Functional study of the human genome

Li, Y.

Chapter 1

Introduction

Introduction

The human genome contains 3 billion base pairs. Since the release of the first draft human reference genome in 2000, the details of our own blueprint have been getting more and more accurate[1,2]. According to the most recent telomere-to-telomere sequencing release, one human genome contains 3,054,815,472 bp of nuclear DNA, plus a 16,569-bp mitochondrial genome with a total of 63,494 genes, of which 19,969 are predicted to code proteins[3]. If we consider the human genome as an instruction book to make a human being, we have a reasonably good grasp of the letters in this book. However, we are still far from understanding the content of this instruction book. Proteins are the main building blocks of cells and the body. Even for the proteins, new functions are constantly updated. Yet the sequence coding the proteins only represents less than 2% of the human genome[3].

The rest of the genome is the protein non-coding genome, which contains mainly repeat regions, transposons, gene introns, non-coding RNAs or other intergenic regions[6]. In fact, although the majority of the human genome contains non-coding sequences, many of them have certain functions and play important roles in regulating the transcription of genes. Depending on the function of the non-coding regulatory regions (NCREs), they can be categorized into a few major classes, such as promoters (where transcriptional machinery assemble and drive the transcription), enhancers (which boost transcription), insulators (which bridge distal NCREs with promoters or serve as the barriers), or silencers (which repress transcription)[7-11]. Following the human genome project, we also started to catalog our genome, not only based on its intrinsic genomic information but also on epigenomic build-ups. International consortium projects like ENCODE and Roadmap Epigenomics have made enormous contributions to profile the non-coding part of the genome based on their epigenetic modifications[12-17], hoping to locate and define the potential functions of the NCREs. Besides computational prediction, recent developments of massively parallel reporter systems also facilitate the direct biological activity measurement and identifications of promoters, enhancers and silencers[6,18,19].

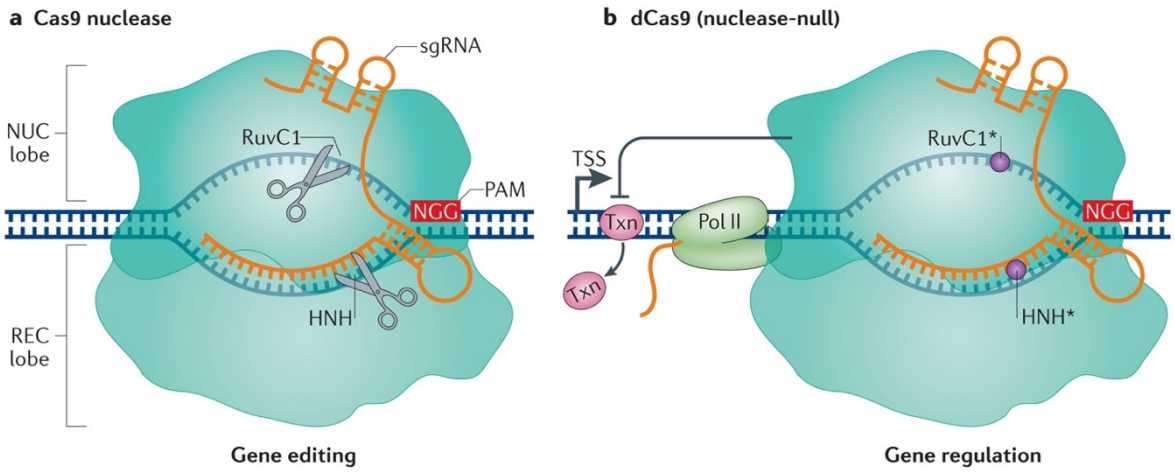

Figure 1: Mechanisms of Streptococcus pyogenes Cas9 and dCas9
a. The *S. pyogenes* Cas9 endonuclease consists of a nuclease (NUC) domain and a recognition (REC) domain. Cas9 is targeted to specific DNA sequences by direct pairing of the chimeric single guide RNA (sgRNA) with the target DNA. **b.** The *S. pyogenes* dCas9 protein contains mutations in its RuvC1 (D10A) and HNH (H841A) domains, which inactivate its nuclease function (circles). dCas9 retains the ability to target specific sequences through the sgRNA and PAM[5].

To understand the content of the human genome, it is key first to understand the function of their key elements—either genes or NCREs, before diving into their complex interactions to form the complicated human bodies. To study the function of genes, many tools have been developed during the past few decades to study the function of genes. The cloning of genes in the 70s led to the explosion of our understanding of gene functions[20]. In the following years, the discovery of siRNA and shRNAs provided a new way to study the functions of genes by reducing the expression levels of genes in cells[21]. Further development of these techniques allows for studying all potential genes in a genome-wide and systematic fashion focusing on distinct phenotypes[22-24]. The past decade witnessed the application of another revolutionary genetic tool to study the function of the genome, as exemplified by the CRISPR-Cas9-mediated genome editing systems[25-27] (Figure 1). Modification of the CRISPR-Cas9 systems not only allows the study of functions genes (in terms of silencing, upregulation and repression), but for the first time also allows the study of NCREs in their endogenous loci[5,28-31] (Figure 2).

