UNIVERSITY OF CALIFORNIA, MERCED

Quantitative and Systems Biology

# A computational approach for microbial genome editing

By

Thaddeus D. Seher

A dissertation

submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Quantitative and Systems Biology

Awarded in the year

2021

Committee:

Professor Aaron D. Hernday, Chair

Professor Chris T. Amemiya, Member

Professor Clarissa J. Nobile, PhD Advisor

Professor Suzanne S. Sindi, PhD Advisor

Professor Zhong Wang, Member

A computational approach for microbial genome editing

Portions of this dissertation also appear in
“AddTag, a two-step approach with supporting software package that facilitates
CRISPR/Cas-mediated precision genome editing”

## Signatures

The dissertation of Thaddeus D. Seher, titled "A computational approach for microbial genome editing," is approved and is acceptable on quality and form for publication.

*"I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy."*

**Clarissa J. Nobile**
Advisor

Date

**Suzanne S. Sindi**
Advisor

Date

**Chris T. Amemiya**
Member

Date

**Zhong Wang**
Member

Date

**Aaron D. Hernday**
Chair

Date

UNIVERSITY OF CALIFORNIA, MERCED

2021

# Dedication

To God

and

to my parents.

## Table of Contents

## List of abbreviations

| | | |
|---|---|---|
| +gDNA | – Reference, genomic DNA | 28 |
| $\Delta G$ | – Change in Gibbs' free energy | 105 |
| ΔgDNA | – Intermediary, genomic DNA | 28 |
| ABgDNA | – Add-back, genomic DNA | 28 |
| AEJ | – Alternative end joining | 48 |
| AIR | – Aminoimidazole ribotide | 42 |
| bp | – Base pairs | 35 |
| bs | – Binding site | 23 |
| CAIR | – 1-(5-Phospho-D-ribosyl)-5-amino-4-imidazolecarboxylate | 42 |
| Cas | – CRISPR-associated family of genes | 3 |
| CDF | – Cumulative distribution function | 79 |
| CDS | – Coding sequence | 30 |
| CRE | – Cis-acting regulatory element | 30 |
| CRISPR | – Clustered, regularly-interspaced, short, palindromic repeats | 3 |
| dDNA | – Donor DNA | 18 |
| DNA | – Deoxyribonucleic acid | 2 |
| DS | – Downstream DNA | 19 |
| DSB | – DNA Double-stranded break | 4 |
| eFeature | – Expanded Feature | 65 |
| gDNA | – Genomic DNA | 18 |
| gRNA | – Guide RNA | 3 |
| HA | – Homology arm | 4 |
| HDR | – Homology-directed repair | 18 |
| LOH | – Loss of heterozygosity | 45 |
| NCBI | – National Center for Biotechnology Information | 6 |
| NHEJ | – Non-homologous end joining | 18 |
| nt | – Nucleotides | 3 |
| PAM | – Protospacer-adjacent motif | 18 |
| PCR | – Polymerase chain reaction | 3 |
| RNA | – Ribonucleic acid | 3 |
| RGN | – RNA-guided (endo)nuclease | 3 |
| SynFrag | – Synthetic DNA fragment | 125 |
| $T_m$ | – Melting temperature | 106 |
| US | – Upstream DNA | 19 |
| vPCR | – Verification PCR | 93 |

## List of figures

## List of tables

## List of equations

## Acknowledgements

This is the section of my writing where I get to thank all the remarkable people who have touched my life in this pursuit of scientific excellence—my mentors, friends, and companions. As you suspect, this dissertation didn't will itself into existence. It is the product of years of dedication other people have invested in me. For over a decade, obtaining a PhD has been my dream. In the rest of this section, I'll attempt the impossible: to express my gratitude for the people who literally made my dream a reality.

First, I would like to thank Dr. Mose Durst. You were the first person in my life who introduced himself as a PhD. For years, I didn't know what "PhD" meant, but I did see what kind of person you were. You carried yourself with constant dignity (and you still do). You were patient in the extreme. You listened. You thought. That first impression of what it means to be a PhD has stuck with me. You conferred exceptional deference to everyone, even young kids like me. Being a PhD isn't just what you accomplish—it's how you treat others in the process. From you, I learned respect before scholarship.

Next, I would like to thank Dr. Charles H. Langley for being willing to take a chance on an overly-eager undergraduate student who loved the concept of genomics without necessarily understanding all the details. Dr. Charis M. Cardeno was the perfect lab manager—always professional, even when her biting sarcasm tore my soul for not turning the lights off when leaving the lab that one time. I credit Marc Crepau for encouraging me to engage with professors who I thought were far above me. You were also willing to let me do library preparations alongside you, even though it wouldn't increase your output. Many thanks.

Thank you Dr. Artyom Kopp for inviting me into your science family. The culture of the lab was amazing to behold. Everyone was so, so different. We were spread across different floors and buildings, yet your constant thoughtfulness always encouraged us to be engaged in the matters important to each other. I hope I can unify people as you do—even a fraction would be huge.

I want to convey immense gratitude to Dr. William B. Ludington. You kindled my dedication to the broader, ecological and societal implications of genomics. You gave me the greatest support by encouraging me to apply for graduate studies. You gave me a space where I could grow and learn. Just as you gave me ground to stand on, I too want to uplift other scientists.

I am indebted to Dr. Gregory M. Barton, who seemed content let me struggle through many things, but was also always waiting patiently for me to succeed. Your strong desire to know "why" still inspires me. Everyone in the lab worked together, and I could be a part of an ideal example of how collaboration should work in the sciences. Additionally, Dr. Meghan A. Koch embodied the quintessential qualities of team leadership, which I continue to endeavor emulating with every new team I join.

I thank my doctoral advisors, Dr. Clarissa J. Nobile and Dr. Suzanne S. Sindi, who were with me through everything. Every time there was an impending deadline, you went out of your way to set aside time just for me so we could talk about any issues that arose. You enthusiastically supported every one of my ambitions, and you placed great importance on each of my academic pursuits. Clarissa, I will always admire your ability to be compassionate and encouraging. Suzanne, you repeatedly asked me to question my

assumptions, which really helped me identify and correct weaknesses in our experiments. Also, you were not afraid to set aside stale conversation topics for gelastic ones. Clarissa and Suzanne, you showed me that consideration of others is essential to become a PhD. You both invested in me—nay, sacrificed for my sake—and I am endlessly grateful. Together, the Clarissa and Suzanne duo are the Best Advisors of All Time™.

My advisory committee was stalwart throughout my PhD studies. You have my profuse thanks for your kindness, patience, and thoughtfulness. Dr. Kirk Jensen gave me quiet but firm support whenever I needed it. You are an outstanding mentor in both religious and scientific pursuits. I was overwhelmed with kindness when Dr. Chris T. Amemiya came onto my committee during my final semester of studies. You are constantly managing conflicting responsibilities, and you spent your precious time with me. I am grateful. Dr. Zhong Wang always had brilliant questions after every committee meeting, and you bestowed sharp comments (the good kind) on the AddTag manuscript. Dr. Aaron D. Hernday was insightful when we talked about molecular biology. Life threw us in the same situation—each with a newborn of our own—and it was nice to have a different kind of solidarity with you. You originated of many of the ideas I developed in this dissertation. You worked tirelessly to make sure our research was awesome. I hope to be as creative and far-thinking as you in my future endeavors.

I would like to thank the other co-authors on the AddTag project. Firstly, Namkha Nguyen for his experimental abilities. You performed a significant number of PCR experiments and sequenced the edited loci, and for that I am grateful. Next, I thank Diana Ramos for dedicating a large portion of her undergraduate research career to our work together. You did incredible tapping the minds of the many different personalities in the labs in order to teach yourself molecular cloning. I also recognize Dr. Priyanka Bapat for her vital contributions in performing the biofilm experiments for the project.

Throughout the course of my PhD studies, I have been blessed to be around friends that didn't mind that all I wanted to do was talk about science. There are too many people to name, but there are a few that kept me moving forward toward my goals that I would like to acknowledge: Dr. David H. Ardell, Dr. A. Carolin Frank, Dr. Miriam Barlow, Dr. Michael D. Cleary, Dr. Emilia Huerta-Sanchez, Dr. Glenda Polack, Dr. Megha Gulati, Dr. Melanie Ikeh, Dr. Travis J. Lawrence, Noelle Anderson, Mohammad Qasim, Ashley Valle Arevalo, Akshay Paropkari, Morgan Quail, Craig Ennis, Austin Perry, Diana Rodriguez, and Clement Laksana.

I'd like to give big thanks to the School of Natural Sciences and the Graduate Division at the University of California, Merced for taking me on as a PhD student. Being a Teaching Assistant helped me realize that instruction is both an art and a science. I hope I didn't cause too much trouble? I don't actually want to know the answer to that. Special thanks to Dr. Christopher T. Kello, Dr. Marjorie S. Zatz, Cassie Gunter, Janice Zarate, Rita Guel, and Joy Sanchez-Bell.

Finally, I would like to recognize Aquina, my loving wife, who accompanied me on this journey through graduate studies. She has been wonderfully caring and supportive of my research and teaching.

## Curriculum Vita

# *Thaddeus D. Seher*

## Education

**PhD in Quantitative and Systems Biology** Aug 2015 – Aug 2021
UC MERCED

**BS in Genetics** Sep 2006 – Jun 2010
Minor in Quantitative Biology and Bioinformatics
UC DAVIS

## Publications

**Thaddeus D. Seher**, Namkha Nguyen, Diana Ramos, Priyanka Bapat, Clarissa J. Nobile, Suzanne S. Sindi, Aaron D. Hernday. AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing. *G3 Genes | Genomes | Genetics* (June 2021). doi: 10.1093/g3journal/jkab216, available from: <https://github.com/tdseher/addtag-project>.

William Ludington, **Thaddeus D. Seher**, Olin Applegate, Xunde Li, Joseph Kliegman, Charles Langelier, Edward Atwill, Thomas Harter, Joseph DeRisi. Assessing biosynthetic potential of agricultural groundwater through metagenomic sequencing: A diverse anammox community dominates nitrate-rich groundwater. *PLoS ONE*: 12(4), e0174930 (December 2016), doi: 10.1371/journal.pone.0174930, PMID: 28384184.

Amir Yassin, Emily Delaney, Adam Reddiex, **Thaddeus D. Seher**, Héloïse Bastide, Nicholas Appleton, Justin Lack, Jean David, Stephen Chenoweth, John Pool, Artyom Kopp. The *pdm3* locus is a hotspot for recurrent evolution of female-limited color dimorphism in *Drosophila*. *Current Biology*: pii S0960-9822(16)30767-9 (August 2016),
doi: 10.1016/j.cub.2016.07.016, PMID: 27546577.

Meghan Koch, Gabrielle Reiner, Kyler Lugo, Lieselotte Kreuk, Alison Stanbery, Eduard Ansaldo, **Thaddeus D. Seher**, William Ludington, Gregory Barton. Maternal IgG and IgA antibodies dampen mucosal T helper cell responses in early life. *Cell*: 165(4), 827-841 (May 2016), doi: 10.1016/j.cell.2016.04.055, PMID: 27153495.

Sarah Signor, **Thaddeus D. Seher**, Artyom Kopp. Genomic resources for multiple species in the *Drosophila ananassae* species group. *Fly*: 7(1), 47-56 (January 2013), doi: 10.4161/fly.22353, PMID: 23639891.

**Thaddeus D. Seher**, Chen Siang Ng, Sarah Signor, Ondrej Podlaha, Olga Barmina, Artyom Kopp. Genetic basis of a violation of Dollo's law: re-evolution of rotating sex combs in *Drosophila bipectinata*. **Front cover** *Genetics*: 192(4), 1465-1475 (December 2012), doi: 10.1534/genetics.112.145524, PMID: 23086218.

## Abstracts and posters

**Thaddeus D. Seher**, Namkha Nguyen, Diana Ramos, Priyanka Bapat, Clarissa Nobile, Suzanne Sindi, Aaron Hernday. A computational approach for genome editing using CRISPR/Cas. Society for Mathematical Biology: Abstract + Poster (June 2021).

Meghan Koch, Gabrielle Reiner, Kyler Lugo, **Thaddeus D. Seher**, William Ludington, Gregory Barton. Anti-commensal IgG3 antibodies reinforce intestinal immune homeostasis. Inflammatory Bowel Diseases—Mucosal Inflammation Program: Abstract + Poster (December 2014), doi: 10.1097/01.MIB.0000456934.57591.13.

Meghan Koch, Gabrielle Reiner, Kyler Lugo, **Thaddeus D. Seher**, William Ludington, Gregory Barton. TLR-induced anti-commensal antibodies regulate intestinal homeostasis. Kenneth Rainin Foundation's 2014 Innovations Symposium—Taming the Microbiome: Abstract + Poster (July 2014).

**Thaddeus D. Seher**, Ted Kim, Carolyn Elya, Javier Navarro, Eoin Brodie, Michael Eisen, William Ludington. Ecological epistasis in the fly gut on spatial, temporal, and evolutionary dimensions. 2013 U.C. San Francisco/U.C. Berkeley Immunology Retreat: Poster (September 2013).

**Thaddeus D. Seher**, Ondrej Podlaha, Artyom Kopp. Evolution of abdominal pigmentation in the montium subgroup of *Drosophila*. 2011 Genetics Society of America *Drosophila* Research Conference: Abstract + Poster (March 2011).

## Research positions

**Graduate Student** Jan 2016 – Aug 2021
Nobile Laboratory (MCB), UC Merced

**Graduate Student** Jan 2016 – Aug 2021
Sindi Laboratory (Applied Math), UC Merced

**Computer Resource Specialist** May 2013 – Jul 2015
Ludington Laboratory (MCB), UC Berkeley

**Computer Resource specialist** May 2013 – Jul 2015
Barton Laboratory (MCB), UC Berkeley

**Junior Specialist** Oct 2020 – Oct 2012
Kopp Laboratory (Evolution), UC Davis

**Animal Technician and Pipeline Engineer** Dec 2008 – Oct 2010
Langley Laboratory (Evolution), UC Davis

## Teaching positions

**Teaching Assistant** Spring 2021
BIO-180: Mathematical Modeling for Biology, UC Merced

**Teaching Assistant** Fall 2020
MATH-012: Calculus II, UC Merced

**Teaching Assistant** Summer 2020
MATH-012: Calculus II, UC Merced

**Teaching Assistant** Spring 2020
MATH-011: Calculus I, UC Merced

**Teaching Assistant** Fall 2019
MATH-011: Calculus I, UC Merced

**Teaching Assistant** Summer 2019
MATH-011: Calculus I, UC Merced

**Statistics Tutor for Graduate Students and Postdocs** Spring 2019
Graduate Division, UC Merced

**Teaching Assistant** Fall 2018
BIO-018: Statistics for Scientific Data Analysis, UC Merced

**Statistics Tutor for Graduate Students and Postdocs** Fall 2017
Graduate Division, UC Merced

**Teaching Assistant** Fall 2017
MATH-015: Introduction to Scientific Data Analysis, UC Merced

**Teaching Assistant** Spring 2017
BIO-184: Object Oriented Programming for Biologists, UC Merced

**Teaching Assistant** Fall 2016
MATH-015: Introduction to Scientific Data Analysis, UC Merced

**STEM Instructor** Summer 2016
Summer workshop for pre-K to 5th grade, Harvest Park Education Center

**Teaching Assistant** Spring 2016
BIO-175: Biostatistics, UC Merced

**Teaching Assistant** Fall 2015
BIO-001: Contemporary Biology, UC Merced

**Mentor** Spring 2014
Undergraduate Research Apprentice Program (L&S), UC Berkeley

## Invited lectures

**Thaddeus Seher**. Making hard genome editing easy with CRISPR/Cas-induced homology-directed repair. Quantitative and Systems Biology seminar series, UC Merced (March 2021).

**Thaddeus Seher**. Subtleties of BLAST parameters. BIO-184: Object Oriented Programming for Biologists, UC Merced (April 2017).

**Thaddeus Seher**. NumPy data structures and Matplotlib recipes. BIO-184: Object Oriented Programming for Biologists, UC Merced (April 2017).

**Thaddeus Seher**. Practical steps to forge a career in science (for undergraduates). Society for Advancement of Chicanos/Hispanics and Native Americans in Science, UC Merced (April 2016).

**Thaddeus Seher**. $\chi^2$ tests: contingency tables, directionality, & estimators. BIO-175: Biostatistics, UC Merced (April 2016).

**Thaddeus Seher**. t-tests: $\alpha$, confidence intervals, & hypothesis directionality. BIO-175: Biostatistics, UC Merced (March 2016).

## Honors and awards

**Remote Teaching and Research Fellowship** April 2021
Quantitative and Systems Biology, UC Merced

**Dissertation Incentive Award** April 2020
School of Natural Sciences, UC Merced

**Summer Teaching Assistant Top-off Award** March 2020
School of Natural Sciences, UC Merced

**Qualifying Exam Award** April 2019
Quantitative and Systems Biology, UC Merced

**Honorable Mention, Graduate Research Fellowship Program** March 2017
National Science Foundation

**Summer Research Fellowship** March 2017
Quantitative and Systems Biology, UC Merced

**Program for Excellence in Science 3-year Sponsorship** August 2016
American Association for the Advancement of Science

**Honorable Mention, Graduate Research Fellowship Program** March 2016
National Science Foundation

**Graduate Award** April 2016
UC Merced

**Summer Fellowship** March 2016
Quantitative and Systems Biology, UC Merced

**Graduate Fellowship Incentive Program Award** Jan 2016
UC Merced

**Honors Student** Fall 2006 – Spring 2009
Davis Honors Challenge, UC DAVIS

## Professional enterprises

**President** 2018 – 2019
Graduate Student Association, UC MERCED

**Founder & Chief Technology Officer** 2016 – 2018
RadioBio Podcast, UC MERCED

**Vice President** 2015 – 2017
Quantitative Project, UC MERCED

**Founder & Committee Member** 2012 – 2018
BAFC Scholars College Endowment Fund

# Abstract

Thaddeus D. Seher's "A computational approach for microbial genome editing" is submitted in partial fulfillment of the Doctor of Philosophy degree in Quantitative and Systems Biology at the University of California, Merced, awarded in 2021. The committee in charge is Dr. Aaron D. Hernday (Chair), Dr. Chris T. Amemiya (Member), Dr. Zhong Wang (Member), Dr. Suzanne S. Sindi (Advisor), and Dr. Clarissa J. Nobile (Advisor).

This dissertation describes the work that I performed with a team to develop the AddTag method for genome editing. First, I introduce genome editing through CRISPR/Cas-induced homology-directed repair, and I introduce the *Candida albicans* biological system (Chapter 1). Next, I describe how AddTag editing utilizes CRISPR/Cas-induced homology-directed repair to edit the *C. albicans* genome, and then the process of validating the edits through phenotyping, sequencing, and PCR (Chapter 2). I introduce the ADDTAG software which assists with AddTag editing. First, I describe how ADDTAG identifies genome targets for RNA-guided nucleases (Chapter 3). Then I demonstrate how the software constructs artificial sequences for use as genome repair templates. Lastly, I explain a computational method for producing a set of verification PCR primers for determining if genome edits are successful (Chapter 4).

# Chapter 1 Background

## 1.1 Introduction

This dissertation focuses on a new method of genome editing called AddTag. In Chapter 1, I first introduce the historic and technical concepts of genome editing (1.2). Next, I detail the components for CRISPR/Cas genome editing (1.3) used in Chapter 2, Chapter 3, and Chapter 4. Then, I describe the *Candida albicans* model system—how *C. albicans* affects human health, and why we chose to edit the *C. albicans* genome (1.4). Chapter 2 focuses on editing the *Candida albicans* genome using AddTag. In this dissertation, I present the ADDTAG software for use with designing AddTag genome editing experiments. Chapter 3 describes the computational method for obtaining oligonucleotide sequences used in AddTag editing. Finally, Chapter 4 communicates how ADDTAG identifies PCR primers used for validating genome edits.

## 1.2 Genome editing is fundamental to microbial genetics

In Chapter 2 of this manuscript, we edit the genome of the fungus *Candida albicans*. Genome editing has been used extensively in yeast model systems for various pursuits: economic (agricultural, pharmaceutical, and biotechnological) [1-5], medical [6], and environmental [7]. Genome editing in model organisms is also used to attain knowledge of cellular functions of other organisms [8]. Genome editing is a field of study focused on changing the heritable genetic information, encoded in chromosomes made of deoxyribonucleic acid (DNA), within living organisms. Genome editing is applied to large-scale manufacturing of biomolecules, therapeutic treatment of genetic diseases, environmental remediation, and crop improvement. Some examples, in no particular order, include: creation of caffeine-free coffee beans [9, 10], increased standing variation in banana cultivar genomes [11, 12], deactivation of cancer cells [13], treatment of muscular dystrophy [14], opioid biosynthesis in yeast [15], and addition of a visual marker for the sex of chicken eggs [16]. Artificial genome editing has been applied across all domains of life (bacteria [17], archaea [18, 19], amoebozoa [20, 21], fungi [22], plants [23-27], and animals [28, 29]).

There are two broad categories of genome editing: (1) inserting exogenous DNA into a heritable DNA element (such as a chromosome), and (2) inducing an organism to modify its own genome. Many researchers use a combination of these two techniques to bring about an intended change in the genome.

In the first method (1), exogenous DNA is inserted into the cell; the exogenous DNA is incorporated into the genome; and finally, the modified individuals are assayed to determine if editing was successful. One example is using electroporation to insert a plasmid into *C. albicans*, the cells propagate the plasmid, and the resultant cells are evaluated for transformation-specific selectable markers [30]. Examples of DNA delivery methods include microinjection [31], chemical induction [32], electroporation [33], viral transduction [34, 35], heat-shock, bacterial conjugation, natural uptake (transformation) [36, 37], liposome/micelle transmission [38-40], and biolistic particle delivery [41-43]. In general, these methods all enable artificial horizontal gene transfer. Following transfer of experimental DNA into the cell, that DNA is incorporated into the genome. There are several means to induce an organism to modify its genome using the exogenous DNA. In this dissertation, we use a nuclease to induce DNA repair [44, 45], but other methods include hijacking the synthesis portion of the cell cycle (recombineering) [46, 47], and recombination [48-51].

In the second category of genome editing (2), individuals are subjected to an environment that drives them to use their endogenous cellular machinery to modify their own genomes. Certain carbon sources can induce C. *albicans* to duplicate portions of its chromosomes [52]. Another example is that growing C*andida albicans* under non-physiological temperatures can induce chromosomal loss of heterozygosity (LOH) at the mating type like (*MTL*) locus [53, 54].

Following either genome editing method (1 or 2), the experimental organisms are assayed for genomic modifications. Broadly speaking, these methods are polymerase chain reaction (PCR)-based, sequencing-based, and linkage-based. Examples include differential real-time PCR [55], whole genome sequencing, or phenotyping by linkage with a selectable marker. Many genome edits are phenotypically associated with ribonucleic acid (RNA) transcript production. Thus, RNA-based assays—like quantitative, reverse transcription PCR (RT-qPCR) and transcriptome sequencing—can be appropriate as well.

In this dissertation, I describe the results of delivering a DNA repair template into C*andida albicans* cells through electroporation, and I report the results of incorporating that exogenous template into the genome through a programmable nuclease and endogenous homology-directed repair (HDR) (Chapter 2). Then, I show the genomes were edited as intended using both an amplicon sequencing-based method as well as a diagnostic PCR-based method (Chapter 2). We also use three separate phenotype-linked assays to evaluate edits to the C. *albicans ADE2*, *BRG1*, *EFG1*, *ZAP1*, and *ZRT2* loci (Chapter 2). I next describe how the nuclease was programmed (Chapter 3), and how the PCR primers were identified (Chapter 4).

## 1.3 CRISPR/Cas-directed homology-directed repair is used to edit genomes

To edit the C. *albicans* genome (Chapter 2), we use a molecular technique based on Clustered, regularly-interspaced, short, palindromic repeat (CRISPR) sequences found in bacteria and archaea. A subset of the family of genes found to be CRISPR-associated (Cas) are RNA-guided (endo)nucleases (RGNs), which typically couple a nuclease, helicase, or polymerase domain with a poly-nucleotide binding domain [56]. RGNs are a class of programmable nucleases, which include zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) [57]. RGN examples include both single-molecule proteins like Cas9 and Cas12a, as well as protein complexes like Cas3. In this dissertation, we use Cas9 to induce double-stranded breaks (DSB) in the C. *albicans* genome at specific locations called “Targets.”

To cut the chromosomal DNA, the RGN first forms a complex with a guide RNA (gRNA). The gRNA is a single RNA molecule whose sequence is composed of two parts: a Spacer region and a Scaffold region [58]. This gRNA is a synthetic fusion of the crRNA and tracrRNA sequences [59] typically found in bacteria, and provides both targeting specificity (Spacer) and RGN binding ability (Scaffold) for nuclease activity [58, 60-62]. Spacers used for artificial genome manipulation are commonly 20 nucleotides (nt) in length [63, 64]. Using a gRNA simplifies the components needed for genome editing. Once the gRNA:RGN complex forms, it preferentially associates with genome segments containing a protospacer-adjacent motif (PAM) sequence [65]. The PAM sequence is a few nucleotides that serve as a binding signal for the RGN [60], and its presence is a strict requirement for most RGN-mediated DNA cleavage. For Cas9 the gRNA’s Spacer segment hybridizes with the Target, and the

PAM sequence is on the chromosome just outside the 3' end of the hybridization zone. If both the RGN:PAM interface and the Spacer:Target interface bind strongly enough, the RGN will cleave the chromosomal DNA.

The subsequent DSB is either fatal, mutagenic, or corrected through homology-directed repair (HDR). When two segments of DNA share an abundance of identically-ordered base pairs, they are referred to as sequence homologs. HDR is a DNA repair mechanism that uses a template to repair nicks or double-stranded breaks (DSBs). Most molecular characterizations of HDR present it as an error-free process that removes sequence around the DNA break, then replaces it with sequence from the template DNA through the process of recombination. Recombination is the process of transferring nucleotide segments between DNA molecules. In this dissertation, I term the sequence that is removed from the chromosome the "Feature." Features are specific chromosomal locations intended to be engineered, and are each defined by a name, contig, strand, start position, and end position. A Target is considered on-target if it lies within the bounds of the Feature, and it is considered off-target if it lies outside a Feature's bounds.

In the experiments presented in this dissertation, chromosomal DSBs induce DNA recombination between chromosomal DNA and an artificially-introduced donor DNA (dDNA). The sequences on either end of the dDNA, called homology arms (HAs), contain DNA identical to the DNA flanking the Feature on the genome. The homology allows the endogenous DNA repair machinery to use the dDNA as a template to repair the artificially cut chromosomes, thereby replacing the entire sequence of the Feature with the dDNA. Therefore, following restriction at the Target by the RGN, the middle of the dDNA sequence is added to the genome (knocked in) using endogenous cellular homologous recombination machinery.

This dissertation presents a novel computational method for choosing genomic Targets and dDNA sequences based on arbitrary genomic Features (Chapter 3). There are three specific problems with typical genome editing [66] through CRISPR/Cas-induced HDR that we address. First, because the dDNA sequence must omit the original Target [67, 68], the Target needs to be disrupted so it does not exist in the edited genome. Second, the genomic Feature and exogenous dDNA must not share homology, and the HAs shared between the dDNA and Feature should not repeat throughout either of them. Third, since edits must include sequence changes to the Target, and the dDNA must not have homology in excess of the flanking arms, the Target needs to be within or adjacent to the Feature.

## 1.4 The *Candida albicans* biological system

In Chapter 2 of this dissertation, we edit the genome of the fungus *Candida albicans* using the novel AddTag method we describe. Editing the genome of C. *albicans* has the potential to provide insight into several aspects of human health. C. *albicans* has a close association with humans—it is found on the skin [69-71], ocular surface [72], the oropharyngeal tissues [73, 74], the gut [75-77], and the genitourinary system [78, 79] (Figure 1.1).

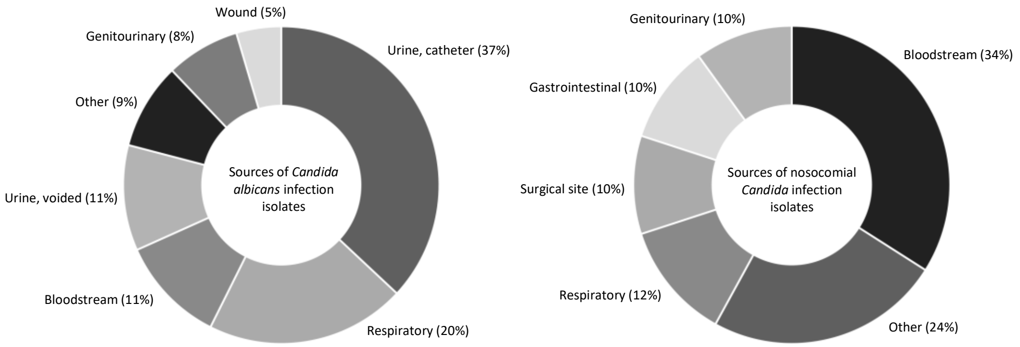


**Figure 1.1 – Sources of *Candida albicans* and other *Candida* species infection isolates**

Two representative studies with body locations of *Candida* and C. *albicans* isolates. Source locations of nosocomial and non-nosocomial C. *albicans* isolates are from 1,005 patients from a Korean hospital [80] (left). Source locations of healthcare associated infections attributed to *Candida* species are from 32 patients in a survey of 11,282 patients across 183 hospitals in the United States [81] (right).

C. *albicans* typically exist in a commensal state with their human hosts—neither harming, nor measurably helping the person. In individuals who are immunocompromised or immune deficient, C. *albicans* can become pathogenic [82]. When a patient acquires an infection during their hospital stay, medical practitioners record this as a nosocomial infection, which make up a large fraction of observed C. *albicans* isolates. Bloodstream (systemic) C. *albicans* infections are considered the most dangerous, and have a 40-65% lethality rate, even with comprehensive medical treatment [83-87].

Most human-associated *Candida* samples are isolated when a person demonstrates abnormal symptoms, and then a culture is taken at a medical clinic. When the symptoms are attributed to fungal overgrowth, the isolate is considered an invasive fungal isolate. *Candida* species consistently account for ~95% of invasive fungal infections in humans [88-90], with C. *albicans* consistently comprising ~65% or more of invasive isolates from hospital settings (Figure 1.2, right) [85]. However, not all fungal isolates are associated with disease (Figure 1.2, left).

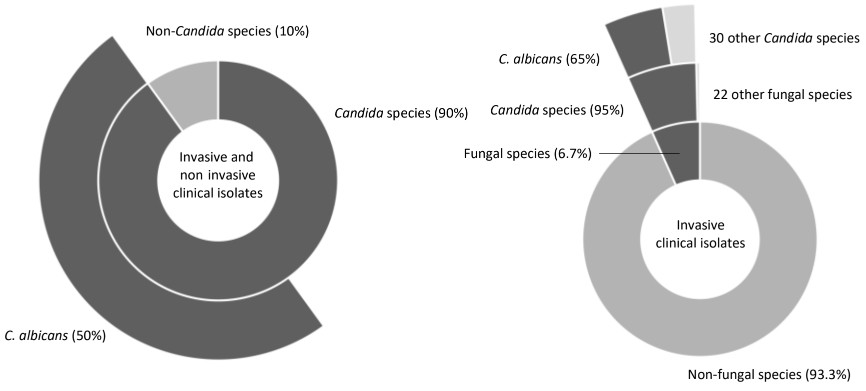


**Figure 1.2 – Proportion of fungal, clinical isolates that are *Candida albicans*.**

(Left) Frequency of C. *albicans* in fungal clinical isolates. (Right) Percentage of invasive clinical isolates that are C. *albicans*. The percentages are obtained from a 10-year global survey spanning 256,882 fungal isolates from 142 sites in 41 countries [80, 90, 91].

One reason C. *albicans* is pervasive is because it can form biofilms—structured cellular communities that are resistant to stressors like antimicrobial compounds [92] and immune cells [93]. Biofilms are encased in a protective substance, called an extracellular matrix, and are often cellular reservoirs for reinfection [94-96]. C. *albicans* biofilms contain a network of different C. *albicans* cell types [97] that can rapidly adapt to evade the human immune system [98]. Because C. *albicans* can form biofilms by itself (monomicrobial) or with other species (polymicrobial) [99], it is well-adapted to living in human hosts.

Modifications to the C. *albicans* genome can potentially reveal details about the biological mechanisms within the fungus that lead to human disease. In 2.4.3, we edit the *EFG1* and *BRG1* genes that directly influence C. *albicans* hyphal growth [100], which is essential for robust biofilm development. In 2.4.1 we edit the *WOR1* [101] and *WOR2* [102] genes involved in cell type switching between the "white" and "opaque" states [103]. Finally, in 2.4.2 and 2.4.4 we edit the *ZAP1* and *ZRT2* genes involved in cellular response to zinc, which is implicated in regulation of the "goliath" cell type [104] as well as the biofilm's protective extracellular matrix [105]. By editing the C. *albicans* genome, we demonstrate the utility of the AddTag approach.

At the time of this writing, the National Center for Biotechnology Information (NCBI)—one central, public repository for sequence databases—lists 70 C. *albicans* genome assemblies and contains information on 91 different strains [106]. This is an under-representation of the total genome data available for C. *albicans* across the literature. The first published C. *albicans* whole-genome sequence assembly was of strain SC5314 in 2004 because of its already widespread use in molecular biology labs [107]. Over time, that assembly has been revised, and it now represents a full haplotype-resolved (phased) genome assembly [108, 109]. SC5314 has subsequently maintained a high frequency of use; and SC5314 is considered the most commonly-accessible, described C. *albicans* strain. Since its isolation [110], SC5314 has been manipulated in the lab to introduce or remove biological functions in order to make it more amenable to genetic manipulations and laboratory conditions. Therefore, we use the genome information for SC5314 as the reference for our bioinformatics-based approach, and we leverage the assembly completeness to provide increased software prediction accuracy.

C. *albicans* is frequently a diploid yeast [111-113], although it can be induced to stably exist in haploid [113-117] or tetraploid [118-122] states [111, 112]. While this dissertation demonstrates only how to perform homozygous genome edits (Chapter 2), it also describes how the AddTag system handles polyploid genomes and can perform allele-specific genome edits (Chapter 3).

## 1.5 References

1. Parolini G. Building human and industrial capacity in European biotechnology: The Yeast Genome Sequencing Project (1989–1996). 2018. DOI: 10.14279/depositonce-6693.
2. Dzialo MC, Park R, Steensels J, Lievens B, Verstrepen KJ. Physiology, ecology and industrial applications of aroma formation in yeast. *FEMS Microbiology Reviews*. 2017; 41(Supp_1):S95-S128. DOI: 10.1093/femsre/fux031, PMID: 28830094.

3. Banat IM, Marchant R. Characterization and potential industrial applications of five novel, thermotolerant, fermentative, yeast strains. *World Journal of Microbiology and Biotechnology*. 1995; 11(3):304-6. DOI: 10.1007/BF00367104, PMID: 24414653.
4. Morata A, Loira I. Yeast: Industrial Applications: BoD–Books on Demand; 2017 November 8, 2017. DOI: 10.5772/intechopen.69360.
5. Satyanarayana T, Kunze G. Yeast biotechnology: Diversity and applications. Netherlands: Springer; 2009. DOI: 10.1007/978-1-4020-8292-4.
6. Boroumand Moghaddam A, Namvar F, Moniri M, Md. Tahir P, Azizi S, Mohamad R. Nanoparticles biosynthesized by fungi and yeast: A review of their preparation, properties, and medical applications. *Molecules*. 2015; 20(9):16540-65. DOI: 10.3390/molecules200916540, PMID: 26378513.
7. Bahafid W, Joutey NT, Asri M, Sayel H, Tirry N, El Ghachtouli N, Sayel N. Yeast biomass: An alternative for bioremediation of heavy metals. *Yeast: Industrial Applications*. 2017:269-89. DOI: 10.5772/intechopen.70559.
8. Baghban R, Farajnia S, Rajabibazl M, Ghasemi Y, Mafi A, Hoseinpoor R, Rahbarnia L, Aria M. Yeast expression systems: Overview and recent advances. *Molecular Biotechnology*. 2019; 61(5):365-84. DOI: 10.1007/s12033-019-00164-8, PMID: 30805909.
9. Tran HTM, Ramaraj T, Furtado A, Lee LS, Henry RJ. Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. *Plant Biotechnology Journal*. 2018; 16(10):1756-66. DOI: 10.1111/pbi.12912, PMID: 29509991.
10. Breitler J-C, Dechamp E, Campa C, Zebral Rodrigues LA, Guyot R, Marraccini P, Etienne H. CRISPR/Cas9-mediated efficient targeted mutagenesis has the potential to accelerate the domestication of *Coffea canephora*. *Plant Cell, Tissue and Organ Culture (PCTOC)*. 2018; 134(3):383-94. DOI: 10.1007/s11240-018-1429-2.
11. Naim F, Dugdale B, Kleidon J, Brinin A, Shand K, Waterhouse P, Dale J. Gene editing the phytoene desaturase alleles of Cavendish banana using CRISPR/Cas9. *Transgenic Research*. 2018; 27(5):451-60. DOI: 10.1007/s11248-018-0083-0, PMID: 29987710.
12. Tripathi L, Ntui VO, Tripathi JN. CRISPR/Cas9-based genome editing of banana for disease resistance. *Current Opinion in Plant Biology*. 2020; 56:118-26. DOI: 10.1016/j.pbi.2020.05.003, PMID: 32604025.
13. Chen Z-H, Yu YP, Zuo Z-H, Nelson JB, Michalopoulos GK, Monga S, Liu S, *et al*. Targeting genomic rearrangements in tumor cells through Cas9-mediated insertion of a suicide gene. *Nature Biotechnology*. 2017; 35(6):543-50. DOI: 10.1038/nbt.3843, PMID: 28459452.
14. Amoasii L, Hildyard JCW, Li H, Sanchez-Ortiz E, Mireault A, Caballero D, Harron R, *et al*. Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. *Science (New York, NY)*. 2018; 362(6410):86-91. DOI: 10.1126/science.aau1549, PMID: 30166439.
15. Galanie S, Thodey K, Trenchard IJ, Filsinger Interrante M, Smolke CD. Complete biosynthesis of opioids in yeast. *Science (New York, NY)*. 2015; 349(6252):1095-100. DOI: 10.1126/science.aac9373, PMID: 26272907.
16. Cremer J. How CRISPR can create more ethical eggs: *Alliance for Science, Cornell University*; 2021 [updated March 2, 2021; cited 2021 March 10, 2021]. Available from: https://allianceforscience.cornell.edu/blog/2021/03/how-crispr-can-create-more-ethical-eggs/.

17. Vento JM, Crook N, Beisel CL. Barriers to genome editing with CRISPR in bacteria. *Journal of Industrial Microbiology and Biotechnology*. 2019; 46(9-10):1327-41. DOI: 10.1007/s10295-019-02195-1, PMID: 31165970.

18. Gophna U, Allers T, Marchfelder A. Finally, Archaea get their CRISPR-Cas toolbox. *Trends in Microbiology*. 2017; 25(6):430-2. DOI: 10.1016/j.tim.2017.03.009, PMID: 28391963.

19. Nayak DD, Metcalf WW. Cas9-mediated genome editing in the methanogenic archaeon *Methanosarcina acetivorans*. *Proceedings of the National Academy of Sciences*. 2017; 114(11):2976-81. DOI: 10.1073/pnas.1618596114, PMID: 28265068.

20. Sekine R, Kawata T, Muramoto T. CRISPR/Cas9 mediated targeting of multiple genes in *Dictyostelium*. *Scientific Reports*. 2018; 8(1):8471. DOI: 10.1038/s41598-018-26756-z, PMID: 29855514.

21. Muramoto T, Iriki H, Watanabe J, Kawata T. Recent advances in CRISPR/Cas9-mediated genome editing in *Dictyostelium*. *Cells*. 2019; 8(1):46. DOI: 10.3390/cells8010046, PMID: 30642074.

22. El-Sayed ASA, Abdel-Ghany SE, Ali GS. Genome editing approaches: Manipulating of lovastatin and taxol synthesis of filamentous fungi by CRISPR/Cas9 system. *Applied Microbiology and Biotechnology*. 2017; 101(10):3953-76. DOI: 10.1007/s00253-017-8263-z, PMID: 28389711.

23. Feng Z, Zhang B, Ding W, Liu X, Yang D-L, Wei P, Cao F*, et al*. Efficient genome editing in plants using a CRISPR/Cas system. *Cell Research*. 2013; 23(10):1229-32. DOI: 10.1038/cr.2013.114, PMID: 23958582.

24. Bortesi L, Fischer R. The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances*. 2015; 33(1):41-52. DOI: 10.1016/j.biotechadv.2014.12.006, PMID: 25536441.

25. Jaganathan D, Ramasamy K, Sellamuthu G, Jayabalan S, Venkataraman G. CRISPR for crop improvement: An update review. *Frontiers in Plant Science*. 2018; 9(985). DOI: 10.3389/fpls.2018.00985, PMID: 30065734.

26. Vats S, Kumawat S, Kumar V, Patil GB, Joshi T, Sonah H, Sharma TR, Deshmukh R. Genome editing in plants: Exploration of technological advancements and challenges. *Cells*. 2019; 8(11):1386. DOI: 10.3390/cells8111386, PMID: 31689989.

27. Arora L, Narula A. Gene editing and crop improvement using CRISPR-Cas9 system. *Frontiers in Plant Science*. 2017; 8(1932). DOI: 10.3389/fpls.2017.01932, PMID: 29167680.

28. Zarei A, Razban V, Hosseini SE, Tabei SMB. Creating cell and animal models of human disease by genome editing using CRISPR/Cas9. *The Journal of Gene Medicine*. 2019; 21(4):e3082. DOI: 10.1002/jgm.3082, PMID: 30786106.

29. Birling M-C, Herault Y, Pavlovic G. Modeling human disease in rodents by CRISPR/Cas9 genome editing. *Mammalian Genome*. 2017; 28(7):291-301. DOI: 10.1007/s00335-017-9703-x, PMID: 28677007.

30. Bijlani S, Thevandavakkam MA, Tsai H-J, Berman J. Autonomously replicating linear plasmids facilitate the analysis of replication origin function in *Candida albicans*. *bioRxiv*. 2019:551127. DOI: 10.1101/551127.

31. Diacumakos EG. Chapter 15 – Methods for Micromanipulation of Human Somatic Cells in Culture. In: Prescott DM, editor. Methods in Cell Biology. 7: Academic Press; 1974; p. 287-311. DOI: 10.1016/S0091-679X(08)61783-5.

32. Graham FL, van der Eb AJ. A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology*. 1973; 52(2):456-67. DOI: 10.1016/0042-6822(73)90341-3, PMID: 4705382.

33. Neumann E, Schaefer-Ridder M, Wang Y, Hofschneider PH. Gene transfer into mouse lyoma cells by electroporation in high electric fields. *The EMBO journal*. 1982; 1(7):841-5. DOI: 10.1002/j.1460-2075.1982.tb01257.x, PMID: 6329708.

34. Mulligan RC, Howard BH, Berg P. Synthesis of rabbit β-globin in cultured monkey kidney cells following infection with a SV40 β-globin recombinant genome. *Nature*. 1979; 277(5692):108-14. DOI: 10.1038/277108a0, PMID: 215915.

35. Hamer DH, Leder P. Splicing and the formation of stable RNA. *Cell*. 1979; 18(4):1299-302. DOI: 10.1016/0092-8674(79)90240-X, PMID: 229971.

36. Griffith F. The significance of pneumococcal types. *Journal of Hygiene*. 1928; 27(2):113-59. Epub 2009/05/15. DOI: 10.1017/S0022172400031879, PMID: 20474956.

37. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J Exp Med*. 1944; 79(2):137-58. DOI: 10.1084/jem.79.2.137, PMID: 19871359.

38. Fraley R, Subramani S, Berg P, Papahadjopoulos D. Introduction of liposome-encapsulated SV40 DNA into cells. *Journal of Biological Chemistry*. 1980; 255(21):10431-5. DOI: 10.1016/S0021-9258(19)70482-7, PMID: 6253474.

39. Tai-Kin W, Nicolau C, Hofschneider PH. Appearance of β-lactamase activity in animal cells upon liposome-mediated gene transfer. *Gene*. 1980; 10(2):87-94. DOI: 10.1016/0378-1119(80)90126-2, PMID: 6248423.

40. Schaefer-Ridder M, Wang Y, Hofschneider P. Liposomes as gene carriers: Efficient transformation of mouse L cells by thymidine kinase gene. *Science (New York, NY)*. 1982; 215(4529):166-8. DOI: 10.1126/science.7053567, PMID: 7053567.

41. Klein TM, Wolf ED, Wu R, Sanford JC. High-velocity microprojectiles for delivering nucleic acids into living cells. *Nature*. 1987; 327(6117):70-3. DOI: 10.1038/327070a0, PMID: 1422046.

42. Sanford JC, Klein TM, Wolf ED, Allen N. Delivery of substances into cells and tissues using a particle bombardment process. *Particulate Science and Technology*. 1987; 5(1):27-37. DOI: 10.1080/02726358708904533.

43. Segelken R. Biologists invent gun for shooting cells with DNA. Cornell Chronicle. 1987 May 14, 1987.

44. Nguyen N, Quail MMF, Hernday AD. An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. *mSphere*. 2017; 2(2). DOI: 10.1128/mSphereDirect.00149-17, PMID: 28497115.

45. Vyas VK, Barrasa MI, Fink GR. A *Candida albicans* CRISPR system permits genetic engineering of essential genes and gene families. *Science advances*. 2015; 1(3):e1500248. DOI: 10.1126/sciadv.1500248, PMID: 25977940.

46. Moerschell RP, Tsunasawa S, Sherman F. Transformation of yeast with synthetic oligonucleotides. *Proceedings of the National Academy of Sciences*. 1988; 85(2):524-8. DOI: 10.1073/pnas.85.2.524, PMID: 2829192.

47. Mosberg JA, Lajoie MJ, Church GM. Lambda red recombineering in *Escherichia coli* occurs through a fully single-stranded intermediate. *Genetics*. 2010; 186(3):791-9. Epub 2010/09/02. DOI: 10.1534/genetics.110.120782, PMID: 20813883.

48. Gu P, Yang F, Su T, Wang Q, Liang Q, Qi Q. A rapid and reliable strategy for chromosomal integration of gene(s) with multiple copies. *Scientific Reports*. 2015; 5(1):9684. DOI: 10.1038/srep09684, PMID: 25851494.

49. Senecoff JF, Bruckner RC, Cox MM. The FLP recombinase of the yeast 2-micron plasmid: Characterization of its recombination site. *Proceedings of the National*

*Academy of Sciences*. 1985; 82(21):7270-4. DOI: 10.1073/pnas.82.21.7270, PMID: 2997780.

50. Martin RM, Ikeda K, Cromer MK, Uchida N, Nishimura T, Romano R, Tong AJ*, et al*. Highly efficient and marker-free genome editing of human pluripotent stem cells by CRISPR-Cas9 RNP and AAV6 donor-mediated homologous recombination. *Cell Stem Cell*. 2019; 24(5):821-8.e5. DOI: 10.1016/j.stem.2019.04.001, PMID: 31051134.

51. Rogers GL, Chen H-Y, Morales H, Cannon PM. Homologous recombination-based genome editing by clade F AAVs is inefficient in the absence of a targeted DNA break. *Molecular Therapy*. 2019; 27(10):1726-36. DOI: 10.1016/j.ymthe.2019.08.019, PMID: 31540849.

52. Kabir MA, Ahmad A, Greenberg JR, Wang Y-K, Rustchenko E. Loss and gain of chromosome 5 controls growth of *Candida albicans* on sorbose due to dispersed redundant negative regulators. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(34):12147-52. DOI: 10.1073/pnas.0505625102, PMID: 16099828.

53. Wu W, Pujol C, Lockhart SR, Soll DR. Chromosome loss followed by duplication is the major mechanism of spontaneous mating-type locus homozygosis in *Candida albicans*. *Genetics*. 2005; 169(3):1311-27. DOI: 10.1534/genetics.104.033167, PMID: 15654090.

54. Lachke SA, Lockhart SR, Daniels KJ, Soll DR. Skin facilitates *Candida albicans* mating. *Infection and Immunity*. 2003; 71(9):4970-6. DOI: 10.1128/IAI.71.9.4970-4976.2003, PMID: 12933839.

55. Li B, Ren N, Yang L, Liu J, Huang Q. A qPCR method for genome editing efficiency determination and single-cell clone screening in human cells. *Scientific Reports*. 2019; 9(1):18877. DOI: 10.1038/s41598-019-55463-6, PMID: 31827197.

56. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*. 2014; 513:569. DOI: 10.1038/nature13579, PMID: 25079318.

57. Li H, Yang Y, Hong W, Huang M, Wu M, Zhao X. Applications of genome editing technology in the targeted therapy of human diseases: Mechanisms, advances and prospects. *Signal Transduction and Targeted Therapy*. 2020; 5(1):1. DOI: 10.1038/s41392-019-0089-y, PMID: 32296011.

58. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD*, et al*. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, NY)*. 2013:1231143. DOI: 10.1126/science.1231143, PMID: 23287718.

59. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, NY)*. 2012; 337(6096):816-21. Epub 2012/06/28. DOI: 10.1126/science.1225829, PMID: 22745249.

60. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*. 2014; 507:62. DOI: 10.1038/nature13011, PMID: 24476820.

61. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. RNA-guided human genome engineering via Cas9. *Science (New York, NY)*. 2013; 339(6121):823-6. DOI: 10.1126/science.1232033, PMID: 23287722.

62. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology*. 2013; 31(3):230-2. DOI: 10.1038/nbt.2507, PMID: 23360966.

63. Friedland AE, Baral R, Singhal P, Loveluck K, Shen S, Sanchez M, Marco E*, et al*. Characterization of *Staphylococcus aureus* Cas9: A smaller Cas9 for all-in-one

adeno-associated virus delivery and paired nickase applications. *Genome Biology*. 2015; 16(1):257. DOI: 10.1186/s13059-015-0817-8, PMID: 26596280.

64. Tycko J, Barrera LA, Huston NC, Friedland AE, Wu X, Gootenberg JS, Abudayyeh OO*, et al*. Pairwise library screen systematically interrogates Staphylococcus aureus Cas9 specificity in human cells. *Nature Communications*. 2018; 9(1):2962. DOI: 10.1038/s41467-018-05391-2, PMID: 30054474.

65. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, NY)*. 2007; 315(5819):1709-12. DOI: 10.1126/science.1138140, PMID: 17379808.

66. Cai P, Gao J, Zhou Y. CRISPR-mediated genome editing in non-conventional yeasts for biotechnological applications. *Microbial Cell Factories*. 2019; 18(1):63. DOI: 10.1186/s12934-019-1112-2, PMID: 30940138.

67. Satomura A, Nishioka R, Mori H, Sato K, Kuroda K, Ueda M. Precise genome-wide base editing by the CRISPR nickase system in yeast. *Scientific Reports*. 2017; 7(1):2095. DOI: 10.1038/s41598-017-02013-7, PMID: 28522803.

68. Satomura A, Nishioka R, Mori H, Sato K, Kuroda K, Ueda M. Erratum: Precise genome-wide base editing by the CRISPR Nickase system in yeast. *Scientific Reports*. 2017; 7(1):12354. DOI: 10.1038/s41598-017-09606-2, PMID: 28955053.

69. López-García B, Lee PHA, Yamasaki K, Gallo RL. Anti-fungal activity of cathelicidins and their potential role in *Candida albicans* skin infection. *Journal of Investigative Dermatology*. 2005; 125(1):108-15. DOI: 10.1111/j.0022-202X.2005.23713.x, PMID: 15982310.

70. Kashem SW, Kaplan DH. Skin immunity to *Candida albicans*. *Trends in Immunology*. 2016; 37(7):440-50. DOI: 10.1016/j.it.2016.04.007, PMID: 27178391.

71. Kühbacher A, Burger-Kentischer A, Rupp S. Interaction of *Candida* species with the skin. *Microorganisms*. 2017; 5(2):32. DOI: 10.3390/microorganisms5020032, PMID: 28590443.

72. Shivaji S, Jayasudha R, Sai Prashanthi G, Kalyana Chakravarthy S, Sharma S. The human ocular surface fungal microbiome. *Investigative Ophthalmology & Visual Science*. 2019; 60(1):451-9. DOI: 10.1167/iovs.18-26076, PMID: 30703210.

73. Solis NV, Park Y-N, Swidergall M, Daniels KJ, Filler SG, Soll DR. *Candida albicans* white-opaque switching influences virulence but not mating during oropharyngeal candidiasis. *Infection and Immunity*. 2018; 86(6):e00774-17. DOI: 10.1128/iai.00774-17, PMID: 29581190.

74. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLOS Pathogens*. 2010; 6(1):e1000713. DOI: 10.1371/journal.ppat.1000713, PMID: 20072605.

75. Kennedy MJ, Volz PA. Ecology of *Candida albicans* gut colonization: Inhibition of *Candida* adhesion, colonization, and dissemination from the gastrointestinal tract by bacterial antagonism. *Infection and immunity*. 1985; 49(3):654-63. DOI: 10.1128/iai.49.3.654-663.1985, PMID: 3897061.

76. Prieto D, Correia I, Pla J, Román E. Adaptation of *Candida albicans* to commensalism in the gut. *Future Microbiology*. 2016; 11(4):567-83. DOI: 10.2217/fmb.16.1, PMID: 27070839.

77. Kumamoto CA, Gresnigt MS, Hube B. The gut, the bad and the harmless: *Candida albicans* as a commensal and opportunistic pathogen in the intestine. *Current Opinion in Microbiology*. 2020; 56:7-15. DOI: 10.1016/j.mib.2020.05.006, PMID: 32604030.

78. Odds FC, Webster CE, Mayuranathan P, Simmons PD. *Candida* concentrations in the vagina and their association with signs and symptoms of vaginal candidosis. *Journal of Medical and Veterinary Mycology*. 1988; 26(5):277-83. DOI: 10.1080/02681218880000391, PMID: 3236147.
79. Bertholf ME, Stafford MJ. Colonization of *Candida albicans* in vagina, rectum, and mouth. *J Fam Pract*. 1983; 16(5):919-24. Epub 1983/05/01. PMID: 6341500.
80. Kim G-Y, Jeon J-S, Kim JK. Isolation frequency characteristics of *Candida* species from clinical specimens. *Mycobiology*. 2016; 44(2):99-104. DOI: 10.5941/MYCO.2016.44.2.99, PMID: 27433120.
81. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R*, et al*. Multistate point-prevalence survey of health care-associated infections. *New England Journal of Medicine*. 2014; 370(13):1198-208. DOI: 10.1056/NEJMoa1306801, PMID: 24670166.
82. Yamaguchi N, Sonoyama K, Kikuchi H, Nagura T, Aritsuka T, Kawabata J. Gastric colonization of *Candida albicans* differs in mice fed commercial and purified diets. *The Journal of Nutrition*. 2005; 135(1):109-15. DOI: 10.1093/jn/135.1.109, PMID: 15623841.
83. Gudlaugsson O, Gillespie S, Lee K, Berg JV, Hu J, Messer S, Herwaldt L*, et al*. Attributable mortality of nosocomial candidemia, revisited. *Clinical Infectious Diseases*. 2003; 37(9):1172-7. DOI: 10.1086/378745, PMID: 14557960.
84. Colombo AL, Guimarães T, Sukienik T, Pasqualotto AC, Andreotti R, Queiroz-Telles F, Nouér SA, Nucci M. Prognostic factors and historical trends in the epidemiology of candidemia in critically ill patients: An analysis of five multicenter studies sequentially conducted over a 9-year period. *Intensive Care Medicine*. 2014; 40(10):1489-98. DOI: 10.1007/s00134-014-3400-y, PMID: 25082359.
85. Lortholary O, Renaudat C, Sitbon K, Madec Y, Denoeud-Ndam L, Wolff M, Fontanet A*, et al*. Worrisome trends in incidence and mortality of candidemia in intensive care units (Paris area, 2002–2010). *Intensive Care Medicine*. 2014; 40(9):1303-12. DOI: 10.1007/s00134-014-3408-3, PMID: 25097069.
86. Bassetti M, Merelli M, Righi E, Diaz-Martin A, Rosello EM, Luzzati R, Parra A*, et al*. Epidemiology, species distribution, antifungal susceptibility, and outcome of candidemia across five sites in Italy and Spain. *Journal of Clinical Microbiology*. 2013; 51(12):4167-72. DOI: 10.1128/JCM.01998-13, PMID: 24108614.
87. Leroy O, Bailly S, Gangneux J-P, Mira J-P, Devos P, Dupont H, Montravers P*, et al*. Systemic antifungal therapy for proven or suspected invasive candidiasis: The AmarCAND 2 study. *Annals of Intensive Care*. 2016; 6(1):2. DOI: 10.1186/s13613-015-0103-7, PMID: 26743881.
88. Calandra T, Roberts JA, Antonelli M, Bassetti M, Vincent J-L. Diagnosis and management of invasive candidiasis in the ICU: An updated approach to an old enemy. *Critical Care*. 2016; 20(1):1-6. DOI: 10.1186/s13054-016-1313-6, PMID: 27230564.
89. Pfaller MA, Diekema DJ, Rinaldi MG, Barnes R, Hu B, Veselov AV, Tiraboschi N*, et al*. Results from the ARTEMIS DISK Global Antifungal Surveillance Study: A 6.5-year analysis of susceptibilities of *Candida* and other yeast species to fluconazole and voriconazole by standardized disk diffusion testing. *Journal of Clinical Microbiology*. 2005; 43(12):5848-59. DOI: 10.1128/jcm.43.12.5848-5859.2005, PMID: 16333066.
90. Pfaller MA, Diekema DJ, Gibbs DL, Newell VA, Meis JF, Gould IM, Fu W*, et al*. Results from the ARTEMIS DISK Global Antifungal Surveillance Study, 1997 to 2005: An 8.5-year analysis of susceptibilities of *Candida* species and other yeast species to fluconazole and voriconazole determined by CLSI standardized disk

diffusion testing. *Journal of Clinical Microbiology*. 2007; 45(6):1735-45. DOI: 10.1128/jcm.00409-07, PMID: 17442797.

