The zebra (mussel) genome

A developing tool for research on prevention and control

Michael A. McCartney, Ph.D.,
Sophie Mallez, Ph.D.

University of Minnesota, Twin Cities
mmccartn@umn.edu
612-301-7703

UMISC 2018
Mayo Civic Center
Rochester, MN
October 16, 2018
Zebra genomes

Image: M. Manceau

2018

2010

Not yet

Image National Geographic

Image M. Manceau
The Zebra Mussel Genome Project

Keegan Lund, MN DNR
Photo: A Souffle, Minneapolis Star Tribune
Why sequence the zebra mussel genome?

1. To isolate 1000’s of markers to study invasion pathways

2. To create a resource for study of one the world’s most damaging AIS

3. To characterize “invasiveness” genes and target genes for biocontrol

Photo: T. Lafrancois
Regions of rapid spread between clustered lakes

Emerging regions of spread

2017
Minnesota’s rate of new inland invasions is now among the highest in the US

Data from USGS NAS program; MN DNR

We can still prevent the infestation of many prized water bodies!

- Prevention must be targeted by
  - Mapping transport pathways, identifying invasion sources and vectors (boats, docks, lifts...)
    - Overland & downstream spread studies (empirical, modeling)
    - Genetics/genomics: “invasion forensics”
Could invasion genetics identify routes and source water bodies at the scale of a US state?

Mallez S, McCartney MA (2018) *Biological Invasions*
Sampling infested waters 2014-2017

Microsatellites
- 9 markers/mussel
- 76 sites
- 48 water bodies
- 2257 mussels

SNPs
- > 6000 markers/mussel
- 91 sites
- 73 water bodies
- 1445 mussels
Sampling zebra mussels

Potential source waters throughout the Great Lakes
MN inland lakes show strong genetic structure

9 microsatellite loci

K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10

Mille Lacs
Prior
Alexandria area
Brainerd area
Pelican Rapids area
MN inland lakes show strong genetic structure

Some well-defined genetic clusters allow invasion model testing
The role of hub lakes in MN invasions
Mille Lacs Lake: a super-spreader?
• Highly connected by boater traffic to other lakes
• Infested early (2005)

Mille Lacs was not a source for 35 lakes that were infested post-2005
1. Overland transport from outside the region
2. Local spread

Results guide local management: (community based social marketing, identify short-distance spread vectors, regional inspection strategies...?)
But the sources for introductions from outside these regions remained obscure.
The Zebra Mussel Genome Project:

A reference genome to score Single Nucleotide Polymorphisms (SNPs) to study invasion pathways

Illumina sequencing only
N50: 2.2 kb
Contigs: ~500,000
Coverage: 100x

Juan Abrahante
Population genomics to resolve invasion sources

Sequence-Based Genotyping

**Step 1**
Construct reduced representation libraries (RRLs) by digesting each DNA sample with a restriction enzyme (SbfI).

**Step 2**
Ligate custom ‘barcoded’ adaptors to sticky ends of restriction site. Each sample has its own unique barcode sequence.

**Step 3**
Pool digested and barcoded DNA into a single tube. Perform PCR amplification, library preparation, and sequencing on Illumina platform.

**Step 4**
Use barcodes to assign sequences to samples. Produce a file of DNA sequence data for each sample.

Modified from Myles (2013)
Outcome 1: SNPs detect genetic differences between source populations that microsatellites do not.

Genetic structure plots from *Admixture*. From Mallez et al. (in prep.)
Outcome 2: SNPs implicate Great Lakes, large rivers as sources for inland lake invasions

Genetic structure plots from Admixture. 64 MN inland lakes, 5 rivers and streams
Outcome 3: Local sources for rapid spread in clustered lake regions is confirmed.
Outcome 4: Emerging regions: multiple pathways in...
Why sequence the zebra mussel genome?

2. To create a resource for study of one the world’s most damaging AIS
   - Of IUCN’s “100 World Worst” invasive species
     - 29 have complete genomes of varying quality
     - 9 were sequenced to study invasion biology *per se*

3. To characterize “invasiveness” genes and target genes for biocontrol
   - Byssal thread synthesis and attachment
   - Shell formation, mineralization, low Ca\(^{2+}\) tolerance (an Achilles' heel?)