CRISPR–Cas9 recognizes 20-bp genomic regions followed by PAM (5'-NGG-3') and typically introduces genetic changes of a few nucleotide deletions or insertions around the targeting sites in the genome[32]. When these mutations are introduced in the gene coding regions, it would render the proteins not functional anymore. On the other hand, NCREs often range from 50-200 bp in length, with multiple transcription factor (TF) binding sites[33], a single guide RNA mediated CRISPR editing may not completely destroy the TF binding sites[34,35]. Due to such limitations, single guide RNAs tiling an entire testing region were often used to study NCREs regulating a few important genes[28]. The modified CRISPR system that uses catalytic inactive Cas9 proteins (dCas9), either linked to a transcription activation or repression system, can also be used to study the functions of enhancers and insulators[35,36]. However, the design of tiling RNAs in selected few regions is also required. Therefore, systematic and genome-wide studies on the NCREs have not been performed yet.

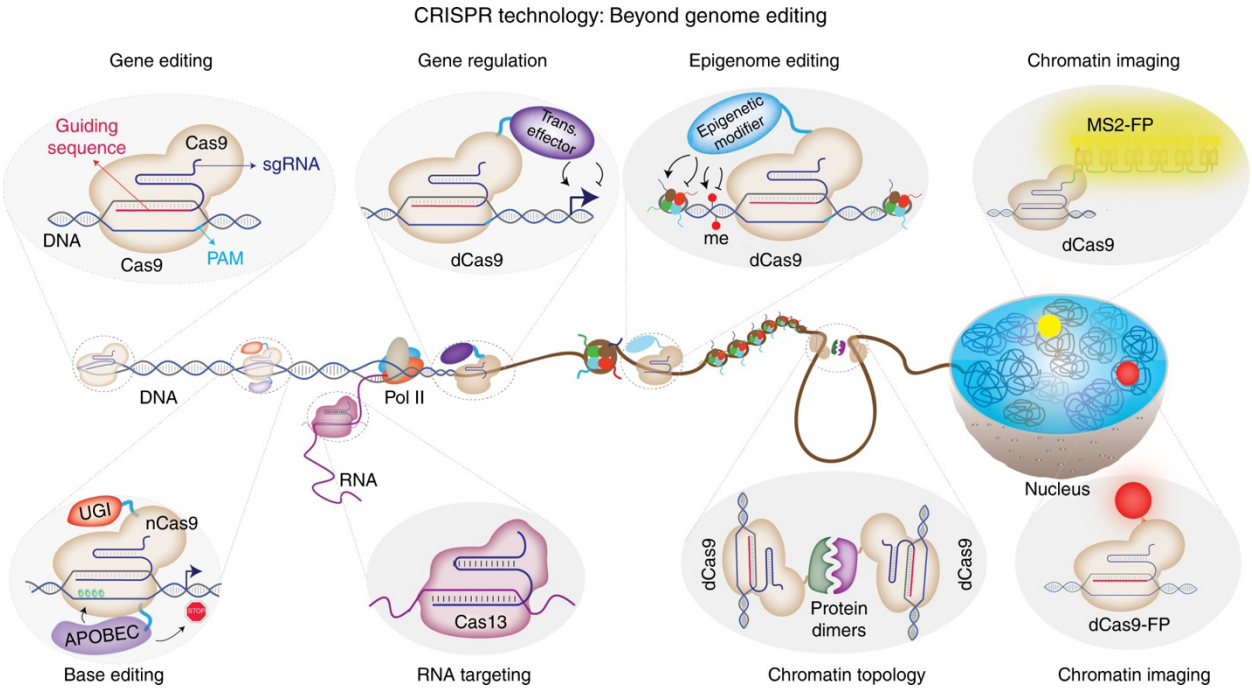

Figure 2: New applications based on CRISPR-Cas systems.
The CRISPR-Cas systems can be repurposed to achieve targeted gene regulation, epigenome editing, chromatin imaging, and chromatin topology manipulations[4].

In this thesis, research was performed to study the function of both the genes and the NCREs of the human genome in a systematic way, using different custom-designed CRISPR screening systems. In **Chapter 2**, aiming to understand the drug importing machinery, custom CRISPR CRISPR–Cas9 knockout (CRISPRko, to induce loss of function of endogenous genes) and CRISPR/dCas9 activation (CRISPRa, to up-regulate endogenous genes) libraries targeting all known transporters from the human genome were made to identify transporters of doxorubicin, one of the most commonly used chemotherapeutic drugs in the clinic. By directly monitoring the doxorubicin drug uptake as the screening readout, both previously confirmed drug exporters of doxorubicin such as ABCB1 and ABCG2 genes, and a novel doxorubicin importer SLC2A3 (GLUT3). The SLC2A3 gene could potentially serve as a new marker for doxorubicin response and for identifying subtypes of patients with tumors that could benefit from doxorubicin treatment.

