Nucleic Acids Research, 2025, **53**, gkaf783
https://doi.org/10.1093/nar/gkaf783
Methods

Kingdom-wide CRISPR guide design with ALLEGRO

Amirsadra Mohseni[1,†], Reyhane Ghorbani Nia[2,†], Aida Tafrishi[2], Mario León López[3], Xin-Zhan Liu[4,5], Jason E. Stajich[4], Stefano Lonardi[1,*], Ian Wheeldon[2,*]

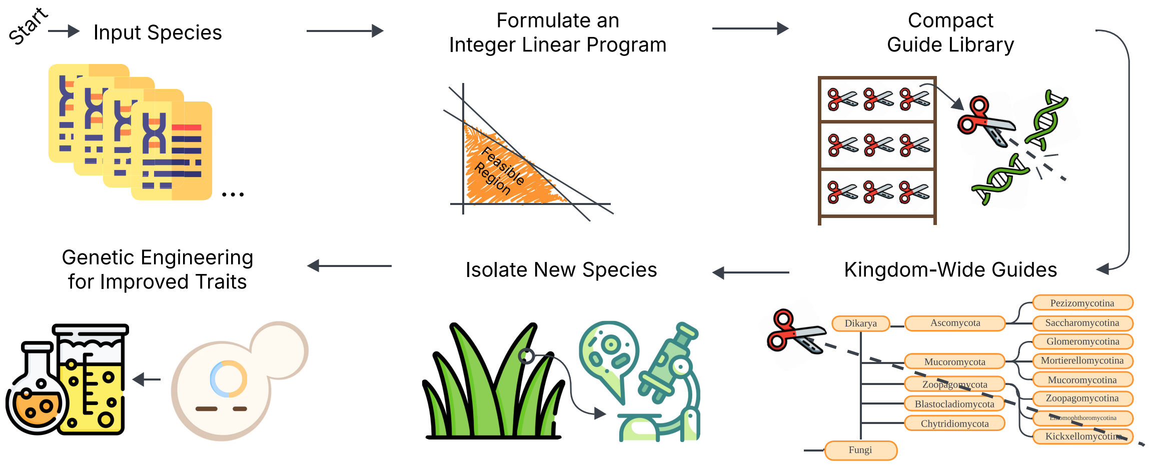
[1]Computer Science and Engineering, University of California, Riverside, CA 92521, United States
[2]Chemical and Environmental Engineering, University of California, Riverside, CA 92521, United States
[3]Bioengineering, University of California, Riverside, CA 92521, United States
[4]Microbiology and Plant Pathology, University of California, Riverside, CA 92521, United States
[5]Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

[*]To whom correspondence should be addressed. Email: stelo@cs.ucr.edu
Correspondence may also be addressed to Ian Wheeldon. Email: iwheeldon@engr.ucr.edu
[†]Contributed equally to this work.

Abstract

Designing CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) single guide RNA (sgRNA) libraries targeting entire kingdoms of life will significantly advance genetic research in diverse and underexplored taxa. Current sgRNA design tools are often species-specific and fail to scale to large, phylogenetically diverse datasets, limiting their applicability to comparative genomics, evolutionary studies, and biotechnology. Here, we introduce ALLEGRO, a combinatorial optimization algorithm designed to compose minimal, yet highly effective sgRNA libraries targeting thousands of species at the same time. Leveraging integer linear programming, ALLEGRO identified compact sgRNA sets simultaneously targeting multiple genes of interest for over 2000 species across the fungal kingdom. We experimentally validated sgRNAs designed by ALLEGRO in *Kluyveromyces marxianus*, *Komagataella phaffii*, *Yarrowia lipolytica*, and *Saccharomyces cerevisiae*, confirming successful genome edits. Additionally, we employed a generalized Cas9–ribonucleoprotein delivery system to apply ALLEGRO's sgRNA libraries to untested fungal genomes, such as *Rhodotorula araucariae*. Our experimental findings, together with cross-validation, demonstrate that ALLEGRO facilitates efficient CRISPR genome editing, enabling the development of universal sgRNA libraries applicable to entire taxonomic groups.

Graphical abstract



Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) and their associated proteins, known as CRISPR–Cas systems, are innate bacterial and archaeal defense mechanisms that use CRISPR RNAs (crRNAs) to detect and induce double-stranded breaks into foreign nucleic acids, consequently silencing them [1–3]. Over the past decade, these defense mechanisms have been "hijacked" into genome editing systems to great success. In engineered systems, the crRNA and its *trans*-activating partner tracrRNA are typically fused

Received: February 7, 2025. **Revised:** July 24, 2025. **Editorial Decision:** July 28, 2025. **Accepted:** July 29, 2025

into a single guide RNA (sgRNA, or simply "guide"), which directs the Cas nuclease to a specific genomic target. However, most research has largely focused on model organisms, leaving a gap in accessibility for researchers working on diverse taxonomic groups and nonmodel organisms. Additionally, current sgRNA design tools are typically species-specific, with limited flexibility to address the challenges of cross-species variation in target and background sequences [4]. These constraints hinder the development of scalable solutions for editing multiple genes across diverse genomes, particularly when addressing large-scale studies or species outside of the conventional models. The ability to design effective CRISPR libraries across species is critical for advancing genome editing in underexplored taxa. Expanding CRISPR applications to diverse species enables the integration of unique organisms into engineering and discovery pipelines, unlocking novel genetic and enzymatic pathways previously inaccessible with traditional approaches. However, designing and cloning individually tailored sgRNAs for multiple genes across numerous genomes is experimentally burdensome and inefficient, both in cost and synthesis time, making large-scale genomic screens laborious and impractical. This highlights the need for design tools that can generate a minimal set of guides capable of efficiently covering all desired targets across multiple species.

This problem can be formulated as a "Set Covering Problem," where the objective is to identify the smallest set of sgRNAs that collectively target all desired genes across a group of organisms. In this combinatorial optimization framework, the input consists of a universe of elements (e.g. genes or genomes) and a collection of sets (each representing an sgRNA targeting a subset of those elements). The goal is to select the minimal number of sgRNAs whose combined targets cover the entire set of desired genes. This problem has been extensively studied in the field of Computer Science, and is known to be NP-hard, meaning that finding an optimal solution quickly becomes computationally intractable as the problem size grows [5].

Designing a compact guide library across many genomes has been previously studied and a few tools have been designed for this purpose. For instance, Endo *et al.* leveraged the mismatch tolerance of Cas9 guides in the protospacer adjacent motif (PAM)–distal region to design an sgRNA capable of targeting homologous genes of interest in the rice genome [6]. CRISPR MultiTargeter similarly identifies common guides using multiple sequence alignments as input, though it does not optimize for the smallest guide set [7]. CRISPys is another tool designed for creating guide libraries using various strategies, including one functionality that identifies minimal sets of guides targeting small-sized gene sets [8]. Thus far, the tool that most effectively addressed the guide design challenge was MINORg [9], which identifies minimal sets of guides for each input sequence but is limited to relatively small gene and guide sets. While the algorithms employed by the available methods can effectively design minimal guide libraries for relatively small datasets, they prove to be computationally inefficient and fail to scale to larger, more complex datasets.

Somewhat related to the problem we address here is multiplexed CRISPR genome editing, which has been widely explored in the literature (e.g. [10–13]). However, these studies focus on selecting multiple sgRNAs within a single genome to enable simultaneous editing at multiple loci. They do not attempt to design a minimal set of sgRNAs that can target specific loci across a large number of species—a significantly more complex optimization problem that we seek to address here.

To facilitate the design of guide libraries that cover a multitude of diverse species, here we present ALLEGRO (Algorithm for a Linear program Enabling Guide RNA Optimization), a time- and memory-efficient algorithm capable of designing a minimal guide library for thousands of organisms within minutes. ALLEGRO harnesses combinatorial optimization and integer linear programming to tackle the challenge of scalable, cross-species sgRNA design. Designed for flexibility and ease of use, ALLEGRO enables users to (i) choose between two sgRNA design strategies, called *tracks*, (ii) filter guides based on sequence features, and (iii) cluster sgRNAs to further reduce the library size. These features, among others, enable the composition of guide libraries applicable across entire biological kingdoms.

To demonstrate the scalability of ALLEGRO, we conducted a large-scale computational experiment to generate minimal sgRNA libraries targeting the complete transcriptome of 1000 species from the *Ascomycota* phylum. We found that a library comprising only nine sgRNAs—each with hundreds of shared targets—was sufficient to introduce cuts across these species. Furthermore, ALLEGRO was able to design compact sgRNA libraries using two design strategies ("tracks") to target the coding sequences of >2000 fungal species.

To experimentally validate the sgRNA libraries generated by ALLEGRO, we performed a series of CRISPR–Cas9 knockout experiments across several fungal species, including *Kluyveromyces marxianus*, *Komagataella phaffii*, *Yarrowia lipolytica*, *Saccharomyces cerevisiae*, and *Rhodotorula araucariae*. The validated sgRNAs were chosen to target auxotrophy-associated genes, enabling straightforward screening of gene disruptions. Our results confirmed that ALLEGRO constructs highly efficient and specific sgRNA libraries, consistently yielding robust gene-editing outcomes across all tested species. In addition, we developed a generalizable protocol for applying ALLEGRO-designed sgRNA libraries to other fungal genomes by combining Cas9–ribonucleoprotein (RNP) delivery with protoplast transformation. These findings highlight ALLEGRO's effectiveness in designing functional sgRNAs for genome engineering across a broad range of fungi, advancing applications in both microbiology and synthetic biology.

Materials and methods

Computational arrangement

The input to ALLEGRO is a set of species and a set of genes of interest. In this work, we apply ALLEGRO to a large group of fungal species, but the algorithm can be used on any group of organisms. First, ALLEGRO determines orthogroups (set of genes descended from a single gene in the last common ancestor of the species) across all input species for the genes of interest. These orthogroups are determined using reciprocal best hits with DIAMOND [14]. ALLEGRO then (optionally) computes predicted efficiency scores for all Cas9 guides in the orthogroups using the uCRISPR guide design algorithm [15]. Finally, ALLEGRO determines the smallest library of guides that maximizes the cutting efficiency while targeting all genes in the set. An overview of the workflow is illustrated in Fig. 1.

