

TAGS FOR ENHANCED EXPRESSION OF RECOMBINANT PROTEINS

FIELD OF THE INVENTION

The present invention is in the field of recombinant biotechnology, in particular
5 in the field of protein expression. The invention generally relates to tag sequences
encoding peptide tags capable of increasing the yield and/or titer and/or solubility of
at least one protein of interest in a host cell, in particular difficult-to-express recom-
binant proteins. More specifically, the invention relates to a peptide tag encoded by
a novel nucleotide sequence comprising selected codons, including rare codons,
10 and to host cells and expression vectors comprising such nucleotide sequence. The
invention further relates to a method of producing at least one protein of interest in
a recombinant host cell using said tag sequences.

BACKGROUND OF THE INVENTION

15 *Escherichia coli* is one of the most commonly used host organisms for the pro-
duction of biopharmaceuticals, as it allows for cost-efficient and fast recombinant
protein expression. Today, about 30% of approved biopharmaceuticals are pro-
duced by *E. coli*, and despite the recent advance of other production systems such
as mammalian cell culture, new product approvals, like Beovu® and Cablivi^R (2019),
20 are constantly being granted, allowing *E. coli* to compete with both mammalian and
non-mammalian hosts (McElawain et al. 2022, Baeshen et al. 2015). However, *E.*
coli is not equally suited for the expression of all proteins, as it lacks the ability for
several key posttranslational modifications such as glycosylation, cytosolic disul-
phide bond formation and proteolytic protein maturation. Additionally, the expres-
25 sion of complex recombinant proteins in *E. coli* can lead to accumulation of the pro-
tein of interest (POI) in inclusion bodies (IB) or low expression titers.

To obtain an efficient recombinant protein production process, several factors
play a key role, such as the basic choice of the expression system design, like plas-
mid-based or genome-integrated expression cassettes, or the choice of promoter
30 and many more. The effects of codon usage of the respective transcripts on protein
expression efficiency have also been considered as central to enhance protein ex-
pression efficiency. Early on, the codon adaption index (CAI), a measure to calculate
similarity between the codon usage of a gene sequence and the codon preference

of a host organism, has been developed as a tool for the prediction of the expression level of a gene (Sharp & Li, 1987) and its optimization is since widely used for the adaption of heterologous gene sequences. However, this view has been recently challenged. By introducing multiple silent mutations in a reporter gene, Kudla et al. 5 (2009) could not identify an influence of the CAI or the frequency of optimal codons on reporter gene expression when considering the whole coding sequence as well as only the first 42 nucleotides, nor was the frequency of rare codons or number of pairs of consecutive rare codons significantly correlated with reporter gene expression.

10 There are continuous efforts to enhance solubility of proteins, e.g. by the N-terminal fusion of peptide extensions with large net negative charge, so called solubility tags, to a target protein, wherein the solubility tag increases the electrostatic repulsion between nascent polypeptides and thereby limits protein aggregation (Cserjan-Puschmann et al. 2020). For example, WO2021/028590A1 describes the 15 CASPON™ platform technology, employing a solubility tag to facilitate a generic combinatorial fusion tag-based platform process for the high-titer soluble expression of biopharmaceuticals including a standardized downstream processing and highly specific enzymatic cleavage of the fusion tag, leaving the POI with its native N-terminus (Lingg et al. 2022, Köppl et al. 2022, Kröß et al. 2021, Cserjan-Puschmann 20 et al. 2020). This technology has been proven to be very well suited for a variety of different biopharmaceutical model proteins, outperforming previously published recombinant model protein titers. However, POI-dependent differences in expression levels can be observed (Köppl et al. 2022).

Various attempts were made in the art for improving tag sequences to aid in the 25 production of a protein of interest, most requiring multiple and laboursome modifications, including modifications in the target sequence.

Goodman et al. (2013), screened over 14,000 combinations of promoters, ribosome binding sites and N-terminal peptide sequences comprising rare codons, driving the expression of a codon optimized super-folder GFP and found that increased 30 secondary structure was correlated with decreased expression.

US 10,385,350 B2 discloses a method of preparing an expression vector, comprising a 5' UTR sequence followed by a peptide tag sequence, wherein both

sequences are modified to minimize RNA secondary structure within and/or between these two regions.

However, there remains an urgent need for improved, efficient and hence more profitable tools for protein production.

5

BRIEF SUMMARY OF THE INVENTION

The production of recombinant proteins is currently still limited by various aspects, including protein solubility and translation rate. This holds especially true for complex recombinant proteins, where currently available methods still suffer from low expression titers.

10

It is the objective of the present invention to overcome current limitations in protein biosynthesis by introducing a novel nucleic acid sequence encoding for a peptide tag sequence for the high-titer expression of proteins.

The objective is solved by the present invention.

15

According to the invention, there is provided an isolated nucleic acid sequence encoding a peptide tag comprising amino acid sequence SEQ ID NO:1 for expression of a protein of interest (POI) in a host cell, wherein said nucleic acid sequence comprises SEQ ID NO:2 comprising at least 3 silent mutations at positions selected from the group consisting of:

20

a. codon encoding Pro at position 1 of SEQ ID NO:1, wherein CCG has been modified to CCC or CCA;

b. codon encoding Arg at position 3 of SEQ ID NO:1, wherein CGC has been modified to AGG, AGA or CGA;

25

c. codon encoding Asn at position 4 of SEQ ID NO:1, wherein AAC has been modified to AAT;

d. codon encoding Glu at position 6 of SEQ ID NO:1, wherein GAG has been modified to GAA;

e. codon encoding Arg at position 7 of SEQ ID NO:1, wherein CGA has been modified to AGG, or AGA; and

30

f. codon encoding Lys at position 8 of SEQ ID NO:1, wherein AAG has been modified to AAA.

According to a specific embodiment the peptide tag comprises SEQ ID NO:4 and the nucleic acid sequence described herein comprises SEQ ID NO:5 comprising at least 3 silent mutations selected from the group consisting of:

- a. codon encoding Leu at position 1 of SEQ ID NO:4, wherein CTG has been modified to CTA, TTA or TTG;
- b. codon encoding Glu at position 2 of SEQ ID NO:4, wherein GAG has been modified to GAA;
- c. codon encoding Pro at position 4 of SEQ ID NO:4, wherein CCG has been modified to CCC or CCA;
- d. codon encoding Arg at position 6 of SEQ ID NO:4, wherein CGC has been modified to AGG, AGA or CGA;
- e. codon encoding Asn at position 7 of SEQ ID NO:4, wherein AAC has been modified to AAT;
- f. codon encoding Glu at position 9 of SEQ ID NO:4, wherein GAG has been modified to GAA;
- g. codon encoding Arg at position 10 of SEQ ID NO:4, wherein CGA has been modified to AGG, or AGA; and
- h. codon encoding Lys at position 11 of SEQ ID NO:4, wherein AAG has been modified to AAA.

Specifically, the nucleic acid sequences described herein, comprise at least 4 silent mutations, preferably at least 5 silent mutations and even more preferably at least 6 silent mutations.

Specifically, the nucleic acid sequences described herein, comprise at least 7 silent mutations, preferably at least 8 silent mutations.

According to a further specific embodiment described herein, the nucleic acid sequences described herein encode a peptide tag for expression of a protein of interest (POI) in a host cell, wherein the host cell is a eukaryotic or prokaryotic host cell, preferably a yeast cell, a mammalian cell or a bacterial cell.

Specifically, the host cell described herein is a bacterial cell, preferably *E. coli*.

Specifically, the host cell described herein is a yeast cell, preferably of the genus *Pichia*, even more preferably *Komagataella phaffii* or *Komagataella pastoris*.

Specifically, the host cell described herein is a mammalian cell, preferably a CHO cell.

According to a further specific embodiment described herein, the nucleic acid sequences described herein comprise a nucleic acid sequence selected from the group consisting of SEQ ID NO:3, SEQ ID NO:6, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20 and SEQ ID NO:42.

According to a further specific embodiment described herein, the nucleic acid sequences described herein comprise a nucleic acid sequence selected from the group consisting of SEQ ID NO:3, SEQ ID NO:6, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20 and SEQ ID NO:42.

Further specifically provided herein is a peptide tag encoded by any one of the nucleic acid sequences described herein.

As described herein, the nucleic acid sequence of the invention encodes at least part of a peptide tag, that can be used to facilitate expression of a recombinant POI. Said peptide tag comprises at least SEQ ID NO:1. The peptide tag may comprise additional elements N-terminal or C-terminal of SEQ ID NO:1.

In a specific example, the peptide tag comprises at least 3 additional amino acids N-terminal of SEQ ID NO:1. Specifically, the 3 additional amino acids are L, E and D, resulting in the amino acid sequence of SEQ ID NO:4.

In a specific example, the peptide tag encoded by the nucleic acid sequence of the invention comprises further amino acids C-terminal of SEQ ID NO:1 or SEQ ID NO:4. Specifically, the peptide tag comprises at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, or 40 additional amino acids C-terminal of SEQ ID NO:1 or SEQ ID NO:4. In a specific example, the peptide tag comprises 11 additional amino acids, resulting for example in the amino acid sequence of SEQ ID NO:7 or SEQ ID NO:12.

Specifically, the nucleic acid sequence of the invention, encoding a peptide tag of SEQ ID NO:7, is selected from the group consisting of SEQ ID NO:9, SEQ ID NO:10 and SEQ ID NO:11.

Specifically, the nucleic acid sequence of the invention, encoding a peptide tag of SEQ ID NO:12, is selected from the group consisting of SEQ ID NO:14, SEQ ID NO:15 and SEQ ID NO:16.

The peptide tag described herein may further comprise, C-terminal or N-terminal of SEQ ID NO:1 or SEQ ID NO:4, any one or more or all of the following elements:

- a. one or more affinity tags,
- b. one or more solubility enhancement tags,
- 5 c. one or more monitoring tags,
- d. one or more linkers,
- e. one or more protease recognition and/or cleavage sites.

Specifically, the affinity tag is selected from the group consisting of poly-histidine tag, poly-arginine tag, peptide substrate for antibodies, chitin binding domain, 10 RNase S peptide, protein A, β -galactosidase, FLAG tag, Strep II tag, streptavidin-binding peptide (SBP) tag, calmodulin-binding peptide (CBP), glutathione S-transferase (GST), maltose-binding protein (MBP), S-tag, HA tag, c-Myc tag, SUMO tag, E.coli thioredoxin, NusA, chitin binding domain CBD, chloramphenicol acetyl transferase CAT, LysRS, ubiquitin, calmodulin, and lambda gpV, specifically the tag is a 15 His tag comprising one or more His, more specifically it is a hexahistidine tag (6-His or 6H).

Specifically, the solubility enhancement tag is selected from the group consisting of T7C, T7B, T7B1, T7B2, T7B3, T7B3, T7B4, T7B5, T7B6, T7B6, T7B7, T7B8, 20 T7B9, T7B10, T7B11, T7B12, T7B13, T7A, T7A1, T7A2, T7A3, T7A4, T7A5, T7AC, T3, N1, N2, N3, N4, N5, N6, N7, calmodulin-binding peptide (CBP), poly Arg, poly Lys, G B1 domain, protein D, Z domain of Staphylococcal protein A, DsbA, DsbC and thioredoxin.

Specifically, the monitoring tag is selected from the group consisting of m-Cherry, GFP, YFP and f-Actin.

25 According to a specific embodiment, the peptide tag described herein comprises more than one additional tags, specifically it comprises an affinity tag and a solubility enhancement tag. Specifically, it comprises an affinity tag, a solubility enhancement tag and a monitoring tag. Specifically, it comprises more than one tag of the same functionality, specifically it comprises more than one affinity tag, more than one solubility enhancement tag and/or more than one monitoring tag, and any combination 30 thereof. Specifically, the peptide tag described herein comprises a C-terminal and an N-terminal tag, each comprising one or more tag sequences, preferably selected from affinity tag, solubility enhancement tag and monitoring tag.

According to a further specific embodiment, the peptide tag described herein comprises one or more linker sequences, specifically consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 or even more amino acid residues. Specifically, the linker sequence comprises glycine, alanine and/or serine residues. Specifically, the linker
5 comprises at least one glycine and serine residue, more specifically the linker is GS, GSG, GSGGG, GSGSGSGS and/or GSAGSAAGSG.

In a specific example, the peptide tag described herein comprises preferably in the following order, a T7AC tag, a 6-His-tag, a GSG linker and a caspase-2 recognition and cleavage site, preferably a caspase-2 recognition and cleavage site com-
10 prising the amino acid sequence of VDVAD.

In a further specific example, the peptide tag described herein comprises, preferably in the following order, a T7AC tag, a 6-His-tag, optionally an SA linker, optionally a Strep II tag, a GSG linker and a caspase-2 recognition and cleavage site. Specifically, the peptide tag comprises SEQ ID NO:40, which sequence includes a
15 starting methionine. The nucleic acid sequence of the invention encoding a peptide tag of SEQ ID NO:40, thus may be SEQ ID NO:42, which comprises SEQ ID NO:9.

Further specifically provided herein is an expression vector comprising any one of the nucleic acid sequences described herein.

Specifically described herein is an expression vector comprising any one of the
20 nucleic acid sequences described herein encoding a peptide tag, optionally a protease recognition site and/or cleavage site, optionally a linker, and a gene of interest (GOI).

According to a further embodiment described herein, the expression vector comprises: a promoter sequence, a nucleic acid sequence encoding a 5' untranslated region of an expressed mRNA that comprises a ribosome binding site (RBS),
25 any one of the nucleic acid sequences described herein encoding a peptide tag as described herein, optionally a linker, optionally a protease recognition site and/or cleavage site, and a cloning site, wherein the cloning site enables a gene of interest (GOI) encoding a POI to be inserted into the vector in-frame with any one of the
30 nucleic acid sequences described herein to encode a fusion protein comprising said peptide tag, optionally a linker and/or a protease recognition and/or cleavage site, and said POI. According to a specific embodiment, the expression vector described

herein comprises a GOI, preferably inserted at the cloning site in-frame with the nucleic acid sequence described herein to encode a fusion protein.

Specifically, the expression vector comprises a nucleic acid sequence as described herein encoding a peptide tag comprising, preferably in the following order,
5 a T7AC tag, a 6-His tag, optionally an SA linker, optionally a Strep II tag, followed by a GSG linker and a protease recognition and cleavage site, which preferably is the caspase-2 recognition and cleavage site VDVAD (SEQ ID NO:49). Specifically, said protease recognition and cleavage site is directly linked to the 5' end of the GOI encoding a POI.

10 Further specifically provided herein is a host cell comprising any one of the expression vectors described herein.

Further specifically described herein is a host cell comprising any one of the nucleic acid sequences, expression cassettes and/or expression vectors described herein integrated into its genome.

15 Further specifically described herein is a host cell comprising any one of the nucleic acid sequences, expression cassettes and/or expression vectors described herein integrated into its chromosome(s).

Specifically, said host cell is a eukaryotic or prokaryotic host cell, preferably a yeast cell or a bacterial cell. Specifically, the host cell is a bacterial cell, preferably
20 *E. coli*. Specifically, the host cell is a yeast cell, preferably of the genus *Pichia*, even more preferably *Komagataella phaffii*. Specifically, the host cell is a mammalian cell, preferably a CHO cell.

Further specifically provided herein is a method for the production of a recombinant protein of interest (POI), comprising culturing any one of the host cells described herein comprising a GOI for a period of time under conditions permitting
25 expression of a POI in the form of a fusion protein comprising the peptide tag described herein. According to a specific embodiment, the method further comprises any one or more of the steps of:

- a. recovering the fusion protein and/or POI,
- 30 b. purifying the fusion protein and/or POI,
- c. removing the peptide tag,
- d. further purifying the fusion protein and/or POI,
- e. formulating the fusion protein and/or POI,

f. modifying the fusion protein and/or POI.

