Evolution of a new testis-specific functional promotor within the highly conserved *Map2k7* gene of the mouse

Tobias Heinen[2], Chen Xie, Maryam Keshavarz[1,3], Dominik Stappert[3], Sven Künzel[1], Diethard Tautz[1]*

[1]Max-Plank Institute for Evolutionary Biology, August-Thienemann-Strasse 2, 24306 Plön, Germany

* Corresponding author: tautz@evolbio.mpg.de

[2] Uniklinik Köln, Köln, Germany

[3] Deutsches Zentrum für Neurodegenerative Erkrankungen e. V. (DZNE), Bonn, Germany

Keywords: regulatory evolution, new promotor, kinase, testis, sperm maturation

Abstract

***Map2k7* (synonym *Mkk7*) is a conserved regulatory kinase gene and a central component of the JNK signaling cascade with key functions during cellular differentiation. It shows complex transcription patterns and different transcript isoforms are known in the mouse (*Mus musculus)*. We have previously identified a newly evolved testis specific transcript for the *Map2k7* gene in the subspecies *M. m. domesticus*. Here, we identify the new promotor that drives this transcript and find that its transcript codes for an open reading frame (ORF) of 50 amino acids. The new promotor was gained in the stem lineage of closely related mouse species, but was secondarily lost in the subspecies *M. m. musculus* and *M. m. castaneus*. A single mutation can be correlated with its transcriptional activity in *M. m. domesticus* and cell culture assays demonstrate the capability of this mutation to drive expression. A mouse knock-out line in which the promotor region of the new transcript is deleted reveals a functional contribution of the newly evolved promotor to sperm motility and to the spermatid transcriptome. Our data show that a new functional transcript (and possibly protein) can evolve within an otherwise highly conserved gene, supporting the notion of regulatory changes contributing to the emergence of evolutionary novelties.**

Introduction

Mitogen activated protein kinase (MAPK) pathways are highly conserved throughout eukaryotes and trigger multistep signaling cascades mediating transcriptional response upon reception of outside stimuli (Chang and Karin 2001, English et al. 1999). *Map2k7* belongs to the JNK group of kinases and acts as its specific activator (Fleming et al. 2000, Holland et al. 1997, Kishimoto et al. 2003, Takekawa et al. 2005, Tournier et al. 2001, Tournier et al. 1997, Wang et al. 2007). Cellular stresses like UV and gamma irradiation, osmotic shock and drug treatments on the one hand and different inflammatory cytokines, such as tumor necrosis factor, interleukin-1 or interleukin-3 on the other hand, lead to JNK pathway activation (Chang and Karin 2001, Foltz et al. 1998, Moriguchi et al. 1997, Nishina et al. 2004). Downstream targets of JNK include transcription factors (Yang et al. 2003) as well as other proteins, for example microtubule-associated proteins (Chang et al. 2003) and members of the *Bcl2* family (Deng et al. 2003, Lei et al. 2002). The JNK pathway has several known functions in the immune system, in apoptosis and in developmental processes (Dong et al. 2000, Nishina, Wada and Katada 2004, Sabapathy et al. 1999, Wada et al. 2004, Wang, Destrument and Tournier 2007). Double mutant mice lacking the JNK1 and JNK2 isoforms, as well as *Map2k7* total knockout mice, lead to embryonic lethality (Wang, Destrument and Tournier 2007)..

In their report on the first identification of *Map2k7* in mice, Tournier and colleagues (Tournier, Whitmarsh, Cavanagh, Barrett and Davis 1997) showed by Northern blotting the expression of a long transcript in various organs, plus an additional shorter transcript specific to the testis. While all the further studies concentrated on the longer transcript, the origin and function of the shorter one was left unaccounted for. In a systematic study for differentially expressed genes in mouse populations, we identified *Map2k7* to be differentially expressed in testis between wild populations of *M. m. domesticus* and *M. m. musculus*, with strongly elevated expression in *M. m. domesticus* (Harr et al. 2006). A cis-trans test via allele specific expression analysis in F1 hybrids of both subspecies demonstrated that the expression change is caused by a cis-acting sequence. It turned out that the elevated expression level can be correlated with the additional testis specific transcript in the *M. m. domesticus* subspecies which is absent in *M. m. musculus*. This suggested that a new testis-specific promotor had evolved within the *Map2k7* gene (Harr, Voolstra, Heinen, Baines, Rottscheidt, Ihle, Mueller, Bonhomme and Tautz 2006). Given the highly conserved nature of the *Map2k7* gene, such an evolution of a strong new promotor is of special interest. We present here comparative and functional data that allow inferences on the evolutionary history of the new promotor, which includes both a new origination event, as well as a secondary loss event triggered by a single mutation in one of the subspecies. The knockout analysis proofs that the new promotor has assumed a new function in the maturation of spermatids and the regulation of the transcriptome during this phase.

Materials and Methods

Ethics statement

The work did not involve in vivo experiments with animals. Mouse samples were taken from mice derived from the maintenance of the mouse strain collections at our institute (Harr et al. 2016). Maintenance and handling of mice in the facility were conducted in accordance with German animal welfare law (Tierschutzgesetz) and FELASA guidelines. Permits for keeping mice were obtained from the local veterinary office 'Veterinäramt Kreis Plön' (permit number: 1401–144/PLÖ-004697). The respective animal welfare officer at the University of Kiel was informed about the sacrifice of the animals for this study.

In situ hybridization and Northern blotting

In situ detection of Map2k7 RNA was performed by hybridization with a digoxigenin (DIG) labeled probe (TAUTZ and PFEIFLE 1989). For probe generation, a fragment spanning Map2k7 exons 5-10 was amplified from testis C57Bl6 cDNA with primers P49 and P50 and cloned into a PCR cloning vector. The DNA fragment was reamplified from a pure plasmid clone. Reverse transcription to generate a DIG labeled probe was set up by adding 200 ng of purified PCR product to 2 µL DIG RNA Labeling Mix, 2 µL transcription buffer, 2 µL T7 polymerase, 0.5 µL RNase inhibitor (Roche, Basel). Pure water was added to the reaction mix to obtain a final volume of 20 µL. The reaction mix was incubated for 2 h at 37°C followed by a treatment with 1µL Turbo DNAse for 15 min at 37°C to remove the DNA template. The probe was precipitated with salt and alcohol, washed and re-suspended in 40 µL of 50% formamide diluted in nuclease free water (Applied Biosystems / Ambion, Austin).

All buffers and tools that were used for the following procedure were kept RNAse free. Paraffinized sections were dewaxed in xylene for 2x 10 min, washed for 5 min in ethanol, rehydrated in a series of decreasing ethanol concentration (95%, 90%, 70%, 30%; 3 min each) and washed for 5 min in PBS before postfixing them for 1 h in 4% PFA. After postfixation the tissue was washed in PBS for 2x 5 min and partially digested with 10 µg/mL proteinase K in 100 mM Tris-HCl pH 7.5 for 10 min at 37°C.

Digestion was stopped with 0.2% glycine in PBS. 2x 5 min washing in PBS was followed by 15 min incubation in 0.1 N HCl and another 2x 5 min washing in PBS was performed previous to blocking of positively charged amino acids by 0.25% acetic anhydride in 0.1 M triethanolamine pH 8.0 for 10 min. Afterwards slides were washed for 5 min in PBS and for 5 min in pure water before prehybridization for 2 h at 65°C (50% formamide, 5x SSC, 1x Denhardt's, 0.1% Tween-20). 1 µL of DIG labeled probe was diluted in 100 µL prehybridization buffer containing 400 ng tRNA (Sigma-Aldrich, St. Louis) and denatured at 70°C for 5 min. The hybridization mix was applied to the sections and covered with coverslips. Slides were incubated over night at 65°C in a moist chamber. Next day, the sections were washed in 50% formamide containing 5x SSC and 1% SDS at 70°C for 30 min and subsequently with 50% formamide containing 2x SSC and 0.2% SDS for another 30 min at 65°C. Afterwards the sections were washed for 3x 5 min in MABS (100 mM maleic acid, 150 mM NaCl, 0.1% Tween-20 and 2 mM levamisole; adjusted to pH 7.5 with NaOH).

Samples were blocked with 1% blocking reagent (Roche, Basel) in MABS. Anti-DIG-AP antibody was applied in 1% blocking reagent in MABS by overnight incubation at 4°C. Next day, the sections were first washed 3x 10 min and then 3x 30 min in MABS. Subsequently, pH was adjusted by incubating for 3x 10 min in NTMLT buffer (100 mM Tris-HCl pH 9.5, 50 mM $MgCl_2$, 100 mM NaCl, 100 mM levamisole, 0.1% Tween-20). BM purple solution (Roche, Basel) was applied as substrate for the alkaline phosphatase coupled with the anti-DIG antibody. Tissue was stained until the desired degree of signal was observed. Slides were washed 1 min in water and mounted with Kaisers glyceringelatine.

Detection of RNA in Northern Blotting (Alwine et al. 1977) was performed with radioactively labeled probes generated from the same clone that was used for in situ hybridization. Probes were labled with ^{32}P-dCTP (Hartmann Analytic, Braunschweig) by the use of the Rediprime II DNA Labeling Kit (GE Healthcare Life Science, Little Chalfont) according to the manufacturer's manual. Labeled probes were cleaned up with MicoSpin S-200 HR columns (GE Healthcare Life Science, Little Chalfont) according to the manufacturer's manual.

10 µg of total RNA per sample were diluted in 15 µL nuclease free pure water (Applied Biosystems/Ambion, Austin) and mixed with 10 µL sample buffer (50% formamide, 5.18% formaldehyde, 2.5x MOPS, 0.1 mg/ml ethidiumbromide and 2.5x blue marker). Samples were heat-denatured for 5 min at 70°C and separated on an agarose gel (1.2% agarose, 0.666% formaldehyde, 1x MOPS). The RNA lanes were blotted through classic upward blot onto a Amersham Hybond N+ membrane (GE Healthcare Life Science, Little Chalfont) by neutral transfer (20x SSC) over night. Membranes were baked for 2 h at 80°C and prehybridized in ExpressHyb (Clontech, Mountain View) at 65°C for 1h. Radioactively labeled probe was added to the prehybridized blot and hybridization took place over night at 65°C in a rotating oven. Next day, the blots were washed 10 – 40 min in 2x SSC containing 0.05% SDS at RT and subsequently washed for 5 – 30 min with 0.1x SSC containing 0.1% SDS at 50°C. After washing, the blots were dipped in 2x SSC, sealed in a plastic bag and analyzed via autoradiography using Kodak Biomax-MS films (Kodak, Rochester).

Promotor tests in cell culture

The promotor tests in cell culture required to set up an appropriate expression system. The details on the construction and testing of this system are described in (Heinen 2008). It resulted in the construction of a "Luciflip plasmid" that contains the following elements in the given order: PGK-promoter, ATG, FRT, splice acceptor, double polyA signal, EcoRI site, Kozak, Luc-MYC, intron, polyA signal and was the basis for the Map2k7 alpha reporter assay. For this, fragments spanning -487 to +43 relative to the transcription start of the Map2k7 alpha promoter were amplified from genomic DNA of *M. m. musculus* and *M. m. domesticus* using the primer pair P318 / P319. A two-step PCR strategy was pursued to generate fragments with a deleted insulator motive. Two separate PCRs with the primer pairs P318 / P320 and P321 / P319 were run on top of the cloned promoter fragment. The primers P319 and P320 bind just right upstream and downstream of the insulator sequence and were tailed with a sequence stretch which is homologous to the sequence on the opposite part exactly beyond the insulator. The other primer is one of the primers that were used in the first PCR. Thus, the promoter fragment is divided into two fragments each defined by an inner and an outer primer. The inner edges overlapped, but were lacking the insulator. Both PCR products were cleaned up and included into another PCR without primers. After 5 cycles, the outer primers P318 and P319 were added to the reaction and PCR continued as usual. The resulting product was cloned into a PCR cloning vector and sequenced with M13 primers.

