phyloTAGs Pipeline Manual (Version 1, August 1, 2015)

Alejandro Caro-Quintero & Howard Ochman

1. General description

The *phyloTAGs* pipeline is a set of PERL scripts that assist in the selection of low-degeneracy primers used to amplify single-copy genes from bacteria representing a set of taxonomic groups, selected according to their evolutionary relatedness (*i.e,* their level of 16s rDNA divergence). The scripts expedite the process of alignment, search and assessment of polymorphic regions within the targeted protein-coding gene gene. The *phyloTAGs* pipeline is composed of the following scripts: (1) **01_phyloTAGs_align.pl** and (2) **02_phyloTAGs_primer.pl**, which, respectively, align genes by codons and design primers along the alignment.

Input files: As examples of the primer design strategy, we provide two sets of orthologous single-copy gene sequences in FASTA format, *gyrB* and *recA* (*e.g.,* **gyrB_genes.fas**), which have been extracted from the completed genomes database at NCBI. A file containing pairwise 16S rDNA identities of the same set of genomes is also provided as a measurement of evolutionary relatedness (*i.e.,* **16S_rDNA_id.txt).** Additionally, a text file matches identifiers to species and genus names (*i.e.,* **NC_names.txt**). All files use the RefSeq NC numbers (complete genomic molecules) as organism/genome identifiers.

Group targeting and reference selection: Prior to running the scripts, the reference organism and 16S rDNA % identity range are selected. A RefSeq NC number is used to specify the representative reference organism from the taxonomic group of interest, and the 16S rDNA identity range delineates the degree of 16S divergence of the taxonomic group of interest. Examples of identity ranges for targeting specific taxonomic ranks are as follows:

- For bacterial species: 97 to 99 16S rDNA gene identity (%).
- For bacterial genera: 95 to 99 16S rDNA gene identity (%).
- For bacterial families: 90 to 99 16S rDNA gene identity (%).

2. Scripts and parameters

00_phyloTAGs_NC.pl description**:** This script facilitates the identification of the NC numbers, it takes as a parameter the genus and/or species that will be used as a reference and retrieves all possible NC numbers associated to the specified parameters. It can also print the organisms and identity of the taxonomic group that will be target for primer designing. To do this, the 16S rDNA identity range has to be specified. This information can be used to established, (1) if the required reference is in the database, (2) the number of related organisms within the targeted taxonomic range and (3) their divergence with respect to the reference. All this information can be used to establish the level of representation of the targeted taxonomic group and the possible biases on the primer design.

Parameters and descriptions:

-genus	(reference genus, required)
-species	(reference species, optional)

01_phyloTAGs_align.pl description**:** This script extracts a subset of orthologous gene sequences from a FASTA file (*e.g., gyrB* genes from complete genomes) using the user-specified reference organism (NC number) and 16S rDNA identity range, and performs a codon-by-codon alignment of the extracted sequences. This script conceptually translates gene sequences into protein sequences, which are aligned using ClustalW (Thompson *et al.,* 1994, Larkin *et al.,* 2007). With this protein-sequence alignment as a guide, the corresponding nucleotide sequences are aligned codon-by-codon with PAL2NAL (Suyama *et al.,* 2006). [Other distance measurements, such as the Average Amino-acid Identity (AAI), can also be applied if the files are submitted in the same format.] If the initial extraction step is not required (*e.g.,* in cases where the orthologues have been previously selected by the user), the "no_distance" option can be used, in which case, only the alignment process is performed.

Parameters and descriptions:

-in	(file with sequences in fasta format for primer designing)
-no_distance	(option for alignment by codons with no sequence selection using divergence)
-distance	(file with the pairwise relatedness distance)
-reference	(identifier of reference sequence, e.g., NC_012345)
-idlow	(lower value of pairwise distance to be evaluated)
-idhigh	(higher value of pairwise distance to be evaluated)
-help	(prints help)

