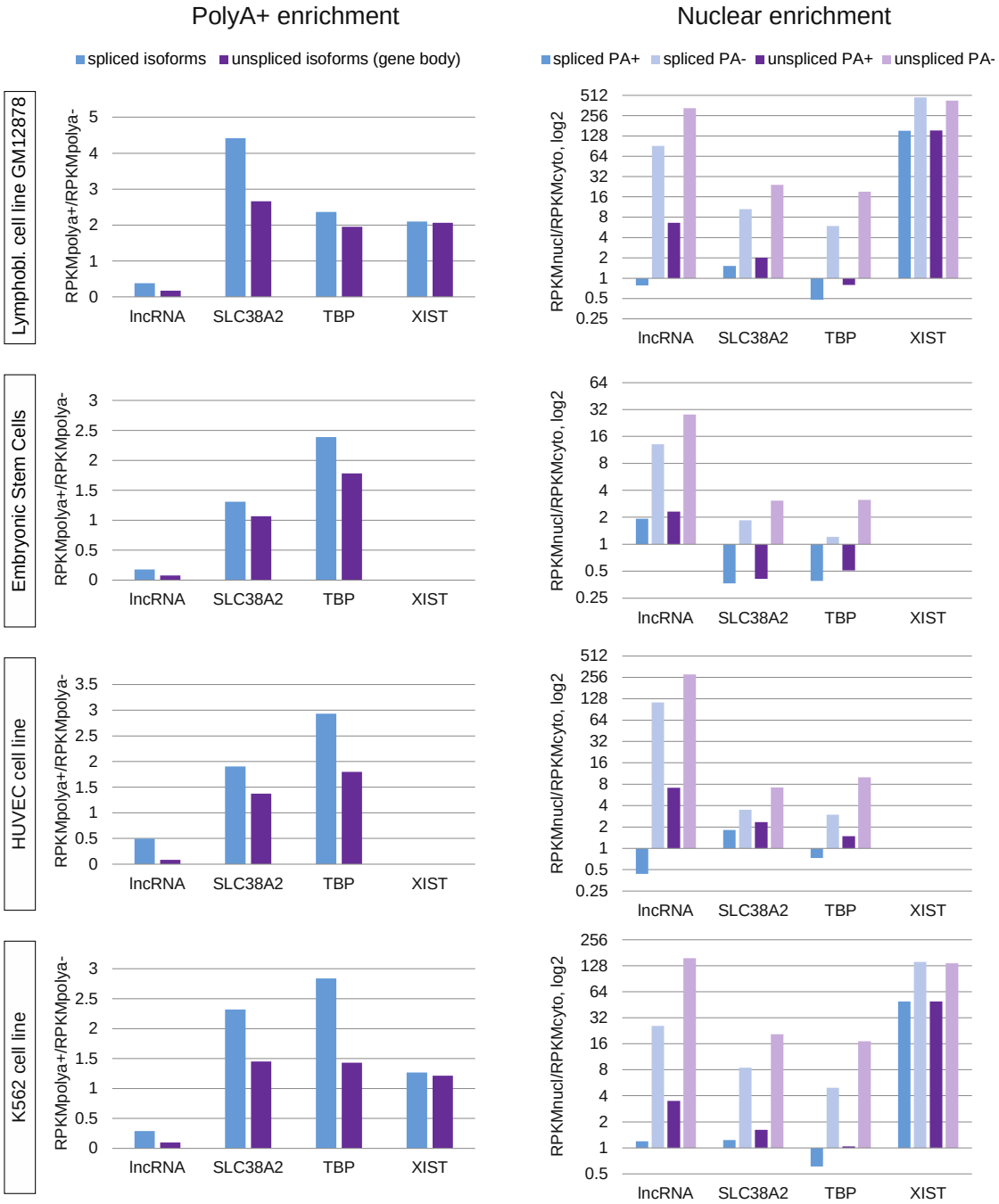
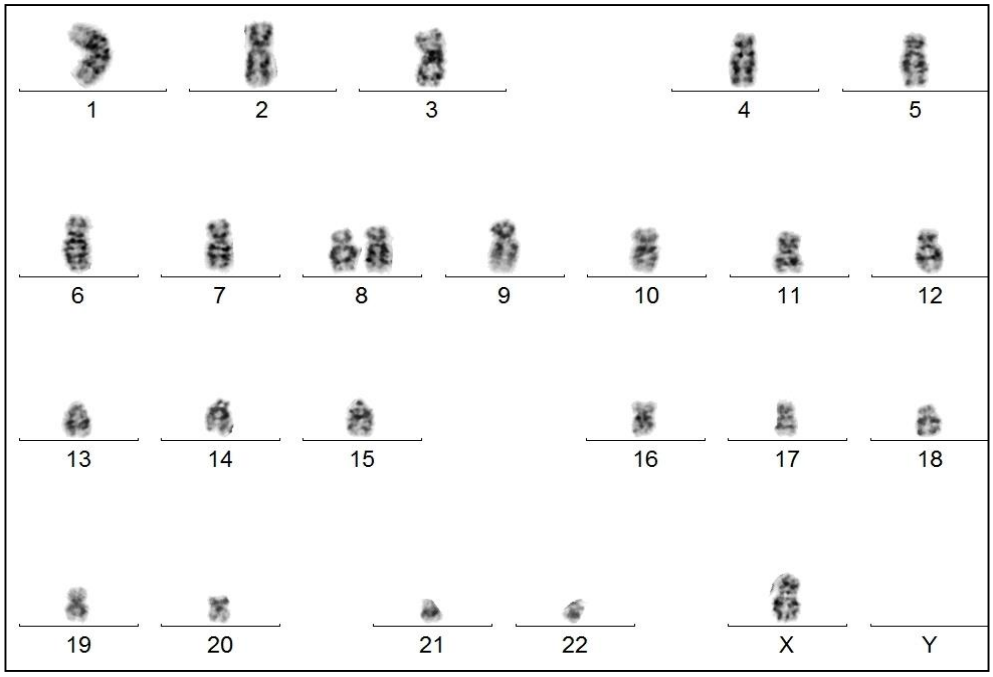
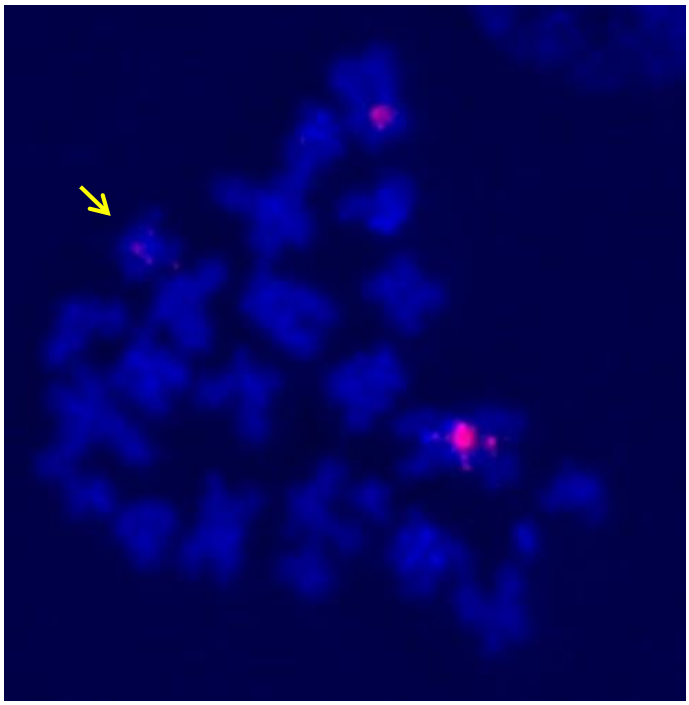
Supplemental Figure 1. Analysis of *LOC100288798* processing in additional cell lines.



