OmicsLink™ Expression Clones
ORFExpress ™ Shuttle clones

Large Collection of
Full Length ORF of Human Genes

GeneCopoeia, Inc.
www.genecopoeia.com
19520 Amaranth Drive
Germantown, MD 20874
Tel: 301-515-6982
Fax: 301-515-6983

Contents

1. Clone collection, curation and construction

a) GeneCopoeia's human full length gene product line consists of three sets of over 15,000 high fidelity and highly versatile full length coding ORF clones in three unique vector systems: one shuttle vector system (ORFExpress™-Shuttle vector) and two expression vector systems (OmicsLink™-CMV mammalian expression vector system and OmicsLink™-T7 bacterial expression vector system)

b) GeneCopoeia's collection of over 15,000 full-length Human genes was chosen through a strict selection process. This process included the extraction, comparison, and validation of, gene sequence and annotation information from multiple public and private sources; the clustering and reduction of redundant gene sequences; filtering out error sequences or sequencing errors, plus several other manual curation steps. Once this stringent selection process was completed, the entire coding open reading frames (ORFs) of these genes were obtained by utilizing a high fidelity polymerase chain reaction (PCR). The templates were selected from sequence-verified full-length cDNA clones or plasmids from high quality human tissue cDNA libraries.

c) The ORFExpress™-Shuttle clones were manufactured to be compatible with Invitrogen Corporation's Gateway® Technology, a universal cloning system developed by Invitrogen Corp. This makes it possible the rapid and simple transfer of the coding ORFs into any Gateway® Expression Vector for the expression and functional analysis of target proteins in many hosts such as *E. coli*, yeast, baculovirus, CHO and mammalian cell lines. See Appendix I for more details about Gateway® Technology and its compatibility with ORFExpress™-Shuttle clones.

d) OmicsLink™ expression clone product line offers Human full length ORFs in two expression vectors, T7 and CMV promoter driven vectors with optimal translation signals (Shine-Dalgarno and Kozak) for *E. coli* and mammalian cell expression systems, respectively. Proteins are expressed with unique tag that facilitates purification and

down-stream analysis (see Figure 1). These two sets of ORF expression clones are available as ready-for-shipping catalog products.

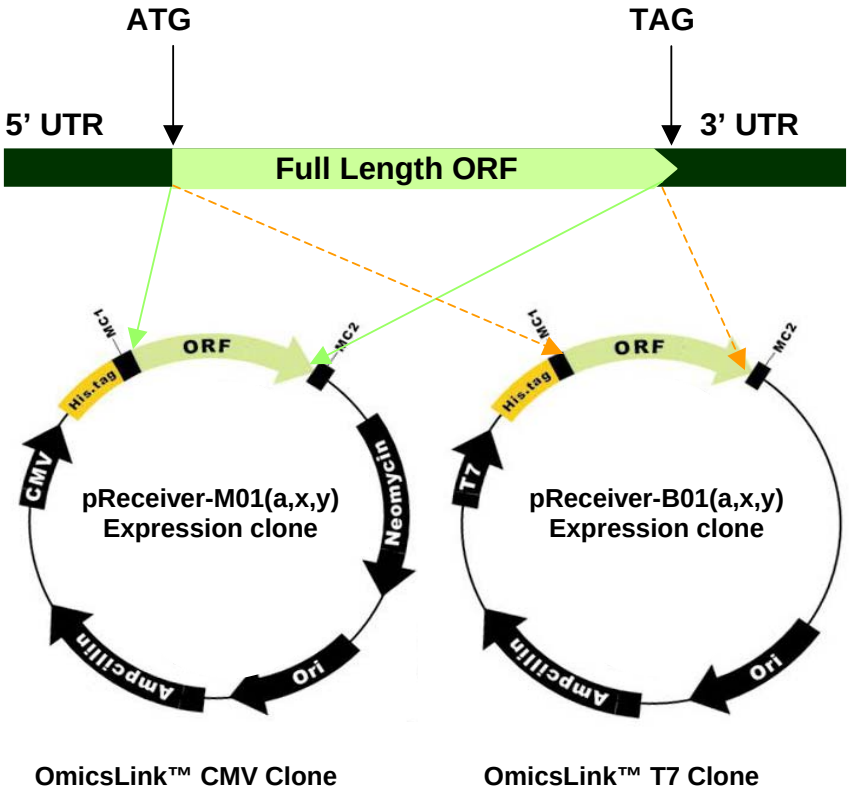

Figure 1. Examples of OmicsLink™ expression clones: OmicsLink™ bacterial (T7 promoter driven) and mammalian (CMV promoter driven) expression clones

