מכון ויצמן למדע

WEIZMANN
INSTITUTE
OF SCIENCE

Thesis for the degree
Master of Science

עבודת גמר (תזה) לתואר
מוסמך למדעים

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Donya Khoury

מאת
דוניא ח'ורי

חקירת מנגנוני ההורשה של תכונות כמותיות בשמרים באמצעות הזדווגות מינית

Investigating the Inheritance Mechanisms of Quantitative Traits in *Saccharomyces cerevisiae* Through Sexual Mating

Advisor:
Prof.Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

February 2025

שבט ה'תשפ"ה

1. Abstract

The inheritance of quantitative traits upon sexual mating is influenced by multiple genes and environmental factors and is a cornerstone of genetics and evolution. This study investigates the mechanisms underlying quantitative trait inheritance in *Saccharomyces cerevisiae*, focusing on traits such as gene expression of foreign proteins and cell size. Using over 100 natural yeast isolates genetically modified with unique barcodes, fluorescent markers (GFP and RFP), and resistance genes, we conducted high-throughput mating experiments employing "all-against-all" and "one-against-one" strategies. Fluorescence-Activated Cell Sorting (FACS) was used to analyze inheritance patterns in over 3,000 offspring combinations. The results revealed complex inheritance patterns for fluorescent protein expression upon sexual mating, which were influenced by both parental traits and potentially an additional regulatory mechanism. Notably, this study introduces the concept of "parental and non-parental inheritance," identifying distinct patterns of inheritance that shed light on how specific traits are transmitted across generations. In addition to characterizing inheritance patterns, this study explored applied applications by identifying strains with interesting inheritance and expression patterns. Selected strains were engineered with foreign genes commonly used in industrial processes to evaluate their efficiency. Using pan-transcriptomic data and machine learning models, we explored correlations between gene expression profiles, genetic variation, and observed traits, uncovering key predictors of trait variability. These findings demonstrate the power of *S. cerevisiae* as a model for studying quantitative trait inheritance and offer a robust framework for optimizing yeast strains for industrial applications, including foreign protein production, bioproduction, and bioengineering.

2. Table of contents

1. Abstract....1

2. Table of contents 3

3. List of abbreviations 5

4. Introduction....6

5. Goals....9

6. Material and methods....10

6.1. Yeast strains and growth conditions....10

6.2. FACS measurements of Fluorescent Protein Levels in Parental Strains 11

6.3 Sonication Test to Ensure Single cell measurement 12

6.4. "All-Against-All" Mating, FACS Sorting, Library Preparation, and Barcode Sequencing 12

6.4.1. Mating and Post-Mating Selection.... 12

6.4.2. FACS Sorting.... 13

6.4.3. Library preparation and sequencing.... 14

6.5. "One-Against-One" Mating and FACS Screening....14

6.5.1. Mating the strains.... 14

6.5.2. Verifying Offspring Identity.... 15

6.5.3. FACS Screening.... 15

6.6. Measuring Antibiotic Resistance levels in engineered strains 16

6.6.1. Measuring Antibiotic Resistance in Un-Engineered Strains.... 17

6.7. Comparative Analysis of Protein Production in Yeast Strains 17

6.7.1. Transformation of Foreign Protein Constructs into Yeast Strains.... 17

6.7.2. Western Blot Analysis of Protein Production Efficiency.... 18

6.8. Data Analysis 20

6.8.1. Machine Learning models and Analysis 20

7. Results....22

7.1. FACS Screen for Parents to Determine Fluorescent Protein Levels....22

7.1.1. Sonication test to eliminate clumps 23

7.1.2. Computational Analysis of Transcriptomic Data.... 26

7.2. Offspring Screen to Study Fluorescent Protein Inheritance Mechanisms....27

7.2.1. "All-Against-All" Mating and FACS Sorting.... 27

7.2.2. "One-Against-One" 29

7.3. Comparing Offspring and Parental Fluorescence Expression Levels....30

7.3.1. Investigating Inheritance Patterns Based on Parental Expression Properties.... 31

7.3.2. Patterns of Parental and Non-Parental Inheritance 32

7.4. Comparing the Results of Both Approaches 37

7.5. Noise in Fluorescent Protein Expression Across Strains and Generations: Scaling with Mean Expression and Inheritance Patterns 38

7.5.1. Residual noise calculation.. *39*

7.6. Cell Size in Parents and Offspring and Its Inheritance 41

7.7. Correlation Between Cell Size and Fluorescent Protein Production.................43

7.8. Machine Learning Analysis for Predicting Inherited Features 45

7.8.1. Predicting Offspring's Fluorescent Intensity *45*

7.8.2. Predicting Offspring's Noise Residual... *46*

7.8.3. Identifying Connections Between Genes with High Predictive Importance and Their Target Outcomes *46*

7.9. Measuring Antibiotic Resistance levels 49

7.10. Western Blot Analysis of Protein Production Efficiency 51

8. Discussion...53

8.1. Insights from Parental Strain Screening...53

8.2. High-Throughput Screening of Offspring for Trait Inheritance Patterns 53

8.3. Exploring Inheritance Patterns Between Parents and Offspring.......................54

8.4. Parental and Non-Parental Inheritance Theory 55

8.5. Predicting Expression Levels Using Transcriptomic Data 55

8.6. Heritability of Noise in Protein Expression Upon Sexual Mating....................56

8.7. The Inheritance of Cell Size..57

8.8. Summary 57

8.8. Future Directions...58

9. Acknowledgments 59

10. Literature...60

11. Appendix...63

3. List of abbreviations

S. cerevisiae – Saccharomyces cerevisiae

GFP – Green Fluorescent Proteins

RFP – Red Fluorescent Proteins

FACS – Fluorescence-Activated Cell Sorting.

Hyg – Hygromycin

G418 – kanamycin

Zeo – Zeocin

Nat – Nourseothricin

OD – Optical Density

FS – Forward Scatter

Doxy – Doxycycline

NGS – Next Generation Sequencing

FGF2 - Fibroblast Growth Factor 2

4. Introduction

The inheritance of traits across generations of sexually mating organisms is a cornerstone of biological research, shaping our understanding of genetics, evolution, and the subtle forces that govern life's diversity. While Mendel's laws laid the foundation for inheritance patterns of binary traits, the inheritance patterns become far more complex when it comes to quantitative traits. These traits, influenced by multiple genes and environmental factors, include characteristics like height and metabolic rates (1). For instance, a long-standing question in genetics asks: *If two individuals with different heights reproduce, how tall will their offspring be?*

Such questions underscore the complexity of quantitative traits, which do not adhere to simple Mendelian ratios but instead result from the interplay of multiple loci. In this study, we explore quantitative traits such as the gene expression of foreign proteins, antibiotic resistance, and cell size in *Saccharomyces cerevisiae*. Understanding the mechanisms of inheritance for these quantitative traits is not only critical for advancing scientific knowledge but also holds immense potential for agricultural, medical and industrial applications. Insights into the inheritance of quantitative traits can pave the way for more efficient production of foreign proteins, revolutionizing industries such as pharmaceuticals, bioengineering, and biotechnology, where optimizing protein yield and function is paramount.

To expand our understanding of the inheritance of quantitative traits, the choice of an appropriate model organism is essential. Throughout history, organisms such as *E. coli*, *C. elegans*, *Drosophila*, and *Saccharomyces cerevisiae* have played pivotal roles in uncovering the principles of inheritance and gene function (2). For this study, *Saccharomyces cerevisiae*, commonly known as baker's yeast, was selected as the model organism for investigating the genetic basis of quantitative traits. Its unicellular nature, the availability of hundreds of natural strains that are fully sequenced and have well-annotated genome, rapid life cycle, and ease of genetic manipulation make it an ideal system for addressing complex genetic questions (3,4). Furthermore, *S. cerevisiae* has been widely used to study interactions between genetic and environmental factors (5), which are central to understanding the variability of traits influenced by multiple genes and environmental contexts.

A unique advantage of *S. cerevisiae* as a model organism lies in its unique reproductive biology. Unlike many other organisms, yeast can reproduce both sexually and asexually. Asexual reproduction occurs through mitotic division, allowing the yeast to clone itself while maintaining a haploid or diploid. In its haploid form, it exists as one of two mating types, "*A*" or "*alpha*," which function as the sexes in yeast reproduction. Mating occurs between haploids of opposite mating types through a well-studied process mediated by pheromones and their specific receptors (6). During mating, the genetic material from two parents combines, resulting in a diploid organism that inherits traits from both mating types. The ability to alternate between haploid and diploid states offers unparalleled flexibility for genetic studies, allowing the examination of how specific traits are inherited and how genetic recombination contributes to phenotypic diversity (7).

To study quantitative traits such as gene expression, antibiotic resistance, and cellular variability, we used a diverse collection of *S. cerevisiae* strains and their offspring. Specifically, we utilized over 100 genetically manipulated natural isolates, selected from an extensive collection of 1,011 strains generously provided by Prof. Gianni Liti from IRCAN, a leading expert in yeast population genomics and phylogeny (8). These strains were obtained from diverse geographic regions and niches, both wild and domesticated, representing a rich genomic resource. This diversity captures the natural variability observed in wild yeast populations and provides a robust platform for studying the inheritance of quantitative traits.

To enable the study of quantitative traits such as gene expression and antibiotic resistance, these *S. cerevisiae* strains were engineered using advanced genetic tools (9). In a recent study conducted in our lab by Sivan Kaminsky, over 100 strains were genetically modified to include constructs carrying foreign genes, such as antibiotic resistance markers and fluorescent proteins. Each mating type was engineered with distinct genetic markers: mating type *A* strains were modified to express hygromycin (Hyg) resistance cassettes, a BleoR resistance marker that enabled growth on zeocin (Zeo), and a green fluorescent protein (GFP). Mating type *alpha* strains were engineered to carry nourseothricin (Nat) resistance cassettes, a KanMX resistance marker that enabled growth on geneticin (G418), and a red fluorescent protein (RFP). After mating, the resulting diploid offspring inherited both sets of markers from their respective parents. In addition to analyzing population gene expression levels and antibiotic resistance, we investigated noise in protein expression as a quantitative trait. Noise, defined as the variability in expression levels relative to the mean among genetically identical cells, arises from both intrinsic factors, such as random transcriptional and translational events, and extrinsic factors, including differences in cellular states or environments (10). It can be calculated for both parents and offspring based on fluorescent protein expression in order to follow its inheritance patterns.

Fluorescence-Activated Cell Sorting (FACS) played a crucial role in this study, enabling the precise quantification of GFP and RFP levels in both parents and offspring. This technique also allowed for the sorting of cells based on their fluorescent properties, facilitating the isolation of subpopulations with specific traits (11). Additionally, FACS was used to measure cell size by analyzing the forward scatter area (FS) parameter, a proxy for cellular dimensions (12). Its high-throughput, single-cell resolution proved particularly advantageous for studying quantitative traits and their inheritance, allowing for precise measurement of gene expression and cellular phenotypes within heterogeneous populations of parents and their offspring (13). Complementing FACS, optical density (OD) measurements were employed to quantify antibiotic resistance by measuring growth levels under various antibiotic conditions. Together, these methodologies provided robust and complementary tools for quantifying and analyzing the traits of both parents and offspring and modes of their inheritance.

To deepen our understanding of the genes and mechanisms involved in the inheritance of quantitative traits, we incorporated data from a recently published pan-transcriptome analysis of strains in our yeast collection. This comprehensive study identified 4,977 core genes that are shared among nearly all strains and 1,468 accessory genes that are present in some of the strains. RNA seq provided expression levels of each gene in each strain, providing a detailed map of gene expression variability across the natural yeast isolates (14). From this dataset, we focused on ~120 strains that overlapped with our engineered strains, leveraging their transcriptomic profiles to explore potential links between gene expression patterns, genetic distance, fitness, and fluorescent protein production in both parents and offspring. To analyze these complex relationships, we employed machine learning techniques. By correlating gene expression levels with additional parameters, such as genetic distance between parents and fitness, we aimed to identify factors that play important roles in influencing the inheritance of quantitative traits.

This study is distinctive in its approach, combining the use of over 100 genetically manipulated sexually reproducing strains from an extensive collection of 1,011 isolates offering an unparalleled opportunity to explore the natural variability inherent in wild yeast populations. These isolates, collected from diverse geographic regions and niches, represent a broad genomic resource that has been largely underutilized in studying inheritance of quantitative traits. Moreover, the integration of genetic constructs carrying foreign genes, such as antibiotic resistance markers and fluorescent proteins, provides a novel framework for tracking inheritance patterns and paves the way for the development of methodologies to use in the industry of foreign protein production. By leveraging these unique resources and tools, this study addresses gaps in our understanding of the mechanisms governing quantitative trait inheritance, offering insights with both scientific and industrial relevance.

5. Goals

The goal of this study is to investigate the mechanisms governing the inheritance of quantitative traits in *Saccharomyces cerevisiae* upon sexual mating focusing on traits such as gene expression of foreign proteins, antibiotic resistance, and cell size. Leveraging a diverse collection of genetically engineered natural isolates and their sexual mating offspring combinations, this research aims to uncover the contributions of genetic variation and other factors to the variability of these traits across generations. The findings aim to advance our understanding of quantitative trait inheritance and explore practical applications in optimizing foreign protein production within industrial biotechnology.

To achieve this, the study is structured around the following aims:

Aims:

1. **Establishing a Comprehensive Offspring Library**. Developing an extensive library of over 3,000 diploid offspring combinations, enabling the measurement and quantification of traits using advanced methods such as Fluorescence-Activated Cell Sorting (FACS) and optical density (OD) assays.
2. **Investigating Gene Expression and Cell Size Quantitative Trait Inheritance in Natural Isolates**. Utilizing over 100 genetically engineered natural isolates from a diverse collection of 1,011 strains to study how genetic variation influences the inheritance of foreign protein expression and cell size in diploid offspring.
3. **Identifying Genes That Govern Patterns of Inheritance**. Analyzing the relationships between parental and offspring traits by computationally comparing expression levels. Integrating data from a published pan-transcriptome analysis to identify potential links between natural genes' expression profiles, parental genetic distance, fitness, cell size, expression noise and the inheritance of quantitative traits.
4. **Exploring Applications in Selective Breeding**. Assessing the feasibility of employing selective breeding techniques to enhance yeast strains, leveraging insights into quantitative trait inheritance to optimize foreign protein production for industrial applications such as pharmaceuticals and bioengineering.

6. Material and methods

Please note that all listed primers and PCR protocols can be found in the Appendix in Table 1 and Table 2.

6.1. Yeast strains and growth conditions

The yeast strains used in this study were derived from wild-type *Saccharomyces cerevisiae* strains, part of a collection of natural isolates curated in (8). In a previous experiment conducted in our laboratory, approximately 100 strains from this collection were selected and transformed with a genomic construct inserted into the HO locus. This construct included a unique 20-bp barcode specific to each strain, resistance genes, fluorescent markers, and a system based on Cre-lox recombination. The Cre-lox system facilitates barcode fusion following sexual mating, allowing parental identification of the two parents of each offspring through sequencing of the fused barcode. In addition to the natural isolates, two other strains were engineered to serve as comparative models: the lab strain BY4741/2 (*MATa/alpha his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) and the industrial strain CEN.PK.1C/1D (*MATa/alpha ura3-52 trp1-289 leu2-3,112 his3Δ1 MAL2-8c SUC2*). Both mating type *A* and mating type *alpha* cells from each strain were transformed with the same construct to ensure consistency in the experimental design. After transformation, diploid cells carrying the constructs were selected and induced to sporulate. Haploid spores of both mating types were isolated using haploid-specific antibiotics. In mating type *A* strains, the construct introduced green fluorescent protein (GFP) and hygromycin (Hyg) resistance cassettes, along with a BleoR resistance marker that enabled growth on zeocin (Zeo). In mating type *alpha* strains, the construct introduced red fluorescent protein (RFP) and nourseothricin (Nat) resistance cassettes, along with a KanMX resistance marker that enabled growth on geneticin (G418). The haploid-specific promoters driving BleoR and KanMX facilitated efficient selection of haploids. Following mating and recombination, offspring inherited resistance to both Hyg and Nat, driven by diploid-specific promoters, and exhibited dual fluorescence with both GFP and RFP markers. These modifications facilitated strain-specific identification, offspring tracking, and selection under various growth conditions. This system was originally developed and is comprehensively described in the doctoral thesis of Sivan Kaminski Strauss (15). (Figure 1, Appendix).

Media and antibiotics used in this study:

YPD (Rich Media): Contains 10 g/L yeast extract, 20 g/L peptone, and 20 g/L glucose.

SD Comp (Synthetic Defined Media): Contains 1.7 g/L nitrogen base (without amino acids and ammonium sulfate), 1 g/L monosodium glutamic acid (MSG), 1.5 g/L amino acid mix, and 20 g/L glucose.

SD Comp + DOXY: Same as SD, with the addition of 10 µg/mL doxycycline.

The antibiotics Hyg, Zeo, Nat, and G418 were added as necessary to any media that does not include ammonium sulfate. The concentrations of these antibiotics varied depending on the specific experiment. They were used for the selection of mating types, diploids, and haploids, facilitating the maintenance and tracking of genetically modified yeast strains.

Antibiotic basal concentration and initials:

Hygromycin (Hyg) - 300ug/ml

Zeocin (Zeo) - 150ug/ml

Geneticin (G418) - 200ug/ml

Nourseothricin (Nat) - 100ug/ml

6.2. FACS measurements of Fluorescent Protein Levels in Parental Strains

Starters were prepared from frozen stock and grown for 48 hours in YPD media supplemented with antibiotics to prevent contamination. Two 96-well plates were prepared for mating type *A* (77 strains in total), and one plate was prepared for mating type *alpha* (42 strains in total). Antibiotics were added as follows:

Mating type *A*: 50 µg/mL Zeo, prepared from a stock solution of 100 mg/mL.

Mating type *alpha*: 200 µg/mL G418, prepared from a stock solution of 200 mg/mL.

After reaching the stationary phase (10^8 cell/ml), confirmed by cell counting under a microscope, the cells were diluted 1:10 with SD Comp media supplemented with the corresponding antibiotics. Three technical replicates were prepared for each plate. The cells were then allowed to grow to mid-log phase (~4.5 hours).

To prepare for Fluorescence-Activated cell sorting (FACS) analysis, the cells were diluted at a 1:10 ratio into PBS with EDTA at a final concentration of 0.05 M to reduce cell clumping. The plate was shaken for 2 minutes to ensure proper mixing, and then it was loaded into the FACS plate reader for analysis. (FACS) measurements were performed using an Attune NxT Flow Cytometer equipped with 405-, 488-, 561-, and 638-nm lasers. GFP was detected by excitation at 561 nm and emission collection using a 530/30 BP filter, while RFP (mCherry) was detected by excitation at 561 nm and emission collection using a 620/15 BP filter. The threshold was set on forward scatter (FSC) at 7.0×10^3 to minimize background noise and capture relevant events. Data were acquired for 50,000 cells per well, measuring side scatter (SSC) for granularity, FSC for cell size, and fluorescence intensity for GFP and RFP. These optimized settings ensured high-quality data collection and robust gating, allowing the analyzed population to consist exclusively of single, viable cells (Figure 2, Appendix).

6.3 Sonication Test to Ensure Single cell measurement

To ensure that FACS measurements accurately reflect single-cell fluorescence and not the fluorescence intensity of clumps, and to enable accurate comparisons between the fluorescence intensity of wild strains and the non-clumpy engineered lab strains and CEN.PK, a sonication test was performed on selected strains. Eight mating type *A* strains and 11 mating type *alpha* strains were chosen from the "natural strain collection", along with both mating types of the lab strains and CEN.PK. Starters were prepared from frozen stocks and grown for 48 hours in YPD media supplemented with antibiotics to prevent contamination. Mating type *A* strains were cultured in media containing Hyg at a final concentration of 0.3 mg/mL, while mating type *alpha* strains were cultured in media containing Nat at a final concentration of 0.1 mg/mL.

The same protocol outlined in the preceding chapter was followed, with one key modification: the cells were subjected to sonication immediately prior to FACS measurement. Sonication was performed using a sonication device, "Bioruptor Plus," set to low intensity. Each sample underwent three rounds of sonication for 30 seconds, with 30-second intervals between rounds to prevent overheating and maintain cell viability. In parallel, the fluorescence intensity of the same strains from the same batch was measured without sonication for comparison. For each strain and treatment, three technical repeats were performed to ensure consistency and reproducibility.

6.4. "All-Against-All" Mating, FACS Sorting, Library Preparation, and Barcode Sequencing

6.4.1. Mating and Post-Mating Selection

Starters were prepared from frozen stocks and grown for 20 hours in SD Comp media supplemented with antibiotics to prevent contamination by bacteria and other nonresistant microorganisms. Two 96-well plates were prepared for mating type *A* (96 strains in total), and one 96-well plate was prepared for mating type *alpha* (48 strains in total). For mating type *A*, 50 µg/mL Zeo was added, prepared from a stock solution of 100 mg/mL. For mating type *alpha*, 200 µg/mL G418 was added, prepared from a stock solution of 200 mg/mL. After reaching stationary phase, the cells were diluted 1:50 into SD Comp media containing Zeo and G418 at the same concentrations. After 12 hours, the cells were diluted 1:5 into fresh SD Comp media with antibiotics and allowed to grow for 4.5 hours until they reached mid-log phase, confirmed by counting two representative strains.

To ensure equal representation of cells from each mating type, OD values were measured to determine the appropriate volumes for mixing each strain. Most strains had an OD value around 0.15, and a cutoff of 0.8 was used for volume adjustments. For mating type *A*, 10 µL was taken from strains with OD values above 0.15, and 14 µL from strains with OD values below 0.15, resulting in a mixture where each mating type *A* strain was represented by approximately $5*10^5$ cells. For mating type *alpha*, 18 µL was taken from strains with OD values above 0.15, and 29 µL from strains with OD values below 0.15, generating a mixture where each mating type *alpha* strain was represented by approximately $5*10^5$ cells. To facilitate "all-against-all" mating, 50 µL of each mixture was transferred into five wells of a 24-well plate. Each well contained 900 µL of SD Comp media supplemented with 10 µg/mL doxycycline (Doxy) to activate the barcode fusion system. The plates were incubated without shaking at 25°C for 20 hours.

Following incubation, to eliminate any remaining haploid (parents) cells that did not undergo mating, the cells were diluted 1:200 into SD Comp media containing 10ug/ml Doxy and basal concentrations of both Hyg and Nat. This selective pressure ensured that only offspring that inherited resistance to both antibiotics survived, while parental strains - resistant to either Hyg (mating type *A*) or Nat (mating type *alpha*) - were eliminated. The antibiotic concentrations used were 300 µg/mL Hyg, prepared from a stock solution of 500 mg/mL, and 200 µg/mL Nat, prepared from a stock solution of 100 mg/mL. Cultures were incubated for 24 hours at 30°C with shaking.

6.4.2. FACS Sorting

Prior to sorting, cells were prepared to ensure optimal conditions. Cultures were confirmed to be in the stationary phase, and to eliminate clumps and ensure uniformity, samples underwent sonication at a voltage level of 2 for 30 seconds. This process was repeated twice, with cells placed on ice between sonication cycles to prevent overheating and to maintain viability. Sorting of offspring displaying both GFP and RFP markers was performed using a FACS Aria Fusion instrument (BD Biosciences) equipped with 405, 488, 561, and 640 nm lasers and a 100 µm nozzle, controlled by BD FACS Diva software v8.0.1 (BD Biosciences) at The Weizmann Institute of Science Flow Cytometry Core Facility. Double-positive cells, displaying fluorescence for both GFP and RFP, accounted for approximately 85% of the total population, with GFP-RFP fluorescence levels differing by approximately 1-fold.

Cells were sorted into three distinct populations: (Figure 3, Appendix):

1. **Sorted High**: The top 10% of cells based on GFP and RFP fluorescence intensity, a total of ~ 2 million cells.
2. **Sorted Low**: The bottom 10% of cells based on GFP and RFP fluorescence intensity, a total of ~ 2 million cells.
3. **Total**: A middle population including cells from both high and low fractions, a total of ~ 2 million cells.

Each sorting was performed in three biological replicates. The sorted cells were frozen for downstream processing.

6.4.3. Library preparation and sequencing

DNA was extracted by boiling the cells for 20 minutes in 50 µL solution of NaOH (20 mM). The samples were centrifuged for 1.5 minutes, and the supernatant was transferred to a new tube. The barcode region was amplified using PCR (See primers 4-5 and PCR 1), with four technical replicates prepared for each sample to reduce PCR biases.

The PCR products were cleaned using SPRI beads. For cleaning, 66 µL of SPRI beads was added to 44 µL of the sample, mixed thoroughly by slow pipetting, and allowed to sit at room temperature for 7 minutes. The samples were then placed on a magnetic stand for 5 minutes to separate the beads, and the clear supernatant was carefully removed. The beads were washed twice with 150 µL of 70% ethanol, and after the final wash, the plate was spun down to remove residual ethanol. The beads were allowed to dry for 2–5 minutes, and the DNA was eluted in 25 µL of DDW. Elution was performed sequentially for each technical replicate. After amplifying the barcode region, the Illumina (Next-generation sequencing) indexes were added using customized I5 and I7 index primers (See primers 6-7 and PCR 2). The resulting PCR product was cleaned again using the same SPRI bead method described above. The prepared library was sequenced using Illumina. After initial de-multiplexing by the Illumina platform, generating paired end reads with a read length of 25 bp. All reads were subsequently processed by Cutadapt to retain only the barcode region in the output files (16). Alignment to the reference barcode database was performed using Bowtie2 (17). To recover read counts per fused barcode, we utilized an in-house script (15). The average coverage per sample was approximately $3.3*10^6$ read counts.

6.5. “One-Against-One” Mating and FACS Screening

6.5.1. Mating the strains

In this experiment, we mated each pair of strains in a separate well, creating all possible diploid offspring combinations. Starters were prepared from frozen stocks and grown for 72 hours in YPD media supplemented with antibiotics to prevent contamination. Two 96-well plates were prepared for mating type *A* (96 strains in total), and one 96-well plate was prepared for mating type *alpha* (48 strains in total). For mating type *A*, 300 µg/mL Hyg was added, prepared from a stock solution of 500 mg/mL. For mating type *alpha*, 200 µg/mL Nat was added, prepared from a stock solution of 100 mg/mL.

After reaching stationary phase, the cells were diluted 1:1000 into SD Comp media containing Nat and Hyg at the same concentrations listed above and incubated for 12 hours until they reached mid-log phase. To ensure equal representation of cells from each mating type, OD values were measured, and a volume equivalent to $5×10^5$ cells was taken from each mating type. A total of 35 96-well plates were prepared, with each well containing 130 µL of SD Comp media supplemented with doxycycline (Doxy) to activate barcode fusion. The final Doxy concentration was 10 µg/mL, prepared by dissolving 2 mg of Doxy in 200 mL of SD Comp media. To each well, 5×10^5 cells from one strain of mating type *A* and 5×10^5 cells from one strain of mating type *alpha* were added. Each well had a final volume of 140 µL.

The plates were gently shaken for 5 minutes to ensure proper mixing, then incubated at 25°C without shaking for 20 hours to allow mating and barcode fusion to occur. This setup ensured accurate mating for all strain combinations while maintaining separation for downstream assays.

To eliminate any remaining haploid (parental) cells that did not undergo mating, the cells were diluted 1:50 into SD Comp media supplemented with Doxy and basal concentrations of both Hyg and Nat. This selective pressure ensured that only the offspring, which inherited resistance to both antibiotics, survived, while parental strains - resistant to either Hyg (mating type *A*) or Nat (mating type *alpha*) - were eliminated. The antibiotic concentrations used for this step were 300 µg/mL Hyg, prepared from a stock solution of 500 mg/mL, and 200 µg/mL Nat, prepared from a stock solution of 100 mg/mL. The selection process for diploids under double antibiotic conditions was repeated twice, with a 24-hour interval between each step, to ensure complete elimination of haploid parental cells and to strengthen the selection of diploid offspring**.** After completing the selection process, all plates containing the 3,024 offspring were stored in a 30% glycerol stock at -80°C for future use.

