Cell Reports, Volume *35*

Supplemental information

Parallel genetics of regulatory sequences using scalable genome editing *in vivo*

Jonathan J. Froehlich, Bora Uyar, Margareta Herzog, Kathrin Theil, Petar Glažar, Altuna Akalin, and Nikolaus Rajewsky

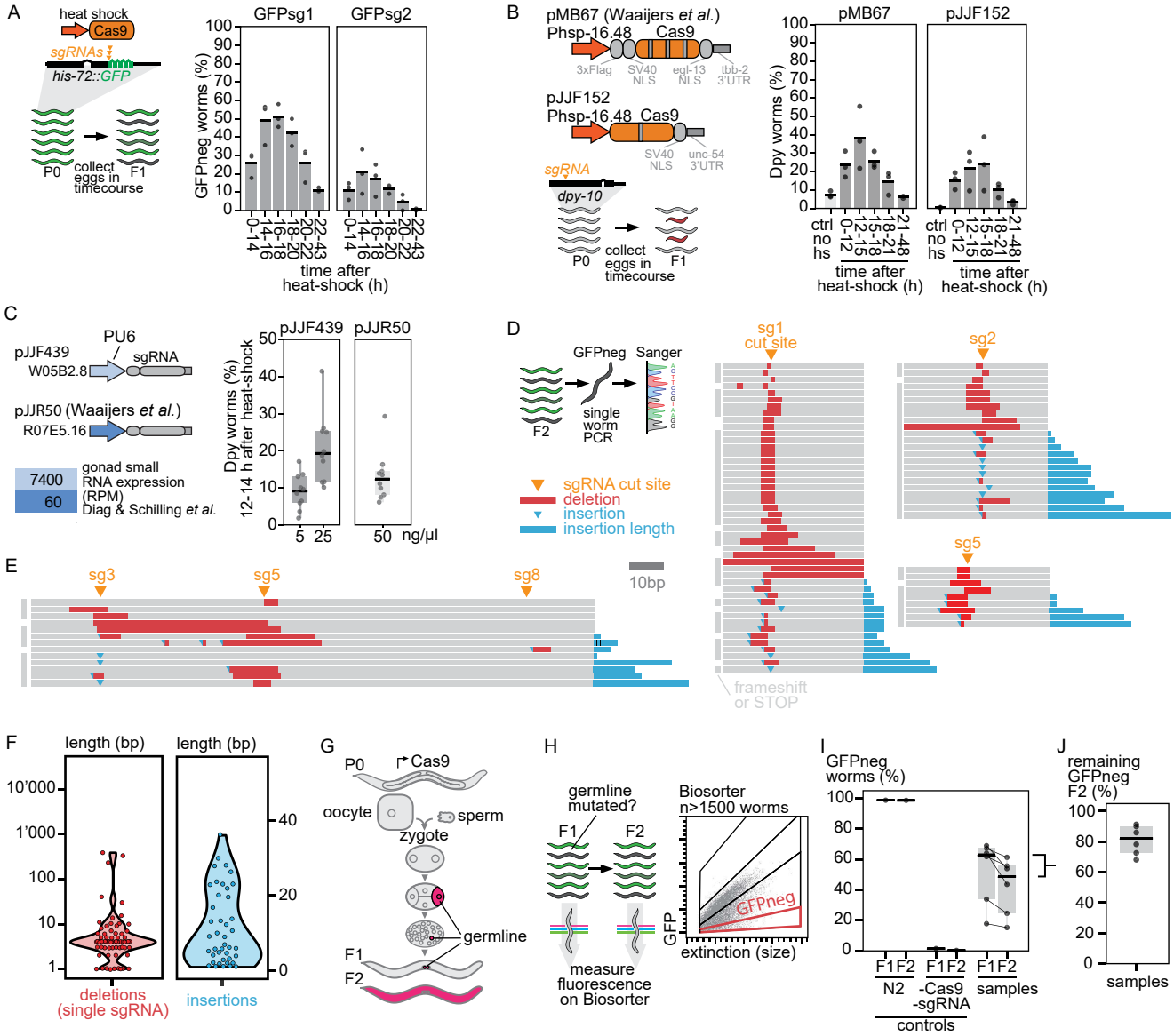

Figure S1. Transiently induced Cas9 expression creates germline indel mutations. Related to Figures 1 and 2. (A) Defining the temporal dynamics of Cas9 induction. An endogenously tagged *his-72::GFP* was targeted with two different sgRNAs. After a two-hour heat shock, eggs were collected in a time course and GFP-negative animals were counted. Experiment was conducted with 3 independent lines (n=3). The eggs collected 14 – 16 hours after heat shock produced the most GFP-negative animals. **(B)** Comparison of two different plasmids for heat shock inducible Cas9, pMB67 (Waaijers *et al.*, 2013) and pJJF152 (this study). *Dpy-10* coding sequence was targeted with a sgRNA ("dpy-10_CDS_sg1", pJJF449), time course was performed as in A) and Dpy progeny were counted. Experiment was conducted with 3 independent lines (n=3). Eggs collected 12 – 14 hours after heat shock produced the most Dpy animals. **(C)** Comparison of two different U6 promoters for sgRNA expression, in backbone plasmids pJJR50 (Waaijers *et al.