Applied Biosystems™ Eureka™ Genotyping 555-plex panel - understanding migration and harvest of residual tetrasomic Chum Salmon sub-populations

H. Koshinsky, J.D. Curry, M.A. Patil, A. Pirani, V. Missiran, V. Joshi
Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA, USA, 95051

SUMMARY

Using a low density genotyping panel for surveillance is critical to the management of coastal fishery populations. The amount of genetic information required depends on the application and can range from a few hundred markers to full genome sequence for more complicated questions. For example, a panel that interrogated a small number of markers could be used to determine the likely country of origin of a catch in question, trace migration routes or understand bycatch.

While determining the genotype of many diploid agriculturally important species is relatively straightforward, this is not necessarily the case with polyploid species such as chum salmon. At any bi-allelic locus, a diploid has three possible genotypes and a tetraploid has five possible genotypes.

The Applied Biosystems™ Eureka™ Genotyping 550+plex panel was developed to trace migration and harvest of sub-populations in residual tetrasomic chum salmon. This panel includes five markers that have five possible genotypes. Additionally, approximately 95% of the markers segregate as diploid and 5% of the markers segregate as compressed diploid (the genotype is in the background of a non-segregating allele). All three types of loci are interrogated in a single well and data is generated on a single next-generation sequencing (NGS) run without the need for extra sequencing depth. The availability of over 4,000 barcodes enables processing of over 4,000 samples in a single sequencing run for fast turnaround time.

Across 96 samples and 516 diploid markers, the total call rate (the percent of markers generating a genotyping call) was 99.7% and the total relative concordance to an orthogonal truth was 98.5%. Fifty-three of the diploid markers in the panel either had call rate below 95% or concordance below 98%, but clear evidence that the truth could likely be incorrect. The data demonstrates that Eureka Genotyping panels and workflow can be used with residual tetrasomic species to provide accurate genetic information that can be used for surveillance and management.

INTRODUCTION

Gene sequence similarity is caused by duplication either at a local level (gene duplication) or a genome level. Sequences in these duplications confound the determination of genotypes and its associated assessment of genetic variation. Genotyping chum salmon (a residual tetrasomic species) is impacted by ancestral duplication.

We developed a targeted Eureka Genotyping Chum Salmon panel to monitor chum salmon sub-populations. The chum salmon panel includes loci that are diploid, compressed diploid and tetraploid in a single workflow and next generation sequencing (NGS) run.

The Applied Biosystems™ Eureka™ Genotyping platform is a high throughput, flexible and simple workflow with automated data analysis.

MATERIALS / METHODS

Design: Eureka myDesign™ was used to generate a 550+plex chum salmon panel and obtain genotypes on DNA extracted from fin clips of 192 samples.

Wet Lab: Eureka™ Genotyping Solution uses barcodes to identify both allele and locus and sample indices for sample identification. Each locus is interrogated with a probe triplet (left, left-prime, and right hybridization sequences). Tens to thousands of triplets are combined to create a Eureka genotyping panel. The Eureka genotyping workflow is shown in **Figure 1**. Sample DNA/crude lysate is mixed with the Eureka genotyping panel (Figure 1a) and allowed to hybridize overnight (Figure 1b), followed by a ligation reaction (Figure 1c). Use of crude lysate reduce the cost and time needed to purify genomic DNA. The completed ligation reaction is used as template for the sample indexing PCR reaction which adds sequencing reaction compatible adapters as well as a unique sample ID barcode (Figures 1d and e). Amplified samples are combined, purified on a silica-based column, quantified, and sequenced (Figure 1f).

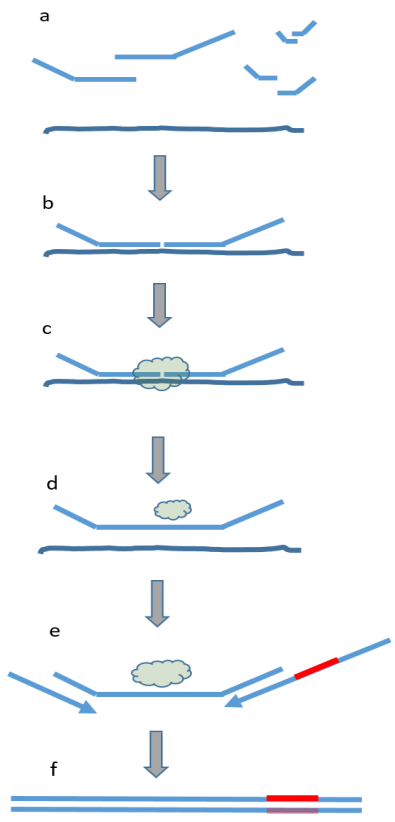

Figure 1: Eureka Genotyping Workflow.

