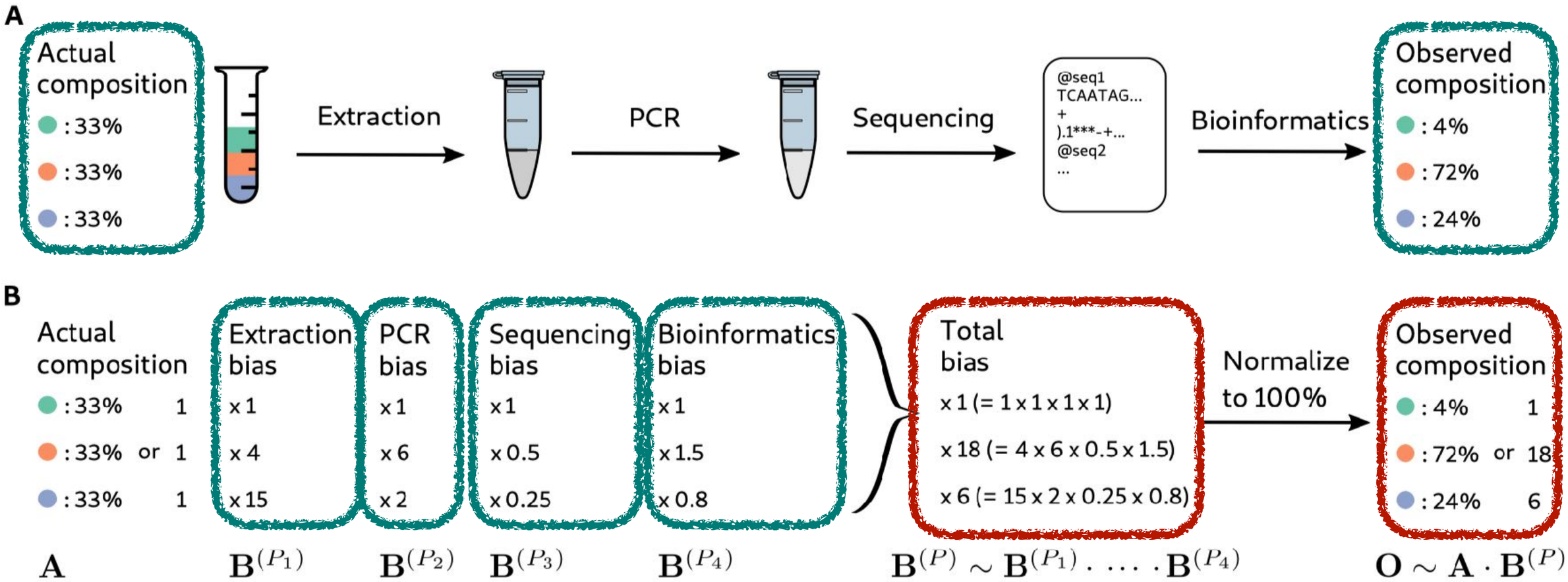
Estimating abundance from DNA-based data

Douglas Yu

Across-species quantification is difficult



eLIFE elifesciences.org

eLife 8:e46923.

Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

Across-species quantification is really difficult!

eLIFE elifesciences.org

Consistent and correctable bias in metagenomic sequencing experiments

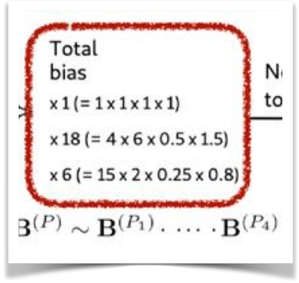
Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

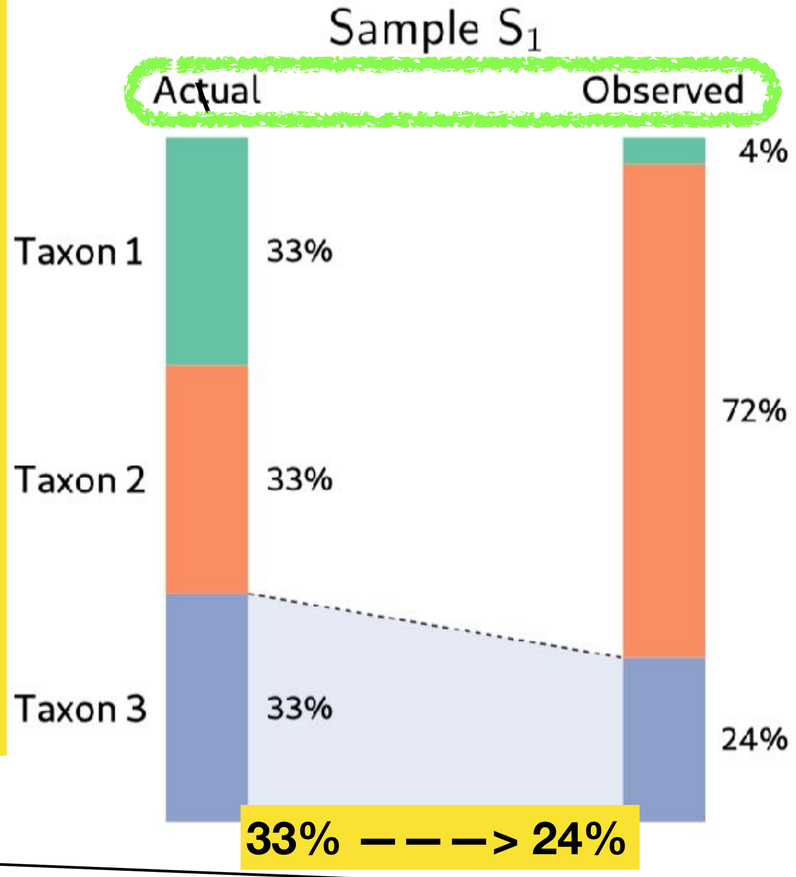
Even with the same underlying biases, it is possible for a given species to become **more** or **less** abundant in a sample after metabarcoding.

Imagine this is a diet study. You only see the Observed proportions. Your conclusions about the relative abundances of prey items would be incorrect.

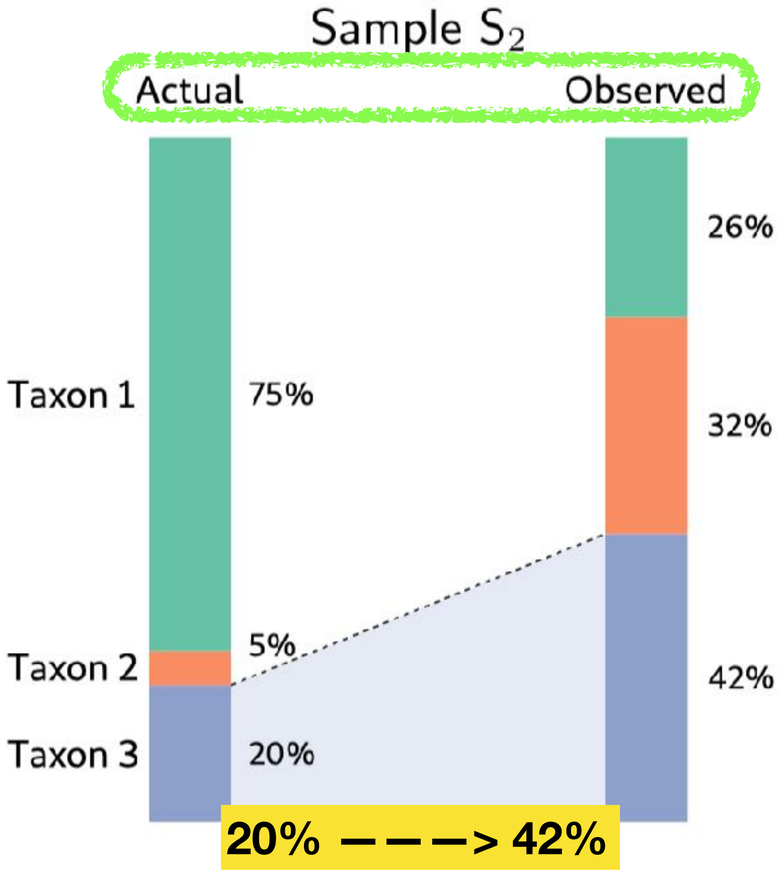
Diet studies try to achieve "across-species" quantification:

"Taxon 3 has higher/lower relative abundance than Taxon 2".





33% ———> 24%



20% ———> 42%

Ratios to Taxon 1

$$\underset{\mathbf{A}(S_1)}{\begin{pmatrix}1\\1\\1\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_1)}{\begin{pmatrix}1\\18\\6\end{pmatrix}} \sim \begin{pmatrix}0.04\\0.72\\0.24\end{pmatrix}$$

$$\underset{\mathbf{A}(S_2)}{\begin{pmatrix}1\\1/15\\4/15\end{pmatrix}} \cdot \underset{\mathbf{B}}{\begin{pmatrix}1\\18\\6\end{pmatrix}} = \underset{\mathbf{O}(S_2)}{\begin{pmatrix}1\\18/15\\24/15\end{pmatrix}} \sim \begin{pmatrix}0.26\\0.32\\0.42\end{pmatrix}$$

Methods to extract abundance information from DNA data

- Single-species quantitative PCR (qPCR)
- Multiplexed specimen barcoding (mBRAVE)
- Mitogenomics and DNA spike-in (SPIKEPIPE)
- Metabarcoding and DNA spike-in (qSeq)
- Reverse metagenomics (RevMet)

Methods to extract abundance information from DNA data

- **Single-species quantitative PCR (qPCR)**
- Multiplexed specimen barcoding (mBRAVE)
- Mitogenomics and DNA spike-in (SPIKEPIPE)
- Metabarcoding and DNA spike-in (qSeq)
- Reverse metagenomics (RevMet)

DOI: 10.1111/1755-0998.12987

RESOURCE ARTICLE

2019

WILEY MOLECULAR ECOLOGY RESOURCES

Environmental DNA for the enumeration and management of Pacific salmon

Taal Levi[1] | Jennifer M. Allen[1] | Donovan Bell[2] | John Joyce[2] | Joshua R. Russell[2] | David A. Tallmon[3] | Scott C. Vulstek[2] | Chunyan Yang[4] | Douglas W. Yu[4,5,6]

>>$500 million / year industry

6000 streams used by the 5 salmon species

1000 streams used for spawning

4 - 9 streams used to enumerate Coho salmon

~14 streams for Sockeye salmon

0 for Pink and Chum salmon (main export species)

0 125 250 500 Kilometers

Alaskan salmon fishery

The ‘escapement’ is the breeding population

Is the fishery allowing enough salmon to escape?

Commercial Fishing

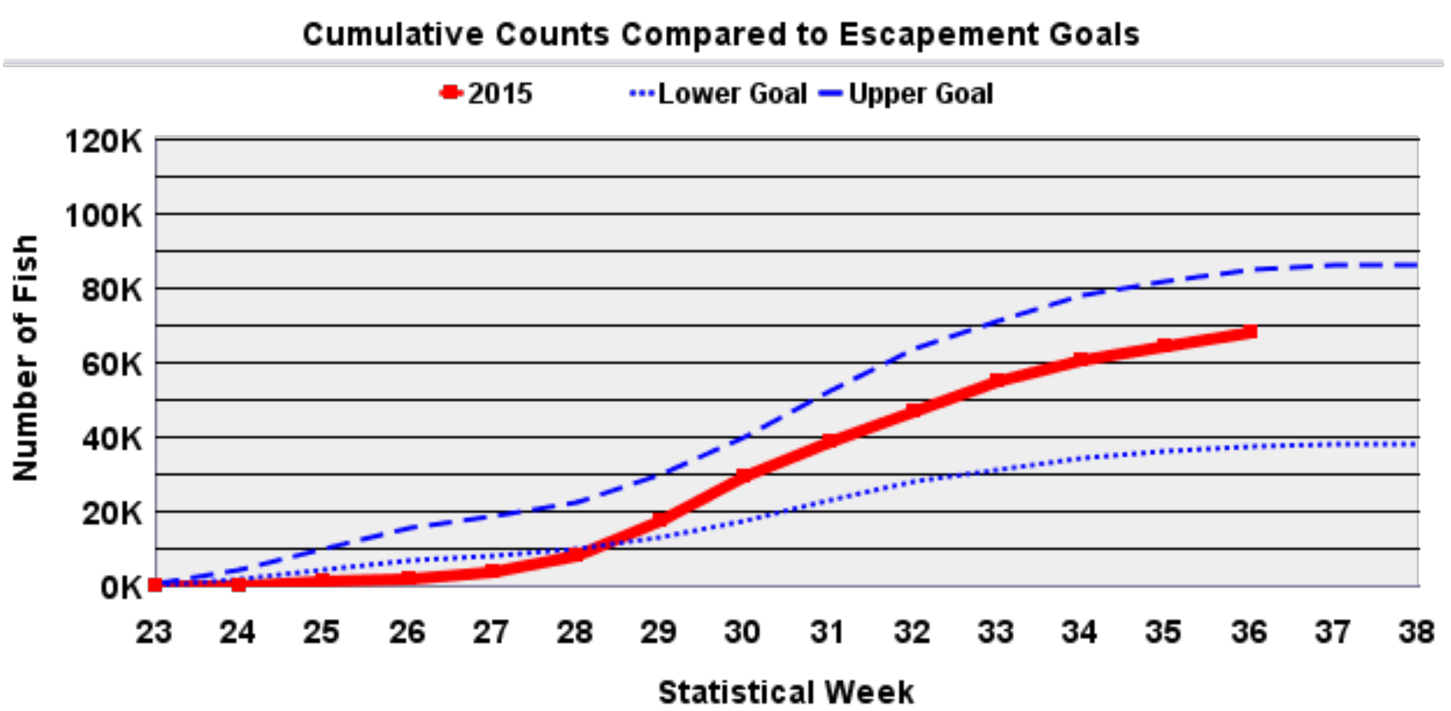
Commercial Fishing Home
News Releases & Announcements
Information By Area
- Arctic-Yukon-Kuskokwim
  - Kuskokwim
  - Arctic Area
  - Norton Sound & Kotzebue
  - Yukon
- Central Region
  - Bristol Bay
  - Copper River
  - Cook Inlet
    - Lower Cook Inlet
    - Upper Cook Inlet
  - Prince William Sound
- Southeast Region
- Westward Region
  - Alaska Peninsula
  - Bering Sea & Aleutian Islands
  - Chignik
  - Kodiak Island

ADF&G Home » Fishing » Commercial » Information By Area » Southeast

Like ? Share

Commercial Salmon Fisheries

Chilkoot Lake Weir - Sockeye Counts



ALASKA DEPARTMENT OF FISH AND GAME
DIVISION OF COMMERCIAL FISHERIES
NEWS RELEASE

Sam Cotten, Commissioner
Jeff Regnart, Director

Contact:
Gordie Woods

Phone: (907) 784-3255
Fax: (907) 784-3254

Yakutat Area Office
P.O. Box 49
Yakutat, AK, 99689

Date: September 4, 2015
Time: 8:30 a.m.

YAKUTAT COMMERCIAL SET GILLNET OPENING ANNOUNCEMENT

Dangerous River: will be open from 12:01 p.m., Sunday, September 6 through 12:00 noon, Wednesday, September 9.

Akwe River: will be open from 12:01 p.m., Sunday, September 6 through 12:00 noon, Wednesday, September 9.

Lost River: will remain closed until further notice.

Situk-Ahrnklin Inlet: will be open from 12:01 p.m., Sunday, September 6 through 12:00 noon, Wednesday, September 9 with the following restrictions:

North Bank of Situk-Ahrnklin Inlet: commercial set gillnet fishing will be prohibited along the north bank of the Situk-Ahrnklin Inlet between two ADF&G regulation markers located 500 yards above and 500 yards below the confluence of the Situk-Ahrnklin Inlet and the Lost River;

Validation data from Auke Creek Station in Juneau, Alaska

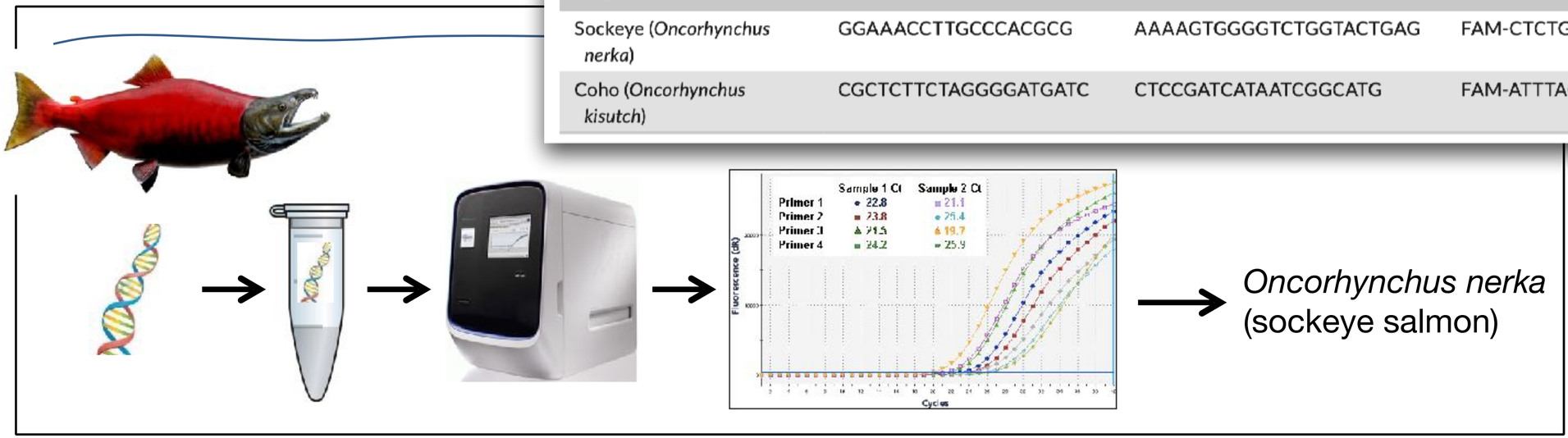
Daily counts of

- in-migrating Coho and Sockeye **adults**
- out-migrating Sockeye **smolts (juveniles)**

Does environmental DNA (eDNA) contain enough information to estimate salmon escapement sizes?

TABLE 1 Species-specific primers and probes used in this study (Rasmussen Hellberg et al., 2010)

Target species	Forward Primer (5'–3')	Reverse Primer (5'–3')	Probe (5'–3')
Sockeye (*Oncorhynchus nerka*)	GGAAACCTTGCCCACGCG	AAAAGTGGGGTCTGGTACTGAG	FAM-CTCTGTTGACTTAACCATC-MGB
Coho (*Oncorhynchus kisutch*)	CGCTCTTCTAGGGGATGATC	CTCCGATCATAATCGGCATG	FAM-ATTTACAACGTAATCGTC-MGB



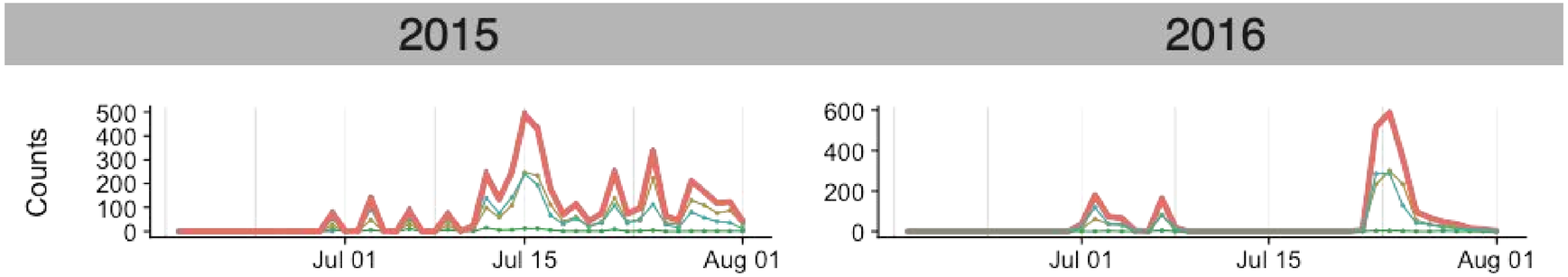
Oncorhynchus nerka (sockeye salmon)

