Tephritid Metabarcoding workshop

Alexander Piper

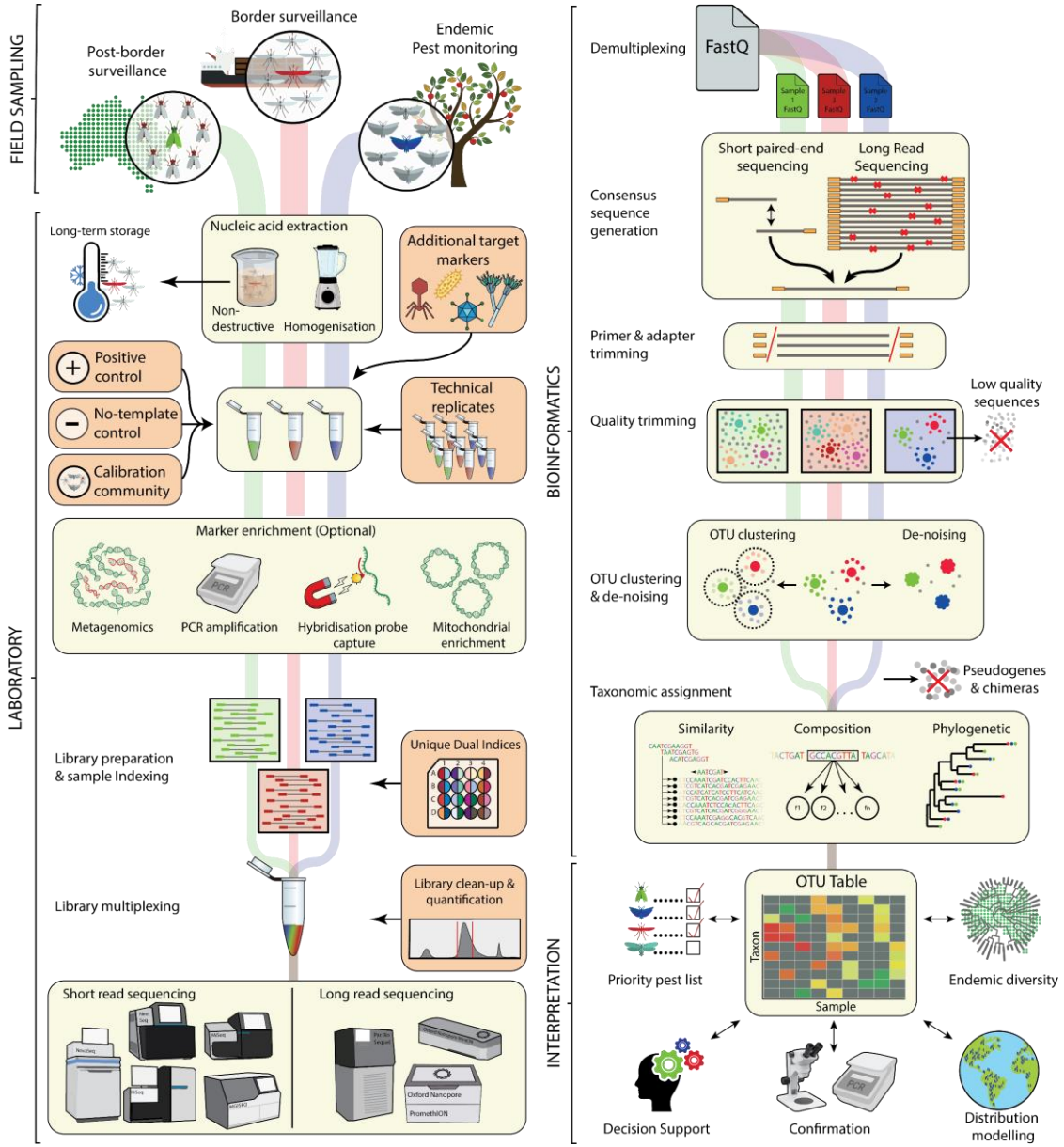
AGRICULTURE VICTORIA

Workshop overview

Tuesday 27th September - OPTIONAL	**2:00-2:30**	**Arrive at AgriBio, get visitors pass**
	2:30-5:00	Lab: First PCR (incubate overnight)
Wednesday 28th	9:30-10:30	Presentation: Overview of metabarcoding lab work
	10:30-12:30	Lab: Run gel, Clean-up & Normalise amplicons
	12:30-1:30	Lunch
	1:30-3:30	Lab: Indexing PCR
	3:30-4:00	Afternoon tea
	4:00-5:30	Lab: Clean-up & Normalise indexed amplicons (incubate overnight)
Thursday 29th	9:30-10:30	Presentation: Overview of bioinformatics
	10:30-12:30	Bioinformatics: Get set up with R, Run pipeline
	12:30-1:30	Lunch
	1:30-3:00	Lab: Finish normalisation, pooling
	3:00-3:30	Afternoon tea
	3:30-5:30	Bioinformatics: look at outputs, quality control
Friday 30th	9:30-12:30	Lab: Final clean-up & Quantification
	12:30-1:30	Lunch
	1:30-3:00	Lab: Load & run MiSeq
	3:30-5:00	Final Discussions

AGRICULTURE VICTORIA

Alex's phone number: 0488 040 119



Piper AM, Batovska J, Cogan NOI, Weiss J, Cunningham JP, Rodoni BC, MJ Blacket (2019). **Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance**. Gigascience

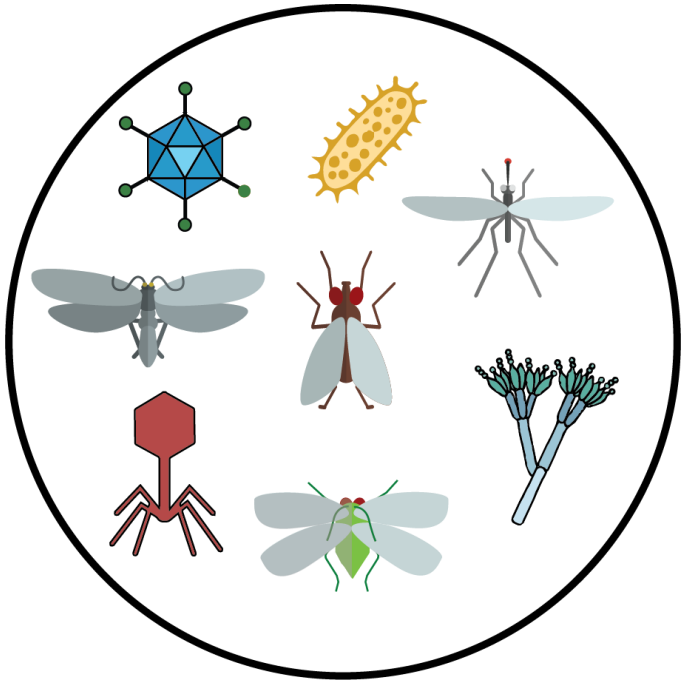
AGRICULTURE VICTORIA

Assay development

AGRICULTURE VICTORIA

Scope of the test

- Metabarcoding assays are generic - can detect a broad range of taxa
- Important to define a clear scope for the intended use of the test
- This scope informs primer design, laboratory protocol, and bioinformatic parameters
- 3 scopes of targets:
    1. Lure responding Dacinae
    2. Other Dacinae
    3. Other Tephritidae



AGRICULTURE VICTORIA

Target loci

- The diversity of species detected depends on how conserved the target locus is across taxa
- The resolution of the assay depends on the amount of nucleotide variation contained within the target locus
- Diagnostic loci for Dacinae were developed through the previous PBCRC project
- However HTS platforms have read length limits which may change their performance
- First step was to validate their performance across the different scopes and design appropriate metabarcoding primers

Table 1. Genetic loci that best separate easily confused species

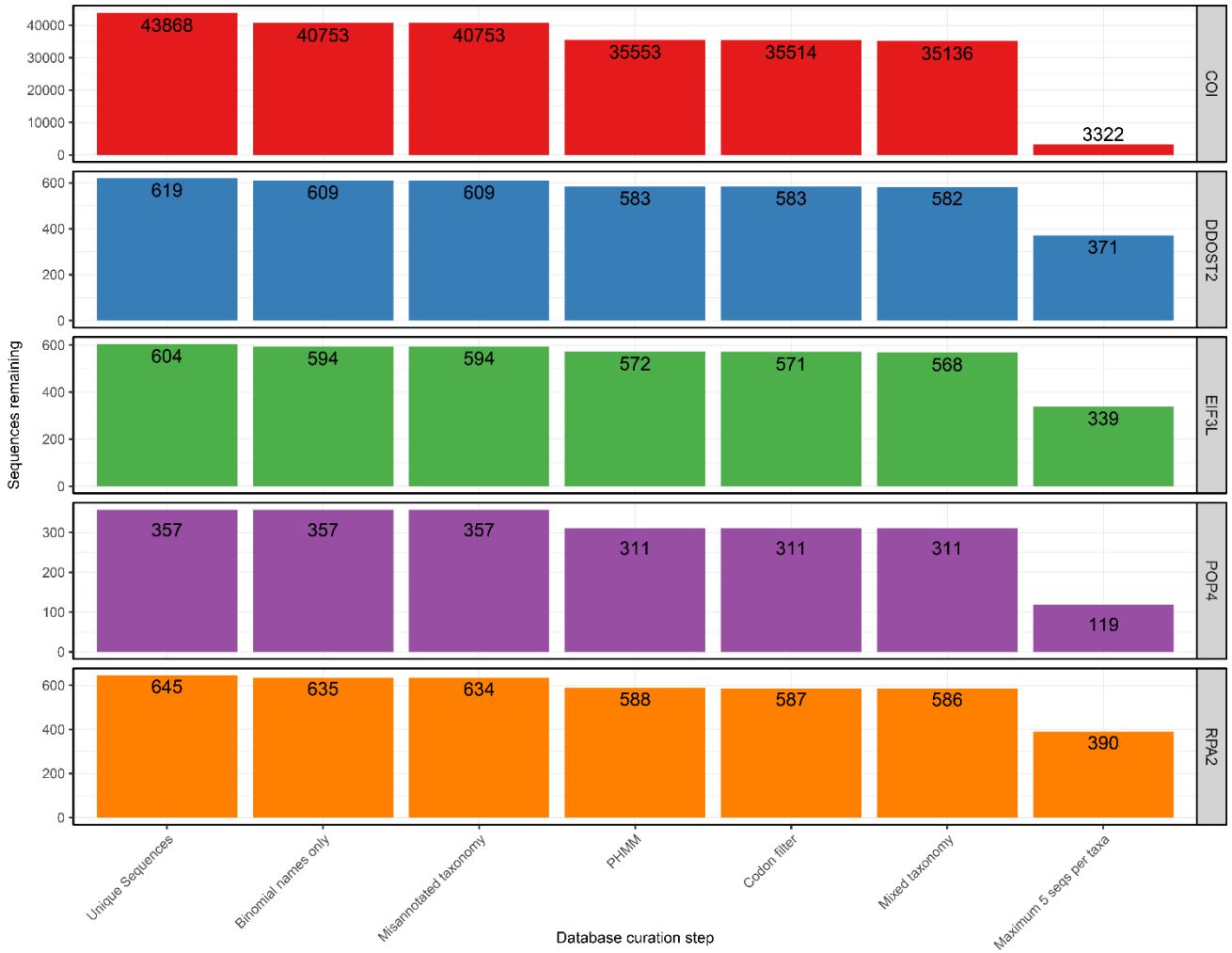
SPECIES PAIR	MOLECULAR DIAGNOSTIC LOCUS				
	COI	POP4	EIF3L	RPA2	DDOSTS2
B. dorsalis / B. carambolae	✓	✓	✓	✓	✓
B. dorsalis / B. kandiensis	✓	✓	✓	✓	✓
B. dorsalis / B. musae	✓	✓	✓	inconclusive	✓
B. dorsalis / B. endiandrae	✓	✓	✓	✓	✓
B. dorsalis / B. cacuminata	✓	✓	✓	inconclusive	✓
B. dorsalis / B. occipitalis	✓	N/A	✓	✓	inconclusive
B. dorsalis / B. opiliae	✓	N/A	✓	✓	✓
B. dorsalis / B. latifrons	✓	✓	✓	✓	✓
B. carambolae / B. kandiensis	✓	✓	✓	✓	✓
B. carambolae / B. opiliae	✓	N/A	✓	✓	✓
B. carambolae / B. musae	✓	✓	✓	✓	✓
B. carambolae / B. occipitalis	✓	inconclusive	✓	✓	✓
B. musae / B. opiliae	✓	N/A	✓	inconclusive	✓
B. musae / B. latifrons	✓	✓	✓	✓	✓
B. musae / B. endiandrae	✓	✓	✓	✓	✓
B. opiliae / B. endiandrae	✓	N/A	✓	✓	✓
B. musae / B. bancroftii	✓	inconclusive	✓	✓	✓
B. zonata / B. correcta / B. pallida / B. jarvisi	✓	✓	✓	✓	✓
B. kraussi / B. tryoni	✓	✓	✓	✓	✓
B. tryoni complex	✗	✗	✗	✗	✗
B. trivialis / B. rufofuscula	✓	N/A	N/A	N/A	N/A
B. trivialis / B. breviaculeus	✓	N/A	N/A	N/A	N/A
B. rufofuscula / B. breviaculeus	inconclusive	N/A	✗	N/A	inconclusive
B. decurtans / B. pallida	✓	N/A	N/A	✓	✓
B. passiflorae / B. facialis	✓	N/A	✓	N/A	N/A
B. frauenfeldi / B. albistrigata	✗	✗	✓	✗	✗
Z. atrisetosus / Z. cucumis	✓	N/A	N/A	N/A	N/A
Z. choristus / Z. cucurbitae	✓	✓	✓	✓	✓
Z. tau / Z. cucurbitae	✓	✓	✓	✓	✓
Z. depressus / Z. tau	✓	N/A	✓	✓	✓

Reference sequences

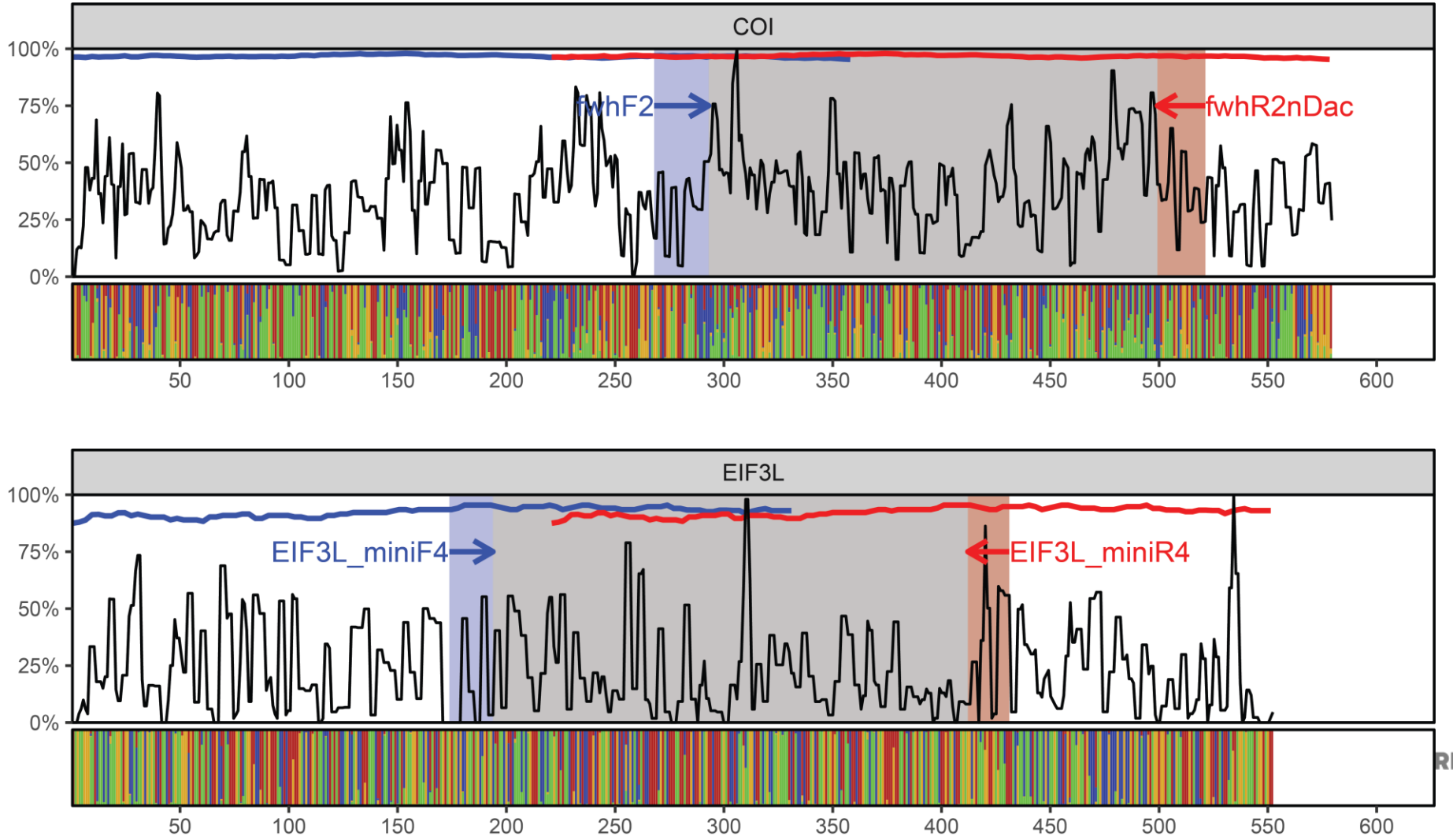
	COI	DDOST2	EIF3L	POP4	RPA2
FruitFlyID	316	168	162	172	174
Melissa thesis	273	273	273		272
Genbank	21,337	178	169	185	199
BOLD	21,478				

Public sequence curation

1. Sequences with identical haplotypes deduplicated
2. Homology verified using a reference Profile Hidden Markov Model (PHMM)
3. Filtered for stop-codons or frameshift mutations.
4. Sequences clustered at 97% identity and flagged if taxonomy disagreed with more than 60% of other sequences within its cluster.
5. Sequence flagged if its intraspecific distance was >3 standard deviations from the mean, or >5% diverged from others
6. Flagged sequences were manually verified using BLAST searches and neighbour-joining trees
7. Database was then pruned to a maximum five representative sequences per species

