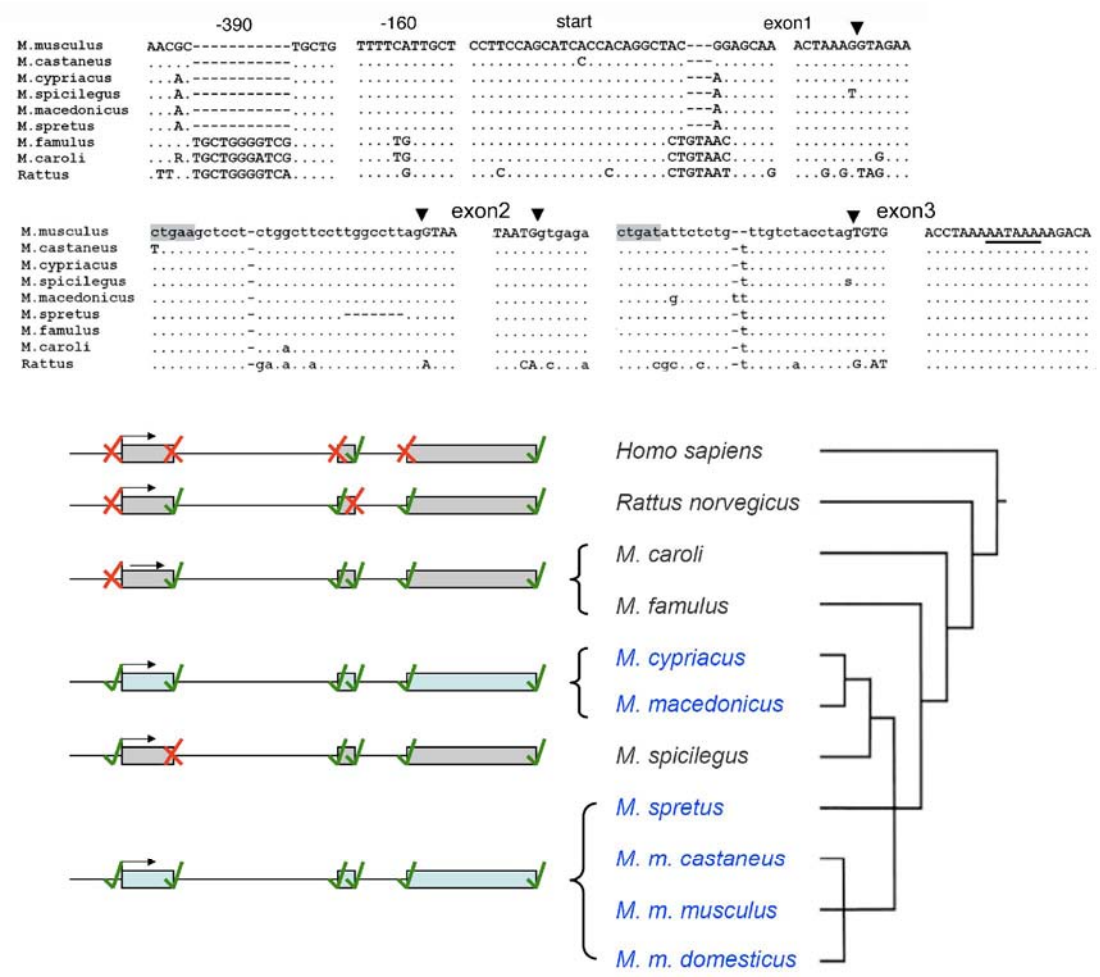


Figure 3

Emergence pattern of functional regions around the *Poldi* gene. (Top) Sequence alignments of upstream and processing regions of *Poldi* between various mouse species and rat. Branch point regions in the introns are shaded in grey; intron-exon junctions are marked with black triangles; the polyadenylation signal is underlined. (Bottom) Sketch showing the changes of the functional regions in the context of the phylogeny. Green checks represent presence of the signal, red crosses represent absence. Exons and introns are not drawn to scale.

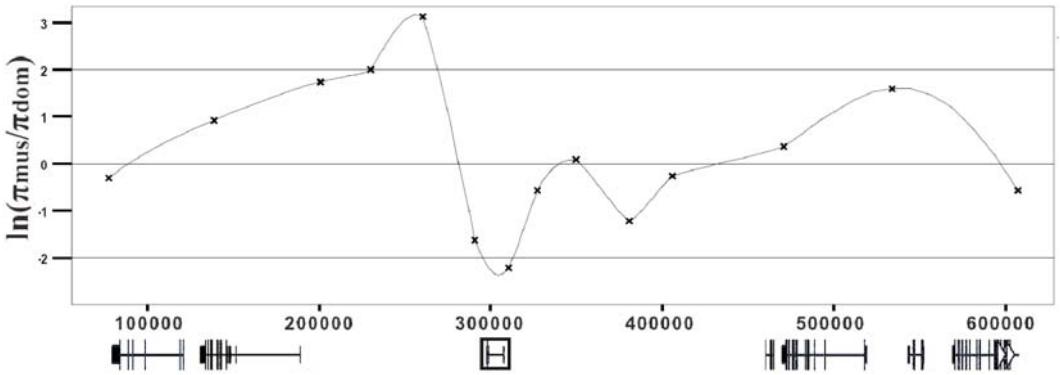
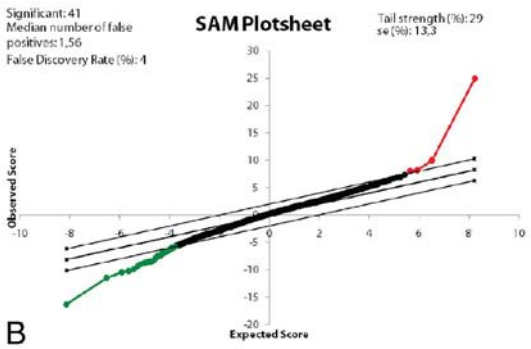


Figure 4

Extended genomic region around the *Poldi* transcript (boxed) and nucleotide variability levels between *M. m. musculus* and *M. m. domesticus* populations. Estimates for π were obtained from 14 sequenced fragments (550bp on average) for eleven animals from each of two populations of both subspecies. π was averaged for both populations and the logarithm of the ratio is plotted on the y-axis. A signature of a sweep in the *M. m. musculus* populations is evident in the region of the *Poldi* transcript. A calculation of Tajima´s D for each population shows a significant negative value for the two fragments in the *Poldi* region in the Kazahkstan (*M. m. musculus*) population ($D = -1.97$, $p < 0.05$ and $D = -1.86$, $p < 0,05$). The corresponding values for the second *M. m. musculus* population (Czech) are higher because of the low numbers of segregating sites (suppl. Table 3), which reduces the power of the test.

