Rational design and whole-genome predictions of single guide RNAs for efficient CRISPR/Cas9-mediated genome editing in *Ciona*

Shashank Gandhi[1], Lionel Christiaen[2] and Alberto Stolfi[2]

Center for Developmental Genetics, Department of Biology, New York University, New York, USA

[1] present address: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA.[2] corresponding authors: lc121@nyu.edu, ass8@nyu.edu

Abstract

The CRISPR/Cas9 system has emerged as an important tool for a wide variety of genome engineering applications, including reverse genetic screens. Previously, we described the implementation of the CRISPR/Cas9 system to induce tissue-specific mutations at targeted locations in the genome of the sea squirt *Ciona* (STOLFI et al. 2014). In the present study, we designed 83 single guide RNA (sgRNA) vectors targeting 23 genes expressed in the cardiopharyngeal progenitors and surrounding tissues in the *Ciona* embryo and measured their mutagenesis efficacy rates by massively parallel indel detection at the targeted loci using high-throughput sequencing. We show that the combined activity of two highly active sgRNAs allows us to generate large (>3 kbp) deletions of intervening genomic DNA in somatic cells of electroporated embryos, permitting tissue-specific gene knockouts. Additionally, we employed L1-regularized regression modeling to develop an optimal sgRNA design algorithm (TuniCUT), based on correlations between target sequence features and mutagenesis rates. Using this algorithm, we have predicted mutagenesis rates for sgRNAs targeting all 4,853,589 sites in the *Ciona* genome, which we have compiled into a "**C**RISPR/Cas9-**i**nduced ***Ci****ona* **K**nock-**O**ut" (Ci^2KO) sgRNA sequence library. Finally, we describe a new method for the assembly of sgRNA expression cassettes using a simple one-step overlap PCR (OSO-PCR) protocol. These cassettes can be electroporated directly into *Ciona* embryos as unpurified PCR products to drive sgRNA expression, bypassing the need for time-consuming cloning and plasmid DNA preparations. We anticipate that this method will be used in combination with genome-wide sgRNA predictions to systematically investigate tissue-specific gene functions in *Ciona.*

Introduction

A platform for targeted mutagenesis has been recently developed based on the prokaryotic immune response system known as CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-Associated) (BARRANGOU *et al.* 2007). In its most common derivation for genome engineering applications, the system makes use of a short RNA sequence, known as a single guide RNA (sgRNA) to guide the Cas9 nuclease of *Streptococcus pyogenes* to a specific target DNA sequence. (JINEK *et al.* 2012; CONG *et al.* 2013; JINEK *et al.* 2013; MALI *et al.* 2013). Although initial Cas9 binding requires a Protospacer Adjacent Motif (PAM) sequence of "NGG", the high specificity of this system is accounted for by Watson-Crick base pairing between the 5' end of the sgRNA and an 17-20bp "protospacer" sequence immediately adjacent to the PAM (FU *et al.* 2014). Upon sgRNA-guided binding to the intended target, Cas9 generates a double stranded break (DSB) within the protospacer sequence. Imperfect repair of these DSBs by non-homologous end joining (NHEJ) often results in short insertions or deletions (indels) that may disrupt the function of the targeted sequence. Numerous reports have confirmed the high efficiency of CRISPR/Cas9 for genome editing purposes (DICKINSON *et al.* 2013; HWANG *et al.* 2013; WANG *et al.* 2013a; SHALEM *et al.* 2014; GANTZ AND BIER 2015).

The tunicate *Ciona* has emerged as a model organism for the study of chordate-specific developmental processes (SATOH 2013). The CRISPR/Cas9 system was recently adapted to induce site-specific DSBs in the *Ciona* genome (SASAKI *et al.* 2014). Using electroporation to transiently transfect *Ciona* embryos with plasmids encoding CRISPR/Cas9 components, we were able to generate clonal populations of somatic cells carrying loss-of-function mutations of *Ebf*, a transcription-factor-coding gene required for muscle and neuron development, in F0-generation embryos (STOLFI *et al.* 2014). By using developmentally regulated *cis*-regulatory elements to

drive expression of Cas9 in specific cell lineages or tissue types, we were thus able to control the disruption of *Ebf* function with spatiotemporal precision.

We sought to expand the strategy to target more genes, with the ultimate goal of building a genome-wide library of sgRNAs for systematic genetic loss-of-function assays. However, CRISPR/Cas9-based genome engineering is still in its infancy, and few guidelines exist for the rational design of highly active sgRNAs, which are critical for rapid gene disruption in F0. Individual studies have revealed certain protospacer-targeting sequence features that correlate with high sgRNA expression and/or activity in CRISPR/Cas9-mediated DNA cleavage (DOENCH *et al.* 2014; GAGNON *et al.* 2014; REN *et al.* 2014; CHARI *et al.* 2015; MORENO-MATEOS *et al.* 2015), but these have been performed in different organisms, using a variety of sgRNA and Cas9 delivery methods.

Given the uncertainty regarding how sgRNA design principles gleaned from experiments in other species might be applicable to CRISPR/Cas9 efficacy in *Ciona*, we tested a collection of sgRNAs using our own modified tools for tissue-specific CRISPR/Cas9-mediated mutagenesis in *Ciona* embryos. We describe here the construction and validation of this collection using high-throughput sequencing of PCR-amplified target sequences. This dataset allowed us to develop a pipeline for machine-learning-optimized design and efficient assembly of sgRNA expression constructs for use in *Ciona.* Using these methods, we have pre-emptively designed and predicted the mutagenesis efficacy rates of sgRNAs targeting all 4,853,589 viable Cas9 target sites (4,249,756 of them unique) identified in the entire *Ciona* genome. This "CRISPR/Cas9-Induced *Ciona* Knock-Out" (Ci^2KO) sgRNA database should facilitate the widespread use of CRISPR/Cas9 by the *Ciona* research community.

Results

High-Throughput sequencing to estimate sgRNA-specific mutagenesis rates

Previous studies using CRISPR/Cas9-based mutagenesis in *Ciona* suggested that individual sgRNAs define the level of endonuclease activity of Cas9/sgRNA complexes, therefore varying mutagenesis rates at distinct target sites (SASAKI *et al.* 2014; STOLFI *et al.* 2014). In order to test a larger number of sgRNAs and search for parameters that may influence mutagenesis efficacy, we constructed a library of 83 sgRNA expression plasmids targeting a set of 23 genes (**Table 1**). The majority of these genes are transcription factors and signaling molecules of potential interest in the study of cardiopharyngeal mesoderm development. The cardiopharyngeal mesoderm of *Ciona*, also known as the Trunk Ventral Cells (TVCs), are multipotent cells that invariantly give rise to the heart and pharyngeal muscles of the adult (HIRANO AND NISHIDA 1997; STOLFI *et al.* 2010; RAZY-KRAJKA *et al.* 2014), thus sharing a common ontogenetic motif with the potentially homologous cardiopharyngeal mesoderm of vertebrates (WANG *et al.* 2013b; DIOGO *et al.* 2015).

We followed a high-throughput-sequencing-based approach to quantify the efficacy of each sgRNA (**Figure 1a**). The 83 sgRNA plasmids were individually co-electroporated with *Eef1a1>nls::Cas9::nls* plasmid into pools of *Ciona* zygotes, which were then grown at 18°C for 16 hours post-fertilization (hpf; embryonic stage 25)(HOTTA *et al.* 2007). Targeted sequences were individually PCR-amplified from each pool of embryos, as well as from "negative control" embryos (electroporated with *Eef1a1>nls::Cas9::nls* and "*U6>Negative Control*" sgRNA vector) grown in parallel to each batch of electroporated embryos. Agarose gel-selected and purified amplicons (varying from 108 to 350 bp in length) were pooled in a series of barcoded

Table 1. Genes targeted for CRISPR/Cas9-mediated mutagenesis

The 23 genes targeted in the initial screen, each identified here by official gene symbol, aliases, and KyotoHoya identifier.

#	Gene Symbol	Aliases	2012 KyotoHoya ID
1	*Bmp2/4*	*Bone morphogenetic protein 2/4*	KH.C4.125
2	*Ddr*	*Discoidin Domain Receptor Tyrosine Kinase 1/2*	KH.C9.371
3	*Ebf*	*Collier/Olf/EBF; COE*	KH.L24.10
4	*Eph.a*	*Ephrin type-A receptor.a; Eph1*	KH.C1.404
5	*Ets.b*	*Ets/Pointed2*	KH.C11.10
6	*Fgf4/6*	*Fibroblast growth factor 4/6; FGF, unassigned 1*	KH.C1.697
7	*Fgf8/17/18*	*Fibroblast growth factor 8/17/18*	KH.C5.5
8	*Fgfr*	*Fibroblast growth factor receptor*	KH.S742.2
9	*Foxf*	*FoxF*	KH.C3.170
10	*Foxg-r*	*Foxg-related; Orphan Fox-4; Ci-ZF248*	KH.C5.74
11	*Fzd5/8*	*Frizzled5/8*	KH.C9.260
12	*Gata4/5/6*	*GATA-a*	KH.L20.1
13	*Hand*	*Heart And Neural Crest Derivatives Expressed 1/2*	KH.C14.604
14	*Hand-r*	*Hand-related; Hand-like; NoTrlc*	KH.C1.1116
15	*Htr7-r*	*5-Hydroxytryptamine Receptor 7-related*	KH.S555.1
16	*Isl*	*Islet1/2*	KH.L152.2
17	*Lef1*	*Lef/TCF*	KH.C6.71
18	*Mrf*	*Muscle regulatory factor; MyoD*	KH.C14.307
19	*Neurog*	*Neurogenin; Ngn*	KH.C6.129
20	*Nk4*	*Nkx2-5; Tinman*	KH.C8.482
21	*Rhod/f*	*RhoD/F; Rif*	KH.C1.129
22	*Tle.b*	*Groucho2*	KH.L96.50
23	*Tll*	*Tolloid-like; Tolloid*	KH.C12.156

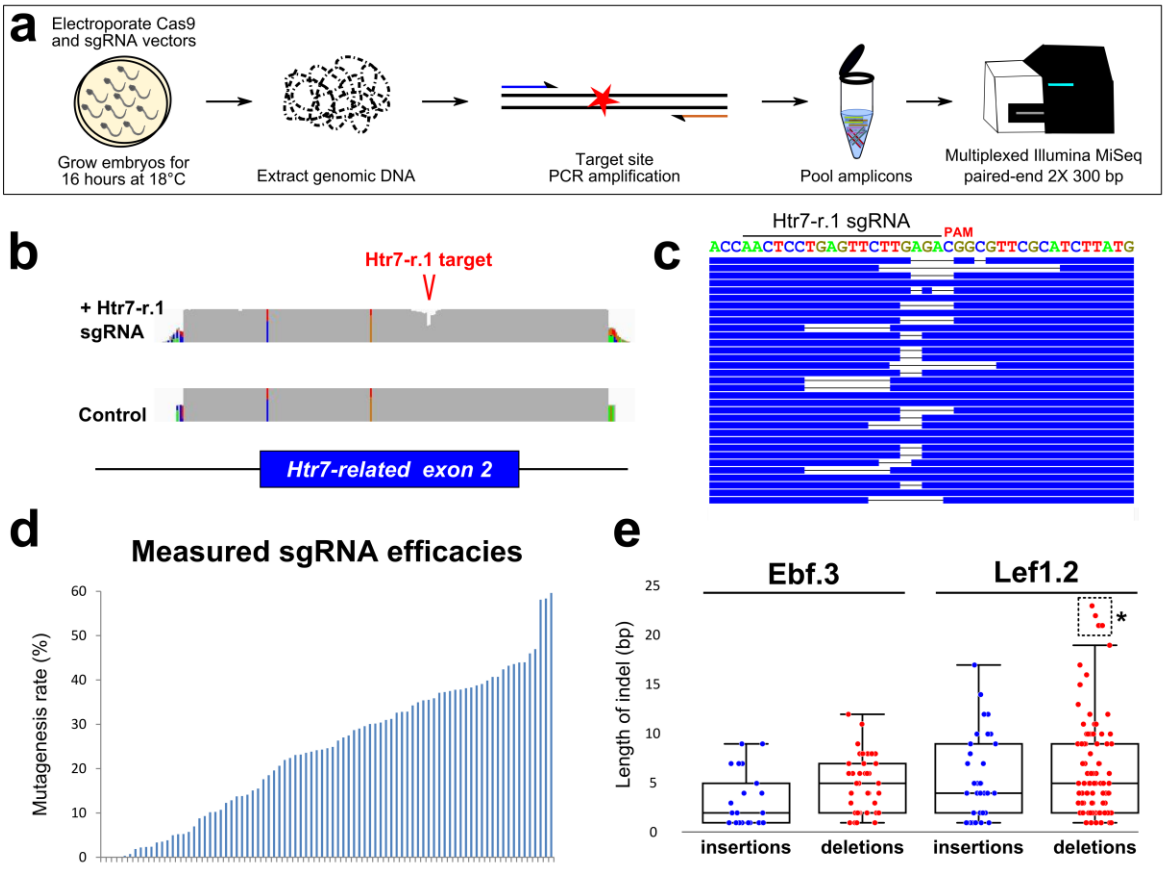


Figure 1. Next-Generation Sequencing approach to validating sgRNAs for use in *Ciona* embryos

a) Schematic for next-generation sequencing approach to measuring mutagenesis efficacies of sgRNAs expressed in F0 *Ciona* embryos. See results and materials and methods for details. **b)** Representative view in IGV browser of coverage (grey areas) of sequencing reads aligned to the reference genome. "Dip" in coverage of reads from embryos co-electroporated with *Eef1a1>Cas9* and *U6>Htr7-r.1* sgRNA vector indicates CRISPR/Cas9-induced indels around the sgRNA target site, in the 2nd exon of the *Htr7-related* (*Htr7-r)* gene. Colored bars in coverage indicate single-nucleotide polymorphisms/mismatches relative to reference genome. **c)** Diagram representing a stack of reads bearing indels of various types and sizes, aligned to exact target sequence of Htr7-r.1 sgRNA. **d)** Distribution of mutagenesis efficacy rates measured for each sgRNA, ordered from lowest (0%) to highest (59.63%). **e)** Box-and-whisker plots showing the size distribution of insertions and deletions caused by Ebf.3 or Lef1.2 sgRNAs. Dashed box with asterisk indicates outliers.