The second version of all 4 fragments has an additional upstream CMV enhancer. For this, the CMV enhancer was amplified with the primers P322 and P323 using the phrGFPII-1 plasmid (Agilent Technologies, Santa Clara) as template. The resulting product was cloned into a PCR cloning vector and validated by M13 primer sequencing. Both CMV primers and the upstream primer that was used for promoter fragment generation (P318) were tailed with an XhoI restriction site overhang. CMV enhancer fragments were retrieved by XhoI digestion and ligated into the XhoI site of all 4 Map2k7 promoter fragments. The assembled fragments were cut out by EcoRI digestion and cloned into the EcoRI site of Luciflip. The orientation of the inserts was controlled with an XhoI digest. All 8 Luci-flip constructs were sequenced with the primers P293, P358 and P199 to control the inserts. No mutations were found.

The 8 expression constructs were transfected into NIH/3T3 cells. For this, NIH/3T3 fibroblast cells were grown in DMEM medium containing sodium pyruvate, non-essential amino acids, L-glutamine penicillin/streptomycin (all from Invitrogen, Carlsbad) and 10% fetal calf serum (PAN, Aidenbach) at 37°C incubation maintaining 5% CO_2 concentration. One day before transfection, 3 x 10^3 NIH/3T3 cells were seeded with 70 µL medium into each well of 96-well plates and grown over night. Cells were co-transfected with Luciflip plasmid and a pGL3 plasmid (Promega, Mannheim) which contains firefly luciferase under the control of SV40 promoter. Firefly luciferase was used to normalize transfection efficiency. Therefore, 0.18 µL Fugene 6 reagent (Roche, Basel) was added to 4.82 µL serum free medium and incubated for 5 min at RT. Subsequently, 30 ng of Luciflip and 30 ng of pGL3 DNA were mixed and added to the medium containing Fugene 6. The mixture was incubated for 20 min and added to one well of the 96 well plate containing NIH/3T3. The transfection was performed for the Luciflip-CMV construct and for an empty Luciflip plasmid as blank control. Each transfection was performed in 8 replicates in parallel. Cells were incubated over night. Next day, firefly and renilla luciferase substrates were applied using the Dual-Glo Luciferase Assay System (Promega, Mannheim) according to the manufacturer’s manual. Relative light units were measured with a Mithras LB 940 Luminometer (Berthold Technologies, Bad Wildbad). For every well, the renilla luciferase signal was divided by the firefly luciferase signal to normalize transfection efficiency. For every construct, median and standard deviation was calculated from the 8 individual replicates.

Construction of knockout mice

The general scheme for the construction of the promotor knockout is shown in suppl. file S2. It consisted two steps. In the first step, in neomycin cassette was inserted at position chr8:4,239,573 (mm10) via homologous recombination in embryonic stem cells, whereby

593 bp of the promotor region, including the start sites, was deleted. These cells were then used to generate transgenic mice via injection into blastocycsts. In the second step, the neomycin cassette was removed via flp recombination at the FRT sites in the mice. The annotated wildtype sequence in this region, as well as the neomycin cassette and the final sequence after recombination are given in suppl file S2. The generation of the KO mice was done by inGenious Targeting Laboratory (iTL), Stony Brook. The mice were then transferred into our facility and backcrossed against C57Bl6/J until final analysis after about 15 generations.

Sperm analysis

Testis and epididymis were dissected from the right side of each mouse and weight was measured. The cauda epididymis was excised, immediately transferred in 250 µl human tubular fluid medium (Millipore, Billerica), punctured with a needle and placed at 37°C, 5% CO_2 for 20 minutes. After 5 min of incubating a 6 µl drop of medium with dispersed spermatozoa was transferred onto a warmed glass slide and covered with a 20 x 20 mm coverslip. Progressive motility was estimated by phase contrast microscopy at a magnification of x 200 according to WHO (World Health Organization, 1992) in a Neubauer chamber. At least 200 sperm cells (but up to 700) in different areas of the slide were counted per animal and the percentage of different classes of motility was calculated per animal.

RNA-Seq and data analysis

The testis tissues of 8 WT and 8 knockout mice were carefully collected and immediately frozen in liquid nitrogen. Total RNA was purified using QIAGEN RNeasy Microarray Tissue Mini Kit (Catalog no. 73304), and prepared using Illumina TruSeq Stranded mRNA HT Library Prep Kit (Catalog no. RS-122–2103), and sequenced using Illumina NextSeq 500 and NextSeq 500/ 550 High Output v2 Kit (150 cycles) (Catalog no. FC-404–2002). Raw reads in FASTQ format were trimmed with Trimmomatic (0.38) (Bolger et al. 2014), and only the reads left in pairs were used for further analysis. The trimmed reads were mapped to the mouse reference genome GRCm39 (Howe et al. 2021, Waterston et al. 2002) with HISAT2 (2.2.1) (Kim et al. 2015) and SAMtools (1.9) (Li et al. 2009), and the mouse gene annotation in Ensembl (Version 104) was used for indexing the genome, i.e., the options “--ss” and “--exon” were used for command “hisat2-build”. The numbers of fragments uniquely mapped to the genes annotated in Ensembl (Version 104) were calculated with featureCounts (2.0.3) (Liao et al. 2014). Principle component analysis on variance stabilizing transformed fragment counts and differential expression analysis on raw counts were performed with DESeq2 (1.30.1) (Love et al. 2014, Zhu et al. 2019).

Primer list

All primers were obtained from Metabion (Martinsried). Sequences are listed 5`> 3`.

P49	AATTAACCCTCACTAAAGGGGAGCATCGAGATTGACCAGA

P50	TAATACGACTCACTATAGGGGCTCGGATGTCATAGTCAGG
P318	CTCGAGTGACCAACTACTTTTCACTATTGCTG
P319	CAAGCTGTGAAGGTCAGTCAGG
P320	TGGTGGACAAGCTGGATCTAGAAAGGAAGAGGAAGCACT
P321	CTCTTCCTTTCTAGATCCAGCTTGTCCACCATGACC
P322	CTCGAGCGCGTTACATAACTTACGGTAAA
P323	CTCGAGCAAAACAAACTCCCATTGACG

Results

Comparison of testis expression of *Map2k7* in two mouse subspecies (*Mus musculus domesticus*, and *Mus musculus musculus*) via Northern blotting and in situ hybridization shows a major difference between the two subspecies (Figure 1). We included in our analysis testis samples from the inbred strain C57Bl6/J (derived from *M. m. domesticus* (Frazer et al. *2007)*), as well as from wild caught mice that were kept under outbreeding conditions (Harr, Karakoc, Neme, Teschke, Pfeifle, Pezer, Babiker, Linnenbrink, Montero, Scavetta, Abai, Molins, Schlegel, Ulrich, Altmuller, Franitza, Buntge, Kunzel and Tautz 2016). Northern blotting revealed a weak 3.5 kb and a strong 1.6 kb band in *M. m. domesticus* (Figure 1A), but only the 3.5kb band in *M. m. musculus*, with the 1.6kb band completely missing. For the inbred strain this confirms the previous observations by (Tournier, Whitmarsh, Cavanagh, Barrett and Davis 1997) and for the wild type strains the observation by (Harr, Voolstra, Heinen, Baines, Rottscheidt, Ihle, Mueller, Bonhomme and Tautz 2006).

To assess the stage of spermatid development at which *Map2k7* is expressed, we used in situ hybridization on testis sections from *M. m. domesticus* and *M. m. musculus*. Testis tissue mainly consists of seminiferous tubules, which are the location of spermatogenesis. Spermatogonial stem cells adjacent to the inner tubule wall divide and form spermatocytes which undergo meiosis. After meiosis the spermatocytes develop into spermatids and change morphologically from round spermatids to elongated spermatids before the generation of mature spermatozoa is completed. The three main stages, spermatogonia, spermatocytes and spermatids, are classified into further substages (Russell et al. 1990). Sperm precursor cells are embedded in Sertoli cells which define the shape of the spermatogenic epithelium and support the germ cells. Through the influence of Sertoli cells, developing sperm precursor cells proceed towards the lumen of seminiferous tubules according to their degree of maturation. Terminal spermiation releases the sperm cells into the luminar fluid of the tubules that transfers them to the epididymis. Hence, ring-shaped zones representing different cell stages can be distinguished in a transverse section of seminiferous tubules.

In situ hybridization results on testis sections using the same probe that had been used for the Northern blotting are shown in Figure 1B. In *M. m. domesticus*, we find a strong signal in post-meiotic spermatid stages. In contrast, *Map2k7* expression pattern in *M. m. musculus* is very weak and becomes only visible after several days of color development. This signal is restricted to earlier stages and we interpret it as the expression of the long transcript. This expression would be expected to be present also in *M. m. domesticus*, but the sections are over-stained after several days of incubation, making it impossible to visualize this directly.

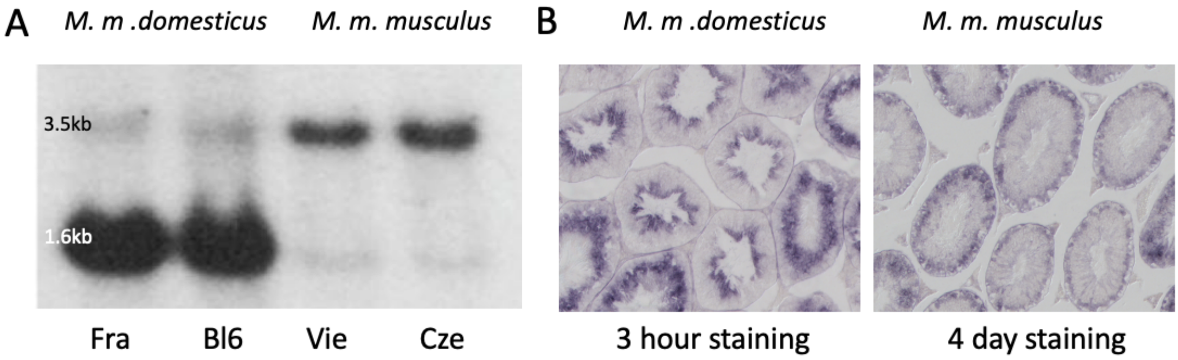


Figure 1: ***Map2k7*** **expression in** ***M. m. domesticus*** **versus** ***M. m. musculus.*** (A) Northern blots with testis RNA from individuals of a wild population from France (Fra), the laboratory inbred strain C57Bl6/J (Bl6), a wild caught population from Vienna (Vie) and a wild caught population from the Czech Republic (Cze). The strongly expressed 1.6kb band is only visible in the two individuals from the *M. m. domesticus* sub-species. (B) In situ hybridization with the *Map2k7* probe on cross sections of seminiferous tubules of an individual from *M. m. domesticus* (left) and an individual from *M. m. musculus* (right). Note that the left signal developed already after 3 hours of color incubation, while the right signal developed only after 4 days of color incubation.

To better resolve the transcript structures and to assess the origin of the new promotor, we made use of the transcriptome data described in (Harr, Karakoc, Neme, Teschke, Pfeifle, Pezer, Babiker, Linnenbrink, Montero, Scavetta, Abai, Molins, Schlegel, Ulrich, Altmuller, Franitza, Buntge, Kunzel and Tautz 2016). Figure 2 shows the read coverage as browser tracks aligned to the annotated versions of the *Map2k7* transcripts in the UCSC browser. Note that there are two versions of the 3.5kb transcript, differing by the inclusion/exclusion of a small exon. Further, there are four versions of the 1.6kb transcript, whereby only two are due to the new promotor, while the other ones are splice variants of the longer transcript (Figure 2 - top panel). The set of browser tracks representing the transcriptomes from 10 different tissues of individuals from a *M. m. domesticus* population show that the new exon (highlighted in grey) is only expressed in the testis (Figure 2 - middle panel). The second set of browser tracks shows testis transcriptomes from different populations of the sub-species, as well as closely related species. The new exon is only present in *M. m. domesticus* populations (DOM), as well as in the sister species *M. spretus* and *M. spicilegus*, but absent in the subspecies *M. m. musculus*, *M. m. castaneus* as well as the further distant mouse species *M. matheyi* and the wood mouse *Apodemus* (Figure 2 - lower panel).