		Δ *Poldi*	Wt Bl/6
sperm motility [%]	rapid	21.6 ± 10.2	29 ± 11.1
	medium	4.8 ± 4	5.5 ± 3.3
	slow	9.8 ± 3.7	10.3 ± 7
	static	63.8 ± 15.3	55.2 ± 18.5
testis weight [mg]		84.26 ± 8.03	94.26 ± 9.46

A



B

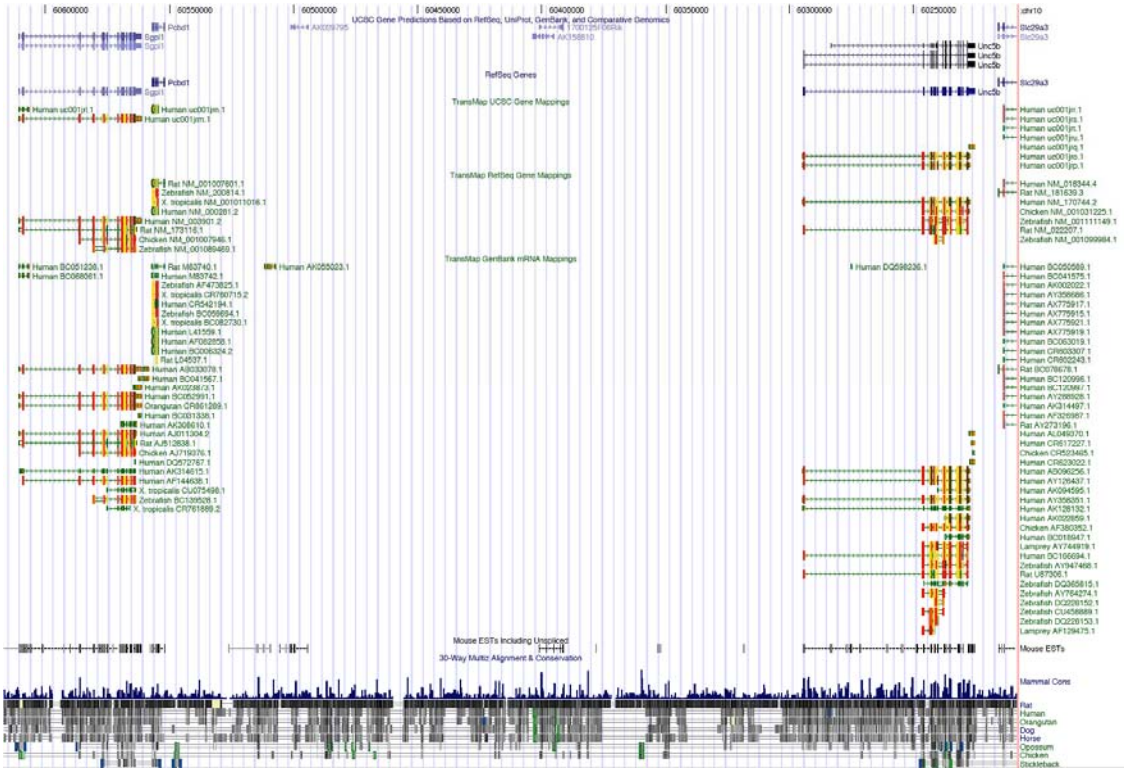
Figure 5

Phenotypic analysis of *Poldi* knock out mice. A) Sperm motility and testis weight. Different sperm motility classes are compared between Δ *Poldi* (n=23) and Bl/6 wild-type (n=21) animals and standard deviations are provided. The percentage of rapidly progressing sperms is significantly reduced in knock out animals (t-test: p=0.0298) while at the same time the number of static cells is increased. Testes of knock out mice (n=19) show significantly reduced weight compared to wild-type animals (n=23) (t-test: p=0.001). B) Microarray analysis. Four knock out animals were compared with four wild-type animals. 37 genes are upregulated in the knock out mice (green) and four genes are down-regulated (red - note that this includes the *Poldi* transcript which is absent in the knock out mice - see suppl. Fig. 5).

Supplementary material

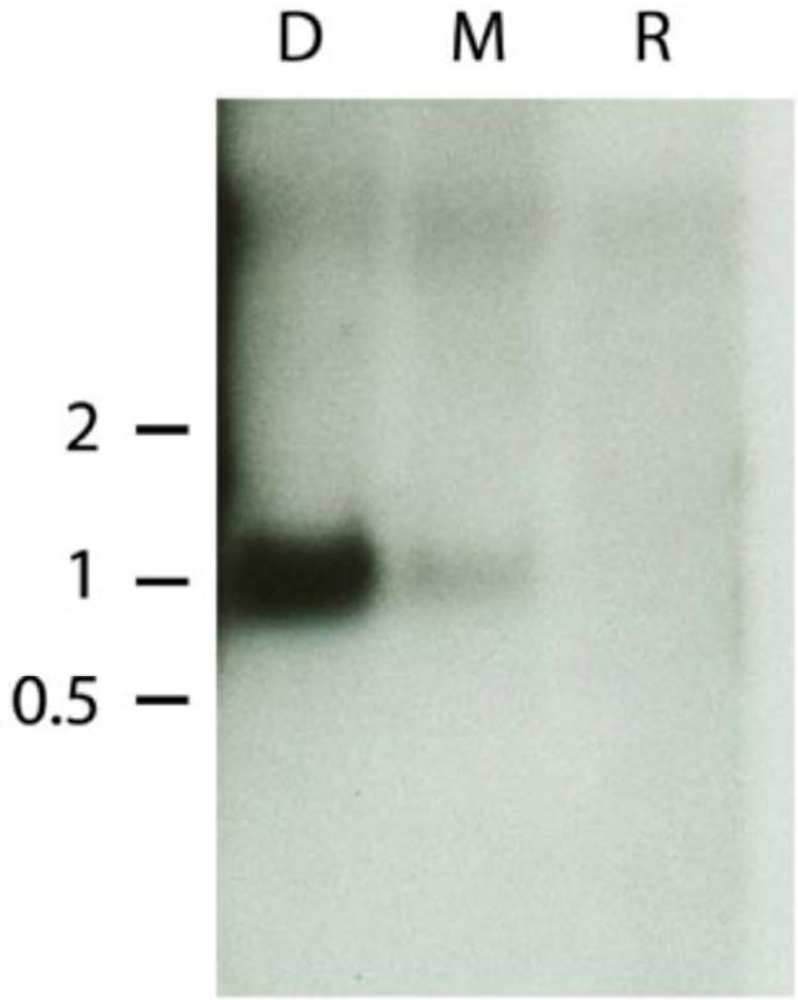
Emergence of a new gene function from an intergenic region

