Optimization and Performance Testing of a Sequence Processing Pipeline Applied to Early Detection of Nonindigenous Species

Ryan Scott[1], Robin Gras[1,2], Emily A. Brown[3], Frédéric J.J. Chain[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

- Moving toward high-throughput (aka next-gen) sequencing methods for early detection of AIS
  - Using metabarcoding
    - Bulk sampling or eDNA sampling
  - Better detection of rare or cryptic taxa
- High-throughput sequencing
  - Fast
  - Relatively inexpensive
  - Extremely sensitive or exposing artefacts?

Sequencing and Clustering in Biological Invasions

Typical Pipeline:

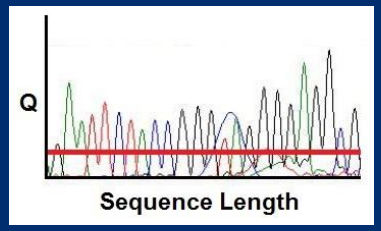
Obtain Sample (bulk or eDNA)

Extract DNA

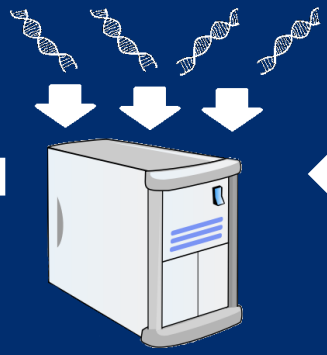
PCR

Sequencing

Data Preprocessing



BLAST



Identify Invaders

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	✖

Project Goals

- Determine how parameters interact (previous studies only vary 1-2 parameters, low resolution)
  - We tested 1050 parameter combinations total
- Determine usable parameters for two different research objectives:
  - Estimation of species richness by metabarcoding of metazoan bulk samples
  - Early detection of invasive species by metabarcoding of metazoan bulk samples
- Determine effectiveness of this pipeline in detection of AIS computationally inoculated into real zooplankton community bulk samples

Methods: Dataset D1

Taxon	Sequences	Q = 10	Q = 20	MEE = 1
Artemia	2145	0.9920	0.0490	0.8015
Balanus	14732	0.9910	0.1310	0.8128
Brachionus	207	0.9950	0.0000	0.0483
Cancer	1629	0.9940	0.1040	0.7185
Carcinus	200	1.0000	0.1750	0.9400
Cercopagis	1222	0.9920	0.0110	0.7709
Corbicula	46915	0.9900	0.2980	0.8952
Daphnia	706	0.9750	0.0160	0.6232
Diacyclops	812	0.9900	0.0090	0.7106
Dreissena	200	1.0000	0.1550	0.9450
Echinogammarus	7337	0.9820	0.2430	0.8327
Epischura	10002	0.9900	0.1400	0.8465
Leptodiaptomus	5461	0.9890	0.0790	0.7539
Mesocyclops	1055	0.9910	0.0000	0.2812
Microsetella	814	0.9950	0.0530	0.8136
Oikopleura	3545	0.9940	0.1090	0.8434
Palaemonetes	5170	0.9930	0.3630	0.9154
Pleuroxus	644	0.9800	0.0080	0.6182
Senecella	348	0.9970	0.0140	0.4580
Themisto	4269	0.9830	0.5000	0.9311

Methods: Dataset D2

Location	Sequences	Q = 10	Q = 20	MEE = 1
Churchill	684163	0.2290	0.0000	0.0809
Halifax	877078	0.2480	0.0000	0.0477
Hamilton	686064	0.2660	0.0230	0.1750
Hawkesbury	444315	0.6370	0.1110	0.5076
Nanaimo	406215	0.6240	0.0200	0.4074
Nanticoke	480962	0.5820	0.0570	0.4305
Sept Iles	249663	0.9550	0.1900	0.8645
Thunder Bay	556984	0.6910	0.1170	0.5798
Vancouver	1008358	0.2670	0.0020	0.1359
Victoria	456391	0.5720	0.0310	0.3976

Methods:

- First part: Optimization
  - Use D1 (metazoan 18S) sequences to optimize sequence processing pipeline
    - USEARCH (using selected combinations of parameters) + BLASTn (with 97% ID threshold, keeping all hits)
    - Search the space of parameter sets for those that minimize error, prioritizing false negative error