6.5.2. Verifying Offspring Identity

To ensure the integrity of the offspring and confirm that no contamination or errors occurred during the mating process, a verification test was performed on 56 randomly selected offspring. Sanger sequencing was used to analyze the fused barcodes, enabling identification of the parental strains. A sample was taken from the plate prior to FACS, and a starter culture was prepared in YPD. DNA was extracted by boiling the cells for 20 minutes in a 50 µL solution of 20 mM NaOH. The samples were then centrifuged for 1.5 minutes, and the supernatant was transferred to a new tube for downstream analysis. The verification process involved two separate PCR reactions: To identify the *A* parent (See primers 1-2 and PCR 2) and (See primers 3-4 and PCR 2) to identify the *alpha* parent.

6.5.3. FACS Screening

A FACS machine was used to screen all offspring for GFP and RFP fluorescence levels. Starters were prepared from frozen stock and grown for 48 hours in YPD media supplemented with double antibiotics to prevent contamination and eliminate any remaining haploids. The antibiotic concentrations used were 300 µg/mL Hyg, prepared from a stock solution of 500 mg/mL, and 200 µg/mL Nat, prepared from a stock solution of 100 mg/mL.

After reaching stationary phase (~10^8 cells/mL), confirmed by cell counting under a microscope, the cells were diluted 1:10 into SD Comp media supplemented with the corresponding antibiotics and allowed to grow for ~4.5 hours to reach mid-log phase.

For FACS analysis, the cells were further diluted 1:10 with a 1:10 EDTA solution prepared in PBS to minimize cell clumping. The plate was gently shaken for 2 minutes to ensure proper mixing being loaded into the FACS plate reader for analysis of GFP and RFP fluorescence levels, using the same parameters and gating system described in the section "FACS Measurements of Fluorescent Protein Levels in Parental Strains".

6.6. Measuring Antibiotic Resistance levels in engineered strains

Starters were prepared from frozen stock and grown for 24 hours in YPD media in three 96-well plates: two plates for mating type *A* (77 strains in total) and one plate for mating type *alpha* (42 strains in total), with each well containing a specific strain. Basal concentrations of antibiotics were added to prevent contamination. For mating type *A*, 50 µg/mL zeocin (Zeo) was added, prepared from a stock solution of 100 mg/mL. For mating type *alpha*, 200 µg/mL Geneticin (G418) was added, prepared from a stock solution of 200 mg/mL.

The cells were incubated at 30°C with shaking. After reaching the stationary phase (10^8 cell/ml), the cells were diluted 1:100 into SD comp media and grown under different antibiotic conditions for 24 hours to reach stationary phase. Two concentrations of each antibiotic were tested for each mating type, along with a no-antibiotic control. Each plate setup was conducted in three biological replicates to ensure the reliability and reproducibility of the results. The specific experimental setups are outlined below:

Mating Type *A*:

Plate 1: SD comp (no antibiotic)

Plate 2: SD comp + Hyg 0.3 mg/mL

Plate 3: SD comp + Hyg 3 mg/mL

Plate 4: SD comp + Zeo 0.05 mg/mL

Plate 5: SD comp + Zeo 0.35 mg/mL

Mating Type *alpha*:

Plate 1: SD comp (no antibiotic)

Plate 2: SD comp + NAT 0.1 mg/mL

Plate 3: SD comp + NAT 0.8 mg/mL

Plate 4: SD comp + G418 0.2 mg/mL

Plate 5: SD comp + G418 1.5 mg/mL

To confirm that the cells have reached the stationary phase (10^8 cell/ml), we measured OD levels for representative plates, making sure all strains were around 1. Then the plates were diluted 1:10 into the same media and antibiotic conditions and allowed to grow for approximately 4 hours to reach mid-log phase (10^7 cell/ml). Subsequently, OD measurements were performed on all plates (600 wavelength using Tecan Spark plate reader) to determine growth levels for each strain under the specified conditions.

6.6.1. Measuring Antibiotic Resistance in Un-Engineered Strains

To investigate the potential for innate resistance mechanisms, eight un-engineered (i.e. without inserted antibiotics resistance gene) natural isolates in their diploid form (lacking resistance gene constructs) were tested for growth under varying concentrations of each of the antibiotics. Four strains (ANE, BQG, AIK, and BBA) were grown in media containing Hyg and Zeo, while four other strains (AIM, BKR, BRB, and BTH) were grown in media containing Nat and G418. Strains were first grown for 24 hours in YPD media to reach the stationary phase. Once stationary phase was reached (10^8 cell/ml), the cultures were diluted 1:100 into SD Comp media supplemented with specific concentrations of antibiotics. Since these strains lacked antibiotic resistance genes, lower antibiotic concentrations were used in this experiment to assess their innate resistance. The antibiotic concentrations tested are summarized as follows:

1. **For strains grown in Hyg and Zeo**:

- Hyg concentrations: 0.0015 mg/mL, 0.03 mg/mL, 0.09 mg/mL, 0.2 mg/mL, and 0.3 mg/mL.
- Zeo concentrations: 0.0075 mg/mL, 0.015 mg/mL, 0.03 mg/mL, 0.07 mg/mL, and 0.15 mg/mL.

2. **For strains grown in Nat and G418**:

- Nat concentrations: 0.005 mg/mL, 0.01 mg/mL, 0.03 mg/mL, 0.06 mg/mL, and 0.1 mg/mL.
- G418 concentrations: 0.01 mg/mL, 0.03 mg/mL, 0.07 mg/mL, 0.13 mg/mL, and 0.2 mg/mL.

OD values of all strains grown in the presence of antibiotics were compared to OD values of a control plate containing SD Comp media without antibiotics. Each condition was done in 3 technical replicates. To monitor growth, a robotic system was employed to maintain optimal growth conditions. The robot kept the plates shaking at 30°C and measured the OD levels of the cells every 1.5 hours for 10-time points, allowing for detailed tracking of growth kinetics under the various antibiotic conditions. This setup ensured precise and consistent monitoring of growth patterns, providing reliable data for growth rate analysis.

6.7. Comparative Analysis of Protein Production in Yeast Strains

6.7.1. Transformation of Foreign Protein Constructs into Yeast Strains

To explore the industrial application of this project, four yeast strains were utilized: two wild strains from the “natural strain collection”, BMB (*MATa ho::construct Hyg, Zeo, GFP, Δura3::Z3TF-G418*) and AKP (*MATa ho::construct Hyg, Zeo, GFP, Δura3::Z3TF-G418*); the lab strain BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0* Δ*ura3::Z3TF-G418*); and the industrial strain CEN.PK (*MATa trp1-289 leu2-3,112 his3Δ1 MAL2-8c SUC2* Δ*ura3::Z3TF-G418*). The term "construct" refers to the engineered genetic sequence integrated into the *ho* locus of the BMB and AKP strains, as depicted in (Figure 1, Appendix). All four strains were subsequently transformed with an additional genetic cassette carrying an estrogen-inducible transcription factor (*Z3TF*) (18) kindly provided by Maya Schuldiner's lab. Next, these strains were transformed with a construct designed by Noa Hefetz-Aharon that includes an α-factor signal peptide (MFα1-sp) and a FLAG tag and encodes either Albumin or FGF2 under the control of an estrogen-inducible promoter. Both proteins are of significant industrial relevance.

Starter cultures were prepared by plating frozen stocks on YPD agar and incubating for 48 hours. Single colonies were then inoculated into approximately 1 mL of YPD medium supplemented with strain-specific antibiotics, as follows:

- BY4741 – YPD + 0.1 mg/ml G418
- CEN.PK.1C – YPD + 0.1 mg/ml G418
- BMB – YPD + 0.1 mg/ml G418 + 0.3 mg/ml HYG
- AKP – YPD + 0.1 mg/ml NAT + 0.3 mg/ml HYG

To transform the plasmid containing the construct into the cells, we started with 10^8 cells, from culture of 1-2*10^7 cells/ml. Transformation solutions were freshly prepared to ensure optimal conditions. The TE+LiAc solution consisted of 0.1 M Lithium Acetate and 1X TE buffer, while the PEG+LiAc solution contained 40% (w/v) Polyethylene Glycol (PEG), 0.1 M Lithium Acetate, and 1X TE buffer. Salmon sperm DNA (7.5 mg/mL) was boiled for 5 minutes at 100C◦ and cooled on ice for 3 minutes prior to use. Cells were counted to ensure a concentration of 1*10^7 and processed through two rounds of centrifugation (3000g, 3 minutes) and resuspension in TE+LiAc. For the first and second round, cells were resuspended in 3 mL and 1 mL of TE+LiAc, respectively. For each transformation reaction, 50 µL of cells containing 10^8 cells were mixed with 1 mg DNA (or 45 µL PCR product), 5 µL boiled salmon sperm DNA, and 350 µL PEG+LiAc. The mixture was vortexed to ensure homogeneity and incubated at 30◦C for 30 minutes. Followed by heat shock at 42 ◦C for 40 minutes. Following transformation, cells were centrifuged at 11,000 rpm for 1 minute, and the toxic PEG-containing supernatant was aspirated. The pellet was resuspended in 150-200 µL DDW and plated on SD-URA selective media for growth. To ensure transformation success we performed PCR on Amp gene present in the construct (See primers 8-9 and PCR 3).

6.7.2. Western Blot Analysis of Protein Production Efficiency

To assess protein production efficiency, we performed Western blot analysis on four engineered strains: BY4741, 1C, and two strains from our collection, AKP and BMB. These strains were engineered with a construct containing the FGF2 and Albumin genes. The strains were cultured in SD-URA media, and protein production was induced when the cultures reached a density of 3×10^6 cells/mL. Induction was initiated by adding 100 nM estrogen to the media, followed by incubation for 12 hours at 30°C with shaking. After incubation, a volume of cells corresponding to 10 OD was transferred to a new tube and centrifuged at 3000 × g for 5 minutes to pellet the cells. The medium was removed, and the pellet was washed with 1 mL of TEx1 buffer. The cells were centrifuged again, the medium was discarded, and the pellet was resuspended in 300 µL of lysis buffer containing 8 M urea, 40 mM Tris-HCl (pH 6.8) and o.1 nM DTT. Glass beads were added to the lysate until almost all the liquid was surrounded by beads, and the samples were vigorously shaken at 4°C for 10 minutes. The supernatant was separated from the glass beads using centrifugation at 750 × g for 3 minutes in a cold centrifuge. The supernatant was vortexed briefly, and 150 µL was transferred to a new tube to ensure the inclusion of all protein content. To this , 30 µL of a 6X SDS+DTT solution containing 1X β-mercaptoethanol was added to the sample, resulting in a final volume of 180 µL and a 1:6 dilution of the original supernatant. The SDS+DTT solution was prepared by warming SDS X6 and adding 60 mM DTT (final concentration 10 mM). The samples were vortexed briefly, heated at 95°C for 5 minutes, and stored at 4°C until loading. The protein samples were loaded into a 10% gel prepared with two phases: a lower resolving phase to separate proteins based on molecular weight and an upper stacking phase to align proteins before entry into the resolving phase. The lower phase (10% gel) was prepared with 4 mL H_2O, 3.3 mL 30% acrylamide mix, 2.5 mL 1.5 M Tris (pH 8.8), 0.1 mL 10% SDS, 0.1 mL 10% ammonium persulfate, and 4 µL TEMED, for a total volume of 10 mL. This solution was pipetted into the gel plates, and 400 µL of isopropanol was layered on top to prevent air exposure and ensure even polymerization. After 20 minutes, the isopropanol was removed. The stacking phase (3 mL gel) was prepared with 2.1 mL H_2O, 0.5 mL 30% acrylamide mix, 0.38 mL 1.0 M Tris (pH 6.8), 0.03 mL 10% SDS, 0.03 mL 10% ammonium persulfate, and 3 µL TEMED. This solution was layered on top of the resolving phase, and a comb was inserted to form wells, ensuring no air bubbles were trapped. After polymerization, the prepared protein samples were loaded into the gel, and electrophoresis was performed to separate the proteins. The gel chamber was filled with 10% TG-SDS running buffer, and the gel was placed in a plastic box to ensure a tight seal. A 3 µL protein marker and 35 µL of each sample were loaded into the wells. The gel was run at 80 V for 20 minutes until the samples reached the boundary between the stacking and resolving phases. The voltage was then increased to 200 V, and the gel was run for an additional 30 minutes. Following electrophoresis, the gel was removed from the chamber and placed in DDW. Proteins were transferred onto nitrocellulose membrane using a semi-dry (BioRad) protocol. The membrane was then blocked for 1hr while shaking at room temperature in Odyssey Blocking Buffer diluted 1:2.5 in PBS. then incubated with an Anti-FLAG antibody diluted 1:1000 in Blocking Buffer (diluted 1:5 in PBS) for 1 hour at room temperature with shaking. Following incubation, the membrane was washed three times with 1X TBST, shaking for 5 minutes at room temperature, and the wash solution was discarded after each wash. Next, the membrane was incubated with the secondary antibody (Sigma Goat anti-Mouse antibody) diluted 1:10,000 in Blocking Buffer (1:5 in PBS) for 1 hour at room temperature with shaking. After incubation, the same washing protocol was applied three times with 1X TBST to remove unbound secondary antibodies. Finally, the membrane was imaged to detect the target protein using the LI-COR Odyssey Imaging System.

6.8. Data Analysis

All data analyses in this study were conducted using Python programming language. The analyses encompassed data preprocessing, visualization, and statistical testing to ensure the reliability and reproducibility of results. Raw data from sequencing, FACS, and growth assays were preprocessed and organized using libraries such as pandas for data manipulation and numpy for numerical operations. Statistical analyses, including correlation tests, regression models, and enrichment index calculations, were performed using scipy and statsmodels. Principal Component Analysis (PCA) was conducted using scikit-learn to explore variations in transcriptomic data and their relationship to fluorescent protein expression.

At times, ChatGPT Large Language Model (LLM) was used to draft Python scripts. Typical prompts included a general description of the computational task, followed by iterations of running suggested scripts, modifying them as needed, and debugging.

6.8.1. Machine Learning models and Analysis

To investigate the relationship between fluorescent gene expression and transcriptomic profiles, we used data from a study analyzing the pan-transcriptome of approximately 1,000 natural yeast isolates, encompassing 4,977 core genes and 1,468 accessory genes (14). Focusing on about 120 strains, we aimed to uncover potential links between natural genes' expression and fluorescent protein production, and to develop models predicting both GFP and RFP expression levels in offspring and their noise residual. We considered several predictive features for this model: for each offspring we considered expression level of all natural genes in each of its two parents (about 2*6,000 features), along with parental GFP and RFP expression levels, genetic distance between parental strains (8), fitness of both parents and offspring (15), and forward scatter measurements (used as a proxy for cell size), which were obtained via fluorescence-activated cell sorting (FACS) alongside fluorescent protein intensity. To predict our target features, we applied multiple machine learning models; LightGBM (LGB) (19), XGBoost (20), GradientBoostingRegressor (21) and DoubleLearning CatBoostRegressor (22). We chose to use the DoubleLearning CatBoostRegressor model based on its high R-squared score, indicating that it provided strong predictive accuracy for our target features. To evaluate the model's performance, we used a customized cross-validation approach. In this approach, we repeatedly partitioned the data into training and test sets by randomly selecting a subset of parents along with all their offspring for the training set. After training, we identified the 20 most predictive features based on their feature importance using SHAP (SHapley Additive exPlanations). SHAP is a method that helps explain the output of machine learning models. It works by fairly attributing the contribution of each feature (or group of features) to the model’s predictions. SHAP values provide insights into how individual features impact the model’s prediction for each instance, allowing us to understand the relative importance of different features, such as gene expression levels, genetic distance, and fitness.

7. Results

To investigate the inheritance of quantitative traits, we focused on five distinct phenotypic categories: antibiotic resistance, gene expression of fluorescent genes, noise in gene expression and cell size. These traits were chosen due to their measurable and potentially heritable properties, making them ideal for studying the mechanisms underlying quantitative trait inheritance in Saccharomyces cerevisiae.

7.1. FACS Screen for Parents to Determine Fluorescent Protein Levels

In this study, the expression of fluorescent genes was used as a quantitative trait. Using a FACS machine, we screened all engineered strains containing the foreign fluorescent genes GFP (in mating type *A*) and RFP (in mating type *alpha*). A wide variation in fluorescent protein expression levels was observed across the different strains (Figure 1), supporting our hypothesis that fluorescent gene expression behaves as a quantitative trait that may span a range between natural isolates of the species. To evaluate the statistical significance of differences in gene expression between the highest- and lowest-expressing strains, we performed a rank-sum test. This analysis was conducted using data from a single replicate, focusing on all cells measured for each strain. The results demonstrated a significant difference in gene expression (p-value = 1.544e-91) between BMB (the highest GFP-expressing strain) and BLS (p-value < 5e-324) (the lowest GFP-expressing strain). Similarly, in mating type *alpha*, BNI was identified as the highest-expressing strain, and AKB as the lowest-expressing strain, with a statistically significant difference observed (p-value < 0.001) (Figure 2). To further explore the distribution of fluorescence intensities, we analyzed the GFP and RFP expression levels for each strain, including the lab strain (BY4741/2) and industrial strain (CEN.PK.1C/1D). Both linear and log-transformed distributions revealed that the fluorescence intensities did not conform to a normal distribution (Figure 5, Appendix), Next, we tested the correlation between fluorescent gene expression in mating types *A* (GFP) and *alpha* (RFP) within the same strain. We note though that only a limited number of strains had both mating type a and mating type alpha partners, thus correlation could only be measured on a small subset of 32 strains. No significant correlation was observed (R = 0.02, p-value = 0.2674, Figure 3A).

However, when strains were color-coded according to their zygosity index as defined in (8), distinct patterns emerged. For homozygous strains, a positive but non-significant correlation was observed (R = 0.34, p-value = 0.1558, Figure 3B). In contrast, heterozygous strains exhibited no correlation (R = -0.02, p-value = 0.9578, Figure 3C), highlighting the variability in gene expression across different zygosity levels. To ensure accurate comparisons between

GFP and RFP levels, which are not on the same scale due to differences in fluorescence intensity captured using FACS, fluorescence values were normalized to the median of each experiment. This normalization was consistently applied across all experiments, facilitating both easier visualization and more accurate analyses. The graphs presented in Figure 5 reflect these normalized values.

7.1.1. Sonication test to eliminate clumps

Many of the strains exhibited significant clumping, as observed under the microscope (Figure 6, Appendix), with most strains forming clumps of varying sizes. To ensure that the FACS machine accurately measured fluorescent intensity of viable single cells, we conducted a test using 18 strains from the collection, representing both extremes of the fluorescence intensity spectrum, with two different treatments: sonicated and non-sonicated.

The results showed that sonication had a minimal impact on the fluorescence intensity ranking of strains. In mating type *A*, the highest-expressing strains consistently remained the highest, and the lowest-expressing strains remained the lowest both with and without sonication. For mating type *alpha*, the highest and lowest strains generally maintained their positions; however, there was some shuffling among the intermediate strains within the highest- and lowest-expressing groups (Figure 4). Additionally, we compared the fluorescence intensity of the sonicated strains to the engineered lab strain (BY4741/2) and industrial strain (CEN.PK.1C/1D) containing the GFP/RFP constructs. These strains, which are not clumpy, were consistently positioned at the bottom of the fluorescence intensity spectrum in both mating type *A* and mating type *alpha* under the sonicated treatment, as illustrated in Figure 4 (B) and (D). This result could be taken to indicate that clumping of some of the natural strains results in inflated expression estimation.

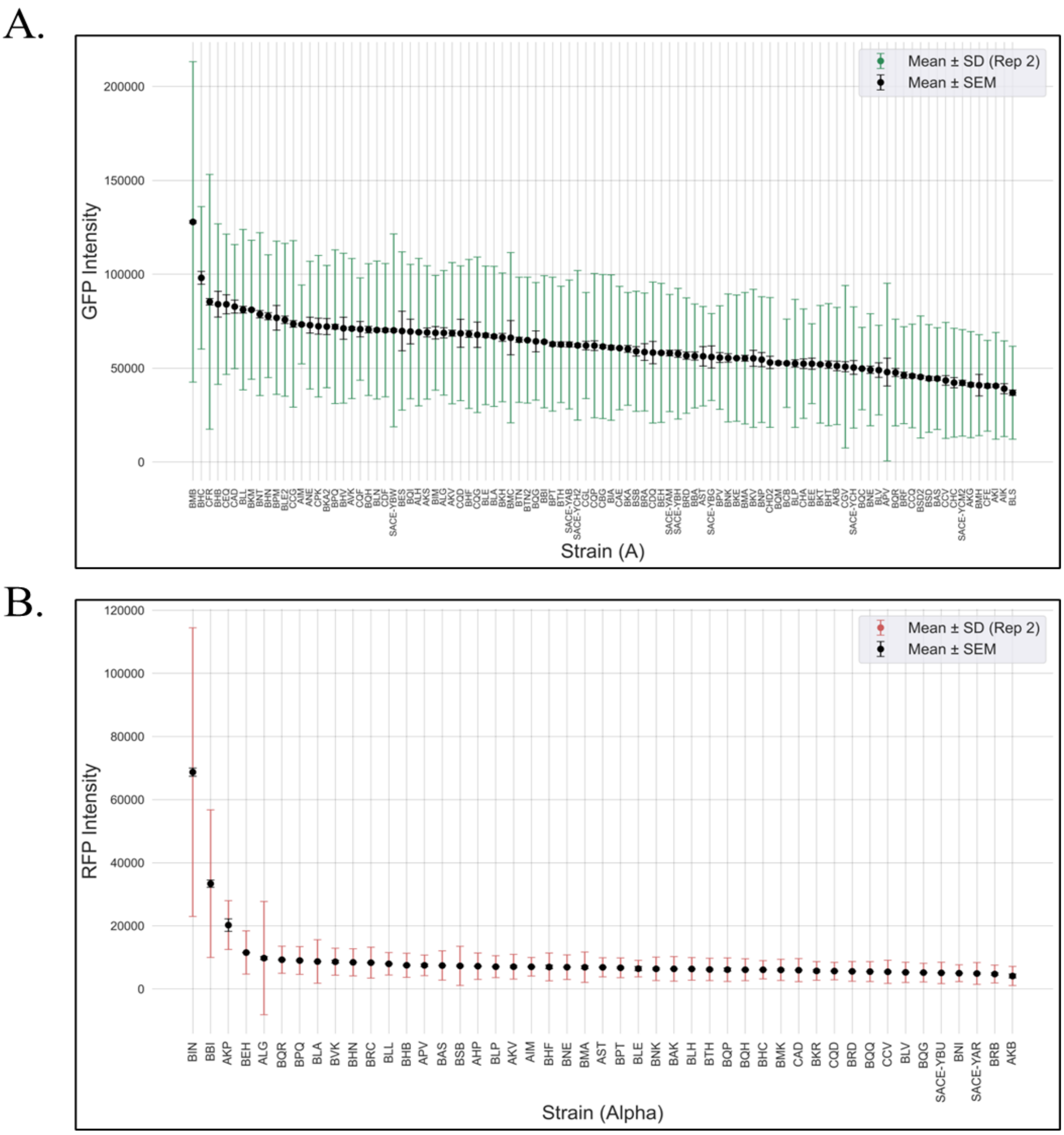


Figure 1 | GFP and RFP levels in parent strains. This figure presents GFP and RFP fluorescence intensity levels across 96 mating type *A* strains (A) and 48 mating type *alpha* strains (B), measured in three technical replicates. For each strain, black markers indicate the mean fluorescence intensity, with black error bars representing the standard error of the mean (SEM) across replicates. The green lines (A) and red lines (B) represent the standard deviation (SD) of the first replicate, calculated from a population of cells. The FACS machine measured fluorescence intensity for 50,000 cells per strain, with a rigorous three-step gating system applied to exclude clumped cells and debris, ensuring the analyzed population consisted of single, viable cells ending up with an average of 25000 cells per strain. SD values across replicates showed strong and statistically significant correlations for both mating types. For mating type *A* strains: (rep1 vs. rep2) r = 0.793, p < 0.001; (rep1 vs. rep3) r = 0.657, p < 0.001; and (rep2 vs. rep3) r = 0. 916, p < 0.001. For mating type *alpha* strains: (rep1 vs. rep2) r = 0.996, p < 0.001; (rep1 vs. rep3) r = 0.995, p < 0.001; and (rep2 vs. rep3) r = 0.992, p < 0.001.

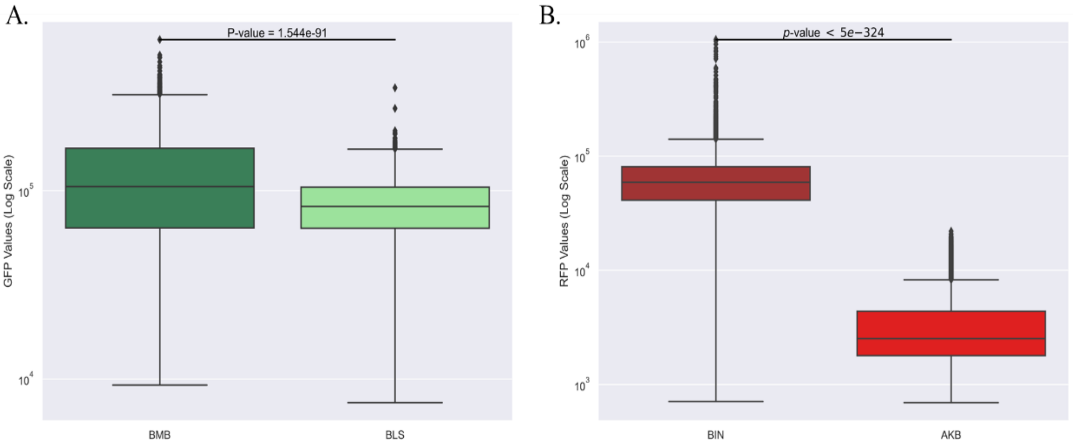


Figure 2 | Box plot of GFP and RFP intensities (log scale) for strains representing the extremes of the fluorescence spectrum. (A) GFP intensity of cells from mating type A, comparing BMB (the highest-expressing strain) to BLS (the lowest-expressing strain). A Ranksum test confirmed a significant difference in expression between the strains (p-value < 0.001). (B) RFP intensity of cells from mating type α, comparing BIN (the highest-expressing strain) to AKB (the lowest-expressing strain). A Ranksum test also showed a significant difference in expression between these strains (p-value < 0.001).