ALLEGRO applies combinatorial optimization techniques to efficiently identify the minimal set of sgRNAs needed to

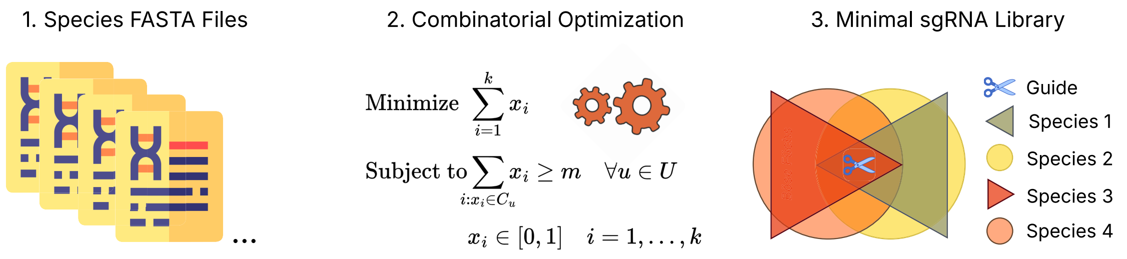

1. Species FASTA Files
...
2. Combinatorial Optimization
$\text{Minimize} \sum_{i=1}^{k} x_i$
$\text{Subject to} \sum_{i:x_i \in C_u} x_i \geq m \quad \forall u \in U$
$x_i \in [0,1] \quad i = 1, \ldots, k$
3. Minimal sgRNA Library
Guide
Species 1
Species 2
Species 3
Species 4

Figure 1. ALLEGRO's workflow. Step (1): Given the gene sequence or the genome of hundreds to thousands of input species, ALLEGRO extracts Cas9 target sequences. Step (2): ALLEGRO builds and solves an (integer) linear program involving millions of variables. Step (3): The optimal solution of the linear program determines the sgRNA library with minimal size that covers all targets.

target a given set of genes across diverse species. The set of guides can be designed under user-defined constraints, which we call *tracks*. ALLEGRO allows users to choose two types of tracks, namely *A* for *any* and *E* for *each*. Tracks are also further characterized by the multiplicity factor *m*. When users choose track A_{m}, ALLEGRO's solution guarantees that any of the genes in every species is targeted at least *m* times by the guides in the library. When users choose track E_m, ALLEGRO's solution guarantees that each gene must be targeted at least *m* times in each species. A detailed description of the mathematical formulation, tracks, and multiplicity is provided in Supplementary Notes, and a more detailed overview is provided in Supplementary Fig. S1.

Experimental validation

Microbial strains and culturing

Yarrowia lipolytica PO1f (*MatA*, *leu2-270*, *ura3-302*, *xpr2-322*, and *axp-2*), *Kluyveromyces marxianus* CBS6556 (*ura3Δ*), *Komagataella phaffii* GS115 (*his4::CAS9*), *Saccharomyces cerevisiae* BY4741 *(MATa, his3Δ1, leu2Δ0, met15Δ0,* and *ura3Δ0)*, and *Rhodotorula araucariae* NRRL Y-17376 were used in ALLEGRO guide validation experiments (Supplementary Table S1). Unless stated otherwise, yeast cultures were grown in 14 mL of polypropylene tubes or 250 mL baffled flasks at 30 °C or 37 °C with shaking at 225 rpm. Under nonselective conditions, yeast strains were cultivated in YPD medium (1% Bacto yeast extract, 2% Bacto peptone, and 2% glucose).

Y. lipolytica and *K. marxianus* transformants were initially propagated at 30 °C in Synthetic Defined medium lacking uracil (SD–ura) composed of 0.67% Difco yeast nitrogen base (YNB) without amino acids, 0.069% Complete Supplement Mixture without uracil (CSM–ura; Sunrise Science, San Diego, CA), and 2% glucose. *K. phaffii* transformants were selected in Synthetic Defined media lacking histidine (SD–his; 0.67% Difco YNB without amino acids, 0.069% CSM–his; Sunrise Science, San Diego, CA, and 2% glucose) at 30 °C.

R. araucariae and *K. marxianus* protoplasts were mixed with regeneration agar solution and incubated at 30 °C for 5–6 days to allow for genome edits to occur. All transformations were performed in a minimum of three independent biological replicates.

Plasmid construction

pSC012 [16], pIW601 [17], pCRISPRpp [18], and pCAS [19] were used as the backbone plasmids to express Cas9 and sgRNAs in *Y. lipolytica*, *K. marxianus*, *K. phaffii*, and *S. cerevisiae*. Cloning sgRNAs in the first three plasmid backbones was carried out following the method in [20]. Each sgRNA was ordered as a primer from Integrated DNA Technology (IDT) with 20 bp of homology upstream and downstream of the AvrII, PspXI, and BbvCI cut sites in pSC012, pIW601, and pCRISPRpp plasmids, respectively. 60–bp top and bottom strands were ordered and annealed together. The annealed strand and digested plasmid were assembled using Gibson Assembly in a 3:1 molar ratio (insert:vector). To construct the sgRNA plasmids for *S. cerevisiae*, 33 bp–long single-stranded oligo pairs were ordered from IDT and designed to reconstitute part of the gRNA scaffold in the pCAS derivative pMLL033 as well as having 4 bp–long 5′ overhangs when the oligos were annealed together. These duplexed oligos were inserted into pMLL033 via Golden Gate using NEBridge® Golden Gate Assembly Kit (BsaI-HF® v2; New England Biolabs, Ipswich, MA, USA). For the Golden Gate reaction, ∼55 fmol of backbone was mixed with 1 µl on 100 µM duplexed oligos and incubated at 37 °C for 1 h.

The assembled plasmids were then transformed directly into electrocompetent *Escherichia coli* TOP10 cells. *E. coli* cultures were grown in Luria Broth (LB; Sigma–Aldrich) with either 100 mg/l ampicillin or 50 mg/l kanamycin at 37 °C in 14 mL of polypropylene tubes, at 225 rpm. Plasmid isolation was performed using the Zymo Research Plasmid Miniprep Kit. All primers used for colony polymerase chain reaction (PCR), as well as plasmids and sgRNAs used in this work, are listed in Supplementary Tables S2–S4.

Ribonucleoprotein complex assembly

For the generation of sgRNAs to form Cas9–RNP complexes, the EnGen® sgRNA Synthesis Kit, *S. pyogenes* (NEB #E3322), was used according to the manufacturer's instructions. Target-specific DNA oligonucleotides were designed using the EnGen sgRNA Template Oligo Designer and checked for the presence of a "G" at the 5′ end. If absent, a "G" was manually added. The resulting templates were then used for *in vitro* transcription with the EnGen® sgRNA Synthesis Kit. The transcribed sgRNA was treated with DNase I (NEB #E3322) and purified using the RNA Clean-Up Kit (NEB #T2040). Before RNP assembly, the sgRNA was denatured and refolded as described in [21].

RNPs were assembled following [22] with slight modifications. Cas9–RNPs were immediately assembled before experiments in a 50 µl reaction in buffer B (25 mM $CaCl_2$, 10 mM Tris–HCl, 1 M sorbitol, pH 7.5) containing 5 µl of 1X NEBuffer™ r3.1 (100 mM NaCl, 50 mM Tris–HCl, 10 mM $MgCl_2$, and 100 µg/mL Recombinant Albumin, pH 7.9), and 1:1 molar ratio of sgRNA:Cas9 at 37 °C water bath for 10 min.

CRISPR genome editing with ALLEGRO-designed sgRNAs

We instructed ALLEGRO to design an sgRNA library with minimal size targeting a counter-selectable marker gene set (*S. cerevisiae* orthologs of *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3*) across 2263 fungal species using track E_1, meaning that each gene in the set must be targeted at least once in each species (list of genes in Supplementary Table S5, and list of species in Supplementary Data S1). To validate the on-target activity of this sgRNA library, we conducted CRISPR–Cas9 gene knockout experiments across four yeast species, namely *K. marxianus*, *K. phaffii*, *Y. lipolytica*, and *S. cerevisiae*. Gene sets tested in each species included *CAN1*, *GAP1*, *LYS2*, and *TRP1* in *Y. lipolytica*; *CAN1*, *FCY1*, and *URA3* in *K. marxianus*; and *CAN1* and *FCY1* in both *K. phaffii* and *S. cerevisiae*. The experimental workflow began with the selection of individual sgRNAs from the E_1 library, targeting each set of auxotrophy-related genes in the desired genome and cloning into pIW601, pCRISPRpp, pSC012, and pCAS plasmid backbones, respectively, or assembled into Cas9–RNP complexes. To compare the efficiency of the sgRNAs, a set of active sgRNAs from previously designed and validated libraries [16, 18, 23] were selected as controls. All control sgRNAs are listed in Supplementary Table S4. These plasmids and RNP complexes were subsequently transformed into yeast cells and plated on synthetic defined media (SD) with and without the inhibitor of the targeted counter-selectable genes. To identify gene mutations, the chemical inhibitors specific to the auxotrophy-associated genes were added to the SD plates (0.67% Difco YNB without amino acids, 0.079% CSM; Sunrise Science, San Diego, CA, 2% glucose, and 2% agar). To control for false-positive colonies, yeast transformants or protoplasts were also transformed without the respective sgRNA, using either an empty vector or a Cas9 without a guide. Colonies appearing on the chemical inhibitor plates were randomly picked and analyzed for frameshift or INDEL mutations. The list of chemical inhibitors associated with each gene is as follows: *CAN1*: L-Canavanine, *FCY1*: 5-fluorocytosine (5-FC), *GAP1*: L-Histidine, *LYS2*: α-aminoadipic acid, *TRP1*: 5-fluoroanthranilic acid (5-FAA), and *URA3*: 5-fluoroorotic acid (5-FOA). Selections for the loss of *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3* genes on SD plates containing L-canavanine, 5-FC, minimal L-proline+D-histidine medium (MPDHis), α-aminoadipic acid, 5-FAA, and 5-FOA, respectively, were carried out using the protocols described in [24–29].

Yeast transformation and screening

Y. lipolytica transformation was conducted by using the protocol [16]. Briefly, a single colony of the PO1f strain was grown in 2 mL of YPD liquid culture in a 14 mL culture tube at 30 °C with shaking at 225 rpm for 18 h (final OD600 ∼30). A total of 350 µl of the culture was pelleted by centrifugation at 4000 × *g* for 2 min and resuspended in 300 µl of transformation buffer [45% PEG 4000, 0.1 M lithium acetate (LiAc), and 100 mM dithiothreitol (DTT)]. Next, 500 ng of plasmid DNA and 80 µg of 10 mg/mL Salmon Sperm DNA (ssDNA, Agilent) were added to the mixture by thoroughly pipetting. Following incubation for 1 h at 39 °C, 1 mL of water was added, and the cells were pelleted and redistributed in 2 mL of SD-ura. After 2 days of growth, the cells were plated at a 10^{-2} dilution on SD + L-Canavanine (60 mg/l), SD + MPDHis (1.5516 g/l), SD + α-aminoadipic acid (4.36 g/l), and SD + 5-FAA (1.5 g/l) to counter-select for *CAN1*, *GAP1*, *LYS2*, and *TRP1* knockouts, respectively. Colonies were randomly picked from plates, and the targeted sequence of the gene was PCR amplified. Gene knockout was verified by Sanger sequencing.

