Decode-seq: a practical approach to improve differential gene expression analysis

Supplementary Figures

Yingshu Li[1,2,3], Hang Yang[1,2,3], Hujun Zhang[1,2,3], Yongjie Liu[1,2,3], Hanqiao Shang[1,2], Herong Zhao[1,2,3], Ting Zhang[1,2], and Qiang Tu[1,2,3,*]

[1] *State Key Laboratory for Molecular and Developmental Biology, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China*

[2] *Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China.*

[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

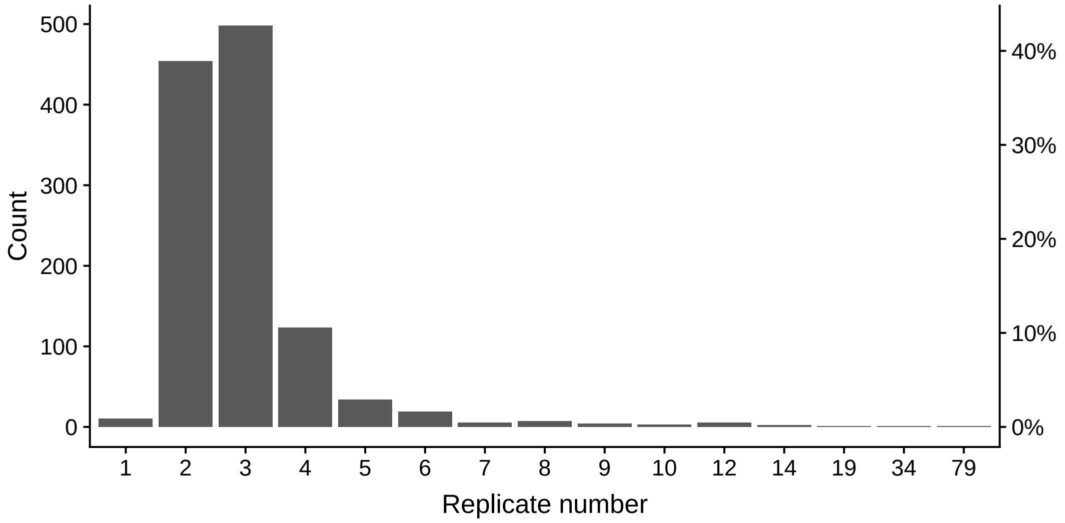

Fig S1: Distribution of replicate numbers employed in surveyed GEO studies. From NCBI GEO (Gene Expression Omnibus) database, we extracted 1,167 expression profiling studies which cited edgeR or DESeq2 and named the associated samples like 'rep1' or 'replicate 1' etc. We used these sample names to estimate the replicate numbers employed in these studies. Among them, 39% used only 2 replicates, 43% used 3 replicates, while only 18% used 4 or more replicates.