Tobias J.A.J. Heinen, Fabian Staubach, Daniela Häming and Diethard Tautz



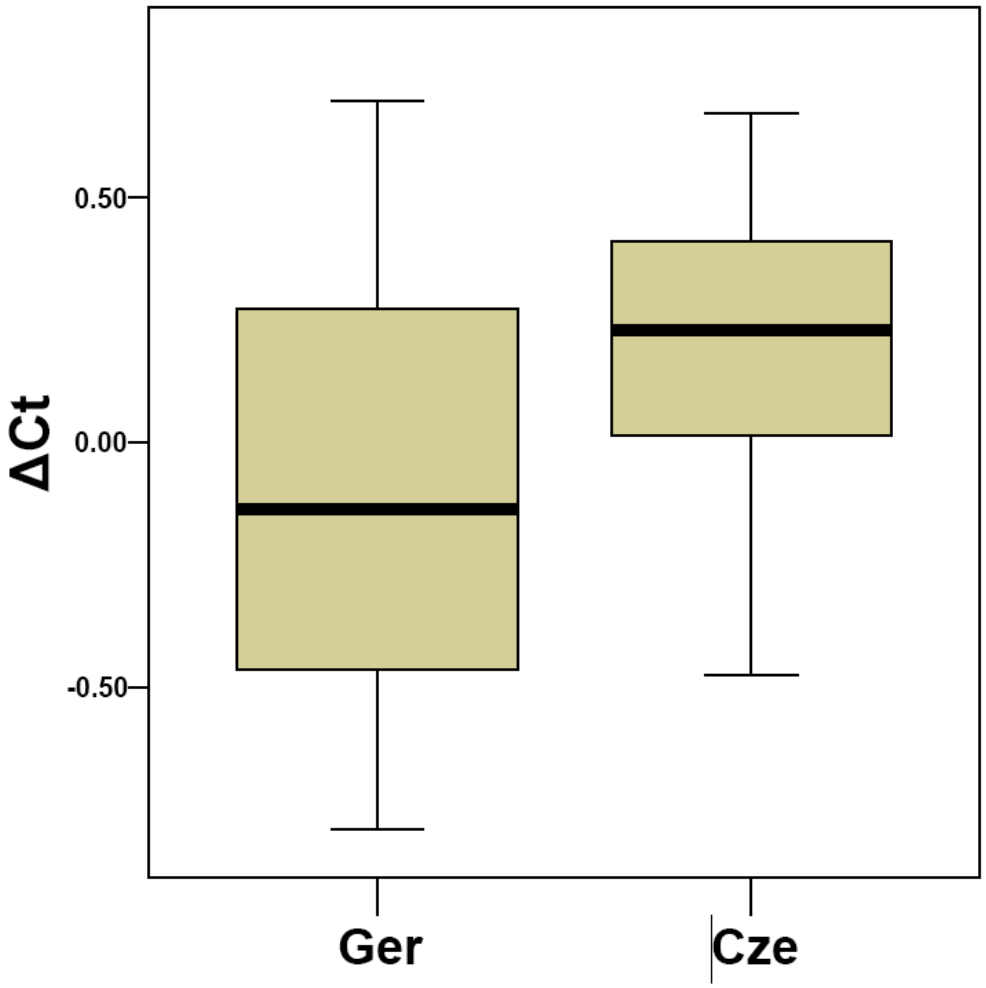
Suppl. Figure 1

UCSC genome browser window of the region on chromosome 10 that includes the *Poldi* transcript (labelled as 1700125F08Rik - top track). The following geneprediction tracks are shown: UCSC Genes, RefSeq Genes, TransMap and Mouse ESTs. Genome conservation tracks across vertebrate genomes are displayed at the bottom. It is evident that a region of ca. 200kb is free of transcripts in other genomes, but alignable. In the mouse, the only other transcripts are AK158810 (Mus musculus visual cortex cDNA, RIKEN full-length enriched library, clone:K430330N23 product:unclassifiable, full insert sequence) and AK029795 (Mus musculus adult male testis cDNA, RIKEN full-length enriched library, clone:4930565A17 product:unclassifiable, full insert sequence). AK158810 overlaps the *Poldi* transcript and is further discussed in the text and in suppl. Fig. 9. AK029795 is approximately 100kb upstream of *Poldi* and was only recently added to the data base - hence we have so far not studied it.



