Evaluation and rational design of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in Ciona

Shashank Gandhi, Maximilian Haeussler, Florian Razy-Krajka, Lionel Christiaen, Alberto Stolfi

Department of Biology, New York University, New York, USA
Santa Cruz Genomics Institute, University of California, MS CBSE, Santa Cruz, USA

ARTICLE INFO

Keywords:
- sgRNA
- Cardiopharyngeal mesoderm
- Tunicate

ABSTRACT

The CRISPR/Cas9 system has emerged as an important tool for various genome engineering applications. A current obstacle to high-throughput applications of CRISPR/Cas9 is the imprecise prediction of highly active single guide RNAs (sgRNAs). We previously implemented the CRISPR/Cas9 system to induce tissue-specific mutations in the tunicate Ciona. In the present study, we designed and tested 83 single guide RNA (sgRNA) vectors targeting 23 genes expressed in the cardiopharyngeal progenitors and surrounding tissues of Ciona embryo. Using high-throughput sequencing of mutagenized alleles, we identified guide sequences that correlate with sgRNA mutagenesis activity and used this information for the rational design of all possible sgRNAs targeting the Ciona transcriptome. We also describe a one-step cloning-free protocol for the assembly of sgRNA expression cassettes. These cassettes can be directly electroporated as unpurified PCR products into Ciona embryos for sgRNA expression in vivo, resulting in high frequency of CRISPR/Cas9-mediated mutagenesis in somatic cells of electroporated embryos. We found a strong correlation between the frequency of an Ebf loss-of-function phenotype and the mutagenesis efficiency of individual Ebf-targeting sgRNAs tested using this method. We anticipate that our approach can be scaled up to systematically design and deliver highly efficient sgRNAs for the tissue-specific investigation of gene functions in Ciona.

1. 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 direct the Cas9 nuclease of Streptococcus pyogenes to a specific target DNA sequence. (Jinek et al., 2012, 2013; Cong et al., 2013; Mali et al., 2013). Although initial Cas9 binding requires a Protospacer Adjacent Motif (PAM) sequence, most commonly “NGG”, the high specificity of this system is accounted for by Watson–Crick base pairing between the 5′ end of the sgRNA and a 17–20 bp “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; H. Wang et al., 2013; Koike-Yusa et al., 2014; Paix et al., 2014; Shalem et al., 2014; Wang et al. 2014; Gantz and Bier, 2015; Sanjana et al., 2016).

The tunicate Ciona is a model organism for the study of chordate-specific developmental processes (Satoh, 2013). The CRISPR/Cas9 system was adapted to induce site-specific DSBs in the Ciona genome (Sasaki et al., 2014; Stolfi et al., 2014). Using electroporation to transiently transflect 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 Cas9 expression in specific cell lineages or tissue types, we were thus able to control the disruption of Ebf function with spatiotemporal precision. Following
this proof-of-principle, tissue-specific CRISPR/Cas9 has rapidly propagated as a simple yet powerful tool to elucidate gene function in the Ciona embryo (Abdul-Wajid et al., 2015; Cota and Davidson, 2015; Segade et al., 2016; Tolkin and Christiaen, 2016).

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 in Ciona embryos. However, not all sgRNAs drive robust CRISPR/Cas9-induced mutagenesis, and few guidelines exist for the rational design of highly active sgRNAs, which are critical for rapid gene disruption in F0. This variability and unpredictable efficacy demands experimental validation of each sgRNA tested. Individual studies have revealed certain nucleotide sequence features that correlate with high sgRNA expression and/or activity in CRISPR/Cas9-mediated DNA cleavage (Doench et al., 2014, 2016; Gagnon et al., 2014; Ren et al., 2014; Wang et al., 2014; Chari et al., 2015; Fusi et al., 2015; Housden et al., 2015; Moreno-Mateos et al., 2015; Wong et al., 2015; Xu et al., 2015). These studies have been performed in different organisms using a variety of sgRNA and Cas9 delivery methods and show varying ability to predict sgRNA activities across platforms (Haessler et al., 2016).

Given the uncertainty regarding how sgRNA design principles gleaned from experiments in other species might be applicable to 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 practical pipeline for optimal design and efficient assembly of sgRNA expression constructs for use in Ciona.

2. Results

2.1. High-Throughput sequencing to estimate sgRNA-specific mutagenesis rates

Previous studies using CRISPR/Cas9-based mutagenesis in Ciona revealed that different sgRNAs have varying ability to induce mutations (Sasaki et al., 2014; Stolfi et al., 2014). In order to test a larger number of sgRNAs and identify 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 cardiopharyngeal mesoderm of vertebrates (W. Wang et al., 2013; Diogo et al., 2015; Kaplan et al., 2015).