According to a specific example, the method of producing a POI thus may comprise the following steps, preferably in that order:

a. recovering the fusion protein,

5 b. purifying the fusion protein,

c. removing the peptide tag to release the POI, preferably with an authentic N-terminus,

d. optionally further purifying the POI,

e. optionally modifying the POI, and

10 f. formulating the POI, e.g. with a pharmaceutically acceptable carrier.

BRIEF SUMMARY OF THE DRAWINGS

Figure 1: Schematic overview of tested constructs. **A:** General structure of a bacterial 5'UTR in the context of the fusion protein. The T7 promoter is depicted as arrow while the ribosomal binding site is illustrated as rectangle. **B:** N-terminal variants of the T7AC tag and their minimal free folding energies of nucleotides -4 to +37. The original T7AC tag sequence is depicted as well as the T7AC_{new} version differing in codon usage and minimal free folding energy.

Figure 2: Growth behaviour and soluble product formation kinetics of T7AC and T7AC_{new} tagged model proteins in identical carbon-limited fed-batch fermentations (fermentation strategy I: $\mu = 0.1 \text{ h}^{-1}$ for 3.6 generations; production during the last 1.2 generations). **A to C** show the cellular growth behaviour of the model proteins PTH, hFGF2 and TNF α respectively. **D to E** depict the soluble product formation kinetics of the model proteins PTH, hFGF2 and TNF α respectively.

Figure 3: Direct comparison of specific recombinant protein titres at the end of fermentations. Solid bars represent the T7AC tagged variants, while hatched bars depict their T7AC_{new} tagged counterparts.

Figure 4: Growth behaviour and soluble product formation kinetics of T7AC and T7AC_{new} tagged TNF α in identical carbon-limited fed-batch fermentations (fermentation strategy II: $\mu = 0.1 \text{ h}^{-1}$ for 1.6 generations followed by 0.05 h^{-1} for 2.0 generations; production during the last 1.7 generations). **A:** cellular growth behaviour. **B:** soluble product formation kinetics.

Figure 5: Schematic overview of tested T7AC and T7ACnew N-terminal variant constructs. N-terminal variants of the tags and their minimal free folding energies of nucleotides -4 to +37.

Figure 6: Recombinant protein production kinetics for the PTH fusion proteins (A and B) and the hFGF2 fusion proteins (D and E) and biomass formation for the PTH fusion proteins (C) and the hFGF2 fusion proteins (F) over the time course of the fermentations. For amino acid sequences and nucleotide sequences of the fusion constructs see Table 14.

Figure 7: Direct comparison of specific end-of-fermentation titres for all tested constructs. For amino acid sequences and nucleotide sequences of the fusion constructs see Table 14.

Figure 8: Soluble product formation kinetics for PTH fusion proteins. Volumetric titer [g/L] of fermentations of PTH fusion proteins expressed from nucleic acid sequences comprising nucleic acid sequences of the present invention compared to the original nucleic acid sequence (T7AC) encoding the T7AC tag.

Figure 9: Soluble product formation kinetics for TNF α fusion proteins. Volumetric titer [g/L] of fermentations of TNF α fusion proteins expressed from nucleic acid sequences comprising nucleic acid sequences of the present invention compared to the original nucleic acid sequence (T7AC) encoding the T7AC tag.

Figure 10: Amino acid and nucleotide sequences referred to herein.

DETAILED DESCRIPTION OF THE INVENTION

A primary objective of the present invention is to enhance expression, solubility and/or proper folding of recombinant proteins of interest expressed in host cells. Accordingly, the present invention enables the production of biologically active non-prokaryotic and prokaryotic proteins within host cells in sufficient quantities, *e.g.* for industrial and/or medical applications. As demonstrated in the Examples provided herein, prokaryotic cells (*e.g.* *E. coli*) represent an important example of host cells, however, low titres, solubility problems and the formation of inclusion bodies are well-known issues faced in eukaryotic host cells as well.

The present invention is based on the surprising finding that introducing modifications in the nucleotide sequence encoding a peptide tag comprising the amino acid sequence SEQ ID NO:1 significantly enhances expression of target proteins

fused to such peptide tag. Said modifications in the nucleotide sequence of the peptide tag are silent mutations including the introduction of rare codon(s).

Use of the novel nucleotide sequences described herein significantly increased expression of several model proteins as is shown in the Examples provided herein.

5 It was particularly unexpected that the use of rare codons in said peptide tag could so strongly enhance the expression of several POIs because it had been previously thought that low translation rates caused by the use of rare codons lead to slow and inefficient product formation (Dana & Tuller 2014; Sørensen et al. 1989; Kane 1995). Furthermore, it had previously been reported that the introduction of rare codons

10 into highly expressed genes induces the depletion of rare tRNA molecules and retention of ribosomes, therefore inducing changes in translation efficiency of numerous genes and consequently having detrimental effects on cellular fitness (Kudla et al. 2009, Frumkin et al. 2018). In the context of protein biosynthesis, a reduction of protein product formation and/or cellular fitness imposes a severe reduction in production efficiency and profitability. Contrary to these studies, however, cell growth

15 behaviour upon use of the novel tag sequences described herein was not adversely affected, despite the use of rare codons in the tag sequences, in fact, it was even better at later time points (Figure 2 A-C).

The present invention thus provides means to efficiently improve existing strategies of recombinant protein production and to obtain a target protein at a high yield,

20 such that active protein required for biotechnology and pharmaceuticals may be supplied in mass-production.

Tag sequences of the invention

25 Provided herein is a nucleic acid sequence encoding a peptide tag comprising the amino acid sequence PERNKERK (SEQ ID NO:1), which peptide tag is useful for expression of a protein or peptide of interest (POI) in a host cell.

As used herein, the term "tag sequence" refers to the nucleic acid sequence of the invention. The nucleic acid sequence according to the invention encodes at least

30 the amino acid sequence PERNKERK (SEQ ID NO:1) and comprises SEQ ID NO:2 comprising at least 3 silent mutations at positions selected from the group consisting of:

- a. codon encoding Pro at position 1 of SEQ ID NO:1, wherein CCG has been modified to CCC or CCA;
- b. codon encoding Arg at position 3 of SEQ ID NO:1, wherein CGC has been modified to AGG, AGA or CGA;
- 5 c. codon encoding Asn at position 4 of SEQ ID NO:1, wherein AAC has been modified to AAT;
- d. codon encoding Glu at position 6 of SEQ ID NO:1, wherein GAG has been modified to GAA;
- e. codon encoding Arg at position 7 of SEQ ID NO:1, wherein CGA has been modified to AGG, or AGA; and
- 10 f. codon encoding Lys at position 8 of SEQ ID NO:1, wherein AAG has been modified to AAA.

In a specific example the nucleic acid sequence according to the invention encodes at least the amino acid sequence PERNKERK (SEQ ID NO:1) and comprises
15 SEQ ID NO:2 comprising at least 3 silent mutations at positions selected from the group consisting of:

- a. codon encoding Pro at position 1 of SEQ ID NO:1, wherein CCG has been modified to CCC or CCA;
- b. codon encoding Arg at position 3 of SEQ ID NO:1, wherein CGC has been modified to AGG or CGA;
- 20 c. codon encoding Asn at position 4 of SEQ ID NO:1, wherein AAC has been modified to AAT;
- d. codon encoding Glu at position 6 of SEQ ID NO:1, wherein GAG has been modified to GAA;
- 25 e. codon encoding Arg at position 7 of SEQ ID NO:1, wherein CGA has been modified to AGG, or AGA; and
- f. codon encoding Lys at position 8 of SEQ ID NO:1, wherein AAG has been modified to AAA.

Specifically, the nucleic acid sequence described herein comprises SEQ ID
30 NO:2, wherein at least 4, 5, 6, 7, or 8 codons have been modified to introduce silent mutations. Such modifications may be achieved by introducing single or multiple nucleotide substitutions.

In a specific example, the first codon of SEQ ID NO:2, which encodes Pro at position 1 of SEQ ID NO:1, has been modified. Specifically, CCG has been replaced by CCC or CCA, e.g. by substituting the base G with C or A. Thus, in this example, instead of CCG, the nucleic acid sequence of the invention comprises CCC or CCA at this position.

In another specific example, the third codon of SEQ ID NO:2, which encodes Arg at position 3 of SEQ ID NO:1, has been modified. Specifically, CGC has been replaced by AGG, AGA or CGA. Thus, in this example, instead of CGC, the nucleic acid sequence of the invention comprises AGG, AGA or CGA at this position.

In another specific example, the third codon of SEQ ID NO:2, which encodes Arg at position 3 of SEQ ID NO:1, has been modified. Specifically, CGC has been replaced by AGG or CGA. Thus, in this example, instead of CGC, the nucleic acid sequence of the invention comprises AGG or CGA at this position.

In another specific example, the fourth codon of SEQ ID NO:2, which encodes Asn at position 4 of SEQ ID NO:1, has been modified. Specifically, AAC has been replaced by AAT, e.g. by substituting the base C with T. Thus, in this example, instead of AAC, the nucleic acid sequence of the invention comprises AAT at this position.

In another specific example, the sixth codon of SEQ ID NO:2, which encodes Glu at position 6 of SEQ ID NO:1, has been modified. Specifically, GAG has been replaced by GAA, e.g. by substituting the base G with A. Thus, in this example, instead of GAG, the nucleic acid sequence of the invention comprises GAA at this position.

In another specific example, the seventh codon of SEQ ID NO:2, which encodes Arg at position 7 of SEQ ID NO:1, has been modified. Specifically, CGA has been replaced by AGG or AGA. Thus, in this example, instead of CGA, the nucleic acid sequence of the invention comprises AGG or AGA at this position.

In another specific example, the eighth codon of SEQ ID NO:2, which encodes Lys at position 8 of SEQ ID NO:1, has been modified. Specifically, AAG has been replaced by AAA, e.g. by substituting the base G with A. Thus, in this example, instead of AAG, the nucleic acid sequence of the invention comprises AAA at this position.

Various combinations of the listed codon substitutions are possible. Preferably, more than one of the silent mutations result in rare codons. Rare codons typically are codons with a codon usage of less than 1% in the host cell, or less than 10 in 1000 codons.

5 In a specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:3.

In a further specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:17.

10 In yet a further specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:18.

The peptide tag encoded by the nucleic acid sequence described herein may be longer than the sequence of SEQ ID NO:1. In a specific embodiment the peptide tag encoded by the nucleic acid sequence of the invention comprises the amino acid sequence LEDPERNKERK (SEQ ID NO:4). The nucleic acid sequence according to the invention encoding such amino acid sequence LEDPERNKERNK (SEQ ID NO:4) comprises SEQ ID NO:5 comprising at least 3 silent mutations selected from the group consisting of:

- a. codon encoding Leu at position 1 of SEQ ID NO:4, wherein CTG has been modified to CTA, TTA or TTG;
- 20 b. codon encoding Glu at position 2 of SEQ ID NO:4, wherein GAG has been modified to GAA;
- c. codon encoding Pro at position 4 of SEQ ID NO:4, wherein CCG has been modified to CCC or CCA;
- d. codon encoding Arg at position 6 of SEQ ID NO:4, wherein CGC has been modified to AGG, AGA or CGA;
- 25 e. codon encoding Asn at position 7 of SEQ ID NO:4, wherein AAC has been modified to AAT;
- f. codon encoding Glu at position 9 of SEQ ID NO:4, wherein GAG has been modified to GAA;
- 30 g. codon encoding Arg at position 10 of SEQ ID NO:4, wherein CGA has been modified to AGG, or AGA; and
- h. codon encoding Lys at position 11 of SEQ ID NO:4, wherein AAG has been modified to AAA.

- In a specific embodiment the peptide tag encoded by the nucleic acid sequence of the invention comprises the amino acid sequence LEDPERNKERK (SEQ ID NO:4). The nucleic acid sequence according to the invention encoding such amino acid sequence LEDPERNKERNK (SEQ ID NO:4) comprises SEQ ID NO:5 comprising at least 3 silent mutations selected from the group consisting of:
- a. codon encoding Leu at position 1 of SEQ ID NO:4, wherein CTG has been modified to CTA, TTA or TTG;
 - b. codon encoding Glu at position 2 of SEQ ID NO:4, wherein GAG has been modified to GAA;
 - 10 c. codon encoding Pro at position 4 of SEQ ID NO:4, wherein CCG has been modified to CCC or CCA;
 - d. codon encoding Arg at position 6 of SEQ ID NO:4, wherein CGC has been modified to AGG or CGA;
 - e. codon encoding Asn at position 7 of SEQ ID NO:4, wherein AAC has been modified to AAT;
 - 15 f. codon encoding Glu at position 9 of SEQ ID NO:4, wherein GAG has been modified to GAA;
 - g. codon encoding Arg at position 10 of SEQ ID NO:4, wherein CGA has been modified to AGG, or AGA; and
 - 20 h. codon encoding Lys at position 11 of SEQ ID NO:4, wherein AAG has been modified to AAA.

Specifically, the nucleic acid sequence of the invention comprises SEQ ID NO:5, wherein at least 4, 5, 6, 7, 8, 9, 10 or 11 codons have been modified to introduce silent mutations. In a specific example, the first codon of SEQ ID NO:5, which encodes Leu at position 1 of SEQ ID NO:5, has been modified. Specifically, CTG has been replaced by CTA, TTA or TTG. Thus, in this example, instead of CTG, the nucleic acid sequence of the invention comprises CTA, TTA or TTG at this position.

In another specific example, the second codon of SEQ ID NO:5, which encodes Glu at position 2 of SEQ ID NO:5, has been modified. Specifically, GAG has been replaced by GAA, e.g. by substituting the base G with A. Thus, in this example, instead of GAG, the nucleic acid sequence of the invention comprises GAA at this position.

In a specific example, the fourth codon of SEQ ID NO:5, which encodes Pro at position 4 of SEQ ID NO:5, has been modified. Specifically, CCG has been replaced by CCC or CCA, e.g. by substituting the base G with C or A. Thus, in this example, instead of CCG, the nucleic acid sequence of the invention comprises CCC or CCA at this position.

In another specific example, the 6th codon of SEQ ID NO:5, which encodes Arg at position 6 of SEQ ID NO:5, has been modified. Specifically, CGC has been replaced by AGG, AGA or CGA. Thus, in this example, instead of CGC, the nucleic acid sequence of the invention comprises AGG, AGA or CGA at this position.

In another specific example, the 6th codon of SEQ ID NO:5, which encodes Arg at position 6 of SEQ ID NO:5, has been modified. Specifically, CGC has been replaced by AGG or CGA. Thus, in this example, instead of CGC, the nucleic acid sequence of the invention comprises AGG or CGA at this position.

In another specific example, the 7th codon of SEQ ID NO:5, which encodes Asn at position 7 of SEQ ID NO:5, has been modified. Specifically, AAC has been replaced by AAT, e.g. by substituting the base C with T. Thus, in this example, instead of AAC, the nucleic acid sequence of the invention comprises AAT at this position.

In another specific example, the 9th codon of SEQ ID NO:5, which encodes Glu at position 9 of SEQ ID NO:5, has been modified. Specifically, GAG has been replaced by GAA, e.g. by substituting the base G with A. Thus, in this example, instead of GAG, the nucleic acid sequence of the invention comprises GAG at this position.

In another specific example, the 10th codon of SEQ ID NO:5, which encodes Arg at position 10 of SEQ ID NO:5, has been modified. Specifically, CGA has been replaced by AGG or AGA. Thus, in this example, instead of CGA, the nucleic acid sequence of the invention comprises AGG or AGA at this position.

