FEATURE

FishGen.net: An Online Genetic Repository for Salmon and Steelhead Genetic Baselines

Jesse McCane | Pacific States Marine Fisheries Commission, Idaho Department of Fish and Game, Eagle Fish Genetics Lab, 1800 Trout Rd., Eagle, ID 83616. E-mail: jesse.mccane@idfg.idaho.gov

Chris Adam, Bill Fleming, and Mark Bricker | Resource Data Incorporated, Boise, ID

Matthew R. Campbell | Idaho Department of Fish and Game, Eagle Fish Genetics Lab, Eagle, ID

Photo credit: Steve Martarano/USFWS.

DOI: 10.1002/fsh.10105

FishGen is a final repository for Pacific salmon *Oncorhynchus* spp. and steelhead *O. mykiss* genetic data generated as part of the genetic stock identification and parentage-based tagging projects in the Columbia River basin and throughout the Pacific Coast of North America. Resource Data, Inc., developed this web-based, GIS-interfaced software, which is freely available to the public, with funding from the Pacific Coastal Salmon Recovery Fund and Bonneville Power Administration. FishGen currently houses genetic stock identification baselines for both Chinook Salmon *O. tshawytscha* and steelhead in the Columbia and Snake river basins, as well as hatchery, parentage-based, tagging baselines for both species in the Snake River basin. Because it has a user-friendly interface and protocol for submitting and storing standardized genetic and sample metadata, it is an excellent tool for supporting genetic research and monitoring projects throughout the region.

BACKGROUND

Every year federal, state, tribal, and university laboratories generate hundreds of thousands of genotypes (genetic data) used to construct and maintain genetic baselines for Pacific salmon *Oncorhynchus* spp. and steelhead *O. mykiss* along the Pacific Coast of the United States and Canada (IEAB 2013). A genetic baseline is a sample of fish from a specific population or populations that allows one to assign individuals from a mixed stock to their population of origin based on the observed genetic structure or genetic relationships. One of the most common applications of genetic baselines is for genetic stock identification (GSI) to delineate and assess mixed-stock fisheries. To help manage commercial, recreational, and tribal harvest allocations, GSI baselines have been developed for every species of Pacific salmon across the Pacific Coast: Chinook Salmon *O. tshawytscha* (Satterthwaite et al. 2015), Sockeye Salmon *O. nerka* (Gilk-Baumer et al. 2015), Pink Salmon *O. gorbuscha* (Araujo et al. 2014), Chum Salmon *O. keta* (Beacham et al. 2009), and Coho Salmon *O. kisutch* (Van Doornik et al. 2007).

More recently, GSI baselines have also been used for viability status assessments of salmon and steelhead listed on the U.S. Endangered Species Act (ESA) by contributing to a better understanding of population genetic structure (Narum et al. 2010; Hess and Matala 2014), estimating abundance and productivity (Hamazaki and DeCovich 2014; Hess et al. 2014; Bellinger et al. 2015), and estimating genetic diversity and life history characteristics of individual stocks (Van Doornik et al. 2011; Campbell et al. 2012).

While stock-level genetic baselines have a long history of use by fish researchers and managers, genetic baselines are increasingly being used to implement parentage-based evaluations of the reproductive success of hatchery fish compared to that of wild fish (Christie et al. 2014), estimate spawner abundance (Rawding et al. 2014) and manage conservation broodstocks (O'Reilly and Kozfkay 2014). In addition, with recent advances in both time and cost efficiency of genotyping technologies and statistical software for handling large data sets, managers are beginning to use parentage analyses on a large scale for the purpose of tagging hatchery stocks. For example, since 2008, the Idaho Department of Fish and Game (IDFG) and the Columbia River Inter-Tribal Fish Commission have implemented a parentage-based tagging (PBT) program in the Snake River basin (Steele et al. 2013a, 2013b). All steelhead and Chinook Salmon hatchery broodstocks in the Snake River basin are sampled to obtain tissue and genotyped (Steele et al. 2013b). This results in a baseline of about 10,000 adult hatchery spawning spring and summer Chinook Salmon and about 5,000 adult steelhead being sampled each year, which allows users to assign any of their offspring back to a specific parent pair using parentage analysis.

