Supporting Information 1

Machine learning of designed translational control allows predictive pathway optimisation in *Escherichia coli*

Adrian J. Jervis[†], Pablo Carbonell[†], Maria Vinaixa[†], Mark Dunstan[†], Katherine Hollywood[†], Nicholas Rattray[‡], Christopher J. Robinson[†], Cunyu Yan[†], Neil Swainston[†], Andrew Currin[†], Rehana Sung[†], Helen Toogood[†], Sandra Taylor[†], Jean-Loup Faulon[†§], Rainer Breitling[†], Eriko Takano[†], Nigel S. Scrutton[†*]

Contents

S1. *In vivo* monoterpenoid screening pipeline

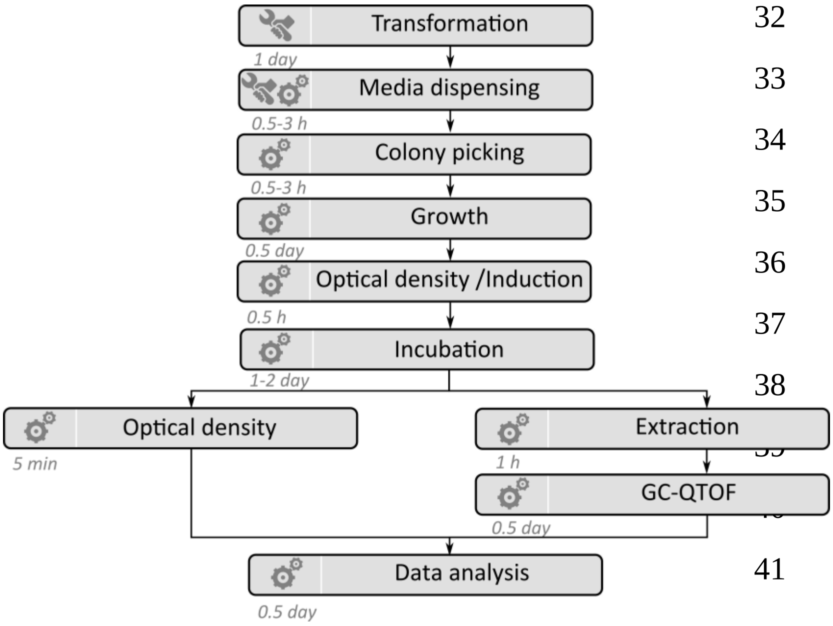

Figure S1. Schematic of the *in vivo* monoterpenoid screening pipeline for *E. coli*. The pipeline uses 96-well plates and is designed to employ robotic liquid handling platforms to process samples simultaneously. Symbols indicate steps which are carried out manually (hand/spanner) or automated (cogs). Approximate time requirements for each step are also indicated. Typically 3 plates (288 samples) can be processed per day, per robot.

S2. Identification of optimal inducer concentrations

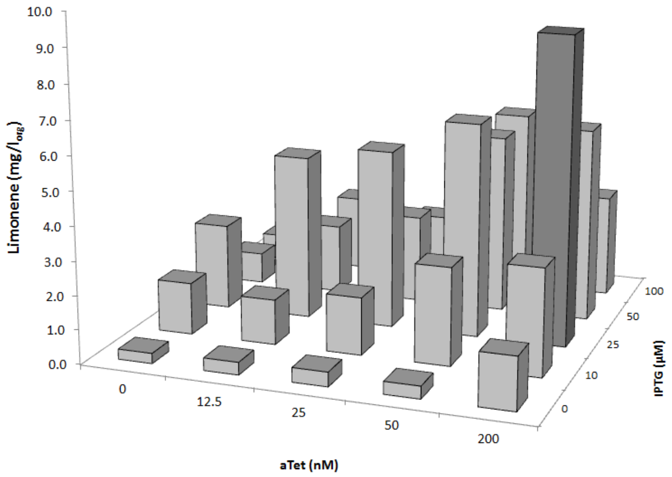

Figure S2. Identification of optimal of inducer concentrations for limonene production during plate-based fermentation. *Escherichia coli* DH10β cells transformed with plasmids pMVA and pGL were grown in a 96-deepwell block in a biphasic media as described in the Methods section. Triplicate wells were induced with different concentrations of the inducers Isopropyl-β-D-thiogalactoside (IPTG) and anhydro-tetracycline (aTet). Titres are the average of 3 biological replicates and are displayed as the concentration in the organic phase.

