Parallel Genetics of Gene Regulatory Sequences in *Caenorhabditis elegans*

D i s s e r t a t i o n
zur Erlangung des akademischen Grades
d o c t o r r e r u m n a t u r a l i u m
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M.Sc. Jonathan Johannes Froehlich

Kommissarischer Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Peter Frensch

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Dr. Christian Ulrichs

Gutachter:
1. Prof. Dr. Nikolaus Rajewsky
2. Prof. Dr. Baris Tursun
3. Dr. Christine Mayr

Tag der mündlichen Prüfung: 13.05.2022

Abstract

How regulatory sequences control gene expression is fundamental for explaining phenotypes in health and disease. The function of regulatory sequences must ultimately be understood within their genomic environment and development- or tissue-specific contexts. Because this is technically challenging, few regulatory elements have been characterized *in vivo*. Here, we use inducible Cas9 and multiplexed guide RNAs to create hundreds of mutations in enhancers/promoters and 3′ UTRs of 16 genes in *C. elegans*. We quantify the impact of mutations on expression and physiology by targeted RNA sequencing and DNA sampling. When applying our approach to the *lin-41* 3′ UTR, generating hundreds of mutants, we find that the two adjacent binding sites for the miRNA *let-7* can regulate *lin-41* expression largely independently of each other, with indications of a compensatory interaction. Finally, we map regulatory genotypes to phenotypic traits for several genes. Our approach enables parallel analysis of gene regulatory sequences directly in animals.

Zusammenfassung

Wie regulatorische Sequenzen die Genexpression steuern, ist von grundlegender Bedeutung für die Erklärung von Phänotypen in Gesundheit und Krankheit. Die Funktion regulatorischer Sequenzen muss letztlich in ihrer genomischen Umgebung und in entwicklungs- oder gewebespezifischen Zusammenhängen verstanden werden. Da dies eine technische Herausforderung ist, wurden bisher nur wenige regulatorische Elemente *in vivo* charakterisiert. Hier verwenden wir Induktion von Cas9 und multiplexed-sgRNAs, um hunderte von Mutationen in Enhancern/Promotoren und 3′ UTRs von 16 Genen in *C. elegans* zu erzeugen. Wir quantifizieren die Auswirkungen von Mutationen auf Genexpression und Physiologie durch gezielte RNA- und DNA-Sequenzierung. Bei der Anwendung unseres Ansatzes auf den 3′ UTR von *lin-41*, bei der wir hunderte von Mutanten erzeugen, stellen wir fest, dass die beiden benachbarten Bindungsstellen für die miRNA *let-7* die *lin-41*-Expression größtenteils unabhängig voneinander regulieren können, mit Hinweisen auf eine mögliche kompensatorische Interaktion. Schließlich verbinden wir regulatorische Genotypen mit phänotypischen Merkmalen für mehrere Gene. Unser Ansatz ermöglicht die parallele Analyse von genregulatorischen Sequenzen direkt in Tieren.

Authorship

Parts of this thesis – most of the results and methods – have been published in:

Parallel genetics of regulatory sequences using scalable genome editing *in vivo.*
Jonathan J. Froehlich*, Bora Uyar*, Margareta Herzog, Kathrin Theil,
Petar Glažar, Altuna Akalin, Nikolaus Rajewsky

Cell Reports, April 13th, 2021.
DOI: 10.1016/j.celrep.2021.108988. PMID: 33852857.

Contributions

J.J.F. and N.R. developed concepts and methodology and discussed the data. J.J.F. performed investigations and experimental work with help from M.H., and K.T. J.J.F. and B.U. performed validation, formal analysis, curation, and visualization of data. B.U. wrote the software with input from J.J.F. and A.A., and P.G. contributed to the software. A.A. and N.R. contributed resources, supervision, project administration, and funding acquisition.

Table of Contents

List of Figures and Tables

Introduction

Gene regulatory sequences are essential for genome function of multicellular life. Their function and logic can be studied in model organisms. However, current genetic methods are time-intensive and large-scale studies *in vivo* are still rare. This work describes a simple approach to test many gene regulatory sequences in *C. elegans* in parallel using inducible CRISPR-Cas genome editing.

Background: Studying gene regulatory sequences in animals

Genomes of multicellular animals contain protein coding DNA sequences (CDS), but mostly consist of non-coding DNA. This includes "gene regulatory sequences", stretches of hundreds, sometimes thousands of nucleotides, that determine the spatial- and temporal expression of proteins[1]. This function is usually encoded by 5-15 nt binding sites that recruit and coordinate regulatory factors. Promoters, enhancers, and silencers are bound by transcription factors (TFs) and increase or decrease transcriptional output[2–4]. 5′- and 3′ untranslated regions (UTRs) of messenger RNA (mRNA) can increase or decrease gene expression and affect sub-cellular localization and protein interaction[5,6]. This post-transcriptional regulation is often mediated by RNA binding proteins (RBPs) or microRNAs (miRNAs)[7–9] (Extended Background, p.53: details and illustrations related to study background).

Mutations in gene regulatory sequences can have dramatic consequences for animal phenotype and evolution[1,10–12]. This can affect animal morphology, development, behavior, and physiology. Regulatory mutations can also cause monogenetic disorders or contribute to complex diseases[13–16]. New therapeutic approaches can even target gene regulation using small molecules, therapeutic RNAs, or genome editing[17–22]. Therefore, to understand life, but also for medicine, it is necessary to predict phenotypic consequences of specific gene regulatory mutations. This includes identifying functional sequences, but also better understanding their logic, robustness, and connection to phenotype (**Figure 1A**).

However, predicting gene regulatory activity of a given sequence is hard. Nucleotide conservation has limited utility, because it can show little correlation to function at individual sequences[23–27]. Functional binding sites *in vivo* can diverge from optimal binding site motifs, with specificity contributed by their number and arrangement[28–31]. Often, multiple binding sites, their combination and arrangement determine the final activity of a gene regulatory sequence[4,32–35]. Separate sequences may cooperate ("AND" logic) or act compensatory ("OR"), mediated either directly (e.g., through interaction of bound factors) or indirectly (e.g., through feedback loops)[36–39]. For 5′ and 3′ UTRs, RNA properties such as copy number, secondary structure, and subcellular localization can determine interaction with gene regulatory factors and regulatory outcome[5,6,40].

Genetic methods have been developed in cell lines to successfully identify and increase our understanding of gene regulatory sequences *in vitro*. These assays usually follow a similar design: 1. Different mutations are randomly introduced into different cells in parallel by transduction or transfection. The goal is to create a high diversity of cells with distinct sequence mutations. 2. Cells are then selected by expression level of a reporter gene or cellular fitness by competitive growth. 3. Finally, the identity of the underlying genetic perturbation is read out by deep sequencing of cells selected by phenotype. This allows to link genotype and phenotype and establishes sequence–expression or sequence–fitness relationships. When the genetic perturbation and expression level can be determined from RNA directly, step 2 can be skipped (**Figure 1B**)[41–43].

Such approaches can be described as "parallel genetics" which come in several variations[43]. Massively parallel reporter assays (MPRAs) produce quantitative sequence-expression relationships for millions of sequences in reporter genes[41–44]. More recent genetic methods target native genomic

sequences and measure effects on expression or cellular fitness. The CRISPR-nuclease Cas9 can be directed to genomic sequences with a single guide RNA (sgRNA), where it may create a double strand DNA break that is repaired by cellular pathways, which can result in insertion-, deletion- mutations (indels), or a combination of both (complex indels)[45–47] (see also Extended Background and Extended Background Figure 5C,D). Using parallel delivery of different sgRNAs along a region of interest, CRISPR tiling screens can identify functional sequences in genomic regions up to the kilobase scale[48–55]. In a related method, 10-20 multiplexed sgRNAs create many diverse deletions at one 3′ UTR for which mRNA levels are then measured using targeted mRNA sequencing[56–58]. Endogenous variant testing introduces programmed point mutations using more advanced CRISPR-Cas-based techniques with homology-dependent repair, base-, or prime editing[59–62]. While all these "parallel" approaches can determine sequences with regulatory activity in high throughput, they are so far mostly restricted to cell lines and yeast.

There are two good reasons to study gene regulatory sequences in living animals (*in vivo*). First, gene regulatory activity often depends on cell type and developmental stage, which can be studied simultaneously in developing animals[1]. Second, most aspects of phenotype are only observable *in vivo*, in complete organisms interacting with their surroundings. This extends to the study of the diverse mechanisms that can affect how a genotype translates to phenotype and its penetrance, expressivity, or plasticity[63–71] (List of Terminology, p.51).

Several genetic methods can test mutations of gene regulatory sequences in animals, but most of these methods cover only a few sequence variants at a time. Apart from using single reporter genes one by one, around eighteen studies use parallel reporter approaches to test several thousand sequences *in vivo* (**Figure 1C**). However, to evaluate phenotype and preserve natural regulation, mutations of the native genomic sequence (endogenous mutations), are required. Animals carrying alleles from forward genetics screens can be used, but they are difficult to isolate and maintain in large numbers. CRISPR-Cas systems have revolutionized genome editing[72], but most studies of gene regulatory sequences still rarely examine more than a handful mutations (**Figure 1C**). Larger animals like *M. musculus* and *D. rerio* require long generation times, large culturing spaces, ethical considerations; as well as work-intensive injection, line generation, and maintenance. This limits observations often to a particular tissue or developmental stage and makes it difficult to study phenotypes across the whole life cycle or many individuals (Supplemental Table 1 and Supplemental Table 2). For smaller animals, like *D. melanogaster* and *C. elegans*, current limitations are mainly due to the injection of genome editing reagents and missing parallel workflows.

Caenorhabditis elegans is a small nematode worm, of which millions can be routinely cultured for large-scale experiments across its complete 3-5 day life cycle[73]. Gene expression can be analyzed simultaneously across diverse cell types, specialized tissues, and developmental processes[73–76]. It is useful for genetic screens of behavior, morphology, and numerous other phenotypes; which has enabled many discoveries of metazoan gene regulation – in particular post-transcriptional regulation; while excellent genome-wide resources and annotations are available (Supplemental Table 3 and Supplemental Table 4)[77,78]. Despite this potential, current genetic methods in *C. elegans* can only test tens to a hundred mutations one by one: several orders of magnitude fewer than parallel approaches in other animals or cell lines. ♦

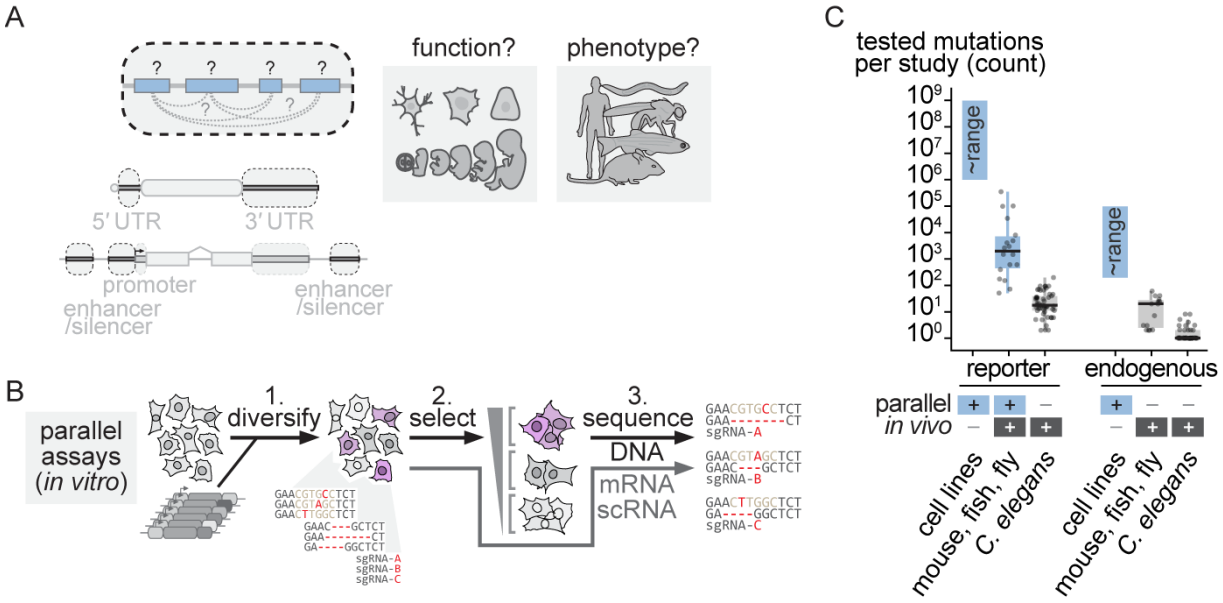


Figure 1. Limited scale of genetic methods *in vivo*

(A) Functions of gene regulatory sequences are difficult to predict and depend on cell type, tissue, development, and environment. Impact on phenotype additionally needs to be tested *in vivo*. **(B)** Basic workflow of parallel assays (like MPRAs, CRISPR tiling, or variant testing). Second step can be skipped with RNA sequencing to associate RNA levels with genotypes. **(C)** Number of tested mutations per study. "~range" for cell lines indicates the typical range of such studies without references. *In vivo* primary studies are listed in Supplemental Tables 1, 2, 5, 6 (number of studies: reporter-based, "mouse, …" n=18, "*C. elegans*" n=50; endogenous, "mouse, …" n=15, "*C. elegans*" n=40).

Research focus: Genetics of gene regulatory sequences in *C. elegans*

In *C. elegans* the activity of gene regulatory sequences is usually tested with fluorescent reporter genes in the translucent animal. Producing transgenic *C. elegans* requires microinjection of DNA into the gonad[79]. This results in a fraction of transgenic progeny carrying the injected DNA as extra-chromosomal arrays, which must then be selected with fluorescent markers, antibiotic resistance or treated further for genomic integration[80]. Endogenous mutations of regulatory sequences can be obtained in two ways. Either using random chemical mutagenesis, selection by phenotype, and genetic mapping (forward genetics). Or by genome editing of pre-determined sequences with CRISPR-Cas systems (reverse genetics)[81]. Forward genetics however is inefficient, and only around 35 non-coding mutations have been isolated in the last ~40 years[82]. Genome editing is commonly performed by microinjection of Cas9- and sgRNA plasmids or the assembled ribonucleoprotein, followed by separation of resulting progeny and genotyping the targeted loci to identify mutants with indel mutations[81]. Because microinjection is time-consuming, and the extra work on individually separated animals, current studies in *C. elegans* rarely cover more than a handful endogenous mutations (**Figure 1C**) (Supplemental Table 5 and Supplemental Table 6). This means that very few of all possible variants are tested, which limits the possibility to discover rules and new mechanisms. There are currently also no methods to automate or circumvent injections at sufficient scale[83–85]. In addition, no targeted sequencing is established to measure expression or competitive fitness of specific gene regulatory sequences from bulk populations[86]. Altogether these factors currently restrict larger, parallelized studies of gene regulatory sequences and their connection to phenotype in *C. elegans*. ♦

Aim

This work describes the development of an approach that can create many diverse mutations along a specified gene regulatory region in *C. elegans* populations, and link these to function and phenotype by deep sequencing or manual analysis.

Significance

Applied in *C. elegans* such a systematic and explorative method would increase our understanding of fundamental functions of gene regulatory sequences and their relationship to animal physiology. More specifically, this work will directly benefit *C. elegans* researchers that study gene regulation or wish to manipulate the genome at large scales. Furthermore, this study systematically analyzes the endogenous *lin-41* 3′ UTR and its two *let-7* miRNA binding sites, explores gene regulatory mutations that affect animal morphology, and provides a large dataset of *in vivo* indel mutations.

Hypothesis

To introduce many different, targeted mutations in *C. elegans*, we decided to use inducible expression of Cas9 and multiplexed guide RNAs along a pre-selected, regulatory genomic region. This would require few injections to create transgenic animals and allow maintenance and population expansion without Cas9 activity. Induction of Cas9 in a large sample could then lead to offspring with different indel mutations by the stochastic combinatorial activity of multiplexed guide RNAs. The number of animals with mutations would be mainly limited by culturing approaches (e.g., ~10^6), multiplied by the achieved mutagenesis efficiency. Mutations could then be connected to phenotype or reporter activity, or directly to RNA levels and competitive fitness using targeted sequencing in bulk (**Figure 2**).

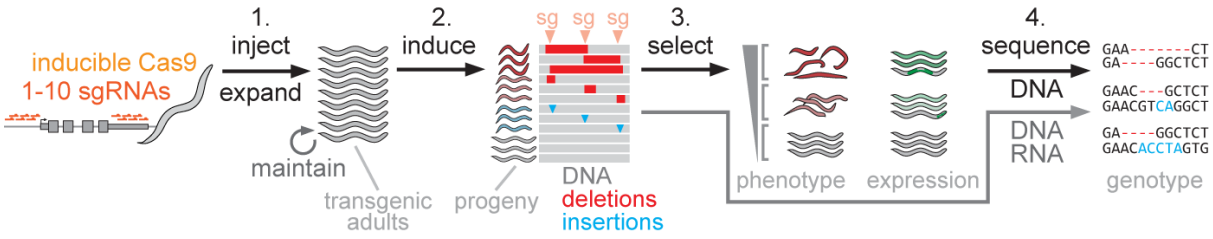


Figure 2. An approach for scalable parallel genetics in *C. elegans*
Heat-shock inducible Cas9 expression would allow targeted mutagenesis of large populations in parallel. Mutated populations could then be used to systematically link phenotype and gene expression to individual genotypes.

Specific objectives and questions

- test and optimize heat-shock induced Cas9 mutagenesis in *C. elegans* populations.
- analyze characteristics and diversity of mutations.
- use targeted RNA- and DNA sequencing to measure expression and fitness of 3′ UTR mutations.
- map gene regulatory mutations to morphological phenotypes.

~ does transient heat-shock-induced Cas9 expression create targeted indel mutations?
~ which sequence diversity do mutations create along the targeted regions?
~ can targeted RNA- and DNA sequencing measure expression and fitness from bulk populations?
~ can gene regulatory mutations systematically be mapped to morphological phenotypes?

Outline

The following results chapter is structured into four sections according to the four research objectives. First, heat-shock induced targeted mutagenesis with Cas9 is demonstrated and optimized. Second, large DNA amplicon sequencing and a computational pipeline are established and used to measure characteristics of 12,700 indel mutations from many experiments. Third, with over 900 different deletions along the *lin-41* 3′ UTR, targeted RNA and DNA sequencing, we show that two *let-7* miRNA binding sites interact compensatory in regulating mRNA levels and phenotype. Fourth and final, we target regulatory regions of 8 genes, isolate 57 mutations using associated morphological defects, and show genetic interaction within non-coding regions.

Results

Cas9 induction for parallel and targeted mutagenesis in *C. elegans*

As an initial test, we generated transgenic lines with plasmids for heat shock-driven Cas9 expression and one- or multiple sgRNAs targeting a ubiquitously expressed single-copy GFP reporter. After a transient heat shock, we could observe GFP-negative animals in culture, indicating activity of Cas9. We performed a two-hour heat shock induction of Cas9 in the parents (P0) and collected progeny (F1) in a time course experiment. The highest fractions of mutants were obtained 14 – 16 hours after heat shock, with approximately 50% (sg1) and 20% (sg2) of eggs producing GFP negative animals (**Figure 3A**). We obtained similar results when we targeted the *dpy-10* gene and counted the characteristic Dumpy (Dpy) phenotype comparing two plasmids for heat shock-induced expression of Cas9. The eggs collected 12 – 15 hours after heat-shock produced around 20 – 35% Dpy animals with both plasmids (**Figure 3B**). This time window overlaps with observations from genome editing by ZFN and TALEN (10-14 hrs), and is close to the time of extra-chromosomal array formation (8-12 hrs)[87,88]. We also tested a sgRNA U6 promoter with a reported higher gonad expression[89] and found that this resulted in a larger number of Dpy progeny on average (**Figure 3C**).

Characteristic CRISPR-Cas9-induced mutations from 91 GFP negative animals consisted of deletions or insertions (indels) or a combination of both and originated from sgRNA cut sites (**Figure 4A**). When we used three sgRNAs within the same transgenic line, targeting adjacent positions, deletions appeared around one cut site or spanned between two cut sites (**Figure 4B**). This indicated that pools of sgRNAs could lead to more diverse genotypes and cover more nucleotides. Most deletions induced by a single sgRNA were between 3 – 10 bp long and we observed insertion lengths between 1 – 30 bp (**Figure 4C**).

Homozygous animals would be produced in the F2 by heterozygous self-fertilizing F1. Additionally, since Cas9 induced in the P0 could still be active after fertilization, F1 animals could be mosaic with a wild-type germline and mutant somatic cells (**Figure 5A**). We therefore wanted to assess how many germline mutations were generated. For this we analyzed the inheritance of the GFP negative phenotype from F1 to F2 generations using an automated flow system and found that ~80% of mutations were indeed germline mutations (**Figure 5B–D**). For the rest of our work, we used such non-mosaic F2, generated by F1 germline mutations.

Homology-dependent repair would allow to install programmable mutations (or random nucleotides) in parallel. To test if inducible Cas9 also supported this we designed an experiment to restore the coding frame of a non-functional GFP. The non-fluorescent *his-72::GFP* allele was obtained in our previous experiments that resulted in ~24% (11/46) of GFP negative mutants with thw same 4 bp deletion ("4delGFP") (**Figure 4A**). We prepared HDR templates using PCR from the original *his-72::GFP* locus and used a sgRNA specific to the mutant 4delGFP allele. We also tested if availability of the repair template from extra-chromosomal arrays could be enhanced with sgRNA-targeted overhangs (**Figure 6A**). We observed restored GFP fluorescence in 1.5 and 4.9 % of F2 progeny (with and without overhangs respectively) (**Figure 6B**). After sorting animals with restored GFP levels we evaluated if the expression pattern matches the known *his-72* expression and determined genotypes by Sanger sequencing. SgRNA-targeted overhangs drastically increased correct editing. Most GFP positive animals from the experiment with the blunt HDR template carried new indels in the 4delGFP coding frame and had ectopic expression patterns which suggested possible random GFP insertions (**Figure 6C–E**). We did not measure the dynamics of this in a time course, but studies indicate HDR-edited progeny would peak later than indel mutations, >24 hrs after injection[47,87,90,91].

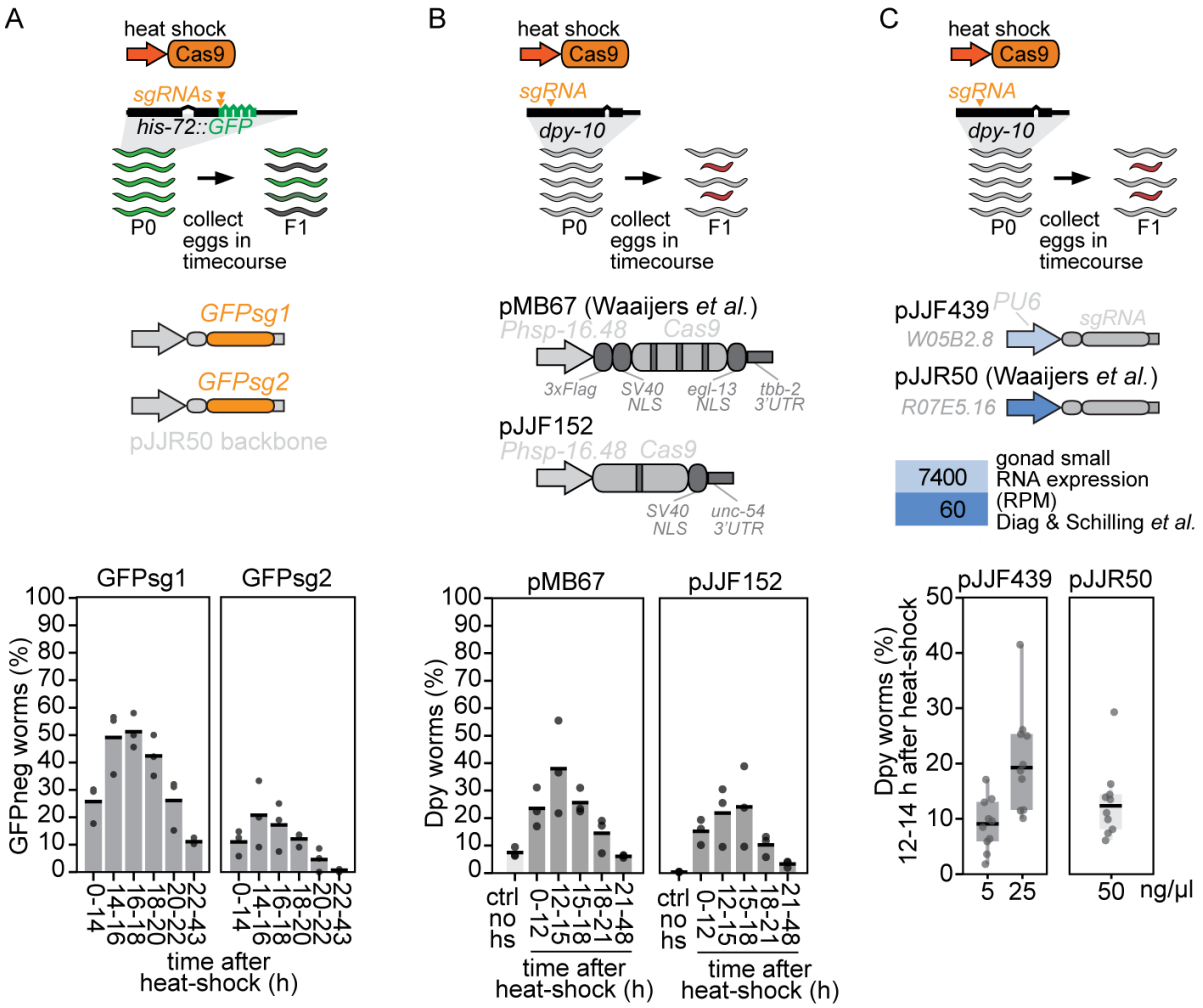


Figure 3. Optimization of workflow and plasmids for induced Cas9 mutagenesis

(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[92] 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[93] and pJJF439[(this study)], used at 5, 25 or 50 ng/µL in the injection mix. *Dpy-10* coding sequence was targeted with sgRNA "dpy-10_CDS_sg6". Data from two experiments using 5 independent lines (n=10). Expression of U6 small RNAs in reads per million (RPM) was obtained from ref.[89].

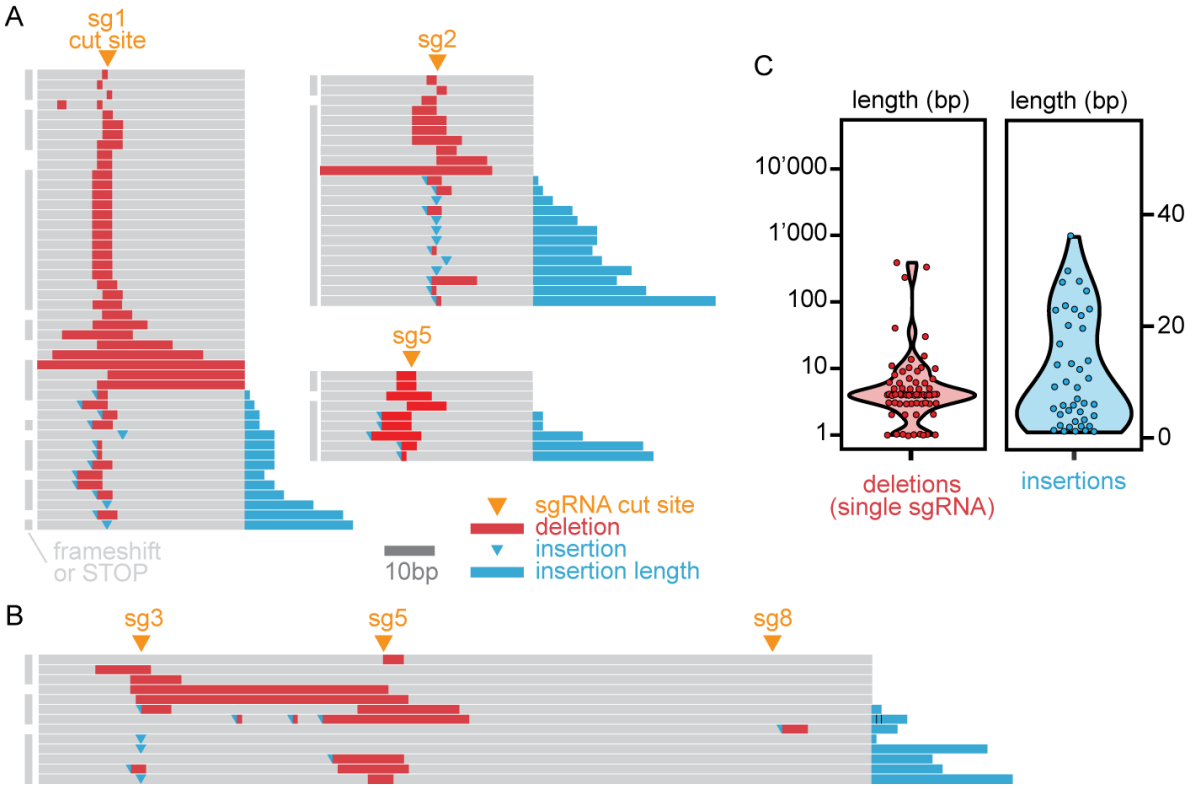


Figure 4. Indel mutations created by transiently induced Cas9

(A) Indel mutations detected by Sanger sequencing of individual GFP-negative animals after targeting *his-72::GFP* with sgRNAs. **(B)** Sanger sequencing of indel mutations created by a pool of three sgRNAs. **(C)** 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.

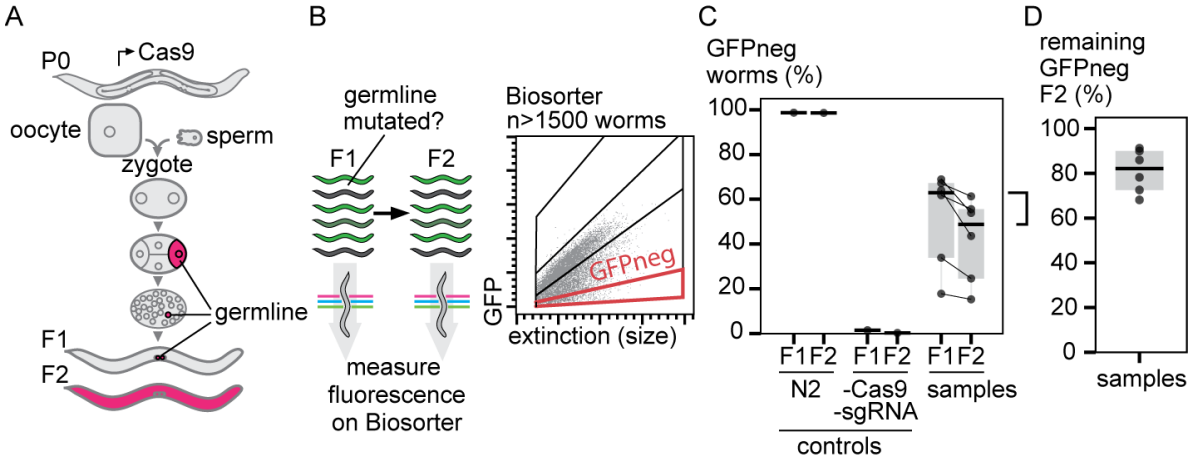


Figure 5. >80% indels are germline mutations

(A) 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. **(B)** Scheme showing automated fluidics measurement of F1 and F2 GFP negative animals to determine the amount of germline mutations. **(C)** 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. **(D)** Difference in the number 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 mutation

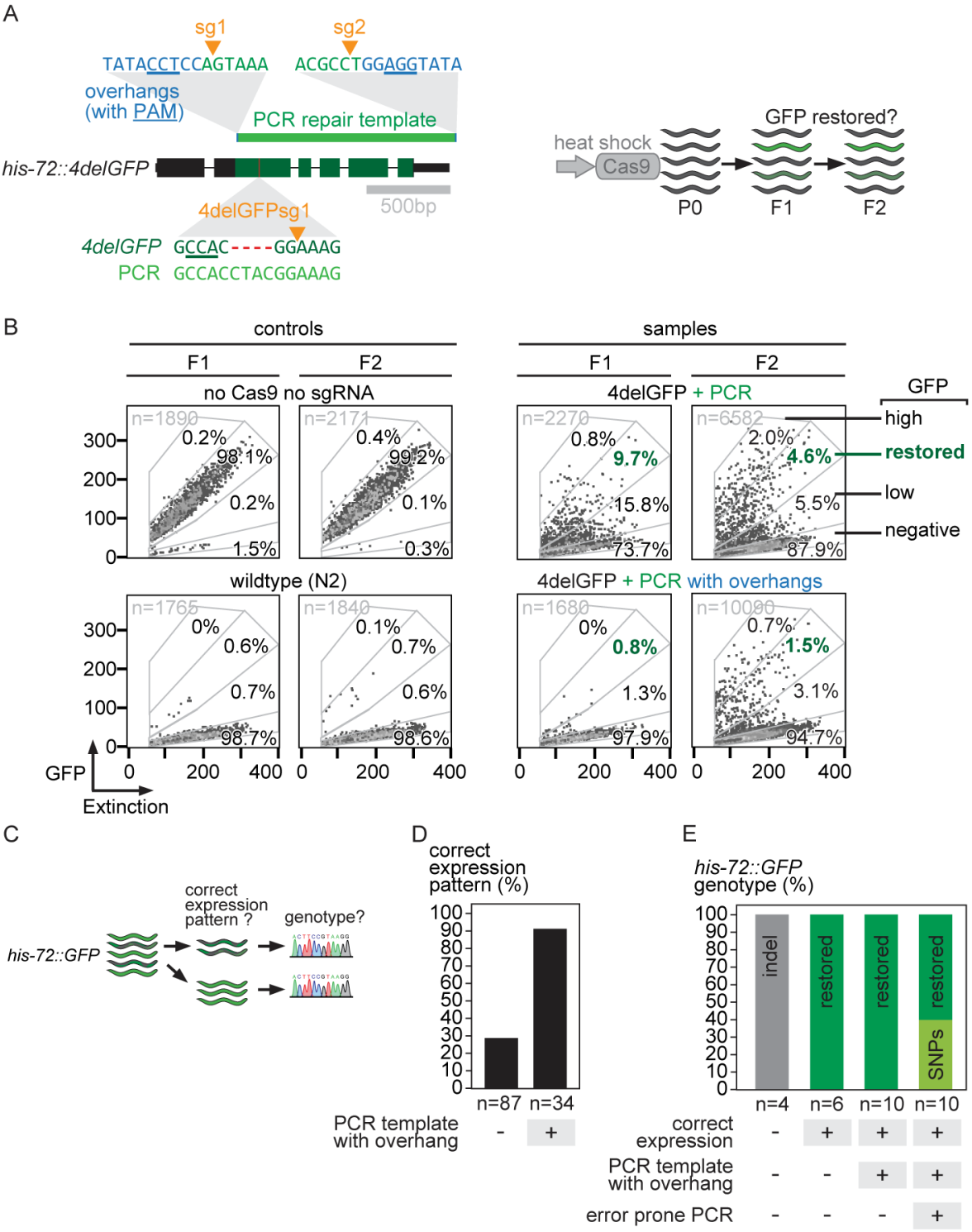


Figure 6. Induced homology-dependent repair

(A) Experimental design to restore a non-fluorescent GFP (4delGFP) using inducible Cas9 and homology-dependent repair. **(B)** Scatter plots of Biosorter analysis and sort the "GFP restored" gate. **(C)** Scheme of experimental follow-up. **(D)** Percent of animals that show a GFP expression pattern matching the known *his-72* expression ("correct"). **(E)** Genotype analysis of different fractions of animals, based on experimental condition and "correct" expression pattern.

To analyze large populations of mutated *C. elegans* in bulk, we established a targeted sequencing protocol based on long 0.5-3 kb PCR amplicons. This allowed us to sequence the complete targeted locus, to place most primers >300 bp away from the nearest sgRNA cut site to avoid deletion of primer binding sites, and to capture very large deletions. Barcoding samples enabled combined sequencing on the same flow cell (**Figure 7A**). To handle targeted sequencing data of such amplicons and analyze the contained mutations we created the software pipeline "crispr-DART" ("CRISPR-Cas Downstream Analysis and Reporting Tool") (https://github.com/BIMSBbioinfo/crispr_DART)[94]. The pipeline extracts and quantifies indels from various targeted sequencing technologies, single- or multiple regions of interest and single- or multiplexed sgRNAs. The output contains html reports of coverage, mutation profiles, sgRNA efficiencies and optional comparisons between pairs of samples. Processed genomics files from the output can then be used for more in-depth custom analyses with additionally supplied R scripts[94] (**Figure 7B**).

To test our approach in larger scale, we induced Cas9 in 50,000 P0 animals by heat shock, and amplicon-sequenced the mutated locus from bulk samples of 400,000 F2 progeny. Deletions per genomic base-pair peaked sharply around sgRNA cut sites (**Figure 7C**). Pools of multiplexed sgRNA plasmids resulted in deletions spanning two or several sgRNAs ("multi cut") in addition to smaller deletions surrounding single sgRNAs ("single cut") (**Figure 7C, bottom**). Insertions occurred within a few nucleotides to cut sites and were less frequent than deletions (~1/2 – 1/10th) (**Figure 7D**). We observed background mutations of short 1 bp deletions and insertions that were also present in similar abundance in isogenic wild type controls and that occurred independent of sgRNA cut sites. These could have been caused by biological (e.g., DNA modifications, natural mutations) and technical factors (e.g., during or after extraction of genomic DNA, PCR, sequencing errors). Such mutations were absent in genotyping by Sanger sequencing, and we later established computational filters to separate these from CRISPR-Cas9-induced mutations.

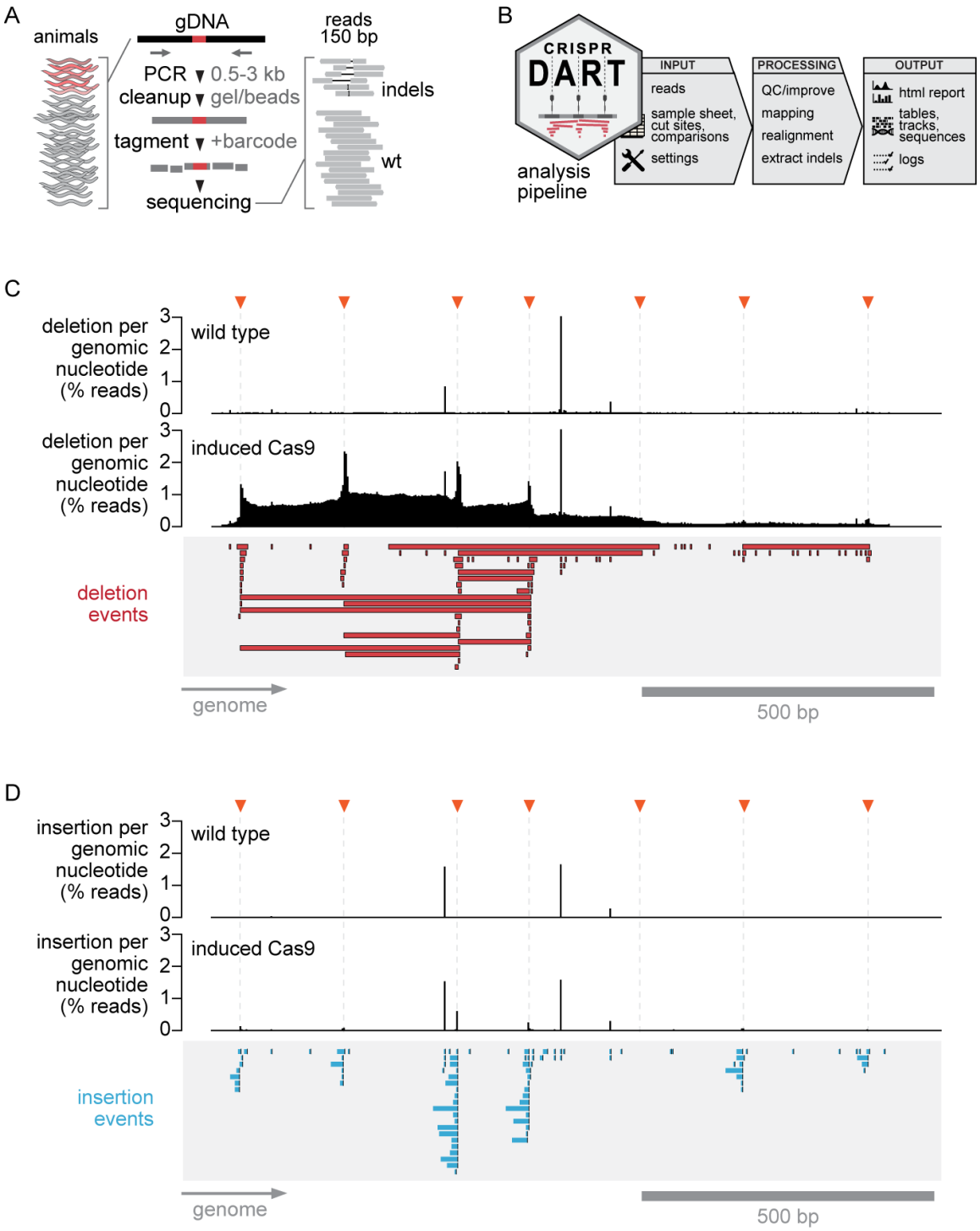


Figure 7. Long amplicon sequencing to detect indels

(A) Scheme showing our long PCR amplicon sequencing approach. **(B)** 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 more information see ref.[94]. **(C)** Example of the complete spectrum of observed mutations after targeting a locus. The percentage of DNA sequencing reads containing deletions with respect to the total read coverage is plotted at the corresponding genomic position. Bulk worm samples were sequenced, thus 2% deletions per genomic nucleotide refers to approximately 2% of worms with a deletion at the respective nucleotide. Orange triangles: sgRNA cut sites. Individual deletion events below in red. **(D)** Same analysis as in C) but for insertion events.

Features of 12,700 CRISPR-Cas9-induced indels

To understand gene regulatory logic, ideally many different variants are produced at high efficiency, which can then be tested for their effects *in vivo*. We set out to analyze the efficiency and characteristics of mutations produced with our approach. We targeted 16 genes at different regions with 1-9 sgRNAs per transgenic line. These genes were selected for different downstream experiments and contained one gene with a known miRNA interaction, 8 genes with known reduction-of-function phenotypes, and 7 essential genes. After Cas9 heat shock-induction, we sequenced bulk genomic DNA from 400,000 F2 animals with long amplicon sequencing. Together with wild type controls this produced data for 36 samples, 127 sgRNAs and 12,700 indels (**Figure 8A**) (Supplemental Table 7). Large amplicons of 0.5-3 kb allowed us to place most primers more than 300 bp from the next sgRNA cut site, to amplify DNA also from animals with larger deletions (**Figure 8B–C**). To measure sgRNA efficiencies, we counted all reads with deletions overlapping +/- 5bp of a given sgRNA cut site and normalized this value by the number of total reads at that position. The median efficiency was 1.4% with most sgRNAs showing efficiencies 0 – 6.3% (95% CI) (**Figure 8D**). 1.4% corresponded to approximately 5,600 mutant animals per sgRNA in our samples. The overall lower efficiency of mutagenesis in our large-scale experiments, compared to our initial small-scale experiments, might be due to less well optimized heat shock conditions on large plates, but is similar to efficiency of heat-shock induced *Mos1* transposase in a previous study[95–97].

We then compared observed sgRNA efficiencies to published efficiency prediction scores but found no score with significant predictive power (**Figure 9A**). Possible reasons for this could be that these scores were obtained in other experimental models, mostly human cell lines, or that sequence-independent factors were dominating in our system. Also, injected plasmid concentrations, used to generate transgenic lines, were not correlating with efficiency (**Figure 9B**). We found however, that sgRNAs for target sites with GG preceding the PAM ("GGNGG") were significantly more efficient, as previously described for *C. elegans*[98] (**Figure 9C**). SgRNA efficiencies were likely not confounded by lethal phenotypes - by depleting for animals with efficient sgRNAs - because sgRNAs targeting essential genes did not show reduced efficiencies compared to other sgRNA (**Figure 9D**).