Analyzing ENCODE RNA-seq from different cell and RNA fractions confirms inefficient processing and distinct biology of spliced and unspliced isoforms of *LOC100288798* (marked as "lncRNA") in multiple cell types (Results, Fig. 1E,F): from top to bottom - Lymphoblastoid Cell line GM12878, Human Embryonic Stem Cells, HUVEC cell line and K562 cell line. Bar plots on the left show PolyA+ enrichment as described for Figure 1E and bar plots on the right show nuclear enrichment as described for Figure 1F for the four genes in the four cell lines. Embryonic stem cells and HUVEC do not express *XIST* lncRNA and thus there are no bars corresponding to *XIST* for these two cell lines

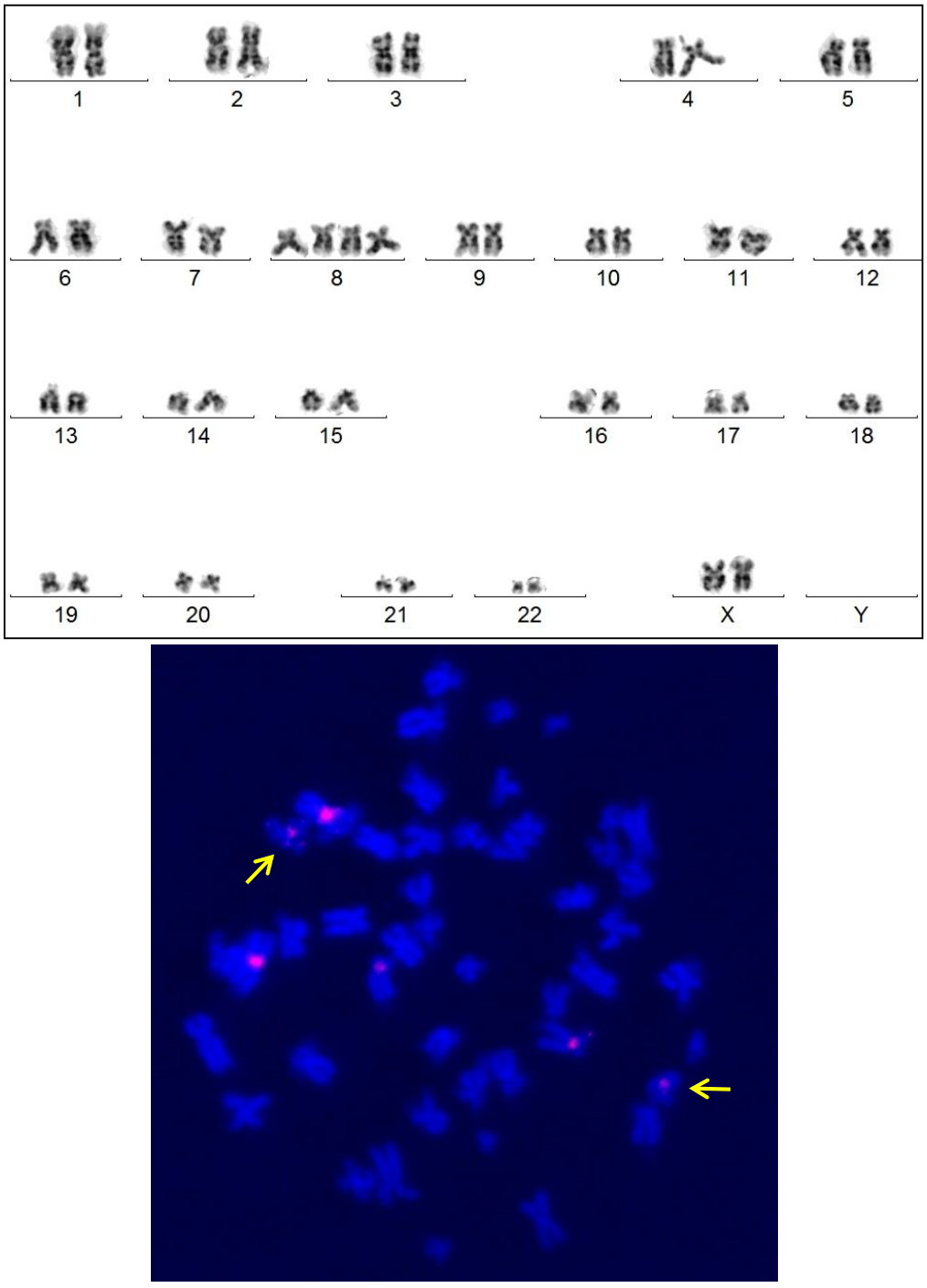
Supplemental Figure 2. Chromosome analysis of WT2 KBM7 cell line





