Synthetic core promoters for *Pichia pastoris*

Supporting Information

Supporting information

Methods

Chemicals and media

Oligonucleotides were ordered from Integrated DNA Technologies (Leuven, Belgium), see Supplementary Table 1 for the sequences. Plasmid isolations were performed using a GeneJET Plasmid Miniprep Kit by Thermo Fisher Scientific (Waltham, MA, USA). Gel fragments, PCRs and restriction digests were purified using a Wizard SV Gel and PCR Clean-Up System by Promega (Fitchburg, WI, USA). Restriction enzymes, Phusion DNA Polymerase and other DNA modifying enzymes were purchased from Thermo Fisher Scientific. Plasmids were sequenced using Sanger sequencing services by LGC Genomics GmbH, Berlin. Miscellaneous chemicals were purchased from Fresenius Kabi Austria (Graz, Austria), Carl Roth (Karlsruhe, Germany), Becton, Dickinson and Company (Franklin Lakes, NJ, USA).

P. pastoris strains were grown on full medium (yeast extract, peptone, 2% glucose (YPD)), buffered minimal dextrose with 1% glucose (BMD1) and buffered minimal methanol medium with 0,5% methanol (BMM) as described by Weis *et al*. [16]. *Escherichia coli* strains were grown on standard LB-medium containing 25 µg/ml Zeocin (Life Technologies, Carlsbad, CA, USA). For transformations and re-streaking *P. pastoris* selection was performed using YPD agar plates containing 100 µg/ml Zeocin.

Strains and plasmids

In this study, the *P. pastoris* CBS7435 wildtype strain was used, cloning work was performed using an *E. coli* Top10 F' strain. The expression plasmid used in this study is based on the pPpT4_S *P. pastoris*/*E. coli* shuttle vector (GenBank accession number JQ519690.1) reported by Näätsaari *et al*. [17]. A plasmid map is shown in S 3 and the plasmid sequences are provided in Supporting File 2 in GenBank format. At first the gene coding for an enhanced variant of green fluorescent protein (here termed eGFP, containing mutations F64L, S65T, F199S, M153T, V163A [18] resulting in a higher quantum yield and shorter folding time) was cloned into pPpT4_S resulting in pPpT4_S-eGFP. Therefore the eGFP gene was PCR amplified using primers eGFPfwdEcoRIKozak and eGFPrevNotI (Supplementary Table 1). pPpT4_S and the eGFP PCR product were *Eco*RI and *Not*I digested, ligated and verified by sequencing. This vector contains already the wildtype *AOX1* promoter. However, the pCore*AOX1* variants we aimed to compare were also fused to the upstream p*AOX1* region. Therefore they could only be distinguished by sequencing from the wildtype *AOX1* promoter already present in the vector. To circumvent this issue, we designed at first an entry vector with a truncated version of p*AOX1*, thereby correct insertion of a synthetic promoter could be easily checked by the plasmid size after linearization. The pPpT4_S vector contains also only a blunt end restriction enzyme site (*Swa*I) upstream of the promoter. As sticky end ligations work more efficiently than blunt end ligations, we added a *Bgl*II site downstream of the *Swa*I site. This entry vector facilitating cloning was termed pPpT4_SB_truncatedpAOX1-eGFP. The vector was obtained by PCR amplifying the truncated p*AOX1* using primers pAOX1SwaIBglIIfwd and pAOX1truncatedEcoRIrev and subsequent *Swa*I and *Not*I digestion of the PCR product and the vector pPpT4_S-eGFP. After ligation, pPpT4_SB_truncatedpAOX1-eGFP was confirmed by sequencing. Subsequently the synthetic core promoter variants Sync1 to Sync6ins, pCore1, pCore11 and the wildtype core promoter pCore*AOX1* were cloned into this vector via the *Bgl*II and *Eco*RI sites.

Concerning the synthetic core promoters pCore1, pCore11 and variants Sync1 to Sync6ins, the multiple sequence alignments shown were obtained using MultAlin (applying standard settings) [12]. TFBS predictions were performed using MatInspector [13,19] using the software and settings previously applied by Hartner *et al*. [2].

