Comparative analysis reveals genomic features of stress-induced transcriptional readthrough

Anna Vilborg[1*], Niv Sabath[2], Yuval Wiesel[2], Jenny Nathans[1], Flonia Levy-Adam[2], Therese A. Yario[1], Joan A. Steitz[1] and Reut Shalgi[2*]

1 Department of Molecular Biophysics and Biochemistry, Howard Hughes Medical Institute, Boyer Center for Molecular Medicine, Yale University School of Medicine, 295 Congress Avenue, New Haven, CT 06536, USA

2 Department of Biochemistry, Rappaport Faculty of Medicine, Technion – Israel Institute of Technology, Haifa 31096, Israel

* To whom correspondence should be addressed. Email: reutshalgi@technion.ac.il, vilborg.anna@gmail.com

Supplementary Information Appendix

Materials and Methods

Cell culture, transfections and treatments

NIH3T3 (mouse embryonic fibroblasts) cells were cultured in a 1:1 mix of Dulbecco's Modified Eagle Medium and F-12 Nutrient's mixture (both Gibco) supplemented by 10% Fetal Bovine Serum (Atlanta Biologicals), 1% L-glutamine (Gibco) and 1% Penicillin/Streptomycin (Sigma-Aldrich). For transfections, media without antibiotics were used. siRNAs were purchased from Qiagen (accession numbers are found in Table S3) and were used at 20 nM final concentration. Transfections were performed using RNAiMAX (Invitrogen) according to the manufacturer's recommendations on NIH3T3 cells plated at 55,000 cells/well in 6-well plates and cultured for 24 h. For stress treatments, cells were cultured at 44°C (heat shock), or in the presence of 200 mM KCl (osmotic stress), or of 0.2 mM H_2O_2 (oxidative stress) for 2 h if not otherwise stated. Unless otherwise stated, NIH3T3 cells were plated at 500,000 cells/well in 6-well plates and cultured for 24 h before treatment. For pretreatments with inhibitors, cells were treated with 50 mM BAPTA (Abcam) or 100 μM 2-ABP (Sigma-Aldrich) for 30 min before stress treatment, or with 0.05 ng/ml cycloheximide (CHX) for 15 min before stress treatment.

RNA-preparation and qRT-PCR

All primers are listed in Supplementary Table S3.

If not otherwise stated, total RNA was used and was prepared by harvesting cells in Trizol (Ambion) following the manufacturer's instructions. For preparations of nuclear-enriched RNA, cells were grown to confluency on 15 cm plates (Falcon), treated with

stress, washed in PBS, scraped in PBS, pelleted by centrifugation at 3000 x g and resuspended in 200 µl of RLB buffer (10 mM Tris, pH 7.5, 140 mM NaCl, 1.5 mM $MgCl_2$, 0.5% Nonidet P-40, 1mM DTT and 100U/ml RNaseIN) (Roche). Samples were incubated on ice for 5 min and then layered over 600 µl of RLB buffer containing 24% (wt/vol) sucrose. Samples were centrifuged at 13,000 x g for 10 min at 4°C. After centrifugation, supernatants (cytoplasmic extracts) were discarded and nuclear pellets were resuspended in 200 µl buffer with sucrose, and RNA was extracted by the addition of 1 ml Trizol. RNA (total or nuclear) was then treated with DNase RQ (Promega) using 0.5-1 units DNase/µg RNA following the manufacturer's instructions. RNA was recovered by phenol/chloroform/isoamyl alcohol (PCA) extraction and ethanol precipitation, following standard protocols. For sequencing, RNA was subjected to additional rounds of purification by Qiagen RNaEasy kit followed by PCA. RNA concentrations were analyzed on a NanoDrop 8000 (Thermo Scientific) and 1-4 µg was used for cDNA synthesis using SuperScript III and random primers (Invitrogen) according to the manufacturer's protocol. Minus-RT controls were included and analyzed alongside other samples to ensure efficient DNase treatment. qPCR was performed on a Light Cycler 96 instrument (Roche) with FastStart Essential DNA Green Master Mix (Roche) according to the manufacturer's recommendation. Relative RNA levels were calculated by the ΔΔCT method and targets of interest were normalized to the mean value of two control genes (*GAPDH* and *18S*). All primer pairs used for qRT-PCR were validated by a) agarose gel electrophoresis of the qRT-PCR products to ensure expected product size and absence of other amplification products and b) sequencing of the qRT-PCR product to ensure amplification of the expected target. For sequencing, RNA was analyzed on agarose gels to verify the quality of the RNA.

Library preparation and sequencing

RNA quality was assessed by NanoDrop 8000 (Thermo Scientific) and agarose gel electrophoresis. RNA quality was further established at the Yale Center for Genome Analysis (YCGA) by running an Agilent Bioanalyzer gel. Ribosomal RNA was then depleted using the Ribo-Zero Gold Kit (MRZG12324 Epicentre) and remaining RNA fragmented by incubation at 94°C. Strand-specific libraries were prepared using Illumina reagents. Following first-strand synthesis with random primers, second-strand synthesis was performed with dUTP for generating strand-specific sequencing libraries. The cDNA library was end-repaired and A-tailed, adapters were ligated and second-strand digestion was performed by Uricil-DNA-Glycosylase. Quality of indexed libraries was validated by quantification by qRT-PCR (KAPA Biosystems) and by insert size distribution determination (LabChip GX). Two samples were multiplexed per sequencing lane, and samples were sequenced using 75 bp paired-end sequencing on an Illumina HiSeq 2000 according to Illumina protocols. Signal intensities were converted to individual base calls during the run using the system's Real-Time Analysis (RTA) software. Sample de-multiplexing and splitting of the paired-end reads were done using Illumina's CASAVA 1.8.2 software suite.

Analysis of readthrough in published RNA-seq datasets

We re-analyzed our previously published RNA-sequencing data from heat-shock-treated NIH3T3 mouse fibroblasts (1). As this dataset was generated from polyA-selected RNA, and we previously observed DoGs to be only partially polyadenylated (2), we expected the signal to be lower than observed in our previous study. Thus, we examined the read density in the 5 kb immediately downstream of all known gene ends, and normalized it to the read density in the corresponding upstream gene (normalized readthrough, Fig. S1a). Comparing the normalized readthrough in heat shock (2 h at 44^{o}C) to control cells showed a marked shift above the diagonal, demonstrating a widespread readthrough

induction. A similar analysis was performed for a dataset of H_2O_2 treatment in human fibroblasts (3), revealing the same trend (Fig. S1b).

DoG Discovery pipeline

Data preprocessing

1. For every treatment, we aligned the resulting paired-end reads using Tophat2 (4). Since readthrough detection varies with library depth, it was important that the number of aligned reads in every treatment be approximately equal. Thus, we down-sampled Tophat output bam files using samtools so that the number of reads in every bam file in each of the biological replicate experiments was equal.

2. To use the paired-end reads correctly, we split the bam files into two separate bam files according to the DNA strand they aligned to, positive or negative. We did this by splitting sam files according to the flag field values.

3. In the discovery pipeline it was crucial to include all known annotated gene regions to avoid false discovery of UTRs (untranslated regions), which could be annotated in some but not other databases, as DoGs. To achieve this, we downloaded from the UCSC database mm9 gene annotations using three different tracks: Ensembl genes, UCSC genes and RefSeq genes. To construct a global annotation file, which we termed gene loci annotation, we included in each track the official symbol for each gene.

4. We then generated a unified gene loci annotation file by combining the annotations of all transcripts belonging to each gene from all the databases, according to their official symbol and overlapping genomic positions. For our gene loci annotation file, we noted the minimal start coordinate and the maximal end coordinate out of all transcripts belonging to a gene. We used this annotation file (loci annotation file) for all subsequent steps of the pipeline.

5. The next step was to filter out all reads mapping to annotated genes in order not to assign these to readthrough transcripts. We excluded these reads by using bedtools intersect -v with the loci annotation file and the positive strand or negative strand bam files (separately), resulting in new bed files with only non-genic reads.

DoG discovery

The non-genic read bed file and the loci annotation file were used as inputs.

1. In order to identify genes with potential readthrough we first generated a list of DoG candidates. First, a new annotation file was generated using bedtools flank to move all loci annotation 4 kb downstream of gene ends.

2. We then used bedtools coverage to calculate read coverage of the 4 kb downstream annotation and filtered out annotations that had less than 80% coverage in this area. We kept DoG candidates that had at least 80% coverage over the first 4 kb downstream of gene ends.

3. To find the DoG end positions, we elongated the DoG ends in moving windows of 200 bases requiring at least 80% coverage of the entire region downstream to the gene 3′ end. When coverage was below 80%, or when DoG coordinates overlapped the coordinates of a neighboring gene, DoG elongation was terminated. We ran the pipeline for every treatment, resulting in a DoG annotation file for each treatment.

4. To generate a single DoG annotation file, we combined the DoG annotations from all treatments and created two new annotation files: “Shortest DoG annotations,” which used the most proximal DoG endpoint found in any one individual DoG annotation file, and “Longest DoG annotations,” which used the most distal DoG endpoint found in any one individual DoG annotation file.

5. We then used these two annotation files to calculate two types of RPKMs (Reads Per Kilobase per Million mapped reads): “RPKM Short” – for the “Shortest” annotation (see

Fig. S1f) and “RPKM Long” for the Longest annotation (Fig. 1d) for each treatment. Read counts were obtained using bedtools coverage in each condition.

DoG and gene filtering

To calculate gene expression levels, we ran Cufflinks (4) with the union of all three gene annotation files described above (data preprocessing, step 3), and used the gene expression estimated by Cufflinks (a combination of all transcripts belonging to the same gene locus). To make sure we did not discover DoGs that are in fact known isoforms or overlap known lncRNAs, we initially used the combined annotation loci to eliminate all known annotated transcripts from the DoG discovery pipeline. However, for all subsequent analyses we excluded genes and DoGs derived from genes that were shorter than 200 bp, expressed at levels lower than RPKM of 4 in control conditions, or did not overlap any RefSeq gene.

Robustness of pipeline parameter choice

To ensure that our results are not sensitive to the choice of the pipeline parameters, we ran the discovery pipeline with several input parameters, and performed all subsequent analyses presented in this study. We re-ran the pipeline with a minimal DoG length of 4.5 and 5 kb. We also tested an alternative approach in which the elongating window is required to have at least 20% coverage without the requirement for 80% coverage over the entire DoG region, and applied this approach using minimal DoG lengths of 4.5 and 5 kb. For all parameter choices, the pipeline resulted in very similar DoG sets affecting up to ~7% of our pan-stress annotated DoGs. Importantly, all subsequent results remained the same.

We note that DoG ends are dependent on sequencing depth. Therefore, the different parameter choices resulted in slightly varying DoG ends, as did the two

biological replicate experiments, which were different in sequencing depth. Nevertheless, all of the subsequent results presented in the paper were the same in all cases.

Hierarchical clustering analysis

Hierarchical clustering analysis was performed using average linkage hierarchical clustering in MATLAB on DoG readthrough scores: the log2 ratio of readthrough transcription (DoG RPKM / gene RPKM) in each stress condition, normalized to that of the control), calculated based on the long DoG annotations. To increase the distinction between clusters, we included only DoGs with 2-fold readthrough score change in at least one treatment. We then used three clusters with stress-specific high readthrough levels (Fig. 1e, green, red, and turquoise clusters) to examine the sequence composition of stress-specific DoGs (see below).