*, 2016) and pJJF439 (this study), used at 5, 25 or 50 ng/ul in the injection mix. *Dpy-10* coding sequence was targeted with sgRNA "dpy-10_CDS_sg6". Eggs were collected 12 – 14 hours after heat shock and Dpy progeny were counted. Data from two experiments using 5 independent lines (n=10). Expression of U6 snRNAs in reads per million (RPM) was obtained from Diag & Schilling *et al.*, 2018. **(D)** Indel mutations detected by Sanger sequencing of individual GFP-negative animals after targeting *his-72::GFP* with sgRNAs. **(E)** Sanger sequencing of indel mutations created by a pool of three sgRNAs. **(F)** Length distribution of the indels from individual GFP-negative worms. Deletion length is shown only for the two lines with a single sgRNA. Insertion length is shown for all three lines including the line with a pool of sgRNAs. **(G)** A scheme showing the germline lineage in *C. elegans*. F2 animals are created by a germline cell which is determined in the F1 4-cell embryo. **(H)** Scheme showing automated fluidics measurement of F1 and F2 GFP negative animals to determine the amount of germline mutations. **(I)** Amount of GFP-negative F1 and F2 animals in control strains and after targeting his-72::GFP with sg1, sg2, pool1 or pool2. N = 1,662 - 21,983 analyzed worms per sample. **(J)** Difference in the amount of GFP-negative animals between F1 and F2 generation. Almost the same amount (80%) of GFP-negative animals in the F2 generations indicates high germline transmission of mutations.

