







# An in vivo gene amplification system for high level expression in *Saccharomyces cerevisiae*

Bingyin Peng <sup>1,2,3,4</sup>✉, Lygie Esquirol <sup>1,5</sup>, Zeyu Lu<sup>1,3,4</sup>, Qianyi Shen<sup>1,3,4</sup>, Li Chen Cheah <sup>1,3</sup>, Christopher B. Howard<sup>1</sup>, Colin Scott <sup>2,6</sup>, Matt Trau <sup>1,7</sup>, Geoff Dumsday<sup>8</sup> & Claudia E. Vickers <sup>2,3,4,5</sup>✉

Bottlenecks in metabolic pathways due to insufficient gene expression levels remain a significant problem for industrial bioproduction using microbial cell factories. Increasing gene dosage can overcome these bottlenecks, but current approaches suffer from numerous drawbacks. Here, we describe HapAmp, a method that uses haploinsufficiency as evolutionary force to drive in vivo gene amplification. HapAmp enables efficient, titratable, and stable integration of heterologous gene copies, delivering up to 47 copies onto the yeast genome. The method is exemplified in metabolic engineering to significantly improve production of the sesquiterpene nerolidol, the monoterpene limonene, and the tetraterpene lycopene. Limonene titre is improved by 20-fold in a single engineering step, delivering  $\sim 1 \text{ g L}^{-1}$  in the flask cultivation. We also show a significant increase in heterologous protein production in yeast. HapAmp is an efficient approach to unlock metabolic bottlenecks rapidly for development of microbial cell factories.

<sup>1</sup> Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, Brisbane, QLD 4072, Australia. <sup>2</sup> CSIRO Synthetic Biology Future Science Platform, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Black Mountain, ACT 2601, Australia. <sup>3</sup> ARC Centre of Excellence in Synthetic Biology, Queensland University of Technology, Brisbane, QLD 4000, Australia. <sup>4</sup> Centre of Agriculture and the Bioeconomy, School of Biology and Environmental Science, Faculty of Science, Queensland University of Technology, Brisbane, QLD 4000, Australia. <sup>5</sup> Griffith Institute for Drug Discovery, Griffith University, Brisbane, QLD 4111, Australia. <sup>6</sup> Biocatalysis and Synthetic Biology Team, CSIRO Land and Water, Black Mountain Science and Innovation Park, Canberra, ACT 2061, Australia. <sup>7</sup> School of Chemistry and Molecular Biosciences (SCMB), The University of Queensland, Brisbane, QLD 4072, Australia. <sup>8</sup> CSIRO Manufacturing, Clayton, VIC 3169, Australia. ✉email: [bingyin.peng@qut.edu.au](mailto:bingyin.peng@qut.edu.au); [claudia.vickers@qut.edu.au](mailto:claudia.vickers@qut.edu.au)

To achieve economically viable rates, yields and titres for a given product in microbial cell factories, it is commonly necessary to increase expression of introduced genetic constructs<sup>1,2</sup>. This is typically achieved by manipulating transcription levels via transcriptional control elements (promoters and other genetic sequences)<sup>3</sup>. However, this approach is subject to thresholds on individual constructs. This often means that expression levels are insufficient for a desired application. For example, enzymes with poor catalytic properties that cannot be improved by enzyme engineering represent significant flux bottlenecks in metabolic engineering<sup>4</sup>. In addition, where extremely high product levels are required (e.g., protein production systems), very high expression can deliver a direct economic benefit to the bioprocess. Increasing the gene dosage can be used to overcome transcriptional thresholds and increase expression levels.

The brewer's yeast *Saccharomyces cerevisiae* is a eukaryotic model organism and an important industrial microorganism for production of biofuels, biochemicals, and biopharmaceuticals. In *S. cerevisiae*, multi-copy yeast episomal plasmids or genome integration into ribosomal DNA (rDNA) sites are typically used to increase gene dosage<sup>5–8</sup>. However, these approaches are not stable in the absence of selection pressure, and plasmids can suffer from copy number instability leading to variable expression levels<sup>5–8</sup>. In addition, use of selection systems in industrial processes adds additional costs and often is not scalable<sup>9,10</sup>. To stabilise strains without the need for selective antibiotic or auxotrophy systems, auto-selection markers such as glycolytic genes (*FBA1*, fructose-bisphosphate aldolase; *POT1/TPI1*, triosephosphate isomerase) can be used<sup>5,11,12</sup>. However, this requires the background strains to have the correct genotype for knock-out. Transposable elements can also be used for multi-copy integration, however variable copies are integrated at random loci on genome, which means integrated components cannot be removed to facilitate future engineering steps (for example, swapping terpenoid synthases for different terpenoid production platforms)<sup>13–17</sup>. A method overcoming all these limitations is highly desirable.

Gene amplification commonly happens in nature during cell proliferation, as part of molecular evolution, as well as in some laboratory experiments<sup>2,18–23</sup>. In yeast, tandem amplification of fitness-associated genes on the genome permits improved survival and propagation of cells under new or changing conditions<sup>18–20</sup>. For example, amplification of the xylose isomerase, cellobiose-utilisation, and copper resistance (*CUP1*) genes occurs over prolonged adaptive cultivation on xylose<sup>19,20</sup>, cellobiose<sup>24</sup>, and copper ions<sup>25</sup>, respectively. Another example is the amplification of tandem repeated rDNA under some conditions<sup>26</sup>. These examples demonstrate that if the expression level of a gene product is tightly linked to growth fitness and cannot meet the needs for maximum growth, gene amplification can occur through adaptive evolution.

In diploids, haploinsufficiency describes a state whereby one allele at a heterozygous locus provides little or no product, and the combined product from both alleles is insufficient to deliver the wild type phenotype<sup>27</sup>. Expression dosage of haploinsufficient genes links tightly with the growth fitness in yeast<sup>28</sup>. This can be explored as an evolutionary force to drive gene amplification and as a selection pressure for maintenance of the amplified constructs under normal cultivation conditions.

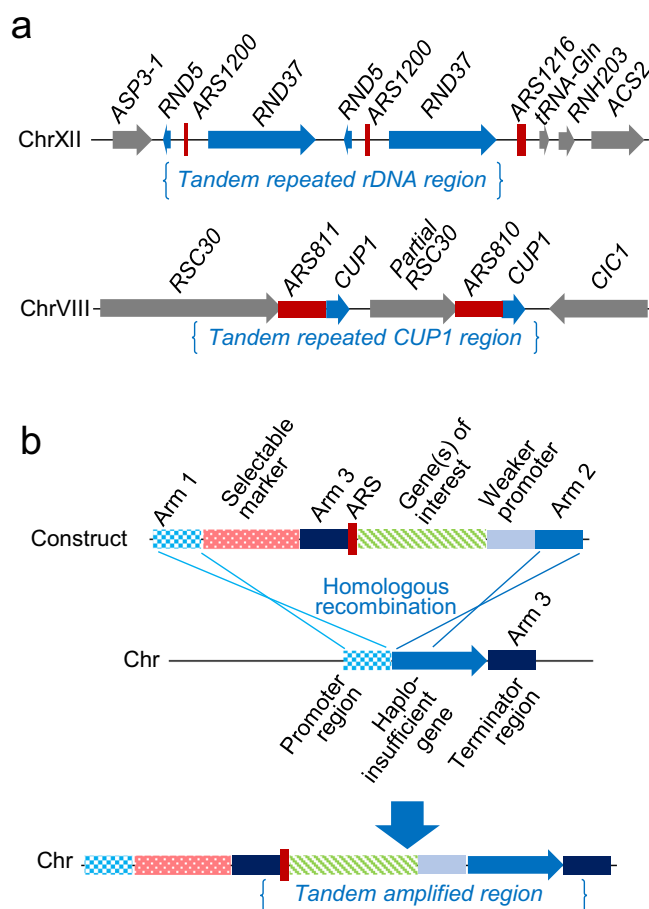
Here, we design an artificial genetic structure that enables amplification of a haploinsufficient gene through tuning of its promoter strength or translational efficiency (HapAmp). This structure is incorporated into genetic vectors which can be used to introduce multiple copies of linked heterogeneous genes on the genome. We exemplify the applications of this technique by

developing yeast factories for improved production of terpenes by metabolic engineering and for high production of pharmaceutically relevant proteins.

## Results

**Construct design for in vivo gene amplification.** Two elements are required for gene amplification to occur: (1) a gene linked to cell fitness, and (2) homologous DNA sequences to support recombination<sup>20</sup>. In addition, a strong replication origin can promote amplification<sup>29–31</sup>. These three elements exist in tandem repeat in the rDNA region and the *CUP1* region in the yeast genome (Fig. 1a).

We designed a genetic structure for gene amplification in yeast (Fig. 1b). The construct has recombination arms at each end. Arm 1 is homologous to the promoter region of a haploinsufficient gene, and Arm 2 is homologous to the initial part of the haploinsufficient gene open reading frame. This allows insertion of the construct into the genome by homologous recombination. Downstream of Arm 1 are a selectable marker for transformation selection and homologous Arm 3, which is homologous to the terminator region of the haploinsufficient gene. Between Arm 3 and Arm 2, there are an autonomous replicating sequence (ARS) and a promoter. The promoter is weaker than the native promoter of the haploinsufficient gene and positioned such that integration results in substitution of the native promoter of the haploinsufficient gene with the weaker



**Fig. 1** Design of in vivo gene amplification. **a** Natural genome structures at the ribosomal DNA (rDNA) locus on chromosome XII and the *CUP1* locus on chromosome VII. **b** Construct design for in vivo gene amplification (HapAmp). ARS autonomous replicating sequence.

promoter. Genes of interest, to be expressed heterologously, can be inserted between Arm 3 and the weaker promoter.

Driving expression through a weaker promoter attenuates the protein yield from each copy of the haploinsufficient gene. This, in turn, is expected to decrease the growth rate in yeast. Native amplification of the region between homologous Arm 3 will then occur as yeast evolves towards faster growth.

