Published online 10 December 2021

Nucleic Acids Research, 2022, Vol. 50, No. 1 ***561–578***
https://doi.org/10.1093/nar/gkab1214

Rational design and construction of multi-copy biomanufacturing islands in mammalian cells

Raffaele Altamura, Jiten Doshi and Yaakov Benenson*

Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, Basel, 4058, Switzerland

Received July 06, 2021; Revised November 21, 2021; Editorial Decision November 22, 2021; Accepted November 26, 2021

ABSTRACT

Cell line development is a critical step in the establishment of a biopharmaceutical manufacturing process. Current protocols rely on random transgene integration and amplification. Due to considerable variability in transgene integration profiles, this workflow results in laborious screening campaigns before stable producers can be identified. Alternative approaches for transgene dosage increase and integration are therefore highly desirable. In this study, we present a novel strategy for the rapid design, construction, and genomic integration of engineered multiple-copy gene constructs consisting of up to 10 gene expression cassettes. Key to this strategy is the diversification, at the sequence level, of the individual gene cassettes without altering their protein products. We show a computational workflow for coding and regulatory sequence diversification and optimization followed by experimental assembly of up to nine gene copies and a sentinel reporter on a contiguous scaffold. Transient transfections in CHO cells indicates that protein expression increases with the gene copy number on the scaffold. Further, we stably integrate these cassettes into a pre-validated genomic locus. Altogether, our findings point to the feasibility of engineering a fully mapped multi-copy recombinant protein 'production island' in a mammalian cell line with greatly reduced screening effort, improved stability, and predictable product titers.

INTRODUCTION

Cell line development is a critical step in the establishment of a biopharmaceutical manufacturing process (1). A robust manufacturing process requires a producer cell line that supports high and stable product expression, while ensuring consistency with respect to product quality. Mammalian expression systems are the platform of choice for the production of many high-value biologics, such as recombinant monoclonal antibodies and other recombinant therapeutic proteins, due to their inherent capacity for complex protein folding and post-translational modifications (2).

The traditional and still most widely adopted mammalian cell line development scheme hinges on the random integration of a gene of interest (GOI) and a selection marker, together with their regulatory elements, into a host cell line's genome, followed by selection for the populations of survivor cells that have stably integrated the selection marker and GOI (3). For the purpose of industrial cell line development, metabolic selection is preferred over antibiotic selection, with the glutamine synthetase (GS) and dihydrofolate reductase (DHFR) coding sequences serving as the most common selection markers, normally employed on *gs*- and *dhfr*-null cellular backgrounds (4,5). Frequently, to increase the stringency of metabolic selection, GS and DHFR inhibitors such as methionine sulfoximine (MSX) and methotrexate (MTX), respectively, are administered in one or multiple rounds, resulting in marker and GOI copy number amplification and, concomitantly, higher product titers due to GOI dosage increase (6). Following selection, single-cell-derived clonal lines are obtained and expanded, and their specific productivity and growth characteristics evaluated. Due to variations in the GOI integration sites and copy number profiles exhibited by different clones, the random integration and selection approach results in high clonal heterogeneity and unpredictability with respect to GOI expression, growth behaviour, genetic stability and adaptability to manufacturing processes, requiring long and laborious screening procedures before an at-scale bioprocess can be established (7–13).

In an effort to streamline mammalian cell line generation, recent research has focused on developing strategies for the targeted integration of one or more GOIs into transcriptionally active sites of the host genome. In the context of biomanufacturing, a number of studies have shown that both designer nucleases (e.g. Crispr/Cas9) (14–16) and site-specific recombinases (17,18) can successfully be employed to catalyze the integration of gene expression cassettes at defined loci in the genome. When compared with the traditional random integration approach, site-specific integration has been associated with reduced clonal heterogeneity and improved clonal stability over time with respect to product expression (16,19–20).

*To whom correspondence should be addressed. Tel: +41 61 387 33 38; Email: kobi.benenson@bsse.ethz.ch

While site-specific integration platforms have already found applications in industry in connection with the accelerated development of stable producer cell lines for the early molecular assessment of candidate protein therapeutics, product titers can seldom match those obtained with the best clones derived from a random integration and selection approach, where multiple GOI copies are normally integrated and active in the host genome (21). With the objective of narrowing the gap in product titers between the random and targeted integration platforms, there have been attempts at increasing gene dosage by simultaneously targeting more than one genomic site (22), by concatenating multiple GOI copies (23–25), or through a combination of both approaches, that is, targeting multiple GOI copies to two or more loci (26,27). Despite the promise held by these approaches, they are limited with respect to scalability, as evidenced by a recent study reporting GOI loss and, concomitantly, a lack of product increase when four or more GOI-bearing, identical gene expression cassettes were integrated at a single genomic site, potentially owing to recombination between repeated sequences (27).

In this study, we present a novel strategy for GOI dosage increase based on the rapid assembly of genetic constructs that carry multiple GOI copies and can be site-specifically integrated into a mammalian host cell line's genome, where they serve as 'biomanufacturing islands'. Importantly, all gene expression cassettes forming a production island consist of a GOI and regulatory elements that are unique in their genetic sequence. Recent work established a computational framework for efficiently designing thousands of nonrepetitive regulatory elements in bacteria and yeast and showed experimentally that the use of nonrepetitive promoter elements enhances the genetic stability of a two-protein system in *Escherichia coli* (28). Here, we introduce a new *in silico* method that enables the automated design of both degenerate GOI coding sequences and synthetic regulatory elements for expression in a mammalian host. We show that the use of genetically unique gene expression cassettes - all expressing the same user-defined protein but built from unique coding and regulatory elements – greatly expedites the assembly of multiple gene constructs carrying up to ten gene cassettes, while minimising the risk of gene recombination upon delivery of the multiple gene constructs into the mammalian host cell line of choice.

MATERIALS AND METHODS

Design of degenerate GOI coding sequences

A gene diversifier and optimizer software tool, comprising a series of scripts written in the Python programming language (Python version 2.7.10), was created in order to automate the design of a user-defined number of gene variants, starting from the amino acid sequence of a protein of interest. The software pursues three main sequence design objectives. Firstly, genetic variation within an output set of degenerate gene sequences must be maximized, as measured by the average pairwise dissimilarity between sequences. Secondly, genetic variation must be distributed homogeneously across coding sequences. This is in order to avoid the existence of long, uninterrupted stretches of sequence homology with high recombinogenic potential. Finally, a sequence optimization module tunes synonymous codon usage, Guanine-Cytosine (GC) content and, optionally, mRNA folding propensity around the start codon in an attempt to increase gene expression levels. Therefore, given the amino acid sequence of a protein of interest, our gene diversifier and optimizer tool automates the design of a set of degenerate coding sequences that diverge from one another in terms of genetic composition without affecting protein expression in a specified mammalian host cell line.

Briefly, gene sequence diversification proceeds as follows: (i) an amino acid sequence and the number of coding sequence variants (n) to be generated are specified by the user, along with the desired expression host selected among Chinese hamster ovary (CHO), mouse and human cells; (ii) codon usage and GC content are tuned by the user by setting two parameters, the minimum relative synonymous codon usage (RSCU) value for a codon to be deemed usable, $RSCU_{min}$, and the minimum RSCU for a codon ending in Adenine (A) or Thymine (T), $RSCU_{min\_AT}$ (codons characterized by scores below the set $RSCU_{min}$ and $RSCU_{min\_AT}$ are not used in any gene variant; for instance, setting $RSCU_{min} = 0.5$ results in excluding those codons whose observed usage frequency in the selected host is below half of the frequency expected if all synonymous variants were used equally; note that a codon with an RSCU >1.0 has a positive codon usage bias and RSCU can take on a real value between 0 and the number of synonymous codons for a given amino acid; see also further below); (iii) n sequences are initialized at the first ('*ATG*') and second codon positions. For the initialization of the second codon, synonymous codon variants, fulfilling the RSCU criteria outlined above, are randomly assigned to the n growing sequences, while being used in equal proportions; (iv) codon addition proceeds so as to break up uninterrupted stretches of homology between growing sequences, on the basis of the following evaluation: starting from just the final triplet in all sequences (i.e. the last added codon, 3 nucleotides), sequences are clustered into groups based on the nature of their last codon (that is, all sequences ending in the same codon are put into the same group/cluster); the procedure is then repeated considering the last two triplets in all sequences (that is, the last 6 bp of all sequences, including the final triplet already evaluated but stretching backwards by one additional triplet) and, if 6-bp homologies are found, the corresponding sequences are 'elevated' within the cluster they belong to; the procedure is then repeated for longer DNA stretches (9, 12 bp, etc.) until no uninterrupted homologies can be found anymore (i.e. all the sequences belonging to the same cluster eventually diverge). Having completed this evaluation, synonymous codons are added in such a way so as to break up the longer homologies in each cluster, to the extent that this is allowed by the number of available synonymous variants; (v) the procedure of 'sequence clustering' just described is repeated before every new codon addition until the end of the amino acid sequence is reached, at which point sequences are returned; (vi) optionally, attempts are made *a posteriori* at minimising mRNA secondary structure around the start codon of every sequence; this is done by integrating the command line tool mRNA Optimizer (29) into the sequence construction

pipeline. Only the exact number of user-defined sequences are generated, making sequence generation very rapid.

High-level pseudocode for the program is given below, outlining how input parameters are processed by the sequence design module, where sequence diversification and optimization takes place. The core computational procedure of the sequence-building algorithm is illustrated in Supplementary Figure S1 (see also Results).

- # Input parameters
- ENTER protein_of_interest (amino acid sequence)
- ENTER number_of_coding_sequences
- ENTER expression_host (CHO, mouse or human)
- ENTER RSCU_min and RSCU_min_AT
- # Design sequences
- ASSIGN codon_usage_table based on expression_host
- UPDATE codon_usage_table based on RSCU_min and RSCU_min_AT values
- FOR every amino acid in protein_of_interest:
- COMPUTE synonymous_codons_list for every amino acid
- INITIALIZE library_array (size: number_of_coding_sequences X length protein_of_interest)
- INITIALIZE all sequences in library_array at positions 1 and 2
- FOR amino acid in protein_of_interest (from third amino acid onwards):
- # Starting from the last codon column in library_array, check sequence homologies
- # moving backward
- INITIALIZE idents_dictionary to store sequence identity tree
- WHILE homologies are found:
- UPDATE idents_dictionary with sequence homology relations
- MOVE to previous codon column
- BUILD homology_clusters based on idents_dictionary
- ASSIGN codons from synonymous_codon_list to homology_clusters
- NEXT
- RETURN library_array
- IF 5′ region optimization = = TRUE:
- CALL mRNA_Optimizer
- OPTIMIZE codons in 5′ region (maximize MFE) of sequences in library_array
- UPDATE library_array
- # Sequence analysis
- COMPUTE hamming_distance_matrix for library_array
- COMPUTE longest_homology_stretch
- COMPUTE CAI_values
- COMPUTE 5′_folding_energies

CHO, mouse and human codon usage tables were obtained from the Kasuza database (http://www.kasuza.or.jp) and are provided as part of the github package accompanying this work. The relative synonymous codon usage (RSCU) for every codon is calculated as the ratio of the observed frequency of a given codon j encoding an amino acid i (observed frequency denoted $X_{i,j}$) to the expected frequency of the same codon under the assumption of equal usage of all the codons specifying i:

$$RSCU_{i,j} = \frac{X_{i,j}}{1/n_i},$$

where n_i is the number of synonymous codons encoding amino acid i.

The codon adaptation index (CAI) (30) for a gene sequence was used as a metric for coding sequence fitness in a given expression host. To calculate CAI scores, the relative adaptiveness of a codon (denoted $w_{i,j}$) is first calculated as the RSCU value for a given codon divided by the RSCU of the optimal codon encoding the same amino acid:

$w_{i,j} = \frac{RSCU_{i,j}}{RSCU_{i,max}}$.

Using the relative adaptiveness values, the codon adaptation index (CAI) for a gene is calculated as the geometric mean of the relative adaptiveness of each codon over the length of the gene sequence, L:

$$CAI = \left(\prod_{i=1}^{L} w_i\right)^{1/L}$$

For the minimum free energy (MFE) calculations of predicted mRNA secondary structures around the start codon, the ViennaRNA core package (31) was installed and configured for use within a Python environment through the freely available RNAlib library (version 2.4.14). mRNA folding near the translation start site has been implicated in gene expression in both eukaryotic (32,33) and prokaryotic (34,35) organisms. Coding sequence optimization in order to minimize mRNA folding propensity was performed with the freely available mRNA Optimizer (29) command-line tool (version 1.0), which was interfaced with the Python programming pipeline. Default parameters in the mRNA Optimizer tool were used, without constraints on maximum optimization time, number of iterations or GC content in the optimized sequence.