We followed a high-throughput-sequencing-based approach to quantify the efficacy of each sgRNA, i.e. its ability to cause CRISPR/Cas9-induced mutations in targeted sequences in the genome (Fig. 1a–c). The 83 sgRNA plasmids were co-electroporated with Ef1α > nls::Cas9::nls plasmid. The ubiquitous Ef1α promoter is active in all cell lineages of the embryo and has been used to express a variety of site-specific nucleases for targeted somatic knockouts in Ciona (Sasakura et al., 2010; Kawai et al., 2012; Sasaki et al., 2014; Stolfi et al., 2014; Treen et al., 2014). Each individual sgRNA + Cas9 vector combination was electroporated into pooled Ciona zygotes, which were then grown at 18 °C for 16 hours post-fertilization (hp; embryonic stage 25) (Hotta et al., 2007). Targeted sequences were individually PCR-amplified from each pool of embryos. Each target was also amplified from “negative control” embryos grown in parallel and electroporated with Ef1α > nls::Cas9::nls and “Ub > Negative Control” sgRNA vector. Agarose gel-selected and purified amplicons (varying from 108 to 350 bp in length) were pooled in a series of barcoded “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 3899 reads each (Supplementary Table 1). 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 and PAM. The ratio of [indels]/[total reads] represents our estimation of the mutagenesis efficacy of the sgRNA (Fig. 1b–d, Supplementary Table 1). For each mutation, an unknown number of cell divisions occur between the moment Cas9 induces a DSB and the time of sample collection and genomic DNA extraction. This prevents us from quantifying the true mutagenesis rates. However, we can surmise that this applies comparably to all sgRNAs, such that our values still represent an accurate ranking of mutagenesis efficacy.

For simplicity, we did not count single nucleotide point mutations, 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 efficiencies varying from 0.05% (Ef4.4) to 59.63% (Htr7-r.z2). Although each sgRNA was tested only once, we did not observe any evidence of electroporation variability or batch effects that may have confounded our estimates (Supplementary Figure 1).

This conservative approach most likely underestimates the actual mutagenesis rates. First, we excluded point mutations 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 of 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

### Table 1

<table>
<thead>
<tr>
<th>#</th>
<th>Gene Symbol</th>
<th>Aliases</th>
<th>2012 KyotoHoya ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bmp2/4</td>
<td>Bone morphogenetic protein 2/4</td>
<td>KH.C1.125</td>
</tr>
<tr>
<td>2</td>
<td>Ddr</td>
<td>Discladin Domain Receptor</td>
<td>KH.C9.371</td>
</tr>
<tr>
<td>3</td>
<td>Ef4</td>
<td>Collar/09/EBF; COE</td>
<td>KH.C12.9</td>
</tr>
<tr>
<td>4</td>
<td>Eph-a</td>
<td>Ephrin type-A receptor; Eph1</td>
<td>KH.C1.404</td>
</tr>
<tr>
<td>5</td>
<td>Ets.b</td>
<td>Ets/Pointed2</td>
<td>KH.C1.10</td>
</tr>
<tr>
<td>6</td>
<td>Fgk6/6</td>
<td>Fibroblast growth factor 4/6; FGF, unassigned 1</td>
<td>KH.C5.697</td>
</tr>
<tr>
<td>7</td>
<td>Fgk8/17/18</td>
<td>Fibroblast growth factor 8/17/18</td>
<td>KH.C5.5</td>
</tr>
<tr>
<td>8</td>
<td>Fgf</td>
<td>Fibroblast growth factor receptor</td>
<td>KH.S7.42</td>
</tr>
<tr>
<td>9</td>
<td>Foxf</td>
<td>FoxF</td>
<td>KH.C3.170</td>
</tr>
<tr>
<td>10</td>
<td>Foxg-r</td>
<td>Foxg-related; Orphan Fox-4; Cz-ZF248</td>
<td>KH.C5.74</td>
</tr>
<tr>
<td>11</td>
<td>Fad5/5</td>
<td>Fad5/5</td>
<td>KH.C9.260</td>
</tr>
<tr>
<td>12</td>
<td>Gate4/5/6</td>
<td>GATA-a</td>
<td>KH.C1.120</td>
</tr>
<tr>
<td>13</td>
<td>Hand</td>
<td>Heart And Neural Crest Derivatives Expressed 1/2</td>
<td>KH.C1.606</td>
</tr>
<tr>
<td>14</td>
<td>Hand-r</td>
<td>Hand-related; Hand-like; NoTrbc</td>
<td>KH.C1.1116</td>
</tr>
<tr>
<td>15</td>
<td>Htr7-r</td>
<td>5-Hydroxytryptamine Receptor 7-related</td>
<td>KH.S5.55</td>
</tr>
<tr>
<td>16</td>
<td>Id1</td>
<td>Idlet1/2</td>
<td>KH.L1.522</td>
</tr>
<tr>
<td>17</td>
<td>Lef1</td>
<td>LeF1/TCF</td>
<td>KH.C6.71</td>
</tr>
<tr>
<td>18</td>
<td>Myf</td>
<td>Muscle regulatory factor; MyoD</td>
<td>KH.C1.307</td>
</tr>
<tr>
<td>19</td>
<td>Neurog</td>
<td>Neurogenin; Ngn</td>
<td>KH.C1.129</td>
</tr>
<tr>
<td>20</td>
<td>Nk4</td>
<td>Nk4-2/5; Tninn</td>
<td>KH.O8.482</td>
</tr>
<tr>
<td>21</td>
<td>Rhod/f</td>
<td>Rhod/F; Rhf</td>
<td>KH.C1.129</td>
</tr>
<tr>
<td>22</td>
<td>Tle.b</td>
<td>Groucho2</td>
<td>KH.L9.50</td>
</tr>
<tr>
<td>23</td>
<td>Tll</td>
<td>Tolloid-like; Tolloid</td>
<td>KH.C1.156</td>
</tr>
</tbody>
</table>

