**Summary**

This document collates information related to design and validation of the SFGL Albacore Tuna GTseq panel in full detail. There are still some open questions that need to be resolved. These are highlighted in red text. Known issues with resolutions are presented in purple text. The document is currently formatted as a narrative, simultaneously describing methods results and discussion to facilitate review.

**Outline**

**Overview**

Flowchart describing entire panel development

**Intro / Goals**

What is the motivation for the panel?

**Ascertainment Data**

Description of the RADseq data used to ascertain SNPs for feature selection

**Feature Selection**

What genomic features were selected for inclusion in the panel, and why?

**Primer Design**

Detailed methods from feature selection to primer (and probe) sequences.

**Panel Validation 1**

Initial test library results

**Panel Validation 2**

Second test library

**Panel\_289**

Description of the final panel

# Panel Development Overview

Diagram

Description automatically generated

**Figure 0:** Overview of panel development. Two ascertainment datasets were used to select three sets of genomic features for inclusion in the panel. Consensus sequences surrounding the target features were inferred using the ascertainment data and an assembly of the Bluefin Tuna genome. Primers were designed from these consensus sequences and the panel was optimized using two validation libraries.

\* There are 170 spatial outliers, but 59 overlap with the FST outliers.

\* There are 384 individual DNA libraries in validation library 2, but only the best 100 were used.

# Introduction

# Ascertainment Data

## ddRADseq

The GTseq panel was developed using an ascertainment dataset of SNPs identified among 361 Albacore Tuna sampled throughout the Pacific ocean by Vaux *et al* (2021). Vaux *et al* used ddRADseq with *SphI* and *PstI.* Raw reads were processed using *process\_radtags* from STACKS v2.41, and processed RADtags were mapped to a Pacific Bluefin genome assembly (RefSeq: GCA\_000418415.1) using BWA-MEM. SNPs were called using ref\_map pipeline from STACKS v2.41 with variant sites and genotypes called at a significance level of 5% in *gstacks* (--var\_alpha 0.05 --gt-alpha 0.05).

308 individuals genotyped successfully. Both ascertainment datasets described below contained genotypes from the same 308 individuals from the North and South Pacific Oceans using ddRADseq. Sample sizes and locations are in Table 1 below.

|  |  |  |
| --- | --- | --- |
| Location | Abbreviation | n |
| North Pacific |  |  |
| Baja California | BA | 4 |
| Southern California | CS | 39 |
| Northern California | CN | 19 |
| Oregon | OR | 31 |
| Washington | WA | 27 |
| British Columbia | BC | 10 |
| Hawaii | HW | 25 |
| NW Hawaii | HN | 28 |
| Japan | JP | 22 |
| Philippine Sea | PH | 29 |
|  |  |  |
| South Pacific |  |  |
| New Caledonia | NC | 54 |
| Tasmania | TS | 20 |

**Table 1:** Sampling locations and sample size in VauxSNP and MultiSNP datasets

## North and South Pacific Dataset

Further downstream processing of the resulting RADtag catalog was completed in the *populations* module of STACKS and produced multiple datasets. In panel development we use the “North and South Pacific” dataset, as described in Vaux *et al* (2021). This filtering approach considered all North Pacific samples a single population, and all South Pacific samples a second, single population. To be retained in the dataset, a RADtag must be present in at least 95% of individuals within a population (Stacks Populations -r 0.95), and have a minor allele frequency (MAF) of at least 5%. Only the first SNP per RADtag was retained (*populations* --write-single-snp). After filtering in *populations*, SNPs significantly outside Hardy-Weinberg proportions within each group (North and South Pacific) were excluded, followed by potential paralogues identified using HDplot and Paralog-Finder, SNPs in LD within each population, low coverage SNPs (minimum depth per individual < 10), SNPs with ≥20% missing data among all individuals, and SNPs that were considered coverage depth outliers. We refer to this SNP dataset as “North and South Pacific dataset” to maintain consistency with Vaux *et al* (Table 2). This dataset was used as the input for identification of FST and spatial outliers during feature selection.