Using the diversifier and optimizer tool, ten degenerate mCitrine coding sequences (denoted c2.x) and twenty-one Interferon-gamma (IFNg) coding sequences (denoted IFNx) were designed. Tunable parameters were as follows: host = CHO, RSCU_min = 0.5, RSCU_min_AT = 0.9, 5′ coding region optimization (first 6 codons) = TRUE for mCitrine genes, and host = CHO, RSCU_min = 0.5, RSCU_min_AT = 0.8, 5′ coding region optimization (first six codons) = TRUE for IFNg genes. In order to assess the impact of the software's sequence optimization module, a library of 37 degenerate mCitrine coding sequences (c1.x) was designed by random codon assignment with no RSCU or GC (%) constraints, and no mRNA sequence optimization around the start codon, while still ensuring that sequences differed from each other, on average, by at least 20% of their nucleotides. Direct comparison between c1.x and c2.x helped to validate the use of the gene diversifier tool for coding sequence design. cOpt, also used for comparison, is optimized with the 'one amino acid-one codon' rule, that is, for a given amino acid, only the codon with the highest CAI is used.

The gene diversifier and optimizer software is freely available for download at https://github.com/altamurr/Gene-Diversifier.

Design of synthetic regulatory elements

The wild type human elongation factor 1-alpha (hEF1-a) promoter (human chr6: 73520048–73521229) was chosen as a scaffold to construct a small library of synthetic promoters. A number of hEF1a promoter deletions, denoted pΔSacII, pΔMluI-AgeI, pΔMluI-AgeI-SacII, pΔApaI-FseI, pΔbstAPI-SacI, pΔFseI-BstAPI, pΔSacI-KasI, pΔApaI-KasI, were generated by digestion of pRA52 (hEF1aP-mCitrine-RbG p(A)) with restriction enzymes specified in their mutants' denomination and religation. Four additional mutants, denoted cm1-cm4.hEF1a, were created by mutating 10-bp windows in the core promoter region. cm1-cm4.hEF1a were synthesized as synthetic gene fragments (gBlocks, IDT). The exact nature of the mutations is given in the Supplementary Information (Supplementary Table S1).

21 synthetic promoters, named P1-P21.hEF1a (Supplementary Table S2), were built as follows. For the construction of P7-P18.hEF1a, four wild type hEF1a promoter regions were preserved, denoted c1–c4 (Supplementary Table S3). c1 contains the EFP1 and EFP2 regulatory elements (36); c2 harbours the TATA box (5′-TATATAA-3′) and Initiator element (5′-TCTTTTT-3′); c3 and c4 include the 5′ donor splice site (5′-CAGGTAAGT-3′) and 3′ acceptor splice site (5′-CAGG-3′). A 6-bp and a 10-bp spacer between c1 and c2 and between c2 and c3, respectively, were fully randomized to introduce sequence variation between promoters. 11 and 12 transcription factor binding sites, randomly drawn from a pool containing Activator Protein-1 (AP-1), GC box, Enhancer Box (E-box) and nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) binding sites, were placed upstream of c1 and between c3 and c4, respectively, separated by random 6-bp spacers. Transcription factor binding site sequences were AP1: 5′ – AGTGACTCA – 3′, NF-kB: 5′ – GGGACTTTCC – 3′, e-Box: 5′ – CCACGTGATC – 3′ and gc-Box: 5′ – TGGGCGGGAT – 3′, found to be overrepresented in promoter sequences showing high activity in CHO cells, according to ref. (37). Promoters P1-P6 and P19-P21 were built as just described, with the sole difference that wild type hEF1a sequence was preserved between regions c1 and c2 - resulting in a conserved c1/2 core region - and 12, rather than 11, randomly drawn transcription factor binding sites were inserted upstream of c1/2.

A library of 67 synthetic 3′ untranslated region (UTR) and polyadenylation elements was built on the scaffold of the wild type rabbit beta-globin 3′ -UTR and polyadenylation sequence, denoted RbG polyA (rabbit Chr1, 146,236,634–146,237,211) (Supplementary Table S4). To build the library, the polyadenylation signal (PAS, 5′-AATAA-3′) and two GC-rich downstream elements (DSE1, 5′-GTGTGTTGG-3′; DSE2, 5′-TTTTTTGTGT-3′) were preserved, and randomized nucleotide sequences were added between them while preserving wild type spacing (23-bp random linkers were used between PAS and DSE1; random dinucleotides separate DSE1 and DSE1; 20- and 21-bp random DNA stretches were added upstream of PAS and downstream of DSE2, respectively). A minimal, 48-bp RbG polyA signal (the Levitt signal (38)) comprising only the wild type sequence stretching from PAS to DSE2, without dinucleotide spacer between DSE1 and DSE2, was used in preliminary analyses to assess the impact on reporter expression of 3′-UTR and polyadenylation compacting.

Cloning of coding and regulatory sequence libraries

c1.x, c2.x and IFNx DNA sequences were synthesized as gBlock gene fragments (Integrated DNA Technology) (Supplementary Tables S5-S7). The sequence 5′-AGTTCTGAATTCCGTTCGCTAGCGCCACC– 3′, containing an EcoRI cloning site as well as a consensus Kozak signal, was included directly upstream of the ATG start site of every coding sequence; the sequence 5′-TCTAGAGTTACA-3′, containing an XbaI cloning site, was annexed downstream of the stop codon. Gene fragments were digested with EcoRI and XbaI restriction enzymes (New England Biolabs, NEB) and cloned into the pRA52 expression vector, between the hEF1a promoter and a RbG 3′-UTR and polyadenylation sequence. Cloning procedures were partially automated on a EVO200 liquid handling robot.

Promoter library components, denoted pX.hEF1a, were ordered as gBlock fragments, containing MluI and EcoRI cloning sites. After digestion with MluI and EcoRI restriction enyzmes (New England Biolabs), the promoters were cloned into a pRA52 expression vector, upstream of the mCitrine fluorescent reporter and RbG 3′-UTR.

To produce a library of synthetic 3′-UTR sequences, denoted in the text as pAx, complementary forward and reverse oligo nucleotides were designed such that, after annealing, double stranded fragments were ready for cloning into a pRA52 expression vector digested with XbaI and XmaI (NEB).

Assembly of single-gene expression vectors and multiple gene constructs

Expression cassettes encoding either the mCitrine fluorescent protein or IFNg were constructed from arrays of insulator elements chosen from a previously published set (39) (Supplementary Table S8), synthetic promoters and 3′-UTR sequences, as well as mCitrine and IFN coding sequence variants designed with the diversifier software. Expression vectors (also denoted as first level expression vectors), each harbouring a genetically unique expression cassette, were created by isothermal (Gibson) assembly (40). Briefly, each expression cassette was subdivided into three component blocks: (i) insulator together with a promoter, (ii) a gene sequence, (iii) a 3′-UTR sequence followed by an adapter element homologous to the first 80 nucleotides of a different cassette's insulator sequence (required for multiple gene construct assembly, see further below). These three blocks, designed to have 24–30 bp overlaps for isothermal assembly, were ordered as gBlock fragments (Integrated DNA Technologies); 100 ng of each together with 100 ng of linear pRA52 backbone, cut at SalI and HindIII restric-

tion sites, were assembled using a Gibson Assembly Master Mix (NEB). Using this cloning strategy, nine first level expression vectors encoding either mCitrine or IFNg were constructed. Additionally, two 'sentinel' cassettes were created, containing either a promoter-less mCerulean reporter or the coding sequence for the mCherry fluorescent protein together with a promoter (wild type hEf1a promoter) and 3′-UTR (wild type RbG p(A)). Importantly, expression cassettes with first level vectors were designed to occupy a specific position within larger, multi-cassette assemblies, the position being specified by the identity of their adapter.

Using single-gene expression cassettes as building blocks, multiple gene constructs harbouring 1, 3, 6 or 9 gene expression cassettes (all encoding either mCitrine or IFNg), as well as one sentinel cassette, were constructed in yeast by homologous recombination. Terminal overlaps between recombining fragments (gene expression cassettes) were 80 bp in length, while fragment-to-backbone overlaps spanned 40 bp of sequence homology. The pYES1L shuttle vector (ThermoFisher, A13287), bearing a BAC origin for replication in bacteria and a low-copy CEN/ARS origin of replication for propagation in yeast, was used to produce all the multiple gene constructs subsequently delivered to CHO cells. For the assemblies, 500 ng of each PCR-amplified expression cassette (Phusion Polymerase, NEB) were used together with 100 ng of the pYES1L vector. Prior to transformation, purified DNA preparations were obtained by running PCR reactions on a 1% agarose gel, followed by DNA band excision and clean up with a gel extraction clean-up kit (Qiagen, 28704). Yeast transformations were then performed as follows. pYES1L vector backbone (100 ng) and DNA fragments with compatible terminal overlaps (500 ng) were mixed in a 1.5 mL microcentrifuge tube, ensuring the final volume did not exceed 10 μl. 100 μl of thawed MaV203 competent yeast cells (ThermoFisher, Cat# 11445012) were added to the DNA mix, followed by the addition of 600 ul PEG/Lithium Acetate solution (provided with GeneArt assembly kit, ThermoFisher, Cat# A13286). After mixing by inversion, tubes were incubated at 30^{o}C for 30 min. 35.5 μl DSMO (Sigma, D8418) were added to the DNA-cell mix prior to heat-shock transformation (incubation at 42^{o}C for 20 min). Cells were then pelleted by centrifugation at 300 × g for 5 min and resuspended in 1 ml 0.9% NaCl in water. 150 μl of the cell solution were plated on CSM minus tryptophan plates (ThermoFisher, A13292) and placed in an incubator at 30°C for 48 h. Transformant colonies were visible after ∼48 h. Candidate multi-copy gene constructs were transferred from yeast to *E. coli* as follows. Individual yeast colonies were picked from the plates using a toothpick and transferred to 0.2 ml PCR tubes (one colony per tube) containing 4–5 1 mm zirconium beads (Sigma, Z763780) and 15 μl lysis buffer (provided as part of ThermoFisher, A13286). (Note that a lysis buffer solution can also be prepared with Tris-HCl (pH - 8) to a final concentration of 20 mM, $CaCl_2$ to a final concentration of 1 mM, 0.1% SDS and 40 U/ml Proteinase K (NEB, P8107S)). The tubes were vortexed for 5 min at medium speed (PCR tube strips were taped to the vortex mixer). 1.5 μl lysate were used to electroporate OneShot Top10 Electrocompetent *E. coli* cells (Invitrogen, Cat# C404052). Note that for larger constructs a higher number of yeast colonies were processed and evaluated (after plasmid rescue from *E. coli*) than for smaller constructs (up to 15 yeast colonies for 10-cassette multiple gene constructs). In order to validate assembly products, minipreparations were obtained from *E. coli* using a BAC DNA miniprep kit (ZymoResearch, D4048). DNA preparations were screened by restriction digestion and further product validation was obtained via junction PCR (jPCR) analysis. For the jPCRs, primer pairs (annealing temperature = 60°C) were designed to amplify amplicons of sizes between 100 and 1000 bp. PCRs were set up with Taq Polymerase (NEB, M0273) in a total volume of 25 μl, using ∼0.5–1 ng plasmid DNA as substrate and 0.5 μl of 10 μM forward and reverse primers. Cycling parameters were 95°C for 30 s, then 30 cycles of 95°C for 20 s, 60°C for 30 s and 68°C for 60 s, followed by a 5-min incubation at 68°C. To multiplex jPCRs, up to 4 primer pairs with different amplicon lengths were routinely added to the same PCR reaction mixture, using the same parameters and reagent concentrations indicated above for single jPCRs.

Three plasmid series were obtained: pYES1L-mCherry-mCitrine.*x* and pYES1L-mCherry-IFN.*x*, for the assessment of gene dosage on protein expression in transient transfection assays; and pYES1L-attB/B'-mCitrine.*x*, for stable integration experiments (Supplementary Table S9). After assembly product confirmation, larger DNA preparations were obtained with a NucleoBond Xtra BAC kit (Macherey-Nagel, 740437) and purified from endotoxin with an endotoxin purification kit (Norgen, 22700).

Cell culture and transfections

Adherent CHO/dhfr- cells (ATCC-CRL-9096) were cultured in Iscove's modified Dulbecco's medium, containing 4 mM L-glutamine, 4500 mg/l glucose and 1500 mg/l sodium bicarbonate (IMDM, ATCC 30-2005), and supplemented with 10% Fetal Bovine Serum (Life Technologies, 10270106), 1% Penicillin-Streptomycin (Sigma, P4333), 0.1 mM hypoxanthine and 0.016 mM thymidine (Hypoxanthine-Thymidine, ATCC 71-X). Cultures were maintained at 37°C with 5% CO_2 in the air and passaged every 2–3 days using 0.25% trypsin–EDTA (ThermoFisher Scientific, 25200072) with a sub-cultivation ratio of 1:6–8.