Comparing pan-stress and highly-induced DoGs

We defined sets of highly-induced DoGs, as those with readthrough score log2 fold changes 2 standard deviations above the mean of each stress condition distribution, resulting in 68, 158 and 115 in heat shock, osmotic and oxidative stress, respectively. We then used the hypergeometric test to compare the intersection between pan-stress DoGs (total of 1556) and highly-induced DoGs. The number of highly-induced pan-stress DoGs is 37 (54.4%) in heat shock, 37 (32.2%) in oxidative stress, and 24 (15.2% out of the highly-induced DoGs) in osmotic stress. In all conditions, highly-induced DoGs are significantly depleted from pan-stress DoGs (p-value=$1.5*10^{-8}$ in heat shock and osmotic stress and 0.001 in oxidative stress).

Single molecule RNA FISH

We custom-designed Stellaris probes for doHnrnpa2b1 and doHspa8 (using the Biosearch design tools for the 9 kb downstream of the end of Hnrnpa2b1 and 12 kb downstream of Hspa8 as input) labeled with Quasar 570, and custom-designed Stellaris probes for intron 1 of Hnrnpa2b1 (using the Biosearch design tools and the entire sequence of Hnrnpa2b1 intron 1 input) labeled with FAM (Biosearch Technologies). FISH was performed on NIH3T3 cells either untreated or treated with 200 mM KCl or heat shock at 44°C for 2 h prior to fixation using standard Stellaris protocols (Biosearch Technologies). FISH staining was visualized on a ZEISS Axiovert 200 inverted fluorescent microscope with appropriate filters and processed with Fiji.

Ribosome footprint density analysis

To examine the possibility that DoGs exit the nucleus, we used ribosomal profiling data of heat shock and control samples from Shalgi et al. (5) to search for potential evidence of translation in the first 4 kb of heat-shock DoGs using bedtools coverage. For ~80% of the heat-shock DoGs (1492), we did not find any footprint reads. We found 379 DoGs with non-zero footprints, but only four show RPKM greater than one. Manual examination of these four regions showed that in all cases, footprint reads mapped to transposons embedded within the DoG region.

Pan-stress DoG and non-DoG groups

In several analyses, we characterized DoG properties through comparison with genes that show no sign of readthrough. We thus defined two groups: pan-stress-DoGs – a group of DoGs that were discovered in all three stress conditions; and non-DoGs – a group of transcripts with very low readthrough levels in all conditions. Specifically, we included in the non-DoG group only genes in which the RPKM of the 4 kb downstream region (the maximum of the three stresses) was lower than the minimal RPKM of pan-

stress DoGs (Fig. S7a, left panel). Because the pan-stress- and non-DoG groups differ in size and gene expression (see Fig. S7a right panel, RPKMs of DoG and non-DoG-associated genes in untreated cells), we constructed equal size expression-matched subsets of pan-stress-DoGs and non-DoG by randomly sampling an equal number of genes in each group from bins of equal expression in the control (Fig. S7a). We repeated the sampling procedure many times to obtain a representative expression-matched comparison. This approach was used in Figures 4a, 6, 7c-d, S7b-c, and S11c-e.

Pol-II genome-wide occupancy (PRO-seq) analyses

To explore the landscape of RNA polymerases in readthrough regions before and after heat shock, we examined a recently published high-resolution dataset of Pol-II genome-wide occupancy (PRO-seq) performed in MEFs (6). The data included normalized reads mapped to mm10 in BED format, with two replicates of heat shock (1 h) and untreated cells, which were obtained from DB Mahat from the Lis lab. We first generated the mapped data in mm9 using UCSC liftOver, and performed down-sampling for each replicate using samtools to obtain similar depth files for control and heat shock. We then quantified Pol-II occupancy in the different regions using bedtools coverage. For each dataset, we estimated RPKM values for each gene along four regions: (1) gene body, using bedtools coverage with the entire gene locus from our loci annotation file as input (containing all exons and introns), (2) 5 kb downstream of gene end (according to the loci annotation file described above), (3) 5-10 kb downstream of gene end, (4) 10-15 kb downstream of gene end. To avoid promoters of neighboring genes, if a region overlapped the 1 kb upstream of the same-strand downstream neighboring gene, the overlapping part was trimmed. A threshold of 0.01 was set for all RPKM values. We compared the mean RPKM values between pan-stress- and non-DoGs using 1000 expression-matched sub-samples generated through the same sampling procedure

described above. To ensure that results are not biased by the differences between NIH3T3 and MEF cells, we conducted the expression-matched comparison three times: (1) matching the expression of genes according to our RNA-seq untreated cells RPKM data (Fig. S7b), (2) matching the expression of genes according to PRO-seq untreated cells RPKM data (Fig. 4a) and (3) matching the expression of genes according to PRO-seq heat-shock treatment RPKM data (Fig. S7c). Significant differences were marked if the FDR-corrected 95th percentile of these 1000 p-values was lower than 0.05. To distinguish between naturally occurring and stress-induced readthrough, we compared the normalized Pol-II occupancy, i.e. PRO-seq RPKM values normalized by gene PRO-seq RPKM, in DoGs that were found in heat shock and not in untreated cells (Fig. 4b), as well as DoGs that were found in both heat shock and untreated cells (Fig. S7d) in different regions in the different conditions.

HSF1 ChIP-seq data integration

To examine the potential involvement of a transcription factor in DoG induction, we analyzed a published dataset of a genome-wide HSF1 ChIP-seq, performed in MEFs, before and after heat shock (6). We downloaded BED files of control and heat-shock HSF1 binding peaks of two different antibodies from NCBI GEO datasets. For each sample, we used the bedtools intersect command to generate a stringent set of HSF1 binding sites with peaks common to both antibodies. Following Mahat et al. (6), we removed HSF1 peaks that were present in Hsf1-/- MEFs, assuming that these peaks were false positives. We used the bedtools intersect command to sum the HSF1 binding peaks scores for each gene along a 1-kb region (1) upstream of the promoter, (2) upstream of the gene end and (3) downstream of gene end. Fig. S8a,b (SI Appendix) show little HSF1 binding beyond gene promoters. We therefore decided to focus on HSF1 binding within gene promoters. We defined HSF1 binding score in gene promoters

as the difference between the sum of HSF1 peak scores in heat shock and the sum of HSF1 peak scores in control samples. Finally, we only considered peaks with a score of at least 200 in both antibodies. A similar analysis was performed using another HSF1 heat shock ChIP-seq dataset (7) and resulted in the same trends (Fig. S8c).

Sequence motifs (6-mers) analysis

To test whether DoGs contain unique sequence features, we examined the occurrence of all possible 6-mers in the first 500 or 1000 bases (in a strand-specific manner) of pan-stress-DoGs and compared it to the concordant regions immediately downstream of non-DoG genes. We defined the log ratio of 6-mer occurrence in these two groups as the 6-mer score. To assess the significance of enrichment or depletion of a 6-mer, we compared the distribution of 6-mer scores from 10,000 sub-samples of expression-matched pan-stress- and non-DoG groups (as described above) to a random distribution generated by random shuffling of the gene lists in each sub-sample. Significant enrichment was assigned when the 2.5 percentile score of the 10,000 sub-samples was found to be higher than the 97.5 percentile score of the random distribution. Similarly, significant depletion was assigned when the 97.5 percentile score was found to be lower than the 2.5 percentile of the random distribution (see Dataset S2). We note that this analysis ignores the contribution of specific genes to the observed enrichment/depletion of each 6-mer. The same approach was used to search for 6-mers significantly enriched or depleted in stress-specific DoG clusters, defined by Fig. 1e: The green, red and turquoise clusters, representing oxidative stress, osmotic stress, and heat shock specific DoGs, respectively.

We conducted an additional gene-centric procedure in which the 6-mer score (presence enrichment score) was defined as the log ratio of the number of genes

containing at least one copy of the 6-mer in pan-stress DoGs versus non-DoGs (Dataset S2d). These two lists of motifs significantly overlap, as shown in Dataset S2e. All 6-mers were compared to known RNA binding protein sites from (8) by PSSM comparisons (Position Specific Scoring Matrices, performed as in Shalgi et al. (1)). A match was considered significant if it passed the 99.9th percentile cutoff of all possible matching scores.

Functional enrichment analysis

We used g:Profiler (9) to examine if there is significant enrichment of GO terms in various gene sets. First, we examined pan-stress DoGs and non-DoGs. As a background set, we used all expressed (RPKM>=4) genes. We first examined the entire set of 1556 pan-stress DoG genes and 1710 non-DoG genes, and found 83 significant terms for pan-stress DoGs and none for non-DoGs. We next repeated this analysis with 100 expression-matched sets of DoG and non-DoG genes. The results were inconsistent with the initial analysis; for pan-stress DoGs, there were no GO terms that were significant in more than 38 sets. In non-DoGs, there were no GO terms that were significant in more than 64 sets. We therefore conclude that, by properly controlling for expression, pan-stress DoGs and non-DoGs show no significant functional enrichment. The same procedure was applied to examine the set of genes associated with highly-induced DoGs, those with readthrough score fold change above 2 standard deviations from the mean in at least one condition (68, 115, and 158 for heat shock, H_2O_2, and KCl, respectively, 302 genes in total).

Chromatin environment analysis

To examine the chromatin environment of DoGs, we used several published datasets.

1. Histone modification marks (H3K4me1, H3K4me3 and H3K27ac), and CTCF (insulator) ChIP-seq, performed in MEF cells (from mouse ENCODE (10, 11)).
2. Histone modification marks (H3ac, H3K27me3, H3K36me3, H3K79me2 and H3K4me3), performed in C2C12 cells (from mouse ENCODE, generated by the Wold lab (10)).
3. DNase-Seq data, generated in NIH3T3 cells (from mouse ENCODE (10, 12)).
4. ATAC-Seq data generated in MEF cells (13).

For each dataset, we summed the peak scores (narrow peaks downloaded from the mouse ENCODE website, or from GEO) for each gene along three regions: (1) 1 kb upstream of transcription starts, i.e. the promoter region, (2) the last 1 kb of the gene (upstream of gene end, according to the loci annotation file), and (3) 5 kb downstream of gene end (according to the loci annotation file). We then compared the frequency of chromatin marks using 1000 expression-matched equal size sub-samples generated using the same sampling procedure described above. Importantly, to avoid promoters of neighboring genes, only genes with a same-strand downstream neighbor that is at least 10 kb away from the gene end were considered for the expression-matched sets. In each sub-sample, we applied the Fisher exact test on a contingency table with the number of pan-stress DoGs versus non-DoG regions with or without ChIP-seq peaks of the specific chromatin mark/binding event/accessibility mark. Significant difference was assigned if the 95th percentile of these 1000 p-values was lower than 0.05 (Fig. 7c, 7d, Fig. S11d).

We further conducted an additional permutation test, and defined a score for each mark in each region, as the difference in the number of regions containing each mark between pan-stress DoG and non-DoG genes, and generated a distribution of these scores with 1000 sub-samples of expression-matched equal size pan-stress DoG

and non-DoG sets. We then used the 1000 shuffled expression-matched subsets (as described above) to generate a background distribution for each of these scores. Significant difference was assigned if the 5th percentile of the observed score distribution was greater than 95% of the permutation-based background score distribution (as in Fig. S11c, S11e). The results of the permutation test were similar to those of the Fisher exact test, or slightly more significant in a couple of cases (H3K27ac promoter, Fig. S11c, DNase-seq Rep1 end of gene, Fig. S11e).

