Molecular Chipper: Functional Mapping of the Non-Coding Genome with CRISPR

Jijun Cheng[1,2*], **Wen Pan**[1,2,] **Christine Roden**[1,2,3], **Zhuo Chen**[4] **and Jun Lu**[1,2,5,6]

[1]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA

[2]Yale Stem Cell Center, Yale Cancer Center, New Haven, Connecticut, 06520, USA

[3]Graduate Program in Biological and Biomedical Sciences, Yale University, New Haven, Connecticut 06510, USA

[4]Department of Biomedical Engineering, Yale University, New Haven, Connecticut 06520, USA

[5]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

[6]Yale Center for RNA Science and Medicine, New Haven, Connecticut, 06520, USA

*Corresponding author: Jijun Cheng, Department of Genetics, Yale University School of Medicine, 10 Amistad Street, Room 214K, New Haven, CT 06520-8005, USA, Tel: +2037374652; Fax: +2037854305; E-mail: j.cheng@yale.edu

Abstract

Only ~2% of the human genome encodes proteins. The functions of the vast majority of non-coding sequences, transcribed or non-transcribed, remain largely unknown. Powered by next generation sequencing and evolving high through-put biochemical technologies, continuing efforts have been devoted to the biochemical annotation of genomic regions based on their transcription activity, DNase-I hypersensitivity, transcription factor binding, chromatin modifications and replication domains. However, biochemical annotation does not equal functional roles in molecular control or cellular behavior. While these genome-scale biochemical data pave the way for deciphering the non-coding genome, high-throughput mapping technologies are required to assign functional roles to specific noncoding elements. Here, we review the use of CRISPR-based technologies in functional screens of noncoding genomic elements, with a focus on the Molecular Chipper technology.

Keywords: Non-coding genome; CRISPR-Cas9; sgRNA library; Molecular chipper; Reporter screen; Next generation sequencing

Forward Genetics and CRISPR-Cas9 Screening

Earlier efforts in dissecting genome-wide loss-of-function biological phenotypes involved screening shRNA libraries that targeting >16,000 human and mouse genes, aided by microarray or next generation sequencing [1-8]. This approach is limited by the effectiveness of targeting by shRNA, and not all target genes can be nearly completely knocked-down. The recent clustered regularly-inter spaced palindromic repeats (CRISPR)-Cas9 technology, using computationally designed and chemically synthesized single-guide-RNA (sgRNA) libraries of coding and non-coding genes, can completely inactivate both alleles of target genes, thus leading to new possibilities in functional screenings [9-11].

The sgRNAs in these libraries (~1×10^5 in total number) have been rationally designed to target N-terminal of ~20,000 known protein-coding genes and ~1,000 miRNAs. Such libraries were generated by chemical synthesis using high throughput microarray-based oligonucleotide synthesis, followed by cloning into viral vectors [9-11]. Functional screens have been carried out by infecting mammalian cells to screen for mutants affecting a phenotype, recovering sgRNAs in mutants, and deep sequencing to reveal sgRNA enrichment in the genes that the sgRNAs were designed against (Figure 1A).

One limitation in these libraries seems to be sgRNAs targeting the N-terminal of proteins, generating in-frame functional variants to obscure genetics screening. An improved design of sgRNAs targeting functional domains in target proteins increased null mutant rates and improved negative-selection screens, utilizing a large number of sgRNAs per gene [12]. To go beyond mutating only one gene in each cell, a recent study screened a library of 23,409 barcoded dual sgRNAs, generated by combinations out of 153 gRNAs targeting 50 epigenetic genes, and identified gene pairs that inhibited ovarian cancer cell growth [13].

