AMPLICONprocessing: Pipeline for 16S rRNA datasets analysis

Umer Zeeshan Ijaz

The workflow of the pipeline is given on the next page. It is useful for analyzing 16s RNA amplicons to generate taxonomic level breakdowns. The pipeline has the following steps:

- Trim forward and reverse sequences
- Generate fastqc and SeqQA stastistics
- Overlap forward and reverse sequences (using PANDASEQ)
- De-replicate overlapped sequences while keeping a record of their abundances (using UCHIME)
- Remove chimeric sequences (using UCHIME)
- Search non-chimeric sequences through CREST software (`http://code.google.com/p/lcaclassifier/`) or RDP Classifier (`http://sourceforge.net/projects/rdp-classifier/files/rdp-classifier/`) to generate taxonomic assignments
- Collate the results together to generate frequency tables for PHYLUM, CLASS, ORDER, FAMILY and GENUS level assignments for all samples

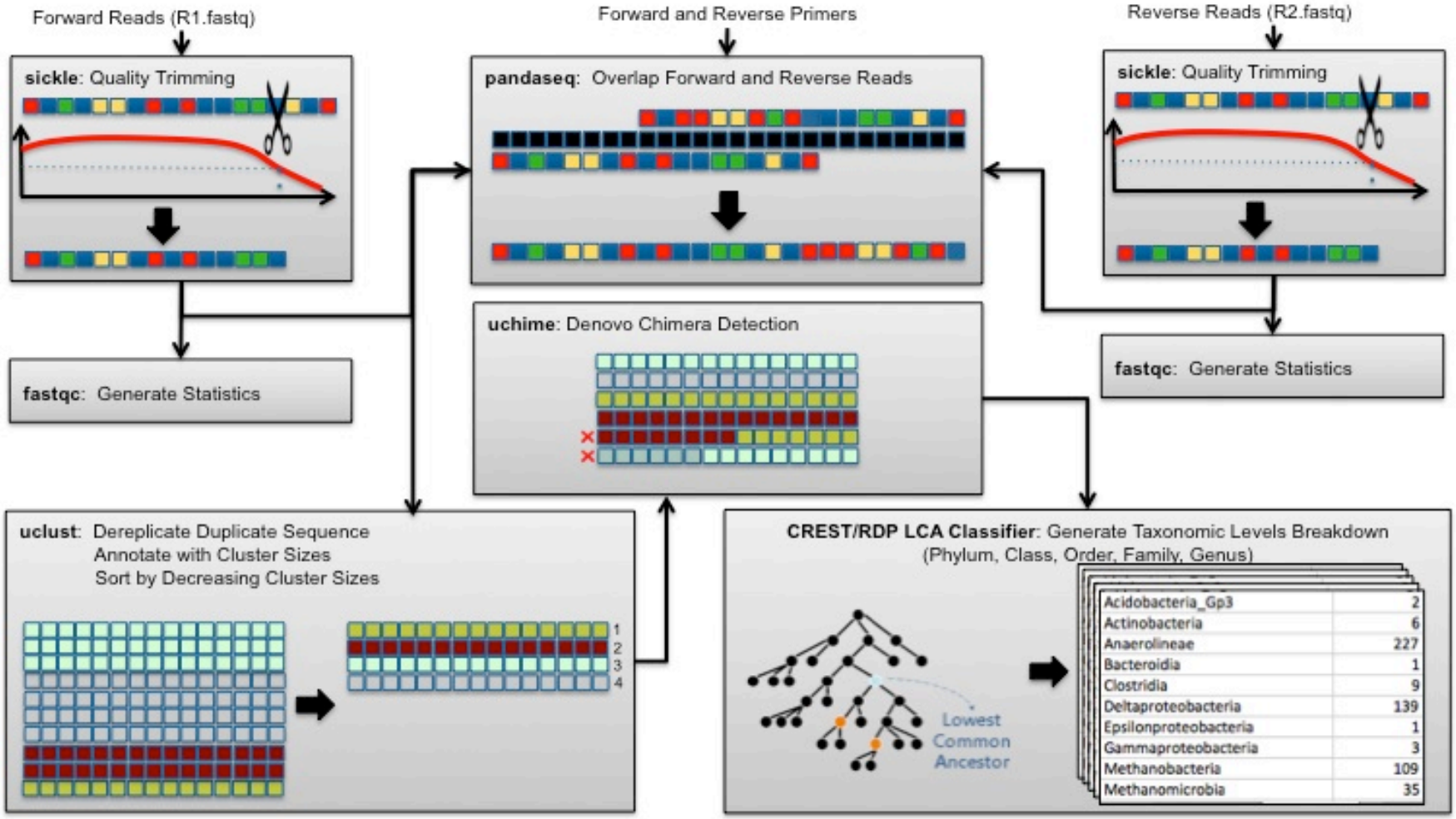
You can run the pipeline without any arguments to get the usage information:

```
[uzi@quince-srv2 ~]$ bash /home/opt/AMPLICONprocessing_v0.2/AMPLICONprocessing.sh
Script to analyse 16S rRNA datasets

Usage:
    bash AMPLICONprocessing.sh -f <forward.fastq> -r <reverse.fastq> [options]
Options:
    -n Flag to eliminate all sequences with unknown nucleotides in the output (pandaseq) default: off
    -p Forward primer if available
    -q Reverse primer if available
    -o Overlap base pairs between paired-end reads (default: 50)
    -t Type of quality values (solexa (CASAVA < 1.3), illumina (CASAVA 1.3 to 1.7), sanger (which is CASAVA >= 1.8)
       default: sanger
    -c Type of classifier (CREST,RDP) default: CREST
```