91. Pfaller MA, Diekema DJ, Gibbs DL, Newell VA, Ellis D, Tullio V, Rodloff A*, et al.* Results from the ARTEMIS DISK Global Antifungal Surveillance Study, 1997 to 2007: A 10.5-year analysis of susceptibilities of *Candida* species to fluconazole and voriconazole as determined by CLSI standardized disk diffusion. *Journal of Clinical Microbiology*. 2010; 48(4):1366-77. DOI: 10.1128/jcm.02117-09, PMID: 20164282.

92. Sherry L, Rajendran R, Lappin DF, Borghi E, Perdoni F, Falleni M, Tosi D*, et al.* Biofilms formed by *Candida albicans* bloodstream isolates display phenotypic and transcriptional heterogeneity that are associated with resistance and pathogenicity. *BMC Microbiology*. 2014; 14(1):182. DOI: 10.1186/1471-2180-14-182, PMID: 24996549.

93. Sandai D, Tabana YM, Ouweini AE, Ayodeji IO. Resistance of *Candida albicans* biofilms to drugs and the host immune system. *Jundishapur J Microbiol*. 2016; 9(11):e37385. Epub 2016-09-02. DOI: 10.5812/jjm.37385, PMID: 28138373.

94. Mukherjee PK, Zhou G, Munyon R, Ghannoum MA. *Candida* biofilm: A well-designed protected environment. *Medical mycology*. 2005; 43(3):191-208. Epub 2005/07/14. DOI: 10.1080/13693780500107554, PMID: 16010846.

95. Uppuluri P, Chaturvedi AK, Srinivasan A, Banerjee M, Ramasubramaniam AK, Köhler JR, Kadosh D, Lopez-Ribot JL. Dispersion as an important step in the *Candida albicans* biofilm developmental cycle. *PLOS Pathogens*. 2010; 6(3):e1000828. DOI: 10.1371/journal.ppat.1000828, PMID: 20360962.

96. Chandra J, Kuhn DM, Mukherjee PK, Hoyer LL, McCormick T, Ghannoum MA. Biofilm formation by the fungal pathogen *Candida albicans*: Development, architecture, and drug resistance. *Journal of Bacteriology*. 2001; 183(18):5385-94. DOI: 10.1128/JB.183.18.5385-5394.2001, PMID: 11514524.

97. Noble SM, Gianetti BA, Witchley JN. *Candida albicans* cell-type switching and functional plasticity in the mammalian host. *Nature Reviews Microbiology*. 2017; 15(2):96-108. DOI: 10.1038/nrmicro.2016.157, PMID: 27867199.

98. Sasse C, Hasenberg M, Weyler M, Gunzer M, Morschhäuser J. White-opaque switching of *Candida albicans* allows immune evasion in an environment-dependent fashion. *Eukaryotic Cell*. 2013; 12(1):50-8. DOI: 10.1128/EC.00266-12, PMID: 23125350.

99. Nobile CJ, Johnson AD. *Candida albicans* biofilms and human disease. *Annual Review of Microbiology*. 2015; 69(1):71-92. DOI: 10.1146/annurev-micro-091014-104330, PMID: 26488273.

100. Nobile Clarissa J, Fox Emily P, Nett Jeniel E, Sorrells Trevor R, Mitrovich Quinn M, Hernday Aaron D, Tuch Brian B*, et al.* A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell*. 2012; 148(1):126-38. DOI: 10.1016/j.cell.2011.10.048, PMID: 22265407.

101. Huang G, Wang H, Chou S, Nie X, Chen J, Liu H. Bistable expression of *WOR1*, a master regulator of white-opaque switching in *Candida albicans*. *Proceedings of the National Academy of Sciences*. 2006; 103(34):12813-8. DOI: 10.1073/pnas.0605270103, PMID: 16905649.

102. Zordan RE, Miller MG, Galgoczy DJ, Tuch BB, Johnson AD. Interlocking transcriptional feedback loops control white-opaque switching in *Candida albicans*. *PLOS Biology*. 2007; 5(10):e256. DOI: 10.1371/journal.pbio.0050256, PMID: 17880264.

103. Hernday AD, Lohse MB, Nobile CJ, Noiman L, Laksana CN, Johnson AD. Ssn6 defines a new level of regulation of white-opaque switching in *Candida albicans*

and is required for the stochasticity of the switch. *mBio*. 2016; 7(1):e01565-15. DOI: 10.1128/mBio.01565-15, PMID: 26814177.

104. Malavia D, Lehtovirta-Morley LE, Alamir O, Weiß E, Gow NAR, Hube B, Wilson D. Zinc limitation induces a hyper-adherent goliath phenotype in *Candida albicans*. *Front Microbiol*. 2017; 8(2238). DOI: 10.3389/fmicb.2017.02238, PMID: 29184547.

105. Nobile CJ, Nett JE, Hernday AD, Homann OR, Deneault J-S, Nantel A, Andes DR, *et al*. Biofilm matrix regulation by *Candida albicans* Zap1. *PLOS Biology*. 2009; 7(6):e1000133. DOI: 10.1371/journal.pbio.1000133, PMID: 19529758.

106. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, *et al*. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database (Oxford)*. 2020; 2020. Epub 2020/08/08. DOI: 10.1093/database/baaa062, PMID: 32761142.

107. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, *et al*. The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(19):7329-34. Epub 2004/05/03. DOI: 10.1073/pnas.0401648101, PMID: 15123810.

108. Muzzey D, Schwartz K, Weissman JS, Sherlock G. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biology*. 2013; 14(9):R97. DOI: 10.1186/gb-2013-14-9-r97, PMID: 24025428.

109. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The *Candida* Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Research*. 2016; 45(D1):D592-D6. DOI: 10.1093/nar/gkw924, PMID: 27738138.

110. Gillum AM, Tsay EYH, Kirsch DR. Isolation of the *Candida albicans* gene for orotidine-5′-phosphate decarboxylase by complementation of *S. cerevisiae ura3* and *E. coli pyrF* mutations. *Molecular and General Genetics MGG*. 1984; 198(1):179-82. DOI: 10.1007/BF00328721, PMID: 6394964.

111. Olaiya AF, Sogin SJ. Ploidy determination of Canadida albicans. *Journal of bacteriology*. 1979; 140(3):1043-9. DOI: 10.1128/JB.140.3.1043-1049.1979, PMID: 391796.

112. Suzuki T, Nishibayashi S, Kuroiwa T, Kanbe T, Tanaka K. Variance of ploidy in *Candida albicans*. *Journal of bacteriology*. 1982; 152(2):893-6. DOI: 10.1128/jb.152.2.893-896.1982, PMID: 6752122.

113. Hickman MA, Zeng G, Forche A, Hirakawa MP, Abbey D, Harrison BD, Wang Y-M, *et al*. The 'obligate diploid' *Candida albicans* forms mating-competent haploids. *Nature*. 2013; 494(7435):55-9. DOI: 10.1038/nature11865, PMID: 23364695.

114. Truong T, Suriyanarayanan T, Zeng G, Le TD, Liu L, Li J, Tong C, *et al*. Use of haploid model of *Candida albicans* to uncover mechanism of action of a novel antifungal agent. *Frontiers in Cellular and Infection Microbiology*. 2018; 8(164). DOI: 10.3389/fcimb.2018.00164, PMID: 29938200.

115. Seneviratne CJ, Zeng G, Truong T, Sze S, Wong W, Samaranayake L, Chan FY, *et al*. New "haploid biofilm model" unravels *IRA2* as a novel regulator of *Candida albicans* biofilm formation. *Scientific Reports*. 2015; 5(1):12433. DOI: 10.1038/srep12433, PMID: 26202015.

116. Segal ES, Gritsenko V, Levitan A, Yadav B, Dror N, Steenwyk JL, Silberberg Y, *et al*. Gene essentiality analyzed by *in vivo* transposon mutagenesis and machine learning in a stable haploid isolate of *Candida albicans*. *mBio*. 2018; 9(5):e02048-18. DOI: 10.1128/mBio.02048-18, PMID: 30377286.

117. Zeng G, Wang Y-M, Chan FY, Wang Y. One-step targeted gene deletion in *Candida albicans* haploids. *Nature Protocols*. 2014; 9(2):464-73. DOI: 10.1038/nprot.2014.029, PMID: 24481273.
118. Bennett RJ. The parasexual lifestyle of *Candida albicans*. *Current Opinion in Microbiology*. 2015; 28:10-7. DOI: 10.1016/j.mib.2015.06.017, PMID: 26210747.
119. Lockhart SR, Daniels KJ, Zhao R, Wessels D, Soll DR. Cell biology of mating in *Candida albicans*. *Eukaryotic Cell*. 2003; 2(1):49-61. DOI: 10.1128/ec.2.1.49-61.2003, PMID: 12582122.
120. Hubbard MJ, Poulter RT, Sullivan PA, Shepherd MG. Characterization of a tetraploid derivative of *Candida albicans* ATCC 10261. *Journal of bacteriology*. 1985; 161(2):781-3. DOI: 10.1128/jb.161.2.781-783.1985, PMID: 3881412.
121. Hilton C, Markie D, Corner B, Rikkerink E, Poulter R. Heat shock induces chromosome loss in the yeast *Candida albicans*. *Molecular and General Genetics MGG*. 1985; 200(1):162. DOI: 10.1007/BF00383330, PMID: 3897792.
122. Bennett RJ, Johnson AD. Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *The EMBO Journal*. 2003; 22(10):2505-15. DOI: 10.1093/emboj/cdg235, PMID: 12743044.

# Chapter 2 Editing the *Candida albicans* genome with the AddTag method

## 2.1 Abstract

CRISPR/Cas-induced genome editing is a powerful tool for genetic engineering; however, targeting constraints limit which loci are editable with this method. Since the length of a DNA sequence impacts the likelihood it overlaps a unique target site, precision editing of small genomic features with CRISPR/Cas remains an obstacle. We introduce AddTag—a novel genome editing strategy that virtually eliminates CRISPR/Cas targeting constraints and facilitates precision genome editing of elements as short as a single base-pair at virtually any locus in any organism that supports CRISPR/Cas-induced genome editing. This two-step approach first replaces the locus of interest with an `addtag` sequence, which is subsequently replaced with any engineered sequence, and thus circumvents the need for direct overlap with a unique CRISPR/Cas target site. In this study, we demonstrate the feasibility of our approach by editing transcription factor binding sites within *Candida albicans* that could not be targeted directly using the typical gene editing approach. We also demonstrate the utility of the AddTag approach for combinatorial genome editing and gene complementation analysis, and we present the ADDTAG software package that automates editing designs (https://github.com/tdseher/addtag-project).

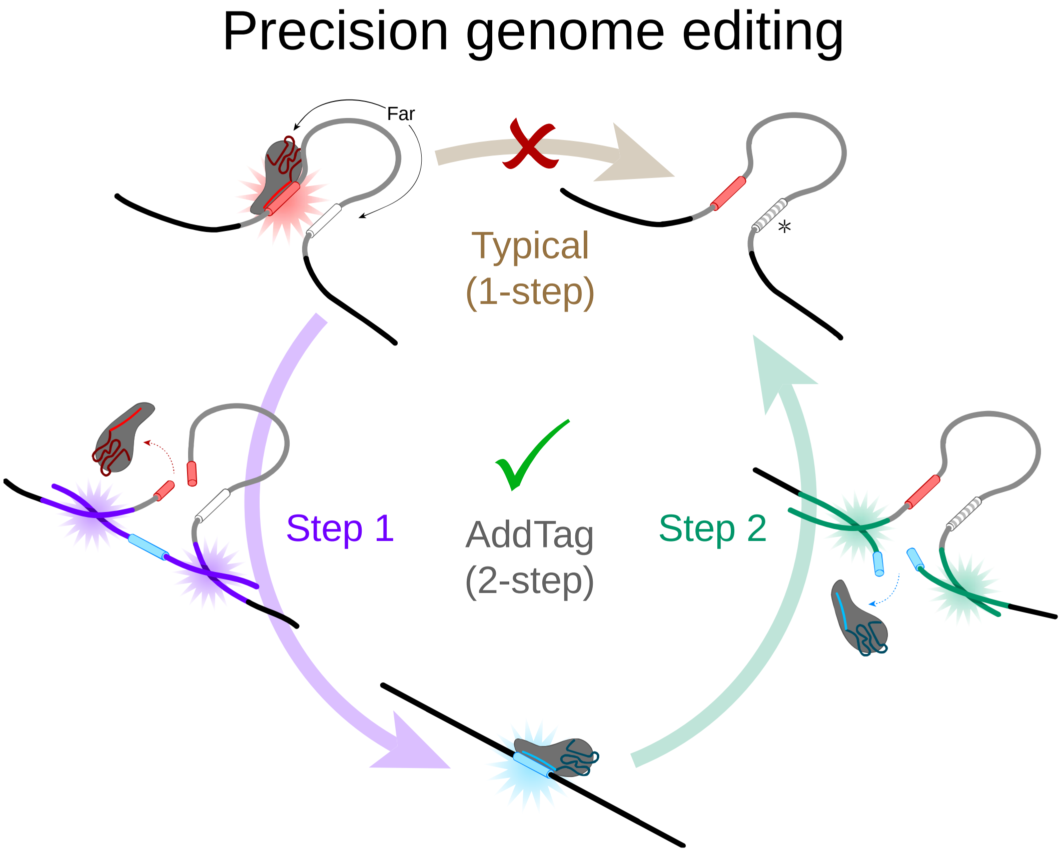


**Figure 2.1 – Graphical abstract**

Typical (1-step) CRISPR/Cas-mediated genome editing (tan) often will fail to edit non-coding genetic elements or loci far from the targeted restriction site. The AddTag (2-step) method surmounts these issues by using two editing events (violet, green).

## 2.2 Preface

This chapter is adapted from the manuscript authored by myself, Namkha Nguyen, Diana Ramos, Priyanka Bapat, Clarissa J. Nobile, Suzanne S. Sindi, and Aaron D. Hernday, titled "AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing," which was published by the peer-reviewed journal *G3 Genes|Genomes|Genetics* [1].

## 2.3 Introduction

### 2.3.1 Description of the Typical (1-step) genome editing method

RNA-guided nucleases (RGNs) such as Cas9 have revolutionized genome editing by enabling the introduction of targeted double stranded breaks (DSBs) within the genomes of living organisms (The process where an RGN cuts DNA is also called DNA restriction/cleavage). These DSBs create a powerful selection for DNA repair, which can be harnessed to promote integration of engineered exogenous "donor" DNA (dDNA) sequences in place of a cut target site, resulting in precision genome edits such as insertions, deletions, or substitutions in the genomes of organisms that support efficient homology directed repair (HDR). Since RGNs can be directed to introduce DSBs at specific user-defined genomic locations through the use of synthetic guide RNAs (gRNAs), this system represents a powerful customizable platform for genome editing [2-4]. Certain constraints, however, limit the flexibility of this technology, particularly when attempting to edit segments of genomic DNA (Features) that are short.

The primary limit is the need for a specific protospacer adjacent motif (PAM) sequence, which is recognized directly by the RGN protein, to be immediately adjacent to the user-defined target site [2, 5-7]. For instance, the commonly used *Streptococcus pyogenes* Cas9 protein recognizes a 3'-adjacent NGG PAM sequence, thus limiting the extent to which A/T-rich sequences, such as non-coding DNA regions, can be targeted. Furthermore, the engineered dDNA must avoid including sequence containing the user-defined RGN target site (defined by the gRNA's spacer region and adjacent PAM sequence) in order to prevent repeated cutting of the repaired target locus [8], which would otherwise result in uncontrolled mutations via non-homologous end joining (NHEJ) (Figure 2.3, Figure 2.4, Figure 2.28) rather than the intended precision editing via HDR (Figure 2.31) [9]. In practicality, successful RGN-mediated genome editing requires the Feature being modified to contain, or overlap, the user-defined RGN target site; and the intended genome edit(s) must ablate, or substantially change, the sequence region of the genome that defines the RGN target (Figure 2.2) [9-11]. While these constraints are generally not significant in the context of deleting or inserting large genetic elements, such as entire protein-coding genes, they substantially limit the number of loci that can be modified with small-scale edits, like cis-acting regulatory elements [12].

**Common Figure Description 1**

Thick horizontal lines represent DNA, with genomic DNA (gDNA) terminating in helices, and donor DNA (dDNA) terminating in blunt ends. The sequence intervening between the genomic region selected for editing (Feature) and the RNA-guided nuclease (RGN) restriction site (Target) is colored gray. The superscript X ($^{X}$) indicates there are unintended genomic changes in the terminal gDNA. The stopwatch (⏱) indicates a transient sequence that is not heritably maintained.

Stretched rectangles labeled with upstream (US) or downstream (DS) indicate regions of homology between different DNA molecules and represent the intended recombination events during HDR.

Rectangles with internal labels represent annotated regions. Rectangles with staggered edges represent DNA breaks at the edge of the annotated region. Rectangles with uncolored, dashed lines indicate the bounds of an expanded Feature. Annotations with striped shading and labels preceded by an asterisk (✻) represent modified sequences. Annotations with effectively identical sequences are shaded the same color. Spacer regions of the guide RNAs (gRNAs) are color-matched with the genomic regions the RNA-guided nuclease (RGN) is programmed to cleave. In general, the Target site that corresponds to the reference (+) genome, and hence the wild-type Feature, is colored red. The insert encoding for either an `addtag` or `mintag` is colored cyan, with the Target denoted by a colored rectangle with dashed lines. Allelic variants on DNA are colored orange, and fixed variants are colored dark grey.

Black arrows represent specific biological processes with substrates at the tail and products at the head. Half-arrows pointing right represent annealing positions of "forward" primers, and half-arrows pointing left denote annealing positions of "reverse" primers. Genomic sites where primers anneal are color-matched to their respective primers.

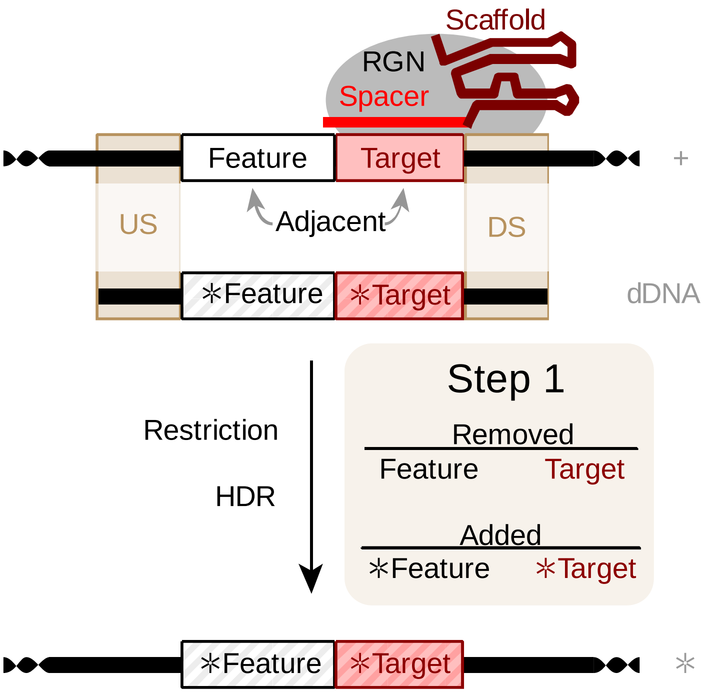


**Figure 2.2 – Direct (1-step) typical method**

The direct (1-step) "typical" method turns the reference (+) genome into the modified (✻) genome by incorporating a single dDNA. Direct editing requires the RGN restriction site (Target) to be disrupted. The modified target site (✻Target) must be sufficiently different from the target sequence to prevent RGN restriction with the same gRNA. Therefore, unless the Feature and Target are largely overlapping, the final modified genomic sequence

(✻gDNA) must contain modifications outside of the modified feature (✻Feature). Any intervening sequence between the feature and target must also be short to ensure that the feature is replaced with the intended ✻Feature. If the ✻Target and ✻Feature sites are not overlapping or closely adjacent, then HDR at the cut target site can result in ✻Target incorporation without ✻Feature incorporation (Figure 4.2).

The grey arrows connecting "Adjacent" to the Feature and Target are for clarity only, and do not represent a biological process. Vertical black arrows represent RGN-mediated cutting of the target locus, followed by dDNA incorporation via HDR. More detail can be found in **Common Figure Description 1**.

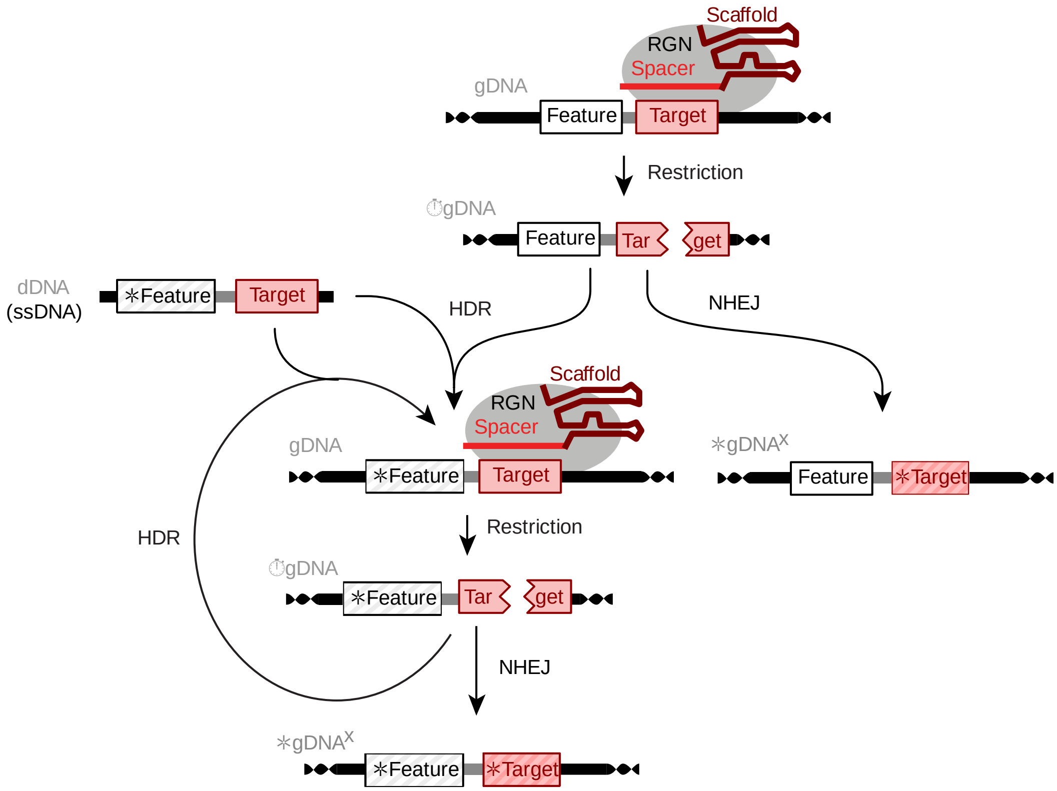


**Figure 2.3 – When single-stranded dDNA contains an identical Target, edits are possible, but the Target is disrupted**

Genomic DSB must be repaired to maintain cell viability. If the input is single-stranded DNA (ssDNA), and the RNA-guided nuclease (RGN) is double-stranded DNA (dsDNA)-specific, then the modified Feature (✻Feature) can be integrated, but the Target is still disrupted through uncontrolled mutation. Otherwise, the Target is disrupted, and the Feature is not modified as intended.

Black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by either homology-directed repair (HDR) as the process by which dDNA is incorporated into the gDNA, or double-stranded break repair through non-HDR methods (NHEJ) (Figure 2.28) representing uncontrolled mutations to maintain cell viability. More detail can be found in **Common Figure Description 1**.

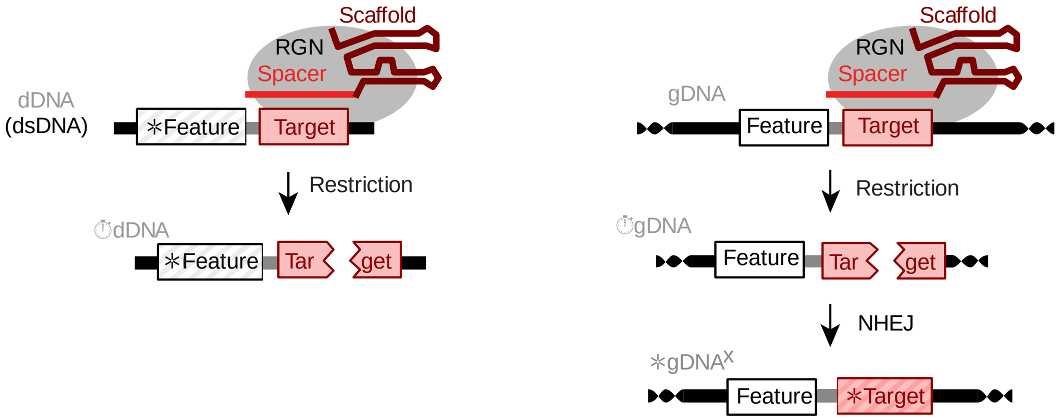


**Figure 2.4 – When double-stranded dDNA contains an identical Target, edits are not possible, and the Target is disrupted**

If the Target sequence is on the donor DNA (dDNA) (left), then it will be cleaved by the gRNA:RGN complex, thereby preventing dDNA incorporation into the genomic DNA (gDNA). Additionally, the Target on the gDNA (right) is disrupted, then repaired erroneously.

Black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by double-stranded break repair through non-homology-directed repair methods (NHEJ) (Figure 2.28) representing uncontrolled mutations to maintain cell viability. More detail can be found in **Common Figure Description 1**.

### 2.3.2 Description of the AddTag (2-step) genome editing method

We present AddTag (Figure 2.5), a powerful 2-step genome editing method that bypasses the targeting constraints that limit typical (1-step) RGN-mediated genome editing approaches and thus enables precision editing of practically any genetic locus, independent of its size (Figure 2.6). In 2-step editing, the genomic region to be edited (Feature) does not need to overlap a unique RGN restriction site (Target); instead, a user can utilize any potential RGN target site that is within the vicinity of the feature to be edited. In Step 1, an RGN is directed to cut at a user-defined target sequence that is near the genomic feature to be edited, and both the Target and the Feature, along with any intervening sequence, are replaced by a unique addtag sequence that contains a new Target. In Step 2, RGN-mediated cutting of the addtag sequence enables the introduction of almost any DNA sequence of choice in place of the genomic region that was originally deleted in the first step. By decoupling the Feature to be edited from the Step 2 Target, and thus removing the need to disrupt the original Step 1 Target, this 2-step methodology virtually eliminates typical RGN targeting constraints and enables genome edits that would otherwise not be possible. This 2-step methodology is uniquely effective at introducing small scale edits into genomic features that cannot be directly targeted by RGNs. In addition, this general strategy also enables a wide range of reverse genetic approaches including the introduction of targeted deletion or substitution mutations, as well as the reintroduction of the gene of interest (i.e., complementation).

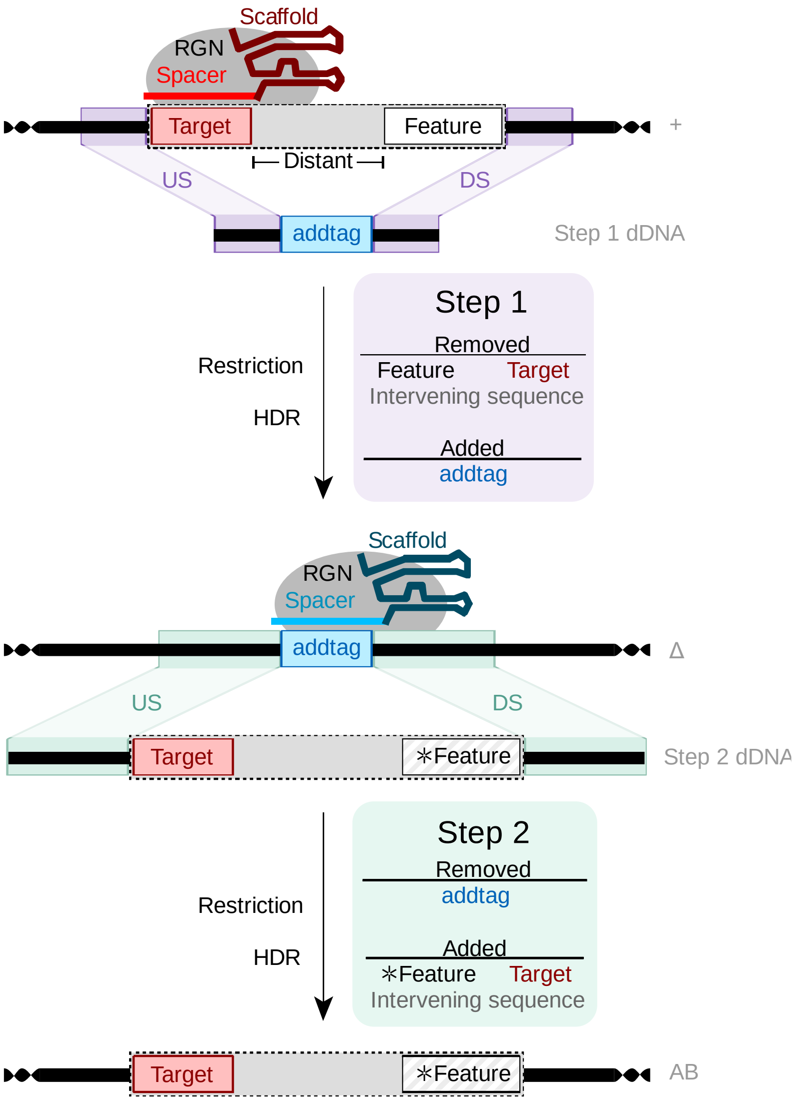


**Figure 2.5 – Indirect (2-step) AddTag method**

The indirect (2-step) AddTag method first changes the reference (+) genome into an intermediate (Δ) genome, and then turns the Δ genome into the add-back (AB) genome. AddTag enables precision feature editing without the need for a proximal target or any modifications outside of the intended ∗feature. Step 1 removes the Feature and Target, along with any intervening sequence, and replaces them with a unique engineered RNA-guided nuclease (RGN) Step 2 Target site. Step 2 uses RGN cutting of the Step 2 Target to enable re-introduction of the previously removed intervening sequence (grey) and target, along with a modified ∗feature (or even the unmodified feature). Since the target sequence is not cut during Step 2, modifications to the target, or any other portion of the previously deleted locus, are not required. Vertical black arrows represent RGN-mediated cutting of the target locus, followed by donor DNA (dDNA) incorporation via HDR. See **Common Figure Description 1** for more details.

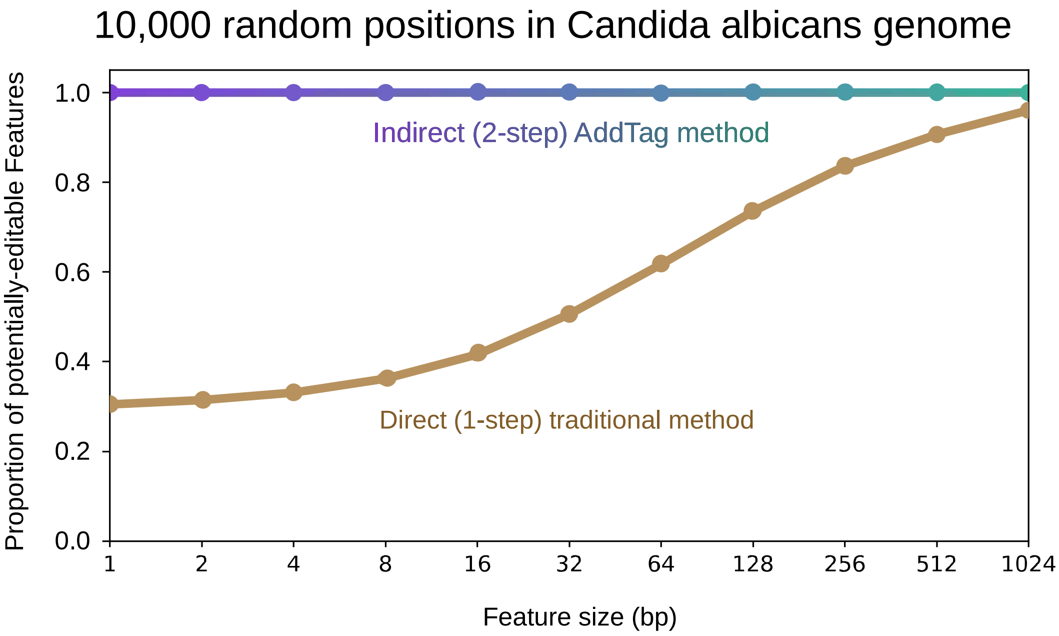


**Figure 2.6 – AddTag method outperforms Typical method in C. *albicans***

The 2-step AddTag methodology enables precision genome editing of genetic loci that would not be possible using traditional (1-step) methods. Each curve shows the proportion of genomic loci (vertical axis) that are potentially editable with either the Direct (1-step) typical method (tan line) or Indirect (2-step) AddTag method (violet-green line), as a of size. 10,000 C. *albicans* genomic loci were uniformly, randomly selected, and the Feature at each locus was varied across 11 sizes, ranging from 1 to 1024 bp (horizontal axis). For Direct editing, sites were considered potentially editable if there was at least 1 bp of overlap between the Feature and a Cas9 target motif ('N{17}|N{3}>NGG'). For Indirect editing, sites were considered potentially editable if a Cas9 target motif was found within a maximum expanded feature size of 4096 bp (3.4.1). For both Direct and Indirect editing, targets were required to pass the default AddTag quality controls (3.3.2): polyT ≤ 4.5, 25 ≤ GC ≤ 75, post-alignment Errors ≤ 5, AZIMUTH on-target ≥ 45, and HSU-ZHANG off-target ≥ 90.

To demonstrate the utility of the AddTag approach, we performed a series of genome edits in the diploid human fungal pathogen *Candida albicans*. First, we show that it is possible to edit small genomic features, such as transcription factor binding sites (bs), that could not be edited using the typical (1-step) approach due to RGN targeting constraints (Figure 2.19, Figure 2.20, Figure 2.21). Second, we demonstrate that the AddTag approach can be used to generate a matrix of isogenic strains to investigate the effects of combinatorial mutations in neighboring genomic features (Figure 2.23). We also highlight the advantage of using the AddTag approach for gene complementation analyses by completely restoring the wild-type phenotype of gene deletion strains where previous approaches had failed to achieve full phenotypic restoration (Figure 2.24).

We also provide a custom software package that automates the extensive manual design work that would otherwise be necessary to implement the AddTag approach to genome editing (Figure 2.7, Figure 2.8, Figure 2.9, Figure 2.10, Figure 2.11). This ADDTAG software introduces new functionality unavailable in competing gRNA design utilities (Table 3.2). ADDTAG not only automates RGN target selection and dDNA design for both steps of the AddTag approach (Chapter 3), it also designs an integrated set of PCR primers for validation of the intended genome edits after each step (Figure 2.11, Figure 2.26, Chapter 4). These integrated experimental and computational tools greatly expand the range and feasibility of precision genome editing applications in any organism that supports *in vivo* genome editing via RGN-mediated HDR. ADDTAG requires the following inputs: a reference genome sequence (+) with the annotated Feature to be edited, a ✳Feature sequence with which to replace the wild-type feature, and one or more RGN Target motifs, representing the specific RGN(s) being used for HDR-mediated genome editing. The following figures illustrate the internal process that the ADDTAG software uses to produce its output.

Features that are difficult to modify with 1-step genome editing either lack a Target or have excess heterozygosity. ADDTAG solves these issues by facilitating removal of more DNA than what is defined by the bounds of the Feature (Figure 2.7). If necessary, ADDTAG will expand the bounds of a user-defined feature to include an RGN target that meets or exceeds gRNA quality control filters. The Feature expansion process also ensures that flanking homology arms manifest an acceptable level of polymorphism (Figure 3.5). The software also automatically generates the dDNA sequences used in Steps 1 and 2, based on specific user inputs and the outcome of the Feature expansion process.

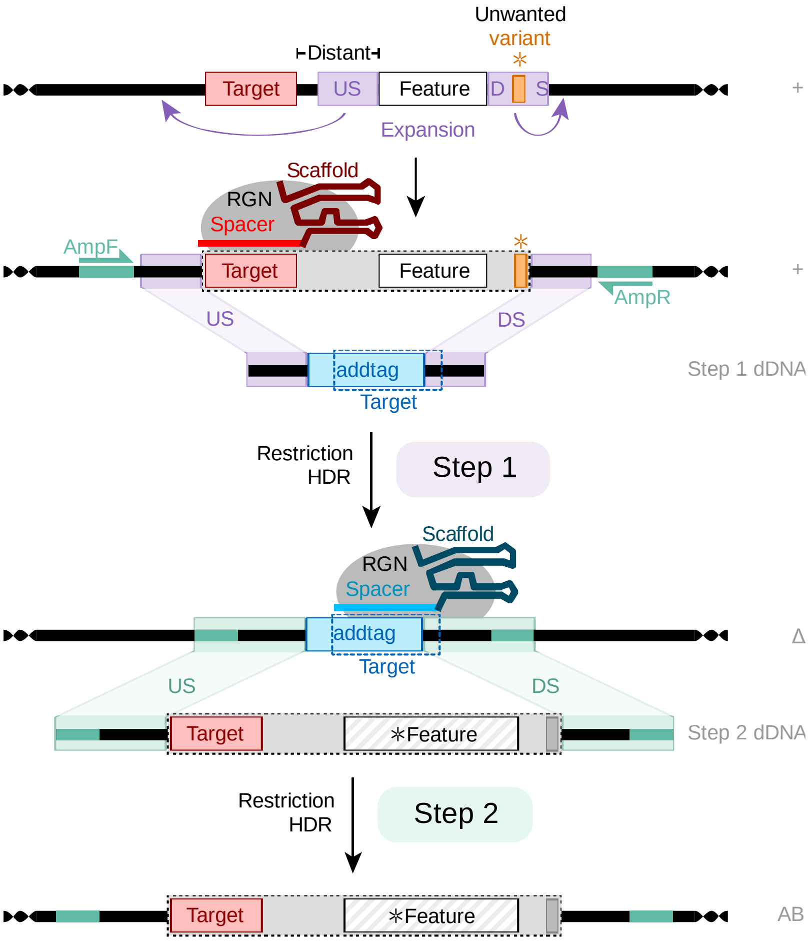


**Figure 2.7 – AddTag (2-step) genome editing with Feature expansion**

ADDTAG automatically identifies the Step 1 Target sequence (red) and designs the Step 2 Target sequence (cyan) used for RGN-mediated cutting of genomic DNA in Steps 1 and 2, respectively. In this example, the Feature is expanded in both directions (violet arrows) such that the Target and variant are enclosed within the bounds of the expanded Feature.

The Step 1 donor DNA (dDNA) consists of upstream and downstream homology sequences (violet US and DS regions) derived from the reference gDNA (+gDNA) sequences that flank the expanded feature, combined with the addtag insert to create a new RGN target (box with dark blue, dashed lines). Step 2 dDNA consists of expanded upstream and downstream homology regions (green US and DS regions) flanking the expanded genomic feature that was removed in Step 1 (grey box with dashed line border). A wild-type version of the Step 2 add-back dDNA (not shown) can be amplified from +gDNA using the output AmpF/AmpR primer pair (green), while Step 2 dDNAs with modified sequences (*Feature) can be generated by stitching PCR or DNA synthesis (not shown).

Violet arrows are for clarity and do not represent biological processes. Vertical black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by homology-directed repair (HDR) as the process by which dDNA is incorporated into the gDNA. See **Common Figure Description 1** for more details.

Next, AddTag systematically searches for candidate primer sequences to populate a standardized set of verification PCR (vPCR) primers that can be used for dDNA amplification and genotype confirmation for each genome editing step (Figure 2.8). Some regions, such as the far upstream and far downstream (blue), are shared among all gDNAs (+, Δ, and AB genotypes). Other regions are genome-specific, like the Step 2 Target (cyan) and expanded feature (grey) regions. The software uses a sliding window approach to identify all potential primers within each region, but for simplicity only a few are depicted for each region.

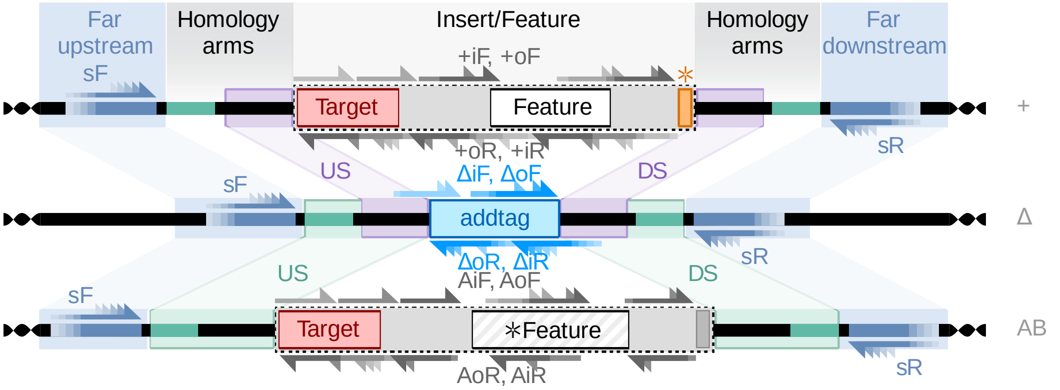


**Figure 2.8 – Identify potential primers within diagnostic regions**

Alignment of the experimental locus from the reference (+), intermediary (Δ), and add-back (AB) genomes. AddTag searches for forward primers in the "Far upstream" and "Insert/Feature" regions and for reverse primers in the "Insert/Feature" and "Far downstream" regions. When the Insert is small, such as the 23 bp addtags used in this study, primer tails are allowed to overlap with adjacent homology arms. More detail can be found in **Common Figure Description 1**.

AddTag then identifies candidate primers that would be suitable for the generation of PCR amplicons (Figure 2.9), indicated by dark grey bars. These include: the sF/sR primer pair, which spans all genomes (+, Δ, and AB gDNAs), and 3 primer pairs which are specific to each genome, labeled sF/oR, oF/sR, and iF/iR. The software assigns a weight to each primer (4.4.4), then it evaluates the compatibility of every pairwise combination of forward and reverse primers for each amplicon and assigns a weight to that primer pair (4.4.5). If the (possibly expanded) Feature or addtag are small, then usable iF/iR primer pairs might not be found within those regions.

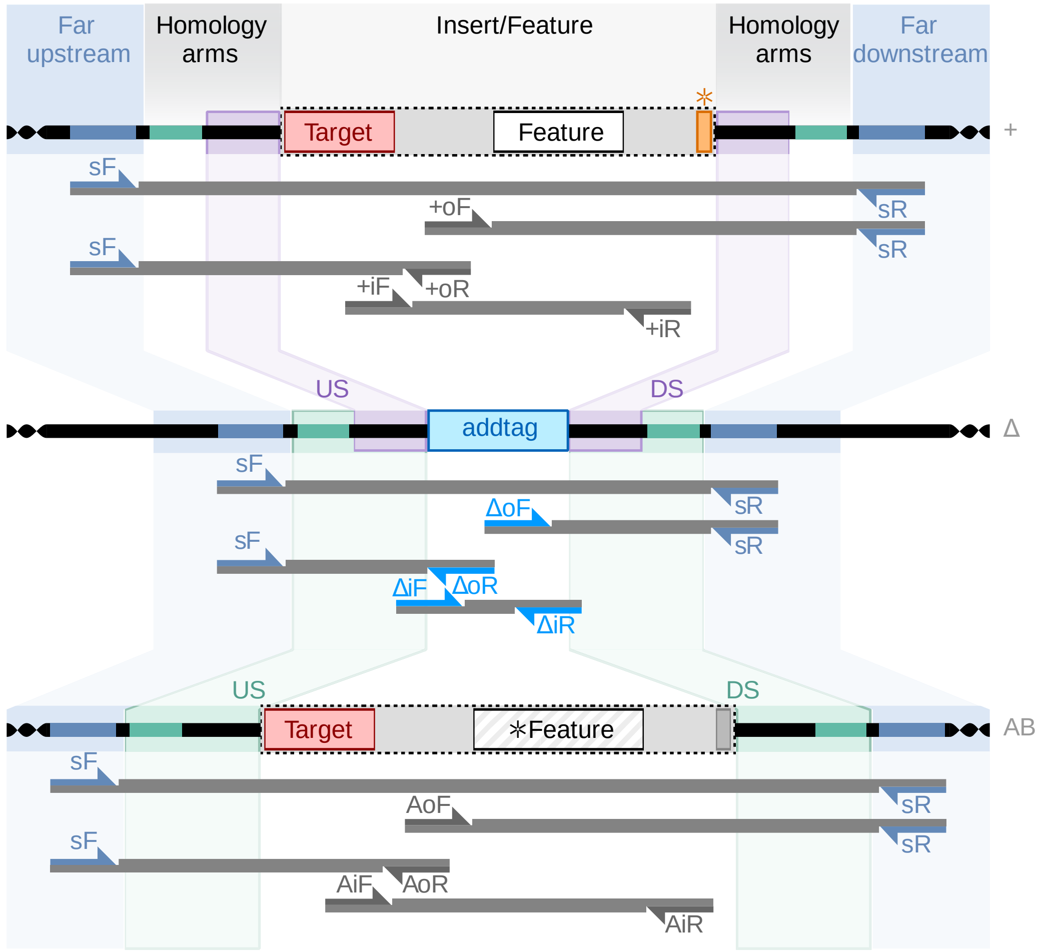


**Figure 2.9 – Evaluate amplicons from different combinations of vPCR primers**

The vPCR primer pairs, and amplicons they form, are shown below the aligned genome from each editing step. In this example, the +oF, +oR, +iF, and +iR primers anneal to identical locations as the AoF, AoR, AiF, and AiR primers, respectively. More detail can be found in **Common Figure Description 1**.

ADDTAG selects an optimal integrated set of PCR primers from the pool of potential primer pairs (Figure 2.10). Each color-coded stack of primers represents an arbitrarily large set of primers identified through the sliding window approach for that like-colored region in Figure 2.8. Simulated annealing identifies the set of primer pairs with highest compatibility (black outline) (4.4.3).

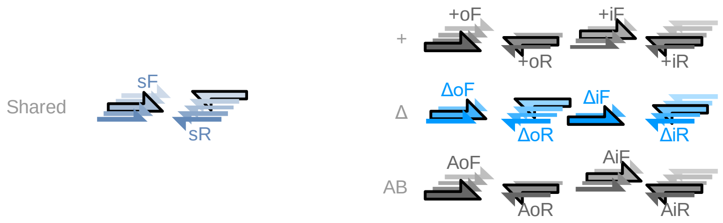
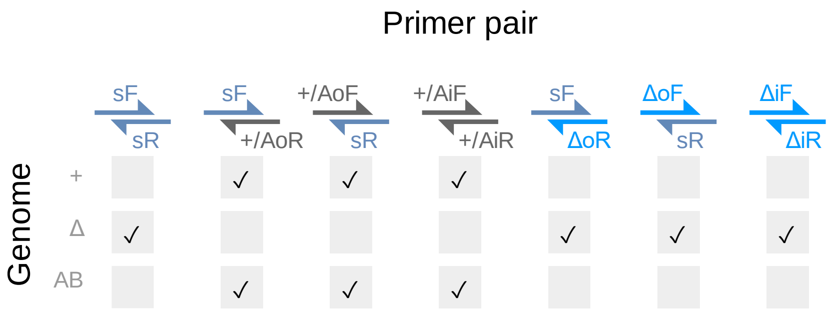


**Figure 2.10 – Pick set of primers with highest compatibility across all amplicons in all genomes**

Primers identified from each stranded region (stacks of half-arrows with identical orientation) at the experimental locus in each genome (Figure 2.8) are subjected to a process of optimization (4.4.3). AddTag predicts the set of primers that are most compatible with each other for use with verification PCR (vPCR). One primer is selected (black outline) from each stranded region (Figure 2.8) for inclusion in the vPCR design.

AddTag calculates expected vPCR amplification for each gDNA using *in silico* PCR (Figure 2.11). Given restrictive vPCR conditions (pictured), the sF/sR pair is expected to amplify only the intermediary, genomic DNA (ΔgDNA), and fail to amplify the reference, genomic DNA (+gDNA) and add-back, genomic DNA (ABgDNA). Alternatively (unpictured), if the feature size is small, amplification should occur at all gDNAs and band migration on a gel should indicate the successful Step 1 dDNA integration. Amplification of the sF/oR and oF/sR pairs across the gDNAs indicates the Feature or insert is present at the expected locus (it is possible dDNA may incorporate at an unintended locus). In this example where none of the optimal primers overlap with the feature or *feature, several primers are identical: +/AoF (+oF and AoF), +/AoR (+oR and AoR), +/AiF(+iF and AiF), and +/AiR (+iR and AiR). The iF/iR pairs amplify if the feature or insert exists anywhere in the gDNA, regardless of its locus.


Primer pair
Genome

| | sF / sR | sF / +/AoR | +/AoF / sR | +/AiF / +/AiR | sF / ΔoR | ΔoF / sR | ΔiF / ΔiR |
|---|---|---|---|---|---|---|---|
| + | | ✓ | ✓ | ✓ | | | |
| Δ | ✓ | | | | ✓ | ✓ | ✓ |
| AB | | ✓ | ✓ | ✓ | | | |

**Figure 2.11 – Calculate expected verification PCR amplification across all genomes**

Each column displays the expected amplification of the indicated verification PCR primer pair (two stacked half-arrows pointing in opposite directions). Each row represents a different PCR template DNA—the reference (+), the intermediate (Δ), and the add-back (AB) genomes. A check (✓) symbol indicates that amplification is expected when using that column's primer pair with that column's genome. Empty boxes mean no amplification is expected with the primer pair for that genome.

To circumvent the above limitations of 1-step CRISPR/Cas-induced HDR, several 2-step methods have been proposed [13] involving a counter-selectable marker at the locus of interest [14], a random transient Target in the intermediary genome [11], the use of transposons [15], single-stranded dDNA [16], and a defined Target in the intermediary genome [17]. AddTag is a generalized approach for developing 2-step CRISPR/Cas genome engineering experiments that expands on these previous approaches. After each step, the genomes are assayed for CRISPR/Cas-induced recombination events with a PCR-based assay [18]. Most software packages that choose gRNA Targets and design verification primers are confined to a single engineering step and genome [19], but ADDTAG primer design can span any number of serial genome editing steps and can apply to any sequenced genome.

### 2.3.3 The AddTag method simplifies and improves gene complementation analyses

Reverse genetics is the process by which known genotypes are assayed for unknown phenotypic differences (Figure 2.12) [20]. Genetic deletion and complementation are fundamental tools for evaluating the relationship between a locus and a phenotype [21]. These methods can be used to determine if alleles are dominant or recessive and ascertain where a gene product lies within a biological pathway. An elemental step in complementation is either adding a Feature into a genome lacking any similar ones, or removing a Feature from a genome, then identifying any differences in phenotype between the genotypes. If a Feature is removed from a genome, thereby breaking a certain biological function, then that Feature is considered necessary for that function. Conversely, if a Feature is added to an organism lacking similar ones, and a new biological function is observed, then that Feature is considered sufficient for that function in that genetic background.

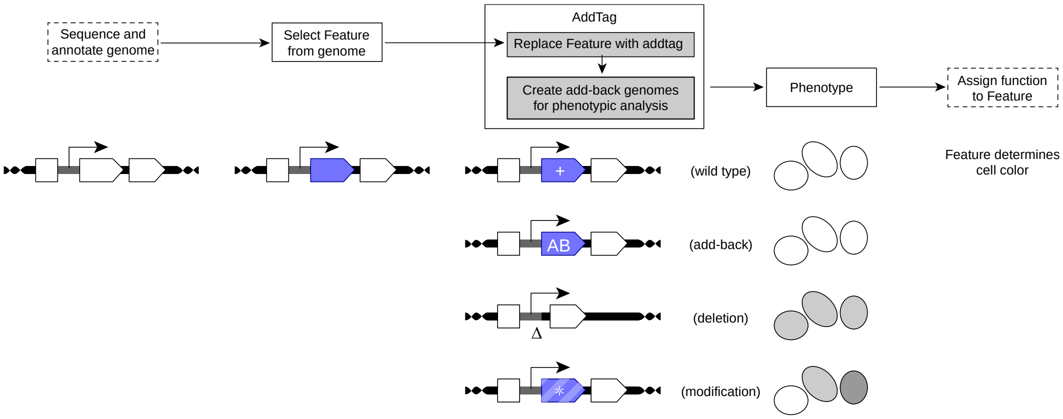

Sequence and annotate genome
Select Feature from genome
AddTag
Replace Feature with addtag
Create add-back genomes for phenotypic analysis
Phenotype
Assign function to Feature
Feature determines cell color
+
(wild type)
AB
(add-back)
Δ
(deletion)
*
(modification)

**Figure 2.12 – Reverse genetics approach using AddTag genome editing**

A general workflow of a reverse genetics experiment, moving from the annotated genome through to the function assignment. Rectangles with dashed borders are processes that are not addressed in this study. Below some processes, an example segment of a genome is given. Arrows bent at right angles indicate transcription at the locus. More detail on the genome fragments below the processes can be found in **Common Figure Description 1**.

A common reverse genetics paradigm is to remove a Feature from a parent strain (+/+), thereby creating a deletion strain (Δ/Δ), then re-introduce the removed Feature back into the deletion strain (AB/AB). Often, this re-introduction happens at a different locus than where the Feature originated—an Auxiliary, non-native locus (Figure 2.15, Figure 2.17).

Because the Native and Auxiliary loci have different genomic contexts, they can be subject to different regulatory processes and result in different measurable phenotypes. For instance, the chromosomal position of *URA3* in the C. *albicans* genome can have dramatic effects on virulence (filamentation phenotype), mRNA expression, and cellular traits (growth, reproduction, survival, and morphology) [22-27]. One of the advantages of the AddTag (2-step) approach is that it facilitates returning wild-type alleles to the Native locus (Figure 2.14, Figure 2.16), thereby avoiding the partial phenotypic restoration of Auxiliary loci. We select well-characterized genes in C. *albicans* that show a pattern of insufficient complementation at Auxiliary loci; and we use the AddTag method to (1) demonstrate the necessity of each gene to confer a phenotype by removing its coding sequence (CDS), and to (2) demonstrate the sufficiency of each gene to confer the phenotype by returning it to its Native locus.

It is possible that any observed phenotype changes could be either due to the intended modification or due to some different, unintended genomic alteration, denoted by an asterisk (*). If a deleted Feature is re-introduced into the genome at the Native locus and the wild-type phenotype is recovered, then this outcome provides evidence that the intended deletion is responsible for the change in phenotype. If the deleted Feature is re-introduced into the genome at an Auxiliary locus and the wild-type phenotype is recovered, then this result offers additional information about cis-regulatory elements and epistasis. However, if the Feature incorporated at the Auxiliary locus fails or only partially-recovers the wild-type phenotype, then it is unclear if the Native locus, the Auxiliary locus, or unobserved genomic changes cause the phenotype.