In another specific example, the 11th codon of SEQ ID NO:5, which encodes Lys at position 11 of SEQ ID NO:5, has been modified. Specifically, AAG has been replaced by AAA, e.g. by substituting the base G with A. Thus, in this example, instead of AAG, the nucleic acid sequence of the invention comprises AAA at this position.

In a specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:6.

In a further specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:19.

In yet a further specific embodiment, the nucleic acid sequence of the invention thus comprises or consists of SEQ ID NO:20.

5 The peptide tag encoded by the nucleic acid sequence of the invention may be longer than the sequence of SEQ ID NO:1 or SEQ ID NO:4. In a specific embodiment, the peptide tag comprises an amino acid elongation at its C-terminus. Specifically, the peptide tag comprises at least 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or 20, or 30, or 40 or more additional amino acids, preferably at its C-terminus.

10 For example, the peptide tag comprises or consists of SEQ ID NO:7 and the nucleic acid sequence of the invention encoding such tag may comprise or consist of SEQ ID NO: 9, SEQ ID NO:10 or SEQ ID NO:11.

 According to another example, the peptide tag comprises or consists of SEQ ID NO:12 and the nucleic acid sequence of the invention encoding such tag may com-
15 prise or consist of SEQ ID NO:14, SEQ ID NO:15 or SEQ ID NO: 16.

 The tag sequence described herein may further comprise nucleic acid sequences encoding one or more affinity tags, one or more solubility enhancement tags, one or more monitoring tags, one or more linkers, and/or one or more protease recognition and/or cleavage sites as further described herein.

20 In a specific example, the tag sequence described herein encodes a peptide tag comprising, preferably in the following order, a T7AC tag, a 6-His-tag, optionally an SA linker, optionally a Strep II tag, a GSG linker and a caspase-2 recognition and cleavage site. Specifically, the peptide tag comprises SEQ ID NO:40, which sequence includes a starting methionine. The nucleic acid sequence of the invention
25 encoding the peptide tag of SEQ ID NO:40, thus may be SEQ ID NO:42.

Used Terms and Definitions

 Unless indicated or defined otherwise, all terms used herein have their usual meaning in the art, which will be clear to the skilled person. Reference is for example
30 made to standard handbooks, such as Sambrook et al, "Molecular Cloning: A Laboratory Manual" (2nd Ed.), Vols. 1–3, Cold Spring Harbor Laboratory Press (1989); Lewin, "Genes IV", Oxford University Press, New York, (1990), and Janeway et al,

"Immunobiology" (5th Ed., or more recent editions, Garland Science, New York, 2001).

The subject matter of the claims specifically refers to artificial products or methods employing or producing such artificial products, which may be variants of native (wild-type) products. Though there can be a certain degree of sequence identity to the native structure, it is well understood that the materials, methods and uses of the invention, e.g., specifically referring to isolated nucleic acid sequences, amino acid sequences, fusion constructs, expression constructs, transformed host cells and modified proteins, are "man-made" or synthetic, and are therefore not considered as a result of "laws of nature".

It must be noted that as used herein, the singular forms "a", "an" and "the" include plural references and vice versa unless the context clearly indicates otherwise. Thus, for example, a reference to "a host cell" or "a method" includes one or more of such host cells or methods, respectively, and a reference to "the method" includes equivalent steps and methods that could be modified or substituted known to those of ordinary skill in the art. Similarly, for example, a reference to "methods" or "host cells" includes "a host cell" or "a method", respectively.

Unless otherwise indicated, the term "at least" preceding a series of elements is to be understood to refer to every element in the series. Those skilled in the art will recognize or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the present invention.

The term "and/or" wherever used herein includes the meaning of "and", "or" and "all or any other combination of the elements connected by said term". For example, A, B and/or C means A, B, C, A+B, A+C, B+C and A+B+C.

The term "about" or "approximately" as used herein means within 20%, preferably within 10%, and more preferably within 5% of a given value or range. It includes also the concrete number, e.g., about 20 includes 20.

The term "less than", "more than" or "larger than" includes the concrete number. For example, less than 20 means ≤ 20 and more than 20 means ≥ 20 .

Throughout this specification and the claims, unless the context requires otherwise, the word "comprise" and variations such as "comprises" and "comprising" will be understood to imply the inclusion of a stated integer (or step) or group of integers

(or steps). It does not exclude any other integer (or step) or group of integers (or steps). When used herein, the term "comprising" can be substituted with "containing", "composed of", "including", "having" or "carrying" and vice versa, by way of example the term "having" can be substituted with the term "comprising". When used
5 herein, "consisting of" excludes any integer or step not specified in the claim/item.

It should be understood that this invention is not limited to the particular methodology, protocols, material, reagents, and substances, *etc.*, described herein. The terminologies used herein are for the purpose of describing particular embodiments only and are not intended to limit the scope of the present invention, which is defined
10 solely by the claims.

All publications and patents cited throughout the text of this specification (including all patents, patent applications, scientific publications, manufacturer's specifications, instructions, *etc.*), whether *supra* or *infra*, are hereby incorporated by reference in their entirety. Nothing herein is to be construed as an admission that the
15 invention is not entitled to antedate such disclosure by virtue of prior invention. To the extent the material incorporated by reference contradicts or is inconsistent with this specification, the specification will supersede any such material.

The term "nucleic acid" refers to deoxyribonucleic acid (DNA, *e.g.* a cDNA or genomic DNA), ribonucleic acid (RNA, *e.g.* a mRNA), or a DNA or RNA analog and
20 polymers thereof, in either single- or double-stranded form, but preferably is double-stranded DNA, made of monomers (nucleotides) containing a sugar, phosphate and a base that is either a purine or pyrimidine.

A "polynucleotide" as used herein, refers to nucleotides, either ribonucleotides (RNA, *e.g.*, an mRNA) or deoxyribonucleotides (DNA, *e.g.*, a cDNA) or a combination of both, in a polymeric unbranched form of any length. Preferably, a polynucleotide refers to deoxyribonucleotides in a polymeric unbranched form of any length. Here, nucleotides consist of a pentose sugar (deoxyribose), a nitrogenous base (adenine, guanine, cytosine or thymine) and a phosphate group. The terms "polynucleotide(s)", "nucleic acid sequence(s)" are used interchangeably herein. Unless otherwise indicated, a particular nucleic acid sequence also encompasses complementary sequences.
25
30

The term "isolated," as used herein, refers to material that is removed from its original or native environment (*e.g.* the natural environment if it is naturally

occurring). For example, a naturally-occurring nucleic acid molecule or polypeptide present in a living animal is not isolated, but the same nucleic acid molecule or polypeptide, separated by human intervention from some or all of the co-existing materials in the natural system, is isolated. Such nucleic acid molecules could be part of a vector and/or such nucleic acid molecules or polypeptides could be part of a composition, and still be isolated in that such vector or composition is not part of the environment in which the nucleic acid molecule or the polypeptide is found in nature. For example, a nucleic acid molecule or polypeptide is considered to be "(in) isolated (form)" when, compared to its native biological source and/or the reaction medium or cultivation medium from which it has been obtained, it has been separated from at least one other component with which it is usually associated in said source or medium, such as another nucleic acid molecule, another polypeptide, another biological component or macromolecule or at least one contaminant, impurity or minor component.

Amino acid residues will be indicated according to the standard three-letter or one-letter amino acid code, as generally known and agreed upon in the art. There are twenty known naturally occurring amino acids encoded by sixty-one triplet codons. These 20 amino acids can be split into those that have neutral charges, positive charges, and negative charges:

The "neutral" amino acids are shown below along with their respective three-letter and single-letter code and polarity:

Alanine: (Ala, A) nonpolar, neutral;
Asparagine: (Asn, N) polar, neutral;
Cysteine: (Cys, C) polar, neutral;
Glutamine: (Gln, Q) polar, neutral;
Glycine: (Gly, G) nonpolar, neutral;
Isoleucine: (Ile, I) nonpolar, neutral;
Leucine: (Leu, L) nonpolar, neutral;
Methionine: (Met, M) nonpolar, neutral;
Phenylalanine: (Phe, F) nonpolar, neutral;
Proline: (Pro, P) nonpolar, neutral;
Serine: (Ser, S) polar, neutral;
Threonine: (Thr, T) polar, neutral;

Tryptophan: (Trp, W) nonpolar, neutral;

Tyrosine: (Tyr, Y) polar, neutral;

Valine: (Val, V) nonpolar, neutral; and

Histidine: (His, H) polar, positive (10%) neutral (90%).

5 The "positively" charged amino acids are:

Arginine: (Arg, R) polar, positive; and

Lysine: (Lys, K) polar, positive.

The "negatively" charged amino acids are:

Aspartic acid: (Asp, D) polar, negative; and

10 Glutamic acid: (Glu, E) polar, negative.

The terms "peptide", "polypeptide" and "protein" are used herein interchangeably and refer to the arrangement of amino acid residues in a polymer (polypeptide chain). A peptide, polypeptide or protein can be composed of the standard 20 naturally occurring amino acids, and may include rare amino acids and synthetic amino acid analogs (e.g. non-canonical amino acids). A protein can be a monomer, a dimer
15 or a higher -mer comprising one, two or more polypeptide chains. The protein can be a homo-di- or higher -mer comprising two or more identical polypeptide chains or a hetero-di- or higher -mer comprising two or more different polypeptide chains or a mixture thereof, such as a heteromer of a homomer or a homomer of a heteromer.
20 The protein and the polypeptide chains can be any chain of amino acids, regardless of length or post-translational modification (e.g. glycosylation or phosphorylation). The peptide, polypeptide, protein or polypeptide chain(s) described herein includes recombinantly or synthetically produced peptide tags described herein, either individually or as part of a fusion with e.g. a POI.

25 The genetic code translates mRNA nucleotide sequences to amino acid sequences. Genetic information is coded using this process with groups of three nucleotides along the mRNA which are commonly known as "codons". The set of three nucleotides almost always produce the same amino acid, with a few exceptions like UGA which typically serves as the stop codon but can also encode tryptophan in
30 mammalian mitochondria. Most amino acids are specified by multiple codons demonstrating that the genetic code is degenerate—different codons result in the same amino acid. Codons that code for the same amino acid are termed synonyms or synonymous codons.

A point mutation is particularly understood as the engineering of a polynucleotide that results in the expression of an amino acid sequence that differs from the non-engineered amino acid sequence in the substitution or exchange, deletion or insertion of one or more single (non-consecutive) or doublets of amino acids for
5 different amino acids.

As used herein, the term "silent mutation(s)" refers to base substitutions that result in no change of the amino acid or amino acid functionality when the altered messenger RNA (mRNA) is translated. Specifically, it refers to a substitution of at least one nucleotide in a codon that does not result in a change of the respective
10 amino acid. For example, if the codon AAA is altered to become AAG, the same amino acid – lysine - will be incorporated into the peptide chain.

The term "recombinant" as used herein shall mean "being prepared by or the result of genetic engineering". A "recombinant cell" or "recombinant host cell" refers to a cell or host cell that has been genetically altered to comprise a nucleic acid
15 sequence which was not native to said cell. A recombinant host specifically comprises a recombinant expression vector or cloning vector, or it has been genetically engineered to contain a recombinant nucleic acid sequence in its genome, such as its chromosome.

As used herein, the term "homologous" means derived from the same cell or
20 organism with the same genomic background. As used herein, the term "heterologous" means derived from a cell or organism with a different genomic background, or an artificial sequence not found in nature. Thus, a heterologous nucleotide sequence, *e.g.* a DNA, can be a DNA encoding an artificial protein not found in nature.

As used herein the term "endogenous" means originating inside the organism
25 or cell. Specifically, use of an endogenous DNA, *e.g.* for expression of a protein, means that the DNA originating inside the organism or cell is used for expression. As used herein the term "exogenous" means originating outside the organism or cell. Specifically, use of an exogenous sequence, such as DNA, for expression of a protein, means that the sequence is introduced into the cell, where expression from
30 this sequence takes place.

A heterologous or a homologous gene for expression of a heterologous or homologous protein can be introduced into the cell by integrating a vector carrying the gene (there are one or more copies of the vector and thus one or more copies of the

gene in the host cell) and being expressed from the vector or one or more copies of the gene can be integrated into the genome or chromosome of the host cell, from where the gene is expressed. A homologous protein can also be expressed or over-expressed in a host cell from an endogenous gene. In this case, the tag sequence
5 described herein and optionally a promoter, is integrated into the genome or chromosome of the host cell such that it is operably linked to the endogenous gene.

A recombinant protein also may be a homologous protein. In this case, one or more copies of the polynucleotide encoding the homologous protein are introduced into the host cell by genetic manipulation.

10 Specifically, heterologous nucleotide sequences are those not found in the same relationship to a host cell in nature (i.e., "not natively associated"). Any recombinant or artificial nucleotide sequence is understood to be heterologous.

The term "increasing the expression of at least one protein of interest" means that the yield and/or the titer and/or secretion and/or the soluble yield and/or soluble
15 titer of a POI which is expressed by a host cell is increased, when said POI is expressed from a fusion construct comprising a tag sequence of the invention compared to expression by the same host cell but from a fusion construct comprising SEQ ID NO:2 or 5 instead of the tag sequence of the invention.

The terms "yield" or "specific titer" also refers to the amount of POI(s) as described herein per cell or biomass and may be presented by mg POI/g biomass
20 (biomass being measured as dry cell weight or cell dry mass (CDM) or wet cell weight (WCW), preferably measured as CDM so that "yield" is present by mg POI/g (CDM) of a host cell. The term "titer" or "titre" or "volumetric titer" when used herein refers similarly to the amount of produced POI or model protein(s) as described
25 herein per volume and may be presented as mg POI/L culture supernatant or whole cell broth. The expressed POI(s) can be located within the cytosol and/or periplasm (in soluble form or in insoluble form, specifically as Inclusion Bodies ("IB")) of the host cell and/or the supernatant of the cell culture or cell broth. Thus, the terms "titer" (volumetric titer) and "yield" (specific titer) can be used to refer to "soluble titer" /
30 "soluble volumetric titer" or "soluble yield" / "soluble specific titer" for the part of the POI or fusion protein which, is expressed solubly and "total titer" / "total volumetric titer" or "total yield" / "total specific titer" for the whole amount of the POI or fusion protein including soluble and insoluble fractions, and "IB titer" / "IB volumetric titer"

or "IB yield" / "IB specific titer" for the part of the POI or fusion protein which is in the form of IBs.

With regard to the protein or polypeptide of interest (POI) there are no limitations. More specifically, the protein may either be a polypeptide not naturally occurring in the host cell, *i.e.* a heterologous protein. The POI may further be a homologous protein to the host cell, but is produced, for example, upon integration by recombinant techniques of one or more copies of the nucleic acid sequence encoding the homologous POI into a host cell. The POI may be produced upon integration of the nucleic acid sequence encoding the POI into the genome or chromosome of the host cell. According to a further example, the POI can also be expressed in a host using a vector, more specifically a plasmid. Thus, the POI can be expressed from a plasmid or from the chromosome or genome of the host cell. In case the POI is homologous to the host cell, a nucleic acid sequence comprising the tag sequence of the invention, and, optionally, a promoter different to the native promoter of the endogenous gene, may be integrated into the host cell so that it is operably linked to the endogenous gene encoding the homologous POI.

The POI can be a monomer, dimer or multimer, it can be a homomer or heteromer or a mixture thereof. The POI can be any peptide, protein or polypeptide, whether naturally occurring or not. The POI can also be a synthetic peptide, protein or polypeptide. Examples for proteins that can be produced by the method of the invention are, without limitation, enzymes, regulatory proteins, receptors, growth factors, hormones, peptides, *e.g.* peptide hormones, cytokines, membrane or transport proteins. The POIs may also be antigens as used for vaccination, vaccines, antigen-binding proteins, immune stimulatory proteins, interleukins, interferons, allergens, full-length antibodies or antibody fragments or derivatives thereof or affinity scaffolds. Antibody derivatives may be for example, but not limited to single chain variable fragments (scFv), Fab fragments or single domain antibodies or camelid antibodies or heavy chain antibodies or derivatives thereof such as V_{HH} fragments or the like. The POI can be an artificial protein comprising polypeptide chains comprising natural as well as artificial amino acid sequences and/or comprising non-canonical amino acids. The POI can also exclusively comprise artificial sequences or can comprise more than one artificial polypeptide chains.