The 23 genes targeted in the initial screen, each identified here by official gene symbol, aliases, and KyotoHoya identifier.
eacy induced by the Ebf.3 sgRNA was 66% in sorted sample versus 45% in mixed sample. This suggests the actual e
ffi
cacies of some sgRNAs may be up to 1.5-fold higher than their estimated rates.

Analysis of unique indels generated by the activity of two e
ffi
cient sgRNAs, Ebf.3 and Lef1.2, indicated a bias towards deletions rather than insertions, at a ratio of roughly 2:1 deletions:insertions (Fig. 1e). However, these two sgRNAs generated different distributions of indel lengths, indicating indel position and size may depend on locus-specific repair dynamics as previously shown (Bae et al., 2014).

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 Section 3 for details). To test the assumption that off-target DSBs are unlikely at 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 5570 and 6690 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.

2.2. Identifying sequence features correlated with sgRNA eacy

We analyzed our dataset for potential correlations between target sequence composition and sgRNA-specific mutagenesis rate (excluding the Bmp2/4.1 sgRNA because only two reads mapped to its target sequence). We hypothesized that, if mutagenesis eacy can be predicted by nucleotide composition at defined positions in the protospacer and flanking sequences, then comparing the target sequences of the most or 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 eacy (Fig. 2; 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 (Fig. 2). 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 experimental systems (Gagnon et al., 2014; Ren et al., 2014; Wu et al., 2014; Chari et al., 2015; Moreno-Mateos et al., 2015; Haeussler et al., 2016).

Several studies in different model organisms have similarly examined the nucleotide preferences amongst sgRNAs inducing high or low rates of mutagenesis, and put forth predictive heuristics and/or
algorithms for the rational design of highly active sgRNAs (Doench et al., 2014, 2016; Gagnon et al., 2014; Ren et al., 2014; Wang et al., 2014; Chari et al., 2015; Fusi et al., 2015; Housden et al., 2015; Moreno-Mateos et al., 2015; Wong et al., 2015; Xu et al., 2015). These various algorithms have been evaluated, summarized, and aggregated by the CRISPR sgRNA design web-based platform CRISPOR (Haeussler et al., 2016). According to this meta-analysis, the “Fusi/Doench” algorithm (“Rule Set 2”) (Fusi et al., 2015; Doench et al., 2016) is the best at predicting the activities of sgRNA transcribed in vivo from a U6 small RNA promoter transcribed by RNA Polymerase III (PolIII), while CRISPRscan is recommended for sgRNAs transcribed in vitro from a T7 promoter (Moreno-Mateos et al., 2015).

Ciona rearing conditions differ from most other model systems, being a marine invertebrate that develops optimally at 16–24 °C (Bellas et al., 2003). 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 promoter (Nishiyama and Fujiwara, 2008). We used CRISPOR to calculate predictive scores for all sgRNAs in our data set and compare these scores to our mutagenesis efficacy measurements for each (Fig. 3a, Supplementary Table 1). We found that indeed the Fusi/Doench score best correlated with our measured sgRNA efficacies, with a Spearman’s rank correlation coefficient (rho) of 0.435 (p=3.884e-05) (Fig. 3a). CRISPRscan, the algorithm based on in vitro-, T7-transcribed sgRNAs injected into zebrafish, was less predictive (rho=0.344, p=0.001435) (Supplementary Table 1), supporting the conclusion that sgRNA expression method (e.g. U6 vs. T7) accounts for an important parameter when choosing an sgRNA design algorithm. Scores computed by other published algorithms available on the CRISPOR platform did not yield good correlations with our measurements (Supplementary Table 1), indicating these are perhaps not suited for predicting sgRNA activity in Ciona.