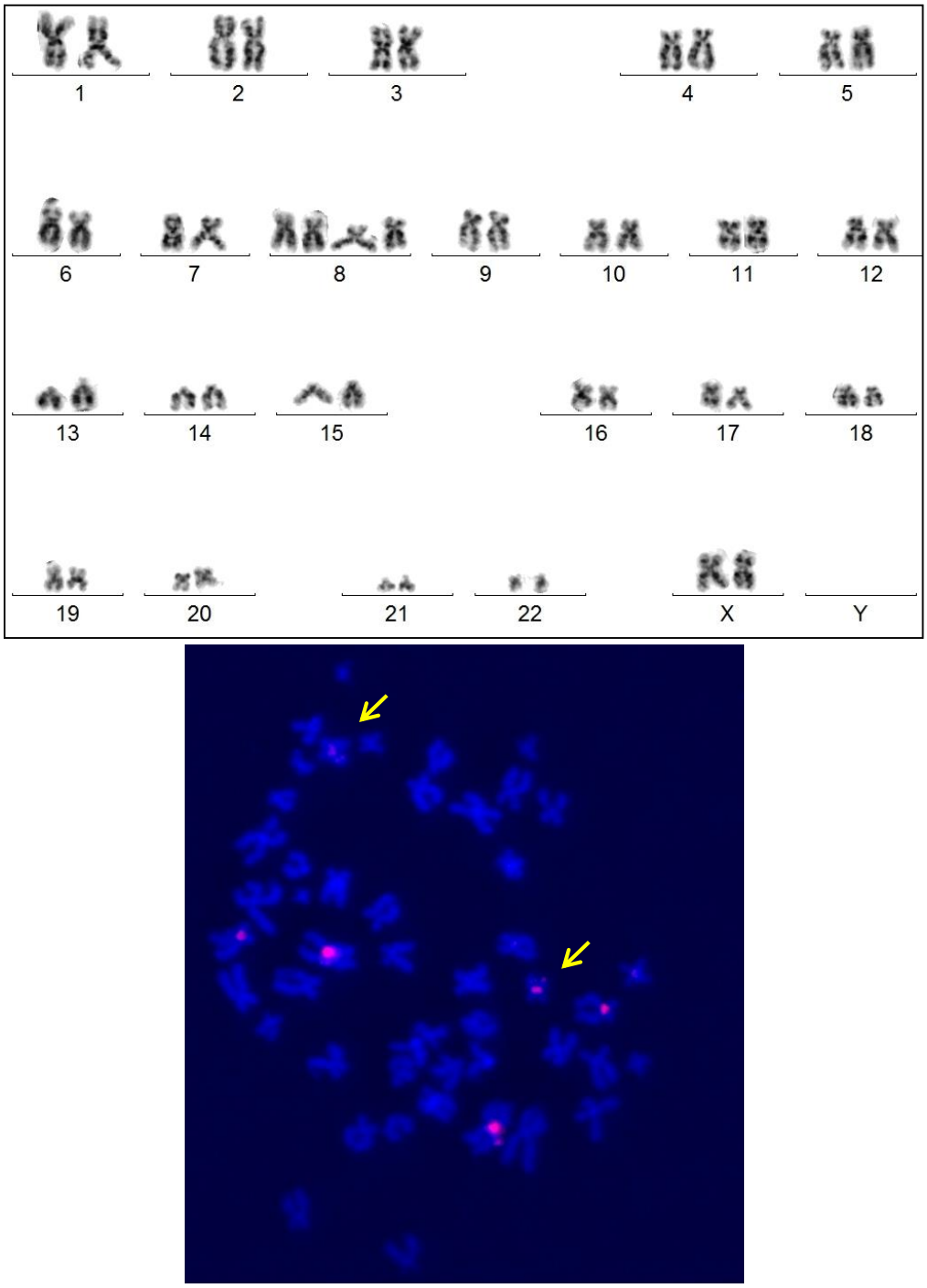
Top: Karyogram showing a haploid set of chromosomes (without the Y-chromosome), except for chromosome 8 which is disomic as reported before (details see text). Bottom: Because chromosomes 16 and 19 display a similar size in "G-bands produced with trypsin and Giemsa" (GTG) banding analysis we performed FISH analysis (Fluorescent In Situ Hybridisation) on metaphase chromosomes using a probe mix that label the centromere regions of chromosomes 1, 5 und 19. The result indicates the presence of both chromosomes 19 and 16. Yellow arrow indicates chromosome 19.

