**Supplementary Information: Analysis of *Mater* duplication**

*Mater* is a maternal effect gene that is essential for early embryonic development in mice [25]. A detailed analysis of *Mater* also has clinical relevance, as it is the predominant autoantigen in mouse autoimmune oophoritis, a model for autoimmune premature ovarian failure in women [24]. *Mater* is expressed exclusively in oocytes and accumulates in the egg during oogenesis. Eggs from mice lacking *Mater* are fertilized but are unable to progress past the 2-cell stage [25].

A 42 kb segmental duplication involving two duplicons (DUP1 – where *Mater* is located; and DUP2 – where a novel *Mater2* is located) reside about 5 Mb apart and are situated in an inverted orientation. Furthermore, an intron-less *Mater* pseudogene (*MaterP*), sharing 87 percent DNA sequence identity to *Mater*, is located 10 Mb proximal to *Mater*. DUP1 and DUP2 are on average 91.1 percent identical over the entire 57 kb genomic region with a 96.6 percent average in the exonic regions. DUP2 contains similarity to all *Mater* exons except for exon 1 and a deletion corresponding to *Mater* exon 8, which encodes the leucine-rich domain (Fig. A). One full length cDNA sequence (AK016782) and three separate EST sequences (BB645510, BB624558, and AV367637) align unambiguously to DUP2 supporting a putative *Mater2* transcript sequence that is 1878 bp in length, with an open reading frame of 269 amino acids. This open reading frame shares 95 percent amino acid identity with the C-terminal end of *Mater*. These *Mater2* transcripts are derived from testis, adipose, and lung tissue libraries suggesting it may have a different expression pattern than the oocyte-specific *Mater* [40]. It is still unclear whether *Mater2* represents a pseudogene or a functional gene that has evolved a different expression pattern and function.

The characterization of the regulatory elements surrounding duplicated gene loci are important for sub-functionalization hypotheses, as it is believed that mutations in these regions may alter expression patterns that can allow for gene specialization (reviewed by [3]). While examining conserved non-coding sequences (CNS) between *Mater* and the related loci, we found two EST sequences (BB608289 and BB666495) suggesting the presence of an additional exon 18 kb upstream of previously labeled exon 1 [41]. This new exon, labeled as exon 1 in Fig. A, is found conserved in human but not in the *Mater2* locus. While no expression data supports the existence of this exon 1 in human, its conservation further supports *MATER* and *Mater*'s proposed orthologous relationship [5]. Using rVISTA [42] we looked for conserved regulatory elements and identified a conserved CCAAT motif 59 bp upstream of exon 1. This conserved non-coding segment (CNS) may contain the proximal promoter responsible for the oocyte specific expression observed in both species. We utilized Laj (Local Alignments with Java; [43]) to interactively display the annotation, genomic alignments and coordinates, and can be found at http://marcie.ceh.uvic.ca/~mdwilson/mater.html. Similar analyses and tools will be useful for preliminary analyses of other duplicated loci in the mouse genome.