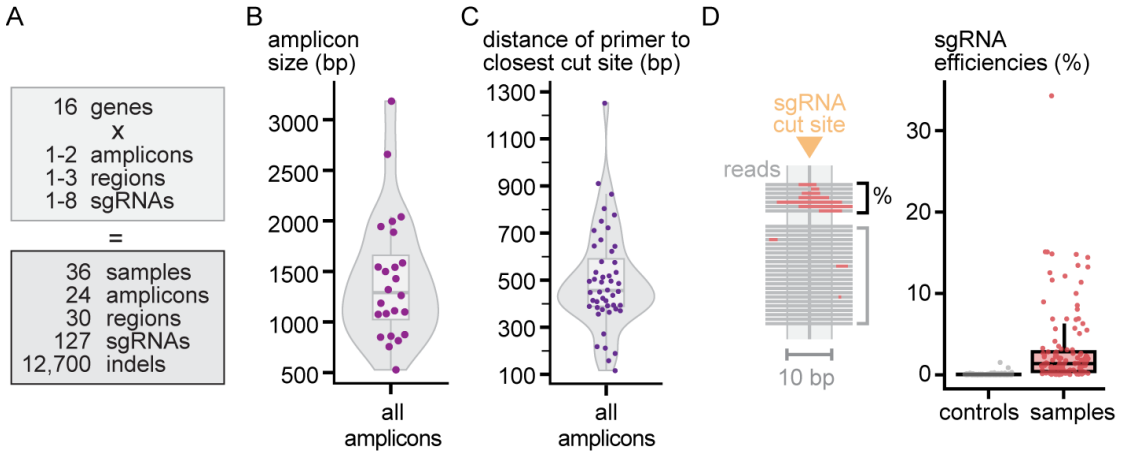


Figure 8. Targeting regulatory regions of 16 genes

(A) Overview of data collected from targeting regulatory regions of 16 genes with multiplexed sgRNAs. **(B)** Size of amplicons used for targeted DNA sequencing (n=24 amplicons). **(C)** Distance of primers to closest sgRNA cut site (n=48 primers, two primers per amplicon). **(D)** Estimation of sgRNA efficiencies (n=127 sgRNAs) (n=24 wild type controls, n=36 samples with induced Cas9). Each sample expresses 1-8 sgRNAs targeting one region among 16 genes.

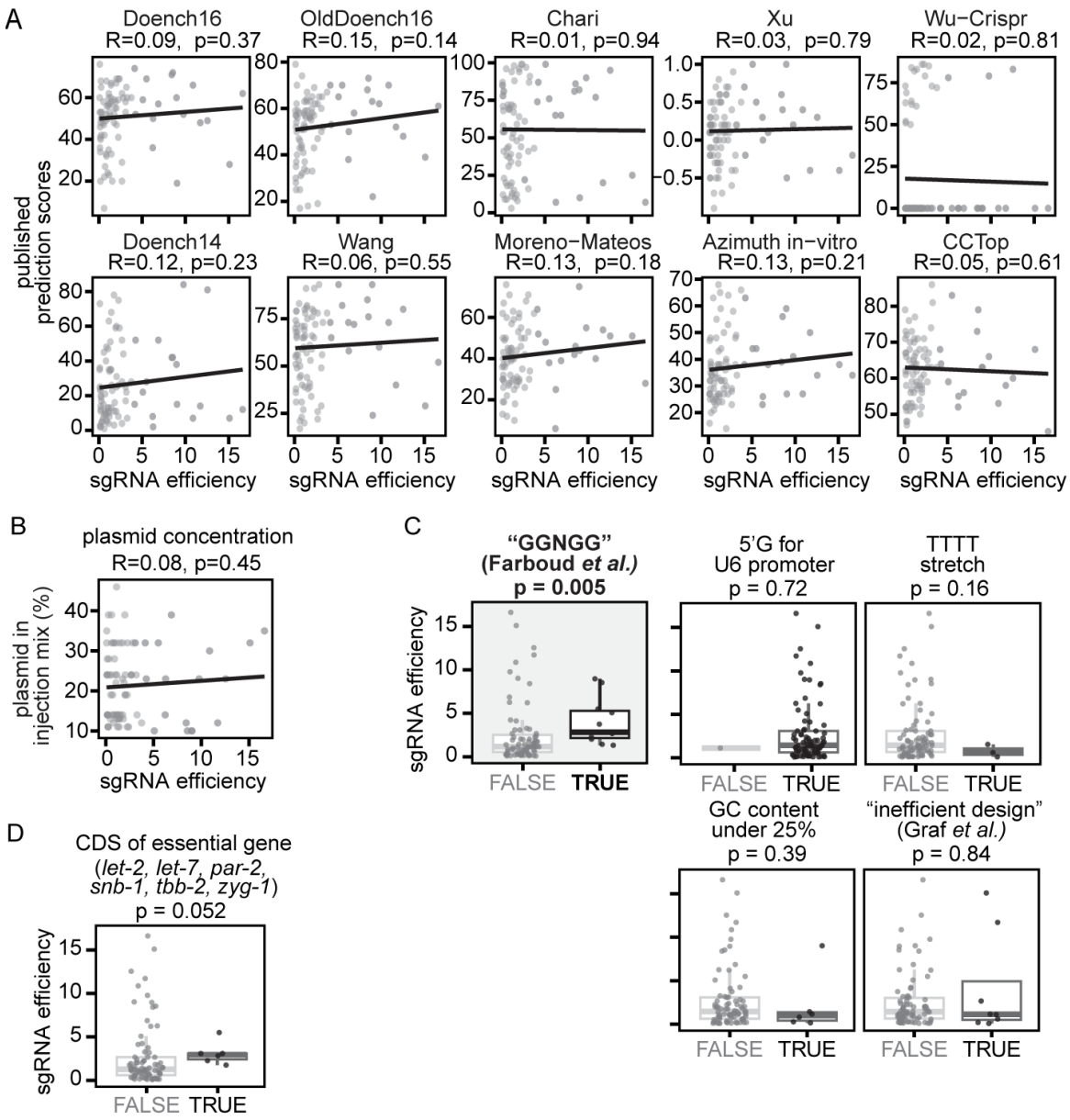


Figure 9. sgRNA efficiencies are independent of published prediction scores

(A) Correlation of various published sgRNA efficiency prediction scores and our observed sgRNA efficiency (n=91 sgRNAs). **(B)** Correlation of the percentage of plasmid in the original injection mix and the observed sgRNA efficiency. **(C)** Comparison of sgRNA efficiency for different sgRNA features. Categories were compared using the Wilcoxon signed-rank test. **(D)** 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.

We used the detected mutations to characterize CRISPR-Cas9-induced dsDNA-break repair outcomes in the *C. elegans* germline. On average, samples contained 57.9% deletions, 22.9% insertions and 19.3% complex events (combination of insertions, deletions, or substitutions) (**Figure 10A**). These proportions are similar for naturally occurring germline indels in *C. elegans* (75% deletions, 25% insertions)[99] and human (50% deletions, 35% insertions)[100].

The targeted sequencing approach resulted in a uniform read coverage per amplicon between 200,000 – 800,000-fold. We empirically determined general read thresholds to detect mutations robustly in treated samples while observing few mutations in the isogenic wild type controls. An insertion or deletion (indel) had to be supported by at least 0.001 % reads mapped to a position, at least 5 reads and overlap with a sgRNA cut site +/- 5bp. We excluded complex events (combinations of insertions, deletions, or substitutions) from the rest of our analyses to be more certain about the resulting sequences. 100 ng of genomic DNA was used as input for our sequencing protocol, representing more than 90 million genomes, enough to cover all animals in our samples. With the assumption that animals contributed equally to the extracted genomic DNA, we estimated that 4 – 10 mutants among 400,000 animals were sufficient to detect a mutation, depending on the amplicon coverage between 1,200,000 – 200,000 (**Table 1**).

Using these thresholds, we detected exactly 12,700 indels in our samples. We computationally separated deletions into those originating from a single sgRNA ("single-cut") or from two or more sgRNAs ("multi-cut") based on overlap with cut sites (**Figure 10B**). The length of single-cut deletions ranged from 1 to over 100 bp, with the majority being around 5 – 25 bp. Because larger deletions have a higher chance of overlapping with a second sgRNA cut site, this is likely an underestimation. Multi-cut deletions were larger, mostly several hundred bp, as expected from the spacing between multiplexed sgRNAs (**Figure 10B**). Most (>90%) insertions were 1 – 20 bp long although we could find insertions up to 45 bp (**Figure 10D**). These length distributions were similar to our observations previously made by Sanger sequencing (**Figure 4C**).

Inspection of individual genotypes revealed that most insertions contained short sequences also found close to the insertion position (**Figure 11A**). Using our deep sequencing data, we systematically analyzed such microhomologous matches between insertions and the surrounding regions. 5-mers from insertions matched to sequences in a window roughly +/- 13 bp around the insertion position and only in the same orientation (**Figure 11B–D**). Thus, our data indicate that many insertions are duplications of surrounding microhomologous sequences occurring mainly in the same orientation. Such templated insertions can likely be explained by activity of microhomology-mediated end joining (MMEJ), or its sub-process, theta-mediated end joining (TMEJ), that use 5–25 bp microhomology, and which has been reported as the main dsDNA-break repair pathway in *C. elegans*[47,101,102]. Independent of this, we saw very few 1 bp templated insertions, which in human cells originate from staggered Cas9 cutting[103]. This can likely be explained by the absence of a *C. elegans* homolog for the required Polymerase lambda[82,104].

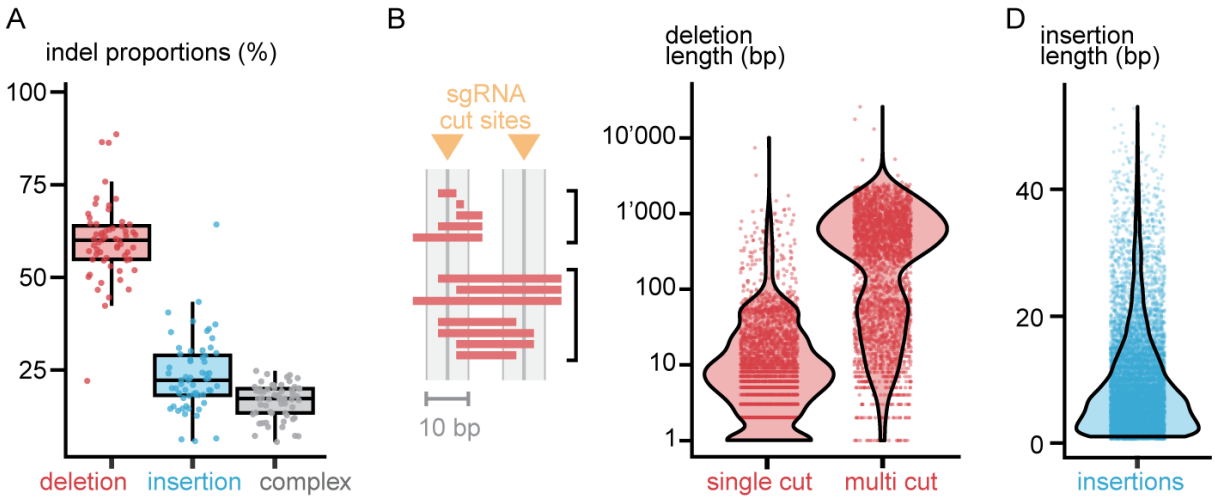


Figure 10. Proportion and length of different indels
Pooled data from 60 experiments, each sample expressing 1-8 sgRNAs targeting one region among 16 genes (n=24 wild type controls, n=36 samples with induced Cas9). **(A)** Proportions of reads with different types of mutations detected in each experiment (n=60 experiments). "Complex": reads with more than one insertion or deletion, or additional adjacent substitutions. **(B)** Length distribution of deletions found in all treated samples (n=2,915 multi cut, 3,169 single cut deletions). **(C)** Length distribution of insertions found in all experiments (n=6,616 insertions).

mean coverage /amplicon *(reads)*	animals /sample *(count)*	mean coverage /animal *(reads)*	"0.001 %" threshold = *(reads)*	"5 reads" threshold = *(%)*	more stringent threshold	minimum observations to call indel *(reads)*	minimum animals to call indel *(count)*
200,000	400,000	0.5	2	0.0025	5 reads	5	10
400,000	400,000	1	4	0.0013	5 reads	5	5
800,000	400,000	2	8	0.0006	0.001 %	8	4
1,200,000	400,000	3	12	0.0004	0.001 %	12	4
1,200,000	200,000	6	12	0.0004	0.001 %	12	2

Table 1. Estimated sensitivity of amplicon sequencing
(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.

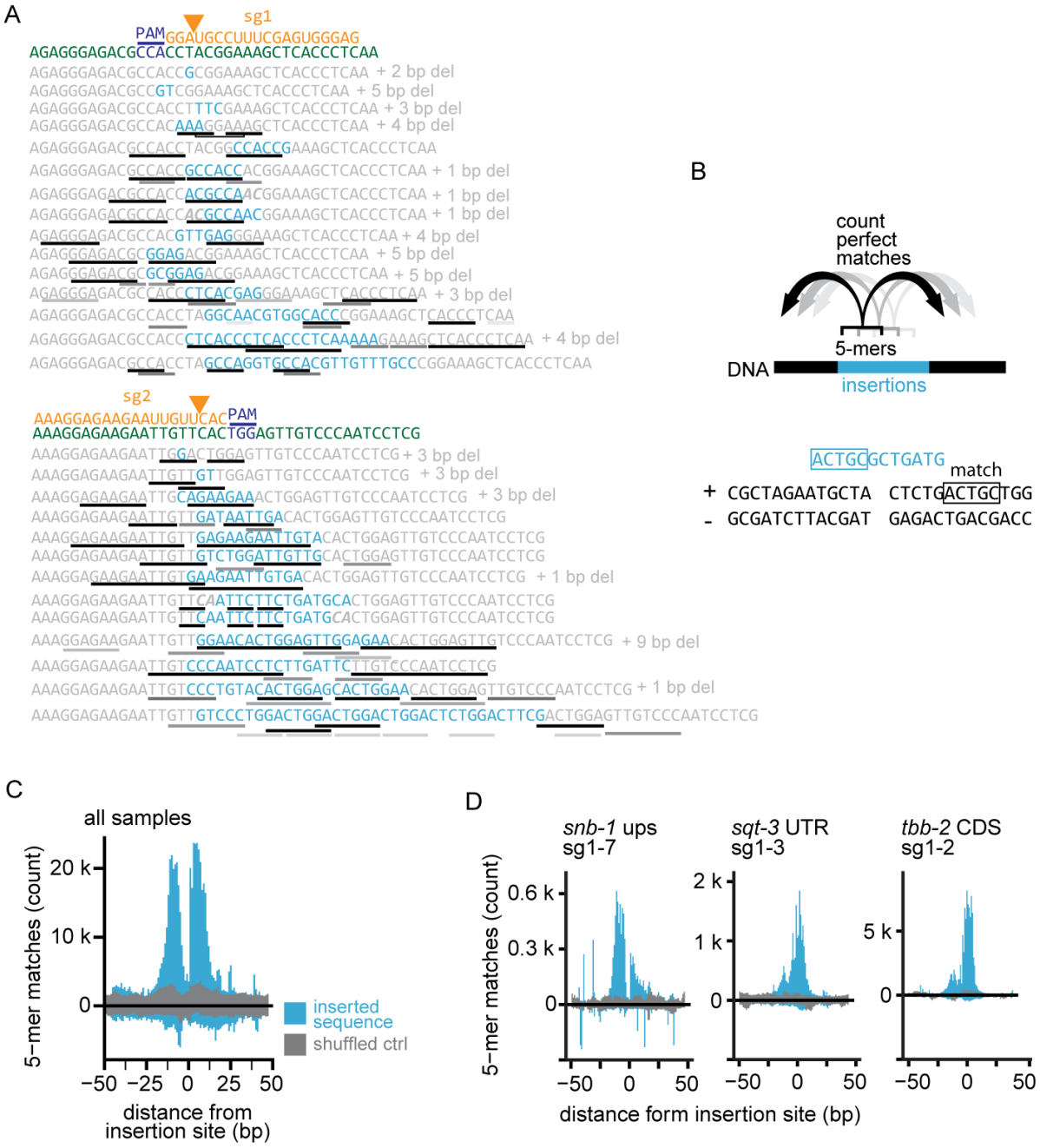


Figure 11. Insertions are templated from surrounding sequences

(A) Examples of microhomology observed between insertions and surrounding regions in genotypes of GFP-negative *his-72::GFP* animals. **(B)** Scheme showing the analysis approach which matches all possible 5-mers from an insertion to the surrounding sequence. **(C)** Matches of 5-mers from insertions (blue) to surrounding sequence (+/- 50 bp) in 34 samples. Randomly shuffled insertion sequences as controls (grey). **(D)** Same analysis as in C) for three different samples.

Finally, we assessed the genotype diversity generated by indels. We considered each unique deletion or each insertion-sequence a genotype, given that they reached the filtering thresholds defined before (0.001% reads, 5 reads, cut site overlap). We started by counting the number of unique deletions per base pair. We first studied deletions created by single-cut events for each sgRNA and found that highly active sgRNAs could generate up to 150 unique deletion genotypes and the highest diversity close to cut sites (rows in **Figure 12A**). Most of these genotypes defined by deletions covered a 10 – 12 bp region surrounding the cut sites. On average, every sgRNA could generate around 15 different genotypes per bp at the center of the cut site and up to 5 different genotypes per bp, 5 bp away from the cut site. On average 80% of nucleotides affected by deletions from a single sgRNA were +/- 5 bp within the cut site (black line profile in **Figure 12A**). We then studied multi-cut events. Here we found up to 200 unique deletion genotypes per base pair and on average around 20 per sgRNA covering a region more than 500 bp surrounding each cut site (**Figure 12B**). When counting the number of genotypes generated by one sgRNA, one sgRNA created 50 deletion- and 10 insertion genotypes on average. However, some sgRNAs created up to 400 genotypes (**Figure 12C**). Since we used several sgRNAs per transgenic line, we observed a median of 162 insertion- and 190 deletion genotypes per sample and in the most efficient lines 1833 deletion- and 1213 insertion genotypes (**Figure 12D**). More efficient sgRNAs resulted in a higher number of new genotypes (**Figure 12E**). Transgenic lines expressing more sgRNAs showed more unique deletion genotypes, possibly because of an increased chance of containing efficient sgRNAs and the combined activity of multiple sgRNAs creating combinatorial deletions (**Figure 12F**).

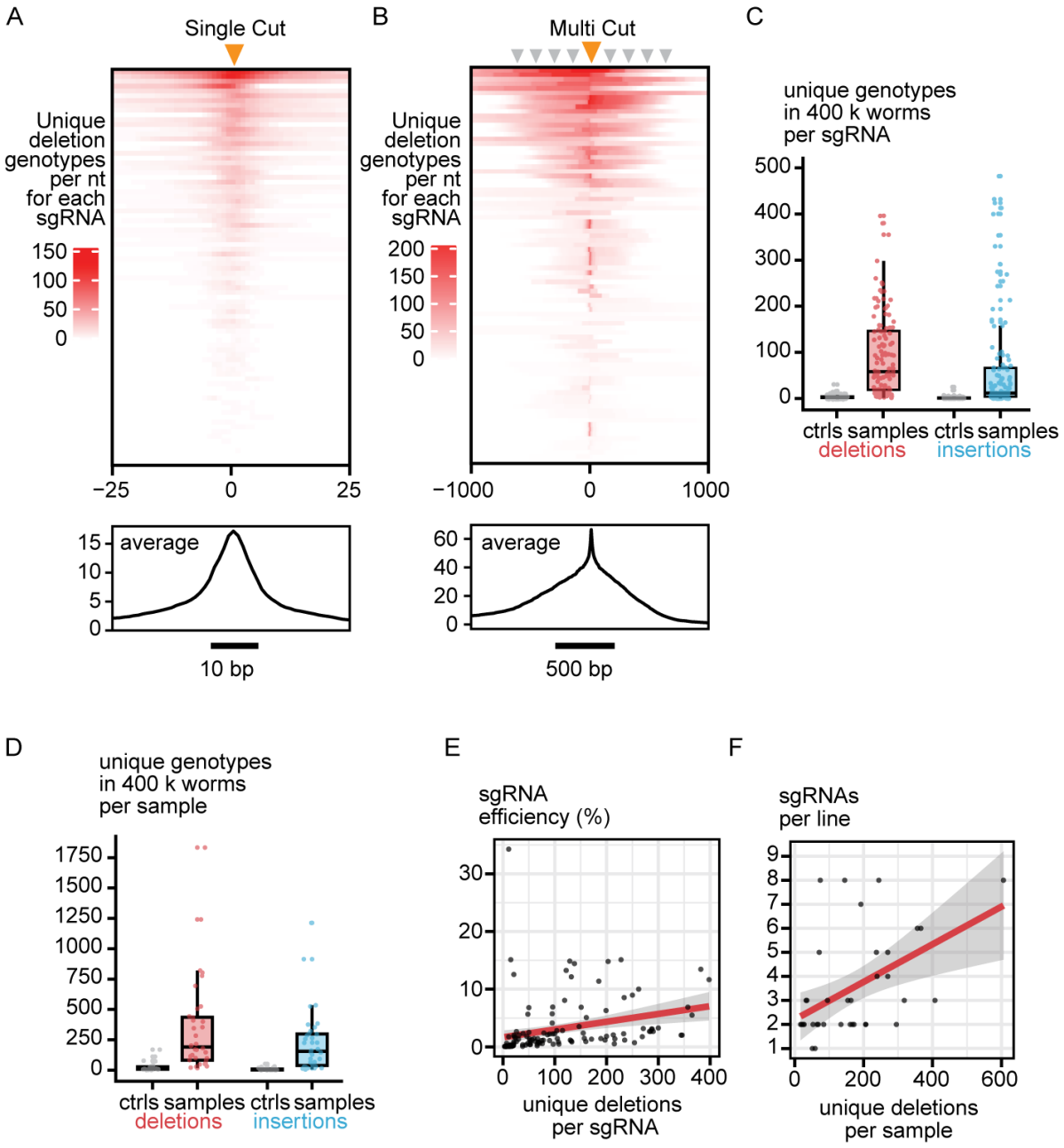


Figure 12. Indels create diverse genotypes along the targeted regions

Pooled data from 60 experiments, each sample expressing 1-8 sgRNAs targeting one region among 16 genes (n=24 wild type controls, n=36 samples with induced Cas9). **(A and B)** Unique deletion genotypes per nucleotide for each sgRNA centered at cut sites. Each row shows the count of distinct genotypes per nucleotide for one sgRNA (n=86 sgRNAs); black curve on the bottom: average unique deletion genotypes per bp. **(C)** Unique genotypes detected per sgRNA in 400,000 sequenced worms (n=76 ctrls cut sites, n=86 samples cut sites) (Wilcoxon, $p < 2.2e{-}16$ for deletions, $p < 2.2e{-}16$ for insertions). **(D)** Unique genotypes created per sample by deletions or insertions (n=24 ctrls, n=36 samples) (Wilcoxon, $p = 1.7e{-}08$ for deletions, $p = 4.7e{-}09$ for insertions). **(E)** Correlation between sgRNA efficiency and the created unique deletions per sgRNA per sample (n=91 sgRNAs). **(F)** Correlation between the amount of different sgRNAs in a transgenic line and the created unique deletions per sample (n=6,084 unique deletions, n=36 treated samples).

Regulation of *lin-41* expression and phenotype by *let-7* miRNA binding sites

A major challenge to the understanding of gene regulation is the interaction of different regulatory elements. Especially in 3′ UTRs, which can act on all levels of gene expression, this can be difficult. To simultaneously measure mRNA levels for all generated 3′ UTR deletions within large *C. elegans* populations, we developed a targeted RNA sequencing strategy. As a proof of principle, we tested it on a microRNA-regulated mRNA. The *lin-41* mRNA is regulated by *let-7* microRNAs which bind two complementary sites in the 1.1 kb long 3′ UTR (site1 and site2, 22 and 20 nucleotides long, separated by a 27 nt spacer)[105–109] (**Figure 13A**). Although studies with reporter plasmids showed that each binding site could not function on its own[109], other studies concluded that each site could recapitulate wild-type regulation when present in three copies[110]. We wanted to explore the function and interaction of the two binding sites in the native sequence context and at natural expression levels. Therefore, we targeted the *lin-41* 3′ UTR with a pool of 8 sgRNAs or, individually, two different pairs of sgRNAs close to the *let-7* binding sites (**Figure 13B**). We then sequenced *lin-41*-specific cDNA with long reads to cover the complete 3′ UTR (**Figure 13C**). Each read contained full information on any deletion in the RNA molecule, while the number of reads supporting each deletion could be used to estimate RNA expression level. *Lin-41* down-regulation occurs with *let-7* expression in the developmental stages L3-L4[105,107,108,111]. To measure *let-7*-dependent regulation, we collected RNA from mutated F2 generation bulk worms at L1 and L4 stages. We extracted L4 stage RNA after complete *lin-41* mRNA downregulation by *let-7*[112] and before the occurrence of the lethal vulva bursting phenotype[106] (**Figure 13D**). To determine *let-7*-dependent effects, we then analyzed how different deletions affected RNA abundance at L4-, relative to L1 stage.

We observed an average of more than 4-fold up-regulation of *lin-41* mRNA at larvae stage L4 when both *let-7* miRNA seed sites were affected by deletions (**Figure 13E**). A 4-fold regulatory effect is consistent with the known magnitude of down-regulation in the natural context[105,108,109] or the up-regulation when disrupting both *let-7* interactions (2 – 4-fold)[106,113,114]. A weak but significant up-regulation was observed for deletions overlapping with the site1 seed. We obtained fewer deletions for the site2 seed and therefore did not have the statistical power to rule out a similar weak impact.

As an independent approach and to measure the effect of genotypes with multiple deletions per animal, we used unsupervised clustering of long cDNA reads using the k-mer content of reads to obtain clusters representing similar genotypes. These data also suggest that RNA molecules transcribed from genotypes with deletions overlapping both sites were detected with more reads in L4 stage compared to L1 stage animals (see cluster 1-4, 7-8, 11-13 in **Figure 14A–C**). Additionally, this analysis revealed two other areas that affected mRNA in the opposite way by either increasing levels at L1- or decreasing levels at L4 stages, which could be further investigated in the future (clusters 5 and 10 in **Figure 14C**).

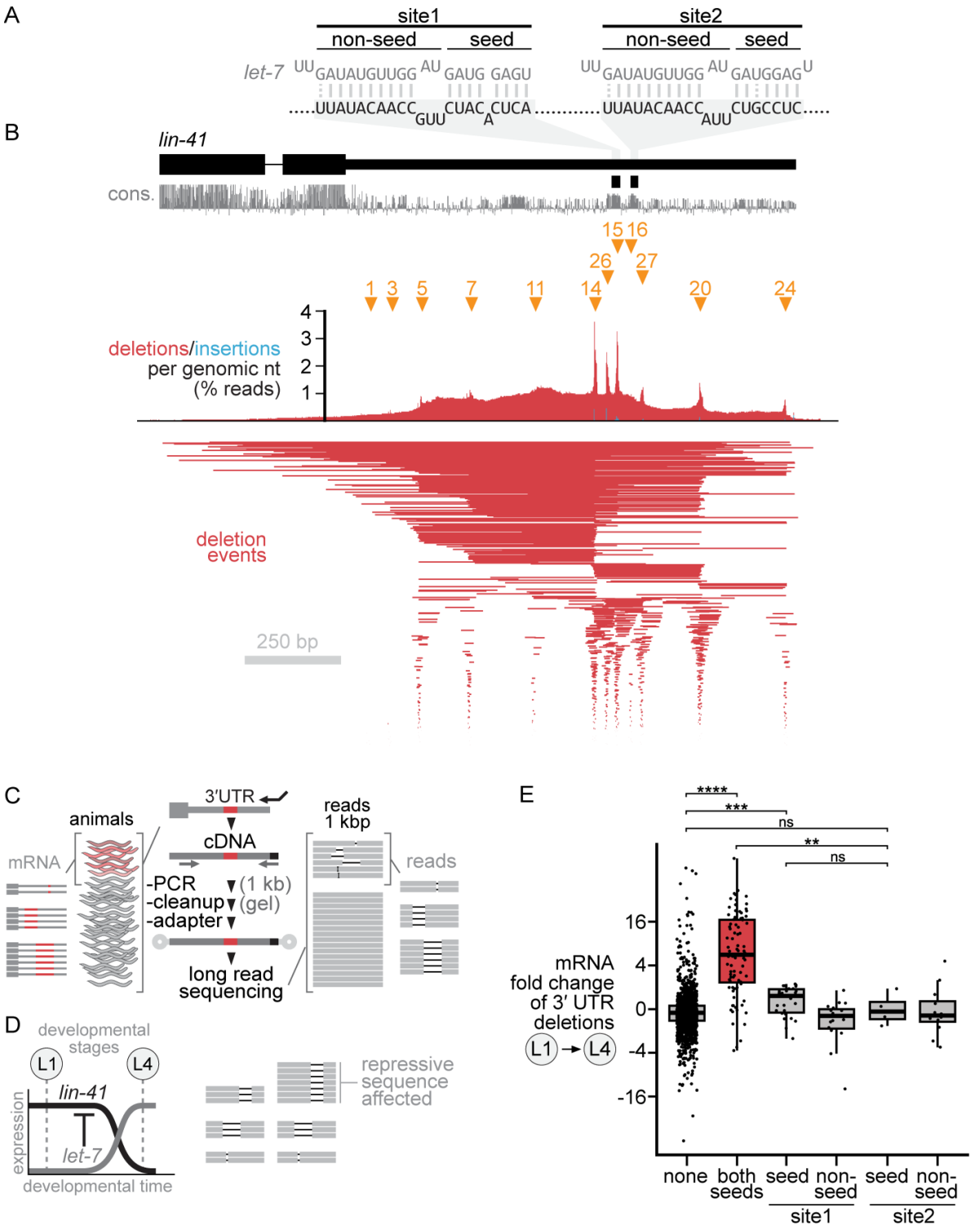


Figure 13. Each *let-7* site alone maintains near-wild-type repression of *lin-41*

(A) Diagram showing the two *let-7* complementary sites in the *lin-41* 3′ UTR. **(B)** The *lin-41* 3′ UTR locus after targeted mutagenesis with three different lines (sg pool, sg15+sg16, sg26+sg27, sgRNA cut sites indicated by orange triangles). Deletions of three lines were pooled and analyzed together (n>900 deletion events). **(C)** Diagram of the targeted RNA sequencing strategy. cDNA was amplified using a large amplicon and sequenced using the Pacbio long read workflow. **(D)** Diagram of *lin-41* and *let-7* developmental expression and time points of RNA extraction. **(E)** Relative fold change of deletions detected in targeted full-length sequencing of cDNA between L1 and L4 developmental stages. Deletions are classified by their unique overlap with regions of interest. "Seed" and "non-seed" as depicted in A). (Wilcoxon rank-sum test, ns $p > 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ and $^{****}p < 0.0001$).

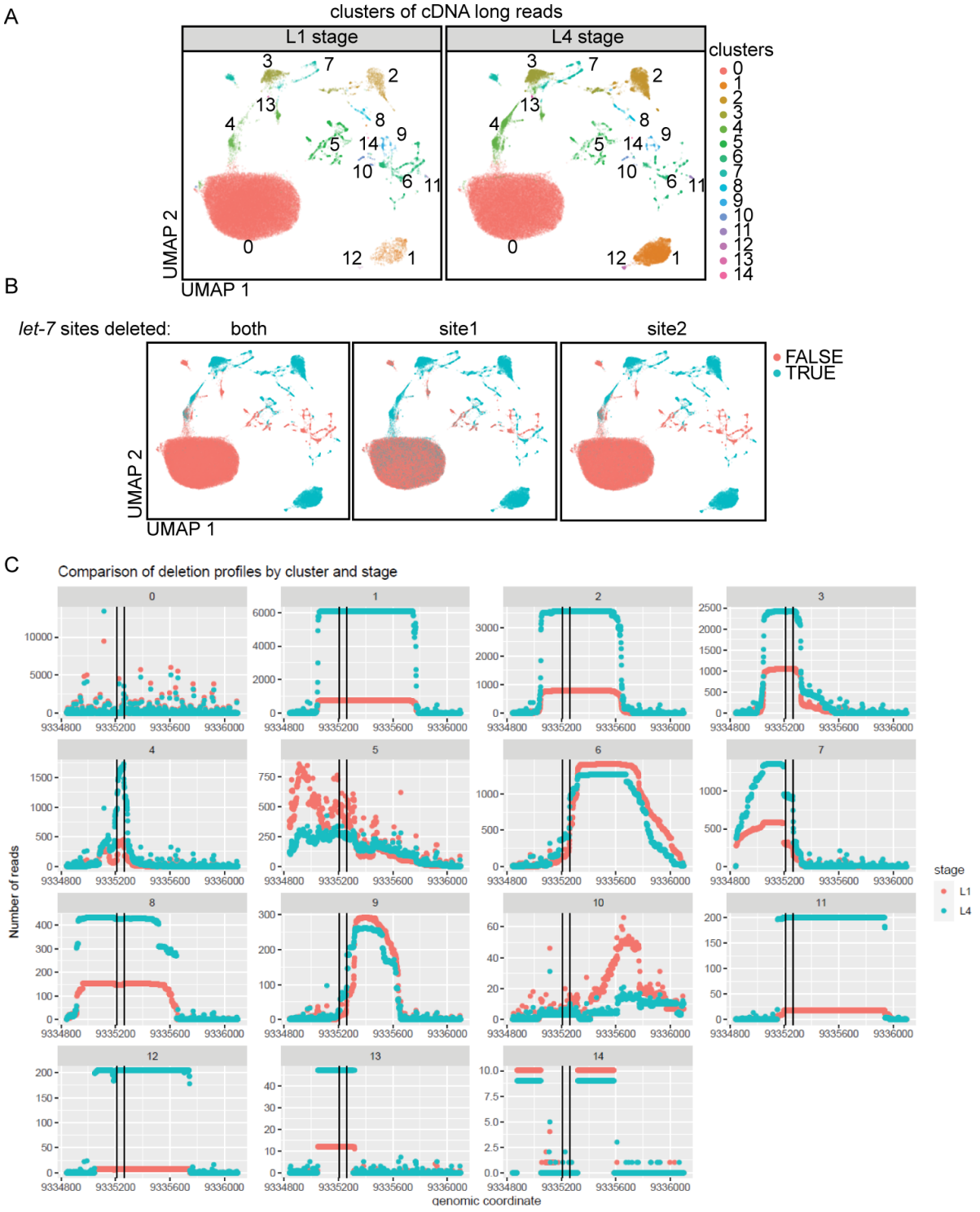


Figure 14. Genotypes that delete both *let-7* sites upregulate *lin-41*

(A) 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. **(B)** Status of overlap with *let-7* sites for each read. **(C)** 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). Note that *lin-41* lays on the minus-strand and the transcript 3′-end is left on these plots.

Disrupting *let-7* regulation of *lin-41* mRNA is known to result in lethal developmental defects[106–108,113,115]. To assign fitness to individual mutations in a controlled environment, we established measurements on genotype abundance over several generations. For this, we sampled genomic DNA of consecutive generations (**Figure 15A–B**). We performed this analysis starting at the F1 generation because also mosaic animals would be expected to show a phenotype with a fitness disadvantage. Deletions in the *lin-41* 3′ UTR which overlapped both seeds quickly disappeared from the population already after one generation. Consistent with the effect on RNA expression, deletions of both seeds were strongly depleted, while deletions affecting either one of the two sites alone were depleted only slightly compared to control deletions not overlapping with any features ("none") (**Figure 15C–D**). This also indicated that deletions with stronger effects were possibly already missing in the mRNA analysis that we performed in the F2 generation.

While deletion of both *let-7* binding sites is reported to be lethal[106], our results showed that deletions of one site could be tolerated. To validate these findings, we created two seed-disrupting deletions for each site (**Figure 16A**). We compared *lin-41* mRNA expression and phenotypes of homozygous mutants with wild type animals. To disrupt both *let-7* interactions simultaneously, we used the temperature sensitive *let-7(n2853)* allele[107]. At 50 hours into development adult animals with mutations in site2 displayed a normal wild type phenotype, while site1 mutants were visibly sick, but were still laying eggs. *Let-7(n2853)* mutants were not alive anymore (**Figure 16B**). We quantified the lethal vulval bursting phenotype and found that although 98% of *let-7* mutants were dead or had burst, only 3% of site1 and none of site2 mutants showed this phenotype (**Figure 16C**). At the L4 developmental stage, *lin-41* mRNA was strongly up-regulated in *let-7(n2853)* (8-fold), slightly in site1 mutants (3-fold), and very little in site2 mutants (1.5-fold) (**Figure 16D**). This could indicate that our high-throughput bulk mRNA measurements, which showed less strong effects, were biased towards deletions with smaller effects, possibly due to the dropout of animals after the F1 generation.

Because inactivation of *let-7* binding site2 seemed to be tolerated, we hypothesized that an equal level of repression could be explained by higher levels of *let-7* acting on the remaining site1 in compensation. Indeed, we found that site2 mutants (little *lin-41* upregulation and a normal phenotype), displayed 3-4-fold higher levels of *let-7*, while site1 mutants (stronger *lin-41* upregulation and a slightly sick phenotype), showed *let-7* levels like wild type animals (**Figure 16E**).

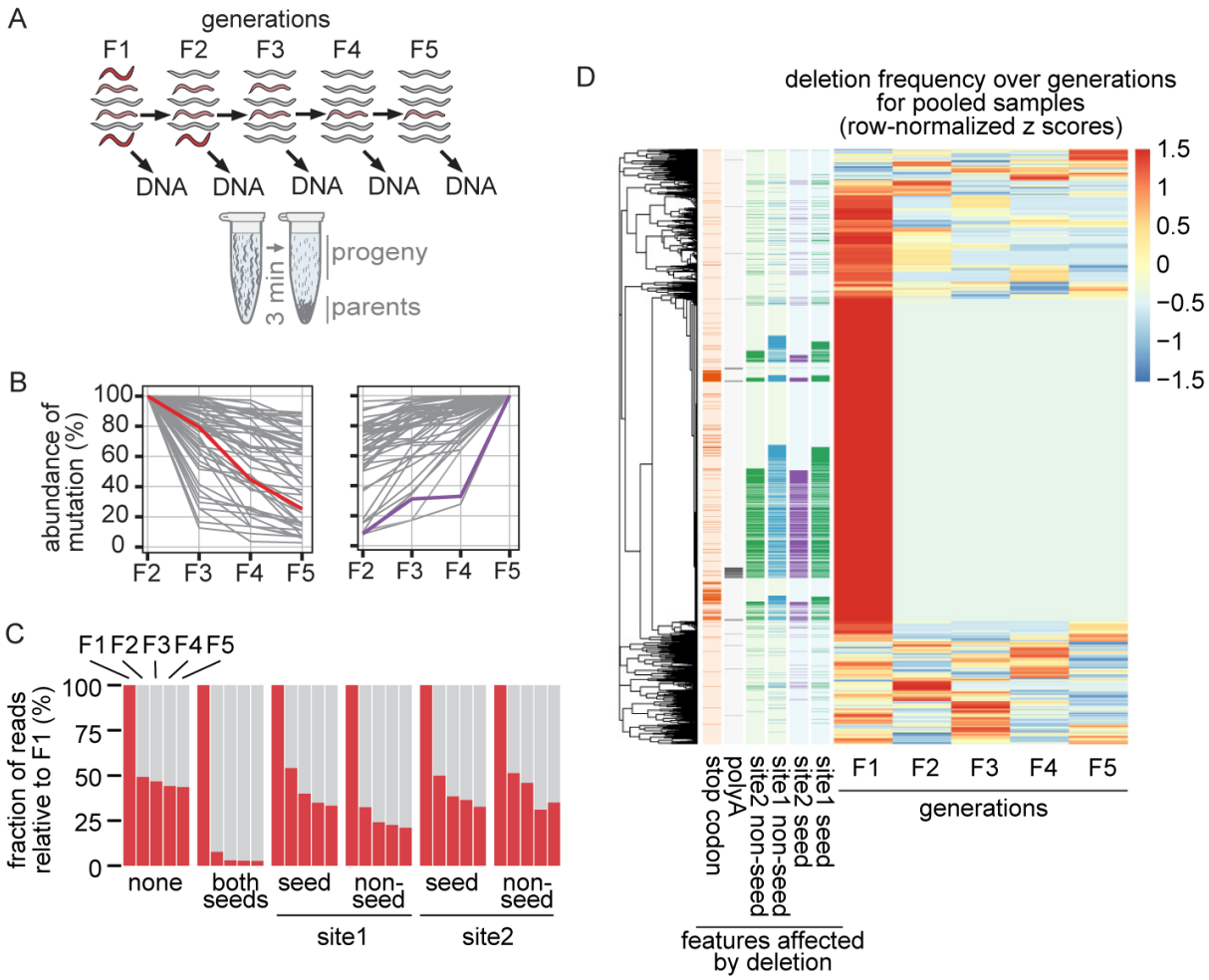


Figure 15. Each *let-7* site alone maintains near-wild-type generational fitness
(A) Experimental outline. **(B)** Example of mutations that decrease or increase in relative abundance over several generations. **(C)** Fraction of reads supporting deletions in bulk genomic DNA of consecutive generations, relative to the first (F1) generation. Deletions from six samples were pooled for this analysis (sg pool, sg15+sg16, sg26+sg27 grown at 16°C and 24°C). **(D)** 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).

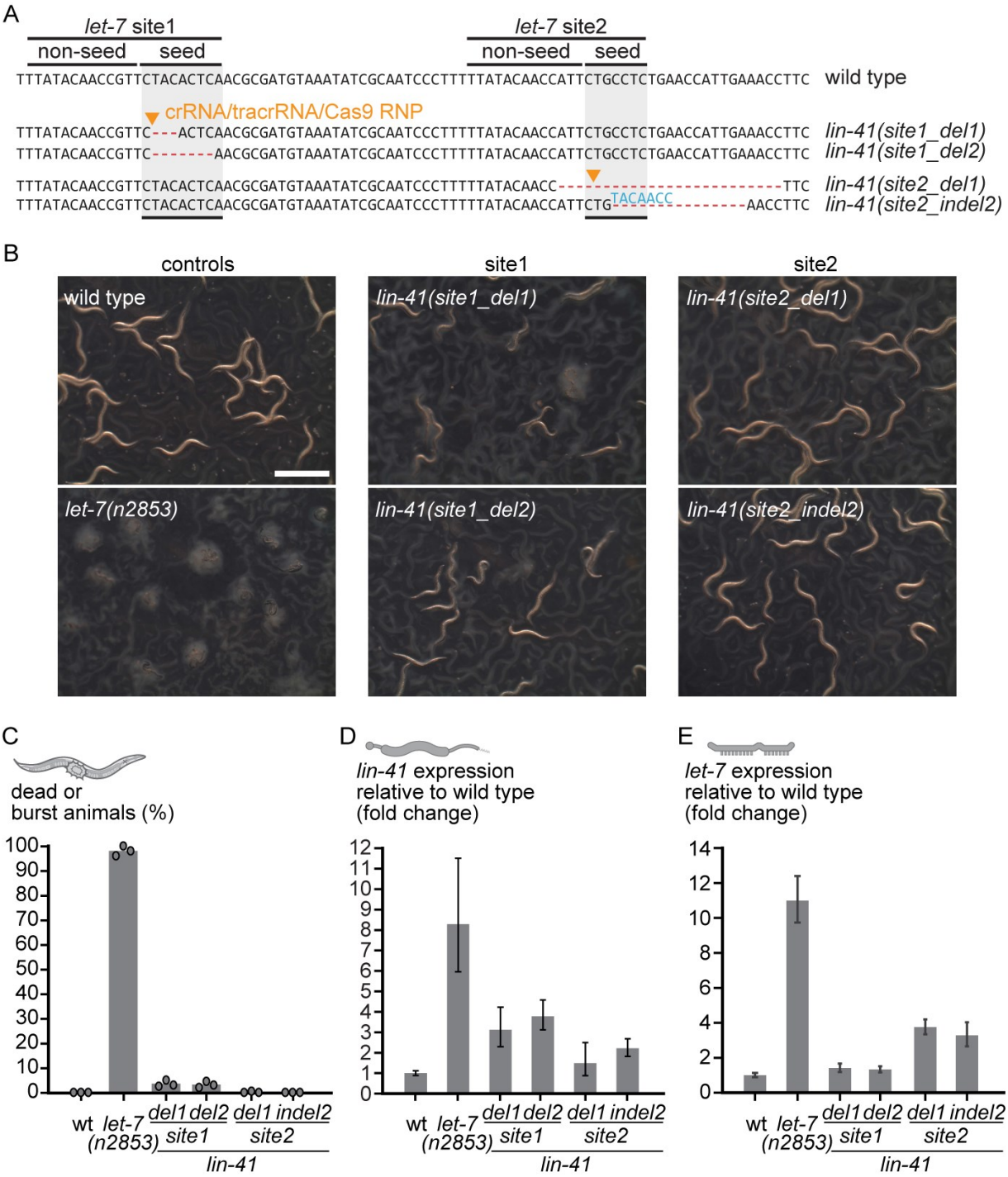


Figure 16. Single *let-7* site mutants are viable and show slight *lin-41* upregulation.
(A) Genotypes of strains with deletions in *let-7* complementary site1 and site2 in the *lin-41* 3′ UTR. **(B)** Phenotype of *lin-41* site1 and site2 mutant strains compared to wild type and *let-7(n2853)*, 50 hours into synchronized development at 24°C. Scale: 1 mm. **(C)** Dead or burst animals at 50 hours into synchronized development at 24°C from three plates (n=3) and scoring 200 animals. **(D)** *lin-41* mRNA levels in the *let-7* mutant allele *let-7(n2853)* and in *lin-41* strains with deletions affecting site1 or site2 relative to wild type levels, quantified by qPCR. One experiment with 7000 animals, 30 hours into synchronized development at 24°C. Bars represent mean with error bars +/- standard deviation. **(E)** *let-7* miRNA levels, relative to wild type, quantified by Taqman assay, same experiment as in D).