e) GeneCopoeia's human full length ORF clones are available now in Invitrogen's Gateway® compatible shuttle vector (ORFExpress™ clones) and 7 expression-ready vectors (OmicsLink™ expression clones). See table 1 for details.

Table 1: Various expression vectors and their features

Vector Catalog #	Promoter	Host Cell	Stable Transfection	Tag
ORFExpress™ Shuttle Clone	Promoter less	N/A	N/A	N/A
pReceiver-M01	CMV	Mammalian cell	Yes	N-His
pReceiver-M02	CMV	Mammalian cell	Yes	None
pReceiver-M03	CMV	Mammalian cell	Yes	C-GFP
pReceiver-B01	T7	*E. coli*	N/A	N-His
pReceiver-B02	T7	*E. coli*	N/A	None
pReceiver-I01	AcMNPV polyhedrin	Insect	N/A	N-His
pReceiver-Y01	GAL1	Yeast	N/A	C-His

2. Applications of OmicsLink™ and ORFexpress™ clones

a) OmicsLink™ clones are expression ready and can be directly used for expression of proteins in bacterial or mammalian cells without further sub-cloning procedures.
b) OmicsLink™ clones in mammalian expression vectors can be used for both transient transfection and stable transfection of mammalian cells.
c) OmicsLink™ clones can be used to produce proteins in cell-free translation systems
d) Expressed proteins with OmicsLink™ clones have unique tags that could be used for concentration of lowly expressed proteins, studies such protein-molecule interactions and cellular localization.
e) GeneCopoeia's ORF clones can also be used to produce templates for esiRNA and generate probes for *in situ* hybridization

3. Advantages of GeneCopoeia's full length ORF clones over whole transcript cDNA clones from other vendors (see table 2 for summary)

a) GeneCopoeia's full-length genes were selected through a careful and rigorous curation and selection process including steps to ensure the most current valid version of RefSeq sequences used in cloning.
b) In GeneCopoeia' clones, only the validated full length coding ORFs were cloned into carefully constructed and highly versatile vector systems that are designed for easy down stream assays and functional analyses, while clones from other vendors are selected by sequencing individual clones from libraries constructed with the whole sequences of unprocessed full length genes (some are partial length genes) with 5' and 3' UTRs intact.
c) QC/QA of ORF only clones are relatively easy and take fewer resources. Sequencing validation covers the most important coding regions. Specially designed cloning primers and enzyme digestion sites at the clone insertion sites make it easy for re-clone, size validation and enzyme digestion validation.
d) ORF only clones have fewer potential artificial mutations and genetic variation related issues.
e) Full-length cDNA clones including long 3' and 5' UTRs increase the difficulty for cloning and the instances of artificial mutations.
f) Partially sequencing validations at 5' and 3' ends may not cover any ORF region due to long UTRs of many genes in other vendor's clones.
g) GeneCopoeia's ORF clones were constructed using GeneCopoeia's patented proprietary EnzyStart™ amplification and licensed recombinational cloning system, which generates high fidelity and low mutation rate PCR product. See Appendix II for details about this technology.

