Determining Best-Case Effectiveness of a Molecular Method for Detection of Aquatic Invasive Species

Ryan Scott[1], Robin Gras[1,2], Emily A. Brown[3], Melania E. Cristescu[3], Aibin Zhan[4], and Hugh J. MacIsaac[2]

[1]University of Windsor, School of Computer Science

[2]Great Lakes Institute for Environmental Research (GLIER)

[3]McGill University

[4]Research Center for Eco-Environmental Sciences, Beijing

Sequencing and Clustering in Biological Invasions

- Move toward high-throughput next-gen sequencing methods for early detection of AIS
  - Make use of eDNA (environmental DNA)
  - Better detection of rare or cryptic taxa
- High-throughput next-gen sequencing
  - Fast
  - Relatively inexpensive
  - Extremely sensitive or exposing artefacts?

Sequencing and Clustering in Biological Invasions

- Move toward high-throughput next-gen sequencing methods for early detection of AIS
  - Make use of eDNA (environmental DNA)
  - Better detection of rare or cryptic taxa
- High-throughput next-gen sequencing
  - Fast
  - Relatively inexpensive
  - Extremely sensitive or exposing artefacts?

Sequencing and Clustering in Biological Invasions

Typical Pathway:

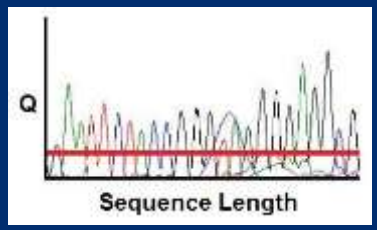
Obtain eDNA Sample

Extract DNA

PCR

Sequencing

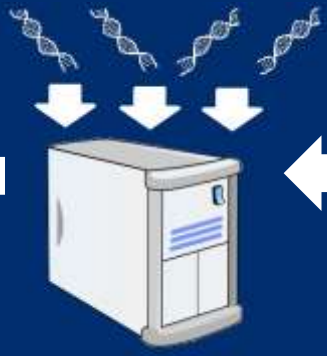
Data Preprocessing



Clustering (define OTUs)

(Optional)

BLAST



Identify Invaders

Sequencing and Clustering in Biological Invasions

Typical Pathway:

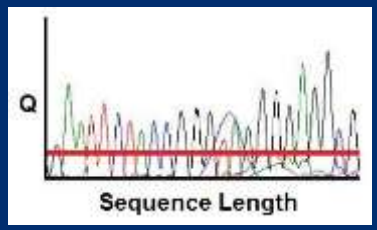
Obtain eDNA Sample

Extract DNA

PCR

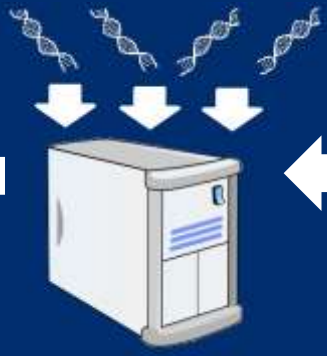
Sequencing

Data Preprocessing



Clustering (define OTUs)

BLAST



Identify Invaders

(Optional)

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU
A	…ATGCA…	→	1
A	…ATAAA…		

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU	
A	…ATGCA…	→	1	
A	…ATAAA…	→	2	✖

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU	
A	…ATGCA…	→	1	
A	…ATAAA…	→	2	✖

Type II Error – false negative

Species	Sequence		OTU
A	…ATGCA…	→	1
B	…TGGCC…		

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU	
A	…ATGCA…	→	1	
A	…ATAAA…	→	2	✖

Type II Error – false negative

Species	Sequence		OTU	
A	…ATGCA…	→	1	
B	…			✖

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU	
A	…ATGCA…	→	1	
A	…ATAAA…	→	2	✖

Type II Error – false negative

Species	Sequence		OTU
A	…ATGCA…	→	1
B	…TGGCC…		

Parameter Selection and Error Manifestations

- Parameter selection in processing affects results

Type I Error – false positive

Species	Sequence		OTU	
A	…ATGCA…	→	1	
A	…ATAAA…	→	2	✖

Type II Error – false negative

Species	Sequence		OTU	
A	…ATGCA…	→	1	
B	…TGGCC…	→	1	✖

Parameter Selection and Error Manifestations

"Parameter Set" Definition

Parameter	Synopsis	Possible Values
Sequence Length	Trim length of all sequences	150, 175, 200, 225, 250, 300, 325, 350, 400
Minimum Phred Score (Q)	Minimum quality score per base call	None, 10, 15, 20, 25, 30
Maximum Expected Error	Sequence-wide expected error score	None, 0.1, 0.25, 0.5, 0.75, 1
Similarity Threshold	Intraspecific genetic similarity threshold	93% - 99%, 0.5% increments
Singletons	Do we keep sequences that are unique	Yes, No

Project Goals

- Determine how parameters interact
- Aid in determining usable parameters with or without clustering with application to early detection of invasive species
  - Focus on minimizing type II errors
  - Try to minimize type I errors as well if possible
- Determine effectiveness of this pathway in detection of an invader

Methods:

Part 1

- 115,902 labeled sequences across 19 taxa of varied relatedness
- Use these 18S sequences to optimize pipeline
  - USEARCH + BLASTn
  - Search the space of parameter sets for those that minimize error, particularly type II error (falsely excluding a species or grouping it with another)
  - Cached BLAST results for fast computing

