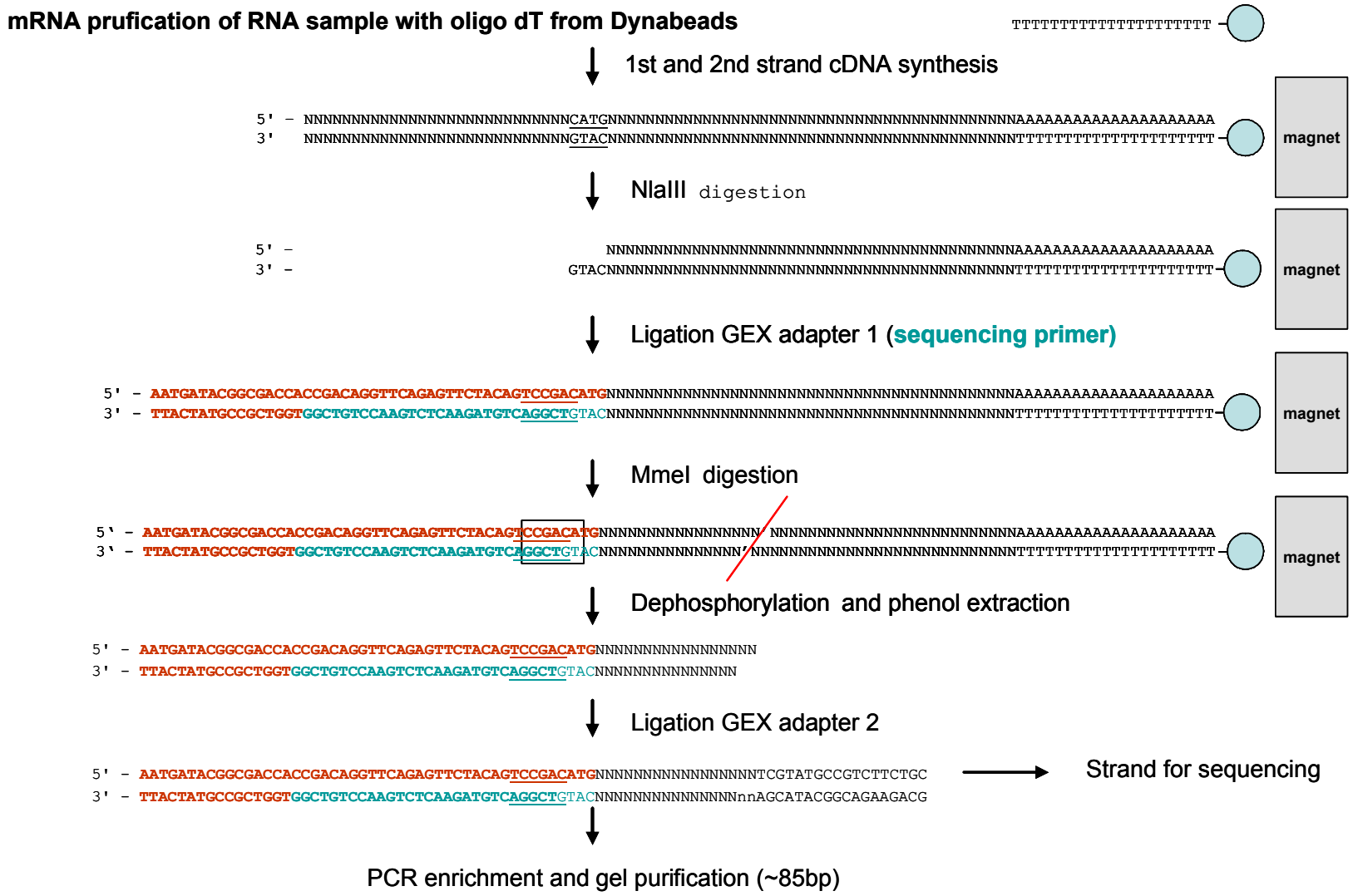
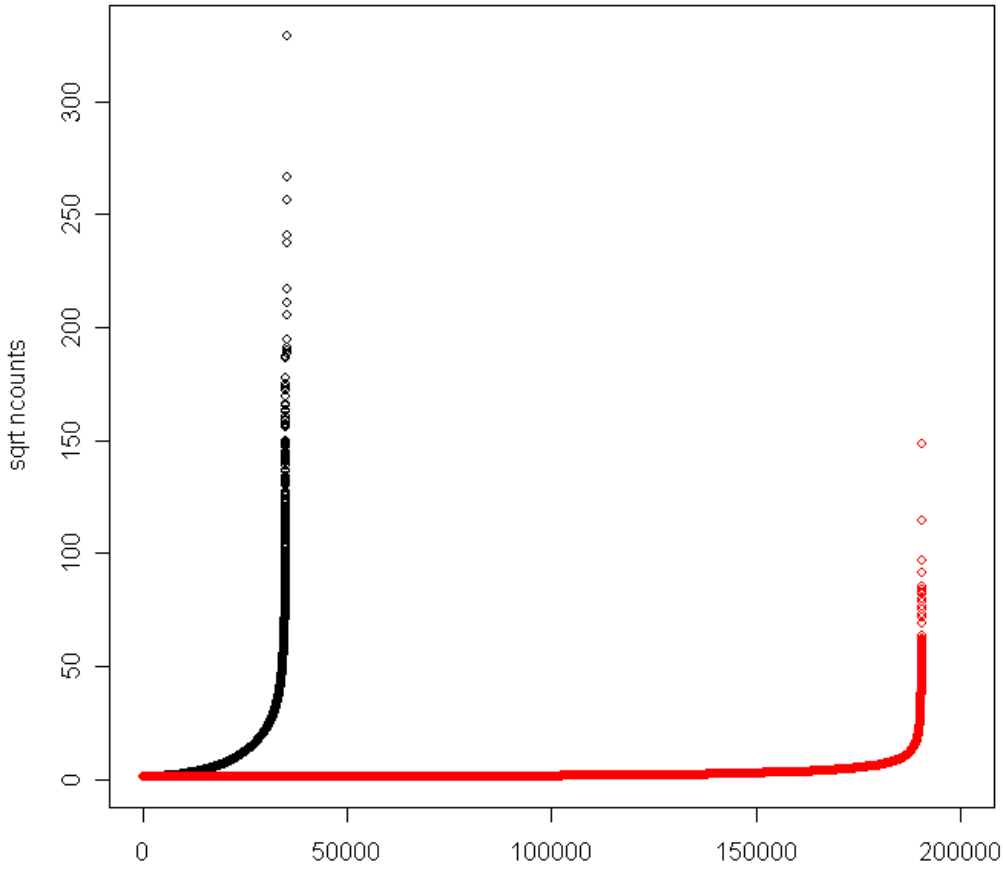
Supplementary Figure 1: Schematic overview of the procedure for target preparation