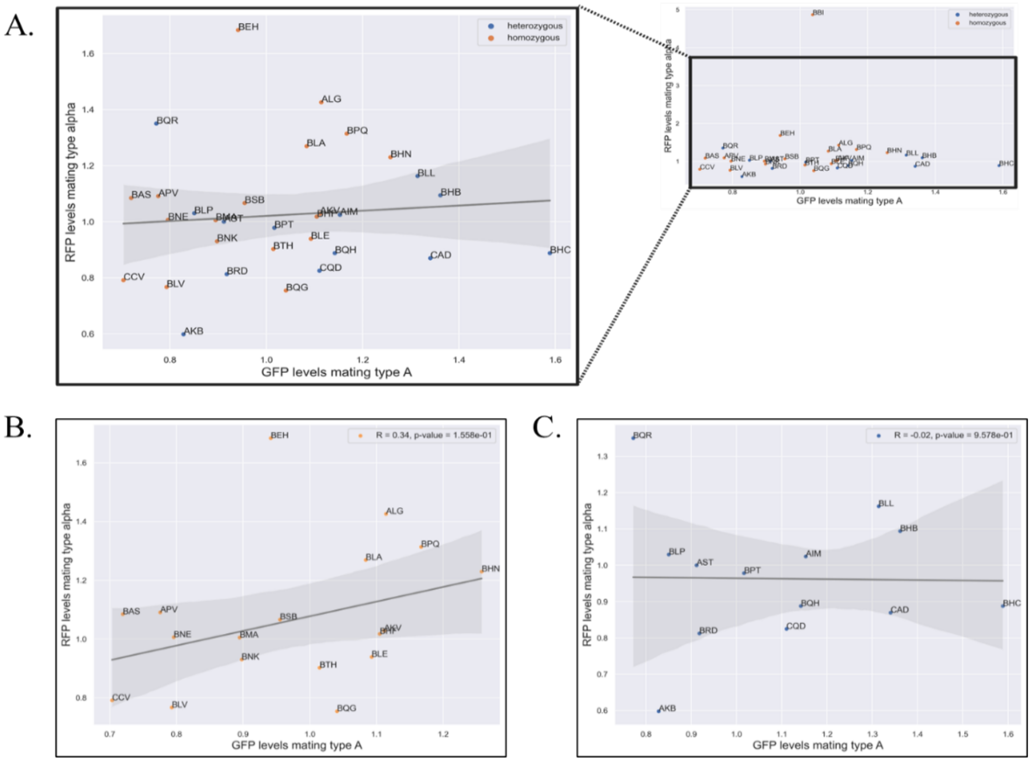


Figure 3 | Correlation analysis of GFP and RFP fluorescence intensities across mating types of the same strain and zygosity levels, normalized to the median. (A) Scatter plot showing the correlation between GFP (mating type *A*) and RFP (mating type *alpha*) fluorescence intensities across 32 strains. Each point represents a single strain and is color-coded by zygosity (orange for homozygous strains and blue for heterozygous strains). No significant correlation was observed (R = 0.02, p-value = 0.2674). (B) Scatter plot of GFP and RFP fluorescence intensities for homozygous strains. A positive but non-significant correlation was observed (R = 0.34, p-value = 0.1558). (C) Scatter plot of GFP and RFP fluorescence intensities for heterozygous strains. No correlation was observed (R = -0.02, p-value = 0.9578)

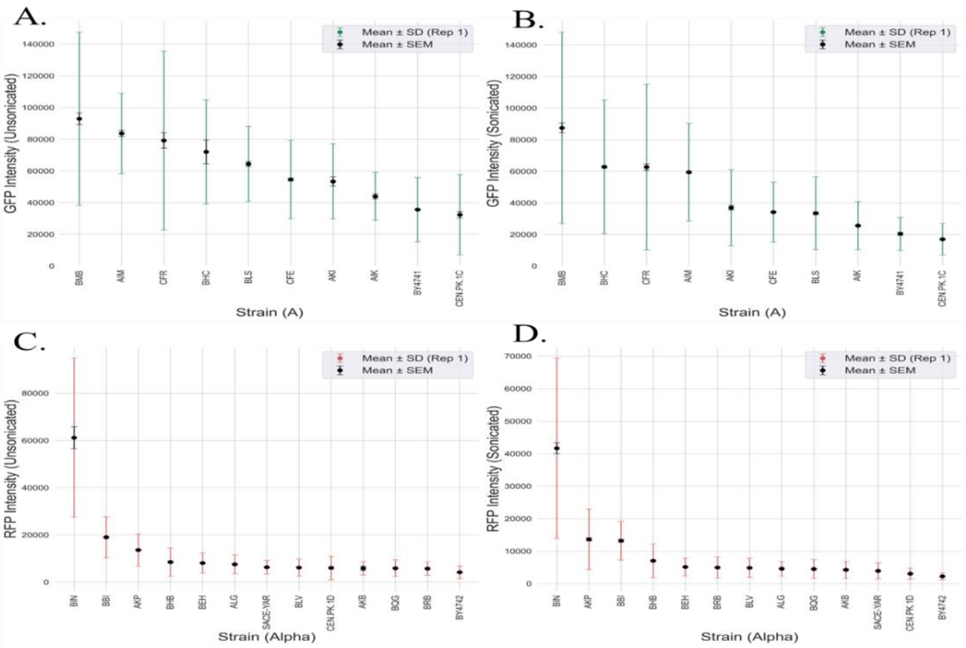


Figure 4 | GFP and RFP intensity in 18 strains from the collection under sonicated and unsonicated treatments. (A) and (B) display GFP intensity for 8 mating type *A* strains under unsonicated and sonicated conditions, respectively. (C) and (D) show RFP intensity for 10 mating type *alpha* strains under unsonicated and sonicated conditions, respectively. In both fluorescence channels (GFP and RFP) and treatments, the lab strain (BY4741) and the industrial strain (CEN.PK.1C) consistently appear at the lower end of the intensity spectrum.

7.1.2. Computational Analysis of Transcriptomic Data

To investigate the relationship between fluorescent gene expression and transcriptomic profiles, we used data from a recently published study analyzing the pan-transcriptome of most (approximately 1,000) of the yeast natural isolates, covering 4,977 core and 1,468 accessory genes (14) . For this study, we focused on our ~120 strains from the dataset to identify potential links between gene expression and fluorescent protein production. We reasoned that certain genes from the natural genome of yeast could affect the expression level of the fluorescent protein and we wished to identify these genes. Principal Component Analysis (PCA) was performed on the transcriptomic data to identify differences in the transcriptome of the various strains Figure 5 illustrates the distribution of strains based on PC1 and PC2. The wide range of PC1 values indicates wide differences in gene expression profiles across the strains, with extreme strains highlighted. To explore the functional contributions of genes to PC1, we identified the top 10% of genes contributing to either end of the PC1 axis and performed Gene Ontology (GO) analysis to categorize these genes into functional groups. On the negative side of PC1, genes associated with **translation**, **glycolytic processes**, and **metabolic processes** were highly represented. These findings suggest that strains at this end of the spectrum exhibit enhanced metabolic and biosynthetic activity. In contrast, on the positive side of PC1, genes involved in **respiratory electron transport** and **response to abiotic stimulus** were more prevalent, suggesting that strains with higher and positive PC1 values focus on stress responses and respiratory activity. Additionally, no significant correlation was found between PC1 values and GFP or RFP expression levels (Figure 5C). These results reveal distinct transcriptomic profiles associated with the extreme ends of PC1, reflecting possible biological differences in metabolic and stress-related gene expression.

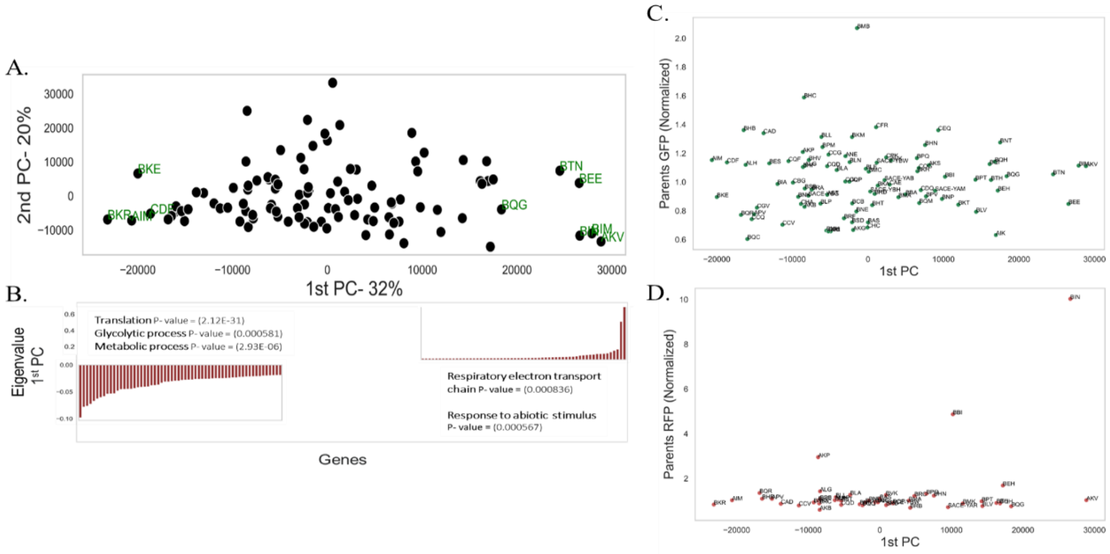


Figure 5 | PCA analysis of transcriptomic data and correlation with parental fluorescence levels. This figure presents an analysis of transcriptomic data using Principal Component Analysis (PCA) and scatter plots examining the relationship between PC1 and parental GFP and RFP expression levels. **(A)** The PCA plot illustrates the variance in the transcriptomic dataset, with the first principal component (PC1) explaining 32% of the variance and the second principal component (PC2) explaining 20%. Strains with extreme PC1 values are highlighted in green. **(B)** GO analysis of the genes contributing to the extremes of PC1 reveals functional differences. On the negative side of PC1, genes associated with translation, glycolytic processes, and metabolic processes (p-value < 0.001) are highly expressed. On the positive side of PC1, genes involved in the respiratory electron transport chain and response to abiotic stimulus (p-value < 0.001) are more prevalent. **(C)** The scatter plot shows normalized parental GFP expression levels plotted against PC1, and **(D)** displays normalized parental RFP expression levels against PC1. No significant correlation is observed.

7.2. Offspring Screen to Study Fluorescent Protein Inheritance Mechanisms

To investigate the inheritance mechanisms of fluorescent proteins, we generated and screened offspring combinations of the parental strains using two distinct approaches, called “all against all” and “one-against one.”

7.2.1. “All-Against-All” Mating and FACS Sorting

In this experiment, all parental strains were combined, allowing for free mating among them. After mating, offspring were isolated using FACS based on the fluorescence intensity of both parental markers, GFP and RFP. The top and bottom 10th percentiles of offspring, representing the highest and lowest fluorescence intensities, were sorted. Following mating barcode recombination is induced to generate fused barcodes whose sequence reveals the identity of the two parents of each offspring. To determine the parental origins of the sorted offspring, Next-Generation Sequencing (NGS) was performed on strain-specific genomic barcodes. Sequencing results provided read-per-million (RPM) values for each offspring in the sorted fractions. An enrichment index score was calculated by comparing the presence of each offspring in the high and low fluorescence fractions, reflecting their association with either extreme of the fluorescence intensity distribution (Figure 3, Appendix). We defined the $Enrichment\ Index = \frac{Sorted\ High\ (RPM)}{Sorted\ Low\ (RPM)}$. By plotting the standard deviation between repeats against the enrichment index for each offspring, we identified strains with both low standard deviation and high enrichment index values. The cutoff for low standard deviation was defined by identifying strains with an enrichment index greater than 20,000 and a standard deviation more than 100 units below the mean. Notably, certain parental strains appeared multiple times in the highlighted offspring. For example, SACE-YBW (mating type *A*) and BHB (mating type *A*) each appeared in three of the 12 highlighted offspring, and AKP (mating type *alpha*) appeared four times, suggesting a high positive contribution of these parental strains to offspring with high fluorescence intensity on both GFP and RFP channels (Figure 6). To visualize the strains with low enrichment index, we plotted a scatter plot with a log scale and identified strains that were located close to zero. These strains expressed GFP and RFP very poorly. Interestingly, the strain CQD (mating type *alpha*) appeared five times with five different partners, suggesting a negative contribution of this strain to the offspring's fluorescent expression (figure 7).

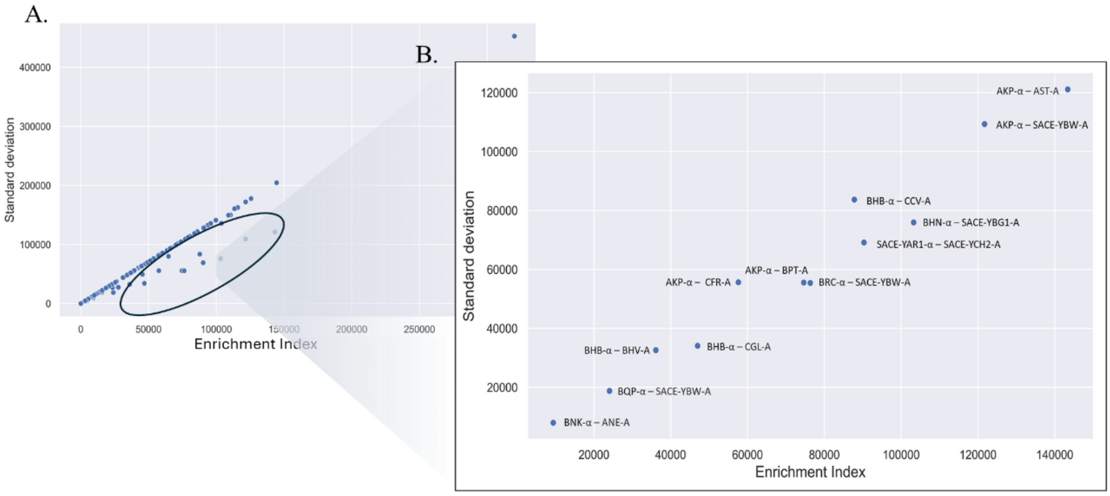


Figure 6 | Identification of strains with high enrichment index and low standard deviation. This figure highlights strains with high enrichment index and low standard deviation across three biological repeats. (A) Scatter plot displaying enrichment index versus standard deviation for various offspring. The area highlighted by the ellipse indicates a cluster of strains with both high enrichment index and relatively low standard deviation, suggesting a consistent presence in the high fluorescence fraction across repeats. (B) Close-up of the highlighted region. Notably, AKP (alpha) appeared 4 times, BHB (*A*) and SACE-YBW (*A*) appeared 3 times.

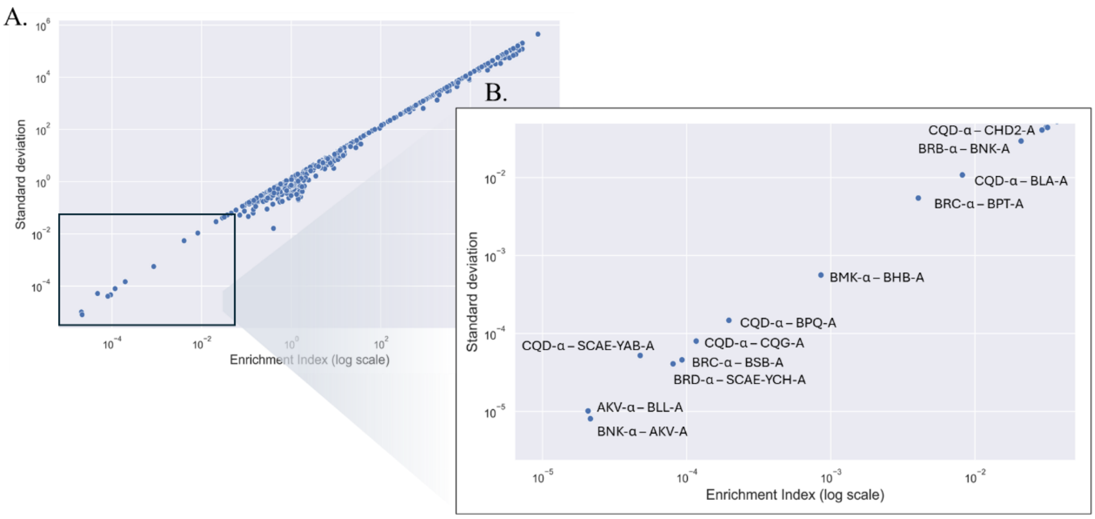


Figure 7 | Identification of strains with low enrichment index. This figure highlights strains with low enrichment index. (A) Scatter plot (log scale) displaying enrichment index versus standard deviation for various offspring. The area highlighted by a square indicates a cluster of strains with low enrichment index. (B) Close-up of the highlighted region. Notably, CQD (alpha) appeared five times, each with a different partner.

7.2.2. "One-Against-One"

In the previous method, some offspring combinations were missing due to variations in fitness and mating affinity among strains, as well as the loss of offspring before reaching the final sorting stage. To overcome this limitation, we generated a complementary method to measure GFP and RFP, culminating in a comprehensive library of 3,002 offspring, encompassing all possible successful mating combinations. These were arranged in 32 96-well plates. Using a FACS plate reader, we screened these offspring for GFP and RFP expression levels, revealing significant variability in fluorescence intensity across the library. To ensure the integrity of the high-throughput analysis and confirm minimal contamination, we randomly selected 68 offspring from the library for Sanger sequencing of the barcodes to verify parental identities. Of these, 63 offspring (92.6%) were confirmed to have the expected parental combinations, demonstrating the reliability of the mating and selection process (Table 4, Appendix). Notably, barcode recombination was performed, and the fused barcode underwent sequencing to confirm parental identity. Additionally, to rule out batch effects, we plotted a heatmap showing GFP and RFP levels for the offspring, organized by plate. The analysis revealed no evidence of batch effects or plate "geographical" effect in any of the plates (Figure 7, Appendix). Offspring from AKP (*alpha*) and BBI (*alpha*) consistently showed high expression levels of the parental trait RFP, highlighted in red and yellow, reflecting parental inheritance as these strains are among the top two RFP producers among the parents (Figure 8). Furthermore, microscope analysis of multiple offspring revealed that most, if not all, do not form aggregates like their haploid parents. Instead, they exhibit behavior more similar to the lab strain BY4741/2, which does not display aggregate formation (Figure 8, Appendix).

A comparative analysis was performed to evaluate the results of both approaches, and the findings are presented in chapter 7.4.

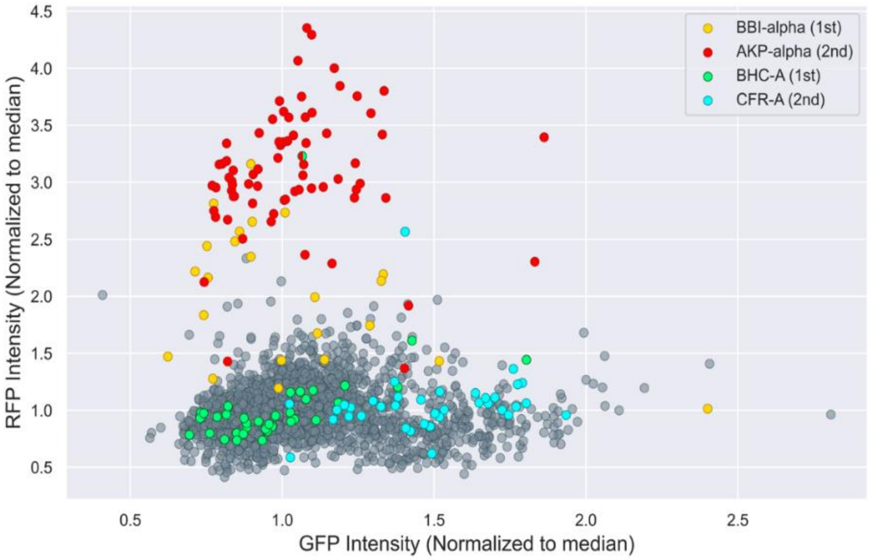


Figure 8 | Scatter plot of GFP and RFP expression levels in offspring. This scatterplot visualizes data from 3002 offspring derived from 100 different natural isolates. Showing GFP and RFP expression levels, normalized to the median. The offspring of the top two *alpha* strains in RFP production, BBI and AKP, are highlighted in yellow and red, respectively, indicating their offspring's elevated RFP expression levels compared to other strains. For comparison, the offspring of the top two *A* strain GFP producers, BHC and CFR, are highlighted in green and blue, respectively.

7.3. Comparing Offspring and Parental Fluorescence Expression Levels

An initial comparative analysis between GFP and RFP expression levels in parents and offspring revealed a surprising trend: parental expression levels were almost two-fold higher than those observed in the offspring. Specifically, a 2.27-fold difference was noted on average in GFP levels between mating type *A* parents and their offspring, and a 2.1-fold difference in RFP levels between mating type *alpha* parents and their offspring. This difference in expression levels contradicted our initial expectations, as diploid cells typically exhibit higher protein expression levels due to their larger size. Instead, the opposite pattern was observed, suggesting an unexpected mechanism influencing fluorescent protein expression in the offspring (Figure 9).

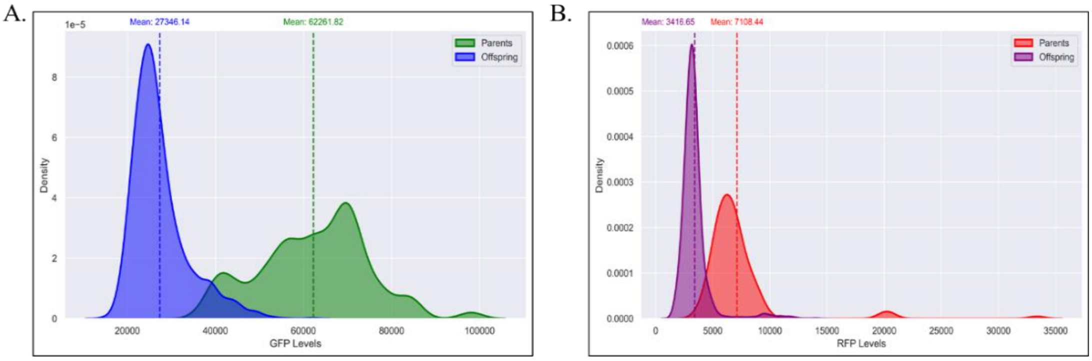


Figure 9 | Relationship between noise and mean expression levels of GFP and RFP in offspring (log scale). (A) Scatter plot showing the negative correlation between noise (σ^2/μ^2) and mean GFP expression levels. Each point represents an offspring strain. (B) illustrates the negative correlation between noise ($\sigma 2/\mu 2$) and mean RFP expression levels. While most offspring follow the expected trend of higher noise at lower mean expression levels, some strains diverge from the overall pattern.

7.3.1. Investigating Inheritance Patterns Based on Parental Expression Properties

To explore inheritance patterns, we investigated various approaches to model parental contributions to offspring fluorescence traits. Specifically, we calculated three parental metrics for GFP and RFP expression levels: the **average**, the **minimum**, and the **maximum** of parental values. These parental metrics were then compared to the offspring's GFP and RFP expression levels to assess which approach showed the strongest correlation between each pair of parents and their offspring. Our analysis revealed that all three parental metrics (average, minimum, and maximum) exhibited significant but weak positive Pearson correlations with offspring expression levels (Figure 9, Appendix). Focusing on the average parental values, the correlation between parental average expression and offspring GFP expression across all offspring was R = 0.04, with a p-value of 4.44e-02. Similarly, the correlation between parental average expression and offspring RFP expression was R = 0.65, with a p-value of 5.26e-308. After excluding outliers, specifically offspring of alpha mating type AKP and BBI, the correlation between average parental expression and offspring GFP expression increased slightly to R = 0.11, with a p-value of 3.374e-08. In contrast, the correlation between average parental expression and offspring RFP expression decreased to R = 0.12, with a p-value of 2.318e-09 (Figure 10). These results suggest that inheritance patterns for fluorescent protein expression are more complex than a simple averaging of parental traits. Multiple additional factors likely contribute to the observed patterns of fluorescent protein expression in offspring, indicating that the inheritance of these traits involves intricate mechanisms beyond straightforward parental. The correlation between the minimum parental value and offspring GFP expression was R = 0.12 (p-value = 4.37e-10). Similarly, for offspring RFP expression, the correlation was R = 0.21 (p-value = 2.85e-26) before outlier removal and R = 0.12 (p-value = 1.17e-08) after removal. For the maximum parental value, the correlation with offspring GFP expression was

R = 0.01 (p-value = 7.56e-01) before outlier removal and R = 0.06 (p-value < 5e-342) after removal. Likewise, for offspring RFP expression, the correlation was R = 0.68 (p-value < 5e-342) before outlier removal and R = 0.08 (p-value < 5e-342) after removal.

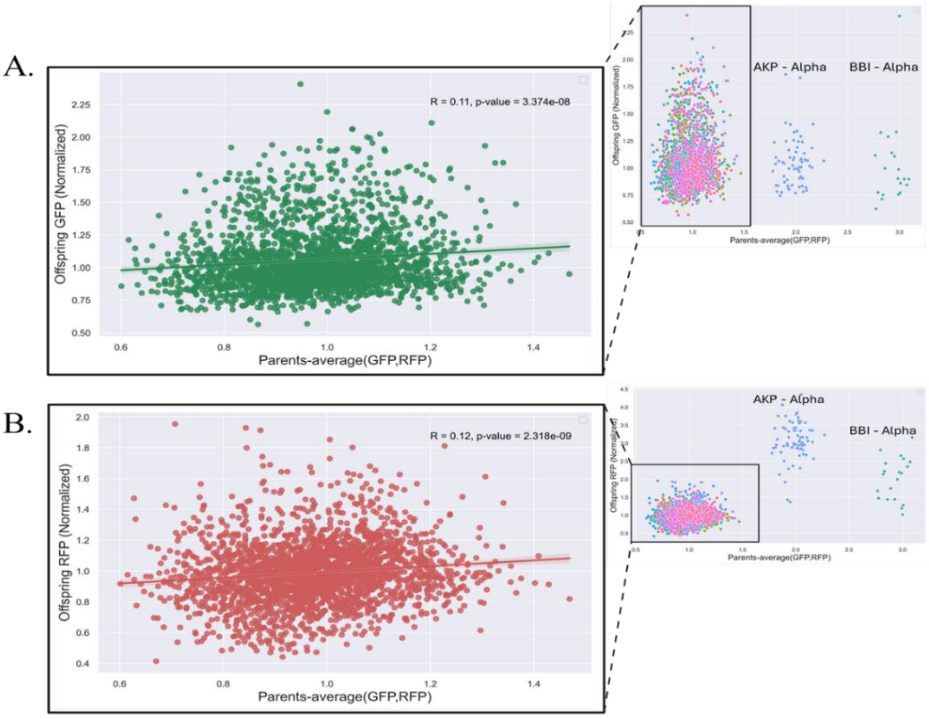


Figure 10 | Correlation between parental fluorescent averages and offspring expression levels. This figure illustrates the relationship between the average parental GFP and RFP expression levels and the offspring's corresponding expression values. (A) a weak positive correlation is observed between the offspring's GFP levels and the parental average (R = 0.11, p-value <0.001). (B) a similar weak positive correlation is seen for RFP levels (R = 0.12, p-value < 0.001). Both graphs display these correlations after excluding the outlier offspring of the *alpha* strains AKP and BBI.

7.3.2. Patterns of Parental and Non-Parental Inheritance

During our analysis, we identified an intriguing phenomenon that we termed parental and non-parental inheritance. Non-parental inheritance occurs when a parent influences the expression levels of the other parent's fluorescent color in the offspring. For instance, a mating type *A* parent, which contributes the GFP gene, consistently produced offspring with high expression levels of RFP, the fluorescent protein associated with the *alph*a mating type. In contrast, parental inheritance refers to cases where the parent directly influences the expression levels of its own fluorescent marker. For example, siblings from a mating type *alpha* parent exhibited similar levels of RFP expression, reflecting the parent's contribution. This observation highlights an intriguing inheritance pattern, where parental influence extends beyond its direct genetic contribution, suggesting the involvement of additional regulatory mechanisms.

(Figure 11) illustrates this phenomenon. To visualize this effect, we used two complementary approaches. In the first approach, we created heatmaps of GFP and of RFP expression levels in the offspring, organized by sibling groups, with *A* type parents arranged vertically and *alpha* type parents arranged horizontally. This layout allowed for direct comparison of siblings and facilitated the identification of distinct inheritance patterns. Certain strains consistently produced offspring with high expression levels, either for the parental or non-parental color. These patterns are highlighted by black rectangles (Figure 12). In the second approach, we calculated the mean expression level of the parental color for all offspring derived from the same parent and plotted these values in a box plot, comparing them to the parent's own expression level. For mating type A parents and their offspring's GFP mean levels, we observed that some parents consistently produced high-expressing offspring, despite having average GFP levels themselves. This pattern, highlighted by a black circle, suggests a pattern of non-parental inheritance (Figure 13 (A-B)). For *alpha* type parents, we conducted a similar analysis using the parental color RFP. Here, distinct patterns of parental inheritance emerged, particularly in strains like AKP and BBI. These strains were both high-expressing parents and consistently produced high-expressing offspring, indicating a strong parental inheritance (Figure 13 (C-D)). Further analysis of the box plot revealed that certain parental strains consistently appeared at the extremes of the distribution, producing offspring with either low or high expression levels of the non-parental color. For example, *alpha* type strains BHB and BRB were frequently observed at the lower and upper ends of the distribution of GFP expression levels, respectively. Similarly, *A* type strains AKI and BMA exerted a similar influence on RFP expression levels of the offspring (Figure 14). Additionally, by plotting the average expression level of the parental color in offspring from the same parent against the parent's own expression level, we found that in mating type *A*, no significant correlation was observed (R = 0.09, p-value = 9.813e-06), although some strains showed an improvement in GFP expression levels as diploids. In contrast, in mating type *alpha* there is a stronger positive correlation (R = 0.37, p-value = 6.675e-77), after outlier removal and (R = 0.80, p-value < 5e-324) after removal, with certain strains exhibiting similar improvement as diploids (Figure 15).

