Parallel genetics of regulatory sequences in vivo

Jonathan Froehlich1,3, Bora Uyar2,3, Margareta Herzog1, Kathrin Theil1, Petar Glažar1, Altuna Akalin2, Nikolaus Rajewsky1

Abstract
Understanding how regulatory sequences control gene expression is fundamental to explain how phenotypes arise in health and disease. Traditional reporter assays inform about function of individual regulatory elements, typically in isolation. However, regulatory elements must ultimately be understood by perturbing them within their genomic environment and developmental- or tissue-specific contexts. This is technically challenging; therefore, few regulatory elements have been characterized in vivo. Here, we used inducible Cas9 and multiplexed guide RNAs to create hundreds of mutations in enhancers/promoters and 3’ UTRs of 16 genes in C. elegans. To quantify the consequences of mutations on expression, we developed a targeted RNA sequencing strategy across hundreds of mutant animals. We were also able to systematically and quantitatively assign fitness cost to mutations. Finally, we identified and characterized sequence elements that strongly regulate phenotypic traits. Our approach enables highly parallelized, functional analysis of regulatory sequences in vivo.

Introduction
Understanding gene regulation is fundamental for understanding development and tissue function in health and disease. Animal genomes contain diverse regulatory sequences, which are organized in contiguous stretches of genomic DNA, ranging from a few to hundreds or thousands of bases. Promoters, enhancers, and silencers act mainly on transcription while transcribed sequences (including 5’ and 3’ UTRs) regulate splicing, export, localization, degradation, and translation of mRNAs. Many regulatory sequences encode multiple regulatory functions which can cooperate, compensate and antagonize each other1–3. Understanding this logic requires combinatorial perturbations. Moreover, a single binding site, due to fuzzy recognition motifs, may tolerate certain mutations4–6. The interaction of effectors with regulatory elements can additionally be modulated by sequence structure2,5,7, co-factors2,3, chemical modifications and the temporal order of binding1,3. Sequence activity is therefore dependent on native sequence context, cell type, tissue, development and the environment. Accordingly, phenotypic consequences of mutations in regulatory regions are difficult to predict. To understand biological functions, an approach to target regulatory sequences in vivo with many different mutations is required.

Although parallel interrogation of regulatory sequences has been developed in cell lines and yeast8–12; only a few in vivo approaches have been achieved in animal models. These use integration of reporters13,14 or injection of RNA libraries15,16 and therefore do not evaluate endogenous phenotypes, or are restricted to one stage of the animal life cycle. Classical genome editing by injection, now widely accessible due to CRISPR-Cas17,18, has enabled endogenous functional tests, but is still work-intensive and limited in scalability19,20.

Here we use inducible expression of Cas9 and multiplexed single guide RNAs in C. elegans populations to generate hundreds of targeted mutations in parallel. We targeted different regulatory regions across 16 genes and analyzed >12,000 Cas9-induced mutations to first describe characteristics of dsDNA break repair in the C. elegans germline and the genotype diversity that this introduces at the targeted loci. We then applied our mutagenesis approach to generate hundreds of deletions along the well-studied lin-41 3’ UTR, which is targeted by the miRNA let-71,24. We developed an RNA-sequencing-based strategy to quantify the impact of each mutation on lin-41 RNA level. Using RNA sequencing we followed the relative abundance of these different mutations over several generations to infer their phenotype. This analysis revealed a previously undescribed compensatory function between the two let-7 miRNA binding sites. Finally, we couple the targeted mutagenesis of regulatory sequences to selection of phenotypic traits. We isolate 57 reduction-of-function alleles in 3 genes that show strong changes in phenotype, mediated by mutations in enhancer, TATA-box, and 3’ UTRs which act by both loss and gain of gene regulation.

Results
Inducible Cas9 to generate hundreds of mutants in parallel. To introduce many different, targeted mutations in vivo, we developed an approach in C. elegans using inducible expression of Cas9 and several multiplexed “single guide RNAs” (sgRNAs). This required only few initial injections to create transgenic animals, allowed maintenance without mutagenesis, and enabled time-controlled creation of mutated populations in parallel, with sizes only limited by culturing approaches (up to ~107 in our case). Mutant populations could then be used for various purposes. For example, they could be selected by phenotype or reporter activity, or directly analyzed by targeted sequencing to measure the impact of mutations on RNA regulation or fitness (Fig. 1a).

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

1 Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str 28, 10115 Berlin, Germany.
2 Bioinformatics and Omics Data Science Platform, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Hannoversche Str 28, 10115 Berlin, Germany.
3 These authors contributed equally. *e-mail: rajewsky@mdc-berlin.de
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 ~50% (sg1) and ~20% (sg2) of eggs producing GFP negative animals (Supplementary Fig. 1a). We obtained similar results when targeting the dpy-10 gene and counting the characteristic Dumpy (Dpy) phenotype with the highest fractions 12 – 15 hours after heat-shock, and ~20 – 35% of eggs producing Dpy animals (Supplementary Fig.1b).

Characteristic CRISPR-Cas9-induced mutations from 91 GFP negative animals consisted of deletions, insertions or a combination of both and originated from sgRNA cut sites (Supplementary Fig. 1c, d). 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 (Supplementary Fig. 1d). 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 (Supplementary Fig. 1e).

Homozygous animals will 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 (Supplementary Fig. 1f). We therefore wanted to assess how many germline mutations were generated. For this we analyzed the inheritance of GFP negative animals from F1 to F2 generations using an automated flow system and found that ~80% of mutations were indeed germline mutations (Supplementary Fig. 1g-i). For the rest of our work we used such non-mosaic F2, generated by F1 germline mutations.

To analyze large mutated C. elegans populations in bulk, we established a targeted sequencing protocol based on long 0.5-3 kb PCR amplicons. It allowed us to sequence the complete targeted locus, to capture large deletions and to multiplex samples for sequencing (Supplementary Fig. 2a). This resulted in ~200,000 – 800,000-fold read coverage at the targeted regions. We also created a software pipeline (“crispr-DART” for “CRISPR-Cas Downstream Analysis and Reporting Tool”) to handle targeted sequencing data of such amplicons and analyze the contained indels. The pipeline works with various targeted sequencing
technologies and extracts and quantifies indels. The output contains html reports of coverage, mutation profiles, sgRNA efficiencies and optional comparisons between two samples. Processed genomics files from the output can then be used for more in-depth custom analyses with additionally supplied R scripts (Supplementary Fig. 2b-d) (“Code Availability” in Supplementary Methods).

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. 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”) (Fig. 1b). These occurred within a few nucleotides to cut sites and were less frequent than deletions (~1/2 – 1/10th) (Fig. 1c).