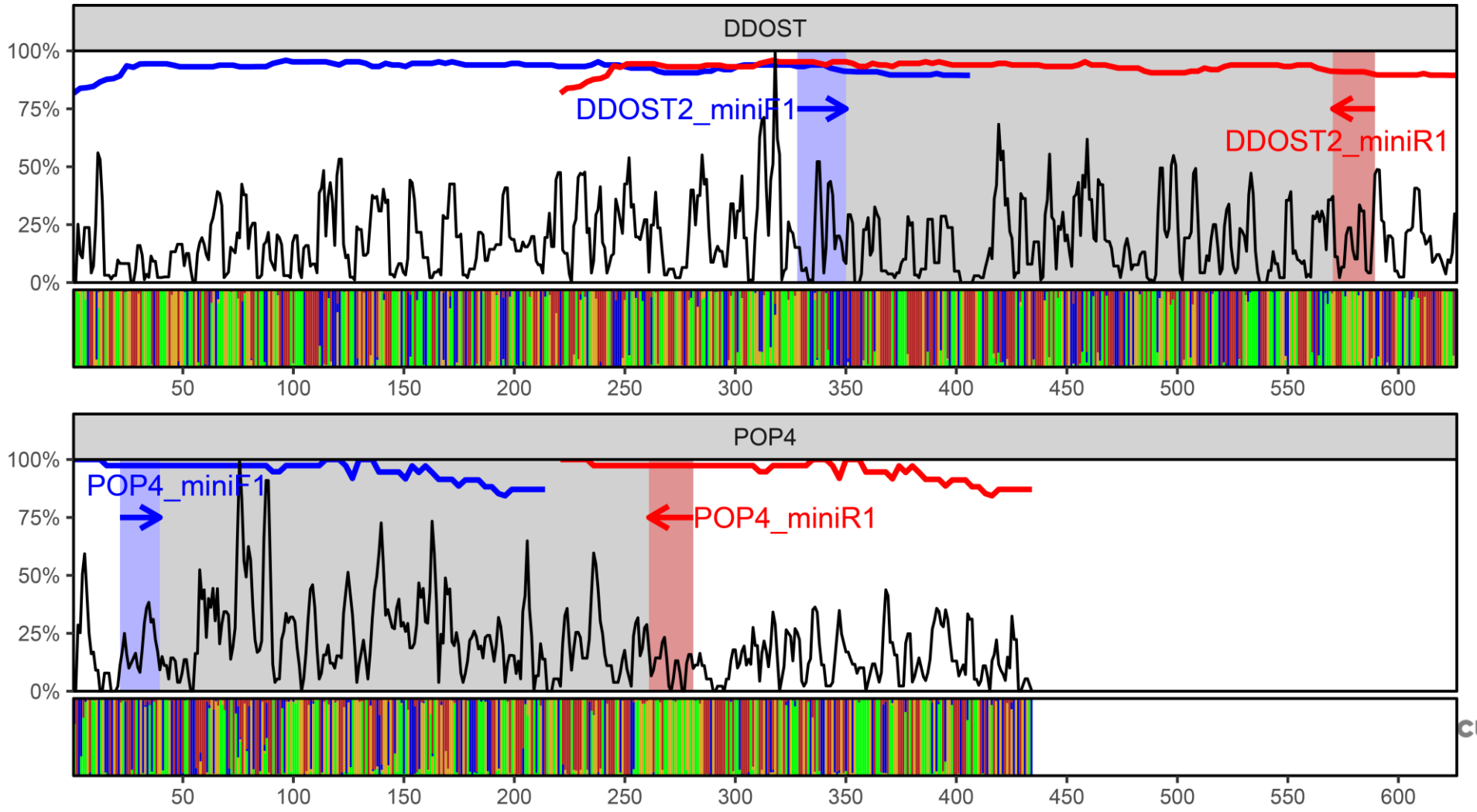


Finding optimal barcode region



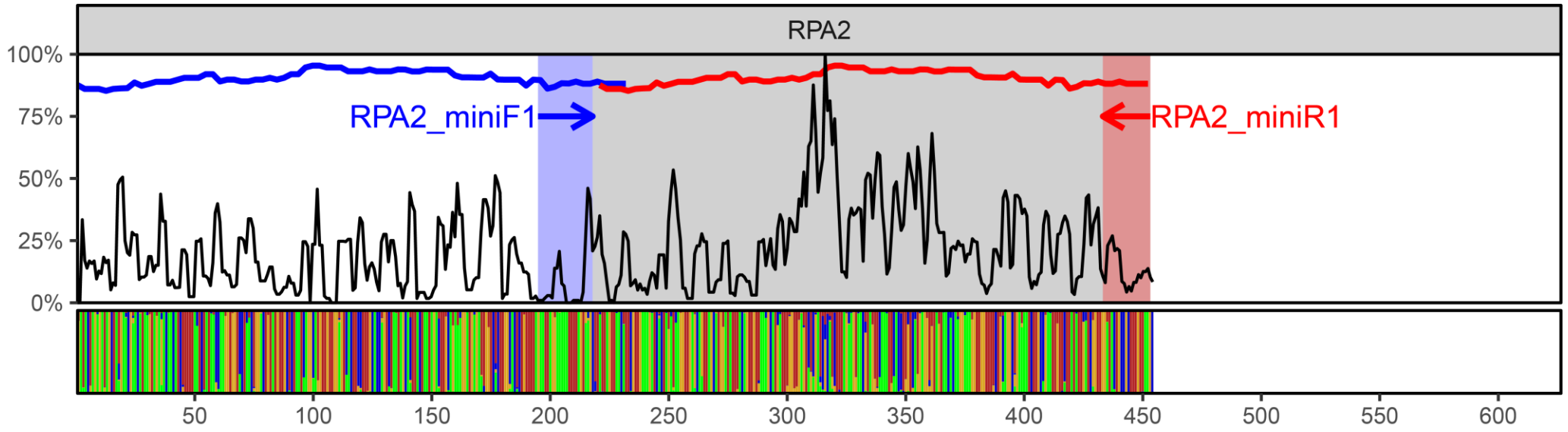
RICULTURE VICTORIA

Finding optimal barcode region

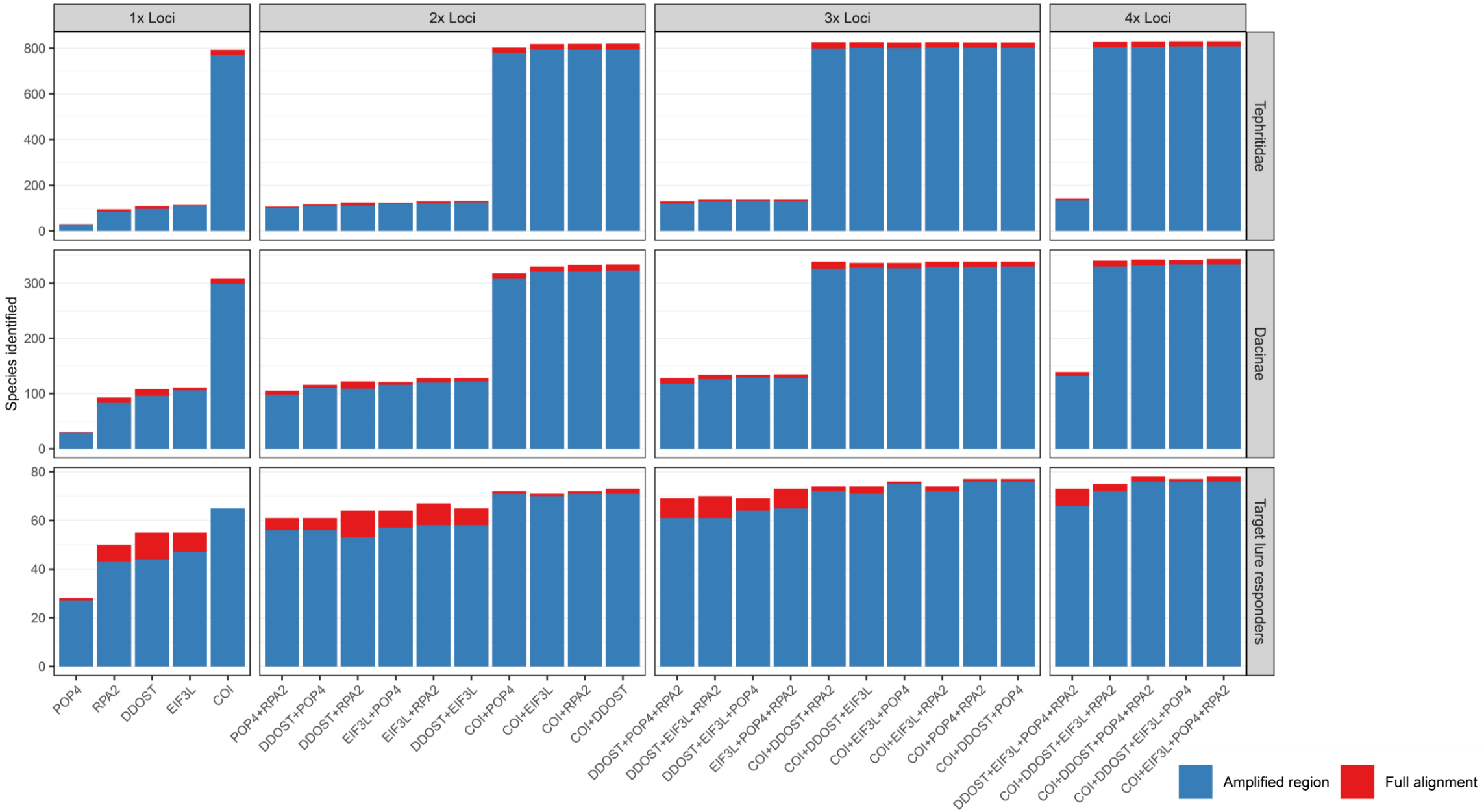


CULTURE VICTORIA

Finding optimal barcode region

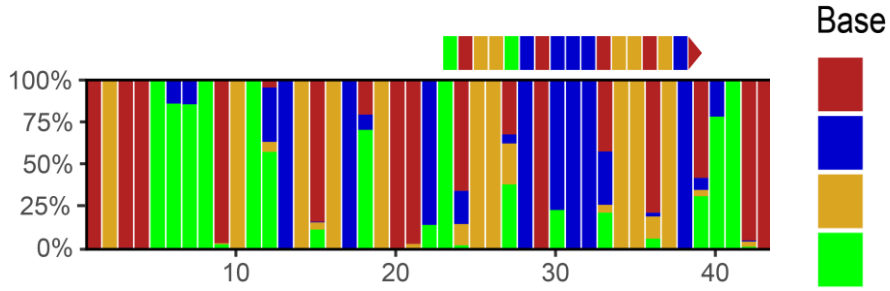


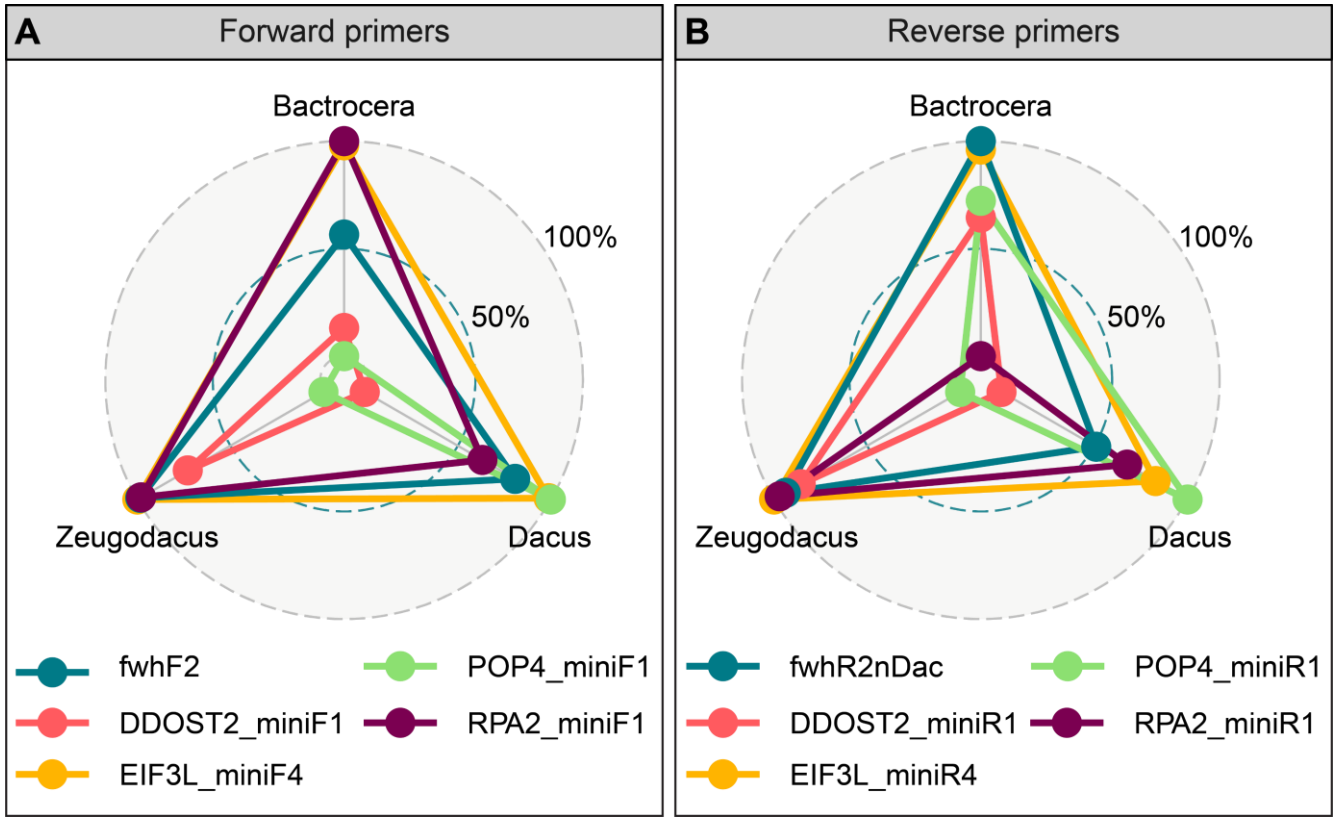
AGRICULTURE VICTORIA


1x Loci
2x Loci
3x Loci
4x Loci
Tephritidae
Dacinae
Target lure responders
Species identified
0
200
400
600
800
100
300
20
40
60
80
POP4
RPA2
DDOST
EIF3L
COI
POP4+RPA2
DDOST+POP4
DDOST+RPA2
EIF3L+POP4
EIF3L+RPA2
DDOST+EIF3L
COI+POP4
COI+EIF3L
COI+RPA2
COI+DDOST
DDOST+POP4+RPA2
DDOST+EIF3L+RPA2
DDOST+EIF3L+POP4
EIF3L+POP4+RPA2
COI+DDOST+RPA2
COI+DDOST+EIF3L
COI+EIF3L+POP4
COI+EIF3L+RPA2
COI+POP4+RPA2
COI+DDOST+POP4
DDOST+EIF3L+POP4+RPA2
COI+DDOST+EIF3L+RPA2
COI+DDOST+POP4+RPA2
COI+DDOST+EIF3L+POP4
COI+EIF3L+POP4+RPA2
Amplified region
Full alignment

PCR primers & Bias

- Differences in primer-binding efficiency can bias read counts towards certain species
- In-silico analyses with the ‘PrimerMiner’ package were used to predict amplification efficiency across all species
- Degenerate nucleotide bases were included where target sequences were variable





Final combination

- COI (205bp):
- fwhF2: GGDACWGGWTGAACWGTWTAYCCHCC
- fwhR2n_dac: GTRATWGCHCCIGCTAADACHGG

- EIF3L (217bp):
- EIF3L_miniF4: GATGCGYCGTTATGCYGATGC
- EIF3L_miniR4: TTRAAYACTTCYARATCRCC

Chosen as they show good diagnostic performance, low predicted bias, and lots of reference sequences available

In combination, uniquely identifies:

- 79.5% of Cue lure / ME responding species
- 73.6% of all Dacinae
- 80.6% of all Tephritidae

AGRICULTURE VICTORIA

Problem species

Lure responders

- B. *tryoni/neohumeralis/aquilonis* (CUE)
- D. *newmani/bellulus* (CUE)
- B. *albistrigata/frauenfeldi* (CUE)
- B. *rufofuscula/peninsularis* (CUE)
- B. *alyxiae/repanda* (CUE)
- B. *tenuifascia/mayi* (ME)
- B. *pallida* (ME)
- B. *recurrens* (CUE)

Exotics

- B. *dorsalis* complex
- B. *nigrifacia* complex
- Ceratitis FAR complex
- A. *fraterculus* complex
- Z. *tau* complex (No EIF3L)

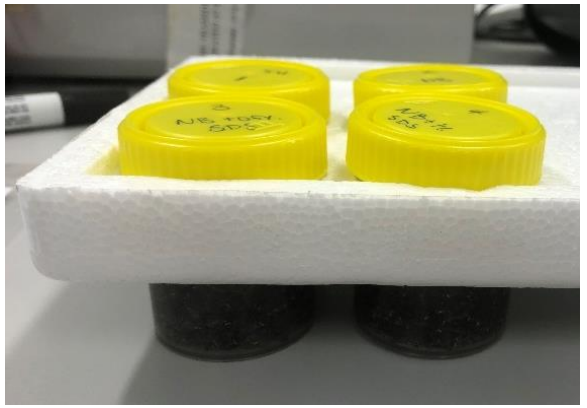
In all these cases, COI were mixed between 'species', and no EIF3L available for some or all

Laboratory steps

AGRICULTURE VICTORIA

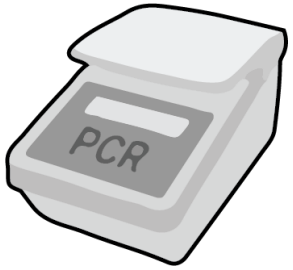
Non-destructive extraction

- DNA is extracted from specimens using HotShot chemical lysis buffer
- Extra clean-up step using Qiagen DNEasy columns as the EDTA inhibits the MYFI polymerase
- Specimens can be retained for morphological confirmation or inclusion in collections
- Specimens contribute DNA depending on their ‘hardness’, rather than biomass



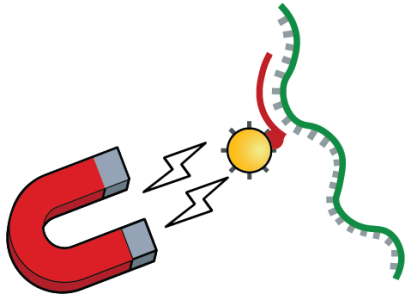
Options: Enriching for target loci

PCR amplification



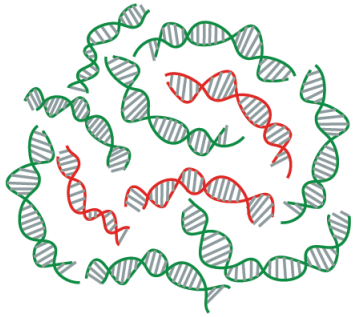
- Standard metabarcoding approach
- Cheap & well established protocols
- PCR amplification can introduce bias
- Requires careful primer design

Hybridisation probes



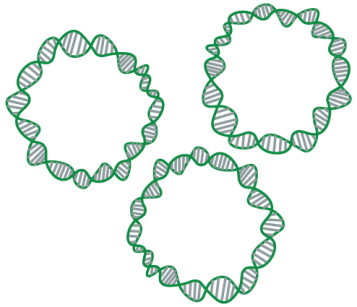
- Flexible length and number of target loci
- Less amplification bias
- Less affected by divergent binding regions

No enrichment (Metagenomics)



- Can return whole genomes
- Can provide functional information
- No amplification bias
- Expensive for large genomes
- More complex analysis & larger datasets

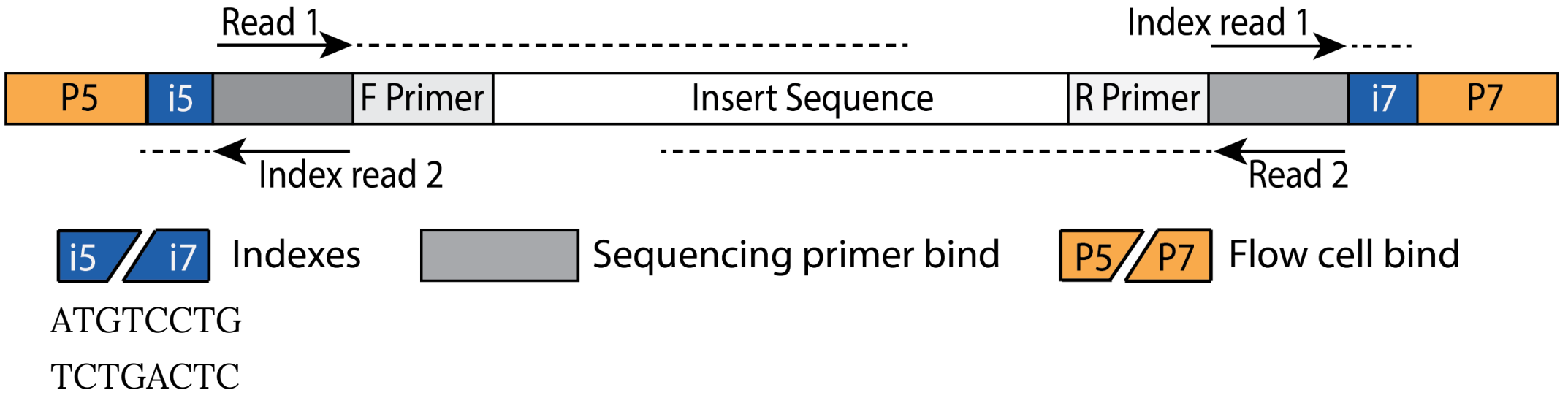
Mitochondrial enrichment (Mitogenomics)