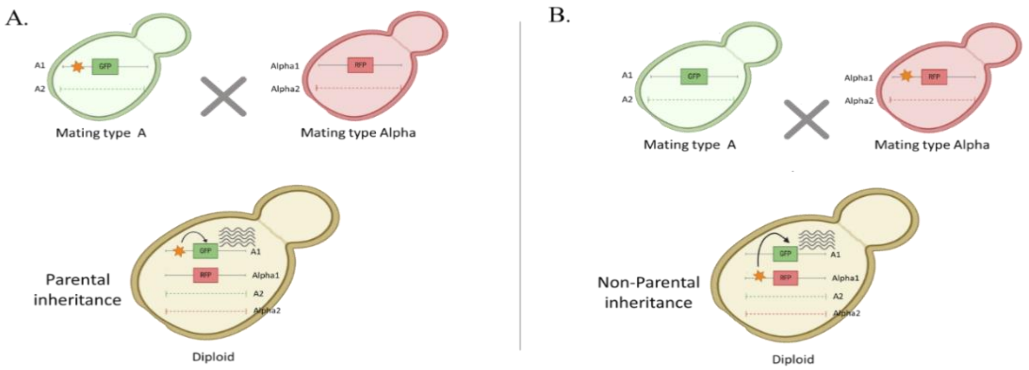


Figure 11 | Illustration of parental and non-parental inheritance. This figure illustrates examples of parental and non-parental inheritance patterns. (A) **Parental inheritance**: The offspring (diploid) resulting from the mating of a random *A* strain and a random *alpha* strain inherits a variation from the *A* parent, which specifically influences the expression levels of the parental color, GFP. (B) **Non-parental inheritance**: The offspring inherits a variation from the *alpha* parent that unexpectedly impacts the expression levels of the non-parental color, GFP, highlighting the influence of the *alpha* parent on a trait it does not directly contribute.

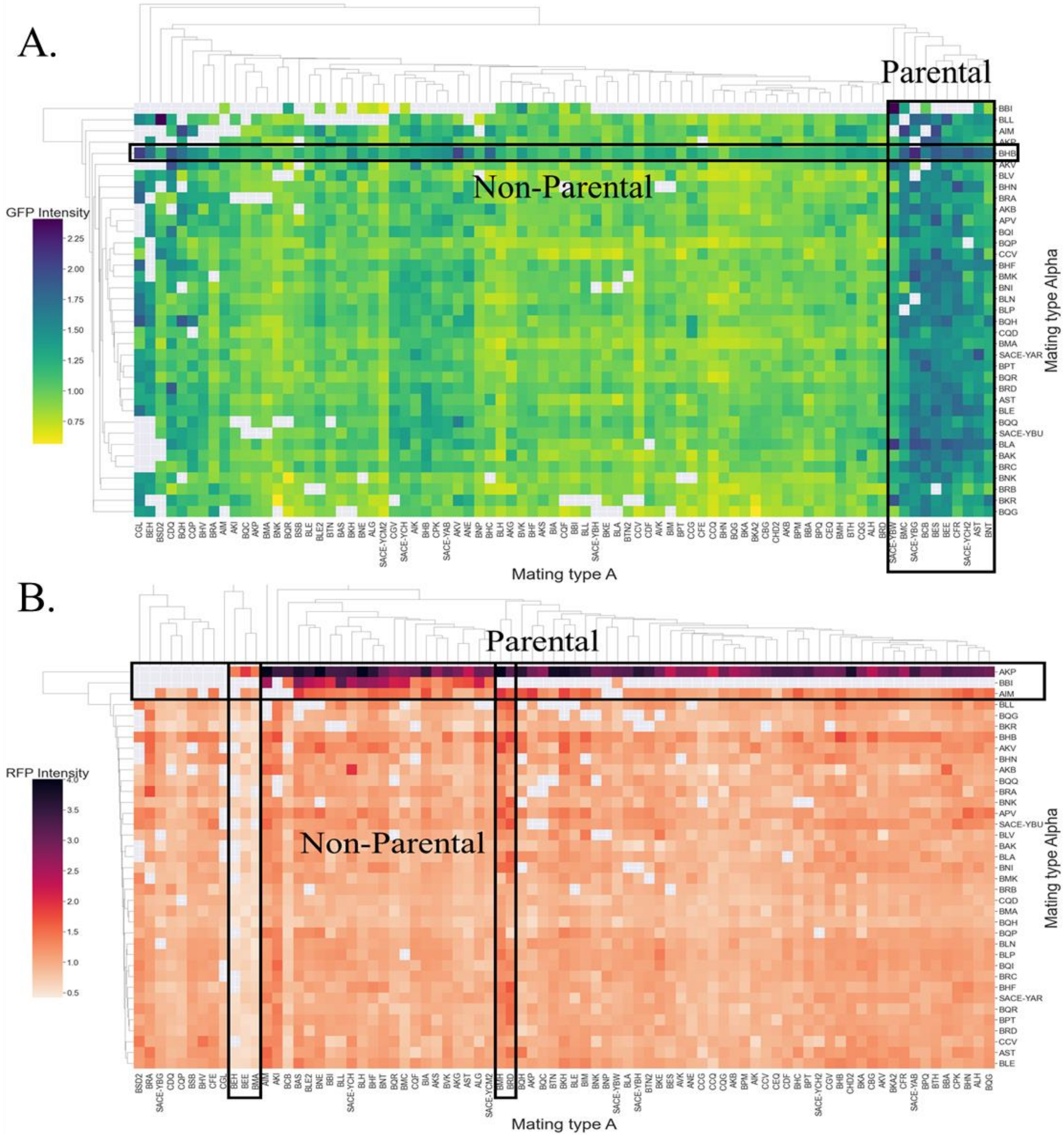


Figure 12 | Heatmap of GFP and RFP expression in offspring. These clustered heatmaps show GFP and RFP expression levels in the offspring, organized by comparison to siblings of the same mating type *A* (vertically) and *alpha* (horizontally). In (A) The color scale represents GFP expression levels, and in (B) RFP levels, with darker colors indicating higher intensities. Black rectangles highlight specific strains that consistently produce offspring with similar expression levels, either high or low, of the parental trait or the non-parental trait, as indicated on the map.

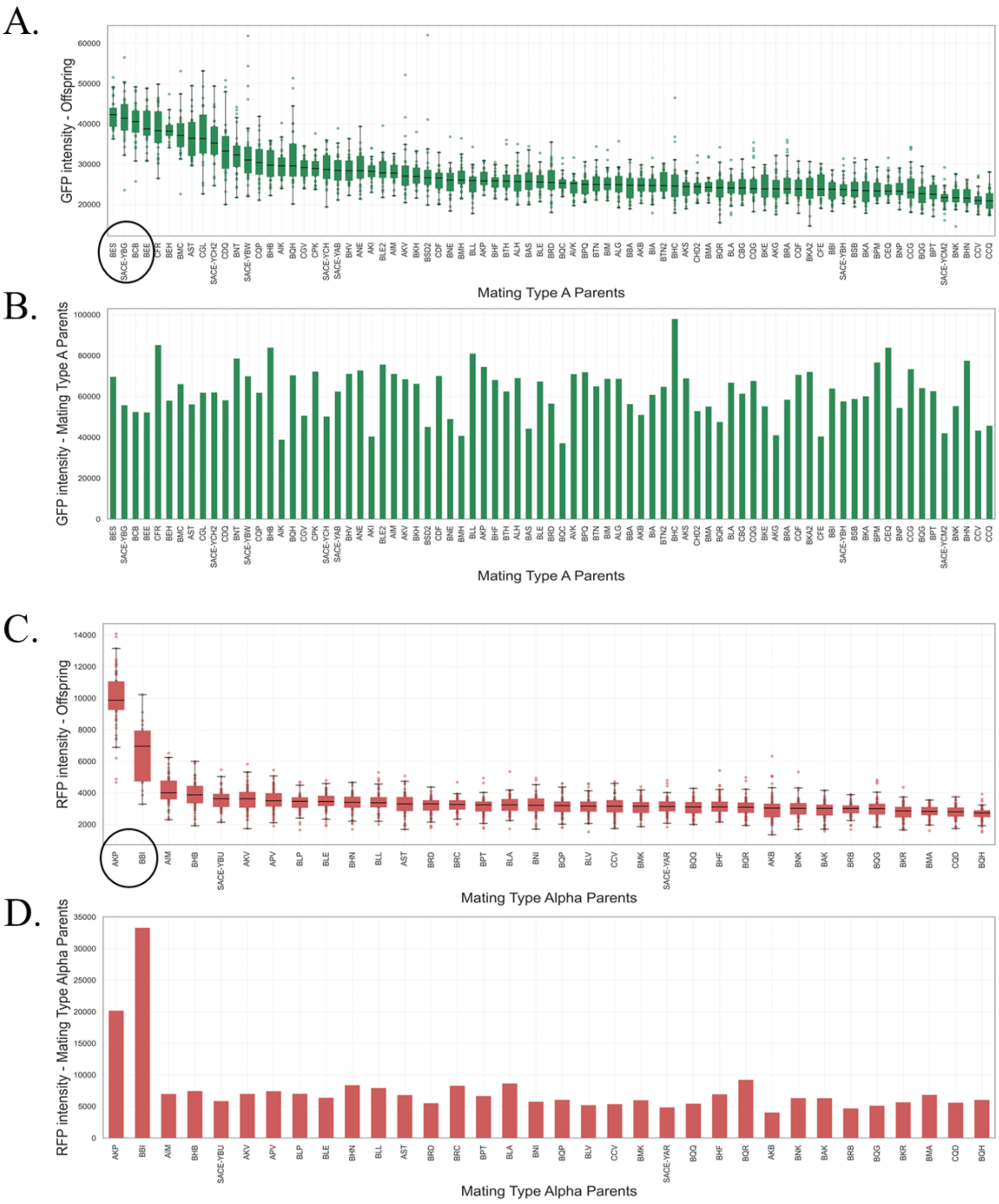


Figure 13 | Comparison of offspring and parental expression levels. This figure presents box plots comparing the expression levels of offspring with their parents. In (A) box plots of GFP expression levels of offspring derived from each *A* mating type parent, with the average expression for each sibling group shown within each box. (B) GFP expression levels of the *A* mating type parents for comparison. Similarly, in (C), the box plots illustrate the RFP expression levels of offspring from each *alpha* mating type parent, with the average RFP level for each sibling group shown in each box. (D) RFP expression levels of the *alpha* mating type parents.

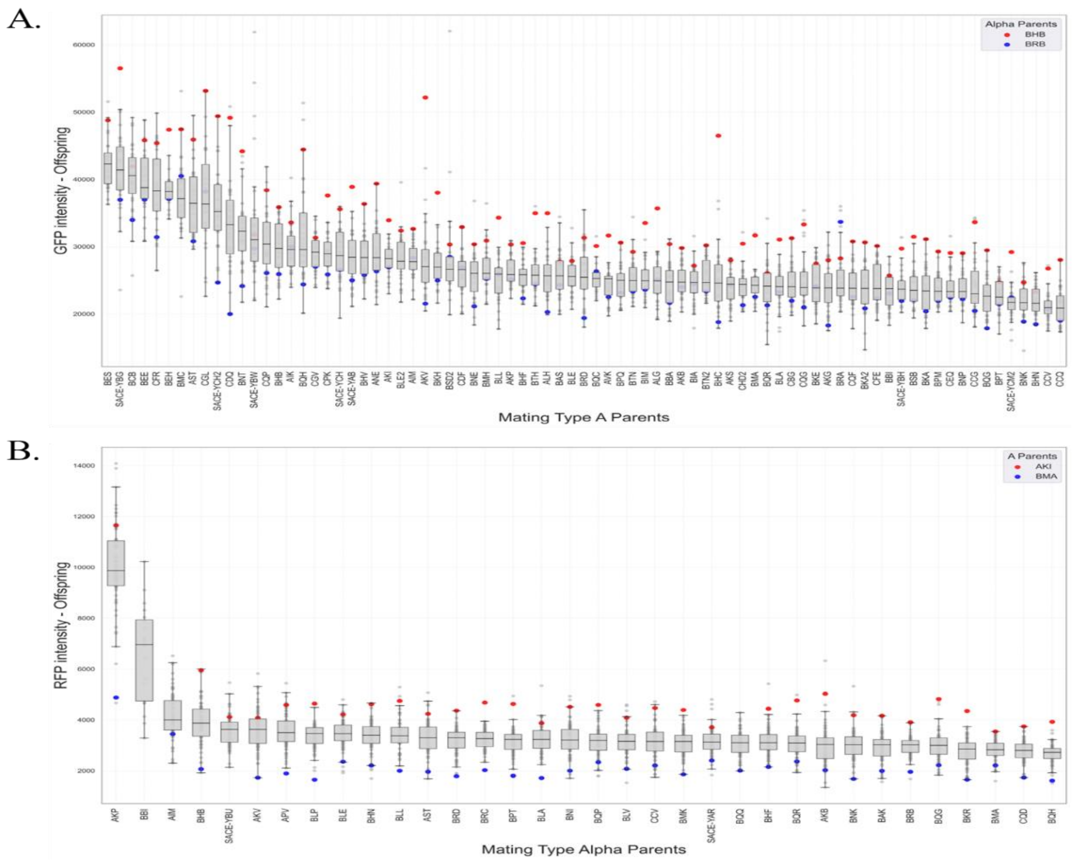


Figure 14 | Identification of parents with non-parental color influence. This figure presents box plots showing GFP and RFP expression levels in offspring, organized by parents in descending order of intensity. Each box displays the distribution of offspring expression levels for each parent, with the average level marked within the box. Blue and red dots indicate specific parents whose offspring appear at the extremes of the distribution, influencing the expression levels of the non-parental color. (A), BHB and BRB *alpha* parents that affect their offspring's GFP levels, while in (B), *A* mating type parents AKI and BMA influence their offspring's RFP levels.

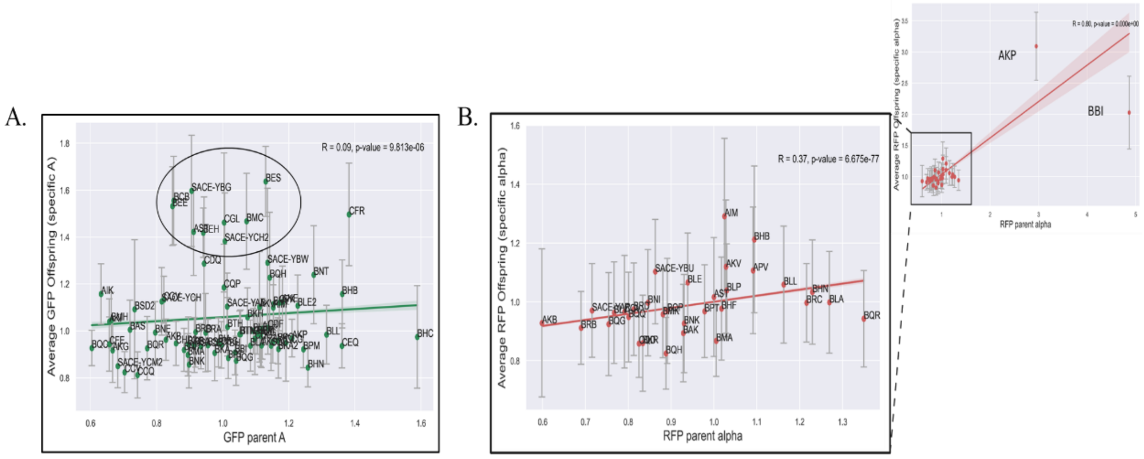


Figure 15 | Scatter plot showing the parental effect. Scatter plots comparing the average fluorescence levels of offspring to their respective parent's expression level. In (A), the scatter plot shows the average GFP expression of all offspring (siblings) derived from a specific *A* parent plotted against the GFP level of that *A* parent. No correlation is observed (R = 0.09, p-value < 0.001), although some strains exhibit, on average, higher GFP levels as diploids, highlighted with a black circle. (B) a similar plot is shown for RFP expression, comparing the average RFP levels of offspring to the RFP level of the *alpha* parent. A stronger positive correlation is observed (R = 0.37, p-value < 0.001) after removing outliers.

7.4. Comparing the Results of Both Approaches

We then turned to compare the results of the two approaches: “All-against-all” and “One-against-one”. Certain strains exhibited consistent patterns across both methods, reinforcing their significance in offspring inheritance. AKP *(alpha)* was a standout strain in both approaches in conferring high expression even upon its offspring with multiple partners. In the “All-against-all” method, it was identified as a parent to four different offspring out of 12 with a high enrichment index and low standard deviation (Figure 6B). Also, in the parental vs non-parental analysis these strains stood out as a strong contributor of parental contribution to expression of RFP (Figure 12B).

Another notable strain, SACE-YBW *(A)*, displayed similar consistency across both methods. In the “All-against-all” approach, it appeared as a parent to three different offspring out of 12 with a high enrichment index and low standard deviation (Figure 6B). In the “One-against-one” approach, SACE-YBW *(A)* exhibited strong parental effects, consistently producing offspring with high GFP expression levels (Figure 12A). Finally, BHB *(alpha)* emerged as an interesting strain with unique behavior. In the “All-against-all” approach, it appeared three times paired with different mating type *A* strains out of 12, positioned below the line in the standard deviation vs enrichment index graph (Figure 6B). In parental vs. non-parental analysis, these strains stood out as strong contributors to both parental and non-parental inheritance of GFP and RFP expression, respectively (Figure 12). When examining low-expressing offspring, CQD (*alpha*) consistently produced offspring with low GFP and RFP expression levels. In “All-against-all” it appeared five times paired with different mating type *A* strains with very low enrichment index (Figure 7B). And in “One-against-one” method, it repeatedly produced offspring with low RFP expression, as shown in the siblings' box plot (Figure 13), which appears at the far right, indicating low expression levels (Figure 13B).

To better visualize the comparison, strains that stood out in the “All-against-All” approach were highlighted in the "One-against-One" GFP vs. RFP density plot. Strains with a high enrichment index are represented by circles, while those with a low enrichment index are marked with triangles. Nine out of the 12 high-enrichment strains, highlighted with dashed circles, are located at the higher end of the graph, exhibiting high expression levels of either GFP or RFP (Figure 16). These strains are positioned outside the dense regions where the majority of offspring are clustered, further emphasizing their unique and distinct expression patterns.

In contrast, strains with a low enrichment index do not exhibit high levels of GFP or RFP and are instead concentrated within the dense central region of the plot, exhibiting low expression levels of both GFP and RFP.

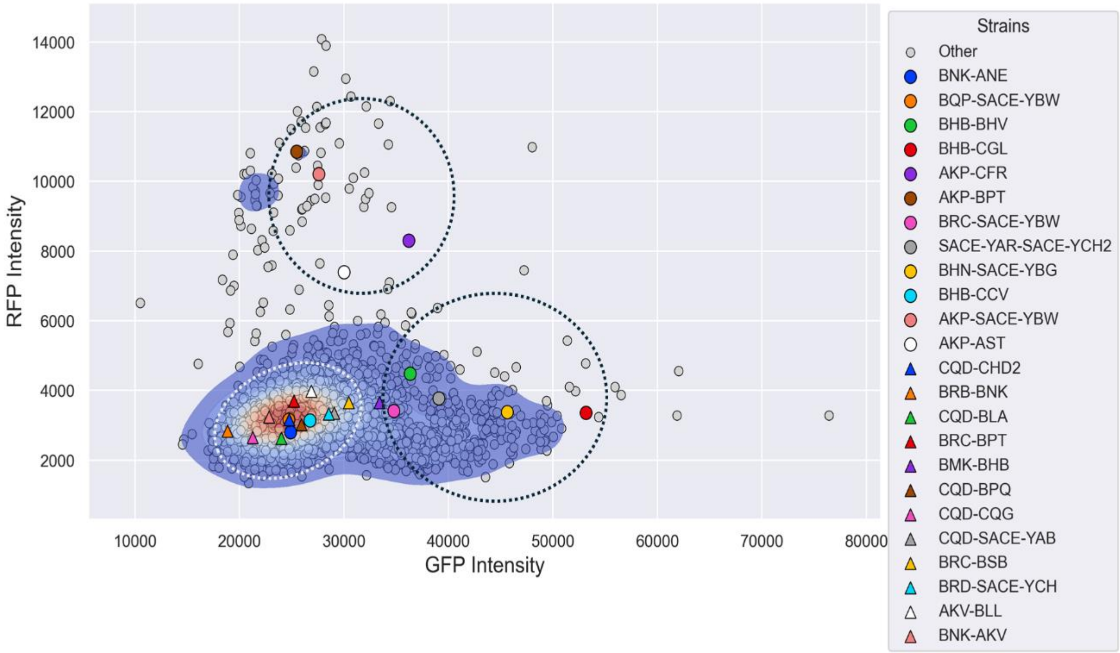


Figure 16 | Density plot of GFP vs. RFP intensities of offspring from the "one-against-one" approach. This plot illustrates the distribution of GFP and RFP intensities across offspring strains, visualized as a density plot. Highlighted are strains identified as interesting in the "All-against-All" approach, categorized into two groups based on their enrichment index values: **High Enrichment Index Strains** (circles): These strains exhibit high expression of either GFP or RFP and are mostly located outside the regions with the highest density of offspring. Notably, 9 out of the 12 strains highlighted with black dashed circle, exhibit high expression of either GFP or RFP. **Low Enrichment Index Strains** (triangles): These strains are associated with low expression levels of both GFP and RFP and are concentrated within the white dashed circle, aligning with the regions of highest offspring density.

7.5. Noise in Fluorescent Protein Expression Across Strains and Generations: Scaling with Mean Expression and Inheritance Patterns

To investigate variability in protein expression, we calculated the noise (σ^2/μ^2) in GFP and RFP production across different *S. cerevisiae* strains. Noise was measured for mating type *A* strains (GFP expression) and mating type *alpha* strains (RFP expression). Consistent with findings reported by (10), which demonstrated that low-abundance proteins exhibit higher noise, while high-abundance proteins show lower noise, following a predictable scaling law, a negative correlation between noise and mean fluorescence was observed across strains in both mating types and their offspring (Figure 17). To determine the general scaling relationship, we calculated the fitted line for the parents after excluding outliers using the interquartile range (IQR) method, a systematic approach ensuring unbiased outlier removal based on statistical thresholds. This resulted in slopes of -0.93 for GFP and -0.82 for RFP. Notably, certain parent strains diverged from this trend, including SACE-YBW and CFR (*A*), which exhibited higher noise relative to GFP expression, and AKP and BBI (*alpha*), which displayed similar divergences in RFP noise (Figures 17A and 17B). For the offspring, based on observations in the parents, the offspring of SACE-YBW and CFR (*A*), as well as AKP and BBI (*alpha*), were manually excluded from the general population to assess whether they displayed similar behavior. A fitted line on the remaining general population revealed slopes of -0.22 for GFP and -0.52 for RFP, highlighting the negative correlation between noise and mean fluorescence. Notably, the excluded offspring (SACE-YBW and CFR from *A*, and AKP and BBI from *alpha*) exhibited elevated noise levels relative to their mean expression, mirroring the distinctive patterns observed in their parents (Figures 17C and 17D).

7.5.1. Residual noise calculation

To evaluate the noise properties of yeast strains independent of expression levels, we performed a "noise residual" analysis, defined as the vertical distance between observed noise values and the fitted line in log-log space (corresponding to the difference between observed noise level and the expected level given the mean expression and the fitted line). We analyzed both parent strains and their offspring, focusing on noise residuals for GFP and RFP fluorescence markers. Specifically, we averaged the noise residuals of offspring from mating type *A* parents and plotted these against corresponding parent noise residuals. For mating type *A* strains, we observed a significant positive correlation in GFP noise residuals (R = 0.55, p-value = 1.990e-07). A similar analysis of mating type *alpha* strains revealed a non-significant positive correlation in RFP noise residuals (R = 0.41, p-value = 1.568e-02). These results suggest that intrinsic noise characteristics can be transmitted across generations in these yeast strains, with varying degrees of statistical significance between mating types and fluorescence markers (Figure 18). To further investigate the relationship between parental noise residuals and offspring. First, we calculated the average noise residuals of both parents and plotted these against the GFP and RFP noise residuals of the offspring. For GFP, we observed a significant positive correlation (R = 0.22, p-value = 1.81e-29). For RFP, we also observed a significant positive correlation (R = 0.16, p-value = 2.264e-16) (Figure 19A and 19B).

Furthermore, we plotted the offspring GFP and RFP noise residuals against the maximum and minimum parental noise residuals. For the maximum parental noise residuals, we observed a significant positive correlation with offspring GFP noise residuals (R = 0.38, p-value = 3.29e-86), while no correlation was found for RFP (R = 0.06, p-value = 2.17e-03) (Figure 19C and 19D). As for the minimum parental noise residuals, we observed no correlation with offspring GFP noise residuals (R = 0.02, p-value = 2.86e-01) and a weak positive correlation for RFP (R = 0.16, p-value = 3.93e-16) (Figure 19E and 19F).

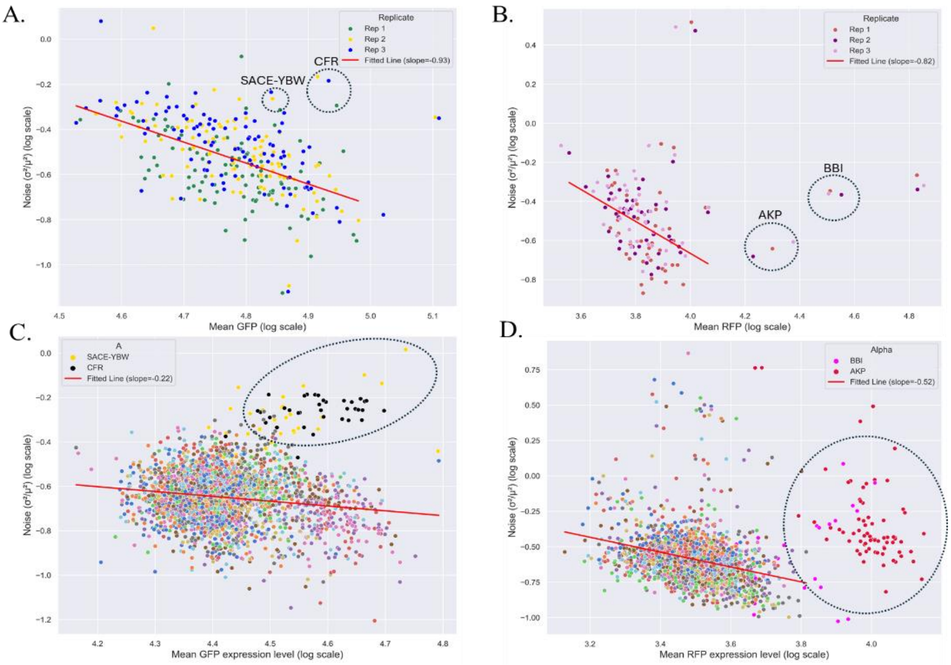


Figure 17 | Relationship between noise and mean expression levels of gfp and rfp in parents and offspring (log scale). (A) Scatter plot of GFP noise versus mean GFP expression in parent strains of mating type *A* of 3 biological repeats. A fitted line calculated after excluding outliers using the interquartile range (IQR) method shows a negative slope (-0.93). (B) Scatter plot of RFP noise versus mean RFP expression in parent strains of mating type *alpha*. of 3 biological repeats. The fitted line calculated after excluding outliers using the IQR method has a slope of -0.82. (C) Scatter plot of GFP noise versus mean GFP expression in offspring of mating type *A*. Data points are color-coded based on parental identity, with outlier strains SACE-YBW (black) and CFR (yellow) excluded from the fitted line calculation (slope = -0.22). The remaining data reveal a weak negative correlation. (D) Scatter plot of RFP noise versus mean RFP expression in offspring of mating type *alpha*. Data points are color-coded based on parental identity, with outlier strains AKP (red) and BBI (pink) excluded from the fitted line calculation (slope = -0.52). The general population shows a negative correlation, while the excluded strains maintain elevated noise levels relative to their mean expression.