**Features of CRISPR-Cas9-induced indels and genotype diversity.** To understand gene regulatory logic, ideally many different variants are produced and tested for their effects in vivo. For this we set out to analyze the genotype diversity 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 strong organismal phenotypes, and 7 genes with lethal phenotypes. After Cas9 heat shock-induction we sequenced bulk genomic DNA from resultant 400,000 F2 animals with long amplicon sequencing. Together with 50,000 P0 animals by heat shock, and amplicon sequenced with lethal phenotypes. After Cas9 heat shock-induction we sequenced bulk genomic DNA from resultant 400,000 F2 animals with long amplicon sequencing. Together with wild type controls this produced 60 samples and data of 12,000 indel mutations for 91 sgRNAs (Supplementary Table 1). To measure sgRNA efficiencies, we quantified deletions +/-5bp bracketing a given sgRNA cut site. The median efficiency was ~1.4% with most sgRNAs showing efficiencies ~0 – 6.3% (95% CI) (Fig. 2a). 1.4% corresponded to approximately 5,600 mutant animals per sgRNA in our samples. We could not determine factors associated with the efficiency of sgRNAs in our system but found that lethal phenotypes were not confounding this analysis (Supplementary Fig. 3a-d). We then used the detected >12,000 indel mutations to characterize CRISPR-Cas9-induced dsDNA break repair outcomes in the *C. elegans* germline. We analyzed the proportion of mutation types for each sample. On average, samples contained 60% deletions, 25% insertions and 15% complex events (Fig. 2b). These proportions are similar for naturally occurring indels in *C. elegans* (75% deletions, 25% insertions) and human (50% deletions, 35% insertions), suggesting that natural and CRISPR-Cas9-induced dsDNA breaks undergo similar repair mechanisms.

We computationally separated deletions into those originating from a single sgRNA (“single-cut”) or from two or more sgRNAs (“multi-cut”) (Supplementary Fig. 3e). The length of single-cut deletions ranged from 1 to over 100 bp, with the majority being around 5 – 25 bp. Multi-cut deletions were larger, mostly several hundred bp, as expected from the spacing between multiplexed sgRNAs (Fig. 2c). Most (>90%) insertions were 1 – 20 bp long although we could find insertions up to 45 bp (Fig. 2d). These length distributions were similar to our observations made by Sanger sequencing (Supplementary Fig. 1e).

Insertion and deletion lengths reported from human cell lines are dramatically shorter, with the majority being 1 or 2 bp27-30. Longer indels in our samples can likely be explained by a higher activity of microhomology-mediated end joining (MMEJ) which uses 5 – 25 bp microhomologies and which has been reported as the main dsDNA break repair pathway in *C. elegans*31.

Inspection of insertions from individual genotypes revealed that most contained short sequences found in close vicinity to the insertion position (Supplementary Fig. 4a). Using our deep sequencing data, we systematically analyzed matches between insertions and the surrounding regions. 5-mers from insertions matched to sequences in a window +/- 13 bp around the insertion position and only in the same orientation (Supplementary Fig. 4b-d). Thus, our data indicate that many insertions are duplications of surrounding microhomologous sequences in the same orientation, likely the product of dissociation and re-annealing during MMEJ (Supplementary Fig. 4e).

Finally, we assessed the diversity of the generated mutations. 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 deletions (rows in Fig. 2e). Most deletions covered a region 10 – 12 bp surrounding the cut sites. On average, every sgRNA could generate 15 unique deletions per bp at the center of the cut site and up to an additional 5 unique deletions 5 bp away from the cut site (black line profile in Fig. 2e). We then studied multi-cut events. Naturally, these deletions spanned much larger regions, as defined by the spacing between two or more sgRNAs. Here we found up to 200 unique deletions per base pair and on average 20 unique deletions per sgRNA covering a region >500 bp surrounding each cut site (Fig. 2f). We then considered each unique indel a genotype (>0.001% frequency and >5 reads). On average, one sgRNA created 50 deletion genotypes and 10 insertion genotypes. However, some sgRNAs created up to 400 genotypes (Fig. 2g). Since we used several sgRNAs per transgenic line, we observed on average ~200 insertion and deletion genotypes per sample and in efficient lines up to 1700 deletion and 1200 insertion genotypes (Supplementary Fig. 5a). More efficient sgRNAs resulted in a higher number of unique deletions (Supplementary Fig. 5b). 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 (Supplementary Fig. 5c).

Together these data show that inducible expression of Cas9 with multiplexed sgRNAs can induce a variety of mutations to study hundreds of regulatory variants in parallel. This includes small deletions to target individual elements at nucleotide resolution, large deletions to interrogate combinatorial interactions and insertions to change spacing between sites or duplicate existing sequences.

---

*Froehlich, Uyar et al. | bioRxiv | July 2020 | 3*
Fig. 2 | Features of sgRNA-induced mutations. Pooled data from 60 experiments (24 wild type controls, 36 samples with induced Cas9), each expressing 1-8 sgRNAs targeting one region among 16 genes (for an overview of samples see Supplementary Table 1). a, Efficiency measured for each sgRNA (n=127) per experiment. This is calculated by counting the fraction of reads containing deletions +/- 5 bp around the sgRNA cut site. b, 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 substitutions which suggest a combination of multiple events. c,d, Length distribution of mutations supported by >5 reads and >0.001% frequency, found in all experiments. Deletions were classified as multi cut deletions (n=2915) when a deletion overlapped with more than one sgRNA cut site +/- 5bp or otherwise were classified as single cut deletions (n=3169) (for a scheme describing this see Supplementary Fig. 2). Insertions supported by at least 5 reads with the exact insertion sequence at the same genomic coordinate and a frequency >0.001% (n=6616). e,f, Unique deletions (>0.001% frequency and >5 reads) per bp. Each row shows deletions per bp for one sgRNA (n=86 sgRNAs); the black curve on the bottom represents average unique deletions per nt and sgRNA. g, Unique genotypes detected per sgRNA in 400 k sequenced worms per sgRNA. Ctrls n=76 cut sites, samples n=86 cut sites. Wilcoxon, p < 2.2e−16 for deletions, p < 2.2e−16 for insertions.
Quantifying mutant gene expression and fitness. A major challenge to understand gene regulation is the interaction of different 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 (LCS1 and LCS2, 22 and 20 nucleotides long, separated by a 27 nt spacer)21,22,24,32,33 (Supplementary Fig. 6a). Although studies with reporter plasmids showed that each binding site could not function on its own22, other studies concluded that each site could recapitulate wild-type regulation when present in three copies23. Prior to our work, compensation had not been tested for each site, in the native sequence context, or at natural expression levels. Therefore, we targeted the lin-41 3′ UTR with a pool of 8 sgRNAs and two sgRNA pairs close to the LCSs (Fig. 3a). Lin-41 down-regulation occurs with let-7 expression in the developmental stages L3-L421,32–34. To measure let-7-dependent regulation, we collected RNA from bulk worms at L1 and L4 stages. We extracted L4 stage RNA after complete lin-41 mRNA downregulation by let-7 and before the occurrence of the lethal vulva bursting phenotype24 (Supplementary Fig. 6b) (see Methods). We then sequenced lin-41-specific cDNA with long reads covering the complete 3′ UTR (Supplementary Fig. 6c). Each read contained full information on any deletion present in the RNA molecule, while the number of reads supporting each deletion could be used to estimate RNA expression level. To determine let-7-dependent effects, we then analyzed how different deletions affect RNA abundance at L4- relative to L1 stage.