Before using the pipeline, make sure **sickle, usearch, pandaseq, fastqc**, **RDP & CREST Classifiers** are installed and available in $PATH. Furthermore, edit AMPLICONprocessing.sh, find the parameters section and set the directories for both RDP & CREST databases

```
# = Parameters to set ============== #
LOGFILE="`pwd`/AMPLICONprocessing.log" # Where to save the log
CREST_DATABASE_DIR="/home/opt/LCAClassifier/parts/flatdb/silvamod"
CREST_MEGABLAST_DIR="/home/opt/LCAClassifier/bin"
CREST_CLASSIFY_DIR="/home/opt/LCAClassifier/bin"
RDP_CLASSIFY_DIR="/home/opt/rdp_classifier_2.6/dist"
FORWARD_FASTQ=""
REVERSE_FASTQ=""
FORWARD_PRIMER=""
REVERSE_PRIMER=""
QUALITY_TYPE="sanger"
OVERLAP_BASEPAIRS=50
N_FLAG=0
CLASSIFIER_TYPE="CREST"
# =/Parameters to set ============== #
```



Workflow of **`AMPLICONprocessing`** v0.2

To process the data, say I have a PROJECT_BACKUP folder where I have paired-end reads for several samples with names starting with 1-1*.fastq, 2-1*.fastq and so on. To process 1-1*.fastq, we run the following steps

```
[uzi@quince-srv2 ~/PROJECT]$ mkdir 1-1
[uzi@quince-srv2 ~/PROJECT]$ cd 1-1
[uzi@quince-srv2 ~/PROJECT/1-1]$ cp /home/uzi/PROJECT_BACKUP/1-1* .
[uzi@quince-srv2 ~/PROJECT/1-1]$ gunzip *
[uzi@quince-srv2 ~/PROJECT/1-1]$ nohup bash /home/opt/AMPLICONprocessing_v0.1/AMPLICONprocessing.sh -f
*_R1_001.fastq -r *_R2_001.fastq &
[uzi@quince-srv2 ~/PROJECT/1-1]$ cat AMPLICONprocessing.log
[2013-07-31 08:41:37] AMPLICONprocessing v0.1. Copyright (c) 2013 Computational Microbial Genomics Group, University
of Glasgow, UK
[2013-07-31 08:41:37] Using  sickle
[2013-07-31 08:41:37] Using  usearch
[2013-07-31 08:41:37] Using  pandaseq
[2013-07-31 08:41:37] Using  fastqc
[2013-07-31 08:41:37] Using  /home/opt/AMPLICONprocessing_v0.2/scripts/Convert.py
[2013-07-31 08:41:37] Using  /home/opt/AMPLICONprocessing_v0.2/scripts/SeqQA.pl
[2013-07-31 08:41:37] STEP 1: Trimming fastq files.
[2013-07-31 08:41:38] 1-1_S1_L001_R1_001.fastq and 1-1_S1_L001_R2_001.fastq are not trimmed. Using sickle with
sanger as quality type, quality window of 20, and minimum length of 10 to keep
[2013-07-31 08:41:38] 1-1_S1_L001_R1_001.fastq, 1-1_S1_L001_R2_001.fastq, and 1-1_S1_L001_singlet.fastq generated
successfully!
[2013-07-31 08:41:38] STEP 2: Generating SeqQA statistics for 1-1_S1_L001_R1_trim_001.fastq.
[2013-07-31 08:41:52] 1-1_S1_L001_R1_trim_001.SeqQA generated successfully.
[2013-07-31 08:41:52] STEP 3: Generating SeqQA statistics for 1-1_S1_L001_R2_trim_001.fastq.
[2013-07-31 08:42:05] 1-1_S1_L001_R2_trim_001.SeqQA generated successfully.
[2013-07-31 08:42:05] STEP 4: Overlapping forward and reverse reads using pandaseq.
[2013-07-31 08:42:05] Forward and reverse primers not found! Running pandaseq without them.
[2013-07-31 08:42:07] 1-1_S1_L001_R12.fastq generated successfully.
[2013-07-31 08:42:07] STEP 5: Convert 1-1_S1_L001_R12.fastq into 1-1_S1_L001_R12.fasta
[2013-07-31 08:42:08] 1-1_S1_L001_R12.fasta and 1-1_S1_L001_R12.qual generated successfully.
[2013-07-31 08:42:08] STEP 6: Generating fastqc statistics.
[2013-07-31 08:42:12] 1-1_S1_L001_R1_trim_001_fastqc.zip generated successfully.
[2013-07-31 08:42:15] 1-1_S1_L001_R2_trim_001_fastqc.zip generated successfully.
```