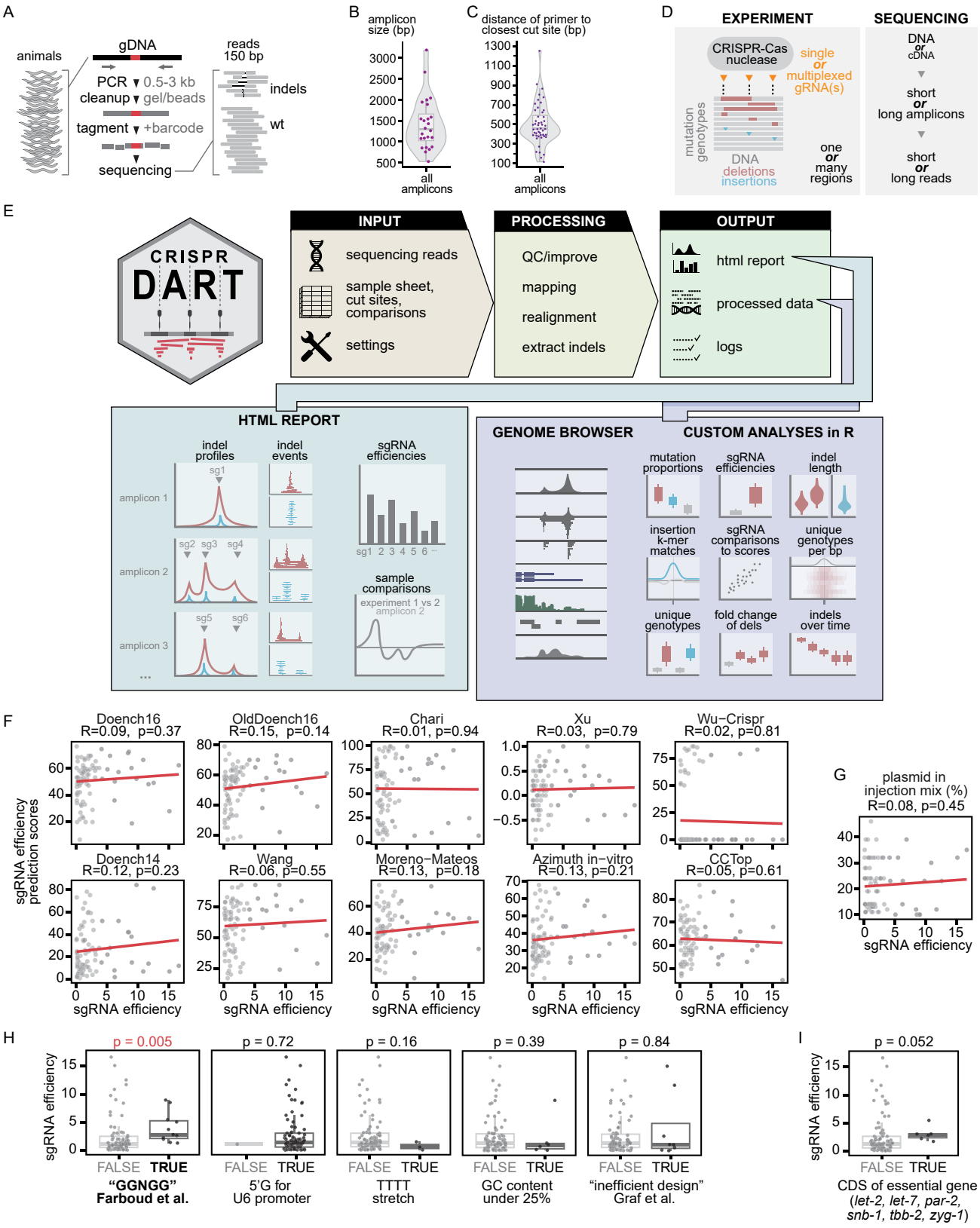

Figure S2. Software pipeline "crispr-DART" and sgRNA efficiency characteristics. Related to Figures 1 - 3. (A) Scheme showing our long PCR amplicon sequencing approach. **(B)** Size of the amplicons used for targeted DNA sequencing. **(C)** Distance between the amplicon PCR primers and the closest sgRNA cut site. **(D)** Examples to show the versatility of experimental data that can be analyzed with crispr-DART. **(E)** The software pipeline "crispr-DART". The user provides input files, and the pipeline produces processed genomic files and html reports. Custom analyses for this study were then performed with R scripts using the processed genomic files as input. For crispr-DART and R scripts see "Code Availability" in the STAR Methods. **(F)** Correlation of various prediction scores for sgRNA efficiency and our observed sgRNA efficiency (n=91 sgRNAs). **(G)** Correlation of the percentage of plasmid in the original injection mix and the observed sgRNA efficiency. **(H)** Comparison of sgRNA efficiency for different sgRNA features. Categories were compared using the Wilcoxon signed-rank test. **(I)** Comparison of sgRNA efficiency for sgRNAs targeting the coding sequence of essential genes and all other sgRNAs. Categories were compared using the Wilcoxon signed-rank test.