While most fish genetics labs have databases to store and manage genetic data produced in their own lab, none are suitable to act as a long-term, shared, data repository. For example, the IDFG genetics lab relies on Progeny software, which is critical for securely cataloging samples and managing the genetic data that is produced. However, Progeny software is neither web-based nor GIS-enabled and could not function as a final data repository that can easily be accessed and shared by multiple agencies and institutions. Most data sharing among laboratories is presently accomplished via methods such as manually cutting and pasting information to and from text files and spreadsheets. Despite the extreme care taken by individuals performing such work, with large amounts of data errors are unavoidable.

In recent years, international, multilaboratory projects have been completed to construct standardized genetic baselines for Chinook Salmon (Genetic Analysis of Pacific Salmonids [GAPS]; Seeb et al. 2007) and steelhead (Stevan Phelps Allele Nomenclature [SPAN]; Stephenson et al. 2009). These efforts demonstrate the conservation and management benefits of merging regional data sets from multiple laboratories, the results of which created the first range-wide genetic baselines for these species. However, these baselines were specific to one type of genetic marker (microsatellites), and the construction of a dynamic, evolving database was outside the scope of these projects. In addition, other genetic repositories such as the Dryad Digital Repository (datadryad.org) lack marker validation procedures and GIS interfaces. Accordingly, agencies across the Columbia River basin and the Pacific Coast partnered to create just such a tool, FishGen, to ensure long-term accessibility and security of the growing volume of genetic data being generated in this region and to facilitate collaboration among the contributors.

CONSTRUCTION AND CONTENT

The partners and development team identified a series of design priorities to address the limitations of current data management systems and to help solicit and guide technical support to create FishGen. These included the need for a web-based platform and the ability to evolve over time to accept new types of genetic data. The initial design process began once funding was secured from the Pacific Coastal Salmon Recovery Fund and the Bonneville Power Administration. After a bidding process, Resource Data, Inc., (RDI) was chosen to lead the construction of FishGen because of their extensive experience with software development, system integration, and GIS, including work on many fisheries-related projects: AKFIN (akfin.org), PacFin (pacfin.psmfc.org), and eLandings (elandings.alaska.gov).

Throughout the design process, a large amount of feedback was requested and received from various genetics laboratories pertaining to field requirements and validation, marker standardization, and general ease of use features. The various laboratories that were consulted in regards to FishGen design included the U.S. Fish and Wildlife Service Abernathy

Fish Technology Center, the Columbia River Inter-Tribal Fisheries Commission Hagerman Laboratory, the University of Washington Seeb Lab, and the National Oceanic and Atmospheric Administration Northwest Fisheries Science Center and Southwest Fisheries Science Center. We also carefully reviewed the GAPS and SPAN databases for ideas on marker standardization as well as ways to improve on these previous data repositories.

FishGen was built as a web application using Microsoft technologies. The front end of the site is built using ASP.Net with a combination of Web Forms and AJAX with jQuery hosted on Internet Information Services. The back end of the system, used for processing file uploads, is a Windows service. The system is built on SQL Server 2014 and Windows Server 2012. FishGen is currently hosted at Amazon Web Services (AWS) on an M3 large instance. The database is fully backed up on a nightly basis; the previous five backups are archived on the server, and additional backups are kept in Amazon S3 storage for recovery purposes. This results in several redundant means of maintaining, protecting, and archiving the data saved in FishGen to prevent data loss in the event of any unforeseen circumstances.

To initiate the flow of data in to FishGen, the user uploads definitions of their genetic markers and subsequently uploads individual fish and associated metadata and genotypes. A genetic marker is a location on an individual's genome that can be interrogated and used for various analytical purposes. FishGen supports two types of genetic markers—microsatellites and single nucleotide polymorphisms (SNPs)—with the flexibility to support new types of genetic markers in the future. A SNP (the primary type of genetic data housed in FishGen) refers to a single, variable base pair at a specific location on an individual's genome. Complex analyses, such as genetic stock identification and parentage-based tagging, are possible when panels of hundreds of SNPs are genotyped in tandem. Within SNPs, three definition subtypes are supported: TaqMan, GTSeq, and RAD. These subtypes delineate the various laboratory methods used to interrogate the SNPs within an individual's genome. A problem that has arisen in the past with genetic repositories is that of marker standardization. With FishGen, we have used varying levels of marker definition validation to help ensure marker standardization within the data set. These include comparing primer and probe sequences of markers being uploaded to markers already in FishGen to ensure that the same marker cannot be uploaded under two different marker names. Currently FishGen only looks for an exact match when comparing primer and probe sequences, but in the future, we plan on making these algorithms more robust to find matches with a threshold for percent similarity of sequences. The marker name field is also validated to ensure that two different markers cannot be uploaded under the same name. FishGen supports "marker synonyms," which allows laboratories to upload additional marker definitions (as well as lab-specific marker aliases) for loci already defined in FishGen. FishGen allows for the creation of marker sets, which allow users to create custom groups of markers that may be project- or species-specific. These marker sets can then be exported from FishGen, along with all included marker fields such as primer, probe, and allele information.

