Title

Understanding the connection between genotypes and phenotypes using linkage analysis and CRISPR genetic engineering

Author

Gou, Liangke

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Understanding the connection between genotypes and phenotypes using linkage analysis and CRISPR genetic engineering

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Human Genetics

by

Liangke Gou

2019

ABSTRACT OF THE DISSERTATION

Understand the connection between genotypes and phenotypes

using linkage analysis and CRISPR genetic engineering

by

Liangke Gou

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2019

Professor Leonid Kruglyak, Chair

The fundamental goal of genetics is to understand the functional effect of DNA sequence variations on a wide range of phenotypes, from basic biology to genetic diseases. Broadly, there are two major strategies to approach this goal: the first one is to find natural genetic variants underlying the trait of interest through linkage or association studies; the other is experimentally introducing genetic perturbations and assaying the effects of the perturbations in a high-throughput manner.

In this dissertation, both approaches were employed to understand the effect of genetic variants. Following the first approach, we used linkage analysis to find the genetic basis of mutation rate variation in yeast. We developed a high-throughput

fluctuation assay to enable quantification of spontaneous mutation rate in hundreds of yeast for the first time. We measured the mutation rate of 1040 yeast segregants from a cross between two diverge yeast strains, BY and RM. Combined with the genotype data, we performed linkage analysis in the segregants and identified four quantitative trait loci (QTLs) that contribute to the mutation rate variation in the cross. We fine-mapped two QTLs to the underlying causal genes, *RAD5* and *MKT1*, that contribute to mutation rate variation.

For the second approach, we developed three different systems to study the effect of natural variants using the genetic engineering tool CRISPR-Cas9. We constructed ten different CRISPR-Cas9 base editor systems for yeast, aiming to expand the targetable regions and the base converting types by using different base editors. We measured the efficiency of ten base editors in yeast from amplicon sequencing results at ten different sites along the genome and found one base editor that recognized the protospacer adjacent motif (PAM) site NGA with high efficiency. In addition to CRISPR base editor, we constructed a precise genome editing system with trackable genome integrated barcode using CRISPR-Cas9 with gRNA and donor DNA pairs. The integrated barcode enables precise tracking of edited strains with sequencing, ensuring robust downstream phenotyping. We also worked toward developing a CRISPR-directed mitotic recombination mapping panel in human cell lines to narrow down mapped out regions to causal genes by targeted creation of DNA double strand breaks along the chromosome.

The dissertation of Liangke Gou is approved.

Hilary Ann Coller

Bogdan Pasaniuc

Karen Reue

Leonid Kruglyak, Committee Chair

University of California, Los Angeles

2019

To Mom and Dad

and Zijun

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my unwavering gratitude to my thesis advisor, Leonid Kruglyak, who has been extremely supportive to my research and career development over the past five years. His critical thinking and scientific insight have been invaluable in cultivating me to become a quantitative scientist. His advice and support paved the way to my future career.

I owe a special thanks to my husband, Zijun Zhang, for his encouragement, support and love. He is always at my side and gave me endless support during all the hard times and he has devoted wholeheartedly to our family. I could only imagine how difficult it would be without him.

I cannot express enough appreciation for my committee members: Hilary Coller, Karen Reue and Bogdan Pasaniuc. They were incredibly supportive for my projects and had given me generous help and advice. I was lucky to have them on my committee and I owe them a debt of gratitude.

I would also like to thank both current and previous members of the Kruglyak lab: Joshua Bloom, Olga Schubert, Tzitziki Lemus Vergara, Matthieu Delcourt, Meru Sadhu, Alejandro Burga Ramos, James Boocock, Longhua Guo, Eyal Ben David, Oliver Brandenberg, Daniel Leighton, Frank Albert, Danny Zeevi, Laura Day, Kelly Tagami and Elise Pham, for helpful and productive intellectual discussions, and more importantly, for the colorful lives that they brought to my life as a Ph.D. student. I am particularly grateful

to Joshua Bloom, who took me under his wing during my early days in the Kruglyak lab. Josh's guidance has undoubtedly helped me develop into a better researcher.

Last but not least, I would like to thank my parents for their unconditioned love and endless support during my life. I could not imagine completing the Ph.D. journey without the support from my parents.

VITA

Education

2014-2019 Graduate Student Researcher, Department of Human Genetics, University of California, Los Angeles

2017 Teaching Assistant, LS4: Genetics University of California, Los Angeles

2016 Teaching Assistant, MCDB 104AL: Developmental Biology Lab University of California, Los Angeles

2010-2014 B.S. in Biological Sciences, College of Life Sciences, Sichuan University

Publications

Liangke Gou, Joshua S Bloom, Leonid Kruglyak (2019). The genetic basis of mutation rate variation in yeast. *Genetics*, 211 (2), 731-740.

Li-Chun Jiang**, Liang-Ke Gou,** Xin Zhang, Qing-Mei Zhao, Shuai Tan, Rui Peng, Yu-Qing Wei, Fang-Dong Zou (2015). Complete mitochondrial genome of a new subspecies of the blue sheep, Pseudois nayaur (Cetartiodactyla: Caprinae) from Helan Mountain in China. *Mitochondrial DNA*. 26 (5), 797-798.

Chapter 1 Introduction

The fundamental goal of genetics is to understand the effect of genetic variation on trait variation across a wide range of phenomena. Knowing how individual DNA polymorphisms combine to create diversity in phenotypes will advance our understanding of evolutionary process, as well as the mechanisms of genetic diseases [1]. There are two major strategies to approach this goal, one is to find the link between natural genetic variants and the trait of interest in a population. Traditionally the effects of natural genetic variation have been studied in this approach via quantitative trait locus (QTL) mapping or genome-wide association study (GWAS). In model organisms such as *Saccharomyces cerevisiae*, QTL mapping is performed within hundreds or thousands of offspring from a cross of one particular pair of parents [2]. The genomic regions that genotypes correlate with a phenotype of interest indicate the harboring of one or more causal variants. QTL mapping has successfully identified regions with causal genes for many traits [3]. In humans, QTL analyses are applied to family-pedigree based mapping, and GWAS is widely used to study complex traits, such as schizophrenia [4], using the genotype and phenotype data from a large cohort.

An alternative approach to study the link between genetic variants and phenotype variation is to experimentally introduce genetic perturbations into genomes and assay the consequences of the perturbations. One example of the genetic perturbation approach is the ‘deep mutational scanning’ approach, and this was used to study the functional consequences of possible single amino acid changes in a protein in yeast [5]. Until very recently, Zinc Finger Nuclease (ZFN) and Transcription-Activator

Like Effector Molecules (TALEN) were the two major techniques for targeted genome editing in mammalian systems [6]. With the development of the new genetic engineering tool CRISPR-Cas9, it is easier to engineer variants in a high-throughput manner. Sadhu *et al.* introduced premature stop codons into the yeast essential genes by genome-wide variant engineering using CRISPR-Cas9 with donor DNAs [7]. Sharon *et al.* measured the fitness consequences of thousands of natural genetic variants in yeast using CRISPR-Cas9-based high-throughput genome editing [8]. Different CRISPR/Cas9 editing systems are discovered in recent years, such as the CRISPR base editor and the Cpf1 proteins [9]. CRISPR-Cas9 based high-throughput genetic engineering tools with more robust downstream phenotyping will potentially allow rapid, base-pair level investigation of a wide range of traits and diseases.

In this dissertation, we tried to understand the effect of genetic variants on phenotypes via both approaches. We identified the underlying causal genes for spontaneous mutation rate variation in yeast using linkage analysis. We also worked toward developing three different library-based high-throughput CRISPR-Cas9 genetic engineering systems to study the effect of genetic variants via the experimental engineering approach.

In general QTL mapping approach, two or more strains of organisms that differ genetically with regard to the trait of interest are crossed to generate heterozygous individuals (F1), and these individuals are then crossed to generate F2. In model organisms like yeast, the F2 step was replaced by tetrads dissection during sporulation.

The phenotypes and genotypes of the derived (F2) population are scored using a wide variety of assays. The correlation between phenotype and genotype was calculated to identify QTL regions that are expected to contain one or more variants contributing to the phenotype. In **Chapter 2**, we developed a high-throughput fluctuation assay to quantify the spontaneous mutation rate of 1040 segregants from a cross between two diverse yeast strains: a lab strain BY, and a wild wine strain RM. Combined with the whole genome sequencing data of these segregants, we performed linkage analysis and identified four QTLs underlying mutation rate variation in the cross. We also fine-mapped two QTLs to the underlying causal genes *RAD5* and *MKT1*, that contribute to mutation rate variations.

An alternative method to understand the relationship between genotype and phenotype variation is to query the effect of variants directly from experiments. The promising avenue of generating precise mutations is genetic editing with the CRISPR-Cas9 system. One new genome-editing approach is called base editing, which can introduce precise point mutations into DNA without generating double stranded DNA breaks (DSBs). Base editors compromise a catalytically disabled Cas9 nuclease and a deaminase enzyme, which can introduce DNA base pair changes of C to T or A to G depending on the enzyme type. The directly converting of DNA bases into desired bases will largely prevent undesired editing byproducts from DSBs [10]. In **Chapter 3**, we constructed ten base editor systems for yeast, measured the editing efficiency and analyzed the editing patterns for those ten different base editors that could recognize different protospacer adjacent motif (PAM) sites. We found the NGG PAM base editors

BE3, BE4 and BE4-Gam had highest editing efficiencies at our tested sites. We identified one base editor that recognizes the NGA PAM site also had high editing efficiency, similar to the efficiency of BE3 in yeast. This base editor empowers us to edit 1.7-fold more regions in the genome using base editors compared to using NGG PAM site base editor BE3.

Another approach to generate precise variant edits in the genome is providing a donor DNA containing the mutation of interest following a Cas9-mediated double-strand break. However, this process is usually low-throughput and the editing efficiency varies at different targeting sites. To boost editing efficiency and enable high-throughput variants effect screens, we derived an approach that contained two gRNA and donor DNA repair template pairs in one plasmid. One pair of gRNA and donor DNA was designed to introduce the desired mutation, while the other pair of gRNA and donor DNA targeted gene *CAN1* to insert an integrated genomic barcode to the genome. The construction and efficiency analysis for this approach is discussed in **Chapter 4**. This plasmid construction method enables trackable genomic integrated cellular barcoding, thus ensuring robust phenotyping. Furthermore, this designed system can enrich functional CRISPR-Cas9 and homology-directed repair (HDR) by selecting cells resistant to canavanine, given the fact that cells repair the *CAN1* region cutting through HDR can block gene function and survive canavanine.

In addition to precise genome editing, an application of CRISPR-Cas9 system is to build mapping panels with targeted recombination events to narrow candidate regions for trait

of interest to causal genes or variants. Linkage and association studies rely on meiotic recombination events that break up linkage of genetic markers on the chromosome. However, the spatial resolution of genetic mapping is limited by the recombination rate. To address this problem, one method is to use CRISPR-Cas9 to create targeted recombination events by generating double strand break (DSB) and repairing with homologous recombination (HR). In heterozygous individual, cell division can create cells with a genotype that is completely homozygous from the DSB site to the telomere, which referred as loss of heterozygosity (LOH). **In Chapter 5**, we sought to generate this type of LOH mapping panel in human stem cell lines.

References

1. Kruglyak L. The road to genome-wide association studies. Nat Rev Genet. Nature Publishing Group; 2008;9: 314–318. doi:10.1038/nrg2316
2. Rockman M V., Kruglyak L. Genetics of global gene expression. Nature Reviews Genetics. 2006. pp. 862–872. doi:10.1038/nrg1964
3. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. Nature Publishing Group; 2013;494: 234–237. doi:10.1038/nature11867
4. Europe PMC Funders Group. Biological Insights From 108 Schizophrenia-Associated Genetic Loci. Nature. 2014;511: 421–427. doi:10.1038/nature13595.Biological
5. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. NIH Public Access; 2014;11: 801–7. doi:10.1038/nmeth.3027
6. Gupta D, Bhattacharjee O, Mandal D, Sen MK, Dey D, Dasgupta A, et al. CRISPR-Cas9 system: A new-fangled dawn in gene editing. Life Sci. 2019; 116636. doi:10.1016/j.lfs.2019.116636
7. Sadhu MJ, Bloom JS, Day L, Siegel JJ, Kosuri S, Kruglyak L. Highly parallel genome variant engineering with CRISPR-Cas9. Nat Genet. 2018;50: 510–514. doi:10.1038/s41588-018-0087-y
8. Sharon E, Chen SAA, Khosla NM, Smith JD, Pritchard JK, Fraser HB. Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell. 2018;175: 544–557.e16. doi:10.1016/j.cell.2018.08.057
9. System CC, Zetsche B, Gootenberg JS, Abudayyeh OO, Regev A, Koonin E V.,

et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015; 1–13. doi:10.1016/j.cell.2015.09.038

10. Rees HA, Liu DR. Base editing: precision chemistry on the genome and transcriptome of living cells. Nature Reviews Genetics. 2018. doi:10.1038/s41576-018-0059-1

Chapter 2 The genetic basis of mutation rate variation in yeast

Abstract

Mutations are the root source of genetic variation and underlie the process of evolution. Although the rates at which mutations occur vary considerably between species, little is known about differences within species, or the genetic and molecular basis of these differences. Here we leveraged the power of the yeast *Saccharomyces cerevisiae* as a model system to uncover natural genetic variants that underlie variation in mutation rate. We developed a high-throughput fluctuation assay and used it to quantify mutation rates in 7 natural yeast isolates and in 1040 segregant progeny from a cross between BY, a lab strain, and RM, a wine strain. We observed that mutation rate varies among yeast strains and is heritable (H^2=0.49). We performed linkage mapping in the segregants and identified four quantitative trait loci (QTLs) underlying mutation rate variation in the cross. We fine-mapped two QTLs to the underlying causal genes, *RAD5* and *MKT1*, that contribute to mutation rate variation. These genes also underlie sensitivity to the DNA damaging agents 4NQO and MMS, suggesting a connection between spontaneous mutation rate and mutagen sensitivity.

Introduction

Mutations are permanent changes to the genome of an organism that can result from DNA damage that is improperly repaired, from errors in DNA replication [1], or from

the movement of mobile genetic elements. Mutations give rise to genetic variants in populations and are the wellspring of evolution [2]. Mutations also play a major role in both inherited diseases and acquired diseases such as cancer [3].

The mutation rate can be defined as the number of mutational events per cell division, generation, or unit of time [4]. Mutation rates tend to be approximately 10^{-9} to 10^{-10} mutations per base pair, per cell division, for most microbial species [5], making them difficult to measure and compare across individuals. As a consequence, the effects of genetic background differences on mutation rates have only been investigated on a small scale [6]. Two types of experimental approaches have been used to measure mutation rates in yeast. The first is the fluctuation assay [7]. This method requires a gene with a selectable phenotype such that loss-of-function mutations in the gene enable the mutants to grow in the corresponding selective conditions. Spontaneous mutation rate is then estimated from the distribution of mutant numbers in parallel cultures. Lang and Murray applied the fluctuation assay to *S. cerevisiae* and estimated the per-base-pair mutation rate in yeast [8]. A second method tracks mutation accumulation during experimental evolution and uses whole-genome sequencing to estimate mutation rates [9]. This approach also provides information on the number, locations and types of spontaneous mutations. However, this assay requires growing the mutation accumulation lines over hundreds of generations, as well as sequencing many genomes. Although the fluctuation assay is faster and cheaper, the need for many parallel cultures makes it laborious to extend it to many different strains.

Here we developed a modified version of the fluctuation assay to enable higher-throughput measurements of spontaneous mutation rates. We used the new assay to quantify mutation rates across genetically distinct yeast strains and observed considerable variation. To find the genes underlying the observed variation, we applied the modified fluctuation assay to a large panel of 1040 segregants from a cross between the laboratory strain BY4724 (hereafter referred to as BY) and the vineyard strain RM11-1a (hereafter referred to as RM). We identified four loci associated with mutation rate variation and narrowed the two loci that contributed the most to mutation rate variation to missense variants in the genes *RAD5* and *MKT1*. We also found interactions between alleles of *RAD5* and *MKT1*.

Materials and Methods

Yeast strains and media

Seven natural *S. cerevisiae* strains (Table Supplementary 2.3) were used in this study. The 1040 segregants derived from BY4724 (MATa) and RM11-1a (MATa, *MKT1*-BY, hoΔ::HphMX, flo8Δ::NatMX) were generated, genotyped and described previously [10]. The RM::*MKT1*-BY strain was made previously by our lab. The BY::*RAD5*-RM strain and the *RAD5* variants substitution strains (Table 2.1) were from Demogines et al [11]. For fluctuation assay, yeast was grown in synthetic complete liquid medium without arginine (SC-Arg) before plating onto selective plates. For DNA damaging agents sensitivity assays, yeast were grown in rich YPD medium (1% yeast extract, 2% peptone and 2% glucose) before plating onto YPD agar plates with DNA damaging

agents. SC-Arg and YPD liquid media and agar plates were made according to Amberg et al [12].

Selection agar plate construction

Selective canavanine plates were made from arginine minus synthetic complete agar medium with 60mg/liter L-canavanine (Sigma C1625). The canavanine plates were dried by incubating at 30ºC overnight. Selective plates for the DNA damaging agents sensitivity assay were made with YPD agar medium containing the respective agents at the concentrations indicated in Table 2.2. 50ml of the agar medium was poured into each Nunc OmniTray plates (Thermo Scientific 264728) and placed on a flat surface to solidify. Each experiment was performed with the same batch of selection plates. The concentrations for 4NQO (Sigma N8141), MMS (Sigma 64382) and H_2O_2 (Sigma 216763) were 0.1µg/ml, 0.01% and 4mM. These concentrations capture the sensitivity difference between the segregants, while maintaining enough colony growth for QTL mapping.

Fluctuation assays

To begin the fluctuation assay, yeast was grown in synthetic complete medium without arginine (SC-Arg) in 96-well plates (Costar 3370) for ~48 hours to saturation. Saturated cultures were diluted and pinned into a new 96-well plate with liquid SC-Arg medium. This step ensured a small number of ~1000 yeast cells in the initial inoculum. Plates were sealed with a Breath-Easy sealing membrane (Sigma Z380059) to prevent

evaporation and incubated at 30ºC with shaking for ~48 hours. 100µl saturated cultures were spot-plated onto canavanine plates in a four by six configuration using a Biomek FX^P automated workstation. Plates with spot-plated yeast culture were dried in the laminar flow hood (Nuair) for half an hour or until dry, and incubated at 30ºC for ~48 hours. We imaged the plates using an imaging robot (S&P Robotics BM3-SC), and the number of colonies in each spot was manually counted from the images. An example of the imaged plate is shown in Supplementary Figure 2.5.

Mutation rate was estimated using the Ma-Sandri-Sarkar Maximum Likelihood Method where the numbers of observed colonies on canavanine plates was fitted into the Luria-Delbrück distribution and a single parameter m was calculated [13]. The parameter m represents the expected number of mutation events per culture. For the natural isolates and engineered strains, the mutation rate was calculated from the equation $\mu = m/N$, where *N* is the average number of cells per culture (as a proxy for the number of cell divisions given the starting inoculum is much smaller than *N*). In the segregant panel, we defined a mutation rate score that was calculated as the residual phenotype after regressing out the effect of average number of cells per strain (N) from the estimate of m per strain across all of the segregants.

For each of the seven natural isolate strains, we performed ninety-six replicates of the fluctuation assay, which means we had ninety-six estimations of mutation rate. In each replicate three cultures were plated onto canavanine plates, and the number of resistance colonies in these three plates were fitted into the Luria-Delbrück distribution

to estimate the mutation events per culture (m). One culture was diluted and plated onto YPD to determine the number of cells per culture (N) in each replicate. Given the number of replicates used for estimate m and N were limited, the mutation rate estimation for the seven natural isolate strains had large variance. For the BYxRM segregants panel, twelve independent replicate cultures were plated onto canavanine plates for every segregant. The number of canavanine resistant colonies in these twelve plates was fitted into the Luria-Delbrück distribution to calculate the number of mutations per culture (m), and one culture was diluted and plated on the YPD plates to determine the number of cells (N). Given only one culture was used to estimate the number of cells (N) for each segregants, which means the mutation rate estimation from the equation $\mu = m/N$ would be largely affected by that one measure of N. The number of cells per culture (N) was measured by the number of cells growing on an YPD non-selective plate after a 10^5 dilution. The dilution scale was picked from pilot experiments where different dilution scales were tested. This 10^5 dilution was applied to all segregants before plating on YPD to get the measure of N. However, this degree of dilution was not suitable for all individual segregants. Any segregants that have too many colonies (>70) or have no colony (<1) growing on the YPD plates cannot give us an accurate measure of N. Thus we remove these 197 segregants from the downstream QTL analysis: only 843 segregants with confident measure of N were used. To minimize the noise driven by the measure of N, we defined a mutation rate score that regressed out the effect of N instead of the division. We computed a linear model that included the number of cells per culture (N) and the plate effect as additive covariates on the number of mutations per culture (m). The residuals from this linear

model were called the ‘mutation rate score’ and used in downstream analyses. We used the mutation rate score of segregants for later QTL mapping. For each allele replacement strain (Table 2.1), ninety-six replicates of fluctuation analysis were performed, providing us ninety-six estimations of mutation rate. In each replicate, twelve cultures were plated onto canavanine plate to estimate the number of mutations per culture (m), and three cultures were pooled, diluted and plated on YPD plates to determine the number of cells per culture (N).

Yeast growth measurement for DNA damaging agents sensitivity assay

The segregant panel were originally stored in 96-well plates (Costar 3370). During the DNA damaging agents sensitivity assay, individual segregants were inoculated in two plate configurations in 384-well plates (Thermo Scientific 264574) with YPD and grown for ~48 hours in a 30ºC incubator without shaking. Saturated cultures were mixed for 1min at 2,000 r.p.m. using a MixMate (Eppendorf) before pinning. The colony handling robot (S&P Robotics BM3-SC) was used to pin segregants onto selective agar plates with 384 long pins. The plates were incubated at 30ºC for ~48 hours and imaged by the colony handling robot (S&P Robotics BM3-SC). Custom R code [10] was used to determine the size of each colony and the size was used as a proxy for growth in the presence of the DNA damaging agents.

QTL mapping and detecting QTL-QTL interactions

In order to control for intrinsic growth rate differences and plate position effects, we normalized the traits for growth by fitting a regression for growth of the yeast that were in the same layout configuration on control plate (YPD agar plates for mutagen sensitivity assay). Residuals were used for QTL mapping. We tested for linkage by calculating logarithm likelihood ratio (LOD scores) for each genotypic marker and trait as $-n(\ln(1-r^2)/(2\ln(10))$, where r is the Pearson correlation coefficient between the segregant genotypes and the segregant mutation rate or DNA damaging agents sensitivity. The threshold declaring the significant QTL effect was calculated from the empirical null distribution of the maximum LOD score determined from 1,000 permutations [14]. The estimated 5% family-wise error rate significance thresholds were 3.52, 3.62, 3.61 and 3.64 for mutation rate, mutagen sensitivity for 4NQO, MMS and H_2O_2 respectively. The 95% confidence intervals were determined using a 1.5 LOD score drop. The code and the data for QTL mapping is available at https://github.com/gouliangke/Mutation-rate/tree/master/qtl_mapping

We tested for interactions between each QTL by comparing a model that includes an interaction term of two QTLs, $y = ax + bz + cxz + d$, with a model that does not, $y = ax + bz + d$ using the add1 function in R and calculating an F-statistic. Here y is residuals vector for mutation rate score after fitting the additive QTL model, x is the genotype vector at one QTL in the genome, z is the genotype vector at another QTL in the genome, and a, b, c and d are estimated parameters specific to each marker pair.

Calculating heritability

Broad-sense heritability was calculated using the natural isolate data and a random effect analysis of variance. The variance structure of the phenotype is $V = {\sigma^2}_G \, ZZ' + {\sigma^2}_E I_m$, where Z is an incidence matrix mapping phenotypes to strain identity and I_m is the identity matrix. The broad-sense heritability was estimated as $\frac{{\sigma^2}_G}{{\sigma^2}_G + {\sigma^2}_E}$, where the ${\sigma^2}_G$ is the genetic variance due to genetic difference and ${\sigma^2}_E$ is the error variance. Standard errors of variance component estimates were calculated as the square root of the diagonal of the Fisher information matrix from the iteration at convergence of the AI-REML algorithm [15].

Amplicon sequencing of the CAN1 region in segregants

1040 segregants were assigned into four groups according to their alleles at gene *RAD5* and *MKT1* (Supplementary Table 2.4). We collected the cananvanine resistant colonies from the canavanine plates that we used to measure the mutation rate of segregants in the previous fluctuation analysis. A single canavanine resistant colony (if any) was picked from each segregant and the picked colonies from the same group were pooled together for DNA extraction. DNA was extracted from the pool using the Qiagen DNeasy Blood & Tissue Kit. The *CAN1* region was amplified from the DNA of four groups using the Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) and eight pairs of designed primers. The amplicon sequencing library was prepared using the Illumina Nextera DNA Library Prep Kit with the adjusted protocols to skip the nextera treatment. The library was then sequenced on the MiSeq platform

using the MiSeq Reagent V2 Nano Kit. As shown in Supplementary Table 2.5, the original aligned read counts for each library varies largely. In order to eliminate the bias caused by the read counts variation, we adjusted the number read counts for each sample by randomly down-sampling the reads in the fastq files to be the same for all samples using custom Python codes. Then we processed reads with the following pipeline: read pairing (PEAE), read trimming (Trimmomatic-0.36), and read alignment (BWA) to the reference-targeting region using the down-sampled fastq files. The adjusted read counts for each sample is shown in the Supplementary Table 2.5. Custom R codes were used to detect the mutation rate spectrum of each group. The code for the mutation rate spectrum analysis is available at https://github.com/gouliangke/Mutation-rate/tree/master/mutation_spectrum

Results

High-throughput fluctuation assay for measuring mutation rates

The fluctuation assay for measuring mutation rate involves growing many parallel cultures, each starting from a small number of cells, under non-selective conditions, followed by plating to selective medium to identify mutants. The number of mutations that occurs in each culture should follow the Poisson distribution, as mutations arise spontaneously. However, the number of mutant cells that survive on the selective plates can vary greatly, because early mutations are inherited by all offspring of the mutant. This leads to the “jackpot” effect, in which some cultures contain a large number of mutant individuals. The number of observed mutant cells per culture follows the Luria-Delbrück distribution [7], and the Ma-Sandri-Sarkar maximum likelihood method can be

used to estimate the expected number of mutations per culture from the observed numbers of mutants [13]. The underlying mutation rate is then calculated by dividing the number of mutations per culture by the average number of cells per culture [1,4]. Here we measured rare spontaneous loss-of-function mutations in the gene *CAN1*, which encodes an arginine permease. Yeast cells carrying loss-of-function mutations in *CAN1* can grow on canavanine, an otherwise toxic arginine analog. Typically, fluctuation assays are labor-intensive and have limited throughput, because a large number of parallel cultures is required for estimating the mutation rate in each assay, and several replicate assays are needed for a robust measurement of the mutation rate in each strain [16]. We modified the fluctuation assay into a high-throughput method for measuring mutation rates in many strains in parallel. We grew cultures in 96-well plates, automated the spotting of cultures, and used a plate-imaging robot to capture images of the mutant colonies on plates (Methods, Figure 2.1A). The automated spotting process for 96 strains took only approximately twenty minutes, and the imaging process required even less time. These improvements enabled us to measure the spontaneous mutation rates in the hundreds of strains necessary for genetic mapping.

Spontaneous mutation rate varies among yeast isolates

To investigate mutation rate variation among *S. cerevisiae* strains, we measured the spontaneous mutation rate of seven yeast isolates using the high-throughput fluctuation assay (Supplementary Table 2.3). The seven strains span a large range of yeast genetic diversity [17]. We found that the mutation rates of these strains range from 1.1×10^{-7} to 5.8×10^{-7} mutations per gene per generation, with a median of 1.7×10^{-7}

(Supplementary Table 2.3, Supplementary Figure 2.6). The median mutation rate was very similar to the previously reported mutation rate at *CAN1* [8]. In particular, the mutation rate we observed for the BY strain (1.7×10^{-7}) is very similar to the previously reported rate, which was measured in strain W303 (1.5×10^{-7}) [8], consistent with the fact that W303 shares a large fraction of its genome with BY [18,19]. An analysis of variance (ANOVA) showed that strain identity explained a significant fraction of the observed variance in mutation rates ($F=69.9$, $df=6$, $p<2\times10^{-16}$) (Supplementary Figure 2.6). The fraction of total variance in mutation rates explained by the repeatability of measurements for each strain, 49% (SE=0.29), serves as an upper bound for the estimate of the total contribution of genetic differences between strains to trait variation (broad-sense heritability or H^2). We observed that RM, a vineyard strain, had a mutation rate higher than all other strains (Supplementary Figure 2.6).

Four QTLs explain the majority of observed mutation rate variation

In order to find the genetic factors underlying the difference in mutation rate between BY and RM, we performed quantitative trait locus (QTL) mapping in 1040 genotyped haploid segregants from a cross between these strains [10]. We measured the mutation rate of each segregant using the high-throughput fluctuation assay (Methods). We estimated the fraction of phenotypic variance explained by the additive effects of all segregating markers (narrow-sense heritability) to be 30% (Methods) [20]. This sets an upper bound for the expectation of the total amount of additive genetic variance that could be explained with a QTL-based model. QTL mapping in the segregant panel identified significant linkage at four distinct loci (Figure 2.1B). At two of

the QTLs, on chromosomes XII and I, the RM allele conferred a higher mutation rate, consistent with the higher mutation rate of this strain. At the other two QTLs, on chromosomes XIV and V, the BY allele conferred a higher mutation rate (Supplementary Figure 2.7), showing that a strain with lower trait value can nevertheless harbor trait-increasing alleles. The four detected QTLs explained 20.7% of the phenotypic variance, thus accounting for 69% of the estimated additive heritability. The loci on chromosomes XII, XIV, I and V explained 8.8%, 6.1%, 3.1% and 2.6% of the variance, respectively. We tested the four identified QTLs for pairwise interactions and found a significant interaction between the QTL on chromosome XII and the QTL on chromosome XIV that explained 1% of the phenotypic variance (F=8.41, df=1, Bonferroni-corrected p=0.023).