A

animals in sample	400,000	400,000	400,000	400,000	200,000
amplicon read coverage	200,000	400,000	800,000	1,200,000	1,200,000
coverage per animal	0.5	1	2	3	6
0.001% reads threshold	2 reads	4 reads	8 reads	12 reads	12 reads
5 reads threshold	0.0025%	0.0013%	0.0006%	0.0004%	0.0004%
acting threshold	5 reads	5 reads	8 reads	12 reads	12 reads
minimum animals to support one indel	**10**	**5**	**4**	**4**	**2**

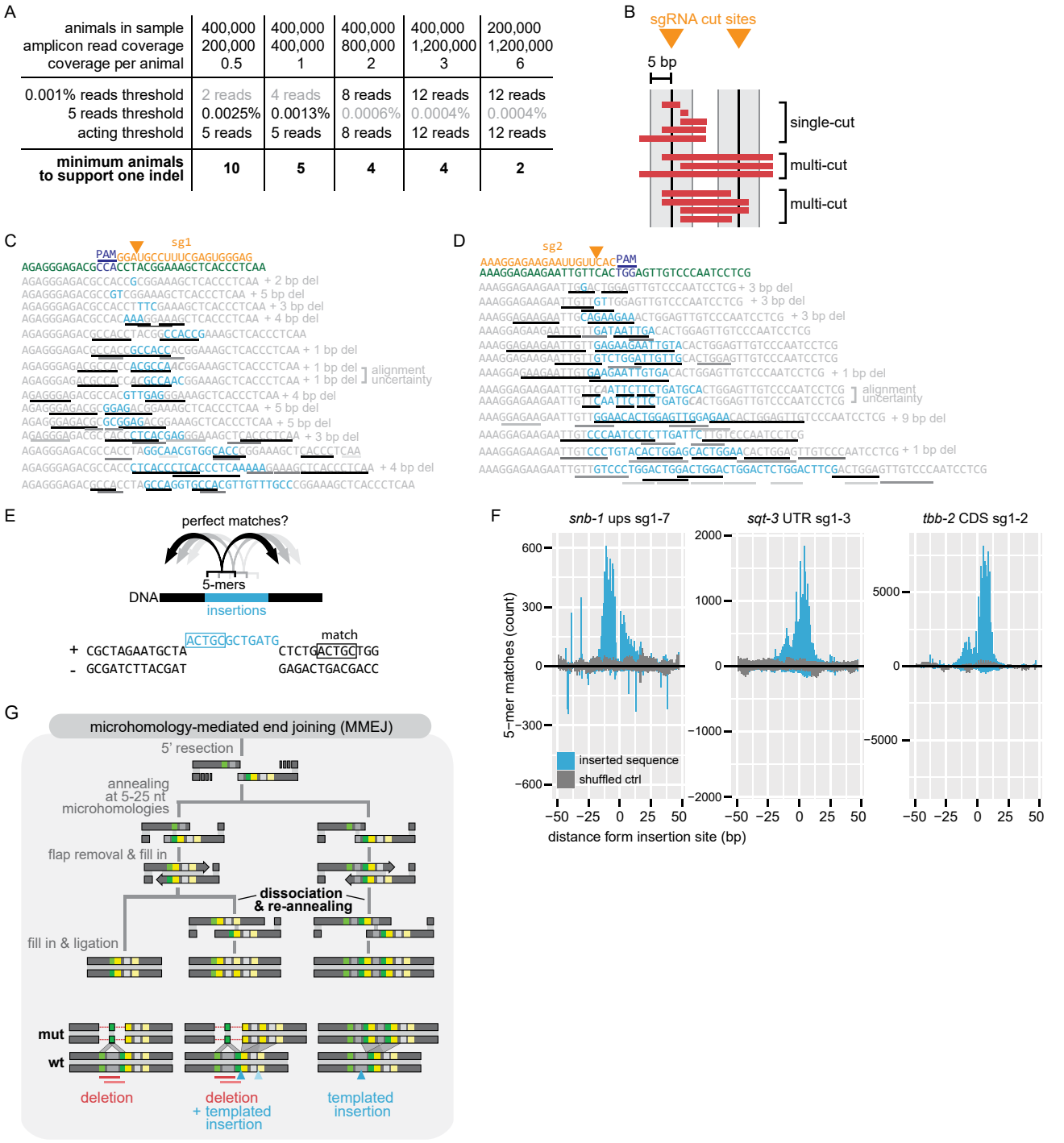

Figure S3. Indel detection and templated insertions. Related to Figures 2 and 3. (A) A table estimating the sensitivity of calling one indel present in the sequenced animal populations. In samples of lower coverage (e.g. 200,000-fold), the threshold of 5 reads acts, while for samples with higher coverage (e.g.800,000-fold) the threshold of 0.001 % reads acts. This results in usually 4-10 animals required to call an indel in our samples with 400,000 animals. **(B)** Scheme which illustrates how single-cut or multi-cut deletions were categorized computationally. **(C and D)** Examples of microhomology observed between insertions and surrounding regions in genotypes of GFP-negative his-72::GFP animals. **(E)** Scheme showing the analysis approach which matches all possible 5-mers from an insertion to the surrounding sequence. **(F)** Matches of 5-mers from insertions to surrounding sequence (+/- 50bp), shown for three samples. **(G)** Diagram showing mechanistic steps of dsDNA repair by microhomology-mediated end joining. Highlighted is the dissociation & re-annealing step which could lead to the templated insertions observed in our data.