S3. Re-screening of pGLlib high producers

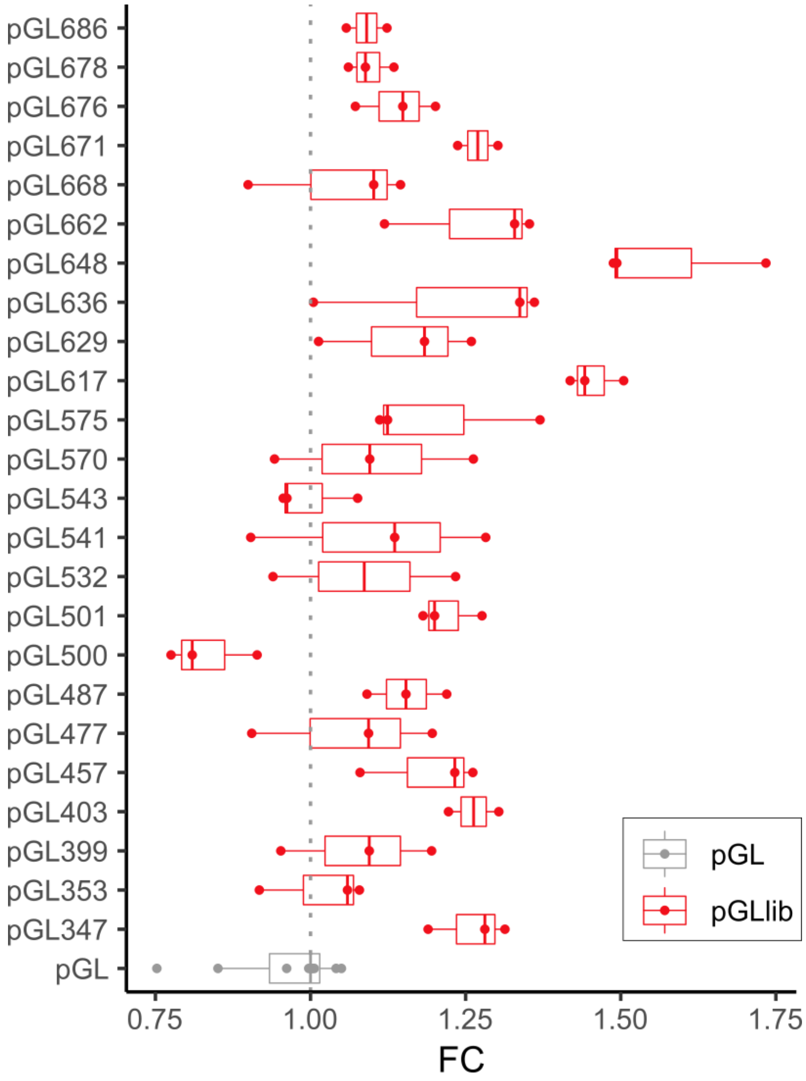

Figure S3. Re-screening of the high-producing variants from pGLlib. Plasmid DNA from the top 24 producers in the pGLlib screen (Fig. 1) was purified and co-transformed with pMVA into fresh cells for testing in triplicate. The median fold-change in limonene production and upper/lower quartiles are indicated.

S4. Comparison of RBS sequences for pGLlib high/low producers

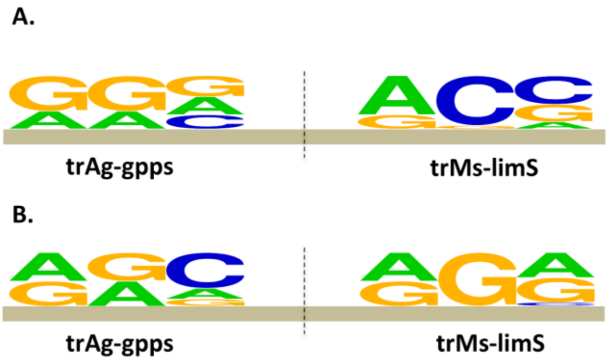

Figure S4. Comparison of the frequency of the variable nucleotides of RBSs in library pGLlib. The sequence patterns are displayed for the A) low producers (bottom quartile) and B) high producers (upper quartile) for the two genes *trAg-gpps* and *trMs-limS*. Sequence logos were generated through the weblogo service (weblogo.berkeley.edu).