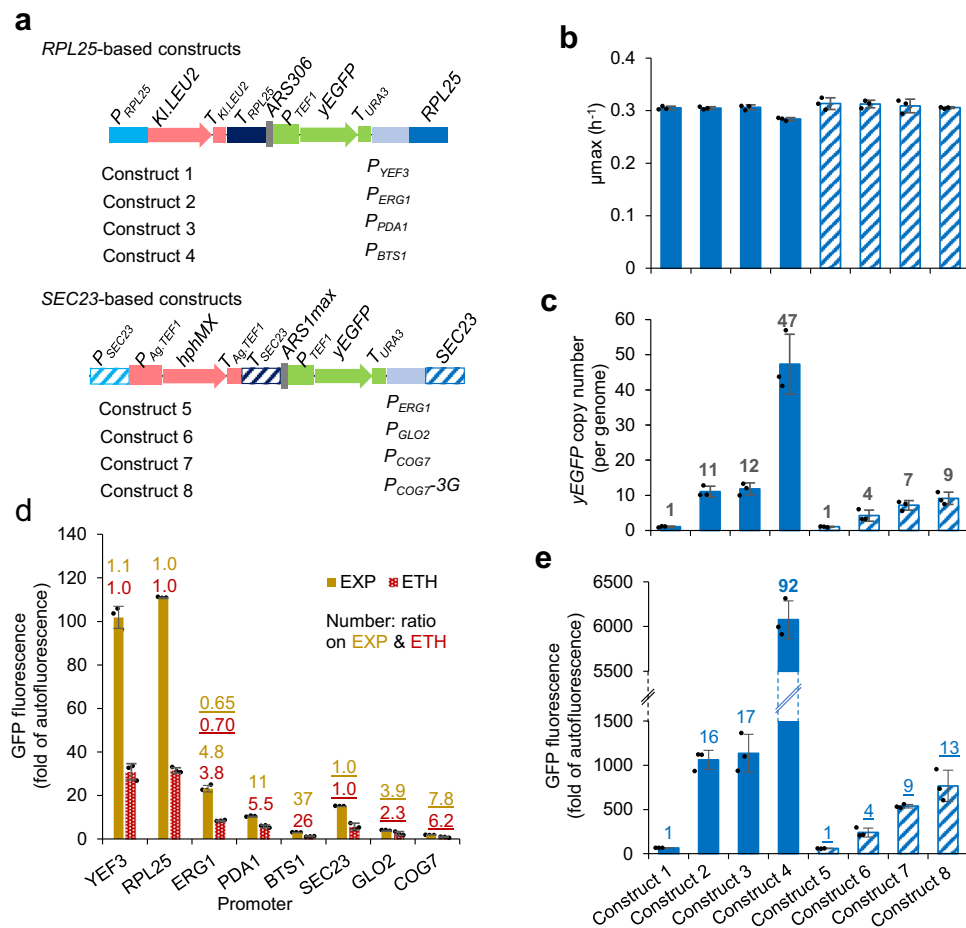
**Using *RPL25* or *SEC23* haploinsufficient gene loci to drive amplification.** The effect of haploinsufficient genes on growth fitness has been characterised previously<sup>28</sup>. We used the ribosomal 60S subunit protein L25 (*RPL25*) and the *SEC23*-encoding component of the Sec23p-Sec24p heterodimer of the COPII vesicle coat. These two genes have the strongest fitness effect in rich medium and in minimal mineral medium<sup>28</sup>. We developed four constructs with *RPL25* as the driving gene, *LEU2* as selection marker, and an early-firing ARS *ARS306*<sup>32</sup> to facilitate amplification; and three constructs with *SEC23* as the driving gene, hygromycin B resistant gene *hphMX* as selection marker, and the strong *ARS1max* ARS<sup>33</sup> to facilitate amplification (Fig. 2a).

To identify promoters with suitable expression strengths, promoters were selected from the wide variety of promoters we previously analysed<sup>34</sup>, to test with each target locus (Fig. 2a, d).

For the *RPL25* constructs we used the *YEF3* promoter (which has similar strength to the *RPL25* promoter; Construct 1) and the *ERG1*, *PDA1*, or *BTS1* promoters (all with multiple-fold weaker expression than *RPL25* promoter; Constructs 2–4). For the *SEC23* constructs, we used the *ERG1* promoter (stronger than the *SEC23* promoter; Construct 5), the *GLO2* promoter, or the *COG7* promoter (both multiple-fold weaker than the *SEC23* promoter; Constructs 6 and 7). An eighth promoter construct was designed and tested later (see below). We used yeast-enhanced green fluorescent protein (yEGFP) under the control of the *TEF1* promoter and the *URA3* terminator as the gene of interest and as a reporter for proof of concept.

The seven constructs were transformed into *S. cerevisiae* CEN.PK strains. Transformation plates were screened by imaging yEGFP fluorescence under blue light (Supplementary Fig. 1a, c) and colonies were selected for increased fluorescence. For each construct, six strongly fluorescing clones were selected. Visual observation after sub-culturing demonstrated an inverse correlation between promoter strength (Fig. 2d) and GFP fluorescence (Supplementary Fig. 1b). Three clones with similar fluorescence were selected for quantitative characterisation for each construct.

Where promoter strength was similar or greater than the native promoter, yEGFP was found at a single copy on the genome (Fig. 2c: Constructs 1 and 5), and fluorescence (Fig. 2e:



**Fig. 2** Design and characterisation of gene amplification constructs for haploinsufficient target genes *RPL25* or *SEC23*. **a** Schematic of gene amplification constructs. **b, c, e** Maximum growth rate, yEGFP (yeast-enhanced green fluorescent protein) gene copy number, and yEGFP fluorescence in strains transformed with the constructs in **a**. Strains were selected by brightness of yEGFP fluorescence (Supplementary Fig. 1). **d** Promoter characterisation using yEGFP as the reporter in the cells at the exponential growth phase (EXP) and the post-diauxic-shift growth phase (ETH) when ethanol was used as the carbon source. Yeast cells were grown in microplates in **d** and in flasks in **b, c, e**. yEGFP fluorescence is expressed as percentage of exponential-phase auto-fluorescence of the reference strain. The numbers were calculated by dividing the mean value for *RPL25* or *SEC23* (underlined) by the mean value. Mean values ± standard deviations are shown (N = 3 independent biological replicates). Source data are provided as a Source Data file.

Constructs 1 and 5) was similar to fluorescence we observed previously in strains with a single copy of the *P<sub>TEF1</sub>-yEGFP-T<sub>URA3</sub>* construct<sup>3</sup>. *yEGFP* gene copy number and fluorescence both increased where the native promoter was substituted for weaker promoters (Fig. 2c, e: Constructs 2–4, 6, 7). Copy number increased from 4-fold to 47-fold, whereas fluorescence increase was 4-fold to 92-fold. There was a strong positive correlation between copy number and fluorescence ( $r^2 = 0.985$ ), and a weak negative correlation between fluorescence and promoter strength/copy number ( $r^2 = 0.376$  and  $0.694$  respectively). The most remarkable result was where the *RPL25* promoter was substituted for the *BTS1* promoter; this resulted in ~47 copies of *yEGFP* per genome and a ~92-fold increase *yEGFP* fluorescence (Fig. 2c, e).

To further increase copy number at the *SEC23* locus, we attenuated translation by making a construct with three non-preferred glycine codons (GGA) inserted following the start codon of *SEC23* under the control of the *COG7* promoter (Fig. 2a: Construct 8), which delivered the most gene amplification in the first round (9 copies). A slight increase in gene copy and fluorescence was obtained (Fig. 2c, e). Translational down-regulation by use of non-preferred codons provides a second mechanism to drive an increase in copy number for genes at haploinsufficient gene loci.

In the initial design (Fig. 1), we include ARS in the module basing on the genetic features at naturally amplified genomic loci. To confirm the role of ARS in the current system, we removed the ARS sequence in the Construct 3. The ARS-removed construct could lead to the formation of the very fluorescent colonies after transformation (Supplementary Fig. 1). This indicates that ARS may not be essential for HapAmp.

Increased copy number did not negatively impact the growth rate of any of the strains except for clones with the *P<sub>BTS1</sub>-RPL25* construct (Fig. 2b), which had an exceptionally high integration copy number (Fig. 2c). This strain showed an ~7% decrease in growth rate (two-tailed *t*-test  $p = 0.001$ ).

Long-read sequencing on strains containing Constructs 3 and 4 confirmed that the constructs were integrated into the *RPL25* (YOL127W) locus and that *yEGFP-RPL25* sequences were amplified in tandem repeat structures (Supplementary Figs. 2 and 3–5). The strain expressed the highest level of *yEGFP* (Construct 4) was sub-cultured in yeast extract-peptone-glucose medium for ~48 generations for stability test (Supplementary Fig. 6). GFP fluorescence levels and population homogeneity did not change, indicating that HapAmp is genetically stable.

**Improving heterologous production of the sesquiterpene trans-nerolidol.** We examined the performance of the HapAmp method using sesquiterpene ( $C_{15}$ ; trans-nerolidol) production. We used a background strain with an upregulated mevalonate pathway for production of terpene precursors (o401R)<sup>35–38</sup>. In this strain, the *GAL80* repressor gene is disrupted allowing diauxic induction of *GAL* promoters, which are used to control transgenes.

We constructed a reference strain N401-1 harbouring a multi-copy 2 $\mu$  plasmid pJT9RFR<sup>39</sup> (Fig. 3a) with overexpression cassettes for farnesyl pyrophosphate synthase (*ERG20*) and nerolidol synthase (*Ac.NES1*). The nerolidol synthase cassette includes a fluorescence-activating and absorption-shifting tag (*Y-FAST*)<sup>40</sup> and a 2A peptide from Equine rhinitis B virus 1<sup>41</sup> fused to the N-terminus of nerolidol synthase. This allows *Y-FAST* fluorescence to be used as a proxy for nerolidol synthase expression<sup>39</sup>.

The nerolidol synthase expression cassette (*Y-FAST-2A-Ac.NES1*) was cloned into the *RPL25* insertion vector in the amplification region with three different promoters for replacement

of the *RPL25* promoter; the *ERG20* expression cassette was cloned at the non-amplification region (Fig. 3b). Colonies with bright *Y-FAST* fluorescence were selected from the transformation plates. This delivered strains N401-2, N401-3, & N401-4 (promoters *P<sub>ERG1</sub>*, *P<sub>PDA1</sub>*, and *P<sub>BTS1</sub>*, respectively).