One person, one sample /day

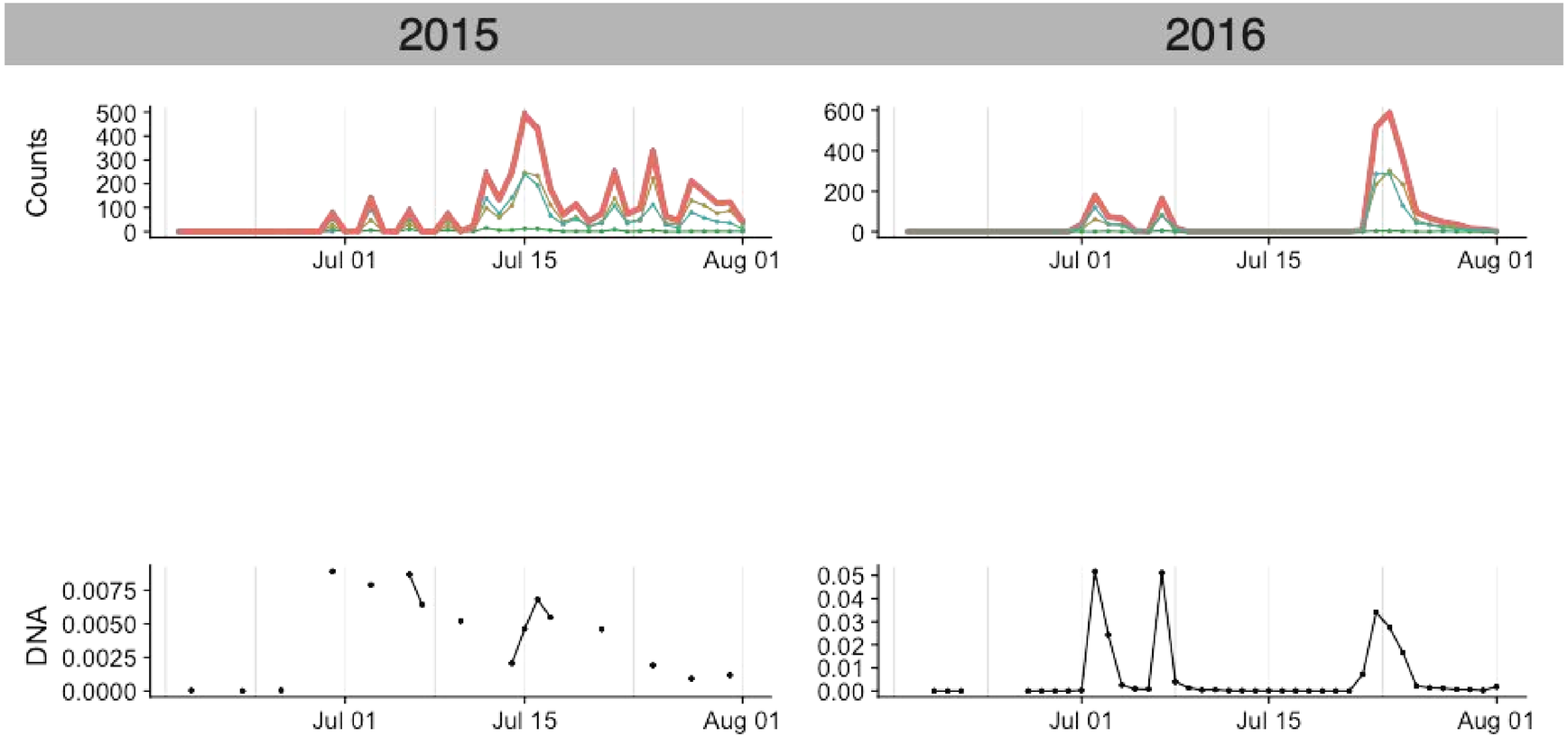
=

Many people, all day

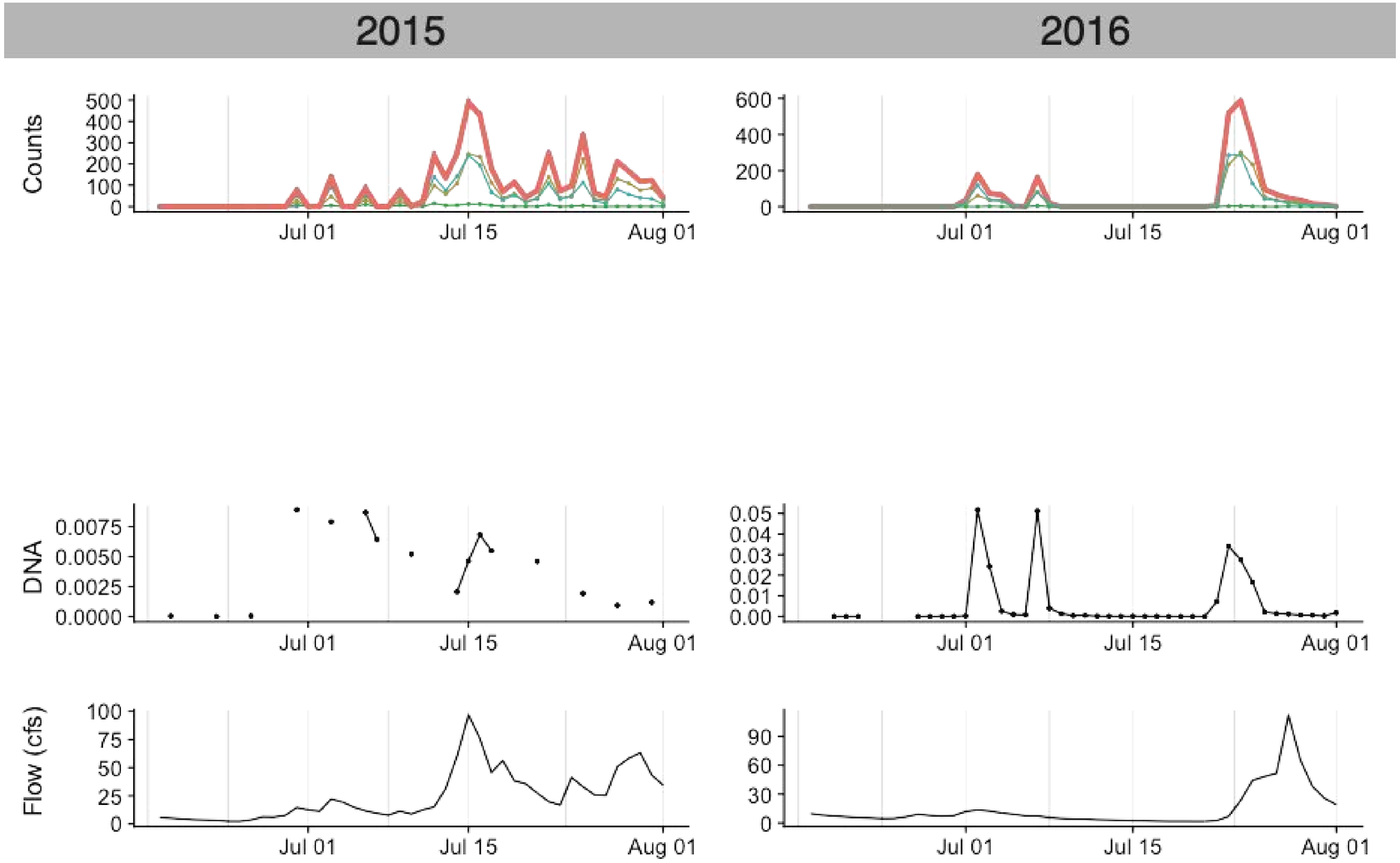
Sockeye in-migrating adults



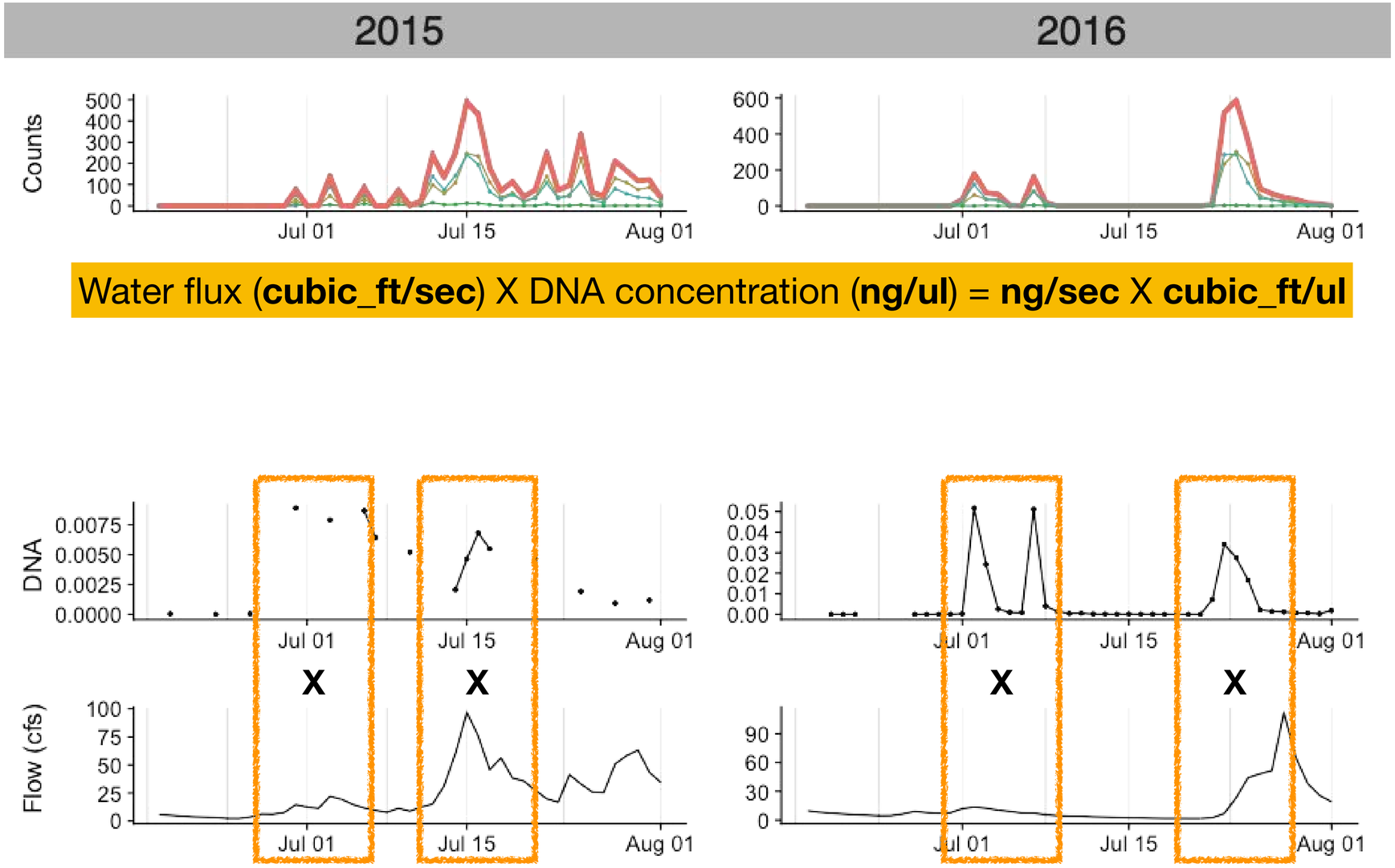
Sockeye in-migrating adults



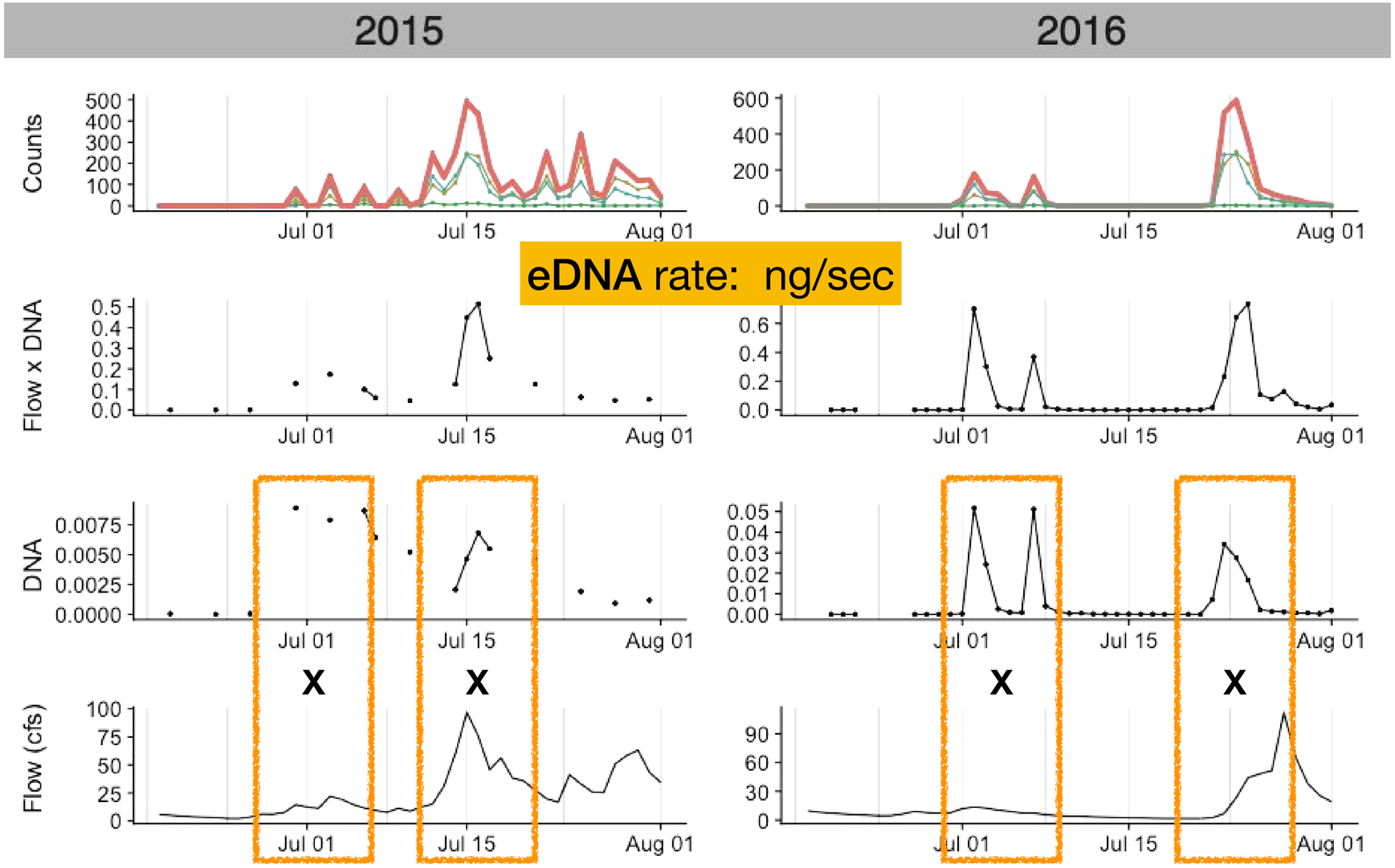
Sockeye in-migrating adults



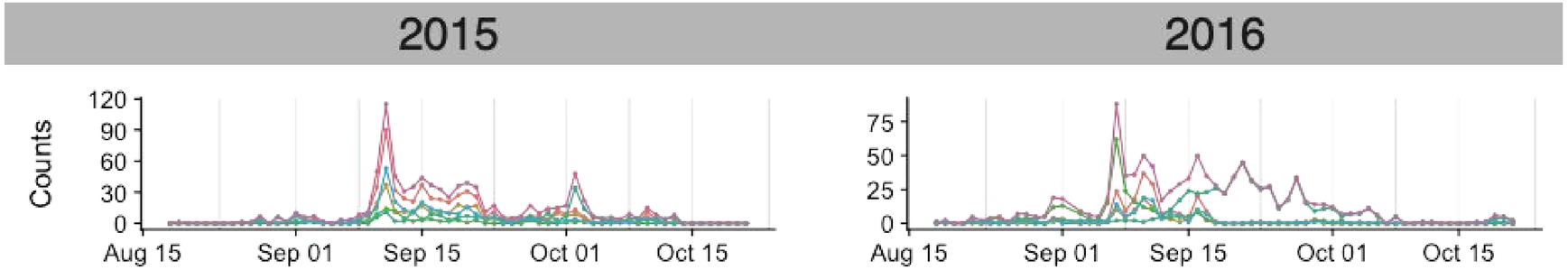
Sockeye in-migrating adults



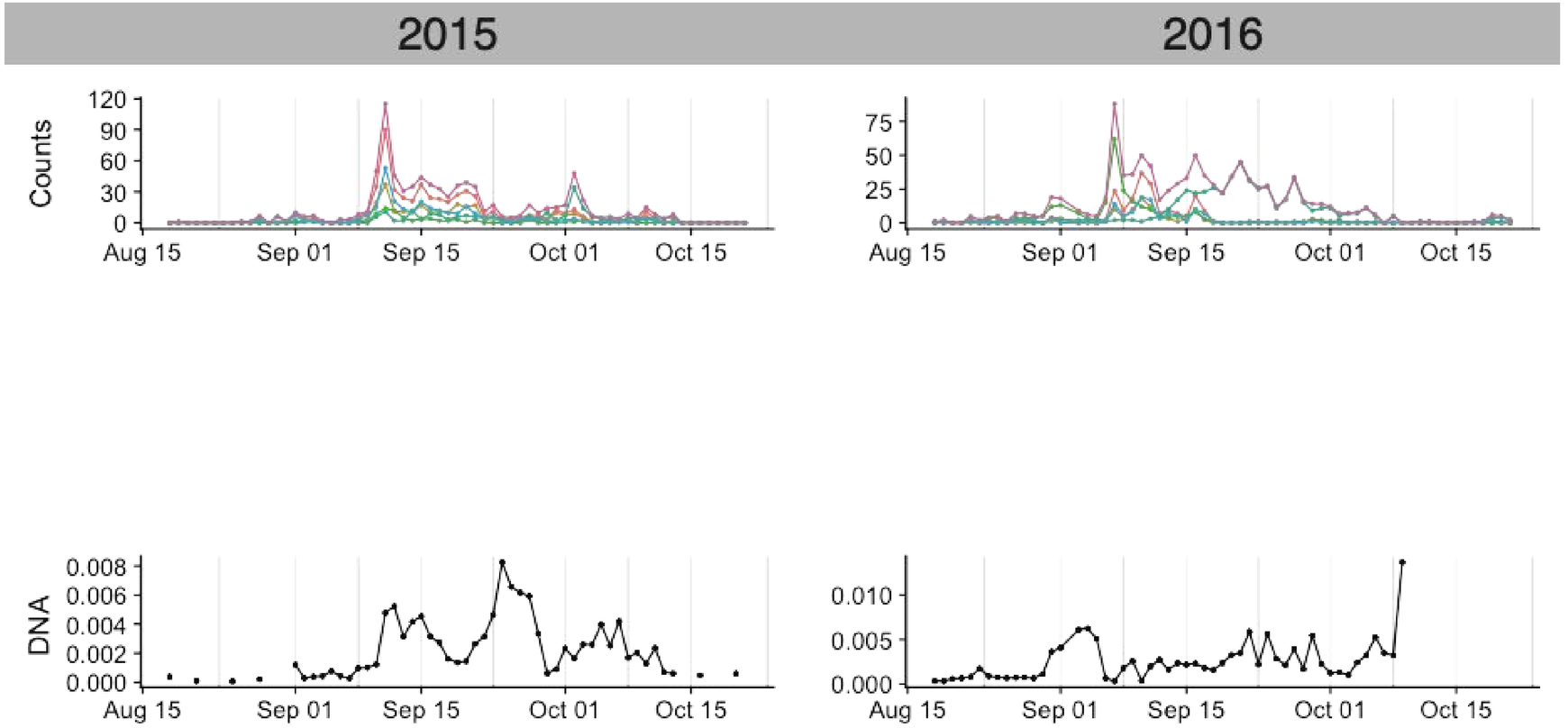
Sockeye in-migrating adults



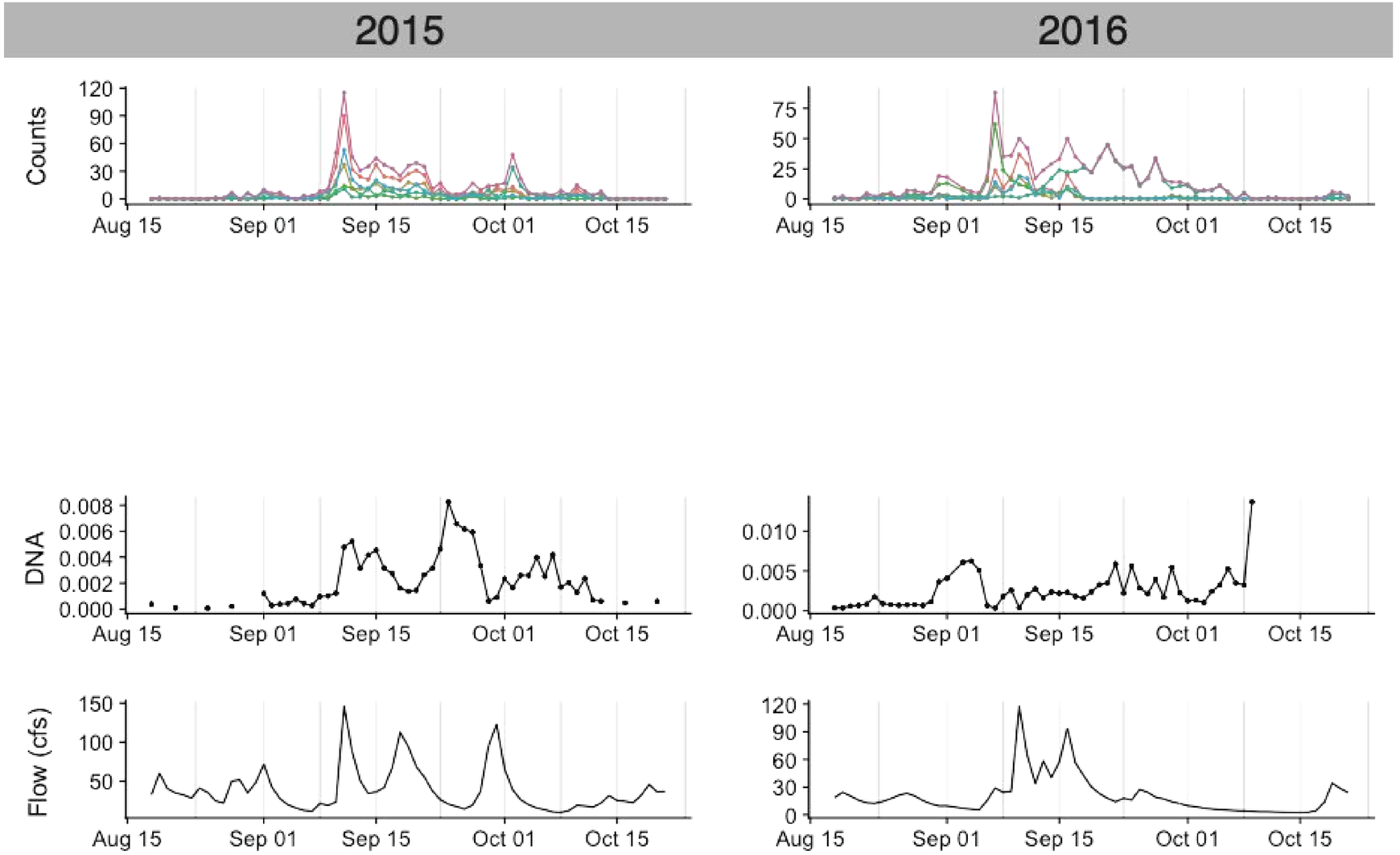
Coho in-migrating adults



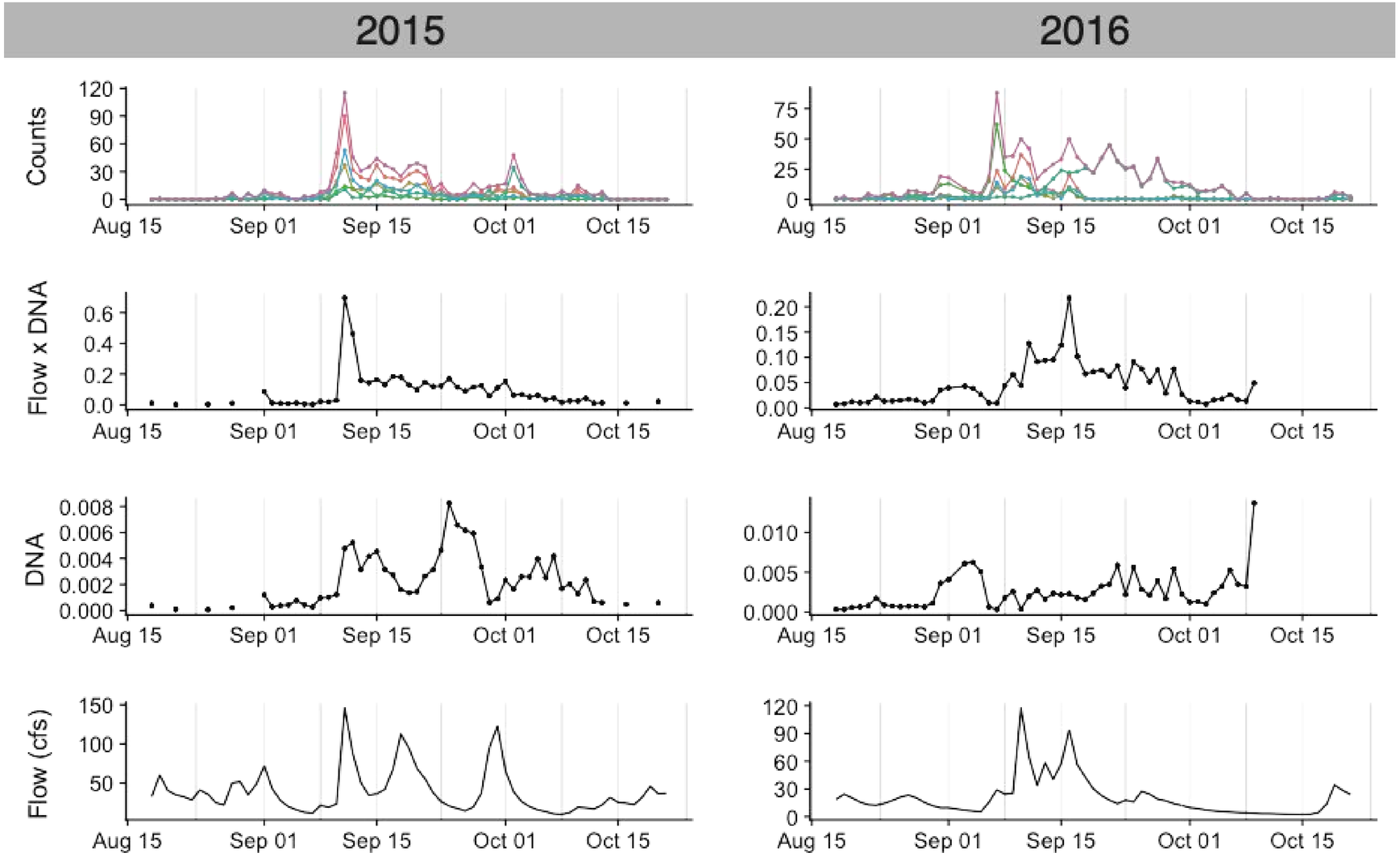
Coho in-migrating adults



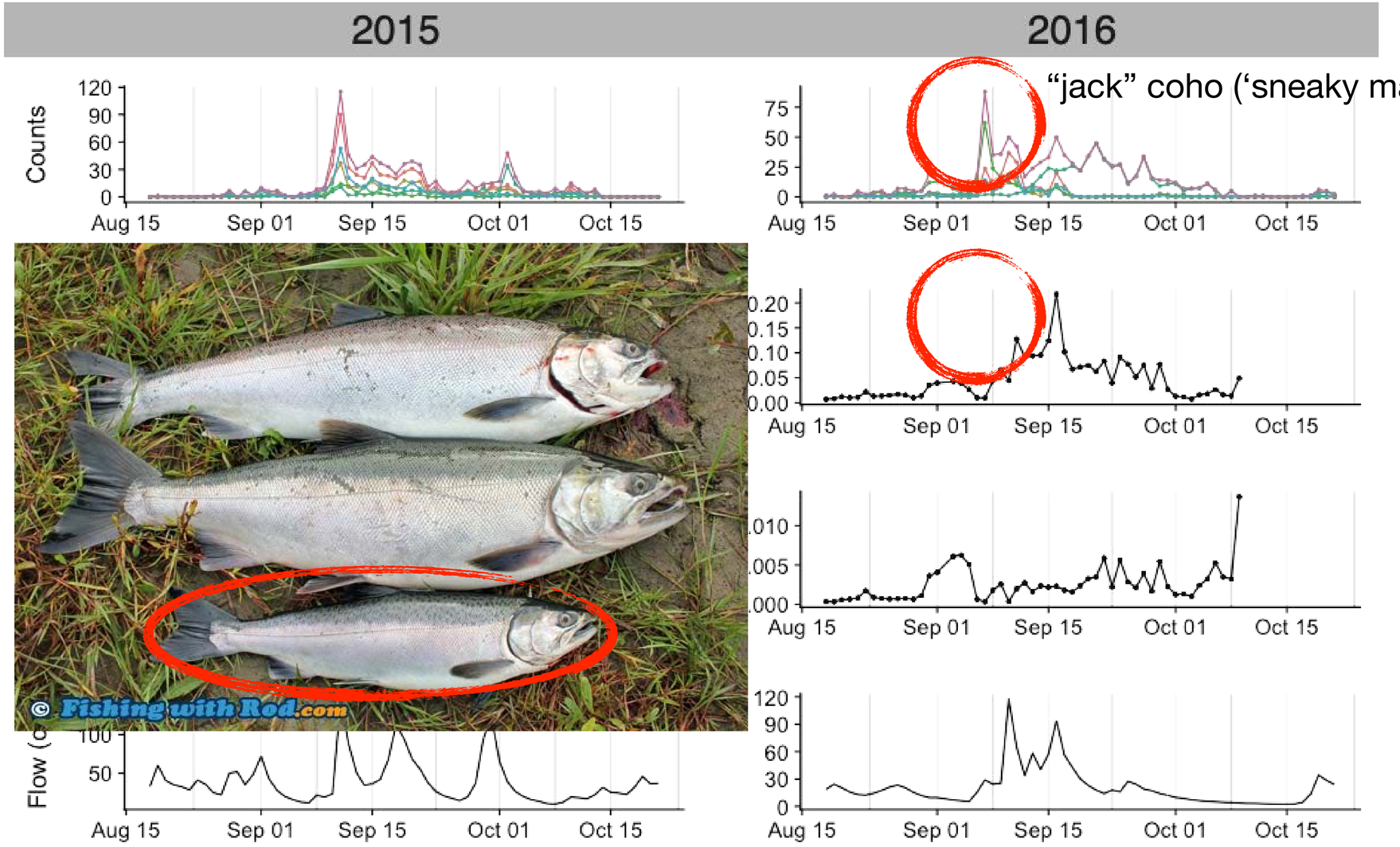
Coho in-migrating adults



Coho in-migrating adults

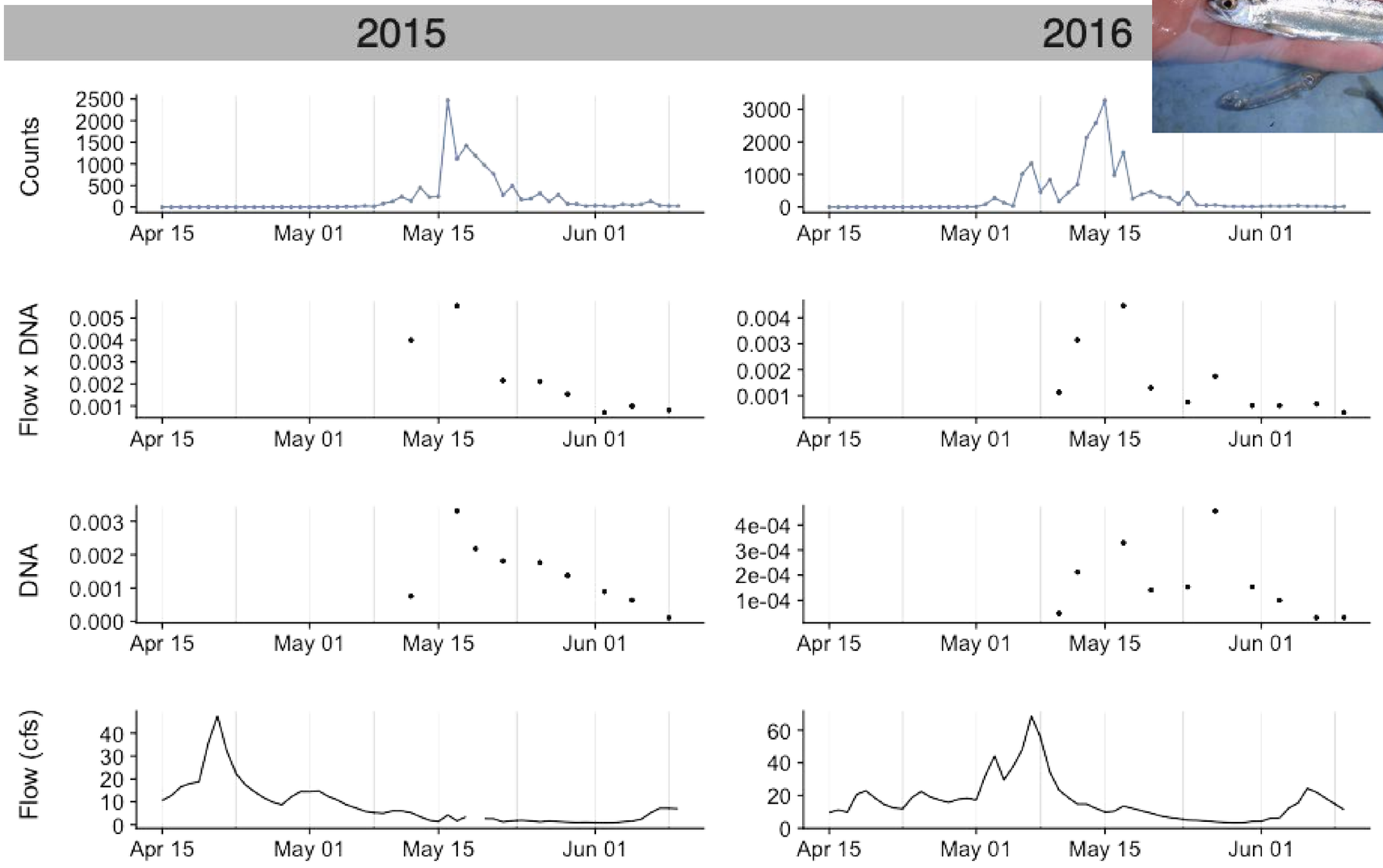


Coho in-migrating adults



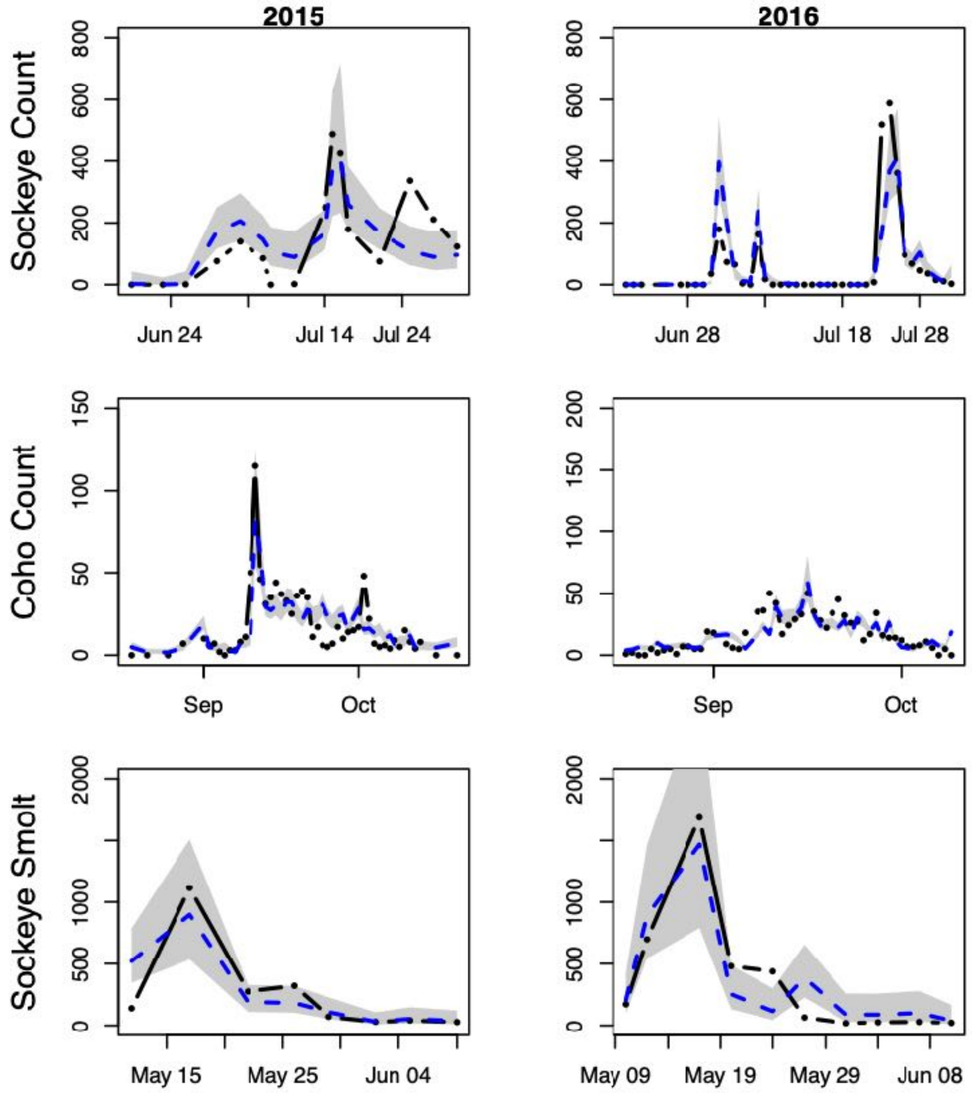
“jack” coho (‘sneaky males’)

Sockeye out-migrating smolts (juveniles)



Fitted models (blue) vs. Data (black)

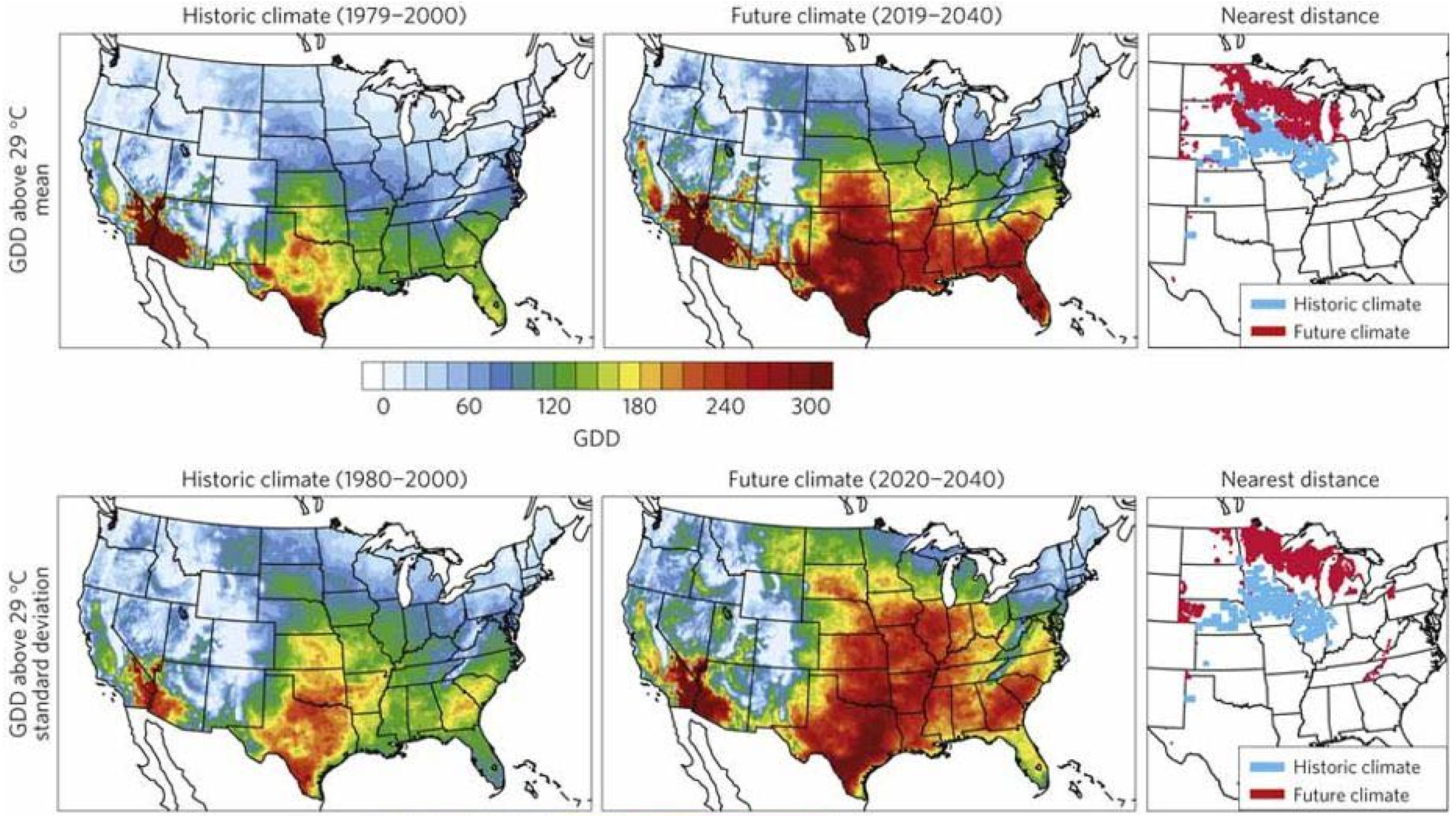
- `glm(log(count) ~ log(eDNA_rate), quasipoisson)`
- All highly significant
- Models not significantly different **between years**
- Small signal of salmon one day ago,
  No effect of salmon two days ago



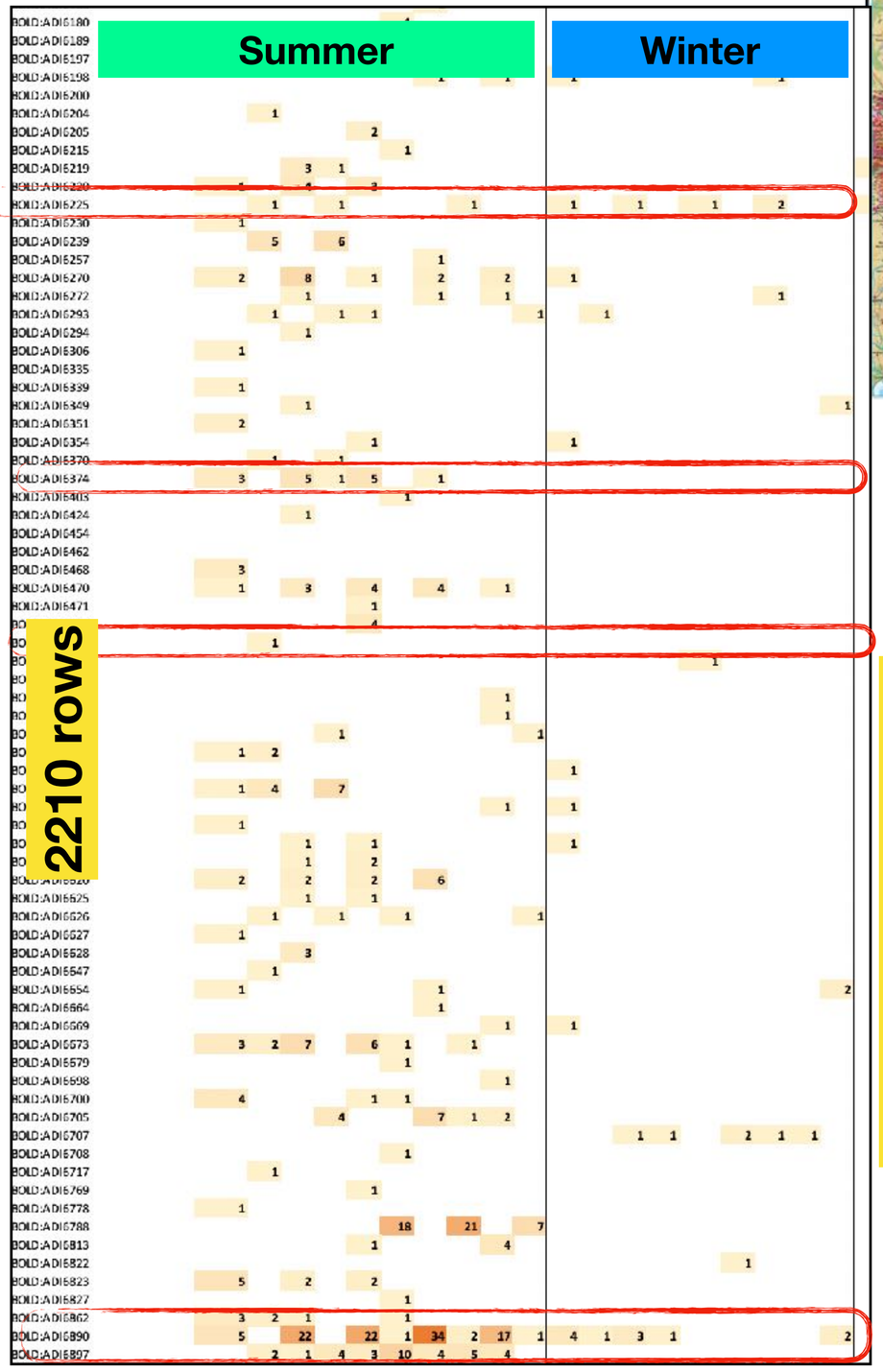
Methods to extract abundance information from DNA data

- Single-species quantitative PCR (qPCR): *control for eDNA transport dynamics*
- **Multiplexed *individual* barcoding (mBRAVE)**
- Mitogenomics and DNA spike-in (SPIKEPIPE)
- Metabarcoding and DNA spike-in (qSeq)