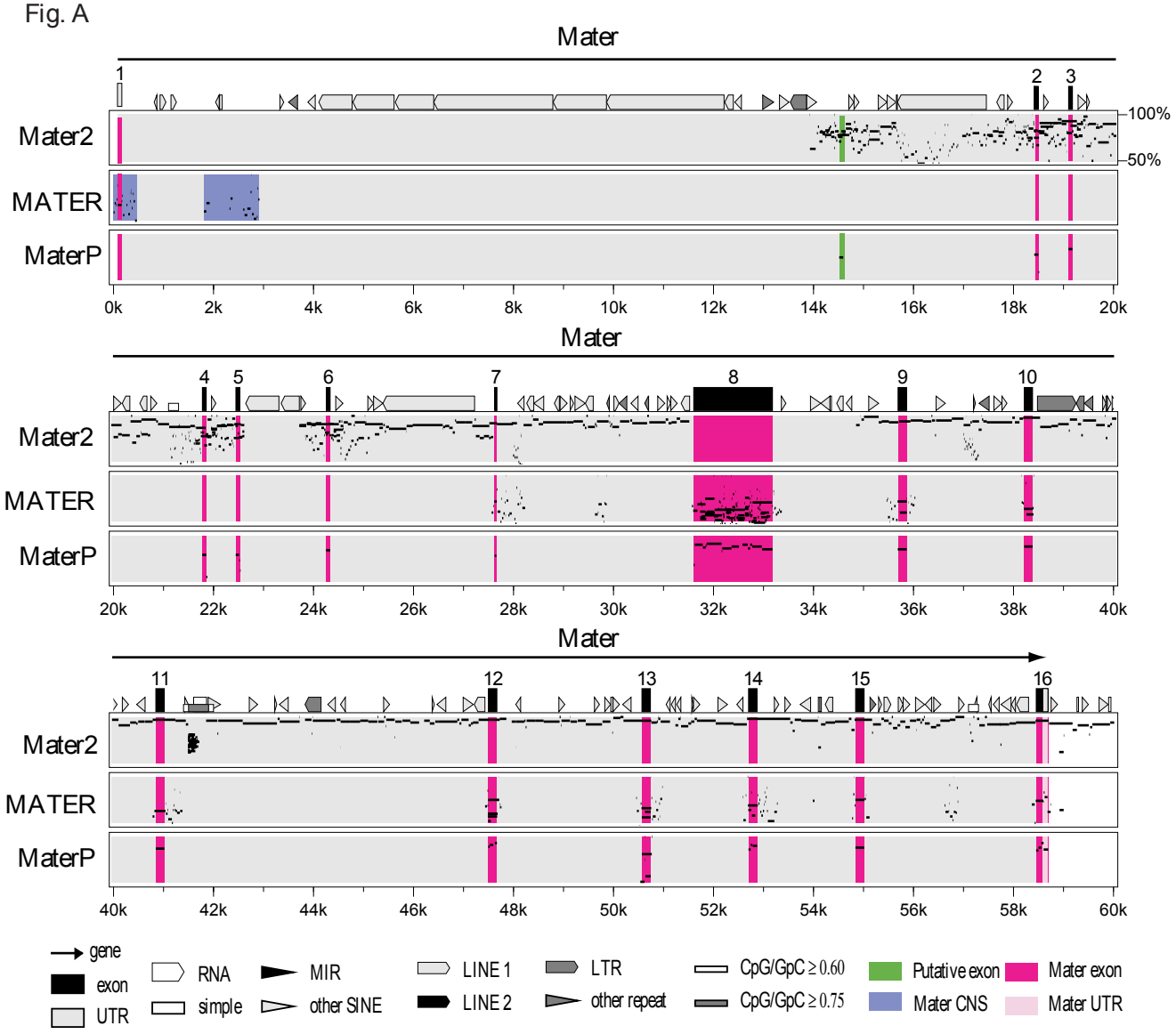
**Methods:**

**Comparative sequence analysis of *Mater*.** Repeat elements were characterized using RepeatMasker (Smit and Green unpublished). Comparative sequence alignments were generated using PipMaker [44] and VISTA [45]. Conserved regulatory elements were examined using regulatory VISTA [42]. DNA and protein sequences were compared using BLAST-2-Seq [46].

**Figure Legend:**

**Fig. A.** Multiple percent identity plot of *Mater* versus: *Mater2*, *MATER* (human), and *MaterP*. 60 kb of sequence from the mouse *Mater* genomic locus (found at chromosome 7, position 15617364-15677363 on the February 2002 mouse genome assembly) is compared to 500 kb surrounding mouse *Mater2*; 509 kb surrounding human *MATER*; and 42 kb surrounding the intron-less pseudogene *MaterP*. The percent identity, which ranges from 50 percent to 100 percent is indicated on the y-axis and is drawn as a series of horizontal lines indicating similarity between the 60 kb *Mater* genomic region and each target sequence. Discontinuity in the horizontal lines and dots along the x-axis corresponds to gaps in the alignment. Stacking of these lines indicates more than one section of the target sequence aligns to the same position on *Mater*. Purple shading indicates regions corresponds to *Mater* exons, blue shading corresponds to conserved non-coding sequence (CNS) between *Mater* and *MATER*, and green shading corresponds to a putative exon containing splice acceptor/donor sites that are highly similar to *Mater* exons 2 through 6 (all of which appear to have arisen through duplication events).

Fig. A


Mater
1
2
3
100%
50%
Mater2
MATER
MaterP
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
Mater
4
5
6
7
8
9
10
Mater2
MATER
MaterP
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
Mater
11
12
13
14
15
16
Mater2
MATER
MaterP
40k
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
gene
exon
UTR
RNA
simple
MIR
other SINE
LINE 1
LINE 2
LTR
other repeat
CpG/GpC ≥ 0.60
CpG/GpC ≥ 0.75
Putative exon
Mater CNS
Mater exon
Mater UTR