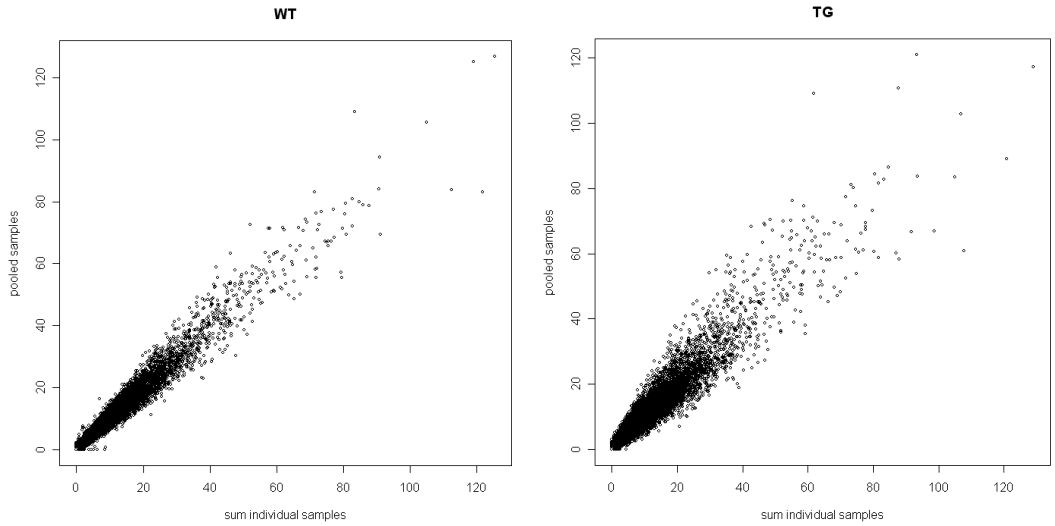
Supplementary Figure 2: Non-canonical tags are generally less abundant than canonical tags