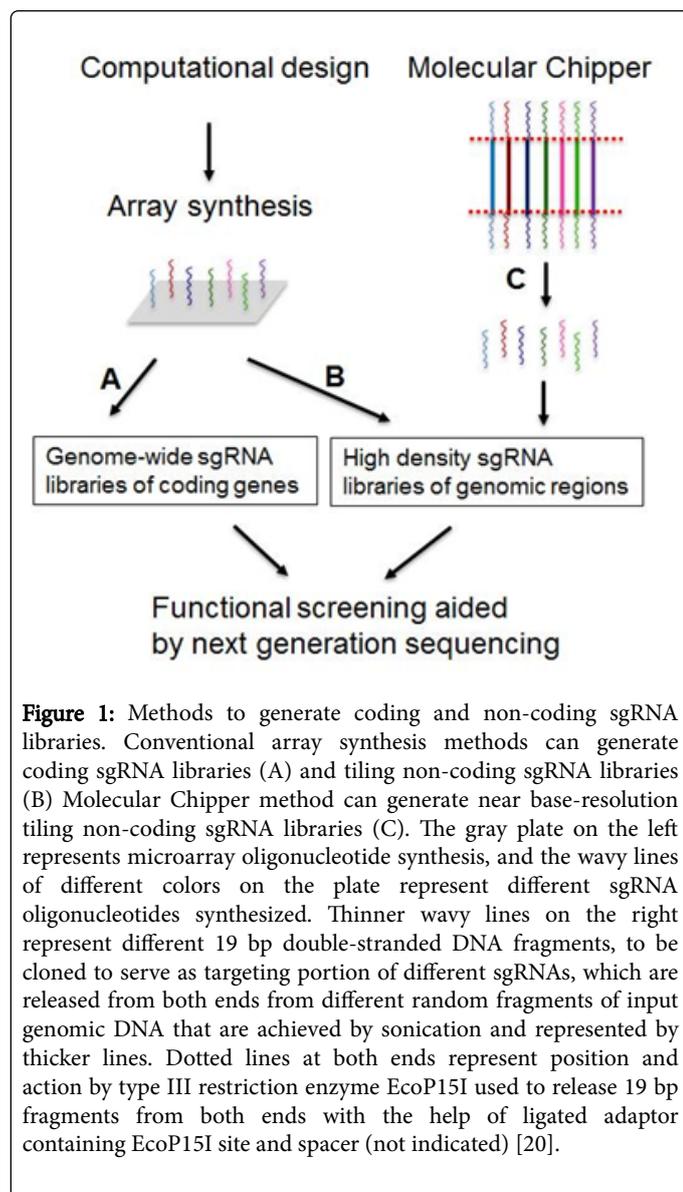## Screens of Functional Non-coding Elements by CRISPR-Cas9

Aided by CRISPR-Cas9-based screens, recent attempts have taken on the challenge of interrogating functional non-coding genomic elements (Figure 1B). Earlier work by Qi et al. showed that specific sgRNAs blocking key transcriptional factor binding sites can interfere with gene activation when a catalytically dead Cas9 was used [14]. Using a library of 702 sgRNAs covering a 12 kb region in the enhancer of region of human BCL11A, critical minimal features and discrete vulnerabilities for the expression of fetal hemoglobin in erythroid cells were identified [15].

Employing ~18,000 sgRNAs targeting a total >700 kb of noncoding sequences flanking three cancer resistance genes, specific noncoding locations near genes were identified to modulate drug resistance when mutated [16]. An elegantly designed multiplexed editing regulatory assay (MERA), developed to ensure only one sgRNA expressed per mutant cell and with knocked-in green fluorescence protein (GFP) as a readout, was able to screen 3,908 sgRNAs tiled across ~40 kb cis-regulatory regions of Tdgf1a and identify functional proximal and distal regulatory elements [17].

Genome-wide transcription factor binding data or genetic traits such as SNP association data can aid the design of sgRNA libraries to target and study the detailed functions of a sub set of genomic sites [18,19]. For example, screening 1,116 sgRNAs targeting 685 p53 genomic DNA binding sites, functional enhancers were identified to regulate p53, including two sites near CDKN1A, a well-known mediator of p53-induced cell cycle arrest [18].

CRISPR-Cas9 Molecular Chipper

Due to cost and special bioinformatics expertise involved in conventional sgRNA library generation by microarray synthesis, only a limited number of labs can make customized sgRNA libraries adapted to customized screenings for functional non-coding regions [19].

Figure 1: Methods to generate coding and non-coding sgRNA libraries. Conventional array synthesis methods can generate coding sgRNA libraries (A) and tiling non-coding sgRNA libraries (B) Molecular Chipper method can generate near base-resolution tiling non-coding sgRNA libraries (C). The gray plate on the left represents microarray oligonucleotide synthesis, and the wavy lines of different colors on the plate represent different sgRNA oligonucleotides synthesized. Thinner wavy lines on the right represent different 19 bp double-stranded DNA fragments, to be cloned to serve as targeting portion of different sgRNAs, which are released from both ends from different random fragments of input genomic DNA that are achieved by sonication and represented by thicker lines. Dotted lines at both ends represent position and action by type III restriction enzyme EcoP15I used to release 19 bp fragments from both ends with the help of ligated adaptor containing EcoP15I site and spacer (not indicated) [20].