- More diagnostic resolution for eukaryotes
- Less amplification bias
- More complex lab procedures

AGRICULTURE VICTORIA

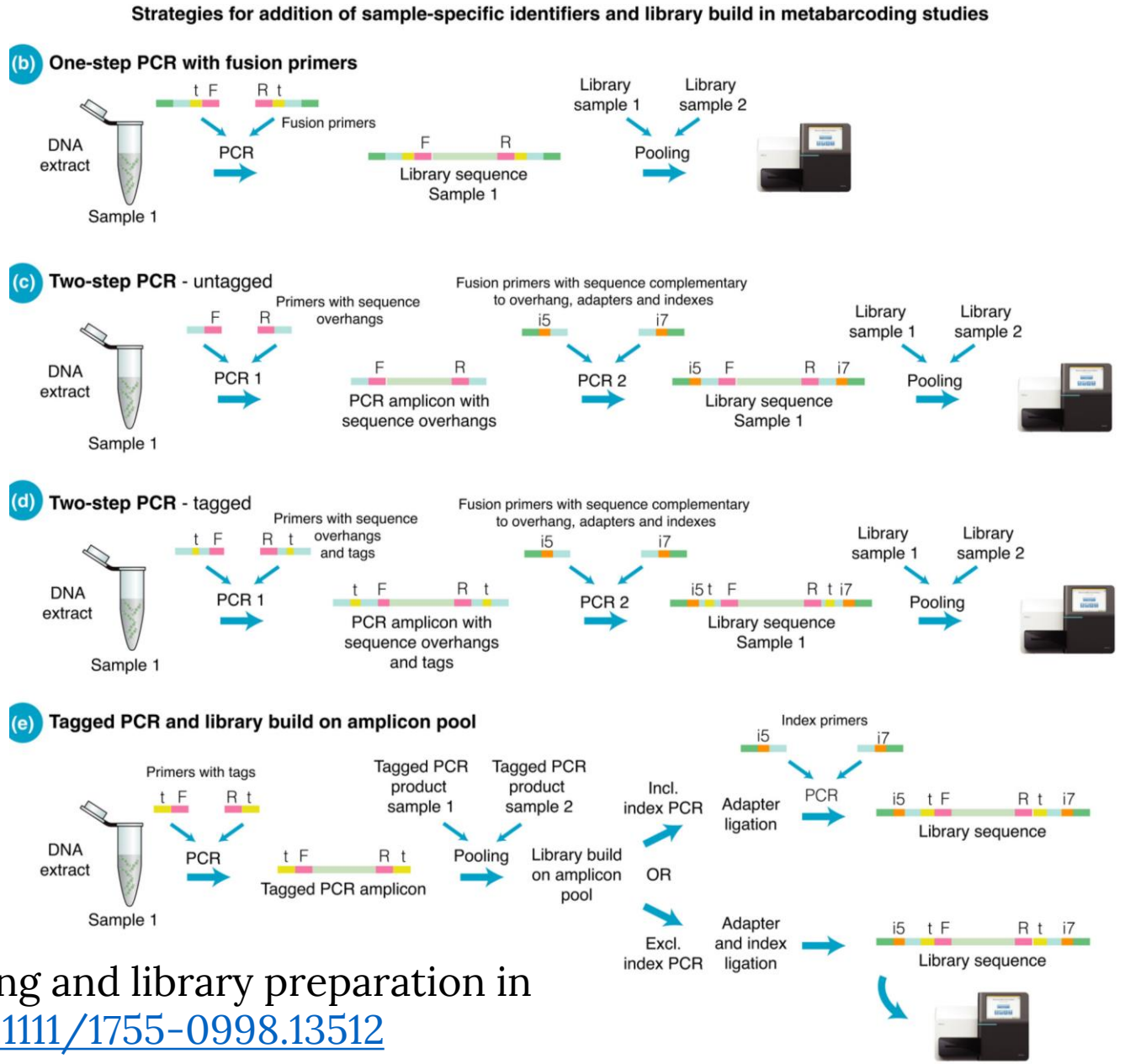
Library preparation (Illumina)



- Adapters are platform-specific nucleotide sequences that enable the target fragments to anchor to the sequencer and start the sequencing process.
- Unique dual indices are used for multiplexed runs to link each nucleic acid fragment to the sample it originated from

Library preparation

- Labelling samples earlier with indexes reduces risk of cross-contamination
- However, this requires much longer PCR primers (>80bp), reducing amplification efficiency and increasing bias
- 2-step PCR approaches use 2 sets of ~40bp primers instead
- Comes at the expense of an extra PCR step



Bohmann et al 2021 - Strategies for sample labelling and library preparation in DNA metabarcoding studies. https://doi.org/10.1111/1755-0998.13512

PCR 1

- Initial PCR enriches for target loci and adds partial Illumina adapter sequences – approx. 30 cycles
- PCRs are conducted in tandem rather than multiplexed
- This modular workflow means that alternative barcodes/primer sets can be integrated in future

fwhF2: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGDACWGGWTGAACWGTWTAYCCHCC-3'

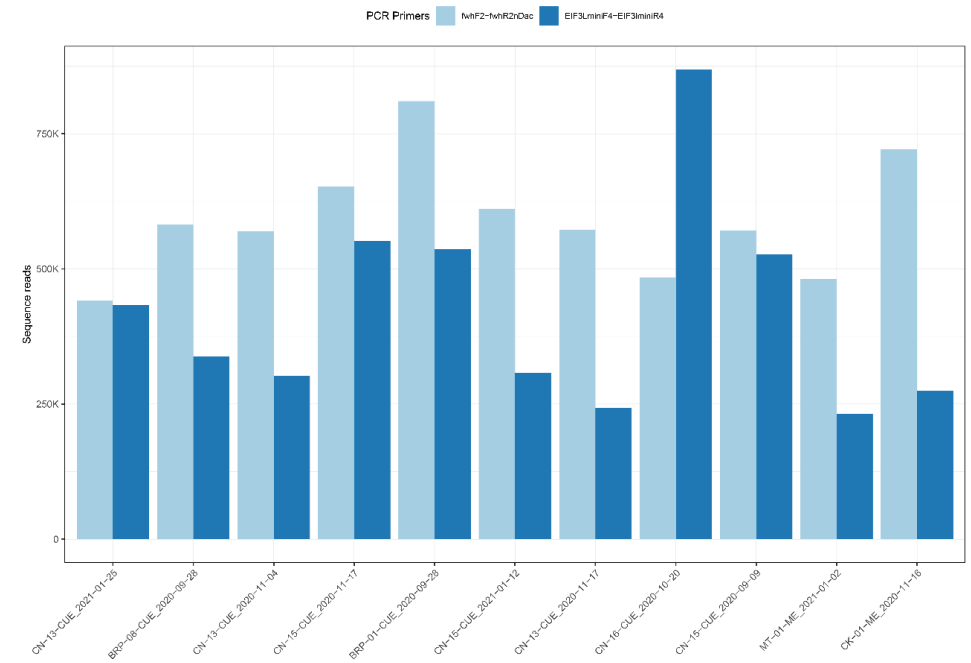
fwhR2nDac: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTRATWGCHCCIGCTAADACHGG-3'

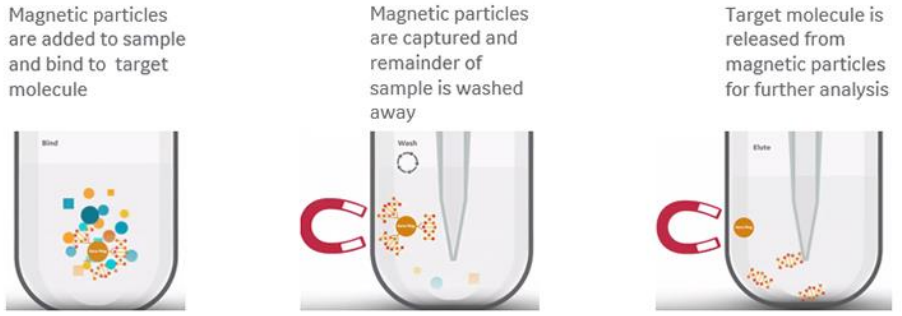
EIF3L_miniF4: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGCGYCGTTATGCYGATGC-3'

EIF3L_miniR4: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTRAAYACTTCYARATCRCC-3'

Normalise and combine barcodes

- Quantifying many samples for normalisation can be expensive and time consuming
- The quickest and most convenient methods (e.g. nanodrop) can be inaccurate
- The most accurate methods (e.g. qPCR, Qubit) are time consuming for many samples.
- Magnetic bead-based normalization is quick and high-throughput
- A specific volume of beads can bind a consistent quantity of DNA - if there are enough molecules in each library to saturate the beads, an equimolar quantity of library fragments will bind and be retained from each sample





PCR 2

- PCR 2 adds the remainder of the Illumina adapter sequences including 8bp unique-dual indexes
- Reactions are conducted in qPCR machine and stopped before plateau to prevent over-amplification artefacts

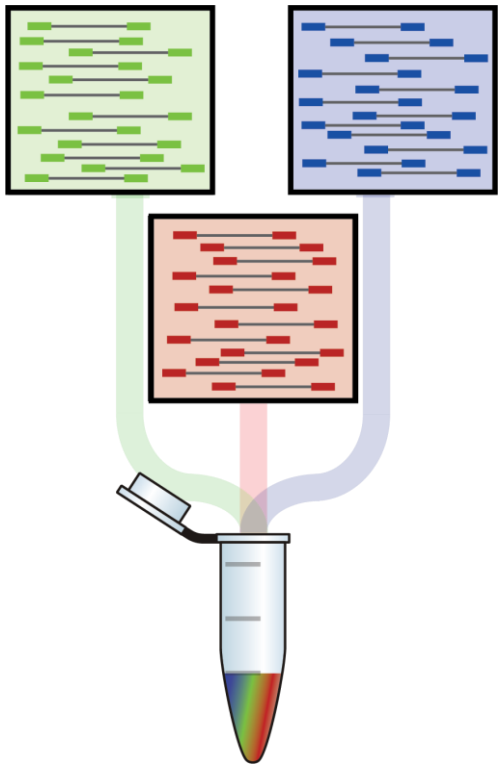
```
                                        ACACTCTTTCCCTACACGACGCTCTTCCGATCT-{barcode}
AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT

                                        GACTGGAGTTCAGACGTGTGCTCTTCCGATCT-{barcode}
CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```

AGRICULTURE VICTORIA

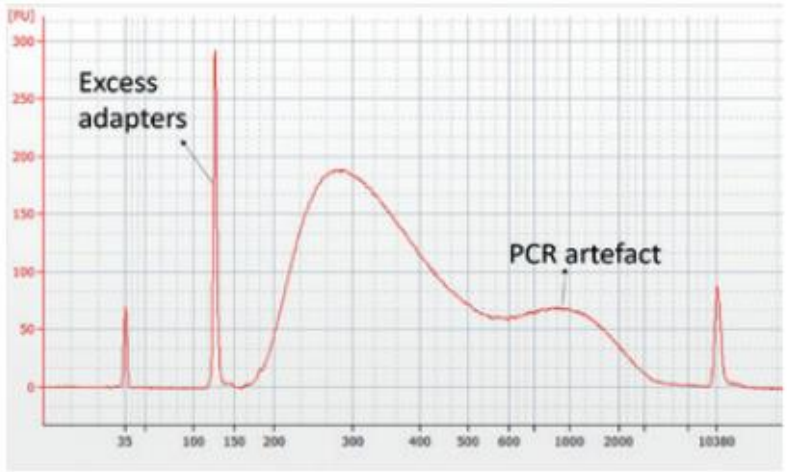
Normalise and pool libraries

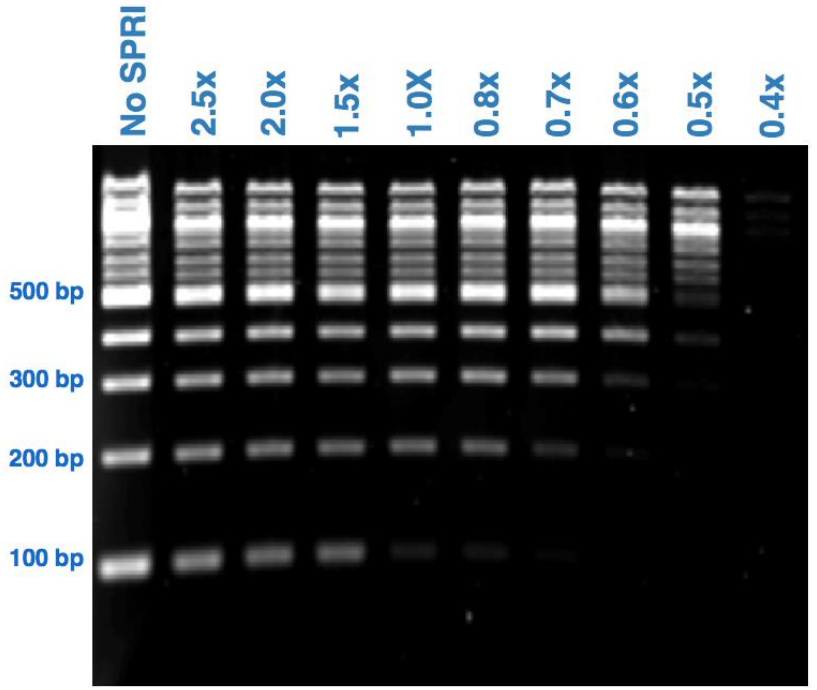
- Separately prepared and indexed libraries are pooled together into a single tube for sequencing
- Libraries need to be normalise dto balance sequencing effort between samples
- Again this step uses the SequelPrep normalisation plates



Library quality check

- Check pooled library for any excess adapters, primer dimers, or other artefacts
- Verify fragment sizes are correct using TapeStation or BioAnlyzer
- Quantify the libraries using TapeStation or Qubit
- If any adapters are present, libraries can be further cleaned using SPRI beads
- Libraries need to be at 2nM for sequencing protocol
- Libraries below 2nM can be concentrated using SPRI beads and eluting in smaller volumes





Sequencing platforms

Illumina MiSeq

- Most widely adopted platform
- 2x300bp max read length
- Requires ~50 samples to fill flow cell
- Expensive (>$50) per sample
- ~48hr runtime
- Expensive purchase (~$250k)

Illumina NovaSeq

- Very high throughput (hundreds of samples) 2x250bp max read length
- Very low cost per sample (<$10)
- Very expensive purchase (~$800k)

PacBio Sequel 2

- Moderate throughput
- Highly accurate long reads
- ~20kb max read length
- Low cost per sample ($20)
- Very expensive purchase (~$800k)

Oxford Nanopore MinION

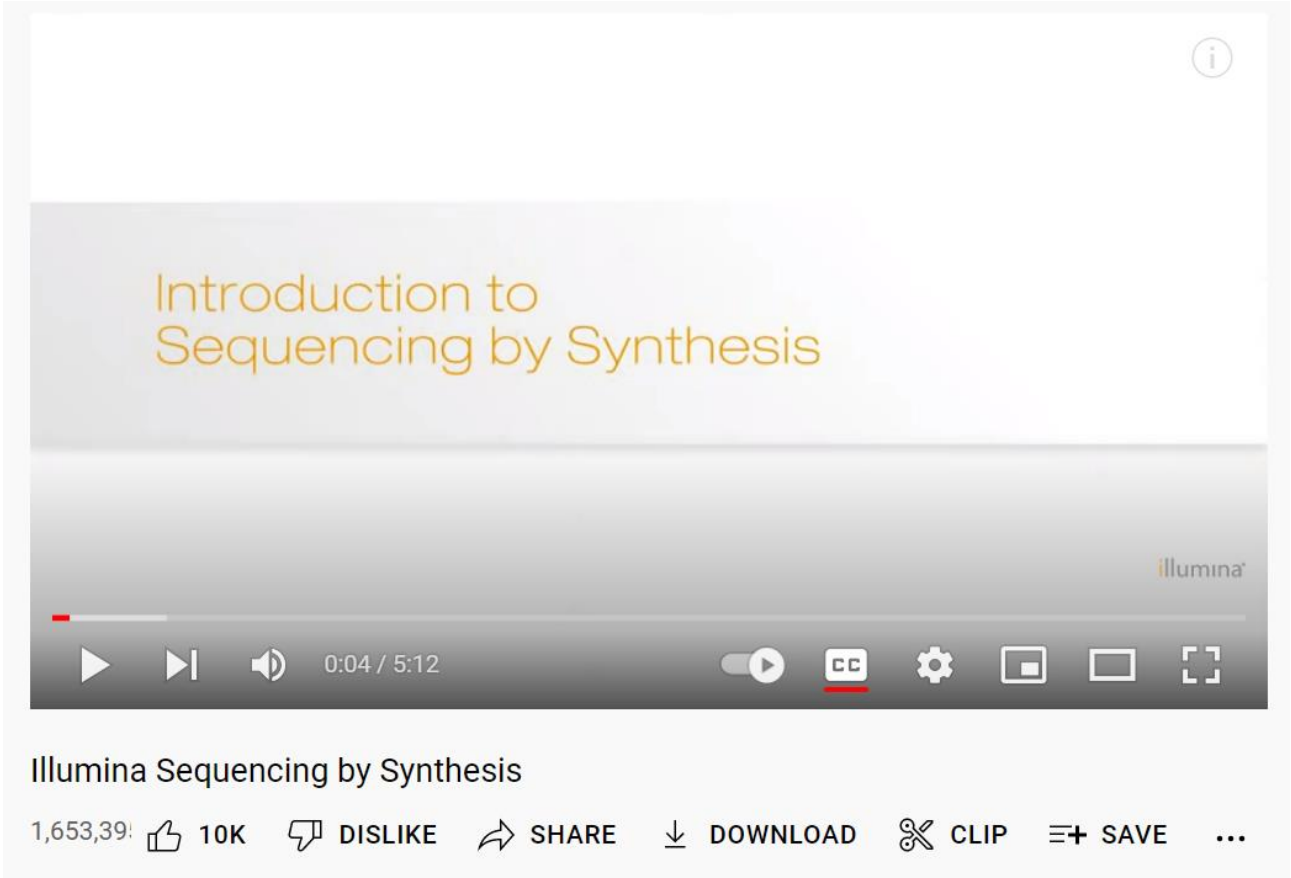
- Moderate throughput
- Very long reads (20kb+)
- High error rates that can exceed intraspecific differences
- Custom library preparation methods required to overcome errors
- Low cost per sample ($20)
- Very cheap purchase (~$1000)

Sample data sheet

	A	B	C	D	E	F	G	H	I	J
1	[Header]									
2	IEMFileVersion	4								
3	Investigator Name	Alexander Piper								
4	Project Name	Pathogens								
5	Experiment Name	K739J_tephritid_metabarcoding								
6	Date	23/05/2022								
7	Workflow	GenerateFASTQ								
8	Application	FASTQ Only								
9	Assay	Metabarcoding								
10	Description									
11	Chemistry	Amplicon								
12										
13	[Reads]									
14	251									
15	251									
16										
17	[Settings]									
18	Adapter	CTGTCTCTTATACACATCT								
19										
20	[Data]									
21	Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
22	Trap8	Trap8	1	A1	AVR_DUI_i7_001	GAGACGAT	AVR_DUI_i5_001	GTTCTCGT	Pathogens	
23	Trap7	Trap7	1	B1	AVR_DUI_i7_002	ACGGAACA	AVR_DUI_i5_002	CGCTCTAT	Pathogens	
24	Trap6	Trap6	1	C1	AVR_DUI_i7_003	CTTAGGAC	AVR_DUI_i5_003	TGGTAGCT	Pathogens	
25	Trap5	Trap5	1	D1	AVR_DUI_i7_004	TACGCCTT	AVR_DUI_i5_004	ACAGCTCA	Pathogens	
26	Trap4	Trap4	1	E1	AVR_DUI_i7_005	CTACTTGG	AVR_DUI_i5_005	GCAAGATC	Pathogens	
27	Trap3	Trap3	1	F1	AVR_DUI_i7_006	TGATACGC	AVR_DUI_i5_006	GATAGCGA	Pathogens	
28	Trap2	Trap2	1	G1	AVR_DUI_i7_007	TGCGTAGA	AVR_DUI_i5_007	TTGGTGAG	Pathogens	
29	Trap1	Trap1	1	H1	AVR_DUI_i7_008	ACTCGTTG	AVR_DUI_i5_008	CACCACTA	Pathogens	
30	Trap16	Trap16	1	A2	AVR_DUI_i7_009	ACTCTCGA	AVR_DUI_i5_009	AACCGTTC	Pathogens	

RICULTURE VICTORIA

Sequencing process



https://www.youtube.com/watch?v=fCd6B5HRaZ8

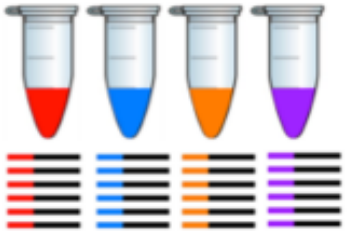
AGRICULTURE VICTORIA

Bioinformatics

AGRICULTURE VICTORIA

Demultiplexing

Barcoded per-sample



Pool and sequence samples

Track per-sample barcodes (e.g., in spreadsheet)

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAG
CGTACGTAGGCGGTTAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`[M]a`[VYa^_X^_
Z]NZ\`]TY\]_^RVH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB
```