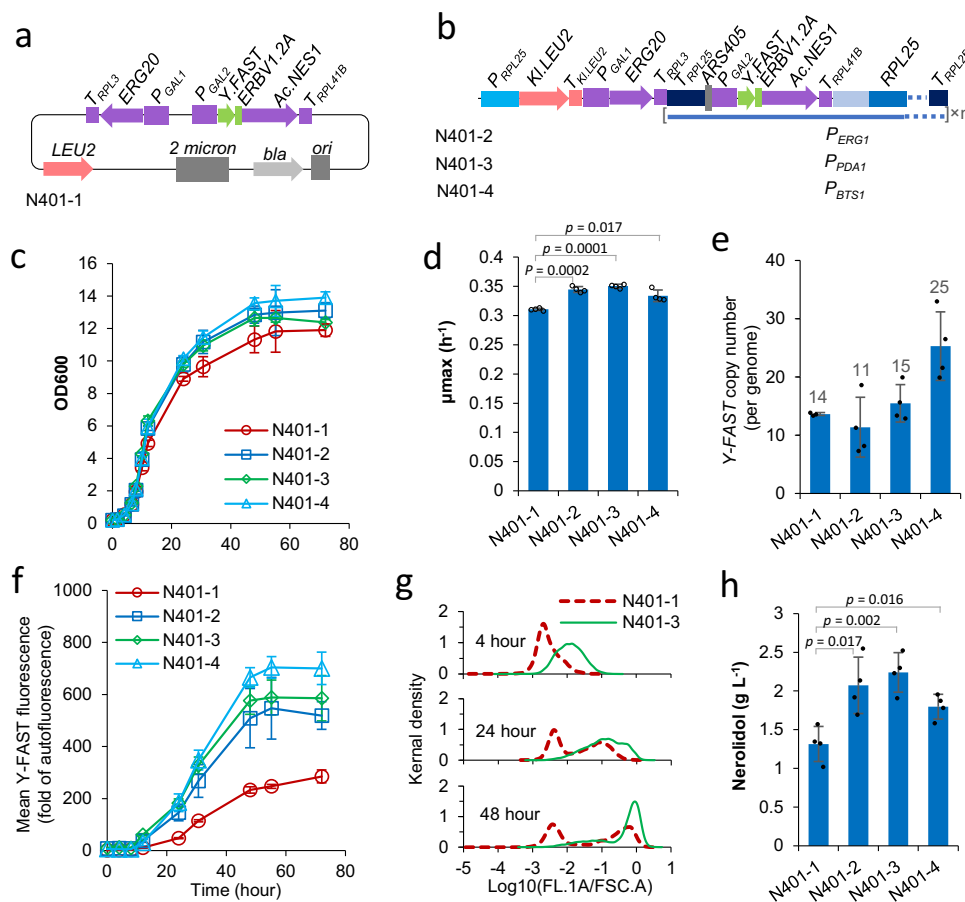
Compared to the reference strain N401-1, these three strains exhibited faster growth (Fig. 3c, d), higher *Y-FAST* fluorescence (Fig. 3f), and higher nerolidol production (Fig. 3h). The *Y-FAST-2A-Ac.NES1* cassette was successfully amplified in vivo in the three test strains (Fig. 3e).

The reference 2 $\mu$  plasmid strain harboured 14 copies of the *Y-FAST-2A-Ac.NES1* construct, similar to strain N401-3, and higher than that in strain N401-2. However, N401-1 had the lowest *Y-FAST* fluorescence (Fig. 3f). The discrepancy between copy number and fluorescence was due to lack of induction of *Y-FAST* expression in a large proportion of N401-1 cells (Fig. 3g). In contrast to the 2 $\mu$  plasmid strain, the strains harbouring the in vivo amplification constructs showed better synchronicity for *Y-FAST* induction (Fig. 3g N401-3; others not shown). This may contribute to the improved production.

**Improving heterologous production of the monoterpene limonene.** We next tested the system on production of monoterpenes ( $C_{10}$ ). Monoterpene production requires introduction of a dedicated  $C_{10}$  geranyl pyrophosphate (GPP) synthase<sup>42</sup>. We have previously used an *Erg20p<sup>N127W</sup>* mutant<sup>42</sup>, which excludes the  $C_{15}$  chain from the active site to generate a GPP pool, in combination with targeted degradation of the endogenous  $C_{15}$  synthase *Erg20p* via protein degron tags<sup>35,39</sup> to decrease competition at the  $C_{10}$  node by *Erg20p* and redirect GPP towards monoterpene production. In mevalonate pathway-enhanced strains, this approach delivered less than 100 mg l<sup>-1</sup> monoterpene—an order of magnitude below the levels achieved for sesquiterpene engineering.

We used a mevalonate pathway-enhanced strain with the endogenous *Erg20p* under an auxin-inducible protein degradation mechanism<sup>39</sup> as a background strain to minimise flux competition through the native sterol pathway. Two different promoter constructs were developed for amplification of the limonene synthetic module (Fig. 4a). The amplified region contained a fusion of multiple genes: *Y-FAST-2A*<sup>39</sup>, the maltose-binding protein from *E. coli* for improved solubility<sup>43</sup>, a short linker, limonene synthase from *Citrus limon*<sup>35</sup>, a 6\*glycerine linker, and the *Erg20p<sup>N127W F96W</sup>* mutant<sup>42</sup> (which has a higher specific GPP production rate than the *Erg20p<sup>N127W</sup>* mutant) as a GPP synthase. This fusion construct was under the control of the *GAL2* promoter from *S. kudriavzevii*<sup>44</sup>. The two constructs were transformed into the *RPL25* locus in the background strain, delivering strains LIM141M (*P<sub>PDA1</sub>*) and LIM141MH (*P<sub>BTS1</sub>*).

For the reference strain, the construct was introduced into the background strain via a 2 $\mu$  plasmid (Fig. 4a). We characterised four biological replicates (LIM141R representing three biological replicates and LIM141R2 representing one biological replicate; Fig. 4). In this case, 2 $\mu$  plasmid delivered ~2 copies per genome of the limonene synthase/*Y-FAST* module (shown by *Y-FAST* copy number; Fig. 4c). LIM141R, the three biological replicates produced ~40 mg l<sup>-1</sup> limonene (Fig. 4f), the titre same to a previous strain LIM141 expressing limonene synthase and *Erg20p<sup>N127W</sup>* without gene fusion<sup>39</sup>. However, one biological replicate (LIM141R2, Fig. 4) produced ~300 mg l<sup>-1</sup> limonene. LIM141R2 exhibited faster growth and higher *Y-FAST* fluorescence levels than other three biological replicates (LIM141R, Fig. 4b, d, e). The improvement in LIM141R2 may be caused by unintended genetic variations.



**Fig. 3** Characterisation of nerolidol-producing strains, harbouring nerolidol synthetic genes on a 2 $\mu$  plasmid (N401-1) or integrated at amplified RPL25 locus (N401-2, N401-3, and N401-4). **a, b** Schematic map of genetic vectors used to introduce nerolidol synthetic genes into yeast. **c–h** Strain characterisation in two-phase flask cultivation with 20 g l<sup>-1</sup> glucose and dodecane overlay. Y-FAST (fluorescence-activating and absorption-shifting tag) fluorescence was measured after 4-hydroxy-3-methylbenzylidene rhodanine (HMBR) with final concentration 20  $\mu$ M was added to the yeast samples before flow cytometry assay, and is expressed as fold-change of exponential-phase auto-fluorescence of the reference strain GH4<sup>3</sup>. Nerolidol production at 72 h was shown. Kernel density was calculated with bandwidth equal to 0.05. Mean values  $\pm$  standard deviations are shown (**c–f, h**;  $N = 4$  independent biological replicates). Two-tailed Welch's  $t$ -test was used for comparing two groups, and  $p$  values were shown in **d, h**. Source data are provided as a Source Data file.

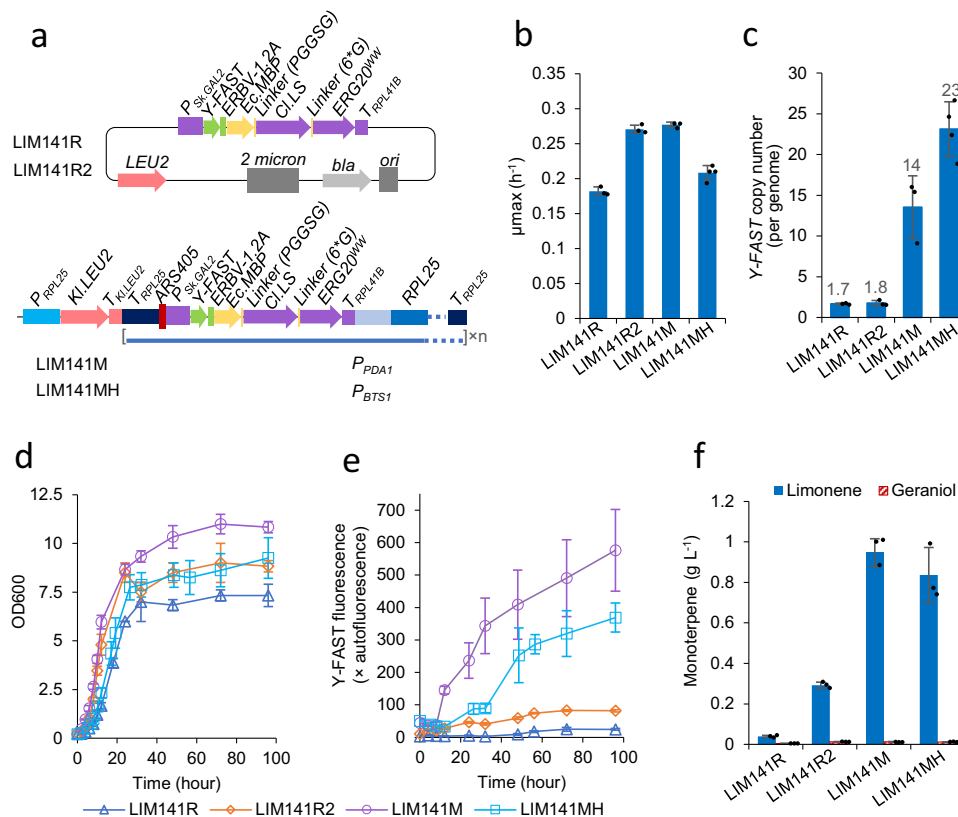
Harbouring HapAmp limonene synthetic module, both strains LIM141M and LIM141MH produced an order of magnitude more limonene than LIM141R and previous efforts using 2 $\mu$  plasmids<sup>35,39</sup>, with the best production,  $\sim 0.95$  g l<sup>-1</sup> limonene at 96 h, by strain LIM141M (Fig. 4f). This titre is 5.6-fold higher than the previous highest titre ever obtained in yeast<sup>45</sup>, and  $\sim 2$ -fold higher than the best titres achieved in batch cultivation in *E. coli*<sup>46,47</sup>. Strain LIM141MH showed a slower exponential growth and the lower levels of Y-FAST fluorescence compared to strain LIM141M (Fig. 4b, d, e), despite having more copies of the limonene synthase/Y-FAST module (shown by Y-FAST copy number; Fig. 4c). Both strains also accumulated  $\sim 12$  mg l<sup>-1</sup> of the monoterpene alcohol geraniol, which is commonly produced by yeast with an increased GPP pool<sup>35,39</sup>. No farnesol (C<sub>15</sub> alcohol) or geranylgeraniol (C<sub>20</sub> alcohol) were accumulated by the strains, indicating that subcellular pools of FPP and the C<sub>20</sub> geranylgeranyl pyrophosphate (GGPP) were low, and that amplification of limonene synthetic module led to significant redirection of the carbon flux towards monoterpene production.