Supplementary Figures

Figure S1

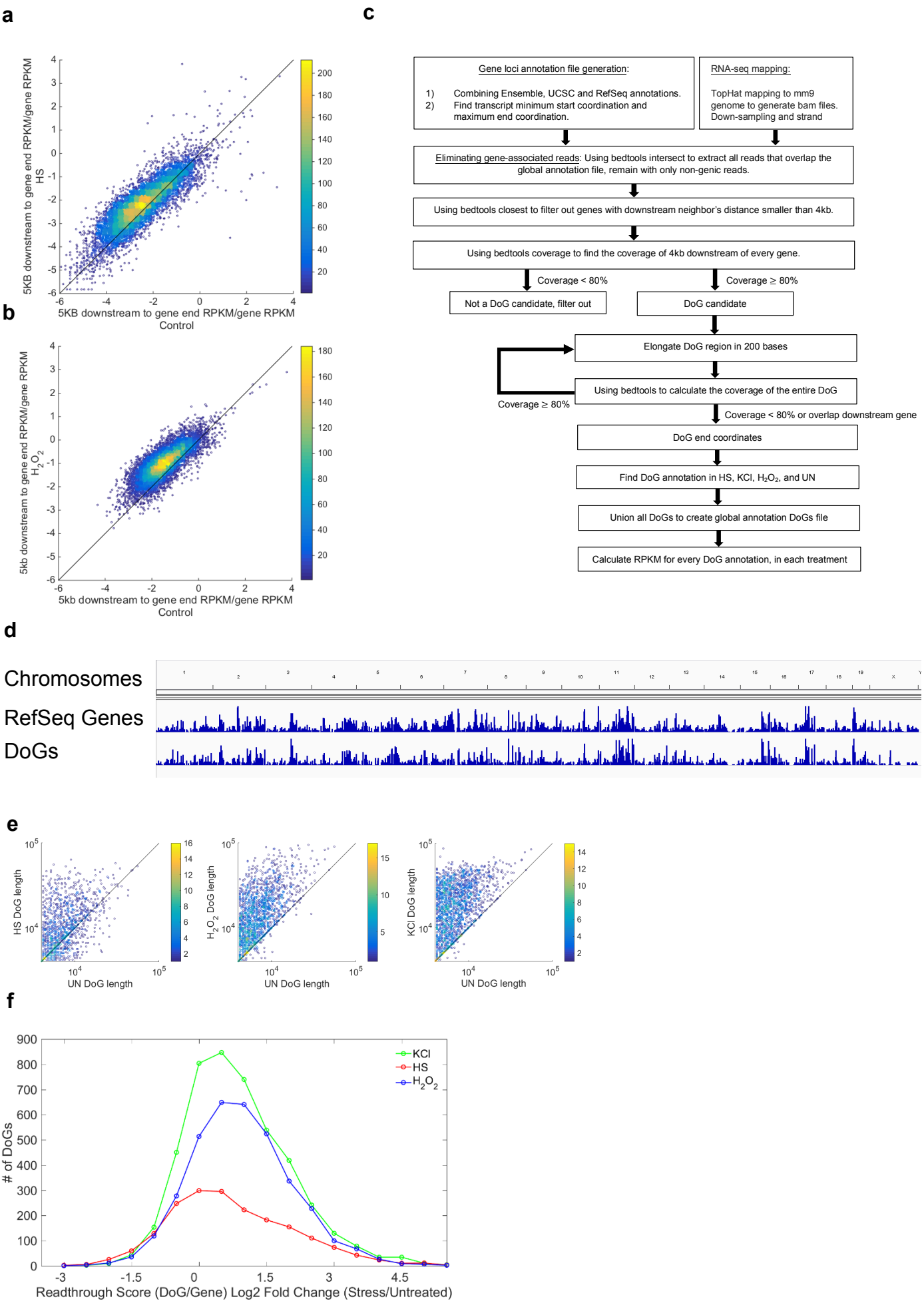

a
5KB downstream to gene end RPKM/gene RPKM HS
5KB downstream to gene end RPKM/gene RPKM Control
b
5kb downstream to gene end RPKM/gene RPKM H2O2
5kb downstream to gene end RPKM/gene RPKM Control
c
Gene loci annotation file generation:
1) Combining Ensemble, UCSC and RefSeq annotations.
2) Find transcript minimum start coordination and maximum end coordination.
RNA-seq mapping:
TopHat mapping to mm9 genome to generate bam files. Down-sampling and strand
Eliminating gene-associated reads: Using bedtools intersect to extract all reads that overlap the global annotation file, remain with only non-genic reads.
Using bedtools closest to filter out genes with downstream neighbor's distance smaller than 4kb.
Using bedtools coverage to find the coverage of 4kb downstream of every gene.
Coverage < 80%
Coverage ≥ 80%
Not a DoG candidate, filter out
DoG candidate
Elongate DoG region in 200 bases
Using bedtools to calculate the coverage of the entire DoG
Coverage ≥ 80%
Coverage < 80% or overlap downstream gene
DoG end coordinates
Find DoG annotation in HS, KCl, H2O2, and UN
Union all DoGs to create global annotation DoGs file
Calculate RPKM for every DoG annotation, in each treatment
d
Chromosomes
RefSeq Genes
DoGs
e
HS DoG length
H2O2 DoG length
KCl DoG length
UN DoG length
f
# of DoGs
KCl
HS
H2O2
Readthrough Score (DoG/Gene) Log2 Fold Change (Stress/Untreated)

Figure S1: Readthrough data analysis and identified DoGs.

(a) Heat-shock-induced readthrough in 5 kb downstream of gene loci from previously published NIH3T3 RNA-seq data (1). Ratio of RPKM in 5 kb downstream of gene over RPKM of the gene (as calculated by Cufflinks) is shown for heat shock (y-axes, log scale) against control (x-axes, log scale). (b) DoGs are induced by oxidative stress (0.2 mM H_2O_2) in human skin fibroblasts. Analysis of data from Giannakakis et al. (3), depicted as in (a), except that the RPKM over the 5 kb downstream of the end-of-gene is normalized to the RPKM of the last 1 kb of the gene. (c) DoG discovery pipeline flowchart. (d) DoG distribution throughout the mouse genome. (e) DoG length scatterplot in the three treatments versus control (log10 scale). (f) Readthrough scores (DoG RPKM/gene RPKM), log2 fold change in the three stress conditions – according to the shortest DoG definition. DoGs were discovered independently in all four conditions (three stresses and untreated cells), and the shortest DoG definition was noted. Then RPKMs were calculated according to this definition in all conditions. The histogram shows that, even with this definition, DoGs show a marked induction in all three conditions compared to control conditions.

Figure S2

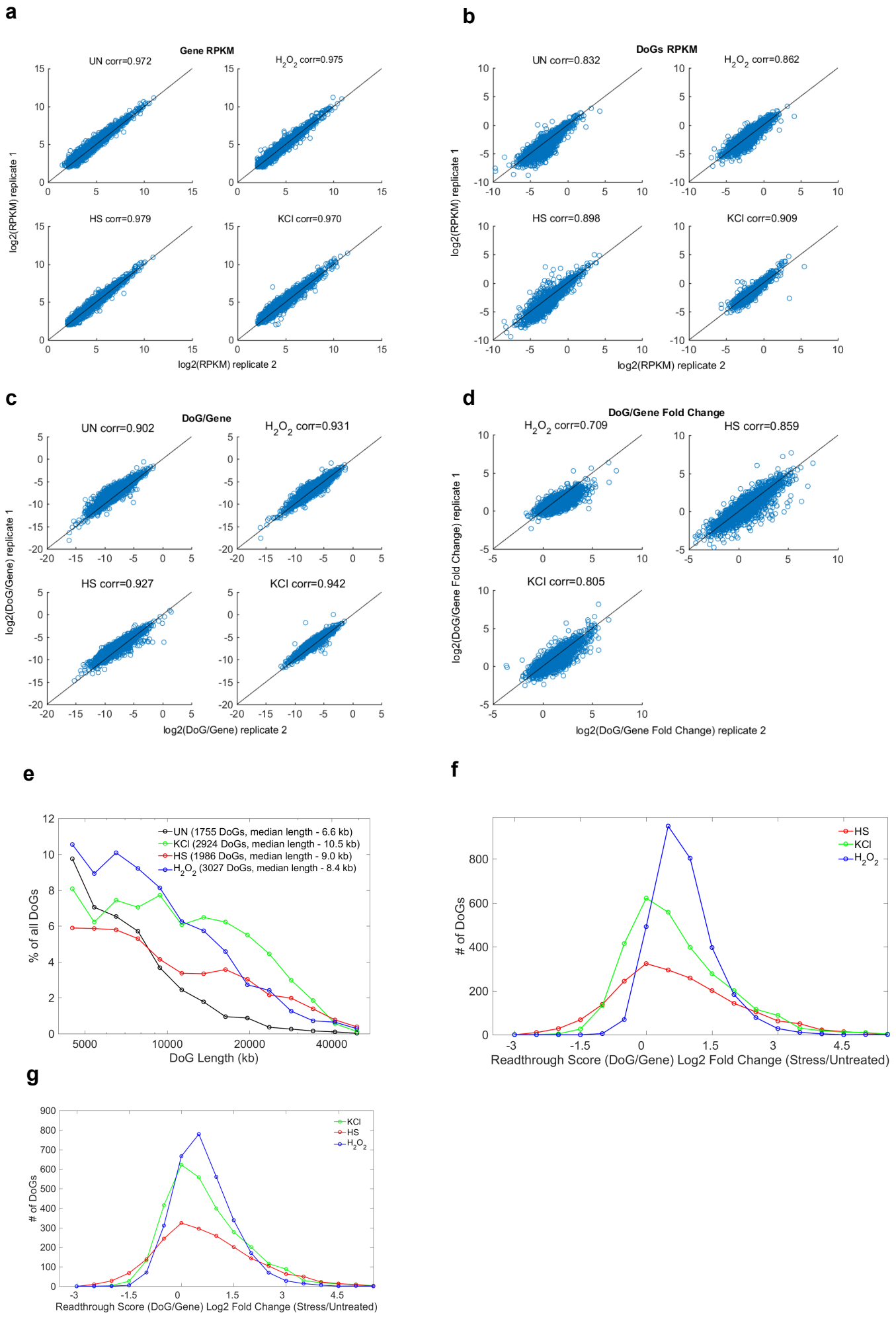

a
Gene RPKM
UN corr=0.972
H2O2 corr=0.975
HS corr=0.979
KCl corr=0.970
log2(RPKM) replicate 1
log2(RPKM) replicate 2
b
DoGs RPKM
UN corr=0.832
H2O2 corr=0.862
HS corr=0.898
KCl corr=0.909
log2(RPKM) replicate 1
log2(RPKM) replicate 2
c
DoG/Gene
UN corr=0.902
H2O2 corr=0.931
HS corr=0.927
KCl corr=0.942
log2(DoG/Gene) replicate 1
log2(DoG/Gene) replicate 2
d
DoG/Gene Fold Change
H2O2 corr=0.709
HS corr=0.859
KCl corr=0.805
log2(DoG/Gene Fold Change) replicate 1
log2(DoG/Gene Fold Change) replicate 2
e
UN (1755 DoGs, median length - 6.6 kb)
KCl (2924 DoGs, median length - 10.5 kb)
HS (1986 DoGs, median length - 9.0 kb)
H2O2 (3027 DoGs, median length - 8.4 kb)
% of all DoGs
DoG Length (kb)
f
HS
KCl
H2O2
# of DoGs
Readthrough Score (DoG/Gene) Log2 Fold Change (Stress/Untreated)
g
KCl
HS
H2O2
# of DoGs
Readthrough Score (DoG/Gene) Log2 Fold Change (Stress/Untreated)

Figure S2: Readthrough data – biological replicate

A biological replicate experiment was performed and subjected to the pipeline described above. The figure shows the reproducibility of (a) gene RPKMs (b) DoG RPKMs (c) Readthrough scores (DoG RPKM/Gene RPKM) and (d) readthrough score fold changes in stress versus control. Correlation coefficients (corr) between the data in the two biological replicates are noted in the plots. (e) DoG length distributions of the biological replicate, as in Figure 1b. (f) DoG readthrough score fold changes (in log2) distributions, for the longest DoG definitions, as in Fig. 1d. (g) DoG readthrough score fold changes (in log2) distributions, for the shortest DoG definitions, as in Figure S1f.