Gene deletion and complementation are two fundamental techniques in the reverse genetics approach to understanding gene function [20, 21]. However, typical complementation (add-back) approaches rely on gene expression from a non-native locus, often in single copy, and thus are prone to issues with partial complementation or inconclusive results (Figure 2.15, Figure 2.17) [22-27].

To conceptually demonstrate the utility of an add-back at the Native locus compared to an Auxiliary locus, consider a loss of function gene. Each additional locus involved in phenotyping confers a combinatorial expansion for the number of interactions. Therefore, using the Native locus eliminates the potential for interference by cis-acting regulatory elements (CREs), and reduces the potential for interference by unintended mutations. Figure 2.14 and Figure 2.15 depict examples where the phenotype is determined wholly by expression of the gene, denoted as the CDS. Figure 2.16 and Figure 2.17 depict examples where the phenotype is determined by expression of some unknown gene at the other locus. Figure 2.14 and Figure 2.16 show likely results from add-back at the Native locus, while Figure 2.15 and Figure 2.17 show likely results from add-back at an Auxiliary locus. While the AddTag (2-step) approach removes possible interference by CREs, it does not guarantee unintended mutations will not occur. Therefore, experimental design utilizing blocking, repetition, and randomization should be used where possible, and orthogonal experiments that test the link between the gene and phenotype should be performed.

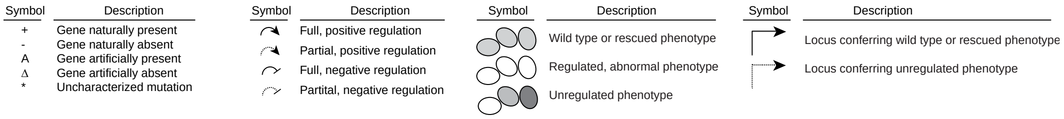
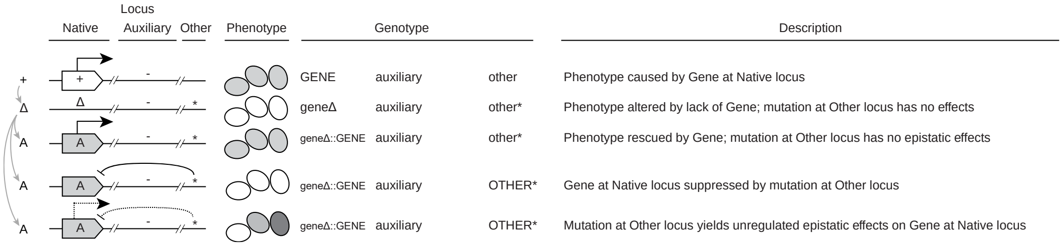

Symbol
Description
+
Gene naturally present
-
Gene naturally absent
A
Gene artificially present
Δ
Gene artificially absent
*
Uncharacterized mutation
Symbol
Description
Full, positive regulation
Partial, positive regulation
Full, negative regulation
Partital, negative regulation
Symbol
Description
Wild type or rescued phenotype
Regulated, abnormal phenotype
Unregulated phenotype
Symbol
Description
Locus conferring wild type or rescued phenotype
Locus conferring unregulated phenotype

**Figure 2.13 – Legend for gene complementation at Native and Auxiliary loci**

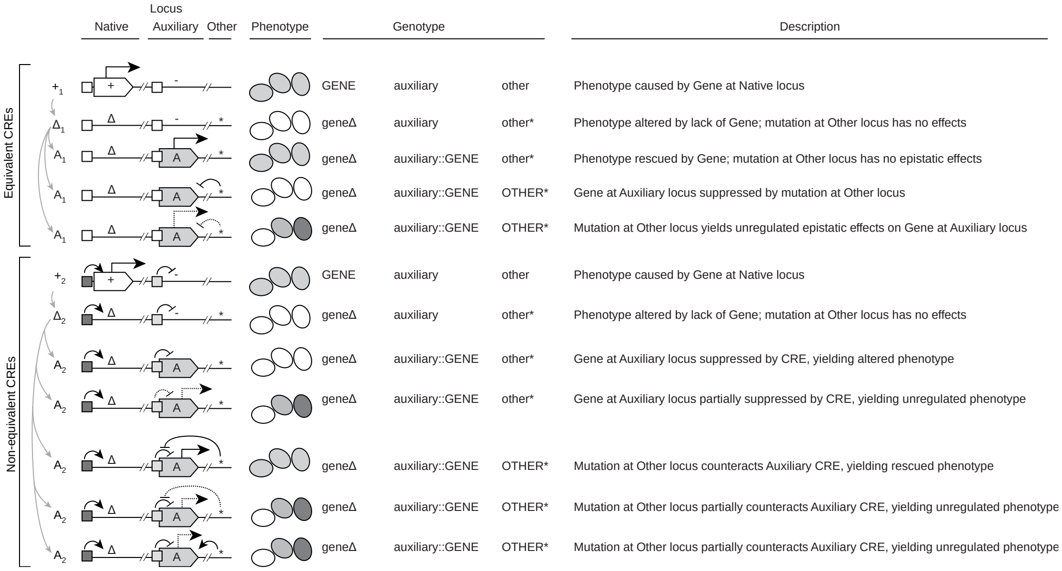
Row labels (+, ∆, A) represent different genotypes, with numeric subscripts denoting alternative genotypes. Grey arrows show the progression from removing a gene from its native locus, then re-introducing that gene to either its Native locus, or an Auxiliary locus. Rows with repeated labels show alternative genic regulations that could be possible given the genotype and phenotype. The genotypes are provided using C. *albicans* notation (Table 0.1). Each genotype is composed of three loci (Native, Auxiliary, Other). If an unintended mutation occurs during a genome editing event, then an asterisk (*) is added to the Other locus. Genes conferring a dominant phenotype are written in upper-case letters, and genes conferring recessive phenotypes are written in lower-case letters.



| | Genotype | | | Description |
|---|---|---|---|---|
| + | GENE | auxiliary | other | Phenotype caused by Gene at Native locus |
| Δ | geneΔ | auxiliary | other* | Phenotype altered by lack of Gene; mutation at Other locus has no effects |
| A | geneΔ::GENE | auxiliary | other* | Phenotype rescued by Gene; mutation at Other locus has no epistatic effects |
| A | geneΔ::GENE | auxiliary | OTHER* | Gene at Native locus suppressed by mutation at Other locus |
| A | geneΔ::GENE | auxiliary | OTHER* | Mutation at Other locus yields unregulated epistatic effects on Gene at Native locus |

**Figure 2.14 – AddTag method (add-back at Native locus); phenotype caused by Gene**

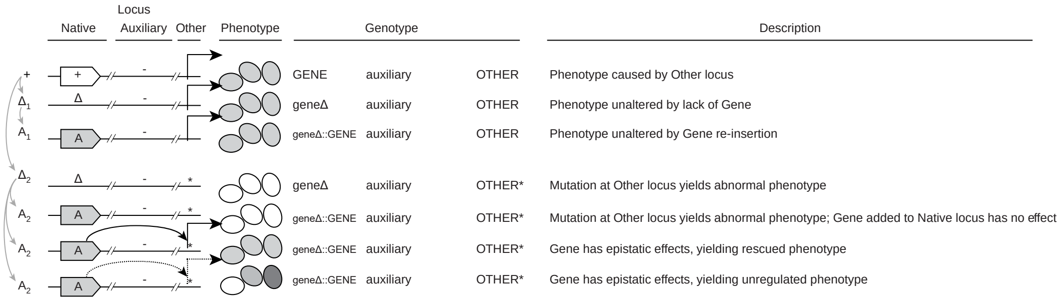
When using the AddTag method to return GENE to its native locus, unintended mutations have only limited possibilities for affecting the phenotype. Either the phenotype is restored, it remains abolished, or it becomes dysregulated. For more details, please refer to Figure 2.13.



| | Genotype | | | Description |
|---|---|---|---|---|
| $+_1$ | GENE | auxiliary | other | Phenotype caused by Gene at Native locus |
| $\Delta_1$ | geneΔ | auxiliary | other* | Phenotype altered by lack of Gene; mutation at Other locus has no effects |
| $A_1$ | geneΔ | auxiliary::GENE | other* | Phenotype rescued by Gene; mutation at Other locus has no epistatic effects |
| $A_1$ | geneΔ | auxiliary::GENE | OTHER* | Gene at Auxiliary locus suppressed by mutation at Other locus |
| $A_1$ | geneΔ | auxiliary::GENE | OTHER* | Mutation at Other locus yields unregulated epistatic effects on Gene at Auxiliary locus |
| $+_2$ | GENE | auxiliary | other | Phenotype caused by Gene at Native locus |
| $\Delta_2$ | geneΔ | auxiliary | other* | Phenotype altered by lack of Gene; mutation at Other locus has no effects |
| $A_2$ | geneΔ | auxiliary::GENE | other* | Gene at Auxiliary locus suppressed by CRE, yielding altered phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | other* | Gene at Auxiliary locus partially suppressed by CRE, yielding unregulated phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Mutation at Other locus counteracts Auxiliary CRE, yielding rescued phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Mutation at Other locus partially counteracts Auxiliary CRE, yielding unregulated phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Mutation at Other locus partially counteracts Auxiliary CRE, yielding unregulated phenotype |

**Figure 2.15 – Traditional method (add-back at Auxiliary locus); phenotype caused by Gene**

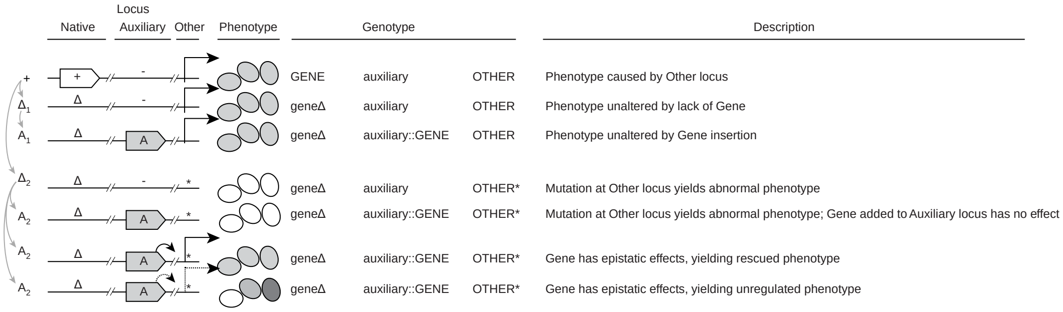
The Traditional add-back method returns the deleted gene to the Auxiliary locus. When the genomic context differs between the Native and Auxiliary loci, there are more regulatory possibilities, and thus more uncertainty in the phenotype. Whenever there are multiple potential explanations for a phenotype, additional experiments are needed to determine the true explanation. CREs are depicted as squares adjacent to the loci they regulate. White squares indicate effectively identical CREs between loci which drive equivalent regulation patterns. Grey squares indicate CREs that affect GENE differently. For brevity, only a limited number of interactions are depicted for the CREs and the unintended mutation (*) at the Other locus. For more details, please refer to Figure 2.13.



| | Genotype | | | Description |
|---|---|---|---|---|
| + | GENE | auxiliary | OTHER | Phenotype caused by Other locus |
| $\Delta_1$ | geneΔ | auxiliary | OTHER | Phenotype unaltered by lack of Gene |
| $A_1$ | geneΔ::GENE | auxiliary | OTHER | Phenotype unaltered by Gene re-insertion |
| $\Delta_2$ | geneΔ | auxiliary | OTHER* | Mutation at Other locus yields abnormal phenotype |
| $A_2$ | geneΔ::GENE | auxiliary | OTHER* | Mutation at Other locus yields abnormal phenotype; Gene added to Native locus has no effect |
| $A_2$ | geneΔ::GENE | auxiliary | OTHER* | Gene has epistatic effects, yielding rescued phenotype |
| $A_2$ | geneΔ::GENE | auxiliary | OTHER* | Gene has epistatic effects, yielding unregulated phenotype |

**Figure 2.16 – AddTag method (add-back at Native locus); phenotype caused by Other locus**

If the locus that is intentionally modified (Native) does not directly control the phenotype, then the deletion and complementation will not affect the phenotype (+, $\Delta_1$, $A_1$). If an unintended mutation affects the causative locus, then there is a chance that the experimental results will resemble Figure 2.14 (+, $\Delta_2$, $A_2$), where the GENE at the Native locus confers the phenotype. For more details, please refer to Figure 2.13.



| | Genotype | | | Description |
|---|---|---|---|---|
| + | GENE | auxiliary | OTHER | Phenotype caused by Other locus |
| $\Delta_1$ | geneΔ | auxiliary | OTHER | Phenotype unaltered by lack of Gene |
| $A_1$ | geneΔ | auxiliary::GENE | OTHER | Phenotype unaltered by Gene insertion |
| $\Delta_2$ | geneΔ | auxiliary | OTHER* | Mutation at Other locus yields abnormal phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Mutation at Other locus yields abnormal phenotype; Gene added to Auxiliary locus has no effect |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Gene has epistatic effects, yielding rescued phenotype |
| $A_2$ | geneΔ | auxiliary::GENE | OTHER* | Gene has epistatic effects, yielding unregulated phenotype |

**Figure 2.17 – Traditional method (add-back at Auxiliary locus); phenotype caused by Other locus**

If the two intentionally-manipulated loci (Native, Auxiliary) do not directly control the phenotype, then deletion and complementation will not affect the phenotype (+, $\Delta_1$, $A_1$). If an unintended mutation affects the causative locus, then there is a chance the experimental results will resemble Figure 2.15 ($+_1$, $\Delta_1$, $A_1$), where GENE at the Auxiliary locus confers the phenotype. For more details, please refer to Figure 2.13.

## 2.4 Results

### 2.4.1 The AddTag method enables precision genome editing of small features that do not overlap RGN target sites

To demonstrate how the 2-step AddTag approach enables precision editing of small genomic features that cannot be targeted directly by typical (1-step) methods, we modified three independent DNA binding sites for the Wor1 transcriptional regulator in C. *albicans*: *$WOR2_{DS}$* (Figure 2.19), *$WOR1_{USd}$* (Figure 2.20), and *$WOR1_{USp}$* (Figure 2.21). C. *albicans* can exhibit multiple morphologies, depending on environmental, genetic, and epigenetic factors [28, 29]. Each morphology is characterized by a different transcriptional regulatory network that defines the profile of gene expression. Two of the morphologies, "white" and "opaque," possess distinct regulatory networks from each other [30]. The *WOR1* and *WOR2* genes are involved in the white/opaque morphological switch.

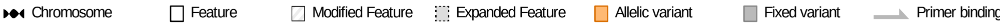


**Figure 2.18 – Common legend for AddTag (2-step) sequencing results**

The AddTag method was used to edit short genomic features, including those that lacked overlapping RGN targets. For Step 1 (violet), the wild-type (+/+) genome was turned into the intermediary (Δ/Δ) genome. For Step 2 (green), the intermediary genome is turned into one or more an add-back (AB/AB) genomes.

Segments of Sanger sequencing chromatogram traces are depicted for the experimental Target and Feature at the edited locus. Grey bars in the traces represent the Phred [31, 32] quality score from 0 (low) to 62 (high).

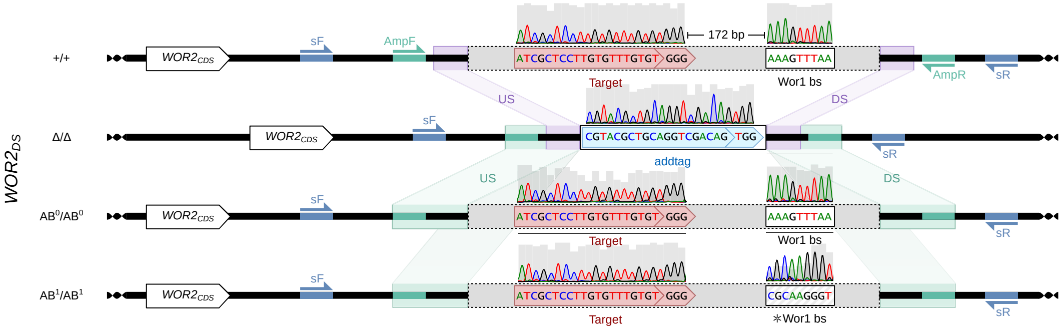


**Figure 2.19 – Sequencing confirms intended, precise editing of *WOR2DS* using AddTag (2-step) method**

A 9 bp Wor1 binding site (Wor1 bs) that is located downstream of the *WOR2* coding sequence (*WOR2CDS*) and lacks an overlapping RGN target site was edited via the AddTag method. In Step 1, both the Wor1 bs and an RGN target 172 bp upstream, along with intervening and flanking sequences included in the expanded feature, were replaced with a Step 2 Target to create the intermediate *wor2DS* Δ/Δ genotype. Two parallel Step 2 transformations converted the intermediary genome into either an add-back genome containing the wild-type Wor1 bs (AB$^0$/AB$^0$) or an add-back genome containing an edited Wor1 bs (AB$^1$/AB$^1$). All sequences outside of the Wor1 bs that were deleted in Step 1 were subsequently restored to their wild-type state in Step 2.

More details can be found in **Common Figure Description 1** and Figure 2.18.

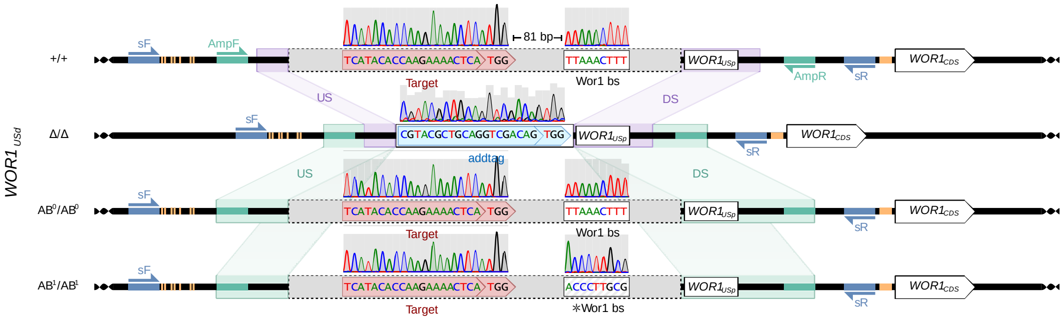


**Figure 2.20 – Sequencing confirms intended, precise editing of *WOR1*$_{USd}$ using AddTag (2-step) method**

A 9 bp Wor1 binding site (Wor1 bs) that is located upstream of the *WOR1* coding sequence (*WOR1*$_{CDS}$) and lacks an overlapping RGN target site was edited via the AddTag method. In Step 1, both the Wor1 bs and an RGN target 81 bp upstream, along with intervening and flanking sequences included in the expanded feature, were replaced with a Step 2 Target to create the intermediate *wor1*$_{USd}$ Δ/Δ genotype. Two parallel Step 2 transformations converted the intermediary genome into either an add-back genome containing the wild-type Wor1 bs (AB$^0$/AB$^0$) or an add-back genome containing an edited Wor1 bs (AB$^1$/AB$^1$). All sequences outside of the Wor1 bs that were deleted in Step 1 were subsequently restored to their wild-type state in Step 2.

More details can be found in **Common Figure Description 1** and Figure 2.18.

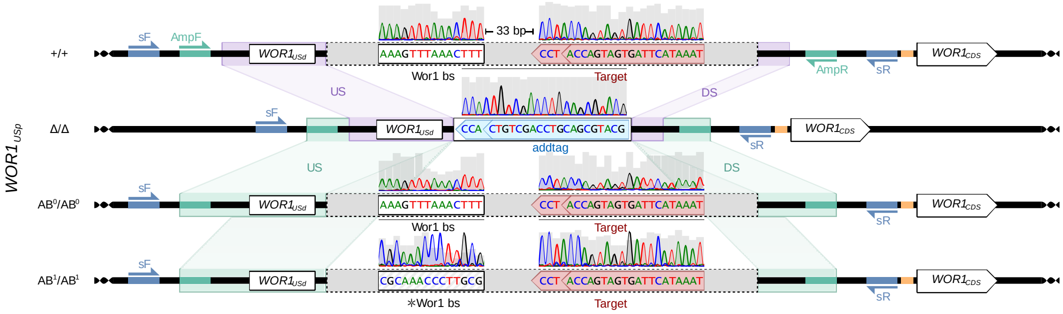


**Figure 2.21 – Sequencing confirms intended, precise editing of *WOR1*$_{USp}$ using AddTag (2-step) method**

A 14 bp Wor1 binding site (Wor1 bs) that is located upstream of the *WOR1* coding sequence (*WOR1*$_{CDS}$) and lacks an overlapping RGN target site was edited via the AddTag method. In Step 1, both the Wor1 bs and an RGN target 33 bp downstream, along with intervening and flanking sequences included in the expanded feature, were replaced with a Step 2 Target to create the intermediate *wor1*$_{USp}$ Δ/Δ genotype. Two parallel Step 2 transformations converted the intermediary genome into either an add-back genome containing the wild-type Wor1 bs (AB$^0$/AB$^0$) or an add-back genome containing an edited Wor1 bs (AB$^1$/AB$^1$). All sequences outside of the Wor1 bs that were deleted in Step 1 were subsequently restored to their wild-type state in Step 2.

More details can be found in **Common Figure Description 1** and Figure 2.18.

Wor1 regulates transcription of the *WOR1* and *WOR2* genes. Wor1 has a 6 bp core TAAACT/AGTTTA consensus binding motif, and is known to bind DNA up to 14 bp in length [33-35]. Since this consensus sequence lacks an NGG PAM sequence, and the binding targets for Wor1 fall within A/T-rich intergenic regions, genomic Wor1-bound sites are either challenging or impossible to edit using typical (1-step) genome editing methods. We selected 3 Wor1 binding sites that lack any significant overlap with potential Cas9 target sites. Only one of these three Wor1 binding sites (*WOR1*$_{USd}$) has any overlap with a potential Cas9 target site. However, the overlap lies 16 base pairs (bp) away from the PAM, and thus would likely require additional genome edits beyond the boundaries of the Wor1 binding site to enable direct (1-step) editing [36, 37]. The targeting constraints we observe with the three selected Wor1 binding sites are representative of all predicted Wor1-bound sites genome wide. Of the 352 predicted Wor1-bound sites, 217 (61.6%) have at least a single base pair of overlap with a potential Cas9 target site (20 bp gRNA target plus the NGG PAM). However, most of these binding sites lack sufficient overlap with the potential Cas9 targets to enable precision editing that bypasses unwanted substitutions outside of the Wor1 binding sites. Upon filtering for sufficient overlap between the Cas9 target site and the Wor1-bound sites, as well as applying gRNA quality control thresholds to maximize on-target cutting and reduce off-target cutting, the number of Wor1-bound sites that could practically be edited by the direct (1-step) method is reduced to 0/352 (0%). This observation highlights the difficulty of performing targeted precision genome editing of this type of small A/T-rich genomic feature using typical (1-step) methods.

We performed AddTag (2-step) genome editing on the three selected loci containing a Wor1 binding site. In each, we first deleted the genomic region including the binding site using a nearby, high-quality, Step 1 Target, and replaced it with an `addtag` that encodes a unique Step 2 Target (CGTACGCTGCAGGTCGACAGTGG) (Table 0.1, Table 0.4). In a subsequent round of transformations, the Step 2 Target sites were cut with Cas9 and the previously deleted regions were restored with either the wild-type genomic sequence (complementation) or a modified version in which the Wor1 consensus binding motif was replaced with a scrambled sequence (ACCCTTGCG/CGCAAGGGT). In all three cases, Sanger sequencing of PCR products spanning the edited loci revealed complete restoration of the wild-type sequence (complementation) or precise editing of the Wor1 binding motif (modification) without any unintended changes to the surrounding genomic DNA that was deleted and subsequently restored. Thus, we successfully demonstrated the ability of the AddTag methodology to precisely edit genomic loci that could not be edited via typical (1-step) methods.

### 2.4.2 The AddTag method can be used to streamline combinatorial editing of neighboring sites

To further exhibit the utility of the AddTag strategy, we performed combinatorial editing of a pair of Zap1 transcription factor binding sites that are separated from each other by 645 bp within the upstream intergenic region of *ZRT2* in C. *albicans* (Figure 2.23). Since these two binding sites are not immediately adjacent to each other, it would be extremely difficult to simultaneously edit both sites using typical (1-step) genome editing methods without also altering the RGN Targets (Figure 2.22).

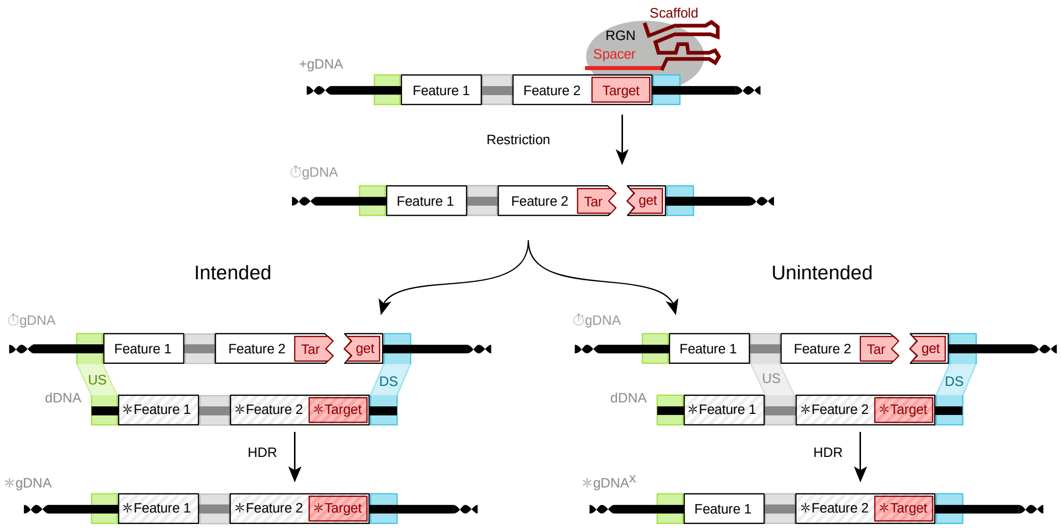


**Figure 2.22 – Excessive homology in long dDNA reduces efficiency of combinatorial, direct (1-step) typical editing**

An example of combinatoric editing under the 1-step method. Here, two Features (Feature 1 and Feature 2) should both be edited. Because they are separated by an intervening sequence (grey) with significant homology to the interior of the dDNA, HDR might not replace Feature 1 with ✻Feature 1 in the modified genome.

Black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by either homology-directed repair (HDR) to incorporate dDNA into the gDNA. More detail can be found in **Common Figure Description 1**.

Zap1 is a well-characterized zinc-finger transcriptional regulator that binds to DNA with the 11 bp consensus motif ACCTTNAAGGT/ACCTTNAAGGT [38-40]. Two instances of this motif are found upstream of the *ZRT2* gene, and those instances have an empirically-significant amount of Zap1 binding [38]. We defined the input Feature for ADDTAG to contain no more than the Zap1 binding site(s) and any intervening sequence. Since the two Zap1 binding sites upstream of *ZRT2* are separated by 645 bp, we deleted a minimal 668 bp region that encompassed both Zap1 binding sites and replaced this region with an addtag containing a Step 2 Target (CGTACGCTGCAGGTCGACAGTGG). The first step of genome editing removed both Zap1 bs and their intervening sequence, creating the deletion strain (*zrt2*$_{US}$ $\Delta/\Delta$). We next designed synthetic dDNA sequences to edit one, the other, or both of the Zap1 binding sites without altering any of the intervening sequence. In separate but parallel operations, we transformed these three independent mutant dDNAs, as well as a wild-type add-back version, into the *zrt2*$_{US}$ $\Delta/\Delta$ base strain. The resulting set of *ZRT2*$_{US}$ add-back strains successfully restored the full wild-type sequence (*ZRT2*$_{US}$ $AB^{00}/AB^{00}$), mutated both the Zap1 binding sites (*ZRT2*$_{US}$ $AB^{11}/AB^{11}$), or individually mutated the CDS-proximal (*ZRT2*$_{US}$ $AB^{01}/AB^{01}$) or CDS-distal (*ZRT2*$_{US}$ $AB^{10}/AB^{10}$) Zap1 binding sites. *ZRT2* encodes a major zinc transporter. Phenotypic assessment of these mutant strains revealed subtle yet consistent alterations in growth between each genotype that suggests that the promoter proximal site is required for Zap1-mediated activation of *ZRT2* on both zinc-sufficient and zinc-deficient media (2.4.4).

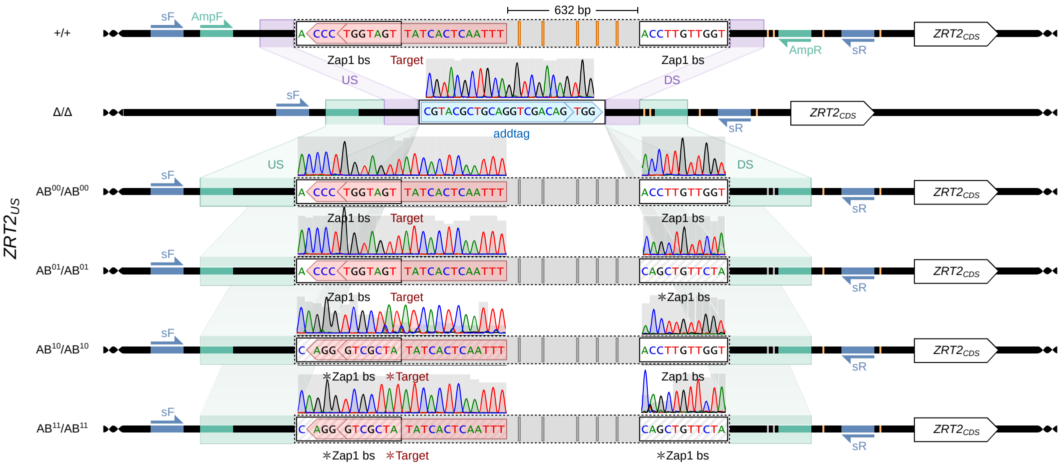


**Figure 2.23 – Sequencing confirms intended, precise editing of *ZRT2*$_{US}$ using AddTag (2-step) method**

AddTag method was used to perform combinatorial editing of two 11 bp Zap1 binding sites (Zap1 bs) that are located 645 bp apart upstream of the *ZRT2* coding sequence (*ZRT2*$_{CDS}$). In Step 1, the two Zap1 bs sequences, along with the intervening sequence, were replaced with an AddTag. Four parallel Step 2 transformations produced add-back genomes with neither ($AB^{00}/AB^{00}$), either ($AB^{01}/AB^{01}$ and $AB^{10}/AB^{10}$), or both (AB11/AB11) Zap1 bs sequences edited. Genomic positions within the feature and homology arms containing heterozygous allelic variants (orange) in the wild-type genomic DNA (+) became fixed in the homozygous state (dark grey) in each add-back genome.

More detail can be found in **Common Figure Description 1** and Figure 2.18.

### 2.4.3 Genetic complementation of *EFG1* and *BRG1* with the AddTag method enables full biofilm phenotype restoration

To highlight the utility of our AddTag approach for gene complementation studies, and to demonstrate the power of creating homozygous gene add-backs at Native loci, we performed gene deletions and add-backs for two key biofilm regulators in C. *albicans*. Biofilm formation is an important virulence trait of C. *albicans* that allows the fungus to successfully colonize host mucosal layers and cause local and disseminated disease in the host [41]. We deleted and subsequently restored the *EFG1* and *BRG1* CDSs that encode master biofilm transcriptional regulators [42]. Under standard biofilm inducing conditions, strains with homozygous deletions (Δ/Δ) of either *EFG1* or *BRG1* have notably impaired biofilm growth [42], while strains that are heterozygous (Δ/+) for either of these genes form biofilms that are intermediary between those produced by +/+ and Δ/Δ strains [43]. Previous studies showed that the traditional add-back approach, using a single-copy of either *EFG1* or *BRG1* integrated at a non-native locus, failed to fully restore the wild-type biofilm phenotype, thus generating partial gene complementation results [42].

Independent *EFG1* and *BRG1* homozygous gene deletion strains (*efg1* Δ/Δ or *brg1* Δ/Δ) were generated by replacing each CDS with unique minimal Step 2 Target sequences—called `mintag`s—which were automatically designed by the AddTag software (Table 0.1, Table 0.4). Homozygous gene complementation strains were subsequently generated using gRNAs that direct Cas9 to cut the Taget encoded by the mintag, along with *EFG1 or BRG1* Step 2 dDNA sequences that were derived from PCR amplification of wild-type genomic DNA.

To assess the phenotypes of the *EFG1* and *BRG1* deletion and add-back strains, relative to their wild-type counterparts, we performed a standard 24-hour biofilm growth assay [44, 45] that assesses the extent of biofilm formation by optical density readouts (Figure 2.24). The $OD_{600}$ growth of *BRG1* and *EFG1* is a highly reliable phenotyping method [44, 45]. The wild-type parental strain formed biofilms with the expected average $OD_{600}$ of 0.69 ± 0.12 (mean ± standard deviation), while the *EFG1* and *BRG1* homozygous deletion strains yielded biofilms with $OD_{600}$ values of 0.17 ± 0.02 and 0.29 ± 0.06, respectively, revealing severely compromised biofilm growth. Upon homozygous add-back of *EFG1* or *BRG1* into the respective deletion strains, robust biofilm growth that is statistically indistinguishable from that of the wild-type parental strain was observed (0.85 ± 0.06 for *EFG1*$_{CDS}$ AB/AB, and 0.83 ± 0.15 for *BRG1*$_{CDS}$ AB/AB). We note that the wild-type phenotype observed with these homozygous gene add-back strains stands in contrast to the previously reported add-back strains (single-copy add-back at a non-native locus), which failed to fully complement the wild-type phenotype [42]. Together, these results demonstrate that the AddTag method can facilitate the generation of a complete set of matched isogenic strains to conclusively assess the phenotypic effects of specific gene deletions, without the ambiguity of partial complementation.

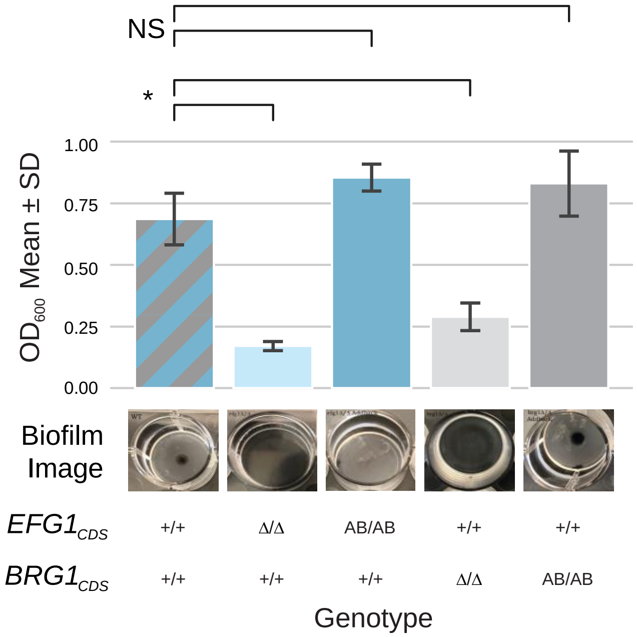


**Figure 2.24 – AddTag-mediated homozygous gene restoration at native loci confers full complementation of the wild-type phenotype**

Strains with *EFG1* or *BRG1* restored at their native loci are indistinguishable from the original wildtype strain background in which the *efg1* or *brg1* deletion strains were engineered. Each column represents a different genotype, with a representative image and the $OD_{600}$ of its biofilm depicted as a bar above. For each genotype, two independently derived strains were cultured in a 24-hour biofilm assay at n = 4 wells. Brackets at the top represent Student t-tests with unequal variance (Welch) where “NS” means p-value > 0.05 and “*”

indicates p-value < 0.05. For both *EFG1* and *BRG1* loci, the Δ/Δ genotype shows a biofilm growth defect, and the AB/AB genotype shows full phenotypic restoration. Bars are colored according to the experimental loci, with *EFG1* in blue, *BRG1* in grey, and both indicated by a bar striped with both blue and grey.

C. *albicans* biofilms develop through a process that goes through four distinct phases [46]. The population of cells in the biofilm is composed of several specific morphologies. Hyphal cells are one of the essential morphologies for complex biofilms. Hyphal formation is an intricate cellular process, involving endocytic and vesicle transport [47]. Hyphal formation is regulated differently at the early and late stages of biofilm development, and it incorporates genes responsible for polarized growth and cell-separation suppression [48]. Hyphal growth in C. *albicans* results from the competitive balance between positive and negative filamentous growth regulator genes [49]. *EFG1* is a key initiator of the hyphal biogenesis pathway. Mechanisms that these growth regulators target include mRNA translational regulation [50], mRNA stability and transport [51], and differential chromatin accessibility [52-54]. Under normoxia, *EFG1* promotes hyphal growth, and under certain conditions such as low-temperature [55] or hypoxia, *EFG1* represses hyphal growth [56-59]. Therefore, *EFG1* behavior differs depending on the growth conditions. On the other hand, *BRG1* expression levels have a dominant influence on whether a cell will filament or not, regardless of the cell's environment [60]. *BRG1* is also important for maintaining hyphal growth after initiation [42].

Expression at the *EFG1* locus depends on CREs—for instance, the *LEU2* locus confers a different expression level than the *EFG1* locus [61]. Therefore, *EFG1* is an appropriate candidate for genetic complementation at the native locus. *EFG1* is a transcription factor that binds to DNA upstream of genes, thereby acting as an activator or a repressor that modulates filamentous growth and biofilm development [42, 62]. *EFG1* influences the expression of hundreds of genes, from cell wall organization [63] and adhesion proteins [64] to other DNA-binding transcription factors known to induce hyphal growth [65].

Like *EFG1*, the *BRG1* gene encodes for a transcription factor protein that helps induce and sustain filamentous growth. *BRG1* is essential for hyphal development in log phase cells (in N-Acetylglucosamine) because deletion of *BRG1* abolishes hyphal growth under these conditions [66, 67]. Strains without a functional *BRG1*$_{CDS}$ have previously been shown to form biofilms, but they have significant reduction in hyphae [42]. Constitutively expressed *BRG1* induces *NRG1* down-regulation and increases hyphal growth. *BRG1* does this by reducing *NRG1* mRNA stability [60]. *HDA1* restricts *NRG1*'s ability to repress downstream hyphae inducing genes [68]. Ectopic *BRG1* expression (under the promotor found at the *MAL2* locus, and not the native promoter) cannot induce hyphal growth, but it can maintain it [68], thus providing only partial phenotypic restoration.

Environmental signals such as temperature, pH, and carbon source all influence C. *albicans* growth morphology. Hyphal growth, a key factor in biofilm formation, can be induced through high temperature, hypoxia, high $CO_2$, starvation, N-acetylglucosamine, serum, and attachment to solid surfaces [69]. Under the growth conditions we used (Chapter 1B.7), both *EFG1*$_{CDS}$ and *BRG1*$_{CDS}$ enhance C. *albicans* biofilm growth.

### 2.4.4 Phenotypic characterization of engineered *ZAP1*$_{US}$ and *ZRT2*$_{US}$ strains

To demonstrate the power of 2-step editing for the introduction of small-scale genome edits, including at sites that do not overlap a PAM sequence, we modified the DNA binding sites for the zinc responsive transcriptional regulator Zap1 upstream of the *ZAP1* and *ZRT2* genes in C. *albicans*. A single Zap1 binding site (bs) exists upstream of the *ZAP1* gene, and two Zap 1 bs exist upstream of *ZRT2* [38]. For both loci (*ZAP1US* and *ZRT2US*, respectively), we directed ADDTAG to use a predefined addtag insert containing a known Step 2 Target within the Step 1 dDNA, and ADDTAG identified a suitable Step 1 Target and vPCR primers (For *ZRT2US*, see 2.4.2). For *ZAP1US*, ADDTAG identified the optimal Step 1 Target as one that directly overlapped the wild-type Zap1 bs ACCTTGGTGGT/ACCACCAAGGT. ADDTAG used Feature expansion to identify an expanded Feature containing both the Zap1 bs and the Step 1 Target. To ensure vPCR success, the expanded Feature extended about 200 bp on either side of the Feature. ADDTAG then safeguarded dDNA integration by ensuring dDNA homology arms contained no allelic variation (Table 0.4). The straightforward genome editing and subsequent output of expected phenotypes is a proof of principle that demonstrates AddTag (2-step) genome editing can enable the introduction of targeted genome edits at sites that do not overlap an RGN Target sequence. While we opted to edit 11 bp cis-acting regulatory elements, one should be able to use this same approach to change as little as a single base pair at nearly all site within the genome.

We modified the DNA upstream of the *ZAP1* CDS using the AddTag 2-step method, then we investigated the phenotypic effects of the Zap1 binding site mutants by assessing colony growth on zinc-sufficient and zinc-deficient media (Figure 2.25). Under zinc-deficient conditions, the *ZAP1* upstream deletion strain (*zap1US* $\Delta/\Delta$) showed reduced growth compared to the wild-type reference (*ZAP1US* +/+), while there was no difference between these strains under zinc-sufficient conditions, indicating that the upstream region is required for robust growth on zinc-deficient media. Reintegrating the wild-type *ZAP1* upstream region (*ZAP1US* AB$^{0}$/AB$^{0}$) restored wild-type levels of growth, while introduction of the *ZAP1* upstream region containing a mutated Zap1 binding site (*ZAP1US* AB$^{1}$/AB$^{1}$) resulted in an intermediate growth defect. The *zap1US* $\Delta/\Delta$ strain phenocopies a strain that is lacking the entire *ZAP1* CDS (*zap1CDS* $\Delta/\Delta$), indicating that deletion of the upstream regulatory region prevents *ZAP1* expression. Together, these results indicate that the Zap1 bs upstream of *ZAP1* does play a role in the autoregulatory induction of *ZAP1* under zinc-deficient conditions, and other elements within the *ZAP1* upstream region support at least a basal level of *ZAP1* expression under zinc-deficient conditions.

Like the *ZAP1* strains, we assessed growth of the modified *ZRT2* strains on zinc-sufficient and zinc-deficient media (Figure 2.25). We observed subtle yet consistent differences in growth between each genotype that were comparable under both growth conditions, implying that changes in environmental zinc levels might not influence Zap1-dependent regulation of *ZRT2*. Under both growth conditions, the *ZRT2US* AB$^{00}$/AB$^{00}$ and *ZRT2US* AB$^{10}$/AB$^{10}$ strains exhibited growth that was indistinguishable from the wild-type *ZRT2*US +/+ strain, while the *zrt2US* $\Delta/\Delta$, *ZRT2US* AB$^{01}$/AB$^{01}$, and *ZRT2US* AB$^{11}$/AB$^{11}$ strains showed slightly reduced growth relative to the wild-type reference strain. This suggests that the CDS-proximal Zap1 binding site upstream of *ZRT2* is predominantly responsible for Zap1-mediated induction of *ZRT2*, and that the CDS-distal Zap1 binding site may not play a significant role under the growth conditions tested. Interestingly, disruption of the CDS-proximal Zap1 binding site upstream of *ZRT2* resulted in a similar growth defect as complete deletion of the entire region that encompasses both Zap1 binding sites, again highlighting the importance of the CDS-proximal Zap1 binding site. We note that although the observed growth defects are relatively subtle, this is not entirely unexpected. C. *albicans* has several genes that work in concert to maintain homeostasis of zinc levels. One explanation for why *ZRT2*$_{US}$ +/+, *ZRT2*$_{US}$ AB$^{00}$/AB$^{00}$, and *ZRT2*$_{US}$ AB$^{10}$/AB$^{10}$ show similar growth phenotypes is that other zinc transporter genes, such as *ZRT1*, are compensating for any deficiencies to *ZRT2* [70].

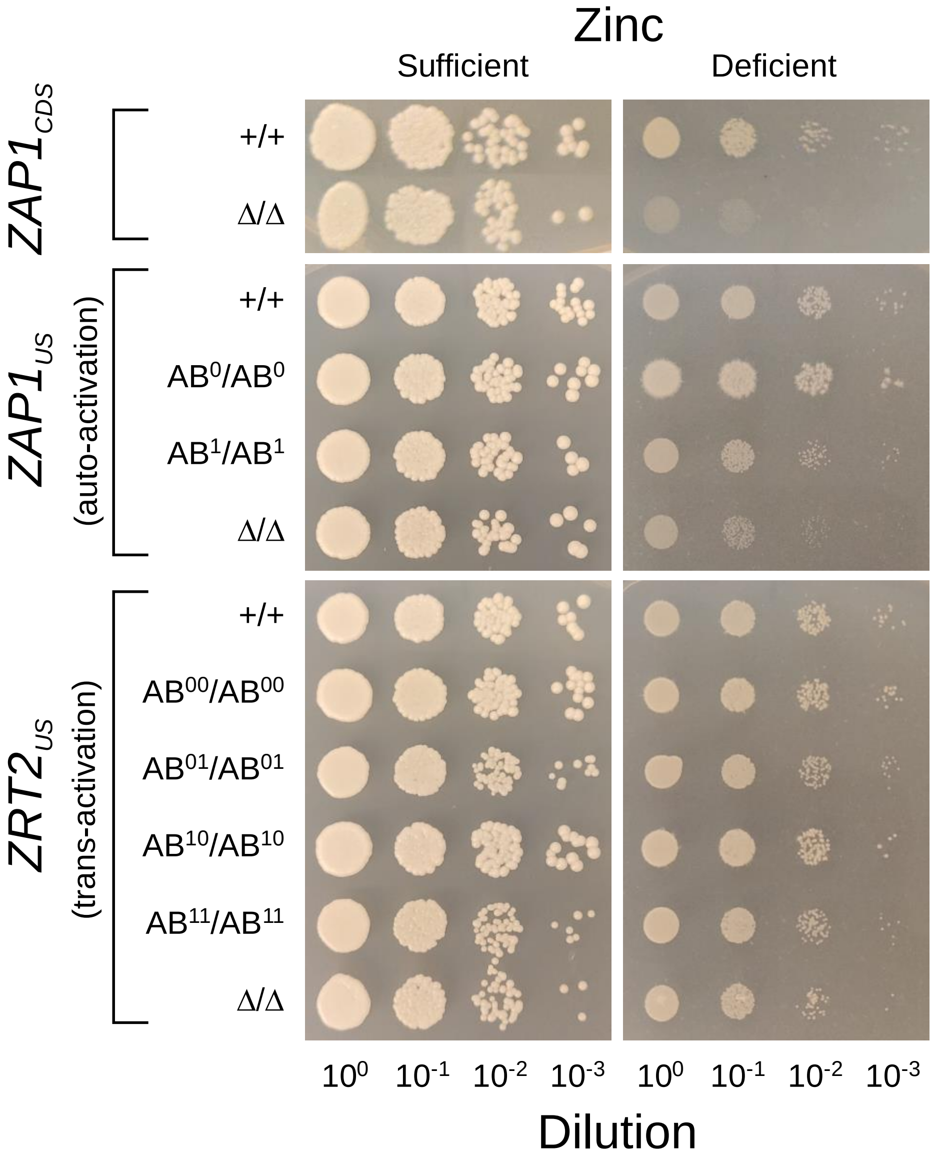


**Figure 2.25 – Zap1 binding site add-backs reveal importance of cis-regulatory elements for *ZAP1* and *ZRT2***

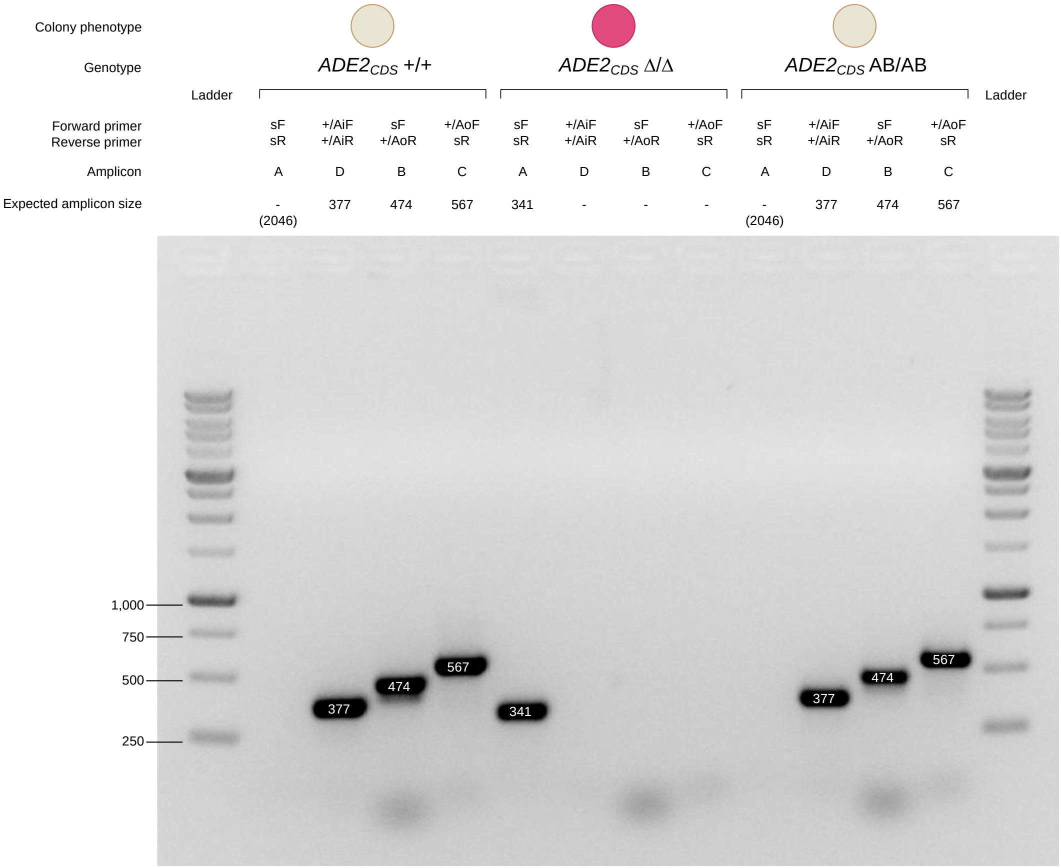
Cells from each genotype were cultured under zinc-sufficient and zinc-deficient conditions. 2 independent biological derivations for each genotype were plated twice, and representative plate images at 48 hours were selected for depiction. Each row is a different genotype. Each spot originated from 5 µL of culture. Spots on the left-most column of each condition came from cultures with OD$_{600}$ of 0.3, which were serially diluted by a factor of $10^{-1}$ in each successive column to the right. Spots were digitally aligned to a grid undistorted, maintaining original sizes and average spacings.

### 2.4.5 *ADE2$_{CDS}$* is required for purine biosynthesis

When performing experiments, it is important to have a positive control that can be used to diagnose unexpected results. If the RGN is somehow deficient, it may fail to cut the Target. We address this possibility by engineering the *ADE2* locus. The *ADE2$_{CDS}$* gene has been well-studied in several yeast systems [71-73] and is a staple control for yeast genetic engineering experiments [74-78]. *ADE2* is commonly used as a genetic marker for spontaneous and directed recombination. A benefit of using the *ADE2* locus as a control is that C. *albicans* colonies have a visually-discernable phenotype, depending on the functionality of the locus.

In C. *albicans*, *ADE2$_{CDS}$* encodes for a 568 amino acid phosphoribosylaminoimidazole carboxylase protein that turns aminoimidazole ribotide (AIR) into 1-(5-Phospho-D-ribosyl)-5-amino-4-imidazolecarboxylate (CAIR). *ADE2* genotypes are assayed on adenine-deficient media. If a yeast has at least one functional copy of *ADE2$_{CDS}$*, it will process AIR into CAIR, which is required for purine biosynthesis. If the yeast has no functional copies of *ADE2$_{CDS}$*, then it will fail to convert AIR into the necessary purine precursors. Instead, AIR is conjugated with a targeting compound that allows for shuttling into the central vacuole [71]. There, the AIR is separated from the targeting compound [79-81], and AIR derivatives polymerize into the red pigment [82, 83]. The accumulation of these AIR-derived compounds makes the colony appear reddish on adenine-deficient media. Besides color, AIR-derived compounds display cytotoxicity [83, 84] and can alter DNA cleavage [85].

The ADDTAG software predicts the expected amplicon size for each vPCR primer pair. As an example of how the ADDTAG predictions compare to reality, we executed *ADE2$_{CDS}$* vPCR in parallel for all primer pairs across the +/+, Δ/Δ, and AB/AB genotypes (Figure 2.26). We used agarose gel electrophoresis to visualize the vPCR products. On the gel, the amplicon band migration for each primer pair matched ADDTAG's determination for successful editing. We performed a 2-step genome editing procedure to delete, and subsequently restore, the CDS of *ADE2*. C. *albicans* strains that lack Ade2 enzyme activity accumulate AIR, which is ultimately converted into a red pigment [82, 83]. *ADE2* genotypes are easily differentiable by assessing colony colors—with mutant and wild-type colonies showing red and white colors, respectively. In the first transformation, *ADE2$_{CDS}$* was replaced by a `mintag(CC)` sequence, yielding *ade2$_{CDS}$* Δ/Δ with the expected pattern of vPCR verification bands and a red colony phenotype. Upon reintegration of the *ADE2* CDS at the native locus, using a gRNA that targets the `mintag` generated in Step 1, the expected pattern of vPCR amplicon bands was observed and the white colony phenotype was fully restored. This experiment serves to validate the efficacy of the ADDTAG-designed gRNAs, dDNAs, and vPCR primers, and it demonstrates a proof of principle for the homozygous add-back approach. However, *ADE2* add-back even at a non-native locus has previously been shown to restore the white colony phenotype, so this example does not fully illustrate the advantage of homozygous gene add-back at the native locus.



## Figure 2.26 – vPCR amplification shows *ADE2*$_{CDS}$ was edited as intended

*Candida albicans ADE2*$_{CDS}$ +/+ was replaced with a `mintag(CC)` to create the Δ/Δ genotype. The `mintag(CC)` introduced the PAM site for the intermediary Step 2 Target, which enabled editing a second time to create the AB/AB genotype when inserting the +gDNA-amplified *ADE2*$_{CDS}$ Feature into the native locus. For this locus, ADDTAG returned an optimal primer design lacking ΔiF, ΔiR, ΔoF, or ΔoR primers. Genomic template from colony lysates were added directly to the PCR mix. PCR was conducted for 30 amplification cycles. Hyphens (-) indicate no expected amplification. The sF/sR primer pairs did not amplify in the +/+ and AB/AB genotypes because the extension step was not a sufficient time duration.

vPCR reaction products were subjected to agarose gel electrophoresis. At the top of the gel image are a row of wells that initially contained the input vPCR reaction products. After applying a charge differential across the gel, DNA migrated from the well down toward the bottom of the image. Low molecular weight DNA migrated faster, and is farther from the wells at the top, and high molecular weight DNA migrated slower, and is closer to the wells at the top. Along the left and right side of the image are bands representing DNA of known size and molecular weight. The size, in bp, are written on experimental bands.

### 2.4.6 The ADDTAG software automates experimental designs for 2-step genome editing

To facilitate implementation of this genome editing methodology, we developed the ADDTAG software package which automates numerous critical experimental considerations that are necessary for successful 2-step genome editing (Chapter 3) and genotype verification (Chapter 4). Users are required to make only broad decisions to provide the framework by which the program automates the experimental design process, thus decreasing the trial and error associated with gRNA target identification, dDNA design, and PCR primer selection. The software automatically identifies high quality RGN targets in the vicinity of the genomic feature to be edited (Figure 3.12), expands the selected feature to encompass an optimal RGN target site (Figure 3.2, Figure 3.3, Figure 3.4, Figure 3.6), and designs dDNA fragments for both the first (Figure 3.13) and second steps of editing. Furthermore, ADDTAG automatically generates an integrated minimal set of PCR primers that enable unambiguous genotypic verification at each step of the genome editing process (Figure 2.8, Figure 2.9, Figure 2.10, Figure 4.6). All genome edits described in this study were successfully performed using ADDTAG-generated gRNA, dDNA, and PCR primer sequences, thus validating the utility of this automated design software. For a representative example of the PCR-based genotype verification assay, see Figure 2.26.

