Supplemental Material

Origin of chemical diversity in the *Prochloron*-tunicate symbiosis

Zhenjian Lin, Joshua P. Torres, Ma. Diarey Tianero, Jason C. Kwan, and Eric W. Schmidt[*]

Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT 84112, USA

*ews1@utah.edu

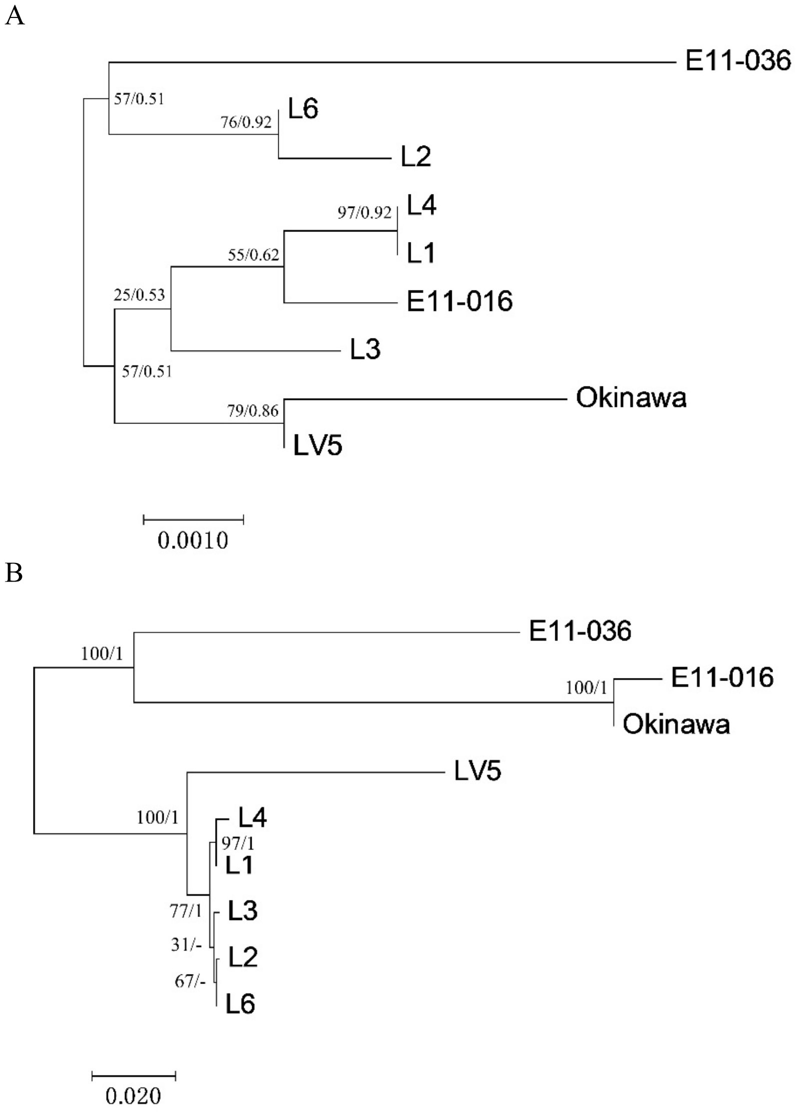


Figure S1. Comparison of 16srRNA gene tree with the 18S tree topology. The maximum likelihood (ML) tree from MEGA 7.0 analysis of the **(A)** 16S rRNA nucleotide sequences of *Prochloron* and **(B)** 18S rRNA nucleotide sequences of tunicates. Maximum likelihood (ML) bootstrap values and Bayesian clade credibility values are indicated at the nodes (bootstrap values/clade credibility values).