- Reverse metagenomics (RevMet)

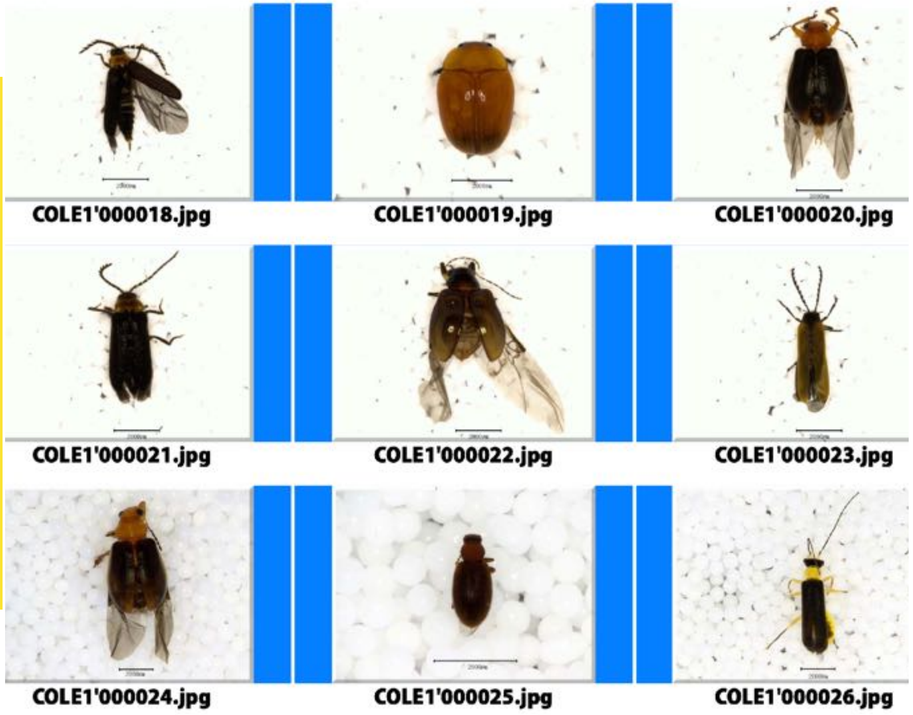
How much biodiversity will we lose because of climate change?

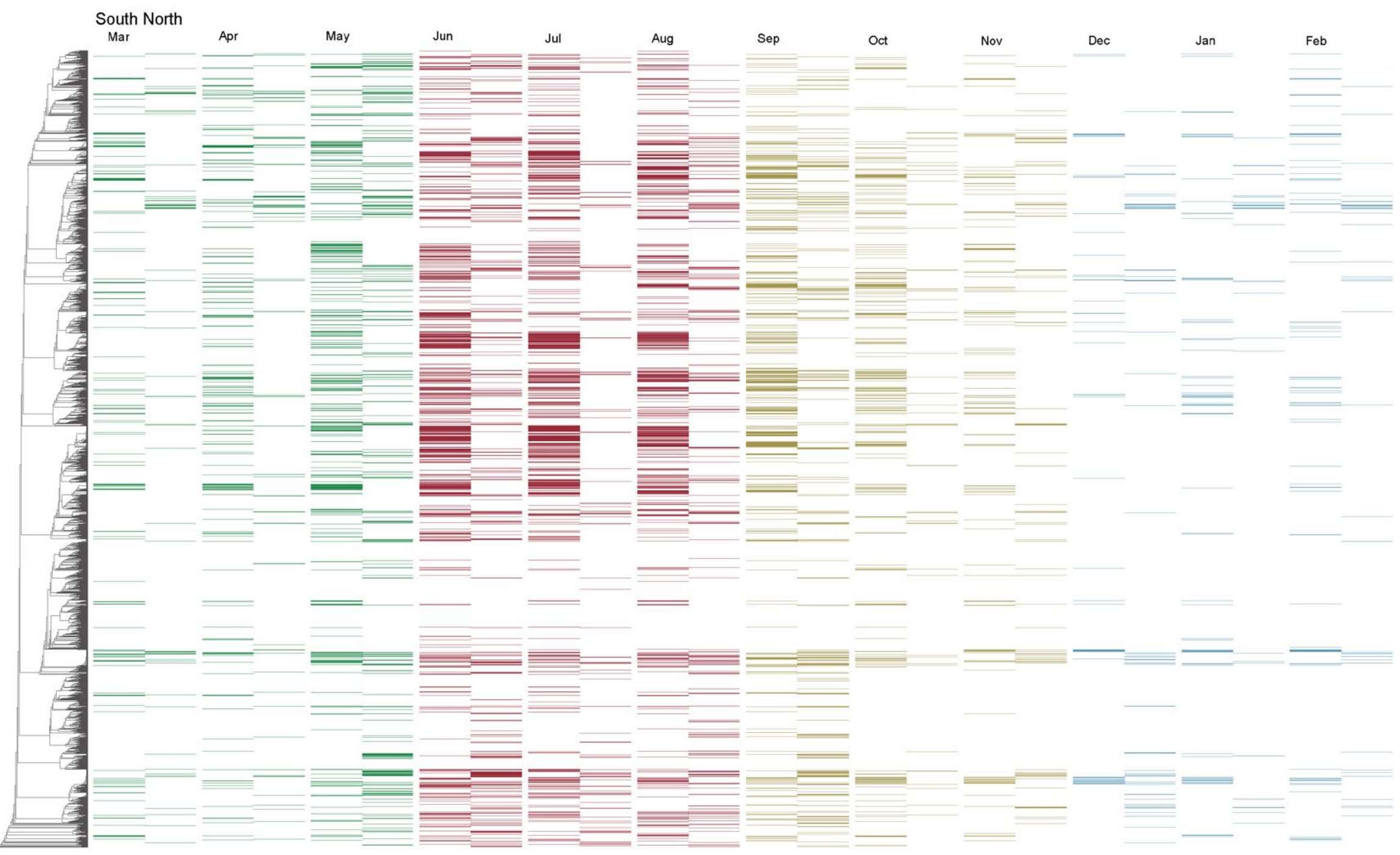


Changes in distribution due to changes in climate envelopes

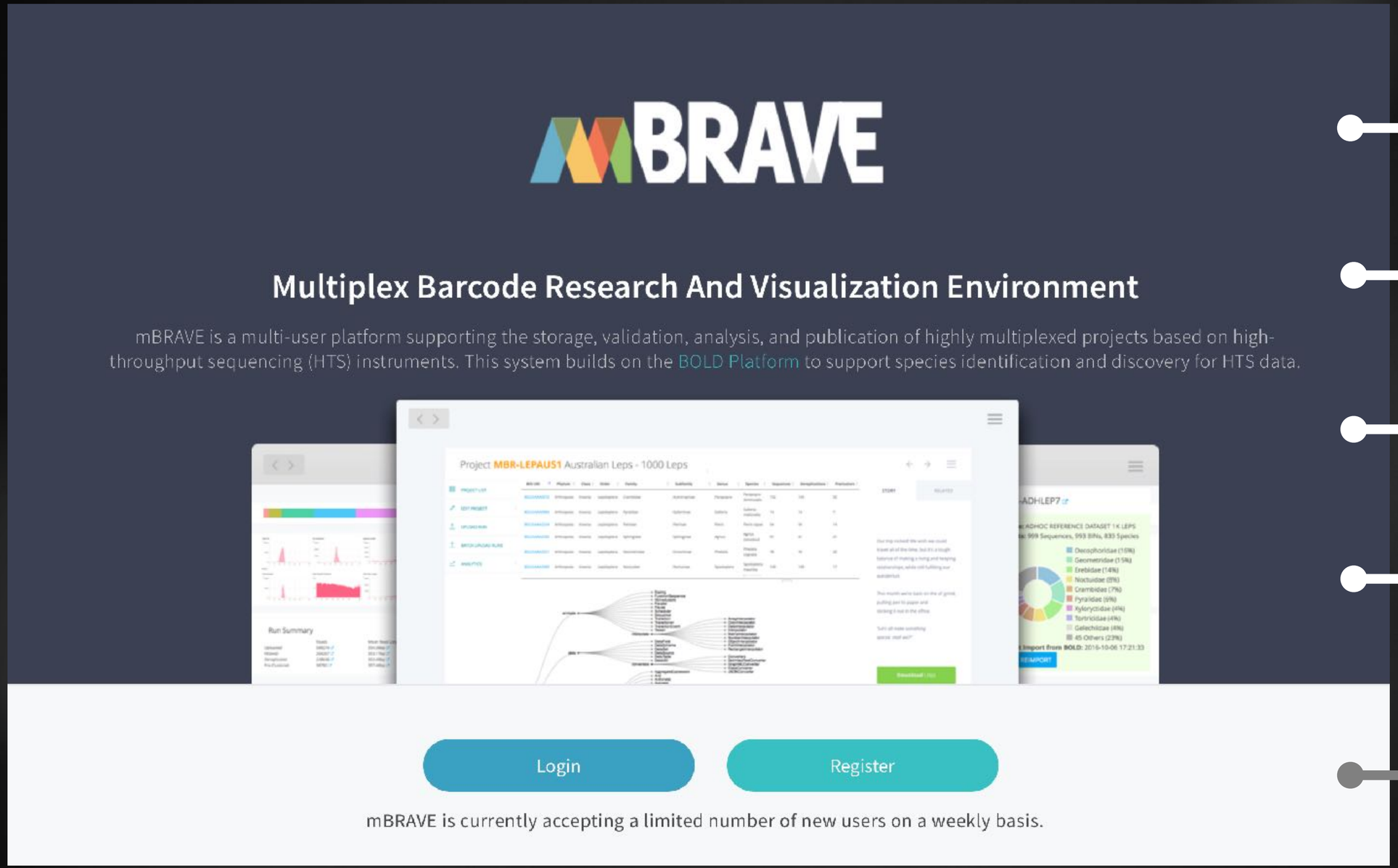


- **10 462 beetles**
- collected biweekly over winter and summer
- alpine and subtropical sites
- **individually DNA barcoded**
- 2210 beetle *BINs (“species”)*
  - Barcode Identification Numbers





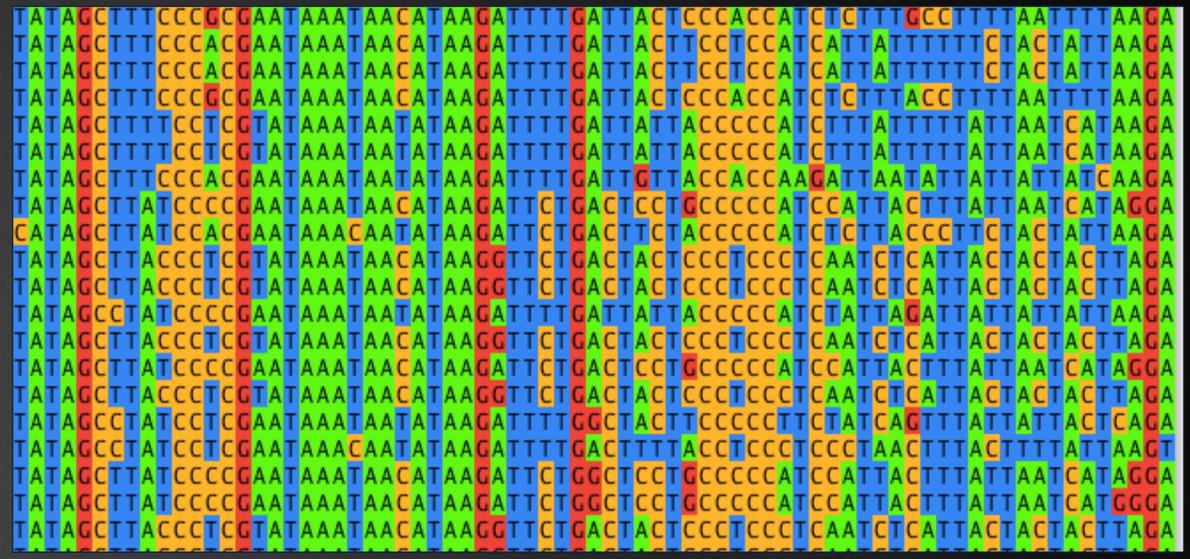
- **10 462 beetles**
- collected biweekly over winter and summer
- alpine and subtropical sites
- **individually DNA barcoded**
- 2210 beetle *BINs ("species")*
  - Barcode Identification Numbers



- Storage
- Analysis
- Indexing
- Discovery
- Standards

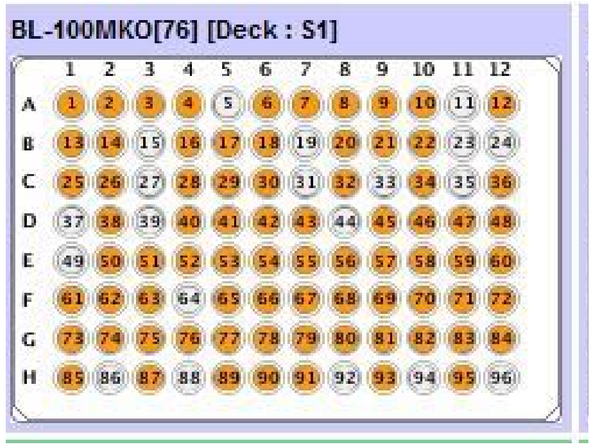
mbrave.net

Massively parallel individual barcoding
Multiplexed individual barcoding

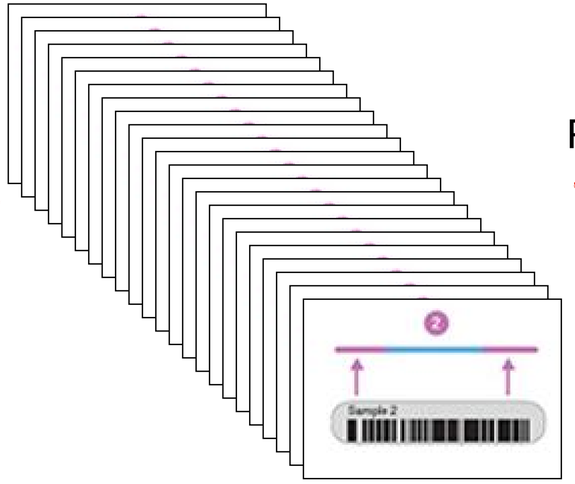


TATACTTATCCCCGAATAAATAACATAAGATTTGACTCCCCCATCATCAAATTATAAT

DNA from 96 beetles in 96 wells



$$** Add 96 COI primer pairs with 96 different twin tags **$$



Tag	F Primer	barcode sequence	R Primer	Tag

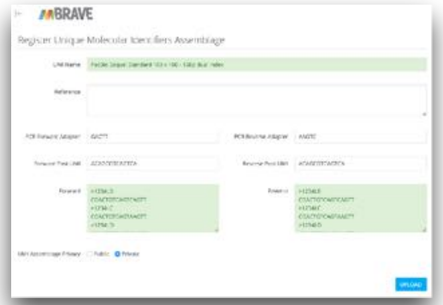
Pool amplicons

Illumina library prep

Illumina sequence

Upload fastq files to mBRAVE

Upload tag information to mBRAVE

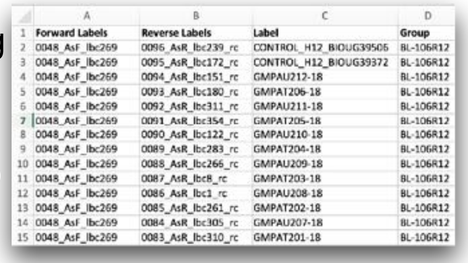


Demultiplex by tag

Keep the most abundant read per tag (= well = beetle), filter out the rest, and Assign a taxonomy to using BOLD