"targeted" and "negative control" Illumina sequencing libraries and sequenced using the Illumina MiSeq platform.

Alignment of the resulting reads to the reference genome sequence (SATOU *et al.* 2008) revealed that targeted sites were represented on average by 16,204 reads, with a median of 3,899 reads each (**Table 2**; note that we excluded the Bmp2/4.1 sgRNA from further analyses because only two reads mapped to its target sequence). The ability of each sgRNA to guide Cas9 to induce DSBs at its intended target was detected by the presence of insertions and deletions (indels) within the targeted protospacer + PAM. The simple ratio of [indels]/[total reads] represents the mutagenesis efficacy of the sgRNA (**Figure 1b-d, Table 2, Supplementary Table 1**). For simplicity, we did not count single nucleotide polymorphisms (SNPs), even though a fraction of them may result from NHEJ-repair of a DSB event. Our data indicated that all sgRNAs (with the exception of Neurog.2) were able to induce DSBs, with estimated efficacies varying from 0.05% (Ebf.4) to 59.63% (Htr7-r.2).

This conservative approach most likely underestimates the actual mutagenesis rates. First, we excluded SNPs potentially resulting from imperfect DSB repair. Second, but more importantly, amplicons from transfected cells are always diluted by wild-type sequences from untransfected cells in the same sample, due to mosaic incorporation of sgRNA and Cas9 plasmids. Indeed, we previously observed an enrichment for mutated sequences amplified from reporter transgene-expressing cells isolated by magnetic-activated cell sorting (representing the transfected population of cells) relative to unsorted cells (representing mixed transfected and untransfected cells) (STOLFI *et al.* 2014). In that particular example, the estimated mutagenesis efficacy induced by the Ebf.3 sgRNA was 66% in sorted sample versus 45% in mixed sample. This

Table 2. All sgRNAs and their mutagenesis efficacy rates (Mut%) assayed by next-generation sequencing

Protospacer and PAM sequences, number of reads sequenced and indels detected for each sgRNA in this study. Ebf.3 was tested either as a plasmid or unpurified traditional PCR and OSO-PCR cassettes (see text for details).

gRNA	Protospacer (N19) + PAM	Reads	Indels	Mut%
Bmp2/4.1	CTGCATAATACGCGGGACCTGG	2	0	0.00*
Bmp2/4.2	CTAGAAGTTATCACCACGAAGG	8032	2001	24.91
Bmp2/4.3	ATGTGGTTGCTCGGCATCCCGG	4821	620	12.86
Bmp2/4.4	GAGCTTCTCCTGCATCGAGAGG	20319	7686	37.83
Ddr.1	CTACAGCACAAATAGATACWGG	1146	44	3.84
Ddr.2	CAATGCTATCCATTGGGGCAGG	2597	1101	42.40
Ddr.3	GAGCGTCCGCAGTTGTCGCTGG	270	9	3.33
Ddr.4	CATCCACTGGTGCAGGGGTTGG	854	401	46.96
Ddr.5	TGAGCCTTATGTCTCTGCATGG	2118	695	32.81
Ebf.1	TACGACAGACAAGGGCAGCTGG	923	227	24.59
Ebf.2	GCATCCATCCTCTCACTGCCGG	825	41	4.97
Ebf.3	CTGAGGGTTGGACAACAGGWGG	923	344	37.27
Ebf.3 (PCR)	CTGAGGGTTGGACAACAGGWGG	346167	104397	30.16
Ebf.3 (OSO-PCR)	CTGAGGGTTGGACAACAGGWGG	9264	3174	34.26
Ebf.4	TTGGAGAAAATTTCTTTGACGG	10184	5	0.05
Eph.a.1	ATTCACGATAAGGTAAGACGGG	1528	354	23.17
Eph.a.2	GGCTGCAATCGTATCAACCTGG	915	320	34.97
Eph.a.3	AATTGGGGACACATTGTCCTGG	3703	1374	37.11
Ets.b.1	ATATCTCGCCCACAAATGGAGG	1740	162	9.31
Ets.b.2	ATCTTGATAGGTTGCAGTCAGG	23999	1373	5.72
Ets.b.3	GAGTGGTCCAATCCAACTGTGG	975	373	38.30
Ets.b.4	CTTACCTTGCTACTTCATCTGG	916	1	0.12
Ffg4/6.1	CCCGCAATTCAGTACCTGTCGG	6540	2453	37.51
Ffg4/6.2	AGTCTTTATTCCGTAGCTCGGG	5014	982	19.59
Ffg4/6.3	CATCTCGGTGGACAATATGTGG	57606	11894	20.65
Ffg4/6.4	TGTTCATTTAGGCTTACCATGG	8182	838	10.24
Fgf8/17/18.1	CTCGGAGACATAGCCAGCGGGG	3899	1554	39.86
Fgf8/17/18.2	GGACGACTAGTTGGAAAGGTGG	6193	1933	31.21
Ffgr.1	CAACACACGTCTTACCTTCTGG	1209	292	24.15
Ffgr.2	CAAACGGAGCACCAAAAAGTGG	1744	530	30.39
Ffgr.3	TTGAACTGCACATCTAAACTGG	1459	34	2.33
Ffgr.4	TACATCCAGTTCGTAAGTATGG	6117	1486	24.29
Foxf.1	CCATTCGCTGCGACCGCTGCGG	2420	703	29.05
Foxf.2	ACTCTGCCCATCCCGCCAAAGG	27427	4830	17.61
Foxf.3	ACAGCCACCTCGCTTATGAAGG	5147	1477	28.70
Foxf.4	GTATATATAAGGGGCTCAGCGG	5820	2257	38.78
Foxg-r.1	AGAGAGGTCCCAATACAAAAGG	3559	67	1.87
Foxg-r.2	GTTGAGACTGAGATTGTGACGG	5472	2087	38.14

Foxg-r.3	ATTGGACCATGACGAGAGAGGG	4005	407	10.16
Foxg-r.4	TCAAAAGGAGACGAAGACGAGG	36132	124	0.34
Fzd5/8.1	TGCGAATGCACTTACCGACTGG	3624	547	15.09
Fzd5/8.2	CTTAGTCAAAGGGCCGAATTGG	3948	941	23.83
Fzd5/8.3	ACACCCCGATCTGTATACCTGG	33293	9853	29.59
Fzd5/8.4	GGTCATTCTGTCACTGACTTGG	459	85	18.52
Gata4/5/6.1	CTACGGTAGGGGTAGTAGTAGG	3475	1502	43.22
Gata4/5/6.2	GTAACGGTTGTGCTACAGTTGG	2621	360	13.74
Gata4/5/6.3	AGAAGTCCGGACGGAAACCCGG	30165	12270	40.68
Gata4/5/6.4	CGTAGAGGCTGACGTCACGAGG	1686	597	35.41
Hand.1	TTTGTATCCGACGGTACGTTGG	183	65	35.52
Hand.2	CCCGTACAGTCGCCGGTATCGG	20560	6369	30.98
Hand.3	GGGTTCCGCGCCGACTTAAAGG	7507	392	5.22
Hand.4	TACTTCGGTGTAAGTCAATTGG	2224	513	23.07
Hand-r.1	GCAACCGAAAACCCACACATGG	5118	628	12.27
Hand-r.2	AACATTGGGAGGGTAGCGGGGG	37245	8178	21.96
Hand-r.3	GAACCCTTGTTTGATAACTTGG	3602	317	8.80
Hand-r.4	CATCGTTAATTCTACTCTGYGG	7760	58	0.75
Htr7-r.1	AACTCCTGAGTTCTTGAGACGG	1812	1053	58.11
Htr7-r.2	TGATCTGGCAATCACAGCATGG	20142	12011	59.63
Htr7-r.3	ACACAGAAAGACACTGGGGTGG	1048	461	43.99
Htr7-r.4	AAGCAAAGGGGGACTCCCTGGG	1023	367	35.87
Islet.2	GATGCGGCAGGGGTCTGTAAGG	2390	1050	43.93
Lef1.1	AGAATGCCTCAGTTAAACTCGG	3547	1067	30.08
Lef1.2	CGTAAGATTGTCGCCTGTGTGG	1588	692	43.58
Mrf.1	ATGGGGTTTGTGGCGTAACTGG	2163	335	15.49
Mrf.2	GACGAACTGAAGAGATGCGTGG	23595	7750	32.85
Mrf.3	TAGTGTTGCCGCCCTCCGTCGG	5544	1808	32.61
Mrf.4	CTTAAGGGCGTCGTACGCTTGG	36222	1902	5.25
Neurog.1	CTTACCATTACGTCTTGTGAGG	2456	579	23.57
Neurog.2	TTAACTTGTTAATTGTCACAGG	5585	0	0.00
Nk4.1	GGAAGTGATTCATCCGTACAGG	5088	719	14.13
Nk4.2	AATCAGTTACTCAAGTCACGGG	41237	18969	46.00
Nk4.3	AACCAGATCTTGACCTGGGTGG	3125	1181	37.79
Nk4.4	CAAGACAAAACCTTAGAGCTGG	37168	5134	13.81
Rhod/f.1	TAGTTGGGGATGGTGGATGCGG	2882	1682	58.36
Rhod/f.2	TCAAAGGTGTACACGCCCACGG	3755	841	22.40
Rhod/f.3	TAACATTGTGGGATACGGCGGG	23845	9707	40.71
Rhod/f.4	CCAGAACGTAGAAATCCGATGG	625	22	3.52
Tle.b.1	TAGACTTACCTCTTGTGACAGG	5808	1570	27.03
Tle.b.2	ACCCGACAGCTCCCATACCYGG	3933	1035	26.32
Tle.b.3	GGAGGGAGTCCTGGAGGGTGGG	30004	11731	39.10
Tle.b.4	TTCCCTCCTCCTCCGCAGCCGG	288073	20041	6.96

Tll.1	CACCGATTGGGGACCGGGTTGG	546	150	27.47
Tll.2	CGTATTGGTACGGTTGCCTTGG	7606	180	2.37
Tll.3	AGCAAGGGCGGGTCGCAGAAGG	4367	97	2.22
Tll.4	GAATAGACACAGGCGACTTCGG	3243	347	10.70

suggests the actual efficacies of some sgRNAs may be up to 1.5-fold higher than their measured rates.

Analysis of unique indels generated by the activity of two different sgRNAs, Ebf.3 and Lef1.2, indicated a bias towards deletions rather than insertions, at a ratio of roughly 2:1 deletions;insertions (**Figure 1d**). However, these two sgRNAs generated different distributions of indel lengths, indicating indel position and size may depend on locus-specific repair dynamics.

Numerous studies have reported the potential off-target effects of CRISPR/Cas9 in different model systems (FU *et al.* 2013; HSU *et al.* 2013; PATTANAYAK *et al.* 2013; CHO *et al.* 2014). For this study, we were able to mostly select highly specific sgRNAs, owing to the low frequency of predicted off-target sequences in the small, A/T-rich *Ciona* genome (see **Discussion** for details). To test the assumption that off-target DSBs are unlikely for partial sgRNA seed-sequence matches, we analyzed the mutagenesis rates at two potential off-target sites that matched the protospacer at the 10 and 8 most PAM-proximal positions of the Ebf.3 and Fgf4/6.1 sgRNAs, respectively. We did not detect any mutations in 5,570 and 6,690 reads mapped to the two loci, respectively, suggesting high specificity of the sgRNA:Cas9 complex to induce DSBs only at sites of more extensive sequence match in developing *Ciona* embryos.

Identifying sequence features correlated with sgRNA efficacy

Recent studies in different model organisms have examined the nucleotide preferences amongst sgRNAs inducing high or low rates of mutagenesis (DOENCH *et al.* 2014; GAGNON *et al.* 2014; REN *et al.* 2014; CHARI *et al.* 2015; MORENO-MATEOS *et al.* 2015). However, *Ciona* rearing conditions differ from those of other model systems, being a marine invertebrate that develops optimally at 13-20°C. Moreover, most CRISPR/Cas9-based experiments in *Ciona* rely on *in vivo* transcription of sgRNAs built with a modified "F+E" backbone (CHEN *et al.* 2013) by a *Ciona*-specific U6 small RNA promoter transcribed by RNA Polymerase III (PolIII)(NISHIYAMA AND FUJIWARA 2008). Therefore, in order to identify potential sequence features in target sequences that might contribute towards a higher or lower mutagenesis rate specifically in *Ciona* embryos, we analyzed our dataset for potential correlations between target sequence composition and sgRNA-specific mutagenesis rate.

We hypothesized that, if mutagenesis efficacy can be predicted by nucleotide composition at defined positions in the protospacer and flanking sequences, then comparing the target sequences of the most and least active sgRNAs in our dataset should reveal features that affect CRISPR/Cas9 efficacy in *Ciona*. To that effect, we performed nucleotide enrichment analyses for the top and bottom 25% sgRNAs ranked by measured mutagenesis efficacy (**Figure 2a,b, Supplementary Figure 1**)(SCHNEIDER AND STEPHENS 1990; CROOKS *et al.* 2004). For the top 25% sgRNAs, guanine was overrepresented in the PAM-proximal region, while the ambiguous nucleotide of the PAM ('N' in 'NGG') was enriched for cytosine. We also observed an overall depletion of thymine in the protospacer sequence for the top 25% sgRNAs, likely due to premature termination of PolIII-driven transcription as previously demonstrated (WU *et al.* 2014). Among the bottom 25% sgRNAs, we observed a higher representation of cytosine at

nucleotide 20 of the protospacer (**Figure 2b, Supplementary Figure 1**). All these observations are consistent with the inferences drawn from previous studies, suggesting that certain sgRNA and target sequence features that influence Cas9:sgRNA-mediated mutagenesis rates are consistent across different metazoans (GAGNON *et al.* 2014; CHARI *et al.* 2015; MORENO-MATEOS *et al.* 2015).