Coding and regulatory components' performance was evaluated in transient transfection experiments. Transfections of the c1.x (pRAc1.x), c2.x (pRAc2.x), IFNx (pRA.IFNx) and pAx plasmid libraries were performed as follows: CHO cells were seeded ∼24 h prior to transfection on uncoated 96-well plates (ThermoScientific, 1161093) at a density of 10 000 cells/well in 100 μl culture medium. At the time of transfection, culture medium was replaced with fresh medium. 100 ng of pRAc1.x, pRAc2.x or pRA.IFNx plasmids and 30 ng of pEF1a-mCherry-RbGpA (pKH26) (for the mCitrine libraries only), used as a transfection control, were diluted in 20 μl Opti-Mem (Gibco, 31965–062), and 0.5 μl Lipofectamine 2000 (Thermo, 11668030) was added to each plasmid mixture. The pX.hEF1a promoter library was transfected as described above, but in 24-well plates, seeding 70000 cells 24h before transfection, using 450 ng of pX.hEF1a construct and 50 ng of pEF1a-mCherry-RbG reporter diluted in 100 μl Opti-MEM and with the addition of 2 μl of Lipofectamine 2000.

For transient transfections of the pYES1L-mCherry-mCitrine.*x* and pYES1L-mCherry-IFN.*x*, CHO/dhfr- cells were seeded at a density of 100 000/well 24 h prior to transfection. Transfections were performed in 24-well plates. 1μg BAC dna for the mCitrine.*x* series and 0.5 μg DNA for the IFN.*x* series (concentration estimated with a NanoDrop One spectrophotometer and ascertained on agarose gels) were diluted in 40 μl Opti-MEM and 3 μl (mCitrine.*x*) or 1.5 μl (IFN.*x*) of Fugene transfection reagent (Promega, E2311) were added to the DNA solution. Mixing was by gentle pipetting (15 times).

For stable integration experiments with the pYES1L-attB/B'-mCitrine.*x* series, transfection was as for transient transfections but with the addition of 200 ng Bxb1 Int (pEL215) (control transfections without Bxb1 Int were also routinely set up). Forty-eight hours post-transfection, three quarters of each transfected culture was transferred to a T25 flask and the remaining culture analyzed by flow cytometry with a BD Fortessa to estimate transfection efficiencies. Growing cultures were expanded up to T75 flasks and maintained in culture for 10 days before flow cytometry analysis.

Construction of landing pad CHO cell lines for RMCE

For the construction of a lentiviral vector, third-generation transfer plasmid pJD17 was obtained by modifying pFUGW (41) (Addgene # 14883) to carry an mCherry coding sequence flanked upstream and downstream by attP and attP' (the landing pad sites), respectively, and controlled by a hEF1a promoter, while pMDLg/RRE (Addgene # 12251), pMD2.G (Addgene # 12259), pRSV-Rev (Addgene # 12253) were used for packaging (42). For the production of lentiviral particles, 5×10^6 HEK293T cells (ATCC, Cat # CRL-11268) in 9 ml DMEM media (ThermoFisher, Cat# 41966029) were plated into 60 cm^2 dishes (Cat# 93100, TPP). After three hours, cells were transfected with 15 μg pJD17, 10 μg pMDLg/RE, 2 μg pMD2.G and 1 μg pRSV-Rev using the calcium phosphate transfection method as previously described (43). Forty-eight hours after transfection, medium containing the viral particles was harvested, filtered through a 0.45 um syringe, aliquoted and stored at –80°C. To create our landing pad cell lines, CHO/dhfr- cells were transduced with the lentiviral particle-containing solution at a multiplicity of infection (M.O.I.) of ∼ 0.05. Transduced cultures were monitored weekly by flow cytometry analysis. Three weeks after transduction, mCherry-positive cells were bulk-sorted and kept in culture for ∼6 months to ensure the stability of transgene expression. Finally, twenty-four single cells sorted from low, intermediate or high mCherry expression gates were expanded to generate landing pad (LP)-bearing mCherry lines (LP1–LP24), ready for RMCE.

Flow cytometry

For all transient transfection experiments, cell culture samples were analyzed 48 h post-transfection with a BD Fortessa Cell Analyzer. For mCherry, a 561 nm excitation laser, 600nm long-pass filter and 610/20 emission filter were used. For mCitrine, a 488 nm laser, 505-nm long pass filter and 610/20 emission filter were used. Flow cytometry data were analyzed with FlowJo (https://www.flowjo.com/). Live cells were gated based on forward and side scattering, negative controls (no transfected plasmids) were used for gating on mCherry and mCitrine axes, such that 99.9% of the negative control fell into the negative bin for each colour. mCitrine output was normalized to the mCherry control using the following formula:

$$mCitrine(n.u.) = \frac{mean\ (mCitrine(+)) \times frequency\ (mCitrine(+))}{mean\ (mCherry(+)) \times frequency(mCherry(+))}$$

where mCitrine(+) and mCherry(+) populations denote populations that express the respective fluorescent marker above control level

For integration (RMCE) experiments, mCherry and mCitrine recordings were performed as described above. For mCerulean, a 445-nm laser and a 473/10 emission filter were used. Live gates were based on forward and side scattering. Negative controls (no transfected plasmids) were used for gating on mCherry, mCitrine and mCerulean axes, such that 99.9% of the negative control fell into the negative bin for each colour. mCitrine fluorescence calculations were based on mCitrine expression of mCitrine(+), mCherry(–) and mCerulean(+) cell populations.

ELISA

IFNg measurements were performed as follows: culture medium was collected 48 hours post-transfection and diluted 1:1000 (with additional medium). A human IFNg ELISA kit (Invitrogen, EHIFNG) was used according to the manufacturer's specifications. Briefly, 50 μl of biotinylated antibody reagent was added to each well of a flat-bottom transparent 96-well plate (provided with the kit); 50 μl diluted samples or standards were added in triplicates to separate wells in the plate and incubated for 2 h at room temperature; after three plate washes with Wash Buffer, 100 μl of Streptavidin-HRP solution was added to each well, followed by incubation at room temperature for 30 min and three additional washes; 100 μl of Tetramethylbenzidine (TMB) substrate was added to each well and a color reaction was allowed to develop for 30 min at room temperature in the dark; the reaction was stopped with the addition of 100 μl Stop Reaction Solution to each well. Standards were resuspended in ultrapure water (Thermo, 10977015) and standard curves were obtained from serial dilutions of the standards. IFNg signal intensity was estimated from absorbance readings at 450 nm (on a Tecan M1000 Pro Reader) and final IFNg concentration estimates were obtained by extrapolation from the standard curve.

RESULTS

Coding sequence diversification: generation of multiple GOI sequences

The novel strategy for cell line development put forward in this study centres on the *in silico* design of multiple gene expression cassettes that are all unique in terms of their DNA sequences but express the same protein of interest. The unique genetic makeup of each gene expression cassette enables the use of a homologous recombination-based method for the rapid assembly of multiple-gene constructs

and, further, it serves to minimize the risk of recombination after construct delivery to a mammalian host cell line of choice (Figure 1).

Multiple GOI variants encoding the same protein can be generated by virtue of the degeneracy of the genetic code, whereby many of the twenty amino acids are specified by more than one codon. To automate the design of sets of degenerate coding sequence, a software tool was built that combines gene diversification and gene optimization to produce arrays of gene sequence variants encoding the same protein. The kernel of the diversification module is a novel algorithm that generates divergent sequences codon by codon, while dynamically monitoring sequence identity relations within the sequence set. Specifically, before a new synonymous codon is added to the sequences, information is gathered on the extent to which sequence homology in the set is accumulating, in terms of the distance between the 'last' codon mismatch and the current codon position between every pair of sequences. Synonymous codons are then allocated in such a way so as to prioritize the breaking of the longest stretches of growing homologies in the set (Supplementary Figure S1). This dynamic sequence clustering and homology-breaking procedure allows to maximize genetic variation within the resulting set of degenerate coding sequences, and concomitantly contributes to distributing such variation evenly among the sequences by reducing the lengths of worst-case contiguous homologous sequence between different sequences in the set by about 2-fold compared to random codon assignment (Figure 2A).

Together with the diversification module just described, the software includes a number of sequence optimization features. To increase sequence fitness, a relative synonymous codon usage (RSCU) threshold can be set below which codons are not utilized (the RSCU_min parameter), resulting in an increase of the median codon adaptation index (CAI) (30) in the resulting sets from 0.7 to 0.84, largely due to the elimination of rare codons (Figure 2B). Furthermore, as intragenic GC density has been found to positively correlate with protein expression in mammalian organisms through increased mRNA transcription rates (44,45), a second parameter, RSCU_min_AT, was included that allows the GC content of the sequence pool to be independently modulated. RSCU_min_AT sets the RSCU threshold for codons ending with an adenine or a thymidine base and, when set higher than RSCU_min, results in a further increase in GC content. In addition, further sequence optimization can be performed on the output set of diversified gene sequences through the redesign of the sequences' 5′-coding region aimed at minimising their mRNA folding propensity around the translation start site. This feature is included in our software in an attempt to leverage the observation that weak 5′-mRNA secondary structure may favor translation initiation and thereby protein expression in prokaryotes and eukaryotes (32,35,46), and is implemented through the use of the third party software tool 'mRNA optimizer' (29), resulting in the decrease of the mRNA folding energy of the gene set (Figure 2C). It must be noted that sequence pool diversity in the optimized region (five codons after the translation start site) is greatly reduced after optimization as all sequences converge on a limited number of variants. This is evidenced by the highly clustered appearance of the histogram for the optimized sequences in Figure 2C. The interplay of the sequence diversification and optimization modules is shown diagrammatically in Figure 3. Furthermore, it must be noted that, due to the flexible structure of the optimization module, additional optimization features and constraints on codon choice can be easily implemented, for instance the preferential usage of certain codon variants in order to limit the occurrence of amino acid misincorporation due to mistranslation (47).

To put the sequence diversifier and optimizer tool to the test and to appreciate the impact of sequence optimization on protein expression, two sets of gene variants encoding the mCitrine fluorescent marker protein were designed. The first set comprised 37 mCitrine variants (denoted c1.x) generated without the use of the diversifier by randomly selecting synonymous codons for each amino acid in the mCitrine coding sequence (synonymous codons were selected with equal probability without any constraints on codon usage, GC content or mRNA folding energy). Sequences were selected so as to satisfy an average Hamming distance criterion (>20%), therefore ensuring sufficient genetic variation among c1.x sequences. The second library, c2.x, comprising 10 mCitrine variants, was designed *in silico* with the gene diversifier tool. As a result of applying sequence constraints for both codon adaptation and GC content ($RSCU_{min}$ = 0.5 and $RSCU_{min\_AT}$ = 0.9) as well as redesigning the first six codons of the output sequences to lower their mRNA folding propensity around the start site, the average Hamming distance in the c2.x sequence set was slightly lower than for the c1.x library (∼20% versus ∼24%). Figure 4A shows that mCitrine expression levels in CHO cells of both c1.x and c2.x libraries, normalized to the expression level of cOpt, ('one amino acid-one codon' mCitrine-coding sequence, see Methods), positively correlate with sequence CAI (Pearson's $r = 0.84$, P-value = 8.6e–14), resulting in, approximately, a 3-fold increase in median expression between the two sets. However, since the more frequently used mammalian codons tend to end in G or C, contributions to gene expression from CAI and GC content are difficult to disentangle. Indeed, sequence composition (GC %) is also a good predictor of mCitrine fluorescence (Supplementary Figure S2A). Further, the observed correlations are lost if not enough variability in the predictor variable is allowed: the correlation between mCitrine expression from library c1.x only and CAI, for instance, is not significant ($r = 0.25$, P-value = 0.14), whereas exploring a broader CAI space results in a higher signal-to-noise ratio and the large correlation value reported above.

