



Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular

MATHEUS DE CASTRO LEITÃO

**IDENTIFICAÇÃO DE REGIÕES GENÔMICAS SEGURAS PARA APLICAÇÕES DE
BIOLOGIA SINTÉTICA**

Orientador: Prof. Dr. Georgios Joannis Pappas Jr
Coorientadora: Profa. Dra. Cíntia Marques Coelho

Brasília-DF

2024

MATHEUS DE CASTRO LEITÃO

**IDENTIFICAÇÃO DE REGIÕES GENÔMICAS SEGURAS PARA APLICAÇÕES DE
BIOLOGIA SINTÉTICA**

Dissertação de mestrado apresentada ao
Programa de Pós-Graduação em Biologia
Molecular da Universidade de Brasília como
requisito para obtenção do título de mestre.

Orientador: Prof. Dr. Georgios J. Pappas Jr

Coorientadora: Profa. Dra. Cíntia M. Coelho

Brasília-DF

2024

FICHA CATALOGRÁFICA

MATHEUS DE CASTRO LEITÃO

**IDENTIFICAÇÃO DE REGIÕES GENÔMICAS SEGURAS PARA APLICAÇÕES DE
BIOLOGIA SINTÉTICA**

Dissertação de mestrado apresentada ao
Programa de Pós-Graduação em Biologia
Molecular da Universidade de Brasília como
requisito para obtenção do título de mestre.

Data de aprovação: 29 / 01 / 2024

Prof. Dr. Georgios Joannis Pappas Jr
Universidade de Brasília
Presidente da Banca

Prof. Dr. Marcelo de Macedo Brigido
Universidade de Brasília
Membro Titular Interno

Dr. João Ricardo Moreira de Almeida
Embrapa Agroenergia
Membro Titular Externo

Profª. Dra. Ildinete Silva Pereira
Universidade de Brasília
Membro Suplente

Aos meus pais, Christiani e Márcio Leitão,
que são meus maiores incentivadores,
referência e fonte de inspiração.

"What I cannot create, I do not understand."

Richard Feynman, 1988

AGRADECIMENTOS

Agradeço a todos que me ajudaram de uma forma ou de outra em minha carreira acadêmica, contribuindo para o desenvolvimento e a conclusão deste trabalho.

RESUMO

Desde o início da engenharia genética, a inserção de genes exógenos em células hospedeiras apresentou desafios, seja pelo aumento do tamanho das inserções, seja pelo desafio de direcionamento preciso para integração em regiões do genoma permissivas para transcrição. Com o advento da Biologia Sintética, há um aumento da complexidade para a inserção, não apenas de um gene, mas de diversos genes componentes de vias metabólicas e circuitos genéticos, além da necessidade de estudos mais refinados de organização, estruturação e regulação genômica. A fim de evitar integração aleatória no genoma que possa gerar expressões instáveis ou fenótipos imprevisíveis, é demandada integração específica sem modificação da expressão de genes endógenos codificadores de proteínas essenciais. Uma abordagem comum é direcionar as inserções para regiões intragênicas ou intergênicas capazes de acomodar a expressão esperada do inserto sem alterações adversas ao hospedeiro, denominadas Porto Genômico Seguro (do inglês, *Genomic Safe Harbor*, GSH). GSHs foram encontrados em humanos, camundongos, macacos e *Cryptococcus*, porém em sua maioria estão em regiões com alta densidade gênica e no caso de humanos próximos a genes relacionados ao câncer. Com o objetivo de expandir o número de GSHs disponíveis em diferentes organismos, a presente proposta visou o desenvolvimento de um programa de computador de identificação de prováveis GSHs com uma abordagem otimizada seguindo critérios definidos para refinar a busca por novas regiões extragênicas de inserção em três organismos eucarióticos: *Homo sapiens*, *Mus musculus* e *Saccharomyces cerevisiae*. Esse *software*, SHIP (do inglês, *Safe Harbor Identification Program*), foi desenvolvido na linguagem Python e utilizou como entrada arquivos de anotação gênica do organismo estudado aliado a informações de características regulatórias de expressão gênica, como potencial de metilação do DNA, locais de ligação de fator de transcrição e sítios de alterações de histonas, identificando 6 pGSHs (GSHs putativos) em *S. cerevisiae*, 16 em *H. sapiens* e 11 em *M. musculus*. Adicionalmente, essa proposta validou *in vivo* 5 GSHs identificados pelo *software* SHIP no organismo modelo *S. cerevisiae*. Desta maneira, o *software* SHIP é o primeiro algoritmo, disponível e validado experimentalmente, capaz de identificar pGSHs em qualquer organismo Eucarioto com anotação genômica.

Palavras-chave: Porto Seguro Genômico, Python, *S. cerevisiae*, Biologia Sintética

ABSTRACT

Since the beginning of genetic engineering, the insertion of exogenous genes into host cells has presented challenges, whether due to the increase in the size of the insertions or the challenge of precise targeting for integration into regions of the genome that are permissive for transcription. With the advent of Synthetic Biology, there has been an increase in the complexity of inserting not just one gene, but several genes that are components of metabolic pathways and genetic circuits, as well as the need for more refined studies of genomic organization, structuring, and regulation. Specific integration without modifying the expression of endogenous genes coding for essential proteins is required to avoid random integration into the genome that could generate unstable expression or unpredictable phenotypes. A common approach is to target insertions to intragenic or intergenic regions capable of accommodating the expected expression of the insert without adverse changes to the host, known as the Genomic Safe Harbor (GSH). GSHs have been found in humans, mice, monkeys, and *Cryptococcus*, but most of them are in regions with high gene density and, in the case of humans, close to cancer-related genes. To expand the number of GSHs available in different organisms, this proposal aimed to develop a program to identify probable GSHs with an optimized approach following defined criteria to refine the search for new extragenic insertion regions in three eukaryotic organisms: human, mouse, and *Saccharomyces cerevisiae*. This program, SHIP (Safe Harbor Identification Program), was developed in the Python language and used gene annotation files from the studied organism as input, together with information on gene expression regulatory characteristics, such as DNA methylation potential, transcription factor binding sites, and histone alteration sites, identifying six pGSHs in *Saccharomyces cerevisiae*, 16 in *Homo sapiens* and 11 in *Mus musculus*. In addition, this proposal validated *in vivo* 5 GSHs identified by the SHIP program in the model organism *S. cerevisiae*. Thus, the SHIP program is the first experimentally validated algorithm capable of identifying pGSHs in any eukaryotic organism with genomic annotation.

Keywords: Genomic Safe Harbour, Python, *S. cerevisiae*, Synthetic Biology

LISTA DE FIGURAS

Figura 1 -	Uma breve história da biologia sintética.....	03
Figura 2 -	GSHs intragênicos comumente utilizados em estudos com células humanas.....	06
Figura 3 -	Esquema representativo do <i>software</i> SHIP.....	12
Figura 4 -	Visão geral da estratégia do sistema de identificação e validação funcional de porto seguro genômico em eucariotos.....	14
Figura 5 -	Representação esquemática da estratégia de montagem dos BioBricks utilizados para transformação genética de <i>S. cerevisiae</i>	17
Figura 6 -	Histograma da distribuição genômica das regiões intergênicas e ideograma da localização genômica dos pGSHs identificados.....	26
Figura 7 -	Identificação dos clones com inserção correta na região de cada pGSH por análise de PCR.....	31
Figura 8 -	Sequenciamento de Sanger dos clones com inserção correta na região de cada pGSH.....	31
Figura 9 -	Teste de estabilidade genômica de cada pGSH após 10 dias de cultivo contínuo.....	33
Figura 10 -	Curva de crescimento para cada linhagem pGSH.....	34
Figura 11 -	Validação funcional das linhagens celulares por citometria de fluxo realizada em triplicata técnica e biológicas em três dias independentes....	35
Figura 12 -	Dinâmica de expressão dos genes vizinhos das linhagens celulares pGSHs por RT-qPCR.....	37
Figura 13 -	Dinâmica de expressão genômica das linhagens celulares pGSHs por RNA-Seq.....	39

Figura 14 -	Diagramas de Venn do número de genes diferencialmente expressos únicos e compartilhados entre as linhagens pGSHs.....	40
Figura 15 -	Ideograma mostrando a localização de regiões genômicas com genes diferencialmente expressos na linhagem S2U.....	41
Figura 16 -	Análise de Enriquecimento de ontologia gênica dos genes compartilhados diferencialmente regulados.....	43
Figura 17 -	Vulcanos <i>Plots</i> da expressão diferencial das linhagens celulares pGSHs por RNA-Seq.....	44
Figura 18 -	Gráfico <i>Box Plot</i> da alteração de $\log_2FC$ dos genes diferencialmente expressos nas construções pGSH.....	46
Figura 19 -	Balloon plot da alteração de $\log_2FC$ dos 10 genes diferencialmente regulados, em termos do valor de log fold change, nas construções pGSH.....	48

LISTA DE TABELAS

Tabela 1 -	Critérios e justificativas para escolha dos potenciais GSHs (pGSHs) avaliados pelo <i>software</i> SHIP.....	13
Tabela 2 -	BioBricks utilizados para validação funcional do <i>software</i> SHIP.....	15
Tabela 3 -	Oligonucleotídeos utilizados para montagem dos BioBricks.....	15
Tabela 4 -	Ciclo utilizado para montagem dos BioBricks com a <i>Platinum™ Taq DNA Polymerase High Fidelity</i>	16
Tabela 5 -	Oligonucleotídeos genômicos utilizados para identificação dos clones com inserção correta dos BioBricks U _k G e URA3 em cada pGSH.....	18
Tabela 6 -	Ciclo utilizado para identificação dos clones com inserção correta com a <i>GoTaq® G2 DNA Polymerase</i>	19
Tabela 7 -	Oligonucleotídeos para análise de expressão gênica por RT-qPCR.....	22
Tabela 8 -	Ciclo utilizado para para análise de expressão gênica por RT-qPCR com o <i>Syber Green Master Mix</i>	23
Tabela 9 -	Coordenadas genômicas, genes vizinhos e tamanho dos pGSH identificadas pelo SHIP em <i>S. cerevisiae</i>	29
Tabela 10 -	Tabela 10. Resumo quantitativo da expressão diferencial significativa dos genes vizinhos nas linhagens pGSHs.....	38
Tabela 11 -	Descrição e características dos genes diferencialmente expresso mais próximo com $\log_2FC > 2$ de cada pGSH.....	42
Tabela 12 -	Genes caracterizados mais diferencialmente expressos em todos os experimentos.....	47

LISTA DE SIGLAS

a.C	Antes de Cristo
AAVS1	Vírus adeno-associado 1
API	Interface de Programação de Aplicação
Cas9	Proteína 9 Associada ao CRISPR
cDNA	Ácido desoxirribonucleico complementar
CRISPR	Repetições palindrômicas curtas interespaçadas e regularmente agrupadas
DBTL	<i>Design - Build - Test - Learn</i>
DHS	Sítio hipersensível à DNase I
DNA	Ácido desoxirribonucleico
DSB	Quebras de fita dupla de Ácido desoxirribonucleico
EST	Etiqueta de sequência expressa
EUROSCARF	Arquivo Europeu de <i>Saccharomyces cerevisiae</i> para análise funcional
FITC	Isotiocianato de fluoresceína
FSC-A	Área de dispersão frontal
FSC-H	Altura de dispersão direta
GFF	Formato de característica geral
GRAS	Geralmente reconhecido como seguro
GSH	Porto genômico seguro
HIV	Vírus da imunodeficiência humana
INPI	Instituto Nacional da Propriedade Industrial
LiAc	Acetato de lítio
lncRNA	Ácido ribonucleico não codante
LTR	Repetição terminal longa
miRNA	Ácido ribonucleico micro

mRNA	Ácido ribonucleico mensageiro
NCBI	Centro Nacional de Informações sobre Biotecnologia
OD ₆₀₀	Densidade óptica em um comprimento de onda de 600 nm
ORF	Fase de leitura aberta
pb	Pares de Base
PBS	Solução salina tamponada com fosfato
PCR	Reação em Cadeia da Polimerase
PEG	Polietileno glicol
pGSH	Potencial porto genômico seguro
pMEIs	Inserções de elementos móveis polimórficos
rDNA	DNA ribossômico
RNA	Ácido ribonucleico
RNA-Seq	Sequenciamento de larga escala de ácido ribonucleico
RT-qPCR	Transcrição reversa seguida de reação em cadeia da polimerase quantitativa em tempo real
SC2.0	Projeto Genoma Sintética de Levedura
SGD	Banco de dados do genoma de <i>Saccharomyces</i>
SHIP	<i>Safe Harbor Identification Program</i>
SS-DNA	Ácido desoxirribonucleico de fita simples
SSC-A	Área de dispersão lateral
tRNA	Ácido ribonucleico transportador
Ty	Retrotransposons de levedura
UCSC	Universidade da Califórnia Santa Cruz
ymUKG1	UKG1 monomérico otimizado códon para <i>S. cerevisiae</i>
YPD	Meio de extrato de levedura, peptona e dextrose

SUMÁRIO

1	INTRODUÇÃO.....	1
1.1	Justificativa.....	1
1.2	Biologia Sintética.....	2
1.3	Portos Seguros Genômicos.....	4
1.4	Saccharomyces cerevisiae.....	8
2	OBJETIVOS.....	10
2.1	Objetivo geral.....	10
2.2	Objetivos específicos.....	10
3	METODOLOGIA.....	11
3.1	Desenvolvimento do software SHIP.....	11
3.2	Validação funcional do software SHIP.....	14
3.2.1	Validação experimental in vivo.....	14
3.2.2	Montagem dos BioBricks.....	14
3.2.3	Purificação e concentração dos BioBricks.....	17
3.2.4	Transformação genética.....	17
3.2.5	Extração de DNA genômico.....	18
3.2.6	Identificação dos clones com inserção correta.....	18
3.2.7	Sequenciamento.....	19
3.2.8	Estabilidade genômica dos pGSH.....	20
3.2.9	Análise do acúmulo das proteínas repórteres e taxa de crescimento.....	20
3.2.10	Extração de RNA total.....	21
3.2.11	Quantificação de transcritos via RT-qPCR.....	21
3.2.12	RNA-Seq.....	23
4	RESULTADOS.....	25
4.1	Implementação do SHIP e avaliação geral dos critérios de seleção de GSHs.....	25
4.2	SHIP identifica potenciais GSHs em genomas de eucariotos usando dados de características gerais.....	25
4.3	Geração de linhagens individuais de S. cerevisiae contendo os BioBricks em cada GSH.....	30
4.4	GSHs identificados apresentam estabilidade genômica ao longo de 10 dias sem afetar o crescimento da levedura.....	32
4.5	Validação funcional das linhagens celulares por citometria de fluxo.....	35
4.6	Análise da expressão dos genes vizinhos aos GSHs.....	36
4.7	Análise global de expressão gênica.....	38
4.7.1	Análise de enriquecimento de ontologia gênica dos genes diferencialmente regulados.....	43
4.7.2	Análise individual dos genes mais diferencialmente expressos.....	44
5	DISCUSSÃO.....	49
5.1	Pipelines semelhantes descritos na literatura.....	49

5.2	Software SHIP e dados da validação experimental dos pGSHs.....	50
5.3	pGSHs identificados pelo software SHIP já utilizados na literatura.....	52
5.4	Sumário crítico do contexto dos genes diferencialmente expressos do RNA-Seq.....	53
6	CONCLUSÕES E PERSPECTIVAS.....	54
	REFERÊNCIAS.....	56
	Apêndice A - Tabela com as descrição dos genes vizinhos dos pGSHs de <i>S. cerevisiae</i>	64
	Apêndice B - Tabela com as características genômicas dos pGSHs identificadas em <i>H. sapiens</i>	65
	Apêndice C - Tabela com as características genômicas dos pGSHs identificadas em <i>M. musculus</i>	66
	Apêndice D - Tabela com a lista dos pacotes requeridos para a execução do SHIP...	67
	Apêndice E - Ideograma mostrando os locais dos genes diferencialmente expressos	68
	Apêndice F - Tabela com a lista de genes diferencialmente regulados positivamente	70
	Apêndice G - Tabela com a lista de genes diferencialmente regulados negativamente	71
	Apêndice H - Registro do software SHIP pelo INPI.....	72

1. INTRODUÇÃO

1.1. Justificativa

Com o surgimento da tecnologia do DNA recombinante e da transformação genética, o local de inserção de genes de interesse (transgenes) surgiu como uma importante preocupação. Esse local de inserção deveria garantir a transcrição do transgene sem efeitos indesejáveis ao fenótipo ou ao genoma do organismo alvo. O advento da Biologia Sintética, que almeja inserção de vias metabólicas ou circuitos genéticos, demanda inserção simultânea de vários genes, o que reforça a preocupação inicial. Assim, faz-se necessário a identificação de locais de integração seguros no genoma (do inglês, *Genomic Safe Harbours*, GSHs).

Há, ainda hoje, uma grande atenção relacionada à expressão dos próprios transgenes, pois esses são majoritariamente inseridos em regiões selecionadas de forma empírica ou mesmo ao acaso. Por conta disso há uma necessidade de busca, entre os organismos transformados, por aqueles exibindo o padrão desejado de expressão dos transgenes, gerando a necessidade de considerável tamanho amostral para que essa seleção seja atingida, a depender do organismo alvo.

A maioria dos candidatos à GSHs usados atualmente foram identificados por local de inserção viral, análise de perda de função gênica ou por semelhança com GSHs de outros organismos (Sadelain; Papapetrou; Bushman, 2012). Estas regiões estão, em sua maioria, em locos intragênicos, o que traz importantes implicações com relação ao efeito da inserção de genes de interesse na expressão de genes próximos ao local de inserção. Observou-se que, mesmo nos estudos que elencaram critérios para identificação de GSHs no genoma humano, a adesão aos mesmos não pôde ser integral, além de não ter havido uma análise ampla da expressão gênica (Arras *et al.*, 2015; Pellenz *et al.*, 2019; Sadelain; Papapetrou; Bushman, 2012). Assim, apesar de terem sido denominadas *à priori* como GSHs, ainda persiste a ausência de avaliação ou avaliação parcial dos efeitos das inserções na expressão dos genes endógenos dos organismos geneticamente modificados.

Portanto, há necessidade de identificação de GSHs que sigam critérios específicos para os diversos organismos, respeitando as suas particularidades em termos de tamanho do genoma, incidência de elementos repetitivos e marcas epigenéticas. Além disso, a ruptura

bialélica, que tem probabilidade de ocorrer quando o direcionamento é mediado por endonucleases, pode resultar em fenótipos indesejados por mutação nula homozigota quando ocorre inserção em GSH localizado em região intragênica (Pellenz *et al.*, 2019; Sadelain; Papapetrou; Bushman, 2012). Levando em consideração o contexto apresentado, bem como a falta de descrição na literatura de estratégias consolidadas para realizar buscas sistemáticas em genomas por GSHs, torna-se relevante o desenvolvimento de abordagens computacionais para caracterizar e identificar estas regiões em genomas conhecidos, bem como derivar regras genéricas visando expandir a aplicação para outros organismos.

Para tanto, propõe-se o desenvolvimento de um *software* de computador, denominado SHIP (do inglês, *Safe Harbor Identification Program*), que visa a identificação de GSHs a partir de genomas sequenciados. O SHIP preencheria uma lacuna atual na literatura e serviria como uma ferramenta para guiar experimentos de biologia sintética a partir de critérios pré-definidos, baseado no ciclo de desenho, construção (síntese), teste (validação *in vivo*) e aprendizado (do inglês, *Design - Build - Test - Learn*, DBTL). Com isso, o presente trabalho pode contribuir para a identificação racional de GSHs e, no futuro, para o desenvolvimento orientado de organismos contendo circuitos genéticos complexos ou vias metabólicas completas, de forma segura e rápida.

1.2. Biologia Sintética

O desenvolvimento da engenharia genética baseou-se principalmente na tecnologia do DNA recombinante e na análise de Reação em Cadeia da Polimerase (do inglês, *Polymerase Chain Reaction*, PCR) das décadas de 1970 e 80 (Cameron; Bashor; Collins, 2014). Desde esse período, a expressão do transgene introduzido no organismo-alvo, sem que houvesse outras interferências na fisiologia celular, era o objetivo central da área. Com o advento da biologia sintética, que visa o desenho ou redesenho de sistemas biológicos para que o organismo alvo possa desempenhar uma nova função ou ser mais bem sucedido no desempenho de uma função já existente (Haseloff; Ajioka, 2009), tornou-se mais desafiador o alcance do objetivo central.

Seguindo a ideia de circuitos genéticos introduzida pela primeira vez com o estudo do operon da lactose por Jacob e Monod (1961), a biologia sintética iniciou a montagem de circuitos gênicos simples, como osciladores (Gardner; Cantor; Collins, 2000) e

repressiladores (Elowitz; Leibler, 2000), com perspectivas de evolução para circuitos mais complexos (Figura 1). Com isso, houve a necessidade de expansão da manipulação genética de características de interesse, migrando de monogênicas para poligênicas, como vias metabólicas completas e circuitos genéticos multimodais, as quais requerem a inserção de vários genes codificadores (Naseri; Koffas, 2020).

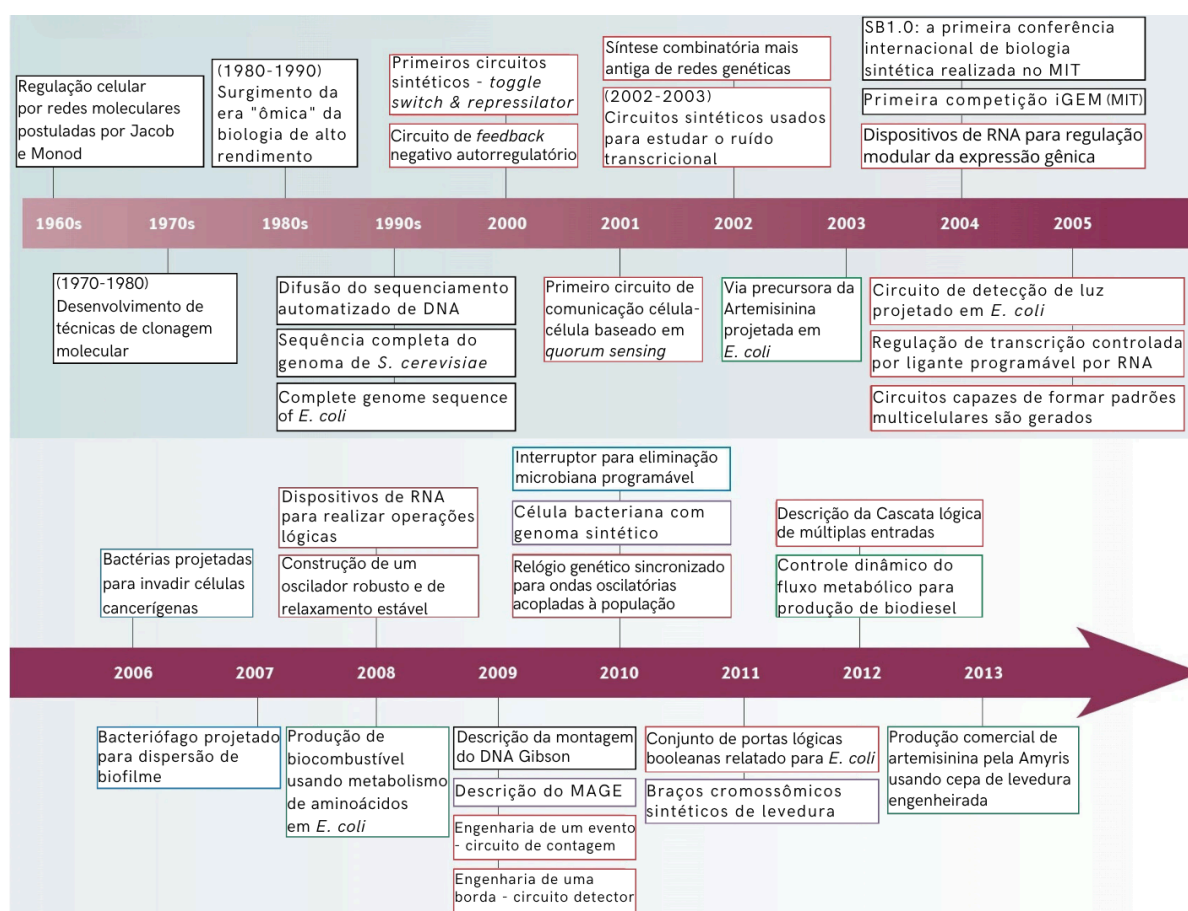


Figura 1 - Uma breve história da biologia sintética. Caixas pretas indicam marcos técnicos ou culturais; caixas vermelhas indicam engenharia de circuitos; biologia sintética em engenharia metabólica (verde); aplicações terapêuticas (azul); engenharia de genoma completo (roxo). Siglas: iGEM (do inglês, *International Genetically Engineered Machine*); MAGE (do inglês, *multiplex automated genome engineering*); SB1.0 (do inglês, *Synthetic Biology 1.0*). Retirado e adaptado de Cameron; Bashor; Collins, 2014.

Além disso, a biologia sintética também trouxe o desenvolvimento de ferramentas de edição gênica, como o sistema formado por Repetições palindrômicas curtas interespaçadas e regularmente agrupadas (do inglês, *Clustered Regularly Interspaced Short Palindromic Repeats*, CRISPR) com a Proteína Associada ao CRISPR (do inglês, *CRISPR-associated*

Protein 9, Cas9) (Jinek *et al.*, 2012). Esse sistema permite a introdução de genes em locais específicos no genoma alvo posicionado por complementaridade via um RNA-guia associado a nuclease Cas9, que quebra a dupla fita do DNA e induz o reparo por recombinação homóloga resultando na integração da sequência de DNA doador, ou seja, do transgene (Cong *et al.*, 2013).

A Biologia sintética também trouxe o conceito BioBrick que, utilizado pela primeira vez em 2003, consiste de uma peça individual de DNA (Knight, 2003). Apesar de simples, essa abstração, oriunda da engenharia, para modularizar partes biológicas traz uma sistematização para que tais "blocos de construção" sejam utilizados para projetar e montar sistemas genéticos e vias metabólicas mais refinadas e complexas (Stephanopoulos, 2012). A padronização e sistematização também contribuiu para disseminar e facilitar o uso de inúmeros BioBricks de maneira fácil e replicável entre diferentes grupos de pesquisa (Shetty; Endy; Knight, 2008) além de proporcionar o surgimento do Registro de Peças Biológicas Padrão (Galdzicki *et al.*, 2011), agora hospedado pelo grupo iGEM (do inglês, *International Genetically Engineered Machine*), e da Linguagem Aberta de Biologia Sintética (do inglês, *Synthetic Biology Open Language*, SBOL) (Galdzicki *et al.*, 2014).

Com o notável avanço das técnicas de edição gênica e da sistematização para a montagem de circuitos genéticos e vias metabólicas complexas dentro de células surge a necessidade de identificar locais de inserção que permitam a expressão dos transgenes, sem, contudo, interferir significativamente na fisiologia celular, evitando fenótipos não preditos.

1.3. Portos Seguros Genômicos

No contexto de contemplar as potencialidades da biologia sintética, vislumbra-se um cenário ideal onde gene(s) possam ser integrados em regiões específicas do genoma, sem, contudo, afetar a expressão de genes endógenos codificadores de proteínas essenciais. Uma abordagem comum para satisfazer este desafio seria direcionar os locais de integração para regiões intragênicas ou intergênicas, capazes de acomodar a expressão esperada do inserto, sem alterações adversas ao hospedeiro. Estes locos distintivos são denominados de Portos Seguros Genômicos (do inglês, *Genomic Safe Harbours*, GSHs) descritos em sua maioria para os organismos modelos como *H. sapiens* (Sadelain; Papapetrou; Bushman, 2012) e *Mus*

musculos (Clark; Dinsmore; Soriano, 2020) pela sua importância para estudos genéticos e clínicos.