Table 2. Advantages of OmicsLink™'s Full Length ORF Expression Clones

Clone Characteristics	GeneCopoeia's ORF Clones	cDNA Clones from Other Vendors
Gene Transcript	ORFs (Start to stop only)	Whole transcripts including 5' and 3' UTRs
Expression Ready	Yes	Some are, most will need further subcloning procedures to remove 5' and 3' UTRs
Mammalian cell transfection **- Transient Transfection** **- Stable Transfection**	 Yes Yes	 Only a small number of clones No
Unique protein tags - Protein-molecule interactions - Concentration of lowly expressed proteins	Yes Yes Yes	No No No
Produce proteins via *in vitro* cell-free translation systems	Yes	No
Produce templates for esiRNA	Yes	No
Probe generation for *in situ* hybridization	Yes	No

4. QC/QA process for GeneCopoeia's full length gene clones

a) PCR amplification with Gene specific primers to do size validation.
b) Restriction enzyme digestion analysis.
c) 5' and 3' end sequencing validation.
d) Sequence analysis of artificial mutations vs. genetic variations (splicing and SNPs), see Appendix III for details.

5. Difficulties and limitations for do-it-yourself cloning.

For most researchers/investigators to clone full-length ORF clones, generating full length ORF by PCR is still the best methodology to use for economic and technical reasons. They may face some or all of the following factors and difficulties:

a) RNA secondary structure interference. Even for many genes with ORFs of a few hundred base pairs, it is not uncommon that secondary structures of their mRNAs make

it difficulty to produce first strand cDNA by reverse transcriptase, a method that are commonly used to generate ORFs via RT-PCR

b) GC content issues: both overall GC content of a full length ORF and those of local regions within full length ORF affect significantly the success rates of PCR used to generate full length ORFs of genes of interest. . Negative effect of regional high GC% within a full length ORF on PCR reactions is usually overlooked yet they are present in significant number human genes (see figure 2 and table 3 for examples of genes with regional high GC%) .