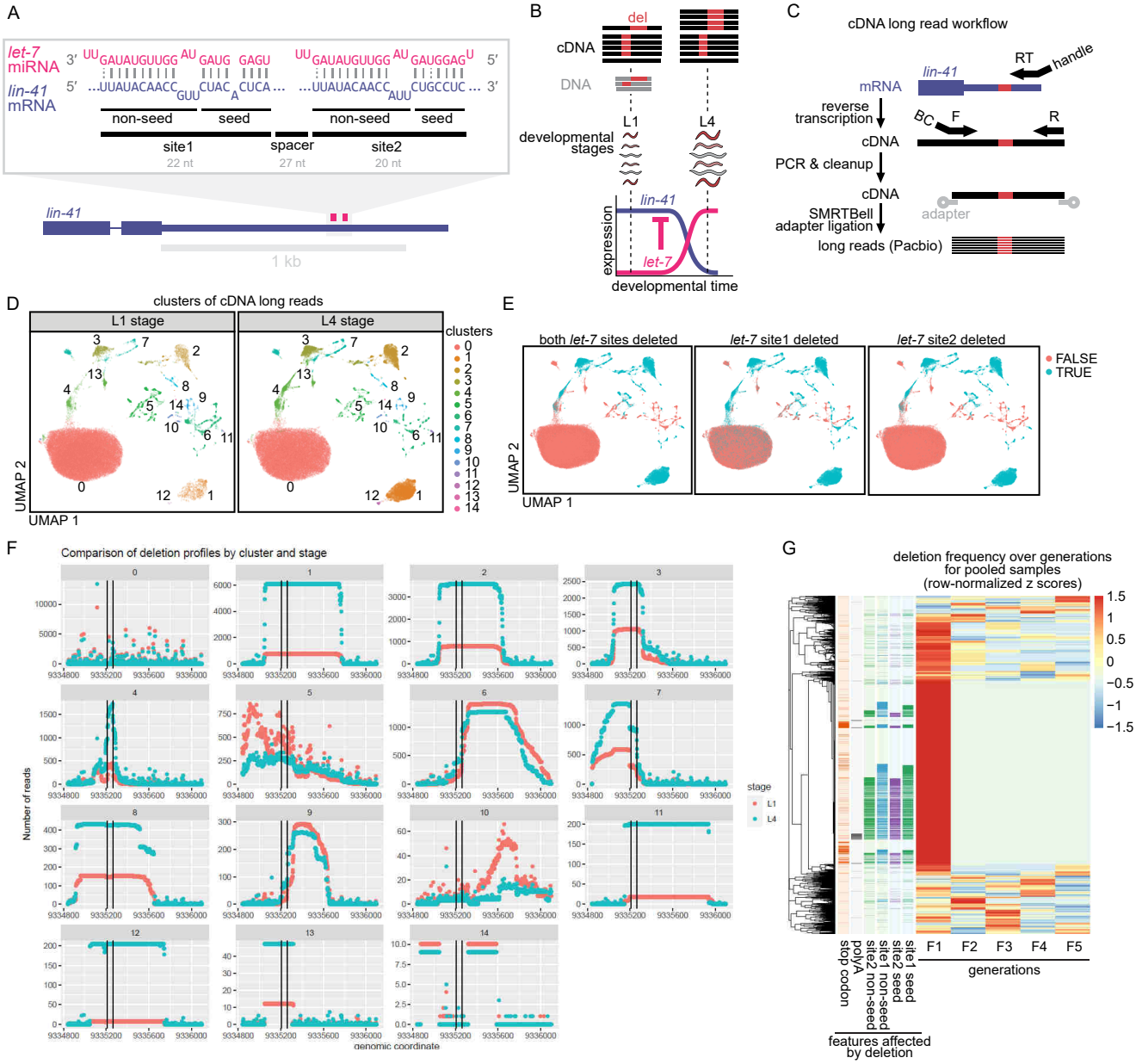

Figure S4. Impact of *lin-41* 3′ UTR deletions on RNA levels. Related to Figure 4. (A) Diagram showing *let-7* complementary sites site1, site2 in the *lin-41* 3′ UTR. **(B)** Diagram of *lin-41* and *let-7* developmental expression and time points of RNA extraction. **(C)** Diagram of the targeted RNA sequencing strategy. cDNA was amplified using a large amplicon and sequenced using the Pacbio long read workflow. **(D)** UMAP clusters of long reads covering the complete *lin-41* 3′ UTR, detected in cDNA from L1 or L4 developmental stages. Each dot represents one read. **(E)** Status of overlap with *let-7* sites for each read. **(F)** Number of detected reads with a deletion (y-axis) per genomic nucleotide (x-axis). Reads are separated by cluster (sub-panels) and developmental stage (L1=red, L4=green). The two vertical black lines indicate the location of the two *let-7* complementary sites (site1 and site2). **(G)** Heatmap displaying the frequency of deletions (on rows) scaled by row over multiple generations (columns). The annotation columns display which deletions overlap different features (e.g. *let-7* binding sites, polyA signal, stop codon). **(H)** Genotypes of strains with deletions in *let-7* complementary site1 and site2 in the *lin-41* 3′ UTR.