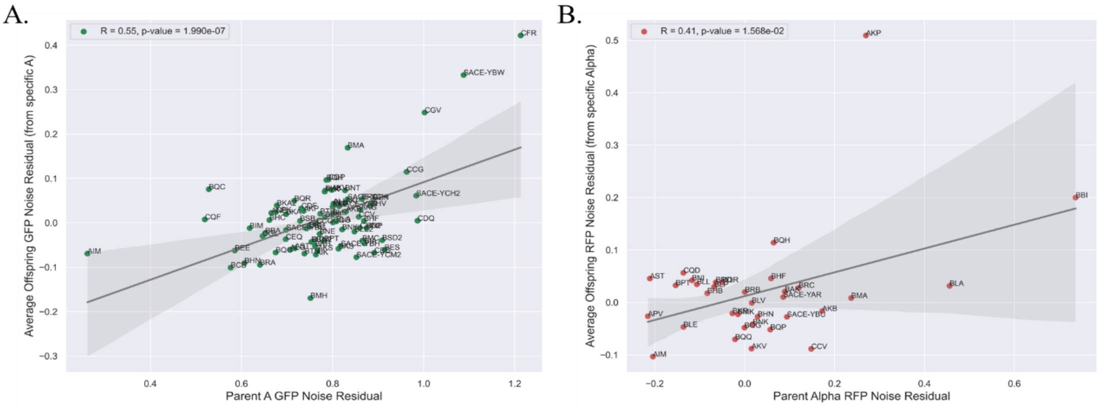


Figure 18 | Relationship between noise residuals of parents and their offspring. Scatter plots illustrate the relationship between the noise residuals of parents and their respective offspring. In (A), the average GFP noise residual of all offspring (siblings) derived from a specific mating type *A* parent is plotted against the GFP noise residual of that parent. A significant positive correlation is observed (R = 0.55, p-value < 0.001). In (B), the average RFP noise residual of offspring is plotted against the RFP noise residual of the corresponding mating type *alpha* parent. A non-significant positive correlation is observed (R = 0.41, p-value = 1.568e-02).

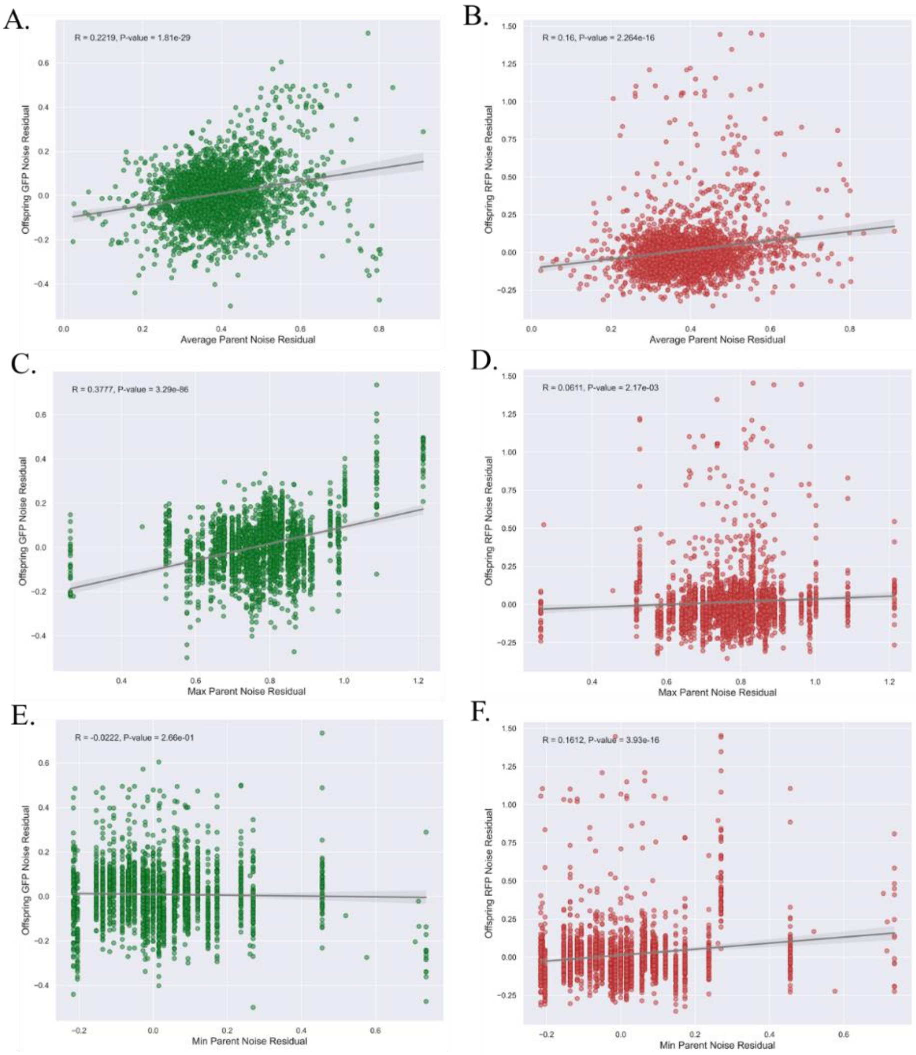


Figure 19 | Correlation between parental noise residuals and offspring noise residuals across different measures. (A)-(B) Scatter plots showing the relationship between the average parental noise residual and offspring GFP and RFP noise residual, respectively. A significant positive correlation is observed in both markers (R = 0.22, p-value = 1.81e-29) and (R = 0.16, p-value = 2.26e-16). (C)-(D) Scatter plots showing the relationship between the maximum parental noise residual and offspring GFP and RFP noise residual, respectively. A strong positive correlation is observed for GFP (R = 0.38, p-value = 2.32e-86), while no significant correlation is observed for RFP (R = 0.06, p-value = 2.17e-03). (E)-(F) Scatter plots showing the relationship between the minimum parental noise residual and offspring GFP and RFP noise residual, respectively. No correlation is observed for GFP (R = 0.02, p-value = 2.86e-01), while a weak positive correlation is observed for RFP (R = 0.16, p-value = 3.39e-16).

7.6. Cell Size in Parents and Offspring and Its Inheritance

Cell size was another quantitative trait used to study inheritance patterns. Alongside screening for fluorescent protein levels, we used the FACS machine to collect forward scatter area (FSC-A) measurements, a parameter serving as a proxy for cell size. Since larger cells scatter more light, FSC-A values provide an estimate of cell size, enabling us to analyze its inheritance patterns. By measuring FS in both parents and offspring, we were able to compare offspring values to their respective parents, investigating potential inheritance patterns for this trait.

Additionally, we compared cell sizes among siblings to assess variation within offspring groups, providing further insight into cell size inheritance. The initial analysis to visualize this data was through a heatmap similar to the one used for fluorescent protein levels; all offspring from a single *A* parent are arranged vertically, and offspring from a single *alpha* parent are arranged horizontally, and the color scale reflects cell size. This visualization reveals that certain strains consistently produce offspring with similar cell sizes, suggesting the possibility of an underlying mechanism of inheritance for cell size (Figure 20). Following the analysis of sibling groups, we proceeded to examine the correlation between offspring and parental cell sizes. To do this, we averaged the cell size for each sibling group and plotted this average against the cell size of their respective parents. A positive correlation was observed between the cell sizes of siblings and their *alpha* parents (R = 0.41, p-value = 7.304e-100) and between siblings and their *A* parents (R = 0.25, p-value = 7.047e-38), suggesting that parental cell size influences offspring cell size to some extent (Figure 21). To further evaluate whether averaging the cell sizes of both parents could serve as a reliable predictor for offspring cell size, we plotted the average parental cell size against the offspring cell size. This analysis revealed a weak but significant positive correlation (R = 0.23, p-value = 1.421e-32) (Figure 22). Additionally, we explored the minimum and maximum parental cell sizes as predictors, but these measures showed even weaker correlations with offspring cell size (Figure 10, Appendix).

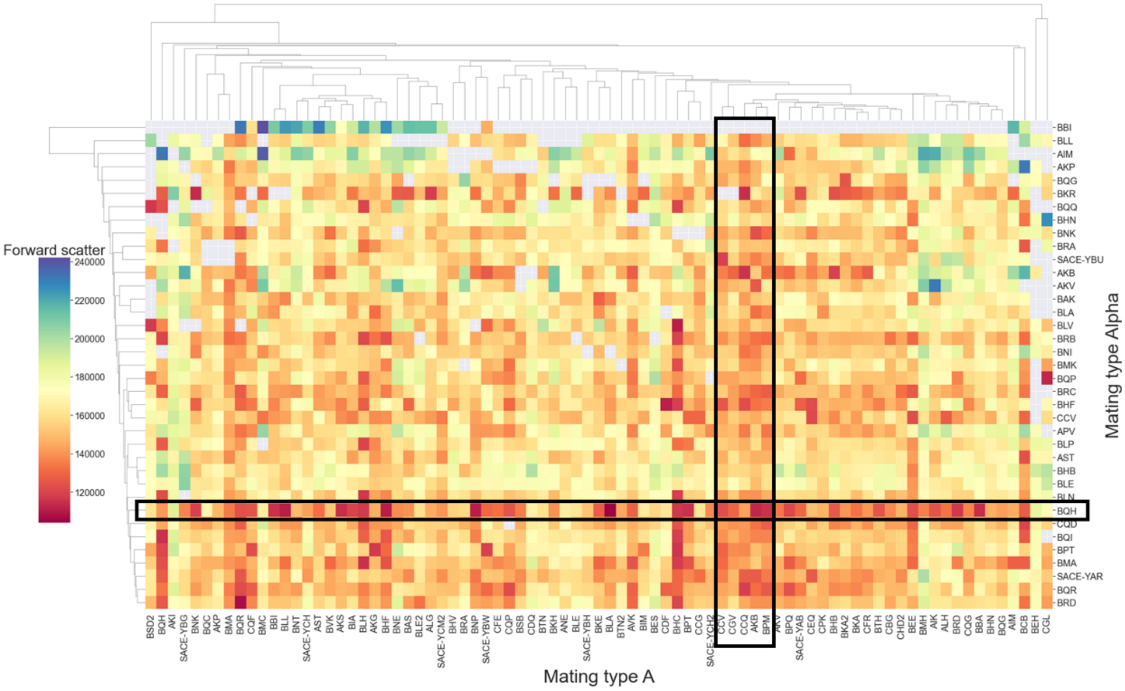


Figure 20 | Heatmap of offspring cell size. This heatmap displays FS values representing cell sizes of offspring, allowing comparison among siblings from the same parent. Offspring from *A* mating type parents are arranged vertically, while offspring from *alpha* mating type parents are arranged horizontally. The color scale reflects FSC values, with darker colors indicating larger cells. Black rectangles highlight a few strains that consistently produce offspring with similar size.

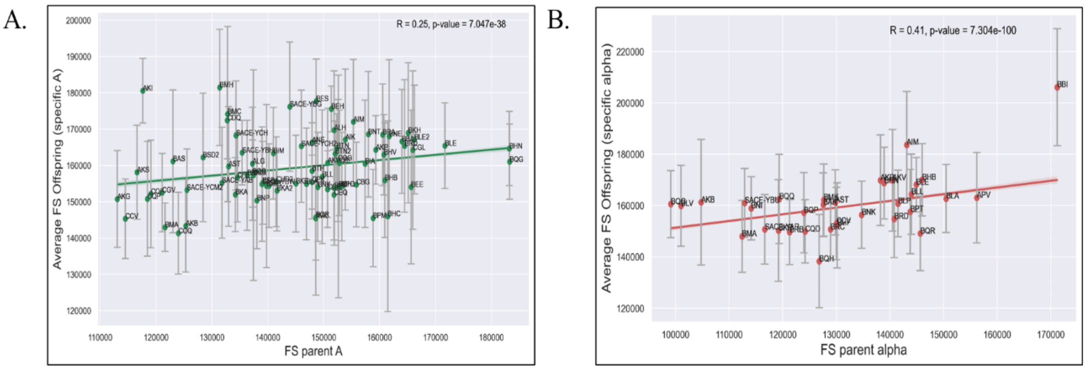


Figure 21 | scatter plot of average offspring cell size against parental cell size. This figure presents scatter plots showing the relationship between the average cell size of all offspring from a specific parent and the parent's cell size. In (A) the scatter plot shows the correlation between cell sizes of offspring and their *A* parents, displaying a positive correlation (R = 0.25, p-value < 0.001). In (B) the scatter plot shows the correlation between cell sizes of offspring and their *alpha* parents, displaying an even stronger positive correlation (R = 0.42, p-value < 0.001).

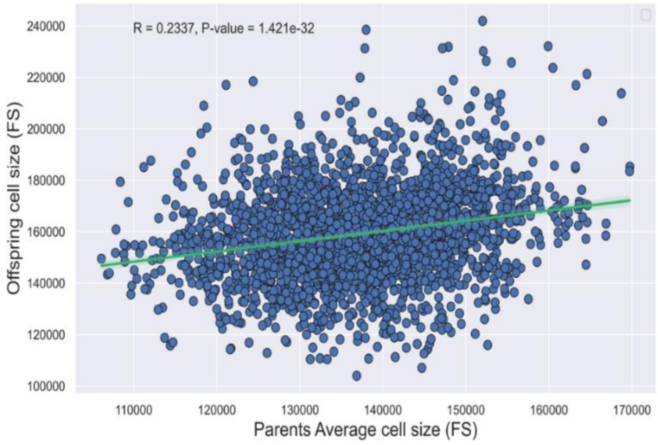


Figure 22 | Relationship between parental average cell size and offspring cell size. This scatter plot illustrates the relationship between the average cell size of both parents (FS values) and the measured cell size of the offspring. A weak but significant positive correlation is observed (R = 0.24, p-value < 0.001).

7.7. Correlation Between Cell Size and Fluorescent Protein Production

In this section of the study, I explored the relationship between cell size and fluorescent protein production among the offspring. To identify the best model, I applied polynomial regression of degrees 1 (linear), 2 (quadratic), and 3 (cubic) and compared their fits.

The second-degree polynomial provided the best fit, as determined by equation inspection RMSD values. A quadratic regression was applied, revealing a positive correlation between cell size and both GFP and RFP production, with RMSD values of 5852.17 (p-value = 9.14e-83) and 1257.69 (p-value = 1.39e-95), respectively. These results indicate that larger cells generally produced more fluorescent protein, suggesting that cell size serves as a predictive factor for protein production across most strains. However, an exception was observed in the offspring of AKP *alpha* type. This strain produced offspring with high levels of RFP that were not proportional to their cell size, thus diverging from the general trend observed in other offspring. Instead, their fluorescent protein production followed a unique trend, which was better described by a first-degree polynomial fit (RMSD = 1516.43, p-value = 1.78e-05) (Figure 23). Of note, this strain also featured strong RFP expression as a parent and exhibited strong parental inheritance effect. These findings highlight a general relationship between cell size and fluorescent protein production, while also emphasizing the existence of strain-specific deviations, such as those observed in the AKP *alpha* strain.

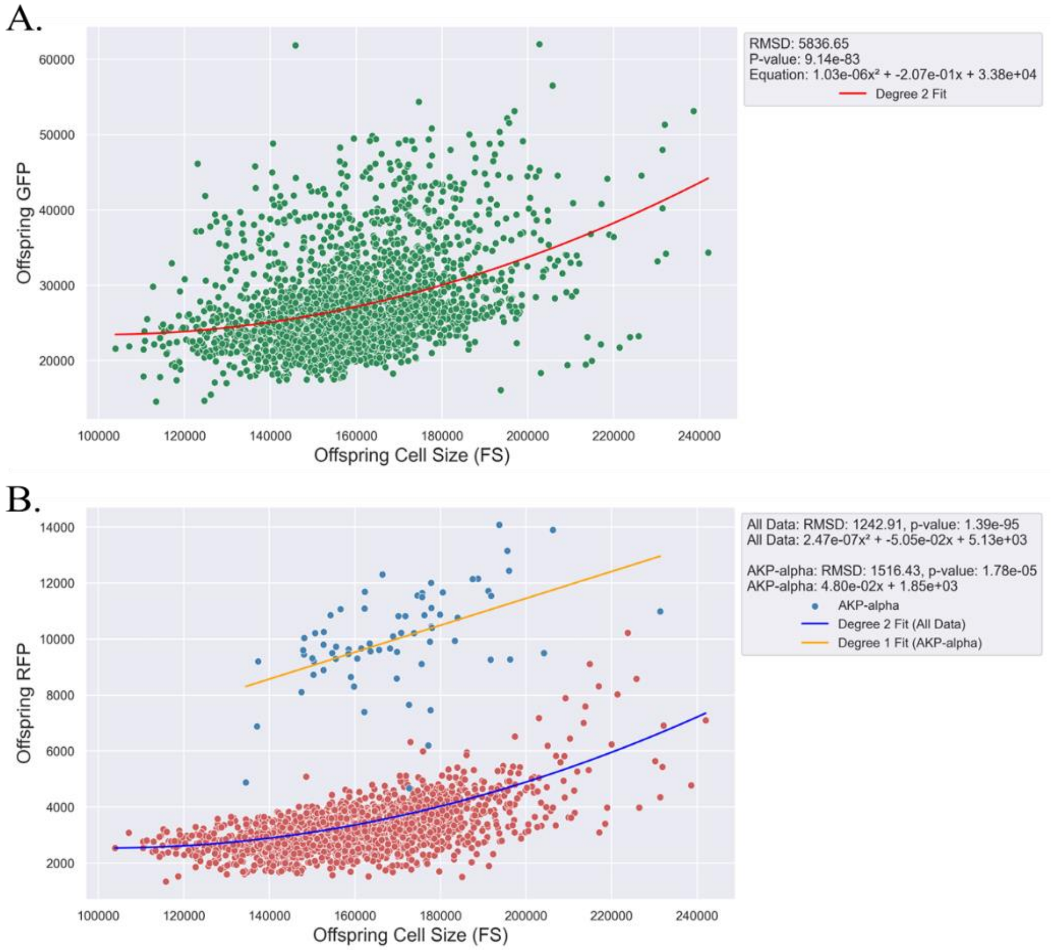


Figure 23 | Correlation between cell size and fluorescent protein production. This scatterplot visualizes the relationship between cell size and fluorescent protein production in offspring. The x-axis represents cell size, while the y-axis shows fluorescent protein production intensity. In (A), a significant second-degree polynomial fit is observed between GFP production and cell size (RMSD = 5852.17, p-value < 0.001). In (B), a similar second-degree polynomial fit is evident for RFP production and cell size (RMSD = 1257.69, p-value < 0.001). Blue dots represent all progeny of the AKP *alpha* strain, which deviate from the general trend. These progenies exhibit a unique relationship, forming a first-degree polynomial fit distinct from other offspring (RMSD = 1516.43, p-value < 0.001).

7.8. Machine Learning Analysis for Predicting Inherited Features

Using the DoubleLearningCatBoostRegressor model and SHAP (SHapley Additive exPlanations), we aimed to identify the key features that most influence the prediction of GFP and RFP expression levels in our offspring, as well as their noise residuals. The model's performance was evaluated using Pearson correlation and R-squared values. For offspring's GFP prediction, the model achieved a Pearson correlation of 0.9980 and an R^2 value of 0.9956 on the training set. On the test set, the Pearson correlation was 0.8453, with an R^2 value of 0.7083. For offspring's RFP prediction, the training set showed a Pearson correlation of 0.9990 and an R^2 value of 0.9980. On the test set, the Pearson correlation was 0.9575, and the R^2 value was 0.9094. For offspring's GFP noise residual prediction, the model achieved a Pearson correlation of 0.9988 and an R^2 value of 0.9974 on the training set. On the test set, the Pearson correlation was 0.9422, and the R^2 value was 0.8791. Finally, for offspring's RFP noise residual prediction, the training set showed a Pearson correlation of 0.9804 and an R^2 value of 0.9569. On the test set, the Pearson correlation dropped to 0.6195, with an R^2 value of 0.3787. Plots for predicted vs actual values for each target are presented in Appendix, Figure 11. To visualize the results, SHAP values provided a comprehensive understanding of how each feature contributes to the model's predictions. Our analysis revealed that specific genes and genetic factors, such as cell size, play significant roles in determining fluorescent protein production.

7.8.1. Predicting Offspring’s Fluorescent Intensity

When running the model to predict the offspring’s GFP and RFP mean expression, we observed a strong parental effect. 11 out of the 12 genes selected as the most important predictors of GFP expression levels came from the mating type *A* parent, the same mating type contributing to the GFP marker. A similar result was observed for RFP, where 12 out of the 15 genes selected were from the *alpha* mating type parent. Interestingly, cell size (forward scatter) of the offspring emerged as the top feature predicting the fluorescent markers in both GFP and RFP.

For offspring’s GFP level, the gene with the highest prediction importance was YAR023C (DFP1), a gene of uncharacterized function from the A parent. Additionally, the offspring’s RFP mean intensity and standard deviation, along with the A parent’s GFP noise residual, were also important features predicting the offspring’s GFP expression (Figure 24A).

For offspring’s RFP level, the top predictor gene was YKL155C (RSM22), which encodes a mitochondrial ribosomal protein of the small subunit. This gene is also predicted to function as an S-adenosylmethionine-dependent RNA methyltransferase and CoA synthetase. Similar to the GFP prediction, the offspring’s GFP expression levels, as well as the parent's RFP, were key features influencing RFP expression (Figure 24B).

7.8.2. Predicting Offspring’s Noise Residual

When predicting noise residuals for both GFP and RFP, we observed an intriguing result: the noise and noise residual of one marker was an important feature to predict the noise residual of the other marker. Specifically, the offspring’s GFP noise and noise residual were important features to predict offspring RFP noise residual, and vice versa. Cell size (forward scatter) appeared as an important feature for predicting noise residuals. For GFP noise residual, the mating type *A* parent’s GFP noise residual was a significant predictor, highlighting a strong parental effect. In terms of genes, we observed a strong parental effect for GFP noise residual, with 11 out of 14 genes coming from the mating type *A* parent. However, for RFP noise residual, 9 out of 14 genes coming from the *A* parent, which did not contribute to the RFP marker (Figure 25).

7.8.3. Identifying Connections Between Genes with High Predictive Importance and Their Target Outcomes

When comparing the genes predicting each target and looking for connections, we observed some interesting results. Hem13 (YDR044W) appeared as a predictive feature for both offspring’s RFP mean expression levels (Figure 24B) and RFP noise residuals (Figure 25B), Hem13 encodes coproporphyrinogen III oxidase, an oxygen-requiring enzyme that catalyzes the sixth step in the heme biosynthetic pathway.

SPP1 (YPL138C), SPP2 (YOR148C), and SAS4 (YDR181C) were identified as predictive features for both offspring’s mean GFP expression (Figure 24A) and GFP noise residuals (Figure 25A). SPP2 is required for telomeric transcriptional silencing and promoting meiotic double-strand break formation, while SPP1 is an essential protein that facilitates the first step of splicing and is required for final spliceosome maturation and activation. SAS4 acetylates free histones and nucleosomes, playing a role in regulating transcriptional silencing. These three genes are involved in chromatin modification and silencing, an interesting finding since they play a role in transcription regulation.

FAA2 (YER015W) appeared as a predictive feature for both offspring’s mean RFP and RFP noise residuals (Figure 24B and Figure 25B). It encodes a medium-chain fatty acyl-CoA synthetase, involved in activating imported fatty acids. Surprisingly, SOR1 (YDL246C) and SOR2 (YJR159W), which encode sorbitol dehydrogenases, were identified as genes predicting offspring’s GFP expression (Figure 24A). According to SHAP values, low expression levels of these genes lead to more accurate predictions of offspring’s GFP. Additionally, two genes, HOT13 (YKL084W), a member of the zinc cluster family of proteins, and RDS2 (YPL133C), a zinc-binding mitochondrial intermembrane space (IMS) protein, were present as features predicting offspring’s GFP noise residual and RFP noise residual, respectively (Figure 25). All the information about gene functions was obtained from the Saccharomyces Genome Database (SGD) (23).

Note: Offspring plate appeared as features in the prediction of offspring's GFP expression, but was placed at the very bottom of the feature importance graph. In another instance, the another plate appeared as predictive features for offspring's RFP noise residual. The cause of this result is unclear, but it could be due to a batch effect, with specific strains from that plate affecting the results. This issue needs further investigation.

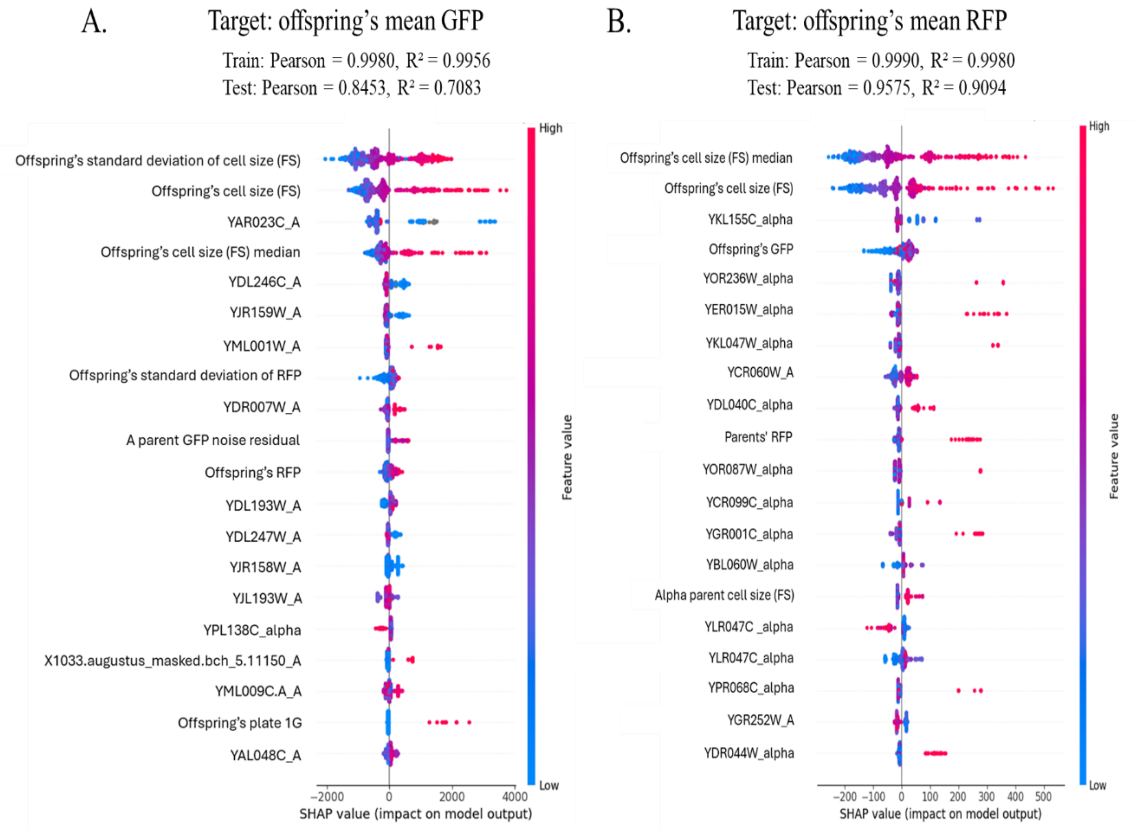


Figure 24 | SHAP plot visualizing the 20 most important features for predicting GFP and RFP values in the offspring. The x-axis represents the feature value, which is the actual value of each feature for a specific observation. This could be the raw value of the gene expression or a measured feature. The SHAP value on the horizontal axis represents how much each feature value contributes to the model's predicted outcome. (A) The top 20 features for predicting offspring's GFP levels, with offspring's cell size and YAR023C_A gene at the top of the list. 11 out of the 12 important genes come from the mating type *A* parent, highlighting a parental effect. The model successfully explained 70.83% of the variance in GFP prediction (Pearson = 0.8453, R^2 = 0.7083). (B) The top 20 features predicting offspring's RFP levels, we again observe a parental effect, with 12 out of the 15 genes coming from the *alpha* parent. The model successfully explained 90.94% of the variance in RFP prediction (Pearson = 0.9575, R^2 = 0.9094). Offspring's cell size emerged as an important feature in predicting both markers.