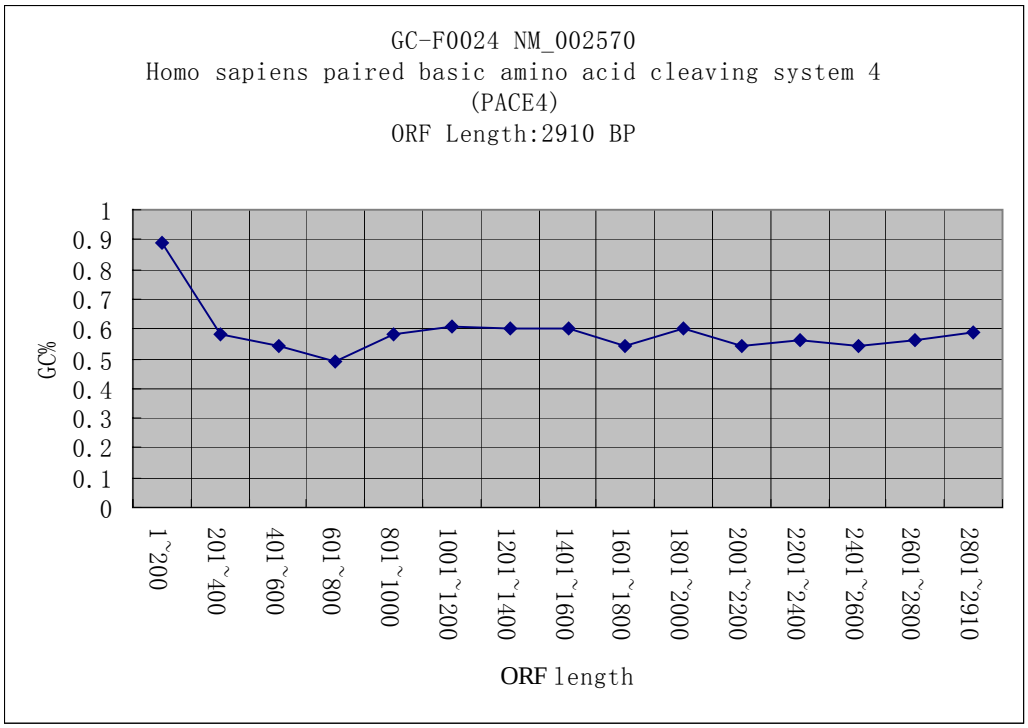

Figure 2. An example of a gene with high regional GC%

Table 3. More examples of ORFs with high regional GC%

Gene_ID	ORF Length	1-150bp	150-300bp	300-450bp	450-600bp	600-750bp	750-900bp	900-1050bp
GC-E0521	1248	0.84	0.67	0.61	0.60	0.64	0.67	0.67
GC-T4157	870	0.84	0.58	0.58	0.46	0.54	0.43	
GC-T7234	1092	0.84	0.42	0.44	0.44	0.35	0.42	0.45
GC-W0082	1155	0.83	0.76	0.66	0.74	0.80	0.80	0.75
GC-U0310	894	0.83	0.74	0.74	0.57	0.56	0.60	
GC-M0876	657	0.83	0.76	0.58	0.52	0.20		
GC-V1209	2364	0.82	0.74	0.66	0.61	0.65	0.57	0.55
GC-Q0263	1467	0.82	0.58	0.55	0.64	0.63	0.58	0.56
GC-U0111	1224	0.82	0.69	0.71	0.60	0.62	0.68	0.60
GC-T1024	2634	0.82	0.77	0.79	0.63	0.60	0.60	0.59
GC-C0759	1452	0.82	0.77	0.56	0.48	0.58	0.56	0.58
GC-G0716	1365	0.82	0.66	0.58	0.47	0.58	0.60	0.48
GC-T1365	1053	0.82	0.76	0.58	0.57	0.57	0.56	0.60
GC-F0121	954	0.82	0.41	0.48	0.57	0.56	0.52	0.20
GC-Q0071	1881	0.82	0.58	0.57	0.61	0.55	0.60	0.63
GC-U1028	1089	0.82	0.77	0.40	0.46	0.54	0.48	0.44
GC-W1349	2052	0.82	0.74	0.59	0.62	0.52	0.52	0.62
GC-M0001	1503	0.82	0.60	0.46	0.34	0.52	0.52	0.52
GC-U1140	1611	0.82	0.68	0.65	0.60	0.52	0.54	0.61
GC-U0120	687	0.82	0.66	0.48	0.40	0.27		
GC-T5655	474	0.82	0.68	0.52	0.05			
GC-T2253	360	0.82	0.73	0.28				
GC-C0648	909	0.81	0.60	0.66	0.62	0.68	0.56	0.02
GC-T0552	1695	0.81	0.52	0.43	0.45	0.64	0.48	0.42
GC-Z0210	858	0.81	0.60	0.66	0.70	0.64	0.36	

c) High global and regional GC and/or AT rich contents usually result in high mutation rates by PCR and increase the difficulty in the sequencing validation of full length ORF clones
d) Primers designed for cloning full length ORFs are of fewer choices and not primer-optimized for PCR reaction.
e) The fidelity of PCR product is significantly lower using commonly available commercial PCR reaction systems compared with those generated using GeneCopoeia's patented high fidelity and high yield PCR system.
f) With regular commercial PCR reaction components, it has been reported that PCR failure rate could be up to 25%.
g) Hard to keep up with sequence version revised in public database such as NCBI

6. Service for whole ORF sequencing validation and mutation correction

For a minimal fee to cover reagents, GeneCopoeia offers to sequence and validate the entire length of ORF, and/or correct/change any variations/mutations observed ORF sequence to versions of customer's choosing.

APPENDIX I. ORFExpress™-Shuttle Clones and Gateway® Technology

The Gateway® Cloning Technology is based on the site-specific recombination reaction of phage lambda, an efficient biochemical process that conserves genetic information. The transfer of an ORF from ORFExpress™-Shuttle Clones to the Destination Vector is a one step simple and rapid operation (60 minutes at room temperature), which saves a significant amount of time and effort from the tedious, multiple step, sub-cloning process involving enzyme digestions, ligations, gel electrophoresis and purification (see Figure 3). Furthermore, a family of Destination Vectors that contain different promoters and fusion-tags that meet the distinct requirements for production and purification of desired recombinant proteins in several expression hosts is available directly from Invitrogen Corp. For more detailed information on the Gateway® Cloning Technology, please visit Invitrogen Corporation's web site at www.invitrogen.com.