Polymorphisms in genes RAD5 and MKT1 underlie the major QTLs on chromosomes XII and XIV

Ten genes fell within the confidence interval of the QTL on chromosome XII. A strong candidate was *RAD5*, based on previous studies which showed that natural variants in *RAD5* contribute to sensitivity to the mutagen 4-nitroquinoline 1-oxide (4NQO) [11]. *RAD5* encodes a DNA repair protein involved in the error-free DNA damage tolerance (DDT) pathway [21,22]. The DDT pathway promotes the bypass of single-stranded DNA lesions encountered by DNA polymerases during DNA replication, thus preventing the stalling of DNA replication [23]. *RAD5* plays a crucial role in one branch of the DDT pathway called template switching (TS), in which the stalled nascent strand

switches from the damaged template to the undamaged newly synthesized sister strand for extension past the lesion [23]. Two non-synonymous substitutions exist between BY and RM strains in *RAD5* (Figure 2.2A), at amino acid positions 783 (glutamic acid in BY and aspartic acid in RM) and 791 (isoleucine in BY and serine in RM). According to Pfam alignments [24], *RAD5* contains a HIRAN domain, an SNF2-related N-terminal domain, a RING-type zinc finger domain, and a helicase C-terminal domain (Figure 2.2A). Both non-synonymous polymorphisms mapped to the helicase domain of *RAD5* (Figure 2.2A), and no other sequenced strains of *S. cerevisiae* in the 1002 Yeast Genomes Project contain the aspartic acid 783 and serine 791 variants that are private to the RM strain [25]. We used protein variation effect analyzer (PROVEAN) [26] to predict whether the two non-synonymous substitutions have an impact on the biological function of the protein. PROVEAN showed the I791S substitution (score –5.4) might have a strong deleterious effect, while the E783D variant (score –1.8) was not predicted to have a strong effect.

Nineteen genes fell within the confidence interval of the QTL on chromosome XIV. A strong candidate was *MKT1*, which was also reported to affect 4NQO sensitivity [11]. *MKT1* encodes an RNA-binding protein that affects multiple traits and underlies an eQTL hotspot in yeast [27]. The RM allele of *MKT1* increases sporulation rate [28] and improves survival at high temperature [29], in low glucose [30], after exposure to DNA-damaging agents [11], and in high ethanol levels [31]. The coding region of the BY and RM alleles of *MKT1* differs by one synonymous polymorphism and two non-synonymous substitutions. *MKT1* has an XPG domain, which is relevant to DNA repair,

and an *MKT1* domain, which is related to the maintenance of K2 killer toxin [32]. One non-synonymous variant is in the XPG domain at amino acid position 30 (aspartic acid in BY and glycine in RM), while the other non-synonymous variant is in the *MKT1* domain at position 453 (lysine in BY and arginine in RM). PROVEAN predicted a large effect of the D30G variant (score 6.7) on the function of *MKT1*, and this variant was previously found to influence sporulation rate [28], mitochondrial genome stability [33] and survival at high temperature [30]. The other variant (K453R) was not predicted to have a strong effect (score 0.6).

We tested whether *RAD5* and *MKT1* alleles caused differences in mutation rate by using the fluctuation test on allele replacement strains [11,34] (Table 2.1). The BY strain carrying the RM allele of *RAD5* (BY::*RAD5*-RM) had a higher mutation rate than the BY strain (permutation t-test, mean difference=2.9×10^{-7}, $p<1 \times 10^{-4}$), demonstrating that the RM *RAD5* allele increases mutation rate (Figure 2.3A). This result is consistent with the observed difference between segregants grouped by parental allele at *RAD5* (mean difference=2.3×10^{-7}). The RM strain carrying the BY allele of *MKT1* (RM::*MKT1*-BY) had a higher mutation rate than the RM strain (permutation t-test, mean difference=6.1×10^{-7}, $p<1 \times 10^{-4}$), showing that the BY *MKT1* allele increases mutation rate (Figure 2.3A), consistent with the direction of effect observed in the segregants.

To gain a finer-level understanding of the two missense variants between BY and RM in the gene *RAD5,* we tested strains [11] in which these sites in BY were individually replaced with the RM alleles (Table 2.1) by site-directed mutagenesis. Strains with

either variant had a higher mutation rate than BY (permutation t-test, mean difference=$0.9x10^{-7}$, $p<1x10^{-4}$ for BY::*RAD5*-I791S; mean difference=$0.3x10^{-7}$, $p=6x10^{-4}$ for BY::*RAD5*-E783D) (Figure 2.2B), suggesting that both variants contribute to the higher mutation rate. The BY strain with the I791S substitution had a higher mutation rate than the BY strain with the E783D substitution (permutation t-test, mean difference=$0.6x10^{-7}$, $p<1x10^{-4}$) (Figure 2.2B), consistent with the PROVEAN prediction of a stronger effect for the I791S variant. However, neither variant alone nor the additive effect of the two variants fully recapitulated the increase in mutation rate that we observed when replacing the entire coding region of *RAD5* in BY with the RM allele ($F=67.6$, $df=1$, $p=3.3x10^{-15}$), suggesting an interaction between the two variants.

Mutation rate shares two large effect QTLs with growth on DNA damaging agents 4NQO and MMS

Deficiencies in DNA repair can increase mutation rate [35,36] and increase sensitivity to DNA damaging agents such as alkylating compounds and UV light [37,38]. We hypothesized that genetic variants that cause deficiencies in DNA repair may underlie QTLs for both mutation rate variation and sensitivity to DNA damaging agents. Previously, Demogines *et al.* identified a large-effect QTL on chromosome XII for MMS and 4NQO sensitivity in a panel of 123 segregants from a cross between BY and RM [11]. Additionally, they identified a QTL on chromosome XIV for 4NQO sensitivity by using backcrossing and bulk segregant analysis. These QTLs overlapped with the major QTLs that we identified for mutation rate variation, and the underlying causal genes for 4NQO sensitivity were also *RAD5* and *MKT1*.

To follow up on these results, we measured sensitivity to three different DNA damaging agents in our panel of 1040 segregants (Table 2.2). The compounds assayed included methyl methanesulfonate (MMS), an alkylating agent that induces DNA double strand breaks and stalls replication forks [39], 4NQO, an ultraviolet light mimetic agent [39] and hydrogen peroxide (H_2O_2), a compound that induces DNA single and double strand breaks [39]. We observed that segregants with higher mutation rate, and presumably less efficient DNA repair systems, were more sensitive to MMS, 4NQO and H_2O_2 (Supplementary Figure 2.8), consistent with our hypothesis that deficiencies in DNA repair increase the rate of spontaneous mutations and increase sensitivity to DNA damaging agents. We identified two large-effect QTLs for 4NQO and MMS sensitivity that overlapped with the major QTLs for mutation rate (Figure 2.4A and B). The QTLs on chromosome XII and XIV were still observed in the linkage mapping for H_2O_2, but they had small effects (Supplementary Figure 2.9). The large effect QTLs detected for H_2O_2 sensitivity on other chromosomes likely reflects trait-specific effects of variants acting on sensitivity to H_2O_2 (Supplementary Figure 2.9).

Similar mutation spectra in segregants with different RAD5 and MKT1 genotypes

In order to gain a better understanding of how genetic variation in *RAD5* and *MKT1* might influence the DNA damage repair process, we characterized the spontaneous mutation spectrum at *CAN1* in the segregants. We divided 1040 segregants into four groups based on their genotypes at *RAD5* and *MKT1* and

sequenced pools of *CAN1*-resistant mutants from each group (Supplementary Table 2.4; Supplementary Table 2.5). The mutation spectra of the four groups are shown in Supplementary Table 2.6 and Supplementary Figure 2.10. C:G > T:A transitions were the most frequently observed mutations. A:T > G:C was the rarest transition, and A:T > T:A was the rarest transversion. The spectra for single base pair substitutions observed here (Supplementary Figure 2.11) are similar to previous observations based on whole-genome sequencing of mutation accumulation strains [9]. While there were some differences in the relative frequencies of specific mutation types (for instance, more C:G > G:C transversions in segregants with the RM *MKT1* allele and more A:T > C:G transversions in segregants with BY *RAD5* allele), these mutation differences were not statistically significant after correction for multiple testing.

Discussion

We developed and implemented a high-throughput fluctuation assay to directly measure mutation rates in yeast. We used this assay to map four QTLs that influence differences in the spontaneous mutation rate, and narrowed the two QTLs with the largest effects to causal genes and variants. We attempted to gain insight into how these variants might affect the mutation rate by comparing mutational spectra of segregants grouped by genotype, but the differences we observed did not reach statistical significance.

We identified *RAD5* as the gene underlying the QTL with the largest effect on mutation rate. *RAD5* encodes a DNA helicase and ubiquitin ligase involved in error-free

DNA damage tolerance (DDT), a pathway that facilitates chromosome replication through DNA lesions [40,41]. Previous work showed that Rad5 is a structure-specific DNA helicase that is able to carry out replication fork regression [21], a process of remodeling the replication fork into four-way junctions when replication perturbations are encountered [42]. This process was hypothesized to promote DNA damage tolerance and repair during replication [42]. We showed that two non-synonymous variants between BY and RM in the helicase domain affect mutation rate. The RM allele of *RAD5* increases the sensitivity of yeast to 4NQO and MMS [43], probably due to a defect in replication fork regression. Thus the RM allele of *RAD5* causes both decreased growth in mutagenic conditions and a higher mutation rate in non-stressful normal conditions.

We furthermore showed that polymorphisms in *MKT1* contribute to mutation rate variation. *MKT1* is a highly pleiotropic gene that has been shown to affect levels of transcript and protein abundance for numerous genes [34] [44], as well as numerous cellular phenotypes [11,27–31,45]. The BY and RM alleles of *MKT1* differ by two non-synonymous substitutions. One variant (K453R) is located in the MKT1 domain, which is required for activity of the Mkt1 protein in maintaining K2 killer toxin [46]. Another variant (D30G) localizes to the XPG-N (the N-terminus of XPG) domain. Four other yeast proteins contain this domain: Exo1, Din7, Rad27 and Rad2. All of these proteins have functions related to DNA repair and cellular response to DNA damage, including DNA double-strand break repair (Exo1) [47], DNA mismatch repair (Exo1, Din7) [48,49], nucleotide excision repair (Rad2) [50], ribonucleotide excision repair (Rad27) [51] and

large loop repair (LLR) (Rad27) [52]. The internal XPG (XPG-I) domain, together with XPG-N, forms the catalytic domain of the Xeroderma Pigmentosum Complementation Group G (XPG) protein. The XPG protein has well-established catalytic and structural roles in nucleotide excision repair, a DNA repair process, and acts as a cofactor for a DNA glycosylase that removes oxidized pyrimidines from DNA [53]. In humans, mutations in the XPG protein commonly cause Xeroderma Pigmentosum, which often leads to skin cancer [54]. We hypothesize that Mkt1 has a previously unknown function in DNA damage repair, mediated through its XPG domain.

We found that variants in *RAD5* and *MKT1* contribute to both mutation rate variation and mutagen sensitivity. These results suggest that spontaneously occurring mutations may have a similar mutation spectrum to those created by 4NQO and MMS, and are potentially repaired by the same mechanisms. Deficient DNA repair can lead to increased sensitivity to agents such as alkylating compounds and UV light [37,38,55] and to higher mutation rates at sites that are less accessible to the DNA repair system [35]. Because mutation rates can be difficult to measure, sensitivity to mutagens may serve as a useful proxy.

Recently, Jerison et al. reported heritable differences in adaptability in 230 yeast segregants from the same cross we studied here [56]. They measured adaptability as the difference in fitness between a given segregant (‘founder’) and a descendant of that founder after 500 generations of experimental evolution. Interestingly, *RAD5* fell within one of the QTLs found to influence adaptability. Together with our observation that

RAD5 influences mutation rate, this finding suggests that differences in mutation rate can affect the adaptability of organisms.

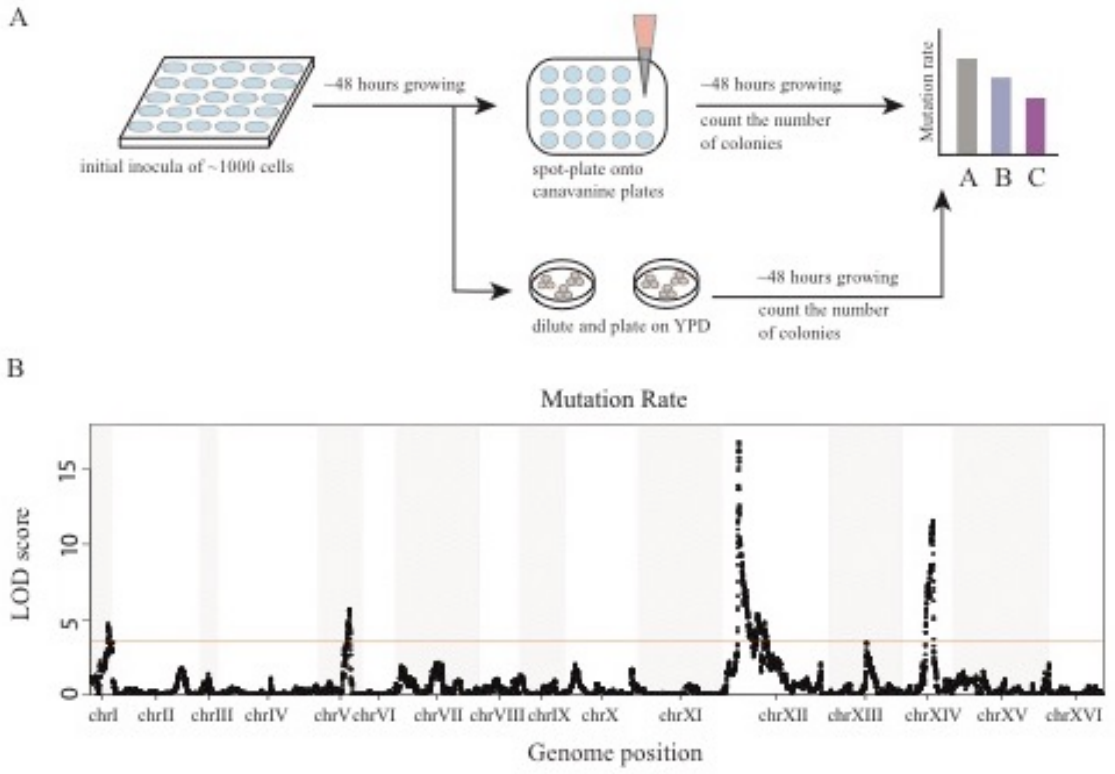


Figure 2.1 Linkage analysis identified four loci underlying mutation rate variation.

(A) The fluctuation assay was performed as shown in the workflow. The assay started with a small number of cells growing in 96-well plates in liquid SC-Arg medium for ~48 hours, followed by plating onto selective agar plates with canavanine. A proportion of the cultures were diluted to measure the number of cells per culture (Methods). Plates were imaged two days after spot-plating, and the number of colonies on canavanine plate was counted. (B) LOD score for mutation rate variation is plotted against the genetic map. The 4 significant QTLs explain 20.7% of the phenotypic variance. The red line indicates a 5% FWER significance threshold (LOD =3.52).

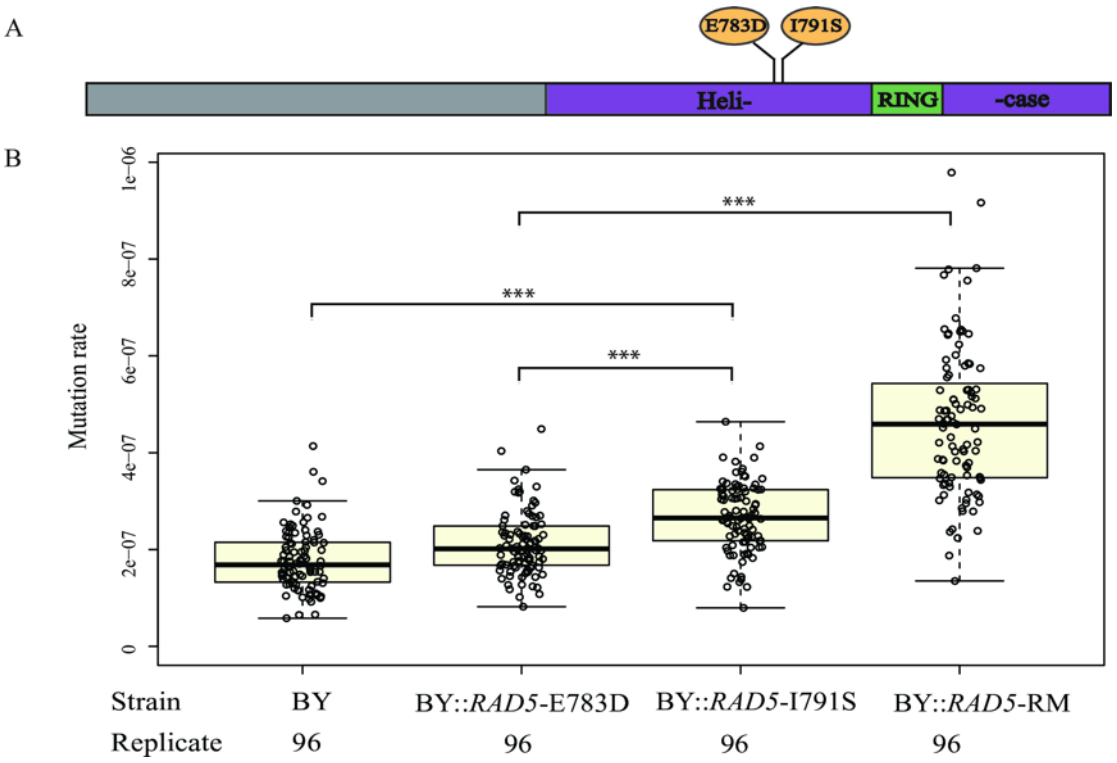


Figure 2.2 Polymorphisms in *RAD5* underlie mutation rate variation.

(A) *RAD5* polymorphisms between BY and RM are located in the helicase region. The first letter for each polymorphism indicates the BY polymorphisms (E783, I791) and the second letter indicates the RM polymorphisms (D783, S791). (B) The effect of single *RAD5* polymorphism and *RAD5* whole gene replacement was tested in the BY strain background for mutation rate. For each strain, the mutation rates of ninety-six replicates were measured. Bold lines show the mean. Boxes show the interquartile range. Statistical significance was tested using a permutation t-test. Permutation p value < 0.001 is shown as ***.

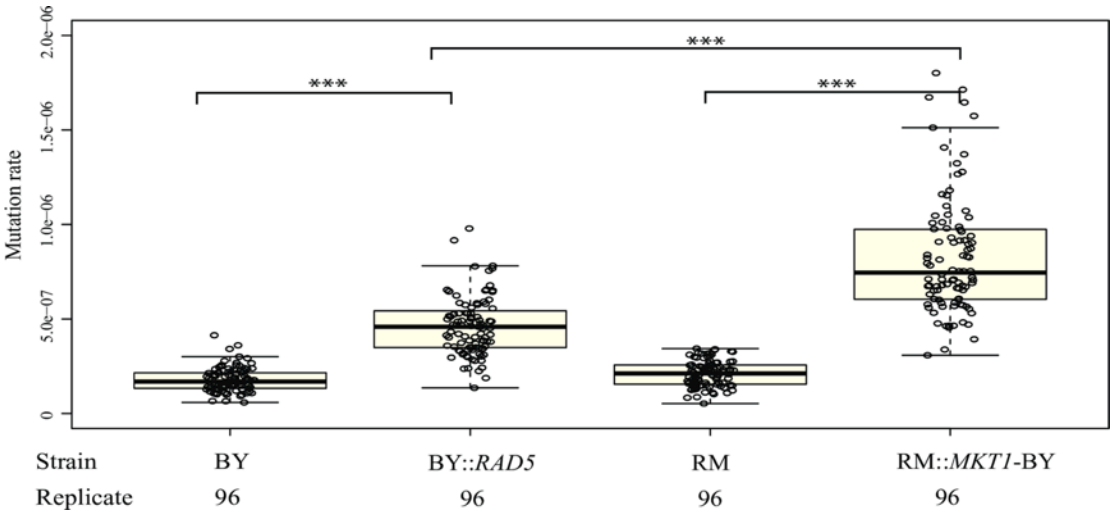


Figure 2.3 The RM allele of *RAD5* and BY allele of *MKT1* increase mutation rate.

The mutation rate of two allele replacement strains, the BY strain and the RM strain are measured and compared. For each strain, ninety-six replicate measurement for mutation rate was performed. Bold lines show the mean. Boxes show the interquartile range. Statistical significance was tested using permutation t-test. Permutation p value < 0.001 is shown as ***.

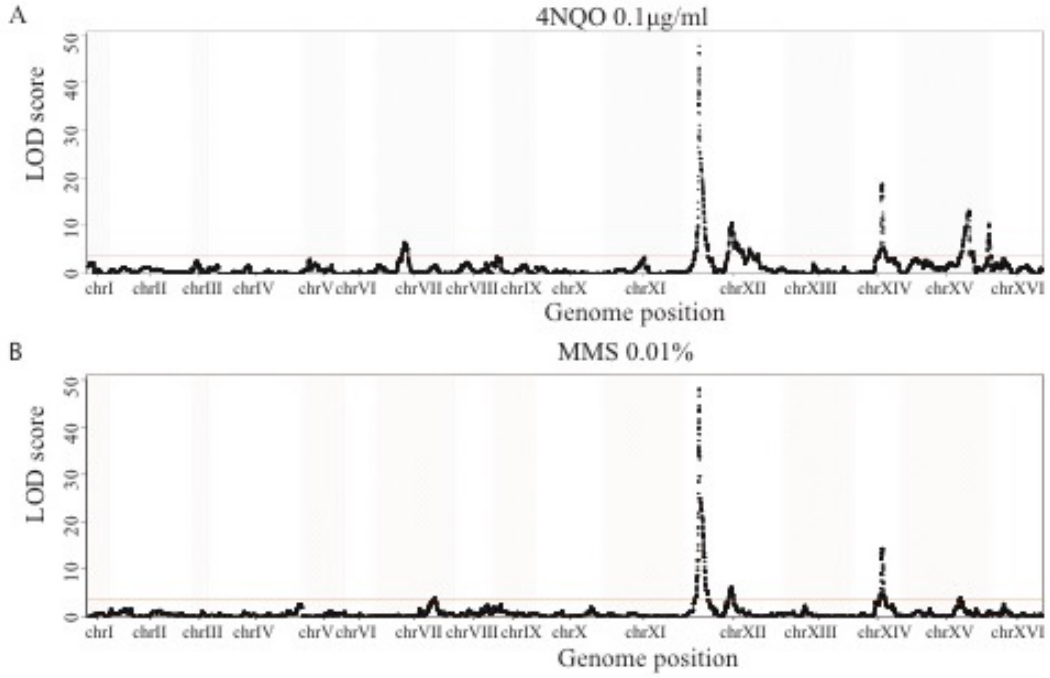


Figure 2.4 Loci underlying mutation rate variation, 4NQO sensitivity and MMS sensitivity are overlapped.

(A-B) The LOD scores for 4NQO (0.1 μg/ml) sensitivity and MMS (0.01%) sensitivity are plotted against the genetic map. The red line indicates a 5% FWER significance threshold (LOD=3.62 for 4NQO and LOD=3.61 for MMS).

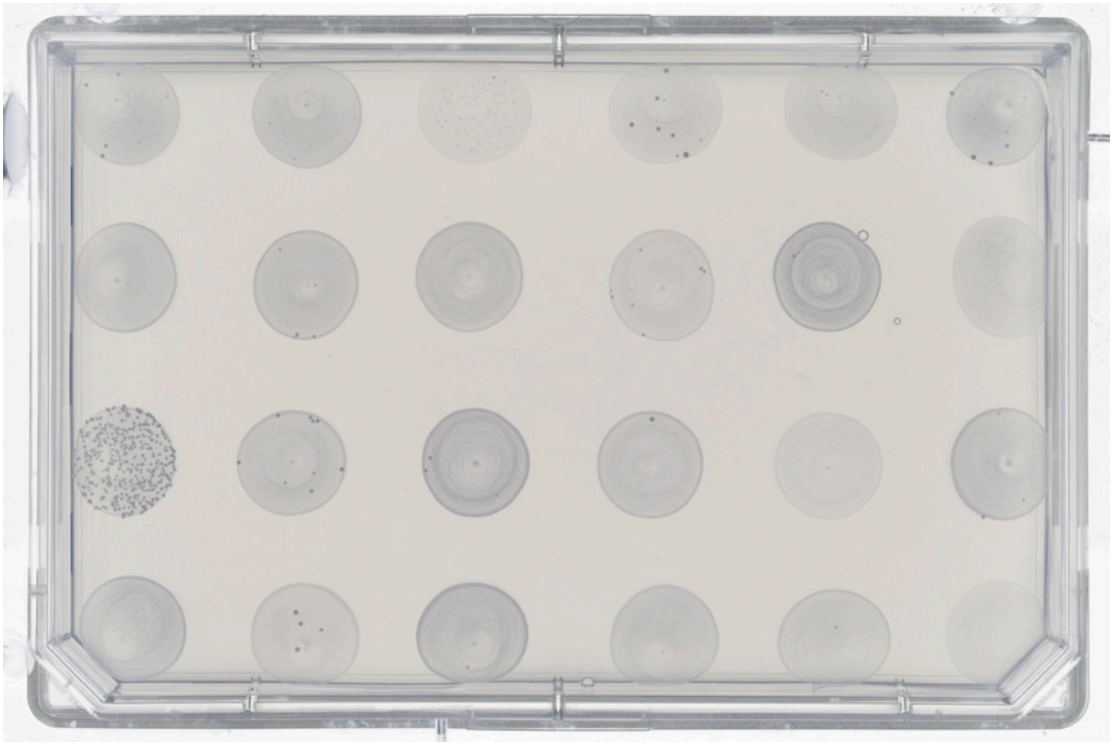
Table 2.1 The allele replacement strains and variant substitution strains

Strain	Background	Relevant Genotype	Source
YLK802	RM	*MATa*, *MKT1*-BY, ho∆::HphMX, flo8∆::NatMX	Smith *et al*., 2008
EAY1463	BY	*MATα*, *lys2*Δ, *RAD5*-RM::NatMX	Demogines *et al*., 2008
EAY1471	BY	*MATα*, *lys2*Δ, *RAD5*-I791S::KanMX	Demogines *et al*., 2008
EAY2169	BY	*MATα*, *lys2*Δ, *RAD5*-E783D::KanMX	Demogines *et al*., 2008

Table 2.2 DNA damaging agents used for the sensitivity assay

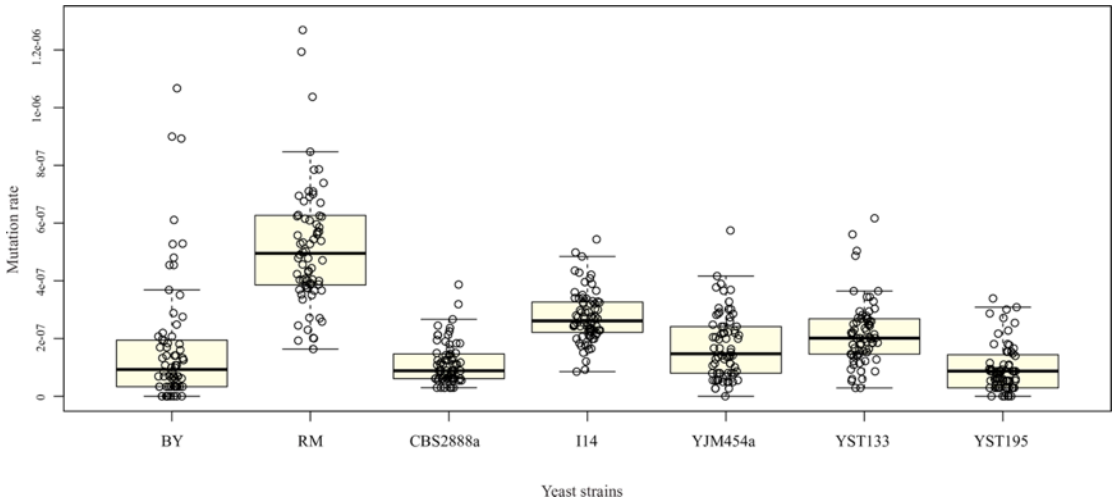
Agent	Agent characteristic
Hydrogen peroxide (H_2O_2)	Altering DNA structure
Methyl methane sulfonate (MMS)	Altering (alkylating) DNA bases
4-nitroquinoline 1-oxide (4NQO)	UV mimetic

Supplementary information



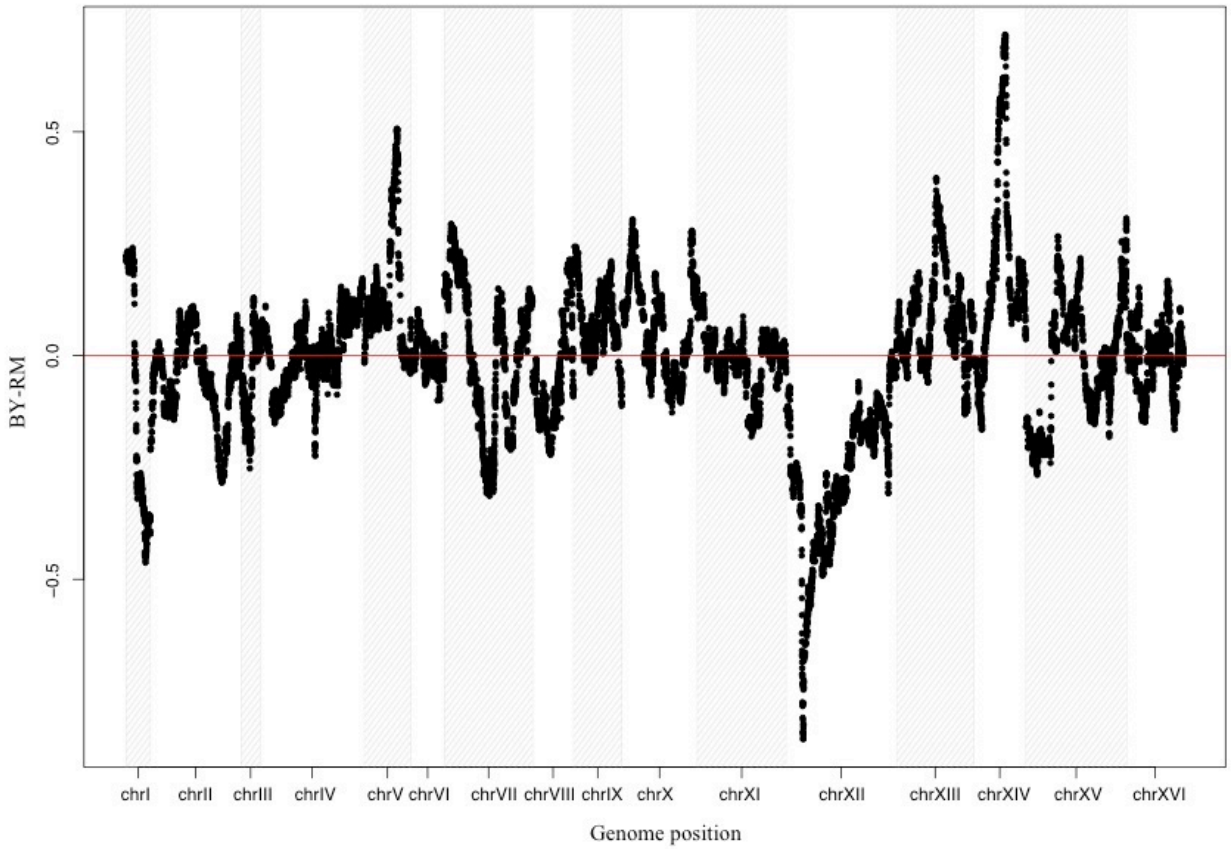
Supplementary Figure 2.5 An example of the fluctuation analysis canavanine plate.

