Supplemental information

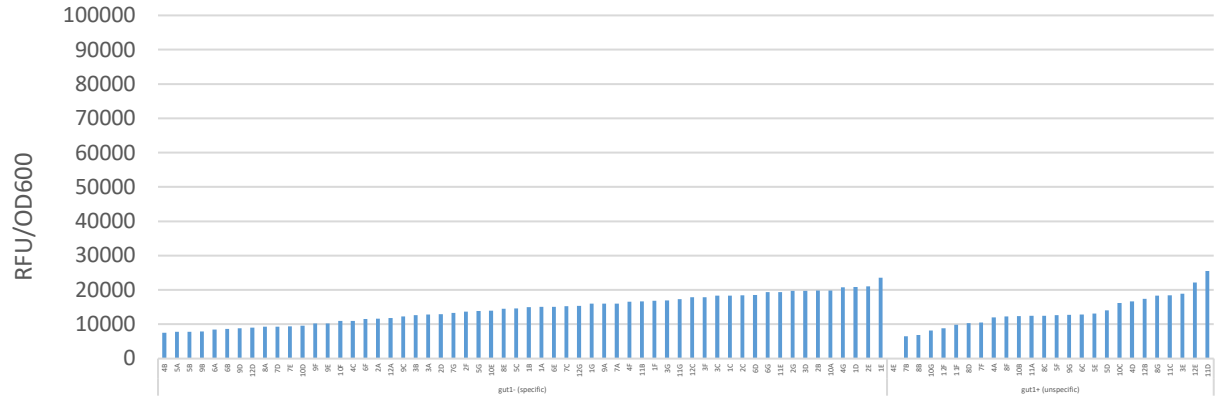
S 1

S 1: Complete screening data for 755 transformants from three plasmid/linearization combinations and transformation performed in triplicate. Cells were pre-grown on glucose for 60 h and subsequently induced with methanol for 48 h. On the x-axis the well numbers of each clone is provided. *GUT1* targeting *Swa*I linearized vectors were replica-plated in glycerol containing media after growth on glucose for 60 h to test for specific/non-specific integration. Wells H1-3 of each plate were loaded with the wildtype strain as a negative control and wells H4-12 left empty as sterile controls.

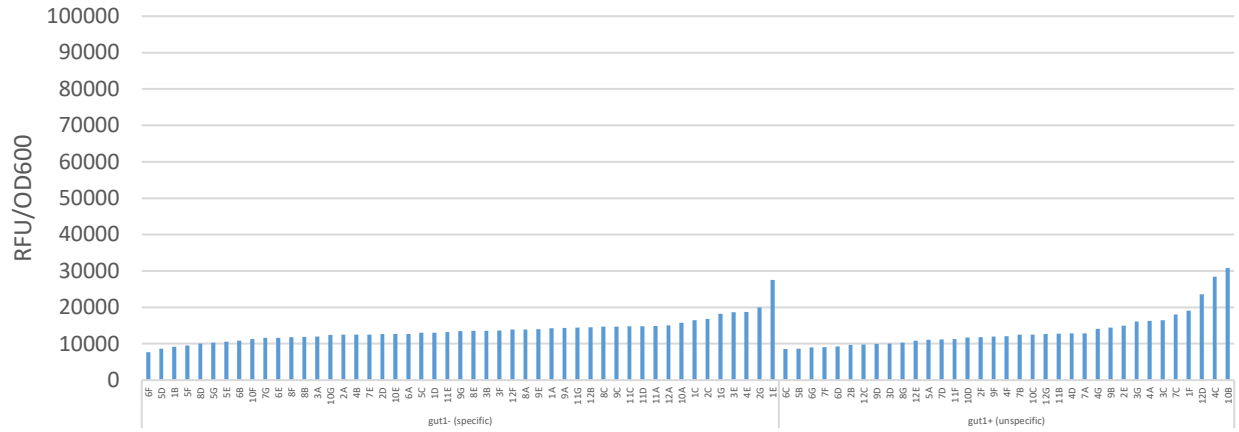
One transformant of GUT1-*Sac*I replicate 2 did not grow on glucose and was hence omitted from the analysis. As the transformant grew as a colony on agar plates after transformation this result is hard to rationalize but may be explained by a rare integration event (Schwarzhans J-P, Wibberg D, Winkler A, Luttermann T, Kalinowski J, Friehs K. Sci Rep 6:38952, 2016).