02_phyloTAGs_primer.pl description: This script retrieves every combination of potential forward and reverse primers of a specified length and calculates the total number of degeneracies per primer. As input, this script takes a multiple sequence alignment in FASTA format, and the input alignment is searched in short blocks of a specified window size, using a sliding window of one nucleotide, to identify regions of the gene suitable for designing low degeneracy primers. To locate such regions, a consensus nucleotide is assigned to each nucleotide position using the following criteria: (*i*) when one of the four nucleobases is present in at least 80% of the aligned sequences (note that this stringency threshold can be modified), we assign that nucleobase as the consensus; (*ii*) for sites with higher levels of polymorphism, we combined all nucleobases occurring at frequencies over 20% and assigned a single-letter nucleotide in accordance with the standard degeneracy code (Table 1). For each window of specified length, the script calculates the total number of degeneracies. Results are saved in a tab-delimited output file “phyloTAGs_primers_table.txt”. (An example of this output file is presented below as Table 2).

Parameters and descriptions:

-in	(file with multiple sequence alignment)
-format	(file in "fasta" or "olt" format, “olt” by default)
-codons	(primer length in codons, e.g., 7 codons=21 nucleotides [7*3])
-slide	(sliding window size in bp)
-consensus	(selection of percentage for calculating consensus nucleotide in each position)
-greedy	(explores all windows of primer length)
-help	(prints help)

3. Dependencies:

The phyloTAGs pipeline relies on **ClustalW** and **PAL2NAL** for the codon alignment. The program ClustalW can be automatically installed in Debian/Ubuntu using apt-get from the command line as follows:

```
> sudo apt-get update
> sudo apt-get install clustalw
```

Or download and install the appropriate version for your operating system available at http://www.clustal.org/clustal2/. If after installation the programs require execution from a specific folder, please modify the ClustalW path (at line 54 in the 01_phyloTAGS_align.pl) and add the path to the file where the executable is installed. PAL2NAL can be downloaded from http://www.bork.embl.de/pal2nal/, and the extracted folder needs to be saved and extracted within the “phyloTAGS_pipeline” folder.

```
>perl  00_phyloTAGs_NC.pl  –genus Escherichia –species coli –NC NC_names.txt
```

4. Using the *phyloTAGs* pipeline: An example

NC number identification. The script will print a list of identifiers and will ask if the user wants to find the NC identifiers of related organisms, if “yes” is specify then the program will ask for extra information such as the desired NC reference and the ranges of 16S rDNA identity.

```
##############################
 List of identifiers
##############################

NC_000913   Escherichia_coli_K_12_substr__MG1655_uid57779
NC_002655   Escherichia_coli_O157_H7_EDL933_uid57831
NC_002695   Escherichia_coli_O157_H7_uid57781
..
Do you want to find the NC and identities associated to the reference? (y/n):y

1. Please specified which identifier you want to use : NC_000913
2. Please specified the distance file ? : 16S_rDNA_id.txt
3. Lower identity range? :99
4. Higher identity range? :100
5. Output file :output.txt
```

A few lines from the output.txt are presented below: column 1, the NC number of the reference organism; column 2, the NC number of a related organism within the identity range; column 3, the % 16S rDNA sequence identity value between the reference and related organisms; column 4, the name of the reference organism; column 5, the name of the related organism.

NC_000913 NC_017626 99.6 Escherichia_coli_K_12_substr__MG1655_uid57779 Escherichia_coli_042_uid161985

NC_000913 NC_008253 99.6 Escherichia_coli_K_12_substr__MG1655_uid57779 Escherichia_coli_536_uid58531

NC_000913 NC_011748 99.5 Escherichia_coli_K_12_substr__MG1655_uid57779 Escherichia_coli_55989_uid59383

Extraction and alignment by codons.

```
>perl 01_phyloTAGs_align.pl -in gyrB_all_genomes.fas –distance 16SrDNA_identity.txt
–reference NC_012345 –idlow 95 –idhigh 99
```

Primer search.

```
>perl 02_phyloTAGs_primer.pl –in alignment.fas -format fasta -codons 7 -consensus 90 -greedy
```

The script generates the file “phyloTAGs_primers_table.txt”, which can be opened as a spreadsheet (Figure 1) to identify conserved and polymorphic regions, and to select forward and reverse primers.