K. marxianus transformations were carried out using a modified protocol described in [17]. Briefly, a single colony of *K. marxianus* CBS6556 *ura3Δ* strain was inoculated into 2 mL of YPD medium in a 14 mL tube and incubated overnight. This culture was used to inoculate a 250 mL baffled shake flask containing 50 mL of fresh YPD at a starting OD of 0.05, which was then grown for ∼13 h (final OD ∼15–18). For each transformation, 7×10^8 cells were collected and centrifuged at 4000 × *g* for 1 min at 4 °C (1 OD600 = 1.4×10^7 cells/mL). The cells were washed three times with 1 mL of the wash buffer (1 mM EDTA and 0.1 M LiAc) and then resuspended in 500 µl of the transformation buffer (38% PEG 4000, 1 M LiAc, 10 mM DTT, 10 mM Tris–HCl, and 10 mM EDTA). Then, 4 µg of plasmid DNA and 150 µg ssDNA were added to the mixture, and the cells were incubated at room temperature for 15 min. Subsequently, the mixture was heat shocked in a 47 °C water bath for 9 min. Following the heat shock, the cells were pelleted, the supernatant was removed, and cell pellets were resuspended into 2 mL of SD-ura selective media for 2 days at 30 °C. After 2 days, the cells were plated at a 10^{-2} dilution on SD + L-Canavanine (50 mg/l) and SD + 5-FC (0.129 g/l) to counter-select for *CAN1* and *FCY1*, respectively. After 2–3 days of incubation at 30 °C, colony PCR was performed to amplify the targeted region, which was followed by Sanger sequencing to identify frameshift mutations leading to gene knockouts.

K. phaffii transformation was carried out using the method in [18]. Briefly, 2 mL of YPD was inoculated with a single colony of *K. phaffii* GS115 *his4::CAS9* and incubated overnight. A total of 4×10^7 cells were transferred to 150 mL of YPD in a 500 mL baffled shake flask and grown for ∼14 h when the final OD600 was ∼1.8. 100 mL of cells were chilled on ice for 1.5 h, washed three times with 1 M ice-cold sorbitol, incubated with 25 mL of pretreatment solution (0.1 M LiAc, 30 mM DTT, 0.6 M sorbitol, and 10 mM Tris–HCl, pH 7.5) for 30 min at room temperature, and washed three more times with 1 M ice-cold sorbitol. For each transformation, 8×10^8 cells were mixed with 2 µg of plasmid to a final volume of 80 µl, incubated on ice for 15 min, and pulsed at 1.5 kV with Bio-Rad MicroPulser Electroporator in an ice-cold 0.2-cm-gap cuvette (1 OD600 = 5×10^7 cells/mL). Immediately after electroporation shock, 1 mL of ice-cold solution of YPD, and 1 M sorbitol was added to each cuvette. Cells were transferred to 1 mL of YPD and 1 M sorbitol in 14 mL tubes, incubated for 3 h at 30 °C and 225 rpm for recovery, washed with 1 mL of room-temperature autoclaved water to dispose of the excess plasmid DNA in samples, and transferred to SD-his selective media for 3 days at 30 °C. Cells reached confluency after 3 days and were transferred into 2 mL of fresh SD-his media with a starting OD600 of 1 to perform outgrowth experiments and were allowed to grow for 3 more days. After 6 days of growth, the cells were plated at OD600 = 1 on SD + L-Canavanine (50 mg/l) and SD + 5-FC (0.129 g/l) to counter-select for *CAN1* and *FCY1*, respectively. Colonies were randomly picked from plates, and the targeted sequence of the gene was PCR amplified, and gene knockout was verified by Sanger sequencing. Centrifugation was done at 3000 × *g* for 5 min at 4 °C.

Transformation of *S. cerevisiae* was carried out using a standard LiAc/SS carrier DNA/PEG protocol [30]. Briefly, 5 mL of 2X YPD was inoculated with a single colony of *S. cerevisiae* BY4741 and grown overnight at 30 °C with shaking at 220 rpm. This culture was used to inoculate a 125 mL of baffled flask containing 50 mL of fresh 2X YPD at a starting OD of 0.5 and grown for 3–4 h until it reached an OD of 2. Cells were washed, centrifuged, and aliquoted. The competent cells were transformed using 2 μg of DNA and heat shocked at 42 °C for 1 h. Immediately after heat shock, cells were centrifuged at 13 000 × *g* for 30 s, the supernatant was discarded and the cells were resuspended in 1 mL of YPD. After 2 h of outgrowth at 30 °C, cells were transferred to a 14 mL of culture tube containing 1 mL of YPD and 400 μg/mL of G418, resulting in a 2 mL culture with a G418 concentration of 200 μg/mL. After 3 days of growth, cells were diluted to an OD of 0.01. Hundred microliters of the dilution was then plated on SD + L-canavanine (60 mg/l) and SD + 5-FC (150 mg/l) to counter-select for *CAN1* and *FCY1*, respectively. After 2–3 days of incubation at 30 °C, colony PCR was performed to amplify the targeted region which was followed by Sanger sequencing to identify mutations leading to gene knockouts.

Protoplasts preparation and transformation

Protoplasts of *K. marxianus* CBS6556 were prepared according to a previously described method [31] with slight modifications. Briefly, three fresh colonies of *K. marxianus* CBS6556 were scraped from the YPD plate and grown at 30 °C overnight in a 250 mL flask containing 50 mL of YPD medium. The following day, the overnight culture was transferred to 50 mL of fresh YPD medium to prepare a culture with a starting OD600 of 0.05. The cells were grown at 30 °C for 10–18 h until OD600 of 6–8. From the culture, 25 mL was transferred into a 50 mL conical tube and then centrifuged at 1000 × *g* for 5 min at 5 °C. The supernatant was discarded, and the cells were resuspended in 30 mL of sterile water. Cells were centrifuged again at 3000 × *g* for 5 min at 5 °C, the supernatant was discarded, and cells were resuspended in 20 mL of citrate phosphate buffer (10 mM citrate phosphate, 1.5 M sorbitol, pH 6.8). Then, 70 μl of 10 mg/mL Zymolyase 20T solution (#E1005, Zymo Research), 25% w/v glycerol, and 50 mM Tris–HCl, pH 7.5), and 40 μl of β-mercaptoethanol (BME) were added. The samples were mixed gently by vortexing and then incubated at 30 °C for 45 min. Protoplast formation was monitored by measuring the OD600 of samples diluted 1:10 in 1.5 M sorbitol and comparing it to the OD600 in 5% sodium dodecyl sulfate. Protoplast preparation was terminated for transformation when the ratio reached 5. After incubation at 30 °C, protoplasts were collected by centrifugation at 600 × *g* for 10 min at 5 °C. After centrifugation, the supernatant was discarded, and the cells were resuspended in 10 mL of 1.5 M sorbitol with a wide-bore 5 mL pipette tip and then supplemented with more 1.5 M sorbitol to achieve a final volume of 30 mL. Next, protoplasts were centrifuged at 700 × *g* for 10 min at 5 °C, and the washing step with 30 mL of 1.5 M sorbitol was repeated once more. Finally, protoplasts were resuspended in STC solution (1.5 M sorbitol, 10 mM Tris–HCl, 50 mM $CaCl_2$, pH 7.5) to a final concentration of 3–8 × 10^8 cells/mL.

For the *K. marxianus* CBS6556 protoplast transformation, 200 μl of protoplast suspension was mixed with 50 μl of the RNP mix and Triton X-100 [0.006% (w/v) final concentration in transformation reaction) in a 14 mL tube to enhance cell membrane permeability [32]. The mixture was incubated on ice for 25 min, as prolonged incubation time may improve editing efficiency [33]. Then, protoplasts were supplemented with 1 mL of transformation buffer (40% PEG 4000, 10 mM Tris–HCl pH 7.5, and 50 mM $CaCl_2$). Samples were mixed thoroughly by flipping over 5–10 times and then incubated on ice for 1 h before centrifugation at 500 × *g* for 5 min at 5 °C. Supernatants were discarded, and pellets were resuspended in 800 μl of STC solution gently and were mixed with protoplast regeneration agar, which consisted of 6.9 g/l yeast nitrogen base without amino acids, 2% glucose, 0.8 M sorbitol, 3% agar, 20 mg/l L-adenine hemisulfate, 20 mg/l L-arginine·HCl, 20 mg/l L-Histidine·HCl, 30 mg/l L-Isoleucine, 100 mg/l L-leucine, 30 mg/l L-lysine·HCl, 20 mg/l L-methionine, 50 mg/l L-phenylalanine, 200 mg/l L-threonine, 20 mg/l L-tryptophan, 30 mg/l L-tyrosine, 20 mg/l uracil, and 150 mg/l L-valine. The regeneration agar was melted and equilibrated at 47 °C in advance. Transformed protoplasts were mixed with 10 mL of regeneration agar and 5-FOA (1 g/l) in a 15 mL falcon tube. The tubes were flipped three to five times to mix the contents, and the mixture was poured onto the surface of SD + 5-FOA plates (1 g/l). Plates were incubated at 30 °C until transformants formed, and then colony PCR and Sanger sequencing were performed to find frameshift mutation gene knockouts.

Protoplast transformation was also performed for the *R. araucariae* NRRL Y-17376 strain. We used the frozen protoplast protocol following a previously described method [34] with slight modifications. Briefly, single colonies of NRRL Y-17376 were separately inoculated in 50 mL of YPD medium at 30 °C for 15 h. Cells were harvested at 3000 rpm for 10 min and suspended in 20 mL autoclaved H_2O. Cells were harvested again, gently resuspended in 10 mL of 1 M sorbitol, and subsequently harvested. Finally, the cells were suspended in 10 mL SCEM (1 M sorbitol, 0.1 M sodium citrate, 10 mM EDTA, and 30 mM 2-mercaptoethanol, pH 5.8). Then, cells were mixed with 40 μl of lyticase solution (25 000 U/mL, Sigma–Aldrich) and incubated at 30 °C for 1 h. Following lyticase digestion, the cells were harvested by centrifugation at 1200 × *g* for 10 min and resuspended in SCEM to a final concentration of 10^9 cells/mL. Subsequently, 0.5 mL of lytic enzyme solution [1.5% (w/v) Zymolyase Ultra 2000U) was added to 1 mL of the cell suspension, and the mixture was incubated overnight at 30 °C. Next, the cells were centrifuged gently at 300 × *g* for 5 min at 4 °C in round-bottom plastic tubes and suspended in 10 mL of 1 M sorbitol by gently tapping the tube. Then, cells were centrifuged at 300 × *g* for 5 min. This procedure was repeated two more times to remove lyticase thoroughly, with the supernatant being discarded at each step. Finally, cells were suspended in 2 mL of CaST solution (1 M sorbitol, 10 mM Tris–HCl, 10 mM $CaCl_2$, pH 7.5) along with 2 mL of cell storage solution, and the protoplasts were stored at −80 °C.