Supplemental Figure 3. Chromosome analysis of C1 KBM7 cell line



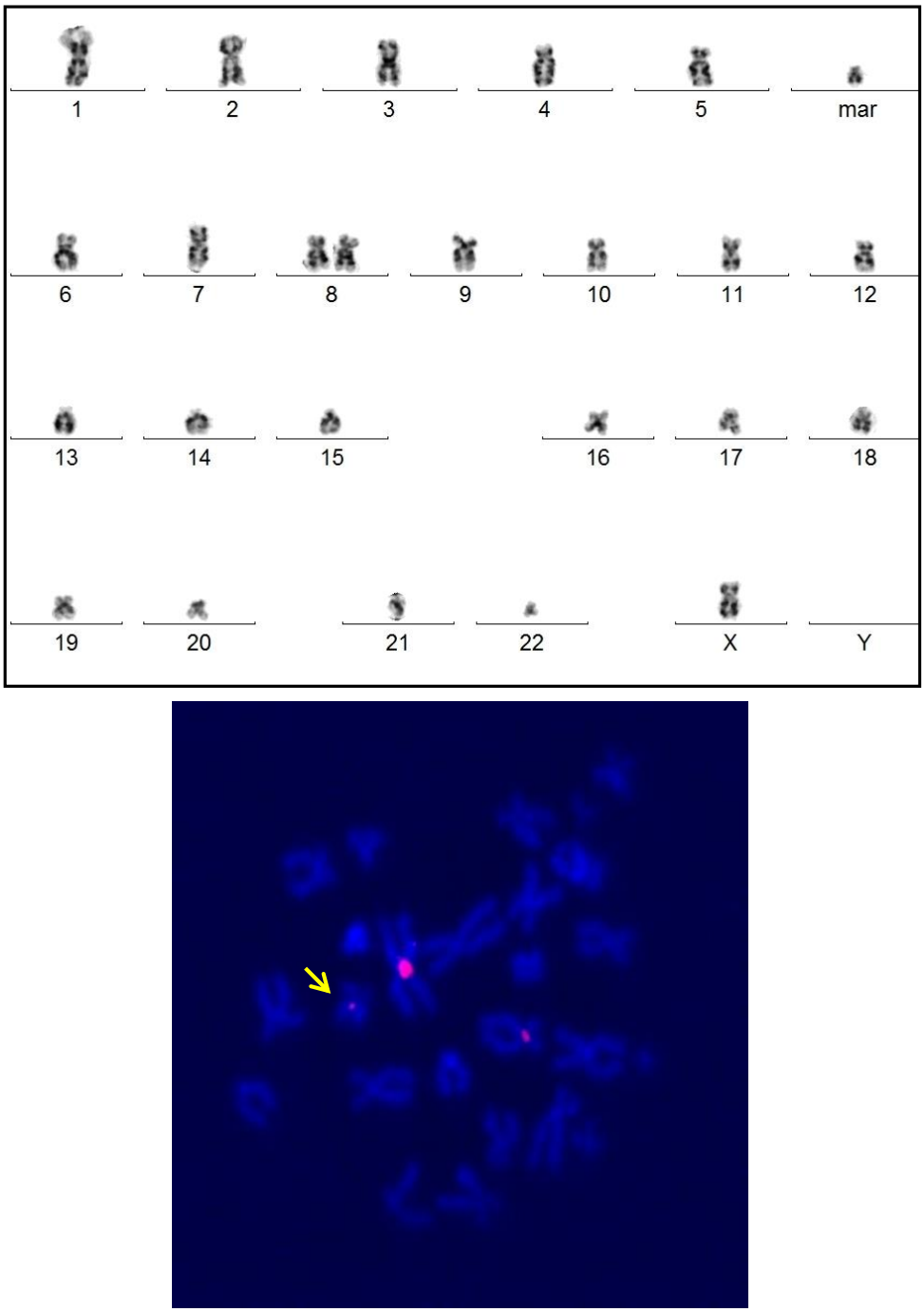
Top: Karyogram showing diploid chromosome set including all chromosomes, except Y. Note that chromosome 8 is tetrasomic, as diploid KBM7 cells result from endoreduplication of haploid KBM7 cells, where chromosome 8 is disomic (details see text). Bottom: As chromosomes 16 and 19 could not be visually distinguished by GTG-banding analysis we performed FISH analysis of metaphase chromosomes with centromeric probes for chromosomes 1, 5 und 19. This analysis indicated the presence of both chromosomes 19 and 16. Yellow arrows indicate chromosomes 19.