```
[2013-07-31 08:42:18] 1-1_S1_L001_R12_fastqc.zip generated successfully.
[2013-07-31 08:42:18] STEP 7: Dereplicate duplicate sequences, annotate with cluster sizes and sort by decreasing
cluster sizes.
[2013-07-31 08:42:18] 1-1_S1_L001_R12.derep.fasta generated successfully.
[2013-07-31 08:42:18] STEP 8: Sort reads in order of decreasing abundance
[2013-07-31 08:42:18] 1-1_S1_L001_R12.sorted.derep.fasta generated successfully.
[2013-07-31 08:42:18] STEP 9: De novo chimera detection using the UCHIME algorithm
[2013-07-31 08:42:22] 1-1_S1_L001_R12.nonchim.sorted.derep.fasta and 1-1_S1_L001_R12.chim.sorted.derep.fasta
generated successfully.
[2013-07-31 08:42:22] STEP 10(a): Classify 1-1_S1_L001_R12.nonchim.sorted.derep.fasta using CREST - Run megablast
[2013-07-31 08:51:55] 1-1_S1_L001_R12.nonchim.sorted.derep.xml generated successfully.
[2013-07-31 08:51:55] STEP 10(b): Classify 1-1_S1_L001_R12.nonchim.sorted.derep.fasta using CREST - Run classify
[2013-07-31 08:55:23] 1-1_S1_L001_R12.nonchim.sorted.derep_Assignments.txt, 1-
1_S1_L001_R12.nonchim.sorted.derep_Assignments.fasta and 1-1_S1_L001_R12.nonchim.sorted.derep_Composition.txt
generated successfully.
[2013-07-31 08:55:23] STEP 11: Generate Taxonomic levels breakdown for 20348 non-chimeric sequences
[2013-07-31 08:55:23] 1-1_S1_L001_R12.nonchim_[PHYLUM/CLASS/ORDER/FAMILY/GENUS].csv generated successfully.
[2013-07-31 08:55:23] Finished processing!
```

Please note that you can use either of the classifiers by using -c switch. i.e. -c CREST or -c RDP
The only difference is in step 10 of the pipeline. For example, when processing CICRA sample 10_a through both CREST and RDP, the output is as follows

```
[uzi@quince-srv2 ~/PROJECT/10-1]$ bash ~/AMPLICONprocessing_v0.2/AMPLICONprocessing.sh -f 10-1_S10_L001_R1_001.fastq
-r 10-1_S10_L001_R2_001.fastq -o 50 -c CREST

[2013-10-21 13:59:07] STEP 10(a): Classify 10-1_S10_L001_R12.nonchim.sorted.derep.fasta using CREST - Run megablast
[2013-10-21 14:08:10] 10-1_S10_L001_R12.nonchim.sorted.derep.xml generated successfully.
[2013-10-21 14:08:10] STEP 10(b): Classify 10-1_S10_L001_R12.nonchim.sorted.derep.fasta using CREST - Run classify
[2013-10-21 14:11:15] 10-1_S10_L001_R12.nonchim.sorted.derep_Assignments.txt, 10-
1_S10_L001_R12.nonchim.sorted.derep_Assignments.fasta and 10-1_S10_L001_R12.nonchim.sorted.derep_Composition.txt
generated successfully.
[2013-10-21 14:11:15] STEP 11: Generate Taxonomic levels breakdown for 10077 non-chimeric sequences
[2013-10-21 14:11:15] 10-1_S10_L001_R12.nonchim_[PHYLUM/CLASS/ORDER/FAMILY/GENUS].csv generated successfully.

[uzi@quince-srv2 ~/PROJECT/10-1]$ bash ~/AMPLICONprocessing_v0.2/AMPLICONprocessing.sh -f 10-1_S10_L001_R1_001.fastq
```

```
-r 10-1_S10_L001_R2_001.fastq -o 50 -c RDP

[2013-10-21 14:13:46] STEP 10: Classify 10-1_S10_L001_R12.nonchim.sorted.derep.fasta using RDP
[2013-10-21 14:14:17] 10-1_S10_L001_R12.nonchim.sorted.derep_Assignments.txt generated successfully.
[2013-10-21 14:14:17] STEP 11: Generate Taxonomic levels breakdown for 10077 non-chimeric sequences
[2013-10-21 14:14:17] 10-1_S10_L001_R12.nonchim_[PHYLUM/CLASS/ORDER/FAMILY/GENUS].csv generated successfully.
[2013-10-21 14:14:17] Finished processing!
```

