PART I Builing a reference database

#Build a reference database (EMBL)

#Download the whole set of EMBL sequences

wget ftp://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/std/rel_std_pln*.dat.gz

gunzip *.gz

#Download the NCBI taxonomy

wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz

tar -xzvf taxdump.tar.gz

#Format them into the ecoPCR format

obiconvert --embl -t ../../taxo/ --ecopcrdb-output=pln_r143 --skip-on-error ../raw/pool*.dat

#Use ecoPCR to simulate amplification and build a reference database

ecoPCR -d ~/ngs/database/embl/ecopcr/pln_r143 -e 3 -l 100 -L 250 GTACACACCGCCCGTC TGATCCTTCTGCAGGTTCACCTAC > ~/ngs/database/embl/ecopcr_Euk_V9/18S.pln.r143.ecopcr

#Clean the database

#filter sequences so that they have a good taxonomic description at the species, genus, and family levels

obigrep -d pln_r143 --require-rank=species --require-rank=genus --require-rank=family pool*.ecopcr > pool*.clean.fasta

#remove redundant sequences

obiuniq -d pln_r143 pool*.clean.fasta > pool*.uniq.fasta

#ensure that the dereplicated sequences have a taxid at the family level

obigrep -d pln_r143 --require-rank=family pool*.uniq.fasta > pool*.taxid.fasta

#ensure that sequences each have a unique identification

obiannotate --uniq-id pool*.taxid.fasta > db.pool*.fasta

PART II Processing raw data

#Decompress raw data

gunzip pool*-R*.fastq.gz

#QC Analysis

fastqc pool*-R1.fastq

fastqc pool*-R2.fastq

#Alignment F/R

illuminapairedend --score-min=32 -r pool*2.fastq pool*1.fastq > pool*.align.fastq

#Filtering Unpaired

obigrep -p 'mode!="joined"' pool*.align.fastq > pool*.grep.fastq

#Removing tags & primers

ngsfilter -t pool*.ngsfilter.txt pool*.grep.fastq > pool*.tag.fastq

#Remove repetetive sequences

obiuniq -m sample pool*.tag.fastq > pool*.uniq.fasta

#OBICount

obigrep -l 100 -L 250 -p 'count>=10' pool*.uniq.fasta > pool*.c10.l100.L250.fasta

obigrep -l 100 -L 250 -p 'count<10 and count>=1' pool*.uniq.fasta > pool*.c1-10.l150.L300.fasta

#Count statistics

obistat -c count pool*.uniq.fasta

#Clean PCR/Sequencing Errors

obiclean -r 0.05 -H pool*.uniq.c110.l100.L250.fasta > pool*.uniq.clean.c110.l100.L250.fasta

obiclean -r 0.05 -H pool*.uniq.c1-10.l100.L250.fasta > pool*.uniq.clean.c1-10.l100.L250.fasta

#Taxonomic assignment

#via downloaded EMBL database (offline)

ecotag -d database/pln_r143 -R database/db.pool*.fasta pool*.uniq.clean.c110.l100.L250.fasta >pool*.uniq.clean.c110.l100.L250.fasta.l100.ecotag.fasta

ecotag -d database/pln_r143 -R database/db.pool*.fasta pool*.uniq.clean.c1-10.l100.L250.fasta >pool*.uniq.clean.c1-10.l100.L250.ecotag.fasta

#via remote BLAST (online)

blastn -query pool*.uniq.clean.c110.l100.L250.fasta -db nt -remote -max_target_seqs 3 -outfmt "6 qseqid stitle length pident ssciname" -out pool*.uniq.clean.c110.l100.L250.rblast.results

blastn -query pool*.uniq.clean.c1-10.l100.L250.fasta -db nt -remote -max_target_seqs 3 -outfmt "6 qseqid stitle length pident ssciname" -out pool*.uniq.clean.c1-10.l100.L250.rblast.results

#Annotating/Exporting to Spreadsheet

obiannotate --delete-tag=scientific_name_by_db --delete-tag=obiclean_samplecount --delete-tag=obiclean_count --delete-tag=obiclean_singletoncount --delete-tag=obiclean_cluster --delete-tag=obiclean_internalcount --delete-tag=obiclean_head --delete-tag=taxid_by_db --delete-tag=obiclean_headcount --delete-tag=id_status --delete-tag=rank_by_db --delete-tag=order_name --delete-tag=order pool*.uniq.clean.c110.l100.L250.ecotag.fasta > pool*.uniq.clean.c110.l100.L250.ecotag.ann.fasta

obiannotate --delete-tag=scientific_name_by_db --delete-tag=obiclean_samplecount --delete-tag=obiclean_count --delete-tag=obiclean_singletoncount --delete-tag=obiclean_cluster --delete-tag=obiclean_internalcount --delete-tag=obiclean_head --delete-tag=taxid_by_db --delete-tag=obiclean_headcount --delete-tag=id_status --delete-tag=rank_by_db --delete-tag=order_name --delete-tag=order pool*.uniq.clean.c1-10.l100.L250.ecotag.fasta > pool*.uniq.clean.c1-10.l100.L250.ecotag.ann.fasta

#Sorting Read Numbers

obisort -k count -r pool*.uniq.clean.c110.l100.L250.ecotag.ann.fasta > pool*.uniq.clean.c110.l100.L250.sort.fasta

obisort -k count -r pool*.uniq.clean.c1-10.l100.L250.ecotag.ann.fasta > pool*.uniq.clean.c1-10.l100.L250.sort.fasta

#Converting to a tab-delimited/excel file

obitab -o pool*.uniq.clean.c110.l100.L250.sort.fasta > pool*.uniq.clean.c110.l100.L250.sort.fasta.sort.tab

obitab -o pool*.uniq.clean.c1-10.l100.L250.sort.fasta > pool*.uniq.clean.c1-10.l100.L250.sort.fasta.sort.tab

obitab -o pool*.uniq.clean.c110.l100.L250.sort.fasta > pool*.uniq.clean.c110.l100.L250.sort.xls

obitab -o pool*.uniq.clean.c1-10.l100.L250.sort.fasta > pool*.uniq.clean.c1-10.l100.L250.sort.xls