Tournier and colleagues have earlier identified the different isoforms of *Map2k7* by screening testis cDNA clones from laboratory mice and implemented a nomenclature (Tournier et al. 1999). Consistent with the data described above, they found three different 5'-versions (named α, β and γ isoforms) and two 3'-variants (named 1 and 2 isoforms) whose combinations are obtained by alternative splicing. The nomenclature can be complemented by a 3'-variant named 3 that defines the long ~3.5 kb transcript. Northern blotting and qPCR using different exon-specific probes and primers (Suppl. file S1) lead us to the conclusion that the highly expressed testis specific ~1.6 kb fragment corresponds to the transcript that was called Mkk7-α1 in (Tournier, Whitmarsh, Cavanagh, Barrett and Davis 1999) and that we will call Map2k7-α1 in the following to account for the change in the official gene nomenclature.

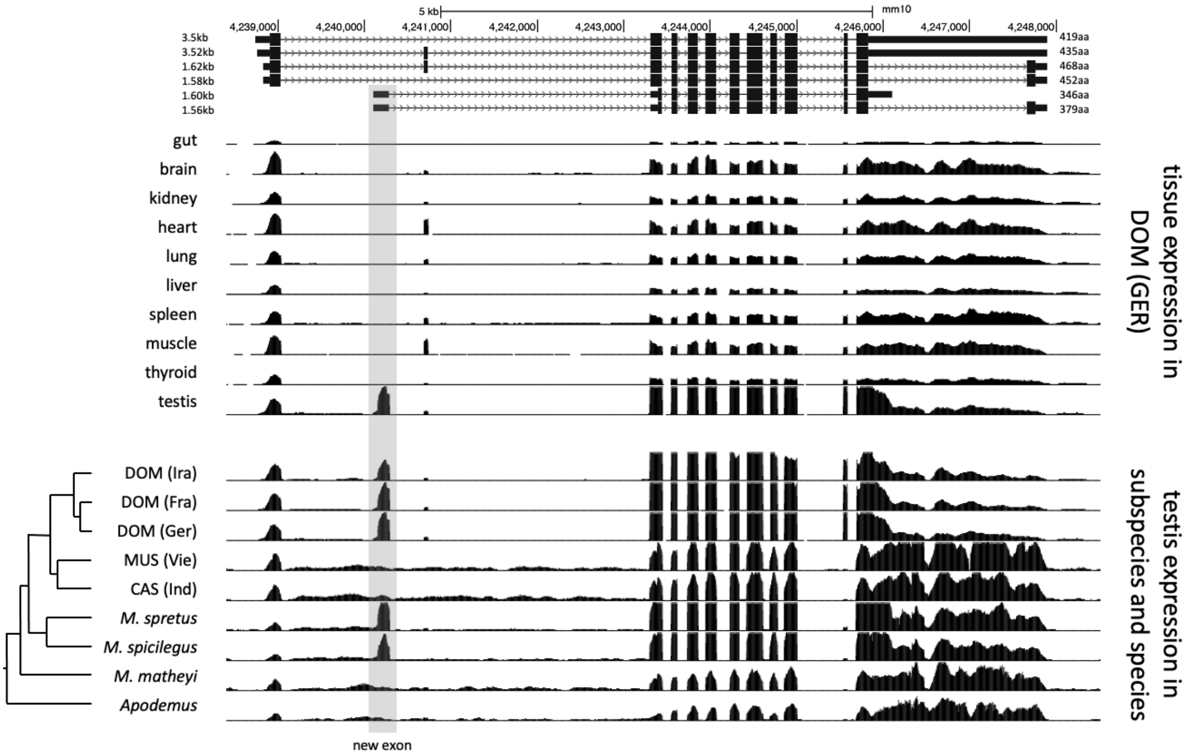


Figure 2: ***Map2k7*** **transcript variants and transcriptome read coverage.** The figure is based on UCSC genome browser tracks. The top panel shows the different transcript and splice versions from the mouse reference genome, which reflects *M. m. domesticus* and therefore includes the new promotor/exon (highlighted by grey shading). The middle panel is based on data from (Harr, Karakoc, Neme, Teschke, Pfeifle, Pezer, Babiker, Linnenbrink, Montero, Scavetta, Abai, Molins, Schlegel, Ulrich, Altmuller, Franitza, Buntge, Kunzel and Tautz 2016) and shows transcriptome read mapping tracks for different tissues. The lower panel is based on transcriptome data from (Harr, Karakoc, Neme, Teschke, Pfeifle, Pezer, Babiker, Linnenbrink, Montero, Scavetta, Abai, Molins, Schlegel, Ulrich, Altmuller, Franitza, Buntge, Kunzel and Tautz 2016) and (Neme and Tautz 2016) and shows transcriptome read mapping tracks for different populations, sub-species and species. The phylogenetic relationships are depicted to the left. DOM: *M. m. domestics*, MUS: *M. m. musculus*, CAS: *M. m. castaneus*.

The transcripts Map2k7-β3 and Map2k7-γ3 (γ includes an additional small exon without an alteration of the reading frame), include the JNK-binding site (D-domains) in the first exon (Figure 3). They are about ~3.5 kb in size and can be found in all analyzed populations and species. The newly evolved transcription start is situated within the first intron of the conserved transcripts. Its transcript does not include the exon with the D-domains, but encodes potentially a protein with the kinase and DVD domain. (Tournier, Whitmarsh, Cavanagh, Barrett and Davis 1999) have shown that this truncated protein has a detectable, but very weak kinase activity when expressed from an expression vector in cell culture. However, the first AUG in the new transcript is before this long ORF and in a different reading frame. It codes for a novel 50aa protein (Figure 3) and the nucleotides surrounding the start codon of this new ORF (UGGCCAACG <u>AUG</u> G) match much better to the Kozak-consensus-sequence (Kozak 1987) than the nucleotides surrounding the start codon of the remaining reading frame of the Map2k7-α1 transcript (CCCCGCCAC AUG C). A purine at position -3 and a guanine at position +4 are the most important sequence elements for the initiation of translation. It is therefore questionable whether the shortened form of *Map2k7* is translated at all under natural conditions.

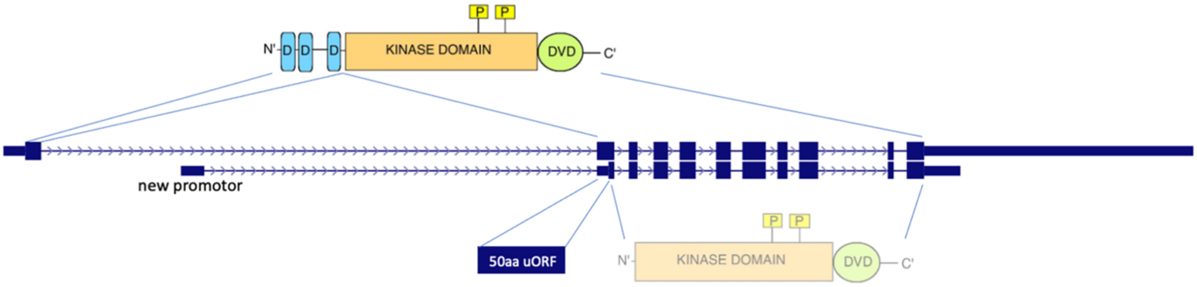


Figure 3: Comparison of the conserved and the new transcripts and their coding potentials. Transcript depictions are taken from the UCSC browser annotations (see also Figure 2), whereby only the two main transcripts are shown, the conserved one (upper) and the new one starting from the new promotor (lower). The protein domains of the *Map2k7* functional kinase are depicted on the top, including the JNK-binding sites (D-domains in blue), the kinase domain (in orange) with its functional phosphorylation sites (yellow), as well as the C-terminal DVD domain (green). The first AUG in the new transcript is in the second exon and it would result in a 50aa protein, which represents a *de novo* generated protein without known domain. If this AUG would not be used, the next AUG in a different frame could potentially lead to the expression of a truncated version of the *Map2k7* protein, containing only the kinase and the DVD domains.

Based on the phylogenetic tree of wild mice (Chevret et al. 2005, Guenet and Bonhomme 2003), we can infer that the new 1.6 kb transcript has arisen before the branching of *M. spicilegus* and *M. spretus* at least 2 million years ago, but not more than 6 million years ago, since it is not present in the outgroups *M. matheyi* and *Apodemus* (Figure 2). Note that while these outgroups show some general level of intron transcription in the respective area, this does not reflect the clear exon structure that is seen in the other species. The new promotor has secondarily disappeared in the lineage of *M. m. musculus* and *M. m. castaneus*, apparently because of a crucial T/A substitution (see below).

Promotor analysis

To further understand the cis regulation of the spermatid specific 1.6 kb Map2k7α1 transcript, we analyzed genomic sequences in a 500 bp window upstream of the transcription start site in different wild populations, subspecies and species of *M. m. domesticus, M. m. musculus, M. m. castaneus, M. spretus and M. spicilegus* (Figure 4). Only one SNP at -84 bp (with respect to the start site of transcription - see also legend Figure 4 for alternative start sites) is correlated with the expression of the spermatid specific isoform in the different populations. An adenine is found in populations with expression of the new transcript, whereas thymine is present in those without expression (Figure 4).

Empirical data provide evidence that specific transcription activity in reproductive tissues and particularly in spermatogenesis is regulated by very short proximal promoters (Blaise et al. 2001, Han et al. 2004, Li et al. 1998, Reddi et al. 1999, Scieglinska et al. 2004, Topaloglu et al. 2001, Zambrowicz et al. 1993). Those studies demonstrated that proximal promoters shorter than 300bp, or even less than 100bp, are sufficient to drive spermatid specific expression in mice. Unique mechanisms of gene regulation are postulated to exist in post-meiotic cells (Acharya et al. 2006, Somboonthum et al. 2005). Additionally, it was shown that a 5'-CACACA motive ~170 bp upstream of the transcription start serves as an insulator in the spermatid specific expression of the SP-10 gene (acrosomal vesicle protein 1 - Acrv1) and it was suggested that insulators might generally play an important role in maintaining spermatid specific transcription (Abhyankar et al. 2007, Acharya, Govind, Shore, Stoler and Reddi 2006, Reddi et al. 2003, Reddi et al. 2007).

Such a 5'- CACACA motif can be found at around -259 base pairs upstream the transcription start site of the testis specific Map2k7-α1 RNA (Figure 4). These considerations raise the question, whether a short sequence carrying the -84 A mutation in combination with the -259 5'- CACACA motif would meet the requirements to serve as a testis specific promoter. Therefore, we tested a fragment representing the genomic sequence between -487/+43 of the new *Map2k7* testis promoter in cell culture-based luciferase expression assays (Figure 4).

The expression of most interest in this context is restricted to late spermatids. Culturing this type of cells is very difficult due to its haploid post meiotic stage with condensed chromatin. A well-established spermatid cell culture model is not available and alternative cell lines have the disadvantage that they will most likely not recognize the spermatid specific *Map2k7* promotor. In the absence of better options, we have chosen the widely used NIH/3T3 fibroblast cell line for this experiment.