S5. Validation of machine learning for pGLlib

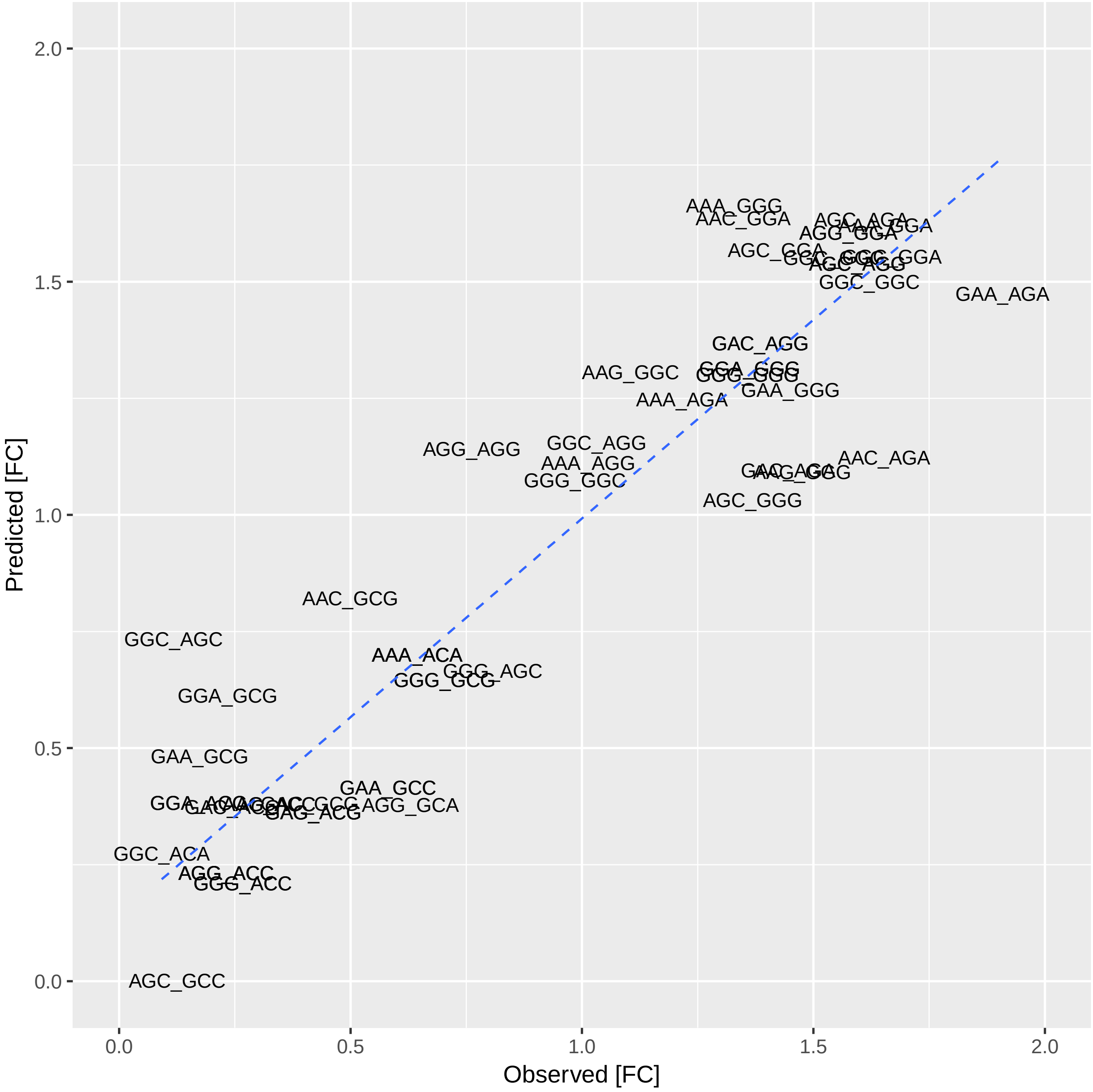

Figure S5. Validation of the machine learning model for library pGLlib. The observed levels of limonene production (x- axis) are compared with predicted levels (y- axis) in fold change (FC) for each RBS combination using leave-one-out cross-validation with a resulting performance of $Q^2 = 0.87$. The blue dashed line is the linear fitting of the points.

S6. Genetic organisation of the pMVA series

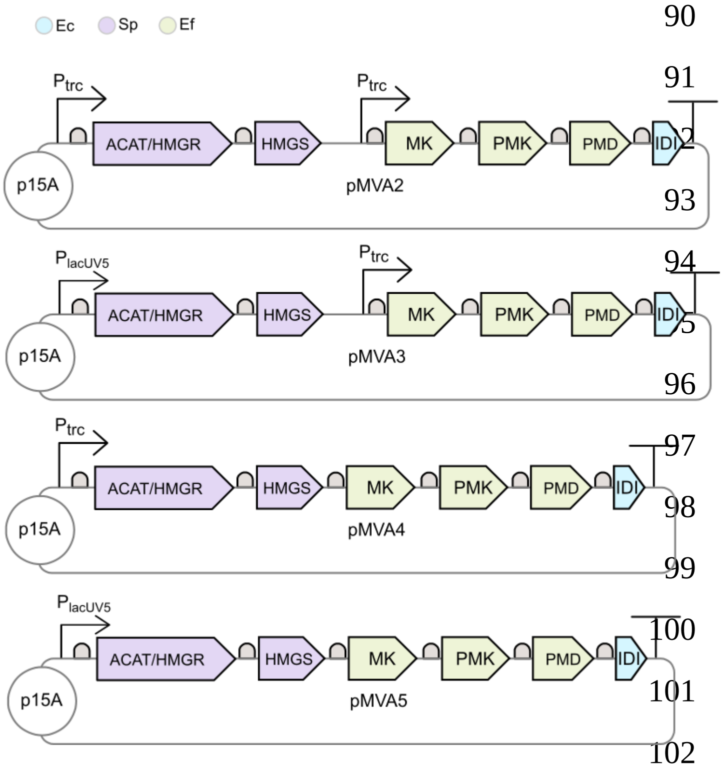

Figure S6. Genetic organisation of the pMVA series of plasmids. Members differ through the position and identity of gene promoters.

S7. pMVA plasmid series (*S*)-limonene production

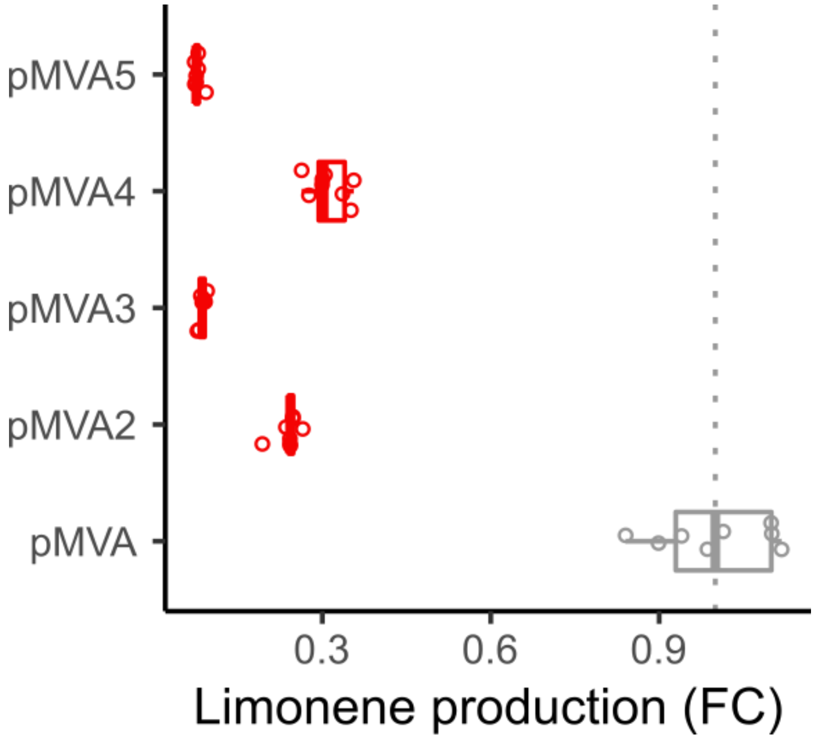

Figure S7. Screening of the pMVA series of plasmids co-expressed with plasmid pGL403. Limonene production titres from triplicate cultures containing the indicated pMVA variant (red) compared to the original plasmid pMVA (gray) are displayed as fold-change (FC). Boxplots indicate the median, and upper/lower quartiles.