Note: Genes ending with _A indicate they come from the *A* parent, while those ending with _alpha are from the *alpha* parent. The transcriptomic data regarding gene X1033.augustus_masked was predicted using the AUGUSTUS program.

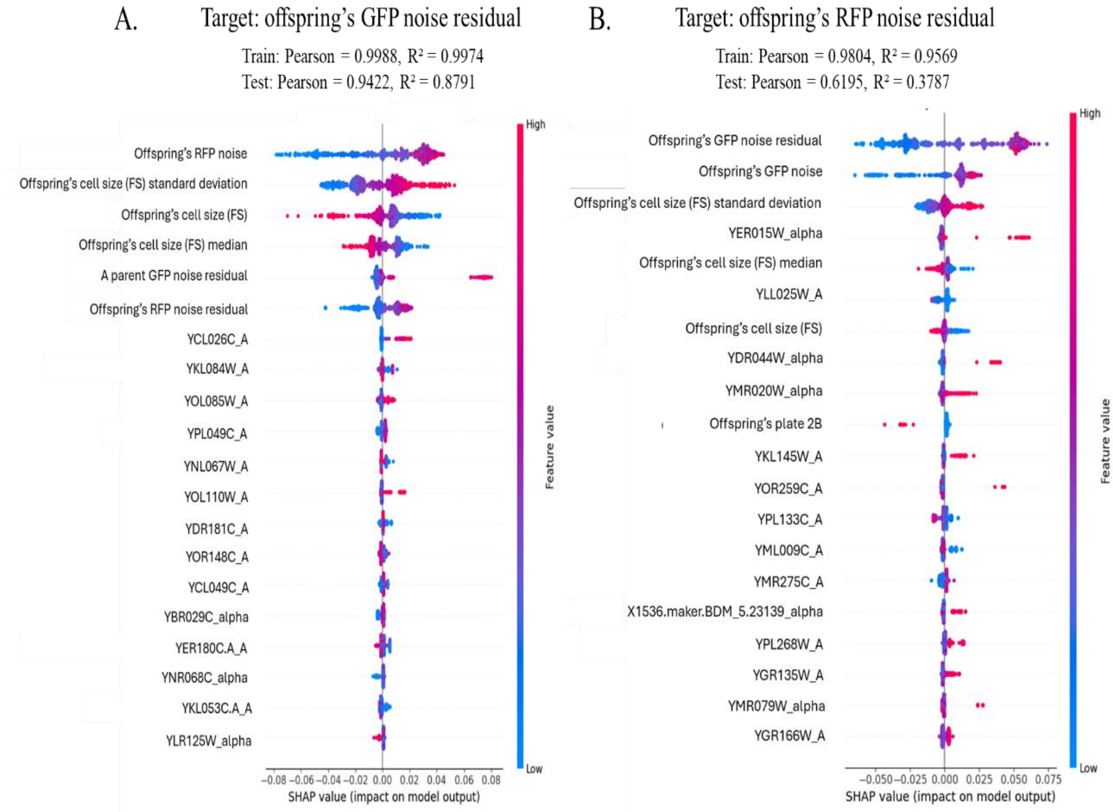


Figure 25 | SHAP plot visualizing the 20 most important features for predicting the noise residuals of GFP and RFP in the offspring. The x-axis represents the feature value, which is the actual value of each feature for a specific observation. This could be the raw value of the gene expression or a measured feature. The SHAP value on the horizontal axis represents how much each feature value contributes to the model's predicted outcome. (A) The top 20 features for predicting offspring's GFP noise residual. (B) The top 20 features for predicting offspring's RFP noise residual. For both targets, the noise residual of one marker was an important feature for predicting the noise residual of the other marker, as well as cell size (forward scatter). Noise residual of mating type *A* parent was a strong predictor for offspring's GFP noise residual, highlighting a strong parental effect. In terms of genes, we observed a strong parental effect for GFP noise residual, with 11 out of 14 genes coming from the mating type *A* parent. However, for RFP noise residual, 9 out of 14 genes came from the *A* parent, which did not contribute to the RFP marker. The model successfully explained 87.91% of the variance in offspring's GFP noise residual prediction (Pearson = 0.9422, R² = 0.8791), and 37.87% of the variance in offspring's RFP noise residual prediction (Pearson = 0.6195, R² = 0.3787).

Note: Genes ending with _A indicate they come from the *A* parent, while those ending with _alpha are from the *alpha* parent. The transcriptomic data regarding gene X1536.maker.BDM was predicted using the MAKER program.

7.9. Measuring Antibiotic Resistance levels

The major aim of my thesis was to quantify inheritance of gene expression level as a quantitative trait. Beyond the fluorescent proteins used here we wanted to take advantage of the fact that our strains were also engineered to carry antibiotics resistance genes, two for each mating type. We hypothesized that if various strains have a different general capacity to encode such foreign genes, then perhaps such variation in expression level will also manifest in corresponding variation in resistance to these drugs, namely strains with a general high capacity for heterologous gene expression will resist the two drugs better. For that I began by measuring for parental strains their resistance to the drugs.

To test this, I first measured the resistance levels of the parental strains by assessing their growth rates under different antibiotic concentrations using optical density (OD) measurements. Growth rates were then compared to control conditions without antibiotics to determine the degree of resistance. Mating type *A* strains were tested for resistance to Hygromycin (Hyg) and Zeocin (Zeo), while mating type *alpha* strains were tested for resistance to Nourseothricin (Nat) and Kanamycin (G418). A significant positive correlation was observed between the growth levels of the strains under both antibiotic conditions, indicating that some strains consistently exhibit higher resistance for more than one drug (Figure 26). Additionally, we examined the relationship between resistance levels and growth rate, and fitness as measured in (15). Since different antibiotics affect cellular processes in distinct ways (as detailed in Table 4, Appendix), we hypothesized that fitness could influence resistance. However, no significant correlation was observed between antibiotic resistance and fitness levels across the strains (Figure 4, Appendix). One possibility to explain the high correlation between the resistance to the two different drugs in each strain is that certain strains express better the two resistance genes (but potentially also other heterologous genes too), while others produce such genes less efficiently. An alternative to this is that certain strains have a better general antibiotic resistance capacity than others, which is independent of their ability to express and utilize the antibiotic resistance genes that they were engineered to encode in their genome. We wished to distinguish between these two possibilities. In particular, to explore the possibility of "innate resistance," e.g., due to factors beyond the introduced resistance genes influencing antibiotic resistance, we tested eight randomly selected strains in their un-engineered versions, lacking the resistance gene construct. Four strains were selected to represent each mating type (*A* and *alpha*). These un-engineered strains, when exposed to baseline antibiotic concentrations, exhibited growth rates similar to those observed in the no-antibiotic condition. These concentrations were equivalent to the baseline levels used for strains containing the resistance gene and are detailed in the Methods section. This observation suggests the possibility of an innate resistance mechanism (Figure 27). From these experiments, we hypothesize that unquantifiable factors, such as the presence of a multidrug resistance (MDR) system, could contribute to antibiotic resistance.

MDR systems, such as efflux pumps in the cell wall, could actively expel antibiotic molecules, explaining the correlation in resistance to two different antibiotics. Indeed, MDR has been documented in the yeast *S. cerevisiae* (24) and could hence be manifested to various levels between strains in this collection. This finding complicates the reliability of antibiotic resistance as a quantitative trait model for studying the capacity to express heterologous genes and their modes of inheritance, as these additional mechanisms introduce variability unrelated to the genetic constructs used in the strains.

In conclusion, our analysis of drug resistance revealed a general correlation across strains, where certain strains exhibited higher resistance to antibiotics, even in the absence of an antibiotic resistance gene. This finding led us to conclude that, unlike the fluorescence assay, drug resistance is unlikely to serve as a reliable measure of the capacity to express a foreign gene. As a result, we have decided not to pursue this direction further in the current thesis.

However, the innate ability of certain strains to resist multiple antibiotics remains a noteworthy observation and may warrant further investigation into the underlying mechanisms of this unexpected resistance.

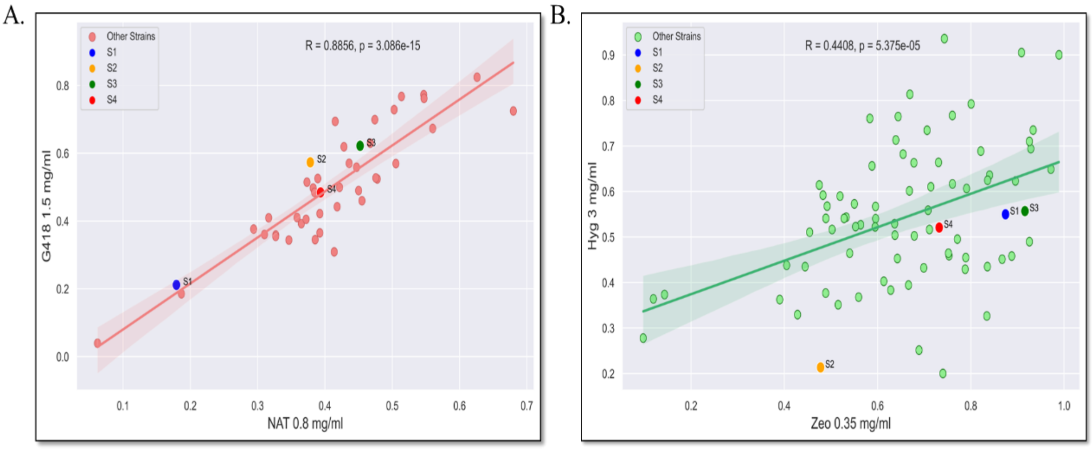


Figure 26 | Scatter plot of OD levels under antibiotic conditions. This figure displays scatter plots of OD levels for strains grown under different antibiotics, averaged across three highly correlated technical replicates. (A) OD levels of 42 *alpha* mating type strains grown under the antibiotics NAT and G418 also exhibit a strong positive correlation (R = 0.8856, p-value < 0.001). (B) OD levels of 77 *A* mating type strains grown under the antibiotics Hyg and Zeo show a positive correlation (R = 0.4408, p-value < 0.001). OD levels were measured while yeast cells were in mid-log phase.

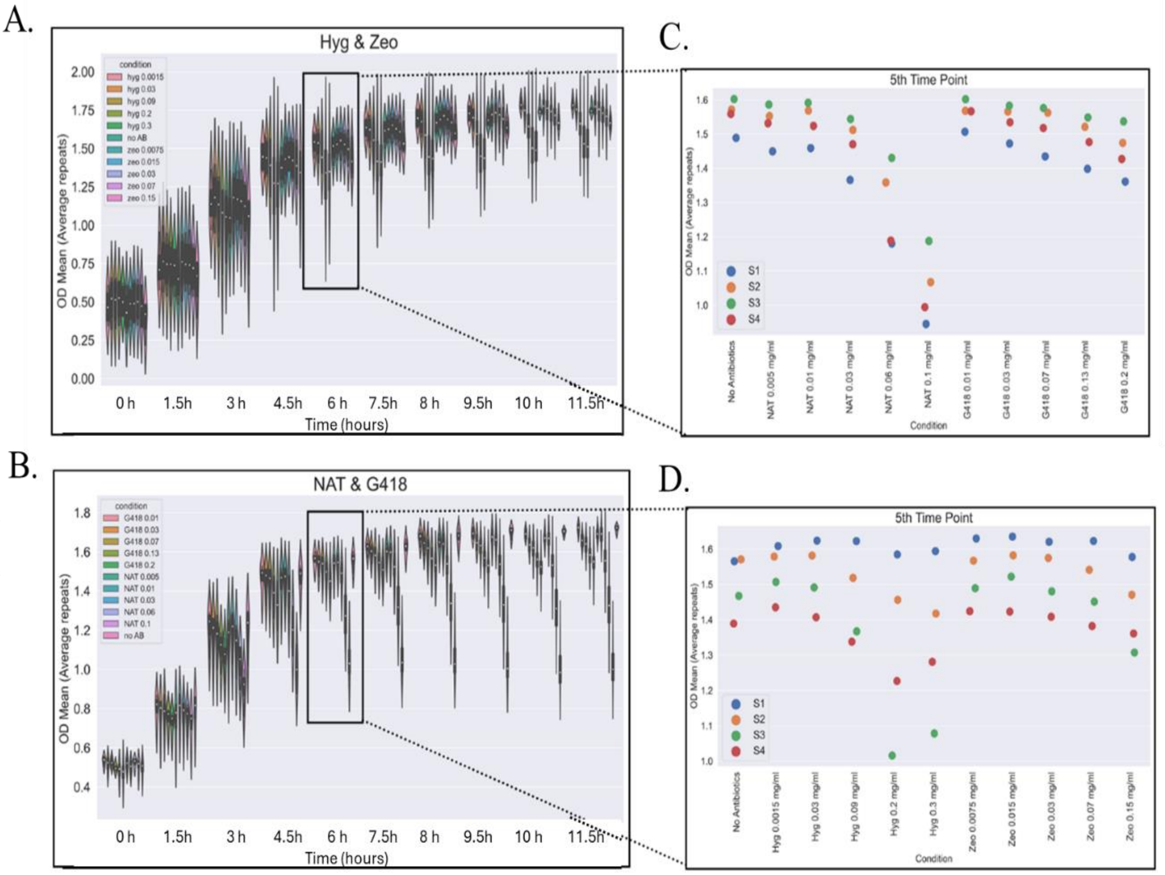


Figure 27 | Growth rate of un-engineered strains. This figure shows the growth levels of four randomly selected strains from the "natural strain collection," measured across three biological replicates. OD levels were recorded at 10 time points, each spaced 1.5 hours apart, to capture growth patterns over time. The high consistency among replicates is demonstrated by Pearson correlation coefficients: (rep 1 vs. rep 2) $r = 0.9643$, $p < 0.001$; (rep 3 vs. rep 2) $r = 0.9651$, $p < 0.001$; and (rep 1 vs. rep 3) $r = 0.9215$, $p < 0.001$. (A) and (B) display violin plots of OD levels (averaged across replicates) across all time points for different concentrations of antibiotics. (A) OD levels under varying concentrations of Hyg and Zeo, (B) illustrates the response to NAT and G418. (C) and (D) provide close-ups of OD levels at time point 5, highlighting the growth for each of the four strains under each condition.

7.10. Western Blot Analysis of Protein Production Efficiency

To assess the efficiency of foreign protein expression in yeast, we first transformed the four selected strains: BY4741, CEN.PK, AKP (α), and BMB (A) with a construct encoding FGF2 under an estrogen-inducible promoter. Following successful integration, we performed Western blot analysis to compare FGF2 protein expression levels across different strains. Additionally, we included two control samples: BY4741 without the construct (negative control) and A positive control expressing a FLAG-tagged protein of size 35kDa. The Western blot revealed specific bands around ~30kDa in three of the four transformed strains (BY4741, CEN.PK, and BMB), while AKP (*alpha*) showed no detectable expression, suggesting that the estrogen-inducible system may not be functioning as expected in this strain. However, an unexpected finding was that our target protein, FGF2 (~17kDa), appeared to migrate at a higher molecular weight (~30kDa) (Figure 28). At this stage, we cannot conclusively determine whether the detected band corresponds to FGF2, and further analysis is required to clarify its identity. Additional experiments, such as mass spectrometry, will be necessary to confirm the presence of FGF2.

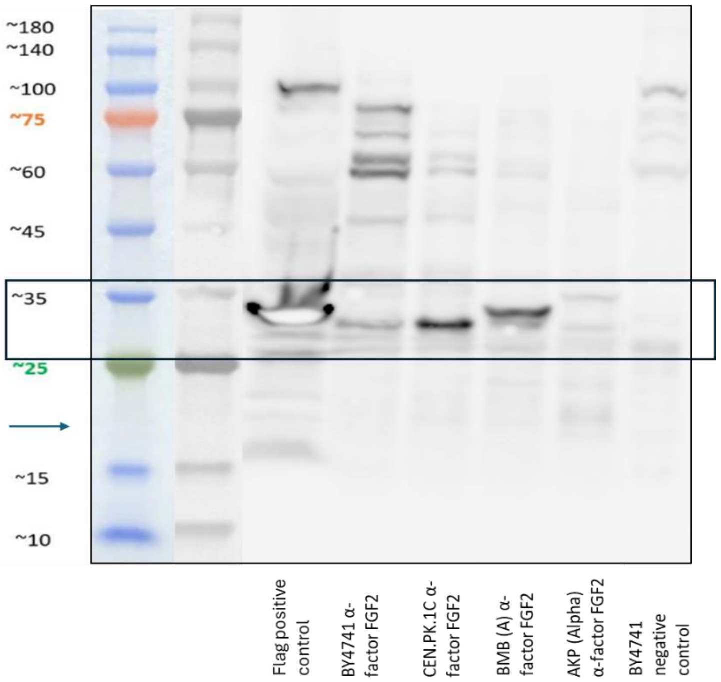


Figure 28 | Western blot analysis of FGF2 expression in yeast strains BY4741, CEN.PK, AKP *(alpha)*, and BMB *(A)* transformed with an estrogen-inducible construct encoding FGF2. A specific band around ~30kDa is observed in three of the four transformed strains (BY4741, CEN.PK, and BMB), whereas AKP (*alpha*) shows no detectable expression, suggesting that the estrogen-inducible system may not be functioning properly in this strain. The expected molecular weight of FGF2 is ~17 kDa, but a higher-than-expected band is detected. A negative control (BY4741 without the construct) and a positive control expressing a FLAG-tagged protein 35kDa were included for reference. Molecular weight markers (left) indicate protein sizes in kDa. The arrow highlights the area where FGF2 expression is expected. And the box highlights the area where we got a specific band. The image was edited using PowerPoint to overlay the molecular weight ladder for clarity.

8. Discussion

In this study, we investigate the complex mechanisms underlying quantitative trait inheritance in *Saccharomyces cerevisiae* using a comprehensive, high-throughput approach applied to a diverse collection of wild yeast strains. Sourced from a wide range of geographical and ecological niches, we examined the inheritance patterns of key quantitative traits, including fluorescent protein expression, cell size, and protein expression noise. By systematically analyzing trait inheritance across over 3000 offspring combinations derived from more than 100 genetically diverse parent strains we provide insights that connect fundamental research with practical applications, particularly in optimizing yeast strains for industrial protein production.

8.1. Insights from Parental Strain Screening

Starting with the screening of haploid strains "the parents", we observed significant variation in fluorescent protein expression across strains of both mating types. This variation was particularly notable when comparing the highest- and lowest-producing strains, which exhibited statistically significant differences in fluorescence intensity. Further comparisons of these extreme strains with the laboratory strain (BY4741) and the industrial strain (CEN.PK) revealed that both the lab and industrial strains showed lower fluorescence intensities for GFP and RFP than even the lowest-performing wild strains from the "natural strain collection". Strikingly, the industrial and laboratory strains consistently ranked at the bottom of the fluorescence intensity spectrum, indicating a wide room for improvement in their capacity for foreign protein production.

An additional analysis comparing the mating types of the strains highlighted the importance of heterozygosity in achieving a correlation between GFP and RFP expression levels across the two mating types. Specifically, homozygous strains exhibited a positive correlation between GFP and RFP expression, suggesting a more stable relationship in expression levels. In contrast, heterozygous strains showed no correlation between the two, reflecting higher variability in expression patterns. These results highlight the critical role of heterozygosity in regulating gene expression consistency between mating types and provide insights into how genetic diversity impacts trait expression.

8.2. High-Throughput Screening of Offspring for Trait Inheritance Patterns

To uncover trait inheritance patterns, we generated over 3,000 offspring combinations using two distinct high-throughput approaches: "all-against-all" and "one-against-one." The results of these approaches were complementary, as each provided unique insights into inheritance mechanisms. In the “all-against-all” approach, we utilized FACS sorting to isolate offspring with extreme fluorescence intensities. Meanwhile, the “one-against-one” approach employed FACS screening to systematically measure fluorescence levels across all possible offspring combinations. Together, these approaches highlighted strains such as AKP (*alpha*) and BHB (*alpha*), which consistently produced offspring with unique and interesting inheritance patterns.

8.3. Exploring Inheritance Patterns Between Parents and Offspring

An initial comparison between parents and offspring revealed an unexpected result: instead of observing increased expression levels in diploid offspring, as might be expected due to their larger size and dual gene copies, we found the opposite. Parental strains in both mating types displayed higher expression levels of GFP (mating type *A*) and RFP (mating type *alpha*), with approximately a two-fold difference compared to their diploid offspring. One possible explanation for this phenomenon lies in promoter competition. In the haploid parents, the promoters driving GFP and RFP expression operate independently within their respective genomes. Upon mating, the diploid inherits both constructs, potentially leading to competition for transcriptional resources between the two markers, thereby reducing their overall activity. However, since the promoters driving GFP and RFP are not identical, this competition cannot simply be attributed to shared transcription factors. Instead, it may involve a broader limitation in the availability of transcriptional machinery or other regulatory factors. Furthermore, the regulatory environment in the diploid state may introduce crosstalk or epigenetic modifications that reset or weaken promoter activity, further contributing to the observed reduction. An alternative explanation involves what we term the "distant promoter effect" In the haploid parents, GFP and RFP are driven by constitutive promoters; however, these strains also contain haploid-specific promoters, for Zeo in mating type A and for G418 in mating type alpha, that not only regulate antibiotic resistance genes but may also indirectly enhance the transcription of GFP and RFP. In the diploid offspring, these haploid-specific promoters are absent, likely leading to reduced transcriptional activity of GFP and RFP. This suggests that the haploid-specific environment plays a critical role in the observed expression levels, which is lost in the diploid state. Despite these potential explanations, we currently lack a definitive understanding of this unusual phenomenon. The observed reduction in GFP and RFP expression in diploid offspring raises several intriguing hypotheses regarding the interplay between promoter activity, transcriptional machinery, and cellular regulatory environments. Further research is needed to investigate these mechanisms in greater depth.

Another analysis was conducted aimed to determine whether the offspring's expression levels could be predicted by averaging the expression levels of the parents. This analysis revealed a very weak positive correlation, suggesting that the mechanism governing trait inheritance is far more complex than simple parental averaging.

8.4. Parental and Non-Parental Inheritance Theory

In this study, we identified an intriguing phenomenon that we term **parental and non-parental inheritance.** In cases of parental inheritance, the parent contributing a specific fluorescent marker directly influences its expression level in the offspring. This pattern is observed consistently across offspring derived from the same parent. Conversely, non-parental inheritance refers to instances where the parent not directly contributing a given fluorescent marker influences its expression in the offspring. For example, the mating type *alpha* parent, which contributes RFP, systematically affects the GFP expression levels in its offspring.

A similar effect has been observed in *Drosophila* male sterility, where mothers transmit cytoplasmic factors that influence male fertility without passing the corresponding nuclear genes (25). This results in the maternal inheritance of a male-specific trait, demonstrating how a parent's contribution can shape an offspring's phenotype without directly transmitting the associated genetic determinant. This parallel suggests that, as in *Drosophila*, additional regulatory mechanisms beyond direct genetic inheritance may be influencing fluorescence expression in yeast.

Interestingly, we also observed cases where offspring exhibited higher gene expression levels than their parents. This phenomenon, where sexual mating resulted in offspring surpassing parental expression levels, highlights the potential for using sexual mating as a strategy to improve strains. Such observations suggest that the regulatory network within the diploid state introduces additional layers of complexity, likely involving epistatic interactions or crosstalk between genetic elements. Epistasis, defined as the interaction between genes where the effect of one gene is modified by one or more other genes (26), provides a potential explanation for these findings. Such interactions significantly influence phenotypic expression and complicate the relationship between genotype and phenotype. In the context of our study, the observed non-parental inheritance patterns, where a parent not directly contributing a specific fluorescent marker influences its expression in the offspring, may be attributed to epistatic interactions. In *Saccharomyces cerevisiae*, epistasis has been shown to play a crucial role in various genetic pathways (27). The results of this study suggest that we may be observing epistatic interactions in action, highlighting the complexity of gene regulation within diploid states.

8.5. Predicting Expression Levels Using Transcriptomic Data

To further investigate the mechanisms influencing fluorescent protein inheritance, we initiated a machine learning project incorporating gene expression profiles of both parents and offspring, along with additional features such as genetic distance, fitness, cell size, and noise. The goal was to identify key predictors of GFP and RFP expression levels in offspring, as well as their noise residuals. Our results revealed a strong parental effect on both GFP and RFP expression levels, with most top predictive genes originating from the mating type contributing the corresponding fluorescent marker. Offspring cell size consistently emerged as a significant predictor for both expression levels and noise residuals, reinforcing the observed correlation between cell size and fluorescent intensity. Interestingly, many of the highly ranked predictive genes had no identified function, highlighting the potential for novel discoveries. Additionally, some genes such as sorbitol dehydrogenase genes were found to be predictive without an obvious biological connection to fluorescent protein expression. These findings underscore the complex interplay between genetic factors, cellular characteristics, and parental contributions in shaping fluorescent protein expression and variability in offspring. This analysis provides a foundation for further exploration of gene-specific contributions and the regulatory mechanisms underlying inherited traits.

8.6. Heritability of Noise in Protein Expression Upon Sexual Mating

In our investigation of protein expression noise as a quantitative trait, we observed a negative correlation between noise and mean fluorescence across various strains in both mating types and their offspring. This aligns with previous findings from our lab, which demonstrated that low-abundance proteins exhibit higher noise levels, while high-abundance proteins show lower noise, following a predictable scaling law (10). Notably, certain strains deviated from the expected noise-to-mean fluorescence relationship. These deviations were consistently inherited by their offspring, suggesting a heritable component to expression noise.

To further explore this, we conducted a noise residual analysis, comparing the residuals of parents and their offspring. The analysis revealed a positive correlation between parental and offspring noise residuals in both fluorescent markers, indicating that noise is a heritable quantitative trait. In addition, our finding that the average parental noise residual could predict offspring noise residual strengthens the case that noise is heritable upon sexual mating. This suggests that parental contributions to offspring noise are not random but are systematically influenced by quantitative genetic factors. Additionally, the observation that the maximum parental noise residual is a strong predictor of offspring GFP noise highlights a potential dominance effect in noise heritability. Specifically, if one parent exhibits high noise in protein expression, this trait is likely to be inherited by the offspring regardless of the partner’s noise level. Conversely, if a parent is not noisy, its influence on the offspring's noise levels is relatively weak. This suggests an asymmetry in how parental noise traits contribute to offspring noise, pointing to a potential dominance mechanism at play. This observation is supported by previous research demonstrating that natural sequence variants can influence cell-to-cell expression variability. For example, specific genetic loci in yeast have been identified as contributors to elevated variability in gene expression (28). Additionally, regulatory mechanisms, such as methyltransferase Hmt1, have been shown to play a role in buffering gene expression noise (29). These findings underscore the genetic basis of expression noise and highlight how both genetic and regulatory factors can modulate variability. Collectively, these studies suggest that noise in protein expression is not merely a stochastic phenomenon but may also be inherited across generations upon sexual mating, contributing to phenotypic diversity within populations.

