Development and Characterization of The O_my_50K SNP Array

Yniv Palti
USDA ARS
National Center for Cool and Cold Water Aquaculture
Leeown, West Virginia

Why SNP Chip?
• More SNPs than RADs, faster and easier to process and call genotypes and much more standardized (easier to compare between populations, studies and labs).
• More expensive, but cost coming down (pending volume of samples).
• Useful to screen SNPs for population traceability, GWAS, QTL mapping, genome-wide enabled selection and for generating high density genome maps.

Our SNP Discovery Approach Designed to Filter-Out PSVs and MSVs*

Phase 1: SbfI RADs of 19 DH lines in one indexed library pool. Sequenced on one Illumina HiSeq lane
Phylogenetic analysis using 1,971 SbfI RAD SNPs
Selected 11 lines representing the nodes of the phylogenetic tree

Phase 2: PstI RADs of the 11 selected DH lines in two indexed libraries. Sequenced on four Illumina HiSeq lanes
Produced a database resource of 145,168 RAD SNPs for rainbow trout

* Palti et al. 2014, Molecular Ecology Resources 14: 588-596

SNP Sources and Validation Rate

<table>
<thead>
<tr>
<th>Category</th>
<th>DH RADs</th>
<th>Re-Seq</th>
<th>Other</th>
<th>Total*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotypes</td>
<td>57,501</td>
<td>57,501</td>
<td>57,501</td>
<td>57,501</td>
</tr>
<tr>
<td>Call Rate (&gt;97%)</td>
<td>999</td>
<td>1,366</td>
<td>434</td>
<td>2,130</td>
</tr>
<tr>
<td>No Minor Homozygotes</td>
<td>526</td>
<td>1,369</td>
<td>459</td>
<td>1,942</td>
</tr>
<tr>
<td>Off Target Variant</td>
<td>139</td>
<td>291</td>
<td>73</td>
<td>415</td>
</tr>
<tr>
<td>Other</td>
<td>1,084</td>
<td>1,889</td>
<td>615</td>
<td>2,892</td>
</tr>
<tr>
<td>Monomorphic High Resolution</td>
<td>223</td>
<td>420</td>
<td>118</td>
<td>654</td>
</tr>
<tr>
<td>Polymorphic High Resolution</td>
<td>27,061</td>
<td>33,824</td>
<td>9,265</td>
<td>49,468</td>
</tr>
</tbody>
</table>
* Overlap between sources. ~40% of the SNPs were discovered by two sources.
* Palti et al. 2015, Molecular Ecology Resources E-Published

Fish Included in The Validation Panel
• A total of 960 samples were included in the genotype validation panel.
• 19 DH lines as previously described (Palti et al., 2014).
• 265 unrelated(?) fish representing 18 populations (Aquaculture and Wild).
• Full-sib families from 10 of the NCCCWA disease resistance breeding population with 39-90 offspring per family.

Polymorphism in Farmed vs. Wild Populations Sampled
• The array is very useful for rainbow trout aquaculture populations with more than 40,900 polymorphic markers per population.
• For wild populations that were confounded by a smaller sample size the number of polymorphic markers was between 10,577 and 24,330.

Polymorphism in Farmed vs. Wild Populations Sampled

Fish Included in The Validation Panel
• A total of 960 samples were included in the genotype validation panel.
• 19 DH lines as previously described (Palti et al., 2014).
• 265 unrelated(?) fish representing 18 populations (Aquaculture and Wild).
• Full-sib families from 10 of the NCCCWA disease resistance breeding population with 39-90 offspring per family.
Population Traceability:
Example from Sashin Steelhead vs. Rainbow from Hale and Nichols

<table>
<thead>
<tr>
<th>Traceability</th>
<th>Gtype1</th>
<th>Gtype2</th>
<th>Nf1</th>
<th>Nf2</th>
<th>Geno1</th>
<th>Geno2</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:1</td>
<td>AA</td>
<td>BB</td>
<td>0</td>
<td>5</td>
<td>AA AA AA AA AA</td>
<td>BB BB BB BB BB</td>
</tr>
<tr>
<td>00:2</td>
<td>AA</td>
<td>BB</td>
<td>0</td>
<td>3</td>
<td>AA AA AA AA AA</td>
<td>BB BB BB BB BB</td>
</tr>
<tr>
<td>00:3</td>
<td>AA</td>
<td>BB</td>
<td>0</td>
<td>5</td>
<td>AA AA AA AA AA</td>
<td>BB BB BB BB BB</td>
</tr>
<tr>
<td>00:4</td>
<td>BB</td>
<td>AA</td>
<td>1</td>
<td>5</td>
<td>BB BB BB BB BB</td>
<td>AA AA AA AA AA</td>
</tr>
<tr>
<td>00:5</td>
<td>BB</td>
<td>AA</td>
<td>2</td>
<td>1</td>
<td>BB BB BB</td>
<td>AA</td>
</tr>
<tr>
<td>00:6</td>
<td>BB</td>
<td>AA</td>
<td>3</td>
<td>5</td>
<td>BB BB BB BB BB</td>
<td>AA AA AA AA AA</td>
</tr>
<tr>
<td>00:7</td>
<td>AA</td>
<td>BB</td>
<td>5</td>
<td>5</td>
<td>AA AA AA AA AA</td>
<td>BB BB BB BB BB</td>
</tr>
</tbody>
</table>