Gene regulatory mutations that change morphological phenotype

Next, we wanted to directly map regulatory sequence variants to phenotypic traits. This could be useful to discover functional elements, provide starting points to study regulatory mechanisms, and to explore phenotypic plasticity in animals. Such an approach would also capture any functional sequences regardless of the type, time, or place of regulation. We targeted a predicted enhancer[116], three promoters, and all 3′ UTRs of 8 genes and manually screened 35,000 animals for each of these regions. Loss-of-function or reduction-of-function of the screened genes are known to result in strong organismal defects in animal movement and body shape (Unc, Slu, Rol, Dpy). We proceeded to select worms based on these phenotypes and to identify the causative mutations (**Figure 17A**). Although we screened for all defects in movement and body shape, our approach was therefore biased towards finding reduction- and loss-of-function mutations. To determine which mutations were initially present in the screened population, we performed targeted sequencing on siblings. Initially, we isolated several mutants with large deletions (>500 bp) that disrupted the coding sequence or the polyadenylation signal (AATAAA) (**Figure 17B, C**). Similar large-scale, on-target deletions have also been described in cell lines and mice[51,117,118]. We also found large insertions (up to 250 bp) which originated from within +/-1 kb of the targeted region, or from loci on other chromosomes (**Figure 17B, C**). We found such large deletions or insertions in 5 out of 8 screened genes, demonstrating that for these genes our screen was sensitive enough to detect animals with affected phenotypes (**Table 2**).

From the screen we isolated 57 alleles for 3 genes (*egl-30*, *sqt-2*, *sqt-3*) and none for the other 5 genes (*dpy-2, dpy-10, rol-6, unc-26, unc-54*) (**Table 2**). All alleles showed phenotypic defects previously described for a reduction-of-function of the affected genes. Deletions, insertions, and complex mutations (combination of insertions and deletions) were represented equally among isolates (**Figure 17D**). The observed phenotypic traits showed complete penetrance and we scored their expressivity which differed between mutations. We found that several mutations in the 3′ UTR of *egl-30* resulted in the Sluggish (Slu) phenotype which is characterized by slow movement. In 7/11 mutants, a region around 100 bp downstream of the STOP codon was affected and the smallest deletion was 6 bp (**Figure 17E**). We found mutations overlapping a putative *sqt-2* enhancer predicted from chromatin accessibility profiling[116] with a Roller (Rol) phenotype where animals rotate around their body axis and move in circles (**Figure 17F**). This was the only region for which penetrance varied between different mutations. We also targeted *sqt-3*, a gene associated with three distinct morphological traits (Dpy, Rol and Lon)[119,120]. 13 Rol mutations upstream of *sqt-3* likely affected transcriptional initiation, with 11/13 overlapping the predicted TATA-box (**Figure 17G**). In line with the Rol phenotype, which indicates a reduction-of-function, we later showed that pre-mRNA and mRNA levels were both reduced to around half in one TATA-box-deficient mutant (next paragraph **Figure 18C**). This suggests that *sqt-3* transcription is only partially dependent on the TATA-box.

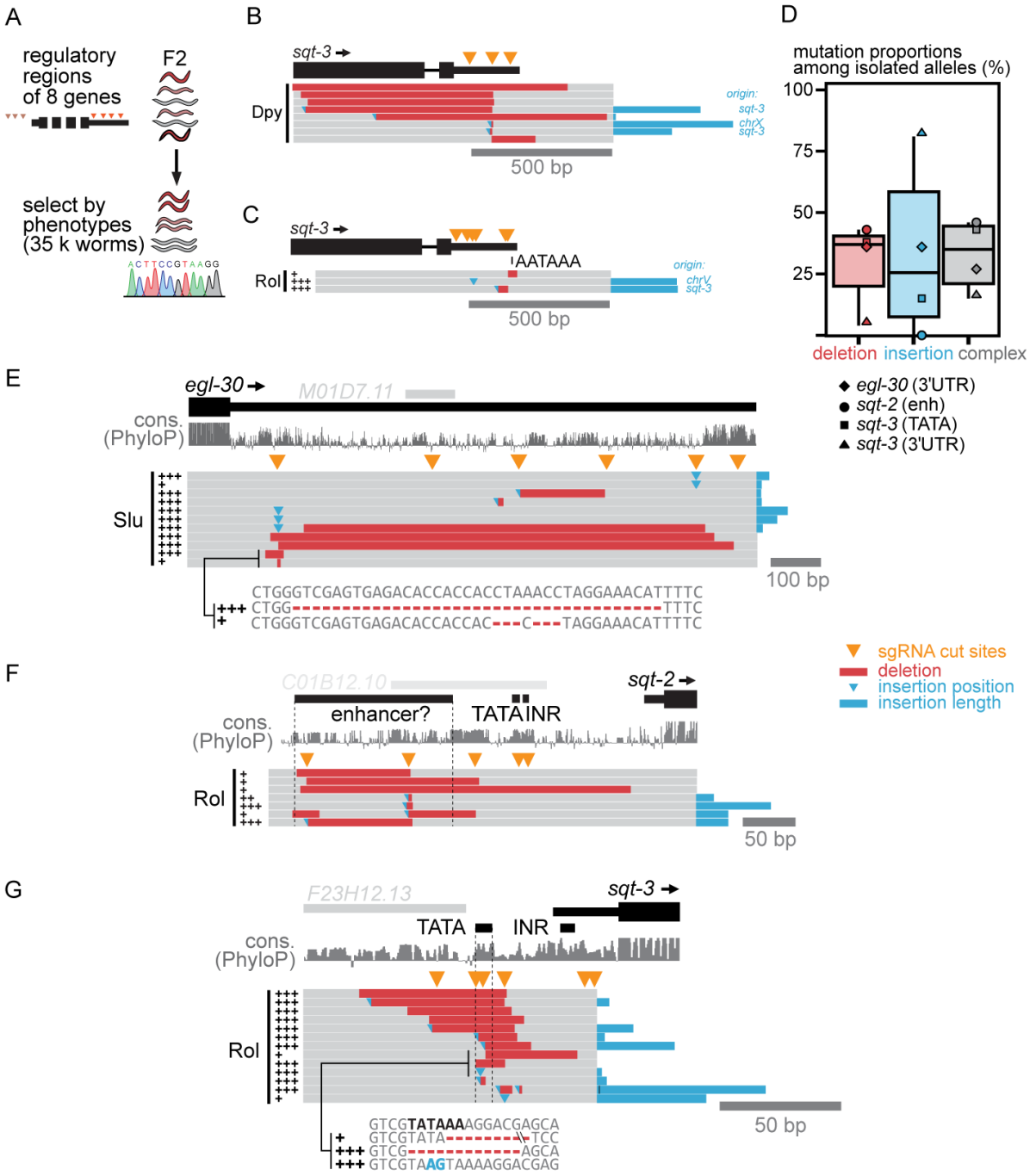


Figure 17. 57 gene regulatory mutations in 3 genes that change phenotype

Shown are genotypes of strains which were isolated according to phenotypic traits after targeting regulatory regions. Phenotypes showed complete penetrance (n>300 animals) and expressivity was scored as indicated by +, ++, or +++ (n>300 animals). **(A)** Outline of the screen. 8 genes were targeted by pools of 2-6 sgRNAs in different regulatory regions (some enhancer, promoter, all 3′ UTR) resulting in 21 samples. 35,000 F2 animals were screened manually for morphological traits. **(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)** Eleven mutations along the *egl-30* 3′ UTR which show slight or strong Sluggish (Slu) phenotypes. No canonical polyadenylation signal present. **(F)** 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%). **(G)** Thirteen mutations upstream of *sqt-3* which show a Roller (Rol) phenotype.

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

Table 2. Isolated mutants
"dels into coding" indicates whether mutants were found that carried deletions into the coding sequence, an indication that our screen was sensitive enough to detect mutants for these genes.

The 26 other isolated *sqt-3* alleles were 3′ UTR mutations. Almost all (25/26) were insertions or insertions combined with deletions, originating at sg2 (**Figure 18A**). The only pure deletion (that did not contain any insertions) overlapped with a canonical polyadenylation signal (AATAAA). We knew from amplicon sequencing of siblings that sg2 was very efficient (~25%) and that various deletions covering the complete 3′ UTR were present in the screened samples. We therefore used direct PCR screening to isolate non-Rol mutants. 24/96 (25%) genotyped animals contained mutations, thus also confirming the estimation from amplicon sequencing (~25%). Despite containing 13 distinct deletions or insertions originating at the efficient sg2, these animals showed the wild type non-Rol trait (**Figure 18B**). We did follow-up experiments with one of the 25 insertion alleles, *sqt-3(ins)*, and determined that mRNA levels were reduced post-transcriptionally to around 50% (**Figure 18C, D**). Since deletions and some insertions in this region were well tolerated (non-Rol), we concluded that the isolated Rol mutations likely resulted from a gain of repressive sequence which led to the observed reduction of mRNA. The polyA mutant *sqt-3(polyA)*, for which mRNA levels were equally reduced to 50%, showed a weaker Rol phenotype, with only slight bending of the head (**Figure 18D, E**). This suggests that additional mechanisms besides mRNA down-regulation might further reduce protein output in *sqt-3(ins)*.

To define the repressive sequence elements, we targeted the inserted sequence with several sgRNAs and screened for revertants, in which the wild type non-Rol trait was restored by intragenic suppressor mutations. 12/13 revertants contained deletions overlapping with the insertion, with the smallest being 5 bp (**Figure 19A**). A restored wild type trait likely resulted from restored expression levels. Indeed, mRNA levels in two independent revertants were restored to normal (**Figure 19B**).

To discover other genetically interacting sequences, we had included sgRNAs for the remaining 3′ UTR. This revealed a compensatory deletion upstream of the insertion, which was able to revert the Rol phenotype. We isolated two more additional revertants after using sgRNAs specific for this region (**Figure 19A, C**). Surprisingly, mRNA levels were not restored ("revertant3", **Figure 19B**). This points to an alternative mechanism of restored gene function, for example on translational level, or affecting mRNA at a different developmental time point.

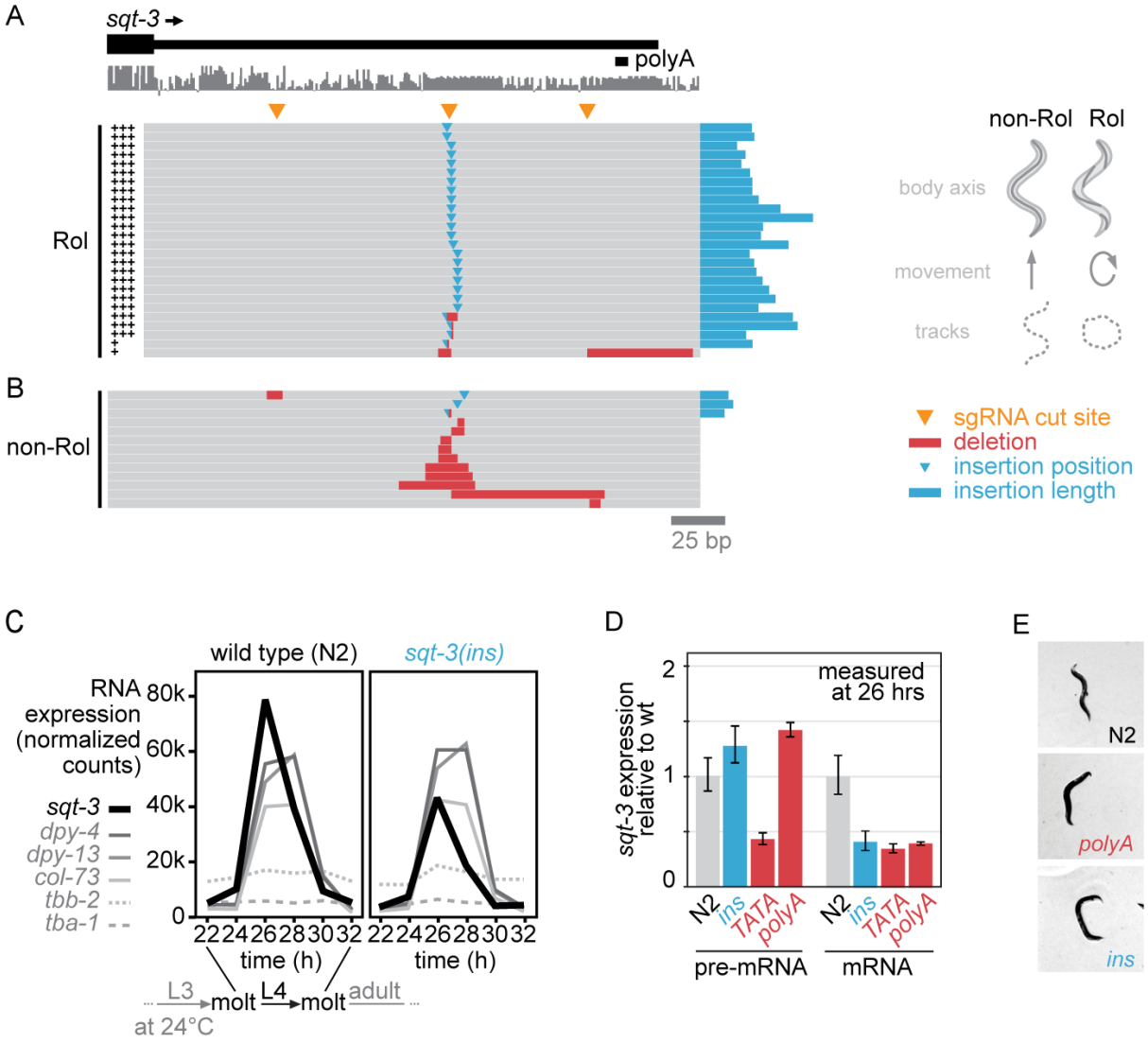


Figure 18. 25 semi-random insertions in the 3′UTR repress *sqt-3*

(A) Mutations in the *sqt-3* 3′ UTR which show a Rol phenotype. "polyA": canonical polyadenylation signal AATAAA. **(B)** Mutations which were tolerated (non-Rol). **(C)** 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. **(D)** *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. **(E)** 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.

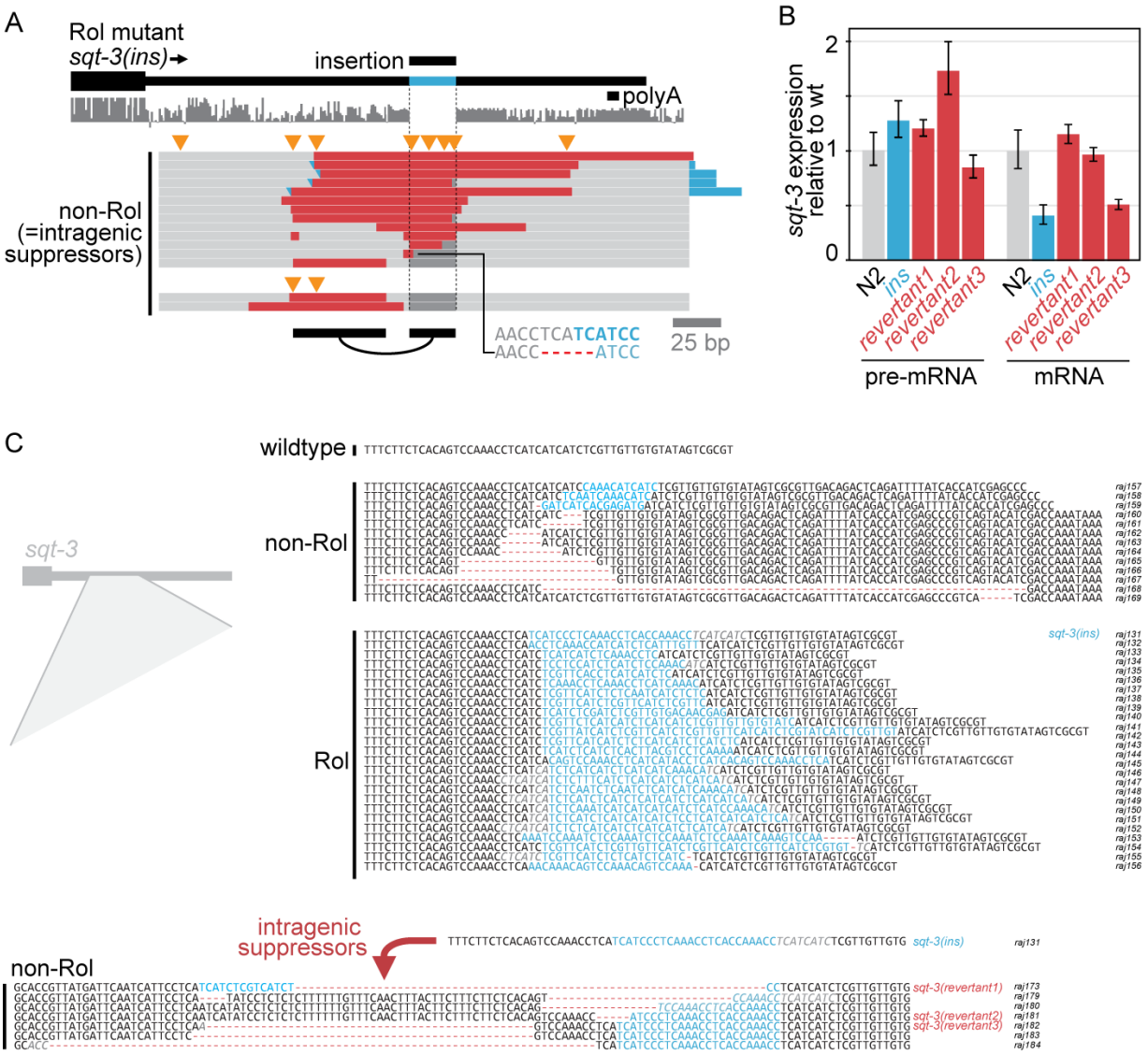


Figure 19. Intragenic suppressor deletions revert the RoI phenotype of *sqt-3(ins)*

(A) Fifteen mutations, mostly deletions, which suppressed the RoI phenotype of one insertion allele *sqt-3(ins)*. Black bars on the bottom: uncovered compensatory interaction by intragenic suppressor mutations. **(B)** 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. **(C)** Nucleotide sequences of relevant 3′ UTR regions in RoI, non-RoI and revertant mutants showing the inserted and deleted nucleotides.

Even though the different Rol insertions were not present in nature, we wanted to understand how insertions could mechanistically reduce *sqt-3* expression post-transcriptionally. We determined that *sqt-3(ins)* was a recessive mutation (Supplemental Figure 5). Overall, predicted RNA secondary structures did not change between the *sqt-3(ins)* and *sqt-3(revertant2)* alleles, suggesting other factors than mRNA structure and accessibility (**Figure 20A**). Newly created splicing acceptors could lead to skipping of the coding sequence of the last exon, with subsequent mRNA decay and reduction of protein function. However, we could not find any additional splice isoforms in three tested alleles (**Figure 20B**). We also could not detect increased small RNAs along the mutant *sqt-3(ins)* allele, that would indicate a siRNA or piRNA -dependent mechanism (**Figure 20C**). We performed *in vivo* targeted mRNA pull-down ("viPR", ref.[121]) and identified mRNA-bound microRNAs and proteins, but could not find any significant differences in factors binding to wild type and mutant mRNA (**Figure 20D–G**). We performed sequence transplantations into the 3′ UTR of *dpy-10* and *unc-22*, of which *unc-22* showed the characteristic reduction-of-function Twitcher phenotype (**Figure 20H**). This indicates that the repressive sequence might also function in other sequence contexts, but more experiments would be needed to test this thoroughly. Also, because *unc-22* is a neuronally expressed gene, the mechanism is not specific to hypodermis, the main tissue of *sqt-3* expression[82]. ♦

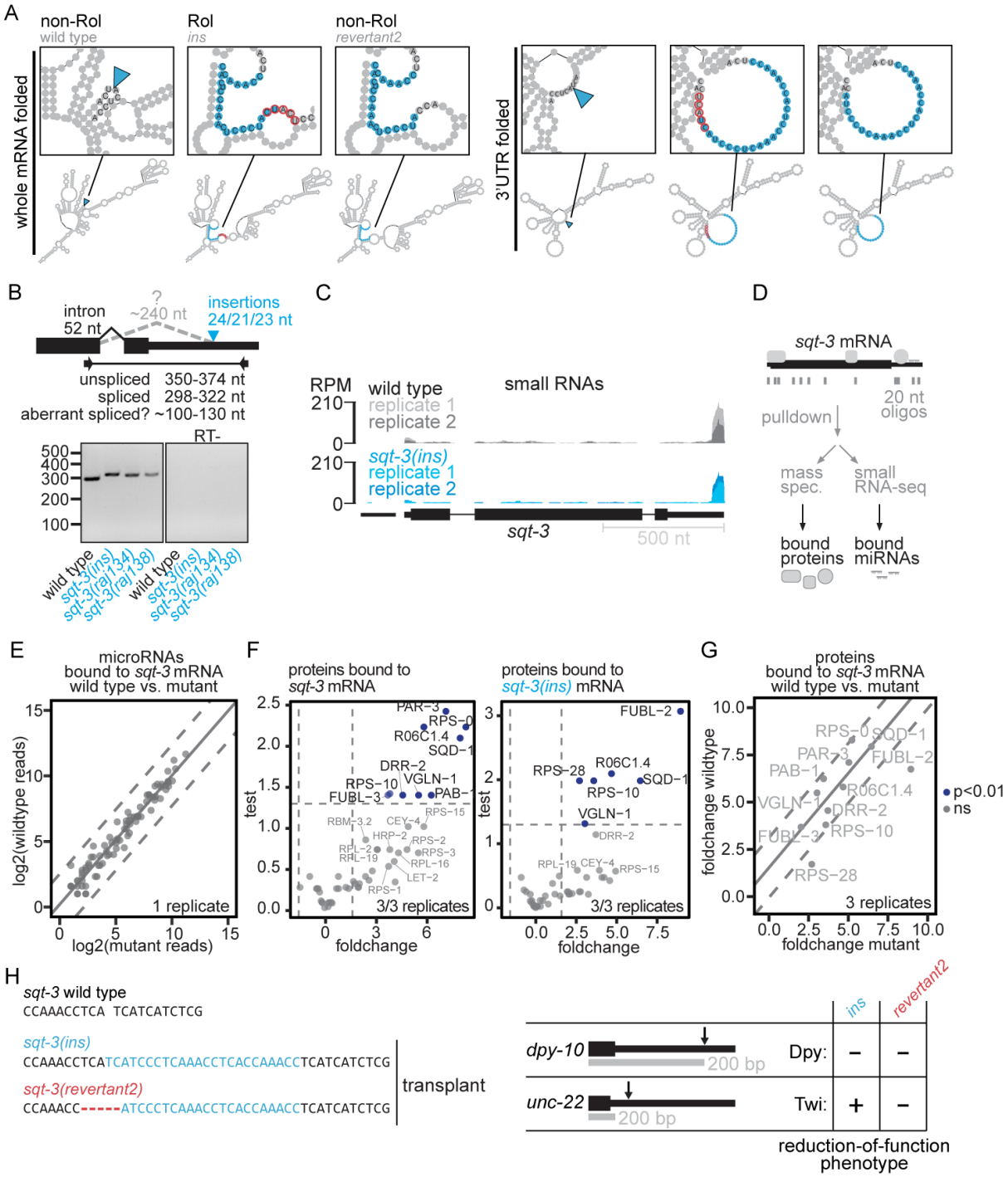


Figure 20. Molecular biology of *sqt-3(ins)* mRNA

(**A**) Predicted RNA secondary structures of wild type, insertion mutant and revertant allele. Predictions were made for the whole mRNA or only the 3′ UTR using RNAfold[122]. **(B)** Scheme of primers, size of expected DNA bands and the observed gel for a PCR to test for aberrant splicing. **(C)** Small RNA reads mapping to the *sqt-3* locus in wild type and *sqt-3(ins)* mutant. **(D)** Scheme of *sqt-3* mRNA pulldown to measure mRNA-bound proteins and miRNAs ("viPR", ref.[121]). **(E)** mRNA-bound microRNAs in wild type compared to *sqt-3(ins)* mutants. **(F)** mRNA-bound proteins in wild type or *sqt-3(ins)* mutants. **(G)** comparison of mRNA-bound proteins in wild type or *sqt-3(ins)* mutants. For all pulldown experiments: significantly (p<0.01) enriched proteins or miRNAs in blue. P-values were determined with a moderated t-test and corrected for multiple comparisons by the Benjamini-Hochberg procedure. **(H)** Transplantation experiments. Sequences from *sqt-3(ins)* or *sqt-3(revertant2)* were knocked-in at the *dpy-10* or *unc-22* 3′ UTR and the known reduction-of-function phenotypes were evaluated.

Discussion and Conclusion

Main findings

The aim of this study was to develop a method that could introduce many different mutations along a non-coding region and measure the impact on gene expression and phenotype. We demonstrated that heat-shock induction of Cas9 and multiplexed sgRNAs create diverse indel mutations at the targeted DNA in expanded *C. elegans* populations, circumventing the need for individual microinjections. Using targeted sequencing of samples with diverse individual deletions along the *lin-41* 3′ UTR, we showed that the two *let-7* miRNA binding sites are not cooperating, in contrast to previous studies. We also targeted regulatory sequences of genes with known functions in establishing morphology. Screening for changes in phenotype allowed us to isolate dozens of alleles, to infer functional regulatory sequences and uncover genetic interaction between sequences.

Interpretation of results

Characteristics of Cas9-induced indel mutations

We showed that heat-shock-induced Cas9 expression in *C. elegans* populations can produce progeny with indel mutations at the targeted genomic sequences. We then characterized the allele diversity created by these mutations. Our results suggest that a single sgRNA can create unique deletions at nucleotide resolution within a window of ~10 bp, and some larger deletions at lower frequency, which together can be useful to interrogate single binding sites and their immediate vicinity. Multiple sgRNAs can introduce larger deletions of up to a thousand bp (depending on the sgRNA target sites), thereby allowing to test genetic interaction between separated gene regulatory sequences. Insertions create semi-random sequences *in situ* and thus essentially can be used as a massively parallel reporter assay. On average, our samples contained proportions of 57.9% deletions, 22.9% insertions and 19.3% complex events, which resembles naturally occurring germline indels in *C. elegans* (75% deletions, 25% insertions)[99] and human (50% deletions, 35% insertions)[100]. This suggests that our method could simulate the impact of natural indel mutations for a site of interest. Compared to Cas9-outcomes in human cells, we found longer indels and many insertions templated from the surrounding sequence[103,123–127]. This can likely be explained by a higher activity of microhomology-mediated end joining (MMEJ), or its sub-process, theta-mediated end joining (TMEJ), that use 5–25 bp microhomology[47,101]. TMEJ has been previously reported as the main dsDNA-break repair pathway in *C. elegans*[102]. Some models of MMEJ/TMEJ include *cis* intramolecular synthesis ("snap-back"), which would result in inverted repeat insertion and k-mer matches of insertions to the opposite strand[101,128–130]. Our data, shows k-mer matches mainly to the same strand, which suggests that *trans* intermolecular synthesis is the main mechanism of MMEJ/TMEJ dsDNA-break repair in the *C. elegans* germline (Supplemental Figure 2)[88,131].

Activity and interaction of *let-7* miRNA binding sites

We then showed that mutated *C. elegans* populations can be analyzed with targeted sequencing to measure the impact of gene regulatory mutations on gene expression and phenotype. We measured the expression of more than 900 different deletion alleles along the *lin-41* 3′ UTR by targeted RNA sequencing. These data were complemented by competitive genotype fitness derived from DNA sampling over generations. The 4-fold upregulation of *lin-41* mRNA only after simultaneous deletion of both *let-7* binding sites is consistent with the amount of natural down-regulation by *let-7*[105,108,109] and the de-repression observed when disrupting both *let-7* interactions (2–4-fold)[106,113,114]. Surprisingly, we found that each binding site could function on its own. Previous studies, together with their proximity

of 27 bp, suggested that both *let-7* sites may act cooperatively[109,132]. This was based on findings that each site alone in one-, two-, or three copies does not show repressive activity[109,133] – although others reported that three copies can show wildtype activity[110] – and that the identity of the intervening sequence is important[109,110]. All these experiments used lacZ reporter overexpression and possibly lacked the sensitivity to detect the activity of each single site. Systematic *in vitro* experiments have previously determined that similar miRNA sites - extensively paired with a bulge - would not cooperate at distances over 13 nt[134,135]. On the other hand, miRNA cooperativity has been observed *in vivo* at distances of 27 nt or more (for *mir-35*)[136,137].

We found a stronger effect on RNA regulation and phenotype when disrupting *let-7* binding site1 compared to site2. Site1 has a stronger seed pairing, which might indicate a stronger repressive activity[138]. Unexpectedly *let-7* levels were elevated in site1-, and even more in site2 mutants. This is consistent with target-dependent miRNA degradation (TDMD)[139–141], in particular for site2, which is also supported by computational predictions (A. Filipchyk, personal communication). Up-regulation of *lin-41* after deletion of each site separately adds up to less de-repression than the complete loss of *let-7*. This points to a compensatory interaction between the sites, which could be involve feedback through TDMD (Supplemental Figure 3A). Such a negative feedback loop could tighten the time-window of *lin-41* repression during development (Supplemental Figure 3B,C). Future studies could investigate this further (Supplemental Figure 4A-C). In most cases the depletion of different *let-7* target genes suppresses vulval bursting in *let-7(n2853)* mutants, which raises the possibility that these targets also reduce *let-7* levels[114,142–148]. The *C. elegans* homolog for the human ubiquitin E3 ligase that mediates TDMD, *ebax-1*, plays a role in reducing levels of the *mir-35* family during development, but shows no strong loss-of-function phenotypes[82,139,141]. It is possible that TDMD function expanded to other E3 ligases, especially in light of other gene expansions in *C. elegans* (e.g., 19 Argonautes)[149]. TDMD pathway genes could be found among suppressors of the *let-7* bursting phenotype[114,142–148,150] or using a forward genetics screen (Supplemental Figure 4D).

Isolation of *reduction-of-function* mutations using phenotypes

Finally, we demonstrated that our method can be used to discover functional gene regulatory sequences that affect animal physiology. For this we targeted regulatory regions of genes with known morphological phenotypes and manually isolated 57 mutants with phenotypic defects. These various hypomorphic alleles were used to identify the gene regulatory sequences underlying the phenotype. For the cuticle component *sqt-3*, many templated insertions in its 3′ UTR led to the *reduction-of-function* Rol phenotype, while deletions were tolerated. We tried to understand these unexpected sequence constraints by identifying the mechanism but were unsuccessful. This could be investigated further in the future (Supplemental Figure 6A-D). Gain and loss of regulatory sequences are important processes in the determination and evolution of phenotype[1,10,12,151–153]. Although rarer than point mutations, templated insertions might be underestimated in their ability to generate or multiply functional sequences with immediate phenotypic consequences.

Using iterative mutagenesis and selection by phenotype, we uncovered unexpected intragenic suppressor mutations. This approach could be extended to other genes, to explore the general potential of genetic interactions within gene regulatory regions or their potential to suppress phenotype defects from coding mutations[77,119,120,154]. The multidimensional phenotypes affected by these genes (length, thickness, curvature, movement) can be analyzed with automated imaging setups[155–157]. Alternatively, many other phenotypes can be selected in automated and quantitative ways (Supplemental Table 4)[158].♦

Implications

This study demonstrates a generalized workflow for parallel, targeted mutagenesis screens in *C. elegans*, to systematically explore function of gene regulatory sequences and their impact on phenotype. Single aspects of this workflow will be useful to measure sequence function in bulk populations, explore phenotype formation and -plasticity, discover interactions, rules, and constraints of gene regulatory sequences, or create dense genotype-expression-phenotype maps (**Figure 21A-E**). Our results for the *lin-41* 3'UTR indicate a compensatory genetic interaction of *let-7* miRNA binding sites that might provide a biological function for target-dependent miRNA degradation.

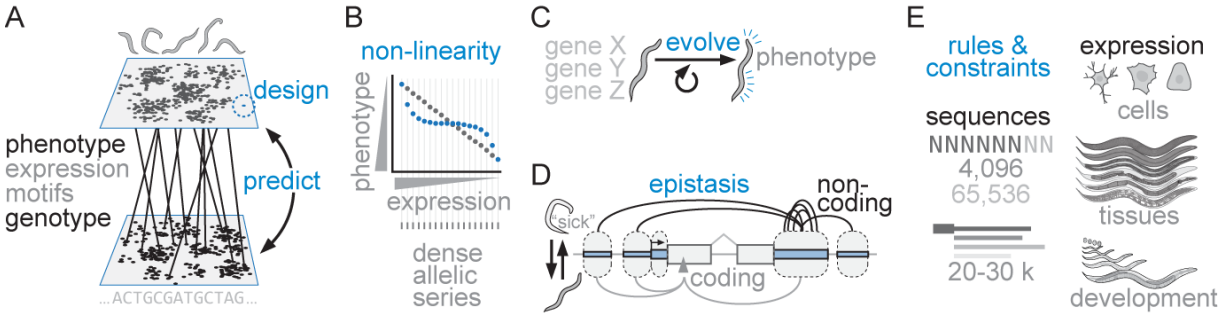


Figure 21. Parallel genetics to study gene regulatory sequences in *C. elegans*
Parallel genetics can be used to study the relationship between regulatory sequences, expression, and phenotype in several ways, for example: **(A)** Dense genotype-phenotype maps to learn rules and design new extreme phenotypes. **(B)** Studying non-linear expression-phenotype relationships[68]. **(C)** Testing evolvability towards phenotypes (longevity, behavior, morphology, resistance). **(D)** Analyzing epistasis (genetic interaction) between non-coding-, and between coding- and non-coding sequences. **(E)** Creating parallel reporter assays to study rules and constraints of regulatory sequences.

Limitations

As we have demonstrated in this study, sgRNA efficiencies around 1.5 % are sufficient to analyze effects of mutations on gene regulation and phenotype, when coupled with deep sequencing and manual- or automated selection of animals from large populations. However, in many cases this efficiency is impractical for comprehensive analyses, for instance, to exhaustively determine all non-functional mutations in a regulatory sequence. Cas9 RNP injections or integrated Cas9 lines can now reach 10-80% editing efficiency[87,159–161]. This is immediately advantageous for small scale applications, for example to isolate 10-100 mutations, and could also be combined with population-scale bulk RNA-seq to analyze thousands of non-lethal mutations.

CRISPR-Cas9 naturally comes with several limitations. Its NGG PAM prevents dense tiling in many regions which restricts the resolution. Furthermore, apart from sgRNA-dependent off-target mutagenesis, dsDNA-breaks have non-specific effects on cell cycle, chromatin, cellular physiology[162–164], and can result in large chromosomal mutations[165–167] – necessitating careful experimental design and validation.

Our large amplicon sequencing protocol does not incorporate UMIs as they would be separated from the initial molecule by the tagmentation step. For the targeted full-length RNA sequencing we added 8 nt UMIs, but the high read coverage would have required significantly longer UMIs to be uniquely assignable to individual molecules. Because of this, we simply used high DNA input and PCR amplicons that were ten times longer than the largest possible deletion expected from sgRNA target sites. We also validated the main results with independent methods, such as indel length estimations with Sanger sequencing or *lin-41* expression with qPCR measurements on individual mutant strains.

Along the 1.2 kb *lin-41* 3′ UTR we did not detect other sequences with a significant effect on mRNA expression apart from the *let-7* binding sites. Because we measured changes from L1 to L4 developmental stage, other parts of the 3′ UTR might be functional at other time points and environments, or they could be completely dispensable.

For the phenotype screen we relied on manual selection of mutants, which likely misses subtle phenotypic defects. Together with the relatively low mutagenesis efficiency we likely missed many functional mutations; thus, the absence of mutations cannot imply non-functionality for these regions.

Recommendations

There are several technical recommendations for future parallel genetics approaches. A significantly improved efficiency could likely be achieved with more advanced expression systems and optimizations for germline expression[80,160,168,169]. Alternative induction systems[80,170–172] would enable continuous germline-specific Cas9 expression to further increase efficiency, and allow synthetic evolution experiments[173]. For denser tiling of targeted regions, CRISPR-nucleases with dispensable PAM requirements could be used (Supplemental Figure 7A,B)[45,174,175]. Inducible sgRNA expression might be possible with PolII-compatible architectures or an RNA-based induction system (Supplemental Figure 7C)[171]. Although we focused on native loci, Cas9 could also produce semi-random sequences at reporter genes *in situ* for massively parallel reporter assays. For point mutagenesis, inducible homology-directed repair would be suitable[61], while base editors[176–180] or programmable prime editors[60,181,182] have the advantage of functioning without dsDNA break, but are possibly less efficient and not yet established in *C. elegans*. To reduce PCR biases for long read sequencing, unique molecule counting methods could be incorporated[183,184]. More established protocols are available for 100-300 bp target regions, which are recommended if possible[125].

Implementations with alternative genetic tools could be based on inducible expression of FLP/Cre recombinases[170,185], Mos1 transposon[96,186], or serine integrases[187] (Supplemental Figure 8). Not all could replace endogenous sequences to evaluate phenotype, but all would be good candidates for massively parallel reporter assays, and likely would also support multiplexing insertions per animal to increase throughput. For this, one injection would produce several extra-chromosomal array lines with each likely fitting a DNA library of 1-2 Mbp (e.g., ~200-400 copies of 3,000 bp)[79,188,189], which would allow delivery of diverse libraries with few injections. Expression could be measured by bulk sequencing and by microfluidic worm sorting[190], or at single-cell level using combinations of FACS-sorting and targeted scRNA[74,191–197].

Concluding summary

We have shown that CRISPR-Cas-based mutagenesis can be parallelized in *C. elegans* to study gene regulatory sequences *in vivo*. This can be achieved with existing tools combined with a powerful experimental workflow, that was so far mostly restricted to cell culture systems. We have shown that this method creates diverse mutations which can be used to identify functional gene regulatory sequences and connect these directly to animal physiology. Along the way, we made insights into genome editing and DNA repair mechanisms in *C.elegans*. Our results for the well-studied *lin-41* 3' UTR indicate a compensatory genetic interaction of *let-7* miRNA binding sites that might involve target-dependent miRNA degradation. The findings for the *sqt-3* 3' UTR show that templated insertions from dsDNA break repair can result in strong regulatory changes with phenotype defects, and that genetic interaction within a 3' UTR has the potential to revert such defects. Together with future improvements of efficiency, the genetic approach presented here suggests a layout for explorative and comprehensive studies to better understand fundamental functions of gene regulatory sequences in animals.

Methods

Establishing the approach for parallel genetics in *C. elegans*

Caenorhabditis elegans culture

The wild-type strain N2 Bristol[198] was used to create transgenic lines for experiments. In a screen for phenotypes, we isolated several mutants and revertants for different regulatory regions. For initial tests we generated a *his-72* c-terminal GFP knock-in strain (NIK123) which we crossed into a strain expressing *Peft-3:tdTomato:H2B* from a single copy insertion (EG7927)[96] resulting in a GFP/tdTomato expressing strain (NIK124) for automated quantifications and sorting using the Copas Biosorter. A complete list of strains can be found in Table S3 of ref.[94].

Animals were maintained on NGM plates with *Escherichia coli* OP50 as originally described[77], at 16, 20 or 24 °C. Plates for hygromycin resistant transgenic animals were modified by adding working stock solution of 5 mg/mL Hygromycin B (Thermo Fisher) in water onto plates before use, to a final concentration of 75 µg/mL NGM. For standard 6 cm plates with 10 mL NGM that would be 150 µL of 5 mg/mL Hygromycin working stock solution.

Plasmid construction

A list of all plasmids created or used in this study can be found in Table S3 of ref.[94]. The plasmid for heat-shock inducible *Streptococcus pyogenes* Cas9 expression (pJJF152) was created by Gibson assembly[199] of a previously published *C. elegans* optimized SpCas9[200] ("Friedland Cas9"), with the *hsp-16.48* heat-shock promoter and the *unc-54* 3′ UTR using HiFi DNA Assembly Master Mix (NEB). The plasmid backbone for sgRNA expression (pJJF439) was created by PCR amplification of the U6 promoter of *W05B2.8* and replacing the promoter of pJJR50, using restriction digest and Gibson assembly.

Plasmids for sgRNA expression were cloned as previously described using one of two published backbones, pMB70[92], pJJR50[93] or pJJF439[(this study)]. For this, 5-10 µg of backbone was digested using 1 µL Fastdigest Eco31I (aka BsaI, Thermo Fisher) or Fastdigest BpiI (aka BbsI, Thermo Fisher) at 37°C for 2-6 hrs, separated from undigested plasmid on a 1.5% Agarose/TAE gel, and extracted using the Zymoclean Gel DNA Recovery Kit (Zymo), according to the instruction manual. Two complementary DNA oligonucleotides containing the spacer sequence, plus an optional 5' G for optimal U6 promoter expression, and 4 nucleotide overhangs for ligation into the backbone were phosphorylated and annealed in a thermocycler. This reaction contained 1 µL of each oligo (at 100 µM), 1 µL of 10x T4 DNA ligase buffer (Thermo Fisher), 1 µL T4 PNK (Thermo Fisher) and 6 µL water and was incubated 37°C 30 min, 95°C 5 minutes and cooled down at -0.1 °C/second to 25°C. Sample was diluted 1:200 in water and 1 µL was used for ligation with 70-130 ng of linearized backbone, 1 µL of 10x T4 DNA ligase buffer and 1 µL of T4 DNA ligase (Thermo Fisher) and water to a volume of 10 µL. Ligation was performed at room temperature for 1 hr or overnight. 5 µL were then transformed.

The HDR repair template plasmid used for the *his-72::GFP* knock-in was prepared as described previously[201].