VICTORIA

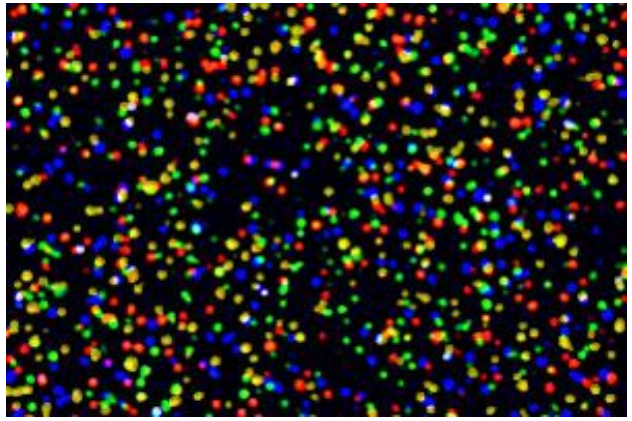
FastQ file format

- The header contains metadata about the sequencer, read number, lane, position of read on flow cell etc.
- Quality scores are associated with each base call depending on how accurately that base was determined
- These quality scores are ASCII encoded for data compression purposes

9872.fastq(.gz)

```
@M03633:513:000000000-JK3B3...
GGGACAGGTTGAACAGTATATCCACCT...
+
CCCCCCGGGGGCGGAFGCFCEFGDFGG...
@M03633:513:000000000-JK3B3...
GGGACTGGTTGAACTGTTTATCCTCCT...
+
CCCCCFGGGGGGGGDGGGGGEGGFGGG...
```

- Header
- Nucleotide Sequence
- Second Header (not used)
- Phred Quality Score (ASCII encoded)



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Bioinformatic pipelines

- A range of open-source and proprietary bioinformatic pipelines are available for processing metabarcoding data
- Many of these developed for comparing microbiome diversity
- Diagnostics generally has different goals – detecting target species or strains
- Important to validate that the chosen pipeline is fit-for-purpose before use in diagnostics
- May need to develop custom pipeline

Mothur

USEARCH

OBITools

JAMP

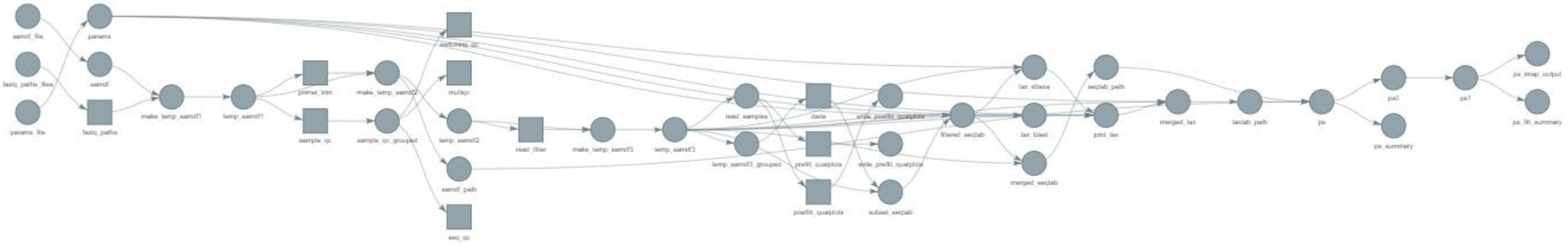
Just Another Metabarcoding Pipeline

mTAGs

PEMA

AGRICULTURE VICTORIA

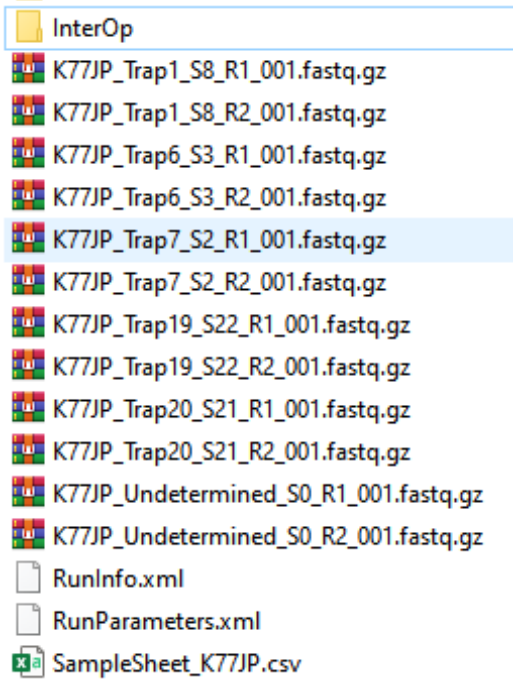
Our bioinformatics pipeline



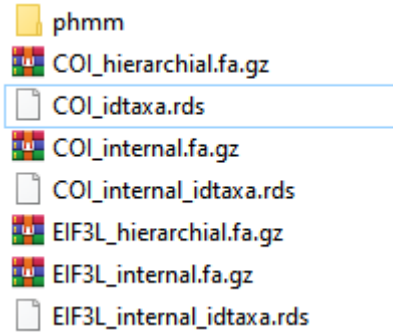
- Series of R and command line tools
- Based upon the DADA2 default pipeline
- Optimised for species level detection
- Implemented in the ‘targets’ automated pipeline framework
- Automatically tracks number of samples, and can resume from the previous step if interrupted

AGRICULTURE VICTORIA

Pipeline inputs



- Sequencing reads
- Sample sheet
- RunInfo
- RunParameters
- InterOp



- Trained models for IDTAXA
- Fasta files for BLAST
- PHMM models for cleaning ASVs

Locus parameters file

	A	B	C	D	E	F	G	H
1	pcr_primers	target_gene	phmm	ref_db	blast_db	exp_length	genetic_code	coding
2	fwhF2-fwhR2nDac	COI	reference/phmi	reference/COI_i	reference/COI_inte	205	SGC4	TRUE
3	EIF3LminiF4-EIF3lminiR4	EIF3L	reference/phmi	reference/EIF3L	reference/EIF3L_in	217	SGC0	TRUE

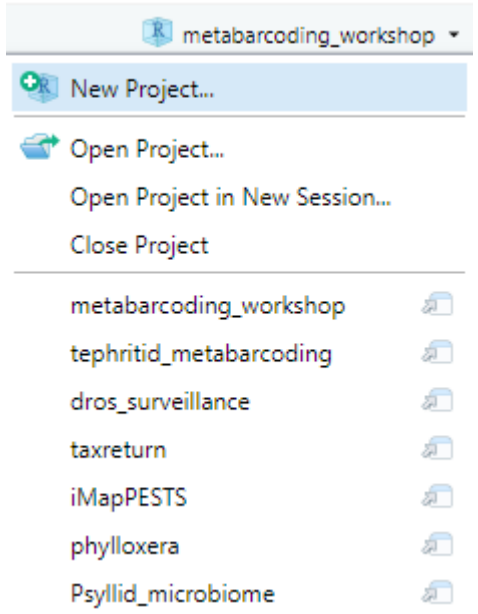
AGRICULTURE VICTORIA

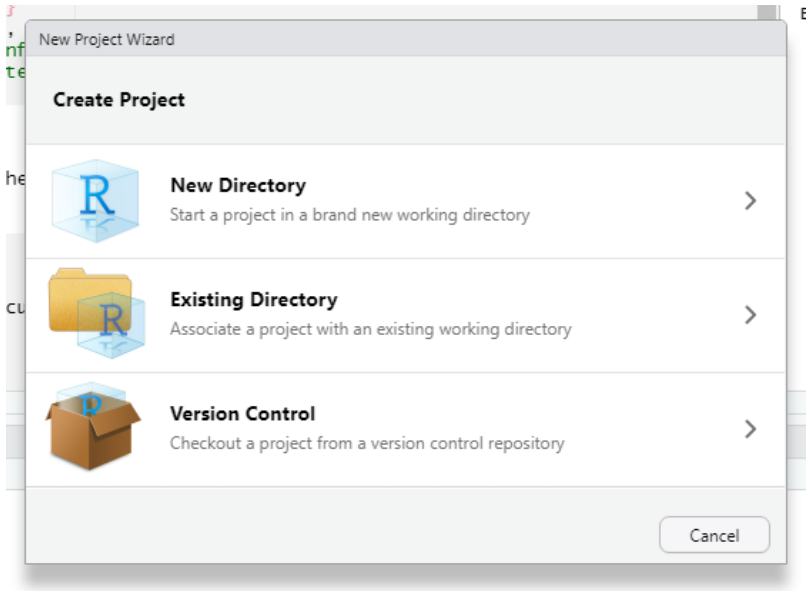
Bioinformatics materials

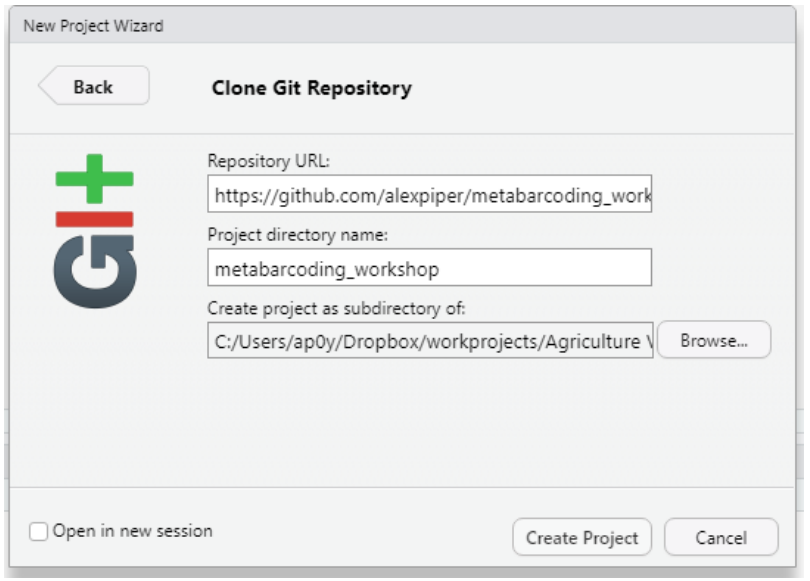
- Code and instructions: https://alexpiper.github.io/metabarcoding_workshop/bioinformatics_targets.html

- Data files: https://zenodo.org/record/7112162#.YzP2wddByck

Clone the github repository in Rstudio







AGRICULTURE VICTORIA

Barcode demultiplexing

- A second round of demultiplexing is conducted to split the samples into individual barcodes
- This uses the primer sequences as an ‘index’ to split the data

fwhF2: 5'–ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGDACWGGWTGAACWGTWTAYCCHCC–3’

fwhR2nDac: 5'–GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTRATWGCHCCIGCTAADACHGG–3'

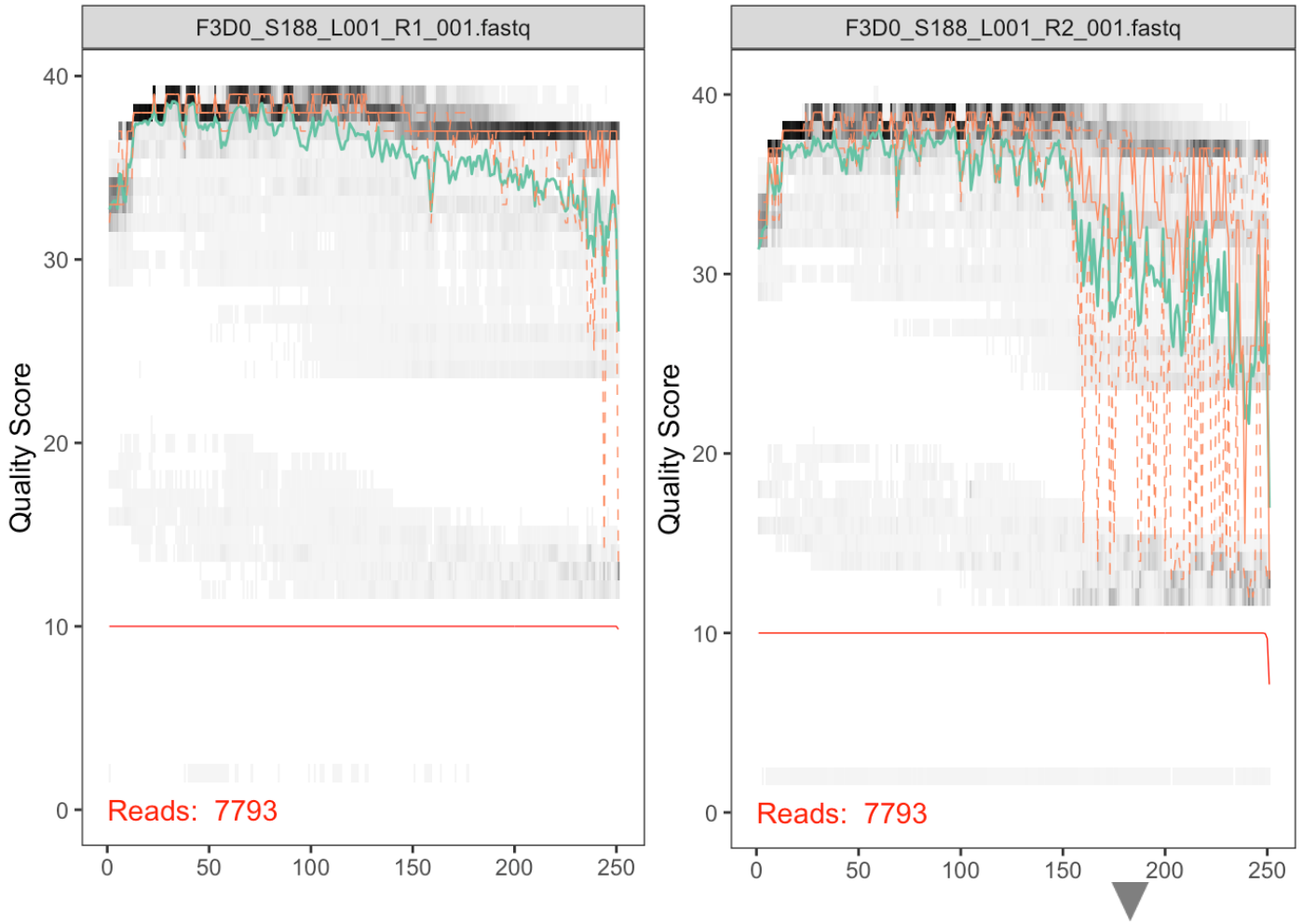
EIF3L_miniF4: 5'–ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGCGYCGTTATGCYGATGC–3'

EIF3L_miniR4: 5'–GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTRAAYACTTCYARATCRCC–3'

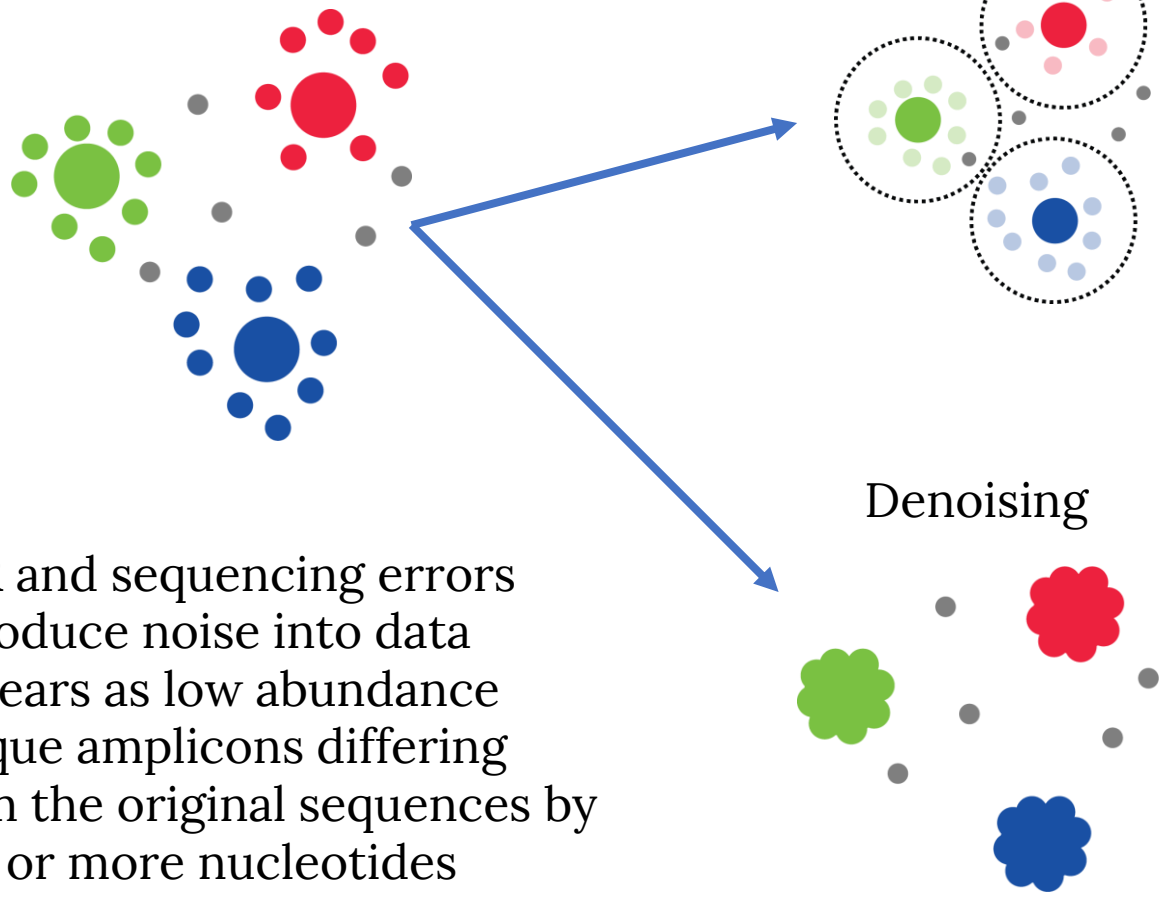
AGRICULTURE VICTORIA

Primer trimming & Quality filtering

- Remove non-biological nucleotides (i.e. PCR primers, adapters)
- Filter sequences by average quality
- Quality scores crash towards end of reads – particularly the reverse reads
- Reads can be trimmed of remove bad quality ends
- This will increase the amount of sequences passing average quality score filters – increasing retained data
- When trimming paired-end reads, maintain enough base overlap (~20bp) to accurately merge reads



