

Gene Name	Gene Description & Selection Reasoning
<p>BIRC5 aa: 166 bp: 498</p>	<p>BIRC5 (Baculoviral IAP repeat-containing 5) is a multitasking protein that has roles in preventing apoptosis and cell proliferation (PubMed:<a href="#">9859993</a>, PubMed:<a href="#">21364656</a>, PubMed:<a href="#">20627126</a>, PubMed:<a href="#">25778398</a>, PubMed:<a href="#">28218735</a>). In <a href="#">Reference Paper</a>, BIRC5 expression levels were analyzed in <i>E. coli</i>. Sequence optimization (generated by using a sliding window → narrowed down) was used for this human gene to be expressed in <i>E. coli</i>.</p>
<p>CAV1 aa: 179 bp: 537</p>	<p>CAV1 (Caveolin 1, caveolae protein) is involved in promoting cell cycle progression. It's a gene of interest because it is a tumor suppressor candidate making it potentially useful in recombinant expression. In <a href="#">Reference Paper</a>, CAV1 expression levels were analyzed in <i>E. coli</i>.</p>
<p>PTP4A3 aa: 187 bp: 561</p>	<p>Protein tyrosine phosphatase 4A3 (PTP4A3/PRL-3) is a well-documented protein that can be produced recombinantly in <i>E. coli</i>. This protein has been identified as a potential target to treat some cancers. It has been studied in (<a href="#">codon optimization papers</a>).</p>
<p>PA aa: 187 bp: 561</p>	<p>Polymerase acidic protein (PA) plays a role in viral RNA transcription and replication. It is from the Influenza A virus. It has been studied in (<a href="#">codon optimization papers</a>).</p>
<p>hPDF aa: 205 bp: 615</p>	<p>Human peptide deformylase (hPDF) is a target for cancer therapeutics. However, its expression is not very efficient in <i>E. coli</i>. This serves as a good target for benchmarks as past <a href="#">studies</a> have noted the valuable potential of codon optimization of this gene.</p>
<p>EMG1 aa: 245 bp: 735</p>	<p>EMG1 (Protein: <a href="#">nucleolar protein homolog (<i>S. cerevisiae</i>)</a>) is a human-based recombinant protein expressed in <i>E. coli</i>. This gene encodes a protein that methylates pseudouridine and is an essential eukaryotic protein. In <a href="#">Reference Paper</a>, EMG1 expression levels were analyzed in <i>E. coli</i>.</p>
<p>CDK1 aa: 298 bp: 894</p>	<p>CDK1 (Cyclin Dependent Kinase 1) codes for a protein which is essential for G1/S and G2/M phase transitions. It has been used in calculating prognostic value in human cancer (<a href="#">Reference Paper</a>). Higher expression levels correlated with a more-advanced tumor. Another reference paper analyzing the importance of this gene looked at expression and structure through recombinant expression in <i>E. coli</i> (<a href="#">Reference Study</a>).</p>
<p>CD80 aa: 307 bp: 921</p>	<p>Cd80 (Mus musculus CD80 antigen) is a protein-coding gene whose protein, once activated, induces T-cell proliferation and cytokine production. In <a href="#">Reference Paper</a>, Cd80 expression levels were analyzed in <i>E. coli</i>.</p>

TAS2R10 <b>aa: 308</b> <b>bp: 924</b>	Taste receptor, type 2, member 10 (TAS2R10) encodes proteins expressed in taste receptor cells. In <a href="#">Reference Paper</a> , TAS2R10 expression levels were analyzed in <i>E. coli</i> .
PIM1 <b>aa: 314</b> <b>bp: 942</b>	Pim-1 oncogene (PIM1) encodes a protein that plays a role in signal transduction in blood cells. The gene is expressed primarily in B-lymphoid and myeloid cell lines. It is found to be overexpressed in hematopoietic malignancies and in prostate cancer. In <a href="#">Reference Paper</a> , PIM1 expression levels were analyzed in <i>E. coli</i> .
FALVAC-1 <b>aa: 324</b> <b>bp: 972</b>	FALVAC-1 is a vaccine against Plasmodium Falciparum. This is a good benchmark as a synthetic gene of codon optimization tools because it is a real candidate vaccine that is recombinantly produced ( <a href="#">Reference Paper</a> ).