To circumvent this challenge, we recently reported a new technique, named Molecular Chipper, that was based on standard molecular biology procedures to generate a tiling sgRNA library from a total of 10 kb input genomic DNA of covering miRNA genomic regions [20]. We used this library and identified novel cis-regulatory elements controlling miR-142 biogenesis [20]. The high-density and nearly base-resolution of the sgRNA library was achieved by random fragmentation of input genomic DNA piece(s), and the 19 bp ends of the DNA fragments are released to serve as targeting region of sgRNAs through adaptor ligation and type III restriction enzyme digestion. These ends were further ligated to sgRNA backbone and cloned into a viral expression vector (Figure 1C). Using a cell line with constitutive Cas9 expression and a GFP reporter for the loss of endogenous miR-142 expression, we infected the library and FACS-sorted GFP+ miR-142-inactvated cells, recovered sgRNAs from the genomic DNA, and identified enriched sgRNAs that led to miR-142 inactivation. We successfully identified a 5' and a 3' cis flanking regions as candidate regulatory regions for miR-142 expression. Subsequent mechanistic studies revealed these two cis regions control pri-miR-142 processing. To facilitate the analysis of sgRNA enrichment in noncoding regions, we also designed an algorithm called ESC Scanner to scan DNA regions of interest for clusters of enriched sgRNAs. This approach may be widely applicable for the identification of functional noncoding genomic elements.

The Future Perspectives of Molecular Chipper

The success of screening selected noncoding regions on the genome raised the possibility that true-genome-wide noncoding region screens could be performed at high resolution using the CRISPR system. However, achieving such a goal requires new advances to overcome key technical bottlenecks. One of the limiting bottlenecks is the large numbers of sgRNAs required for the genome-coverage. For the human genome of 3.2×10^9 bp [1], assuming that the protospacer adjacent motif (PAM) of nucleotide sequence "NGG" for Streptococcus pyogenes (sp) Cas9 is present in the double-stranded genome at an average of 8 bp (16 bp for each strand), 1X coverage for the genome will require $\sim 4 \times 10^8$ sgRNAs [21]. Due to cost problem, it is currently near impossible to use microarray-based synthesis of such a large number of sgRNAs. An alternative approach is the synthesis of random 19mers for generating completely randomized sgRNAs. This approach, although feasible by chemical synthesis, suffers from a huge amount of inactive sgRNAs- it would contain 4^{19} or 2.75×10^{11} sgRNAs, with only 0.145% sgRNAs having NGG as the PAM sequence and thus >99% are not useful by Sp-Cas9. At this moment, Molecular Chipper-based sgRNA library generation holds the potential for achieving high-density genome-wide sgRNA libraries. This method can be theoretically scaled using longer input DNA [20]. We are modifying the Molecular Chipper technology to clone only NGG-PAM specific sgRNAs from the input DNA. In this regard, because each human cell has 6.96 pg of genomic DNA, 4×10^8 19 bp sgRNA DNA fragments from the human genomic DNA only weighs 6.96*19/16 or 8.27 pg, and cloning of 8.27 ng of the DNA fragments would easily cover the genome at 1000X. However, a number of technical challenges need to be resolved before this goal can be realized.

One limitation of the published Molecular Chipper approach is that only one-sixteenth of the sgRNAs in this base resolution library are followed by PAM sequence NGG and thus useful in reporter cell line screening with wild type Cas9. Interestingly, we showed that the VQR (mutations D1135V, R1335Q, and T1337R) Cas9 variant using nucleotide sequence NAG as PAM sequence is compatible with this library, and indeed reporter cell line expressing both the wild type and the VQR variant was capable of generating more GFP+ mutants when infected with the library [22]. This suggest that, given multiple Cas9 variants with altered PAM or a Cas9 variant requiring no PAM, most if

not all of the sgRNAs could be functional to generate mutants. This would be particularly useful for identifying very small cis DNA elements. In addition to using multiple Cas9 variants, efforts are underway to generate sgRNA libraries consisting only NGG-PAM sgRNAs by modifying the current Molecular Chipper technology. Such future improvements can help the cloning of sgRNA libraries for long input DNA, such as eventually toward the genome-scale.