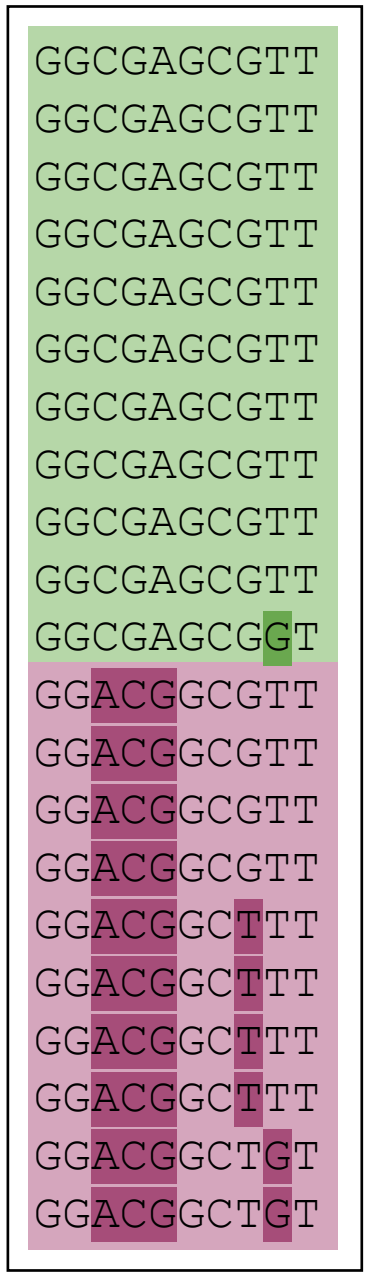
OTU tables



- PCR and sequencing errors introduce noise into data
- Appears as low abundance unique amplicons differing from the original sequences by one or more nucleotides

- Cluster sequences that fall within a fixed similarity threshold that approximate species (i.e. 97%)
- Creates molecular operational taxonomic units (mOTUs)
- Optimal threshold can vary across taxa
- Can result in over-clustering (grouping different species together in one cluster)
- Or under-clustering (splitting one species into separate clusters)

- Resolve sequencing errors from true variants using similarity, quality scores, and co-occurrence patterns
- Creates Amplicon Sequence Variants (ASVs)
- Single nucleotide resolution
- More reproducible (doesn't depend on what else is in the dataset)



90% OTU Clustering

OTU table

	OTU1	OTU2
Sample 1	100	79
Sample 2	88	35
Sample 3	86	51
Sample 4	12	87

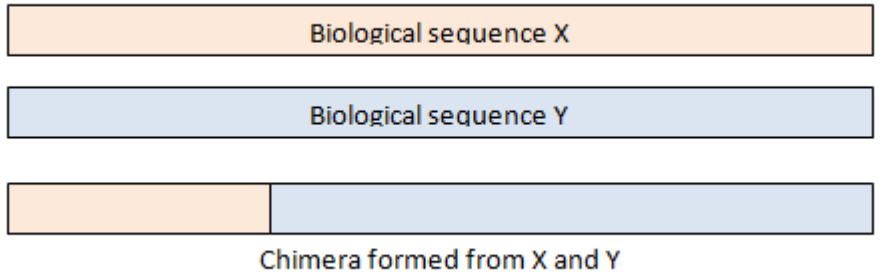
Denoised Sequence Variants
(100% OTU Clustering)

OTU table

	SV1	SV2	SV3	SV4	SV5
Sample 1	42	0	37	99	1
Sample 2	12	1	22	88	0
Sample 3	25	3	23	86	0
Sample 4	0	0	87	12	0

OTU/ASV curation

- PCR generated chimeras are those where one part of the sequences comes from one organism and the second part comes from a separate organism in the sample
- Chimeras are identified by aligning sequences and seeing if there are any lower-abundance sequences that can be made exactly by mixing left and right portions of two more abundant ones.
- Pseudogenes and off-target amplification can be removed by aligning to a reference sequence or 'profile' (i.e, hidden Markov model) for the target locus
- Expect to lose lots of unique sequences / OTUs, but the majority of reads should be retained



AGRICULTURE VICTORIA

Taxonomic assignment

OTU / ASV table

```
> OTU1
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
> OTU2
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
> OTU3
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
> OTU4
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
> OTU5
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

Reference sequence[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGC                                    CGGATGGACAAGTCTGATGTGAAAGG
CTGGGGCTC                                    ATCACTGGGCATAAAGCGCGCGTAGG
TGGCTTGGT
>referenc
TGGGTATGC
AGGCACTAT
TTGGTAAGT
>referenc
TGGGTATGC
AGGCACTAT
TTGGTAAGT
```

FeatureData[Taxonomy]

```
reference-sequence-1    Insecta; Diptera;
Tephritidae; Bactrocera; Bactrocera tryoni

reference-sequence-2    Insecta; Diptera;
Tephritidae; Bactrocera; Bactrocera melas

reference-sequence-3    Insecta; Diptera;
Tephritidae; Bactrocera; Bactrocera neohumeralis
```

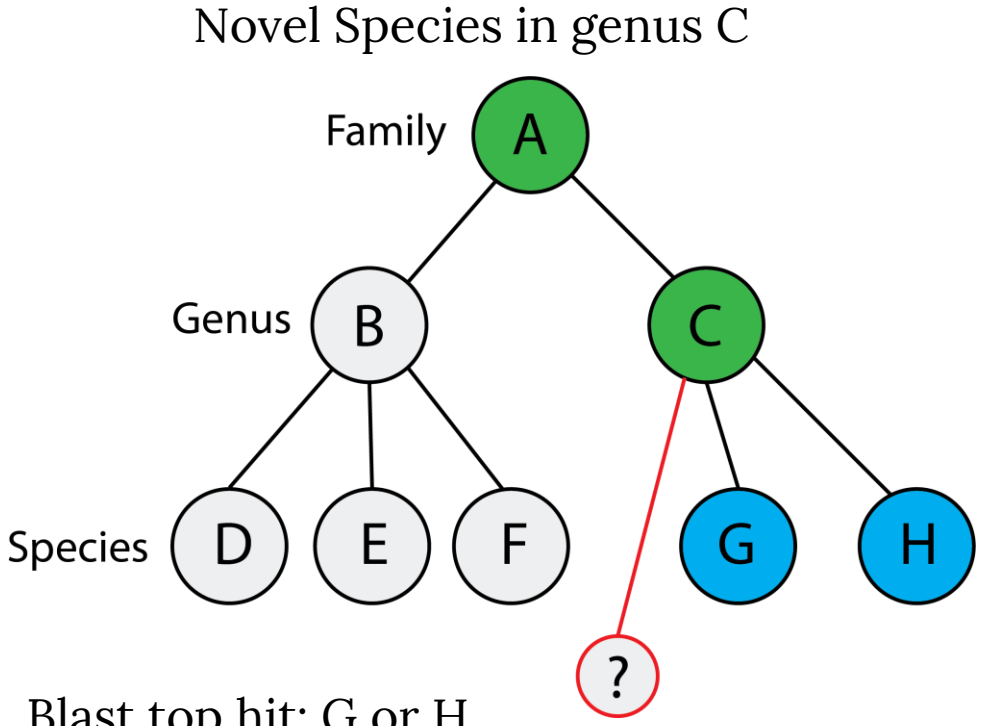
Compare observed sequences to annotated reference sequences to make taxonomic assignments.

Taxonomy table

```
OTU1     Insecta; Coleoptera
OTU2     Insecta; Diptera; Drosophilidae
OTU3     Insecta; Diptera; Tephritidae; Bactrocera; Bactrocera_tryoni
OTU4     Insecta; Diptera; Tephritidae; Bactrocera; Bactrocera_melas
OTU5     Insecta; Diptera; Tephritidae; Bactrocera;
```

Taxonomic assignment –why not just BLAST?

- ‘Top-hit’ searches such as BLAST only consider the ‘tips’ of the taxonomic tree
- Works well when species represented in reference database
- But risks over classification when species aren’t in reference database i.e. novel or un-sequenced taxon
- Tree-descent approaches “descend” the taxonomic tree until they cant pass confidence threshold (normally from bootstrap)



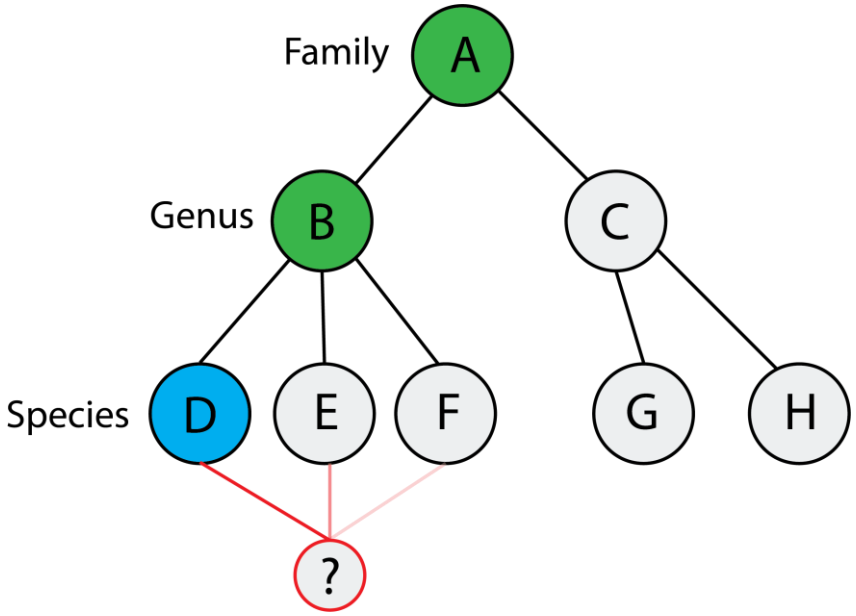
Blast top hit: G or H
Tree descent: Genus only assignment

Software suggestions: IDTAXA, QIIME2 Feature classifier, RDP classifier

Risk of under-classification

- Sometimes for closely related species complexes, tree descent approaches don't have enough confidence to assign to species level
- Combining multiple tools can improve accuracy
- i.e. using BLAST or exact sequence matching to get more identifications at the species level
- Accepting only assignments where the genus BLAST predicts matches the tree descent

Species D but closely related to D, E, F



Blast top hit: D
Tree descent: Not enough confidence for species, chooses genus B
Consensus: Species D

Iterative taxonomic assignment

- Sequences first assigned against in-house references from FruitFlyID alignments and melissas thesis
- Anything left unclassified to the species level then assigned against the curated public reference sequences

CCTGTCATCTATTATTGCTCACGGAGGAGCATCAGTTGATCTGGCTAT
TTTTTCTCTTCACTTAGCCGGTATTTCCTCAATTTTGGGAGCTGTTAAT

↓

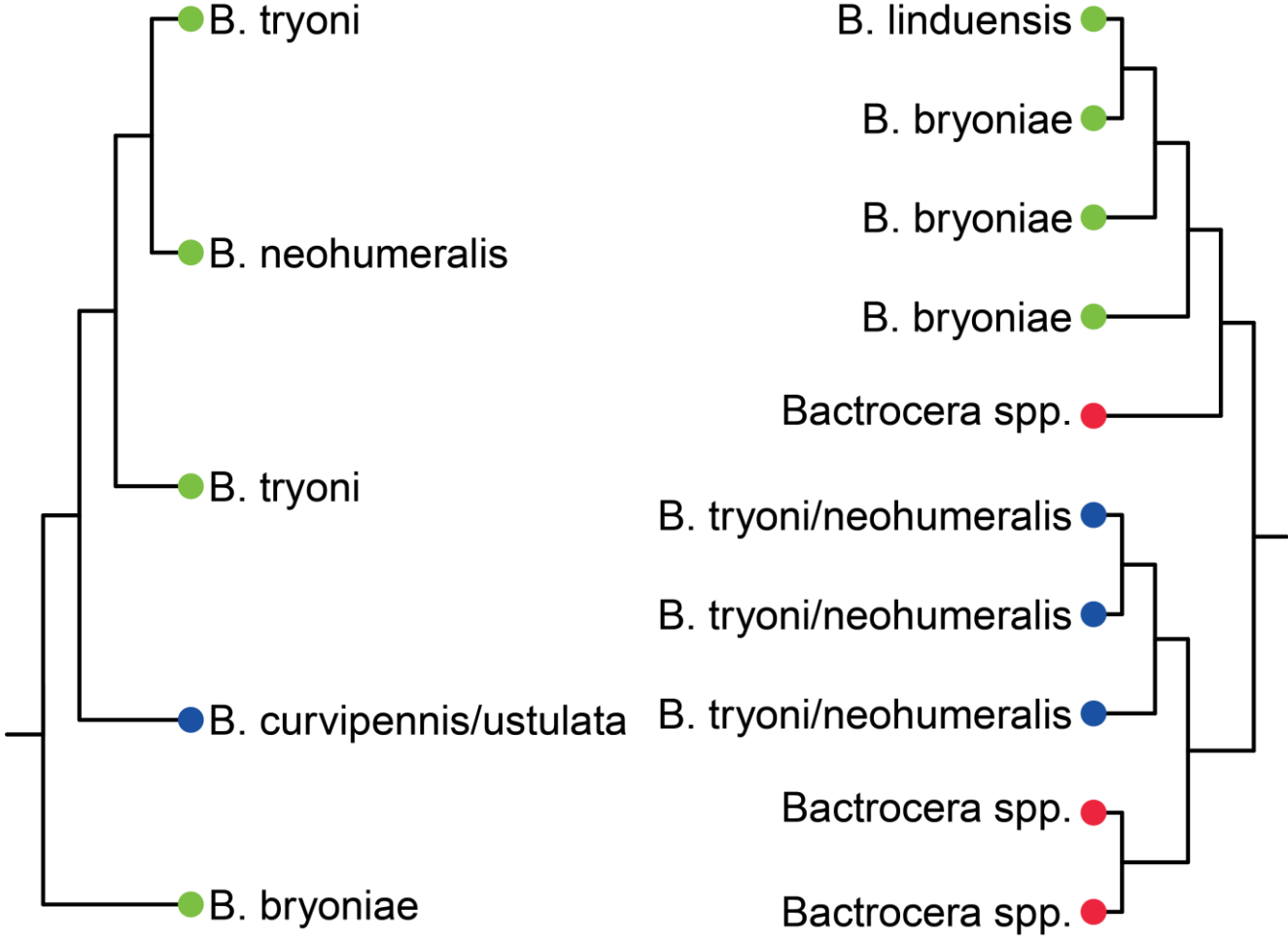
Diptera;Tephritidae;Bactrocera;NA

↓

Diptera;Tephritidae;Bactrocera;Bactrocera_aeroginosa

Multi-locus metabarcoding issues

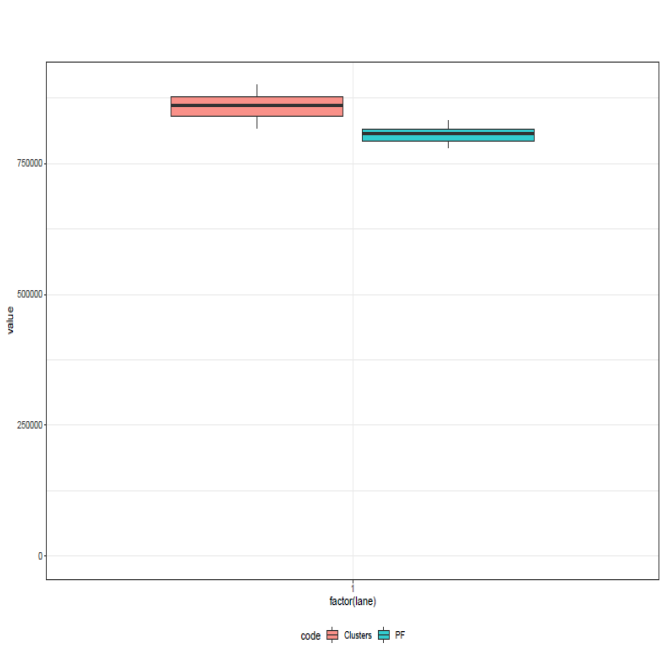
- No way to link distinct loci to an origin specimen in a mixed sample
- Loci cannot be simply concatenated to increase species resolution like single-specimen barcoding
- Difficult to integrate multiple loci when species lists are discordant due to different resolution/reference databases
- How to calculate a consensus abundance estimate in this case?
- Can expand scope of targets, and provide more confidence in species detection when a taxon is detected by more than one loci

