Clone and Vector Annotation Details

General Comments

- The more information you provide, the more information we can provide to other researchers. Please be as comprehensive as possible.
- Attached in an Excel spreadsheet containing forms to help you fill out information about the Vectors and Clones (vector + inserts) that you are submitting.

File Details

- We have a "comments" field in the database that you can use to put any additional information about the clone. For example, if you would like to include the quantification for the protein yields for individual clones, information about small scale vs. large scale testing or details about solubility, you are welcome to use the comments field. This field is not searchable, but users will be able to see it on the website for each clone in order to get more detailed information about the clone.
- Anything with multiples should be separated with semicolons.

Key for Fields in Clone and Vector Data:

Column Header
Required
Required if applicable
Optional

Clone Data

CloneID		
Column Header	*Description*	*Example*
UniqueCloneID	Enter the unique identifier for the clone. This ID will be used as the main cross referencing ID for this clone	U1914HC100-12, ccsbBroad304_050

Container		
PlateLabel	Enter the label on the plate or tube containing the clone	U397PFC030-plate1, U452KFE220_7
PlateWell	Enter the well location. This is a plate specific field. Leave empty for tubes. A01 to H12 format	B02, E11

Vector		
Vector	Enter the vector name here and corresponding information on the Vector_Data sheet. Provide annotation for each vector as a new entry in Vector_Data	pLenti6.3/V5-DEST, pDONR221

CloneInsert		
InsertNTSeq	Enter the actual physical nucleotide sequence of the clone insert, including the STOP codon if present. This field is NOT for the reference sequence	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTG
CloningFormat	Enter the Cloning Format (CLOSED or FUSION). If a STOP codon is present in the insert sequence, then this is a CLOSED format. If a STOP codon is not present, then this is a FUSION format	CLOSED or FUSION

Linker		
5pLinkerSequence	If there is an extra sequence before the insert sequence listed in Column E that is not included in the vector sequence, please enter this sequence here. Please see the Definitions for annotating CDS sequences info below for more information about linkers. The linker sequences are typically about 8 -30 nucleotides that immediately flank the target sequence. This sequence would be the same for a family of clones that were constructed using the same cloning strategy/vector. Linker sequences will be analyzed during sequencing, *but only at the nucleotide level*. Whereas sequence evaluation of the **relevant CDS** includes both nucleotide and amino acid level evaluation	acctttcttctttcaacgggg
5pLinkerName	If applicable, please enter the name associated with the 5' Linker Sequence	5p pMCSG28
3pLinkerSequence	See above Note that the precise definition of 3p_Linkers depends on whether the clones are in the Closed or Fusion Format	ggggccttgagaaagtccgcg Please see the Definitions for annotating CDS sequences info below for precise definitions of linker sequences
3pLinkerName	If applicable, please enter the name associated with the 5' Linker Sequence	3p pMCSG28

CloneAnnotation		
MutationsNT	Enter any nucleotide discrepancies from the reference sequence (e.g. intended mutations, SNPs, allelic variation). Semicolon-separated list of expected mutations, deletions and insertions; this includes all mutations compared to the wildtype sequence; REQUIRED field if there are known mutations	Preferred format: “g3t; del@56, 20; ins@89, 32” for single nt change g to t at position 3; and deletion where nt 55 is present and 56 is not there and the deletion size is 20 bp; and insertion at position 89 where 89 is there like normal wild-type and after that a 32 bp insertion is present. If you use a different format, please explain it. The numbering of the nucleotide mutation should start from the CDS start (where the A or the ATG would equal 1)
MutationsAA	Enter any amino acid discrepancies from the reference sequence (e.g. intended mutations, SNPs, allelic variation). Semicolon-separated list of expected mutations, deletions and insertions	
InsertSource	Enter the physical source from which the insert was cloned or amplified	LaBaer lab, Gene synthesis by GenScript
IsCodonOptimized	Enter (yes/no) if this sequence was codon optimized	Yes or No
PurposeOfCodonOptimization	If yes, select the purpose for codon optimization	For gene expression, for gene synthesis
CodonOptimizedTo	If the purpose is for gene expression, enter the NCBI species name (genus species) of the organism codon optimized to	*Drosophila melanogaster*, *Vibrio cholerae* 01 biovar eltor
CloneNucleotideAccession	Enter the NCBI GenBank Nucleotide Accession number for the clone (if submitted)	NM_015974.3

	The nucleotide number is preferred but is not available for all organisms	
CloneProteinAccession	Enter the NCBI GenBank Protein Accession number for the clone (if submitted). This field is preferred but is not available for all organisms	NP_001341569.1

InsertAnnotation		
GenusSpecies	Genus species (plus strain, serovar, etc. if applicable). Please use the FULL NAME of species. No abbreviations	*Drosophila melanogaster*, *Vibrio cholerae* 01 biovar eltor
LocusTag	Enter the NCBI CDS Locus Tag	CIMG_00911
TaxonomyID	Enter the NCBI taxon (https://www.ncbi.nlm.nih.gov/taxonomy), if available	9606 for *Homo sapiens*, 246410 for *Coccidioides immitis RS*, etc.
CloneType	Enter the clone type	cDNA, shRNA, synthetic construct
GeneSymbol	Enter the official Entrez NCBI gene symbol. Include this column OR the GeneID.	TP53
GeneID	Enter the Entrez (NCBI gene ID. Include this column OR the GeneSymbol	3043, 4567420
GeneSynonym	Enter any aliases or gene synonyms	For HBB, ECYT6; CD113t-C; beta-globin