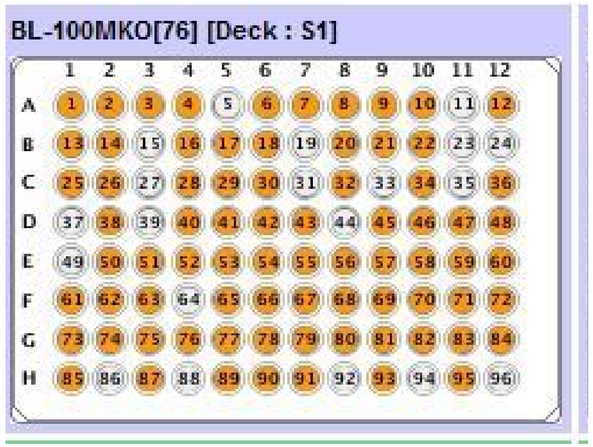
A consensus COI sequence for each beetle, plus metadata, in fasta file: **count data!!**

```
>ANML_DLJ06_E12|run:ANML_DLJ06_dtd_0|contig-id:1|rep_count:1169|id_similarity:98.
8950276243|c_count:1|cmxd:1.1|cmnd:0.1116338751|cnnd:None|p:None|c:None|o:None|f:None|g:
None|s:None|otu:BOLD:AAB7930|date:2019-08-13
AACTCTTTATTTTATTTTTGGTGCTTGGGCTGGAATAGTTGGTATATCCCTTAGTATACTGATTCGAGTTGAATTAAGAACACCTGGTA
CTCTACTAGGTGACGACCAAATATATAATGTAATCGTTACAGCTCATGCATTTATCATAATTTTCTTTATAGTTATACCTATTGTAATT
GGT
>ANML_DLJ04_C04|run:ANML_DLJ04_dtd_0|contig-id:1|rep_count:419|id_similarity:91.
1602209945|c_count:3|cmxd:1.7|cmnd:0.2888305489|cnnd:3.3|p:Arthropoda|c:Insecta|o:None|f:
None|g:None|s:None|otu:OTU14|date:2019-08-13
AACACTATATTTCATTTTTGGAGCTTGAGCAGGTATAGTTGGGACATCACTGAGATTACTAATTCGAGCAGAATTAGGAAACCCCGGAT
CTCTAATTGGTGATGACCAAATCTATAACGTAATTGTAACAGCCCATGCATTTGTAATAATTTTTTTTATAGTTATACCTATTGTAATT
GGT
>ANML_2909_D06|run:ANML_2909_dtd_0|contig-id:1|rep_count:20|id_similarity:nan|c_count:2|
cmxd:1.1|cmnd:0.255|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU37|date:
2019-08-01
TACTTTGTATTTCATTTTTGGTGCTTGAGCTGGAATAGTAGGAACCTCTTTAAGCCTTCTAATTCGAGCAGAATTAGGGAATCCTGGTT
CCCTAATTGGTGATGATCAGATTTATAATGTAATTGTAACTGCACATGCTTTCATTATAATTTTTTTTATAGTTATACCAATTGTTATT
GGA
>ANML_DLJ05_H08|run:ANML_DLJ05_dtd_0|contig-id:1|rep_count:6196|id_similarity:nan|c_count
:6|cmxd:1.7|cmnd:0.103518399|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU2|
date:2019-08-14
CACTCTTTATTTTATCTTTGGAGCATGGGCTGGAATACTTGGAACATCCTTAAGACTCCTCATCCGAGCAGAACTTGGAAACCCGGGTT
CATTAATCGGAAACGATCAAATTTACAACGTAATTGTTACTGCCCACGCTTTCATTATAATTTTCTTCATAGTAATACCAATCATAATT
GGT
>ANML_2922_F09|run:ANML_2922_dtd_0|contig-id:1|rep_count:1342|id_similarity:nan|c_count:2
|cmxd:1.7|cmnd:0.1236214605|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU7|
date:2019-08-02
ATCCTTATACTTTCTTTTTGGTATATGATCAGGAATAGTTGGAACCTCATTAAGATTACTAATTCGCCTAGAATTAGGTAATCCTGGGT
CTTTAATTGGAGACGACCAAATTTATAATGTAATTGTAACTGCCCATGCATTTATTATAATTTTCTTTATAGTTATACCTGTAATAATA
GGA
>ANML_DLJ01_F05|run:ANML_DLJ01_dtd_0|contig-id:3|rep_count:10|id_similarity:nan|c_count:3
|cmxd:0.6|cmnd:0.24|cnnd:26.1|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU115|date:
2019-08-13
AACCCTTTACTTCATTTTCGGAGCATGAAGAGGTATAATCGGAACCTCCATAAGATTGATGATTCGAACAGAATTAGGAACAGCTGGCT
CTCTAATTGGAGATGATCAAATTTATAATGTAATTGTAACAGCACATGCTTTTATCATAATTTTTTTCATAGTAATACCCATCATGATA
GGA
```

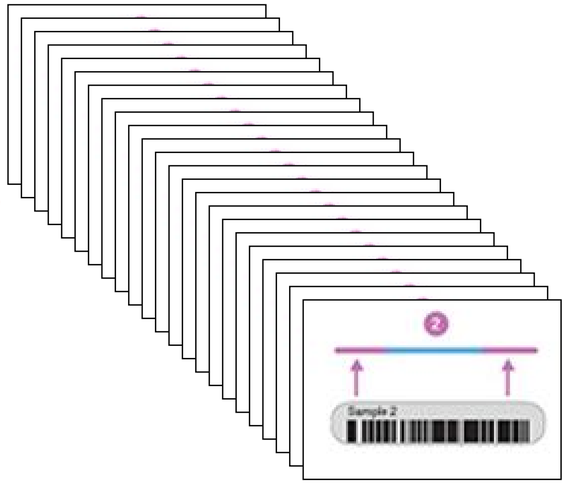


	A	B	C	D
1	Forward Labels	Reverse Labels	Label	Group
2	0048_AsF_lbc269	0096_AsR_lbc239_rc	CONTROL_H12_BIOUG39505	BL-106R12
3	0048_AsF_lbc269	0095_AsR_lbc172_rc	CONTROL_H12_BIOUG39372	BL-106R12
4	0048_AsF_lbc269	0094_AsR_lbc151_rc	GMPAU212-18	BL-106R12
5	0048_AsF_lbc269	0093_AsR_lbc180_rc	GMPAT206-18	BL-106R12
6	0048_AsF_lbc269	0092_AsR_lbc311_rc	GMPAU211-18	BL-106R12
7	0048_AsF_lbc269	0091_AsR_lbc354_rc	GMPAT205-18	BL-106R12
8	0048_AsF_lbc269	0090_AsR_lbc122_rc	GMPAU210-18	BL-106R12
9	0048_AsF_lbc269	0089_AsR_lbc283_rc	GMPAT204-18	BL-106R12
10	0048_AsF_lbc269	0088_AsR_lbc266_rc	GMPAU209-18	BL-106R12
11	0048_AsF_lbc269	0087_AsR_lbc8_rc	GMPAT203-18	BL-106R12
12	0048_AsF_lbc269	0086_AsR_lbc1_rc	GMPAU208-18	BL-106R12
13	0048_AsF_lbc269	0085_AsR_lbc261_rc	GMPAT202-18	BL-106R12
14	0048_AsF_lbc269	0084_AsR_lbc305_rc	GMPAU207-18	BL-106R12
15	0048_AsF_lbc269	0083_AsR_lbc310_rc	GMPAT201-18	BL-106R12

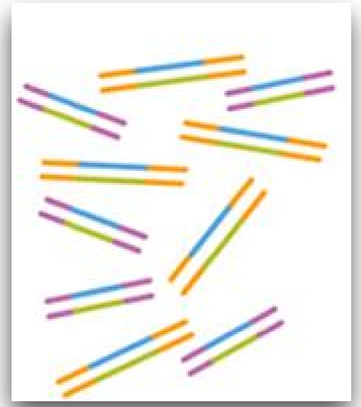
DNA from 96 beetles in 96 wells



Add 96 COI primer pairs with 96 different twin tags



Pool amplicons

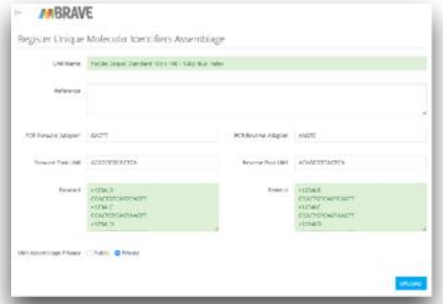


Illumina library prep

Illumina sequence

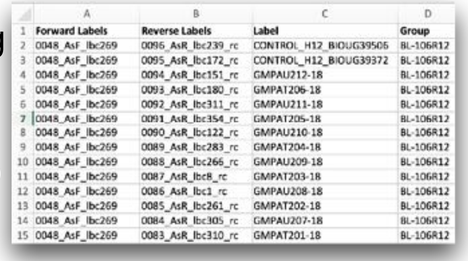
Upload fastq files to mBRAVE

Upload tag information to mBRAVE



Demultiplex by library and tag

Keep the most abundant read per tag (= well = beetle), filter out the rest, and Assign a taxonomy to using BOLD



	A	B	C	D
1	Forward Labels	Reverse Labels	Label	Group
2	0048_AsF_lbc269	0096_AsR_lbc239_rc	CONTROL_H12_BIOUG39505	BL-106R12
3	0048_AsF_lbc269	0095_AsR_lbc172_rc	CONTROL_H12_BIOUG39372	BL-106R12
4	0048_AsF_lbc269	0094_AsR_lbc151_rc	GMPAU212-18	BL-106R12
5	0048_AsF_lbc269	0093_AsR_lbc180_rc	GMPAT206-18	BL-106R12
6	0048_AsF_lbc269	0092_AsR_lbc311_rc	GMPAU211-18	BL-106R12
7	0048_AsF_lbc269	0091_AsR_lbc354_rc	GMPAT205-18	BL-106R12
8	0048_AsF_lbc269	0090_AsR_lbc122_rc	GMPAU210-18	BL-106R12
9	0048_AsF_lbc269	0089_AsR_lbc283_rc	GMPAT204-18	BL-106R12
10	0048_AsF_lbc269	0088_AsR_lbc266_rc	GMPAU209-18	BL-106R12
11	0048_AsF_lbc269	0087_AsR_lbc8_rc	GMPAT203-18	BL-106R12
12	0048_AsF_lbc269	0086_AsR_lbc1_rc	GMPAU208-18	BL-106R12
13	0048_AsF_lbc269	0085_AsR_lbc261_rc	GMPAT202-18	BL-106R12
14	0048_AsF_lbc269	0084_AsR_lbc305_rc	GMPAU207-18	BL-106R12
15	0048_AsF_lbc269	0083_AsR_lbc310_rc	GMPAT201-18	BL-106R12

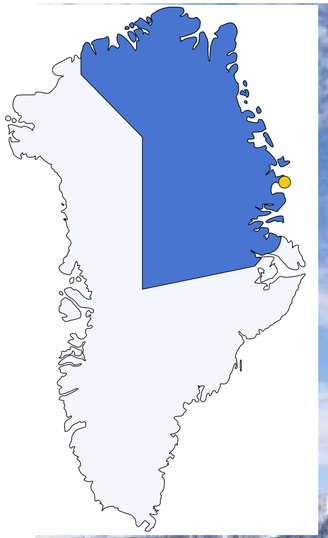
A consensus COI sequence for <u>each beetle</u>, plus metadata, in fasta file: **count data!!**

```
>ANML_DLJ06_E12|run:ANML_DLJ06_dtd_0|contig-id:1|rep_count:1169|id_similarity:98.
8950276243|c_count:1|cmxd:1.1|cmnd:0.1116338751|cnnd:None|p:None|c:None|o:None|f:None|g:
None|s:None|otu:BOLD:AAB7930|date:2019-08-13
AACTCTTTATTTTATTTTTGGTGCTTGGGCTGGAATAGTTGGTATATCCCTTAGTATACTGATTCGAGTTGAATTAAGAACACCTGGTA
CTCTACTAGGTGACGACCAAATATATAATGTAATCGTTACAGCTCATGCATTTATCATAATTTTCTTTATAGTTATACCTATTGTAATT
GGT
>ANML_DLJ04_C04|run:ANML_DLJ04_dtd_0|contig-id:1|rep_count:419|id_similarity:91.
1602209945|c_count:3|cmxd:1.7|cmnd:0.2888305489|cnnd:3.3|p:Arthropoda|c:Insecta|o:None|f:
None|g:None|s:None|otu:OTU14|date:2019-08-13
AACACTATATTTCATTTTTGGAGCTTGAGCAGGTATAGTTGGGACATCACTGAGATTACTAATTCGAGCAGAATTAGGAAACCCCGGAT
CTCTAATTGGTGATGACCAAATCTATAACGTAATTGTAACAGCCCATGCATTTGTAATAATTTTTTTTATAGTTATACCTATTGTAATT
GGT
>ANML_2909_D06|run:ANML_2909_dtd_0|contig-id:1|rep_count:20|id_similarity:nan|c_count:2|
cmxd:1.1|cmnd:0.255|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU37|date:
2019-08-01
TACTTTGTATTTCATTTTTGGTGCTTGAGCTGGAATAGTAGGAACCTCTTTAAGCCTTCTAATTCGAGCAGAATTAGGGAATCCTGGTT
CCCTAATTGGTGATGATCAGATTTATAATGTAATTGTAACTGCACATGCTTTCATTATAATTTTTTTTATAGTTATACCAATTGTTATT
GGA
>ANML_DLJ05_H08|run:ANML_DLJ05_dtd_0|contig-id:1|rep_count:6196|id_similarity:nan|c_count
:6|cmxd:1.7|cmnd:0.103518399|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU2|
date:2019-08-14
CACTCTTTATTTTATCTTTGGAGCATGGGCTGGAATACTTGGAACATCCTTAAGACTCCTCATCCGAGCAGAACTTGGAAACCCCGGGTT
CATTAATCGGAAACGATCAAATTTACAACGTAATTGTTACTGCCCACGCTTTCATTATAATTTTCTTCATAGTAATACCAATCATAATT
GGT
>ANML_2922_F09|run:ANML_2922_dtd_0|contig-id:1|rep_count:1342|id_similarity:nan|c_count:2
|cmxd:1.7|cmnd:0.1236214605|cnnd:2.2|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU7|
date:2019-08-02
ATCCTTATACTTTCTTTTTGGTATATGATCAGGAATAGTTGGAACCTCATTAAGATTACTAATTCGCCTAGAATTAGGTAATCCTGGGT
CTTTAATTGGAGACGACCAAATTTATAATGTAATTGTAACTGCCCATGCATTTATTATAATTTTCTTTATAGTTATACCTGTAATAATA
GGA
>ANML_DLJ01_F05|run:ANML_DLJ01_dtd_0|contig-id:3|rep_count:10|id_similarity:nan|c_count:3
|cmxd:0.6|cmnd:0.24|cnnd:26.1|p:None|c:None|o:None|f:None|g:None|s:None|otu:OTU115|date:
2019-08-13
AACCCTTTACTTCATTTTCGGAGCATGAAGAGGTATAATCGGAACCTCCATAAGATTGATGATTCGAACAGAATTAGGAACAGCTGGCT
CTCTAATTGGAGATGATCAAATTTATAATGTAATTGTAACAGCACATGCTTTTATCATAATTTTTTTCATAGTAATACCCATCATGATA
GGA
```

Methods to extract abundance information from DNA data

- Single-species quantitative PCR (qPCR)
- Multiplexed individual barcoding (mBRAVE): *individual count data in the 1000s*
- **Mitogenomics and DNA spike-in (SPIKEPIPE)**
- Metabarcoding and DNA spike-in (qSeq)
- Reverse metagenomics (RevMet)

Zackenberg Research Station, Greenland



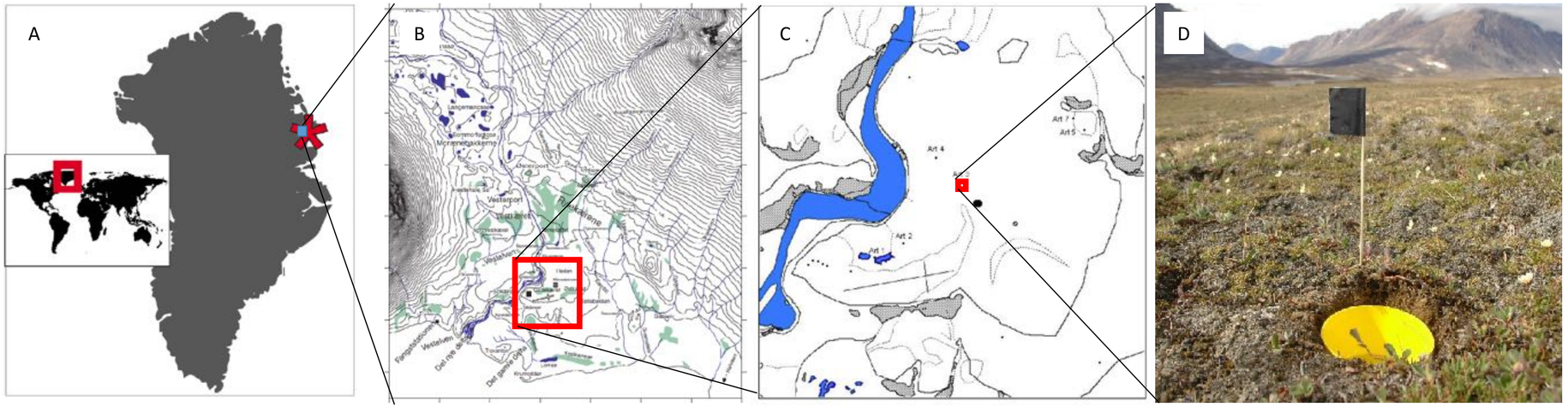
DOI: 10.1111/1755-0998.13057

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES

SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes

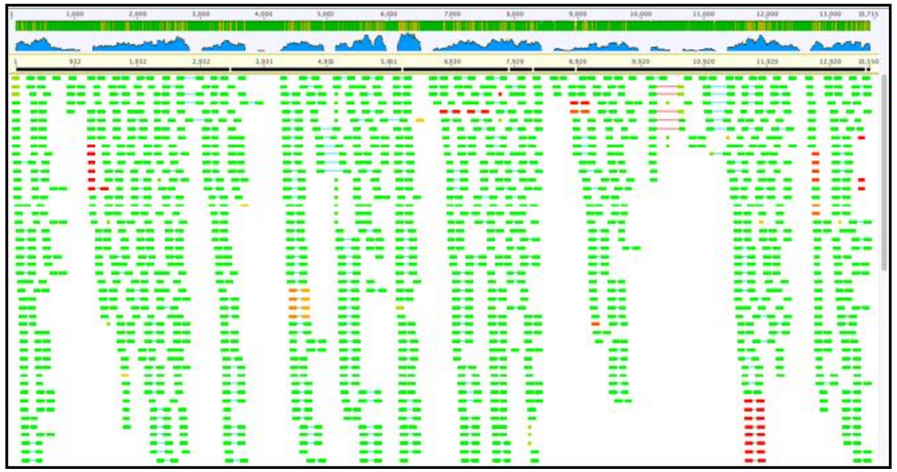
Yinqiu Ji[1*] | Tea Huotari[2*] | Tomas Roslin[2,3] | Niels Martin Schmidt[4,5] | Jiaxin Wang[1] | Douglas W. Yu[1,6,7] | Otso Ovaskainen[8,9]


A
B
C
D

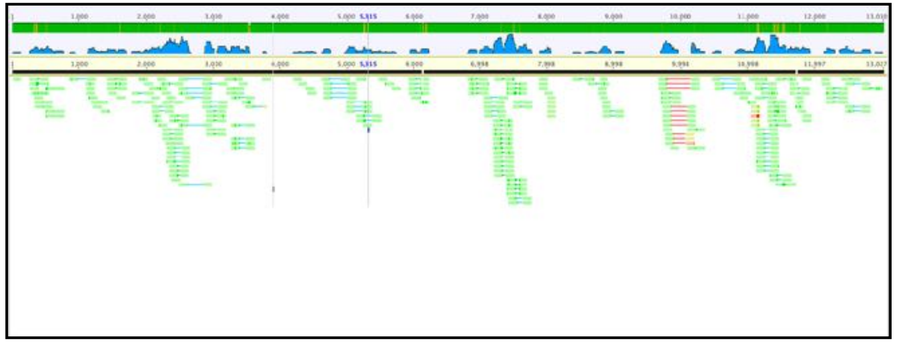

PÅFYLDT
JAN 2015

- entire aboveground arthropod community **~375 species**
- **>760,000 arthropods** collected in weekly samples and multiple pan traps from **1996-2013** (and ongoing)
- We assembled **308 mitogenome** sequences
- We shotgun-sequenced **~750 samples**: **3 samples per week from 1997-2013** + technical replicates

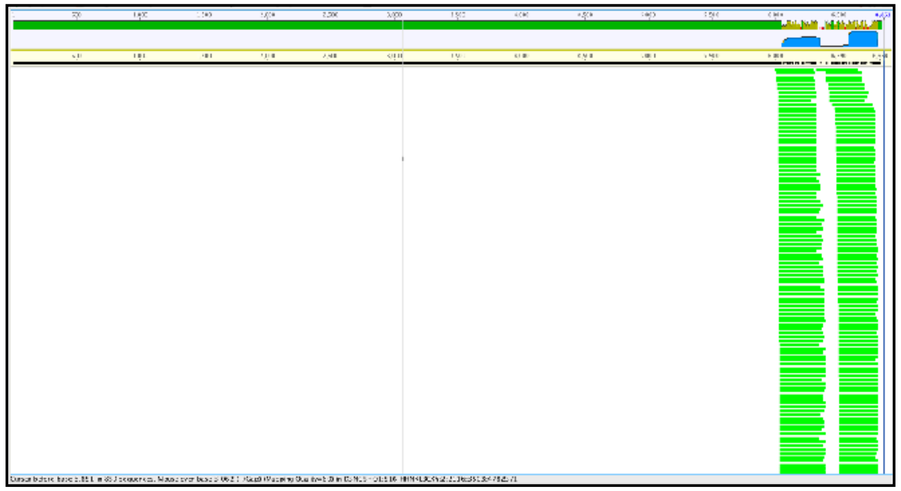
We mapped each sample's short reads to 308 mitogenomes



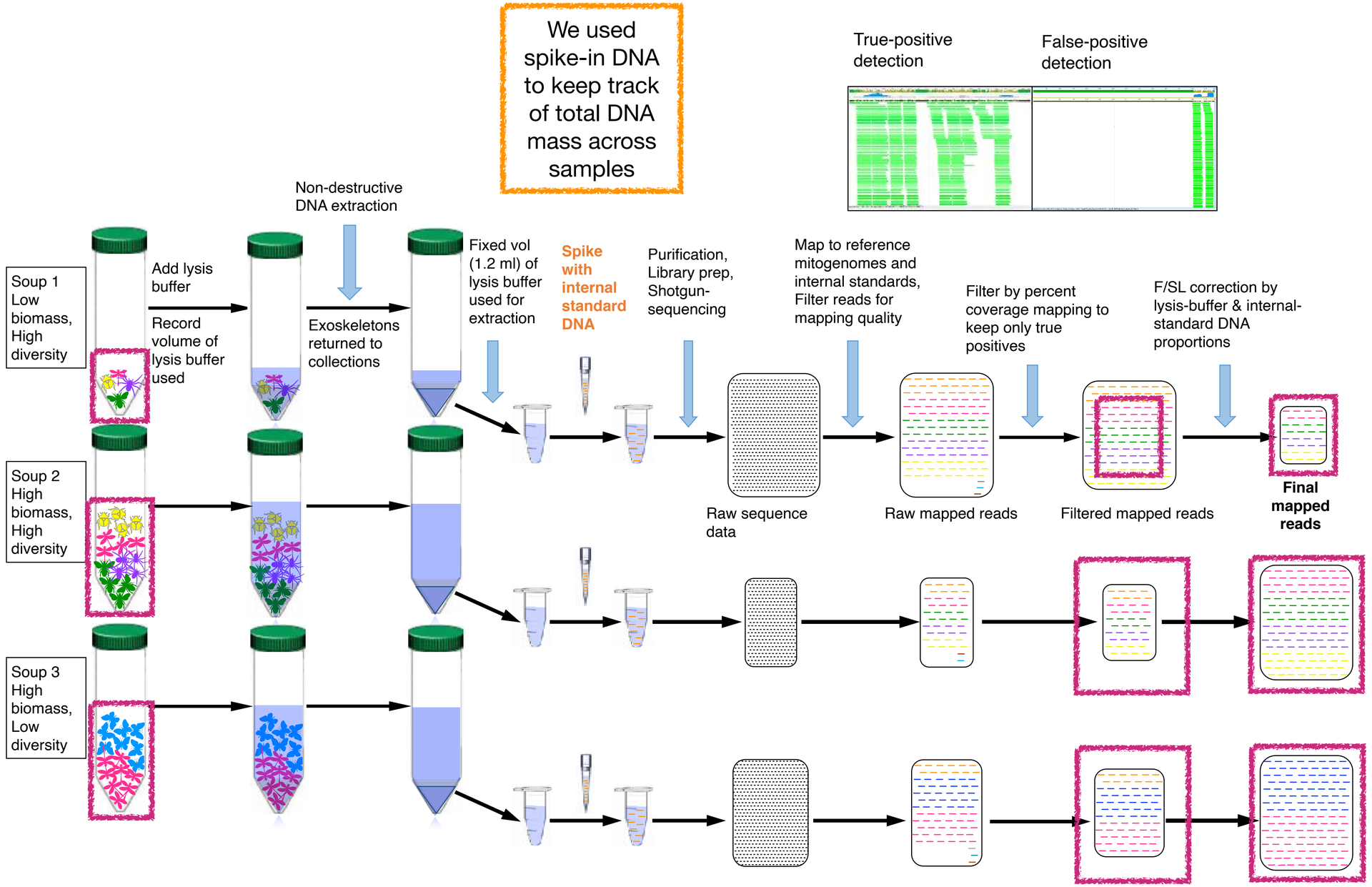
This mitogenome (a species) is confidently present in high biomass