## 2.5 Discussion

### 2.5.1 Summary

The novel 2-step genome editing methodology presented here is highly flexible and enables a wide-range of genome editing applications that would be difficult, if not impossible, to accomplish using traditional approaches. Perhaps the most striking capability of this methodology is the introduction of small-scale precision genome edits at loci that cannot be directly targeted by Cas9 (or your RGN of choice). Using the human fungal pathogen C. *albicans* as a test case, we demonstrate this capability by introducing targeted substitutions within three independent 9 bp transcription factor binding sites, all of which lack the necessary overlap with potential Cas9 target sites for typical (1-step) editing. We also demonstrate that the AddTag approach enables facile combinatorial editing of loci that are proximal to each other in the genome, without the need to modify the intervening sequence. Furthermore, we restored previously deleted genes to their native loci, demonstrating how the AddTag approach facilitates improved gene complementation analyses without the need for molecular cloning. This native locus gene add-back approach resulted in complete restoration of the wild-type phenotype for the two gene deletion strains tested, whereas traditional methods had previously failed to achieve full phenotypic complementation for the same genes, further highlighting the advantages of the AddTag-mediated complementation method. While these examples are by no means exhaustive representations of the potential applications supported by the AddTag approach (other applications include, but are not limited to, the construction of translational fusions, construction of large mutant libraries, and modification of the lengths of repetitive elements), they highlight how this 2-step process facilitates a wide-range of reverse genetic experiments by enabling seamless and efficient deletion and subsequent complementation or modification of virtually any locus in an organism that supports efficient RGN-mediated genome editing.

Although we opted to edit 9 bp cis-regulatory motifs as a proof of concept for introducing small-scale targeted edits at sites that do not overlap RGN target sites, it should be possible to use the AddTag approach to change as little as a single base pair at virtually any site within the genome. While this methodology is highly flexible and facilitates otherwise challenging or impossible precision genome edits, there are some caveats and limitations that are worth considering when implementing this strategy.

### 2.5.2 Loss of heterozygosity

A caveat of the AddTag approach is that heterozygosity within the region being edited (including any sequences that lie between the feature and target loci) can be lost. Indeed, we observed loss of heterozygosity within the 645 bp region between the two Zap1 binding sites during our combinatorial editing of *ZRT2*$_{US}$. However, the potential effects of this loss of heterozygosity can be controlled for by performing Sanger sequencing of the affected region and selecting a matched set of wild-type add-back and mutant strains which are homozygous for the same allelic variant.

Often, loss of heterozygosity (LOH)—a term that to encompass various molecular processes including gene conversion [86] and specific instances of mitotic recombination [87]—can convert a heterozygous locus into a homozygous one. LOH demonstrates that a single dDNA integration is often all that is necessary to produce a homozygous edit. Enzymatic kinetics supports that only a minority of gDNA RGN associations result in restriction, and furthermore, only a minority of dDNA-gDNA interactions result in integration events. This phenomenon means that unless there are significant differences in either the specificity of allele cutting, or in the specificity of dDNA homology regions, if alleles are similar enough, then gene conversion will occur (Figure 2.27). LOH is a common phenomenon in fungal species, including the C. *albicans* model we use. One paramount implication of LOH is that if a researcher wishes to target only a specific allele for editing, it may be difficult if there are insufficient polymorphisms nearby. ADDTAG does allow the user to preferentially design against polymorphic sites, but it does not estimate LOH frequency.

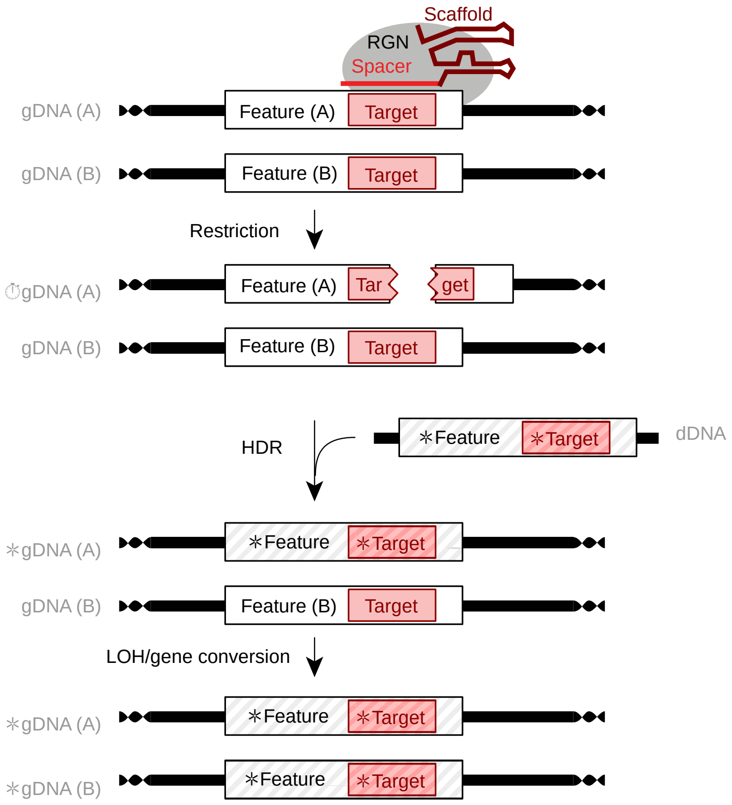


**Figure 2.27 – Loss of heterozygosity is possible using CRISPR/Cas-induced HDR**

One allele (A) of an arbitrary locus undergoes CRISPR/Cas-induced HDR. Without RGN restriction of the other allele (B), the modified Feature and Target (✻Feature, ✻Target) are copied to the locus. More detail can be found in **Common Figure Description 1**.

### 2.5.3 Reducing off-target effects

The AddTag method can utilize any number of serial genome edits. In this manuscript, we specifically describe editing in 2 steps. Each step of editing risks introducing unintended (off-target) mutations into the experimental organism. However, if one is using the addtag sequence as the launching point for subsequent edits, then each subsequent strain created will be subject to the same laboratory manipulations. Regardless of the modifications introduced at the loci of your choice, each strain is subject to the same selective pressures (the selection is only for the genes encoding for the gRNA and Cas9 enzyme), with specific differences due to only the dDNAs used. Thus, mutations that happen in some Step 2 reactions but not others are the ones to be wary of. Using this design (like what was done with the *ZRT2*$_{US}$ *and ZAP1*$_{US}$ loci), fixation of unintended mutations is equal to a typical, 1-step genome edit.

CRISPR/Cas technologies demonstrate variable levels of unintended mutagenicity driven by off-target cutting, which is both genome- and cell type-dependant [88-90]. The AddTag approach requires cutting at two distinct RGN target sites during two sequential steps of genome editing, whereas the typical approach requires only one round of cutting and HDR repair. While the extra round of cutting and repair increases the opportunity for unintended off-target cutting, we note that the AddTag approach and accompanying software enables the use of highly stringent gRNA selection criteria, which should significantly mitigate this risk. In contrast, when using the typical (1-step) approach, particularly in the context of small-scale edits, one can often be faced with the decision of whether to proceed with a poor-quality gRNA or forego the desired experiment altogether. We also note that the overall genome editing strategy implemented in the AddTag approach controls for the potential effects of any off-target RGN cutting; since homozygous, wild-type, add-back strains can be generated in parallel with the desired modified strains, using the same pair of RGN target sites and the same base strain, any phenotypic effects of unintended off-target cutting should be apparent in both the mutant and add-back strains when compared against the original wild-type strain.

There are several types of errors inherent in CRISPR/Cas genome engineering experiments that produce undesirable outcomes [88, 91-93]. One possible off-target effect of AddTag genome editing occurs when the RGN restricts an unintended location in the genome. Several engineered RGNs have been developed to address this phenomenon. Other than the use of a sole Cas9, there are alternative methods that yield lower off-target rates. One is a mutated Cas9 that cleaves only one of the two strands (Cas9n), and is thus a nickase [94]. Two different gRNA, each directing the Cas9n to different strands of DNA, thus together creating a DSB. This technique has shown increased Target specificity at the cost of less total DSB [95, 96].

### 2.5.4 AddTag (2-step) editing efficiency

For serial genome engineering protocols, like the targeted knock-out then knock-in performed in this study, final engineering efficiency is the product of each step's efficiency. We observe editing efficiency of each Step closely parallels the on-target score, and efficiency is equivalent to what was previously reported [74]. The ADDTAG software uses gDNA masking to addresses factors that complicate efficiency, such as RGNs failing to cut the Target due to differential genome accessibility [97].

Sometimes genome editing experiments do not work. Here are three things to do if dDNA does not integrate properly into the genome. (1) Obtain fresh reagents, and then try again. Often, large differences in editing efficiencies are attributable to different buffer batches. (2) As a control, design a gRNA to target a site far away from where your dDNA is intended to incorporate. Add a selectable marker to the insert of your dDNA. Create transformants, then phenotype them. If the dDNA phenotype is expressed, then this gives the rate at which off-target integrations happens (i.e. spontaneously). (3) Alternatively, you can directly measure the spontaneous rate of dDNA integration. Perform experiment sans gRNA and Cas, and observe incidence of dDNA integration due to naturally-occurring genomic double-stranded breaks.

### 2.5.5 Applicability of AddTag (2-step) editing to other organisms

The practical application of CRIPSR/Cas genome editing described in this study is predicated on stable expression of the gRNA and the RGN. As in the C. *albicans* CRISPR/Cas framework [74], this requires stable integration into the host genome. This might not be applicable to all practices, as constitutive expression of Cas genes can cause uncontrolled phenotypic changes [98]. The C. *albicans* system circumvents this by leveraging spontaneous excision of the transgenic gRNA and RGN genes [74]. Many CRISPR/Cas systems, such as those used in other organisms, rely on transient RGN and gRNA expression [77, 99], or a transfected RGN and gRNA [100], which can provide the same utility but with attenuated efficiency.

AddTag was designed primarily for use in cultivable organisms with a reference genome, short generation times, and high levels of HDR for repairing DSB. HDR is known to vary by species [101] or even cell type [102], and it plays a large role in the overall success of these types of CRISPR/Cas-induced genome editing experiments. Some molecules have been shown to increase activity/efficiency of the RGN complex, such as dimethylsulfoxide (DMSO) [103]. Certain small molecules have been found to shift the ratio of HDR/non-HDR repair, such as Caffeine [104, 105], NU7026 [106], Azidothymidine [107], Trifluridine [107], and others [108, 109].

Several classes of mechanisms have been described to repair CRISPR/Cas-induced double-stranded breaks in chromosomes. Besides HDR, non-homologous end joining (NHEJ) and alternative end joining (AEJ) are methods cells use to repair their genomic DNA. Figure 2.28, Figure 2.29, Figure 2.30, and Figure 2.31 depict the first step of genome editing using the AddTag method, where the Feature and Step 1 Target are replaced with the `addtag`, but each depicts the DNA repaired through a different process. Figure 2.32 summarizes the expected vPCR amplification of the resulting ΔgDNA.

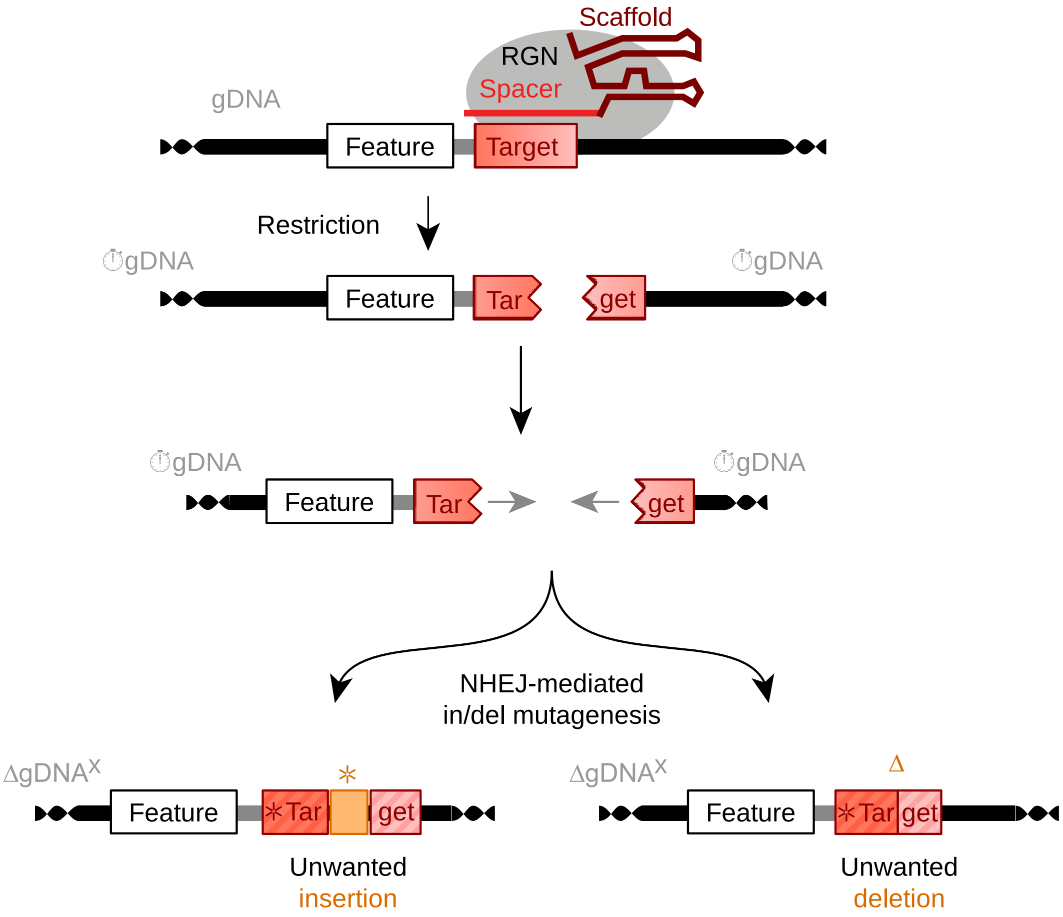


**Figure 2.28 – Double-stranded break repair through NHEJ-mediated in/del mutagenesis during AddTag Step 1**

Genomic double-stranded breaks may be repaired through non-homologous end joining (NHEJ). This figure depicts the genomic DNA repairing itself through NHEJ, thereby preventing the AddTag Step 1 dDNA from integrating. The NHEJ process commonly introduces insertion and deletion mutations. NHEJ-mediated non-insertion is observable when no dDNA sequence incorporates into the cut locus. More detail can be found in **Common Figure Description 1**.

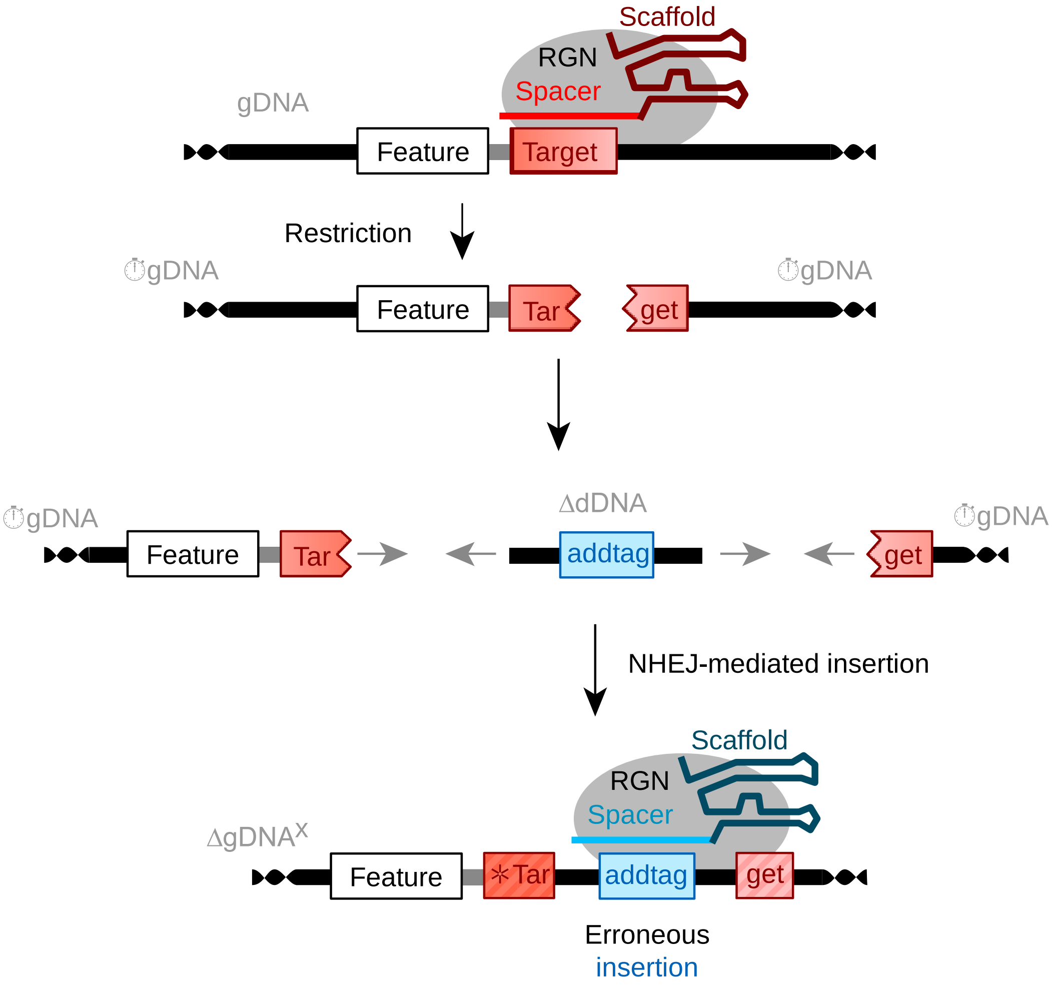


**Figure 2.29 – Double-stranded break repair through NHEJ-mediated insertion during AddTag Step 1**

The gRNA:RGN restricts the genomic DNA (gDNA), and then the gDNA is repaired by erroneously incorporating the dDNA at the site of the restriction. Non-homologous end joining (NHEJ)-mediated insertion is detectable by the presence of the dDNA homology arms inserted at the cut locus. More detail can be found in **Common Figure Description 1**.

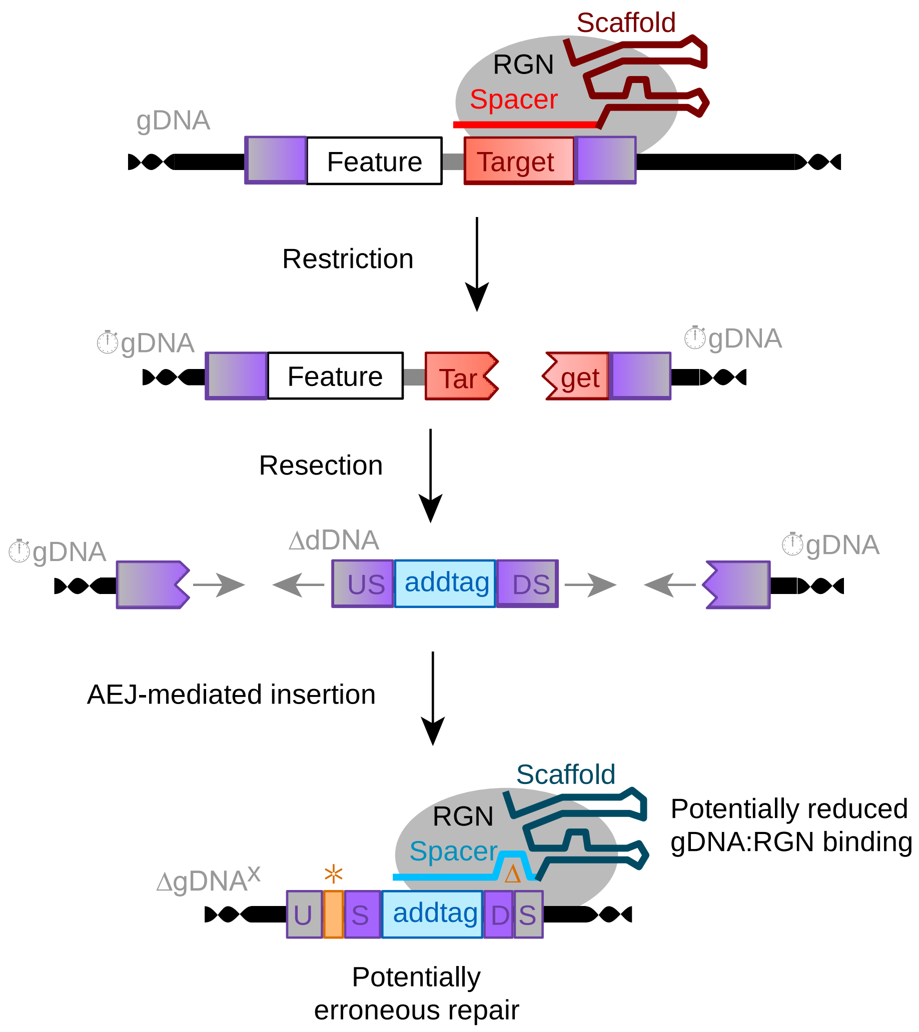


**Figure 2.30 – Double-stranded break repair through AEJ-mediated insertion during AddTag Step 1**

When alternative end joining (AEJ) processes repair the genomic double-stranded break induced by the gRNA:RGN complex, errors might occur (orange). These types of errors can be observed by sequencing the restricted locus, and designing PCR primers against the introduced variation. Black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by a DNA repair process to incorporate dDNA into the gDNA. More detail can be found in **Common Figure Description 1**.

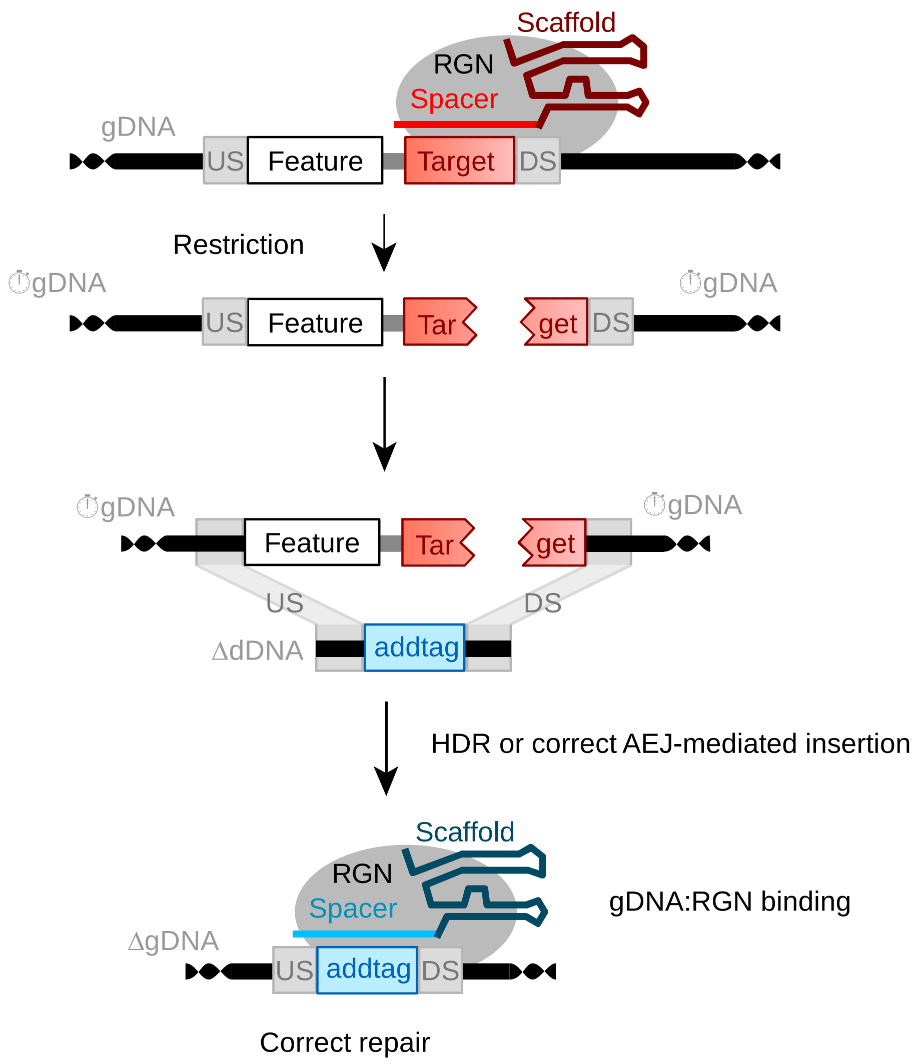
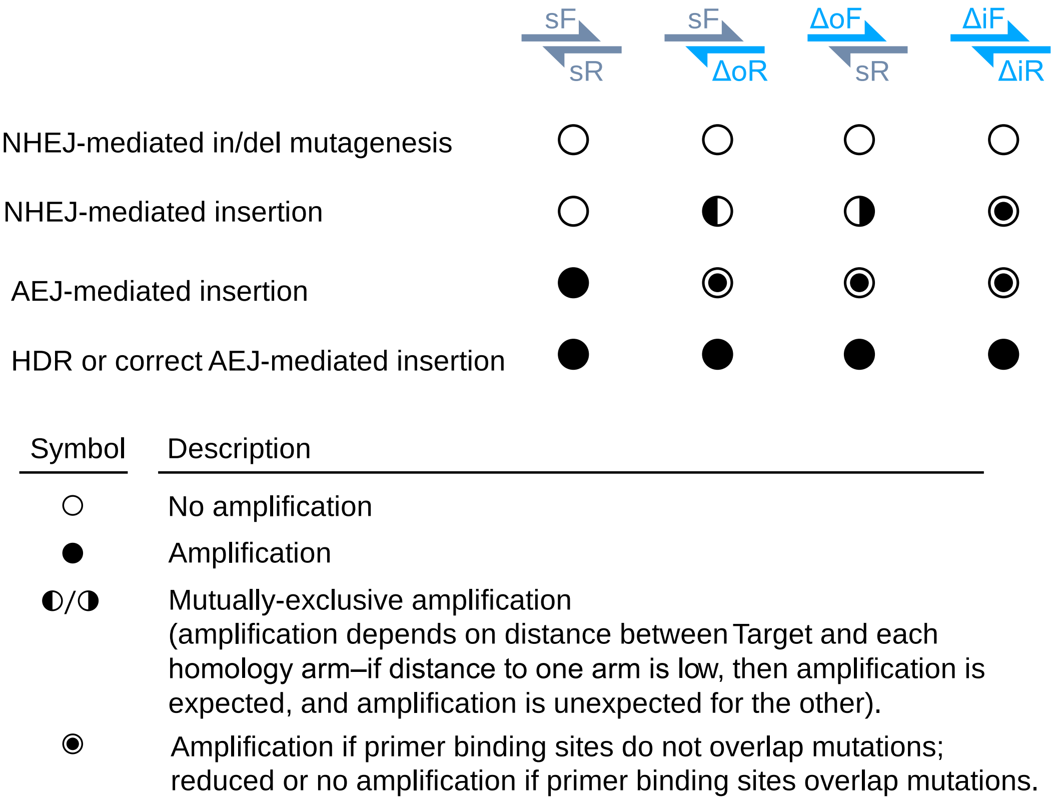


**Figure 2.31 – Double-stranded break repair through HDR or correct AEJ-mediated insertion during AddTag Step 1**

Both homology-direced repair (HDR) and alternative end joining (AEJ) processes can correctly insert the `addtag` from the dDNA into the restricted locus as intended. Black arrows represent gRNA and RGN complex association and restriction of gDNA, followed by a DNA repair process to incorporate dDNA into the gDNA. More detail can be found in **Common Figure Description 1**.



| | sF / sR | sF / ΔoR | ΔoF / sR | ΔiF / ΔiR |
|---|---|---|---|---|
| NHEJ-mediated in/del mutagenesis | ○ | ○ | ○ | ○ |
| NHEJ-mediated insertion | ○ | ◐ | ◑ | ◉ |
| AEJ-mediated insertion | ● | ◉ | ◉ | ◉ |
| HDR or correct AEJ-mediated insertion | ● | ● | ● | ● |

| Symbol | Description |
|---|---|
| ○ | No amplification |
| ● | Amplification |
| ◐/◑ | Mutually-exclusive amplification<br>(amplification depends on distance between Target and each homology arm–if distance to one arm is low, then amplification is expected, and amplification is unexpected for the other). |
| ◉ | Amplification if primer binding sites do not overlap mutations;<br>reduced or no amplification if primer binding sites overlap mutations. |

**Figure 2.32 – Expected vPCR amplification of gDNA following Step 1**

The DNA repair mechanism that handles the double-stranded break introduced by the RGN can affect the results of vPCR verification. Therefore, vPCR is useful for diagnosing which repair mechanism was used for Step 1 DNA repair. NHEJ-mediated in/del mutagenesis (Figure 2.28) and NHEJ-mediated insertion (Figure 2.29) result in diagnostically-different PCR results than expected through HDR or correct AEJ-mediated insertion (Figure 2.31). However, AEJ-mediated insertion (Figure 2.30) may produce unintended mutations that are not distinguishable from correct editing. For more information on the mutually-exclusive amplification, please see 4.3.1 and Table 4.4.

AddTag suffers from many of the same biases as other genome editing methods. For instance, multinucleate cells, which are common in fungi like *Sclerotinia sclerotiorum*, require either complete editing in each nucleus or additional culturing steps to select for non-chimeric cells [110]. Nevertheless, ADDTAG formalizes the process for modifying the same locus multiple times, editing genomic sites without direct Targets, and performing effective genotype validation through vPCR. For these reasons, AddTag is a useful advance in precision genome editing.

## 2.6 References

1. Seher TD, Nguyen N, Ramos D, Bapat P, Nobile CJ, Sindi SS, Hernday AD. AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing. *G3 Genes|Genomes|Genetics*. 2021. DOI: 10.1093/g3journal/jkab216.
2. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial

immunity. *Science (New York, NY)*. 2012; 337(6096):816-21. Epub 2012/06/28. DOI: 10.1126/science.1225829, PMID: 22745249.

3. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. 2014; 157(6):1262-78. DOI: 10.1016/j.cell.2014.05.010, PMID: 24906146.
4. Adli M. The CRISPR tool kit for genome editing and beyond. *Nature Communications*. 2018; 9(1):1911. DOI: 10.1038/s41467-018-04252-2, PMID: 29765029.
5. Satomura A, Nishioka R, Mori H, Sato K, Kuroda K, Ueda M. Precise genome-wide base editing by the CRISPR nickase system in yeast. *Scientific Reports*. 2017; 7(1):2095. DOI: 10.1038/s41598-017-02013-7, PMID: 28522803.
6. Satomura A, Nishioka R, Mori H, Sato K, Kuroda K, Ueda M. Erratum: Precise genome-wide base editing by the CRISPR Nickase system in yeast. *Scientific Reports*. 2017; 7(1):12354. DOI: 10.1038/s41598-017-09606-2, PMID: 28955053.
7. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*. 2014; 513:569. DOI: 10.1038/nature13579, PMID: 25079318.
8. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Research*. 2013; 41(7):4336-43. DOI: 10.1093/nar/gkt135, PMID: 23460208.
9. Mans R, van Rossum HM, Wijsman M, Backx A, Kuijpers NGA, van den Broek M, Daran-Lapujade P*, et al*. CRISPR/Cas9: A molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae*. *FEMS Yeast Research*. 2015; 15(2):fov004-fov. DOI: 10.1093/femsyr/fov004, PMID: 25743786.
10. Horwitz Andrew A, Walter Jessica M, Schubert Max G, Kung Stephanie H, Hawkins K, Platt Darren M, Hernday Aaron D*, et al*. Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR-Cas. *Cell Systems*. 2015; 1(1):88-96. DOI: 10.1016/j.cels.2015.02.001, PMID: 27135688.
11. Biot-Pelletier D, Martin VJJ. Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. *Journal of Biological Engineering*. 2016; 10(1):6. DOI: 10.1186/s13036-016-0028-1, PMID: 27134651.
12. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*. 2012; 13(1):59-69. DOI: 10.1038/nrg3095, PMID: 22143240.
13. Lee ME, DeLoache WC, Cervantes B, Dueber JE. A highly characterized yeast toolkit for modular, multipart assembly. *ACS Synthetic Biology*. 2015; 4(9):975-86. DOI: 10.1021/sb500366v, PMID: 25871405.
14. Storici F, Lewis LK, Resnick MA. *In vivo* site-directed mutagenesis using oligonucleotides. *Nature Biotechnology*. 2001; 19(8):773-6. DOI: 10.1038/90837, PMID: 11479573.
15. Xie F, Ye L, Chang JC, Beyer AI, Wang J, Muench MO, Kan YW. Seamless gene correction of β-thalassemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Research*. 2014; 24(9):1526-33. DOI: 10.1101/gr.173427.114, PMID: 25096406.
16. Vicencio J, Martínez-Fernández C, Serrat X, Cerón J. Efficient generation of endogenous fluorescent reporters by nested CRISPR in *Caenorhabditis elegans*. *Genetics*. 2019; 211(4):1143-54. DOI: 10.1534/genetics.119.301965, PMID: 30696716.

17. Elison GL, Song R, Acar M. A precise genome editing method reveals insights into the activity of eukaryotic promoters. *Cell Reports*. 2017; 18(1):275-86. DOI: 10.1016/j.celrep.2016.12.014, PMID: 28052256.
18. Kim H-S, Smithies O. Recombinant fragment assay for gene targetting based on the polymerase chain reaction. *Nucleic Acids Research*. 1988; 16(18):8887-903. DOI: 10.1093/nar/16.18.8887, PMID: 3174435.
19. Rodríguez-López M, Cotobal C, Fernández-Sánchez O, Borbarán Bravo N, Oktriani R, Abendroth H, Uka D*, et al*. A CRISPR/Cas9-based method and primer design tool for seamless genome editing in fission yeast. *Wellcome Open Research*. 2017; 1(19). DOI: 10.12688/wellcomeopenres.10038.3, PMID: 28612052.
20. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Reverse Genetics. An introduction to genetic analysis. 7th ed. New York: W. H. Freeman; 2000.
21. Griffiths AJF, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Complementation. An introduction to genetic analysis. 7th ed. New York: W. H. Freeman; 2000.
22. Cheng S, Nguyen MH, Zhang Z, Jia H, Handfield M, Clancy CJ. Evaluation of the roles of four *Candida albicans* genes in virulence by using gene disruption strains that express *URA3* from the native locus. *Infection and Immunity*. 2003; 71(10):6101-3. DOI: 10.1128/iai.71.10.6101-6103.2003, PMID: 14500538.
23. Brand A, MacCallum DM, Brown AJP, Gow NAR, Odds FC. Ectopic expression of *URA3* can influence the virulence phenotypes and proteome of *Candida albicans* but can be overcome by targeted reintegration of *URA3* at the *RPS10* locus. *Eukaryotic Cell*. 2004; 3(4):900-9. DOI: 10.1128/ec.3.4.900-909.2004, PMID: 15302823.
24. Staab JF, Sundstrom P. *URA3* as a selectable marker for disruption and virulence assessment of *Candida albicans* genes. *Trends in Microbiology*. 2003; 11(2):69-73. DOI: 10.1016/S0966-842X(02)00029-X, PMID: 12598128.
25. Freire-Benéitez V, Price RJ, Tarrant D, Berman J, Buscaino A. *Candida albicans* repetitive elements display epigenetic diversity and plasticity. *Scientific reports*. 2016; 6:22989-. DOI: 10.1038/srep22989, PMID: 26971880.
26. Freire-Benéitez V, Price RJ, Buscaino A. The chromatin of *Candida albicans* pericentromeres bears features of both euchromatin and heterochromatin. *Front Microbiol*. 2016; 7:759-. DOI: 10.3389/fmicb.2016.00759, PMID: 27242771.
27. Burrack LS, Hutton HF, Matter KJ, Clancey SA, Liachko I, Plemmons AE, Saha A*, et al*. Neocentromeres provide chromosome segregation accuracy and centromere clustering to multiple loci along a *Candida albicans* chromosome. *PLoS genetics*. 2016; 12(9):e1006317-e. DOI: 10.1371/journal.pgen.1006317, PMID: 27662467.
28. Lohse MB, Ene IV, Craik VB, Hernday AD, Mancera E, Morschhäuser J, Bennett RJ, Johnson AD. Systematic genetic screen for transcriptional regulators of the *Candida albicans* white-opaque switch. *Genetics*. 2016; 203(4):1679-92. DOI: 10.1534/genetics.116.190645, PMID: 27280690.
29. Gow NAR. A developmental program for *Candida* commensalism. *Nature Genetics*. 2013; 45(9):967-8. DOI: 10.1038/ng.2737, PMID: 23985683.
30. Hernday AD, Lohse MB, Nobile CJ, Noiman L, Laksana CN, Johnson AD. Ssn6 defines a new level of regulation of white-opaque switching in *Candida albicans* and is required for the stochasticity of the switch. *mBio*. 2016; 7(1):e01565-15. DOI: 10.1128/mBio.01565-15, PMID: 26814177.
31. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred: I. Accuracy assessment. *Genome Research*. 1998; 8(3):175-85. DOI: 10.1101/gr.8.3.175, PMID: 9521921.

32. Ewing B, Green P. Base-calling of automated sequencer traces using Phred: II. Error probabilities. *Genome Research*. 1998; 8(3):186-94. DOI: 10.1101/gr.8.3.186, PMID: 9521922.

33. Hernday AD, Lohse MB, Fordyce PM, Nobile CJ, DeRisi JL, Johnson AD. Structure of the transcriptional network controlling white-opaque switching in *Candida albicans*. *Molecular Microbiology*. 2013; 90(1):22-35. DOI: 10.1111/mmi.12329, PMID: 23855748.

34. Lohse MB, Zordan RE, Cain CW, Johnson AD. Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in *Candida albicans*. *Proceedings of the National Academy of Sciences*. 2010; 107(32):14105-10. DOI: 10.1073/pnas.1005911107, PMID: 20660774.

35. Zhang S, Zhang T, Yan M, Ding J, Chen J. Crystal structure of the WOPR-DNA complex and implications for Wor1 function in white-opaque switching of Candida albicans. *Cell Research*. 2014; 24(9):1108-20. DOI: 10.1038/cr.2014.102, PMID: 25091450.

36. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I*, et al*. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech*. 2016; 34(2):184-91. DOI: 10.1038/nbt.3437, PMID: 26780180.

37. Zheng T, Hou Y, Zhang P, Zhang Z, Xu Y, Zhang L, Niu L*, et al*. Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. *Scientific Reports*. 2017; 7(1):40638. DOI: 10.1038/srep40638, PMID: 28098181.

38. Nobile CJ, Nett JE, Hernday AD, Homann OR, Deneault J-S, Nantel A, Andes DR*, et al*. Biofilm matrix regulation by *Candida albicans* Zap1. *PLOS Biology*. 2009; 7(6):e1000133. DOI: 10.1371/journal.pbio.1000133, PMID: 19529758.

39. Zhao H, Butler E, Rodgers J, Spizzo T, Duesterhoeft S, Eide D. Regulation of zinc homeostasis in yeast by binding of the *ZAP1* transcriptional activator to zinc-responsive promoter elements. *Journal of Biological Chemistry*. 1998; 273(44):28713-20. DOI: 10.1074/jbc.273.44.28713, PMID: 9786867.

40. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM*, et al*. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004; 431(7004):99-104. DOI: 10.1038/nature02800, PMID: 15343339.

41. Nobile CJ, Johnson AD. *Candida albicans* biofilms and human disease. *Annual Review of Microbiology*. 2015; 69(1):71-92. DOI: 10.1146/annurev-micro-091014-104330, PMID: 26488273.

42. Nobile Clarissa J, Fox Emily P, Nett Jeniel E, Sorrells Trevor R, Mitrovich Quinn M, Hernday Aaron D, Tuch Brian B*, et al*. A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell*. 2012; 148(1):126-38. DOI: 10.1016/j.cell.2011.10.048, PMID: 22265407.

43. Glazier VE, Murante T, Murante D, Koselny K, Liu Y, Kim D, Koo H, Krysan DJ. Genetic analysis of the *Candida albicans* biofilm transcription factor network using simple and complex haploinsufficiency. *PLOS Genetics*. 2017; 13(8):e1006948. DOI: 10.1371/journal.pgen.1006948, PMID: 28793308.

44. Gulati M, Lohse MB, Ennis CL, Gonzalez RE, Perry AM, Bapat P, Arevalo AV*, et al*. *In vitro* culturing and screening of *Candida albicans* biofilms. *Current Protocols in Microbiology*. 2018; 50(1):e60. DOI: 10.1002/cpmc.60, PMID: 29995344.

45. Lohse MB, Gulati M, Valle Arevalo A, Fishburn A, Johnson AD, Nobile CJ. Assessment and optimizations of *Candida albicans in vitro* biofilm assays. *Antimicrobial Agents and Chemotherapy*. 2017; 61(5):e02749-16. DOI: 10.1128/aac.02749-16, PMID: 28289028.

46. Gulati M, Nobile CJ. *Candida albicans* biofilms: Development, regulation, and molecular mechanisms. *Microbes and Infection*. 2016; 18(5):310-21. DOI: 10.1016/j.micinf.2016.01.002, PMID: 26806384.
47. Bar-Yosef H, Vivanco Gonzalez N, Ben-Aroya S, Kron SJ, Kornitzer D. Chemical inhibitors of *Candida albicans* hyphal morphogenesis target endocytosis. *Scientific Reports*. 2017; 7(1):5692. DOI: 10.1038/s41598-017-05741-y, PMID: 28720834.
48. Sudbery PE. Growth of *Candida albicans* hyphae. *Nature Reviews Microbiology*. 2011; 9(10):737-48. DOI: 10.1038/nrmicro2636, PMID: 21844880.
49. Kadosh D. Regulatory mechanisms controlling morphology and pathogenesis in *Candida albicans*. *Current Opinion in Microbiology*. 2019; 52:27-34. DOI: 10.1016/j.mib.2019.04.005, PMID: 31129557.
50. Desai PR, Lengeler K, Kapitan M, Janßen SM, Alepuz P, Jacobsen ID, Ernst JF. The 5′ untranslated region of the *EFG1* transcript promotes its translation to regulate hyphal morphogenesis in *Candida albicans*. *mSphere*. 2018; 3(4):e00280-18. DOI: 10.1128/mSphere.00280-18, PMID: 29976646.
51. Kadosh D. Control of *Candida albicans* morphology and pathogenicity by post-transcriptional mechanisms. *Cellular and Molecular Life Sciences*. 2016; 73(22):4265-78. DOI: 10.1007/s00018-016-2294-y, PMID: 27312239.
52. Kim J, Lee J-E, Lee J-S. Histone deacetylase-mediated morphological transition in *Candida albicans*. *Journal of Microbiology*. 2015; 53(12):805-11. DOI: 10.1007/s12275-015-5488-3, PMID: 26626350.
53. Hnisz D, Bardet AF, Nobile CJ, Petryshyn A, Glaser W, Schöck U, Stark A, Kuchler K. A histone deacetylase adjusts transcription kinetics at coding sequences during *Candida albicans* morphogenesis. *PLOS Genetics*. 2012; 8(12):e1003118. DOI: 10.1371/journal.pgen.1003118, PMID: 23236295.
54. Tebarth B, Doedt T, Krishnamurthy S, Weide M, Monterola F, Dominguez A, Ernst JF. Adaptation of the Efg1p morphogenetic pathway in *Candida albicans* by negative autoregulation and PKA-dependent repression of the *EFG1* gene. *Journal of Molecular Biology*. 2003; 329(5):949-62. DOI: 10.1016/S0022-2836(03)00505-9, PMID: 12798685.
55. Veri AO, Miao Z, Shapiro RS, Tebbji F, O'Meara TR, Kim SH, Colazo J, *et al*. Tuning Hsf1 levels drives distinct fungal morphogenetic programs with depletion impairing Hsp90 function and overexpression expanding the target space. *PLOS Genetics*. 2018; 14(3):e1007270. DOI: 10.1371/journal.pgen.1007270, PMID: 29590106.
56. Setiadi ER, Doedt T, Cottier F, Noffz C, Ernst JF. Transcriptional response of *Candida albicans* to hypoxia: Linkage of oxygen sensing and Efg1p-regulatory networks. *Journal of Molecular Biology*. 2006; 361(3):399-411. DOI: 10.1016/j.jmb.2006.06.040, PMID: 16854431.
57. Doedt T, Krishnamurthy S, Bockmühl DP, Tebarth B, Stempel C, Russell CL, Brown AJP, Ernst JF. APSES proteins regulate morphogenesis and metabolism in *Candida albicans*. *Molecular Biology of the Cell*. 2004; 15(7):3167-80. DOI: 10.1091/mbc.e03-11-0782, PMID: 15218092.
58. Rastogi SK, van Wijlick L, Ror S, Lee KK, Román E, Agarwal P, Manzoor N, *et al*. Ifu5, a WW domain-containing protein interacts with Efg1 to achieve coordination of normoxic and hypoxic functions to influence pathogenicity traits in *Candida albicans*. *Cellular Microbiology*. 2020; 22(2):e13140. DOI: 10.1111/cmi.13140, PMID: 31736226.
59. Riggle PJ, Andrutis KA, Chen X, Tzipori SR, Kumamoto CA. Invasive lesions containing filamentous forms produced by a *Candida albicans* mutant that is

defective in filamentous growth in culture. *Infection and Immunity*. 1999; 67(7):3649-52. DOI: 10.1128/IAI.67.7.3649-3652.1999, PMID: 10377153.

60. Cleary IA, Lazzell AL, Monteagudo C, Thomas DP, Saville SP. *BRG1* and *NRG1* form a novel feedback circuit regulating *Candida albicans* hypha formation and virulence. *Molecular Microbiology*. 2012; 85(3):557-73. DOI: 10.1111/j.1365-2958.2012.08127.x, PMID: 22757963.

61. Zavrel M, Majer O, Kuchler K, Rupp S. Transcription factor Efg1 shows a haploinsufficiency phenotype in modulating the cell wall architecture and immunogenicity of *Candida albicans*. *Eukaryotic Cell*. 2012; 11(2):129-40. DOI: 10.1128/ec.05206-11, PMID: 22140230.

62. Stoldt VR, Sonneborn A, Leuker CE, Ernst JF. Efg1p, an essential regulator of morphogenesis of the human pathogen *Candida albicans*, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. *The EMBO Journal*. 1997; 16(8):1982-91. DOI: 10.1093/emboj/16.8.1982, PMID: 9155024.

63. Sohn K, Urban C, Brunner H, Rupp S. *EFG1* is a major regulator of cell wall dynamics in *Candida albicans* as revealed by DNA microarrays. *Molecular Microbiology*. 2003; 47(1):89-102. DOI: 10.1046/j.1365-2958.2003.03300.x, PMID: 12492856.

64. Li F, Palecek SP. *EAP1*, a *Candida albicans* gene involved in binding human epithelial cells. *Eukaryotic Cell*. 2003; 2(6):1266-73. DOI: 10.1128/ec.2.6.1266-1273.2003, PMID: 14665461.

65. Desai PR, van Wijlick L, Kurtz D, Juchimiuk M, Ernst JF. Hypoxia and temperature regulated morphogenesis in *Candida albicans*. *PLOS Genetics*. 2015; 11(8):e1005447. DOI: 10.1371/journal.pgen.1005447, PMID: 26274602.

66. Su C, Yu J, Sun Q, Liu Q, Lu Y. Hyphal induction under the condition without inoculation in *Candida albicans* is triggered by Brg1-mediated removal of *NRG1* inhibition. *Molecular Microbiology*. 2018; 108(4):410-23. DOI: 10.1111/mmi.13944, PMID: 29485686.

67. Hanumantha Rao K, Paul S, Ghosh S. N-acetylglucosamine signaling: Transcriptional dynamics of a novel sugar sensing cascade in a model pathogenic yeast, *Candida albicans*. *Journal of Fungi*. 2021; 7(1):65. DOI: 10.3390/jof7010065, PMID: 33477740.

68. Lu Y, Su C, Liu H. A GATA transcription factor recruits Hda1 in response to reduced Tor1 signaling to establish a hyphal chromatin state in *Candida albicans*. *PLOS Pathogens*. 2012; 8(4):e1002663. DOI: 10.1371/journal.ppat.1002663, PMID: 22536157.

69. Desai JV. *Candida albicans* hyphae: From growth initiation to invasion. *J Fungi (Basel)*. 2018; 4(1):10. DOI: 10.3390/jof4010010, PMID: 29371503.

70. Crawford AC, Lehtovirta-Morley LE, Alamir O, Niemiec MJ, Alawfi B, Alsarraf M, Skrahina V*, et al*. Biphasic zinc compartmentalisation in a human fungal pathogen. *PLOS Pathogens*. 2018; 14(5):e1007013. DOI: 10.1371/journal.ppat.1007013, PMID: 29727465.

71. Chaudhuri B, Ingavale S, Bachhawat AK. apd1+, a gene required for red pigment formation in *ade6* mutants of *Schizosaccharomyces pombe*, encodes an enzyme required for glutathione biosynthesis: A role for glutathione and a glutathione-conjugate pump. *Genetics*. 1997; 145(1):75-83. PMID: 9017391.

72. Poulter RT, Rikkerink EH. Genetic analysis of red, adenine-requiring mutants of *Candida albicans*. 1983; 156(3):1066-77. DOI: 10.1128/jb.156.3.1066-1077.1983, PMID: 6358187.

73. Zonneveld BJM, van der Zanden AL. The red *ade* mutants of *Kluyveromyces lactis* and their classification by complementation with cloned *ADE1* or *ADE2* genes from

*Saccharomyces cerevisiae*. *Yeast*. 1995; 11(9):823-7. DOI: 10.1002/yea.320110904, PMID: 7483846.

74. Nguyen N, Quail MMF, Hernday AD. An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. *mSphere*. 2017; 2(2). DOI: 10.1128/mSphereDirect.00149-17, PMID: 28497115.

75. Vyas VK, Barrasa MI, Fink GR. A *Candida albicans* CRISPR system permits genetic engineering of essential genes and gene families. *Science advances*. 2015; 1(3):e1500248. DOI: 10.1126/sciadv.1500248, PMID: 25977940.

76. Vyas VK, Bushkin GG, Bernstein DA, Getz MA, Sewastianik M, Barrasa MI, Bartel DP, Fink GR. New CRISPR mutagenesis strategies reveal variation in repair mechanisms among fungi. *mSphere*. 2018; 3(2):e00154-18. DOI: 10.1128/mSphere.00154-18, PMID: 29695624.

77. Min K, Ichikawa Y, Woolford CA, Mitchell AP. *Candida albicans* gene deletion with a transient CRISPR-Cas9 system. *mSphere*. 2016; 1(3). DOI: 10.1128/mSphere.00130-16, PMID: 27340698.

78. Shahana S, Childers DS, Ballou ER, Bohovych I, Odds FC, Gow NAR, Brown AJP. New clox systems for rapid and efficient gene disruption in *Candida albicans*. *PLOS ONE*. 2014; 9(6):e100390. DOI: 10.1371/journal.pone.0100390, PMID: 24940603.

79. Park J, McCormick SP, Cockrell AL, Chakrabarti M, Lindahl PA. High-spin ferric ions in *Saccharomyces cerevisiae* vacuoles are reduced to the ferrous state during adenine-precursor detoxification. *Biochemistry*. 2014; 53(24):3940-51. Epub 2014/06/11. DOI: 10.1021/bi500148y, PMID: 24919141.

80. Toledano MB, Delaunay-Moisan A, Outten CE, Igbaria A. Functions and cellular compartmentation of the thioredoxin and glutathione pathways in yeast. *Antioxid Redox Signal*. 2013; 18(13):1699-711. Epub 2013/02/05. DOI: 10.1089/ars.2012.5033, PMID: 23198979.

81. Morgan B, Ezeriņa D, Amoako TNE, Riemer J, Seedorf M, Dick TP. Multiple glutathione disulfide removal pathways mediate cytosolic redox homeostasis. *Nature Chemical Biology*. 2012; 9:119. DOI: 10.1038/nchembio.1142, PMID: 23242256.

82. Nevzglyadova OV, Kuznetsova IM, Mikhailova EV, Artamonova TO, Artemov AV, Mittenberg AG, Kostyleva EI*, et al*. The effect of red pigment on the amyloidization of yeast proteins. *Yeast*. 2011; 28(7):505-26. DOI: 10.1002/yea.1854, PMID: 21547947.

83. Bharathi V, Girdhar A, Prasad A, Verma M, Taneja V, Patel BK. Use of *ade1* and *ade2* mutations for development of a versatile red/white colour assay of amyloid-induced oxidative stress in *Saccharomyces cerevisiae*. *Yeast*. 2016; 33(12):607-20. DOI: 10.1002/yea.3209, PMID: 27654890.

84. Ugolini S, Bruschi CV. The red/white colony color assay in the yeast *Saccharomyces cerevisiae*: Epistatic growth advantage of white *ade8-18*, *ade2* cells over red *ade2* cells. *Current Genetics*. 1996; 30(6):485-92. DOI: 10.1007/s002940050160, PMID: 8939809.

85. Meškauskas A, Ksenzenko V, Shlyapnikov M, Kryukov V, Čitavičius D. 'Red pigment' from *ADE-2* mutants of *S. cerevisiae* prevents DNA cleavage by restriction endonucleases. *FEBS Letters*. 1985; 182(2):413-4. DOI: 10.1016/0014-5793(85)80344-6, PMID: 2984046.

86. Hastings PJ, Rosenberg SM. Gene Conversion. In: Delves PJ, editor. Encyclopedia of Immunology. 2nd ed. Oxford: Elsevier; 1998; p. 969-73. DOI: 10.1006/rwei.1999.0252.

87. Holsclaw JK, Hatkevich T, Sekelsky J. Chapter 9 – Meiotic and Mitotic Recombination: First in Flies. In: Kovalchuk I, Kovalchuk O, editors. Genome

Stability. Boston: Academic Press; 2016; p. 139-54. DOI: 10.1016/B978-0-12-803309-8.00009-4.

88. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Molecular Therapy - Nucleic Acids*. 2015; 4:e264. DOI: 10.1038/mtna.2015.37, PMID: 26575098.

89. Duan J, Lu G, Xie Z, Lou M, Luo J, Guo L, Zhang Y. Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Research*. 2014; 24:1009. DOI: 10.1038/cr.2014.87, PMID: 24980957.

90. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N*, et al*. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotech*. 2015; 33(2):187-97. DOI: 10.1038/nbt.3117, PMID: 25513782.

91. Roth TL, Puig-Saus C, Yu R, Shifrut E, Carnevale J, Li PJ, Hiatt J*, et al*. Reprogramming human T cell function and specificity with non-viral genome targeting. *Nature*. 2018. DOI: 10.1038/s41586-018-0326-5, PMID: 29995861.

92. Cho SW, Kim S, Kim Y, Kweon J, Kim HS, Bae S, Kim J-S. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Research*. 2014; 24(1):132-41. DOI: 10.1101/gr.162339.113, PMID: 24253446.

93. Kim D, Bae S, Park J, Kim E, Kim S, Yu HR, Hwang J*, et al*. Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods*. 2015; 12(3):237-43. DOI: 10.1038/nmeth.3284, PMID: 25664545.

94. Shen B, Zhang W, Zhang J, Zhou J, Wang J, Chen L, Wang L*, et al*. Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nature Methods*. 2014; 11(4):399-402. DOI: 10.1038/nmeth.2857, PMID: 24584192.

95. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology*. 2013; 31(9):833-8. DOI: 10.1038/nbt.2675, PMID: 23907171.

96. Ran FA, Hsu Patrick D, Lin C-Y, Gootenberg Jonathan S, Konermann S, Trevino AE, Scott David A*, et al*. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*. 2013; 154(6):1380-9. DOI: 10.1016/j.cell.2013.08.021, PMID: 23992846.

97. Yarrington RM, Verma S, Schwartz S, Trautman JK, Carroll D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 *in vivo*. *Proceedings of the National Academy of Sciences*. 2018; 115(38):9351-8. DOI: 10.1073/pnas.1810062115, PMID: 30201707.

98. Cho S, Choe D, Lee E, Kim SC, Palsson BO, Cho B-K. High-level dCas9 expression induces abnormal cell morphology in *Escherichia coli*. *ACS Synthetic Biology*. 2018; 7(4):1085-94. DOI: 10.1021/acssynbio.7b00462, PMID: 29544049.

99. Shaw S, Knüsel S, Hoenner S, Roditi I. A transient CRISPR/Cas9 expression system for genome editing in *Trypanosoma brucei*. *BMC Research Notes*. 2020; 13(1):268. DOI: 10.1186/s13104-020-05089-z, PMID: 32493474.

100. Aird EJ, Lovendahl KN, St. Martin A, Harris RS, Gordon WR. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Communications Biology*. 2018; 1(1):54. DOI: 10.1038/s42003-018-0054-2, PMID: 30271937.

101. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Research*. 1999; 27(5):1223-42. DOI: 10.1093/nar/27.5.1223, PMID: 9973609.

102. Haapaniemi E, Botla S, Persson J, Schmierer B, Taipale J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nature Medicine*. 2018. DOI: 10.1038/s41591-018-0049-z, PMID: 29892067.