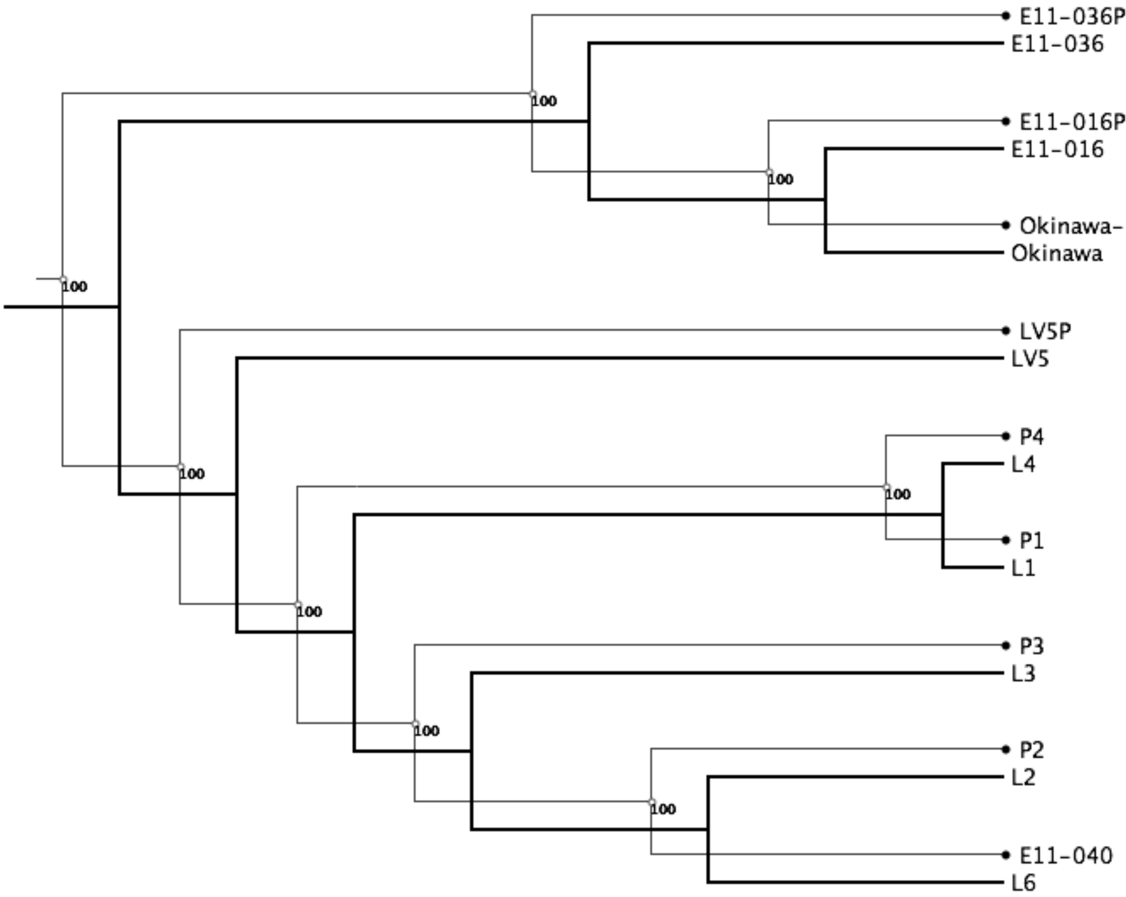