The DNA molecule encoding the protein of interest is also termed "gene of interest" or "GOI". The gene of interest encoding the POI can be a naturally existing DNA sequence or a non-natural DNA sequence. One or more GOI(s) can be under the control of one promoter. Alternatively, each gene of interest is under the control of its own promoter. If the POI is a heteromer, all different polypeptide chains may be under the control of the same promoter, or each polypeptide chain may be under the control of its own promoter or a mixture thereof. The genes of interest resp. the genes encoding the polypeptide chains of a heteromer may all be on the same expression cassette or on multiple expression cassettes. The POI can be modified in any way. Non-limiting examples for modifications can be insertion or deletion of post-translational modification sites, insertion or deletion of targeting signals (e.g.: leader peptides), fusion to tags (e.g. improving expression and/or solubility of the POI, affinity tags such as 6His-tag for IMAC purification, detection tags such as GFP and/or any tag having other functions), protease recognition and/or cleavage sites, proteins or protein fragments facilitating purification or detection, mutations affecting changes in stability or changes in solubility, or any other modification known in the art. In certain embodiments of the invention the recombinant protein is a biopharmaceutical product, which can be any protein suitable for therapeutic or prophylactic purposes in mammals.

As used herein the term "host cell" refers to a cell which is capable of protein expression and optionally protein secretion. Such host cell is applied in the methods of the present invention. For that purpose, for the host cell to produce a recombinant POI, a tag sequence as described herein is introduced or present in the cell. Typically, the term refers to viable cells, capable of growing in a cell culture. Common cells that serve as hosts for expression of recombinant genes are prokaryotic and eukaryotic cells, such as bacteria, yeast, or mammalian cells. Examples of eukaryotic cells include, but are not limited to, vertebrate cells, mammalian cells, human cells, animal cells, invertebrate cells, plant cells, nematodal cells, insect cells, stem cells, fungal cells, or yeast cells.

Examples of bacterial cells include, but are not limited to *Escherichia coli*, *Bacillus species*, *Corynebacterium species*, *Pseudomonas species*, *Salmonella species* and *Streptomyces species*. Specific examples of *E. coli* strains include but are not limited to B strains and K strains, such as BL21, BL21 (DE3), HMS 174 (DE3).

The eukaryotic host cell may be a fungal cell. More preferred is a yeast host cell. Examples of yeast cells include but are not limited to the *Saccharomyces* genus (e.g. *Saccharomyces cerevisiae*, *Saccharomyces kluyveri*, *Saccharomyces uvarum*), the *Komagataella* genus (*Komagataella pastoris*, *Komagataella pseudo-*
5 *pastoris* or *Komagataella phaffii*), *Kluyveromyces* genus (e.g. *Kluyveromyces lactis*, *Kluyveromyces marxianus*), the *Candida* genus (e.g. *Candida utilis*, *Candida ca- caoi*), as well as *Hansenula polymorpha*.

In a preferred embodiment, the genus *Pichia* is of particular interest. *Pichia* comprises a number of species, including the species *Pichia pastoris*, *Pichia methanol-*
10 *ica*, *Pichia kluyveri*, and *Pichia angusta*. Most preferred is the species *Pichia pas- toris*. The former species *Pichia pastoris* has been divided and renamed to *Koma- gataella pastoris*, *Komagataella phaffii* and *Komagataella pseudopastoris*. There- fore, *Pichia pastoris* is a synonymous for both *Komagataella pastoris*, *Komagataella phaffii* and *Komagataella pseudopastoris*.

15 Non-limiting examples of mammalian cells include, without being limiting, hu- man, mice, rat, monkey and rodent cells lines. Specific mammalian cell lines avail- able as host cells for expression are well known in the art and include, inter alia, Chinese hamster ovary (CHO) cells, NS0, SP2/0 cells, HeLa cells, baby hamster kidney (BHK) cells, monkey kidney cells (COS), human carcinoma cells (e.g., Hep
20 G2 and A-549 cells), 3T3 cells or the derivatives/progenies of any such cell line.

Appropriate culture media and conditions for the above-described host cells are known in the art.

Alternatively, the gene can be translated into protein using cell free translation systems, possibly coupled to an *in vitro* transcription system. These systems provide
25 all steps necessary to obtain protein from DNA by supplying the necessary enzymes and substrates in an *in vitro* reaction. In principle, any living cell or organism can provide the necessary enzymes for this process and extraction protocols for obtain- ing such enzyme systems are known in the art. Common systems used for *in vitro* transcription/translation are extracts or lysates from reticulocytes, wheat germ or
30 *Escherichia coli*.

The term "expression" is understood in the following way. Nucleic acid mole- cules containing a desired coding sequence of an expression product such as e.g., a peptide tag, a POI or a fusion protein as described herein, may be used for

expression purposes. Hosts transformed or transfected with or containing these sequences are capable of producing the encoded proteins. In order to effect transformation, the expression system may be included in a vector; however, the relevant DNA may also be integrated into the host chromosome or genome. Specifically, the term refers to a host cell and compatible vector under suitable conditions, *e.g.*, for the expression of a protein coded for by foreign DNA carried by the vector and introduced to the host cell.

A POI or a fusion protein can be expressed as soluble protein within the cytoplasm or as soluble protein within the periplasm (*e.g.* upon fusion of a leader or signal sequence to the POI or fusion protein). A POI or a fusion protein can also be expressed in soluble form and released to the culture supernatant by active transport of the POI or fusion protein from the cytosol, periplasm, or endoplasmic reticulum to the supernatant. For transport of the POI or fusion protein through the cell membrane(s) or membrane(s) of the involved cell organelles, a leader or signal sequence (specifically in bacterial expression systems) or a pre and/or pro sequence (specifically in yeast expression systems) may be fused to the N-terminus of the POI or fusion protein. The POI or fusion protein can also be released from the periplasm to the supernatant upon cell lysis or leakiness of the cell membrane. This may be achieved by addition of enzymes inducing lysis of the cells or by inducing expression of a cell lytic enzyme within the host cell during or after fermentation. A POI or fusion protein can also be expressed as inclusion bodies within the cytosol or, for example when an N-terminal signal or leader peptide is used, within the periplasm of the host cell.

The term "operably linked" as used herein refers to the association of nucleotide sequences on a single nucleic acid molecule, *i.e.* the vector, in a way such that the function of one or more nucleotide sequences is affected by at least one other nucleotide sequence present on said nucleic acid molecule. For example, a promoter is operably linked with a coding sequence encoding the protein of interest or a fusion protein comprising the POI, when it is capable of effecting the expression of that coding sequence. Specifically, such nucleic acids operably linked to each other may be immediately linked, *i.e.* without further elements or nucleic acid sequences in between or may be indirectly linked with spacer sequences or other sequences in between. Specifically, in the context of a lac operator being operably linked to a

promoter refers to the ability of the lac operator to regulate the ability of the promoter to control expression of the coding sequence under specific conditions. Such as the ability of the lac operator to inhibit promoter-dependent expression of the gene of interest when lac repressor protein is bound thereto.

5 The term “vector” as used herein includes autonomously replicating nucleotide sequences as well as genome integrating nucleotide sequences. A common type of vector is a “plasmid”, which generally is a molecule of double-stranded DNA that can readily accept additional (foreign) DNA and which can readily be introduced into a suitable host cell. A plasmid vector often contains coding DNA and promoter DNA
10 and has one or more restriction sites suitable for inserting foreign DNA. Specifically, the term “vector” or “plasmid” refers to a vehicle by which a DNA or RNA sequence (e.g., a foreign gene) can be introduced into a host cell, so as to transform the host and promote expression (e.g., transcription and translation) of the introduced sequence. The plasmid or expression vector can also be the 2μ plasmid for use in
15 yeasts from which the POI or fusion protein is expressed.

 “Expression vectors” or “vectors” as used herein are defined as DNA sequences that are required for the transcription of cloned recombinant nucleotide sequences, i.e. of recombinant genes and the translation of their mRNA in a suitable host or
20 organism. To obtain expression, a sequence encoding a desired expression product, such as e.g. the fusion protein described herein, is typically cloned into an expression vector that contains a promoter to direct transcription. Suitable bacterial and eukaryotic promoters are well known in the art. The promoter used to direct expression of a nucleic acid depends on the particular application. For example, an inducible or regulatable promoter is typically used for expression and purification of
25 proteins such as e.g. fusion proteins. Expression vectors comprise the expression cassette and additionally usually comprise an origin for autonomous replication in the host cells or a genome or chromosome integration site, one or more selectable markers (e.g., an amino acid synthesis gene or a gene conferring resistance to antibiotics such as zeocin, kanamycin, G418 or hygromycin), a number of restriction
30 enzyme cleavage sites, a suitable promoter sequence, a cloning site or multiple cloning site and a transcription terminator, which components are operably linked together.

Examples for inducible promoters for use in bacteria, include but are not limited to the T7, A1 (also termed "T7A1" or "T7_{A1}") and T5 promoter systems, being inducible in connection with e. g. one or more lacO repressor binding sites, the Pm promoter/operator or the pBAD promoter/operator. Examples for inducible promoters
5 for use in yeast, include for example the AOX 1 promoter or derivatives thereof, such as those published in WO2006/089329A1.

The promoter can also be a constitutive promoter, such as but not limited to the GAP or ADH1 promoter for yeast, or a promoter which is repressible by a metabolite or carbon source and is activated upon depletion of the metabolite or carbon source.
10 Examples include the FMD promoter or the modified AOX 1 promoters, which are modified in that they are repressible (see e.g. WO2006/089329A1). The promoter can be any promoter with and without additional regulatory sequences (such as e.g. repressor binding sites) facilitating translation of the GOI.

An "expression cassette" refers to a DNA coding sequence or segment of DNA
15 coding for an expression product that can be inserted into a vector at defined restriction sites. The cassette restriction sites are designed to ensure insertion of the cassette in the proper reading frame. Generally, DNA is inserted at one or more restriction sites of the vector DNA, and then is carried by the vector into a host cell along with the transmissible vector DNA. A segment or sequence of DNA having
20 inserted or added DNA, such as an expression vector, can also be called a "DNA construct".

Preferably, an expression cassette is a DNA construct comprising essentially a promoter, a gene of interest, and upstream of the gene of interest a Shine-Dalgarno (SO) sequence, also termed ribosome binding site (RBS).

25 The term "expression cassette" may also refer to a linear or circular DNA construct to be integrated into the host genome, such as the bacterial chromosome. As a result of integration, the expression host cell has an integrated expression cassette. If the expression cassette is to be inserted into the genome of a host cell, the expression cassette preferably also comprises two terminally flanking regions which
30 are homologous to a genomic region and which enable homologous recombination. In addition, the cassette may contain other sequences such as for example sequences coding for antibiotic selection markers, prototrophic selection markers or fluorescent markers, markers coding for a metabolic gene, genes which improve

protein expression or two flippase recognition target sites (FRT) which enable the removal of certain sequences (e.g. antibiotic resistance genes) after integration. The use of a linear expression cassette provides the advantage that the genomic integration site can be freely chosen by the respective design of the flanking homologous regions of the cassette. Thereby, integration of the linear expression cassette allows for greater variability with regard to the genomic region. Expression products, such as the fusion protein described herein, can be expressed from an autonomously replicating nucleotide sequence, or from nucleotide sequences stably integrated into the genome or chromosome of a host cell.

Any of the known procedures for introducing foreign nucleotide sequences into host cells may be used. These include the use of calcium phosphate transfection, polybrene, protoplast fusion, electroporation, nucleofection, liposomes, microinjection, naked DNA, plasmid vectors, viral vectors, both episomal and integrative, and any of the other well-known methods for introducing cloned genomic DNA, cDNA, synthetic DNA or other foreign genetic material into a host cell (see, e.g., Sambrook et al.).

The term "sequence identity" as used herein is understood as the relatedness between two amino acid sequences or between two nucleotide sequences and described by the degree of sequence identity or sequence complementarity. The sequence identity of a variant, homologue or orthologue as compared to a parent nucleotide or amino acid sequence indicates the degree of identity of two or more sequences. Two or more amino acid sequences may have the same or conserved amino acid residues at a corresponding position, to a certain degree, up to 100%. Two or more nucleotide sequences may have the same or conserved base pairs at a corresponding position, to a certain degree, up to 100%.

Sequence similarity searching is an effective and reliable strategy for identifying homologs with excess (e.g., at least 50%) sequence identity. Sequence similarity search tools frequently used are e.g., BLAST, FASTA, and HMMER.

Sequence similarity searches can identify such homologous proteins or polynucleotides by detecting excess similarity, and statistically significant similarity that reflects common ancestry. Homologues may encompass orthologues, which are herein understood as the same protein in different organisms, e.g., variants of such protein in different organisms or species.

To determine the % complementarity of two complementary sequences, one of the two sequences needs to be converted to its complementary sequence before the % complementarity can then be calculated as the % identity between the first sequence and the second converted sequences using the above-mentioned algorithm.

"Percent (%) identity" with respect to an amino acid sequence, homologs and orthologues described herein is defined as the percentage of amino acid residues in a candidate sequence that are identical with the amino acid residues in the specific polypeptide sequence, after aligning the sequence and introducing gaps, if necessary, to achieve the maximum percent sequence identity, and not considering any conservative substitutions as part of the sequence identity. Those skilled in the art can determine appropriate parameters for measuring alignment, including any algorithms needed to achieve maximal alignment over the full length of the sequences being compared.

For purposes described herein, the sequence identity between two amino acid sequences can be determined using NCBI BLAST, specifically NCBI BLAST + 2.9.0 program version (Apr-02-2019).

"Percent (%) identity" with respect to a nucleotide sequence *e.g.*, of a nucleic acid molecule or a part thereof, in particular a coding DNA sequence, is defined as the percentage of nucleotides in a candidate DNA sequence that is identical with the nucleotides in the DNA sequence, after aligning the sequence and introducing gaps, if necessary, to achieve the maximum percent sequence identity, and not considering any conservative substitutions as part of the sequence identity. Alignment for purposes of determining percent nucleotide sequence identity can be achieved in various ways that are within the skill in the art, for instance, using publicly available computer software. Those skilled in the art can determine appropriate parameters for measuring alignment, including any algorithms needed to achieve maximal alignment over the full length of the sequences being compared.

Optimal alignment may be determined with the use of any suitable algorithm for aligning sequences, non-limiting examples of which include the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, algorithms based on the Burrows-Wheeler Transform (*e.g.*, the Burrows Wheeler Aligner), ClustalW, Clustal X, BLAT, Novoalign (Novocraft Technologies; available at novocraft.com), ELAND (Illumina,

San Diego, CA), SOAP (available at soap.genomics.org.cn), and Maq (available at maq.sourceforge.net).

As used herein the term "fusion protein" refers to a POI comprising, preferably at its N-terminus, an engineered fusion sequence comprising the peptide tag(s) described herein. Such engineered fusion sequence may comprise a protease recognition and/or cleavage site, *e.g.* a caspase recognition and/or cleavage site or a TEV-protease recognition and/or cleavage site, to facilitate separation of the POI and the fusion sequence. The fusion protein may be cleaved at the protease cleavage site by addition of a protease such as caspases or TEV-protease, *e.g.* to the cell culture, to remove the peptide tag from the POI. In a specific aspect, the engineered fusion sequence described herein comprises at least one caspase recognition site, one or more peptide tags as described herein and optionally one or more linkers. According to a specific example, the fusion protein comprises one or more peptide tags, optionally linked via linker sequences, one or more caspase recognition sites and one or more POIs.