Data Analysis: Applied Biosystems™ Eureka™ Analysis Suite software performs all steps to convert sequence reads to genotypes. (1) Each sequence read is assigned to appropriate sample index and allele + locus barcode (or not assigned if there is no match). (2) The tabulated reads are normalized, outlier samples are removed and reads are plotted in size vs contrast space. (3) The genotype of each sample and locus is inferred with our proprietary genotype caller which is an extension of the BRLMM-P clustering algorithm that adapts pre-positioned genotype cluster locations called “priors” to the sample data in a Bayesian step and computes three posterior cluster locations (**Figures 2 and 3**). The genotypes for loci that have more than three clusters are computed with fitTetra (Voorrips *et al.,* 2011) with examples shown in **Figure 4**.

Using a modified panel for genotyping in chum salmon, a same day workflow that goes from DNA to genotypes in 12 hours was also recently demonstrated (Figure 5).

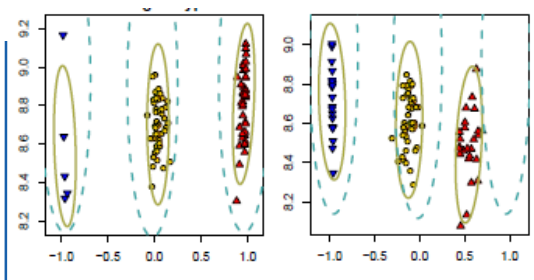

Figure 2. Typical cluster plots for diploid loci. Each plot is a single locus and each sample is a single point in the plot. The x-axis is contrast and the y-axis is size. The BB (blue), AB (yellow), or AA (red) genotype of each sample is shown.

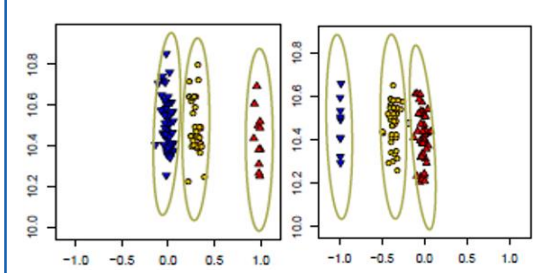

Figure 3. Cluster plots for compressed diploid loci. Each plot is a single locus and each sample is a single point in the plot. The x-axis is contrast and the y-axis is size. The BB (blue), AB (yellow), or AA (red) genotype of each sample is shown. In contrast to the cluster plots in **Figure 2**, in these plots the three clusters are compressed to the right (fixed AA background for this locus) or the left (fixed BB background for this locus).

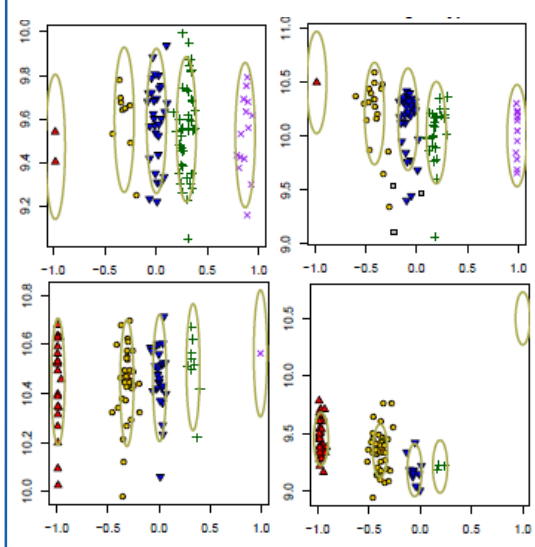

Figure 4. Cluster plots for tetraploid loci. Each plot is a single locus and each sample is a single point in the plot. The x-axis is contrast and the y-axis is size. The BBBB (red), ABBB (yellow), AABB (blue), AAAB (green) or AAAA (purple) genotype of each sample is shown.