GUT1 integration vector linearized with *Swa*I

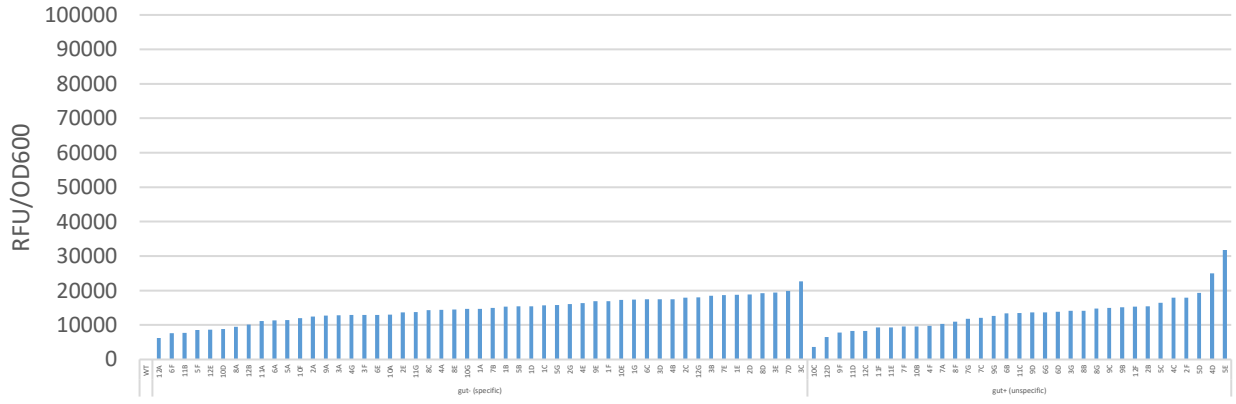
Replicate 1



Replicate 2

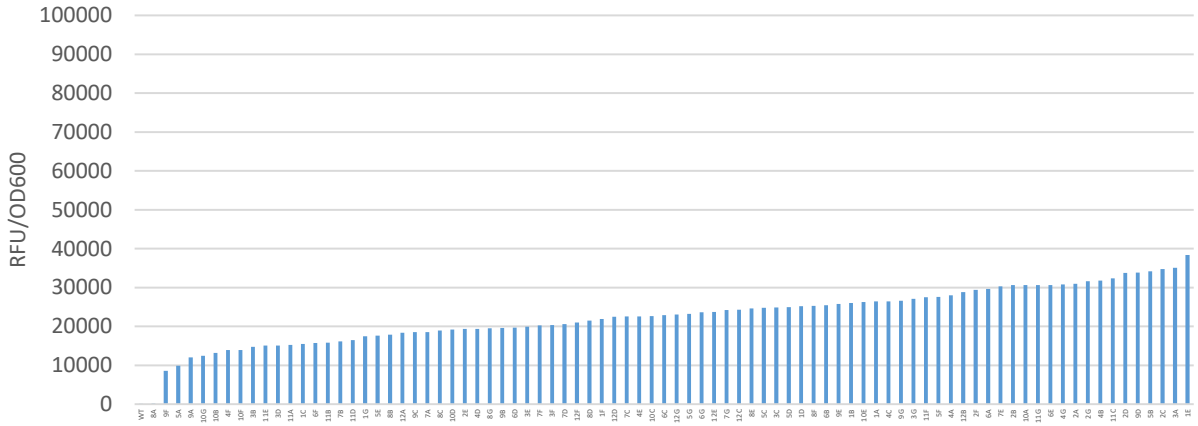


Replicate 3

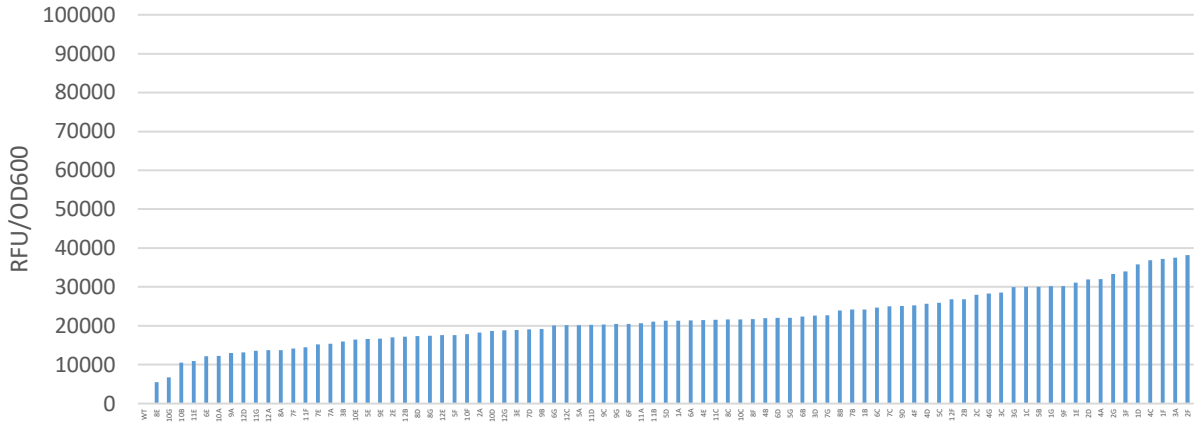


GUT1 integration vector linearized with *Sac*I

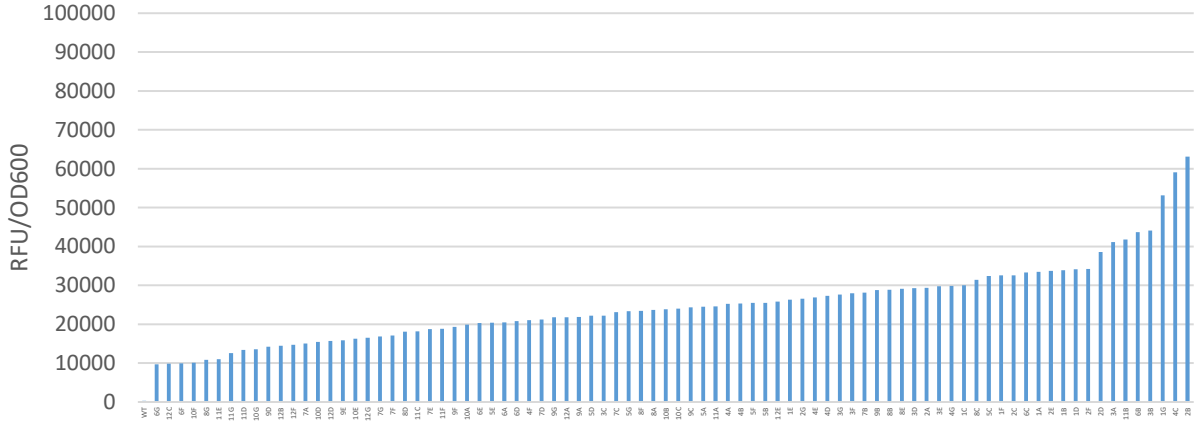
Replicate 1



Replicate 2

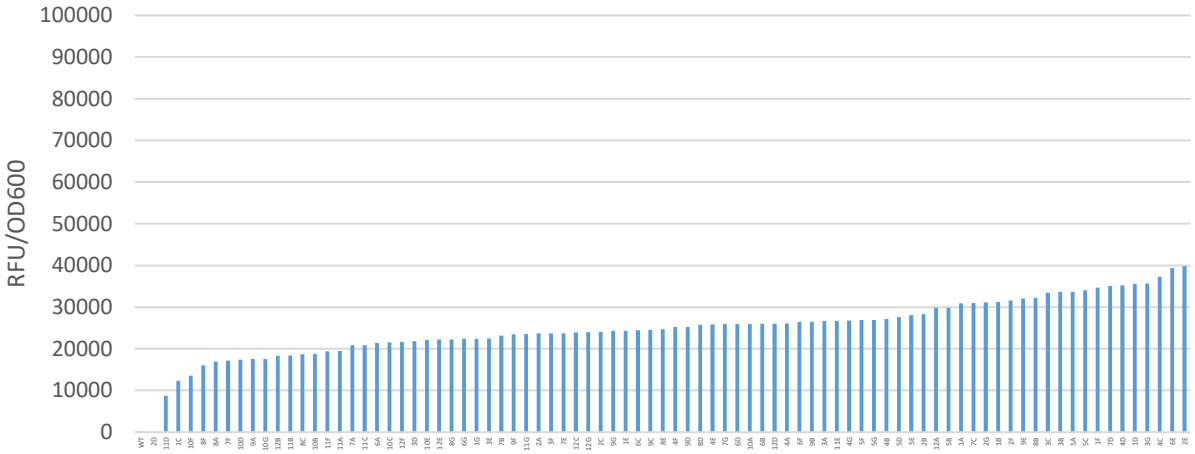


Replicate 3

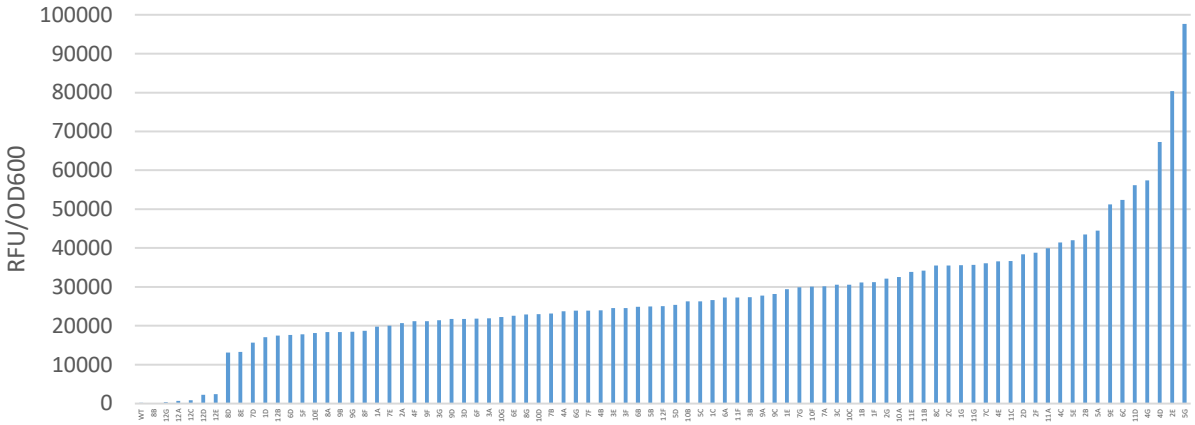


STD vector without integration sequences linearized with *Sac*I

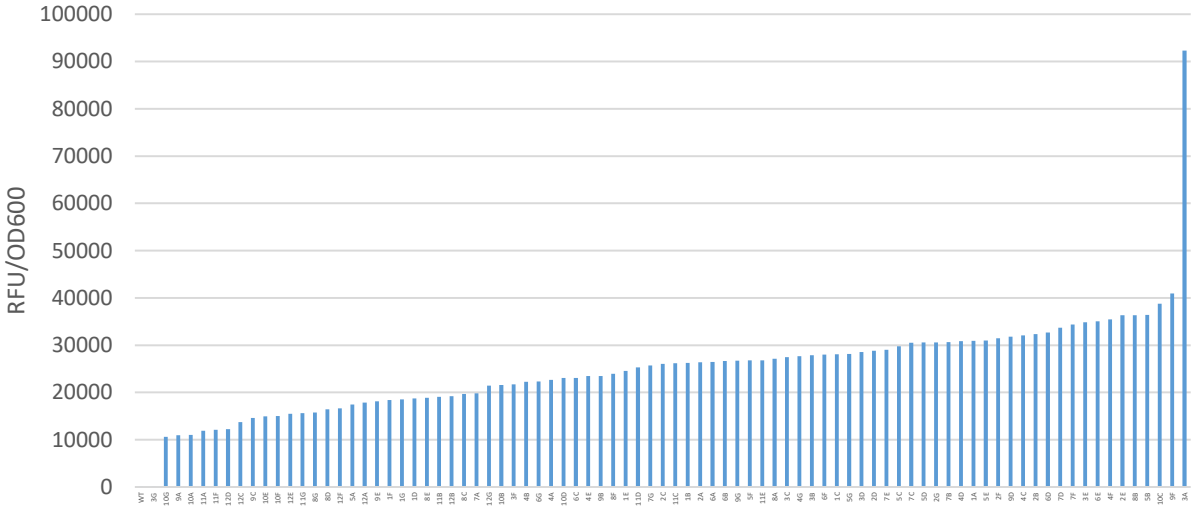
Replicate 1



Replicate 2



Replicate 3



S 2

S 2: Integration rates of specific and non-specific GUT1-SwaI transformants from different replicates.

GUT1 targeting *Swa*I linearized vectors were replica-plated in glycerol containing media after growth on glucose for 60 h to test for specific/non-specific integration.

Transformation replicate (number)	Integration event (number of colonies)		Total (colonies)	% specific integration
	Specific (no growth on glycerol)	Non-specific (still growth on glycerol)		
1	58	26	84	69.0
2	49	35	84	58.3
3	51	33	84	60.7
Sum	158	94	252	**62.7**

S 3

S 3: Rescreening results of 44 selected transformants.

The summary table shows a comparison of screening, rescreening results and which transformants were used for whole genome sequencing. Reporter fluorescence measurements are shown in separate diagrams for each plasmid/linearization.

The strains were streaked as single colonies from glycerol stocks. Cells were pre-grown on glucose for 60 h and subsequently induced with methanol for 48 h. Mean values of biological four-fold replicates are shown. For GUT1-*Swa*I constructs transformants were replica-plated in glycerol containing media after growth on glucose for 60 h to confirm specific/non-specific integration.