According to a specific embodiment, the fusion protein provided herein comprises a first part, comprising one or more peptide tags, optionally linked via linker sequences, a second part, comprising a recognition site for target-specific proteolytic cleavage, *e.g.* by caspases, cp-caspase, caspase-2 or cp caspase-2, or variants thereof having one or more modifications of the amino acid sequence and a third part, comprising a POI. Specifically, the second part may be part of the peptide tag. Specifically, the fusion protein described herein may comprise each part more than once and in different order. For example, the fusion protein provided herein may comprise a first part comprising a tag, a second part comprising a caspase recognition site, another first part comprising the same or a different tag sequence, another second part comprising the same or a different recognition site and a third part comprising a POI. According to a further example, the fusion protein described herein may comprise more than one POI separated by one or more fusion sequences comprising one or more recognition sites. The fusion protein described herein is encoded by a heterologous gene which is engineered in such a way that it is translated into protein by a host organism. Specifically, the heterologous gene comprises the tag sequence of the invention.

As described herein, the tag sequence of the invention encodes at least part of a peptide tag, that can be used to facilitate expression of a POI. Said peptide tag comprises at least SEQ ID NO:1. The peptide tag may comprise additional elements N-terminal or C-terminal of SEQ ID NO:1.

5 In a specific example, the peptide tag comprises at least 3 additional amino acids N-terminal of SEQ ID NO:1. Specifically, the 3 additional amino acids are L, E and D, resulting in the amino acid sequence of SEQ ID NO:4.

In a specific example, the peptide tag comprises further amino acids C-terminal of SEQ ID NO:1 or SEQ ID NO:4. Specifically, the peptide tag comprises at least
10 about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, or 40 additional amino acids C-terminal of SEQ ID NO:1 or SEQ ID NO:4. In a specific example, the peptide tag comprises 11 additional amino acids, resulting for example in the amino acid sequence of SEQ ID NO:7 or SEQ ID NO:12.

15 In addition, the peptide tag described herein may comprise one or more affinity tags, one or more solubility enhancement tags, one or more monitoring tags, one or more linkers, and/or one or more protease recognition and/or cleavage sites.

Specifically, the protease recognition and/or cleavage site refers to those recognized and/or cleaved by TEV protease or a caspase, preferably caspase-2.

20 Affinity tags are amino acid sequences that can be used for example for the purification of proteins where they are attached to (fusion proteins with affinity tag e.g. at its N-terminus). These affinity tags have high affinity to appropriate ligands of a solid support, like chromatography resins or directly to the resins. By selectively binding of the fusion protein having the affinity tag to the particular resin, the fusion
25 protein can be purified highly effective by only one chromatography step. According to a specific embodiment, affinity tag sequences used herein are selected from histidine (His) tag, specifically a poly-histidine tag, arginine-tag, specifically a poly-arginine tag, peptide substrate for antibodies, chitin binding domain, RNase S peptide, protein A, β -galactosidase, FLAG tag, Strep II tag, streptavidin binding peptide
30 (SBP) tag, calmodulin-binding peptide (CBP), glutathione S-transferase (GST), maltose-binding protein (MBP), S-tag, HA tag, or c-Myc tag or any other tag known to be useful for the efficient purification of a protein it is fused to. Preferably, the affinity tag is a His tag comprising one or more H, specifically a hexahistidine tag.

Specifically, fusion proteins comprising a poly-, or hexa-histidine tag (His-tag) can be captured and purified by IMAC, preferably using a Ni-NTA chromatography material.

Solubility enhancement tags used herein are selected from calmodulin-binding peptide (CBP), poly Arg, poly Lys, G B1 domain, protein D, Z domain of Staphylococcal protein A, and thioredoxin or any other tag known to improve the solubility of the protein it is fused to e.g. during expression in a host cell. Preferably the solubility tag is based on highly charged peptides of bacteriophage genes, for example such as those listed in US 8,535, 908 B2. Specifically, the solubility enhancement tag is selected from the group consisting of T7C, T7B, T7B1, T7B2, T7B3, T7B4, T7B5, T7B6, T7B7, T7B8, T7B9, T7B10, T7B11, T7B12, T7B13, T7A, T7A1, T7A2, T7A3, T7A4, T7A5, T7AC T3, N1, N2, N3, N4, N5, N6, N7, calmodulin-binding peptide (CBP), poly Arg, poly Lys, G B1 domain, protein D, Z domain of Staphylococcal protein A, DsbA, DsbC and thioredoxin.

According to a further specific embodiment, the monitoring tag sequence used herein is m-Cherry, GFP or f-Actin or any other tag useful for detection or quantification of the POI and/or the fusion protein during production of the fusion protein including fermentation, isolation and purification by simple in-situ, inline online or atline detectors, like UV, IR, Raman, Fluorescence and the like.

According to a specific embodiment, the peptide tag described herein comprises more than one additional tags, specifically it comprises an affinity tag and a solubility enhancement tag. Specifically, it comprises an affinity tag, a solubility enhancement tag and a monitoring tag. Specifically, it comprises more than one tag of the same functionality, specifically it comprises more than one affinity tag, more than one solubility enhancement tag and/or more than one monitoring tag, and any combination thereof. Specifically, the peptide tag described herein comprises a C-terminal and an N-terminal tag, each comprising one or more tag sequences, preferably selected from affinity tag, solubility enhancement tag and monitoring tag.

The term "linker" as used herein refers to any amino acid sequence that does not interfere with the function of elements being linked. Linkers may connect e.g., nucleotide sequences, or amino acid sequences. Linkers can be used between the tag and the POI or between tag sequences. The linkers may be used to engineer appropriate amounts of flexibility. Preferably, the linkers are short, e.g., 1–20

nucleotides or amino acids or even more and are typically flexible. Amino acid linkers commonly used consist of a number of glycine, serine, and optionally alanine, in any order. Such linkers usually have a length of at least any one of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or 20 amino acids, as required. Preferably, the linker comprises 1 to 12 amino acid residues, preferably it is a short linker. Preferably the linker is a GS, GSG, GGSGG (SEQ ID NO:21), GSAGSAAGSG (SEQ ID NO:22), (GS)_n, GSGSGSG (SEQ ID NO:23), GSG or GGGGS (SEQ ID NO:24) linker or any combination thereof. In some embodiments, the linker comprises one or more units, repeats or copies of a motif, such as for example GS, GSG or G₄S.

Further provided herein is a method of producing a POI using the tag sequence described herein. Specifically, the fusion protein described herein, comprising at least one POI and the peptide tag encoded by a nucleotide sequence comprising the tag sequence of the invention, is cloned into an expression vector under operable linkage to a promoter. Said expression vector is integrated into a host cell and the host cell is cultured under conditions allowing expression of the fusion protein, optionally following a growth phase for the accumulation of biomass before the recombinant protein is expressed. The POI may be produced employing a fed-batch process as described herein, comprising an expression phase as described herein and optionally a growth phase as described herein.

According to a specific embodiment of the method of producing a POI as described herein, the fusion protein, comprising a protease recognition/ cleavage site, is contacted with a protease after expression, to cleave off the peptide tag herein and produce a POI comprising the desired N-terminus, *i.e.* the natural or designed N-terminus without any unwanted amino acids attached. Specifically, the fusion protein is contacted with the protease enzyme after isolation of the fusion protein from the host cell culture. Specifically, the protease recognition site is recognized by a caspase, preferably caspase-2 and the fusion protein is contacted with the corresponding caspase, preferably caspase-2 or a variant thereof such as circularly permuted (cp) caspase-2 or mutated variants thereof having a higher P1'prime tolerance.

After production of the POI according to the method described herein, the POI may be further modified, purified and/or formulated.

In this context, the term “modifying the at least one protein of interest” is meant that the POI may be chemically, physically or enzymatically modified. There are many methods known in the art to modify proteins. Proteins can be coupled to carbohydrates or lipids. The POI may be PEGylated (the POI chemically coupled to polyethyleneglycole) or HESylated (the POI is chemically coupled to hydroxyethyl starch) for half-life extension. The POI may also be coupled with other moieties such as affinity domains for e.g. human serum albumin for half-life extension. The POI also may be treated by a protease, e.g. such as described above, or under hydrolytic conditions for cleavage to form the active ingredient from a pre-sequence or to cleave off a tag such as an affinity tag for purification. The POI may also be coupled to other moieties such as toxins, radioactive moieties or any other moiety. The POI may further be treated under conditions to form dimers, trimers and the like.

Additionally, the term “formulating the at least one protein of interest” refers to bringing the POI to conditions, where the POI can be stored for a longer time and/or for optimized pharmaceutical form and/or pharmaceutical application and/or to better adjust the concentration and/or to provide higher concentrations in liquid formulations. Many different methods known in the art are available to stabilize proteins. By exchanging the buffer in which the POI is existent after purification and / or modification, the POI can be brought under conditions, where it is more stable. Different buffer substances and additives, such as sucrose, mild detergents, stabilizer and the like, known in the art can be used. The POI can also be stabilized by lyophilization. For some POIs formulations can be done by formation of complexes of the POI with lipids or lipoproteins, such as polyplexes, and the like. Some protein may be co-formulated with other proteins.

The methods described herein specifically refer to the production of heterologous compounds. Such term used with respect to a nucleotide or amino acid sequence or protein, refers to a compound which is either foreign, *i.e.* “exogenous”, such as not found in nature, to a given host cell; or that is naturally found in a given host cell, e.g., is “endogenous”, however, in the context of a heterologous construct, e.g., employing a heterologous nucleic acid, thus “not naturally-occurring”. The heterologous nucleotide sequence as found endogenously may also be produced in an unnatural, e.g., greater than expected or greater than naturally found, amount in the cell. The heterologous nucleotide sequence, or a nucleic acid comprising the

heterologous nucleotide sequence, possibly differs in sequence from the endogenous nucleotide sequence but encodes the same protein as found endogenously. Specifically, heterologous nucleotide sequences are those not found in the same relationship to a host cell in nature (*i.e.*, “not natively associated”). Any recombinant or artificial nucleotide sequence is understood to be heterologous.

Recombinant host cells according to the present invention can be obtained by introducing a vector or plasmid (such as an expression vector as mentioned above) comprising the target polynucleotide sequences, including the tag sequences described herein, into the cells. Techniques for transfecting or transforming eukaryotic cells or transforming prokaryotic cells are well known in the art. These can include lipid vesicle mediated uptake, heat shock mediated uptake, calcium phosphate mediated transfection (calcium phosphate/DNA co-precipitation), viral infection, particularly using modified viruses such as, for example, modified adenoviruses, microinjection and electroporation. For prokaryotic transformation, techniques can include heat shock mediated uptake, bacterial protoplast fusion with intact cells, microinjection and electroporation. The DNA can be single or double stranded, linear or circular, relaxed or supercoiled DNA. For various techniques for transfecting mammalian cells, see, for example, Keown et al. (1990) *Processes in Enzymology* 185:527-537. Procedures used to manipulate polynucleotide sequences, *e.g.* coding for the proteins of the present invention and/or the POI, the promoters, enhancers, leaders, etc., are well known to persons skilled in the art, *e.g.* described by J. Sambrook et al., *Molecular Cloning: A Laboratory Manual* (3rd edition), Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, New York (2001).

Recombinant host cells according to the present invention may also comprise tag sequences described herein stably integrated into their genome. A foreign or target polynucleotide such as the nucleic acid sequence encoding a peptide tag as described herein can be inserted into the chromosome by various means, *e.g.*, by homologous recombination or by using a hybrid recombinase that specifically targets sequences at the integration sites. The foreign or target polynucleotide comprising a tag sequence as described herein and/or a GOI is typically present in a vector (“inserting/integration vector”), preferably within an expression cassette. These vectors are typically circular and linearized before they are used for homologous recombination. As an alternative, the foreign or target polynucleotides may be

DNA fragments joined by fusion PCR or synthetically constructed DNA fragments which are then recombined into the host cell. In addition to the homology arms, the vectors may also contain markers suitable for selection or screening, an origin of replication, and other elements. It is also possible to use heterologous recombination which results in random or non-targeted integration. Heterologous recombination refers to recombination between DNA molecules with significantly different sequences. Methods of recombinations are known in the art and for example described in Boer et al., Appl Microbiol Biotechnol (2007) 77:513–523. One may also refer to Principles of Gene Manipulation and Genomics by Primrose and Twyman (7th edition, Blackwell Publishing 2006) for genetic manipulation of yeast cells.

The phrase “culturing said host cell under conditions permitting” expression of a target protein refers to maintaining and/or growing host cells under conditions (e.g., but not limited to temperature, aeration, pressure, pH, induction, growth rate, culture medium, nutrients duration of the cultivation, mode of nutrient feed(s) etc.) appropriate or sufficient to obtain production of the desired compound (fusion protein, POI).

A host cell may preferably first be cultivated at conditions to grow efficiently to a large cell number without the burden of expressing a protein. When the cells are prepared for POI expression, suitable cultivation conditions are selected and optimized to produce the POI.

An inducible promoter may be used that becomes activated as soon as an inductive stimulus is applied to direct transcription of the gene under its control. An inductive stimulus is preferably the addition of an appropriate agents (e.g. methanol for the AOX-promoter or IPTG or lactose for promoters under the control of lac operators (binding sites for the lac repressor) such as the lac, T7, or A1 promoters) or the depletion of an appropriate nutrient (e.g., methionine for the MET3-promoter or phosphate for the phoA promoter). Also, the addition of ethanol, methylamine, cadmium or copper as well as heat or an osmotic pressure increasing agent can induce the expression depending on the promoters operably linked to the proteins of the invention and the POI(s). Some inducible or de-repressible promoters can be induced by depletion of a certain medium component or metabolite, such as phosphate for the phoA promoter. An inductive stimulus can also be a pH, temperature or osmotic pressure change. Some promoters are activated by a combination of an

inductive stimulus and de-repression by depletion or limitation of certain metabolites.

It is preferred to cultivate the host cell(s) according to the invention in a bioreactor under optimized growth conditions to obtain a cell density of at least 1 g/L, preferably at least 10 g/L cell dry weight, more preferably at least 50 g/L cell dry weight. It is advantageous to achieve such yields of biomolecule production not only on a laboratory scale, but also on a pilot or industrial scale.

Preferably, the host cells are cultivated in a minimal medium with a suitable carbon source, thereby further simplifying the isolation process significantly. By way of example, the minimal medium contains a utilizable carbon source (e.g. glucose, glycerol, ethanol or methanol), salts containing the macro elements (potassium, magnesium, calcium, ammonium, chloride, sulphate, phosphate) and trace elements (copper, iodide, manganese, molybdate, cobalt, zinc, and iron salts, and boric acid).

In the case of yeast cells, the cells may be transformed with one or more of the above-described expression vector(s) and cultured in conventional nutrient media, optionally modified as appropriate for inducing promoters, selecting transformants or amplifying the genes encoding the desired sequences. A number of minimal media suitable for the growth of yeast are known in the art. Any of these media may be supplemented as necessary with salts (such as sodium chloride, calcium, magnesium, and phosphate), buffers (such as HEPES, citric acid and phosphate buffer), nucleosides (such as adenosine and thymidine), antibiotics, trace elements, vitamins, and glucose or an equivalent energy source. Any other necessary supplements may also be included at appropriate concentrations that would be known to those skilled in the art. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression and are known to the ordinarily skilled artisan. Cell culture conditions for other type of host cells are also known and can be readily determined by the artisan. Descriptions of culture media for various microorganisms are for example contained in the handbook "Manual of Methods for General Bacteriology" of the American Society for Bacteriology (Washington D.C, USA, 1981).