Finally, no significant correlation between minimum folding energy (MFE), used as a proxy for the strength of 5′-mRNA secondary structure, and mCitrine expression was observed (Supplementary Figure S2B). MFE values were calculated over the mRNA region spanning 23 nucleotides upstream and 18 nucleotides (six codons) downstream of the translation initiation site, that is, the region whose folding propensity was minimized by codon optimization with the mRNA Optimizer tool. A sliding window analysis covering a broader sequence region around the translation initiation site also does not suggest a correlation between relaxed mRNA secondary structure and increased translation efficiency (Supplementary Figure S2C). Since sequence

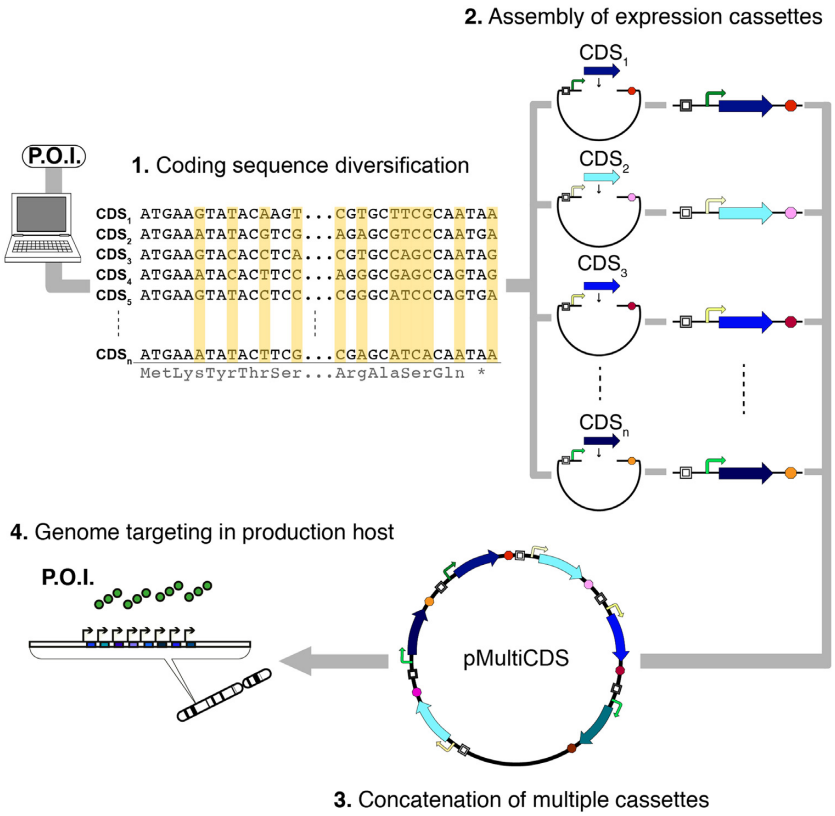

Figure 1. A novel strategy for cell line development. Using custom-built software, multiple coding sequence (CDS) variants encoding a protein of interest (P.O.I.) are obtained. These sequences are optimized for protein expression in a mammalian production host of choice, while their DNA sequences are designed to be highly divergent from each other. Together with libraries of regulatory components (insulators, promoters, 3′-UTRs), coding sequence variants are used to construct a number of expression cassettes, all expressing the P.O.I. but each unique in its genetic makeup. Multiple expression cassettes are subsequently assembled in a single step via homologous recombination in yeast. A string of concatenated expression cassettes is integrated, without any trailing vector sequences, into the host chromosome at a pre-validated site via recombinase-mediated cassette exchange, giving rise to a rationally engineered bioproduction island.

variation was confined to the coding region, it is possible that not enough dispersion in the predictor variable (MFE) was generated for a potential effect on expression to be detected (low signal-to-noise ratio) or, alternatively, that the thermal stability of the mRNA sequences in our dataset may not be sufficiently large to significantly inhibit translation. (MFE values in the sequence window spanning positions −23 to + 18 relative to translation start ranged between −17 and −2 kcal/mol, while RNA hairpins with a thermal stability of at least -30 Kcal/mol were previously reported to be required to severely inhibit translation (48).) Finally, as it has been previously suggested, higher eukaryotes may be more sensitive to mRNA structures in the proximity of the 5′ capping site rather than the translation start site (48).

Next, the diversifier software was utilized for the design of a small library of IFNg genes, optimized for expression in CHO cells. Twenty-one IFNg coding sequence variants were designed *in silico* with the diversifier software. The resulting library has an average CAI of 0.84 and an average GC content of 46% (for comparison, wild type human IFNg has a CAI score of 0.71 in the hamster host and a GC content of 38%). Upon transient transfection in CHO cells, most gene variants produce levels of secreted IFNg that are comparable to those attained from wild type IFNg, and a few even outperform wild type levels (Figure 4B), with sequences differing on average by 20% (Figure 4C). Taken together, these results show that it is possible to design gene sequences that substantially diverge in their sequence composition without having a large negative impact on protein expression. In connection with the use of multiple degenerate coding sequences for a protein production campaign, the individual assessment of computationally-generated coding sequences can help identify high-performers for inclusion in large multi-copy constructs, thereby positively affecting final protein titer.

Regulatory sequences diversification: generation of libraries of promoters and 3′-UTRs

The barebones of a mammalian gene expression cassette are a coding sequence and gene regulatory sequences, namely a promoter, a 5′-UTR and a 3′-UTR with a polyadenylation signal. For the purpose of designing multiple gene expression cassettes all encoding the same protein but also unique in their non-coding genetic makeup, small collections of synthetic promoters and 3′-UTRs with embedded polyadenylation signals were created. Unlike coding sequences, these *cis*-regulatory elements can be re-used for the expression of different recombinant proteins, thus becom-

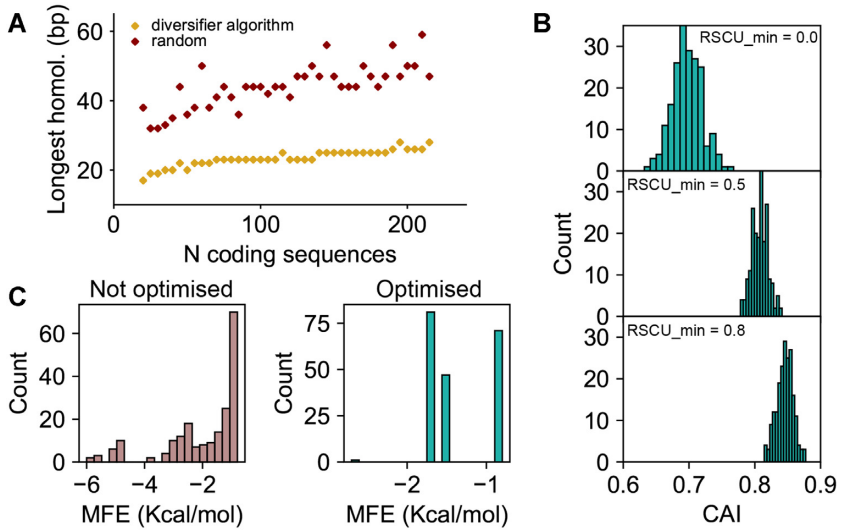

Figure 2. Coding sequence diversification and optimization. (**A**) Harmonization of genetic divergence. Sets of 20, 25, 30, .., 220 degenerate IFNg sequences were generated with the diversifier algorithm (orange diamonds) or by assigning random synonymous codons (red diamonds, at every codon position, synonymous codons are chosen with equal probability and therefore tend to be used with similar frequencies). For every set of sequences, the longest continuous stretch of homology between any two sequences in the set is plotted against the size of the set. (**B**) The effect of tuning $RSCU_{min}$ on CAI. Three sets of 200 IFNg sequences were generated *in silico* with the diversifier algorithm, with the $RSCU_{min}$ paramter set to 0.0, 0.5 or 0.8, thus restricting the available codon space to those codons with an RSCU higher than the threshold value (a value of 0.0 means that all codons are accepted) The codon adaptation index (CAI) for every sequence was computed and CAI values in each of the three IFNg sets were binned to construct the histograms shown in the figure. **(C)** 5′-sequence re-design with mRNA Optimizer. 200 IFNg degenerate sequences were constructed with the diversification algorithm. Sequence regions from -23 to + 18 around the ATG start site were fed to the optimizer and the codons following the start site were optimized so as to increase mRNA folding energies. Note that sequences located 5′ of the start site (containing the Kozak region and restriction sites for cloning) are not modified by the optimizer. The distributions of Minimum Folding Energies (MFE) before (not optimized) and after optimization (optimized) are shown as histograms.

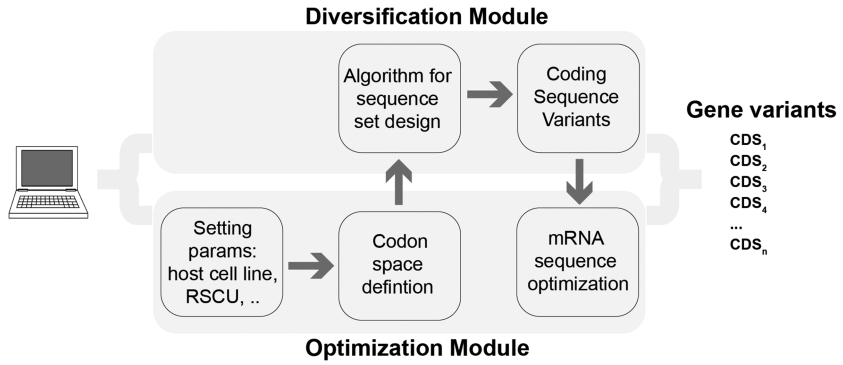

Figure 3. Structure of the purposely-built software used to obtain multiple gene coding sequence (CDS) variants from the amino acid sequence of a protein-of-interest. The software consists of two interconnected modules for sequence optimization and diversification. Through the sequence optimization module, the user is able to select a mammalian expression host (CHO, human or mouse) and set codon usage thresholds (RSCU_min, RSCU_min_AT), thus boosting the output sequences' codon adaptation index (CAI). Through the diversification routines, sequence variation between CDSs is maximized and homogenized so as to avoid long stretches of homology between sequences. Optionally, DNA sequence composition around the start site can be optimized using third-party software (mRNA Optimizer) so as to minimize the mRNA folding propensity in this region. Output CDSs thus designed are ready for synthesis and downstream applications.

ing part of a toolkit for the rapid assembly of various multi-copy gene cassettes. The synthetic promoter design is based on the hEF1a promoter architecture, a strong constitutive promoter that drives high gene expression across many cell types (49). Synthetic 3′-UTRs, harboring a polyadenylation signal sequence (PAS), are built on the Rabbit beta-globin (RbG) 3′-UTR scaffold, which is well-characterized and supports strong gene expression (38).

In an attempt to identify regions within the hEF1a promoter that might be amenable to mutation without major loss of gene expression, a series of deletion mutants was created, spanning mainly intronic sequences (Supplementary Figure S3A). These mutant promoter sequences were placed upstream of a fluorescent reporter in a plasmid which was transiently transfected in CHO cells to assess the impact on protein expression. The majority of the deletions had a relatively small negative effect on the expression of the fluorescent reporter. Additionally, a large 350 bp deletion mutation located between two SacII restriction sites had almost no effect on expression, in agreement with previously published data (36) (Supplementary Figure S3A). To evaluate whether deletions in the core regions could be tolerated,

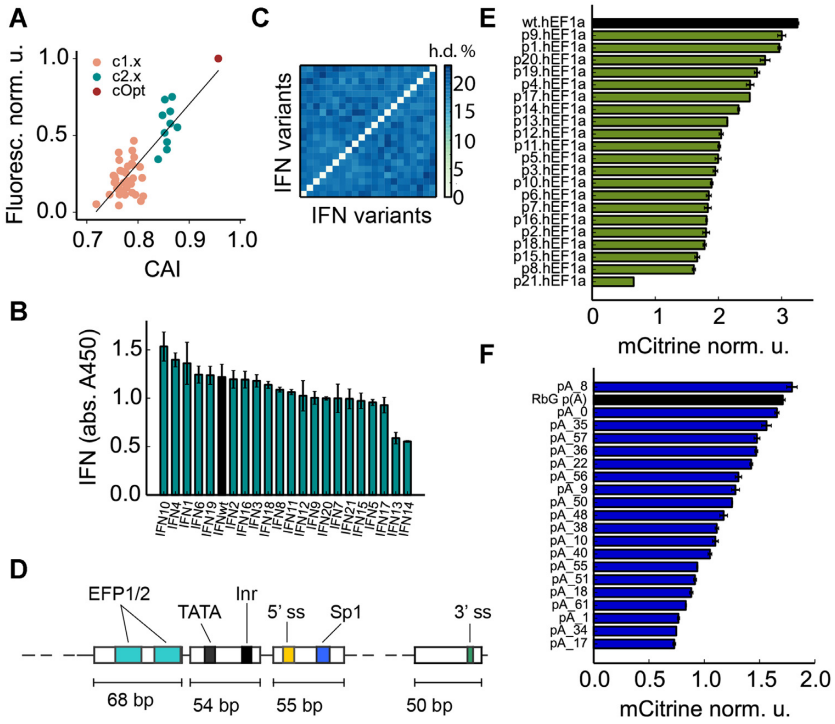

Figure 4. Experimental testing of coding and regulatory elements. (**A**) The relationship between codon adaptation index (CAI) and mCitrine expression. Fluorescence values for libraries c1.x (n = 37) and c2.x (*n* = 10) are normalized to cOpt (obtained with the 'one amino acid-one codon' optimization rule, i.e. only the most frequent codon for every amino acid is used) and are plotted against sequence CAI. The best-fit line is shown. (**B**) IFNg expression from 21 synthetic coding sequences designed with the gene diversifier software tool. The black bar shows expression from the wild type human IFNg coding sequence. (**C**) Heat map showing the hamming distance (h.d.) matrix for the IFNg library, where the color of each square represents the extent of sequence divergence between two sequences in the library. (**D**) Map showing the wild type human EF1a promoter sequence regions that were preserved across our synthetic promoter library. The length of such regions is indicated at the bottom. TATA = TATA box; Inr = Initiator element; 5′ ss = 5′ splice site; 3′ ss = 3′ splice site. (**E**) Reporter (mCitrine) expression driven by synthetic promoters p1.hEF1a-p21.hEF1a as well as the wild type hEF1a promoter sequence (black bar). mCitrine expression, evaluated after transient expression in CHO cells, was normalized to a transfection control (mCherry). (**F**) A library of 3′ UTRs (91 bp in length) was built by sequence randomization around the RbG poly(A) functional elements. The 3′ UTR library was cloned downstream of an mCitrine fluorescent reporter and transfected in CHO cells together with an mCherry reporter plasmid (used as a transfection control). The twenty strongest library members are shown in the bar chart. (The black bar indicates wild type RbG 3′-UTR performance.)