For the *R. araucariae* protoplast transformation, 200 μl of protoplast suspension was mixed with 150 μl of the RNP mix and Triton X-100 [0.006% (w/v) final concentration in transformation reaction] in a 2 mL tube. The mixture was kept on ice for 25 min. Then, the cells were transferred to an ice-cold 0.2-cm-gap aluminum cuvette and electroporation was performed (1.2 kV or 400 V, 400 Ω and 25 μF capacitance), using a Bio-Rad MicroPulser Electroporator. After electroporation, the cells were resuspended in 1 mL of ice-cold YPD and transferred into new tubes on ice for 30 min, and then incubated at 30 °C for 4 h. Protoplasts were mixed with protoplast regeneration agar consisting of 5 g/l glucose, 4 g/l KH_2PO_4, 0.5 g/l Na_2HPO_4, 3.0 g/l NH_4C_1, 0.5 g/l NaCl, 0.4 g/l $MgSO_4·7H_2O$, 0.01 g/l $CaCl_2·2H_2O$, 0.008 g/l $FeCl_3·6H_2O$, 0.0001 g/l $ZnSO_4·7H_2O$, 0.069% CSM, pH 5.5, supplemented with 1.5 M sorbitol, and 2.5% agar. The regeneration agar was melted and equilibrated at 47 °C in advance. Transformed protoplasts were mixed with 10 mL of regeneration agar, respective concentrations of α-aminoadipic acid and 5-FC in a 15 mL Falcon tube. The tubes were flipped over three to five times quickly and then the mixture was poured on the top of SD + α-aminoadipic acid, and SD + 5-FC plates. Plates were incubated at 30 °C until transformants formed, and then colony PCR, and Sanger sequencing were performed to find frameshift mutation gene knockouts.

Results

To create minimal Cas9 sgRNA libraries using ALLEGRO's two design tracks across the fungal kingdom, we downloaded the genomes, protein sequences, and GFF files for 2263 fungal species from NCBI [35], FungiDB [36], EnsemblFungi [37, 38], and MycoCosm [39], the list of which is provided in Supplementary Data S1. We extracted the transcriptome of each genome using the corresponding GFF and recorded the intron–exon boundaries. An illustrative tree that shows the composition of the dataset is provided in Supplementary Fig. S2. The scripts and datasets, along with the documentation on how we carried out this step, can be found at https://github.com/ucrbioinfo/fugue.

Nine guides target anywhere in the transcriptome of one thousand *Ascomycota*

In this first experiment, we set out to test ALLEGRO's scalability by designing a minimal sgRNA library that targets at least once anywhere across the entire transcriptome of a large number of fungal species. Out of the 2263 fungal genomes in our dataset, the phylum *Ascomycota* had the greatest number of representative species. We therefore selected 1000 *Ascomycota* species for this proof-of-concept experiment, as this densely sampled and phylogenetically diverse clade provided a rigorous test case for ALLEGRO's performance. Given this dataset, ALLEGRO computed the smallest possible sgRNA library that ensures at least one cut anywhere across the entire transcriptome of each 1000 *Ascomycota* species (the full species list is provided in Supplementary Data S2).

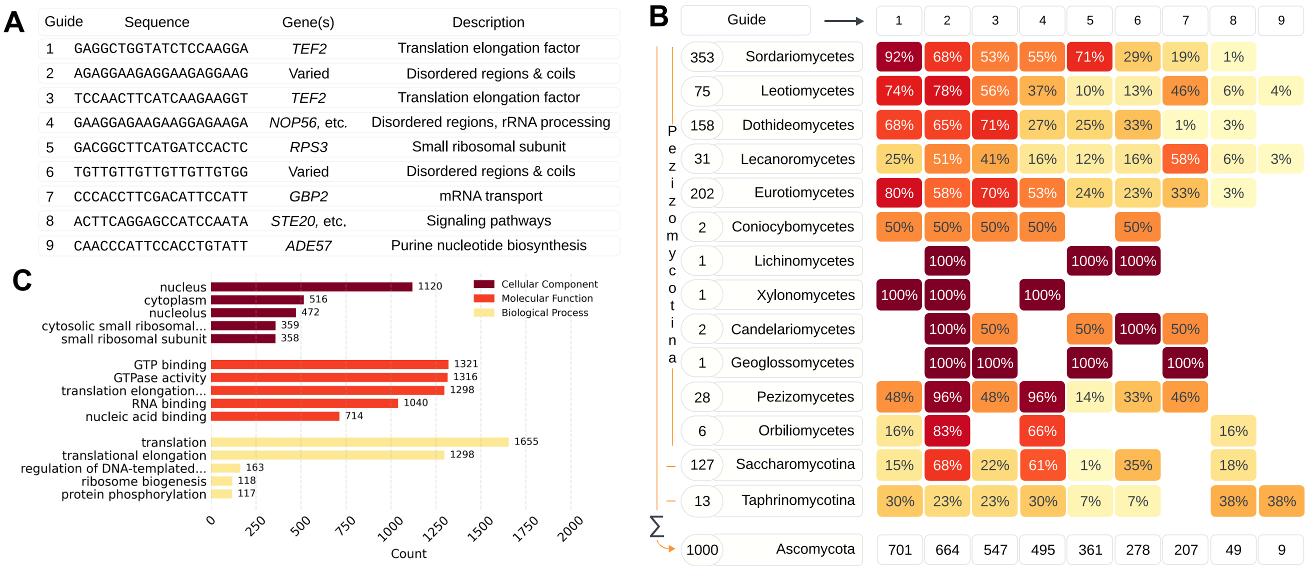
Configured with track A_1, and given the entire transcriptome of 1000 *Ascomycota* species, ALLEGRO considered more than a billion candidate guides with unique or overlapping targets. This high number of guides remains in the set even after removing guides containing homopolymers of length four or higher (e.g. TTTT), and guides with a GC content outside of the 40%–60% range, factors known to affect sgRNA efficiency [40–42]. ALLEGRO, however, can safely ignore millions of potential guides without affecting the quality or size of the final library by employing a heuristic which we call the *redundancy threshold* r. For example, in A_m we can safely discard all except m guides that target up to the exact same r species without affecting the final size of the library. This concept extends to track E_m as well. In the E_m case, we discard all but m guides that target the exact same set of genes of size up to r (if configured to use scores, we would keep the best m guides). We outline a more detailed description of this heuristic in Supplementary Notes Section S1.6. Using a redundancy threshold of 1000—equal to the number of input species for A_1—ALLEGRO reduced the candidate guide set from over one billion to ∼8.2 million, and solved the resulting linear program with 8.2 million variables in ∼13.5 h using roughly 400 GB of RAM.

ALLEGRO produced a notably small solution for the A_1 track for the 1000 *Ascomycota*. While a naive approach would select one guide per species totaling 1000 guides, ALLEGRO revealed that a library of only nine guides can target at least one gene in every species. Each guide, its sequence, most targeted gene, and description of gene function are shown in Fig. 2A. Figure 2B illustrates the 1000 *Ascomycota*, and the number of species in each group targeted by each guide. The first few guides target the majority of the species, while the last two guides are needed to achieve the full coverage of the 1000 species, as counted in the bottom row of Fig. 2B.

Given the small number of guides needed to cover all species, we examined whether these guides were more likely to target species that are closely related in terms of evolution. We found that many of the shared targets corresponded to highly conserved functions such as translation and RNA processing. While this was not used as a formal validation metric during guide selection, the functional enrichment served as a post hoc sanity check, supporting the biological plausibility of the selected guides, which increased our confidence in ALLEGRO's performance.

Guide 1, which targets exclusively *TEF2* orthologs, affects mostly fungal species in the *Pezizomycotina* family, except for those in *Lecanoromycetes* and *Orbiliomycetes*. In fact, 92% of all *Sordariomycetes* and 80% of *Eurotiomycetes* in the dataset are targeted by this guide, i.e. they share this exact DNA sequence in their *TEF2* orthologs. Guides 2 and 4 target almost all *Pezizomycetes* and more than half of *Sordariomycetes*, *Eurotiomycetes*, and *Saccharomycotina*. Guide 7 mostly targets the *GBP2* gene in almost half of the species from *Leotiomycetes*, *Lecanoromycetes*, and *Pezizomycetes*, and about one-third of those in *Eurotiomycetes*, with fewer targets in *Sordariomycetes*. The majority of the targets of Guide 5, which targets the orthologs of *RPS3*, are in *Sordariomycetes*, with a quarter of *Dothideomycetes* and *Eurotiomycetes* also targeted. Guides 8 and 9, having the fewest targets, cut mostly in *Saccharomycotina* and *Taphrinomycotina*, the two subdivisions in this dataset separate from the rest of the *Pezizomycotina*. Supplementary Figure S3 shows the number of species targeted for any subset of these guides. The upset plot shows a strong overlap between the sets: for instance, while Guide 1 targets 701 species overall, it targets only 17 species that no other guide targets.

Figure 2C illustrates a Gene Ontology (GO) enrichment analysis for genes targeted by the library guides. Notably, Guides 1 and 3 exclusively target orthologs of *TEF2*, which, according to [43], is a conserved gene crucial for translation elongation and accuracy. These guides specifically target the "Translation factor GTPase" family, focusing on the "Translational (tr)-type GTP-binding domain." Guide 4 primarily targets disordered regions of orthologs such as *NOP56*, *NOP6*, *NHP2*, *CBF5*, and *DBP3*, which are involved in ribosome biogenesis and rRNA processing, localized to the nucleus [44, 45]. Guide 5 exclusively targets orthologs of *RPS3*, an essential gene encoding proteins of the cytosolic small ribosomal sub-



Guide	Sequence	Gene(s)	Description
1	GAGGCTGGTATCTCCAAGGA	*TEF2*	Translation elongation factor
2	AGAGGAAGAGGAAGAGGAAG	Varied	Disordered regions & coils
3	TCCAACTTCATCAAGAAGGT	*TEF2*	Translation elongation factor
4	GAAGGAGAAGAAGGAGAAGA	*NOP56*, etc.	Disordered regions, rRNA processing
5	GACGGCTTCATGATCCACTC	*RPS3*	Small ribosomal subunit
6	TGTTGTTGTTGTTGTTGTGG	Varied	Disordered regions & coils
7	CCCACCTTCGACATTCCATT	*GBP2*	mRNA transport
8	ACTTCAGGAGCCATCCAATA	*STE20*, etc.	Signaling pathways
9	CAACCCATTCCACCTGTATT	*ADE57*	Purine nucleotide biosynthesis

	Guide →	1	2	3	4	5	6	7	8	9
353	Sordariomycetes	92%	68%	53%	55%	71%	29%	19%	1%	
75	Leotiomycetes	74%	78%	56%	37%	10%	13%	46%	6%	4%
158	Dothideomycetes	68%	65%	71%	27%	25%	33%	1%	3%	
31	Lecanoromycetes	25%	51%	41%	16%	12%	16%	58%	6%	3%
202	Eurotiomycetes	80%	58%	70%	53%	24%	23%	33%	3%	
2	Coniocybomycetes	50%	50%	50%	50%		50%			
1	Lichinomycetes		100%			100%	100%			
1	Xylonomycetes	100%	100%		100%					
2	Candelariomycetes		100%	50%		50%	100%	50%		
1	Geoglossomycetes		100%	100%		100%		100%		
28	Pezizomycetes	48%	96%	48%	96%	14%	33%	46%		
6	Orbiliomycetes	16%	83%		66%				16%	
127	Saccharomycotina	15%	68%	22%	61%	1%	35%		18%	
13	Taphrinomycotina	30%	23%	23%	30%	7%	7%		38%	38%
1000	Ascomycota	701	664	547	495	361	278	207	49	9