Extended discussion

In general, the initial screening results were reproduced in the rescreening: Outliers that had shown low or no fluorescence (*e.g.* GUT1 *Swa*I clones R1-4E [*i.e.* replicate 1, well 4E]/QTV84 and R3-10C/QTV85) yielded similar results. Transformants showing increased expression also yielded reproducible results (*e.g.* GUT1 *Sac*I clones R1-1E/QTV92 and R3-4C/QTV93 or STD *Sac*I R2-2E/QTV95, R2-5G/QTV96 and R3-3A/QTV97). Transformants, that in the initial screening had shown only moderately reduced or increased expression, showed mostly average expression in the rescreening (*i.e.* similar reporter protein fluorescence as specifically integrated cassettes). This result was expected since we had initially sampled a large number of transformants. Even for a single strain, according to a normal distribution, a certain number of higher/lower expressing sample points would be expected. When these strains were now measured in biological replicates, this distribution issue was accounted for. Notably, for specifically integrated clones (of the GUT1 *Swa*I plasmid), we did not find any clones with clearly elevated or reduced expression, which is consistent with the boxplot analysis where only one outlier was apparent (Fig. 2D).

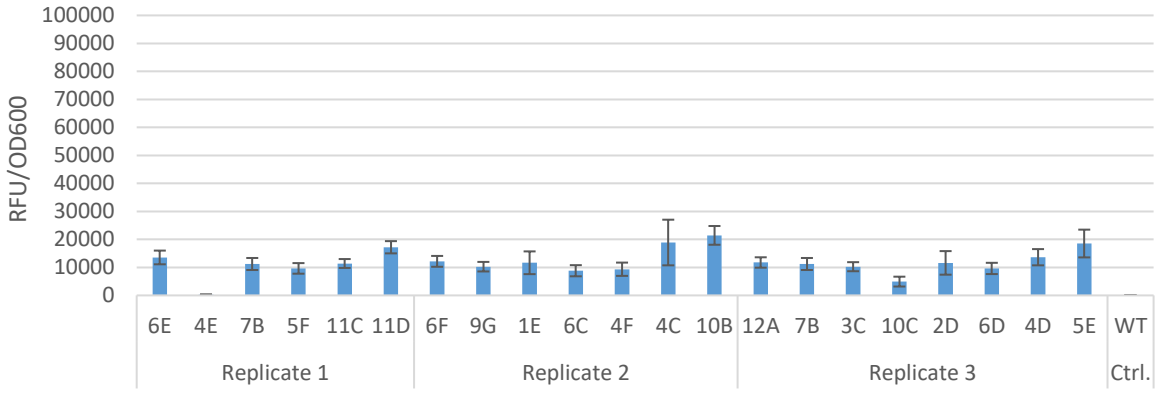
Summary table of rescreening clones

Vector & linearization	Transformant			Expression			Selected for genome sequencing	Identifier assigned
	Replicate	Identifier	Integration	Screening expression	Rescreening	Screening vs. rescreening		
GUT1 *Swa*I	1	6E	specific	average	average	confirmed	Yes	QTV76
		4E	non-specific	no expression	no expression	confirmed	Yes	QTV84
		7B	non-specific	low	average	dissimilar	No	n.a.
		5F	non-specific	average (rather low)	average	similar	No	n.a.
		11C	non-specific	average (rather high)	average	similar	Yes	QTV79
		11D	non-specific	~high	~high	confirmed	Yes	QTV82
	2	6F	specific	~low	average	dissimilar	No	n.a.
		9G	specific	average	average	confirmed	Yes	QTV77
		1E	specific	~high	~average	dissimilar (large SD)	No	n.a.
		6C	non-specific	low	average	dissimilar	Yes	QTV80
		4F	non-specific	average	average	confirmed	No	n.a.
		4C	non-specific	~high	~high (large SD)	similar (large SD)	No	n.a.
		10B	non-specific	~high	~high	confirmed	Yes	QTV83
	3	12A	specific	low-average	average	dissimilar	No	n.a.
		7B	specific	average	average	confirmed	Yes	QTV78
		10C	non-specific	low	low	confirmed	Yes	QTV85
		3C	specific	average-high	average	dissimilar	No	n.a.
		2D	specific	low-average	average	dissimilar	No	n.a.
		6D	non-specific	average	average	confirmed	Yes	QTV81
		4D	non-specific	~high	~average	dissimilar	No	n.a.
		5E	non-specific	~high	~average-high (SD)	~dissimilar	No	n.a.
GUT1 *Sac*I	1	9F	-	low-average	low-average	confirmed	Yes	QTV86
		7G	-	average	average	confirmed	Yes	QTV89
		1E	-	average-high	average-high	confirmed	Yes	QTV92
	2	8E	-	low	low-average	~similar	Yes	QTV87
		10G	-	low	low-average	~similar	No	n.a.
		11C	-	average	average	confirmed	No	n.a.
		2F	-	average-high	average	dissimilar	Yes	QTV90

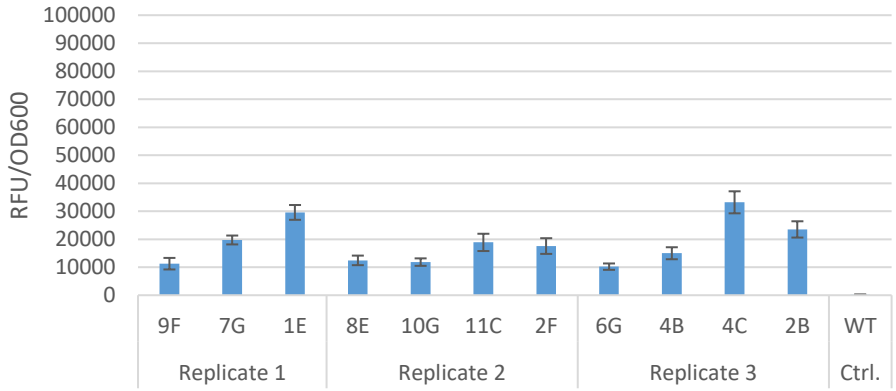
		6G	-	low-average	low-average	confirmed	Yes	QTV88
		4B	-	average	average	confirmed	Yes	QTV91
		4C	-	high	high	confirmed	Yes	QTV93
	3	2B	-	high	high	confirmed	Yes	QTV94
STD *Sac*I	1	11D	-	low-average	average	dissimilar	No	n.a.
		4F	-	average	average	confirmed	Yes	QTV98
		2E	-	average-high	average	dissimilar	No	n.a.
	2	12C	-	low	average	dissimilar	No	n.a.
		12E	-	low	average	dissimilar	Yes	QTV99
		8D	-	low-average	average	no, average	No	n.a.
		6A	-	average	low-average (SD)	~dissimilar (SD)	No	n.a.
		2E	-	high	high	confirmed	Yes	QTV95
		5G	-	high	high	confirmed	Yes	QTV96
	3	10G	-	low-average	low-average	confirmed	No	n.a.
		6B	-	average	average	confirmed	Yes	QTV100
		3A	-	high	high	confirmed	Yes	QTV97