Figure S3

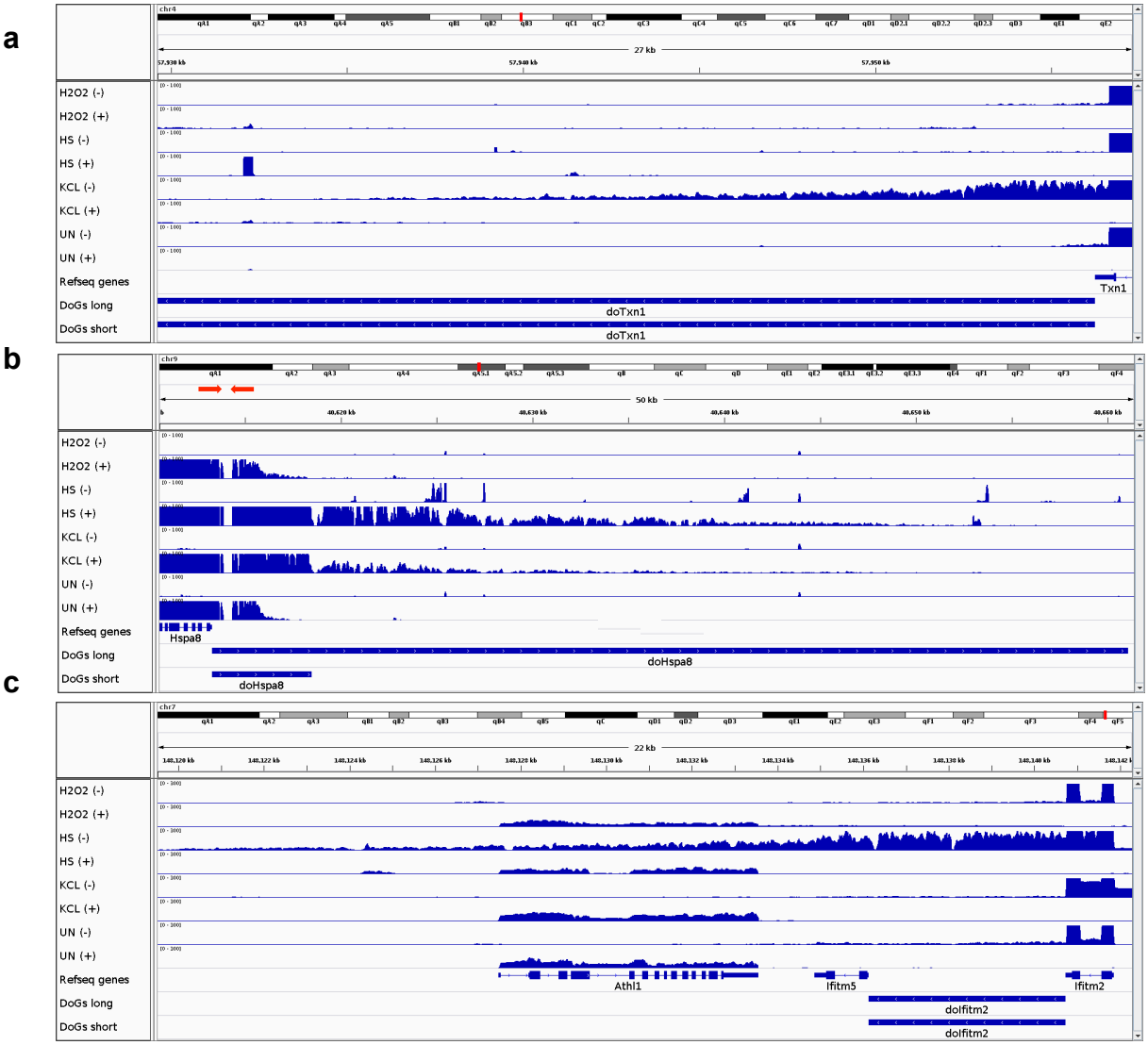

a
H2O2 (-)
H2O2 (+)
HS (-)
HS (+)
KCL (-)
KCL (+)
UN (-)
UN (+)
Refseq genes
DoGs long
DoGs short
Txn1
doTxn1
doTxn1
b
Hspa8
doHspa8
doHspa8
c
Athl1
Ifitm5
Ifitm2
dolfitm2
dolfitm2
d
564 nt
24 27 30 33
untreated
44°C
KCl
H2O2
e
relative RNA levels (log scale)
Untreated
44° 2h
44° 4h
f
Untreated
KCl 1h
g
relative RNA levels
doHnrnpa2b1
doHspa8
dolfitm2
Untreated
H2O2 2h

Figure S3: Additional analyses of DoG induction by multiple stresses

(a-c) Mapped read density for three DoGs: doTxn1 (a), doHspa8 (b), and doIfitm2 (c) in all stress conditions. Top panel shows genomic coordinates. Middle panel shows mapped reads, with reads mapping to the forward and reverse strands separately shown (as "+" and "-", respectively). The bottom panel shows RefSeq genes, and the longest DoG definition. Red arrows in panel (b) mark an observed gap in doHspa8, most likely due to RNA-seq mapping issues in that region. (d) doHspa8 is continuous over the observed gap. PCR was performed using primers flanking the observed gap in mapped RNA-seq reads (as indicated by red arrows in Fig. S3b) on cDNA made from NIH3T3 cells treated with 2 h of heat shock, osmotic stress or oxidative stress. PCR made in several cycles shows that the region marked by red arrows in (b) is highly induced in heat shock and that doHspa8 is continuous in all conditions. (e-g) DoG induction by qPCR. doHnrnpa2b1, doHspa8, and doIfitm2 are confirmed by qRT-PCR to be induced by heat shock (e), osmotic stress (f), and (g) oxidative stress. * denotes $p<0.05$, ** denotes $p<0.01$ and *** denotes $p<0.001$. (e-f) Figure shows mean and standard deviations of 4 biological replicate experiments; data are shown on a logarithmic scale. (g) Figure shows mean and standard deviations of 3 biological replicate experiments.

Figure S4

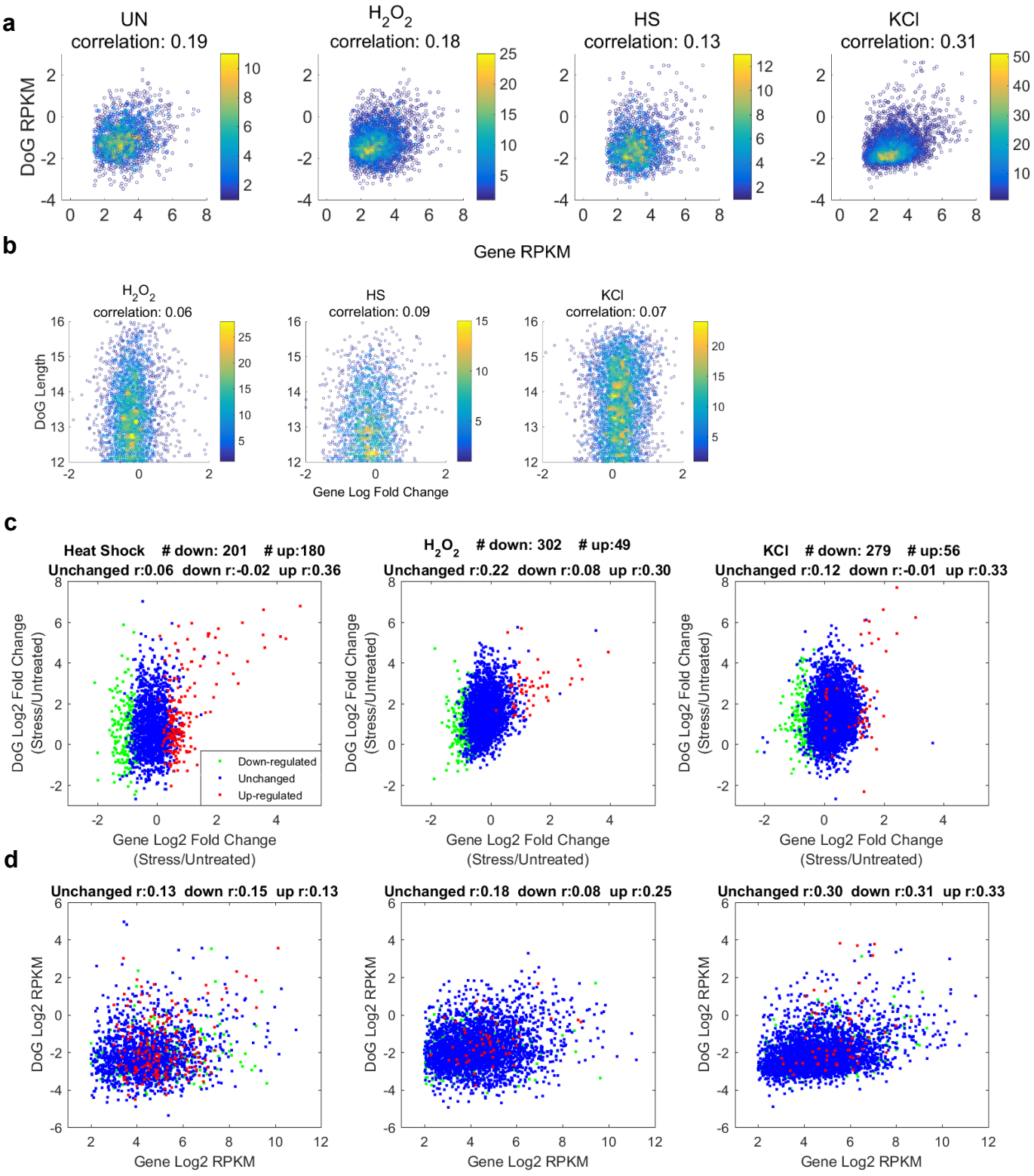

a
UN
correlation: 0.19
H2O2
correlation: 0.18
HS
correlation: 0.13
KCl
correlation: 0.31
DoG RPKM
Gene RPKM
b
H2O2
correlation: 0.06
HS
correlation: 0.09
KCl
correlation: 0.07
DoG Length
Gene Log Fold Change
c
Heat Shock # down: 201 # up:180
Unchanged r:0.06 down r:-0.02 up r:0.36
H2O2 # down: 302 # up:49
Unchanged r:0.22 down r:0.08 up r:0.30
KCl # down: 279 # up:56
Unchanged r:0.12 down r:-0.01 up r:0.33
DoG Log2 Fold Change (Stress/Untreated)
Gene Log2 Fold Change (Stress/Untreated)
Down-regulated
Unchanged
Up-regulated
d
Unchanged r:0.13 down r:0.15 up r:0.13
Unchanged r:0.18 down r:0.08 up r:0.25
Unchanged r:0.30 down r:0.31 up r:0.33
DoG Log2 RPKM
Gene Log2 RPKM

Figure S4: *Comparative analysis of DoG and gene expression*

(a) DoG versus gene expression levels (RPKM, both in log2 scale) in all conditions show low correlation. (b) DoG length (in kb, log2 scale) versus gene expression fold change (log2 scale) show no correlation. (c,d) Down-regulated (green), unchanged (blue), and up-regulated (red) genes were identified by DESeq2 (14). (c) DoG RPKM log2 fold change versus Gene RPKM log2 fold change in the three treatments. (d) DoG log2 RPKM versus Gene log2 RPKM in the three treatments.