Zebra mussels near threshold (14 mg/L Ca\(^{2+}\)) in Lake Superior grow thin, fragile shells. Photo T. Lafrancois
Target genes for genetic biocontrol

McCartney et al. (in review)
<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>Commercial interest</th>
<th>% genome in DNA repeats</th>
<th>Assembly level</th>
<th>Number of contigs</th>
<th>Contig N50 (bp)</th>
<th>Genome length (Mb)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathymodiolus platifrons</td>
<td>Hydrothermal vent mussel</td>
<td>None</td>
<td>47.9</td>
<td>Scaffold</td>
<td>272,497</td>
<td>12,602</td>
<td>1,658.2</td>
<td>Sun et al. 2017</td>
</tr>
<tr>
<td>Chlamys farreri</td>
<td>Zhikong (Chinese) scallop</td>
<td>Wild harvest and culture</td>
<td>32</td>
<td>Scaffold</td>
<td>148,999</td>
<td>21,500</td>
<td>779.9</td>
<td>Li et al. 2017</td>
</tr>
<tr>
<td>Crassostrea gigas</td>
<td>Pacific oyster</td>
<td>Hatchery culture—leads aquatic animals in global harvest</td>
<td>36</td>
<td>Scaffold</td>
<td>30,460</td>
<td>31,239</td>
<td>557.7</td>
<td>Zhang et al. 2012</td>
</tr>
<tr>
<td>Crassostrea virginica</td>
<td>Eastern oyster</td>
<td>Wild harvest and hatchery culture</td>
<td></td>
<td>Chromosome</td>
<td>669</td>
<td>1,971,208</td>
<td>684.7</td>
<td>Gómez-Chiarri et al. 2015</td>
</tr>
<tr>
<td>Mizuhopecten (Patinopectin) yessoensis</td>
<td>Yesso scalllop</td>
<td>Culture from wild seed</td>
<td>39</td>
<td>Scaffold</td>
<td>120,022</td>
<td>65,014</td>
<td>987.6</td>
<td>Wang et al. 2017</td>
</tr>
<tr>
<td>Modiolus philippinarum</td>
<td>Phillipine horse mussel</td>
<td>None</td>
<td>62</td>
<td>Scaffold</td>
<td>301,873</td>
<td>18,389</td>
<td>2,629.6</td>
<td>Sun et al. 2017</td>
</tr>
<tr>
<td>Pinctada martensii</td>
<td>Akoya pearl oyster</td>
<td>Cultured pearls</td>
<td>50</td>
<td>Chromosome</td>
<td>85,944</td>
<td>21,518</td>
<td>991.0</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

From McCartney et al. (in review)
Zebra mussel genome project strategy

- Use long-read (PacBio) sequencing to improve assembly of repeat-DNA regions
- Hi-C to place the assembled genome onto chromosomes
- Sequence RNAs expressed by
  - Byssal thread genes (foot)
  - Shell formation/biomineralization genes (mantle)
  - Thermal tolerance genes (gill)
  - Developmental and sex determination genes (gonad, embryos, larvae)
- Automated annotation, comparative genomics, homology searching for target genes
N50: A Measure of Assembly Contiguity

Definition: 50% of the genome is in contigs larger than N50.

Example: 1 Mb genome

Contigs

1000 kb

50% (N50 = 40 kb)

Context
Human Genome Project (2001) N50: 82 kb
Human Latest Release GRCh38 N50: 67,794 kb (67 Mb)

Daryl Gohl, UMGC
Improving the zebra mussel assembly with long reads

**Illumina sequencing only**
- N50: 2.2 kb
- Contigs: ~500,000
- Coverage: 100x

**PacBio sequencing**
- N50: 1.1 Mb
- Contigs: 2,683
- Coverage: ~80x

Kevin Silverstein – RIS, MSI
Ying Zhang – RIS, MSI
Serge Koren – U. Maryland
Hi-C to Map Genome onto Chromosomes

Ben Auch
Zebra Mussel Genome Project – Next Steps

1. Canu assembly
2. Polish with Illumina reads
3. Scaffold using Hi-C data
4. PacBio IsoSeq Data?
5. Analyze RNA-Seq data
6. Annotate genome (gene models)
7. Comparative genomics, homology searches

- Genome Assembly
- Genome Annotation
- Genome Analysis

We are here...

Minnesota Supercomputing Institute

UNIVERSITY OF MINNESOTA GENOMICS CENTER

University of Minnesota
Thanks: genetics and genomics

**MnDNR** Dan Swanson, Rich Rezanka, Keegan Lund, Mark Ranweiler, and several others for help collecting

**USGS** Wendylee Stott, Mary Anne Evans **NOAA** Ashley Baldridge Elgin; **INHS** Jeremy Tiemann **NPS** B. Karns, B Moraska Lafrancois, T. Lafrancois, M. Hove (collaborators on APIS project) and many and others for collecting from the Great Lakes Region

**McCartney Lab** Dr. Sophie Mallez, Melody Truong

**UMN Genomics Center** K Beckman, D Gohl, S Anderson, J Garbe, B Auch

**MN Supercomputing Institute** K Silverstein, Y Zhang, T Kono

**University of MD** S Koren for assembly

**Funding:** Clean Water, Land and Legacy Fund (2014-2016); Environment and Natural Resources Trust Fund (current); MAISRC, Gull Chain of Lakes Association, Pelican Lakes Association, Lake Washington Improvement Association.