Machine learning reveals rational design principles for highly active sgRNAs

CRISPRScan is an online tool for rational sgRNA design, based on large-scale sgRNA validation in zebrafish embryos (MORENO-MATEOS *et al.* 2015). However, we reasoned that fundamental differences in sgRNA backbone and delivery method between CRISPR/Cas9 experiments in zebrafish and *Ciona* would necessitate a novel algorithm to identify and construct sgRNAs with high activity specifically in *Ciona.*

In order to uncover additional sequence-based determinants of sgRNA activity, especially synergistic effects of nucleotide features at disjointed positions in and around the targeted sequence, we used a linear regression-based approach and trained an L1 regularized LASSO (Least Absolute Shrinkage and SelectiOn) regression model to identify which single- and dual nucleotide features of the protospacer, PAM, and flanking sequences affected mutagenesis efficacy the most (**Supplementary Table 3,4**)(TIBSHIRANI 1996). We used L1 regularization in our model to efficiently perform feature selection in a sparse feature space. We identified sequence feature effects by fitting a unified training set comprising the top and bottom 25% sgRNAs. This regression model, which we named "TuniCUT", was then used to predict the efficacy rates of the remaining sgRNAs that were left out of the training set. TuniCUT predictions were plotted against the actual measured efficacy rates to visualize the accuracy of

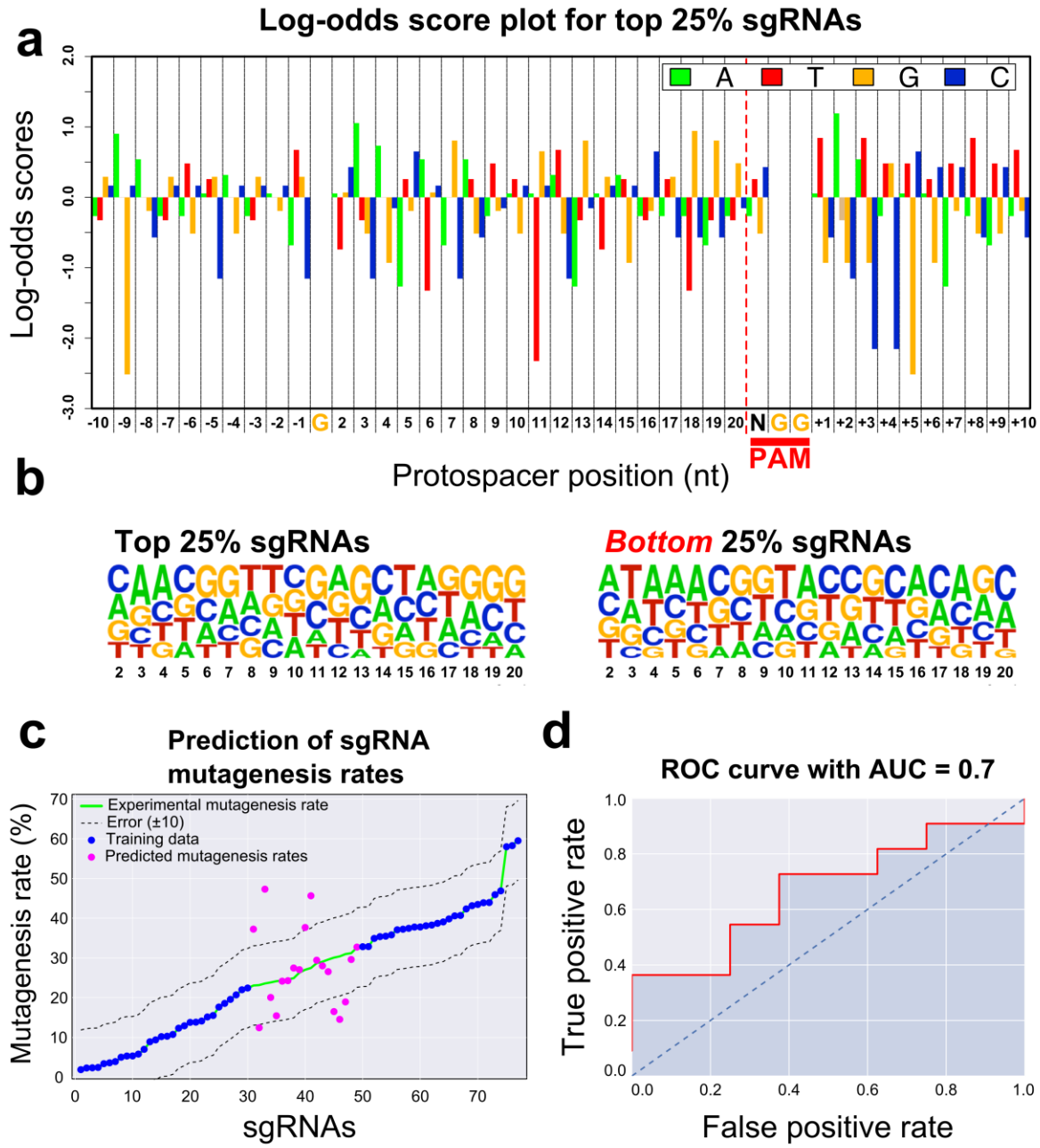


Figure 2. Correlations between sgRNA sequence composition and mutagenesis efficacy

a) Log-odds scores depicting the frequency of occurrence for nucleotides in the top 25% most effective sgRNAs, at all positions of the protospacer, PAM, and flanking regions. Position "1" of the protospacer has been omitted from the analysis, due to this always being "G" for PolIII-dependent transcription of U6-promoter-based vectors. Likewise, the "GG" of the PAM has also been omitted, as this sequence is invariant in all targeted sites. **b**) WebLogos representing the nucleotide composition at each variable position of the protospacer (nt 2-20, X axis), in top 25% and bottom 25% performing sgRNAs. **c**) All sgRNAs in this study plotted by their mutagenesis efficacy rates (Y axis). The top 25%

and bottom 25% sgRNAs constituted the training set (blue dots) for the TuniCUT prediction model. TuniCUT was then used to predict the efficacies of the remaining sgRNAs (magenta dots). **d**) The performance of TuniCUT was tested by plotting the Area Under the Curve of the Receiver Operator Characteristic (AUC-ROC). The area was calculated based on the discrimination ability of a classifier which was fitted to the training data using key parameters identified by Lasso regression.

our model (**Figure 2c**). Based on variable uptake of plasmid DNA by electroporation (ZELLER *et al.* 2006; STOLFI *et al.* 2014), we accommodated an arbitrary error of ±10%.

In a previous study, highly penetrant, tissue-specific loss-of-function phenotypes in F0 embryos were elicited using Ebf.3 sgRNA, which had a measured mutagenesis efficacy of 37% (STOLFI *et al.* 2014). We thus reasoned that an sgRNA with an efficacy of ~25% would be acceptable for loss-of-function assays. Based on this rationale, we defined a threshold of 25%, which was close to the median value of measured mutagenesis rates (26%), and the training set was binned, with values 0 and 1 representing sgRNAs with an estimated efficacy of less than 25% ("bad") and greater than or equal to 25% ("good"), respectively. We fit a logistic regression classifier using this transformed training set, and used an L2 regularization to minimize prediction error (NG 2004). Once the model was fitted to the data, we tested the performance of this prediction model by plotting sensitivity (True Positive Rate) versus specificity (False Positive Rate) as the Receiver Operator Characteristic (ROC) curve (HANLEY AND MCNEIL 1982). The area under this curve (AUC) yielded a moderately predictive value of 0.7 (**Figure 2d**). In other words, our model has a 70% probability of discriminating between a "good" (efficacy ≥ 25%) and a "bad" (efficacy < 25%) sgRNA, which represents a ~40% increase in the likelihood of identifying efficient sgRNAs as compared to chance.

Multiplexed targeting with CRISPR/Cas9 generates large deletions in the *Ciona* genome

Large deletions of up to 23 kb of intervening DNA resulting from NHEJ between two CRISPR/Cas9-induced DSBs have been reported in *Ciona* (ABDUL-WAJID *et al.* 2015). For functional analyses of protein-coding genes, such deletions would more likely produce null mutations than small deletions resulting from the action of lone sgRNAs. To test whether we could cause tissue-specific, homozygous, large deletions in F0 embryos, we targeted the forkhead/winged helix transcription-factor-encoding gene *Foxf* (**Figure 3a**), which contributes to TVC migration in *Ciona* (BEH et al. 2007). We co-electroporated *Eef1a1>nls::Cas9::nls* with sgRNA vectors Foxf.4 and Foxf.2 (with induced mutagenesis rates of 39 and 18%, respectively; **Table 2**, **Figure 3**). We extracted genomic DNA from electroporated embryos and PCR-amplified the sequence spanning both target sites. We obtained a specific ~300 bp PCR product corresponding to the amplified region missing the ~2.1 kbp of intervening sequence between the two target sites. Cloning the deletion band and sequencing individual clones confirmed that the shorter PCR product corresponds to a deletion of most of the *Foxf* first exon and 5' *cis*-regulatory sequences (BEH *et al.* 2007). We did not detect this deletion using genomic DNA extracted from embryos electroporated with either sgRNA alone. Similar deletion PCR products were observed, cloned, and sequenced for other genes including *Nk4, Fgfr, Mrf, Htr7-related, Bmp2/4,* and *Hand,* using pairs of highly mutagenic sgRNAs (**Supplementary Figure 2**). The largest deletion recorded was ~3.6 kbp, with sgRNAs Nk4.2 (46% efficacy) and Nk4.3 (38% efficacy), entirely removing the sole intron of *Nk4* and small portions of the flanking exons.

Foxf is expressed in TVCs and head epidermis, the latter of which is derived exclusively from the animal pole (NISHIDA 1987; IMAI *et al.* 2004; PASINI *et al.* 2006; BEH *et al.* 2007). Because

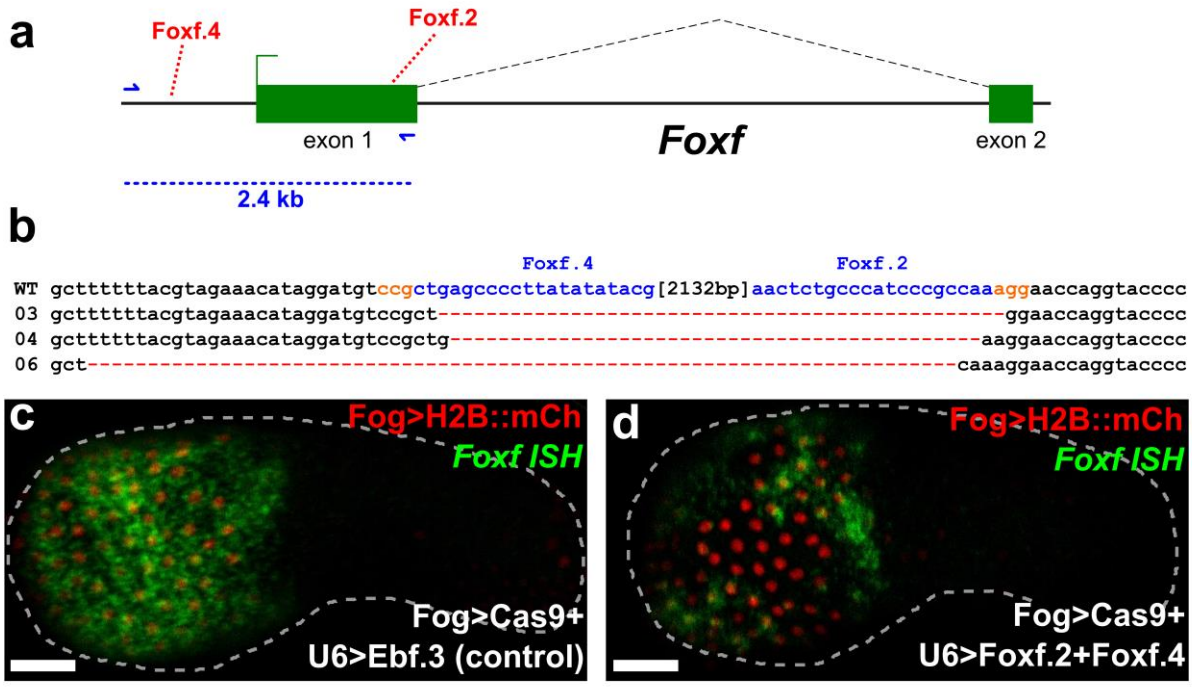


Figure 3. Combinatorial targeting of *Foxf* results in large deletions

a) Diagram of *Foxf* locus, showing positions targeted by *Foxf.4* and *Foxf.2* sgRNAs. *Foxf.4* targets a non-coding, *cis*-regulatory sequence 881 base pairs (bp) upstream of the transcription start site of *Foxf. Foxf.2* targets a coding sequence in exon 1 of *Foxf.* The distance between the target sites is 2132 bp, and encompasses most of exon 1, the core promoter, and *cis*-regulatory modules that drive *Foxf* expression in the head epidermis and trunk ventral cells (TVCs) (BEH *et al.* 2007). Blue arrows indicate primers used to amplify the region between the target sites. In wild-type embryos, the resulting PCR product is ~2.4 kilobase pairs (kbp). **b**) Alignment of cloned PCR products amplified using the primers indicated in (a), from wild-type (wt) embryos, and from embryos electroporated with 25 µg *EF1α>nls::Cas9::nls* and 50 µg each of *U6>Foxf.2* and *U6>Foxf.4.* Colonies 03, 04, and 06 shown containing large deletions between the approximate sites targeted by the two sgRNAs, indicating non-homologous end-joining (NHEJ) repair from two separate double stranded break events as a result of combinatorial action of *Foxf.2* and *Foxf.4* sgRNAs. **c**) *In situ* hybridization for *Foxf* (green) showing strong expression throughout the head epidermis in embryos electroporated with 10 µg *Fog>H2B::mCherry* (red), 50 µg *Fog>nls::Cas9::nls* and 45 µg of *U6>Ebf.3. Foxf* expression is essentially wild-type, as Ebf function is not required for activation of *Foxf* in the epidermis. **d**) *In situ* hybridization for *Foxf* (green) showing patchy expression in the head epidermis of embryos electroporated with 10 µg *Fog>H2B::mCherry* (red), 50 µg *Fog>nls::Cas9::nls* and 45 µg each of *U6>Foxf.2* and *U6>Foxf.4*. Loss of

in situ signal in some transfected head epidermis cells indicates loss of *Foxf* activation, presumably through deletion of all or part of the upstream *cis*-regulatory sequences by CRISPR/Cas9. Scale bars = 25 μm.

the ~2.1 kbp deletion introduced in the *Foxf* locus excised the epidermal enhancer and basal promoter (BEH *et al.* 2007), we sought to examine the effects of these large deletions on *Foxf* transcription. We used the *cis*-regulatory sequences from *Zfpm* (also known as *Friend of GATA*, or *Fog*, and referred to as such from here onwards) to drive Cas9 expression in early animal pole blastomeres (ROTHBÄCHER *et al.* 2007). We electroporated *Fog>nls::Cas9::nls* together with Foxf.2 and Foxf.4 sgRNA vectors and *Fog>H2B::mCherry,* and raised embryos at 18°C for 9.5 hpf (early tailbud, embryonic stage 20). We performed whole mount mRNA *in situ* hybridization to monitor *Foxf* expression, expecting it to be silenced in the epidermis by tissue-specific CRISPR/Cas9-induced homozygous deletions of the *Foxf cis*-regulatory sequences (**Figure 3a**). Indeed, we observed patches of transfected head epidermal cells (marked by H2B::mCherry) in which *Foxf* expression was reduced or eliminated (**Figure 3d**). This was in contrast to the uniform, high levels of *Foxf* expression observed in "control" embryos electroporated with Ebf.3 sgRNA (*Ebf* is unlikely to be involved in *Foxf* regulation in the epidermis where it is not expressed, **Figure 3c**). Taken together, these results indicate that, by co-electroporating two or more sgRNAs targeting neighboring sequences, one can routinely generate homozygous, large deletions in the *Ciona* genome in a tissue-specific manner.