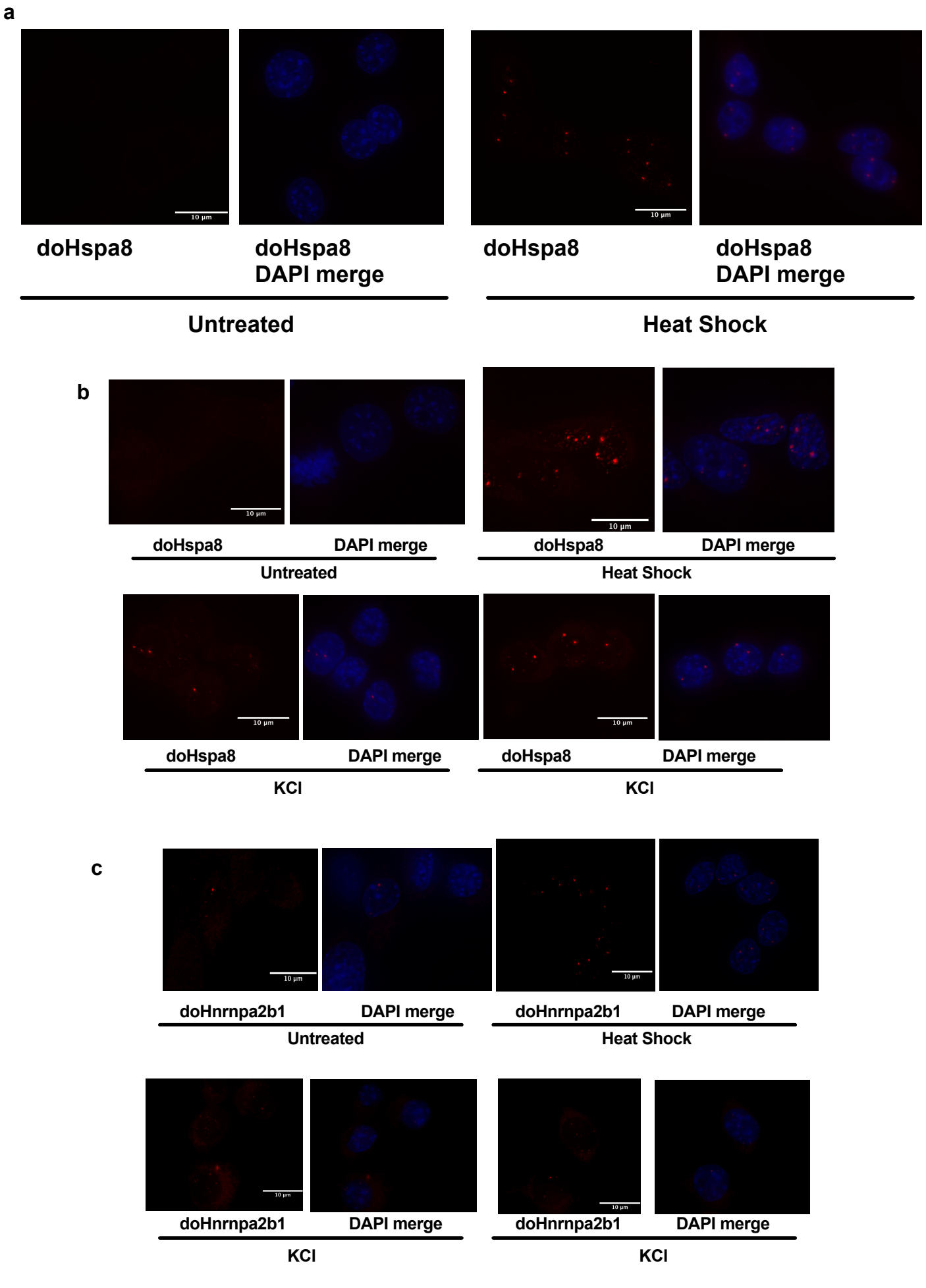
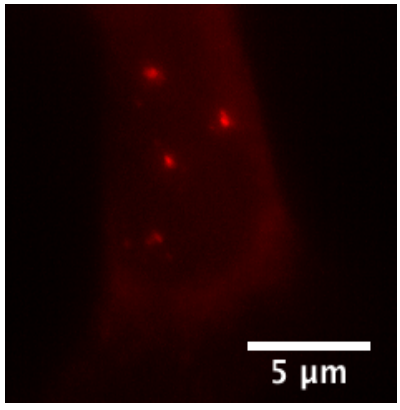

a
10 µm
doHspa8
doHspa8
DAPI merge
doHspa8
doHspa8
DAPI merge
Untreated
Heat Shock
b
10 µm
doHspa8
DAPI merge
doHspa8
DAPI merge
Untreated
Heat Shock
10 µm
doHspa8
DAPI merge
doHspa8
DAPI merge
KCl
KCl
c
10 µm
doHnrnpa2b1
DAPI merge
doHnrnpa2b1
DAPI merge
Untreated
Heat Shock
10 µm
doHnrnpa2b1
DAPI merge
doHnrnpa2b1
DAPI merge
KCl
KCl

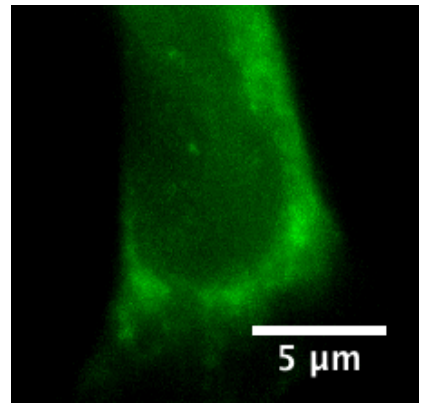
Figure S5: Single molecule RNA-FISH confirms induction of nuclear DoGs

Single molecule Stellaris RNA-FISH was used to confirm the heat shock (a-b) and KCl-mediated induction (b) of doHspa8 (a-b) and doHnrnpa2b1 (c) in NIH3T3 cells. In both cases, DoGs appear as nuclear punctae. Scale bars indicate 10 μm.

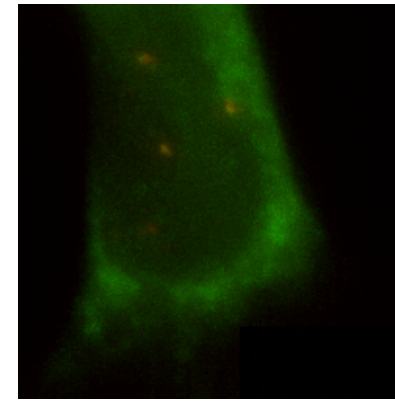
a



doHnrnpa2b1

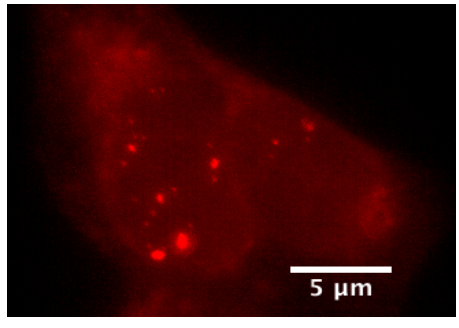


Hnrnpa2b1
intron

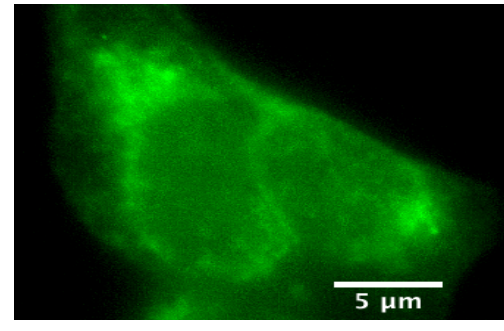


merge

b

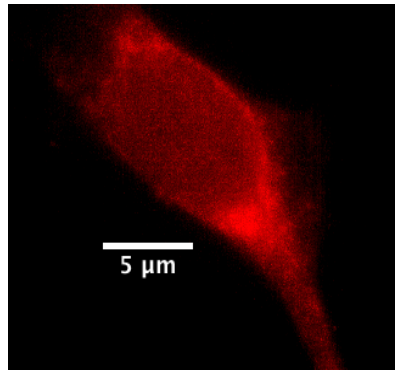


doHnrnpa2b1

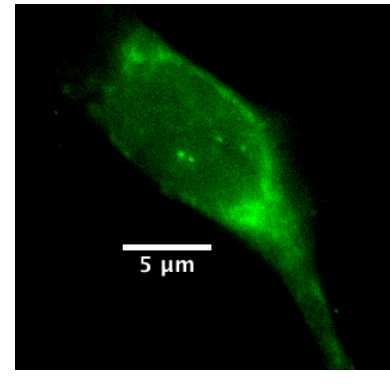


No probe

c



No probe



Hnrnpa2b1
intron

Figure S6: Single molecule RNA-FISH confirms DoG localization to their sites of transcription

Co-staining with the doHnrnpa2b1 probe and a probe targeting the first intron of doHnrpna2b1 in heat-shock-treated NIH3T3 cells confirms that some of the nuclear punctae with doHnrnpa2b1 staining represent Hnrnpa2b1 transcription sites (a). The specificity of the staining is confirmed by the observation of doHnrnpa2b1 (red) but not intron (green) staining when the doHnrnpa2b1 probe (red) only was used (b), and vice versa when the intron probe only (green) was used (c). Scale bars indicate 5 μm.

Figure S7

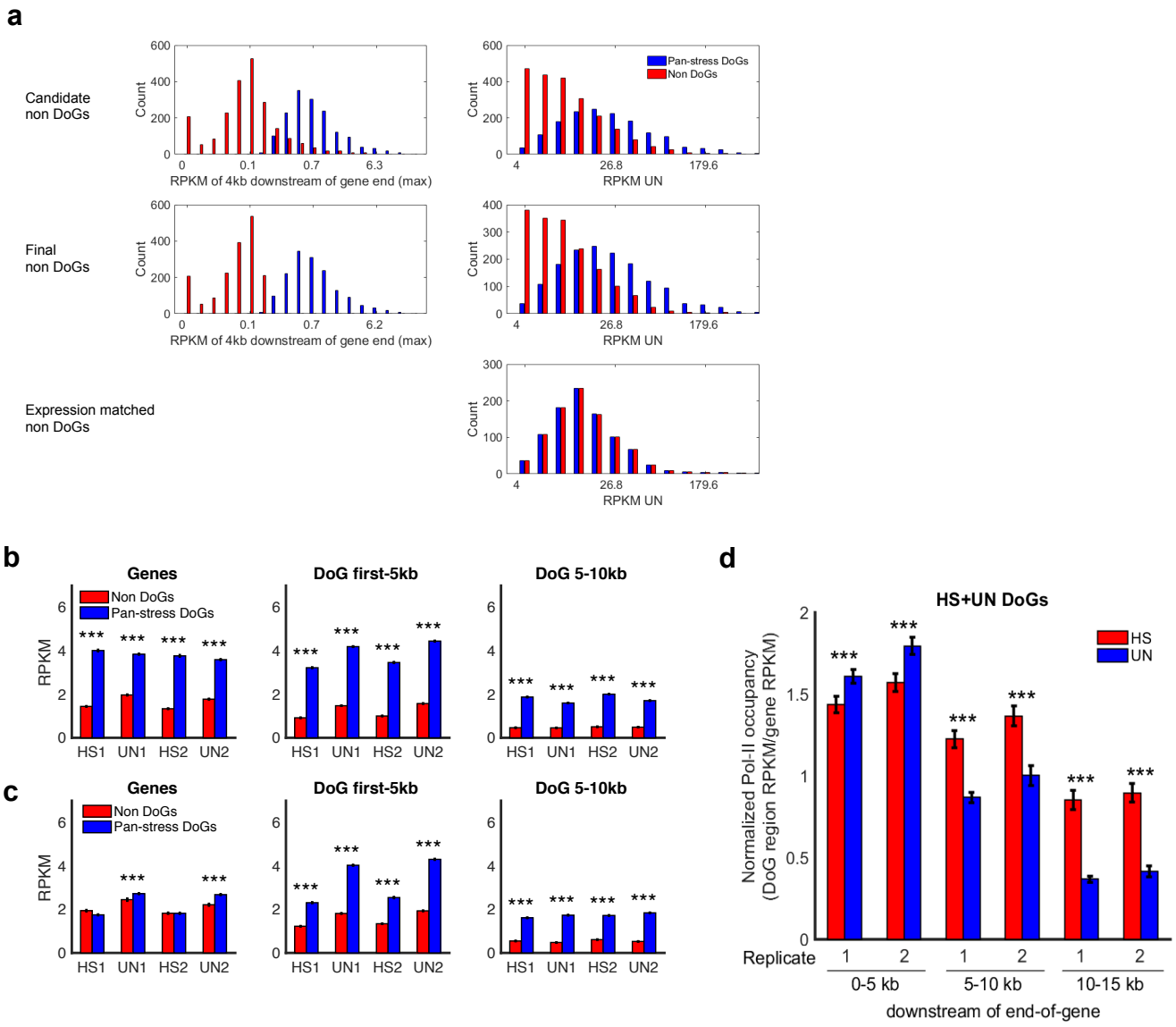

a
Candidate non DoGs
Final non DoGs
Expression matched non DoGs
Count
RPKM of 4kb downstream of gene end (max)
RPKM UN
Pan-stress DoGs
Non DoGs
b
Genes
DoG first-5kb
DoG 5-10kb
RPKM
HS1 UN1 HS2 UN2
c
d
HS+UN DoGs
HS
UN
Normalized Pol-II occupancy (DoG region RPKM/gene RPKM)
Replicate
0-5 kb
5-10 kb
10-15 kb
downstream of end-of-gene