CREB1 <b>aa: 328</b> <b>bp: 984</b>	CREB1 ( <a href="#">cAMP responsive element binding protein 1</a> ) mutations can increase risk for certain diseases. Its most common variant is actually from a tumor. It is from a family of transcription factors that is expressed in the brain. In <a href="#">Reference Paper</a> , CREB1 expression levels were analyzed in <i>E. coli</i> .
JUN <b>aa: 332</b> <b>bp: 996</b>	JUN (Jun Proto-Oncogene, AP-1 Transcription Factor Subunit) can often be mutated to cause a cell to become a tumor cell. JUN is the “putative transforming gene” of avian sarcoma virus 17 ( <a href="#">Reference</a> ). It was expressed as a “transcription factor” in this <a href="#">Reference Paper</a> in <i>E. coli</i> ; its expression was measured in vivo.
CSNK1A1 <b>aa: 338</b> <b>bp: 1014</b>	CSNK1A1 ( <a href="#">Casein Kinase 1 Alpha 1</a> ) belongs to the protein kinase superfamily and is involved in phosphorylating numerous proteins involved in cellular functions. Recombinant human CSNK1a1 protein, fused to His-tag at N-terminus, was expressed in <i>E. coli</i> by <a href="#">CSNK1A1, 1-337aa Human, His tag, E.coli (GWB-ATG0D4)</a> .
MAPK1 <b>aa: 361</b> <b>bp: 1083</b>	Mitogen-Activated Protein Kinase 1 (MAPK1) is a part of the MAP kinase signal transduction pathway. It is of special interest because in this <a href="#">Reference Paper</a> , when expressed in <i>E. coli</i> , there was a significant difference in the protein yield for their wild-type and optimized genes: 24.3 and 11.5 mg/L respectively.
OPRM1 <b>aa: 401</b> <b>bp: 1203</b>	“This gene encodes one of at least three opioid receptors in humans; the mu opioid receptor (MOR). The MOR is the principal target of endogenous opioid peptides and opioid analgesic agents such as beta-endorphin and enkephalins. The MOR also has an important role in dependence to other drugs of abuse, such as nicotine, cocaine, and alcohol via its modulation of the dopamine system” ( <a href="#">Study</a> ). In <a href="#">Reference Paper</a> , OPRM1 expression levels were analyzed in <i>E. coli</i> .
LAMP1 <b>aa: 418</b> <b>bp: 1254</b>	Lysosomal-associated membrane protein 1 (LAMP1) is a protein coding gene that is associated with <a href="#">Chediak-Higashi Syndrome</a> and <a href="#">Gaucher's Disease</a> . This is a gene of interest because it may play a role in tumor cell metastasis. It was

	expressed as a “transcription factor” in this <a href="#">Reference Paper</a> in <i>E. coli</i> ; its expression was measured in vivo.
<b>NGFR</b> <b>aa: 428</b> <b>bp: 1284</b>	Nerve growth factor receptor (NGFR) is a protein coding gene associated with important pathways within the cell. Diseases associated with NGFR include <a href="#">Prurigo Nodularis</a> and <a href="#">Infiltrative Basal Cell Carcinoma</a> . In <a href="#">Reference Paper</a> , NGFR expression levels were analyzed in <i>E. coli</i> .
<b>GSK3B</b> <b>aa: 434</b> <b>bp: 1302</b>	GSK3B (Glycogen Synthase Kinase 3 Beta) encodes a protein that is a negative regulator of glucose homeostasis. Mutations that lead to defects in this gene are associated with Parkinson disease and Alzheimer disease. This is a gene of interest because it is found to be an important enzyme in glycogen metabolism. In <a href="#">Reference Paper</a> , GSK3B expression levels were analyzed in <i>E. coli</i> .
<b>CLN3</b> <b>aa: 439</b> <b>bp: 1317</b>	CLN3 (Ceroid-lipofuscinosis, neuronal 3) is a gene that encodes a protein involved in lysosomal function. This is an interesting gene of study because it can cause neurodegenerative diseases once mutated (Batten disease). In <a href="#">Reference Paper</a> , CLN3 expression levels were analyzed in <i>E. coli</i> which makes it a good benchmark gene to compare to.