Figure 2. ALLEGRO scales to thousands of species for genome-wide library design. (**A**) ALLEGRO produced a library of nine guides that target at least one gene in 1000 *Ascomycota* species. The sequence of the guides in the library with a brief description of their functions are shown. (**B**) An illustration of the 1000 input *Ascomycota*, in which the numbers to the left of each group's name indicate how many species from that group are included in the dataset. Percentages in each column indicate the portion of each species group targeted by each guide. The numbers in the bottom row indicate the total number of targets by each guide. (**C**) The top five GO terms and their counts for each functional category for genes targeted by the nine guides, with each category representing biological process, molecular function, or cellular component.

unit (40S), which localize to both the nucleus and cytosol [46]. Guide 7 targets orthologs of *GBP2* and its paralog *HRB1*. These genes encode RNA-binding proteins involved in mRNA transport from the nucleus to the cytoplasm, playing a key role in mRNA export, especially under stress conditions. They are crucial surveillance factors for the selective export of spliced mRNAs in yeast, interacting with the nuclear RNA export factor *MEX67*. While nonessential, *GBP2* and *HRB1* are important for efficient RNA processing and stress response [47]. Guide 8 targets orthologs such as *STE20*, *KIC1*, and *CLA4*, members of the PAK (p21-activated kinase) family involved in signaling pathways regulating cell growth, polarity, and responses to environmental cues. *STE20* plays a key role in the mating and filamentous growth pathways, activated by GTP-bound *CDC42*, a small GTPase, to transduce signals for mating and growth [48]. *KIC1* is involved in cell wall integrity and polarity via the RAM pathway, regulating cell separation and polarized growth [49]. *CLA4*, also *CDC42*-activated, is essential for cytokinesis and polarity maintenance [50]. Lastly, guide 9 targets orthologs of *ADE57*, a gene involved in the *de novo* purine nucleotide biosynthesis pathway. This bifunctional enzyme, comprising aminoimidazole ribotide synthase and glycinamide ribotide synthase activities, is essential for purine synthesis [51].

Further analysis of the target sites revealed that Guides 2 and 6 do not target a specific gene or gene family but instead a diverse set of genes, predicted to target disordered protein regions (Supplementary Data S3). These disordered regions contain shared sequences across many genes, explaining the broad targeting of these guides. Additionally, guides 2, 4, and 6 exhibit significant similarity and high levels of repetition. According to [52], simple amino acid sequences are the most frequently occurring protein fragments in *S. cerevisiae*. Moreover, some studies (see e.g. [53–55]) suggest that intrinsically disordered protein regions are often composed of sequences that are compositionally biased and low-complexity, corroborating our findings.

ALLEGRO scales to auxotrophy-associated genes in over two thousand fungal species

A key application of ALLERGO is designing CRISPR guides that target many different species across a large number of organisms from the same kingdom. In this section, we demonstrate ALLEGRO's capability in designing sgRNA libraries targeting auxotrophy-related genes in various combinations (tracks and multiplicities) for 2263 fungal species where track A_1 designs a library of sgRNAs targeting any of the genes of interest at least once. Track A_6 is similar to A_1 but targets anywhere across the six genes at least six times. Track E_1 targets each individual gene at least once. By leveraging the counter-selectability of these genes and the sgRNA libraries designed by ALLEGRO, we aim to facilitate the domestication and onboarding of any fungal species.

We grouped the protein-coding genes for 2263 fungal species representing isolates from across the fungal kingdom into orthogroups using DIAMOND [14]. We used the protein sequences of *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3* in *S. cerevisiae* to identify the orthogroups for these genes. We anticipated that some species might lack one or more of the six auxotrophy-associated genes. In such cases, where a given gene has no identifiable ortholog in a target species, the corresponding sgRNAs are not included in the output sgRNA library. Therefore, guides for genes absent in a species are naturally excluded from the library without affecting the overall performance or applicability of the method. As such, all input species contain at least one gene with a best reciprocal hit to *S. cerevisiae* and ≥30% protein sequence identity, a threshold informed by prior studies on structural similarity [55–57].

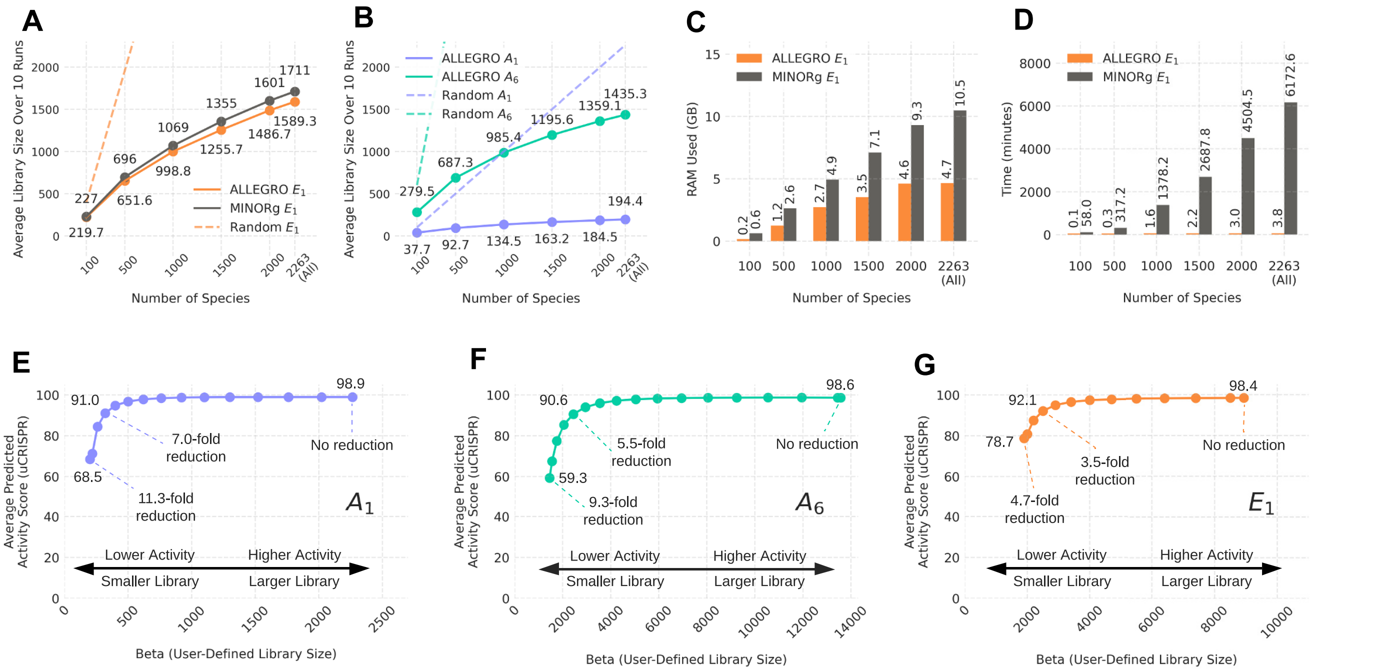


Figure 3. ALLEGRO enables the design of fungal kingdom-wide sgRNAs. (**A**) Library sizes generated by MINORg and ALLEGRO track E_1 as a function of the number of species. In this comparison, we allowed guides crossing intron–exon boundaries in ALLEGRO as it is done in MINORg and configured ALLEGRO with a redundancy threshold set to the maximum number of genes available. Random E_1 refers to a naive sgRNA selection strategy where a random guide per gene is chosen. (**B**) Library sizes for ALLEGRO tracks A_1 and A_6 as a function of input size. Here, the redundancy threshold was set to the maximum number of input species. (**C**) The amount of memory used (in gigabytes) by MINORg and ALLEGRO (track E_1) as a function of the number of input species. (**D**) Running time (in minutes) of ALLEGRO (track E_1) and MINORg as a function of the number of input species. (**E–G**) The beta parameter controls the trade-off between guide library size and the guides' average scores (computationally predicted by uCRISPR).

Library size as a function of the number of species

The objective of this experiment was to study the size of ALLEGRO's libraries for tracks A_1, A_6, and E_1 as a function of the number of input species. We also wanted to compare the size of the library produced by ALLEGRO in track E_1 to the size of the libraries produced by MINORg. Both ALLEGRO and MINORg were executed on the same inputs and used the same parameters, including the use of multi-threading, not checking for off-targets, allowing no mismatches, and allowing only guides with a GC content between 30% and 70%. The inputs and commands used to run MINORg, and the configuration files for ALLEGRO can be found in Supplementary Data S4.

We randomly sampled $x = \{100, 500, 1000, 1500, 2263\}$ species from the pool of 2263, each with at least one gene out of the six auxotrophy-associated genes. For each choice of x, we created ten samples of x species to reduce the sampling bias. ALLEGRO was executed on ten samples for all tracks, while MINORg was run on a single sample due to its high running time. To ensure fairness, for each species set, the sample where ALLEGRO's library size was closest to the average across the ten samples was selected as the input for MINORg. This approach was used to minimize variability and provide a consistent benchmark for comparing the performance of both tools. For both tools, we recorded the size of the library as well as the time/memory used. The results presented in Fig. 3A illustrate the average library sizes across the ten runs for ALLEGRO and the library sizes generated by MINORg. Supplementary Figurre S4 shows the means and standard deviations of ALLEGRO's library sizes, indicating stable and minimal variance between different runs of the algorithm.

Figure 3A shows the size of the library as a function of the number of species x. To establish a baseline, a naive algorithm, Random E_1, was also included. Random E_1 generates libraries by selecting one random sgRNA for each targeted gene in each species, ultimately dropping any potential duplicates. ALLEGRO's library size is smaller than MINORg's for all choices of x. Figure 3B shows the average sizes of the A_1 and E_6 libraries as a function of the number of species x. Figure 3C shows the peak RAM usage of MINORg and ALLEGRO in gigabytes. ALLEGRO uses almost half as much memory as MINORg given the largest available dataset. Additionally, ALLEGRO completes its tasks in minutes while MINORg may take up to multiple days to complete the calculation (Fig. 3D).

In the next ensemble of experiments, we added two extra criteria to our ALLEGRO libraries where we (i) allowed up to a single mismatch between the sgRNA sequence and its target in the PAM–distal region and (ii) did not allow sgRNAs that crossed an intron–exon boundary. We allowed the first criterion because we found that the library size can be significantly reduced if a mismatch is allowed. According to previous studies [1, 58–61], a single mutation in the 8 bases of the 5′ seed-distal region is tolerated, while mismatches in the first 12 bases upstream of the 3′ after the NGG PAM abolish Cas9 cleavage. The second criterion is important because an sgRNA that overlaps a splice site—while appearing to target a coding sequence—may be ineffective due to the discontinuity of genomic sequence after splicing.

Designing a library to target each auxotrophy-associated gene across all species at least once (track E_1) would triv-

Table 1. The size of ALLEGRO's libraries that target six auxotrophy-associated genes in 2263 fungal species for tracks A_1, A_6, and E_1. Allowing one mismatch in the PAM–distal region of the guides significantly reduces the library sizes. These experiments exclude sgRNA spanning intron–exon boundaries

Track	Library size	Library size (allowing one mismatch)
A_1	195	151
A_6	1436	872
E_1	1809	1485

ially require 13 578 sgRNAs (2263 species × 6 genes, 1 guide per gene). However, with ALLEGRO's E_1, this number can be drastically reduced to 1809 sgRNAs (a reduction of 86%) if single mismatches are not allowed and to 1485 sgRNAs (a reduction of 89%) if mismatches are allowed. The library sizes for tracks A_1, A_6, and E_1 for ALLEGRO can be found in Table 1.