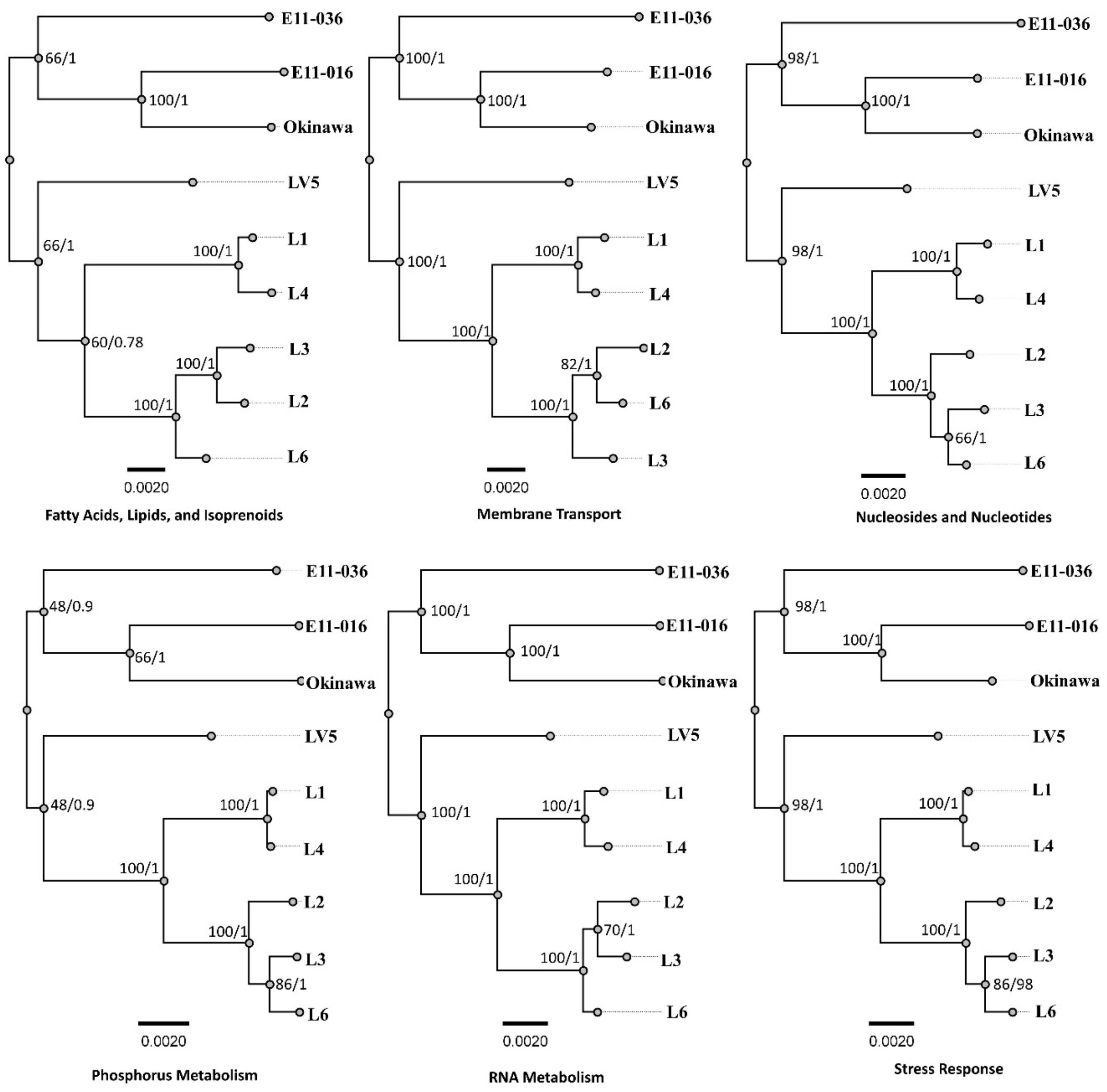