S8. Comparison of RBS sequences for pMVA2lib1 high/low producers

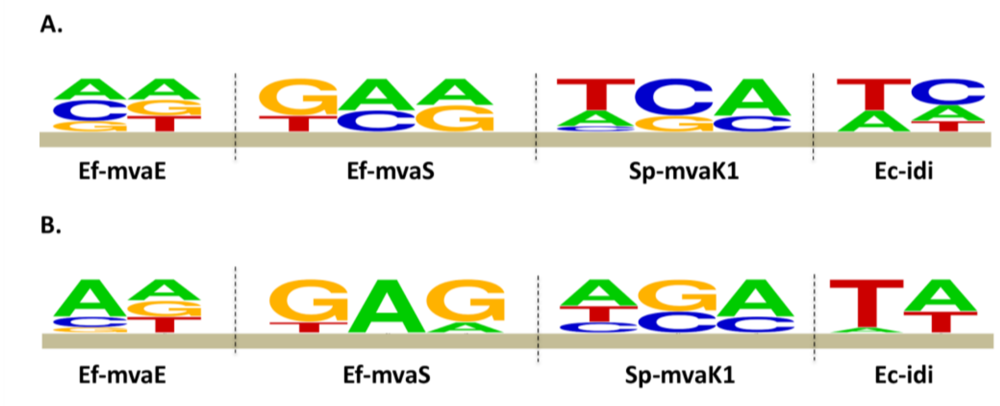

Figure S8. Comparison of the frequency of variable nucleotides distribution at RBS positions in the sequenced high and low producer variants of pMVA2lib1. A) Distribution for the *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1* and *Ec-idi* RBSs for low producers in the pMVA2 library (FC < lower quartile); B) Distribution for the *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1* and *Ec-idi* RBS sequences for high producers in the pMVA2 library (FC > upper quartile). Sequence logos were generated through the weblogo service (weblogo.berkeley.edu).

S9. Validation of machine learning for pMVA2lib1

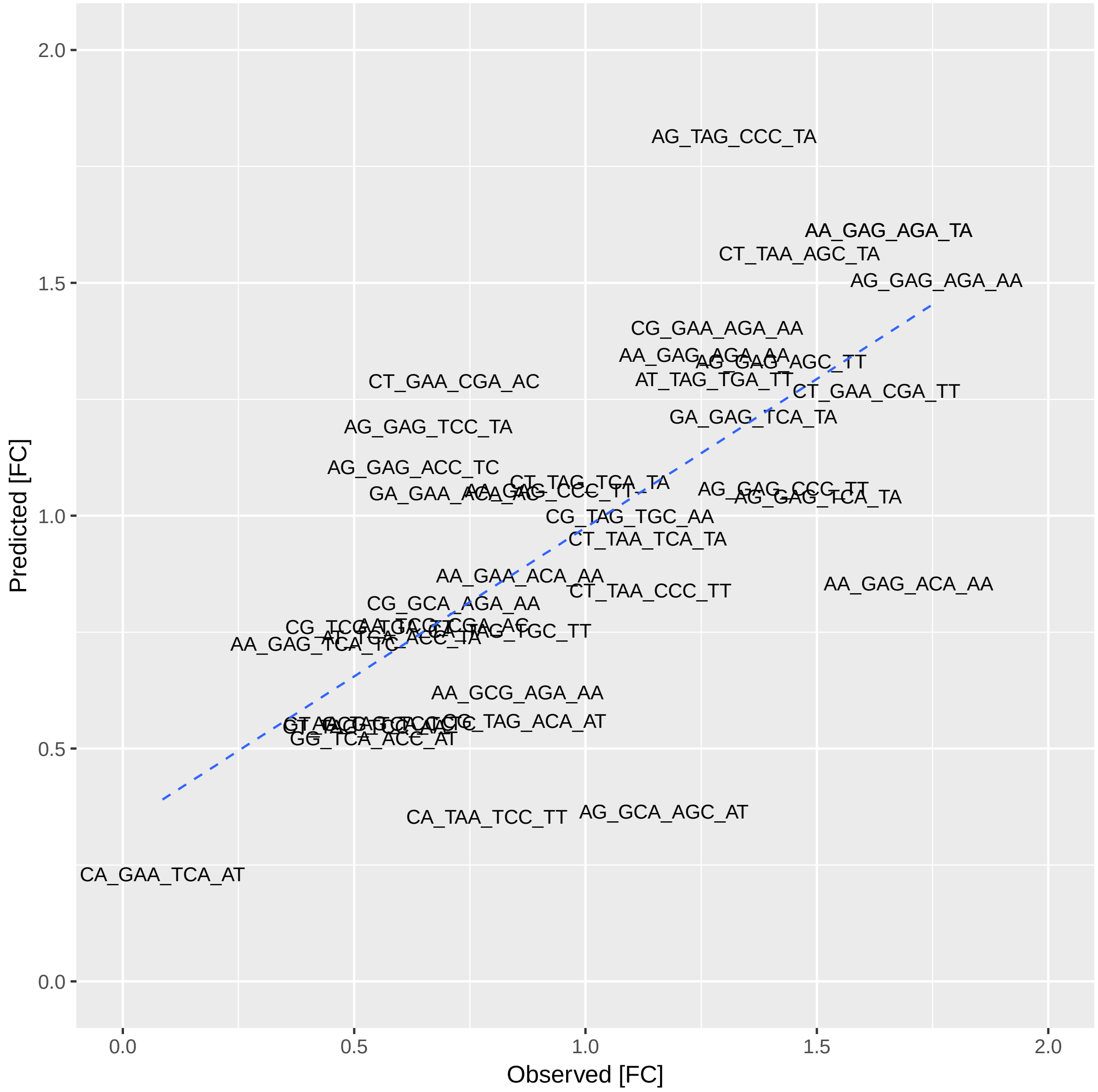

Figure S9. Model validation for library pMVA2lib1. The observed levels of limonene production (horizontal axis) are compared with predicted levels (vertical axis) in fold change (FC) for each RBS combination using leave-one-out cross-validation with a resulting performance of $Q^2 = 0.48$. The blue dashed line is the linear fitting of the points.