Supplemental Figure 4. Chromosome analysis of 3kb2 KBM7 cell line



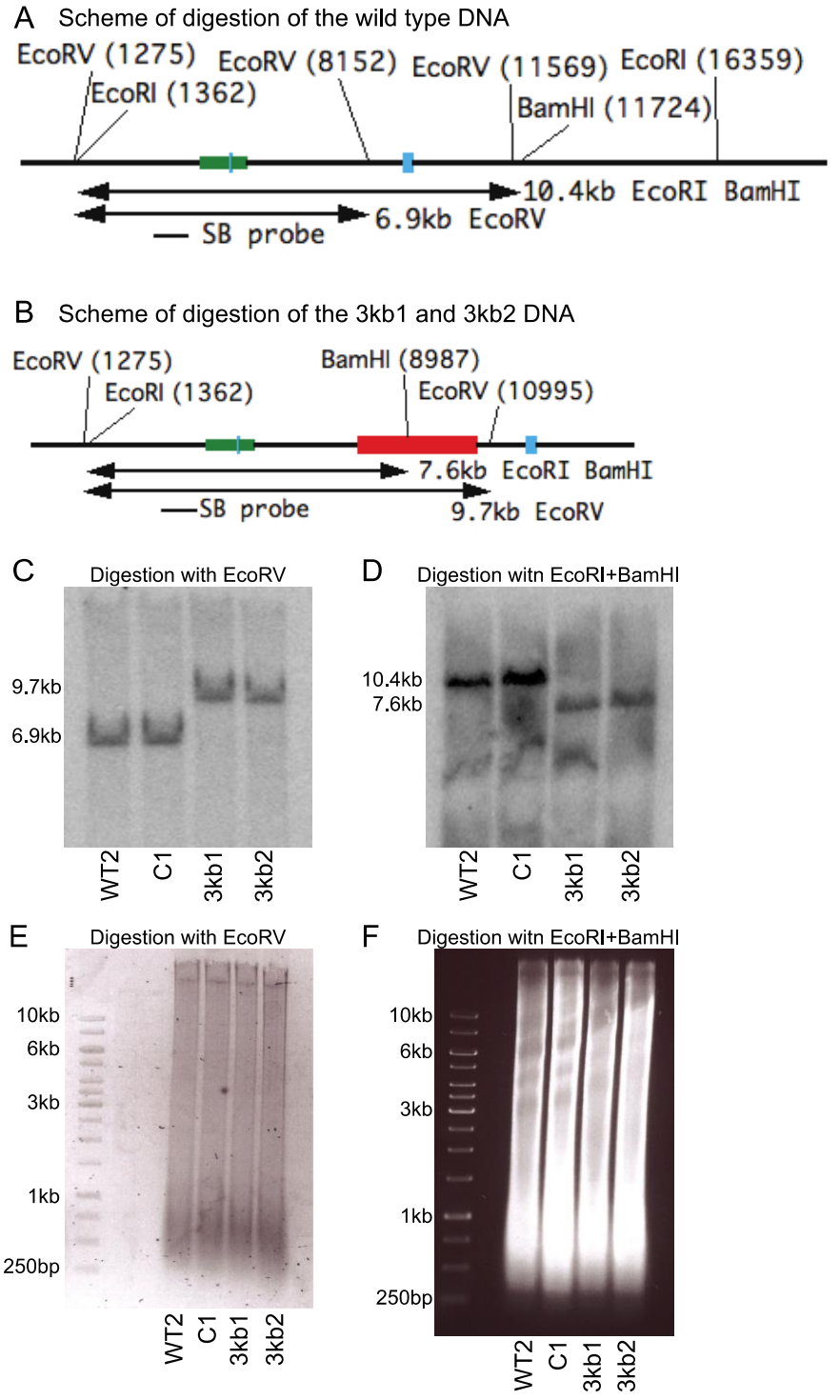
Top: Karyogram showing diploid chromosome set including all chromosomes, except Y. Note that chromosome 8 is tetrasomic, as diploid KBM7 cells result from endoreduplication of haploid KBM7 cells, where chromosome 8 is disomic (details see text). Bottom: As chromosomes 16 and 19 could not be visually distinguished by GTG-banding analysis we performed FISH analysis of metaphase chromosomes with centromeric probes for chromosomes 1, 5 und 19. This analysis indicated the presence of both chromosomes 19 and 16. Yellow arrows indicate chromosomes 19.

