scientific reports



OPEN SHIP identifies genomic safe harbors in eukaryotic organisms using genomic general feature annotation

Matheus de Castro Leitão^{1,2}, Letícia Sousa Cabral¹, Luiza Cesca Piva², Pedro Felipe de Sousa Queiroz¹, Taísa Godoy Gomes³, Rosângela Vieira de Andrade⁴, Ana Laura Alfonso Perez², Karen Letycia Rodrigues de Paiva², Sônia Nair Báo², Viviane Castelo Branco Reis², Lídia Maria Pepe Moraes², Roberto Coiti Togawa⁵, Leila Maria Gomes Barros⁵, Fernando Araripe Gonçalves Torres², Georgios Joannis Pappas Júnior^{2 \vee} & Cíntia Marques Coelho^{1 \vee}

Integrating foreign genes into loci, allowing their transcription without affecting endogenous gene expression, is the desirable strategy in genomic engineering. However, these loci, known as genomic safe harbors (GSHs), have been mainly identified by empirical methods. Furthermore, the most prominent available GSHs are localized within regions of high gene density, raising concerns about unstable expression. As synthetic biology is moving towards investigating polygenic modules rather than single genes, there is an increasing demand for tools to identify GSHs systematically. To expand the GSH repertoire, we present SHIP, an algorithm designed to detect potential GSHs in eukaryotes. Using the chassis organism Saccharomyces cerevisiae, five GSHs were experimentally curated based on data from DNA sequencing, stability, flow cytometry, gPCR, electron microscopy, RT-gPCR, and RNA-Seq assays. Our study places SHIP as a valuable tool for providing a list of promising candidates to assist in the experimental assessment of GSHs in eukaryotic organisms with available annotated genomes.

Since the beginning of genetic engineering, scientists have aimed to target heterologous DNA to genomic regions that allow transcription without perturbing endogenous gene expression. The emergence of synthetic biology has further driven the development of strains that can express polygenic traits of interest, often needing the insertion of complete metabolic pathways and genetic circuits. Consequently, researchers seek to preferentially direct foreign DNA to landing pads known as Genomic Safe Harbors¹ (GSHs), Genomic Safe Havens², or Neutral Sites³. GSHs are intragenic or intergenic regions expected to accommodate and transcribe foreign DNA inserts with no or minimal perturbation on the general gene expression levels of the host genome¹. The concept is that GSHs provide protected locations for inserting DNA, reducing unpredictable phenotypes.

The first studies identifying GSHs arose for higher eukaryote genomes, traditionally relying on empirical methods such as viral insertion sites⁴, analysis of gene function loss⁵, or gene orthology⁶. Notably, many GSHs are in intragenic loci and genomic regions of high gene density, often near oncogenes in the context of the human genome¹, raising concerns about their impact on neighboring gene expression. With the advent of gene editing tools such as zinc-finger nucleases⁷ and CRISPR-Cas systems^{8,9} or DNA integrases¹⁰, which enable biallelic transgene insertion, the risk of functional knockouts is further aggravated.

Several studies have proposed criteria to characterize and systematically identify GSHs in human cells¹, rice¹¹, Schistosoma mansoni¹², Cryptococcus neoformans², Saccharomyces cerevisiae¹³, Komagataella phaffii (Pichia pastoris)¹⁴, and Aspergillus fumigatus¹⁵. Despite their increased importance, all sites currently used as GSHs have yet to be entirely validated¹⁶, and this identification is not standardized. The orientation of neighboring genes, the length of intergenic regions, and off-target insertions have not been verified, and validation is limited to cell growth and indirect expression of the insert, with no assessment of interference in genomic expression.

¹Department of Genetics and Morphology, University of Brasilia, Brasilia, Brazil. ²Department of Cell Biology, University of Brasilia, Brasilia, Brazil. ³Department of Microbiology, University of Brasilia, Brasilia, Brazil. ⁴Catholic University of Brasilia, Brasilia, Brazil. ⁵Genetic Resources and Biotechnology, CENARGEN, Embrapa, Brasilia, Brazil. [™]email: gpappas@unb.br; cintiacoelhom@unb.br

These concerns challenge the development of new rational, standardized, and systematic strategies, leading to the emergence of bioinformatics programs to meet these demands. However, fewer than a handful are available to identify these sites, and two are restricted to the human genome^{17,18}, with the vast majority of the identified GSH candidates not fully experimentally validated^{17–19}. A common concern is that none of the cited works assessed the number of copies of the reporter gene in the genome and the possibility of off-target insertions. Thus, any expression characterization may come from insertions in other locations. They also showed that heterologous DNA insertions lead to some genomic expression perturbation^{17,18}. Furthermore, the definition of GSH criteria is challenging when considering eukaryotic organisms due to their genomic differences, making it difficult to develop generic programs for predicting GSHs. This challenge is further compounded by the potential variation in GSH characteristics among distantly related species.

To address the need to expand GSHs to a broad range of organisms, the Safe Harbor Identification Program (SHIP) provides a graphical user interface that enables the prediction of GSH candidate regions in any annotated eukaryotic genome by incorporating defined criteria for refining the search for new safe integration sites. SHIP brings unique advantages by identifying all GSHs in intergenic regions, while allowing users to select the intergenic length range and neighboring genes orientation. To assist this selection, the software displays the distribution of intergenic regions throughout the genome. Users have the flexibility to choose the genetic parts and regulatory characteristics of interest. The output includes a comprehensive report detailing the regulatory characteristics of each GSH candidate, assisting the researcher in selecting the most suitable candidates for their studies. As proof of concept, we validated the GSHs identified by SHIP in the chassis organism *S. cerevisiae* through reporter gene expression evaluation (Fig. 1A).

Results

SHIP identifies putative GSHs in eukaryote genomes using general feature data

This work evaluated three eukaryotic model organisms: *S. cerevisiae* (R64), *Homo sapiens* (GRCh38), and *Mus musculus* (GRCm38). As canonical criteria, all identifiable GSH regions are located in intergenic regions. Other criteria, such as region length and neighborhood gene orientation, may be tailored to fit the specifications of the target organism genome (Fig. 1B). Afterwards, the program applies several qualitative filters based on the length of intergenic regions and the orientation of the flanking genes. Using annotations from external databases (UCSC), it selects intergenic regions that do not overlap with annotated features such as genes, regulatory regions or transposable elements, and telomeres or centromeres (see the Materials and methods section). The output is ordered by the genomic coordinate of putative GSH (pGSH) and is presented without any particular ranking. Notably, the convergent orientation of the neighboring genes is strongly recommended to reduce possible interference in the flanking promoter regions once the majority of regulator factors are upstream sequences^{20,21}. As eukaryotic genomes present a wide range of sizes and gene distributions, GSH length is expected to change with the target organism. Just as the insertion of transgenes within very short GSHs could lead to neighborhood gene interference, very long GSHs could contain an uncharacterized gene or other genetic parts, possibly leading to genomic expression perturbation. To minimize these unwanted events, a GSH length range was established depending on the organism's average gene length distribution².

As previously shown, *S. cerevisiae* presents a condensed genome²², with most intergenic regions shorter than 1 kb with flaking genes in tandem (Fig. 1C). The yeast genome has an average of 1 gene for every 1,728 bp, directing the putative GSH (pGSH) size range between 1.2 and 1.7 kb. The *H. sapiens* (GRCh38.p13) pGSHs range was defined as 50,300–50,800 bp, and *M. musculus* (GRCm38.p6), with a pGSH size range of 48,700–49,200 bp. Notably, even for organisms with knowingly not uniform gene distribution, such as humans and mice, the canonical previously established criteria that locate GSHs in intergenic regions allowed the joint use of the average gene length distribution, expanding the identification parameters.