	A	B	C	D	E
1	##-- phyloTAGs run on 08/10/2015 --				
2	## Authors: Caro-Quintero A. and Ochman H. 2015				
3					
4	Position_window_(bp)	Degeneracies	Bases_per_position	Forward_primer	Reverse_primer
5	165_185	768	(GTCA)(G)(T)(TGC)(G)(A)(CT)(A)(A)(C)(TG)(C)(GC)(AG)(T)(ACTG)(G)(A)(C)(G)(A)	NGTBGAYAACKCSRTNGACGA	TCGTCNAYSGMGTTRTCVACN
6	166_186	192	(G)(T)(GTC)(G)(A)(CT)(A)(A)(C)(TG)(C)(CG)(AG)(T)(GTAC)(G)(A)(C)(G)(A)(G)	GTBGAYAACKCSRTNGACGAG	CTCGTCNAYSGMGTTRTCVAC
7	167_187	192	(T)(TGC)(G)(A)(CT)(A)(A)(C)(TG)(C)(GC)(AG)(T)(GTAC)(G)(A)(C)(G)(A)(G)(G)	TBGAYAACKCSRTNGACGAGG	CCTCGTCNAYSGMGTTRTCVA
8	168_188	192	(CGT)(G)(A)(CT)(A)(A)(C)(GT)(C)(CG)(GA)(T)(GTCA)(G)(A)(C)(G)(A)(G)(G)(C)	BGAYAACKCSRTNGACGAGGC	GCCTCGTCNAYSGMGTTRTCV
9	170_190	576	(A)(TC)(A)(A)(C)(TG)(C)(GC)(AG)(T)(ACGT)(G)(A)(C)(G)(A)(G)(G)(C)(ACG)(TAC)	AYAACKCSRTNGACGAGGCVH	DBGCCTCGTCNAYSGMGTTRT
10	171_191	576	(TC)(A)(A)(C)(TG)(C)(GC)(GA)(T)(TGAC)(G)(A)(C)(G)(A)(G)(G)(C)(GAC)(CAT)(T)	YAACKCSRTNGACGAGGCVHT	ADBGCCTCGTCNAYSGMGTTR
11	172_192	288	(A)(A)(C)(GT)(C)(GC)(AG)(T)(CATG)(G)(A)(C)(G)(A)(G)(G)(C)(GCA)(TAC)(T)(G)	AACKCSRTNGACGAGGCVHTG	CADBGCCTCGTCNAYSGMGTT
12	173_193	288	(A)(C)(GT)(C)(GC)(AG)(T)(ACTG)(G)(A)(C)(G)(A)(G)(G)(C)(ACG)(CAT)(T)(G)(G)	ACKCSRTNGACGAGGCVHTGG	CCADBGCCTCGTCNAYSGMGT
13	174_194	288	(C)(TG)(C)(GC)(AG)(T)(ACTG)(G)(A)(C)(G)(A)(G)(G)(C)(CAG)(TAC)(T)(G)(G)(C)	CKCSRTNGACGAGGCVHTGGC	GCCADBGCCTCGTCNAYSGMG
14	175_195	864	(TG)(C)(GC)(GA)(T)(GTAC)(G)(A)(C)(G)(A)(G)(G)(C)(CAG)(CAT)(T)(G)(G)(C)(GTC)	KCSRTNGACGAGGCVHTGGCB	VGCCADBGCCTCGTCNAYSGM
15	176_196	432	(C)(GC)(GA)(T)(CAGT)(G)(A)(C)(G)(A)(G)(G)(C)(CAG)(CAT)(T)(G)(G)(C)(TGC)(G)	CSRTNGACGAGGCVHTGGCBG	CVGCCADBGCCTCGTCNAYSG