Reporter fluorescence measurements

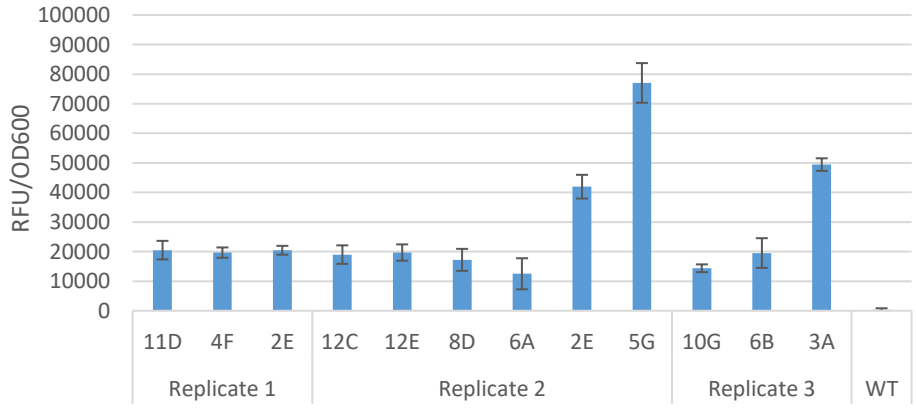
GUT1 integration vector linearized with SwaI



GUT1 integration vector linearized with SacI



STD vector without integration sequences linearized with SacI



S 4

S 4: Summary of statistics from Illumina MiSeq sequencing of 25 transformants (QTV76-100) and two controls (QTV19 and mut^S) and mapping to CBS7435, the mitochondrial genome and respective plasmid. Statistics were taken from BAM QC analysis files generated with Qualimap 2.2.

Line	Number of reads	Mapped paired reads (%)	Mean coverage (x times)	Mitochondrial genome coverage (x times)	Mean mapping quality	General error rate (%)	GC content (%)
QTV19	4,372,628	91.36	62.56	779.20	39.2	0.94	40.48
mut^S	8,128,680	97	123.88	635.02	40.51	0.59	41.33
QTV76	3,555,518	90	50.30	503.25	39	1.08	40.65
QTV77	3,092,864	90.21	43.92	457.58	38.99	1.10	40.62
QTV78	3,584,154	90.69	51.18	560.01	39.12	1.04	40.62
QTV79	3,146,518	88.82	43.89	433.56	38.76	1.15	40.70
QTV80	4,665,052	90.16	66.15	613.13	38.94	1.08	40.86
QTV81	4,825,938	91.64	69.34	494.37	39.32	0.95	41.16
QTV82	4,840,258	92.44	70.34	573.92	39.47	0.90	41.04
QTV83	3,359,162	91.2	47.98	427.96	39.24	0.91	40.77
QTV84	10,477,258	97.79	161.02	488.41	40.55	0.54	41.34
QTV85	3,227,268	90.86	46.13	281.66	39.21	0.97	40.90
QTV86	3,998,764	93.29	58.66	444.50	39.69	0.80	40.76
QTV87	4,974,690	89.64	69.86	600.19	38.9	1.05	40.59
QTV88	5,872,234	91.43	84.65	641.39	39.32	0.98	40.67
QTV89	5,452,592	91	78.23	601.78	39.21	1.03	40.82
QTV90	6,080,906	91.71	87.97	632.90	39.35	0.98	40.88
QTV91	5,793,460	88.91	80.72	757.23	38.67	1.13	40.76
QTV92	13,395,412	95.94	201.04	657.43	40.35	0.60	41.38
QTV93	4,291,388	92.87	62.54	485.41	39.52	0.84	40.67
QTV94	5,509,700	92.49	79.18	640.72	39.48	0.80	40.70
QTV95	5,669,660	92.23	81.98	650.28	39.49	0.87	40.72
QTV96	6,924,978	89.45	96.88	779.59	38.87	1.06	40.64
QTV97	9,552,652	94.95	142.88	346.94	40.13	0.77	41.58
QTV98	6,227,676	96.05	93.44	625.46	40.25	0.64	40.77
QTV99	5,032,884	92.37	72.51	405.23	39.49	0.89	40.84
QTV100	16,303,798	98.14	249.54	165.29	40.85	0.51	41.51

S 5

S 5: Example of sequence mapping to identify the insertions sites listed in Tab. 1 in the main manuscript.