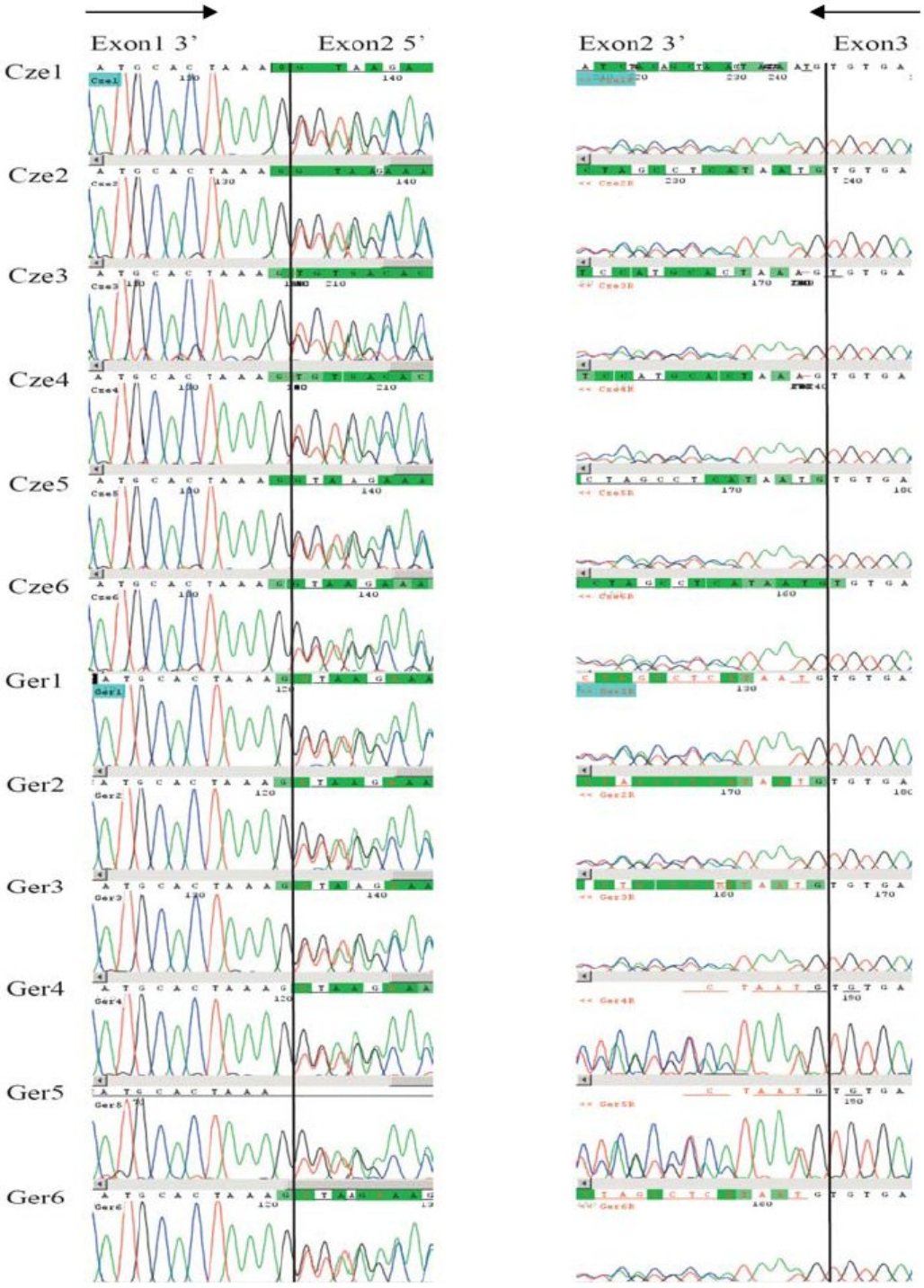
Suppl. Figure 2

Northern blot with testis RNA from *M. m. domesticus* (D), *M. m. musculus* (M) and *Rattus norvegicus* (R) hybridized with a probe derived from the homologous genomic region of the rat. The experiment shows that lack of hybridization in rat is not due to divergence at the probe level. Note that the lower expression level in *M. m. musculus* is within the range of expression polymorphism of this gene in different individuals (compare suppl. Fig. 3).



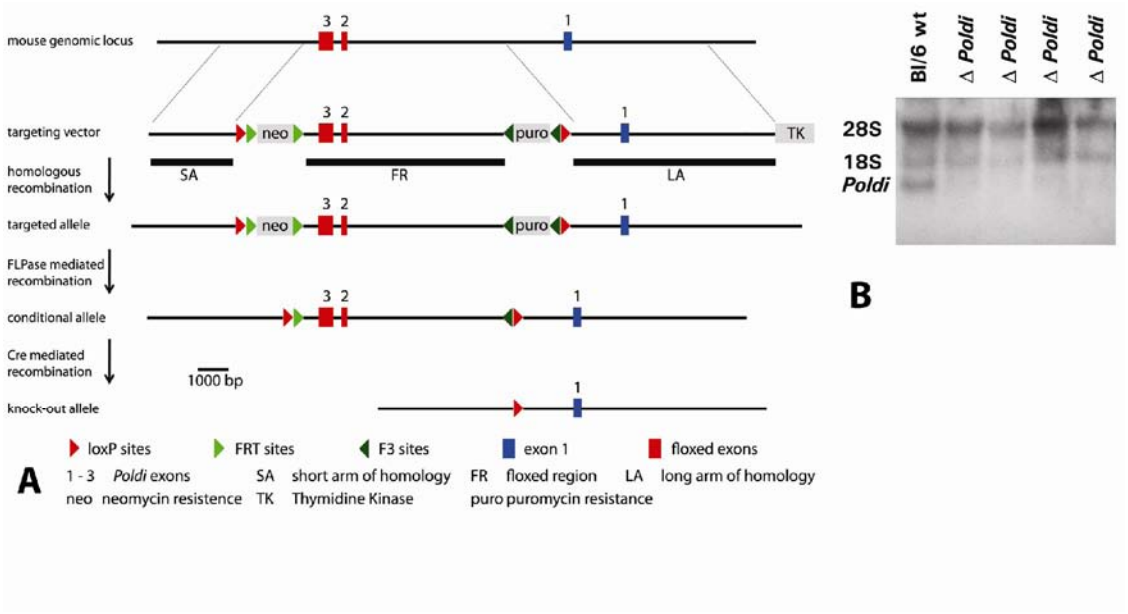
Suppl. Figure 3

Quantitative PCR assays to evaluate the expression level of *Poldi* in six unrelated individuals from *M. m. domesticus* (Ger), *M. m. musculus* (Cze) populations. Although the medians suggest a higher expression (lower ΔCt corresponds to higher expression in RT-PCR) in *M. m. domesticus*, the variance is so high that this is not a statistically significant difference (p = 0.49, Wilcoxon W-Test).