Figure 3. ORFExpress™-Shuttle Clones

APPENDIX II. GeneCopoeia's patented proprietary EnzyStart™ amplification/cloning system

1) The system incorporated a new high fidelity PCR composition that improves fidelity of generating ORFs by 10 fold (See the assay method in Yang, SW, (2002) *Nucleic Acids Research* Vol. 30 No. 19 P4314-4320)
2) EnzyStart™'s ultra-high sensitivity and specificity minimize the PCR cycles needed to clone ORF fragments, thus, minimize the mutations caused by the amplification process (see Figure 4 for the amplification result of our proprietary system compared to commonly used ones from other commercial sources)
3) The system utilizes recombination cloning protocol that requires less molecules of ORFs than traditional restriction and ligation method, thus further reduces the number of PCR cycles, which in turn reduces PCR induced mutations
4) It has become known that errors introduced by PCR primers (0.5% error rate in synthesis by generally accepted estimation) are very significant. EnzyStart™ system also incorporated a patent pending method to reduce the mutation in primer regions of clones.

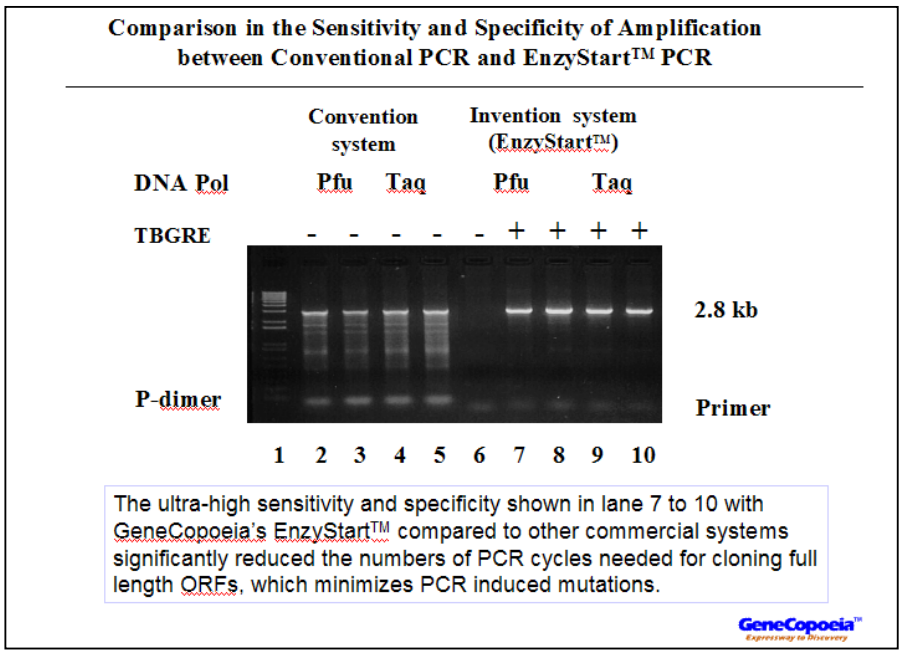

Figure 4. Comparison in the sensitivity and specificity of amplification between EnzyStart™ and conventional PCR system

APPENDIX III. Sequence analysis of artificial mutations vs genetic variations (splicings and SNPs)

1) Sequence analysis shows that over 50% of GeneCopoeia's full length ORF clones have no discrepancies compared with original sequence records in sequence verified regions
2) Observed discrepancies or variations fall into the following categories:
   a) About 50% are multiple silent variations, which are more likely to be attributed to SNP variations than PCR mutations, see Case 1 in figure 5. Multiple silent variations (17 in this clone) were often observed. Although no SNPs have been reported in these locations, non-randomness of these 17 variations makes it very unlikely the result of random mutations caused by PCR. Many of these variations are also seen in finished genome sequence, which is not shown in this figure.