For *S. cerevisiae*, the program predicted six pGSHs distributed throughout five chromosomes (Fig. 1D and Table S1). The SHIP program identified 16 pGSHs in *H. sapiens* and 11 in *M. musculus* (Fig. S1, S2). These results included the name, description, and cross-references of neighboring genes and the FASTA sequence of pGSHs (see the Data availability section). For multicellular organisms, as practical examples, *H. sapiens* and *M. musculus*, along with the pGSH list, the program also returns the pGSHs' regulatory features in several cell types for the given species. In addition, parameters already defined by previous studies^{1,17} must be taken into consideration. Importantly, all regions were manually curated, confirming the algorithm's functionalities and precision.

Supplementary Table S1 presents detailed information about each pGSH identified in *S. cerevisiae*. pGSH-1 is on chromosome 2, and it is 3,780 bp from the telomere. pGSH-2 is on chromosome V, 7,895 bp away from the telomere. From the Saccharomyces Genome Database (SGD)²³, this region has the epigenetic mark H3K4me3, which is related to high transcriptional levels²⁴. pGSHs 3 and 4 belong to chromosome V. pGSH-3 is 12,337 bp from the telomere, presenting DNAseI hypersensitive sites. pGSH-4 is located in the middle of the left arm of the chromosome. pGSH-5 is on chromosome XI, 14,902 bp from the telomere. pGSH-6 is at the middle region of chromosome XV, marked by H3K4me3. Notably, pGSH-2, 3, and 5, in addition to being close to subtelomeric regions, one of the flanking genes belongs to gene families.

Generation of S. cerevisiae strains with reporter genes in each GSH

Five of the six pGSHs identified by SHIP were selected for *in vivo* validation. Due to its relative proximity to the telomeric region, pGSH-1 was discarded from the experimental analysis. The BioBricks (*ymUKG1* and *URA3*) were directed centrally within each pGSH to reduce the probability of gene expression perturbations. Furthermore, the BioBricks were aimed at sites with favorable epigenetic marks for open chromatin states, such as H3K4me3 and DNase I hypersensitive sites, to improve insertion efficiency and expression. These BioBricks were delivered as separate DNA pieces, simulating the assembling approaches to build metabolic pathways and



Fig. 1. SHIP identification of putative genomic safe harbors (pGSHs). (**A**) Strategy overview of the eukaryotic genomic Safe Harbor Identification Program. The first step involved the design of the Safe Harbor Identification Program (SHIP), an algorithm for searching genomic safe harbors in eukaryotes from general feature data (Design); *S. cerevisiae* was chosen for the *in vivo* validation of SHIP, resulting in its genetic transformation of two overlapping BioBricks composed of a reporter gene and an auxotrophic marker (Test); followed by the analysis steps (Analysis). (**B**) Representative scheme of the SHIP software. As inputs, a genomic annotation (.gff3), a regulatory annotation (.gff3), and two files (.json) with the indication of genetic parts to be considered for pGSH selection. As output files, the algorithm returns a table and a graph (.png) with the distribution of intergenic distances and a file (.txt) with the list of intergenic regions and regulatory aspects. (**C**) Histogram with the genomic distribution of the intergenic regions between the three possible arrangements of the flanking genes. (**D**) Chromosome number, coordinates, neighboring genes, and size of the intergenic regions identified as pGSH. (**E**) Ideogram marking the identified pGSHs generated with Ideogram.js⁷⁶.

genetic circuits. Following PCR identification of the clones with correct BioBricks insertion (Fig. S3), three clones were selected for each pGSH, and their PCR amplicon was sequenced to confirm their insertion (Fig. 2 and Fig. S4). Each pGSH clone exhibited the expected insertion site, confirming their suitability for further validation as true GSHs (see the Data availability section).

Identified GSHs present genomic stability without fitness loss and an increasing copy number near subtelomeric regions

One of the primordial characteristics of a GSH is its stability and capacity to keep the exogenous insertion through generations with no adverse fitness effects. Thus, the strains transformed with the reporter genes were cultivated in liquid YPD for ten days continuously, and the BioBrick's presence within each GSH was verified by PCR using genomic primers (Fig. 3A). All the GSHs presented genomic stability after approximately 100 mitotic generations (Fig. 3B and Fig. S5).

The growth curve for each GSH strain was comparable but distinct from the control (untransformed BY4741) after 36 h in YPD medium (Fig. 3C, D). The calculated growth rates were between 0.35 and 0.39 for the GSH lines and 0.30 for the control (Fig. S6), indicating a detectable difference in adaptive value between them, being higher for the lines where uracil prototrophy was restored and demonstrating no fitness loss.

Previous studies have demonstrated that subtelomeric regions are double-strand break hotspots with a high recombination rate^{25,26}. This activity contributes to the emergence and evolution of gene families^{27,28}. Since SHIP identified three GSHs near these regions, we assessed the copy number of the reporter transgene using qPCR. The results showed that GSH-4 and GSH-6 had a single copy at the central region of their chromosome arms.



Fig. 2. Sanger sequencing of the clones demonstrates correct insertion in the region of each GSH in the 5 lineages. (**A**) Representation of PCR amplification for sequencing using primers targeting the genomic region outside homology arms (HR) marked in yellow. *ymUkG1* is marked in green and *URA3* in pink. (**B**) S2U. (**C**) S3U. (**D**) S4U. (**E**) S5U. (**F**) S6U. Multisequence alignment was performed with the MAFFT v7 program⁷⁷ using the following parameters: Gap opening penalty of 1.53; Gap extension penalty equal to 0 and quick direction adjustment function enabled.



Fig. 3. GSHs genomic characterization and Growth curve for each GSH cell line. (**A**) Schematic strategy of approximately 100 mitotic generations. (**B**) PCR analysis of each one of the five GSH cell lines after approximately 100 generations with genomic primers specific for complete amplification of the insert. For each biological triplicate, fifteen colonies were randomly collected. BY4741 as positive control (C+) and H₂O as negative control (C–). A cut of each gel is shown, removing wells and unused parts. Full-length gels are included in the Data availability section. (**C**) Schematic strategy for the growth rate analysis of the five GSH cell lines compared to the BY4741, as control. (**D**) Average and standard deviations of growth curves measured at 0, 2, 4, 6, 12, 24, and 36 h with an initial OD₆₀₀ of 0.1. (**E**) Copy number analysis of the *ymUkG1* gene in 3 clones of each GSH cell line. (**F**) Graphic bars showing the mean and standard deviation of the percentage of cells expressing *ymUKG1* on all GSH cell lines. Experiments performed in technical and biological replicates on three independent days. Bioicons from Servier Medical Art licensed under CC BY 4.0.

Scientific Reports | (2025) 15:7193

The GSHs located at subtelomeric regions presented more than one copy of the transgene (Fig. 3E). GSH-2, the one closest to the telomere (7.9 kb away), presented an average of seven copies. In contrast, GSH-3 (12.3 kb) had three copies, and GSH-5 (14.9 kb) presented two copies, potentially indicating to the user that targeting inserts to GSHs close to subtelomeric regions can lead to the insertion of more than one copy into the genome.