We observed an average 4-fold up-regulation of lin-41 at larval stage L4 only when both let-7 seed sites were mutated by deletions (Fig. 3b). No significant effect was observed from deletions affecting either LCS1 or LCS2 seeds alone. The 4-fold regulatory effect is consistent with the known magnitude of down-regulation in the natural context22,32,33 or the up-regulation when disrupting both let-7 interactions24,36,37 (2 – 4-fold). A very slight and non-significant up-regulation was observed for LCS1. Interestingly, LCS1 has a longer seed pairing (8 nt) than LCS2 (6 nt and a G-U pair) and was detected with more reads in an in vivo miRNA proximity ligation approach38.

Fig. 3 | Analysis of the lin-41 3′ UTR shows redundancy of the two let-7 miRNA binding sites for RNA regulation and fitness. a, The lin-41 3′ UTR locus after targeted mutagenesis with three different lines (sg pool, sg1516, sg2627). Deletions of three lines were pooled and analyzed together. Deletion events (n>900) supported by >0.001% reads. b, Relative fold change of deletions detected in lin-41 3′ UTR full-length cDNA between L1 and L4 developmental stages. We considered only deletions that are detected in both L1 and L4 RNA samples. c, Generations were separated by sedimentation and worms were sampled for DNA extraction. d, Reads supporting deletions in bulk genomic DNA of consecutive generations, relative to the first (F2) generation. At each generation shown are all six samples (sg pool, sg1516, sg2627 grown at 16 and 24°C).

---

Froehlich, Uyar et al. | bioRxiv | July 2020 | 5
addition 3xLCS1 acted stronger than 3xLCS2 in a reporter assay\cite{33}. As an independent approach, we used UMAP on the k-mer content of long cDNA reads to obtain clusters of genotypes\cite{34}. These data also suggest that RNA molecules transcribed from clusters with large deletions overlapping both LCS seeds were detected with more reads in L4 stage compared to L1 stage animals (Supplementary Fig. 6d-f).

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 (Fig. 3c). Disrupting let-7 regulation of lin-41 is known to result in lethal developmental defects\cite{35,36,37,38,39}. Deletions in the lin-41 3' UTR which overlapped with both LCS seeds quickly disappeared from the population and were mostly absent already after one generation (Fig. 3d). Consistent with our observed absence of an effect on RNA expression, deletions affecting either one of the two LCS alone were not significantly depleted. Our results indicate that the two let-7 complementary sites can largely compensate each other’s loss under laboratory conditions. We conclude that parallel deletions affecting either one of the two LCS alone were not significantly depleted. Our results indicate that the two let-7 complementary sites can largely compensate each other’s loss under laboratory conditions. We conclude that parallel deletions affecting either one of the two LCS alone were not significantly depleted. Our results indicate that the two let-7 complementary sites can largely compensate each other’s loss under laboratory conditions.

Gene regulatory sequences which affect phenotypic traits. Next, we wanted to isolate regulatory sequence variants which affected animal phenotype. This could be useful to discover functional elements, provide starting points to study mechanisms, and to obtain animals with desired phenotypic traits. Such an approach would also capture any functional sequences regardless of the level, time or tissue of regulation, which can be hard to predict. We proceeded to select worms based on strong organismal phenotypes affecting animal movement and body shape (Unc, Rol, Dpy) and to identify the causative mutations. We targeted a predicted enhancer\cite{40}, three promoters and all 3' UTRs of 8 genes with known loss-of-function phenotypes and screened 35,000 – 45,000 animals for each of these regions (Supplementary Table 2). To determine which mutations were initially present in the screened population, we performed targeted sequencing on siblings (Supplementary Fig. 8a, b). Initially, we isolated several mutants with large deletions (>500 bp) that disrupted the coding frame or the polyadenylation signal (AATAAA) (Supplementary Fig. 8c, d). Similar large-scale, on-target deletions have previously been described in cell lines\cite{41,42} and mice\cite{43}. We also found large insertions (<250 bp) which originated from within +/-1 kb of the targeted region, or from loci on other chromosomes (Supplementary Fig. 8c, d). 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

![Fig. 4](image-url) A screen finds regulatory mutations strongly affecting 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, Eleven mutations along the egl-30 3' UTR which show slight or strong Uncordinated (Unc) phenotypes. b, Thirteen mutations upstream of sqt-3 which show a Roller (Rol) phenotype. Most (11/13) overlap with the predicted TATA-box. c, Mutations in the sqt-3 3' UTR which show a Rol phenotype or which are tolerated (non-Rol). Almost all (25/26) Rol mutants contained insertions. d, Fifteen mutations, mostly deletions, which rescue the Rol phenotype of one insertion allele sqt-3(ins). While most overlapped with the insertion isolated for a Rol phenotype, three deletions affected a region upstream (black bars on the bottom indicate the uncovered compensatory interaction).
animals with affected phenotypes (Supplementary Table 2).

From the screen we isolated 57 reduction-of-function alleles in 3 genes (egl-30, sqt-2, sqt-3) and none from the other 5 genes (dpy-2, dpy-10, rol-6, unc-26, unc-54) (Supplementary Table 2). Deletions, insertions and complex mutations were represented equally among isolates (Supplementary Fig. 8e). 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 lead to Uncordinated (Unc) phenotypes. In 7/11 mutants, a region circa 100 bp downstream of the STOP codon was affected. The smallest deletion was 6 bp (Fig. 4a) (Supplementary Fig. 8g). We also found mutations with Roller (Rol) phenotypes overlapping a putative sqt-2 enhancer predicted from chromatin accessibility profiling51 (Supplementary Fig. 8f). We also targeted sqt-3, a gene associated with three distinct morphological traits52,53 (Dpy, Rol and Lon). In total we isolated 39 alleles. 13 mutations upstream of sqt-3 likely affected transcriptional initiation, with 11/13 overlapping the predicted TATA-box (Fig. 4b). In line with the Rol phenotype, which indicates reduction-of-function, pre-mRNA and mRNA levels were both reduced to ~50% (Supplementary Fig. 9d). The remaining sqt-3 alleles were 3′ UTR mutations. Almost all (25/26) were insertions or complex mutations originating at sg2 (Fig. 4c). The only deletion overlapped with the polyadenylation signal (AAUUAAA). We knew from sequencing siblings that sg2 was very efficient (~25%) and that in our samples various deletions covering the 3′ UTR were present. We therefore isolated 13 distinct non-Rol mutants using direct PCR screening (Supplementary Fig. 9a, b). Despite containing deletions or insertions originating at the efficient sg2, these animals showed the wild type non-Rol trait (Fig. 4d). 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 ~50% (Supplementary Fig. 9c, d). Since deletions 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. Interestingly the polyA mutant sqt-3(polyA), for which mRNA levels were equally reduced to 50%, showed a much weaker Rol phenotype, with only slight bending of the head (Fig. 4e, Supplementary Fig. 9d, e). This suggests that additional mechanisms besides mRNA down-regulation might reduce protein output in sqt-3(ins).