Saturated yeast cultures were spot-plated onto canavanine plates in a four by six configuration using the automated workstation. Plates with spot-plated yeast culture were dried and incubated at 30ºC for ~48 hours. Images of the plates were taken by an imaging robot. The above plate was spot-plated with 24 independent yeast cultures. The little dark dots in the spot are the canavanine resistant colonies. The number of observed resistant colonies varies between different cultures.



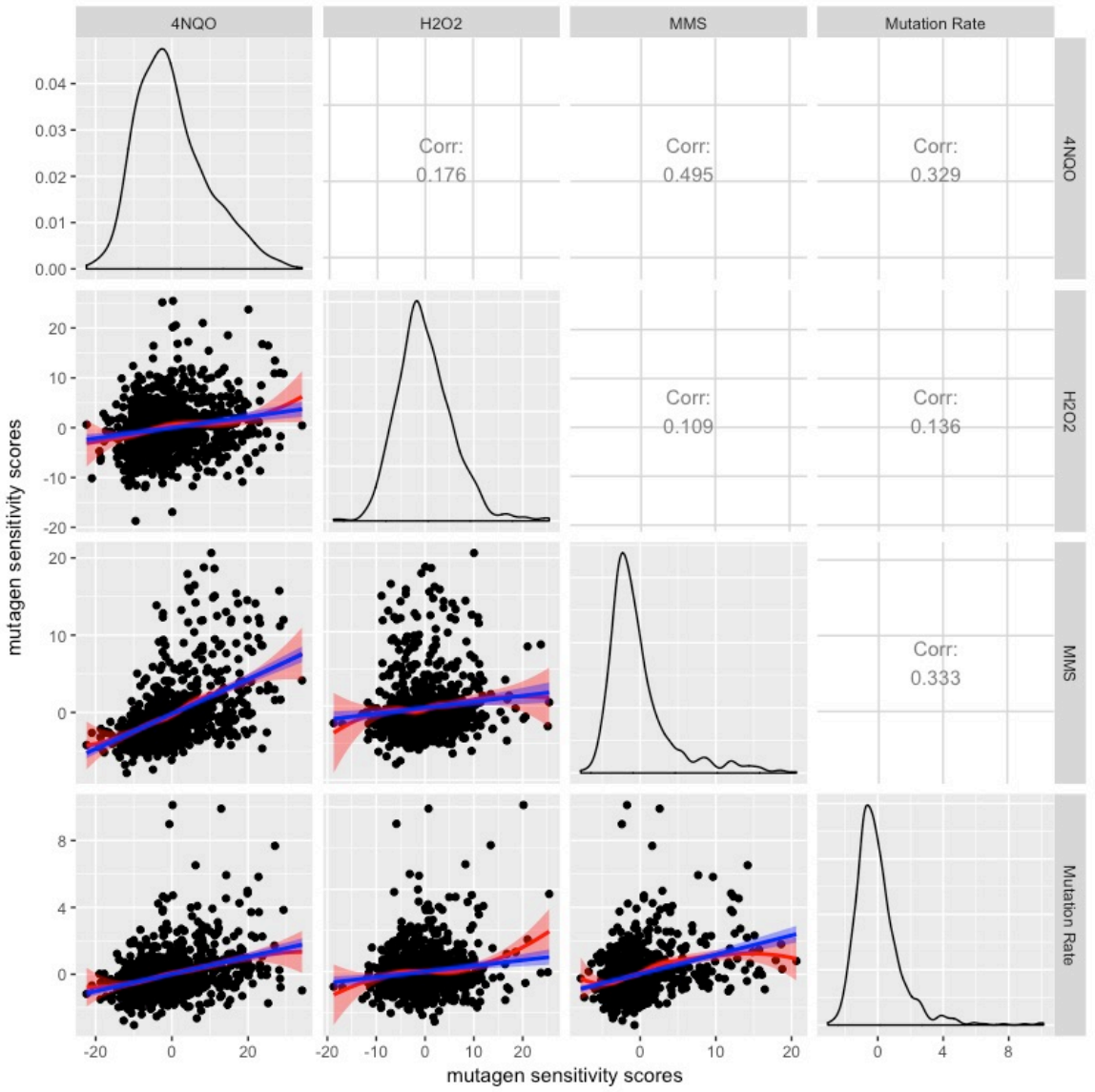
Supplementary Figure 2.6 Mutation rate differs between seven natural yeast strains.

Ninety-six measure of mutation rate was performed for each strain. Means of the mutation rate are plotted as the line. Boxes show the 25%-75% percentile.



Supplementary Figure 2.7 Loci on chromosome XII and XIV have large effects on mutation rate.

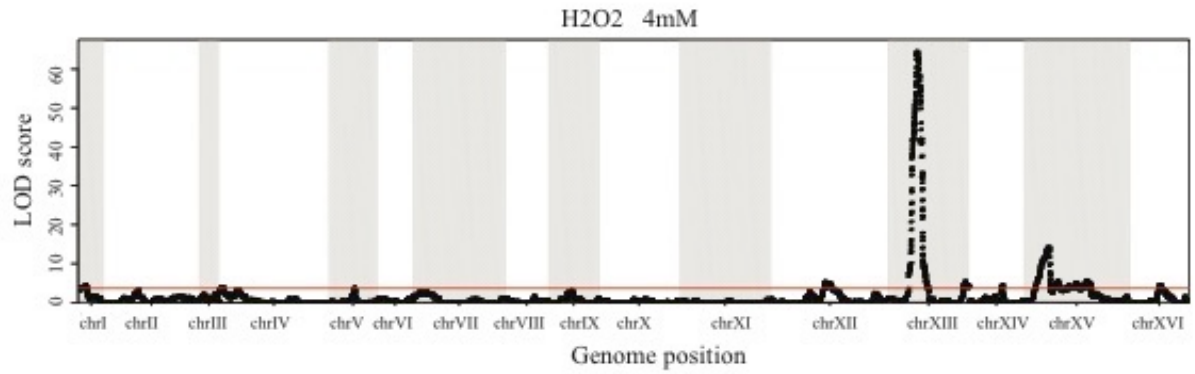
Effect size of genetic markers along the genome shows the BY alleles on chromosome XIV and V increase the mutation rate, while the RM alleles on chromosome XII and I increase the mutation rate.



Supplementary Figure 2.8 Mutation rate is positively correlated with 4NQO, MMS and H_2O_2 sensitivity in the segregant panel.

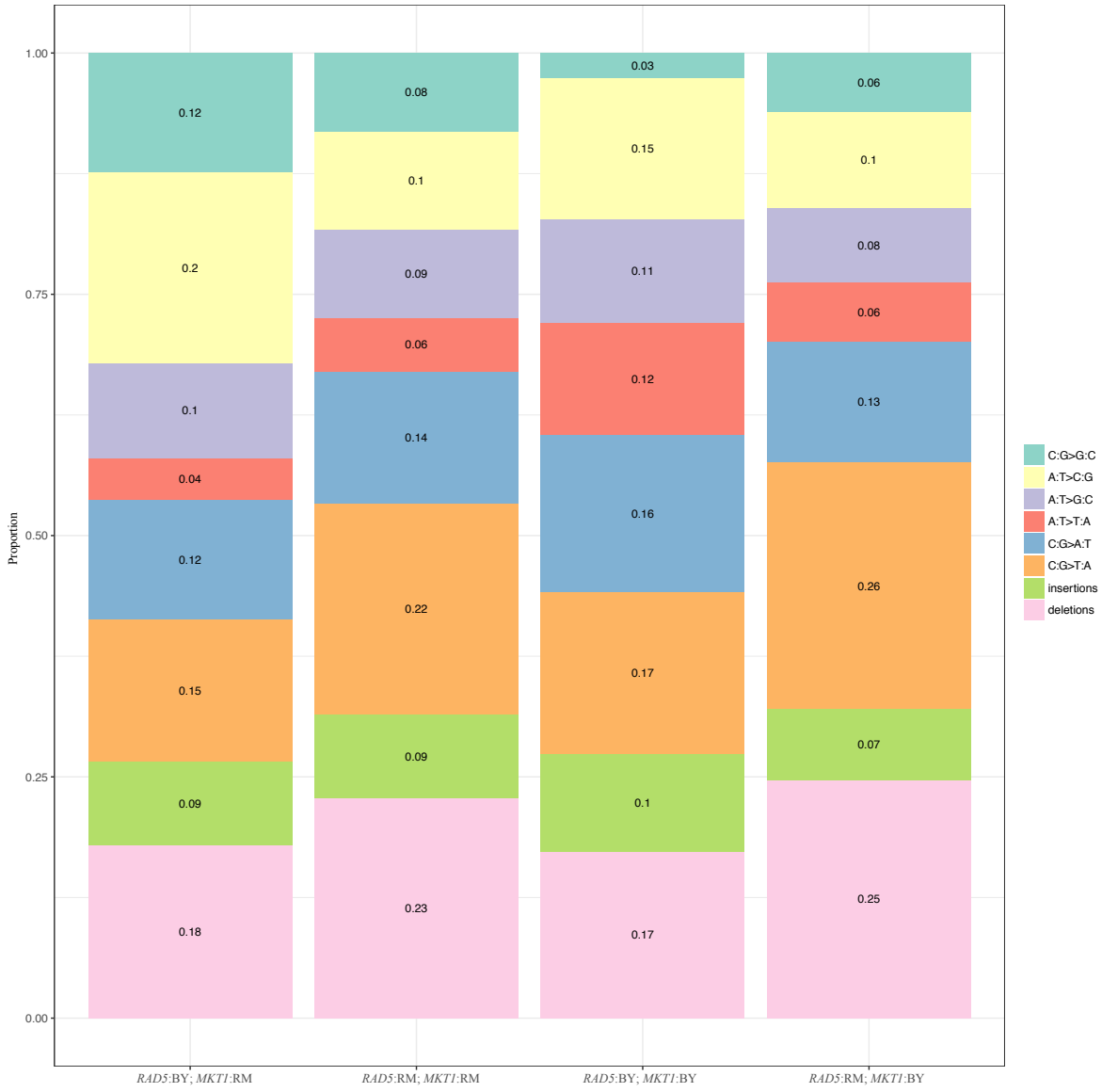
Trait values for mutagens are proxy measurements for mutagen resistance. As shown in the last row, mutation rate is negatively correlated with 4NQO, MMS and H_2O_2

resistances, meaning mutation rate is positively correlated with the sensitivity of these mutagens. Data are displayed in the lower triangle and the linear Pearson correlation values are shown in the upper triangle. The blue lines show the linear regression fit for the points. The red lines show the locally weighted scatterplot smoothing (LOWESS) fit for the points. The slopes of the lines indicate the correlation between the sensitivity of different mutagens and mutation rate.



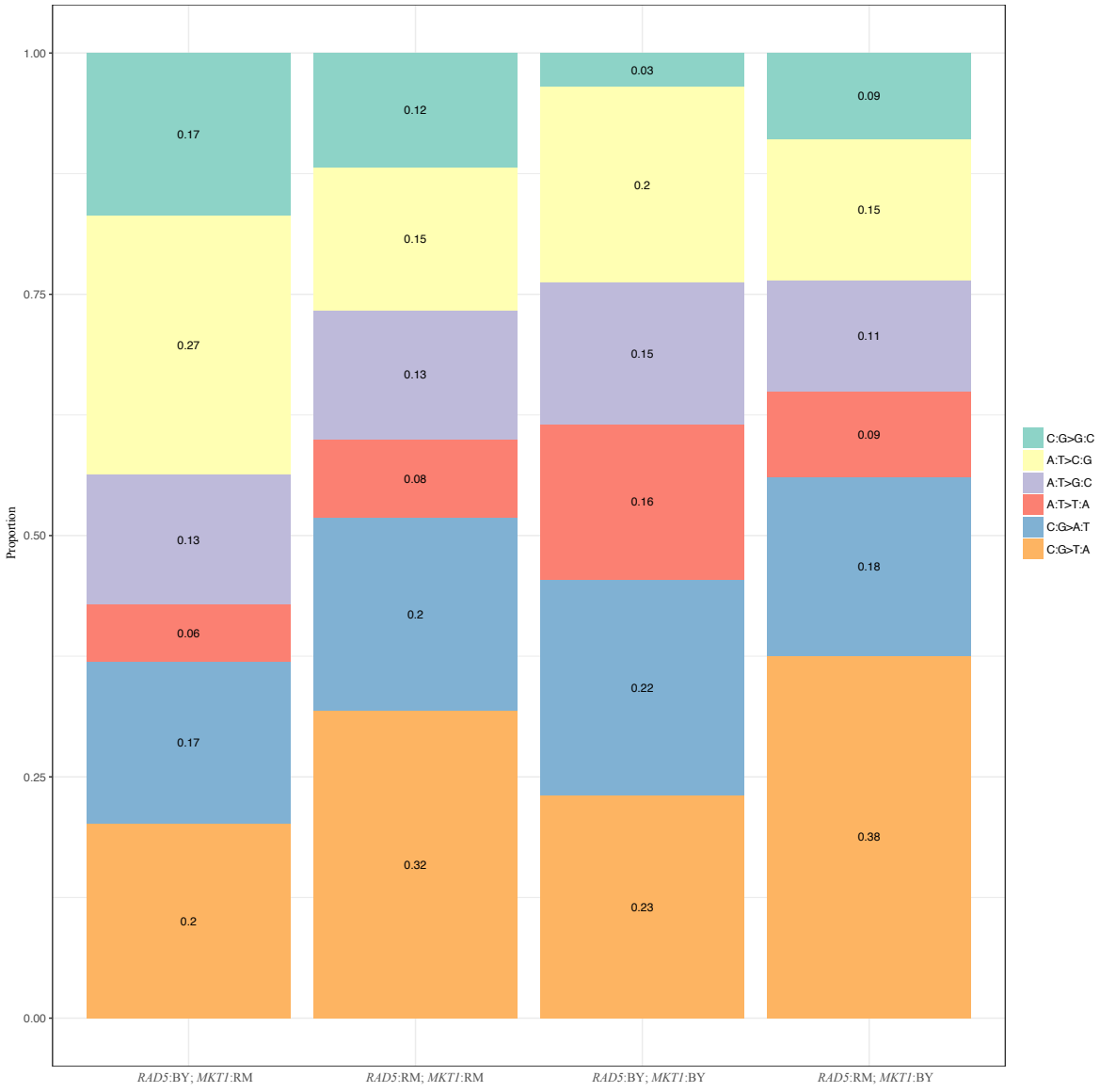
Supplementary Figure 2.9 Loci underlie the H_2O_2 sensitivity.

LOD scores of sensitivity for H_2O_2 (4mM) are plotted against the genetic map. The red line indicates the significant threshold (3.64) from 1000 permutations.



Supplementary Figure 2.10 The proportions of the possible base pair substitution types and indels in different segregant groups.

The color represents six different types of base pair substitutions, and the proportion of each substitution type or indels is labeled on the figure.



Supplementary Figure 2.11 The proportions of the possible base pair substitution types in different segregant groups.

The color represents six different types of base pair substitutions, and the proportion of each type is labeled on the figure.

Supplementary Table 2.3 The mutation rate of seven natural yeast strains.

Strain	Mutation rate	Std. Error
BY4724	1.7×10^{-7}	2.2×10^{-7}
RM11-1a	5.8×10^{-7}	4.0×10^{-7}
CBS2888a	1.1×10^{-7}	7.2×10^{-8}
I14	2.8×10^{-7}	9.0×10^{-8}
YJM454a	1.7×10^{-7}	1.1×10^{-7}
YST133	2.2×10^{-7}	1.1×10^{-7}
YST195	1.7×10^{-7}	5.3×10^{-7}

Mutation rate shown in the table is the mean of ninety-six replicates.

Supplementary Table 2.4 The number of segregants and the allele at gene *RAD5* and *MKT1* of each group.

Allele at *RAD5*	RM	RM	BY	BY
Allele at *MKT1*	RM	BY	RM	BY
Number of segregants	281	230	252	277

We divided 1040 segregants into four groups based on their genotypes at *RAD5* and *MKT1*. The genotype and the number of segregants within each group is shown in the above table.

Supplementary Table 2.5 The *CAN1* region amplicon sequencing read counts of segregants in four groups.

Allele at *RAD5*	RM	RM	BY	BY
Allele at *MKT1*	RM	BY	RM	BY
Original read counts	91722	248508	190548	160182
Adjusted read counts	91722	92433	88963	89090

Segregants were assigned into four groups based on their alleles at gene *RAD5* and *MKT1* (Supplementary Table 2.4). The *CAN1* coding region of the segregants in each group was amplified and sequenced. The number of the original aligned read counts and the adjusted read counts for each library is shown in the table. The read counts were adjusted by a down-sampling process that is described in the methods section.

Supplementary Table 2.6 The mutation spectra of the four groups.

Type of mutation	Number of mutations detected			
Allele				
RAD5	RM	RM	BY	BY
MKT1	RM	BY	RM	BY
Transition				
C:G → T:A	43	59	24	33
A:T → G:C	18	18	16	21
Transitions total	61	77	40	54
Transversion				
C:G → A:T	27	29	20	32
C:G → G:C	16	14	20	5
A:T → T:A	11	14	7	23
A:T → C:G	20	23	32	29
Transversions total	74	80	79	89
One base pair indels				
Insertions	17	17	14	20
Deletions	45	57	29	34

Indels total	62	74	43	54
Total	197	231	162	197

Supplementary Table 2.7 The primers used for amplifying the *CAN1* gene region.

Eight primers were used to amplify the coding region of gene *CAN1*, each primer has the linked MiSeq adapter sequence.

Primers used to amplify the *CAN1* region.

The primers have the miseq adapter overhangs

Forward overhang: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

Reverse overhang: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Primer name	Sequence
CAN1-1F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGACGCCGACATAGAGGAGA
CAN1-1R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ACCAAGGGCAATCATACCAA
CAN1-2F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTGGTATGATTGCCCTTGGT
CAN1-2F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AAGTTCCAGGGCAAAAGTGA
CAN1-3F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCACTTTTGCCCTGGAACTT
CAN1-3R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCACCTGGGTTTCTCCAATA

CAN1-4F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTTTGTGGTGCTGGGGTTAC
CAN1-4F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TCTTGGAACGGATTTTCTGG
CAN1-5F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TAACGCTGCCTTCACATTTC
CAN1-5R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCGGCAGAAATAATGGTTGT
CAN1-6F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG ATTTCTGCCGCAAATTCAAA
CAN1-6F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCCACGGTATTTCAAAGC
CAN1-7F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AAATACCGTGGCATCTCTCG
CAN1-7R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TTCCCATACAATTGCCTCAA
CAN1-8F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TTTGCACCAAAATTCAATGG
CAN1-8R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TGAGGGTGAGAATGCGAAAT

References

1. Lieber MR. The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. Annu Rev Biochem. 2010;79: 181–211. doi:10.1146/annurev.biochem.052308.093131
2. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: Glimpses from the young and old. Nature Reviews Genetics. 2003. pp. 865–875. doi:10.1038/nrg1204
3. Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci U S A. 1996;93: 14800–14803. doi:10.1073/pnas.93.25.14800
4. Baer CF, Miyamoto MM, Denver DR. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet. 2007;8: 619–631. doi:10.1038/nrg2158
5. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics. 1998;148: 1667–1686. doi:citeulike-article-id:610966
6. Demerec M. Frequency of spontaneous mutations in certain stocks of drosophila melanogaster. Genetics. 1937;22: 469–478.
7. Luria S, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 1943;28: 491–511. doi:10.1038/nature10260
8. Lang GI, Murray AW. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics. 2008;178: 67–82. doi:10.1534/genetics.107.071506
9. Zhu YO, Siegal ML, Hall DW, Petrov D a. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A. 2014; 1–9.

doi:10.1073/pnas.1323011111

10. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. Nature Publishing Group; 2013;494: 234–237. doi:10.1038/nature11867
11. Demogines A, Smith E, Kruglyak L, Alani E. Identification and dissection of a complex DNA repair sensitivity phenotype in baker's yeast. PLoS Genet. 2008;4. doi:10.1371/journal.pgen.1000123
12. Amberg DC, Burke DJ, Strathern JN. Methods in yeast genetics: a cold spring harbor laboratory course manual [Internet]. A Cold Spring Harbor Laboratory Course Manual. 2005. Available: http://www.amazon.com/Methods-Yeast-Genetics-Spring-Laboratory/dp/0879697288
13. Sarkar S, Ma WT, Sandri GH. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. Genetica. 1992;85: 173–179. doi:10.1007/BF00120324
14. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. Genetics. 1994;138: 963–971. doi:10.1534/genetics.107.080101
15. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun. 2015;6: 8712. doi:10.1038/ncomms9712
16. Lang GI, Murray AW. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. Genetics. 2008;178: 67–82. doi:10.1534/genetics.107.071506
17. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L. Comprehensive

polymorphism survey elucidates population structure of Saccharomyces cerevisiae. Nature. 2009;458: 342–345. doi:10.1038/nature07670

18. Matheson K, Parsons L, Gammie A. Whole-Genome Sequence and Variant Analysis of W303, a Widely-Used Strain of *Saccharomyces cerevisiae*. G3: Genes|Genomes|Genetics. 2017;7: 2219–2226. doi:10.1534/g3.117.040022
19. Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L. Genome-wide analysis of nucleotide-level variation in commonly used Saccharomyces cerevisiae strains. PLoS One. 2007;2. doi:10.1371/journal.pone.0000322
20. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Genetics and Analysis of Quantitative Traits. 1998. p. 980. doi:10.1086/318209
21. Blastyák A, Pintér L, Unk I, Prakash L, Prakash S, Haracska L. Yeast rad5 protein required for postreplication repair has a DNA helicase activity specific for replication fork regression. Mol Cell. 2007;28: 167–175. doi:10.1016/j.molcel.2007.07.030
22. Torres-Ramos CA, Prakash S, Prakash L. Requirement of RAD5 and MMS2 for postreplication repair of UV-damaged DNA in Saccharomyces cerevisiae. Mol Cell Biol. 2002;22: 2419–2426. doi:10.1128/MCB.2419
23. Bi X. Mechanism of DNA damage tolerance. World J Biol Chem. 2015;6: 48. doi:10.4331/wjbc.v6.i3.48
24. Sonnhammer ELL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins Struct Funct Genet.

1997;28: 405–420. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L

25. Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, et al. Genome evolution across 1,011 Saccharomyces cerevisiae isolates. Nature. Nature Publishing Group; 2018;556: 339–344. doi:10.1038/s41586-018-0030-5
26. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. Oxford University Press; 2015;31: 2745–7. doi:10.1093/bioinformatics/btv195
27. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16: 197–212. doi:10.1038/nrg3891
28. Deutschbauer AM, Davis RW. Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet. Nature Publishing Group; 2005;37: 1333–1340. doi:10.1038/ng1674
29. Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, et al. Dissecting the architecture of a quantitative trait locus in yeast. Nature. 2002. pp. 326–330. doi:10.1038/416326a
30. Parreiras LS, Kohn LM, Anderson JB. Cellular Effects and Epistasis among Three Determinants of Adaptation in Experimental Populations of Saccharomyces cerevisiae□??? Eukaryot Cell. 2011;10: 1348–1356. doi:10.1128/EC.05083-11
31. Swinnen S, Schaerlaekens K, Pais T, Claesen J, Hubmann G, Yang Y, et al. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. Genome Res. Cold Spring Harbor Laboratory Press; 2012;22: 975–84.

doi:10.1101/gr.131698.111

32. Wickner RB. Plasmids controlling exclusion of the K2 killer double-stranded RNA plasmid of yeast. Cell. 1980;21: 217–226. doi:10.1016/0092-8674(80)90129-4

33. Dimitrov LN, Brem RB, Kruglyak L, Gottschling DE. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. Genetics. 2009;183: 365–383. doi:10.1534/genetics.109.104497

34. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. PLoS Biol. 2008;6: 810–824. doi:10.1371/journal.pbio.0060083

35. Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature. 2016;532: 264–267. doi:10.1038/nature17661

36. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. Nature Research; 2015;521: 81–84. doi:10.1038/nature14173

37. Frankfurt OS. Inhibition of DNA repair and the enhancement of cytotoxicity of alkylating agents. Int J cancer. 1991;48: 916–23. Available: http://www.ncbi.nlm.nih.gov/pubmed/1907257

38. Sun Y, Moses RE. Reactivation of psoralen-reacted plasmid DNA in Fanconi anemia, xeroderma pigmentosum, and normal human fibroblast cells. Somat Cell Mol Genet. 1991;17: 229–238. Available: http://www.ncbi.nlm.nih.gov/pubmed/2047939

39. Hampsey M. A review of phenotypes in Saccharomyces cerevisiae. Yeast. 1997.

pp. 1099–1133. doi:10.1002/(SICI)1097-0061(19970930)13:12<1099::AID-YEA177>3.0.CO;2-7

40. Unk I, Hajdú I, Blastyák A, Haracska L. Role of yeast Rad5 and its human orthologs, HLTF and SHPRH in DNA damage tolerance. DNA Repair. 2010. pp. 257–267. doi:10.1016/j.dnarep.2009.12.013

41. Hishida T, Kubota Y, Carr AM, Iwasaki H. RAD6-RAD18-RAD5-pathway-dependent tolerance to chronic low-dose ultraviolet light. Nature. 2009;457: 612–615. doi:10.1038/nature07580

42. Neelsen KJ, Lopes M. Replication fork reversal in eukaryotes: from dead end to dynamic response. Nat Rev Mol Cell Biol. 2015;16: 207–220. doi:10.1038/nrm3935

43. Demogines A, Wong A, Aquadro C, Alani E. Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. PLoS Genet. 2008;4: e1000103. doi:10.1371/journal.pgen.1000103

44. Foss EJ, Radulovic D, Shaffer SA, Goodlett DR, Kruglyak L, Bedalov A. Genetic variation shapes protein networks mainly through non-transcriptional mechanisms. PLoS Biol. 2011;9. doi:10.1371/journal.pbio.1001144

45. Wang X, Kruglyak L. Genetic basis of haloperidol resistance in Saccharomyces cerevisiae is complex and dose dependent. PLoS Genet. 2014;10. doi:10.1371/journal.pgen.1004894

46. Vermut M, Widner WR, Dinman JD, Wickner RB. XIV. Yeast sequencing reports. Sequence of MKT1, needed for propagation of M2 satellite dsRNA of the L-A

virus of Saccharomyces cerevisiae. Yeast. 1994;10: 1477–1479. doi:10.1002/yea.320101111

47. Tkach JM, Yimit A, Lee AY, Riffle M, Costanzo M, Jaschob D, et al. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat Cell Biol. 2012;14: 966–976. doi:10.1038/ncb2549

48. Tran PT, Fey JP, Erdeniz N, Gellon L, Boiteux S, Liskay RM. A mutation in EXO1 defines separable roles in DNA mismatch repair and post-replication repair. DNA Repair (Amst). 2007;6: 1572–1583. doi:10.1016/j.dnarep.2007.05.004

49. Koprowski P, Fikus MU, Dzierzbicki P, Mieczkowski P, Lazowska J, Ciesla Z. Enhanced expression of the DNA damage-inducible gene DIN7 results in increased mutagenesis of mitochondrial DNA in Saccharomyces cerevisiae. Mol Genet Genomics. 2003;269: 632–639. doi:10.1007/s00438-003-0873-8

50. Northam MR, Robinson HA, Kochenova O V., Shcherbakova P V. Participation of DNA polymerase□?? in replication of undamaged DNA in Saccharomyces cerevisiae. Genetics. 2010;184: 27–42. doi:10.1534/genetics.109.107482

51. Sparks JL, Chon H, Cerritelli SM, Kunkel TA, Johansson E, Crouch RJ, et al. RNase H2-Initiated ribonucleotide excision repair. Mol Cell. 2012;47: 980–986. doi:10.1016/j.molcel.2012.06.035

52. Sommer D, Stith CM, Burgers PMJ, Lahue RS. Partial reconstitution of DNA large loop repair with purified proteins from Saccharomyces cerevisiae. Nucleic Acids Res. Oxford University Press; 2008;36: 4699–707. doi:10.1093/nar/gkn446

53. Clarkson SG. The XPG story. Biochimie. 2003. pp. 1113–1121.

doi:10.1016/j.biochi.2003.10.014

54. O'Donovan A, Scherly D, Clarkson SG, Wood RD. Isolation of active recombinant XPG protein, a human DNA repair endonuclease. J Biol Chem. 1994;269: 15965–15968.

