Identifying improved sites for heterologous gene integration using ATAC-seq

Joseph R. Brady[1,2], Melody C. Tan[1], Charles A. Whittaker[1], Noelle A. Colant[1,2], Neil C. Dalvie, Kerry Routenberg Love[1], J. Christopher Love[1,2]*

1. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA
2. Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA

* Corresponding author:

J. Christopher Love
77 Massachusetts Avenue / 76-253
Cambridge, MA 02139
United States

phone: (617) 324-2300
email: clove@mit.edu

SUPPORTING INFORMATION

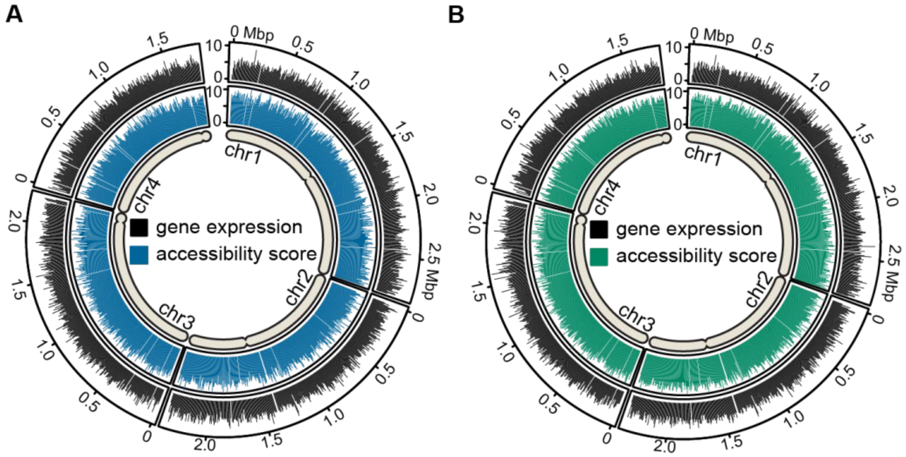

Figure S1. Gene expression measured as log$_2$(TPM +1) and log$_2$(accessibility score) for 7.5 kbp intervals across each chromosome for A) glycerol or B) methanol conditions.

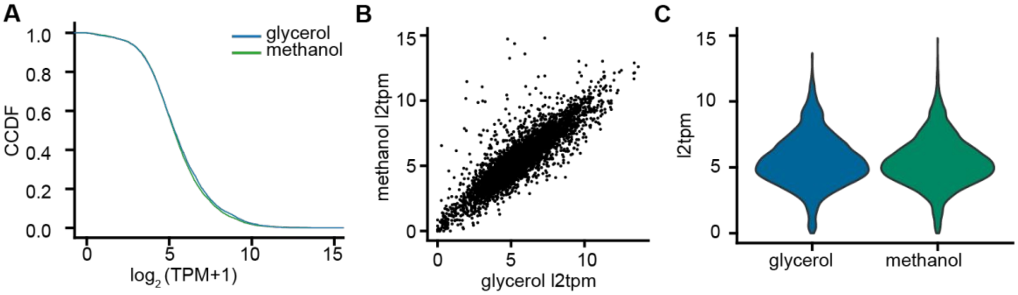

Figure S2. Gene expression in *K. phaffii* across carbon sources. A) Complementary cumulative distribution function (CCDF) of gene expression measured as $\log_2$ (TPM+1) or l2tpm for all genes in glycerol or methanol conditions. B) Gene expression in methanol condition vs. glycerol condition. C) Distributions of gene expression in methanol and glycerol conditions.