The generated taxonomic level breakdowns are as follows:

Phylum (RDP)

```
"Actinobacteria",99
"Bacteroidetes",12
Firmicutes,7275
"Proteobacteria",1390
unclassified_Bacteria,1267
__Unknown__,32
"Verrucomicrobia",2
```

Phylum (CREST)

```
Actinobacteria,100
Bacteroidetes,10
Firmicutes,8397
Proteobacteria,1390
Tenericutes,2
__Unknown__,176
Verrucomicrobia,2
```

Class (RDP)

```
Actinobacteria,99
Alphaproteobacteria,1
Bacilli,56
"Bacteroidia",7
Clostridia,7085
Erysipelotrichia,42
Flavobacteria,2
```

Gammaproteobacteria,1389
Negativicutes,12
unclassified_"Bacteroidetes",3
unclassified_Firmicutes,80
__Unknown__,1299
Verrucomicrobiae,2

Class (CREST)
Actinobacteria (class),9
Bacilli,56
Bacteroidia,8
Clostridia,8297
Coriobacteriia,91
Erysipelotrichi,44
Flavobacteria,2
Gammaproteobacteria,1390
Mollicutes,2
__Unknown__,176
Verrucomicrobiae,2

Order (RDP)
Actinomycetales,1
"Bacteroidales",7
Bifidobacteriales,7
Clostridiales,7084
Coriobacteriales,91
"Enterobacteriales",1386
Erysipelotrichales,42
"Flavobacteriales",2
Lactobacillales,56
Pseudomonadales,1
Rhizobiales,1
Selenomonadales,12
unclassified_Clostridia,1
unclassified_Gammaproteobacteria,2
__Unknown__,1382

Verrucomicrobiales,2

Order (CREST)
Actinomycetales,1
Bacteroidales,8
Bifidobacteriales,8
Clostridiales,8296
Coriobacteriales,91
Enterobacteriales,1388
Erysipelotrichales,44
Flavobacteriales,2
Lactobacillales,56
RF9,2
Unknown Clostridia order,1
__Unknown__,178
Verrucomicrobiales,2

Family (RDP)
Actinomycetaceae,1
Bacteroidaceae,3
Bifidobacteriaceae,7
Clostridiaceae 1,70
Coriobacteriaceae,91
Enterobacteriaceae,1386
Erysipelotrichaceae,42
Eubacteriaceae,1
Flavobacteriaceae,2
Hyphomicrobiaceae,1
Lachnospiraceae,5650
Peptostreptococcaceae,94
Pseudomonadaceae,1
"Rikenellaceae",4
Ruminococcaceae,1198
Streptococcaceae,54
unclassified_Clostridiales,71
unclassified_Lactobacillales,2
__Unknown__,1385

Veillonellaceae,12
Verrucomicrobiaceae,2

Family (CREST)
Actinomycetaceae (Actinomycetales),1
Akkermansiaceae,2
Bacteroidaceae,1
Bifidobacteriaceae,7
Clostridiaceae,69
Coriobacteriaceae,80
Enterobacteriaceae,1385
Enterococcaceae,1
Erysipelotrichaceae,42
Family XIII Incertae Sedis,2
Family XI Incertae Sedis (Clostridiales),2
Flavobacteriaceae,2
Lachnospiraceae,5996
Peptostreptococcaceae (Clostridiales),94
Rikenellaceae,7
Ruminococcaceae,2100
Streptococcaceae,54
Unknown Bifidobacteriales family,1
Unknown Clostridiales family,17
Unknown Coriobacteriales family,11
Unknown Enterobacteriales family,3
Unknown Erysipelotrichales family,2
__Unknown__,186
Veillonellaceae,12