GSH cell lines expressed the reporter gene in more than 90% of the population, preserving the morphological characteristics

Flow cytometry analysis revealed that the GSH lineages expressing *ymUkg1* exhibited an average detectable green fluorescence in more than 90% of the population (Fig. 3F). This percentage was achieved and maintained from 2 to 36 h of growth. No detectable variation was observed in the fluorescent population throughout cultivation and among GSHs (Fig. S7 and Fig. S8), indicating that each strain was capable of expressing and maintaining stable and predictable expression of the transgene (see the Data availability section).

Scanning electron microscopy (SEM) was used to investigate any morphological impact on the GSH cell lines. Compared to BY4741, the GSH strains did not display relevant cell surface changes, being indistinguishable from the control and from each other (Fig. S9 and Fig. S10).

Differentially expressed genes are not related to the insertion site in GSH cell lines

One of the fundamental characteristics of a GSH is its capacity to ensure predictable BioBrick transcription with no or minimal interference on the local and global gene expression. Therefore, we analyzed the expression of the transgene and of genes directly in contact with the GSHs (Neighboring Genes) using RT-qPCR (Fig. 4A). The results demonstrate an accumulation of the *ymUKG1* transcript for all GSH strains. pGSH-5 which has 2 copies of the transgene, had the highest *ymUKG1* transcript relative expression, followed by pGSH-4 with one copy. Although pGSH-2 and 3 (the ones closest to the telomeres) have the highest number of *ymUKG1* copies, their relative expression is equivalent to pGSH-6, with a single copy (Fig. S11). Notably, regardless of the GSH used to host the transgene, four neighboring genes across all the predicted GSHs exhibited persistent variation in expression, which was not correlated to the specific GSH employed (Fig. 4B–F).

Finally, we assessed possible interferences from BioBrick insertion in overall host genomic expression. Global gene expression analysis was performed using RNA-Seq on one clone of each GSH cell line once the insertion occurred precisely at the same site. We analyzed the transcriptome of each GSH construct in biological replicates against the control strain BY4741. All GSH cell lines exhibited around 16% of genes differentially expressed compared to the control, displaying a high degree of similarity among themselves (Fig. 5A and Fig. S12). Both RT-qPCR and RNA-Seq analysis corroborated previous studies in which differential expression (DE) analysis revealed significant differences for several genes between GSH and control human cell lines, and similarity among them¹⁷.

URA3, intentionally recovered as one of the reporter genes, was the most positively regulated gene, whereas *URA1* was the most negatively regulated in all experimentsFig. . 5B–F), as expected^{29,30}. Moreover, it is important to report that according to previous works, the recovery of auxotrophy vias affects genomic expression, again corroborating our DE analysis data¹⁷.

Most of the differentially expressed genes are shared among all GSH cell lines

Considering that the expression profile of \overline{GSHs} was similar between cell lines and that the BioBricks were in different loci and chromosomes, the reintroduction of *URA3* could cause this apparent genomic perturbation, as already observed by previous studies^{31–33}. To elucidate this possibility, the level of sharing of differentially expressed genes between the GSHs and whether, among the shared genes, there is some functional enrichment, the results of the DE analysis of the RNA-seq experiment for each GSH were processed. For all GSH lines, we performed a paired DE analysis against the untransformed control BY4741.

The group of upregulated genes from all experiments totaled 1257 upregulated genes (Fig. S13A). Of these, 423 (33.65%) had increased DE in all experiments, being the most frequent subgroup of genes among the clone combinations. It is worth mentioning that the GSH-6 and 4 clones presented 120 and 83 upregulated genes, respectively, which were not shared with the other lineages. In all experiments, 1,086 unique genes showed a significant decrease in expression (Fig. S13B). Of these, 257 (23.66%) were downregulated in all cell lines, constituting the most frequent subgroup of genes.

To verify whether these groups of shared genes, positively or negatively regulated, were related to GSHs, the data was mapped to chromosomal regions. The sites with differentially expressed genes were spread throughout the yeast genome, showing no positional correlation with any of the identified GSHs (Fig. S13, C–G). Moreover, there is no overlap between these regions of differentially expressed genes and the GSH loci, further corroborating previous studies^{17,18} and indicating no link between transgene insertion and differential expression.

Lack of gene ontology enrichment in differentially expressed genes

A gene ontology enrichment analysis of the differentially expressed genes shared among all strains was performed to verify possible functional perturbations in the host genome. The upregulated gene set displayed significant enrichment for the term cellular component (CC) of the ribosomal unit and molecular function (MF) related to the structural component of the ribosome (Fig. S14A, B). In general, this might suggest more prominent translational activity. There was no enrichment of gene ontology terms for downregulated genes shared across all GSH constructs (Fig. S14C).

Even without substantial ontological enrichment, the most differentially expressed genes relative to the untransformed control were analyzed. They showed again that gene expression distribution was generally affected, with *URA3* being the most strongly upregulated and *URA1* being the most strongly downregulated across all the experiments (Fig. S15). However, most of the genes were unannotated, putative, or with dubious



Fig. 4. Expression dynamics of neighboring genes of GSHs cell lines by RT-qPCR. (**A**) Schematic representation of the genes analyzed. Histogram of the relative expression of all neighboring genes for each GSH lineage. (**B**) S2U. (**C**) S3U. (**D**) S4U. (**E**) S5U. (**F**) S6U. Green column indicates expression of the *ymUKG1* gene. Colored columns highlight neighboring genes (Purple at 5' and Blue at 3') of the GSH of the analyzed strain and gray columns show the neighboring genes of the other GSHs. The cracked columns represent values from the untransformed control (BY4741). Relative expression on the Y axis and genes analyzed on the X axis. Asterisks indicate significant changes in the analysis of variance (ANOVA). (*) p < 0.05, (**) p < 0.01, (***) p < 0.001 and (****) p < 0.0001.

ORFs, comprising 73.33% of the positively regulated genes and 80% of the negatively regulated ones (Table S3 and Table S4), suggesting that fundamental biological processes and essential metabolic and stress response pathways were not affected.

Generation of triplex GSH cell line expressing ymUKG1, ymBeRFP, and α -amylase

As proof that the predicted GSHs can serve as landing pads for the simultaneous expression of entire biological pathways or complex genetic circuits, we constructed a triplex strain expressing *ymUKG1*, *ymBeRFP*³⁴, and α -*AMY*³⁵. To construct the triplex, the strain containing *ymUKG1* within GSH-2 (Strain S2U) was transformed with α -*AMY* targeting GSH-5 (Strain S2U5A), and sequentially, *ymBeRFP* was inserted in GSH-6 (Strain S2U5A6B) (Fig. 6A).

S2U5A6B presented a capacity to degrade starch (Fig. 6B and Fig. S16) despite a modest amylolytic activity (Fig. 6C). It also exhibited expression of α -*AMY* and *ymBeRFP*, without affecting the neighboring genes (Fig. 6D). Most importantly, the result showed that it was possible to simulate the construction of polygenic circuits of interest using the GSHs described by SHIP.



Fig. 5. Genomic expression dynamics of GSHs cell lines by RNA-Seq. (**A**) PCA graph of the differential gene expression of the GSHs lines (S2U, S3U, S4U, S5U and S6U) in relation to the untransformed control BY4741. Vulcanos Plots of differential expression of GSHs cell lines by RNA-Seq. (**B**) S2U. (**C**) S3U. (**D**) S4U. (**E**) S5U. (**F**) S6U. Colored dots represent differentially downregulated (blue) and upregulated (red) genes. Described genes are identified with their standard name and undescribed or putative genes are identified with their systematic name.