It cannot be expected that the tested fragment is sufficient to drive luciferase expression in non-spermatid cells, but it is likely that the expression level can be raised by deleting the 5'-CACACA motive at -259, if the assumption is correct, that this sequence maintains spermatid specific transcription by acting as an insulator in other cells. Thus, -487/+43 fragments lacking the 5'-CACACA motive at -259 were generated as well. It can be assumed that the -487/+43 fragments do not contain enhancer elements which promote expression in fibroblasts. Therefore, a CMV enhancer was ligated upstream to both versions. All four constructs (wild type, wild type with deleted insulator, CMV enhancer + wild type, CMV enhancer + wild type with deleted insulator) were created as *M. m. domesticus* variants with an adenine at position -84 and as *M. m. musculus* variants with a thymine at position -84. We find that the *M. m. domesticus* variant generates significantly higher signals compared to *M. m. musculus* in every combination. The different replicates are consistent, indicated by relatively small standard deviations (Figure 4). Deletion of the 5'-CACACA motive at -259 indeed increases the expression strength in both variants. The presence of an enhancer potentiates the effects as expected. These data provide strong evidence that the adenine at position -84 enhances the activity of the basal promotor. For this reason, it can be supposed that a major contribution of this mutation to the expression difference between *M. m. domesticus* and *M. m. musculus* in late spermatids is likely. The sequence 5'-CACACA represses the action of an adjacent enhancer to a certain extent. This finding supports the hypothesis that it acts as an insulator in the spermatid specific *Map2k7* promotor.

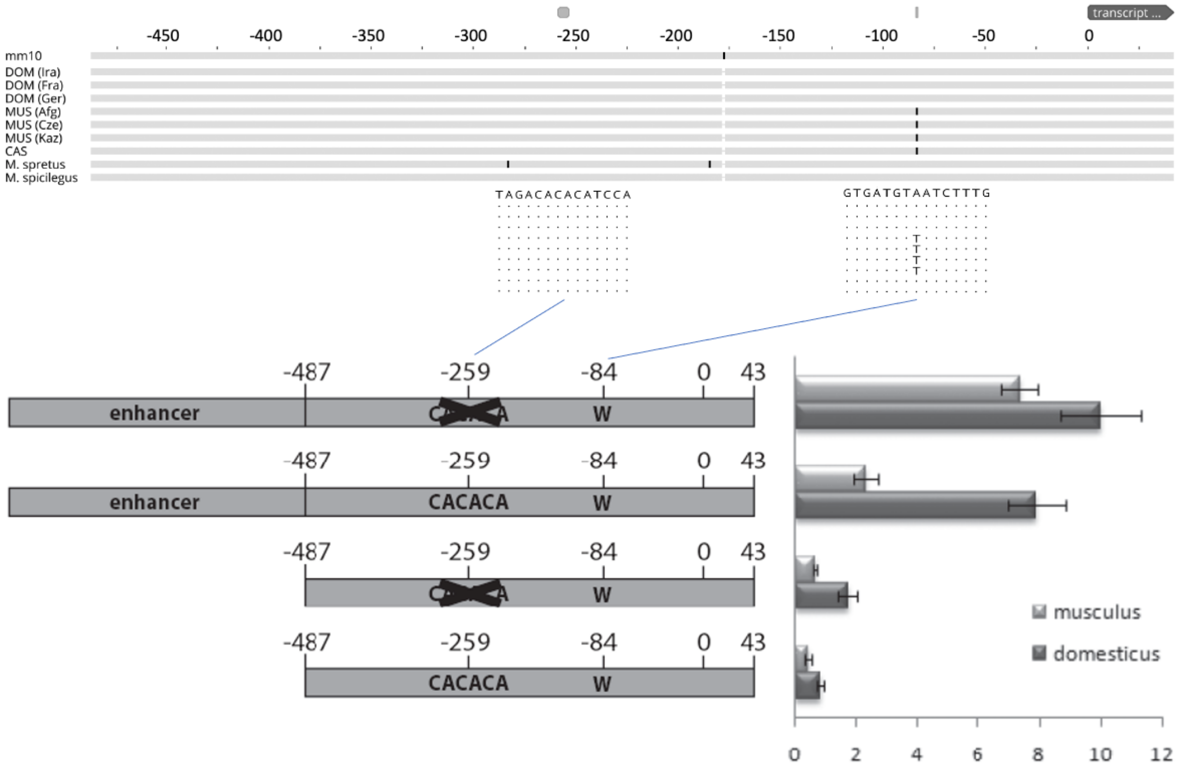


Figure 4: Functional test of the promotor region driving the new transcript. The top shows the alignment of genomic sequences from populations, sub-species and species based on the genome data from (Harr, Karakoc, Neme, Teschke, Pfeifle, Pezer, Babiker, Linnenbrink, Montero, Scavetta, Abai, Molins, Schlegel, Ulrich, Altmuller, Franitza, Buntge, Kunzel and Tautz 2016), aligned to the mouse mm10 reference sequence. The fragment shown represents the one used for the promotor studies - only replacements with respect to the reference are marked, the two relevant regions discussed in the text are enlarged with their respective sequences. DOM represents *M. m. domesticus* populations, MUS represents *M. m. musculus* populations, CAS represents a *M. m. castaneus* population. T/A is the only mutation that correlates with the expression of the new promotor. Note that the transcriptional start site, marked as "0", is located between the annotated site for this transcript, which would be 31bp further upstream and the site from which the bulk of the transcripts generated in the RNASeq experiment starts, which is 8bp downstream (see suppl. file S2 for a corresponding sequence depiction of this region). The bottom shows the scheme of the four constructs that were tested in cell culture, as well as their expression levels measured as fluorescence intensity (see Methods). Error bars indicate standard deviations from eight replicates each.

Functional analysis

To assess a possible functional role of the new spermatid specific *Map2k7* promotor, a knock-out mouse was generated in which the promotor was deleted in the *M. m. domesticus* background (see Methods and targeting strategy in suppl. file S2). The knock-out was designed in a way that it should not interfere with the conserved *Map2k7* transcript.

The knockout animals were fully viable and showed no overt phenotype. KO animals are on average a bit heavier, but have lower normalized testis weights (Table 1). Given the specific expression of the new transcripts during a crucial phase of sperm maturation, we assessed also sperm motility phenotypes. KO animals have fewer motile sperm and fewer progressive sperm (Table 1). All differences are significant at p< 0.05 (t-test, 2 sided).

Table 1. Sperm phenotypes

genotype		mouse weight (g)	normalized testis weight (g)	motile sperm (%)	progressive sperm (%)
WT (N=10)	averages:	25.14	0.20	48	26
	SD:	*1.42*	*0.03*	*6*	*5*
KO (N=18)	averages:	27.59	0.17	42	22
	SD:	*2.07*	*0.01*	*7*	*5*
	P-values:	0.003	0.001	0.021	0.031

RNASeq analysis

A comparative RNASeq analysis with RNA from knock-out mice versus wild type mice was used to find out whether the 1.6 kb Map2k7-α1 testis specific transcript is influencing the expression of other genes. The RNA was collected from three different tissues of the male reproductive organs, the testis, the caput epididymis and the cauda with eight biological replicates each. The testis is the place of the primary sperm production. The sperm from the testis move through the caput epididymis where they mature and are eventually stored in the cauda. While the chromatin of post meiotic sperm is condensed, there is still some transcriptome turnover (Ren et al. 2017) and the epididymal cells contribute to this transcriptome turnover as well (Shi et al. 2021).

The overall analysis of the transcriptome data in the PCA analysis shows that the samples from each of the three tissues are very different, implying that there is indeed a major turn-over of RNA between these stages, either due to differential stability, or new transcription. On the other hand, differences between wild type and knockout are much smaller (Figure 5A). Still, since we used eight replicates for each tissue, we have a very high sensitivity to detect even small transcriptome changes (Xie et al. 2020). Accordingly, we find thousands of genes with significant expression differences (i.e., p_{adj} values <0.05 in the DSeq2 analysis), but mostly with relatively low log2fold-changes (Figure 5B-D). Interestingly, however, the cauda samples show a set of genes with very high positive log2fold changes (see set of dots on the top of the panel for "cauda in Figure 5D).

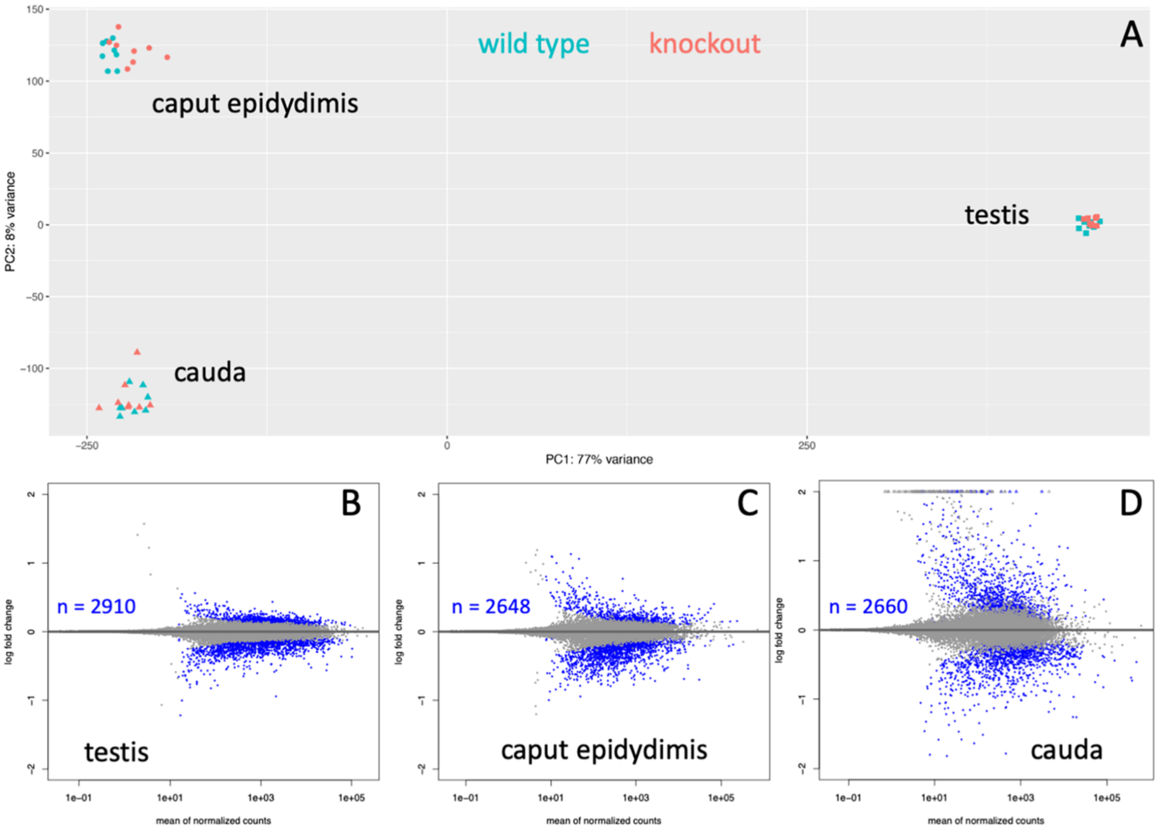


Figure 5: Whole transcriptomes analysis of three tissues from wild type versus knockout mice. (A) Overall PCA comparison. Strong differentiation is seen between the tissue samples, implying that that transcript sets are very different. The differences between wild type and knouts are much smaller for each tissue. (B-D) Significantly differentially expressed genes. Each gene is represented by a dot, genes with $p_{adj} < 0.05$ values are plotted as blue dots. The number of significantly differentially expressed genes is provided as inset for each tissue.