For transformation and amplification, we used DH5alpha Mix & Go Competent Cells (Zymo) in all the above clonings except for the *his-72::GFP* repair template which required ccdB resistant bacteria for which we used One Shot ccdB Survival (Thermo Fisher). DNA extractions by miniprep were done with the ZymoPURE Plasmid Miniprep kit (Zymo) and elution with water.

sgRNA design

Most sgRNAs were designed using the CRISPOR web application (http://crispor.tefor.net/)[202]. Some sgRNAs were designed manually using the plasmid editor Ape (A plasmid Editor, M.W. Davis, unpublished, https://jorgensen.biology.utah.edu/wayned/ape/). All sgRNAs were designed for *C. elegans* genome version ce11 and we evaluated all sgRNAs using the E-CRISP web application (http://www.e-crisp.org/E-CRISP)[203]. For regulatory regions of interest, we aimed at a regular spacing between target sites, dense coverage and as little as possible predicted off-targets with less than three mismatches. A detailed list of sgRNA sequences, together with their characteristics, efficiency prediction scores and predicted off-targets can be found in Table S3 of ref.[94].

Generation of transgenic *C. elegans*

Extra-chromosomal array transgenes were generated by standard procedure using micro-injection into the gonad[204]. A detailed list of injection mixes and their composition can be found in Table S3 of ref[94]. The injection mix usually contained plasmids for heat-shock inducible Cas9, pMB67[92] or pJJF152 (this study) at 50 ng/µL, 1-

10 sgRNAs using the backbones pMB70[92], pJJR50[93] or pJJF439[(this study)] at 10-50 ng/µL, a visual co-injection marker expressing mCherry in the pharynx, pCFJ90[95] at 5 ng/µL, and hygromycin resistance IR98[205] at 3 ng/µL. For large scale experiments followed by targeted DNA sequencing we used pMB67 for Cas9 expression and sgRNAs cloned into the pJJR50 backbone. Independent lines were created from F1 animals selected for pharynx expression of the mCherry co-injection marker. Lines were maintained on Hygromycin as described above.

C-terminal GFP knock-in of *his-72*

C-terminal GFP knock-in of *his-72* was performed as described previously using a self-excising selection cassette[201].

Biosorter

Automated measurement of GFP negative animals in F1 and their F2 progeny. *His-72::GFP* was targeted with sg1, sg2, pool1 (sg2, 3, 4, 6, 8) or pool2 (sg3, 5, 8). F1 generation was collected by bleaching 12 hrs after heat-shock. These were either measured on the Biosorter flow system at larvae stage L3 or grown to adulthood to collect F2 generation which was then also measured at larvae stage L3. The number of analyzed worms per sample was between 1,662 and 21,983 worms.

Small-scale Cas9 induction and time course

20-40 egg-laying adults were transferred to small 6cm NGM plates with OP50 *Escherichia coli* and without Hygromycin. Plates were placed in a programmable incubator "Innova 42" (New Brunswick Scientific/Eppendorf) at 20°C. Heat shock was applied for 2 hours at 34°C, followed by 20°C. For time course experiments adults were transferred to new plates using a picking tool at regular time intervals (14, 16, 18, 20, 22, 43 or 12, 15, 18, 21, 48 hrs) after heat shock to analyze eggs laid within each interval. PCR genotyping of GFP-negative *his-72::GFP* knock-in animals was done as described further below under "PCR genotyping".

Induced homology-dependent repair of non-fluorescent GFP

From targeting the GFP in *his-72::GFP* and selecting GFP-negative animals, one allele was occurring more frequently, pointing to a nonrandom indel outcome that favors this 4 bp deletion (Figure 2) (11/46 of GFP-negative mutations, 24%). Such a *his-72::nfGFP4del* strain was injected with mixes containing a PCR product from the original, intact, *his-72::GFP*, in addition to the usual plasmids for heat-shock inducible Cas9, sgRNA(s) and co-injection markers. All conditions contained a sgRNA specific for the "nfGFP4del". In one condition ("with overhangs") the PCR amplicon contained additional overhangs including a PAM sequence, that allowed targeting the ends with a specific sgRNA. The rationale was, that the PCR template for homology-dependent repair might not be accessible for DNA repair if it is present in concatenated extra-chromosomal arrays. In another condition the PCR amplicon with overhangs was prepared with two rounds of error prone PCR, using the "GeneMorph II EZClone Domain Mutagenesis Kit" (Agilent Technologies), to introduce random point mutations.

Developmental synchronization

Synchronized L1s were obtained by bleaching, as previously described[206]. Egg-laying animals were washed off plates in 50 mL M9 buffer (42 mM Na_2HPO_4, 22 mM KH_2PO_4, 86 mM NaCl, 1 mM $MgSO_4$) and settled for 10 minutes. M9 was aspirated until a remaining volume of 7.5 mL. Then 1 mL 12% NaClO and 1 mL 5 M NaOH were added. Worms were incubated under gentle rotation, vortexed briefly after 4 minutes and incubated under constant observation for another 3 minutes. Bleaching was stopped by addition of 40 mL M9 when circa 50% of animals were dissolved. Eggs were then pelleted by centrifugation at 1,200 g for 1.5 minutes and washed two more times using M9, centrifugation and decanting. Finally, eggs were resuspended in circa 4 mL M9 and left shaking at 16 °C overnight for at least 12 hours to allow hatching and developmental arrest of L1 larvae. Larvae concentration was then counted in triplicates and the desired amount was dispensed on plates with food to begin synchronized development.

Large-scale experiments, targeted DNA sequencing and analysis

Large-scale Cas9 heat shock induction

Before the experiments, animals were maintained 5-25 generations in culture under Hygromycin selection to ensure expression of transgenes. Expression was indicated by Hygromycin resistance, and the visual mCherry co-injection marker expressed in the pharynx. For all experiments three independent lines from the same injection mix were used. For transient heat shock induction of Cas9, synchronized populations were seeded on large 15 cm NGM plates with food and without Hygromycin. Plates with egg-laying adults (P0) were placed in a programmable incubator "Innova 42" (New Brunswick Scientific/Eppendorf) at 20°C and 34°C heat shock was applied for 2 hours. Because the heat shock conditions were optimized with small 6 cm NGM plates, 2 hours at 34°C might not be long enough for optimal induction of the large 15 cm plates that contain ~3-times more NGM. Plates were kept at 20°C for 12 hrs and eggs were collected by bleaching as described above for developmental synchronization. Hatched larvae, arrested at the L1-stage, the first generation after Cas9 induction (F1), were then again seeded on large NGM plates with food for synchronized development until egg-laying, to collect the next generation (F2) by bleaching. We used this F2 generation for all experiments to ensure non-mosaic animals generated by F1 germline mutations. We seeded 50,000 P0 for Cas9 induction at 24°C on Hygromycin (25,000 / big plate), and 100,000 F1 at 16°C (25,000 / big plate). 400,000 F2 were frozen for genomic DNA extraction to determine introduced indel mutations. The remaining F2 were used for experiments described below.

Genomic DNA extraction

Genomic DNA was obtained using worm lysis, phenol-chloroform extraction, and ethanol precipitation. Worms were washed once in 50 mL M9 buffer and frozen in 1 mL M9. After thawing, M9 was removed and 100 µL of TENSK buffer (50mM Tris pH 7.5, 10 mM EDTA, 100 mM NaCl, 0.5% SDS. 0.1 mg/mL proteinase K, 0.5% ß-Mercaptoethanol) was added. Sample was incubated for 1.5 hrs at 60 °C while shaking at 1,000 rpm on a benchtop heating block. 300 µL of water was added, followed by 400 µL phenol/chloroform/isoamylacohol pH 8.0 (Carl Roth). Sample was mixed by shaking the tube and centrifuged for 10 min. at 15'000 g at room temperature. The upper aqueous phase, circa 350 µL, was transferred to a new tube and an equal volume of chloroform was added. After additional centrifugation 10 min. at 15,000 g at 4°C, the upper aqueous phase was transferred to a new tube, and 2 µL glyco blue added. This was followed by addition of 30 µL 3M NaAc (pH 5.2-6) and 1 mL pure ethanol. Samples were centrifuged for 10 min. at full speed and 4°C in benchtop centrifuge. Pellet was washed once with 70% ethanol and resuspended in 25 µL water at 50°C for 30 min. Then 0.25 µL RNAse I (10 U/µL, Thermo Fisher) was added and incubated for 30 min. at 37°C. DNA concentration was determined on a Nanodrop ND-1000 (Thermo Fisher) and diluted to 50-200 ng/µL in water. Since we did not test other protocols or commercial solutions to extract genomic DNA, likely there are alternatives that could be quicker and better in preserving DNA integrity.

DNA long amplicon sequencing

Amplicons were designed so that they contained all the regions of a gene targeted in our experiments. 0.5 – 3 kb amplicons were large enough that deletions between the outermost sgRNAs would not change the amplicon size by more than 10% to avoid more efficient amplification of templates with large deletions. Furthermore, large amplicons should capture the reported large deletions missed by 100-300 bp amplicons of other workflows. Primers used for amplification together with annealing temperature and resulting amplicon sizes can be found in Table S3 of ref.[94]. Genomic DNA concentration was fluorimetrically quantified using Qubit dsDNA HS kit (Thermo Fisher). For PCR reactions we used 100 ng template DNA. We calculated that 100 ng of genomic DNA equals more than 90 million *C. elegans* genomes and therefore represented all animals in our samples that contained for most samples 400,000 and maximal (for DNA sampling over generations) 2,000,000 animals.

50 µL PCR the reactions were set up as follows. Phusion HF polymerase (NEB) 0.2 µL, 5X HF buffer 10 µL, dNTP mix 1 µL, forward and reverse oligos at 10 µM 5 µL, water 32 µL, and template DNA. Samples were incubated at 98°C 3 min, followed by 35 cycles of 98°C 15 sec, 58-72 °C 30 sec, 72 °C for 7 min with a final elongation at 72 °C for 7 min. PCR reactions were analyzed on agarose gels to ensure successful amplification.

Cleanup was then done by either agarose gel or SPRI beads. For gel-based cleanup 1.5 % Agarose/TAE gels were run and bands were excised with circa +/-500 bp, to also include products with deletions or insertions. DNA was recovered from agarose gel using the Zymoclean Gel DNA Recovery Kit (Zymo). For SPRI beads cleanup

and no size selection we used AMPure XP Reagent (Beckman Coulter). 0.8 x volume of beads were added to PCR reactions, incubated 2 min at room temperature, washed twice with freshly prepared 80 % EtOH using a magnetic rack, and eluted with water.

DNA was quantified by Nanodrop, diluted to 5 ng/µL, quantified by Qubit, diluted to 0.4 ng/µL, quantified by Qubit and diluted to 0.2 ng/µL for library preparation. Library preparation was done with the Nextera XT DNA kit (Illumina) which fragments input DNA and adds sample-specific barcodes by tagmentation. Although we used one barcode per sample, it is also possible to pool amplicons before library preparation and use the same barcode for multiple samples provided that samples don't need to be identified individually or that reads for each sample can be distinguished after mapping (e.g., non-overlapping amplicons from different genes). Libraries were analyzed with a Tapestation D1000 ScreenTape system (Agilent) or Bioanalyzer HS DNA kit (Agilent) and showed an average fragment size of around 500 bp (range 400 – 600 bp). Average fragment size, together with the DNA concentration measured with Qubit, was used to determine molarity and an equimolar pool of libraries was prepared. This pool was again analyzed using Tapestation or Bioanalyzer, measured by Qubit and diluted to 2 nM as input for the Illumina sequencing workflow. The library pool was then sequenced using 150 bp reads with a Miniseq Mid Output kit, 2x150 cycles (Illumina), or a Nextseq 500 V2 Mid Output kit, 150 cycles (Illumina).

Established protocols with UMIs are available for shorter, 100-300 bp, PCR amplicons and are recommended if the experimental design allows[125]. Alternative sequencing platforms or protocols that achieve longer read lengths might be able to cover PCR amplicons up to 600 bp. For approaches that skip the tagmentation step and directly use long read sequencing (e.g., Pacbio or Nanopore) or synthetic long reads (e.g., Illumina or 10x) to cover longer PCR amplicons, unique molecule counting methods (using UMIs) should be incorporated to reduce PCR biases[183,184].

Analysis of targeted sequencing using crispr-DART

Primary analysis of targeted sequencing data was done using the crispr-DART software. A detailed description of crispr-DART can be found in the associated publication[94] and at https://github.com/BIMSBbioinfo/crispr_DART. The pipeline accepts short or long reads from different platforms as input and has no technical limit on the size and number of target regions or number of sgRNAs. Sequence of processing steps is shown in Figure 4. The resulting html reports give an intuitive overview of quality and efficiency of experiments, and characteristics of the introduced indel mutations. Additionally, the output consists of BAM files, bigwig files, BED files, and different tables. These were used for the downstream analyses that are described further below. R scripts for these analyses can be found at https://github.com/BIMSBbioinfo/froehlich_uyar_et_al_2020.

Browser shots

Browser shots were compiled using indel profiles and top indels provided by the computational pipeline crispr-DART as BigWig and BED files and loading them into the UCSC genome browser[207] or the IGV browser[208] followed by export as vector graphics compatible format. We used *C. elegans* genome version ce11/WBcel235 including 26 species base-wise conservation (PhyloP).

sgRNA efficiency comparisons

Crispr-DART calculates the efficiency of a sgRNA as the ratio of the number of reads with an insertion/deletion that start or end at +/- 5bp of the intended cut-site to the total number of reads aligned at this region. For untreated wild type control samples, we used all cut sites present in any of the treated samples of the same amplicon. For comparing observed efficiencies to published prediction scores and other sgRNA characteristics[98,209–216], these scores were manually extracted from the CRISPOR web application (http://crispor.tefor.net/)[202] for each sgRNA and compared to the sgRNA efficiencies determined by crispr-DART.

Indel characteristics

For indel proportions, the fraction of reads containing insertion, deletion or complex events was determined per sample. Complex events were defined as reads containing more than one event. These could be either insertions, deletions or additional substitutions which suggested a combination of multiple events.

For the distribution of indel lengths we considered all deletions or insertions supported by at least 0.001% of reads at that position, at least 5 reads and overlapping with any cut site +/- 5 bp. Deletions were further classified as "multi cut" deletions when a deletion overlapped with more than one sgRNA cut site +/- 5 bp or otherwise were classified as "single cut" deletions when they only overlapped with one cut site.

For the analysis of insertion origin, all 5-mers from insertions were extracted. Then matches to the surrounding sequence +/- 50 bp of the insertion position were counted on the forward and reverse complement strand. As control sequences nucleotides of insertions were shuffled randomly.

R scripts to reproduce these analyses and figures are available at the Github repository (see Data and Code Availability section).

Genotype diversity

For genotype diversity we considered indels supported by at least 0.001% of reads at that position, at least 5 reads and overlapping with any cut site +/- 5 bp. Each deletion, defined by start and end coordinates, irrespective of its abundance (except reaching the threshold defined above) was considered as one unique deletion genotype. Each insertion genotype was defined by position and by considering the inserted sequence. For untreated wild type control samples, we used all cut sites present in any of the treated samples of the same amplicon.

For the plots of "unique deletions per nucleotide by sgRNA", each deletion was assigned to a sgRNA when it was overlapping with its cut site +/- 5 bp.

R scripts to reproduce these analyses and figures are available at the Github repository (see Data and Code Availability section).

Analysis of the *lin-41* 3′ UTR and the *let-7* miRNA binding sites

Targeted mRNA sequencing

Mutated F2, arrested at the L1 developmental stage, were obtained from Cas9-induced P0 as described above. 40,000 were directly frozen for genomic DNA extraction. 80,000 were directly frozen for RNA extraction by adding 1 mL TRIzol reagent (Thermo Fisher), homogenization with a Precellys 24 tissue homogenizer (Bertin Instruments) and storage at -80°C. 5,000 L1s were seeded on large 15 cm NGM plates at 24°C and collected 32 hours later, at late-L4 stage, and prepared for RNA extraction like the L1 sample. At 32 hours, *lin-41* mRNA is fully downregulated[112], while the lethal vulva bursting occurs later, after molting, in the adult stage[106].

RNA was chloroform-extracted as follows. Samples were thawed, 0.2 mL of chloroform added, incubated for 3 minutes, and centrifuged for 15 minutes at 12,000 x g at 4°C. The upper aqueous phase was transferred to a new tube, 2 µL GlycoBlue (30 µg) were added, 500 µL of isopropanol were added and sample was incubated for 10 minutes. Sample was centrifuged 10 minutes at 12,000 x g at 4°C, supernatant discarded, and 1 mL of 75% EtOH was added. Sample was centrifuged for 5 minutes at 7,500 x g at 4°C, supernatant removed, pellet air-dried and resuspended in 20 µL RNase-free water. RNA concentrations ranged between 1,000 - 2,000 ng/µL, as determined on a Nanodrop ND-1000. Sample was diluted to 300 ng/µL and used for reverse transcription.

RNA was reverse transcribed using Maxima H Minus Reverse Transcriptase (Thermo Fisher). A reaction containing 11.5 µL RNA (3.45 µg), 2 µL gene-specific RT primer at 10 µM (oJJF890 "3'end", containing a UMI and PCR handle), 1 µL dNTP Mix (10 mM each), was incubated 5 minutes at 65°C. Then 4 µL 5X RT buffer, 0.5 µL RiboLock RNase inhibitor, and 1 µL (200 U) Maxima H Minus reverse transcriptase were added and the reaction was incubated for 30 minutes at 60°C, and 5 minutes at 85°C.

PCR was performed with a *lin-41*-specific primer containing a sample-specific barcode (oJJF1140-1147 for samples N2, 1516, 2627, pool3 at L1 and L4 stages) binding in the second last exon and a primer (oJJF960) binding the PCR handle introduced by the reverse transcription primer. 2 µL of each RT reaction was used as template in 4 PCR reactions, each containing 10 µL 5X HF buffer, 1 µL dNTP mix (10 mM each), 5 µL F+R primer mix (10 µM), 0.2 µL Phusion polymerase, 32 µL water and 2.5 µL DMSO (5% final). Samples were incubated at 98°C 3 min, followed by 35 cycles of 98°C 10 sec, 69 °C 20 sec, 72 °C for 1 min with a final elongation at 72 °C for 7 min. PCR was then analyzed on an agarose gel and DNA was cleaned up using Ampure XP beads (Beckman Coulter). For this the four PCR reactions were pooled resulting in 100 µL. 80 µL beads were added, incubated for 5 min at room temperature, washed once with 70% ethanol, and DNA was eluted in 10 µL water. This resulted in concentrations between 40-110 ng/µL. All samples were diluted to 40ng/µL and then

pooled. 32 µL of this pool (1280 ng) was then used as the input for SMRTBell (Pacbio) library preparation according to the instruction manual and sequenced using a Pacbio Sequel I sequencer.

In the future unique molecule counting methods could be incorporated to reduce PCR biases[183,184]. UMIs would need to be longer than usual, because of the high coverage from targeted sequencing, and to allow error correction.

RNA analysis of *lin-41* 3′ UTR deletions

Deletions supported by at least 5 Pacbio reads from L1 and L4 stage samples were filtered to keep only those deletions detected in both samples. No read percentage threshold was applied in this analysis. Each deletion was categorized based on their overlap with important sites in the 3′ UTR of *lin-41*.

Seed region of the first *let-7* complementary site (site1) ("LCS1_seed"): chrI:9335255-9335263
Seed region of the second *let-7* complementary site (site2) ("LCS2_seed"): chrI:9335208-9335214
Non-seed of the first *let-7* complementary site (site1) ("LCS1_3compl"): chrI:9335264-9335276
Non-seed of the second *let-7* complementary site (site2) ("LCS2_3compl"): chrI:9335215-9335227

Deletions were further categorized based on whether they overlap both *let-7* microRNA seed regions ("both"), and those that don't overlap any of these defined regions ("none").
Deletion frequency values were computed and the ratio of deletion frequencies between L4 stage and L1 stage samples were computed in log2 scale. For each category of deletions, a one-sided Wilcoxon rank-sum test was computed to test the null hypothesis that the stage specific abundance of deletions that overlap a *let-7* binding site is not greater from those deletions that don't overlap any of these sites.

RNA analysis by unsupervised clustering of long reads

Only Pacbio reads from both L1 and L4 stage *lin-41* RNA samples that covered the complete region between chrI:9334840-9336100 (the region from the beginning of the amplified segment up to the first intron) were selected, to make sure that all reads that go into analysis are covering the whole segment. For each read, the alignment of the read (including the inserted sequences) was obtained and all combinations of k-mers (k=5) were counted within these alignments allowing for up to 1 mismatch using Biostrings package[217]. Seurat package[218] was used to process the k-mer count matrix to do scaling, dimension reduction (PCA and UMAP) and network-based spectral clustering. The clustering of long PacBio reads covering the region enabled us to cluster reads into genotypes, thus taking advantage of the length of the reads while also allowing for the high rate of indels in the PacBio reads (compared to Illumina reads).

DNA sampling over generations

Mutated F1 samples were obtained as described above using large-scale mutagenesis by Cas9 heat shock induction. For this we used N2 as control and 3 lines with sgRNAs against the *lin-41* 3' UTR (sg15 and sg16, sg26 and sg27, sg pool). We conducted the experiment at 16°C and 24°C. 3,000 L1 stage animals (F1 generation) were seeded on medium plates with OP50. After egg laying and hatching of the next generation (F2) after 3 or 5 days (24°C or 16°C) F1 and F2 were separated. For this, animals were washed from plates in a final volume of 2 ml M9 buffer into 2 ml Eppendorf tubes. Adult animals sink faster and after circa 2-5 minutes are collected at the bottom of the tube, while L1 animals still swim. This was carefully monitored visually. When most adults (95%) had sunken to the bottom, supernatant M9, containing L1 stage animals, was removed to a separate tube. This was repeated three times by adding 2 ml M9 and separation by sinking. Adult animals were frozen for genomic DNA extraction in circa 20 uL M9. For generations F2-F4, 2,000 L1 were seeded on new medium plates, and frozen as adults after separation from the next generation. Generation F5 was frozen at L1 stage. Genomic DNA extraction and targeted large amplicon sequencing was performed as described above.

Fitness analysis of *lin-41* 3′ UTR deletions

For this analysis, we used *lin-41* DNA samples sequenced with Illumina single-end sequencing from multiple generations from F1 to F5 of the same pool of animals treated with sgRNA guides "sg15 and sg16", "sg26 and sg27" or "sg pool". Deletions were considered for this analysis when they were supported in the F1 samples by at

least 0.001% of reads at that position and at least 5 reads. The important sites considered for this analysis were the following.

Seed region of the first *let-7* complementary site (site1) ("LCS1_seed"): chrI:9335255-9335263
Seed region of the second *let-7* complementary site (site2) ("LCS2_seed"): chrI:9335208-9335214
Non-seed of the first *let-7* complementary site (site1) ("LCS1_3compl"): chrI:9335264-9335276
Non-seed of the second *let-7* complementary site (site2) ("LCS2_3compl"): chrI:9335215-9335227
Poly-adenylation signal: chrI:9334816-9334821
Stop-codon: chrI:9335965-9335967

We wanted to address the question whether the deletions that exist at F1 were exposed to purifying selection over generations if they overlapped the important sites in the 3′ UTR region of *lin-41*. We did this analysis in two ways. First, we counted the deletions categorized by their overlap (or non-overlap) with the important sites that existed in F1 generation and analyzed how many of them still existed in later generations. Second, we did the same analysis at the level of reads: we counted the reads with deletions that overlapped or did not overlap the important sites from generations F1 to F5. When comparing the number of reads, the read counts were normalized by the library sizes (total number of reads in the sample).

Lin-41 strains with site1 or site2 deletions

We generated mutant strains by targeting either site1 or site2 using Cas9/tracRNA/crRNA RNP injections. Injection mix contained 0.3 µg/µl Cas9 protein (Alt-R Cas9 V3 from IDT), 0.12 M KCl, 8 nM Hepes pH 7.4, 8 µM tracrRNA (Alt-R from IDT), 8 µM crRNA (custom crRNA, Alt-R from IDT), 5 ng/µl pCFJ90 (RFP co-injection marker), in duplex buffer (IDT). To prepare injection mixes, Cas9 protein was mixed with KCl and Hepes. crRNA and tracrRNA were annealed in duplex buffer for 5 min at 95 °C and ramp down to 25°C and added. Cas9/tracRNA/crRNA mix was incubated at 37°C for 10 min. F1 progeny positive for the pharynx expressed RFP co-injection marker were singled, allowed to lay eggs at 16°C, then genotyped using single worm lysis followed by Sanger sequencing of PCR amplicons. We observed mutations in 12/24 (50%) (site1) and 15/32 (47%) (site2) genotyped animals. For each site we kept the two strains with the biggest disruption of the seed regions. We maintained these strains at 16°C. Strains were bleached for developmental synchronization as described above. Three 10cm plates with egg-laying adults were bleached for each strain. For the strain MT7626 *let-7(n2853),* which shows developmental defects, six plates were bleached. L1 larvae hatched overnight at 16°C.

For RNA quantifications, 7000 L1 larvae were seeded onto medium 10cm plates and cultured at 24°C. 30 hours into synchronized development animals were collected using M9. After settling 200 uL were added to 1 mL of TRIzol reagent, homogenized in a Precellys 24 tissue homogenizer (Bertin Instruments) and stored at -80°C. Samples were thawed and RNA was chloroform-extracted as follows. 0.2 mL of chloroform were added, incubated for 3 minutes, and centrifuged for 15 minutes at 12,000 x g at 4°C. The upper aqueous phase was transferred to a new tube, 2 µL GlycoBlue (30 µg) were added, 500 µL of isopropanol were added and sample was incubated for 10 minutes. Sample was centrifuged 10 minutes at 12,000 x g at 4°C, supernatant discarded, and 1 mL of 75% EtOH was added. Sample was centrifuged for 5 minutes at 7,500 x g at 4°C, supernatant removed, pellet air-dried and resuspended in 20 µL RNase-free water. RNA concentrations ranged between 1,000 - 4,000 ng/µL, as determined on a Nanodrop ND-1000.

To quantify mRNA by qPCR, the total RNA was diluted to 150 ng/µL and used for reverse transcription. RNA was reverse transcribed using Maxima H Minus Reverse Transcriptase (Thermo Fisher). A reaction containing 10 µL RNA (1. 5 µg), 2.5 µL water, 1 µL of random hexamer primer at 5 ng/µL, 1 µL dNTP Mix (10 mM each), was incubated 5 minutes at 65°C. Then 4 µL 5X RT buffer, 0.5 µL RiboLock RNase inhibitor, and 1 µL (200 U) Maxima H Minus reverse transcriptase were added and the reaction was incubated for 30 minutes at 60°C, and 5 minutes at 85°C. Quantitative real-time PCR (qPCR) was then performed using 10 µL SYBR green 2x (with 35 µL ROX/ 1mL), 2 µL forward and reverse primer mix (5 µM each), and 8 µL cDNA (10 ng/µL) (80ng total). Primers were tested using a stepwise four-fold dilution series for efficiency and melting curves for specificity. Reactions were performed in technical triplicates, water and RT- reactions served as controls for contamination and genomic DNA amplifications respectively. Differences in RNA/cDNA input were normalized using the tubulin gene *tbb-2* and fold changes were calculated relative to wild type (N2) samples.

To quantify microRNAs with Taqman assays, the total RNA was diluted to a final concentration of 2.5 ng/µL. Then the protocol of the TaqMan ® Universal Master Mix II (Applied Biosystems) was followed with modifications. Briefly, RT master mix was prepared slightly modified: 1.5 µL 10mM dNTPs, 0.25 µL Superscript III (200 U/µL), 3 µL 5x buffer, 0.2 µL Ribolock (40 U/µL), 2.15 µL H2O. Then 7.7 µL RT master mix were combined with 5.5 µL of total RNA concentrated at 2.5 ng/µL (total of 12.5 ng). 20x miRNA-specific "RT primer" were diluted to 5x in 0.1x TE buffer. Then 12 µL of RT master mix, containing total RNA from the previous step, were combined with 3 µL of 5x "RT primer". For miRNA-specific reverse transcription, samples were incubated in a thermal cycler 30 min. at 16°C, 30 min. at 42°C, 5 min. at 85°C, and kept at 4°C afterwards. For the real-time PCR amplification reactions were set up as follows. 10 µL TaqMan Universal Master Mix II, no UNG, 7.67 µL H2O were mixed. 1 µL of 20x TaqMan microRNA Assay mix was added. 1.33 µL of the RT product from the RT reaction was added. Real-time PCR system was then programmed for the following cycling parameters. 10 min at 95°C, then 40 cycles of 15 sec. at 95°C, 60 sec. at 60°C. Measurements were done in triplicates. RT primers for *let-7*, *mir-1*, and *U6* were used. "No target controls" were used to control for any non-specific amplifications, that were not observed. Comparison of let-7 levels between mutants and wild type were then calculated by first normalizing *let-7* levels to RNA input using *mir-1* (which is expected to be unaffected by mutating the *let-7* binding sites in *lin-41*), and then by calculating *let-7* abundance in mutants relative to N2 wild type.

For quantification of lethal and bursting phenotype, 200 L1 larvae were seeded onto small 6cm plates and cultured at 24°C. Photos and videos were taken at 50 hours into synchronized development. We then scored dead animals or animals that had burst (with the intestine exiting the body cavity through the vulva) by examining 200 animals per plate and 3 plates for each strain.

Isolation of regulatory mutations that change animal phenotype

Screen for regulatory sequences by phenotype

We targeted 8 genes with known RNAi-phenotypes (*dpy-2*, *dpy-10*, *egl-30*, *rol-6*, *sqt-2*, *sqt-3*, *unc-26*, *unc-54*) using different sets of sgRNAs against regulatory regions. We used lines in which we targeted the 3′ UTR and for some genes we used additional lines targeting predicted enhancer, TATA-box, initiator (INR) and upstream/promoter regions. A list with all samples can be found in Table 1.

For each transgenic line (injection mixes imJJF181-215) we screened 35,000 F2 animals produced from P0 with large-scale induced Cas9 expression as described above. Animals were seeded onto NGM plates with food at a concentration of 15,000 per 15 cm plates or at 2,500 - 5,000 per 10 cm plates. Plates were kept at 16°C or 24°C. We then directly screened these plates by eye. Additionally, we collected worms in M9 and dispensed worms in drops on an empty plate. We then observed worms moving in M9 and moving away after M9 was dried (<1 min.). Dpy, Unc, and Rol worms were identified by morphology, their movement in M9 or slow and otherwise impaired movement away from the spot of dispension. Potential mutants were then picked and kept on plates for 2 to 4 generations at 24°C to achieve homozygosity. Animals were then singled again by phenotype and genotyped. This resulted in isolation of several mutant strains with the same genotype. We could not distinguish between cousins/siblings coming from the same F1/F2 or independent mutants coming from independently mutated F1s. In these cases, we kept one representative strain. We determined that penetrance was complete for all alleles except for the *sqt-2* enhancer locus (n>300 animals). For sqt-2 the penetrance varied between 10-100%. We scored the expressivity of the phenotypes into three categories (+, ++, +++) (n>300 animals). All the reported phenotypes have been determined and validated for several generations at 24°C. We also validated the absence of the extra-chromosomal transgenes judged by the red fluorescent co-injection marker. For *sqt-3* all isolated Dpy animals, characteristic for complete loss-of-function, contained large mutations affecting the coding frame. We therefore screened mainly for reduction-of-function alleles by screening for Rol animals. Non-Rol revertants of the *sqt-3(ins)* Rol animals were isolated using the small-scale approach on 6 cm plates (see above) with injection mixes imJJF215 or imJJF230.

PCR Genotyping

Single worms were picked using a platin wire picking tool and immersed in 10 µL of worm lysis buffer (WLB) (10mM Tris pH 8.3, 2.5 mM $MgCl_2$, 50mM KCl, 0.45% NP-40, 0.45% Tween-20, 0.01% gelatine, and freshly added 100 µg/mL proteinase K). Samples were frozen at -80°C for at least 10 minutes, incubated at 60°C for 30-

60 minutes, and 95°C for 15-30 minutes in a thermocycler. 1 µL of lysate was used as template in the following PCR. 25 µL PCR reactions were set up as follows. Phusion HF polymerase (NEB) 0.1 µL, 5X HF buffer 5 µL, dNTP mix 0.5 µL, forward and reverse oligos at 10 µM 2.5 µL, water 16 µL, and template DNA. 98°C 3 min, followed by 35 cycles of 98°C 15 sec, 58-72 °C 30 sec, 72 °C for 7 min with a final 7 min at 72 °C. 2 µL of the reaction was then analyzed on an agarose gel. DNA was then cleaned up using AMPure XP Reagent (Beckman Coulter) by adding 0.8 x volume of beads to 23 µL PCR reaction, 2 min at room temperature, washed twice with freshly prepared 80 % EtOH using a magnetic rack, and eluted with 18 µL water. DNA was then either analyzed by T7 nuclease assay or directly sent to Sanger sequencing. T7 nuclease assay was performed on cleaned up DNA using T7 endonuclease. Sanger sequencing traces were aligned to genomic loci using Snapgene (GSL Biotech) and linear maps were exported as svg vector files to create figures.

mRNA pull downs

mRNA pull down was conducted as previously described ("in vivo Interactions by Pulldown of RNA", "viPR", ref.[121]). Pull down probes for *sqt-3* were designed using an initial set from an online probe designer for Stellaris RNA fish (www.biosearchtech.com/stellaris-designer) 20 nt long, with >2 nt spacing and "masking level" 5. From this list, probes were chosen with a GC content >0.4, and no or few off-targets based on a BLAST search. We excluded probes where off-targets were likely expressed more than 5% of the target transcript (only one probe, five probes with off-targets expressed at 0.5-0.01%), based on expression information in Wormbase. The final twelve probes were distributed along the *sqt-3* transcript and did not overlap the location of the insertion mutations. SsDNA HPLC-grade oligonucleotides with a 3' TEG-Biotin were ordered from Metabion in the dry 0,04 µmol scale, resuspended and mixed to a final pool of 100 µM (8.3 µM each).

Experiments were performed with 400,000 worms per sample. Worms were grown on *E.coli* OP50 (40,000 worms/15 cm plate), harvested in L3 stage (after ~25 h at 24 °C), washed three times in 0.1 M NaCl (600 x g for 2 min centrifugation), transferred to non-seeded NGM plates and crosslinked at 254 nm (1 J cm^{-2}). Worm pellets were frozen in liquid nitrogen and the pulldown procedure was performed as described before[121].

For identification by mass spectrometry, protein input and pulldown samples were processed using the Single-Pot Solid-Phase-enhanced Sample Preparation (SP3) protocol[219]. Resulting peptide mixtures were analyzed by LC-MS/MS technology as follows: Peptides were separated on a 20 cm reversed-phase column (75 µm inner diameter, packed with ReproSil-Pur C18-AQ,1.9 µm, Dr. Maisch GmbH)[220] using a 90 min gradient with a 250 nL/min flow rate of increasing acetonitril concentration (from 2% to 60%, in 0.1% formic acid) on a High Performance Liquid Chromatography (HPLC) system (Easy-nLC, ThermoScientific). Eluting peptides were analyzec on an Q Exactive HF-X mass spectrometer (Thermo Fisher Scientific). The mass spectrometer was operated in the data dependent mode with a 60K resolution, m/z 350-1800, 3 x 10^6 ion count target and maximum injection time 10 ms for the full scan, followed by Top 20 MS/MS scans using HCD (15K resolution, 1 x 10^5 ion count target, isolation width, 1.3 m/z, normalized collision energy of 27, maximum injection time of 22 ms. Each replicate was injected and measured twice. Ions with an unassigned charge state and singly charged ions were rejected. Dynamic exclusion was set to 30 s.

Raw data were analyzed using MaxQuant (1.6.7.0) with standard settings, unless stated otherwise in the following. Search parameters included two missed cleavage sites, fixed cysteine carbamidomethyl modification, and variable modifications including methionine oxidation, N-terminal protein acetylation, and asparagine/glutamine deamidation. The "match between runs" option was enabled. Database search was performed using Andromeda against the UniProt/Swiss-Prot worm database (April 2017) with common contaminants. Protein quantification was done based on razor and unique peptides and the label-free algorithm[221] was used with a minimum LFQ ratio of 1. Known contaminants, proteins only identified by site, and reverse mappings were filtered out from MaxQuant output. For pulldown samples, we additionally removed proteins with gene ontology (GO) terms related to "biotin" (PYC-1, MCCC-1, POD-2, PCCA-1, BPL-1), which are expected to enrich unspecifically during the procedure. Imputation of missing intensities was done with the Perseus software package (1.5.6.0), after log2-transformation of LFQ values (normal distribution, width: 0.3; shift: 1.8). To determine significance of proteins identified in triplicate, we calculated p-values with a moderated t-test, implemented in the Bioconductor LIMMA package, and corrected for multiple comparisons by the Benjamini-Hochberg procedure.

Analysis of small RNA sequencing was performed as follows. 3′ adapter, poly(A)-tails and 5′ overhang trimming was performed with custom scripts. Reads were aligned to the ce11 genome using Bowtie2. miRNAs were quantified using miRDeep2 with hairpin and mature strand sequences extracted from miRBase v21.

mRNA quantifications by Nanostring or qPCR

10 k L1-arrested synchronized animals were dispensed on 10 cm NGM plates with *Escherichia coli* OP50 at 24 °C. Worms were then collected at different time points (22, 24, 26, 28, 30, 32 hrs), washed once with M9 and homogenized in 1 mL of TRIzol reagent (Thermo Fisher) using a Precellys 24 tissue homogenizer (Bertin Instruments). RNA was isolated by standard phenol-chloroform extraction. RNA expression was quantified using an nCounter (Nanostring) which measures absolute RNA amounts using a set of gene-specific probes. Raw counts were normalized using reference genes ("house-keeping"). For quantitative real-time PCR (qPCR) of pre-mRNA and mRNA we used RNA from the 26 hrs time point where *sqt-3* expression peaked. Pre-mRNA was specifically detected using intron-overlapping primers, while mRNA primers overlapped with exon-exon junctions. Controls without reverse transcriptase ("RT-") were done to ensure specific amplification of cDNA and no amplification from potential contaminating genomic DNA. Final values were obtained by normalizing to pre-mRNA or mRNA of *tbb-2* and presented relative to N2 wild-type controls. QPCR was performed using Blue S´Green qPCR Kit following the instruction manual and quantification on a StepOnePlus real-time PCR system. Probes and primers can be found in Table S3 of ref.[94].

Sequence transplantations into the *dpy-10*, *unc-22* 3′ UTRs

Knock-in animals were produced using Cas9/tracRNA/crRNA RNP injections with ssDNA oligo repair templates. Injection mixes contained: 0.3 µg/µl Cas9 protein (Alt-R Cas9 V3 from IDT), 0.12 M KCl, 8 nM Hepes pH 7.4, 8 µM tracrRNA (Alt-R from IDT), 8 µM crRNA (custom crRNA, Alt-R from IDT), 3.15 ng/µl pJJF062 (GFP co-injection marker), 3.15 ng/µl pIR98 (HygroR), 0.75 µM of a ssDNA oligo repair template, in duplex buffer (IDT). To prepare injection mixes, Cas9 protein was mixed with KCl and Hepes. crRNA and tracrRNA were annealed in duplex buffer for 5 min at 95 °C and ramp down to 25°C and added. Cas9/tracRNA/crRNA mix was incubated at 37°C for 10 min. Then plasmids and ssDNA repair template were added and 10 P0 animals were injected. For each injection mix 8 F1s positive for the co-injection marker were picked and genotyped using two PCR reactions (one primer pair flanking the insertion, the other with one primer binding in the insertion).

APPENDIX

List of Terminology

General

Gene regulatory sequence	Genomic sequence that can increase or decrease the amount; or affect localization or interactions of a gene's mRNA or protein.
Genetic interaction	Relationship between two genomic sequences that affects each other's activity or that connect their activity to the same biological process.
Robustness	Ability to maintain function under perturbation
In vivo	In living multicellular organisms, which can replicate in nature (e.g., animals).*
In vitro	In biological systems that need to be maintained in the lab and cannot replicate in nature (e.g., cell lines, organoids), or which are naturally single-celled (e.g., yeast, bacteria), or mixtures of biological components outside of their biological systems (e.g., RNA, DNA, protein in the test tube).*
Multiplexing	Multiple signals combined within one readout, for example multiple sgRNAs or sequences per analyzed unit (e.g., per animal or per cell).
Forward genetics	Random mutations are introduced into an animal population, by physical or chemical mutagens, and screening for individuals with altered phenotypic trait. Followed by identification of underlying mutation one-by-one ("mapping").
Reverse genetics	Targeted mutagenesis or perturbation of a pre-determined sequence or gene, usually with the intention to inactivate it, followed by analysis of the functional or phenotypic consequence one-by-one.
Parallel genetics	Many targeted perturbations are introduced and analyzed in parallel, with one perturbation per analyzed unit (e.g., per animal or cell). Analysis of many units and perturbations in parallel.

Phenotype

Phenotypic trait	An observable characteristic of an organism (e.g., the hair color brown). Can be monogenetic (determined by one gene) or more commonly multigenetic (determined by several genes). Usually determined by genetic- together with environmental factors.
Phenotype	Combined observable characteristics ("traits") of an organism. For example, related to physiology, morphology, or behavior.
-penetrance	Fraction of individuals with the same genotype showing a trait.
-expressivity	Degree to which a trait is displayed.
-plasticity	Capacity of a trait to change in response to the environment.
Rol	Roller, a phenotypic trait of *C. elegans* where a defective cuticle results in rotation along the body axis and movement in circles.
Dpy, Slu, Egl, Twi, Unc	Other *C. elegans* phenotypic traits: Dumpy – short and thick. Sluggish – slow movement. Egg laying defective – eggs retained in gonad. Twitching – tremors across the body. Uncoordinated – paralysis.

*Definitions may differ – this is the one used in this study.

CRISPR-Cas genome editing

CRISPR-Cas	From "clustered regularly interspaced short palindromic repeats" -CRISPR, and "CRISPR-associated proteins" -Cas. Diverse microbial defense systems that have been repurposed for genome editing across the tree of life.
Cas9	RNA-guided DNA nuclease from a particular CRISPR-Cas system (CRISPR-Cas9), which creates dsDNA-breaks at a genomic site specified by the sgRNA.
sgRNA	"single-guide RNA", It guides Cas9 to its genomic target site. Engineered fusion-RNA of the 42 nt crRNA ("crispr RNA") containing 20 target-specific nucleotides, and the 85 nt structural tracrRNA ("trans-activating crispr RNA").
Protospacer	Target-specific 20 nucleotides of the sgRNA (or crRNA).
Cut site	Position of DNA-break by Cas9. Determined by sgRNA base-pairing with target genomic DNA. For Cas9 this is commonly between nucleotides 3 and 4 from the PAM.
PAM	"protospacer adjacent motif". Sequence that needs to be present in the genomic sequences for a CRISPR-nuclease to be activated. For Cas9 this is commonly the nucleotides "NGG".
Indel	insertion OR deletion mutation *
Complex indel	combination of insertion AND deletion *

Genetics

Wild type, wt	Typical form of a species, regarding phenotype and genotype. The reference genome can be considered the wild-type genotype (e.g., Ensembl WBcel235).
Genotype	The variants or alleles at a particular genomic sequence. Also: complete genetic information of an individual.
P0, F1, F2, F3, …	Generations. Parents "Parental" and progeny "first Filial", "second Filial", …
Reduction-of-function, hypomorphic	Partial loss of gene function (as opposed to complete loss in a "*loss-of-function*"). Coding mutations that impair protein function or non-coding mutations that reduce expression.
Gain-of-function, hypermorphic	Increase in protein activity or expression level. Coding mutations that increase protein function or non-coding mutations that increase expression.
Suppressor- or enhancer mutation	A suppressor mutation "suppresses"-, while an enhancer mutation "enhances" the effect of another mutation. This genetic interaction can occur within a gene (intragenic) or across the genome (extragenic) and can be useful to understand underlying biological connections and mechanisms.
Allele	Specific variant of a gene.
Heterozygous	Having two different alleles of a gene (e.g., one wt- and one mutated allele). Homozygous: having two of the same alleles of a gene.
xyz-1	Gene.
XYZ-1	Protein.
xyz-1(raj123)	A specific allele that differs from the wild-type sequence (e.g., an indel mutation). "*raj*" indicates the lab in which the allele was generated.
xyz-1(blue)	Recognizable in-text description for a specific allele.