Supplemental Figure 5. Chromosome analysis of 100kb1 KBM7 cell line



Top: Karyogram showing a haploid set of chromosomes (without the Y-chromosome), except for chromosome 8 which is disomic as reported before (details see text). Bottom: Because chromosomes 16 and 19 display a similar size in GTG banding analysis we performed FISH analysis (Fluorescent In Situ Hybridisation) on metaphase chromosomes using a probe mix that label the centromere regions of chromosomes 1, 5 und 19. The result indicates the presence of both chromosomes 19 and 16. Yellow arrow indicates chromosome 19.

Supplemental Figure 6. Integrity of the locus remains upon 3kb1 and 3kb2 gene trap insertions



DNA-blot assay to validate the integrity of the genomic locus after the gene trap insertion in 3kb1 and 3kb2 KBM7 clones.

DNA-blot assay design: (**A**) wild type allele (displayed region - chr12:46,772,535-46,791,476), (**B**) 3kb gene trap insertion allele (displayed region - chr12:46,772,535-46,784,103). Genomic regions (h19) downloaded from the UCSC genome browser are shown with the positions (relative to the region start) and

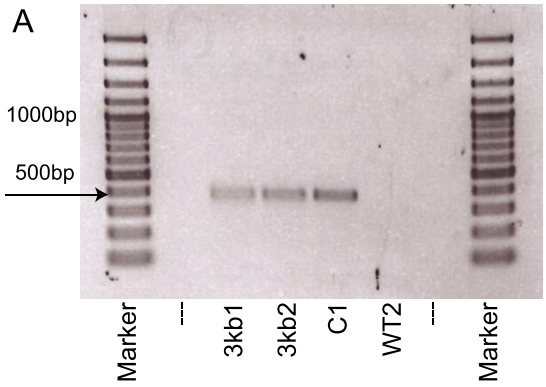
names of the selected restriction enzyme cutting sites. Below the scheme of the region the sizes of restriction fragments relevant for the assay are shown together with the position of the 865bp Southern blot probe that was produced with primers SLC38ASBPF/R (forward SLC38ASBPF: CCTTTTCATTTGACCCTGGA, reverse SLC38ASBPR: ACTCAAAGGGGGTTGTTGTG). Green box: CpG island promoter of *SLC38A4-AS*, light blue boxes: exons 1 and 2 of *SLC38A4-AS*. Red box: gene trap cassette sequence (only present in **B** that describes the 3kb truncation allele).

(**C**) DNA blot of genomic DNA from WT2, C1, 3kb1 and 3kb2 cell lines digested with EcoRV enzyme, transferred to a membrane and hybridized with the probe indicated in **A** and **B**.

(**D**) DNA blot of genomic DNA from WT2, C1, 3kb1 and 3kb2 cell lines digested with EcoRI and BamHI enzyme, transferred to a membrane and hybridized with the probe indicated in **A** and **B**.