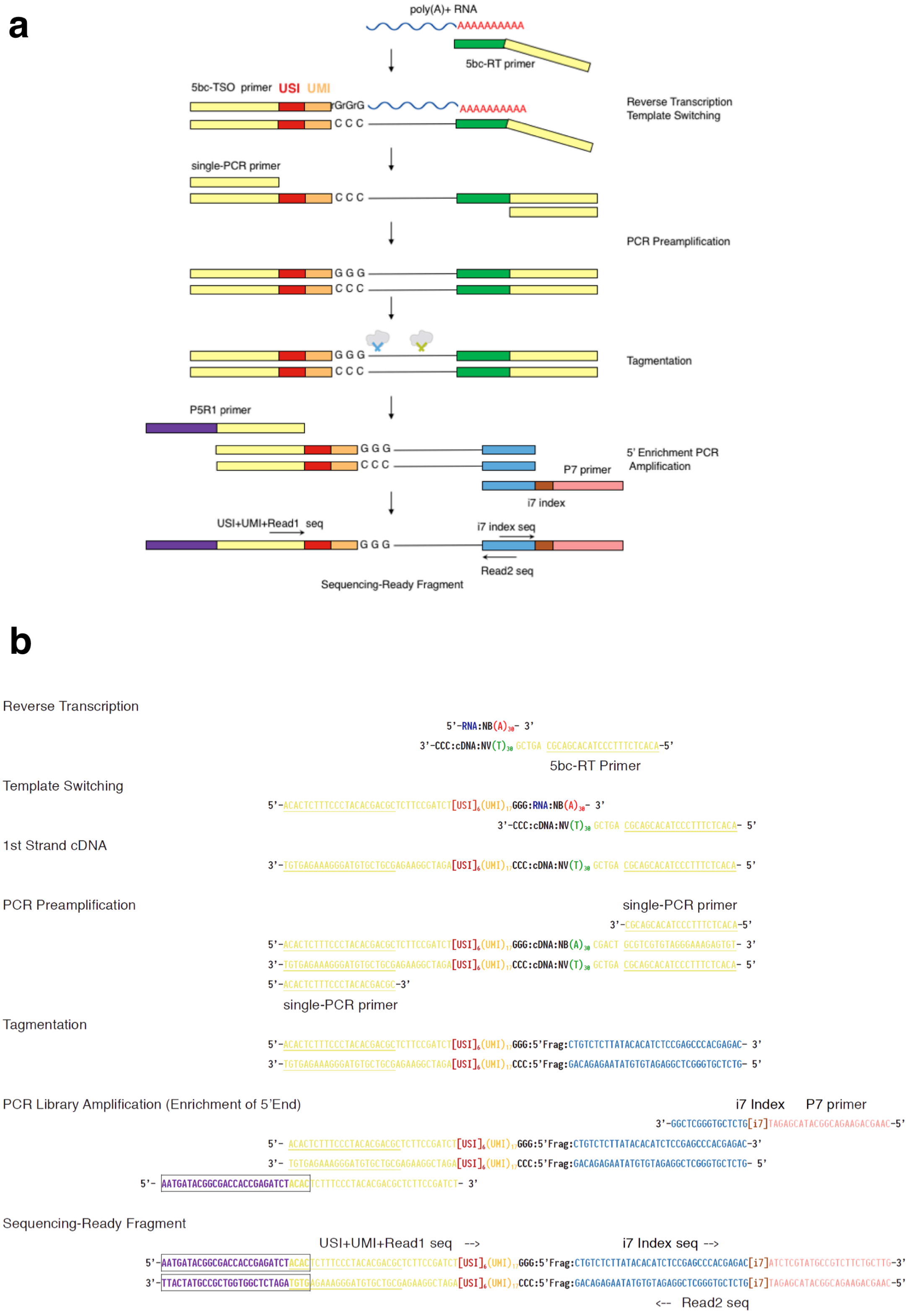
a
poly(A)+ RNA
AAAAAAAAAA
5bc-RT primer
5bc-TSO primer USI UMI
rGrGrG
AAAAAAAAAA
C C C
Reverse Transcription
Template Switching
single-PCR primer
C C C
PCR Preamplification
G G G
C C C
G G G
C C C
Tagmentation
P5R1 primer
G G G
C C C
5’ Enrichment PCR
Amplification
P7 primer
i7 index
USI+UMI+Read1 seq
i7 index seq
G G G
Read2 seq
Sequencing-Ready Fragment
b
Reverse Transcription
5’-RNA:NB(A)$_{30}$- 3’
3’-CCC:cDNA:NV(T)$_{30}$ GCTGA CGCAGCACATCCCTTTCTCACA-5’
5bc-RT Primer
Template Switching
5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCT[USI]$_{6}$(UMI)$_{17}$GGG:RNA:NB(A)$_{30}$- 3’
3’-CCC:cDNA:NV(T)$_{30}$ GCTGA CGCAGCACATCCCTTTCTCACA- 5’
1st Strand cDNA
3’-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[USI]$_{6}$(UMI)$_{17}$CCC:cDNA:NV(T)$_{30}$ GCTGA CGCAGCACATCCCTTTCTCACA- 5’
PCR Preamplification
single-PCR primer
3’-CGCAGCACATCCCTTTCTCACA-5’
5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCT[USI]$_{6}$(UMI)$_{17}$GGG:cDNA:NB(A)$_{30}$ CGACT GCGTCGTGTAGGGAAAGAGTGT- 3’
3’-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[USI]$_{6}$(UMI)$_{17}$CCC:cDNA:NV(T)$_{30}$ GCTGA CGCAGCACATCCCTTTCTCACA- 5’
5’-ACACTCTTTCCCTACACGACGC-3’
single-PCR primer
Tagmentation
5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCT[USI]$_{6}$(UMI)$_{17}$GGG:5’Frag:CTGTCTCTTATACACATCTCCGAGCCCACGAGAC- 3’
3’-TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[USI]$_{6}$(UMI)$_{17}$CCC:5’Frag:GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG- 5’
PCR Library Amplification (Enrichment of 5’End)
i7 Index
P7 primer
3’-GGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC-5’
5’- ACACTCTTTCCCTACACGACGCTCTTCCGATCT[USI]$_{6}$(UMI)$_{17}$GGG:5’Frag:CTGTCTCTTATACACATCTCCGAGCCCACGAGAC-3’
3’- TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[USI]$_{6}$(UMI)$_{17}$CCC:5’Frag:GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG- 5’
5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT- 3’
Sequencing-Ready Fragment
USI+UMI+Read1 seq -->
i7 Index seq -->
5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT[USI]$_{6}$(UMI)$_{17}$GGG:5’Frag:CTGTCTCTTATACACATCTCCGAGCCCACGAGAC[i7]ATCTCGTATGCCGTCTTCTGCTTG-3’
3’-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA[USI]$_{6}$(UMI)$_{17}$CCC:5’Frag:GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG[i7]TAGAGCATACGGCAGAAGACGAAC-5’
<-- Read2 seq