The synthetic variants were ordered as long primers Sync1_rv to Sync6_ins_rv and pCore1_rv (see Supplementary Table 1) and attached by PCR to the upstream sequence of p*AOX1* using the forward primer pAOX1SwaIBglIIfwd.
pCore11 was added to the upstream p*AOX1* region by overlap-extension PCR (olePCR). The pCore11 fragment was amplified using primers pCore11linkerFWDolePCR and pCore11linkerREVolePCR, the p*AOX1* upstream region was amplified using the plasmid pPpT4_S as a template and primers pAOX1SwaIBglIIfwd and pAOX1linkerREVolePCR. The two fragments were gel-purified, mixed in an equimolar ratio and used for olePCR using primers pAOX1SwaIBglIIfwd and pCore11linkerREVolePCR. The wildtype promoter (pCore*AOX1*) was amplified using primers pAOX1SwaIBglIIfwd and pAOX1wtEcoRIrev. All promoter PCR products and pPpT4_SB_truncatedpAOX1-eGFP were digested with *Bgl*II and *Eco*RI, ligated and verified by sequencing. The long primers contained occasionally point mutations, requiring sequencing of multiple clones.

Transformation, cultivation conditions and fluorescence measurements

Transformations of *P. pastoris* were performed using the condensed protocol of Lin-Cereghino *et al*. [20]. To avoid biasing of the results by transformant to transformant variability, we transformed only low amounts of the *Swa*I-linearized expression cassettes (0.5 to 1 µg of DNA) to avoid multi copy integration and screened subsequently a landscape of more than 80 transformants in a high-throughput deep well plate system [16]. In short, the cells were grown for 60h on 250µl BMD1 and subsequently induced with methanol (250µl BMM2 [1% methanol] at 60h and 50µl BMM10 [5% methanol] at 72h). We checked the landscape for uniformity and rescreened three transformants from the linear range of the landscape for uniform, reproducible expression (assayed by eGFP fluorescence). One representative transformant per construct was confirmed to be single copy by qPCR (as described by Abad *et al*. [21], data not shown) and then used for the further work and comparison of the variants under the same growth conditions. Mean values and standard deviations shown Fig. 2 B represent biological replicates from four-fold cultivations of the same transformant.
eGFP fluorescence (excitation at 488 nm and emission at 507 nm) and absorption at 600 nm (OD600, optical density 600) were measured in micro titer plates (Nunc MicroWell 96-Well Optical-Bottom Plates with Polymer Base, Black; Thermo Fisher Scientific) using a Synergy MX plate reader (Biotek, Winooski, VT, USA). 10 µl from the cultures were diluted with 190 µl of water to be within the linear range of the absorption readings of the plate reader. For both, eGFP fluorescence and OD600, the background measurements of diluted medium were subtracted. Subsequently the relative fluorescence units were normalized per OD600.