<b>MAPKAPK5</b> <b>aa: 472</b> <b>bp: 1416</b>	Mitogen-activated protein kinase-activated protein kinase 5 (MAPKAPK5) is a gene that encodes for a tumor suppressor. It is activated by the MAPK1 kinase and also plays an important role in the MAP kinase signal transduction pathway. It is of special interest because in this <a href="#">Reference Paper</a> , when expressed in <i>E. coli</i> , there was a significant difference in the protein yield for their wild-type and optimized genes: 14.9 and 80 mg/L respectively.
<b>SMARCD1</b> <b>aa: 475</b> <b>bp: 1425</b>	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1 (SMARCD1) is a gene that encodes a protein which is thought to regulate transcription. It has been found to inhibit certain human glioblastoma cells, sparking interest in the overexpression of this gene. In <a href="#">Reference Paper</a> , SMARCD1 expression levels were analyzed in <i>E. coli</i> .
<b>AKT1</b> <b>aa: 481</b> <b>bp: 1443</b>	AKT serine/threonine kinase 1, naturally found in humans, encodes one of three kinases which are referred to as protein kinase B alpha, beta, and gamma. In the <a href="#">Reference Paper</a> , AKT1 was found to have a reduction in optimized heterologous expression by 10% compared to the wild-type gene. This makes it a particularly interesting benchmark gene to gauge the performance of codon optimization tools.
<b>LCK</b> <b>aa: 510</b> <b>bp: 1530</b>	Lymphocyte-specific protein tyrosine kinase ( <a href="#">LCK</a> ) is a gene that encodes a protein that is a key signaling molecule “...in the selection and maturation of developing T-cells” ( <a href="#">Reference</a> ). In <a href="#">Reference Paper</a> , LCK expression levels were analyzed in <i>E. coli</i> . It was found that LCK showed a 500-fold increase in mRNA transcripts for the sequence-optimized gene.

RPS6KB1 <b>aa: 526</b> <b>bp: 1578</b>	Recombinant Human Ribosomal protein S6 kinase beta-1 (RPS6KB1) promotes cell proliferation and protein synthesis. It has been extensively studied in yeast and <i>E. coli</i> expression systems.
PAK1 <b>aa: 546</b> <b>bp: 1638</b>	p21 protein (Cdc42/Rac)-activated kinase 1 (PAK1) is a gene that encodes a family member of p21-activating kinases. Overexpression of this gene can result in cancer growth and mutations in this gene are associated with diseases such as macrocephaly. In <a href="#">Reference Paper</a> , PAK1 expression levels were analyzed in <i>E. coli</i> .
PLK1 <b>aa: 604</b> <b>bp: 1812</b>	Polo like kinase 1 (PLK1) is a human gene that has been targeted in cancer therapy. A reduction in PLK1 levels is associated with the inhibition of cell proliferation and induced apoptosis. In <a href="#">Reference Paper</a> , PLK1 expression levels were analyzed in <i>E. coli</i> .
PEA <b>aa: 613</b> <b>bp: 1839</b>	Pseudomonas aeruginosa exotoxin A (PEA) is an important pathogenic factor <a href="#">1</a> . It retains high immunogenicity even after detoxification, enabling its use as vaccine adjuvants and vaccine carriers. This is a good benchmark as it can be produced for a low-cost at a large-scale in <i>E. coli</i> . Further, the ( <a href="#">Reference Paper</a> ) finds that codon optimization enhances expression of PEA in <i>E. coli</i> -- thus, if the tool presented in this research can achieve similar/better results than the paper and/or other approaches, it will be considered an improvement.
TAP1 <b>aa: 749</b> <b>bp: 2247</b>	Transporter 1, ATP-binding cassette, sub-family B (MDR/TAP) (TAP1) encodes a protein known to be involved in molecular transport and drug resistance. In <a href="#">Reference Paper</a> , TAP1 expression levels were analyzed in <i>E. coli</i> .
NOC2L <b>aa: 750</b> <b>bp: 2250</b>	Nucleolar complex associated 2 homolog ( <i>S. cerevisiae</i> ) (NOC2L) can control major aspects of transcriptional regulation. It is an RNA/Ribosomal protein. In <a href="#">Reference Paper</a> , NOC2L expression levels were analyzed in <i>E. coli</i> .