Rapid generation of sgRNA expression cassettes ready for embryo transfection

CRISPR/Cas9 is an efficient and attractive system for targeted mutagenesis in *Ciona,* but cloning individual sgRNA vectors is a labor-intensive, rate-limiting step. To further expedite

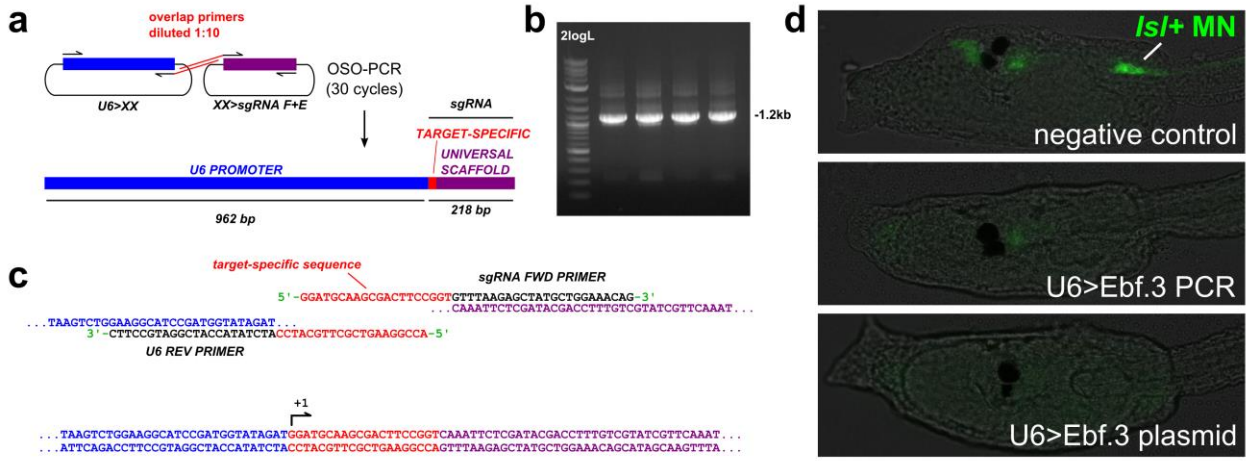


Figure 4. One-step Overlap Polymerase Chain Reaction (OSO-PCR) for the high-throughput construction of sgRNA expression cassette libraries

a) Diagram of OSO-PCR for amplification of *U6>sgRNA* expression cassettes in which the target-specific sequence of each (red) is encoded in complementary overhangs attached to universal primers. 1:10 dilution of these primers ensures that the overlap product, the entire U6>sgRNA cassette, is preferentially amplified (see methods for details). **b**) Agarose gel electrophoresis showing products of four different *U6>sgRNA* OSO-PCRs. The desired product is ~1.2 kilobase pairs (kbp) long. 2logL = NEB 2-Log DNA ladder. **c**) Detailed diagram of how the overlap primers form a target-specific bridge that fuses universal U6 promoter and sgRNA scaffold sequences. **d**) Larvae co-electroporated with *Sox1/2/3>nls::Cas9::nls, Islet>eGFP,* and either 25 µl (~2.5 µg) unpurified *U6>NegativeControl* PCR (top panel), 25 µl (~2.5 µg) unpurified *U6>Ebf.3* PCR (middle panel), or 25 µg *U6>Ebf.3* plasmid (bottom panel). *Islet>eGFP* reporter plasmid is normally expressed in the A10.57 motor neuron ("Isl+ MN", green), which is dependent on Ebf function. *Islet>eGFP* was expressed in MNs in 75 of 100 embryos. In embryos electroporated with unpurified *U6>Ebf.3* PCR products or *U6>Ebf.3* plasmid, only 16 of 100 and 17 of 100 embryos, respectively, had *Islet>eGFP* expression in MNs. This indicates similar loss of Ebf function *in vivo* by either unpurified PCR or purified plasmid sgRNA delivery method.

CRISPR/Cas9 experiments, we adapted a one-step overlap PCR (OSO-PCR) protocol to generate U6 promoter>sgRNA expression "cassettes" for direct electroporation without purification (**Figure 4a-c, Supplementary Figure 3,** see **Materials and Methods** and **Supplementary Protocol** for details). We tested the efficacy of sgRNAs expressed from these unpurified PCR products, by generating such expression cassettes for the validated Ebf.3 sgRNA (**Table 2**)(STOLFI *et al.* 2014). We electroporated *Eef1a1>nls::Cas9::nls* and 25 µl (corresponding to ~2.5 µg, see **Materials and Methods** and **Supplementary Figure 4**) of unpurified, *U6>Ebf.3* sgRNA OSO-PCR or *U6>Ebf.3* sgRNA traditional PCR products (total electroporation volume: 700 µl). Next-generation sequencing of the targeted Ebf.3 site revealed mutagenesis rates similar to those obtained with 75 µg of *U6>Ebf.3* sgRNA plasmid (**Table 2**). This was surprising given the much lower total amount of DNA electroporated from the PCR reaction relative to the plasmid prep (2.5 µg vs. 75 µg), and suggests that our sequencing assay is measuring the near-maximal mutagenesis efficacy of each sgRNA.

To assess whether unpurified sgRNA PCR cassettes could be used in CRISPR/Cas9-mediated loss-of-function experiments in F0 embryos, we assayed the expression of an *Islet>GFP* reporter in A10.57 motor neurons, which depends upon Ebf function (STOLFI *et al.* 2014). Indeed, *Islet>GFP* expression was downregulated in embryos electroporated with *Sox1/2/3>nls::Cas9::nls* and 25 µl of unpurified *U6>Ebf.3* traditional PCR or 25µg *U6>Ebf.3* plasmid, but not with 25 µl (~2.5 µg) of unpurified *U6>Negative Control* sgRNA PCR cassette (**Figure 4d**). Taken together, these results indicate that unpurified PCR products can be used in lieu of plasmids to express sgRNAs for tissue-specific CRISPR/Cas9-mediated mutagenesis in *Ciona* embryos.

Pre-emptive design and scoring of all predicted sgRNAs for the *Ciona* genome

While *Ciona* researchers will find the TuniCUT algorithm and the OSO-PCR method useful for sgRNA expression cassette assembly, we hoped to further empower the community by pre-emptively designing all possible sgRNAs targeting at least one site in the *Ciona* genome (excluding those containing the PolIII-terminating sequence of more than three "T" in a row). We computationally identified 4,853,589 such sgRNAs (4,249,756 of them targeting a single site in the genome). We called this collection of predictions the **C**RISPR/Cas9-**i**nduced ***Ci****ona* **K**nock-**O**ut (**Ci^{2}KO**) sgRNA Library. We used TuniCUT to predict the mutagenesis efficacy rates of each sgRNA in this library, and further pre-designed the requisite oligonucleotides to be used as primers for OSO-PCR to construct each sgRNA expression cassette. We have compiled all sgRNA sequences, their predicted efficacy rates, and corresponding OSO-PCR oligonucleotide sequences and have made this available in several formats (**Figure 5**, see **Materials and methods** for details), including as a genome annotation track hosted on the ANISEED *Ciona* genome browser (BROZOVIC *et al.* 2015).

Discussion

We have built a library of 83 plasmid vectors for the *in vivo* expression of sgRNAs targeting 23 genes expressed in the cardiopharyngeal mesoderm and surrounding tissues, mostly hypothesized to be involved in regulating the specification of heart and/or pharyngeal muscles in *Ciona*, even though many have complex expression patterns and probably pleiotropic functions. We have also established a reliable protocol for the validation of sgRNA efficacy in electroporated *Ciona* embryos by next-generation sequencing. This allowed us to estimate the activity of most of these sgRNAs, which are ready to be used for future functional studies.

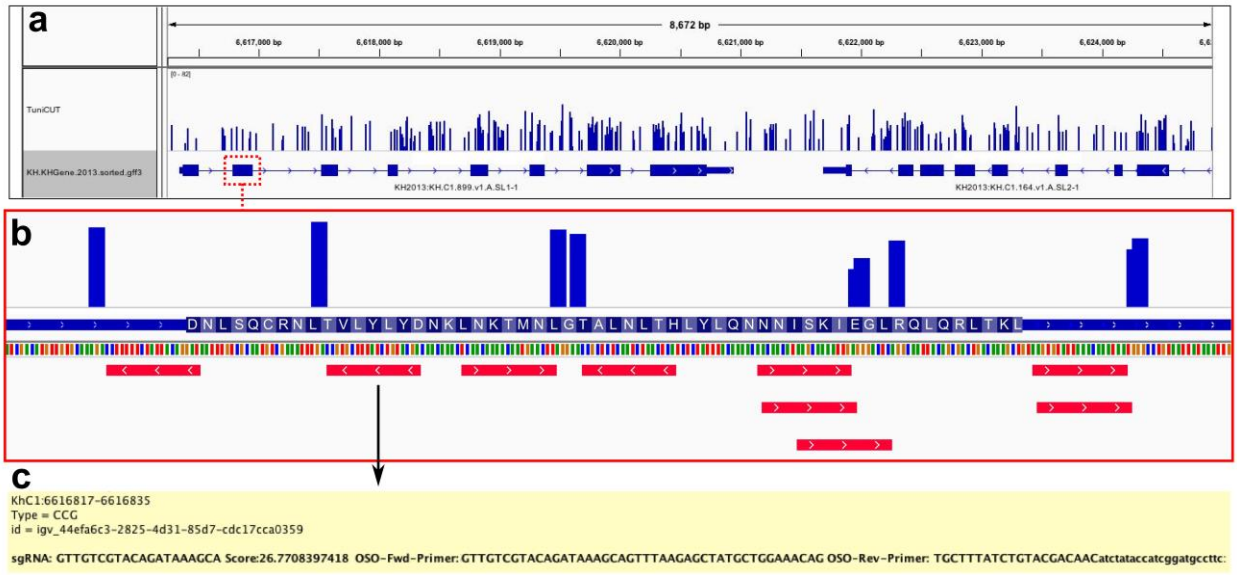


Figure 5. Whole-genome prediction and scoring of sgRNAs

a) Screen shot of whole-genome pre-designed sgRNAs aligned to a part of the *Ciona* genome, viewed in the IGV browser. Bars in "TuniCUT" track represent each sgRNA, relative to KyotoHoya (KH) gene models depicted in bottom track. Height of sgRNA bars indicates predicted TuniCUT score. **b)** Zoom-in on the sequence of exon 2 (red dashed box in panel "a") of the KH.C1.899 gene (*Protein Phosphatase 1, Regulatory Subunit 42,* or *Ppp1r42*). Underneath the representation of *Ppp1r42* exon 2 and its amino acid sequence is the DNA sequence (small multicolored vertical bars) and the position all sgRNAs (red horizontal bars) predicted in this region (red horizontal bars). On top are TuniCUT scores (blue vertical bars) for each sgRNA, centered on each PAM. **c)** Detailed information relating to each sgRNA is immediately available, including protospacer sequence, genomic position, TuniCUT score, and sequences of primers for sgRNA cassette construction by OSO-PCR. These sequences can be copied by right-clicking on the desired sgRNA, and selecting "Copy Details to Clipboard".

By analyzing correlations between protospacer and flanking sequence nucleotide composition and sgRNA mutagenesis efficacy, we revealed principles that may contribute to Cas9:sgRNA activity. This prompted us to develop TuniCUT, an algorithm to predict relative sgRNA mutagenesis efficacy rates in *Ciona.* We demonstrate that TuniCUT functions well as a classifier

that allowed researchers to select, with greater confidence, potentially "good" sgRNAs with enough mutagenic activity for functional studies in F0.