Extended Background

Gene regulatory sequences and gene expression

Gene expression

Protein coding genes are transcribed by Polymerase II (Pol II), introns are removed by splicing, 5′ ends are modified by a 7-methylguanosine ("capped"), and 3′ ends are modified by stretches of adenosines ("polyadenylated"). The resulting messenger RNA (mRNA) is exported to the cytoplasm, where it is translated into proteins. It can be modified, degraded, stabilized, and localized at various steps of this process. Altogether these steps determine the amount, place, and activity of the produced proteins. This regulatory process depends on non-coding DNA and is crucial for cell function and animal phenotype (see Extended Background Figure 1A-D). Some very rough median numbers for human cells: transcription progresses at ~60 nts/s, splicing takes 5–10 min., translation takes ~5 aa/s; half-life is often between 5–25 hrs for mRNA and 10–70 hrs for protein; a 3000 μm^3 cell contains in total roughly $10^{5–6}$ mRNAs and 10^9 proteins (thus roughly 1000 proteins/mRNA)[222]. Individual genes can be expressed at ranges of $10^{-1–3}$ mRNA molecules and $10^{0–7}$ protein molecules per cell[223].

The genomes of multicellular organisms encode protein sequences (coding DNA sequence – CDS). But around 75 or 97.5 % of the *C. elegans* and human genome is not encoding protein sequence (and thus "non-coding DNA")[224]. Parts of this function as gene regulatory regions, typically stretches of a few hundred nucleotides, sometimes up to a few kilobases. These can contain sequences that increase, reduce, or localize gene expression, thus called "gene regulatory sequences" (also called cis-regulatory modules – CRMs). These in turn consist of one or more cis-regulatory sequence elements of about 5–15 nt. These sequence elements are binding sites for *trans*-factors such as transcription-factors (TFs), which bind to regulatory DNA sequences that determine the transcriptional activity of genes. In metazoan animals most mature mRNAs also contain regulatory regions, flanking the coding DNA sequence (CDS) on both sides. These 5′ and 3′ untranslated regions (UTRs) are part of the single-stranded mRNA molecule and therefore interact with a different set of *trans*-factors, such as RNA-binding proteins (RBPs) and microRNAs (miRNAs). The following paragraphs define regulatory regions, focusing on their sequence properties. As *trans*-factors are also genes, the genome can encode complex transcriptional states and transitions between them, with genetic feedback loops, circuits, and switches[1,69].

Promoters

The core promoter, +/- 50 bp of the transcription start site (TSS), carries sequences bound by general transcription factors and supports the assembly of the Pol II pre-initiation complex [2]. Examples for such sequence elements are the TATA-box (TATAWA motif), initiator sequence INR, or downstream promoter element DPE[2,225]. Their sequence motifs are usually 5-7 bp long. Core promoters are sufficient for transcription initiation, but have low basal activity without activity from other regulatory sequences[2]. The "proximal promoter", ~250 bp upstream of the TSS, may contain such additional regulatory sequences, which are usually binding sites for development-, cell type-, or tissue-specific transcription factors; this proximal promoter thus functions like an enhancer/silencer[2].

Enhancers and silencers

Enhancers and silencers can be found up to 10 kb (worm) or 1,000 kb (=1 Mb) (human) from the regulated gene, but also very close, like in introns. They typically consist of ~100 – 1,000 bp of contiguous sequence, with several TF binding sites that together determine the regulatory output[4,32,34,35].

Enhancers activate or elevate transcriptional output, likely often by increasing transcription initiation[226]. They can be densely covered with functional binding sites *in vivo*, for example with 20 binding sites in a 300 bp enhancer[23]. Enhancers often regulate their closest gene and data on chromatin contacts improves enhancer-target predictions because chromatin looping can bring enhancer DNA close to the promoter DNA[227–229]. On average a gene is regulated by more than one enhancer. Different enhancers and promoters are usually compatible and activate genes multiplicatively, but some context specific restrictions occur[230]. Silencers abolish or weaken transcriptional output[231]. Enhancer and silencer sequences can overlap, with activity depending on cellular context[232].

Introns and coding sequence

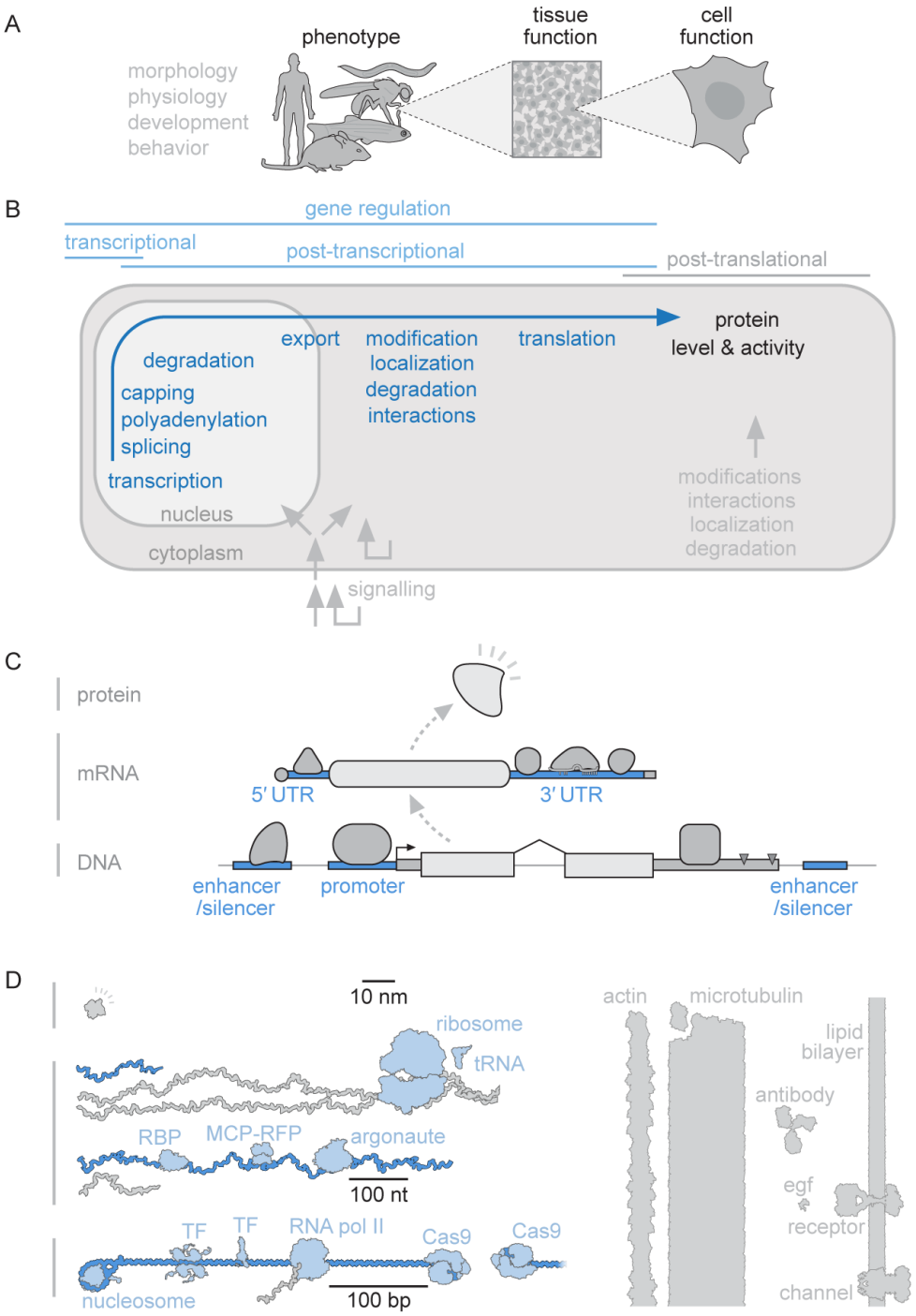
Introns can contain enhancers and silencers but more importantly regulate splicing dynamics and alternative splicing. Protein coding DNA sequence (CDS) can also regulate gene expression. It can affect mRNA stability[233–237]. It has also been shown to direct subcellular localization of mRNA in some cases[238–240].

5′ UTRs and 3′ UTRs

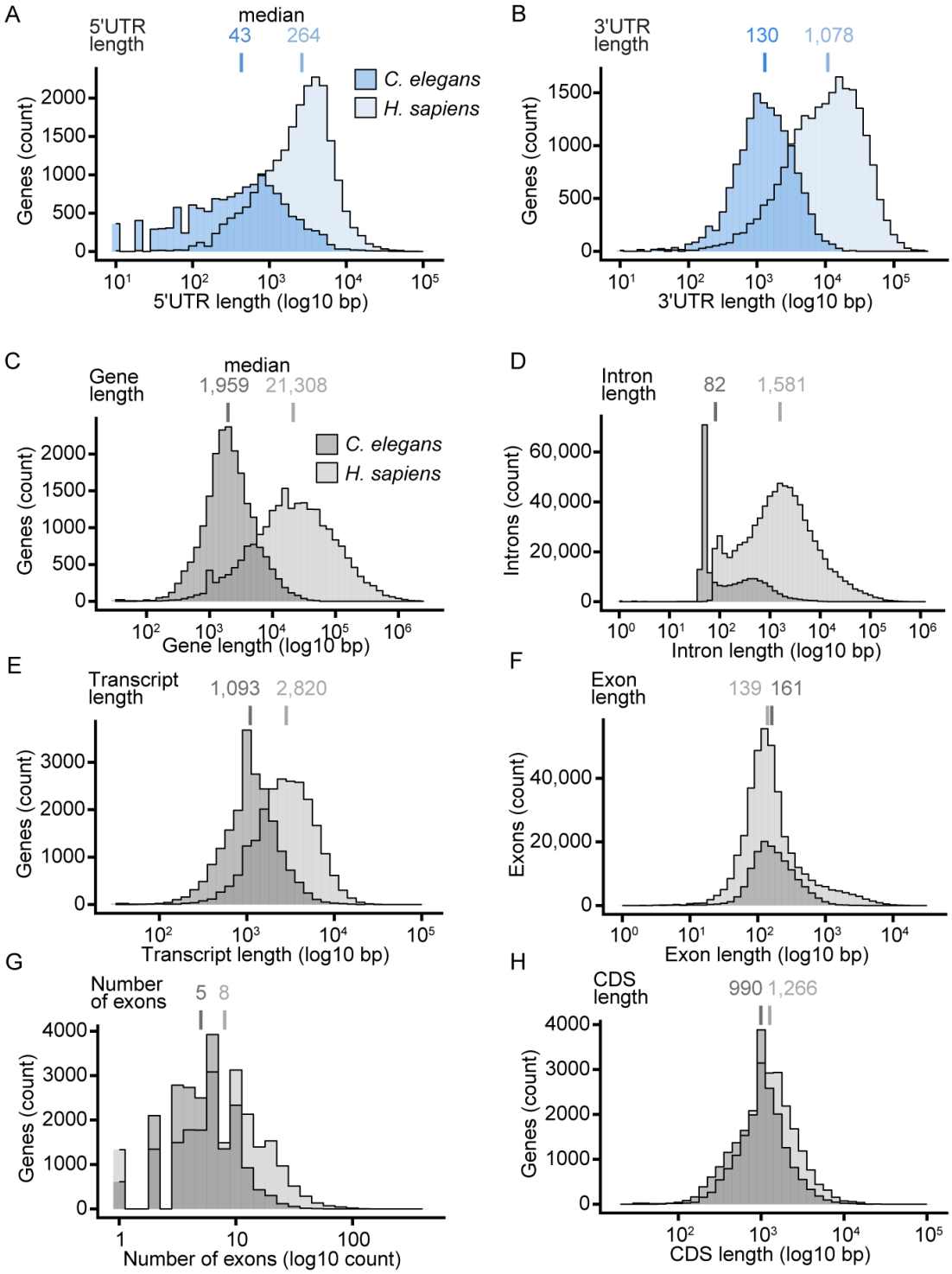
mRNA untranslated regions are preceding (5′ UTR) or succeeding (3′ UTR) the protein coding sequence. UTRs can regulate all steps of gene expression; the levels, localization and interactions of mRNA and its translated proteins[5,6].

5′ UTRs have a median length of ~40–260 bp (*C. elegans – H. sapiens*). They can contain several species-specific nucleotides surrounding the start codon that promote translation initiation, called Kozak consensus sequence. Additionally, 5′ UTRs might contain binding sites for regulatory factors or regulatory small open reading frames (sORFs). While the beginning of 5′ UTRs in human is defined by the transcription start site, in *C. elegans* trans-splicing adds a 21–23 nt constant RNA spliced leader at the UUuCAG motif, while the upstream sequence is removed at around 70% of genes[241]. Trans-splicing acts in operons to process primary transcripts into separate mRNAs. Around 16% of *C. elegans* genes are in operons with 2 (~8%) or 3–8 genes (~8%)[241].

3′ UTRs in human can be several thousand bp long, with a median length of over 1,000 bp[6]. *C. elegans* 3′ UTRs are shorter with a median length of around ~130-150 bp[242] (see Extended Background Figure 2B). Genes usually have relatively clearly defined 3′ UTR isoforms, but their selection can be regulated. Around 60% of *C. elegans* genes encode a single-, ~30% encode two-, 8% three-, 2% encode four or more 3′ UTR isoforms[243]. One element found in most 3′ UTRs is the canonical polyadenylation signal AATAAA (AAUAAA in RNA) that determines the position of cleavage and polyadenylation, and thus the 3′ UTR end. 60% of worm 3′ UTRs are defined by this AATAAA, the remaining 40% by slight variations (e.g., AATGAA, TATAAA, CATAAA, …)[242]. Apart from this, 3′ UTRs might contain one or multiple important regulatory elements[6]. These can be binding sites for RBPs, miRNAs, or other ncRNAs, as well as structural elements like riboswitches. 3′ UTRs are generally AT rich (~70% AT)[243].



Extended Background Figure 1. Gene expression and gene regulatory regions
(A) Animal phenotype depends on tissue- and cell function. **(B)** Hierarchy and subcellular location of gene regulatory processes. **(C)** Scheme of a gene with gene regulatory regions marked in blue. **(D)** DNA, mRNA, and proteins at scale. Actin mRNA transcript is shown with different sections (5′UTR, CDS, 3′UTR, polyA) true to length. One actin protein shown on the top-left in grey. Gene regulatory regions marked in blue. Composition based on figures from refs.[222,244,245].



Extended Background Figure 2. Features of protein coding transcripts, *C. elegans* and *H. sapiens*
Plots based on protein coding transcripts from the Ensembl Transcriptomes (WBcel235, GRCh38.p13; retrieved 30.11.2021). Medians indicated on the top. **(A)** 5′UTR length. Longest per gene, 0 bp features removed. **(B)** 3′UTR length. Longest per gene, 0 bp features removed. **(C)** Gene length. **(D)** Intron length. All transcript isoforms considered. **(E)** Transcript length. Longest per gene. **(F)** Exon length. All transcript isoforms considered. **(G)** Number of exons. From transcripts with most exons per gene. **(H)** CDS length. Longest per gene.

Nucleotide contacts of regulatory factors

Transcription factors

Transcription factors (TFs) bind DNA of promoter, enhancer, or silencer regions. The human and worm genomes encode roughly 1,600 and 900 TFs[3,246]. Amino acid side chains can make direct hydrogen bonds, water-mediated hydrogen bonds, and hydrophobic contacts with nucleotide edges. This allows nucleotide-specific (A/C/T/G) pairing in the major- and base-pair-specific (W/S) pairing in the minor groove of dsDNA[247]. Except for zinc finger TFs which usually contain several homotypic repeats, ~95% TFs of the remaining families contain one DNA binding domain[3]. Eukaryotic transcription factors contact dsDNA along 5 – 15 bp with an average length of around 10 bp (estimated by ref.[248]) (see also Extended Background Figure 3).

RNA binding proteins

RNA-binding proteins (RBPs) bind RNA. Human and worm genomes encode roughly 1,500-1,900 and 600-900 RBPs[9,249,250]. Roughly half of human RBPs bind to mRNAs, while the rest bind non-coding RNAs[9]. ~99% of RBPs bind ssRNA and their secondary structures[224]. Even the same domain can contact different RNA sequences and structures (see Extended Background Figure 3). For example, one RRM domain often recognizes 2-8 nt of RNA by Lysine/Arginine salt bridges and aromatic amino acid nucleotide stacking[8,251]. KH domains bind mainly by hydrogen-, electrostatic bonding and shape-complementarity to recognize ~5 nt, often expanded to ~10 nt by a second KH domain[8,251]. The exceptionally uniform structure of the Pumilio domain allows programmable binding to 8-10 nt[252,253]. More than half of RBPs contain multiple RNA binding domains[9], domain linkers can also contact RNA[8,251], and RNA secondary structure is often important for binding. Therefore, the actual RNA sequence involved can be longer than 5-15 nt and may include uncontacted but structural important RNA.

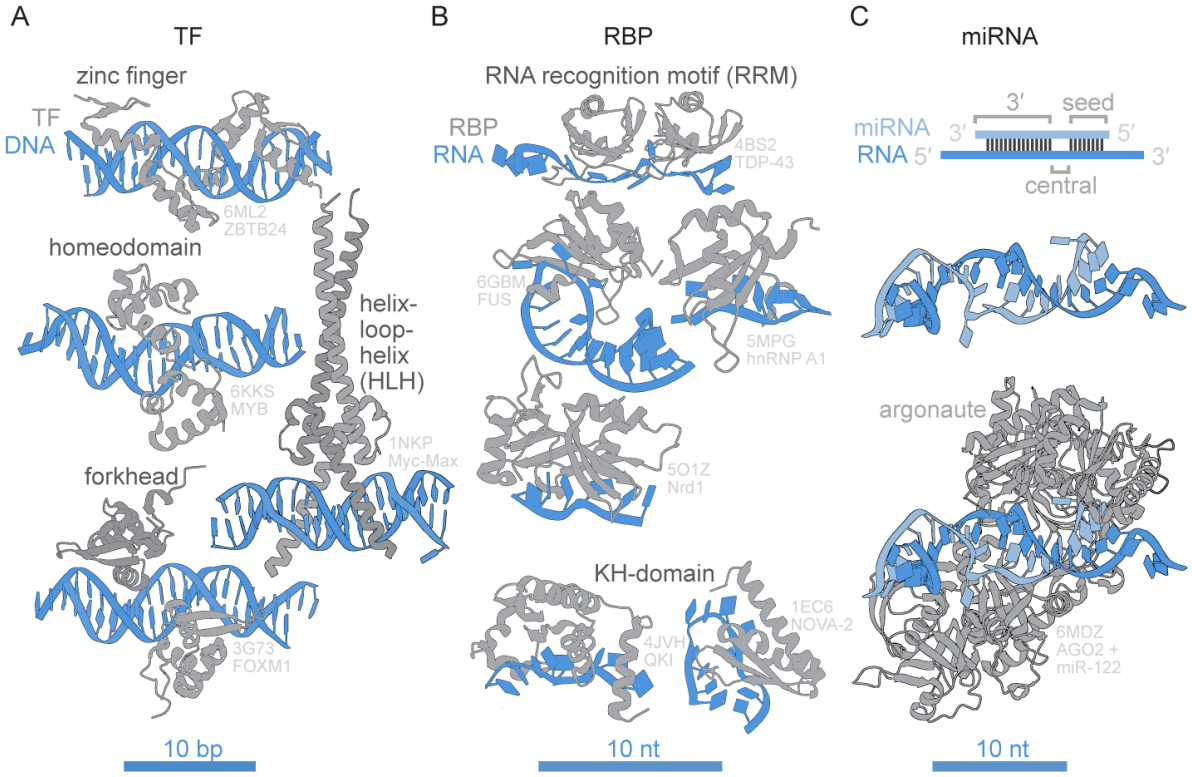
Non-coding RNAs

Gene regulatory non-coding RNAs are diverse. Small 20-30 nt RNAs like microRNAs, siRNAs, and piRNAs typically repress mRNA post-transcriptionally by 5-30 nt base-pairing and defined pathways. Long non-coding RNAs can target RNA and DNA by diverse mechanisms[254,255]. Circular RNAs can indirectly affect gene expression by sponging microRNAs[256–258] and biogenesis can affect host mRNA processing and expression[259,260]. Other ncRNAs like tRNAs, rRNAs, snRNAs and snoRNAs are mostly serving constitutive functions in translation and splicing but can also act regulatory in some cases.

MicroRNAs

MicroRNAs (miRNAs) are conserved ~22 nucleotide RNAs that bind mRNAs and repress target genes post-transcriptionally[7,10,261,262]. Human and worm genomes encode roughly 556 and 145 miRNAs[263–265], or 1,917 and 253 miRNAs under less strict criteria (miRbase release 22.1, Oct. 2018)[266]. MiRNAs repress genes by mRNA degradation as well as translational repression[7,262]. Their half-life can range from 2 to >48 hrs, with a median half-life of 10 – 30 hrs in different cell lines[267,268]. Active degradation by target-dependent miRNA degradation (TDMD) can be triggered by a specific architecture of extensive miRNA-target pairing[139–141]. MiRNAs and their binding sites can be divided into three architectural regions, the seed region (nts 2-8), the central region (nts 9-11), and the 3′ region (nts 12-22) (see Extended Background Figure 3). Usually six to eight seed nucleotides are base pairing, with more base-pairs and a target adenosine downstream of the seed resulting in stronger repression [7]. Additionally, miRNAs can show extended 3′ region pairing. Here, the central region (nts 9-11) forms a loop with unpaired miRNA- and target nucleotides, followed by paired nucleotides, often with essential- and intermediate function of nucleotides 11-13 and 14-16[269,270]. With such additional base-pairing the

seed can tolerate 1 nt bulges, mismatches, or GU-pairing. MiRNAs with identical seeds are grouped into families and differences in their 3′ regions confer differential target specificity[113,138]. Genome-wide experiments have found that around 40-80% of Argonaute-miRNA-mRNA interactions show extended 3′ pairing[138,271–273]. Furthermore, many physiologically important miRNA binding sites show 3′ pairing, for example the *lin-4*, *let-7*, and *lsy-6* miRNA binding sites in the *lin-14*, *lin-41*, and *cog-1* 3′ UTRs of *C. elegans*[107,108,274–277]. Whether such sites with extended 3′ pairing also trigger miRNA degradation by TDMD, in addition to their gene regulatory role, has not been shown.



Extended Background Figure 3. Nucleotide contacts of gene regulatory factors
Contacts of regulatory factors with their bound target molecule (DNA or RNA) in dark blue. **(A)** Transcription factor domains. **(B)** RNA binding protein domains. **(C)** MicroRNA and argonaute.
PDB IDs in light grey. Modified from RCSB PDB (rcsb.org), original structures from refs.[278–288].

Complexity of gene regulatory sequences and phenotype

Biochemical variables affecting sequence activity

Binding of gene regulatory factors and the regulatory output is affected by several biochemical variables (see Extended Background Figure 4A)[4,33]. Some of these can be recapitulated with extrachromosomal reporters[289]. However, most depend on the native genomic sequence context[235,290], cell type, and development.

Definition of binding sites

A position weight matrix of the consensus "motif" of a binding site can be constructed using probabilistic models (Extended Background Figure 4B)[247,291]. However, their use to identify functional sequences can be limited. Interdependency of nucleotides or wider sequence context is not included, and motifs are often shorter (3-8 bp) than contacted nucleotides *in vivo* (10 bp and more). Also, factors can bind many related sequences with similar affinity[292–294], and biochemical variables influence binding (e.g., cofactors, chemical modifications, structure, sequence context)[295–297]. Therefore, flexible nucleotides of "fuzzy" or "degenerate" motifs may still be sensitive to mutations at individual sequences. Furthermore, functional sequences *in vivo* can diverge from optimal binding site motifs and have low, suboptimal affinity with specificity contributed by multiple binding sites and their relative arrangement[28–31]. Although functional nucleotides have a higher genome-wide average conservation compared to non-functional nucleotides, such analysis can have limited utility at individual sequences where conservation sometimes shows little correlation to function[23–27].

Arrangement and logical interaction of binding sites

Binding sites and their arrangement within a regulatory unit determine the interpretation of gene regulatory information (Extended Background Figure 4C)[4,32,33,35]. Gene regulatory output can depend on these sequence characteristics to very different degrees, ranging from a flexible to a very rigid identity and arrangement of binding sites (e.g., "Billboard", "Enhanceosome" models)[4,32,34]. These aspects are mainly studied for enhancer DNA and might be different for RNA, which has a structurally more flexible backbone. For RNA, the molecular abundance (copy number) is an important parameter to consider, because it affects regulator-target ratio, which is crucial for regulatory outcome[40]. Binding sites can generally also influence each other's activity, either by direct biochemical interaction (e.g., binding sites that allow factors to interact physically) or by indirect logical interaction (e.g., through feedback loops, redundancy, or competition for *trans*-factors) (Extended Background Figure 4D)[4,39,298,299]. Quantitative measurements are necessary to determine the type of interaction and competitive interactions will necessarily occur at certain concentrations of binding site and regulator[40,300].

Dependency on cell type and development

Specialized signaling pathways and gene regulatory networks have evolved in parallel to cell specification and developmental processes[1,10,11,152,301,302]. Activity of gene regulatory sequences is often specific to tissue- and cell identity (Extended Background Figure 4E)[303–306]. Spatial identity is established by signaling pathways and gene regulatory factors, often by spatially co-occurring activity (spatial intersection), that can also be separated temporally along the developmental cell lineage (temporal intersection)[307,308]. Together, spatial-, tissue-, cell type- and terminal- "selectors" determine the final cell identity[309–311]. Consequently, TF repertoire can be used to define cell identity[312–314]. Besides depending on *space*, activity of gene regulatory sequences also depends on *time* (Extended Background Figure 4E). Most dynamic processes result in dynamic activity of gene regulatory factors and their

target sequences (e.g., cell differentiation, embryogenesis, regeneration, reproduction, feeding, circadian oscillations, or mechano-transduction)[1,315–323].

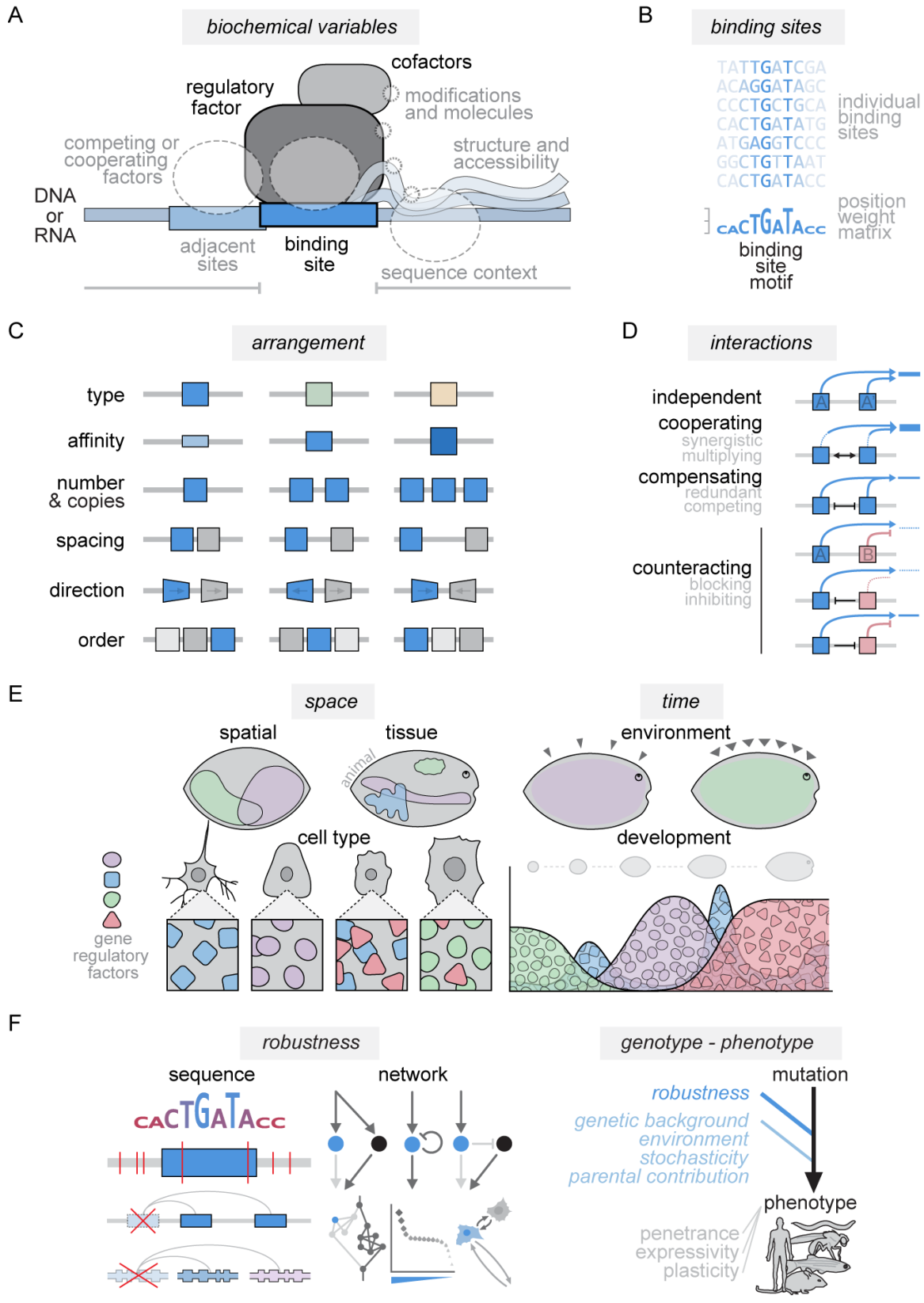
Phenotype

Phenotype describes the combined properties of an organism, which includes for example morphology, physiology, developmental processes, and behavior. There can be variability in the proportion of genetically identical individuals that show a phenotype ("penetrance") and the degree to which a phenotype is displayed by an individual ("expressivity"). Also, the capacity to change phenotype in response to the environment ("plasticity") may vary.

Translation of gene regulatory genotype to phenotype

Regulatory sequences can be robust to mutations due to four properties: 1.) Mutations within a binding site might not affect the interaction with the regulatory factor. 2.) Surrounding sequences may be mutated without consequence. 3.) Redundant or compensatory binding sites may replace functions, for example within enhancers[36,38] or UTRs[37,39]. Several low-affinity sites for the same factor can increase overall robustness[28,29,31]. 4.) The loss of a whole region can be compensated by another region (e.g., shadow enhancers) (Extended Background Figure 4F)[298,324,325].

Beyond this, biological systems also show robustness of phenotype because of higher-order network properties[68,326]. Fluctuations or loss of gene expression can be compensated within gene networks, for example by paralog gene redundancy, rewiring of redundant network modules[68,326], autoregulatory feedback loops[69,327], and transcriptional compensation[67,328]. Furthermore, non-linearity can result in phenotype robustness within a wide range of gene dose (e.g., half the dose of a recessive allele does not alter phenotype)[68]. At the physiological level, additional robustness can be provided by cell type plasticity and signaling feedback between cells or tissues. Even more factors influence the final phenotype, but can be usually controlled in the lab: environment and genetic background[66,71], parental contribution and transgenerational inheritance[63,329], and stochasticity of gene expression and development (Extended Background Figure 4F)[64,65,70,330,331].



Extended Background Figure 4. Complexities of gene regulation

(A) Variables that can act at the biochemical level to determine contacts of regulator factors with their targets and its regulatory outcome. **(B)** Individual binding sites can be used to determine a binding site motif. **(C)** Arrangement of binding sites can affect their ability to interact logically. **(D)** Types of interactions between binding sites. **(E)** Examples of spatial and temporal depending on activity of gene regulatory factors. **(F)** Robustness at sequence and network level (see text) affects how gene regulatory mutations may affect phenotype.

Genetic methods to analyze gene regulatory sequences

Parallel and high throughput

Parallel high-throughput methods in cell lines usually follow a similar design: cells are altered in parallel by large-scale delivery (e.g., virus or transfection), cells are binned or selected by expression or fitness, and finally both parameters are linked by deep sequencing[41–43] (see Extended Background Figure 5A). Despite their potential to systematically map gene regulatory sequences to animal phenotype, few comparable approaches have been demonstrated *in vivo* yet.

Massively parallel reporter assays (MPRAs): can produce quantitative sequence-expression relationships for millions of sequences in cell lines, yeast, or bacteria[41–44]. MPRAs can identify functional sequences genome-wide and recently has produced further insights into gene regulatory logic when combined with machine learning[332–337]. Still, very few related approaches have been achieved in animal models and these are usually limited to specific organs, developmental stages, and low numbers of tested variants[338] (Supplemental Table 1).

CRISPR tiling screens: identify functional sequences in genomic regions up to the kilobase scale, but only realized in cell lines. Studies have focused on identifying enhancer regions with parallel delivery of different sgRNAs and Cas9[48,49,51–55] or dCas9-KRAB repressor (CRISPRi)[50]. Identity of sgRNAs is used to infer perturbed sequences and genomic DNA is only sequenced for selected samples. In a related method, 10-20 sgRNAs can be multiplexed to create diverse deletions at a sequence of interest. Mutations in 3′ UTRs can then be associated with mRNA levels using targeted mRNA sequencing[56–58].

Endogenous variant testing: these can assign functions to hundreds and thousands of programmed variants, but have focused on protein coding sequences, and are so far limited to cell lines. Edits can be made in parallel with ss- or dsDNA template libraries and Cas9[61,339,340], template libraries produced *in situ* in prime editing[60,182], or random point mutations by hyperactive deaminases[176–178,180].

Classic techniques in animals

Gene regulation can be studied across cell types, tissues, and development using animal models that also display complex phenotypes. Classic model systems are listed in Supplemental Table 3. Several genetic methods to perform perturbations *in vivo* are established (Extended Background Figure 5B). Nevertheless, parallelized high-throughput methods are still missing, and experiments are often limited in throughput by microinjection, culturing capacities, and manual scoring steps.

Systematic reporter deletions: regulatory activity of distinct sub-fragments is analyzed using reporter plasmids (“bashing”) (Supplemental Figure 1). Manual microinjection and genotyping limits the number of examined mutations per study to ~10 – 100, with few exceptions (Supplemental Table 1 and Supplemental Table 5). Phenotype is not examined, and overexpression limits sensitivity to endogenous regulatory mechanisms.

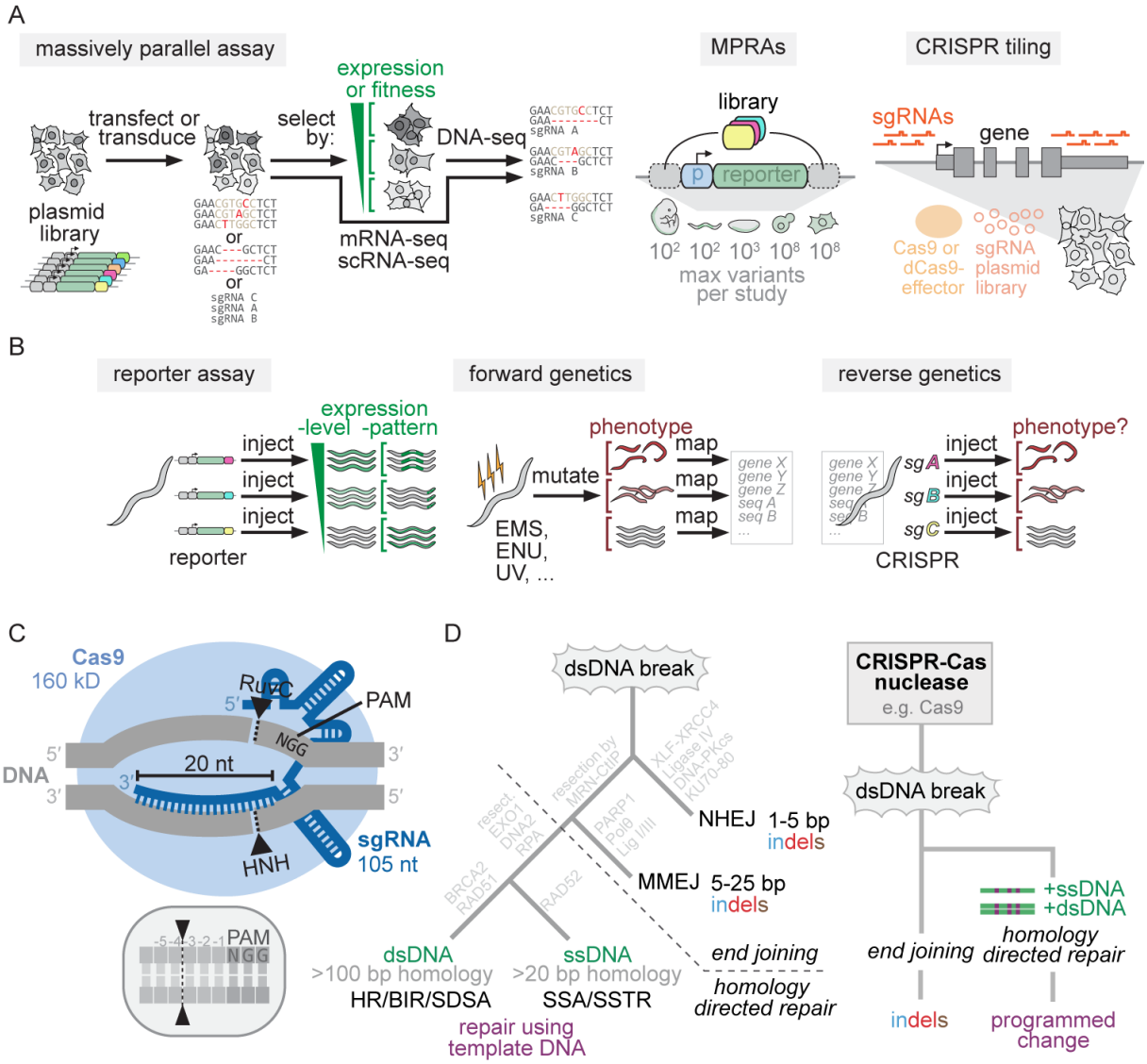
Forward genetics: organisms are subjected to random mutagenesis, individuals with altered phenotype are isolated, and the genetic lesion is determined. Most common mutagens are EMS or ENU that introduce predominantly point mutations[341,342]. This approach is cumbersome and

regulatory alleles are rarely isolated. In *C. elegans*, only 35 non-coding alleles have been described in the last ~40 years[82].

Reverse genetics: a specific gene or sequence is altered, and the phenotype examined. Until recently, most studies were testing gene functions. With genome editing, regulatory sequences can be studied. However, manual microinjection and genotyping limits the number of examined mutations per study to ~10 – 100 (Supplemental Table 2).

Genome editing with CRISPR-Cas

Genome editing can test regulatory sequences in their native context and analyze the impact on phenotype. It became widely accessible with the discovery and engineering of RNA-guided CRISPR-Cas nucleases[45,46,343–345]. Cas9 from the bacterium *Streptococcus pyogenes* (SpyCas9 or SpCas9), forms a ribonucleoprotein complex with the target-specific crRNA and structural tracrRNA, that can be fused into a "single guide RNA" (sgRNA)[45]. The target-specific 20 nt "protospacer", base-pairs with the target DNA sequence to define the position of nuclease cleavage to create a double-strand DNA break. This also depends on the presence of a genomic "protospacer adjacent motif" (PAM) -sequence of NGG[45] (Extended Background Figure 5C). Breaks in genomic dsDNA are repaired by different cellular pathways that result in insertions or deletions (indels), combinations thereof, or perfect repair. Cas9 can thus be applied to introduce indel mutations or precise edits[47] (Extended Background Figure 5D). Most indels occur within 5 bp of the dsDNA break and the majority is between 1 – 25 bp long[346]. Although the precise sequence outcome of indels is not deterministic, it depends on the surrounding sequence, which makes some outcomes more likely than others, and thus "nonrandom"[347]. This enables indel outcome predictions based on sequences surrounding the sgRNA cut sites, in *in vitro* systems where indel outcomes can be measured in high-throughput (HEK293, K562, mESC, U2OS, primary T-cells etc.)[123–127,348–350]. Pairs of sgRNAs can be used to create larger deletions in the range of 10s to 1,000s bp. For precise programmed edits, DNA templates with the desired changes are added, which are then used in different repair pathways, the most common being homology-dependent repair (HDR). A multitude of other CRISPR-based tools have been discovered and engineered, for example nucleases with different size or PAM properties, base- or prime editors to edit sequences without dsDNA breaks, CRISPRi or CRISPRa to modify chromatin and transcription, or proteins to target RNA[45,345].



Extended Background Figure 5. Genetic methods to study gene regulatory sequences
(A) High throughput approaches possible in cell lines. **(B)** Classical approaches used in model organisms. **(C)** Diagram of CRISPR-Cas9 with sgRNA and location of dsDNA break. **(D)** Main repair pathways of dsDNA breaks with typical outcomes[47] and application of CRISPR-nucleases to introduce indel mutations or programmed changes.

The model organism *Caenorhabditis elegans*

Life, development, tissues, cells, genome

The roundworm *C. elegans* is useful for genetic screens of behavior or morphology phenotypes[77,78,351]. Due to its microscopic size of maximal ~1 mm, high fecundity of ~300 progeny per adult, and a rapid life cycle, millions of worms can easily be cultured within days and used for large-scale experiments[73]. *C. elegans* can be phylogenetically classified to the phylum *Nematoda* and its 100 mio bp genome is around 30-times more gene-dense than the human genome[222,352], from which it diverged roughly 800 mio years ago[353]. In nature it lives in soil around rotting vegetable and fruit, where its bacterial food source is abundant - in the laboratory it can be grown at the scale of millions, on agarose plates or in liquid culture, with *E. coli* bacteria[73,77,354]. Useful for genetics, 99.9% of animals are self-fertilizing hermaphrodites, while rare meiotic non-disjunction of the X chromosome produces males, which can be used for genetic crosses[73]. The developmental life cycle from egg to adult takes 3-5 days (at 25-16 °C) and includes embryonic development, several molts and four larval stages (L1-L4) [73] (Extended Background Figure 6A). Adult hermaphrodites have 959 somatic cells organized in different tissues and individual cell types have diverse functions and morphologies (Extended Background Figure 6B). Several thousand scientists, likely more than 5,000, work with *C. elegans* (2,252 participants at the "2021 International *C. elegans* Conference", retrieved 03.11.2021, https://genetics-gsa.org/celegans-2021/whos-attending-2) (1,362 labs registered at the strain repository "CGC", retrieved 03.11.2021, https://cgc.umn.edu/laboratories).