103. Stratigopoulos G, De Rosa MC, LeDuc CA, Leibel RL, Doege CA. DMSO increases efficiency of genome editing at two non-coding loci. *PLOS ONE*. 2018; 13(6):e0198637. DOI: 10.1371/journal.pone.0198637, PMID: 29864154.

104. Tsabar M, Eapen VV, Mason JM, Memisoglu G, Waterman DP, Long MJ, Bishop DK, Haber JE. Caffeine impairs resection during DNA break repair by reducing the levels of nucleases Sae2 and Dna2. *Nucleic Acids Research*. 2015; 43(14):6889-901. DOI: 10.1093/nar/gkv520, PMID: 26019182.

105. Selby CP, Sancar A. Molecular mechanisms of DNA repair inhibition by caffeine. *Proceedings of the National Academy of Sciences*. 1990; 87(9):3522-5. DOI: 10.1073/pnas.87.9.3522, PMID: 2185474.

106. Kostyushev D, Kostyusheva A, Brezgin S, Zarifyan D, Utkina A, Goptar I, Chulanov V. Suppressing the NHEJ pathway by DNA-PKcs inhibitor NU7026 prevents degradation of HBV cccDNA cleaved by CRISPR/Cas9. *Scientific Reports*. 2019; 9(1):1847. DOI: 10.1038/s41598-019-38526-6, PMID: 30755668.

107. Yu C, Liu Y, Ma T, Liu K, Xu S, Zhang Y, Liu H*, et al*. Small molecules enhance CRISPR genome editing in pluripotent stem cells. *Cell Stem Cell*. 2015; 16(2):142-7. DOI: 10.1016/j.stem.2015.01.003, PMID: 25658371.

108. Li G, Zhang X, Zhong C, Mo J, Quan R, Yang J, Liu D*, et al*. Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells. *Scientific Reports*. 2017; 7(1):8943. DOI: 10.1038/s41598-017-09306-x, PMID: 28827551.

109. Riesenberg S, Maricic T. Targeting repair pathways with small molecules increases precise genome editing in pluripotent stem cells. *Nature Communications*. 2018; 9(1):2164. DOI: 10.1038/s41467-018-04609-7, PMID: 29867139.

110. Li J, Zhang Y, Zhang Y, Yu P-L, Pan H, Rollins JA, Vidaver AK. Introduction of large sequence inserts by CRISPR-Cas9 to create pathogenicity mutants in the multinucleate filamentous pathogen *Sclerotinia sclerotiorum*. *mBio*. 2018; 9(3):e00567-18. DOI: 10.1128/mBio.00567-18, PMID: 29946044.

# Chapter 3 AddTag software Target identification and dDNA generation

## 3.1 Abstract

The AddTag (2-step) method enables scientists to serially edit a locus that displays redundant homology, poor quality genome binding sites (Targets) for RNA-guided nucleases (RGNs), and excessive allelic variation using CRISPR/Cas-induced homology-directed repair (HDR). I developed the ADDTAG software (https://github.com/tdseher/addtag-project) to identify useable Targets in or near the DNA chosen for editing (Feature), and to construct donor DNA sequences (dDNAs) to replace the Feature. The process has six parts: (A) Feature expansion and Step 1 Target identification, (B) Step 1 dDNA construction and Step 2 Target identification, (C) determination of Step 1 Target quality, (D) Step 2 dDNA construction, (E) determination of Step 2 Target quality, and (F) Final Target and dDNA ranking. Numerous methods for evaluating Target quality are supported and are user-selectable. I introduce a novel mathematical framework for comparing Target qualities across genomes and RGN molecules. Several methods are included for ensuring the Step 1 dDNA includes a Step 2 Target. As a whole, ADDTAG significantly simplifies and clarifies the decisions scientists need to make in order to develop successful CRISPR/Cas-induced HDR editing experiments.

## 3.2 Preface

This chapter is adapted from the manuscript authored by myself, Namkha Nguyen, Diana Ramos, Priyanka Bapat, Clarissa J. Nobile, Suzanne S. Sindi, and Aaron D. Hernday, titled "AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing," which was published by the peer-reviewed journal *G3 Genes|Genomes|Genetics* [1].

## 3.3 Introduction

### 3.3.1 ADDTAG simplifies decisions needed for genome editing experiments

In this chapter I describe the operations the ADDTAG software performs to simplify experimental design of AddTag genome editing experiments. I report how ADDTAG overcomes the obstacles with 1-step genome editing (2.3.1) through computational processes. I describe the interdependence between the Step 1 Target, Step 1 dDNA, Step 2 Target, and Step 2 dDNA of the AddTag (2-step), indirect editing method (2.3.2). For introductory information about genome editing through CRISPR/Cas-induced homology-directed repair (HDR), please see 1.3 and 2.3.

The objective of creating the ADDTAG software was not to create a new or better method for identifying RGN targets on a genome—rather, it was to integrate this process with the dDNA generation. Current CRISPR/Cas software facilitates two practical approaches toward designing genome editing experiments. First, in the name of efficiency, researchers often assume the genomes of their experimental organisms are accurately reflected in curated genome databases. The CRISPR/Cas design software they use are tailored to intentionally-limited sets of genomes, and pre-computed enzyme binding computations. These researchers prefer interactive tools with illustrated depictions of sequence alignments and annotations. A workable experimental scheme requires user guidance at multiple steps throughout the design process. While useful, these approaches are cumbersome when needing to consider large sets of Feature edits. In contrast, ADDTAG follows the second practical approach that maximizes customizability by front-loading all decisions, then performs the calculations without user supervision. ADDTAG is a general tool that allows for any genome to be used, rather than limited to a specific set of organisms. Additionally, any arbitrary Target motif (spacer constraints, cutting arrangement, and PAM sequence) can be used, even if only substandard scoring Algorithms are available for them.

This makes AddTag ideal for labs with a high-volume of genome editing demands, labs that edit non-model species or species with significant divergence from reference genomes, and labs that are using novel RGNs.

### 3.3.2 Overview of computational workflow

AddTag implements a core subroutine (Table 3.1) designed for Target identification (3.4.2) and dDNA generation (3.4.3). For both Step 1 and Step 2 Target evaluation, all potential Target sequences passing the prefilter (3.4.8) are aligned to the intended gDNA and dDNA sequences. Then those alignments are used to calculate off-target Algorithm scores (3.4.10), and the remaining Algorithm scores are calculated (3.4.9, 3.4.10). Next, the Targets are ranked according to their Algorithm weights (3.4.11).

**Table 3.1 – Computational workflow for identifying Targets and generating dDNA**

| | | # | Task |
|---|---|---|---|
| (A) | Expand Feature | (1) | (Expand Feature to include potential Step 1 Target sequences) |
| | | (2) | (Expand Feature to account for genome variation) |
| | | 3 | Identify potential Step 1 Target sequences |
| (B) | Construct Step 1 dDNA | 4 | Concatenate HAs with Insert |
| | | 5 | Identify potential Step 2 Target sequences |
| (C) | Evaluate Step 1 Targets | 6 | Pre-alignment filter of Step 1 Target sequences |
| | | 7 | Align Step 1 Target to genome and Step 1 dDNA |
| | | 8 | Post-alignment filter of potential off-targets |
| (D) | Construct Step 2 dDNA | 9 | Identify AmpF and AmpR primers |
| | | (10) | (Concatenate HAs with ✳Feature) |
| (E) | Evaluate Step 2 Targets | 11 | Pre-alignment filter of Step 2 Target sequences |
| | | 12 | Align Step 2 Target sequences to genome and Step 2 dDNA |
| | | 13 | Post-alignment filter of potential off-targets |
| (F) | Rank | 14 | Calculate Algorithm scores and weights for Step 1 Target sequences |
| | | 15 | Calculate Algorithm scores and weights for Step 2 Target sequences |
| | | 16 | Calculate primer and primer pair attribute scores and weights |

Parentheses indicate optional tasks. Homolog arm (HA). Feature expansion and AmpF and AmpR primers are outlined in Figure 2.7.

Following complete identification of all potential Step 1 dDNAs, Step 1 Targets, Step 2 dDNAs, and Step 2 Targets, AddTag outputs design chains that connect these components together in a logical fashion. Each Step 2 Target is linked directly with a Step 1 dDNA, and that Step 1 dDNA is liked with a Step 1 Target as well as a Step 2 dDNA.

### 3.3.3 AddTag introduces new features unavailable with other gRNA design software

There are many software solutions available to select an RGN Target for genome editing (Table 3.2). However, there are experimental protocols that do not have computational solutions to design them. We present AddTag to fulfill these needs. The best software packages let users specify the full set of gDNA sequences in their organism, accounting for the ploidy and uncertainty of complete genome coverage; let users specify the specific RGN molecules their biological system uses (for instance, 5'-adjacent PAM or 3'-adjacent PAM); let users choose Target scoring Algorithms compatible with their biological system and chosen RGN; and assist users in strain validation. AddTag incorporates all these useful enhancements into one software package. No other software suite incorporates all these into a single tool.

### Table 3.2 – Comparison of gRNA design software

| Name <URL> [CITATION] | Spacer motifs | 3'-adjacent PAM motifs | 5'-adjacent PAM motifs | Command line | Web | Interactive output | Integrated ranking | Scoring algorithms | Genomes | vPCR primers | Ambiguous bases | Knock-in design |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ADDTAG (This study) | Any | Any | Any | ✓ | ✗ | ✗ | ✓ | 13 | Any | ✓ | ✓ | ✓ |
| BENCHLING [2] | N{16,24} | N{2,8} | N{2,8} | ✗ | ✓ | ✓ | ✗ | 2 | 163 | Limited | ✗ | ✗ |
| BREAKING-CAS [3] | N{18,25} | Any | Any | ✗ | ✓ | ✓ | ✗ | 1 | 1457 | ✗ | Limited | ✗ |
| CAS9 TARGET FINDER [4] | GN{19} | NGG | ✗ | ✗ | ✓ | ✗ | ✗ | 1 | ✗ | ✗ | Limited | ✗ |
| CAS-DESIGNER [5] | Any | Any | Any | ✓ | ✓ | ✗ | ✗ | 4 | Any* | ✗ | Limited | ✗ |
| CASFINDER [6] | N{1,} | Any | ✗ | ✓ | ✗ | ✗ | ✗ | 1 | Any | ✗ | Limited | ✗ |
| CCTOP [7, 8] | Any | Any | ✗ | ✓ | ✓ | ✗ | ✗ | 2 | Any | ✗ | Limited | ✗ |
| CHOPCHOP [9, 10] | N{1,}* | Any* | Any* | ✓ | ✓ | ✓ | ✓ | 9 | Any* | ✓ | Limited | ✗ |
| CRISPOR [11] | N{20,24) | 20 | 13 | ✗ | ✓ | ✓ | ✗ | 16 | Any* | ✓ | ✗ | ✗ |
| CRISPR4P [12] | N{20} | NGG | ✗ | ✓ | ✓ | ✗ | ✗ | 0 | 1 | ✓ | ✗ | ✓ |
| CRISPR DESIGN TOOL [13] & CRISPR SPECIFICITY ANALYSIS [14] | N{18,22} | Any | Any | ✗ | ✓ | ✓ | ✗ | 1 | 39 | ✗ | ✗ | ✗ |
| CRISPRDIRECT [15] | N{20} | Any | ✗ | ✗ | ✓ | ✓ | ✓ | 4 | 671 | ✗ | ✗ | ✗ |
| CRISPR-P [16, 17] | N{20,22} | 6 | 8 | ✗ | ✓ | ✓ | ✗ | 4 | 75 plants | ✗ | ✗ | ✗ |
| CRISPRSCAN [18] | 2 | NGG | 2 | ✗ | ✓ | ✗ | ✗ | 2 | 14 animals | ✗ | ✗ | ✗ |
| GUIDESCAN [19] | N{20} | NGG | TTTN | ✓ | ✓ | ✗ | ✗ | 2 | Any* | ✗ | Limited | ✗ |
| SGRNA DESIGN TOOL [20, 21] | 2 | 2 | 2 | ✗ | ✓ | ✗ | ✓ | 6 | 4 | ✗ | ✗ | ✗ |
| SYNTHEGO CRISPR DESIGN TOOL | N{20} | NGG | ✗ | ✗ | ✓ | ✓ | ✗ | 1 | >120,000 | ✗ | ✗ | ✗ |
| VARSCOT [22] | N{20} | NGG | ✗ | ✓ | ✗ | ✗ | ✗ | 3 | Any | ✗ | ✓ | ✗ |

In the columns containing motif descriptions or quantifiers, the text "Any" refers to motifs >0 nt in length. Motifs are defined by a string of (possibly ambiguous) nucleotide characters. Repeated characters are followed by braces ({}) containing comma-separated numbers representing the minimum and maximum number of repeats, respectively. If only one number is listed, then the minimum and maximum are equal. If the second number is omitted, then no maximum exists. Motifs are specified, when possible. Otherwise a number representing the total number of different motifs allowed is given. The asterisk (*) indicates that this element differs between the web and command line versions. The column "Ambiguous bases" refers to the ability to identify Targets from regions in the genome that contain ambiguous nucleotides. The column "Integrated ranking" means that results are ordered based on a combination of metrics, and not just one algorithm score.

Many software packages exist for facilitating CRISPR/Cas-based genome edits with varying versatility [23, 24] (Table 3.2). Some provide interactive graphical interfaces [3], but require user intervention at multiple steps in the design process [25]. Some CRISPR/Cas design programs are tailored to specific sets of genomes (Table 3.2, "Genomes"), and pre-computed enzyme binding computations [19, 26-28]. The programs with the simplest outputs order candidate Targets based on a single scoring algorithm (Table 3.2, "Integrated ranking"). Many let users choose a Target scoring algorithm and PAM compatible with their biological system and chosen RGN [5, 9-11]. Several tools simultaneously calculate multiple scoring algorithms [9, 10, 25], most notably CRISPOR [11]. CRISPR4P [12] introduced limited forms of automated knock-in dDNA and vPCR primer design. VARSCOT [22] introduced ambiguous nucleotide compatibility and allows for uncertainty in genome coverage and ploidy. CRISPRDIRECT [15] introduced integrated Target ranking. We have taken many of the best features of all these software programs and combined them into a single, universally-applicable tool.

One feature that distinguishes ADDTAG from other Target identification software is that ADDTAG de-couples the PAM sequence identification from the score evaluation. For instance, the typical *in silico* Cas9 gRNA design uses NGG for searching (on-targets), and NRG for scoring (off-targets). Few gRNA design software programs provide the functionality to search for arbitrary PAM motifs (Table 3.2), largely because scoring algorithms have not been empirically verified for these. However, ADDTAG can identify useful gRNA Target sites for any set of arbitrary PAM sequences by adapting existing algorithms. Thus, experimenters can take into account the known flexibility of their chosen RNA-guided nucleases, or use the software without modification to predict binding sites of new RGNs. Because ADDTAG can query a genome with any number of Target motifs simultaneously, researchers can identify which RGNs would be most useful to edit the genome of their chosen biological system.

I developed ADDTAG with the goal of maximizing utility through computational flexibility without requiring continual user input. ADDTAG incorporates many of the previously described software features into a single tool: any genome can be used, and any arbitrary Target motif can be used.

## 3.4 Results

### 3.4.1 Identifying targets and Feature expansion

With 2-step genome editing facilitated by ADDTAG, users need not be concerned if their genomic Feature contains a suitable Target sequence for their experimental RGN. Users can simply define the bounds of the Feature they want altered, and provide the sequence to which they wish to change the Feature; and then ADDTAG will design all the dDNA, primer, and gRNA sequences necessary to achieve this goal. At the heart of this capability is a powerful Feature selection and expansion utility which expands a user-defined Feature to enable editing of a Feature, via 2-step editing, that cannot be efficiently edited via direct (1-step) editing (Figure 2.6). Even if the Feature does contain a quality RGN Target sequence, expansion may still be needed if the flanking homology regions contain allelic polymorphisms that surpass user-defined thresholds (Figure 3.5). To widen the bounds of a Feature, ADDTAG creates an expanded Feature (eFeature) by incorporating additional upstream and/or downstream sequences until the criteria for RGN Target quality and flanking homology are met (Figure 3.1). Then this expanded Feature is used to determine the necessary dDNA sequences for deletion and restoration, or modification, of the original user-defined Feature. The Feature expansion utility is highly configurable with options such as the maximum eFeature size, the directionality of expansion, and constraints preventing expansion into neighboring annotations.

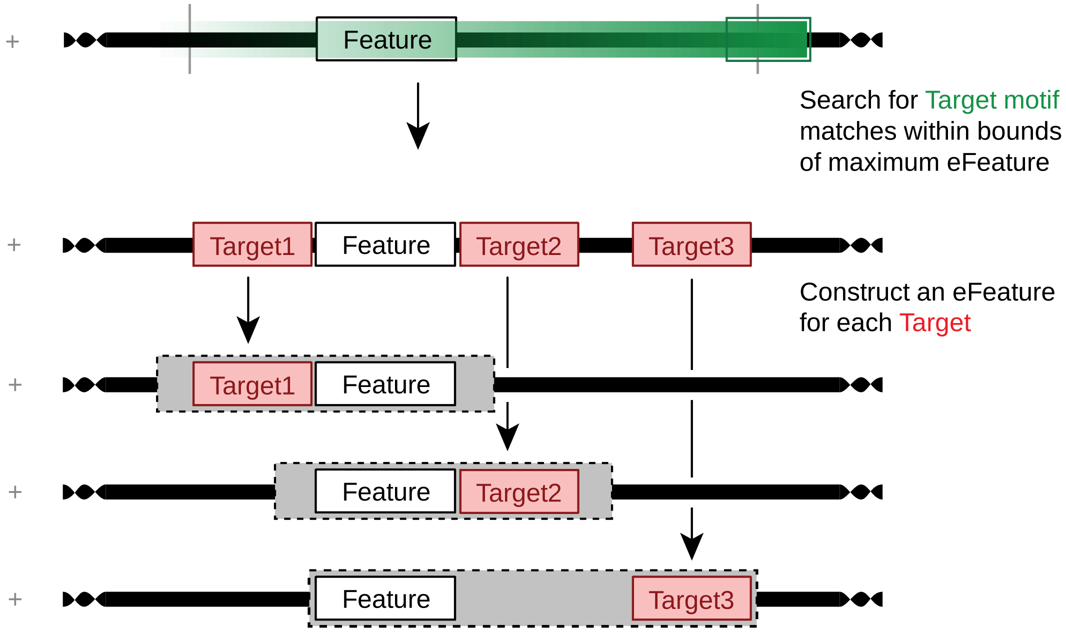

+
Feature
Search for Target motif matches within bounds of maximum eFeature
Target1
Feature
Target2
Target3
Construct an eFeature for each Target
Target1
Feature
Feature
Target2
Feature
Target3

**Figure 3.1 – ADDTAG evaluates many expanded Features for every user-defined Feature**

First, ADDTAG scans the Feature and neighboring DNA for matches to the Target motif. Each match is used as input, as well as the Feature itself, to construct expanded Features (eFeatures), respecting allelic specificity (Figure 3.2, Figure 3.3, Figure 3.5) and size (Figure 3.6) and format (Figure 3.4) restrictions. More detail can be found in **Common Figure Description 1**, Chapter 2.

ADDTAG's objective is to identify gDNA targets within or near to specific genomic Features that can be cut with a given RGN, and then produce dDNAs that will replace the Features. If a Feature does not contain a Target, or the user wishes to attempt to find a higher scoring one, the user can direct ADDTAG to expand the Feature until it finds a suitable Step 1 Target (Figure 2.7). The bounds of the Feature will be spread up to a defined number of nucleotides in both the up- and down-steam directions. All Step 1 Targets identified within each eFeature derived from the input Feature will be scored. This will generate a 2-step genome editing design in which the first-round Step 1 dDNA excises more than just the input Feature from the genome, and the Step 2 dDNA re-introduces the extemporaneously-subtracted DNA, along with any intended Insert or modification.

ADDTAG allows the user to specify if the Step 1 Target identified should exist on all homologs (multi-allelic), a single homolog (allele-specific), or any number of homologs (allele-agnostic). Multi-allelic Targets comprise of invariant sites within the Feature, allele-specific Targets match allelic variants, and allele-agnostic Targets do not check Feature homology. Polymorphism-aware identification of Targets and dDNA flanking homology regions is a five-step process (Figure 3.2, Figure 3.3, Figure 3.4, Figure 3.5, Figure 3.6). Here I present an example with three homologous instances of a Feature, designated (A), (B), and (C), which are shown in a vertical alignment. First, all Targets matching the user-provided Target motif are identified across each homologous Feature within the maximum bounds of an expanded Feature, designated by the vertical dashed lines (Figure 3.2). Within the full bounds of the potentially expanded Feature, each allele has four Targets.

Within the user-specified Feature, there is one multi-allelic set of Targets (Target A2, Target B2, and Target C2).

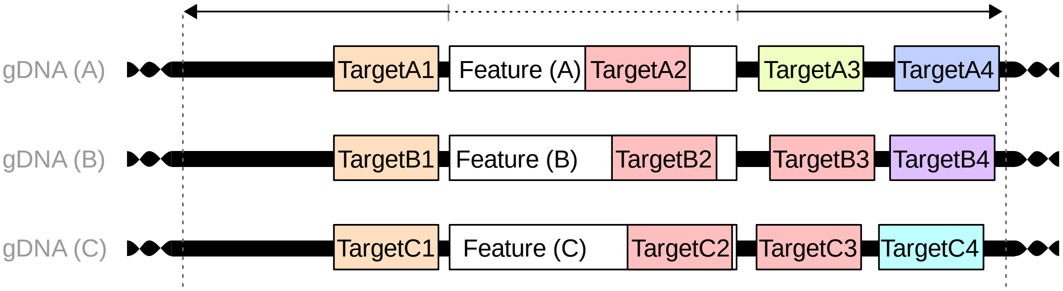


**Figure 3.2 – Identify all Targets within a distance from the Feature on all homologs**

Each homologous Feature, and its surrounding sequence, are scanned for matches to the user-defined Target motif (3.4.2). Targets of the same color have equivalent sequence identity. Different colored Targets have divergent sequences. More detail can be found in **Common Figure Description 1**, Chapter 2.

An equivalence group is a set of Targets with one site per homologous locus. Each equivalence group can be considered multi-allelic, allele-specific, or neither (Figure 3.3). Based on sequence identity, Targets are grouped into either multi-allelic or allele-specific equivalence groups.

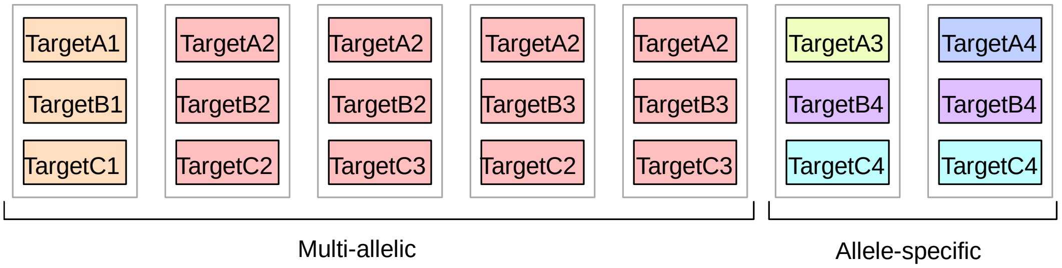


**Figure 3.3 – Identify Targets with desired allelic specificity**

In this example, there are five multi-allelic equivalence groups. Because the red Target has one copy in gDNA (A), but two copies each in gDNAs (B) and (C), there are four Target equivalence groups containing red Targets. This is a number equal to the permutations with replacement. A Target whose color is shared between at least two gDNAs precludes it from being allele specific. Thus, there are two allele-specific equivalence groups.

In order to facilitate diagnostic vPCR amplification, the size of the expanded Feature must exceed a minimal size. For each Target equivalence group, if necessary, the Feature is expanded according to the user-selected expansion method (Figure 3.4). AddTag provides 5 expansion methods, with “center_both” as the default.

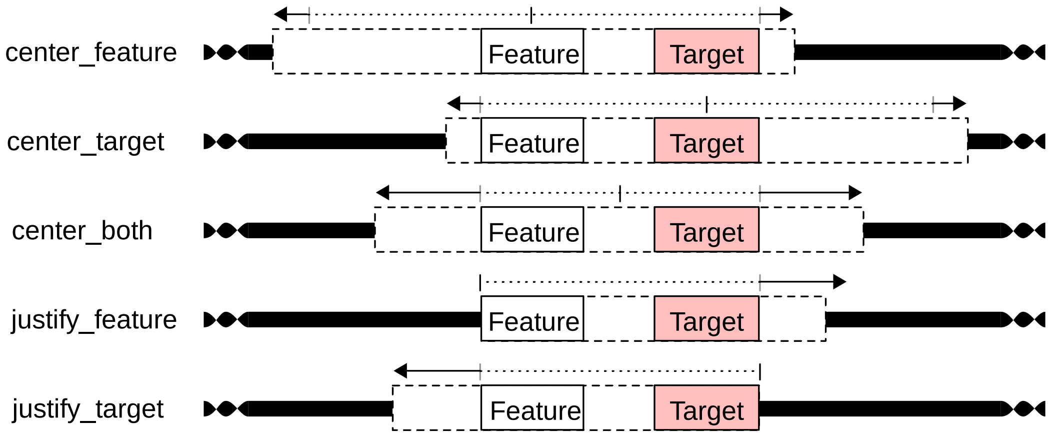


**Figure 3.4 – Feature expansion formats**

The black vertical line above each annotation represents the origin of expansion. The grey vertical line represents the minimum edge of the eFeature. The horizontal arrows are dashed where they cover the minimum eFeature size, and are solid where they represent further potential expansion. The methods "center_feature", "center_target", and "center_both" expand the eFeature in both upstream and downstream directions. The "justify_feature" and "justify_target" methods only expand in one direction. More detail can be found in **Common Figure Description 1**, Chapter 2.

ADDTAG allows the user to preferentially design homology arms and Targets against polymorphic sites. ADDTAG also supports editing multiple Features with the same dDNA when the user inputs a list of homologous loci. ADDTAG will then determine if the putative HAs, adjacent to each homologous Feature, contain variants when aligned with MAFFT [29]. The user can direct ADDTAG to use this information in one of three ways:

(1) ADDTAG can require each homologous locus to have a distinct nucleotide sequence, thereby creating allele-specific dDNAs;

(2) ADDTAG can require all homologous loci to have identical (or a maximal level of variants within the) flanking HAs, thereby creating multi-allelic dDNAs.

(3) Finally, ADDTAG can forgo any alignment of flanking regions, thereby creating allele-agnostic dDNAs.

If the homology arms of putative dDNAs would have undesired polymorphism levels, then the Feature can be expanded further to delimit new flanking regions (Figure 3.5). The expanded feature can take one of several formats (Figure 3.4). Therefore, Feature expansion directly informs the HAs of Step 1 and Step 2 dDNAs.

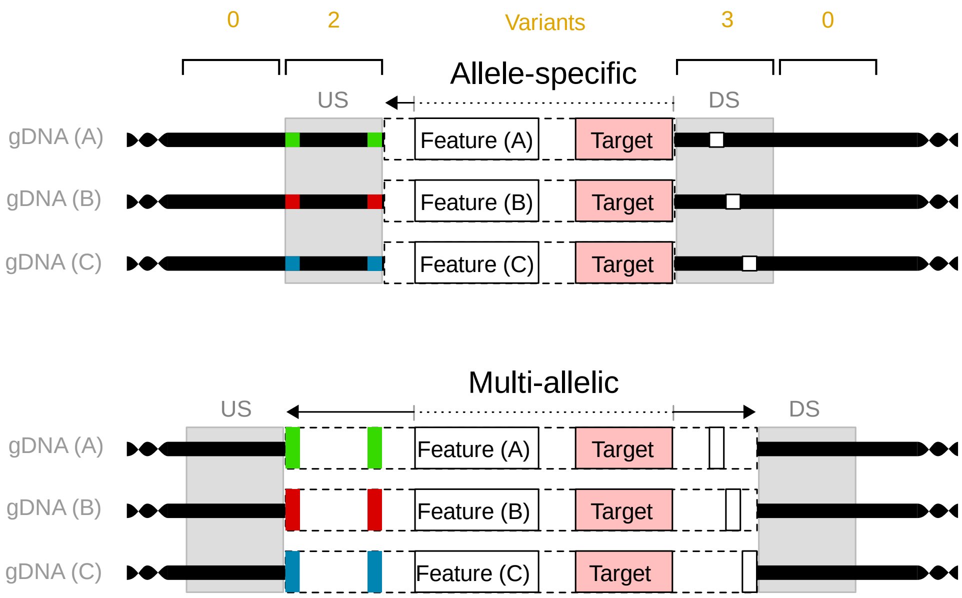


**Figure 3.5 – Expand Feature such that flanking homology regions meet allelic specificity**

Homology arm (HA) determination for allele-specific (top) and multi-allelic (bottom) dDNAs. In the top example, HAs must have a minimum of 2 variants, so the Features are expanded to the left until the US region overlaps with 2 variants. The DS region is not shifted right because it already contains 3 variants. In the bottom example, HAs must have a maximum of 0 variants. ADDTAG expands the Feature on the left and right until no variants exist within the HAs. Feature expansion is triggered when the number of variants in a HA does not meet the requirement. The solid red, green, and blue squares on the gDNAs represent a polymorphism at the same genomic position, such as an alignment mismatch. The white squares on the gDNA represent polymorphisms at different genomic positions, such as an alignment gap. The grey shaded regions on either side of the eFeature represent the flanking homology arms that will be present in the Step 1 dDNAs. More detail can be found in **Common Figure Description 1**, Chapter 2.

If the user desires allele-specific Step 1 dDNAs, ADDTAG will expand the Feature such that the flanking homology regions contain an adequate number of polymorphisms (Figure 3.5). In this example, the chromosome alignments outside the right side of each Feature contains an allele-specific gap. Homology arms (HAs) are the up- and down-stream regions flanking a Feature. For allele-specific HAs, these 3 variants exceed the minimum number of 2 variants required, so no expansion is necessary. For multi-allelic HAs, these 3 variants exceed the maximum number of 1, so expansion occurs to the right. Outside the left of the user-defined Feature, there are two allele-specific variants on each chromosome. Without expansion, the HA on the left side would contain 1 polymorphic site. Therefore, for allele-specific expansion requiring a minimum of 2 variants, the Feature is expanded to the left to encompass 2 sites that are polymorphic across all 3 chromosomes. If the user desires multi-allelic Step 1 dDNAs, then the left HA is expanded past both variant sites. In this example, both the left and right sides of the Feature are expanded so there are no polymorphisms in either flanking homology arm.

Because the vPCR amplification needs to be diagnostically useful (4.3.1), the amount of DNA replaced with an insert in Step 1 is limited to a fixed range. The final step ADDTAG performs in eFeature creation is filtering potential sets based on their lengths (Figure 3.6). In this example, the homologous expanded Features with red multi-allelic Targets all fall within the accepted size range, so it is included in downstream analyses. The putative set of expanded Features with orange Targets has at least one expanded Feature that violates the size thresholds, so this set will be excluded from downstream analyses.

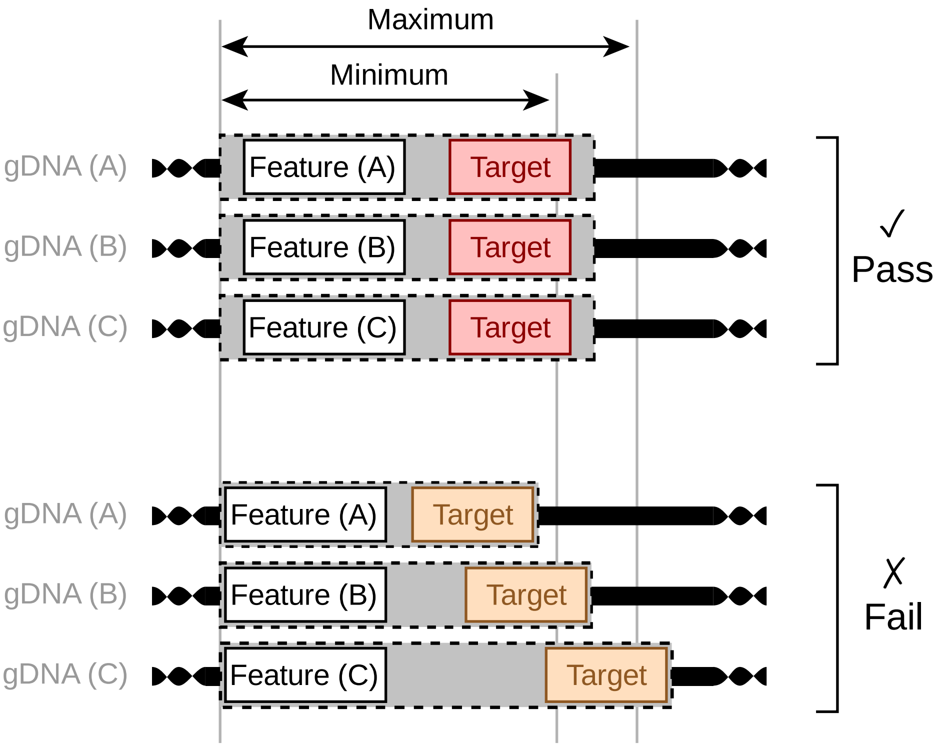


**Figure 3.6 – Enforce size limits on expanded Feature**

ADDTAG enforces size limits on expanded Features (eFeatures). Homologs with eFeatures smaller than the minimum, or larger than the maximum, are discarded. More detail can be found in **Common Figure Description 1**, Chapter 2.

After Feature expansion produces a set of eFeatures, with each linked to a set of equivalent Targets, each candidate Target must then pass through a pre-alignment filter (prefilter) before its Algorithm scores are computed (3.4.8).

### 3.4.2 Target identification

ADDTAG identifies genome positions that can be restricted by RGNs by using a Target motif. The Target motif is a string of characters that models the template sequence on the genome, and not the gRNA sequence that the RGN uses. The Target motif thus describes the effective spacer, homologous to the genomic DNA when experimental conditions are met (e.g. temperature and pH), and not the full-length gRNA sequence that may include complex secondary structures where only a portion of the sequence is homologous to the genome [30]. Candidate Target sites are identified within Features using regular expression-like syntax. The user must include one or more Target motifs specific to the RGN proteins being used. This option takes a string of characters as input, written in the 5'→3' direction. The greater than “>” and less than “<” characters point toward the PAM, and delimit the PAM from the region that corresponds to the Spacer. Thus, Targets with 5’-adjacent PAM sequences like Cas12a would encode this information with the “<” character, and Targets with 3’-adjacent PAM sequences like Cas9 would use the “>” character. The vertical bar “|” represents a double-strand cut. A slash “/” represents a forward strand (sense) cut. Backslashes “\” represent reverse strand (antisense) cuts. Full nucleotide ambiguity codes specified by the IUBMB/IUPAC are supported [31]. Open “{“ and close “}” braces surrounding a number or a comma-separated pair of numbers represent quantifiers. A period “.” represents a base used for positional information, but not enzymatic recognition. ADDTAG affords two options that use this syntax: motifs to design gRNA targets against, and motifs to include for off-target calculations only. Any number of motifs can be specified for each.

For example, a typical genome editing experiment using Cas9 (derived from *Streptococcus pyogenes*) would use the Target motif `'N{17}|N{3}>NGG'`. ADDTAG has the flexibility to deal with the PAM upstream of the spacer, such as the Cas12a (derived from *Acidaminococcus/Lachnospiraceae* species) [32], which uses the canonical Target motif `'TTTN<N{19}/.{4}\'`.

Native microbial gRNA components typically encode for spacer sequences >30 nt in length [33, 34]. However, early experiments found that shortening the spacer length can increase specificity without severely impacting efficiency. Precedent has set 20 nt length spacer as the standard, but shorter 17-19 nt lengths can be used just as effectively [35]. For ease of use, ADDTAG includes a list of the most commonly used Target motifs identified across the entire family of Cas RGN molecules (such as the one provided in [36], but with a deeper sampling of the literature). ADDTAG thus has the most extensive compendium of all identified RGN Target motifs [37-39], which is accessible on the command line. It describes the characteristics for each RGN that have been published, including the highest-efficiency on-target Target motif, the lengths of the spacers tested, the empirical biological system, and any restriction positions determined. ADDTAG establishes a unified framework for working with both 3’-adjacent and 5’-adjacent PAM sequences, and both blunt and staggered cuts, such as with Cas9 and Cas12a.

### 3.4.3 Donor DNA generation

A critical element of successful RGN-mediated genome editing is an effectively designed dDNA sequence. ADDTAG automatically designs dDNA sequences for each genome editing step. Each dDNA sequence has three elements in its basic structure: a region of homology to the gDNA upstream (US) of the Feature, the insert (e.g. `addtag`), and a region of homology to the gDNA downstream (DS) of the Feature.

The ADDTAG software gives several options for inserts in the Step 1 dDNA. ADDTAG can construct Step 1 dDNAs with unique (`addtag`) or identical (`unitag`), full-length Step 2 Targets, so experimental loci can be edited in isolation or in parallel. ADDTAG can construct Step 1 dDNAs with minimal extrinsic DNA (`mintag`) while also ensuring a high-quality Step 2 Target for efficient second step RGN restriction. ADDTAG also supports typical (1-step) genome editing using any user-defined dDNA insert.

Step 2 dDNA construction relies upon identifying the AmpF/AmpR PCR primer pair that surrounds the Feature on the reference genome. This primer pair can be used to either amplify the reference genome template to produce wild-type, add-back, Step 2 dDNA (for instance, the *ADE2*$_{CDS}$, *EFG1*$_{CDS}$, and *BRG1*$_{CDS}$ loci described in 2.4.5 and 2.4.3), or to amplify custom, synthetic DNA fragments for edited Features (for instance, the *ZAP1*$_{US}$, *ZRT2*$_{US}$, *WOR1*$_{USp}$, *WOR1*$_{USd}$, and *WOR2*$_{DS}$ loci described in 2.4.2 and 2.4.1).

Users can specify whether dDNA homology arms should avoid or require allelic variation (polymorphisms) among input homologous Features (3.4.1). For allele-specific dDNAs, ADDTAG finds HAs that are unique for each homologous Feature. Primer amplicons will either be diagnostically-different sizes, or primer sequences themselves will be different. For multi-allelic dDNAs, ADDTAG will minimize the number of variants in HAs. This maximizes the sequence identity of DNA adjacent to all homologous Features. For allele-agnostic dDNA, HAs and primer pairs are calculated for each Feature independently. If the dDNA design for a given Feature fails to pass these criteria, then the Feature will be expanded as described above.

Following dDNA generation, users can direct ADDTAG to perform *in silico* recombination for determining if dDNA homology arms possess significant similarity to non-target locations in the gDNA (4.4.1).

### 3.4.4 Generating Step 1 dDNA inserts that encode Step 2 Targets

The ADDTAG software generates Step 1 dDNA sequences with inserts that facilitate several genetic and molecular biology techniques. The general label for these inserts is `addtag`. ADDTAG is able to use any arbitrary user-defined sequence as the insert, as well as create several types of specific `addtag`s as follows in this subsection. ADDTAG first generates a number of potential dDNAs, and subsequently only keeps dDNAs that encode for Step 2 Targets (Figure 3.13).

What separates ADDTAG from other programs that identify RGN binding sites is its ability to create unique gRNA Targets at the site of cleavage. If `mintag` (also called "mAT" or "mini-add-tag" [40]) is selected as the insert type for the first round of editing, then ADDTAG generates the Step 1 dDNA by stitching the immediately-adjacent upstream and downstream regions of each Feature or eFeature together into one concatemer (Figure 3.7). If the junction sequence is not unique in the genome, then it uses a combination of 3 additional adjustments to generate a unique site:

(1) additional bases upstream of the Feature can be trimmed (thereby effectively expanding the Feature),

(2) bases can be added,

(3) and additional bases downstream of the Feature can be trimmed (also expanding the Feature).

Users can use a command line argument to specify each of these as well as the final fragment size of generated Step 1 dDNAs. Default `mintag` implementation in ADDTAG uses "brute force" calculation of all k-mers possible for the insert if the query insert size (k) is less than 5 nt. Otherwise, it samples k-mers uniformly to obtain putative sequences for constituting a Target site.

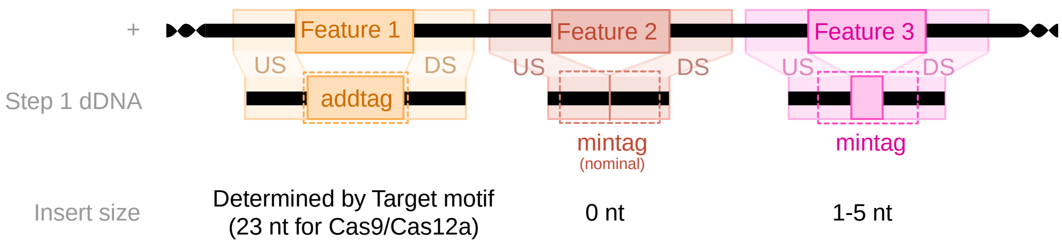


**Figure 3.7 – Types of locus-specific insert sequences generated for AddTag Step 1 dDNA**

The full-length `addtag` and small `mintag` inserts are unique to each locus. Step 2 Target sequences are indicated by rectangles with dashed borders. More detail can be found in **Common Figure Description 1**, Chapter 2.

AddTag derives its name from the `addtag` insert type (also called "AT" or "add-tag") for the first round of editing. The Step 1 dDNA is a concatenation of the upstream flanking region, an explicit Target sequence that matches the Target motif, and the downstream flanking region (Figure 3.7). As I describe next, ADDTAG will procedurally generate this insert so its nucleotide composition differs as much as possible (within a stochastic sampling distribution) from the rest of the genome, thereby minimizing the likelihood the generated Target sequence exists elsewhere, and maximizing the off-target score. This is especially useful when the input genome sequence represents only a portion of the true DNA within the biological system.

First the unstranded tetranucleotide distribution of the reference genome is computed. A stranded tetranucleotide distribution has 256 elements, but the unstranded one has 136 because 240 tetranucleotides reduce to 120 when they are merged with their reverse complement (TGGT = rc(ACCA)), and there are 16 palindromic tetranucleotides (AATT = rc(AATT)). The count distribution is normalized. Next, random sequences matching each Target motif are generated. ADDTAG then calculates the likelihood of observing the sequence, given the genomic tetranucleotide distribution, and selects the sequences with the minimum likelihood. ADDTAG uses a "batches of batches" approach, where the minimum likelihood of 100 random sequences is selected, and up to 10,000 selected sequences are generated and used as potential dDNA inserts.

The `unitag` insert type is used for generating dDNAs that contain a single instance of a Target motif for all edited loci (Figure 3.8). Because the same insert sequence is added to every dDNA, each Feature edited in the intermediate genome (ΔgDNA) will contain identical `unitag` sequences. This allows a single gRNA to target every locus in subsequent genome editing steps. Like the `addtag` insert, the `unitag` is a random sequence that is generated from the complement composition of the genome, thereby increasing the probability of specific RGN activity, even in the absence of a complete genome sequence as input.

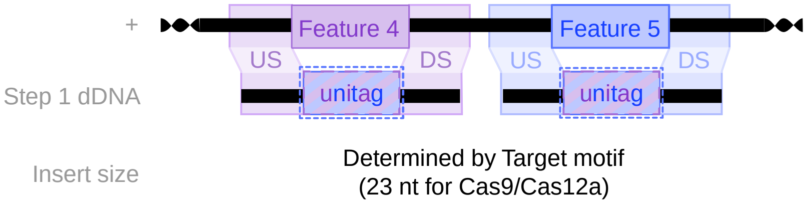


**Figure 3.8 – Types of locus-shared insert sequences generated for AddTag Step 1 dDNAs**

The `unitag` insert is shared among loci. Each locus is given a different color (purple, blue), but because the `unitag` sequence is inserted into both loci, it is colored both purple and blue. The Target the Step 1 dDNA insert encodes for is indicated by a rectangle with a dashed border. More detail can be found in **Common Figure Description 1**, Chapter 2.

The `bartag` inserts are short sequences that are unique across engineered strains. If the user wishes to add a 'bartag' at the Feature, then first round dDNAs will contain a sequence derived from the user-defined Bartag motif (similar syntax to the Target motif). The Bartag motif is text describing the nucleotide ambiguity allowed within the `bartag` sequences. Each `bartag` sequence ADDTAG generates is guaranteed a user-configurable minimum edit distance from all other `bartag` sequences. Typically, they are used for assigning molecular barcodes to various microbial species that are grown as a community. As an end-point measurement for the experiment, amplicon sequencing is performed on the `bartag`, thereby revealing the relative frequencies of all strains in the community.

### 3.4.5 Generating Step 2 dDNA with edited Features

If Feature expansion is used, each Feature can have one or more eFeatures. This shifts the start and end positions on the chromosomes of the regions to be removed during Step 1 editing. Therefore, each Step 2 dDNA can have a different US HA and a different DS HA. The US HA begins with the annealing site of AmpF, and the DS HA ends with the AmpR annealing site on the reverse strand. For each eFeature, ADDTAG identifies a collection of AmpF/AmpR primer pairs, and ranks them with weight calculations (4.4.5). Following HA identification, the insert for Step 2 dDNA is calculated. The US region trimmed in excess of the user-specified Feature is concatenated with either the reference Feature, or any user-specified ✳Feature, and the DS region trimmed in excess of the Feature. Finally, the US HA, insert, and DS HA are concatenated together. ADDTAG assigns a rank to the Step 2 dDNA using the AmpF/AmpR primer pair weight. Because each input Feature can result in numerous potential Step 2 dDNAs due to predicted amplification efficiencies, ADDTAG encourages the user to use the dDNA with the highest weight.

### 3.4.6 ADDTAG Aligner interface

ADDTAG relies on the ability to align short RNA sequences to whole genomes. Several short-read aligners have been repurposed for use with aligning Spacer and PAM sequences to genomes [41] such as BOWTIE [42] in CRISPOR [11] and CHOPCHOP [9, 10]. In order to facilitate comparisons with other software, and to simplify updating ADDTAG with more sophisticated aligners in the future, we present a unified, general Aligner interface.

The Aligner class has three primary functions: indexing, aligning, and parsing. Indexing and aligning conceptually mirror those same methodologies used by short read aligners. The index function serves as a staging ground for alignment, where scoring matrices are loaded and genome hash indexes are calculated. Following indexing, the align function identifies the genomic coordinates that with similarity to the query RGN Target. ADDTAG finally parses the alignment output and converts it into an internally-usable data structure. ADDTAG supports the general alignment format SAM [43] and several other program-specific formats. ADDTAG includes wrappers for the following programs, with all parameters preset to appropriate values: NCBI BLAST+ [44], BOWTIE 2 [45], BWA [46], and CAS-OFFINDER [47]. For the exact shell commands, please refer to either the ADDTAG source code or an output log file after running ADDTAG.

### 3.4.7 ADDTAG Algorithm interface

Because ADDTAG uses a variety of scoring algorithms (`Algorithm` in the source code) for evaluating Target suitability, it implements a flexible computational interface for dealing with arbitrary requirements. We distinguish the Algorithm type by whether it requires 1 or 2 sequences to be input (`SingleSequenceAlgorithm` and `PairedSequenceAlgorithm` in the source code). 1-sequence Algorithms compare the candidate Target sequence to a model trained on empirical gRNA experiments. 2-sequence algorithms directly compare a gRNA Spacer sequence to a Target sequence, or compare two competing Target sequences.

Any 1-sequence Algorithm can be used to calculate an on-target score. Additionally, several authors claim their Algorithms are appropriate to use for off-target scoring as well. Any 2-sequence Algorithm may be used to calculate off-target scores. After candidate Targets are aligned, each selected Algorithm is used to calculate a score. 1-sequence Algorithms use the match sequence as input. 2-sequence algorithms use the query and the match sequences as input. In addition, new scoring algorithms can be implemented by creating an `Algorithm` subclass in the `source/algorithms` subdirectory that utilizes the "universal" Algorithm interface developed for ADDTAG. The minimal interface is defined as the tuple (sequence, side, target, pam, upstream, downstream), with any number of additional, optional parameters.

Nearly all published implementations of CRISPR/Cas scoring algorithms lack flexibility in varying spacer lengths and alternative PAM sequences. We modified the scoring algorithms presented in these papers to allow for assessment of spacers less than or greater than the typical 20 nt length. Additionally, they have been expanded to include scoring of ambiguous characters (using the unweighted average score, a subsampled average, and sometimes the maximum score). For full information on how each Algorithm was adjusted, please refer to the source code in the `source/algorithms` subdirectory.

### 3.4.8 Target filtering

In order to score how efficient a spacer sequence is at directing gDNA cutting, Targets identified within Features or Inserts are filtered in two steps. First, the prefilter checks the quality of candidate Target sequences before aligning to the genome and potential dDNA sequences. This reduces the total number of Targets that need to be aligned to the genome and then evaluated by scoring Algorithms. After aligning, the Targets are subjected to a postfilter that takes individual Algorithm scores into account. Only Targets that pass both the prefilter and postfilter are reported to the user.

ADDTAG implements the following optional prefilters:

- Upper/lower case masking (ignore, upper-only, lower-only, mixed-lower, mixed-upper, mixed-only). Users can apply several masks to the genome, which can prevent selecting Targets from these masked regions.

- Process ambiguous characters (discard, keep, disambiguate, exclusive). Any ambiguous characters can be equivalently masked. If the user intends to target a region containing ambiguous characters, then ADDTAG will optionally disambiguate potential spacers.

- Target motif sanity check. Some potential gRNAs derived from ambiguous character expansion may violate the initial motif. This filter ensures that none of these enter into downstream calculations.

- Maximum consecutive T residues. Sequences containing consecutive T residues may cause polymerase termination [48].

- Upper/lower %GC content thresholds. The GC content of Cas targets may affects binding specificity [49].

- Proximal G, which evaluates if the single nucleotide of the spacer adjacent to the PAM is a guanine.

Users can create any number of additional prefilters by subclassing `Algorithm` in the `source/algorithms` subdirectory, and setting its prefilter attribute to `True`.

Then, after aligning Target queries to the gDNA and dDNA sequences and scoring them according to the user-selected Algorithms, the postfilter removes poor quality Targets. After alignment, each alignment match is scored by all selected Algorithms. Then any off-target Algorithm scores are calculated. Each Algorithm is given a minimum and maximum cutoff value. If the Algorithm score for that match is outside these bounds, then it fails the postfilter. For each Algorithm designated as a postfilter, if the match passes the postfilter criteria, then it is included as a potential on/off-target. By default, PAM-identity and the number of substitutions, insertions, deletions, and errors [50] are included in the postfilter.

### 3.4.9 On-target calculations

ADDTAG evaluates the predicted joint binding and cutting efficiency of a gRNA:RGN complex for the Target sequences it identifies. These are “on-target” scores, and they rely on analyzing the Target site with a model. ADDTAG implements these as 1-sequence Algorithms. There are 93 known Cas protein families, spread across 394 PSSMs, 2 classes, 6 types, and more than 16 subtypes [51-54]. ADDTAG implements the field-standard scoring schemes for Cas9 and Cas12a (also called Cpf1). Several scoring algorithms are provided for use with uncharacterized Cas proteins, such as the “linear” score. However, these are founded on unsophisticated assumptions, such as SPACER to target homology lengths and positions of errors within the alignment relative to the PAM site.

By default, ADDTAG uses the AZIMUTH on-target score for use with Cas9 motifs [55]. The AZIMUTH algorithm takes as input a genomic Target (including the spacer and PAM sequence plus a few nucleotides up- and down-stream of it), and compares it to a gradient-boosted regression trees model trained on cutting efficiency of over 4000 individual sgRNAs targeting sites in 17 genes in human A375 cells. Cas9 cutting efficiency in yeast [40] as well as mouse, worm, and fly cells [11] mirror these predicted values. For Cas12a motifs, ADDTAG includes the CINDEL/DeepCpf1 [56, 57] Algorithm, which predicts the likelihood of getting a Cas12a-induced in/del at the target locus [56, 57]. Thus, it serves as a decent proxy for an on-target score. CINDEL is a logistic regression classifier trained on 938 Spacer-Target pairs in HEK293T cells. Additional on-target scores implemented are CRISPRater [58], Doench-2014 [21], Housden [59], and Moreno-Mateos [18].

The structure of the template DNA, such as its chromatin packing, can influence an RGN's ability to bind and cleave [60-66]. Therefore, the on-target score of a single locus can differ greatly between cells that are transcriptionally active at that locus, and cells that are transcriptionally quiescent. When calculating Algorithm scores, AddTag specifically ignores chromatin-based DNA accessibility. However, if users have nuclease, transposase, or nucleosome footprint information, they can mask nuclease-inaccessible regions of the genome, and then they can have AddTag use Feature expansion to obtain Targets in regions known to be accessible.

### 3.4.10 Off-target calculations

The "off-target" score represents the predicted fraction of events that the gRNA:RGN complex will associate with the intended Feature (region of DNA to be targeted) and restrict it compared to all sites with similar homology to the gRNA spacer. Spacers with high off-target scores are preferred, and indicate that unwanted restriction events are unlikely to occur.

Each Target motif $T$ is composed of the spacer sequence (SPACER), the restriction sites (✂), the PAM, and the relative position of the PAM to the spacer sequence ($>$ or $<$), defined by the following syntax:

$$T = \{\text{SPACER}, \gtrless, \text{PAM}, \text{✂}\}.$$

The Target motif is encoded by the user as a string of characters (3.4.2).

The off-target calculation we use is the general-purpose MIT Guide Score [67] $S$ (Equation 1), which can be applied to the results from any number of scoring Algorithms. We calculate it through the following steps.

First, align the motif $T$ to the Feature (or expanded Feature), which is the region to be disrupted, cut, or edited. Motif alignment, $align$, occurs through a regular expression text search using the Regex Python 3 package (Figure 3.12). Each substring in the Feature matching the motif is considered a query $q$, and the set of all matches for the Target motif is $Q$ such that $q \in Q$. Thus

$$Q_T = align(T, \text{Feature}).$$

Next, each identified query $q$ is aligned exhaustively across the gDNA and dDNA expected to be present using the user-specified Aligner, $align$ (3.4.6). Each substring in the genome with homology to the query that passes the postfilter (3.4.8) is called a match $m$, and these are put into two categories. Those that lie within the Feature are on-target matches (denoted by superscript $on$), and those that lie outside the Feature are off-target matches (denoted by superscript $off$). Thus, each query has a set of within-Feature and outside-Feature matches. Alignment links each query $q$ with a set of matches $M$

$$M_q = align(q, \text{gDNA} \cup \text{dDNA}).$$

Furthermore, each match is classified mutually exclusively as either within-Feature or outside Feature such that

$$M_q = M_q^{on} \cup M_q^{off},$$

where the set of within-Feature matches is $m \in M_q^{\mathrm{on}}$ and the set of outside-Feature matches is $m \in M_q^{\mathrm{off}}$.

Each algorithm $a$ is contained within the set of all algorithms $A$ chosen to evaluate the Targets: $a \in A$; and each algorithm can take either 1 or 2 sequences as input. The algorithm assigns each match a score $s$ such that

$$s_a(q,m) = \begin{cases} a(m|T) & \text{1-sequence Algorithm} \\ a(q,m|T) & \text{2-sequence Algorithm} \end{cases}.$$

The final off-target score $S$ for any particular query and algorithm pair is calculated as

$$S_a(q) = \frac{\sum_{m \in M_q^{on}} s_a(q,m)}{\sum_{m \in M_q^{on}} s_a(q,m) + \sum_{m \in M_q^{off}} s_a(q,m)}. \qquad \textbf{Equation 1 – MIT Guide Score}$$

For each off-target compatible algorithm, this procedure is followed to calculate an off-target score.

The final off-target score $S$ for any particular algorithm is thus a ratio between the sum of targeting efficiency for acceptable sites and the sum of targeting efficiency across all sites. A higher score means more of the predicted targeting is at the intended Feature. A lower score means more predicted targeting outside of the Feature. In other words, the off-target score $S$ represents the frequency a gRNA:RGN complex will target a correct site in the genome. AddTag reports final off-target scores as percentages on a scale from 0 to 100 ($100 \cdot S$).

AddTag uses the MIT Guide Score [67] due to its wide adoption rather than the alternative Stemmer [7] method for calculating off-target scores. The Stemmer off-target score scales relatively to the number of matches, and thus requires additional computations to compare across experiments. The MIT Guide Score, contrarily, returns a value constrained by probability, so scores can be compared across motifs, Features, and genomes. Of note is that the off-target specificity is dependent on the number of errors (mismatches, inserts, deletions) permitted by the Aligner used. In the genome editing experiments presented in this study, we assume that the Aligner finds all relevant matches. AddTag implements the following algorithms for off-target scoring: CFD [55], Hsu-Zhang [68], a simple linear model (this study), CRISPRater [58], Doench-2014 [21], Housden [59], and Moreno-Mateos [18].