Genus (RDP)
Actinomyces,1
Akkermansia,2
Alistipes,4
Alloscardovia,1
Anaerorhabdus,2
Anaerostipes,4

Bacteroides,1
Bifidobacterium,6
Blautia,1074
Butyricicoccus,8
Clostridium IV,11
Clostridium sensu stricto,67
Clostridium XI,94
Clostridium XlVa,15
Clostridium XVIII,26
Collinsella,2
Coprobacillus,6
Coprococcus,8
Dorea,2476
Eggerthella,60
Erysipelotrichaceae_incertae_sedis,6
Escherichia/Shigella,143
Eubacterium,1
Faecalibacterium,11
Flavobacterium,2
Gemmiger,1
Gordonibacter,26
Holdemania,2
Lachnospiracea_incertae_sedis,782
Lactococcus,2
Lactonifactor,6
Pseudomonas,1
Roseburia,202
Ruminococcus,467
Streptococcus,52
Turicibacter,2
unclassified_Clostridiaceae 1,3
unclassified_Coriobacteriaceae,3
unclassified_Enterobacteriaceae,1243
unclassified_Lachnospiraceae,1083
unclassified_Ruminococcaceae,701
__Unknown__,1458
Veillonella,12

Genus (CREST)
Actinomyces,1
Akkermansia,2
Alistipes,7
Alloscardovia,1
Anaerostipes,4
Anaerotruncus,2
Bacteroides,1
Bifidobacterium,6
Blautia,964
Catabacter,1
Clostridium (Clostridiaceae),69
Coprococcus,13
Dorea,2
Eggerthella,49
Epulopiscium,17
Erysipelotrichaceae Incertae Sedis,28
Escherichia-Shigella,1374
Family XIII Incertae Sedis Incertae Sedis,1
Flavobacterium,2
Gordonibacter,26
Holdemania,2
Lachnospiraceae Incertae Sedis,1242
Lactococcus,2
Marvinbryantia,1
Peptoniphilus,2
Peptostreptococcaceae Incertae Sedis,94
Pseudobutyrivibrio,175
Roseburia,41
Ruminococcaceae Incertae Sedis,244
Ruminococcus,485
Streptococcus,52
Subdoligranulum,1205
Turicibacter,2
Unknown Enterococcaceae genus,1
Unknown Erysipelotrichaceae genus,1

```
Unknown Lachnospiraceae genus,9
Unknown Ruminococcaceae genus,11
__Unknown__,3926
Veillonella,12
```

Once you have generated the results in all the folders, you can collate them together. Say you have processed all your samples, and you have generated the following subfolders each containing data generated by the pipeline:

```
[uzi@quince-srv2 ~/PROJECT/1-1]$ cd ..
[uzi@quince-srv2 ~/PROJECT]$ ls
100-2  106-2  111-2  118-2  125-2  131-2  138-2  144-2  15-1   157-2  163-2  22-1  29-1  35-1  41-1  48-1  54-1  6-1
67-1  73-1  80-1  86-1  92-1  99-2
10-1   107-2  112-2  119-2  126-2  132-2  139-2  145-2  151-2  158-2  17-1   23    30-1  36-1  42-1  49-1  55-1  61-
1  68-1  74-1  8-1   87-1  93-1
101-2  108-2  113-2  120-2  127-2  133-2  140-2  146-2  152-2  159-2  18-1   24-1  3-1   37-1  43-1  50-1  56-1  62-
1  69-1  75-1  81-1  88-1  94
102-2  109-2  114-2  12-1   128-2  134-2  14-1   147-2  153-2  160-2  19-1   25-1  31-1  38-1  44-1  5-1   57-1  63-
1  70-1  76-1  82-1  89-1  95
103-2  1-1    115-2  121-2  129-2  135-2  141-2  148-2  154-2  16-1   20-1   26-1  32-1  39-1  45-1  51-1  58-1  64-
1  7-1   77-1  83-1  90-1  96
104-2  110-2  116-2  122-2  130-2  136-2  142-2  149-2  155-2  161-2  2-1    27-1  33-1  40-1  46-1  52-1  59-1  65-
1  71-1  78-1  84-1  9-1   97-2
105-2  11-1   117-2  124-2  13-1   137-2  143-2  150-2  156-2  162-2  21-1   28-1  34-1  4-1   47-1  53-1  60-1  66-
1  72-1  79-1  85-1  91-1  98-2
```