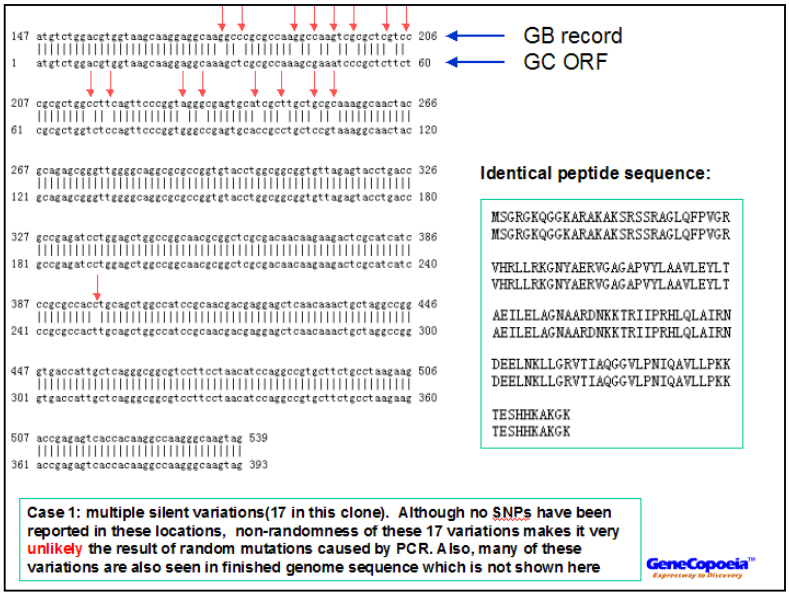

Figure 5. Example of silent variations

   b) Vast majority of coding variations (amino acid changing, frame-shifting, etc.) can be attributed to SNP variations, likely errors in original public sequence records, and alternate splicing events seen in figure 6, and 7.

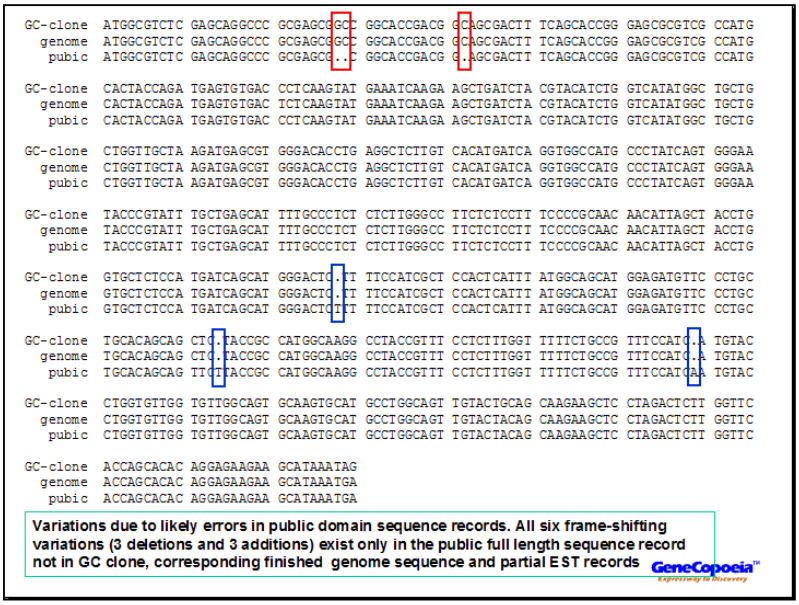

Figure 6. Example of frame-shifting changes due to errors, splicing variations and SNPs in public sequences. GC clone sequence is consistent with genome sequence while GenBank sequence contains frame-shifting variations and significant substitutions that are not present in genome

Figure 7. Example of amino acid substitution variations due to likely SNPs. Almost all substitutions can be seen in at lease two or more sequences of different origins: two forms of publicly submitted mRNAs, one genome sequence and GeneCopoeia clone.

c) In rare cases (such as number 7 variation in figure 7), variations were only observed in GC clones but not in public domain sequences, which could be PCR induced mutations and/or new SNPs.