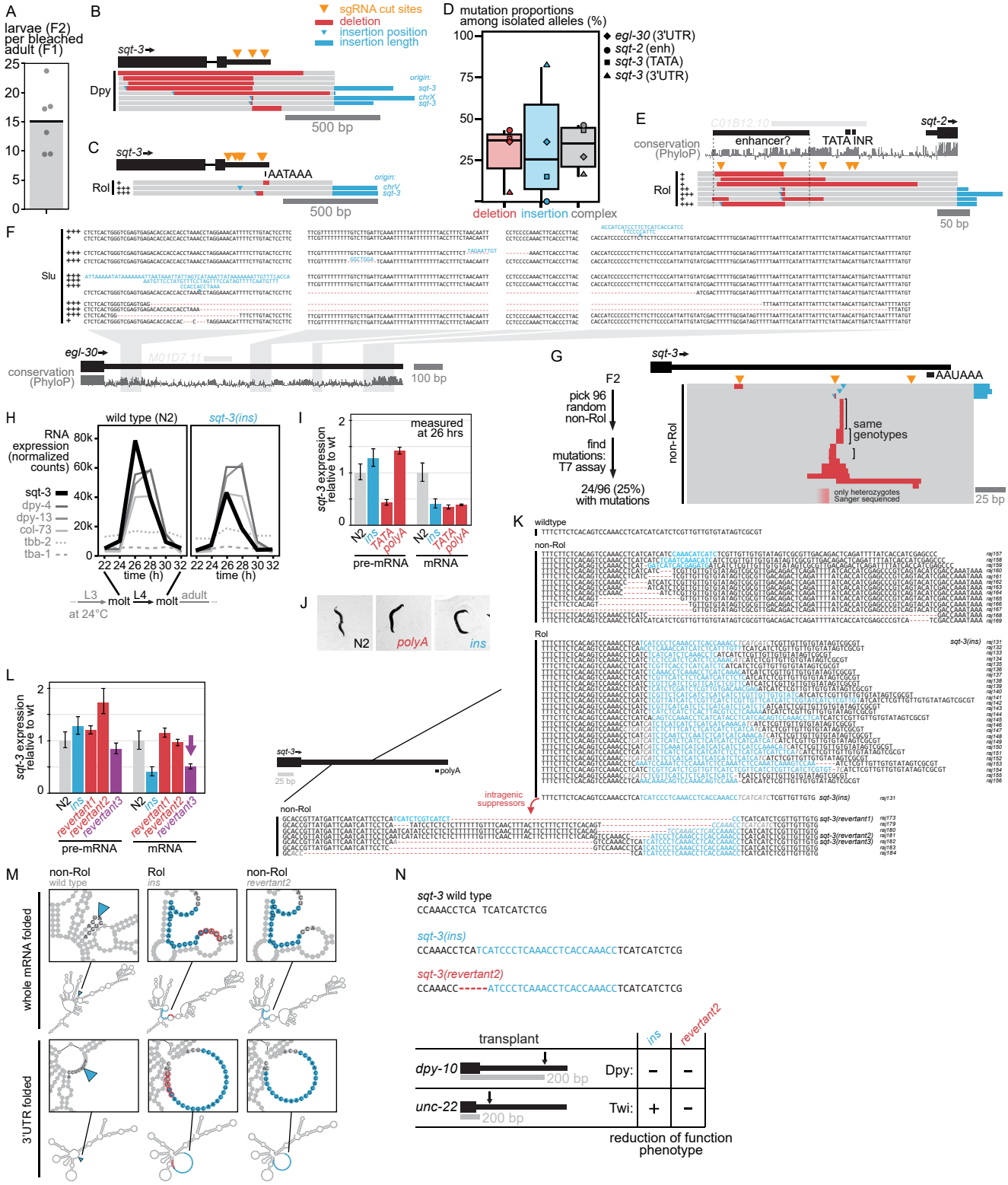

Figure S5. Screening for functional regulatory sequences using morphological phenotypes. Related to Figure 5. (A) Amount of F2 progeny obtained per bleached adult of the F1 generation which allowed us to perform targeted sequencing on the siblings of the screened animals. **(B)** Location and extent of mutations affecting the coding sequence in Dpy *sqt-3* mutants. For long insertions the origin was determined by BLAT. **(C)** Rol mutations isolated after targeting the *sqt-3* 3′ UTR without sg2. **(D)** Proportion of mutation types in the isolated reduction-of-function alleles from four targeted regions (*egl-30* 3′ UTR, *sqt-2* enhancer, *sqt-3* TATA-box, and *sqt-3* 3′ UTR). "Complex": alleles with a combination of insertion and deletion. **(E)** Indels affecting a putative enhancer region (Jänes et al. 2018) of *sqt-2*. +, ++, +++ indicate the expressivity of the trait. This was the only region for which penetrance was not complete (10-100%). **(F)** Sequences of the mutations in Uncoordinated *egl-30* mutants. **(G)** Isolation strategy of non-Rol mutants. Indels in non-Rol animals determined by Sanger sequencing. **(H)** Quantification of *sqt-3* RNA expression along development during L4 stage in wild type (N2) and sqt-3(ins) mutant. Worms were synchronized by bleaching and RNA was quantified