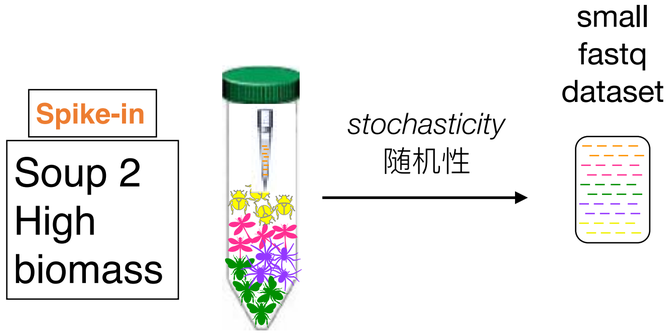


This species is confidently present in low biomass



This species is not present, despite the many mapped reads (they mapped only to a conserved sequence)


We used spike-in DNA to keep track of total DNA mass across samples
True-positive detection
False-positive detection
Non-destructive DNA extraction
Soup 1 Low biomass, High diversity
Soup 2 High biomass, High diversity
Soup 3 High biomass, Low diversity
Add lysis buffer
Record volume of lysis buffer used
Exoskeletons returned to collections
Fixed vol (1.2 ml) of lysis buffer used for extraction
Spike with internal standard DNA
Purification, Library prep, Shotgun-sequencing
Map to reference mitogenomes and internal standards, Filter reads for mapping quality
Filter by percent coverage mapping to keep only true positives
F/SL correction by lysis-buffer & internal-standard DNA proportions
Raw sequence data
Raw mapped reads
Filtered mapped reads
Final mapped reads



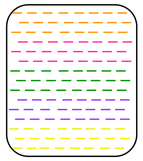
Sample 1	Reads_Raw
Taxon1	**1000**
Taxon2	**1200**
Taxon3	**1600**
Taxon4	**2000**
Taxon5	**400**
Taxon6	**200**
Sample 2	Reads_Raw
Taxon1	500
Taxon2	600
Taxon3	800
Taxon4	1000
Taxon5	200
Taxon6	100

Spike-in

Soup 1
Low
biomass

stochasticity
随机性

big
fastq
dataset



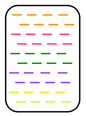
Sample 1	Reads_Raw
OTU1	**1000**
OTU2	**1200**
OTU3	**1600**
OTU4	**2000**
OTU5	**400**
OTU6	**200**
OTU_SPIKE	**40**

Spike-in

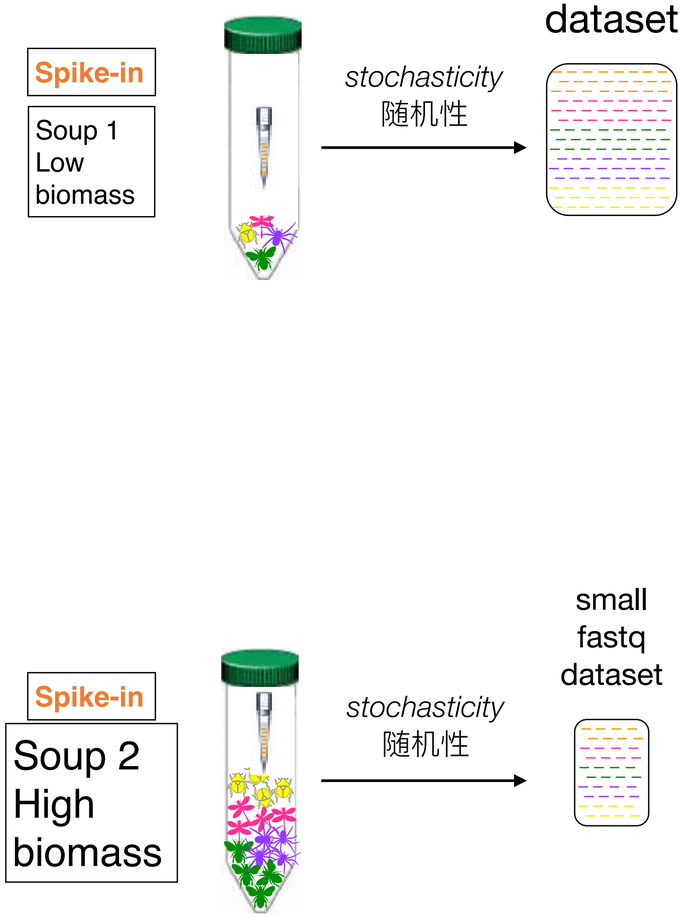
Soup 2
High
biomass

stochasticity
随机性

small
fastq
dataset



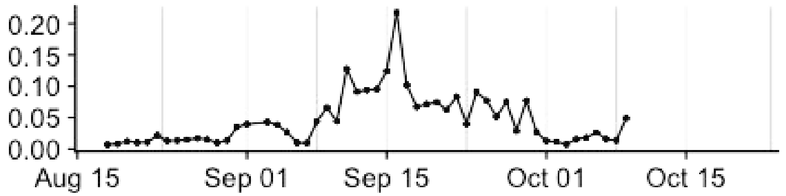
Sample 2	Reads_Raw
OTU1	500
OTU2	600
OTU3	800
OTU4	1000
OTU5	200
OTU6	100
OTU_SPIKE	**10**



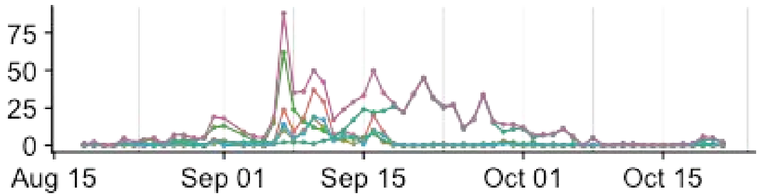
Sample 1	Reads_Raw	Reads_Corr	
OTU1	**1000**	**25**	**1000/40 = 25**
OTU2	**1200**	**30**	**1200/40 = 30**
OTU3	**1600**	**40**	**1600/40 = 40**
OTU4	**2000**	**50**	**2000/40 = 50**
OTU5	**400**	**10**	**400/40 = 10**
OTU6	**200**	**5**	**200/40 = 5**
OTU_SPIKE	**40**		
Sample 2	Reads_Raw	Reads_Corr	
OTU1	500	**50**	
OTU2	600	**60**	
OTU3	800	**80**	
OTU4	1000	**100**	
OTU5	200	**20**	
OTU6	100	**10**	
OTU_SPIKE	**10**		

Two kinds of quantification

- **Within-species quantification**: **“Species A is more abundant in these samples, less abundant in those samples.” <u>Follow one species.</u>**

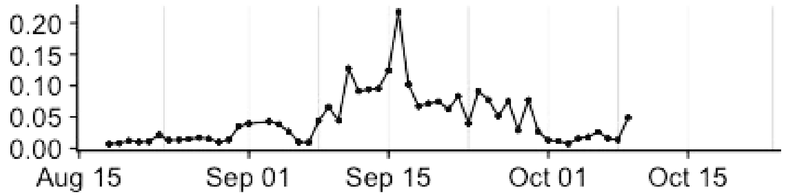


- ~~*Across-species quantification*~~: “Species A is more abundant than Species B in this sample” <u>Compare two species.</u>



- We tested quantification by making mock samples with 20 species, with different DNA quantities

How can within-species be quantification useful?



- **Within-species quantification**: **“Species A is more abundant in these samples, less abundant in those samples.” <u>Follow one species.</u>**

We can rescale within-species abundances to [0,1], known as quasiprobability

Information loss with Presence-Absence [0 or 1]

Potential importance of QP data, which preserves within-species frequency information

A. QP distribution data for one species

Species shows a preference for green habitat

		0.1
0.1	1.0	0.1
0.1	0.7	
0.1	0.1	0.1

optimal env conditions

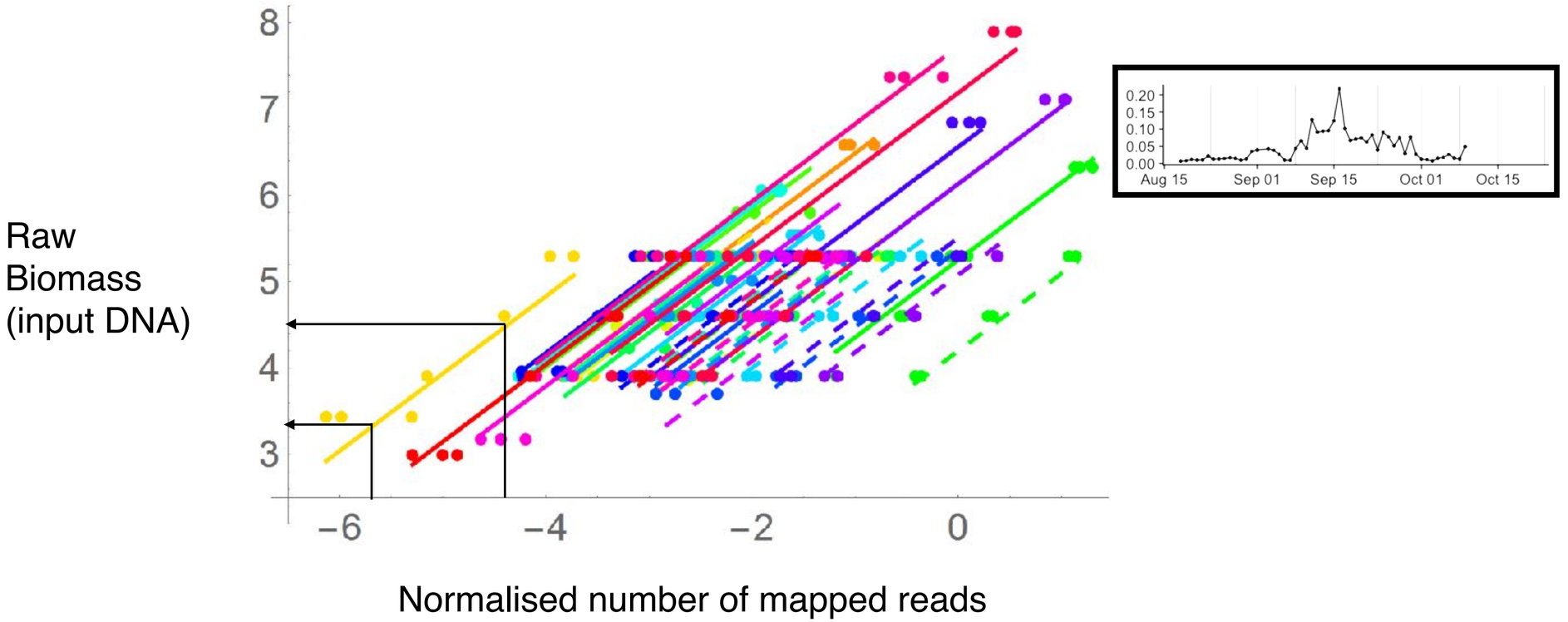
suboptimal env conditions

B. PA distribution data for one species

Species incorrectly shows no habitat preference

		1
1	1	1
1	1	
1	1	1

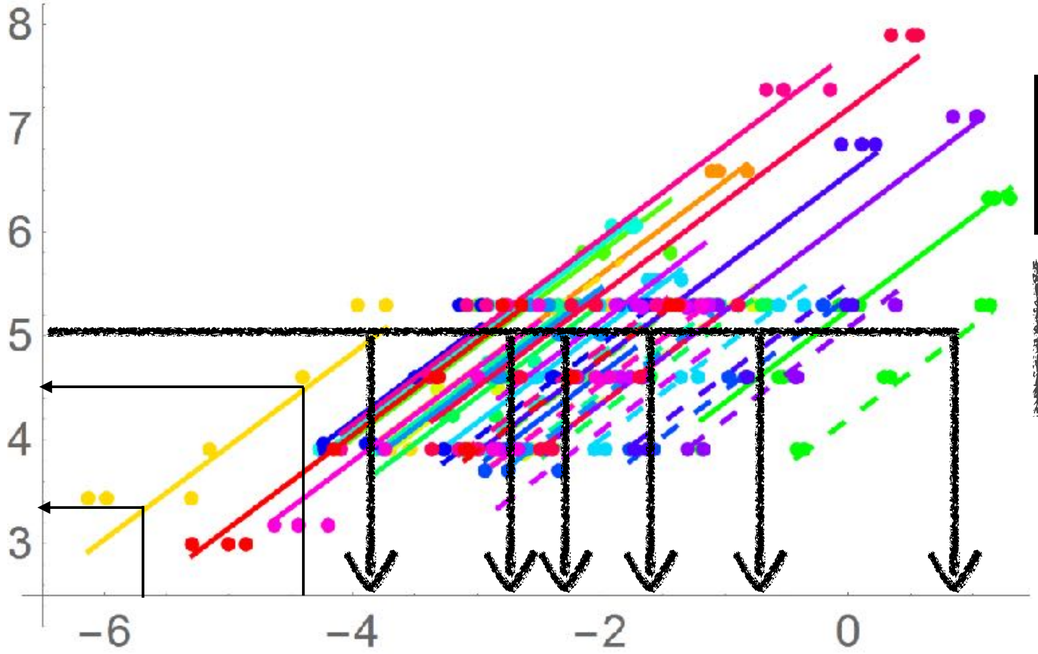
- **Within-species quantification** works well <u>when you use a spike-in</u>
  - Each colour is a different species. More mapped reads (x-axis) correctly predict more input DNA (y-axis)
  - Thus, we can observe how each species' population dynamics has changed from 1997 to 2013



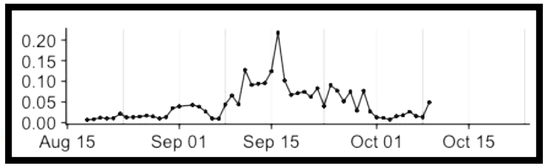
- **Within-species quantification** works well. Each species is a different colour. More mapped reads (x-axis) correctly predicts more input DNA (y-axis). So we can observe how each species' population dynamics has changed from 1997 to 2013

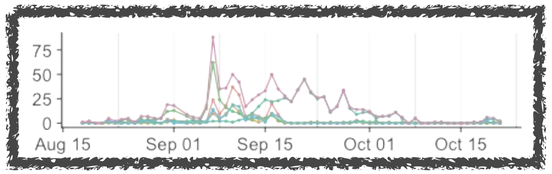
- ***Across-species quantification*** *does not work well. (we would need to have species correction factors)*

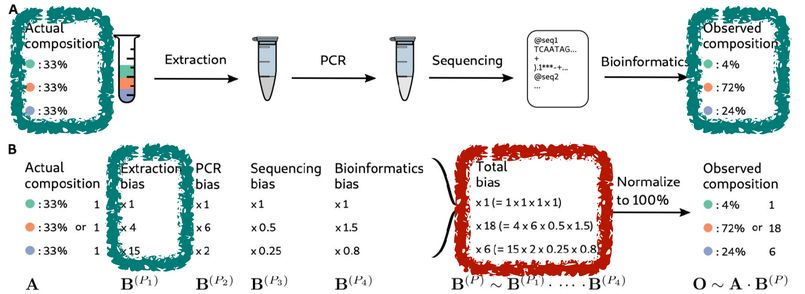
Raw Biomass (input DNA)



Normalised number of mapped reads







Example from pa_vs_qp_tables_20210402.xlsx

1. True OTU table

	OTU01	OTU02	OTU03	OTU04	spikeOTU	True rowSum
Sample 1	10	20	0	0	20	**30**
Sample 2	0	100	20	50	20	**170**
Sample 3	40	40	5	50	20	**135**
Sample 4	60	0	30	100	20	**190**

Metabarcoding pipeline

500 = 10 * 5 * 10

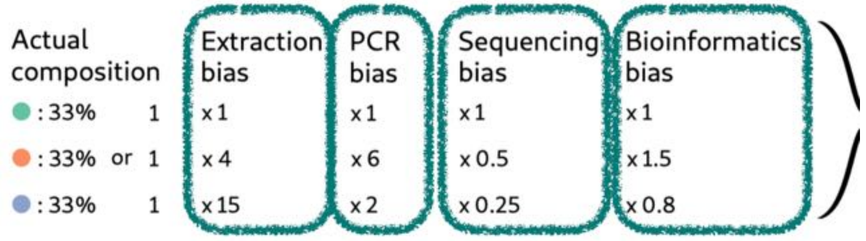
50 = 20 * 5 * 0.5

2. Observed OTU table with row noise and species-specific biases

	OTU01	OTU02	OTU03	OTU04	spikeOTU	row noise
Sample 1	500	200	0	0	50	5
Sample 2	0	1000	300	1000	50	5
Sample 3	1200	240	45	600	30	3
Sample 4	600	0	90	400	10	1
species-specific biases	10	2	3	4	0.5	

Laboratory stochasticity causes row noise
(e.g. from pooling amplicons for library prep)

Species-specific biases are caused by species properties

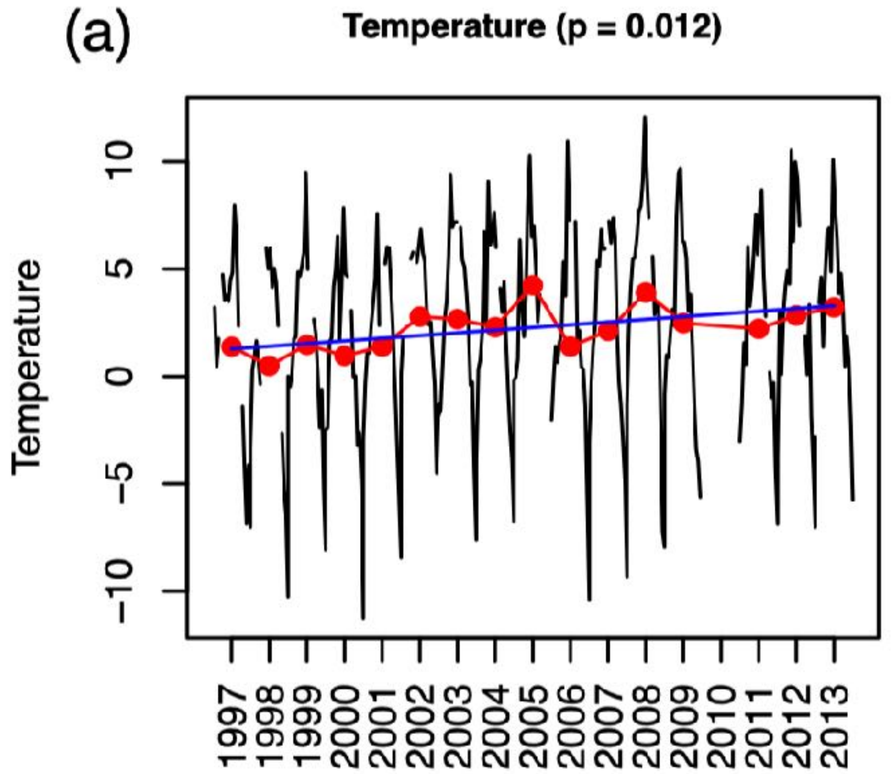


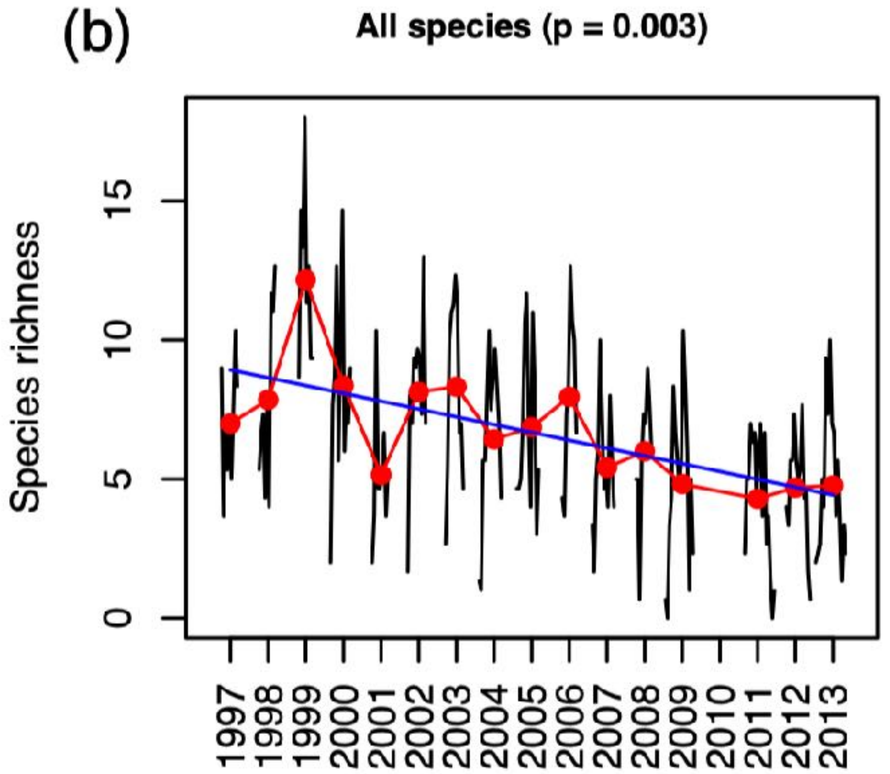
Research

Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change

Nerea Abrego, Tomas Roslin, Tea Huotari, Yinqiu Ji, Niels Martin Schmidt, Jiaxin Wang, Douglas W. Yu and Otso Ovaskainen

Ecography
44: 1–12, 2021
doi: 10.1111/ecog.05547





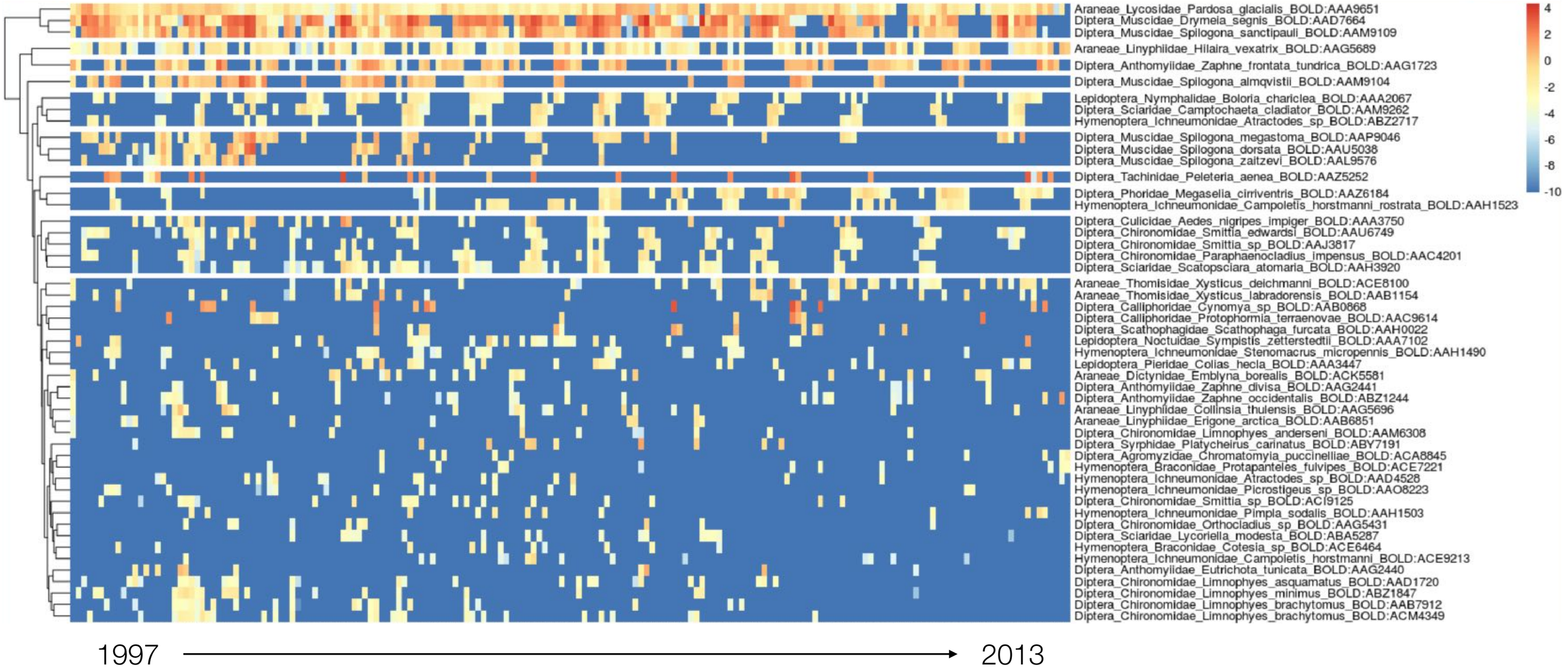
52 species in
542 trap-week samples

average summer temperature has increased by 2.0 °C (from 1.28 °C to 3.28 °C)

(observed) species richness has declined by half (8.9 to 4.4 species per trap-week, decline is mostly in the predator species)