Fig S2. Reconciliation between *Prochloron* and tunicate phylogenies. One of 63 solutions, all of which show eight cospeciations and no duplications, host switches, or losses (total cost = 0). Reconciliation of *Prochloron* and host trees was generated with Jane v.4. Black and bold black lines represent *Prochloron* and their tunicate hosts, respectively. Empty circles represent cospeciations; supporting value was labeled at each node. (The *Prochloron* tree was generated from the concatenation of the 1851 conserved genes and the tunicate tree was generated from 18s rRNA gene. P1-P4, E11-040, E11-036P, E11-016P and Okinawa-P are the *Prochlorons* corresponding to their hosts)



Fatty Acids, Lipids, and Isoprenoids

Membrane Transport

Nucleosides and Nucleotides

Phosphorus Metabolism

RNA Metabolism

Stress Response

Fig S3. The maximum likelihood (ML) trees derived from individual functional categories of genes showed similar topology to 18s rRNA tree in FigS1. Maximum likelihood (ML) bootstrap values and Bayesian clade credibility values are indicated at the nodes (bootstrap values/clade credibility values).

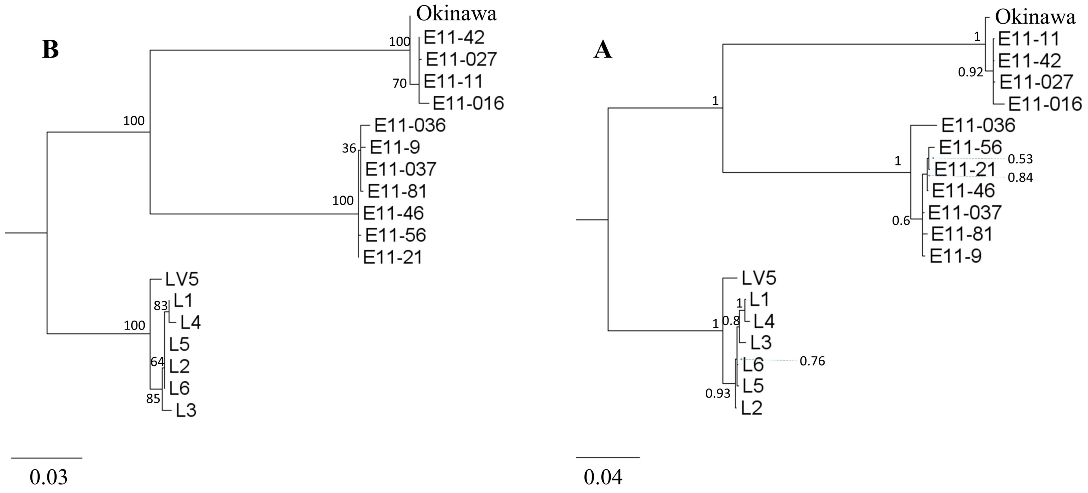


Fig S4. Phylogenetic trees from (A) Maximum Likelihood analysis (by MEGA 7.0) and (B) Bayesian analysis (by Mrbayes) of 18s rRNA gene. Maximum likelihood (ML) bootstrap values and Bayesian clade credibility values are indicated at the nodes.

Table S1. Collection location (coordinates) for each sample.

07°15', 134°15'	L1
-17°55', 177°16'	L2
-08°57'21.0000", 159°15'41.4000"	L3
-04°45'36.0600", 151°25'19.2000"	L4
-10°16'07.9248", 145°38'45.4128"	L6, E11-036, L5, E11-037
-04°07'60.0000", 151°34'00.0000"	LV5
-10°10'12.6077", 145°42'53.8812"	E11-021
-10°10'34.8744", 145°33'13.1904"	E11-046, E11-042
-10°02'13.4826", 145°46'03.8676"	E11-09, E11-011, E11-016
-10°11'56.4015", 145°38'31.4016"	E11-027
-10°02'36.3552", 145°33'23.1480"	E11-056
-10°00'21.3372", 145°43'59.6352"	E11-081
Hateruma Island, Okinawa	Okinawa

Table S2. Metagenome data obtained in this study (blue) and in previous studies (gray).

host tunicate	sample collection	metagenome sequencing	metagenome assembly	metagenome binning	*Prochloron* genome identification	*Prochloron* genome annotation
L1	▨	▨	▨	▨	▨	▨
L2	▨	▨	▨	▨	▨	▨
L3	▨	▨	▨	▨	▨	▨
L4	▨	▨	▨	▨	▨	▨
L5	▨	▨	🟦	🟦	🟦	🟦
L6	▨	▨	🟦	🟦	🟦	🟦
LV5	🟦	🟦	🟦	🟦	🟦	🟦
E11-036	🟦	🟦	🟦	🟦	🟦	🟦
E11-037	🟦	🟦	🟦	🟦	🟦	🟦
E11-016	🟦	🟦	🟦	🟦	🟦	🟦
Okinawa	🟦	🟦	🟦	🟦	🟦	🟦

Table S3. Metagenome sequencing and assembly information.