Parameter Selection and Error Manifestations

"Parameter Set"

Parameter	Synopsis	Values Tested
Sequence Length	Length of sequences	300, 325, 350, 375, 400
Minimum Phred Score (Q)	Minimum quality score per base call	10, 20, 30
Maximum Expected Error (MEE)	Sequence-wide expected error score	1.0, 1.5, 2.0, 2.5, 3.0
Clustering Identity Threshold (Optional)	Intraspecific genetic identity threshold	97%, 98%, 99%
Denoising Minimum Abundance Threshold (Optional)	Minimum abundance of a sequence to not be considered noise	2, 4, 8
Singletons	Do we keep unique sequences	Yes, no

Methods:

- Second Part: Performance Testing
  - Using selected optimal parameter sets, tried to detect a “target” species
    - Used subsamples from the 20 taxa in D1 as “targets”
    - Spiked sequences of target into samples from D2 (1-50 sequences)
    - Iterated through all potential targets and communities

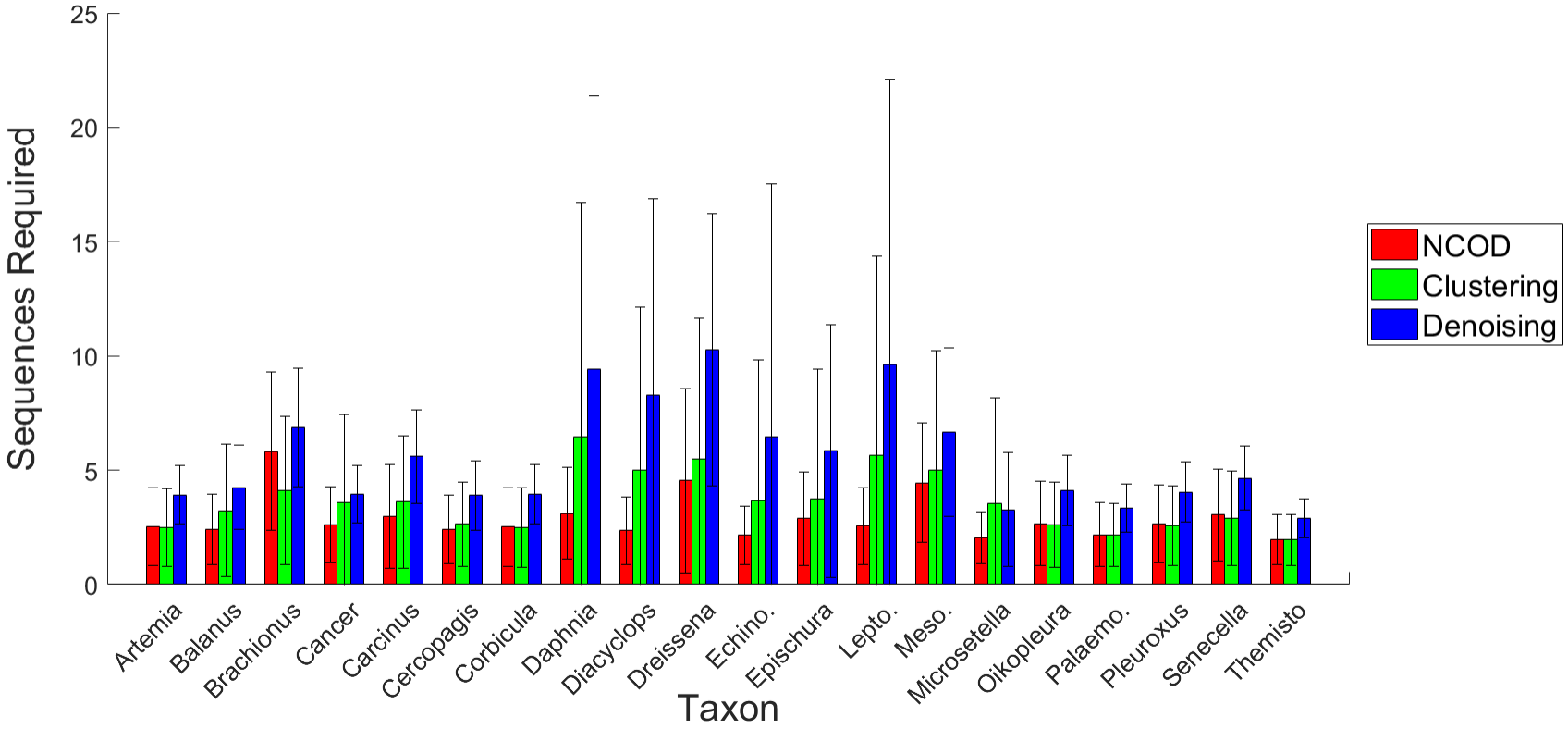
Optimal Parameter Choice

Optimized for Early Detection of AIS							
Trim Length (bp)	**Q Filter**	**MEE Filter**	**Processing Method**	**ID or Min. Abundance**	**Allow Singles?**	**Correct + Ambiguous OTUs**	**Incorrect OTUs**
400	10	3	D	2	No	20	0.2
400	10	2.5	D	2	Yes	20	0.3
400	10	2.5	NCOD	N/A	No	20	1.0
375	10	2	NCOD	N/A	No	20	2.3
400	10	1.5	NCOD	N/A	Yes	20	8.2
375	10	2.5	NCOD	N/A	Yes	20	22.6
400	10	2.5	C	99	Yes	19.8	5.8
375	10	3	D	2	Yes	19.75	0.4
375	10	2.5	D	2	No	19.73	0.4
375	10	3	C	99	Yes	19.6	9.1
400	10	3	C	99	No	19.3	0.0
375	10	2.5	C	99	No	19.0	0.3