Methods:

Part 2

- Use optimal parameter sets, try to detect a “target” species
  - Use the same sequences from 19 taxa as targets
  - Spike 2-50 sequences of target into samples from 12 real communities
  - Record when the target is observed
  - Iterate through all potential targets and communities
- Do the same without clustering to determine which option is better in this context

Methods:

Target Taxon	Sequences	Ratio of Sequences Kept At Length 400							
		Q = 5	Q = 10	Q = 15	Q = 20	MEE = 1	MEE = 0.5	MEE = 0.25	MEE = 0.1
Artemia	2141	0.9360	0.9360	0.3650	0.0350	0.7170	0.4517	0.1738	0.0126
Balanus	14732	0.9890	0.9890	0.4620	0.1140	0.7492	0.4716	0.1821	0.0069
Brachionus	207	0.9950	0.9950	0.0140	0.0000	0.0435	0.0193	0.0000	0.0000
Cancer	1563	0.9450	0.9450	0.2530	0.0520	0.6379	0.3397	0.1260	0.0038
Cercopagis	1222	0.9890	0.9890	0.2010	0.0050	0.6457	0.2700	0.0270	0.0000
Corbicula	46911	0.9640	0.9640	0.4870	0.1930	0.8319	0.6328	0.3523	0.0729
Ciona	3994	0.9400	0.9400	0.2460	0.0530	0.5626	0.2694	0.0874	0.0043
Daphnia	706	0.9550	0.9550	0.0990	0.0080	0.3782	0.1204	0.0312	0.0000
Diacyclops	812	0.9890	0.9890	0.1320	0.0070	0.6860	0.2635	0.0222	0.0000
Echinogammarus	7335	0.9690	0.9690	0.4470	0.1040	0.7286	0.4539	0.1840	0.0100
Epischura	9876	0.9890	0.9890	0.3420	0.0880	0.8019	0.5018	0.2132	0.0060
Leptodora	5436	0.9760	0.9760	0.3240	0.0620	0.6382	0.3891	0.1818	0.0166
Mesocyclops	1010	0.9910	0.9910	0.0150	0.0000	0.1772	0.0099	0.0000	0.0000
Microstella	810	0.9740	0.9740	0.1890	0.0410	0.6988	0.3395	0.0889	0.0000
Oikopleura	3543	0.8910	0.8910	0.0810	0.0040	0.4417	0.1087	0.0164	0.0003
Palaemonetes	5168	0.9770	0.9770	0.5850	0.2500	0.8611	0.6683	0.4435	0.0851
Pleuroxus	639	0.9800	0.9800	0.0690	0.0060	0.5227	0.1596	0.0172	0.0000
Senecella	345	0.9830	0.9830	0.0900	0.0060	0.4087	0.1101	0.0087	0.0000
Themisto	4266	0.9770	0.9770	0.7450	0.4330	0.9027	0.7935	0.6139	0.2956
Average	110716	0.9689	0.9689	0.2708	0.0769	0.6018	0.3354	0.1458	0.0271

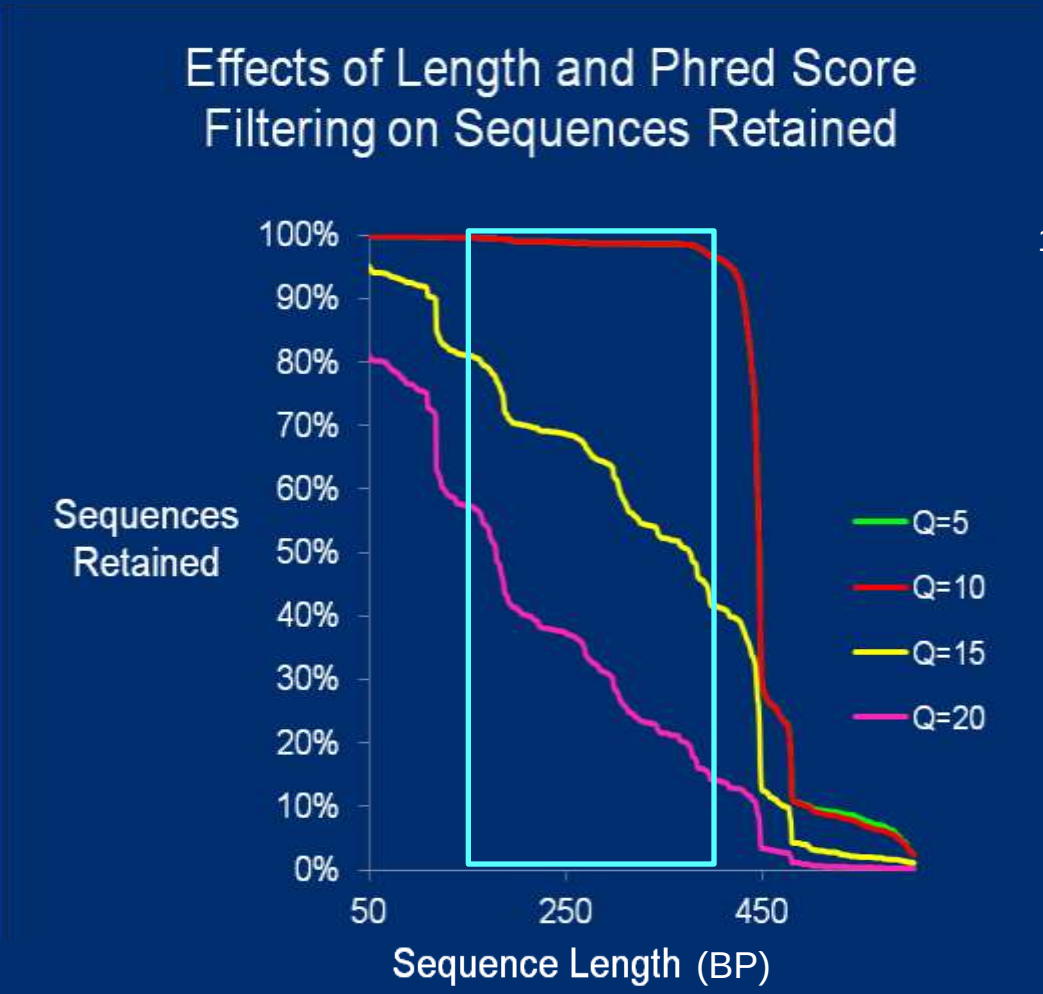
Methods:

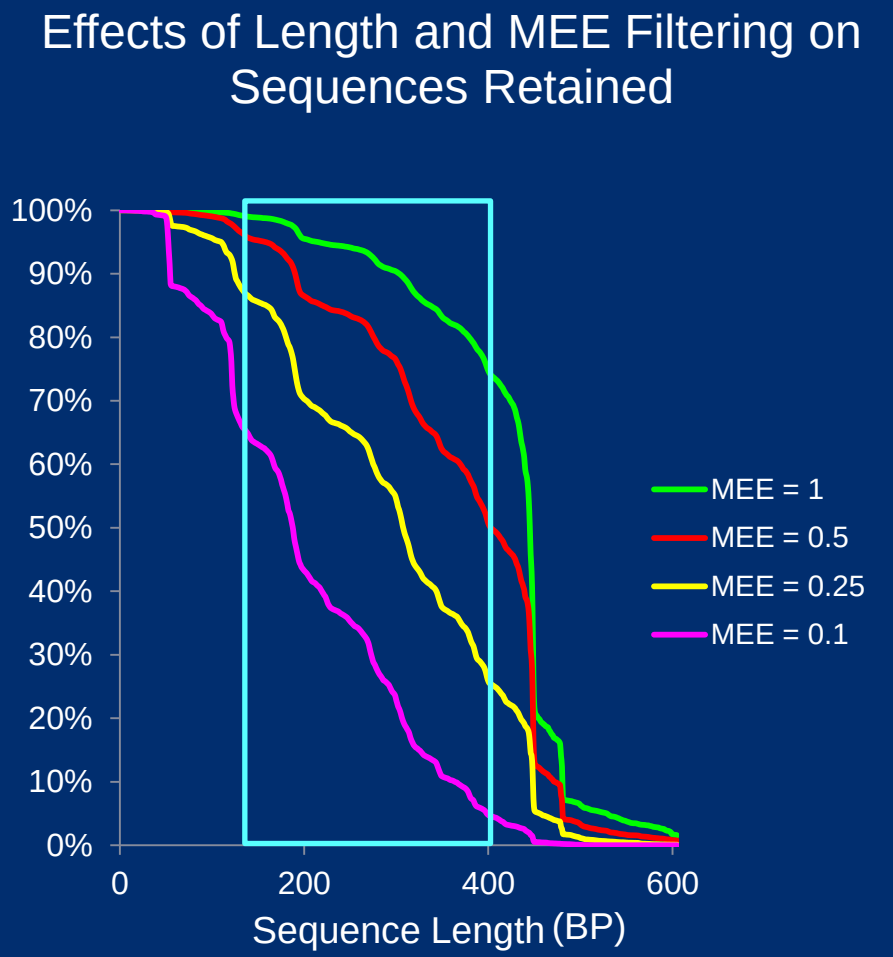
Location	Sequences	Ratio of Sequences Kept At Length 400							
		Q = 5	Q = 10	Q = 15	Q = 20	MEE = 1	MEE = 0.5	MEE = 0.25	MEE = 0.1
Artificial - Amplicon	95237	0.9470	0.9220	0.3390	0.0960	0.7467	0.5151	0.2828	0.0230
Artificial - Shotgun	237146	0.0000	0.0000	0.0000	0.0000	0.4810	0.0680	0.0000	0.0000
Churchill	684163	0.2420	0.2120	0.0210	0.0000	0.0632	0.0235	0.0042	0.0000
Halifax	877078	0.2670	0.1690	0.0020	0.0000	0.0191	0.0009	0.0000	0.0000
Hamilton	686064	0.2680	0.2500	0.0590	0.0090	0.1506	0.0828	0.0327	0.0023
Hawkesbury	444315	0.6360	0.6200	0.2820	0.1070	0.4881	0.3727	0.2340	0.0377
Nanaimo1	406215	0.6160	0.5980	0.1360	0.0170	0.3775	0.2008	0.0714	0.0044
Nanaimo2	383190	0.4510	0.4170	0.1090	0.0290	0.2546	0.1514	0.0742	0.0095
Nanticoke	480962	0.5910	0.5650	0.1770	0.0510	0.4033	0.2573	0.1312	0.0254
Sept Iles - Amplicon	249663	0.9740	0.9540	0.4590	0.1760	0.8175	0.6306	0.4075	0.0319
Sept Iles - Shotgun	502688	0.5970	0.5610	0.2130	0.0710	0.3995	0.2812	0.1630	0.0080
Thunder Bay	556984	0.6950	0.6740	0.3100	0.1080	0.5550	0.4273	0.2582	0.0513
Vancouver	1008358	0.2740	0.2480	0.0400	0.0020	0.1171	0.0469	0.0107	0.0001
Victoria	456391	0.5620	0.5460	0.1470	0.0280	0.3641	0.2149	0.0906	0.0095
Contrived	115902	0.9670	0.9670	0.4150	0.1410	0.7471	0.5067	0.2584	0.0475