Some of the predictive sequence features have been identified in previous CRISPR/Cas9-mediated mutagenesis screens performed in other metazoan model organisms, suggesting that these are determined by the intrinsic properties of sgRNAs and/or Cas9 (GAGNON *et al.* 2014; CHARI *et al.* 2015; MORENO-MATEOS *et al.* 2015). For instance, mutagenesis rate increases with increasing guanine content in the PAM-proximal nucleotides of the sgRNA, postulated to be due to increased sgRNA stability by G-quadruplex formation (MORENO-MATEOS *et al.* 2015). This would explain the specific enrichment for guanine but not cytosine, even if both could in theory augment sgRNA folding or binding to target DNA. Additionally, we encountered a depletion of thymine and cytosine in the PAM-proximal nucleotides of the protospacers for highly active sgRNAs, which has also been reported in other organisms. The strong negative correlation between sgRNA efficacy and thymine content of the protospacer is easily attributed to our use of the PolIII-dependent U6 promoter to express our sgRNAs. It has been shown that termination of transcription by PolIII is triggered by a string of thymines, and a high number of thymines clustered in the protospacer could result in lower sgRNA expression level due to premature termination of sgRNA transcription (STOLFI *et al.* 2014; WU *et al.* 2014). Likewise, depletion of adenine has been attributed to lower stability of sgRNAs (MORENO-MATEOS *et al.* 2015), suggesting that CRISPR/Cas9 mutagenesis efficacies are primarily determined by sgRNA transcription and degradation rates, which will vary depending on the species studied and the mode of sgRNA delivery (e.g. *in vitro* vs. *in vivo* synthesis).

Despite these general trends, several highly active sgRNAs had sequence features that defied the consensus, for instance by having more C than G at a PAM-proximal position. This suggests that there are multiple, possibly additive or synergistic factors that determine the mutagenesis efficacy, only one of which is primary sequence composition of the sgRNA or target. What those additional factors are will be an important topic of study as CRISPR/Cas9-based approaches are expanded to address additional questions in basic research as well as for therapeutic purposes.

Despite legitimate concerns about potential off-target effects for functional studies, we were not able to detect CRISPR/Cas9-mediated mutagenesis at two potential off-target sites for the sgRNAs Ebf.3 and Fgf4/6.1. For the remainder of the sgRNAs, we purposefully selected those with zero predicted off-targets. This was possible in *Ciona* due to two factors. First, the *Ciona* genome is significantly smaller than the human genome and most metazoans, resulting in a lower number of identical protospacer seed sequences. Second, the GC content of the *Ciona* genome is only 35% as compared to 65% in humans, which results in a lower frequency of PAMs. Based on these considerations, we predict off-target effects to be less pervasive in *Ciona* than in other model organisms with more complex, GC-rich genomes.

We designed all possible sgRNAs targeting the compact *Ciona* genome and calculated their predicted relative mutagenesis rates, which we have made available either as downloadable files for visualizing locally on the IGV browser (ROBINSON *et al.* 2011), or as a track on the ANISEED genome browser (BROZOVIC *et al.* 2015). This will allow researchers to locally browse for sgRNAs with predicted high activity targeting their loci of interest and select the corresponding pre-designed OSO-PCR oligonucleotide primers for rapid, efficient synthesis and

transfection. We expect this resource to facilitate the scaling of CRISPR/Cas9-mediated targeted mutagenesis and enable genome-wide screens for gene function in *Ciona.*

Materials and Methods

Target sequence selection and sgRNA design

23 genes from *Ciona robusta* (formerly *Ciona intestinalis* type A)(HOSHINO AND TOKIOKA 1967; BRUNETTI *et al.* 2015) hypothesized to be important for cardiopharyngeal mesoderm development were shortlisted (**Table 1**) and one to four sgRNAs targeting non-overlapping sequences per gene were designed, for a total of 83 sgRNA vectors (**Table 2**). Two sgRNAs were designed to target the neurogenic bHLH factor Neurogenin, a gene that is not hypothesized to be involved in cardiopharyngeal development. Target sequences were selected by searching for N_{19} + NGG (protospacer + PAM) motifs and screened for polymorphisms and off-target matches using the GHOST genome browser and BLAST portal (SATOU *et al.* 2005; SATOU *et al.* 2008). Potential off-targets were also identified using the CRISPRdirect platform (NAITO *et al.* 2015). sgRNA expression plasmids were designed for each of these protospacers and constructed using the *U6>sgRNA(F+E)* vector as previously described (STOLFI *et al.* 2014), as well as a "Negative Control" protospacer that does not match any sequence in the *C. robusta* genome (5'-GCTTTGCTACGATCTACATT-3'). Stretches of more than four thymine bases (T) were avoided due to potential premature transcription termination. Candidate sgRNAs with a partial PAM-proximal match of 13 bp or more were also discarded due to off-target concerns. All sgRNAs were designed to target protein-coding, splice-donor, or splice-acceptor sites, unless specifically noted. We preferred more 5' target sites, as this provides a greater probability of generating loss-of-function alleles.

Electroporation of *Ciona* embryos

DNA electroporation was performed on pooled, dechorionated zygotes (1-cell stage embryos) from *C. robusta* adults collected from San Diego, CA (M-REP) as previously described (CHRISTIAEN *et al.* 2009). All sgRNA plasmid maxipreps were individually electroporated at a final concentration of 107 ng/µl (75 µg in 700 µl) concentration together with *Eef1a1>nls::Cas9::nls* plasmid (STOLFI *et al.* 2014) at 35.7 ng/µl (25 µg in 700 µl) concentration. For testing *U6>Ebf.3* PCR or OSO-PCR, 25 µl was used instead of sgRNA plasmid. Embryos were then rinsed once in artificial sea water, to remove excess DNA and electroporation buffer, and grown at 18°C for 16 hours post-fertilization.

Embryo lysis

After 16 hpf, each pool of embryos targeted with a single sgRNA + Cas9 combination was washed in one sea water exchange before lysis, to remove excess plasmid DNA, and transferred to a 1.7 ml microcentrifuge tube. Excess sea water was then removed and embryos were lysed in 50 µl of lysis mixture prepared by mixing 500 µL of DirectPCR Cell Lysis Reagent (Viagen Biotech Inc., Los Angeles, CA, Cat # 301-C) with 1 µl of Proteinase K (20 mg/ml, Life Technologies, Carlsbad, CA). The embryos were thoroughly mixed in lysis mixture and incubated at 68°C for 15 minutes, followed by 95°C for 10 minutes.

PCR amplification of targeted sequences

Targeted sequences were individually PCR-amplified directly from lysate from embryos targeted with the respective sgRNA, and from "negative control" lysate (from embryos electroporated with *Eef1a1>nls::Cas9::nls* and *U6>Negative Control* sgRNA vector). Primers (**Supplementary Table 2**) were designed to flank target sites as to obtain PCR products in the

size range of 108-290bp with an exception of the sequence targeted by Ebf.3 ("Ebf.774" in Stolfi et al. 2014) and Ebf.4 sgRNAs, for which the designed primers resulted in a product size of 350 bp. Potential off-target sites predicted for sgRNAs Ebf.3 (CTCGCAACGGGGACAACAGGGGG, genome position KhC8:2,068,844-2,068,866) and Fgf4/6.1 (TATTTTAATTCTGTACCTGTGGG, genome position KhC9:6,318,421-6,318,443) were amplified to test for off-target CRISPR/Cas9 activity with the primers: 5'-CCAGCACTTCAGAGCAATCA-3' and 5'- TGACGTCACACTCACCGTTT-3' (Ebf.3), and 5'-AACGATTGTCCATACGAGGA-3' and 5'-ACTTCCCAACAGCAAACTGG-3' (Fgf/6.1).

For each targeted sequence, 12.5 µL PCR reactions were set up with final concentrations of 600 nM each primer, 300 µM dNTPs, 1 mM $MgSO_4$, 2X buffer, and 0.05 U/µl Platinum Pfx DNA polymerase (Life Technologies), and subjected to the following PCR program: an initial cycle of 10 minutes at 95°C, followed by 30 cycles of 30 seconds at 94°C, 30s at 60°C and 30s at 68°C, and a final cycle of 3 minutes at 68°C. PCR reactions were quickly checked on an agarose gel for the presence/absence of amplicon. Those that resulted in a single band were not initially purified. For those reactions with more than one band, the correct amplicon (selected based on expected size) was gel purified using a Nucleospin Gel Clean-up Kit (Macherey-Nagel, Düren, Germany). Purified and unpurified PCR products were then pooled for subsequent processing. The majority of PCR products amplified from larvae treated with Cas9 + gene-targeting sgRNA were pooled in Pool 1. All products from larvae treated with Cas9 + "negative control" sgRNA were pooled in Pool 2. For those sequences targeted by distinct sgRNAs but amplified using the same set of flanking primers, their PCR products were split into separate pools, as to allow for separate efficacy estimates.

Sequencing library preparation

The PCR product pools were electrophoresed on ethidium bromide-stained, 1% agarose gel in 0.5X Tris-Acetate-EDTA (TAE) buffer and a band of ~150-300 bp was excised (Nucleospin gel and PCR cleanup kit, Macherey-Nagel). 102-235 ng of each pool was used as a starting material to prepare sequencing libraries (protocol adapted from http://wasp.einstein.yu.edu/index.php/Protocol:directional_WholeTranscript_seq). Ends were repaired using T4 DNA polymerase (New England Biolabs, Ipswich, MA) and T4 Polynucleotide Kinase (New England Biolabs), and then A-tailed using Klenow fragment (3'→5' exo-) (New England Biolabs) and dATP (Sigma-Aldrich, St. Louis, MO). Each pool was then ligated to distinct barcoded adapters. (NEXTflex DNA Barcodes - BioO Scientific Cat# 514101). The six barcodes used in this study were: CGATGT, TGACCA, ACAGTG, GCCAAT, CAGATC and CTTGTA. At this step, the adapter-ligated DNA fragments were purified twice using Ampure XP beads (Beckman Coulter, Brea, CA). The final amplification, using primers included with NEXTflex adapters, was done using the PCR program: 2 minutes at 98°C followed by 8 cycles of 30 seconds at 98°C; 30 seconds at 60°C; 15 seconds of 72°C, followed by 10 minutes at 72°C. Ampure XP bead-based selection was performed twice, and the libraries were quantified using qPCR. The libraries were then mixed in equimolar ratio to get a final DNA sequencing library concentration of 4 nM. The multiplexed library was sequenced by Illumina MiSeq V2 platform (Illumina, San Diego, CA) using 2x250 paired end configuration.

Next generation sequencing data analysis

FastQ files obtained from sequencing were de-multiplexed and subjected to quality control analysis. FastQ reads were mapped to the 2008 KyotoHoya genome assembly (SATOU *et al.*

2008) by local alignment using Bowtie2.2 (LANGMEAD AND SALZBERG 2012). Single end reads were also mapped to a reduced genome assembly consisting of only those scaffolds containing the targeted genes. This allowed for a much faster and accurate alignment using Bowtie2.2. The SAM file generated was converted into a BAM file using *samtools* (LI *et al.* 2009). The BAM file was sorted and indexed to visualize reads on Integrative Genomics Viewer (IGV) (ROBINSON *et al.* 2011). Most mutagenesis rates were obtained by counting indels in IGV. For some targets with partially overlapping aplicon sequences, custom Python scripts were written to parse the BAM file to get estimated rate of mutagenesis. Since a successful CRISPR/Cas9-mediated deletion or insertion should eliminate or disrupt all or part of the protospacer + PAM sequence (jointly termed the "pineapple"), we simply looked for mapped reads in which the pineapple was not fully present. When appropriate, the rate of naturally occurring indels around each target, as detected in reads from "negative control" embryos, was subtracted from the raw efficacy rates.

Regression model for predicting sgRNA mutagenesis efficiency

We based the sequence feature space on a 43-mer target site consisting of the protospacer sequence, PAM, and a 10-nucleotide contextual region flanking the protospacer to get a comprehensive set of features to inform a predictive mathematical model (10 nt 5' flanking + Protospacer + PAM + 10 nt 3' flanking). PAM-proximal G or C content, 43 single-nucleotide and 903 dual nucleotide (two nucleotides in all possible combinations at 43 positions) features were one-hot encoded using the Python scikit-learn machine learning module as a binary vector with a value of 1 representing the presence of a particular feature, while a value of 0 representing the absence of one. An L1 regularized lasso regression model was used to select 75 features that had the maximum effect on mutagenesis efficiency. The top and bottom 25% data points were

then used to train the model. Cross-validation was performed with 200 iterations but was not included in the analysis due to overfitting.

Python scripts

Custom python scripts were written to perform next-generation sequencing data analysis, regression modeling, and whole-genome sgRNA design. These are available upon request. Matplotlib (http://matplotlib.org) was used for plotting, Numpy (http://numpy.org) and Pandas (http://pandas.pydata.org) were used for data mining, and scikit-learn (http://scikit-learn.org) was used for machine learning.

Combinatorial sgRNA electroporation to induce large deletions

Embryos were electroporated with 25 µg *Eef1a1>nls::Cas9::nls* and two vectors from the set of validated sgRNA plasmids for each targeted gene (50 µg per sgRNA vector). Embryos were grown for 12 hpf at 18°C, pooled, and genomic DNA extracted from them using QIAamp DNA mini kit (Qiagen). Deletion bands were amplified in PCR reactions using Pfx platinum enzyme as described above (see "**PCR amplification of targeted sequences**") and a program in which the extension time was minimized to 15 seconds only, in order to suppress the longer wild-type amplicon and promote the replication of the smaller deletion band. Primers used were immediately flanking the sequences targeted by each pair of sgRNAs (**Supplementary Table 2**). Products were purified from agarose gels, A-overhung and TOPO-cloned. Colonies were picked, cultured, prepped and sequenced.