**Improving heterologous tetraterpenoid lycopene production in yeast.** A three-gene lycopene synthetic module controlled by *GAL* promoters was previously constructed in a 2 $\mu$  plasmid<sup>37</sup> (Fig. 5a). This construct includes the farnesyl pyrophosphate mutant gene

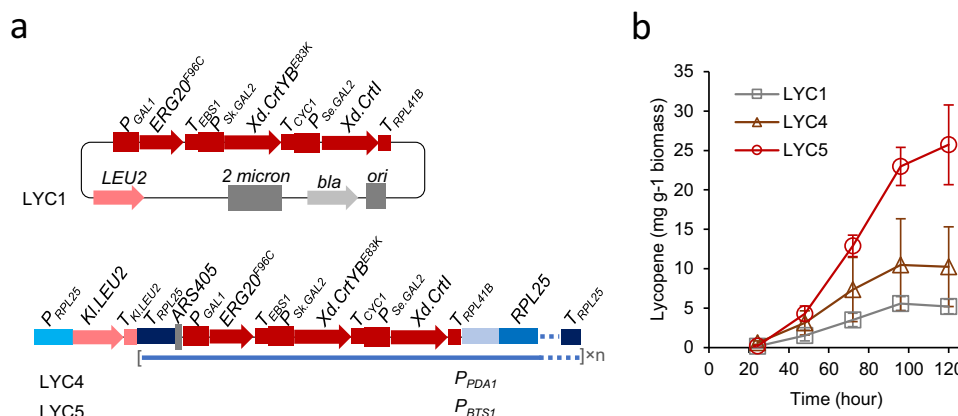
*ERG20*<sup>F96C</sup> which produces GGPP<sup>48</sup>, a phytoene synthase<sup>49,50</sup>, and a lycopene-forming phytoene desaturase mutant<sup>50</sup>. This plasmid was transformed into a mevalonate pathway-enhanced background strain, generating strain LYC1<sup>37</sup>. This strain accumulated  $\sim 5$  mg lycopene per gram of biomass in 120-h flask cultivation (Fig. 5b).

The lycopene synthetic module was sub-cloned into both the *PDA1* and *BTS1* promoter RPL25-driving HapAmp vectors (Fig. 5a). The resulting constructs were transformed into the same background strain, generating strains LYC4 and LYC5, respectively. Strain LYC4 (*P<sub>PDA1</sub>-RPL25*) accumulated slightly more lycopene than strain LYC1, although the increase was not significant (Fig. 5b). Strain LYC5 accumulated  $\sim 25$  mg lycopene per gram of biomass, five-fold higher than strain LYC1 (Fig. 5b).

**High-level expression of heterologous proteins in yeast.** *S. cerevisiae* can be used as a platform organism for protein production, including production of pharmaceutical proteins. However, a notorious disadvantage is that heterologous proteins production is not as high as what is achievable with *E. coli* expression systems. The high-level expression in *E. coli* can be attributed to the usage of high-copy-number plasmids (such as the common pET vectors with copy number about  $\sim 15$ – $20$ ) and the use of a very strong inducible



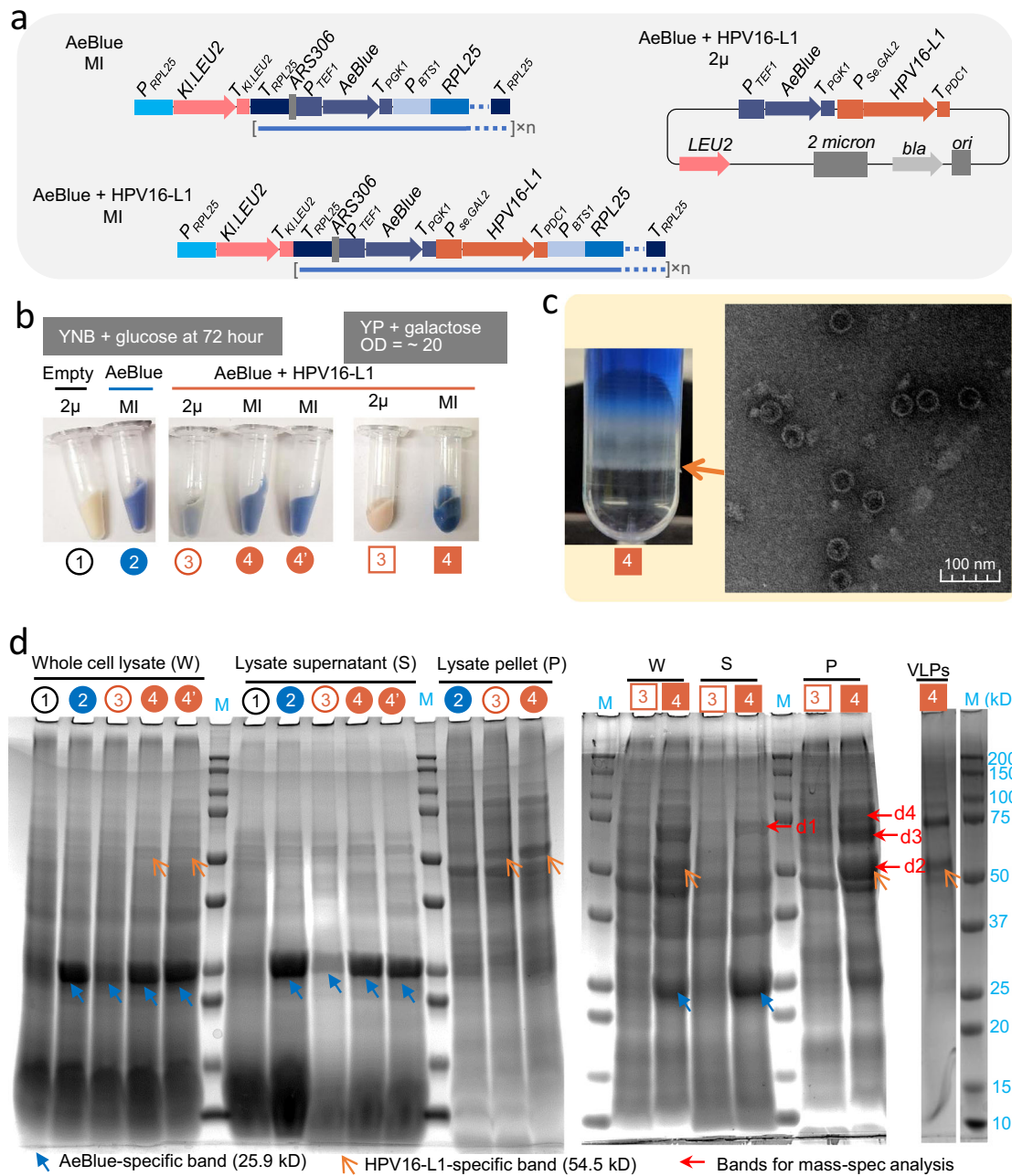
**Fig. 4** Characterisation of limonene-producing strains with limonene synthetic genes on a 2 $\mu$  plasmid (LIM141R and LIM141R2) or integrated at amplified RPL25 locus (LIM141M and LIM141MH). LIM141R2 is one of LIM141R biological replicates. **a** Schematic map of genetic vectors used to introduce limonene synthetic genes into yeast. **b–f** Strain characterisation in two-phase flask cultivation with 20 g l<sup>-1</sup> glucose and dodecane overlay. Synthetic auxin 1-Naphthaleneacetic acid (NAA) was added to 1 mM at the late exponential growth phase (OD > 4). Y-FAST fluorescence was measured after 4-hydroxy-3-methylbenzylidene rhodanine (HMBR) with final concentration 20  $\mu$ M was added to the yeast samples before flow cytometry assay and is expressed as fold-change of exponential-phase auto-fluorescence of the reference strain GH4<sup>30</sup>. Limonene and geraniol production at 96 h was shown. Mean values  $\pm$  standard deviations are shown ( $N = 3$  independent biological replicates for LIM141R, LIM141M and three independent cultures for LIM141R2 in **b–f**.  $N = 4$  independent biological replicates in **b–e** and three independent biological replicates in **f** for LIM141MH). Source data are provided as a Source Data file.



**Fig. 5** Characterisation of lycopene-producing strains with lycopene synthetic genes integrated at amplified RPL25 locus. **a** Schematic maps of genetic vectors used to introduce lycopene synthetic genes into yeast. **b** Lycopene production in flask cultivations. Yeast cells in exponential growth was inoculated into 20 ml MES-buffered YNB medium with 20 g l<sup>-1</sup> glucose in 125 ml Erlenmeyer flask to start a culture at OD<sub>600</sub> = 0.2. Mean values  $\pm$  standard deviations are shown ( $N = 4$  independent biological replicates). Source data are provided as a Source Data file.

promoter<sup>51</sup>. We used the  $P_{BTS1}$ -RPL25-driving HapAmp constructs to introduce the *AeBlue* chromoprotein gene<sup>52</sup> (Fig. 6a) or the *EforRed* chromoprotein gene<sup>53</sup>. Blue or pink colonies were obtained on the transformation plates (Supplementary Fig. 7), indicating high-level expression of the chromoproteins.