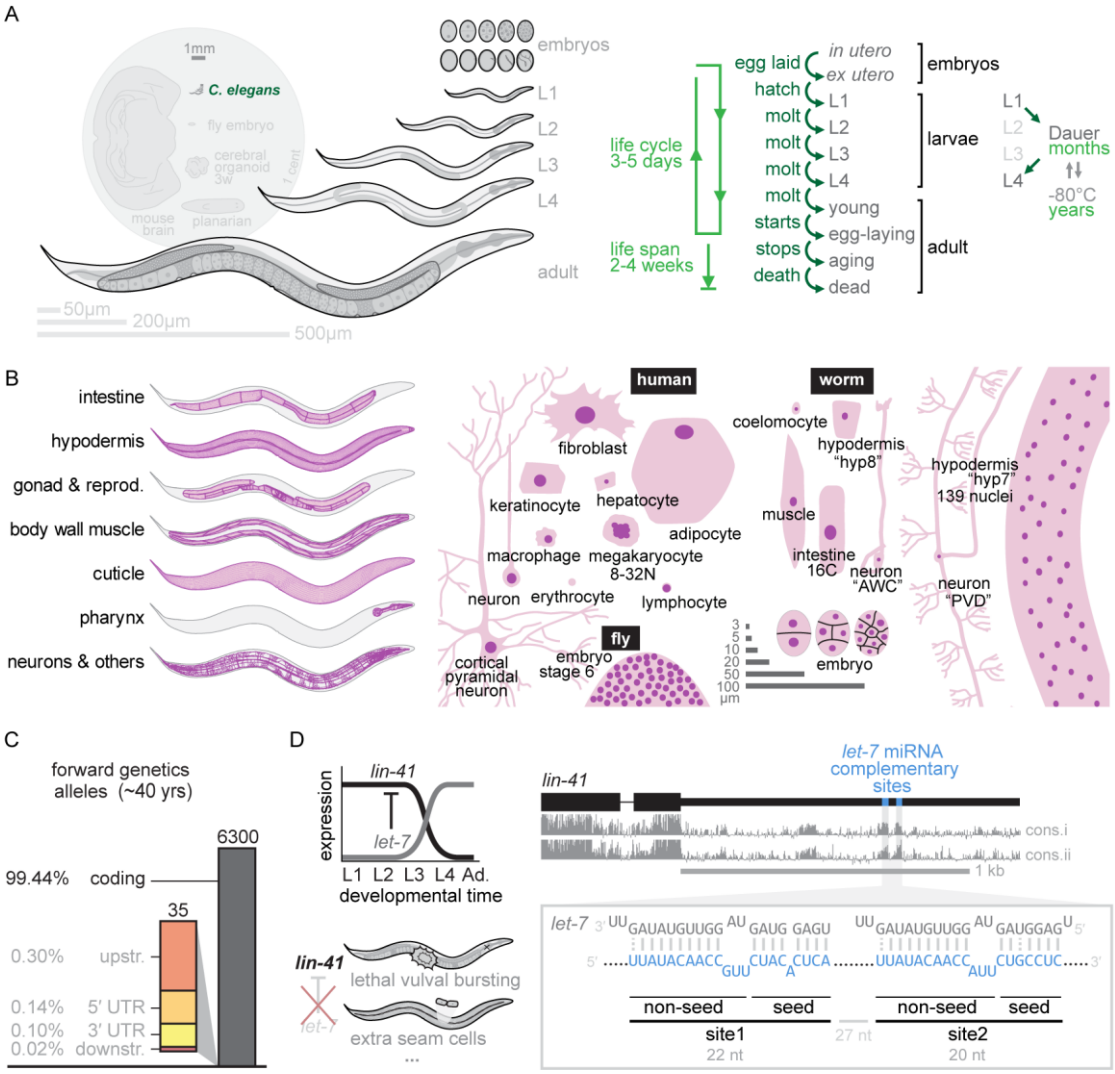
Limitations of available methods

C. elegans has contributed to the discovery of important gene regulatory mechanisms, and advanced methods are available for genome-wide insights and manual or automated phenotyping (Supplemental Table 4). Available methods to test function and phenotypes of gene regulatory sequences are still limited in throughput. Because *C. elegans* is translucent, activity of gene regulatory sequences can be directly tested *in vivo* with fluorescent tags and systematic deletion analysis ("bashing"). However, because of manual microinjection and genotyping, the number of tested sequences is limited to around 10-100 per study (Supplemental Table 5). There are several approaches to study phenotypes of genomic sequences. Forward genetic screens have isolated 6,288 alleles in the last ~40 years (Wormbase.org "classical alleles", retrieved 01.09.2017)[82]. But only 35 alleles (0.56%) of these reside in regulatory sequences (Extended Background Figure 6C) (Supplemental Table 6). This relative low number cannot be solely explained by the genomic fraction of non-coding DNA[355] when considering the probability of nonsense mutations (data not shown). Without data from alternative methods, it is currently hard to estimate to what extent this might be due to regulatory sequence robustness (forward genetics alleles are ~85% point mutations), experimental biases (historic mapping techniques focusing on coding sequence), or few functional regulatory nucleotides. Other methods like population genetics approaches have limited resolution, usually at the scale of several genes, and can be limited by sparse natural variation along individual regulatory sequences[86,356]. Reverse genetics with genome editing is well established in *C. elegans*, but manual microinjection and genotyping still limits the number of tested mutations to around 10-100 per study[81] (Supplemental Table 6). This means that new approaches would be necessary to test function and phenotypes of gene regulatory sequences at larger scales.

Regulation of *lin-41* by the miRNA *let-7*

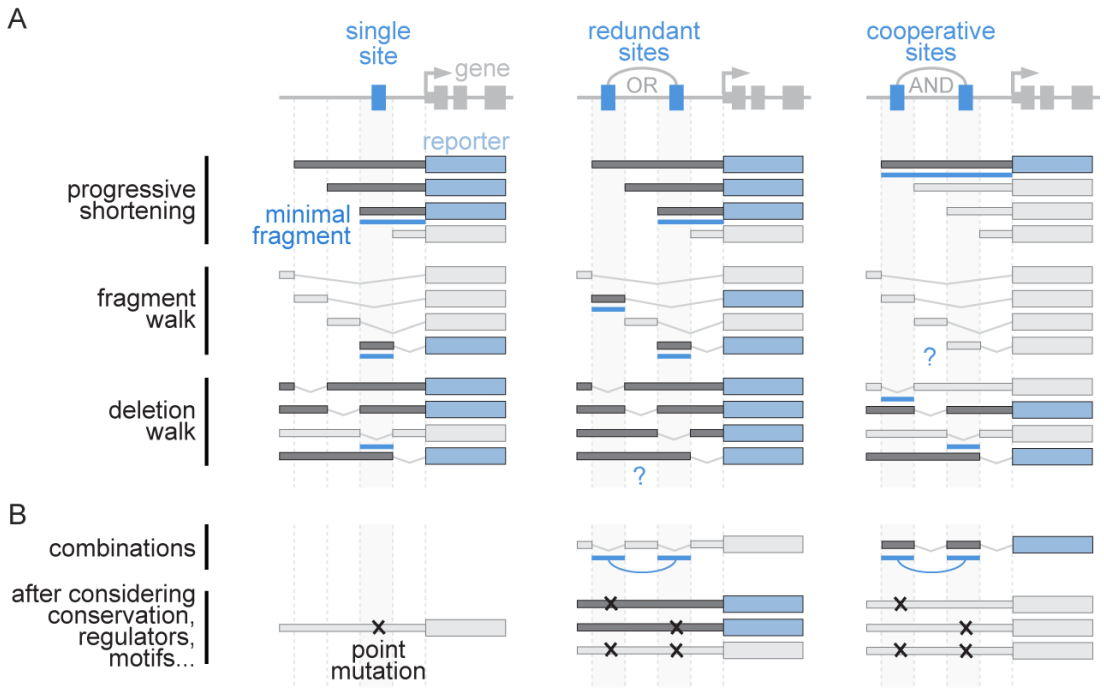
An interesting regulatory unit is contained in the *lin-41* 3′ UTR. It consists of two complementary sequences to the microRNA *let-7,* which repress *lin-41* post-transcriptionally[107–109]. The temporal aspect of this regulation is crucial for development: *Let-7* becomes expressed between L2 and late L3

stage[107,111,357,358]. As a result, LIN-41 protein expression is reduced post-transcriptionally by 3-4 fold, starting around early L3 stage and being completed by late L4 stage[105,108,109,112]. Several phenotypic changes can be observed when the natural *let-7*-mediated regulation is disrupted. Subtle morphogenesis defects and decreased vulva integrity results in the drastic and lethal vulval bursting[106,107]. A retarded terminal differentiation and one additional cell division of seam precursor cells results in two extra seam cells[107,359] (Extended Background Figure 6D). Further defects are also found in the differentiation of sexually dimorphic neurons and the morphogenesis of male tail retraction [360–362]. *Let-7* regulation of LIN-41 during cell differentiation and development is conserved in human and *let-7* exists with the identical seed in most animals[276,357,363]. The two 22 and 20 nt long *let-7* complementary sequences in the *lin-41* 3′ UTR are separated by 27 nt[109] (Extended Background Figure 6D). The extended 3′ complementarity to *let-7* is crucial for function and specificity[113,138,269]. Studies with reporter plasmids have found that each site alone in one, two, or three copies does not show repressive activity[109,133] – although others found that three copies can show wildtype activity[110] – and that the identity of the intervening sequence is important[109,110]. These results, together with the sites proximity, suggested that both *let-7* sites may act cooperatively[109,132]. Although the 1.2 kb 3′ UTR is almost ten times longer than the median *C. elegans* 3′ UTR, no other functional sequences have been identified so far.

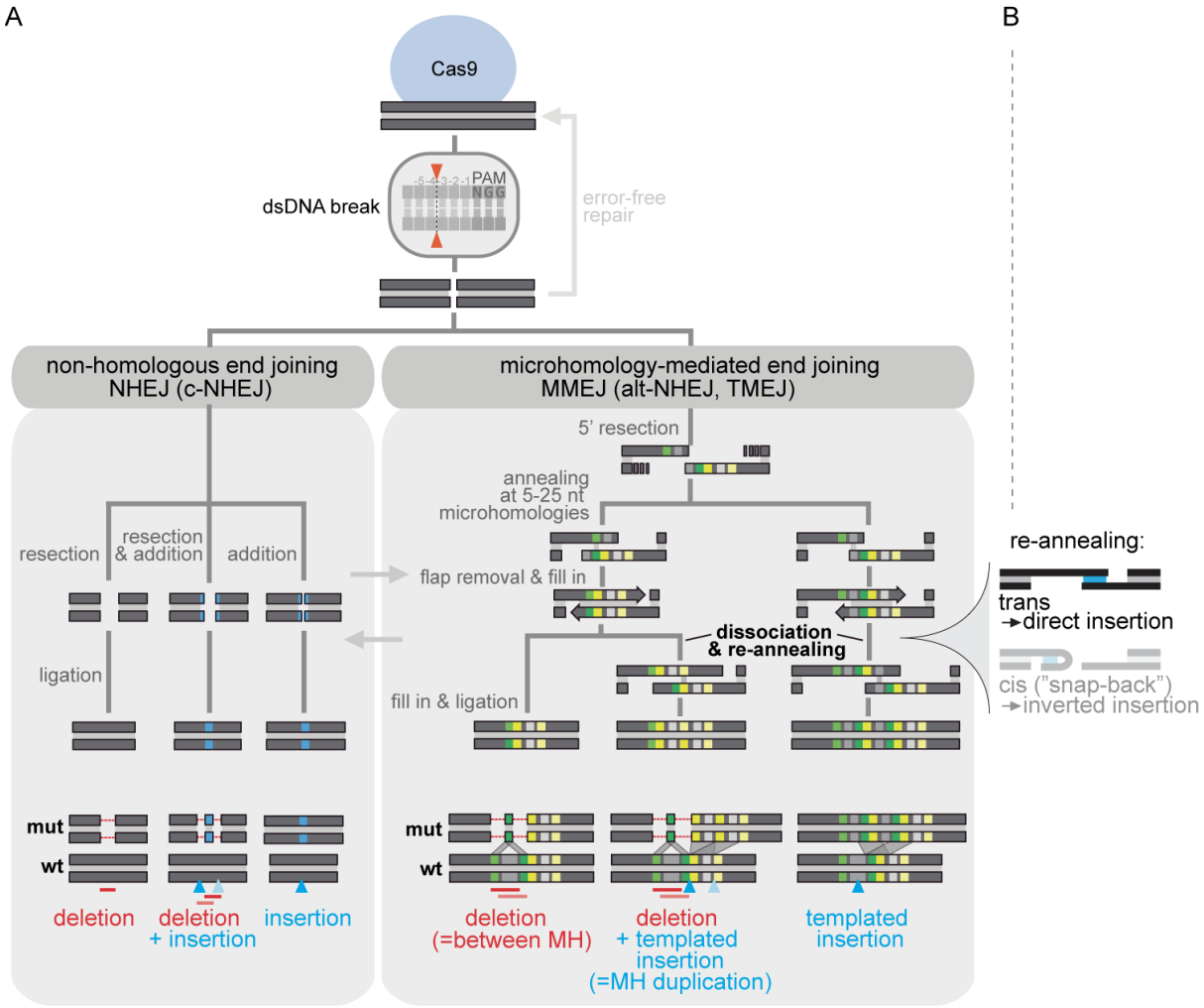


Extended Background Figure 6. Properties of *Caenorhabditis elegans*
(A) Morphology, sizes, and developmental stages. **(B)** All tissues and several typical cells, roughly true to scale[75]. **(C)** Number of alleles with phenotypic defects isolated in classical forward genetics screens, classified by affected genomic regions (Wormbase.org "classical alleles", retrieved 01.09.2017, ref.[82]. **(D)** Regulation of *lin-41* by the miRNA *let-7*.

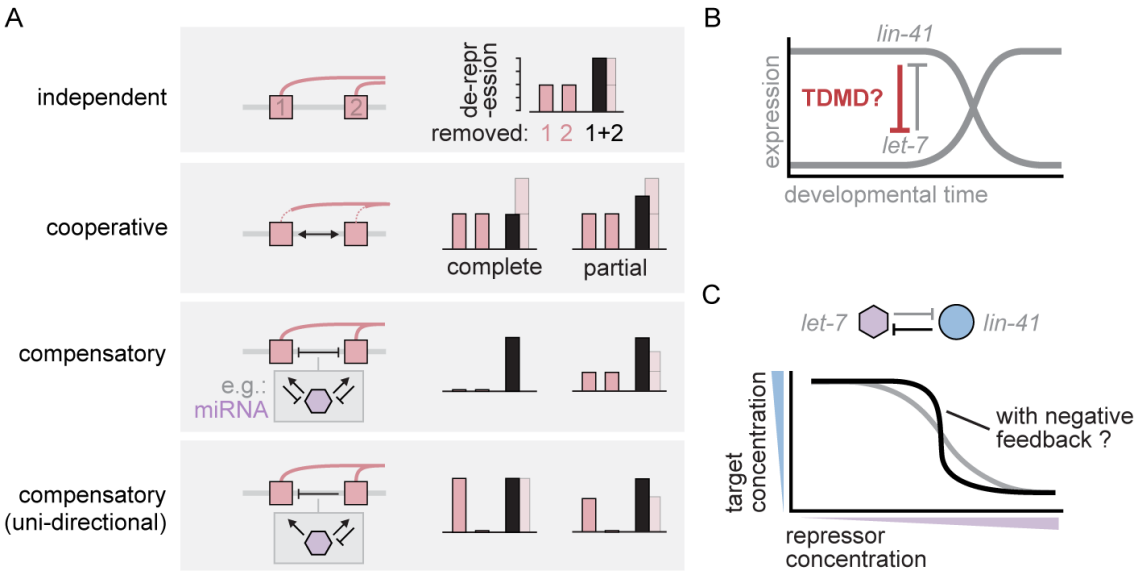
Supplemental Figures and Tables



Supplemental Figure 1. Systematic deletion analysis of regulatory sequences ("bashing").
(A) Three systems of analysis and their ability to detect gene regulatory elements in three scenarios (single site, two sites interacting either redundantly or cooperatively). **(B)** Possible follow-up experiments based on the initial analysis.

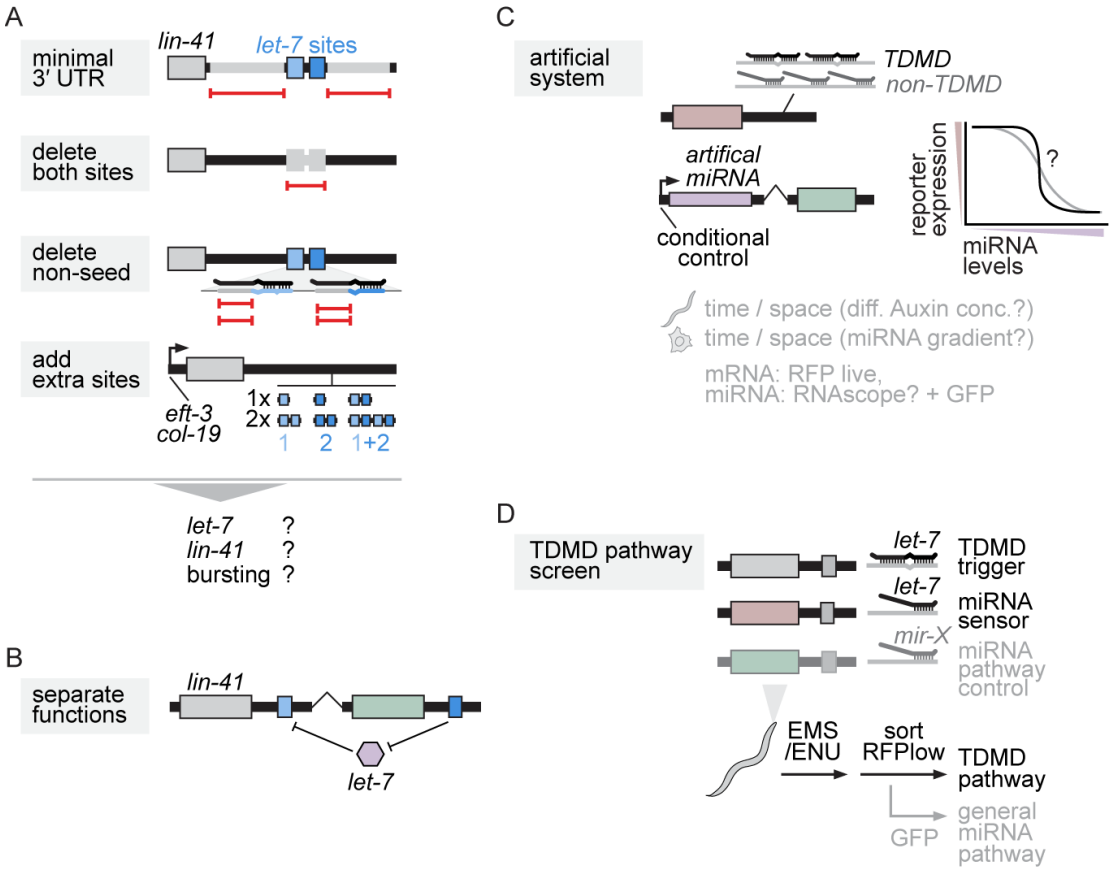


Supplemental Figure 2. End joining DNA repair outcomes and mechanisms
(A) NHEJ and MMEJ pathways. For MMEJ the dissociation & re-annealing step is highlighted that could lead to the templated insertions observed in our data. **(B)** Diagram showing the two possible mechanistic causes of insertions: *trans*-intermolecular, or *cis*-intramolecular synthesis[101,130].



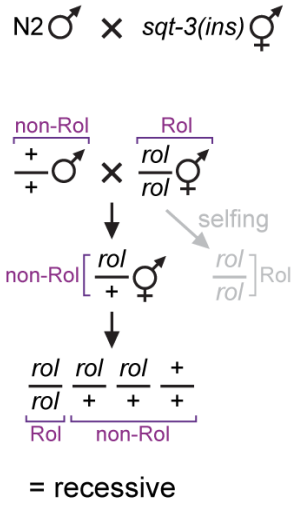
Supplemental Figure 3. Interaction of *let-7* binding sites and possible feedback loop

(A) Definition of different types of interactions between two repressive binding sites. **(B)** Diagram showing the possible repression of *let-7* via TDMD by its target sites. **(C)** Diagram illustrating the possible difference of dynamics that negative feedback could cause.

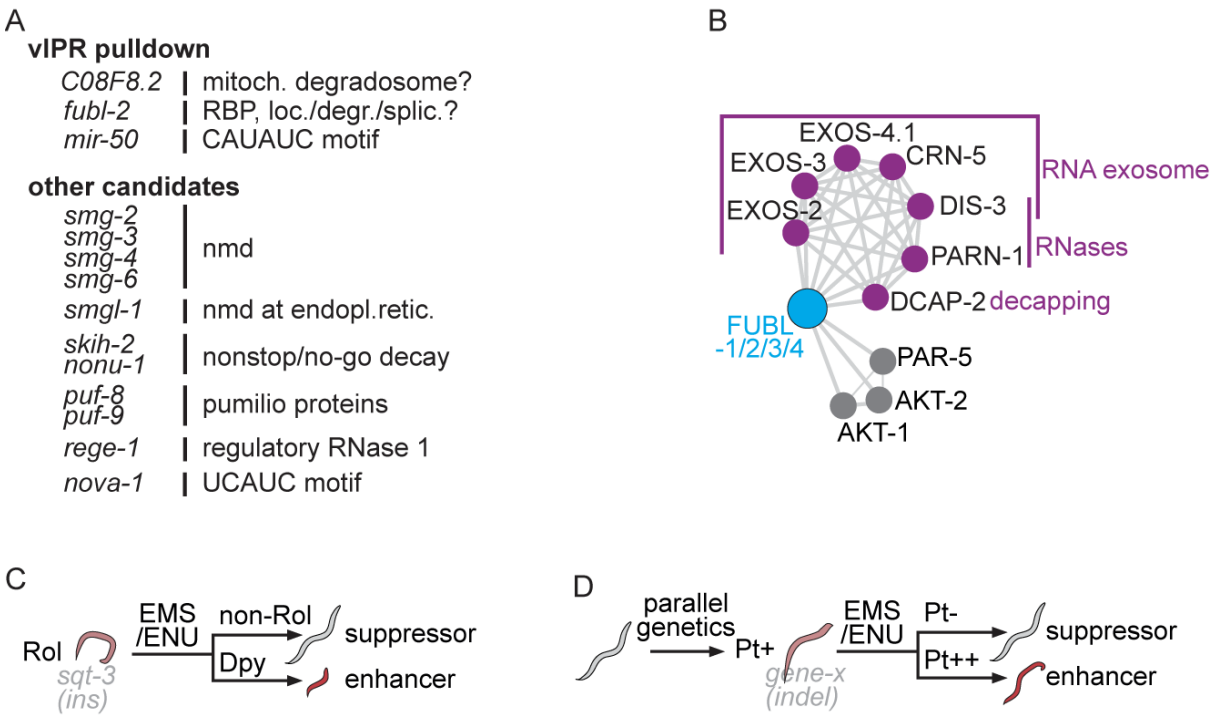


Supplemental Figure 4. Possible experiments to study TDMD at the *let-7* sites

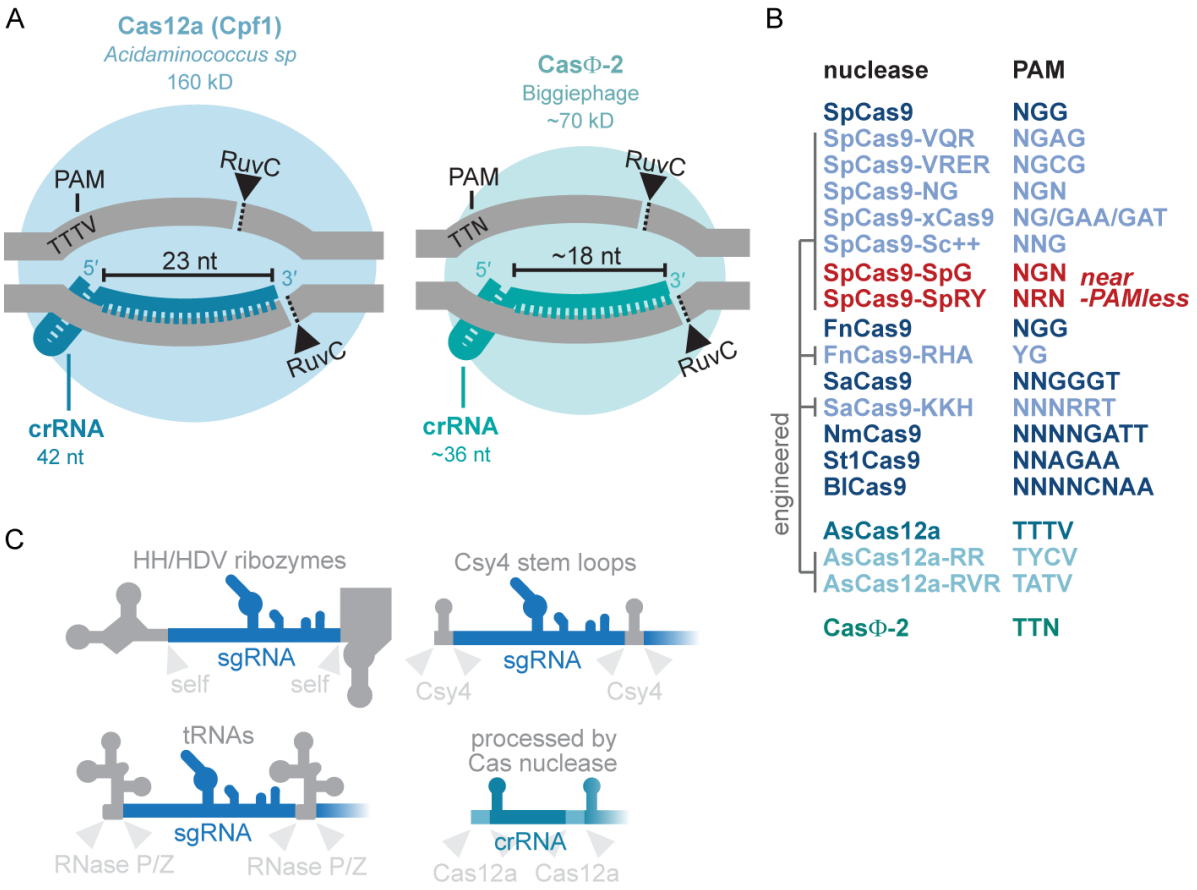
(A) Different mutations could be introduced using genome editing to observe effects on *lin-41*, *let-7* expression, and bursting phenotype. Extra sites (potentially destructive sites) could be expressed from single-copy insertions. **(B)** *Let-7* sites could be separated (e.g., using genome editing to add a *gpd-2* operon). **(C)** An artificial system could be constructed to test if negative feedback by TDMD leads to a difference in spatial or temporal dynamics of target gene regulation. **(D)** A possible forward genetics screen for TDMD pathway genes leading to *let-7* degradation.



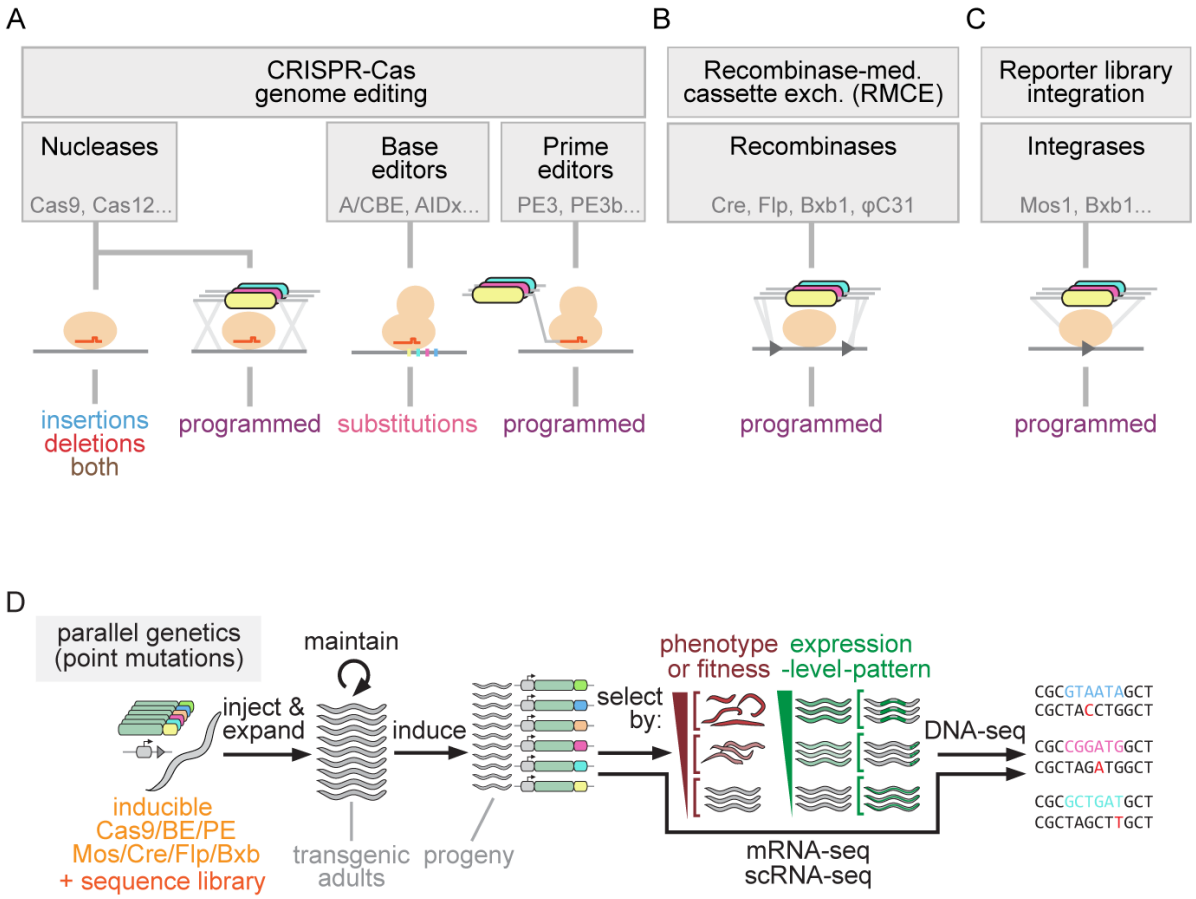
Supplemental Figure 5. Crossing scheme to determine the dominance of *sqt-3(ins)*



Supplemental Figure 6. Genes that may be responsible for repressing *sqt-3(ins)*.
(A) Candidates that may be involved in repressing *sqt-3(ins)*. **(B)** Interaction network of FUBL-1/2/3/4 proteins from STRINGdb[364]. *Fubl-2* was previously shown to also bind *lin-41*, *gld-1*, and *alg-1* mRNAs[121]. **(C)** Possible genetic screen to find suppressors (genes responsible for repressing *sqt-3(ins)*) or enhancers. **(D)** General outline to use a similar approach to find genetic interactors (supr./enh.) for any gene ("gene-x") with a screen-able phenotype ("Pt").



Supplemental Figure 7. Alternative CRISPR-Cas nucleases and sgRNA expression
(A) Alternative CRISPR-nucleases to Cas9 with different properties. **(B)** Various CRISPR-nucleases and their PAM requirements. **(C)** Alternative systems for guide RNA expression from PolII promoters and multiple guide RNAs processed from one transcript (for Csy4, tRNA, Cas12a).



Supplemental Figure 8. Alternative genetic systems that allow programmed edits
(A) CRISPR-Cas systems. **(B)** Recombinase-based systems. **(C)** Integrase-based systems. **(D)** Possible workflow for future parallel genetics approaches. Targeted FACS-based sorting and scRNA-seq could be used to measure tissue- or cell type-specific expression of massively parallel reporter assays.

reference	year	model	*in vivo*	endogenous	readout: level	readout: pattern	readout: phenotype	life cycle	tissue	regions	type of variants	variants tested
several	-	yeast	-	-	+	-	-	-	-	all	all	50 k, 100 mio fragm.
several	-	cell lines	-	-	+	-	-	-	-	all	all	6 mio, billion fragm.
several	-	worm	+	-	+	+	-	all	all	all	all	<100 / region
several	-	fly	+	-	+	+	-	embryo	all	all	all	<30 / region
Gisselbrecht	2013	fly	+	-	+	-	-	embryo	all	enhancer	fragment	600
Kvon	2014	fly	+	-	+	+	-	embryo	all	enhancer	fragment	8 k
Crocker	2015	fly	+	-	+	+	-	embryo	all	enhancer	point, fragment	50
Gisselbrecht	2020	fly	+	-	+	-	-	embryo	all	enhancer	fragment	600
Fuqua	2020	fly	+	-	+	+	-	embryo	all	enhancer	point (most)	300 (2.6 k*)
Farley	2015	seasquirt	+	-	+	-	-	larvae	all	enhancer	point	100 k
Smith	2013	zebrafish	+	-	+	-	-	embryo	all	enhancer	fragment	180 (4.1 k*)
Yartseva	2016	zebrafish	+	-	+	-	-	embryo	all	3'UTRs	fragment, tiling	nd (350 k*)
Rabani	2017	zebrafish	+	-	+	-	-	embryo	all	3'UTRs	fragment, tiling	35 k
Pennacchio	2006	mouse	+	-	+	+	-	embryo	all	enhancer	fragment	170
Visel	2009	mouse	+	-	+	+	-	embryo	all	enhancer	fragment	70
Patwardhan	2012	mouse	+	-	+	-	-	adult	liver	3 enhancers	point	100 k
Smith	2013	mouse	+	-	+	-	-	adult	liver	enhancer	synth.fragment	5 k
Shen	2016	mouse	+	-	+	-	-	adult	brain/retina	enhancer	fragment	3 k
Kvon	2020	mouse	+	-	+	+	-	embryo	all	enhancer	point	150 (1.5 k*)
Snetkova	2021	mouse	+	-	+	+	-	embryo	all	23 enhancers	point	70 (1.5 k*)
Lagunas	2021	mouse	+	-	+	-	-	adult	brain	3'UTR	point	500
Lambert	2021	mouse	+	-	+	-	-	adult	brain	enhancer	fragment	400

*total # of substitutions (each "variant" contained several substitutions)

Supplemental Table 1. Reporter studies of regulatory sequences in animals.
References[23,25,28,29,232,365–372,372–376].

reference	year	model	*in vivo*	endogenous	readout: level	readout: pattern	readout: phenotype	life cycle	tissue	regions	type of variants	variants tested
several	-	yeast	-	+	+	-	+	-	-	all	all	15 k
several	-	cell lines	-	+	+	-	+	-	-	all	all	4 k, or 200 k sgRNAs
several	-	worm	+	+	+	+	+	all	all	all	all	<10 / region
several	-	fly	+	+	+	+	+	all	all	all	all	<10 / region
Parsch	2000	fly	+	+	+	-	+	all	all	3'UTR	indel	2
Chen	2020	fly	+	+	-	-	+	all	all	5'UTR	indel	25
Burger	2016	zebrafish	+	+	-	-	+	all	all	genes*	indel	-*
Kroll	2021	zebrafish	+	+	-	-	+	all	all	genes*	indel	-*
Shaw	2021	zebrafish	+	+	-	-	+	all	all	genes*	indel	-*
Rodríguez-Leal	2017	tomato	+	+	-	-	+	all	all	promoter	indel	30
Hendelman	2021	tomato	+	+	-	-	+	all	all	promoter	indel	40
Wang	2021	tomato	+	+	-	-	+	all	all	promoter	indel	60
Kvon	2020	mouse	+	+	-	-	+	embryo	all	enhancer	point	20
Hay	2016	mouse	+	+	+	+	+	embryo	all	enhancer	indel	7
Osterwalder	2018	mouse	+	+	+	+	+	embryo	all	enhancer	indel	23
Hörnblad	2021	mouse	+	+	+	+	+	embryo	all	enhancer	indel	40
Anania	2021	mouse	+	+	+	+	+	embryo	all	insulator	indel	20
Perry	2012	mouse	+	+	+	+	+	adult	all	3'UTR, isoforms	fragment	3
Gruner	2019	mouse	+	+	+	+	+	adult	all	3'UTR, isoforms	fragment	3
Miller	2002	mouse	+	+	+	+	+	adult	all	3'UTR, complete	fragment	2
Terenzio	2018	mouse	+	+	+	+	+	adult	all	3'UTR, complete	fragment	2
Labi	2019	mouse	+	+	+	+	+	adult	all	3'UTR, miRNAs	point	2
-this study-	2021	worm	+	+	+	+	+	all	all	all	indel	30-1,000
V2.0	?	worm	+	+	+	+	+	all	all	all	all	1,000-10,000

*no regulatory regions were targeted but approach could be adapted

Supplemental Table 2. Mutagenesis studies of endogenous regulatory sequences in animals.
Mutations or genetic variants tested at the native genomic locus (recent studies using CRISPR-Cas9).
References[24,94,367,377–392].

model	*in vivo*	cell types	develop-ment	parallel DNA delivery	reporter assay	forward genetics	reverse genetics (RNAi)	reverse genetics (CRISPR)	individuals per experiment (10^x)
cell lines, yeast	-	1s	-	+	+	(+)	+	+	8
iPSCs, ESCs	-	1s	+	+	+	(+)	+	+	7
organoids	-	10s	+	+	+	-	+	+	8 (cells), 2
Mus musculus	+	1000s	+++	-	+	(+)	+	+	1
Danio rerio	+	100s	+++	-	+	+	+	+	3 (eggs), 1
Xenopus tropicalis	+	100s	+++	-	+	+	+	+	1
Drosophila melanogaster	+	100s	+++	-	+	+	+	+	3 (embryos), 2
Caenorhabditis elegans	+	100s	+++	-	+	+	+	+	6
Arabidopsis thaliana	+	100s	++	+	+	+	+	+	2
Schmidtea mediterranea	+	100s	++	-	-	-	+	-	2
Ciona intestinalis	+	10s	++	+	+	+	+	+	3 (larvae), 1

Supplemental Table 3. *C. elegans* compared to other model systems.

year	discoveries in transcriptional regulation	references
1991	*ges-1* promoter encodes spatial regulation, first promoter bashing	Aamodt, '91
2009	*CHE-1 activates *cog-1*	O'Meara, '09
2011	reprogramming by CHE-1, chromatin barrier	Tursun, '11
2020	temporal intersection of TBX-37/38+CHE-1 activates *lsy-6*	Charest, '20
2020	neuronal homebox TF code	Reilly, '20
	discoveries in post-transcriptional regulation	
1993	*miRNA *lin-4* represses *lin-14*	Lee, '93; Wightman, '93
1993	*genes of the nonsense-mediated decay pathway	Pulak and Anderson, '93
1997	*RBP FBF-1/2 (Pumilio) translationally represses *fem-3*	Zhang, '97
1998	dsRNA-triggered RNA interference (parallel work in plants and fungi)	Fire, '98
1999	*RBP GLD-1 (QKI) translationally represses *tra-2*	Jan, '99
2000	*miRNA *let-7* represses *lin-41*	Reinhart, '00
2003	*miRNA *lsy-6* represses *cog-1*	Johnston and Hobert, '03
2008	3′ UTRs primary regulators in the gonad	Merrit, '08
2018	interaction of 5'- and 3' UTR binding sites (GLD-1)	Theil, '18
	*discovery enabled by phenotypic non-coding mutants	
	methods for genome-wide insights	
2003	phenotypes annotated by RNAi	Kamath, '03; Kemphues, '05; Sönnichsen, '05
2006	miRNA binding sites, predicted	Lall, '06
2007	RBP binding sites	Zisoulis, '10; Kershner, '10; Jungkamp, '11; Rybak-Wolf, '14
2010	transcriptomics, *development*	Gerstein, '10; Spencer, '11; Grün, '14
2010	3' UTRomes, *development*	Mangone, '10; Jan, '11
2010	TF binding sites/chromatin states, *development*	Gerstein, '10; Niu, '11; Araya, '14
2011	transcriptomics, *single-cell*	Spencer,'11/14; Cao,'17; Kaletsky,'18; Packer,'19; Taylor,'21
2014	miRNA binding sites	Grosswendt, '14; Broughton, '16
2015	spatial transcriptomics, *early embryo/gonad*	Hashimshony, '15; Diag, '18
2017	regions with accessible chromatin, *development*	Daugherty, '17; Ho, '17; Jänes, '18
2021	regions with accessible chromatin, *single-cell*	Durham, '21
2021	eQTLs, *single-cell*	Ben-David, '21
	methods to select for phenotypes	
-	manual: morphology, behavior, cell lineage, sex development	Ambros and Horvitz, '84; Brenner, '74; Cox, '80; Hodgkin and Brenner, '77; Horvitz and Sulston, '80; Kusch and Edgar, '86
-	manual w. fluorescent tags: expression pattern/level, cell/pathway reporter	Chang, '03; Troemel, '99; Tursun, '11; Winston, '02
-	by sensitivity or resistance: temperature, hypoxia, toxic/chemotactic compounds	Driscoll, '89; Friedman and Johnson, '88; Kemphues, '88; Lewis, '80; McGrath, '11; Mori, '99; O'Rourke, '11; Zhou, '11
-	by microfluidics devices: size, shape, fluorescence	Andersen, '15; Crane, '12; Doitsidou, '08; Timmermeyer, '19

Supplemental Table 4. *C. elegans*, gene regulation discoveries and methods.
References[39,74,77,89,107,116,119,120,138,147,190,197,271,274,275,277,303,307,311,356,393–430,430–434].

location	gene	year ↑	reference	variants tested	type	mechanism explained	evidence	regulatory factors	expression in mutant
ups.	*ges-1*	1991	Aamodt	95	del	no	-		reduced / higher
ups.	*mec-3*	1991	Way	11	indel, point	no	-		reduced / higher
ups.	*vit-2*	1992	MacMorris	13	del, point	no	-		null / reduced
ups.	*3 genes*	1993	Okkema	95	del	no	-		reduced
ups./5'/3'/int.	*unc-54*	1993	Okkema	30	del	no	-		reduced
ups.	*hlh-1*	1994	Krause	45	del	no	-		reduced / higher
ups.	*ges-1*	1995	Egan	35	del, point	for some	genetics	GATA, nd	reduced / higher
ups.	*col-19*	1995	Liu	6	del	no	-		reduced
ups.	*mec-3*	1996	Wang	5	del	yes	genetics	UNC-86	null / reduced
ups.	*dpy-7*	1997	Gilleard	16	del	yes	motif	GATA	reduced
ups.	*ace-1*	1999	Culetto	10	del	no	-		null / reduced
ups.	*mtl-1, mtl-2*	1999	Moilanen	20	del, point	for some	genetics	ELT-2, nd	null / reduced
ups.	*lin-48*	2001	Johnson	14	del	yes	genetics	EGL-38	reduced
ups.	*lin-48*	2002	Wang	12	del	yes	motif	CES-2	reduced
ups.	*egl-1*	2003	Thellmann	6	del	yes	binding	HLH-2/HLH-3	reduced
ups.	*let-7*	2003	Johnson	16	del	no	-		reduced
ups.	*egl-17*	2003	Cui	35	del	no	-		null / reduced
ups.	*3 genes*	2003	Kirouac	65	del	no	-		reduced
ups./intron	*8 genes*	2004	Wenick	80	del, point	yes	binding	CEH-10, TTX-3	reduced
ups./intron	*lin-3*	2004	Hwang	30	del, point	yes	binding	NHR-25, HLH-2	reduced
ups.	*egl-5*	2004	Teng	40	del	no	-		reduced / higher
ups.	*end-1, end-3*	2005	Broitman-Maduro	2	point	yes	binding	MED-1	reduced
ups.	*end-1*	2005	Maduro	18	del, point	yes	genetics	POP-1, SKN-1	reduced
ups./intron	*lin-39*	2006	Wagmaister	35	del	for some	genetics	LIN-1, LIN-31	reduced / higher
ups.	*aqp-8*	2007	Mah	17	del, point	yes	genetics	CEH-6	reduced
ups.	*20 genes*	2007	Etchberger	95	del, point	yes	binding	CHE-1	null / reduced
ups.	*lin-11*	2009	Marri	17	del	yes	genetics	FOS-1, LAG-1, ...	null / reduced
ups.	*ttx-3, ceh-10*	2009	Bertrand	20	del	yes	genetics	REF-2, HLH-2	reduced
ups.	*4 genes*	2009	Etchberger	90	del	no	-		reduced / higher
ups.	*5 genes*	2009	Flames	45	del, point	yes	genetics	AST-2	reduced
ups.	*let-7*	2013	Kai	14	del	for some	genetics,		reduced
ups.	*eat-4*	2013	Serrano-Saiz	15	del, point	yes	genetics		reduced
ups.	*col-41*	2015	Yin	8	del	no	-		reduced
ups.	*26 genes*	2015	Stefanakis	195	del, point	for some	some	several	reduced
ups.	*4 genes*	2018	Lloret-Fernández	65	del, point	for some	motif, gen.	several	reduced
ups.	*7 genes*	2020	Serrano-Saiz	70	del, point	for some	some	several	reduced
ups.	*lin-4 miRNA*	2020	Stec	6	del	yes	binding	BLMP-1	reduced
ups/3'/intr.	*ceh-6*	2020	Ahier	11	del	no	-		reduced
3'UTR	*cog-1*	2006	Didiano	25	point	for some	some		higher
3'UTR	*hbl-1*	2007	Nolde	5	del	yes	motif	PUF-9	higher
3'UTR	*30 genes*	2008	Merritt	30+	fragment	for some	genetics	FBF-1/2, MEX-3, ...	higher
3'UTR	*cog-1*	2008	Didiano	70	del, point	for some	some	some	higher
3'UTR	*cebp-1*	2009	Yan	2	point	no	-		reduced / higher
3'UTR	*6 genes*	2010	Merritt	6	fragment	yes	genetics	FBF-1/2	higher
3'UTR	*die-1*	2010	Didiano	35	del, point	no	-		reduced / higher
3'UTR	*par-5*	2014	Mikl	3	fragment	no	-		higher
3'UTR	*cebp-1*	2017	Sharifnia	15	del	no	-		higher
3'/5'UTR	*4 genes*	2018	Theil	18	point	yes	genetics	GLD-1	higher
3'UTR	*lin-41*	2018	Brancati	12	point	yes	genetics	*let-7*	higher
3'UTR	*cyb-2.1*	2019	Wang	2	point	yes	genetics	FBF-1/2	higher