What we get is a ***community time series***

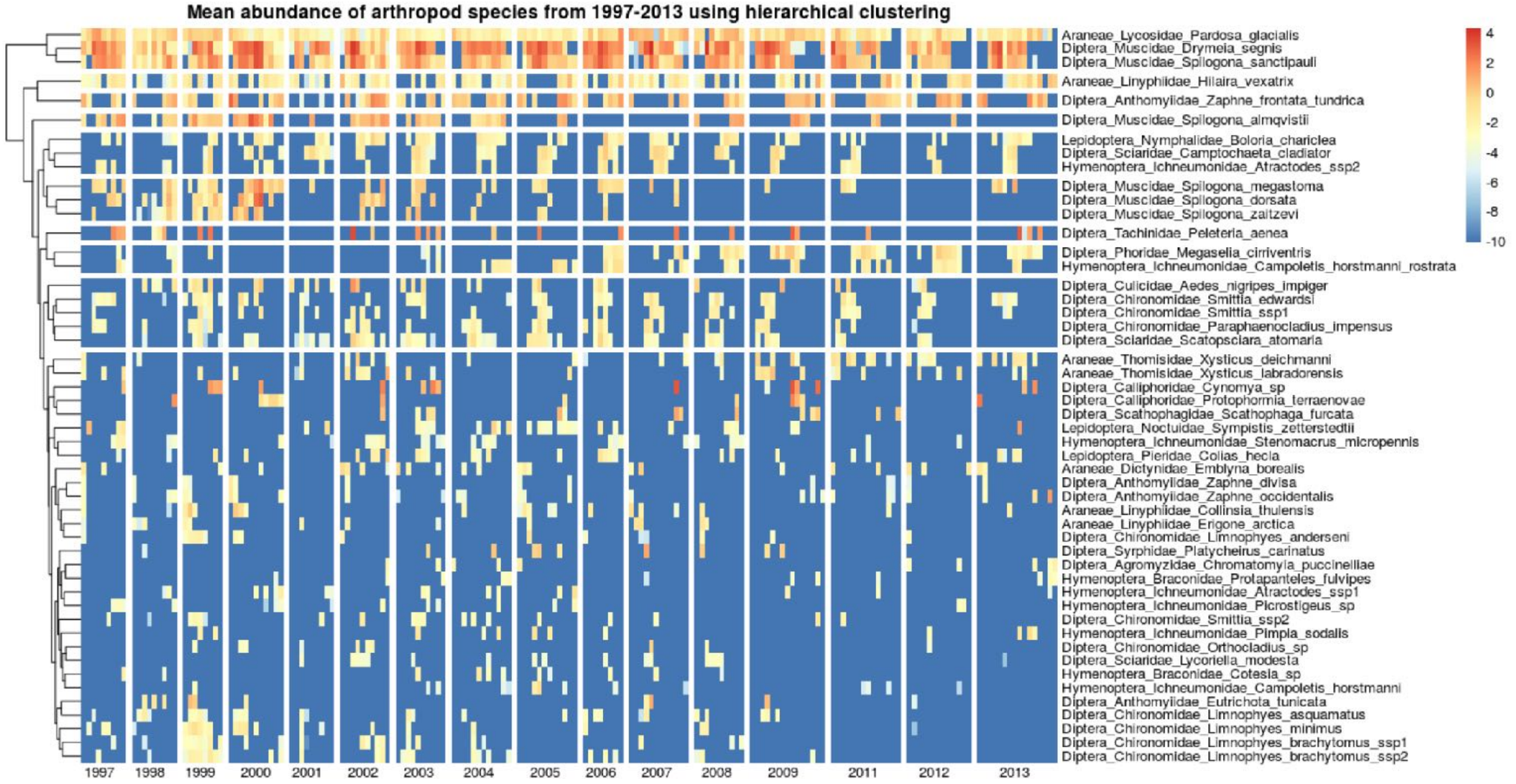
(red = highest **within-species** abundance, blue = absence)



more blue cells (more absences) toward 2013

What we get is a ***community time series***

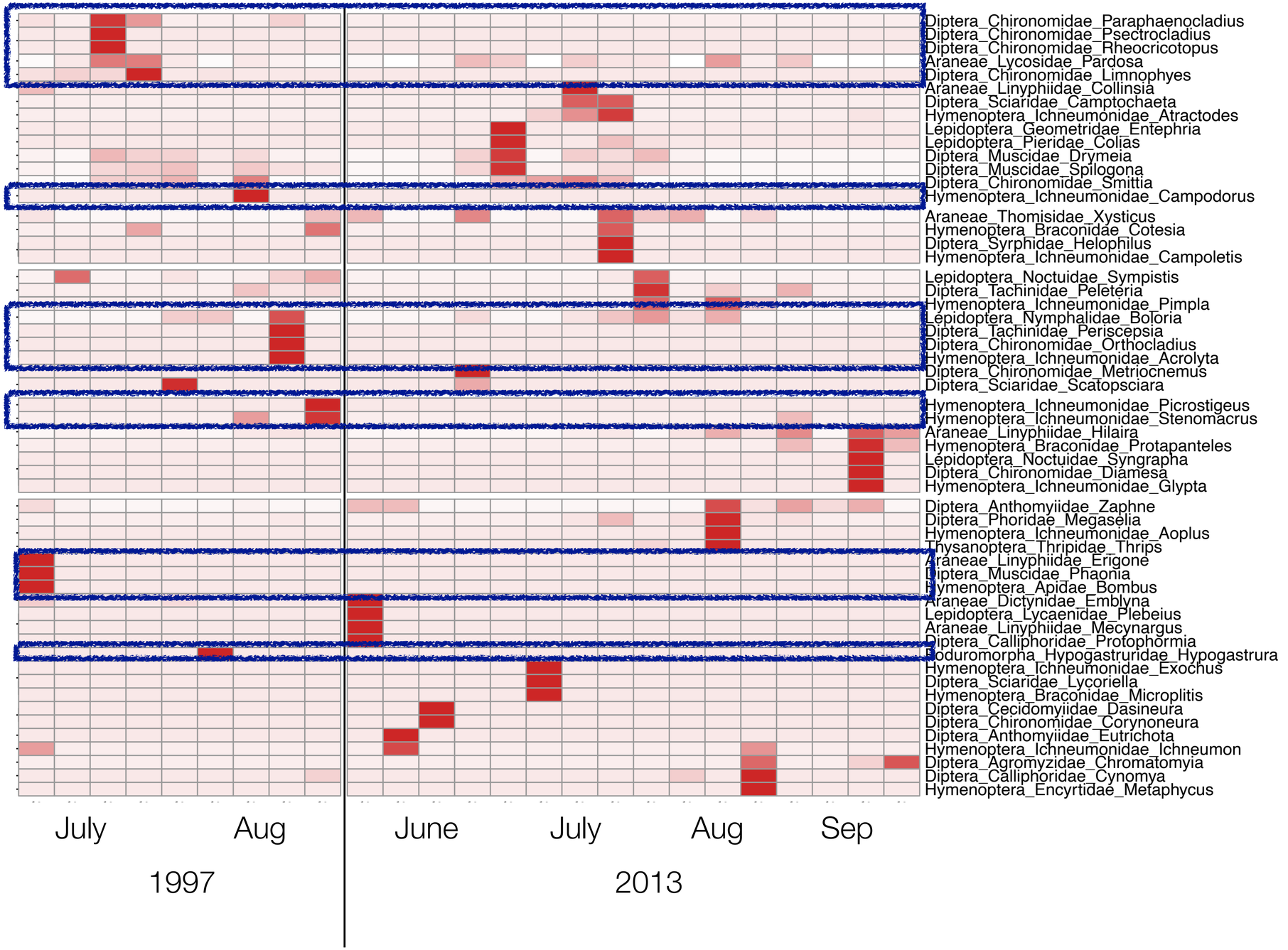
(red = highest **within-species** abundance, blue = absence)



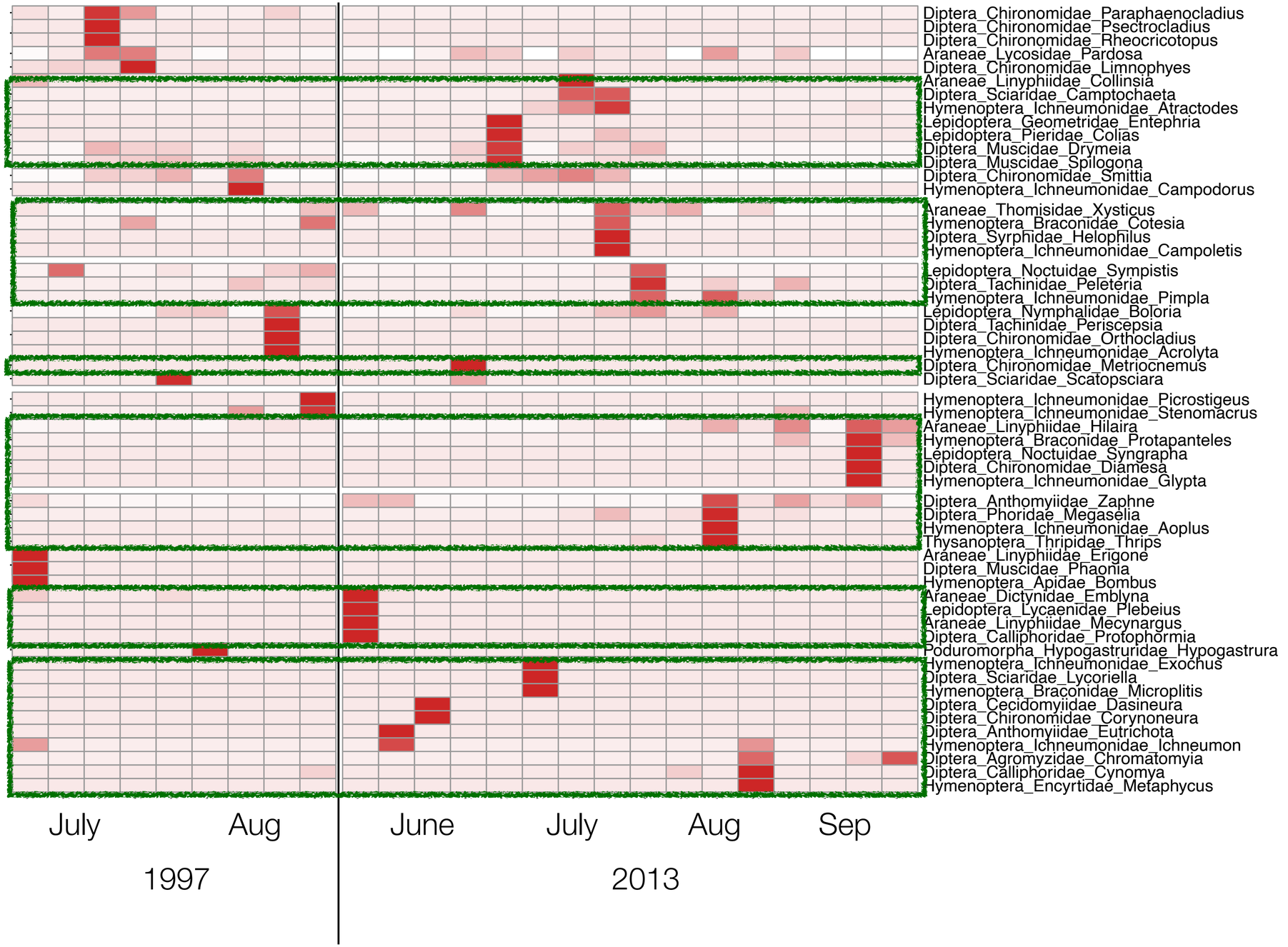
Summer has approx doubled in length in 16 years

Let’s look at 2 years of real data (dark red = more biomass of that species)

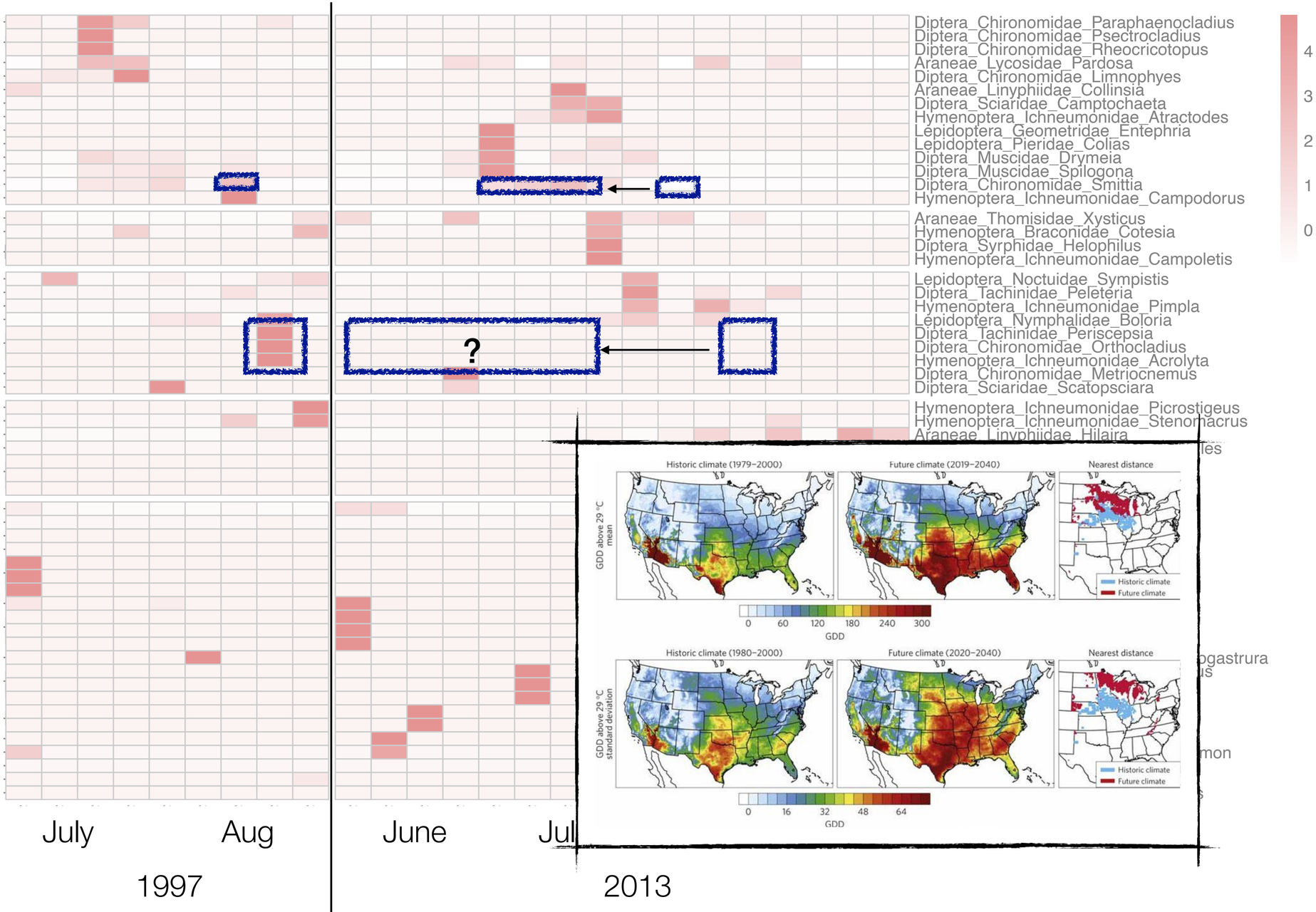
Many species that exhibited activity peaks in 1997, now no longer do



Many species that showed no activity peaks in 1997, now do have activity peaks!



We *rarely* see simple shifts in phenology (timing of peaks): the climate envelope idea

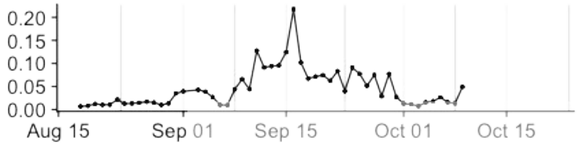


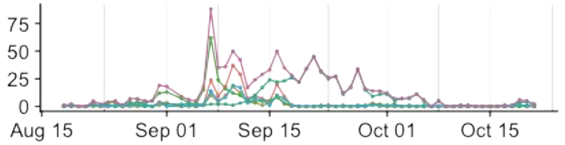
Methods to extract abundance information from DNA data

- Single-species quantitative PCR (qPCR)
- Multiplexed individual barcoding (mBRAVE)
- Mitogenomics and DNA spike-in (SPIKEPIPE)
- **Metabarcoding and DNA spike-in (qSeq)**
- Reverse metagenomics (RevMet)

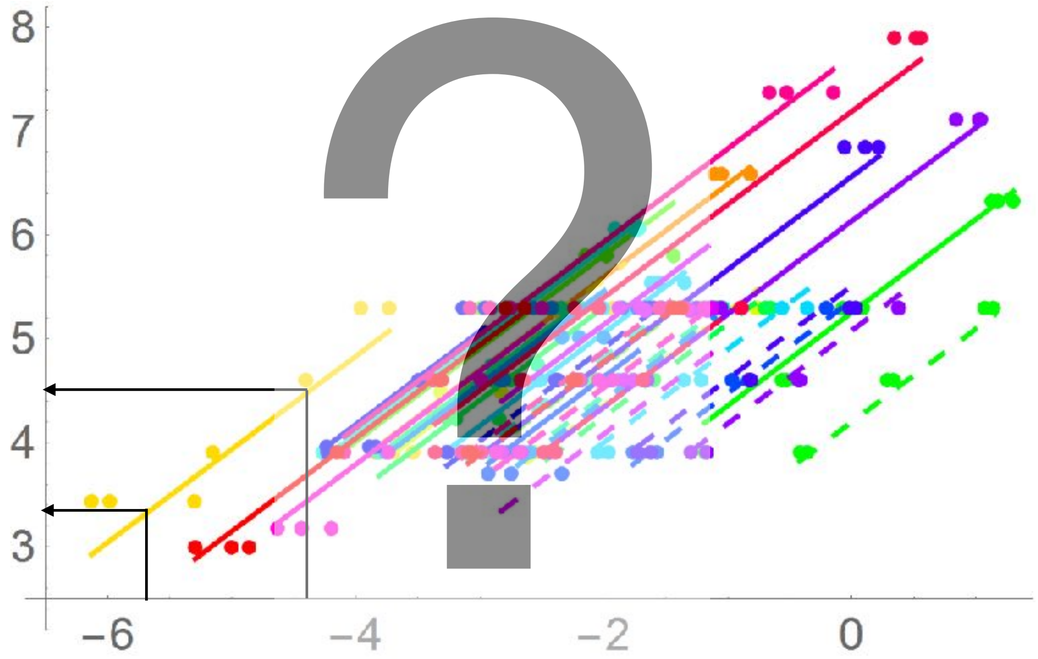
Is quantitative **metabarcoding** possible (qSeq)?

Mingjie Luo, Yinqiu Ji, Yuanheng Li, Douglas Yu





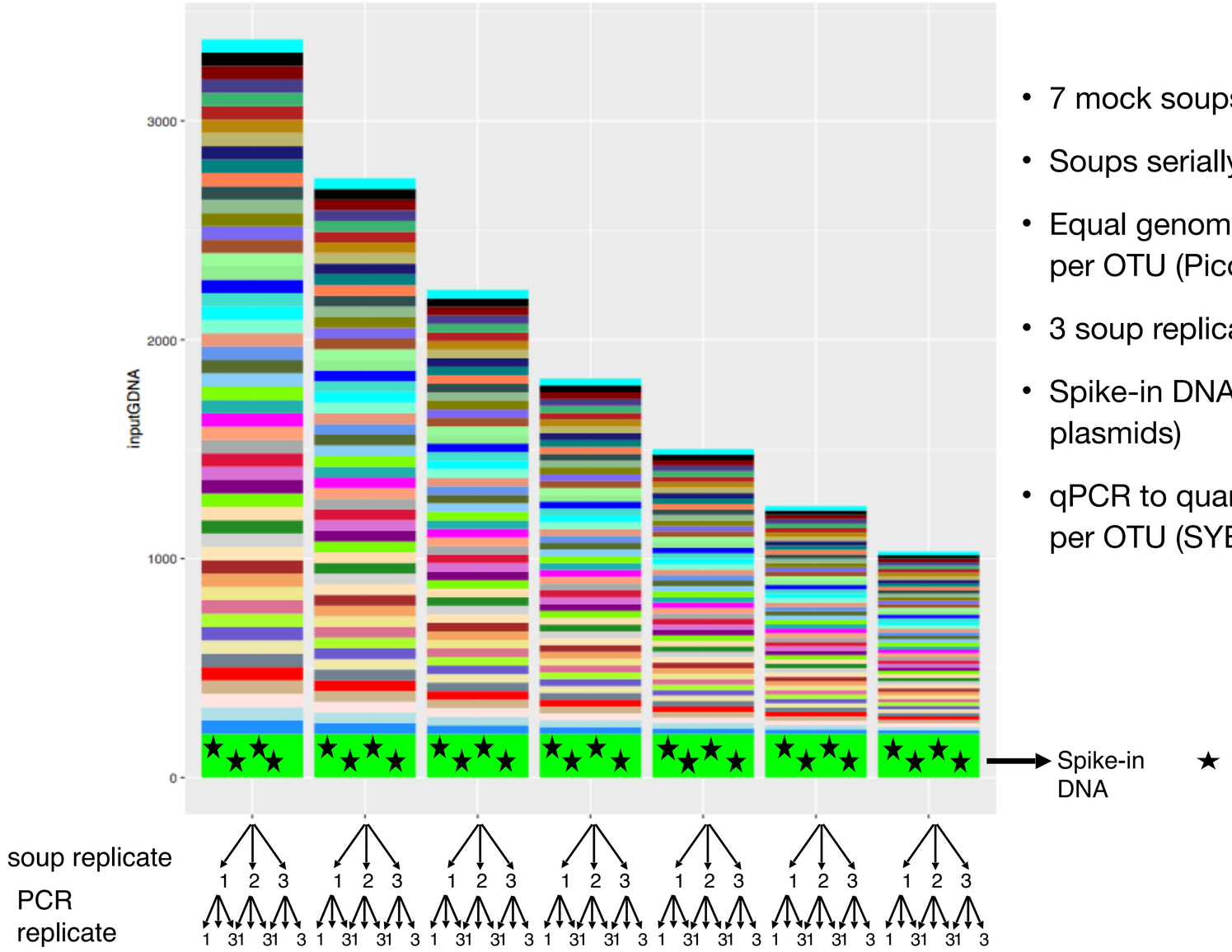
Raw
Biomass
(input DNA)



Normalised number of mapped reads

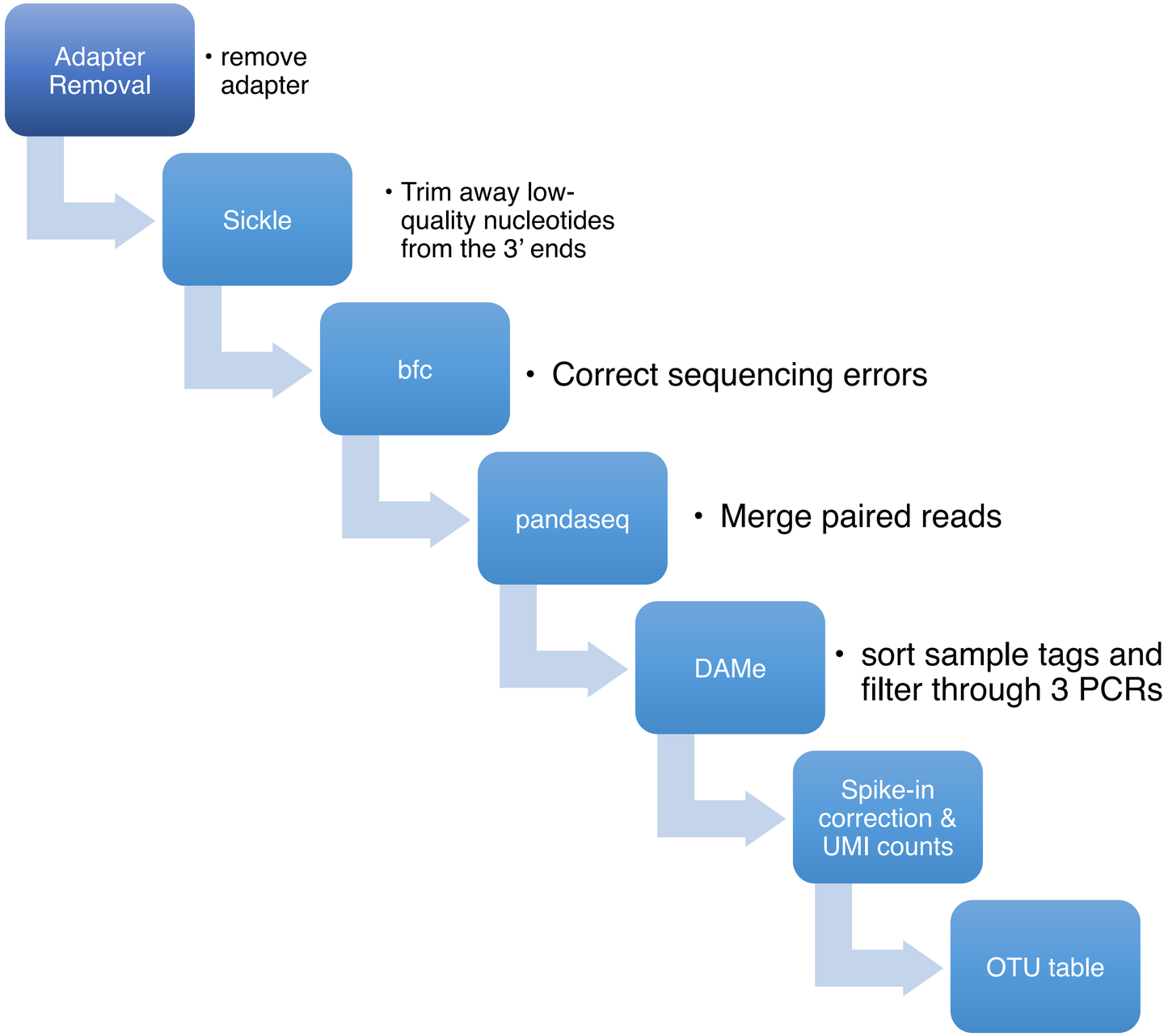
With metabarcoding, we could avoid the initial cost of creating a mitogenome dataset

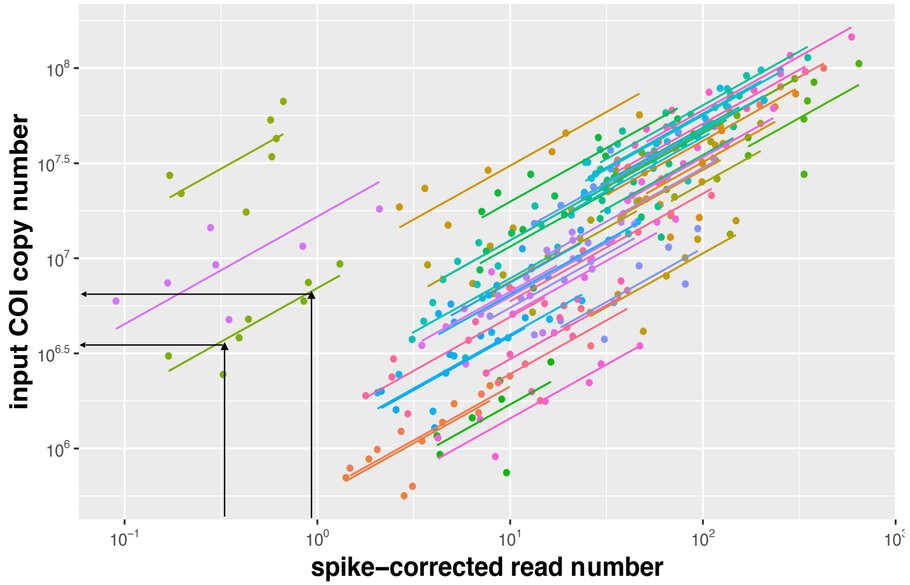
qSeq Experiment



- 7 mock soups, 52 insect OTUs.
- Soups serially diluted by 0.8X
- Equal genomic DNA concentration per OTU (PicoGreen)
- 3 soup replicates, 3 PCR replicates.
- Spike-in DNA (3 DNA barcodes in plasmids)
- qPCR to quantify COI copy number per OTU (SYBR green)

Bioinformatic Pipeline



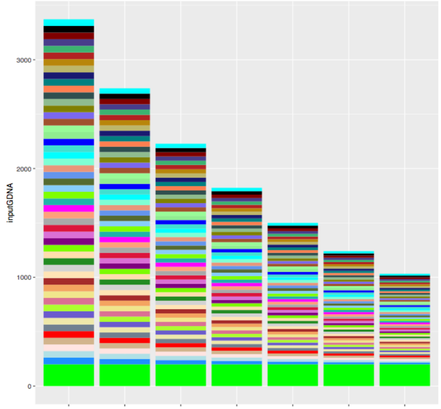


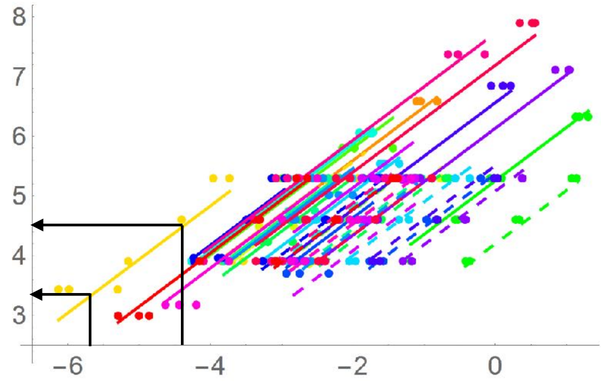
It works!