Synthesis and electroporation of unpurified sgRNA PCR expression cassettes

U6>Ebf.3 and *U6>Negative Control* sgRNA expression cassettes were amplified from their respective plasmids using the primers U6 forward (5'-TGGCGGGTGTATTAAACCAC-3') and sgRNA reverse (5'-GGATTTCCTTACGCGAAATACG-3') in reactions of final concentrations of 600 nM each primer, 300 µM dNTPs, 1 mM $MgSO_4$, 2X buffer, and 0.05 U/µl Platinum Pfx DNA polymerase (Life Technologies), and subjected to the following PCR program: an initial cycle of 3 minutes at 95°C, followed by 30 cycles of 30 seconds at 94°C, 30s at 55°C and 2 minutes at 68°C, and a final cycle of 5 minutes at 68°C. *U6>sgRNA(F+E)::eGFP* (STOLFI *et al.* 2014) was amplified as above but using Seq forward primer (5'-AGGGTTATTGTCTCATGAGCG-3') instead. For phenotyping Ebf-dependent expression of *Islet* reporter in A10.57 motor neurons (STOLFI *et al.* 2014), embryos were co-electroporated with 35 µg of *Sox1/2/3> nls::Cas9::nls,* 5 µg of *Sox1/2/3>H2B::mCherry,* 30 µg of *Isl>eGFP ,* and either 25 µg of *U6>Ebf.3* plasmid or 25 µl (~2.5 µg) of unpurified PCR product.

sgRNA expression cassette assembly by One-step Overlap PCR (OSO-PCR)

sgRNA PCR cassettes were constructed using an adapted One-step Overlap PCR (OSO-PCR) protocol (URBAN *et al.* 1997). Basically, a desired protospacer sequence is appended 5' to a forward primer (5'-GTTTAAGAGCTATGCTGGAAACAG-3') and its reverse complement is appended 5' to a reverse primer (5'-ATCTATACCATCGGATGCCTTC-3'). These primers are then added to a PCR reaction at limiting amounts, together with U6 forward (5'-TGGCGGGTGTATTAAACCAC-3') and sgRNA reverse (5'-GGATTTCCTTACGCGAAATACG-3') primers and separate template plasmids containing the U6 promoter (*U6>XX*) and the sgRNA scaffold (*XX>sgRNA F+E*). Plasmids are available from Addgene (https://www.addgene.org/Lionel_Christiaen/). The complementarity between the 5'

ends of the inner primers bridges initially separate U6 and sgRNA scaffold sequences into a single amplicon, and because they are quickly depleted, the entire cassette is preferentially amplified in later cycles by the outer primers (see **Figure 4** and **Supplementary Protocol** for details). Final, unpurified reactions should contain PCR amplicon at ~100 ng/µl, as measured by image analysis after gel electrophoresis (**Supplementary Figure 4**).

TuniCUT and Ci2KO sgRNA Library

The TuniCUT script and supporting databases can be downloaded from github (https://github.com/shashank357/TuniCUT) and run in Python 3.5. Whole-genome sgRNA predictions and oligo designs ("Ci2KO" library) are available for download as a GFF3 file or a BED file available upon request, for local browsing in the IGV browser. To get the sequences of primers for OSO-PCR in IGV, right-click on the desired sgRNA and select "Copy Details to Clipboard". Ci2KO can also be visualized in the genome browser at ANISEED (http://www.aniseed.cnrs.fr/fgb2/gbrowse/ciona_intestinalis/).

Embryo imaging

Fluorescent *in situ* hybridization of *eGFP* or *Foxf* coupled to immunohistochemsitry was carried out as previously described (BEH *et al.* 2007; STOLFI *et al.* 2014). Images were taken on a Leica Microsystems inverted TCS SP8 X confocal microscope or a Leica DM2500 epifluorescence microscope. Mouse monoclonal anti-β-Gal Z3781 (Promega, Madison, WI) was used diluted at 1:500. Goat anti-Mouse IgG (H+L) Secondary Antibody Alexa Fluor 568 conjugate (Life Technologies) was used diluted at 1:500.

Acknowledgments

We are grateful to Florian Razy-Krajka, Farhana Salek, and Aakarsha Pandey for discussions and technical assistance; Tara Rock for advice on next-generation sequencing; Rahul Satija for sequencing the libraries and for his invaluable insights into the sgRNA sequence analysis; Shyam Saladi, Elena K. Perry, and the High Performance Computing team at NYU for their help troubleshooting the bioinformatic analysis. This work was funded by the NIH R01 GM096032 award to L.C., an NYU Biology Masters Research Grant to S.G., while A.S. was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology [NSF-1161835].

Supplementary Table 1. All sgRNAs sorted by decreasing mutagenesis efficacy (Mut%)

sgRNA	Mut%
Htr7-r.2	59.63
Rhod/f.1	58.36
Htr7-r.1	58.11
Ddr.4	46.96
Nk4.2	46.00
Htr7-r.3	43.99
Islet.2	43.93
Lef1.2	43.58
Gata4/5/6.1	43.22
Ddr.2	42.40
Rhod/f.3	40.71
Gata4/5/6.3	40.68
Fgf8/17/18.1	39.86
Tle.b.3	39.10
Foxf.4	38.78
Ets.b.3	38.30
Foxg-r.2	38.14
Bmp2/4.4	37.83
Nk4.3	37.79
Ffg4/6.1	37.51
Ebf.3	37.27
Eph.a.3	37.11
Htr7-r.4	35.87
Hand.1	35.52
Gata4/5/6.4	35.41
Eph.a.2	34.97
Mrf.2	32.85
Ddr.5	32.81
Mrf.3	32.61
Fgf8/17/18.2	31.21
Hand.2	30.98
Ffgr.2	30.39
Lef1.1	30.08
Fzd5/8.3	29.59
Foxf.1	29.05
Foxf.3	28.70
Tll.1	27.47
Tle.b.1	27.03
Tle.b.2	26.32
Bmp2/4.2	24.91
Ebf.1	24.59
Ffgr.4	24.29
Ffgr.1	24.15
Fzd5/8.2	23.83
Neurog.1	23.57
Eph.a.1	23.17
Hand.4	23.07
Rhod/f.2	22.40
Hand-r.2	21.96
Ffg4/6.3	20.65
Ffg4/6.2	19.59
Fzd5/8.4	18.52
Foxf.2	17.61
Mrf.1	15.49
Fzd5/8.1	15.09
Nk4.1	14.13
Nk4.4	13.81
Gata4/5/6.2	13.74
Bmp2/4.3	12.86
Hand-r.1	12.27
Tll.4	10.70
Ffg4/6.4	10.24
Foxg-r.3	10.16
Ets.b.1	9.31
Hand-r.3	8.80
Tle.b.4	6.96
Ets.b.2	5.72
Mrf.4	5.25
Hand.3	5.22
Ebf.2	4.97
Ddr.1	3.84
Rhod/f.4	3.52
Ddr.3	3.33
Tll.2	2.37
Ffgr.3	2.33
Tll.3	2.22
Foxg-r.1	1.87
Hand-r.4	0.75
Foxg-r.4	0.34
Ets.b.4	0.12
Ebf.4	0.05
Neurog.2	0.00
Bmp2/4.1	0.00

Supplementary Table 2. List of Primers (5' -> 3') used to amplify loci for deep sequencing.

Forward Primer (5' -> 3')	Reverse Primer (5' -> 3')	gRNAs assayed
GGACGCGATAGTCAACGAAT	GGCTCGCAATATCTTCATGC	Bmp2/4.1
CGACTACTACGAGCCCGAAC	GCCGAGAAAACAAAATTTTCAA	Bmp2/4.2
TCATTAGCGCGATGGATGT	ATCGTCCGGAGGCATTTT	Bmp2/4.3, Bmp2/4.4
AAACGTCAACGCTGAAAAGC	GTATTGAAGCCGCCAAGAAA	Ddr.1
TGTAGTTGCTCCTCCACACG	CACACCATCAGTCAACTGTCC	Ddr.2
CACATAATATGTGCCAACTCG	CCCTAATACAACACGAACCTC	Ddr.3, Ddr.4
TGACGTAGCCACACTTATAGG	AACTGGGTTACACATGTCAC	Ddr.5
CCGCCAAACAACCTTAGAAA	GTGTCCCATCTTACCCCACA	Ebf.1
AAGTAATATCCAAATGGCAACAATG	AACTTAACACAAATATCGGTCAGATT	Ebf.2
CATTGACCATCCTGAACGAA	ACGGGAAAACGAAATGAACA	Ebf.3, Ebf.4
TTCCAAGTTTGTCACATTCGAT	TGAAAACGTGTCTCATCTTACCC	Eph.a.1
GCGAGCACGAGGTGTATGTA	TGTATAATCCTCTGCTTCTGATCG	Eph.a.2
TGCATATCCTCCACACAGGA	CAGTCTCGCAATAACAACAAGC	Eph.a.3
GCATCCCATCTTTGGATGAA	TCCTCGTGTCATTTCCACAT	Ets.b.1, Ets.b.2
AGCTCTGCCTAAGTCTATTACCG	CATCTTAAACTCCCAGCCATC	Ets.b.3
GTGGTCCAATCCAACTGTGG	TCGATTCTAAACTCCAGCAATG	Ets.b.4
GCTGTGGATTACTATAAATAGCACTGT	CTGTTTTCCACAGCAGCAGA	Fgf4/6.1
GAAGATGGTATAAGCACTCAGCAA	TACACCGTAGGACGAGCAAG	Fgf4/6.2, Fgf4/6.3
GGAAATGAGAAGCTTCGAAAGA	AATTCCGATGGAAGGAGGTT	Fgf4/6.4
TTTCTCACAATGATCGGTATACAAAC	TCCCAGAAAAGACCTCGTTG	Fgf8/17/18.1
GAATCCCCGATCCCCATA	ATATAACTGTAGCCTTGAGACTC	Fgf8/17/18.2
GAATGAAACCAAACCCCTCA	AGGAACGAATAAACAATGCTGA	Fgfr.1
CCCCAACGTATCCCATCTTA	GACGATCCGTAGTTGTAGATGC	Fgfr.2
TCTTGGGTAATGGCCAACTC	CCCCCAGTCTTCCATTCTTC	Fgfr.3
TTGGAGTCGGTGATAATGTCC	ACATGCTACCTTTGTGCTAGTGA	Fgfr.4
ACCACCGACCCAACTTAATG	TGGACAGAAGTGTTTGTTCCAG	Foxf.1, Foxf.2
GAATCCCCGATCCCCATA	ATCAATATGGCGGAAAACGA	Foxf.3
TGGAAATGGGAAAGGCTTAC	TTAAAGCGCTGCTCTCTCG	Foxf.4
CCGAACAAAACAGTCGTTTC	TAATTAGATGCCCGGGACTG	Foxg-r.1
CACACACATACCCCGCATTA	TCATGATAAGCGGGAAACAA	Foxg-r.2
GGAGAATCGACCAACGTGAG	CGTTGACATGTTTGTACCTTCG	Foxg-r.3, Foxg-r.4
GCGAACCTATCCAAGTACCG	CCATTAAACGGTTTGTCTAAGAATG	Fzd5/8.1
TATGTCGAAGTTCGCGGTTG	TTCCTAAACCTATGATTTAACTGACCT	Fzd5/8.2, Fzd5/8.3
GGTATTCCATTTTACCCCACA	TGCTTTTTAACGCTGGGATA	Fzd5/8.4
CGTTCCAGATTCATCCCATC	CCAAGTTTTGCTGTGTGACG	Gata4/5/6.1
CAGCATGACTAACCTTGTATTCCA	CACGCTAACGCAAAGTAGCC	Gata4/5/6.2, Gata4/5/6.3
CCCGCAAAGATACGACTACC	ATGCTTCCGGCTGTAGTGTT	Gata4/5/6.4
GTAACGTCCGGCAGAAAGAA	CTTTCTACGTTCGCGGGATT	Hand.1, Hand.2
TGGCGTCTTCCTATATTGCTT	TGGTTTAAAAGCGCTTCATT	Hand.3

GACGCTTCTGCAACTGCTTA	TCAGTTTATCACCGCAGCAG	Hand.4
AACGTGGCGGATTCACAAT	CGACGACTCATACGGCTCTT	Hand-r.1, Hand-r.2
CTCCCTTATACCCCACACAGTT	CCAGCTAGCGGTGGTGTAGT	Hand-r.3
TCCGAGTCACCCGTATTATCA	ATTCTATCGCGCAGATCGTC	Hand-r.4
GCGGTTGCGACACTAAAGAT	CCCACCCTCTATTCCGAAAC	Htr7-r.1, Htr7-r.2
TAACGCTGTGTGAACGCACT	AAGGTGTCCCATCTTTCCCTA	Htr7-r.3
CGACTCCCATAGACCGTTG	ACAAAGATGGCGGTCGATAC	Htr7-r.4
CGAATCATTCGCCATTTTCG	TCAACCCTGGCCTGATTTC	Isl.1
TTACTTGCGCGTGTTTTCAG	AGCATGCCACTCAAGGTTG	Isl.2
GCGCGTACTGGATTACTTGG	TCTTCTTCCTCGACTAAATCACC	Lef1.1
AAAGGGTAGGCGAAGAGGAC	ACACCAACTAAAACGCTAATATGTAA	Lef1.2
GGAGCTCGACCTCTCTTCAA	ATATCGCTCTTTTCCGCATC	Mrf.1, Mrf.2
CAAACGGCCATCAATGTTTAG	TGTTGGTCGCCATACAACATA	Mrf.3, Mrf.4
CTTCTGTTGTTCCGACAAGC	AGTCATACCAAAATGAGCCAC	Neurog.1
TCCAAAAAGACACAGTCAACAG	ACCTTAATTCTTCTAGTGCGTC	Neurog.2
GTGGTCGTCGATTTTCCATC	GCTAATTCCCATTCCGCTTA	Nk4.1, Nk4.2
CTGTTCTCCCAGGCACAAGT	CGCGAACTAAAACAGGAACTG	Nk4.3, Nk4.4
AACTTGTAAAGCCAAAACTATACACAA	TTAAACGTAAACTTACTTCAGGGAAT	Rhod/f.1
TAGGGCGAGCAGGATAATTG	CACTTCCATGCTAACCGAAA	Rhod/f.2, Rhod/f.3
CAAATGTTTACTTGTACCAGGTCAG	ATCCGAGTTCACCTGAGACC	Rhod/f.4
TCAACGTCCTTTGTTTGTGG	GGTGCCAGCTTCGACCTAT	Tle.b.1
TGGAACATTATGAGATTAGGACTTCA	CGTTTAACGGATGAGGAGGT	Tle.b.2, Tle.b.3, Tle.b.4
TGGCGACATTGCTTTAGATG	ACTTTCTGTTCCGCCGTCT	Tll.1
CCCCATTCTTATCAGTTGCTTT	CGTCGTCTGGATACCTCTCG	Tll.2
ACCACCGGCGATACATTAAG	AACTGATGTTTAACTAATACCTTTTCG	Tll.3
TTGGTTAGGGATTGTCGTTTT	ATGGGGCCATTTTCTCTCTT	Tll.4