## MultiSNP Dataset

A second ascertainment dataset drawn from the same ddRADseq library was used to select neutral SNPs and for primer design. The MAF filtering cutoff (5%) and the --write-single-snp option present significant challenges when attempting to use the North and South Pacific dataset for these purposes. These challenges are elaborated below:

1. **Neutral SNP Feature Selection:**

During pilot analyses of the raw ddRADseq data, the estimated site frequency spectrum (SFS) indicated that the majority of the SNPs among individuals were substantially below the MAF cutoff used for the North and South Pacific dataset. Given the assumption that albacore have a large effective population size (*Ne*), this result is hardly surprising. Drift should be less effective at eliminating rare alleles from the population over time compared to a population with smaller Ne. Therefore, while the MAF cutoff used for the Vaux et al (2021) manuscript may have been appropriate for adaptive outlier scans between North and South Pacific, it likely obscures neutral structure that may be present at finer spatial scales (De la Cruz & Raska 2014). Or, from a coalescent perspective, the 5% MAF filter likely truncates the evolutionary history that can be inferred in the data to only ancient events (Linck & Battey 2019). In short, different allele frequency bins reveal different aspects of past demographic and selective events. Therefore, biasing the site frequency spectrum towards common SNPs will limit the future utility of neutral markers in the panel to infer basic population genomic parameters and population structure.

1. **Primer Design**

The --write-single-snp option from the *populations* module of STACKS keeps only the first SNP along the RADtag and discards the rest. In the case of common SNPs (i.e. MAF > 5%), this presents a major problem for primer design because unaccounted for variation within the regions of the genome covered by marker primers or probes will lead to allele dropout in the resulting GTseq panel.

To solve the problems presented by the filtering approach of the North and South Pacific dataset to both neutral feature selection and primer design, a new dataset (referred to as the MultiSNP dataset, Table 2) was developed from the same set of 31,523 RADtags used to develop the North and South Pacific dataset. The same set of filtering criteria was used with two exceptions: instead of a MAF filter, we used a minor allele count filter of three (to exclude singleton alleles), and we retained all SNPs in the RADtag. Filtering for HWE, paralogs, LD, coverage and depth was applied at the RADtag level and used the same list of excluded RADtags as the North and South Pacific dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset Name | n SNPs | Filtering Summary | Applications in Panel Development |
| North and South Pacific | 12,872 | -r 0.95  --min-maf 0.05  --write-single-SNP  HWE: FDR p-value 0.05  Paralogs (HDplot)  LD  min-depth 10  max-missing 0.2  high coverage | Feature Selection: FST Outlier  Feature Selection: Spatial Outlier |
|  |  |  |  |
| MultiSNP | 166,248 | -r 0.95  --min-mac 3  HWE: FDR p-value 0.05  Paralogs (HDplot)  LD  min-depth 10  max-missing 0.2  high coverage | Feature Selection: Neutral  Primer Design |

**Table 2:** Ascertainment datasets, number of SNPs, summary of filtering steps with differences bolded, and the applications of each dataset during panel development.

# Feature Selection

## Overview

Feature selection is the process of identifying genetic variants (in this case SNPs) to develop into markers for the initial panel. Most other amplicon panels have identifying structure among known populations, analysis of markers underlying a trait of interest or under strong selection, or conducting parentage analysis as their priorities. But, (a) the only structure we know confidently is that North Pacific individuals are weakly differentiated from South Pacific individual and that this structure can be captured with just 84 FST outliers, (b) we do not have any traits of interest or preceding GWAS, and (c) the likelihood we'll ever use accurate parentage assignment is extremely low given the life history of albacore. Given constraints on primer interactions and read depth, we would like select approximately 500 genomic features for initial primer design. So, assuming we want to retain the 84 known North-South FST outliers, the question for feature selection is what to do with the remaining 426 markers.