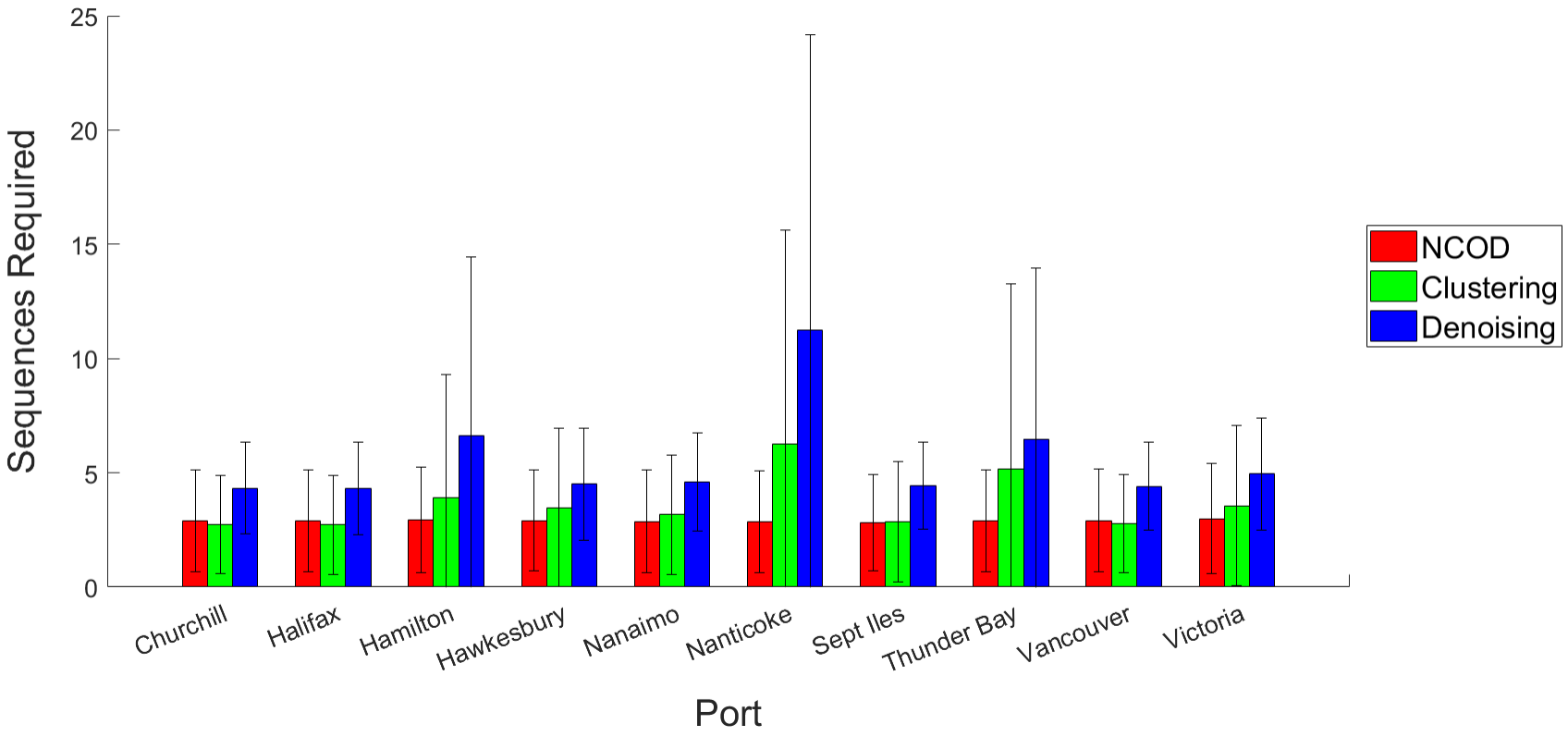
Optimal Parameter Choice

- Amplified fragment was 400-600bp in our taxa
- Longer sequences with weaker filtering results in more sequences kept and high taxonomic resolution
- No clustering or denoising = more false positives