Supplementary Table 1

Supplementary Table 1: Primer sequences used in this study

Name	Sequence (5'→3')
eGFPfwdEcoRIKozak	gcccgaattccgaaacgatggctagcaaaggagaagaacttttc
eGFPrevNotI	gatcgcggccgcttacttgtacaattcatccatgccatg
pAOX1SwaIBglIIfwd	tgacatttaaatagatctaacatccaaagacgaaaggttgaatgaaac
pAOX1truncatedEcoRIrev	actggaattcgtcagttttgggccatttggggaac
Sync1_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattatgaa agtaagctaataatgatgataaaaaaaaaaggtttaaaaagaggaaaggtgcttttgtattatatatgctgtcaagtaggggttagaacag
Sync2_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattatgaa agtaagctaataatgatgatcgaaaagaaaggagaaaaagaggaagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync3_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattttcttg agactgattagattatgatcgaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync4_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtccaatttccattacttctggtttcttgag acgctaataatgatgataaaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync5_rv	tttgaattctttcaataattagttgttttttgatcttctctagtaagattgtatggtcttaagttgtgaaagcttgtcaattggaaccagtcgcaattatgaaa gtaagctaataatgatgataaaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync6_rv	tttgaattctttgaagtttgatatatctagtaagagttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattatgaaagtaagctaat aatgatgataaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync3_ins_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattttgaga ctgattagattatgatgacgaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync4_ins_rv	tttgaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgttaaaatcaaaagcttgtccattacttctggtttcaattttgagact gagctaataatgatgataaaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
Sync6_ins_rv	tttgaattctttcaatttgaagtttgatatatctagtaacaagttgtcgttaaaagtcgttaaaatcaaaagcttgtcaattggaaccagtcgcaattatga aagtaagctaataatgatgataaaaaaaaaaggtttaagacagggcagcttccttctgtttatatattgctgtcaagtaggggttagaacag
pCore1_rv	gtttcggaattctttgaagtttgacattgtttttagttgttataagattgtatgcgtcttaagttgtattaatatttatcttgtgagaagaaaccagaagtaat ggaaattgaggtgaaagtgatgatgaaaaaaaaggagaaaaagaggaaaggtgcttttgttttatatagatatgttggtcgtgctcaaaacgttag
pCore11linkerFWDolePCR	gattctggtgggaatactgctgatagcctaacgttttgagcacgaccaacacatc
pCore11linkerREVolePCR	ttcggaattctttgaagtttgatatatctagtaagattgtatg
pAOX1linkerREVolePCR	gtcgtgctcaaaacgttaggctatcagcagtattcccaccagaatc
pAOX1wtEcoRIrev	ttcggaattctttcaataattagttgttttttgatcttctcaagttgtcgttaaaagtcgtt

Supplementary Figures

S 1

S 1: Detailed description of the design of the synthetic promoters pCore1 and pCore11.

- General outline and sequence selection

  We aimed to design a controllable, generally applicable synthetic *P. pastoris* core promoter able to initiate transcription and suitable for the attachment of URS. Therefore, the core promoter sequences of four differently regulated promoters were aligned for consensus finding. The core promoter sequence is usually defined between +40/-40 in relation to the transcription start site (TSS) [22], but was in this study extended to approximately +110/-70 according to the previously applied p*AOX1* core promoter sequence [2,23]. This sequence selection and synthetic promoter design includes also the 5'UTR, as it is challenging to determine the TSS and separate the core promoter from the 5' UTR *in silico.*
- Information on the natural core promoters used for the alignment

  Core promoter sequences of the following promoters were used for the alignment: p*AOX1,* p*GAP*, p*HIS4* and p*ScADH2*. p*AOX1,* p*GAP* (promoter of the glyceraldehyde-3-phosphate dehydrogenase gene) and p*HIS4* (promoter of a gene involved in histidine biosynthesis) are *P. pastoris* endogenous promoters.

  p*AOX1* is repressed on glucose, glycerol and ethanol and strongly induced on methanol. p*GAP* is the promoter of a constitutively expressed housekeeping gene, although there is an influence of the carbon source [9,24]. p*ScADH2* is the promoter from the *S. cerevisiae alcohol dehydrogenase 2* gene that was found to be also functional in *P. pastoris* providing sufficient expression of a resistance marker gene [25]. The *ADH2* promoter from *S. cerevisiae* is repressed in the presence and induced in the absence of glucose. To compensate the shorter length of pCore*GAP*, the most 5' and 3' sequences of the core promoter were aligned independently. Therefore the top row of Fig. 1 B shows an alignment of the 5', the bottom row of the 3' regions of the respective core promoters.
- Detailed information on pCore1 and pCore11 design
  - pCore1

    The alignment of the four natural promoters resulted in the finding of two roughly defined consensus boxes located between -18/-48 and +33/+42 relative to the p*AOX1* TSS (-113 from ATG). For the other sequence parts a 50% identity rule was applied. Thus, if two out of four bases were similar the resulting consensus was used for the design of pCore1. Sequence gaps without consensus were filled randomly. Finally, pCore1 was 62% identical to pCore*AOX1.*
  - pCore11

    The four natural promoters used for the initial consensus design and pCore1 were analyzed regarding putative TFBSs using the program MatInspector (Genomatix Software Inc., USA) [13]. All TFBSs found in pCore*AOX1*, pCore*GAP*, pCore*HIS4*, pCore*ScADH2* and pCore1 were grouped by their number of occurrence (see Supporting file 1 for the raw data and a summary table). Predicted TFBSs appearing in at least two natural core promoters were manually incorporated into pCore1 and superfluous sites removed, resulting in pCore11. To the best of our knowledge there is no algorithm available that would allow automation of this step.