Methods:


Ratio of Inconsistent BLAST Results
Ratio Inconsistent
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
150
175
200
225
250
275
300
325
350
375
400
Sequence Length
Artemia
Balanus
Brachionus
Cancer
Cercopagis
Ciona
Corbicula
Daphnia
Diacyclops
Echinogammarus
Epischura
Leptodora
Mesocyclops
Microstella
Oikopleura
Palaemonetes
Pleuroxus
Senecella
Themisto

Effects of Parameter Choice (Part 1)





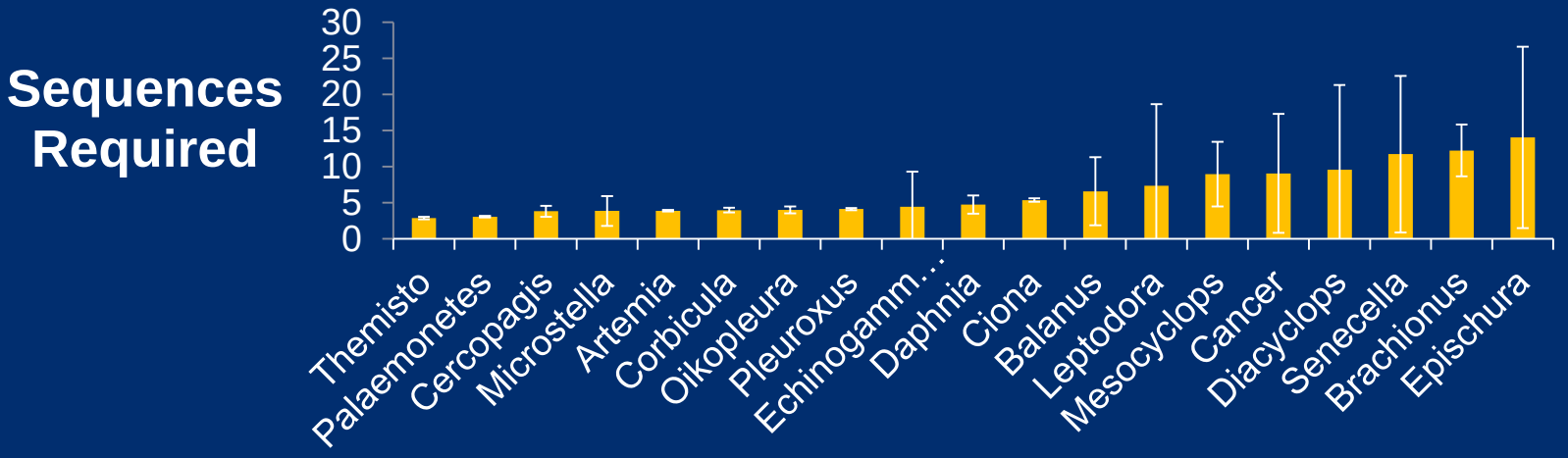
Effects of Parameter Choice (Part 1)

- Type II error reduced by using longer sequences, little or no filtering, and lower similarity thresholds
  - Longer sequences with less stringent filtering results in more sequences kept
- Lower similarity threshold?
  - Suggested to use 97%, but here we see 93-96% typically yields best results
  - Relatively high dissimilarity among taxa in the sample used for optimization
    - Redundancy was a factor that hurt the score for a given parameter set