Distribution of the square root of the number of counts per unique tag (canonical tags in black and non-canonical tags in red), sorted from high to low.



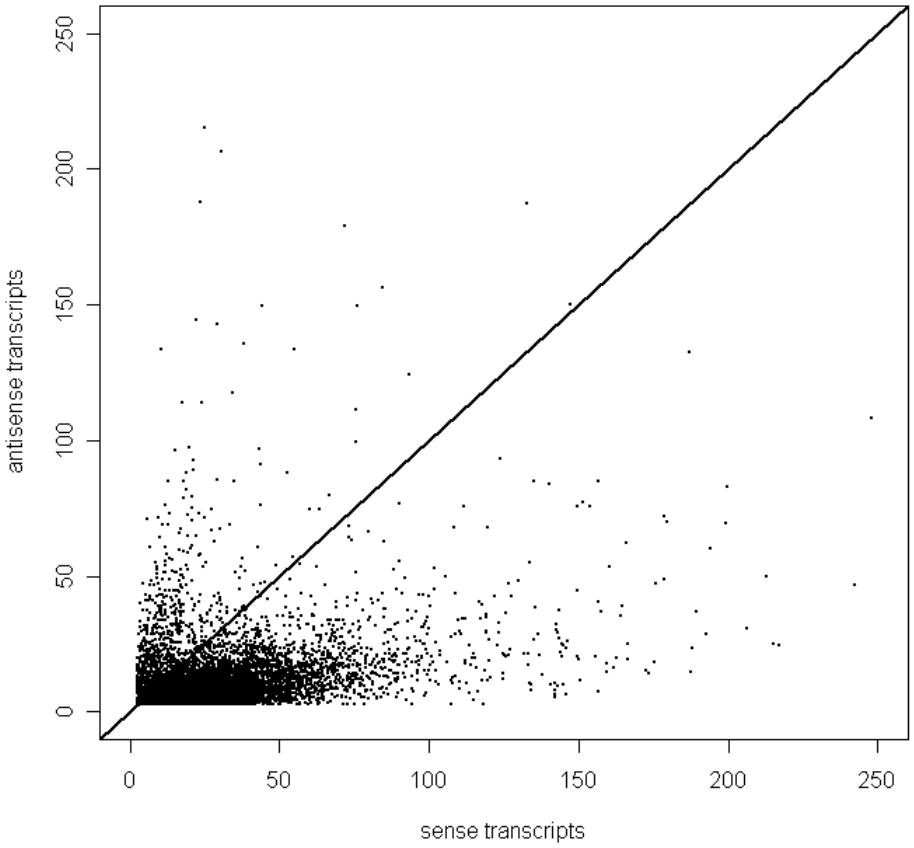
Supplementary Figure 3: Cross-laboratory consistency of Solexa MPSS results

Pooled wild-type samples were analyzed at the site of Illumina, while individual samples were analyzed at the Leiden Genome Technology Center. On the x-axis, we plotted summed scaled and square-root transformed number of counts for all wild-type (left panel) or transgenic (right panel) samples from the Leiden experiment, while on the y-axis the scaled and square-root transformed number of counts in the pool of wild-type (left panel) or transgenic (right panel) samples is plotted.