To, generate the frequency table for phylum level assignments, we will run the collateResults.pl script on these folders

```
[uzi@quince-srv2 ~/PROJECT]$ perl /home/opt/AMPLICONprocessing_v0.2/scripts/collateResults.pl -f ~/PROJECT -p
_PHYLUM.csv
Samples,52-1,24-1,19-1,139-2,150-2,140-2,23,141-2,132-2,43-1,7-1,156-2,42-1,147-2,25-1,74-1,8-1,86-1,60-1,51-1,49-
1,71-1,159-2,108-2,2-1,81-1,36-1,73-1,68-1,32-1,45-1,9-1,118-2,38-1,46-1,129-2,70-1,17-1,104-2,50-1,82-1,130-2,3-
1,97-2,163-2,31-1,41-1,113-2,91-1,11-1,145-2,133-2,55-1,117-2,63-1,85-1,103-2,83-1,135-2,62-1,75-1,109-2,119-2,88-
1,157-2,101-2,107-2,158-2,28-1,149-2,138-2,162-2,110-2,96,98-2,21-1,37-1,10-1,146-2,99-2,26-1,148-2,121-2,72-1,125-
2,44-1,142-2,160-2,116-2,93-1,127-2,18-1,39-1,131-2,64-1,106-2,84-1,58-1,155-2,61-1,13-1,69-1,15-1,153-2,66-1,53-
```

1,90-1,102-2,47-1,30-1,77-1,22-1,100-2,95,89-1,48-1,1-1,6-1,120-2,14-1,136-2,114-2,151-2,124-2,137-2,112-2,78-1,128-
2,4-1,79-1,5-1,152-2,35-1,143-2,33-1,111-2,34-1,105-2,94,27-1,29-1,59-1,115-2,87-1,122-2,134-2,80-1,20-1,12-1,67-
1,76-1,54-1,161-2,144-2,154-2,126-2,65-1,56-1,57-1,16-1,92-1
Actinobacteria,392,0,4297,15510,313,8084,149,55,1942,1149,0,1045,3128,1401,9904,2785,173,13194,4066,2326,2511,8435,1
6643,1759,1321,1459,20187,5083,1081,1342,1061,212,16508,19,719,6387,146,1232,811,3869,1168,1485,906,0,1097,30,0,230,
90,1384,1483,4486,617,109338,212,1,46,500,1098,378,2201,1832,3223,1258,2503,1097,3563,5007,930,1985,1182,3824,1409,1
676,0,1006,2,100,434,6,9689,118,39,2542,322,2045,1463,4445,10984,957,31590,739,0,737,96,5670,683,4901,1390,3061,1594
7,27,8838,9529,2150,382,1025,56,16162,23,2943,398,13615,222,2913,10642,3328,2,1812,1353,665,2032,320,398,421,1164,10
42,6539,492,1633,485,1020,6907,5504,20266,835,2041,13316,2702,773,17,5294,1392,3936,6674,9112,1761,8315,306,36762,50
27,2171,4976,7856,903,516,1345,7977,2881,7463,56
Bacteroidetes,2,0,79,17,35,9842,2091,417,40,121,0,4346,481,154,131,19,3,685,49,2141,15,5,249,37,13,58,105,427,2438,4
,15,7,830,34,2293,672,1454,12,33,62,8,167,18,0,12,11,0,187,544,26,679,4146,3443,2214,58,0,4,1048,1686,869,92,10,531,
13,3858,27,95,307,38,1,25,29,286,22,2,12071,0,10,169,3,119,1144,6,50,38,5,104,151,97,18,75,423,0,3508,1,4132,90,190,
1994,87,65,420,118,181,3193,15,277,9,444,34,5443,1617,1777,22,49,434,23,0,12,2,26,29,7,14,64,9,9,368,29,139,14,177,6
,2526,12,29,0,500,34,2216,27,76,15,1057,1,147,409,125,81,609,8702,189,68,1783,2093,4,49,87,106,56,45
Firmicutes,3168,2,4569,11919,4909,53128,4584,1093,10293,1726,0,7613,13586,3018,36849,4174,15548,23459,2964,26508,612
3,4538,24652,2862,14921,6760,22954,5746,11003,982,3122,5005,43039,5343,7478,12360,2002,2013,22103,13640,7848,9880,17
424,4,7274,3068,8,1328,9501,27334,3053,31551,10929,78957,2161,0,10589,3949,14319,4258,5069,2816,7806,10633,18225,158
03,16848,10312,3835,5155,2032,10202,1601,6044,2,20531,14,8397,1586,26,9692,911,376,9122,10593,3281,4574,20590,19706,
3537,38902,7201,6,11830,189,10275,2131,16193,9281,10172,14816,168,26298,9576,19763,2686,12828,15441,93798,11227,1648
4,11145,46180,1459,15885,60204,15294,13,4044,966,2601,4558,7349,5372,2422,1551,10064,12228,8187,12512,5974,33223,205
97,50068,30870,3075,939,50904,10272,15600,3610,8263,952,19619,1757,18074,5089,43074,1371,63100,47945,22364,13970,234
46,13407,620,19631,22949,9818,32692,1199
Proteobacteria,5,2,20,29,305,1416,1818,65,18,170,0,74,475,14,65,59,151,2155,27,153,1635,23,12,16,11,10,117,46,804,12
1,159,43,70,4568,1472,69,864,8,19,12,3,52,11,0,11,3134,2,50,332,5880,11,69,69,2306,238,0,163,34,62,664,2,34,119,106,
354,5448,10,27,1033,11,17,23,7,8,0,9389,2,1390,6,2,223,10,510,22,8,3,12,18,35,19,33,92,1,94,9,22,40,25,92,39,34,149,
19,15,713,1,8173,85,725,1588,975,14095,52,25,204,417,15,0,8,1,51,6,679,7,54,1,864,31,5,137,26,19,1362,67,180,11,5,11
8,7,3600,3028,259,7,56,1,293,236,1296,34,242,3795,11,21,117,133,4,85,167,51,25,1529
__Unknown__,117,1,1190,1091,196,7670,487,231,423,261,2,682,389,484,786,936,378,534,130,2544,1818,473,1713,368,911,58
7,685,1723,1065,378,348,782,1468,311,886,383,775,1009,313,976,377,922,1902,3,640,357,4,31,1015,575,376,574,1897,7193
,193,2,424,226,606,854,393,471,1421,961,1535,1339,430,1474,391,201,2107,2035,267,939,0,761,3,176,279,42,1004,920,336
,767,650,369,2018,1968,2653,1790,1980,932,3,740,158,337,159,1884,289,218,559,370,252,1298,783,721,1794,621,1106,662,
587,275,1357,812,1451,838,1673,0,1134,66,1142,472,369,46,881,683,355,658,627,393,315,658,471,3024,1949,732,168,714,5
89,350,207,663,749,2113,100,938,188,1001,33,1544,1643,1219,531,1379,585,449,729,1157,2083,2022,952
Euryarchaeota,0,0,6,3,0,11152,0,24,2,455,0,0,0,1,1,4,0,3,0,9,7,1,778,0,1,113,0,4,2,1,0,1,2,2,1,2067,4,2,0,2,0,2,5,0,
12,0,0,0,2,1,0,1,8,21,1,0,1,16,2,2,1,1,4,5,1,67,0,4,0,2,52,7,2,256,0,2,0,0,0,0,3,1,4,4,4,1,7,231,7,7,5,1,0,2,0,0,318