55. O'Driscoll M, Macpherson P, Xu YZ, Karran P. The cytotoxicity of DNA carboxymethylation and methylation by the model carboxymethylating agent azaserine in human cells. Carcinogenesis. 1999;20: 1855–1862. doi:10.1093/carcin/20.9.1855

56. Jerison ER, Kryazhimskiy S, Mitchell J, Bloom JS et al. Genetic variation in adaptability and pleiotropy in budding yeast. Elife. eLife Sciences Publications Limited; 2017;6: 1–38. doi:10.1101/121749

57. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol. 2016;34: 339–344. doi:10.1038/nbt.3481

58. Yang D, Scavuzzo MA, Chmielowiec J, Sharp R, Bajic A, Borowiak M. Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases. Sci Rep. 2016;6. doi:10.1038/srep21264

59. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol. 2015;33: 543–548. doi:10.1038/nbt.3198

60. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL.

Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat Biotechnol. 2015;33: 538–542. doi:10.1038/nbt.3190

61. Song J, Yang D, Xu J, Zhu T, Chen YE, Zhang J. RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun. 2016;7. doi:10.1038/ncomms10548

62. Lin S, Staahl BT, Alla RK, Doudna JA. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife. 2014;3: e04766. doi:10.7554/eLife.04766

Chapter 3 Increasing the genome-targeting scope and type of base editing with engineered CRISPR-Cas9 base editors

Abstract

The CRISPR base editors are programmable DNA editing systems that induce single-nucleotide changes in the DNA using a fusion protein containing a catalytically defective Cas9, a cytidine or adenine deaminase, and an inhibitor of base excision repair. This genome editing approach has the advantage that it does not require generation of double-stranded DNA breaks or a donor DNA template. Adenine and cytidine deaminases convert their target nucleotides to other DNA bases, enabling versatile DNA editing. Base editors with natural or engineered Cas9 variants can target genomes at different protospacer-adjacent motif (PAM) sites, which significantly expand the number of sites that can be targeted by base editing. The application of these base editors will largely expand the targeting scope of base editing. However, a systematic analysis of the performance of different base editors has not been conducted. Here we systematically evaluated the editing efficiency of ten different base editors in the model organism *Saccharomyces cerevisiae* using targeted sequencing results at the editing region. The efficiency at ten different regions along the genome showed that three NGG PAM base editors BE3, BE4 and BE4-Gam had the highest editing efficiency, followed by the NGA PAM base editor VQR with editing efficiency around 50%. We observed that most of the base editors had a preference for editing cytosines within the position window 4 to 8 on gRNA, while in the VQR, EQR, VRER base editors the editing window

was wider, around positions 4 to 11. Our study suggests the high editing efficiency of VQR base editor in yeast and its capability to extend the genome-targeting scope for base editors.

Introduction

Genome engineering via the CRISPR (clustered regularly interspaced short palindromic repeat)/Cas9 system has revolutionized biology, therapeutics, and biotechnology. CRISPR/Cas9 is an adaptive immune system used to protect bacterial and archaeal species against invasion by foreign DNA and phages [1][2]. CRISPR is a family of DNA sequences derived from DNA fragments from viruses that have previously infected the bacteria and archaea. Sequences remained in the cells and used to detect and destroy DNA from similar viruses during subsequent infections. Cas9 ("CRISPR-associated protein 9") is an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific strands of DNA that are complementary to the CRISPR sequences. Cas9 activity depends on the presence of a protospacer adjacent motif (PAM) sequence in the target DNA, thereby enabling the CRISPR/Cas9 machinery to recognize self from non-self DNA and can specifically targeting a designed genomic region.

A different version of Cas9 is a catalytically inactive endonuclease, which is called dead Cas9 or dCas9. Dead Cas9 can be used as a DNA targeting tool to tether different enzymatic activities to specific DNA sequences, resulting gene repression, activation and epigenome editing. Interestingly, dCas9 itself could repress the gene

expression by strongly binding to the DNA targeting sequences and interfering with the activity of other DNA binding proteins such as endogenous transcription factors and RNA Polymerase II [3]. Fusing the dCas9 to a stronger repressor complex such as Kruppel-associated Box (KRAB) will have stronger and more specific repression [4], and the gene repression system is called CRISPRi. Similarly fusing dCas9 to a tripartite transactivation domain of the mammalian NF-κB transcription factor could achieve robust gene induction, called as the CRISPRa system. Dead Cas9 is also used in the epigenetic editing field by fusing to the endogenous DNA demethylation. Proteins TET1, TET2 and TET3 play a critical role in dynamic epigenetic regulation that mediates cell type-specific gene expression programming. The dCas9-TET1 fusion complex resulted in DNA demethylation in up to 90% of local CpG dinucleotides and a consequential increase in mRNA expression at the target sites [5].

A specific base editor is invented by fusing a catalytically dead dCas9 to a cytidine deaminase protein, and it can alter DNA bases without inducing a DNA break. Base editors convert C->T (or G->A on the opposite strand) within a small editing window specified by the gRNA [6]. Two classes of DNA base editor have been invented so far: cytosine base editors (CBEs) convert a C•G base pair into a T•A base pair [6], and adenine base editors (ABEs) convert an A•T base pair into a G•C base pair [7].

The first CRIPSR base editor was developed by Liu's group that they engineered fusions of an inactivate version of CRISPR/ Cas9 and a cytidine deaminase enzyme that retain the ability to be programmed with a guide RNA, do not induce dsDNA breaks,

and mediate the direct conversion of cytidine to uridine, thereby effecting a C→T (or G→A) substitution [6]. The resulting 'base editors' convert cytidines within a window of approximately five nucleotides and can efficiently correct a variety of point mutations relevant to human disease without double strand breaks. This was the most commonly-used version of base editor and was known as BE3. Later Liu's group further established five C to T (or G to A) base editors that use natural and engineered Cas9 variants with different protospacer-adjacent motif (PAM) specificities to expand the number of sites that can be targeted by base editing 2.5-fold [8]. Later the same group described adenine base editors (ABEs) that can mediate the conversion of A:T to G:C in genomic DNA [7]. In the same year, an improved version of the original NGG PAM site base editor (BE3) was developed and named as BE4. This version increased the efficiency of C:G to T:A base editing by approximately 50%, while halving the frequency of undesired by-products compared to BE3. This was further improved by fusing BE4 to Gam, a bacteriophage Mu protein that binds DSBs greatly reduces indel formation during base editing (BE4-Gam) [9]. The information of ten base editors that tested in this assay was shown in Table 3.1.

To our knowledge, the efficiency and editing preferences for different base editors in yeast has not been systematically studied before. Here we comprehensively analyzed the targeting efficiency and editing patterns of ten different base editors by analyzing the amplicon sequencing results from ten targeted regions in the genome. We found out BE3, BE4 and BE4-Gam had similar efficiency and editing patterns, consistent with the fact that they are different improved versions of the C to T base

editor with NGG PAM recognition site. Another base editor (VQR) with the NGA PAM site also performed similarly efficient as the BE3, which was able to expand the targeting-scope for base editors by 1.7 folds.

Materials and Methods

Yeast strains and media

Saccharomyces cerevisiae BY4741 (haploid, *MAT* a, *his3*Δ*1, leu2*Δ*0, met15*Δ*0, ura3*Δ*0*) was used as host strain for genome editing. Cells were grown non-selectively in YPAD medium (2% Bacto peptone, 1% Bacto yeast extract, 2% glucose). Selection of yeast transformants based on the *URA*3 and *LEU*2 markers was done on a synthetic complete (SC-U-L) medium (6.7 g/L of Yeast Nitrogen Base, 2% glucose, 0.54g/L Complete Supplement Mixture-Ura-Leu). For culture in Petri dishes, the medium was solidified with 2% agar. Yeast strains were incubated at 30 °C and liquid culture were growing on a shaker. Selective canavanine plates were made from arginine minus synthetic complete agar medium with 60mg/liter L-canavanine (Sigma C1625).

Molecular cloning for plasmids

We have constructed nine CRIPSR base editors that could work in the yeast background by transferring the base editor functional region of existing mammal plasmids into the yeast plasmid pBEVY-GL (pBEVY-GL was a gift from Charles Miller; Addgene plasmid # 51225 ; http://n2t.net/addgene:51225)[8], which contains a galactose inducible promoter. The functional region was amplified using Phusion High-

Fidelity DNA Polymerase (Thermo Fisher) with the primers that contained 30bp-overlapping region around the SacI restriction enzyme digestion region (enzyme recognize site: gagctc). The PCR product was digested by Dpn1 restriction enzyme for 30min to remove the circular plasmid template given the methylated characteristic of bacteria DNA. The PCR product was cleaned up using the Qiagen PCR cleanup kit. The pBEVY-GL plasmid was digested by SacI and the product was cleaned up by gel purification. Gibson ligation was used to ligate the plasmid backbone and the PCR amplification products to construct the base editors. The mammal plasmids for nine different CRISPR base editor are listed in Table 3.1.

gRNA plasmids construction

To generate edit-directing plasmids, we synthesized DNA fragments carrying the desired gRNA as well as the sites for *BbsI* restriction enzyme cutting as DNA oligos (Integrated DNA Technologies), which were then cloned into pOS05a (pRS426 SNR52p-lentiCR2-gRNA-filler-SUP4t), a plasmid for expressing gRNAs under an SNR52 promoter, through T4 ligation. We designed and synthesized 10 gRNAs that targeted three different genes in the genome: 4 gRNAs targeting gene *CAN1* on chromosome V, 3 gRNAs targeting gene *CBS1* on chromosome IV, 3 gRNAs targeting gene *CCR4* on chromosome I.

The detail steps for targeting sequence cloning is as the follows [9][10]:

To clone the guide sequence into the sgRNA scaffold, synthesize two oligos of the form:

5’ – CACCGNNNNNNNNNNNNNNNNNN – 3’

3’ – CNNNNNNNNNNNNNNNNNNNCAAA – 5’

1) Digest 1ug of pOS05a plasmid with *BbsI* for 30 min at 37°C:

Plasmid	1 ug
BbsI – HF (NEB)	1ul
10X NEB 2.1 Buffer	5 ul
ddH2O	X ul
Total	50 ul

Incubate at 37ºC for 1 hour

2) Gel purification digested plasmid using QIAquick Gel Extraction Kit and elute in EB.

3) Phosphorylate and anneal each pair of oligos:

Oligo1 (100uM)	1 ul
Oligo2 (100uM)	1 ul
10X T4 ligation buffer (NEB)	1 ul
ddH2O	6.5 ul
T4 PNK (NEB)	0.5 ul
Total	10 ul

4) Set up ligation reaction and incubate at room temperature for 10 min:

BbsI digested plasmid from step2 (50ng)	X ul
Phosphorylate and annealed oligo duplex from step 3 (1:200 dilution)	1ul
2X quick ligation Buffer (NEB)	5 ul

ddH2O	X ul
subtotal	10 ul
Quick ligase (NEB)	1 ul
Total	11 ul

5) Transformation

Yeast transformation and genomic DNA extraction

Yeast cells were transformed with the LiAc/SS carrier DNA/PEG method using 0.5–1 μg plasmid DNA[11]. Transgenic clones were selected on SC-U-L media. Single colonies were picked and grown in galactose selective liquid medium (6.7 g/L of Yeast Nitrogen Base, 2% galactose, 0.54g/L Complete Supplement Mixture-Ura-Leu) overnight to induce Cas9 base editor expression. Colonies were allowed to grow for approximately 24 h in the galactose selective medium. Genomic DNA was extracted from 1ml of the harvested cells from galactose culture with a DNeasy Blood and Tissue kit (Qiagen). In parallel, 1 ml of the culture was plated onto the canavanine plates to gain an estimate of number of colonies growing. canavanine plates. Plates were incubating at 30ºC for two days and images of plates were taken using the scanner.

Amplicon sequencing and data analysis

Yeast colonies harboring plasmids expressing base editors and sgRNAs were picked from SC-L-U plates, suspended in 3 mL SC-L-U medium with 2% glucose, and grown to a stationary phase. The cultures were then washed twice to remove residual

glucose, resuspended in 5 mL SC-L-U medium with 2% galactose and 1% raffinose to an OD600 of 0.3, and incubated for 20 h at 28 °C on a rotary shaker. Genomic DNA was extracted from culture samples of 0.5 mL volume, and the regions targeted by base editing were amplified by PCR with primer pairs containing index tags for sample multiplexing. PCR amplification was performed with the Phusion High-Fidelity DNA Polymerase (Thermo Fisher) according to the manufacturer's protocol, followed by product purification with the PCR Clean-up kit (Qiagen). The purified index-labeled PCR products were pooled at equal amount. Pooled library was gel purified and sequenced using MiSeq v2 PE150X2. Then we processed sequencing output reads with the following pipeline: read pairing (PEAE), read trimming (Trimmomatic-0.36), and read alignment (BWA) to the reference-targeting region using the down-sampled fastq files. The editing efficiency of a specific site was calculated as ratio of the reads number contained the desired edits and the total reads number that was the sum of edited and unchanged reads.

Results

Yeast plasmids construction of ten base editors

Plasmids for ten different base editors were all designed for the application of mammalian cell lines expression, where they contain a CMV promoter and selectable markers for cell lines. In order to apply the base editor system in yeast, as well as optimizing it into an inducible system, I amplified the functional region of the plasmids and reconstruct them into the yeast pBEVY-GL backbone, which contains a Gal inducible promoter, a 2micro origin of replication (2u ori) and a Leu2 selectable marker.

The amplified active functional region for BE3 contains the deaminase APOBEC1, the nicase Cas9 (Cas9 D10A) and the UGI sites, similar to other base editor versions. Examples of the structure of newly constructed plasmids was demonstrated in Figure 3.1. The amplified regions were ligated to the backbone through Gibson assembly and the ligated regions were verified by Sanger sequencing (Figure 3.2).

Design of multiple gRNAs targeting genomic DNA

In order to have a systematic view of base editor efficiency across the genome, we designed ten gRNAs targeting ten genomic regions on three different genes. The genes we chose are *CBS1*/YDL069C on chromosome IV, mitochondrial translational activator of the COB mRNA and a membrane protein that interacts with translating ribosomes; *CCR4*/YAL021C on chromosome I, which is a component of the CCR4-NOT transcriptional complex that is involved in regulation of gene expression, as well as a component of the major cytoplasmic deadenylase, which is involved in mRNA poly(A) tail shortening; *CAN1*/YEL063C on chromosome V, which is a plasma membrane arginine permease and loss of function mutations in this gene can cause resistance to canavanine. All three genes are non-essential genes, which enables us to extract DNA after the genetic editing and analyze the editing pattern by amplicon sequencing the targeting region. We chose these genes from the yeast non-essential genes data base from the *Saccharomyces* genome deletion project.

(http://www-sequence.stanford.edu/group/yeast_deletion_project/downloads.html#instru)

Optimize the high-throughput transformation in yeast

An urgent call for the high-throughput transformation pipeline in yeast arised for this project, because we need to test the editing efficiency for ten base editors at ten different genomic sites. A well-controled experiment would be independently performed the transformation step where each yeast culture obtained a specific pair of gRNA and base editor. Here we optizied the LiAc/SS carrier DNA/PEG methods into a large-scale high-efficient yeast transformation protocol using previous researches as reference [14][15]. The process for the high-throughput transformation was shown in Figure 3.3.

During the high-throughput transformation process, two parameters required optimization. One is the amount of plasmids put into the transformation, the other is the time length for 42ºC heatshock. Bigger amount of plasmids will increase the transformation efficiency, but more plasmids demand higher concentration of plasmids. It is important to test out the balance point of the tradeoff. Therefore we performed a gradient of conditions for four different levels of plasmids amount (100ng, 200ng, 400ng and 600ng) and a gradient of conditions for three different incubation time at 42ºC (1 hour, 2 hours and 3 hours). For each condition, we preformed five independent replications of each condition, as well as two negative controls. The result showed that 400ng plasmids with 2 hours heatshock incubation time is the best combination (Figure 3.4). The high-throughput transformation enables us to test the efficiency of ten base editors at the same time, controlling the batch effect of different experiments.

Base editor editing pattern was observed using amplicon sequencing

We applied the optimized protocol with the tested best condition to ten base editors with ten different gRNAs. In general the transformation effiicency is high (Figure 3.5). The base editor plasmids contain the Leu selective marker while the gRNA plasmids contain the Ura selective marker. Transformed yeast were grown in liquid medium that lacked of Leucine and Uracil overnight. Yeast was pinned into the liquid medium with galactose while keeping the selection of Leu and Ura. This step turned on the CRISPR base editor system. DNA was extracted from the culture after 24 hours. Primers were designed and synthesized to amplified the gRNA targeting region. Amplicon sequencing library of the PCR products was sequenced using MiSeq (Figure 3.6).

We used high-throughput DNA sequencing to quantify base-editing efficiency in ten base editors. BE3, BE4-Gam and BE4 very efficiently edited target Cs at most of the targeted loci, with conversion efficiencies of ~50–75% of total DNA sequences converted from C to T, without enrichment for edited cells (Figure 3.7). The efficiency of VQR on NGAN-containing target sites in general performed as well as that of BE3 on NGG-containing target sites, with the conversion efficiencies of ~50-60%. xCas9 had a wide range of PAM recognition sites (NGG, NG, GAA and GAT), and it also had the efficiency expanded a large range, with every around 0 to above 50%. Perhaps due to its preferences of the PAM site or the gRNAs design. EQR and VRER, with the NGAG and NGCG PAM sites respectively, had conversion efficiency around 0-25% at the multiple sites tested in this assay. Sa-BE3 and SaKKH-BE3 recognized largely different

PAM sites (NNGRRT and NNNRRT), but unfortunately not much editing was observed for these two base editors, probably some mutations in the plasmids that decreased the efficiency of the base editors. The ABE base editor that convert A to G instead of C to T had the efficiency around 10%, and optimized gRNA design may help improve the editing efficiency (Figure 3.7).

To further study the targeting window of different base editors, we investigated the editing sites and pattern of the targeted region. Consistent with the previous literatures, we observed editing enriched between position 4 to 8 on gRNAs for base editors BE3, BE4 and BE4-Gam. VQR resulted in detectable base editing at target Cs at positions outside of the canonical BE3 activity window (Figure 3.8), indicating larger number of potential editing sites. xCas9 had the similar editing window as BE3, and the window for other base editors were hard to determined due to the lack of efficiency in general.

We looked into a few gRNAs examples and farther understood the editing efficiency for C at different positions of the gRNA when multiple C appeared in gRNA. In Table 3.2, we showed the efficiency of two example gRNAs for BE3 where one gRNA contained four Cs and the other gRNA had nine Cs along the gRNA sequence. The editing window for BE3 was the C at position 4 to 8, which was consistent with previous studies. For the first gRNA, we found the C at position 5 had the highest efficiency that 57.7% of the C at this position were converted into T, followed by the C at position 7. In

the case of gRNA that contained nine Cs, the highest efficiency was achieved at position four with the efficiency of 70.3%.

We performed the same analysis for the NGA PAM base editor VQR to explore the editing pattern of gRNAs contained multiple Cs. When the gRNA had three Cs at position 4,5 and 10, both the Cs at position 4 and 5 were edited with efficiency around 50%. In the other example where nine Cs in the gRNA, the Cs at positions 6,7 and 8 had high editing efficiencies. Over 50% of the Cs at positions 6 and 7 were converted into T (Table 3.3).

Given the high editing efficiency of the NGA PAM base editor VQR observed in yeast, we measured the number of regions in yeast genome that can be designed into gRNAs based on the PAM sites. We gained 1461302 gRNAs with the NGA PAM sites, while sites with the NGG PAM sites was 805045, meaning the regions with NGA was 1.81-fold more than the NGG sites (Figure 3.9A). We improved the estimate of targetable regions by taking the editing window of BE3 and VQR into account, where we required at least one C in the window 4-8 for BE3 and at least one C in the window 4-11 for VQR. We received 459621 gRNA sites with NGA PAM this time, while the number was 266724 for NGG PAM when considering the BE3 editing window, revealing the NGA PAM base editor VQR had 1.72-fold of the editable sites (Figure 3.9B). The analyses showed VQR could largely extend the editable scope in yeast genome.

Discussion

CRISPR-Cas9 base editors have recently become a powerful tool in genome editing, owing to their capability of producing precise editing without creating DNA double strand breaks [16]. Several base editors have derived from the original *Streptococcus pyogenes* Cas9 (spCas9) base editor BE3 to target different PAM sites or to increase targeting efficiency via the engineered mutations in Cas9. Three or four engineered variants in Cas9 could build new recognition PAM sites, such as the base editor with NGAN, NGAG and NGCG PAM sites, which are referred as VQR (D1135V/R1335Q/T1337R), EQR (D1135E/R1335Q/T1337R) and VRER (D1135V/G1218R/R1335E/T1337R) [17]. The Cas9 homolog from *Staphylococcus aureus* (SaCas9) can mediate efficient genome editing that requires an NNGRRT PAM [18]. SaBE3 is built by replacing the *Streptococcus pyogenes* Cas9 (spCas9) in BE3 to the SaCas9 [8]. Researchers further relax the PAM requirement to NNNRRT by engineering three mutations into SaCas9 and generate the SaKKH-BE3 [8]. XCas9 contains multiple mutations that make it eligible to recognize a broad range of PAM site sequences including NG, GAA and GAT [19]. BE4 and BE4-Gam are the fourth-generation of base editors that showed increasing efficiency of C to T base editing in mammalian cell lines [9]. The ABE base editor which can generate A to G edits was created by replacing the APOBEC1 component of BE3 with a natural adenine deaminase *Escherichia coli* TadA [7]. The above base editors largely extend the genome-targeting scope of base editors.

Here we constructed ten base editors for yeast system and measured the editing efficiency and editing preferences for each of them at ten different sites in the genome. When measuring the editing efficiency of ten different base editors, we optimized a high-throughput yeast transformation protocol, which could be a quicker, easier and cost-effective approach for high-throughput genetic analysis. We found the editing efficiency varies at different positions in the genome, consistent with the previous observation of regular CRISPR-Cas9. Our results showed that in general BE3, BE4, BE4-Gam and VQR had the highest editing efficiency among the base editors tested. The editing efficiency of XCas9 was largely depended on the targeting region of gRNAs. BE3, BE4, BE4-Gam and xCas9 had the preference to edit the C in the position window of 4 to 8 on gRNAs, while VQR had wider editing window preference of 4 to 11, which is also consistent with previous reports [8]. The VQR base editor with NGA PAM site had a high editing efficiency, enabling extended editing scope of base editors in yeast. We observed low editing efficiency for base editors SaBE3, SAKKH-BE3 and ABE, which could be caused by some spontaneous mutations in the plasmid during the construction, which potentially reduces the functional activity of the base editor.

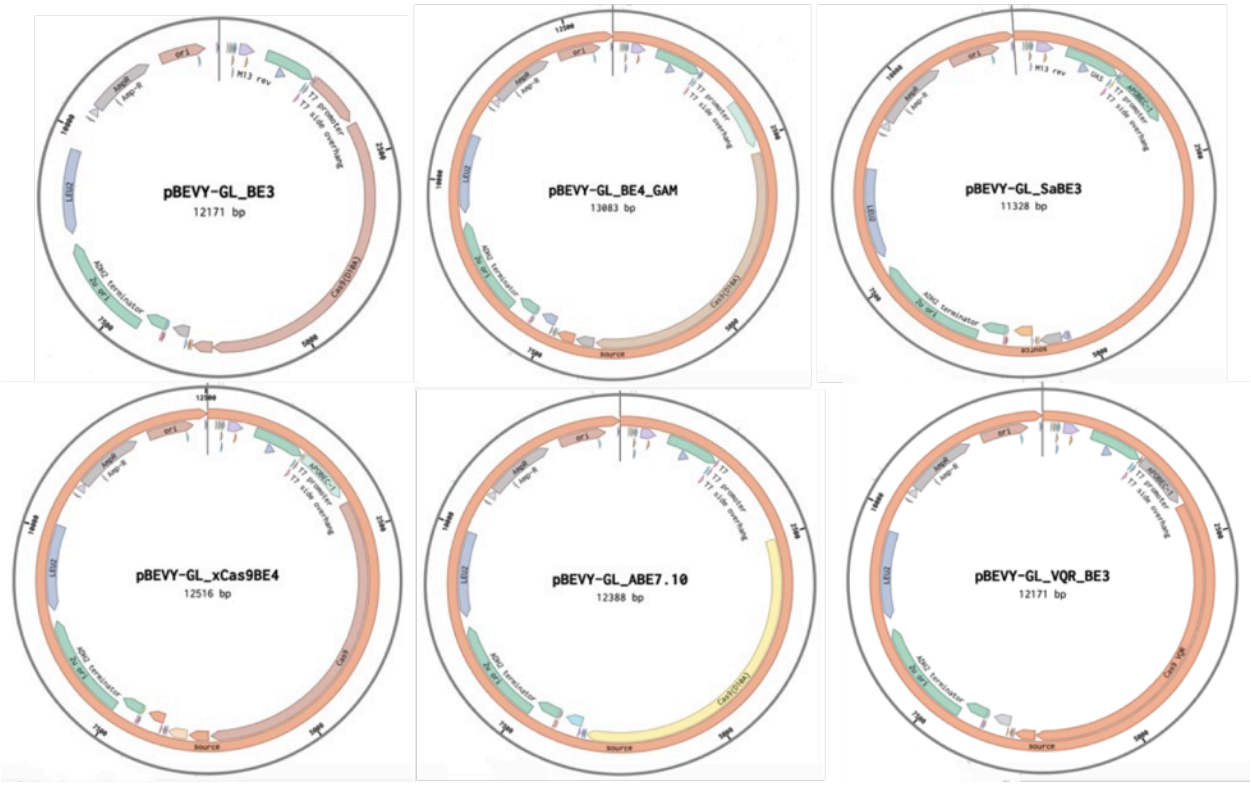


Figure 3.1 Six examples of the plasmid structure of different base editors.

Every base editor plasmid constitutes of a Gal inducible promoter, an Ampicillin resistant marker, a Leu2 auxotrophic selective marker, and a Cas9 base editor functioning part.

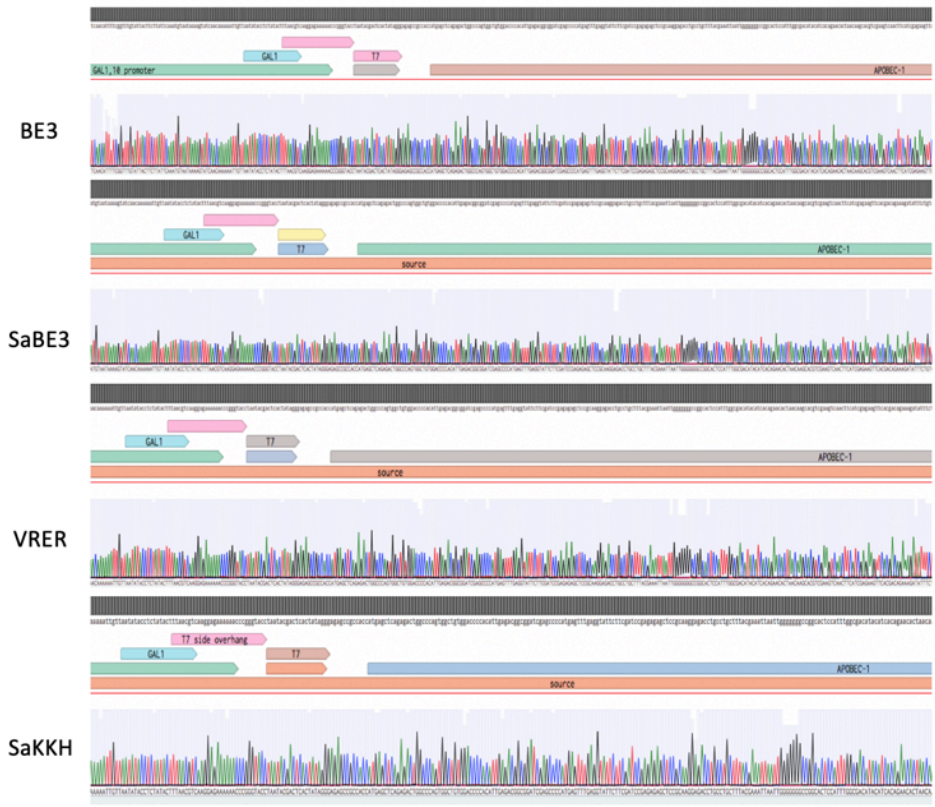


Figure 3.2 Four examples of the confirmation for plasmid construction using sanger sequencing.