- Clustering or denoising = more false negatives
- Best parameter sets for species richness estimates were different
  - Length ≥ 350bp, MEE ≥ 1, denoising minimum abundance = 8

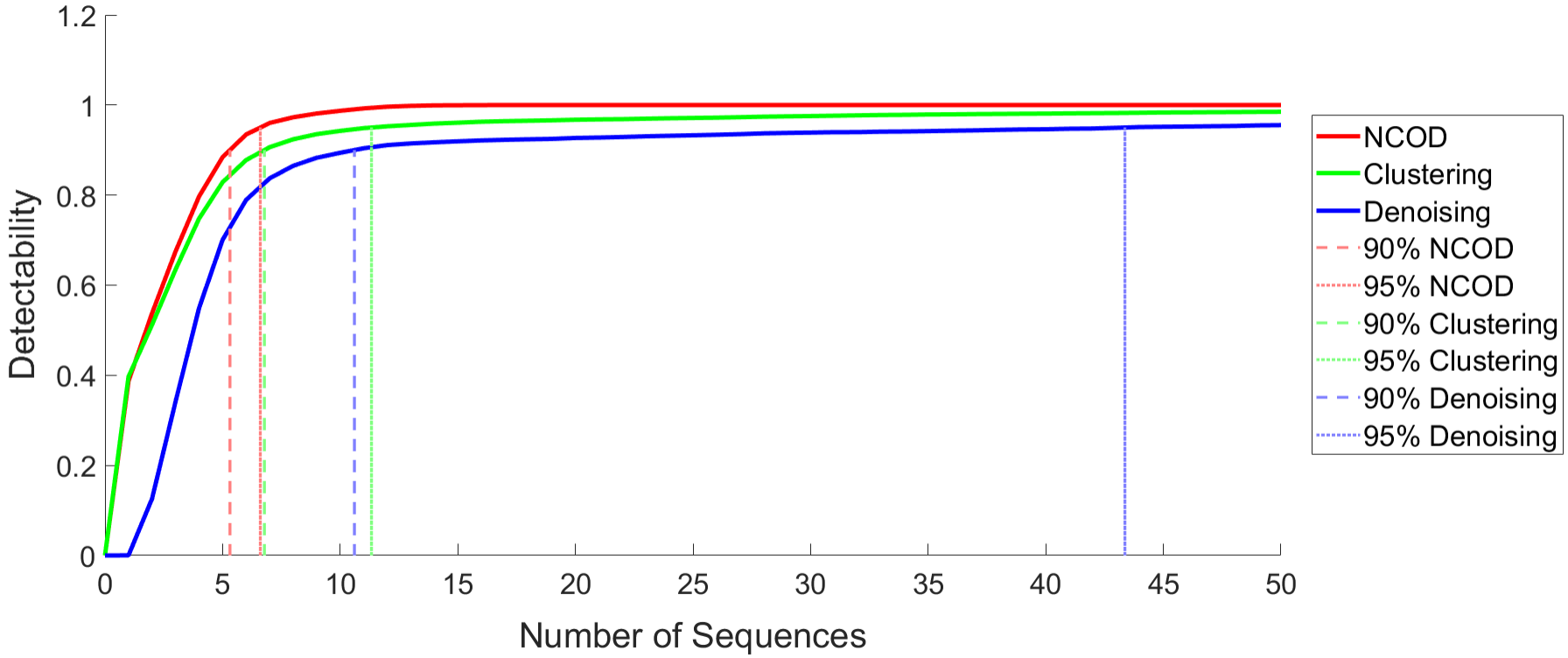
Performance:
Sequences Required By Taxon