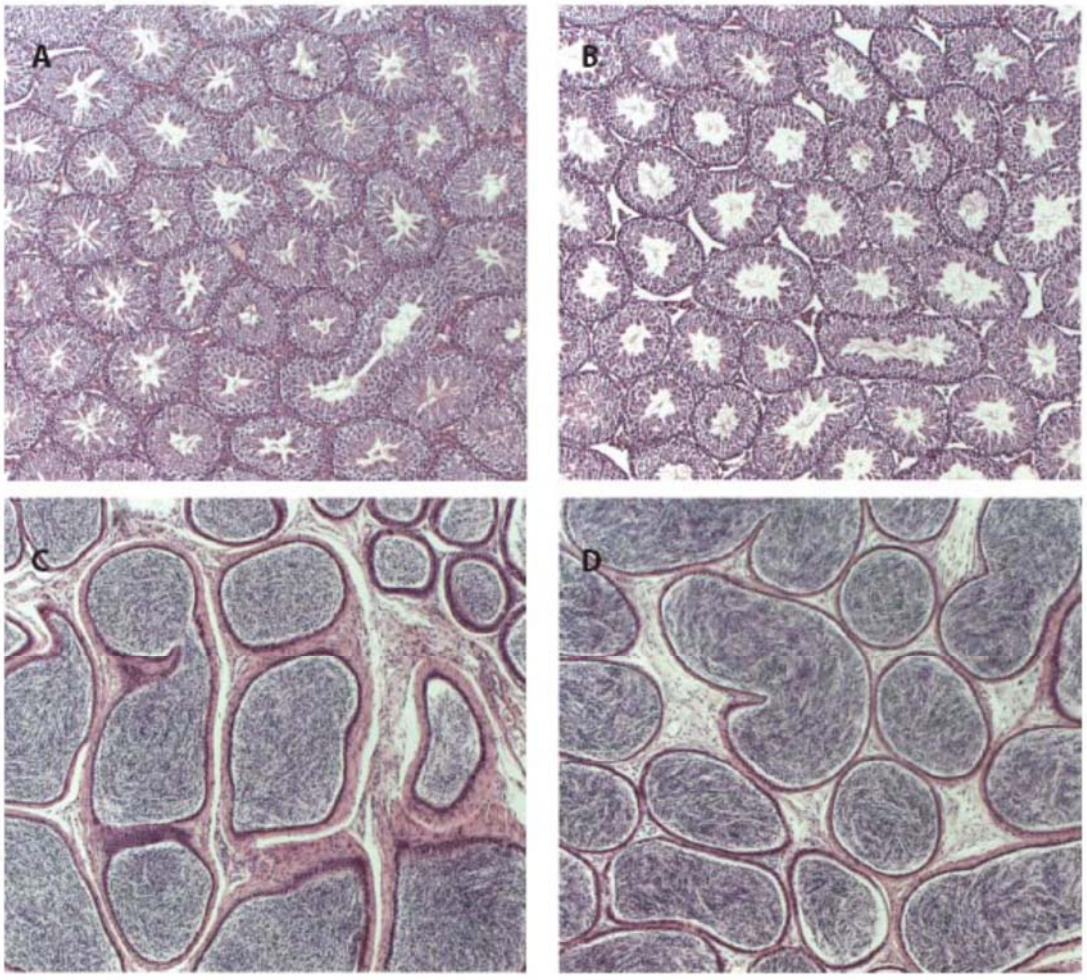
Suppl. Figure 4

Electropherograms of sequencing reactions across the splice junctions of the second *Poldi* exon in six individuals from *M. m. domesticus* (Ger) and *M. m. musculus* (Cze). Double reads are evident from the exon junction onwards, which suggest alternative splicing of this exon in all individuals.



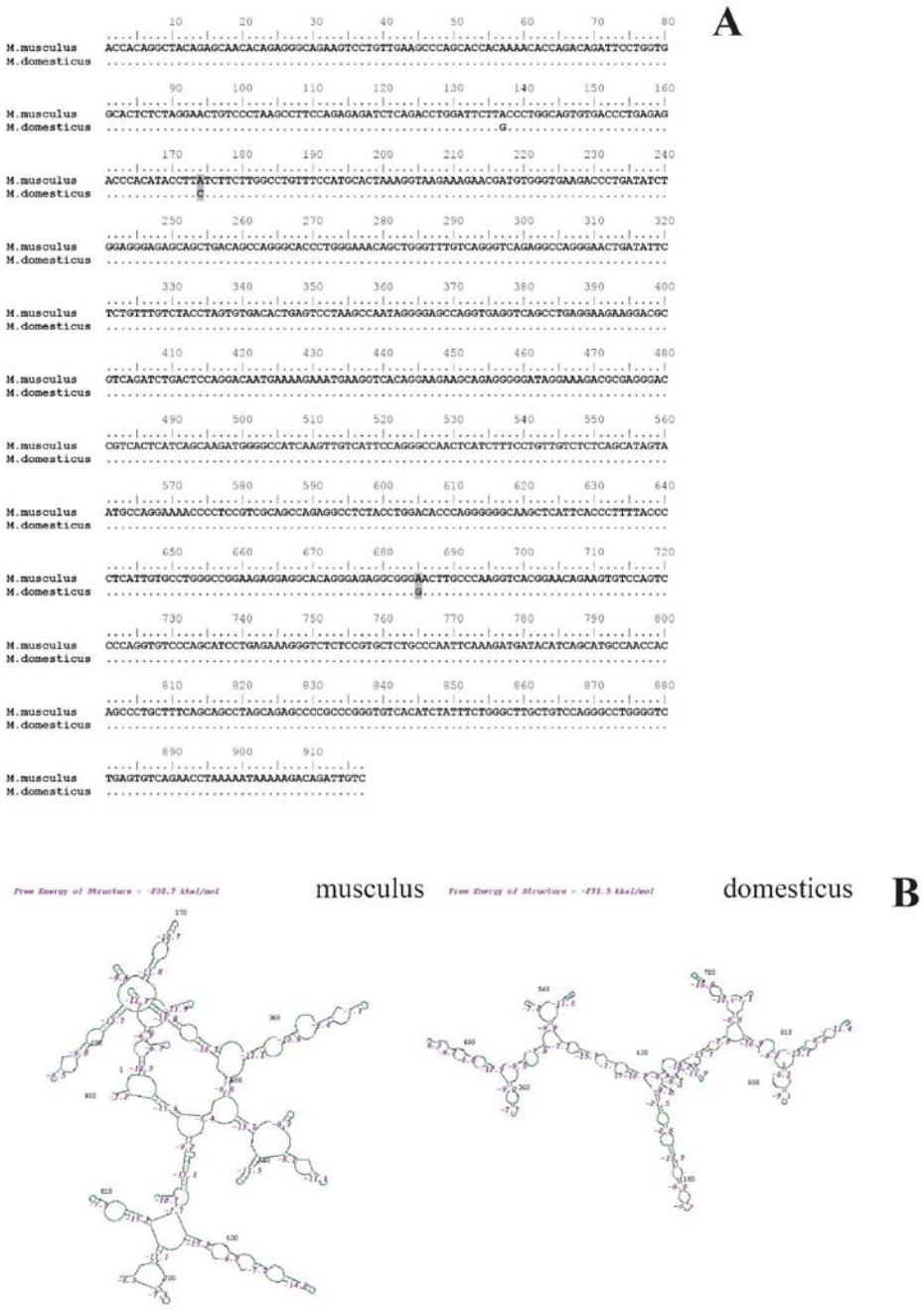
Suppl. Figure 5

Conditional knock-out of *Poldi* in the Bl/6 strain. A) Targeting strategy. A 7887 bp fragment was flanked by loxP sites. One loxP site was placed downstream of the last *Poldi* exon. The second loxP site was placed into the first intron of *Poldi*. The targeting vector contained a FRT-flanked neomycin resistance cassette and a F3-flanked puromycin cassette in order to increase the co-recombination frequency of both loxP sites after selection. Thymidine kinase served as negative selection marker. The neomycin and puromycin resistance cassettes were removed by transient transfection of homologous recombinants with FLPase. One FRT and one F3 site each remained in the genome of the conditional allele. Cre mediated recombination leads to deletion of *Poldi* exons 1 and 2 resulting in a loss of a large region of the RNA including both hypothetical ORFs. B) Validation of knock out. Northernblot with testis RNA from wild-type control animal and four deleted *Poldi* animals hybridized with a *Poldi* cDNA probe. The *Poldi* transcript is absent in knock out individuals. Low stringency washing was performed to visualize ribosomal RNA as internal control.