Fig S2: Design of Decode-seq. (a) Overview of Decode-seq workflow. (b) Detailed sequences used in each step.

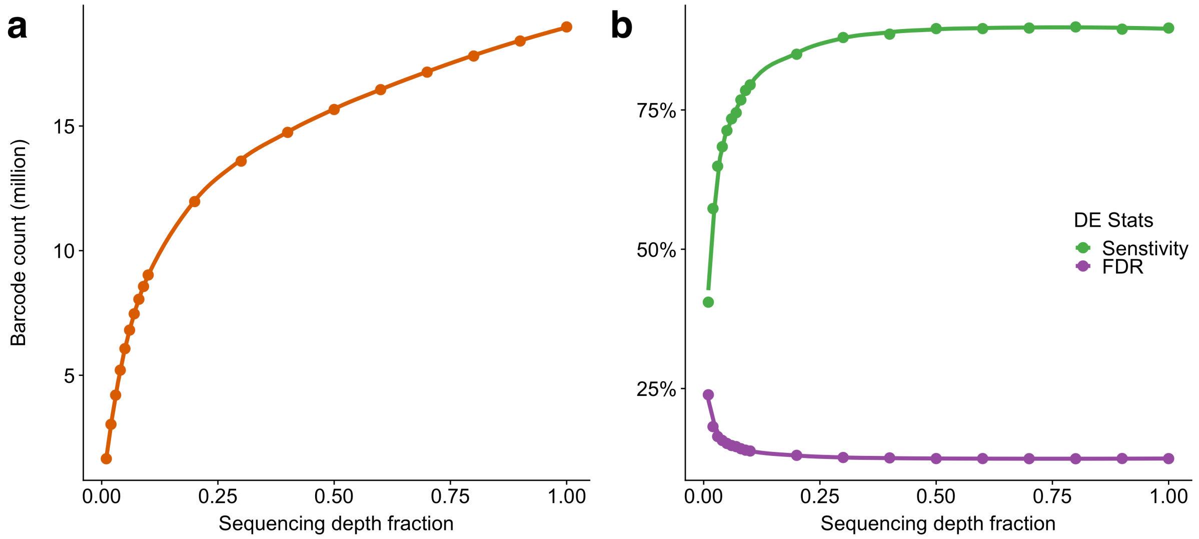

Fig S3: Downsampling of sequencing depth. From reads generated from the human/mouse control experiment, a fraction of reads was randomly selected and used in downstream analysis. The results showed that the sequencing depth used in this experiment was close to saturation. (a) Numbers of barcodes detected from a fraction of reads. (b) Differential expression analysis stats (sensitivity and false discovery rate) calculated from a fraction of reads.