Having confirmed that the chromoproteins were effective markers, we then inserted a human papillomavirus (HPV) 16 major capsid protein L1 gene after the *AeBlue* expression cassette (Fig. 6a) to test the system for production of a pharmaceutical protein. For a reference, we cloned *AeBlue*-and-HPV16-L1



**Fig. 6 Characterisation of the expression of heterologous proteins (AeBlue and HPV16 capsid L1) via multi-copy genome integration (MI) using  $P_{BTSR}$ - $RPL25$ -driven *in vivo* gene amplification.** **a** Schematic of genetic vectors used to express AeBlue and HPV16 L1. **b** cells harbouring an empty 2 $\mu$ , the amplifiable AeBlue construct (MI), AeBlue-and-HPV16-L1 2 $\mu$  plasmid, and amplifiable AeBlue-and-HPV16-L1 construct (MI). Cells were grown in MES-buffered YNB medium with 20 g l<sup>-1</sup> glucose and collected at 72 h, or were grown in YP medium with 20 g l<sup>-1</sup> galactose to OD<sub>600</sub> = ~20. **c** Ultracentrifugation of the supernatant on an iodixanol gradient to separate a band containing HPV16-L1 virus-like particles (shown by orange arrow), and transmission electron microscopy confirming the presence of HPV16-L1 virus-like particles (VLPs). **d** SDS-PAGE (sodium dodecyl sulphate-polyacrylamide gel electrophoresis) for whole-cell lysates, lysate supernatant, and lysate pellets of yeast samples in **b**, and VLPs sample from **c**. Experimental repetition is not done for **c** and **d**. Numbers in **b-d** are for sample cross-reference. The bands d1, d2, d3, and d4 are analysed using a LC-MS/MS-based proteomic method (Supplementary Method 1), and the data are available in Supplementary Data 1 (d1), Supplementary Data 2 (d2), Supplementary Data 3 (d3), and Supplementary Data 4 (d4). Source data for VLPs in **d** are provided as a Source Data File.

expression cassettes into a yeast 2 $\mu$  plasmid (Fig. 6a). To compare the efficiency of protein production in different systems, an empty 2 $\mu$  plasmid, the AeBlue-and-HPV16-L1 2 $\mu$  plasmid, the  $RPL25$ -amplifiable AeBlue construct, and the  $RPL25$ -amplifiable AeBlue-and-HPV16-L1 construct were transformed individually into CEN.PK (*gal80 $\Delta$* ). The four resulting strains were grown in MES-buffered YNB medium with 20 g l<sup>-1</sup> glucose aerobically for

72 h. Cells with multi-copy integration of the AeBlue expression cassette showed a strong Tibetan blue colour, while cells with an empty cassette were milky white colour (Fig. 6b). The cells with 2 $\mu$  plasmid containing AeBlue + HPV-L1 expression cassettes were a faint blue colour, whereas the cells with multi-copy integration of AeBlue + HPV-L1 expression cassettes displayed the strong Tibetan blue colour (Fig. 6b). This indicated superior

expression capacity from the *in vivo* amplification method for multi-copy genome integration, compared to conventional 2 $\mu$  plasmid method.

SDS-PAGE of whole-cell and soluble protein extracts showed bands at ~25 kD (AeBlue molecular weight) in all samples, with much stronger bands observed in the multi-copy integration strain samples than in the 2 $\mu$  plasmid strain samples (Fig. 6d). In the multi-copy integration strains, these bands represented ~3% of whole-cell protein, suggesting heterologous protein expression in yeast may reach the levels often obtained in *E. coli*.

A second strong band at ~50 kD band (HPV16-L1 molecular weight) was observed in samples from cells expressing HPV-L1, although it was not as distinct at the putative AeBlue band (Fig. 6d). This may be due to the use of the *Se.GAL2* promoter, which is not fully induced in the ethanol phase, in these constructs compared to the constitutive *ALD6* promoter used for the AeBlue expression cassette. Again, the bands in the multi-copy integration strain samples were stronger than the 2 $\mu$  plasmid samples. Surprisingly, considering that HPV16-L1 is a soluble protein<sup>54</sup>, these bands were not distinguishable in lysate supernatant samples.

To fully induce the *Se.GAL2* promoter for HPV16-L1 expression, we attempted to grow the plasmid and integration strains harbouring HPV16-L1 in synthetic minimal medium (YNB) with ethanol or galactose as the carbon source. However, these cultivation conditions were lethal for the multi-copy-integration cells. We then grew the cells in rich (yeast-peptone (YP)) medium with 20 g l<sup>-1</sup> galactose as the carbon source. Under these conditions, AeBlue expression from 2 $\mu$  plasmid was not observable by visual examination (Fig. 6b) or SDS-PAGE (Fig. 6d). This may be due to loss of 2 $\mu$  plasmid in the rich medium. In contrast, strong AeBlue-specific and HPV16-L1-specific bands were seen in whole-cell lysate and lysate supernatant samples from the cells with multi-copy integration constructs. This further confirmed that HPV16 L1 capsid protein is insoluble in yeast in our system. Attempts to solubilise HPV16-L1 capsid protein were unsuccessful (data not shown). Despite being unable to detect HPV16-L1-specific bands in lysate supernatant (Fig. 6d), we could still separate properly assembled virus-like particles (VLPs) by ultracentrifugation of lysate supernatant (Fig. 6c). SDS-PAGE examination of VLP components purified from ultracentrifugation showed a HPV16-L1-specific band at ~50 kD (Fig. 6d; Lane VLPs:4). TEM images of the VLPs showed their diameter was around 40 nm (Fig. 6c), consistent with previous literature<sup>55</sup>.

In SDS-PAGE results, we observed strong bands in the lysate supernatant sample (band d1) and lysate pellet samples (bands d2, d3, and d4) (Fig. 6d). LC-MS/MS-based proteomics was used to analyse the protein composition in these four bands (Supplementary Data 1–4). The top hit protein in the ~50 kD band (band d2) was the HPV16 L1 capsid. Interestingly, the top hit proteins in other three bands (d1, d3 and d4) were yeast chaperones. In bands d1 and d3, the top hit proteins were HSP70 family chaperone Ssa1, and in bands d4, the top hit protein was HSP90 family chaperone Hsc82. We therefore hypothesised that insoluble expression of HPV16-L1 caused upregulation of yeast chaperones, and HPV16-L1, HSP70 chaperones, and HSP90 chaperones might exist in insoluble forms. However, it would require further systematic examination to get a better understanding of these phenomena.

In summary, although some insoluble expression of the HPV16 L1 was observed, our results both with chromoprotein AeBlue and the HPV16 L1 showed that multi-copy gene integration via HapAmp method can lead to heterologous protein overexpression in yeast to the high levels that are commonly seen in *E. coli* expression systems.

## Discussion

Here, we developed a genetic engineering method to integrate multiple copies of heterologous gene(s) into the yeast genome using *in vivo* gene amplification driven by a haploinsufficient gene (HapAmp). The functional strength per copy of a haploinsufficient gene is strongly associated with growth fitness, which can be exploited as an evolutionary force to drive gene amplification. Decreased expression level provides an evolutionary force that drives amplification of linked haploinsufficient and heterologous genes, so that cells are growth competitive. We exemplified the application of this method to improve production of different types of terpene products. We also showed that our method enabled high-level expression of heterologous protein in yeast, at levels similar to that achieved in *E. coli* for protein production.

This method presents three main advantages for the introduction of heterologous genes via genome integration. Firstly, integration copy number can be titrated by altering expression dosage per copy of haploinsufficient gene. Expression level can be reduced by a variety of methods. Here, we tested two approaches: (1) replacing the gene promoter with a weaker promoter (Figs. 2–4), and (2) using non-preferred codons (Fig. 2). In these experiments, we observed a range of between 4 and 47 copies, with an inverse relationship between promoter strength and copy number. We characterised a range of weak promoters here (Supplementary Fig. 8) and in previous work<sup>3</sup> that can be applied to decrease gene dosage. In addition to promoter strength and codon usage, other approaches could be used to decrease expression dosage, including engineering the kozak sequence and/or the 5'-mRNA structure. These genetic tools add engineering flexibility to modify copy number for this HapAmp method in yeast.

Secondly, the maintenance of integration is auto-selectable: selection pressure is provided from the dosage sensitivity of the haploinsufficient gene, which is linked to the gene of interest and is maintained to support normal growth rates. This means that no antibiotics or modification of other environmental conditions in the culture are required to provide ongoing selection pressure for maintenance of the gene of interest. Compared to use of a 2 $\mu$  plasmid, this method provides more stable expression of heterologous proteins in yeast (Fig. 6b). In addition, it does not require chemical induction for amplification<sup>2,15</sup>.

Thirdly, the presence of multiple haploinsufficient genes means that many different loci are available for engineering gene amplification. We demonstrated the method using *RPL25* and *SEC23* as the driving gene. We further characterised the promoter strength of fifteen additional haploinsufficient genes (Supplementary Fig. 8) that can also be used to drive gene amplification.

Initial integration of the genes of interest uses standard yeast transformation procedures by selection of an auxotrophic or antibiotic marker (e.g. *LEU2* or *hphMax* in Figs. 2–6). Upon transformation, we observed a variable proportion of false clones (not expressing the gene of interest) on the transformation plates (Supplementary Figs. 1 and 5). We presume that, in these cases, either spontaneous mutations have provided the yeast an alternative mechanism to recover growth rate, or the gene of interest was not correctly integrated into the target locus. Use of visual markers (fluorescent proteins or chromoproteins) can facilitate the selection of correct clones with amplified constructs. In the absence of such visual markers, characterisation or verification of a pool of clones would be necessary to select clones with multi-copy integration of heterologous genes. Further optimising the genetic background of the yeast strains used, such as eliminating the non-homologous end joining mechanism to decrease non-homologous gene integration, might be useful to eliminate rate of false positives for the current method.



The HapAmp method successfully improved production of heterologous terpenes including the  $C_{15}$  sesquiterpene nerolidol (Fig. 3), the  $C_{10}$  monoterpene limonene (Fig. 4), and the  $C_{40}$  tetraterpene lycopene (Fig. 5). Production of  $C_{15}$  terpenes in yeast is typically relatively straightforward, with gram per litre titres achievable<sup>39,56</sup>. This is likely because the  $C_{15}$  precursor, FPP, is produced in yeast naturally to deliver sterol pathway products required for yeast growth. In addition, sesquiterpene synthases have reasonably good catalytic properties, making them more competitive to access FPP. Production of  $C_{10}$  monoterpenes, however, has historically been very challenging. This is due to both a dearth of  $C_{10}$  precursors<sup>57</sup> and the poor catalytic properties of many monoterpene synthases<sup>45,58</sup>. These limitations have previously restricted published titres of monoterpenes to  $\text{mg l}^{-1}$  in flask cultivation<sup>35,39,45,59</sup>. Here, we have achieved  $\text{g l}^{-1}$  titres (Fig. 4) in a single engineering step using a high mevalonate pathway flux strain with an introduced GPPS and targeted degradation of FPPS to decrease competition at the  $C_{10}$  pathway node. We believe this is the highest titre achieved in metabolically engineered microbes in a flask cultivation with  $20 \text{ g l}^{-1}$  glucose as carbon source reported to date.