Host cells can be cultured (e.g., maintained and/or grown) in liquid media and preferably are cultured, either continuously or intermittently, by conventional

culturing methods such as standing culture, test tube culture, shaking culture (e.g., rotary shaking culture, shake flask culture, etc.), aeration spinner culture, or fermentation. In some embodiments, cells are cultured in shake flasks or deep well plates. In yet other embodiments, cells are cultured in a bioreactor (e.g., in a bioreactor cultivation process). Cultivation processes include, but are not limited to, batch, fed-batch and continuous methods of cultivation. The terms "batch process" and "batch cultivation" refer to a closed system in which the composition of media, nutrients, supplemental additives and the like is set at the beginning of the cultivation and not subject to alteration during the cultivation; however, attempts may be made to control such factors as pH and oxygen concentration to prevent excess media acidification and/or cell death. The terms "fed-batch process" and "fed-batch cultivation" refer to a batch cultivation with the exception that one or more substrates and/or supplements and/or inducers are added (e.g., added in increments or continuously) as the cultivation progresses in one or more feed solutions/suspensions. The terms "continuous process" and "continuous cultivation" refer to a system in which a defined cultivation media is added continuously to a bioreactor and an equal amount of used or "conditioned" media is simultaneously removed, for example, for recovery of the desired product. A variety of such processes has been developed and is well-known in the art.

In some embodiments, host cells are cultured for about 12 to 24 hours, in other embodiments, host cells are cultured for about 24 to 36 hours, about 36 to 48 hours, about 48 to 72 hours, about 72 to 96 hours, about 96 to 120 hours, about 120 to 144 hours, or for a duration greater than 144 hours. In yet other embodiments, culturing is continued for a time sufficient to reach desirable production yields of POI.

The above-mentioned methods may further comprise a step of isolating the expressed at least one target protein, e.g. the POI or fusion protein described herein, from the cell culture and optionally followed by a step of purifying the at least one POI or fusion protein. If the target protein is secreted from the cells, it can be isolated and then purified from the culture medium using state of the art techniques. Secretion of the target protein from the cells is generally preferred, since the products are recovered from the culture supernatant rather than from the complex mixture of proteins that results when cells are disrupted to release intracellular proteins. A protease inhibitor, such as phenyl methyl sulfonyl fluoride (PMSF) or any other means

may be useful to inhibit proteolytic degradation during purification, and antibiotics may be included to prevent the growth of adventitious contaminants. The composition may be concentrated, filtered, dialyzed, etc., using methods known in the art. The cell culture after fermentation / cultivation can be centrifuged using a separator
5 or a tube centrifuge to separate the cells from the culture supernatant or the cells can be separated from the supernatant by filtration, e.g. depth or tangential flow filtration. The supernatant can then be filtered and/or concentrated by using a tangential flow filtration. Alternatively, cultured host cells may also be ruptured sonically or mechanically (e.g. high pressure homogenisation), enzymatically or chemically
10 or by heat to obtain a cell extract containing the desired POI, from which the POI may be isolated and purified.

Isolation and purification methods, also referred to as methods of recovery, for obtaining the POI may be based on methods utilizing difference in solubility, such as salting out, solvent precipitation, heat precipitation, methods utilizing difference
15 in molecular weight, such as size exclusion chromatography, ultrafiltration and gel electrophoresis, methods utilizing difference in electric charge, such as ion-exchange chromatography, methods utilizing specific affinity, such as affinity chromatography, methods utilizing difference in hydrophobicity, such as hydrophobic interaction chromatography and reverse phase high performance liquid chromatog-
20 raphy, methods utilizing difference in isoelectric point, such as isoelectric focusing may be used and methods utilizing certain amino acids, such as IMAC (immobilized metal ion affinity chromatography. If the POI is expressed as inactive and soluble Inclusion Bodies the solubilized Inclusion Bodies need to be refolded and may be purified. When the POI is fused to a His-tag such as e.g. the 6 His-tag, which can
25 be included in the peptide tag described herein, IMAC may be used for capturing and purification of the fusion protein. After cleavage of the fusion protein to release the POI, the POI can be separated by IMAC in the flow through mode from the cleaved tag, un-cleaved fusion protein and the protease, if the protease also comprised a His-tag.

30 The isolated and purified POI can be identified by conventional methods such as Western Blotting or specific assays for POI activity. The structure of the purified POI can be determined by amino acid analysis, amino-terminal peptide sequencing, primary structure analysis for example by mass spectrometry, RP-HPLC, ion

exchange-HPLC, ELISA and the like. It is preferred that the POI is obtainable in large amounts and in a high purity level, thus meeting the necessary requirements for being used as an active ingredient in pharmaceutical compositions or as feed or food additive.

5 The following examples are put forth to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the subject invention and are not intended to limit the scope of what is regarded as the invention and defined in the claims. Efforts have been made to ensure accuracy with respect to the numbers used (e.g. amounts, temperature, concentrations, etc.) but some
10 experimental errors and deviations should be allowed for.

EXAMPLES

The examples below will demonstrate that the newly identified tag sequence(s) increase(s) the titer (product per volume in mg/L) and the yield (product per biomass
15 in mg/g biomass measured as dry cell weight), respectively, of recombinant proteins upon their expression. As an example, the titre of recombinant proteins (TNF α , PTH and hFGF2) in the bacteria *E. coli* are increased.

Example 1: General Material and Methods

20 1.1. *Escherichia coli* strains

Strains used for recombinant protein expression as well as for cloning purposes were acquired from New England Biolabs (NEB, Ipswich, MA, USA). Chemically competent *E. coli* BL21(DE3) cells were used for all protein expression experiments, while chemically competent NEB-5 α cells were applied for cloning purposes. Trans-
25 formations and subsequent cultivations for cloning purposes were performed according to the manufacturer's instructions.

1.2. Generation of expression constructs

Expression vectors were created using backbones from existing pET30acer
30 plasmids from a previous study (Köppl et al. 2022). Q5® High-Fidelity DNA Polymerase, BsaI-HF®v2, DpnI and T4 DNA Ligase were purchased from NEB. Construction of the vector plasmids followed a standard cloning protocol. In brief, site-directed mutagenesis was carried out utilizing primers carrying the desired

mutations. A polymerase chain reaction (PCR) was performed employing Q5® High-Fidelity DNA Polymerase to amplify the original plasmid backbone and simultaneously insert the wanted mutations. After amplification, the linear DNA fragment was purified using the Monarch® PCR & DNA Cleanup Kit (NEB) followed by restriction digest with BsaI-HF®v2 and DpnI (37°C, overnight). After preparative agarose gel (2% agarose) purification (90 V, 90 min), the correctly sized band was excised and dissolved using the Monarch® DNA Gel Extraction Kit (NEB). Ligation was performed at 16 °C overnight using T4 DNA Ligase (NEB) and the ligation mix was directly transformed into chemically competent cells according to the manufacturer's instructions. A comprehensive list of all primers and DNA fragments used can be found in Table 1. All oligonucleotides labelled with the letter P (being the primers for site directed mutagenesis - P3 and P4 - and subsequent screening and sequencing – P1 and P2) were bought from Sigma Aldrich (Taufkirchen, Germany). Only 4 primers were used for the cloning procedures. Two primers for inserting the silent mutations were designed in a way that they only bind to the fusion tag and can therefore be applied universally to every POI. The remaining two primers bind outside of the POI entirely and were used for screening and sequencing. After cloning, all constructs were sequenced to confirm their correctness.

Table 1. Primers used in the Examples. Lower case letters denote BsaI cleavage sites with accompanying inserted bases for specific overhang creation.

Primer/ number	Name	Sequence (5'-3')
P1	Seq-MCS-for	GAAGCAGCCCAGTAGTAG (SEQ ID NO:45)
P2	Seq-MCS-rev	CGGATATAGTTCCTCCTTTCAG (SEQ ID NO:46)
P3	T7AC-NEW-for	ggtctcaGAAAGGAATAAAGAAAGGAAAGAAGCTGAGTT GCAAG (SEQ ID NO:47)
P4	T7AC-NEW-rev	ggtctcttttcGGGATCTTCTAGCATATGTATATCTCCTTC (SEQ ID NO:48)

1.3. Laboratory scale fed-batch fermentations

All fermentation procedures, as well as media preparation and composition were performed according to Fink et al. (2021a, 2021b).

In each experimental series, the same fermentation process was repeated in an identical fashion for each construct. Briefly, cells were cultivated in the DASGIP

parallel bioreactor system (Eppendorf SE, Hamburg, Germany) using a vessel with 2.1 L volume which allows for a maximum working volume of 1.8 L. To control relevant process parameters, the bioreactors were equipped with pH probes (Hamilton, Bonaduz, GR, Switzerland), optical dissolved O₂ probes (Hamilton, Bonaduz, GR, Switzerland) and temperature probes. The dissolved oxygen concentration in the media was kept constantly at >30% and the pH was sustained at 7.0 ± 0.2 by the addition of 12.5% ammonia solution. Temperature of the cultivation media was maintained at 37 ± 0.2 °C during the batch phase and shifted to 30 ± 0.2 °C at the beginning of the feed. Induction was facilitated by the introduction of Isopropyl- β -D-thiogalactopyranoside (IPTG). All fermentation parameters not mentioned here are listed in Table 2 and 3.

Table 2. Fermentation strategy I for initial T7AC and T7AC_{new} experiments using PTH, hFGF2 and TNF α as model proteins (production during the last 1.2 generations).

	Temp. [°C]	CDM [g/L]	Vol. [mL]	Duration [h]	μ [h⁻¹]	Generations
Batch	37	6.7	500	7.5	-	-
Fed- Batch	30	40.4	1014	21.0	0.1	3.6

Induction of fermentations according to Table 2 was carried out at feed hour 17 with an IPTG pulse added directly into the bioreactor corresponding to 2 μ mol IPTG per gram of biomass at the end of fermentation.

Table 3. Fermentation strategy II for additional T7AC and T7AC_{new} experiments using TNF α as model protein (production during the last 1.7 generations).

	Temp. [°C]	CDM [g/L]	Vol. [mL]	Duration [h]	μ [h⁻¹]	Generations
Batch	37	6.7	500	7.5	-	-
Fed-Batch 1	30	16.9	598	11.0	0.1	1.6
Fed-Batch 2	30	38.2	1018	27.0	0.05	2.0

Induction of fermentations according to Table 3 was carried out at feed hour 15 with an IPTG pulse added directly into the bioreactor corresponding to 15 μ mol IPTG per gram of biomass at the end of fermentation.

1.4. Cell Lysis

Cell lysis and separation of the soluble und insoluble intracellular protein fractions was carried out according to Fink et al. (2019). Additionally, the cell lysis buffer
5 contained 4 mmol L⁻¹ of NuPAGE Sample Reducing Agent (10x) (Invitrogen, Waltham, MA, USA).

1.5. Recombinant protein quantification

To estimate the recombinant protein concentration present in the cell lysate,
10 reducing SDS-polyacrylamide gel electrophoresis (PAGE) was performed as described by Stargardt et al. (2020) and Cserjan-Puschmann et al. (2020). Bovine Serum Albumin heat shock fraction (Sigma Aldrich, St. Louis, MO, USA) in the concentrations of 25, 50 and 75 µg/mL was used as quantification standard. A brief description of the exact analytical procedure has been published by Köppl et al.
15 (2022).

Example 2: Novel tag sequence enhances protein expression

2.1 Generation of Fusion constructs

Recently, the development of a universally applicable, generic fusion tag, the
20 CASPON™ tag, has been reported. This tag does not only facilitate high soluble titres but also enhances recombinant protein production in *E. coli* (Köppl et al. 2022). The CASPON™ tag comprises SEQ ID NO:40, including a T7AC tag at its N-terminus.

To investigate the influence of silent mutations, including the introduction of multiple rare codons, at the 5' end of a protein tag's mRNA on recombinant protein
25 expression and cellular growth kinetics, the nucleotide sequence of the previously published T7AC tag was modified (see Table 4 and Fig. 1B). 8 silent mutations were introduced in the first 12 codons of the nucleic acid encoding the tag, resulting in the nucleotide sequence of SEQ ID NO:9. The codon usage frequency per 1000 codons
30 (in *E. coli* BL21) of the silent mutations encoded by rare codons – codon usage frequency lower than 10 per 1000 codons, i.e. 1% or lower (Daniel et al. 2015) – is as follows: CTA – 3.85, CCC – 5.38 and AGG – 1.02 (<https://dnahive.fda.gov>). The

codon usage frequency of introduced silent mutations that are not rare codons (in *E.coli* BL21) is as follows: GAA – 39.91, AAT – 17.38 and AAA – 33.64. Thus, 4 of the 8 modifications resulted in the introduction of rare codons.

The amino acid sequence of the tag remained unchanged (SEQ ID NO:7).

5

Table 4. Tag sequences, modifications in bold and underlined (T7AC-tag means the T7AC-tag encoded by SEQ ID NO: 8; T7ACnew-tag means the T7AC-tag encoded by SEQ ID NO: 9)

Tag	Amino acid sequence	Nucleotide sequence
T7AC	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO:7)	CTG GAG GAT CCG GAA CGC AAC AAA GAG CGA AAG GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:8)
T7ACnew	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO:7)	<u>CTA</u> <u>GAA</u> GAT <u>CCG</u> GAA <u>AGG</u> <u>AAT</u> AAA GAA <u>AGG</u> <u>AAA</u> GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:9)

The free folding energy at the mRNA nucleotide positions -4 to +37 of the tag sequences was calculated using the method developed by Zuker et al. 1980, resulting in a $\Delta G_{\min}$ of -7.50 kcal/mol for T7AC and -1.50 kcal/mol for T7ACnew (see Fig. 1B).

The different T7AC nucleotide sequences were then introduced into the CAS-PON™ tag, fused to three different pharmaceutically relevant model proteins: parathyroid hormone (PTH), human fibroblast growth factor 2 (hFGF2) and TNF α and tested in a series of carbon limited fed-batch fermentations. For a general structure of the expression constructs see Fig. 1A. For sequences of specific constructs tested in this Example see Table 5 and the respective sequences in Fig. 10.

Table 5. Sequences of fusion constructs

Construct name	Amino acid sequence	Nucleotide Sequence
T7AC-PTH	SEQ ID NO: 37	SEQ ID NO: 38
T7ACnew-PTH	SEQ ID NO: 37	SEQ ID NO: 39
T7AC-hFGF2	SEQ ID NO:32	SEQ ID NO:33
T7ACnew-hFGF2	SEQ ID NO:32	SEQ ID NO:34
T7AC-TNF α	SEQ ID NO:27	SEQ ID NO:28
T7ACnew-TNF α	SEQ ID NO:27	SEQ ID NO:29

2.2 Cellular fitness of fusion constructs

The above constructs were cloned into the plasmid pET30acer to drive the expression of the fusion proteins, and the resulting plasmids were transformed into the host strain BL21(DE3). In identical laboratory-scale carbon-limited fed-batch cultivations (according to Table 2) all constructs were investigated regarding their influence on production strain growth (Figure 2, A to C) and expression of fusion protein (Figure 2 D to F) immediately before and at four time points after induction. The fitness of the production hosts during the fermentations was comparable, as all variants almost reached the theoretical values of the calculated cell dry mass (CDM).

Cellular growth of both tag variants, T7AC and T7ACnew, regardless of the expressed model protein was highly similar in all fermentations. Biomass formation followed the predicted values reasonably well with a trend to reduced cellular growth at the end of fermentations, but higher growth rates for the T7ACnew tagged POIs at the end of fermentation (Figure 2 A to C). The theoretical CDM at fermentation end is 40.4 g/L. Measured CDM concentrations are summarized in Table 6 below.