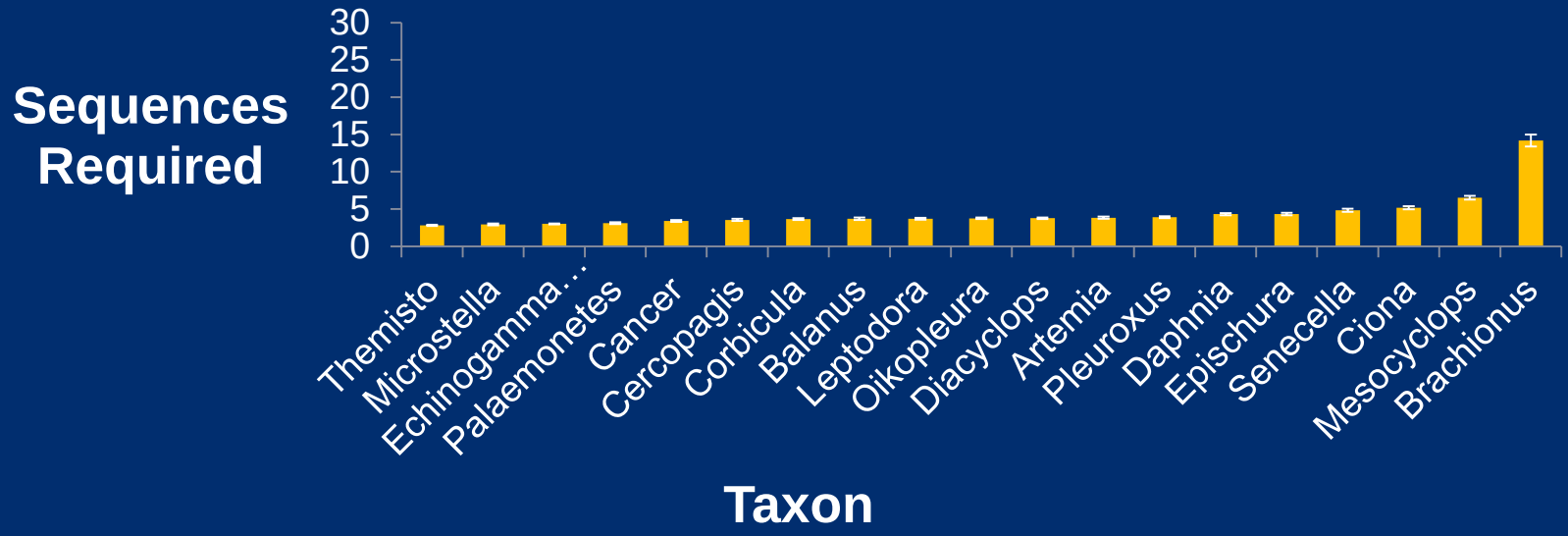
Sequences Required Per Taxon

(Part 2)