UBTF <b>aa: 765</b> <b>bp: 2295</b>	Upstream Binding Transcription Factor (UBTF) encodes a protein important for ribosomal RNA transcription. The UBTF studied originates from human genes. UBTF is well-known as a recombinant protein and has been noted as useful as a blocking peptide for certain antibodies.
BRAF1 <b>aa: 767</b> <b>bp: 2301</b>	BRAF1 (v-raf murine sarcoma viral oncogene homolog B1) is a protein kinase that transduces mitogenic signals from the cell membrane to the nucleus. This kinase phosphorylates MAP2K1 which activates the MAP kinase signal pathway (PubMed: <a href="#">21441910</a> , PubMed: <a href="#">29433126</a> ). This gene is a gene of interest because it is frequently mutated and allows a cell to become a tumor cell. In <a href="#">Reference Paper</a> , BRAF1 expression levels were analyzed in <i>E. coli</i> .
FGFR4 <b>aa: 803</b>	FGFR4 (Fibroblast growth factor receptor 4) comes under the kinases category and the protein encoded by this gene is a tyrosine kinase and cell surface receptor

<b>bp: 2409</b>	for fibroblast growth factors. Diseases associated with FGFR4 include <a href="#">Prostate Cancer</a> and <a href="#">Rhabdomyosarcoma</a> . In <a href="#">Reference Paper</a> , FGFR4 expression levels were analyzed in <i>E. coli</i> .
<b>LEMD3</b> <b>aa: 912</b> <b>bp: 2736</b>	LEM domain containing 3 (LEMD3) is a human gene that is of interest. Mutations in this gene have been associated with osteopoikilosis, Buschke-Ollendorff syndrome and melorheostosis. It was expressed as a “membrane protein” in this Reference Paper in <i>E. coli</i> ; its expression was measured in vivo.
<b>MMLP3</b> <b>aa: 945</b> <b>bp: 2835</b>	Proteins associated with the MMP family are known to be involved in the breakdown of extracellular matrices in physiological processes within the cell. The MMP3 or MMLP3 gene encodes an enzyme that degrades glycoproteins such as fibronectin. It has been studied in ( <a href="#">codon optimization papers</a> ).
<b>CEBPZ</b> <b>aa: 1055</b> <b>bp: 3165</b>	CEBPZ ( <a href="#">CCAAT/enhancer binding protein zeta</a> ) is a human protein-encoding gene that plays a role in responding to environmental stimuli (heat). In <a href="#">Reference Paper</a> , CEBPZ expression levels were analyzed in <i>E. coli</i> which makes it a good benchmark gene to compare to.
<b>KIF11</b> <b>aa: 1057</b> <b>bp: 3171</b>	KIF11 (Kinesin family member 11) was expressed recombinantly in <i>E. coli</i> in this <a href="#">study</a> . “This gene encodes a motor protein that belongs to the kinesin-like protein family. Members of this protein family are known to be involved in various kinds of spindle dynamics.” ( <a href="#">KIF11 Gene (Protein Coding)</a> ).
<b>NPR1</b> <b>aa: 1062</b> <b>bp: 3186</b>	NPR1 ( <a href="#">Natriuretic Peptide Receptor 1</a> ) is a gene that encodes a peptide receptor that is located in the kidney, lungs, and adipocytes. NPR1 is associated with diseases including congestive heart failure and malt worker’s lung. In <a href="#">Reference Paper</a> , NPR1 expression levels were analyzed in <i>E. coli</i> .
<b>FLT1</b> <b>aa: 1339</b> <b>bp: 4017</b>	FLT1 ( <a href="#">Fms-related tyrosine kinase 1</a> ) is a gene that encodes a member of the vascular endothelial growth factor receptor family. In <a href="#">Reference Paper</a> , FLT1 expression levels were analyzed in <i>E. coli</i> .
<b>PDCD11</b> <b>aa: 1872</b> <b>bp: 5616</b>	Programmed Cell Death 11 (PDCD11) is a useful binding protein that colocalizes with U3 RNA (MIM 180710) in the nucleolus and is required for rRNA maturation and generation. It is important because, as a plasma protein, it is within the nucleolus of the cell and is a ribosomal protein. In <a href="#">Reference Paper</a> , PDCD11 expression levels were analyzed in <i>E. coli</i> .