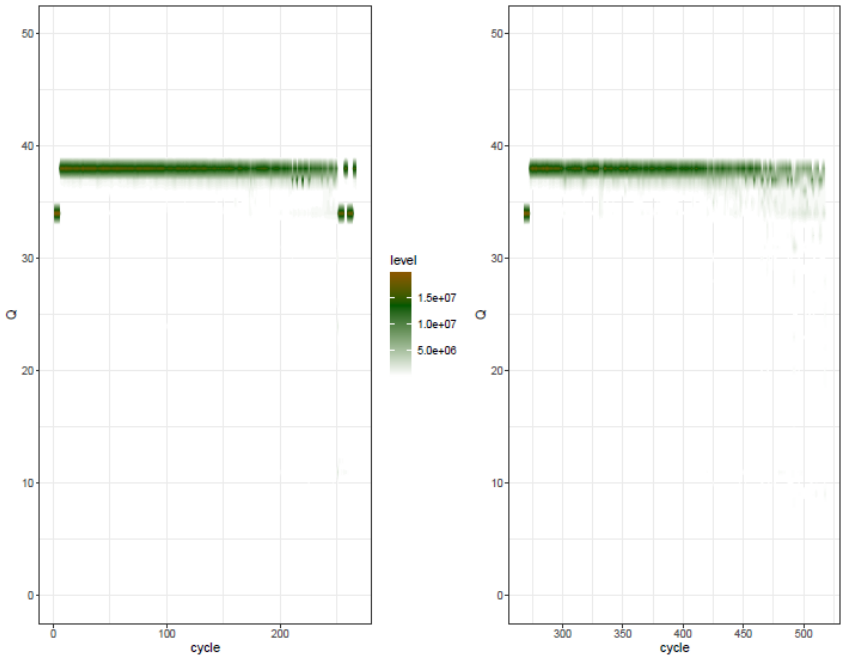


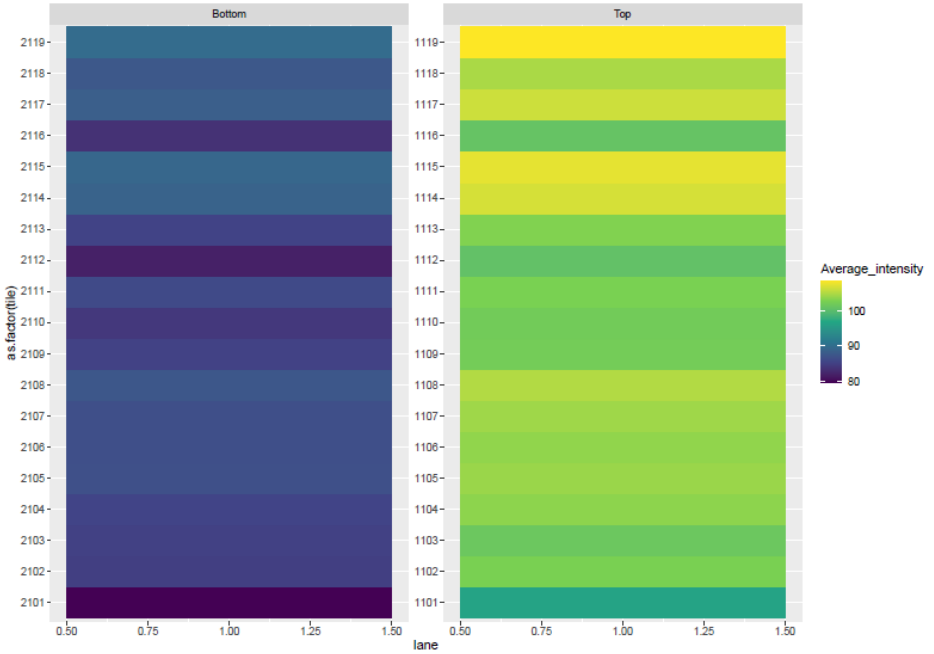
Quality control checkpoints

AGRICULTURE VICTORIA

Sequencing run quality





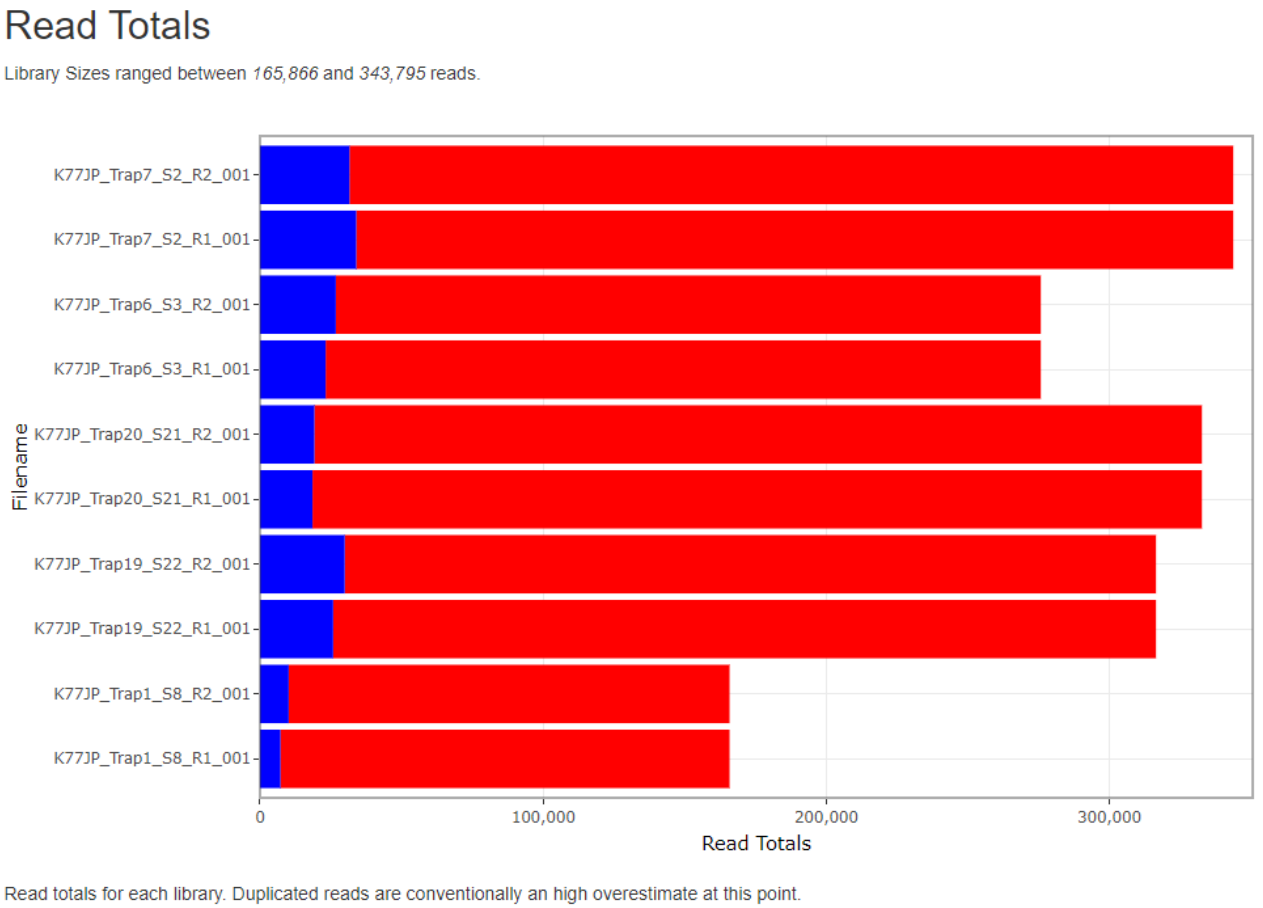


output/logs/fcid/PFclusters.pdf
output/logs/fcid/Qscore_L1.pdf
output/logs/fcid/avg_intensity.pdf

AGRICULTURE VICTORIA

Reads obtained per sample

- For amplicon sequencing, lots of ‘duplicated reads’ are expected

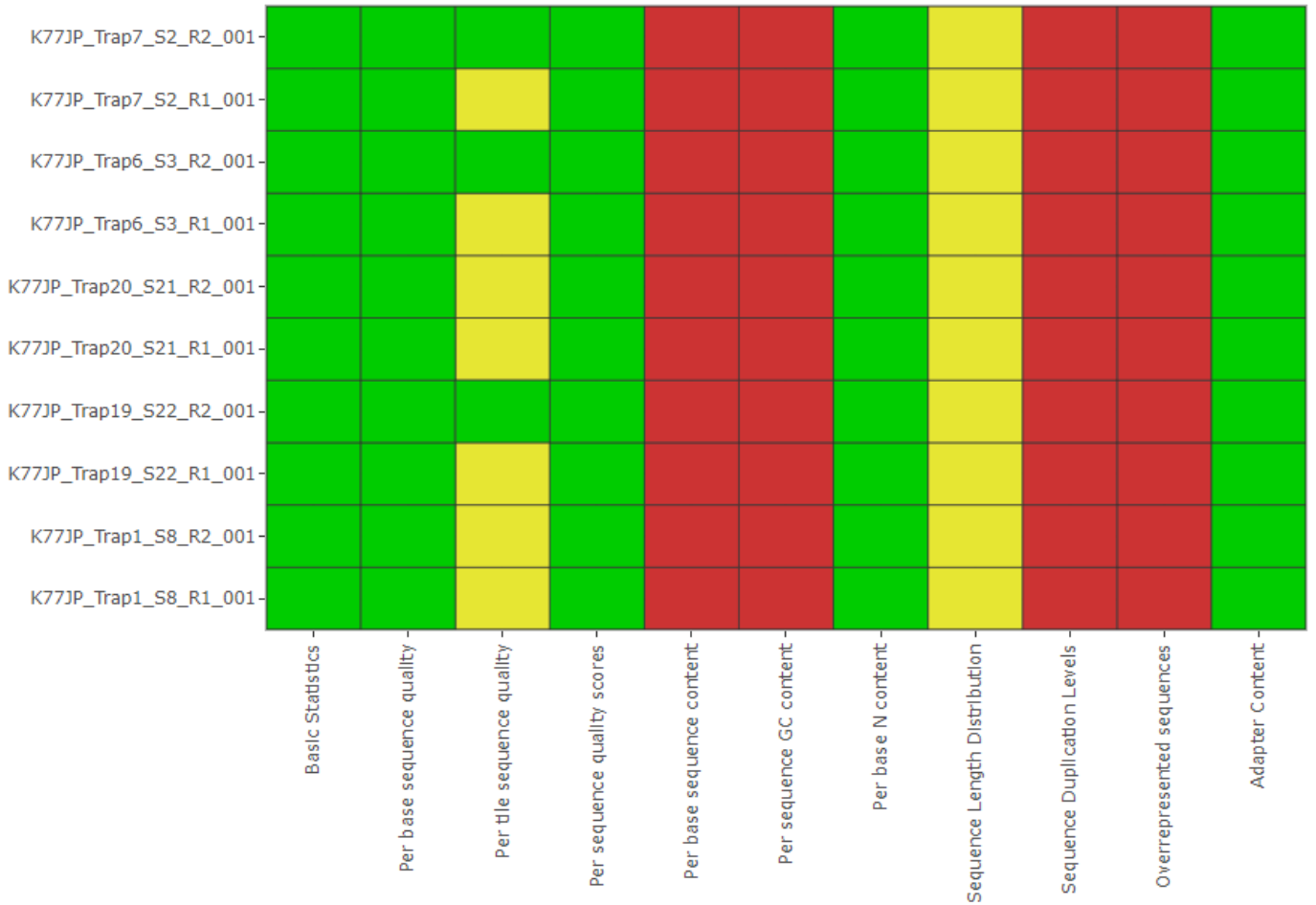


output/logs/fcid/ngsReports_Fastqc.html

AGRICULTURE VICTORIA

Sample quality overview

- Basic statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content (Ignore)
- Per sequence GC content (Ignore)
- Per base N content
- Sequence length distribution
- Sequence duplication levels (Ignore)
- Overrepresented sequences (Ignore)
- Adapter content

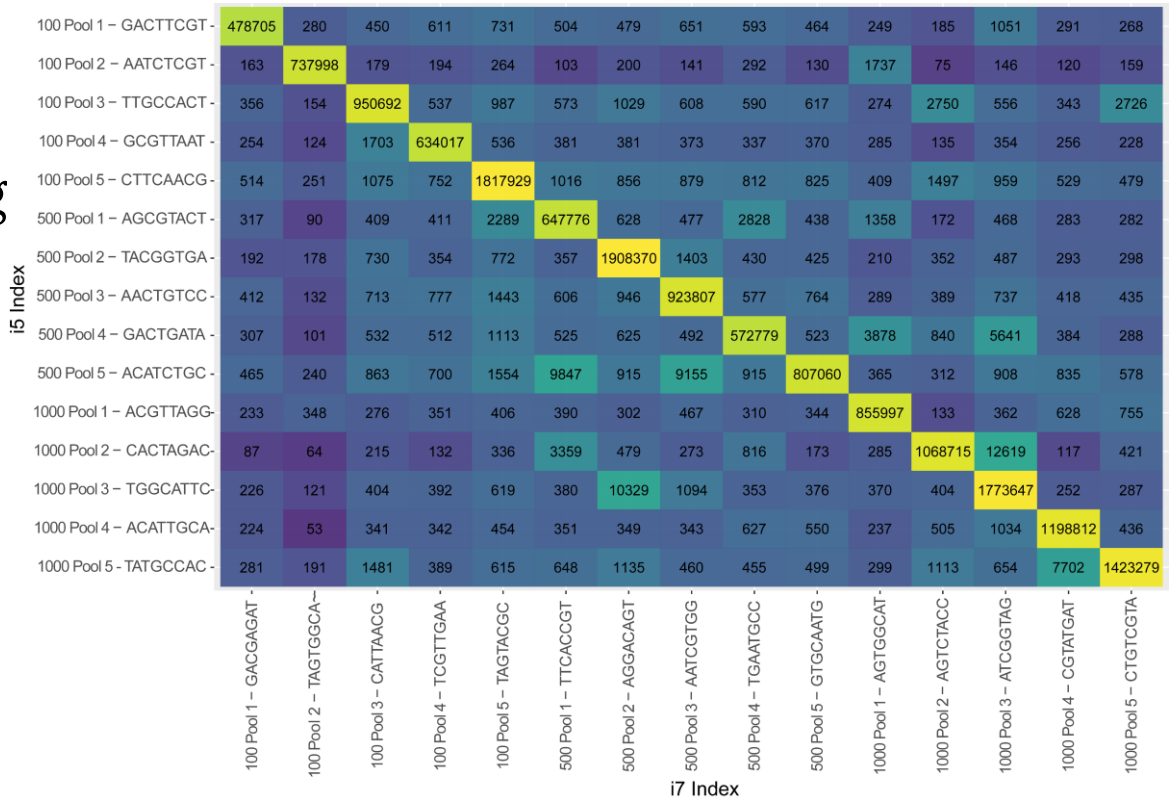


output/logs/fcid/ngsReports_Fastqc.html

Per sample plots: output/logs/fcid/FASTQC

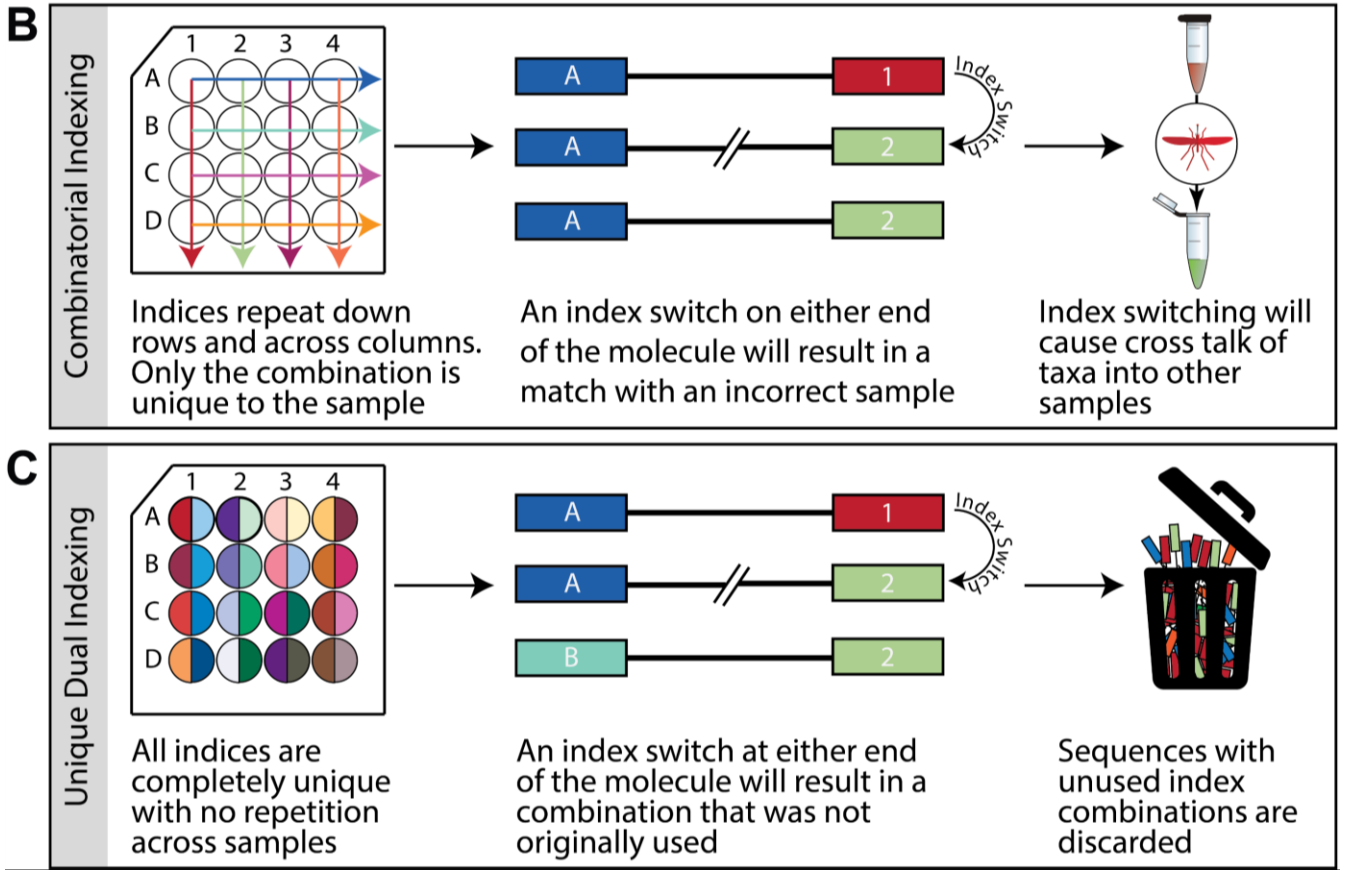
Index switch rate

- Metabarcoding is a highly sensitive technique and prone to detecting contaminants
- Index-switching occurs due to recombination between indexed molecules during sequencing
- Cross-contamination will only result in false positives for target taxa if that taxa is present in another sample at high abundance
- Index-switching can be measured and filtered using positive controls, or the ratio of applied index combinations (correctly demultiplexed reads), to unapplied combinations (found in undetermined reads file)
- Selecting a filtering threshold is a tradeoff between false-positives and false-negatives - 0.01% is a good heuristic



output/logs/fcid/index_switching.pdf

Combinatorial vs Dual unique Indexing



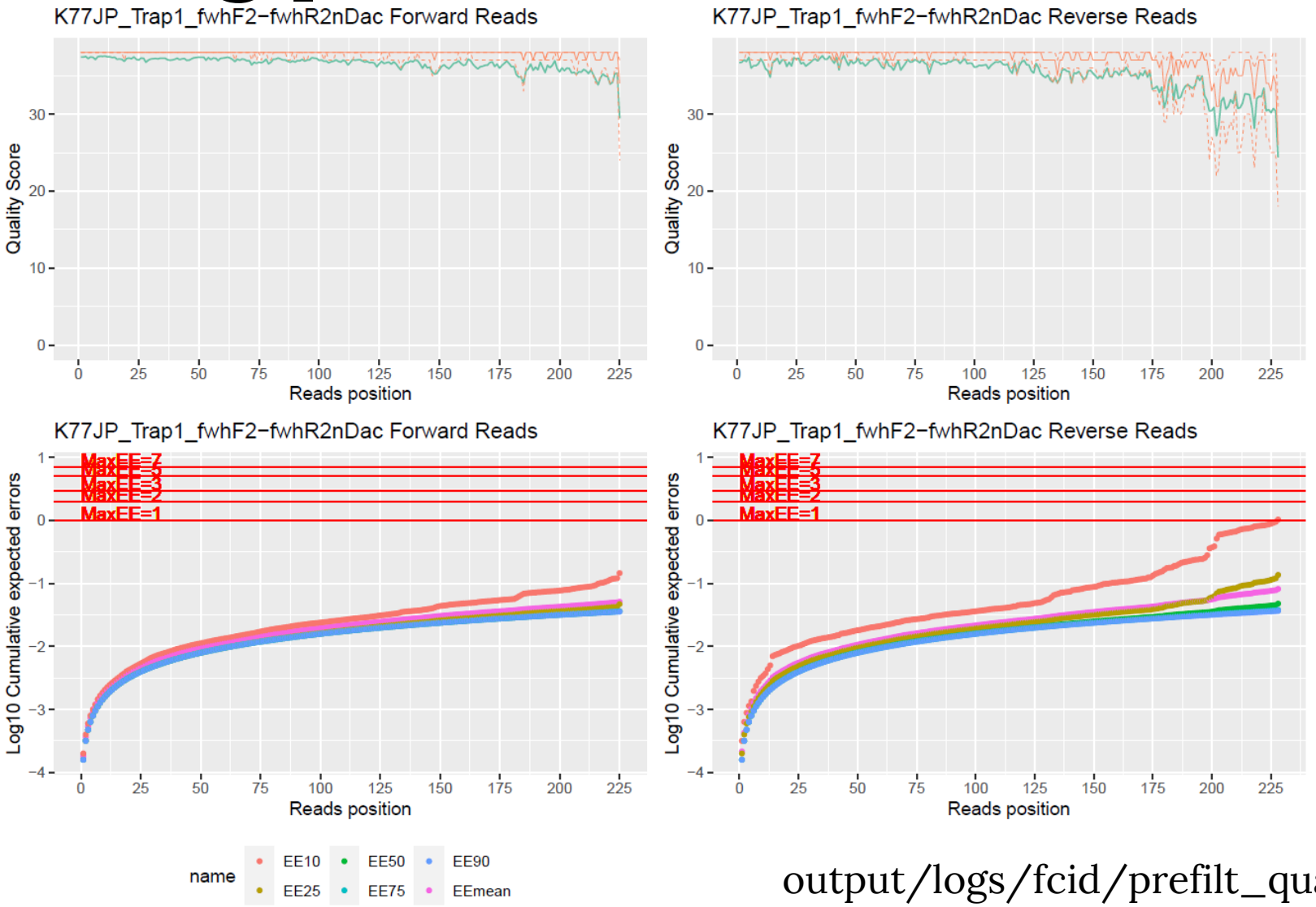
Combinatorial Indexing				
Pool 1	**Pool 2**	**Pool 3**	**Pool 4**	**Pool 5**
854981	599	656838	683030	501069
1710	13	85731	227	9
0	0	0	0	0
44	2165	7404	5763	46
1071	43316	24814	106802	265978
928	866820	85935	58857	98242