Fig. 6. Functional and expression analysis of the S2U5A6M triplex GSH cell line. (**A**) Sanger sequencing of the S2U5A6M strain. (**B**) Growth and amylase activity of S2U5A6M clones and growth of S2U6M (no-amylase control) on minimal medium containing soluble starch. The plate was stained with iodine vapor and clear halos indicate starch hydrolysis. (**C**) Amylase enzyme activity assay. All clones significantly differed from BY4741 according to a Mann-Whitney U test but did not show statistically significant differences between themselves. (**) p < 0.01 and (***) p < 0.001. (**D**) Expression dynamics of neighboring genes by RT-qPCR. Green, yellow, and red columns respectively indicate expression of the *ymUKG1*, *ymBeRFP*, and α -amylase (α -AMY). Colored columns highlight neighboring genes (Blue at 5' and Purple at 3') of the GSH used and gray columns show the neighboring genes of the other GSHs. The cracked columns represent values from the untransformed control (BY4741). Relative expression on the Y axis and genes analyzed on the X axis. Asterisks indicate significant changes in the analysis of variance (ANOVA). (***) p < 0.001 and (****) p < 0.0001.

Scientific Reports | (2025) 15:7193

nature portfolio

Discussion

We used the program SHIP to identify six regions with characteristics of GSHs in S. cerevisiae and validated five of them (GSH-2 to 6) in vivo (Fig. 1). These fulfilled fundamental criteria for a safe landing pad:¹ accessibility to transgene integration (Fig. 2)², genomic stability through 100 mitotic generations (Fig. 3B) without reducing cell viability (Fig. 3D and Fig. S6)³, predictable and constant expression of the transgene (Fig. 3F and Fig. S7), and⁴ minimal interference with genomic and flanking gene expression (Figs. 4 and 5). As GSH cell lines restored the uracil biosynthetic pathway by reintroducing URA3, a higher growth rate than the control strain was expected (Fig. 3D and fig. S6)^{31,32}. Flow cytometry data showed that approximately 90% of the cells accumulated the reporter protein (Fig. 3F and Fig. S7), demonstrating high transgene expression with no morphological effects (Figs. S9, S10). qPCR results revealed that targeted insertion of BioBricks into GSHs near telomeric regions (GSH-2, 3, and 6) resulted in more than one gene copy. In contrast, those in the middle of the chromosome arms (GSH-4 and 6) resulted in one copy (Fig. 3E). These additional inserts found in clones with GSH near the telomeric regions can be considered off-targets. Although these follow biological characteristics of subtelomeric regions^{25,26} further investigations will be needed to identify the locations of these insertions and to determine whether they occur during or after the transformation process. Analysis of the four annotated genes flanking GSH-2, 3, and 5 revealed that three belong to gene families (PUG1³⁶, ARN1³⁷, and COS9³⁸). However, the number of copies did not correlate to the ymUKG1 transcript accumulation (Fig. S11) since GSHs-2 and -3, probably due to their telomere proximity, had the lowest relative expression profile. This information gives the user of the SHIP tool flexibility in selecting putative GSHs based on the intended transcript accumulation, allowing finetuning of gene expression control.

RT-qPCR and RNA-Seq results from our study corroborate findings that transgene insertion affects the cell's transcriptional program^{17,18}. Previous studies aimed to develop software to identify GSHs in humans detected significant differential expression of several genes spread across different chromosomes. Furthermore, most DE genes are shared between cell lines, even at different integration sites. Approximately 33% of genes were upregulated, and around 25% were downregulated. Among the differentially expressed genes, most were detected across all GSH cell lineages (Fig. S13A, B), most being unannotated, putative, or with dubious ORFs.

Additionally, no functional enrichment was observed among these, without indication of a coordinated cellular response, such as a response to cellular stress or cell division reprogramming (Fig. S14). Considering that the expression profiles of GSH cell lines were very similar and the BioBricks were at different loci, the genomic perturbations did not follow a clear trend linked to the transgene insertion. Restoring uracil metabolism led to higher growth rates than the control and possibly contributed to this differentiation in genomic expression, as shown by previous studies^{31–33}. As demonstrated by Alam, auxotrophic lineages presented 85% of the transcriptome with differential expression when compared to prototrophic lineage³³. Nevertheless, this does not exclude the possibility of long-term effects of stress, particularly regarding epigenetic changes that may modulate genomic expression^{39–42}.

Corroborating our data, three of the five GSHs identified by the SHIP program have been previously used as regions for transgene insertion in *S. cerevisiae*, indicating consistency in our predictions. Previous studies that identified these regions searched for intergenic sequences but did not conduct further analyses to validate the use of GSH-2⁴³, GSH-5⁴⁴, and GSH-6⁴⁵.

Furthermore, the GSHs described in this study supported multiplex insertion, paving the way for polygenic assemblies (Fig. 6). Despite this, the insertion of an entire metabolic pathway is a more complex approach that will demand deeper investigation related to the length limit of each BioBrick inserted in each GSH to avoid flanking gene perturbations and possible variation in insert expression depending on the GSH's tridimensional position in the nucleus.

Finally, the SHIP program will significantly contribute to the rational and systematic identification of GSHs for eukaryotic organisms with an annotated genome, returning an overview of the genomic density and a detailed description of each region, leading to time and cost reduction for obtaining lineages of interest. It is worth highlighting that the algorithm deals with the genome as a conjunction of annotated parts. As it does not make any priori predictions based on the DNA sequence, the SHIP program has some shortcomings that may affect its applicability to other species. It depends on high-quality genome assemblies, ideally at the chromosome level, as well as accurate annotations. A lack of assembly contiguity and imprecision in gene, repeats, or regulatory region prediction can directly impact the results. Additionally, specific parameter adjustments are necessary to account for the characteristics of the taxa being studied, such as the range of intergenic lengths. The crossspecies applicability of the software also requires careful analysis, considering the specific characteristics of each genome. Therefore, as new species are analyzed, it will be essential to incorporate specific modules in future updates. Furthermore, there may be cases where predicted GSHs do not pass experimental validation or exhibit behaviors that were not anticipated due to the presence of unannotated genetic parts. In addition, future improvements should include aspects of genome 3D structure and epigenetic marks related to chromatin states as an option in the SHIP software. In conclusion, the SHIP returns a list of putative GSHs requesting a critical analysis to choose the pGSH that best meets the users' needs, providing the cornerstone to characterize and explore such genomic regions.

Materials and methods Experimental design

The objectives of this work were the development of a software for the identification GSH in eukaryote genomes and the experimental validation of this tool in *S. cerevisiae*. The first step was the design of the Safe Harbor Identification Program (SHIP), an algorithm for identifying genomic safe harbors in eukaryotes using general feature data (Design). *S. cerevisiae* was selected for *in vivo* validation, leading to its genetic transformation with

two overlapping BioBricks containing a reporter gene and an auxotrophic marker (Test), followed by subsequent results analysis (Analysis) (Fig. 1A).

Rules for predicting GSHs

The rule for predicting GSHs must change depending on the target genome, but two characteristics are constant: the orientation of neighboring genes and the length of the intergenic region. The orientation of neighboring genes was convergent for all three organisms analyzed in this study. The length of the intergenic region was determined as the rounded value of the average gene per bp as the upper limit and 500 bp less as the lower limit.