Fig. 2. Correlations between sgRNA sequence composition and mutagenesis efficacy. a) 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. b) Log-odds scores depicting the frequency of occurrence for nucleotides in the top 25% and bottom 25% 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. Positive and negative log-odds scores reflect abundance and depletion, respectively, of a specific nucleotide at a given position relative to its random occurrence in our sample space. See Section 4 for details.
Fig. 3. Using Fusi/Doench scores to predict mutagenesis efficacies of sgRNAs in Ciona. a) Mutagenesis efficacy rate of each sgRNA plotted against the same sgRNA’s Fusi/Doench algorithm score, obtained from CRISPOR. The Spearman’s rank correlation coefficient (rho) is 0.435 (p=3.88e-05). b) sgRNAs grouped by sorted Fusi/Doench predicted scores (left: > 60; right: < 50). 18 of 23 sgRNAs of Fusi/Doench score over 60 showed a measured mutagenesis efficacy (MUT%) over 24%, classified as “good” (shaded green). In contrast, only 4 from the same set had a MUT% under 24%, classified as “bad”. “Good” and “bad” classifications were based on phenotype frequency in F0 (see text for details). Out of 25 sgRNAs with Fusi/Doench score under 50, 17 were “bad”, while only 8 were “good”. c) A Receiver Operating Characteristic (ROC) curve assesses the credibility of using Fusi/Doench score cutoff (from 0 to 100) to classify sgRNAs as either “good” (> 24.5% efficacy) or “bad” (<24.5% efficacy). Using Fusi/Doench cutoffs as such a classifier returns an AUC of 0.77 (black line), while an AUC of 0.5 (dashed red line) represents the performance of a classifier solely based on random chance. The optimal Fusi/Doench cutoff (above which a score is likely to indicate “good” sgRNAs) was found to be between 50 and 55. See Section 4 and Supplementary Table 2 for details.

In a previous study, highly penetrant, tissue-specific loss-of-function phenotypes in F0 embryos were elicited using the Ebf.3 sgRNA (Stolfi et al., 2014), which in our current study had a measured mutagenesis efficacy of 37%. Further comparison of measured mutagenesis efficacy and mutant phenotype frequency for a series of Ebf-targeting sgRNAs revealed that an estimated sgRNA efficacy at least 25% correlated with over 70% reduction in the frequency of embryos expressing an Islet reporter construct (see Section 2.5 and Fig. 6). We thus reasoned that a mutagenesis efficacy of ~25% would be the minimum threshold of acceptable sgRNA activity for loss-of-function assays in F0. Within our dataset, among the sgRNAs that had a Fusi/Doench score > 60, 18 of 23 (78%) had a mutagenesis efficacy rate over 24% (Fig. 3b). In contrast, only 8 of 25 (32%) of sgRNAs with a Fusi/Doench score < 50 had an efficacy over 24% (Fig. 3b). Indeed, a receiver operating characteristic (ROC) analysis showed an area under the curve (AUC) of 0.77 when using Fusi/Doench score as a classifier of “good” (> 24.5% efficacy) vs. “bad” (<24.5%) sgRNAs (Fig. 3c). The most accurate Fusi/Doench score cutoffs appeared to be between 50 and 55, when taking both specificity and sensitivity into account (Supplementary Table 2, Supplementary Figure 2). However, if the number of candidate sgRNAs is not limiting, it may be more desirable to use a Fusi/Doench score cutoff of 60 in order to avoid false positives (i.e., sgRNAs that are predicted as “good” when in reality they are “bad”). Thus, for Ciona, we recommend using CRISPOR to select sgRNAs with Fusi/Doench scores > 60 and avoid those with Fusi/Doench < 50. We believe this approach will significantly streamline the search for suitable U6 promoter-driven sgRNAs targeting one’s gene of interest.

2.3. 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 large deletions in F0 embryos, we employed the forkhead/winged helix transcription-factor-encoding gene Foxf (Fig. 4a), which contributes to cardiopharyngeal development 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). We extracted genomic DNA from electroporated embryos and PCR-amplified the sequence spanning both target sites (Fig. 4b). 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 (Fig. 4b). Cloning the deletion band and sequencing individual clones confirmed that the short 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 3). 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
regulatory sequences from 2007). We electroporated expression in early animal pole blastomeres (Rothbächer 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 Fig. 4. 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). Nk4 and small portions of the flanking exons. The sgRNAs targeting Mzf were shown to inhibit its function and subsequent siphon muscle development (Tolkin and Christiansen, 2016).

Foxf is expressed in TVCs and head epidermis (Fig. 4c), 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 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 Foxf > nls::Cas9::nls together with Foxf.2 and Foxf.4 sgRNA vectors and Foxf > 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 some epidermal cells by the plasmid prep (2.5 µg vs. 75 µg). This discrepancy could indicate a lower efficiency of transcription of linear vs. circular transfected DNA, though more thorough analyses are warranted to investigate the behavior of linear DNA in Ciona embryos.

2.4. 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 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 (Fig. 5, Supplementary Figure 4, see Section 4 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. We electroporated Ebf.3 with U6 > Ebf.3 and 25 µl 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 (Supplementary Table 1). 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). This discrepancy could indicate a higher efficiency of expression of linear vs. circular transfected DNA, though more thorough analyses are warranted to investigate the behavior of linear DNA in Ciona embryos.

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 reporter transgene in MN2 motor neurons (Ryan et al., 2016), 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 25 µg U6 > Ebf.3 plasmid, but not with 25 µl (~2.5 µg) of unpurified U6 > Negative Control sgRNA PCR cassette (Fig. 6a–d). 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.
2.5. Correlation between sgRNA efficacy and mutant phenotype penetration

Since we measured a wide range of mutagenesis efficacies across our library of 83 sgRNAs, we wanted to assess how this variability correlates with mutant phenotype penetrance. One possibility is that sgRNA efficacies and frequencies of mutant phenotypes in F0 are linearly correlated. Alternatively, mutant phenotypes might only be observed at distinct thresholds of sgRNA activity.