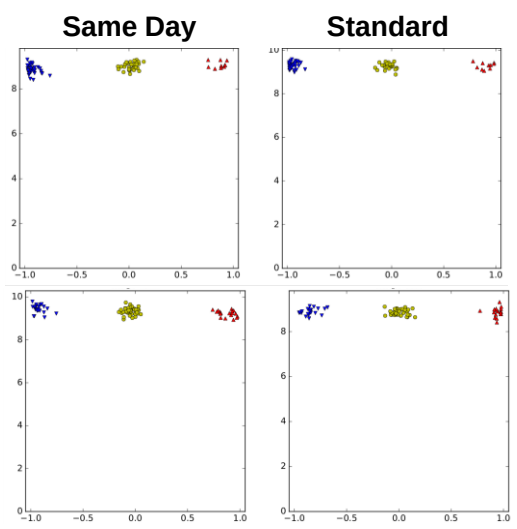

Figure 5. Cluster plots same day workflow and standard Eureka Genotyping workflow. Left panel is the same day workflow and right panel is the standard workflow for the same locus; top and bottom panels are two different loci. Each plot is a single locus and each sample is a single point in the plot. The x-axis is contrast and the y-axis is size. The BB (blue), AB (yellow), or AA (red) genotype of each sample is shown.

RESULTS

The 550+plex panel has a high total call rate (99.7%). All samples tested had a genotype on > 97.5% of the loci. The total relative concordance was (98.5%) from 96 samples and 516 diploid markers. Over fifty of the markers provided additional information not captured in the orthogonal truth, such as tetraploid information on a locus previously thought to be diploid (**Figure 4**), bi-allelic information on a locus previously thought to be monomorphic or clearer separation of the genotypes (**Figure 6**).

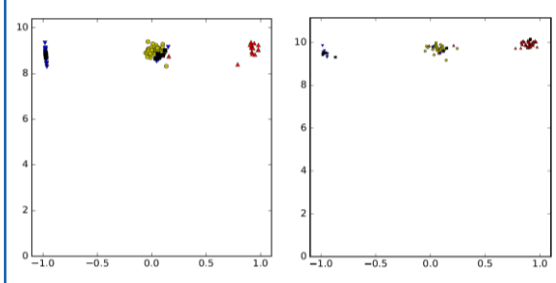

Figure 6. Cluster plots for diploid loci where genotypes from Eureka Genotyping provide clearer genotype resolution than the orthogonal truth. Each plot is a single locus and each sample is represented by a single point in the plot. The x-axis is contrast and the y-axis is size. The BB (blue), AB (yellow), or AA (red) provided orthogonal truth genotype of each sample is shown.

CONCLUSIONS

The Eureka Genotyping Solution workflow and Chum Salmon panel provide a streamlined, cost-effective method to generate SNP information on greater than 550 loci. 3500 samples (or samples in nine 384-well plates) can be pooled into a single high throughput NGS run. Clear genotype clusters were obtained for loci that presented as **standard diploid**, for loci that presented as **compressed diploid** and for loci that presented as **tetraploid**. The Eureka Genotyping platform has been applied to cattle, sheep, barley, corn, soy and wheat.

The Eureka Chum Salmon panel generated genotypes that can be used in applications such as:

- Monitoring of migration in transboundary rivers or high seas
- Assessment of bycatch
- Determination of country of origin of pirated fish

The recently demonstrated same day workflow that allows mid-plex high sample throughput of DNA to genotypes in 12 hours opens further applications.

REFERENCES

Voorrips, R.E., G. Gort, & B. Vosman. 2011. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12: 172-183.

ACKNOWLEDGEMENTS

We would like to thank James Seeb and his lab, and Garrett McKinney (College of Environment, University of Washington), Christopher Habicht (Alaska Department of Fish and Game) for supplying gDNA samples and SNPs of interest. We would also like to acknowledge Maria Shin at Thermo Fisher Scientific for sequence data generation and Dan Fitzpatrick for data analysis.

FURTHER INFORMATION

heather.koshinsky@thermofisher.com