sample	Illumina libraries	sequencer	sequencing runs	sequencing depth reads	*Prochloron* reads	assemble program	total length of contigs	number of contigs	N50 stats	GC %
L5	350 bps	Illumina HiSeq 2000 sequencer	101 bp paired-end	244718082	NA	IDBA_ud	685252988	253830	6939	32.1
L6	350 bps	Illumina HiSeq 2000	101 bp paired-end	347679462	2797494	IDBA_ud	712541213	241649	10445	31.9
LV5	400 bps	Illumina HiSeq 2000	101 bp paired-end	629369536	13607781	IDBA_ud	756714778	380574	5973	32.8
E11-036	350 bps	Illumina MiSeq	251 bp paired-end	17849726	249387	IDBA_ud	716585695	490682	3303	34.9
E11-037	350 bps	Illumina HiSeq 2000	101 bp paired-end	369180626	NA	IDBA_ud	944670545	473951	6648	35.5
E11-016	280/800bps	Illumina HiSeq 2000	125 bp paired-end	174618110	8443432	IDBA_ud	957775449	944402	2460	34.1
Okinawa	350 bps	Illumina HiSeq 2000	125 bp paired-end	190421681	3657600	IDBA_ud	932843649	607266	3155	33.9

Table S4. Primers used in this study.

Primer Name	Sequence
AscF2new	CAAGGAAGGCAGCAGGCGCGCAAAT
AscR5New	GCGGTGTGTACAAAGGGCAGGGA
DiplosomaF	GCGTTCGAAGCAGTCTTG
DiplosomaR	GATCGCCTTTTCGTCGGA
LissoclinumF	TAACGACACTGCGAAAGGC
LissoclinumR	GCTCGATCCCCGAAGGAC
DidemnumF	GTCTACGTGGTCGTGCGGCGACG
DidemnumR	TTCACGAGCCTCTCGGCCCGC
201_Fwd	ATGAAAAGATTGGTGAGCGTA
201_Rev	TTAGTAAACGCCAATATTAAAGC

Table S5. Branch lengths comparison of maximum likelihood (ML) tree derived from 1,851 conserved *Prochloron* genes (comparison tree) and 18S rRNA gene sequences (reference tree) from the same samples.

Brl_ref_tree	Brl_comp_tree	Brl_comp_tree_K	Partition
0.03032	0.00127	0.00668	(E11-036, Okinawa, E11-016) / (LV5, L4, L1, L3, L2, L6, root)
0.11493	0.00424	0.02221	(Okinawa, E11-016) / (E11-036, LV5, L4, L1, L3, L2, L6, root)
0.03032	0.00127	0.00668	(LV5, L4, L1, L3, L2, L6) / (E11-036, Okinawa, E11-016, root)
0.00546	0.00363	0.01906	(L4, L1, L3, L2, L6) / (E11-036, Okinawa, E11-016, LV5, root)
0.00154	0.00501	0.02626	(L4, L1) / (E11-036, Okinawa, E11-016, LV5, L3, L2, L6, root)
0.00094	0.00408	0.02142	(L3, L2, L6) / (E11-036, Okinawa, E11-016, LV5, L4, L1, root)
0.00064	0.00131	0.00689	(L2, L6) / (E11-036, Okinawa, E11-016, LV5, L4, L1, L3, root)
0.09245	0.01324	0.06945	(E11-036) / (Okinawa, E11-016, LV5, L4, L1, L3, L2, L6, root)
0	0.00741	0.03888	(Okinawa) / (E11-036, E11-016, LV5, L4, L1, L3, L2, L6, root)
0.01159	0.00712	0.03735	(E11-016) / (E11-036, Okinawa, LV5, L4, L1, L3, L2, L6, root)
0.06176	0.00815	0.04273	(LV5) / (E11-036, Okinawa, E11-016, L4, L1, L3, L2, L6, root)
0.00304	0.00149	0.00779	(L4) / (E11-036, Okinawa, E11-016, LV5, L1, L3, L2, L6, root)
0	0.00114	0.00596	(L1) / (E11-036, Okinawa, E11-016, LV5, L4, L3, L2, L6, root)
0.00118	0.00215	0.01129	(L3) / (E11-036, Okinawa, E11-016, LV5, L4, L1, L2, L6, root)
0.00061	0.00342	0.01794	(L2) / (E11-036, Okinawa, E11-016, LV5, L4, L1, L3, L6, root)
0	0.00393	0.02061	(L6) / (E11-036, Okinawa, E11-016, LV5, L4, L1, L3, L2, root)

Brl_ref_tree: Branch length of this partition on the reference tree.
Brl_comp_tree: Branch length of this partition on the original comparison tree.
Brl_comp_tree_K: Branch length of this partition on the comparison tree after scaling.
Partition: Tip nodes that constitute this partition.