To test this, we designed 7 additional sgRNAs (Ebf.A through Ebf.G) targeting the IPT and HLH domain-coding exons of Ebf (Fig. 6a,e, Supplementary Table 3). We generated these sgRNA expression cassettes by OSO-PCR, and co-electroporated 25 µl of each reaction with either 25 µg EefIa1 > nlsp:Cas9::nls (for mutagenesis efficacy sequencing) or 40 µg Sox1/2/3 > nlsp:Cas9::nls +15 µg Sox1/2/3 > H2B::eGFP +40 µg Islet > mCherry (for phenotype assay). sgRNA efficacies were measured by direct Sanger sequencing of the target region PCR-amplified from larvae electroporated with the former combination (see Section 4 for details), and Islet > mCherry expression in H2B:eGFP+ motor neurons was scored in larvae electroporated with the latter mix.

Quantification of indel-shifted electrophoresis chromatogram peaks (“Peakshift” assay) revealed sgRNA mutagenesis efficacies ranging from 5% to 43% (Fig. 6e, Supplementary Table 3). Of note, the sgRNA with the lowest efficacy (Ebf.E) was clearly hampered by a naturally occurring SNP that eliminates the PAM (NGG to NGA) in a majority of haplotypes. In parallel, the proportion of transfected embryos Islet > mCherry expression was observed to vary between 0% and 94%. When this was plotted against the measured mutagenesis efficacies, we observed a nearly perfect linear correlation between the mutant phenotype (loss of Islet > mCherry expression) and Ebf mutagenesis (Fig. 6e, Supplementary Table 3). Taken together, these results suggest that highly efficient sgRNAs can be generated from OSO-PCR cassettes and validated for loss-of-function studies by Sanger sequencing. The Sanger sequencing-based peakshift assay is highly reproducible and approximates the efficacy rates estimated by NGS (Supplementary Table 6). As such, we recommend the peakshift assay as a cheap, fast alternative for testing sgRNA activity in Ciona embryos.

2.6. Genome-wide design and scoring of sgRNAs for Ciona

While Ciona researchers will find the Fusi/Doench scoring system and the OSO-PCR method useful for sgRNA expression cassette design and assembly, we hoped to further empower the community by preemptively designing all possible sgRNAs with a unique target site within 200 bp of all Ciona exonic sequences. We computationally identified 3,596,551 such sgRNAs. We have compiled all sgRNA sequences and their corresponding specificity and efficiency scores (including Fusi/Doench), they are available as a UCSC Genome Browser track for the C12 genome assembly (http://genome.ucsc.edu/; with links to CRISPOR for off-target prediction and automated OSO-PCR primer design. This track is freely available for download to use on other browsers, like those of specific interest to the tunicate community such as ANISEED (Brozovic et al., 2015) or GHOST (Satou et al., 2005).

3. 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 reliable protocols for the validation of sgRNA efficacy in electroporated Ciona embryos by either next-generation sequencing or Sanger sequencing. This has allowed us to estimate the activity of all these sgRNAs, which are ready to be used for ongoing and future functional studies.

By analyzing correlations between target sequence nucleotide composition and sgRNA mutagenesis efficacy, we identified sgRNA sequence features that may contribute to Cas9:sgRNA activity. Some of these 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 (Doench et al., 2014; Gagnon et al., 2014; Ren et al., 2014; Chari et al., 2015; Moreno-Mateos et al., 2015). For instance, sgRNA efficacy is correlated with increased...
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. We also encountered a depletion of thymine and cytosine in the PAM-proximal nucleotides of the protospacers for highly active sgRNAs. 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 can be promoted by degenerate poly-dT tracts (Nielsen et al., 2013). A high number of non-consecutive thymines clustered in the protospacer could thus result in lower sgRNA expression level due to premature termination of sgRNA transcription (Stolli et al., 2014; Wu et al., 2014). Similarly, adenines are thought to contribute to the instability of sgRNAs (Moreno-Mateos et al., 2015), suggesting that CRISPR/Cas9 mutagenesis efficiencies might be 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 the lack of biological replicates for the NGS-based measurements, we have presented different data suggesting that sgRNA mutagenesis efficiencies are reproduceable between experiments, and not greatly affected by electroporation variability. First and foremost, we found a high correlation (Spearman's rank correlation coefficient of 0.435, p=3.884e-05) between our NGS data and scores from the predictive Fusi/Doench algorithm (Fig. 3). One would expect little correlation between our NGS data and predicted scores (by any algorithm) if the variation observed in our data were caused by technical variability and not innate sgRNA efficacy. Second, all three Ebf3 sgRNA vector formats (plasmid, traditional PCR cassette, OSO-PCR cassette) resulted in very similar mutagenesis efficiencies by NGS and Ebf loss-of-function phenotypes, which were assayed in two separate sets of experiments (Fig. 6e). Lastly, there was little variability in efficacy between biological replicates of a subset of sgRNAs as measured by peakshift (Supplementary Table 6).