Gene Description	If applicable, enter the best available description of the gene product. This will help users know what kind of protein this is.	Phosphatase, Cdk-activating kinase 1At (cak1At), pyrophosphate-dependent phosphofructo-1-kinase-like protein
Clone Description	If applicable, enter the best available description associated with this clone	
InsertNucleotideAccession	Enter the transcript or genome NCBI RefSeq or GenBank Accession number for the reference sequence.	NM_032517.6
InsertProteinAccession	Enter the NCBI RefSeq or GenBank Protein Accession number for the reference sequence.	NP_060585.2
ProteinName	If different from the reference gene ID/symbol, enter the protein name. This would be the case for some viral polyproteins.	nsp2 protein in CoV-2 ORF1ab gene
ProteinDescription	If applicable, enter a description for the protein	

Authors		
Authors	Enter the associated author for this clone	Arizona State University, John Doe

Publications		
AssociatedPublications	Enter any publications associated with this clone	PMID: 16512675 Title: Functional proteomics approach to investigate the biological activities of cDNAs implicated in breast cancer.

Vector Data

Vector_Data		
Column Header	*Description*	*Example*
VectorName	Enter the name of the empty plasmid vector or vector with insert	pLenti6.3/V5-DEST
VectorUse	Enter the primary use of this plasmid	Mammalian expression, bacterial expression
SelectionMarker	Enter the antibiotic selection marker(s)	Ampicillin, kanamycin and chloramphenicol
CloningMethod	Enter the cloning method associated with this vector	Gateway Cloning, InFusion, Restriction Enzyme
5pForwardPrimerName	Enter the name of the 5' forward sequencing primer	M13F, CMV forward
5pForwardPrimerSequence	Enter the sequence for the 5' forward sequencing primer	GTAAAACGACGGCCAGT for M13F
3pReversePrimerName	Enter the name of the 3' reverse sequencing primer	M13R, Nap138R
3pReversePrimerSequence	Enter the sequence for the 3' reverse sequencing primer	CAGGAAACAGCTATGACC for M13R
Authors	Enter the author or company source for this vector	Arizona State University, John Doe or ThermoFisher
VectorSequence	Enter the sequence for this vector	CTTTCCTGCGTTATCCCCTGATTCT GTGGATAACCGTATTACCGCT…

MapFileName	Enter the name of the annotated map file	GenBank or SnapGene file
SequenceFileName	Enter the name of the provided sequence file	fasta or text file containing the sequence of the empty vector
GrowthTemp	Enter the optimal growth temperature for bacterial growth using this vector	37 degrees centigrade
GrowthStrain	Enter the growth strain of the bacterial host	DH5alpha, survival cells
Significant Features*	***Fill out the Vector_Features sheet if an annotated map file is not provided	See **Vector_Features** sheet
VectorDescription	If applicable, enter any additional description associated with this vector	
EmptyVector	Indicate whether the vector is empty or not	Yes or No

Vector_Features*		
Column Header	*Description*	*Example*
VectorName	Enter the name of the vector entered in Column A of the **Vector_Data** sheet	pLenti6.3/V5-DEST
FeatureType	Feature type refers to the general category that a vector feature belongs to	gene insert, selectable marker, protease cleavage site, recombination site, etc
FeatureName	Feature name refers to the specific feature of the vector you are describing	AmpR for selectable marker, TEV for protease cleavage site
FeatureDescription	Enter a description for the feature	ampicillin resistance for AmpR, TEV protease cleavage site for TEV
StartPosition	Enter the start position for the feature. This is a numeric value	
EndPosition	Enter the end position for the feature. This is a numeric value	

Directionality	Indicate the directionality of the feature as 0, 1, or -1. Nondirectional is 0. Forward is 1. Reverse is -1	0, 1, -1

CDS and Linker Definitions

These definitions are intended to help you fill in the correct information for linker sequences in the '**Clone Information File**'. The automated sequencing software that we use relies on precise definitions of the CDS and linker sequences. Without a clear definition of the *expected sequence*, it is impossible to determine if the sequence is correct.

1. Defining the Linker Sequences

In the context of this analysis, "Linkers" refers to nucleotide sequences that flank the **relevant CDS** that will be evaluated on the nucleotide level but not at the amino acid level. From a molecular biology perspective, these are often thought of as "junction sequences". Some investigators wish to confirm flanking nucleotide sequences that might have been accidentally altered during the cloning process (e.g. PCR primer). For example, sequencing would be advised to detect possible mutations due to PCR errors in the 5' sequence of a Gateway cloning vector, because such mutations could insert 5' stop codons or prevent subsequent Gateway cloning reactions. Any sequences for which the user wants/needs the *amino acids* to be analyzed should be included as part of the **relevant CDS** sequence.

Linker sequences are typically between 6 and 40 bases. If there are no sequences that flank the **relevant CDS** that need to be analyzed at the nucleotide level, it is sufficient to indicate "N/A". It is also worth noting that any sequences outside of the linker sequences will be masked out and not analyzed.

5' Linker – any sequences upstream of the **relevant CDS** for which the user needs nucleotide (but not amino acid) analysis. The last nucleotide of the 5' linker should be the nucleotide that immediately precedes the CDS Start.
3' Linker – any sequences downstream of the **relevant CDS** for which the user needs nucleotide (but not amino acid) analysis. The first base of the 3' linker must be the base immediately following the last base of the last codon of the gene of interest for the fusion format or the last base of the relevant STOP codon for "closed" format.