For the *S. cerevisiae* genome, the criteria for predicting GSHs were the absence of annotated gene, ncRNA, pseudogene, centromere, telomere, long terminal repeat (LTR), mobile genetic element, origin of replication, transposable element gene, meiotic recombination region, sequence annotated feature, autonomously replicating sequence, telomeric region sequence, mating-type cassette, retrotransposon mRNA, expressed sequence tag (EST), and predicted transcription factor binding sites. Additionally, for *H. sapiens* excluded regions with transfer RNA (tRNAs) and long intergenic non-coding RNAs (lincRNAs), and for *M. musculus*, transfer RNA (tRNAs), protein-coding and non-protein-coding genes (ncbiRefSeq). The SHIP algorithm included these rules for predicting pGSHs as Features and Tracks files.

SHIP algorithm

The Safe Harbor Identification Program (SHIP) is written in Python and searches genomic safe harbors in eukaryotes with complete and annotated genome sequences available in public databases (NCBI or Ensembl).

The inputs to SHIP are¹ a genomic annotation in general feature format file (.gff);² an optional regulatory annotation in (.gff) to characterize the GSH;³ two files (.json) with the indication of genetic parts to be analyzed (Features) and to be considered as GSH exclusion filter after verification of its presence in the University of California Santa Cruz (UCSC) Genome Browser⁴⁶ (Tracks);⁴ characteristics desired by the user when searching for GSHs (Intergenic size range (bp), Neighboring genes orientation, and Analysis options). As outputs for sequential analysis, SHIP returns tabular and graphical reports with the distribution of intergenic distances and frequency of neighboring genes, the list of GSHs regions with the neighboring genes with their information from several databases, such as Uniprot⁴⁷, RefSeq⁴⁸, GO⁴⁹, UniParc⁵⁰, String, and among others. The program is available at www.github.com/MCLeitao/Ship.

The modular auxiliary files (Features and Tracks) are required to expand the algorithm's scope and give flexibility based on user necessity, annotation quality, and genomic particulars of each organism. Features and Tracks used for each organism are listed in the supplementary (Table S2).

The SHIP algorithm has the following steps:

- 1. Organize the genome. Sort and filter the GFF, select the parts considered for the next steps (indicated by the user through the features.json file) and remove the completely overlapping annotation.
- 2. Table assembly and data charting. At this point, the algorithm identifies the intergenic regions by calculating the difference between the downstream gene start and the end of the upstream gene. A table is returned for the user with the number of chromosomes and genes identified, the number of intergenic intervals in total and within each neighboring gene orientation (Tandem, Divergent, and Convergent), and the disregarded amount of totally overlapping genes. The data is charted to facilitate the user visualization of the target genome distribution and organization.
- 3. Defining the GSHs characteristics. With the chart containing the target genome distribution and organization, the user can select the flanking gene orientation and the minimum and maximum size of the GSHs. Suppose the user has a GFF file with all regulatory features, such as DNA methylation, histone modifications, chromatin conformation, transcription factor binding sites, enhancers, repressors, and microarray annotations. In that case, the features within each GSH will be recorded for further processing.
- 4. Annotation tracks analysis. Based on the track file, the algorithm looks for these annotation tracks (indicated by the user) within each selected GSH through the genomic coordinate data from the UCSC Genome database. If the presence has been confirmed, the GSH is discarded.
- 5. Flanking genes analysis. Using the Ensembl REST API⁵¹, all genetic features linked to the neighborhood genes ID and all external references are recorded. It includes information related to the genes' transcripts and translations. Especially when dealing with a multicellular organism, the algorithm also searches for the Regulatory Feature recorded in each Epigenome for the given species.

Strains and media

S. cerevisiae BY4741 (MATa, met15 Δ , his3 Δ , ura3 Δ , leu2 Δ)⁵² was obtained from the EUROSCARF collection (Y00000), and it was used to generate all the yeast cell line strains described in this study. Yeast was routinely grown in YPD medium (10 g L' yeast extract, 20 g L' peptone, and 20 g L' glucose). The solid medium used 2% agar. MD medium used 0.34% Yeast Nitrogen Base (YNB), 1% (NH4)2SO4, 2% glucose and was supplied with 0,002% methionine, 0.002% histidine, 0,002% uracil or 0,01% leucine when needed.

DNA cloning of the single-copy reference gene *TAF10*, constructed from plasmid pPCV, was performed using chemically competent *Escherichia coli* XL-10 Gold (Agilent Technologies) grown in Luria-Bertani (LB) medium (5 g L' yeast extract, 10 g L' peptone, and 10 g L' NaCl, pH 7.2). When needed, agar was added to a final concentration of 1.5%.

BioBricks assembly and yeast transformation

BioBricks (Table S5) design followed the homologous recombination strategy of 60 bp overlap arms and convergent gene orientation (see Table S6 for a list of primers). BioBrick one was composed of the monomeric

version of a coral-derived green fluorescent protein coding gene⁵³, codon-optimized for *S. cerevisiae* (*ymUKG1*)⁵⁴, driven by *TEF1* constitutive promoter. The second one was the auxotrophic marker *URA3* from the plasmid pYC210⁵⁵ with its promoter and terminator. Following the manufacturer's recommendations, both were amplified using Platinum⁵⁴ Taq DNA Polymerase High Fidelity (Invitrogen). PCR reactions were purified using the ReliaPrep⁵⁴ DNA Clean-Up and Concentration System (Promega), concentrated in the Speed Vacuum, quantified in NanoDrop⁵⁴ OneC Spectrophotometer (Thermo Fisher), and visualized on an agarose gel to verify integrity, size, and corroborate quantification.

S. cerevisiae BY4741 was used as the parental strain to generate the yeast strains with the BioBricks within each tested genomic safe harbor. BioBricks were assembled by *in vivo* homologous recombination through a transformation of yeast cells using the following protocols: the High-efficiency transformation by electroporation⁵⁶ and the LiAc/SS-DNA/PEG transformation⁵⁷ with the modifications⁵⁸.

PCR identification and Sanger sequencing

After transformation, colonies were randomly selected to identify the correct insertion clones. Each transformed cell was grown in YPD medium overnight and had the genomic DNA extracted using the Wizard[®] Genomic DNA Purification Kit (Promega) following the manufacturer's recommendations. The samples were amplified by PCR with the GoTaq[®] G2 DNA Polymerase (Promega) using genomic primers (Table S7) for complete amplification of the insert. Three clones of each GSH were selected for posterior analyses.

For the samples with correct insertion, PCR replicates were performed with the Platinum[™] Taq DNA Polymerase High Fidelity (Invitrogen), genomic oligonucleotides, and following the manufacturer's instructions. Then, the samples were analyzed using 1% agarose gel electrophoresis to confirm the size of the amplicon and separate any non-specific or partial amplifications. All image acquisitions were made using the Gel Doc EZ System 1,708,270 (Bio-Rad) and the Image Lab[™] software with the default parameters. The gels were not processed. The bands of the expected size were cut from the gel, and the DNA was purified using the QIAquick PCR & Gel Cleanup Kit (Qiagen). Sanger sequencing was performed by ACTGene Análises Moleculares (Brazil) using the AB 3500 Genetic Analyzer (Thermo Fisher).

Genomic stability test

The genomic stability analysis of the GSHs regions after BioBricks insertion was performed using the continuous cell passage method⁵⁸. The colonies were initially inoculated in 5 mL of YPD and grown overnight. Then, 5 μ L of this culture was transferred to 5 mL of fresh YPD medium. This cycle was repeated for ten days, resulting in approximately 100 generations. Subsequently, the cells were plated on YPD, and fifteen clones were randomly selected and analyzed by PCR using the genomic primers (Table S7) for complete amplification of the insert for each technical triplicate.