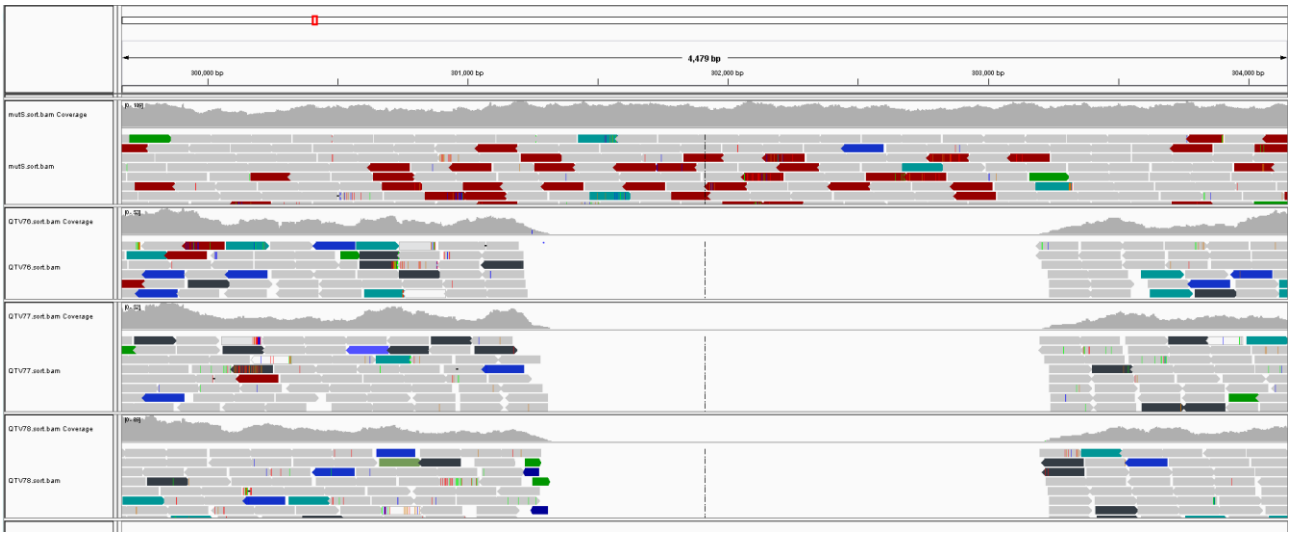
(A) Example 1: IGV window of mut^S (wild type), QTV76, QTV77 and QTV78 reads mapped against the reference. In this zoom in on chromosome 4, the deletion of the *GUT1* gene (1866 bp from 301,326 to 303,192) can be clearly seen by the absence of mapped reads in the transformants.

(B) Example 2: The BLAST readwalking method pinpointed the putative insertion site on chromosome 2 in QTV80. In this zoom in on IGV, the BLAST reads are seen in the bottom panel, above the genome annotation. The mapping is disrupted in QTV80 (top panel) and the reads that did map contain parts of the plasmid and genome. It is compared to the perfect mapping of the wildtype (mut^S) in the second panel.

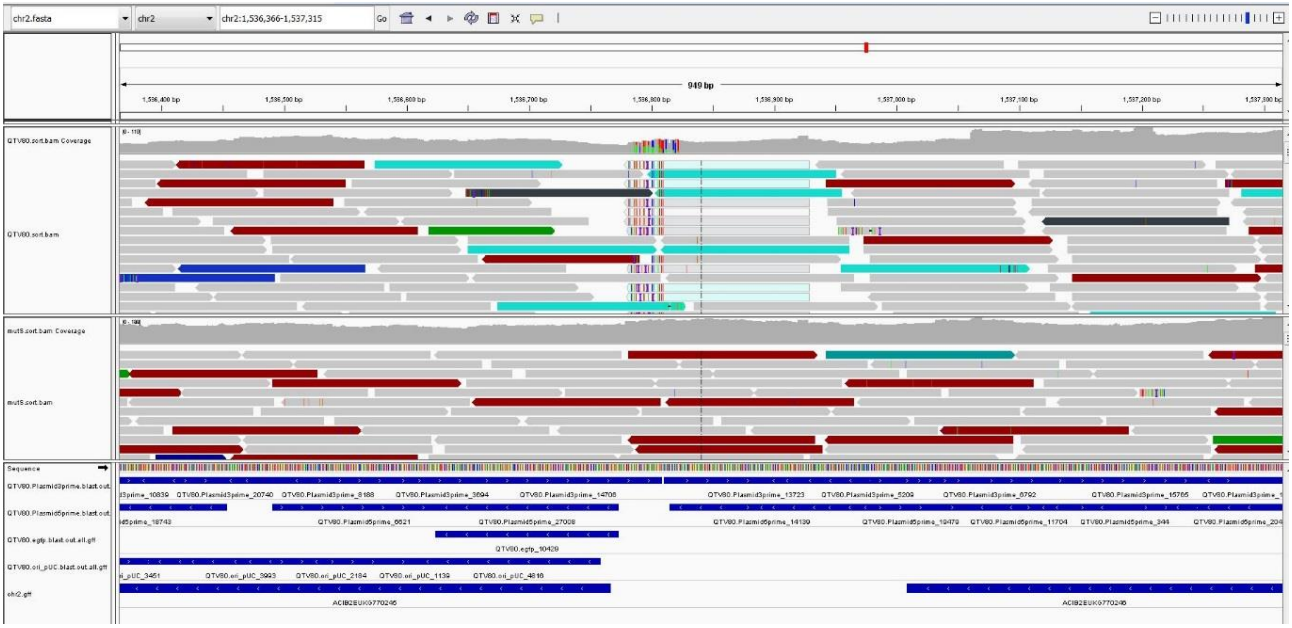
(C) Example 3: The BLAST readwalking approach pinpointed a deletion on chromosome 1 in QTV85 which is a potential plasmid insertion site. In the zoom in on IGV, the BLAST reads are in the bottom panel, just above the annoted genes. The deletion can be clearly seen in QTV85 (top) panel compared to the wildtype (mut^S) mapping.

(D) Example 4: The BLAST readwalking approach (BLASTed reads in bottom panel) pinpointed an area of disrupted mapping in the AOX1 promoter on chromosome 4 in QTV86 (top panel) compared to the wildtype (mut^S, second panel) and other lines (here QTV88, third panel).