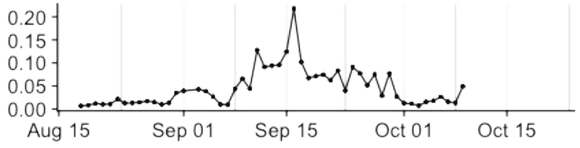
input_COI_copy_num ~ **readNum_spikeCorrected** + otuID

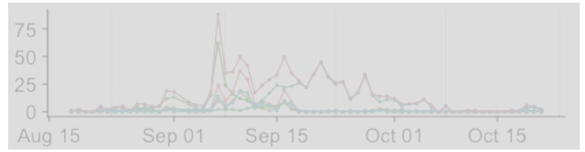
slope= 0.56, R^2=0.95 (soup and PCR replicates averaged)

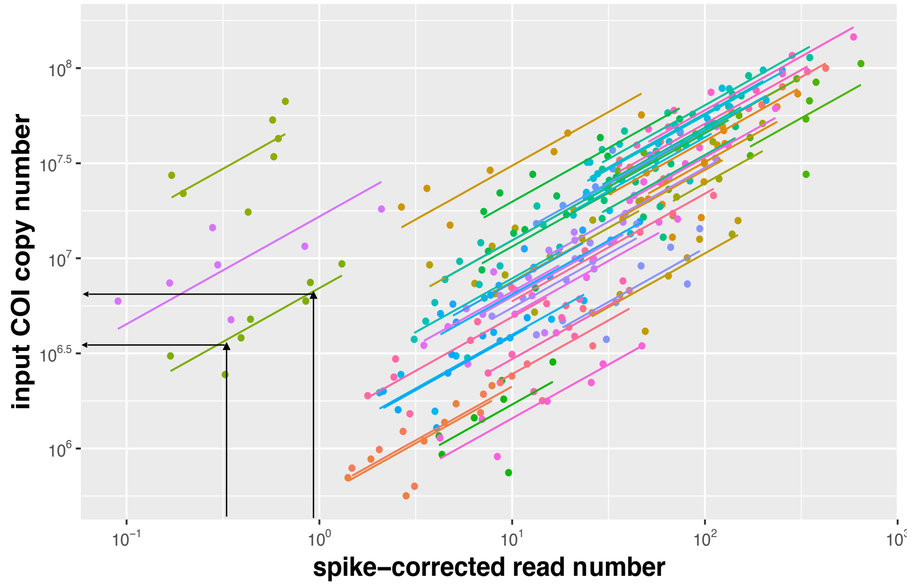
As with mitogenomics, we can estimate **within-species** change in abundance (DNA mass)









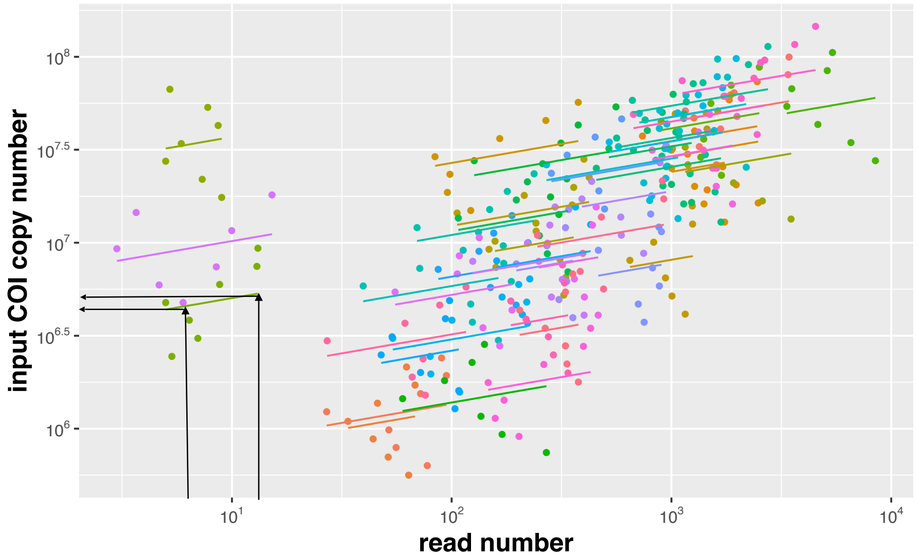


It works!

input_COI_copy_num ~ **readNum_spikeCorrected** + otuID

slope= 0.56, R^2=0.95 (soup and PCR replicates averaged)

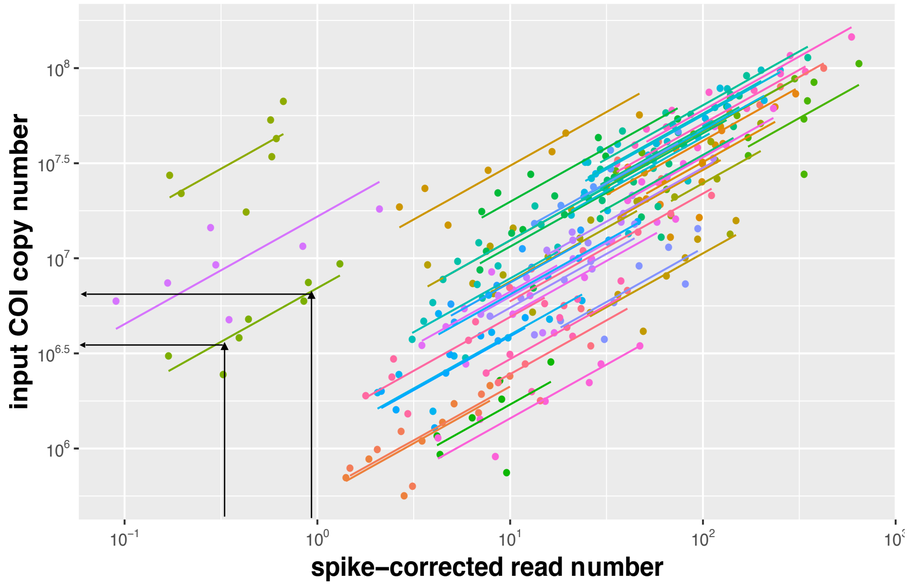
We can estimate **within-species** change in abundance (COI copy number)



Spike correction is necessary: increases slope and R^2

input_COI_copy_num ~ **readNum** + otuID

slope= 0.2, R^2=0.85 (soup and PCR replicates averaged)



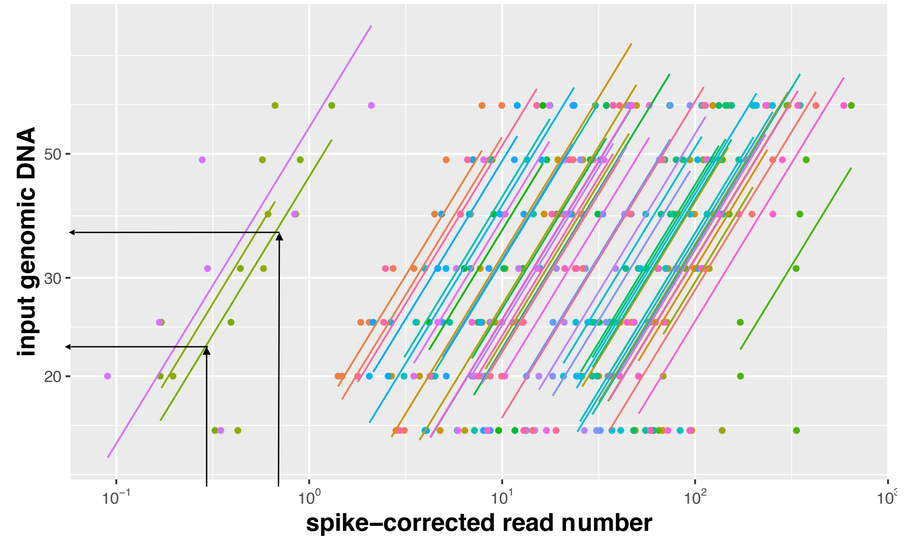
It works!

input_COI_copy_num ~ **readNum_spikeCorrected** + otuID

slope= 0.56, R^2=0.95 (soup and PCR replicates averaged)

We can estimate **within-species** change in abundance (COI copy number)

Spike correction is necessary: increases slope and R^2

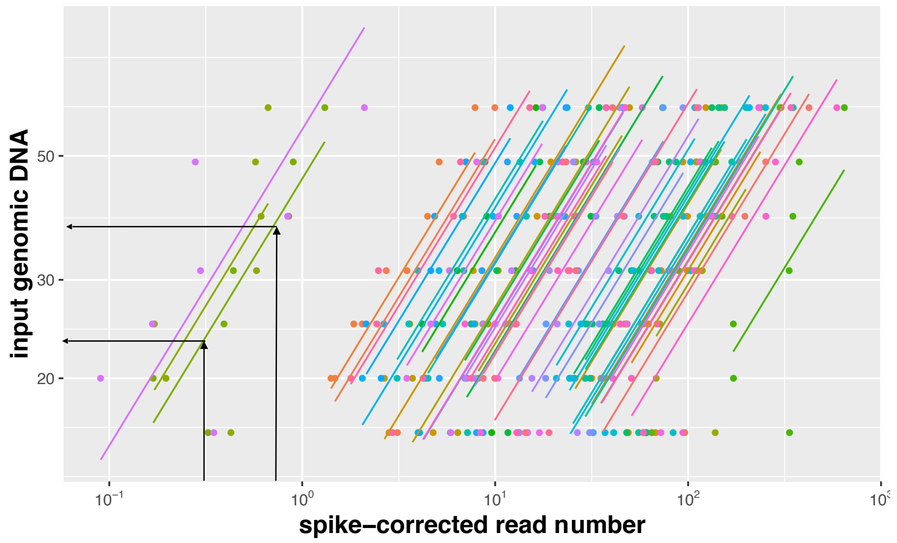


We can also predict input genomic DNA (our real goal)

input_gDNA ~ readNum_SpikeCorr + otuID

slope= 0.57, R^2=0.62 (soup and PCR replicates averaged)

Note that the y-axis scale is different (ng, not copy number)



It works!

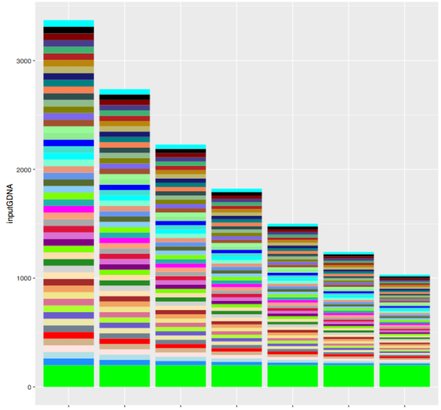
input_gDNA ~ readNum_SpikeCorr + otuID

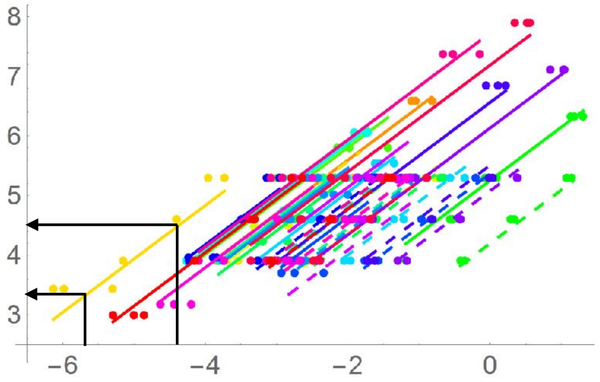
slope= 0.57, R^2=0.62 (soup and PCR replicates averaged)

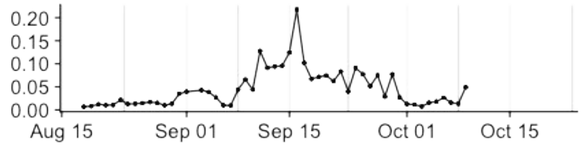
We can estimate **within-species** change in abundance (as genomic DNA == our real goal)

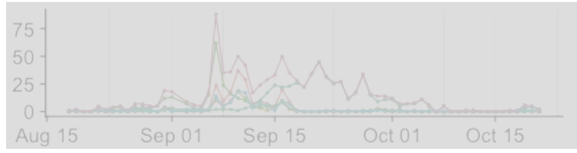
Spike correction is necessary: increases slope and R^2

We can indeed use metabarcoding to estimate within-species change in abundance (gDNA mass)









Methods to extract abundance information from DNA data

- Single-species quantitative PCR (qPCR)
- Multiplexed individual barcoding (mBRAVE)
- Mitogenomics and DNA spike-in (SPIKEPIPE)
- Metabarcoding and DNA spike-in (qSeq)
- **Reverse metagenomics (RevMet)**

We want to identify pollen to species **with across-species quantification** (which pollen species is more abundant in the honeybee's diet?)

DOI: 10.1111/2041-210X.13265

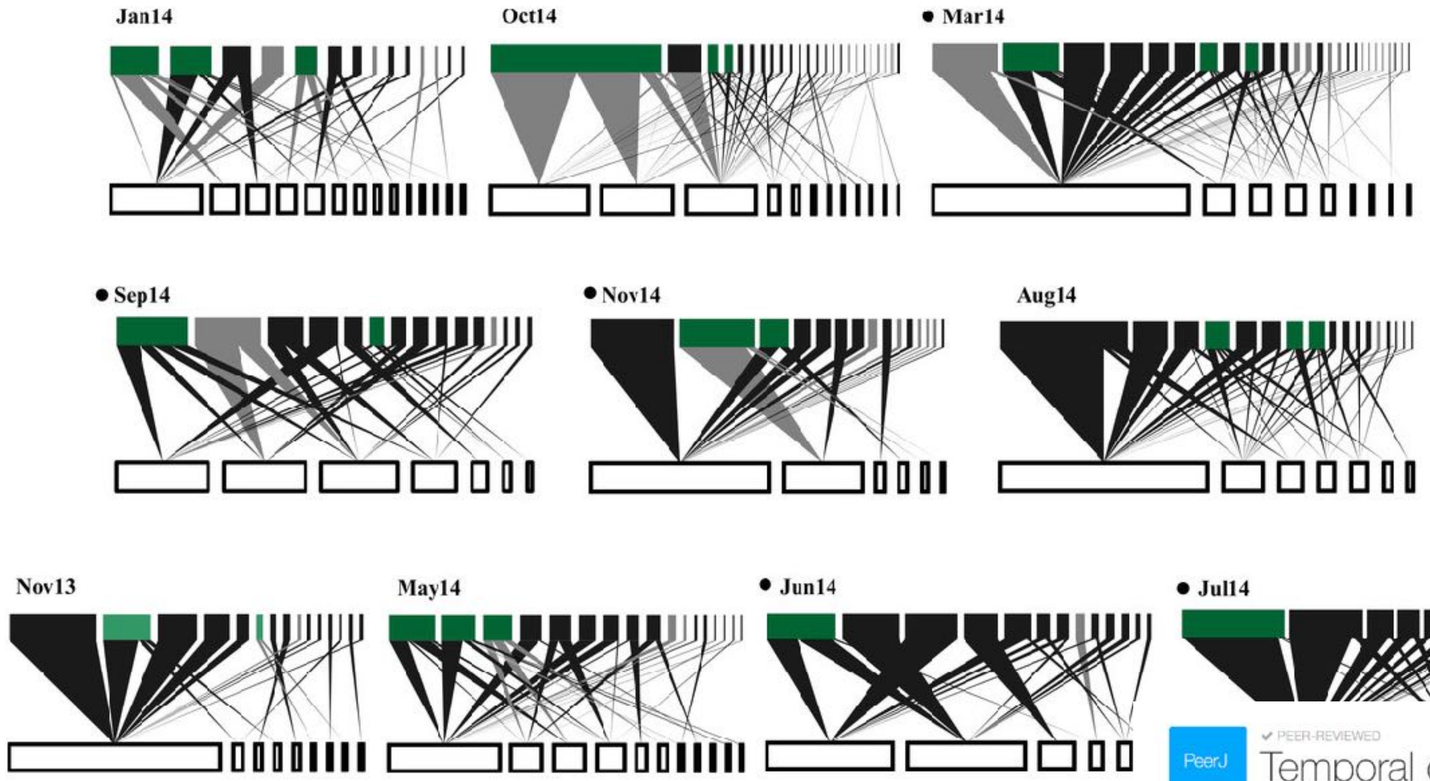
RESEARCH ARTICLE

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet)

Ned Peel[1,2] | Lynn V. Dicks[2] | Matthew D. Clark[1,3] | Darren Heavens[1] | Lawrence Percival-Alwyn[1] | Chris Cooper[4] | Richard G. Davies[2] | Richard M. Leggett[1] | Douglas W. Yu[2,5,6]

Basic: Construct quantitative pollinator-plant networks



PeerJ

✔ PEER-REVIEWED

Temporal changes in the structure of a plant-frugivore network are influenced by bird migration and fruit availability

Related research

Biodiversity | Biogeography | Conservation Biology | Ecology | Zoology

Michelle Ramos-Robles[1], Ellen Andresen[2], Cecilia Díaz-Castelazo[1]

Published June 8, 2016 PubMed 27330852

Applied: Diet analysis to identify preferred plant species for bees: **need across-species quantification**

Crop plants require a pulse of pollination, which requires a large wild bee population

Preferred plants differ across bee species and might differ over time

Our goal – a protocol that:

- Does not need specialist molecular-lab skills
- Can be used on small amounts of pollen

Plastid:Nuclear DNA ratio might vary across pollen grains and plant species

- Uses the **nuclear genome**, not (just) the chloroplast genome

Ease of building references

- Does **NOT require assembled genomes** for the reference database

We want the **dominant** pollen

- Is '**semi-quantitative**' (low vs. high biomass frequency)
- Can **identify pollen to plant species**

MinION sequencer

- **long reads** (1000s of bp per read)
- ~ 85-95% accuracy rate
- low-input DNA kit (200 ng)

SQK-RBK001

Rapid Barcoding Kit

$672.00

− 1 + Add to cart

Simple and rapid library preparation, with barcoding for up to 12 gDNA samples

gDNA | 200 ng | < 10 min | No PCR | Speed / simplicity

More information

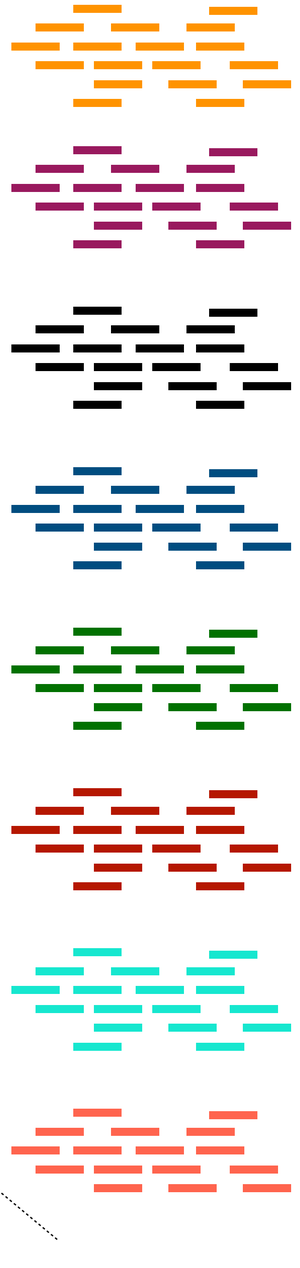
MinION metagenomics

Reference dataset

Reference database: each species is a 0.5X **'genome skim': a cloud of short reads**

Illumina sequencing

100-250 bp long



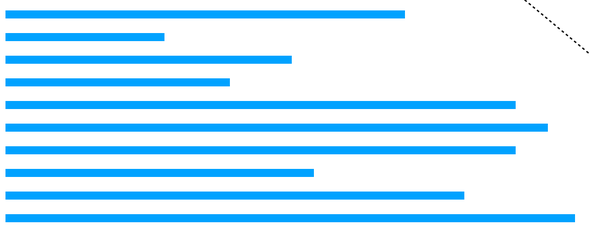
Query ('question') dataset

Sample database: each pollen sample is sequenced without PCR

Pollen sample

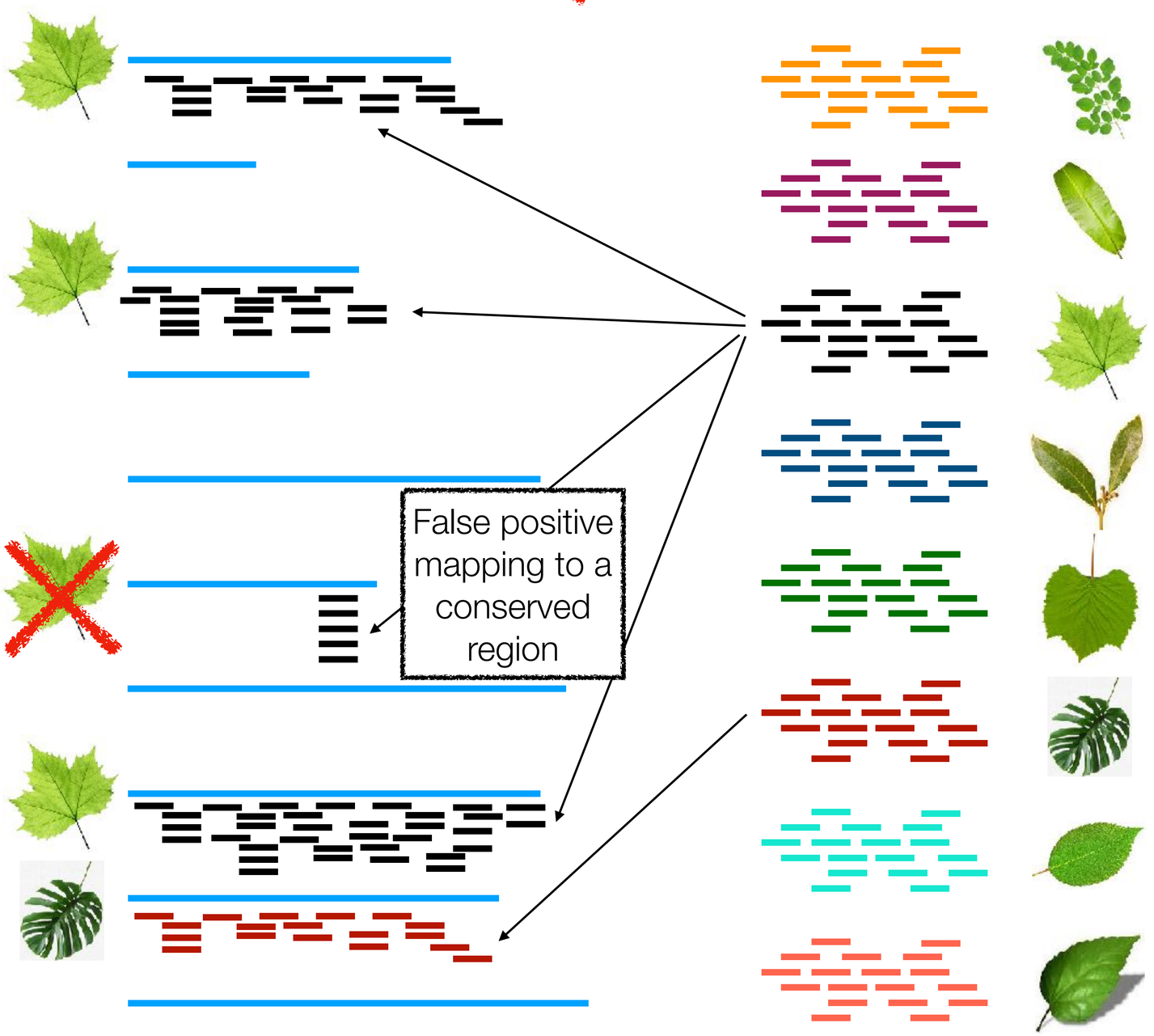
MinION sequencing

100s to 1000s bp long

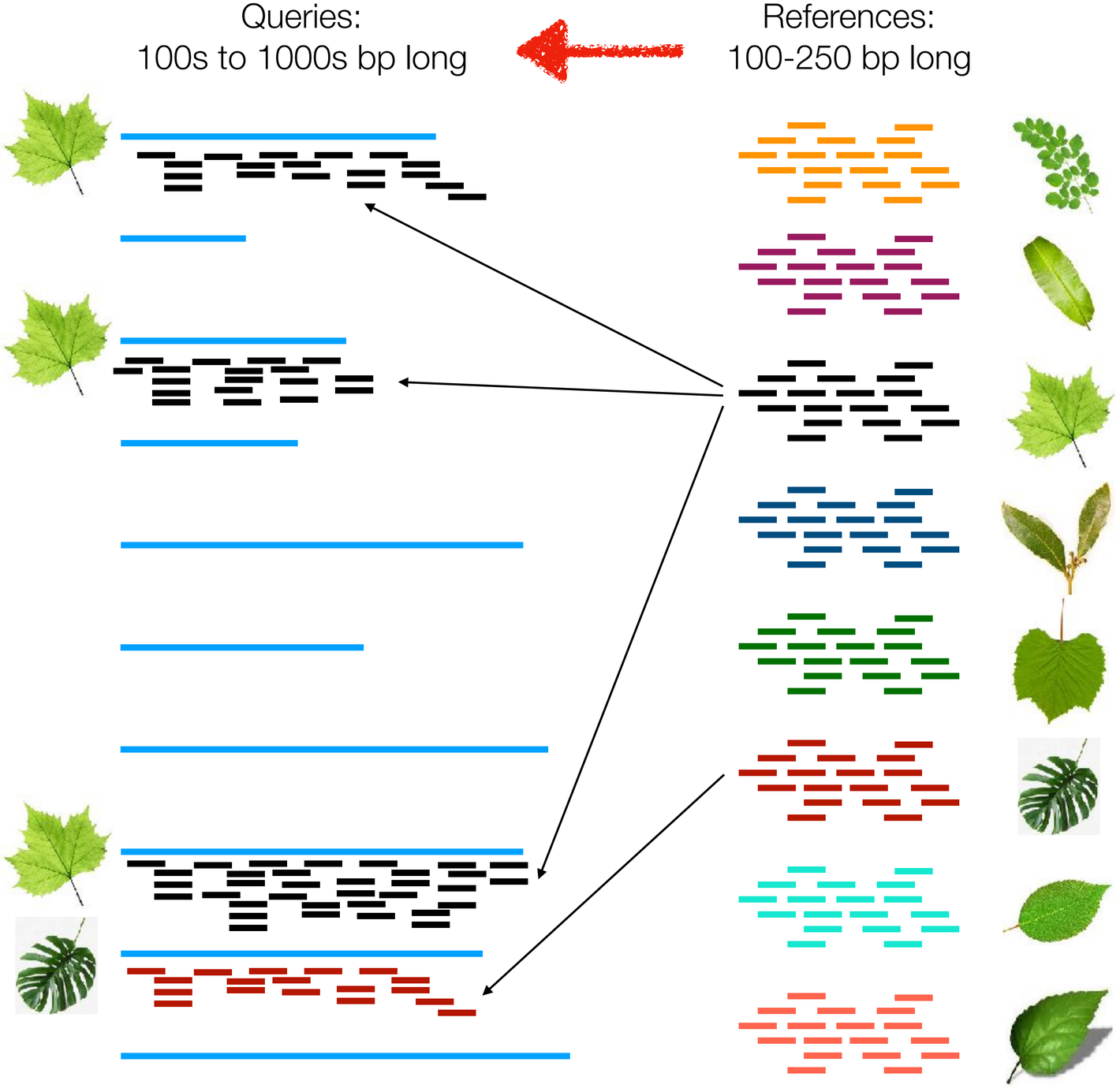


.....