Na literatura foram estabelecidos alguns critérios, principalmente em *H. sapiens*, para que regiões genômicas possam ser consideradas GSHs: (1) distanciamento de mais de 300 kb de genes relacionados ao desenvolvimento do câncer, de microRNAs e de RNAs não codificadores funcionais; (2) distanciamento de mais de 50 kb da região 5' de demais genes, qualquer origem de replicação e de regiões ultra-conservadas; (3) exclusão de regiões com baixa atividade transcricional e de regiões de número de cópia variável no genoma; (4) priorização de regiões genômicas de única cópia e com estrutura aberta de cromatina (Pellenz *et al.*, 2019; Sadelain; Papapetrou; Bushman, 2012).

Os GSHs mais comumente utilizados em estudos com células humanas são: (1) o lóco do sítio de integração do Vírus adeno-associado 1 (AAVS1) no gene PPP1R12C, cromossomo 19 (Kotin *et al.*, 1991; Samulski *et al.*, 1991); (2) o lóco CCR5 localizado no cromossomo 3 que codifica um dos principais receptores do vírus HIV (Liu *et al.*, 1996) e (3) o lóco ROSA26 também localizado no cromossomo 3, identificado no genoma humano por homologia de sequência a partir da identificação desse lóco em camundongo. No entanto, singularmente esses três lócos são encontrados em regiões com alta densidade gênica e próxima a genes relacionados ao câncer (Henckaerts; Linden, 2010; Sadelain; Papapetrou; Bushman, 2012), o que faz com que estes não cumpram os critérios acima (Figura 2). Além disso, nenhum estudo mais aprofundado foi realizado para verificar o impacto da inserção de transgenes na estrutura, regulação e funcionalidade genômica do organismo alvo.

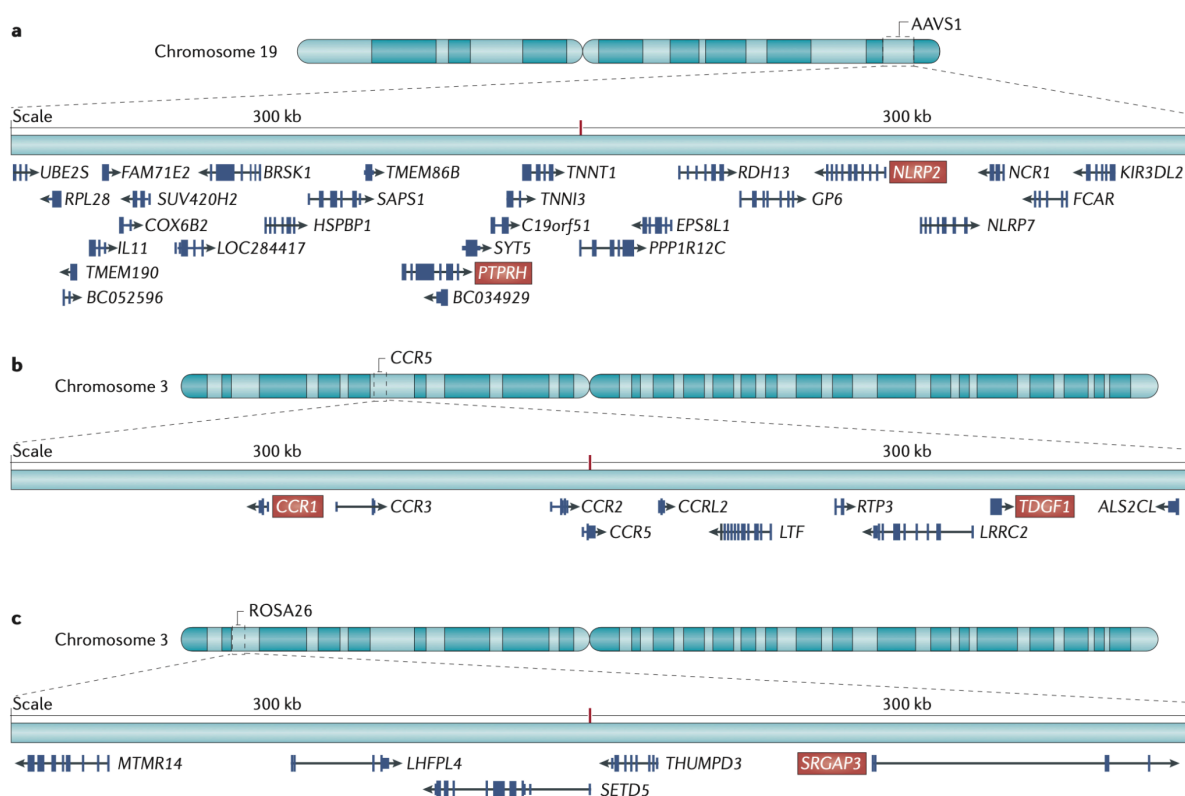


Figura 2 - GSHs intragênicos comumente utilizados em estudos com células humanas. Cada local (identificado por uma linha vermelha vertical) são mostrados para a) AAVS1 b) CCR5 e c) ROSA26. Os oncogenes estão ilustrados em caixas vermelhas. As setas indicam a direção da transcrição. Retirado de Sadelain; Papapetrou; Bushman, (2012).

Com extensa anotação genômica e devido às suas características biológicas, *S. cerevisiae* é um ótimo modelo para validação funcional e estudos de GSHs. Duas estratégias de integração nessa levedura se destacam, a de (1) integração múltipla que direciona o inserto para regiões repetidas no genoma como retrotransposons de levedura (do inglês, *Yeast retrotransposons*, Ty) (Boeke *et al.*, 1985; Lee; Silva, 1997; Sakai; Shimizu; Hishinuma, 1990) e DNA ribossômico (rDNA) (Dai *et al.*, 2013; Lopes *et al.*, 1989, e de (2) integração simples que direcionam para regiões intragênicas de genes de cópia única, como o gene HO (Voth *et al.*, 2001), genes codificadores de enzimas ativas em vias biossintéticas de aminoácidos ou nucleotídeos, considerados marcadores auxotróficos (Da Silva; Srikrishnan, 2012) e locais denominados sítios neutros (do inglês, *neutral sites*) (Kong *et al.*, 2022; Mikkelsen *et al.*, 2012) que consistem de regiões não codantes. Tais regiões carecem de análises completas *in vivo* e de critérios específicos para serem utilizadas na acomodação de

vias metabólicas completas e circuitos genéticos complexos, levando à necessidade da identificação de GSHs em *S. cerevisiae*.

GSHs em outros organismos também foram relatados, como os supracitados, lóco do tipo AAVS1 no genoma do macaco-rhesus, lóco ROSA26 em camundongos (Sadelain; Papapetrou; Bushman, 2012), bem como outras regiões em *Oryza sativa* (Sun *et al.*, 2023), *Schistosoma mansoni* (Ittiprasert *et al.*, 2023), *Caenorhabditis elegans* (Mouridi; Alkhaldi; Frokjar-Jensen, 2022), *Nannochloropsis oceanica* (Südfeld *et al.*, 2022), *Aspergillus fumigatus* (Furukawa *et al.*, 2022; Pham; Xie; Lin, 2020), *Cryptococcus neoformans* (Arras *et al.*, 2015), *Saccharomyces cerevisiae* (Liu *et al.*, 2022) e *Komagataella phaffi* (*Pichia pastoris*) (Gao *et al.*, 2022; Liu *et al.*, 2019; Nishi *et al.*, 2022). Apesar da sua crescente importância como plataforma segura para o direcionamento genético, todos os locais atualmente utilizados como porto seguro genômico ainda não foram totalmente validados (Pavani, 2020).

Mesmo em estudos que propuseram seguir critérios para identificação de GSHs no genoma humano, a adesão aos mesmos não foi completa, bem como não foram realizadas análises globais da expressão gênica no organismo alvo após a inserção de transgenes (Arras *et al.*, 2015; Pellenz *et al.*, 2019; Sadelain; Papapetrou; Bushman, 2012). Há, ainda hoje, uma grande preocupação relacionada à expressão dos próprios transgenes, pois esses são majoritariamente inseridos em regiões selecionadas de forma empírica ou mesmo ao acaso. Por conta disso há uma necessidade de selecionar, entre os organismos transformados, aqueles exibindo o padrão desejado de expressão dos transgenes, gerando a necessidade de considerável tamanho amostral para que esse objetivo seja atingido, a depender do organismo alvo. Outra preocupação também relevante é quanto ao efeito da inserção do transgene na expressão dos genes vizinhos e mesmo na expressão genômica.

Atualmente estão disponíveis dois *pipelines* computacionais para identificação de pGSHs. O pipeline de predição GSH-1 que segue critérios anteriormente definidos para *H. sapiens* (Sadelain; Papapetrou; Bushman, 2012) e também critérios novos (Aznauryan *et al.*, 2022), são eles: 50 kb de distância de genes; 300 kb de distância de oncogenes; 150 kb de distância de RNAs longos não-codantes (do inglês, *Long non-coding RNAs*, lncRNAs) e transportadores (do inglês, *Transfer RNA*, tRNAs); 300 kb de distância de micro RNAs (miRNA), telômeros e centrômeros; e 20 kb de distância de regiões potenciadoras (do inglês, *enhancers*). No mesmo ano, é publicado o *software* de computador GEG-SH mapper

(Shrestha *et al.*, 2022) para a identificação de pGSHs em sítios de inserções de elementos móveis polimórficos (pMEIs) em regiões intrônicas. Ambos trazem a desvantagem de serem restritos para o genoma humano, e apesar de apresentarem abordagens diferentes para identificar tais regiões, ambos ainda tiveram apenas um entre todos os pGSHs identificados validados completamente.

Portanto, há necessidade de identificação de GSHs que sigam critérios específicos para os diversos organismos, respeitando as suas particularidades em termos de tamanho do genoma, incidência de elementos repetitivos e essenciais, e alterações epigenéticas. Além disso, rupturas bialélicas geradas pelo direcionamento mediado por endonucleases, como o sistema CRISPR-Cas9, pode resultar em fenótipos indesejados por mutação nula homozigota quando ocorre inserção em GSH localizado em região intragênica (Pellenz *et al.*, 2019; Sadelain; Papapetrou; Bushman, 2012). Levando em consideração o contexto apresentado, bem como a falta de descrição na literatura de estratégias consolidadas para realizar buscas sistemáticas em genomas por GSHs, torna-se relevante o desenvolvimento de abordagens computacionais para caracterizar e identificar estas regiões em genomas conhecidos, bem como derivar regras genéricas visando expandir a aplicação para outros organismos.

1.4. *Saccharomyces cerevisiae*

S. cerevisiae, também conhecida como levedura da cerveja ou de padeiro, foi domesticada há pelo menos 5000 anos (Cavalieri *et al.*, 2003) e desde então seu genoma tem passado por intensa seleção artificial para a produção de alimentos e bebidas (pão, cerveja, vinho e destilados), biocombustíveis, compostos farmacêuticos e um extensa lista de outros compostos bioquímicos (Nielsen, 2019). Como o primeiro genoma eucariótico totalmente sequenciado (Goffeau *et al.*, 1996), essa levedura é um organismo modelo para estudos moleculares e genômicos de eucariotos por características como: (1) tempo de geração curto, (2) alta taxa de recombinação homóloga, (3) um organismo não patogênico e seguro (do inglês, *Generally recognized as safe*, GRAS), (4) diversos processos celulares conservados em relação às células humanas (Kachroo *et al.*, 2015), e (5) variedade de ferramentas disponíveis para sua manipulação genética (Guirimand *et al.*, 2021; Malcı *et al.*, 2022).

Essa levedura foi previamente utilizada como organismo chassi para produção de importantes biomoléculas, como nos casos do precursor da droga antimalária artemisinina

(Paddon *et al.*, 2013) e de opióides (Galanie *et al.*, 2015). No entanto, mesmo com a preocupação dos pesquisadores em direcionar as inserções dos genes da via metabólica dos opióides em *S. cerevisiae* para regiões genômicas que atuem minimamente como um GSH, a baixa titulação do produto final tornou a produção economicamente inviável (Galanie *et al.*, 2015). Assim, é plausível que a produção poderia ser aumentada pelo direcionamento da inserção em locais validados corretamente como GSHs, já que os autores não avaliaram viabilidade celular e taxa de expressão dos transgenes inseridos, esses poderiam ser gargalos para a produção das biomoléculas de interesse.

Há aproximadamente uma década, o consórcio internacional do Projeto Genoma Sintético de Levedura (do inglês, *Synthetic Yeast Genome Project*, SC2.0) (Richardson *et al.*, 2017) vêm desenvolvendo linhagens celulares contendo um dos cromossomos sintéticos desse organismo. Em novembro de 2023 esse grupo de pesquisadores atingiu um novo marco, a criação de uma cepa de *S. cerevisiae* com mais de 50% do genoma composto por DNA sintético (Zhao *et al.*, 2023), demonstrando o potencial de engenharia genômica desse organismo e, portanto, sua continuada importância como organismo chassi na área de biologia sintética para produção de biomoléculas de interesse (Galanie *et al.*, 2015; Paddon *et al.*, 2013).

Adicionalmente, estudos recentes demonstraram que essa levedura compartilha diversos processos celulares com o ser humano (Kachroo *et al.*, 2015) possibilitando que a *S. cerevisiae* se tornasse um organismo modelo para estudos de doenças e processos celulares humanos. Os trabalhos aplicam diversas abordagens com a identificação de genes ortólogos humanos (Laurent *et al.*, 2020; Steinmetz *et al.*, 2002; Zhang *et al.*, 2003), inserção de genes humanos relacionados a doenças (Coelho; Pinto; Murray, 2019; Hamza *et al.*, 2015; Sun *et al.*, 2016) além da humanização de vias metabólicas da *S. cerevisiae* (Boonekamp *et al.*, 2022; Garge *et al.*, 2020; Hamilton *et al.*, 2006; Hamza *et al.*, 2020; Lee *et al.*, 2021; Wildt; Gerngross, 2005).

2. OBJETIVOS

2.1. Objetivo geral

Desenvolver um *software* computacional que permita a identificação de GSHs em diferentes organismos eucarióticos, seguindo critérios específicos e anotações genômicas curadas dos organismos analisados.

2.2. Objetivos específicos

- Delinear critérios para refinar a identificação de sítios de GSHs para a inserção genética.
- Escrever o programa computacional em linguagem Python e de fácil utilização para pesquisadores da área.
- Avaliar o sistema *in silico* usando três modelos eucariotos: *Saccharomyces cerevisiae*, *Mus musculus* e *Homo sapiens*, com o *software* desenvolvido.
- Realizar a avaliação funcional desses potenciais GSHs (pGSHs) no organismo modelo *S. cerevisiae* com a criação e inserção de BioBricks nos diversos GSHs preditos permitindo.
- Analisar a expressão gênica local de genes vizinhos aos sítios de inserção (qPCR) e global via RNA-seq.

3. METODOLOGIA

3.1. Desenvolvimento do *software* SHIP

O *software* SHIP foi escrito na linguagem de programação Python 3.9 utilizando bibliotecas padrão, o que implica em um processo de instalação simplificado e na sua utilização transparente em múltiplas plataformas, como Linux e Windows.

A organização geral em termos de entradas e saídas do *software* encontra-se delineada na Figura 3. A entrada consiste no arquivo de anotação do genoma do organismo estudado, no formato GFF3 (do inglês, *Generic Feature Format*), que é o padrão em projetos genoma.

Para a identificação de GSHs baseada em premissas biológicas, uma série de atributos de anotação foram selecionados para direcionar uma busca por regiões com um tamanho mínimo (ver abaixo) e desprovidas de genes, mas também que possam albergar contextos regulatórios, como sítios de ligação a fatores de transcrição, hipermetilação do DNA e estado de alterações da cromatina e histonas, os quais são avaliados por meio de informações coletadas nos bancos de dados externo por meio da Interface de Programação de Aplicação (API) *Entrez Programming Utilities* (E-utilities) do Centro Nacional de informações sobre biotecnologia dos Estados Unidos (do inglês, *National Center for Biotechnology Information*, NCBI) (Sayers *et al.*, 2023), API REST do Ensembl (Martin *et al.*, 2023) e o UCSC Genome Browser (Nassar *et al.*, 2023). O *software* realiza uma análise combinatória para a identificação de regiões intergênicas e direção dos genes flanqueadores das mesmas. Posteriormente, são excluídas regiões intergênicas que possuem elementos genéticos como pseudogenes transcritos, elementos de transposição, origens de replicação, sequência de replicação autônoma, região telomérica, cassete do tipo acasalamento, repetição terminal longa (do inglês, *Long Terminal Repeat*, LTR), retrotransposon, centrômero, etiqueta de sequência expressa EST e locais de ligação do fator de transcrição previstos.

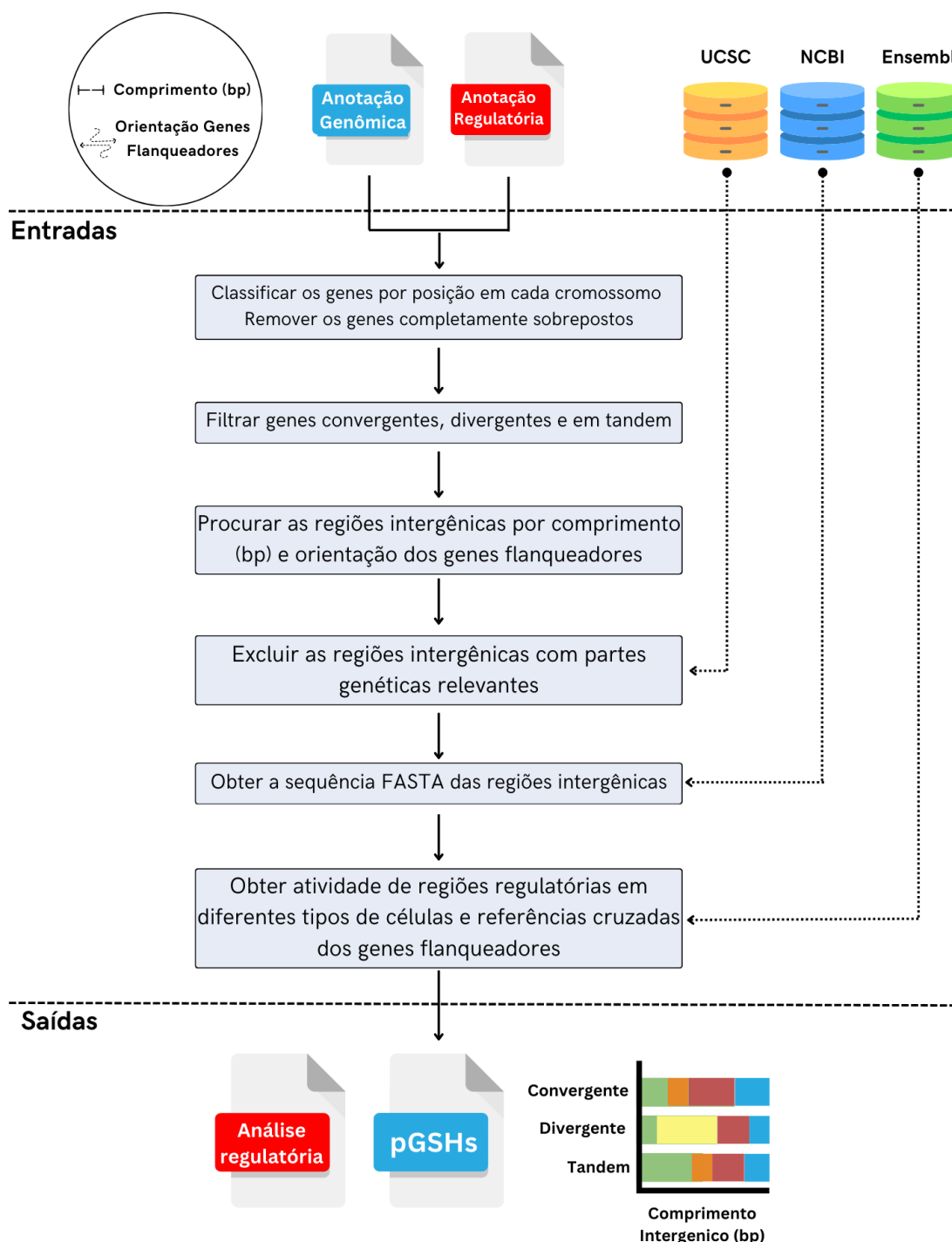


Figura 3 - Esquema representativo do software SHIP. Como entrada, foi usado o arquivo em formato de GFF3 de anotação genômica do organismo alvo. Como opção de execução pode ser utilizado também um arquivo de características regulatórias, ainda no formato GFF3. Após a escolha do usuário do tamanho da região intergênica e da direção da transcrição dos genes que flanqueiam a região intergênica, o *software* classifica e organiza os genes desde o início dentro de cada cromossomo. Em casos de sobreposições completas de genes, o menor gene de sobreposição é excluído. Com a seleção de cada GSH, há acesso por meio da Interface de Programação de Aplicação (API) *Entrez Programming Utilities* (E-utilities) do *National Center for Biotechnology Information* (NCBI) e a API REST do Ensembl, que permite a obtenção da sequência de nucleotídeos de cada região intergênica e informações adicionais sobre seus genes flanqueadores contidos no próprio Ensembl e em outros bancos de dados. Como arquivos de saída, o algoritmo retorna uma tabela e um gráfico no formato PNG com a distribuição de distâncias intergênicas e um arquivo em formato txt com a lista de regiões intergênicas e aspectos regulatórios, se houverem.

Além da localização de componentes relevantes, o *software* adota uma série de critérios que permitem ao usuário refinar a busca por GSHs potenciais (pGSHs). Estes critérios, bem como a sua justificativa encontram-se listados na Tabela 1.

Tabela 1. Critérios e justificativas para escolha dos potenciais GSHs (pGSHs) avaliados pelo *software* SHIP.

Critérios	Justificativa
Tamanho da região intergênica	O usuário pode escolher o tamanho do GSH a ser identificado pelo <i>software</i> SHIP considerando o tamanho do genoma e densidade gênica do organismo alvo, tamanho do transgene a ser inserido. Valor padrão: 1500 - 2000 pb.
Orientação dos genes flanqueadores dos GSHs	Os genes que flanqueiam a região GSH podem estar in <i>tandem</i> (Mesma direção), divergentes (Direções opostas com a região promotora mais próxima ao GSH), ou convergentes (Regiões terminadoras mais próximas ao GSH). Valor padrão: Convergente.
Distância dos GSHs de genes flanqueadores considerados essenciais para viabilidade celular e de genes codificadores de proteínas relacionados a câncer (para organismo como humano)	Para evitar ou minimizar que a inserção de transgenes nos GSHs identificados afete a expressão dos genes vizinhos, flanqueadores, levando à fenótipos adversos indesejados, o indicado nessa proposta foi escolher os GSHs que não sejam flanqueados por esses genes.

Como saída, o *software* SHIP retorna uma série de relatórios tabulares e gráficos com a distribuição de distâncias intergênicas e frequência de genes vizinhos no genoma analisado. Um arquivo em formato texto é gerado com a lista de regiões intergênicas contendo os genes flanqueadores, com informações de diferentes bancos de dados e sequências FASTA dos pGSHs. Além disso, se fornecida no início, o *software* retorna a lista das sequências regulatórias que estão dentro da região que engloba os pGSHs bem como os genes vizinhos (Figura 3).

3.2. Validação funcional do *software* SHIP

3.2.1. Validação experimental *in vivo*

Para verificar se os GSHs identificados pelo *software* SHIP funcionam como um local seguro para ancoragem de sequências de interesse, genes repórteres foram inseridos no organismo modelo *S. cerevisiae* (Figura 4). Os critérios de seleção foram: comprimento do pGSH (Ex: entre 1 e 1,5 kb); direção convergente dos genes flanqueadores à região; e filtragem de pGSHs excluindo-se as regiões não-gênicas supracitadas. A construção do cassete de expressão dos genes repórteres foi feita por meio da estratégia de BioBricks detalhada a seguir.

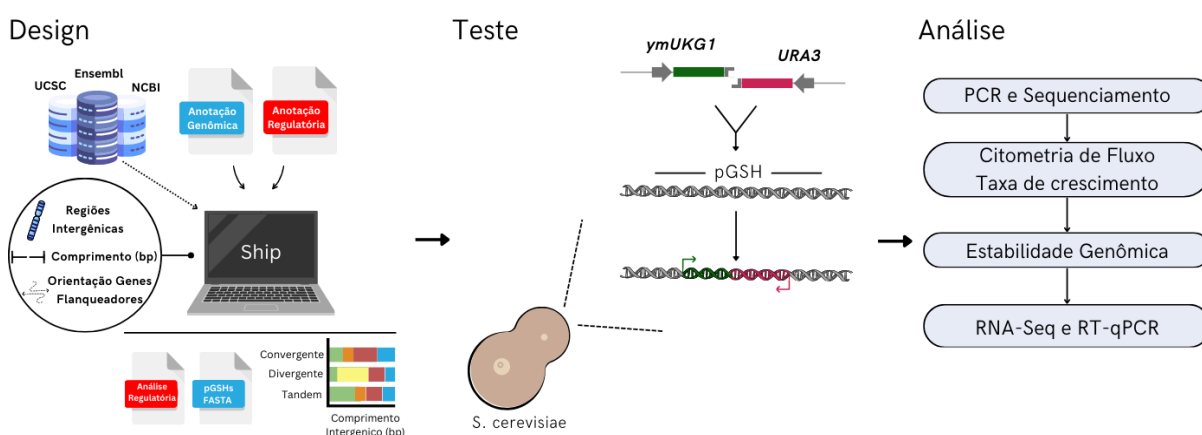


Figura 4 - Visão geral da estratégia do sistema de identificação e validação funcional de porto seguro genômico em eucariotos. A primeira etapa (*Design*) envolveu o desenvolvimento do Programa de Identificação de Porto Seguro (SHIP), algoritmo para pesquisar portos seguros genômicos em eucariotos a partir de dados de características gerais. *Saccharomyces cerevisiae* foi escolhido como organismo chassi para a validação *in vivo* do SHIP, resultando em sua transformação genética com dois BioBricks sobrepostos compostos pelo gene repórter de fluorescência verde ymUKG1 e URA3, como marcador auxotrófico (Teste). Segue-se a etapa de análise que consiste em PCR e análise de sequenciamento para verificar a correta inserção dos BioBricks, teste de estabilidade genômica e análise fenotípica, como citometria de fluxo, análise de taxa de crescimento e análise de RNA-Seq e RT-qPCR.

3.2.2. Montagem dos BioBricks

O BioBrick 1 foi composto pela versão monomérica de um gene codificador de proteína fluorescente verde derivado de coral (Tsutsui *et al.*, 2008), códon otimizado para *S. cerevisiae* (ymUKG1) (Kaishima *et al.*, 2016), sob controle do promotor constitutivo TEF1. O

segundo foi o gene marcador auxotrófico URA3 (nome sistemático: YEL021W) do plasmídeo pYC210 (Olesen *et al.*, 2000) com promotor e terminador próprios. Todos os BioBricks foram amplificados utilizando a *Platinum™ Taq DNA Polymerase High Fidelity* (Invitrogen, 11304011). A Tabela 2 descreve genes, plasmídeos de origem e sequências regulatórias, como promotor e terminador de cada um dos BioBricks usados neste estudo. Tabela 3 e Tabela 4 descrevem os Oligonucleotídeos (Oligos) e a programação dos ciclos utilizados nessas reações, respectivamente.

Tabela 2. BioBricks utilizados para validação funcional do software SHIP.