Table 6. Measured CDM concentrations, soluble recombinant protein titer (volumetric (vol.) titer) and yield (specific (spec.) recombinant protein titer) at the end of fermentations.

	CDM [g/L]	spec. titer [mg/g CDM]	vol. titer [g/L]
T7AC-PTH	38.8	46.2	1.8
T7ACnew-PTH	38.7	97.2	3.6
T7AC-hFGF2	36.0	56.4	2.0
T7ACnew-hFGF2	37.5	72.7	2.7
T7AC-TNF α	36.8	31.2	1.1
T7ACnew-TNF α	39.2	164.8	6.5

2.3 Increased POI production

Recombinant protein formation was higher with the T7ACnew-tag compared to the unchanged tag for every tested model protein. The comparative increase in recombinant soluble protein titre at the end of fermentation is dependent on the model protein used. For PTH and hFGF2 increases in specific titre of 2.1-fold and 29 % were observed respectively, while TNF α showed the highest increase in specific soluble titre with the T7ACnew-tag producing 5.3-fold as much TNF α as its

reference (Fig. 2 D to F and Figure 3). Measured specific titers (yields) and volumetric recombinant protein titres (titers) are summarized in Table 6 above.

5 The results of Examples 2.2 and 2.3 were confirmed using different fermentation conditions (see Table 3). Figure 4 shows that also at different fermentation conditions, TNF α fused to either T7AC or T7ACnew reaches similar cell mass (Figure 4A) but the construct comprising POI fused to T7ACnew instead of T7AC reaches significantly higher protein titers (Figure 4B).

10 Example 2 shows, that although the productivity for the fusion proteins of PTH, hFGF2 and TNF α is very high as well as the specific titers (about 80 up to > 300 mg/g CDM), the whole amount of the fusion proteins is expressed solubly in the cytosol of E. coli without the formation of inclusion bodies. Use of the new tag sequence did not negatively impact cell activity and cell growth, despite the high
15 productivity: at the end of fermentation, when using the T7ACnew tag sequence (SEQ ID NO:6), the growth rate is even higher compared to using the original nucleotide sequence of the T7AC part (SEQ ID NO:5) of the peptide tag. This is surprising since high productivity often leads to high cell stress, e.g. caused by depletion of t-RNAs resp. charged t-RNAs, which further leads to decreased growth or
20 even breakdown of the cells.

It had previously been reported that the introduction of rare codons into highly expressed genes induces the depletion of rare tRNA molecules and retention of ribosomes, therefore inducing changes in translation efficiency of numerous genes
25 and consequently having detrimental effects on cellular fitness (Kudla et al. 2009, Frumkin et al. 2018). In the context of protein biosynthesis, a reduction of protein product formation and/or cellular fitness imposes a severe reduction in production efficiency and profitability. Contrary to these studies, however, cell growth behaviour was not adversely affected and even better when using the T7ACnew tag sequence
30 (see Example 2.2).

Example 3: N-terminal DBv1 and DBv2 T7AC and T7ACnew peptide tag variants

Two N-terminal peptide sequence variants, named DBv1 (SEQ ID NO:50) and DBv2 (SEQ ID NO:51), were incorporated N-terminally of SEQ ID NO: 1 and the performance of the respective tags was assessed.

3.1 Materials and Methods

3.1.1 Escherichia coli strains

Strains used for recombinant protein expression as well as for cloning purposes were acquired from New England Biolabs (NEB, Ipswich, MA, USA). Chemically competent *E. coli* BL21(DE3) cells were used for all protein expression experiments, while chemically competent NEB-5 α cells were applied for cloning purposes. Transformations and subsequent cultivations for cloning purposes were performed according to the manufacturer's instructions.

3.1.2 Generation of expression systems

Expression vectors were created using backbones from existing pET30acer plasmids from a previous study by (Köppl et al. 2022). Q5® High-Fidelity DNA Polymerase, BsaI-HF®v2, DpnI and T4 DNA Ligase were purchased from NEB. Construction of the vector plasmids followed a standard cloning protocol. In brief, site-directed mutagenesis was carried out utilizing primers carrying the desired mutations. A polymerase chain reaction (PCR) was performed employing Q5® High-Fidelity DNA Polymerase to amplify the original plasmid backbone and simultaneously insert the wanted mutations. After amplification, the linear DNA fragment was purified using the Monarch® PCR & DNA Cleanup Kit (NEB) followed by restriction digest with BsaI-HF®v2 and DpnI (37°C, overnight) to create specific overhangs for the following ligation step and cut the template plasmid used for PCR amplification. After preparative agarose gel (2% agarose) purification (90 V, 90 min), the correctly sized band was excised and dissolved using the Monarch® DNA Gel Extraction Kit (NEB). Ligation to generate circular plasmid DNA was performed at 16 °C overnight using T4 DNA Ligase (NEB) and the ligation mix was directly transformed into chemically competent cells according to the manufacturer's instructions. A comprehensive list of all primers and DNA fragments used can be found in Table 7. All

oligonucleotides labelled with the letter P (being the primers for site directed mutagenesis – P3 and P5 to P10 - and subsequent screening and sequencing – P1 and P2) were bought from Sigma Aldrich (Taufkirchen, Germany). Primers for site directed mutagenesis were used as follows: P5 and P6 were used to create the DBv1-T7AC construct, primers P7 and P8 were used to create the DBv2-T7AC constructs, P9 and P3 were used to create the DBv1-T7AC_{new} construct and P10 and P3 were used to create the DBv2-T7AC_{new} constructs. After cloning, all constructs were sequenced to confirm their correctness.

Table 7. Primers used in the Examples. Lower case letters denote BsaI cleavage sites with accompanying inserted bases for specific overhang creation.

Primer number	Name	Sequence (5'-3')
P5	DBv1-rev	ggtctctagggCATATGTATATCTCCTTCTTAAAGTTA (SEQ ID NO:67)
P6	DBv1-for	ggtctcacccttggtcCCGGAACGCAACAAAG (SEQ ID NO:68)
P7	DBv2-rev	ggtctctgtggCATATGTATATCTCCTTCTTAAAGTTA (SEQ ID NO:69)
P8	DBv2-for	ggtctcaccacacagcCCGGAACGCAACAAAG (SEQ ID NO:70)
P9	DBv1-NEW-rev	ggtctcttttcgGGGACCAAGGGCATATG (SEQ ID NO:71)
P10	DBv2-NEW-rev	ggtctcttttcgGGGCTGTGTGGCATAT (SEQ ID NO:72)
P3	T7AC-NEW-for	ggtctcaGAAAGGAATAAAGAAAGGAAAGAAGCTGAGTT-GCAAG (SEQ ID NO:47)
P1	Seq-MCS-for	GAAGCAGCCCAGTAGTAG (SEQ ID NO:45)
P2	Seq-MCS-rev	CGGATATAGTTCCTCCTTTCAG (SEQ ID NO:46)

All GOIs were expressed under the T7 promoter/operator system.
The resulting fusion proteins are shown in Table 10.

3.1.3 Laboratory scale fed-batch fermentations

All fermentation procedures, as well as media preparation and composition were performed according to Fink et al. 2019, 2021.

In each experimental series, the same fermentation process was repeated in an identical fashion for each construct. Briefly, cells were cultivated in the DASGIP parallel bioreactor system (Eppendorf SE, Hamburg, Germany) using a vessel with 2.1 L volume which allows for a maximum working volume of 1.8 L. To control relevant

process parameters, the bioreactors were equipped with pH probes (Hamilton, Bonaduz, GR, Switzerland), optical dissolved O₂ probes (Hamilton, Bonaduz, GR, Switzerland) and temperature probes. The dissolved oxygen concentration in the media was kept constantly at >30% and the pH was sustained at 7.0 ± 0.2 by the addition of 12.5% ammonia solution. Temperature of the cultivation media was maintained at 37 ± 0.2 °C during the batch phase and shifted to 30 ± 0.2 °C at the beginning of the feed. Induction was facilitated by the introduction of Isopropyl- β -D-thiogalactopyranoside (IPTG). All fermentation parameters not mentioned here are listed in Table 8.

10

Table 8. Fermentation strategy for all DB-T7AC and DB-T7AC_{new} experiments using PTH and hFGF2 as model proteins (production during the last 1.2 generations).

	<i>Temp. [°C]</i>	<i>CDM [g/L]</i>	<i>Vol. [mL]</i>	<i>Duration [h]</i>	<i>μ [h⁻¹]</i>	<i>Generations</i>
Batch	37	6.7	500	7.5	-	-
Fed- Batch	30	40.4	1014	21.0	0.1	3.6

Induction of fermentations according to Table 8 was carried out at feed hour 17 with an IPTG pulse added directly into the bioreactor corresponding to 2 μ mol IPTG per gram of biomass at the end of fermentation.

15

3.1.4 Cell Lysis

Cell lysis and separation of the soluble und insoluble intracellular protein fractions was carried out according to Fink et al. (2019). Additionally, the cell lysis buffer contained 4 mmol L⁻¹ of NuPAGE Sample Reducing Agent (10x) (Invitrogen, Waltham, MA, USA).

20

3.1.5 Recombinant protein quantification

To estimate the recombinant protein concentration present in the cell lysate, reducing SDS-polyacrylamide gel electrophoresis (PAGE) was performed as described by Stargardt et al. (2020) and Cserjan-Puschmann et al. (2020). Bovine Serum Albumin heat shock fraction (Sigma Aldrich, St. Louis, MO, USA) in the concentrations of 25, 50 and 75 μ g/mL was used as quantification standard. A brief

25

description of the exact analytical procedure has been published by Köppl et al. (2022).

3.2Generation of Fusion constructs

5 Recently, the development of a universally applicable, generic fusion tag, the CASPON™ tag, has been reported. This tag does not only facilitate high soluble titres but also enhances recombinant protein production in *E. coli* (Köppl et al. 2022). The CASPON™ tag comprises SEQ ID NO:40 (here shown with a starting methio-

10 nine), including a T7AC tag at its N-terminus.

Two N-terminal peptide sequence variants, named DBv1 (SEQ ID NO:50) and DBv2 (SEQ ID NO:51), were incorporated at the N-terminus of the T7AC and the T7ACnew tags at amino acid positions +2 to +4 (see Figure 5).

Furthermore, performance of the novel peptide tags was assessed. The resulting amino acid and nucleotide sequences of the tags are shown in Table 9.

15 Table 9. N-terminal DBv1 and DBv2 T7AC and T7ACnew peptide tag variant sequences (DBv1-T7ACnew and DBv2-T7ACnew comprise SEQ ID NO: 1 encoded by SEQ ID NO: 3. DBv1-T7AC and DBv2-T7AC comprise SEQ ID NO: 1 encoded by SEQ ID NO: 2. Differences are highlighted in bold and underlined).

Tag	Amino acid sequence	Nucleotide sequence
DBv1-T7AC	PLVPERNKERKEAEL-QAQTAEQ (SEQ ID NO:52)	<u>CCC TTG GTC</u> CCG GAA CGC AAC AAA GAG CGA AAG GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:54)
DBv1-T7ACnew	PLVPERNKERKEAEL-QAQTAEQ (SEQ ID NO:52)	<u>CCC TTG GTC</u> <u>CCC</u> GAA <u>AGG</u> <u>AAT</u> AAA GAA <u>AGG</u> <u>AAA</u> GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:55)
DBv2-T7AC	PHSPERNKERKEAEL-QAQTAEQ (SEQ ID NO:53)	<u>CCA CAC AGC</u> CCG GAA CGC AAC AAA GAG CGA AAG GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:56)
DBv2-T7ACnew	PHSPERNKERKEAEL-QAQTAEQ (SEQ ID NO:53)	<u>CCA CAC AGC</u> <u>CCC</u> GAA <u>AGG</u> <u>AAT</u> AAA GAA <u>AGG</u> <u>AAA</u> GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO:57)

20 The free folding energy at the mRNA nucleotide positions -4 to +37 of the tag sequences was calculated using the method developed by Zuker et al. 1980,

resulting in e.g. a $\Delta G_{\min}$ of -4.20 kcal/mol for DBv1-T7AC and -1.50 kcal/mol for DBv2-T7ACnew (see Fig. 5).

The different tag sequence variants were then introduced into the CASPON™ tag, fused to two different pharmaceutically relevant model proteins: parathyroid hormone (PTH) and human fibroblast growth factor 2 (hFGF2) tested in a series of carbon limited fed-batch fermentations.

For a general structure of the expression constructs see Fig. 5. For sequences of specific constructs tested in the Examples see Table 10 and the respective sequences in Fig. 10.

Table 10. Sequences of fusion constructs

Construct name	Amino acid sequence	Nucleotide Sequence
T7AC-PTH	SEQ ID NO:37	SEQ ID NO:38
T7ACnew-PTH	SEQ ID NO:37	SEQ ID NO:39
DBv1-T7AC-PTH	SEQ ID NO:58	SEQ ID NO:59
DBv1-T7ACnew-PTH	SEQ ID NO:58	SEQ ID NO:60
DBv2-T7AC-PTH	SEQ ID NO:61	SEQ ID NO:62
DBv2-T7ACnew-PTH	SEQ ID NO:61	SEQ ID NO:63
T7AC-hFGF2	SEQ ID NO:32	SEQ ID NO:33
T7ACnew-hFGF2	SEQ ID NO:32	SEQ ID NO:34
DBv2-T7AC-hFGF2	SEQ ID NO:64	SEQ ID NO:65
DBv2-T7ACnew-hFGF2	SEQ ID NO:64	SEQ ID NO:66

3.3Cellular fitness of fusion constructs

The above constructs were cloned into the plasmid pET30acer to drive the expression of the fusion proteins, and the resulting plasmids were transformed into the host strain BL21(DE3). In identical laboratory-scale carbon-limited fed-batch cultivations (according to Table 8) all constructs were investigated regarding their influence on production strain growth (Figure 6 C and F) and expression of fusion protein (Figure 6 A, B, D and E) immediately before and at four time points after induction. The fitness of all production hosts during the fermentations was comparable, as all variants almost reached the theoretical values of the calculated cell dry mass (CDM).

Cellular growth of all tag variants, regardless of the expressed model protein was highly similar in all fermentations. Biomass formation followed the predicted

values reasonably well with a trend to reduced cellular growth at the end of fermentations (Figure 6 C and F). The theoretical CDM at fermentation end is 40.4 g/L. Measured CDM concentrations are summarized in Table 11 below.

- 5 Table 11. Measured CDM concentrations, soluble recombinant protein titer (volumetric (vol.) titer) and yield (specific (spec.) recombinant protein titer) at the end of fermentations.

	CDM [g/L]	spec. titer [mg/g CDM]	vol. titer [g/L]
T7AC-PTH	38.8	46.2	1.8
T7ACnew-PTH	38.7	97.2	3.6
DBv1-T7AC-PTH	35.9	58.0	2.1
DBv1-T7ACnew-PTH	36.0	232.0	8.6
DBv2-T7AC-PTH	41.1	145.2	5.9
DBv2-T7ACnew-PTH	39.6	195.3	7.7
T7AC-hFGF2	36.0	56.4	2.0
T7ACnew-hFGF2	37.5	72.7	2.7
DBv2-T7AC-hFGF2	37.1	102.4	3.8
DBv2-T7ACnew-hFGF2	34.9	127.5	4.7

3.4 Increased POI production by T7ACnew comprising tag variants

- Recombinant protein formation was higher with the DBv1-T7ACnew-tag and the DBv2-T7ACnew-tag compared to the DBv1-T7AC-tag or the DBv2-T7AC-tag, respectively, for every tested model protein (see Table 11 and Fig. 7). The comparative increase in recombinant soluble protein titre at the end of fermentation is dependent on the fusion tag variant used. With PTH increases in specific titre of 4-fold and 1.35-fold have been observed for DBv1-T7ACnew and DBv2-T7ACnew, respectively, compared to DBv1-T7AC and DBv2-T7AC. hFGF2 showed an increase in titre of 1.25-fold comparing DBv2-T7ACnew with DBv2-T7AC. Measured specific and volumetric recombinant protein titres are summarized in Table 11 above.