Trade-off between set size and guide activity

In the experiments carried out so far, we have not considered guide activity. The objective of minimizing the size of the library competes against the objective of maximizing the predicted activity scores for the library. Allowing the library size to grow gives ALLEGRO the flexibility to choose low activity guides that may cover more targets and vice versa. This trade-off is controlled by the beta parameter, the library size budget. By setting beta to be higher than the smallest library size possible, we allow for the average predicted activity scores for the library to increase. In these experiments, we computationally predicted the cutting activity of our guides using the uCRISPR method [15]. Briefly, this unified, physical model predicts guide editing efficacy using the energetics of R-loop formation under Cas9 binding, the effect of the PAM sequence, and the folding stability of the whole sgRNA.

The trade-off between the library's size and its average predicted activity score is illustrated in our experiments for tracks A_1, A_6, and E_1 on the six auxotrophic marker genes *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3* from 2263 fungal species (Fig. 3E–G). For track A_1, the largest possible beta (corresponding to the maximum size of the library) is 2263 when choosing the highest activity guide per species. As beta decreases, the library size becomes smaller at the expense of the average activity score. With a beta of 320, the library contains 320 guides (a 7-fold reduction) and the average activity score for the library is still quite high at 91.0. The smallest library possible is for beta = 212 which gives an 11.3-fold reduction from the original library size, but the average predicted score is low at 68.7. We observe a similar trade-off for track A_6 (Fig. 3F) where for beta = 2450, the library reduced 5.5-fold from the maximum size while maintaining a high average guide activity score of 90.6. The smallest beta for this experiment is 1450 (a 9.3-fold reduction) with a low average guide score of 59.2. A similar trend is observed in track E_1 (Fig. 3G) where with a 3.5-fold reduction at beta = 2500, the average guide score is 92.1. With a 4.7-fold reduction at beta = 1900, we have a library with an average guide score of 78.7. For A_6 the largest beta is 2263 × 6 = 12 330 where each species may be targeted six times. For E_1, the largest beta is 8932 which is the total number of available orthologous genes across the 2263 species. All data used to produce Fig. 3 may be found in Supplementary Data S4.

Track E_1 achieves validation-grade efficiency in fungi

To demonstrate ALLEGRO's capability in designing active sgRNAs, we experimentally validated the targeting efficiency of sgRNAs designed for the track E_1 library that target the counter-selectable marker gene set (*CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3*). In these experiments, four yeast species—*Kluyveromyces marxianus*, *Komagataella phaffii*, *Yarrowia lipolytica*, and *Saccharomyces cerevisiae*—were transformed with plasmid DNA consisting of Cas9 and sgRNA expression cassettes optimized for each species, with an illustration shown in Fig. 4A. Colonies on selective media with chemical inhibitors, which required gene disruption for growth, were analyzed for knockouts (Fig. 4B). These chemical inhibitors—L-Canavanine, 5-FC, MPDHis, α-aminoadipic acid, 5-FAA, and 5-FOA—were used to counter-select for functional disruptions in *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3*, respectively. Figure 4C illustrates the cutting efficiency results for the sgRNAs designed by ALLEGRO across a diverse set of auxotrophy-related genes in the *K. marxianus*, *K. phaffii*, and *Y. lipolytica* genomes. Notably, ALLEGRO-designed sgRNAs exhibited high cutting efficiency—achieving 90%–100% efficiency when targeting *CAN1* and *TRP1* in *Y. lipolytica*, *CAN1* and *FCY1* in *K. marxianus*, and *FCY1* in *K. phaffii*—and outperformed the control sgRNAs across the targeted genes. ALLEGRO was also successful in designing effective guides for editing in *S. cerevisiae*, the species that was used to train the guide design algorithm (Supplementary Fig. S5 and Supplementary Table S4). Additional data regarding the number of colonies observed on both control and selective media are shown in Supplementary Figs S5–S8. Aside from targeting specific genes in *K. marxianus*, *K. phaffii*, *Y. lipolytica*, and *S. cerevisiae*, Supplementary Data S5 provides a phylogenetic analysis of the broader fungal groups targeted by track E_1 sgRNAs validated in these species. This analysis highlights the evolutionary relationships and demonstrates the broader applicability of the designed sgRNAs across related fungal groups.

The next validation experiments assessed the broader applicability of ALLEGRO's sgRNAs to enable genome editing in fungi using a universal transformation method. This approach combined Cas9–RNP complexes with protoplast transformation. Initially, we demonstrated the feasibility of protoplast transformation in the *K. marxianus* genome by targeting the *CAN1* gene with plasmid DNA. We then showed that this method, coupled with RNP complexes, efficiently targets the *CAN1* gene (Fig. 4D). Notably, plasmid-based protoplast transformation yielded lower editing efficiency than the Cas9–RNP approach, possibly due to Cas9 toxicity from prolonged expression and the reduced uptake of large Cas9-expressing plasmids (8–12 kb) [62, 63]. We also demonstrated that Cas9–RNPs complexed with ALLEGRO-designed sgRNAs efficiently knockout the *URA3* gene with a cutting efficiency of ∼60% in *K. marxianus*. Supplementary Figure S9 provides additional data on the number of colonies observed on both control and selective media for *K. marxianus* protoplast transformation targeting the *CAN1* gene.

Kingdom-wide guide design with ALLEGRO

One of the goals of our work is to design a set of universal guides that would target a given set of genes of interest in every fungal species. Since our genomic knowledge of fungi is limited to the ≈2000 species whose genomes have been sequenced

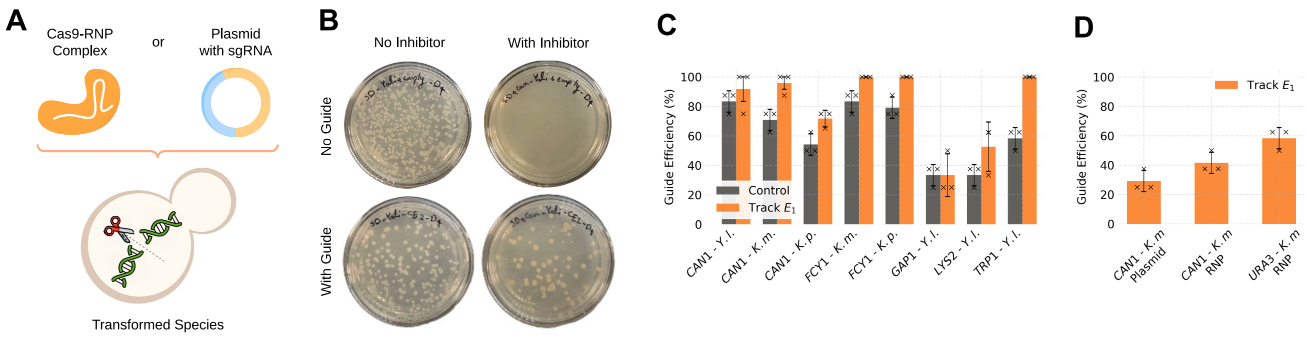


Figure 4. Experimental validation for ALLEGRO's track E_1. (**A**) Illustration of fungal transformation using either a Cas9–RNP complex or a plasmid expressing the sgRNA. (**B**) CRISPR–Cas9 knockout screening. Track E_1 and control sgRNAs targeting auxotrophy-related genes are cloned into plasmid backbones and transformed into the yeast of interest. Transformants are plated on standard control media (SD) and inhibitor-containing media, with experiments conducted in triplicate. The inhibitors associated to each gene are *CAN1*:L-Canavanine, *FCY1*:5-FC, *GAP1*:MPDHis, *LYS2*:α-Aminoadipic acid, *TRP1*:5-FAA, and *URA3*:5-FOA. Colonies growing on inhibitor media are validated using colony PCR. Selected images are an illustration of control and sample plates for targeting the *CAN1* gene in *Yarrowia lipolytica*. (**C**) CRISPR–Cas9 cutting efficiency scores of ALLEGRO-designed sgRNAs targeting auxotrophy-associated genes in *Yarrowia lipolytica* (Y. l), *Kluyveromyces marxianus* (K. m.), and *Komagataella phaffii* (K. p.). From each of the three biological replicates, eight colonies appearing on inhibitor-containing plates are randomly selected and sequence analyzed to confirm the presence of mutations. (**D**) CRISPR–Cas9 cutting efficiency scores of ALLEGRO-designed sgRNAs targeting *CAN1* and *URA3* genes in *K. marxianus* genome using protoplast transformation with plasmid or Cas9–RNP complexes. The error bars represent the standard deviation and bars represent the mean. Individual data points for all experiments are also shown.

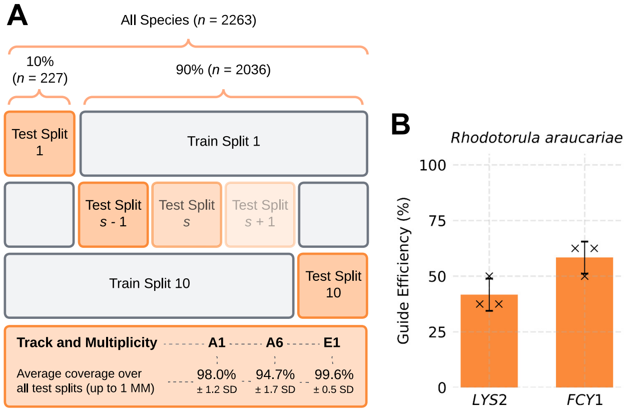


Figure 5. Robustness Analysis of ALLEGRO. (**A**) Cross-species validation results. 10% of the species were excluded as a test set, and the pipeline was 'trained' on the remaining 90%. The sgRNA library size and average coverage were calculated for the excluded test splits. The results demonstrate high coverage and highlight ALLEGRO's ability to design efficient sgRNAs across diverse species. (**B**) ALLEGRO's performance on fungal species not included in the training set. sgRNAs from track E_1 library were tested for targeting *FCY1* and *LYS2* in *R. araucariae*.

and annotated, in this experiment we attempt to quantify how guides designed for a subset of species would also work for other species not included in ALLEGRO's training set.

We evaluated the generalizability of ALLEGRO's sgRNA libraries using 10-fold cross-validation experiments for tracks A_1, A_6, and E_1 on the six counter-selectable marker genes *CAN1*, *FCY1*, *GAP1*, *LYS2*, *TRP1*, and *URA3*. In each calculation, we split the 2263 species into 90% training (2036 species) and 10% testing (227 species), as illustrated in Fig. 5A. We ran ALLEGRO on the training set of each split and determined whether the guides in the library would also cut these six genes in the 10% test species. Guide design was successful if the guide matched the gene sequence with up to one mismatch in the seed-distal region. We measured the performance of ALLEGRO on each test set by counting the number of species that (i) would have every gene cut somewhere at least once (track E_1), (ii) would have any of the six genes cut somewhere at least once (track A_1), or (iii) would have the set of genes cut somewhere at least six times (track A_6). If a sufficient number of cuts were produced in a species to satisfy the track conditions, we said that the species was covered.