Although more replicates and better statistical support would be required to develop a Ciona-specific sgRNA prediction algorithm de novo, we believe our data are accurate enough to support using the Fusi/Doench algorithm as the gold-standard in sgRNA design for Ciona (Fig. 3). Among the various published sgRNA prediction algorithms, Fusi/Doench (Fusi et al., 2015; Doench et al., 2016) functions well as a classifier for “good” (>24.5% mutagenesis efficacy) and “bad” (<24.5%) sgRNAs. Despite these general trends, several sgRNAs defied this algorithm-based prediction. 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. Other factors that can influence
Cas9 binding include chromatin accessibility and nucleosome occupancy (Wu et al., 2014; Hinz et al., 2015; Horbeck et al., 2016; Isaac et al., 2016). What other additional factors contribute to variability of in vivo sgRNA mutagenesis efficiencies 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.

While an optimized predictive algorithm for Ciona-specific sgRNA design remains a desirable goal, our current approach should help other researchers to identify, with greater confidence, which sgRNAs are likely to confer enough mutagenic activity for functional studies in F0. We have shown, using a series of Ebf-targeting sgRNAs and an Ebf loss-of-function readout assay (Islet reporter expression) that a linear correlation exists between sgRNA mutagenesis efficacy and the probability of somatic gene knockout in F0. In this case, we measured the frequency of a binary loss-of-function assay (Islet ON or OFF) in a large population of embryos. Therefore it is expected that the probability a homozygous Ebf knock out (resulting in Islet OFF phenotype) should be correlated to the observed frequency of Ebf mutations in somatic cell populations. While Ebf function can be disrupted by small changes to its crucial IFT/HLH domains, other genes may prove harder to disrupt with the short indels generated by CRISPR/Cas9. However, we show that the combinatorial action of two or more sgRNAs can result in high frequency of large deletions spanning many kbp, which should help generate loss-of-function alleles for such recalcitrant genes.

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 no strongly 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 similar protoscaler sequences. Second, the GC content of the Ciona genome is only 35% compared to 65% in humans, which should result in a lower overall frequency of canonical PAMs (NGG). 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.

Even with improved prediction of sgRNA efficacy and specificity, there is still a need to test several sgRNAs to identify the optimal one targeting a gene of interest. To this end, we have developed a cloning-free OSO-PCR method for the rapid assembly of sgRNA expression cassettes. We have shown that these unpurified PCR cassettes can be directly electroporated into Ciona embryos and screened by either Sanger sequencing of target sequences or mutant phenotype frequency in F0. The automated design of primers for Ciona-specific sgRNA cassette OSO-PCR has been implemented in the latest version (v4.0) of CRISPOR (http://crispor.tefor.net/).

Finally, we have pre-emptively designed all possible sgRNAs targeting exonic sequences in the compact Ciona genome and calculated their specificity and efficiency by various predictive algorithms. This track is available online on the UCSC genome browser, but is also freely available for download and use with other genome browsers. This allows researchers to locally browse for sgRNAs with predicted high activity targeting their loci of interest. Integration with the CRISPOR website further allows SNP and off-target prediction, and pre-designed OSO-PCR oligonucleotide primers for rapid, efficient synthesis and transfection. We expect these resources to facilitate the scaling of CRISPR/Cas9-mediated targeted mutagenesis and enable genome-wide screens for gene function in Ciona.

4. Materials and methods

4.1. 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 (Supplementary Table 3). Two sgRNAs were designed to target the neurogenic bHLH factor Neurogenin, a gene that is not expressed in the cardiopharyngeal mesoderm and is not thought to be involved in cardiopharyngeal development. Target sequences were selected by searching for N19, NGG (proteasome + PAM) motifs and screened for polymorphisms and off-target matches using the GHOST genome browser and BLAST portal (Satou et al., 2005, 2008). Potential off-targets were also identified using the CRISPRdirect platform (Naito et al., 2015). sgRNA expression plasmids were designed for each of these protoscaler and constructed using the U6 > sgRNA(F+E) vector as previously described (Stolli et al., 2014), as well as a “Negative Control” protoscaler 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 to multiple loci 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.

4.2. 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 ng in 700 µl) concentration together with Eef1a1 > nls::Cas9::nls plasmid (Stolli et al., 2014) at 35.7 ng/µl (25 µg in 700 µl) concentration. For testing U6 > Ebf3 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 h post-fertilization.

4.3. 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 min, followed by 95 °C for 10 min.

4.4. 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 4) were designed to flank target sites as to obtain PCR products in the size range of 108–290 bp with an exception of the sequence targeted by Ebf3 (‘Ebf3’ Genome Coordinates reference: KhC2:8,066,844–2,068,866) and Fgf4/6.1 (TTATTTATCTGTACATACATT-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 to multiple loci 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.