16	177_197	432	(GC)(GA)(T)(TGAC)(G)(A)(C)(G)(A)(G)(G)(C)(ACG)(TCA)(T)(G)(G)(C)(CTG)(G)(G)	SRTNGACGAGGCVHTGGCBGG	CCVGCCADBGCCTCGTCNAYS
17	178_198	864	(AG)(T)(CAGT)(G)(A)(C)(G)(A)(G)(G)(C)(ACG)(ACT)(T)(G)(G)(C)(CGT)(G)(G)(TGCA)	RTNGACGAGGCVHTGGCBGGN	NCCVGCCADBGCCTCGTCNAY
18	179_199	864	(T)(GTCA)(G)(A)(C)(G)(A)(G)(G)(C)(GAC)(CAT)(T)(G)(G)(C)(CGT)(G)(G)(TGAC)(CT)	TNGACGAGGCVHTGGCBGGNY	RNCCVGCCADBGCCTCGTCNA
19	180_200	1728	(GTAC)(G)(A)(C)(G)(A)(G)(G)(C)(GCA)(TAC)(T)(G)(G)(C)(TGC)(G)(G)(GTAC)(CT)(GA)	NGACGAGGCVHTGGCBGGNYR	YRNCCVGCCADBGCCTCGTCN
20	181_201	864	(G)(A)(C)(G)(A)(G)(G)(C)(GAC)(CAT)(T)(G)(G)(C)(GTC)(G)(G)(TGCA)(CT)(GA)(TC)	GACGAGGCVHTGGCBGGNYRY	RYRNCCVGCCADBGCCTCGTC
21	182_202	1728	(A)(C)(G)(A)(G)(G)(C)(GAC)(CAT)(T)(G)(G)(C)(TGC)(G)(G)(ACTG)(TC)(GA)(TC)(TG)	ACGAGGCVHTGGCBGGNYRYK	MRYRNCCVGCCADBGCCTCGT
22	183_203	3456	(C)(G)(A)(G)(G)(C)(GAC)(ACT)(T)(G)(G)(C)(TGC)(G)(G)(CATG)(TC)(AG)(CT)(GT)(CG)	CGAGGCVHTGGCBGGNYRYKS	SMRYRNCCVGCCADBGCCTCG
23	184_204	10368	(G)(A)(G)(G)(C)(GCA)(ACT)(T)(G)(G)(C)(TGC)(G)(G)(ACGT)(TC)(GA)(CT)(TG)(GC)(GTC)	GAGGCVHTGGCBGGNYRYKSB	VSMRYRNCCVGCCADBGCCTC
24	185_205	20736	(A)(G)(G)(C)(CAG)(CAT)(T)(G)(G)(C)(CTG)(G)(G)(CAGT)(TC)(GA)(TC)(GT)(CG)(CTG)(AG)	AGGCVHTGGCBGGNYRYKSBR	YVSMRYRNCCVGCCADBGCCT
25	186_206	20736	(G)(G)(C)(GAC)(TAC)(T)(G)(G)(C)(GTC)(G)(G)(CAGT)(CT)(AG)(TC)(GT)(GC)(TGC)(AG)(A)	GGCVHTGGCBGGNYRYKSBRA	TYVSMRYRNCCVGCCADBGCC
26	187_207	20736	(G)(C)(GCA)(ACT)(T)(G)(G)(C)(CTG)(G)(G)(CATG)(TC)(AG)(TC)(TG)(CG)(CTG)(AG)(A)(C)	GCVHTGGCBGGNYRYKSBRAC	GTYVSMRYRNCCVGCCADBGC
27	188_208	20736	(C)(CAG)(CAT)(T)(G)(G)(C)(TGC)(G)(G)(CAGT)(CT)(GA)(TC)(GT)(GC)(GTC)(AG)(A)(C)(G)	CVHTGGCBGGNYRYKSBRACG	CGTYVSMRYRNCCVGCCADBG