First, we asked whether kinase signaling processes are specifically affected, which would suggest that the shortened protein of the Map2k7-α1 transcript could be involved in signaling. However, the top biological process GO terms among the significant genes do not include "kinase signaling", "signal transduction" or "Jnk cascade" in either of the data sets (based on a GO analysis with Panther (Mi et al. 2013) - see suppl Table 1). Instead, the top GO terms indicate an involvement in meiotic division and chromosome segregation for testis, an involvement in extracellular matrix organization for the caput epididymis and an involvement in peptide biosynthetic processes for the cauda (suppl Table 1b, d, f). Hence, it is unlikely that the primary function of the Map2k7-α1 transcript is related to residual kinase signaling activity.

To get a further insight into the functional changes in the knockout animals, we focused on the genes that are most highly expressed in the respective tissues (based on the length-normalized baseMean counts of the RNASeq data), since small concentration changes in such genes could have a more marked influence on the phenotype.

For most of the highly expressed genes that we identify in the significant gene lists for testis and cauda epididymis, one can retrieve functional information from knockout experiments in mice and almost all of these find an effect on sperm maturation and/or sperm mobility (Table 2). Hence, while the relative expression changes are not large, it is well possible that the effect of the Map2k7-α1 transcript knockout is mediated via these genes. Interestingly, for the cauda, we find a rather different pattern. Most of the top expressed genes in the list are not specific for the cauda, but are more broadly expressed (e.g., an enzyme, actin and a ribosomal protein). Interestingly, several code for immunity proteins, including sperm-specific antimicrobial peptides. Another major difference in the cauda transcriptome is a set of 37 genes with very high log2fold changes, i.e., expression at a much higher level in the knockout than in the wild type. Intriguingly, most of these are generally expressed motor proteins and the role of such proteins in spermiogenesis has only recently been fully recognized (Wu et al. 2021).

Table 2. List of top significant genes in the RNASeq analysis.

tissue	gene name	description	log2 fold-change	padj	expression	function	literature
testis (top 10 most highly expressed genes that are specifically expressed in testis)#1							
	Tnp2	Nuclear transition protein 2	-0.37	0.0021	testis	conversion of nucleosomal chromatin in spermatids, involved in sperm motility	(Adham et al. 2001)
	Smcp	Sperm mitochondrial-associated cysteine-rich protein	-0.25	0.0006	testis	involved in sperm motility and fertilization	(Nayernia et al. 2002)
	Gapdhs	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic	-0.12	0.0206	testis	required for normal sperm motility and male fertility,	(Huang et al. 2017)
	Fhl4	Four and a half LIM domains 4	0.09	0.0186	testis	may affect sperm maturation and morphology	(Lardenois et al. 2009)
	Tuba3a	tubulin, alpha 3A, spermatogenic	-0.14	0.0027	testis	required for the production of normal spermatozoa.	(Akter et al. 2021)
	Crisp2	Cysteine-rich secretory protein 2	0.10	0.0444	testis	regulates calcium fluxes during sperm capacitation, essential for fertility	(Curci et al. 2020)
	Akap4	A-kinase anchor protein 4	0.13	0.0037	testis	major structural component of sperm fibrous sheath, plays a role in sperm motility	(Zhang et al. 2021)
	Tuba3b	tubulin, alpha 3B, spermatogenic	-0.11	0.0036	testis	required for the production of normal spermatozoa.	(Akter, Hada, Shikata, Watanabe, Ogura and Matoba 2021)
	Cox8c	cytochrome c oxidase subunit 8C	-0.26	0.0000	testis	required for the production of normal spermatozoa.	(Akter, Hada, Shikata, Watanabe, Ogura and Matoba 2021)
	H1fnt	testis specific H1.7 linker histone	-0.15	0.0013	testis	required for cell restructuring and DNA condensation during spermiogenesis	(Martianov et al. 2005)
caput epididymis (top 10 most highly expressed genes that are specifically expressed in testis/epididymis) #2							
	Lcn8	Epididymal-specific lipocalin-8	0.15	0.0493	epididymis	involved in sperm maturation and motility	(Wen et al. 2021)
	Lcn9	lipocalin 9	0.14	0.0450	epididymis	may act redundantly to Lcn8	(Wen, Liu, Zhu, Sun, Xiao, Lin, Zhang, Ye and Gao 2021)
	Epp13	Epididymal protein 13	-0.13	0.0378	epididymis	no fecundity defect in knockout mice	(Noda et al. 2019)
	Cst8	cystatin 8 (cystatin-related epididymal spermatogenic)	0.24	0.0016	epididymis	capacitation of spermatozoa	(Chau and Cornwall 2011)
	Rnase10	Ribonuclease-like protein 10	0.40	0.0000	epididymis	required for post-testicular sperm maturation	(Krutskikh et al. 2012)
	Adam7	Disintegrin and metalloproteinase domain-containing protein 7	0.15	0.0108	epididymis	required for epididymal integrity, sperm morphology and motility	(Choi et al. 2015)
	Teddm1a	Transmembrane epididymal family member 1a	0.28	0.0035	epididymis	no information on sperm effects	
	Teddm2	Transmembrane epididymal family member 2	0.34	0.0001	epididymis	no information on sperm effects	
	Spink12	serine peptidase inhibitor, Kazal type 12	-0.31	0.0003	epididymis	part of Spink family with redundant functions in sperm maturation	(Jeong et al. 2019)
	Defb12	defensin beta 12	0.20	0.0033	epididymis	immunity protein	

cauda (top 10 most highly expressed genes)[#3]							
	Crisp1	Cysteine-rich secretory protein 1	-0.46	0.0013	testis, epididymis	required to optimize sperm flagellum waveform	(Gaikwad et al. 2021)
	Cd52	CAMPATH-1 antigen	-0.69	0.0005	broad	immunity protein	(Yamaguchi et al. 2008)
	Gpx3	Glutathione peroxidase 3	-0.73	0.0043	broad	protects cells from oxidative damage, involved in the maturation of sperm cells	(Noblanc et al. 2011)
	Defb28	Defensin beta 28	-0.37	0.0404	testis, epididymis	immunity protein	
	Spink8	Serine protease inhibitor Kazal-type 8	-0.44	0.0009	testis, epididymis	part of Spink family with redundant functions in sperm maturation	(Jeong, Lee, Kim, Hong, Kim, Choi, Cho and Cho 2019)
	B2m	Beta-2-microglobulin	-0.39	0.0035	broad	component of the class I major histocompatibility complex (MHC)	
	Wfdc15b	WAP four-disulfide core domain protein 15B	-0.75	0.0002	testis, epididymis	Wfdc family members act in innate immune responses during epididymitis	(Andrade et al. 2021)
	Defb2	defensin beta 2	-0.38	0.0336	testis, epididymis	immunity protein	
	Acta2	Actin, aortic smooth muscle	0.44	0.0409	broad	cell motility, may affect contractability of seminiferous tubules	(Uchida et al. 2020)
	Rplp1	ribosomal protein, large, P1	-0.54	0.0110	broad	ribosomal protein	
cauda (top 10 genes with highest log2fold changes)[#3]							
	Pvalb	parvalbumin	8.93	0.0158	broad	no information on sperm effects	
	Myl2	myosin, light polypeptide 2, regulatory, cardiac, slow	8.36	0.0338	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Myh7	myosin, heavy polypeptide 7, cardiac muscle, beta	8.04	0.0405	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Myl1	myosin, light polypeptide 1	7.26	0.0170	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Atp2a1	ATPase, Ca++ transporting, cardiac muscle, fast twitch 1	7.05	0.0008	broad	Ca-gradient regulation, male fertility	(Prasad et al. 2004)
	Myh1	myosin, heavy polypeptide 1, skeletal muscle, adult	6.73	0.0404	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Actn2	actinin alpha 2	6.40	0.0380	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Sln	sarcolipin	6.01	0.0307	broad	no information on sperm effects	
	Tnnt3	troponin T3, skeletal, fast	5.86	0.0050	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)
	Mylk4	myosin light chain kinase family, member 4	5.85	0.0138	broad	motorprotein involved in spermiogenesis	(Wu et al. 2021)

[#1] selected from the top 17 genes in the whole list (see suppl. Table1b)

[#2] selected from the top 27 genes in the whole list (see suppl. Table1d)

[#3] represent the top 10 genes from the list (see suppl. Table1f)

Discussion

Based on comparative genomic and functional analysis, we show here that a new intra-intronic promotor has arisen in the mouse lineage 2-6 million years ago and has led to the evolution of a functionally new transcript within an otherwise highly conserved gene. The transcript is specific to the testis and knockout combined with transcriptome analysis shows that it is functionally involved in sperm maturation. Interestingly, it got also secondarily lost in some mouse lineages, apparently due to the acquisition of a disabling mutation in the promotor region.

The emergence of evolutionary novelties out of regulatory changes is by now well documented in many species (see (Carroll 2008, He et al. 2021, Mattioli et al. 2020, Osada et al. 2017, Romero et al. 2012, Signor and Nuzhdin 2018, Tautz 2000, Wray 2007) for a subset of relevant papers and reviews). In fact, it is so abundant that it constitutes often the first measurable differences in population and species divergence, including diverging mouse populations (Bryk et al. 2013), which raises the question whether much of is initially neutrally evolving or could be functional (Fay and Wittkopp 2008, Hill et al. 2021, Hodgins-Davis et al. 2019, Staubach et al. 2010). Given the evolutionary volatility of the Map2k7-α1 transcript, with its fast secondary loss after its initial emergence, one would normally have considered it to be mostly neutral and therefore subject to random fixation or loss. However, our data show that it has a clear functional role in spermatogenesis. The high evolutionary dynamics of this transcript is therefore more likely explained by the general effects of sexual selection that would be particularly effective in the germline and the gonads (Kleene 2005).

Possible function of the Map2k7-α1 transcript

There are several possibilities of how the Map2k7-α1 transcript could function. The first is that it leads to the translation of a truncated protein that codes only for *Map2k7* kinase, but does not bind specifically to JNK. It could therefore phosphorylate other signaling proteins, but in an unspecific manner. This would likely be detrimental, rather than advantageous for the cells. Also, since we do not find GO terms that relate to signaling processes in the transcriptome analysis of knockout mice, we assume that the truncated protein is not expressed, or at least not funtional.

The second possibility is that Map2k7-α1 acts as a non-coding RNA. There are multiple ways of how non-coding RNAs can regulate other genes or gene complexes (Gil and Ulitsky 2020, Statello et al. 2021) and some have been implicated in male infertility (Joshi and Rajender 2020). In a previous study, we identified a testis specific new transcript that has also emerged via a new promotor acquisition and for which we could infer that it acts as lncRNA in spermiogenesis (Heinen et al. 2009). However, given that most of the Map2k7-α1 RNA overlaps with the functional *Map2k7* transcripts, it would seem unlikely that it could have assumed such a function as non-coding RNA as a whole, since most of its RNA is actually potentially coding. Only the new exon that emerged out of intronic sequences might have such a function.

On the other hand, the first AUG in this new transcript is embedded in an optimal Kozak-consensus-sequence (Kozak 1987) and one would therefore expect that it leads to the translation of a 50aa ORF. The resulting peptide does not match with any other protein or domain in the data bases, since it is actually produced out of a previously non-coding intron sequence.

While it has long been thought that proteins that emerge out of such more or less random sequences would not be functional, it has by now become clear that the *de novo* evolution of proteins is well possible (Tautz and Domazet-Loso 2011, Van Oss and Carvunis 2019). In fact, we have recently described such a case of a very recent *de novo* emergence of a protein that regulates pregnancy cycles in mice (Xie et al. 2019). In that case, we could proof a direct function of the protein in a knockout mouse which carried only a frameshift mutation in the protein. In the current study, we have deleted the whole transcript, implying that it is not fully proven that it is really the translated peptide that conveys the function. But studies with random peptide sequences in *E. coli* and in plants have shown that a substantial fraction of them can have a direct positive effect on their hosts (Bao et al. 2017, Neme et al. 2017). It seems therefore possible that the new peptide that is encoded by the Map2k7-α1 transcript is indeed a *de novo* protein with a function. Hence, this would be a case where a pre-existing potentially functional sequence was "waiting" for a promotor emergence to allow it to become functional.