### 3.4.11 Weighing Target sequences

A goal of the AddTag software is to output simple recommendations for Target and dDNA sequences. AddTag thus assigns ranking to Targets and dDNA in a method similar to how it assigns primer weights (4.4.4). This subsection describes how default Algorithm weights are calculated.

For simplicity, weights of on-target Algorithms are treated differently than weights of off-target Algorithms. On-target weights are sigmoid approximations of their cumulative distribution functions (CDFs). However, off-target weights are calculated based on *Candida albicans*' genome-specific off-target density distributions. We used sigmoidal functions to approximate the CDFs describing the possible scores for each Algorithm (Figure 3.9). The sigmoid function is useful because the shape of the curve is easily interpretable from a small number of parameters ($\theta$) to turn the score ($s$) into the weight ($w$):

$$\theta = \{x, slope\} \text{ and}$$

$$w(s|\theta) = \frac{1}{1+slope^{x-s}}. \quad \textbf{Equation 2 – Algorithm weight}$$

The cumulative distribution serves as a means for relative ranking of targets. Where a Target's score lies on the CDF indicates its quality. The distance between two targets (vertical axis) represents the proportion of all possible Targets that lie between them.

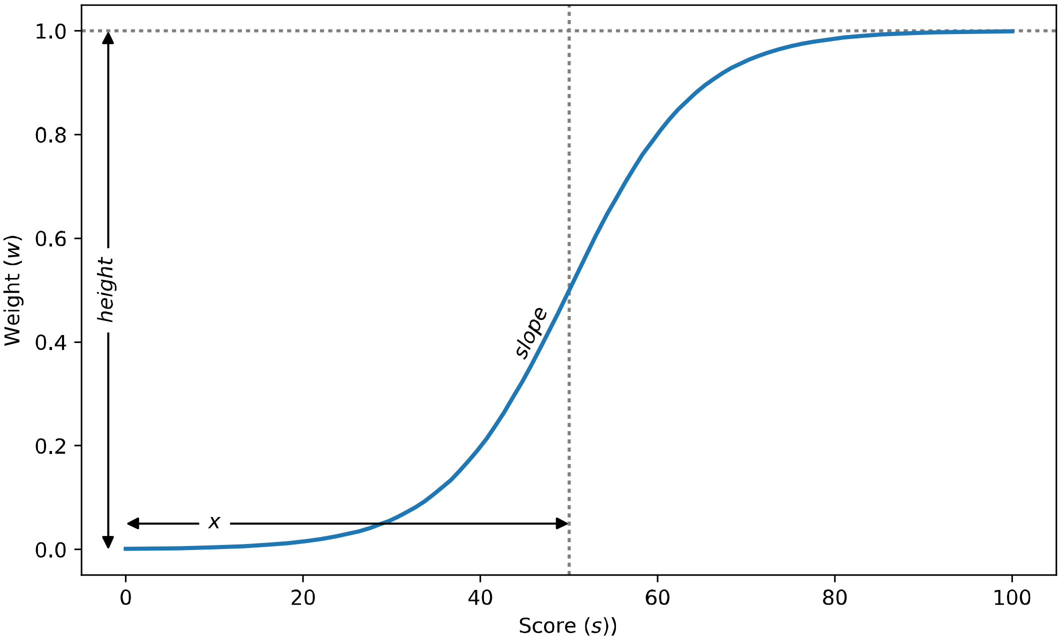


**Figure 3.9 – Description of unisigmoidal function parameters**

ADDTAG converts Algorithm scores (horizontal axis) to weights (vertical axis) using sigmoid approximations. Each sigmoid function is defined by an inflection point ($x$) and a steepness ($slope$). ADDTAG sets the final parameter $height = 1.0$ for simplicity. The $slope$ can be either positive (depicted) or negative (not depicted).

Like CRISPOR [11], ADDTAG calculates multiple types of Algorithm scores to evaluate how appropriate a Target is for genome editing. This raises the obstacle of effectively ranking Targets in a useful manner that takes the different scores into account. For our solution, we propose each Algorithm score is given a weight function that transforms the raw score into a weighted one. Thus each Target sequence ($seq$) will have a score for each algorithm ($a$). Then the final weight ($W$) is the product of all weighted scores ($w$) across the set of selected Algorithms ($a \in A$)

$$W(seq) = \prod_{a \in A} w_a(s_a(seq)|\theta_a).$$ **Equation 3 – Target weight**

By default, ADDTAG utilizes sigmoidal weight functions that estimate the cumulative distribution of scores from random Target sequences that match the Target motif. The parameters of the sigmoidal function utilized are the following: the magnitude of change, or height (importance of the score); the slope of the change (how quickly the weight changes based on the score); and the center of the point of inflection (where the threshold of importance is) (Figure 4.8). If needed, the user can apply a different weight function or altered weight parameters to each scoring Algorithm using command line parameters. This provides an easily tunable mechanism to specify which scores are most important for any CRISPR/Cas application. For example, the AZIMUTH [55] Algorithm returns a score in the inexact domain from 0 to 90, and its corresponding weight function converts the scores to weights on the range from 0 to 1 (Figure 3.10). The *Candida albicans*-specific off-target Algorithm weight for HSU-ZHANG [68] scores reveal that most uniformly-sampled 23 nt sequences within the genome yield high scores (Figure 3.11).

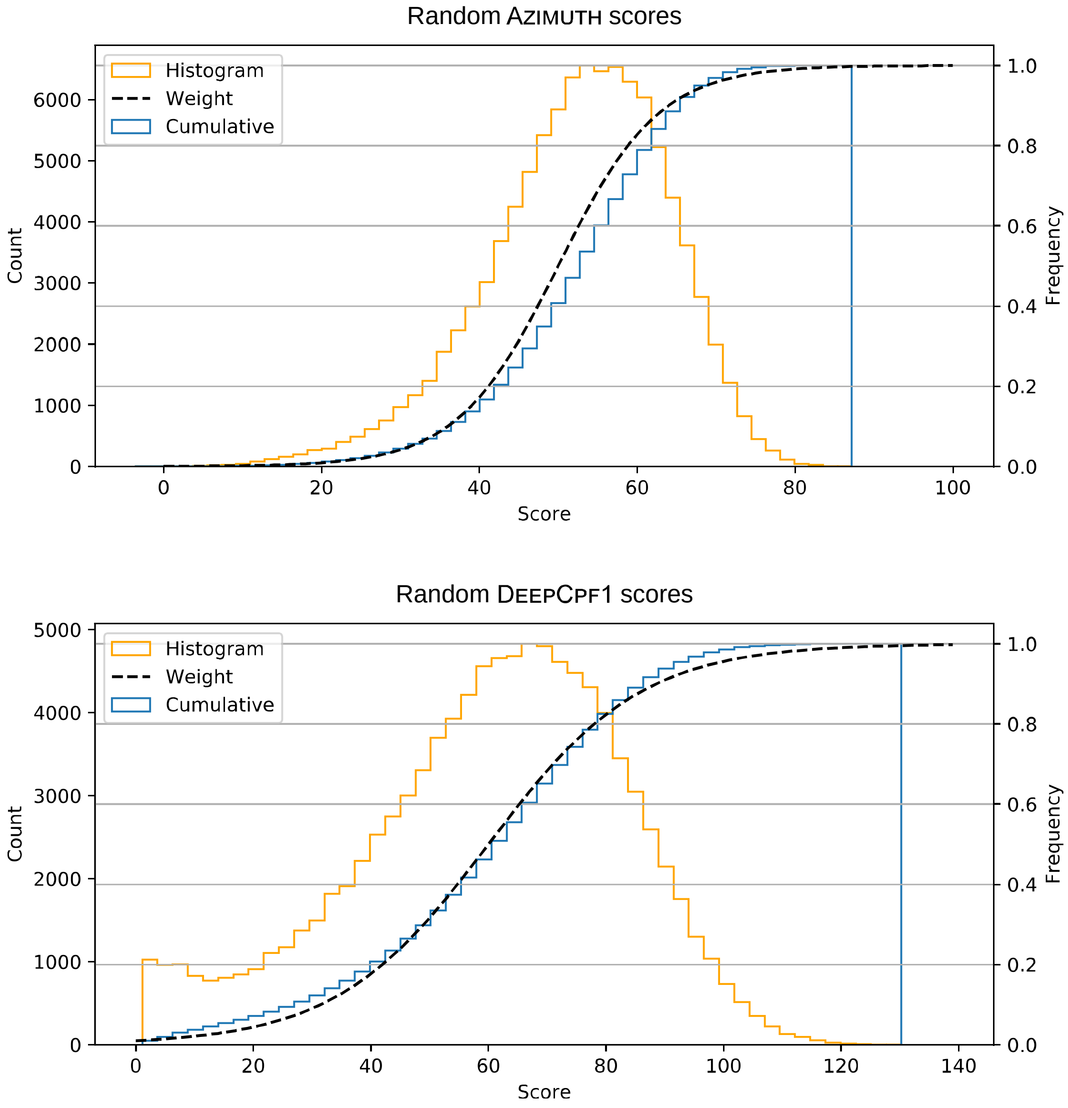


**Figure 3.10 – *A priori* weight functions for on-target algorithm scores**

By default, AddTag encodes weight functions for on-target Algorithm scores that approximate their cumulative distribution functions (CDFs). The graphs show the Azimuth [55] and CINDEL/DeepCpf1 [56, 57] scores for 100,000 random sequences (orange), used for evaluating Cas9 and Cas12a respectively. The CDF histogram frequency (blue) on right vertical axis also corresponds with the weight value (black). Users can adjust the weight functions for each Algorithm with command line parameters.

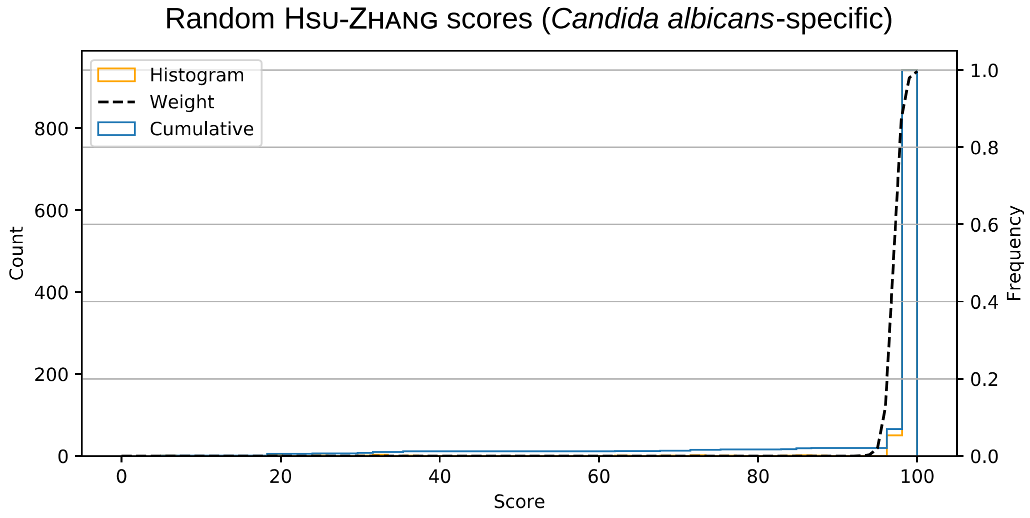


**Figure 3.11 - *A priori* weight functions for off-target algorithm scores**

By default, AddTag encodes weight functions for off-target Algorithm scores that approximate their cumulative distribution functions (CDFs). The graph shows the HSU-ZHANG [68] scores for 1,000 random sequences drawn from the *Candida albicans* genome (orange), used for evaluating Cas9. The CDF histogram frequency (blue) on right vertical axis also corresponds with the weight value (black). Users can adjust the weight functions for each Algorithm with command line parameters.

For each candidate Target, the weights of all selected Algorithms are multiplied together, which gives the final, reported product weight (Equation 3). All candidate Targets are then re-arranged in decreasing order from the highest weight to the lowest weight.

Because the sigmoidal weight function approximates the cumulative distribution of scores, the transformed weight represents the percentile of the input score. Thus, the raw score of each Algorithm is converted into a weight that represents how good that score is compared to all other possible scores. This means weights of different Algorithms are comparable, and weights of different Target sequences are comparable as well. Because the on-target distributions were generated with random sequences using the uniform distribution, on-target weights are comparable across genomes. However, with the current implementation, off-target weights cannot be used to compare Targets on different genomes. At the time of publication, each on-target Algorithm implemented displays a unimodal score distribution (Figure 3.10, orange "Histogram" line), which is required for the sigmoidal calculation to approximate the CDF. If the score distribution was multimodal, one additional sigmoidal product should be added for each mode. Thus, the sigmoidal transformation of the Algorithm scores is sufficient for Target ranking.

The Target weight (Equation 3) is an end-point value intended to make Target selection more intuitive for ADDTAG users. ADDTAG calculates Target weight for both the candidate Step 1 Targets (Figure 3.12) and the candidate Step 2 Targets (Figure 3.13). Targets with higher weights are predicted to have more success with genome editing.

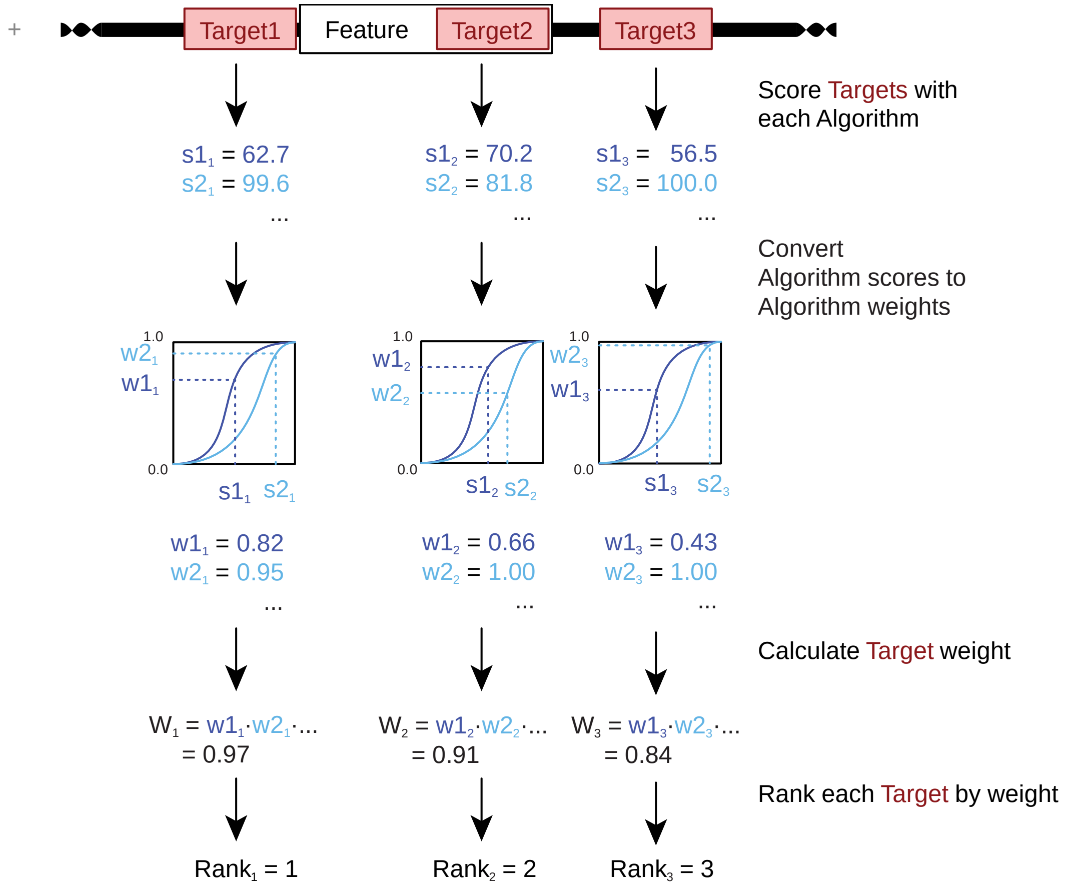


**Figure 3.12 – AddTag ranks Step 1 Targets by combining multiple Algorithm scores into an aggregate, Target weight**

All relevant Step 1 Targets within or near the Feature are scored with each user-selected Algorithms (s1, s2, …). Then the Algorithm scores are converted to weights (w1, w2, …). Note that potential Step 1 dDNA sequences are included in off-target calculations. Finally, Targets are ranked according to their weight W, which is a combination of the individual Algorithm weights. Black arrows represent computational, instead of biological, progression. More detail can be found in **Common Figure Description 1**, Chapter 2.

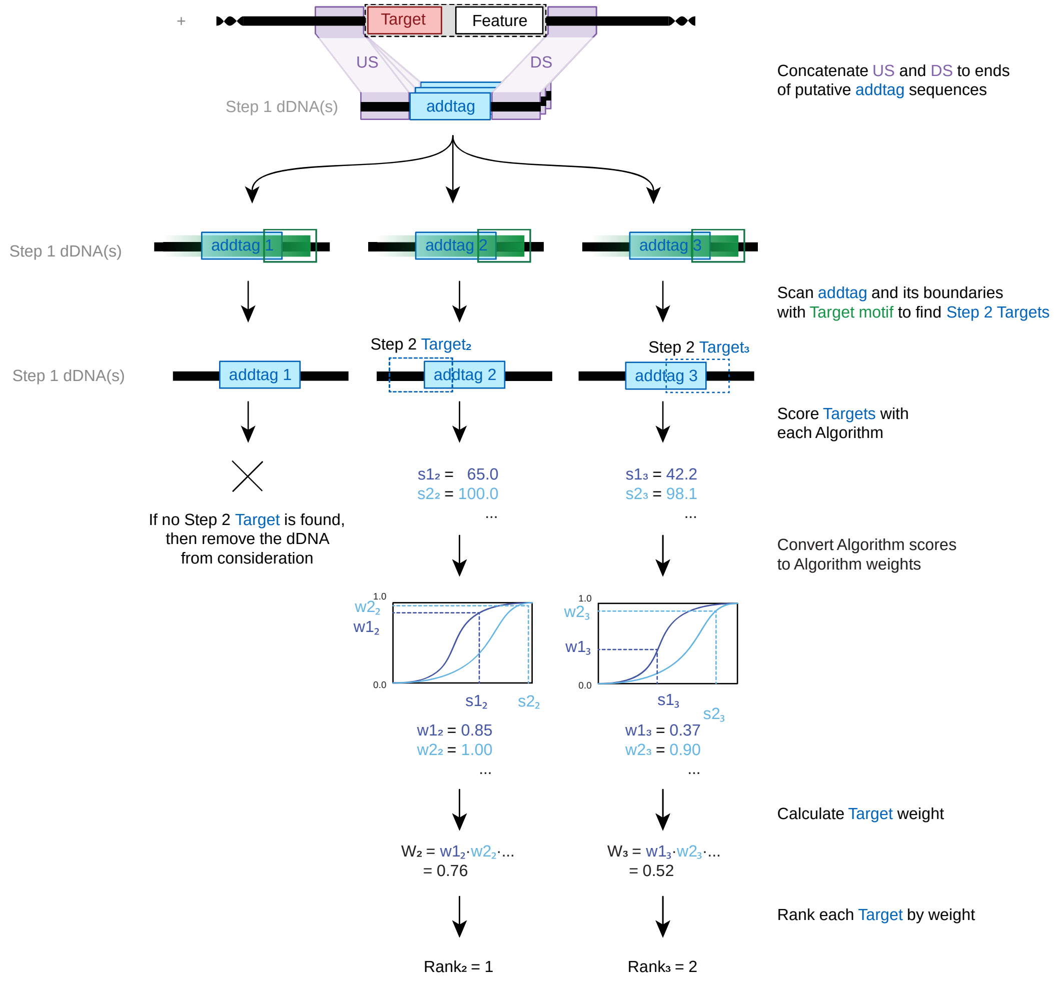


**Figure 3.13 – AddTag ranks Step 2 Targets within the Step 1 dDNA by combining multiple Algorithm scores into a single weight calculation**

The upstream (US) and downstream (DS) sequences flanking each expanded Feature, which contains both the Feature and the Step 1 Target, can produce multiple Step 1 dDNAs—each with a different addtag sequence. ADDTAG scans these potential Step 1 dDNAs for Step 2 Target sequences (green) using the user-provided Target motif(s). Step 1 dDNAs lacking a Step 2 Target are removed from further consideration (╳). Step 1 dDNAs that contain valid Step 2 Targets are scored and weighed in a manner identical to Step 1 Targets (Figure 3.12), except that Step 2 dDNA is included in off-target calculations. Black arrows represent computational progression. More detail can be found in **Common Figure Description 1**, Chapter 2.

## 3.5 Discussion

### 3.5.1 Summary

ADDTAG automates the process of identifying quality Targets that make RGN-induced DSBs within, or adjacent to, the user-defined genomic Feature to be edited. ADDTAG searches both strands of the input or generated DNA for a sequence matching the input Target motif (Figure 3.12). Then ADDTAG evaluates the predicted efficiency and specificity of the Target using any number of on-target (3.4.9) or off-target (3.4.10) scoring Algorithms

(3.4.7). However, unlike most utilities for evaluating RGN Targets, ADDTAG contains a holistic method for ranking Targets based on the scores resulting from each of these Algorithms (Figure 3.10, Figure 4.8, Figure 3.12). ADDTAG implements a breadth of functions to provide a level of flexibility and robustness in gRNA design that is unmatched (Table 3.2).

### 3.5.2 Utility of Algorithm weights

The ADDTAG software introduces the concept of relative ranking of Target efficiency and specificity (for on-target and off-target scores, respectively) through use of the Algorithm weight (Equation 2). The weight of a Target is determined by comparing the Algorithm's score for that Target with the scores of a large sample of other, potential Targets. This process mitigates Algorithm-specific score biases. Thus, the Target weight can be compared across different genomes, across Algorithms, and across RGNs. The Target weight is a novel and important tool for determining which RGN should be used to perform a genome edit.

For example, if a researcher has two organismal strains—one that supports editing through Cas9, and one that supports editing through Cas12a—by comparing the weights between a Cas9 Target and a Cas12a Target, a scientist can definitively determine which strain would have a higher predicted editing success. Without the use of a Target's Algorithm weight, the scientist would consider their Algorithm scores, but discover that they are not comparable. For instance, random Targets give the AZIMUTH and DEEPCPF1 different score distributions (Figure 3.10). It is not obvious if an AZIMUTH score of 60 is better, or worse, than a DEEPCPF1 score of 60. However, the Algorithm weight, which is a continuous approximation of the cumulative distribution function, reveals that a Target with an AZIMUTH score of 60 is better than around the 85% of possible Cas9 Targets, and Target with a DEEPCPF1 score of 60 is better than around only 50% of possible Cas12a Targets. Therefore, using Cas9 has a higher predicted relative editing efficiency than Cas12a.

Please note that the Algorithm weight therefore reflects the relative score of the Target, and does not necessarily reflect actual editing efficiency or specificity. Presumably, the Algorithm score attempts to model the probability of RGN binding and cleaving, but ADDTAG does not guarantee this. Therefore, the AZIMUTH and DEEPCPF1 weights for a Target should not be treated as the probability of successful editing. Additional information linking an Algorithm's score with the probability of successful editing could theoretically be added. However, the current trend, and more likely scenario, is that a new version of the Algorithm will be released attempting to output scores as probabilities between 0 and 100%. When this occurs, the Algorithm weights will achieve a probabilistic interpretation.

### 3.5.3 Accuracy of scoring Algorithms

Ideally, any program that evaluates gRNA specificities would have definitive accuracy for ex *vivo*, *in vivo*, and *in vitro* studies across all experimental species. However, until a more thorough model of extracting informative parameters from arbitrary combinations of gRNA, RGN, gDNA, and cell state is created, ADDTAG relies on Algorithms developed to address only specific combinations of these. ADDTAG incorporates field-standard methods for predicting RGN binding and cutting efficiencies. It requires the user assume that the biological conditions used to train the scoring Algorithms are general enough that they are transferrable to their experiments. In order to obtain the most accurate results, the user must select the appropriate Algorithms trained on their chosen species and RGN protein. Additionally, there is no definitive way to claim that one scoring Algorithm outperforms another scoring Algorithm unless the biological conditions used to train each are comparable. While we have used probability theory as the foundation for the computational assumptions, there are immeasurable numbers of biological conditions that can violate them. We present software that is useful specifically for C. *albicans* editing, but can be applied to other organisms with different genome engineering systems.

### 3.5.4 Reducing off-target effects

With any genome editing process, unintended (off-target) changes to the genome are possible. The default assumption is that the largest source of unintended mutations is the directed chromosomal restriction by the gRNA:RGN complex. One way to determine if off-target mutations occur is to sequence the genome of the experimental organism following genome editing. Future experiments may discover locations in the genome that are more prone to edits (attractors). Assessing the PCR, sequencing, and phenotyping signatures of those attractors would be important for reducing unintended mutagenicity. The ADDTAG software provides the ability to use a case-masked reference genome as input. Therefore, if problematic regions are specified, ADDTAG will avoid identifying primers and Targets there.

To facilitate genome editing experiments with reduced off-target mutagenesis, the ADDTAG software allows the user to specify any number of Target motifs for intended edits as well as any number of Target motifs for unintended edits (3.4.2). ADDTAG supports the use of any arbitrary Target motif (Table 3.2), even if it cannot precisely calculate an on-target or off-target score (when there is no Algorithm for that specific RGN-organism pairing). The off-target scores reflect the specified Target motifs for unintended edits. Therefore, if users suspect their RGN will cleave in unintended but predictable locations, they can specify Target motifs for unintended edits, and select only Targets with high off-target scores. Additionally, because ADDTAG permits the full genome and any ancillary sequences to be used as input, it can do strain-specific off-target calculations.

### 3.5.5 Future directions

The flexible Algorithm interface and Target motif definition provides for forward compatibility with advances. Since the initial development of ADDTAG, several other Algorithms have rose to prominence—most notably ELEVATION [69] as a complement to AZIMUTH on-target scores. Interestingly, finding an ELEVATION off-target score requires a different formula than the MIT Guide Score (Equation 1), which is what ADDTAG uses. When ADDTAG implements the ELEVATION score, it will potentially determine higher-accuracy weights for Cas9 Targets. On another note, designs using the `bartag` dDNA insert type have not been evaluated in actual biological systems. Hence, additional biological validation of the computational methods could be performed. Finally, evaluation of excess sequence repetition in the dDNAs is deferred until *in silico* recombination is performed. This step could be incorporated into the dDNA generation steps to improve computational efficiency, and provide more-immediate user feedback on potential designs.

## 3.6 References

1. Seher TD, Nguyen N, Ramos D, Bapat P, Nobile CJ, Sindi SS, Hernday AD. AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing. *G3 Genes|Genomes|Genetics*. 2021. DOI: 10.1093/g3journal/jkab216.
2. Benchling 2019. Available from: https://benchling.com/.
3. Oliveros JC, Franch M, Tabas-Madrid D, San-León D, Montoliu L, Cubas P, Pazos F. Breaking-Cas—interactive design of guide RNAs for CRISPR-Cas experiments

for ENSEMBL genomes. *Nucleic Acids Research*. 2016; 44(W1):W267-W71. DOI: 10.1093/nar/gkw407, PMID: 27166368.

4. Kondo S, Ueda R. Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics*. 2013; 195(3):715-21. DOI: 10.1534/genetics.113.156737, PMID: 24002648.
5. Park J, Bae S, Kim J-S. Cas-Designer: A web-based tool for choice of CRISPR-Cas9 target sites. *Bioinformatics*. 2015; 31(24):4014-6. DOI: 10.1093/bioinformatics/btv537, PMID: 26358729.
6. Aach J, Mali P, Church GM. CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes. *bioRxiv*. 2014. DOI: 10.1101/005074.
7. Stemmer M, Thumberger T, del Sol Keyer M, Wittbrodt J, Mateo JL. CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLOS ONE*. 2015; 10(4):e0124633. DOI: 10.1371/journal.pone.0124633, PMID: 25909470.
8. Stemmer M, Thumberger T, del Sol Keyer M, Wittbrodt J, Mateo JL. Correction: CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLOS ONE*. 2017; 12(4):e0176619. DOI: 10.1371/journal.pone.0176619, PMID: 28426791.
9. Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. CHOPCHOP v2: A web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Research*. 2016; 44(W1):W272-W6. DOI: 10.1093/nar/gkw398, PMID: 27185894.
10. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: A CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Research*. 2014; 42(W1):W401-W7. DOI: 10.1093/nar/gku410, PMID: 24861617.
11. Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud J-B, Schneider-Maunoury S*, et al*. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*. 2016; 17(1):148. DOI: 10.1186/s13059-016-1012-2, PMID: 27380939.
12. Rodríguez-López M, Cotobal C, Fernández-Sánchez O, Borbarán Bravo N, Oktriani R, Abendroth H, Uka D*, et al*. A CRISPR/Cas9-based method and primer design tool for seamless genome editing in fission yeast. *Wellcome Open Research*. 2017; 1(19). DOI: 10.12688/wellcomeopenres.10038.3, PMID: 28612052.
13. CRISPR Design Tool: *Horizon Discovery, Ltd*. Available from: https://horizondiscovery.com/en/products/tools/CRISPR-Design-Tool.
14. CRISPR Specificity Analysis: *Horizon Discovery, Ltd*. Available from: https://horizondiscovery.com/en/products/tools/crispr-specificity-analysis.
15. Naito Y, Hino K, Bono H, Ui-Tei K. CRISPRdirect: Software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics*. 2014; 31(7):1120-3. DOI: 10.1093/bioinformatics/btu743, PMID: 25414360.
16. Lei Y, Lu L, Liu H-Y, Li S, Xing F, Chen L-L. CRISPR-P: A web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Molecular Plant*. 2014; 7(9):1494-6. DOI: 10.1093/mp/ssu044, PMID: 24719468.
17. Liu H, Ding Y, Zhou Y, Jin W, Xie K, Chen L-L. CRISPR-P 2.0: An improved CRISPR-Cas9 tool for genome editing in plants. *Molecular Plant*. 2017; 10(3):530-2. DOI: 10.1016/j.molp.2017.01.003, PMID: 28089950.
18. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ. CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. *Nature Methods*. 2015; 12:982. DOI: 10.1038/nmeth.3543, PMID: 26322839.
19. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, Ventura A. GuideScan software for improved single and paired CRISPR guide RNA

design. *Nature Biotechnology*. 2017; 35:347. DOI: 10.1038/nbt.3804, PMID: 28263296.

20. Broad Institute. sgRNA Design Tool. 2015.
21. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, *et al*. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotech*. 2014; 32(12):1262-7. DOI: 10.1038/nbt.3026, PMID: 25184501.
22. Wilson LOW, Hetzel S, Pockrandt C, Reinert K, Bauer DC. VARSCOT: Variant-aware detection and scoring enables sensitive and personalized off-target detection for CRISPR-Cas9. *BMC Biotechnology*. 2019; 19(1):40. DOI: 10.1186/s12896-019-0535-5, PMID: 31248401.
23. Ding Y, Li H, Chen L-L, Xie K. Recent advances in genome editing using CRISPR/Cas9. *Frontiers in Plant Science*. 2016; 7(703). DOI: 10.3389/fpls.2016.00703, PMID: 27252719.
24. Hanna RE, Doench JG. Design and analysis of CRISPR-Cas experiments. *Nature Biotechnology*. 2020. DOI: 10.1038/s41587-020-0490-7, PMID: 32284587.
25. Hough SH, Kancleris K, Brody L, Humphryes-Kirilov N, Wolanski J, Dunaway K, Ajetunmobi A, Dillard V. Guide Picker is a comprehensive design tool for visualizing and selecting guides for CRISPR experiments. *BMC Bioinformatics*. 2017; 18(1):167. DOI: 10.1186/s12859-017-1581-4, PMID: 28288556.
26. Hodgkins A, Farne A, Perera S, Grego T, Parry-Smith DJ, Skarnes WC, Iyer V. WGE: A CRISPR database for genome engineering. *Bioinformatics*. 2015; 31(18):3078-80. DOI: 10.1093/bioinformatics/btv308, PMID: 25979474.
27. Park J, Kim J-S, Bae S. Cas-Database: Web-based genome-wide guide RNA library design for gene knockout screens using CRISPR-Cas9. *Bioinformatics*. 2016; 32(13):2017-23. DOI: 10.1093/bioinformatics/btw103, PMID: 27153724.
28. Rauscher B, Heigwer F, Breinig M, Winter J, Boutros M. GenomeCRISPR – A database for high-throughput CRISPR/Cas9 screens. *Nucleic Acids Research*. 2016; 45(D1):D679-D86. DOI: 10.1093/nar/gkw997, PMID: 27789686.
29. Katoh K, Kuma K-i, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*. 2005; 33(2):511-8. DOI: 10.1093/nar/gki198, PMID: 15661851.
30. Kocak DD, Josephs EA, Bhandarkar V, Adkar SS, Kwon JB, Gersbach CA. Increasing the specificity of CRISPR systems with engineered RNA secondary structures. *Nature Biotechnology*. 2019; 37(6):657-66. DOI: 10.1038/s41587-019-0095-1, PMID: 30988504.
31. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Nucleic acids research*. 1985; 13(9):3021-30. DOI: 10.1093/nar/13.9.3021, PMID: 2582368.
32. Zetsche B, Gootenberg Jonathan S, Abudayyeh Omar O, Slaymaker Ian M, Makarova Kira S, Essletzbichler P, Volz Sara E*, et al*. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*. 2015; 163(3):759-71. DOI: 10.1016/j.cell.2015.09.038, PMID: 26422227.
33. Barrangou R, Marraffini Luciano A. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Molecular Cell*. 2014; 54(2):234-44. DOI: 10.1016/j.molcel.2014.03.011, PMID: 24766887.
34. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, Boyaval P*, et al*. Diversity, Activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology*. 2008; 190(4):1401-12. DOI: 10.1128/jb.01415-07, PMID: 18065539.

35. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology*. 2014; 32(3):279-84. DOI: 10.1038/nbt.2808, PMID: 24463574.

36. Swarts DC, Jinek M. Cas9 versus Cas12a/Cpf1: Structure-function comparisons and implications for genome editing. 2018; 9(5):e1481. DOI: 10.1002/wrna.1481, PMID: 29790280.

37. Raikwar SP, Kikkeri NS, Sakuru R, Saeed D, Zahoor H, Premkumar K, Mentor S*, et al*. Next generation precision medicine: CRISPR-mediated genome editing for the treatment of neurodegenerative disorders. *Journal of Neuroimmune Pharmacology*. 2019; 14(4):608-41. DOI: 10.1007/s11481-019-09849-y, PMID: 31011884.

38. Xu H, Xiao T, Chen C-H, Li W, Meyer C, Wu Q, Wu D*, et al*. Sequence determinants of improved CRISPR sgRNA design. *Genome Research*. 2015. DOI: 10.1101/gr.191452.115, PMID: 26063738.

39. Leenay RT, Beisel CL. Deciphering, communicating, and engineering the CRISPR PAM. *Journal of Molecular Biology*. 2017; 429(2):177-91. DOI: 10.1016/j.jmb.2016.11.024, PMID: 27916599.

40. Nguyen N, Quail MMF, Hernday AD. An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. *mSphere*. 2017; 2(2). DOI: 10.1128/mSphereDirect.00149-17, PMID: 28497115.

41. Liu G, Zhang Y, Zhang T. Computational approaches for effective CRISPR guide RNA design and evaluation. *Computational and Structural Biotechnology Journal*. 2020; 18:35-44. DOI: 10.1016/j.csbj.2019.11.006, PMID: 31890142.

42. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009; 10(3):R25. DOI: 10.1186/gb-2009-10-3-r25, PMID: 19261174.

43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G*, et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25(16):2078-9. DOI: 10.1093/bioinformatics/btp352, PMID: 19505943.

44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009; 10(1):421. DOI: 10.1186/1471-2105-10-421, PMID: 20003500.

45. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012; 9:357. DOI: 10.1038/nmeth.1923, PMID: 22388286.

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14):1754-60. DOI: 10.1093/bioinformatics/btp324, PMID: 19451168.

47. Bae S, Park J, Kim J-S. Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*. 2014; 30(10):1473-5. DOI: 10.1093/bioinformatics/btu048, PMID: 24463181.

48. Braglia P, Percudani R, Dieci G. Sequence context effects on oligo(dT) termination signal recognition by *Saccharomyces cerevisiae* RNA Polymerase III. *Journal of Biological Chemistry*. 2005; 280(20):19551-62. DOI: 10.1074/jbc.M412238200, PMID: 15788403.

49. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, Wile BM*, et al*. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research*. 2014; 42(11):7473-85. DOI: 10.1093/nar/gku402, PMID: 24838573.

50. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970; 48(3):443-53. DOI: 10.1016/0022-2836(70)90057-4, PMID: 5420325.

51. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, *et al*. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology*. 2011; 9:467. DOI: 10.1038/nrmicro2577, PMID: 21552286.
52. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, *et al*. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*. 2015; 13:722. DOI: 10.1038/nrmicro3569, PMID: 26411297.
53. Shmakov S, Abudayyeh Omar O, Makarova Kira S, Wolf Yuri I, Gootenberg Jonathan S, Semenova E, Minakhin L, *et al*. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Molecular Cell*. 2015; 60(3):385-97. DOI: 10.1016/j.molcel.2015.10.008, PMID: 26593719.
54. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, *et al*. Diversity and evolution of class 2 CRISPR-Cas systems. *Nature Reviews Microbiology*. 2017; 15:169. DOI: 10.1038/nrmicro.2016.184, PMID: 28111461.
55. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, *et al*. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech*. 2016; 34(2):184-91. DOI: 10.1038/nbt.3437, PMID: 26780180.
56. Kim HK, Song M, Lee J, Menon AV, Jung S, Kang Y-M, Choi JW, *et al*. *In vivo* high-throughput profiling of CRISPR-Cpf1 activity. *Nature Methods*. 2016; 14:153-9. DOI: 10.1038/nmeth.4104, PMID: 27992409.
57. Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y, Lee S, *et al*. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nature Biotechnology*. 2018; 36(3):239-41. DOI: 10.1038/nbt.4061, PMID: 29431740.
58. Labuhn M, Adams FF, Ng M, Knoess S, Schambach A, Charpentier EM, Schwarzer A, *et al*. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Research*. 2017; 46(3):1375-85. DOI: 10.1093/nar/gkx1268, PMID: 29267886.
59. Housden BE, Valvezan AJ, Kelley C, Sopko R, Hu Y, Roesel C, Lin S, *et al*. Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Science signaling*. 2015; 8(393):rs9. DOI: 10.1126/scisignal.aab3729, PMID: 26350902.
60. Chen Y, Zeng S, Hu R, Wang X, Huang W, Liu J, Wang L, *et al*. Using local chromatin structure to improve CRISPR/Cas9 efficiency in zebrafish. *PLOS ONE*. 2017; 12(8):e0182528. DOI: 10.1371/journal.pone.0182528, PMID: 28800611.
61. Uusi-Mäkelä MIE, Barker HR, Bäuerlein CA, Häkkinen T, Nykter M, Rämet M. Chromatin accessibility is associated with CRISPR-Cas9 efficiency in the zebrafish (*Danio rerio*). *PLOS ONE*. 2018; 13(4):e0196238. DOI: 10.1371/journal.pone.0196238, PMID: 29684067.
62. Jensen KT, Fløe L, Petersen TS, Huang J, Xu F, Bolund L, Luo Y, Lin L. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Letters*. 2017; 591(13):1892-901. DOI: 10.1002/1873-3468.12707, PMID: 28580607.
63. Barkal AA, Srinivasan S, Hashimoto T, Gifford DK, Sherwood RI. Cas9 functionally opens chromatin. *PLOS ONE*. 2016; 11(3):e0152683. DOI: 10.1371/journal.pone.0152683, PMID: 27031353.
64. Singh R, Kuscu C, Quinlan A, Qi Y, Adli M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Research*. 2015; 43(18):e118-e. DOI: 10.1093/nar/gkv575, PMID: 26032770.

65. Verkuijl SAN, Rots MG. The influence of eukaryotic chromatin state on CRISPR-Cas9 editing efficiencies. *Current Opinion in Biotechnology*. 2019; 55:68-73. DOI: 10.1016/j.copbio.2018.07.005, PMID: 30189348.

66. Daer RM, Cutts JP, Brafman DA, Haynes KA. The impact of chromatin dynamics on Cas9-mediated genome editing in human cells. *ACS Synthetic Biology*. 2017; 6(3):428-38. DOI: 10.1021/acssynbio.5b00299, PMID: 27783893.

67. Massachusetts Institute of Technology. Optimized CRISPR Design: Aggregate Scores by Guide 2014 [cited 2017 2017-06-22]. Available from: http://crispr.mit.edu/about.

68. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, *et al*. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotech*. 2013; 31(9):827-32. DOI: 10.1038/nbt.2647, PMID: 23873081.

69. Listgarten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J, Gao K, *et al*. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*. 2018; 2(1):38-47. DOI: 10.1038/s41551-017-0178-6, PMID: 29998038.

# Chapter 4 ADDTAG software primer design

## 4.1 Abstract

The AddTag method enables a single locus to be edited multiple times. Verification polymerase chain reaction (vPCR) is a primary way to assess if genome modifications occurred according to, or contrary to, what was intended. These vPCR designs assess if the original DNA (Feature) or modified DNA (Insert) is present in the genome and positioned at the experimental locus. We programmed the ADDTAG software to identify sets of maximally-reusable primers for vPCR that can be assayed in parallel using identical amplification conditions. First, ADDTAG uses the reference genome and dDNA sequences to predict the sequences after each editing step. Next, ADDTAG identifies primer pairs within the genome sequences of each editing step that assess editing success or failure at the locus. Finally, ADDTAG calculates a set of optimal primer pairs that are compatible through a simple genetic algorithm. By providing automated vPCR calculations, ADDTAG streamlines the process of validating edited genomes (https://github.com/tdseher/addtag-project).

## 4.2 Preface

This chapter is adapted from the manuscript authored by myself, Namkha Nguyen, Diana Ramos, Priyanka Bapat, Clarissa J. Nobile, Suzanne S. Sindi, and Aaron D. Hernday, titled "AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing," which was published by the peer-reviewed journal *G3 Genes|Genomes|Genetics* [1].

## 4.3 Introduction

### 4.3.1 Verification PCR primers can be used to evaluate if a genome is edited as intended

Each time a locus is edited, there is a possibility unforeseen errors can arise. With the AddTag (2-step) genome editing method (2.3.2), no direct selectable marker is used for determining if genome editing is successful. For many genome editing procedures, it is common to insert a selectable marker, such as an antimycotic resistance gene like $KAN^R$ [2] or a fluorescent reporter like *GFP* [3, 4], at the edited locus. These genes enable researchers to streak out potentially edited cells, and assess single colonies based on growth or light reactivity. Since the AddTag method does not use selectable markers at the edited locus, an alternative method for evaluating genome edits is necessary, such as sequencing or verification PCR (vPCR). This chapter describes how the ADDTAG software automates design of vPCR primers for evaluating serial genome edits at an experimental locus.

Following editing, vPCR should tell if the resultant genomes are as expected. If we consider a hypothetical AddTag experiment (Figure 4.1), each gDNA (+, Δ, $AB^0$, $AB^1$, $AB^\Delta$) has unique sequences at the edited locus, and each has shared sequences (Table 4.1). We expect the Feature to exist in the +gDNA and $AB^0$gDNA, but not on other gDNAs. We expect the Step 1 dDNA insert (often, the `addtag`) to exist in only the ΔgDNA, and any DNA trimmed in excess of the Feature during Feature expansion to exist in all gDNA except ΔgDNA. Finally, we expect modified DNA introduced in *Feature to be present only in $AB^1$gDNA. ADDTAG identifies primer sequences within or overlapping these regions to make the vPCR design.

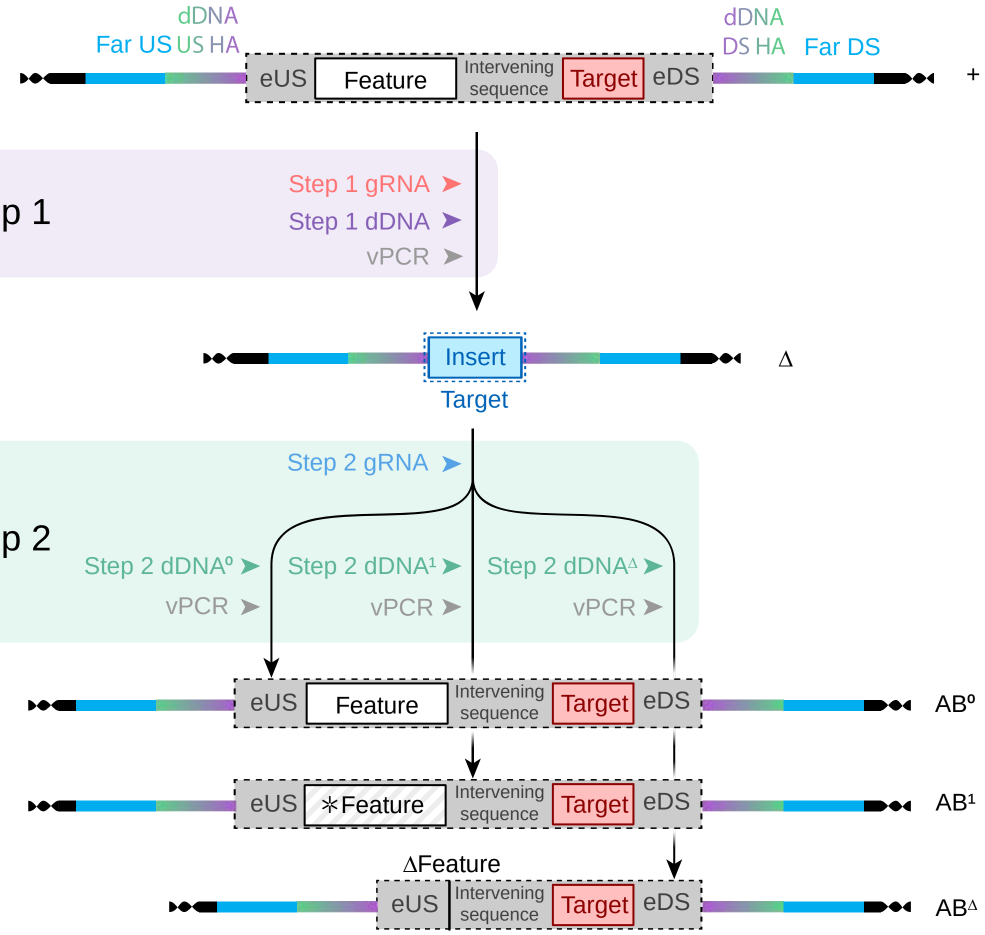


**Figure 4.1 – Indirect (2-step) AddTag editing facilitates artifact-free Feature delations**

An AddTag experiment that performs three parallel add-backs (AB$^{0}$, AB$^{1}$, AB$^{\Delta}$) using the same intermediary (Δ) genome. AB$^{0}$ contains the wild-type Feature; AB$^{1}$ contains the modified ✱Feature; and AB$^{\Delta}$ has the original Feature removed, but DNA removed in excess of Feature during Step 1 is re-incorporated, thereby restoring the Step 1 Target. The grey-colored eUS, intervening sequence, and eDS regions, as well as the Target, are removed during editing Step 1. The sequence identical to the upstream (US) and downstream (DS) dDNA homology arms (HAs) for both Step 1 and Step 2 are shown as the gradient between green and violet. More detail can be found in **Common Figure Description 1**, Chapter 2.

**Table 4.1 – DNA regions expected to exist in edited genomes**

| | | dDNA | | | Intervening | | | | | dDNA | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **gDNA** | **Far US** | **US HA** | **eUS** | **Feature** | **sequence** | **Target** | **Insert** | **✱Feature** | **eDS** | **DS HA** | **Far DS** |
| + | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | ✓ | ✓ |
| Δ | ✓ | ✓ | X | X | X | X | ✓ | X | X | ✓ | ✓ |
| AB$^{0}$ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | ✓ | ✓ |
| AB$^{1}$ | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | ✓ | ✓ | ✓ | ✓ |
| AB$^{\Delta}$ | ✓ | ✓ | ✓ | X | ✓ | ✓ | X | X | ✓ | ✓ | ✓ |

Rows represent genomic DNA (gDNA), and columns are DNA sequences. “eUS” and “eDS” represent the extra sequence on either end of the Feature that may have been trimmed during Step 1 editing. The check (✓) indicates the DNA should exist in that genome, and the cross (✗) indicates the DNA should not exist in that genome. Cells with dark grey background indicate regions that have been excluded from calculations, and cells with light grey background were regions not explicitly identified for vPCR design.

These genome regions need to be somehow translated into diagnostic primer annealing sites. We can reduce the total number of regions by considering which should be present, and which should be absent, in each genome (Table 4.2).

**Table 4.2 – Theoretical, presence of mutually exclusive regions on gDNA**

| | | **eUS/Intervening** | | |
|---|---|---|---|---|
| **gDNA** | **Feature** | **sequence/Target/eDS** | **Insert** | **✳Feature** |
| + | Y | Y | N | N |
| Δ | N | N | Y | N |
| AB$^{0}$ | Y | Y | N | N |
| AB$^{1}$ | N | Y | N | Y |
| AB$^{\Delta}$ | N | Y | N | N |

When evaluating successful editing, the "Y" character indicates vPCR should indicate presence of this region at the locus, and "N" means vPCR should reflect this region is absent from the locus. Cells with dark grey background indicate regions that have been excluded from calculations, and cells with light grey background were regions not explicitly identified for vPCR design.

Each row represents the genomic DNA being tested, with the reference genome (+) edited in Step 1 to make the intermediary genome (Δ). In independent, parallel procedures, the ΔgDNA is converted to each of AB$^{0}$, AB$^{1}$, and AB$^{\Delta}$. The columns "Feature", "eUS/eDS", "Insert", and "✳Feature" represent mutually exclusive genetic regions spread across the gDNAs at the experimental locus.

When a PCR reaction fails to amplify the template, it does not definitively mean the primers failed to hybridize. There are alternative explanations, such as an incorrect annealing temperature, or too much salt or protein contamination in the sample. We therefore programmed ADDTAG to identify one primer pair that would amplify if the sequence of interest is present in the template DNA, and a different primer pair that would amplify if the sequence of interest is absent (Table 4.3 "Deleted" and "Present" columns). Also, it is possible a genome edit may occur, but instead of at the intended locus, the edit happens somewhere else in the genome. To address this, we programmed ADDTAG to find primer pairs with one primer within the Feature, Step 1 dDNA insert (Insert), or ✳Feature, and the other at the locus (Table 4.3 "Present at locus" columns).

**Table 4.3 – Theoretical, expected amplification of mutually exclusive regions on gDNA**

| gDNA | Feature | | | | eUS/Intervening sequence/Target/eDS | | | | Insert | | | | ✻Feature | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Deleted | Deleted from locus | Present | Present at locus | Deleted | Deleted from locus | Present | Present at locus | Deleted | Deleted from locus | Present | Present at locus | Deleted | Deleted from locus | Present | Present at locus |
| + | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Δ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ |
| AB$^{0}$ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ |
| AB$^{1}$ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| AB$^{\Delta}$ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ |

Each cell in Table 4.2, which indicates either presence or absence of the mutually exclusive region, is expanded to a vector of "Deleted", "Deleted from locus", "Present", and "Present at locus". When evaluating successful editing, the check (✓) indicates positive vPCR amplification should be observed, and the cross (✗) indicates no amplification should be observed. Cells with dark grey background indicate reactions that have been excluded from calculations.

Each row represents the genomic DNA being tested, with the reference genome (+) edited in Step 1 to make the intermediary genome (Δ). In independent, parallel procedures, the ΔgDNA is converted to each of AB$^{0}$, AB$^{1}$, and AB$^{\Delta}$. The columns "Feature", "eUS/eDS", "Insert", and "✻Feature" represent mutually exclusive genetic regions spread across the gDNAs at the experimental locus.

In order to simplify programming, we removed the mutual exclusivity requirement between Feature, Insert, and ✻Feature. Therefore, primers were designed against the entire eFeature (composed of eUS, Feature, intervening sequence, Target, and eDS) for the +gDNA, and the entire edited eFeature (✻eFeature, composed of eUS, ✻Feature, intervening sequence, Target, and eDS) for the ABgDNA. Also, experiments producing a Feature deletion, while simultaneously utilizing Feature expansion, were excluded from consideration (AB$^{\Delta}$gDNA row with grey background). Because we decided to not distinguish between eFeatures and Features, the eUS/intervening sequence/Target/eDS regions are not explicitly used for diagnosing correct genome editing through vPCR (light grey cells in Table 4.1, Table 4.2, and Table 4.3).

With these simplifications (Table 4.4), ADDTAG scans each region for potential primer sequences to serve for vPCR (Figure 2.8). The sF primers are found in the forward strand of the Far US region and the sR primers are found in the reverse strand of the Far DS region. ADDTAG scans the eFeature on both strands to identify potential +oF, +oR, +iF, and +iR primers. Similarly, ADDTAG determines the ΔoF, ΔoR, ΔiF, and ΔiR primers by searching the Step 1 dDNA insert (containing the `addtag`, `mintag`, etc), and the AoF, AoR, AiF, and AiR primers by searching the Step 2 dDNA insert (containing the ✳eFeature). ADDTAG identifies a single sF/sR primer pair that surrounds the locus on each genome, and then identifies a sF/oR, oF/sR, and iF/iR pairs for each editing step (Figure 2.9).

**Table 4.4 – Pattern of expected vPCR amplification given correct editing**

| | sF/sR | sF/+oR | +oF/sR | +iF/+iR | sF/ΔoR | ΔoF/sR | ΔiF/ΔiR | sF/AoR | AoF/sR | AiF/AiR |
|---|---|---|---|---|---|---|---|---|---|---|
| **gDNA** | **U** | **eFeature** | | | **Insert** | | | **✳eFeature** | | |
| + | ○ | ● | ● | ● | ○ | ○ | - | ◉ | ◉ | ◉ |
| Δ | ● | ○ | ○ | ○ | ◐ | ◑ | - | ○ | ○ | ○ |
| AB$^0$ | ○ | ◉ | ◉ | ◉ | ○ | ○ | - | ● | ● | ● |
| AB$^1$ | ○ | ◉ | ◉ | ◉ | ○ | ○ | - | ● | ● | ● |

Cells contain symbols representing no amplification (○), amplification (●), mutually exclusive amplification (◐/◑), potential amplification (◉), and no valid primer pair identified (-). Each row represents the genomic DNA being tested, with the reference genome (+) edited in Step 1 to make the intermediary genome (Δ). In independent, parallel procedures, the ΔgDNA is converted to each of AB$^0$ and AB$^1$.

The columns “eFeature”, “Insert”, and “✳eFeature” represent regions taken from the +, Δ, and AB gDNAs, respectively (Figure 4.1). The primers +oF, +oR, +iF, and +iR are all designed against the eFeature found on +gDNA; the primers ΔoF, ΔoR, ΔiF, and ΔiR are all designed against the Insert found on ΔgDNA; and the primers AoF, AoR, AiF, and AiR are all designed against the ✳eFeature found on ABgDNA. The union (U) column containing sF/sR represents a primer pair designed for use with all regions. Because the sequences of eFeature and ✳eFeature are not mutually exclusive, the same primers might be identified for them, and thus amplification may not be mutually exclusive.

With the `addtag`, `unitag`, and especially `mintag` Step 1 dDNA insert types (3.4.3), the length of the Insert sequence is quite small. This leaves very few, if any, primers of usable quality for vPCR. To broaden the number of potential primers, junctions between regions are interrogated. The junction between a region unique to a particular gDNA and its neighboring region may present unique sequences that can be used for vPCR design. For example, the *EFG1*$_{CDS}$ locus (2.4.3) potentially has usable primer annealing sites that all overlap the US HA, the `mintag(CCTC)`, and the DS HA regions of the Step 1 dDNA.

The sF/sR primer pair will not amplify when their annealing sites are too far apart, therefore yielding positive amplification when the eFeature is absent from the locus (Table 4.4, “sF/sR” column). The iF/iR primer pair will only amplify when the sequence region between them exists (Table 4.4 “eFeature”, “Insert”, “✳eFeature” columns). However, iF/iR amplification will arise even if the eFeature is at an unintended location in the genome. For example, if the Insert from the Step 1 dDNA incorporates into an off-target locus, then ΔiF/ΔiR may still amplify. The sF/oR and oF/sR primers hybridize both inside the sequence region (Feature, Insert, ✳Feature) and outside at the edited locus (Far US, Far DS), and thus serve to indicate if the presence of the region at the locus. In other words, upon successful reintegration of the Step 2 dDNA into the locus, the sF/AoR or AoF/sR primer pairs should yield a positive PCR product that is indicative of the intended genomic edit.