We chose to supplement the FST outliers with a complementary analysis that attempts to identify markers that reveal finer scale population structure (spatial outliers), and a set of neutral loci that can be used for demographic inference or an unbiased estimate of population genetic parameters (neutral markers). In the event that the spatial analysis reveals spurious structure, the spatial outliers will redound to additional neutral markers, albeit ones not drawn randomly from the empirical SFS.

## FST Outliers

Detailed methods for the selection of the FST outliers is available in Vaux *et al* (2021). In brief, the data was organized into North and South Pacific populations and four outlier scans were applied to identify putatively adaptive SNPs: FSTHET, OUTFLANK, BAYESCAN, and PCADAPT. Any SNPs identified in at least three of these four outlier scans were retained as putatively adaptive FST outliers (n = 84, Supplemental Table 4)

## Spatial Outliers - Rationale

Given that effective management of the albacore fishery requires characterizing the demographic connectivity among stocks rather than genetic isolation among populations, the large effective population size of albacore may create a disconnect between what a panel based solely on FST outliers can say and how it is interpreted. As Ne becomes very large, even highly demographically independent populations will demonstrate low genetic differentiation (Gagnaire & Gaggiotti 2016; Gagnaire *et al.* 2015),because FST varies as a function of the product of Ne and the migration rate ( FST ≈ 1/(1 + 4Nem) ). In this scenario, the number of migrants that will produce apparent panmixia is substantially fewer than the number needed to ensure the demographic rescue of different stocks. There is a troubling dearth of research or discussion about the disconnect between migration rates needed to create genetic panmixia and dispersal rates needed for demographic coupling of fishery stocks, but some have put the value needed for demographic coupling as high as 10% (Waples *et al.* 2008), which will produce apparent panmixia in all but the smallest marine populations (Ne < 1000). Therefore, our tolerance of type II error for the detection of population structure or identification of putatively adaptive variation should be very low.

In contrast, the methods in Vaux et al (2021) are geared (correctly) towards reducing type I error: the spurious identification of putatively adaptive SNPs. For example, type I error is reduced in the 2021 study by requiring that a SNP be identified by three of four outlier scans methods before describing it as a putatively adaptive outlier. Beyond the conservative requirement of identification by three of four outlier scans, there are two additional reasons to believe that type II error may be high with such an approach:

1. **We should expect a polygenic architecture of adaptation characterized by subtle allele frequency shifts**: If we assume that albacore population genetics are typical of highly migratory pelagic fish, (1) large Ne should bias local adaptation towards polygenic architectures and/or soft sweeps (Hermisson & Pennings 2005; Höllinger *et al.* 2019), and (2) a high migration rate should bias local adaptation towards transient covariance of adaptive alleles and accompanying small allele frequency changes, or the evolution of inversions/supergenes (Tigano & Friesen 2016; Yeaman 2015; Yeaman & Whitlock 2011). Except in the last case (evolution of inversions/supergenes), the expectation is that adaptation will be driven by subtle allele frequency shifts across many loci. FST outlier scans are very poorly suited detect this signature of selection.
2. **Conducting FST outlier scans requires binning individuals into *a priori* defined, discrete populations so that FST can be estimated for each locus**:
   1. **Insufficient power:** If we bin samples into populations according to their smallest scale sampling location (e.g. North California or Japan), sample size is too low to detect population structure given the Ne and the number of markers. Vaux et al state that “the only prevailing genetic pattern across all three loci datasets was the separation between the North and South Pacific” based on the absence of clustering in the PCA, and therefore only conduct outlier scans using the North and South Pacific dataset, which includes all individuals sampled in the North Pacific as one population and all individuals sampled in the South Pacific as a second. I agree that this conservative approach was appropriate for deciding how to parse individuals into groups for FST outlier scans, but it is important to remember that the detection structure using PCA is effectively impossible when FST < , where *n* is sample size and *p* is number of SNPs, and sample size is equal across populations (Patterson *et al.* 2006). So, given the size of sample groups, the number of markers in the ascertainment dataset and Ne, it is unlikely PCA will detect structure at sub-gyre levels, even if it is present. As an example, if we assume a modestly large to very large range of Ne (5k - 160k) and (generously) the two largest sampling locations (SC and OR), PCA will fail to detect structure with 12,872 markers between these sampling locations even at modest migration rates, and certainly below the level of dispersal needed for demographic coupling (Figure 1 below).