Copy number analysis

The number of copies of the reporter gene integrated into each GSH was analyzed by qPCR using primers for the *ymUKG1* gene and the single-copy reference gene $TAF10^{59}$ (Table S8). qPCR was performed using Applied Biosystems' Fast SYBR Green Master Mix and the QIAGEN Rotor Gene Q thermal cycler. Each qPCR reaction used 10 ng of genomic DNA. The number of copies of *ymUKG1* in each genome was analyzed by absolute quantification using plasmid-based standard curves for calculating the absolute number of copies of each gene (*ymUKG1* and reference gene *TAF10*) in the samples, according to the thermal cycler manufacturer's protocol⁶⁰. The standard curve for *ymUKG1* used plasmid YCUkGy, constructed in this study and 6972 bp in size, while the standard curve for *TAF10* used plasmid pPCV-TAF10, constructed in this study from plasmid pPCV and 3,012 bp in size⁶¹. The number of copies of *ymUKG1* in each genome was obtained by normalizing the results for the reporter and reference genes.

Flow cytometry and growth rate analysis

The accumulation of *ymUkg1* protein was analyzed by flow cytometry using the BD FACSVerse[™] Cell Analyzer (BD Biosciences) configuration 4-Blue 2-Red (ROU) and flow between 500 and 900 events per second. Preinoculated clones grown overnight were inoculated in 20 mL of YPD in 125 mL Erlenmeyers with an initial OD_{600} of 0.1. Growth was measured at 0, 2, 4, 6, 12, 24, and 36 h using the NanoDropTM OneC (Thermo Fisher). For each evaluated time, 1×10^6 cells were collected and fixed with 4% Formaldehyde and 1% Phosphate-buffered saline (PBS) for 30 min at room temperature and stored at 4 °C until use. Analysis of the cytometer occurred on the same day of fixation or within two days. Each assay was conducted using technical and biological triplicates, analyzing 25,000 events. Data was analyzed using the FlowJo[™] v.10.9.0 software (BD Biosciences). Data were validated by FlowAI⁶² to select events with SSC-A/FSC-A patterns compatible with viable cells, then cells with FITC-A channel. Once these analysis gates were determined, all samples were selected and processed under the same gates in the same workspace.

Electron microscopy

The GSH cell lines were analyzed under a scanning electron microscope (SEM) to visualize possible cell shape and surface changes. The cells were fixed with Karnovsky fixative solution (2% glutaraldehyde, 2% paraformaldehyde, and 3% sucrose in sodium cacodylate buffer 0.1 M pH 7.2) overnight at 4 °C. Cell cells were washed with 0.1 M sodium cacodylate buffer, pH 7.2 the following day. A drop of the cells was seeded on 18×18 mm coverslips covered with Poly-D-Lysine for 3 min. After adhesion, the coverslips with cells were incubated in vapor of sodium tetroxide 2% osmium for 30 min and then washed with distilled water. Dehydration was carried out in

increasing series with acetone (50–100%) and, finally, drying to a critical point using CPD 030 (Balzers) and SCD 500 metalization (Leica), to be analyzed in a JSM 7001 F scanning electron microscope (15 kV) (Jeol).

Total RNA extraction

For each of the GSH cell lines and the control BY4741, one clone was selected to extract total RNA in biological replicates performed on independent days. The strains were cultivated in 20 mL of YPD within 125 mL Erlenmeyers for 16 h. Afterward, 2 mL of each cultivation was employed for total RNA extraction with the RNeasy Plus Micro Kit (Qiagen), following the manufacturer's instructions. The extracted RNAs were quantified using the Qubit 2.0 Fluorometer (Thermo Fisher) and a quality check using electrophoresis and the 2100 Bioanalyzer (Agilent). Finally, a purity analysis was conducted with NanoDropTM OneC (Thermo Fisher) to detect possible DNA contamination.

RT-qPCR

For the RT-qPCR, the cDNA library was assembled with the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Random Primers and 2 μ g of each RNA sample as templates were used, heeding the kit protocol. The cDNA was diluted 10x, and 2 μ L of this dilution was used for each qPCR reaction in a final volume of 10 μ L. A negative control was included for each gene analyzed (NTCs). qPCR reactions were performed using Syber Green Master Mix (Thermo Fisher) following the manufacturer's instructions in the QuantStudio 3 Real-Time System (Thermo Fisher). Replicates were run on the same plate for each sample, and to avoid systematic differences within the PCR procedure, samples were randomly distributed on the plate. Genes *ACT1*, *ALG6*, and *TAF10* were used as internal controls to normalize the data⁵⁹. The Ct value was used to estimate the specificity of the amplified products. The primers used and cycling conditions are listed in (Table S8).

RNA-Seq sequencing and data processing

RNA-Seq analysis of each of the GSH cell lines and BY4741 was performed by GenOne Biotechnologies (Brazil) using the NovaSeq 6000 Sequencing System (Illumina). Supplementary Table S9 presents the total count for each yeast cell line analyzed in this study. All analyses were performed using the reference genome of S. cerevisiae S288C (R64.5.1).

For sequence quality control, the fastp v0.23.4 software was used for adapter removal and low-quality sequencing trimming using Q30 as the quality threshold⁶³. Sequence mapping against the reference genome was done with the STAR 2.7.11b software⁶⁴, and the software htseq-count 2.0.3⁶⁵ was used to estimate the transcript counts. Differential expression estimation was performed in the R environment with the libraries edgeR 4.2.0⁶⁶ and DESeq2 1.44.0⁶⁷. The samples were filtered with a value of statistics p < 0.05 and adjusted to control the false discovery rate (FDR) the Benjamini-Hochberg method (R package stats 3.6.2) by setting FDR < 0.001. The data was visualized (Heatmaps and PCAs) using the ClustVis library 0.7.7⁶⁸ and Volcanos Plots with ggplot 2 3.5.1⁶⁹, with a limit of log2FC±0.6 to highlight the significant differences. We used pheatmap 1.0.12⁷⁰ for visualization and plotting.

Ontology enrichment analysis

Gene ontology enrichment analysis of the set of differentially expressed genes shared among all GSHs lineages was carried out using the program g: $Profiler^{71}$, with a significance threshold of FDR < 0.01.

Amylase enzyme activity

The transformed strains and BY4741 were grown in YPD medium containing 0.4% of Asp-Glu buffer for 72 h (with 2% glucose feeds every 24 h) at 28 °C and 200 rpm. 1 mL of each culture was centrifuged at 4,000 rpm for 2 min, and the supernatant was used in the enzyme activity assay. Enzyme activity was measured using a variation of the starch-iodine method for quantifying soluble starch as the remaining substrate after amylase activity⁷². 100 μ L of a 0.5% starch solution was heated at 40 °C along with 40 μ L of 0.5 M sodium acetate buffer. 60 μ L of the supernatant of each sample was added to the mixture and incubated at 10–30 min at 40 °C. After incubation, 200 μ L of 1 M acetic acid was added to each reaction, along with 200 μ L of an iodine solution (0.2% I2, 2% KI). 4.4 mL of water were added to each reaction and the results were read in a Spectramax M5 Spectrophotometer at 660 nm. A standard curve of starch concentration was constructed, ranging from 0.1 to 0.5 mg, and one enzyme unit was defined as the amount of enzyme necessary to hydrolyze 0.1 mg of starch per minute.