BioBricks	Tipo	Plasmídeo de Origem	Promotor	Terminador	Tamanho Total
ymUkG1	Proteína Fluorescente	YCUkGy	pTEF1	PGK1t	1402 pb
URA3	Marca Auxotrófica	pYC210	pURA3	URA3t	1169 pb

Tabela 3. Oligonucleotídeos utilizados para montagem dos BioBricks.

pGSH	Oligonucleotídeo	Sequência
	JR-UkG	TCGTAATGATTTTTATAATGACGAAAAAAAAAAAAATTGGAAAGAA AAAGCTGGGCGCGCCTAACGAACGCAGAATTTTCGAG
	JF-URA3	TCTATGTTTCGGGTTTCAGCGTATTTTAAGTTTAATAACTCGAAAATT CTGCGTTTCGTTAGGCGCGCCCAGCTTTTTCTTTCC
2	S2F-UkG	AATAAAGTATGTCACCTGAATTGAAGACTACTGAACCATGGCAGCT AATACAATCACGCCCCCCCCACACCCATAGCTTCA
	S2R-UkG	TATTAGGGCCGTCATTACGACTTAAGAGCACTTTGAAATCAGTAC CCTCTGTTCAACGTGTTCAATTCATCATTTTTTTTTTTATTC
3	S3F-UkG	TTGCCTTGGGATGACAAACAGCTGGATATTATTATTATTGGCAGTA GCCCTAGTATGTTGCCCCACACACCATAGCTTCA
	S3R-URA3	AATGTTTATTTGTGCCTTTAGAACTTTCCCCTTGAGTCCAACGTT ATAACGTGCCCTCGTTCAATTCATCATTTTTTTTTTTATTC
4	S4F-UkG	CTTCGAGAAGGCCATAAAAAGAACCGTGACAGGGCCAAACAGA AAAATAATGAAGACAAGCCCCACACACCATAGCTTCA
	S4R-URA3	AACCCAGAATATTTTCAGTGCAAAAAGCTTCCGATCCTCAAATAC AGTGAGAGAAAAAGCTTCAATTCATCATTTTTTTTTTTATTC
5	S5F-UkG	TTGAACATAACCATGTAACTCTTTTCCCGTCTACTGTTAATCAAGG GAACATGTTGAAAGCCCCACACACCATAGCTTCA
	S5R-URA3	ACGATGTATTACTTAGTGCTCTTTTGTAGCAATAAATTCGCGTGAA

pGSH	Oligonucleotídeo	Sequência
		ATACTCAACACATGTTCAATTCATCATTTTTTTTTTTATTC
6	S6F-UkG	CTCATAGGCAACCGGTACTTATCATAAGTACGTAATAAGAAAAGT ATGTCGCAGGACGTCCCCACACACCATAGCTTCA
	S6R-URA3	ATCAGCTTCAGTGCCGATATCCTATCAGGCTCTAGTAGGGCTGAA AATTAGGTGGAGGCGTTCAATTCATCATTTTTTTTTTTATTC

Tabela 4. Ciclo utilizado para montagem dos BioBricks com a *Platinum™ Taq DNA Polymerase High Fidelity*.

	Fase	Temperatura	Tempo
	Desnaturação Inicial	94 °C	00:30
	Desnaturação	94 °C	00:30
35	Anelamento	55 °C	00:30
	Extensão	68 °C	1 min/kb
	Extensão Final	68 °C	10:00
	Espera	4 °C	∞

A inserção dessas sequências no genoma de *S. cerevisiae* foi realizada por recombinação homóloga direta. Para tanto, foi necessário incorporar 60 pb na região 5' de sequência homóloga à região do pGSH no transgene codificador de ymUKG1 e também 60 pb na região 3' de sequência homóloga ao cassete de expressão da marca de seleção. A incorporação das sequências de homologia foi realizada com diferentes oligonucleotídeos iniciadores para formação do primeiro BioBrick. O segundo BioBrick foi obtido a partir de amplificação por PCR de sequências de seleção, como o marcador auxotrófico URA3. Essa sequência também foi amplificada com o uso de oligonucleotídeos iniciadores diferentes, cada um contendo 60 pb de sequência homóloga ao transgene e à região do pGSH, 5' e 3', respectivamente. A Figura 5 descreve a estratégia de montagem dos dois BioBricks.

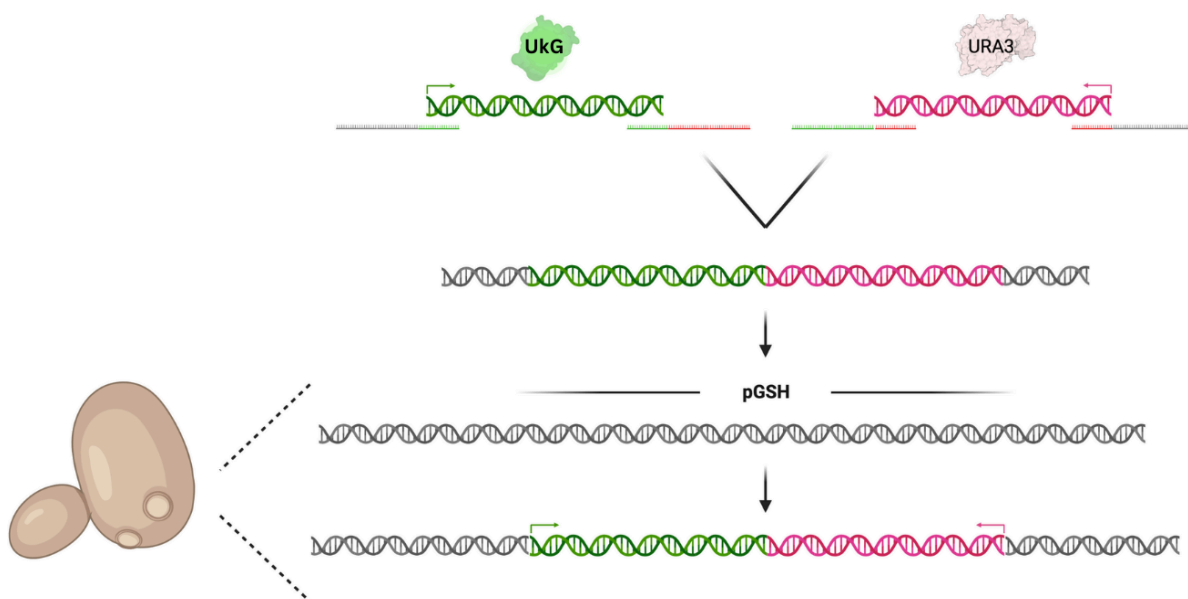


Figura 5 - Representação esquemática da estratégia de montagem dos BioBricks utilizados para transformação genética de *S. cerevisiae*. O primeiro BioBrick foi obtido a partir de amplificação do cassete de expressão do gene codificador de ymUKG1 (UkG) com oligonucleotídeos que apresentam 60 pb de homologia com o 5' pGSH e com o cassete de expressão da marca de seleção, representado por ranhuras verticais. O segundo BioBrick foi amplificado para conter o cassete de seleção flanqueado por 60 pb de sequência homóloga (ranhuras verticais) ao cassete de expressão de UkG e ao 3' do pGSH.

3.2.3. Purificação e concentração dos BioBricks

As reações de de PCR foram diretamente purificadas e concentradas utilizando o *ReliaPrep™ DNA Clean-Up and Concentration System* (Promega, A2893) seguindo o protocolo do fabricante. Várias replicatas de um mesmo BioBrick foram concentrados em conjunto no *Speed Vacuum*, quantificados no espectrômetro *NanoDrop™ OneC* (Thermo Fisher Scientific, EUA) e visualizados em gel de agarose 1% para verificar integridade, tamanho e corroborar quantificação.

3.2.4. Transformação genética

Os BioBricks foram montados por recombinação homóloga *in vivo* através da transformação celular da cepa *S. cerevisiae* BY4741 (MATa, met15Δ, his3Δ, ura3Δ, leu2Δ) (Brachmann *et al.*, 1998) oriunda do Arquivo Europeu de *Saccharomyces cerevisiae* para análise funcional (do inglês, *European Saccharomyces cerevisiae Archive for Functional analysis*, EUROSCARF) utilizando o método de transformação de alta eficiência por

eletroporação com acetato de lítio (Becker; Guarente, 1991) e a transformação LiAc/SS-DNA/PEG (Gietz *et al.*, 1995) com as modificações descritas por (Annaluru *et al.*, 2014).

3.2.5. Extração de DNA genômico

Extração do DNA genômico foi realizada utilizando o *Wizard® Genomic DNA Purification Kit* (Promega, A1120) seguindo as recomendações do fabricante. Os seguintes passos foram otimizados: células cresceram por 20 horas em 1,1 mL de meio YPD (do inglês, *yeast extract peptone dextrose*) em Placas de 96 poços de poço profundo; 30 µL de *Lysing Enzymes from Trichoderma harzianum* (Sigma-Aldrich, EUA) na concentração inicial igual a 20 mg/mL foi adicionado inicialmente para ruptura de parede celular; ao final as amostras foram ressuspensas em 30 µL de Água MilliQ.

3.2.6. Identificação dos clones com inserção correta

Para identificação dos clones com a inserção correta dos BioBricks, 48 transformantes foram selecionados ao acaso para a extração do DNA genômico e a amplificação completa do inserto foi realizada com a *GoTaq® G2 DNA Polymerase* (Promega, M7841) utilizando oligonucleotídeos que se anelam na região genômica fora da área de recombinação homóloga (Oligonucleotídeos genômicos). Tabela 5 e Tabela 6 apresentam os oligonucleotídeos utilizados neste estudo e as condições de amplificação por PCR, respectivamente.

Tabela 5. Oligonucleotídeos genômicos utilizados para identificação dos clones com inserção correta dos BioBricks UkG e URA3 em cada pGSH.

pGSH	Oligonucleotídeo	Sequência	Tamanho Amplicon
2	gS2-R	AGGCTCTGTGTACTCTACAGCCG	3047 pb
	gS2-R	AGGCTCTGTGTACTCTACAGCCG	
3	gS3-F	GCCTGCATTTGGAAACCTGCTGCG	2930 pb
	gS3-R	CATTGGCTGACGTCGCGCTTTCC	
4	gS4-F	GATGAGACCTGCAAGGATTTCTGTGTGAGG	3009 pb

pGSH	Oligonucleotídeo	Sequência	Tamanho Amplicon
	gS4-R	ACTGAAAACTCCGCTACCTGCTGTAACAGC	
5	gS5-F	TCCAATTGATTGTGCAACAGTTTGATCTTCC	2970 pb
	gS5-R	GTGTCAAGATATTCGCAGCAGCCAC	
6	gS6-F	CCTCTCAAATAAGTGGCGCAGATGCTCTAG	3235 pb
	gS6-R	ACCTGAGAGCTCCGATGATGAACTACC	

Tabela 6. Ciclo utilizado para identificação dos clones com inserção correta com a *GoTaq® G2 DNA Polymerase*.

	Fase	Temperatura	Tempo
	Desnaturação Inicial	95 °C	02:00
45x	Desnaturação	95 °C	00:30
	Anelamento	55 °C	00:30
	Extensão	72 °C	1 min/kb
	Extensão Final	72 °C	05:00
	Espera	4 °C	∞

3.2.7. Sequenciamento

Para as amostras com inserção correta, foi realizado replicatas de PCR com a *Platinum™ Taq DNA Polymerase High Fidelity* (Invitrogen, EUA), oligonucleotídeos genômicos (Tabela 5) e ciclo apresentado na Tabela 4 seguindo orientações do fabricante. Em seguida as amostras foram analisadas por eletroforese em gel de agarose 1% para confirmar tamanho do amplicon e separar possíveis amplificações inespecíficas ou parciais. As bandas de tamanho esperado foram cortadas do gel e o DNA purificado seguindo os mesmos passos do item 3.2.5. O sequenciamento pelo método de Sanger foi realizado pela ACTGene Análises Moleculares (Rio Grande do Sul, Brasil) com o equipamento *AB 3500 Genetic Analyzer* (ThermoFisher, EUA).

3.2.8. Estabilidade genômica dos pGSH

A análise de estabilidade genômica das regiões pGSHs após inserção de BioBricks foi realizada utilizando o método de passagem celular contínua (Annaluru *et al.*, 2014). As colônias foram inicialmente inoculadas em 5 mL de meio YPD e cultivadas durante a noite. Em seguida, 5 μ L de cada cultura foram transferidos para um meio YPD fresco de 5 mL. Esse ciclo se repetiu durante dez dias. Posteriormente, as células foram plaqueadas em YPD e cinco clones foram selecionados aleatoriamente e analisados por PCR utilizando os oligonucleotídeos genômicos para amplificação completa do inserto. Feito em triplicata biológica para os três clones de cada linhagem pGSH.

3.2.9. Análise do acúmulo das proteínas repórteres e taxa de crescimento

Para cada uma das 5 linhagens celulares contendo os BioBricks no pGSH selecionado, o acúmulo da proteína UkG1 foi analisado por citometria de fluxo utilizando o citômetro *BD FACSVerser™ Cytometer* (BD Biosciences, EUA), configuração 4-Blue 2-Red (ROU) e fluxo entre 500 e 900 eventos por segundo. Células foram pré-inoculadas em 5 mL de YPD, e incubadas durante a noite a 30 °C a 200 rpm. Depois do crescimento, as células foram inoculadas em 20 mL de YPD em frascos de Erlenmeyer de 125 mL com OD₆₀₀ inicial de 0,1. Em cada etapa de 0, 2, 4, 6, 12, 24 e 36 horas, as linhagens foram avaliadas quanto ao perfil de crescimento com medição da OD₆₀₀ utilizando 1 mL de cultura com o espectrômetro *NanoDrop™ OneC*. Para aferir OD₆₀₀ maiores que 3, amostras foram diluídas 10x com água destilada. 1 x 10⁶ células foram coletadas, fixadas com formaldeído 4% e solução salina tamponada com fosfato (PBS, do inglês, *Phosphate-Buffered Saline*) por 30 minutos à temperatura ambiente e armazenadas em PBS 1% a 4 °C até o uso. Análises no citômetro ocorreram no mesmo dia de fixação ou em até 2 dias seguindo os parâmetros da Tabela 6. Cada ensaio foi realizado em triplicata técnica e biológica com análise de 25.000 eventos. Os dados foram processados utilizando o *software FlowJo™ v.10.9.0* (BD Biosciences, EUA). Amostras passaram pela análise do FlowAI (Monaco *et al.*, 2016) para selecionar os eventos com padrões SSC-A/FSC-A compatíveis com células viáveis, em seguida células com padrões FSC-H/FSC-A de células únicas e não sobrepostas foram analisadas quanto à detecção de fluorescência pelo canal FITC-A. Uma vez determinado essas portas de análise, todas as amostras foram selecionadas e processadas sob mesmas portas na mesma área de trabalho.

3.2.10. Extração de RNA total

Para cada uma das 5 linhagens celulares obtidas, um clone de cada foi selecionado para a extração de RNA total em replicatas biológicas realizadas em dias independentes. As células foram cultivadas em 20 mL de YPD em frascos de Erlenmeyer de 125 mL por 16 horas. Em seguida, 2 mL de cada cultivo foram utilizados para extração de RNA total com o *RNeasy Plus Micro Kit* (Qiagen, 74034), seguindo as instruções do fabricante. Os RNAs extraídos foram quantificados com o *Qubit 2.0 Fluorometer* (ThermoFisher, EUA), após uma verificação de qualidade usando eletroforese e o *2100 Bioanalyzer* (Agilent, EUA). Por fim, foi realizada uma análise de pureza com o NanoDrop para detectar possível contaminação de DNA.

3.2.11. Quantificação de transcritos via RT-qPCR

A quantificação de transcritos foi feita utilizando a técnica de PCR quantitativo (qPCR) associado a transcrição reversa (RT-qPCR). A biblioteca de cDNA foi montada com o *High-Capacity cDNA Reverse Transcription Kit* (Applied Biosystems, 4368814). Foram usados oligonucleotídeos aleatórios e 2 µg de cada amostra de RNA, seguindo o protocolo do fabricante. O cDNA foi diluído 10X e 2 µL dessa diluição foram usados para cada reação de qPCR em um volume final de 10 µL.

Foi incluído um controle negativo para cada gene analisado (NTCs). As reações de qPCR foram realizadas com o *SYBR™ Green PCR Master Mix* (Thermo Fisher Scientific, 4309155), seguindo as instruções do fabricante no *QuantStudio 3 Real-Time System* (Thermo Fisher Scientific, A28567). As réplicas foram executadas na mesma placa para cada amostra e, para evitar diferenças sistemáticas no procedimento de PCR, as amostras foram distribuídas aleatoriamente na placa. Os genes ACT1, ALG9 e TAF10 foram usados como controles internos para normalizar os dados (Teste *et al.*, 2009). O valor Ct foi usado para estimar a especificidade dos produtos amplificados. Os oligonucleotídeos usados para analisar a expressão dos genes flanqueadores de cada pGSH e as condições de ciclo estão listados nas Tabela 7 e Tabela 8, respectivamente.

Tabela 7. Oligonucleotídeos utilizados para análise de expressão gênica por RT-qPCR.

pGSH	Gene Vizinho	Oligo	Sequência	Amplicon	Referência
2	PUG1	PUG1-S2F PUG1-S2R	GCAACTAAAAATCACACCCACAGT TCCCGATATACCCGACAGCTT	83 pb	Protchenko <i>et al.</i> , 2008
	YER186C	YER-S2F YER-S2R	TGCCTTCCGCACTAAGAAGC TGAAACCGTCGCACTCTCC	137 pb	Este Trabalho
3	YHL041W	YHL-S3F YHL-S3R	CCGTTCTAGAGAACTGCCGC TTGCAGAAACGATAGACCGCA	141 pb	Este Trabalho
	ARN1	ARN1-S3F ARN1-S3R	AACCCATCAAAAGGACACCA TTGAAGGGCCAGAAGTCCTA	100 pb	Demuyser <i>et al.</i> , 2017
5	COS9	COS9-S5F COS9-S5R	TGTGGGGGAGGTTTTCTCCT TCGCACCACTTTCTTGCTCA	121 pb	Este Trabalho
	SRY1	SRY1-S5F SRY1-S5R	ATATGTGCTCCGTAACCGGC ACCATGCACAGGCTATTGCT	118 pb	Este Trabalho
6	ULS1	ULS1-S6F ULS1-S6R	AGAACCAGAAGCGATGTCCG GGATAGCAAATGCCCCACCT	131 pb	Este Trabalho
	THI72	THI72-S6F THI72-S6R	GCCGACCTTCCAGTCATAGC GTGGATAGGCGGTTCTGCTG	136 pb	Este Trabalho
BioBricks	UkG	UkG-SF UkG-SR	GAAGTGGGAACCATCCACCG TGATGACCACCACTTCCAAC	96 pb	Este trabalho
	URA3	URA3-SF URA3-SR	TGTCGCTCTTCGCAATGTCA GTTAGCGGTTTGAAGCAGGC	139 pb	Este Trabalho
Genes endógenos	ACT1	ACT1-SF ACT1-SR	GTTACGTCGCCTTGGA CTTC TGGGGCTCTGAATCTTTCGT	125 pb	Holmes <i>et al.</i> , 2015
	ALG9	ALG9-SF ALG9-SR	AATCAAGGTGGTGTGAAGGCAC GATTATCTGGCAGCAGGAAAGAAC	91 pb	Bosio <i>et al.</i> , 2017
	TAF10	TAF10-SF TAF10-SR	ATATTCCAGGATCAGGTCTCCGTAGC GTAGTCTTCTCATTCTGTTGATGTTGTTG	141 pb	Ottillie <i>et al.</i> , 2022

Tabela 8. Ciclo utilizado para análise de expressão gênica por RT-qPCR com o *Syber Green Master Mix*.

	Fase	Temperatura	Tempo
40x	Ativação Polimerase	95 °C	10:00
	Desnaturação	95 °C	00:15
	Anelamento & Extensão	60 °C	01:00
Curva de Dissociação	Desnaturação	95 °C	00:15
	Anelamento	60 °C	01:00
	Dissociação	95 °C	--

A análise de expressão relativa dos genes candidatos foi obtida pelo software Biogazelle qBase+ 3.0. A normalização da expressão gênica foi realizada utilizando os genes referência mais estáveis determinados pelo algoritmo geNorm (Vandesompele *et al.*, 2002). O geNorm foi empregado para calcular a variação pareada ($V_n / n + 1$) entre cada gene de referência, para indicação do número ótimo de genes de referência para normalização da expressão gênica. Os resultados de expressão relativa foram analisados utilizando o programa REST© 2009 (do inglês, *Relative expression programa tool*) (Pfaffl; Horgan; Dempfle, 2002) e a análise de variância (ANOVA) para comparar as diferenças entre as médias e gerar um nível de significância entre os grupos.

3.2.12. RNA-Seq

O experimento de RNA-Seq de cada uma das linhagens pGSHs e BY4741 foi realizada pela GenOne Soluções em Biotecnologia (Rio de Janeiro, Brasil) usando o *NovaSeq 6000 Sequencing System* (Illumina, EUA). Todas as análises foram realizadas utilizando o genoma de referência de *S. cerevisiae* S288C.

Para controle de qualidade das sequências, fragmentos suscetíveis a alinhamentos incorretos foram eliminados com o limite Q30 e utilizando o *software* fastp (Chen *et al.*, 2018). O mapeamento foi feito com o *software* STAR (Dobin *et al.*, 2013) e o *software* htseq-count (Anders; Pyl; Huber, 2015) foi utilizado para contar o número de leituras associadas a cada gene resultando em um arquivo tabular contendo o código correspondente à região ou gene, juntamente com a quantidade de leituras para cada amostra.

As análises foram iniciadas no ambiente R (R: The R Project for Statistical Computing) com as bibliotecas edgeR (Robinson; McCarthy; Smyth, 2010), DESeq2 (Love; Huber; Anders, 2014) e pheatmap (Kolde, 2012). As amostras foram filtradas com valor de estatística $p < 0,05$ e submetidas a normalização. Os dados foram visualizados (Heatmaps e PCAs) utilizando a biblioteca ClustVis (Metsalu; Vilo, 2015) e os Volcanos Plots com ggplot2 (Wickham, 2016), com limite de $\log_2FC \pm 0,6$ e $FDR < 0,001$ para destacar as diferenças significativas. Análise de enriquecimento do conjunto de genes diferencialmente expressos e compartilhados entre todas as linhagens pGSHs foi realizada utilizando o programa g:Profiler (g:Profiler – a web server for functional enrichment analysis and conversions of gene lists, [s. d.]; Kolberg *et al.*, 2023). Foi utilizado o Limite de significância de Benjamini-Hochberg FDR com limite de significância de 0.01.

4. RESULTADOS

4.1. Implementação do SHIP e avaliação geral dos critérios de seleção de GSHs

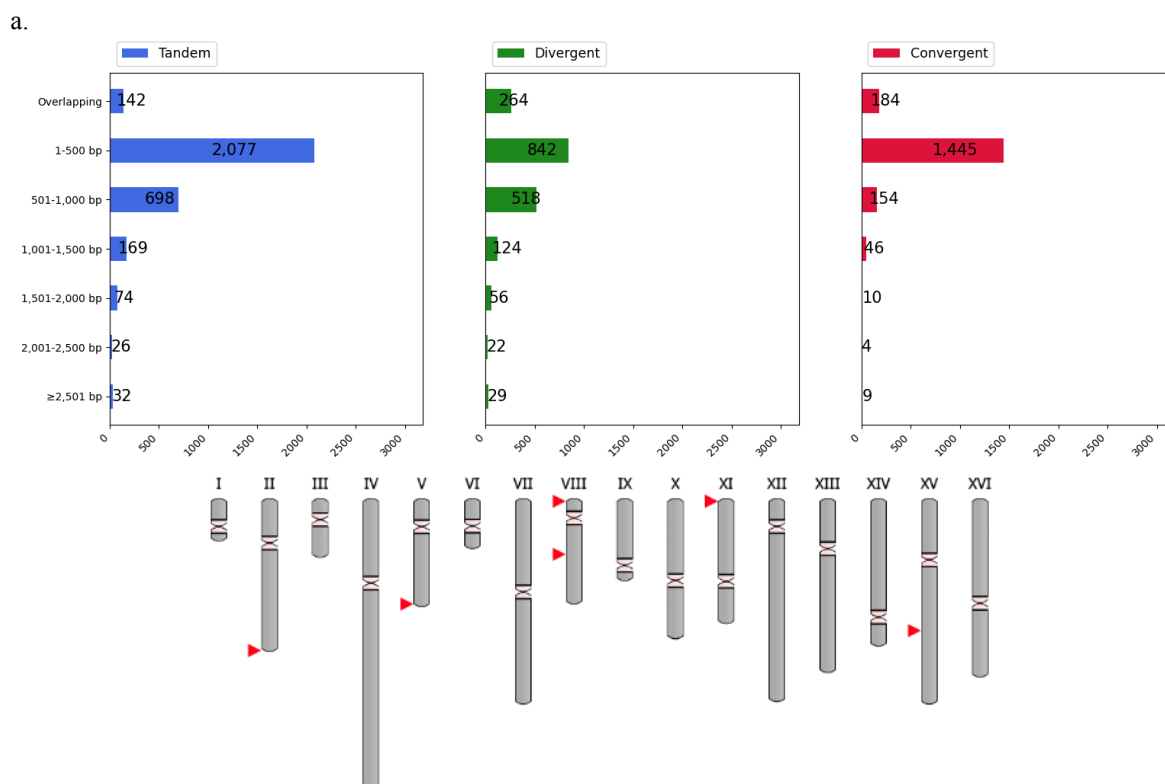
Em uma visão geral, a estratégia do algoritmo do *software* SHIP para identificação de pGSHs consiste na execução de 5 etapas principais. O SHIP (1) Seleciona e organiza os genes dentro de cada cromossomo; (2) Monta tabela e gráfico com a distribuição das regiões intergênicas do genoma alvo; (3) O usuário seleciona a orientação dos genes flanqueadores e a variação de comprimento dos pGSHs; (4) Acesso à bancos de dados externos (UCSC Genome Browser, NCBI e Ensembl) para obtenção de informações cruzadas; e (5) Geração da lista com informações detalhadas de cada pGSH e dos genes vizinhos, bem como sua sequência FASTA. Baseado nos critérios descritos para seleção de GSHs em diferentes organismos, SHIP leva em consideração apenas regiões intergênicas, a orientação dos genes vizinhos e a utilização da densidade genômica do organismo alvo como parâmetros para a escolha do comprimento em pb do GSH.

Os requerimentos básicos para a execução do *software* SHIP são: a disponibilidade de um genoma completo, preferencialmente com poucas lacunas de montagem, e de uma anotação genômica de alta qualidade, incluindo regiões codificadoras e não-codificadoras de proteínas, como RNAs não-codificadores e elementos transponíveis. O tempo de execução varia de acordo com o tamanho do genoma a ser analisado, da qualidade da internet e do tráfego de acesso aos bancos de dados externos. Assim, entre os organismos analisados nesse estudo, nenhuma execução durou mais que 10 minutos utilizando um MacBook Pro 2015, sistema operacional macOS, processador Intel Core i5 e 8 GB de memória RAM (do inglês, *Random Access Memory*). Vale ressaltar que o SHIP pode ser executado em computadores de uso cotidiano.

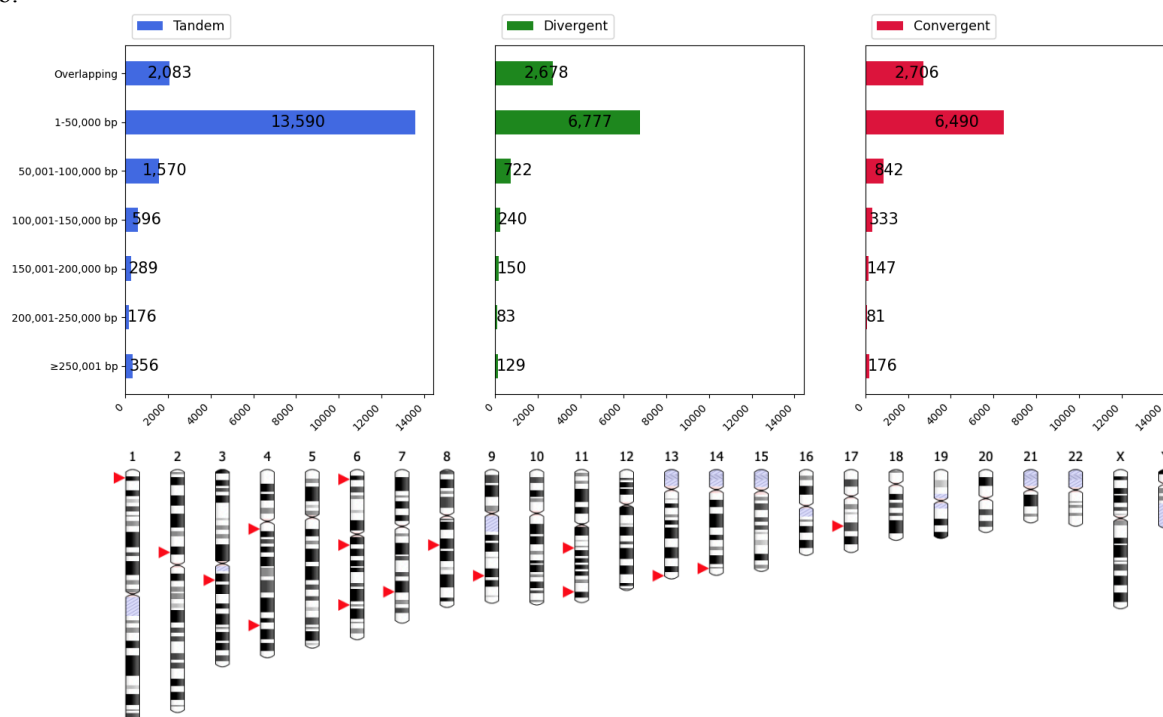
4.2. SHIP identifica potenciais GSHs em genomas de eucariotos usando dados de características gerais

Este trabalho avaliou os genomas de três organismos eucarióticos modelos: *Saccharomyces cerevisiae* (R64), *Homo sapiens* (GRCh38) e *Mus musculus* (GRCm38), que foram escolhidos devido à sua importância para a indústria, saúde e pesquisa acadêmica.