4.2. 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 ng in 700 µl) concentration together with Eef1a1 > nls::Cas9::nls plasmid (Stolli et al., 2014) at 35.7 ng/µl (25 µg in 700 µl) concentration. For testing U6 > Ebf3 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 h post-fertilization.

4.3. 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 min, followed by 95 °C for 10 min.

4.4. 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 4) were designed to flank target sites as to obtain PCR products in the size range of 108–290 bp with an exception of the sequence targeted by Ebf3 (‘Ebf3’ Genome Coordinates reference: KhC2:8,066,844–2,068,866) and Fgf4/6.1 (TTATTTATCTGTACATACATT-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 to multiple loci 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.
TGAGGTACCACTACCGTTT-3' (Eef3), and 5'-AAGCA TTTGCTTACAGGAGA-3' and 5'-ACTTCCCAACGCAAATGGG-3' (Fgf6.1).

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 MgSO4, 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 min at 95 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at 60 °C and 30 s at 68 °C, and a final cycle of 3 min 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 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.

4.5. 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 barcode adapters. (NEXTflex DNA Barcodes - BioO Scientific Cat# 514101). The six barcodes used in this study were: CGATGT, TGACCA, ACAGTG, GCCAAT, CAGATC and CTGGTA. At this step, the adapter-ligated DNA fragments were purified twice using Ampure XP beads (Beckman Coulter, Brea, CA). The final amplification, with primers included with NEXTflex adapters, was done using the PCR program: 2 min at 98 °C followed by 8 cycles of 30 s at 98 °C; 30 s at 60 °C; 15 s of 72 °C, followed by 10 min 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 2×250 paired end configuration.

4.6. 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 same 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 apilcone 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. Custom python scripts used 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. All predictive algorithm scores were generated using CRISPOR (http://crispor.tefor.net/) (Haessler et al., 2016).

4.7. Nucleotide enrichment analysis

We used log-odds score as a measure to estimate how enriched each nucleotide was at a given position in the 43 bp region of interest (excluding position 1 of the protospacer, and the ‘GG’ of the PAM). Log-odds score was defined as the base-2 logarithm of the ratio of probability of observing nucleotide ‘N’ at position ‘x,’ and the background probability of observing nucleotide ‘N’ by random chance, given its frequency in our sample space (all sgRNA targets, n=83). A positive or negative log-odds score reflects enrichment or depletion, respectively, of each nucleotide at a given position.

Mathematically,

\[ I(x) = \log_2 \left( \frac{P_p}{P_b} \right) \]

\[ P_p = \text{probability of finding nucleotide ‘N’ at position ‘x’} \]

\[ P_b = \text{background probability of finding nucleotide ‘N’ at position ‘x’ by chance} \]

4.8. Receiver Operating Characteristic (ROC) curve

For a binary classifier (“good” vs. “bad” sgRNA), there are four possible outcomes: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). If the predicted classification of the sgRNA by our model was “good” or “bad”, and it was supported by our experimental data (i.e. the empirical mutagenesis rate was above or below 24.5% respectively), it was marked as a “True Positive” or a “True Negative,” respectively. If the experimental data failed to support it, it was marked as a “False Positive” or a “False Negative,” respectively. The true positive rate (TPR), or “Sensitivity,” was defined as the proportion of empirically good sgRNAs that were correctly predicted as good [TPR = TP / (TP + FN)]. Similarly, the true negative rate (TNR), or “Specificity,” was defined as the proportion of empirically bad gRNAs that were correctly predicted as bad [TNR = TN / (TN + FP)]. The false positive rate (FPR) is 1 - Specificity. In Fig. 3c, we plotted the true positive rates against the false positive rates, all obtained by applying different Fusi/Doench score thresholds (ranging from 0 to 100) to the predictions generated by our model (see Supplementary Table 2).

4.9. Combinatorial sgRNA electroporation to induce large deletions

Embryos were electroporated with 25 µg Eef1al > 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 s 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 4). Products were purified from agarose gels, A-overhang and TOPO-cloned. Colonies were picked, cultured, prepped and sequence.

4.10. 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'-TGGCGGTATTTAACAC-C3') and sgRNA reverse (5'-GAGATTTCCATGGAAATACG-3') in reactions of final concentrations of 600 nM each primer, 300 µM dNTPs, 1 mM MgSO4, 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 min at 95 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at 55 °C and 2 min at 68 °C, and a final cycle of 5 min at 68 °C. U6 > sgRNA(P+E)::eGFP (Stolfi et al., 2014) was amplified as above but using Seq forward primer (5'-AGGTATTTGCTCTATGAGCC-3') instead. For phenotyping Ebf-dependent expression of Islet reporter in MN2 motor neurons (Stolfi et al., 2014), embryos were co-electroporated with 35–40 µg of Sox1/2/3 > nls::Cas9::nls, forward (5'-AGTCGCGAGTACCTACGT-3') and reverse (5'-AAGTACGGAGAGCAGAATACC-3') primers and either 25 µg of U6 > Ebf.3 plasmid or 25 µl (~2.5 µg) of unpurified PCR product.