S10. Re-screening of pMVA2lib high producers from 1 ml screen

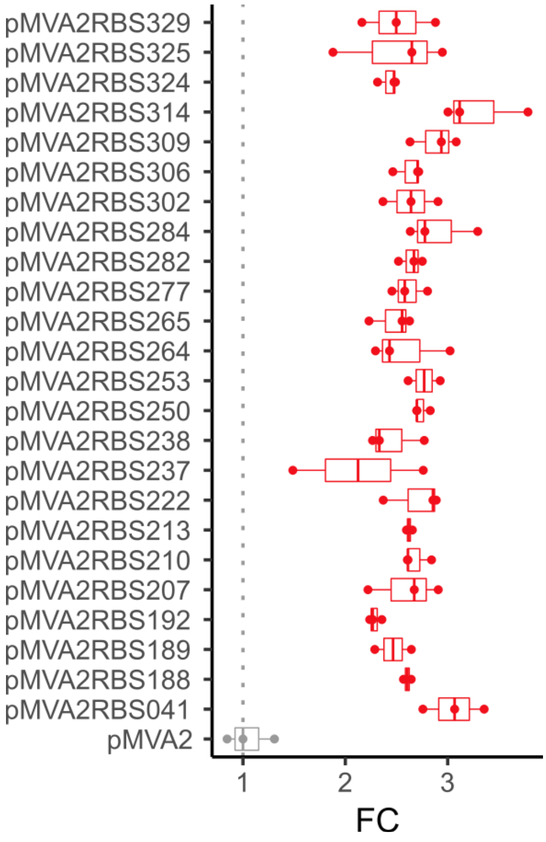

Figure S10. Re-screen of high producer variants from the pMVA2 libraries. High producing variants from the pMVA2lib1/2 were re-introduced into cells, along with pGL403, and grown in triplicate to confirm their production titres relative to starting plasmid pMVA2.

S11. Comparison of production titres from individual clones of pMVA2lib1 at the 1 ml and 5 ml scale

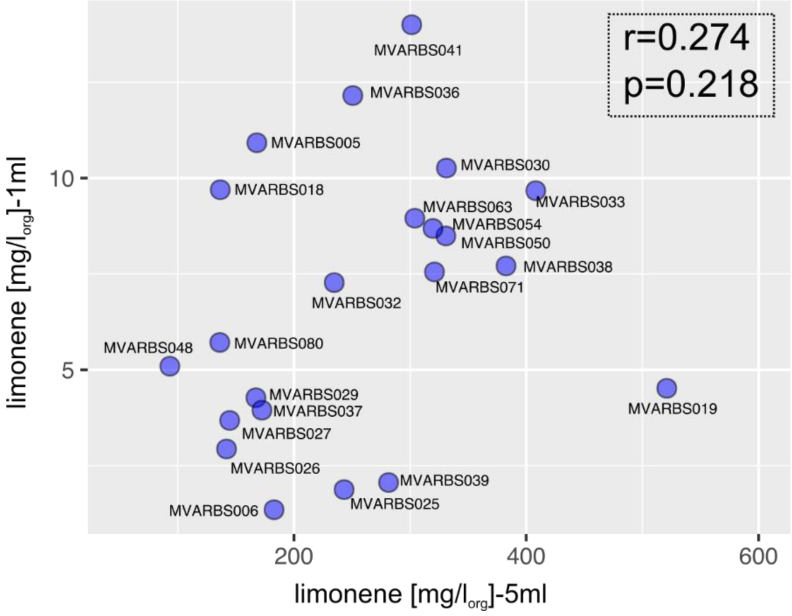

Figure S11. Comparison of limonene production titres for members of pMVA2lib1 at the 1 and 5 ml scaled.

Titres produced by individual clones from RBS library pMVA2lib2 when grown at either the 1 ml or 5 ml scale. As titres for 1ml cultures increase titres at 5ml screening scale tend to increase as well. This positive association however is rather weak as indicated by a non-significant (p=0.218) Pearson coefficient r=0.274.