Primeiro, seus genomas foram analisados pelo *software* SHIP, em seguida uma verificação manual das regiões genômicas de saída (pGSHs) foi realizada em navegadores de banco de dados genômicos para validar *in silico* os resultados gerados. Vale destacar que o algoritmo trata o genoma como um conjunto de partes anotadas e, como não faz nenhuma previsão ou análise *a priori* com base na sequência de DNA, a precisão da análise dependerá da qualidade da anotação do genômica alvo. Como critério canônico para seleção de candidatos à GSH, todos os GSHs identificáveis são regiões intergênicas, ou seja, entre dois genes. Os demais critérios podem, e devem, ser gerenciados para se adequarem às especificações biológicas do genoma do organismo alvo e das necessidades de cada usuário (Figura 6).



b.



c.

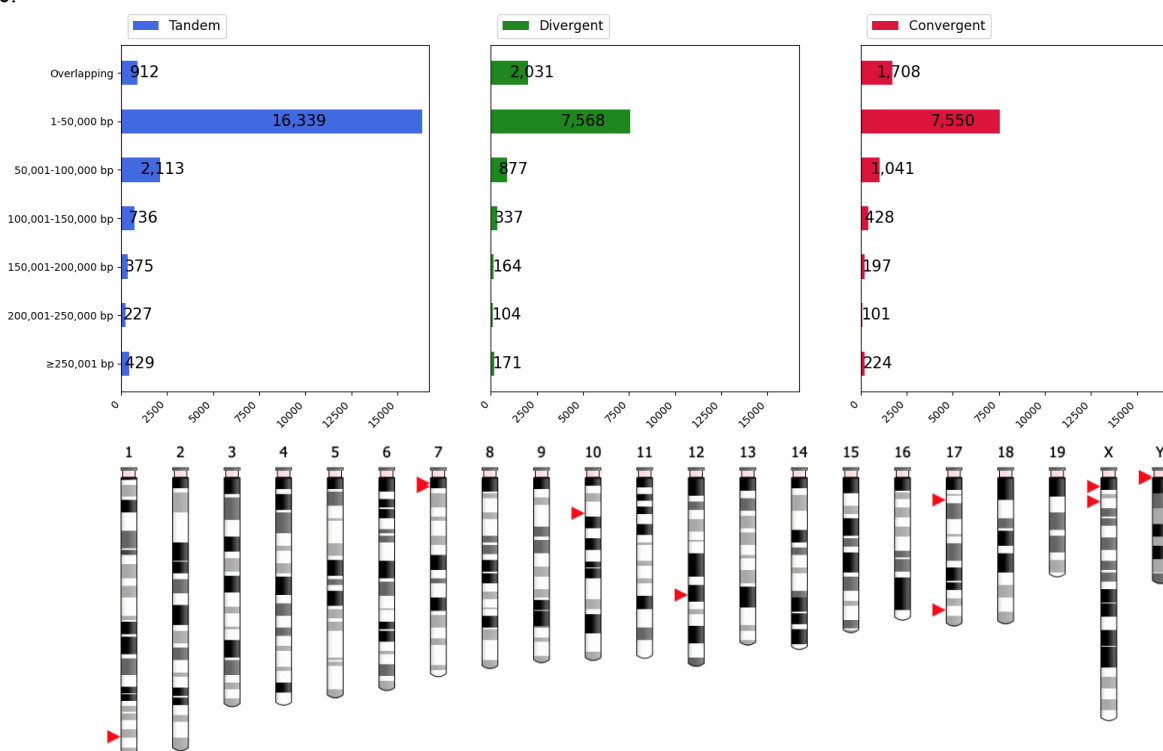


Figura 6 - Histograma da distribuição genômica das regiões intergênicas e ideograma da localização genômica dos pGSHs identificados. Histograma com a distribuição genômica com o comprimento em pb das regiões intergênicas entre as três disposições possíveis dos genes flanqueadores e ideograma com a localização genômica dos pGSHs identificados pelo SHIP considerando genes flanqueadores convergentes dos organismos alvo a) *Saccharomyces cerevisiae* (R64), b) *Homo sapiens* (GRCh38.p13) e c) *Mus musculus* (GRCm38.p6). Setas vermelhas indicam a posição genômica de cada pGSH. Ideogramas gerados com a biblioteca JavaScript para visualização de cromossomos Ideogram.js (Weitz, 2023).

Neste contexto, a orientação dos genes vizinhos e o comprimento da região em pares de base são escolhas do usuário. Como os genomas eucarióticos apresentam uma ampla gama de tamanho e distribuição gênica, o comprimento do GSH deve mudar de acordo com o organismo alvo. Assim como a inserção de transgenes em GSHs muito curtos poderia levar à interferência na expressão gênica da região vizinha, GSHs muito longos poderiam conter genes ou outras partes genéticas não caracterizadas. Para reduzir esses eventos indesejados, foi estabelecida uma faixa de comprimento de GSH dependendo da distribuição média dos genes do organismo, desconsiderando os centrômeros e telômeros (Arras *et al.*, 2015).

Para exemplificar os critérios de seleção de tamanho, no caso do genoma de *S. cerevisiae*, que possui em média 1 gene a cada 1.728 pb, a faixa escolhida para identificação de GSH foi entre 1.200 e 1.700 pb. De acordo com a Figura 6a é possível observar que a maioria das regiões intergênicas possuem menos de 1 kb de comprimento e com a maioria dos genes com orientação em *tandem*. Dessa maneira a faixa de tamanho escolhida é maior que a maioria das regiões intergênicas e apesar de eliminar diversos GSHs potenciais, minimiza possíveis perturbações na expressão gênica de genes vizinhos.

Para a execução do *software* SHIP, além do arquivo com as anotações genômicas do organismo alvo no formato padrão para anotação genômica (GFF3), são também necessários dois arquivos auxiliares contendo a qualidade da anotação e detalhes genômicos de cada organismo (*Features* e *Tracks*), para expandir o escopo do algoritmo e permitir flexibilidade de seleção com base na necessidade do usuário. Ambos os arquivos estão em formato *JSON* (do inglês, *JavaScript Object Notation*) e podem ser editados pelo usuário. O primeiro arquivo, denominado *Features*, indica os tipos de atributos genômicos que serão considerados pelo SHIP para a análise e busca de pGSHs. O segundo arquivo denominado *Tracks* indica os atributos genômicos que serão pesquisados dentro de cada região intergênica de cada candidato à pGSHs. Essa verificação é de caráter eliminatório, ou seja, regiões que apresentarem tais atributos serão eliminadas da lista final de pGSHs.

Para o genoma da levedura, o arquivo *Features* cobriu os seguintes atributos: genes, RNA não-codificador (do inglês, *Non-coding RNA*, ncRNA), pseudogenes, centrômeros, telômeros, repetição terminal longa (do inglês, *Long Terminal Repeats*, LTR), elementos genéticos móveis, origens de replicação, elementos transponíveis e regiões de recombinação meiótica. O conteúdo do arquivo *Tracks* foi: sequências de replicação autônoma, regiões teloméricas, cassetes de acasalamento, LTR, retrotransposons, origens de replicação,

centrômeros, mRNA, marcador de sequência expressa (do inglês, *Expressed Sequence Tag*, EST) e locais de ligação dos fatores de transcrição.

O algoritmo forneceu um arquivo de texto listando seis pGSHs distribuídos por cinco cromossomos de levedura (Tabela 9, Figura 6b). Primeiramente, pode-se observar que na maioria dos cromossomos não foram identificados sítios que acomodariam GSHs de acordo com os critérios estabelecidos, o que na realidade seria esperado e que já foi observado em outros organismos, à exemplo de humano e camundongo (Aznauryan *et al.*, 2022; Clark; Dinsmore; Soriano, 2020). Também pode-se verificar que a maioria dos pGSHs encontram-se nas extremidades dos cromossomos (Figura 6b). O Apêndice A contextualiza as pGSHs em termos dos respectivos genes vizinhos bem como sua anotação.

Tabela 9. Coordenadas genômicas, genes vizinhos e tamanho dos pGSH identificadas pelo SHIP em *S. cerevisiae*.

pGSH	Chromossomo	Coordenadas	Genes flanqueadores	Comprimento
1	II	807.106 - 808.599	MAL32 - YBR300C	1.494 pb
2	V	560.366 - 561.704	PUG1 - YER186C	1.339 pb
3	VIII	17.842 - 19.087	YHL041W - ARN1	1.246 pb
4	VIII	293.438 - 294.668	YHR095W - HXT5	1.231 pb
5	XI	15.709 - 17.358	COS9 - SRY1	1.650 pb
6	XV	697.335 - 698.767	ULS1 - THI72	1.433 pb

Em humanos e camundongos, o SHIP identificou 16 e 11 pGSHs, respectivamente (Figura 6b e 6c; descrição nos Apêndices B e C). Para organismos multicelulares, juntamente com a lista pGSH, o algoritmo também retornou um arquivo de texto com variações de características regulatórias de cada pGSH registradas em cada tipo de célula para uma determinada espécie. Todas as regiões foram revisadas manualmente, confirmando as funcionalidades e os critérios estabelecidos pelo algoritmo.

4.3. Geração de linhagens individuais de *S. cerevisiae* contendo os BioBricks em cada GSH

Por ser um organismo chassi, *S. cerevisiae* foi selecionado para a validação *in vivo* dos pGSHs identificados pelo *software* SHIP. Devido à relativa proximidade com uma região telomérica, o pGSH 1 foi descartado para prosseguir para análise experimental, e portanto, cinco pGSHs foram selecionados. Vale ressaltar que no ideograma da Figura 6a os pGSHs-2, 3 e 5 estão representados em escala próximos aos telômeros, enquanto a distância do pGSH-1 até o telômero é de 4.5 kb, os outros pGSHs tem no mínimo 15 kb de distância.

Com braços de homologia de 60 pb para os pGSHs e sobrepostos entre eles, os BioBricks (ymUKG1 e URA3) foram direcionados o mais centralmente possível dentro de cada pGSH para reduzir a probabilidade de perturbações na expressão de genes flanqueadores. Além disso, foram selecionados locais dentro de cada pGSH com marcas epigenômicas favoráveis, como a presença de H3K4me3 e locais hipersensíveis à DNase I (do inglês, *DNase I Hypersensitive Sites*, DHSs) que indicam cromatina aberta, com a perspectiva de melhorar a eficiência de inserção e de expressão desses transgenes.

Os BioBricks utilizados para transformação genética de *S. cerevisiae* foram usados como partes separadas de DNA, simulando as abordagens de montagem usadas para construir vias metabólicas e circuitos genéticos em organismos hospedeiros. Análises de PCR com oligonucleotídeos genômicos flanqueando os locais de inserção indicaram os transformantes com inserção putativa correta dos BioBricks (Figura 7), gerando as linhagens com inserção de ymUkG1 e URA3 no pGSH-2 (S2U), pGSH-3 (S3U), pGSH-4 (S4U), pGSH-5 (S5U) e pGSH-6 (S6U). Três clones de cada linhagem foram selecionados e seu DNA foi sequenciado para confirmar a inserção adequada. Cada clone pGSH apresentou a inserção exata do sítio (Figura 8), validando-os para continuar a análise funcional dessas regiões putativas como locais genômicos seguros. As próximas etapas investigaram os critérios genômicos, fenotípicos e de expressão genômica necessários para considerar um local genômico como GSH.

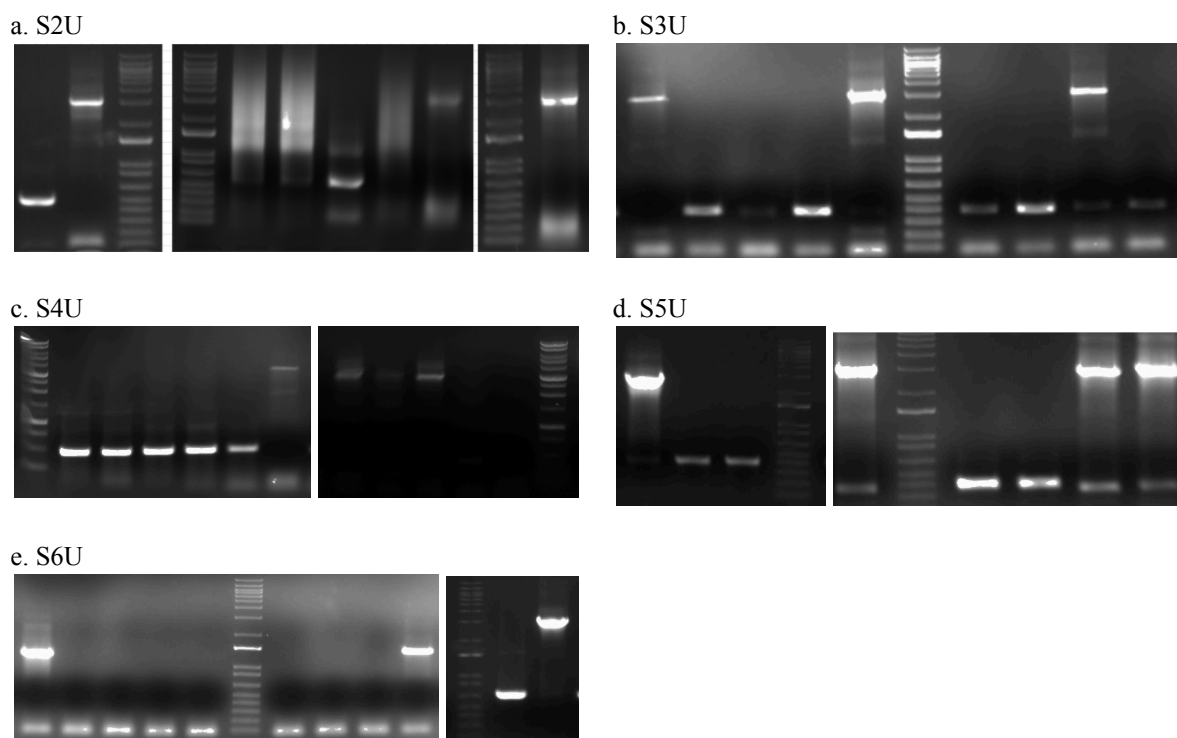
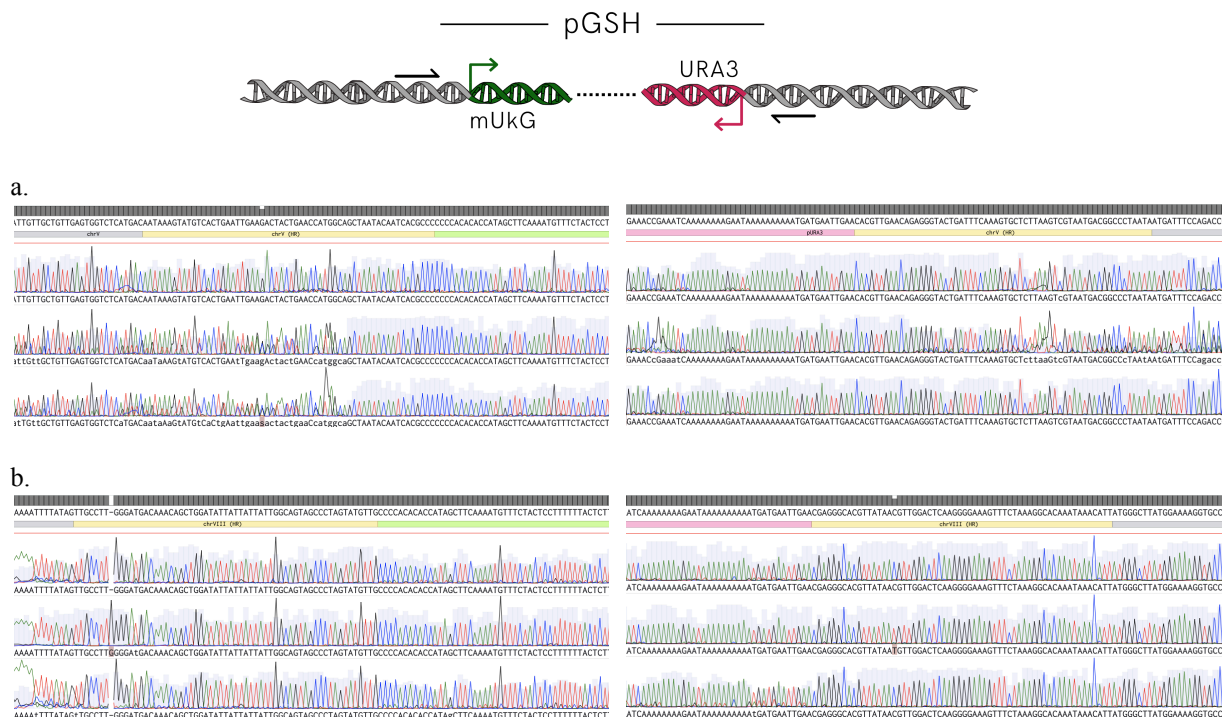


Figura 7 - Identificação dos clones com inserção correta na região de cada pGSH por análise de PCR. Clones com amplicon de tamanho aproximado de 3 kb, amplificado com oligonucleotídeos genômicos que flanqueiam a região de inserção dos BioBricks, indicam inserção correta no a) pGSH-2, b) pGSH-3, c) pGSH-4, d) pGSH-5 e último clone do e) pGSH-6. Nos dois primeiros clones pGSH-6 com oligonucleotídeo reverso complementar a sequência codificadora do gene UKG (Amplicon de 1.5 kb) ao invés de complementar a região genômica flanqueadora. Gel de agarose 1% com brometo de etídio e marcador molecular 1 kb Plus.



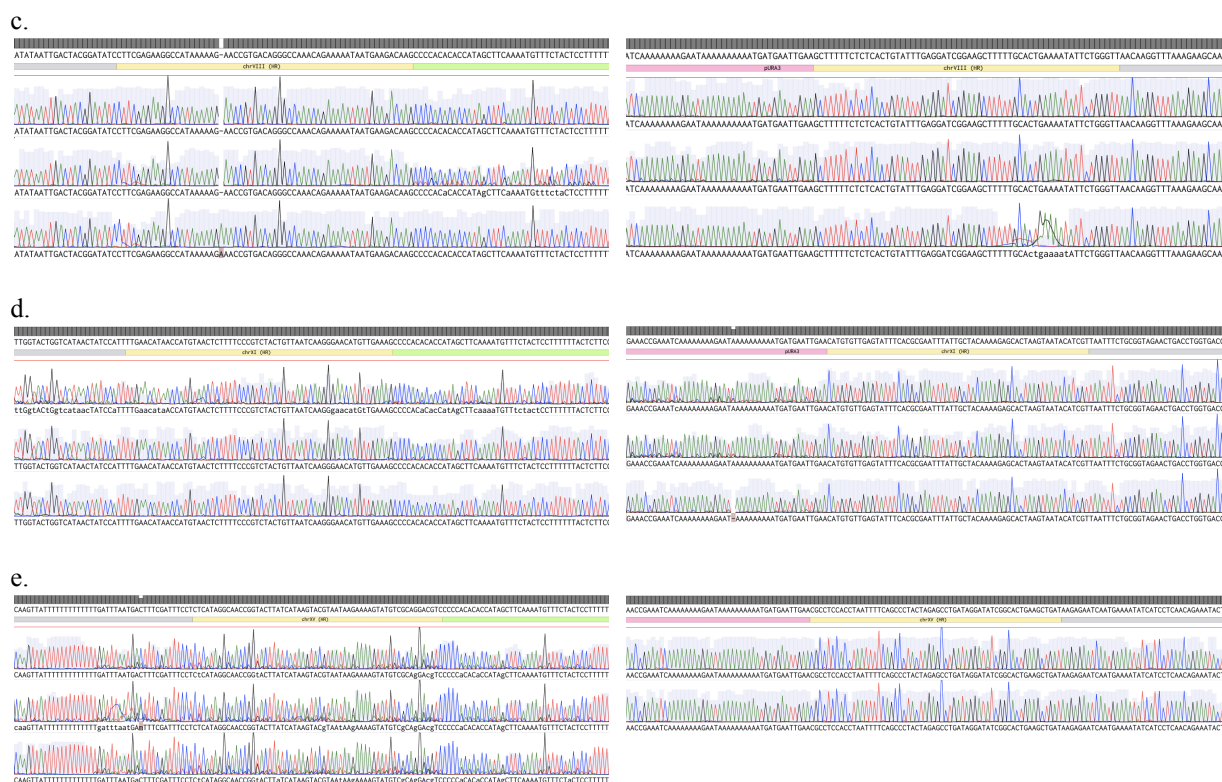


Figura 8 - Sequenciamento de Sanger dos clones demonstrando inserção correta na região de cada pGSH nas 5 linhagens. a) S2U, b) S3U, c) S4U, d) S5U e e) S6U. Alinhamento multisequência foi realizado com o programa MAFFT v7 (Katoh; Standley, 2013) utilizando os seguintes parâmetros: Penalidade de abertura de lacuna de 1,53; Penalidade de extensão de lacuna igual a 0 e a função de ajuste rápido de direção habilitado.

4.4. GSHs identificados apresentam estabilidade genômica ao longo de 10 dias sem afetar o crescimento da levedura

Uma das características primordiais de um GSH é a sua capacidade de manter a inserção exógena através das gerações sem interferência no crescimento celular. Usando o experimento proposto por Annaluru *et al.* (2014), as cinco linhagens de leveduras foram cultivadas em YPD líquido por dez dias continuamente, e a presença dos BioBricks dentro de cada GSH foi verificada por PCR utilizando oligonucleotídeos genômicos.

Todos os pGSHs apresentaram estabilidade genômica após 10 dias de cultivo contínuo (Figura 9), pois todos os 3 clones de cada linhagem apresentaram amplicons de aproximadamente 3 kb, condizente com a inserção dos dois BioBricks usados nesse estudo.

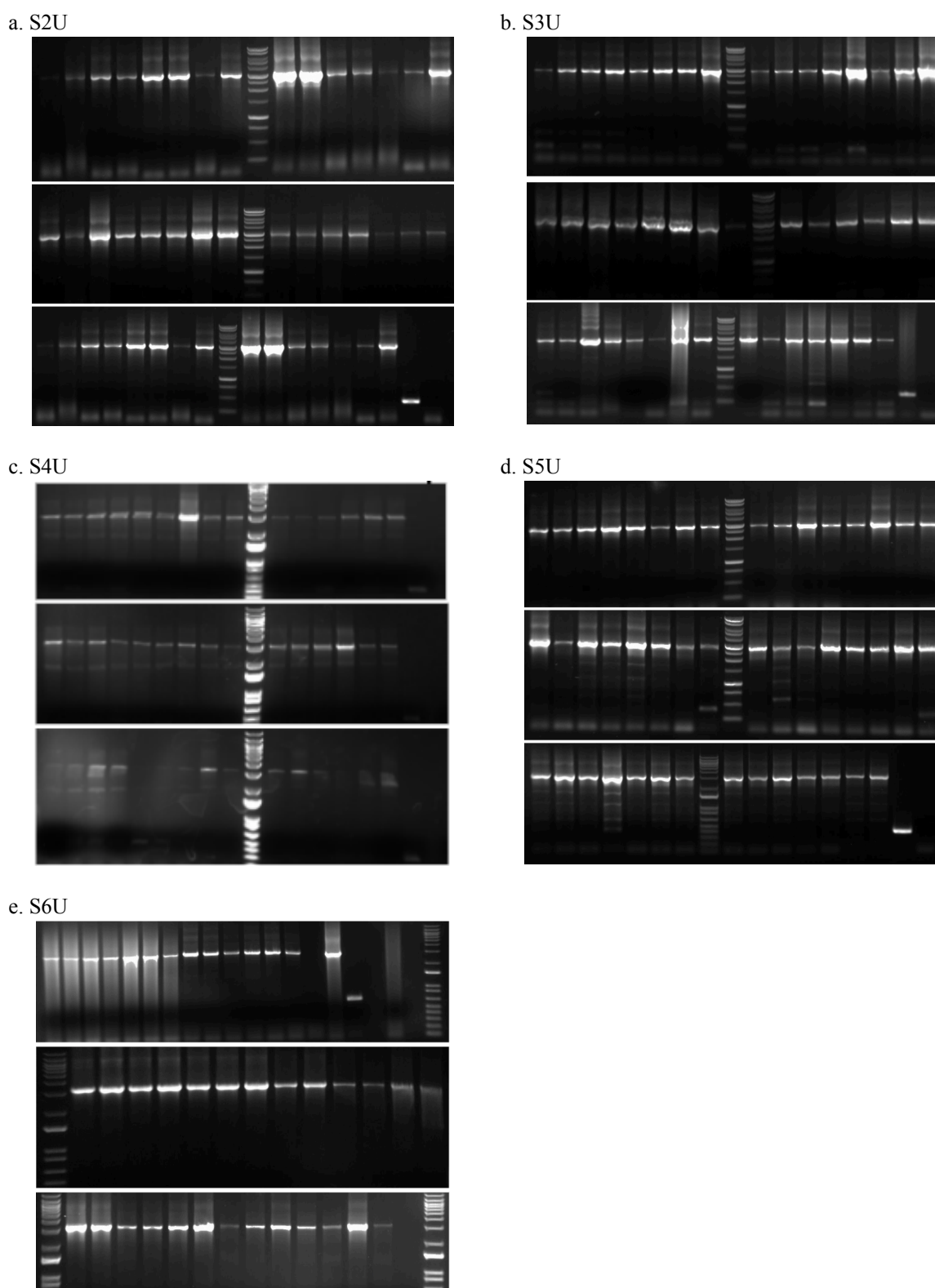


Figura 9 - Teste de estabilidade genômica de cada pGSH após 10 dias de cultivo contínuo. Análise de PCR realizada com oligonucleotídeos genômicos que flanqueiam cada pGSH nas 5 linhagens a) S2U, b) S3U, c) S4U, d) S5U e e) S6U, seguida por eletroforese em gel de agarose 1% obtendo amplicons de aproximadamente 3 kb, tamanho esperado considerando a inserção dos dois BioBricks utilizados neste estudo. Cinco clones de cada triplicata biológica e técnica foram analisados, totalizando 30 clones. Os 3 géis das triplicatas estão um abaixo do outro. Gel de agarose 1% com brometo de etídio com marcador molecular 1 kb Plus, aplicado 1 ou 2 vezes.

A curva de crescimento para cada uma das 5 linhagens pGSH revelou um perfil de crescimento semelhante entre elas, mas diferente ao controle (BY4741 não transformado) após 36 horas de cultivo em meio YPD (Figura 10). As taxas de crescimento calculadas variaram entre 0,35 e 0,39 para as linhagens pGSHs e 0,30 para o controle indicando diferença detectável de valor adaptativo entre eles, sendo mais altas para as linhagens onde a prototrofia de uracila foi restaurada.