Interestingly, one replicate of the monoterpene control strain produced  $\sim 300 \text{ mg l}^{-1}$  limonene, in comparison to the other three replicates which produced  $\sim 40 \text{ mg l}^{-1}$  limonene, despite the plasmid copy number being the same in all four replicates (Fig. 4). This suggests that an unintended mutation has arisen in this strain which affects limonene production positively. The source of this variation will be examined in future work and may form the basis of further engineering efforts.

We observed a tight correlation between gene copy number and GFP fluorescence (Fig. 2); however, this relationship breaks down for the different terpene products, resulting in variable improvement ratios. This is most likely due to the fact that the relationship between the GFP peptide and its fluorescence is very close and does not rely on other factors such as substrate and cofactor availability—whereas the terpene synthases are enzymes and subject to more influences on their behaviour. In addition, variable metabolic burden caused by overexpression of terpene synthases or other physiological perturbations in metabolically engineered systems may affect the relationship between copy number and product titre. For products, limonene production improvement was  $\sim 24$ -fold, whereas nerolidol improvement was 1.7-fold, and lycopene improvement was 5-fold. However, we always obtained a higher titre by *in vivo* gene amplification. In particular, for monoterpenes, insufficient catalytic efficiency of terpene synthase is a significant bottleneck for production of heterologous terpenoids in yeast. Increasing copy number via insertion of tandem repeats at the same locus combined with screening for improved production<sup>56</sup> or introduction of additional expression cassettes at separate loci<sup>43</sup> has been used to overcome this bottleneck previously. However, these approaches require complex cloning and extended experimental timelines to deliver the desired improvements. The HapAmp system provides a faster and simpler method to achieve superior results.

We tested several constructs ranging up to three expression cassettes (lycopene pathway: insert size of 7917 bp). We have not sought to test the maximum cargo size for this approach. However, we did not observe a clear relationship between size of the insert ('cargo') and copy number amplification, suggesting that even larger inserts may be possible for the technique.

In addition to its application in metabolic engineering, we also examined the potential of HapAmp for increasing heterologous protein production. Using chromoprotein AeBlue and the HPV16 L1 capsid protein as examples (Fig. 6), we demonstrated that in *S. cerevisiae*, heterologous protein could be produced at levels commonly seen in *E. coli*. AeBlue was expressed in soluble form,

whereas HPV16 L1 capsid protein was primarily expressed in insoluble form. Insoluble expression of HPV16 L1 capsid protein has been reported in *E. coli*<sup>60–62</sup> but not in *S. cerevisiae*. In *E. coli*, N-terminal truncation<sup>60,61</sup>, use of a fusion partner<sup>61</sup>, and overexpression GroEL/GroES chaperones<sup>62</sup> (which accept broader substrates than cytosolic chaperones in *S. cerevisiae*<sup>63,64</sup>), improved soluble expression of HPV L1 capsid proteins. These strategies might also improve soluble expression of HPV capsid proteins in yeast.

The HapAmp method should be applicable in other industrially relevant chassis organisms that have haploinsufficient genes. A potential haploinsufficient gene may encode essential components of the machineries for protein synthesis and transportation or other essential cell structures<sup>28</sup>. Putative haploinsufficient genes can be identified by comparative genomics and confirmed by testing growth fitness in association with expression dosage of a gene. For diploid organisms, this can be done by disrupting one allele and integrating the amplifiable construct at the other allele locus, or by simultaneously integrating the amplifiable constructs at both alleles. In addition, native non-homologous end joining mechanisms can be diminished/disrupted to improve the successful rate of amplification of genes of interests<sup>65</sup>. A nuclease-mediated DNA double-chain break like CRISPR<sup>66</sup> could also be used to assist the integration of the amplifiable construct. This may avoid the use of a selectable marker in the gene amplification construct.

## Methods

**Plasmid and strain construction.** Plasmids used in this work are listed in Supplementary Data 5, and strains are listed in Supplementary Data 6. Primers used in polymerase chain reaction (PCR) and PCR performed in this work are listed in Supplementary Data 7. Plasmid construction processes are listed in Supplementary Data 8. Yeast strain construction processes are listed in Supplementary Data 9. A LiAc/SS carrier DNA/PEG method<sup>67</sup> was used for yeast transformation.

**Yeast cultivation.** For characterisation of yEGFP-expressing strains, yeast cells from glycerol stocks were streaked on YNB-glucose agar, which comprised of  $6.9 \text{ g l}^{-1}$  yeast nitrogen base without amino acids (YNB, FORMEDIUM#CYN0402) with pH adjusted to 6.0 using sodium hydroxide solution,  $20 \text{ g l}^{-1}$  glucose, and  $20 \text{ g l}^{-1}$  agar. MES-buffered YNB-glucose medium was used in following cultivation, which comprised of  $19.5 \text{ g l}^{-1}$  2-(N-morpholino)ethanesulfonic acid (MES),  $6.9 \text{ g l}^{-1}$  YNB,  $20 \text{ g l}^{-1}$  glucose, and its pH was adjusted to 6.0 with ammonia hydroxide solution. For the growth in flask, seed cultures grown to the exponential phase ( $\text{OD}_{600} \leq 4$ ) were inoculated into 20 ml MES-buffered YNB-glucose medium in 125 ml Erlenmeyer flasks to start the cultivation in a 200 rpm  $30^\circ\text{C}$  incubator. For the growth in 96-well microplate, yeast cells were grown in YNB-glucose medium ( $6.9 \text{ g l}^{-1}$  YNB,  $20 \text{ g l}^{-1}$  glucose, pH 6.0) for about 20 h to stationary phase in a 350 rpm  $30^\circ\text{C}$  incubator to prepare seed culture. Seed culture (5  $\mu\text{l}$ ) was inoculated into 100  $\mu\text{l}$  MES-buffered YNB-glucose medium to prepare Culture 1. Culture 1 (2  $\mu\text{l}$ ) was inoculated into 100  $\mu\text{l}$  MES-buffered YNB-glucose medium to prepare Culture 2. Culture 2 was incubated in a 350 rpm  $30^\circ\text{C}$  incubator overnight for analysis of yEGFP fluorescent in the cells grown to the exponential growth phase, and Culture 1 for two nights for analysis in the cells grown to the ethanol growth phase.

For characterisation of nerolidol/limonene-producing strains, dodecane-overlaid two-phase flask cultivation was used. Yeast cells from glycerol stocks were streaked on YNB-high-glucose agar, which contained  $6.9 \text{ g l}^{-1}$  YNB (pH 6.0),  $200 \text{ g l}^{-1}$  glucose, and  $20 \text{ g l}^{-1}$  agar. Before initiating the two-phase flask cultivation, cells were pre-cultured in MES-buffered YNB- $20 \text{ g l}^{-1}$  glucose to exponential phase ( $\text{OD}_{600}$  between 1 to 4) and collected by centrifugation. Collected cells were then resuspended in fresh fermentation medium. To initiate the cultivation, appropriate volumes of pre-cultured cells were transferred to MES-buffered YNB medium with  $20 \text{ g l}^{-1}$  glucose to an initial  $\text{OD}_{600}$  of 0.2 in a total volume of 23 ml medium in a 250 ml flask, and 2 ml sterile dodecane was added after inoculation. In the first 12 h of cultivation, 3 ml culture was sampled for growth curve measurement. Dodecane was sampled and stored at  $-80^\circ\text{C}$  for terpene analysis.

Flask cultivations for lycopene-producing strains were prepared as the flask cultivation used for yEGFP-expressing strains. For chromoprotein/HPV16 L1-expressing strains, yeast cells grown overnight in 5 ml MES-buffered YNB-glucose medium were inoculated into 20 ml fresh MES-buffered YNB-glucose medium or 20 ml YP-galactose ( $20 \text{ g l}^{-1}$  peptone,  $10 \text{ g l}^{-1}$  yeast extract, and  $20 \text{ g l}^{-1}$  galactose) to start characterisation cultures.

**Flow cytometry.** A BD Accuri™ C6 flow cytometer (BD Biosciences, USA) was used for fluorescence analysis in single cells. Cells expressing yEGFP were sampled and directly used for characterisation of the yEGFP fluorescence. Cells expressing Y-FAST were sampled and mixed with 20 µM HMBR (synthesised and prepared in 2 mM stock in dimethyl sulfoxide<sup>40</sup>) before analysis. Debris particles were excluded through an FSC.H threshold with the threshold value of 250,000. A 488 nm laser was used to excite GFP and Y-FAST fluorescence. The detector equipped with a 530/20 bandpass filter was used to monitor the fluorescence (FL1.A). For each sample, 10,000 events were recorded. A BD C sampler software (BD Accuri C6 software version 1.0.264.21) were used to extract mean values of FSC.A, SSC.A, and FL1.A. The fluorescence level of GFP and Y-FAST was expressed as the fold of a background fluorescence in the exponential grown phase cells of strain GH4<sup>3</sup>.

**Metabolite analysis.** HPLC analysis was performed by the Metabolomics Australia (Queensland node) using a previously described method<sup>68</sup>. In brief, an Agilent 1200 HPLC system and a Thermo Fisher Chromleon Chromatography Data System software were used. Dodecane samples in some cases were diluted with dodecane before HPLC analysis. For HPLC analysis, 5 µl dodecane samples (or standards prepared in dodecane) were mixed with 200 µl ethanol, and 20 µl mixture was injected and separated with a guard column (SecurityGuard Gemini C18, Phenomenex PN: AJO-7597) and a Zorbax Extend C18 column (4.6 × 150 mm, 3.5 µm, Agilent PN: 763953-902). The mixture of solvent A (water) and solvent B (45% acetonitrile, 45% methanol, and 10% water) was used to elute the analytes with a linear gradient (from 0–24 min, 5–100% solvent B; from 24–30 min, 100% solvent B; from 30.1–35 min, 5% solvent B).

For lycopene measurement, yeast cells were collected and resuspended in 200 µl 2 M l<sup>-1</sup> sodium hydroxide and vortexed with 200 mg glass bead and 1 ml hexane for at least 10 min. Lycopene molar extinction coefficient (182 × 10<sup>3</sup>) at 471 nm was used to calculate lycopene concentration<sup>69</sup>. In some cases, lycopene extracts were diluted with hexane to make the absorbance reading <0.6.