With Clustering



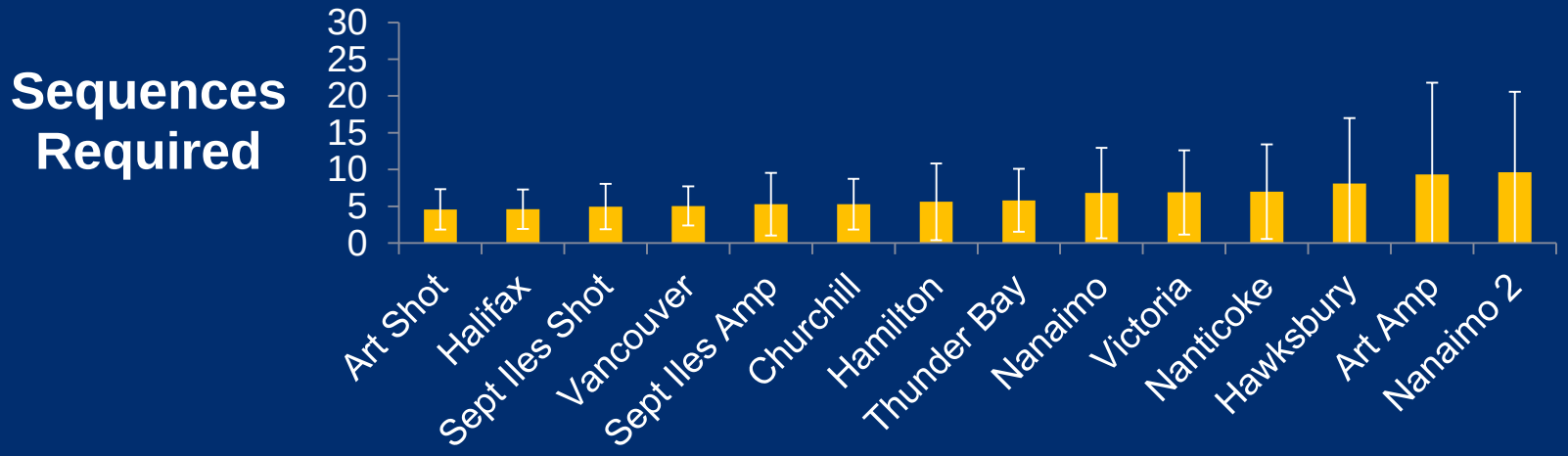
Without Clustering



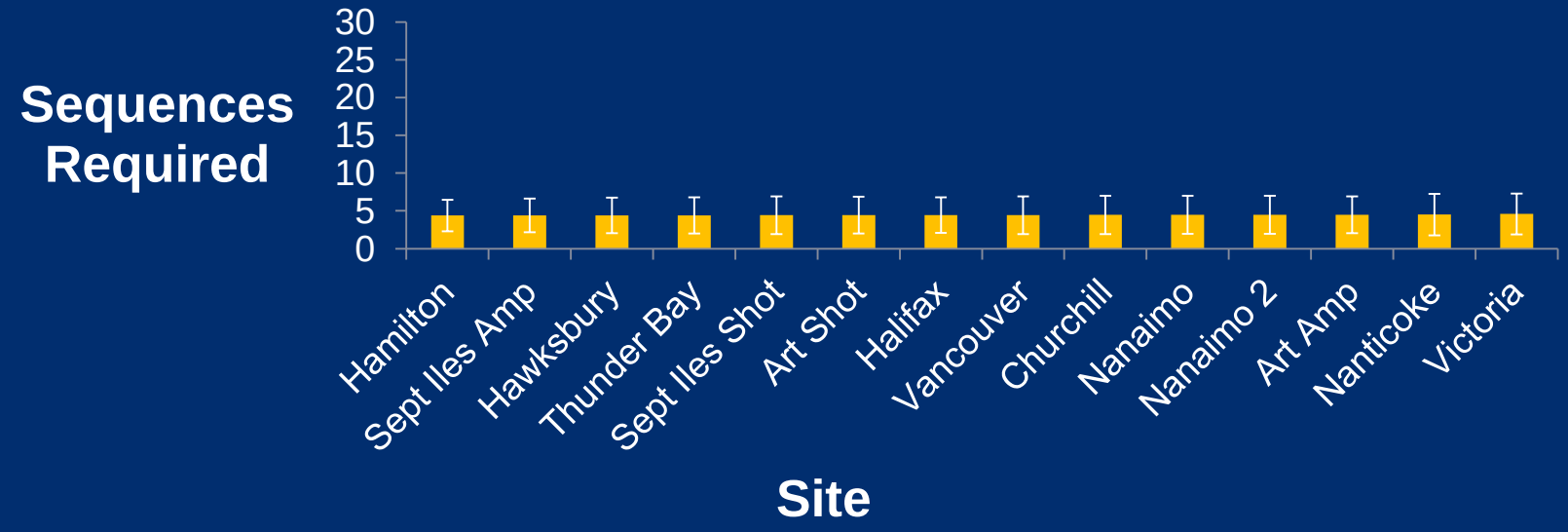
Sequences Required Per Site

(Part 2)