on the Nanostring system. **(I)** *sqt-3* mRNA and pre-mRNA levels in different *sqt-3* alleles at 26 hrs into synchronized development. Levels were quantified by qPCR with primers specific for the spliced or the un-spliced transcript. Barplots show mean +/- standard deviation of technical triplicates. **(J)** Microscope images of the weak Rol phenotype with only slight bending of the head in the *sqt-3(polyA)* mutant and strong characteristic Rol phenotype in the *sqt-3(ins)* mutant. **(K)** Nucleotide sequences of relevant 3′ UTR regions in Rol, non-Rol and revertant mutants showing the inserted and deleted nucleotides. **(L)** mRNA and pre-mRNA levels of *sqt-3* in mutant and revertants at 26 hours into synchronized development. Levels were quantified by qPCR with primers specific for spliced or un-spliced transcript. Barplots show mean +/- standard deviation of technical triplicates. **(M)** Predicted RNA secondary structures of wild type, insertion mutant and revertant allele. Predictions were made for the whole mRNA or only the 3′ UTR. **(N)** Transplantation of mutant sequences into independent 3′ UTRs. Sequence 1 (from insertion mutant) or sequence 2 (from revertant2 of the insertion mutant) were knocked-in at the *dpy-10* or *unc-22* 3′ UTR. Seq1 led to the characteristic reduction-of-function phenotype Twitching (Twi) in *unc-22*.