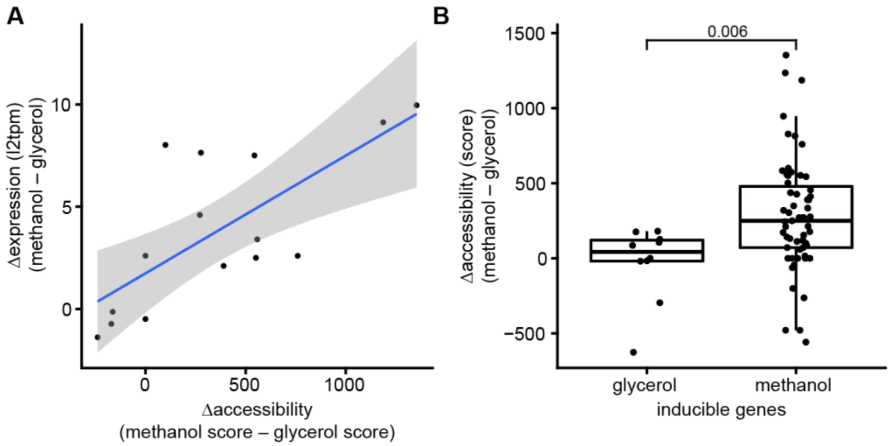

Figure S3. Effect of changes in chromatin accessibility on changes in gene expression. A) Change in gene expression versus change in accessibility score from glycerol to methanol condition for genes in the methanol utilization pathway. B) Distribution of changes in accessibility score from glycerol to methanol for glycerol-specific and methanol-specific genes. Genes with residuals at least three standard deviations from zero on a plot of expression in methanol versus expression in glycerol were deemed carbon-source specific (see Figure S1B). P-value was computed using a student's t-test.

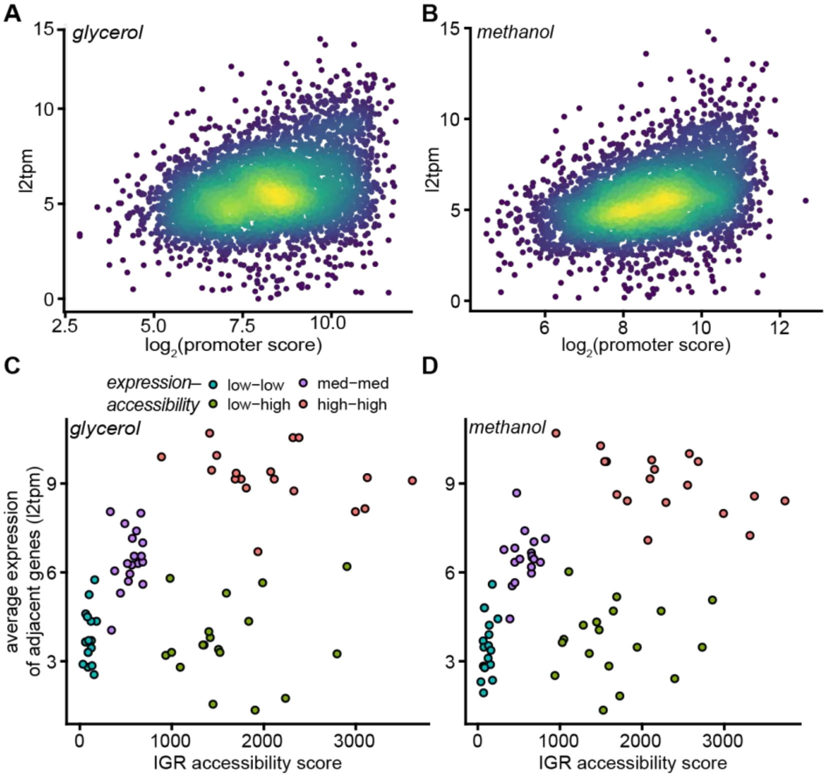

Figure S4. Selection of IGRs with varied adjacent gene expression and accessibility scores. A-B) Gene expression (l2tpm) versus $\log_2$(promoter accessibility score) for all genes in A) glycerol and B) methanol conditions. C-D) Neighboring gene expression versus accessibility score in C) glycerol and D) methanol conditions for selected IGRs used in library. IGRs divided into four categories for expression–accessibility (low-low, low-high, medium-medium, and high-high).