The linearlized plasmind was ligated to the PCR product of functional base editor region using Gibson Assembly. The ligation region was verified by sanger sequencing. Sanger sequencing traces were shown above. The Gal1 is the Galactose inducible promoter on the plasmid backbone, and the APOBEC is the Cas9 base editor region, which is on the PCR product. Sequences were aligned to the expected sequences where two regions were aligned.

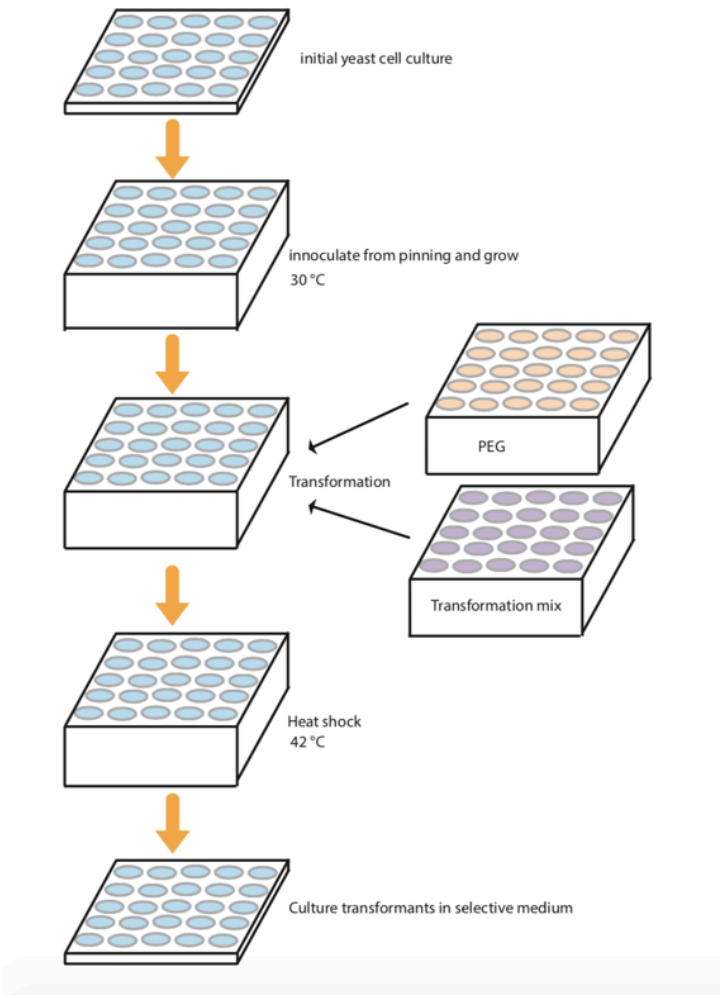


Figure 3.3 The workflow of high-throughput yeast transformation.

The yeast were grown in 96 deep well plates to saturation. PEG, plasmids and transformation reagent mix were added to the plates after centrifuge. After resuspending the yeast using mixer, cells were incubated at 42ºC for 1, 2 or 3 hours for plasmids to get into the cell. After heat shock, cells were transferred to plates with selective medium to grow for 2 days.

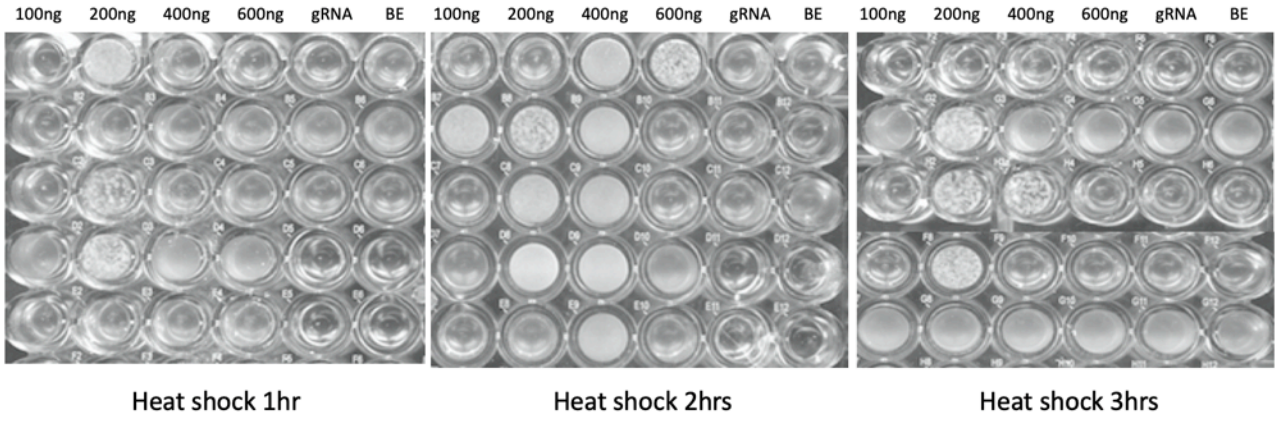


Figure 3.4 The transformation efficiency comparison of 12 tested conditions.

Yeast with -Ura and -Leu genotype was transformed with two plasmids (the gRNA plasmid and the base editor plasmid), each plasmid contains one selective marker. Yeast were growing in liquid selective medium lacking Uracil and Leucine for two days. Only cells successfully obtained both plamsids can survive and grow in the selective medium. The base editor plasmid used was the BE3 plasmid, and was the same for all wells. Each row showed one gRNA plasmids. In the figure we showed the co-transformation efficiency of gRNA 1-5. The columns marked as 100ng, 200ng, 400ng and 600ng represented the amount of plasmids transformed into the well. The gRNA column showed the yeast that was trnasformed the gRNA plamsid only. The BE column showed the yeast that was transformed the base editor plasmids only. For all gRNA plasmids, the efficiency was high when using 400ng plasmids and 2 hours heat shock at 42ºC.

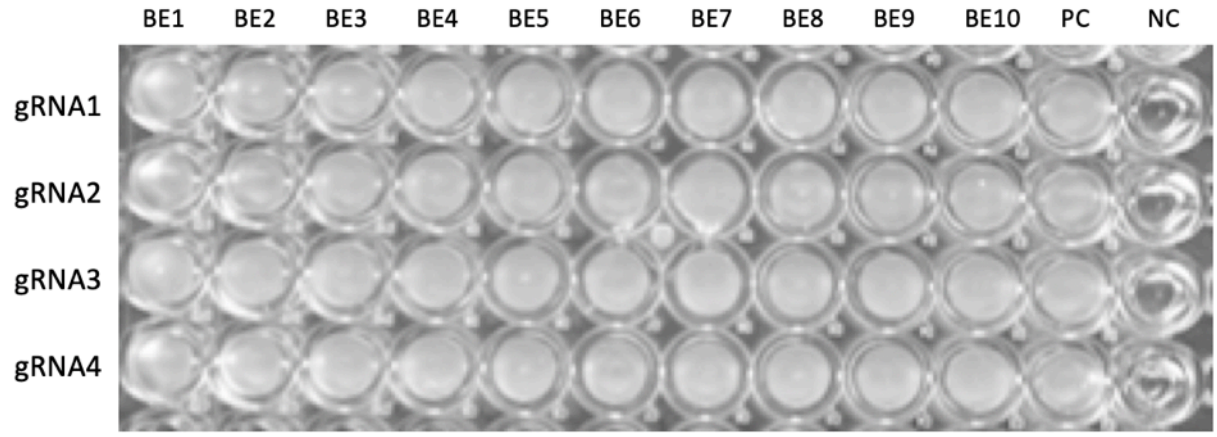


Figure 3.5 High transformation efficiency in 96 well plates was achieved using the optimized conditions.

Yeast that cannot synthesize Uracil and Leucine was transformed with two plasmids, one with the Ura selective marker and the other with the Leu selctive marker. Yeast were growing in liquid selective medium lacking Uracil and Leucine. Only cells successfully obtained both plamsids can survive and grow in the selective medium. Image was taken two days after transformation. PC represents positive control, and NC represents negative control, where just water was transformed into yeast.

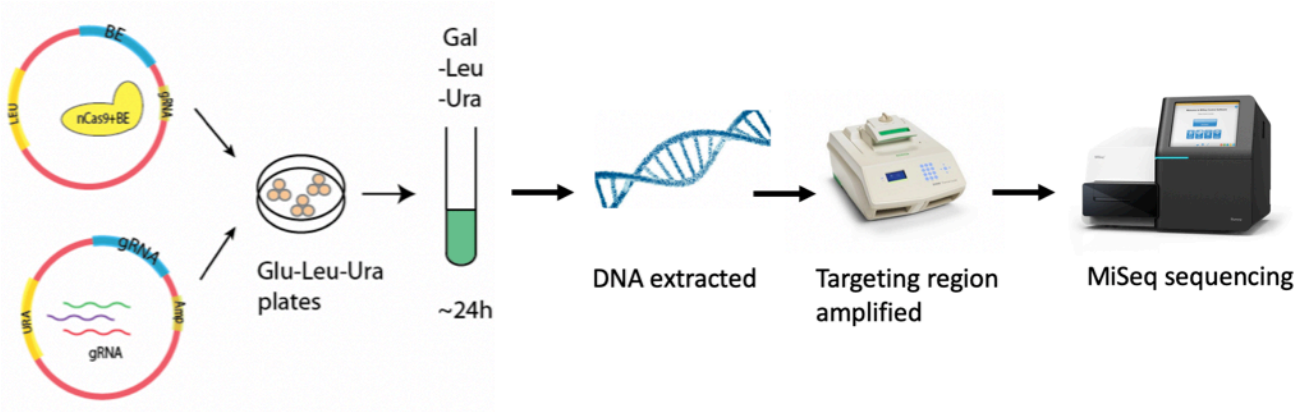


Figure 3.6 The workflow for measuring the base editor editing efficiency through amplicon sequencing.

The gRNA and base editor plasmids were co-transformed into yeast, followed by growing in selective liquid medium. Yeast was incubated in selective medium with galactose for 24 hours before DNA extraction. The targeting regions were amplified and indexed for MiSeq sequencing.

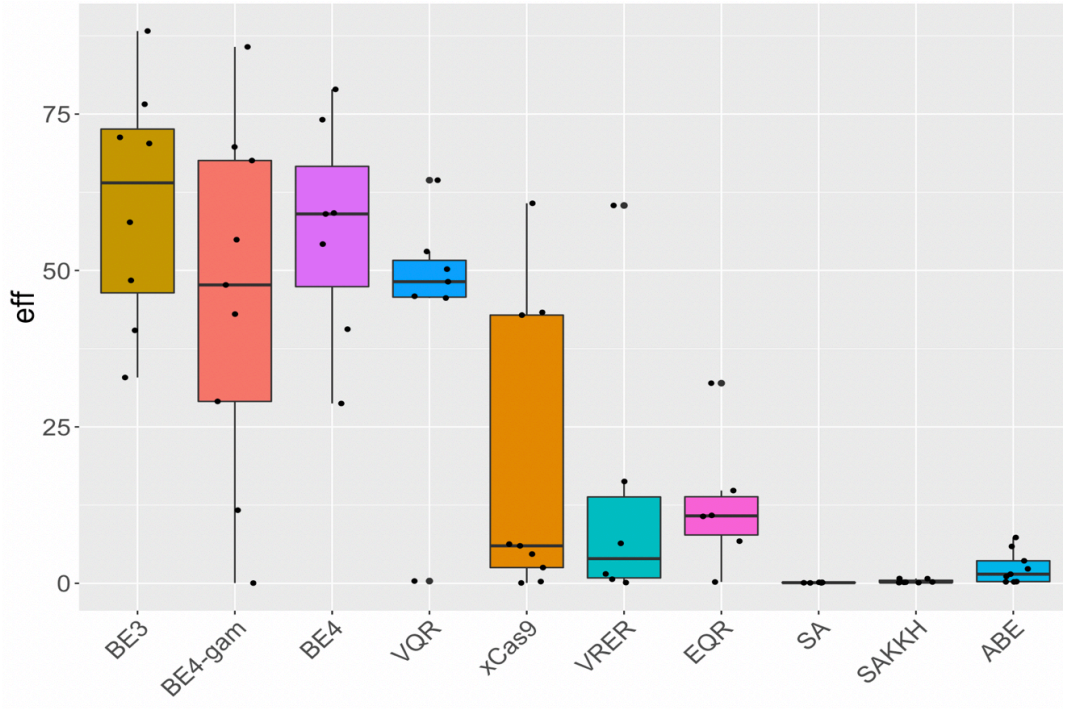


Figure 3.7 The efficiency of ten base editors at the tested positions.

For each base editor, the efficiency of ten gRNAs targeting ten different regions was measured. Each dot represents the editing efficiency of one gRNA. The efficiency information for some gRNAs was missing due to problems during transformation, DNA extraction and PCR amplification. The mean efficiency for the gRNAs was showed as the line, and the box showed the 25 and 75 quantiles. The error bar showed the standard deviation.

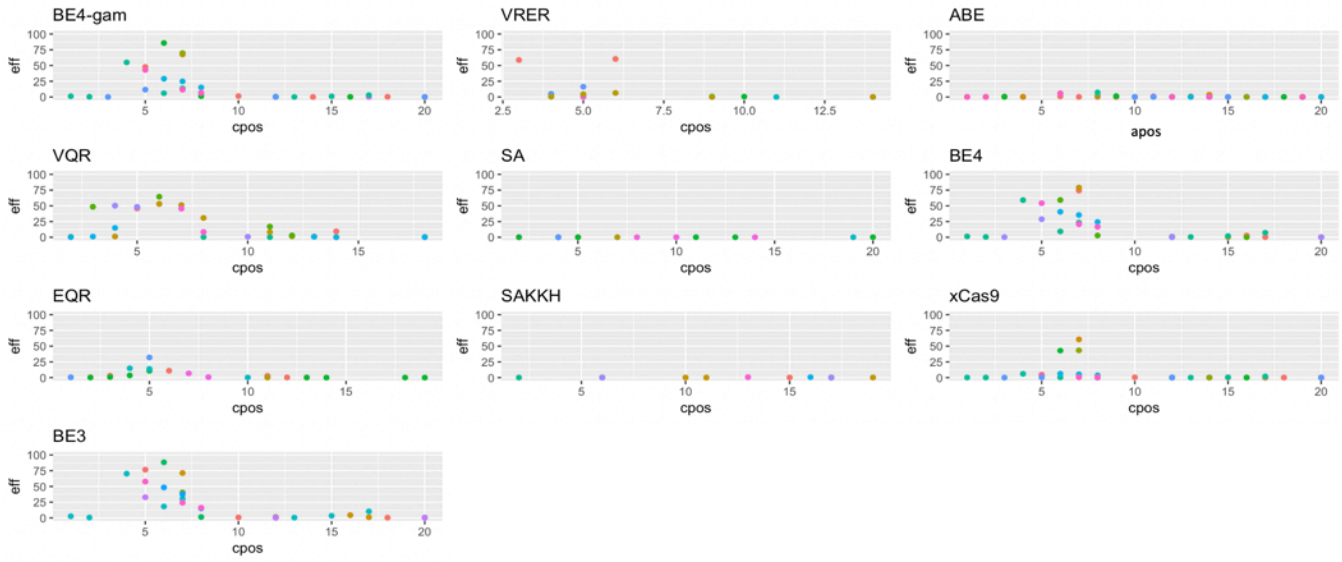


Figure 3.8 The editing efficiency of the C at different positions for ten different base editors.

The x-axis showed the position of C on the gRNA, and the y-axis showed the efficiency of C convert into T at that specific position. For ABE, x-axis is the position of A, and y-axis is the efficiency for A convert into G. Each color represents one gRNA. The dots with the same color in the same base editor mean the multiple C (or A) on that gRNA.

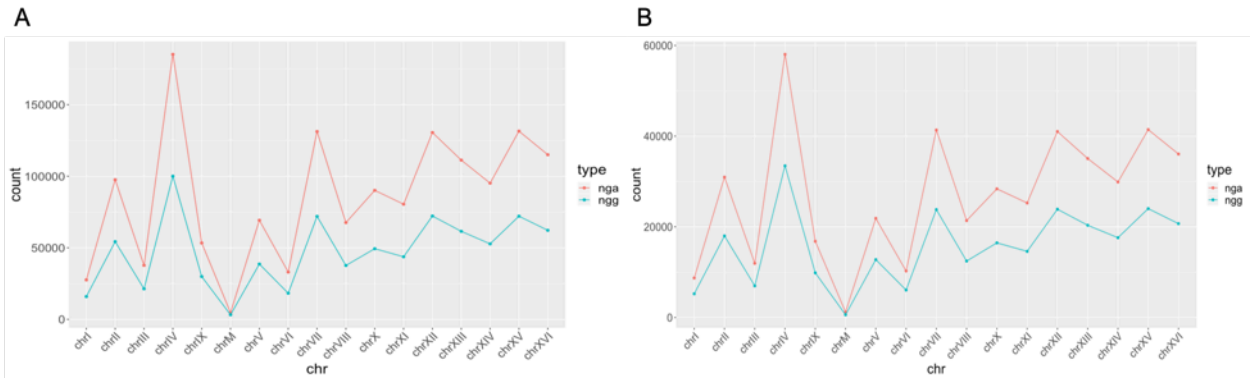


Figure 3.9 The number of targetable regions of NGG and NGA PAM base editors.

(A) The number of sites with NGG PAM (in green) or with NGA PAM (in red) on every chromosome. (B) The number of gRNAs with NGG PAM containing at least one C in the targeting window 4-8 for BE (in green) or with NGA PAM containing at least one C within the targeting window 4-11 for VQR (in red) on every chromosome.

Table 3.1 Different base editors tested and their corresponding recognition PAM sites and base editing characteristics.

BE3	NGG	C to T	Kim et al. (Liu lab)	Nature biotech.
VQR-BE3	NGAN	C to T	Kim et al. (Liu lab)	Nature biotech.
EQR-BE3	NGAG	C to T	Kim et al. (Liu lab)	Nature biotech.
VRER-BE3	NGCG	C to T	Kim et al. (Liu lab)	Nature biotech.
SaBE3	NNGRRT	C to T	Kim et al. (Liu lab)	Nature biotech.
SaKKH-BE3	NNNRRT	C to T	Kim et al. (Liu lab)	Nature biotech.
ABE7.10	NGG	A to G	Gaudelli et al. (Liu lab)	Nature
xCas9(3.7)-BE3	NGG, NG,GAA,GAT	C to T	Hu et al. (Rees lab)	Nature
BE4	NGG	C to T	Komor et al. (Liu lab)	Sci. Adv.
BE4-Gam	NGG	C to T	Komor et al. (Liu lab)	Sci. Adv.

The five columns represent the base editor name, the PAM site, the editing pattern, the authors for the paper discovered or engineered that base editor, the journal first reported that base editor.

Table 3.2 Two example gRNAs with multiple Cs showed the editing efficiency at different positions of BE3.

gRNA with 4Cs	AATTCACCGTAATATTTGACAGG
Position of C on the gRNA	Percentage of C convert into T
5	57.7
7	24.1
8	16.1
20	0.1
gRNA with 9Cs	CCTCGCCATTTACTCTCGTCGGG
Position of C on the gRNA	Percentage of C convert into T
1	2.6
2	0.5
4	70.3
6	18.1
7	30.2
13	0.4
15	3.4
17	10.6
20	0.5

Table 3.3 Two example gRNAs with multiple Cs showed the editing efficiency at different positions of base editor VQR.

gRNA with 3Cs	GGTCCTTGACAGGAATTTAGGAG
Position of C on the gRNA	Percentage of C convert into T
4	50.2
5	47.8
10	0.85
gRNA with 9Cs	CAACACCCGTCCACTTTCTTGGAG
Position of C on the gRNA	Percentage of C convert into T
1	0.0
4	1.3
6	53.0
7	51.2
8	30.8
11	8.2
12	1.5
14	0.0
18	0.5

References

1. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014. pp. 1262–1278. doi:10.1016/j.cell.2014.05.010
2. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014. doi:10.1126/science.1258096
3. Larson MH, Gilbert LA, Weissman JS, Qi LS, Arkin AP, Lim WA, et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell. 2013;152: 1173–1183. doi:10.1016/j.cell.2013.02.022
4. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. XCRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154: 442. doi:10.1016/j.cell.2013.06.044
5. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, et al. Editing DNA Methylation in the Mammalian Genome. Cell. 2016;167: 233-247.e17. doi:10.1016/j.cell.2016.08.056
6. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016; doi:10.1038/nature17946
7. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, et al. Programmable base editing of T to G C in genomic DNA without DNA cleavage. Nature. 2017; doi:10.1038/nature24644
8. Kim YB, Komor AC, Levy JM, Packer MS, Zhao KT, Liu DR. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat Biotechnol. 2017; doi:10.1038/nbt.3803

9. Komor AC, Zhao KT, Packer MS, Gaudelli NM, Waterbury AL, Koblan LW, et al. Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. Sci Adv. 2017; doi:10.1126/sciadv.aao4774

10. Miller CA, Martinat MA, Hyman LE. Assessment of aryl hydrocarbon receptor complex interactions using pBEVY plasmids: Expression vectors with bi-directional promoters for use in Saccharomyces cerevisiae. Nucleic Acids Res. 1998; doi:10.1093/nar/26.15.3577

11. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014; doi:10.1038/nmeth.3047

12. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science (80-). 2014; doi:10.1126/science.1247005

13. Gietz RD, Schiestl RH. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc. 2007;2: 35–37. doi:10.1038/nprot.2007.14

14. Liu G, Lanham C, Buchan JR, Kaplan ME. High-throughput transformation of Saccharomyces cerevisiae using liquid handling robots. PLoS One. 2017;12. doi:10.1371/journal.pone.0174128

15. Gietz RD, Schiestl RH. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc. 2007;2: 38–41. doi:10.1038/nprot.2007.15

16. Rees HA, Liu DR. Base editing: precision chemistry on the genome and transcriptome of living cells. Nature Reviews Genetics. 2018. doi:10.1038/s41576-

018-0059-1

17. Kleinstiver BP, Prew MS, Tsai SQ, Topkar V V., Nguyen NT, Zheng Z, et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature. 2015; doi:10.1038/nature14592

18. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015; doi:10.1038/nature14299

19. Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature. 2018; doi:10.1038/nature26155

Chapter 4 Genome-scale precision engineering of *Saccharomyces cerevisiae* with trackable integrated genomic barcodes

Abstract

Genome-scale engineering enables testing the effect of genetic variants by producing genome-wide mutations in parallel. One strategy for genome-wide engineering is precision editing with CRISPR-Cas9 system and measuring the functional effect of the introduced variants. Recent studies have attempted to use CRISPR libraries to generate single variant mutations genome-wide with fitness phenotyping through plasmid barcode readouts. However, previous assays suffered from noises of variant engineering efficiency variation along the genome, low functional activity of Cas9, as well as the potential plasmid barcode loss during phenotyping. Here we aimed to construct a CRISPR-library-based approach for highly efficient genome-wide variant engineering with trackable genomic integrated barcode and a canavanine selection step to enrich for functional Cas9 system in *Saccharomyces cerevisiae*. Our library construction contained two gRNAs with two repair templates, one pair of gRNA and repair template introduce the designed precision editing into the genome in a library mode, and the other pair of gRNA specifically targeted gene *CAN1* and the repair template introduced a unique barcode to the targeting region for genomic integration. Yeast were selected on canavanine plate, where only yeast that had the successful cutting and template repairing at gene *CAN1* could grow and survive. This step ensured

the functional highly active Cas9 and a functioning repair pathway. This genomic integrated barcode system ensures highly efficient high-throughput genetic engineering, as well as a robust downstream phenotyping.

Introduction

Understanding the functional effects of genetic variation is one of the fundamental challenges in understanding phenotypic diversity, biology, evolution and genetic diseases. Originally the phenotypic effects of natural genetic variation have been studied via genetic mapping within large populations. An important genetic mapping approach is quantitative trait locus (QTL) mapping, where associations between the genetic variant and phenotype variation are measured with the population of hundreds or thousands of offspring from genetic crosses [1][2]. The mapping resolution of genetic approaches is limited by recombination events, and very few phenotypes have been resolved to causal variants. Pinpointing causal variant requires laborious follow-up experiments, and researchers tend to focus on well-annotated genes and missense variants, creating a bias towards the genes and variants with known functions.

In recent years, the CRISPR-Cas9 genetic engineering tool enables us to characterize individual genetic variants via genetic engineering [3]. Precise genetic engineering can be introduced to the genome utilizing a donor DNA template for the homology-directed repair (HDR). A library of donor DNAs, encoding desired mutations, could be co-transformed into yeast cells with the single guide RNAs (sgRNAs). Previous

study had combine the gRNA and donor DNA on the same plasmid to achieve the corresponding pairs during transformation [4]. However, the efficiency of precise editing at the targeting region is still limited and variable across different sites [5].

To boost editing efficiency and enable multiplexed high-throughput screens, we developed a CRISPR-Cas9 library-based method with trackable integrated cellular barcodes and an enrichment selection for functional Cas9 cutting and homology directed repairing (HDR). First, generating uniquely edited cells required the cells to receive the correct pair of gRNA and repair template. We devised an approach that accomplishing such pairing by encoding gRNA sequence and the corresponding repair template on the same plasmid construction. Furthermore, we engineered another pair of gRNA and repair template on the same plasmid to introduce the genomic integrated barcode. This gRNA targets gene *CAN1*, a gene encodes a plasma membrane arginine permease. The repair template contained 40 bp on each side of the homology arm of gene *CAN1* and a 30 bp barcode, which uniquely represent the first pair of gRNA and repair template for variant editing. Integration of the barcode through HDR will disturb the function of gene *CAN1*, making the yeast resistant to the toxic arginine analog canavanine. Transformed yeast is selected for gaining a unique integrated barcode using the canavanine selection. The construction enables precise genome editing with high efficiency and throughput, allowing us to pinpoint the functional effects of variants.

Materials and Methods

Yeast strains and media

Saccharomyces cerevisiae BY4741 (haploid, *MAT* a, *his3*Δ*1, leu2*Δ*0, met15*Δ*0, ura3*Δ*0*) was used as host strain for genome editing. Cells were grown non-selectively in YPAD medium (2% Bacto peptone, 1% Bacto yeast extract, 2% glucose). Selection of yeast transformants based on the *URA*3 and *LEU*2 markers was done on a synthetic complete (SC-U-L) medium (6.7 g/L of Yeast Nitrogen Base, 2% glucose, 0.54g/L Complete Supplement Mixture-Ura-Leu). For culture in Petri dishes, the medium was solidified with 2% agar. Yeast strains were incubated at 30 °C and liquid culture were growing on a shaker. Selective canavanine plates were made from arginine minus synthetic complete agar medium with 60mg/liter L-canavanine (Sigma C1625).

Fixed gRNA targeting CAN1 plasmids construction

The plk88 plasmid was digested by restriction enzyme NaeI and BtgZI. Digested product was heat-treated and gel purified to eliminate the enzyme activity and remove the undigested plasmid. A gblock contained the SNR52 promoter, gRNA and terminator region for a functional gRNA targeting *CAN1* was synthesized with the overlapping region of plk88 for later Gibson assembly purpose (38bp overlapping at the NaeI site and 52 bp overlapping at the BtgZI site). Gblock was amplified using Phusion high fidelity polymerase. Gibson assembly was applied to ligate the gblock with the digested linear plasmid. Ligated product was transformed into bacteria, and plamids were extracted for Sanger sequencing to confirm the construction of *CAN1* gRNA.

Library-based gRNA and donor DNA plasmids construction - first round of cloning

To generate edit-directing plasmids, we synthesized DNA fragments carrying the desired gRNA as well as the sites for *BbsI* restriction enzyme cutting as DNA oligos (Integrated DNA Technologies), which were then cloned into pOS05a (pRS426 SNR52p-lentiCR2-gRNA-filler-SUP4t), a plasmid for expressing gRNAs under an SNR52 promoter, through T4 ligation. We designed and synthesized 10 gRNAs that targeted three different genes in the genome: 4 gRNAs targeting gene *CAN1* on chromosome V, 3 gRNAs targeting gene *CBS1* on chromosome IV, 3 gRNAs targeting gene *CCR4* on chromosome I.

1) Library amplification from chip oligos

2X Master Mix (Kapa LibAmp kit)	200 uL
Synthesized library (oMS263)	0.16ul
Forward primer (oMS246)	1.6 ul
Reverse primer (oMS267)	1.6 ul
ddH2O	196.6 ul
Total	399.96 ul

Split into 20 reactions of 20 uL in a qPCR plate, then amplified with the following reaction:

98 ºC	2 min	1
98 ºC	3 sec	5
50 ºC	20 sec	5

60 ºC	10 sec	5
98 ºC	3 sec	13
54 ºC	20 sec	13
72 ºC	10 sec	13
72 ºC	2min	1

Used the Qiagen PCR purification kit to clean up. Eluted in 40 uL.