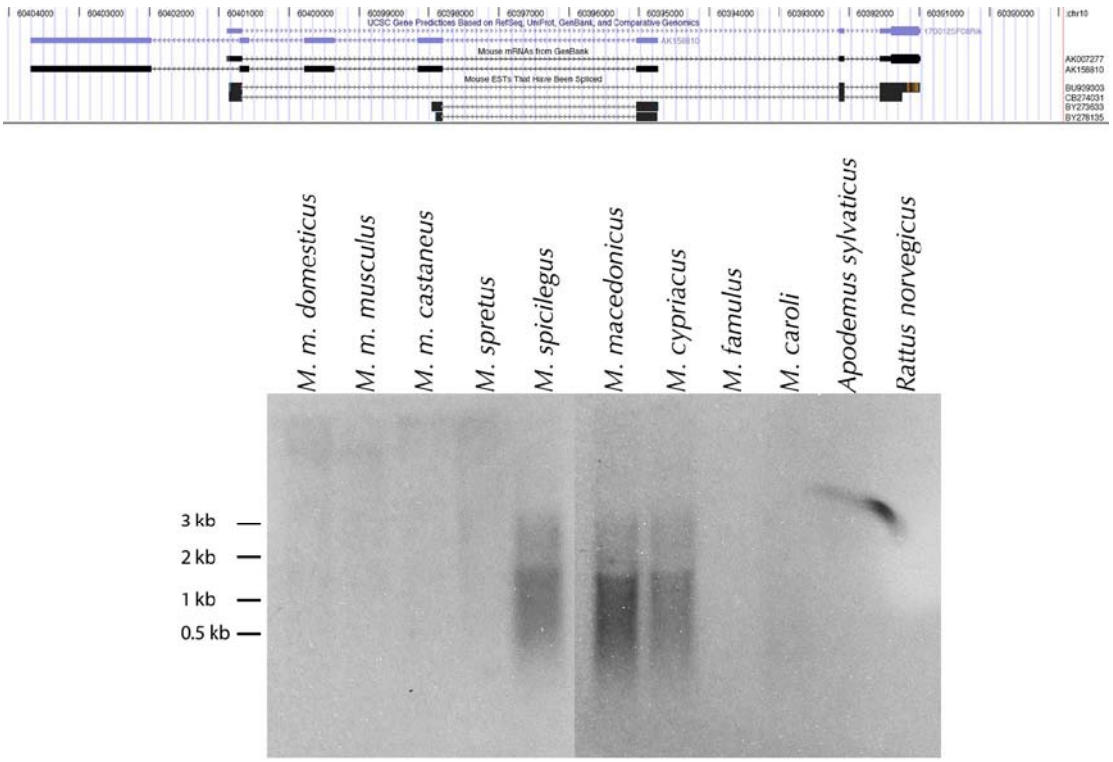
Suppl. Figure 6

Histological analysis of wild type (A+C) and *Poldi* knock out (B+D) mice. The upper two pictures show testis and the lower two show epididymis. Tissue sections were stained with hematoxylin and eosin and photos were taken at 50 fold magnification. No morphological differences can be detected between the wiltype and knock out mice.



Suppl. Figure 7

Comparison of whole trancript sequences (A) and the predicted secondary structures of the *M. m. musculus* and *M. m. domesticus* versions of the *Poldi* transcript, based on the secondary structure prediction module in GeneBee http://www.genebee.msu.su/services/rna2_reduced.html, only fixed differences are considered (B). The two sequences differ only at three fixed derived positions at 137 and 174 in Exon 1 (at 180 and 217), and third fixed dervived position in Exon 3 (pos. 685). The difference at pos. 137 has no influence on the secondary structure.



Suppl. Figure 8

Annotation and expression of the AK158810 transcript. The transcript is transcribed in reverse direction to *Poldi* (1700125F08Rik), but is only known from brain in laboratory mice (*Mus musculus*). The total splice pattern (upper track) is only predicted, but not covered by a full EST. In the testis RNA Northernblot it is only visible in three of the *Mus* species and only as a smear (note that this Northerblot is the same as in Fig. 2, rehybridized after stripping, i.e. the RNA on this blot is intact).

Supplementary Tables

down-regulated in *Poldi* knock out testis						
Gene ID	**Gene Name**	**Score (d)**	**Numerator (r)**	**Denominator (s+s0)**	**Fold Change**	**q-value (%)**
1432460_at	1700125F08Rik (= *Poldi*)	24.80	1076.62	43.41	14.05	0.00
1452534_a_at	Hmgb2	9.94	488.86	49.20	1.16	0.00
1451847_s_at	Arid4b	8.18	107.34	13.13	1.64	2.60
1436886_x_at	Xab2	8.09	179.96	22.26	1.36	2.60