To define the repressive sequence elements, we targeted the inserted sequence with several sgRNAs and screened for non-Rol “revertants” in which the wild type trait was restored. 12/13 revertants contained deletions overlapping with the insertion, with the smallest being 5 bp (Fig. 4d). A restored wild type trait likely resulted from restored expression levels and indeed mRNA levels in two independent revertants were restored to normal (Supplementary Fig. 10b). Overall, predicted RNA secondary structures did not change, suggesting other factors cause the Rol phenotype of sqt-3(ins) (Supplementary Fig. 10c). Finally, we found, using sequence transplantations into hypodermally expressed dpy-10 and neuronaly expressed unc-22, that repression was dependent on 3′ UTR context and observable also in neuronal tissue (Supplementary Fig. 10d). This implies that we isolated a general repressive/destabilizing RNA element. To discover interacting 3′ UTR elements 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 alleles after using sgRNAs specific for this region (Fig. 4d) (Supplementary Fig. 10a). Surprisingly, mRNA levels were not restored (Supplementary Fig. 10b). This points to an alternative mechanism of restored protein function, for example on translational level, or affecting mRNA at a different developmental time point.

Overall, these results demonstrate that parallel genetics and selection by phenotype can be used to obtain specific phenotypic traits, to find functional sequences, and to discover unexpected regulatory interactions in vivo.

Discussion

In this study we develop a general approach for parallel genetics of regulatory sequences in vivo, using inducible expression of a CRISPR-nuclease and multiplexed sgRNAs. Large “diversified” populations can then be used for comprehensive analysis using direct deep sequencing or for selection by phenotypic traits. This allows directly linking regulatory genotypes with phenotypes. We demonstrate this in the model organism C. elegans but believe it is similarly applicable in other animals or plants which allow transgenesis and inducible expression of genome editors.

As we show, 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, higher efficiency would be desired for improved comprehensive testing. This could be achieved with more advanced expression systems57 and optimization of sgRNA promoters for high germline expression. Alternative induction systems (e.g. Auxin, FLP/FRT, Cre/lox, Gal/UAS) could enable continuous and germline-specific Cas9 expression to increase efficiency and allow directed evolution experiments. Our targeted sequencing protocol can capture long deletions, uses the same amplicon for the whole locus, and allows sample multiplexing. Unique molecule counting methods44 for long reads49,50 could be incorporated to reduce PCR biases. Established protocols26 are available for shorter (100-300 bp) target regions. We assumed that each animal in bulk samples contributed equally to the extracted genomic DNA. In the future, barcoding to determine genotypes of individual animals could be developed with plate-based or split-pool51,52 methods. Our method only works at nucleotide resolution close to the sgRNA cut sites. To allow denser titling of regions with mutations, CRISPR-nucleases with alternative or dispensable PAM requirements could be used17,53,54. Although indels are applicable to regulatory regions and even coding sequences55-57, point mutagenesis would enable fine mapping of regulatory nucleotides and amino acids. This exciting possibility could be opened up by implementing hyperactive base58,61- or prime editors62-64. Alternatives to CRISPR-Cas could be developed based on inducible expression of cassette exchange59,66 or targeted transposons67,68. Targeting several independent loci in one
step might be applied to candidate regulatory elements (e.g. miRNA targets, enhancers), screening candidate genes (e.g. from networks, pathways\textsuperscript{6,8}, other assays\textsuperscript{9}), and for synthetic co-evolution of several loci\textsuperscript{1}.

Indel data from high throughput genome editing in human cell lines led to insights into dsDNA break repair outcomes\textsuperscript{27–30,72–75}. Whether these data align with outcomes \textit{in vivo} or in the germline has not been well studied. We found longer indels and less 1nt templated insertions in our data. This can be explained with higher activity of microhomology-mediated end joining (MMEJ), which has been shown to be the main dsDNA break repair pathway in \textit{C. elegans}\textsuperscript{31}. Mutations typical for MMEJ have been implicated in diseases\textsuperscript{6} and microhomologies allow predicting the outcomes of CRISPR-Cas9 editing\textsuperscript{27–29,72–75}. Deep mutation profiles in \textit{C. elegans} could be used to study MMEJ \textit{in vivo}.

In the evolution and determination of phenotypes gain or loss of regulatory sequences is important\textsuperscript{1,77,78} and can be modeled as a gradual process by single nucleotide changes\textsuperscript{6,77–81}. Although indels occur less often than single nucleotide mutations naturally, their impact can be more severe. Deletions have already been found to destroy or generate regulatory elements during evolution\textsuperscript{77,78}. Insertions templated from regions around dsDNA breaks during DNA repair could duplicate functional sequence elements and might be an underestimated force in the evolution of regulatory sequences or disease mutations.

In line with previous observations we found that the two \textit{let-7} complementary sites repress \textit{lin-41} \textit{~4}-fold\textsuperscript{22,24,32,33,36,37}. Furthermore, we uncovered a previously undescribed compensatory interaction. Previous studies concluded either that one LCS could not recapitulate wild-type repression at all\textsuperscript{22} or tested only multiple copies of each site\textsuperscript{23,82}. These studies were done with reporter plasmids and therefore not in the native sequence context (which affects target site accessibility\textsuperscript{22}), or at natural expression levels (which affects miRNA-target ratios\textsuperscript{82}). We found a slight but non-significant effect on RNA regulation when disrupting LCS1. Together with previous studies\textsuperscript{23,28} this indicates that LCS1 could act more strongly than LCS2, and that both sites cannot completely compensate each other’s loss. Since our method might lack the sensitivity to detect more subtle effects from deletion of each LCS alone, future studies would be needed to test this thoroughly.

Our screen for sequences that affect phenotypic traits doubles the phenotypic regulatory alleles registered in the last forty years at Wormbase. Due to the mutagenesis efficiency and because we screened for strong phenotypes, we did not saturate and likely missed many mutations. We expect that higher efficiencies would likely lead to mutations affecting expression and phenotype also for the genes for which we did not isolate any mutations. To determine comprehensively which mutations are tolerated by a locus, even higher efficiencies would be needed. Some mutations were likely affecting known regulatory sequences (\textit{sqt-3}-TATA-box), while for others (\textit{sqt-2 enh, sqt-3 and egl-30 3′ UTR}) we did not determine a mechanism. However, such mutants can be a starting point for future studies on regulatory mechanisms using biochemical, computational and genetic approaches. Additionally, the approach described here can be applied to isolate alleles with altered function (reduction-of-function, gain-of-function) and to obtain animals with desired phenotypic traits. Furthermore we discovered compensatory deletions in the \textit{sqt-3}(ins) 3′ UTR that restored the wild type phenotype. Although not restoring miRNA levels, and likely acting by an alternative mechanism, we uncovered these by screening for the animal phenotype. Our results highlight the possibility to uncover unexpected regulatory interactions using \textit{in vivo} interrogation and readout of phenotype. Such interactions should be detectable, even if they act across different developmental time points, tissues or cell types.

Altogether we believe the approach presented here will help in understanding how gene regulatory logic affects phenotype \textit{in vivo}.

References

21. Reinhart, B. J. et al. The 21-nucleotide let-7 RNA regulates developmental
Gene Editing for applications in Gene Therapy and Functional Genomics. 