Example 4: Additional novel tag sequences enhancing protein expression

- To investigate the influence of alternative silent mutations on recombinant protein expression, further tag constructs comprising alternative silent mutations were tested in high cell density fermentations.

4.1 Materials and Methods

4.1.1 Nucleic acid sequences encoding the alternative T7AC tag variants

Two additional tag constructs, comprising alternative silent mutations in the T7AC tag and termed “T7AC new-alt. 1” and “T7AC new-alt. 2”, respectively, were compared to the original T7AC tag. The respective sequences are shown in table 12.

Table 12: Sequences encoding the T7AC tag

Tag	Amino acid sequence	Nucleic acid sequence
T7AC	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO: 7)	CTG GAG GAT CCG GAA CGC AAC AAA GAG CGA AAG GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO: 8)
T7AC new	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO: 7)	CTA GAA GAT CCC GAA AGG AAT AAA GAA AGG AAA GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO: 9)
T7AC new-alt. 1	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO: 7)	TTA GAA GAT CCA GAA AGA AAT AAA GAA AGA AAA GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO: 10)
T7AC new-alt. 2	LEDPERNKERKEAEL- QAQTAEQ (SEQ ID NO: 7)	TTG GAA GAT CCA GAA CGA AAT AAA GAA AGA AAA GAA GCT GAG TTG CAA GCT CAA ACC GCT GAG CAA (SEQ ID NO: 11)

10

4.1.2 POI fusion constructs

The alternative tag variants were tested in fusion proteins comprising the T7AC tag variants and the pharmaceutically relevant model proteins PTH or TNFα, respectively. The individual constructs are shown in Table 14 and the respective sequences in Fig. 10.

15

Table 13: General amino acid sequence of the fusion proteins used in this experiment

T7AC - 6H - GSG - VDVAD - POI	
T7AC	amino acid sequence of T7AC: SEQ ID NO: 7
6H	amino acid sequence of the His-tag: HHHHHH
GSG	linker with the amino acid sequence GSG
VDVAD	amino acid sequence of the recognition site of a caspase 2
POI	PTH or TNFα

Table 14: Amino acid and nucleic acid sequences of the tested fusion proteins

Name of fusion protein	amino acid sequence	nucleic acid sequence
T7AC - PTH	SEQ ID NO: 73	SEQ ID NO: 74
T7AC new-alt.2 - PTH	SEQ ID NO: 73	SEQ ID NO: 75
T7AC - TNFα	SEQ ID NO: 76	SEQ ID NO: 77
T7AC new - TNFα	SEQ ID NO: 76	SEQ ID NO: 78
T7AC new-alt. 1 - TNFα	SEQ ID NO: 76	SEQ ID NO: 79
T7AC new alt. 2 - TNF	SEQ ID NO: 76	SEQ ID NO: 80

5 **4.1.3 Cell bank generation**

 The nucleic acids encoding the fusion proteins were synthesized and cloned into the respective expression vectors (pET30a-cer derivative). The correct sequences of all expression constructs were confirmed by sequencing. The expression vectors were transformed into *E. coli* BL21(DE3). A research cell bank was produced. The expression of the POI-fusions was under the control of the T7 promoter system and the tZenit terminator (Mairhofer 2014).

4.1.4 Fed batch fermentation

 All production strains were cultivated in a 1 L Multifors bioreactor with a working volume of 0.7 L and a starting batch volume of 0.4 L. The preculture was inoculated with the research cell bank and performed in a shake flask containing preculture medium at 37 °C. The batch medium in the 1 L bioreactor was inoculated with preculture in a ratio of 1:100 when the OD₅₅₀ in the preculture reached a value of 2. Temperature, pH (by addition of an ammonia solution), and dissolved oxygen (pO₂; by stirrer speed and increasing the concentration of O₂ in the in-air) in the bioreactor were controlled. The pO₂ was kept constant at 20 %, the aeration at 1 vvm and

pressure at 1 bar. After an increase of pO₂ in the batch phase the glucose feed was started (exponential feed for 9.5 hours and constant feed for 12 hours). Expression of the POI fusion protein was started 11.5 hours after start of the exponential feed by the addition of 0.179 g IPTG (Isopropyl-β-D-thiogalactopyranoside). The end of fermentation was reached at 21.5 hours after start of glucose feed. The production phase was 10 hours. Temperature was 37 °C in growth phase at 30 °C in production phase. pH was kept at 7.0 +/- 0.2.

4.1.5 Analytics

Samples of the fermentation broth were taken at 5 time points. The WCW [g/L] (wet cell weight) was measured by weighing the pellet of 1 mL of culture broth after centrifugation.

4.1.5.1 Sample preparation

The pellets from the fermentation broth samples were resuspended in Bug-Buster™ from Novagen (including 1 μl Lysonase™ from Novagen) by vortexing until the pellet has been completely resuspended and centrifuged after incubation for 15 minutes. The supernatant was used for the determination of the titer of the intracellular soluble POI fusion protein.

4.1.5.2 Determination of the volumetric titer of the POI fusion proteins

The titer was determined using SDS-Page. scFvM (a single chain antibody) was used as a standard for the calibration curve for quantification.

4.2 Results

Recombinant production of POI fusions (cytosolic, soluble titer) comprising the T7AC tag in high cell density fermentations was significantly increased by using the nucleic acids of the invention (T7ACnew (SEQ ID NO: 9), T7ACnew-alt. 1 (SEQ ID NO: 10), T7ACnew-alt. 2 (SEQ ID NO: 11)) compared to the original nucleic acid sequence (T7AC (SEQ ID NO: 8)) as shown in Figure 8 and 9 and Tables 15 and 16.

Table 15: Expression of PTH fusion proteins from nucleic acid sequences comprising nucleic acid sequences of the present invention compared to the original nucleic acid sequence (T7AC) encoding the T7AC tag.

Tag	T7AC	T7ACnew-alt. 2
volumetric titer [g/L]	12.6	17.6
specific titer [mg/g WCW]	37	50
fold change volumetric titer	-	1.4
fold change specific titer	-	1.4
WCW [g/L]	342	350

- 5 Table 16: Expression of TNF α fusion proteins from nucleic acid sequences comprising nucleic acid sequences of the present invention compared to the original nucleic acid sequence (T7AC) encoding the T7AC tag.

Tag	T7AC	T7ACnew	T7ACnew-alt. 1	T7ACnew-alt. 2
volumetric titer [g/L]	13.4	16.5	21.3	19.8
specific titer [g/L]	41	50	66	66
fold change volumetric titer	-	1.2	1.6	1.5
fold change specific titer	-	1.2	1.6	1.6
WCW [g/L]	328	330	325	298

REFERENCES

- Baeshen MN, Al-Hejin AM, Bora RS, Ahmed MM, Ramadan HA, Saini KS, Baeshen NA, Redwan EM: Production of Biopharmaceuticals in *E. coli*: Current Scenario and Future Perspectives. *J Microbiol Biotechnol* 2015, 25:953-962.
- 5 Cserjan-Puschmann M, Lingg N, Engele P, Kröß C, Loibl J, Fischer A, Bacher F, Frank A-C, Öhlknecht C, Brocard C, et al: Production of Circularly Permuted Caspase-2 for Affinity Fusion-Tag Removal: Cloning, Expression in *Escherichia coli*, Purification, and Characterization. *Biomolecules* 2020, 10.
- Dana A, Tuller T: The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* 2014, 42:9171-9181.
- 10 Fink M, Schimek C, Cserjan-Puschmann M, Reinisch D, Brocard C, Hahn R, Striedner G: Integrated process development: The key to improve Fab production in *E. coli*. *Biotechnology Journal* 2021a, 16:2000562.
- Fink M, Cserjan-Puschmann M, Reinisch D, Striedner G: High-throughput microbioreactor provides a capable tool for early stage bioprocess development. *Scientific Reports* 2021b, 11:2056.
- 15 Fink M, Vazulka S, Egger E, Jarmer J, Grabherr R, Cserjan-Puschmann M, Striedner G: Microbioreactor Cultivations of Fab-Producing *Escherichia coli* Reveal Genome-Integrated Systems as Suitable for Prospective Studies on Direct Fab Expression Effects. *Biotechnology Journal* 2019, 14:1800637.
- 20 Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y: Codon usage of highly expressed genes affects proteome-wide translation efficiency. 2018, 115:E4940-E4949.
- Goodman DB, Church GM, Kosuri S: Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* 2013, 342:475-479.
- 25 Kane JF: Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Current Opinion in Biotechnology* 1995, 6:494-500.
- Köppl C, Lingg N, Fischer A, Kröß C, Loibl J, Buchinger W, Schneider R, Jungbauer A, Striedner G, Cserjan-Puschmann M: Fusion Tag Design Influences Soluble Recombinant Protein Production in *Escherichia coli*. *International Journal of Molecular Sciences* 2022, 23:7678.
- 30

- Kröß C, Engele P, Sprenger B, Fischer A, Lingg N, Baier M, Öhlknecht C, Lier B, Oostenbrink C, Cserjan-Puschmann M, et al: PROFICS: A bacterial selection system for directed evolution of proteases. *Journal of Biological Chemistry* 2021, 297:101095.
- 5 Kudla G, Murray AW, Tollervey D, Plotkin JB: Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science* 2009, 324:255-258.
- Lingg N, Kröß C, Engele P, Öhlknecht C, Köppl C, Fischer A, Lier B, Loibl J, Sprenger B, Liu J, et al: CASPON platform technology: Ultrafast circularly permuted caspase-2 cleaves tagged fusion proteins before all 20 natural
- 10 amino acids at the N-terminus. *New Biotechnology* 2022, 71:37-46.
- Mairhofer Juergen et al.: Preventing T7 RNA Polymerase Read-through Transcription - A Synthetic Termination Signal Capable of Improving Bioprocess Stability. *ACS Synthetic Biology*. 2014, 4/3: 256-273
- Sharp PM, Li WH: The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*
- 15 1987, 15(3):1281-95.
- Sørensen MA, Kurland CG, Pedersen S: Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol* 1989, 207:365-377.
- Sørensen HP, Mortensen KK: Soluble expression of recombinant proteins in the
- 20 cytoplasm of *Escherichia coli*. *Microbial Cell Factories* 2005, 4:1.
- Stargardt P, Feuchtenhofer L, Cserjan-Puschmann M, Striedner G, Mairhofer J: Bacteriophage Inspired Growth-Decoupled Recombinant Protein Production in *Escherichia coli*. *ACS Synthetic Biology* 2020, 9:1336-1348.
- Zuker M and Stiegler P: Optimal computer folding of large RNA sequences using
- 25 thermodynamics and auxiliary information. *Nucleic Acids Res* 1981, 9(1):133-148.

CLAIMS

1. An isolated nucleic acid sequence encoding a peptide tag comprising amino acid sequence SEQ ID NO:1 for expression of a protein or peptide of interest (POI) in a host cell, wherein said nucleic acid sequence comprises SEQ ID NO:2 comprising at least 3 silent mutations at positions selected from the group consisting of:
- a. codon encoding Pro at position 1 of SEQ ID NO:1, wherein CCG has been modified to CCC or CCA;
 - b. codon encoding Arg at position 3 of SEQ ID NO:1, wherein CGC has been modified to AGG, AGA or CGA;
 - c. codon encoding Asn at position 4 of SEQ ID NO:1, wherein AAC has been modified to AAT;
 - d. codon encoding Glu at position 6 of SEQ ID NO:1, wherein GAG has been modified to GAA;
 - e. codon encoding Arg at position 7 of SEQ ID NO:1, wherein CGA has been modified to AGG, or AGA; and
 - f. codon encoding Lys at position 8 of SEQ ID NO:1, wherein AAG has been modified to AAA.
2. The nucleic acid sequence of claim 1, wherein the peptide tag comprises SEQ ID NO:4 and said nucleic acid sequence comprises SEQ ID NO:5 comprising at least 3 silent mutations selected from the group consisting of:
- a. codon encoding Leu at position 1 of SEQ ID NO:4, wherein CTG has been modified to CTA, TTA or TTG;
 - b. codon encoding Glu at position 2 of SEQ ID NO:4, wherein GAG has been modified to GAA;
 - c. codon encoding Pro at position 4 of SEQ ID NO:4, wherein CCG has been modified to CCC or CCA;
 - d. codon encoding Arg at position 6 of SEQ ID NO:4, wherein CGC has been modified to AGG, AGA or CGA;
 - e. codon encoding Asn at position 7 of SEQ ID NO:4, wherein AAC has been modified to AAT;

- f. codon encoding Glu at position 9 of SEQ ID NO:4, wherein GAG has been modified to GAA;
- g. codon encoding Arg at position 10 of SEQ ID NO:4, wherein CGA has been modified to AGG, or AGA; and
- 5 h. codon encoding Lys at position 11 of SEQ ID NO:4, wherein AAG has been modified to AAA.
3. The nucleic acid sequence of claim 1 or 2, comprising at least 4 silent mutations, preferably at least 5 silent mutations and even more preferably at least 6 silent
- 10 mutations.
4. The nucleic acid sequence of claim 1 or 2, comprising at least 7 silent mutations, preferably at least 8 silent mutations.
- 15 5. The nucleic acid sequence of any one of claims 1 to 4, wherein the host cell is a eukaryotic or prokaryotic host cell, preferably a yeast cell, a mammalian cell or a bacterial cell.
6. The nucleic acid sequence of any one of claims 1 to 5, comprising a nucleic acid
- 20 sequence selected from the group consisting of SEQ ID NO:3, SEQ ID NO:6, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19 and SEQ ID NO:20.
- 25 7. A peptide tag encoded by the nucleic acid sequence of any one of claims 1 to 6.
8. An expression vector comprising the nucleic acid sequence of any one of claims 1 to 6.
- 30 9. The expression vector of claim 8, comprising the nucleic acid sequence of any one of claims 1 to 6 encoding a peptide tag, optionally a protease recognition and/or cleavage site, optionally a linker, and a gene of interest (GOI).

10. The expression vector of claim 9, comprising: a promoter sequence, a nucleic acid sequence encoding a 5' untranslated region of an expressed mRNA that comprises a ribosome binding site (RBS), the nucleic acid sequence of any one of claims 1 to 6 encoding a peptide tag, optionally a linker, optionally a protease recognition and/or cleavage site, and a cloning site, wherein the cloning site enables a gene of interest (GOI) encoding a POI to be inserted into the vector in-frame with said nucleic acid sequence of any one of claims 1 to 6 to encode a fusion protein comprising said peptide tag, optionally a linker and/or a protease recognition and/or cleavage site, and said POI.
11. The expression vector of claim 9, comprising a GOI inserted at the cloning site in-frame with the nucleic acid sequence of any one of claims 1 to 6 to encode a fusion protein.
12. A host cell comprising the expression vector of any one of claims 8 to 11.
13. A host cell comprising the nucleic acid sequence of any one of claims 1 to 6 integrated into its genome.
14. A method for the production of a recombinant protein of interest (POI), comprising culturing the host cell of claim 13 or a host cell comprising the expression vector of claim 11 for a period of time under conditions permitting expression of the POI in the form of a fusion protein, preferably wherein said host cell is a yeast cell, a mammalian cell or a bacterial cell.
15. The method of claim 14 further comprising any one or more of the steps of:
- recovering the fusion protein and/or POI,
 - purifying the fusion protein and/or POI,
 - removing the peptide tag,
 - further purifying the fusion protein and/or POI,
 - formulating the fusion protein and/or POI,
 - modifying the fusion protein and/or POI.