Table S6. Topology comparison of maximum likelihood (ML) tree derived with 18S rRNA gene sequences (reference tree) from the same samples.

Trees	K-score	Scale_fac	Symm_dif	N_partitions ref_tree	N_partitions comp_tree
concatenated-1851 genes	0.12215	5.24499	0	16	16
Fatty Acids, Lipids, and Isoprenoids	0.11996	5.35526	2	16	16
Membrane Transport	0.11633	5.85165	0	16	16
Nucleosides and Nucleotides	0.11403	7.43616	2	16	16
Phosphorus Metabolism	0.12765	6.34812	2	16	16
RNA Metabolism	0.12089	6.59201	2	16	16
stress response	0.11418	6.98805	2	16	16

K-score: the minimum branch length distance
Scale_fac: scale factor
Symm_dif: symmetric difference (Robinson-Foulds)
N_partitions_ref_tree: number of partitions of reference tree
N_partitions_comp_tree: number of partitions of comparison tree

Source code

1.fasta_extract_mult.pl

```
#!/usr/bin/perl -w
# fasta_extract_mult.pl - a program that will extract sequences from a
# multifasta file given the sequence headers (without leading ">")
# USAGE fasta_extract_mult.pl --in <input_fasta> --seqs <seq_names>
#                                         --out <output_fasta>
use strict;
use warnings;
use Getopt::Long;
my ($input_fasta, $seq_names, $output_fasta);
GetOptions (
    'in=s'       => \$input_fasta,
    'seqs=s'      => \$seq_names,
    'out=s'      => \$output_fasta,
);
my ($header, $sequence, $switch, %seq_hash);
open (my $seq_names_fh, "<", $seq_names) or die "Can't open $seq_names\n";
while (my $line = <$seq_names_fh>) {
    chomp($line);
    $seq_hash{"\>$line"} = 1;
}
close $seq_names_fh;
open (my $input_fasta_fh, "<", $input_fasta) or die "Can't open $input_fasta\n";
open (my $output_fasta_fh, ">", $output_fasta);
while (my $line = <$input_fasta_fh>) {
    chomp $line;
    if ($line =~ /^>/) {
        print_wrap($header, $sequence);
        ($header, $sequence) = ($line, '');
        }
        else {
                $sequence .= "$line";
        }
        if (eof($input_fasta_fh)) {
                print_wrap($header, $sequence);
        }
}
sub print_wrap {
        my ($header, $sequence) = @_;
        return unless (defined($header) && defined($seq_hash{$header}));
        if ($seq_hash{$header} == 1) {
                print $output_fasta_fh "$header\n";
                foreach (split(/(.{60}.+?)\s/, $sequence)) {
                        print $output_fasta_fh "$_\n" if $_ ne "";
                }
        }
}
```