8.7. The Inheritance of Cell Size

Cell size was examined as a quantitative trait using data extracted from the FACS analysis, where forward scatter (FSC) was used as a proxy for cell size. Analyzing sibling groups revealed similarities in their forward scatter values, indicating that cell size is a heritable trait passed down from the parents.

To further investigate this, we examined the correlation between the cell size of parents and the average cell size of their offspring. A positive correlation was observed in both mating types, reinforcing the heritability of cell size. Additionally, we found a positive polynomial correlation between cell size and fluorescent protein production. Interestingly, the offspring from one specific strain, AKP (*alpha*), deviated from this trend, forming a distinct polynomial fit. These offspring exhibited higher fluorescence levels irrespective of their cell size, in contrast to the rest of the population. This observation suggests that while cell size and fluorescence production are generally correlated, certain genetic or regulatory factors may affect these traits in specific strains.

8.8. Summary

To summarize, this study provides an integrated framework for understanding the inheritance of quantitative traits in *Saccharomyces cerevisiae*. By examining fluorescent protein expression, cell size, and protein expression noise, we revealed distinct patterns of inheritance and regulatory complexity. Each trait highlighted a unique aspect of quantitative inheritance: fluorescence demonstrated the influence of parental and non-parental factors; noise revealed heritable variability shaped by genetic and regulatory mechanisms; and cell size uncovered correlations with other traits, such as protein production. Together, these findings demonstrate how diverse traits can be interconnected through shared genetic and regulatory pathways. Understanding these relationships not only advances our knowledge of quantitative inheritance but also lays the foundation for developing strategies to optimize strains for specific applications. Nonetheless, several challenges and limitations remain. Environmental variability, including differences in growth conditions and nutrient availability, may have influenced trait expression and inheritance patterns, introducing variability into the results. Additionally, the regulatory complexity observed in diploid states, such as epistatic interactions and crosstalk between genetic elements, requires further investigation to fully elucidate the underlying mechanisms. Although we conducted high-throughput analyses, the scalability of this approach may be constrained when applied to more complex traits or larger, more diverse datasets. Addressing these challenges in future studies will be critical for refining our understanding of quantitative inheritance and unlocking its full potential for practical applications.

8.8. Future Directions

Building on the insights gained from this analysis, a promising future direction involves identifying strains with high expression levels that could be utilized in industrial applications. The strains identified through this study can serve as candidates for optimizing foreign protein production, either through direct application or further genetic engineering and evolution. Additionally, the potential of using sexual mating as a method to improve strains, akin to selective breeding, holds great promise. By leveraging the observed inheritance patterns and expression variability, we can design mating strategies to enhance desirable traits, such as high protein expression. To explore this approach, we selected two candidate strains from the "natural strain collection", AKP (mating type *alpha*) and BMB (mating type *A*). These strains were engineered to express foreign genes commonly used in industrial applications: FGF2 (fibroblast growth factor 2) and albumin. We conducted experiments where these genes were successfully inserted into the chosen strains. Protein production efficiency was then quantitatively analyzed using Western blotting, comparing the expression levels of FGF2 in these engineered strains against both the laboratory strain (BY4741) and the industrial strain (CEN.PK). At this stage, the results remain preliminary and inconclusive. While we detected the expression of a FLAG-tagged protein, its observed molecular weight did not match the expected size of FGF2 (~17 kDa). Instead, a higher-than-expected band (~30 kDa) was observed, suggesting potential post-translational modifications, dimerization, or unexpected interactions that may be altering the apparent molecular weight of FGF2. Further analysis is required to confirm the identity of the observed band and determine whether it corresponds to properly expressed and secreted FGF2.

To address this uncertainty, additional experiments will be necessary. Mass spectrometry will be used to verify the presence of FGF2 and identify potential modifications affecting its molecular weight, including partial cleavage of the α-factor signal peptide. Additionally, enzymatic deglycosylation assays could help determine whether glycosylation contributes to the observed size discrepancy. Once the identity of the detected band is confirmed, we will proceed with quantitative assays to accurately compare expression levels between the different strains and assess their potential for industrial applications.

9. Acknowledgments

Despite the challenges, the past two years have been an incredible journey of exploring the fascinating world of yeast and molecular biology. First and foremost, I want to express my deepest gratitude to Prof. Yitzhak Pilpel for his invaluable knowledge, guidance, and support. I will always cherish our stimulating conversations about science and life, as well as your encouragement to challenge myself and grow as a scientist. I am also deeply grateful to Dr. Orna Dahan, who generously shared her expertise and patiently taught me the foundations of working with the yeast. Her invaluable advice and unwavering support were instrumental in bringing this project to life. Additionally, I want to thank Noa Hefetz-Aharon for her significant contributions to this project. Her guidance, patience, and encouragement helped me navigate the most challenging moments of my master's journey. I am especially thankful for her mentorship in coding and data analysis—skills that will stay with me for life. Special thanks to Bar Cohen for contributing to the machine learning project. I would also like to thank all my lab members for their invaluable insight and support, especially Shu-Ting You, whose advice was a constant source of help and encouragement. I extend my gratitude to the core facilities at Weizmann and the bioinformatic unit for their contributions to this work.

Finally, I wish to express my heartfelt appreciation to my friends and family, especially my mom, dad, and my partner, Trivko for their emotional support.

10. Literature

1. Falconer D.S., Mackay T.F.C. (1996) *Introduction to Quantitative Genetics*. 4th Edition. Pearson Prentice Hall, Harlow, England.
2. Müller B., Grossniklaus U. (2010) Model organisms — A historical perspective. *Journal of Proteomics* 73(11), 2054–2063, doi:10.1016/j.jprot.2010.05.003.
3. Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., et al. (1996) Life with 6000 genes. *Science* 274(5287), 546–567, doi:10.1126/science.274.5287.546.
4. Parapouli M., Vasileiadi A., Afendra A.S., Hatziloukas E. (2020) *Saccharomyces cerevisiae* and its industrial applications. *AIMS Microbiology* 6(1), 1–32, doi:10.3934/microbiol.2020.1.1.
5. Cregg J.M., Cereghino J.L., Shi J., Higgins D.R. (2000) Recombinant protein expression in *Pichia pastoris*. *Molecular Biotechnology* 16(1), 23–52, doi:10.1385/MB:16:1:23.
6. Herskowitz I. (1988) Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiological Reviews* 52(4), 536–553.
7. Sherman F. (1991) Getting started with yeast. *Methods in Enzymology* 194, 3–21, doi:10.1016/0076-6879(91)94004-V.
8. Peter J., De Chiara M., Friedrich A., Yue J.X., Pflieger D., Bergström A., et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 556(7701), 339–344, doi:10.1038/s41586-018-0030-5.
9. Cubillos F.A., Louis E.J., Liti G. (2009) Generation of a large set of genetically tractable haploid and diploid *Saccharomyces* strains. *FEMS Yeast Research* 9(8), 1217–1225, doi:10.1111/j.1567-1364.2009.00583.x.
10. Bar-Even A., Paulsson J., Maheshri N., Carmi M., O’Shea E., Pilpel Y., et al. (2006) Noise in protein expression scales with natural protein abundance. *Nature Genetics* 38(6), 636–643, doi:10.1038/ng1807.
11. Herzenberg L.A., Parks D., Sahaf B., Perez O., Roederer M., Herzenberg L.A. (2002) The history and future of the fluorescence-activated cell sorter and flow cytometry: A view from Stanford. *Clinical Chemistry* 48(10), 1819–1827.
12. Wersto R.P., Chrest F.J., Leary J.F., Morris C., Stetler-Stevenson M., Gabrielson E. (2001) Doublet discrimination in DNA cell-cycle analysis. *Cytometry* 46(5), 296–306, doi:10.1002/cyto.1171.
13. Nolan J.P., Sklar L.A. (1998) The emergence of flow cytometry for sensitive, real-time measurements of molecular interactions. *Nature Biotechnology* 16(7), 633–638, doi:10.1038/nbt0798-633.

14. Caudal É., Loegler V., Dutreux F., Vakirlis N., Teyssonnière É., Caradec C., et al. (2024) Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nature Genetics* 56(6), 1278–1287, doi:10.1038/s41588-024-01405-3.
15. Strauss S.K., Golomb R., Sheykhkarimli D., Liti G., Dahan O., Pilpel Y. (2024) Quantitative genetics of natural *Saccharomyces cerevisiae* strains upon sexual mating reveals heritable determinants of cellular fitness. *Under Review*.
16. Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* 17(1), 10, doi:10.14806/ej.17.1.200.
17. Langmead B., Salzberg S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4), 357–359, doi:10.1038/nmeth.1923.
18. McIsaac R.S., Oakes B.L., Wang X., Dummit K.A., Botstein D., Noyes M.B. (2013) Synthetic gene expression perturbation systems with rapid, tunable, single-gene specificity in yeast. *Nucleic Acids Research* 41(4), e57, doi:10.1093/nar/gks1313.
19. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
20. Chen, T., Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, doi:10.1145/2939672.2939785.
21. Chen, T., Guestrin, C. (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, doi:10.1145/2939672.2939785.
22. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018) CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* 31, 6638–6648.
23. Cherry J.M., Hong E.L., Amundsen C., Balakrishnan R., Binkley G., Chan E.T., Christie K.R., Costanzo M.C., Dwight S.S., Engel S.R., Fisk D.G., Hirschman J.E., Hitz B.C., Karra K., Krieger C.J., Miyasato S.R., Nash R.S., Park J., Skrzypek M.S., Simison M., Weng S., Wong E.D. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* 40(D1): D700–D705. DOI: 10.1093/nar/gkr1029.
24. Guinn L, Lo E, Balázsi G. Drug-dependent growth curve reshaping reveals mechanisms of antifungal resistance in *Saccharomyces cerevisiae*. Commun Biol. 2022 Mar 31;5(1):292. doi:10.1038/s42003-022-03228-9.
25. Hoffmann AA, Turelli M, Harshman LG. Factors affecting the distribution of cytoplasmic incompatibility in *Drosophila simulans*. Genetics. 1990 Dec 1;126(4):933–48.

26. Phillips, P.C. (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* 9(11), 855–867, doi:10.1038/nrg2452.
27. Lai-Zhang, J. (1999) Epistatic interactions of deletion mutants in the genes encoding the F1-ATPase in yeast *Saccharomyces cerevisiae*. *The EMBO Journal* 18(1), 58–64, doi:10.1093/emboj/18.1.58.
28. Fehrmann, S., Bottin-Duplus, H., Leonidou, A., Mollereau, E., Barthelaix, A., Wei, W., et al. (2013) Natural sequence variants of yeast environmental sensors confer cell-to-cell expression variability. *Molecular Systems Biology* 9(1), doi:10.1038/msb.2013.53.
29. You, S.T., Jhou, Y.T., Kao, C.F., Leu, J.Y. (2019) Experimental evolution reveals a general role for the methyltransferase Hmt1 in noise buffering. *PLoS Biology* 17(10), e3000433, doi: 10.1371/journal.pbio.3000433.

11. Appendix

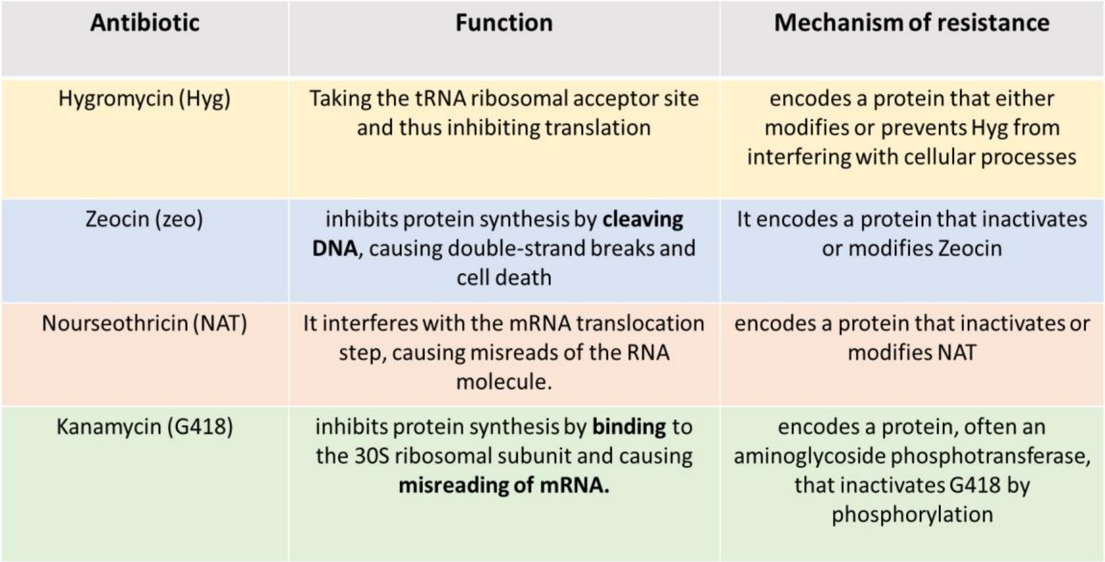
Table 1. List of primers.

ID#	Primer name	Used for	Sequence
1	BFG_K_matA_Rev	Mat A specific primer for barcode identification	CTTGACTGAGCGACTGAGG
2	BFG_H_matA_Fw	Mat A specific primer for barcode identification	CCATACGAGCACATTACGGG
3	BFG_F_alpha_Rev	Mat alpha specific primer for barcode identification	CAGCGGGATAGTGCGATTG
4	BFG_B_alpha_Fw	1. Mat alpha specific primer for barcode identification. 2. Offspring primer for fused barcode identification.	CAGCGGGATAGTGCGATTG
5	BFG_I_matA_Rev	Offspring primer for fused barcode identification.	GTTATCAGAGGTATGCGAGTTAG
6	Illumina_i5_F	Library preparation for barcode sequencing	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
7	Illumina_i7_R	Library preparation for barcode sequencing	CAAGCAGAAGACGGCATACGAGAT [index]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
8	AMP rev	Testing recombination success following transformation	CTGAGAATAGTGTATGCGGCGAC
9	AMP for	Validating recombination success following transformation	CTCACCCAGAAACGCTGGTG

Table 2. List of PCR reactions.

ID#	PCR reaction	Primers used	PCR program	Used for
1	Kapa amplification: Kapa readyMix x2: 12.5 ul F primer: 0.75 ul R primer: 0.75 ul DNA: 2ul DDW: complete volume to 25ul	4-5	1. 95C^{O} 3 min 2. 98 C^{O} 20 sec 3. 60 C^{O} 15 sec 4. 72 C^{O} 15 sec 5. Repeat steps 2-4 for 20 cycles 72 C^{O} 1 min	Barcode regions amplification
2	Kapa amplification: Kapa readyMix x2: 12.5 ul F primer: 0.75 ul R primer: 0.75 ul DNA: 2ul DDW: complete volume to 25ul	6-7	1. 95C^{O} 3 min 2. 98 C^{O} 20 sec 3. 62 C^{O} 15 sec 4. 72 C^{O} 15 sec 5. Repeat steps 2-4 for 20 cycles 72 C^{O} 1 min	Inserting Illumina indexes
3	Kapa amplification: Kapa readyMix x2: 5 ul F primer: 0.3 ul R primer: 0.3 ul DNA: 2ul DDW: complete volume to 10ul	8-9	1. 95C^{O} 5 min 2. 98 C^{O} 20 sec 3. 60 C^{O} 15 sec 4. 72 C^{O} 15 sec 5. Repeat steps 2-4 for 35 cycles 72 C^{O} 1 min	

Table 3. Summary of Antibiotics Used in the Study. This table provides an overview of the antibiotics used in this study, including their primary functions and the mechanisms of resistance introduced into the yeast strains.



Antibiotic	Function	Mechanism of resistance
Hygromycin (Hyg)	Taking the tRNA ribosomal acceptor site and thus inhibiting translation	encodes a protein that either modifies or prevents Hyg from interfering with cellular processes
Zeocin (zeo)	inhibits protein synthesis by **cleaving DNA**, causing double-strand breaks and cell death	It encodes a protein that inactivates or modifies Zeocin
Nourseothricin (NAT)	It interferes with the mRNA translocation step, causing misreads of the RNA molecule.	encodes a protein that inactivates or modifies NAT
Kanamycin (G418)	inhibits protein synthesis by **binding** to the 30S ribosomal subunit and causing **misreading of mRNA.**	encodes a protein, often an aminoglycoside phosphotransferase, that inactivates G418 by phosphorylation

Table 4. List of randomly selected strains used for validating parental identities in the "one-against-one" mating experiment. This table shows the location of each strain, the sequencing results, and whether they matched the expected parental strains. Strains with a green highlight indicate true matches to the expected parents, red indicates contamination, and yellow indicates inconclusive results where one of the parents could not be confidently identified.

Sample Number	Plate	Well	Parent A	Match (True/False)	Parent Alpha	Match (True/False)
1	1A	2F	SACE-YCM	TRUE	AKV	TRUE
2	1B	2F	BHC	TRUE	Inconclusive	
3	1C	2F	AKG	TRUE	Inconclusive	
4	1C	6G	BRD	TRUE	APV	FALSE
5	1D	2F	AKV	TRUE	AKV	TRUE
6	1D	6F	BNP	TRUE	AKV	TRUE
7	1D	7G	BQR	TRUE	BBI	TRUE
8	1E	2F	CFE	TRUE	AKV	TRUE
9	1F	11G	BSB	FALSE	APV	FALSE
10	1F	11H	Inconclusive		APV	TRUE
	1F	2F	BHV	TRUE	AKV	TRUE
11	1F	8H	AST	TRUE	APV	TRUE
13	1G	2F	CEQ	TRUE	AKV	TRUE
14	1G	1A	CPK	TRUE	AKB	TRUE
15	1G	2B	CEQ	TRUE	AKP	TRUE
16	1G	3E	BTN2	TRUE	AIM	TRUE
17	1G	4F	CCG	TRUE	AKV	TRUE
18	1G	5H	BES	TRUE	APV	TRUE
19	2A	2F	SACE-YCM	TRUE	BHB	TRUE
20	2A	3C	BIM	TRUE	BLA	TRUE
21	2B	2F	BHC	TRUE	BHB	TRUE
22	2C	2F	AKG	TRUE	BHB	TRUE
23	2D	2F	AKV	TRUE	BHB	TRUE
24	2D	4C	BHB	TRUE	BLA	TRUE
25	2E	2F	CFE	TRUE	BHB	TRUE
26	2F	2F	BHV	TRUE	BHB	TRUE
27	2F	6C	BRA	TRUE	BLA	TRUE
28	2G	2F	CEQ	TRUE	BHB	TRUE
29	3A	2F	SACE-YCM	TRUE	BQG	TRUE
30	3B	2F	BHC	TRUE	BQG	TRUE
31	3C	2F	AKG	TRUE	BQG	TRUE
32	3D	2F	AKV	TRUE	BQG	TRUE
33	3E	2F	CFE	TRUE	BQG	TRUE
34	3E	9D	CAE	TRUE	BRD	FALSE
35	3F	2F	Inconclusive		BNI	FALSE
36	3F	3E	CQG	TRUE	BNI	TRUE
37	3F	3G	CQG	TRUE	BNK	TRUE
38	3F	7D	BTH	TRUE	BLL	TRUE
39	3G	2F	CEQ	TRUE	BQG	TRUE
40	4A	2F	SACE-YCM	TRUE	BLP	TRUE
41	4B	1B	BKE	TRUE	BRA	TRUE
42	4B	2D	BHC	TRUE	BQI	TRUE
43	4C	2F	AKG	TRUE	BLP	TRUE
44	4D	5E	BNK	TRUE	BRB	TRUE
45	4D	6D	BQR	FALSE	BQI	TRUE
46	4D	7E	BQR	TRUE	BRB	TRUE
47	4E	11F	CGL	TRUE	BLP	TRUE
48	4E	11H	SACE-YAB	FALSE	BRB	FALSE

49	4E	9E	CAE	TRUE	BRB	TRUE
50	4E	9H	CAE	TRUE	CQD	TRUE
51	4F	2F	BHV	TRUE	BLP	TRUE
52	4G	2F	Inconclusive		BLP	TRUE
53	5A	2F	SACE-YCM	TRUE	BQQ	TRUE
53	5B	2F	BHC	TRUE	BQQ	TRUE
54	5C	10A	BCB	FALSE	BLV	FALSE
55	5C	2F	AKG	TRUE	BQQ	TRUE
56	5C	9A	BCB	TRUE	BRD	FALSE
57	5C	9B	BIA	FALSE	BQQ	FALSE
58	5D	2F	AKV	TRUE	BQQ	TRUE
59	5E	2F	Inconclusive		BQQ	TRUE
60	5E	3F	SACE-YBW	TRUE	BQQ	TRUE
61	5F	2F	BHV	TRUE	BQQ	TRUE
62	5F	6A	BRA	TRUE	BPT	TRUE
63	5F	7B	BTH	TRUE	BRD	TRUE
64	5G	2F	CEQ	TRUE	BQQ	TRUE
65	plate 10	6B	BQC	TRUE	BLL	TRUE
66	plate 7	12D	AKI	TRUE	AHP	TRUE
67	plate 7	7G	AKI	TRUE	CCV	TRUE
68	plate 9	2D	BMA	TRUE	AKV	TRUE

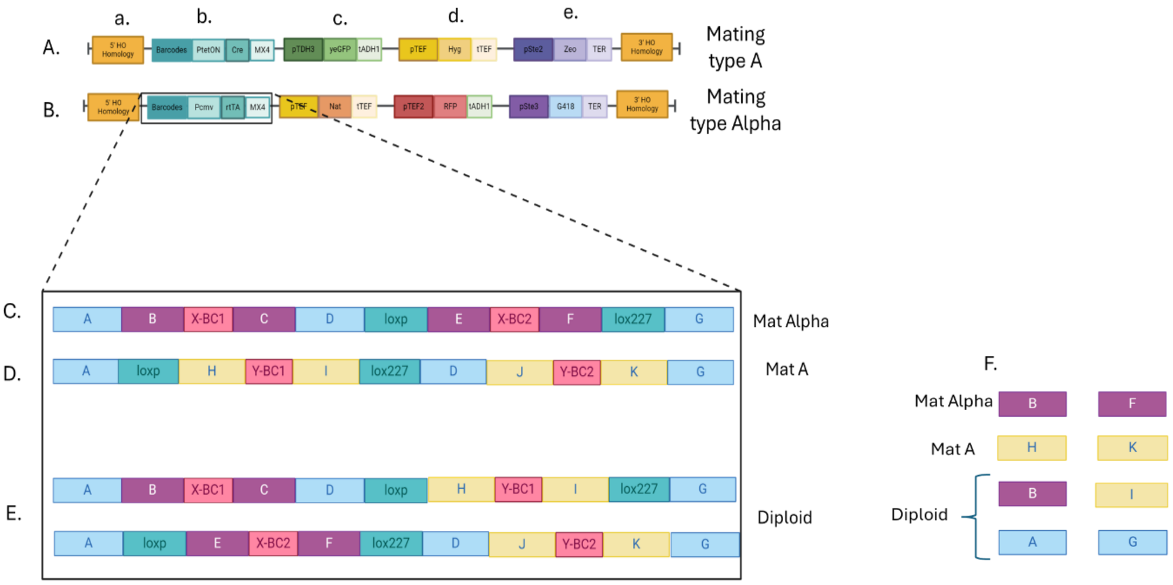


Figure 1 | Construct map that was inserted into the strains. (A) construct design for creating mating type A strains, (B) construct design for creating mating type alpha strains (a) HO homology region that was used for homologous recombination to the HO locus after transformation (b) The Barcode Fusion Genetics system. Composed of the barcodes and lox sequences and the Tet-ON system. Mating type A strains contain the Cre enzyme that is regulated by the rtTA inducer that is found on the mating type alpha design. (c) Constitutive markers. Mating type A contains GFP and Hyg resistance cassettes, while mating type alpha contains RFP and NAT resistance cassette (d) mating type specific markers. Mating type *A* cells induce the BleoR resistant markers and thus can grow on Zeocin, while mating type *Alpha* strains induce the KanMX resistance marker and thus can grow on Geneticin (G418). (C)-(E) each letter corresponds to ~25nt length sequence. Same letter in different panels means that the same sequence is present in both regions. (C) mating type Alpha design, (D) mating type A design, (E) the two fragments created in the offspring after fusion of the barcodes. Different combinations of sequences can be used to amplify specific regions only for NGS.

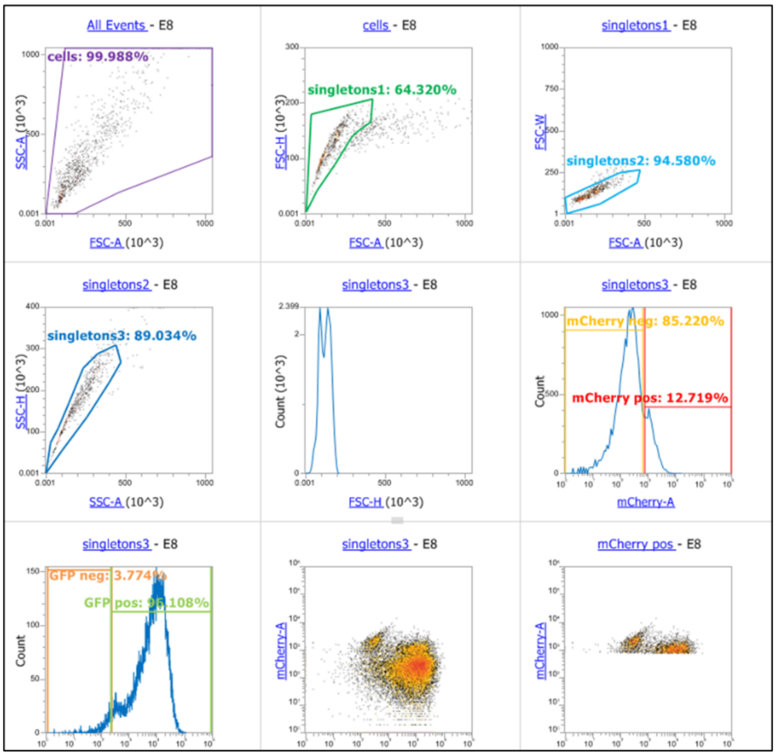


Figure 2 | FACS gating strategy used to measure GFP and RFP fluorescence levels. Forward scatter (FSC) and side scatter (SSC) were used to exclude debris and identify the main cell population. Additional gating removed doublets (singleton gates), and fluorescence intensity for GFP and RFP was measured in the single-cell population.

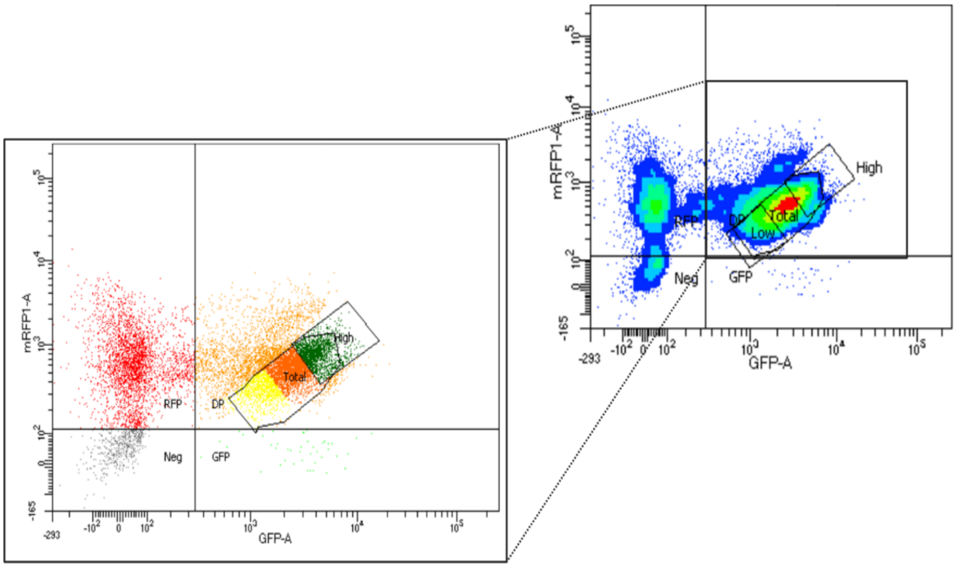


Figure 3 | FACS Sorting of Double-Positive Offspring. This figure shows the results of FACS sorting, where offspring were analyzed based on GFP and RFP fluorescence intensity. Approximately 85% of the population was classified as double-positive, exhibiting both GFP and RFP signals. Within this double-positive population, three groups were sorted: the top 10% with the highest fluorescence intensity, labeled as "High" (green); the bottom 10% with the lowest fluorescence intensity, labeled as "Low" (yellow); and fraction of the double-positive population, labeled as "Total" (orange).