Supplemental Table 5. *C. elegans*, reporter studies of regulatory sequences.
References[37,39,393,435–475].

location	year ↑	ref.	gene	variants tested	mutation type	technique	mechanism explained	expression in mutant
ups.	1988	Trent	*her-1*	1	point	EMS	no	higher
ups.	1992	Zarkower	*tra-1*	1	ins	acetaldehyde	no	reduced
ups.	1994	Perry	*her-1*	3	point	EMS	no	reduced/null
ups.	2000	Wen	*spr-2*	1	del	EMS	no	reduced
ups.	2004	Fay	*pha-1*	1	point	EMS	no	reduced (inferred)
ups.	2004	Garbe	*mat-3*	1	point	gamma irrad.	no	reduced
ups./intron	2004	Hwang	*lin-3*	1	point	EMS	yes, NHR-25 binding	reduced
ups.	2005	Ross	*ref-1*	1	point	EMS	no	nd
ups.	2006	Arata	*psa-3*	1	del	no info	yes, POP-1 motif	reduced
ups.	2006	Thacker	*dpy-5*	1	del	EMS	no	reduced (inferred)
ups.	2006	Harris	*unc-17*	1	del	mutator (mrt-1)	no	reduced (inferred)
ups.	2009	O'Meara	*cog-1*	2	point	EMS	yes, CHE-1 binding	reduced
ups.	2010	Hirose	*egl-1*	1	point	EMS	yes, CEH-34 binding	reduced (inferred)
ups.	2010	Sarin	*ttx-3*	1	del	EMS	yes, REF-2 motif	reduced (inferred)
ups.	2011	Verghese	*nhr-67*	3	del	mutator (dog-1)	yes, HLH-2 motif	reduced (inferred)
ups.	2011	Saffer	*lin-3*	1	point	EMS	no	higher
ups.	2014	Nakagawa	*cnt-1*	1	point	EMS	no	null
ups./intron	2016	Barkoulas	*lin-3*	8	del	CRISPR	no	reduced
ups.	2016	Greene	*srx-44*	1	ins	CRISPR	no	higher (extra cell)
ups.	2020	Serrano-Saiz	*unc-17*	1	del	CRISPR	no	null / reduced
ups.	2021	Lynch	*sygl-1*	8	point	CRISPR	yes, LAG-1 motif	null / reduced
ups.	2021	This study	*sqt-2*	7	indel, cmplx	CRISPR, parallel	no	reduced (inferred)
ups.	2021	This study	*sqt-3*	13	indel, cmplx	CRISPR, parallel	yes, TATA box motif	reduced
5'UTR	1988	Desai	*ham-1*	1	del	no info	yes, SL acceptor site	reduced (inferred)
5'UTR/intron	1993	Clark	*lin-39*	1	point	EMS	no	null / reduced
5'UTR	2000	Furuta	*emb-30*	1	point	EMS	yes, SL acceptor site	null (inferred)
5'UTR	2000	Hong	*mec-4*	1	ins	EMS	yes, creates ups. ATG	null / reduced
5'UTR	2000	Miller	*lin-31*	1	ins	mutator (mut-2)	yes, SL acceptor site	null
5'UTR	2001	Detwiler	*oma-1*	1	point	EMS	yes, creates ups. ATG	null / reduced
5'UTR	2002	Moorman	*acy-1*	1	point	EMS	no	reduced (inferred)
5'UTR	2006	Gleason	*spe-10*	1	point	EMS	yes, creates ups. ATG	null / reduced
5'UTR	2007	Kemp	*zyg-1*	1	point	EMS	no	higher (inferred)
5'UTR	2019	Ilbay	*lin-46*	4	del	CRISPR	yes, LIN-28	higher
3'UTR	1987	Barton	*fem-3*	3	point	EMS	yes, FBF-2 binding (f)	higher
3'UTR	1988	Pulak	*unc-54*	1	del	spontaneous	yes, NMD (f)	reduced
3'UTR	1993	Goodwin	*tra-2*	2	del	Tc1	yes, GLD-1 binding (f)	higher
3'UTR	1994	Zarkower	*tra-1*	1	nd	EMS	no	nd
3'UTR	2007	Sarin	*lsy-6*	1	del	EMS	no	higher
3'UTR	2015	Pagano	*pmk-2*	2	indel	EMS/CRISPR	yes, *miR-58/80-82* sites	higher
3'UTR	2015	Ecsedi	*lin-41*	2	point, del	CRISPR	yes, *let-7*	higher
3'UTR	2017	Sharifnia	*cebp-1*	2	del	CRISPR	no	nd
3'UTR	2021	Albarqi	*mex-3*	5	del	CRISPR	no	higher
3'UTR	2021	This study	*egl-30*	11	indel, cmplx	CRISPR, parallel	no	reduced (inferred)
3'UTR	2021	This study	*sqt-3*	40	indel, cmplx	CRISPR, parallel	no	reduced
3'UTR	2021	This study	*lin-41*	4	del, cmplx	CRISPR	yes, *let-7*	higher
downs	2002	Alper	*ref-2*	1	point	EMS	no	higher

(f): mechanism described in follow-up studies

location	year ↑	ref.	miRNAs	variants tested	mutation type	technique	mechanism explained	expression in mutant
miRNA	1974?	Hodgkin?	*lin-4*	1	del	P^{32}	yes, disrupts miRNA	null / reduced
miRNA	2003	Hanna	*lin-4*	1	point	EMS	yes, disrupts miRNA	null / reduced
miRNA	2010	Ren	*lin-4*	1	point	EMS	yes, affects processing	higher (processing)
miRNA	2000	Reinhart	*let-7*	3	point	EMS?	yes, disrupts miRNA	null / reduced
ups.	2005	Li	*mir-48*	2	point	EMS	no	higher (earlier)
ups.	2007	Sarin	*lsy-6*	1	point	EMS	yes, CHE-1 binding site	null
miRNA	2007	Sarin	*lsy-6*	1	point	EMS	yes, disrupts miRNA	null
downs	2020	Charest	*lsy-6*	1	del	CRISPR	yes, TBX-37/38 binding	null / reduced

Supplemental Table 6. *C. elegans*, mutagenesis studies of endogenous regulatory sequences.
References[107,420,423,447,476–515].

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	-

Supplemental Table 7. Samples

Publications

Parallel genetics of regulatory sequences using scalable genome editing *in vivo.*
Jonathan J. Froehlich*, Bora Uyar*, Margareta Herzog, Kathrin Theil, Petar Glažar, Altuna Akalin, Nikolaus Rajewsky. Cell Reports, 2021. DOI: 10.1016/j.celrep.2021.108988. PMID: 33852857.

Estimation of *C. elegans* cell- and tissue volumes.
Jonathan J. Froehlich, Nikolaus Rajewsky, Collin Ewald. Micropublication Biology, 2021. DOI: 10.17912/micropub.biology.000345. PMID: 33426507.

Conference presentations:

GSA "22nd International *C. elegans* Conference". Selected talk, 2019, UCLA, Los Angeles, USA

GSA "21st International *C. elegans* Conference". Selected talk, 2017, UCLA, Los Angeles, USA

EMBO "Complex Life of mRNA". Poster, 2016, EMBL, Heidelberg, DE

MDC "European Worm Meeting". Poster, 2016, MDC, Berlin, DE

BIMSB "Summer Meeting". Poster, 2015/2016/2017, BIMSB/MDC, Berlin, DE

Declaration of Independent Work

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Jonathan Froehlich
December 2021

Selbstständigkeitserklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsbearbeiterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Jonathan Fröhlich
Dezember 2021

Acknowledgements

I am especially grateful to:

Nikolaus Rajewsky for the deep discussions of results and ideas; advice on decisions, texts, and presentations, and the strong support to follow a path from which I learned so much.

The whole team - current and past - for the discussions on diverse projects, feedback on my science, and most importantly the many uplifting moments we shared.

Margareta Herzog for advice, hundreds of microinjections, help with cultures. Also, for keeping the lab running as lab manager together with the **team of technicians** (current and past).

Alexandra Tschernycheff for helping me kindly and patiently with many administrative issues; for the extra efforts behind the scenes.

Bora Uyar for collaborating on this project, patience, feedback on texts; for what I learned about computation, communication, and project management.

Altuna Akalin for the collaboration and input on computation and writing.

Baris Tursun for very helpful feedback on the project and writings; advice and support during the years; joining advisory and examination committees, and trips to the conferences in Los Angeles.

Sergej Herzog for preparing culture plates and biosorter maintenance.

Luisa Cochella for kind encouragement and very helpful feedback on the publication draft.

Marvin Jens and **David Koppstein** for comments on earlier drafts of the thesis.

Kathrin Theil and **Petar Glažar** for helping with pulldown experiments and kmer match analysis.

Marcel Schilling for writing comradery.

For additional contributions: Claudia Quedenau, Daniele Franze, of the MDC sequencing facility for Pacbio sequencing. Marie Kirchner of the MDC mass spectrometry facility for protein measurements. Seung Joon Kim and Marcel Schilling for small RNA scripts. Salah Ayoub for late-night sequencing runs and advice. João Ramalho, Mike Boxem, Jason Chin, Christian Frøkjaer-Jensen, Erik Jorgensen, Daniel Dickinson, Bob Goldstein for sharing plasmids and strains. For helpful conversations on the topic: Victor Ambros, Ye Duan, João Ramalho, Christian Frøkjaer-Jensen, Matthew Schwartz, Craig Mello, Katherine McJunkin, Mihail Sarov, Philipp Junker, Erik van Nimwegen. For joining my thesis advisory or examination committees: Robert Zinzen, Matthias Selbach, Christian Schmitz-Linneweber, Christine Mayr, Markus Landthaler.

I am extremely grateful for having met so many good **people who are not listed** by name here, at the institute and in the lab. Many also belong to the next group.

Friends, for the many good moments together, help and perspectives.

Moms, dad, grandma, for support and inspiration.

Martina, for best help, advice, and wonderful times.

Bibliography

INTRODUCTION

1. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. (Academic, 2006).
2. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology* 19, 621 (2018).
3. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* 172, 650–665 (2018).
4. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* 167, 1170–1187 (2016).
5. Das, S., Vera, M., Gandin, V., Singer, R. H. & Tutucci, E. Intracellular mRNA transport and localized translation. *Nature Reviews Molecular Cell Biology* 1–22 (2021) doi:10.1038/s41580-021-00356-8.
6. Mayr, C. Regulation by 3'-Untranslated Regions. *Annu Rev Genet* 51, 171–194 (2017).
7. Bartel, D. P. Metazoan MicroRNAs. *Cell* 173, 20–51 (2018).
8. Corley, M., Burns, M. C. & Yeo, G. W. How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell* 78, 9–29 (2020).
9. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature Reviews Genetics* 15, 829–845 (2014).
10. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 8, 93–103 (2007).
11. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* 424, 147–151 (2003).
12. Wittkopp, P. J. & Kalay, G. Cis -regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* 13, 59–69 (2012).
13. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* 16, 197 (2015).
14. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* 577, 179–189 (2020).
15. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* 166, 538–554 (2016).
16. Manning, K. S. & Cooper, T. A. The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology* 18, 102 (2017).
17. Agrawal, S. RNA Therapeutics Are Stepping Out of the Maze. *Trends in Molecular Medicine* 0, (2020).
18. Matharu, N. & Ahituv, N. Modulating gene regulation to treat genetic disorders. *Nature Reviews Drug Discovery* 19, 757–775 (2020).
19. Sheridan, C. First small-molecule drug targeting RNA gains momentum. *Nature Biotechnology* 39, 6–8 (2021).
20. Wang, F., Zuroske, T. & Watts, J. K. RNA therapeutics on the rise. *Nat Rev Drug Discov* 19, 441–442 (2020).
21. Wu, Y. *et al.* Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nature Medicine* 25, 776–783 (2019).
22. Zeng, J. *et al.* Therapeutic base editing of human hematopoietic stem cells. *Nat Med* 1–7 (2020) doi:10.1038/s41591-020-0790-y.
23. Fuqua, T. *et al.* Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* 1–5 (2020) doi:10.1038/s41586-020-2816-5.

24. Hendelman, A. *et al.* Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell* 0, (2021).
25. Snetkova, V. *et al.* Ultraconserved enhancer function does not require perfect sequence conservation. *Nature Genetics* 1–8 (2021) doi:10.1038/s41588-021-00812-3.
26. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet* 26, 66–74 (2010).
27. Wong, E. S. *et al.* Deep conservation of the enhancer regulatory code in animals. *Science* 370, (2020).
28. Crocker, J. *et al.* Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. *Cell* 160, 191–203 (2015).
29. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* 350, 325–328 (2015).
30. Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S. & Levine, M. S. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *PNAS* 113, 6508–6513 (2016).
31. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annual Review of Cell and Developmental Biology* 35, 357–379 (2019).
32. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell* 56, 575–587 (2021).
33. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* 15, 453–468 (2014).
34. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* 13, 613–626 (2012).
35. Zeitlinger, J. Seven myths of how transcription factors read the cis-regulatory code. *Current Opinion in Systems Biology* (2020) doi:10.1016/j.coisb.2020.08.002.
36. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* 122, 205–214 (1996).
37. Didiano, D., Cochella, L., Tursun, B. & Hobert, O. Neuron-type specific regulation of a 3'UTR through redundant and combinatorially acting cis-regulatory elements. *RNA* 16, 349–363 (2010).
38. Ludwig, M. Z., Manu, Kittler, R., White, K. P. & Kreitman, M. Consequences of Eukaryotic Enhancer Architecture for Gene Expression Dynamics, Development, and Fitness. *PLOS Genetics* 7, e1002364 (2011).
39. Theil, K., Herzog, M. & Rajewsky, N. Post-transcriptional Regulation by 3′ UTRs Can Be Masked by Regulatory Elements in 5′ UTRs. *Cell Reports* 22, 3217–3226 (2018).
40. Jens, M. & Rajewsky, N. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nature Reviews Genetics* 16, 113–126 (2015).
41. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nature Protocols* 11, 1782 (2016).
42. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annual Review of Genomics and Human Genetics* 20, 99–127 (2019).
43. Shendure, J. & Fields, S. Massively Parallel Genetics. *Genetics* 203, 617–619 (2016).
44. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* 106, 159–164 (2015).
45. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology* 1–21 (2020) doi:10.1038/s41587-020-0561-9.
46. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337, 816–821 (2012).

47. Yeh, C. D., Richardson, C. D. & Corn, J. E. Advances in genome editing through control of DNA repair pathways. *Nature Cell Biology* 21, 1468–1478 (2019).
48. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 527, 192–197 (2015).
49. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis -regulatory element identification in mammalian cells. *Nature Methods* 14, 629–635 (2017).
50. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* aag2445 (2016) doi:10.1126/science.aag2445.
51. Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am J Hum Genet* 101, 192–205 (2017).
52. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotech* advance online publication, (2016).
53. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotech* advance online publication, (2016).
54. Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* 353, 1545–1549 (2016).
55. Vierstra, J. *et al.* Functional footprinting of regulatory DNA. *Nat Meth* advance online publication, (2015).
56. Litterman, A. J. *et al.* A global map of RNA binding protein occupancy guides functional dissection of post-transcriptional regulation of the T cell transcriptome. *bioRxiv* 448654 (2018) doi:10.1101/448654.
57. Wu, Q. *et al.* In situ functional dissection of RNA cis -regulatory elements by multiplex CRISPR-Cas9 genome engineering. *Nature Communications* 8, 2109 (2017).
58. Zhao, W. *et al.* CRISPR–Cas9-mediated functional dissection of 3′-UTRs. *Nucleic Acids Res* 45, 10800–10810 (2017).
59. Cuella-Martin, R. *et al.* Functional interrogation of DNA damage response variants with base editing screens. *Cell* 184, 1081-1097.e19 (2021).
60. Erwood, S. *et al.* Saturation variant interpretation using CRISPR prime editing. *bioRxiv* 2021.05.11.443710 (2021) doi:10.1101/2021.05.11.443710.
61. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 513, 120–123 (2014).
62. Hanna, R. E. *et al.* Massively parallel assessment of human variants with base editor screens. *Cell* 184, 1064-1080.e20 (2021).
63. Baugh, L. R. & Day, T. Nongenetic inheritance and multigenerational plasticity in the nematode C. elegans. *eLife* 9, e58498 (2020).
64. Burga, A. & Lehner, B. Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *The FEBS Journal* 279, 3765–3775 (2012).
65. Burga, A., Casanueva, M. O. & Lehner, B. Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* 480, 250–253 (2011).
66. Casanueva, M. O., Burga, A. & Lehner, B. Fitness Trade-Offs and Environmentally Induced Mutation Buffering in Isogenic C. elegans. *Science* 335, 82–85 (2012).
67. El-Brolosy, M. A. *et al.* Genetic compensation triggered by mutant mRNA degradation. *Nature* 568, 193 (2019).
68. Félix, M.-A. & Barkoulas, M. Pervasive robustness in biological systems. *Nature Reviews Genetics* 16, 483–496 (2015).

69. Hart, Y. & Alon, U. The Utility of Paradoxical Components in Biological Circuits. *Molecular Cell* 49, 213–221 (2013).

70. MacNeil, L. T. & Walhout, A. J. M. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* 21, 645–657 (2011).

71. Vu, V. *et al.* Natural Variation in Gene Expression Modulates the Severity of Mutant Phenotypes. *Cell* 162, 391–402 (2015).

72. Urnov, F. D. Genome Editing B.C. (Before CRISPR): Lasting Lessons from the "Old Testament". *The CRISPR Journal* 1, 34–46 (2018).

73. Corsi, A. K., Wightman, B. & Chalfie, M. A Transparent Window into Biology: A Primer on Caenorhabditis elegans. *Genetics* 200, 387–407 (2015).

74. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667 (2017).

75. Froehlich, J. J., Rajewsky, N. & Ewald, C. Y. Estimation of C. elegans cell- and tissue volumes. *microPublication Biology* 2021, (2021).

76. Grün, D. *et al.* Conservation of mRNA and protein expression during development of C. elegans. *Cell Rep* 6, 565–577 (2014).

77. Brenner, S. The Genetics of Caenorhabditis Elegans. *Genetics* 77, 71–94 (1974).

78. Yemini, E., Jucikas, T., Grundy, L. J., Brown, A. E. X. & Schafer, W. R. A database of C. elegans behavioral phenotypes. *Nat Methods* 10, 877–879 (2013).

79. Mello, C. C., Kramer, J. M., Stinchcomb, D. & Ambros, V. Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* 10, 3959–3970 (1991).

80. Nance, J. & Frøkjær-Jensen, C. The Caenorhabditis elegans Transgenic Toolbox. *Genetics* 212, 959–990 (2019).

81. Vicencio, J. & Cerón, J. A Living Organism in your CRISPR Toolbox: Caenorhabditis elegans Is a Rapid and Efficient Model for Developing CRISPR-Cas Technologies. *CRISPR J* 4, 32–42 (2021).

82. Harris, T. W. *et al.* WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res* 48, D762–D767 (2020).

83. Félix, M.-A. *et al.* Natural and Experimental Infection of Caenorhabditis Nematodes by Novel Viruses Related to Nodaviruses. *PLOS Biology* 9, e1000586 (2011).

84. Gilleland, C. L., Falls, A. T., Noraky, J., Heiman, M. G. & Yanik, M. F. Computer-Assisted Transgenesis of Caenorhabditis elegans for Deep Phenotyping. *Genetics* 201, 39–46 (2015).

85. Khodakova, A. S., Vilchis, D. V., Amanor, F. & Samuel, B. S. Population scale nucleic acid delivery to Caenorhabditis elegans via electroporation. *bioRxiv* 2020.10.15.340513 (2020) doi:10.1101/2020.10.15.340513.

86. Burga, A., Ben-David, E., Lemus Vergara, T., Boocock, J. & Kruglyak, L. Fast genetic mapping of complex traits in C. elegans using millions of individuals in bulk. *Nat Commun* 10, 2680 (2019).

RESULTS

87. Ghanta, K. S. & Mello, C. C. Melting dsDNA Donor Molecules Greatly Improves Precision Genome Editing in Caenorhabditis elegans. *Genetics* 216, 643–650 (2020).

88. Wood, A. J. *et al.* Targeted Genome Editing Across Species Using ZFNs and TALENs. *Science* 333, 307–307 (2011).

89. Diag, A., Schilling, M., Klironomos, F., Ayoub, S. & Rajewsky, N. Spatiotemporal m(i)RNA Architecture and 3' UTR Regulation in the C. elegans Germline. *Dev. Cell* 47, 785-800.e8 (2018).

90. Fox, P. M. *et al.* Cyclin E and CDK-2 regulate proliferative cell fate and cell cycle progression in the C. elegans germline. *Development* 138, 2223–2234 (2011).
91. Hubbard, E. J. A. & Schedl, T. Biology of the Caenorhabditis elegans Germline Stem Cell System. *Genetics* 213, 1145–1188 (2019).
92. Waaijers, S. *et al.* CRISPR/Cas9-Targeted Mutagenesis in Caenorhabditis elegans. *Genetics* 195, 1187–1191 (2013).
93. Waaijers, S. *et al.* A tissue-specific protein purification approach in Caenorhabditis elegans identifies novel interaction partners of DLG-1/Discs large. *BMC Biology* 14, 66 (2016).
94. Froehlich, J. J. *et al.* Parallel genetics of regulatory sequences using scalable genome editing in vivo. *Cell Reports* 35, 108988 (2021).
95. Frøkjær-Jensen, C. *et al.* Single-copy insertion of transgenes in Caenorhabditis elegans. *Nature Genetics* 40, 1375–1383 (2008).
96. Frøkjær-Jensen, C. *et al.* Random and targeted transgene insertion in Caenorhabditis elegans using a modified Mos1 transposon. *Nature Methods* 11, 529–534 (2014).
97. Zevian, S. C. & Yanowitz, J. L. Methodological considerations for heat shock of the nematode Caenorhabditis elegans. *Methods* 68, 450–457 (2014).
98. Farboud, B. & Meyer, B. J. Dramatic Enhancement of Genome Editing by CRISPR/Cas9 Through Improved Guide RNA Design. *Genetics* genetics.115.175166 (2015) doi:10.1534/genetics.115.175166.
99. Konrad, A., Brady, M. J., Bergthorsson, U. & Katju, V. Mutational Landscape of Spontaneous Base Substitutions and Small Indels in Experimental Caenorhabditis elegans Populations of Differing Size. *Genetics* 212, 837–854 (2019).
100. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* 581, 444–451 (2020).
101. Ramsden, D. A., Carvajal-Garcia, J. & Gupta, G. P. Mechanism, cellular functions and cancer roles of polymerase-theta-mediated DNA end joining. *Nat Rev Mol Cell Biol* 1–16 (2021) doi:10.1038/s41580-021-00405-2.
102. van Schendel, R., Roerink, S. F., Portegijs, V., van den Heuvel, S. & Tijsterman, M. Polymerase Θ is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis. *Nat Commun* 6, (2015).
103. Shou, J., Li, J., Liu, Y. & Wu, Q. Precise and Predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Mol. Cell* 71, 498-509.e4 (2018).
104. Hussmann, J. A. *et al.* Mapping the genetic landscape of DNA double-strand break repair. *Cell* 0, (2021).
105. Bagga, S. *et al.* Regulation by let-7 and lin-4 miRNAs Results in Target mRNA Degradation. *Cell* 122, 553–563 (2005).
106. Ecsedi, M., Rausch, M. & Großhans, H. The let-7 microRNA Directs Vulval Development through a Single Target. *Developmental Cell* 32, 335–344 (2015).
107. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403, 901–906 (2000).
108. Slack, F. J. *et al.* The lin-41 RBCC Gene Acts in the C. elegans Heterochronic Pathway between the let-7 Regulatory RNA and the LIN-29 Transcription Factor. *Molecular Cell* 5, 659–669 (2000).
109. Vella, M. C., Choi, E.-Y., Lin, S.-Y., Reinert, K. & Slack, F. J. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. *Genes Dev.* 18, 132–137 (2004).
110. Long, D. *et al.* Potent effect of target structure on microRNA function. *Nat Struct Mol Biol* 14, 287–294 (2007).

111. Abbott, A. L. *et al.* The let-7 MicroRNA Family Members mir-48, mir-84, and mir-241 Function Together to Regulate Developmental Timing in Caenorhabditis elegans. *Developmental Cell* 9, 403–414 (2005).

112. Aeschimann, F. *et al.* LIN41 Post-transcriptionally Silences mRNAs by Two Distinct and Position-Dependent Mechanisms. *Molecular Cell* 65, 476-489.e4 (2017).

113. Brancati, G. & Großhans, H. An interplay of miRNA abundance and target site architecture determines miRNA activity and specificity. *Nucleic Acids Res* 46, 3259–3269 (2018).

114. Hunter, S. E. *et al.* Functional Genomic Analysis of the let-7 Regulatory Network in Caenorhabditis elegans. *PLOS Genetics* 9, e1003353 (2013).

115. Zhang, H., Artiles, K. L. & Fire, A. Z. Functional relevance of 'seed' and 'non-seed' sequences in microRNA-mediated promotion of C. elegans developmental progression. *RNA* 21, 1980–1992 (2015).

116. Jänes, J. *et al.* Chromatin accessibility dynamics across C. elegans development and ageing. *eLife* 7, e37344 (2018).

117. Adikusuma, F. *et al.* Large deletions induced by Cas9 cleavage. *Nature* 560, E8–E9 (2018).

118. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 36, 765–771 (2018).

119. Cox, G. N., Laufer, J. S., Kusch, M. & Edgar, R. S. Genetic and Phenotypic Characterization of Roller Mutants of Caenorhabditis Elegans. *Genetics* 95, 317–339 (1980).

120. Kusch, M. & Edgar, R. S. Genetic Studies of Unusual Loci That Affect Body Shape of the Nematode Caenorhabditis Elegans and May Code for Cuticle Structural Proteins. *Genetics* 113, 621–639 (1986).

121. Theil, K., Imami, K. & Rajewsky, N. Identification of proteins and miRNAs that specifically bind an mRNA in vivo. *Nat Commun* 10, 1–13 (2019).

122. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 6, 26 (2011).

DISCUSSION & CONCLUSION

123. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology* 37, 64–72 (2019).

124. Chakrabarti, A. M. *et al.* Target-Specific Precision of CRISPR-Mediated Genome Editing. *Mol. Cell* 73, 699-713.e6 (2019).

125. Chen, W. *et al.* Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res* 47, 7989–8003 (2019).

126. Leenay, R. T. *et al.* Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nature Biotechnology* 1 (2019) doi:10.1038/s41587-019-0203-2.

127. Shen, M. W. *et al.* Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* 563, 646 (2018).

128. Carvajal-Garcia, J. *et al.* Mechanistic basis for microhomology identification and genome scarring by polymerase theta. *PNAS* 117, 8476–8485 (2020).

129. Hanscom, T. & McVey, M. Regulation of Error-Prone DNA Double-Strand Break Repair and Its Impact on Genome Evolution. *Cells* 9, 1657 (2020).

130. Schimmel, J., van Schendel, R., den Dunnen, J. T. & Tijsterman, M. Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. *Trends in Genetics* 35, 632–644 (2019).

131. Schendel, R. van, Heteren, J. van, Welten, R. & Tijsterman, M. Genomic Scars Generated by Polymerase Theta Reveal the Versatile Mechanism of Alternative End-Joining. *PLOS Genetics* 12, e1006368 (2016).

132. Mayya, V. K. & Duchaine, T. F. Ciphers and Executioners: How 3′-Untranslated Regions Determine the Fate of Messenger RNAs. *Front. Genet.* 10, (2019).

133. Vella, M. C., Reinert, K. & Slack, F. J. Architecture of a Validated MicroRNA::Target Interaction. *Chemistry & Biology* 11, 1619–1623 (2004).

134. Broderick, J. A., Salomon, W. E., Ryder, S. P., Aronin, N. & Zamore, P. D. Argonaute protein identity and pairing geometry determine cooperativity in mammalian RNA silencing. *RNA* 17, 1858–1869 (2011).

135. Sætrom, P. *et al.* Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res* 35, 2333–2342 (2007).

136. Flamand, M. N., Gan, H. H., Mayya, V. K., Gunsalus, K. C. & Duchaine, T. F. A non-canonical site reveals the cooperative mechanisms of microRNA-mediated silencing. *Nucleic Acids Research* 45, 7212–7225 (2017).

137. Wu, E. *et al.* Pervasive and Cooperative Deadenylation of 3′UTRs by Embryonic MicroRNA Families. *Molecular Cell* 40, 558–570 (2010).

138. Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W. & Pasquinelli, A. E. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell* 64, 320–333 (2016).

139. Han, J. *et al.* A ubiquitin ligase mediates target-directed microRNA decay independently of tailing and trimming. *Science* (2020) doi:10.1126/science.abc9546.

140. de la Mata, M. *et al.* Potent degradation of neuronal miRNAs induced by highly complementary targets. *EMBO Rep.* 16, 500–511 (2015).

141. Shi, C. Y. *et al.* The ZSWIM8 ubiquitin ligase mediates target-directed microRNA degradation. *Science* (2020) doi:10.1126/science.abc9359.

142. Abrahante, J. E. *et al.* The Caenorhabditis elegans hunchback-like Gene lin-57/hbl-1 Controls Developmental Time and Is Regulated by MicroRNAs. *Developmental Cell* 4, 625–637 (2003).

143. Andachi, Y. A novel biochemical method to identify target genes of individual microRNAs: Identification of a new Caenorhabditis elegans let-7 target. *RNA* 14, 2440–2451 (2008).

144. Ding, X. C., Slack, F. J. & Großhans, H. The let-7 microRNA interfaces extensively with the translation machinery to regulate cell differentiation. *Cell Cycle* 7, 3083–3090 (2008).

145. Großhans, H., Johnson, T., Reinert, K. L., Gerstein, M. & Slack, F. J. The Temporal Patterning MicroRNA let-7 Regulates Several Transcription Factors at the Larval to Adult Transition in C. elegans. *Developmental Cell* 8, 321–330 (2005).

146. Johnson, S. M. *et al.* RAS Is Regulated by the let-7 MicroRNA Family. *Cell* 120, 635–647 (2005).

147. Lall, S. *et al.* A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. *Current Biology* 16, 460–471 (2006).

148. Lin, S.-Y. *et al.* The C. elegans hunchback Homolog, hbl-1, Controls Temporal Patterning and Is a Probable MicroRNA Target. *Developmental Cell* 4, 639–650 (2003).

149. Ketting, R. F. & Cochella, L. Chapter Three - Concepts and functions of small RNA pathways in C. elegans. in *Current Topics in Developmental Biology* (eds. Jarriault, S. & Podbilewicz, B.) vol. 144 45–89 (Academic Press, 2021).

150. Rausch, M., Ecsedi, M., Bartake, H., Müllner, A. & Großhans, H. A genetic interactome of the let-7 microRNA in C. elegans. *Developmental Biology* 401, 276–286 (2015).

151. Hardison, R. C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* 13, 469–483 (2012).

152. Wray, G. A. The evolutionary significance of cis -regulatory mutations. *Nature Reviews Genetics* 8, 206–216 (2007).

153. Wray, G. A. *et al.* The Evolution of Transcriptional Regulation in Eukaryotes. *Molecular Biology and Evolution* 20, 1377–1419 (2003).

154. Kramer, J. M. & Johnson, J. J. Analysis of mutations in the sqt-1 and rol-6 collagen genes of Caenorhabditis elegans. *Genetics* 135, 1035–1045 (1993).

155. Cho, J. Y., Choi, T.-W., Kim, S. H., Ahnn, J. & Lee, S.-K. Morphological Characterization of small, dumpy, and long Phenotypes in Caenorhabditis elegans. *Mol Cells* (2021) doi:10.14348/molcells.2021.2236.

156. Javer, A. *et al.* An open-source platform for analyzing and sharing worm-behavior data. *Nature Methods* 15, 645 (2018).

157. Javer, A., Ripoll-Sánchez, L. & Brown, A. E. X. Powerful and interpretable behavioural features for quantitative phenotyping of Caenorhabditis elegans. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373, 20170375 (2018).

158. Evans, K. S., van Wijk, M. H., McGrath, P. T., Andersen, E. C. & Sterken, M. G. From QTL to gene: C. elegans facilitates discoveries of the genetic mechanisms underlying natural variation. *Trends Genet* S0168-9525(21)00146–3 (2021) doi:10.1016/j.tig.2021.06.005.

159. Dokshin, G. A., Ghanta, K. S., Piscopo, K. M. & Mello, C. C. Robust Genome Editing with Short Single-Stranded and Long, Partially Single-Stranded DNA Donors in Caenorhabditis elegans. *Genetics* 210, 781–787 (2018).

160. Schwartz, M. L., Davis, M. W., Rich, M. S. & Jorgensen, E. M. High-efficiency CRISPR gene editing in C. elegans using Cas9 integrated into the genome. *PLOS Genetics* 17, e1009755 (2021).

161. Yang, B., Schwartz, M. & McJunkin, K. In vivo CRISPR screening for phenotypic targets of the mir-35-42 family in C. elegans. *Genes Dev.* (2020) doi:10.1101/gad.339333.120.

162. Enache, O. M. *et al.* Cas9 activates the p53 pathway and selects for p53-inactivating mutations. *Nat Genet* 1–7 (2020) doi:10.1038/s41588-020-0623-4.

163. Geisinger, J. M. & Stearns, T. CRISPR/Cas9 treatment causes extended TP53-dependent cell cycle arrest in human cells. *Nucleic Acids Research* 48, 9067–9081 (2020).

164. Machour, F. E. & Ayoub, N. Transcriptional Regulation at DSBs: Mechanisms and Consequences. *Trends Genet* 36, 981–997 (2020).

165. Boutin, J. *et al.* CRISPR-Cas9 globin editing can induce megabase-scale copy-neutral losses of heterozygosity in hematopoietic cells. *Nat Commun* 12, 4922 (2021).

166. Leibowitz, M. L. *et al.* Chromothripsis as an on-target consequence of CRISPR–Cas9 genome editing. *Nature Genetics* 1–11 (2021) doi:10.1038/s41588-021-00838-7.

167. Papathanasiou, S. *et al.* Whole chromosome loss and genomic instability in mouse embryos after CRISPR-Cas9 genome editing. *Nat Commun* 12, 5855 (2021).

168. Aljohani, M. D., El Mouridi, S., Priyadarshini, M., Vargas-Velazquez, A. M. & Frøkjær-Jensen, C. Engineering rules that minimize germline silencing of transgenes in simple extrachromosomal arrays in C. elegans. *Nature Communications* 11, 6300 (2020).

169. El Mouridi, S., AlHarbi, S. & Frøkjær-Jensen, C. A histamine-gated channel is an efficient negative selection marker for C. elegans transgenesis. *microPublication Biology* 2021, (2021).

170. Hubbard, E. J. A. FLP/FRT and Cre/lox recombination technology in C. elegans. *Methods* 68, 417–424 (2014).

171. Wurmthaler, L. A., Sack, M., Gense, K., Hartig, J. S. & Gamerdinger, M. A tetracycline-dependent ribozyme switch allows conditional induction of gene expression in Caenorhabditis elegans. *Nature Communications* 10, 491 (2019).

172. Zhang, L., Ward, J. D., Cheng, Z. & Dernburg, A. F. The auxin-inducible degradation (AID) system enables versatile conditional protein depletion in C. elegans. *Development* 142, 4374–4384 (2015).

173. Simon, A. J., d'Oelsnitz, S. & Ellington, A. D. Synthetic evolution. *Nature Biotechnology* 37, 730 (2019).

174. Chatterjee, P. *et al.* An engineered ScCas9 with broad PAM range and high specificity and activity. *Nat Biotechnol* 1–5 (2020) doi:10.1038/s41587-020-0517-0.

175. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* (2020) doi:10.1126/science.aba8853.

176. Chen, H. *et al.* Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. *Nat Biotechnol* 1–4 (2019) doi:10.1038/s41587-019-0331-8.

177. Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* 560, 248 (2018).

178. Hess, G. T. *et al.* Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat Meth* 13, 1036–1042 (2016).

179. Li, C. *et al.* Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. *Nat Biotechnol* 1–8 (2020) doi:10.1038/s41587-019-0393-7.

180. Ma, Y. *et al.* Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat Meth* 13, 1029–1035 (2016).

181. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* 175, 544-557.e16 (2018).

182. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 1–1 (2019) doi:10.1038/s41586-019-1711-4.

183. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods* 1–5 (2021) doi:10.1038/s41592-020-01041-y.

184. McCoy, R. C. *et al.* Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLOS ONE* 9, e106689 (2014).

185. Nonet, M. L. Efficient Transgenesis in Caenorhabditis elegans Using Flp Recombinase-Mediated Cassette Exchange. *Genetics* (2020) doi:10.1534/genetics.120.303388.

186. Frøkjær-Jensen, C., Davis, M. W., Ailion, M. & Jorgensen, E. M. Improved Mos1-mediated transgenesis in C. elegans. *Nature Methods* 9, 117–118 (2012).

187. Durrant, M. G. *et al. Large-scale discovery of recombinases for integrating DNA into the human genome*. 2021.11.05.467528 https://www.biorxiv.org/content/10.1101/2021.11.05.467528v1 (2021) doi:10.1101/2021.11.05.467528.

188. Stinchcomb, D. T., Shaw, J. E., Carr, S. H. & Hirsh, D. Extrachromosomal DNA transformation of Caenorhabditis elegans. *Molecular and Cellular Biology* 5, 3484–3496 (1985).

189. Woglar, A. *et al.* Quantitative cytogenetics reveals molecular stoichiometry and longitudinal organization of meiotic chromosome axes and loops. *PLOS Biology* 18, e3000817 (2020).

190. Doitsidou, M., Flames, N., Lee, A. C., Boyanov, A. & Hobert, O. Automated screening for mutants affecting dopaminergic-neuron specification in C. elegans. *Nat Methods* 5, 869–872 (2008).

191. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods* 1–8 (2021) doi:10.1038/s41592-021-01153-z.

192. Gaublomme, J. T. *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun* 10, 2907 (2019).

193. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods* 1–7 (2020) doi:10.1038/s41592-020-0837-5.

194. Steiner, F. A., Talbert, P. B., Kasinathan, S., Deal, R. B. & Henikoff, S. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome Res.* 22, 766–777 (2012).

195. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 14, 865–868 (2017).

196. Sun, H. & Hobert, O. Temporal transitions in the post-mitotic nervous system of Caenorhabditis elegans. *Nature* 1–7 (2021) doi:10.1038/s41586-021-04071-4.

197. Taylor, S. R. *et al.* Molecular topography of an entire nervous system. *Cell* 0, (2021).

METHODS

198. Fatt, H. V. & Dougherty, E. C. Genetic Control of Differential Heat Tolerance in Two Strains of the Nematode Caenorhabditis elegans. *Science* 141, 266–267 (1963).

199. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6, 343–345 (2009).

200. Friedland, A. E. *et al.* Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nature Methods* 10, 741–743 (2013).

201. Dickinson, D. J., Pani, A. M., Heppert, J. K., Higgins, C. D. & Goldstein, B. Streamlined Genome Engineering with a Self-Excising Drug Selection Cassette. *Genetics* genetics.115.178335 (2015) doi:10.1534/genetics.115.178335.

202. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology* 17, 148 (2016).

203. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nature Methods* 11, 122–123 (2014).

204. Mello, C. & Fire, A. DNA transformation. *Methods Cell Biol.* 48, 451–482 (1995).

205. Radman, I., Greiss, S. & Chin, J. W. Efficient and Rapid C. elegans Transgenesis by Bombardment and Hygromycin B Selection. *PLoS ONE* 8, e76019 (2013).

206. Sulston, J. E. & Hodgkin, J. The Nematode Caenorhabditis elegans. in *The Nematode Caenorhabditis elegans* 587–606 (Cold Spring Harbor Laboratory Press, 1988).

207. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* 12, 996–1006 (2002).

208. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).

209. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods* 12, 823–826 (2015).

210. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat Biotechnol* 32, 1262–1267 (2014).

211. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* 34, 184–191 (2016).

212. Graf, R., Li, X., Chu, V. T. & Rajewsky, K. sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. *Cell Reports* 26, 1098-1103.e3 (2019).

213. Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* 12, 982–988 (2015).

214. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR/Cas9 system. *Science* 343, 80–84 (2014).

215. Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biology* 16, 218 (2015).

216. Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* gr.191452.115 (2015) doi:10.1101/gr.191452.115.

217. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings*. (Bioconductor version: Release (3.11), 2020). doi:10.18129/B9.bioc.Biostrings.

218. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21 (2019).

219. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat Protoc* 14, 68–85 (2019).

220. Ishihama, Y., Rappsilber, J., Andersen, J. S. & Mann, M. Microcolumns with self-assembled particle frits for proteomics. *Journal of Chromatography A* 979, 233–239 (2002).

221. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell Proteomics* 13, 2513–2526 (2014).

EXTENDED BACKGROUND

222. Milo, R. & Phillips, R. *Cell Biology by the Numbers*. (Garland Science, 2015).

223. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics* 1–15 (2020) doi:10.1038/s41576-020-0258-4.

224. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nature Reviews Genetics* 11, 559–571 (2010).

225. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annual Review of Biochemistry* 72, 449–479 (2003).

226. Larke, M. S. C. *et al.* Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Molecular Cell* 81, 983-997.e7 (2021).

227. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* 51, 1664–1669 (2019).

228. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03446-x.

229. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *bioRxiv* 2021.04.22.440891 (2021) doi:10.1101/2021.04.22.440891.

230. Bergman, D. T. *et al. Compatibility logic of human enhancer and promoter sequences*. 2021.10.23.462170 https://www.biorxiv.org/content/10.1101/2021.10.23.462170v1 (2021) doi:10.1101/2021.10.23.462170.

231. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat Genet* 1–10 (2020) doi:10.1038/s41588-020-0578-5.

232. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Molecular Cell* 77, 324-337.e8 (2020).

233. Forrest, M. E. *et al.* Codon usage and amino acid identity are major determinants of mRNA stability in humans. *bioRxiv* 488676 (2018) doi:10.1101/488676.

234. Hia, F. *et al.* Codon bias confers stability to human mRNAs. *EMBO reports* 0, e48220 (2019).

235. Mitschka, S. & Mayr, C. Endogenous p53 expression in human and mouse is not regulated by its 3′UTR. *eLife* 10, e65700 (2021).

236. Narula, A., Ellis, J., Taliaferro, J. M. & Rissland, O. S. Coding regions affect mRNA stability in human cells. *RNA* rna.073239.119 (2019) doi:10.1261/rna.073239.119.

237. Wu, Q. *et al.* Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife* 8, (2019).

238. Cohen, B. *et al.* Mitochondria serve as axonal shuttle for Cox7c mRNA through mechanism that involves its mitochondrial targeting signal. *bioRxiv* 2021.05.19.444640 (2021) doi:10.1101/2021.05.19.444640.

239. Harbauer, A. B. *et al.* Neuronal mitochondria transport Pink1 mRNA via Synaptojanin 2 to support local mitophagy. *bioRxiv* 2021.05.19.444778 (2021) doi:10.1101/2021.05.19.444778.

240. Tocchini, C., Rohner, M., Stetina, S. E. V. & Mango, S. E. Translation-dependent mRNA localization to Caenorhabditis elegans adherens junctions. *bioRxiv* 2021.05.20.444977 (2021) doi:10.1101/2021.05.20.444977.