Methods

Maintenance of animals. C. elegans were maintained on NGM plates with Escherichia coli OP50 as originally described83 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 Scientific, cat. 10687010) 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.

Strains. A complete list of strains can be found in a supplementary table. The wild-type strain N2 Bristol46 was used to create transgenic lines for expression in C. elegans. 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 (N1K123) which we crossed into a strain expressing Pef3:3:TuTomato.12 from a single copy insertion (EG9729) resulting in a GFP:RFP expressing strain (N1K124) for automated quantifications and sorting using the Copas Biosorter.

Plasmid construction. A list of all plasmids created or used in this study can be found in a supplementary table. The plasmid for heat-shock inducible Streptococcus pyogenes Cas9 expression (pJJF152) was created by Gibson assembly39 of a previously published C. elegans optimized SpCas9 (“Friedland Cas9”), with the hsp-16.48 heat-shock promoter and the unc-54 3′ UTR.

Plasmids for sgRNA expression were cloned as previously described using one of two published backbones (pMB70 or pJSR502). For this, 5-10 μg of backbone was digested with 1 μL FastDigest Eco31I (aka BsaI, Thermo Fisher Scientific, cat. FD0293) or FastDigest Bpi (aka Thermo Fisher Scientific, cat. FD0114) 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 Research, cat. FD1014) 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 Research, cat. FD1014), according to the instruction manual. Two complementarity 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 and the evolution of gene regulation. Nat Rev Genet 13, 59–69 (2012).


A complete list of strains can be found in a supplementary table. The wild-type strain N2 Bristol46 was used to create transgenic lines for expression in C. elegans. 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 (N1K123) which we crossed into a strain expressing Pef3:3:TuTomato.12 from a single copy insertion (EG9729) resulting in a GFP:RFP expressing strain (N1K124) for automated quantifications and sorting using the Copas Biosorter.

Plasmid construction. A list of all plasmids created or used in this study can be found in a supplementary table. The plasmid for heat-shock inducible Streptococcus pyogenes Cas9 expression (pJJF152) was created by Gibson assembly39 of a previously published C. elegans optimized SpCas9 (“Friedland Cas9”), with the hsp-16.48 heat-shock promoter and the unc-54 3′ UTR.

Plasmids for sgRNA expression were cloned as previously described using one of two published backbones (pMB70 or pJSR502). For this, 5-10 μg of backbone was digested with 1 μL FastDigest Eco31I (aka BsaI, Thermo Fisher Scientific, cat. EL0011), 1 μL of 10x T4 DNA ligase buffer (Thermo Fisher Scientific, cat. EL0011), 1 μL T4 PNK (Thermo Fisher Scientific, cat. EK0031) and 6 μL water and was 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 were 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 larval stage L3 or grown to adulthood to collect F2 generation which was then also measured at larval stage L3. The number of analyzed worms per sample was between 1'662 and 21'983 worms.

Small-scale mutagenesis by Cas9 heat shock induction and time course. 20–40 egg-laying adults were transferred to small fem 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.

Developmental synchronization. Synchronized L1s were obtained by bleeding, as previously described47. Egg-laying animals were washed off plates in 50 mL M9 buffer (42 mM NaH2PO4, 22 mM KH2PO4, 86 mM NaCl, 1 mM MgSO4) 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 centrifuged at 600 rpm at 4°C for 15 minutes and washed two more times using M9, centrifugation and decanting. Finally, eggs were suspended 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 mutagenesis by 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 the good incubator “Innova 42” (New Brunswick Scientific/ Eppendorf) at 20°C and 34°C heat shock was applied for 2 hours. Plates were kept at 20°C for 12 hrs and eggs were collected by bleeding 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 bleeding. 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 protease K, 0.5% B-Mercaptoethanol) was added. Sample was incubated for 1.5 hr at 60°C while shaking at 800 rpm on a benchtop heating block. 300 μL of water was added, followed by 400 μL phenol/chloroform/isoamyl alcohol pH 8.0 (Carl Roth, cat. A156). 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 twice with 70% ethanol and resuspended in 25 μL water at 50°C for 30 min. Then 0.25 μL RNase A (10 μL, Thermo Fisher Scientific, cat. EN0601) was added and incubated for 30 min. at 37°C. DNA concentration was determined on a Nanodrop ND-1000 (Thermo Fisher Scientific) and diluted to 50-200 ng/μL in water.

**DNA large 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 PCR amplification. Genomes were sequenced by Nanopore. DNA large amplicon sequencing was performed 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+sg16, sg26+sg27, sgPool). We conducted the experiment at 16°C and 24°C. 3000 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 and separated by eye. 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.

**DNA concentration measured with Qubit, was used to determine molarity and DNA large amplicon sequencing.** Amplification was then done by either agarose gel or SPRI beads. For gel-based amplification PCR reactions were analyzed on agarose gels to ensure successful amplification. PCR was performed with a lin-41-specific primer containing a sample-specific barcode (oJJF140-1147 for samples N2, 1516, 2627, pool at L1 and L4 stages) binding in the second last exon and a primer (oJJF980) 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 each), 0.2 μL Phusion polymerase, 32 μL water and 2 μL DMSO (5% final). Samples were incubated at 98°C 3 min, followed by 35 cycles of 98°C 15 sec, 72°C 1 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 15 sec, 58-72 °C 30 sec, 72 °C for 7 min. PCR reactions were analyzed on agarose gels to ensure successful amplification. For this, animals were washed from plates in a final volume of 2 ml M9 and separation was performed 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+sg16, sg26+sg27, sgPool). We conducted the experiment at 16°C and 24°C. 3000 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 and separation was performed as described above using large-scale mutagenesis by Cas9 heat shock induction.

**Screen for regulatory sequences affecting phenotype.** We targeted 8 genes known with 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 promoter enhancer, TATA-box, initiator (INR) and upstream/promoter regions. A list with all samples can be found in a supplementary table. For each transgenic line we screened 35,000 - 45,000 F2 animals produced from P0 with transiently 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 as described above on an empty plate. We then observed worms moving in M9 and moving away as described above.

**Screen for regulatory sequences affecting phenotype.** We targeted 8 genes known with 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 promoter enhancer, TATA-box, initiator (INR) and upstream/promoter regions. A list with all samples can be found in a supplementary table. For each transgenic line we screened 35,000 - 45,000 F2 animals produced from P0 with transiently 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 as described above on an empty plate. We then observed worms moving in M9 and moving away as described above.