We averaged the number of test species covered by ALLEGRO's libraries over the 10-fold cross-validation experiment. Track E_1 demonstrated the highest generalizability, with libraries covering an average of 99.6% of test species. Track A_1 also performed strongly, achieving an average coverage of 98.0%. In contrast, the least generalizable design was track A_6, with its libraries covering 94.7% of test species on average. These results suggest that ALLEGRO can reliably generate near-universal sgRNA sets for a given set of target genes, particularly when the training species are representative of broader fungal diversity. Scripts to reproduce the cross-validation experiment are included in Supplementary Data S6.

A central objective of this study is to demonstrate ALLEGRO's capacity to generate effective sgRNA libraries applicable to species not included in the original design set, thus proving the generalizability of the method. To this end, we selected *R. araucariae*, a species that was not part of the training set, for experimental validation to highlight the tool's applicability to diverse and previously untested fungi.

To begin, we performed a BLAST+ [64] analysis using the *R. araucariae* genome and the track E_1 sgRNA library. The analysis revealed that the library successfully targeted *LYS2*, *URA3*, and *FCY1* in *R. araucariae*. No orthologs for *CAN1* and *TRP1* were identified. Additionally, a BLAST search confirmed that no sgRNAs in the E_1 library were available to target *GAP1*. Based on these findings, we proceeded to validate sgRNAs from the E_1 library that specifically target the *FCY1* and *LYS2* genes (Fig. 5B). The sgRNAs were validated through a Cas9-knockout screening using protoplast transformation and RNPs. The list of validated sgRNAs is provided in the Supplementary Table S4. Further data regarding the number of transformants under control and selective conditions for *R. araucariae* transformation are presented in Supplementary Fig. S10. These results confirm ALLEGRO's capability to design active and efficient sgRNAs for genome editing in novel fungal species.

Discussion

In this work, we introduced a CRISPR–Cas9 guide design tool called ALLEGRO capable of efficiently designing guide libraries with minimal cardinality and scaling to two thousand species. ALLEGRO leverages advanced combinatorial optimization tools to find the smallest set of guides that target a set of genes of interest. Previous research has concentrated on designing guide libraries for relatively small sets of species and genes. Based on our experiments, these tools struggle with large datasets, either fail to produce any results, or consume substantial time and computational resources and often generate outputs that are suboptimal.

As a proof of scalability, we showed that ALLEGRO can design guides for all genes (i.e. the full transcriptome) of a thousand *Ascomycota* species. Due to the infeasibility of storing more than a billion candidate guides in main memory, we developed a new heuristic that can drastically reduce the number of candidate guides without affecting the optimality of the solution. Using the new approach, ALLEGRO identified just nine guides capable of targeting anywhere across the transcriptome of the 1000 input species.

We also carried out extensive experiments on a set of six auxotrophy-associated genes on more than two thousand fungal species. We showed that ALLEGRO produces a smaller guide library than a previous method called MINORg, requires less RAM, and is several orders of magnitude faster. We also showed that the size of the required guide library for tracks A_1, A_6, and E_1 tends to plateau as the number of input species increases, which is a required feature of a universal library that would target all the species within a kingdom. To this end, we performed several cross-validation experiments, designing guides for a portion of the input and testing them on a small subset of unseen species. By allowing a single base mismatch in the seed-distal region of the guide and its potential targets, we showed that we could meet the requirements for each track in over ∼95% of the held-out species on average.

A unique feature of ALLEGRO is that it can incorporate the predicted guide activity scores to design the optimal library. We showed that while the smallest guide library is likely to contain low-activity guides, a set with higher predicted activity may be achieved by allowing ALLEGRO to select a slightly larger number of guides. While ALLEGRO is designed to be fast and flexible, large experiments such as our full transcriptome experiment over a thousand species still require a significant amount of primary memory. To further address this bottleneck, new heuristics must be developed to discard guides without compromising the size of the final library. In addition, new linear programming solvers might yield better results in terms of library size, processing time, and memory usage.

We experimentally validated ALLEGRO's ability to design highly efficient and precise sgRNAs for genetic engineering through CRISPR–Cas9 knockout screens. Our results confirmed successful knockouts of auxotrophy-associated genes in multiple industrially relevant yeast species, including *Kluyveromyces marxianus*, *Komagataella phaffii*, *Yarrowia lipolytica*, and *Saccharomyces cerevisiae*, demonstrating the high efficiency and on-target activity of the sgRNAs designed by ALLEGRO. Furthermore, computational cross-validation and successful experimental validation in *Rhodotorula araucariae*, a species not included in the initial input set, underscore ALLEGRO's broad applicability for genome editing across diverse and novel fungal species.

While ALLEGRO tackles key computational limitations in cross-species CRISPR sgRNA design, a major obstacle in practical genome-editing applications is biological variability in transformation efficiencies. Factors such as species-specific defense mechanisms against foreign DNA, variation in cell wall composition, and differences in DNA repair pathways can significantly impact editing outcomes. In our validation experiments with both plasmid-based and Cas9–RNP based delivery systems, we noticed variations in transformation efficiency across fungal species. Addressing these biological barriers will require tailored experimental protocols, including optimization of transformation conditions, enhancing protoplast regeneration methods, and development of species-specific as well as multi-purpose genetic tools. While these biological challenges persist, the robustness of ALLEGRO's sgRNA designs enabled successful genome editing not only in various yeast species, but also in a previously untested fungal genome.

Finally, although ALLEGRO was developed and validated in fungal species, its design principles are broadly applicable to other organisms. The core algorithm operates on user-supplied coding sequences and does not depend on eukaryote-specific genomic features, making it adaptable to nonfungal eukaryotes and even prokaryotes. In prokaryotic genomes, additional considerations may include overlapping genes, operon structures, and organism-specific PAM recognition (e.g. for Cas12a or other non-Cas9 effectors). Adapting ALLEGRO for bacterial systems may therefore involve customizing PAM constraints and incorporating rules to avoid targeting within essential operons. Nonetheless, the underlying logic of optimizing multi-target sgRNA libraries remains generalizable across domains of life.

Acknowledgements

Author contributions: Amirsadra Mohseni: Conceptualization, Software, Methodology, Formal Analysis, Data Curation, Visualization, Writing—Original draft. Reyhane Ghorbani Nia: Conceptualization, Investigation, Methodology, Validation, Data Curation, Visualization, Writing—Original Draft. Aida Tafrishi: Conceptualization, Methodology, Writing—Original Draft. Mario León López: Investigation, Validation. Xin-Zhan Liu and Jason E. Stajich: Resources (*R. araucariae* genome sequence). USDA-ARS Culture Collection (NRRL): Resources (*R. araucariae* strain). Stefano Lonardi and Ian Wheeldon: Conceptualization, Supervision, Project Administration, Funding Acquisition, Visualization, Writing—Original draft. All authors: Writing—Review & Editing.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation, Division of Chemical, Bioengineering, Environmental and Transport Systems (NSF-2225878 to I.W. and S.L.) and by the National Science Foundation under Cooperative Agreement DBI-2400327 to I.W. and J.E.S. J.E.S. is a CIFAR fellow in the program Fungal Kingdom: Threats and Opportunities, and was partially supported by the National Science Foundation grant (EF-2125066) and the U.S. Department of Agriculture, National Institute of Food and Agriculture (CA-R-PPA-211-5062-H). Funding to pay the Open Access publication charges for this article was provided by National Science Foundation, Division of Chemical, Bioengineering, Environmental and Transport Systems (DBI-2400327).

Data availability

ALLEGRO is freely available on GitHub at https://github.com/ucrbioinfo/allegro and on Zenodo at 10.5281/zenodo.14768175. ALLEGRO's documentation is provided on its GitHub Wiki page. All data used in this study can be obtained from ALLEGRO's repository and the Supplementary Data. The pre-processing code for acquiring the input data is available at https://github.com/ucrbioinfo/fugue. The *R. araucariae* genome is available as NCBI BioProject accession PRJNA895933.