,4,0,1,0,3,0,1,2,2,9,0,2,4,0,0,3278,1,2,1,4,0,2,0,5,1,1,8,1,1,0,2245,1,8,0,1,0,12,1,2,0,2,0,0,0,1,2,5,8,1,226,7,0,89
28,3,0,209,7,1500,1,1,1,8,5,1
Verrucomicrobia,0,0,506,12,1,31333,0,11,1,952,0,137,0,1,3,6,0,2,1,456,6,3,0,9,3,976,1,8,5,1558,1,3,4605,2,2,1661,4,3
,0,8,16,147,6,0,6,1,0,0,6,2,10,7,12,280,0,0,2,28,229,5,1,2,24,2930,18,65,1,14,2,2,11,218,5,53,0,4,0,2,1,0,2,4,3,77,1
,6,1,10,22,27,11,2,0,166,1,368,2343,9,378,152,5,4,1,6,1,2,12,1,1,2,1,3,438,2260,8,3,11,0,2432,0,8,116,1,1,12,125,171
,4313,1,65,0,1,927,18,2358,3,313,61,1,1,1,0,2226,9,1225,135,607,9,1,6858,7,4,73,11,1665,3,3,6,20,9,2
Cyanobacteria,0,0,0,0,0,34,0,0,34,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4,
0,0,0,0,0,0,0,0,10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,28,0,0,5,1,4,0,0,0,0,0,0
Tenericutes,0,0,0,0,0,74,0,0,0,5,0,3,0,4,0,0,0,0,0,0,0,0,82,0,0,0,0,34,0,0,0,0,0,0,0,50,0,0,0,204,1,50,0,0,2,0,0,0,0
,1,3,3,0,0,0,0,0,0,6,1,0,0,1,0,0,0,0,0,0,0,18,5,0,51,0,0,0,2,0,0,2,0,0,0,0,0,1,40,2,0,1,0,0,11,0,0,1,0,0,0,0,0,0,57,
0,0,0,0,0,0,0,0,41,0,0,0,0,0,0,0,1,0,0,0,0,1,2,17,0,23,0,0,0,16,1,1,0,3,0,0,0,0,1,0,4,6,1,3,0,167,1,0,26,33,11,0,0,0
,0,0,0
Fusobacteria,0,0,0,0,0,0,233,0,0,0,0,0,0,0,0,0,0,11,0,0,1,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,36,0,0,3,4
,0,1,0,2,0,0,0,0,0,7,0,0,0,0,4,0,0,0,0,0,0,0,0,0,0,174,0,0,0,0,0,267,61,0,1,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0,0,0,0,43
,0,0,0,0,0,3,14,6,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,1,0,10,55,0,0,0,0,1,0,90,0,6,654,0,0,1,0,0,4,0,0,0
,0
RF3,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,2,0,0,0,0,0,0
Mitochondrion,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,
2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Plastid,0,0,0,0,0,0,0,0,0,0,0,0,4,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,1,
0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,4,0,0,0,0,0,1,0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0,25,0,0,0,0,0,1,0,0,0,2,0,0,0,0,0,0
Synergistetes,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,403,0,0,2,0,0,0,0,0,0,0,0,
0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,14,0,0,1,0,0,0,291,0,0,0,0,0,0
Candidate division
TM7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,8,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0
Planctomycetes,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

```
Acidobacteria,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Aquificae,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Spirochaetes,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5,0,0,0,0,0,0,0,0,0,0
[uzi@quince-srv2 ~/PROJECT]$
```