2. proteome_comparison.sh

```
#!/bin/bash
blastp -query combin.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db ../db/P1 -out combin_blast_P1
wait
awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin_blast_P1 > hit
perl ~/scripts/jason/delete_seqs_from_fasta.pl P2-1215.24.faa hit P2-1.fa
perl ~/scripts/jason/delete_seqs_from_fasta.pl P3-1215.23.faa hit P3-1.faa
perl ~/scripts/jason/delete_seqs_from_fasta.pl P4-1215.25.faa hit P4-1.faa
perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-016-1117.47.faa hit E11-016-1.faa
perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-036-1215.20.faa hit E11-036-1.faa
perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-1117.26.faa hit E11-040-1.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-1117.55.faa hit diplosoma-1.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-1117.56.faa hit LV5-1.faa
```

```
cat P3-1.faa P4-1.faa E11-016-1.faa E11-036-1.faa E11-040-1.faa LV5-1.faa diplosoma-1.faa >
combin-1.faa

makeblastdb -in P2-1.faa -dbtype 'prot' -out db/P2-1

blastp -query combin-1.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -
best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db
db/P2-1 -out combin-1_blast_P2-1
```

```
awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-1_blast_P2-1 > hit1

perl ~/scripts/jason/delete_seqs_from_fasta.pl P3-1.faa hit1 P3-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl P4-1.faa hit1 P4-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-016-1.faa hit1 E11-016-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-036-1.faa hit1 E11-036-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-1.faa hit1 E11-040-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-1.faa hit1 diplosoma-2.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-1.faa hit1 LV5-2.faa
```

```
cat  P4-2.faa E11-016-2.faa E11-036-2.faa E11-040-2.faa LV5-2.faa diplosoma-2.faa > combin-
2.faa

makeblastdb -in P3-2.faa -dbtype 'prot' -out db/P3-2

blastp -query combin-2.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -
best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db
db/P3-2 -out combin-2_blast_P3-2

awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-2_blast_P3-2 > hit2

perl ~/scripts/jason/delete_seqs_from_fasta.pl P4-2.faa hit2 P4-3.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-016-2.faa hit2 E11-016-3.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-036-2.faa hit2 E11-036-3.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-2.faa hit2 E11-040-3.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-2.faa hit2 diplosoma-3.faa
perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-2.faa hit2 LV5-3.faa

cat E11-016-3.faa E11-036-3.faa E11-040-3.faa LV5-3.faa diplosoma-3.faa > combin-3.faa

makeblastdb -in P4-3.faa -dbtype 'prot' -out db/P4-3

blastp -query combin-3.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/P4-3 -out combin-3_blast_P4-3

awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-3_blast_P4-3 > hit3

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-016-3.faa hit3 E11-016-4.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-036-3.faa hit3 E11-036-4.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-3.faa hit3 E11-040-4.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-3.faa hit3 diplosoma-4.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-3.faa hit3 LV5-4.faa

cat E11-036-4.faa E11-040-4.faa LV5-4.faa diplosoma-4.faa > combin-4.faa

makeblastdb -in E11-016-4.faa -dbtype 'prot' -out db/E11-016-4

blastp -query combin-4.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/E11-016-4 -out combin-4_blast_E11-016-4

awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-4_blast_E11-016-4 > hit4

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-036-4.faa hit4 E11-036-5.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-4.faa hit4 E11-040-5.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-4.faa hit4 diplosoma-5.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-4.faa hit4 LV5-5.faa

cat E11-040-5.faa LV5-5.faa diplosoma-5.faa > combin-5.faa

makeblastdb -in E11-036-5.faa -dbtype 'prot' -out db/E11-036-5

blastp -query combin-5.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/E11-036-5 -out combin-5_blast_E11-036-5


awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-5_blast_E11-036-5 > hit5

perl ~/scripts/jason/delete_seqs_from_fasta.pl E11-040-5.faa hit5 E11-040-6.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-5.faa hit5 diplosoma-6.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-5.faa hit5 LV5-6.faa
```

```
cat diplosoma-6.faa LV5-6.fa > combin-6.faa

 makeblastdb -in E11-040-6.faa -dbtype 'prot' -out db/E11-040-6

blastp -query combin-6.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/E11-040-6 -out combin-6_blast_E11-040-6


awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }' combin-6_blast_E11-040-6 > hit6

perl ~/scripts/jason/delete_seqs_from_fasta.pl diplosoma-6.faa hit6 diplosoma-7.faa

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-6.faa hit6 LV5-7.faa

cat LV5-7.faa > combin-7.faa

makeblastdb -in diplosoma-7.faa -dbtype 'prot' -out db/diplosoma-7

blastp -query combin-7.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/diplosoma-7 -out combin-7_blast_diplosoma-7


awk ' $3>80 i&& $4/$8>0.8 && $6>80 { print $1 }'  combin-7_blast_diplosoma-7 > hit7

perl ~/scripts/jason/delete_seqs_from_fasta.pl LV5-7.faa hit7 LV5-8.faa


cat P1-1117.48.faa P2-1.faa P3-2.faa P4-3.faa E11-016-4.faa E11-036-5.faa E11-040-6.faa diplosoma-7.faa LV5-8.faa > all-unique.faa


makeblastdb -in all-unique.faa -dbtype 'prot' -out db/all-unique

#############################

wait


blastp -query P1-1117.48.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -out blast_unique/P1_blast_all_unique
blastp -query P2-1215.24.faa -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -out blast_unique/P2_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query P3-1215.23.faa -out blast_unique/P3_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query P4-1215.25.faa -out blast_unique/P4_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query E11-016-1117.47.faa -out blast_unique/E11-016_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query E11-036-1215.20.faa -out blast_unique/E11-036_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query E11-040-1117.26.faa -out blast_unique/E11-040_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query diplosoma-1117.55.faa -out blast_unique/diplosoma_blast_all_unique
blastp -num_threads 128 -max_target_seqs 1 -best_hit_overhang 0.1 -best_hit_score_edge 0.1 -outfmt "6 qseqid sseqid pident length evalue qcovs qlen slen" -db db/all-unique -query LV5-1117.56.faa -out blast_unique/LV5_blast_all_unique

wait

awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' diplosoma_blast_all_unique > diplosoma_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' E11-016_blast_all_unique > E11-016_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' E11-036_blast_all_unique > E11-036_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' E11-040_blast_all_unique > E11-040_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' P1_blast_all_unique > P1_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' P2_blast_all_unique > P2_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' P3_blast_all_unique > P3_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' P4_blast_all_unique > P4_hits
awk ' $3>80 && $4/$8>0.8 && $6>80 { print $2 }' LV5_blast_all_unique > LV5_hits
```