241. Blumenthal, T. *Trans-splicing and operons in C. elegans*. *WormBook: The Online Review of C. elegans Biology [Internet]* (WormBook, 2018).

242. Mangone, M. *et al.* The landscape of C. elegans 3'UTRs. *Science* 329, 432–435 (2010).

243. Steber, H. S., Gallante, C., O'Brien, S., Chiu, P.-L. & Mangone, M. The C. elegans 3' UTRome v2 resource for studying mRNA cleavage and polyadenylation, 3'-UTR biology, and miRNA targeting. *Genome Res.* (2019) doi:10.1101/gr.254839.119.

244. Goodsell, D. S. *et al.* The RCSB PDB "Molecule of the Month": Inspiring a Molecular View of Biology. *PLOS Biology* 13, e1002140 (2015).

245. Goodsell, D. S., Zardecki, C., Berman, H. M. & Burley, S. K. Insights from 20 years of the Molecule of the Month. *Biochemistry and Molecular Biology Education* 48, 350–355 (2020).

246. Reece-Hoyes, J. S. *et al.* A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol* 6, R110 (2005).

247. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences* 39, 381–399 (2014).

248. Stewart, A. J., Hannenhalli, S. & Plotkin, J. B. Why Transcription Factor Binding Sites Are Ten Nucleotides Long. *Genetics* 192, 973–985 (2012).

249. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* 19, 327–341 (2018).

250. Tamburino, A. M., Ryder, S. P. & Walhout, A. J. M. A compendium of Caenorhabditis elegans RNA binding proteins predicts extensive regulation at multiple levels. *G3 (Bethesda)* 3, 297–304 (2013).

251. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology* 8, 479–490 (2007).

252. Adamala, K. P., Martin-Alarcon, D. A. & Boyden, E. S. Programmable RNA-binding protein composed of repeats of a single modular unit. *Proc. Natl. Acad. Sci. U.S.A.* 113, E2579-2588 (2016).

253. Zhao, Y.-Y. *et al.* Expanding RNA binding specificity and affinity of engineered PUF domains. *Nucleic Acids Research* 46, 4771–4782 (2018).

254. Gil, N. & Ulitsky, I. Regulation of gene expression by cis -acting long non-coding RNAs. *Nature Reviews Genetics* 21, 102–117 (2020).

255. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* 22, 96–118 (2021).

256. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388 (2013).

257. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338 (2013).

258. Piwecka, M. *et al.* Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science* 357, (2017).

259. Ashwal-Fluss, R. *et al.* circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* 56, 55–66 (2014).

260. Cao, D. Reverse complementary matches simultaneously promote both back-splicing and exon-skipping. *BMC Genomics* 22, 586 (2021).
261. Dexheimer, P. J. & Cochella, L. MicroRNAs: From Mechanism to Organism. *Front. Cell Dev. Biol.* 8, (2020).
262. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics* 16, 421–433 (2015).
263. Fromm, B. *et al.* A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu Rev Genet* 49, 213–242 (2015).
264. Fromm, B. *et al.* MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Research* 48, D132–D141 (2020).
265. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 469, 97–101 (2011).
266. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Research* 47, D155–D162 (2019).
267. Kingston, E. R. & Bartel, D. P. Global analyses of the dynamics of mammalian microRNA metabolism. *Genome Res.* (2019) doi:10.1101/gr.251421.119.
268. Reichholf, B. *et al.* Time-Resolved Small RNA Sequencing Unravels the Molecular Principles of MicroRNA Homeostasis. *Molecular Cell* 75, 756-768.e7 (2019).
269. Duan, Y., Veksler-Lublinsky, I. & Ambros, V. Critical contribution of 3' non-seed base pairing to the in vivo function of the evolutionarily conserved let-7a microRNA. *bioRxiv* 2021.03.29.437276 (2021) doi:10.1101/2021.03.29.437276.
270. McGeary, S. E., Bisaria, N. & Bartel, D. P. Pairing to the microRNA 3′ region occurs through two alternative binding modes, with affinity shaped by nucleotide identity as well as pairing position. *bioRxiv* 2021.04.13.439700 (2021) doi:10.1101/2021.04.13.439700.
271. Grosswendt, S. *et al.* Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *Molecular Cell* 54, 1042–1054 (2014).
272. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* 153, 654–665 (2013).
273. Moore, M. J. *et al.* miRNA–target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nature Communications* 6, 8864 (2015).
274. Johnston, R. J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* 426, 845–849 (2003).
275. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843–854 (1993).
276. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408, 86–89 (2000).
277. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 75, 855–862 (1993).
278. Beusch, I., Barraud, P., Moursy, A., Cléry, A. & Allain, F. H.-T. Tandem hnRNP A1 RNA recognition motifs act in concert to repress the splicing of survival motor neuron exon 7. *eLife* 6, e25736 (2017).
279. Franco-Echevarría, E. *et al.* The structure of transcription termination factor Nrd1 reveals an original mode for GUAA recognition. *Nucleic Acids Research* 45, 10293–10305 (2017).
280. Lewis, H. A. *et al.* Sequence-Specific RNA Binding by a Nova KH Domain: Implications for Paraneoplastic Disease and the Fragile X Syndrome. *Cell* 100, 323–332 (2000).
281. Littler, D. R. *et al.* Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. *Nucleic Acids Research* 38, 4527–4538 (2010).

282. Loughlin, F. E. *et al.* The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Molecular Cell* 73, 490-504.e6 (2019).
283. Lukavsky, P. J. *et al.* Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nat Struct Mol Biol* 20, 1443–1449 (2013).
284. Nair, S. K. & Burley, S. K. X-Ray Structures of Myc-Max and Mad-Max Recognizing DNA: Molecular Bases of Regulation by Proto-Oncogenic Transcription Factors. *Cell* 112, 193–205 (2003).
285. Ren, R. *et al.* Structural basis of specific DNA binding by the transcription factor ZBTB24. *Nucleic Acids Research* 47, 8388–8398 (2019).
286. Sheu-Gruttadauria, J. *et al.* Structural Basis for Target-Directed MicroRNA Degradation. *Molecular Cell* 75, 1243-1255.e7 (2019).
287. Teplova, M. *et al.* Structure–function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. *Genes Dev.* 27, 928–940 (2013).
288. Wang, B. *et al.* Structural insights into target DNA recognition by R2R3-MYB transcription factors. *Nucleic Acids Research* 48, 460–471 (2020).
289. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature Methods* 1–9 (2020) doi:10.1038/s41592-020-0965-y.
290. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27, 38–52 (2017).
291. Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA–protein interactions. *Nature Reviews Molecular Cell Biology* 16, 533–544 (2015).
292. Badis, G. *et al.* Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* 324, 1720–1723 (2009).
293. Berger, M. F. *et al.* Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell* 133, 1266–1276 (2008).
294. Noyes, M. B. *et al.* Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. *Cell* 133, 1277–1289 (2008).
295. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell* 70, 854-867.e9 (2018).
296. Kribelbauer, J. F. *et al.* Context-Dependent Gene Regulation by Homeodomain Transcription Factor Complexes Revealed by Shape-Readout Deficient Proteins. *Molecular Cell* 78, 152-167.e11 (2020).
297. Michael, A. K. & Thomä, N. H. Reading the chromatinized genome. *Cell* (2021) doi:10.1016/j.cell.2021.05.029.
298. Kvon, E. Z., Waymack, R., Elabd, M. G. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nature Reviews Genetics* 1–13 (2021) doi:10.1038/s41576-020-00311-x.
299. Kwon, B. *et al.* Enhancers regulate polyadenylation site cleavage and control 3′UTR isoform expression. *bioRxiv* 2020.08.17.254193 (2020) doi:10.1101/2020.08.17.254193.
300. Waymack, R., Gad, M. & Wunderlich, Z. Molecular competition can shape enhancer activity in the Drosophila embryo. *bioRxiv* 2021.05.07.443186 (2021) doi:10.1101/2021.05.07.443186.
301. Arendt, D. *et al.* The origin and evolution of cell types. *Nature Reviews Genetics* 17, 744–757 (2016).
302. Gilbert, S. F. & Barresi, M. J. F. *Developmental Biology*. (Sinauer Associates is an imprint of Oxford University Press, 2016).
303. Ben-David, E. *et al.* Whole-organism eQTL mapping at cellular resolution with single-cell sequencing. *eLife* 10, e65857 (2021).

304. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330 (2020).
305. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* 369, (2020).
306. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* 50, 493–497 (2018).
307. Charest, J. *et al.* Combinatorial Action of Temporally Segregated Transcription Factors. *Developmental Cell* 55, 483-499.e7 (2020).
308. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* 25, 2227–2241 (2011).
309. Allan, D. W. & Thor, S. Transcriptional selectors, masters, and combinatorial codes: regulatory principles of neural subtype specification. *WIREs Developmental Biology* 4, 505–528 (2015).
310. Hobert, O. & Kratsios, P. Neuronal identity control by terminal selectors in worms, flies, and chordates. *Current Opinion in Neurobiology* 56, 97–105 (2019).
311. Reilly, M. B., Cros, C., Varol, E., Yemini, E. & Hobert, O. Unique homeobox codes delineate all the neuron classes of C. elegans. *Nature* 584, 595–601 (2020).
312. Morris, S. A. The evolving concept of cell identity in the single cell era. *Development* 146, (2019).
313. various interviewees. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *cels* 4, 255–259 (2017).
314. Xia, B. & Yanai, I. A periodic table of cell types. *Development* 146, (2019).
315. Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet* 8, 450–461 (2007).
316. Goldman, J. A. & Poss, K. D. Gene regulatory programmes of tissue regeneration. *Nature Reviews Genetics* 21, 511–525 (2020).
317. Medioni, C., Mowry, K. & Besse, F. Principles and roles of mRNA localization in animal development. *Development* 139, 3263–3276 (2012).
318. Patke, A., Young, M. W. & Axelrod, S. Molecular mechanisms and physiological importance of circadian rhythms. *Nature Reviews Molecular Cell Biology* 21, 67–84 (2020).
319. Purvis, J. E. & Lahav, G. Encoding and Decoding Cellular Information through Signaling Dynamics. *Cell* 152, 945–956 (2013).
320. Theunissen, T. W. & Jaenisch, R. Mechanisms of gene regulation in human embryos and pluripotent stem cells. *Development* 144, 4496–4509 (2017).
321. Vastenhouw, N. L., Cao, W. X. & Lipshitz, H. D. The maternal-to-zygotic transition revisited. *Development* 146, (2019).
322. Wagh, K. *et al.* Mechanical Regulation of Transcription: Recent Advances. *Trends in Cell Biology* 31, 457–472 (2021).
323. Yosef, N. & Regev, A. Impulse Control: Temporal Dynamics in Gene Transcription. *Cell* 144, 886–896 (2011).
324. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* 466, 490–493 (2010).
325. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow Enhancers Foster Robustness of Drosophila Gastrulation. *Current Biology* 20, 1562–1567 (2010).
326. Kitano, H. Biological robustness. *Nature Reviews Genetics* 5, 826–837 (2004).
327. Ebert, M. S. & Sharp, P. A. Roles for MicroRNAs in Conferring Robustness to Biological Processes. *Cell* 149, 515–524 (2012).
328. Ma, Z. *et al.* PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature* 568, 259 (2019).

329. Perez, M. F., Francesconi, M., Hidalgo-Carcedo, C. & Lehner, B. Maternal age generates phenotypic variation in *Caenorhabditis elegans*. *Nature* (2017) doi:10.1038/nature25012.

330. Ballouz, S., Pena, M. T., Knight, F. M., Adams, L. B. & Gillis, J. A. The transcriptional legacy of developmental stochasticity. *bioRxiv* 2019.12.11.873265 (2019) doi:10.1101/2019.12.11.873265.

331. Raj, A., Rifkin, S. A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* 463, 913–918 (2010).

332. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics* 53, 354–366 (2021).

333. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *bioRxiv* 2021.04.07.438649 (2021) doi:10.1101/2021.04.07.438649.

334. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* 163, 698–711 (2015).

335. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nature Biotechnology* 37, 803–809 (2019).

336. Vaishnav, E. D. *et al.* A comprehensive fitness landscape model reveals the evolutionary history and future evolvability of eukaryotic cis-regulatory DNA sequences. *bioRxiv* 2021.02.17.430503 (2021) doi:10.1101/2021.02.17.430503.

337. Zrimec, J., Buric, F., Kokina, M., Garcia, V. & Zelezniak, A. Learning the Regulatory Code of Gene Expression. *Frontiers in Molecular Biosciences* 8, 530 (2021).

338. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* 106, 185–192 (2015).

339. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217 (2018).

340. Starita, L. M. *et al.* A multiplexed homology-directed DNA repair assay reveals the impact of ~1,700 BRCA1 variants on protein function. *bioRxiv* 295279 (2018) doi:10.1101/295279.

341. Flibotte, S. *et al.* Whole-genome profiling of mutagenesis in Caenorhabditis elegans. *Genetics* 185, 431–441 (2010).

342. Kutscher, L. M. & Shaham, S. Forward and reverse mutagenesis in C. elegans. *WormBook* 1–26 (2014) doi:10.1895/wormbook.1.167.1.

343. Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nature Biotechnology* 34, 933 (2016).

344. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* 578, 229–236 (2020).

345. Zhang, F. Development of CRISPR-Cas systems for genome editing and beyond. *Quarterly Reviews of Biophysics* 52, (2019).

346. Schep, R. *et al.* Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Molecular Cell* 81, 2216-2230.e10 (2021).

347. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell* 63, 633–646 (2016).

348. Ata, H. *et al.* Robust activation of microhomology-mediated end joining for precision gene editing applications. *PLOS Genetics* 14, e1007652 (2018).

349. Bae, S., Kweon, J., Kim, H. S. & Kim, J.-S. Microhomology-based choice of Cas9 nuclease target sites. *Nat Meth* 11, 705–706 (2014).

350. Martinez-Galvez, G., Manduca, A. & Ekker, S. C. MMEJ-based Precision Gene Editing for applications in Gene Therapy and Functional Genomics. *bioRxiv* 2020.04.25.060541 (2020) doi:10.1101/2020.04.25.060541.

351. Sterken, M. G., Snoek, L. B., Kammenga, J. E. & Andersen, E. C. The laboratory domestication of Caenorhabditis elegans. *Trends in Genetics* 31, 224–231 (2015).

352. The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* 282, 2012–2018 (1998).

353. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution* 34, 1812–1819 (2017).

354. Frézal, L. & Félix, M.-A. C. elegans outside the Petri dish. *eLife* 4, e05849 (2015).

355. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Research* 49, D884–D891 (2021).

356. Andersen, E. C. *et al.* A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. *G3* 5, 911–920 (2015).

357. Nelson, C. & Ambros, V. A cohort of Caenorhabditis species lacking the highly conserved let-7 microRNA. *G3 Genes|Genomes|Genetics* 11, (2021).

358. Wynsberghe, P. M. V. *et al.* LIN-28 co-transcriptionally binds primary let-7 to regulate miRNA maturation in Caenorhabditis elegans. *Nat Struct Mol Biol* 18, 302–308 (2011).

359. Rougvie, A. E. Control of developmental timing in animals. *Nat Rev Genet* 2, 690–701 (2001).

360. Aeschimann, F., Neagu, A., Rausch, M. & Großhans, H. let-7 coordinates the transition to adulthood through a single primary and four secondary targets. *Life Science Alliance* 2, (2019).

361. Del Rio-Albrechtsen, T., Kiontke, K., Chiou, S.-Y. & Fitch, D. H. A. Novel gain-of-function alleles demonstrate a role for the heterochronic gene lin-41 in C. elegans male tail tip morphogenesis. *Developmental Biology* 297, 74–86 (2006).

362. Pereira, L. *et al.* Timing mechanism of sexually dimorphic nervous system differentiation. *eLife* 8, e42078 (2019).

363. Ecsedi, M. & Großhans, H. LIN-41/TRIM71: emancipation of a miRNA target. *Genes Dev.* 27, 581–589 (2013).

364. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 49, D605–D612 (2021).

365. Gisselbrecht, S. S. *et al.* Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. *Nature Methods* 10, 774–780 (2013).

366. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* 512, 91–95 (2014).

367. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* 0, (2020).

368. Lagunas, T. *et al.* A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of genetic variants in vivo. *bioRxiv* 2021.05.17.444514 (2021) doi:10.1101/2021.05.17.444514.

369. Lambert, J. T. *et al.* Parallel functional testing identifies enhancers active in early postnatal mouse brain. *bioRxiv* 2021.01.15.426772 (2021) doi:10.1101/2021.01.15.426772.

370. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotech* 30, 265–270 (2012).

371. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502 (2006).

372. Rabani, M., Pieper, L., Chew, G.-L. & Schier, A. F. A Massively Parallel Reporter Assay of 3′ UTR Sequences Identifies In Vivo Rules for mRNA Degradation. *Molecular Cell* 68, 1083-1094.e5 (2017).

373. Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* 26, 238–255 (2016).

374. Smith, R. P. *et al.* A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design. *Genome Biology* 14, R72 (2013).

375. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics* 45, 1021–1028 (2013).

376. Visel, A. *et al.* Functional autonomy of distant-acting human enhancers. *Genomics* 93, 509–513 (2009).

377. Anania, C. *et al.* In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. *bioRxiv* 2021.04.14.439779 (2021) doi:10.1101/2021.04.14.439779.

378. Burger, A. *et al.* Maximizing mutagenesis with solubilized CRISPR-Cas9 ribonucleoprotein complexes. *Development* 143, 2025–2037 (2016).

379. Chen, H.-M. *et al.* CAMIO: a transgenic CRISPR pipeline to create diverse targeted genome deletions in Drosophila. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa177.

380. Gruner, H. N. *et al.* Precise removal of Calm1 long 3′ UTR isoform by CRISPR-Cas9 genome editing impairs dorsal root ganglion development in mice. *bioRxiv* 553990 (2019) doi:10.1101/553990.

381. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics* 48, 895–903 (2016).

382. Hörnblad, A., Bastide, S., Langenfeld, K., Langa, F. & Spitz, F. Dissection of the Fgf8 regulatory landscape by in vivo CRISPR-editing reveals extensive intra- and inter-enhancer redundancy. *Nature Communications* 12, 439 (2021).

383. Kroll, F. *et al.* A simple and effective F0 knockout method for rapid screening of behaviour and other complex phenotypes. *eLife* 10, e59683 (2021).

384. Labi, V. *et al.* Context-specific regulation of cell survival by a miRNA-controlled BIM rheostat. *Genes Dev.* 33, 1673–1687 (2019).

385. Miller, S. *et al.* Disruption of Dendritic Translation of CaMKIIα Impairs Stabilization of Synaptic Plasticity and Memory Consolidation. *Neuron* 36, 507–519 (2002).

386. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* 554, 239–243 (2018).

387. Parsch, J., Russell, J. A., Beerman, I., Hartl, D. L. & Stephan, W. Deletion of a conserved regulatory element in the Drosophila Adh gene leads to increased alcohol dehydrogenase activity but also delays development. *Genetics* 156, 219–227 (2000).

388. Perry, R. B.-T. *et al.* Subcellular Knockout of Importin β1 Perturbs Axonal Retrograde Signaling. *Neuron* 75, 294–305 (2012).

389. Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E. & Lippman, Z. B. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. *Cell* 171, 470-480.e8 (2017).

390. Shaw, D. K. & Mokalled, M. Efficient CRISPR/Cas9 mutagenesis for neurobehavioral screening in adult zebrafish. *bioRxiv* 2021.02.01.429280 (2021) doi:10.1101/2021.02.01.429280.

391. Terenzio, M. *et al.* Locally translated mTOR controls axonal local translation in nerve injury. *Science* 359, 1416–1421 (2018).

392. Wang, X. *et al.* Dissecting cis- regulatory control of quantitative trait variation in a plant stem cell circuit. *Nature Plants* 1–9 (2021) doi:10.1038/s41477-021-00898-x.

393. Aamodt, E. J., Chung, M. A. & McGhee, J. D. Spatial control of gut-specific gene expression during Caenorhabditis elegans development. *Science* 252, 579–582 (1991).

394. Ambros, V. & Horvitz, H. R. Heterochronic mutants of the nematode Caenorhabditis elegans. *Science* 226, 409–416 (1984).

395. Araya, C. L. *et al.* Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* 512, 400–405 (2014).

396. Chang, S., Johnston, R. J. & Hobert, O. A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of C. elegans. *Genes Dev.* 17, 2123–2137 (2003).

397. Crane, M. M. *et al.* Autonomous screening of C. elegans identifies genes implicated in synaptogenesis. *Nat Methods* 9, 977–980 (2012).

398. Daugherty, A. C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res.* 27, 2096–2107 (2017).

399. Driscoll, M., Dean, E., Reilly, E., Bergholz, E. & Chalfie, M. Genetic and molecular analysis of a Caenorhabditis elegans beta-tubulin that conveys benzimidazole sensitivity. *Journal of Cell Biology* 109, 2993–3003 (1989).

400. Durham, T. J. *et al.* Comprehensive characterization of tissue-specific chromatin accessibility in L2 C. elegans nematodes. *Genome Res.* gr.271791.120 (2021) doi:10.1101/gr.271791.120.

401. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 391, 806–811 (1998).

402. Friedman, D. B. & Johnson, T. E. A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics* 118, 75–86 (1988).

403. Gerstein, M. B. *et al.* Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science* 330, 1775–1787 (2010).

404. Hashimshony, T., Feder, M., Levin, M., Hall, B. K. & Yanai, I. Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature* 519, 219–222 (2015).

405. Ho, M. C. W., Quintero-Cadena, P. & Sternberg, P. W. Genome-wide discovery of active regulatory elements and transcription factor footprints in Caenorhabditis elegans using DNase-seq. *Genome Res.* 27, 2108–2119 (2017).

406. Hodgkin, J. A. & Brenner, S. Mutations causing transformation of sexual phenotype in the nematode Caenorhabditis elegans. *Genetics* 86, 275–287 (1977).

407. Horvitz, H. R. & Sulston, J. E. Isolation and genetic characterization of cell-lineage mutants of the nematode Caenorhabditis elegans. *Genetics* 96, 435–454 (1980).

408. Jan, E., Motzny, C. K., Graves, L. E. & Goodwin, E. B. The STAR protein, GLD-1, is a translational regulator of sexual identity in Caenorhabditis elegans. *The EMBO Journal* 18, 258–269 (1999).

409. Jungkamp, A.-C. *et al.* In vivo and transcriptome-wide identification of RNA binding protein target sites. *Mol. Cell* 44, 828–840 (2011).

410. Kaletsky, R. *et al.* Transcriptome analysis of adult Caenorhabditis elegans cells reveals tissue-specific gene and isoform expression. *PLOS Genetics* 14, e1007559 (2018).

411. Kamath, R. S. *et al.* Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 421, 231–237 (2003).

412. Kemphues, K. *Essential genes*. *WormBook: The Online Review of C. elegans Biology [Internet]* (WormBook, 2005).

413. Kemphues, K. J., Priess, J. R., Morton, D. G. & Cheng, N. Identification of genes required for cytoplasmic localization in early C. elegans embryos. *Cell* 52, 311–320 (1988).

414. Kershner, A. M. & Kimble, J. Genome-wide analysis of mRNA targets for Caenorhabditis elegans FBF, a conserved stem cell regulator. *PNAS* 107, 3936–3941 (2010).

415. Lewis, J. A., Wu, C.-H., Berg, H. & Levine, J. H. The Genetics of Levamisole Resistance in the Nematode Caenorhabditis Elegans. *Genetics* 95, 905–928 (1980).

416. McGrath, P. T. *et al.* Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes. *Nature* 477, 321–325 (2011).

417. Merritt, C., Rasoloson, D., Ko, D. & Seydoux, G. 3′ UTRs Are the Primary Regulators of Gene Expression in the C. elegans Germline. *Current Biology* 18, 1476–1482 (2008).

418. Mori, I. Genetics of Chemotaxis and Thermotaxis in the Nematode Caenorhabditis Elegans. *Annu. Rev. Genet.* 33, 399–422 (1999).

419. Niu, W. *et al.* Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. *Genome Res.* 21, 245–254 (2011).

420. O'Meara, M. M. *et al.* Cis-regulatory Mutations in the Caenorhabditis elegans Homeobox Gene Locus cog-1 Affect Neuronal Development. *Genetics* 181, 1679–1686 (2009).

421. O'Rourke, S. M. *et al.* A Survey of New Temperature-Sensitive, Embryonic-Lethal Mutations in C. elegans: 24 Alleles of Thirteen Genes. *PLOS ONE* 6, e16644 (2011).

422. Packer, J. S. *et al.* A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* eaax1971 (2019) doi:10.1126/science.aax1971.

423. Pulak, R. & Anderson, P. mRNA surveillance by the Caenorhabditis elegans smg genes. *Genes Dev.* 7, 1885–1897 (1993).

424. Rybak-Wolf, A. *et al.* A Variety of Dicer Substrates in Human and C. elegans. *Cell* 159, 1153–1167 (2014).

425. Sönnichsen, B. *et al.* Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature* 434, 462–469 (2005).

426. Spencer, W. C. *et al.* A spatial and temporal map of C. elegans gene expression. *Genome Res.* 21, 325–341 (2011).

427. Spencer, W. C. *et al.* Isolation of Specific Neurons from C. elegans Larvae for Gene Expression Profiling. *PLoS ONE* 9, e112102 (2014).

428. Timmermeyer, N. *et al.* An open source microfluidic sorter for Caenorhabditis nematodes. *bioRxiv* 780502 (2019) doi:10.1101/780502.

429. Troemel, E. R., Sagasti, A. & Bargmann, C. I. Lateral Signaling Mediated by Axon Contact and Calcium Entry Regulates Asymmetric Odorant Receptor Expression in C. elegans. *Cell* 99, 387–398 (1999).

430. Tursun, B., Patel, T., Kratsios, P. & Hobert, O. Direct Conversion of C. elegans Germ Cells into Specific Neuron Types. *Science* 331, 304–308 (2011).

431. Winston, W. M., Molodowitch, C. & Hunter, C. P. Systemic RNAi in C. elegans Requires the Putative Transmembrane Protein SID-1. *Science* 295, 2456–2459 (2002).

432. Zhang, B. *et al.* A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line. *Nature* 390, 477–484 (1997).

433. Zhou, K. I., Pincus, Z. & Slack, F. J. Longevity and stress in Caenorhabditis elegans. *Aging (Albany NY)* 3, 733–753 (2011).

434. Zisoulis, D. G. *et al.* Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat. Struct. Mol. Biol.* 17, 173–179 (2010).

435. Ahier, A., Suman, S. K. & Jarriault, S. Gene bashing of ceh-6 locus identifies genomic regions important for ceh-6 rectal cell expression and rescue of its mutant lethality. *microPublication Biology* 2020, (2020).

436. Bertrand, V. & Hobert, O. Linking Asymmetric Cell Division to the Terminal Differentiation Program of Postmitotic Neurons in C. elegans. *Developmental Cell* 16, 563–575 (2009).

437. Broitman-Maduro, G., Maduro, M. F. & Rothman, J. H. The Noncanonical Binding Site of the MED-1 GATA Factor Defines Differentially Regulated Target Genes in the C. elegans Mesendoderm. *Developmental Cell* 8, 427–433 (2005).

438. Cui, M. & Han, M. Cis regulatory requirements for vulval cell-specific expression of the caenorhabditis elegans fibroblast growth factor gene egl-17. *Developmental Biology* 257, 104–116 (2003).

439. Didiano, D. & Hobert, O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology* 13, 849–851 (2006).

440. Didiano, D. & Hobert, O. Molecular architecture of a miRNA-regulated 3' UTR. *RNA* 14, 1297–1317 (2008).

441. Egan, C. R. *et al.* A gut-to-pharynx/tail switch in embryonic expression of the Caenorhabditis elegans ges-1 gene centers on two GATA sequences. *Dev Biol* 170, 397–419 (1995).

442. Etchberger, J. F. *et al.* The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes Dev.* 21, 1653–1674 (2007).

443. Etchberger, J. F., Flowers, E. B., Poole, R. J., Bashllari, E. & Hobert, O. Cis-regulatory mechanisms of left/right asymmetric neuron-subtype specification in C. elegans. *Development* 136, 147–160 (2009).

444. Flames, N. & Hobert, O. Gene regulatory logic of dopamine neuron differentiation. *Nature* 458, 885–889 (2009).

445. Gaudet, J. & McGhee, J. D. Recent advances in understanding the molecular mechanisms regulating C. elegans transcription. *Developmental Dynamics* 239, 1388–1404 (2010).

446. Gilleard, J. S., Barry, J. D. & Johnstone, I. L. cis regulatory requirements for hypodermal cell-specific expression of the Caenorhabditis elegans cuticle collagen gene dpy-7. *Molecular and Cellular Biology* 17, 2301–2311 (1997).

447. Hwang, B. J. & Sternberg, P. W. A cell-specific enhancer that specifies lin-3 expression in the C. elegans anchor cell for vulval development. *Development* 131, 143–151 (2004).

448. Johnson, A. D., Fitzsimmons, D., Hagman, J. & Chamberlin, H. M. EGL-38 Pax regulates the ovo-related gene lin-48 during Caenorhabditis elegans organ development. *Development* 128, 2857–2865 (2001).

449. Johnson, S. M., Lin, S.-Y. & Slack, F. J. The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Developmental Biology* 259, 364–379 (2003).

450. Kirouac, M. & Sternberg, P. W. cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of Caenorhabditis elegans and C. briggsae. *Dev Biol* 257, 85–103 (2003).

451. Krause, M., Harrison, S. W., Xu, S.-Q., Chen, L. & Fire, A. Elements Regulating Cell- and Stage-Specific Expression of the C. elegans MyoD Family Homolog hlh-1. *Developmental Biology* 166, 133–148 (1994).

452. Liu, Z., Kirch, S. & Ambros, V. The Caenorhabditis elegans heterochronic gene pathway controls stage-specific transcription of collagen genes. *Development* 121, 2471–2478 (1995).

453. Lloret-Fernández, C. *et al.* A transcription factor collective defines the HSN serotonergic neuron regulatory landscape. *eLife* 7, e32785 (2018).

454. MacMorris, M. *et al.* Regulation of vitellogenin gene expression in transgenic Caenorhabditis elegans: short sequences required for activation of the vit-2 promoter. *Molecular and Cellular Biology* 12, 1652–1662 (1992).

455. Mah, A. K. *et al.* Transcriptional Regulation of AQP-8, a Caenorhabditis elegans Aquaporin Exclusively Expressed in the Excretory System, by the POU Homeobox Transcription Factor CEH-6 *. *Journal of Biological Chemistry* 282, 28074–28086 (2007).

456. Marri, S. & Gupta, B. P. Dissection of lin-11 enhancer regions in Caenorhabditis elegans and other nematodes. *Dev Biol* 325, 402–411 (2009).

457. Merritt, C. & Seydoux, G. The Puf RNA-binding proteins FBF-1 and FBF-2 inhibit the expression of synaptonemal complex proteins in germline stem cells. *Development* 137, 1787–1798 (2010).

458. Mikl, M. & Cowan, C. R. Alternative 3′ UTR Selection Controls PAR-5 Homeostasis and Cell Polarity in C. elegans Embryos. *Cell Reports* 8, 1380–1390 (2014).

459. Moilanen, L. H., Fukushige, T. & Freedman, J. H. Regulation of metallothionein gene transcription. Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from Caenorhabditis elegans. *J Biol Chem* 274, 29655–29665 (1999).

460. Nolde, M. J., Saka, N., Reinert, K. L. & Slack, F. J. The C. elegans pumilio homolog, puf-9, is required for the 3'UTR mediated repression of the let-7 microRNA target gene, hbl-1. *Dev Biol* 305, 551–563 (2007).

461. Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A. & Fire, A. Sequence requirements for myosin gene expression and regulation in Caenorhabditis elegans. *Genetics* 135, 385–404 (1993).

462. Serrano-Saiz, E. *et al.* Modular Control of Glutamatergic Neuronal Identity in C. elegans by Distinct Homeodomain Proteins. *Cell* 155, 659–673 (2013).

463. Serrano-Saiz, E. *et al.* Modular Organization of Cis-regulatory Control Information of Neurotransmitter Pathway Genes in Caenorhabditis elegans. *Genetics* 215, 665–681 (2020).

464. Sharifnia, P., Kim, K. W., Wu, Z. & Jin, Y. Distinct cis elements in the 3' UTR of the C. elegans cebp-1 mRNA mediate its regulation in neuronal development. *Dev Biol* 429, 240–248 (2017).

465. Stec, N. *et al.* An Epigenetic Priming Mechanism Mediated by Nutrient Sensing Regulates Transcriptional Output during C. elegans Development. *Current Biology* 0, (2020).

466. Stefanakis, N., Carrera, I. & Hobert, O. Regulatory Logic of Pan-Neuronal Gene Expression in C. elegans. *Neuron* 87, 733–750 (2015).

467. Teng, Y., Girard, L., Ferreira, H. B., Sternberg, P. W. & Emmons, S. W. Dissection of cis-regulatory elements in the C. elegans Hox gene egl-5 promoter. *Developmental Biology* 276, 476–492 (2004).

468. Wagmaister, J. A. *et al.* Identification of cis-regulatory elements from the C. elegans Hox gene lin-39 required for embryonic expression and for regulation by the transcription factors LIN-1, LIN-31 and LIN-39. *Developmental Biology* 297, 550–565 (2006).

469. Wang, X. & Chamberlin, H. M. Multiple regulatory changes contribute to the evolution of the Caenorhabditis lin-48 ovo gene. *Genes Dev.* 16, 2345–2349 (2002).

470. Wang, L. & Way, J. C. Promoter sequences for the establishment of mec-3 expression in the nematode Caenorhabditis elegans. *Mechanisms of Development* 56, 183–196 (1996).

471. Wang, X., Ellenbecker, M., Hickey, B., Day, N. J. & Voronina, E. PUF family proteins FBF-1 and FBF-2 regulate germline stem and progenitor cell proliferation and differentiation in C. elegans. *bioRxiv* 825984 (2019) doi:10.1101/825984.

472. Way, J. C., Wang, L., Run, J. Q. & Wang, A. The mec-3 gene contains cis-acting elements mediating positive and negative regulation in cells produced by asymmetric cell division in Caenorhabditis elegans. *Genes Dev* 5, 2199–2211 (1991).

473. Wenick, A. S. & Hobert, O. Genomic cis-Regulatory Architecture and trans-Acting Regulators of a Single Interneuron-Specific Gene Battery in C. elegans. *Developmental Cell* 6, 757–770 (2004).

474. Yan, D., Wu, Z., Chisholm, A. D. & Jin, Y. The DLK-1 kinase promotes mRNA stability and local translation in C. elegans synapses and axon regeneration. *Cell* 138, 1005–1018 (2009).

475. Yin, J., Madaan, U., Park, A., Aftab, N. & Savage-Dunn, C. Multiple cis elements and GATA factors regulate a cuticle collagen gene in Caenorhabditis elegans. *genesis* 53, 278–284 (2015).

476. Alper, S. & Kenyon, C. The zinc finger protein REF-2 functions with the Hox genes to inhibit cell fusion in the ventral epidermis of C. elegans. *Development* 129, 3335–3348 (2002).

477. Arata, Y. *et al.* Wnt Signaling and a Hox Protein Cooperatively Regulate PSA-3/Meis to Determine Daughter Cell Fate after Asymmetric Cell Division in C. elegans. *Developmental Cell* 11, 105–115 (2006).

478. Barton, M. K., Schedl, T. B. & Kimble, J. Gain-of-function mutations of fem-3, a sex-determination gene in Caenorhabditis elegans. *Genetics* 115, 107–119 (1987).

479. Clark, S. G., Chisholm, A. D. & Horvitz, H. R. Control of cell fates in the central body region of C. elegans by the homeobox gene lin-39. *Cell* 74, 43–55 (1993).

480. Desai, C., Garriga, G., Mclntire, S. L. & Horvitz, H. R. A genetic pathway for the development of the Caenorhabditis elegans HSN motor neurons. *Nature* 336, 638–646 (1988).

481. Detwiler, M. R., Reuben, M., Li, X., Rogers, E. & Lin, R. Two Zinc Finger Proteins, OMA-1 and OMA-2, Are Redundantly Required for Oocyte Maturation in C. elegans. *Developmental Cell* 1, 187–199 (2001).

482. Fay, D. S. *et al.* The coordinate regulation of pharyngeal development in C. elegans by lin-35/Rb, pha-1, and ubc-18. *Developmental Biology* 271, 11–25 (2004).

483. Furuta, T. *et al.* EMB-30: An APC4 Homologue Required for Metaphase-to-Anaphase Transitions during Meiosis and Mitosis in Caenorhabditis elegans. *Mol Biol Cell* 11, 1401–1419 (2000).

484. Garbe, D., Doto, J. B. & Sundaram, M. V. Caenorhabditis elegans lin-35/Rb, efl-1/E2F and other synthetic multivulva genes negatively regulate the anaphase-promoting complex gene mat-3/APC8. *Genetics* 167, 663–672 (2004).

485. Gleason, E. J., Lindsey, W. C., Kroft, T. L., Singson, A. W. & L'Hernault, S. W. spe-10 Encodes a DHHC–CRD Zinc-Finger Membrane Protein Required for Endoplasmic Reticulum/Golgi Membrane Morphogenesis During Caenorhabditis elegans Spermatogenesis. *Genetics* 172, 145–158 (2006).

486. Goodwin, E. B., Okkema, P. G., Evans, T. C. & Kimble, J. Translational regulation of tra-2 by its 3′ untranslated region controls sexual identity in C. elegans. *Cell* 75, 329–339 (1993).

487. Harris, J. *et al.* Mutator Phenotype of Caenorhabditis elegans DNA Damage Checkpoint Mutants. *Genetics* 174, 601–616 (2006).

488. Hirose, T., Galvin, B. D. & Horvitz, H. R. Six and Eya promote apoptosis through direct transcriptional activation of the proapoptotic BH3-only gene egl-1 in Caenorhabditis elegans. *PNAS* 107, 15479–15484 (2010).

489. Hodgkin, J. A. Genetic and Anatomical Aspects of the Caenorhabditis elegans Male. (University of Cambridge, 1974).

490. Hong, K., Mano, I. & Driscoll, M. In Vivo Structure–Function Analyses ofCaenorhabditis elegans MEC-4, a Candidate Mechanosensory Ion Channel Subunit. *J Neurosci* 20, 2575–2588 (2000).

491. Hu, P. J., Xu, J. & Ruvkun, G. Two Membrane-Associated Tyrosine Phosphatase Homologs Potentiate C. elegans AKT-1/PKB Signaling. *PLOS Genetics* 2, e99 (2006).

492. Jones, A. R. & Schedl, T. Mutations in gld-1, a female germ cell-specific tumor suppressor gene in Caenorhabditis elegans, affect a conserved domain also found in Src-associated protein Sam68. *Genes Dev.* 9, 1491–1504 (1995).

493. Kemp, C. A., Song, M. H., Addepalli, M. K., Hunter, G. & O'Connell, K. Suppressors of zyg-1 Define Regulators of Centrosome Duplication and Nuclear Association in Caenorhabditis elegans. *Genetics* 176, 95–113 (2007).

494. Li, M., Jones-Rhoades, M. W., Lau, N. C., Bartel, D. P. & Rougvie, A. E. Regulatory Mutations of mir-48, a C. elegans let-7 Family MicroRNA, Cause Developmental Timing Defects. *Developmental Cell* 9, 415–422 (2005).

495. Lynch, T. R., Xue, M., Czerniak, C. W., Lee, C. & Kimble, J. *Notch-dependent DNA cis-regulatory elements and their dose-dependent control of C. elegans stem cell self-renewal*. 2021.11.09.467950 https://www.biorxiv.org/content/10.1101/2021.11.09.467950v2 (2021) doi:10.1101/2021.11.09.467950.

496. Miller, L. M., Hess, H. A., Doroquez, D. B. & Andrews, N. M. Null Mutations in the lin-31 Gene Indicate Two Functions During Caenorhabditis elegans Vulval Development. *Genetics* 156, 1595–1602 (2000).

497. Miska, E. A. *et al.* Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS Genet* 3, e215 (2007).

498. Moorman, C. & Plasterk, R. H. A. Functional Characterization of the Adenylyl Cyclase Gene sgs-1 by Analysis of a Mutational Spectrum in Caenorhabditis elegans. *Genetics* 161, 133–142 (2002).

499. Nakagawa, A., Sullivan, K. D. & Xue, D. Caspase-activated phosphoinositide binding by CNT-1 promotes apoptosis by inhibiting the AKT pathway. *Nat Struct Mol Biol* 21, 1082–1090 (2014).

500. Ogura, K. *et al.* Caenorhabditis elegans unc-51 gene required for axonal elongation encodes a novel serine/threonine kinase. *Genes Dev* 8, 2389–2400 (1994).

501. Oishi, K., Okano, H. & Sawa, H. RMD-1, a novel microtubule-associated protein, functions in chromosome segregation in Caenorhabditis elegans. *J Cell Biol* 179, 1149–1162 (2007).

502. Partridge, F. A., Tearle, A. W., Gravato-Nobre, M. J., Schafer, W. R. & Hodgkin, J. The C. elegans glycosyltransferase BUS-8 has two distinct and essential roles in epidermal morphogenesis. *Developmental Biology* 317, 549–559 (2008).

503. Perry, M. D., Trent, C., Robertson, B., Chamblin, C. & Wood, W. B. Sequenced alleles of the Caenorhabditis elegans sex-determining gene her-1 include a novel class of conditional promoter mutations. *Genetics* 138, 317–327 (1994).

504. Ren, H. & Zhang, H. Wnt signaling controls temporal identities of seam cells in Caenorhabditis elegans. *Developmental Biology* 345, 144–155 (2010).

505. Ross, J. M., Kalis, A. K., Murphy, M. W. & Zarkower, D. The DM Domain Protein MAB-3 Promotes Sex-Specific Neurogenesis in C. elegans by Regulating bHLH Proteins. *Developmental Cell* 8, 881–892 (2005).

506. Saffer, A. M., Kim, D. H., Oudenaarden, A. van & Horvitz, H. R. The Caenorhabditis elegans Synthetic Multivulva Genes Prevent Ras Pathway Activation by Tightly Repressing Global Ectopic Expression of lin-3 EGF. *PLOS Genetics* 7, e1002418 (2011).

507. Sarin, S. *et al.* Genetic screens for Caenorhabditis elegans mutants defective in left/right asymmetric neuronal fate specification. *Genetics* 176, 2109–2130 (2007).

508. Sarin, S. *et al.* Analysis of Multiple Ethyl Methanesulfonate-Mutagenized Caenorhabditis elegans Strains by Whole-Genome Sequencing. *Genetics* 185, 417–430 (2010).

509. Thacker, C., Sheps, J. A. & Rose, A. M. Caenorhabditis elegans dpy-5 is a cuticle procollagen processed by a proprotein convertase. *Cell Mol Life Sci* 63, 1193–1204 (2006).

510. Trent, C., Wood, W. B. & Horvitz, H. R. A novel dominant transformer allele of the sex-determining gene her-1 of Caenorhabditis elegans. *Genetics* 120, 145–157 (1988).

511. Verghese, E. *et al.* The tailless ortholog nhr-67 functions in the development of the C. elegans ventral uterus. *Dev Biol* 356, 516–528 (2011).

512. Wen, C., Levitan, D., Li, X. & Greenwald, I. spr-2, a suppressor of the egg-laying defect caused by loss of sel-12 presenilin in Caenorhabditiselegans, is a member of the SET protein subfamily. *PNAS* 97, 14524–14529 (2000).

513. William F Hanna & Victor R Ambros. Identification of Genes Involved in the Early Heterochronic Genetic Circuit via RNAi-by-Feeding presented in International Worm Meeting. in (2003).

514. Zarkower, D. & Hodgkin, J. Molecular analysis of the C. elegans sex-determining gene tra-1: a gene encoding two zinc finger proteins. *Cell* 70, 237–249 (1992).

515. Zarkower, D., Bono, M. D., Aronoff, R. & Hodgkin, J. Regulatory rearrangements and smg-sensitive alleles of the C. elegans sex-determining gene tra-1. *Developmental Genetics* 15, 240–250 (1994).