Figure S7: Pan-stress DoGs and non-DoG sets and Pol-II occupancy analysis

(a) Generation of expression-matched non-DoG sets. For each DoG-associated and non-DoG gene, the maximum RPKM of the three stresses in the 4 kb downstream of the gene end was calculated. Exclusion of candidate non-DoGs with expression in the 4 kb downstream of the gene end above the minimal RPKM of the first 4 kb of pan-stress DoGs (left upper panel) yielded the final non-DoG set (middle panel). The final non-DoGs were randomly sub-sampled, as are pan-stress DoGs, to generate expression-matched DoGs and non-DoGs sets (lower panel). (b-c) Mean and standard error of Pol-II occupancy (PRO-seq RPKM) in genes and DoG regions (first 5 kb and 5-10 kb) in heat-shock and untreated samples is shown for pan-stress DoGs (blue) versus non-DoGs (red). Significant differences in between pan-stress DoGs and non-DoG downstream regions are shown in expression-matched pan-stress- and non-DoG sets, where expression was matched using either our RNA-seq data (b), or PRO-seq heat-shock data (c). Significance was estimated as the FDR-corrected 95th percentile of 1000 ranksum test p-values for 1000 expression-matched sub-samples (*** - $p < 0.001$). (d) Normalized Pol-II occupancy in all heat-shock DoGs (mean and standard error, PRO-seq RPKM, normalized by PRO-seq RPKM in the corresponding upstream gene) in different DoG regions (first 5 kb, 5-10 kb and 10-15 kb downstream to gene end), display significantly higher occupancy in heat-shocked (red) versus untreated cells (blue), specifically in farther downstream DoG regions (5-10 kb and 10-15 kb downstream of gene ends). Significance was assessed with ranksum test and controlled for false-discovery rate (*** - $p < 0.001$).

Figure S8

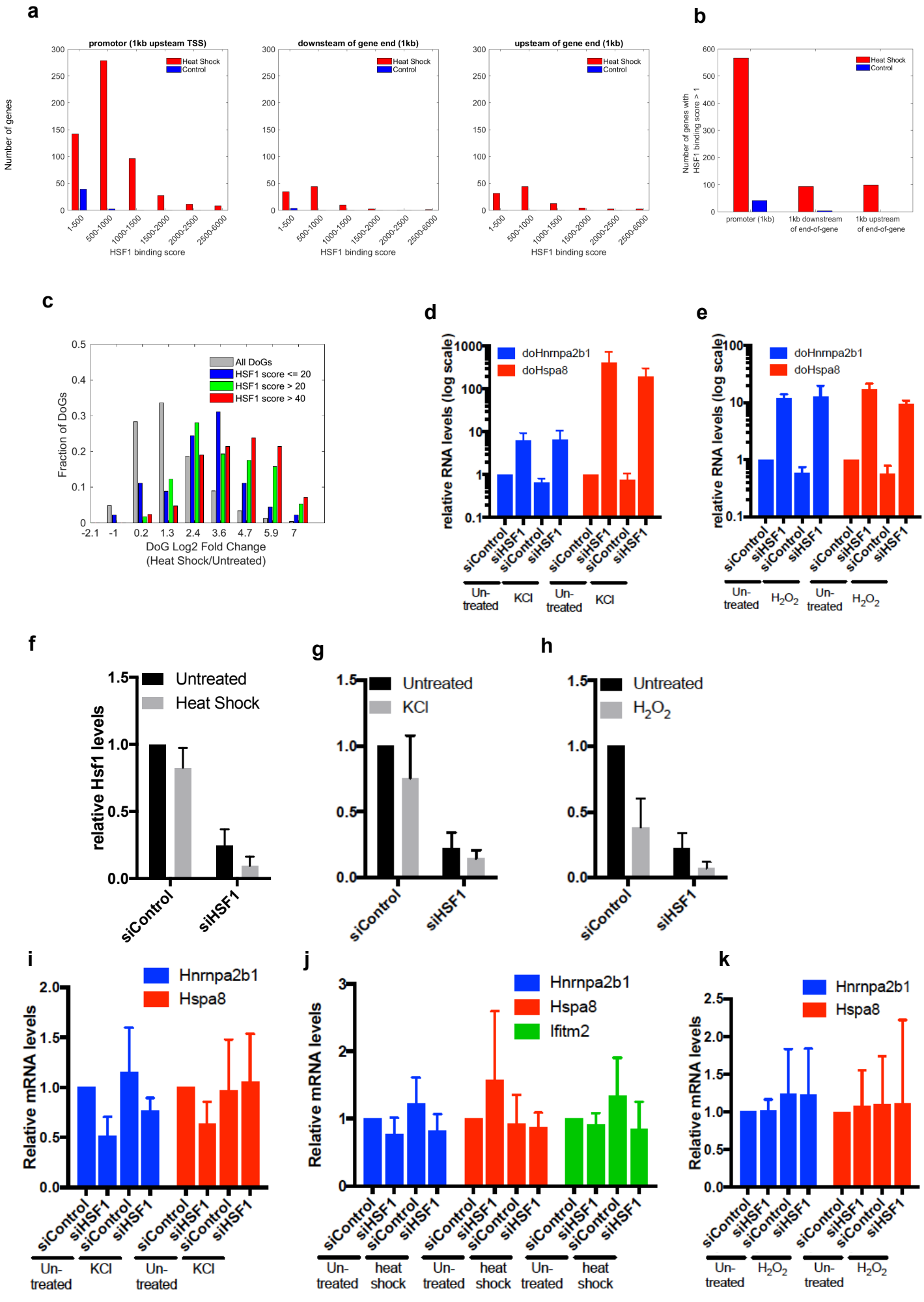

a
promotor (1kb upsteam TSS)
downsteam of gene end (1kb)
upsteam of gene end (1kb)
Heat Shock
Control
Number of genes
HSF1 binding score
1-500
500-1000
1000-1500
1500-2000
2000-2500
2500-6000
b
Number of genes with HSF1 binding score > 1
promoter (1kb)
1kb downstream of end-of-gene
1kb upstream of end-of-gene
c
All DoGs
HSF1 score <= 20
HSF1 score > 20
HSF1 score > 40
Fraction of DoGs
DoG Log2 Fold Change (Heat Shock/Untreated)
d
doHnrnpa2b1
doHspa8
relative RNA levels (log scale)
siControl
siHSF1
Un-treated
KCl
e
H_2O_2
f
Untreated
Heat Shock
relative Hsf1 levels
g
KCl
h
H_2O_2
i
Hnrnpa2b1
Hspa8
Relative mRNA levels
j
Ifitm2
heat shock
k

Figure S8: Impact of Hsf1 knockdown on DoG induction

(a) The number of genes with HSF1 binding peaks taken from Mahat et al. (6) in the 1 kb promoter region (left panel), in the 1 kb downstream (middle panel) and the last 1 kb of gene (right panel) for heat shock and control, shows that HSF1 binds mostly in promoter regions. (b) The number of genes with the sum of HSF1 binding peak scores in each condition (sum of peak binding scores in each region) greater than 20 in the 1 kb promoter region, in the 1 kb downstream of gene, and in the last 1 kb of the gene. (c) HSF1 data from Takii et al. (7) show a similar pattern as in Fig. 5b. DoGs with heat-shock-induced HSF1 promoter binding (binding score is calculated as the difference between the sum of all HSF1 binding peak scores in the 1 kb promoter region of the gene, in heat shock versus control) show increased DoG expression levels, as seen by a shifted distribution of RPKM log2 fold changes in heat shock compared to control. Blue – low binding scores (less than or equal to 20 and greater than zero, mean=2.14), show a marked shift toward induction compared to all DoGs (grey). Green – binding score above 20 (mean=3.15) and red, binding score above 40 (mean=3.63), show even higher levels of induction. (d,e) HSF1 knockdown does not affect DoG induction by KCl (d, n=5) and H_2O_2 (e, n=3) (the same untreated samples were used as in Fig. 4f). (f-h) Efficient knockdown of Hsf1 mRNA in samples transfected with siRNA targeting Hsf1 and treated with heat shock (f), KCl (g) or H_2O_2 (h). (i-k) DoG-generating gene mRNA levels are not significantly altered by Hsf1 knockdown after treatment with KCl (i), heat shock (j) or H_2O_2 (k). See Fig. 5c for details.