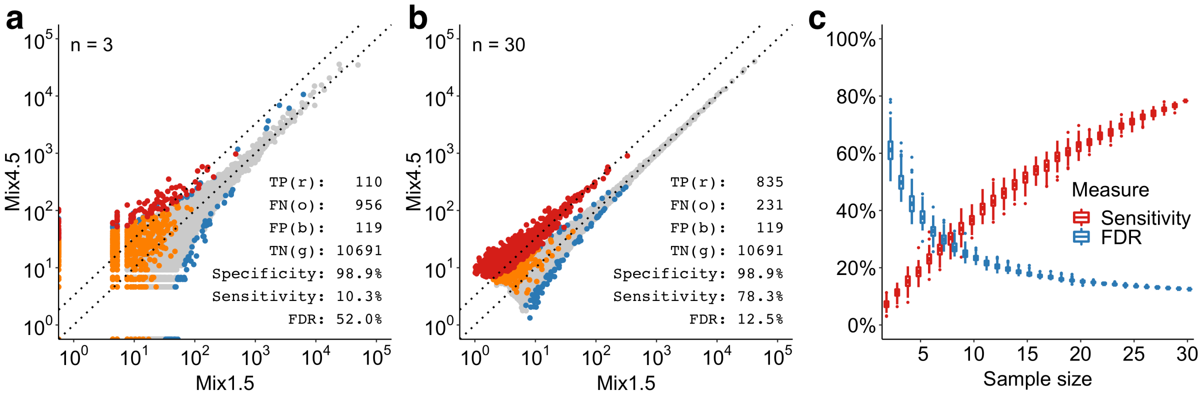

Fig S4: Performance evaluation of Decode-seq at 3-fold change level using human/mouse RNA mixes. (a) Differential expression analysis with 3 pairs of replicates. True positive (TP, red dots): mouse genes which were called DE; false negative (FN, orange dots): mouse genes which were called non-DE; true negative (TN, gray dots): human genes which were called non-DE; false positive (FP, blue dots): human genes which were called DE. Specificity = TN/(TN+FP), and it was fixed to 98.9% in all calculation. Sensitivity = TP/(TP+FN). False discovery rate = FP/(TP+FP). The sensitivity was only 10.3% and the false discovery rate was 52%. (b) DE analysis with 30 pairs of replicates. The sensitivity increased to 78.3% and the false discovery rate dropped to 12.5%. (c) DE performance related to replicate number calculated by random downsampling of 30-pair data. Each replicate number was calculated 100 times. Sensitivity and false discovery rate were improved dramatically when the number of replicates increased.

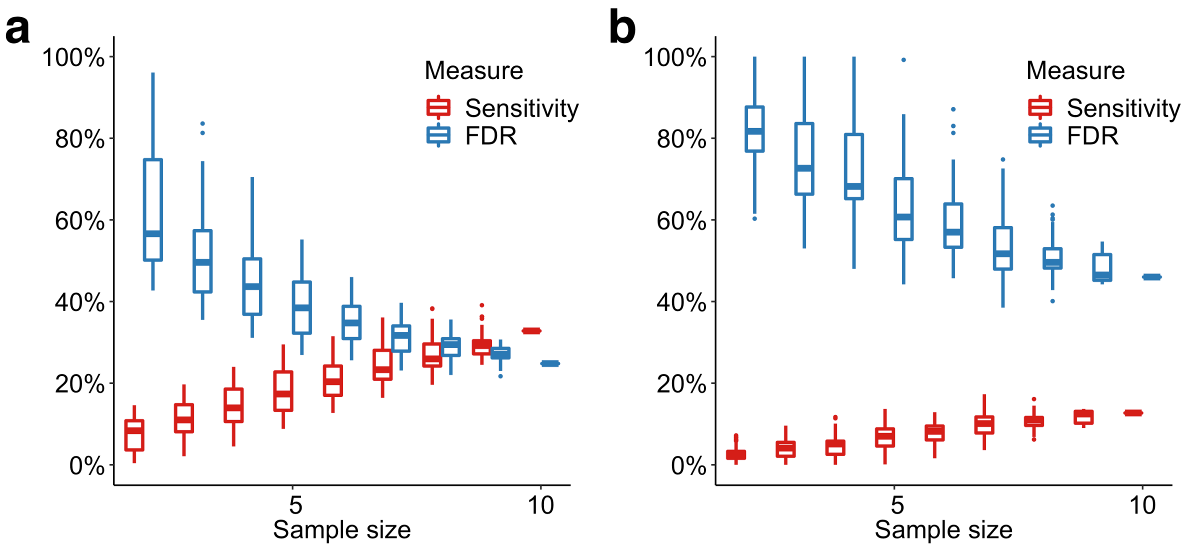

Fig S5: Performance evaluation of Decode-seq with 10 ng and 1 ng RNA at 5-fold change level. 10 pairs of replicates were used in these experiments. The performance was not as good as 100 ng, but the trend was the same, more replicates improved DE analysis. (a) Starting material: 10 ng. (b) Starting material: 1 ng.