Chart, line chart

Description automatically generated

**Figure 1:** FST vs migration rate across a range of large effective population sizes. FST is calculated under the island model (FST = 1 / (1 + Ne\*m)). The red dashed line represents the detection limit for population structure using 12,872 markers and 62 individuals evenly split between two populations according to FST\_limit = 1 / sqrt(n x p), where n is sample size and p is the number of markers. The blue dashed line is the North vs South Pacific FST for reference.

* 1. **Clines:** Binning into discrete groups, in general, also presents challenges and can limit the power of outlier approaches to detect adaptive variants underlying adaptation to environmental gradients that commonly drive selection in the marine environment (e.g. clines) (Dayan 2018; De Mita *et al.* 2013; Jones *et al.* 2013).

Given these challenges, our goal during feature selection should be to find a method to identify putatively adaptive genetic structure that exhibits greater power for detecting subtle allele frequency shifts and does not depend on *a priori* defined, discrete population structure. The desired approach should be complementary to the FST outlier approach. There are many approaches that fit these requirements, but we chose redundancy analysis using spatial eigenvectors that decompose spatial variation at multiple scales among our sampling locations as explanatory variables.

## Spatial Outliers - Methods and Results

Redundancy analysis (RDA) is a multivariate constrained ordination that finds linear combinations of multiple explanatory variables that are redundant with linear combinations of the response variable (Legendre & Legendre 2012). We used redundancy analyses of components (cRDA), followed by a spatial outlier identification using a second RDA directly on SNP data (Capblancq & Forester 2021; Forester *et al.* 2018). The benefit of this two stage approach, is that the cRDA is less prone to overfitting. If the cRDA discovers population substructure, then we can be encouraged to continue to SNP level analyses and identify outliers. For the cRDA, we used principal components of genetic variation in the North vs South Pacific dataset as the response variable, and one set of spatial explanatory variables: distance-based Moran’s eigenvector maps (dbMEMs). We excluded individuals from Baja California because of the high level of missing data and small sample size (n = 4). dbMEMs are capable of describing spatial variation at multiple scales, including spatial autocorrelation as well as local structures, and are suitable as explanatory variables for constrained ordinations (Borcard & Legendre 2002). dbMEMs were calculated using a distance matrix of geodesic distances between approximate sampling locations using the r package *SoDA*. To calculate the dbmems from the distance matrix we used the *dbmem* function of the R package *adespatial*. We retained all dbMEMs with positive values of Moran’s I to capture the effect of positive spatial autocorrelation. We conducted the cRDA using the R package *vegan* using the *rda* command. After fitting the global model and testing for significance using permutation, variable selection was performed used the *ordistep* and *ordiR2step* functions and by forward variable selection using the adjusted cumulated R2 (*forward.sel* function). We then tested the significance of each constrained axis and each explanatory variable after conditioning the model on all other variables by permutation using the *anova.cca* function.

There were four dbMEMs with positive Moran’s I. These dbMEMs describe spatial autocorrelation at 4 orthogonal scales. For example MEM1 captures autocorrelation and the greatest spatial scale such that the two South Pacific sampling locations fall at one extreme, while all eastern North Pacific samples fall at the other extreme. The finest scale dbMEM (MEM4) describes spatial autocorrelation between the relatively densely samples US West coast. The values of the dbMEMs for each of the sampling locations are demonstrated in figure Map

Description automatically generatedMap

Description automatically generated with medium confidenceMap

Description automatically generatedMap

Description automatically generated2.