Once marker definitions have been uploaded into FishGen, the user can upload individual fish information along with metadata and genotypes. In FishGen, individuals are grouped into collections; a collection represents a group of individual fish, sampled on a single day or over a range of dates, from a specific body of water or hatchery. Thus, in FishGen there are collection-level fields such as latitude, longitude, collection method, body of water, and sample year. These are fields that apply to the collection itself and, by extension, all individuals within the collection. There are also individual-level fields such as phenotypic sex, phenotypic species, marks, PIT tag number, and sample date. These are fields that apply to individuals within a collection. A user can upload collections, individuals, and genotypes in a single tab-delimited text file set up in a one-row-per-individual format. Column headers and column ordering is flexible and can be customized at time of upload. There are currently no limitations in terms of the maximum amount of data users can upload either in a single upload or over the course of their use of FishGen.

There are a few different ways to find and extract data from FishGen once it has been uploaded. One method is to use the map search screen. When collections and individuals are uploaded to FishGen, a latitude and longitude for each collection is required (even if they are only approximated). This allows FishGen to automatically calculate a number of GIS-based fields (including state, hydrologic unit codes such as HUC4, HUC6, HUC8, etc.) and visually plot the collection in a map display (Figure 1). On the map search screen, a user can select search criteria from a number of fields in order to search through all the collections in the database. Once search results are returned, the user can choose to display the search results on the map, as well as sort through the results in tabulated form and perform additional searching and filtering on all collection-level fields. Once the user has narrowed down the results to the collections they wish to export, they can download these data from FishGen in either GenePop format or the original format of one row per individual, along with all uploaded genetic data for the individuals in those collections.

The other primary method of exporting data from FishGen is through the use of the saved data set feature. A saved data set can be created by three methods. The first method is at the time of initial upload; the uploading user can select an option to create a saved data set consisting of all the collections and individuals they are currently uploading. The second method is by searching and filtering on the map search screen until the results have been narrowed down to a group of collections with which the user would like to create a saved data set. The final method is by simply providing FishGen with a text file of collection and individual names that have already been uploaded to FishGen, from which the user can create a saved data set. When a saved data set is created, the user can name the saved data set and must also select a relevant marker set for that saved data set. Once a saved data set has been created, the user is provided with a large, formatted text box in which any other notes or annotations can be added that the user may consider useful for other users looking to download and use that saved data set. This allows the user creating the saved data set to provide any extra information that could not be properly captured in the various collection- or individual-level fields. Saved data sets can be searched through and filtered, and when exported they will provide an immutable snapshot of the collection, individual, and genetic data as it was when the saved data set was created. The saved data set feature is not only invaluable to facilitating the easy sharing of data between laboratories, but also extremely useful for referencing data sets found in the literature and publications.