With Clustering



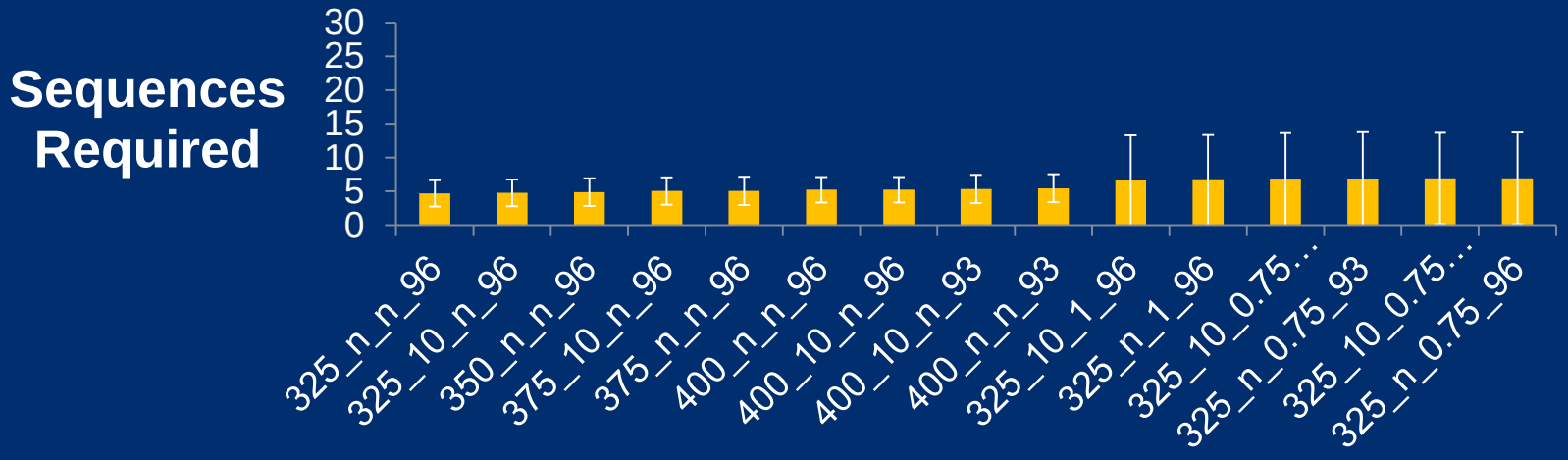
Without Clustering



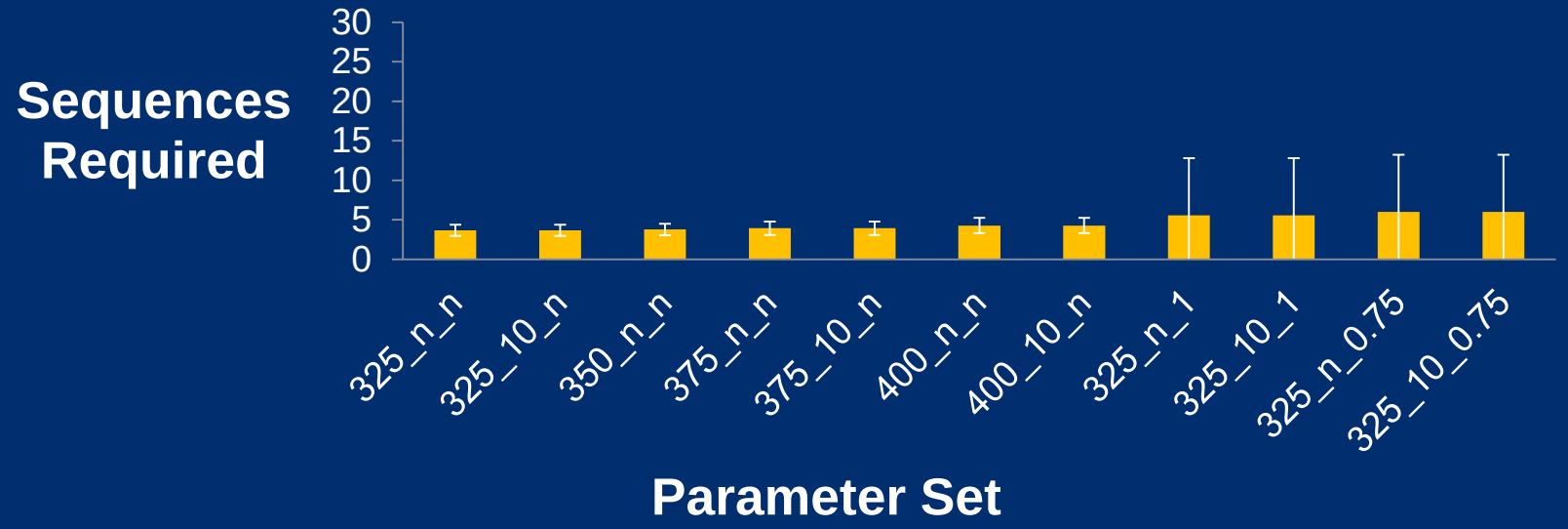
Sequences Required Per Parameters

(Part 2)