    For the manual design of the TFBSs into pCore11, various aspects were taken into consideration:

    a) If the TFBSs predicted in several natural promoters were already present in pCore1, TFBSs predicted only in one of the natural promoters were omitted from the analysis.
    b) TFBSs predicted only in pCore1 were removed or replaced by incorporating TFBSs common between natural core promoters.
    c) The position of the predicted TFBSs in the natural promoters.
    d) Conserved regions between all four natural promoters in the initial consensus sequence design (e.g. the conserved TATA box and short downstream regions) were not changed (see Fig. 1 C).

    When incorporating additional TFBSs, we changed the sequence of pCore1 thereby removing several overlapping TFBSs only predicted in pCore1. Thereby also short deletions were made. Several TFBSs present in the natural promoters had already been present in pCore1 (see the summary table in Supporting file 1). The final difference between pCore1 and pCore11 was mostly the removal of superfluous TFBSs of pCore1 and providing the predicted TFBSs appearing in natural promoters. All TFBSs appearing in 3 or 4 natural promoters are present in pCore11. 4 out of 6 TFBSs predicted in two core promoters are also present (see Supporting File 1). Comparing pCore11 to pCore*AOX1* an identity of 54% was found. For both, pCore1 and pCore11, the transcription start site was adopted from p*AOX1* (ATCA). A comparison of pCore1 and pCore11 is shown in Fig. 1 C.

S 2

S 2: Detailed information on Sync1 to Sync6ins design and sequences of the core promoter variants Sync1 to Sync6ins.

(A) Detailed description of the design of the Sync variants.

We focused on the core promoter region between the core promoter hallmarks TATA box and transcriptional start site (TTS). In this region we incorporated overlapping regions of the pCore11 sequence (24 to 30 bp) into pCore*AOX1* (Sync1 to 4, for synthetic core promoter). Considering possible effects from different spacing (as pCore11 and pCore*AOX1* are of different lengths), Sync3ins and Sync4ins were designed by inserting sequences to fill up gaps in an alignment of pCore11 and pCore*AOX1* (S 2 B below). Namely Sync3ins fills the gap at position 67 and 68 in the alignment shown in S 2 B, Sync4ins fills the gap between positions 91 to 94. The relevant positions are highlighted in yellow in the sequence comparison of all Sync constructs (S 2 C). In addition we tested with Sync5, Sync6 and Sync6ins also possible regulatory regions in the 5'UTR by a similar approach. Accordingly, effects on expression levels might be caused either by direct effects on the core promoter and therefore transcription or by changes in the 5' UTR effecting mRNA stability or translation initiation. This was in consent with our main goal of creating a diversified library of different expression levels by engineering the core promoter/5'UTR region. Sync6ins was included to provide a different spacing and length of the downstream 5'UTR region, as pCore11 is shorter than pCore*AOX1* (S 2 C).

(B) Alignment of pCore11 with pCoreAOX1

The alignment was created using MultAlin [12], the compared sequences reach exactly up to the *Eco*RI site shown in S 2 C.

```
           1       10        20        30        40        50        60        70        80        90        100       110       120       130       140       150       160       170      179
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--------|
pCore11    TTGAGCACGACCA-ACACATCTATATAAAACAAAAGCACCTTTCCTCTTTTTCTCCTTTCTTTTCG--ATCATAATCTAATCAGTCTCAA----GAAACCAGAAGTAATGGAAATTGAGGTGAAAGTCACAACTTAAGACCATACAATCTTACTAGATATATCAAACTTCAAA
pCoreAOX1  CTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAA
```

(C) Sequences of the core promoters used for the Sync studies.

Promoter hallmarks such as the TATA Box, transcriptional start site (TSS, "m" denotes minor, "M" major TSS as determined by [26]) are denoted in the wildtype promoter sequence pCore*AOX1*. Exchanged regions between pCore*AOX1* and pCore11 are highlighted in color; see S 2 A,B for more detailed information. The *Eco*RI site used for cloning is underlined, the start codon is shown in bold face lowercase. The first 25 bp of the sequences were used as a linker to add the p*AOX1* upstream regulatory region (see method section for detailed cloning procedure). The yellow regions in Sync3ins and Sync4ins denote sequences to fill up gaps in an alignment of pCore11 and pCore*AOX1* to maintain the same spacing (see S 2 A for details).