four 10-bp windows located immediately upstream of the EFP1 element, between EFP2 and the TATA box, immediately downstream of the initiator element or the 5′ splice site (5′ ss), were randomized (Supplementary Figure S3B). Three out of four core mutants lost approximately 30% of native expression while one showed no loss in reporter activity (Supplementary Figure S3C), pointing to the possibility of randomising narrow sequence windows within the core promoter region to further increase genetic variability in a promoter library.

Combining the data on promoter activity from these deletion and core mutants (see above) with the previously published findings that identified NF-kB, AP1, E-box and CG-box binding proteins as active transcription factors in CHO cells (37), a library of synthetic promoters was created by randomly inserting binding sites for these regulators upstream the hEF1a core promoter (i.e. upstream of EFP1/2) and within the intronic region, as well as by randomising small sequence windows in the core region (Figure 4D). The key regulatory sites that were preserved were the EFP1 and EFP2 *cis*-elements, TATA box and Initiator (Inr) elements, 5′ and 3′ splice sites (denoted 5′ ss and 3′ ss) as well as a putative binding site for the Sp1 transcription factor located directly downstream of the 5′ splice site. 21 synthetic promoters were designed in total, with a footprint of ∼690 bp per promoter (compared to 1180 of the typically-used wild type hEF1a promoter fragment). The library was found to support reporter gene (mCitrine) expression from nearly wild-type level to below 20% of wild type signal. Ten promoters that retain at least 60% of wild-type activity could be recovered from this small library and subsequently used for the assembly of multi-gene production constructs (Figure 4E).

The RbG p(A) key regulatory elements were previously identified in the 5′- AATAAA – 3′ polyadenylation signal (PAS) and two GC-rich downstream elements (DSE1 and DSE2) (38) (Supplementary Figure S4A). The sequence comprising just these three elements and the intervening region (adding up to a total of 48 bp) generates 75% of wild type expression in a fluorescent reporter assay (Supplementary Figure S4B). This short 48-bp scaffold was used to construct a library of 67 synthetic 3′-UTRs. The region between the functional elements was fully randomized and two additional stretches of 20 and 23 bp of fully randomized sequences were added upstream of the PAS and downstream of DSE2, respectively. The effect of these 91-bp synthetic 3′ UTRs on reporter expression is shown

in Figure 4F and Supplementary Figure S4C: upon transient transfection in CHO cells, reporter expression varied from above wild type levels to almost null. The observation of such large variation in expression, resulting from sequence randomization around the functional elements, suggests that intervening sequences between PAS and DSE1/2 are not neutral with respect to mRNA processing and, ultimately, protein expression. To evaluate whether mRNA secondary structure in the 3′-UTR region may directly impact protein expression, for instance by occluding the target sites required for binding by the cleavage and polyadenylation protein complex, minimum folding energies for the synthetic 3′-UTR library components were calculated using the RNAfold software (31). Only a weak-to-moderate positive correlation between folding propensity and reporter expression was found (Pearson's $r = 0.34$, P-value < 0.005) (Supplementary Figure S5), indicating that while weak RNA structure in the 3′-UTR and polyadenylation sequence may to a certain extent favor expression (50), other factors are likely to be implicated in determining final expression levels, including, for instance, the impact of the 3′-UTR sequences on overall mRNA stability and half-life (51).

Finally, the collection of regulatory sequences was expanded with the addition of a list of published mammalian insulators: these are compact, 200–300 bp genetic elements that are bound by the CTCF transcription factor and have been shown to act as enhancer blockers (52). Together with promoters, coding sequences and 3′-UTRs, these insulators contribute to the basic architecture of the transcriptional units to be used in higher-order assemblies.

Assembly of multiple gene constructs

Single-gene expression cassettes were constructed from arrays of regulatory elements (insulators, promoters and 3′-UTRs) and coding sequence variants (encoding either mCitrine or IFNg). Promoter and 3′-UTR combinations were selected in an attempt to generate cassettes characterized by uniform levels of gene expression, assuming their additive effect on gene expression. Additionally, unique restriction sites were placed between regulatory and coding units in order to allow for their rapid exchange by restriction digestion and ligation cloning. Coding sequences, in particular, can easily be replaced to start a new recombinant protein project, while the infrastructure for multi-gene assembly remains in place (Figure 5A). Each cassette, specified by an insulator, a promoter (including the 5′-UTR), a coding sequence and a 3′-UTR, also bears a position-specific, unique adapter sequence downstream of its 3′-UTR, required for the seamless fusion and higher-order assembly of multiple cassettes via homologous recombination in yeast (53) (Figure 5B). To set up multi-copy cassette assemblies, individual gene expression cassettes are PCR-amplified from first-level vectors to generate linear, overlapping fragments (see Materials and Methods and Figure 5A). Importantly, in a control experiment, the assembly of identical gene expression cassettes bearing unique, compatible terminal overlaps was attempted, consistently giving rise to circular plasmids that were shorter than the desired product and contained only a single gene expression unit, due to recombination between the expression cassettes (Figure 5C). This result reinforced the need for gene cassette diversification within the framework of an *in vivo* homologous recombination-based multiple gene construct assembly: the use of unique regulatory and coding units allows the seamless assembly of multiple expression cassettes (Figure 5D).

Alongside gene cassettes for protein production, two 'sentinel' cassettes were produced, denoted 'transfection sentinel' (hEf1a-mCherry-RbG pA) and 'integration sentinel' (mCerulean-RbG pA), to be used in transient transfection experiments aimed at extracting gene dosage-protein expression relationships, and for targeted integration and stable protein expression, respectively. Single-gene cassettes were first assembled with degenerate mCitrine coding sequences as their gene units. These were subsequently replaced with degenerate IFNg genes via restriction digestion and ligation. Together with the sentinel cassettes, mCitrine and IFNg expression cassettes were used to generate multi-copy constructs in yeast.

Different shuttle vectors—differing in their yeast and bacterial origins of replication as well as in the selection markers they carry—were tested for their capacity to support large plasmid assemblies in the yeast *S. cerevisiae* (up to ten cassettes plus backbone) and subsequent construct propagation in *E. coli*. The choice of an appropriate vector backbone turned out to be of the utmost importance. Shuttles pRG226 and pRG216 (54) (kindly provided by R. Gnügge and F. Rudolf) containing a high-copy 2u-ori and a low-copy CEN/ARS yeast origin of replication, respectively, as well as a high-copy-number bacterial ori (derived from pMB1), only yielded correctly assembled products when 2- and 4-cassette assemblies were performed, but did not support the assembly of 7 and 10 cassettes. Successful assembly of the latter required another shuttle vector, pYES1L (Thermo Fisher), a bacterial artificial chromosome (BAC)—kept at a single copy level in *E. coli*—with a low-copy CEN/ARS origin of replication for propagation in yeast. The list of multi-copy constructs assembled with pYES1L and their destination (gene copy number-expression analysis in transient transfection or genomic integration) is presented in Table 1, while a list of all the elements contained in each plasmid is given in the supplementary material (Supplementary Table S9).

In the best case, a 50% efficiency was attained for a 10-cassette assembly based on restriction digest analysis (Figure 5F), while smaller assembly projects had routinely higher efficiencies (Figure 5E). Screening of correctly assembled constructs was performed by restriction pattern analysis and further product validation was obtained by amplifying fragment junctions (Figure 5G). Since recombination after plasmid transfer into *E. coli* can lead to the recovery of altered product due to instability in the bacterial host, screening was performed directly on bacterial minipreparations rather than on yeast and bacterial colonies. It was found that, by limiting the screening procedure to the final minipreparations obtained from bacterial cultures and by avoiding the preparation of yeast DNA by directly using yeast lysates for *E. coli* transformations, the time required to complete an assembly project could be shortened to 4–5 days (Supplementary Figure S6).

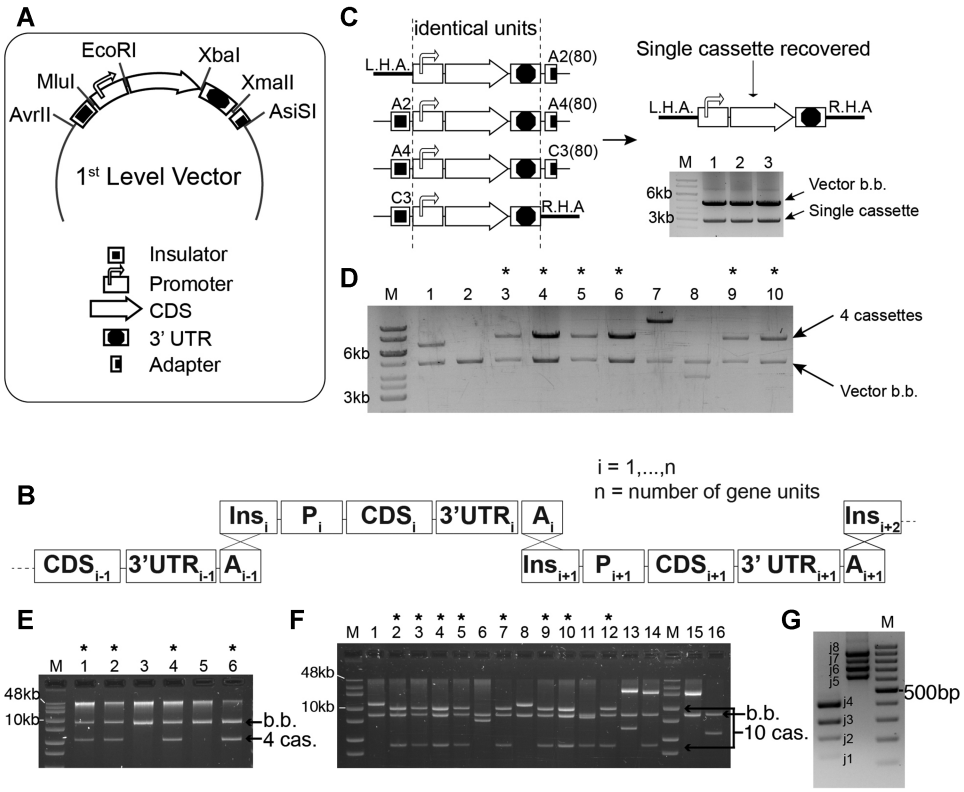

Figure 5. Assembly of multiple gene constructs from individual coding and regulatory elements. (**A**) First Level vector design scheme. Each expression cassette bears an insulator, a promoter, a 3′ UTR and poly(A) signal, and an adapter required for the assembly of multiple cassettes. Cassette components are punctuated by unique restriction enzyme sites, thus allowing for facile component replacement. (**B**) The homology between insulators (Ins) and Adapters (A) guides the assembly of individual gene cassettes into larger multi-gene constructs via homologous recombination in yeast. (**C**) Four cassettes, identical in their hEF1a-mCitrine-RbGpA sequences but with unique, compatible overlaps (see cartoons on the left) were transformed with linearized pRG216 shuttle vector into *S. cerevisiae*. Twelve transformant colonies (3 shown) were analyzed and all were found to contain a single cassette insertion (∼3 kb) between left and right vector homology arms (L.H.A, R.H.A) (see cartoon on the right and agarose gel image). M: 1kb molecular marker (ThermoFisher), Lanes 1–3: plasmid prepared from *E. coli* tranformants and digested with the double-cutter SbfI. Vector b.b. = vector backbone. (**D**) 4-Cassette assembly in vector pRG216 using unique coding and regulatory elements. Agarose gel showing a restriction digestion pattern analysis for 10 candidate assemblies. Lanes with asterisk indicate the correct digestion pattern. M: 1kb molecular marker (ThermoFisher), (**E, F**) agarose gels showing restriction digestion screening results allowing for the identification of plasmids that have correctly assembled 4 (panel E) and 10 (panel F) IFNg gene cassettes together with linearized BAC shuttle vector backbone (b.b.) pYES1L. Lanes are numbered and asterisks indicate lanes where the correct restriction patterns are observed. M = molecular marker, 1 kb extend (NEB). (**G**) Junction PCR analysis indicating the presence of all eight fragment junction amplicons (j1−j8) in a 7-cassette plus backbone assembly. PCR reactions are multiplexed (2 PCR reaction mixes with four PCR reactions per mix). M: 100bp GeneRuler (Thermo Scientific).