With Clustering

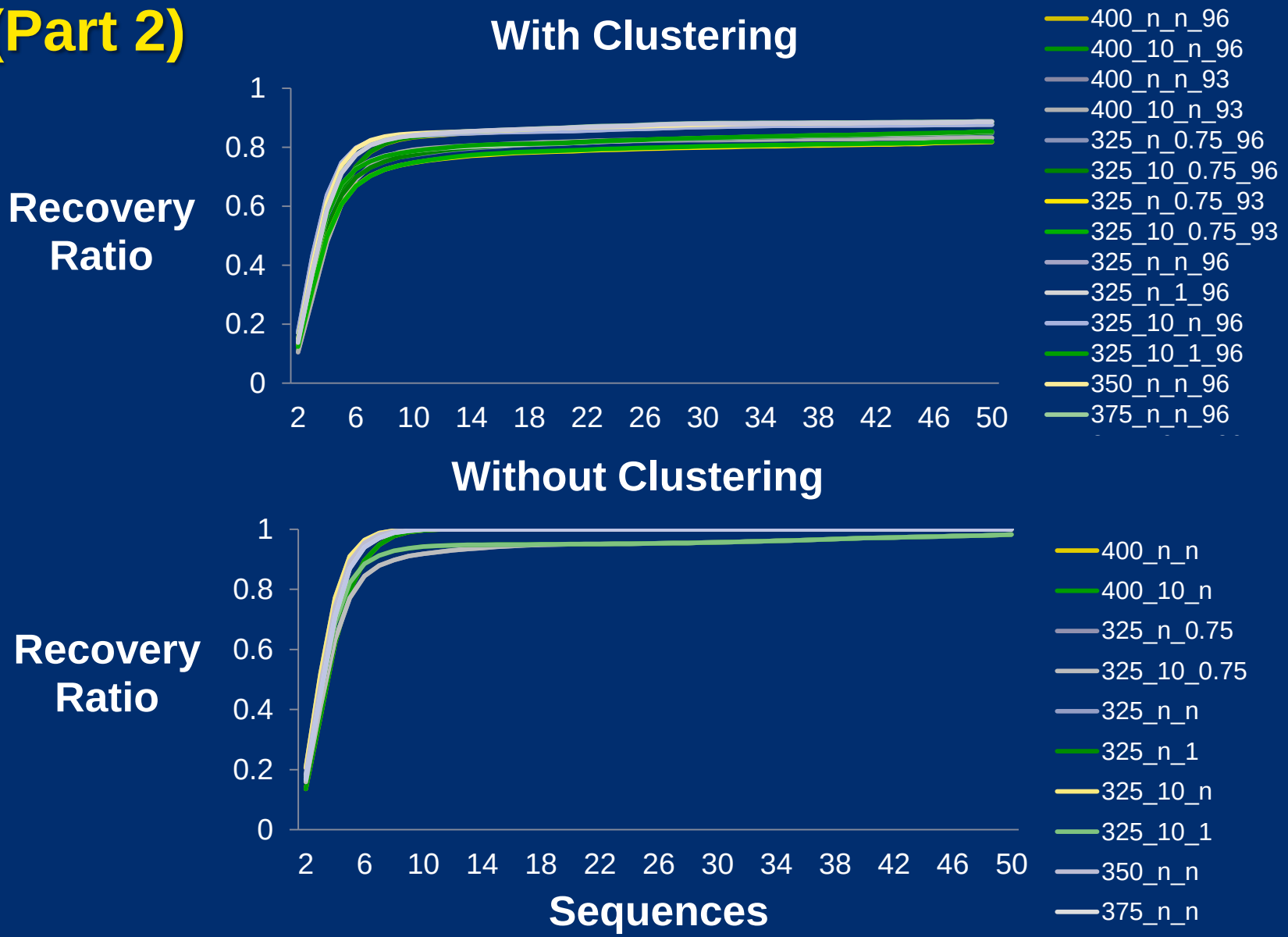


Without Clustering



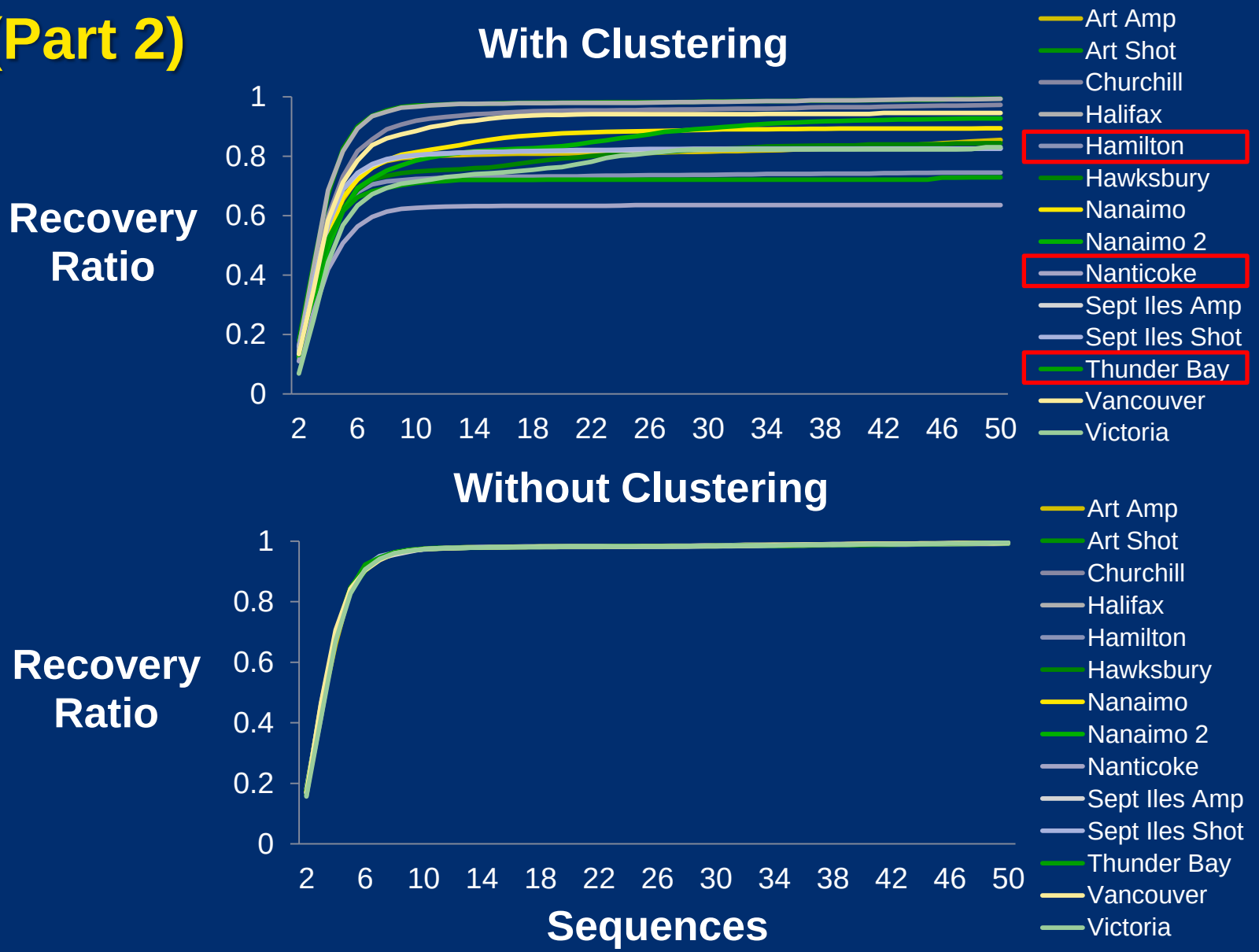
Recovery Ratio Per Parameters

(Part 2)



Ratio of Taxa Recovered

(Part 2)



Conclusions:

- Parameter selection should reflect your goal
  - Biodiversity estimate? More stringent filtering, longer sequences, no singletons, clustering
  - Detecting invaders? Little or no filtering, longer sequences, no singletons, no clustering
- Detectability varies heavily across taxa, site, and parameter selection
  - Brachionus (low quality)
  - Hamilton, Nanticoke, Thunder Bay

Conclusions:

- With clustering: 6.1 sequences required, 87% detection ratio
- Without clustering: 4.5 sequences required, 99.3% detection ratio
- Taking sequences of a known potential threat and an eDNA sample for a particular site, we can determine expected detectability computationally (and the best parameters for it!)
- Molecular detection is NOT perfect

Thank you!


SHARCNET™


Canada Research Chairs
www.chairs-chaires.gc.ca


CAISN
CANADIAN
AQUATIC
INVASIVE
SPECIES
NETWORK


NSERC
CRSNG