This script collates the results produced by AMPLICONprocessing. Say you have a Main_Folder, where each subfolder contains a CSV file. Using -f path_to_Main_Folder -p _PHYLUM.csv, the contents of *_PHYLUM.csv file will be collated together.

BEFORE:

Main_Folder/Folder_1/*_PHYLUM.csv:
species_A,2
species_B,4
species_C,5

Main_Folder/Folder_2/*_PHYLUM.csv:
species_A,3
species_D,5

AFTER:
Samples,Folder_1,Folder_2
species_A,2,3
species_B,4,0
species_C,5,0

species_D,0,5

Similarly, you can get other assignments by typing in the following commands:

```
[uzi@quince-srv2 ~/PROJECT]$ perl /home/opt/AMPLICONprocessing_v0.1/scripts/collateResults.pl -f ~/PROJECT -p
_CLASS.csv
[uzi@quince-srv2 ~/PROJECT]$ perl /home/opt/AMPLICONprocessing_v0.1/scripts/collateResults.pl -f ~/PROJECT -p
_ORDER.csv
[uzi@quince-srv2 ~/PROJECT]$ perl /home/opt/AMPLICONprocessing_v0.1/scripts/collateResults.pl -f ~/PROJECT -p
_FAMILY.csv
[uzi@quince-srv2 ~/PROJECT]$ perl /home/opt/AMPLICONprocessing_v0.1/scripts/collateResults.pl -f ~/PROJECT -p
_GENUS.csv
```

Please note that collateResults.pl is a general purpose script to combine csv files

Example 1 (combining test.csv files in each subfolder within main folder and fixing mac/python bugs):

```
perl /home/opt/AMPLICONprocessing_v0.2/scripts/collateResults.pl -f path_to_main_folder -p _test.csv |
sed -e 's/\r//g' |
awk '{for(i=1; i<=NF; i++) if($i=="") $i=0; else { f=$i; while( gsub(/"/,"&",f)%2 && i<NF) f=f $(++i)}}1' FS=, OFS=,
> output.csv
```

Here, sed is used to fix python/mac '\n' errors, and the awk command is used to replace empty fields with 0

Example 2 (combining either of the [a/b/c/d/e].csv files within main folder):

```
perl /home/opt/AMPLICONprocessing_v0.2/scripts/collateResults.pl -f /home/uzi/TSB_AMPLICONS_ILLUMINA -p
"_a.csv|_b.csv|_c.csv|_d.csv|_e.csv" | sed -e 's/\r//g' > output.csv
```

You can process these frequency tables further in our TAXAenv pipeline: http://quince-srv2.eng.gla.ac.uk:8080/
Tutorial: http://userweb.eng.gla.ac.uk/umer.ijaz/TAXAenv_tutorial.pdf