The ORF is actually already present in the outgroup species, but would not expected to be translated in these species. The fact that it got secondarily lost in the *M. m. musculus* subspecies is in line with the observation of fast gain and loss cycles of *de novo* evolved transcripts and proteins (Neme and Tautz 2014, Palmieri et al. 2014). But our functional data show that despite of this evolutionary instability, the new transcript (and/or peptide) can still be functional.

Acknowledgements

We thank Heike Harre for help with the sperm counting and phenotyping, Cornelia Burghardt for RNA extraction and the mouse team lead by Christine Pfeifle for expert support. This work was funded by institutional funds of the MPG to DT.

Author contributions

T.H.: conceptualization, experimental work, data analysis, paper writing, C.X.: data analysis, M.K.: data analysis, D.S.: experimental work, data analysis, S.K.: experimental work, data analysis, D.T.: conceptualization, data analysis, paper writing. Part of this work was done in the framework of the PhD thesis of the first author (Heinen 2008).

References

Abhyankar MM, Urekar C, Reddi PP. 2007. A novel CpG-free vertebrate insulator silences the testis-specific SP-10 gene in somatic tissues. Journal of Biological Chemistry. Dec;282:36143-36154.

Acharya KK, Govind CK, Shore AN, Stoler MH, Reddi PP. 2006. cis-requirement for the maintenance of round spermatid-specific transcription. Developmental Biology. Jul;295:781-790.

Adham IM, Nayernia K, Burkhardt-Gottges E, Topaloglu O, Dixkens C, Holstein AF, Engle W. 2001. Teratozoospermia in mice lacking the transition protein 2 (Tnp2). Molecular Human Reproduction. Jun;7:513-520.

Akter MS, Hada M, Shikata D, Watanabe G, Ogura A, Matoba S. 2021. CRISPR/Cas9-based genetic screen of SCNT-reprogramming resistant genes identifies critical genes for male germ cell development in mice. Scientific Reports. Jul;11.

Alwine JC, Kemp DJ, Stark GR. 1977. METHOD FOR DETECTION OF SPECIFIC RNAS IN AGAROSE GELS BY TRANSFER TO DIAZOBENZYLOXYMETHYL-PAPER AND HYBRIDIZATION WITH DNA PROBES. Proceedings of the National Academy of Sciences of the United States of America.74:5350-5354.

Andrade AD, Almeida PGC, Mariani NAP, Freitas GA, Kushima H, Filadelpho AL, Spadella MA, Avellar MCW, Silva EJR. 2021. Lipopolysaccharide-induced epididymitis modifies the transcriptional profile of Wfdc genes in mice. Biology of Reproduction. Jan;104:144-158.

Bao ZL, Clancy MA, Carvalho RF, Elliott K, Folta KM. 2017. Identification of Novel Growth Regulators in Plant Populations Expressing Random Peptides. Plant Physiology. Oct;175:619-627.

Blaise R, Guillaudeux T, Tavernier G, Daegelen D, Evrard B, Mairal A, Holm C, Jegou B, Langin D. 2001. Testis hormone-sensitive lipase expression in spermatids is governed by a short promoter in transgenic mice. Journal of Biological Chemistry. Feb;276:5109-5115.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. Aug 1;30:2114-2120.

Bryk J, Somel M, Lorenc A, Teschke M. 2013. Early gene expression divergence between allopatric populations of the house mouse (Mus musculus domesticus). Ecology and Evolution. Mar;3:558-568.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. Cell. Jul;134:25-36.

Chang LF, Jones Y, Ellisman MH, Goldstein LSB, Karin M. 2003. JNK1 is required for maintenance of neuronal microtubules and controls phosphorylation of microtubule-associated proteins. Developmental Cell. Apr;4:521-533.

Chang LF, Karin M. 2001. Mammalian MAP kinase signalling cascades. Nature. Mar;410:37-40.

Chau KM, Cornwall GA. 2011. Reduced Fertility In Vitro in Mice Lacking the Cystatin CRES (Cystatin-Related Epididymal Spermatogenic): Rescue by Exposure of Spermatozoa to Dibutyryl cAMP and Isobutylmethylxanthine. Biology of Reproduction. Jan;84:140-152.

Chevret P, Veyrunes F, Britton-Davidian J. 2005. Molecular phylogeny of the genus Mus (Rodentia : Murinae) based on mitochondrial and nuclear data. Biological Journal of the Linnean Society. Mar;84:417-427.

Choi H, Han C, Jin S, Kwon JT, Kim J, Jeong J, Ham S, Jeon S, Yoo YJ, Cho C. 2015. Reduced Fertility and Altered Epididymal and Sperm Integrity in Mice Lacking ADAM7. Biology of Reproduction. Sep;93.

Curci L, Brukman NG, Munoz MW, Rojo D, Carvajal G, Sulzyk V, Gonzalez SN, Rubinstein M, Da Ros VG, Cuasnicu PS. 2020. Functional redundancy and compensation: Deletion of multiple murineCrispgenes reveals their essential role for male fertility. Faseb Journal. Dec;34:15718-15733.

Deng YB, Ren XY, Yang L, Lin YH, Wu XW. 2003. A JNK-dependent pathway is required for TNF alpha-induced apoptosis. Cell. Oct;115:61-70.

Dong C, Yang DD, Tournier C, Whitmarsh AJ, Xu J, Davis RJ, Flavell RA. 2000. JNK is required for effector T-cell function but not for T-cell activation. Nature. May;405:91-94.

English J, Pearson G, Wilsbacher J, Swantek J, Karandikar M, Xu SC, Cobb MH. 1999. New insights into the control of MAP kinase pathways. Experimental Cell Research. Nov;253:255-270.

Fay JC, Wittkopp PJ. 2008. Evaluating the role of natural selection in the evolution of gene regulation. Heredity. Feb;100:191-199.

Fleming Y, Armstrong CG, Morrice N, Paterson A, Goedert M, Cohen P. 2000. Synergistic activation of stress-activated protein kinase 1/c-Jun N-terminal kinase (SAPK1/JNK) isoforms by mitogen-activated protein kinase kinase 4 (MKK4) and MKK7. Biochemical Journal. Nov;352:145-154.

Foltz IN, Gerl RE, Wieler JS, Luckach M, Salmon RA, Schrader JW. 1998. Human mitogen-activated protein kinase kinase 7 (MKK7) is a highly conserved c-jun N-terminal kinase/stress-activated protein kinase (JNK/SAPK) activated by environmental stresses and physiological stimuli. Journal of Biological Chemistry. Apr;273:9344-9351.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. Aug;448:1050-U1058.

Gaikwad AS, Nandagiri A, Potter DL, Nosrati R, O'Connor AE, Jadhav S, Soria J, Prabhakar R, O'Bryan MK. 2021. CRISPs Function to Boost Sperm Power Output and Motility. Frontiers in Cell and Developmental Biology. Aug;9.

Gil N, Ulitsky I. 2020. Regulation of gene expression by cis-acting long non-coding RNAs. Nature Reviews Genetics. Feb;21:102-117.

Guenet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends in Genetics. Jan;19:24-31.

Han SY, Xie WS, Kim SH, Yue LM, DeJong J. 2004. A short core promoter drives expression of the ALF transcription factor in reproductive tissues of male and female mice. Biology of Reproduction. Sep;71:933-941.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Z, Babiker H, Linnenbrink M, Montero I, Scavetta R, et al. 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Scientific Data. Sep;3.

Harr B, Voolstra C, Heinen TJAJ, Baines JF, Rottscheidt R, Ihle S, Mueller W, Bonhomme F, Tautz D. 2006. A change of expression in the conserved signaling gene MKK7 is associated with a selective sweep in the western house mouse Mus musculus domesticus. Journal of Evolutionary Biology. SEP 2006;19:1486-1496.

He F, Steige KA, Kovacova V, Gobel U, Bouzid M, Keightley PD, Beyer A, de Meaux J. 2021. Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. Nature Communications. Jun;12.

Heinen TJAJ. 2008. Characterization of genes involved in recent adaptation. Universität zu Köln.

Heinen TJAJ, Staubach F, Haeming D, Tautz D. 2009. Emergence of a New Gene from an Intergenic Region. Current Biology. SEP 29 2009;19:1527-1531.

Hill MS, Vande Zande P, Wittkopp PJ. 2021. Molecular and evolutionary processes generating variation in gene expression. Nature Reviews Genetics. Apr;22:203-215.

Hodgins-Davis A, Duveau F, Walker EA, Wittkopp PJ. 2019. Empirical measures of mutational effects define neutral models of regulatory evolution in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America. Oct;116:21085-21093.

Holland PM, Suzanne M, Campbell JS, Noselli S, Cooper JA. 1997. MKK7 is a stress-activated mitogen-activated protein kinase kinase functionally related to hemipterous. Journal of Biological Chemistry. Oct;272:24994-24998.

Howe KL, Achuthan P, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, et al. 2021. Ensembl 2021. Nucleic Acids Research. Jan;49:D884-D891.

Huang ZH, Danshina PV, Mohr K, Qu WD, Goodson SG, O'Connell TM, O'Brien DA. 2017. Sperm function, protein phosphorylation, and metabolism differ in mice lacking successive sperm-specific glycolytic enzymes. Biology of Reproduction. Oct;97:586-597.

Jeong J, Lee B, Kim J, Hong SH, Kim D, Choi S, Cho BN, Cho C. 2019. Expressional and functional analyses of epididymal SPINKs in mice. Gene Expression Patterns. Jan;31:18-25.

Joshi M, Rajender S. 2020. Long non-coding RNAs (lncRNAs) in spermatogenesis and male infertility. Reproductive Biology and Endocrinology. Oct;18.

Kim D, Landmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. Apr;12:357-U121.

Kishimoto H, Nakagawa K, Watanabe T, Kitagawa D, Momose H, Seo J, Nishitai G, Shimizu N, Ohata S, Tanemura S, et al. 2003. Different properties of SEK1 and MKK7 in dual phosphorylation of stress-induced activated protein kinase SAPK/JNK in embryonic stem cells. Journal of Biological Chemistry. May;278:16595-16601.

Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Developmental Biology. Jan;277:16-26.

Kozak M. 1987. AN ANALYSIS OF 5'-NONCODING SEQUENCES FROM 699 VERTEBRATE MESSENGER-RNAS. Nucleic Acids Research. Oct;15:8125-8148.

Krutskikh A, Poliandri A, Cabrera-Sharp V, Dacheux JL, Poutanen M, Huhtaniemi I. 2012. Epididymal protein Rnase10 is required for post-testicular sperm maturation and male fertility. Faseb Journal. Oct;26:4198-4209.

Lardenois A, Chalmel F, Demougin P, Kotaja N, Sassone-Corsi P, Primig M. 2009. Fhl5/Act, a CREM-binding transcriptional activator required for normal sperm maturation and morphology, is not essential for testicular gene expression. Reproductive Biology and Endocrinology. Nov;7.

Lei K, Nimnual A, Zong WX, Kennedy NJ, Flavell RA, Thompson CB, Bar-Sagi D, Davis RJ. 2002. The Bax subfamily of Bcl2-related proteins is essential for apoptotic signal transduction by c-Jun NH2-terminal kinase. Molecular and Cellular Biology. Jul;22:4929-4942.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data P. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Aug;25:2078-2079.

Li SM, Zhou WT, Doglio L, Goldberg E. 1998. Transgenic mice demonstrate a testis-specific promoter for lactate dehydrogenase, LDHC. Journal of Biological Chemistry. Nov;273:31191-31194.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. Apr 1;30:923-930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology.15.