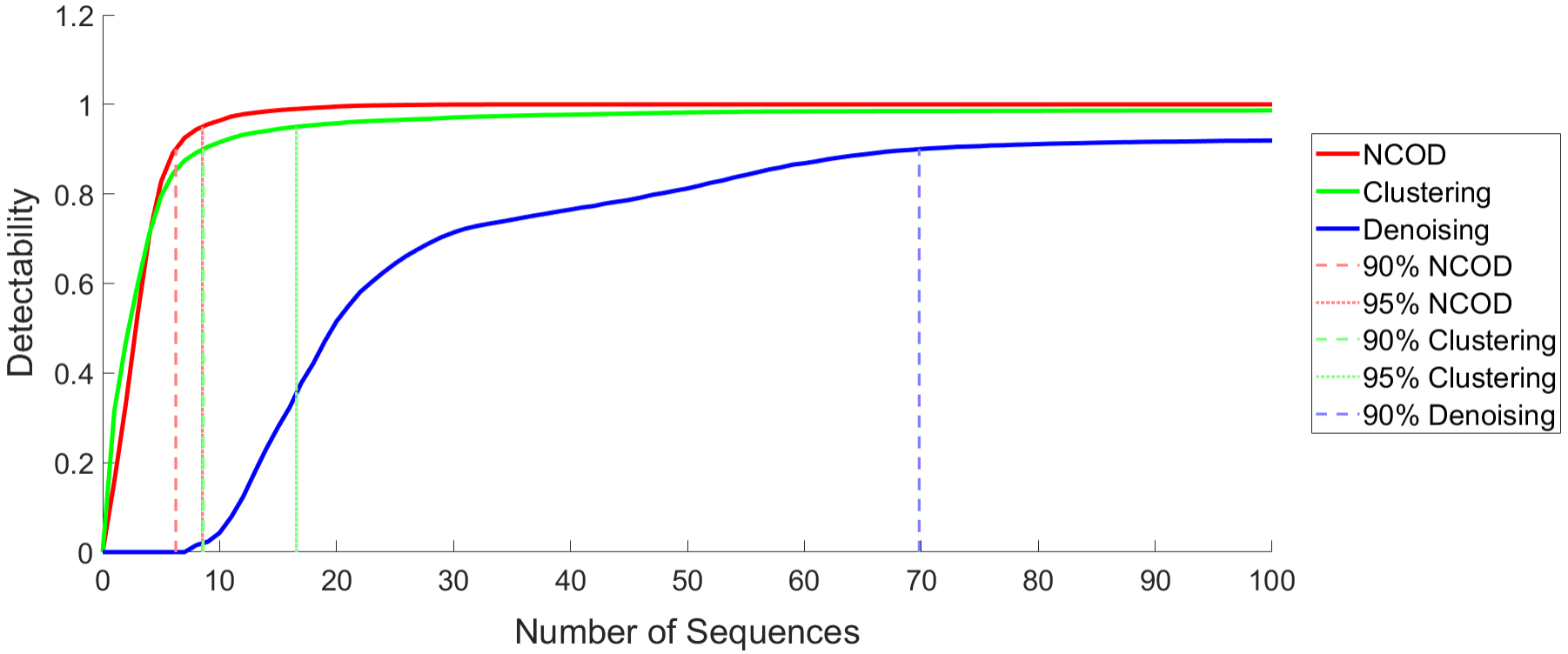
Performance:
Sequences Required By Port