gene	region	number of sgRNAs	amplicon size selected
lin-41	CDS	2	+/-
lin-41	3'UTR	5	+/-
lin-41	3'UTR	2	+/-
lin-41	3'UTR	5	+/-
lin-41	3'UTR	2	+/-
lin-41	3'UTR	8	+/-
lin-41	3'UTR	8	+/-
lin-41	3'UTR	8	+/-
lin-41	downs	2	+/-
dpy-2	3'UTR	2	+
dpy-10	3'UTR	2	+/-
dpy-10	3'UTR	2	-
dpy-10	3'UTR	4	-
egl-30	3'UTR	2	+
egl-30	3'UTR	2	+
egl-30	3'UTR	2	+
rol-6	enh	2	-
rol-6	TATA	2	-
rol-6	INR	1	-
rol-6	3'UTR	2	-
snb-1	ups	7	+
snb-1	CDS	2	+
snb-1	CDS	2	+
snb-1	3'UTR	8	+
snb-1	3'UTR	7	+
sqt-2	ups	3	-
sqt-2	TATA_INR	2	-
sqt-2	3'UTR	3	-
sqt-3	TATA	4	-
sqt-3	INR	2	-
sqt-3	3'UTR	3	+/-
sqt-3	3'UTR	3	-
sqt-3	3'UTR	6	-
sqt-3	3'UTR	9	-
unc-26	3'UTR	2	+
unc-54	3'UTR	3	+
let-2	CDS	2	-
let-2	3'UTR	6	-
let-7	miRNA	2	-
par-2	CDS	2	-
tbb-2	CDS	2	-
tbb-2	3'UTR	3	-
unc-119	CDS	1	+/-
zyg-1	CDS	2	-
zyg-1	3'UTR	3	-

Table S1. Overview of sequenced samples, Related to Figure 2 and 3.

gene	dels into coding	isolated mutants	phenotype	region	deletion	complex /insertion	proportion of deletions (%)
egl-30	-	11	Slu	3'UTR	4	7	36
sqt-2	+	7	Rol	enhancer	3	4	43
sqt-3	+	13	Rol	TATA box	5	8	38
sqt-3	+	26	Rol	3'UTR	1	25	4
dpy-2	+	0	-	3'UTR	-	-	-
dpy-10	+	0	-	3'UTR	-	-	-
rol-6	-	0	-	prom, TATA	-	-	-
rol-6	-	0	-	3'UTR	-	-	-
sqt-2	+	0	-	TATA	-	-	-
sqt-2	+	0	-	3'UTR	-	-	-
unc-26	-	0	-	3'UTR	-	-	-
unc-54	+	0	-	3'UTR	-	-	-
sqt-3(ins)	+	15	Rol->non-Rol	3'UTR	11	4	73

Table S2. Results of the screen for functional regulatory sequences using morphological phenotypes, Related to Figure 5.