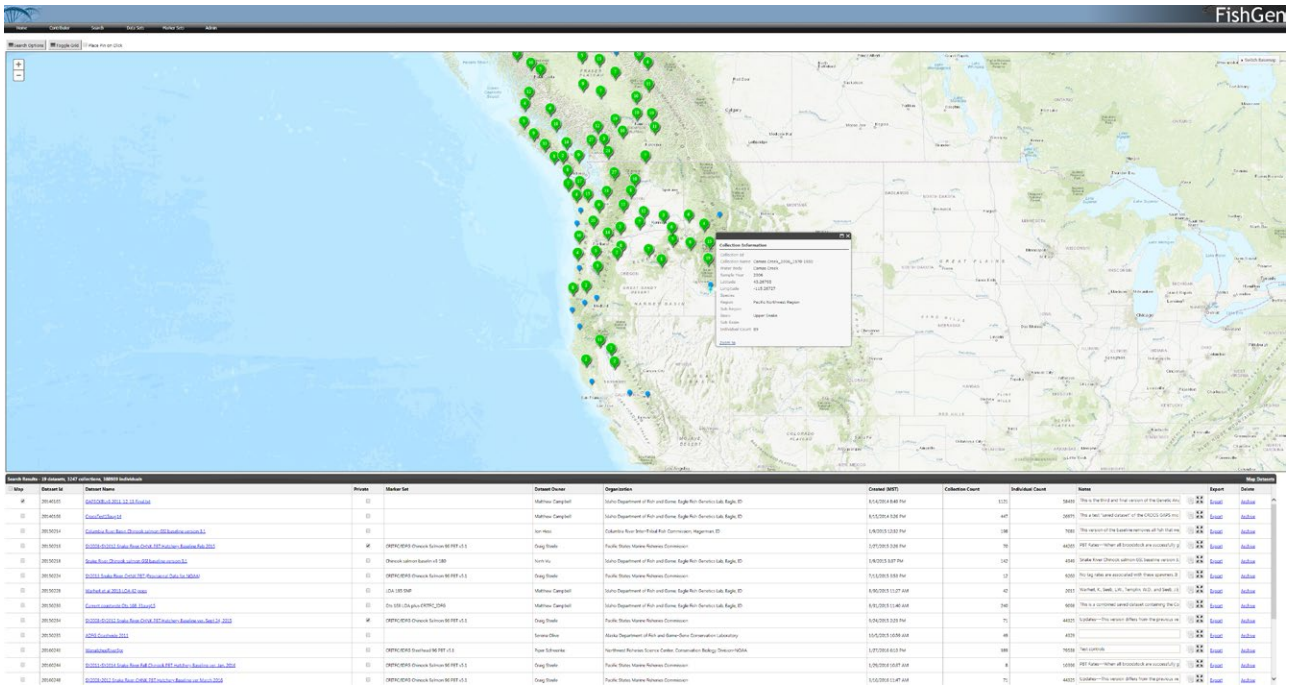

Figure 1. FishGen.net search screen displaying mapped search results. The search options button on the top left allows the user to view a number of different criteria by which they can search and filter. The map view then plots and clusters the search results. Collections within close proximity of each other are clustered together but can be broken out by clicking or zooming in closer on the map. Tabular search results are displayed at the bottom of the screen; users can further filter and refine their search results using the table layout. FishGen features eight different map base layers, including one for HUC designations.

DISCUSSION AND UTILITY

FishGen currently houses over 20 saved data sets created by users from six various agencies and organizations. Although FishGen has primarily been used for steelhead and Chinook Salmon projects, saved data sets now also exist for Burbot *Lota lota*, Brook Trout *Salvelinus fontinalis*, and Bull Trout *S. confluentus*. Some of the larger data sets include the GAPS Chinook baseline (58,439 individuals across 1,121 collections; Seeb et al. 2007), the spawn year (SY) 2008–2012 Snake River steelhead PBT hatchery baseline (37,155 individuals across 67 collections; Steele et al. 2015), the SY2008–SY2012 Snake River Chinook PBT hatchery baseline (44,325 individuals across 71 collections; Steele et al. 2015), and the Columbia River basin steelhead GSI baseline (9,991 individuals across 239 collections; Hess et al. 2016). Baselines stored on FishGen are already contributing to research and monitoring programs throughout the Columbia River basin. For example, Hess et al. (2016) used a Columbia River basin GSI baseline for steelhead (FishGen Dataset identification number [ID] 20150220) and a Snake River steelhead PBT baseline (FishGen Dataset ID 20150221) in conjunction, to estimate the stock-specific abundance and run timing of wild and hatchery steelhead returning to the Columbia River basin during three consecutive migration years (2012–2014). In addition, Hinrichsen et al. (2016) used a Snake River Chinook PBT hatchery baseline (FishGen Dataset ID 20160248) to effectively estimate the proportion of hatchery-origin fish on spawning grounds in the South Fork Salmon River.