Delivery of multi-copy constructs to CHO cells for transient or stable expression

Prior to editing the CHO genome with our multi-copy gene constructs, the relationship between gene dosage and protein expression was assessed in a number of transient transfection experiments. The use of an mCherry sentinel cassette in constructs pYES1L-mCherry-mCitrine.x and pYES1L-mCherry-IFNg.x provides a means for normalising protein expression by allowing for the correction of variations in expression due to differences in transfection efficiency among the multi-copy construct series (as plasmid size increased from 12 to 27 kb, a nearly 3-fold drop in transfection efficiency was observed) (Figure 6A, B). This experimental strategy assumes that reporter expression remains relatively homogenous across the different plasmid vectors. As mCherry was placed between regulatory elements that support robust gene expression (wild type human EF1a promoter and RbG-pA), this assumption should hold within the timeframe of plasmid transfection and protein expression measurement.

Using the internal control strategy just described, the relationship between copy number and protein expression was explored by assessing mCitrine expression from the pYES1L-mCherry-mCitrine.x series, containing 1, 3, 6 or 9 unique mCitrine expression cassettes, as well as the mCherry sentinel expression cassette. mCitrine fluorescence (per unit of mCherry fluorescence) was found to increase with gene copy number, without any observable saturation effect (Figure 6C, D). In order to assess whether individual mCitrine cassettes have an additive effect on final mCitrine fluorescence, mCitrine expression from all nine individual units was measured by flow cytometry and the observed expression trajectory over gene copy number was compared with that expected from an additive model (Figure 6E): joining multiple gene units on the same expression vector produces an increase in final protein expression

Table 1. Multi-copy constructs and their destination. 12 assemblies were performed in total. Multiple gene constructs were designed to express either mCitrine or IFNg, as well as a 'sentinel' fluorophore (mCherry in transient transfection constructs and mCerulean in in the integration plasmids)

Plasmid ID	Number of gene cassettes assembled	Destination
pYES1L-mCherry-mCitrine.1x	2	Transient transfection assay
pYES1L-mCherry-mCitrine.3x	4	Transient transfection assay
pYES1L-mCherry-mCitrine.6x	7	Transient transfection assay
pYES1L-mCherry-mCitrine.9x	10	Transient transfection assay
pYES1L-attB/B'-mCerulean-mCitrine.1x	2	Genome editing (RMCE)
pYES1L-attB/B'-mCerulean-mCitrine.3x	4	Genome editing (RMCE)
pYES1L-attB/B'-mCerulean-mCitrine.6x	7	Genome editing (RMCE)
pYES1L-attB/B'-mCerulean-mCitrine.9x	10	Genome editing (RMCE)
pYES1L-mCherry-IFN.1x	2	Transient transfection assay
pYES1L-mCherry-IFN.3x	4	Transient transfection assay
pYES1L-mCherry-IFN.6x	7	Transient transfection assay
pYES1L-mCherry-IFN.9x	10	Transient transfection assay

that broadly agrees with the expected trend, although the expression is somewhat lower than expected from the additive model, possibly suggesting a modest degree of transcriptional interference. A monotonously increasing relation between gene copy number and protein expression was also found to hold for the pYES1L-mCherry-IFNg.x series, but IFNg expression appears to plateau as the highest gene copy number is approached (Figure 6F). On account of the observation that plateauing in protein expression was not observed with the pYES1L-mCherry-mCitrine.x plasmid series, where mCitrine is expressed intracellularly, saturation of the secretory pathway rather than interference between regulatory components at the transcriptional level (transcriptional interference) seems the more likely cause of the sublinear behaviour in IFNg expression. In this connection, it has been previously shown that components of the secretory pathway can at times become limiting, and their overexpression can improve the secretion efficiency of both 'easy-to-express' and 'difficult-to-express' therapeutic proteins (55,56).

As a strategy to stably integrate our multiple gene constructs into the CHO genome, targeted exchange was preferred over targeted integration, reasoning that the insertion of trailing vector sequences associated with full-vector integration might be detrimental to locus stability. A system with a pair of inversely oriented attachment sites for the Bxb1 recombinase was chosen to construct a 'landing pad' for cassette delivery (57). We leveraged the integrative properties of lentiviral vectors to deliver our landing pad that subsequently served as a platform for recombinase mediated cassette exchange (RMCE). Specifically, two attP attachment sites facing in opposite directions, denoted attP and attP', as well as an mCherry fluorescent reporter enclosed between them, were incorporated into a lentiviral vector, which was then used to infect CHO cells at a low multiplicity of infection (M.O.I. = 0.05). Importantly, the strong hEF1a promoter driving mCherry expression was placed outside of the landing pad region to be targeted for cassette exchange, creating a promoter trap and thus allowing for the easy visualization of recombinant populations: upon successful recombination, a promoter-less 'sentinel' marker (in our setup, the fluorescent reporter mCerulean) replaces the landing pad marker (mCherry), thus allowing the recovery of recombinant subpopulations via single cell sorting (Figure 7A). Production cassettes placed downstream of the sentinel marker on donor DNA are integrated together with the sentinel upon RMCE. It must be noted that, with this set-up, RMCE can lead to the donor material integrating in either a forward or reverse orientation relative to the genomic target. Only recombination in the forward direction activates the promoter trap and leads to mCherry-to-mCerulean marker conversion (refer to Figure 7A).

In order to ensure locus stability at the landing pad's insertion site, an mCherry (+) population obtained after lentiviral transduction and bulk-sorting for mCherry (+) cells was cultured and monitored for continued transgene expression for a period of over 6 months. Phenotypic stability with respect to cellular fluorescence over time was taken as an indicator of the underlying genetic and epigenetic stability of the transgene cassette. A number of single-cell derived populations, denoted LP1-LP24 and characterized by varying degrees of mCherry fluorescence, were obtained from this original polyclonal population (Supplementary Figure S7). A subset of these LP cell lines were tested for their ability to support RMCE with donor plasmid pYES1L-mCerulean-mCitrine.1x: only in the presence of the Bxb1 recombinase was mCherry-to-mCerulean conversion observed (typical results are shown in Figure 7B). Additionally, the recombination profile exhibited by different LP lines upon reacting with pYES1L-mCerulean-mCitrine.1x was informative of whether they harbored one or multiple functional landing pads: cell lines containing a single landing pad site produced only one recombinant population, whereas the presence of two landing pads on the same genome also gave rise to a subpopulation with an intermediate phenotype, in which the coexistence of both landing pad (mCherry) and recombination (mCerulean) markers is observed, and is due to only one out of the two landing pad sites being successfully targeted (Supplementary Figure S8).

As a proof of concept, the multi-copy constructs from the pYES1L-attB/B'-mCerulean-mCitrine.x series, encoding a number of mCitrine-coding cassettes between one and nine, were integrated in LP11. Following transfection of each donor plasmid together with the Bxb1 recombinase-bearing plasmid, cultures were grown and expanded for 10 days, at which point a subpopulation showing mCherry-to-mCerulean conversion, as well as mCtirine expression, could be identified. We observed that as the size of the integrated payload increased, expression from the multi-copy bioproduction island seemed to be affected in a fraction of the population, in which mCitrine expression seemed to be partially silenced. When comparing mCitrine expression levels from the high-expressing subpopulations (in all cases $> 50\%$ of total population) across all four cell lines, a linear gene dosage-protein expression behavior is observed (Figure 7C, D).

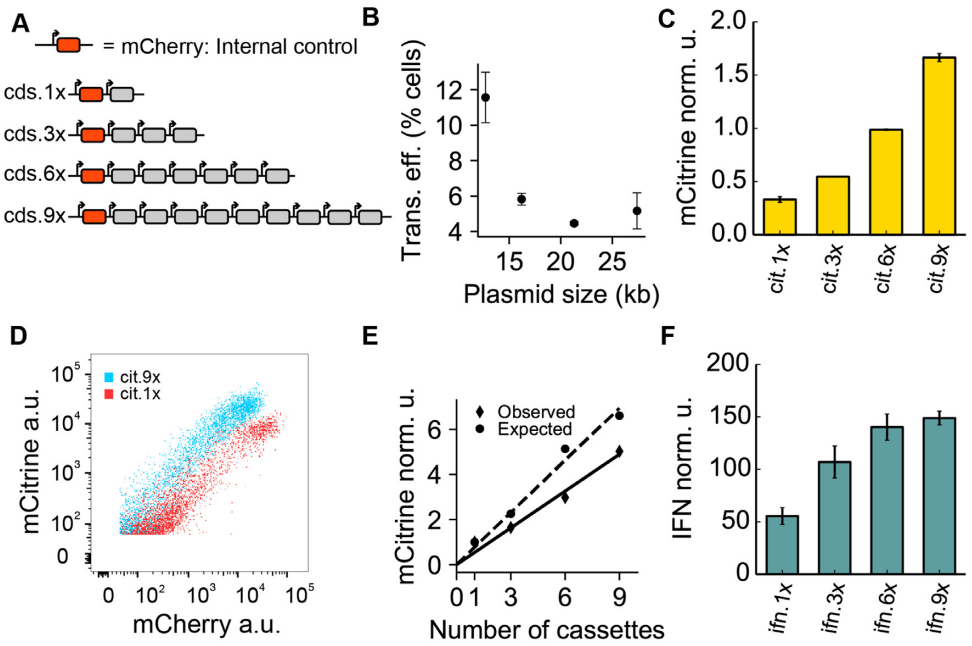

Figure 6. Impact of gene copy number on protein expression in transient transfections. (**A**) Cartoon showing single- and multi-copy constructs designed for assessing the gene dosage-protein expression relationship. Constructs harbor a variable number of gene coding sequences (gray boxes), while all carry the same mCherry expression cassette (the mCherry coding sequence is shown in red), required to normalize protein expression levels. (**B**) The decrease in transfection efficiency as plasmid size increases over the range 12–27 kb is shown. Equimolar amounts of pYES1L-mCherry-mCitrine.1x, 3x, 6x, 9x were transfected into CHO cells and percentages of transfected cells were computed as the number of mCherry positive cells upon flow cytometry analysis. (**C**) Bar chart showing the impact of increasing mCitrine copy number on mCitrine expression. Multi-copy constructs containing a variable number of genetically unique mCitrine expression cassettes (1, 3, 6 or 9) and an mCherry expression cassette were transfected into CHO cells. mCitrine/mCherry ratios are reported in the bar chart. (**D**) Flow cytometry plot showing the shift in mCitrine expression when comparing a single expression cassette (cit.1x) against a multiple gene construct harboring nine gene expression units (cit.9x). (**E**) Expected mCitrine expression values for the four multi-copy constructs (circles) were calculated by adding up expression values from individual mCitrine cassettes. Observed fluorescence values for the pYES1L-mCherry-mCitrine.x series (diamonds), normalized to pYES1L-mCherry-mCitrine.1x, are also shown. The expected (dashed line) and observed (solid line) gene dosage-mCitrine expression trajectories (best-fit lines) are reported. (**F**) Bar chart showing IFNg production levels as the number of IFNg gene copies increases from one to nine. IFNg secretion levels were normalized by mCherry expression levels.