**Protein purification.** Yeast cells were homogenised by vortexing with glass beads for 15 min in phosphate-buffered saline (PBS) buffer plus 2 mM ethylenediaminetetraacetic acid. Whole-cell lysates, lysate supernatants, and lysate pellets were examined by sodium dodecyl sulphate-polyacrylamide gel electrophoresis analysis on Mini-PROTEAN® Precast Gels (Bio-rad).

The lysis was followed by centrifugation at 18,000 × g for 30 min to pellet the cellular debris. The soluble fraction was then loaded on top of a gradient made of 1 ml of 20% Iodixanol/PBS buffer, 1 ml of 30% Iodixanol/PBS and 1 ml of 40% Iodixanol/PBS in a Thinwall Ultra-Clear Tube (Beckman Coulter, Indianapolis, USA) and subjected to ultracentrifugation for 2 h 30 min at 150,000 × g on a SW41 Ti rotor or a using a Beckman Optima L-100XP ultracentrifuge (Beckman Coulter, Indianapolis, USA). A band containing the VLPs encapsulating protein was extracted using a 1 ml syringe by poking a whole through the tube. Bradford was used to measure protein concentration and sample was further examined on TEM and purity confirmed on Mini-PROTEAN® Precast Gels (Bio-rad).

**Transmission electron microscopy.** Samples containing purified VLPs of 0.1 mg ml<sup>-1</sup> were applied to formvar/carbon coated grids (ProSciTech Pty Ltd, Australia) and incubated for 2 min. Grids were then washed with 40 µl of distilled water for 30 s twice, and then stained with 20 g l<sup>-1</sup> uranyl acetate for 1 min, after being blotted on filter paper. Images were taken on a HITACHI HT7700 transmission electron microscope at accelerating voltage of 80 keV at the Centre for Microscopy and Microanalysis.

**Genome sequencing.** Yeast genomic DNA was extracted using MagAttract HMW DNA Kit (Qiagen) with a modified protocol. Yeast cells (20 ml, OD<sub>600</sub> around 10) were washed once using PBS buffer and resuspended in 2 ml 1 M sorbitol solution. Yeast cell walls were digested by adding 30 U Zymolyase-20T (nacalai, Japan; 1 U per µl in 1\* PBS containing 100 mM DTT and 50% v/v glycerol) at 30 °C for 30 min. Yeast protoplast cells were collected and resuspended in 300 µl Buffer AL (MagAttract HMW DNA Kit) by pipetting using wide bore pipette tips, and then 360 buffer ATL (MagAttract HMW DNA Kit) was added and mixed. Following this, protocol provided in MagAttract HMW DNA Kit (Qiagen) was adopted including digestion by Proteinase K and Rnase A and purification using magnetic beads. Genomic DNA was eluted using 400 µl Buffer AE (MagAttract HMW DNA Kit) and treated using 100 µl tris-saturated phenol (pH 8.0, Ameresco) by flickering and 100 µl chloroform was added and mixed. Upper-layer water phase was collected after centrifuging at 17,000 × g for 5 min and mixed with 1 ml ethanol. Magnetic beads (MagAttract HMW DNA Kit) were used to purify genomic DNA with twice 70% ethanol wash and elution in 50 µl water. Concentration of genomic DNA was quantified using Qubit Fluorometer and Qubit dsDNA BR Assay Kit (Thermo Fisher). Genomic DNA (500 ng) was used to prepare genome sequencing library using Rapid Barcoding Kit (SQK-RBK004, Oxford Nanopore) and sequenced using R9 flowcell MIN106D and MinION Mk1C (Oxford Nanopore). High-accurate base-calling was performed using ont-guppy-for-mk1c (version 4.2.3) installed MinION Mk1C (MinKNOW version 20.10.6). Galaxy Australia online server was used for data processing<sup>70</sup>. Collapse Collection (Galaxy Version 5.1.0) was used to combine fastq dataset into a single file. Nanoplot was used for

statistical analysis of MinION reads<sup>71</sup>. Canu assembler was used for genome sequence assembly<sup>72</sup>. Maker (Galaxy Version 2.31.11) was used to collect annotation evidence with input of *S. cerevisiae* gene sequences and heterologous gene sequences as ESTs input file<sup>73</sup>. miniMap2 was used to align trimmed reads outputted by Canu assembler against contigs outputted by Canu assembler<sup>74</sup>. JBrowse (version 1.16.10-desktop)<sup>75</sup> and Integrative Genomics Viewer (version 2.8.13)<sup>76</sup> were used to illustrate genome structure and read alignment.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

MinION whole genome sequencing raw-read data are achieved in NCBI BioProject database with submission ID PRJNA688119. Processed data for MinION genome sequencing are achieved in Zenodo (<https://zenodo.org/record/6378077#.YnPhi9rMI2w>; <https://doi.org/10.5281/zenodo.6378077>). Plasmids used in this study are available on request or on Addgene (Addgene IDs: 185870-185894) ([https://www.addgene.org/Claudia\\_Vickers/](https://www.addgene.org/Claudia_Vickers/)). Source data are provided with this paper.

Received: 11 November 2021; Accepted: 5 May 2022;

Published online: 24 May 2022

## References

- Vickers, C., Blank, L. & Kromer, J. Grand challenge commentary: chassis cells for industrial biochemical production. *Nat. Chem. Biol.* **6**, 875–877 (2010).
- Tyo, K. E., Ajikumar, P. K. & Stephanopoulos, G. Stabilized gene duplication enables long-term selection-free heterologous pathway expression. *Nat. Biotechnol.* **27**, 760–765 (2009).
- Peng, B., Williams, T., Henry, M., Nielsen, L. & Vickers, C. Controlling heterologous gene expression in yeast cell factories on different carbon substrates and across the diauxic shift: a comparison of yeast promoter activities. *Microb. Cell Factories* **14**, 91 (2015).
- Stephanopoulos, G. & Vallino, J. J. Network rigidity and metabolic engineering in metabolite overproduction. *Science* **252**, 1675–1681 (1991).
- Gnugge, R. & Rudolf, F. Saccharomyces cerevisiae Shuttle vectors. *Yeast* **34**, 205–221 (2017).
- Karim, A. S., Curran, K. A. & Alper, H. S. Characterization of plasmid burden and copy number in Saccharomyces cerevisiae for optimization of metabolic engineering applications. *FEMS Yeast Res.* **13**, 107–116 (2013).
- Eguchi, Y. et al. Estimating the protein burden limit of yeast cells by measuring the expression limits of glycolytic proteins. *eLife* **7**, e34595 (2018).
- Lopes, T. S., Hakkaart, G. J. A. J., Koerts, B. L., Raue, H. A. & Planta, R. J. Mechanism of high-copy-number integration of Pmiry-type vectors into the ribosomal DNA of Saccharomyces-Cerevisiae. *Gene* **105**, 83–90 (1991).
- Loison, G., Nguyen-Juillet, M., Alouani, S. & Marquet, M. Plasmid-transformed ura3 fur1 double-mutants of *S. cerevisiae*: an autoselection system applicable to the production of foreign proteins. *Bio/Technology* **4**, 433–437 (1986).
- Yoo, J. I., Seppälä, S. & O'Malley, M. A. Engineered fluoride sensitivity enables biocontainment and selection of genetically-modified yeasts. *Nat. Commun.* **11**, 5459 (2020).
- Song, X. et al. POT1-mediated delta-integration strategy for high-copy, stable expression of heterologous proteins in Saccharomyces cerevisiae. *FEMS Yeast Res.* **17**, fox064 (2017).
- Kawasaki, G. H. & Bell, L. Google Patents (1999).
- Maury, J. et al. EasyCloneMulti: a set of vectors for simultaneous and multiple genomic integrations in Saccharomyces cerevisiae. *PLoS ONE* **11**, e0150394 (2016).
- Yamada, R. et al. Cocktail delta-integration: a novel method to construct cellulolytic enzyme expression ratio-optimized yeast strains. *Microb. Cell Factories* **9**, 32 (2010).
- Lian, J. Z., Jin, R. & Zhao, H. M. Construction of plasmids with tunable copy numbers in Saccharomyces cerevisiae and their applications in pathway optimization and multiplex genome integration. *Biotechnol. Bioeng.* **113**, 2462–2473 (2016).
- Shi, S. B., Liang, Y. Y., Zhang, M. Z. M., Ang, E. L. & Zhao, H. M. A highly efficient single-step, markerless strategy for multi-copy chromosomal integration of large biochemical pathways in Saccharomyces cerevisiae. *Metab. Eng.* **33**, 19–27 (2016).
- Parekh, R. N., Shaw, M. R. & Wittrup, K. D. An integrating vector for tunable, high copy, stable integration into the dispersed Ty δ sites of Saccharomyces cerevisiae. *Biotechnol. Prog.* **12**, 16–21 (1996).