Figure S9

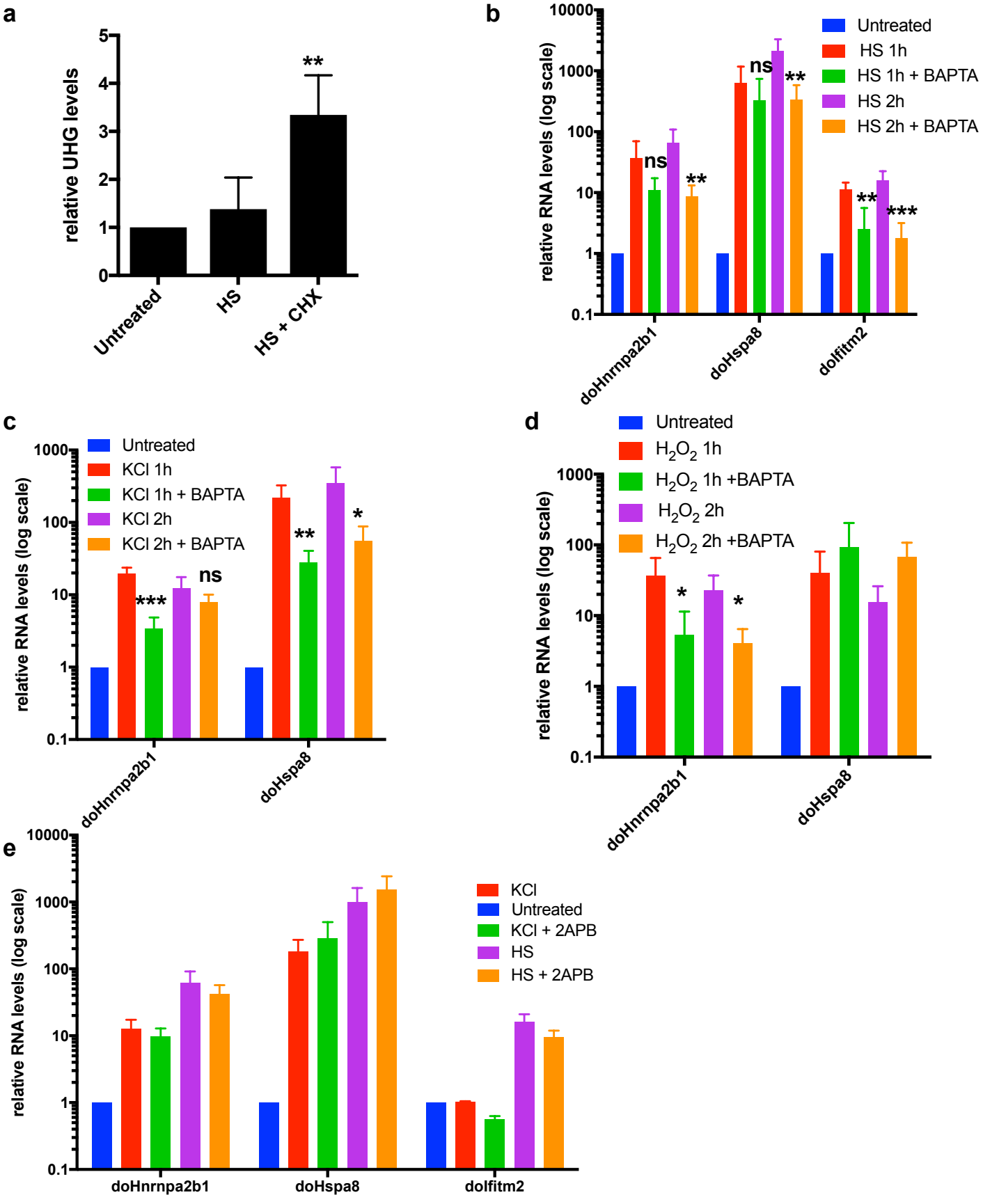
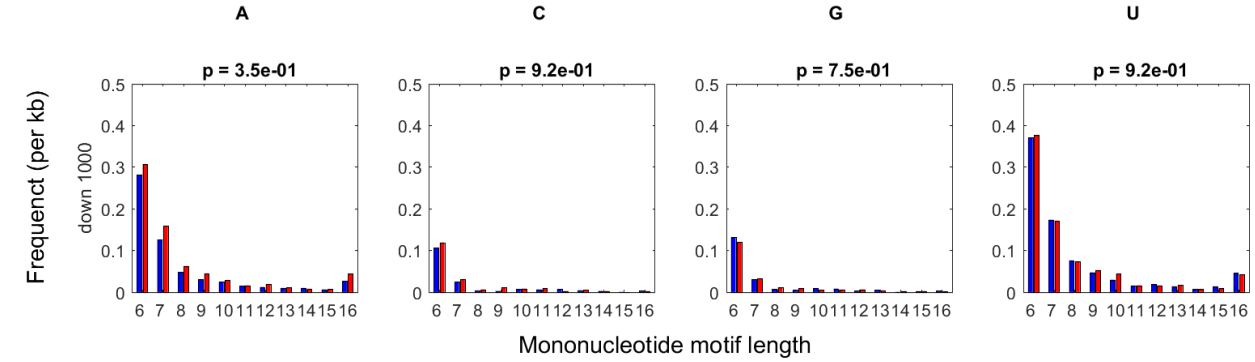
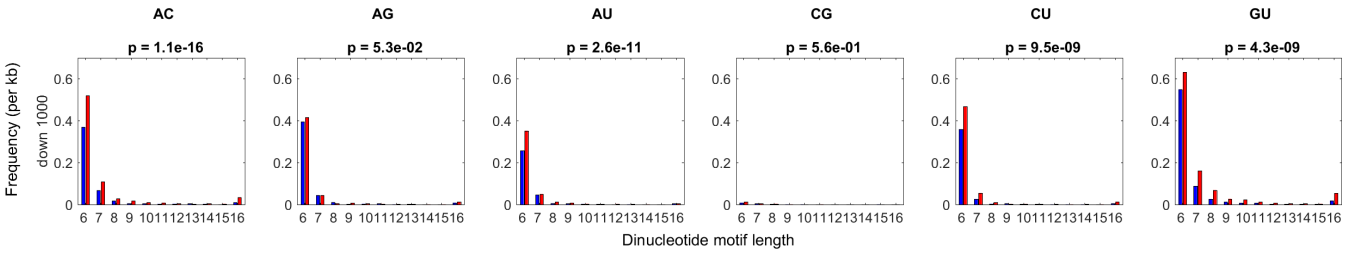

a
relative UHG levels
Untreated
HS
HS + CHX
**
b
relative RNA levels (log scale)
Untreated
HS 1h
HS 1h + BAPTA
HS 2h
HS 2h + BAPTA
ns
**
doHnrnpa2b1
doHspa8
dolfitm2
c
relative RNA levels (log scale)
Untreated
KCl 1h
KCl 1h + BAPTA
KCl 2h
KCl 2h + BAPTA
***
ns
**
*
doHnrnpa2b1
doHspa8
d
relative RNA levels (log scale)
Untreated
H2O2 1h
H2O2 1h +BAPTA
H2O2 2h
H2O2 2h +BAPTA
*
doHnrnpa2b1
doHspa8
e
relative RNA levels (log scale)
KCl
Untreated
KCl + 2APB
HS
HS + 2APB
doHnrnpa2b1
doHspa8
dolfitm2

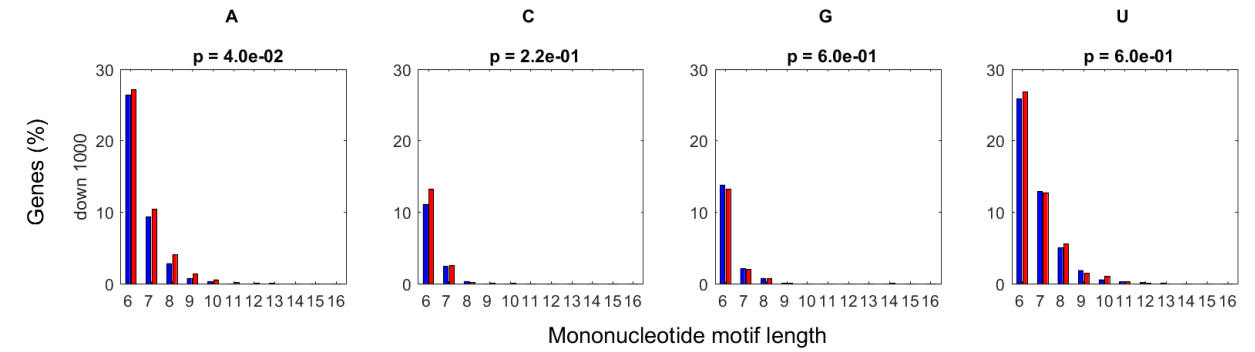
Figure S9: Effect of CHX and calcium signaling on DoG induction

Cycloheximide (CHX) treatment efficiently inhibits translation, as demonstrated by the increase in the NMD-target RNA UHG in response to CHX treatment (a). Pretreating NIH3T3 cells with the calcium chelator BAPTA 30 min ahead of stress reduces DoG induction by heat shock (b, n=7) or KCl (c, n=4) in most cases, but only reduces doHnrnpa2b1 and not doHspa8 induction by H_2O_2 (d, n=5). In some replicates, the same untreated samples were used as references for heat shock and KCl. DoG induction by heat shock is not inhibited by 2APB (e) (n=4).

Figure S10


a
A
C
G
U
p = 3.5e-01
p = 9.2e-01
p = 7.5e-01
p = 9.2e-01
Frequenct (per kb)
down 1000
Mononucleotide motif length


b
AC
AG
AU
CG
CU
GU
p = 1.1e-16
p = 5.3e-02
p = 2.6e-11
p = 5.6e-01
p = 9.5e-09
p = 4.3e-09
Frequency (per kb)
down 1000
Dinucleotide motif length


c
A
C
G
U
p = 4.0e-02
p = 2.2e-01
p = 6.0e-01
p = 6.0e-01
Genes (%)
down 1000
Mononucleotide motif length

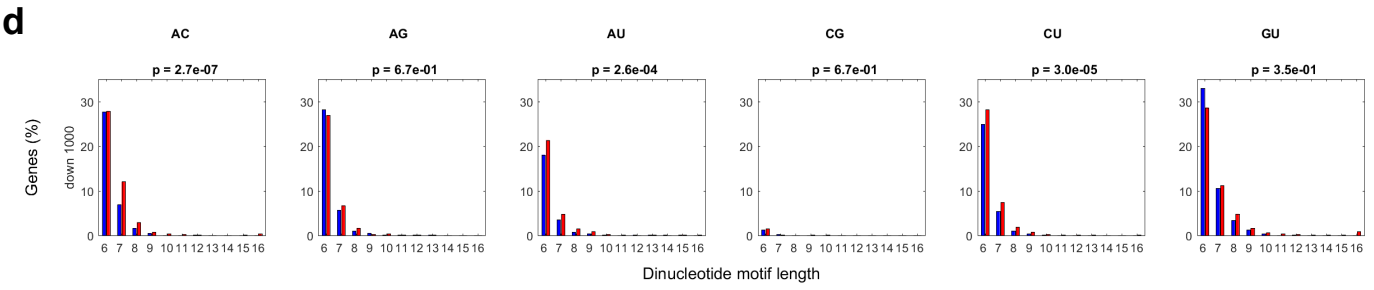

d
AC
AG
AU
CG
CU
GU
p = 2.7e-07
p = 6.7e-01
p = 2.6e-04
p = 6.7e-01
p = 3.0e-05
p = 3.5e-01
Genes (%)
down 1000
Dinucleotide motif length

Figure S10: Readthrough transcripts are depleted of simple repeat sequences

(a-b) The frequency per 1000 bases of repeat (a) mononucleotides and (b) dinucleotides motifs of length six or more, in 1000 base regions downstream of gene end of pan-stress DoGs and non-DoGs. Significance was assessed with ranksum test and p-values are shown. (c-d) The percentage of genes containing repeat motifs (c) mononucleotides and (d) dinucleotides of length six or more, in 1000 base regions downstream of gene end of pan-stress DoGs and non-DoGs. Significance was assessed with ranksum test and p-values are shown.