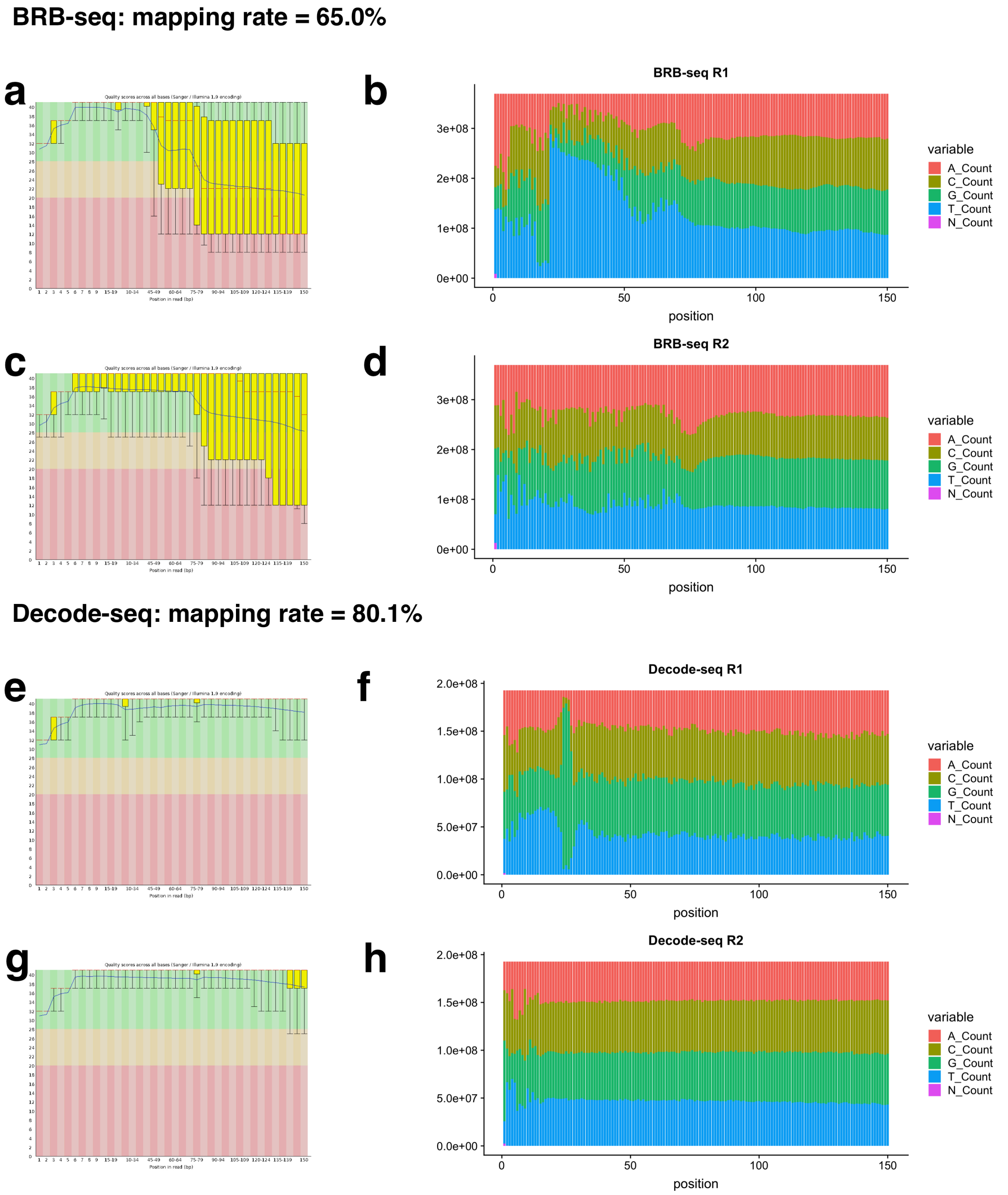

Fig S6: Sequencing quality scores and nucleotide distribution of BRB-seq and Decode-seq. a,c,e,g represent the per base sequence quality of read 1 and read 2 of BRB-seq and Decode-seq. b,d,f,h represent the nucleotide distribution of read 1 and read 2 of BRB-seq and Decode-seq. For BRB-seq, position 1–6 is the USI barcode, position 7–21 is the UMI barcode, and the rest is the cDNA sequence. For Decode-seq, position 1–6 is the USI barcode, position 7–23 is the UMI barcode, position 24–26 is the three Guanines, and the rest is the 5' end cDNA sequence.