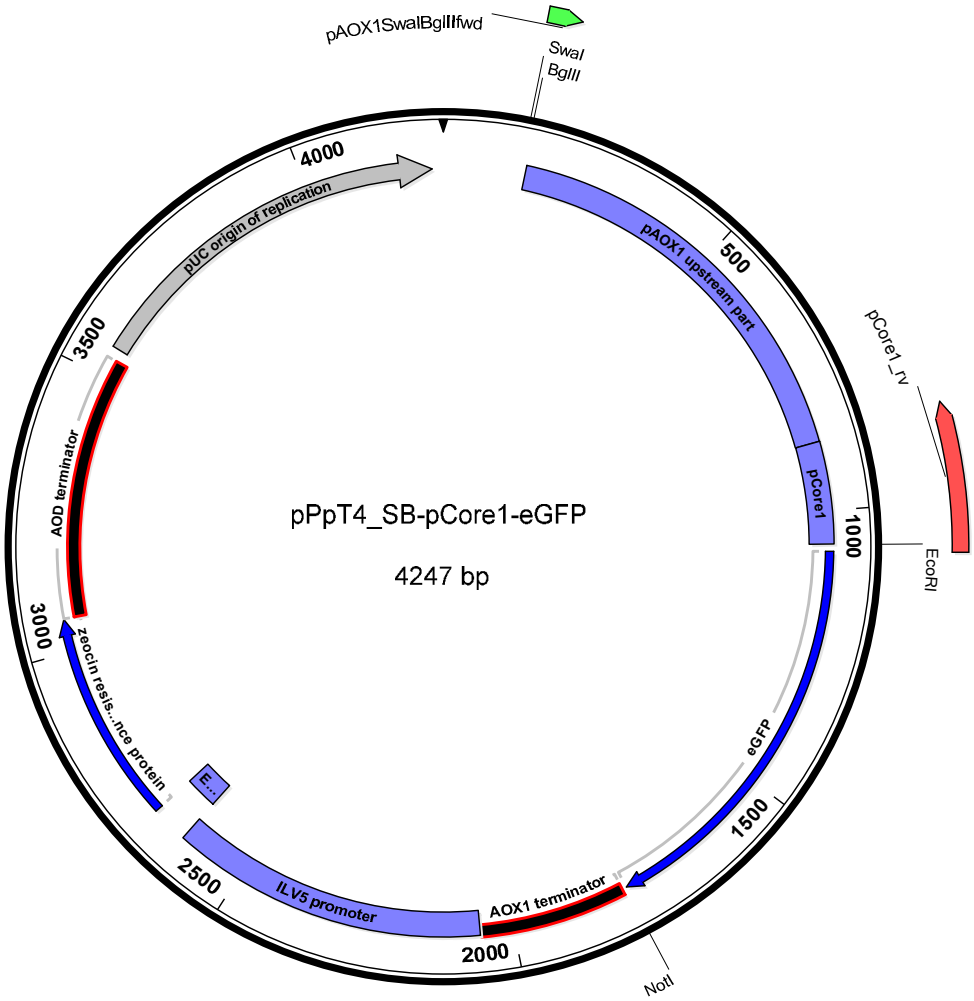
```
                                                                   TSS
          -197                    TATA Box                       m  M  m                                                                                                       EcoRI      -1
pCoreAOX1 CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg

Sync1     CTGTTCTAACCCCTACTTGACAGCATATATAATACAAAAGCACCTTTCCTCTTTTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync2     CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTTCCTCTTTTTCTCCTTTCTTTTCGATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync3     CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTCGATCATAATCTAATCAGTCTCAAGAAAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync4     CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCGTCTCAAGAAACCAGAAGTAATGGAAATTGGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync5     CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTCACAACTTAAGACCATACAATCTTACTAGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync6     CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACTCTTACTAGATATATCAAACTTCAAA            GAATTCCGAAACGatg
Sync3ins  CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTCGTCATCATAATCTAATCAGTCTCAAAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync4ins  CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTCAGTCTCAAAATTGAAACCAGAAGTAATGGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGAGAAGATCAAAAAACAACTAATTATTGAAAGAATTCCGAAACGatg
Sync6ins  CTGTTCTAACCCCTACTTGACAGCAATATATAAACAGAAGGAAGCTGCCCTGTCTTAAACCTTTTTTTTTATCATCATTATTAGCTTACTTTCATAATTGCGACTGGTTCCAATTGACAAGCTTTTGATTTTAACGACTTTTAACGACAACTTGTTACTAGATATATCAAACTTCAAATTGAAAGAATTCCGAAACGatg

             -179                Sync1                                        Sync3                                                                                                -1
   pCore11   TTGAGCACGACCAACACATCTATATAAAACAAAAGCACCTTTCCTCTTTTTCTCCTTTCTTTTCGATCATAATCTAATCAGTCTCAAGAAACCAGAAGTAATGGAAATTGAGGTGAAAGTCACAACTTAAGACCATACAATCTTACTAGATATATCAAACTTCAAAGAATTCCGAAACGatg
                                                     Sync2                                Sync4                                   Sync5                Sync6
```