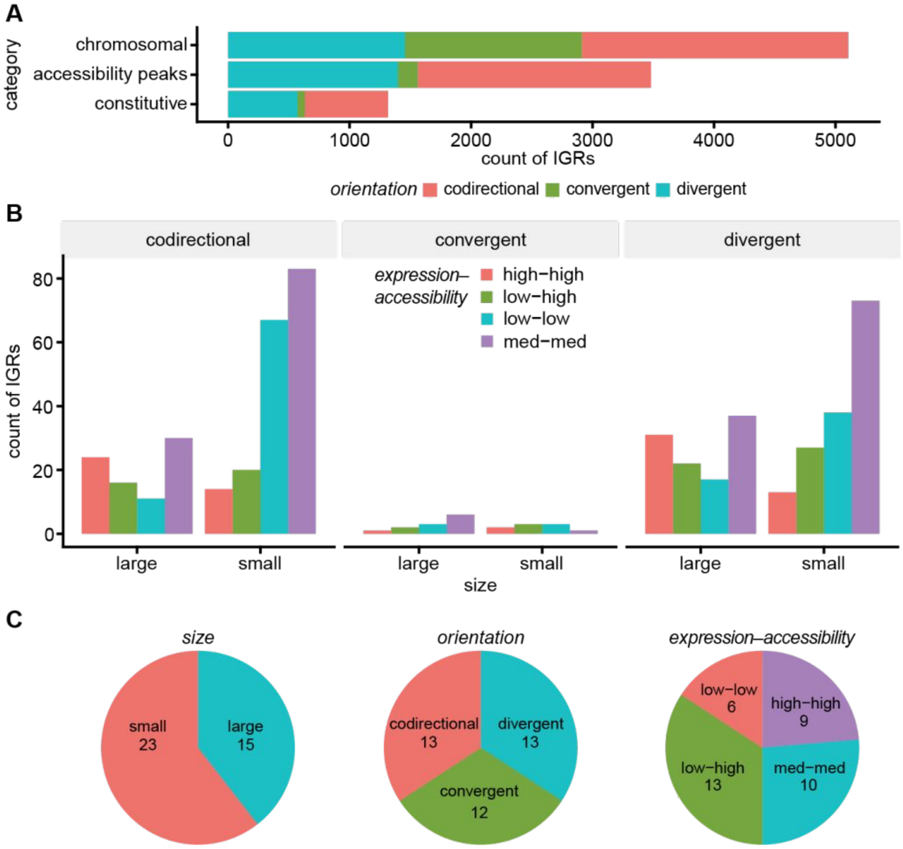

Figure S5. Demographics of IGRs in *K. phaffii*. A) Counts of IGRs separated by the orientation of adjacent genes for each of three increasingly restrictive categories: all chromosomal IGRs, chromosomal IGRs with detected accessibility peaks, and those of the previous category that are constitutive for both gene expression and accessibility across carbon sources. B) Of constitutive IGRs, counts of genomic IGRs within each combination of IGR size, orientation, and expression-accessibility category. C) Representation of sizes, orientations, and expression-accessibility categories across confirmed integrants of IGR library.