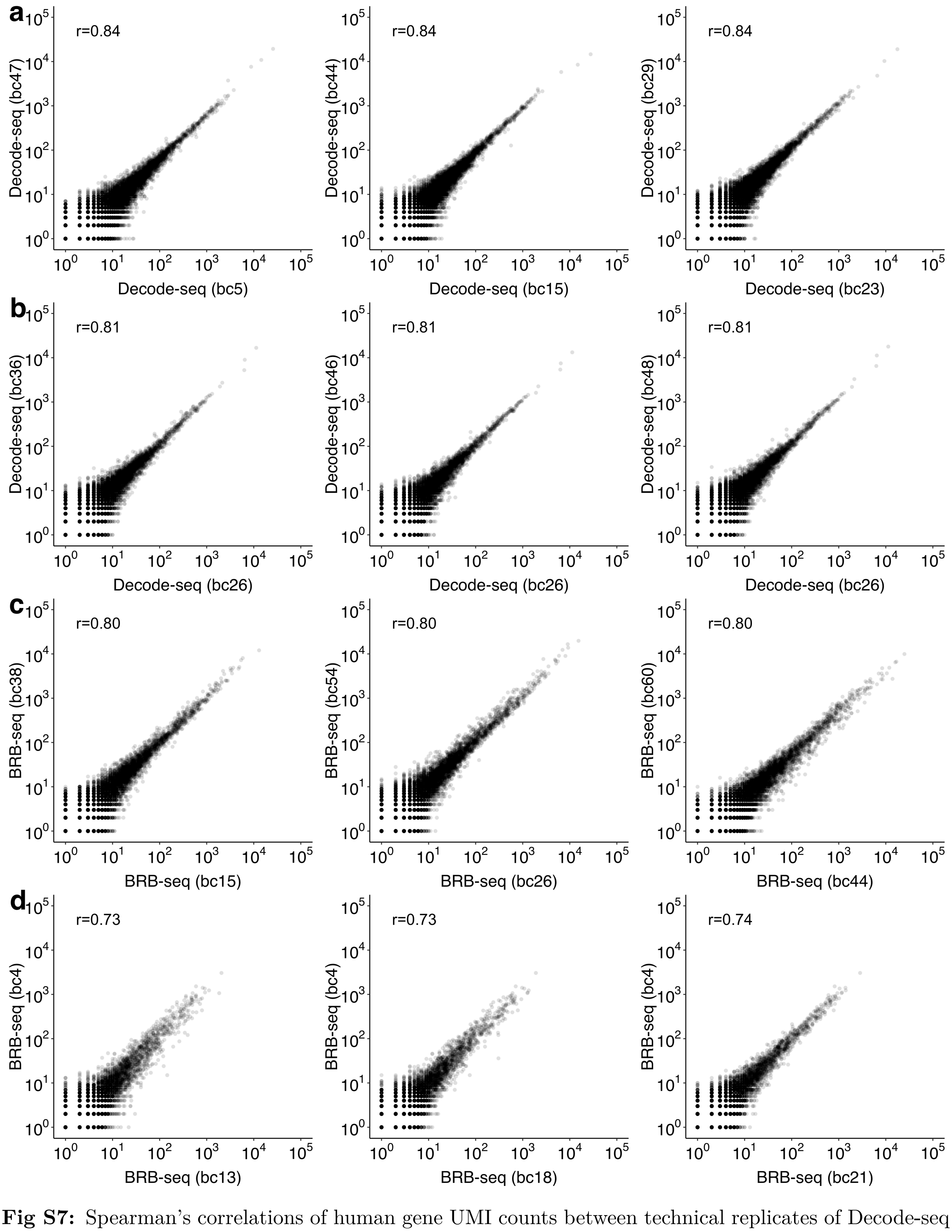

Fig S7: Spearman's correlations of human gene UMI counts between technical replicates of Decode-seq and BRB-seq. (a),(b) Correlation of Decode-seq replicates with three pairs of mean correlation and three of the worst correlation. (c),(d)Correlation of BRB-seq replicates with three pairs of mean correlation and three of the worst correlation.