2) Digesting the amplified library

Purified PCR	40 ul
BstEII	1 ul
EagI	1 ul
CutSmart buffer	5 ul
ddH2O	3 ul
Total	50 ul

Incubate at 37ºC for 30 minutes

Clean up with the Qiagen PCR purification kit, and digest again

Purified PCR	40 ul
BstEII	1 ul
EagI	1 ul
CutSmart buffer	5 ul
ddH2O	3 ul
Total	50 ul

Incubate at 37ºC for 30 minutes

3) Digest the plasmid

Start with Maxiprepped 200 mL of pMS52. Quantify.

pMS52 change it to mine plasmids	20 ug
BstEII	10 ul
EagI	10 ul
CutSmart buffer	50 ul
ddH2O	X ul
Total	500 ul

Incubate at 37ºC for 2 hours

Add 10 uL rSAP, give another 1.5 hrs at 37C

PCR purified in 3x50 uL water to elute, quantify.

4) Set up ligation reaction

digested plasmid from step3	1ug
digested insert from step2	800ng
T4 DNA liagase 10X buffer	20 ul
T4 DNA liagase	4 ul
ddH2O	X ul
Total	200 ul

Incubate at room temperature for 10 mins.

65C 5 min

5) E.coli electroporation transformation

Competent cells (Lucigen Supreme DUO 60080-2) were thawed on ice for about 10 min. Cells were transferred into a cold tube and plasmids were added in and mixed well. Then the plasmid cell mix was transferred into a cuvette for electroporation.

After electroporation with the Ec1 setting, 1ml LB was added into the cuvette to recover the transformed cells, followed by incubation at 37ºC for 1 hour. The culture was dumped into LB+Amp medium to grow later for enough plasmids.

Library-based gRNA and donor DNA plasmids construction – second round of cloning

1) Library amplification from chip oligos

2X Master Mix (Kapa LibAmp kit)	200 uL
Synthesized library (oMS263)	0.16ul
Forward primer (oMS246)	1.6 ul
Reverse primer (oMS267)	1.6 ul
ddH2O	196.6 ul
Total	399.96 ul

Split into 20 reactions of 20 uL in a qPCR plate, then amplified with the following reaction:

98 ºC	2 min	1
98 ºC	3 sec	5
50 ºC	20 sec	5
60 ºC	10 sec	5
98 ºC	3 sec	13
54 ºC	20 sec	13
72 ºC	10 sec	13

72 ºC	2min	1

Used the Qiagen PCR purification kit to clean up. Eluted in 40 uL.

2) Digesting the amplified library

Purified PCR	40 ul
BstEII	1 ul
EagI	1 ul
CutSmart buffer	5 ul
ddH2O	3 ul
Total	50 ul

Incubate at 37ºC for 30 minutes

Clean up with the Qiagen PCR purification kit, and digest again

Purified PCR	40 ul
BstEII	1 ul
EagI	1 ul
CutSmart buffer	5 ul
ddH2O	3 ul
Total	50 ul

Incubate at 37ºC for 30 minutes

3) Digest the plasmid

Start with Maxiprepped 200 mL of pMS52. Quantify.

pMS52 change it to mine plasmids	20 ug
BstEII	10 ul
EagI	10 ul

CutSmart buffer	50 ul
ddH2O	X ul
Total	500 ul

Incubate at 37ºC for 2 hours

Add 10 uL rSAP, give another 1.5 hrs at 37C

PCR purified in 3x50 uL water to elute, quantify.

4) Set up ligation reaction

digested plasmid from step3	1ug
digested insert from step2	800ng
T4 DNA liagase 10X buffer	20 ul
T4 DNA liagase	4 ul
ddH2O	X ul
Total	200 ul

Incubate at room temperature for 10 mins.

65C 5 min

Yeast transformation and genomic DNA extraction

Yeast cells were transformed with the LiAc/SS carrier DNA/PEG method using 0.5–1 μg plasmid DNA[6]. Transgenic clones were selected on SC-U-L media. Single colonies were picked and grown in galactose selective liquid medium (6.7 g/L of Yeast Nitrogen Base, 2% galactose, 0.54g/L Complete Supplement Mixture-Ura-Leu) overnight to induce Cas9 base editor expression. Colonies were allowed to grow for approximately 24 h in the galactose selective medium. Genomic DNA was extracted from 1ml of the harvested cells from galactose culture with a DNeasy Blood and Tissue

kit (Qiagen). In parallel, 1 ml of the culture was plated onto the canavanine plates to gain an estimate of number of colonies growing. canavanine plates. Plates were incubating at 30ºC for two days and images of plates were taken using the scanner.

Canavanine selection for gene *CAN1*

Yeast colonies were picked, suspended in 3 mL SC medium with 2% glucose and without leucine and uracil, and grown to a stationary phase. The cells were then pelleted, washed twice in sterile water, and then resuspended in SC induction medium with 2% galactose and 1% raffinose, but without leucine and uracil, to an OD600 of 0.3. The cells were incubated for 20 h prior to plating on YPAD rich or SC media plates without arginine but with 60 mg/mL l-canavanine (Sigma). After incubation for 3 days, the colony number on each plate was counted. The C-to-T mutation frequency in CAN1 was determined as the ratio of the colony count on canavanine-containing plates to the colony count on YPAD-rich media plates. Each experiment was performed at least three times on different days.

Amplicon sequencing and data analysis

Yeast colonies harboring plasmids expressing base editors and sgRNAs were picked from SC-L-U plates, suspended in 3 mL SC-L-U medium with 2% glucose, and grown to a stationary phase. The cultures were then washed twice to remove residual glucose, resuspended in 5 mL SC-L-U medium with 2% galactose and 1% raffinose to an OD600 of 0.3, and incubated for 20 h at 28 °C on a rotary shaker. Genomic DNA was

extracted from culture samples of 0.5 mL volume, and the regions targeted by base editing were amplified by PCR with primer pairs containing index tags for sample multiplexing. PCR amplification was performed with the Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) according to the manufacturer's protocol. The amplicon sequencing library was prepared using the Illumina Nextera DNA Library Prep Kit with the adjusted protocols to skip the nextera treatment. The library was then sequenced on the MiSeq platform using the MiSeq Reagent V2 PE150X2 Kit. Then we processed reads with the following pipeline: read pairing (PEAE), read trimming (Trimmomatic-0.36), and read alignment (BWA) to the reference-targeting region using the down-sampled fastq files. Custom R codes were used to detect edits at the targeted sites.

Results

Construct a genome editing system with trackable integrated barcodes in yeast

Several methods have begun to explore the high precision genome-wide variant engineering, for example, researchers placed the gRNA and the corresponding repair donor DNA together on a same plasmid in cis on oligonucleotides generated in bulk through high-throughput synthesis to explore the consequences of premature-termination codons (PTCs) at different locations on yeast essential genes [4]. However, the editing efficiency was highly variable at different genomic regions, and confounded by the loss of plasmid barcodes and the plasmids copy number variations. The plasmid

barcode construct introduces a high level of noise into the downstream phenotyping assay, especially when measuring the effects of natural variants, whose effects are expected to be small. Therefore, we aim to construct a system with selective genome integrated barcode to prevent plasmid barcode loss and to enable robust phenotyping.

We optimized from the plasmid structure used in [4], where there was a gRNA driven under the SNR52 promoter and a corresponding donor DNA for it. We constructed a second gRNA structure, which was a fixed gRNA with known high efficiency targeting the gene *CAN1*, on the same plasmid. A yeast Histamine auxotrophic marker, a bacteria Kanamycin resistant marker and a 40bp left homology arm of the *CAN1* targeting region were synthesized using IDT glock. The 40bp right homology arm with 30 bp random barcode were synthesized and linked to the IDT glock using Phusion PCR. The amplified product harboring the region from the Histamine marker to the right homology arm of *CAN1* was cloned into the Ampicillin resistant backbone where one gRNA and donor DNA pair has been embedded (Figure 4.1A). The successfully constructed plasmid was transformed into bacteria and selected with both Ampicillin and Kanamycin. All ligation regions were verified using Sanger sequencing. In general, the complete construction with gRNA library contained two steps of cloning: one step is to insert the synthesized gRNA and donor DNA pairs into the backbone, and the other step is to insert the region with the homology arms, the barcode and two selective genes into the existing system. Plasmids were then extracted from the plates that maintained selection for both AmpR and KanR markers (Figure 4.1B).

Evaluating the editing efficiency in individual gRNA and donor DNA pairs

In order to gain a detailed idea of the editing efficiency using the new system, we chose 18 gRNA and donor DNA pairs by picking single colonies from the previous transformation selective agar plates. All 18 plasmids were sequenced for the repair template region, and the targeting gRNA region was inferred from the design table guided for oligo synthesize. The plasmids were independently transformed into yeast in separate tubes, which already contained a Gal-promoter Cas9 plasmid with the Leucine auxotrophic marker (Figure 4.2A). Yeast successfully transformed were selected on selective plates lacking Leucine, Uracil and Histamine and we observed a large number of yeast colonies at this step, indicating high transformation efficiency. One single colony of each gRNA donor DNA pair was picked into the selective medium with Galactose, followed by pouring onto Canavanine agar plates to enrich for yeast with the *CAN1* gene loss of function from integrated barcode. For 18 tested plasmids, we observed large number of canavanine resistant colonies on the plates, indicating a functional gRNA targeting gene *CAN1* (Figure 4.2B). This step confirmed the working Cas9 system could potentially enrich for yeast cells with editing at the other gRNA targeting site.

High efficiency genome editing achieved at a site dependent manner

To study what is the editing efficiency and editing patterns in selected yeast, we analyzed the amplicon sequencing results from targeted sites in canavanine resistant colonies. The upstream and downstream of gRNA targeting region for each pair of gRNA and donor DNA was amplified and analyzed. For part of our donor DNAs, there were two variants compared to the reference genome: one variant blocking the PAM site to avoid continuous cutting, and the other variant to introduce the desired mutation. In the colonies we gained confident reads half of them had precise variant editing at the desired sites and the PAM site-blocking site. For the examples shown in Figure 4.3, we can see the variant designed to block the PAM was edited at a very high efficiency, and the desired mutation variant occurred at almost the same high efficiency. The editing of both variants at a time indicates low frequency of crossover occurrence between the two designed variants. We observed high efficient precise editing at the donor DNA that only contained one variant to block PAM site (Figure 4.4). There were also some unfortunate cases that no editing was detected (Figure 4.5), which could be due to the low gRNA targeting efficiency for those specific gRNAs, or low HDR rate at the targeted site. In general, more than half of the tested cases had the precise editing at the desired site, as well as gaining intact integrated genomic barcode at the *CAN1* gene region.

Discussion

In order to improve the signal-to-noise properties of high-throughput editing in yeast, we built a CRISPR-Cas9 genetic engineering system with two pairs of gRNA and donor DNA: one aims to introduce the desired mutation via targeting the genome and

replacing with a repair template donor DNA, the other functions as a fixed gRNA targeting gene *CAN1* and introducing a genomic integrated barcode system to the cutting site. This system can be easily applied with DNA oligo library to generate thousands of editing in a pool at a time. To gain a detailed understanding of the editing process, we analyzed the editing efficiency and patterns for 18 pairs of gRNA and donor repair template. Our results show that the improved CRISPR-Cas9 repaired template system with canavanine enrichment process is an effective tool to introduce precise genetic editing in yeast genome to study phenotypic effect of genetic variants.

For some sites tested, we observed that editing efficiency is not as high as we expected, and several reasons could explain this phenomenon. One major reason that showed large effect in multiple previous studies is the gRNA targeting efficiency. It has been known that gRNA efficiency varies at different positions of the genome: for example, the open chromatin regions may have higher targeting efficiency due to easier accessibilities for gRNAs. Another factor that can affect the genetic editing efficiency is the homology directed repair (HDR) rate. Researchers have tried to improve the editing efficiency by improving HDR. In addition to the above two reasons, errors during gRNA and donor DNA synthesis can affect the editing efficiency due to smaller affinity of overlapped regions. Due to the length between the gRNA and donor DNA repair template is more than 2kb, we could not gain the sequence of both from one Sanger sequencing. We inferred the gRNA sequence by looking up the Sanger sequencing results of the donor DNA repair template in the designed table of gRNA and donor DNA pairs. Mistakes in our inference of the gRNA sequences based on the repair template

sequences can also mislead us to check the wrong region for efficiency estimate. It is also possible that a few yeast colonies gained resistance to canavanine due to spontaneous mutations in gene *CAN1*, and these colonies without barcodes at *CAN1* overtaken the canavanine resistant population instead of the yeast with highly active Cas9 system and integrated barcode.

To further investigate how precise and robust the editing system is in high-throughput library screens, we performed library-based genome editing with canavanine plate enrichment. We have picked 192 single colonies from the canavanine plates and perform the whole genome sequencing on the DNA from chosen colonies. The sequencing is in process, and the whole genome sequencing results will inform us about the spectrum of the editing efficiency across the genome.

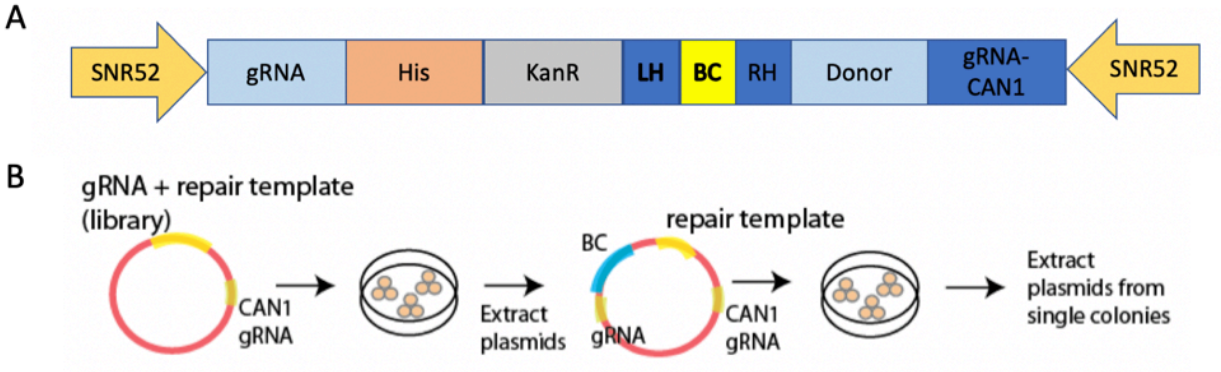


Figure 4.1 Constructing plasmids for genome editing system with trackable genomic barcodes.

(A) The schematic shows the pairing of CRISPR gRNA and repair template on plasmids, the auxotrophic His marker for yeast and the Kanamycin resistant marker for bacteria, the barcoding region with homology arms, and another gRNA targeting gene *CAN1*. LH stands for left homology arm; BC stands for barcode; RH represents right homology arm.

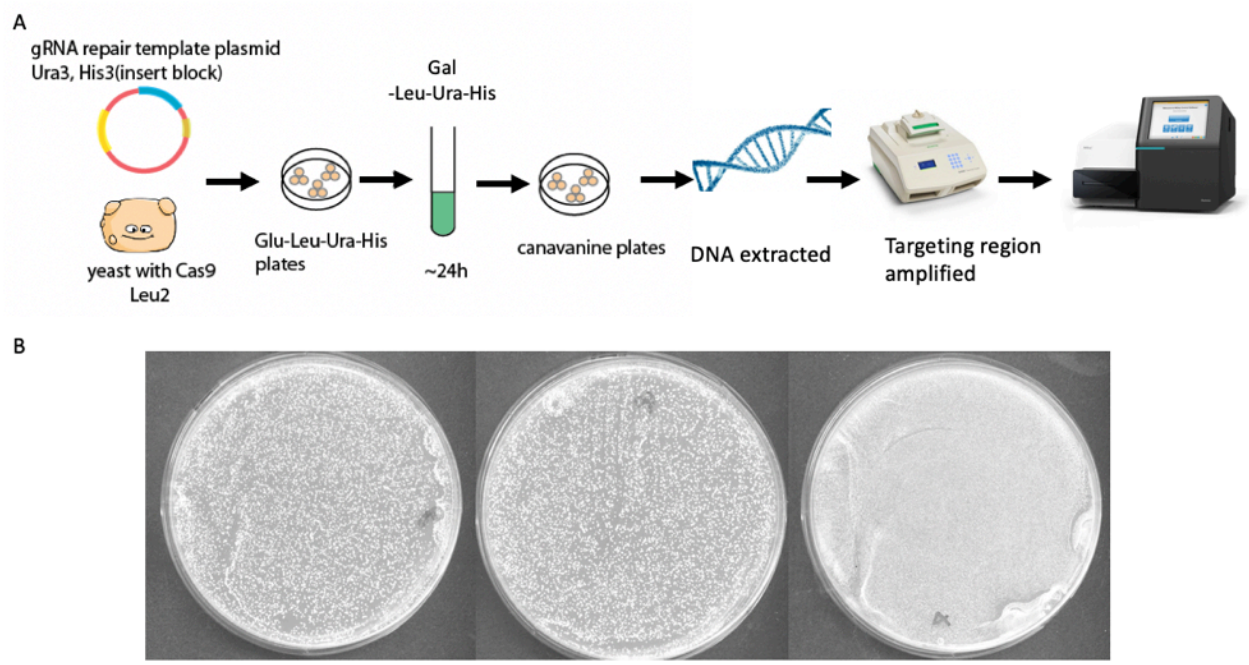


Figure 4.2 Evaluating the editing efficiency of the CRISPR-Cas9 system with the genomic integrated barcoding system.

(A) The schematic shows the transformation, selection, enrichment and amplicon sequencing process for evaluating the editing efficiency. (B) Example plates of yeast colonies resistant to canavanine selection after edited by CRISPR with two gRNAs and two donor DNA pairs, where one donor DNA aimed to insert the barcode.

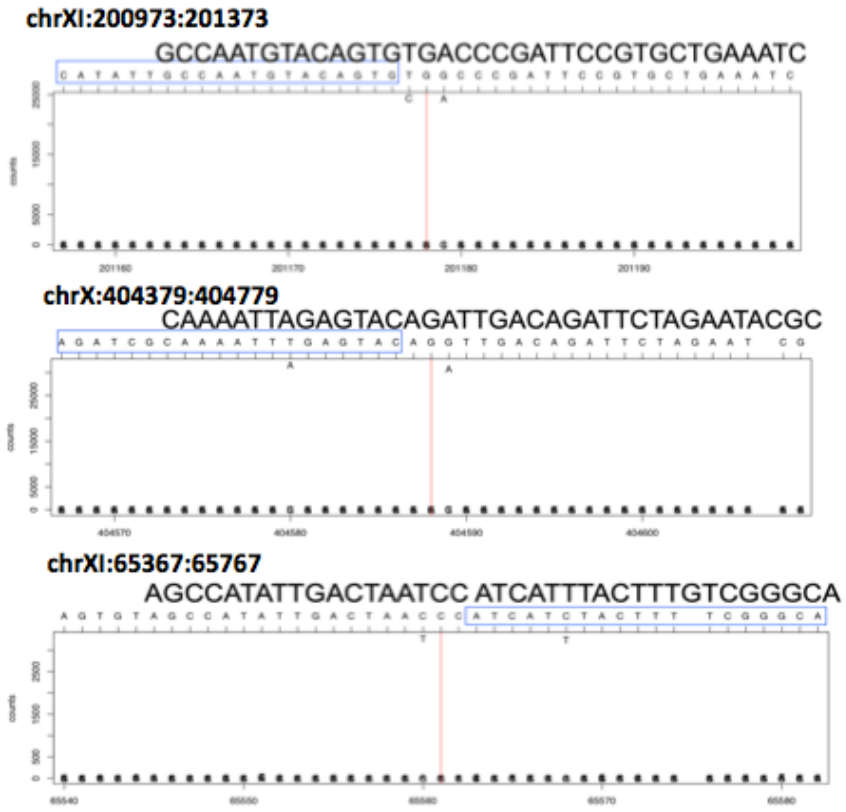


Figure 4.3 Example cases where the editing system is highly efficient and precise for introducing two variants into the genome at a time.

Here we show three examples of the editing patterns at gRNA targeting region. Both the variant to block PAM site and to introduce desired mutation are edited. The red vertical line indicates the PAM site. The gRNA is circled out using the blue rectangular shape. The repair template's sequences are shown above the gRNA sequences.

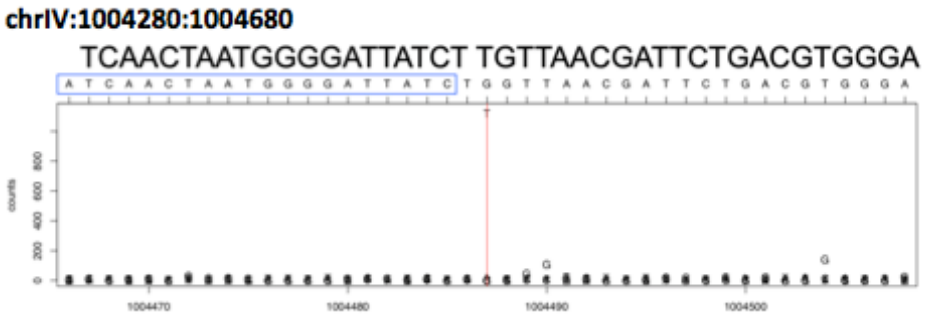


Figure 4.4 An example case where the editing system is highly efficient and precise for introducing one variant into the genome.

This is an example of the case where only the one variant to block the PAM site is successfully edited. The red vertical line indicates the PAM site. The gRNA is circled out using the blue rectangular shape. The repair template's sequences are shown above the gRNA sequences.

Figure 4.5 Example cases of no desired editing occurred.

Here are two examples that no variants are introduced into the targeting region. The red vertical line indicates the PAM site. The gRNA is circled out using the blue rectangular shape. The repair template's sequences are shown above the gRNA sequences.

References

1. Ehrenreich IM, Gerke JP, Kruglyak L. Genetic dissection of complex traits in yeast: Insights from studies of gene expression and other phenotypes in the BYxRM cross. Cold Spring Harbor Symposia on Quantitative Biology. 2009. doi:10.1101/sqb.2009.74.013
2. Bloom JS, Ehrenreich IM, Loo WT, Lite T-LV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. Nature Publishing Group; 2013;494: 234–237. doi:10.1038/nature11867
3. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. Narnia; 2013;41: 4336–4343. doi:10.1093/nar/gkt135
4. Sadhu MJ, Bloom JS, Day L, Siegel JJ, Kosuri S, Kruglyak L. Highly parallel genome variant engineering with CRISPR-Cas9. Nat Genet. 2018; doi:10.1038/s41588-018-0087-y
5. Liu Z, Liang Y, Ang EL, Zhao H. A New Era of Genome Integration - Simply Cut and Paste! ACS Synthetic Biology. 2017. doi:10.1021/acssynbio.6b00331
6. Gietz RD, Schiestl RH. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc. 2007;2: 35–37. doi:10.1038/nprot.2007.14

Chapter 5 The construction of CRISPR-directed mitotic recombination mapping panel in human cell lines

Abstract

Linkage and association studies have mapped out thousands of regions contribute to a wide variety of phenotypic variation, but narrowing these regions down to underlying causal genes remains challenging and requires laborious follow-up experiments. The new genetic engineering tool CRISPR-Cas9 can programmably create DNA double strand breaks (DSBs) in the genome. Followed by DSB, a phenomenon called loss of heterozygosity (LOH), which can generate new genotype that is homozygous from the break all the way to the telomere, can occur at a low frequency. Previous work has proven the feasibility of fine mapping by building a LOH mapping panel in yeast [3]. Implementing the LOH mapping panel in human cell lines will allow rapid fine mapping of a wide range of traits and diseases without a large mapping population and without the limitation of natural recombination rates. The mapping panel also empowers us to identify more complex effects of haplotype blocks and potentially large structural variants that may be inaccessible to knockout or variant editing screens. Here we worked on developing a CRISPR-directed mitotic recombination-mapping panel in human H9 embryonic stem (ES) cell line. We built a heterozygous selective marker on chromosome 16 by knocking out one functional allele of gene *APRT*. LOH events could be enriched by using 2,6-Diaminopurine (2,6-DAP) selection for the homozygous *APRT* marker. We chose eight gRNAs with high targeting

efficiency to generate a second double strand breaks between the selective marker and telomere for LOH panel. None of the cells survived after the selection, indicating the potential low frequency of LOH events in human ES cell lines.

Introduction

Finding the genetic variants underlying trait variation is the central goal of a large segment of basic, medical, and industrial research. The primary tools in this search are genome-wide association studies (GWAS) and linkage mapping. Linkage mapping will find a section of DNA that correlates with variation in a phenotype, which are termed quantitative trait loci (QTLs). QTLs often span many genes, and narrowing them to the specific underlying genes and genetic variants can be difficult. One common approach for fine mapping the variants is to increase the number of samples, as more recombination events are likely to be observed in large sample size. This approach will be limited by the cost, and the genetic loci with low recombination rates will remain low resolution for mapping despite the increase of the sample size. Another approach is to interrogate promising candidate genes based on known gene function. However, studying candidate genes will miss contributions from unexpected sources, and thus is difficult to apply to poorly understood genes and traits.

The recent developed gene-editing tool CRISPR-Cas9 accelerates and simplifies the process of genetic mapping. CRISPR-Cas systems, originally described as adaptive immune systems in bacteria and archaea, have become popular tools for genome modification in recent years [1]. The Cas9 nuclease creates DNA double strand breaks

at the specific locations defined by an RNA duplex, known as guide RNA (gRNA) [2]. gRNAs contains 20 base pairs that matches the desired target sequence follows by a structure sequence called protospacer-adjacent motif (PAM) for Cas9 binding. In heterozygous diploid individuals, if the cutting occurs at one chromosome, the other intact sister chromatid can serve as a template for repair by homologous recombination. This process will generate a cell with new genotype that is completely homozygous from the targeted cutting site to the telomere and unchanged heterozygous else where. The above event is termed as “loss of heterozygosity” (LOH) [3].

Sadhu *et al.* developed a method that uses CRISPR to build mapping panels with targeted recombination events. They tested the method by generating a panel with recombination events spaced along a yeast chromosome arm, mapping trait variation, and then targeting a high density of recombination events to the region of interest. Using this approach, they fine-mapped manganese sensitivity to a single polymorphism in the transporter Pmr1. Targeting recombination events to regions of interest enables a rapidly and systematically way of identifying causal variants underlying trait differences [3].

We aim to develop a novel mapping strategy in which the recombination sites are directly chosen by targeting recombination events with the CRISPR/Cas9 system in human cell lines. Once developed, this method will enable researchers to generate the study populations from cultured cells in vitro, allowing traits to be mapped in a single individual rather than large cohort of samples, which will minimize the confounding

factors that was introduced by differences between individuals regarding to environment and ancestry background. This method could help identify causal variants in recombination cold spots. Furthermore, by targeting multiple sites in a small region of interest will dramatically increase the resolution of mapping without collecting increasingly larger cohorts of humans.

To generate a specific sequence change or a specific loss of heterozygosity (LOH) event, cells must undergo an essential process called homology directed repair (HDR), a distinct cellular DNA repair pathway. We could potentially increase the LOH at the gRNA-targeting region by improving the HDR rate and minimizing the undesired non-homologous end joining (NHEJ). Different strategies have been applied to increase HDR rate in past literatures, including the following: (1) Use single strand DNA as template instead of commonly used double strand DNA templates. Richardson *et al.* increase the rate of HDR in human cells to up to 60% by rationally designing single-stranded DNA (ssDNA) donors of the optimal length complementary to the strand that is released first [4]. (2) Control the timing of CRISPR/Cas9 delivery by cell synchronization. It shows that cells enriched in G2/M phase have a 6-fold increase in targeting (HDR) efficiency compared to cells in G1 phase. Therefore reversible synchronization using Nocodazole or ABT-751 can improve HDR [5]. (3) Suppress the non-homology end joining (NHEJ) pathway. Suppressing the NHEJ key molecules KU70, KU80 or DNA ligase IV by gene silencing can increase HDR. Studies that suppressed KU70 and DNA ligase IV reveal a 4-5 fold HDR efficiency increase [6]. DNA ligase IV inhibitor SCR7 can suppressing the NHEJ pathway and improve HDR rate [6][7]. (4) Enhance HDR.

Previous research show that HDR enhancer RS-1 can increase the knock-in success rates [8]. (5) Use ribonucleoprotein (RNP) complexes instead of regular plasmids based Cas9 transformation. When combining well-established cell cycle synchronization techniques with direct nucleofection of pre-assembled Cas9 ribonucleoprotein (RNP) complexes achieved HDR rates up to 38% in HEK293T cells [9]. (6) Co-express other proteins. Co-expression of adenovirus E1B55K and E4orf6 proteins is reported to increase the HDR efficiency up to eightfold in both human and mouse cell lines[6].

Here we have successfully built up a stable cell line with heterozygous selective marker on chromosome 16 at gene *APRT* by knocking out one functional allele of that gene with CRISPR-Cas9. Eight gRNAs with high targeting efficiency in the test were chosen to create a second double strand break between the selective marker and telomere. We applied 2,6-DAP to enrich for cells undergone LOH events.