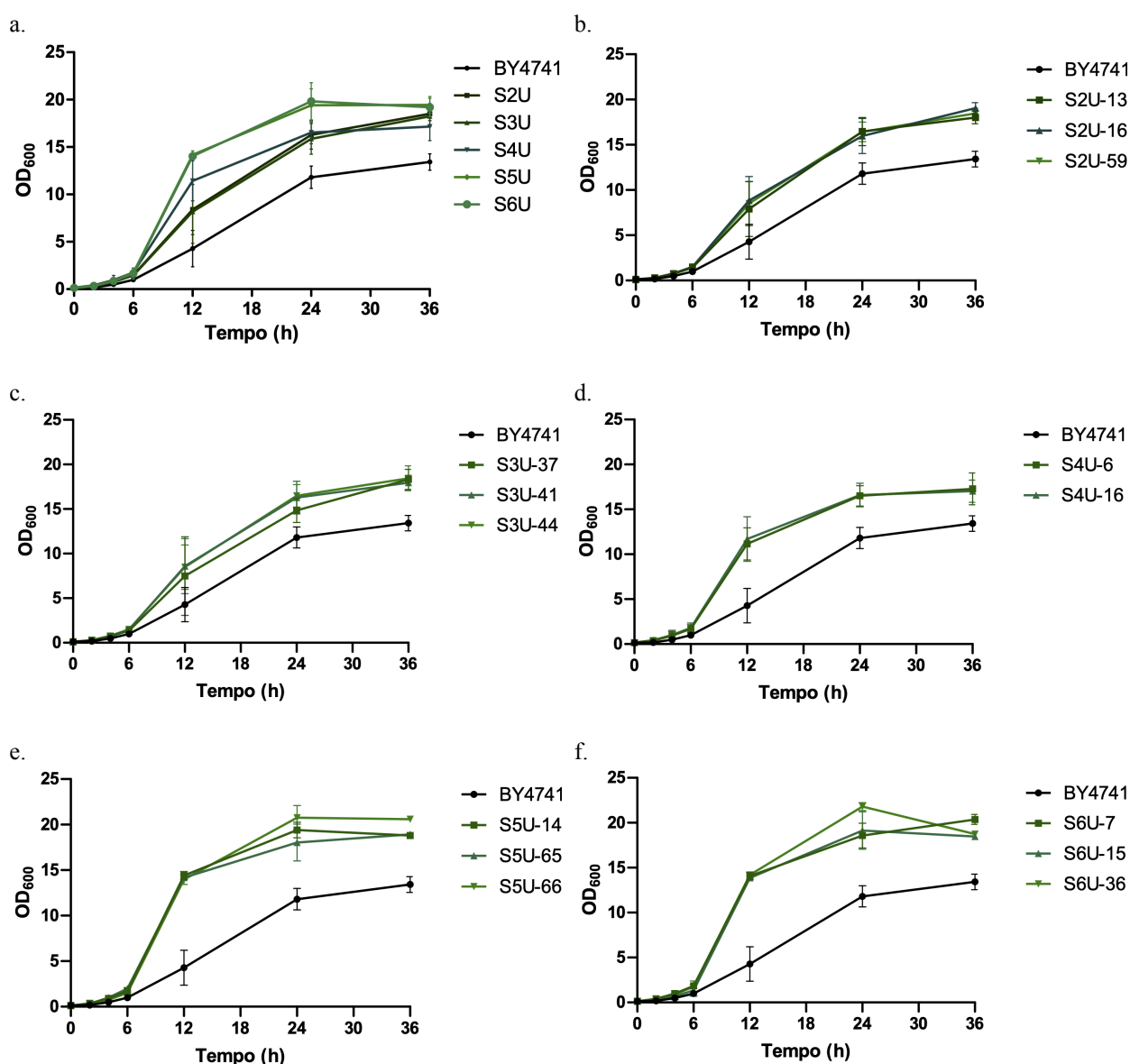
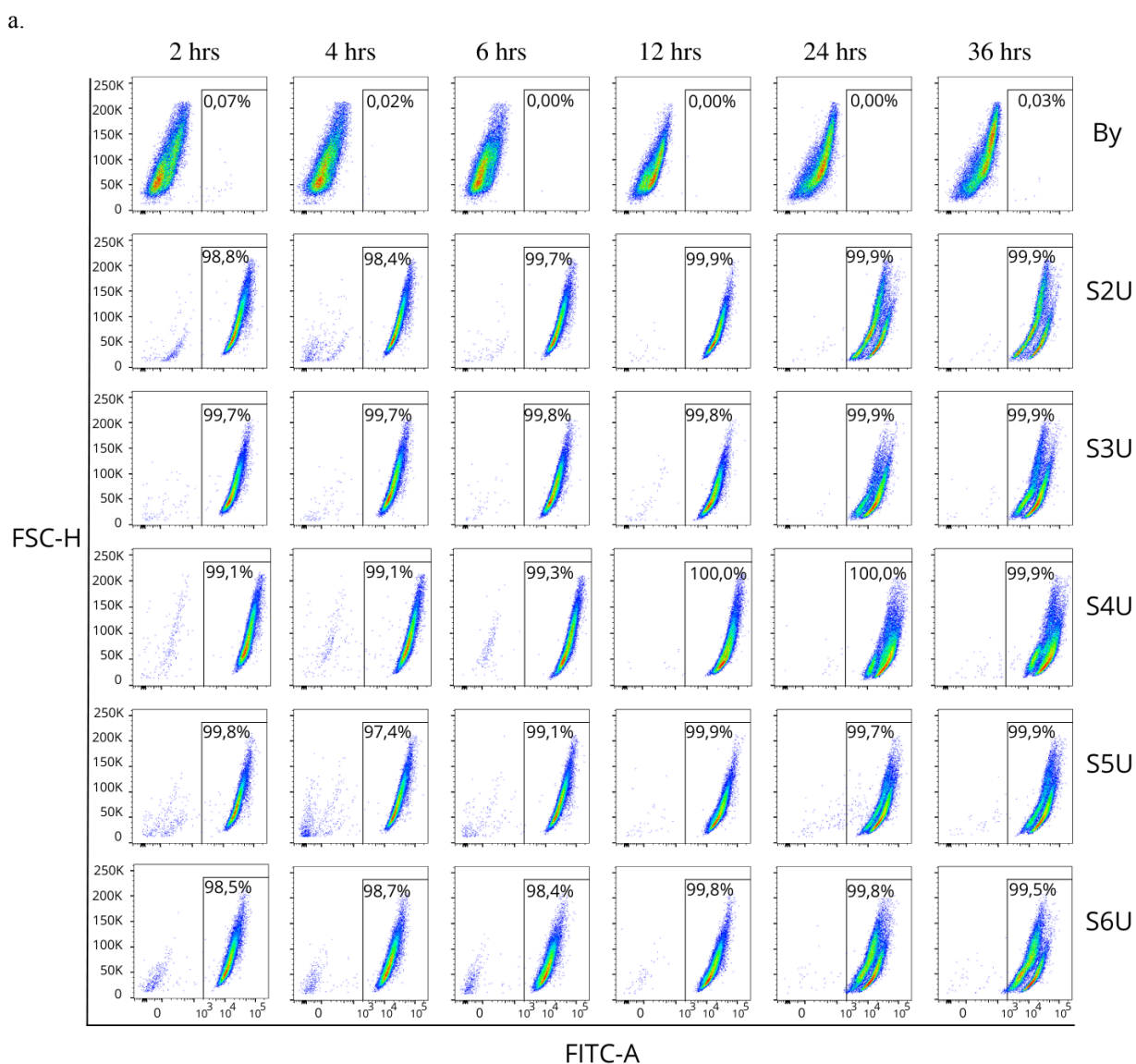


Figura 10 - Curva de crescimento para cada linhagem pGSH. Curvas apresentam a média e os desvios padrões das curvas de crescimento medidas em 0, 2, 4, 6, 12, 24 e 36 horas com uma OD₆₀₀ inicial de 0,1 de cada linhagem a) Média dos três clones de cada linhagem, b) S2U, c) S3U, d) S4U, e) S5U e f) S6U. A cepa BY4741 foi utilizada como controle e ensaios feitos em triplicatas técnicas e biológicas realizadas em três dias independentes. Traços após a identificação da cepa indicam o número do clone.

4.5. Validação funcional das linhagens celulares por citometria de fluxo

Utilizou-se a citometria de fluxo para confirmar a presença da proteína codificada pelo gene repórter em nível populacional para cada uma das 5 linhagens pGSH. A fluorescência média no comprimento de onda emitido pela proteína repórter codificada pelo gene ymUKG1 foi detectável em 90% da população (Figura 11a). Esse percentual foi alcançado e mantido nas primeiras duas horas até às 36 horas de crescimento. Nenhuma variação percentual detectável foi observada na população fluorescente nos diferentes horários de cultivo e entre os pGSHs (Figura 11b), indicando que cada linhagem é capaz de expressar e manter a expressão do transgene.



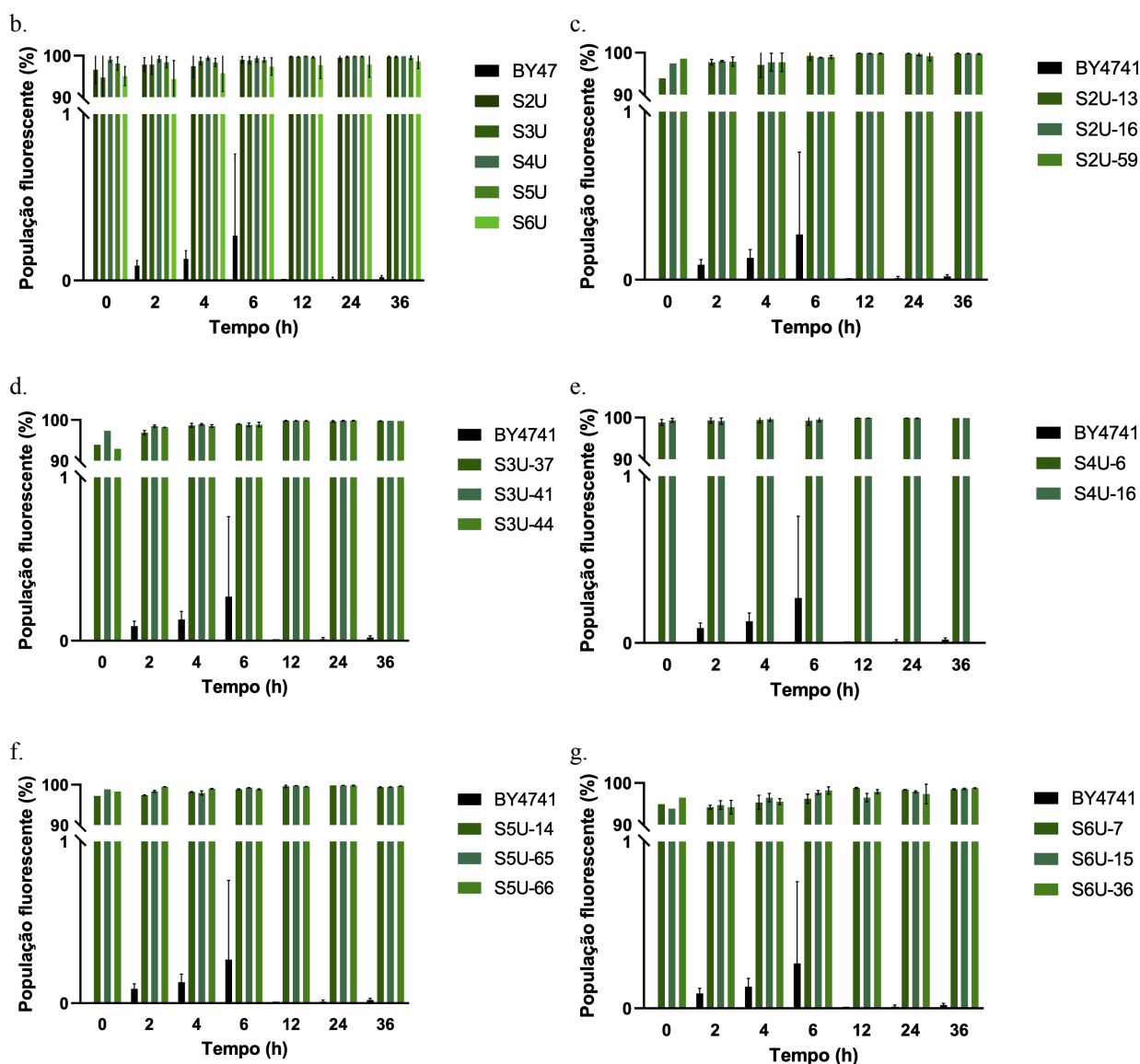


Figura 11 - Validação funcional das linhagens celulares por citometria de fluxo realizada em triplicata técnica e biológicas em três dias independentes. a) Imagem representativa da distribuição de citometria de fluxo das cinco linhagens de células pGSH medidas em 2, 4, 6, 12, 24 e 36 horas; S2U, S3U, S4U, S5U e S6U; BY4741 (By) foi usado como controle. Gráficos mostrando a média e o desvio padrão da porcentagem de células com expressão de ymUKG1. (b) Todas as linhagens, (c) S2U, (d) S3U, (e) S4U, (f) S5U e (g) S6U.

4.6. Análise da expressão dos genes vizinhos aos GSHs

Uma das características fundamentais de um GSH também é sua capacidade de aporte do transgene sem a interferência da expressão local do genoma hospedeiro. Começando pela região diretamente em contato aos GSHs (Genes Vizinhos), a técnica de RT-qPCR foi realizada para verificar a expressão do transgene e possíveis perturbações da expressão dos genes vizinhos a nível transcricional. Os resultados demonstram acúmulo do

transcrito do gene ymUKG1 para todas as linhagens celulares pGSHs. Ao se analisar todos os genes vizinhos individualmente para todas as linhagens, quatro genes apresentaram uma variação persistente da expressão em todas as linhagens celulares pGSHs. Esses genes são: PUG1 (Vizinho do pGSH 2), YHL041W (pGSH 3), YHR095W (pGSH 4) e COS9 (pGSH 5). Tal padrão de expressão sugere uma variação não correlacionada com a localização de cada GSH nas expressões gênicas (Figura 12). Notavelmente, conforme descrição presente na lista de pGSHs resultante do *software* SHIP (Apêndice D), os genes PUG1 e COS9 codificam proteínas de membrana envolvidas no transporte de protoporfirina/heme e na renovação de proteínas de membrana, respectivamente. YHL041W é uma ORF não caracterizada que codifica uma proteína hipotética, e YHR095W é um quadro de leitura aberto não caracterizado funcionalmente.

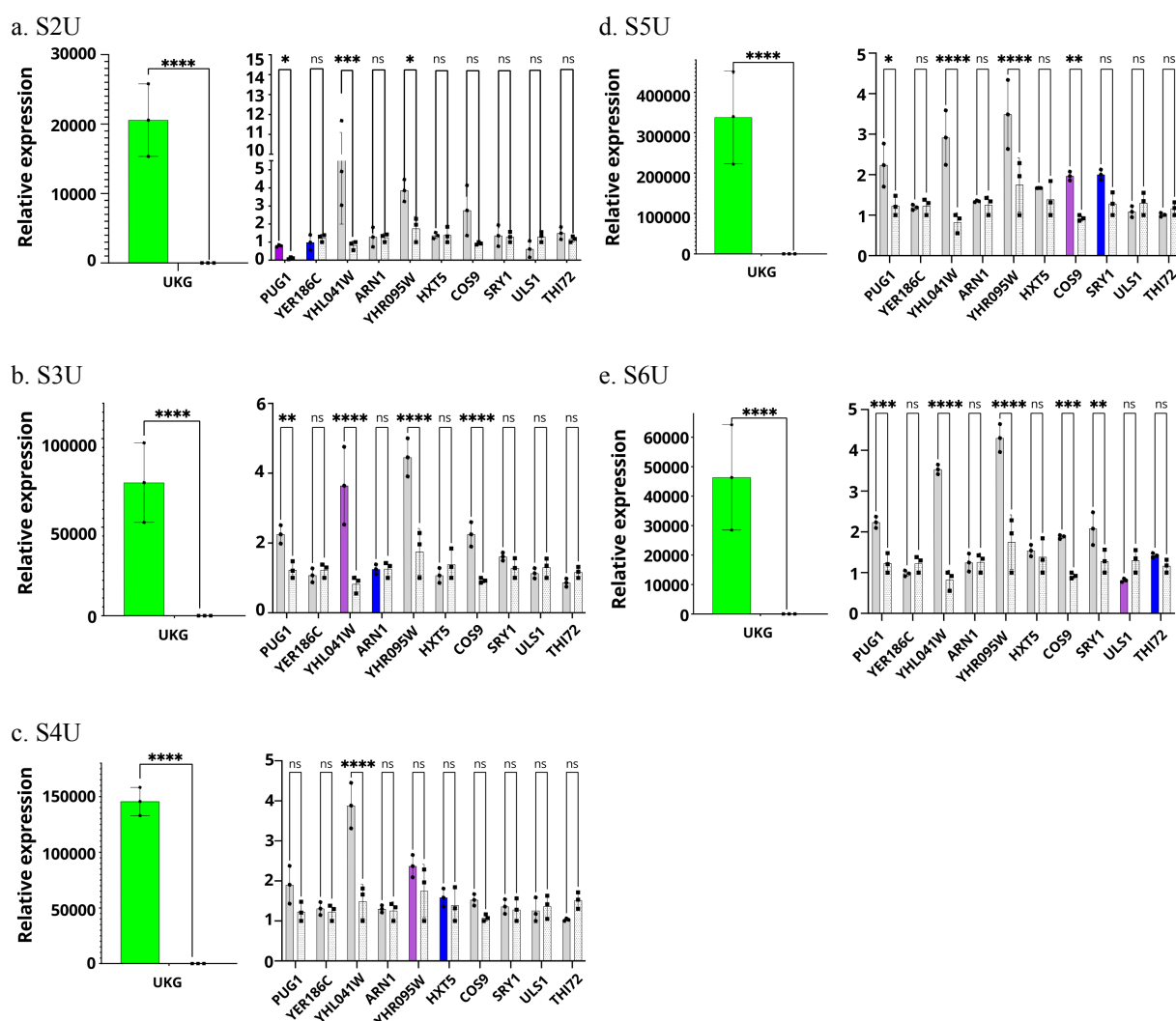


Figura 12 - Dinâmica de expressão dos genes vizinhos das linhagens celulares pGSHs por RT-qPCR. Histograma da expressão relativa de todos os genes vizinhos para cada linhagem pGSH a) S2U, b) S3U, c)

S4U, d) S5U e e) S6U. Coluna verde indica a expressão do gene *ymUKG1* (UKG). Colunas coloridas destacam genes vizinhos (Roxo à 5' e Azul à 3') do pGSH da linhagem analisada e colunas cinzas mostram os genes vizinhos dos demais pGSHs. As colunas rachuradas representam valores do controle não transformado (By4741). Expressão relativa no eixo Y e genes analisados no eixo X. Asteriscos indicam alteração significativa na análise de variância (ANOVA). (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$ e (****) $p < 0.0001$.

Na Tabela 10 é possível encontrar uma simplificação quantitativa dos resultados encontrados nos dados de RT-qPCR presente na Figura 12. Essa tabela exemplifica o efeito não correlacional da variação da expressão diferencial dos genes vizinhos com a localização genômica dos pGSHs.

Tabela 10. Resumo quantitativo da expressão diferencial significativa dos genes vizinhos nas linhagens pGSHs.

pGSH	Gene Vizinho	Expressão Diferencial Significativa na Linhagem				
		S2U	S3U	S4U	S5U	S6U
2	PUG1	X	X		X	X
	YER186C					
3	YHL041W	X	X	X	X	X
	ARN1					
4	YHR095W	X	X		X	X
	HXT5					
5	COS9		X		X	X
	SRY1					X
6	ULS1					
	THI72					

4.7. Análise global de expressão gênica

Finalmente, após análise dos vizinhos, um outro importante critério de identificação de GSH é a sua capacidade de garantir uma transcrição previsível do BioBrick com nenhuma ou mínima interferência na expressão genômica do hospedeiro.

Uma análise global de expressão gênica foi realizada utilizando a técnica RNA-Seq para investigar possíveis interferências da inserção do BioBrick em um clone de cada pGSH, uma vez que a inserção ocorreu precisamente no mesmo local. Cada construção de pGSH teve seu transcrito analisado em duplicada biológica contra a linhagem controle BY4741 (não transformada). Em termos gerais, os clones de pGSHs apresentaram aproximadamente 16% de genes diferencialmente expressos em relação ao controle, mas alta similaridade entre eles (Figura 13).

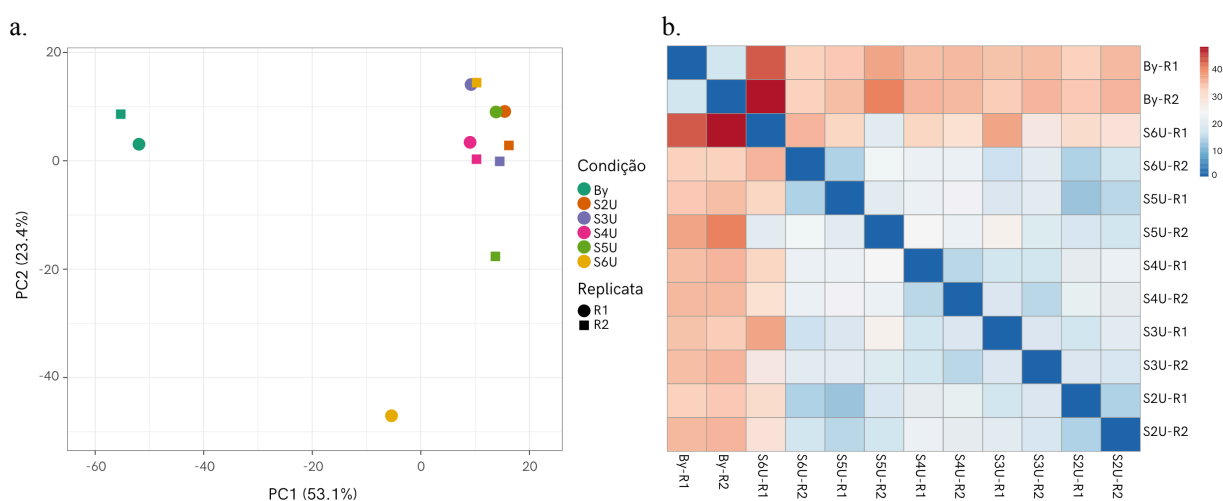


Figura 13 - Dinâmica de expressão genômica das linhagens celulares pGSHs por RNA-Seq. a) Gráfico de PCA da expressão gênica diferencial das linhagens pGSHs (S2U, S3U, S4U, S5U e S6U) em relação ao controle não transformado By4741 (By). b) *Heatmap* da expressão gênica diferencial entre as linhagens pGSHs e o controle By4741 (By).

Investigando essa hipótese, os resultados da análise de expressão diferencial do experimento de RNA-seq para as os pGSHs foram processados de forma a elucidar: (1) o nível de compartilhamento dos genes diferencialmente expressos entre os pGSHs, que estão em sítios distintos localizados em diversas partes do genoma de levedura e se (2) entre os genes compartilhados, existe algum enriquecimento funcional.

Para todas as linhagens pGSHs, foi realizada uma análise de expressão diferencial (DE) pareada contra o controle não transformado. Para cada experimento foi gerada uma lista de genes regulados positivamente (do inglês, *upregulated*) e negativamente (do inglês, *downregulated*). Em seguida, foi realizada uma busca das interseções entre os grupos de genes *up* ou *downregulated*, juntando as listas de genes, em termos dos códigos dos genes

presentes no banco de dados de *Saccharomyces* (do inglês, *Saccharomyces Genome Database*, SGD).

No grupo de genes regulados positivamente de todos os experimentos somam um total de 1.257 genes únicos que tiveram a sua expressão significativamente elevada (Figura 14). Destes, 423 (33,65%) tiveram aumento de expressão ($\log_2$ fold change maior que 2 com p-value menor que 0.05) em todos os experimentos, sendo o subgrupo de genes mais frequente dentre as combinações de clones, ou seja a maioria dos genes com expressão aumentada foram comuns em todos casos testados. Isso pode indicar que a própria transformação genética causou uma perturbação na expressão gênica, independentemente do local de inserção do transgene. Vale ressaltar que os clones S6U e o S4U apresentaram um número razoável de genes regulados positivamente, 120 e 83 respectivamente, que não foram compartilhados com as outras linhagens. Para os genes regulados negativamente de todos os experimentos foi obtido um total de 1.086 genes únicos que tiveram a sua expressão significativamente diminuída. Destes, 257 genes (23,66%) tiveram menor expressão em todos os experimentos comparado com o controle, também que constitui o subgrupo de genes mais frequente, como no caso dos genes com aumento de expressão.

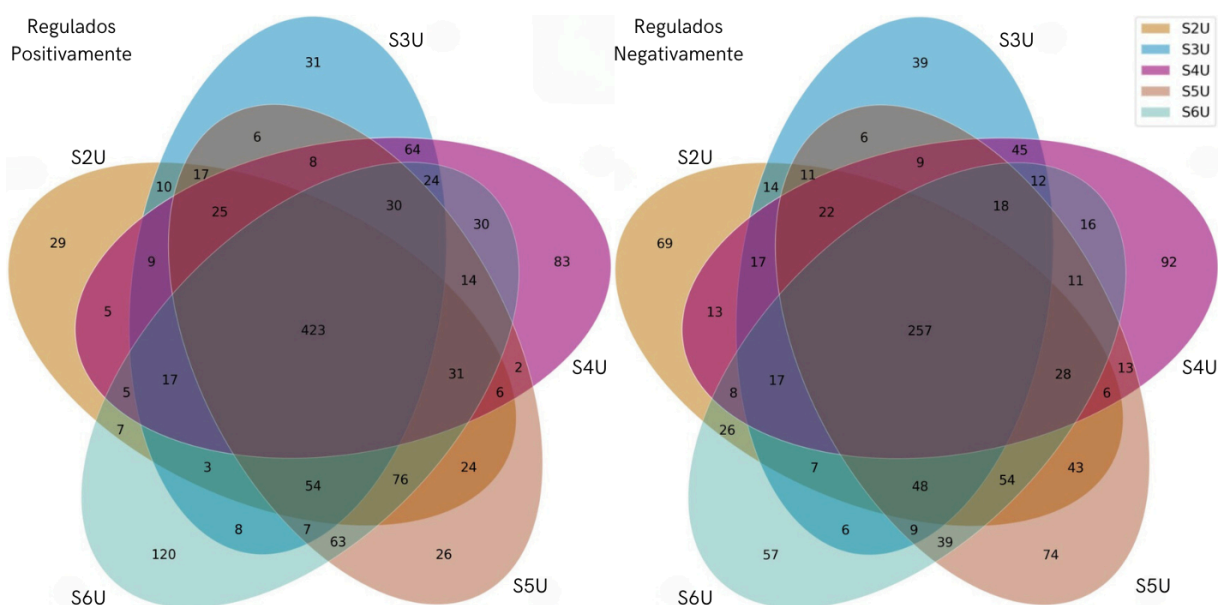


Figura 14 - Diagramas de Venn do número de genes diferencialmente expressos únicos e compartilhados entre as linhagens pGSHs. A elipse amarela representa a S2U, S3U em azul, S4U em roxo, S5U em marrom e S6U em verde piscina. Valores indicam o número de genes diferencialmente expressos.

Para verificar se esses grupos de genes compartilhados, regulados positivamente ou negativamente, estão relacionados com as pGSHs, foi feito o mapeamento desses dados para as regiões cromossômicas. Os dados de RNA-Seq dos genes diferencialmente expressos em relação ao controle foram filtrados para genes com razão $\log_2$ de expressão gênica (do inglês, $\log_2$ *Fold Change*, $\log_2$ FC) maior que 2 e valor de $p < 0,05$, após a estimativa da expressão diferencial. Desta maneira é possível verificar que as regiões com genes diferencialmente expressos estão espalhadas pelo genoma da levedura sem correlação posicional como exemplificado pelo pGSH-2 (Figura 15). Os ideogramas dos demais pGSHs estão nos Apêndices E.