**Figure 2:** dbMEM values for MEM1 to MEM4 among the 11 sampling locations.

The global cRDA model was significant (p = 0.001, 1000 permutations). All model selection procedures suggested retaining MEM1 and MEM2 and excluding MEM4, but there was disagreement on MEM3. We retained MEMs 1 - 3 for the final model. Only the first two RDA axes were significant (i.e. explain significantly more than the equivalent unconstrained axes, Holms adjusted p value < 0.05, Table 3). All explanatory variables (dbMEMs) in the final model are also significant (i.e. significantly greater explanation of variance for the variable, once the data is conditioned on all other variables, Holms adjusted p value < 0.05, Table 4)

|  |  |  |
| --- | --- | --- |
| **Constrained Axis** | **Portion Variance Explained** | **Adjusted P-Value** |
| RDA1 | 0.70% | 0.003\*\* |
| RDA2 | 0.35% | 0.016\* |
| RDA3 | 0.33% | 0.702 |

**Table 3:** Portion of variance explained by each of the redundant axes, and Holm’s adjusted p-value

|  |  |  |
| --- | --- | --- |
| **Explanatory Variable** | **Portion Variance Explained** | **Adjusted P-Value** |
| dbMEM1 | 0.60% | 0.003 |
| dbMEM2 | 0.40% | 0.003 |
| dbMEM3 | 0.30% | 0.012 |

**Table 4:** Portion of variance explained by each of the explanatory variables, and Holm’s adjusted p-value

Collectively these results suggest that there is a significant relationship between genetic variation and spatial autocorrelation (although note this model did not attempt to parse population structure from autocorrelation due to IBD *per se*). The first two constrained axes captured ~ 1% of the total variance in the PCA. This is a large value considering that FST between North and South Pacific samples is only 0.0056.

Diagram

Description automatically generated

**Figure 4:** RDA biplot showing individual scores along the first two redundant axes using type I scaling. Points (individuals) are colored by sampling location. Perpendicular projections of individual points onto black arrows represent approximate spatial variable values for a given individual.

When we examine the spatial structures in the RDA biplot (Figure 4), we can see that MEMs 1 and 2 load strongly onto the first redundant axis, and this axis separates North from South Pacific populations. The second RDA is driven by MEMs 2 and 3 and therefore captures spatial autocorrelation within the North Pacific.

There are some salient features of the RDA that are worth highlighting:

* The total variance constrained by two redundant axes is ~ 1%. The pairwise FST large between North and South Pacific samples in Vaux et al was 0.0056, and the variation explained by the first PC in the Vaux et al PCA which largely separated North from South Pacific samples was 0.8. This suggests that we can identify more structure in the dataset using spatially explicit methods than the first axis of unconstrained ordination, or binning individuals *a priori* into North vs South Pacific populations.
* The global model provides a much better fit to the data than random permutation, and both RDAs are also statistically significant, suggesting that the identification of spatial structure beyond North *vs* South populations is unlikely due to overfitting. Though we did not conduct cross validation, so overfitting remains a possibility.
* The first RDA captures a similar amount of variance in the dataset (0.7%) due to North – South structure (primarily MEM1) as pairwise FST between North and South Pacific samples in Vaux et al (0.0056).
* The second RDA captures about half the variance of the first (0.35%), and suggests tht there is indeed spatial structure within the North Pacific and that the scale of this structure is about half of that between the North and South Pacific.

We fit the RDA on the SNP data using the same methods for the cRDA. To identify spatial outliers, we examined the distribution of SNP loadings on the significant redundant axes and considered any SNP with a z-score greater than three an outlier. The results of the model fit and significance testing procedures are nearly identical to the cRDA, so we do not present them. We retained 69 along RDA1, and 104 along RDA2 (Supplemental Table 1). Note that 59 of the 69 RDA axis 1 SNPs were also FST outliers, and 3 were also RDA axis 2 SNPs. Zero of the 104 RDA axis 1 SNPs overlapped with FST. Therefore spatial analysis contributed 111 (111 = 69 + 104 - 59 - 3 ) additional SNPs to feature selection.