Methods and Material

Feeder-free human embryonic stem cell culture

Feeder-free human embryonic stem cell culture was maintained using mTeSR^{TM}1 medium (STEMCELL, Cat. No. 85850) on Matrigel (Corning Matrigel hESC-qualified Matrix, Cat. No. 354277) coated plates. The Matrigel was aliquot into 0.5mg for one 6 well plate. Thawed Matrigel aliquot was mixed immediately with 6ml DMEM/F-12 medium. Plates were coated with 1ml of Matrigel and DMEM/F12 mix overnight at 37ºC cell culture incubator. ES cells were fed with 2ml mTeSR^{TM}1 for each well in the 37ºC incubator with 5% of CO_2. ES cells were passaged using Versene (Thermo Fisher

scientific, Cat. No. 15040066), an EDTA solution for gentle non-enzymatic cell dissociation reagent. ES cells frozen stocks were generated by storing Versene harvested cells into the mFreSR^{TM} Cryopreservation Medium (STEMCELL, Cat. No. 05855).

When culture fresh thawed H9 ES cells and cells after nucleofection, ROCK inhibitor ((Y-27632 dihydrochloride, STEMCELL, Cat. No. 72302) was added into the mTeSR^{TM}1 medium to improve the attachment and recovery of cells. 10mM working stock solution of ROCK inhibitor was generated by diluting 1mg ROCK inhibitor into 295 μl. 1 μl of the working stock was added for each well of 6-well plate.

Feeder-based human embryonic stem cell culture

For feeder-based culture system, human embryonic stem cell line H9 was maintained and passaged on Mouse Embryonic Fibroblasts (MEF Feeder Cells). MEF were expanded using Dulbecco's modified Eagle's medium (DMEM, Thermo Fisher scientific, Cat. No. 10566016) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher scientific, Cat. No. A31605) and passaged using 0.05% Trypsin-EDTA (Thermo Fisher scientific, Cat. No 15400054). MEF cells were mitotically inactivated using mitomycin C for 4 hours and aliquots were frozen in 60% FBS, 20% DMSO and 20% MEF medium. Plates were pre-coated with 0.1% gelatin overnight at 37ºC incubator before plating MEF cells. MEF cells were plated in the plates at the density of $0.90x10^5$ cells/ml at 2.5 ml per well in a 6 well plate ($2.25x10^5$ cells/well). Stem cells could be plated in the plates one day after plating the MEF. The hES cell medium were

prepared as 800ml DMEM/F12 with Glutamax, 200ml Knockout Serum Replacer (KSR, Thermo Fisher scientific, Cat. No. 10828028), 10ml MEM Non-Essential Amino Acids Solution (Thermo Fisher scientific, Cat. No. 11140050) and 1ml Basic FGFsolution (bFGF, Thermo Fisher scientific, Cat. No. 13256029). H9 ES cells were maintained and fed with 2ml of the hES cell medium for each well. Medium were changed daily to keep the pluripotent character of the cells. H9 ES cells were passaged using 1mg/ml collagenase IV (Thermo Fisher scientific, Cat. No. 17104019) solution.

HEK293 cell culture and transfection

The HEK293 cells were maintained in Dulbecco's modified Eagle's medium (DMEM, Thermo Fisher scientific, Cat. No. 10566016) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher scientific, Cat. No. A31605), and 100 U/ml penicillin-streptomycin (Thermo Fisher scientific, Cat. No. 15140122) in a 5% CO2, humidified incubator. The HEK293 cells were subcultivated with the ratio 1:6 weekly by detaching and dispersing cells using Trypsin-EDTA (Thermo Fisher scientific, Cat. No 15400054). HEK293 cells were frozen in the complete medium with 10% DMSO. Plasmids were transfected into HEK293 using Lipofectamine 3000 Transfection Reagent (Thermo Fisher scientific, Cat. No. L3000008) following the manufacturer protocols.

Human stem cell Nucleofection

Nucleofection was done using Lonza human stem cell nucleofector kit 2 (Cat.No.VPH-5022). Confluent H9 cells was harvested using 1ml Accutase (STEMCELL

Cat.No. 07920) for each well of the 6-well plate and incubating at 37ºC for 5-10 min. Cells were dissociated into a single cell suspension by pipetting the suspension carefully up and down 4-6 times. Cell culture medium was added to stop Accutase. For each Nucleofection, it contained 8X10^5 cells, 1μg plasmid DNA and 100 μl human stem cell nucleofector solution 1 and 2. After harvesting, the cell density was determined by counting an aliquot of the detached cells. Then centrifuge the required number of cells (8 x 10^5 cells per sample) at 115xg for 3 minutes at room temperature. Cells were carefully suspended in 100 μl of Nucleofector solution and transferred into certified cuvette. The Nucleofector Program for the nucleofection was B-16. After transfection cells were incubated in a humidified 37°C/5% CO2 incubator until analysis.

gRNA plasmids construction

The gRNA plasmid used in this study was pSpCas9(BB)-2A-Puro(PX459)V2.0 ((Addgene plasmid # 62988), which was a gift from Feng Zhang lab, and contained a Puro selective marker and a spCas9 under CMV promoter [10].

The detail steps for targeting sequence cloning is as the follows [11][12]:

To clone the guide sequence into the sgRNA scaffold, synthesize two oligos of the form:

5’ – CACCGNNNNNNNNNNNNNNNNNN – 3’

3’ – CNNNNNNNNNNNNNNNNNNNCAAA – 5’

1) Digest 1ug of pOS05a plasmid with *BbsI* for 30 min at 37°C:

Plasmid	1 ug
BbsI – HF (NEB)	1ul

10X NEB 2.1 Buffer	5 ul
ddH2O	X ul
Total	50 ul

Incubate at 37ºC for 1 hour

2) Gel purify digested plasmid using QIAquick Gel Extraction Kit and elute in EB.

3) Phosphorylate and anneal each pair of oligos:

Oligo1 (100uM)	1 ul
Oligo2 (100uM)	1 ul
10X T4 ligation buffer (NEB)	1 ul
ddH2O	6.5 ul
T4 PNK (NEB)	0.5 ul
Total	10 ul

4) Set up ligation reaction and incubate at room temperature for 10 min:

BbsI digested plasmid from step2 (50ng)	X ul
Phosphorylate and annealed oligo duplex from step 3 (1:200 dilution)	1ul
2X quick ligation Buffer (NEB)	5 ul
ddH2O	X ul
subtotal	10 ul
Quick ligase (NEB)	1 ul
Total	11 l

5) Transformation

T7 endonuclease I assay

T7 endonuclease I can recognize the mismatched DNA and make cleavage at the heteroduplex site. The T7 assay can detect heteroduplex DNA that results from annealing DNA stands that have been modified after a sgRNA/Cas9 mediated cut to DNA strands without modifications. This assay can provide us a first estimate of the efficiency of different sgRNAs. This assay is relatively easy, fast and sensitive to detect the indels generated by Cas9 cutting. First, we extracted the genomic DNA from cells after editing using Pure Link Genomic DNA Mini Kit (Thermo Fisher Scientific, Cat. No K182001). The region targeted by sgRNA was amplified using PCR. The PCR product was ideally to be around 600bp to 1000bp, and the sgRNA targeting site was around the middle of the product. The next step aimed to denature and re-annealing the PCR product in order to generate DNA heteroduplex. We added 2ul NEBuffer 2, 200ng of purified PCR product and dH2O to a total of 19μl in a PCR tube. A hybridization reaction in a PCR cycler was run: 5min, 95C; ramp down to 85C at -2C/second; ramp down to 25C at -0.1C/second; hold at 4C. 1μl of T7 endonuclease I (10U) was used to the reaction and incubated at 37ºC for 15 minutes. The reaction was stopped by adding 2 μl of 0.25M EDTA. The digested PCR product, as well as the undigested PCR product, were loaded immediately side by side on a 2% agarose gel for electrophoresis. If sgRNA worked, a band lower than the PCR product was expected to show up.

Cell surface staining for sorting

Cells were dissociated into single cell suspension using Accutase (for hES cells) or Trypsin-EDTA. The sorting buffer was freshly made using 1XPBS, 2% newborn calf serum and 0.1% sodium azide. Cells were washed once in a 12x75 mm polypropylene sample tube. 0.5 ml of the buffer was added to the pellet with 1 μl of the monoclonal antibody and incubated for 30 min at 4 ºC in the dark after brief vertexing. Cells were washed using buffer twice before incubating with the second antibody (if needed) for 30 min at 4 ºC. Cells were washed twice and resuspended in 3 ml buffer before sorting.

Results

H9 human embryonic stem cell was chosen as the working cell line

The optimal cell lines for generating LOH mapping panel requires several features. First of all this cell should be karyotypically normal diploid, or at least diploid at the chromosome that will be targeted by the Cas9 protein. Aneuploidies will make the mapping and LOH generation more complicated. Additionally, the cell line needs to be editable by CRISPR system. Hard-to-transfect cell lines, such as lymphoblastoid and leukemia cell lines, are hard to edit with CRIPSR system. Moreover, as LOH are generated through repair of DSBs by homologous recombination using the homologous chromosome, it is essential that the chosen cell line actively uses homologous recombination to repair DBS lesions at an appreciable frequency. It will be further advantageous if the cell contains a built-in heterozygous selectable marker that can be used to select LOH event.

I have attempted to establish LOH mapping in the human embryonic stem cell (hESC) H9. This cell line is karyotypically normal diploid[13], which simplifies the LOH generation process. This cell line has been used for several CRISPR-Cas9 system involved studies, indicating it is editable by CRISPR[14].

ABO genotyping by polymerase chain reaction amplification of the specific alleles

To enrich for cells that successfully had LOH events, it is desirable to have a negative selectable marker telomere-proximal to the site of LOH. The selectable marker also needs to be heterozygous. The selectable marker can be a drug resistant gene, a fluorescence gene or any gene that has different phenotype under heterozygous and homozygous genotype.

The H9 cell line chosen in my pilot trial is heterozygous at the ABO locus with genotype AO from literature [15]. The heterozygosity at the ABO locus could be very useful for enriching for cells with LOH events, as hES cells express ABO antigens[16]. ABO locus resides close to the telomere on chromosome 9 at the band 9q34.2. In theory, cells that have become homozygous for the O allele of ABO can be selected by flow cytometry with an antibody against the A antigen.

First, I confirmed the genotype at ABO locus of H9 stem cells using he polymerase chain reaction amplification of the specific allele (PASA). PASA

distinguishes the different ABO genotypes on the bases of molecular size of allele-specific amplification products in ABO allelic DNA. For the genotype AA, AO, and OO, two specific bands (379 and 52 bp), three specific bands (379, 104, and 52 bp), and two specific bands (379 and 104 bp) were amplified respectively [17]. The primer sequences for the PASA were shown in Table 5.1. Genomic DNA was extracted from H9 stem cell and human dermal microvascular endothelial cells (HDMECs). HDMECs were used as quality control cells, given the genotype at ABO locus for the cells were known to be AA, and the result from PASA is consistent with the known genotype for HDMECs. Two bands (379bp and 52bp) showed up in the gel of HDMECs specific amplification. The size of those bands correspondence to the AA genotype at ABO locus (Figure 5.1). We observed three specific bands for AO genotype from the allele specific amplification of the H9 stem cell DNA, confirming the genotype of H9 is AO (Figure 5.1).

ABO locus is not a desirable selectable marker for LOH events

H9 stem cell is a diploid cell line, heterozygous at ABO locus (AO) and CRISPR editable. As a stem cell line, H9 needs to be co-cultured with mice embryonic fibroblast (MEF) cells or grow on Matrigel with mTeSR^{TM}1R medium, which prevent stem cell from differentiation. I raised the H9 cells in both conditions above and tested which growing condition worked better.

In order to test whether blood type A antigen is expressed on the surface of H9 cells, flow cytometry with the PE fluorescence conjugated antibody against blood type A antigen was performed. Pacific blue conjugated antibody against the stem cell surface

marker SSEA3 was also used in the flow cytometry as a marker to distinguish MEF and H9. MEF was also used in the flow cytometry as negative control, and endothelial cells with blood type AA was used as positive control for cell surface blood type A antigen. For the blood type A antigen on the cell surface, both H9 stem cell and the endothelial cells were supposed to have signals, which is the PE signal in the FACS set up. For the stem cell specific cell surface marker, only H9 stem cells were expected to have the SSEA3 signals, which is the pacific blue signal in FACS. MEF was assumed to be negative for both fluorescence antibody staining (Table 5.2).

Consistent with the expectation, MEF does not express blood type A antigen nor the SSEA3 on cell surface (Figure 5.2, 5.3). Over 77% H9 cells grown on feeder free plates contain detectable stem cell specific markers. An unforeseen result is that in our H9 cell line, no matter in which medium they grown, a fraction of cells do not have blood type A fluorescence signal (Figure 5.3). This phenotype is also observed in the positive control endothelial cells (Figure 5.2). The endothelial cells are AA blood type, but around 85% of them are not blood type A positive in the flow. The observed results can be explained by literatures about A antigen expression. This discrepancy between our expectation and observation was caused by the mechanism of blood type determination. ABO antigens are not direct products of the genes. ABO genes code for transferase, which cause transfer of monosaccharide molecule onto a precursor substance (called H antigen) on the cell. In other word, the dynamic process of monosaccharide on and off H antigen causes the large variance of A antigen expression in cells.

Construct a stable cell line with heterozygous selection marker at *APRT* locus

Given the ABO locus is unable to be the selective marker to enrich for LOH events, a new heterozygous marker close to the telomere is of high demand. A candidate gene to be engineered into selective marker is gene *APRT*, which is a dual selectable marker located near the telomere on chromosome 16 (Figure 5.4A). This gene codes for an enzyme called adenine phosphoribosyltransferase (APRT). The enzyme is part of the purine salvage pathway, which recycles the purines to make other molecules. This gene was used as a selectable marker for LOH in mice embryonic stem cells [18] and human T lymphocytes [19]. Cells with homologous deficiency in gene *APRT* (*APRT*-/-) can be selected with 2,6-diaminopurine (2,6-DAP). This gene can also be positively selected by medium with aminopterin, a drug that inhibits the de novo purine and pyrimidine biosynthesis. In H9 hES cells, this gene is homozygous dominant (*APRT*+/+). Once a stable H9 hES cell line that is heterozygous at this gene (*APRT*+/-) is constructed, LOH events can be enriched using 2,6-DAP selection for losing the *APRT*+ allele.

Gene *APRT* on chromosome 16 has five exons. In order to break the gene function entirely, the gRNAs are designed to targeting the first two exons (Figure 5.4B), aiming to introduce indels at the beginning of the protein. Because the disruptions at the end of protein may not mess up the protein functions. The gRNA targeting efficiency was tested in HEK293 cells before applied to H9 ES cells. HEK293 cells have higher transfection efficiencies compared to stem cells, making them better for testing

efficiency. The gRNA targeting efficiency was estimated using T7 endonuclease I mismatch assay. During the assay, region surrounding the targeting site was amplified from genomic DNA. If any mutations or indels occur through non-homologous end joining during sRNA-Cas9 cleavage, heteroduplexes of mutant and wildtype PCR will be generated through a denaturing and annealing process. T7 endonuclease I will recognize and cleavage DNA mismatches in those duplexes. The observation on the agarose gel will be two or one bands below the PCR product band.

Five gRNAs were transfected into HEK293 cells and one gRNA targeting exon 2 showed efficient efficiency (Figure 5.5A, 5.5B). The same gRNA was transfected into H9 ES cells using Nucleofection and DNA of H9 stem cells was extracted after cell becoming confluent. Consistent with observed high efficiency in HEK293, H9 ES cells have clear bands indicating the efficiency cleavage at targeting sites (Figure 5.6). T7 endonuclease I assay reflected the cleavage activity in a mixed pool of cells. The cell pool was expected to be made up cells did not have any indels (*APRT* +/+), cells that are knock out at this gene (*APRT* -/-) and with a few cells that were heterozygous (*APRT* +/-), and only the last type was the cells we wanted.

A stable heterozygous cell line was generated through genotype screening

In order to gain a stable cell line that is heterozygous at *APRT*, H9 ES cells after Cas9 cleavage were plated at low density of 1000 cells per well (6 well-plate) to form single clones for genotype screen. ROCK inhibitors were added to improve the

attachment of cells. The optimal colony size for genotype screen is 0.8 mm in diameter and this usually achieves 8 days after plating [20]. Clones derived from single cells were split into two parts under the microscope with the scraping from a P20 pipette. Half of the clone was used for genotyping; the other half was transferred into a clean well in 96 well plates with medium and ROCK inhibitor. Cells for genotyping were lysed and the targeting regions were amplified. T7 endonuclease I assay was applied to the cell clones. Cells in 96 well plates were passaged when confluent.

Forty single clones of H9 ES cells were picked, genotyped and tested using T7 endonuclease I assay. Seven colonies had strong cleavages and two had weak cleavages (Figure 5.7A). The PCR products of the targeting region were sent for Sanger sequencing to detect whether any of them carry the heterozygous indels. Within the cell clones that survived in 96 well plates after splitting and transferring, two single cell clones had homozygous deletion in gene *APRT* and one cell clone had heterozygous deletion (Figure 5.7B). That heterozygous cell clone (*APRT*+/-) was the desired stable cell line we tried to generate, and the homozygous deletion cell lines (*APRT*-/-) could be used as controls in downstream experiments.

2,6-DAP selection condition optimization

The concentration of 2,6-DAP affects the condition of cells after selection: high 2,6-DAP concentration decrease the fitness of cells, including resistant cells; low concentration unable to select cells effectively, causing false positive cells. The concentration of 2,6-DAP was 50 μg/ml to enrich LOH cell clones of mice fibroblast cells

[21]. H9 ES cells are more sensitive to different selection conditions compared to fibroblast. Therefore we performed a drug killing curve similar analysis to test the effect of 2,6-DAP selection at concentrations of 5, 10,20,30, 40 and 50 μg/ml. The *APRT* heterozygous cells could form colonies and survived when the concentration was 5 and 10 μg/ml. However they failed to grow and form colonies at concentrations above 20 μg/ml. The 2,6-DAP resistant cells (*APRT*-/-) could survive and form colonies at concentrations of 5,10, 20 and 30 μg/ml. When the concentration reached to 40 and 50 μg/ml, the resistant cells did not totally die and disassociate the plate bottom, but they showed unhealthy phenotypes and had smaller colonies (Figure 5.8). Taking all the observations into account, 30 μg/ml is a reasonable concentration that could efficiently enrich the resistant *APRT*-/- cells against the heterozygous cells, and we used this concentration for later experiments.

APRT deficient cells resistant to 2,6-DAP selection

We tested the 2,6-diaminopurine (2,6-DAP) selective efficiency using the *APRT* heterozygous and homozygous deficient cells. 2,6-DAP is an adenine analog that is toxic only to cells with APRT enzyme activity [21]. Cells with homologous deficiency in gene *APRT* (*APRT*-/-) is supposed to be resistant to 2,6-DAP, while heterozygous or wild type cells cannot survive. Medium with 2,6-DAP was added into the wells that contained H9 ES cells of *APRT*+/+, *APRT*+/- and *APRT*-/- genotypes independently. Consistent with previous literature, *APRT*+/+, *APRT*+/- cells died from the toxic 2,6-DAP, while *APRT*-/- cells resistant to the selection (Figure 5.9). We confirmed the effective selection of 2,6-DAP in HEK293 cells as well. We grew the HEK293 after the

transfection of sgRNA and Cas9 plasmids for one week before 2,6-DAP selection. Non-resistant cells died during selection and we collected survived cells one week after. Amplification products of the gRNA targeting region were transformed into bacteria for colony sequencing, which confirmed the fact that all sequenced 2,6-DAP resistant cells contained indels in gene *APRT* (Figure 5.9).

Creating LOH events using CRISPR-Cas9 to target regions near the selective marker

The ultimate goal of this project is to generate a panel of human cells, each of them with different proportions of chromosome 16 to be homozygous towards the telomere. We had the stable cell line that is heterozygous at the selective maker. I designed 22 gRNAs targeting the region between the maker and centromere. CRISPR-Cas9 will induce double strand breaks (DSBs) at the targeting region, and DSBs could potentially cause LOH events as a repair approach. 2,6-DAP could be used to enrich for cells undergone LOH. The efficiency of designed gRNAs was tested in HEK293 using T7 endonuclease I assay first. Around forty percent of the gRNAs showed high efficiency and we picked eight gRNAs that had high efficiency to create the DSBs in H9 ES cells. The gRNAs used are listed in Table 5.3, and the distance between targeting region to the selective marker ranged from around 31kb to 1.7Mb. Guide RNAs closer to the marker were expected to have higher possibility of LOH, thus the gRNAs were enriched to be close to the selective marker. Guide RNAs were transfected to H9 ES cells using Nucleofection in separate wells, and we allowed one week of cell growing to facilitate segregation of recombined chromosomes and recovery after transfection. 2,6-

DAP was added to confluent cells one week after nucleofection. Three days after continuing selection, wild type cells and the transfected heterozygous cells died, and disassociated from the bottom, floating in the culture medium (Figure 5.10). The resistant cells remained attaching to the bottom. The death of all cells in the heterozygous cell population indicated extremely low frequency of LOH occurrence during the induced DSBs and stopped us from downstream exploration of building a mapping panel based on cells undergone LOH.

Discussion

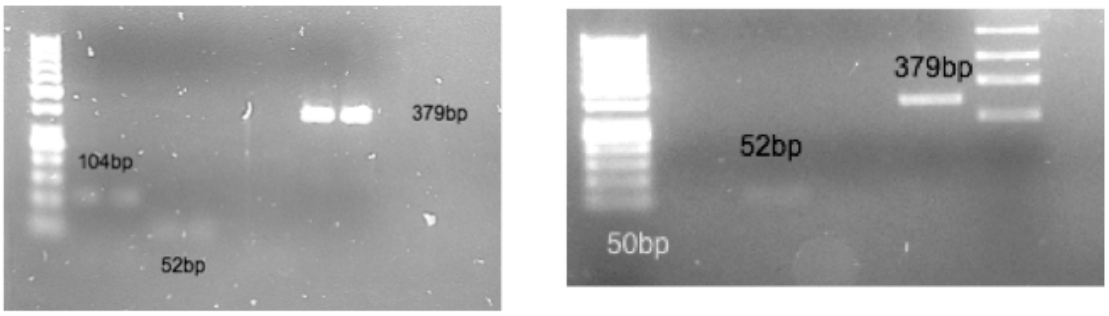
The central goal of genetics is to understand the link between genetic variation and trait variation across the entire range of biological phenomena, from genetic diseases to pathogen virulence. However, current methods for trait mapping often span many genes, and narrowing them to the specific underlying genes and genetic variants can be painstaking. Here we aim to build a fine mapping penal in human cell line using CRISPR-Cas9 to specifically generate DSBs at designed region.

In order to build the mapping panel, first a diploid cell line with relatively high HDR and can be engineered using CRISPR-Cas9 is required. Another requirement for the panel is a heterozygous selective marker close to the telomere, where the selection for homozygous genotype can enrich loss of heterozygosity events. We chose the H9 ES cell line as it is karyotypically normal diploid and has been genome editing through CRISPR before. We generated the heterozygous selectable marker by using CRISPR-Cas9 to knock out one functional allele of gene *APRT*, a gene located on chromosome

16 close to the telomere. We created the stable H9 ES cell line with heterozygous *APRT* selectable marker through genotype screening of single cell clones. By introducing double strand breaks at designed sites along chromosome 16 using CRISPR-cas9, followed by selection for cells undergone loss of heterozygosity (LOH), in theory we will be able to generate a mapping panel of cells with different sizes carrying homozygous recombinant genomes to the telomere. This panel can be used to fine map variants that affect different drug resistance in H9 ES cells, and can also be used to map other traits, such as cell invasion during tumorigeneses. By differentiating the H9 ES cells into other cell types, this panel can be used to fine map more diverse cell specific traits (Figure 5.11).

We introduced double strand breaks at regions between the *APRT* marker and the centromere using CRISPR-Cas9 with eight different gRNAs, whose efficiency were confirmed in HEK293 cells. LOH events were enriched with 2,6-DAP, a reagent that was toxic to cells with APRT enzymatic activity. However heterozygous cells could not survive the selection despite the gRNA transfected. The potential reasons that cause the death of H9 ES cells are: (1) the concentration of 2,6-DAP is too high for resistant cell. (2) the LOH frequency is extremely low. (3) H9 ES cells are vulnerable when not surrounded by healthy cells. The first reason was excluded as the concentration of 2,6-DAP was determined from a killing curve test experiment where the wild type cells will be toxic to death while the resistant cells was fine. The second problem could be solved by treating cells with some reagents to increase LOH rate, but the trade-off is those reagents are usually mutagenic to the whole genome and may affect the downstream

phenotyping. Using another more robust cell line that do not need feeder cells may help to solve the third potential reason for the cell death. Looking ahead, optimizing our current LOH system by increasing HDR rate, boosting LOH rate and improving cell viability could potentially build the LOH mapping panel in mammalian cells, allowing rapidly narrowing down the region of interest to causal genes.



Human H9 stem cell

human dermal microvascular endothelial cells (HDMECs)

Figure 5.1 Electrophoretic patterns of PCR products at the ABO locus for H9 ESCs and HDMECs.

Genomic DNA extracted from H9 ESCs and HDMECs was amplified employing the PASA methods. 50 bp DNA size maker. H9 ESCs has the AO blood type, and is expected to observe the 52bp, 104bp and 379bp bands. HDMECs is AA for the blood type, so only the bands of 52bp and 379bp are expected.

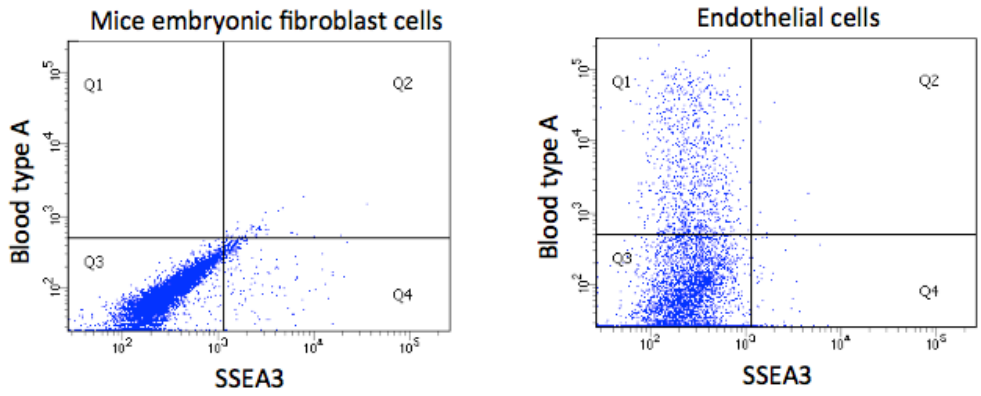


Figure 5.2 FACS result of dual labelling of MEF and endothelial cells for blood type A antigen and stage-specific embryonic antigen 3 (SSEA3).

MEF cells and human dermal microvascular endothelial cells were incubated with PE-conjugated antibody for blood type A antigen and Pacific blue conjugated antibody for cell surface glycosphingolipids SSEA3. The X-axis represents the mean fluorescence intensity (logarithmic scale) for SSEA3 signal, and the Y-axis are the fluorescence intensity for blood type A antigen.

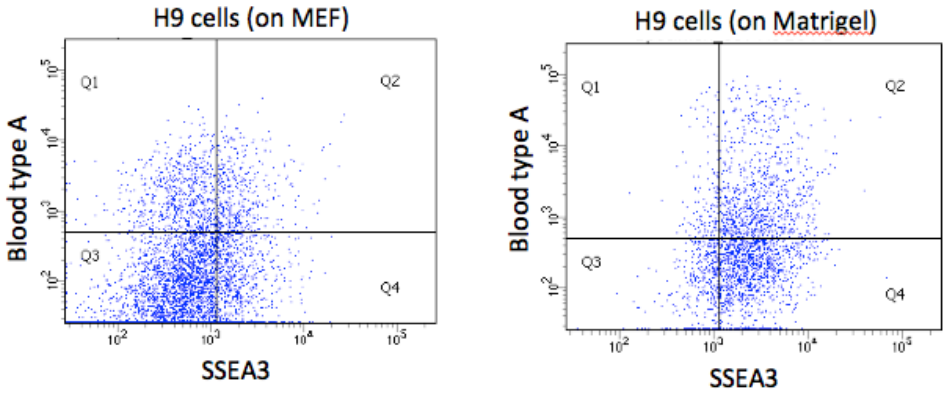


Figure 5.3 FACS result of dual labelling of H9 ES cells on MEF and on Matrigel for blood type A antigen and stage-specific embryonic antigen 3 SSEA3.

H9 ES cells grown on MEF and grown on feeder-free Matrigel were incubated with PE-conjugated antibody for blood type A antigen and Pacific blue conjugated antibody for cell surface glycosphingolipids SSEA3. The X-axis represents the mean fluorescence intensity (logarithmic scale) for SSEA3 signal, and the Y-axis are the fluorescence intensity for blood type A antigen.

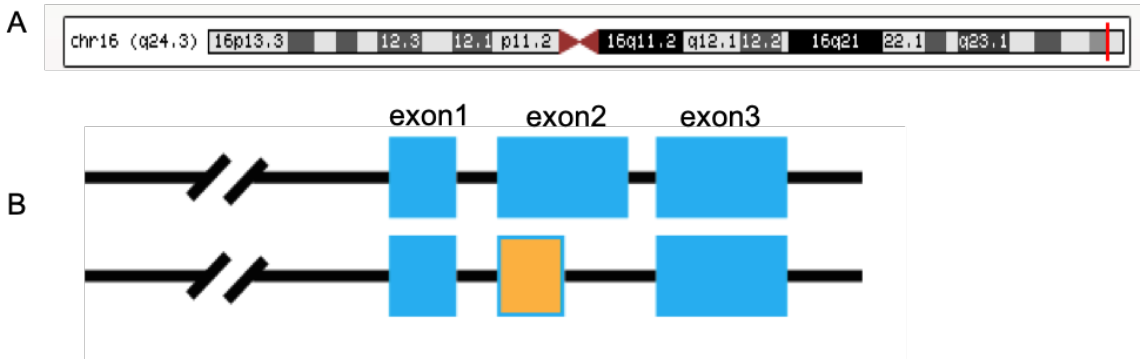


Figure 5.4 The strategy of creating heterozygous selective marker on chromosome 16 by knocking out one allele of gene *APRT*.