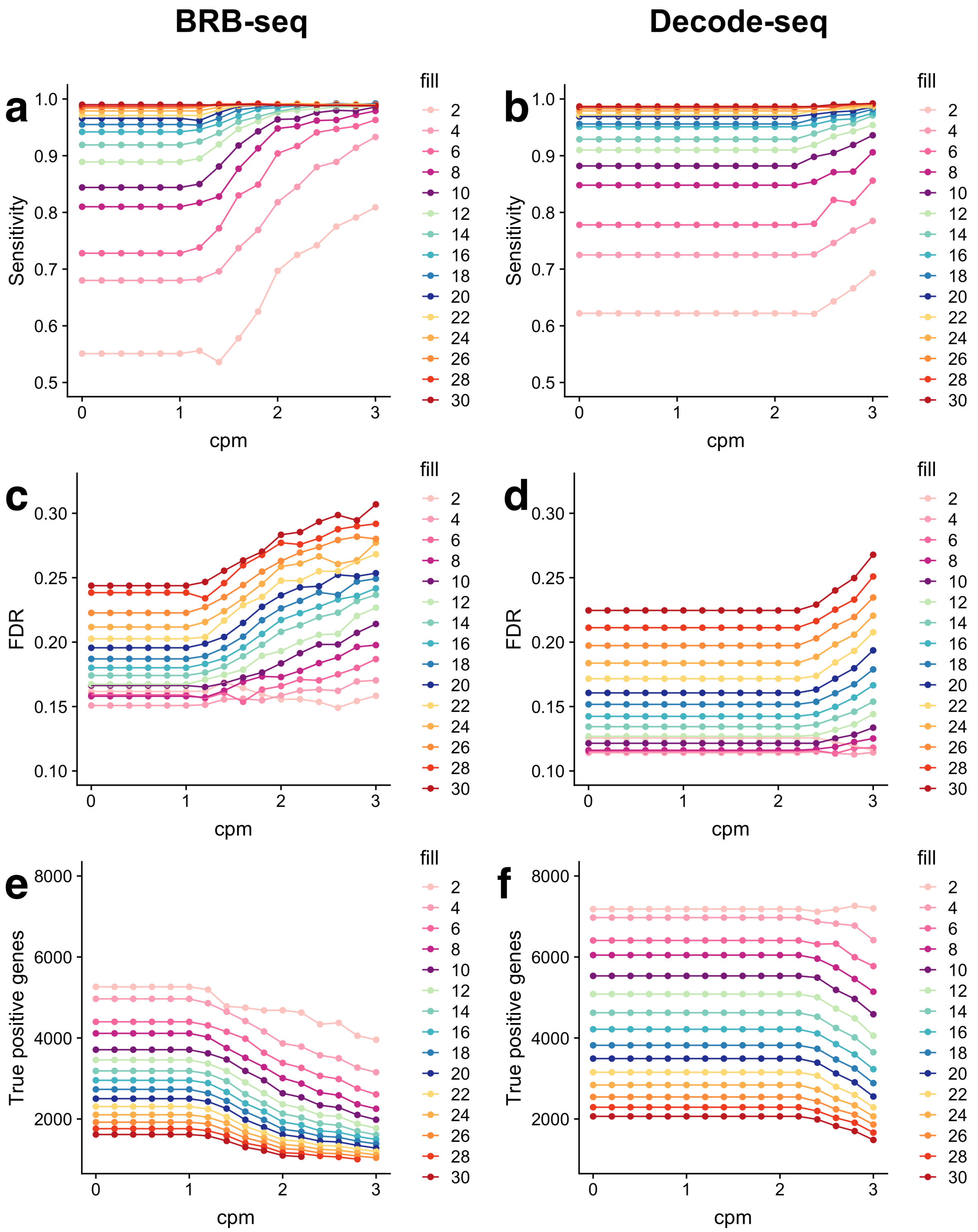

Fig S8: DE performance of BRB-seq and Decode-seq when using different edgeR filtering parameters. The two parameters (cpm and fill) means a gene will be retained if it is only expressed more than a CPM value(cpm) in more than a particular number (fill) of samples in each group. Using various combinations of cpm and fill, the sensitivity, FDR and true positive gene number is calculated.