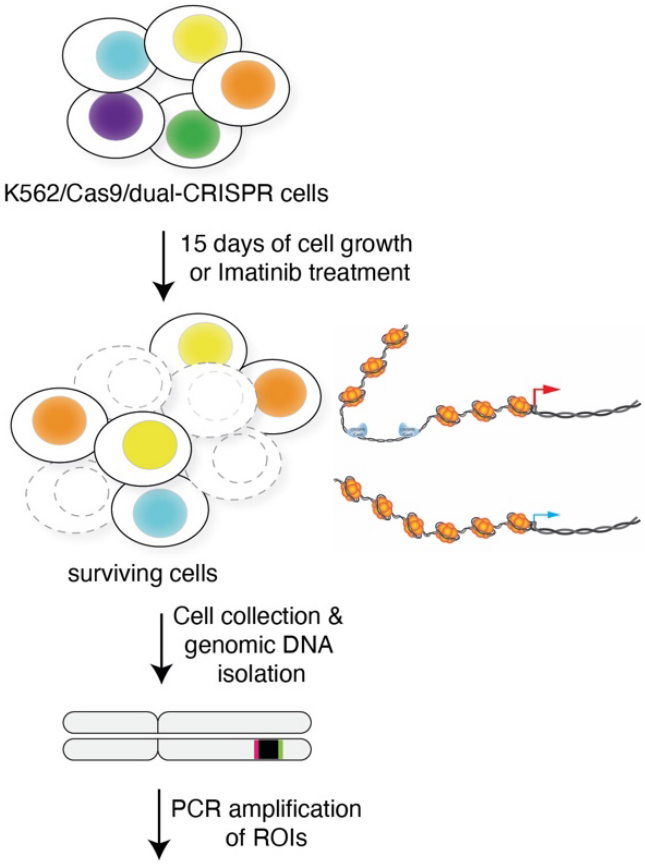

Figure 3: The dual-CRISPR libraries could be used to study different biological pathways.
A pair of guide RNAs within each individual cell will remove one target NCRE. Genomic DNA from different cell populations was then isolated, and the dual-CRISPR libraries could be amplified by direct PCR reactions, which would be ready for next-generation sequencing (NGS). The change in abundance of dual-guide-RNAs will then be calculated to identify potential hits according to the screening phenotype.

In **Chapter 3**, using a new dual-CRISPR screening system developed in our own group (Figure 3), we aim to understand the function of NCREs in cell survival, drug response, and cell differentiation. This system is able to introduce a designed pair of guide RNAs to remove an NCRE in the human genome in a single cell to study the function of the DNA fragment, and tests thousands of such NCREs in one screening. Custom-designed dual-CRISPR libraries targeting all ultraconserved elements in the human genome and all potential enhancer regions in K562 cells were made and used to study these NCREs in a systematic and genome-wide fashion for the first time. Multiple NCREs were identified and validated to affect cell survival or response to imatinib treatment in K562 cells. One ultraconserved element PAX6_Tarzan was also shown to affect cardiomyocyte differentiation using a human embryonic stem cell differentiation model, indicating a potential key function of this ultraconserved element in the evolution. The results from Chapter 3 also indicate that the dual-CRISPR system we developed could be a useful tool for the broad research community to study the functions of the non-coding genome.

References

1 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

2 Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304-1351 (2001).

3 Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44-53 (2022).

4 Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature Communications* **9**, 1911 (2018).

5 Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat Rev Mol Cell Biol* **17**, 5-15 (2016).

6 Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nature Genetics* **52**, 254-263 (2020).

7 Gregory, T. R. Synergy between sequence and size in Large-scale genomics. *Nature Reviews Genetics* **6**, 699 (2005).

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

9 Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics* **15**, 234 (2014).

10 Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5**, 1-1 (2012).

11 Pang, B., van Weerd, J. H., Hamoen, F. L. & Snyder, M. P. Identification of non-coding silencer elements and their regulation of gene expression. *Nature Reviews Molecular Cell Biology* **24**, 383-395 (2023).

12 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

13 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489** (2012).

14 Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).

15 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

16 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

17 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

18 Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074-1077 (2013).

19 van Arensbergen, J. *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotech* **35**, 145-153 (2017).

20 Cohen, S. N., Chang, A. C., Boyer, H. W. & Helling, R. B. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70**, 3240-3244 (1973).

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

22 Johannessen, C. M. *et al.* A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature* **504**, 138-142 (2013).
23 Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550-553 (2002).
24 Paul, P. *et al.* A Genome-wide Multidimensional RNAi Screen Reveals Pathways Controlling MHC Class II Antigen Presentation. *Cell* **145**, 268-283 (2011).
25 Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011).
26 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
27 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).
28 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature biotechnology* **34**, 192 (2016).
29 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).
30 Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509** (2014).
31 Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299-311 (2015).
32 Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32**, 347-355 (2014).
33 Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M. & Gerstein, M. B. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**, 559-571 (2010).
34 Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192-198 (2016).
35 Tycko, J. *et al.* Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat Commun* **10**, 4063 (2019).
36 Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology* **33**, 510-517 (2015).