S12. Comparison of RBS sequences pMVA2lib1 screened at the 1 ml and 5 ml scale

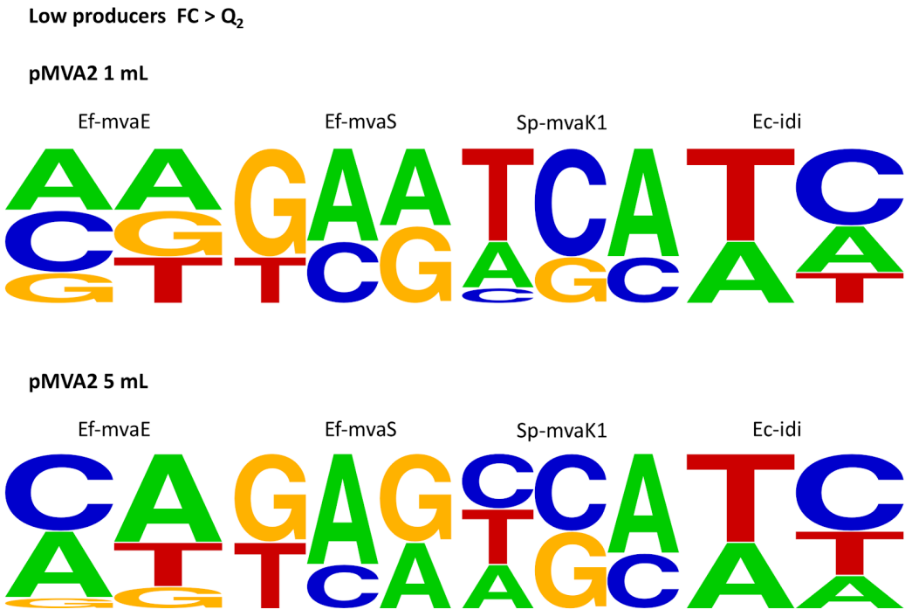

Figure. S12. Comparison of frequency of nucleotides at RBS positions for high producers in the pMVA2lib1 library when screened at 1 mL or 5 mL scale. A) Distribution for the *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1* and *Ec-idi* RBS sequences for high producers in the pMVA2 library (FC > upper quartile); B) Distribution for the *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1* and *Ec-idi* RBS for low producers in the pMVA2 library (FC < lower quartile). Sequence logos were generated through the weblogo service (weblogo.berkeley.edu).

S13. Pathway performance at 25 ml scale in baffled flasks

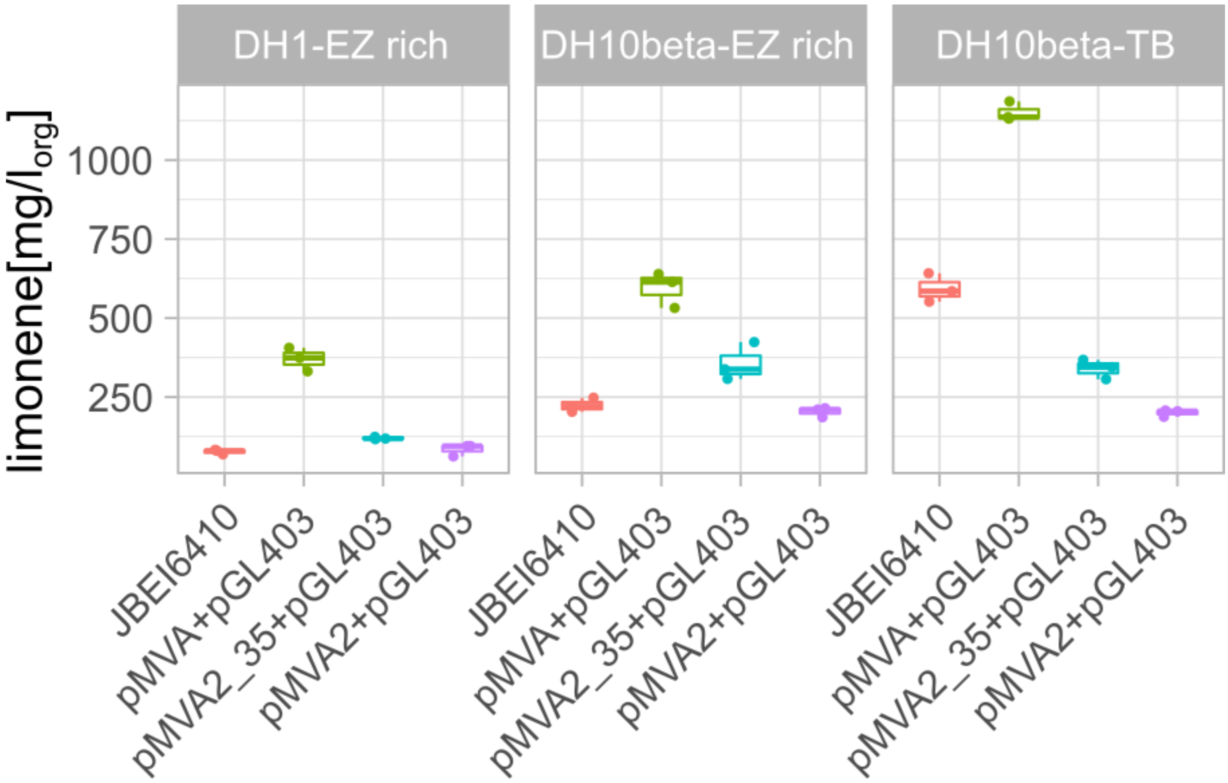

Figure. S13. Comparison of limonene production titres from key pathways in 25 ml shake-flask cultures. The highest limonene producing pathways from each round of translational tuning (pGL403, pMVA2035) plus the previously published pathway (pJBEI6410) were re-tested for limonene production in 25 ml cultures in baffled flasks in either *E. coli* DH1 or DH10β and in EZ Rich defined media or TB. The same relative production tires were seen from the individual pathway in each condition tested, with the highest titers for each pathway observed in *E. coli* DH10β and TB media.