Our vision for FishGen is that of a long-term, relatively curator-free, final genetic repository for various fish genetic projects throughout the entire Pacific Coast of the United States and Canada. Although we had initially envisioned FishGen as a database primarily for steelhead and Chinook Salmon as part of the PBT and GSI projects in the Pacific Northwest, it quickly became apparent that FishGen was versatile and robust enough to house various projects of a much larger scope. Therefore, we expanded the intended use of FishGen to include a wider breadth of projects, such as studies of hybridization or introgression, for a much larger pool of species. We have also expanded the ability of FishGen to support multiple new genetic marker subtypes over the past few years, and plans to support new marker types (such as "haplotypes" or "Kaspar SNPs") are currently being developed. Our hope is that biologists, managers, and geneticists will continue to use FishGen as a one-stop genetic repository for a wide variety of projects throughout the Pacific Coast.

ACKNOWLEDGMENTS

Primary funding for the development of FishGen was provided by the Pacific Coastal Salmon Recovery Fund (webapps.nwfsc.noaa.gov), facilitated through the Idaho Governor's Office of Species Conservation (species.idaho.gov/index.html). Funding for management and hosting costs have been provided by the Bonneville Power Administration. Christian Smith, Brian Leth, Carl Steifel, Mike Ackerman, Craig Steele, Nate Campbell, Shawn Narum, and Jeff Stephenson provided recommendations and ideas on database functionality and design. Other RDI staff including Jennifer Bedient, Will Jensen, and Tiffany Rasmussen, assisted with edits and critiques. There is no conflict of interest declared in this article.

REFERENCES

Araujo, H. A., J. R. Candy, T. D. Beacham, B. White, and C. Wallace. 2014. Advantages and challenges of genetic stock identification in fish stocks with low genetic resolution. Transactions of the American Fisheries Society 143:479–488.

Beacham, T. D., J. R. Candy, S. Sato, S. Urawa, K. D. Le, and M. Wetklo. 2009. Stock origins of Chum Salmon (*Oncorhynchus keta*) in the Gulf of Alaska during winter as estimated with microsatellites. North Pacific Anadromous Fish Commission Bulletin 5:15–23.

Bellinger, M. R., M. A. Banks, S. J. Bates, E. D. Crandall, J. C. Garza, G. Sylvia, and P. W. Lawson. 2015. Geo-referenced, abundance calibrated ocean distribution of Chinook Salmon (*Oncorhynchus tshawytscha*) stocks across the West Coast of North America. PLoS ONE 10(7):e0131276.

Campbell, M. R., C. C. Kozfkay, T. Copeland, W. C. Schrader, M. W. Ackerman, and S. R. Narum. 2012. Estimating abundance and life history characteristics of threatened wild Snake River steelhead stocks by using genetic stock identification. Transactions of the American Fisheries Society 141:1310–1327.

Christie, M. R., M. J. Ford, and M. S. Blouin. 2014. On the reproductive success of early-generation hatchery fish in the wild. Evolutionary Applications 7:883–896.

Gilk-Baumer, S. E., S. D. Rogers Olive, D. K. Harris, S. C. Heinl, E. K. C. Fox, and W. D. Templin. 2015. Genetic mixed stock analysis of Sockeye Salmon harvests in selected northern Chatham Strait commercial fisheries, Southeast Alaska, 2012–2014. Alaska Department of Fish and Game, Fishery Data Series 15-03, Anchorage.

Hamazaki, T., and N. DeCovich. 2014. Application of the genetic mark–recapture technique for run size estimation of Yukon River Chinook Salmon. North American Journal of Fisheries Management 34:276–286.

Hess, J. E., M. W. Ackerman, J. K. Fryer, D. J. Hasselman, C. A. Steele, J. J. Stephenson, J. M. Whiteaker, and S. R. Narum. 2016. Differential adult migration-timing and stock-specific abundance of steelhead in mixed stock assemblages. ICES (International Council for the Exploration of the Sea) Journal of Marine Science 73:2606–2615.

Hess, J. E., and A. P. Matala. 2014. Archival genetic analysis suggests recent immigration has altered a population of Chinook Salmon in an unsupplemented wilderness area. Conservation Genetics 15:387–403.

Hess, J. E., J. M. Whiteaker, J. K. Fryer, and S. R. Narum. 2014. Monitoring stock-specific abundance, run timing, and straying of Chinook Salmon in the Columbia River using genetic stock identification (GSI). North American Journal of Fisheries Management 34:184–201.