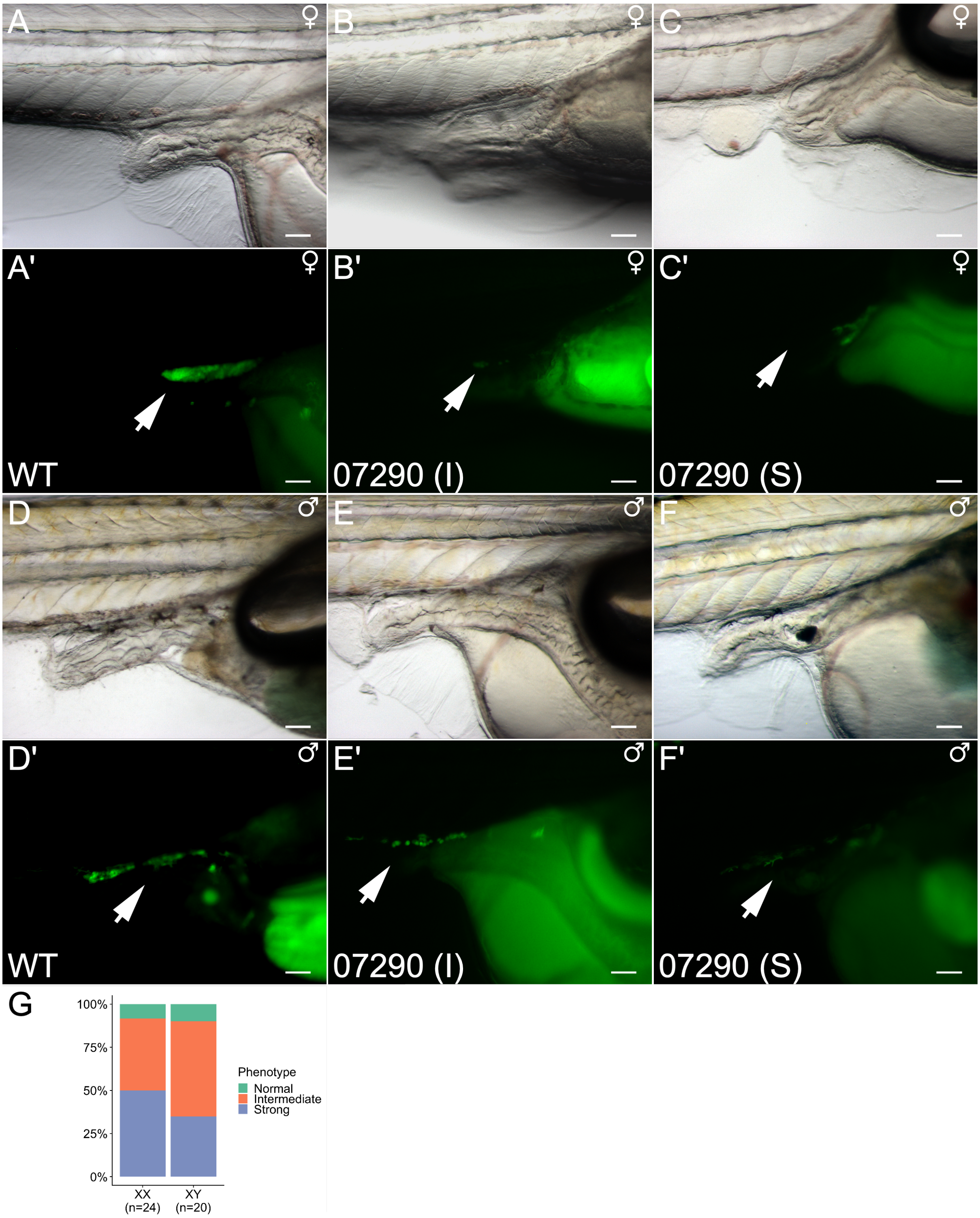

Fig S9: Genetic knockout of ENSORLG00000007290 by four-guide Cas9 RNP. Medaka fry at stage 39 (just hatching) were collected, imaged and scored. Arrows indicate GFP labeled germ cells. The phenotype of mutants was classified into three categories: normal, intermediate, and strong. Scale bar: 100 µm. (A-F) Bright-field images. (A'-F') Fluorescent images of the same specimen. (A-C) Female fry. (D-F) Male fry. (A/D) Wild type fry, clusters of germ cells with GFP signals are visible, and the female carries stronger signal than male. (B/E) Intermediate phenotype, in which germ cells are largely depleted. (C/F) Strong phenotype, in which germ cells are almost completely disappeared. (G) Percentage of three different phenotypes in male and female.