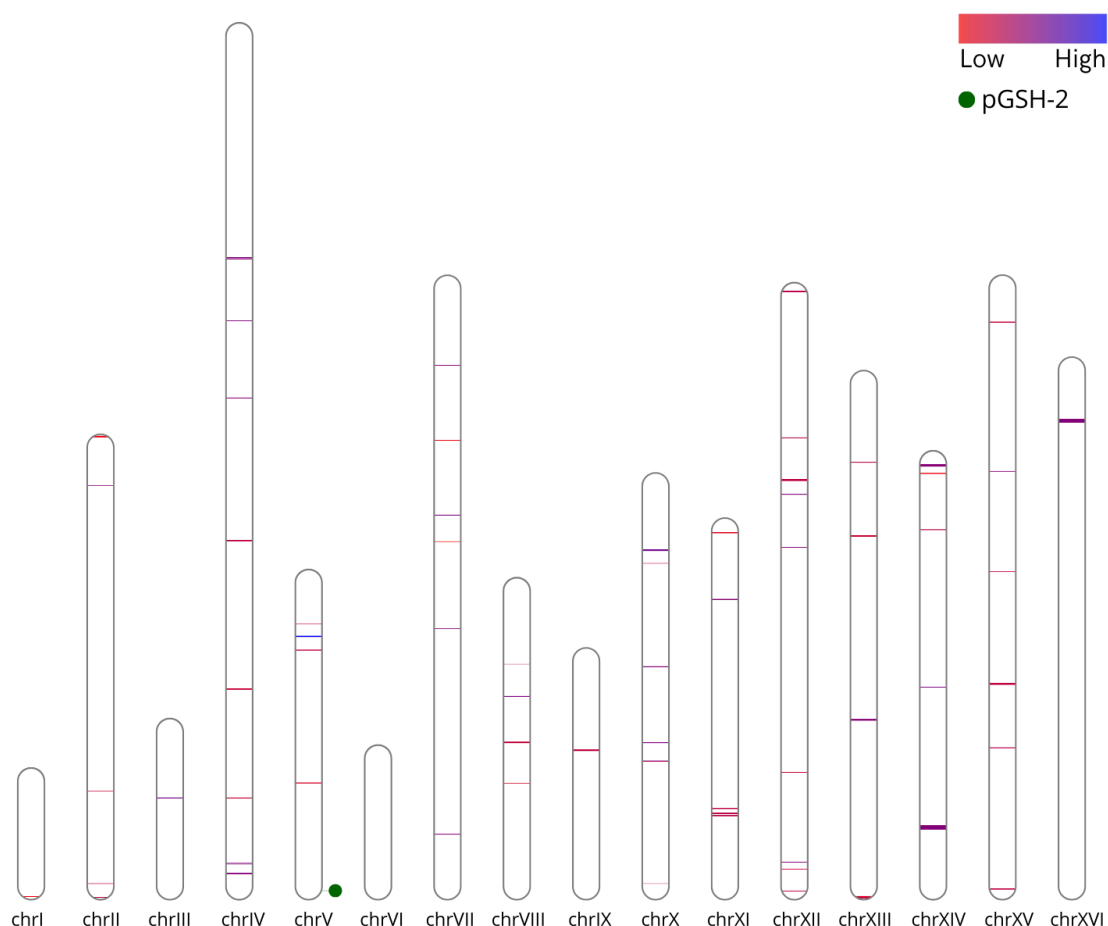


Figura 15 - Ideograma mostrando a localização de regiões genômicas com genes diferencialmente expressos na linhagem S2U. Os genes são coloridos de acordo com a alteração de $\log_2$ FC > 2 . O vermelho indica regulação negativa e o azul, regulação positiva. O loco do pGSH-2 está representado com um círculo verde.

Vale ressaltar também que não há sobreposição dessas regiões de genes diferencialmente expressos com os locos dos pGSHs. Para cada pGSH, foi identificado o gene diferencialmente expresso positivamente ou negativamente mais próximo com $\log_2FC > 2$. Com essas informações é possível inferir a distância da perturbação local na expressão gênica afetada pela inserção do transgene no pGSH. A Tabela 11 abaixo apresenta a lista de genes encontrados, coordenadas e em quais linhagens foi identificado. Assim é possível notar que na maioria das vezes, o $\log_2FC$ é alto independente do pGSH, indicando que não é a proximidade ao pGSH que induz a alteração da expressão.

Tabela 11. Descrição e características dos genes diferencialmente expresso mais próximo com $\log_2FC > 2$ de cada pGSH.

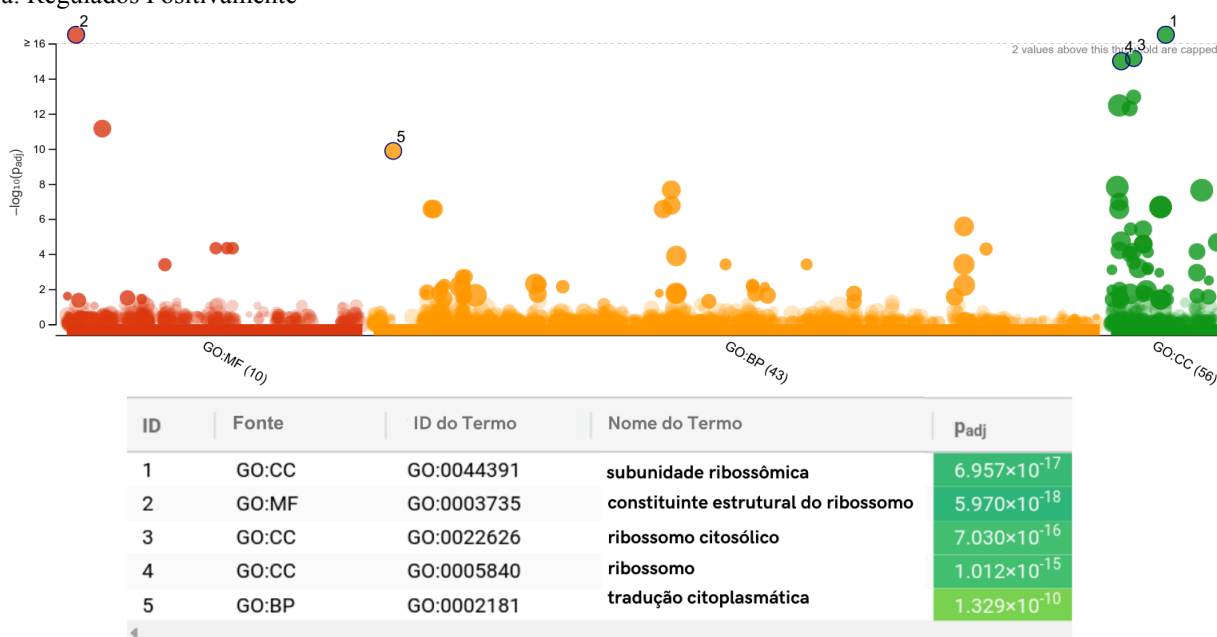
pGSH	Gene	Coordenadas	Linhagem	$\log_2FC$	Anotação (SGD)
2	YER158W-A	V: 491487 - 491702	S3U	2.67	Quadro de leitura aberto duvidoso; é improvável que codifique uma proteína funcional.
			S4U	2.96	
			S5U	1.72	
3	YHL006W-A	VIII: 97732 - 98085	S2U	1.51	Quadro de leitura aberto duvidoso; é improvável que codifique uma proteína funcional, sobrepõe a ORF YHL007C (STE20).
			S3U	2.40	
			S4U	1.51	
			S5U	2.28	
			S6U	1.90	
4	YHR092C (HXT4)	VIII: 287081 - 288811	S2U	-3.04	Transportador de glicose de alta afinidade; membro da superfamília de facilitadores principais, expressão induzida por baixos níveis de glicose e reprimida por altos níveis de glicose; possui um parálogo, o HXT7.
			S3U	-3.67	
			S4U	-3.45	
			S5U	-3.25	
			S6U	-2.46	
5	YKL216W (URA1)	XI: 25215 - 26159	S2U	-5.77	Diidroorotato desidrogenase; catalisa a quarta etapa enzimática na biossíntese de novo de pirimidinas, convertendo o ácido diidroorótico em ácido orótico.
			S3U	-4.97	
			S4U	-5.01	
			S5U	-5.51	
			S6U	-5.47	
6	YOR195W (SLK19)	XV: 712866 - 715331	S2U	-2.62	Proteína associada ao cinetocoro; necessária para a segregação normal de cromossomos na meiose e mitose; componente da rede reguladora FEAR, que promove a liberação de Cdc14p do nucléolo durante a anáfase;
			S3U	-1.82	
			S4U	-1.73	
			S5U	-2.42	
			S6U	-2.30	

4.7.1. Análise de enriquecimento de ontologia gênica dos genes diferencialmente regulados

Uma vez elucidado o efeito posicional da expressão diferencial observado nos dados de RNA-Seq é preciso verificar possíveis perturbações funcionais no genoma do hospedeiro. Para isso foi realizado uma análise de enriquecimento de ontologia gênica dos genes diferencialmente regulados que foram compartilhados entre todas as cepas para obter-se uma visão geral do tipo de perturbação consistente causada na expressão gênica. Os grupos de genes compartilhados foram submetidos a uma análise de enriquecimento do conjunto de genes utilizando o programa g:Profiler (Kolberg *et al.*, 2023).

No caso dos genes regulados positivamente houve apenas um enriquecimento significativo para o termo de componente celular (CC) de unidade ribossômica e para a função molecular (MF) relacionada à componente estrutural do ribossomo (Figura 16a). Em termos gerais, isso indica uma maior atividade traducional. Em termos de metabolismo ou resposta a estresses nada significativo foi encontrado. Notavelmente, para as cepas S6U e o S4U que apresentaram um número razoável de genes regulados positivamente, a análise de enriquecimento desses genes revelou que não existe nenhum termo de ontologia significativamente enriquecido. Para o grupo de genes regulados negativamente compartilhados em todos as construções de pGSHs não houve nenhum enriquecimento de termos de ontologia gênica (Figura 16b).

a. Regulados Positivamente



b. Regulados Negativamente

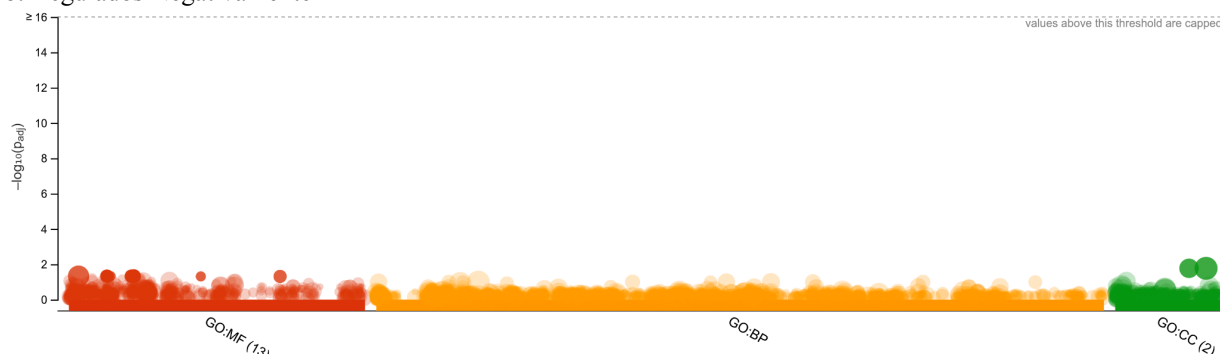
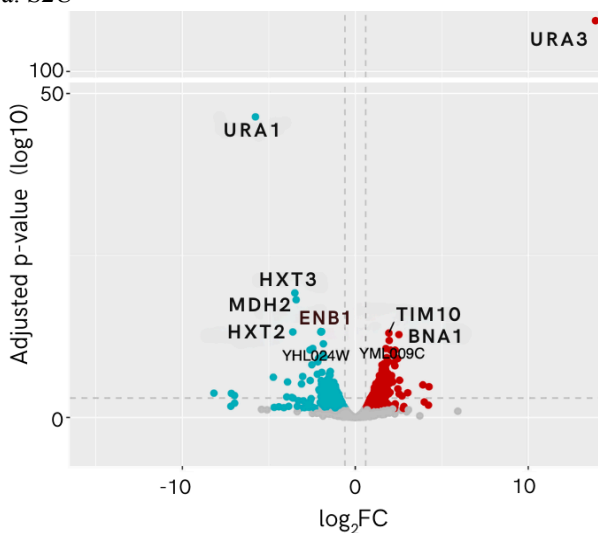


Figura 16 - Análise de Enriquecimento de ontologia gênica dos genes compartilhados diferencialmente regulados. a) Genes regulados positivamente. b) Genes regulados negativamente. Ontologia gênica (do inglês, *Gene Ontology*, GO), Função Molecular (do inglês, *Molecular Function*, MF) (Vermelho), Processo Biológico (do inglês, *Biological Process*, BP) (laranja) e Componente Celular (do inglês, *Cellular Component*, CC) (Verde). A numeração dos círculos corresponde ao ID na tabela de enriquecimento. Linha tracejada em cinza indica nota de corte de enriquecimento significativa.

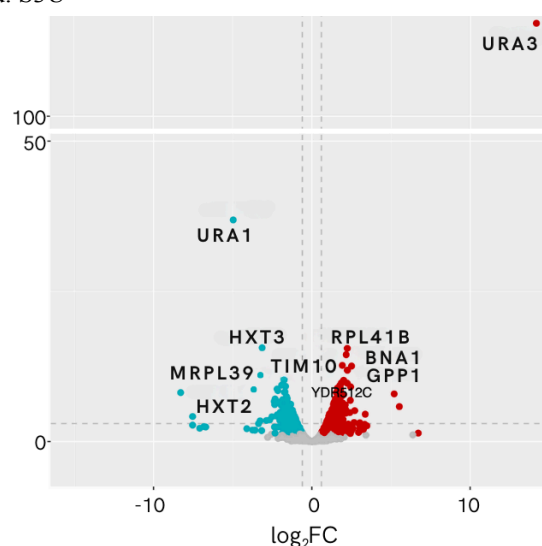
4.7.2. Análise individual dos genes mais diferencialmente expressos

Mesmo sem enriquecimento ontológico substancial, os genes mais diferencialmente expressos foram analisados para se obter um entendimento mais detalhado da expressão diferencial dos clones em relação ao controle não transformado. Em termos gerais, os experimentos de RNA-Seq das linhagens pGSH mostraram que a distribuição da expressão dos genes é afetada mostrando que o gene URA3 (YEL021W) é o mais positivamente regulado e o gene URA1 (YKL216W) o mais negativamente regulado em todos os experimentos (Figuras 17 e 18).

a. S2U



d. S3U



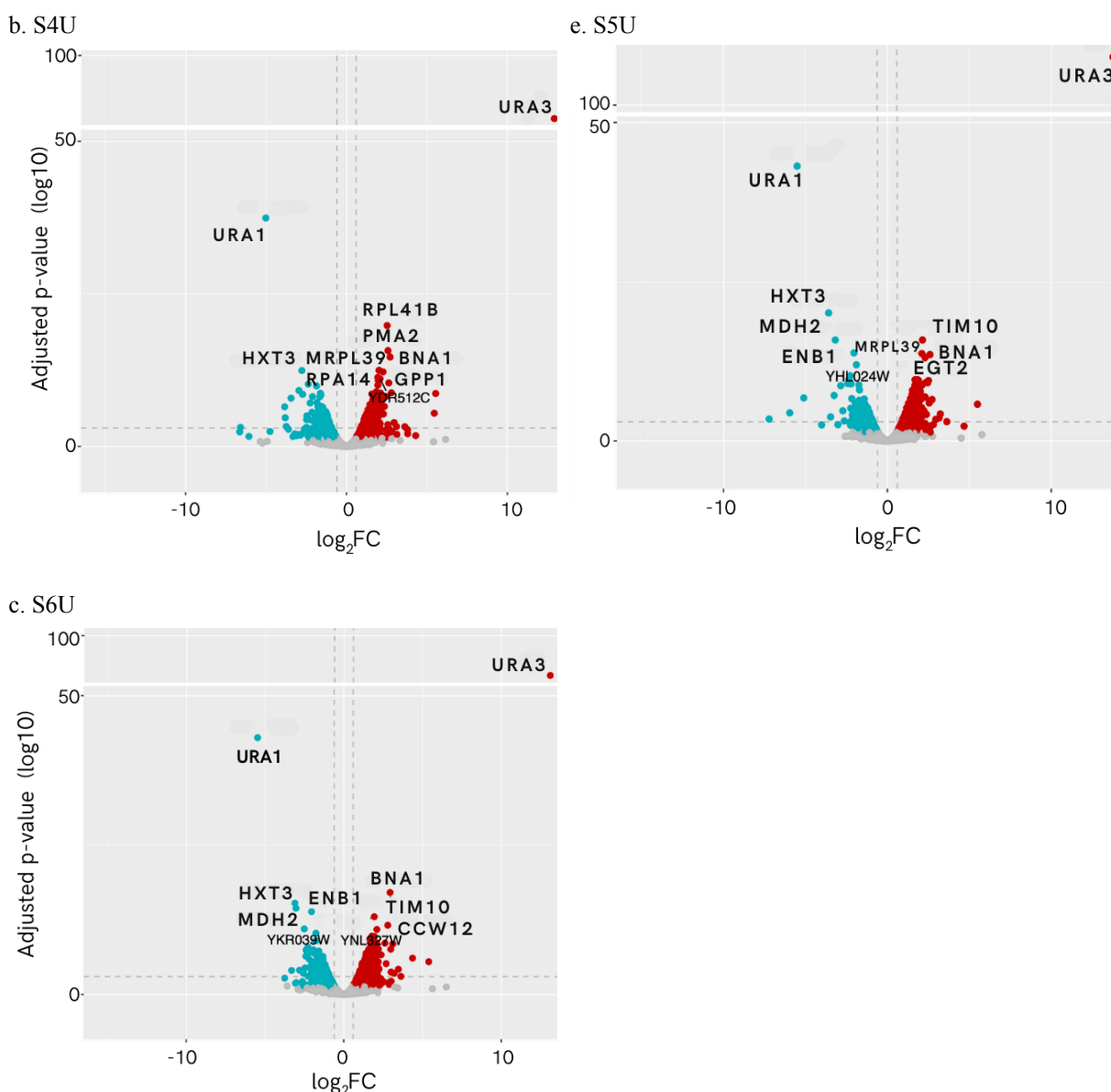


Figura 17 - Vulcanos Plots da expressão diferencial das linhagens celulares pGSHs por RNA-Seq. Pontos coloridos representam genes diferencialmente regulados negativamente (azul) e positivamente (vermelho). Genes descritos estão identificados com seu nome padrão e os genes não descritos ou putativos estão identificados com seu nome sistemático.

Além disso a Figura 17 mostra que os genes TIM 10 (YHR005C-A; Proteína do espaço intermembranar mitocondrial) e BNA1 (YJR025C; Dioxigenase do ácido 3-hidroxiantranílico) são positivamente regulados nas linhagens pGSHs e os genes HXT3 (YDR345C; Transportador de glicose), ENB1 (YOL158C; Transportador transmembranar de enterobactina férrica) e MDH2 (YOL126C, Malato desidrogenase citoplasmática) são negativamente regulados nessas linhagens.

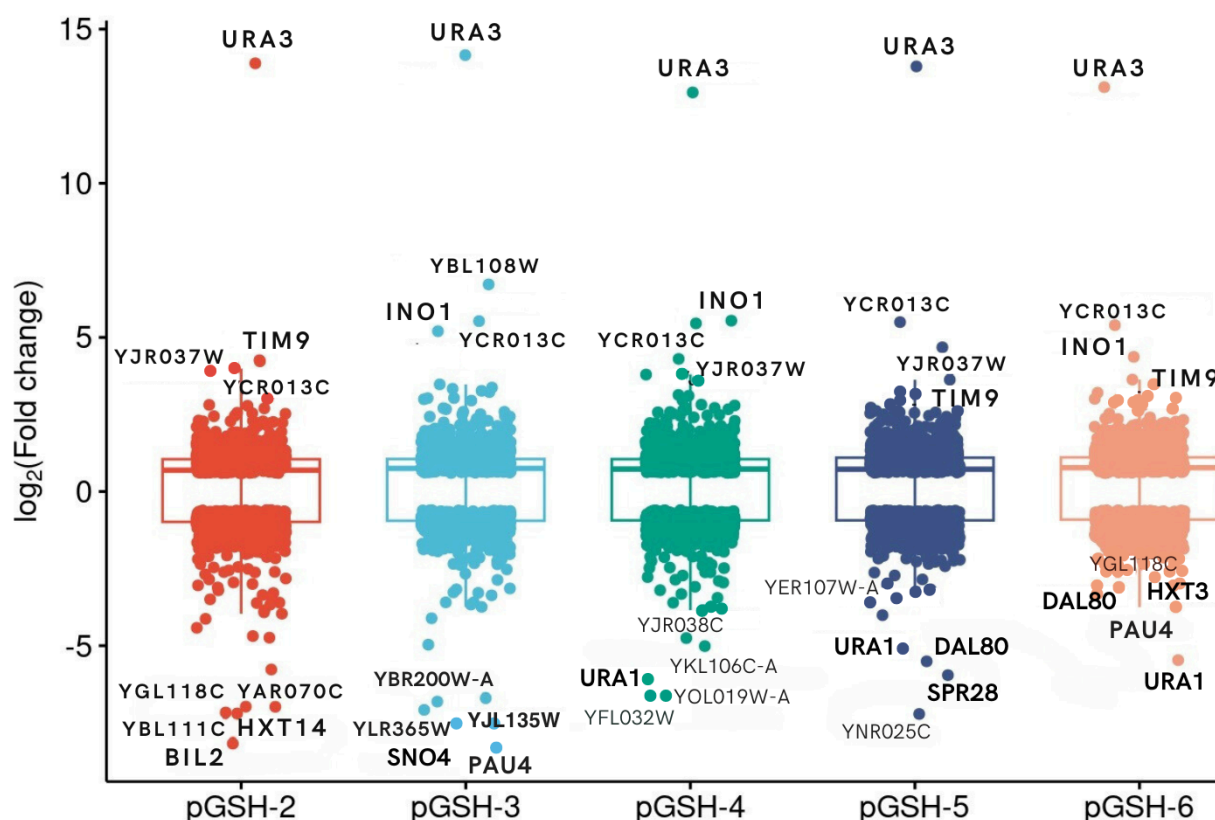


Figura 18 - Gráfico *Box Plot* da alteração de $\log_2FC$ dos genes diferencialmente expressos nas construções pGSH. Cada circula representa um gene. Genes descritos estão identificados com seu nome padrão e os genes não descritos ou putativos estão identificados com seu nome sistemático.

Para ampliar o grupo de genes analisados na Figura 18 foi apenas considerado o $\log_2FC$ acima de 2. Dentro de cada grupo de genes regulados positivamente e negativamente, 15 genes mais diferencialmente expressos foram analisados em ambas as figuras 17 e 18. Em ambos os casos a maior parte dos genes não são anotados, putativos ou com ORF duvidosa, integrando 73.33% dos genes regulados positivamente e 80% dos negativamente (Apêndices F e G). Dos genes regulados positivamente, YCR013C apresentou uma variação persistente da expressão em todas as linhagens celulares pGSHs e YJR037W nas linhagens S2U, S4U e S5U, ambas ORFs duvidosas e improváveis que codifique uma proteína funcional. Entre os genes anotados (Tabela 12), YEL020W-A (TIM9) apresentou uma variação nas cepas S2U, S5U e S6U, e YJL153C (INO1) nas S3U, S4U e S6U. TIM 9 é uma proteína intermembranar mitocondrial e INO1 codifica enzima envolvido na síntese de fosfolipídios.

Nos genes regulados negativamente, não há uma variação persistente da expressão como observado anteriormente para os genes regulados positivamente. Aqui é destacado o gene YGL118C (Proteína putativa de função desconhecida) que apresenta variação nas cepas S2U e S6U.

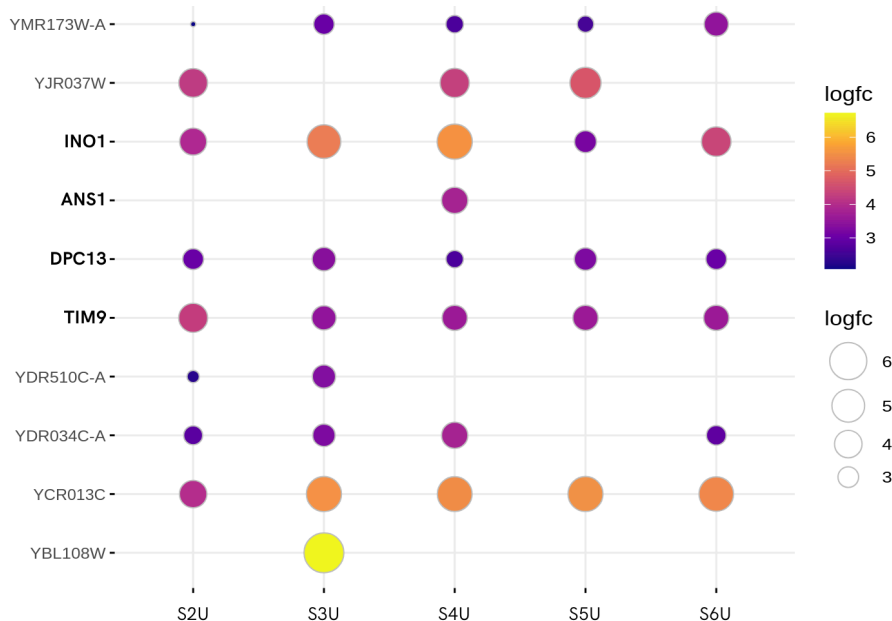
Tabela 12. Genes caracterizados mais diferencialmente expressos em todos os experimentos.

Gene	logfc	Linhagem	Anotação (SGD)
Regulados Positivamente			
YJL153C (INO1)	5.54	S4U	Enzima envolvido na síntese de fosfatos de inositol e fosfolipídios contendo inositol; transcrição coregulada com outros genes biossintéticos de fosfolipídios.
YEL020W-A (TIM9)	4.26	S2U	Proteína essencial do espaço intermembranar mitocondrial; forma um complexo com Tim10p para fornecimento de proteínas hidrofóbicas à membrana interna.
YHR126C (ANS1)	3.80	S4U	Proteína putativa.
YFL010W-A (AUA1)	3.20	S6U	Proteína necessária para a regulação negativa por amônia de Gap1p; Gap1p é uma permease geral de aminoácidos.
YKL164C (PIR1)	3.12	S6U	Proteína glicosilada necessária para a estabilidade da parede celular; expressão regulada pela via de integridade celular e por Swi5p durante o ciclo celular.
Regulados Negativamente			
YLR461W (PAU4)	-7.53	S3U	Membro da família da seripauperina; codificado principalmente em regiões subteloméricas; ativo durante a fermentação alcoólica; regulado por anaerobiose; regulado negativamente por oxigênio; reprimido por heme.
YNL318C (HXT14)	-7.20	S2U	Proteína semelhante aos transportadores de hexose; expressão induzida em baixa glicose e reprimida em alta glicose; detectada em mitocôndrias.
YDR218C (SPR28)	-5.95	S5U	Homólogo específico da esporulação da família de genes CDC3/10-12; septina meiótica expressa em altos níveis durante as divisões meióticas e a formação de ascósporos.