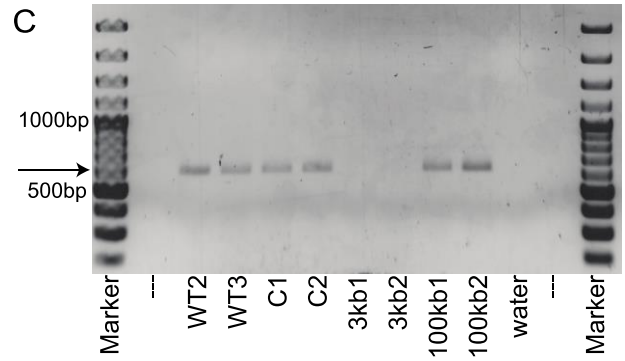
Note that in all cases the expected band sizes were obtained and that the small differences between neighboring bands result from unequal separation of genomic DNA in different wells of the agarose gel as shown by the respective agarose gel picture stained with ethidium bromide, obtained before transfer of the DNA to the membrane (**E** corresponding to **C**, and **F** corresponds to **D**).

Supplemental Figure 7. Splice acceptor sequence validation and wild type cell contamination test in 3kb1 and 3kb2 KBM7 cell lines



B

```
3kb1_trap_genetrap_F   AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
3kb2_trap_genetrap_F   AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
C1_trap_genetrap_F     AAGTTATGATCGCAGGCGCAATCTTCGCATTTCTTTTTTCCAGATGGTGAGCAAGGGCG
splice_acceptor              -----GATCGCAGGCGCAATCTTCGCATTTCTTTTTTC----------
                             **********************************
```

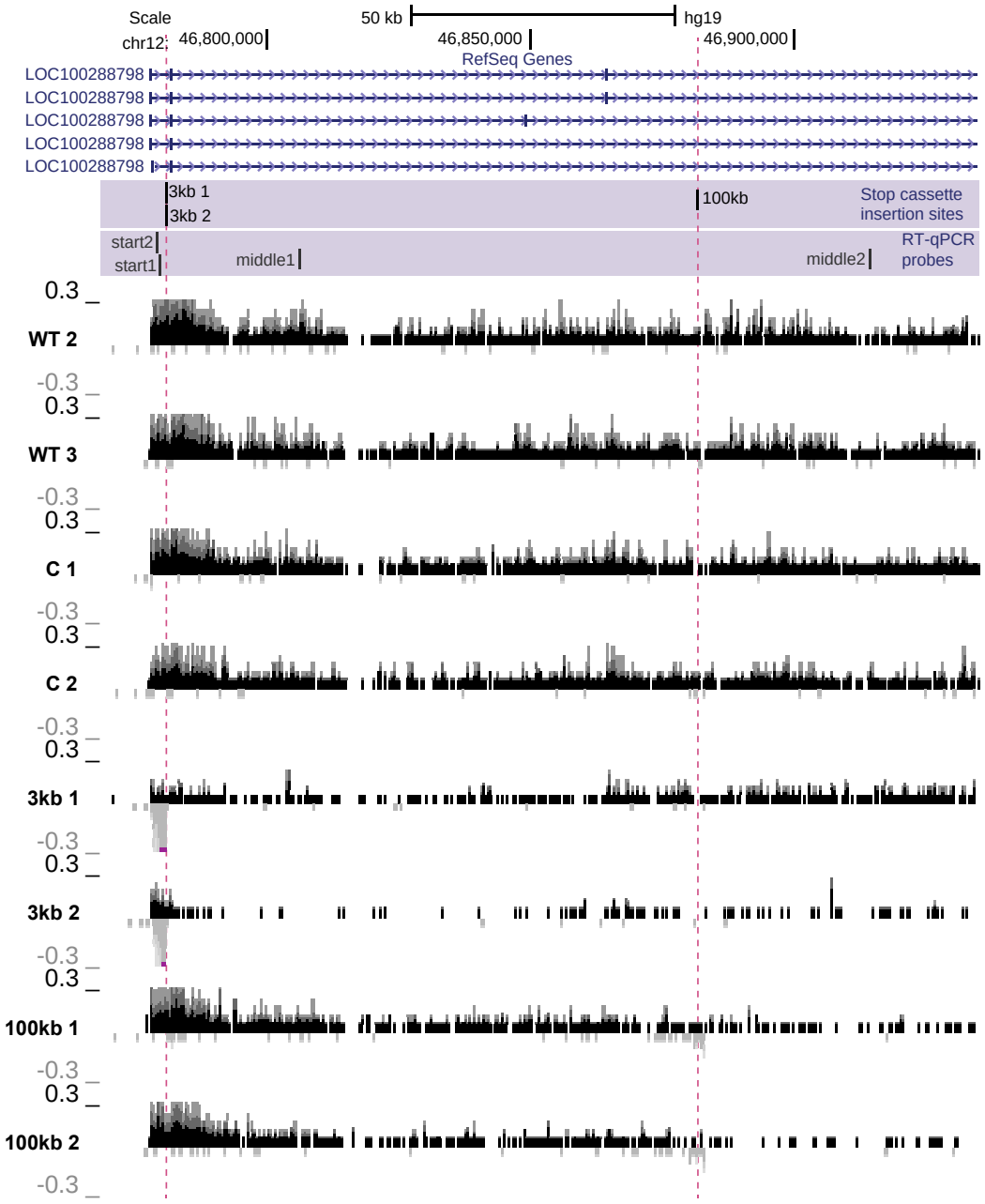


(**A**) PCR amplifying a 385bp long fragment of gene trap cassette containing the splice acceptor site (primers: forward – CCTACAGGTGGGGTCTTTCA, reverse - AAGTCGTGCTGCTTCATGTG) was performed on genomic DNA from 3kb1, 3kb2, C1 and WT2. The resulting fragment amplified by PCR was of the expected size (indicated by the horizontal arrow).