A


4,479 bp
300,000 bp
301,000 bp
302,000 bp
303,000 bp
304,000 bp
mutS.sort.bam Coverage
mutS.sort.bam
QTV76.sort.bam Coverage
QTV76.sort.bam
QTV77.sort.bam Coverage
QTV77.sort.bam
QTV78.sort.bam Coverage
QTV78.sort.bam

B


chr2.fasta
chr2
chr2:1,536,366-1,537,315
Go
949 bp
1,536,400 bp
1,536,500 bp
1,536,600 bp
1,536,700 bp
1,536,800 bp
1,536,900 bp
1,537,000 bp
1,537,100 bp
1,537,200 bp
1,537,300 bp
QTV60.sort.bam Coverage
QTV60.sort.bam
mutS.sort.bam Coverage
mutS.sort.bam
Sequence
QTV60.Plasmid3prime.blast.out
QTV60.Plasmid5prime.blast.out
QTV60.egfp.blast.out.all.gff
QTV60.ori_pUC.blast.out.all.gff
chr2.gff
QTV60.egfp_10429
ACIB2EUK0770246

C


chr1.fasta
chr1
chr1:1,575,627-1,582,113
Go
6,480 bp
1,576,000 bp
1,577,000 bp
1,578,000 bp
1,579,000 bp
1,580,000 bp
1,581,000 bp
1,582,000 bp
QTV85.sort.bam Coverage
QTV85.sort.bam
mut5.sort.bam Coverage
mut5.sort.bam
QTV85.Plasmid5prime.blast.out
QTV85.egfp.blast.out.all.gff
QTV85.ori_pUC.blast.out.all.gff
chr1.gff
10 tracks loaded
chr1:1,577,739
782M of 1,115M

D


chr4noAOX1.fasta
chr4noAOX1
chr4noAOX1:237,558-238,579
Go
1,022 bp
237,600 bp
237,800 bp
238,000 bp
238,200 bp
238,400 bp
QTV86.sort.bam Coverage
QTV86.sort.bam
mut5.sort.bam Coverage
mut5.sort.bam
QTV88.sort.bam Coverage
QTV88.sort.bam
Sequence
QTV86.Plasmid5prime.blast.out
QTV86.Plasmid5prime.blast.out
QTV86.egfp.blast.out.all.gff
11 tracks loaded
chr4noAOX1:237,707
698M of 1,115M

S 6

S 6: List of genes deleted in the approximate 69 kbp fragment lost on chromosome 4 in QTV84.

Gene	Homology identified in annotation or domain found with BLAST
ACIB2EUKG773048	GAL4 transcription factor
ACIB2EUKG773049	Glycoside hydrolase family 3 protein
ACIB2EUKG773050	Glycoside hydrolase family 78 protein
ACIB2EUKG773051	Maltose permease, high-affinity maltose transporter (alpha-glucoside transporter)
ACIB2EUKG773052	Hypothetical protein
ACIB2EUKG773053	Hypothetical protein
ACIB2EUKG773054	Low-affinity Fe(II) transporter of the plasma membrane
ACIB2EUKG773055	Sugar phosphate permease [Carbohydrate transport and metabolism]
ACIB2EUKG773056	High affinity nicotinic acid plasma membrane permease
ACIB2EUKG773057	Proton-coupled oligopeptide transporter of the plasma membrane
ACIB2EUKG773058	Hypothetical protein
ACIB2EUKG773059	Putative flocculin
ACIB2EUKG773060	Hypothetical protein
ACIB2EUKG773061	Methionine sulfoxide reductase
ACIB2EUKG773062	NADPH-dependent medium chain alcohol dehydrogenase
ACIB2EUKG773063	Hypothetical protein
ACIB2EUKG773064	Hypothetical protein

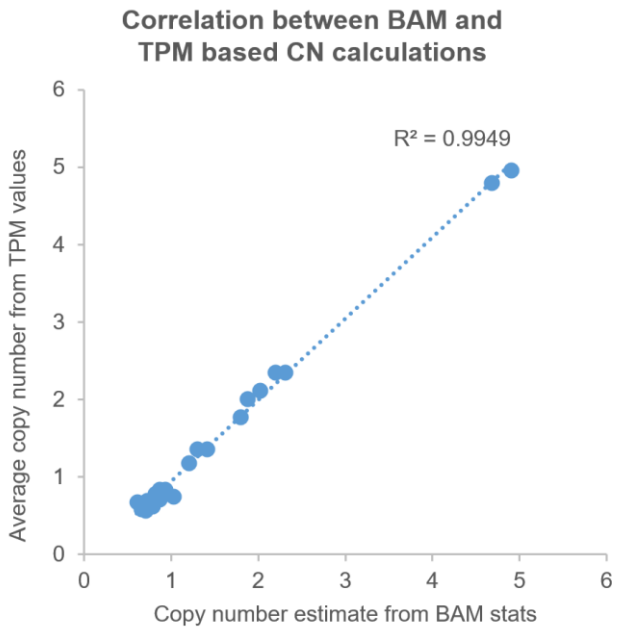
S 7

S 7: Copy number estimates for each transformant. Copy number was calculated in two different ways: 1) as relative coverage from SAMtools (BAM stats) calculated statistics; 2) by averaging the 'transcript per million (TPM) values for eGFP, Zeocin Resistance and PUC-origin sequences. In table A, summary values from supplementary file S 7 are shown and in panel B the BAM stats and TPM obtained CNs are correlated, demonstrating excellent agreement (R^2=0.99). Raw data and calculations are shown in the supplementary Excel file S 7.
TPM represents Salmon's estimate of the relative abundance of this transcript (in units of transcripts per million) and is the recommended relative abundance measure to use for downstream analysis. TPM gives the number of transcripts of a given gene if there were one million genes. It therefore accounts for gene length and library size.
Given that we are looking at genomic reads rather than RNAseq data, we would theoretically expect each of the 5332 *P. pastoris* genes in the transcript file to have the same TPM of just under 200 (there are 5332 genes in the output and 1,000,000/5332 is 187.5). Thus assuming even coverage, one copy will be represented by a TPM of 187.5.