Table S1. Enzymes used in this study

Table S1. Enzymes used in this study**.**

Enzyme ID	**Enzyme activity**	**Source**	**Uniprot ID**	**Ref**
Ef-mvaE	Acetyl-CoA acetyltransferase/HMG-CoA reductase	*Enterococcus faecalis*	Q9FD70	1
Ef-mvaS	Hydroxymethylglutaryl-CoA synthase	*Enterococcus faecalis*	Q9FD71* (A110G)	2
Sp-mvaK1	Mevalonate kinase	*Streptococcus pneumoniae*	A0A0H2UNK6	3
Sp-mvaK2	Phosphomevalonate kinase	*Streptococcus pneumoniae*	A0A0I6R1S1	3
Sp-mvaD	Diphosphomevalonate decarboxylase	*Streptococcus pneumoniae*	A0A0B7L0J5	3
Ec-idi	Isopentenyl-diphosphate Delta-isomerase	*E. coli*	Q46822	4
trAg-GPPS	Geranyl diphosphate synthase	*Abies grandis*	Q8LKJ2	5
trMs-limS	4S-limonene synthase	*Mentha spicata*	Q40322	6

*The enzyme *Ef-mvaS* contains a single mutation of A110G with respect to the wild type. The prefix “tr” indicates that the gpps and limS genes were truncated to remove N-terminal signal peptides.

Table S3. Machine learning validation statistics

Table S3. Coefficients of determination (R^2, Q^2) and root mean square errors (RMSEP, RMSECV) of prediction and 10-fold cross-validation, respectively, and p-value for permutation tests for the predictors of pGLlib and pMVA2lib1/2.

Plasmid	Statistic	Value	p-value
pGL	R^2	0.97	≤ 0.001
pGL	Q^2	0.87	≤ 0.001
pGL	RMSEP	0.09	≤ 0.001
pGL	RMSECV	0.47	≤ 0.001
pMVA2	R^2	0.67	≤ 0.001
pMVA2	Q^2	0.48	≤ 0.001
pMVA2	RMSEP	0.24	≤ 0.001
pMVA2	RMSECV	0.46	≤ 0.021

Table S5. Oligos used in this study

Table S5. Oligonucleotides used in this study.

Oligo name	**Sequence (5’-3’)**
PCR primers	
mvaK1trc-F	TCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCAGACAGGCTCCCATTTAACATAAGG
mvaK1trc-R	CCACACATTATACGAGCCGGATGATTAATTGTCAACAGCTTTAATTACGATAGCTACGCACGGTG
GPPSRBS-F	CCCTATCAGTGATAGAGAAAAGAATTCACGATCTTAAGTARRCGVGGAAAATAATGGAGTTCGACTTCAACAAATAC
GPPSRBS-R	AACATAATCTGCCAGACCCAG
limSRBS-F	GCATTTCGTCAGAATTAAGCTTAGAGTAAAACTAAGCATCTAAGRGSGVTACTAATGGAACGTCGTAGCG
limSRBS-R	TTATGCAAACGGTTCAAACAGGGTACG
pBbSeq-R2	CAGTGTGACTCTAGTAGAGAGCGTTCAC
tetprom2	GACCTCATTAAGCAGCTCTAATGC
pGL-F	GGATCCAAACTCGAGTAAGG
pGL-R	AGTGGTAAAATAACTCTATCAACG
MVAlib1.1-F	GCTAAGGAGTCGCACGAGACGCCAAATWGGGAGGHGGGCGATGCAGACCGAACATG
MVAlib1.1-R	TAACTGGCTTGGAGGAGCGC
MVAlib1.2-F	CCTCGGTTCAAAGAGTTGGTAGC
MVAlib1.2-R	GCGACATCGTATAACGTTACTGG
MVAlib1.3-F	AGAGTATGCCGGTGTCTCTTATCAG
MVAlib1.3-R	GCAGTGCATCAATAATCACCACGGTTTTCATATGTACGTTHCBTCCTTAAAAGATCTTTTGAATTCTGAAATTGTTATCCG
MVAlib2.1-F	GTACCCCGATTGGTAAATACAAAGGTAG
MVAlib2.1-R	GGCCACCACCAATACACAGG
MVAlib2.2-F	TGGGTCTGGCAATGCTGCTG
MVAlib2.2-R	AATTGCCAGGGCACGATCCTG
MVAlib3-F	CTGAATGATCTGCGTAAACAGTAATGATTAGCGACAAAAKATGAGGMGTRCAAAAAATGACCATTGGCATCGACAAAATCAG
MVAlib3-R	GCCTGACCAACACCAACTTTTTTGGTCATCGTATKCCTCSTCDCGTGTTAAATGGGAGCCTGTCTGAAATTG
MVAlib4.1-F	ACATAGCAAAATTATCCTGATTGGTGAAC
MVAlib4.1-R	GATCTTCTGCAACAATACATGCCAG
MVAlib4.2-F	TGGTTCTGTATCAGAGCTTTGATCGTC
MVAlib4.2-R	AACAATCATCCTGGCTCAGATCTTTG
Bridging oligos	

brGL	GCTGCTGGGTCTGGCAGATTATGTTGCATTTCGTCAGAATTAAGCTTAGAGTAAAACTAAGCAT
brLP	CCGTACCCTGTTTGAACCGTTTGCATAAGGATCCAAACTCGAGTAAGGATCTCCAGG
brPG	TGTTGACACTCTATCGTTGATAGAGTTATTTTACCACTCCCTATCAGTGATAGAGAAAAGAATTCACGATCTTAAGT
brMVAlib1.1-1.2	TGACTGCGCTCCTCCAAGCCAGTTACCTCGGTTCAAAGAGTTGGTAGCTCAGAGAACC
brMVAlib1.2-1.3	GGTGAATGTGAAACCAGTAACGTTATACGATGTCGCAGAGTATGCCGGTGTCTCTTATCAGACCGT
brMVAlib1.3-2.1	GTACATATGAAAACCGTGGTGATTATTGATGCACTGCGTACCCCGATTGGTAAATACAAAGGTAGCCTGAGC
brMVAlib2.1-2.2	GCAAGCCTGTGTATTGGTGGTGGCCTGGGTCTGGCAATGCTGCTGGAAC
brMVAlib2.2-3	CAATGAATCAGGATCGTGCCCTGGCAATTCTGAATGATCTGCGTAAACAGTAATGATTAGCGACAAA
brMVAlib3-4.1	CGATGACCAAAAAAGTTGGTGTTGGTCAGGCACATAGCAAAATTATCCTGATTGGTGAACATGCCG
brMVAlib4.1-4.2	TGGGTGATCTGGCATGTATTGTTGCAGAAGATCTGGTTCTGTATCAGAGCTTTGATCGTCAGAAAGC
brMVAlib4.1-1.1	GCAAAACCAAAGATCTGAGCCAGGATGATTGTTGCTAAGGAGTCGCACGAGACGCCA

Table S6. Plasmids and strains used in this study

Table S6. Plasmids and strains used in this study.