Hinrichsen, R. A., C. A. Steele, M. W. Ackerman, M. R. Campbell, S. R. Narum, M. A. Hess, W. P. Young, B. A. Shields, and B. L. Maschhoff. 2016. Maximum likelihood estimation of the proportion of hatchery-origin fish on spawning grounds using coded wire tagging and parentage-based tagging. Transactions of the American Fisheries Society 145:671–686.

IEAB (Independent Economic Analysis Board). 2013. Cost-effectiveness of fish tagging technologies and programs in the Columbia River basin. Northwest Power and Conservation Council, Independent Economic Analysis Board, Document 2013-1, Portland, Oregon.

Narum, S. R., J. E. Hess, and A. P. Matala. 2010. Examining genetic lineages of Chinook Salmon in the Columbia River Basin. Transactions of the American Fisheries Society 139:1465–1477.

O'Reilly, P. T., and C. C. Kozfkay. 2014. Use of microsatellite data and pedigree information in the genetic management of two long-term salmon conservation programs. Reviews in Fish Biology and Fisheries 24:819–848.

Rawding, D. J., C. S. Sharpe, and S. M. Blankenship. 2014. Genetic-based estimates of adult Chinook Salmon spawner abundance from carcass surveys and juvenile out-migrant traps. Transactions of the American Fisheries Society 143:55–67.

Satterthwaite, W. H., J. Ciancio, E. Crandall, M. L. Palmer-Zwahlen, A. M. Grover, M. R. O'Farrell, E. C. Anderson, M. S. Mohr, and J. C. Garza. 2015. Stock composition and ocean spatial distribution inference from California recreational Chinook Salmon fisheries using genetic stock identification. Fisheries Research 170:166–178.

Seeb, L. W., A. Antonovich, A. A. Banks, T. D. Beacham, A. R. Bellinger, S. M. Blankenship, M. R. Campbell, N. A. Decovich, J. C. Garza, C. M. Guthrie, T. A. Lundrigan, P. Moran, S. R. Narum, J. J. Stephenson, K. J. Supernault, D. J. Teel, W. D. Templin, J. K. Wenburg, S. E. Young, and C. T. Smith. 2007. Development of a standardized DNA database for Chinook Salmon. Fisheries 32:540–552.

Steele, C. A., E. C. Anderson, M. W. Ackerman, M. A. Hess, N. R. Campbell, S. R. Narum, and M. R. Campbell. 2013a. A validation of parentage-based tagging using hatchery steelhead in the Snake River Basin. Canadian Journal of Fisheries and Aquatic Sciences 70:1046–1054.

Steele, C. A., M. R. Campbell, M. Ackerman, J. McCane, M. Hess, N. Campbell, and S. Narum. 2013b. Parentage based tagging of Snake River hatchery steelhead and Chinook Salmon, 2013. Idaho Department of Fish and Game, Project Progress Report, Boise.

Steele, C. A., M. R. Campbell, M. Ackerman, J. McCane, M. Hess, N. Campbell, and S. Narum. 2015. Parentage based tagging of Snake River hatchery steelhead and Chinook Salmon, 2015. Idaho Department of Fish and Game, Project Progress Report, Boise.

Stephenson, J. J., M. R. Campbell, J. E. Hess, C. Kozfkay, A. P. Matala, M. V. McPhee, P. Moran, S. R. Narum, M. M. Paquin, O. Schlei, M. P. Small, D. M. Van Doornik, and J. K. Wenburg. 2009. A centralized model for creating shared, standardized, microsatellite data that simplifies inter-laboratory collaboration. Conservation Genetics 10:1145–1149.

Van Doornik, D. M., D. J. Teel, D. R. Kuligowski, C. A. Morgan, and E. Casillas. 2007. Genetic analyses provide insight into the early ocean stock distribution and survival of juvenile Coho Salmon off the coasts of Washington and Oregon. North American Journal of Fisheries Management 27:220–237.

Van Doornik, D. M., R. S. Waples, M. C. Baird, P. Moran, and E. A. Berntson. 2011. Genetic monitoring reveals genetic stability within and among threatened Chinook Salmon populations in the Salmon River, Idaho. North American Journal of Fisheries Management 31:96–105. AFS