**Screen for regulatory sequences affecting phenotype.** We targeted 8 genes known with 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 promoter enhancer, TATA-box, initiator (INR) and upstream/promoter regions. A list with all samples can be found in a supplementary table. For each transgenic line we screened 35,000 - 45,000 F2 animals produced from P0 with transiently 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 as described above on an empty plate. We then observed worms moving in M9 and moving away as described above. F15969-018, homogenization with a Precellys 24 tissue homogenizer (Bertin Instruments) and storage at -80°C. 5,000 L1 were seeded on large 15 cm NFG plates at 24°C and collected 3 hours later, at late-E1 stage, and prepared for RNA extraction like the L1 sample. At 32 hours, lin-41 mRNA is fully downregulated19, while the lethal vulva bursting occurs later after molting, in the adult stage.18
<1 min.). Dpy, Une, and Rol worms were identified by morphology, their movement in M9 or slow and otherwise impaired movement away from the spot of dispersion. Potential mutants were then picked and kept on plates for 2–4 generations at 24°C to achieve homozygosity. Animals were then singed 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 sqrt-2 locus (n=300 animals). For sqrt-2 the penetrance varied between 10–100%. We scored the expressivity of the phenotypes into three categories (++, +++, +). We were able to accommodate different experimental setups, hoping that the pipeline can be useful to the scientific community carrying out genome editing experiments using the CRISPR-based technologies, in particular those that aim to introduce many combinations of mutations in a genome via inducing double-stranded DNA breaks repaired via the non-homologous end joining pathways. The pipeline can handle both short (single- or paired-end Illumina) and long reads (PacBio) from both DNA and RNA, so that the effects in the DNA editing can be also observed in the matching RNA samples. Each sample can contain multiple sgRNAs targeting multiple regions of the genome. A sample might contain reads coming from one or more individual genomes/transcriptomes. However, each sample is treated as if it is a single individual with a mosaic genome. The first purpose of the pipeline is to serve as a quality control/reporting tool to evaluate the genome-editing experiment and address the following questions: Has the CRISPR-Cas9 treatment induced any mutations? If so, how are they distributed in the genome? Do the mutations that are commonly found in many reads originate at the intended cut site based on the designed guide RNA matching sites in the genome? How efficient were different guide designs in inducing DNA damage? Can we capture long deletions if there are multiple sgRNAs used in the same sample targeting nearby sites? How diverse are the deletions or insertions detected at the cut sites? We developed the pipeline to produce many processed files that can be generated based on these intermediate files to address the many custom questions.

**Input description.** crispr-DART is implemented using Snakemake following the practices as implemented for the PiGx pipelines. The input consists of a settings file in yaml format, which contains configurations for the tools used in the pipeline. Moreover, it contains file paths for where the sequencing reads are loaded, the target genome sequence to be used for mapping the reads, the sample sheet file (in comma-separated file format) which contains the experimental design, the file containing the genomic coordinates of the expected sgRNA cut sites in the BED file format, and a table (in tab-separated format) that is needed for when a pair of samples are to be compared (for instance to observe the differences in per-base distribution of deletions detected in a treated sample and an untreated control sample).

**Steps of the pipeline.** The pipeline consists of the following sequence of processing steps (see Supplementary Figure 2b):

**Pre-processing reads**
- Quality control (using fastqc and multiqc and quality improvement of reads (using Trim-Galore))

**Mapping**
- Mapping/alignment of the reads to the genome (using BBMAP). We use BBMAP for read alignment because it can handle both long and short reads, both single-end and paired-end reads, both DNA and RNA reads, and it can help detect both short and long insertion and deletion events.
- Re-alignment of reads with indels using GATK. This step helps reconcile different indel alignments to minimise the noise in alignments.

**Finding indels and generating indel related statistics**
- Extraction of indels from the BAM files using R packages GenomicAlignments and RSamtools, producing:
  - BED files for the genomic coordinates of insertions and deletions

---

**CRISPR-Downstream Analysis and Reporting Tool ("crispr-DART").** The source code along with installation instructions and sample input files can be accessed here: [https://github.com/BIMSBBioinfo/crispr_DART](https://github.com/BIMSBBioinfo/crispr_DART). The source code for reproducing the figures generated based on the pipeline's output can be found here: [https://github.com/BIMSBBioinfo/froehlich_uyar_et_al_2020](https://github.com/BIMSBBioinfo/froehlich_uyar_et_al_2020).
Generating a website of HTML reports

All the pre-processed files from the previous steps are combined to generate interactive (where applicable) HTML reports from all the analysed samples that exist in the input sample sheet. For each targeted region (assuming a region of a few thousand base pairs that is sequenced), currently four different reporting Markdown scripts are run. The resulting HTML files are organised into a website using the `render_site` function of the Rmarkdown package\(^\text{**}\). Thus, all the processed data and outcomes can be quickly browsed through a website. The resulting website that is the output of the pipeline for this study can be browsed here: https://bimsbsstatic.mdc-berlin.de/akalin/buyar/froehlich_uyar_et_al_2020/reports/index.html

How to reproduce the figures in the manuscript

The source code and usage instructions for reproducing the figures in this manuscript, that were made based on the output of the crispr-DART pipeline, can be found here: https://github.com/BIMSBbioinfo/froehlich_uyar_et_al_2020. 

RNA analysis: Comparing deletions in L1 and L4 stages

Deletions in PacBio L1 and L4 samples were filtered to keep only those deletions supported by at least 5 PacBio reads. Each deletion was categorized based on their overlap with important sites in the 3'UTR of lin-41: 

- First let-7 micro-rrna binding site (only the seed region) (LCS1_seed): chrI:9335255-9335263
- Second let-7 micro-rrna binding site (only the seed region) (LCS2_seed): chrI:9335268-9335276
- First let-7 micro-rrna binding site minus the seed region (LCS1_3compl): chrI:9335256-9335259
- Second let-7 micro-rrna binding site minus the seed region (LCS2_3compl): chrI:9335251-9335257
- First let-7 micro-rrna binding site minus the seed region (LCS1_3compl): chrI:9335256-9335259
- Second let-7 micro-rrna binding site minus the seed region (LCS2_3compl): chrI:9335251-9335257

First and second let-7 micro-rrna seed regions, and also those that don’t overlap any of these defined regions.

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 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 different from those deletions that don’t overlap any of these sites.

RNA analysis: Clustering pacbio reads and comparing genotype abundance in L1 and L4 stages

Reads from both L1 and L4 stage lin-41 RNA samples sequenced using PacBio that cover the 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) were obtained and all combinations of k-mers (k=5) were counted within these alignments allowing for up to 1 mismatch using Biostrings package\(^\text{**}\). Seurat package\(^\text{**}\) 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 the Illumina reads).

Fitness analysis

For this analysis, we have utilized lin-41 DNA samples sequenced (Illumina single-end sequencing) from multiple generations from F2 to F5 of the same pool of animals treated with sgRNA guides “sg15 and sg16”, “sg26 and sg27” and “sg pool”.

The important sites considered for this analysis are:
- First let-7 micro-rrna binding site (only the seed region) (LCS1_seed): chrI:9335255-9335263
- Second let-7 micro-rrna binding site (only the seed region) (LCS2_seed): chrI:9335268-9335276
- First let-7 micro-rrna binding site minus the seed region (LCS1_3compl): chrI:9335256-9335259
- Second let-7 micro-rrna binding site minus the seed region (LCS2_3compl): chrI:9335251-9335257
- Poly-adenylylation signal: chrI:9334816-9334821
- Stop-codon: chrI:9335965-9335967

We would like to address the question whether the deletions that exist at F2 are exposed to purified selection over generations if they overlap the important sites on the 3’UTR region of lin-41. We do this analysis in two ways:
- First, we count the deletions categorised by their overlap (or non-overlap) with the important sites that exist in F2 generation and see how many of them still exist in later generations.
- Secondly, we do the same analysis at the level of reads: counting the reads with deletions that overlap or not overlap the important sites from generations F2 to F5. When comparing the number of reads, the read counts are normalized by the library sizes (total number of reads in the sample).

sgRNA efficiency comparisons. For comparing observed efficiencies to published prediction scores and other sgRNA characteristics, these were manually extracted from the CRISPOR web application (http://crispor.tefor.net)\(^\text{**}\). 