Statistical analysis

The candidate gene relative expression analysis was conducted using the Biogazelle qBase+3.0 software⁷³. The normalization of gene expression was carried out using the most stable reference genes identified by the geNorm⁷⁴. geNorm was used to compute the paired variation (Vn/n+1) between each reference gene, determining the optimal number of reference genes required for accurate normalization. The relative expression results were further analyzed using the REST© 2009⁷⁵. GraphPad Prism (version 9.0) was employed to perform a two-way ANOVA with a 95% confidence interval for the differences and a significance level of p < 0.05 between the groups.

Data availability

All data are available in the main text or the supplementary materials. Original photos of agarose gels and SCN files (https://figshare.com/s/c2d9bcc901d4114e11e8).Sanger sequencing (https://figshare.com/s/fc3b89fe7077e dac4e2b). Flow cytometry (https://figshare.com/s/063c0db3f6fd7e6c7f97). List of pGSHs - SHIP Output (https://figshare.com/s/7fcfade8b77f4d41140c). SHIP Code (www.github.com/MCLeitao/Ship). RNA-Seq data deposi

ted in the Sequence Read Archive (SRA) under the BioProject ID PRJNA1156748 (https://www.ncbi.nlm.nih.g ov/bioproject/PRJNA1156748).

Received: 6 November 2024; Accepted: 19 February 2025 Published online: 28 February 2025

References

- Sadelain, M., Papapetrou, E. P. & Bushman, F. D. Safe harbours for the integration of new DNA in the human genome. Nat. Rev. Cancer 12, 51–58 (2012).
- Arras, S. D. M., Chitty, J. L., Blake, K. L., Schulz, B. L. & Fraser, J. A. A genomic safe Haven for mutant complementation in cryptococcus neoformans. *PLoS One* 10, 1–16 (2015).
- 3. Kong, S., Yu, W., Gao, N., Zhai, X. & Zhou, Y. J. Expanding the neutral sites for integrated gene expression in *Saccharomyces cerevisiae*. *FEMS Microbiol. Lett.* **369**, fnac081 (2022).
- 4. Samulski, R. et al. Targeted integration of adeno-associated virus (AAV) into human chromosome 19. *EMBO J.* **10**, 3941–3950 (1991).
- Liu, R. et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86, 367–377 (1996).
- Irion, S. et al. Identification and targeting of the ROSA26 locus in human embryonic stem cells. Nat. Biotechnol. 25, 1477–1482 (2007).
- DeKelver, R. C. et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nucleasedriven transgenesis into a safe harbor locus in the human genome. *Genome Res.* 20, 1133–1142 (2010).
- 8. González, F. et al. An iCRISPR platform for rapid, multiplexable, and inducible genome editing in human pluripotent stem cells. *Cell. Stem Cell.* **15**, 215–226 (2014).
- 9. Jakočiunas, T. et al. CasEMBLR: Cas9-facilitated multiloci genomic integration of in vivo assembled DNA parts in Saccharomyces cerevisiae. ACS Synth. Biol. 4, 1126–1134 (2015).
- 10. Durrant, M. G. et al. Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat. Biotechnol.* **41**, 488–499 (2023).
- 11. Sun, C. et al. Precise integration of large DNA sequences in plant genomes using primeroot editors. *Nat. Biotechnol.* https://doi.or g/10.1038/s41587-023-01769-w (2023).
- 12. Ittiprasert, W. et al. Targeted insertion and reporter transgene activity at a gene safe harbor of the human blood fluke, Schistosoma mansoni. *Cell. Rep. Methods* **3**, 100535–100535 (2023).
- 13. Liu, T. et al. Construction of ajmalicine and sanguinarine de Novo biosynthetic pathways using stable integration sites in yeast. *Biotechnol. Bioeng.* **119**, 1314–1326 (2022).
- 14. Liu, Q. et al. CRISPR-Cas9-mediated genomic multiloci integration in Pichia pastoris. Microb. Cell. Factories 18, 144 (2019).
- 15. Pham, T., Xie, X. & Lin, X. An intergenic safe Haven region in Aspergillus fumigatus. Med. Mycol. 58, 1178-1186 (2020).
- 16. Pavani, G. Targeted gene delivery: where to land. Front. Genome Ed. 2 (2020).
- 17. Aznauryan, E. et al. Discovery and validation of human genomic safe harbor sites for gene and cell therapies. *Cell. Rep. Methods* **2** (2022).
- Shrestha, D. et al. Genomics and epigenetics guided identification of tissue-specific genomic safe harbors. *Genome Biol.* 23 (2022).
 Boob, A. G. et al. CRISPR-COPIES: an in silico platform for discovery of neutral integration sites for CRISPR/Cas-facilitated gene integration. *Nucleic Acids Res.* 52, e30 (2024).
- Charoensawan, V., Wilson, D. & Teichmann, S. A. Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Res. 38, 7364–7377 (2010).
- Weirauch, M. T. & Hughes, T. R. A Catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution in a handbook of transcription factors. (eds T. R. Hughes) 25–73. (Springer, 2011). https://doi.org/10.1007/978-90-4 81-9069-0_3
- 22. Goffeau, A. et al. Life with 6000 genes. Science 274, 546-567 (1996).
- Wong, E. D. et al. Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources. Genetics 224, iyac191 (2023).
- 24. Wang, H. & Helin, K. Roles of H3K4 methylation in biology and disease. Trends Cell. Biol. 0 (2024).
- Barton, A. B., Pekosz, M. R., Kurvathi, R. S. & Kaback, D. B. Meiotic recombination at the ends of chromosomes in *Saccharomyces cerevisiae*. *Genetics* 179, 1221–1235 (2008).
- Blitzblau, H. G., Bell, G. W., Rodriguez, J., Bell, S. P. & Hochwagen, A. Mapping of meiotic single-stranded DNA reveals doublestrand-break hotspots near centromeres and telomeres. *Curr. Biol.* 17, 2003–2012 (2007).
- Brown, C. A., Murray, A. W. & Verstrepen, K. J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* 20, 895–903 (2010).
- Lafontaine, I., Fischer, G., Talla, E. & Dujon, B. Gene relics in the genome of the yeast Saccharomyces cerevisiae. Gene 335, 1–17 (2004).
- 29. Ljungdahl, P. O. & Daignan-Fornier, B. Regulation of amino acid, nucleotide, and phosphate metabolism in Saccharomyces cerevisiae. Genetics 190, 885–929 (2012).
- 30. Lacroute, F. Regulation of pyrimidine biosynthesis in *Saccharomyces cerevisiae*. J. Bacteriol. **95**, 824–832 (1968).
- 31. Pronk, J. T. Auxotrophic yeast strains in fundamental and applied research. Appl. Environ. Microbiol. 68, 2095–2100 (2002).
- 32. Yan, C. et al. Auxotrophs compromise cell growth and fatty acid production in *Saccharomyces cerevisiae*. *Biotechnol. J.* **18**, 2200510 (2023).
- Alam, M. T. et al. The metabolic background is a global player in Saccharomyces gene expression epistasis. Nat. Microbiol. 1, 1–10 (2016).
- 34. Perruca-Foncillas, R., Davidsson, J., Carlquist, M. & Gorwa-Grauslund, M. F. Assessment of fluorescent protein candidates for multi-color flow cytometry analysis of *Saccharomyces cerevisiae*. *Biotechnol. Rep.* 34 (2022).
- 35. de Moraes, L. M. P., Astolfi-filho, S. & Oliver, S. G. Development of yeast strains for the efficient utilisation of starch: evaluation of constructs that express α-amylase and glucoamylase separately or as bifunctional fusion proteins. *Appl. Microbiol. Biotechnol.* 43, 1067–1076 (1995).
- 36. Manente, M. & Ghislain, M. The lipid-translocating exporter family and membrane phospholipid homeostasis in yeast. FEMS Yeast Res. 9, 673-687 (2009).
- 37. Heymann, P., Ernst, J. F. & Winkelmann, G. Identification and substrate specificity of a ferrichrome-type siderophore transporter (Arn1p) in *Saccharomyces cerevisiae*. *FEMS Microbiol. Lett.* **186**, 221–227 (2000).
- Despons, L., Wirth, B., Louis, V. L., Potier, S. & Souciet, J. L. An evolutionary scenario for one of the largest yeast gene families. *Trends Genet.* 22, 10–15 (2006).
- Avramova, Z. Transcriptional 'memory' of a stress: transient chromatin and memory (epigenetic) marks at stress-response genes. *Plant. J.* 83, 149–159 (2015).
- 40. Choi, J. K. & Kim, Y. J. Epigenetic regulation and the variability of gene expression. Nat. Genet. 40, 141-147 (2008).