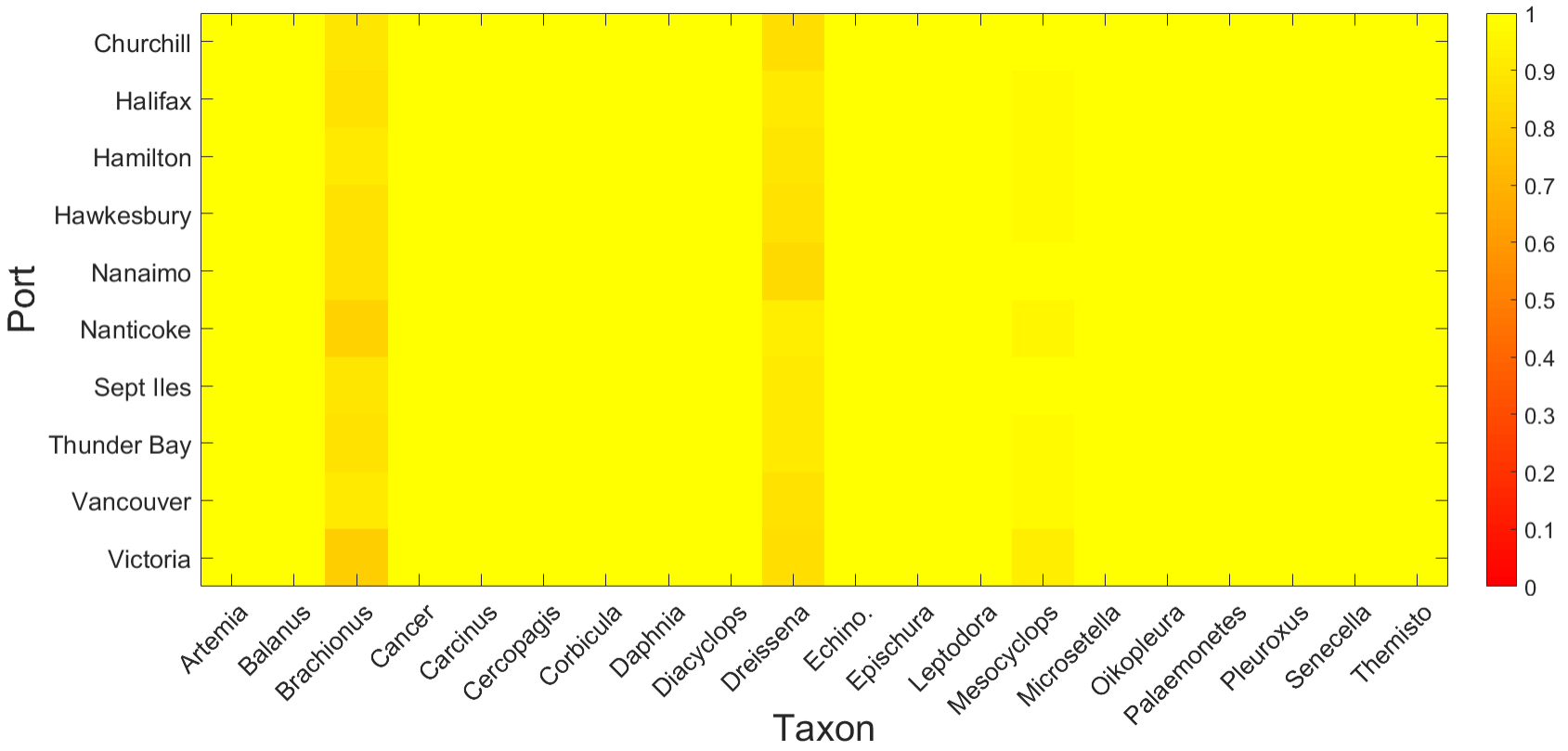
Performance:
Ratio of Taxa Recovered
Optimized for Detection of AIS