A inspeção do comportamento da expressão dos principais genes expressos diferencialmente nas construções pGSH em termos de mudança de *log fold* foi realizada para embasar que a alteração observada até aqui não está relacionada à inserção nos pGSHs. A visão geral qualitativa da alteração de log fold dos principais genes regulados positivamente e negativamente nas construções pGSH é mostrada na Figura 18. Para os genes positivamente

expressos, é possível observar uma distribuição generalizada entre todos os pGSHs com exceções de YHR126C e YBL108W que foram detectados apenas uma vez no S4 e S3 respectivamente (Figura 19a) sendo ambos não caracterizados. Os genes negativamente expressos curiosamente apresentam uma prevalência nas linhagens S2U e S3U mas sem uma correlação clara (Figura 19b, Apêndice F e G).

a. Genes Regulados Positivamente



b. Genes Regulados Negativamente

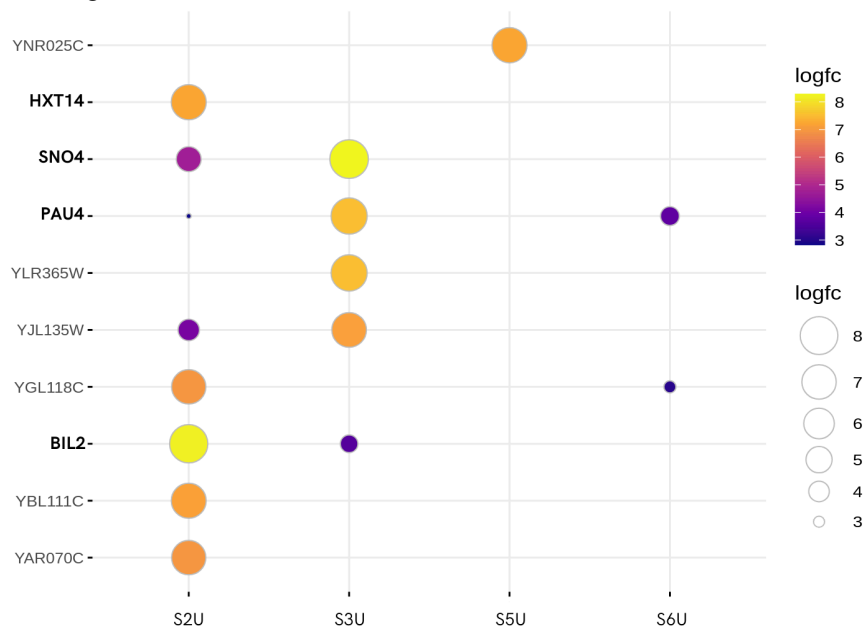


Figura 19 - Balloon plot da alteração de log2FC dos 10 genes diferencialmente regulados, em termos do valor de *log fold change*, nas construções pGSH. a) Genes regulados positivamente. b) Genes regulados negativamente. A linhagem S4U não teve enriquecimento significativo nos genes regulados negativamente. A marca auxotrófica utilizada URA3 não é mostrada no gráfico.

5. DISCUSSÃO

5.1. Pipelines semelhantes descritos na literatura

Dois pipelines de bioinformática estão disponíveis para identificar pGSHs em células humanas. O primeiro pipeline de predição de GSHs do grupo de Aznauryan *et al.* (2022) que apesar de terem sido identificados quase 2.000 pGSHs, algumas características da abordagem de identificação do algoritmo levantam preocupações: (1) não determinar quais são os genes flanqueadores (2) não determinar a orientação dos genes flanqueadores e (3) a faixa de comprimento da região varia de maneira discrepante de 23 pb a 712 kb. Um tamanho pequeno de pGSH, como 23 pb, pode impedir inserções poligênicas, e o tamanho máximo descrito pode conter sequências regulatórias não anotadas. Corroborando essas preocupações, dos 2.000 locais identificados, os autores apenas validaram funcionalmente cinco candidatos à GSHs sendo que somente dois cumpriram minimamente os critérios (Rogi 1 e Rogi 2), um de 40,6 kb e outro de 48 kb de tamanho. A análise de RNA-Seq foi feita apenas para o Rogi 2 demonstrando perturbação genômica semelhante à que descrevemos, provavelmente causada pelo processo de transfecção das células.

O segundo é o *software* de computador GEG-SH mapper que tem uma abordagem diferente para identificar pGSHs específicos de tecidos (Shrestha *et al.*, 2022). O trabalho propõe a integração do cassete de expressão em inserções de elementos móveis polimórficos (pMEIs) e o uso de informações sobre a organização da cromatina em 3D, levando a quase todos os pGSHs em regiões intrônicas. Embora os autores tenham levantado a hipótese de que os pMEIs não levam à instabilidade genômica, as inserções poligênicas em regiões intrônicas têm o potencial de afetar a expressão gênica levando a fenótipo indesejado, assim essa estratégia apresenta uma relevante limitação.

Considerando ainda ambos, outras importantes desvantagens é que esses programas são restritos ao genoma humano, e que esses algoritmos exigem conhecimento de linha de comando para baixar pacotes de pré-requisitos externos e executar os programas, o que restringe sua ampla utilização. Além disso, como essas ferramentas e o SHIP têm abordagens diferentes para identificar pGSHs, elas podem funcionar como ferramentas complementares, dependendo das necessidades do usuário.

5.2. *Software* SHIP e dados da validação experimental dos pGSHs

O presente estudo demonstrou que o *software* SHIP foi capaz de identificar seis regiões com características promissoras para se tornarem GSHs e cinco (pGSH-2 a 6) foram validadas *in vivo* em *S. cerevisiae*. Notavelmente, a orientação convergente dos genes vizinhos é fortemente recomendada para reduzir possíveis interferências nas regiões promotoras e regulatórias a montante, uma vez que a maioria dos fatores reguladores são agentes a montante do gene regulado (Charoensawan; Wilson; Teichmann, 2010; Latchman, 2013; Weirauch; Hughes, 2011).

A reprovação manual do pGSH-1, devido à sua proximidade com uma região telomérica, exemplifica que o SHIP é uma ferramenta de uso geral que fornece uma investigação inicial, reduzindo o número de pGSHs que exigem avaliação. O algoritmo não toma decisões nem elimina pGSHs sem critérios definidos, uma vez que a "proximidade" é um parâmetro subjetivo. Além disso, todos os outros GSHs atenderam aos critérios considerados para uma plataforma de aterrissagem segura para a integração do transgene: (1) acessível à integração do transgene, (2) genomicamente estável, (3) permitindo a expressão previsível e constante do BioBrick e (4) interferindo minimamente ou não interferindo na expressão dos genes genômicos e flanqueadores.

Os dados de PCR, sequenciamento, citometria de fluxo e RT-qPCR mostraram que os BioBrick foram inseridos nos locais exatos identificados pelo SHIP sem causar mutações nas sequências utilizadas como braços de homologia e que as linhagens celulares pGSHs foram capazes de acumular transcrito e proteína fluorescente verde codificada pelo BioBrick ymUKG1, indicando que essas regiões são acessíveis para inserção e expressão do transgene. Adicionalmente, os 5 GSHs demonstraram ser regiões genomicamente estáveis durante 10 dias de cultivo contínuo, acomodando um fenótipo consistente dos BioBricks sem interferir na viabilidade celular.

Considerando a taxa de crescimento maiores observadas entre as linhagens pGSHs comparada ao controle e também a análise de RNA-Seq que demonstrou que mesmo com uma definição rigorosa de critérios e uma anotação consistente do genoma, podem surgir alterações genômicas desencadeadas pelo próprio processo de transformação celular. Os resultados mostraram que há uma expressão gênica diferencial de 16 % do tratamento em relação ao controle (Figura 13a). Considerando que são regiões distintas de inserção em diferentes cromossomos, é muito provável que essa alteração seja uma consequência da

introdução do gene URA3 (Alam *et al.*, 2016; Pronk, 2002; Yan *et al.*, 2023). Ainda nos dados de RNA-Seq, quando a expressão diferencial dos genes vizinhos foi analisada, se nota uma variação não correlacional com a localização de cada GSH nas expressões gênicas.

Em uma primeira avaliação, três genes se destacam em relação à expressão diferencial, PUG1, COS9 e YHL041W. Vizinho 5' do GSH-2, o gene PUG1 apresentou expressão diferencial (RNA-Seq) para todas as regiões com exceção do GSH-5 (Figura 13). Além de apenas não apresentar alteração significativa na expressão relativa (qPCR) na linhagem GSH-4. O gene PUG1 codifica uma proteína de membrana (Pug1) envolvida na captação de protoporfirina IX (PPIX) e no efluxo de heme. Vale destacar que (1) a *S. cerevisiae* não utiliza heme exógeno de maneira eficiente para suprir às necessidades celulares de ferro (Protchenko *et al.*, 2006; Weissman; Shemer; Kornitzer, 2002), (2) Pug1 não é a principal responsável pela absorção de heme ou PPIX, indicando que outros genes codificam a maioria destas atividades (Protchenko *et al.*, 2008) e (3) por ser integrante de uma família de proteínas envolvidas na resistência à xenobióticos e endobióticos, Pug1 pode exercer outras funções relacionadas ao estresse celular (Protchenko *et al.*, 2008).

O gene COS9 é vizinho à 5' do GSH-5 e apresentou expressão diferencial em todas as cepas com exceção do GSH-3 além da alteração significativa na expressão relativa (qPCR) nos GSH-3, 5 e 6. Codifica uma proteína, inferido por homologia, que está envolvida na degradação das proteínas da membrana plasmática para degradação (MacDonald *et al.*, 2015). Além disso, a subfamília DUP380 dos genes COS apresentam elevação da expressão gênica em condições de limitação nutricional (Despons *et al.*, 2006; MacDonald *et al.*, 2015). YHL041W é o gene vizinho à 5' do GSH-3 com expressão diferencial e alteração na expressão relativa em todos os GSHs. Codifica uma proteína putativa de função desconhecida. Assim como esses três genes que foram descritos, a expressão diferencial dos demais também não apresenta um padrão que seja correlacionado com os lócos de inserção, reforçando a hipótese que esta alteração é oriunda do processo de transformação.

5.3. pGSHs identificados pelo *software* SHIP já utilizados na literatura

Interessantemente, dos cinco pGSHs identificadas pelo *software* SHIP, três já foram utilizadas anteriormente como regiões para inserção de transgenes em *S. cerevisiae*, o que indica uma consistência nas predições, além de indicar novos candidatos para estudos futuros. Os trabalhos que identificaram tais regiões buscavam por sequências intergênicas para a inserção dos genes de interesse mas não realizaram estudos para validar o funcionamento ou o impacto da inserção. O pGSH-5 foi utilizado, de forma independente em 2014, para o desenvolvimento de um sistema de integração de múltiplos genes baseados na reciclagem de marcadores com recombinase. Esse GSH, nomeado pelos autores do estudo como inter-ORF I7S, comportou os cassetes de expressão da xilosidase de *Aspergillus oryzae* (XYLA) e da β -glicosidase de *Aspergillus aculeatus* (BGL1) como partes da montagem da via metabólica de xilose na cepa diploide industrial de *S. cerevisiae* sun-49 (Hasunuma *et al.*, 2014).

Em 2016, o GSH-6 foi usado para inserção de 4 genes exógenos na cepa BY4741. A integração resultou na construção de uma linhagem aderente à celulose para produção de biocombustível celulósico, indicando expressão gênica dos insertos (Liu *et al.*, 2016). No trabalho de Mu *et al.* (2020), foi citado que a marca de resistência à nourseotricina (do inglês, *nourseothricin drug-resistant marker*, *natMX4*) foi inserida na coordenada correspondente ao pGSH-2 na linhagem selvagem SK1. Esse estudo avaliou o mecanismo de regulação das quebras de fita dupla de DNA que iniciam a recombinação meiótica. Esses diferentes empregos dos GSHs corroboram os resultados gerados pelo SHIP de que esses são locais de inserção de genes heterólogos que permitem a expressão dos mesmos, um critério fundamental para ser um GSH. Também reforçam a necessidade de programas computacionais para a automação da identificação de tais regiões, já que a validação funcional do SHIP, realizada nesse estudo demonstrou que esses locais também obedecem aos demais critérios para ser GSH.

5.4. Sumário crítico do contexto dos genes diferencialmente expressos do RNA-Seq

Os resultados deste estudo corroboram estudos prévios que a inserção de um transgene causa uma perturbação na programação transcricional da célula. O que observamos é interessante: pelo diagrama de Venn do RNA-Seq, independente do local do pGSH temos um grupo de genes que são DE, ou seja, a própria transformação evoca uma reprogramação da expressão gênica. Destes genes consistentemente DE não existe nenhum enriquecimento funcional que possa indicar uma resposta celular coordenada, como resposta a estresses oxidativos ou reprogramação da divisão celular, dentre outros.

Em relação aos genes diferencialmente expressos compartilhados, aproximadamente 33% dos genes são regulados positivamente e em torno de ~25% genes são regulados negativamente. Sendo que entre os genes diferencialmente expressos, a ampla maioria constitui o grupo de genes compartilhados entre todas as linhagens GSHs. Ou seja, dentre todas as combinações de genes/experimentos, o mais frequente consiste no grupo de genes compartilhados entre todos os experimentos (interseção central nos diagramas de Venn).

Também é importante ressaltar, que apesar dos números de DE genes parecerem altos para o conceito de GSH, quando analisamos as figuras 17 e 18, temos em média 30 a 40 de genes que tem um log FC maior que 2 (4X maior expressão), e a grande maioria tem um aumento mais moderado na expressão, o que apesar de ser estatisticamente significativo, não tem um efeito biológico tão pronunciado. Considerando que o perfil de expressão de pGSHs foi muito semelhante entre as linhagens celulares de GSH e que os BioBricks estão em diferentes locos, o metabolismo de uracila com a recuperação de URA3 pode potencialmente ser a causa dessa diferenciação na expressão genômica (Alam *et al.*, 2016; Pronk, 2002; Yan *et al.*, 2023). Entretanto, isso não exclui possíveis efeitos de estresse a longo prazo, principalmente quanto à alterações epigenéticas que podem atuar na modulação da expressão genômica (Avramova, 2015; Choi; Kim, 2008; Yao *et al.*, 2022; Zhang *et al.*, 2013).

No final, a baixa variação total de apenas 16%, a maioria dos genes DE serem compartilhados entre todas as linhagens e sem enriquecimento ontológico que indique uma resposta celular à inserção dos Biobricks nos GSHs indicam que essas regiões identificadas pelo *software* SHIP constituem alvos que minimizam a perturbação genômica.

6. CONCLUSÕES E PERSPECTIVAS

O algoritmo SHIP é um dos poucos programas desenhados para predição de GSHs e que foi concebido de forma a ser facilmente executável para qualquer organismo eucariótico com genoma anotado. Até o momento apenas dois pipelines estão disponíveis para predição de GSHs (Aznauryan *et al.*, 2022; Shrestha *et al.*, 2022), ambos limitados à *H. sapiens*. SHIP é o primeiro *software* para identificação de pGSHs com amplo espectro de organismos alvos, capaz de retornar uma visão geral da densidade genômica analisada e descrição detalhada de cada região auxiliando na escolha do usuário, além de não ser necessário conhecimento de linha de comando.

Dados de sequenciamento, citometria, taxa de crescimento e estabilidade revelaram regiões promissoras contendo os BioBricks nos pGSHs identificados pelo SHIP, com alto número de células acumulando a proteína fluorescente repórter verde, taxa de crescimento semelhante à cepa não transformada e estabilidade desses BioBricks após dez dias de cultivo. Além disso, análises de RNA-Seq e RT-qPCR mostraram expressão genômica similar entre as linhagens que contêm os BioBricks inseridos em cada um dos pGSHs, indicando que os locais potencialmente seguros identificados pelo SHIP cumprem todos os critérios para serem chamados de GSHs.

No futuro espera-se que essa ferramenta seja utilizada não só para os três organismos-modelo escolhidos para avaliação dessa proposta, mas também para qualquer outro organismo eucarioto cuja anotação genômica curada esteja disponível. Esse é um trabalho inovador já que, até o momento, os GSHs utilizados em diferentes organismos eucariotos foram identificados empiricamente sendo que, apenas em humanos e *C. neoformans*, um dos GSHs identificados em região extragenômica foi avaliado por ensaios *in vivo*.

O *software* SHIP permitirá o desenho racional de inserção dos transgenes em locais que possibilitem a expressão dos mesmos no nível desejado pelo pesquisador, reduzindo, dessa forma, tempo e custo para obtenção do organismo de interesse. Atualmente, os processos seletivos de organismos transformados demandam tamanhos amostrais cada vez mais elevados a depender do número de transgenes a serem inseridos e também do organismo alvo, pois, com os métodos tradicionais, os genes de interesse são inseridos ao acaso no genoma ou seus locais de inserção são selecionados, mas estes não são GSHs e, portanto, não cumprem os critérios de expressão do gene exógeno com mínima ou nenhuma perturbação da expressão genômica.

Assim, não há conhecimento de outro *software* desenvolvido com essa finalidade e flexibilidade de organismos alvos, o que possibilitou que o SHIP fosse registrado pelo Instituto Nacional da Propriedade Industrial (INPI) (Apêndice H). O intuito é que este *software* seja utilizado por pesquisadores e empresas da área da biotecnologia/biologia sintética que visem a construção de circuitos genéticos complexos, ou que visem o desenvolvimento de organismos contendo vias metabólicas completas.

REFERÊNCIAS

- ALAM, Mohammad Tauqeer *et al.* The metabolic background is a global player in *Saccharomyces* gene expression epistasis. **Nature Microbiology**, [s. l.], v. 1, n. 3, p. 1–10, 2016.
- ANDERS, Simon; PYL, Paul Theodor; HUBER, Wolfgang. HTSeq—a Python framework to work with high-throughput sequencing data. **Bioinformatics**, [s. l.], v. 31, n. 2, p. 166–169, 2015.
- ANNALURU, Narayana *et al.* Total synthesis of a functional designer eukaryotic chromosome. **Science**, [s. l.], v. 344, n. 6179, p. 55–58, 2014.
- ARRAS, Samantha D.M. *et al.* A genomic safe haven for mutant complementation in *Cryptococcus neoformans*. **PLoS ONE**, [s. l.], v. 10, n. 4, p. 1–16, 2015.
- AVRAMOVA, Zoya. Transcriptional ‘memory’ of a stress: transient chromatin and memory (epigenetic) marks at stress-response genes. **The Plant Journal**, [s. l.], v. 83, n. 1, p. 149–159, 2015.
- AZNAURYAN, Erik *et al.* Discovery and validation of human genomic safe harbor sites for gene and cell therapies. **Cell Reports Methods**, [s. l.], v. 2, n. 1, 2022.
- BAKER BRACHMANN, Carrie *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**, [s. l.], v. 14, n. 2, p. 115–132, 1998.
- BECKER, Daniel M.; GUARENTE, Leonard. [12] High-efficiency transformation of yeast by electroporation. *In: METHODS IN ENZYMOLOGY*. [S. l.]: Academic Press, 1991. (Guide to Yeast Genetics and Molecular Biology). v. 194, p. 182–187. Disponível em: Acesso em: 13 dez. 2023.
- BOEKE, Jef D. *et al.* Ty elements transpose through an RNA intermediate. **Cell**, [s. l.], v. 40, n. 3, p. 491–500, 1985.
- BOONEKAMP, Francine J. *et al.* Full humanization of the glycolytic pathway in *Saccharomyces cerevisiae*. **Cell Reports**, [s. l.], v. 39, n. 13, 2022. Disponível em: [https://www.cell.com/cell-reports/abstract/S2211-1247\(22\)00799-9](https://www.cell.com/cell-reports/abstract/S2211-1247(22)00799-9). Acesso em: 27 fev. 2024.
- CAMERON, D. Ewen; BASHOR, Caleb J.; COLLINS, James J. A brief history of synthetic biology. **Nature Reviews Microbiology**, [s. l.], v. 12, n. 5, p. 381–390, 2014.
- CAVALIERI, Duccio *et al.* Evidence for *S. cerevisiae* Fermentation in Ancient Wine. *In: JOURNAL OF MOLECULAR EVOLUTION*, 2003. **Anais [...]**. [S. l.: s. n.], 2003.
- CHEN, Shifu *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. **Bioinformatics**, [s. l.], v. 34, n. 17, p. i884–i890, 2018.
- CHOI, Jung Kyoon; KIM, Young-Joon. Epigenetic regulation and the variability of gene expression. **Nature Genetics**, [s. l.], v. 40, n. 2, p. 141–147, 2008.

CLARK, James F.; DINSMORE, Colin J.; SORIANO, Philippe. A most formidable arsenal: genetic technologies for building a better mouse. **Genes & Development**, [s. l.], v. 34, n. 19–20, p. 1256–1286, 2020.

COELHO, Miguel C.; PINTO, Ricardo M.; MURRAY, Andrew W. Heterozygous mutations cause genetic instability in a yeast model of cancer evolution. **Nature**, [s. l.], v. 566, n. 7743, p. 275–278, 2019.

CONG, Le *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. **Science**, [s. l.], v. 339, n. 6121, p. 819–823, 2013.

DA SILVA, Nancy A.; SRIKRISHNAN, Sneha. Introduction and expression of genes for metabolic engineering applications in *Saccharomyces cerevisiae*. **FEMS Yeast Research**, [s. l.], v. 12, n. 2, p. 197–214, 2012.

DAI, Zhubo *et al.* Metabolic engineering of *Saccharomyces cerevisiae* for production of ginsenosides. **Metabolic Engineering**, [s. l.], v. 20, p. 146–156, 2013.

DESPONS, Laurence *et al.* An evolutionary scenario for one of the largest yeast gene families. **Trends in Genetics**, [s. l.], v. 22, n. 1, p. 10–15, 2006.

DOBIN, Alexander *et al.* STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, [s. l.], v. 29, n. 1, p. 15–21, 2013.

ELOWITZ, Michael B.; LEIBLER, Stanislas. A synthetic oscillatory network of transcriptional regulators. **Nature**, [s. l.], v. 403, n. 6767, p. 335–338, 2000.

FURUKAWA, Takanori *et al.* Exploring a novel genomic safe-haven site in the human pathogenic mould *Aspergillus fumigatus*. **Fungal Genetics and Biology**, [s. l.], v. 161, 2022.

GALANIE, Stephanie *et al.* Complete biosynthesis of opioids in yeast. **Science**, [s. l.], v. 349, n. 6252, p. 1095–1100, 2015.

GALDZICKI, Michal *et al.* Standard Biological Parts Knowledgebase. **PLOS ONE**, [s. l.], v. 6, n. 2, p. e17005, 2011.

GALDZICKI, Michal *et al.* The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. **Nature Biotechnology**, [s. l.], v. 32, n. 6, p. 545–550, 2014.

GAO, Jucan *et al.* Synthetic Biology Toolkit for Marker-Less Integration of Multigene Pathways into *Pichia pastoris* via CRISPR/Cas9. **ACS Synthetic Biology**, [s. l.], v. 11, n. 2, p. 623–633, 2022.

GARDNER, Timothy S.; CANTOR, Charles R.; COLLINS, James J. Construction of a genetic toggle switch in *Escherichia coli*. **Nature**, [s. l.], v. 403, n. 6767, p. 339–342, 2000.

GARGE, Riddhiman K *et al.* Systematic Humanization of the Yeast Cytoskeleton Discerns Functionally Replaceable from Divergent Human Genes. **Genetics**, [s. l.], v. 215, n. 4, p. 1153–1169, 2020.

GIETZ, Daniel R *et al.* **Studies on the Transformation of Intact Yeast Cells by the**

LiAc/SS-DNA/PEG Procedure. [S. l.: s. n.], 1995.

GOFFEAU, A. *et al.* Life with 6000 Genes. **Science**, [s. l.], v. 274, n. 5287, p. 546–567, 1996.

G:PROFILER – A WEB SERVER FOR FUNCTIONAL ENRICHMENT ANALYSIS AND CONVERSIONS OF GENE LISTS. [S. l.], [s. d.]. Disponível em: <https://biit.cs.ut.ee/gprofiler/gost>. Acesso em: 14 jan. 2024.

GUIRIMAND, Gregory *et al.* Innovative Tools and Strategies for Optimizing Yeast Cell Factories. **Trends in Biotechnology**, [s. l.], v. 39, n. 5, p. 488–504, 2021.

HAMILTON, Stephen R. *et al.* Humanization of Yeast to Produce Complex Terminally Sialylated Glycoproteins. **Science**, [s. l.], v. 313, n. 5792, p. 1441–1443, 2006.

HAMZA, Akil *et al.* Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. **Genetics**, [s. l.], v. 201, n. 3, p. 1263–1274, 2015.

HAMZA, Akil *et al.* Cross-Species Complementation of Nonessential Yeast Genes Establishes Platforms for Testing Inhibitors of Human Proteins. **Genetics**, [s. l.], v. 214, n. 3, p. 735–747, 2020.

HASELOFF, Jim; AJIOKA, Jim. Synthetic biology: history, challenges and prospects. **Journal of The Royal Society Interface**, [s. l.], v. 6, n. suppl_4, p. S389–S391, 2009.

HASUNUMA, Tomohisa *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. **Microbial Cell Factories**, [s. l.], v. 13, n. 1, 2014.

HENCKAERTS, Els; LINDEN, R Michael. Adeno-associated virus: a key to the human genome?. **Future Virology**, [s. l.], v. 5, n. 5, p. 555–574, 2010.

ITTIPRASERT, Wannaporn *et al.* Targeted insertion and reporter transgene activity at a gene safe harbor of the human blood fluke, *Schistosoma mansoni*. **Cell Reports Methods**, [s. l.], v. 3, n. 7, p. 100535–100535, 2023.

JACOB, François; MONOD, Jacques. On the Regulation of Gene Activity. **Cold Spring Harbor Symposia on Quantitative Biology**, [s. l.], v. 26, p. 193–211, 1961.

JINEK, Martin *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. **Science**, [s. l.], v. 337, n. 6096, p. 816–821, 2012.

KACHROO, Aashiq H. *et al.* Systematic humanization of yeast genes reveals conserved functions and genetic modularity. **Science**, [s. l.], v. 348, n. 6237, p. 921–925, 2015.

KAISHIMA, Misato *et al.* Expression of varied GFPs in *Saccharomyces cerevisiae*: Codon optimization yields stronger than expected expression and fluorescence intensity. **Scientific Reports**, [s. l.], v. 6, 2016.

KATOH, Kazutaka; STANDLEY, Daron M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**,

[*s. l.*], v. 30, n. 4, p. 772–780, 2013.

KNIGHT, Thomas. Idempotent Vector Design for Standard Assembly of Biobricks. [*s. l.*], 2003. Disponível em: <https://dspace.mit.edu/handle/1721.1/21168>. Acesso em: 10 jan. 2024.

KOLBERG, Liis *et al.* g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). **Nucleic Acids Research**, [*s. l.*], v. 51, n. W1, p. W207–W212, 2023.

KOLDE, Raivo. Pheatmap: pretty heatmaps. **R package version**, [*s. l.*], v. 1, n. 2, p. 726, 2012.

KONG, Sijia *et al.* Expanding the neutral sites for integrated gene expression in *Saccharomyces cerevisiae*. **FEMS Microbiology Letters**, [*s. l.*], v. 369, n. 1, p. fnac081, 2022.

KOTIN, Robert M. *et al.* Mapping and direct visualization of a region-specific viral DNA integration site on chromosome 19q13-qter. **Genomics**, [*s. l.*], v. 10, n. 3, p. 831–834, 1991.

LAURENT, Jon M. *et al.* Humanization of yeast genes with multiple human orthologs reveals functional divergence between paralogs. **PLOS Biology**, [*s. l.*], v. 18, n. 5, p. e3000627, 2020.

LEE, Clare S. K. *et al.* Humanizing the yeast origin recognition complex. **Nature Communications**, [*s. l.*], v. 12, n. 1, p. 33, 2021.

LEE, F. W. F.; SILVA, N. A. Da. Improved efficiency and stability of multiple cloned gene insertions at the δ sequences of *Saccharomyces cerevisiae*. **Applied Microbiology and Biotechnology**, [*s. l.*], v. 48, n. 3, p. 339–345, 1997.

LIU, Tengfei *et al.* Construction of ajmalicine and sanguinarine de novo biosynthetic pathways using stable integration sites in yeast. **Biotechnology and Bioengineering**, [*s. l.*], v. 119, n. 5, p. 1314–1326, 2022.

LIU, Qi *et al.* CRISPR–Cas9-mediated genomic multiloci integration in *Pichia pastoris*. **Microbial Cell Factories**, [*s. l.*], v. 18, n. 1, p. 144, 2019.

LIU, Zhuo *et al.* Engineering of a novel cellulose-adherent cellulolytic *Saccharomyces cerevisiae* for cellulosic biofuel production. **Scientific Reports**, [*s. l.*], v. 6, 2016.

LIU, Rong *et al.* Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. **Cell**, [*s. l.*], v. 86, n. 3, p. 367–377, 1996.

LOPES, Teresa S. *et al.* High-copy-number integration into the ribosomal DNA of *Saccharomyces cerevisiae*: a new vector for high-level expression. **Gene**, [*s. l.*], v. 79, n. 2, p. 199–206, 1989.

LOVE, Michael I.; HUBER, Wolfgang; ANDERS, Simon. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, [*s. l.*], v. 15, n. 12, p. 550, 2014.

MACDONALD, Chris *et al.* A Family of Tetraspans Organizes Cargo for Sorting into Multivesicular Bodies. **Developmental Cell**, [*s. l.*], v. 33, n. 3, p. 328–342, 2015.