- 41. Yao, Y. et al. Advances in approaches to study chromatin-mediated epigenetic memory. ACS Synth. Biol. 11, 16-25 (2022).
- 42. Zhang, Q. et al. Stochastic expression and epigenetic memory at the yeast HO promoter. *Proc. Natl. Acad. Sci.* 110, 14012–14017 (2013).
- 43. Mu, X., Murakami, H., Mohibullah, N. & Keeney, S. Chromosome-autonomous feedback down-regulates meiotic DNA break competence upon synaptonemal complex formation. *Genes Dev.* **34**, 1605–1618 (2020).
- 44. Hasunuma, T. et al. Development of a GIN11/FRT-based multiple-gene integration technique affording inhibitor-tolerant, hemicellulolytic, xylose-utilizing abilities to industrial Saccharomyces cerevisiae strains for ethanol production from undetoxified lignocellulosic hemicelluloses. *Microb. Cell. Factor.* **13** (2014).
- Liu, Z. et al. Engineering of a novel cellulose-adherent cellulolytic Saccharomyces cerevisiae for cellulosic biofuel production. Sci. Rep. 6 (2016).
- 46. Nassar, L. R. et al. The UCSC genome browser database: 2023 update. Nucleic Acids Res. 51, D1188–D1195 (2023).
- 47. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. Nucleic Acids Res. 53 D609-D617, (2024).
- Goldfarb, T. et al. Reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. 53, D243-D257 (2025).
- 49. The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. Genetics 224, (2023).
- 50. Leinonen, R. et al. UniProt archive. *Bioinformatics* 20, D3236–D3237, (2004).
- 51. Martin, F. J. et al. Ensembl 2023. Nucleic Acids Res. 51, D933-D941 (2023).
- 52. Brachmann, C. B. et al. Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115–132 (1998).
- Tsutsui, H., Karasawa, S., Okamura, Y. & Miyawaki, A. Improving membrane voltage measurements using FRET with new fluorescent proteins. Nat. Methods 5, 683–685 (2008).
- Kaishima, M., Ishii, J., Matsuno, T., Fukuda, N. & Kondo, A. Expression of varied GFPs in Saccharomyces cerevisiae: codon optimization yields stronger than expected expression and fluorescence intensity. Sci. Rep. 6 (2016).
- 55. Olesen, K. et al. Yeast functional analysis report the pYC plasmids, a series of cassette-based yeast plasmid vectors providing means of counter-selection. *Yeast* **16**, 1035–1043 (2000).
- 56. Becker, D. M. & Guarente, L. High-efficiency transformation of yeast by electroporation. (1984).
- Gietz, D. R., Schiestls, R. H., Willems, A. R. & Woods, R. A. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/ PEG procedure. (1995).
- 58. Annaluru, N. et al. Total synthesis of a functional designer eukaryotic chromosome. Science 344, 55-58 (2014).
- Teste, M. A., Duquenne, M., François, J. M. & Parrou, J. L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in Saccharomyces cerevisiae. BMC Mol. Biol. 10, 99–99 (2009).
- 60. Critical Factors for Successful Real-Time PCR (EN). QIAGEN. https://www.qiagen.com/us/resources/resourcedetail?id=f7efb4 f4-fbcf-4b25-9315-c4702414e8d6&lang=en
- Janner, C. R., Brito, A. L. P., Moraes, L. M. P. & Reis, V. C. Torres, pPCV, a versatile vector for cloning PCR products. SpringerPlus 2, 441 (2013).
- Monaco, G., Chen, H., Poidinger, M., Chen, J. & De Magalhães, J. P. Larbi, FlowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* 32, 2473–2480 (2016).
- 63. Chen, S., Zhou, Y. & Chen, Y. Gu, Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018).
- 64. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21 (2013).
- 65. Anders, S., Pyl, P. T. & Huber, W. HTSeq—A python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).
- 67. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).
- 68. Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.* **43**, W566–W570 (2015).
- 69. Wickham, H. Ggplot2. http://link.springer.com/10.1007/978-3-319-24277-4)Use (Springer International Publishing, 2016).
- 70. Kolde, R. Pheatmap: pretty heatmaps. R Package Version 1, 726 (2012).
- Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P. & Vilo, J. Peterson, g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 51, W207–W212 (2023).
- 72. Fuwa, H. A new method for microdetermination of amylase activity by the use of amylose as the substrate. J. Biochem. 41, 583–603 (1954).
- 73. Biogazelle qBasePlus. Get the software safely and easily. Softw. Informer https://biogazelle-qbaseplus.software.informer.com/ (2024).
- 74. Vandesompele, J. et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research00341 (2002).
- 75. Pfaffl, M. W., Horgan, G. W. & Dempfle, L. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res.* **30**, e36 (2002).
- 76. Weitz, E. ideogram. https://github.com/eweitz/ideogram (2023).
- Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).

Acknowledgements

We thank support agencies, PIBIC/DPG (Programa de iniciação científica / Decanato de Pesquisa e Inovação e Decanato de Pós-Graduação, from University of Brasilia, Brazil), CDT/NUPITEC (Centro de Apoio ao Desenvolvimento Tecnológico, from University of Brasilia, Brazil) to assist with the software registration. We are deeply grateful to Dr. Marie F. Gorwa-Grauslund and Dr. Raquel Perruca Foncillas for the pRP009 plasmid. Finally, we thank Kamilla Tosetto who provided technical support.

Author contributions

Conceptualization: MCL, CMC. Data curation: MCL, GJP, CMC. Formal Analysis: MCL, LCP, PSQ, TGG, RCT, GJP. Funding acquisition: CMC. Investigation: MCL, LSC, LCP, PFQ, TGG, RVA, KLP. Methodology: MCL, LCP, TGG, RVA, SNB, LPM, RCT, FGT, GJP, CMC. Project administration: CMC. Resources: RVA, SNB, VCR, LPM, RCT, LGB, FGT, GJP, CMC. Software: MCL, GJP. Supervision: CMC, GJP. Validation: CMC, GJP. Visualization: MCL, LCP, PSQ, TGG, KLP, GJP, CMC. Writing—original draft: MCL, CMC, GJP. Writing—review & editing: MCL, LSC, LCP, PFQ, TGG, RVA, AAP, KLP, SNB, VCR, LPM, RCT, LGB, FGT, GJP, CMC.

Declarations

Competing interests

The authors declare software registration (Process Number: BR512023002017-6).

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1 0.1038/s41598-025-91249-9.

Correspondence and requests for materials should be addressed to G.J.P. or C.M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025