### 4.3.2 Overview of computational workflow for vPCR primer design

We designed the AddTag software to generate vPCR primers for confirming the presence or absence of intended editing at the chosen locus. The presence of amplification is usable as both a positive and negative control. Most amplicons are between 400 and 800 nt in length so they can be amplified with minimal changes to PCR conditions. The amplicons resulting from sF/oR and oF/sR amplification span the intended locus on at least one of either the upstream junction or the downstream junction, preferably both. This enables diagnosing if the Feature or Insert is at the intended location in the genome. Additionally, these pairs enable discrimination of erroneous within-Feature/Insert HDR events. The iF/iR primer pair will amplify a region within the Feature/Insert, thereby indicating its presence somewhere in the genome. The sF/sR primer pair will amplify only if the Feature is absent at the locus.

AddTag uses a 3-step computational process to find these vPCR primers (Table 4.5). (A) AddTag performs *in silico* recombination to generate the expected genomes given successful dDNA integration, and it links subsequent edits to the same locus together into groups (4.4.1). (B) AddTag delimits 4 discrete regions a primer can be found for each genome in each group, and it identifies all usable primer pairs (4.4.2). (C) AddTag performs simulated annealing to identify the best set of primers for each group of homologous Features (4.4.3).

**Table 4.5 – Computational workflow for identifying vPCR primers**

| | # | Task |
|---|---|---|
| Parse input | 1 | Genome |
| | 2 | Series of dDNA sequences |
| (A) Recombine | 3 | Link each dDNA to the genome by homology |
| | 4 | For Step 1, and then Step 2, generate expected genome sequence given intended editing is successful |
| (B) Primer Calculations | 5 | Identify sF and sR primers shared across all genomes |
| | 6 | For each genome (+, Δ, AB), identify iF, iR, oF, and oR primers |
| | 7 | Weigh Primers and PrimerPairs. |
| (C) Optimization | 8 | Maximize PrimerPair weights through simulated annealing to create PrimerSets |
| | 9 | Rank PrimerSets by number of Primers and PrimerSet weight |

AddTag can perform these steps for each locus in parallel, thereby simulating dDNA incorporation at several loci in the same editing step. For each edited locus, it will identify a separate set of sF and sR primers as well as iF, iR, oF, and oR primers for each editing Step.

There are 3 classes of objects implemented for primer optimization: Primer, PrimerPair, and PrimerSet. A Primer object stores an oligonucleotide sequence and evaluates it based on several thermodynamics-related criteria (4.4.4). PrimerPair objects link two Primer objects and evaluate their amplicon specificity (4.4.5). PrimerSet objects hold a collection of PrimerPair objects and finds their combined weight (4.4.6), or fitness, during the optimization process (4.4.3).

AddTag designs an integrated minimal set of PCR primers with which users can amplify dDNA fragments and verify genome edits throughout 1- or 2-step genome editing applications. The AmpF and AmpR primers are selected to efficiently amplify Step 2 dDNA fragments for native locus add-backs (Figure 2.7). The remaining primers are designed to produce unique PCR products that are indicative of successful deletion and subsequent restoration or modification of the Feature being edited (Figure 2.11).

## 4.4 Results

### 4.4.1 (A) In silico recombination

Before ADDTAG can find the best primer set, it must first identify regions of DNA shared across all editing rounds that flank the intended locus. The initial step to achieving this is by generating the expected genome sequences given successful dDNA integration after each round of genome editing. This process is called *in silic*o recombination, and proceeds in a cyclical manner. The following steps are performed for each round of genome editing:

(1) all dDNAs are aligned to the gDNA;

(2) each gDNA to dDNA alignment is segregated into “Far US”, “US HA”, “Insert”, “DS HA” and “Far DS” regions; and

(3) any sequence on the gDNA that is between the US HA and DS HA is replaced with the Insert region of the dDNA.

This process is performed in the order from higher genomic coordinates to lower coordinates, thereby allowing for multiple loci to be changed with dDNA in a single editing step. After those 3 steps in each round of editing, all dDNA edits that overlap are deemed to be occurring at the same locus, and these are grouped together.

The ADDTAG program naïvely assumes the expected number of alignments parallels the probability of an HDR event. The E-value describes the expected number of alignments by chance given the scoring scheme, nucleotide composition, alignment length, and genome size [5]. We therefore assume that a “significant” alignment is one whose E-value is below 1 (E-value < 1). These are the only alignments that are considered for HDR. As the length and percent identity of an alignment increases, the E-value decreases. Most flanking arms will have smaller subsequences that can align to hundreds of places in any given genome, and thus are potential sources of non-target HDR. However, ADDTAG assumes these hundreds of micro-alignments have negligible impact on the total number of HDR events because these sites are unlikely to have double-stranded breaks. Thus, the E-value of 1 can be thought of as a maximum threshold by which an alignment can be considered significant, and therefore likely to drive HDR.

*In silic*o recombination helps identify problems with dDNA design. Using traditional 1-step CRISPR/Cas-induced HDR to directly edit a Feature may fail because of excess microhomology. Editing may be faulty if homology exists between subsequences within an eFeature HA and a dDNA (Figure 4.2) or between subsequences within the eFeature and a dDNA HA (Figure 4.3). Duplicate instances of HA in either the gDNA or the dDNA thereby decrease genome editing efficiency.

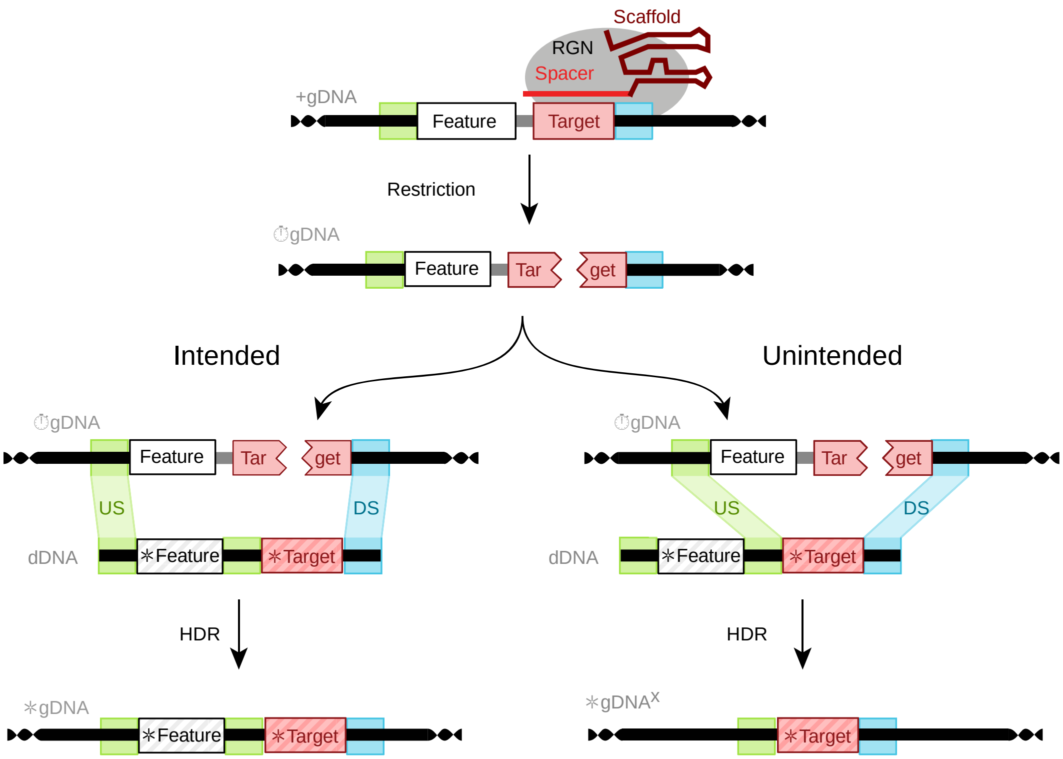


**Figure 4.2 – Excessive homology within dDNA reduces efficiency of intended dDNA incorporation**

In this example, the dDNA contains a subsequence internally that is similar to one of its homology arms (US). Either US region in the dDNA can potentially be selected as a focus for HDR, producing either the intended modification (✻gDNA), or an unintended modification (✻gDNA$^{x}$). The longer the dDNA sequence, the more likely its homology arms share similarity with subsequences inside the homolog arms. Therefore, longer dDNA sequences are less likely to produce correct edits.

More detail can be found in **Common Figure Description 1**, Chapter 2.

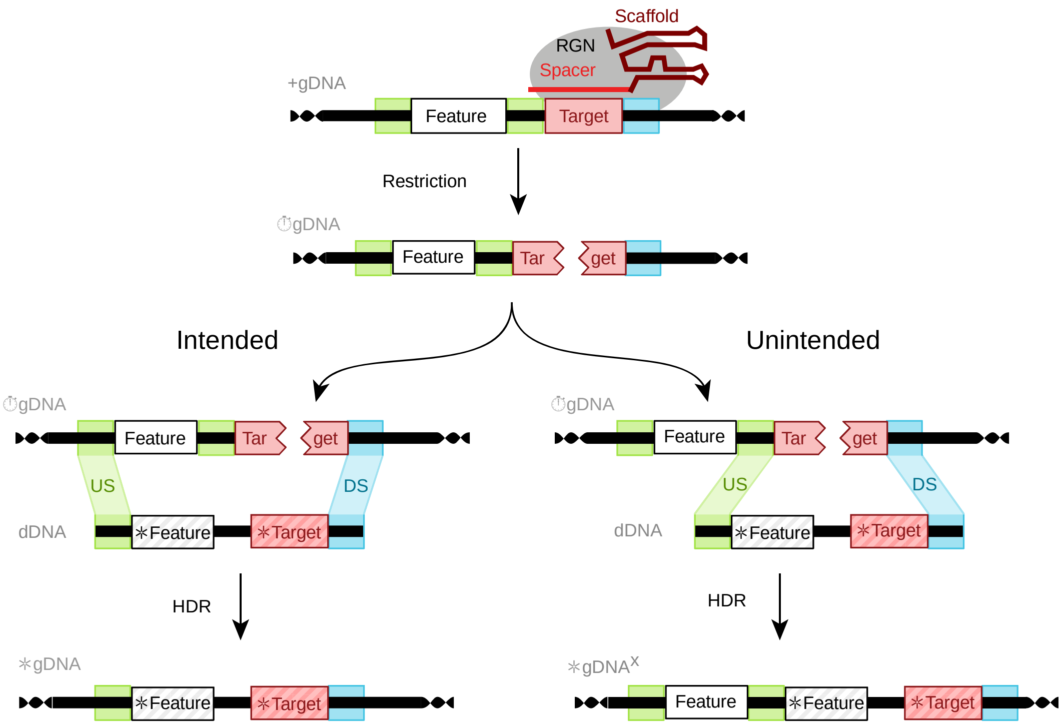


**Figure 4.3 – Excessive homology within gDNA reduces efficiency of inteded dDNA incorporation**

Here, the US flanking homology arm exists multiple times on the genomic DNA (gDNA). Each US instance is a potential focus for HDR, producing either the intended modification (✻gDNA), or an unintended modification (✻gDNA$^X$). The longer the expanded Feature, the more likely repeats are shared between dDNA homology arms and eFeature subsequences. Therefore, longer eFeature sequences are less likely to produce correct edits.

More detail can be found in **Common Figure Description 1**, Chapter 2.

For simplicity, ADDTAG assumes that a single restriction event happens for each pair of dDNA HAs that align. When ADDTAG performs computational recombination, it reports to the user the identified US and DS regions it identifies for each editing Step. When these regions differ from what is expected, then it indicates that excessive homology exists between the gDNA and dDNA, and the user should select a different dDNA for genome editing.

### 4.4.2 (B) Identify primer pairs

After the group of homologous Features is created, AddTag (1) identifies primer sequences, (2) assesses their usability, and (3) assesses the usability of all required primer pairs.

(1) Four strand-specific regions are identified for each locus: “Far US” where sF Primers reside, “Far DS” where sR Primers reside, the forward strand of “Feature/Insert” containing oF and iF Primers, and the reverse strand of “Feature/Insert” containing oR and iR Primers (Figure 2.8). AddTag uses a sliding window to identify all Primers within each region of each gDNA (Figure 4.4). Depending on the desired allelic specificity, certain Primers are

excluded from the list of potential Primers (Figure 4.5). For instance, if the Primers need to be multi-allelic, then Primers with identical sequences existing in all gDNA within the region are kept, and any Primer whose sequence does not exist in all gDNAs is discarded.

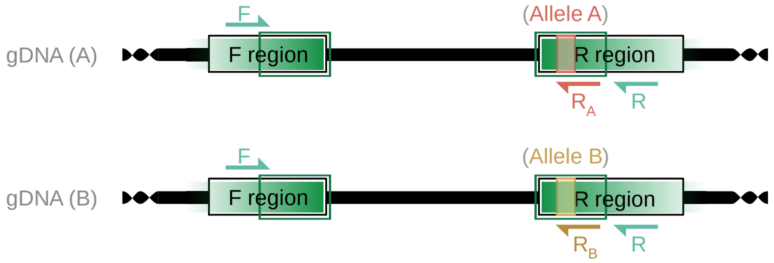
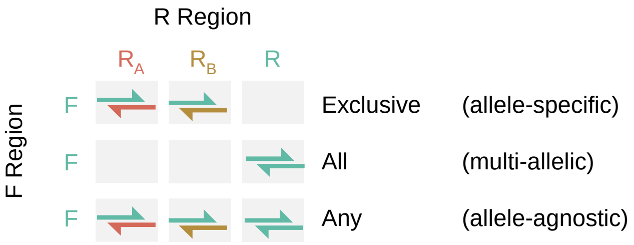


**Figure 4.4 – Find all primers in forward and reverse regions**

Depicted are two regions (F region, R region), each present as homologs on a pair of chromosomes (A, B). There is an allelic variant in the R region, denoted by the colored rectangles (Allele A, Allele B). First, ADDTAG uses a simple sliding window method to identify all potential primers within the genomic regions that should contain the forward (F) and reverse (R) primers. In this example, 1 valid forward primer is identified in the F region, and its sequence is identical in both chromosomes (green color). 3 reverse primers are identified—two allele-specific sequences (red and yellow colors), and one multi-allelic sequence (green color) in the R region. More detail can be found in **Common Figure Description 1**, Chapter 2.



| F Region | R Region: $R_A$ | $R_B$ | R | | |
|---|---|---|---|---|---|
| F | ⇄ | ⇄ | | Exclusive | (allele-specific) |
| F | | | ⇄ | All | (multi-allelic) |
| F | ⇄ | ⇄ | ⇄ | Any | (allele-agnostic) |

**Figure 4.5 – Group each potential primer pair by allelic specificity**

Following sliding window identification of all individual primers (Figure 4.4), all potential pairings of forward and reverse primers are assigned allelic specificity. In this continued example, there are three possible primer pairs.

(2) The Far US and Far DS are processed once for each group of homologous Features, but the stranded Feature/Insert regions are processed once for each gDNA (the reference genome followed by each predicted edited genome). Potential primer sequences are assessed for suitability (4.4.4). Following primer design best-practices [6-9], ADDTAG evaluates primers based on melting temperature; %GC; length; propensity to form hairpins, homodimers, and heterodimers, and several other metrics.

(3) If a primer passes the suitability requirement, then it is considered for pairing. For each locus, a list of potential sF/sR primer pairs are assessed for compatibility. For each gDNA (the reference genome followed by each predicted edited genome) a list of potential sF/oR, oF/sR, and iF/iR pairs is similarly created. Thus, each desired vPCR primer pair (sF/sR for all gDNA and sF/oR, oF/sR, iF/iR for each gDNA) is associated with a list of potential

primer pairs (stored as PrimerPair objects). AddTag calculates the PrimerPair weight (Equation 6) which summarizes the primer pair's ability to perform successful amplification (Figure 4.6).

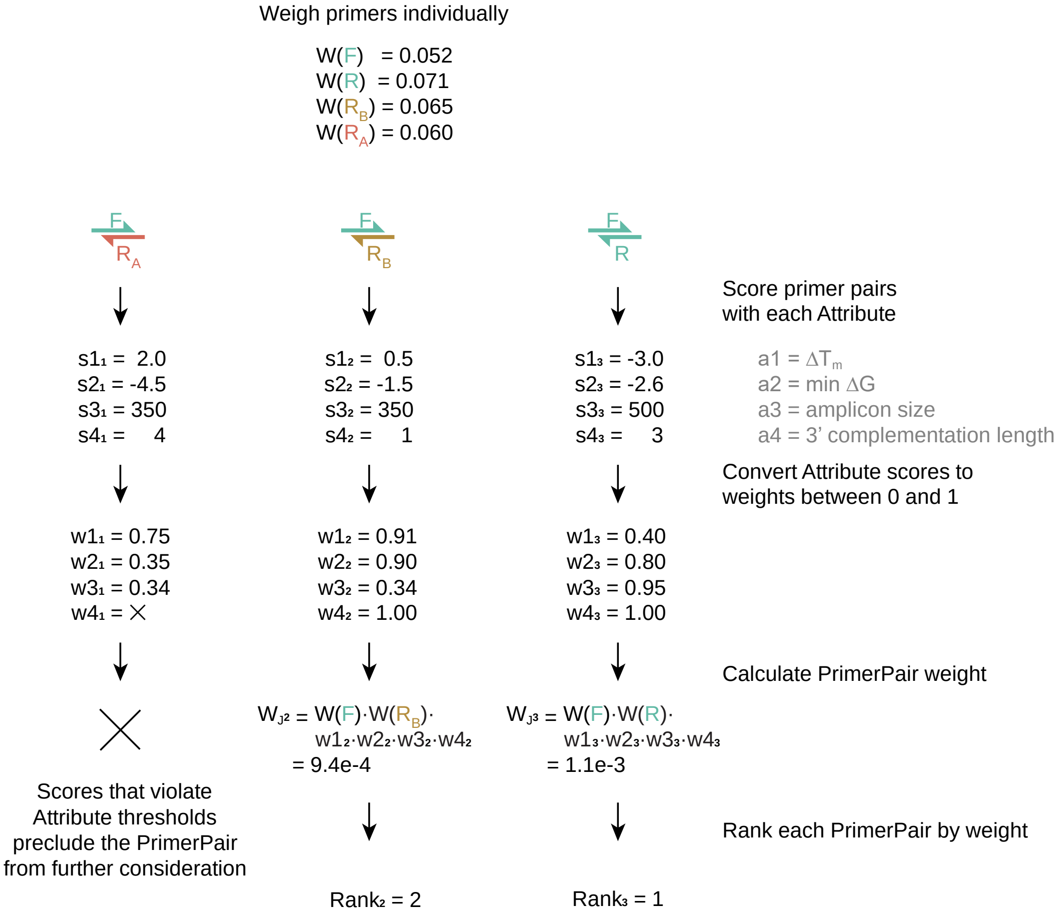

Weigh primers individually
$W(F) = 0.052$
$W(R) = 0.071$
$W(R_B) = 0.065$
$W(R_A) = 0.060$
F
$R_A$
F
$R_B$
F
R
Score primer pairs with each Attribute
$s1_1 = 2.0$
$s2_1 = -4.5$
$s3_1 = 350$
$s4_1 = 4$
$s1_2 = 0.5$
$s2_2 = -1.5$
$s3_2 = 350$
$s4_2 = 1$
$s1_3 = -3.0$
$s2_3 = -2.6$
$s3_3 = 500$
$s4_3 = 3$
$a1 = \Delta T_m$
$a2 = \min \Delta G$
a3 = amplicon size
a4 = 3' complementation length
Convert Attribute scores to weights between 0 and 1
$w1_1 = 0.75$
$w2_1 = 0.35$
$w3_1 = 0.34$
$w4_1 = \times$
$w1_2 = 0.91$
$w2_2 = 0.90$
$w3_2 = 0.34$
$w4_2 = 1.00$
$w1_3 = 0.40$
$w2_3 = 0.80$
$w3_3 = 0.95$
$w4_3 = 1.00$
Calculate PrimerPair weight
$W_{J^2} = W(F)\cdot W(R_B)\cdot w1_2\cdot w2_2\cdot w3_2\cdot w4_2 = 9.4e\text{-}4$
$W_{J^3} = W(F)\cdot W(R)\cdot w1_3\cdot w2_3\cdot w3_3\cdot w4_3 = 1.1e\text{-}3$
Scores that violate Attribute thresholds preclude the PrimerPair from further consideration
Rank each PrimerPair by weight
$Rank_2 = 2$
$Rank_3 = 1$

**Figure 4.6 – Score, weigh, and rank each potential primer pair**

Following identification of forward and reverse primers (Figure 4.4), and Primer weight calculations (Equation 5), each potential primer pair (PrimerPair) has its attributes scored and then weighed (Figure 4.7). Finally, the pairs are ordered by PrimerPair weight (Equation 6), with the highest weight having the best ranking. The final selection of primer pairs for vPCR design proceeds through simulated annealing (Figure 2.10).

After evaluating the potential primer pairs corresponding to sF/sR, sF/oR, oF/sR, and iF/iR across all gDNAs, AddTag performs design optimization (4.4.3).

### 4.4.3 (C) Optimizing sets of primers

After all lists of potential PrimerPairs are created and assessed for compatibility, those pairs are fed into a type of genetic algorithm, called simulated annealing, to identify the best set of compatible primers. Each putative optimization is stored in a PrimerSet data structure. For a typical two round experiment at a single locus, there is a single amplicon A (sF/sR), followed by the amplicons B, C, and D (sF/oR, oF/sR, and iF/iR, respectively) for the +gDNA, then amplicons B, C, and D for the ΔgDNA, and then amplicons B, C, and D for the ABgDNA. This results in a total of 10 PrimerPairs (Figure 2.9). For simplicity, AddTag first fixes the sF/sR pair, then cycles through potential pairs for the remaining PrimerPairs, evaluating their weights each iteration. Note that the AmpF/AmpR PrimerPair determination is separate from the vPCR Primer set determination, and is therefore not included in this section. After using simulated annealing for each sF/sR pair, the results are sorted by first the number of PrimerPairs identified, and second by the weight of the PrimerSet.

Briefly, simulated annealing proceeds as follows. The primer design is composed of several lists of PrimerPairs. The goal is to select one element from each list such that all selected elements produce a high PrimerSet weight (4.4.6). The weight of initial selection of elements for each list is iteratively compared with alternative selection of list elements. The process halts once the number of iterations is reached, or a local optimum is determined. ADDTAG uses a "ranked" simulated annealing process by default, where primer pairs are randomly selected based on their joint weight. A greedy alternative is included where the lowest-weighted primer pair is always swapped with a higher-weighted pair. Early tests indicate this results in the global optimum at a high frequency (>90% of the time). Each potential selection of primer pairs is stored in a PrimerSet object, and the highest-weighted PrimerSet is selected for each fixed sF/sR pair. Simulated annealing allows for automatic determination of near-optimal set of compatible primers.

### 4.4.4 Evaluating single Primers

One central utility of ADDTAG is to choose appropriate primer sequences that bind to genomic templates. There are many possible sequences that can potentially meet the needs of an experiment. However, not all segments of DNA templates yield primers with sequence compositions favorable for PCR. ADDTAG therefore implements a method for evaluating primer sequences. First, ADDTAG independently considers several attributes of the primer sequence, such as its length, melting temperature, and %GC. For each attribute, the primer sequence is scored with model to produce an attribute score. Then, ADDTAG converts each attribute score into a weight. Weights of multiple attributes are aggregated to give information about how well the primer will anneal and amplify the template DNA. The aggregate weight is useful for comparing between primer sequences.

For computational expediency, a primer's attribute scores are calculated successively. If a score is outside usability thresholds (Figure 4.7, vertical lines), then calculation of the remaining attribute scores is halted. Thus, weights are not fully calculated if a usability threshold is violated, and the primer is removed from consideration in downstream analyses. These usability thresholds are tunable by the user on the command line. Looser stringency means more primer sequences are kept for consideration, so later analyses take longer to complete.

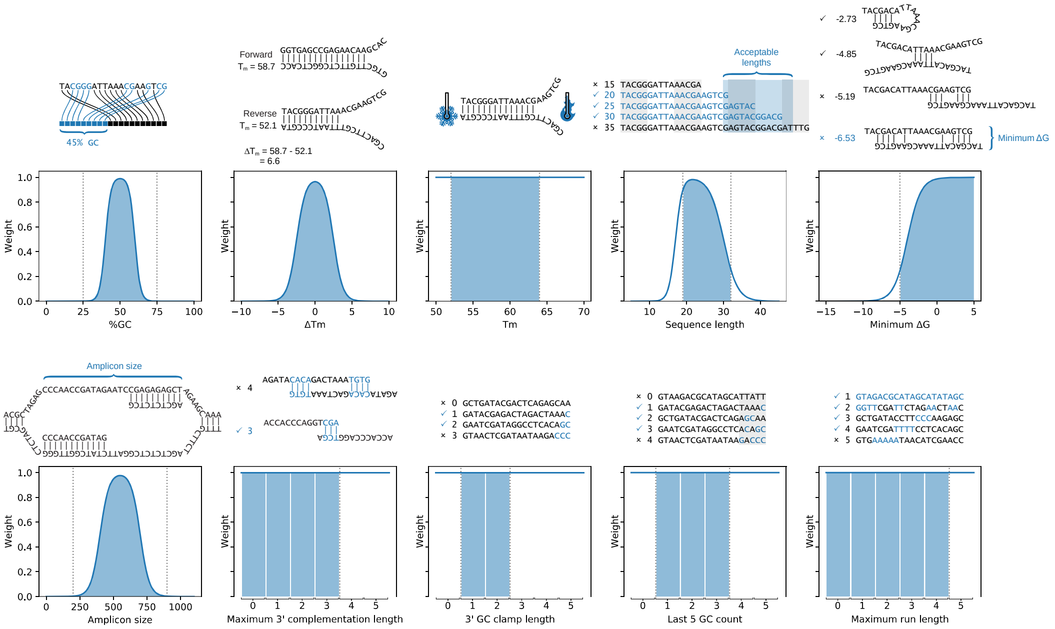


**Figure 4.7 – ADDTAG calculates attribute-specific weights from primer and primer pair scores**

Primer and PrimerPair attribute scores are converted to attribute weights using either uniform, unisigmoidal, or bisigmoidal functions. Minimum and maximum score cutoffs are indicated by vertical dashed lines. Blue areas under the curve represent bounds of the score domain that yield positive weights. Above each graph is a text example illustrating the attribute, where check marks (✓) indicate a positive weight assignment, and cross marks (×), indicate a zero weight; blue text indicates important subsequences for attribute-specific scores.

ADDTAG scores the following attributes of a primer's sequence (Figure 4.7):

(1) the %GC of the primer sequence;

(2) the length of the primer sequence;

(3) the minimum change in Gibbs' free energy (ΔG) of the primer sequence (Although some evidence suggests the effective interference of intended behavior of primers differs based on whether it is a hairpin, homodimer, or heterodimer, we elected to consider only the minimum ΔG reported by the user-selectable software UNAFOLD [10], VIENNARNA [11], or PRIMER3 [12]. DNA oligonucleotides often form unwanted, stable, secondary structures in aqueous solutions. When two molecules of the same oligonucleotide attach to each other, it is called a homodimer (also known as self-dimer). A heterodimer is when DNA molecules with different oligonucleotide sequences hybridize to each other. A hairpin is formed when portions of an oligonucleotide hybridize with itself. Each of these secondary structures has the potential to lower the efficiency of a PCR reaction. The more consecutive nucleotides that hybridize, the stronger the bond. The stronger the bond, the less likely the oligonucleotide will bind with the complementary DNA it is intended for. The ADDTAG software evaluates the minimum ΔG for each of these three secondary

structures. The smaller the ΔG, the more likely the oligonucleotide will form secondary structures and produce a problematic PCR reaction.);

(4) the melting temperature ($T_m$) of the primer as a proxy for annealing temperature ($T_a$) (This holds valid only under the assumption the primer lengths are similar and secondary structures non-existent [7, 13, 14].);

(5) the maximum 3' self-complementation length [7] (Excess 3' complementarity can cause PCR extension of dimerized oligonucleotides. Complementation between oligonucleotides can cause primer dimers, which serve as competitors to the intended template, thus decreasing the efficiency of the intended template amplification [15]. To avoid this, the ADDTAG software implements a maximum 3' complementation length cutoff for both homodimer and heterodimers.);

(6) the 3' GC clamp length as a computationally-efficient proxy for 3' end stability [16, 17];

(7) the number of G and C residues in the last 5 positions of the primer sequence to serve as a heuristic to minimize off-target hybridization by the 3' end of the primer (If the 3' end of an oligonucleotide is rich in G/C, then it might hybridize to a template DNA, even though the 5' end is not hybridized. This is a phenomenon termed "mispriming.");

(8) and the number of consecutive, repeated nucleotides (also known as run length).

A primer attribute $a$ describes an intrinsic property of the primer at experimental conditions (salinities, temperature, and nucleotide concentrations), and is calculated by feeding the primer sequence ($seq$) into a model that produces the attribute score ($s$). Accordingly, $A$ is the set of all primer attribute models, $a \in A$, and $S$ is the set of all attribute scores for a primer sequence

$$S(seq) = [s_a(seq) \text{ for } a \in A].$$

Each primer attribute score $s$ derived from the sequence is passed through a function $w$ to scale its quality on a score from 0 to 1, called its weight

$$0 \leq w \leq 1.$$

For each attribute, $w$ represents the broad probability of successful hybridization to genomic DNA using that sequence as a primer. Higher $w$ correspond to higher likelihood of successful binding and thus amplification. We define the general formula for weight as a sigmoidal (also called logistic) function to model the desired attributes. Sigmoidal functions are useful because they define thresholds. On one side of the threshold, there is a severe penalty, and on the other side, the penalty is light. Additionally, ADDTAG implements hard minimum and maximum cutoffs for each attribute score, outside which the weight is set to 0. Attributes (1) and (2) apply bisigmoidal weight functions; attribute (3) applies a unisigmoidal weight function; and attributes (4), (5), (6), (7), and (8) apply uniform weight functions (Figure 4.7).

Below, we review attribute parameters for a typical bisigmoidal weight function. Each attribute $a$ provides a defined set of parameters $\theta$ to transform the score $s$ into an attribute weight $w$. First, ADDTAG defines two thresholds, which we refer to as $up$ and $down$

$$\theta = \{x_{\mathrm{up}}, slope_{\mathrm{up}}, x_{\mathrm{down}}, slope_{\mathrm{down}}\}.$$

Thus, $x_{up}$ and $x_{down}$ represent the inflection points of the slopes, and $slope_{up}$ and $slope_{down}$ represent the steepness of the function (Figure 4.8).

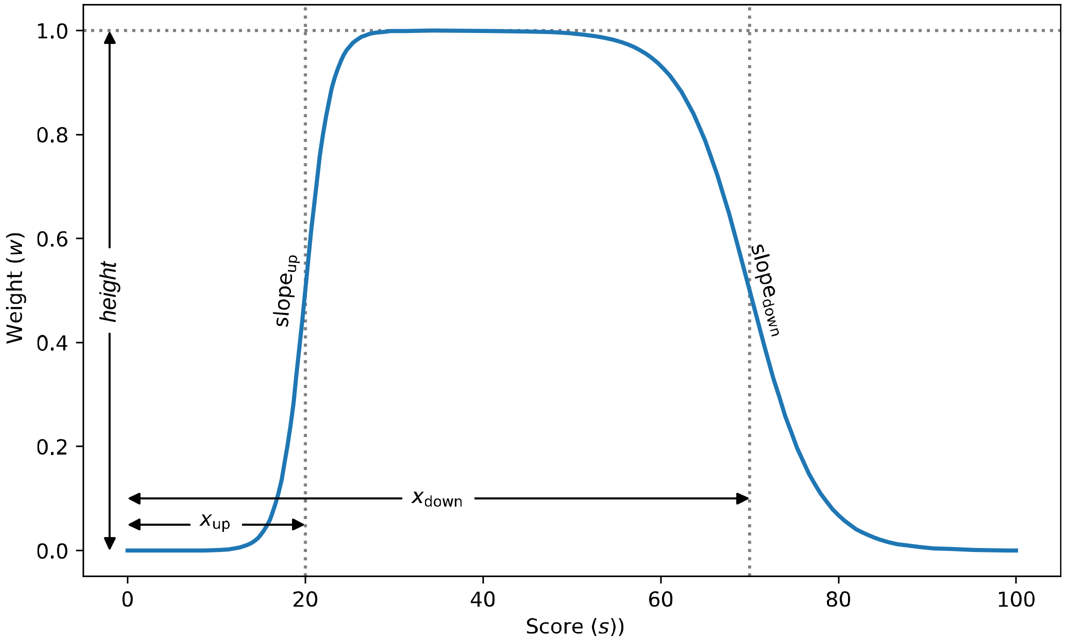


**Figure 4.8 – Description of bisigmoidal function parameters**

Each sigmoid function is defined by an inflection point and its slope. The standard bisigmoidal function contains a first sigmoid term with the inflection point at $x_{\mathrm{up}}$ and a positive steepness of $slope_{\mathrm{up}}$; and a second sigmoid term with the inflection point at $x_{\mathrm{down}}$ and a negative steepness of $slope_{\mathrm{down}}$. ADDTAG sets the final parameter $height = 1.0$ for simplicity. The $x$ value represents the primer attribute score, and the $weight$ value is the corresponding weight for that score.

For each of the two thresholds in this example, we add a sigmoidal factor in the weight calculation with its respective parameters

$$w(s|\theta) = \frac{1}{1+slope_{up}{}^{x_{up}-s}} \cdot \frac{1}{1+slope_{down}{}^{s-x_{down}}}.$$ **Equation 4 – Attribute weight**

Potentially, any number of sigmoid definitions can be multiplied to produce a complex function for converting score into weight.

Finally, the primer weight $W$ for the sequence $seq$ is calculated as the product of all primer attribute weights

$$W(seq) = \prod_{a\in A} w_a(s_a(seq)|\theta_a).$$ **Equation 5 – Primer weight**

Because each $w$ is treated independently, $W$ functions as a proxy for the joint probability of successful hybridization across all attributes. ADDTAG uses $W$ as an essential component for calculating how well a primer pair will amplify as intended (4.4.5).

### 4.4.5 Evaluating PrimerPairs

AddTag has the objective of identifying primer pairs that indicate the presence, or absence, of certain DNA sequences within a genome. Usually there are too many possible forward and reverse primer pairings to consider using a brute force method. For a typical locus, evaluating all possible primer pairs is a computationally expensive number of calculations. To lessen the burden, I implemented an iterative, threshold-relaxation process using the same framework for both Primer and PrimerPair weight evaluations. In the first iteration, AddTag requires PrimerPair attribute scores to be of the highest stringency. Each progressive iteration relaxes the hard thresholds for one or more attributes. These cutoff limitations can be specified by the user through command line options. This means that AddTag will first attempt to complete vPCR primer design using the narrowest attribute cutoffs, which is the least computationally-demanding. Then AddTag will attempt subsequent vPCR designs with lower levels of predicted compatibility, with each attempt taking more computational resources.

A few methods have been proposed to estimate the compatibility of a primer pair, most notably as the sum of weighted primer attribute scores [6, 12]. We chose to implement the novel product strategy over this method because the sum of weighted attribute scores is known to exclude a large proportion of valid primers [18].

To determine how well a pair of primers will amplify, AddTag calculates pair-specific attribute scores using the same principles it does for calculating single attribute scores. Then AddTag converts those scores into weights and combines them with the component Primer weights (Equation 5) to obtain the PrimerPair weight (Equation 6), discussed below. In chemical isolation, several considerations exist for guaranteeing oligonucleotide sequences do not form secondary structures, but can anneal to their template (Figure 4.7):

- the difference in $T_m$ between forward and reverse primers should be minimal;
- the minimum ΔG of the heterodimer should be as large as possible;
- the amplicon size (2 sigmoidal functions/types) should be within thermodynamically achievable range;
- and the maximum 3' heterodimer complementation length should be as short as possible.

Like individual Primer objects, PrimerPairs also have a weight that is the product of their attribute weights. The PrimerPair joint weight $W_J$ is the product of forward Primer weight $W(seq_F)$, the reverse Primer weight $W(seq_R)$, and the pair-specific attribute weights

$$W_J(seq_F, seq_R) = W(seq_F) \cdot W(seq_R) \cdot \prod_{a \in A} w_a(s_a(seq_F, seq_R)|\theta_a). \quad \textbf{Equation 6 – PrimerPair weight}$$

The greater the joint weight, the better the expected amplification. While it has been demonstrated that template sequence composition in between the forward and reverse primers does affect amplification efficiency [19-21], we have omitted calculating this for simplicity. For example, if amplification necessitates strand displacement, efficiency can be lowered [22].

### 4.4.6 Evaluating PrimerSets

Similar to the Primer and PrimerPair data structures, the PrimerSet data structure has specific attributes that are individually weighed and then multiplied together to form the final weight. The first set of attributes regard the non-redundant list of Primer sequences. We make a non-redundant set of Primers $SEQ$ from all $n$ PrimerParis. Each PrimerPair has a forward $seq_F$ sequence and a reverse $seq_R$ sequence.

$$SEQ = \{seq_{F_1}, seq_{R_1}, seq_{F_2}, seq_{R_2}, \dots , seq_{F_n}, seq_{R_n}\}.$$

We calculate the mean melting temperature $\overline{T_m}$, where $|SEQ|$ is the number of elements in $SEQ$

$$\overline{T_m} = \frac{\sum_{seq \in SEQ} T_m(seq)}{|SEQ|}.$$

Next, we calculate $W_P$ as the product of the weight $w$ of the difference between $T_m$ and $\overline{T_m}$ across all elements of $SEQ$

$$W_P = \prod_{seq \in SEQ} w(T_m(seq) - \overline{T_m}\,).$$

Then we calculate one attribute based on the PrimerPair data structure—the product of the joint weights $W_J$, or if $(seq_F, seq_R)$ of a PrimerPair is repeated, the average joint weight $\overline{W_J}$ of those repeated PrimerPairs as a simplification. $PP$ contains the PrimerPair objects

$$[(seq_F, seq_R)_1 \quad (seq_F, seq_R)_2 \quad \cdots \quad (seq_F, seq_R)_n] \in PP,$$

And can be re-written as

$$[pp_1 \quad pp_2 \quad \cdots \quad pp_n] \in PP.$$

The weight representing the PrimerPair, taking redundancy into account, is thus

$$W_{PP} = \prod_{pp \in PP} \overline{W_J(pp)}.$$

Finally, the weight of a PrimerSet $W_{\check{\ss}}$ is calculated as the product of its two components $W_P$ and $W_{PP}$ as follows:

$$W_{\check{\ss}} = W_P \cdot W_{PP}. \quad \textbf{Equation 7 – PrimerSet weight}$$

The weight of the PrimerSet $W_{\check{\ss}}$ therefore incorporates all PrimerPair joint sequence weights with of an additional melting temperature constraint. This process allows for simpler $W_{\check{\ss}}$ calculation by relying on pre-computed $W_J$ at the cost of the one replicated attribute involving $T_m$, thereby artificially increasing its net importance. Because $W_P$ relies on a non-redundant list of sequences, a primer oligonucleotide possessing multiple pairs is only penalized a single time. Also, because $W_{PP}$ relies on a non-redundant list of paired primers, duplicate pairs are effectively counted only once. Together, the $W_{\check{\ss}}$ weight encourages calculating a minimal set of primers that are compatible with each other.

## 4.5 Discussion

### 4.5.1 Summary

AddTag is a generalized approach for developing CRISPR/Cas genome editing experiments. After each step, the genomes are assayed for CRISPR/Cas-induced recombination events with a PCR-based assay [23]. Other software packages that choose gRNA Targets and design verification primers (Table 3.2, "vPCR primers") are confined to a single editing step and genome [24]. Although we specifically limit the discussion in this chapter, as well as the empirical validations (Chapter 2), to 2-step editing in C. *albicans*, ADDTAG primer design can span any number of serial genome editing steps, and is applicable to any sequenced genome.

ADDTAG finds sets of compatible vPCR primers for validating multi-step genome edits. ADDTAG implements a core subroutine designed for vPCR primer design that has 3 parts: (A) predicting the genome sequence after each editing step, (B) identifying primer pairs that would be useful for verification of intended genome edits, and (C) optimization of primer pairs. ADDTAG uses the reference genome and the dDNA sequences as input. ADDTAG includes the ability to discriminate between allele-specific, multi-allelic, and allele-agnostic vPCR amplification designs.

For some loci edited in Chapter 2—*ADE2*$_{CDS}$, *EFG1*$_{CDS}$, *BRG1*$_{CDS}$, *ZRT2*$_{US}$, *WOR1*$_{USd}$, *WOR1*$_{USp}$, and *WOR2*$_{DS}$—no usable ΔoF and ΔoR primers were identified (Table 0.3). However, ADDTAG identified usable ΔoF and ΔoR primers for the *ZAP1*$_{US}$ locus. vPCR Analyses of other loci, such as *FLO8*$_{CDS}$, *RBF1*$_{CDS}$, and *HOT1*$_{CDS}$ revealed that identification of usable ΔoF and ΔoR is generally mutually exclusive (Table 4.4, "Insert" column). Mutually exclusive sF/ΔoR and ΔoF/sR amplification arose due to the hybridization constraints with the sF/sR primer pairing. Of the loci tested, only *ZAP1*$_{US}$ had both an Insert junction forward strand (origin of ΔoF) with sufficient similarity to the far upstream region (origin of sF) and an Insert junction reverse strand (origin of ΔoR) with sufficient similarity to the far downstream region (origin of sR) to allow for identification of sF/ΔoR and ΔoF/sR simultaneously.

In addition to ensuring that each individual primer meets or exceeds all of the evaluation metrics, the ADDTAG ensures the vPCR primers for validating either 1-step or 2-step genome editing experiments are optimized to work under identical conditions. Because all of the selected primers are co-optimized (Figure 2.10), any of the primer pair combinations used in genotype verification (Figure 2.11) can be run in parallel in the same thermal cycler, thus enabling higher throughput. This PCR primer co-optimization step therefore improves efficiency and reduces complexity during the genotype verification. All strains generated in this study (Chapter 2, Table 0.1) were constructed using Targets, dDNAs, and PCR primers that were exclusively designed by the automated ADDTAG software (Table 0.3). In all cases, the desired genome edits were obtained in both biological replicates, and the expected banding patterns were observed from PCR-based genotype verifications (for a representative example see Figure 2.26).

### 4.5.2 Future directions

The ADDTAG software could be improved by increasing its functionality, efficiency, and accuracy. For example, a future version of ADDTAG could function to predict the probability of each dDNA integration into the genome during *in silic*o recombination. For ease of implementation, mutual exclusivity of primer annealing sites was ignored (4.3.1); however, taking this into account could increase the diagnostic power of the vPCR primers. Additionally, ADDTAG relies on exact primer sequence matches in the genome during *in silico* PCR. This means that ADDTAG does not quantify non-specific transient primer-template binding, which can somewhat affect the proportion of time the primer binds to the intended site, and thus the amplification efficiency [25-27]. Also, PCR often demonstrates technical variance in efficiency [28], and ADDTAG makes no attempt to model that variance. Furthermore, ADDTAG would ideally generate both a high-quality Target (Chapter 3) and a high-quality ΔoF or ΔoR primer annealing site on the Step 1 dDNA, thereby requiring vPCR identification to occur concurrently with dDNA generation. This type of improvement would require extensive refactoring. Moreover, replacing the current, brute-force primer calculations, which rely on computational power to evaluate every possibility, with a heuristic could provide vast computer time and memory savings. Finally, ADDTAG accuracy could improve if it used a more nuanced primer compatibility calculation, such as one that incorporates the primer annealing temperature. Despite these shortcomings, ADDTAG still computes quality vPCR primer designs that serve to evaluate genome editing experiments (Chapter 2).

## 4.6 References

1. Seher TD, Nguyen N, Ramos D, Bapat P, Nobile CJ, Sindi SS, Hernday AD. AddTag, a two-step approach with supporting software package that facilitates CRISPR/Cas-mediated precision genome editing. *G3 Genes|Genomes|Genetics*. 2021. DOI: 10.1093/g3journal/jkab216.
2. Storici F, Lewis LK, Resnick MA. *In vivo* site-directed mutagenesis using oligonucleotides. *Nature Biotechnology*. 2001; 19(8):773-6. DOI: 10.1038/90837, PMID: 11479573.
3. Marton T, Maufrais C, d'Enfert C, Legrand M, Mitchell AP. Use of CRISPR-Cas9 to target homologous recombination limits transformation-induced genomic changes in *Candida albicans*. *mSphere*. 2020; 5(5):e00620-20. DOI: 10.1128/mSphere.00620-20, PMID: 32878930.
4. Loll-Krippleber R, Feri A, Nguyen M, Maufrais C, Yansouni J, d'Enfert C, Legrand M. A FACS-optimized screen identifies regulators of genome stability in *Candida albicans*. *Eukaryotic Cell*. 2015; 14(3):311-22. DOI: 10.1128/EC.00286-14, PMID: 25595446.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology*. 1990; 215(3):403-10. DOI: 10.1016/S0022-2836(05)80360-2, PMID: 2231712.
6. Kalendar R, Lee D, Schulman AH. FastPCR software for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. In: Valla S, Lale R, editors. DNA Cloning and Assembly Methods. Totowa, NJ: Humana Press; 2014; p. 271-302. DOI: 10.1007/978-1-62703-764-8_18.
7. Rychlik W. Selection of Primers for Polymerase Chain Reaction. In: White BA, editor. PCR Protocols: Current Methods and Applications. Totowa, NJ: Humana Press; 1993; p. 31-40. DOI: 10.1385/0-89603-244-2:31.
8. Yuryev A. PCR primer design: Springer Science & Business Media; 2007. DOI: 10.1007/978-1-59745-528-2.
9. Rozen S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers. In: Misener S, Krawetz SA, editors. Bioinformatics Methods and Protocols. Totowa, NJ: Humana Press; 1999; p. 365-86. DOI: 10.1385/1-59259-192-2:365.
10. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. In: Keith JM, editor. Bioinformatics: Structure, Function and

Applications. Totowa, NJ: Humana Press; 2008; p. 3-31. DOI: 10.1007/978-1-60327-429-6_1.

11. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. *Algorithms for Molecular Biology*. 2011; 6(1):26. DOI: 10.1186/1748-7188-6-26, PMID: 22115189.
12. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3 – new capabilities and interfaces. *Nucleic Acids Research*. 2012; 40(15):e115-e. DOI: 10.1093/nar/gks596, PMID: 22730293.
13. Rychlik W, Spencer WJ, Rhoads RE. Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Research*. 1990; 18(21):6409-12. DOI: 10.1093/nar/18.21.6409, PMID: 2243783.
14. SantaLucia JJ. Physical principles and Visual-OMP software for optimal PCR design. In: Yuryev A, editor. PCR primer design. 402: Springer Science & Business Media; 2007; p. 3-33. DOI: 10.1007/978-1-59745-528-2_1.
15. Dieffenbach C, Lowe T, Dveksler G. General concepts for PCR primer design. *PCR methods appl*. 1993; 3(3):S30-S7. DOI: 10.1101/gr.3.3.s30, PMID: 8118394.
16. Marky LA, Canuel L, Jones RA, Breslauer KJ. Calorimetric and spectroscopic investigation of the helix-to-coil transition of the self-complementary deoxyribonucleotide ATGCAT. *Biophysical Chemistry*. 1981; 13(2):141-9. DOI: 10.1016/0301-4622(81)80013-0, PMID: 7260332.
17. Breslauer KJ, Frank R, Blöcker H, Marky LA. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*. 1986; 83(11):3746-50. DOI: 10.1073/pnas.83.11.3746, PMID: 3459152.
18. Mann T, Humbert R, Dorschner M, Stamatoyannopoulos J, Noble WS. A thermodynamic approach to PCR primer design. *Nucleic acids research*. 2009; 37(13):e95-e. Epub 2009/06/15. DOI: 10.1093/nar/gkp443, PMID: 19528077.
19. Mamedov TG, Pienaar E, Whitney SE, TerMaat JR, Carvill G, Goliath R, Subramanian A, Viljoen HJ. A fundamental study of the PCR amplification of GC-rich DNA templates. *Computational Biology and Chemistry*. 2008; 32(6):452-7. DOI: 10.1016/j.compbiolchem.2008.07.021, PMID: 18760969.
20. McDowell DG, Burns NA, Parkes HC. Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. *Nucleic Acids Research*. 1998; 26(14):3340-7. DOI: 10.1093/nar/26.14.3340, PMID: 9649616.
21. Rose JA, Komiya K, Yaegashi S, Hagiya M, editors. Displacement Whiplash PCR: Optimized Architecture and Experimental Validation. *International Workshop on DNA-Based Computers*; 2006; Berlin, Heidelberg: Springer. DOI: 10.1007/11925903_31.
22. Ignatov KB, Barsova EV, Fradkov AF, Blagodatskikh KA, Kramarova TV, Kramarov VM. A strong strand displacement activity of thermostable DNA polymerase markedly improves the results of DNA amplification. *BioTechniques*. 2014; 57(2):81-7. DOI: 10.2144/000114198, PMID: 25109293.
23. Kim H-S, Smithies O. Recombinant fragment assay for gene targetting based on the polymerase chain reaction. *Nucleic Acids Research*. 1988; 16(18):8887-903. DOI: 10.1093/nar/16.18.8887, PMID: 3174435.
24. Rodríguez-López M, Cotobal C, Fernández-Sánchez O, Borbarán Bravo N, Oktriani R, Abendroth H, Uka D, *et al*. A CRISPR/Cas9-based method and primer design tool for seamless genome editing in fission yeast. *Wellcome Open Research*. 2017; 1(19). DOI: 10.12688/wellcomeopenres.10038.3, PMID: 28612052.
25. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC*

*Bioinformatics*. 2012; 13(1):134. DOI: 10.1186/1471-2105-13-134, PMID: 22708584.

26. Qu W, Zhou Y, Zhang Y, Lu Y, Wang X, Zhao D, Yang Y, Zhang C. MFEprimer-2.0: A fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Research*. 2012; 40(W1):W205-W8. DOI: 10.1093/nar/gks552, PMID: 22689644.
27. So KYK, Fong JJ, Lam IPY, Dudgeon D. Pitfalls during in silico prediction of primer specificity for eDNA surveillance. *Ecosphere*. 2020; 11(7):e03193. DOI: 10.1002/ecs2.3193.
28. Svec D, Tichopad A, Novosadova V, Pfaffl MW, Kubista M. How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. *Biomolecular Detection and Quantification*. 2015; 3:9-16. DOI: 10.1016/j.bdq.2015.01.005, PMID: 27077029.

# Appendices

# Appendix A Materials

## A.1 Software versions, database resources, and computer specifications

We used ADDTAG r284 and ADDTAG r517 in PYTHON 3.5.3 with the REGEX 2018.2.21 package, and with the following additional software for scoring and aligning: AZIMUTH 2.0 [1] in PYTHON 2.7.13 for on-target scores; CFD [1] and HSU-ZHANG [2] for off-target scores; BLAST+ 2.7.1 [3] for predicting recombination by aligning dDNA to gDNA; BOWTIE 2 2.3.4.1 [4] for aligning target sequences to off-target sites; MAFFT [5] for identifying homologous flanking regions; and UNAFOLD 3.8 [6] for the change in Gibbs' free energy and melting temperature thermodynamics calculations. We used the C. *albicans* assembly 22 sequence and annotations from the *Candida* Genome Database [7] retrieved on February 5, 2017. Oligo designs were computed on a Linux 3.1.0 64-bit Slurm-managed system with 256 Gb RAM and a 12-core (24-logical processors) Intel Xeon E5-2650 v4 @ 2.20 GHz x86 CPU. Analyses were conducted using the BASH shell [8]. Full commands to reproduce the analysis are included in the code repository.