- MALCI, Koray *et al.* Standardization of Synthetic Biology Tools and Assembly Methods for *Saccharomyces cerevisiae* and Emerging Yeast Species. **ACS Synthetic Biology**, [s. l.], v. 11, n. 8, p. 2527–2547, 2022.
- MARTIN, Fergal J. *et al.* Ensembl 2023. **Nucleic Acids Research**, [s. l.], v. 51, n. 1 D, p. D933–D941, 2023.
- METSALU, Tauno; VILO, Jaak. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. **Nucleic Acids Research**, [s. l.], v. 43, n. W1, p. W566–W570, 2015.
- MIKKELSEN, Michael Dalgaard *et al.* Microbial production of indolylglucosinolate through engineering of a multi-gene pathway in a versatile yeast expression platform. **Metabolic Engineering**, [s. l.], v. 14, n. 2, p. 104–111, 2012.
- MONACO, Gianni *et al.* flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. **Bioinformatics**, [s. l.], v. 32, n. 16, p. 2473–2480, 2016.
- MOURIDI, Sonia El; ALKHALDI, Faisal; FROKJAR-JENSEN, Christian. Modular safe-harbor transgene insertion for targeted single-copy and extrachromosomal array integration in *Caenorhabditis elegans*. **G3: Genes, Genomes, Genetics**, [s. l.], v. 12, n. 9, 2022.
- MU, Xiaojing *et al.* Chromosome-autonomous feedback down-regulates meiotic DNA break competence upon synaptonemal complex formation. **Genes and Development**, [s. l.], v. 34, n. 23–24, p. 1605–1618, 2020.
- NASERI, Gita; KOFFAS, Mattheos A. G. Application of combinatorial optimization strategies in synthetic biology. **Nature Communications**, [s. l.], v. 11, n. 1, p. 2446, 2020.
- NASSAR, Luis R. *et al.* The UCSC Genome Browser database: 2023 update. **Nucleic Acids Research**, [s. l.], v. 51, n. 1 D, p. D1188–D1195, 2023.
- NIELSEN, Jens. Yeast Systems Biology: Model Organism and Cell Factory. **Biotechnology Journal**, [s. l.], v. 14, n. 9, 2019.
- NISHI, Teruyuki *et al.* One-Step In Vivo Assembly of Multiple DNA Fragments and Genomic Integration in *Komagataella phaffii*. **ACS Synthetic Biology**, [s. l.], v. 11, n. 2, p. 644–654, 2022.
- OLESEN, Kjeld *et al.* Yeast Functional Analysis Report The pYC plasmids, a series of cassette-based yeast plasmid vectors providing means of counter-selection. **Yeast**, [s. l.], v. 16, n. 11, p. 1035–1043, 2000.
- PADDON, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. **Nature**, [s. l.], v. 496, n. 7446, p. 528–532, 2013.
- PAVANI, Giulia. Targeted Gene Delivery: Where to Land. **Frontiers in Genome Editing**, [s. l.], v. 2, 2020.
- PELLENZ, Stefan *et al.* New Human Chromosomal Sites with “Safe Harbor” Potential for Targeted Transgene Insertion. **Human Gene Therapy**, [s. l.], v. 30, n. 7, p. 814–828, 2019.

PFAFFL, Michael W.; HORGAN, Graham W.; DEMPFLÉ, Leo. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. **Nucleic Acids Research**, [s. l.], v. 30, n. 9, p. e36, 2002.

PHAM, Tuyetnhu; XIE, Xiaofeng; LIN, Xiaorong. An intergenic “safe haven” region in *Aspergillus fumigatus*. **Medical Mycology**, [s. l.], v. 58, n. 8, p. 1178–1186, 2020.

PRONK, Jack T. Auxotrophic Yeast Strains in Fundamental and Applied Research. **Applied and Environmental Microbiology**, [s. l.], v. 68, n. 5, p. 2095–2100, 2002.

PROTCHENKO, Olga *et al.* A screen for genes of heme uptake identifies the FLC family required for import of FAD into the endoplasmic reticulum. **Journal of Biological Chemistry**, [s. l.], v. 281, n. 30, p. 21445–21457, 2006.

PROTCHENKO, Olga *et al.* Role of PUG1 in inducible porphyrin and heme transport in *Saccharomyces cerevisiae*. **Eukaryotic Cell**, [s. l.], v. 7, n. 5, p. 859–871, 2008.

R: THE R PROJECT FOR STATISTICAL COMPUTING. [S. l.], [s. d.]. Disponível em: <https://www.r-project.org/>. Acesso em: 15 jan. 2024.

RICHARDSON, Sarah M *et al.* **Design of a synthetic yeast genome**: Science. [S. l.: s. n.], 2017. Disponível em: <https://www.science.org>. .

ROBINSON, Mark D.; MCCARTHY, Davis J.; SMYTH, Gordon K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, [s. l.], v. 26, n. 1, p. 139–140, 2010.

SADELAIN, Michel; PAPAPETROU, Eirini P.; BUSHMAN, Frederic D. Safe harbours for the integration of new DNA in the human genome. **Nature Reviews Cancer**, [s. l.], v. 12, n. 1, p. 51–58, 2012.

SAKAI, Akira; SHIMIZU, Yuki; HISHINUMA, Fumio. Integration of heterologous genes into the chromosome of *Saccharomyces cerevisiae* using a delta sequence of yeast retrotransposon Ty. **Applied Microbiology and Biotechnology**, [s. l.], v. 33, n. 3, p. 302–306, 1990.

SAMULSKI, R. J. *et al.* Targeted integration of adeno-associated virus (AAV) into human chromosome 19. **EMBO Journal**, [s. l.], v. 10, n. 12, p. 3941–3950, 1991.

SAYERS, Eric W *et al.* Database resources of the National Center for Biotechnology Information in 2023. **Nucleic Acids Research**, [s. l.], v. 51, n. D1, p. D29–D38, 2023.

SHETTY, Reshma P.; ENDY, Drew; KNIGHT, Thomas F. Engineering BioBrick vectors from BioBrick parts. **Journal of Biological Engineering**, [s. l.], v. 2, n. 1, p. 5, 2008.

SHRESTHA, Dewan *et al.* Genomics and epigenetics guided identification of tissue-specific genomic safe harbors. **Genome Biology**, [s. l.], v. 23, n. 1, 2022.

STEINMETZ, Lars M. *et al.* Systematic screen for human disease genes in yeast. **Nature Genetics**, [s. l.], v. 31, n. 4, p. 400–404, 2002.

STEPHANOPOULOS, Gregory. Synthetic biology and metabolic engineering. **ACS**

synthetic biology, [s. l.], v. 1, n. 11, p. 514–525, 2012.

SÜDFELD, Christian *et al.* The nucleolus as a genomic safe harbor for strong gene expression in *Nannochloropsis oceanica*. **Molecular Plant**, [s. l.], v. 15, n. 2, p. 340–353, 2022.

SUN, Song *et al.* An extended set of yeast-based functional assays accurately identifies human disease mutations. **Genome Research**, [s. l.], v. 26, n. 5, p. 670–680, 2016.

SUN, Chao *et al.* Precise integration of large DNA sequences in plant genomes using PrimeRoot editors. **Nature Biotechnology**, [s. l.], 2023.

TESTE, Marie Ange *et al.* Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. **BMC Molecular Biology**, [s. l.], v. 10, p. 99–99, 2009.

TSUTSUI, Hidekazu *et al.* Improving membrane voltage measurements using FRET with new fluorescent proteins. **Nature Methods**, [s. l.], v. 5, n. 8, p. 683–685, 2008.

VANDESOMPELE, Jo *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. **Genome Biology**, [s. l.], v. 3, n. 7, p. research0034.1, 2002.

VOTH, Warren P. *et al.* Yeast vectors for integration at the HO locus. **Nucleic Acids Research**, [s. l.], v. 29, n. 12, p. e59, 2001.

WEISSMAN, Ziva; SHEMER, Revital; KORNITZER, Daniel. Deletion of the copper transporter CaCCC2 reveals two distinct pathways for iron acquisition in *Candida albicans*. **Molecular Microbiology**, [s. l.], v. 44, n. 6, p. 1551–1560, 2002.

WEITZ, Eric. **Ideogram**. [S. l.: s. n.], 2023 [2015]. Disponível em: <https://github.com/eweitz/ideogram>. Acesso em: 19 dez. 2023.

WICKHAM, Hadley. **ggplot2**. Cham: Springer International Publishing, 2016. (Use R!). Disponível em: <http://link.springer.com/10.1007/978-3-319-24277-4>. Acesso em: 15 jan. 2024.

WILDT, Stefan; GERNGROSS, Tillman U. The humanization of N-glycosylation pathways in yeast. **Nature Reviews Microbiology**, [s. l.], v. 3, n. 2, p. 119–128, 2005.

YAN, Chunxiao *et al.* Auxotrophs compromise cell growth and fatty acid production in *Saccharomyces cerevisiae*. **Biotechnology Journal**, [s. l.], v. 18, n. 4, p. 2200510, 2023.

YAO, Yuan *et al.* Advances in Approaches to Study Chromatin-Mediated Epigenetic Memory. **ACS Synthetic Biology**, [s. l.], v. 11, n. 1, p. 16–25, 2022.

ZHANG, Qian *et al.* Stochastic expression and epigenetic memory at the yeast HO promoter. **Proceedings of the National Academy of Sciences**, [s. l.], v. 110, n. 34, p. 14012–14017, 2013.

ZHANG, Nianshu *et al.* Using yeast to place human genes in functional categories. **Gene**, [s. l.], v. 303, p. 121–129, 2003.

ZHAO, Yu *et al.* Debugging and consolidating multiple synthetic chromosomes reveals combinatorial genetic interactions. **Cell**, [s. l.], 2023.

Apêndice A - Tabela com as descrição dos genes vizinhos dos pGSHs de *S. cerevisiae*

pGSH	Gene	Nome Sistemático (SGD)	Anotação (SGD)
1	MAL32	YBR299W	Maltase envolvida na hidrólise dos dissacarídeos maltose, turanose, maltotriose e sacarose.
	YBR300C	YBR300C	ORF potencial; improvável que codifique proteína funcional.
2	PUG1	YER185W	Transportador de protoporfirina e heme; expressão induzida por baixo teor de heme/oxigênio.
	YER186C	YER186C	Proteína putativa de função desconhecida.
3	YHL041W	YHL041W	Proteína putativa de função desconhecida.
	ARN1	YHL040C	Transportador de ferro; relocalização e expressão induzida por estresse da replicação do DNA.
4	YHR095W	YHR095W	ORF potencial; improvável que codifique proteína funcional.
	HXT5	YHR096C	Transportador de hexose; expressão induzida por fontes de carbono não fermentáveis e pela diminuição na taxa de crescimento.
5	COS9	YKL219W	Proteína endossômica envolvida na internalização das proteínas da membrana plasmática para degradação; expressão induzida por limitação de nutrientes.
	SRY1	YKL218C	Desidratase de 3-hidroxiaspartato para formar oxaloacetato e amônia; necessária na presença de hidroxiaspartato.
6	ULS1	YOR191W	Proteína envolvida na inibição de NHEJ nos telômeros; relocaliza-se do núcleo para o citoplasma após estresse de replicação do DNA.
	THI72	YOR192C	Transportador de tiamina; captação de acadesina.

Apêndice B - Tabela com as características genômicas dos pGSHs identificadas em
H. sapiens

pGSH	Chromossomo	Coordenadas	Genes flanqueadores	Comprimento
1	1	10.895.999 - 10.946.470	HSPE1P24 - C1orf127	50.472 pb
2	2	80.648.862 - 80.699.387	CTNNA2 - AC012355.1	50.526 pb
3	3	109.243.126 - 109.293.787	AC063923.1 - DPPA2	50.662 pb
4	4	56.710.615 - 56.760.918	RPL17P20 - AC022483.1	50.304 pb
5	4	151.833.098 - 151.883.710	AC107074.1 - AC112242.2	50.613 pb
6	6	11.811.249 - 11.861.625	AL022724.2 - AMD1P4	50.377 pb
7	6	71.076.376 - 71.126.892	BECN1P2 - U3	50.517 pb
8	6	132.817.565 - 132.868.217	RPS12 - HMGB1P13	50.653 pb
9	7	127.296.204 - 127.346.789	PRELID3BP10 - ZNF800	50.586 pb
10	8	79.520.650 - 79.571.152	AC100870.1 - RPL3P9	50.503 pb
11	9	109.324.797 - 109.375.465	Y_RNA - PTPN3	50.669 pb
12	11	75.018.894 - 75.069.242	NEU3 - AP003175.1	50.349 pb
13	11	121.928.977 - 121.979.570	AP001977.1 - AP001924.1	50.594 pb
14	13	112.434.690 - 112.485.010	SPACA7 - TUBGCP3	50.321 pb
15	14	97.582.147 - 97.632.646	LINC02325 - LINC02291	50.500 pb
16	17	54.703.431 - 54.754.164	AC005951.1 - ARL2BPP8	50.734 pb

Apêndice C - Tabela com as características genômicas dos pGSHs identificadas em

M. musculus

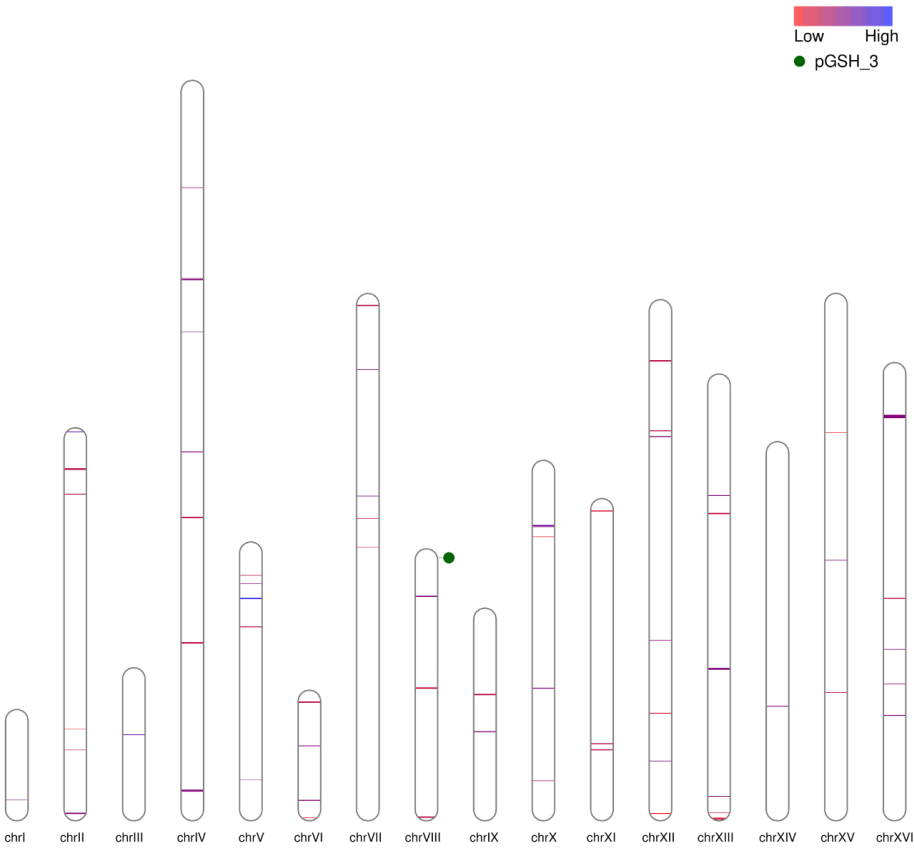
pGSH	Chromossomo	Coordenadas	Genes flanqueadores	Comprimento
1	1	181.460.642 - 181.509.632	Cnih3 - Ccdc121	48.991 pb
2	7	10,909.302 - 10.958.436	Gm45279 - Mir297-1	49.135 pb
3	7	13,670.808 - 13.719.692	Sult2a5 - Tmem167-ps1	48.885 pb
4	10	10.31.640.461 - 31.689.309	Gm8793 - Nkain2	48.849 pb
5	12	77.476.339 - 77.525.281	Fut8 - Gm48177	48.943 pb
6	17	19.938.120 - 19.986.867	Vmn2r-ps124 - Vmn2r-ps125	48.748 pb
7	17	84.853.452 - 84.902.159	Gm9400 - Gm49982	48.708 pb
8	X	12.821.493 - 12.870.268	Gm14634 - Gm25063	48.776 pb
9	X	23.806.026 - 23.854.818	Wdr44 - Gm4789	48.793 pb
10	Y	6.978.214 - 7.027.310	Gm28445 - Gm20788	49.097 pb
11	Y	8.198.620 - 8.247.645	Gm29043 - Gm29044	49.026 pb

Apêndice D - Tabela com a lista dos pacotes requeridos para a execução do SHIP

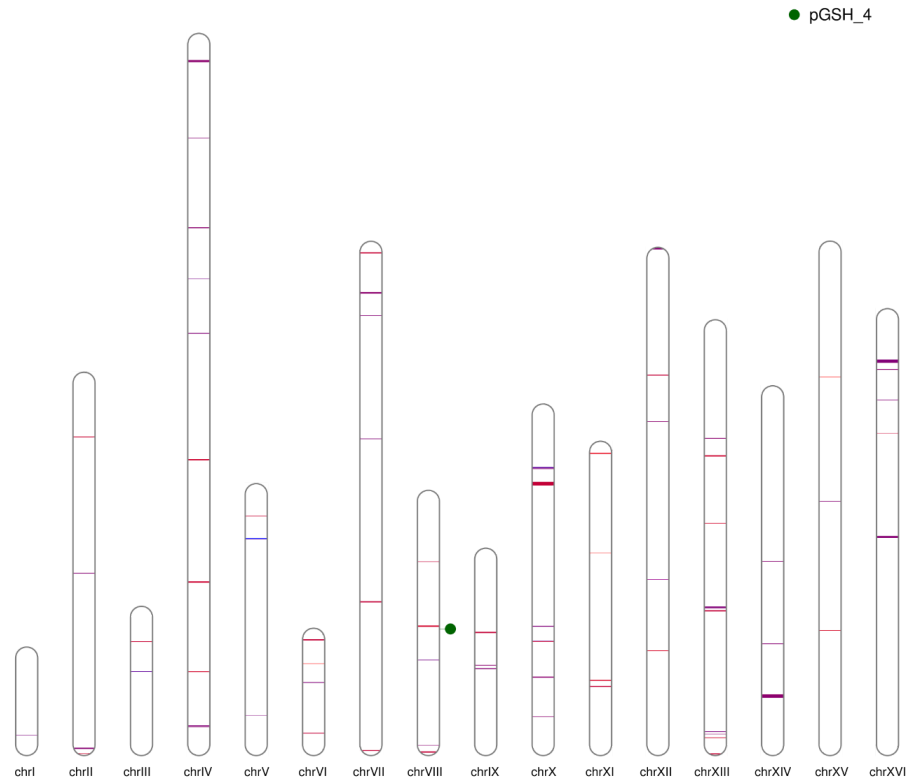
Pacotes	Versão
bio	1.5.9
biopython	1.81
biothings-client	0.3.0
certifi	2023.5.7
charset-normalizer	3.1.0
contourpy	1.0.7
cycler	0.11.0
fonttools	4.39.4
gprofiler-official	1.0.0
idna	3.4
kiwisolver	1.4.4
matplotlib	3.7.1
mygene	3.2.2
numpy	1.24.3
packaging	23.1
pandas	2.0.2
Pillow	9.5.0
platformdirs	3.5.1
pooch	1.7.0
pyparsing	3.0.9
python-dateutil	2.8.2
pytz	2023.3
requests	2.31.0
six	1.16.0
tqdm	4.65.0
tzdata	2023.3
urllib3	2.0.3

Apêndice E - Ideograma mostrando os locais dos genes diferencialmente expressos

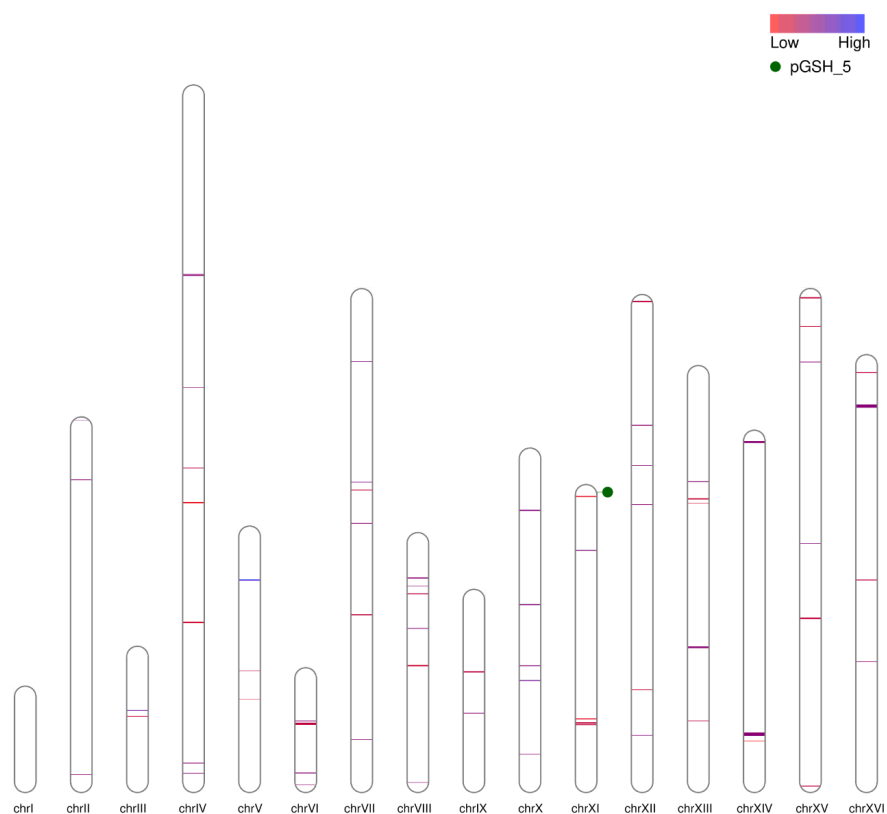
a. S3U



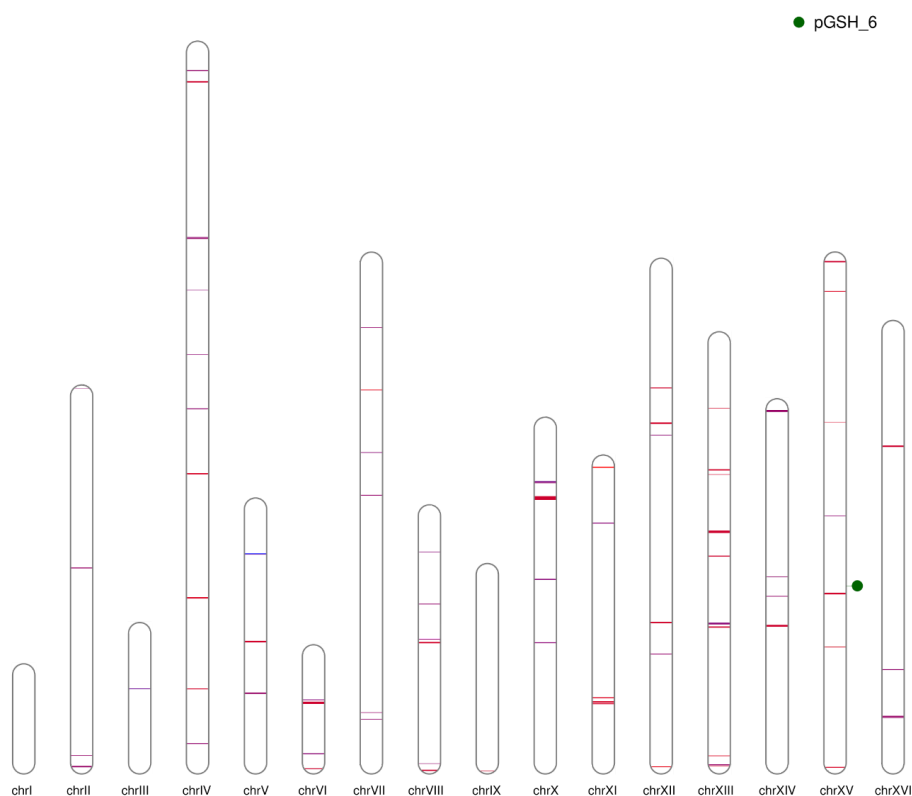
b. S4U



c. S5U



d. S6U



Ideograma mostrando os locais dos genes diferencialmente expressos. a) S3U b) S4U c) S5U e e) S6U. Os genes são coloridos de acordo com a alteração de $\log_2 \text{fold change} > 2$. O vermelho indica regulação negativa e o azul, regulação positiva. O loco do pGSH referente à linhagem analisada é mostrado como um círculo verde.

Apêndice F - Tabela com a lista de genes diferencialmente regulados positivamente

Gene	logfc	logcpm	pvalue	fdr	Experimento
YEL021W	14.16	6.19	0.0000000	0.0000000	S3U
YBL108W	6.72	-1.21	0.0079849	0.0393586	S3U
YJL153C	5.54	10.74	0.0000000	0.0000000	S4U
YCR013C	5.53	0.22	0.0000000	0.0000015	S3U
YJR037W	4.68	-0.87	0.0006270	0.0048588	S5U
YEL020W-A	4.26	0.12	0.0000005	0.0000180	S2U
YHR126C	3.80	-0.18	0.0010637	0.0082241	S4U
YDR034C-A	3.79	0.89	0.0001430	0.0018162	S4U
YMR173W-A	3.47	0.67	0.0000024	0.0000610	S6U
YGL041W-A	3.37	2.12	0.0000007	0.0000277	S3U
YDR510C-A	3.33	0.71	0.0000914	0.0013241	S3U
YLR415C	3.27	-0.10	0.0010496	0.0084947	S3U
YFL010W-A	3.20	0.81	0.0000176	0.0002931	S6U
YAR030C	3.13	-0.61	0.0014750	0.0105582	S4U
YKL164C	3.12	8.70	0.0000000	0.0000000	S6U

Apêndice G - Tabela com a lista de genes diferencialmente regulados negativamente

Gene	logfc	logcpm	pvalue	fdr	Experimento
YMR322C	-8.29	0.42	0.0000000	0.0000000	S3U
YGL015C	-8.18	-0.28	0.0000080	0.0001763	S2U
YLR461W	-7.53	-0.91	0.0000020	0.0000675	S3U
YLR365W	-7.53	-0.14	0.0001399	0.0018274	S3U
YNR025C	-7.22	-0.15	0.0000220	0.0003572	S5U
YNL318C	-7.20	-0.71	0.0032048	0.0187096	S2U
YBL111C	-7.17	-0.03	0.0000090	0.0001925	S2U
YJL135W	-7.08	-0.87	0.0006824	0.0060587	S3U
YGL118C	-6.98	-1.08	0.0000225	0.0004079	S2U
YAR070C	-6.98	-1.05	0.0007250	0.0061361	S2U
YBR200W-A	-6.81	-0.83	0.0002817	0.0031572	S3U
YOL019W-A	-6.70	-0.45	0.0003527	0.0037355	S3U
YFL032W	-6.59	-0.33	0.0000463	0.0007567	S4U
YKL106C-A	-6.07	-0.93	0.0041950	0.0233514	S4U
YDR218C	-5.96	1.65	0.0000014	0.0000396	S5U

Apêndice H - Registro do software SHIP pelo INPI



REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512023002017-6**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 01/09/2020, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

Título: SHIP - Safe Harbor Identification Program

Data de criação: 01/09/2020

Titular(es): FUNDAÇÃO UNIVERSIDADE DE BRASILIA

Autor(es): CÍNTIA MARQUES COELHO; GEORGIOS PAPPAS JR; MATHEUS DE CASTRO LEITÃO

Linguagem: PYTHON

Campo de aplicação: BL-02

Tipo de programa: TC-01

Algoritmo hash: SHA-512

Resumo digital hash:

547BA123FA96D08F38D86EF09F7F3975CA9C3B991FA186D6F41ABBC9653C9288684BAAEEA2159767A38B4B161
5B5A3A9B3C2600DF7F09A08493899FCB6498D12

Expedido em: 18/07/2023

Aprovado por:

Carlos Alexandre Fernandes Silva
Chefe da DIPTO