S 3

S 3: Map of a representative plasmid used in this study. The map of pPpT4_SB-pCore1-eGFP is shown; the synthetic core promoter pCore1 is fused to the upstream part of p*AOX1*. The restriction sites used for cloning the promoter PCR fragment (*Bgl*II and *Eco*RI) and the restriction sites for initially cloning the eGFP reporter gene (*Eco*RI and *Not*I) are shown. Also the primers used for linking pCore1 to the upstream part of p*AOX1* are depicted. The detailed explanation of the cloning strategy is provided in the Materials and Methods section (Strains and plasmids). The sequences of all vectors used in this study are provided in Supporting file 2.



Supplementary Note

The identification of TFBSs in p*AOX1* by DNA/protein interaction studies mentioned in the manuscript focused on the methanol master regulator Mxr1. Lin Cereghino *et al*. [27] identified an approximate binding region. Kranthi *et al*. [28] mapped the exact binding sites and determined a consensus sequence that was later refined with Mxr1 binding results from additional methanol inducible promoters of *P. pastoris* (p*DAS2* and p*PEX8*) [29].

Description of Supporting file 1

Excel file providing the MatInspector predictions of TFBSs, contains two spread sheets:

- Spread sheet 1: Detailed MatInspector matches
  The core promoters used or designed in this study (pCore*AOX1*, pCore*GAP*, pCore*HIS4*, pCore*ScADH2,* pCore1 and pCore11) were analyzed for predicted TFBSs. The excel file contains lists of the TFBSs predicted in each of these promoters. All information available from MatInspector, including graphical illustrations of the binding sites (at the very end of the list) is provided.
  Note that the color code shown in this sheet is unique for each core promoter analyzed. So TFBSs marked in the same color in different core promoters are not necessarily indicating the same TFBSs. See the second sheet "Summary of common TFBSs" for simplified coloring comparing the number of predicted TFBSs in the promoters.
  "Matrix sim." denotes the matrix similarity, a measurement for the identity of the TFBS matrix and the sequence found in the promoter. "Sequence" is the match of the matrix in the promoter sequence, capital letters denote the core sequence of the motif, letters in red are highly conserved. For more detailed information see [13,19].
- Spread sheet 2: Summary of common TFBSs
  Color coded frequencies of TFBSs predicted in the promoters used in this study. For the design of pCore11, all TFBSs found in pCore*AOX1*, pCore*GAP*, pCore*HIS4*, pCore*ScADH2* and pCore1 were grouped by their number of occurrence (see the legend on top). The final predicted sites in pCore11 are also included.
  On the left side of the spreadsheet, the TFBS predictions of all promoters are provided with the color code of the legend.
  On the right side, a separate table is concisely summarizing/comparing the occurrence of the TFBSs in all promoters.