3. delete_seqs_from_fasta.pl

```
#!/usr/bin/perl

# Program to delete sequences from a fasta, when given a list of sequences to delete.
# Especially for large Illumina-derived fastas (e.g. /1 or /2 at end of seqs).
# USAGE delete_seqs_from_fasta.pl source.fasta seqs.list output.fasta

use strict;
use warnings;

unless (defined $ARGV[2])
{
    print "\nNot enough arguments\n";
    exit 1;
}

unless (-e $ARGV[0])
{
    print "\nCouldn't find source .qual file\n";
    exit 1;
}
```

```
unless (-e $ARGV[1])
{
    print "\nCouldn't find list file\n";
    exit 1;
}

my $inputfile = $ARGV[0];
my $inputlist = $ARGV[1];
my $outputfile = $ARGV[2];

# make a hash of sequence names
my %seqs;

open (INPUTLIST, "<$inputlist");
while (<INPUTLIST>)
{
    chomp $_;
    $seqs{ $_ } = 1;
}
close INPUTLIST;

# Parse through fasta file, keeping the lines that belong to the correct sequences
open (INPUTFILE, "<$inputfile");
open (OUTPUTFILE, ">$outputfile");

my $in_sequence = 0; # Boolean to see if we are within a sequence of interest
while (<INPUTFILE>)
{
    my $line = $_;
    my @linearray = split (" ", $line);

    if ($linearray[0] =~ m{>}) # The current line is a sequence header
    {
        my $seq_temp = substr ($linearray[0], 1);
        my @seq_array = split (/\//, $seq_temp); # for Illumina paired end files
        my $seq_name = $seq_array[0];

unless (defined $seqs{$seq_name}) # then we want to keep this sequence
        {
            print OUTPUTFILE $line;
            $in_sequence = 1;
        }
        else # We don't want to keep this sequence
        {
            $in_sequence = 0;
        }
    }
    else
    {
        if ($in_sequence == 1) # we are within a sequence we want to keep
        {
```

```
                        print OUTPUTFILE $line;
                }
        }
}

close INPUTFILE;
close OUTPUTFILE;
```