*Neutral SNPs*

305 SNPs were drawn at random from the MultiSNP dataset (Supplemental Table 3). The rationale is described in detail in the description of the MultiSNP dataset, but briefly, neutral SNPs attempt to capture genome wide variation and should be more suitable for estimating population genetic parameters that may be biased using other sets of SNPs. For example, data from neutral SNPs could be filtered to match other studies and make parameters like Tajima’s D or FST directly comparable. Or data from these SNPs could be used to model demographic history using dadi.

After initial selection of the 305 neutral SNPs, one was removed because it was a FST outlier, and four were removed because they fell on the same RADtag as a second random SNP. A Kolmogorov-Smirnov test (p = 0.26) was used to check that the allele frequency spectrum of the neutral SNPs did not significantly differ from the allele frequency spectrum of the MultiSNP dataset.

*Summary Statistics*

In total there were 495 SNPs selected for primer design: 84 FST outliers, 170 spatial outliers (111 unique), and 300 neutral, supplemental table . Before committing to this set of SNPs, we examined their depths and SFS.

Chart

Description automatically generated

**Figure 3:** Mean depth per locus per individual in the ascertainment dataset for FST outliers, Spatial SNPs, and Neutral SNPs following feature selection (i.e. before primer design – does not reflect the final markers), compared the mean depth per locus per individual among remaining SNPs in the MultiSNP dataset (not\_panel).

Differences in depth distribution between different sets of SNP were apparent (figure 3). SNPs drawn from the multiSNP dataset (neutral SNPs) had lower average depth, than those drawn from the more stringently filter North and South pacific dataset (FST and spatial outliers). However modal and mean depths were high (>50) in all SNP subsets.

There were striking differences in the SFS of different SNP subsets (figure 4), with FST and spatial outliers biased towards more common SNPs, and the neutral SNPs closely following the distribution of the multiSNP dataset.

Chart, line chart

Description automatically generated

**Figure 4:** Folded SFS for different SNP subsets. See figure 3 for details of SNP subsets.

# Primer Design

## Overview

A total of 495 SNPs were selected for primer design. RADtag sequences from Vaux et al (2021) containing the selected features as well as flanking sequence from the 2013 Pacific Bluefin assembly (RefSeq: GCA\_000418415.1) were used to design primers. After several rounds of optimization, primer sets meeting the physical properties for GTseq and filtered for expected performance in multiplex PCR were designed (Nmarkers = 332). Primers for 15 neutral markers with lengths greater than 60bp were excluded to reduce primer costs (Nmarkers = 317).

## Consensus Sequence and Variant Masking

Following feature selection, consensus sequences were inferred for genomic region surrounding the 495 target SNPs using the RADtag catalog from *Vaux et al* (2021). Nucleotide diversity is very high and there are many SNPs on every RADtag. This led to two challenges during consensus sequence inference. First, most FST outliers are very close to the edge of the RADtag because there are many SNPs on a given RADtag and the --write-first-snp used during filtering of the North and South Pacific dataset. Consequently, an initial primer design attempt failed at the majority of FST outliers because there was insufficient flanking sequence around the target SNP to design primers. Second, there is a great deal of variation that might fall under the regions covered by primer and probe sequences and lead to allele dropout if not appropriately masked during consensus sequence inference. To solve these challenges, in addition to the target SNP, any SNP within the RADtag with a MAF > 1% was masked using the IUPAC ambiguity code, and flanking sequence was drawn from the 2013 Pacific Bluefin assembly (RefSeq: GCA\_000418415.1), when needed to design primers. The 1% MAF cutoff for non-target SNPs was chosen after multiple rounds of attempted primer design to balance the tradeoff between allele dropout caused by unaccounted for variation and constraints placed on successful primer design when attempting to account for too many non-target SNPs.

The final FASTA file of variant masked consensus sequences is provided in file albacore\_primer\_design\_input\_1.1.fa in the supplemental file directory.