Description of Supporting file 2

Sequences of the plasmids used in this study in Gene bank format. The plasmids with the different core promoters and the entry vector facilitating cloning are provided. All plasmids are based on pPpT4_S *P. pastoris*/*E. coli* shuttle vector (GenBank accession numbner JQ519690.1) reported by Näätsaari *et al*. [17], the map of this vector is also included. Also the primers used for cloning are shown in the maps.

Supporting references

The supporting references do not appear in the main manuscript text and the main reference list. Yet the numbering of the additional references was continued from the main reference list to avoid confusion by double numbers when citing references from the main list in the supporting information.

(16) Weis, R., Luiten, R., Skranc, W., Schwab, H., Wubbolts, M., and Glieder, A. (2004) Reliable high-throughput screening with Pichia pastoris by limiting yeast cell death phenomena. *FEMS Yeast Res. 5*, 179–89.

(17) Näätsaari, L., Mistlberger, B., Ruth, C., Hajek, T., Hartner, F. S., and Glieder, A. (2012) Deletion of the Pichia pastoris KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. *PLoS One 7*, e39720.

(18) Waldo, G. S., Standish, B. M., Berendzen, J., and Terwilliger, T. C. (1999) Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol. 17*, 691–5.

(19) Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics 21*, 2933–42.

(20) Lin-Cereghino, J., Wong, W. W., Xiong, S., Giang, W., Luong, L. T., Vu, J., Johnson, S. D., and Lin-Cereghino, G. P. (2005) Condensed protocol for competent cell preparation and transformation of the methylotrophic yeast Pichia pastoris. *Biotechniques 38*, 44, 46, 48.

(21) Abad, S., Kitz, K., Hörmann, A., Schreiner, U., Hartner, F. S., and Glieder, A. (2010) Real-time PCR-based determination of gene copy numbers in Pichia pastoris. *Biotechnol. J. 5*, 413–20.

(22) Juven-Gershon, T., and Kadonaga, J. T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol. 339*, 225–9.

(23) Ruth, C., Zuellig, T., Mellitzer, A., Weis, R., Looser, V., Kovar, K., and Glieder, A. (2010) Variable production windows for porcine trypsinogen employing synthetic inducible promoter variants in Pichia pastoris. *Syst. Biol. Synth. Biol. 4*, 181–91.

(24) Waterham, H. R., Digan, M. E., Koutz, P. J., Lair, S. V, and Cregg, J. M. (1997) Isolation of the Pichia pastoris glyceraldehyde-3-phosphate dehydrogenase gene and regulation and use of its promoter. *Gene 186*, 37–44.

(25) Abad, S., Nahalka, J., Bergler, G., Arnold, S. A., Speight, R., Fotheringham, I., Nidetzky, B., and Glieder, A. (2010) Stepwise engineering of a Pichia pastoris D-amino acid oxidase whole cell catalyst. *Microb. Cell Fact. 9*, 24.

(26) Ellis, S. B., Brust, P. F., Koutz, P. J., Waters, A. F., Harpold, M. M., and Gingeras, T. R. (1985) Isolation of alcohol oxidase and two other methanol regulatable genes from the yeast Pichia pastoris. *Mol. Cell. Biol. 5*, 1111–21.

(27) Lin-Cereghino, G. P., Godfrey, L., de la Cruz, B. J., Johnson, S., Khuongsathiene, S., Tolstorukov, I., Yan, M., Lin-Cereghino, J., Veenhuis, M., Subramani, S., and Cregg, J. M. (2006) Mxr1p, a key

regulator of the methanol utilization pathway and peroxisomal genes in Pichia pastoris. *Mol. Cell. Biol. 26*, 883–97.

(28) Kranthi, B. V., Kumar, R., Kumar, N. V., Rao, D. N., and Rangarajan, P. N. (2009) Identification of key DNA elements involved in promoter recognition by Mxr1p, a master regulator of methanol utilization pathway in Pichia pastoris. *Biochim. Biophys. Acta 1789*, 460–8.

(29) Kranthi, B. V., Kumar, H. R. V., and Rangarajan, P. N. (2010) Identification of Mxr1p-binding sites in the promoters of genes encoding dihydroxyacetone synthase and peroxin 8 of the methylotrophic yeast Pichia pastoris. *Yeast 27*, 705–11.