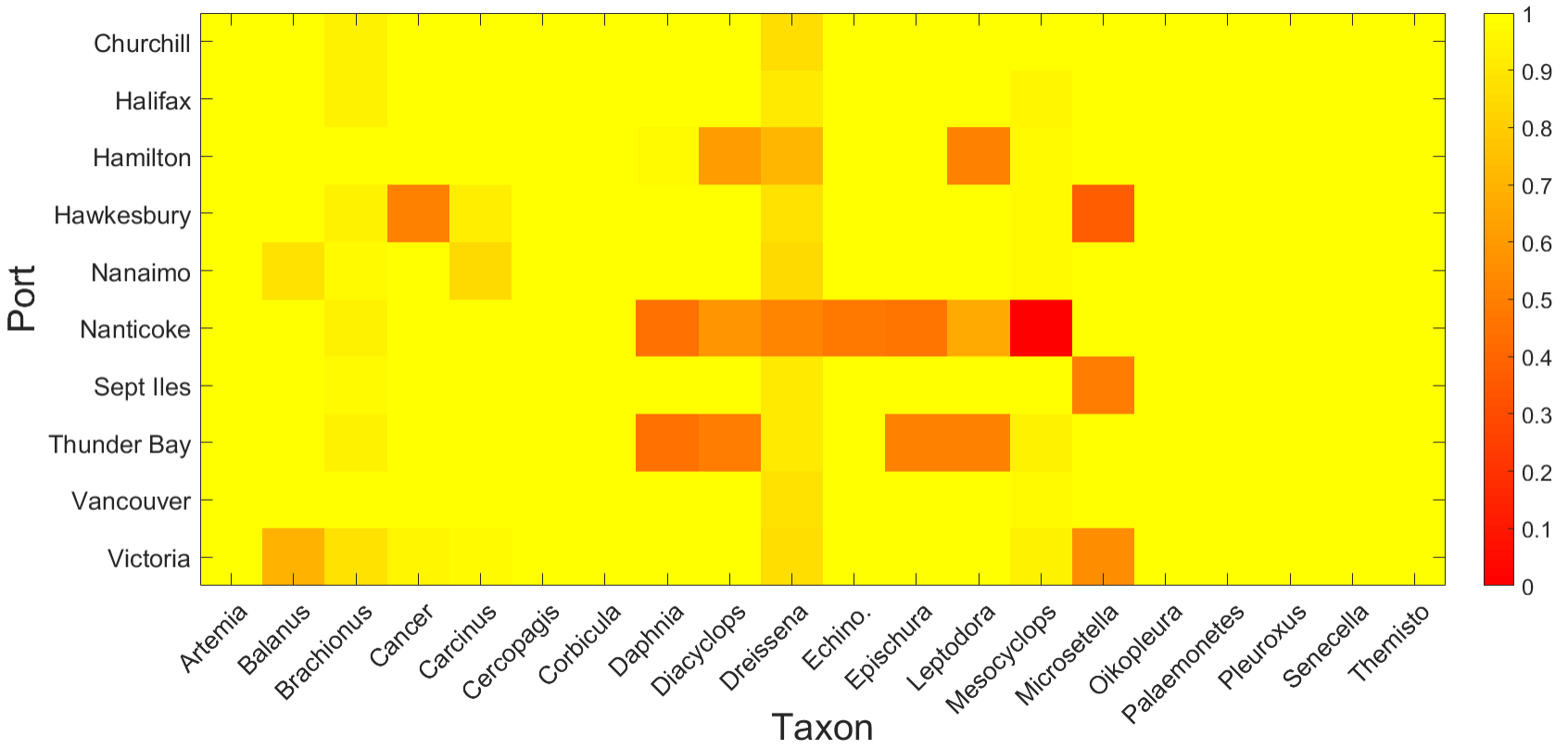
Performance:
Ratio of Taxa Recovered
Optimized for Species Richness Estimates