18. Tumen-Velasquez, M. et al. Accelerating pathway evolution by increasing the gene dosage of chromosomal segments. *Proc. Natl Acad. Sci. USA* **115**, 7105–7110 (2018).
19. Zhou, H., Cheng, J. S., Wang, B. L., Fink, G. R. & Stephanopoulos, G. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by *Saccharomyces cerevisiae*. *Metab. Eng.* **14**, 611–622 (2012).
20. Demeke, M. M., Foulquie-Moreno, M. R., Dumortier, F. & Thevelein, J. M. Rapid evolution of recombinant *Saccharomyces cerevisiae* for xylose fermentation through formation of extrachromosomal circular DNA. *PLoS Genet.* **11**, e1005010 (2015).
21. Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S. K. & Lemos, B. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proc. Natl Acad. Sci. USA* **112**, 2485–2490 (2015).
22. Fischer, U. et al. Gene amplification during differentiation of mammalian neural stem cells in vitro and in vivo. *Oncotarget* **6**, 7023–7039 (2015).
23. Schimke, R. T. Gene amplification in cultured animal-cells. *Cell* **37**, 705–713 (1984).
24. Oh, E. J. et al. Gene amplification on demand accelerates cellobiose utilization in engineered *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **82**, 3631–3639 (2016).
25. Fogel, S. & Welch, J. W. Tandem gene amplification mediates copper resistance in yeast. *Proc. Natl Acad. Sci. Biol.* **79**, 5342–5346 (1982).
26. Jack, C. V. et al. Regulation of ribosomal DNA amplification by the TOR pathway. *Proc. Natl Acad. Sci. USA* **112**, 9674–9679 (2015).
27. Morrill, S. A. & Amon, A. Why haploinsufficiency persists. *Proc. Natl Acad. Sci. USA* **116**, 11866–11871 (2019).
28. Deutschbauer, A. M. et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925 (2005).
29. Ganley, A. R. D., Ide, S., Saka, K. & Kobayashi, T. The effect of replication initiation on gene amplification in the rDNA and its relationship to aging. *Mol. Cell* **35**, 683–693 (2009).
30. Conti, C., Herrick, J. & Bensimon, A. Unscheduled DNA replication origin activation at inserted HPV 18 sequences in a HPV 18/MYC amplicon. *Gene Chromosom. Cancer* **46**, 724–734 (2007).
31. Lu, L., Zhang, H. & Tower, J. Functionally distinct, sequence-specific replicator and origin elements are required for *Drosophila* chorion gene amplification. *Genes Dev.* **15**, 134–146 (2001).
32. Newlon, C. S. et al. Analysis of replication origin function on chromosome III of *Saccharomyces cerevisiae*. *Cold Spring Harb. Symp. Quant. Biol.* **58**, 415–423 (1993).
33. Liachko, I., Youngblood, R. A., Keich, U. & Dunham, M. J. High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res.* **23**, 698–704 (2013).
34. Peng, B. et al. Engineering eukaryote-like regulatory circuits to expand artificial control mechanisms for metabolic engineering in *Saccharomyces cerevisiae*. *Commun. Biol.* **5**, 135 (2022).
35. Peng, B., Nielsen, L. K., Kampranis, S. C. & Vickers, C. E. Engineered protein degradation of farnesyl pyrophosphate synthase is an effective regulatory mechanism to increase monoterpene production in *Saccharomyces cerevisiae*. *Metab. Eng.* **47**, 83–93 (2018).
36. Peng, B., Plan, M. R., Carpenter, A., Nielsen, L. K. & Vickers, C. E. Coupling gene regulatory patterns to bioprocess conditions to optimize synthetic metabolic modules for improved sesquiterpene production in yeast. *Biotechnol. Biofuels* **10**, 43 (2017).
37. Peng, B., Wood, R. J., Nielsen, L. K. & Vickers, C. E. An expanded heterologous GAL promoter collection for diauxie-inducible expression in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **7**, 748–751 (2018).
38. Hayat, I. F. et al. Auxin-mediated induction of GAL promoters by conditional degradation of Mig1p improves sesquiterpene production in *Saccharomyces cerevisiae* with engineered acetyl-CoA synthesis. *Micro. Biotechnol.* **14**, 2627–2642 (2021).
39. Lu, Z., Peng, B., Ebert, B. E., Dumsday, G. & Vickers, C. E. Auxin-mediated protein depletion for metabolic engineering in terpene-producing yeast. *Nat. Commun.* **12**, 1051 (2021).
40. Plamont, M. A. et al. Small fluorescence-activating and absorption-shifting tag for tunable protein imaging in vivo. *Proc. Natl Acad. Sci. USA* **113**, 497–502 (2016).
41. Souza-Moreira, T. M. et al. Screening of 2A peptides for polycistronic gene expression in yeast. *FEMS Yeast Res.* **18**, foy036 (2018).
42. Ignea, C., Pontini, M., Maffei, M. E., Makris, A. M. & Kampranis, S. C. Engineering monoterpene production in yeast using a synthetic dominant negative geranyl diphosphate synthase. *ACS Synth. Biol.* **3**, 298–306 (2013).
43. Wong, J. et al. High-titer production of lathyranes diterpenoids from sugar by engineered *Saccharomyces cerevisiae*. *Metab. Eng.* **45**, 142–148 (2018).
44. Peng, B., Nielsen, L. K. & Vickers, C. E. An expanded heterologous GAL promoter collection for diauxie-inducible over-expression in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **7**, 748–751 (2017).
45. Ignea, C. et al. Orthogonal monoterpenoid biosynthesis in yeast constructed on an isomeric substrate. *Nat. Commun.* **10**, 3799 (2019).
46. Alonso-Gutierrez, J. et al. Metabolic engineering of *Escherichia coli* for limonene and perillyl alcohol production. *Metab. Eng.* **19**, 33–41 (2013).
47. Rolf, J., Julsing, M. K., Rosenthal, K. & Lutz, S. A Gram-scale limonene production process with engineered *Escherichia coli*. *Molecules* **25**, 1881 (2020).
48. Ignea, C. et al. Efficient diterpene production in yeast by engineering Erg20p into a geranylgeranyl diphosphate synthase. *Metab. Eng.* **27**, 65–75 (2015).
49. Xie, W., Lv, X., Ye, L., Zhou, P. & Yu, H. Construction of lycopene-overproducing *Saccharomyces cerevisiae* by combining directed evolution and metabolic engineering. *Metab. Eng.* **30**, 69–78 (2015).
50. Verwaal, R. et al. High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Appl. Environ. Microbiol.* **73**, 4342–4350 (2007).
51. Kuliopulos, A. & Walsh, C. T. Production, purification, and cleavage of tandem repeats of recombinant peptides. *J. Am. Chem. Soc.* **116**, 4599–4607 (1994).
52. Shkrob, M. A. et al. Far-red fluorescent proteins evolved from a blue chromoprotein from *Actinia equina*. *Biochem. J.* **392**, 649–654 (2005).
53. Alieva, N. O. et al. Diversity and evolution of coral fluorescent proteins. *PLoS ONE* **3**, e2680 (2008).
54. Carter, J. J., Yaegashi, N., Jenison, S. A. & Galloway, D. A. Expression of human papillomavirus proteins in yeast *Saccharomyces cerevisiae*. *Virology* **182**, 513–521 (1991).
55. Mach, H. et al. Disassembly and reassembly of yeast-derived recombinant human papillomavirus virus-like particles (HPV VLPs). *J. Pharm. Sci.* **95**, 2195–2206 (2006).
56. Meadows, A. L. et al. Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **537**, 694–697 (2016).
57. Vickers, C. E., Williams, T. C., Peng, B. & Cherry, J. Recent advances in synthetic biology for engineering isoprenoid production in yeast. *Curr. Opin. Chem. Biol.* **40**, 47–56 (2017).
58. Peng, B. Modulating gene expression in yeast to optimize metabolic networks for sesquiterpene and monoterpene production. (2018).
59. Zhao, J. et al. Dynamic control of *ERG20* expression combined with minimized endogenous downstream metabolism contributes to the improvement of geraniol production in *Saccharomyces cerevisiae*. *Microb. Cell Factories* **16**, 17 (2017).
60. Wei, M. X. et al. N-terminal truncations on L1 proteins of human papillomaviruses promote their soluble expression in *Escherichia coli* and self-assembly in vitro. *Emerg. Microbes Infect.* **7**, 160 (2018).
61. Chen, X. J. S., Casini, G., Harrison, S. C. & Garcea, R. L. Papillomavirus capsid protein expression in *Escherichia coli*: purification and assembly of HPV11 and HPV16 L1. *J. Mol. Biol.* **307**, 173–182 (2001).
62. Pan, D., Zha, X., Yu, X. H. & Wu, Y. Q. Enhanced expression of soluble human papillomavirus L1 through coexpression of molecular chaperonin in *Escherichia coli*. *Protein Expr. Purif.* **120**, 92–98 (2016).
63. Xia, P. F. et al. GroE chaperonins assisted functional expression of bacterial enzymes in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **113**, 2149–2155 (2016).
64. Saibil, H. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* **14**, 630–642 (2013).
65. Hefferin, M. L. & Tomkinson, A. E. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair* **4**, 639–648 (2005).
66. DiCarlo, J. E., Chavez, A., Dietz, S. L., Esvelt, K. M. & Church, G. M. Safeguarding CRISPR-Cas9 gene drives in yeast. *Nat. Biotechnol.* **33**, 1250–1255 (2015).
67. Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 38–41 (2007).
68. Peng, B. et al. A squalene synthase protein degradation method for improved sesquiterpene production in *Saccharomyces cerevisiae*. *Metab. Eng.* **39**, 209–219 (2017).
69. Takehara, M. et al. Characterization and Thermal Isomerization of (all-E)-Lycopene. *J. Agric. Food Chem.* **62**, 264–269 (2014).
70. Afgan, E. et al. Genomics virtual laboratory: a practical bioinformatics workbench for the cloud. *PLoS ONE* **10**, e0140829 (2015).
71. De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
72. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
73. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinforma.* **48**, 4.11.11–39 (2014).
74. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

75. Buels, R. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
76. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform* **14**, 178–192 (2013).

### Acknowledgements

B.P. and this research were supported by a CSIRO synthetic biology future science fellowship and the University of Queensland. B.P. and C.E.V. acknowledge current support from Australian Research Council Centre of Excellence in Synthetic Biology and Queensland University of Technology. Metabolite analysis was performed by Dr Manual Plan in Metabolomics Australia (Bioplatform Australia) Queensland Node. HPLC-MS/MS analysis was performed by Dr Lian Liu in Bioplatform Australia Queensland Node. Yeast strains in this study derives from CEN.PK background strains, which were provided by EUROSCARF (Scientific Research and Development GmbH, Germany) under a non-commercial licence. The authors also acknowledge the facilities, and the scientific and technical assistance, of the Australian Microscopy and Microanalysis Research Facility at the Centre for Microscopy and Microanalysis, The University of Queensland.

### Author contributions

B.P. and C.E.V. contributed to the conception of the project. B.P. designed and performed experiments. L.E. participated in protein purification and analysis. Z.L., Q.S., and L.C.C. participated in strain construction and characterisation. B.P. drafted manuscript. C.E.V. revised manuscripts. G.D., M.T., C.S., and C.B.H. provided advice opinions and participated in manuscript revision. C.E.V., G.D., M.T., and C.S. participated in the support and coordination of the project. All authors contributed to result analysis and discussion.

### Competing interests

The University of Queensland has filed two Australian provisional patents on the methods for gene amplification to claim the intellectual property (Inventors: B.P. and

C.E.V. Australian Patent Application numbers: 2022900699 and 2022901094). C.E.V. has a financial interest in Provectus Algae. Other authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-30529-8>.

**Correspondence** and requests for materials should be addressed to Bingyin Peng or Claudia E. Vickers.

**Peer review information** *Nature Communications* thanks Antonios Makris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022