References

1. Jinek M, Chylinski K, Fonfara I *et al.* A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012;**337**:816–21. https://doi.org/10.1126/science.1225829
2. Li T, Yang Y, Qi H *et al.* CRISPR–Cas9 therapeutics: progress and prospects. *Signal Trans Targ Ther* 2023;**8**:36. https://doi.org/10.1038/s41392-023-01309-7
3. Tang Y, Fu Y. Class 2 CRISP–Cas: an expanding biotechnology toolbox for and beyond genome editing. *Cell Biosci* 2018;**8**:59. https://doi.org/10.1186/s13578-018-0255-x
4. Li C, Chu W, Gill RA *et al.* Computational tools and resources for CRISPR–Cas genome editing. *Genom Proteom Bioinform* 2023;**21**:108–26. https://doi.org/10.1016/j.gpb.2022.02.006
5. Karp RM. Reducibility among combinatorial problems. In: *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, 219–41. https://doi.org/10.1007/978-3-540-68279-0_8
6. Endo M, Mikami M, Toki S. Multigene knockout utilizing off-target mutations of the CRISPR–Cas9 system in rice. *Plant Cell Physiol* 2015;**56**:41–7. https://doi.org/10.1093/pcp/pcu154
7. Prykhozhij SV, Rajan V, Gaston D *et al.* CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. *PLoS One* 2015;**10**:e0119372. https://doi.org/10.1371/journal.pone.0119372
8. Hyams G, Abadi S, Lahav S *et al.* CRISPys: optimal sgRNA design for editing multiple members of a gene family using the CRISPR system. *J Mol Biol* 2018;**430**:2184–95. https://doi.org/10.1016/j.jmb.2018.03.019
9. Lee RR, Cher WY, Wang J *et al.* Generating minimum set of gRNA to cover multiple targets in multiple genomes with MINORg. *Nucleic Acids Res* 2023;**51**:e43. https://doi.org/10.1093/nar/gkad142
10. Zou RS, Marin-Gonzalez A, Liu Y *et al.* Massively parallel genomic perturbations with multi-target CRISPR interrogates Cas9 activity and DNA repair at endogenous sites. *Nat Cell Biol* 2022;**24**:1433–44. https://doi.org/10.1038/s41556-022-00975-z
11. Bhagwat AM, Graumann J, Wiegandt R *et al.* multicrispr: gRNA design for prime editing and parallel targeting of thousands of targets. *Life Sci All* 2020;**3**.e202000757. https://doi.org/10.26508/lsa.202000757
12. Wu S, Kyaw H, Tong Z *et al.* A simple and efficient CRISPR–Cas9 system permits ultra-multiplex genome editing in plants. *Crop J* 2024;**12**:569–82. https://doi.org/10.1016/j.cj.2024.01.010
13. Wang L, Deng A, Zhang Y *et al.* Efficient CRISPR–Cas9 mediated multiplex genome editing in yeasts. *Biotechnol Biofuels* 2018;**11**:277. https://doi.org/10.1186/s13068-018-1271-0
14. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods* 2021;**18**:366–8. https://doi.org/10.1038/s41592-021-01101-x
15. Zhang D, Hurst T, Duan D *et al.* Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proc Natl Acad Sci* 2019;**116**:8693–8. https://doi.org/10.1073/pnas.1820523116
16. Robertson NR, Trivedi V, Lupish B *et al.* Optimized genome-wide CRISPR screening enables rapid engineering of growth-based phenotypes in Yarrowia lipolytica. *Metab Eng* 2024;**86**:55–65. https://doi.org/10.1016/j.ymben.2024.09.005
17. Wei S, Li M, Lang X *et al.* Repurposing plant hormone receptors as chemically-inducible genetic switches for dynamic regulation in yeast. *Metab Eng* 2024;**83**:102–9. https://doi.org/10.1016/j.ymben.2024.03.006
18. Tafrishi A, Trivedi V, Xing Z *et al.* Functional genomic screening in *Komagataella phaffii* enabled by high-activity CRISPR–Cas9 library. *Metab Eng* 2024;**85**:73–83. https://doi.org/10.1016/j.ymben.2024.07.006
19. Ryan OW, Skerker JM, Maurer MJ *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *elife* 2014;**3**:e03703. https://doi.org/10.7554/eLife.03703
20. Schwartz C, Wheeldon I. CRISPR–Cas9-mediated genome editing and transcriptional control in *Yarrowia lipolytica..Synth Biol Methods Protoc* 2018;**1772**:327–45. https://doi.org/10.1007/978-1-4939-7795-6_18
21. Pohl C, Mózsik L, Driessen AJ *et al.* Genome editing in *Penicillium chrysogenum* using Cas9 ribonucleoprotein particles. *Synth Biol Methods Protoc* 2018;**1772**:213–32. https://doi.org/10.1007/978-1-4939-7795-6_12
22. Kreuter J, Stark G, Mach RL *et al.* Fast and efficient CRISPR-mediated genome editing in *Aureobasidium pullulans* using Cas9 ribonucleoproteins. *JBiotechnol* 2022;**350**:11–16. https://doi.org/10.1016/j.jbiotec.2022.03.017
23. Robertson NR, Lenert-Mondou C, Leonard AC *et al.* PYR1 biosensor-driven genome-wide CRISPR screens for improved monoterpene production in *Kluyveromyces marxianus*. *ACS Synth Biol* 2025. https://doi.org/10.1021/acssynbio.4c00797
24. Yang Z, Edwards H, Xu P. CRISPR–Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in *Yarrowia lipolytica*. *Metab Eng Commun* 2020;**10**:e00112. https://doi.org/10.1016/j.mec.2019.e00112
25. Smith JD, Schlecht U, Xu W *et al.* A method for high-throughput production of sequence-verified DNA libraries and strain collections. *Mol Syst Biol* 2017;**13**:913. https://doi.org/10.15252/msb.20167233
26. Regenberg B, Hansen J. GAP1, a novel selection and counter-selection marker for multiple gene disruptions in *Saccharomyces cerevisiae*. *Yeast* 2000;**16**:1111–9. https://doi.org/10.1002/1097-0061(20000915)16:12⟨1111::AID-YEA611⟩3.0.CO;2-3
27. Keeney JB, Reed R. A genetics laboratory module involving selection and identification of lysine synthesis mutants in the yeast *Saccharomyces cerevisiae*. *Microbiol Educ* 2000;**1**:26–30. https://doi.org/10.1128/me.1.1.26-30.2000
28. Cheon SA, Han EJ, Kang HA *et al.* Isolation and characterization of the TRP1 gene from the yeast *Yarrowia lipolytica* and multiple

gene disruption using a TRP blaster. *Yeast* 2003;**20**:677–85. https://doi.org/10.1002/yea.987

29. Widlund PO, Davis TN. A high-efficiency method to replace essential genes with mutant alleles in yeast. *Yeast* 2005;**22**:769–74. https://doi.org/10.1002/yea.1244
30. Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* 2007;**2**:31–34. https://doi.org/10.1038/nprot.2007.13
31. Lyu Y, Wu P, Zhou J *et al.* Protoplast transformation of *Kluyveromyces marxianus*. *Biotechnol J* 2021;**16**:2100122. https://doi.org/10.1002/biot.202100122
32. Liu J, Cui H, Wang R *et al.* A simple and efficient CRISPR–Cas9 system using a ribonucleoprotein method for *Flammulina filiformis*. *J Fungi* 2022;**8**:1000. https://doi.org/10.3390/jof8101000
33. Zou G, Xiao M, Chai S *et al.* Efficient genome editing in filamentous fungi via an improved CRISPR–Cas9 ribonucleoprotein method facilitated by chemical reagents. *Microbial Biotech* 2021;**14**:2343–55. https://doi.org/10.1111/1751-7915.13652
34. Pi HW, Anandharaj M, Kao YY *et al.* Engineering the oleaginous red yeast *Rhodotorula glutinis* for simultaneous β-carotene and cellulase production. *Sci Rep* 2018;**8**:10850. https://doi.org/10.1038/s41598-018-29194-z
35. O'Leary NA, Cox E, Holmes JB *et al.* Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Scientific Data* 2024;**11**:732. https://doi.org/10.1038/s41597-024-03571-y
36. Amos B, Aurrecoechea C, Barba M *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res* 2022;**50**:D898–911. https://doi.org/10.1093/nar/gkab929
37. Martin FJ, Amode MR, Aneja A *et al.* Ensembl 2023. *Nucleic Acids Res* 2023;**51**:D933–41. https://doi.org/10.1093/nar/gkac958
38. Yates AD, Allen J, Amode RM *et al.* Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res* 2022;**50**:D996–1003. https://doi.org/10.1093/nar/gkab1007
39. Grigoriev IV, Nikitin R, Haridas S *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* 2014;**42**:D699–704. https://doi.org/10.1093/nar/gkt1183
40. Arimbasseri AG, Rijal K, Maraia RJ. Transcription termination by the eukaryotic RNA polymerase III. *Biochim Biophys Acta* 2013;**1829**:318–30. https://doi.org/10.1016/j.bbagrm.2012.10.006
41. Ui-Tei K, Maruyama S, Nakano Y. Enhancement of single guide RNA transcription for efficient CRISPR–Cas-based genomic engineering. *Genome* 2017;**60**:537–45. https://doi.org/10.1139/gen-2016-0127
42. Konstantakos V, Nentidis A, Krithara A *et al.* CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res* 2022;**50**:3616–37. https://doi.org/10.1093/nar/gkac192
43. Schirmaier F, Philippsen P. Identification of two genes coding for the translation elongation factor EF-1 alpha of *S. cerevisiae*. *EMBO J* 1984;**3**:3311–5. https://doi.org/10.1002/j.1460-2075.1984.tb02295.x
44. Gautier T, Bergès T, Tollervey D *et al.* Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol Cell Biol* 1997;**17**:7088–98. https://doi.org/10.1128/MCB.17.12.7088
45. Bernstein KA, Granneman S, Lee AV *et al.* Comprehensive mutational analysis of yeast DEXD/H box RNA helicases involved in large ribosomal subunit biogenesis. *Mol Cell Biol* 2006;**26**:1195–208. https://doi.org/10.1128/MCB.26.4.1195-1208.2006
46. Koch B, Mitterer V, Niederhauser J *et al.* Yar1 protects the ribosomal protein Rps3 from aggregation. *J Biol Chem* 2012;**287**:21806–15. https://doi.org/10.1074/jbc.M112.365791
47. Hackmann A, Wu H, Schneider UM *et al.* Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. *Nat Commun* 2014;**5**:3123. https://doi.org/10.1038/ncomms4123
48. Peter M, Neiman A, Park H *et al.* Functional analysis of the interaction between the small GTP binding protein Cdc42 and the Ste20 protein kinase in yeast. *EMBO J* 1996;**15**:7046–59. https://doi.org/10.1002/j.1460-2075.1996.tb01096.x
49. Sullivan DS, Biggins S, Rose MD. The yeast centrin, cdc31p, and the interacting protein kinase, Kic1p, are required for cell integrity. *J Cell Biol* 1998;**143**:751–65. https://doi.org/10.1083/jcb.143.3.751
50. Benton BK, Tinkelenberg A, Gonzalez I *et al.* Cla4p, a *Saccharomyces cerevisiae* Cdc42p-activated kinase involved in cytokinesis, is activated at mitosis. *Mol Cell Biol* 1997;**17**.**9**:5067–76. https://doi.org/10.1128/MCB.17.9.5067
51. Henikoff S. The *Saccharomyces cerevisiae* ADE5, 7 protein is homologous to overlapping Drosophila melanogaster Gart polypeptides. *J Mol Biol* 1986;**190**:519–28. https://doi.org/10.1016/0022-2836(86)90238-X
52. Golding G. Simple sequence is abundant in eukaryotic proteins. *Protein Sci* 1999;**8**:1358–61. https://doi.org/10.1110/ps.8.6.1358
53. Romero P, Obradovic Z, Li X *et al.* Sequence complexity of disordered protein. *Proteins Struct Funct Bioinform* 2001;**42**:38–48. https://doi.org/10.1002/1097-0134(20010101)42:1⟨38::AID-PROT50⟩3.0.CO;2-3
54. Simon M, Hancock JM. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol* 2009;**10**:R59. https://doi.org/10.1186/gb-2009-10-6-r59
55. Peterson ME, Chen F, Saven JG *et al.* Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci* 2009;**18**:1306–15. https://doi.org/10.1002/pro.143
56. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94. https://doi.org/10.1093/protein/12.2.85
57. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci* 1998;**95**:6073–8. https://doi.org/10.1073/pnas.95.11.6073
58. Anderson EM, Haupt A, Schiel JA *et al.* Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of off-target activity. *J Biotechnol* 2015;**211**:56–65. https://doi.org/10.1016/j.jbiotec.2015.06.427
59. Jiang W, Bikard D, Cox D *et al.* CRISPR-assisted editing of bacterial genomes. *Nat Biotechnol* 2013;**31**:233. https://doi.org/10.1038/nbt.2508
60. Semenova E, Jore MM, Datsenko KA *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci* 2011;**108**:10098–103. https://doi.org/10.1073/pnas.1104144108
61. Wiedenheft B, van Duijn E, Bultema JB *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci* 2011;**108**:10092–7. https://doi.org/10.1073/pnas.1102716108
62. Rehman L, Su X, Guo H *et al.* Protoplast transformation as a potential platform for exploring gene function in *Verticillium dahliae*. *BMC Biotechnol* 2016;**16**:57. https://doi.org/10.1186/s12896-016-0287-4
63. Foster AJ, Martin-Urdiroz M, Yan X *et al.* CRISPR–Cas9 ribonucleoprotein-mediated co-editing and counterselection in the rice blast fungus. *Sci Rep* 2018;**8**:14355. https://doi.org/10.1038/s41598-018-32702-w
64. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinform* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421

Received: February 7, 2025. **Revised:** July 24, 2025. **Editorial Decision:** July 28, 2025. **Accepted:** July 29, 2025