DISCUSSION

A community effort is ongoing to deconstruct the attributes of optimal producer cell lines and reverse engineer them with the recently developed and refined tools for genome editing and pathway engineering, with the key objective of streamlining the cell line development process. Thus far, work has primarily focused on the identification of transcriptional hotspots in the host cell line's genome (mainly CHO cells) and the establishment of genetic platforms for the targeted integration of transgenes (see (58) for a review). Despite the central role occupied by gene amplification in the current cell line development scheme, more limited has been the progress in the direction of programmed transgene dosage increase (26,27), possibly owing to the difficulties connected with handling large amounts of repetitive sequences in the form of repeated gene expression cassettes. The main objective of this work is to provide a new strategy for achieving controlled transgene dosage increase with the goal of rationally designing, from the bottom up, genomic islands for bioproduction whose territory is fully mapped, in contrast to the current cell line development landscape where neither the integration sites nor transgene copy number are defined, even in the producer clones that are ultimately selected for a production project.

Key to our strategy for overcoming the hurdles connected with the bottom-up design of bioproduction islands - resting primarily with the handling of repetitive sequences - is the diversification of the individual structural elements of a gene expression cassette, whose function—protein expression—must nonetheless be preserved. Dissecting a gene expression cassette to its essential components, we developed a computational workflow for the automated design of both coding units (genes) and regulatory units (promoters and 3′ UTRs). Using these components, it was possible to assemble up to 10 gene expression cassettes in a single, homologous recombination-based reaction, and subsequently deliver our multiple gene constructs to the CHO production vector. As previously mentioned, the need for component diversification stems from the choice of the plasmid assembly pipeline—yeast assembly, previously employed to fuse up to 25 overlapping DNA fragments in a single reaction (59)—and as a requirement for preventing genomic instability that might potentially ensue following integration into the host genome. In an attempt to better define and gauge the need for sequence diversification, we considered each of the biological systems required for multiple gene construct assembly, propagation and expression.

For plasmid assembly in the yeast *S. cerevisiae,* we chose terminal overlaps between recombining fragments that are

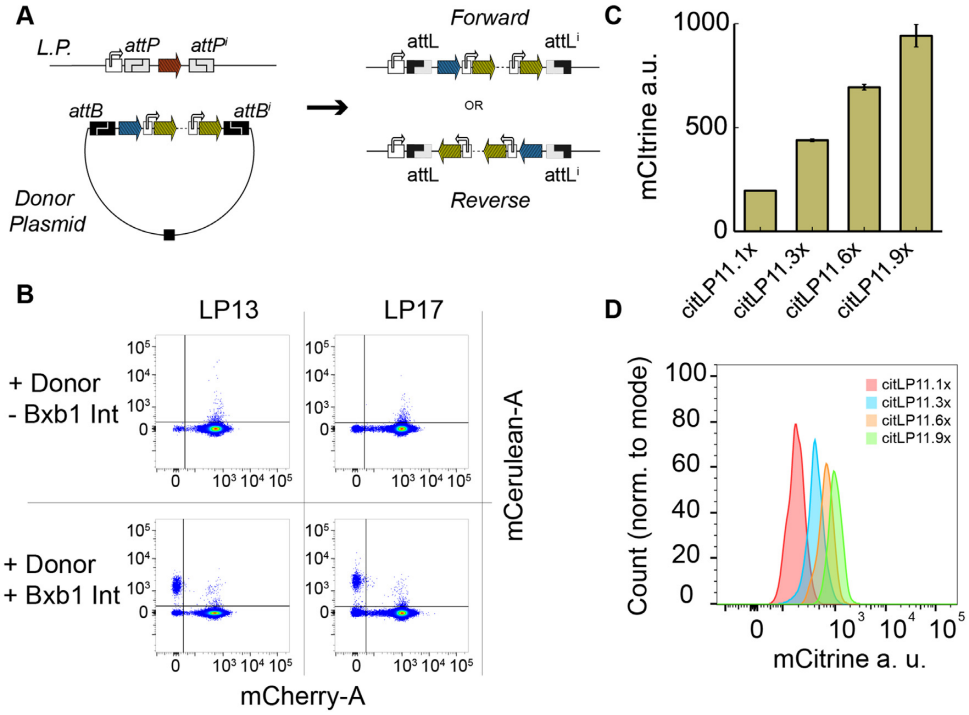

Figure 7. Stable integration of multi-copy constructs into the CHO genome. (**A**) Schematic diagram of the landing pad (LP) and donor plasmid structure (left side of the panel). Two attachment sites (attP and its reverse complement, attP', represented by the gray rectangles) were placed on our landing pad which was integrated into the host chromosome. Donor plasmids are designed with two donor sites facing each other, attB and attB' (black rectangles). When acceptor and donor sites react in the presence of Bxb1, recombination takes place, resulting in targeted exchange and the integration of donor material in either the forward or reverse orientation relative to the landing pad on the genome (right side of the panel). In the diagram, landing pad marker (mCherry), sentinel marker (mCerulean) and production genes (mCitrine genes) are shown with the red, blue and yellow block arrows, respectively. Rectangles with curved arrows represent promoters (3′-UTRs and insulators not shown). (**B**) RMCE experimental design. Flow cytometry diagrams (10 days after transfection of LP lines with donor and recombinase) showing that RMCE requires both donor plasmid DNA and Bxb1 Int (only two LPs are shown here). Conversion of the LP marker (mCherry) to the donor marker (mCerulean) signals a successful genome targeting event. Every RMCE experiment always included a control where no Bxb1 recombinase was added, to confirm that marker conversion was specifically attributable to Bxb1-mediated cassette exchange. (**C**) Bar chart showing average mCitrine fluorescence values of mCerulean(+)/mCherry(-)/mCitrine(+) recombinant populations after targeting of LP11 with the pYES1L-attB/B'-mCerulean-mCitrine1x-9x vector series. Across all recombinant lines, only the largest active subpopulations—that are also characterized by the highest level of mCitrine expression—were considered for this analysis. (**D**) Flow cytometry histograms showing the shift in mCitrine expression as a function of mCitrine copy number.

80 bp in length (40 bp for fragment-backbone overlaps). It was previously shown that, during yeast homologous recombination, internal homologies between fragments can also lead to recombination, but homologous termini have a higher 'recombination potential' than internal stretches of homology (53). This might explain why, even in the presence of rather extensive internal homologies within core promoter regions (required for preserving promoter function) (see Figure 3A), we could still assemble large multiple gene constructs. Only one promoter and one 3′ UTR scaffold were employed, but libraries of regulators built on multiple scaffolds could ensure more genetic variation between different gene expression cassettes and, therefore, improve assembly efficiency and/or support even larger assemblies.

Once correctly assembled plasmids are obtained via yeast homologous recombination, the resulting multiple gene constructs need to be propagated in yeast and subsequently transferred to *E. coli*, from which large quantities of plasmid DNA can be prepared, before genome-editing in the CHO host can be performed. Sequence repeats, on plasmid DNA or in the genome, are associated with instability, in the form of deletions and duplications of the repeated segments, both in prokaryotes and in eukaryotes, with a number of different mechanisms that can underlie rearrangements, including both homologous and nonhomologous recombination, fuelled by slippage and misalignment events during replication (60–62). The connection between replication and the genetic instability of repeated segments may contribute to explaining why low-copy-number plasmid vectors (bacterial artificial chromosomes, BACs) were required for the propagation in yeast and *E. coli* of our largest constructs, while high-copy number vectors became unstable when more than four gene cassettes were fused together (data not shown).

Finally, while studies in mammalian cells on the requirements for intrachromosomal homologous recombination showed that sequence identities of approximately 300 bp are needed for efficient recombination ($\sim 10^{-7}$ events/cell division with 295 bp of homology), with a 100-fold reduction in recombination rates with 95 bp of homology (63), high-order genomic architecture—and not sequence homology alone—also appears to play an important role in chromosomal stability (64). In view of this, we suggest that a number of integration loci be screened for their abil-

ity to support prolonged, stable expression of a large bioproduction island (and not just of a single reporter protein) prior to targeting the host genome with large genetic payloads.

Within the traditional framework for recombinant protein expression, depending on what transfection and amplification methods are chosen, high-producing clones can reportedly bear up to a few hundred copies of the transgene (65,66). However, it is usually unknown how many transgene copies are transcriptionally active and it has been suggested that transcriptional activity serves as a better predictor of protein yield than the number of integrated gene copies, pointing to the possibility that only a moderate number of transgenes integrated into transcriptionally active loci may be required for efficient production (67). With our cassette diversification and assembly strategy, we were able to rapidly assemble up to nine GOI-bearing cassettes and a fluorescent marker-expressing cassette and integrate the resulting multiple gene constructs into a pre-validated CHO locus, without antibiotic or metabolic selection. While a higher number of gene copies may be assembled, not surprisingly RMCE rates were found to drop as payload size increased. In this regard, it is conceivable that improving the transfection efficiency of plasmid donor DNA through, for instance, a thorough evaluation of an array of transfection reagents may help achieve higher rates of genome targeting and recombination, easing cell line development for industrial applications and enabling the integration of larger payloads. A further increase in copy number may also be supported by adopting a hybrid strategy centered on embedding a metabolic selection marker (e.g. the DHFR or GS coding sequence) within a multiple gene construct in order to enrich for recombinant cell populations, but without the subsequent use of gene amplification agents. With this strategy, it may be possible to target more than one landing pad simultaneously, thereby effectively integrating, for instance, 20 or 30 transgene copies into the host genome (if two or three landing pads were to be targeted with a donor construct bearing ten transgene copies). In combination with ongoing cell line engineering efforts aimed at enhancing a host cell's capacity for recombinant protein expression (e.g. through improved protein secretion), the approach presented here may help innovate the cell line development pipeline and reduce the time and effort required to enter the manufacturing phase of a recombinant protein production project.

DATA AVAILABILITY

DNA sequences specifying all coding and regulatory elements used to assemble our multiple gene constructs are provided in the Supplementary Data. Novel sequences have been deposited in the NCBI database. Accession numbers are provided in Supplementary Tables S2-S7.

The gene diversifier and optimizer software is freely available for download at https://github.com/altamurr/GeneDiversifier.

Flow cytometry data has been deposited in FlowRepository. Links to FlowRepository are available in the section "Flow Cytometry Data" in the Supplementary Data.

Microbial and mammalian strains as well as plasmid constructs used to complete this study are available under a Material Transfer Agreement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Prof. Beat Christen, Fabian Rudolf, Dr. Robert Gnügge and Dr. Kristina Elfström for help with setting up the yeast assembly pipeline. We thank Verena Jäggin, Dr. Mariangela Di Tacchio and Dr. Aleksandra Gumienny for help with single cell sorting and Dr. Daniel Meyer for help with plasmid cloning automation.

FUNDING

NCCR Molecular Systems Engineering and by ETH Zurich. Funding for open access charge: SNSF.
Conflict of interest statement. The authors report no conflict of interest regarding the subject matter of this study.

REFERENCES

1. Gronemeyer,P., Ditz,R. and Strube,J. (2014) Trends in upstream and downstream process development for antibody manufacturing. *Bioengineering*, **1**, 188–212.
2. Walsh,G. (2018) Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.*, **36**, 1136–1145.
3. Wurm,F.M. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.*, **22**, 1393–1398.
4. Fan,L., Kadura,I., Krebs,L.E., Hatfield,C.C., Shaw,M.M. and Frye,C.C. (2012) Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnol. Bioeng.*, **109**, 1007–1015.
5. Cacciatore,J.J., Chasin,L.A. and Leonard,E.F. (2010) Gene amplification and vector engineering to achieve rapid and high-level therapeutic protein production using the Dhfr-based CHO cell selection system. *Biotechnol. Adv.*, **28**, 673–681.
6. Kingston,R.E., Kaufman,R.J., Bebbington,C.R. and Rolfe,M.R. (2002) Amplification using CHO cell expression vectors. *Curr. Protoc. Mol. Biol.*, https://doi.org/10.1002/0471142727.mb1623s60.
7. Osterlehner,A., Simmeth,S. and Göpfert,U. (2011) Promoter methylation and transgene copy numbers predict unstable protein production in recombinant chinese hamster ovary cell lines. *Biotechnol. Bioeng.*, **108**, 2670–2681.
8. Lieske,P.L., Wei,W., Crowe,K.B., Figueroa,B. and Zhang,L. (2020) HIF-1 signaling pathway implicated in phenotypic instability in a Chinese hamster ovary production cell line. *Biotechnol. J.*, **15**, e1900306.
9. Kellems,R.E., Wurm,F.M. and Pallavicini,M.G. (2020) Effects of methotrexate on recombinant sequences in mammalian cells. *Gene Amplif. Mamm. Cells*, **10**, 85–94.
10. Bailey,L.A., Hatton,D., Field,R. and Dickson,A.J. (2012) Determination of Chinese hamster ovary cell line stability and recombinant antibody expression during long-term culture. *Biotechnol. Bioeng.*, **109**, 2093–2103.
11. Bandyopadhyay,A.A., O'Brien,S.A., Zhao,L., Fu,H.Y., Vishwanathan,N. and Hu,W.S. (2019) Recurring genomic structural variation leads to clonal instability and loss of productivity. *Biotechnol. Bioeng.*, **116**, 41–53.
12. Kaufman,R.J. and Schimke,R.T. (1981) Amplification and loss of dihydrofolate reductase genes in a Chinese hamster ovary cell line. *Mol. Cell. Biol.*, **1**, 1069–1076.
13. Fann,C.H., Guirgis,F., Chen,G., Lao,M.S. and Piret,J.M. (2000) Limitations to the amplification and stability of human tissue-type plasminogen activator expression by Chinese hamster ovary cells. *Biotechnol. Bioeng.*, **69**, 204–212.