The second challenge is associated with the number of cells required to perform functional screens with libraries of at least 4×10^8 complexity. A 100x coverage would require $>10^{10}$ cells. Nevertheless, it could be possible with hematopoietic cells, which can be cultured at high density. For example, BaF3 cells can be routinely cultured at million/ml. Using large containers with ~1 liter volume, it could be possible to achieve sufficient cell numbers for a true-genome-wide screen.

In general, the CRIPSR-Cas9 coding gene screen relies on stable cell lines expressing sgRNAs and Cas9, high-specific and high-efficient sgRNAs to generate loss-of-function mutations of both alleles, and sgRNA enrichment rather than actual mutations in screening cells and, moreover, the CRIPSR-Cas9 efficiency varies in species, cell types, mutation types and target types [23,24]. Advances in these areas should benefit applications of Molecular Chipper-based CRISPR functional non-coding screens. Although both euchromatin and heterochromatin can be efficiently targeted by sgRNAs [25], recent careful studies indicate that chromatin conformation affect the CRISPR-Cas9 gene editing efficiency by a few to 10-fold. If this is true, it may complicate sgRNA enrichment analysis in screening genome-wide functional non-coding elements. Future Molecular Chipper-based studies involving large genomic DNA input or the whole genome should help to understand how differentially sgRNAs target euchromatin and heterochromatin regions efficiently and if this complication is manageable by separately analyzing sgRNA enrichment in the two types of regions.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

2. Kellisa M, Woldc B, Snyderd MP, Bernsteinb BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. PNAS 111: 6131-6138.

3. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe N, et al. (2003) The transcriptional activity of human chromosome 22. Genes & Dev 2002: 529-540.

4. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, et al. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296: 916-919.

5. Consortium EP (2013) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

6. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, et al. (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature 515: 355-64.

7. Dorsett Y, Tuschl T (2004) siRNAs: applications in functional genomics and potential as therapeutics. Nature Reviews. Drug Discov 3: 318-329.

8. Sims D, Mendes-Pereira AM, Frankum J, Burgess D, Cerone MA, et al. (2011) High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. Genome Biol 12: 1-13.

9. Zhou Y, Zhu S, Cai C, Yuan P, Li C, et al. (2014) High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature 509: 487-491.

10. Shalem O, Sanjana EN, Hartenian E, Shi X, Scott DA, et al. (2014) Genome-Scale CRISPR-Cas9 knockout screening in human cells. Science 343: 84-88.

11. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. Science 343: 80-84.

12. Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, et al (2015) Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nature Biotechnol 33: 661-667.

13. Wong ASL, Choi GCG, Cui CH, Pregernig G, Milani P, et al. (2016) Multiplexed barcoded CRISPR-Cas9 screening enabled by CombiGEM. PNAS 113: 2544-2549.

14. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, et al. (2013) Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell 152: 1173-1183.

15. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, et al. (2015) BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527: 192-197.

16. Sanjana NE (2016) Genome-scale CRISPR pooled screens. Anal Biochem 1-5.

17. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, et al. (2016) High-throughput mapping of regulatory DNA. Nature Biotechnol 34: 167-174.

18. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, et al. (2016) Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nature Biotechnol 34: 192-198.

19. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, et al. (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell 165: 1530-1545.

20. Cheng J, Roden CA, Pan W, Zhu S, Baccei A, et al. (2016) A Molecular Chipper technology for CRISPR sgRNA library generation and functional mapping of noncoding regions. Nat Commun 7: 11178.

21. Wiedenheft B, Sternberg SH, Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature 482: 331-338.

22. Kleinstiver BP, Prew MS, Tsai SQ, Topkar VV, Nguyen NT, et al. (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature 523: 481-485.

23. Xu T, Li Y, Van Nostrand JD, He Z, Zhou J (2014) Cas9-based tools for targeted genome editing and transcriptional control. Appl Environ Microbiol 80: 1544-1552.

24. Sander JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nature biotechnol 32: 347-355.

25. Knight SC, Xie L, Deng W, Guglielmi B, Witkowsky LB, et al. (2015) Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science 350: 823-826.

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

27. Chen X, Rinsma M, Janssen JM, Liu J, Maggio I, et al. (2016) Probing the impact of chromatin conformation on genome editing tools. Nucleic Acids Res 44: 6482-6492.