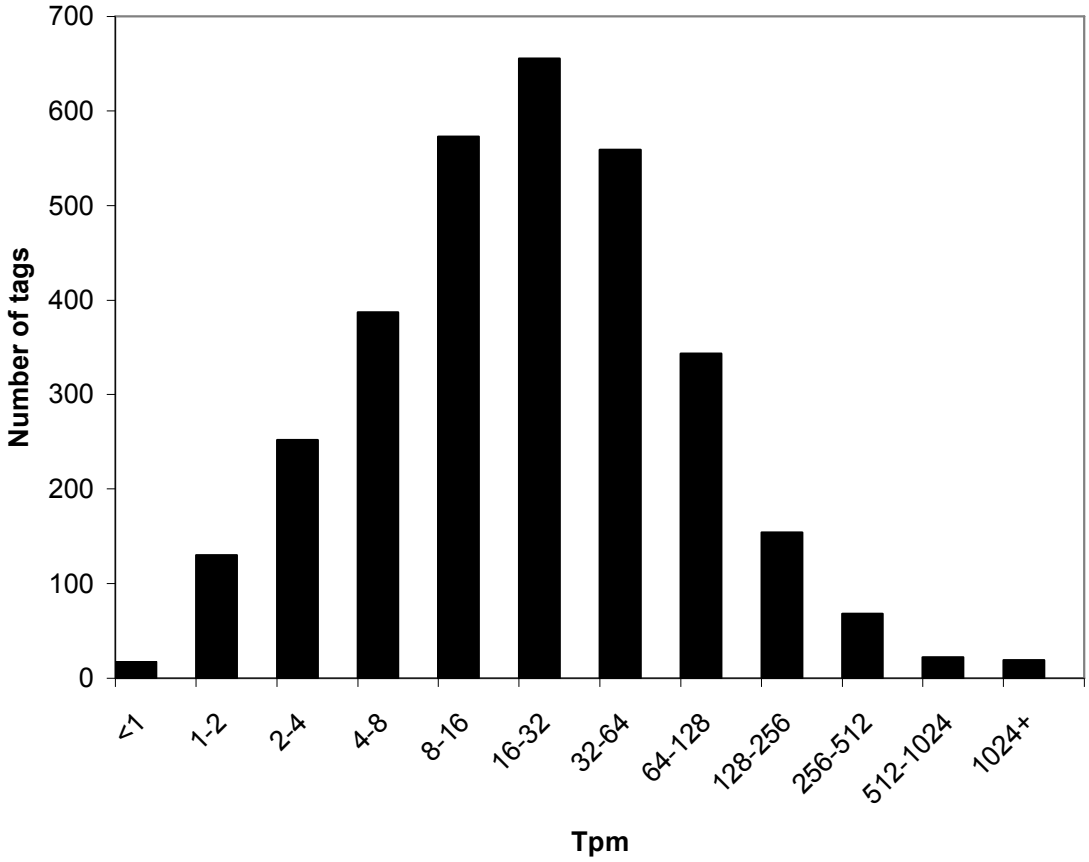


Supplementary Figure 4: Correlation of the abundance of sense and antisense transcripts from the same locus

For each Unigene cluster in which we detected bidirectional transcription at levels >2 tpm (N=9757), all sense tags were summarized into a single value and all antisense tags were summarized into a single value. Subsequently, the scaled and square root transformed abundance of the sense tags (x-axis) was plotted against the scaled and square root transformed abundance of the antisense tags (y-axis). For 1030 Unigene clusters (those left from the diagonal), we found higher abundance of the antisense tag than the sense tag.

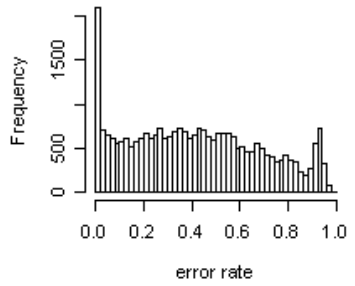


Supplementary Figure 5:
Histogram of the abundance (in transcripts per million, calculated by summing over all samples) of the differentially expressed tags (N=3179). The differentially expressed tags in the two categories with lowest average abundance (between 0.8 and 2 tpm (N=147)) are only detected in one of the groups where they have an average abundance of at least 1.6 tpm.