(**B**) Part of the sequence, obtained by Sanger sequencing of the PCR fragment shown in **A** with the indicated forward primer is displayed together with the splice acceptor sequence used in the gene trap cassette. Stars indicate matched bases.

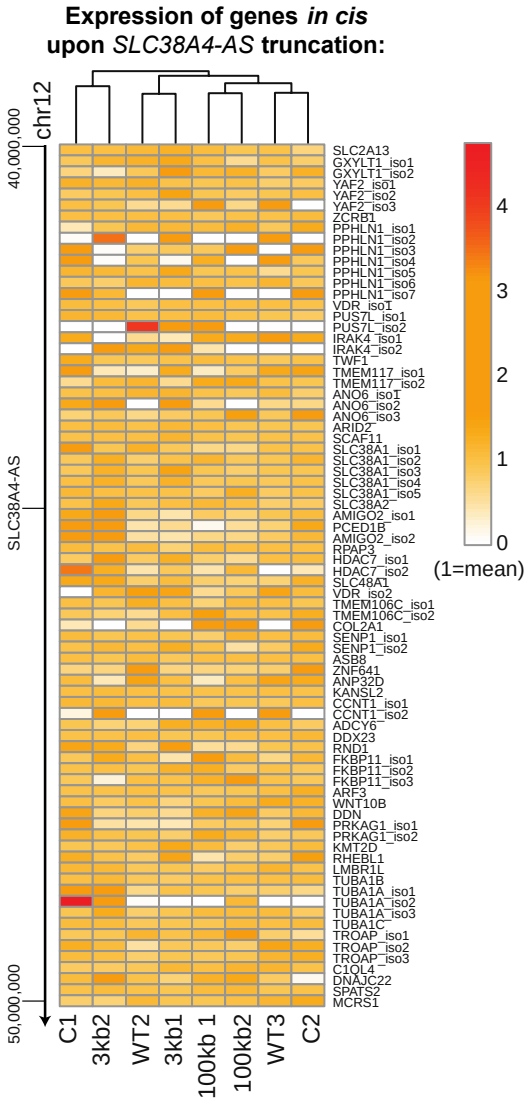
(**C**) PCR flanking the insertion site of the 3kb gene trap cassette insertion (primers: forward – TCAAAGTGTCTGCTGTTAGGTTG, reverse - TATTGCCTCCACAGCTCAAA) was performed on genomic DNA from WT2, WT3, C1, C2, 3kb1, 3kb2, 100kb1 and 100kb2 KBM7 cell lines. The gel picture shows that in all samples, except for 3kb1 and 3kb2, a PCR produced of expected fragment size was obtained: 608bp (indicated by the horizontal arrow). In 3kb1 and 3kb2 cell lines the size of the region targeted by the PCR primers is increased to 3,453bp by the insertion of the gene trap cassette and thus is too long to be amplified in a standard PCR reaction. The absence of signal in 3kb1 and 3kb2 samples indicates lack of detectable wild type cell contamination in both of these cell lines.

Supplemental Figure 8. Emergence of antisense transcription at the site of gene trap cassette insertion



Inspection of RNA-Seq signal of the eight clones reveals emergence of antisense transcription initiating at gene trap insertion sites. Top: chromosome coordinates, RefSeq annotation corresponding to first 150kb of *LOC100288798*, location of gene trap insertion sites, location of RT-qPCR probes. Bottom: RNA-seq signal, normalized to sample read number, pink dots indicate RNA-seq signal that exceeds the range presented inside the box. Name of the cell line is indicated on the left. Vertical dashed red lines indicate position of the 3kb and 100kb stop cassettes. Low density of RNA-seq signal piles indicate low expression and the smallest size corresponds to 1 read.

Supplemental Figure 9. Truncation of *SLC38A4-AS* lncRNA does not affect genes *in cis*



Heat map shows expression level (FPKM, Methods) of genes (name indicated on the right, "iso" stands for isoform when more than one isoform is displayed) in the 10Mbp region around *SLC38A4-AS* lncRNA transcription start site in the four truncation cell lines and the four control cell lines. Expression values are normalized to the mean FPKM among all 8 samples. Mean is set to 1. Only genes with mean FPKM > 1 are displayed: 47 genes (78 isoforms). Heat map color legend is displayed on the right. Heat map was built in R using *pheatmap* function with options *clustering_distance_cols= "canberra", clustering_distance_rows= "euclidean"*.