We map the references to the queries



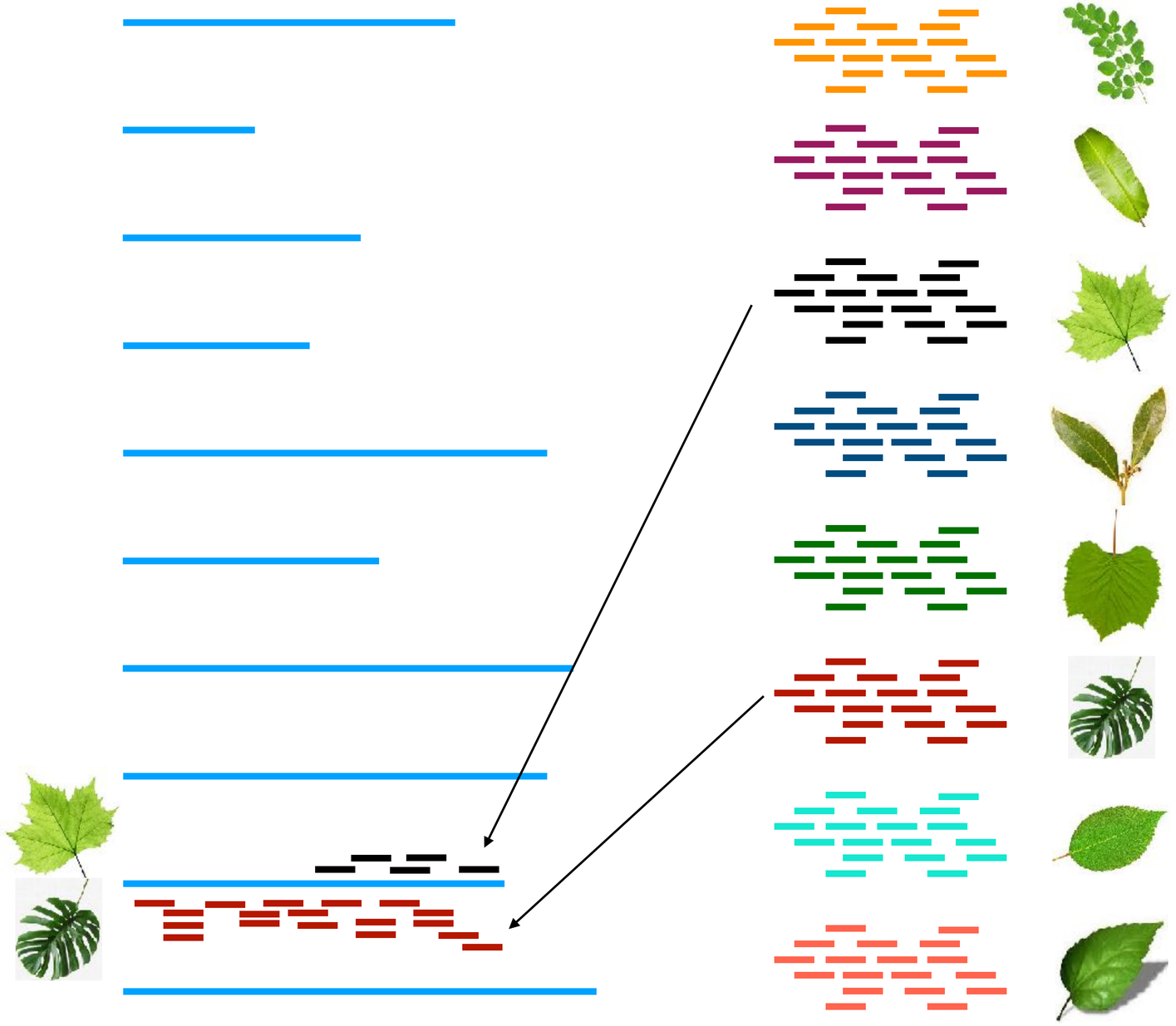
We map the references to the queries



We map the references to the queries

Queries:
100s to 1000s bp long

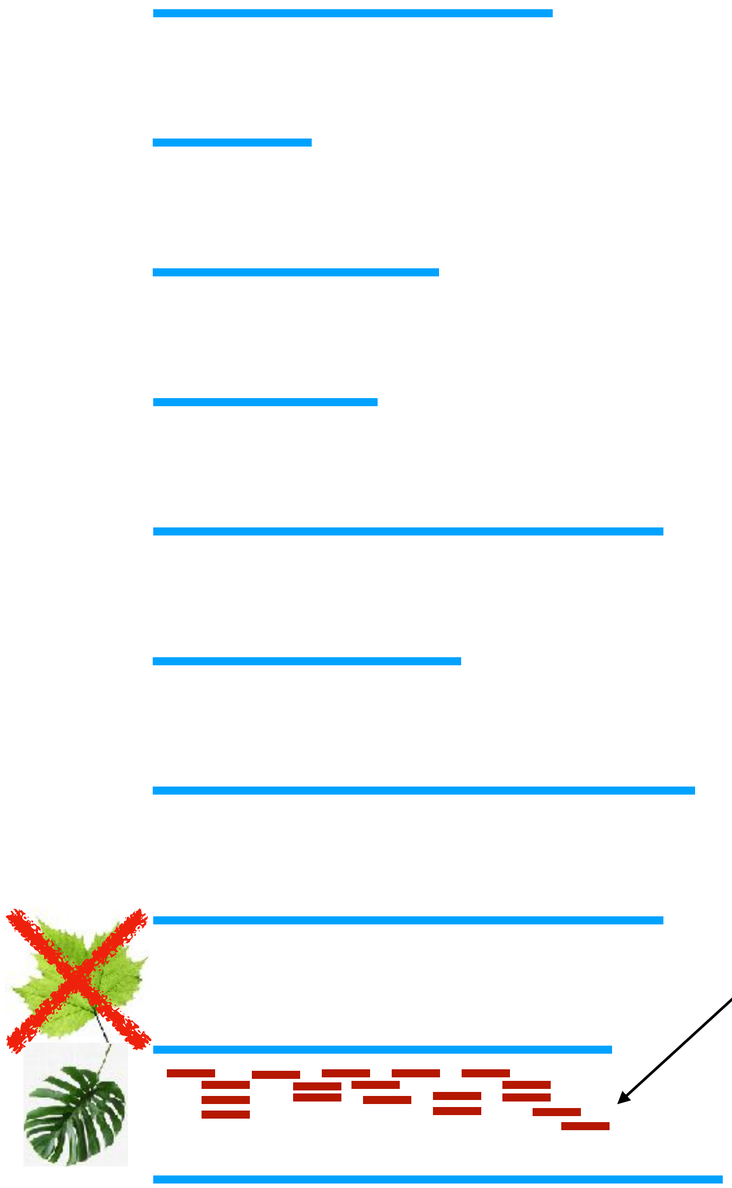
References:
100-250 bp long

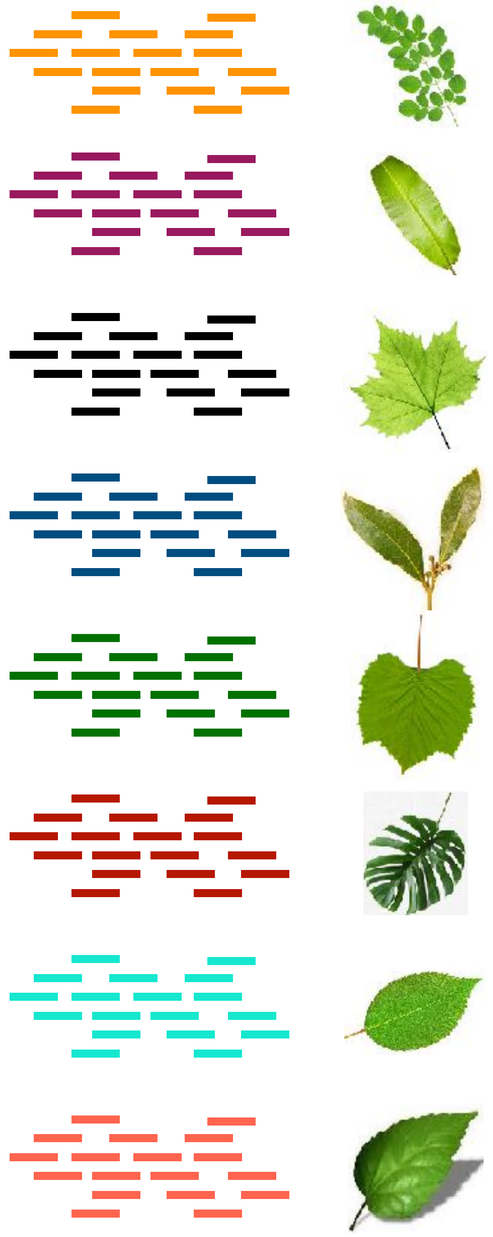


We map the references to the queries

Queries:
100s to 1000s bp long

References:
100-250 bp long



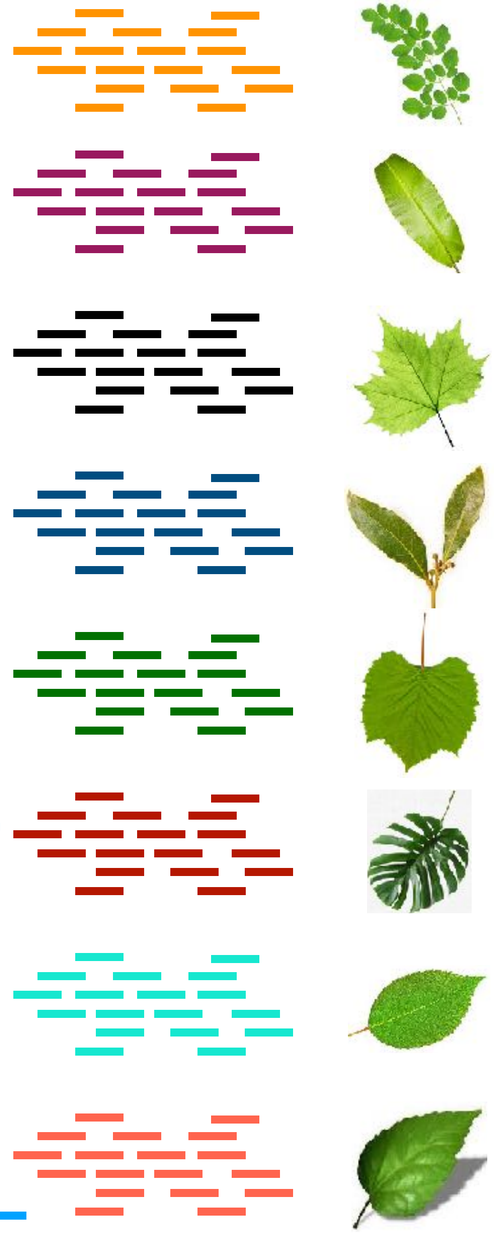


Our goals:

- Does not need specialist molecular-lab skills
- Can be used on small amounts of pollen
- Uses the **nuclear genome**, not (just) the chloroplast genome
- **Does NOT require assembled genomes** for the reference database
- Is '**semi-quantitative**' (low vs. high biomass frequency)
- Can **identify pollen to plant species**

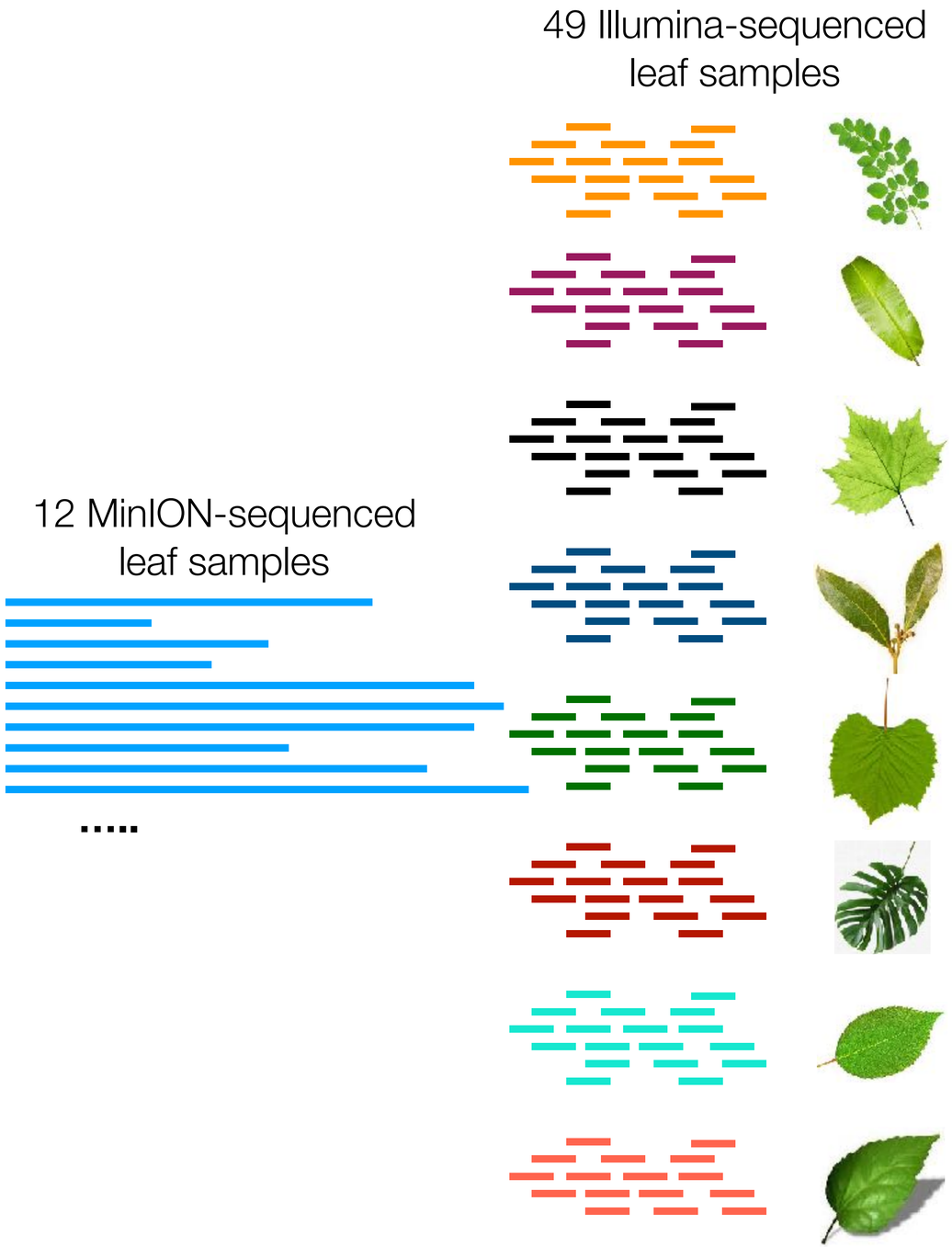
each read is from one cell

References:
100-250 bp long



Our test:

- 49 UK plant species **Illumina shotgun-sequenced**, 250 PE, 0.3-1.0X coverage
- 12 mock leaf samples of known composition, **MinION sequenced**
- **Map all Illumina skims to all MinION reads**
  - 49 X 10 = 490 combinations
  - *minimap2* → *samtools* → *R*
  - filter by depth and evenness of mapping coverage



12 Mock samples – DNA per species

Species_name_Latin	hml - 1	hml - 2	lmh - 1	lmh - 2	hml - 1	hml - 2	Hhl - 1	Hhl - 2	hhll - 1	hhll - 2	llhh - 1	llhh - 2
Knautia arvensis	100	100							100	100	1	1
Galium verum	100	100							100	100	1	1
Crepis capillaris	100	100							100	100	1	1
Papaver somniferum	10	10	1	1			1	1			100	100
Anagallis arvensis	10	10	1	1					1	1	100	100
Sambucus nigra	10	10	1	1			1000	1000	1	1	100	100
Bryonia dioica	1	1	10	10	10	10			1	1	100	100
Ranunculus repens	1	1	10	10	100	100			100	100		
Lotus corniculatus	1	1	10	10			100	100				
Digitalis purpurea	0	0	100	100	10	10						
Leucanthemum	0	0	100	100	100	100						
Stachys sylvatica	0	0	100	100	1	1	100	100	1	1	1	1

+ 37 more species in reference

Expectation is that ONLY Illumina reads from the included plant species should successfully map onto MinION reads in that sample

12 Mock samples – DNA per species

Species_name_Latin	hml - 1	hml - 2	lmh - 1	lmh - 2	hml - 1	hml - 2	Hhl - 1	Hhl - 2	hhll - 1	hhll - 2	llhh - 1	llhh - 2
Knautia arvensis	**100**	**100**							100	100	1	1
Galium verum	**100**	**100**							100	100	1	1
Crepis capillaris	**100**	**100**							100	100	1	1
Papaver somniferum	**10**	**10**	1	1			1	1			100	100
Anagallis arvensis	**10**	**10**	1	1					1	1	100	100
Sambucus nigra	**10**	**10**	1	1			1000	1000	1	1	100	100
Bryonia dioica	**1**	**1**	10	10	10	10			1	1	100	100
Ranunculus repens	**1**	**1**	10	10	100	100			100	100		
Lotus corniculatus	**1**	**1**	10	10			100	100				
Digitalis purpurea	**0**	**0**	100	100	10	10						
Leucanthemum	**0**	**0**	100	100	100	100						
Stachys sylvatica	**0**	**0**	100	100	1	1	100	100	1	1	1	1

Species_name_Latin	hml - 1	hml - 2	lmh - 1	lmh - 2	hml - 1	hml - 2	Hhl - 1	Hhl - 2	hhll - 1	hhll - 2	llhh - 1	llhh - 2
Knautia arvensis	**286**	**267**	0	0	1	0	0	0	139	223	4	2
Galium verum	**127**	**79**	1	1	0	0	1	0	28	40	1	1
Crepis capillaris	**342**	**331**	0	0	1	0	2	0	179	231	1	5
Papaver somniferum	**13**	**10**	1	0	0	0	1	0	0	0	127	88
Anagallis arvensis	**20**	**24**	3	1	0	0	2	0	1	2	370	286
Sambucus nigra	**25**	**25**	1	6	1	6	1225	783	2	3	423	298
Bryonia dioica	**1**	**0**	10	18	54	33	0	0	0	0	205	152
Ranunculus repens	**3**	**3**	21	31	543	516	1	1	180	239	2	2
Lotus corniculatus	**0**	**0**	3	12	0	0	44	30	0	0	0	0
Digitalis purpurea	**0**	**0**	42	70	19	9	0	0	0	0	0	0
Leucanthemum	**1**	**0**	78	143	229	232	0	0	0	1	0	0
Stachys sylvatica	**0**	**1**	299	494	7	6	209	144	5	6	5	3

12 Mock samples – DNA per species

Species_name_Latin	hml - 1	hml - 2	lmh - 1	lmh - 2	hml - 1	hml - 2	HHl - 1	HHl - 2	hhll - 1	hhll - 2	llhh - 1	llhh - 2
Knautia arvensis	**100**	**100**							100	100	1	1
Galium verum	**100**	**100**							100	100	1	1
Crepis capillaris	**100**	**100**							100	100	1	1
Papaver somniferum	**10**	**10**	1	1			1	1			100	100
Anagallis arvensis	**10**	**10**	1	1					1	1	100	100
Sambucus nigra	**10**	**10**	1	1			1000	1000	1	1	100	100
Bryonia dioica	**1**	**1**	10	10	10	10			1	1	100	100
Ranunculus repens	**1**	**1**	10	10	100	100			100	100		
Lotus corniculatus	**1**	**1**	10	10			100	100				
Digitalis purpurea	**0**	**0**	100	100	10	10						
Leucanthemum	**0**	**0**										
Stachys sylvatica	**0**	**0**										

We can see which species are **present** and which are **absent**, and we can see which plant species have **higher biomass**

Species_name_Latin	hml - 1	hml - 2	lmh									
Knautia arvensis	**286**	**267**										
Galium verum	**127**	**79**	1	1	0	0	1	0	28	40	1	1
Crepis capillaris	**342**	**331**	0	0	1	0	2	0	179	231	1	5
Papaver somniferum	**13**	**10**	1	0	0	0	1	0	0	0	127	88
Anagallis arvensis	**20**	**24**	3	1	0	0	2	0	1	2	370	286
Sambucus nigra	**25**	**25**	1	6	1	6	1225	783	2	3	423	298
Bryonia dioica	**1**	**0**	10	18	54	33	0	0	0	0	205	152
Ranunculus repens	**3**	**3**	21	31	543	516	1	1	180	239	2	2
Lotus corniculatus	**0**	**0**	3	12	0	0	44	30	0	0	0	0
Digitalis purpurea	**0**	**0**	42	70	19	9	0	0	0	0	0	0
Leucanthemum	**1**	**0**	78	143	229	232	0	0	0	1	0	0
Stachys sylvatica	**0**	**1**	299	494	7	6	209	144	5	6	5	3

Our goals – a protocol that:

- Does not need specialist molecular-lab skills
- Can be used on small amounts of pollen
- Uses the **nuclear genome**, not (just) the chloroplast genome
- **Does NOT require assembled genomes** for the reference database
- Is '**semi-quantitative**' (low vs. high biomass frequency)
- Can **identify pollen to plant species**

- MinION library prep still complex but getting easier (and NO PCR!)
- ✔
- ✔
- ✔
- ~ Good at identifying dominant constituents, poorer at identifying rare components
- ~ Not great at differentiating closely related congeners (e.g. *Ranunculus acris* and *R. repens*, which also hard to differentiate using barcodes and morphology)

Summary

Single-species quantitative PCR (**qPCR**)

- **use eDNA rate** not concentration! ng/sec not ng/µl

Multiplexed individual barcoding (**mBRAVE**)

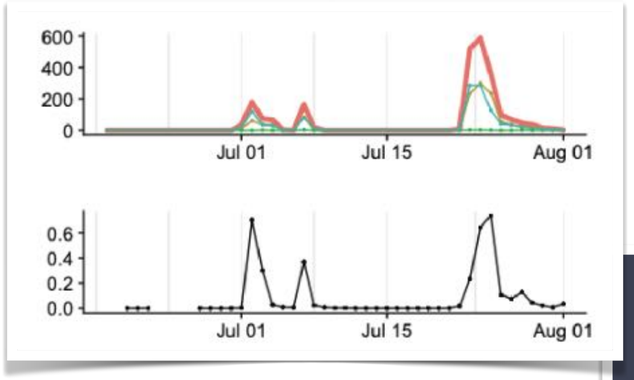
- **count data feasible for many 1000s of individuals**

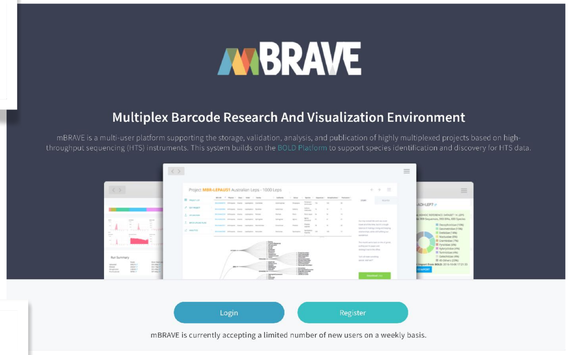
Mitogenomics/Metabarcoding + DNA spike-in (**SPIKEPIPE, qSeq**)

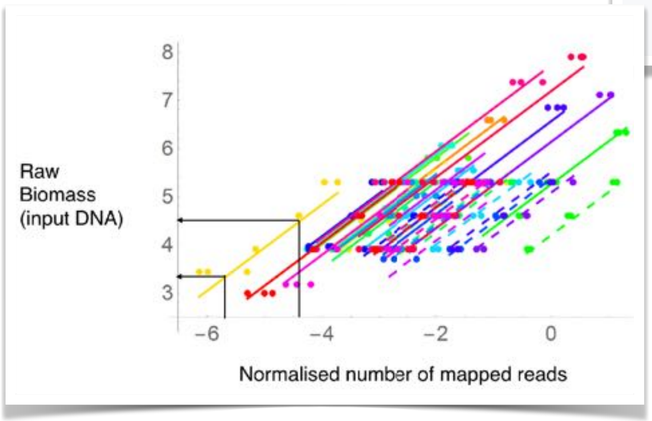
- **For within-species differences across samples** (e.g. time series)
- (difficult to get across-species diffs)

Reverse metagenomics (**RevMet**)

- **For within- and across-species differences within a sample** (e.g. diet analysis)







Species_name_Latin	hmi - 1	hmi - 2
Knautia arvensis	100	100
Galium verum	100	100
Crepis capillaris	100	100
Papaver somniferum	10	10
Anagallis arvensis	10	10
Sambucus nigra	10	10
Bryonia dioica	1	1
Ranunculus repens	1	1
Lotus corniculatus	1	1
Digitalis purpurea	0	0
Leucanthemum	0	0
Stachys sylvatica	0	0

Species_name_Latin	hmi - 1	hmi - 2
Knautia arvensis	286	267
Galium verum	127	79
Crepis capillaris	342	331
Papaver somniferum	13	10
Anagallis arvensis	20	24
Sambucus nigra	25	25
Bryonia dioica	1	0
Ranunculus repens	3	3
Lotus corniculatus	0	0
Digitalis purpurea	0	0
Leucanthemum	1	0
Stachys sylvatica	0	1