Figure 1. Output file “phyloTAGs_primers_table.txt” shown as a spreadsheet.

With this spreadsheet, one can plot the number of degeneracies per position (in blocks that correspond to each primer length) *vs*. the multiple sequence alignment position (as plotted in Figure 2, whose axes correspond to columns A and B in Figure 1). With this information, it is possible to identify conserved regions that are flanked highly polymorphic ones, and therefore suitable for designing conserved primers that flank variable regions within the target gene. As an example, in Figure 2, positions marked A and B are suitable for designing primers that generate short *phyloTAG* amplicons (as appropriate for high-throughput sequencing platforms, *e.g*., Illumina), whereas positions marked A and C are suitable for designing primers that generate longer amplicons (as appropriate for Sanger sequencing applications).

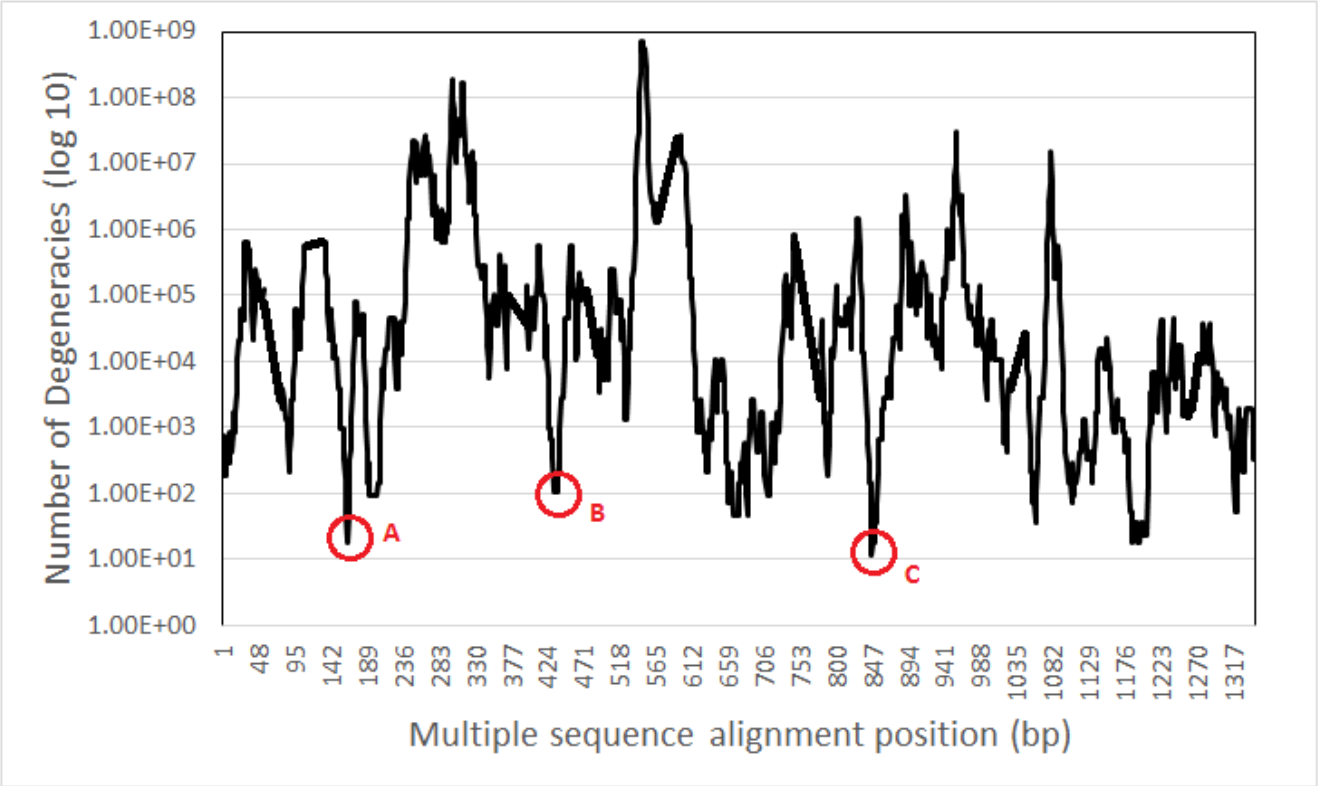

Figure 2. Total number of degeneracies per position in the multiple sequence alignment. A, B and C represent regions of 21 nucleotides that contain less than 100 degeneracies and that flank highly polymorphic regions.

Table 1. DNA Base degeneracy

IUPAC nucleotide code	Nucleobases
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	A or C or G or T

Table 2. Column description of the *phyloTAGs* file

Column	Description
1	Starting and final position in the alignment
2	Total number of degeneracies of the region
3	Diversity of bases per position
4	Forward primer with degenerated code
5	Reverse primer with degenerated code

5. References:

Suyama M, D Torrents, P Bork (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research 34, W609-W612.

Larkin MA, *et al.* (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.