## A.2 Strains used in this study

**Table 0.1 – Strains used in this study**

| Identifier | Strains | Genotype | Source |
|---|---|---|---|
| $ADE2_{CDS}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{ADE2_{CDS}}{ADE2_{CDS}}$ | [9] |
| $ade2_{CDS}$ $\Delta/\Delta$ | AHY1338,<br>AHY1347 | $\frac{leu2\Delta}{LEU2}\ \frac{ade2_{cds}\Delta\text{::mintag(CC)}}{ade2_{cds}\Delta\text{::mintag(CC)}}$ | (This study) |
| $ADE2_{CDS}$ AB/AB | AHY1267,<br>AHY1268 | $\frac{leu2\Delta}{LEU2}\ \frac{ade2_{cds}\Delta\text{::}ADE2_{CDS}}{ade2_{cds}\Delta\text{::}ADE2_{CDS}}$ | (This study) |
| $BRG1_{CDS}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{BRG1_{CDS}}{BRG1_{CDS}}$ | [9] |
| $brg1_{CDS}$ $\Delta/\Delta$ | AHY1219,<br>AHY1220 | $\frac{leu2\Delta}{LEU2}\ \frac{brg1_{cds}\Delta\text{::mintag()}}{brg1_{cds}\Delta\text{::mintag()}}$ | (This study) |
| $BRG1_{CDS}$ AB/AB | AHY1263,<br>AHY1264 | $\frac{leu2\Delta}{LEU2}\ \frac{brg1_{cds}\Delta\text{::}BRG1_{CDS}}{brg1_{cds}\Delta\text{::}BRG1_{CDS}}$ | (This study) |
| $EFG1_{CDS}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{EFG1_{CDS}}{EFG1_{CDS}}$ | [9] |
| $efg1_{CDS}$ $\Delta/\Delta$ | AHY1336,<br>AHY1337 | $\frac{leu2\Delta}{LEU2}\ \frac{efg1_{cds}\Delta\text{::mintag(CCTC)}}{efg1_{cds}\Delta\text{::mintag(CCTC)}}$ | (This study) |
| $EFG1_{CDS}$ AB/AB | AHY1259,<br>AHY1260 | $\frac{leu2\Delta}{LEU2}\ \frac{efg1_{cds}\Delta\text{::}EFG1_{CDS}}{efg1_{cds}\Delta\text{::}EFG1_{CDS}}$ | (This study) |
| $ZAP1_{CDS}$ +/+ | DAY185 | $\frac{ura3\Delta\text{::}\lambda imm434}{ura3\Delta\text{::}\lambda imm434}\ \frac{ARG4\text{::}URA3\text{::}arg4\text{::}hisG}{arg4\text{::}hisG}\ \frac{his1\text{::}hisG\text{::}pHIS1}{his1\text{::}hisG}$ | [10] |
| $zap1_{CDS}$ $\Delta/\Delta$ | CJN1201 | $\frac{ura3\Delta\text{::}\lambda imm434}{ura3\Delta\text{::}\lambda imm434}\ \frac{arg4\text{::}hisG}{arg4\text{::}hisG}\ \frac{his1\text{::}hisG\text{::}pHIS1}{his1\text{::}hisG}\ \frac{zap1\text{::}ARG4}{zap1\text{::}URA3}$ | [11] |
| $ZAP1_{US}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{CSR1_{US}}{CSR1_{US}}$ | [9] |
| $zap1_{US}$ $\Delta/\Delta$ | AHY1348,<br>AHY1349 | $\frac{leu2\Delta}{LEU2}\ \frac{csr1_{us}\Delta\text{::addtag(CGTACGCTGCAGGTCGACAGTGG)}}{csr1_{us}\Delta\text{::addtag(CGTACGCTGCAGGTCGACAGTGG)}}$ | (This study) |
| $ZAP1_{US}$ $AB^0/AB^0$ | AHY1265,<br>AHY1266 | $\frac{leu2\Delta}{LEU2}\ \frac{csr1_{us}\Delta\text{::}CSR1_{US}}{csr1_{us}\Delta\text{::}CSR1_{US}}$ | (This study) |
| $ZAP1_{US}$ $AB^1/AB^1$ | AHY1269,<br>AHY1270 | $\frac{leu2\Delta}{LEU2}\ \frac{csr1_{us}\Delta\text{::}CSR1^{1}_{US}}{csr1_{us}\Delta\text{::}CSR1^{1}_{US}}$ | (This study) |
| $ZRT2_{US}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{ZRT2_{US}}{ZRT2_{US}}$ | [9] |
| $zrt2_{US}$ $\Delta/\Delta$ | AHY1221,<br>AHY1222 | $\frac{leu2\Delta}{LEU2}\ \frac{zrt2_{us}\Delta\text{::addtag(CGTACGCTGCAGGTCGACAGTGG)}}{zrt2_{us}\Delta\text{::addtag(CGTACGCTGCAGGTCGACAGTGG)}}$ | (This study) |
| $ZRT2_{US}$ $AB^{00}/AB^{00}$ | AHY1261,<br>AHY1262 | $\frac{leu2\Delta}{LEU2}\ \frac{zrt2_{us}\Delta\text{::}ZRT2_{US}}{zrt2_{us}\Delta\text{::}ZRT2_{US}}$ | (This study) |

| Identifier | Strains | Genotype | Source |
|---|---|---|---|
| $ZRT2_{US}$ AB$^{01}$/AB$^{01}$ | AHY1271,<br>AHY1272 | $\frac{leu2\Delta}{LEU2}\ \frac{zrt2_{us}\Delta::ZRT2^{01}_{US}}{zrt2_{us}\Delta::ZRT2^{01}_{US}}$ | (This study) |
| $ZRT2_{US}$ AB$^{10}$/AB$^{10}$ | AHY1273,<br>AHY1274 | $\frac{leu2\Delta}{LEU2}\ \frac{zrt2_{us}\Delta::ZRT2^{10}_{US}}{zrt2_{us}\Delta::ZRT2^{10}_{US}}$ | (This study) |
| $ZRT2_{US}$ AB$^{11}$/AB$^{11}$ | AHY1295,<br>AHY1296 | $\frac{leu2\Delta}{LEU2}\ \frac{zrt2_{us}\Delta::ZRT2^{11}_{US}}{zrt2_{us}\Delta::ZRT2^{11}_{US}}$ | (This study) |
| $WOR1_{USd}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{WOR1_{USd}}{WOR1_{USd}}$ | [9] |
| $wor1_{USd}$ Δ/Δ | AHY1447,<br>AHY1448 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usd}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}{wor1_{usd}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}$ | (This study) |
| $WOR1_{USd}$ AB$^{0}$/AB$^{0}$ | AHY1449,<br>AHY1450 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usd}\Delta::WOR1_{USd}}{wor1_{usd}\Delta::WOR1_{USd}}$ | (This study) |
| $WOR1_{USd}$ AB$^{1}$/AB$^{1}$ | AHY1451,<br>AHY1452 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usd}\Delta::WOR1^{1}_{USd}}{wor1_{usd}\Delta::WOR1^{1}_{USd}}$ | (This study) |
| $WOR1_{USp}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{WOR1_{USp}}{WOR1_{USp}}$ | [9] |
| $wor1_{USp}$ Δ/Δ | AHY1453,<br>AHY1454 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usp}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}{wor1_{usp}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}$ | (This study) |
| $WOR1_{USp}$ AB$^{0}$/AB$^{0}$ | AHY1455,<br>AHY1456 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usp}\Delta::WOR1_{USp}}{wor1_{usp}\Delta::WOR1_{USp}}$ | (This study) |
| $WOR1_{USp}$ AB$^{1}$/AB$^{1}$ | AHY1457,<br>AHY1458 | $\frac{leu2\Delta}{LEU2}\ \frac{wor1_{usp}\Delta::WOR1^{1}_{USp}}{wor1_{usp}\Delta::WOR1^{1}_{USp}}$ | (This study) |
| $WOR2_{DS}$ +/+ | AHY940 | $\frac{leu2\Delta}{LEU2}\ \frac{WOR2_{DS}}{WOR2_{DS}}$ | [9] |
| $WOR2_{DS}$ Δ/Δ | AHY1459,<br>AHY1460 | $\frac{leu2\Delta}{LEU2}\ \frac{wor2_{ds}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}{wor2_{ds}\Delta::\text{addtag(CGTACGCTGCAGGTCGACAGTGG)}}$ | (This study) |
| $WOR2_{DS}$ AB$^{0}$/AB$^{0}$ | AHY1461,<br>AHY1462 | $\frac{leu2\Delta}{LEU2}\ \frac{wor2_{ds}\Delta::WOR2_{DS}}{wor2_{ds}\Delta::WOR2_{DS}}$ | (This study) |
| $WOR2_{DS}$ AB$^{1}$/AB$^{1}$ | AHY1463,<br>AHY1464 | $\frac{leu2\Delta}{LEU2}\ \frac{wor2_{ds}\Delta::WOR2^{1}_{DS}}{wor2_{ds}\Delta::WOR2^{1}_{DS}}$ | (This study) |

The identifier is the term used to represent this strain in the manuscript. Each strain name listed for an identifier represents a wholly independent biological derivation of the strain. Genotypes are listed through the standard *Candida* syntax: homologous chromosomes are separated by horizontal lines; linked loci are co-located in the same word, and non-linked loci are separated by white space; loci are represented by the text of their wild-type gene name; wild-type genes are capitalized, and mutant genes are lower-case; and specific modifications to genes at a locus begin with "::". Superscript 0 and 1 numbers represent allelic forms, with 0 being unmodified from wild-type, and 1 being modified. Subscript CDS indicates the Feature of interest is only a portion of the full chromosomal gene, and subscript US represents a section of the upstream intergenic region of the gene. The sequence orientations for `mintag` and `addtag` inserts are indicated with the start as lower contig position and the end as higher contig position, as they appear in the reference genome (i.e. "+" orientation relative to the contig).

## A.3 Plasmids used in this study

### Table 0.2 – Plasmids used in this study

| Identifier | Purpose | Ordering information | Sequence | Source |
|---|---|---|---|---|
| pADH110 | NAT-marked pSNR52 promoter fragment for use in gRNA expression cassette stitching PCR; for use with pADH119 to generate Target-specific gRNA expression cassette. | https://www.addgene.org/90982/ | GBK | [9] |
| pADH119 | NAT-marked "empty" gRNA construct for use in gRNA expression cassette stitching PCR. Use with pADH110 to generate Target-specific part 2 of 2 for C. *albicans* LEUpOUT CRISPR system. | https://www.addgene.org/90985/ | GBK | [9] |
| pADH137 | NAT-marked Cas9 expression construct; part 1 of 2 of C. *albicans* LEUpOUT CRISPR system. Use with pADH118-series gRNA expression constructs. | https://www.addgene.org/90986/ | GBK | [9] |

The identifier is the term used to represent this plasmid in the source manuscript. Each plasmid is available to order from Addgene at the provided web address. Full sequence information is available, but is directly linked in the "Sequence" column.

## A.4 Single stranded oligonucleotide and synthetic DNA sequences

### Table 0.3 – Single stranded oligonucleotide and synthetic DNA sequences

| Identifier | Sequence (5'→3') | Purpose | Source |
|---|---|---|---|
| AHO2144 | CTAAGAAGGGAAAAGCACCAC | *ADE2*$_{CDS}$ AmpF | (This study) |
| AHO2145 | CTCGGTACAATCTTGTCAATGAG | *ADE2*$_{CDS}$ AmpR | (This study) |
| AHO2137 | GTGGTGGATTGGTATTTCTTTCTGTG | *ADE2*$_{CDS}$ sF | (This study) |
| AHO2138 | AAGACCCCAAACATTTTGACTCG | *ADE2*$_{CDS}$ sR | (This study) |
| AHO2142 | CCCCAATGTGTAACAAGTCATCG | *ADE2*$_{CDS}$ +/AoF | (This study) |
| AHO2139 | CATTGCCTGTCATTGGTGTTCC | *ADE2*$_{CDS}$ +/AoR | (This study) |
| AHO2140 | CAGAGTTGTGAGGTCTTGGTG | *ADE2*$_{CDS}$ +/AiF | (This study) |
| AHO2141 | GGCGTATGATGGTAGAGGTAAC | *ADE2*$_{CDS}$ +/AiR | (This study) |
| AHO2135 | CACCATAACGTTTACTTGTTTAATATGCTATTGATATCTATATTTTTTTCCTATGTGTAGTGCTTGTATATGCGTGTGTGATGAGAATAAGATGAATAGA | *ADE2*$_{CDS}$ Step 1 dDNA F | (This study) |
| AHO2136 | TCTATTCATCTTATTCTCATCACACACGCATATACAAGCACTACACATAGGAAAAAAATATAGATATCAATAGCATATTAAACAAGTAAACGTTATGGTG | *ADE2*$_{CDS}$ Step 1 dDNA R | (This study) |
| AHO2134 | CGTAAACTATTTTTAATTTGAACACCAATGACAGGCAATGGTTTTAGAGCTAGAAATAGC | *ADE2*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2143 | CGTAAACTATTTTTAATTTGCATATACAAGCACTACACATGTTTTAGAGCTAGAAATAGC | *ADE2*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| AHO2168 | CACATTAGTTGCTCAGGTCAC | *EFG1*$_{CDS}$ AmpF | (This study) |
| AHO2169 | GTCAATGGATTTGGGAGAAGA | *EFG1*$_{CDS}$ AmpR | (This study) |
| AHO2161 | TTAACCCCTTTGTGTCCCTT | *EFG1*$_{CDS}$ sF | (This study) |
| AHO2162 | CCCAAATAGTATAAATTCGTTCATGTC | *EFG1*$_{CDS}$ sR | (This study) |
| AHO2166 | ACCAATCACCCCAAGTTCAG | *EFG1*$_{CDS}$ +/AoF | (This study) |
| AHO2163 | GCTGTTGTTGTTGTTGTCCT | *EFG1*$_{CDS}$ +/AoR | (This study) |
| AHO2164 | CCCCCATACCTTCCAATTCTAC | *EFG1*$_{CDS}$ +/AiF | (This study) |
| AHO2165 | GACACATTACTGCCACCACTG | *EFG1*$_{CDS}$ +/AiR | (This study) |
| AHO2159 | AACGAATTAAGATTTGTTCTATTTGACTACCAAGAATATAACCCATATTCCTCGTGTACATCACCTTCTGCTTTCTGCCATAAATTCCAAATTAGATTAT | *EFG1*$_{CDS}$ Step 1 dDNA F | (This study) |
| AHO2160 | ATAATCTAATTTGGAATTTATGGCAGAAAGCAGAAGGTGATGTACACGAGGAATATGGGTTATATTCTTGGTAGTCAAATAGAACAAATCTTAATTCGTT | *EFG1*$_{CDS}$ Step 1 dDNA R | (This study) |

| Identifier | Sequence (5'→3') | Purpose | Source |
|---|---|---|---|
| AHO2158 | CGTAAACTATTTTTAATTTGTGGTTGGAATTGCCCCACAGGTTTTAGAGCTAGAAATAGC | *EFG1*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2167 | CGTAAACTATTTTTAATTTGAGCAGAAGGTGATGTACACGGTTTTAGAGCTAGAAATAGC | *EFG1*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| AHO2156 | TATAAATATCAGGTCATAGATCCCTG | *BRG1*$_{CDS}$ AmpF | (This study) |
| AHO2157 | CTGCTACAGTATTGTTGTTTGAAC | *BRG1*$_{CDS}$ AmpR | (This study) |
| AHO2149 | TGCAGCTTTTGTACTACATTTGG | *BRG1*$_{CDS}$ sF | (This study) |
| AHO2150 | CCAGCTCAGGATATAATTTACAGC | *BRG1*$_{CDS}$ sR | (This study) |
| AHO2154 | GTCATTCATCAACCACCACCA | *BRG1*$_{CDS}$ +/AoF | (This study) |
| AHO2151 | ACCTCCACTAATGGTTGATCG | *BRG1*$_{CDS}$ +/AoR | (This study) |
| AHO2152 | CCACCACAACAACCACAATCAG | *BRG1*$_{CDS}$ +/AiF | (This study) |
| AHO2153 | CGACCGTTCTTCCCTTTTGTC | *BRG1*$_{CDS}$ +/AiR | (This study) |
| AHO2147 | GTACTACTGTTCATATTTGATATTTCAACGTTATTTCTCCATCCATACTTCTGGCGGTATTCCTGTTGCTTACCCAACCCAAATTCCTTTAATTCGTCAT | *BRG1*$_{CDS}$ Step 1 dDNA F | (This study) |
| AHO2148 | ATGACGAATTAAAGGAATTTGGGTTGGGTAAGCAACAGGAATACCGCCAGAAGTATGGATGGAGAAATAACGTTGAAATATCAAATATGAACAGTAGTAC | *BRG1*$_{CDS}$ Step 1 dDNA R | (This study) |
| AHO2146 | CGTAAACTATTTTTAATTTGGGGCTAAGTGACGATGCAGGGTTTTAGAGCTAGAAATAGC | *BRG1*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2155 | CGTAAACTATTTTTAATTTGAATACCGCCAGAAGTATGGAGTTTTAGAGCTAGAAATAGC | *BRG1*$_{CDS}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| AHO 2196 | GTGAACCACTCATCATCATTGG | *ZAP1*$_{US}$ AmpF | (This study) |
| AHO 2199 | ACGACTAATGCTATGACTGCTC | *ZAP1*$_{US}$ AmpR | (This study) |
| AHO 2185 | CTGTGATCGTGATTATGAATGTGGC | *ZAP1*$_{US}$ sF | (This study) |
| AHO 2186 | ACGTTGTTCGTCTCAAGCTGG | *ZAP1*$_{US}$ sR | (This study) |
| AHO 2190 | GTTGTCGATGATGATGATGCTGG | *ZAP1*$_{US}$ +/AoF | (This study) |
| AHO 2187 | TACCCAGCATCATCATCATCG | *ZAP1*$_{US}$ +/AoR | (This study) |
| AHO 2188 | CGATGATGATGATGCTGGGTA | *ZAP1*$_{US}$ +/AiF | (This study) |
| AHO 2189 | GTGTTACTTGGTAGCACTTTGATC | *ZAP1*$_{US}$ +/AiR | (This study) |
| AHO 2191 | AGAAAGTGGCGTTTAATAAATACTACGTAC | *ZAP1*$_{US}$ ΔoF | (This study) |
| AHO 2192 | TGGTGTTACTTGGTAGCCCAC | *ZAP1*$_{US}$ ΔoR | (This study) |
| AHO 2183 | GTTTAAATTGATAGTATAATCTAAAATAAGAAAGTGGCGTTTAATAAATACTACGTACGCTGCAGGTCGACAG | *ZAP1*$_{US}$ Step 1 dDNA F | (This study) |
| AHO 2184 | TTGATGCAATATTGTTCTTGTTGAATGTAATGCCGTGTGGTGTTACTTGGTAGCCCACTGTCGACCTGCAGCGTAC | *ZAP1*$_{US}$ Step 1 dDNA R | (This study) |
| SynFrag 1 | GTGAACCACTCATCATCATTGGCATTACCCTTGGTATATCTTTTAGCATATAATGAAGTTTAAATTGATAGTATAATCTAAAATAAGAAAGTGGCGTTTAATAAATACTACCTGAGGAATACGTTTTCTCCTCTTTAAAAATGAAATAAAAAGATCCTCTTATACTATTAAAGAAAAGAAAAAAAGAAAAAATTTCTTTCCAAAAAGTATTATTGTTGTTGTTGTCGATGATGATGATGCTGGGTATAGTATAGTATAGTGATAAATGAATGAAAATTACAACTGTAGGGAAGAAGAAATAATAATTAAAGGTTGCAATGcagctgttctaTATGCCAATTTTCGATTTTGTTCAATTTTTTTTTTCCGGTGCTGGTGGGTGAGAGAGAAGATTAAATTAAATTAAATTGGGTGATTCACTTTTACTTTTACTTTACAATGAATTTTTCTTCTTGTTCTTCTTCTTATTGTTGTTATTGTAAAGGGATTTACTTCTAAATTAAGATACGTCGTGTATAGATGATCAAAGTGCTACCAAGTAACACCACACGGCATTACATTCAACAAGAACAATATTGCATCAAGAAGTTAATATTTTCAAACTTTTCTAAAAGGGGAGCAGTCATAGCATTAGTCGT | *ZAP1*$_{US}$ Step 2 dDNA for AB[1] | (This study) |
| AHO 2182 | CGTAAACTATTTTTAATTTGGAAAATTGGCATAaccaccaGTTTTAGAGCTAGAAATAGC | *ZAP1*$_{US}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO 2195 | CGTAAACTATTTTTAATTTGCGTACGCTGCAGGTCGACAGGTTTTAGAGCTAGAAATAGC | *ZAP1*$_{US}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| AHO2209 | GCATATTTACTTGCTTGCCTG | *ZRT2*$_{US}$ AmpF | (This study) |
| AHO2214 | TTGACAGGAATATGGAGGGTA | *ZRT2*$_{US}$ AmpR | (This study) |
| AHO2203 | GAACCAATCCTTCCACATAGC | *ZRT2*$_{US}$ sF | (This study) |
| AHO2204 | GCTGGGAATTGATAATGAAAGC | *ZRT2*$_{US}$ sR | (This study) |
| AHO2208 | TATTGGTCGGATTGGGTTAC | *ZRT2*$_{US}$ +/AoF | (This study) |
| AHO2205 | TTGCGTTTCGGGTATAATCAC | *ZRT2*$_{US}$ +/AoR | (This study) |
| AHO2206 | GAGAAGAACCATAAAGTCCAAGC | *ZRT2*$_{US}$ +/AiF | (This study) |
| AHO2207 | CACCTCAAACCACACACTAC | *ZRT2*$_{US}$ +/AiR | (This study) |

| Identifier | Sequence (5'→3') | Purpose | Source |
|---|---|---|---|
| AHO2201 | CGATATTGTGTAATTTTACATTTGGGCACAGCATAGCCTGATGCCGTCCGGGTCGTACGCTGCAGGTCGACAG | *ZRT2*$_{US}$ Step 1 dDNA F | (This study) |
| AHO2202 | TGGTGATGGTTTTTATTAGTGGTTACAAAAATGAACAAGAGAAAATTTGCAATACCACTGTCGACCTGCAGCGTAC | *ZRT2*$_{US}$ Step 1 dDNA R | (This study) |
| SynFrag1 | GCATATTTACTTGCTTGCCTGATATCCTCGACTCATATACTTTGTAAATTACCTGTCACGTGTTTTTGTGAACTCCGATATTGTGTAATTTTACATTTGGGCACAGCATAGCCTGATGCCGTCCGGGT**accctggtagt**TATCACTCAATTTTTTTTTGTTTTTCACTGTTTTTCTGTCTTGTTGTTCCAAATAACCACTAATATTTCTCTTATACTTGACGATTTTTGGTGACCTATTATAGCTGGCAAGTGAAAGTGAATTAATAATATGCATTTTATAAAGTAGGCTTATTCATAAAATAATTAATTATTATTCAATCTCTAATTGATGTTCAGAAAATTTTTGGTTTGATGCCATACAAAGCAAAAAAAAAAATAATACATCAAAAATAGAACAAATGTAACTTTATGGTATTAAATCGTAATCATACACTTACTGAGAAGAACCATAAAGTCCAAGCTTTATAGAAAAAAGGCTAATGTTCTTTAGCATATGGTTTTTTTATGTCCTGGATTAACAACGTCCTTGGACTTAAGTACGTATGAAAGAACTAGCTAATAATTTAAAGCCAAACTGAGTCTTTCAACAACTACAATAGTGATTATACCCGAAACGCAAAATAATAAAAACTAATATTGACAATTGAATTATTCATTATTGGTCGGATTGGGTTACATTCAGATTGAAATCACGGTTGTAATTGCCGAATCTCTTTTTCATTGTTGTTCCATTTGTAACATTACCAGCTAGAAATGTAGTGTGTGGTTTGAGGTGCGTTTAGA**cagctgttcta**TATTGCAAATTTTCTCTTGTTCATTTTTGTAACCACTAATAAAAACCATCACCAATTGACAATGAGTAAAAACTTTAAAAAAAAAGTAAAAATTAGAAAGAAAAAGTCAATCTCCCTTTTGTTGTAATTTATTTATAAATACCCTCCATATTCCTGTCAA | *ZRT2*$_{US}$ Step 2 dDNA for AB$^{01}$ | (This study) |
| SynFrag2 | GCATATTTACTTGCTTGCCTGATATCCTCGACTCATATACTTTGTAAATTACCTGTCACGTGTTTTTGTGAACTCCGATATTGTGTAATTTTACATTTGGGCACAGCATAGCCTGATGCCGTCCGGGT**cagggtcgcta**TATCACTCAATTTTTTTTTGTTTTTCACTGTTTTTCTGTCTTGTTGTTCCAAATAACCACTAATATTTCTCTTATACTTGACGATTTTTGGTGACCTATTATAGCTGGCAAGTGAAAGTGAATTAATAATATGCATTTTATAAAGTAGGCTTATTCATAAAATAATTAATTATTATTCAATCTCTAATTGATGTTCAGAAAATTTTTGGTTTGATGCCATACAAAGCAAAAAAAAAAATAATACATCAAAAATAGAACAAATGTAACTTTATGGTATTAAATCGTAATCATACACTTACTGAGAAGAACCATAAAGTCCAAGCTTTATAGAAAAAAGGCTAATGTTCTTTAGCATATGGTTTTTTTATGTCCTGGATTAACAACGTCCTTGGACTTAAGTACGTATGAAAGAACTAGCTAATAATTTAAAGCCAAACTGAGTCTTTCAACAACTACAATAGTGATTATACCCGAAACGCAAAATAATAAAAACTAATATTGACAATTGAATTATTCATTATTGGTCGGATTGGGTTACATTCAGATTGAAATCACGGTTGTAATTGCCGAATCTCTTTTTCATTGTTGTTCCATTTGTAACATTACCAGCTAGAAATGTAGTGTGTGGTTTGAGGTGCGTTTAGA**accttgttggt**TATTGCAAATTTTCTCTTGTTCATTTTTGTAACCACTAATAAAAACCATCACCAATTGACAATGAGTAAAAACTTTAAAAAAAAAGTAAAAATTAGAAAGAAAAAGTCAATCTCCCTTTTGTTGTAATTTATTTATAAATACCCTCCATATTCCTGTCAA | *ZRT2*$_{US}$ Step 2 dDNA for AB$^{10}$ | (This study) |
| SynFrag3 | GCATATTTACTTGCTTGCCTGATATCCTCGACTCATATACTTTGTAAATTACCTGTCACGTGTTTTTGTGAACTCCGATATTGTGTAATTTTACATTTGGGCACAGCATAGCCTGATGCCGTCCGGGT**cagggtcgcta**TATCACTCAATTTTTTTTTGTTTTTCACTGTTTTTCTGTCTTGTTGTTCCAAATAACCACTAATATTTCTCTTATACTTGACGATTTTTGGTGACCTATTATAGCTGGCAAGTGAAAGTGAATTAATAATATGCATTTTATAAAGTAGGCTTATTCATAAAATAATTAATTATTATTCAATCTCTAATTGATGTTCAGAAAATTTTTGGTTTGATGCCATACAAAGCAAAAAAAAAAATAATACATCAAAAATAGAACAAATGTAACTTTATGGTATTAAATCGTAATCATACACTTACTGAGAAGAACCATAAAGTCCAAGCTTTATAGAAAAAAGGCTAATGTTCTTTAGCATATGGTTTTTTTATGTCCTGGATTAACAACGTCCTTGGACTTAAGTACGTATGAAAGAACTAGCTAATAATTTAAAGCCAAACTGAGTCTTTCAACAACTACAATAGTGATTATACCCGAAACGCAAAATAATAAAAACTAATATTGACAATTGAATTATTCATTATTGGTCGGATTGGGTTACATTCAGATTGAAATCACGGTTGTAATTGCCGAATCTCTTTTTCATTGTTGTTCCATTTGTAACATTACCAGCTAGAAATGTAGTGTGTGGTTTGAGGTGCGTTTAGA**cagctgttcta**TAT | *ZRT2*$_{US}$ Step 2 dDNA for AB$^{11}$ | (This study) |

| Identifier | Sequence (5'→3') | Purpose | Source |
|---|---|---|---|
| | TGCAAATTTTCTCTTGTTCATTTTTGTAACCACTAATAAAAACCATCACCAATTGACAATGAGTAAAAACTTTAAAAAAAAAGTAAAAATTAGAAAGAAAAAGTCAATCTCCCTTTTGTTGTAATTTATTTATAAATACCCTCCATATTCCTGTCAA | | |
| AHO2200 | CGTAAACTATTTTTAATTTGAAATTGAGTGATAactaccaGTTTTAGAGCTAGAAATAGC | *ZRT2*$_{US}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2195 | CGTAAACTATTTTTAATTTGCGTACGCTGCAGGTCGACAGGTTTTAGAGCTAGAAATAGC | *ZRT2*$_{US}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| AHO2666 | GTCAGTTTCCCATACACATAAGG | *WOR1*$_{USd}$ AmpF | (This study) |
| AHO2667 | AGCAAGTATAGCCGTCATCT | *WOR1*$_{USd}$ AmpR | (This study) |
| AHO2661 | CTCTCATCAACAACAACGTCA | *WOR1*$_{USd}$ sF | (This study) |
| AHO2662 | AATAGTAGACTCCCTAACAGAGC | *WOR1*$_{USd}$ sR | (This study) |
| AHO2664 | GGAAACTAACCTAACACACAAAC | *WOR1*$_{USd}$ +/AoF | (This study) |
| AHO2663 | GTTTGTGTGTTAGGTTAGTTTCC | *WOR1*$_{USd}$ +/AoR | (This study) |
| AHO2664 | GGAAACTAACCTAACACACAAAC | *WOR1*$_{USd}$ +/AiF | (This study) |
| AHO2665 | TCCCACCCGTCTTTCATAAA | *WOR1*$_{USd}$ +/AiR | (This study) |
| AHO2659 | TAGGGACATTCAATTCGTCTTGAAAATATTAAAATTGACAAGAAAAACTTATTCGTACGCTGCAGGTCGACAG | *WOR1*$_{USd}$ Step 1 dDNA F | (This study) |
| AHO2660 | TGGTAGGTTCTGTCATTTATTGCTCTATTTTATAGTATTTAAAGTTTAAACTTTCCACTGTCGACCTGCAGCGTAC | *WOR1*$_{USd}$ Step 1 dDNA R | (This study) |
| AHO2658 | CGTAAACTATTTTTAATTTGTCATACACCAAGAAAACTCAGTTTTAGAGCTAGAAATAGC | *WOR1*$_{USd}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2195 | CGTAAACTATTTTTAATTTGCGTACGCTGCAGGTCGACAGGTTTTAGAGCTAGAAATAGC | *WOR1*$_{USd}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| SynFrag4 | GTCAGTTTCCCATACACATAAGGGAATGACCACACTCAAAAGTAATATCATAACTACAGGGCATAAAGCATATCACCTAGGGACATTCAATTCGTCTTGAAAATATTAAAATTGACAAGAAAAACTTATTCAAAGGGAGACCAAAAATACAGATTACCAACTATGTACACCCTAGAAAGAACTCAAAAAACGTAACCTTCGTTTCAAGTTGCACTTTAAAACAACAAATCCTGCTTTGATCAGATGAAACTATAATGCACGAATATGGAAACTAACCTAACACACAAACAATATATCATACACCAAGAAAACTCATGGTTTGTTGTTGTTGTTAGTGTATAATGTTAAAAAACTCTATTTTCACAATGACCCAAATAAAACCAAAAAAACACTAAGAAGacccttgcgTTTGAAACTTTTCAAAATGTATAGAGATCCCAAATCTAAAAAATGTTATTCACTATGGTTGTTGTTGTTTATTCAGAATTTAGTTATGGTTATATTAATGAAACTGTAACATAAAAAAAAAACAAGGGAATAATTAGAGTTTTACAAGAAATTTATGAAAGACGGGTGGGAAAAAAGTTTAAACTTTAAATACTATAAAATAGAGCAATAAATGACAGAACCTACCAGTAGTGATTCATAAATTATTATTTCTTGTTATACAATCAAAACCCCAGATATGATAACAGGAAAAAAAAAAAGTACTTATATAGATGACGGCTATACTTGCT | *WOR1*$_{USd}$ Step 2 dDNA for AB[1] | (This study) |
| AHO2664 | GGAAACTAACCTAACACACAAAC | *WOR1*$_{USp}$ AmpF | (This study) |
| AHO2676 | CCCACCTTCTCCCTCTTTC | *WOR1*$_{USp}$ AmpR | (This study) |
| AHO2666 | GTCAGTTTCCCATACACATAAGG | *WOR1*$_{USp}$ sF | (This study) |
| AHO2671 | CTCCCCCAACAACAAGTCTT | *WOR1*$_{USp}$ sR | (This study) |
| AHO2675 | GAGCAATAAATGACAGAACCTACC | *WOR1*$_{USp}$ +/AoF | (This study) |
| AHO2672 | TCCCACCCGTCTTTCATAA | *WOR1*$_{USp}$ +/AoR | (This study) |
| AHO2673 | CACTATGGTTGTTGTTGTTTATTCAG | *WOR1*$_{USp}$ +/AiF | (This study) |
| AHO2674 | AGTAGACTCCCTAACAGAGC | *WOR1*$_{USp}$ +/AiR | (This study) |
| AHO2669 | ACAATGACCCAAATAAAACCAAAAAAACACTAAGAAGTTAAACTTTTTTGAAACCACTGTCGACCTGCAGCGTAC | *WOR1*$_{USp}$ Step 1 dDNA F | (This study) |

| Identifier | Sequence (5’→3’) | Purpose | Source |
|---|---|---|---|
| AHO2670 | ATTTTTGCATGTTCTATTTTTAGTCCATACATAATGTAACGCACACACATTAGACGTACGCTGCAGGTCGACAG | *WOR1*$_{USp}$ Step 1 dDNA R | (This study) |
| AHO2668 | CGTAAACTATTTTTAATTTGATTTATGAATCACTACTGGTGTTTTAGAGCTAGAAATAGC | *WOR1*$_{USp}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2195 | CGTAAACTATTTTTAATTTGCGTACGCTGCAGGTCGACAGGTTTTAGAGCTAGAAATAGC | *WOR1*$_{USp}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| SynFrag5 | GGAAACTAACCTAACACACAAACAATATATCATACACCAAGAAAACTCATGGTTTGTTGTTGTTGTTAGTGTATAATGTTAAAAAACTCTATTTTCACAATGACCCAAATAAAACCAAAAAAACACTAAGAAGTTAAACTTTTTTGAAACTTTTCAAAATGTATAGAGATCCCAAATCTAAAAAATGTTATTCACTATGGTTGTTGTTGTTTATTCAGAATTTAGTTATGGTTATATTAATGAAACTGTAACATAAAAAAAAACAAGGGAATAATTAGAGTTTTACAAGAAATTTATGAAAGACGGGTGGGAAAcgcaaacccttgcgAAATACTATAAAATAGAGCAATAAATGACAGAACCTACCAGTAGTGATTCATAAATTATTATTTCTTGTTATACAATCAAAACCCCAGATATGATAACAGGAAAAAAAAAAGTACTTATATAGATGACGGCTATACTTGCTCAAGTGAGTTTGATGTGATTTTTAACACGCTCTGTTAGGGAGTCTACTATTTTTTTTTCTGGCGATAACAATAAGAAATCTCTAATGTGTGTGCGTTACATTATGTATGGACTAAAAATAGAACATGCAAAAATTGCGAGAAAGAAAGCGAGTGAGTAAGGGCGTGCGTGCGTGCATGAGTGAAAGAGGGAGAAGGTGGG | *WOR1*$_{USp}$ Step 2 dDNA for AB1 | (This study) |
| AHO2686 | ACACACACACACACAATCACAC | *WOR2*$_{DS}$ AmpF | (This study) |
| AHO2687 | TAGCAAGGCAACCATCAAGC | *WOR2*$_{DS}$ AmpR | (This study) |
| AHO2680 | CTACTACTGATGGTCTACTGATGG | *WOR2*$_{DS}$ sF | (This study) |
| AHO2681 | CGTTTGTAGATGGTTCTGGTTTG | *WOR2*$_{DS}$ sR | (This study) |
| AHO2685 | ATCGCTCCTTGTGTTTGTGTG | *WOR2*$_{DS}$ +/AoF | (This study) |
| AHO2682 | CCCACACAAACACAAGGAGC | *WOR2*$_{DS}$ +/AoR | (This study) |
| AHO2683 | TTGTGGAAGTGTAAGAGGGA | *WOR2*$_{DS}$ +/AiF | (This study) |
| AHO2684 | CTGCTTGCTAAACCCAAACC | *WOR2*$_{DS}$ +/AiR | (This study) |
| AHO2678 | CTAAAAACCAACAAGTTACTTGATAGAACCTCGATTTCATTATGAATTCCACACGTACGCTGCAGGTCGACAG | *WOR2*$_{DS}$ Step 1 dDNA F | (This study) |
| AHO2679 | AATACAAATACAGATGACACCAAAAAGAAAAAAGTTAAACTTGTAATAGTTAATCCACTGTCGACCTGCAGCGTAC | *WOR2*$_{DS}$ Step 1 dDNA R | (This study) |
| AHO2677 | CGTAAACTATTTTTAATTTGATCGCTCCTTGTGTTTGTGTGTTTTAGAGCTAGAAATAGC | *WOR2*$_{DS}$ oligo used for stitching spacer to scaffold for Step 1 Target | (This study) |
| AHO2195 | CGTAAACTATTTTTAATTTGCGTACGCTGCAGGTCGACAGGTTTTAGAGCTAGAAATAGC | *WOR2*$_{DS}$ oligo used for stitching spacer to scaffold for Step 2 Target | (This study) |
| SynFrag6 | ACACACACACACACAATCACACAAAATTAGCACTACTAAATGTTTGAGAATGATTCGAATCAAGGGAAACTAAAAACCAACAAGTTACTTGATAGAACCTCGATTTCATTATGAATTCCACATACACAATAATACAGTACCAAAAGTTTAAATTTAAAAAAAAAATCAGCCCATTAGAGAAATCTAGATGTAGATATATTTTGTGGAAGTGTAAGAGGGATAAGCCATTTGTAATTTTACACAATTAATCGCTCCTTGTGTTTGTGTGGGAAAAACTTTGCAATTGGTTGATTGTGCAACAATTGCTAAAACATTGGTTACCCATTTTCCTTTTTTTGCAATTTCCAAATAATAATAATGATAATACTTATCAAAACAAAGAAACAATTAACGAGACAAGTTTAAATCAAACTCAATACAATTCATAAACTCTAACTGGcgcaagggtGTTTTCTATTTTTTGTTTGTGAATGTATTACAATAAATTGAATTTTGATCGAAATATTAATCGGGGCTAGAGTGTGGTTTGGGTTTAGCAAGCAGCTATTGTTTGAAAAAAATTAAAATGACTGCATTAACTATTACAAGTTTAACTTTTTTCTTTTTGGTGTCATCTGTATTTGTATTTATTGCATGGGAAAGACAATACAGTAGTAATAACGAAACTATCAACCACGAAAAGAGGAAATATCCCTCAACTTTCCAAATTTAATTCAAAAGATACTAAAAAAAAACCTTGAGTCAACAATAGAATTTATTGAAACTTAATTCTCCTCATGTGGATTCTTTATTTGCTTGATGGTTGCCTTGCTA | *WOR2*$_{DS}$ Step 2 dDNA for AB1 | (This study) |

| Identifier | Sequence (5’→3’) | Purpose | Source |
|---|---|---|---|
| AHO1096 | GACGGCACGGCCACGCGTTTAAACCGCC | gRNA cassette part 1 F | [9] |
| AHO1098 | CAAATTAAAAATAGTTTACGCAAG | gRNA cassette part 1 R | [9] |
| AHO1099 | GTTTTAGAGCTAGAAATAGCAAGTT | gRNA cassette part 2 F | [9] |
| AHO1097 | CCCGCCAGGCGCTGGGGTTTAAACACCG | gRNA cassette part 2 R | [9] |
| AHO1237 | AGGTGATGCTGAAGCTATTGAAG | gRNA full cassette F | [9] |
| AHO1238 | TGTATTTTGTTTTAAAATTTTAGTGACTGTTTC | gRNA full cassette R | [9] |

Shared upstream and downstream primers sF and sR are colored blue. Sequences homologous to the AmpF and AmpR primers, used to amplify the Step 2 dDNA from wild-type gDNA or synthetic DNA templates, are colored green. Nucleotides that are homologous to the AddTag-selected spacer sequences in the first and second round of genome editing are colored red, and fuchsia, respectively, with the associated PAM sequences in brick and violet. In first round donor DNA sequences (Step 1 dDNA), the upstream and downstream homology regions are given yellow and orange backgrounds, and any addtag insert sequence is given a pink background. **Bold**, lower-case letters are nucleotides that encode for the Zap1 and Wor1 binding sites. Sequences are listed in the 5’ to 3’ orientation. Oligonucleotide sequences used in this study did not require any special modifications or purifications. SynFrag sequences used to generate *ZRT2* $_{US}$, *WOR1*$_{USd}$, *WOR1*$_{USp}$, and *WOR2*$_{DS}$ dDNAs were synthesized as dsDNA fragments.

# Appendix B Methods

## B.1 Plasmids and synthetic DNA

For all genetic modifications in this study, we used the ADDTAG software to automatically select the RGN targets, dDNAs and corresponding Step 2 Targets for Step 1 editing, determine optimal primers (AmpF/AmpR) for amplifying the Step 2 dDNAs, and to pick vPCR primers for validating integration of the intended modified features at the target loci following each step of editing (Table 0.3). We used the previously-published method for C. *albicans* genome editing using CRISPR/Cas-induced HDR [9], with minor modifications (B.3). The gRNA expression cassettes used to make all deletion and complementation/add-back strains were generated via an “All-in-1” PCR stitching approach (B.3). Briefly, linear DNA fragments containing the pSNR52 promoter and the invariable structural component of the gRNA coding sequence were PCR amplified from pADH110 and pADH119, respectively, using AHO1096/AHO1098 and AHO1097/AHO1099 primer pairs and Phusion polymerase (ThermoFisher) (Table 0.2, B.3). The resulting fragments were stitched together in a single reaction using custom target sequence-specific bridging oligos and AHO1237/AHO1238 amplification primers. Linear Cas9 expression cassettes were generated by MssI digestion of pADH137 (Table 0.2), and were transformed along with the stitched custom gRNA expression cassettes and custom dDNA fragments. The Step 1 dDNA fragments for the *ADE2*, *EFG1*, and *BRG1* loci were generated by annealing complementary 100-mer oligonucleotides. We used overlapping primer extension with Phusion polymerase to generate Step 1 dDNAs for the *WOR1* and *ZRT2* loci. Wild-type add-back dDNA fragments were generated by standard PCR amplification of C. *albicans* genomic DNA using ADDTAG-designed amplification primers (AmpF/AmpR) and Phusion polymerase. dDNA fragments containing mutated Zap1 binding sites were first synthesized as full-length synthetic DNA fragments (ThermoFisher) then PCR amplified using the same ADDTAG-designed primers used to amplify the corresponding wild-type dDNA fragments.

## B.2 Summary of polymorphisms in Features and dDNA sequences

**Table 0.4 – Summary of polymorphisms in Features and dDNA sequences**

Polymorphisms/Length (nt)

| | **Step 1 dDNA** | | **eFeature** | | | **Step 1 dDNA** | | | **Step 2 dDNA** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **insert type** | **eUS** | **Feature** | **eDS** | **US** | **Insert** | **DS** | **US** | **✻Feature** | **DS** |
| *ADE2*$_{CDS}$ | mintag | 0/0 | 0/1707 | 0/0 | 0/49 | CC | 0/49 | 0/133 | AB: 0/1707 | 0/140 |
| *BRG1*$_{CDS}$ | mintag | 3/42 | 13/1269,1272 | 0/0 | 0/50 | - | 0/50 | 3/221 | AB: ND | 1/115 |
| *EFG1*$_{CDS}$ | mintag | 0/1 | 6/1658,1653 | 1/12 | 0/49 | CCTC | 0/47 | 1/149 | AB: ND | 2/170 |
| *ZAP1*$_{US}$ | addtag | 3/210,213 | 0/11 | 2/199,198 | 0/53 | CGTACGCTGCAGGTCGACAGTGG | 0/54 | 1/210 | $AB^0$: 0/231<br>$AB^1$: 0/231 | 0/197 |
| *ZRT2*$_{US}$ | addtag | 0/0 | 5/667,666 | 0/0 | 0/53 | CGTACGCTGCAGGTCGACAGTGG | 0/54 | 0/128 | $AB^{00}$: 0/667<br>$AB^{01}$: 0/667<br>$AB^{10}$: 0/667<br>$AB^{11}$: 0/667 | 2/160 |
| *WOR1*$_{USd}$ | addtag | 0/269 | 0/9 | 0/172 | 0/53 | CGTACGCTGCAGGTCGACAGTGG | 0/54 | 0/130 | $AB^0$: 0/450<br>$AB^1$: 0/450 | 0/155 |
| *WOR1*$_{USp}$ | addtag | 0/165 | 0/14 | 0/221 | 0/53 | CCACTGTCGACCTGCAGCGTACG | 0/54 | 0/149 | $AB^0$: 0/400<br>$AB^1$: 0/400 | 0/120 |
| *WOR2*$_{DS}$ | addtag | 0/317 | 0/9 | 0/125 | 0/53 | CGTACGCTGCAGGTCGACAGTGG | 0/54 | 0/122 | $AB^0$: 0/451<br>$AB^1$: 0/451 | 0/239 |

The "US", "DS", "eUS", "eDS", "Feature", and "✻Feature" columns, display an "a/b" format, with "a" representing the number of polymorphisms within "b" contiguous nucleotides. The eFeature columns represent the full input Feature that was expanded to circumvent polymorphisms, with the "eUS" and "eDS" columns representing the lengths of Feature expansion. For the Step 1 dDNA and Step 2 dDNA columns, the "US" and "DS" columns represent the flanking homology arms. Please note that the Step 1 dDNAs do not contain polymorphisms, but the Step 2 dDNAs do. Some genomic regions have alleles with different lengths, which are shown as comma-separated list of lengths. The "ND" indicates that polymorphisms in restored loci were not determined. The sequence orientations for `mintag` and `addtag` inserts are indicated with the start as lower contig position and the end as higher contig position, as they appear in the reference genome (i.e. "+" orientation relative to the contig).

## B.3 "All-in-1" gRNA cassette stitching

This protocol describes the "All-in-1" gRNA expression cassette stitching method, an adaptation of the cloning-free 2-step assembly method previously described by Nguyen, *et al* [9]. The All-in-1 approach assembles two universal gRNA expression cassette fragments (Fragments A and B) into an intact Target-specific gRNA expression cassette in a single reaction using a custom bridging gRNA oligo and conserved amplification primers (Figure 0.1). While the original method requires the generation of a new Fragment B for each unique gRNA target, the All-in-1 simply requires a single unique oligonucleotide and two universal PCR fragments as templates. The All-in-1 approach cuts the time to generate gRNA expression cassettes nearly in half as compared to the traditional method and is as just as efficient in creating gene knockouts.

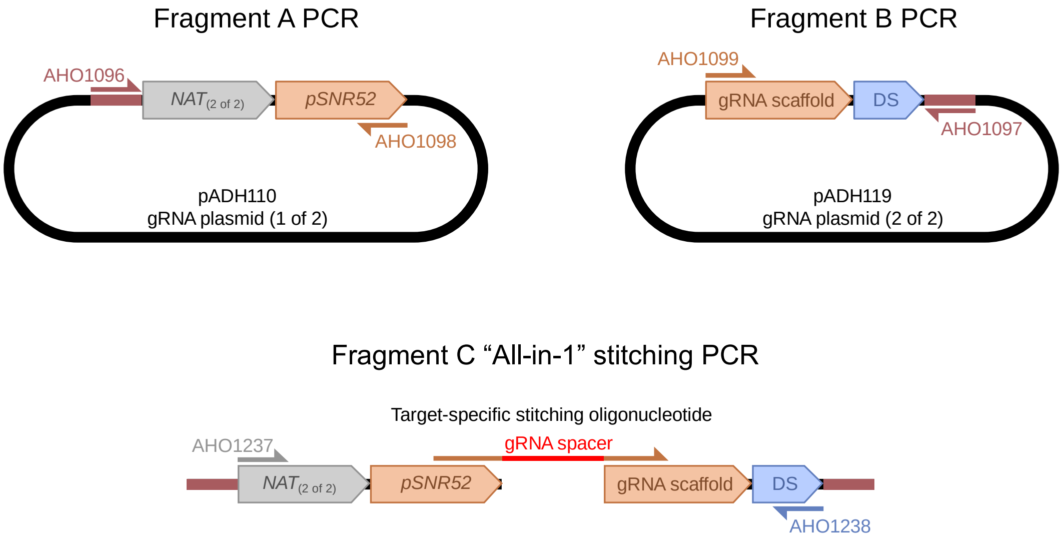


**Figure 0.1 – Schematic of "All-in-1" gRNA stitching methodology**

Arrows indicate PCR primers. The AHO1096/AHO1098 primer pair amplifies the universal "A fragment", and the AHO1099/AHO1097 primer pair amplifies the universal "B fragment". The AHO1237/AHO1238 primer pair is used in conjunction with a "target-specific" gRNA oligo that bridges between the A and B fragments.

Stranded annotations, such as CDS and promoters, are shown as irregular pentagons, pointing toward the right.

Conditions for PCR amplification of the universal "A" fragment from pADH110, gRNA plasmid (1 of 2):

- PCR Mix (makes enough "A" fragment for >75 "C" fragment stitching PCRs):
  - (1) 75.5 µL $H_2O$
  - (2) 20 µL 5× Phusion HF buffer
  - (3) 2 µL dNTP mix (10 mM each dNTP)
  - (4) 1 µL pADH110 (1 ng/µL)
  - (5) 0.5 µL AHO1096 (100 µM)
  - (6) 0.5 µL AHO1098 (100 µM)
  - (7) 0.5 µL Phusion polymerase
- "A" fragment PCR cycling conditions:
  - (1) 98 °C, 30sec
  - (2) 98 °C, 20sec
  - (3) 58 °C, 20sec
  - (4) 72 °C, 30sec
  - (5) Return to step 2 for a total of 30 cycles
  - (6) End

Conditions for PCR amplification of the universal "B" fragment from pADH119, gRNA plasmid (2 of 2):

- PCR Mix (makes enough "B" fragment for >75 "C" fragment stitching PCRs):

(1) 75.5 µL $H_2O$
(2) 20 µL 5× Phusion HF buffer
(3) 2 µL dNTP mix (10 mM each dNTP)
(4) 1 µL pADH119 (1 ng/µL)
(5) 0.5 µL AHO1097 (100 µM)
(6) 0.5 µL AHO1099 (100 µM)
(7) 0.5 µL Phusion polymerase

- "B" fragment touchdown PCR cycling conditions:
  (1) 98 °C, 30 sec
  (2) 98 °C, 20 sec
  (3) 65 °C, 20 sec
  (4) 72 °C, 30 sec
  (5) Return to step 2 for a total of 10 cycles, reducing annealing temperature by 1 °C each cycle.
  (6) 98 °C, 20 sec
  (7) 55 °C, 20 sec
  (8) 72 °C, 30 sec
  (9) Return to step 6 for a total of 25 cycles
  (10) End

Conditions for All-in-1 stitching PCR to amplify unique "C" fragment gRNA expression cassette:

- PCR Mix (makes enough "C" fragment for two transformations):
  (1) 73.5 µL water
  (2) 20 µL 5× Phusion HF buffer
  (3) 2 µL dNTPs 10mM
  (4) 1 µL universal A Fragment (See note 1 below)
  (5) 1 µL universal B Fragment (See note 1 below)
  (6) 0.5 µL AHO1237 100 µM
  (7) 0.5 µL AHO1238 100 µM
  (8) 1 µL custom gRNA oligo 100 nM (See note 2 below)
  (9) 0.5 µL Phusion polymerase
- "C" fragment touchdown PCR cycling conditions:
  (1) 98 °C, 30 sec
  (2) 98 °C, 15 sec
  (3) 60 °C, 15 sec
  (4) 72 °C, 60 sec
  (5) Return to step 2 for a total of 5 cycles, reducing annealing temperature by 1 °C each cycle.
  (6) 98 °C, 15 sec
  (7) 66 °C, 15 sec
  (8) 72 °C, 60 sec
  (9) Return to step 6 for a total of 30 cycles
  (10) End

Notes:

- The A and B fragment PCR products can be added directly to the C fragment PCR reaction without any post-PCR purification.
- It is critical that the custom gRNA oligo is at low concentrations so that the B fragment does not take over the PCR reaction.

### B.4 Using the ADDTAG software to design the genome editing experiments

We used the ADDTAG software to guide the design for two classes of genome edits: coding sequence (CDS) edits and cis-acting regulatory element (CRE) edits (Table 0.4). For CDS edits, we directed ADDTAG to replace the CDS with a `mintag` insert during Step 1, allowing for up to and added 4 bp. For *EFG1* and *BRG1*, we allowed for minor Feature expansion, such that designs would avoid known allelic variations adjacent to the CDSs. For CRE edits, ADDTAG replaced the input protein binding sequence, and some sequence both up- and down-stream of it, with a 23 bp `addtag` insert composed of a 20 nt sequence homologous to the Spacer and a 3 nt PAM sequence, that together encodes a high-efficiency Cas9 binding site. Round 2 dDNA was amplified from either the wild-type (+/+) gDNA or from synthetic DNA fragments (SynFrag) (Table 0.3). For all genome edits, RGN Target sites were chosen that maximized both the Hsu-Zhang and CFD off-target scores. For *ADE2*, *EFG1*, and *BRG1* loci, these were maximal scores of 100%.

### B.5 *Candida albicans* cell culture and transformation

All C. *albicans* strains used in this study (Table 0.1) were derived from strain SC5314 [12]. AHY940 (SC5314 with one allele of *LEU2* deleted) [9] was used as the base strain for all genome editing procedures, and transformations were performed as previously described [9]. Briefly, gRNA and Cas9 expression cassettes, along with dDNA fragments, were transformed into AHY940 or derivative strains via chemical transformation and plated onto YPD agar plates (2% Bacto peptone, 2% dextrose, 1% yeast extract, 2.5% agar) supplemented with 200 µg/ml nourseothricin (NAT; GoldBio). Transformation plates were incubated for two days at 30°C to select for integration of the gRNA and Cas9 expression cassettes [9], and genome editing at the target locus was validated by vPCR using ADDTAG-generated primers. For the experimental validation of each edited locus, whole, individual colonies were selected, and direct cell lysate was used as template input to the vPCR reactions. Subsequent to genotype verification, the gRNA and Cas9 expression cassettes, along with the NAT resistance marker, were subsequently removed via the LEUpOUT method by selection on synthetic defined (SD) agar medium without leucine [9]. Strains that harbored mutated Wor1 or Zap1 binding sites and their wild-type add-back counterparts were further validated at the base pair level via Sanger sequencing [13, 14] of colony PCR products that spanned the engineered loci (Figure 2.19, Figure 2.20, Figure 2.21, Figure 2.23).

### B.6 *Candida albicans* phenotypic assessment of Zap1 binding site mutant strains

Strains with genotypes involving Zap1 binding site manipulations were assayed for their abilities to grow on zinc-sufficient synthetic complete medium (2% dextrose, 0.17% yeast nitrogen base without ammonium sulfate, 0.5% ammonium sulfate, and auxotrophic supplements) and zinc-deficient medium (2% dextrose, 0.17% yeast nitrogen base without either ammonium sulfate or zinc sulfate, 0.2% ammonium sulfate, 2.5 µM EDTA, and auxotrophic supplements) on 2% agar plates [11] (Figure 2.25). These include the *ZAP1*$_{CDS}$ (*ZAP1*$_{CDS}$ +/+, *zap1*$_{CDS}$ Δ/Δ) and *ZAP1*$_{US}$ (*ZAP1*$_{US}$ +/+, *zap1*$_{US}$ Δ/Δ, *ZAP1*$_{US}$ AB$^{0}$/AB$^{0}$, *ZAP1*$_{US}$ AB$^{1}$/AB$^{1}$) strains as well as the *ZRT2* reference strain (*ZRT2*$_{US}$ +/+), *ZRT2* upstream intergenic region deletions (*zrt2*$_{US}$ Δ/Δ), binding site mutants (*ZRT2*$_{US}$ AB$^{01}$/AB$^{01}$, *ZRT2*$_{US}$

AB$^{10}$/AB$^{10}$, *ZRT2$_{US}$* AB$^{11}$/AB$^{11}$), and wild-type add-back strains (*ZRT2$_{US}$* AB$^{00}$/AB$^{00}$). Each strain was grown to saturation via overnight culture in YPD liquid medium at 30 °C with shaking prior to back-dilution. Strains were grown to mid-log, washed, and then serial diluted in sterile 1× phosphate-buffered saline (PBS). Two independent biological replicates for each engineered genotype were assayed twice. Aliquots of each dilution were spotted onto zinc-sufficient and zinc-deficient agar plates and grown at 30°C for 48 hours.

### B.7 *Candida albicans* biofilm phenotype assay

*C. albicans* strains were cultured from cryogenically frozen stocks at 30°C on YPD agar plates (2% bacteriological-grade peptone, 2% dextrose, 1% yeast extract, 2.5% agar) for two days. A single colony of each strain to be tested was grown overnight in liquid YPD medium. Biofilms were grown on the bottoms of 12-well polystyrene plates in Spider medium (1% nutrient broth, 0.2% $K_2HPO_4$, 1% mannitol, pH 7.2) with shaking at 200 rpm at 37°C using an ELMI shaker (ELMI) as described previously [15, 16]. The optical density 600 nm ($OD_{600}$) was measured for each well using a Cytation 5 plate reader (BioTek), and biofilms were imaged. For each genotype, a n=4 number of wells were assayed. Two independent biological replicates for each engineered genotype were assayed twice. Significance levels and confidence intervals were calculated by the Student t statistic with unequal variance using $H_A$: $OD_{600}(+/+) \neq OD_{600}(AB/AB)$ and $H_A$: $OD_{600}(+/+) > OD_{600}(\Delta/\Delta)$.

## Appendix C Appendices references

1. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, *et al*. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech*. 2016; 34(2):184-91. DOI: 10.1038/nbt.3437, PMID: 26780180.
2. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, *et al*. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotech*. 2013; 31(9):827-32. DOI: 10.1038/nbt.2647, PMID: 23873081.
3. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009; 10(1):421. DOI: 10.1186/1471-2105-10-421, PMID: 20003500.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012; 9:357. DOI: 10.1038/nmeth.1923, PMID: 22388286.
5. Katoh K, Kuma K-i, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*. 2005; 33(2):511-8. DOI: 10.1093/nar/gki198, PMID: 15661851.
6. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. In: Keith JM, editor. Bioinformatics: Structure, Function and Applications. Totowa, NJ: Humana Press; 2008; p. 3-31. DOI: 10.1007/978-1-60327-429-6_1.
7. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The *Candida* Genome Database (CGD): Incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Research*. 2016; 45(D1):D592-D6. DOI: 10.1093/nar/gkw924, PMID: 27738138.
8. Foundation FS. GNU Bash. 4.4.7 ed2007.
9. Nguyen N, Quail MMF, Hernday AD. An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. *mSphere*. 2017; 2(2). DOI: 10.1128/mSphereDirect.00149-17, PMID: 28497115.
10. Davis D, Edwards JE, Mitchell AP, Ibrahim AS. *Candida albicans RIM101* pH Response Pathway Is Required for Host-Pathogen Interactions. *Infection and*

*Immunity*. 2000; 68(10):5953-9. DOI: 10.1128/iai.68.10.5953-5959.2000, PMID: 10992507.

11. Nobile CJ, Nett JE, Hernday AD, Homann OR, Deneault J-S, Nantel A, Andes DR, *et al*. Biofilm matrix regulation by *Candida albicans* Zap1. *PLOS Biology*. 2009; 7(6):e1000133. DOI: 10.1371/journal.pbio.1000133, PMID: 19529758.
12. Gillum AM, Tsay EYH, Kirsch DR. Isolation of the *Candida albicans* gene for orotidine-5′-phosphate decarboxylase by complementation of *S. cerevisiae ura3* and *E. coli pyrF* mutations. *Molecular and General Genetics MGG*. 1984; 198(1):179-82. DOI: 10.1007/BF00328721, PMID: 6394964.
13. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*. 1977; 74(12):5463-7. DOI: 10.1073/pnas.74.12.5463, PMID: 271968.
14. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C*, et al*. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986; 321(6071):674-9. DOI: 10.1038/321674a0, PMID: 3713851.
15. Gulati M, Lohse MB, Ennis CL, Gonzalez RE, Perry AM, Bapat P, Arevalo AV*, et al*. *In vitro* culturing and screening of *Candida albicans* biofilms. *Current Protocols in Microbiology*. 2018; 50(1):e60. DOI: 10.1002/cpmc.60, PMID: 29995344.
16. Lohse MB, Gulati M, Valle Arevalo A, Fishburn A, Johnson AD, Nobile CJ. Assessment and optimizations of *Candida albicans in vitro* biofilm assays. *Antimicrobial Agents and Chemotherapy*. 2017; 61(5):e02749-16. DOI: 10.1128/aac.02749-16, PMID: 28289028.

End of document