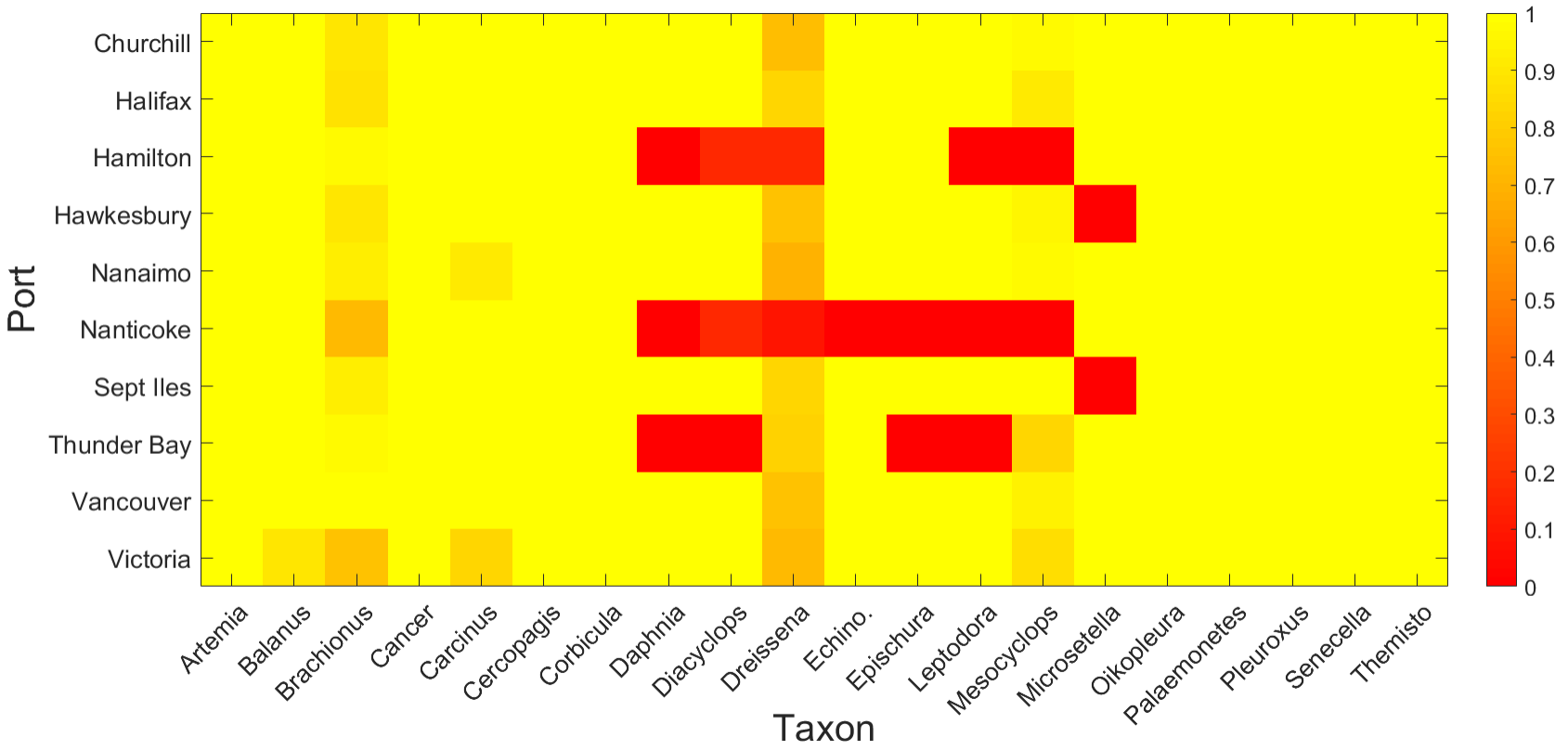
Performance:
Detection Probability (10 sequences)
Without Clustering Or Denoising



Performance:
Detection Probability (10 sequences)
Using Clustering



Performance:
Detection Probability (10 sequences)
Using Denoising



Conclusions:

- Using pre-existing sequence data, we can optimize our pipelines for given taxa/site
- Parameter selection should reflect your goal
  - Estimating Species Richness? Slightly more stringent filtering, longer sequences, no singletons, clustering or denoising is fine
  - Detecting AIS? Slightly relaxed filtering, longer sequences, no singletons, no clustering or denoising
- Detectability and sensitivity varied across taxa and sites, especially with clustering or denoising
  - Sites and taxa that were problematic for clustering were also problematic for denoising

Conclusions:

- We used 454 pyrosequencing
  - The pipeline can easily be adapted for paired reads (i.e. Illumina)
  - Our suggestions for parameter selection can be considered the minimum requirements to get the same performance we had
  - With better sequencing technology, performance using our pipeline will be no worse than what we achieved
- Molecular detection is not perfect
  - Parameter selection is a balancing act between false positive and false negative error

Thank you!


SHARCNET™


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


CAISN
CANADIAN
AQUATIC
INVASIVE
SPECIES
NETWORK


NSERC
CRSNG