## Primer Design

Target capture primers were designed by GTseek LLC using a proprietary pipeline to avoid the formation of primer artifacts during multiplex PCR. In brief, multiple primer sets are designed for each locus surrounding the target variation such that the melt temperatures are between 58 C and 62 C. Once all primer sets are designed, each primer is evaluated for possible primer hetero-dimer formation with any other primers that have passed. When both primers are not predicted to bind to any other primers already added to the multiplex PCR set, the locus primers are “passed” and added to the list. The designed primer sets then have Illumina R1 and R2 sequencing primer sites tagged on to the target-specific sequence and the primers are tested for the formation of stable hairpin structures that can interfere with efficient PCR. Lastly, the amplicon sequences for all loci are evaluated for having potential nested primer sites matching the 3’ ends of any designed primers to avoid concatamer artifacts in the sequencing data. In order to retain as many of the FST outliers as possible, these markers were preferentially retained when our multiplex PCR primer design algorithm found a conflict with another designed primer.

Of the initial 495 target markers, 332passed evaluation. This 332 marker panel and associated data are all labeled **albacore\_gtseq332** in this directory/repository. This includes the final set of target sequences that passed primer design (albacore\_gtseq332\_targets.fa), primer sequences (Supplemental Table 5), and probe sequence file (Supplemental Table 6).

There were 25 markers in the albacore\_gtseq332 panel with primer sequences longer the 60bp. To reduce oligonucleotide synthesis costs, we retained the 2 FST outlier and 8 spatial outlier markers with primers longer than 60bp, but excluded the 15 neutral markers with long primers (Nmarkers = 317). This 317 marker panel and all associated data are labeled **Panel\_1.0** in this directory/repository. This includes a combined panel information table (Supplemental Table 7).

# Panel Validation 1

The first validation library (validation\_library\_1) was sequenced on an Illumina NextSeq500 instrument with 75bp single end chemistry. All associated metadata, library preparation and validation/analysis files can be found in the validation\_library\_1 directory.

Validation\_library\_1 was sequenced on an Illumina NextSeq500 instrument with 75bp single end chemistry (see supplemental file “019 SFGL Illumina OmySFCrook TalTest 2021-05-17.docx”). Validation\_library\_1 used panel\_1.0 which included 317 primer sets, and 148 unique albacore individuals, 138 of which were previously genotyped in Vaux *et al* (2021). The total number of reads in the from the sequencing run was 294,883,769. The total number of reads with an exact match to one of the 317 forward primer sequences was 8,529,829. The total number of reads with both an exact match to the forward primer sequence and one of the two probe sequences was 2,808,094, for an on-target rate of 32%.

Initial analysis of validation\_library\_1 included application of two GTseek LLC scripts to the raw sequence data: SeqTest and Primer-Interaction. The SeqTest script uses a hash file of raw reads and performs and exact matching search for primer sequences, probe sequences and both primer sequences and probe sequences together in the same read. Results of the SeqTest analysis (Supplemental Table 8) revealed 15 markers with no reads, 9 markers with zero probe sequence counts, and 43 markers with poor (<10%) on-target rates. The results from the Primer-Interaction scripts are provided in Supplemental Table 9.

To improve on-target rates, probe sequences were manually revised by GTseek LLC to account for variation within the regions covered by the probe sequences, resulting in revised probe sequence file (Supplemental Table 10). After adjusting probe sequences, 28 markers were identified that contributed most to off-target sequences (Supplemental Table 11). These markers were excluded from the next primer pool used to amplify validation library 2.

# Validation Library 2

Identification of proper amplification from short (e.g. 75bp) single end sequencing data relies on reads containing both the FWD primer sequences and an allele specific probe sequence. Given the high frequency of variation within the genomic regions covered by the probe sequences, it was difficult to discern off-target amplification from probe sequence variation. To confront this challenge, GTseek requested a second validation library using paired-end data. The paired end data could be used to detect properly paired FWD and REV primers, and use this information as a proxy for on-target reads, thereby allowing marker validation and probe sequence revision without perfect *a priori* knowledge of variation within the probe regions.