up-regulated in *Poldi* knock out testis						
Gene ID	**Gene Name**	**Score (d)**	**Numerator (r)**	**Denominator (s+s0)**	**Fold Change**	**q-value (%)**
1419451_at	Fzr1	-16.29	-322.03	19.77	0.86	0.00
1436143_at	4933425L03Rik	-11.47	-444.66	38.75	0.82	0.00
1456823_at	Gm70	-10.44	-739.44	70.80	0.85	0.00
1423444_at	Rock1	-10.24	-311.89	30.45	0.86	0.00
1434782_at	Usp42	-9.91	-1365.43	137.84	0.78	0.00
1423293_at	Rpa1	-9.38	-617.01	65.79	0.88	0.00
1417684_at	Thumpd3	-8.99	-491.80	54.68	0.90	0.00
1417425_at	Prkrip1	-8.76	-172.12	19.65	0.88	0.00
1460000_at	Shisa3	-8.66	-132.23	15.27	0.86	0.00
1423553_at	Dnajb3	-8.58	-1666.54	194.24	0.89	0.00
1418757_at	Trim69	-8.56	-602.46	70.35	0.88	0.00
1416868_at	Cdkn2c	-8.47	-361.79	42.73	0.88	0.00
1451470_s_at	Eif5a	-8.28	-1031.55	124.66	0.95	0.00
1448903_at	Sep15	-7.83	-556.57	71.09	0.88	0.00
1417838_at	Ssty2	-7.52	-2549.77	339.13	0.88	0.00
1452040_a_at	Cdca3	-7.45	-990.17	132.94	0.89	0.00
1432186_at	1700028J19Rik	-7.43	-1325.48	178.42	0.87	0.00
1423801_a_at	Aprt	-7.24	-544.33	75.16	0.86	3.90
1424712_at	Ahctf1	-7.06	-1016.22	144.03	0.86	3.90
1418473_at	Cutc	-7.00	-386.83	55.29	0.80	3.90
1423833_a_at	Brp44	-6.91	-2732.58	395.66	0.87	3.90
1456660_a_at	0610010F05Rik	-6.84	-301.44	44.05	0.89	3.90
1430886_at	1700112E06Rik	-6.68	-171.27	25.64	0.06	3.90
1451337_at	Psmf1	-6.67	-1383.45	207.45	0.86	3.90
1420920_a_at	Arf1	-6.66	-337.16	50.63	0.90	3.90
1420851_at	Pard6g	-6.57	-109.90	16.73	0.77	3.90
1431086_s_at	Pcmt1	-6.34	-422.94	66.76	0.91	4.59
1442871_at	LOC100042492	-6.28	-486.48	77.53	0.54	4.59
1449177_at	Ccna1	-6.24	-556.35	89.17	0.81	4.59
1432353_at	Larp2	-6.21	-141.24	22.75	0.81	4.59
1454726_s_at	Ptpdc1	-6.07	-502.81	82.83	0.91	4.59

1452499_a_at	Kif2a	-6.05	-410.17	67.75	0.86	4.59
1418576_at	Yipf5	-5.97	-465.00	77.93	0.84	4.59
1431871_at	Txndc3	-5.95	-405.44	68.19	0.89	4.59
1416415_a_at	H2afz	-5.88	-2068.67	351.58	0.91	4.59
1422471_at	Pex13	-5.80	-941.83	162.30	0.89	5.57
1437535_at	Ppp3r2	-5.76	-1432.61	248.93	0.89	5.57

Suppl. Table 1

Results of the Affymetrix Genechip experiment comparing testis expression in *Poldi* knock-out mice and wild-type mice. Significantly differentially expressed genes calculated by SAM analysis (FDR=4%. delta=1.98) are presented in SAM output format.

GOID	GO_term	Frequency	Genome frequency	Corrected P-value	Gene(s)
GO:0007126	Meiosis	0.05405	0.00171	0.0843	Fzr1, Rpa1
GO:0051327	M phase of meiotic cell cycle	0.05405	0.00171	0.0843	Fzr1, Rpa1
GO:0051321	meiotic cell cycle	0.05405	0.00173	0.0872	Fzr1, Rpa1
GO:0000279	M phase	0.05405	0.00285	0.2311	Fzr1, Rpa1
GO:0009987	cellular process	0.35135	0.16950	0.2705	Pex13, Cdkn2c, Sep15, Arf1, Ahctf1, Aprt, Rock1, Prkrip1, Ppp3r2, Fzr1, Eif5a, Pcmt1, Rpa1
GO:0022403	cell cycle phase	0.05405	0.00400	0.4437	Fzr1, Rpa1
GO:0044237	cellular metabolic process	0.18919	0.06957	0.5838	Aprt, Pex13, Prkrip1, Ppp3r2, Sep15, Ahctf1, Pcmt1
GO:0022402	cell cycle process	0.05405	0.00494	0.6635	Fzr1, Rpa1
GO:0008152	metabolic process	0.18919	0.07174	0.6832	Aprt, Pex13, Prkrip1, Ppp3r2, Sep15, Ahctf1, Pcmt1
GO:0015031	protein transport	0.05405	0.00547	0.8038	Pex13, Arf1
GO:0045184	establishment of protein localization	0.05405	0.00550	0.8120	Pex13, Arf1
GO:0044238	primary metabolic process	0.16216	0.06483	1.0000	Aprt, Pex13, Ppp3r2, Sep15, Ahctf1, Pcmt1
GO:0008104	protein localization	0.05405	0.00750	1.0000	Pex13, Arf1
GO:0007049	cell cycle	0.05405	0.00753	1.0000	Fzr1, Rpa1
GO:0007010	cytoskeleton organization and biogenesis	0.05405	0.00759	1.0000	Pex13, Rock1
GO:0033036	macromolecule localization	0.05405	0.00785	1.0000	Pex13, Arf1
GO:0007610	behavior	0.05405	0.00920	1.0000	Aprt, Pex13

GO:0043170	macromolecule metabolic process	0.13514	0.05407	1.0000	Pex13, Ppp3r2, Sep15, Ahctf1, Pcmt1
GO:0048523	negative regulation of cellular process	0.08108	0.02408	1.0000	Cdkn2c, Prkrip1, Rock1
GO:0044260	cellular macromolecule metabolic process	0.08108	0.02602	1.0000	Ppp3r2, Sep15, Pcmt1
GO:0048519	negative regulation of biological process	0.08108	0.02637	1.0000	Cdkn2c, Prkrip1, Rock1
GO:0006915	apoptosis	0.05405	0.01426	1.0000	Rock1, Eif5a
GO:0012501	programmed cell death	0.05405	0.01455	1.0000	Rock1, Eif5a
GO:0008219	cell death	0.05405	0.01497	1.0000	Rock1, Eif5a
GO:0016265	death	0.05405	0.01526	1.0000	Rock1, Eif5a
GO:0006996	organelle organization and biogenesis	0.05405	0.01585	1.0000	Pex13, Rock1
GO:0050896	response to stimulus	0.08108	0.03517	1.0000	Aprt, Pex13, Ppp3r2
GO:0044267	cellular protein metabolic process	0.05405	0.02255	1.0000	Sep15, Pcmt1
GO:0019538	protein metabolic process	0.05405	0.02349	1.0000	Sep15, Pcmt1
GO:0006810	transport	0.05405	0.02414	1.0000	Pex13, Arf1
GO:0051234	establishment of localization	0.05405	0.02520	1.0000	Pex13, Arf1
GO:0043283	biopolymer metabolic process	0.08108	0.04666	1.0000	Ppp3r2, Ahctf1, Pcmt1
GO:0031323	regulation of cellular metabolic process	0.05405	0.02876	1.0000	Prkrip1, Ahctf1
GO:0019222	regulation of metabolic process	0.05405	0.02955	1.0000	Prkrip1, Ahctf1
GO:0016043	cellular component organization and biogenesis	0.05405	0.02981	1.0000	Pex13, Rock1
GO:0006139	nucleobase. nucleoside. nucleotide and nucleic acid metabolic process	0.05405	0.03284	1.0000	Aprt, Ahctf1
GO:0048869	cellular developmental process	0.05405	0.03481	1.0000	Pex13, Rock1
GO:0051179	localization	0.05405	0.03561	1.0000	Pex13, Arf1
GO:0032502	developmental process	0.08108	0.06904	1.0000	Pex13, Rock1, Eif5a
GO:0050794	regulation of cellular process	0.10811	0.10294	1.0000	Cdkn2c, Prkrip1, Rock1, Ahctf1
GO:0050789	regulation of biological process	0.10811	0.10649	1.0000	Cdkn2c, Prkrip1, Rock1, Ahctf1
GO:0048856	anatomical	0.05405	0.05425	1.0000	Pex13, Rock1