Martianov I, Brancorsini S, Catena R, Gansmuller A, Kotaja N, Parvinen M, Sassone-Corsi P, Davidson I. 2005. Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. Proceedings of the National Academy of Sciences of the United States of America. Feb;102:2808-2813.

Mattioli K, Oliveros W, Gerhardinger C, Andergassen D, Maass PG, Rinn JL, Mele M. 2020. Cisandtranseffects differentially contribute to the evolution of promoters and enhancers. Genome Biology. Aug;21.

Mi HY, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Research. Jan;41:D377-D386.

Moriguchi T, Toyoshima F, Masuyama N, Hanafusa H, Gotoh Y, Nishida E. 1997. A novel SAPK/JNK kinase, MKK7, stimulated by TNF alpha and cellular stresses. Embo Journal. Dec;16:7045-7053.

Nayernia K, Adham IM, Burkhardt-Gottges E, Neesen J, Rieche M, Wolf S, Sancken U, Kleene K, Engel W. 2002. Asthenozoospermia in mice with targeted deletion of the sperm mitochondrion-associated cysteine-rich protein (Smcp) gene. Molecular and Cellular Biology. May;22:3046-3052.

Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant source of bioactive RNAs or peptides. Nature Ecology & Evolution. Jun;1.

Neme R, Tautz D. 2014. Evolution: Dynamics of De Novo Gene Emergence. Current Biology. MAR 17 2014;24:R238-R240.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife. Feb 2;5.

Nishina H, Wada T, Katada T. 2004. Physiological roles of SAPK/JNK signaling pathway. Journal of Biochemistry. Aug;136:123-126.

Noblanc A, Kocer A, Chabory E, Vernet P, Saez F, Cadet R, Conrad M, Drevet JR. 2011. Glutathione Peroxidases at Work on Epididymal Spermatozoa: An Example of the Dual Effect of Reactive Oxygen Species on Mammalian Male Fertilizing Ability. Journal of Andrology. Nov-Dec;32:641-650.

Noda T, Sakurai N, Nozawa K, Kobayashi S, Devlin DJ, Matzuk MM, Ikawa M. 2019. Nine genes abundantly expressed in the epididymis are not essential for male fecundity in mice. Andrology. Sep;7:644-653.

Osada N, Miyagi R, Takahashi A. 2017. Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. Genetics. Aug;206:2139-2148.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of Drosophila orphan genes. Elife. Feb;3.

Prasad V, Okunade GW, Miller ML, Shull GE. 2004. Phenotypes of SERCA and PMCA knockout mice. Biochemical and Biophysical Research Communications. Oct;322:1192-1203.

Reddi PP, Flickinger CJ, Herr JC. 1999. Round spermatid-specific transcription of the mouse SP-10 gene is mediated by a 294-base pair proximal promoter. Biology of Reproduction. Nov;61:1256-1266.

Reddi PP, Shore AN, Shapiro JA, Anderson A, Stoler MH, Acharya KK. 2003. Spermatid-specific promoter of the SP-10 gene functions as an insulator in somatic cells. Developmental Biology. Oct;262:173-182.

Reddi PP, Urekar CJ, Abhyankar MM, Ranpura SA. 2007. Role of an insulator in testis-specific gene transcription. Testicular Chromosome Structure and Gene Expression.1120:95-103.

Ren XX, Chen XL, Wang ZL, Wang D. 2017. Is transcription in sperm stationary or dynamic? Journal of Reproduction and Development. Oct;63:439-443.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics. Jul;13:505-516.

Russell L, Ettlin R, Sinha Hikim A, Clegg E. 1990. Histological and histopathological evaluation of the testis. St. Louis: Cache River Press.

Sabapathy K, Jochum W, Hochedlinger K, Chang LF, Karin M, Wagner EF. 1999. Defective neural tube morphogenesis and altered apoptosis in the absence of both JNK1 and JNK2. Mechanisms of Development. Dec;89:115-124.

Scieglinska D, Vydra N, Krawczyk Z, Widlak W. 2004. Location of promoter elements necessary and sufficient to direct testis-specific expression of the Hst70/Hsp70.2 gene. Biochemical Journal. May;379:739-747.

Shi JW, Fok KL, Dai PY, Qiao F, Zhang MY, Liu HG, Sang MM, Ye M, Liu Y, Zhou YW, et al. 2021. Spatio-temporal landscape of mouse epididymal cells and specific mitochondria-rich segments defined by large-scale single-cell RNA-seq. Cell Discovery. May;7.

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. Trends in Genetics. Jul;34:532-544.

Somboonthum P, Ohta H, Yamada S, Onishi M, Ike A, Nishimune Y, Nozaki M. 2005. cAMP-responsive element in TATA-less core promoter is essential for haploid-specific gene expression in mouse testis. Nucleic Acids Research.33:3401-3411.

Statello L, Guo CJ, Chen LL, Huarte M. 2021. Gene regulation by long non-coding RNAs and its biological functions. Nature Reviews Molecular Cell Biology. Feb;22:96-118.

Staubach F, Teschke M, Voolstra CR, Wolf JBW, Tautz D. 2010. A TEST OF THE NEUTRAL MODEL OF EXPRESSION CHANGE IN NATURAL POPULATIONS OF HOUSE MOUSE SUBSPECIES. Evolution. Feb;64:549-560.

Takekawa M, Tatebayashi K, Saito H. 2005. Conserved docking site is essential for activation of mammalian MAP kinase kinases by specific MAP kinase kinase kinases. Molecular Cell. Apr;18:295-306.

Tautz D. 2000. Evolution of transcriptional regulation. Current Opinion in Genetics & Development. OCT 2000;10:575-579.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nature Reviews Genetics. Oct;12:692-702.

TAUTZ D, PFEIFLE C. 1989. A NON-RADIOACTIVE INSITU HYBRIDIZATION METHOD FOR THE LOCALIZATION OF SPECIFIC RNAS IN DROSOPHILA EMBRYOS REVEALS TRANSLATIONAL

CONTROL OF THE SEGMENTATION GENE HUNCHBACK. Chromosoma. AUG 1989;98:81-85.

Topaloglu O, Schluter G, Nayernia K, Engel W. 2001. A 74-bp promoter of the Tnp2 gene confers testis- and spermatid-specific expression in transgenic mice. Biochemical and Biophysical Research Communications. Nov;289:597-601.

Tournier C, Dong C, Turner TK, Jones SN, Flavell RA, Davis RJ. 2001. MKK7 is an essential component of the JNK signal transduction pathway activated by proinflammatory cytokines. Genes & Development. Jun;15:1419-1426.

Tournier C, Whitmarsh AJ, Cavanagh J, Barrett T, Davis RJ. 1997. Mitogen-activated protein kinase kinase 7 is an activator of the c-Jun NH2-terminal kinase. Proceedings of the National Academy of Sciences of the United States of America. Jul;94:7337-7342.

Tournier C, Whitmarsh AJ, Cavanagh J, Barrett T, Davis RJ. 1999. The MKK7 gene encodes a group of c-Jun NH2-terminal kinase kinases. Molecular and Cellular Biology. Feb;19:1569-1581.

Uchida A, Sakib S, Labit E, Abbasi S, Scott RW, Underhill TM, Biernaskie J, Dobrinski I. 2020. Development and function of smooth muscle cells is modulated by Hic1 in mouse testis. Development. Jul;147.

Van Oss SB, Carvunis AR. 2019. De novo gene birth. Plos Genetics. May;15.

Wada T, Joza N, Cheng HYM, Sasaki T, Kozieradzki I, Bachmaier K, Katada T, Schreiber M, Wagner EF, Nishina H, et al. 2004. MKK7 couples stress signalling to G2/M cell-cycle progression and cellular senescence. Nature Cell Biology. Mar;6:215-226.

Wang X, Destrument A, Tournier C. 2007. Physiological roles of MKK4 and MKK7: Insights from animal models. Biochimica Et Biophysica Acta-Molecular Cell Research. Aug;1773:1349-1357.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. Dec;420:520-562.

Wen ZZ, Liu DY, Zhu HX, Sun XY, Xiao Y, Lin ZC, Zhang AZ, Ye C, Gao JG. 2021. Deficiency for Lcn8 causes epididymal sperm maturation defects in mice. Biochemical and Biophysical Research Communications. Apr;548:7-13.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics. Mar;8:206-216.

Wu S, Li H, Wang L, Mak N, Wu X, Ge R, Sun F, Cheng C. 2021. Motor proteins and spermatogenesis. In: Molecular Mechanisms in Spermatogenesis. Springer.

Xie C, Bekpen C, Kunzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, Ullrich KK, Tautz D. 2019. A de novo evolved gene in the house mouse regulates female pregnancy cycles. Elife. Aug;8.

Xie C, Bekpen C, Kunzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, Ullrich KK, Zhang WY, Tautz D. 2020. Dedicated transcriptomics combined with power analysis lead to functional understanding of genes with weak phenotypic changes in knockout lines. Plos Computational Biology. Nov;16.

Yamaguchi R, Yamagata K, Hasuwa H, Inano E, Ikawa M, Okabe M. 2008. Cd52, known as a major maturation-associated sperm membrane antigen secreted from the epididymis, is not required for fertilization in the mouse. Genes to Cells. Aug;13:851-861.

Yang SH, Sharrocks AD, Whitmarsh AJ. 2003. Transcriptional regulation by the MAP kinase signaling cascades. Gene. Nov;320:3-21.

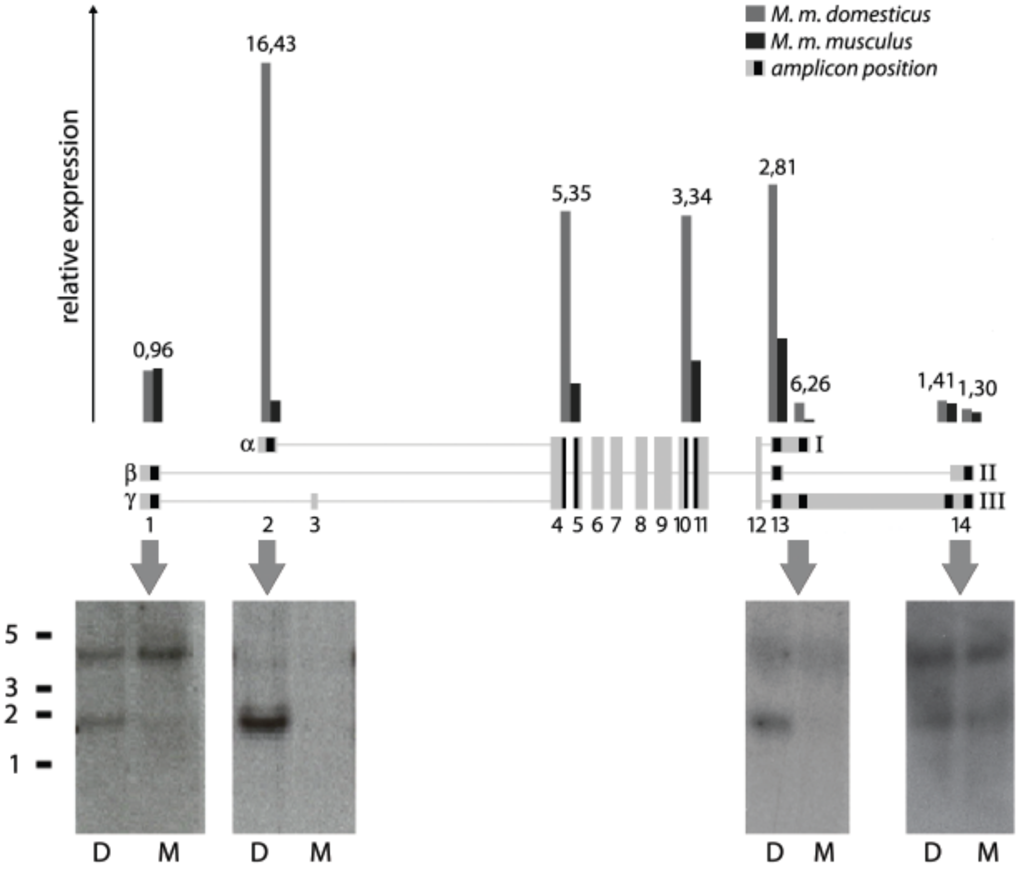
Zambrowicz BP, Harendza CJ, Zimmermann JW, Brinster RL, Palmiter RD. 1993. ANALYSIS OF THE MOUSE PROTAMINE-1 PROMOTER IN TRANSGENIC MICE. Proceedings of the National Academy of Sciences of the United States of America. Jun;90:5071-5075.