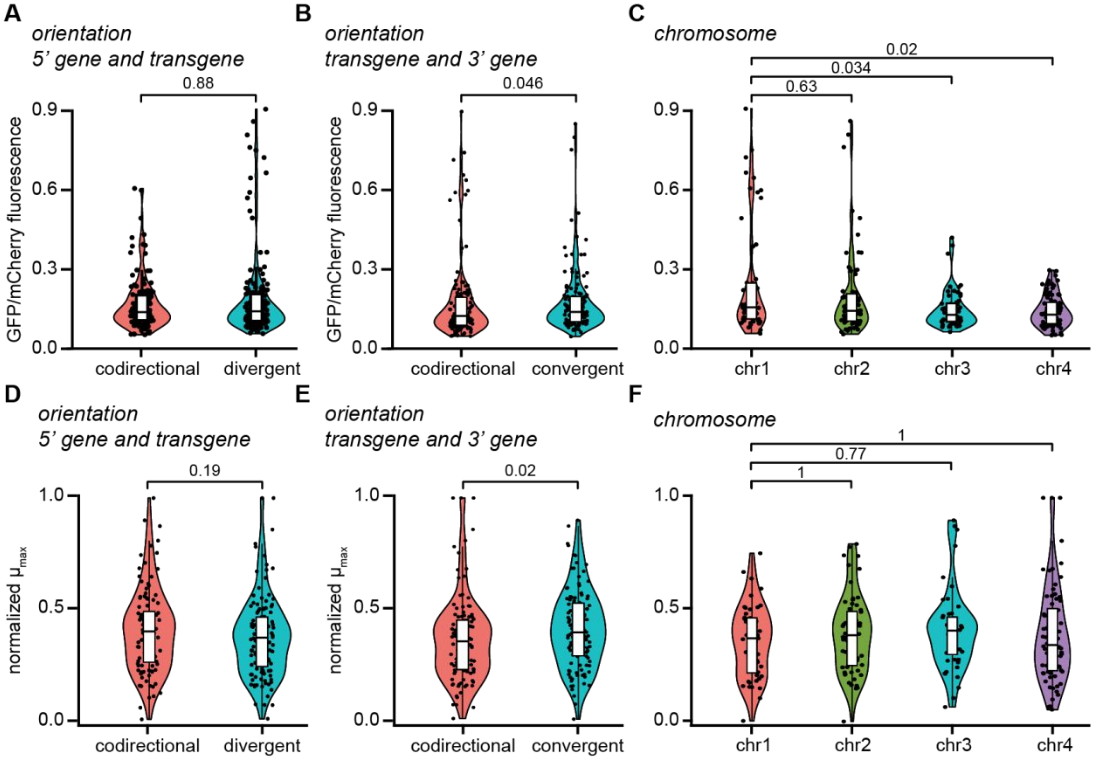

Figure S6. Evaluation of additional genomic properties that impact suitability of a landing site for heterologous gene integration. A-F) Evaluation of IGR properties A,D) orientation of the 5' gene and transgene pair, B,E) orientation of 3' gene and transgene pair, and C,F) chromosome. A-C) GFP fluorescence normalized by mCherry fluorescence and D-F) normalized max growth rate for each biological replicate and IGR library member. Adjusted p-values computed using a Wilcoxon signed-rank test with the Benjamini-Hochberg correction for multiple hypotheses.

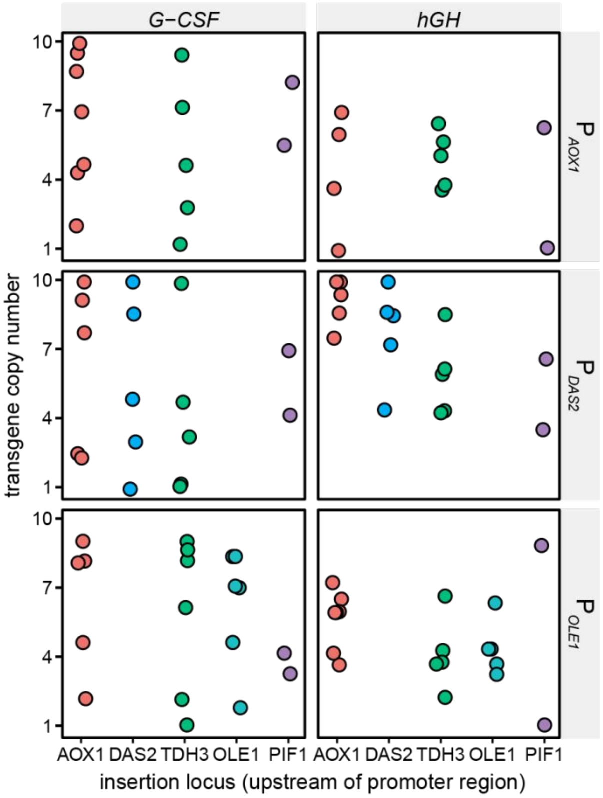

Figure S7. Comparison of transgene copy number among insertion loci across three promoters (P_{AOX1}, P_{DAS2}, and P_{OLE1}) and two heterologous genes (*G-CSF* and *hGH*).