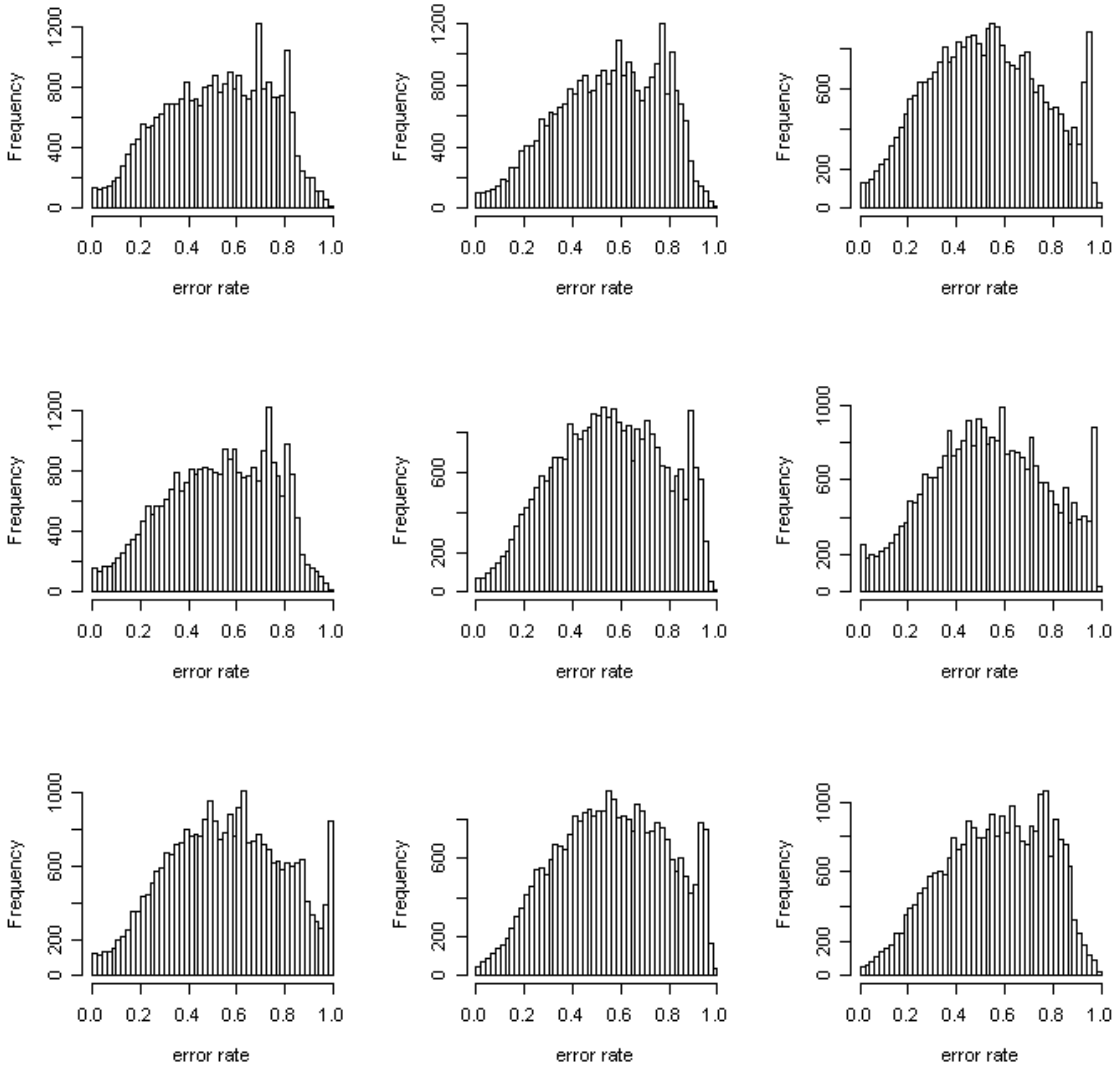


Supplementary Figure 6:
Estimation of the false discovery rate in the list of differentially expressed genes
To estimate the number of false positives in the list of differentially expressed gene obtained with a Bayesian error rate <0.05, we calculated the number of genes below the this error rate between the group of wild-type and transgenic samples (Panel A: true labels), and between all unique pairs of groups, where each group consisted of two wild-type and two transgenic mice (Panel B: permuted labels). We display the distribution of the Bayesian error rates in histograms. A clear difference in the distribution is observed: an extreme enrichment of low Bayesian error rates is observed with the correct labels, whereas with the permutated labels the Bayesian error rates show a nearly Gaussian distribution with a maximum of approximately 0.5.