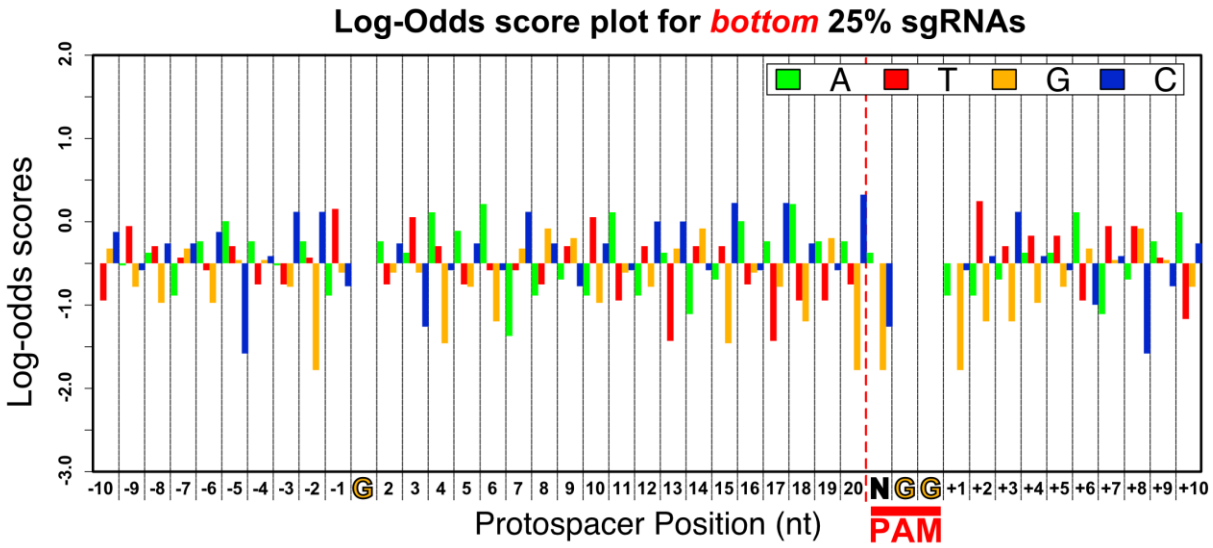
Supplementary Table 3. Contribution of dual nucleotide features to mutagenesis efficacy.
Co-occurrence of specific nucleotides at two given positions (X and Y, e.g. -4C = "C" at position -4. +2G = "G" at position +2) of the 43-mer target site (10 nt 5' flanking + Protospacer + PAM + 10 nt 3' flanking, see **Figure 2** for the exact annotation of positions) and their overall contribution to mutagenesis efficacy (Mut%) of the associated sgRNA, as determined by lasso regression. Contribution is given as percentage points of total mutagenesis efficacy rate. Asterisks indicate those dual nucleotide features that are, in effect, single nucleotide features since one of the positions does not vary in our data set (e.g. all sgRNAs have "G" at position 1, and all PAMs are of the sequence "NGG").

Nucleotide X	Nucleotide Y	Contribution to Mut%
-7C	12G	13.17
-3A	16T	7.74
-10C	+5T	7.23
12A	+8C	6.95
13T	+7C	6.69
-4A	11G	5.76
-1T	+3A	5.19
-1T	+4G	4.16
-10C	+2A	3.85
-8A	5T	3.74
+1T	+3T	3.58
-10T	-5G	3.43
3A	12T	3.14
-8A	+3T	2.75
3A	+2A	2.63
20G	+9C	2.61
3A	17G	2.49
-8A	9G	2.26
-8T	6G	1.79
6A	+8T	1.74
-9T	1G	1.51
1G	17G	1.37*
-9T	PAM[2]G	1.34*
5T	15T	1.26
-9A	10T	1.15
10G	11G	1.07
-10T	+8T	0.82
1G	18G	0.77*
12A	+3T	0.67

10A	+9T	0.57
8T	19G	0.54
13G	+3A	0.44
1G	20G	0.37*
-6T	17G	0.29
1G	8A	0.21*
7G	20G	0.12
-7C	16C	0.11
19G	+3T	0.1
1G	6G	0.02*
6A	PAM[2]G	0.01*
8T	PAM[2]G	0.01*
17A	PAM[2]G	0.01*
6A	10C	-0.02
-3G	PAM[1]T	-0.04
+1T	+4T	-0.04
1G	17C	-0.07*
2G	+3C	-0.1
-7G	14G	-0.13
3T	20C	-0.44
-7G	+2T	-0.79
17C	PAM[1]T	-0.86
-5A	11A	-1.03
1G	8C	-1.03*
5A	8G	-1.07
-2A	12C	-1.19
-8G	+10C	-1.79
-9T	+7G	-1.99
-10C	+2T	-2.16
PAM[1]T	+6A	-2.21
-6T	17A	-2.31
6A	+8G	-4.84
-8T	+10A	-5.14
+1C	+6A	-5.14
PAM[1]T	+7T	-6.49
-2A	3T	-7.45

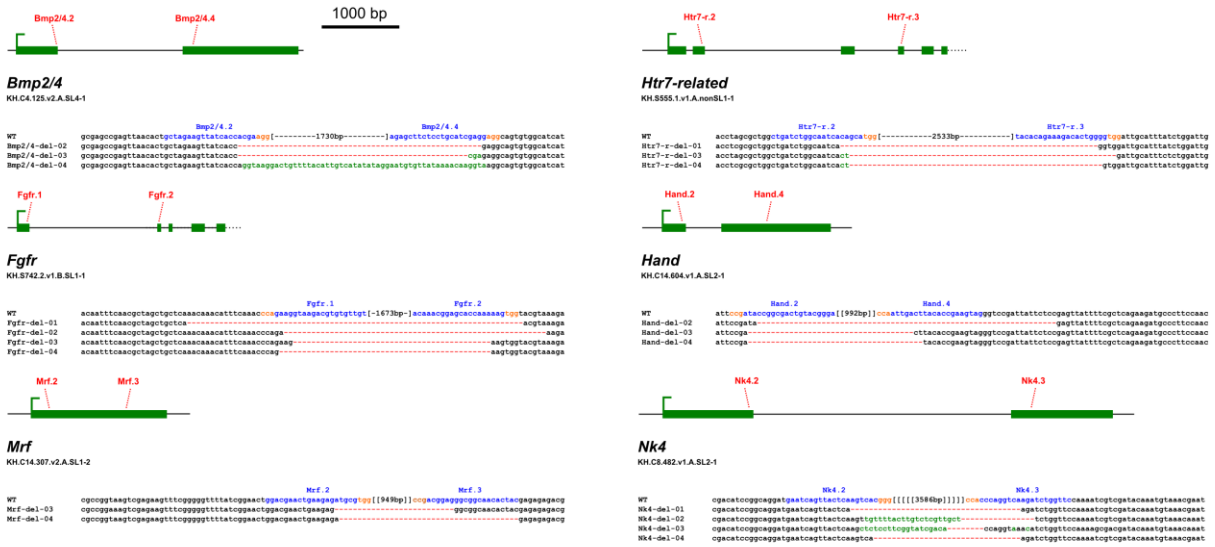
Supplementary Table 4. Contributions of single nucleotide features to mutagenesis efficacy

Nucleotide X	Contribution to Mut%
20G	5.26
18G	3.60
17G	0.77
-9A	0.72
6G	0.34
8A	0.05
17C	-0.44
8C	-1.00
+3C	-4.65



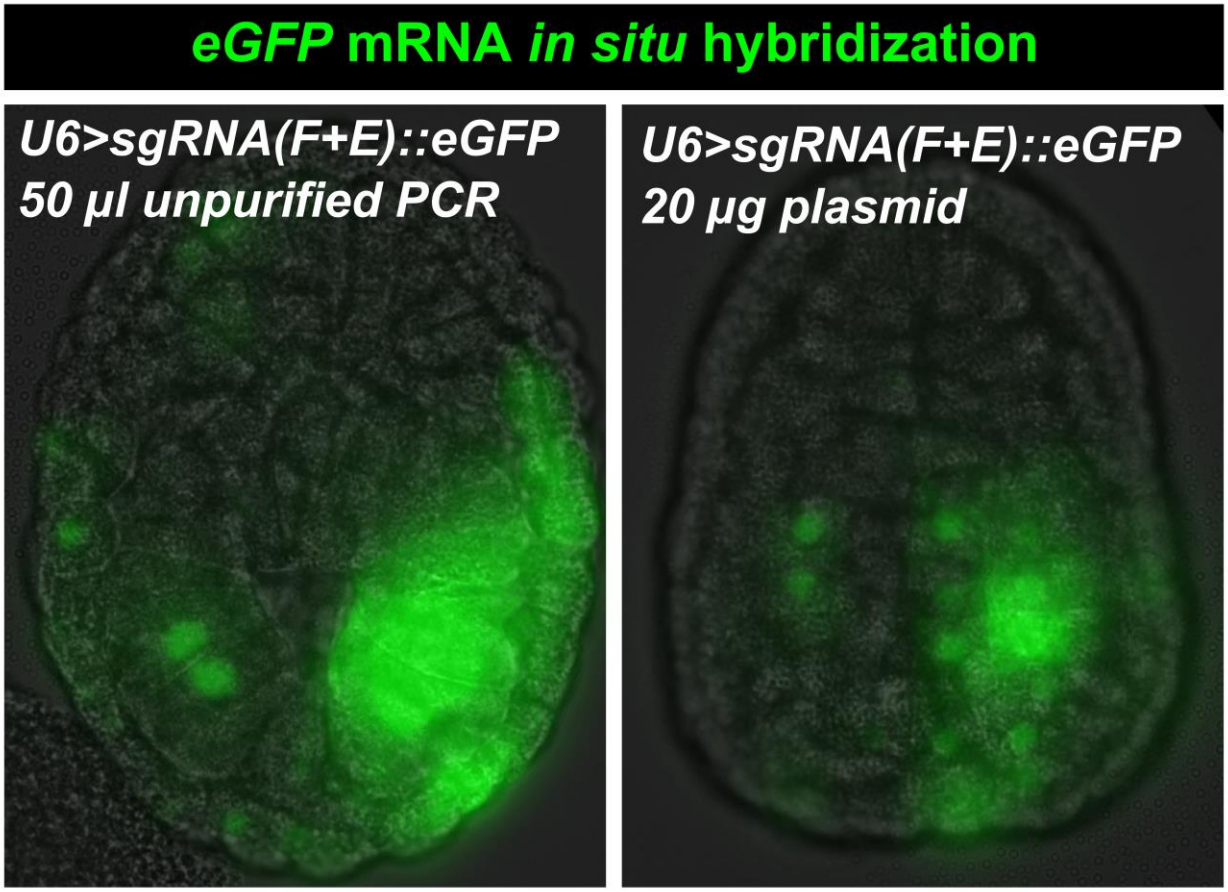
Supplementary Figure 1. Log-odds score plot for bottom 25% sgRNAs

Log-odds scores depicting the frequency of occurrence for nucleotides in the bottom 25% (least effective) sgRNAs, at all positions of the protospacer, PAM, and flanking regions. Position "1" of the protospacer has been omitted from the analysis, due to this always being "G" for PolIII-dependent transcription of U6-promoter-based vectors. Likewise, the "GG" of the PAM has also been omitted, as this sequence is invariant in all targeted sites.



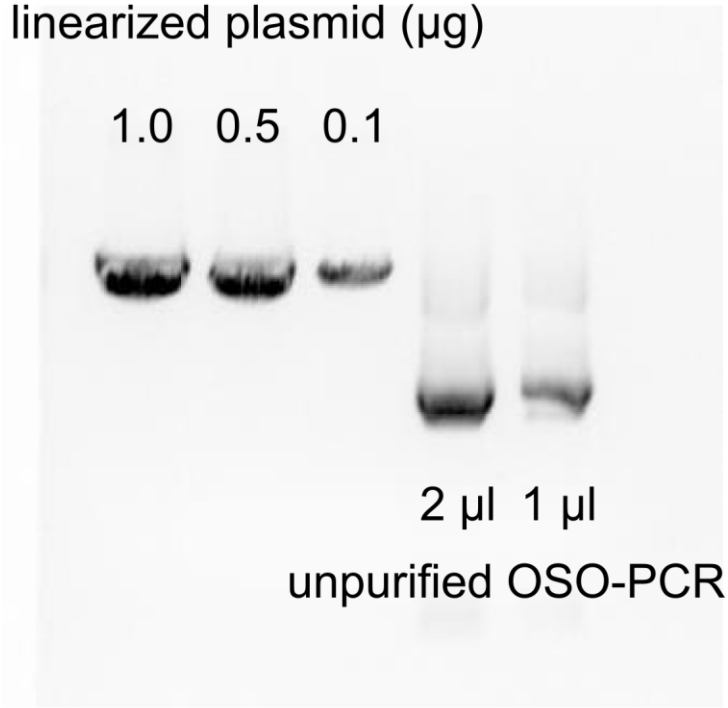
Supplementary Figure 2. Other examples of large deletions obtained by combinatorial action of two sgRNAs

Sequence alignments of clones for each locus, amplified from embryos in which two sgRNAs were used for CRISPR/Cas9-induced site mutagenesis.



Supplementary Figure 3. *In vivo* gene expression from electroporated unpurified PCR products

In situ hybridization of *eGFP* in late gastrula/early neural stage embryos electroporated with either *U6>sgRNA(F+E)::eGFP* plasmid (20 µg) or unpurified PCR product (50 µl, ~5 µg DNA).



Supplementary Figure 4. Quantification of OSO-PCR products

Image of gel electrophoresis of varying amounts of linearized plasmid and unpurified OSO-PCR products. Pixel intensity analysis in ImageJ was performed as previously described (STOLFI *et al.* 2014), and indicated that the sgRNA expression cassette in unpurified OSO-PCR reactions are at a concentration of approximately 100 ng/µl.