A

Line	Copy number SAMtools (BAM stats)	Average copy number from TPM values	Rounded copy number
QTV76	0.6058	0.6768	1
QTV77	0.6608	0.5824	1
QTV78	0.7201	0.6951	1
QTV79	0.7010	0.5707	1
QTV80	0.8522	0.7346	1
QTV81	0.9243	0.8395	1
QTV82	1.8054	1.7745	2
QTV83	1.4162	1.3572	1
QTV84	1.2110	1.1716	1
QTV85	0.7803	0.6157	1
QTV86	1.0203	0.7384	1
QTV87	0.8369	0.7783	1
QTV88	0.8240	0.7028	1
QTV89	0.8657	0.7019	1
QTV90	0.8637	0.7428	1
QTV91	0.8371	0.7704	1
QTV92	1.8785	2.0084	2
QTV93	2.0182	2.1124	2
QTV94	2.2047	2.3392	2
QTV95	2.3142	2.3378	2
QTV96	4.6788	4.7903	5
QTV97	4.9057	4.9626	5
QTV98	0.8635	0.8346	1
QTV99	0.8234	0.7851	1
QTV100	1.3011	1.3548	1

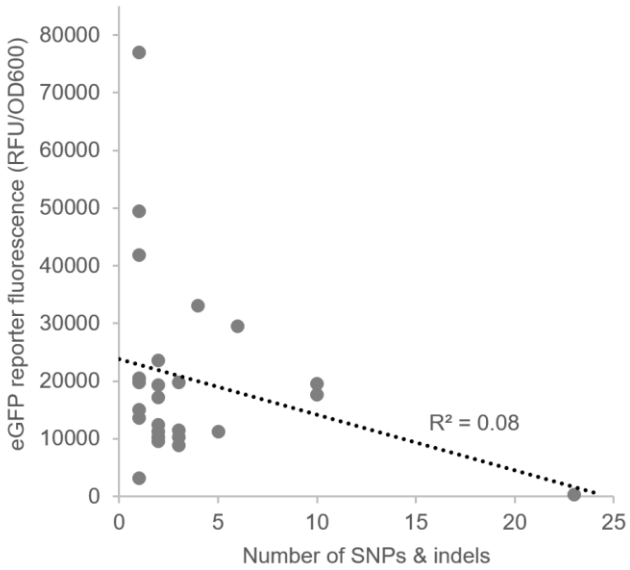
B



S 8

S 8: Raw data on SNP identification after filtering provided as supplementary file (.xlsx table format, each sheet represents a different strain). Each sheet shows the commands used in Bcftools 1.3.1 (Li H, Bioinformatics 27:2987–93, 2011) to filter the list of total ariants identified, the exact SNPS and indels retained after filtering and their effects, when in exons (or very close to exons) as determined using snpEff 4.3p and a custom-built database.

S 9



S 9: The number of sequence variants does not correlate with reporter protein fluorescence. Correlation between sequence variants (SNPs, indels in all four chromosomes and mitochondria, see Tab. 1 for a summary and S 8 for raw data) and eGFP reporter protein fluorescence (normalized per OD600, as obtained from the rescreening and shown in Fig. 3).

S 10

S 10: Extended discussion on effects of GUT1/STD vectors and linearization by *Swa*I/*Sac*I.

In Fig. 3B, QVT86-88 and QVT89-91 are categorized into average (low) and average (high) groups, respectively, showing different expression although their copy numbers are almost the same (S 7).

We noticed this phenomenon when performing the screening/rescreening (Fig. 2, S 1; Fig. 3, S 3). For construct GUT1, linearized with *Sac*I, some of the clones from the middle of the landscape (what we referred to in all other cases as 'average') seemed to represent different populations in the rescreening. Hence we termed them 'average (low)' and 'average (high)', as these two groups had near identical copy number estimates and seemed to perform within each group uniformly (as opposed to high expressers QTV92,93,94 that showed with two copies each clearly elevated expression, S 7). We could only resolve the integration locus of one (QTV86) out of the six strains (QTV86 was found to be correctly integrated to the *AOX1* locus). The other five strains proved to be evasive for identifying the integration site.

This phenomenon is puzzling and we could not come up with a clear mechanistic explanation. We hypothesize that this issue may have to do with the presence of the *GUT1* integration sequences. If it had to do with the linearization (*Swa*I/*Sac*I) and specifically some issue with the *Sac*I targeted integration, one would expect to see the same phenomenon also with the STD plasmid, that was also linearized with *Sac*I (and not only the GUT1 plasmid). But for the STD plasmid, we noticed only one population of average clones. It may be that this effect only occurs in the combination of *GUT1* integration sequence being present and *Sac*I digestion. This might be related to the phenomenon of linearization of the same plasmid (GUT1) with two different restriction endonucleases yielding different expression medians in the screening (Fig. 2C), as discussed in the main manuscript in section 'Effect of plasmid design, vector linearization and type of integration event'.

Maybe a similar proposed effect of the *GUT1* sequence also influences the average low/high phenomenon. A previous report (Schwarzhans J-P, Wibberg D, Winkler A, Luttermann T, Kalinowski J, Friehs K. Microb Cell Fact;15:84, 2016) described recombination events where the *AOX1* terminator recombined with the 3' *AOX1* homologous region, leading to a loss of the gene of interest. It may be that in our setting the *GUT1/AOX1* sequences may recombine in a similar fashion, resulting in a maintenance/loss of the *GUT1* sequence possibly affecting *AOX1* expression. In detail, considering Figure 1A, lower illustration, imagining that only the pAOX1 3' region recombines – then the pAOX1 sequence would not be in proximity with the possibly repressing *GUT1* sequence. But if also the pAOX1 5' region recombines, the *GUT1* region would be adjacent to the pAOX1, as if linearizing with *Swa*I. This notion is supported by the GUT1-*Swa*I linearized average transformants [that always have the *GUT1* integration sequence 5' of the *AOX1* promoter] rather matching the GUT1-*Sac*I average (low) clones [and QTV86, as an 'average (low)' strain being correctly integrated to the *AOX1* locus], whereas the STD-*Sac*I linearized clones [where inherently no GUT1 sequence is present on the vector] rather match the GUT1-*Sac*I average (high) clones]. However, we cannot prove this theory, as the Illumina reads yielded inconclusive results or are too short to cover these extensive sequences/integration events. Future studies with technologies providing longer read lengths (SMRT/Pacbio) may help to resolve these issues.