(A) Gene *APRT* is located close to the telomere site of chromosome 16. The position is shown as the red line. (B) Gene *APRT* constitutes of five exons. Guide RNAs were designed to target the first two exons to eliminate the function of the targeted allele.

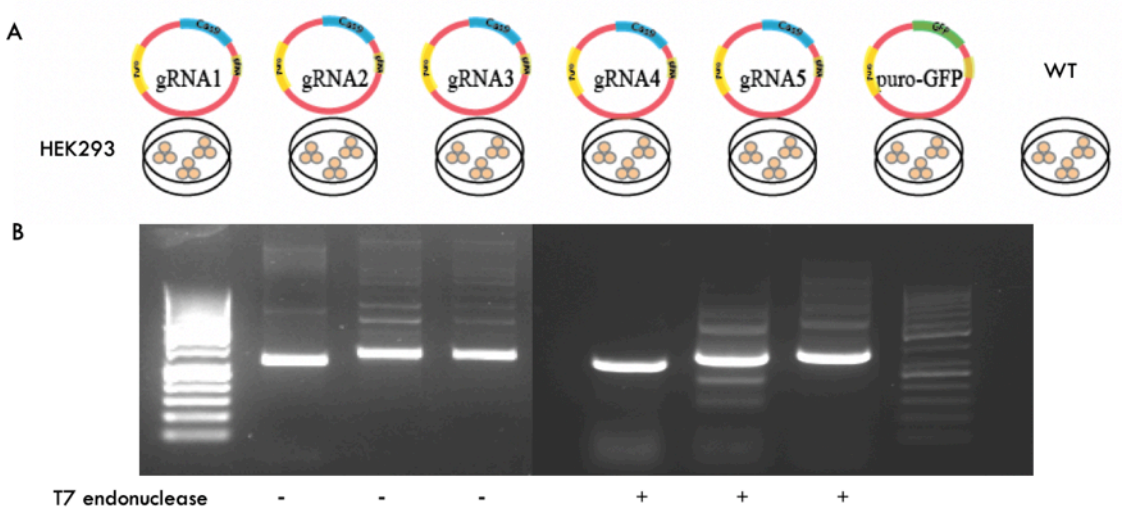


Figure 5.5 The efficiency of gRNAs targeting *APRT* was tested in HEK293.

(A) Five gRNAs were designed to target the first two exons of *APRT* and transfected into HEK293 using lipofectamine 2000. The transfection efficiency was estimated from the positive control where a GFP vector was transfected. (B) T7 endonuclease I assay was applied to the DNA extracted from cells one week after transfection. The gel image showed the assay results for gRNA1-3. For gRNA2, two weak bands under the PCR product band were observed after adding the T7 nuclease, indicating sufficient gRNA efficiency that created indels around the targeting region. The gel was 2% agarose gel. The DNA ladder used is 50bp.

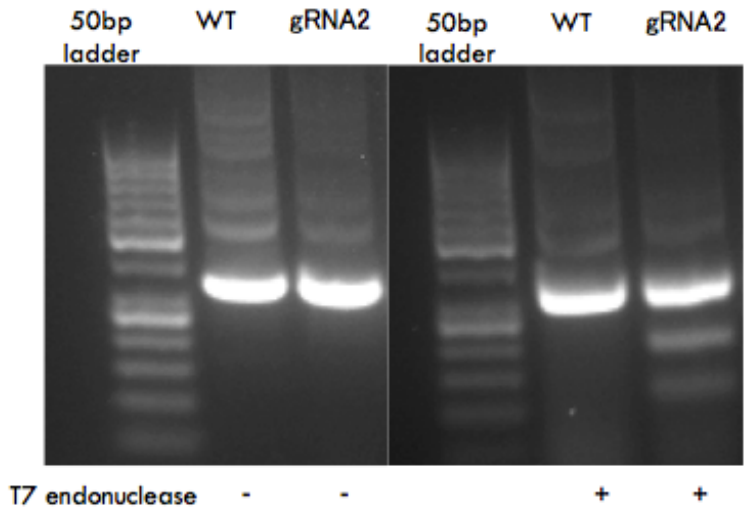


Figure 5.6 The gRNA2 efficiently targeted genome and generated indels in H9 ES cells.

PCR products that amplified the region surrounding gRNA targeting region were run on a 2% agarose gel. The product is 348bp, and the DNA of both wild type H9 and H9 targeted by gRNA2 had the bands corresponded to the right size. After adding T7 endonuclease, the PCR product from H9 ES cells targeted by gRNA2 was cleavage into two smaller bands. The ladder used is 50bp ladder.

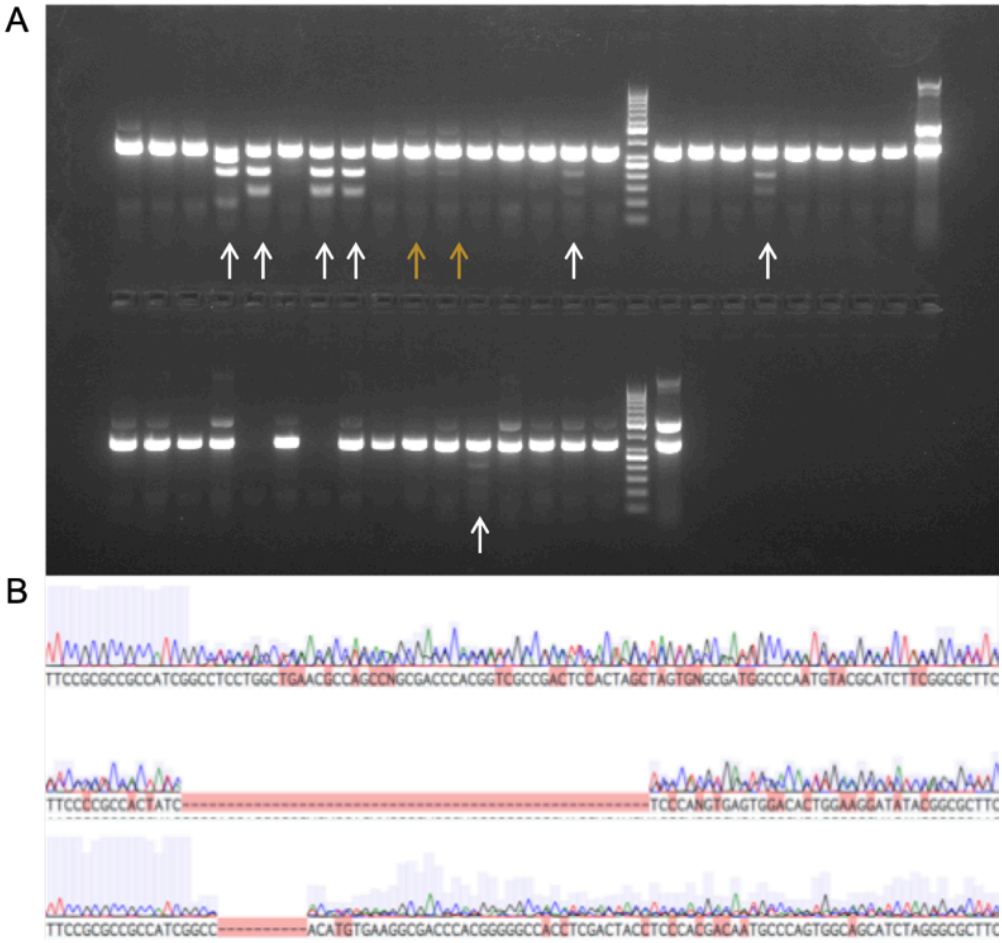


Figure 5.7 T7 endonuclease I assay and Sanger sequencing were applied to single cell clones to identify clones with heterozygous alleles at gene *APRT*.

(A) The amplified product surrounding targeting region from forty single cell clones were incubated with T7 endonuclease I. Cells with clear cleavage were indicated with white arrow, and cells with weak cleavage were pointed with orange arrows. The DNA ladder used is 50 bp. (B) Sanger sequencing traces of the cell clones with heterozygous alleles (top), large homozygous deletion (middle) and small homozygous deletion (bottom). Red shadow shows the base pairs mismatched with the reference genome.

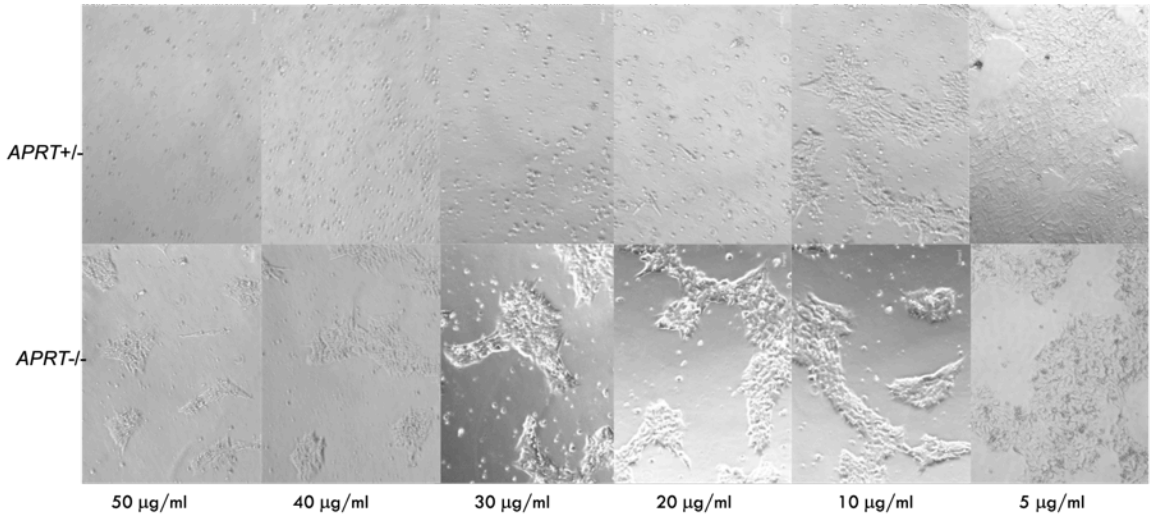


Figure 5.8 H9 ES cell morphology on Matrigel under different concentrations of 2,6-DAP selection.

The upper panels are the *APRT* heterozygous H9 ES cells at six different concentrations of 2,6-DAP. The lower panels are the 2,6-DAP resistant H9 ES cells (*APRT*-/-). Healthy cells form clones on Matrigel, while dead cells disassociate from the bottom and float as single round cells.

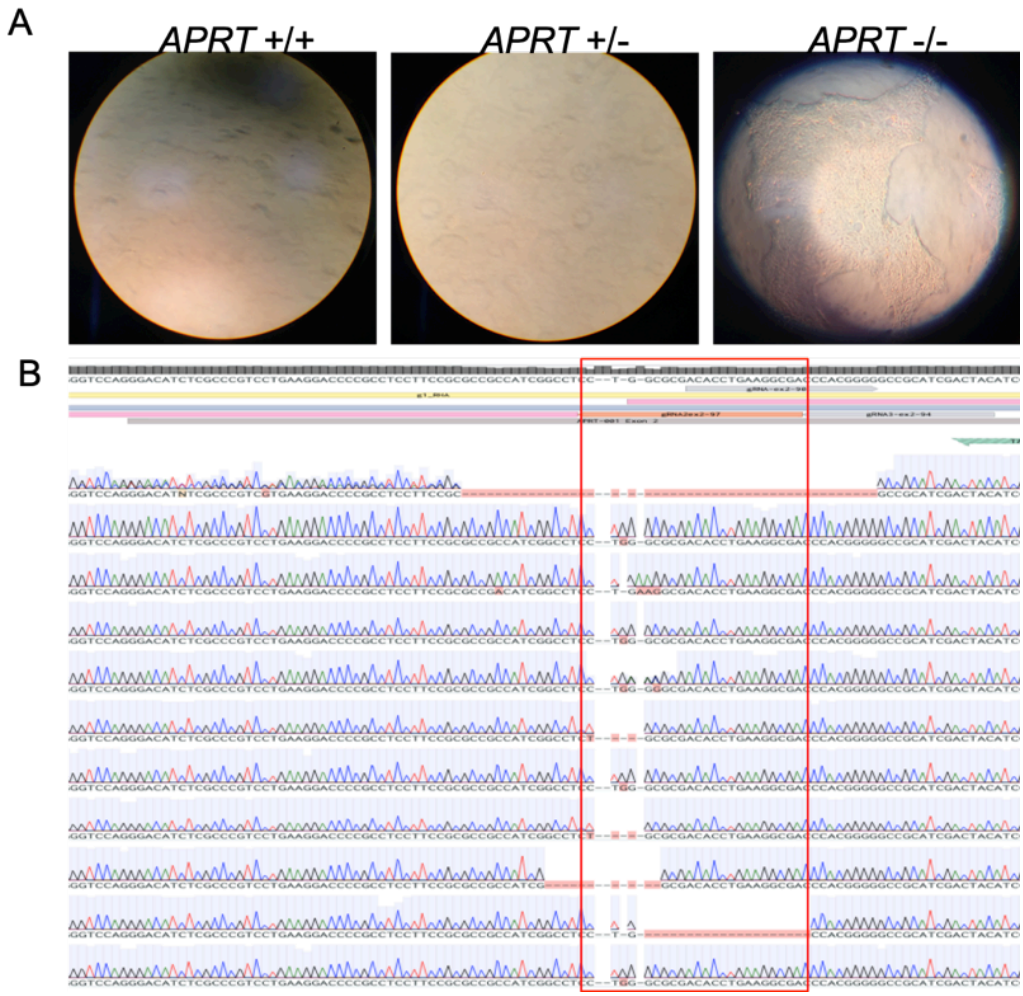


Figure 5.9 2,6-DAP is sufficient to enrich for *APRT*-/- cells.

(A) Wild type H9 ES cells and the *APRT* heterozygous cells died during 2,6-DAP selection. The *APRT*-/- cells survived under the selection. (B) CRISPR-Cas9 was applied to introduce indels at gene *APRT* in HEK293 cells, followed by 2,6-DAP selection. Region surround the targeting region was amplified from the DNA of resistant cells. PCR products were transformed into bacteria and colonies were sanger sequence. Sequence data showed indels around the targeting region. Red square shows the gRNA targeting region.

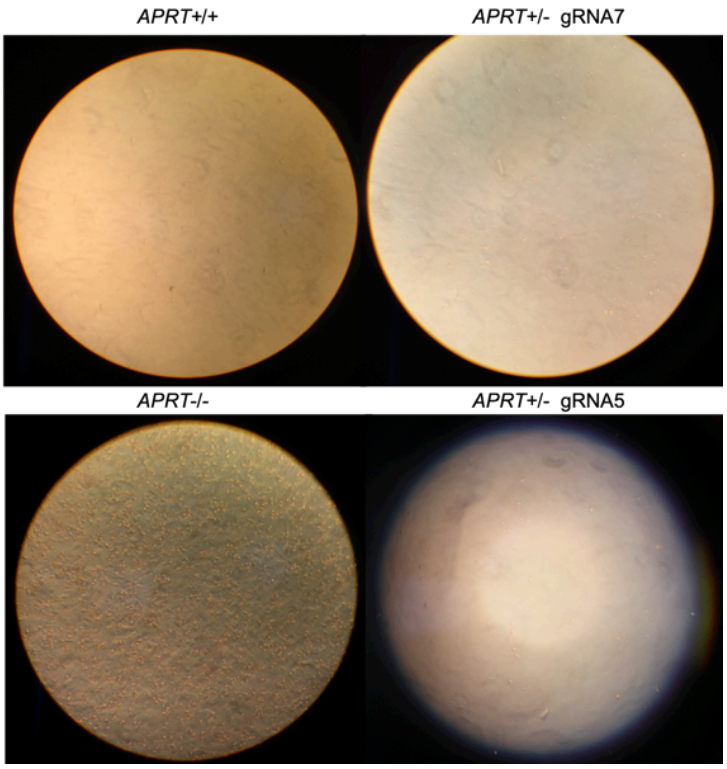


Figure 5.10 H9 *APRT* heterozygous cells failed to survive 2,6-DAP after second DSBs introduced by Cas9 and gRNAs.

Images were taken three days after 2,6-DAP selection. *APRT*-/- H9 ES cells survived, while the *APRT*+/- cells died no matter targeted by a second gRNAs between the selective marker and the centromere or not.

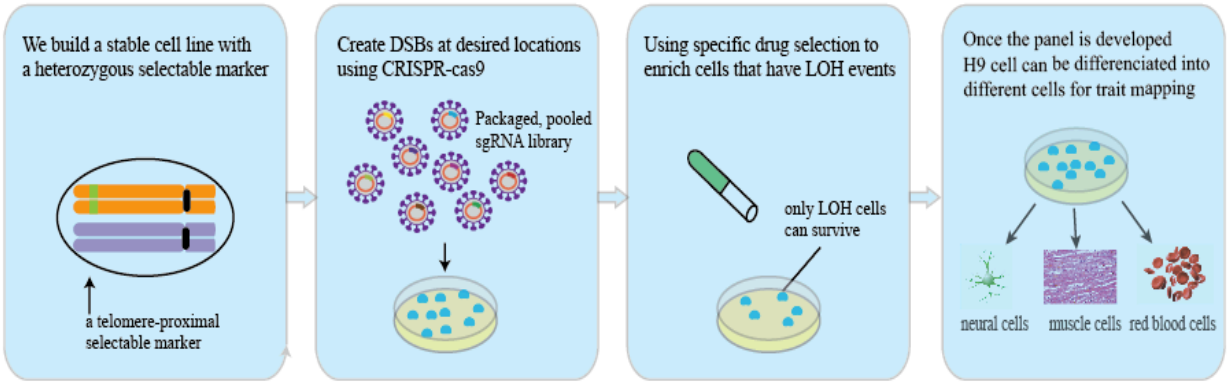


Figure 5.11 Ideal workflow of building the LOH mapping panel in H9 embryonic stem cell.

The idea of building the LOH mapping panel starts with a heterozygous selective marker close to the telomere of one chromosome, followed by a pool of gRNAs to create the DSBs between marker and centromere. Then use a certain selection method to enrich for cells have successfully undergone LOH. Once the panel was built up, we can differentiate the stem cell into different cell types to explore cell-type specific variants effects.

Table 5.1 The sequences of primers used in the PASA.

Name	Allele-specific	Product	Sequences (5'-3')
Primer 1	O-specific	104 bp	GAATTCATGTGGGTGGCACCCTGCCA
Primer 2	O-specific	104 bp	AGACAATGGGAGCCAGCCAAGGGGGA
Primer 3	A,B-specific	379 bp	AGACAATGGGAGCCAGCCAAGGGGTC
Primer 4	A,B-specific	379 bp	GAATTCAGGAAGGATGTCCTCGTGGTG
Primer 5	A,O-specific	52 bp	CAGCTGTCAGTGCTGGAGGTGC
Primer 6	A,O-specific	52 bp	CCGTTGGCCTGGTCGACCATCATGGCC TG

Table 5.2 The expectation of the observed signals of cells tested in FACS.

Cell type		Expected PE signal	Expected pacific blue signal
MEF	Negative control	-	-
Endothelial cell	Blood type A positive control	+	-
H9+MEF		+	+
H9		+	+

Table 5.3 gRNAs used to target the region between centromere and the selectable marker.

Name	Genome position on Chr16	Distance to marker
APRT (marker)	88875877	0
gRNA1	88844927	30950
gRNA2	88844693	31184
gRNA3	88164945	710932
gRNA4	87788864	1087013
gRNA5	87098993	1776884
gRNA6	86339508	2536369
gRNA7	79687003	9188874
gRNA8	66906879	21968998

References

1. Mojica FJM, Montoliu L. On the Origin of CRISPR-Cas Technology: From Prokaryotes to Mammals. Trends in Microbiology. 2016. pp. 811–820. doi:10.1016/j.tim.2016.06.005
2. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science (80-). 2014;346: 1258096–1258096. doi:10.1126/science.1258096
3. Sadhu MJ, Bloom JS, Day L, Kruglyak L. CRISPR-directed mitotic recombination enables genetic mapping without crosses. Science (80-). 2016;352: 1113–1116. doi:10.1126/science.aaf5124
4. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol. 2016;34: 339–344. doi:10.1038/nbt.3481
5. Yang D, Scavuzzo MA, Chmielowiec J, Sharp R, Bajic A, Borowiak M. Enrichment of G2/M cell cycle phase in human pluripotent stem cells enhances HDR-mediated gene repair with customizable endonucleases. Sci Rep. 2016;6. doi:10.1038/srep21264
6. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol. 2015;33: 543–548. doi:10.1038/nbt.3198
7. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by

inhibition of nonhomologous end joining. Nat Biotechnol. 2015;33: 538–542. doi:10.1038/nbt.3190

8. Song J, Yang D, Xu J, Zhu T, Chen YE, Zhang J. RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun. 2016;7. doi:10.1038/ncomms10548
9. Lin S, Staahl BT, Alla RK, Doudna JA. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife. 2014;3: e04766. doi:10.7554/eLife.04766
10. Ann Ran F, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRIPR-Cas9 system. Nature. 2013;
11. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014; doi:10.1038/nmeth.3047
12. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science (80-). 2014; doi:10.1126/science.1247005
13. Thomson JA. Embryonic Stem Cell Lines Derived from Human Blastocysts. Science (80-). 1998;282: 1145–1147. doi:10.1126/science.282.5391.1145
14. Lin S, Staahl B, Alla RK, Doudna J a. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. Elife. 2014;3: 1–13. doi:10.7554/eLife.04766
15. Chen Y-T, Dejosez M, Zwaka TP, Behringer RR. H1 and H9 human embryonic stem cell lines are heterozygous for the ABO locus. Stem Cells Dev. Mary Ann Liebert, Inc.; 2008;17: 853–5. doi:10.1089/scd.2007.0226

16. Molne J, Bjorquist P, Andersson K, Diswall M, Jeppsson A, Strokan V, et al. Blood group ABO antigen expression in human embryonic stem cells and in differentiated hepatocyte- and cardiomyocyte-like cells . Transplantation. 2008;86: 1407–1413. doi:10.1097/TP.0b013e31818a6805

17. Aki K, Izumi A, Hosoi E. The evaluation of histo-blood group ABO typing by flow cytometric and PCR-amplification of specific alleles analyses and their application in clinical laboratories. J Med Investig. 2012;59: 143–151. doi:10.2152/jmi.59.143

18. Cervantes RB, Stringer JR, Shao C, Tischfield JA, Stambrook PJ. Embryonic stem cells and somatic cells differ in mutation frequency and type. Proc Natl Acad Sci U S A. National Academy of Sciences; 2002;99: 3586–90. doi:10.1073/pnas.062527199

19. Gupta PK, Sahota A, Boyadjiev SA, Bye S, Shao C, Patrick O'Neill J, et al. High frequency in vivo loss of heterozygosity is primarily a consequence of mitotic recombination. Cancer Res. 1997;57: 1188–1193.

20. Park CY, Sung JJ, Choi SH, Lee DR, Park IH, Kim DW. Modeling and correction of structural variations in patient-derived iPSCs using CRISPR/Cas9. Nat Protoc. 2016;11: 2154–2169. doi:10.1038/nprot.2016.129

21. Shao C, Deng L, Henegariu O, Liang L, Raikwar N, Sahota A, et al. Mitotic recombination produces the majority of recessive fibroblast variants in heterozygous mice. Proc Natl Acad Sci. 2002; doi:10.1073/pnas.96.16.9230

Chapter 6 Conclusions

Understanding the functional consequences of genetic variants has been a major challenge in genetics for decades. Traditionally, linkage mapping and association studies in natural population are powerful approaches to find out causal genes or regions underlying the trait of interest. In recent year, the programmable genetic engineering tool CRISPR-Cas9 has been widely used, and the utility of this tool in studying the effect of genetic changes has been widely recognized by researchers [1]. A number of genome-wide high-throughput genetic perturbations have been introduced into the genomes of multiple organisms and model systems using CRISPR-Cas9.

In this dissertation, we explored the effect of variants using both linkage analysis and the CRISPR screening. In **Chapter 2**, we studied the genetic basis of mutation rate variation by performing linkage analysis in 1040 segregants from a cross from two diverge yeast strains. We identified four QTLs underlying the mutation rate variation between the cross, and we further fine-mapped two QTLs to the causal genes *RAD5* and *MKT1*. *RAD5* encodes a DNA repair protein involved in the error-free DNA damage tolerance (DDT) pathway, and *MKT1* encodes an RNA-binding protein that affects multiple traits. These genes also underlie sensitivity to the DNA damaging agents 4NQO and MMS, indicating a connection between spontaneous mutation rate and mutagen sensitivity. Mutation rates are difficult to measure, and sensitivity to mutagens may serve as a useful proxy. The two causal variants we identified for mutation rate variation are specific to the BY and RM cross, it will be interesting to explore if other

natural variants present in other yeast isolates affect mutation rate. Furthermore, *RAD5* was reported to fall into one QTL found to influence adaptability [2], indicating the potential effects of spontaneous mutation rate on the adaptability of organisms.

In addition to linkage analysis, in **Chapter 3**, we constructed ten different CRISPR-Cas9 base editor systems for yeast, measured the editing efficiency and tested the editing window for each of them. We observed the three base editors with NGG PAM site (BE3, BE4 and BE4-Gam) and one NGA PAM site base editor (VQR) had the highest editing efficiencies of around 50% in tested base editors. We also observed wider editing window for VQR (edit the C at position 4-11 on gRNA) compared to the regular NGG PAM site base editors (edit the C at position 4-8 on gRNA). Our systematic efficiency analysis of different base editors provided valuable guidance for researchers who plan to apply base editors in yeast and perhaps in other systems. Furthermore, we discovered that a non-NGG PAM site base editor has high editing efficiency in yeast. The NGA PAM site base editor VQR could expand the editable genomic regions for base editor by 1.7-fold.

In **Chapter 4**, we developed a library-based CRISPR-Cas9 editing system with trackable genomic barcodes for high-throughput precise genetic engineering in yeast. The integrated barcode system enables trackable genomic integrated cellular barcoding and robust downstream phenotyping. We tested our system using amplicon sequencing and found more than half of the cases tested had desired precise single variant editing. When applied to the large libraries, both the CRISPR-Cas9 base editor and the donor

DNA repair template system will enable the editing of tens of thousands of genetic variants and will advance our understanding of the effect of genetic variants for a wide range of traits, including growth rate, gene expression levels, industry applications and genetic diseases.

In **Chapter 5**, we worked towards building a CRISPR-directed mitotic recombination mapping panel in human cells, aiming to narrow down the genomic mapped out regions to causal genes. We chose H9 embryonic stem cells based on its diploid and relatively high HDR characteristics. We generated a heterozygous selective marker at gene *APRT* on chromosome 16 close to the telomere by specifically knocking out one functional allele of that gene using CRISPR-Cas9. We tried to enrich for cells undergone the loss of heterozygosity (LOH) process, where the chromosome became homozygous from the double strand break all the way to the telomere. The cell with heterozygous selective marker will be resistant to the 2,6-DAP selection if the LOH occurred. Unfortunately, our cells died when the enrichment selection was applied. It was possible that the LOH happens at extremely low frequency and the resistant cells failed to survive when lacking healthy cells surround them. Potential approaches to enable LOH mapping in human cells could involve increasing HDR rate, boosting LOH rate and improving cell viability. DNA ligase IV inhibitor SCR7 and HDR enhancer RS-1 have been proven to increase HDR rate in human cell lines and could be used here [3][4][5]. LOH frequency could be increased by knocking down the gene *BLM* [6]. If successful, the LOH mapping panel could generate high-density recombination events to a desired region in the genome, enabling fine mapping of disease associated loci.

Altogether, this dissertation aimed to explore the link between genetic variants and phenotypic variation. We approached our goal by using traditional linkage analysis approach and developing library-based high-throughput CRISPR-Cas9 genetic editing tools to perturb the genome. We have mapped out the underlying causal genes for spontaneous mutation rate variation between two diverge yeast strains, and we have constructed three different CRISPR-Cas9 systems for high-throughput genome engineering.

References:

1. Adli M. The CRISPR tool kit for genome editing and beyond. Nature Communications. 2018. doi:10.1038/s41467-018-04252-2
2. Jerison ER, Kryazhimskiy S, Mitchell J, Bloom JS et al. Genetic variation in adaptability and pleiotropy in budding yeast. Elife. eLife Sciences Publications Limited; 2017;6: 1–38. doi:10.1101/121749
3. Chu VT, Weber T, Wefers B, Wurst W, Sander S, Rajewsky K, et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat Biotechnol. 2015;33: 543–548. doi:10.1038/nbt.3198
4. Maruyama T, Dougan SK, Truttmann MC, Bilate AM, Ingram JR, Ploegh HL. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat Biotechnol. 2015;33: 538–542. doi:10.1038/nbt.3190
5. Song J, Yang D, Xu J, Zhu T, Chen YE, Zhang J. RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat Commun. 2016;7. doi:10.1038/ncomms10548
6. LaRocque JR, Stark JM, Oh J, Bojilova E, Yusa K, Horie K, et al. Interhomolog recombination and loss of heterozygosity in wild-type and Bloom syndrome helicase (BLM)-deficient mammalian cells. Proc Natl Acad Sci. 2011; doi:10.1073/pnas.1104421108