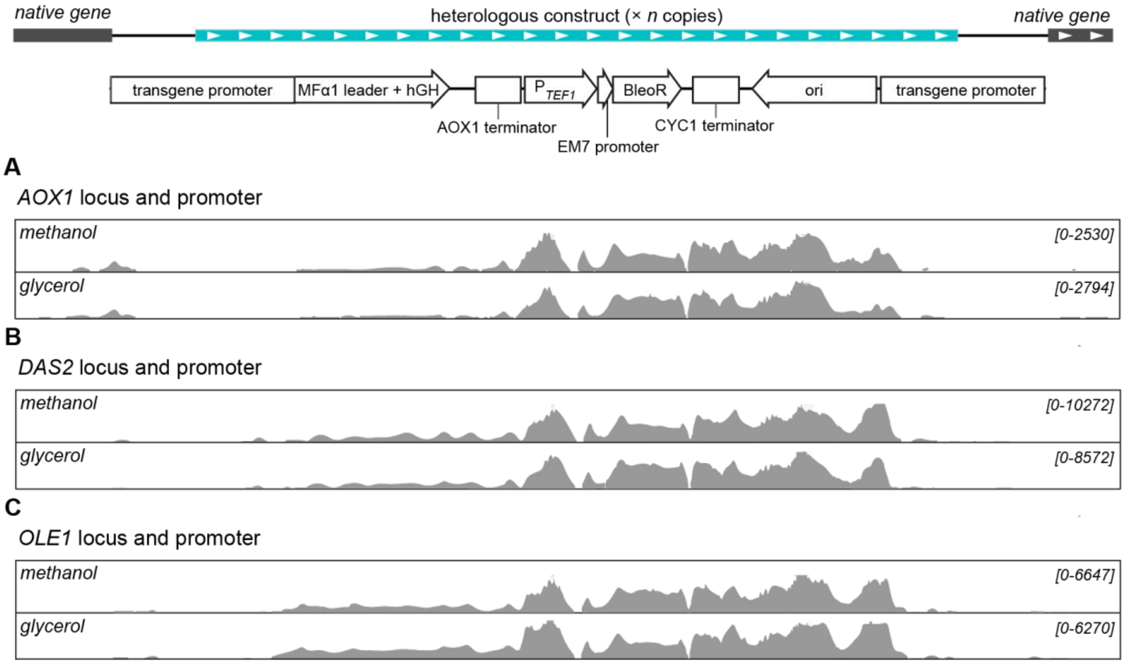

Figure S8. ATAC-seq within three heterologous constructs for methanol and glycerol conditions. Constructs were integrated to obtain tandem, multiple copies just upstream of A) *AOX1*, B) *DAS2*, or C) *OLE1*. The interval is shown for each alignment track, ranging from 0 to the maximum read depth for that track.

Table S1. Expression and accessibility of loci used for multi-copy integration.

locus	acc. score	5' gene expr.	3' gene expr.	condition
AOX1	no peak	5.7	4.8	glycerol
AOX1	1354	5.0	14.7	methanol
DAS2	85	5.7	5.2	glycerol
DAS2	1272	6.0	14.4	methanol
GAPDH	1593	3.6	13.4	glycerol
GAPDH	1521	3.4	11.7	methanol
OLE1	1990	4.5	11.3	glycerol
OLE1	2170	5.0	12.2	methanol
PIF1	875	13.4	3.6	glycerol
PIF1	no peak	11.7	3.4	methanol