The second validation library (validation\_library\_2) was sequenced on an Illumina Miseq v3 instrument with 75bp paired end chemistry (see supplemental file 023 SFGL Illumina TalTest2 2021-07-27.docx). All associated metadata, library preparation and validation/analysis files can be found in the validation\_library\_2 directory. Validation\_library\_2 used a pool of 289 primers for amplification (Supplemental Table 12), and 150 unique albacore individuals, none of which were previously genotyped in Vaux *et al* (2021). The library included 384 unique i7-i5 combinations including replicates and controls. The probe sequence file used to call genotypes is available in Supplemental Table 13. To accommodate variation in amplification success among individuals, the 100 best i7-i5 index combinations were used (Supplemental Table 14). Note that this set of 100 includes many replicates and 2 positive controls, so it is composed of fewer than 100 unique individuals.

The validation library 2 primer analysis is presented in Supplemental Table 15. Unlike the primer analysis conducted for validation library 1, this analysis considers an on-target read as any read with properly paired forward and reverse primer sequences. The Primer-Interaction script was also run using data from the 100 best performing i7-i5 index combinations and results are presented in Supplemental Table 16. Together these analyses suggest ~237 markers are performing adequately. Allele correction values for 4 markers were also determined from this validation library.

The on-target rate for the validation library 2 was 61%. This is an excellent on-target rate. Therefore we chose to not move forward with further optimization of the panel and retained all 289 markers in the pool for validation library 2, including those that had a low amplification rate (Supplemental Table 15) and those that are generating off-target sequence due to primer interactions (Supplemental Table 16).

# Final Panel

# The final panel includes 289 markers (Table 5). Full panel information including primer and probe sequences, alleles, minor allele frequency in the validation library, mapping positions and identity as FST outlier, spatial outlier on RDA axis 1 or axis 2, or neutral marker is provided in the file albacore\_gtseq\_panel\_289 in this repository. The same information is also available on the SFGL GTseq Github repository.

|  |  |
| --- | --- |
| **Marker Type** | **n** |
| FST Outlier | 54 |
| Spatial Outlier | 117\* (81) |
| Neutral | 153 |
| Total | 289 |

**Table 5:** Marker Counts in the final panel (Panel\_289). \*Note that there are 117 spatial outliers, but 36 are also FST outliers

**References**

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling* **153**, 51-68.

Capblancq T, Forester BR (2021) Redundancy analysis: A swiss army knife for landscape genomics. *Methods in Ecology and Evolution* **12**, 2298-2309.

Dayan DI (2018) Clinal adaptation in the marine environment.

De la Cruz O, Raska P (2014) Population structure at different minor allele frequency levels. *BMC Proceedings* **8**, S55.

De Mita S, Thuillet AC, Gay L*, et al.* (2013) Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol Ecol* **22**, 1383-1399.

Forester BR, Lasky JR, Wagner HH, Urban DL (2018) Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Mol Ecol* **27**, 2215-2233.

Gagnaire P-A, Gaggiotti OE (2016) Detecting polygenic selection in marine populations by combining population genomics and quantitative genetics approaches. *Current Zoology* **62**, 603-616.

Gagnaire PA, Broquet T, Aurelle D*, et al.* (2015) Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evol Appl* **8**, 769-786.

Hermisson J, Pennings PS (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335-2352.

Höllinger I, Pennings PS, Hermisson J (2019) Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS genetics* **15**, e1008035.

Jones MR, Forester BR, Teufel AI*, et al.* (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution* **67**, 3455-3468.

Legendre P, Legendre LF (2012) *Numerical ecology* Elsevier.

Linck E, Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol Ecol Resour* **19**, 639-647.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS genetics* **2**, e190.

Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Mol Ecol* **25**, 2144-2164.

Waples RS, Punt AE, Cope JM (2008) Integrating genetic data into management of marine resources: How can we do it better? *Fish and Fisheries* **9**, 423-449.

Yeaman S (2015) Local adaptation by alleles of small effect. *Am Nat* **186 Suppl 1**, S74-89.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution* **65**, 1897-1911.