4.11. 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'-GTTTAAGACGCTGCTGCTACG-3') and its reverse complement is appended 5' to a reverse primer (5'-ACTTATACGCTGCTGCTACG-3'). These primers are then added to a PCR reaction at limiting amounts, together with U6 forward (5'-TGGCGGTATTTAACAC-C3') and sgRNA reverse (5'-GAGATTTCCATGGAAATACG-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_Chiastrien/). 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 Fig. 5 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 5).

4.12. Measuring mutagenesis efficiencies of sgRNAs by Sanger sequencing

OSO-PCR cassettes were designed and constructed for the expression of 7 sgRNA targeting exons 10–12 that code for part of the IFT and HLH domains of Ebf (Fig. 6a, Supplementary Table 3). Embryos were electroporated with 25 µl of unpurified OSO-PCR product and 25 µg Eef1a1 > nls::Cas9::nls, and grown to hatching. For the replication experiment (Supplementary Table 6) 75 µg of sgRNA vectors (Gata4/5/6.2, Gata4/5/6.3, Gata4/5/6.4, Neurog.1, Neurog.2) were each co-electroporated with 25 µg Eef1a1 > nls::Cas9::nls. This was repeated over two separate batches of embryos.

All larvae were allowed to hatch and genomic DNA was extracted using the QiaAmp DNA mini kit (Qiagen). The Ebf target was PCR-amplified using the following primers: (Forward primer: 5'-CTCCACATCGCTAActTTG-3', Reverse primer: 5'-TGTTACGCCAAAATTTGGAAG-3'). For Gata4/5/6 target sequences, the primers used were the corresponding ones from the NGS experiment, while a novel pair of primers was used to amplify the Neurog locus (Forward primer: 5'-AAGTACGGAGAGCAGAATACC-3', Reverse primer: 5'-CTCTTACTGCTCATTAAGACC-3'). PCR was performed in 35 cycles using Platinum PfX DNA polymerase. The resulting PCR products were gel- or column-purified, and sequenced using either a flanking PCR primer or an internal sequencing primer (Ebf internal: 5'-AATTTGGCTGACAGTTGGAG-3', Neurog internal: 5'-GCTCTTGTCTACAAAATGTGG-3').

The resulting ab1 sequencing files were then analyzed by ab1 Peak Reporter webtool (Roy and Schreiber, 2014) (https://apps.thermoshiffer.com/ab1peakreporter/) (Supplementary Table 5). To quantify the peak “shifts” resulting from CRISPR/Cas9-induced short indels (Supplementary Figure 6), we calculated the sum of “maximum signal 7 scan filtered ratio” (MaxSig7Scan Sum) values of minor peaks at each position in a 30 bp window starting from the third bp in the target sequence from the PAM (the most likely Cas9 cut site). The mean MaxSig7Scan Sum was calculated across all 30 bp of this window, resulting in a “raw peakshift score”. The same was repeated for the 30 bp window immediately 5' to the cut site in the sequencing read, for a peakshift “baseline” estimate. This baseline was subtracted from the raw peakshift score to give the “corrected peakshift score”, a relative quantification of indel frequency, and therefore of the mutagenesis efficacy of the sgRNA used.

4.13. Whole-exome sgRNA predictions in Ciona

We used all transcribed regions in the ENSEMBL 65 transcript models (Aken et al., 2016), extended them by 200 bp on each side, searched for all possible NGG-20 bp sgRNA targets in these sequences and ran them through the command-line version of CRISPOR (Haeussler et al., 2016) which aligns 20mers using BWA (Li and Durbin, 2009) in iterative mode, ranks off-targets by CFM or MIT scores and calculates the Fusi/Doench 2016 efficiency scores (Fusi et al., 2015; Doench et al., 2016). Efficiency scores were also translated to percentiles for better ease of use, e.g. a raw Fusi/Doench score of 60 translates to the 80th percentile, meaning that 80% of scored guides are lower than 60. All results are then written to a UCSC BigBed file (Raney et al., 2014) for interactive visualization. The BigBed file can be loaded into all popular genome browsers, like Ensembl (Yates et al., 2016), IGV (Robinson et al., 2011) or GBrowse (Stein et al., 2002). The source code is available as part of the UCSC Genome Browser source tree at: https://github.com/ucscGenomeBrowser/kent/tree/master/src/hg/makeDb/crisprTrack.

4.14. Embryo imaging

Fluorescent in situ hybridization of eGFP or Foxf coupled to immunohistochemistry 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 epi-fluorescence microscope. Mouse monoclonal anti-bGal 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 Farhana Salek, Kristyn Millan, and Aakarsha Pandey for 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; Justin S. Bois, 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 an NIH K99 HD084814 award to A.S., NIH R01 GM096032 award to L.C., award 15CVD01 from the Leucq
Foundation to L.C., and an NYU Biology Masters Research Grant to S.G.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ydbio.2017.03.003.

References