Figure S11

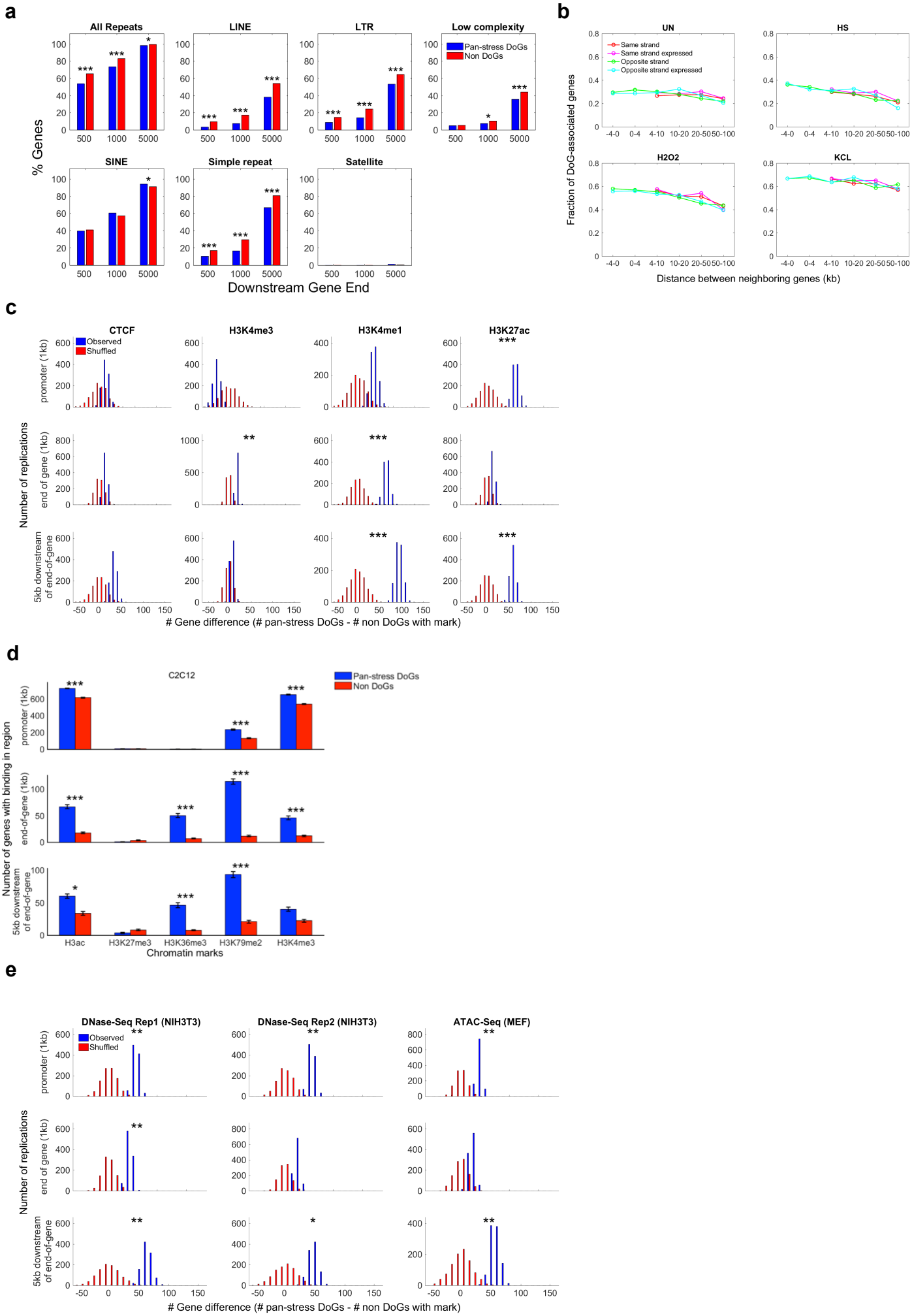

a
All Repeats
LINE
LTR
Low complexity
SINE
Simple repeat
Satellite
Pan-stress DoGs
Non DoGs
% Genes
Downstream Gene End
b
UN
HS
H2O2
KCL
Same strand
Same strand expressed
Opposite strand
Opposite strand expressed
Fraction of DoG-associated genes
Distance between neighboring genes (kb)
c
CTCF
H3K4me3
H3K4me1
H3K27ac
Observed
Shuffled
promoter (1kb)
end of gene (1kb)
5kb downstream of end-of-gene
Number of replications
Gene difference (# pan-stress DoGs - # non DoGs with mark)
d
C2C12
Pan-stress DoGs
Non DoGs
Number of genes with binding in region
H3ac
H3K27me3
H3K36me3
H3K79me2
H3K4me3
Chromatin marks
e
DNase-Seq Rep1 (NIH3T3)
DNase-Seq Rep2 (NIH3T3)
ATAC-Seq (MEF)
Observed
Shuffled
promoter (1kb)
end of gene (1kb)
5kb downstream of end-of-gene
Number of replications
Gene difference (# pan-stress DoGs - # non DoGs with mark)

Figure S11: Chromatin features in stress-induced readthrough regions

(a) The percentage of genes containing annotated repeat sequences (according to Repeat Masker table from UCSC mm9) in 500, 1000 and 5000 bp regions downstream of gene end of pan-stress- and non-DoG genes. Significance was assessed with Fisher exact test and controlled for false-discovery rate (* denotes $p<0.05$, ** denotes $p<0.01$ and *** denotes $p<0.001$). (b) DoG frequency is negatively correlated with the distance to the nearest downstream neighboring gene in DoGs discovered in each of the three stress conditions. The fraction of DoG-associated genes out of all expressed genes is shown against the distance to the downstream nearest neighboring gene, as in Fig. 7a. Gene pairs were grouped according to the strand orientation of the downstream neighbor with respect to the upstream gene (same-strand or opposite-strand) and whether the downstream is expressed (RPKM >=4). (c) Significance assessment of the difference in chromatin accessibility in pan-stress readthrough versus non-readthrough gene regions, using a permutation test. The observed distribution of differences (over 1000 expression-matched sub-samples, in blue) between the number of pan-stress readthrough and non-readthrough gene regions with CTCF binding or each of the three histone marks in MEF cells was compared with a background distribution generated by random permutations (red). Significant differences were assigned if the 5th percentile of the observed difference was greater than the 95th percentile of the permutation-based difference distribution. (d) Number of regions containing histone marks in promoters (1 kb upstream to transcription start site), gene ends (1 kb upstream to gene end), and downstream of gene end (5 kb) in pan-stress readthrough and non-readthrough genes in C2C12 cells (data generated by the Wold lab (10)), as in Fig. 7c,d. Significance is marked by *. (e) Significance assessment of the difference in DNase-seq/ATAC-seq peaks through permutation tests, as in (c).

Table S1

sample	# expressed OS overlapping genes	Total #Genes OS expressed Neighbors (>0)	# DoGs non overlapping expressed OS nearest neighbor (distance to gene>0)	# DoGs overlapping expressed OS nearest neighbor (distance to gene>0)	# DoGs overlapping expressed OS nearest neighbor (0<distance to gene <4kb)	# DoGs overlapping expressed OS nearest neighbor (distance to gene >4kb)	Total #genes OS expressed neighbors (>0 & <4)	Total #genes OS expressed neighbors (>4)
UN	578	1241	111	241	193	48	655	586
HS	526	1208	89	272	203	69	633	575
H2O2	516	1183	127	475	349	126	626	557
KCl	602	1296	203	689	492	197	680	616

% DoGs non overlapping expressed OS nearest neighbor (distance to gene>0)	% DoGs overlapping expressed OS nearest neighbor (distance to gene>0)	# DoGs overlapping expressed OS nearest neighbor (0<distance to gene <4kb)	# DoGs overlapping expressed OS nearest neighbor (distance to gene >4kb)
0.09	0.19	0.29	0.08
0.07	0.23	0.32	0.12
0.11	0.40	0.56	0.23
0.16	0.53	0.72	0.32

Table S1: DoGs increase the potential for antisense between overlapping transcripts from opposite strands (OS).
The upper panel lists the number of expressed overlapping genes, neighboring genes with and without DoGs that overlap neighboring genes, and with respect to distance between genes. Bottom panel lists the fraction of DoGs in each category with respect to number of expressed OS neighbors.

Table S2

			Mouse			Fraction of overlaps with respect to mouse		
			all genes	all DoGs	pan-stress DoGs	all genes	all DoGs	pan-stress DoGs
		Total	23195	4838	1556			
Human	all genes	38476	14346	4194	1386	0.62	0.87	0.89
	all DoGs	2230	1825	1026	427	0.13	0.24	0.31
	DoGs with gene last 1kb RPKM>=4	1332	1143	705	306	0.63	0.69	0.72
	stringent DoGs	281	239	154	70	0.21	0.22	0.23
Fraction of overlaps with respect to human	all genes		0.37	0.29	0.33			
	all DoGs		0.82	0.56	0.42			
	DoGs with gene last 1kb RPKM>=4		0.86	0.62	0.43			
	stringent DoGs		0.85	0.64	0.45			

Table S2: Intersection of human DoGs from Vilborg et al. 2015 and mouse DoGs from the current study.
Highlighted cells are discussed in the text. Significance was calculated using a hypergeometric test. All p-values for overlaps between mouse and human DoG sets are < 2.0E-7.

Table S3

Primer	Target	Sequence	Species	application
Hsf1F	Hsf1	ggccttcctaaccaagctgt	mouse	qPCR
Hsf1R	Hsf1	agccatgttgttgtgcttga	mouse	qPCR
Hspa8F	Hspa8	ccaagaatcaggttgcaatga	mouse	qPCR
Hspa8R	Hspa8	ggaggacacttcctctgggta	mouse	qPCR
Ifitm2F	Ifitm2	agccttcttgtccaccaatg	mouse	qPCR
Ifitm2R	Ifitm2	ttcctgtccctagacttcacaga	mouse	qPCR
Hnrnpa2b1F	Hnrnpa2b1	cgcggaggtctttctcatct	mouse	qPCR
Hnrnpa2b1R	Hnrnpa2b1	atcccgcataaccacacagt	mouse	qPCR
doHspa8F	doHspa8	TGAGGAGACCAGAGCAAGGT	mouse	qPCR
doHspa8R	doHspa8	TCAAGTCTCCCCAAATCAGC	mouse	qPCR
doIfitm2F	doIfitm2	TTCCTCTTTGCCTGCTGTCT	mouse	qPCR
doIfitm2R	doIfitm2	AACCATTGTGGGTCGGTTTA	mouse	qPCR
doHnrnpa2b1F	doHnrnpa2b1	AATGGCAAGGTCCCATCTAAG	mouse	qPCR
doHnrnpa2b1R	doHnrnpa2b1	GCTGGCCTCATTTGCTATGT	mouse	qPCR
UhgF	snhg1	tgtgacaacatgaagagttcgag	mouse	qPCR
UhgR	snhg1	cccacaagtatggcactgct	mouse	qPCR
GapdhF	Gapdh	AGGTCGGTGTGAACGGATTTG	mouse	qPCR
GapdhR	Gapdh	GGGGTCGTTGATGGCAACA	mouse	qPCR
18SF	18S	CGAAAGCATTTGCCAAGAAT	mouse	qPCR
18SR	18S	GCATCGTTTATGGTCGGAAC	mouse	qPCR
doHspa8_592DS_F	doHspa8	GACTGCTCTTCTAGTTCCCTAATGT	mouse	PCR
doHspa8_592DS_R	doHspa8	AGTGATTGTCAGTACTTTCCAGGG	mouse	PCR

siRNA accession numbers		
siHSF1	QIAGEN	SI01071035
AllStars negative control	QIAGEN	SI03650318

Table S3: Primers and oligos used in the study.

References

1. Shalgi R, Hurt JA, Lindquist S, & Burge CB (2014) Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep* 7(5):1362-1370.
2. Vilborg A, Passarelli MC, Yario TA, Tycowski KT, & Steitz JA (2015) Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell* 59(3):449-461.
3. Giannakakis A*, et al.* (2015) Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. *Sci Rep* 5:9737.
4. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511-515.
5. Shalgi R*, et al.* (2013) Widespread regulation of translation by elongation pausing in heat shock. *Molecular cell* 49(3):439-452.
6. Mahat DB, Salamanca HH, Duarte FM, Danko CG, & Lis JT (2016) Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol Cell* 62(1):63-78.
7. Takii R*, et al.* (2015) ATF1 modulates the heat shock response by regulating the stress-inducible heat shock factor 1 transcription complex. *Mol Cell Biol* 35(1):11-25.
8. Ray D*, et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499(7457):172-177.
9. Reimand J*, et al.* (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* 44(W1):W83-89.
10. Yue F*, et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515(7527):355-364.
11. Pope BD*, et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature* 515(7527):402-405.
12. Vierstra J*, et al.* (2014) Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* 346(6212):1007-1012.
13. Maza I*, et al.* (2015) Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. *Nat Biotechnol* 33(7):769-774.
14. Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550.