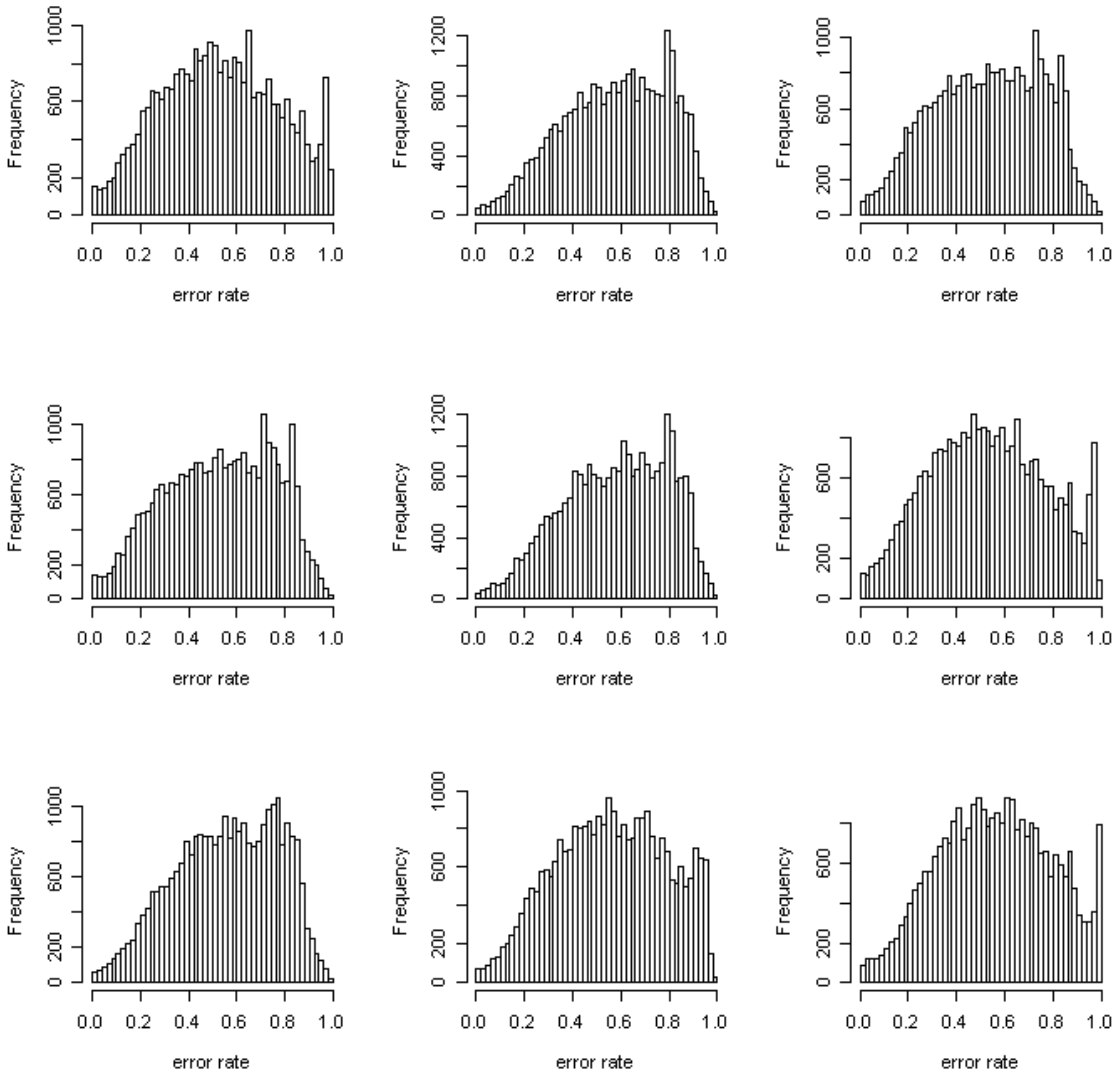
Panel A: true labels



Panel B:
Permutations 1-9



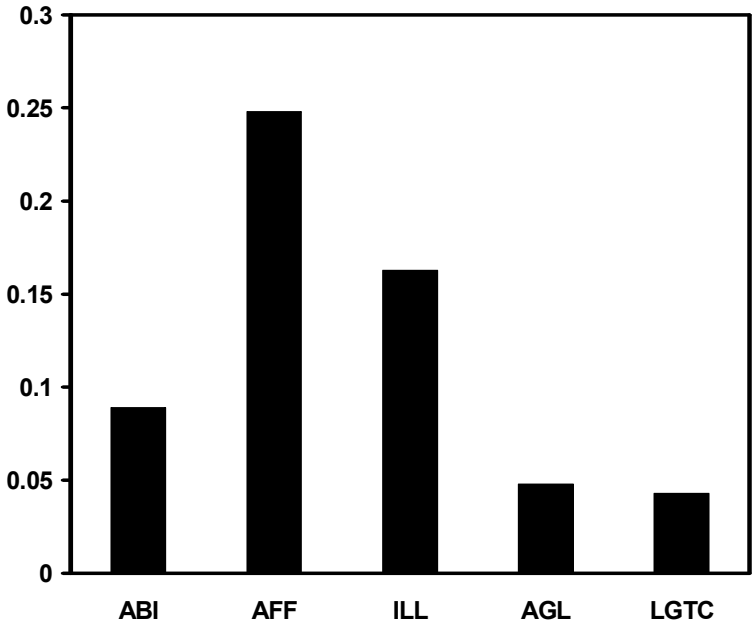
Permutations 10-18



Supplementary Figure 7: Correlation between DGE and microarray ratios

Pearson correlation of the logged ratios of the expression in transgenic vs. wild-type mice obtained by Illumina's Digital Gene Expression tag sequencing assay with those obtained with the different microarray platforms.

ABI: Applied Biosystems; AFF: Affymetrix; ILL: Illumina; AGL: Agilent; LGTC: home-spotted long oligonucleotide arrays.



Supplementary Figure 8:

The genomic location of the 3'-end of the *Gabra2* gene on mouse chromosome 5. The *Gabra2* gene lies on the negative strand and the position of the last exon (exon 10) is indicated. Also indicated the location of microarray probe sequences (AFF=Affymetrix; AGL=Agilent, ILL=Illumina, ABI=ABI, SIGMA=Sigma-Compugen LGTC home-spotted arrays), DGE tags (with strand indication). qPCR primers detecting *Gabra2* expression were designed in exon 9 (forward) and exon 10 (reverse).

Supplementary Figure 9:
Nucleotide composition of sequenced tags depends on transcript abundance

All unique tags obtained were ranked according to their abundance and divided in bins of thousands (*i.e.* set 1 contains the most abundant tags (tags ranked 1-1000), set 230 contains the least abundant tags). Within such a set of thousand tags, the average percentage of A, T, C, and G in each tag is calculated (excluding the CATG recognition sequence). These percentages are plotted for each set of thousand tags, where the set with the highest abundance (set 1) is plotted on the left (index 1, x-axis), and the set with the lowest abundance (index 230, x-axis) is plotted on the right.