Plasmid reference	**Plasmid short and systematic name**	**Description (Origin of replication, Antibiotic marker, Reference(s), Promoters and Operons)**	**Reference**
pJBEI-6410	pBbA5a-MTSAe-T1f-MBI(f)-T1002i-Ptrc-trGPPS(co)-LS	p15A, Ampr, PlacUV5, MTSA, T1, MBI-f, T1002, Ptrc, trGPPS, LS	7
pMVA	pBbA5a-MTSAe-T1f-MBI(f)-T1002i	p15A, Kanr, PlacUV5, MTSA, T1, MBI-f, T1002	8
pMVA2*	pBbA1k-ES-1-K1K2Didi	p15A, Kanr, Ptrc, mvaES, Ptrc, mvaK1K2D, idi	This study
pMVA3	pBbA5k-ES-1-K1K2Didi	p15A, Kanr, PlacUV5, mvaES, Ptrc, mvaK1K2D, idi	This study
pMVA4	pBbA1k-ESK1K2Didi	p15A, Kanr, Ptrc, mvaES, mvaK1K2D, idi	This study
pMVA5	pBbA5k-ESK1K2Didi	p15A, Kanr, PlacUV5, mvaES, mvaK1K2D, idi	This study
pMVA2RBS035*	pBbA1k-ES-1-K1K2Didi	p15A, Kanr, Ptrc, mvaES, Ptrc, mvaK1K2D, idi	This study
pGL*	pBbB2a-trAgGPPS-trMsLS	pBBR, Ampr, Ptet, trAgGPPS-trMsLS	8
pGL403*	pBbB2a-trAgGPPS-trMsLS403	As pGL with optimised RBSs.	This study
Strain	**Alternative designation**	**Genotype**	**Source**
DH10β	NEB 10-beta	Δ(ara-leu) 7697 araD139 fhuA ΔlacX74 galK16 galE15 e14-φ80dlacZΔM15 recA1 relA1 endA1 nupG rpsL (StrR) rph spoT1 Δ(mrrhsdRMS-mcrBC)	New England Biolabs
DH1	ATCC33849	F- supE44 hsdR17 recA1 gyrA96 relA1 endA1 thi-1 lambda-	ATTC

*Plasmids deposited with Addgene

References

1. Wilding, E.I., Brown, J.R., Bryant, A.P., Chalker, A.F., Holmes, D.J., Ingraham, K.A., Iordanescu, S., So, C.Y., Rosenberg, M. and Gwynn, M.N. (2000) Identification, evolution, and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in gram-positive cocci. *J. Bacteriol*. 182, 4319-4327.
2. Yang, J., Xian, M., Su, S., Zhao, G., Nie, Q., Jiang, X., Zheng, Y. and Liu, W. (2012) Enhancing production of bio-isoprene using hybrid MVA pathway and isoprene synthase in E. coli. *PLoS One.* 7, e33509.
3. Yoon, S.H., Lee, S.H., Das, A., Ryu, H.K., Jang, H.J., Kim, J.Y., Oh, D.K., Keasling, J.D. and Kim, S.W. (2009) Combinatorial expression of bacterial whole mevalonate pathway for the production of beta-carotene in E. coli. *J. Biotechnol*. 140, 218-226.
4. Yoon, S.H., Lee, Y.M., Kim, J.E., Lee, S.H., Lee, J.H., Kim, J.Y., Jung, K.H., Shin, Y.C., Keasling, J.D. and Kim, S.W. (2006) Enhanced lycopene production in Escherichia coli engineered to synthesize isopentenyl diphosphate and dimethylallyl diphosphate from mevalonate. *Biotechnol. Bioeng*. 94, 1025-1032.
5. Yang, J., Nie, Q., Ren, M., Feng, H., Jiang, X., Zheng, Y., Liu, M., Zhang, H. and Xian, M. (2013) Metabolic engineering of Escherichia coli for the biosynthesis of alpha-pinene. *Biotechnol. Biofuels*. 6, 60.
6. Colby, S.M., Alonso, W.R., Katahira, E.J., McGarvey, D.J. and Croteau, R. (1993) 4S-limonene synthase from the oil glands of spearmint (Mentha spicata). cDNA isolation, characterization, and bacterial expression of the catalytically active monoterpene cyclase. *J. Biol. Chem*. 268, 23016-23024.
7. Alonso-Gutierrez, J., Chan, R., Batth, T.S., Adams, P.D., Keasling, J.D., Petzold, C.J. and Lee, T.S. (2013) Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production. *Metab. Eng*. 19, 33-41.
8. Leferink, N.G., Jervis, A.J., Zebec, Z., Toogood, H.S., Hay, S., Takano, E. and Scrutton, N.S. (2016) A ‘plug and play’ platform for the production of diverse monoterpene hydrocarbon scaffolds in Escherichia coli. *ChemistrySelect*. 1,1893-1896.