REFERENCES

Abdul-Wajid, S., H. Morales-Diaz, Stephanie M. Khairallah and William C. Smith, 2015 T-type Calcium Channel Regulation of Neural Tube Closure and EphrinA/EPHA Expression. Cell Reports 13**:** 829-839.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval *et al.*, 2007 CRISPR provides acquired resistance against viruses in prokaryotes. Science 315**:** 1709-1712.

Beh, J., W. Shi, M. Levine, B. Davidson and L. Christiaen, 2007 FoxF is essential for FGF-induced migration of heart progenitor cells in the ascidian *Ciona intestinalis*. Development 134**:** 3297-3305.

Brozovic, M., C. Martin, C. Dantec, D. Dauga, M. Mendez *et al.*, 2015 ANISEED 2015: a digital framework for the comparative developmental biology of ascidians. Nucleic acids research**:** gkv966.

Brunetti, R., C. Gissi, R. Pennati, F. Caicci, F. Gasparini *et al.*, 2015 Morphological evidence that the molecularly determined Ciona intestinalis type A and type B are different species: Ciona robusta and Ciona intestinalis. Journal of Zoological Systematics and Evolutionary Research 53**:** 186-193.

Chari, R., P. Mali, M. Moosburner and G. M. Church, 2015 Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nature methods 12**:** 823-826.

Chen, B., L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang *et al.*, 2013 Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell 155**:** 1479-1491.

Cho, S. W., S. Kim, Y. Kim, J. Kweon, H. S. Kim *et al.*, 2014 Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome research 24**:** 132-141.

Christiaen, L., E. Wagner, W. Shi and M. Levine, 2009 The sea squirt *Ciona intestinalis*. Cold Spring Harbor protocols 2009**:** pdb. emo138.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto *et al.*, 2013 Multiplex genome engineering using CRISPR/Cas systems. Science 339**:** 819-823.

Crooks, G. E., G. Hon, J.-M. Chandonia and S. E. Brenner, 2004 WebLogo: a sequence logo generator. Genome research 14**:** 1188-1190.

Dickinson, D. J., J. D. Ward, D. J. Reiner and B. Goldstein, 2013 Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. Nature methods 10**:** 1028-1034.

Diogo, R., R. G. Kelly, L. Christiaen, M. Levine, J. M. Ziermann *et al.*, 2015 A new heart for a new head in vertebrate cardiopharyngeal evolution. Nature 520**:** 466-473.

Doench, J. G., E. Hartenian, D. B. Graham, Z. Tothova, M. Hegde *et al.*, 2014 Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature biotechnology.

Fu, Y., J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon *et al.*, 2013 High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature biotechnology 31**:** 822-826.

Fu, Y., J. D. Sander, D. Reyon, V. M. Cascio and J. K. Joung, 2014 Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nature biotechnology 32**:** 279-284.

Gagnon, J. A., E. Valen, S. B. Thyme, P. Huang, L. Ahkmetova *et al.*, 2014 Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs.

Gantz, V. M., and E. Bier, 2015 The mutagenic chain reaction: A method for converting heterozygous to homozygous mutations. Science 348**:** 442-444.

Hanley, J. A., and B. J. McNeil, 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143**:** 29-36.

Hirano, T., and H. Nishida, 1997 Developmental Fates of Larval Tissues after Metamorphosis in Ascidian *Halocynthia roretzi* I. Origin of Mesodermal Tissues of the Juvenile. Developmental biology 192**:** 199-210.

Hoshino, Z. i., and T. Tokioka, 1967 An unusually robust Ciona from the northeastern coast of Honsyu Island, Japan.

Hotta, K., K. Mitsuhara, H. Takahashi, K. Inaba, K. Oka *et al.*, 2007 A web-based interactive developmental table for the ascidian *Ciona intestinalis*, including 3D real-image embryo reconstructions: I. From fertilized egg to hatching larva. Developmental Dynamics 236**:** 1790-1805.

Hsu, P. D., D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann *et al.*, 2013 DNA targeting specificity of RNA-guided Cas9 nucleases. Nature biotechnology 31**:** 827-832.

Hwang, W. Y., Y. Fu, D. Reyon, M. L. Maeder, S. Q. Tsai *et al.*, 2013 Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature biotechnology 31**:** 227-229.

Imai, K. S., K. Hino, K. Yagi, N. Satoh and Y. Satou, 2004 Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Development 131**:** 4047-4058.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna *et al.*, 2012 A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337**:** 816-821.

Jinek, M., A. East, A. Cheng, S. Lin, E. Ma *et al.*, 2013 RNA-programmed genome editing in human cells. eLife 2.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nature methods 9**:** 357-359.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell *et al.*, 2013 RNA-guided human genome engineering via Cas9. Science 339**:** 823-826.

Moreno-Mateos, M. A., C. E. Vejnar, J.-D. Beaudoin, J. P. Fernandez, E. K. Mis *et al.*, 2015 CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nature methods 12**:** 982-988.

Naito, Y., K. Hino, H. Bono and K. Ui-Tei, 2015 CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics 31**:** 1120-1123.

Ng, A. Y., 2004 Feature selection, L 1 vs. L 2 regularization, and rotational invariance, pp. 78 in *Proceedings of the twenty-first international conference on Machine learning*. ACM.

Nishida, H., 1987 Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme: III. Up to the tissue restricted stage. Developmental biology 121**:** 526-541.

Nishiyama, A., and S. Fujiwara, 2008 RNA interference by expressing short hairpin RNA in the Ciona intestinalis embryo. Development, growth & differentiation 50**:** 521-529.

Pasini, A., A. Amiel, U. Rothbächer, A. Roure, P. Lemaire *et al.*, 2006 Formation of the ascidian epidermal sensory neurons: insights into the origin of the chordate peripheral nervous system. PLoS Biology 4**:** e225.

Pattanayak, V., S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna *et al.*, 2013 High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nature biotechnology 31**:** 839-843.

Razy-Krajka, F., K. Lam, W. Wang, A. Stolfi, M. Joly *et al.*, 2014 Collier/OLF/EBF-Dependent Transcriptional Dynamics Control Pharyngeal Muscle Specification from Primed Cardiopharyngeal Progenitors. Developmental cell 29**:** 263-276.

Ren, X., Z. Yang, J. Xu, J. Sun, D. Mao *et al.*, 2014 Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. Cell reports 9**:** 1151-1162.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nature biotechnology 29**:** 24-26.

Rothbächer, U., V. Bertrand, C. Lamy and P. Lemaire, 2007 A combinatorial code of maternal GATA, Ets and β-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. Development 134**:** 4023-4032.

Sasaki, H., K. Yoshida, A. Hozumi and Y. Sasakura, 2014 CRISPR/Cas9 - mediated gene knockout in the ascidian Ciona intestinalis. Development, growth & differentiation 56**:** 499-510.

Satoh, N., 2013 *Developmental genomics of ascidians*. John Wiley & Sons.

Satou, Y., T. Kawashima, E. Shoguchi, A. Nakayama and N. Satoh, 2005 An integrated database of the ascidian, *Ciona intestinalis*: towards functional genomics. Zoological science 22**:** 837-843.

Satou, Y., K. Mineta, M. Ogasawara, Y. Sasakura, E. Shoguchi *et al.*, 2008 Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. Genome Biology 9**:** R152.

Schneider, T. D., and R. M. Stephens, 1990 Sequence logos: a new way to display consensus sequences. Nucleic acids research 18**:** 6097-6100.

Shalem, O., N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott *et al.*, 2014 Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343**:** 84-87.

Stolfi, A., T. B. Gainous, J. J. Young, A. Mori, M. Levine *et al.*, 2010 Early chordate origins of the vertebrate second heart field. Science 329**:** 565.

Stolfi, A., S. Gandhi, F. Salek and L. Christiaen, 2014 Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. Development 141**:** 4115-4120.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological)**:** 267-288.

Urban, A., S. Neukirchen and K.-E. Jaeger, 1997 A rapid and efficient method for site-directed mutagenesis using one-step overlap extension PCR. Nucleic acids research 25**:** 2227-2228.

Wang, H., H. Yang, C. S. Shivalila, M. M. Dawlaty, A. W. Cheng *et al.*, 2013a One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153**:** 910-918.

Wang, W., F. Razy-Krajka, E. Siu, A. Ketcham and L. Christiaen, 2013b NK4 antagonizes Tbx1/10 to promote cardiac versus pharyngeal muscle fate in the ascidian second heart field. PLoS Biology 11**:** e1001725.

Wu, X., D. A. Scott, A. J. Kriz, A. C. Chiu, P. D. Hsu *et al.*, 2014 Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature biotechnology 32**:** 670-676.

Zeller, R. W., M. J. Virata and A. C. Cone, 2006 Predictable mosaic transgene expression in ascidian embryos produced with a simple electroporation device. Developmental Dynamics 235**:** 1921-1932.

Supplemental Protocol

ONE-STEP OVERLAP PCR (OSO-PCR) TO MAKE READY-TO-ELECTROPORATE SINGLE GUIDE RNA (sgRNA) EXPRESSION CASSETTES – *updated 02/26/2016*

Companion manuscript:

Rational design and whole-genome predictions of single guide RNAs for efficient CRISPR/Cas9-mediated genome editing in *Ciona*

Shashank Gandhi, Lionel Christiaen and Alberto Stolfi

Primers for OSO-PCR ready to be ordered can be copied from each sgRNA entry in the Ci2KO library (in IGV, right-click on the desired sgRNA and select "Copy Details to Clipboard"). Use CRISPRdirect (http://crispr.dbcls.jp/) to cross-references with the JGI v.2 *Ciona intestinalis* genome for potential off-targets. Avoid any sgRNA that has a "12mer+PAM" off-target in the genome. Check for polymorphisms using the Kyoto University Ghost Database genome browser (http://ghost.zool.kyoto-u.ac.jp/cgi-bin/gb2/gbrowse/kh/). To design OSO-PCR primers *de novo,* follow the instructions:

1- Select your target, as identified by online tools such as CRISPRdirect (see above).

target PAM

...TCAACCCAACTGAGGGTTGGACAACAGGTGGAGCAACAGT...

2- A target (the protospacer) is given as N(20). If the target sequence contains too many T's (three or more T's clustered together tend to terminate transcription), or if it spans many known naturally-occurring polymorphisms, or has a high number of potential off-targets, discard it.

3- For transcription initiation from U6 promoter, replace the first base of the target with a "G", to give a G+(N)19 sequence.

GCTGAGGGTTGGACAACAGG

4- Append "GTTTAAGAGCTATGCTGGAAACAG" to the 3' end of the sequence. This entire sequence is now the forward primer used to PCR the sgRNA scaffold part of the cassette ("OSO forward" primer)

GCTGAGGGTTGGACAACAGGGTTTAAGAGCTATGCTGGAAACAG

5- Copy reverse complement of G+N(19), append "ATCTATACCATCGGATGCCTTC" to the 3' end of this. This is now the reverse primer to PCR the U6 promoter part of the cassette ("OSO reverse" primer)

CCTGTTGTCCAACCCTCAGCATCTATACCATCGGATGCCTTC

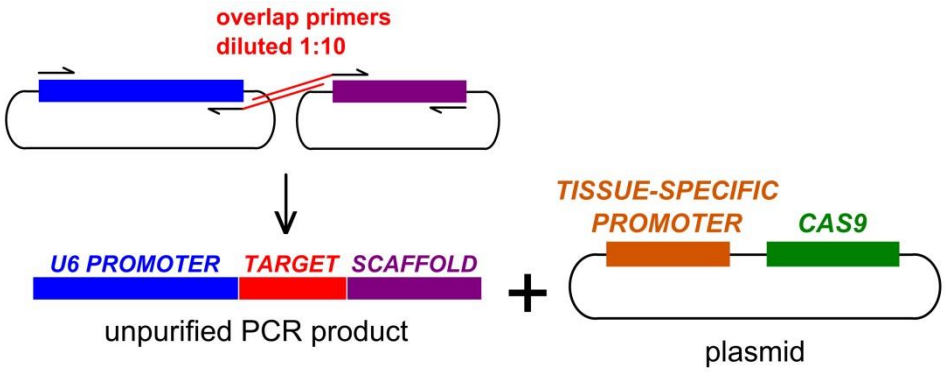
6- Set up a PCR reaction using the following components in the exact amounts described. The amounts/concentrations/proportions are critical for the one-step overlap reaction to occur seamlessly. Also, it is very important to eliminate all sources of contamination, otherwise you may re-amplify sgRNAs already in heavy use in the lab. Template plasmids are available from Addgene (https://www.addgene.org/Lionel_Christiaen/):

For 50 ul reaction:

1.5 ul 10mM dNTPs
1 ul 50mM MgSO4
10 ul 10X Pfx Buffer
1 ul U6>XX plasmid at 15 ng/ul
1 ul X>sgRNA(F+E) plasmid at 15 ng/ul
1.5 ul 20 uM U6 forward primer (5'- TGGCGGGTGTATTAAACCAC -3')
1.5 ul 20 uM sgRNA reverse primer (5'- GGATTTCCTTACGCGAAATACG -3')
1 ul **2 uM** OSO forward primer (designed in step 4, or obtained from TuniCUT/Ci2KO)
1 ul **2 uM** OSO reverse primer (designed in step 5, or obtained from TuniCUT/Ci2KO)
30 ul H2O
0.5 ul Pfx platinum

PCR program:

```
94° - 3'
94° - 30"   |
50° - 30"   | X 30
68° - 3'    |
68° - 5'
```



The 1:10 dilution of your custom overlap target-specific primers will force the "fusion" of the entire cassette later in the reaction, when these primers are depleted from the solution through incorporation into the PCR products.

7- Run 2 ul of the PCR reaction on a gel. There should be a strong band at ~1.2 kbp. If the band is only 1 kbp, the fusion did not occur. The success rate in our hands is ~94%. If possible, run alongside positive control (PCR on verified sgRNA plasmid template using same primers).

OSO-PCR products can be electroporated as is, un-purified. 25 ul appears to be sufficient to recapitulate effects of sgRNAs delivered by traditional plasmid electroporation, but this volume can be adjusted accordingly. If you need to clone the cassette into a plasmid, you can use the product as template for additional PCRs using the outer primers with added overhangs for restriction enzyme or Clontech In-Fusion cloning.