Distribution of MAFs from the rainbow trout 57K SNP chip data.

Distribution of SNPs assigned to rainbow trout chromosomes using two-point linkage analysis.

Accuracy of Genomic Selection for BCWD Resistance in Rainbow Trout

- Presentation and Poster by Vallejo et al.
- Using SNP chip genotypes and historic phenotypes and progeny testing data from the NCCWCA 2005 and 2007 year classes.
Assembly Statistics

<table>
<thead>
<tr>
<th>Species</th>
<th>Number</th>
<th>Size (Mb)</th>
<th>Average (Kb)</th>
<th>N50 (Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. Trout*</td>
<td>Contigs</td>
<td>445,600</td>
<td>1,684</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>79,941</td>
<td>1,877</td>
<td>23.5</td>
</tr>
<tr>
<td>N. Pike</td>
<td>Contigs</td>
<td>94,000</td>
<td>824</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>5,688</td>
<td>874</td>
<td>153</td>
</tr>
<tr>
<td>A. Cod</td>
<td>Contigs</td>
<td>284,239</td>
<td>536</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Scaffolds</td>
<td>6,467</td>
<td>611</td>
<td>94</td>
</tr>
</tbody>
</table>

* The trout’s “chromosomes” assembly which includes all scaffolds and non-scaffolded contigs (>500bp) has a cumulative size of ~2.1 Gb.

Evaluation of Assembly Quality and Coverage

Alignment of our ~50K validated SNPs from the SNP chip to Scaffolds (~1.9 Gb):
- 90% exact match hits on the assembly. Of those, 99.5% are unique hits.
- Where are the PSVs/MSVs?

Anchoring the Assembly to Chromosomes (part 1)

- We have anchored ~45K SNPs from the 57K array onto chromosomes, of which we mapped ~44K to 13,334 scaffolds (1.4Gb; 68% of assembly).

Anchoring the Assembly to Chromosomes (part 2)

- Of the 44K SNPs, ~18K (40%) were mapped to 10,844 scaffolds that were not previously anchored to chromosomes; representing 0.41Gb (20% length of the assembly).
- The remaining 288,141 un-anchored scaffolds and contigs representing 32% of the genome assembly have an average size of only 2,316bp.

Distribution of SNPs on the rainbow trout genome assembly scaffolds

What Next?

1. To anchor the remaining 32% of small scaffolds and contigs we can add more SNPs to the genetic map, but first we will need to increase the N50 scaffold size.
2. We can use mate-pair libraries (Mike Miller) to improve scaffolds size and also plan to test long reads sequencing technologies (Moleculo; PacBio).
3. To order and orient more scaffolds on the chromosomes we will use the HD genetic map produced using genotype data from the O_my_50K SNP array (collaboration with Norway).
So, in the world of Sand Castles: This is the ultimate goal we are aiming to reach.

* Recycling an illustration from John Liu (outlift), who is recycling it from previous frustrated genome assembly project leaders.

“And it looks like we are off to a good start…”

Acknowledgements

USDA-ARS-NCCCWA
Guangtu Gao
Roseanna Long
Caird Rexroad III
Kristy Shewbridge
Sixin Liu
Roger Vallejo
Washington State Univ.
Gary Thorgaard
Paul Wheeler

UC-Davis
Mike Miller
Jeanne Miller
Ming-Cheng Luo
Yuqin Hu

WWU
Jianbo Yao

MTSU
Mohamed Salem

CIGENE
Sigbjørn Lien

Aquagen
Matthew Baranski

NOFIMA
Matthew Kent

Affymetrix

Funding Sources

National Research Initiative
Animal Genome Program

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2011-67015-30091 from the USDA National Institute of Food and Agriculture, and by funds from the USDA Agricultural Research Service in house project no. 1930-31000-009.