Zhu AQ, Ibrahim JG, Love MI. 2019. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. Jun;35:2084-2092.

Supplementary Material

Supplementary file S1

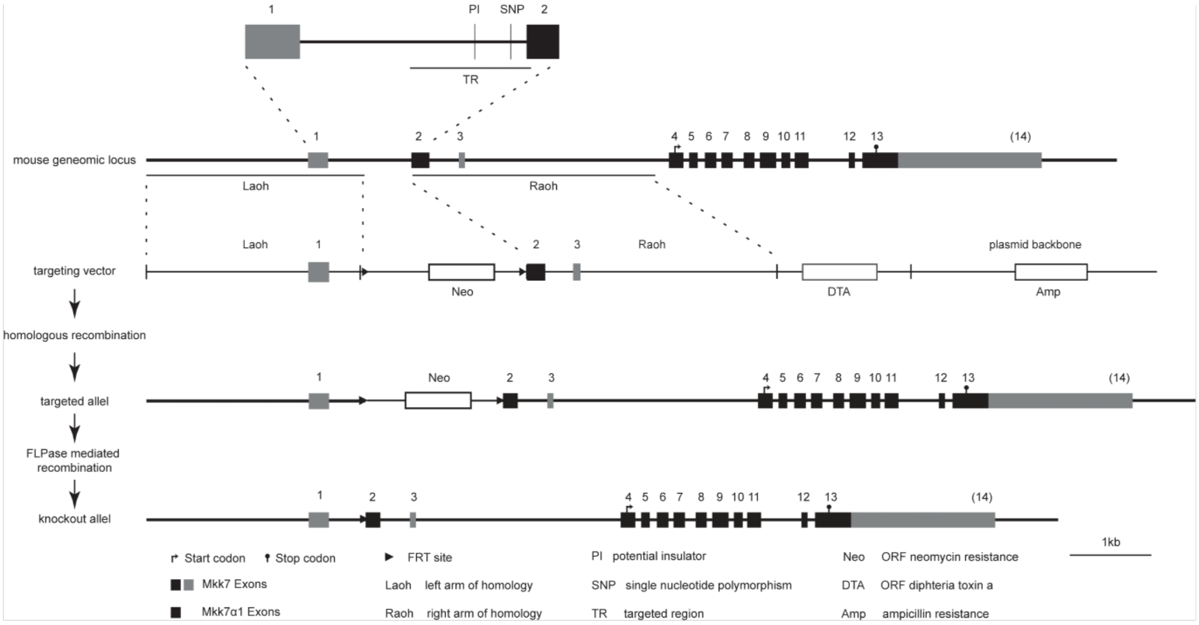
Exon specific Northen-Blotting and qPCR of *Map2k7*



Exon specific *Map2k7* Northern blots and qRT-PCRs comparing *M. m. musculus* and *M. m. domesticus*. Testis RNA from *M. m. domesticus* (D) and *M. m. musculus* (M) was hybridized with different probes against certain parts of *Map2k7*. Probe positions of the respective blots are indicated with grey arrows. The size standard on the left displays kb. Relative expression of different qRT-PCR amplicons (about 100 bp in size) is shown at the top. Two amplicons span neighboring exons (4-5; 10-11). Positions of the different amplicons are indicated by black areas in the exon map. The respective expression values are displayed as bars above. The numbers on top of the bars represent the ratio of the domesticus to the musculus value. The results lead to the assumption of a strong prevalence of Map2k7-α1 in *M. m. domesticus* which is missing in *M. m. musculus*.

Supplementary file S2

Targeting strategy for Map2k7-α1 knock out:



Wildtype sequence

chr8:4,239,310- 4,240,429 in GRCm38/mm10

```
gtcacctgcaccttgcaggctttcagactcgaagctacgccgctgtgactacaaccaagtcttttaactctgcaa
acagttatatctcttctgattcagtggttccactcctgccgagtcagtgactgtcaagaggtcccctcccctagc
agacagtccacacacgtgcatgcctatctgcccatgtaggtcactaagtcctcatctacatccgttttgattgga
                                        |>replaced in knockout
ggcttctatttgacttctttggtcatatcagatggctcc*ttctaagcttggaaggaccttgtcactggacccag
                                         >promotor test fragment
ctcactgcctcctacatacaggggcacctcatatctctaa#tgaccaactacttttcactattgctgtctagccc
tcaggaaacacatagccatctctcagcctggcagcctttgtctacagggctcaagtgactgctactactacagac
acctgtaaccagtgtagtcccatcgagtagaaacgcacctcctttctgagcctactgtctttgggcctgcctcct
                                             -259
gacatacggtcctgtaccaaaagtgcttcctcttcctttctagacAcacatccagcttgtccaccatgaccacag
ttagcatccttgttgtatcccagataccccctgtcccaagttgtctttgctgaaagaatctggctttttttctccc
                                                                     -84
ctctctgtccaacccttcctctgtccctcttgattcagcagaatgtcttctttatatcctctgtgatgtAatctt
                                               L
tggagtatacatactatagttgtctgtgtggtcactatggtaagaggGgaaaggcagcctcctgtaggtgaaaat
   0       B                     <|           <
tctGttcactaCctggccacctggcctgactgac*cttcacagctt#gatcatcttcctgaagaggcattcagga
ttccctccatccctaccccttctggacaaagtcttccacgtttccttcctgggagtttcttccaggaactggaga
          intron donor site
tacccagag>gtggggatgcatttcactgattctgcctgggaccagaggttgggcccctgctggattccagggcc
atccctccacggccctgtggatgagacagggttgggaatagggttctaggtgccataagagctgcctgtccctag
```

Legend: "0" marks the position that is used as reference for the transcription start site in Figure 4, "B" marks the start of the bulk of the transcripts detected in the RNA-Seq experiments, "L" marks the start of the longest annotated transcript in GRCm38/mm10. The region replaced in the knockout construct is indicated by * and red labels, the region used for promotor analysis is indicated by # and blue labels.

Inserted NEO vector

```
gtcacctgcaccttgcaggctttcagactcgaagctacgccgctgtgactacaaccaagtcttttaactctgcaa
acagttatatctcttctgattcagtggttccactcctgccgagtcagtgactgtcaagaggtcccctcccctagc
agacagtccacacacgtgcatgcctatctgcccatgtaggtcactaagtcctcatctacatccgttttgattgga
ggcttctatttgacttctttggtcatatcagatggctcc*caattggataagcttgatatcgaattccGAAGTTC
CTATTCTCTAGAAAGTATAGGAACTTCaggtctgaagaggagtttacgtccagccaagctagcttggctgcaggt
cgtcgaaattctaccgggtaggggaggcgcttttcccaaggcagtctggagcatgcgctttagcagccccgctgg
gcacttggcgctacacaagtggcctctggcctcgcacacattccacatccaccggtaggcgccaaccggctccgt
tctttggtggccccttcgcgccaccttctactcctcccctagtcaggaagttcccccccgccccgcagctcgcgt
cgtgcaggacgtgacaaatggaagtagcacgtctcactagtctcgtgcagatggacagcaccgctgagcaatgga
agcgggtaggcctttggggcagcggccaatagcagctttgctccttcgctttctgggctcagaggctgggaaggg
gtgggtccgggggcgggctcaggggcgggctcaggggcggggcgggcgcccgaaggtcctccggaggcccggcat
tctgcacgcttcaaaagcgcacgtctgccgcgctgttctcctcttcctcatctccgggcctttcgacctgcagcc
tgttgacaattaatcatcggcatagtatatcggcatagtataatacgacaaggtgaggaactaaaccatgggatc
ggccattgaacaagatggattgcacgcaggttctccggccgcttgggtggagaggctattcggctatgactgggc
acaacagacaatcggctgctctgatgccgccgtgttccggctgtcagcgcaggggcgcccggttctttttgtcaa
gaccgacctgtccggtgccctgaatgaactgcaggacgaggcagcgcggctatcgtggctggccacgacgggcgt
tccttgcgcagctgtgctcgacgttgtcactgaagcgggaagggactggctgctattgggcgaagtgccggggca
ggatctcctgtcatctcaccttgctcctgccgagaaagtatccatcatggctgatgcaatgcggcggctgcatac
gcttgatccggctacctgcccattcgaccaccaagcgaaacatcgcatcgagcgagcacgtactcggatggaagc
cggtcttgtcgatcaggatgatctggacgaagagcatcaggggctcgcgccagccgaactgttcgccaggctcaa
ggcgcgcatgcccgacggcgatgatctcgtcgtgacccatggcgatgcctgcttgccgaatatcatggtggaaaa
tggccgcttttctggattcatcgactgtggccggctgggtgtggcggaccgctatcaggacatagcgttggctac
ccgtgatattgctgaagagcttggcggcgaatgggctgaccgcttcctcgtgctttacggtatcgccgctcccga
ttcgcagcgcatcgccttctatcgccttcttgacgagttcttctgaggggatcaattctctagagctcgctgatc
agcctcgactgtgccttctagttgccagccatctgttgtttgcccctcccccgtgccttccttgaccctggaagg
tgccactcccactgtcctttcctaataaaatgaggaaattgcatcgcattgtctgagtaggtgtcattctattct
ggggggtggggtggggcaggacagcaagggggaggattgggaagacaatagcaggcatgctggggatgcggtggg
ctctatggcttctgaggcggaaagaaccagctggggctcgactagagcttgcggaacccttcGAAGTTCCTATTC
TCTAGAAAGTATAGGAACTTCatcagtcaggtacataatatagatct*cttcacagcttgatcatcttcctgaag
aggcattcaggattccctccatccctaccccttctggacaaagtcttccacgtttccttcctgggagtttcttcc
aggaactggagatacccagag>gtggggatgcatttcactgattctgcctgggaccagaggttggcccctgctg
gattccagggccatccctccacggccctgtggatgagacagggttgggaatagggttctaggtgccataagagct
gcctgtccctagagcagagtaagacctggtaaggctagtggctggcaggccccaaggagtcccctgc
```

Sequence after FRT recombination

```
gtcacctgcaccttgcaggctttcagactcgaagctacgccgctgtgactacaaccaagtcttttaactctgcaa
acagttatatctcttctgattcagtggttccactcctgccgagtcagtgactgtcaagaggtcccctcccctagc
agacagtccacacacgtgcatgcctatctgcccatgtaggtcactaagtcctcatctacatccgttttgattgga
ggcttctatttgacttctttggtcatatcagatggctcc*caattggataagcttgatatcgaattccGAAGTTC
CTATTCTCTAGAAAGTATAGGAACTTCatcagtcaggtacataatatagatct*cttcacagcttgatcatcttc
ctgaagaggcattcaggattccctccatccctaccccttctggacaaagtcttccacgtttccttcctgggagtt
tcttccaggaactggagatacccagag>gtggggatgcatttcactgattctgcctgggaccagaggttggcccc
ctgctggattccagggccatccctccacggccctgtggatgagacagggttgggaatagggttctaggtgccata
agagctgcctgtccctagagcagagtaagacctggtaaggctagtggctggcaggccccaaggagtcccctgc
```

vector derived sequences underlined

FRT sites in yellow