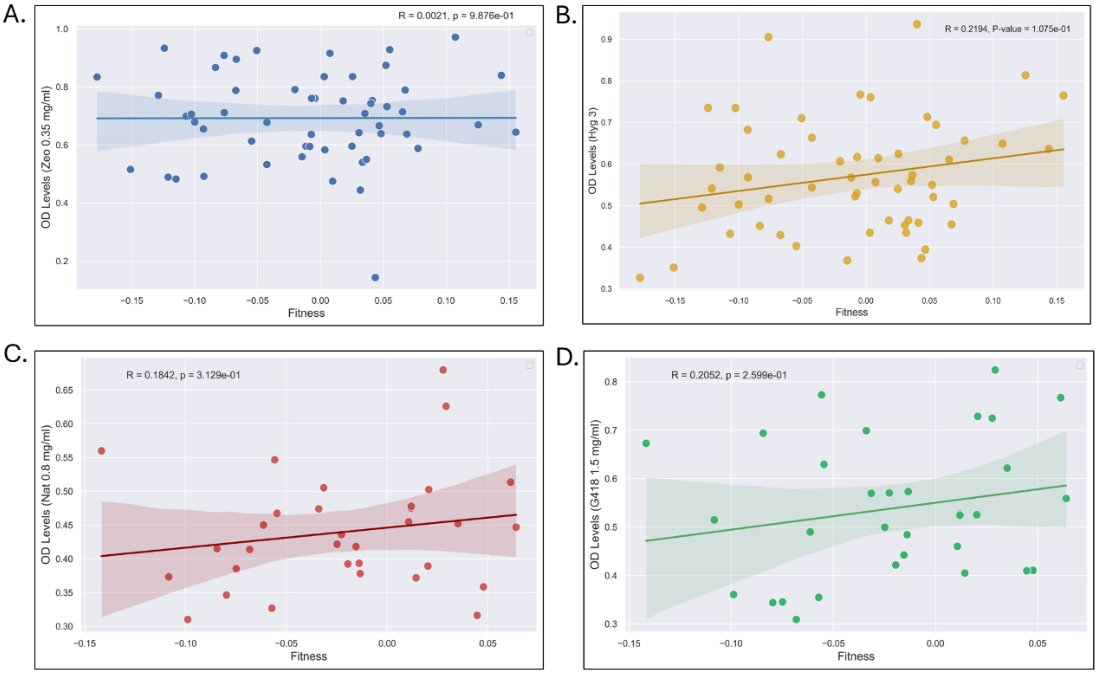


Figure 4 | Lack of Correlation Between Fitness and Antibiotic Resistance. This figure illustrates the absence of significant correlations between fitness and antibiotic resistance across four different antibiotics, as measured by optical density (OD) levels. (A) Zeocin (Zeo), (B) Hygromycin (Hyg), (C) Nourseothricin (Nat), and (D) Kanamycin (G418).

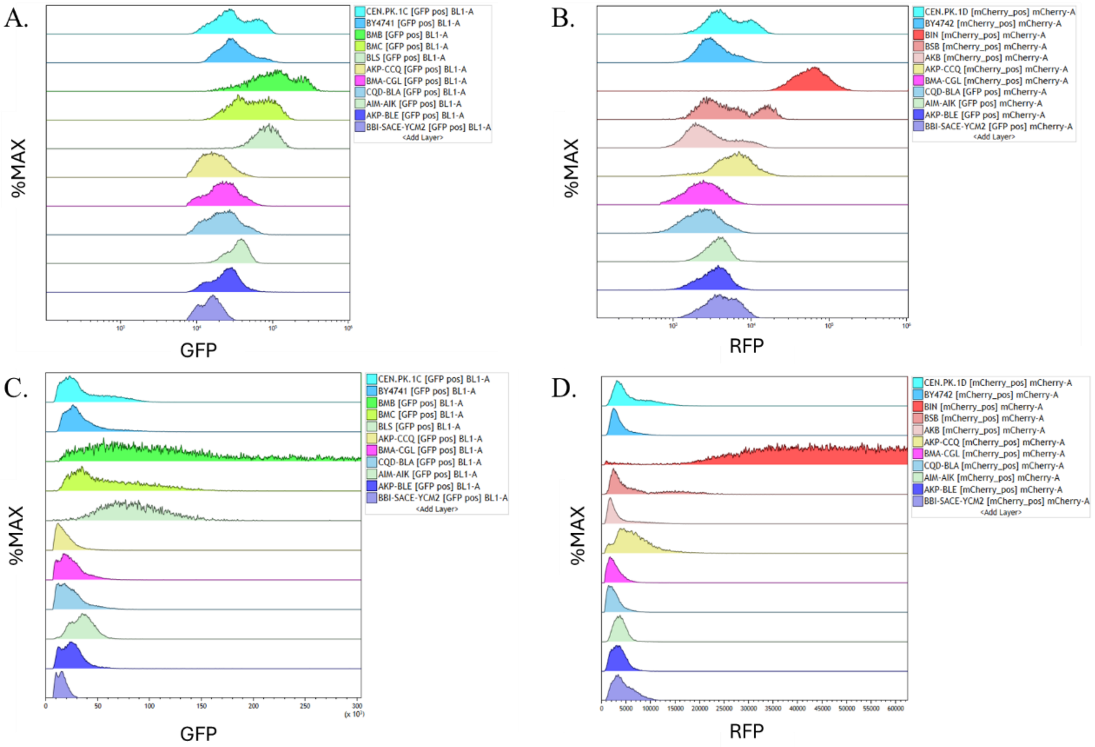


Figure 5 | Linear and log-scale distributions of GFP and RFP expression levels in selected strains. (A) and (B) display log-transformed distributions of GFP and RFP expression levels, respectively, for selected parental strains and offspring. (C) and (D) show the corresponding distributions on a linear scale. These distributions were derived from singleton 3, including only GFP-positive (BL1-A) or RFP-positive (mCherry-A) cells.

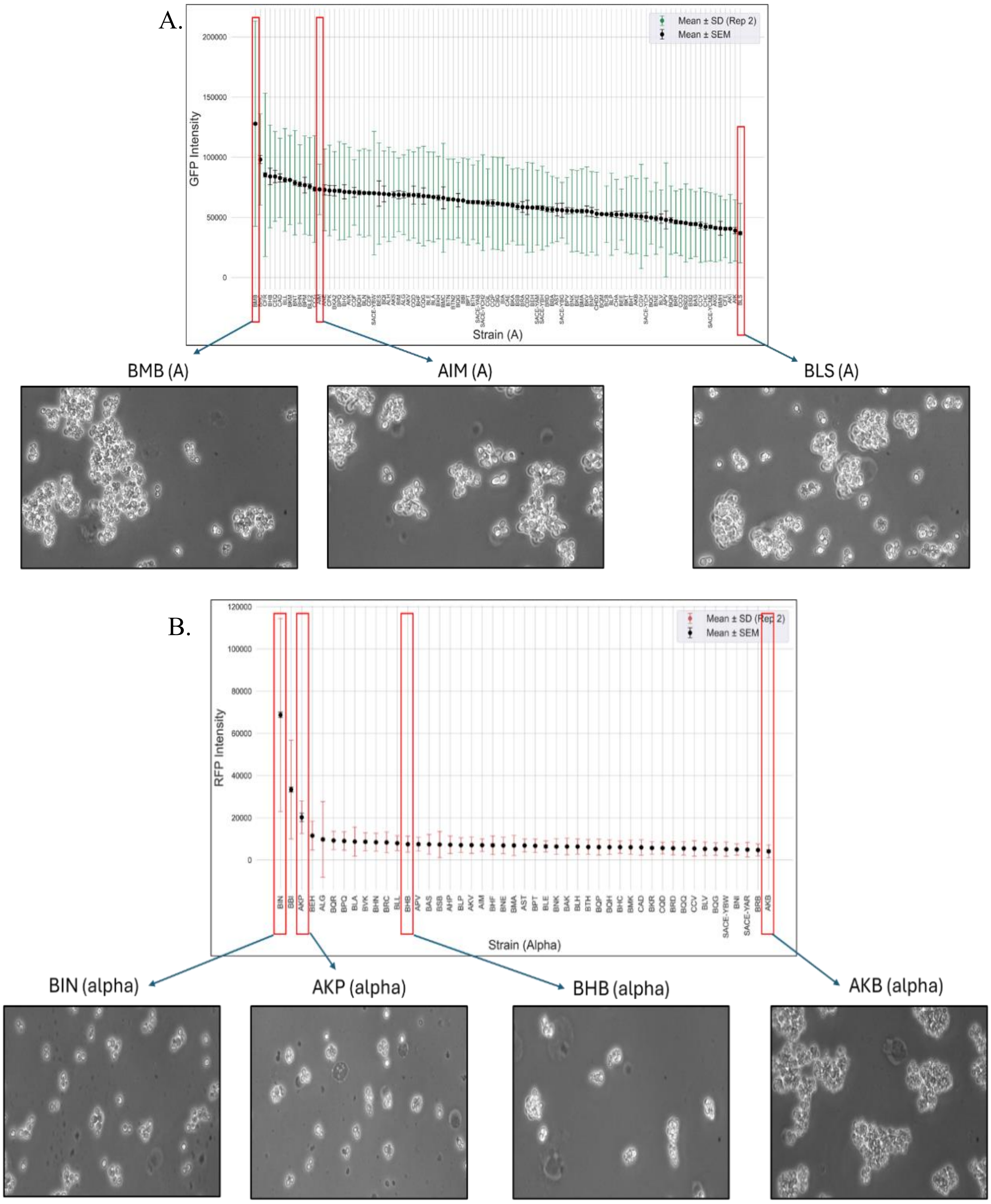


Figure 6 | Microscopic images of selected strains representing different points on the fluorescence intensity spectrum. (A) GFP intensity spectrum of mating type *A* strains: BMB (highest intensity), AIM (middle range), and BLS (lowest intensity). All these strains exhibit aggregate formation. (B) RFP intensity spectrum mating type *alpha* strains: BIN (highest intensity), AKP (third highest), BHB (middle range), and AKB (lowest intensity). BIN and AKP do not form aggregates, whereas BHB shows smaller aggregates, and AKB forms large aggregates. All images were captured using a 40× objective lens with 2× digital zoom.

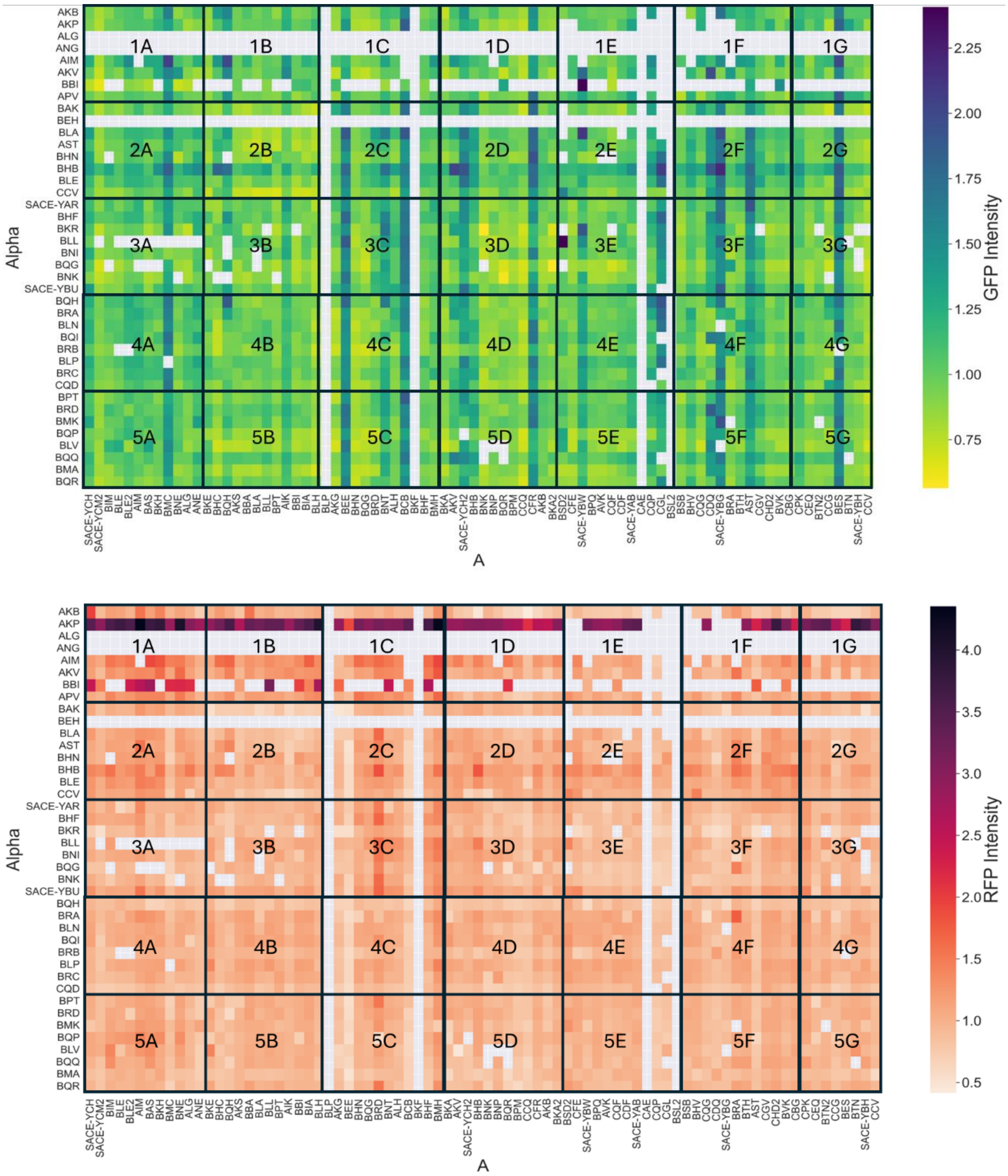


Figure 7 | Heatmap of GFP and RFP expression levels organized by plates. The figure illustrates the plates used for FACS measurements, including both parental strains, to demonstrate that no batch effect was observed across the experimental setup.

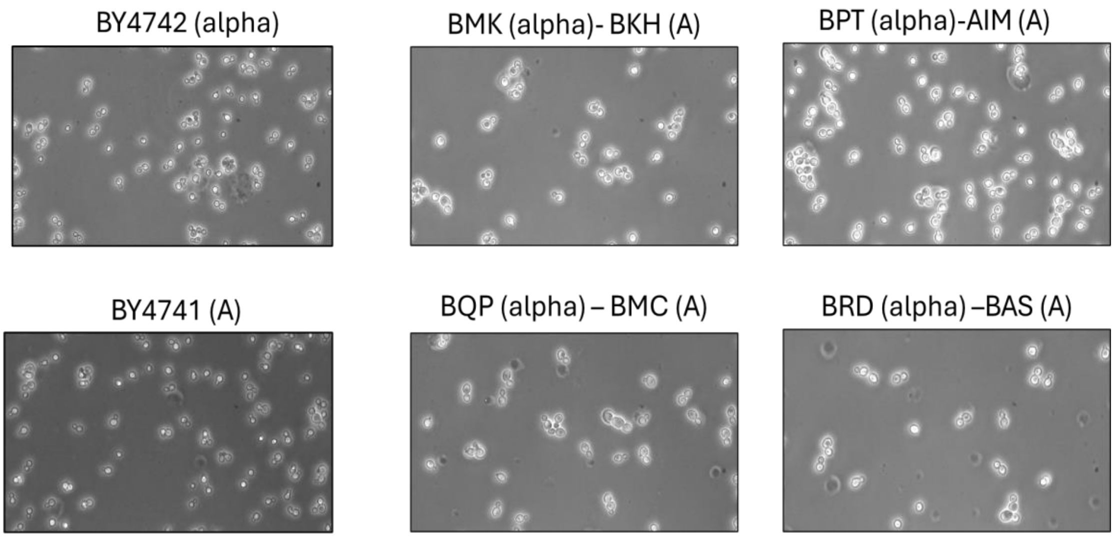


Figure 8 | Microscopic images of randomly selected offspring compared to the lab strain. The left panel shows images of the lab strain BY4741/2 for comparison. The right panel displays images of four randomly selected offspring strains. Unlike their haploid parents, the offspring do not tend to form aggregates and exhibit behavior similar to the lab strain. All images were captured using a 40× objective lens with 2× digital zoom.

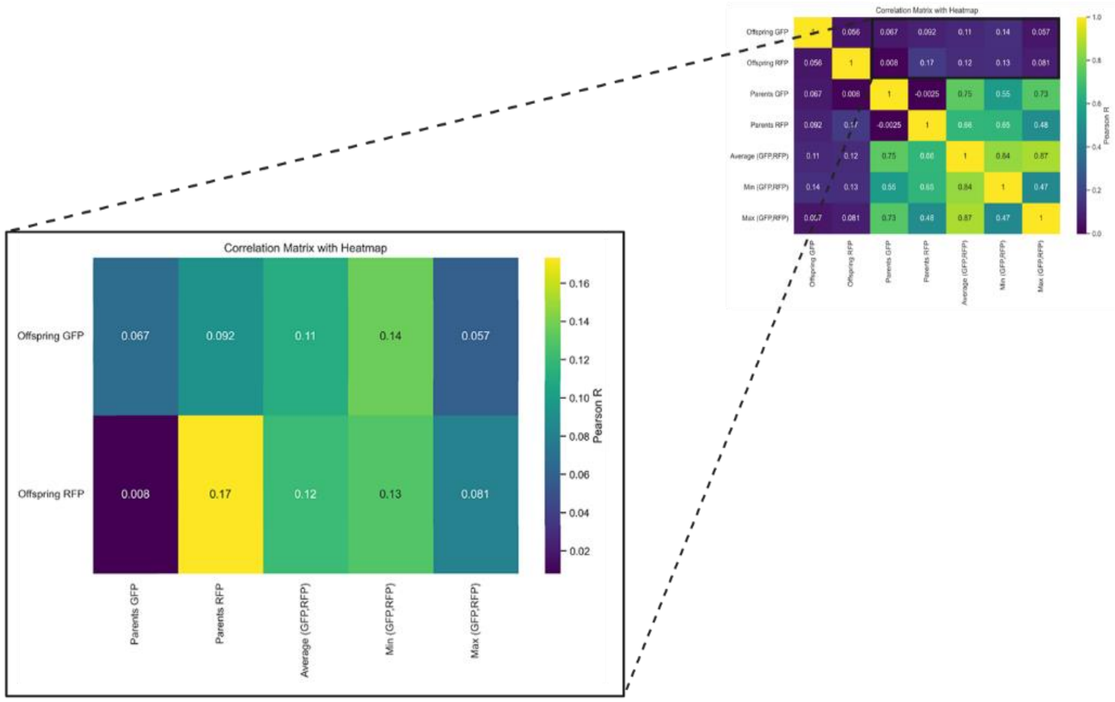


Figure 9 | Correlation map of different inheritance Patterns. This figure presents a correlation map showing Pearson's correlation coefficients (r values) between various inheritance patterns of fluorescent protein expression in parents and their offspring. providing a detailed zoom-in on the average, minimum, and maximum expression levels of GFP and RFP.

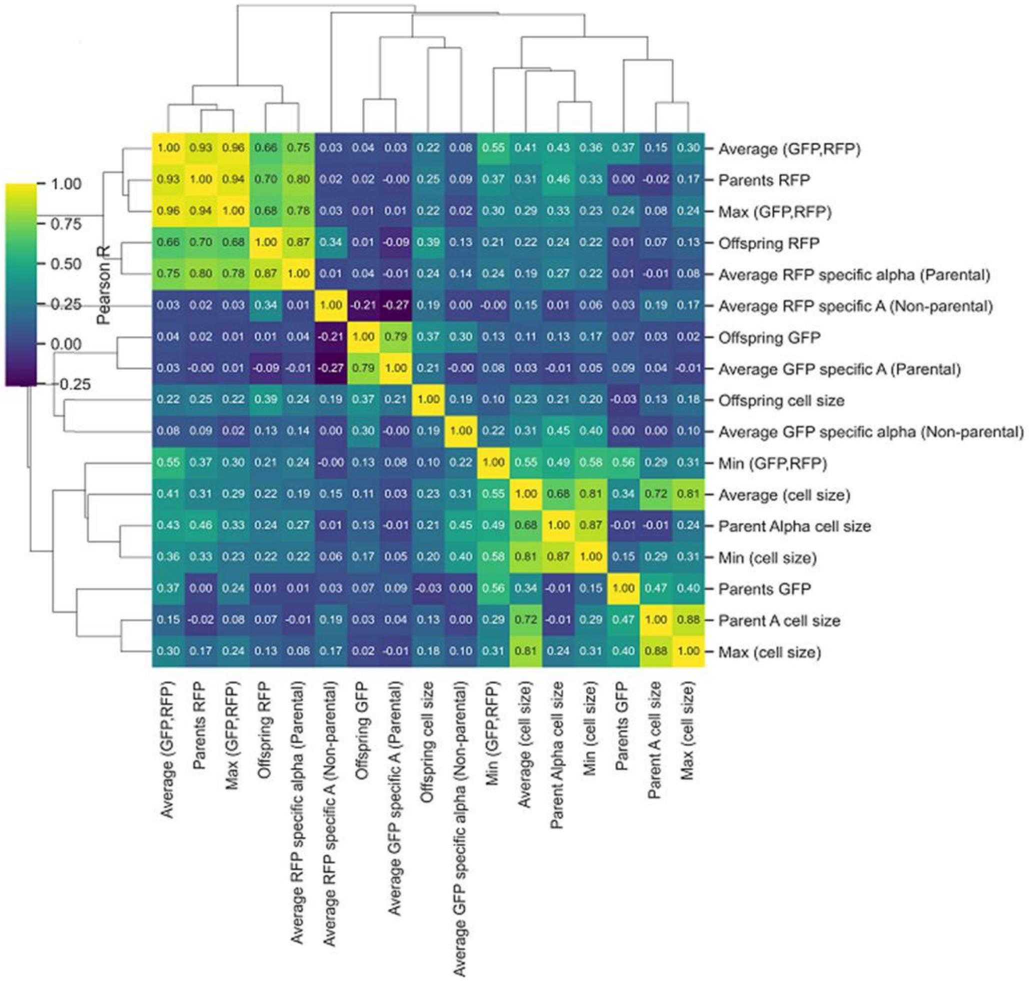


Figure 10 | Correlation map of different inheritance Patterns of both fluorescent proteins and cell size. This figure displays a correlation map illustrating Pearson's correlation coefficients (r values) between different inheritance patterns for both fluorescent protein expression (GFP and RFP) and cell size in parents and offspring.

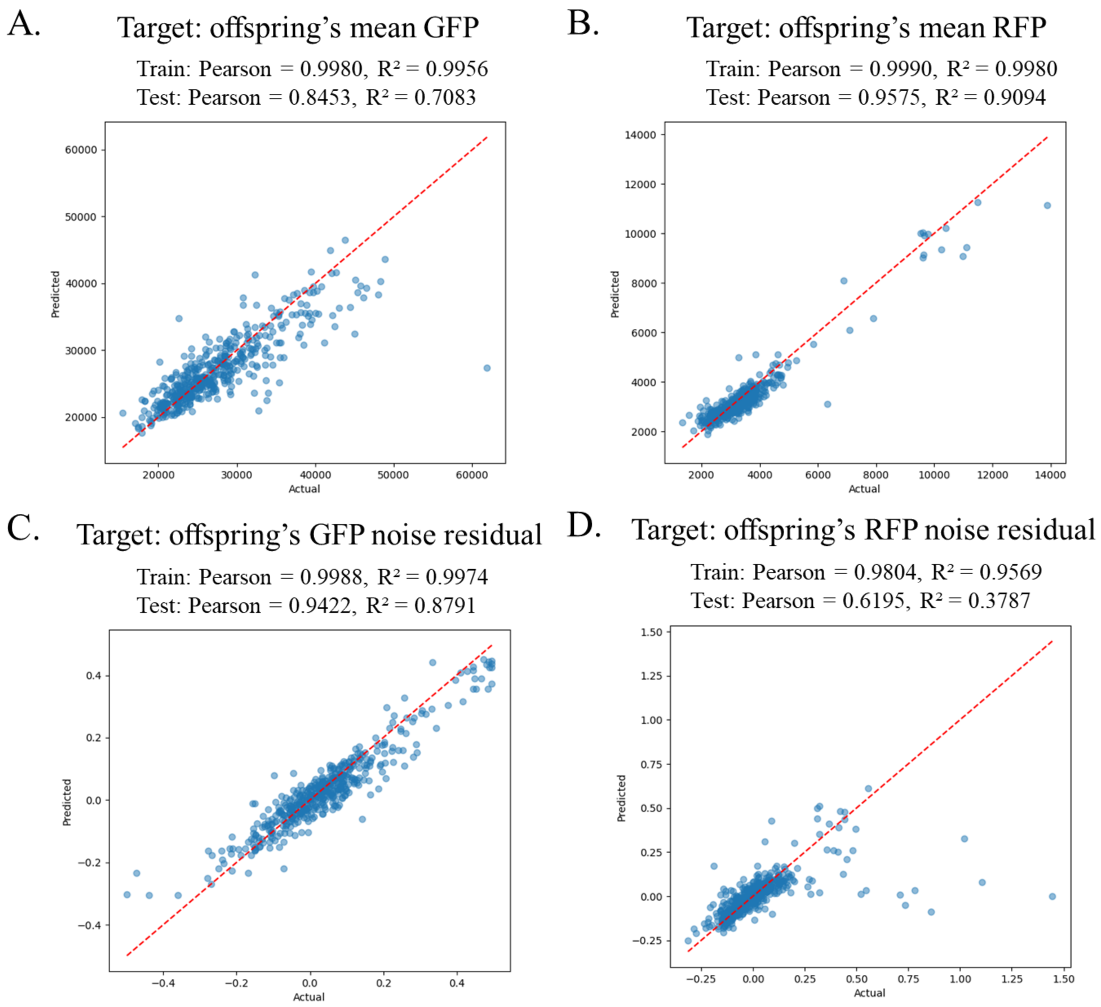


Figure 11 | Predicted vs. Actual values for the double learning CatBoost model's predictions of **GFP and RFP expression levels and noise residuals in offspring**. (A) Predicted vs. Actual values for offspring's GFP expression. The points are closely aligned along the red dashed line, indicating a strong correlation between the predicted and actual values. (B) Predicted vs. Actual values for offspring's RFP expression. The points follow the red dashed line, demonstrating the model's high accuracy in predicting RFP expression levels. (C) Predicted vs. Actual values for offspring's GFP noise residual. The points are generally scattered around the red dashed line, indicating that the residuals are well-distributed, with no obvious patterns. (D) Predicted vs. Actual values for offspring's RFP noise residual. Some points deviate from the red dashed line, reflecting the lower prediction accuracy for this target compared to the others.