Unique Dual indexing				
Pool 1	**Pool 2**	**Pool 3**	**Pool 4**	**Pool 5**
848865	5	662030	677829	506701
1210	0	80686	88	0
0	0	0	0	0
0	2145	6760	6049	42
1	41758	24353	107756	262309
20	867881	79300	61442	95914

Index switching can be reduced further using unique molecular identifiers (labels each molecule amplified, in addition to each sample)

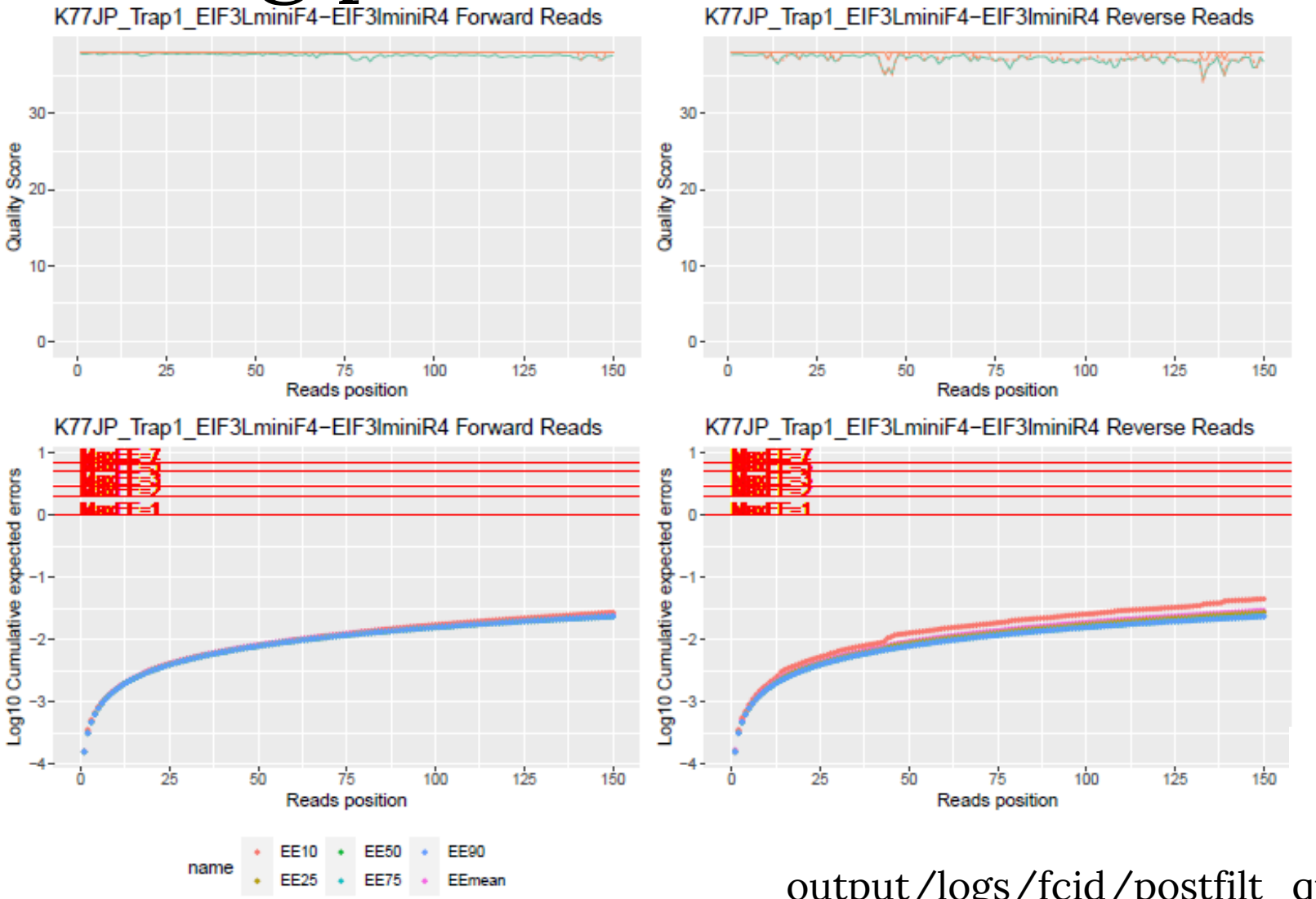
AGRICULTURE VICTORIA

Pre filtering plots



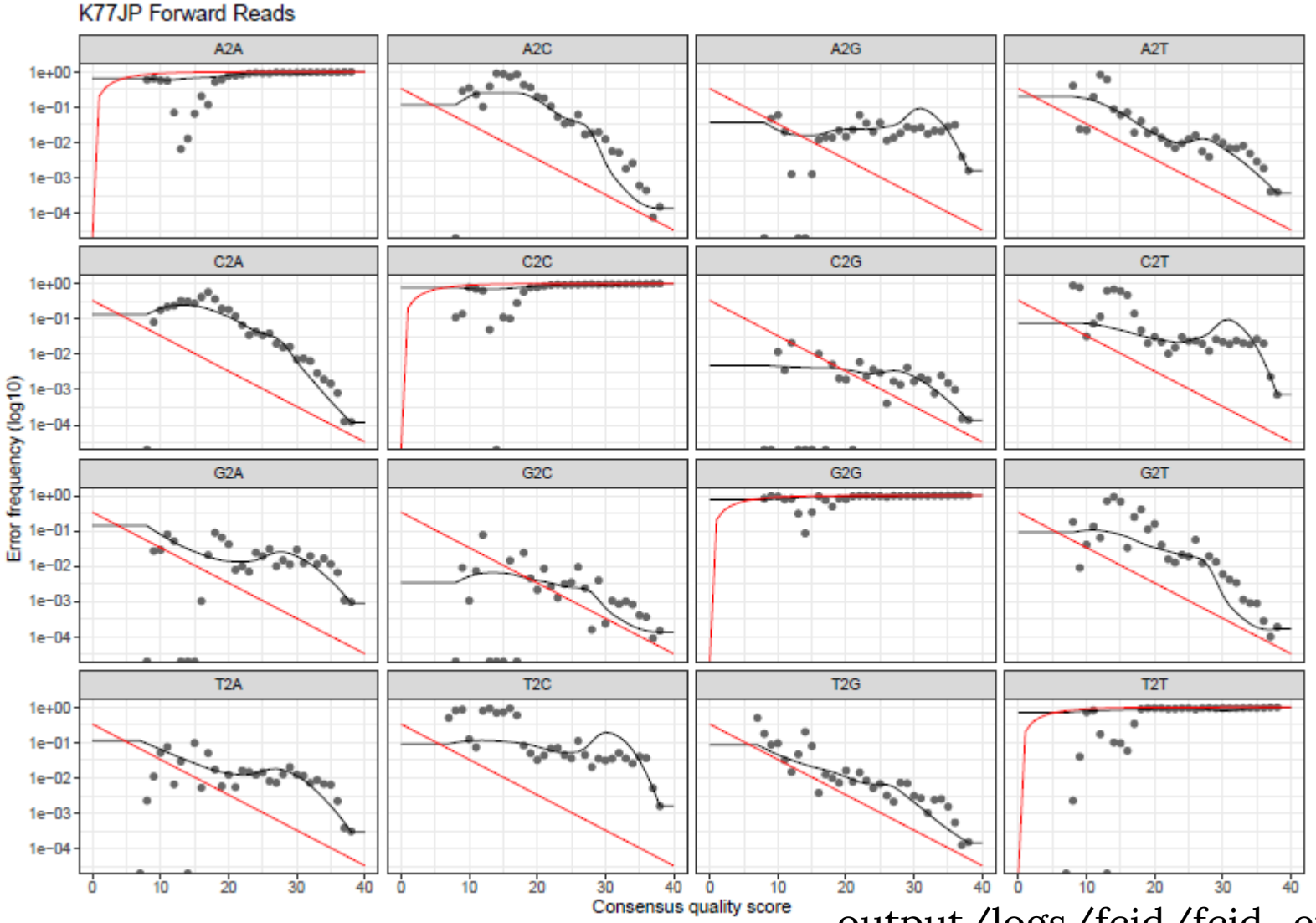
output/logs/fcid/prefilt_qualplots.pdf

Post filtering plots



output/logs/fcid/postfilt_qualplots.pdf

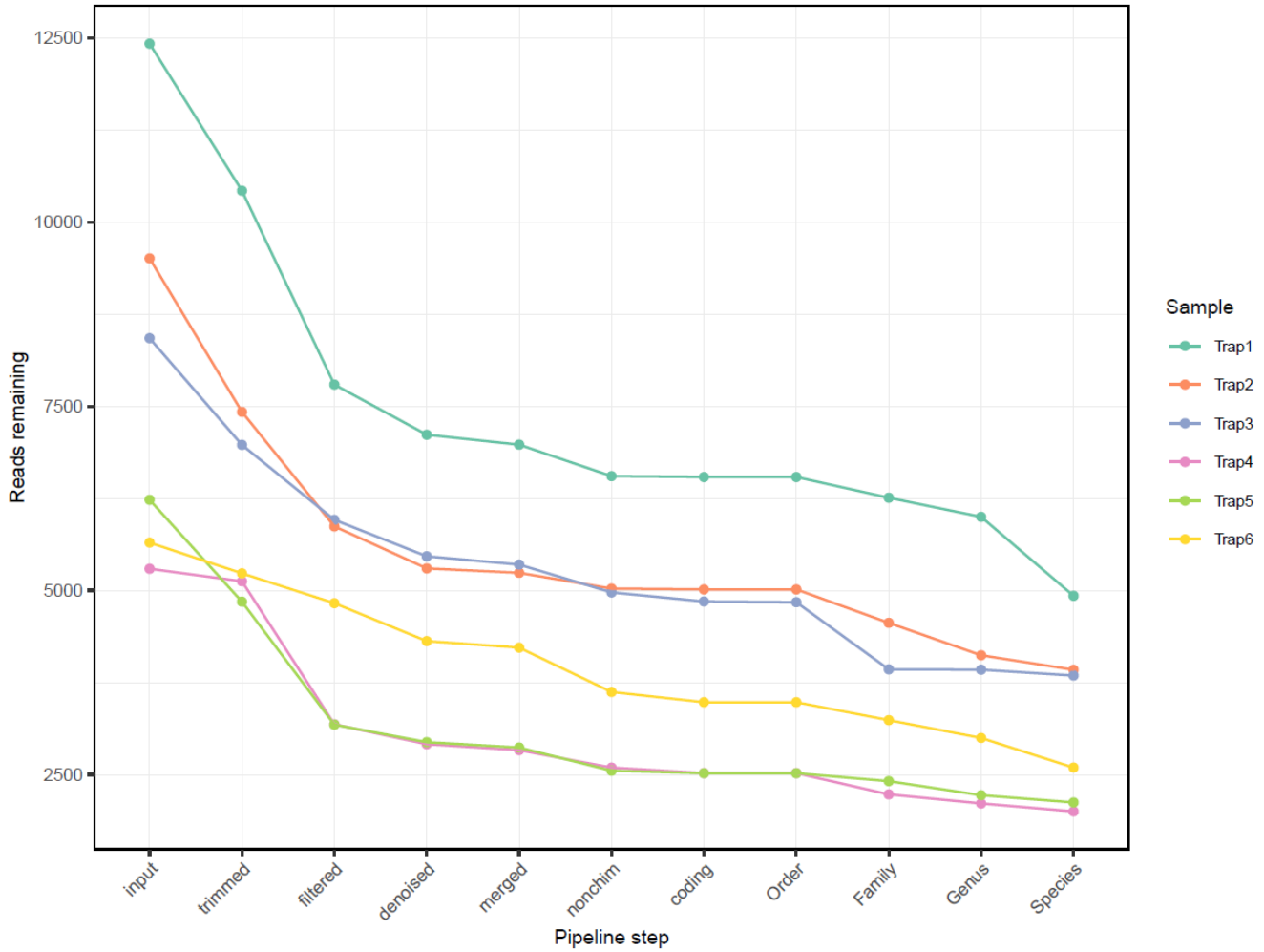
DADA2 Error model



output/logs/fcid/fcid_errormodel.pdf

AGRICULTURE VICTORIA

Number of reads passing through pipeline



Not currently working ☹

AGRICULTURE VICTORIA

Outputs

AGRICULTURE VICTORIA

ASV table

	A	B	C	D	E	F	G	H	I	J	K	L
1	sample_id	TATTCGCA	TATTCGCA	TATTCGCA	TATTCGCA	CCTATCGT	ACTATCAT	CCTATCGT	GCTATCGT	TATTCGCA	TATTCGCA	TATTCGCA T
2	K77JP_Trap1_fwhF2-fwhR2nDac	0	0	0	0	361	0	372	239	0	0	0
3	K77JP_Trap19_fwhF2-fwhR2nDac	0	0	0	0	4765	20217	3151	1776	0	0	0
4	K77JP_Trap20_fwhF2-fwhR2nDac	0	0	0	0	32298	0	21473	530	0	0	0
5	K77JP_Trap6_fwhF2-fwhR2nDac	0	0	0	0	2546	6677	6042	23809	0	0	0
6	K77JP_Trap7_fwhF2-fwhR2nDac	0	0	0	0	8960	17356	8941	11109	0	0	0
7	K77JP_Trap1_EIF3LminiF4-EIF3lminiR4	28188	57022	141	6233	0	0	0	0	5995	666	2419
8	K77JP_Trap19_EIF3LminiF4-EIF3lminiR4	48825	9174	3586	36046	0	0	0	0	14978	2879	14431
9	K77JP_Trap20_EIF3LminiF4-EIF3lminiR4	77167	49382	1404	0	0	0	0	0	0	20426	0
10	K77JP_Trap6_EIF3LminiF4-EIF3lminiR4	29200	6779	45554	14125	0	0	0	0	4883	4233	7233
11	K77JP_Trap7_EIF3LminiF4-EIF3lminiR4	33015	8710	34061	18569	0	0	0	0	6561	3063	5202

output/results/final/seqtab.csv

AGRICULTURE VICTORIA

Taxonomy table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	OTU	Root	Kingdom	Phylum	Class	Order	Family	Genus	Species					
2	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera albistrigata/aquilonis/caledoniensis/neohumeralis					
3	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera erubescentis/ustulata					
4	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera bryoniae					
5	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera frauenfeldi					
6	CCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera neohumeralis					
7	ACTATCAT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera frauenfeldi					
8	CCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera aquilonis/melas/neohumeralis/tryoni					
9	GCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera bryoniae					
10	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera frauenfeldi					
11	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera aquilonis					
12	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera albistrigata/aquilonis/caledoniensis/neohumeralis					
13	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera tryoni					
14	ACTATCAT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera frauenfeldi					
15	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera albistrigata/aquilonis/caledoniensis/neohumeralis					
16	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera neohumeralis/tryoni					
17	GCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera bryoniae					
18	TATTCGCA	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera bryoniae					
19	ACTATCAT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera frauenfeldi					
20	CCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera melas/tryoni					
21	CCTATCGT	Root	Metazoa	Arthropoc	Insecta	Diptera	Tephritida	Bactrocera	Bactrocera tryoni					

output/results/final/taxtab.csv

AGRICULTURE VICTORIA

Sample data table

	A	B	C	D	E	F	G	H	I	J	K	L
1	sample_id	sample_n	extraction	amp_rep	client_nar	experime	sample_ty	collection	collection	lat_lon	environm	collection
2	K77JP_Trap1_fwhF2-fwhR2nDac	Trap1	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
3	K77JP_Trap19_fwhF2-fwhR2nDac	Trap19	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
4	K77JP_Trap20_fwhF2-fwhR2nDac	Trap20	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
5	K77JP_Trap6_fwhF2-fwhR2nDac	Trap6	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
6	K77JP_Trap7_fwhF2-fwhR2nDac	Trap7	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
7	K77JP_Trap1_EIF3LminiF4-EIF3lminiR4	Trap1	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
8	K77JP_Trap19_EIF3LminiF4-EIF3lminiR4	Trap19	NA	NA	Pathogen	K739J_tep	NA	NA	NA	NA	NA	NA
9	K77JP_Trap20_EIF3LminiF4-EIF3lminiR4	Trap20	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
10	K77JP_Trap6_EIF3LminiF4-EIF3lminiR4	Trap6	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA
11	K77JP_Trap7_EIF3LminiF4-EIF3lminiR4	Trap7	NA	NA	Pathogen:	K739J_tep	NA	NA	NA	NA	NA	NA

output/results/final/samdf.csv

AGRICULTURE VICTORIA

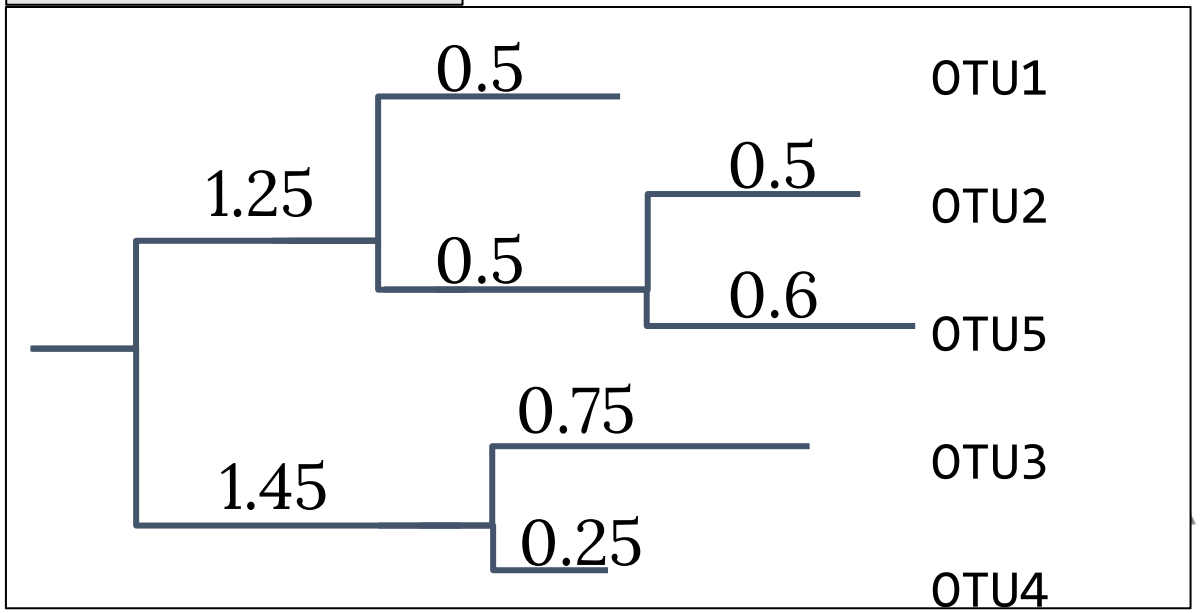
Phylogenetic trees (Not included)