Browser shots. Browser shots were compiled using indel profiles and top indels provided by the computational pipeline as BigWig and BED files and loading them into the UCSC genome browser\(^\text{**}\) or the IGV browser\(^\text{**}\) followed by export as vector graphics compatible format. We used C. elegans genome version cell1/WBce235 including 26 species base-wise conservation (PhyloP).

Graphics. Final figures were assembled using Adobe Illustrator without changing the plotted data, with adjustments to composition, axis labels, line widths, colors, etc.

Manuscript formatting. This manuscript was composed using Adobe InDesign starting with a template from https://github.com/cleterrier/ManuscriptTools. We thank Christophe Leterrier for sharing it.

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Strain and plasmid availability. C. elegans mutant strains will be available through the Caenorhabditis Genetics Center (CGC) as indicated in a supplementary table. Plasmids generated for this work for heat shock Cas9 expression (p1F152) and proof of concept sgRNAs (targeting SECGFP, dpy-10, sgr-3) will be available from Addgene as indicated in a supplementary
References (Methods)


Acknowledgements

We are very grateful to Baris Tursun, Luisa Cochella, David Koppsite, and the Rajewsky, Akalin labs for helpful feedback and discussions and Claudia Quedenau, Daniele Franze, the sequencing facility for Pacbio sequencing and Sergej Herzog for technical assistance. We thank the Caenorhabditis Genetics Center (CGC), Christian Frøkjaer-Jensen, Erik Jorgensen, João Ramalho, Mike Boxem, Daniel Dickinson, Bob Goldstein, and Jason Chin for sharing plasmids and strains. B.U. acknowledges funding by the German Federal Ministry of Education and Research (BMBF) as part of the RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure (de.NBI) (031 A538C RBC).

Author contributions

J.J.F. and N.R. developed concepts, methodology, and discussed the data. 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. P.G. contributed to software. J.J.F., M.H. and K.T. performed investigation and experimental work. J.J.F. wrote the original draft. J.J.F., B.U., A.A., N.R. reviewed and edited the manuscript. A.A. and N.R. contributed with resources, supervision, project administration and funding acquisition.
Supplementary Figure 1. Transiently induced Cas9 expression creates germline indel mutations. 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. The eggs collected 14 – 16 hrs after heat shock produced the most GFP-negative animals. b, Comparison of two different plasmids for heat shock inducible Cas9 in a time course. Dpy-10 was targeted with a sgRNA, time course was performed as in a) and Dpy progeny were counted. Eggs collected 12 – 14 hrs after heat shock produced the most Dpy animals. Both plasmids performed comparably. c, Indel mutations detected by Sanger sequencing of individual GFP-negative animals after targeting his-72::GFP with sgRNAs. d, Sanger sequencing of indel mutations created by a pool of three sgRNAs. e, 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. f, 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. g, Scheme showing automated fluidics measurement of F1 and F2 GFP negative animals to determine the amount of germline mutations. h, 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. i, Difference in the amount of GFP-negative animals between F1 and F2 generation. Almost the same amount (80%) of GFP-negative animals in the F2 generations indicates high germline transmission of mutations.
**Supplementary Figure 2. Software pipeline “crispr-DART”**. a, Scheme showing our long amplicon sequencing approach. b, Diagram of the software pipeline “crispr-DART”. The user provides input files and the pipeline produces processed genomic files and html reports. c, Custom analyses for this study were then performed with R scripts using the processed genomic files as input. d, Screenshots of the html reports provided by crispr-DART. For “crispr-DART” and R scripts see “Code Availability” in the Supplementary Methods.
Supplementary Figure 3. sgRNA efficiency characteristics. a, Correlation of various prediction scores for sgRNA efficiency 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. d, Comparison of sgRNA efficiency for sgRNAs targeting the coding sequence of essential genes and all other sgRNAs. No difference could be observed, indicating that low efficiencies for specific sgRNAs were not due to lethality. e, Scheme which illustrates how single-cut or multi-cut deletions were categorized computationally.
Supplementary Figure 4. 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 to surrounding sequence (+/- 50bp), shown for three samples. d, For data from 34 samples, matches of 5-mers from insertions to surrounding sequence (+/- 50bp). e, Diagram showing mechanistic steps of dsDNA repair via microhomology-mediated end joining. Highlighted is the dissociation & re-annealing step which could lead to the templated insertions observed in our data.

Supplementary Figure 4. 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 to surrounding sequence (+/- 50bp), shown for three samples. d, For data from 34 samples, matches of 5-mers from insertions to surrounding sequence (+/- 50bp). e, Diagram showing mechanistic steps of dsDNA repair via microhomology-mediated end joining. Highlighted is the dissociation & re-annealing step which could lead to the templated insertions observed in our data.
Supplementary Figure 5. **Unique genotypes created by indel mutations correlate with higher sgRNA efficiency and more sgRNAs per line.**

**a,** Unique genotypes created per sample by deletions or insertions. Each genotype is detected by at least 0.001% of reads mapped to the analyzed position.Ctrls n=24, samples n=36. Wilcoxon, p = 1.7e−08 for deletions, p = 4.7e−09 for insertions.