	structure development				
GO:0065007	biological regulation	0.10811	0.11499	1.0000	Cdkn2c, Prkrip1, Rock1, Ahctf1
GO:0032501	multicellular organismal process	0.05405	0.10411	1.0000	Pex13, Ppp3r2
GO:XXXXX	unannotated	0.54054	0.75334	1.0000	0610010F05Rik, Usp42, Pard6g, Brp44, 4933425L03Rik, Dnajb3, 1700028J19Rik, H2afz, Gm70, Kif2a, 1700112E06Rik, Txndc3, Ccna1, Trim69, Shisa3, Larp2, LOC100042492, Ptpdc1, Psmf1, Ssty2
GO:0008150	biological_process	0.10811	0.73391	1.0000	0610010F05Rik, Pard6g, 4933425L03Rik, 1700028J19Rik, Dnajb3, H2afz, Ppp3r2, 1700112E06Rik, Fzr1, Txndc3, Pcmt1, Trim69, Shisa3, Larp2, Cdkn2c, Sep15, Ahctf1, Psmf1, Rock1, Eif5a, Cutc, Yipf5, Usp42, Arf1, Thumpd3, Brp44, Prkrip1, Gm70, Kif2a, Cdca3, Ccna1, Pex13, LOC100042492, Aprt, Ptpdc

Suppl. Table 2

Gene ontology annotations of genes that are significantly higher expressed in testis of *Poldi* knock out mice. Annotation is based on SAM-analysis shown in suppl. Table 1and was obtained using the MGI Gene Ontology Term Finder (http://www.informatics.jax.org/gotools/MGI_Term_Finder.html).

population	position	N (chromosomes)	bp	π (x 10^{-2})	segr sites	θ	Tajima's D
KAZ	77305	22	573	0.181	2	0.00096	2.02478
KAZ	138639	18	616	0.087	2	0.00094	-0.19106
KAZ	200858	18	657	0.098	3	0.00133	-0.71573
KAZ	230367	22	513	0.49	6	0.00321	1.63658
KAZ	260533	22	241	0.345	2	0.00228	1.1667

KAZ	291129	22	561	0.103	2	0.00098	0.11197
KAZ	310483	20	752	0.066	5	0.00187	-1.97429
KAZ	326970	22	625	0.175	9	0.00395	-1.86487
KAZ	350121	22	602	0.12	4	0.00182	-0.95805
KAZ	380921	20	457	0.532	12	0.0074	-1.00912
KAZ	406192	22	607	0.095	2	0.0009	0.11197
KAZ	470828	18	502	0.189	2	0.00116	1.5371
KAZ	533888	22	608	0.246	4	0.0018	1.02513
KAZ	607121	20	396	0.128	1	0.00071	1.43024
CR	77305	22	573	0.183	2	0.00096	2.06053
CR	138639	22	616	0.067	1	0.00045	0.89527
CR	200858	22	657	0.121	3	0.00125	-0.08306
CR	230367	22	519	0.108	2	0.00106	0.04046
CR	260533	22	241	0.519	4	0.00455	0.39361
CR	291129	16	561	0.058	1	0.00054	0.15575
CR	310483	28	752	0.113	3	0.00103	0.25201
CR	326970	22	625	0.249	9	0.00395	-1.23603
CR	350121	22	602	0.397	13	0.00592	-1.16856
CR	380921	22	457	0.241	4	0.0024	0.00584
CR	406192	22	607	0.623	17	0.00768	-0.69334
CR	470828	22	501	0.062	1	0.00055	0.23682
CR	533888	22	608	0.382	7	0.00316	0.67398
CR	607121	18	396	0.464	6	0.00441	0.174
GER	77305	20	573	0.418	5	0.00246	2.13938
GER	138639	20	616	0.031	1	0.00046	-0.59155
GER	200858	14	657	0	0	0	0
GER	230367	19	519	0	0	0	0
GER	260533	20	241	0	0	0	0
GER	291129	20	561	0.434	5	0.00251	2.23079
GER	310483	30	752	0.69	15	0.00537	0.96493
GER	326970	18	625	0.368	5	0.00233	1.84421
GER	350121	18	602	0.195	4	0.00193	0.03489
GER	380921	16	457	2.068	21	0.01385	1.9917
GER	406192	14	607	0.396	6	0.00311	0.98915
GER	470828	18	502	0.044	2	0.00116	-1.50776
GER	533888	18	608	0.086	1	0.00048	1.50518
GER	607121	18	396	0.635	5	0.00367	2.31385
FRA	77305	22	573	0.079	5	0.00239	-1.98725
FRA	138639	22	616	0.03	2	0.00089	-1.51481
FRA	200858	22	657	0	0	0	0
FRA	230367	22	519	0.08	1	0.00053	0.89527
FRA	260533	22	241	0.038	1	0.00114	-1.1624
FRA	291129	18	561	0.393	5	0.00259	1.63074
FRA	310483	18	752	0.937	16	0.00619	1.96488
FRA	326970	22	625	0.391	6	0.00263	1.50617
FRA	350121	22	602	0.283	7	0.00319	-0.35767
FRA	380921	22	457	0.532	11	0.0066	-0.67042
FRA	406192	22	607	0.54	11	0.00497	0.2976
FRA	470828	22	502	0.133	3	0.00164	-0.49124
FRA	533888	22	608	0.041	1	0.00045	-0.17472
FRA	607121	22	396	0.42	5	0.00346	0.63001

Suppl. Table 3

Population genetical data for the region surveyed around the *Poldi* locus. The two *M. m. musculus* populations are KAZ and CR, the two *M. m. domesticus* are GER and FRA. The "position" values refer to

the window shown in Fig. 3. The two fragments spanning the *Poldi* transcript region are at 310483 and 326970.