OTU table

```
> OTU1
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
> OTU2
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
> OTU3
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
> OTU4
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
> OTU5
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

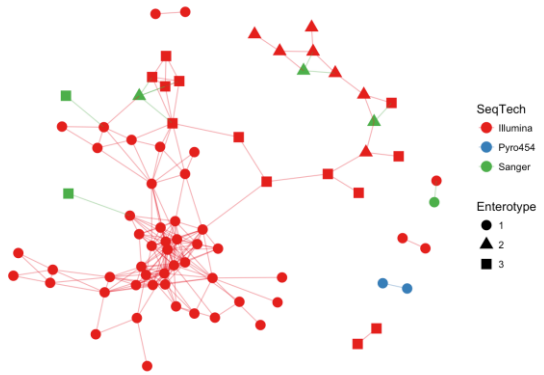
Align sequences, filter highly variable (i.e., randomly evolving) positions, and build phylogenetic tree.

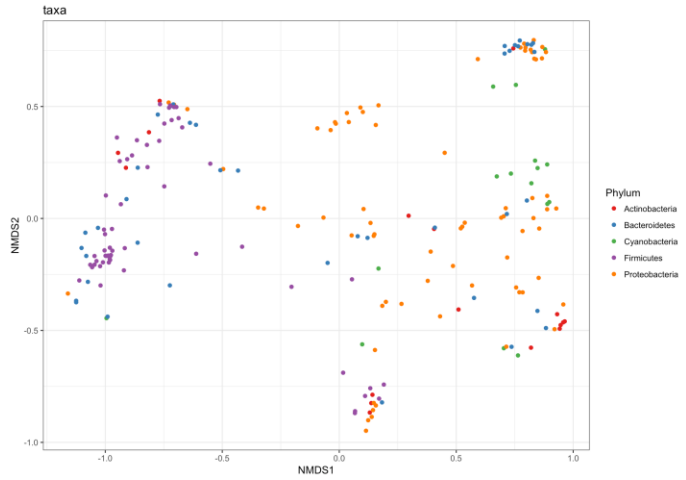
Phylogeny[Rooted]

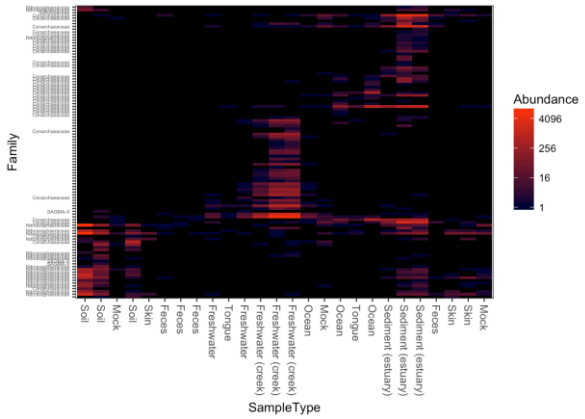


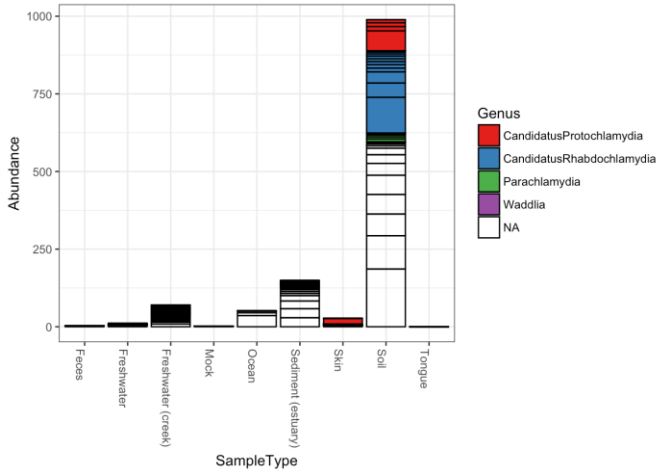
Phyloseq

- The phyloseq R package is a tool to import, store, analyze, and graphically display metabarcoding data
- Uses the 3 tables generated above along with an optional phylogenetic tree



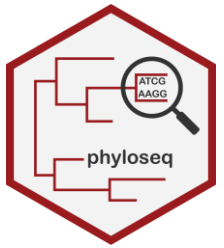







```
> ps <- readRDS("output/rds/ps_filtered.rds")
> ps
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 175 taxa and 10 samples ]:
sample_data() Sample Data:        [ 10 samples by 36 sample variables ]:
tax_table()   Taxonomy Table:     [ 175 taxa by 8 taxonomic ranks ]:
refseq()      DNAStringSet:       [ 175 reference sequences ]
taxa are columns
>
```

output/rds/ps_filtered.rds

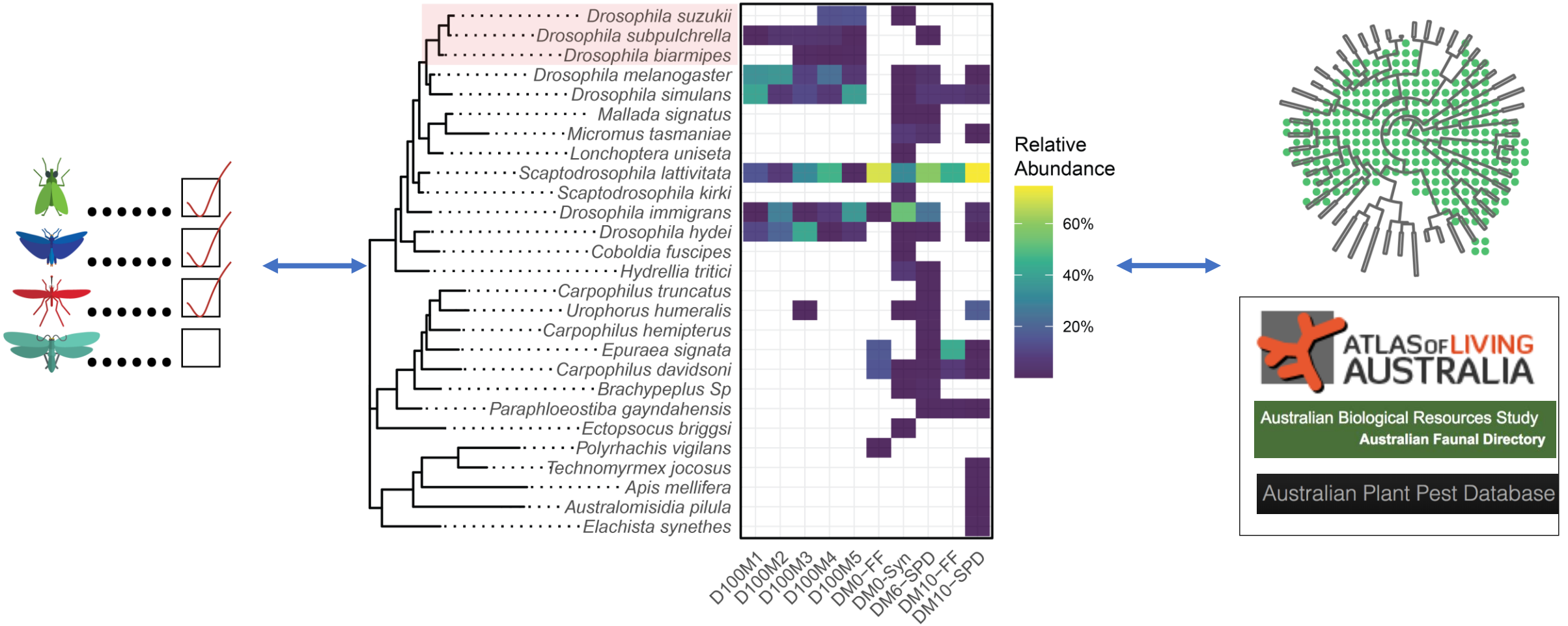


AGRICULTURE VICTORIA

Interpretation

AGRICULTURE VICTORIA

Interpreting species detections

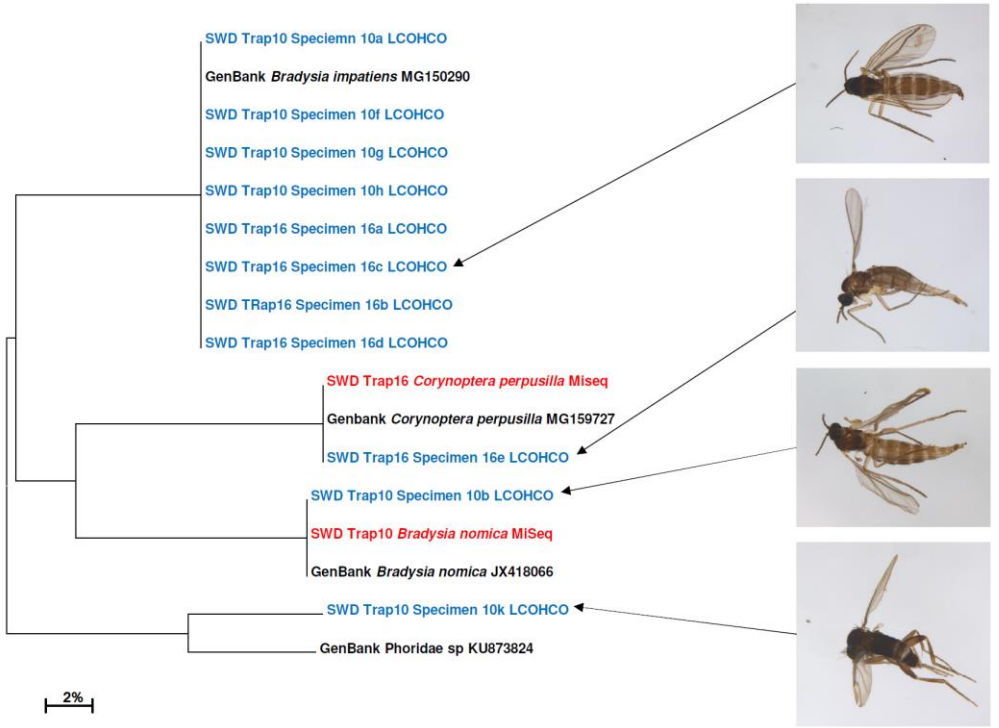


Piper, A. M., Cunningham, J. P., Cogan, N. O., & Blacket, M. J. (2022). **DNA Metabarcoding Enables High-Throughput Detection of Spotted Wing Drosophila (Drosophila suzukii) Within Unsorted Trap Catches**. Frontiers in Ecology and Evolution

Incidental detections

- Metabarcoding is a non-targeted technique using generic primers
- Do not be surprised by first records – lots of Australian biodiversity has not been explored in depth
- Confirmed by revisiting non-destructively extracted specimens
- For destructively extracted samples:
  - Design qPCR primers from the sequences
  - More sampling at the detection site
- Important to have access to group-specific taxonomic expertise for interpretation

Sciaridae flies



Aphid *Pemphigus populivenae*

Exuviae from nymph

AGRICULTURE VICTORIA

Controls

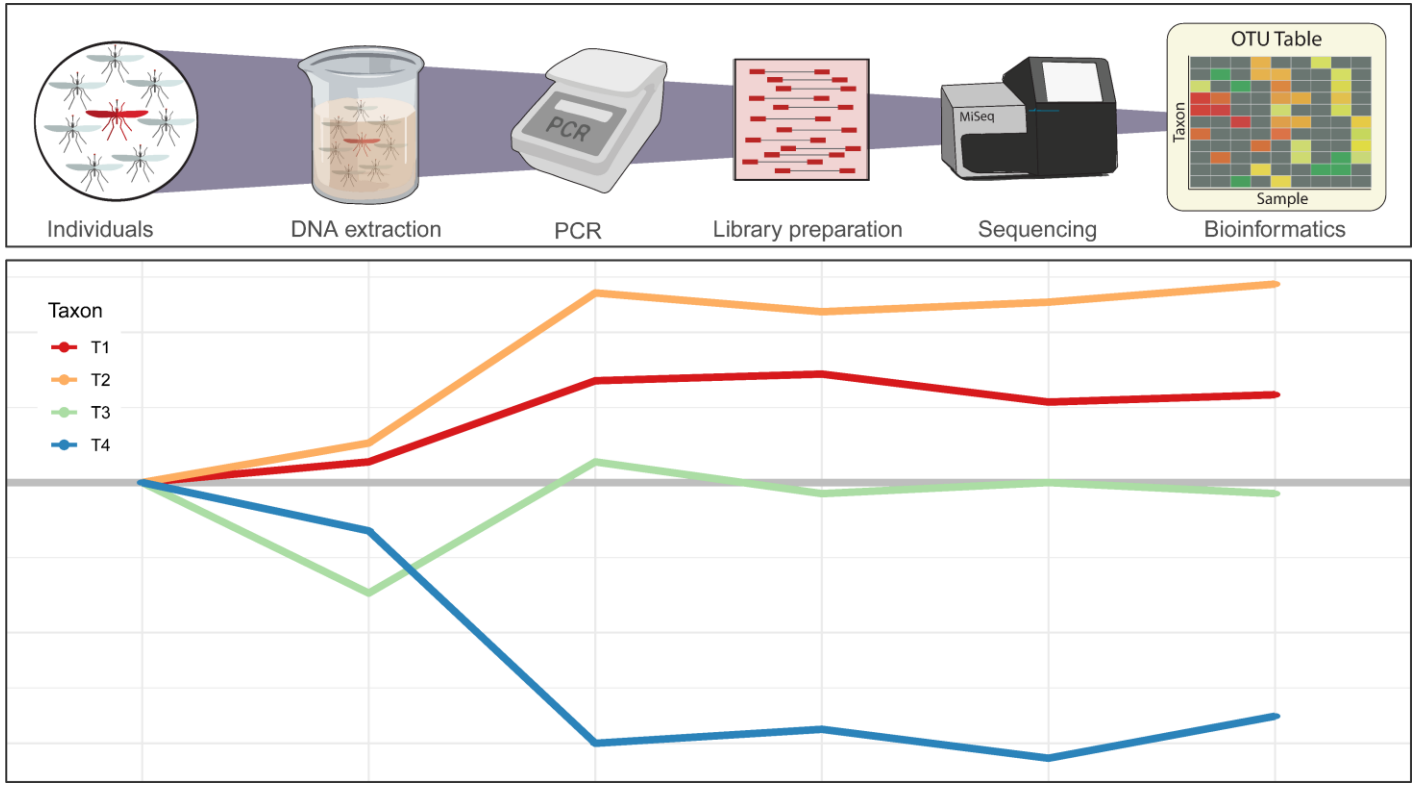
- **Positive controls**
  - A separate sample containing a small but representative fraction of the possible targets the assay could detect
- **Internal positive controls**
  - Spiked synthetic nucleic acids at a low concentration to ensure test has worked correctly
- **Negative controls**
  - Similar to no-template controls in traditional assays
  - Used to monitor for cross contamination and laboratory contamination
- **Alien controls**
  - Contains one or more taxa which belong to the same group as the targets but cannot be present in the samples to be tested (i.e., from different environment)
  - Processed alongside the samples to monitor the detection of targets (role of positive control) and to check for cross contamination between it and the other samples (role of a negative control)

AGRICULTURE VICTORIA

S. Massart et al **Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests**. Peer Community in Infections (In review) https://doi.org/10.5281/zenodo.6637519

Interpreting read counts: Bias

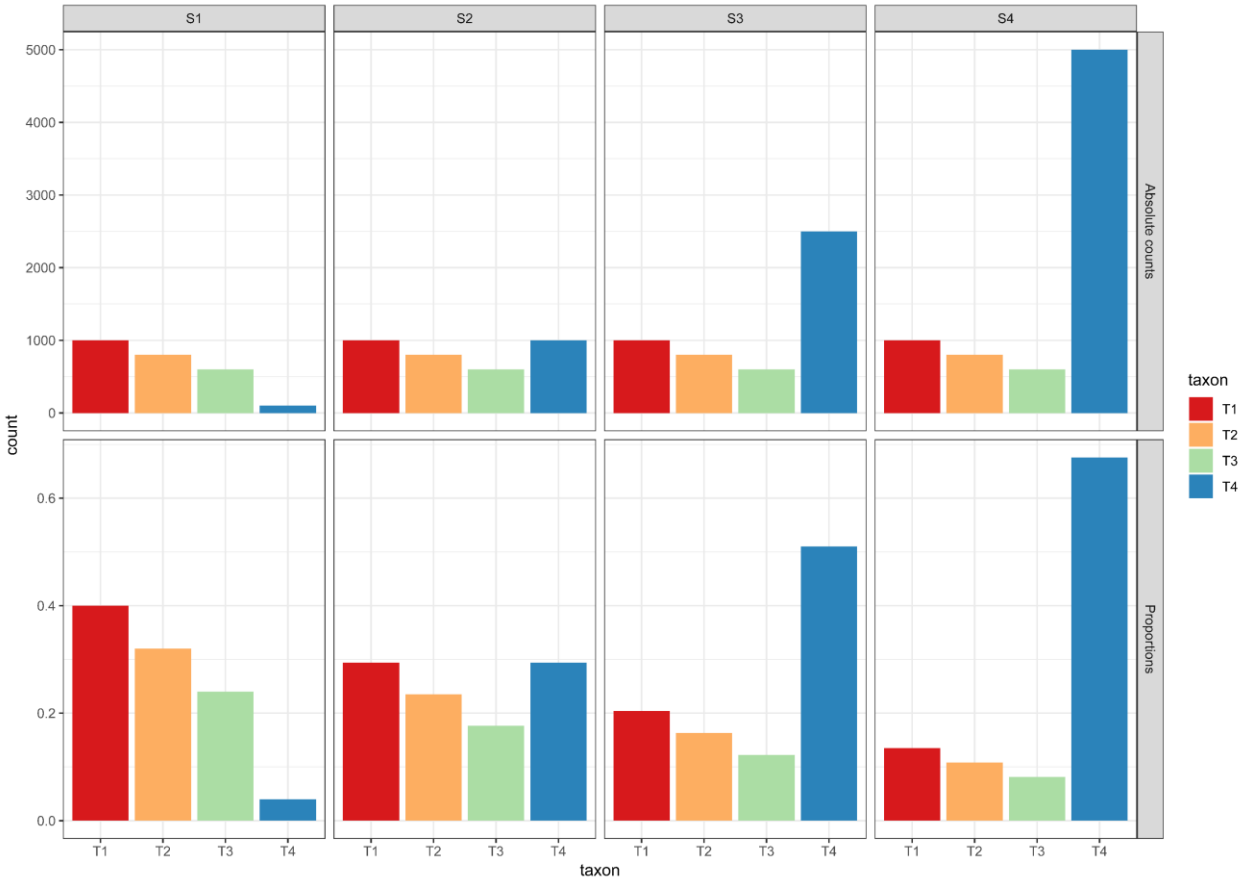
- Metabarcoding provides counts of molecules sequenced
- Read counts only somewhat reflect abundance, due to biases generated through the workflow
- Bias is generally consistent, i.e. Taxon 1 in Sample 1 will have similar bias to Taxon 1 in Sample 2
- Therefore, changes in relative abundances can be compared between treatments or timepoints



AGRICULTURE VICTORIA

Interpreting read counts: Compositionality

- Illumina flow-cells have a fixed output, no matter how much DNA goes in only a fixed sum of reads comes out
- Molecules are competing for limited space on a flow cell
- If more individuals of one species are added to the community, the others will appear to decrease
- If one species is biased upwards, the others will appear to decrease
- Despite looking like ‘counts’ metabarcoding data should be treated as proportions



AGRICULTURE VICTORIA

Summary

- Metabarcoding is high-throughput, sensitive, and broad scope - ideal for a screening / general surveillance tool
- Complex workflow means more opportunities for false positives and negatives to occur
- Appropriate selection of computational tools and parameters is important – critical to have access to bioinformatics expertise
- The ability to detect non-target organisms provides benefits for general surveillance – but must be carefully interpreted
- Abundance information from metabarcoding does not accurately reflect biomass – but is still useful for comparisons

Further reading

- Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance https://doi.org/10.1093/gigascience/giz092
- Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: A step-by-step description. https://doi.org/10.1111/epp.12863
- Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests. https://doi.org/10.5281/zenodo.6637519