**b,** Correlation between sgRNA efficiency and the created unique deletions per sgRNA (n=91) per sample.

**c,** Correlation between the amount of different sgRNAs in a transgenic line and the created unique deletions per sample (n=6084). Unique deletions only from treated samples (n=35), supported by >5 reads and >0.001% frequency.
Supplementary Figure 6. Impact of lin-41 3′ UTR deletions on RNA levels. a, Diagram showing let-7 complementary sites LCS1, LCS2 in the lin-41 3′ UTR. b, Diagram of lin-41 and let-7 developmental expression and time points of RNA extraction. c, Diagram of the targeted RNA sequencing strategy. cDNA was amplified using a large amplicon and sequenced using the Pacbio long read workflow. d, UMAP clusters of long reads covering the complete lin-41 3′ UTR, detected in cDNA from L1 or L4 developmental stages. Each dot represents one read. e, Status of overlap with let-7 sites for each read. f, Number of detected reads with a deletion (y-axis) per genomic nucleotide (x-axis). Reads are separated by cluster (sub-panels) and developmental stage (L1=red, L4=green). The two vertical black lines indicate the location of the two let-7 complementary sites (LCS1 and LCS2).
Supplementary Figure 7. Abundance of \textit{lin-41} 3' UTR deletions over several generations. a, The heatmap (pheatmap package) displays the frequency of deletions (on rows) scaled by row over multiple generations (columns). The annotation columns display which deletions overlap different \textit{let-7} binding sites. Log10\textunderscore read\_count is number of reads in log-10 scale that support the deletions at F2 generation.
Supplementary Figure 8. Screen for regulatory mutations with phenotypes. a, Outline of the screen. 8 genes were targeted by pools of sgRNAs (between 2-6) in different regulatory regions (some in enhancer, promoter, and all in the 3' UTR) resulting in 21 samples. For these, 35,000 F2 animals were screened manually for known RNAi phenotypes. The screen allowed us to sequence F2 siblings to determine the mutations present in the screened population. b, Amount of F2 progeny obtained per bleached adult of the F1 generation. c, Location and extent of mutations affecting the coding sequence in Dpy sqt-3 mutants. For long insertions the origin was determined by BLAT. d, Rol mutations isolated after targeting the sqt-3 3’ UTR without sg2. e, Mutation types found in reduction-of-function alleles isolated for four targeted regions from three genes (egl-30, sqt-2, sqt-3). 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, Sequences of the mutations in Uncoordinated egl-30 mutants.
Supplementary Figure 9. Insertions in the 3′ UTR of sqt-3 lead to post-transcriptional mRNA reduction. **a,** Isolation strategy of non-Rol mutants. 96 random non-Rol animals were picked and T7 assay was performed to screen for mutations. **b,** Indels in non-Rol animals determined by Sanger sequencing. Exact sequence of two mutations could not be resolved due to heterozygosity. **c,** Quantification of sqt-3 RNA expression along development during L4 stage in wild type (N2) and mutant. Worms were synchronized by bleaching and RNA was quantified on the Nanostring system. **d,** SQT-3 mRNA or pre-mRNA levels in different sqt-3 alleles at 26 hrs into synchronized development. Levels were quantified by qRT-PCR with primers specific for the spliced or the un-spliced transcript. **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.
Supplementary Figure 10. Deletions affecting the repressive RNA element in the sqt-3(ins) 3’ UTR can rescue the Rol phenotype. 

a, Nucleotide sequences of relevant 3’ UTR regions in Rol, non-Rol and rescued mutants showing the inserted and deleted nucleotides. 

b, mRNA or pre-mRNA levels of sqt-3 in mutant and revertants at 26 hrs into synchronized development. Levels were quantified by qRT-PCR with primers specific for spliced or un-spliced transcript. 

c, Predicted RNA secondary structures of whole mRNA or pre-mRNA folded. 

d, Transplantation of mutant sequences into independent 3’ UTRs. Sequence 1 (from insertion mutant) or sequence 2 (from rescued insertion mutant) were knocked-in at the dpy-10 or unc-22 3’ UTR. Seq1 led to the characteristic reduction-of-function phenotype Twitching (Twi) in unc-22.
<table>
<thead>
<tr>
<th>gene</th>
<th>region</th>
<th>amount of sgRNAs</th>
<th>amplicon size selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>lin-41</td>
<td>CDS</td>
<td>2</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>5</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>2</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>5</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>2</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>8</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>8</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>3'UTR</td>
<td>8</td>
<td>+/-</td>
</tr>
<tr>
<td>lin-41</td>
<td>downs</td>
<td>2</td>
<td>+/-</td>
</tr>
<tr>
<td>dpy-2</td>
<td>3'UTR</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>dpy-10</td>
<td>3'UTR</td>
<td>2</td>
<td>+/-</td>
</tr>
<tr>
<td>dpy-10</td>
<td>3'UTR</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>dpy-10</td>
<td>3'UTR</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>egl-30</td>
<td>3'UTR</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>egl-30</td>
<td>3'UTR</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>egl-30</td>
<td>3'UTR</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>rol-6</td>
<td>enh</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>rol-6</td>
<td>TATA</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>rol-6</td>
<td>INR</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>rol-6</td>
<td>3'UTR</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>snb-1</td>
<td>ups</td>
<td>7</td>
<td>+</td>
</tr>
<tr>
<td>snb-1</td>
<td>CDS</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>snb-1</td>
<td>CDS</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>snb-1</td>
<td>3'UTR</td>
<td>8</td>
<td>+</td>
</tr>
<tr>
<td>snb-1</td>
<td>3'UTR</td>
<td>7</td>
<td>+</td>
</tr>
<tr>
<td>sqt-2</td>
<td>ups</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>sqt-2</td>
<td>TATA_INR</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>sqt-2</td>
<td>3'UTR</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>TATA</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>INR</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>3'UTR</td>
<td>3</td>
<td>+/-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>3'UTR</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>3'UTR</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>sqt-3</td>
<td>3'UTR</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>unc-26</td>
<td>3'UTR</td>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>unc-54</td>
<td>3'UTR</td>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>let-2</td>
<td>CDS</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>let-2</td>
<td>3'UTR</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>let-7</td>
<td>miRNA</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>par-2</td>
<td>CDS</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>tbb-2</td>
<td>CDS</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>tbb-2</td>
<td>3'UTR</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>unc-119</td>
<td>CDS</td>
<td>1</td>
<td>+/-</td>
</tr>
<tr>
<td>zyg-1</td>
<td>CDS</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>zyg-1</td>
<td>3'UTR</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>

Supplementary Table 1: Overview of samples.
<table>
<thead>
<tr>
<th>gene</th>
<th>dels into coding</th>
<th>isolated mutants</th>
<th>phenotype</th>
<th>region</th>
<th>deletion</th>
<th>complex / insertion</th>
<th>proportion of deletions (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>egl-30</td>
<td>-</td>
<td>11</td>
<td>Unc</td>
<td>3' UTR</td>
<td>4</td>
<td>3' UTR</td>
<td>36</td>
</tr>
<tr>
<td>sqrt-2</td>
<td>+</td>
<td>7</td>
<td>Rol</td>
<td>enhancer</td>
<td>3</td>
<td>3' UTR</td>
<td>43</td>
</tr>
<tr>
<td>sqrt-3</td>
<td>+</td>
<td>13</td>
<td>Rol</td>
<td>TATA box</td>
<td>5</td>
<td>3' UTR</td>
<td>38</td>
</tr>
<tr>
<td>sqrt-5</td>
<td>+</td>
<td>26</td>
<td>Rol</td>
<td>3' UTR</td>
<td>1</td>
<td>3' UTR</td>
<td>25</td>
</tr>
<tr>
<td>dpy-2</td>
<td>+</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dpy-10</td>
<td>+</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rol-6</td>
<td>-</td>
<td>0</td>
<td>prom. TATA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rol-6</td>
<td>-</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sqrt-2</td>
<td>+</td>
<td>0</td>
<td>TATA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sqrt-2</td>
<td>+</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>unc-26</td>
<td>-</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>unc-54</td>
<td>+</td>
<td>0</td>
<td>3' UTR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>sqrt-5(ins)</td>
<td>+</td>
<td>15</td>
<td>Rol-&gt;non-Rol</td>
<td>3' UTR</td>
<td>11</td>
<td>4</td>
<td>73</td>
</tr>
</tbody>
</table>

Supplementary Table 2: Overview of phenotype screen samples.