14. Sakuma,T., Takenaga,M., Kawabe,Y., Nakamura,T., Kamihira,M. and Yamamoto,T. (2015) Homologous recombination-independent large gene cassette knock-in in CHO cells using TALEN and MMEJ-directed donor plasmids. *Int. J. Mol. Sci.*, **16**, 23849–23866.
15. Zhao,M., Wang,J., Luo,M., Luo,H., Zhao,M., Han,L., Zhang,M., Yang,H., Xie,Y., Jiang,H. *et al.* (2018) Rapid development of stable transgene CHO cell lines by CRISPR/Cas9-mediated site-specific integration into C12orf35. *Appl. Microbiol. Biotechnol.*, **102**, 6105–6117.
16. Lee,J.S., Kallehauge,T.B., Pedersen,L.E. and Kildegaard,H.F. (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep.*, **5**, 8572.
17. Inniss,M.C., Bandara,K., Jusiak,B., Lu,T.K., Weiss,R., Wroblewska,L. and Zhang,L. (2017) A novel Bxb1 integrase RMCE system for high fidelity site-specific integration of mAb expression cassette in CHO Cells. *Biotechnol. Bioeng.*, **114**, 1837–1846.
18. Huang,Y., Li,Y., Wang,Y.G., Gu,X., Wang,Y. and Shen,B.F. (2007) An efficient and targeted gene integration system for high-level antibody expression. *J. Immunol. Methods*, **322**, 28–39.
19. Zhang,L., Inniss,M.C., Han,S., Moffat,M., Jones,H., Zhang,B., Cox,W.L., Rance,J.R. and Young,R.J. (2015) Recombinase-mediated cassette exchange (RMCE) for monoclonal antibody expression in the commercially relevant CHOK1SV cell line. *Biotechnol. Prog.*, **31**, 1645–1656.
20. Grav,L.M., Sergeeva,D., Lee,J.S., Marin De Mas,I., Lewis,N.E., Andersen,M.R., Nielsen,L.K., Lee,G.M. and Kildegaard,H.F. (2018) Minimizing clonal variation during mammalian cell line engineering for improved systems biology data generation. *ACS Synth. Biol.*, **7**, 2148–2159.
21. Scarcelli,J.J., Shang,T.Q., Iskra,T., Allen,M.J. and Zhang,L. (2017) Strategic deployment of CHO expression platforms to deliver Pfizer's Monoclonal Antibody Portfolio. *Biotechnol. Prog.*, **33**, 1463–1467.
22. Baser,B., Spehr,J., Büssow,K. and van den Heuvel,J. (2016) A method for specifically targeting two independent genomic integration sites for co-expression of genes in CHO cells. *Methods*, **95**, 3–12.
23. Kameyama,Y., Kawabe,Y., Ito,A. and Kamihira,M. (2010) An accumulative site-specific gene integration system using Cre recombinase-mediated cassette exchange. *Biotechnol. Bioeng.*, **105**, 1106–1114.
24. Kawabe,Y., Makitsubo,H., Kameyama,Y., Huang,S., Ito,A. and Kamihira,M. (2012) Repeated integration of antibody genes into a pre-selected chromosomal locus of CHO cells using an accumulative site-specific gene integration system. *Cytotechnology*, **64**, 267–279.
25. Carver,J., Ng,D., Zhou,M., Ko,P., Zhan,D., Yim,M., Shaw,D., Snedecor,B., Laird,M.W., Lang,S. *et al.* (2020) Maximizing antibody production in a targeted integration host by optimization of subunit gene dosage and position. *Biotechnol. Prog.*, **36**, e2967.
26. Gaidukov,L., Wroblewska,L., Teague,B., Nelson,T., Zhang,X., Liu,Y., Jagtap,K., Mamo,S., Allen Tseng,W., Lowe,A. *et al.* (2018) A multi-landing pad DNA integration platform for mammalian cell engineering. *Nucleic Acids Res.*, **46**, 4072–4086.
27. Sergeeva,D., Lee,G.M., Nielsen,L.K. and Grav,L.M. (2020) Multicopy targeted integration for accelerated development of high-producing Chinese hamster ovary cells. *ACS Synth. Biol.*, **9**, 2546–2561.
28. Hossain,A., Lopez,E., Halper,S.M., Cetnar,D.P., Reis,A.C., Strickland,D., Klavins,E. and Salis,H.M. (2020) Automated design of thousands of nonrepetitive parts for engineering stable genetic systems. *Nat. Biotechnol.*, **38**, 1466–1475.
29. Gaspar,P., Moura,G., Santos,M.A.S. and Oliveira,J.L. (2013) mRNA secondary structure optimization using a correlated stem-loop prediction. *Nucleic Acids Res.*, **41**, 3–7.
30. Sharp,M.P. and Li,W.H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
31. Lorenz,R., Bernhart,S., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P. and Hofacker,I. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
32. Gu,W., Zhou,T. and Wilke,C.O. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **6**, e1000664.
33. Ringnér,M. and Krogh,M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
34. Kudla,G., Murray,A.W., Tollervey,D. and Plotkin,J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.
35. Bentele,K., Saffert,P., Rauscher,R., Ignatova,Z. and Blüthgen,N. (2013) Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.*, **9**, 675.
36. Wakabayashi-Ito,N. and Nagata,S. (1994) Characterization of the regulatory elements in the promoter of the human elongation factor-1α gene. *J. Biol. Chem.*, **269**, 29831–29837.
37. Brown,A.J., Sweeney,B., Mainwaring,D.O. and James,D.C. (2014) Synthetic promoters for CHO cell engineering. *Biotechnol. Bioeng.*, **111**, 1638–1647.
38. Levitt,N., Briggs,D., Gil,A. and Proudfoot,N.J. (1989) Definition of an efficient synthetic poly(A) site. *Genes Dev.*, **3**, 1019–1025.
39. Liu,M., Maurano,M.T., Wang,H., Qi,H., Song,C.Z., Navas,P.A., Emery,D.W., Stamatoyannopoulos,J.A. and Stamatoyannopoulos,G. (2015) Genomic discovery of potent chromatin insulators for human gene therapy. *Nat. Biotechnol.*, **33**, 198–203.
40. Gibson,D.G. (2011) Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol.*, **498**, 349–361.
41. Lois,C., Hong,E.J., Pease,S., Brown,E.J. and Baltimore,d, D. (2002) Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science*, **295**, 868–872.
42. Dull,T., Zufferey,R., Kelly,M., Mandel,R.J., Nguyen,M., Trono,D. and Naldini,L. (1998) A third-generation lentivirus vector with a conditional packaging system. *J. Virol.*, **72**, 8463–8471.
43. Tiscornia,G., Singer,O. and Verma,I.M. (2006) Production and purification of lentiviral vectors. *Nat. Protoc.*, **1**, 241–245.
44. Kudla,G., Lipinski,L., Caffin,F., Helwak,A. and Zylicz,M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.*, **4**, 0933–0942.
45. Newman,Z.R., Young,J.M., Ingolia,N.T. and Barton,G.M. (2016) Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E1362–E1371.
46. Ding,Y., Shah,P. and Plotkin,J.B. (2012) Weak 5′-mRNA secondary structures in short eukaryotic genes. *Genome Biol. Evol.*, **4**, 1046–1053.
47. Guo,D., Gao,A., Michels,D.A., Feeney,L., Eng,M., Chan,B., Laird,M.W., Zhang,B., Yu,X.C., Joly,J. *et al.* (2010) Mechanisms of unintended amino acid sequence changes in recombinant monoclonal antibodies expressed in Chinese Hamster Ovary (CHO) cells. *Biotechnol. Bioeng.*, **107**, 163–171.
48. Babendure,J.R., Babendure,J.L., Ding,J.H. and Tsien,R.Y. (2006) Control of mammalian translation by mRNA structure near caps. *RNA*, **12**, 851–861.
49. Qin,J.Y., Zhang,L., Clift,K.L., Hulur,I., Xiang,A.P., Ren,B.Z. and Lahn,B.T. (2010) Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One*, **5**, 3–6.
50. Hans,H. and Alwine,J.C. (2000) Functionally significant secondary structure of the simian virus 40 late polyadenylation signal. *Mol. Cell. Biol.*, **20**, 2926–2932.
51. Cheng,J.K., Morse,N.J., Wagner,J.M., Tucker,S.K. and Alper,H.S. (2019) Design and evaluation of synthetic terminators for regulating mammalian cell transgene expression. *ACS Synth. Biol.*, **8**, 1263–1275.
52. Liu,M., Maurano,M.T., Wang,H., Qi,H., Song,C.-Z., Navas,P.A., Emery,D.W., Stamatoyannopoulos,J.A. and Stamatoyannopoulos,G. (2015) Genomic discovery of potent chromatin insulators for human gene therapy. *Nat. Biotechnol.*, **33**, 198–203.
53. Ma,H., Kunes,S., Schatz,P.J. and Botstein,d, D. (1987) Plasmid construction by homologous recombination in yeast (Saccharomyces cerevisiae; transformation; plasmid recombination; YCpSO derivatives; YEp420 [previously called 8721 derivatives). *Gene*, **58**, 253–3623.
54. Gnügge,R., Liphardt,T. and Rudolf,F. (2016) A shuttle vector series for precise genetic engineering of Saccharomyces cerevisiae. *Yeast*, **33**, 83–98.
55. Le Fourn,V., Girod,P.-A., Buceta,M., Regamey,A. and Mermod,N. (2014) CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metab. Eng.*, **21**, 91–102.
56. Rahimpour,A., Vaziri,B., Moazzami,R., Nematollahi,L., Barkhordari,F., Kokabee,L., Adeli,A. and Mahboudi,F. (2013) Engineering the cellular protein secretory pathway for enhancement

of recombinant tissue plasminogen activator expression in chinese hamster ovary cells: effects of CERT and XBP1s genes. *jmb*, **23**, 1116–1122.
57. Turan,S. and Bode,J. (2011) Site-specific recombinases: from tag-and-target- to tag-and-exchange-based genomic modifications. *FASEB J.*, **25**, 4088–4107.
58. Hamaker,N.K. and Lee,K.H. (2018) Site-specific integration ushers in a new era of precise CHO cell line engineering. *Curr. Opin. Chem. Eng.*, **22**, 152–160.
59. Gibson,D.G., Benders,G.A., Axelrod,K.C., Zaveri,J., Algire,M.A., Moodie,M., Montague,M.G., Venter,J.C., Smith,H.O. and Hutchison,C.A. (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 20404–20409.
60. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
61. Morag,A.S., Saveson,C.J. and Lovett,S.T. (1999) Expansion of DNA repeats in Escherichia coli: effects of recombination and replication functions. *J. Mol. Biol.*, **289**, 21–27.
62. Bzymek,M. and Lovett,S.T. (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 8319–8325.
63. Liskay,R.M., Letsou,A. and Stachelek,J.L. (1987) Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics*, **115**, 161–167.
64. Stankiewicz,P., Shaw,C.J., Dapper,J.D., Wakui,K., Shaffer,L.G., Withers,M., Elizondo,L., Park,S. and Lupski,J.R. (2003) Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am. J. Hum. Genet.*, **72**, 1101–1116.
65. Kim,S.J. and Lee,G.M. (1999) Cytogenetic analysis of chimeric antibody-producing CHO cells in the course of dihydrofolate reductase-mediated gene amplification and their stability in the absence of selective pressure. *Biotechnol. Bioeng.*, **64**, 741–749.
66. Jiang,Z., Huang,Y. and Sharfstein,S.T. (2006) Regulation of recombinant monoclonal antibody production in Chinese hamster ovary cells: a comparative study of gene copy number, mRNA level, and protein expression. *Biotechnol. Prog.*, **22**, 313–318.
67. Noh,S.M., Shin,S. and Lee,G.M. (2018) Comprehensive characterization of glutamine synthetase-mediated selection for the establishment of recombinant CHO cells producing monoclonal antibodies. *Sci. Rep.*, **8**, 5361.