DR. MARIAH H MEEK (Orcid ID : 0000-0002-3219-4888)
MR. WESLEY ALAN LARSON (Orcid ID : 0000-0003-4473-3401)

Article type : Opinion

The future is now: amplicon sequencing and sequence capture usher in the conservation genomics era

Mariah H. Meek[1,*] and Wesley A. Larson[2,*]

1. Dept. of Integrative Biology and AgBio Research, Michigan State University. Email: mhmeek@msu.edu

2. U.S. Geological Survey, Wisconsin Cooperative Fishery Researcfh Unit, College of Natural Resources, University of Wisconsin-Stevens Point. Email: Wes.Larson@uwsp.edu

*Both authors contributed equally

Running title: Ushering in the conservation genomics era

Article type: News and Views Opinion

Abstract

The genomics revolution has initiated a new era of population genetics where genome-wide data are frequently used to understand complex patterns of population structure and selection. However, the application of genomic tools to inform management and conservation has been somewhat rare outside a few well-studied species. Fortunately, two recently developed approaches, amplicon sequencing and sequence capture, have the

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.12998
This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

potential to significantly advance the field of conservation genomics. Here, amplicon sequencing refers to highly multiplexed PCR followed by high-throughput sequencing (e.g. GTseq), and sequence capture refers to using capture probes to isolate loci from reduced-representation libraries (e.g. Rapture). Both approaches allow sequencing of thousands of individuals at relatively low costs, do not require any specialized equipment for library preparation, and generate data that can be analyzed without sophisticated computational infrastructure. Here, we discuss the advantages and disadvantages of each method and provide a decision framework for geneticists who are looking to integrate these methods into their research program. While it will always be important to consider the specifics of the biological question and system, we believe that amplicon sequencing is best suited for projects aiming to genotype < 500 loci on many individuals (> 10,000) or for species where continued monitoring is anticipated (e.g. long-term pedigrees). Sequence capture, on the other hand, is best applied to projects including fewer individuals or where > 500 loci are required. Both of these techniques should smooth the transition from traditional genetic techniques to genomics, helping to usher in the conservation genomics era.

Key words: amplicon sequencing, sequence capture, RAD sequencing, Rapture, GTseq, conservation genomics

Introduction

Genomics has revolutionized the fields of population genetics and molecular ecology (Andrews *et al.* 2016; Davey *et al.* 2011; Narum *et al.* 2013), and has enormous potential for facilitating similar advances in conservation. However, this potential has yet to be fully realized. Genomic data (defined here as data from 100s to 1000s of sequenced loci sampled across the genome) is now easier to generate and analyze for non-model organisms, and it

This article is protected by copyright. All rights reserved.

can now also be used to answer new questions that traditional genetic markers (limited marker sets such as microsatellites and mtDNA) cannot. This includes gaining an understanding of adaptive genetic variation and genotype-by-environment interactions, in addition to the more traditional analyses of population structure and gene flow (Bernatchez 2016; Luikart *et al.* 2003). Given this potential, several reviews over the last ten years have touted ways that genomic data should revolutionize the field of conservation (e.g. Allendorf, Hohenlohe, Luikart 2010; Bernatchez *et al.* 2017; Funk, McKay, Hohenlohe, Allendorf 2012; Garner *et al.* 2016; Hoffmann *et al.* 2015; Ouborg *et al.* 2010; Supple & Shapiro 2018), but this revolution has not fully come to fruition (Shafer *et al.* 2015). However, we believe we have reached the point where genomics can and should be a part of conservation science and practice. Here, we describe two recently developed genomic techniques, amplicon sequencing and sequence capture, that can help facilitate the transition from conservation genetics to conservation genomics. Additionally, we provide a decision framework for conservation practitioners to decide which of these genomic techniques is best for addressing their conservation questions. Our hope is this commentary will provide a roadmap that can be used by researchers who are attempting to integrate these new techniques into their workflow.

Amplicon sequencing and sequence capture facilitate the transition to conservation genomics

Two recently developed genomic approaches are primed to revolutionize the field of conservation genetics by reducing costs and facilitating more user-friendly and standardized analyses compared to methods such as restriction-site associated DNA (RAD) sequencing. These approaches are amplicon sequencing and sequence capture. Here, we define

Accepted Article

This article is protected by copyright. All rights reserved.

amplicon sequencing as highly-multiplexed PCR followed by high-throughput sequencing with the goal of genotyping of thousands of individuals at hundreds of markers in a single sequencing lane (e.g. Campbell, Harmon, Narum 2015), and sequence capture as the use of capture baits to target and sequence loci identified using reduced-representation approaches, such as RAD sequencing (e.g. Ali *et al.* 2016; Hoffberg *et al.* 2016) (Table 1). Both of these methods facilitate the collection of increased amounts of genomic data at costs that are similar to non-genomic techniques (Fig. 1). Additionally, these methods take advantage of improved and simplified analysis pipelines that are available as a result of the maturation of the population genomics field. The major impediments for most laboratories attempting to implement genomic techniques are cost and lack of bioinformatics expertise or powerful computing resources (Taylor, Dussex, van Heezik 2017). Amplicon sequencing and sequence capture address these hurdles, providing a much easier path to implement genomics than what was available only a few years ago.

Even with these recent advances, the transition from traditional methods (e.g. microsatellites) to genomics is non-trivial. However, we believe that the field of genomics has progressed to the point that there is little reason for conservation genetics laboratories not to begin the genomic transition. Laboratory methods for amplicon sequencing and sequence capture are relatively straight forward (e.g. Ali *et al.* 2016; Campbell *et al.* 2015; Hoffberg *et al.* 2016). Additionally, neither of these approaches require highly specialized equipment for library preparation, and sequencing can be conducted at a core facility, eliminating the need for laboratories to purchase an expensive sequencer.

Methods for analyzing genomic data have also advanced significantly over the last decade. Initially, analysis pipelines and parameters were rarely shared among laboratories, giving the field somewhat of a “wild west” mentality, with everyone developing their own

This article is protected by copyright. All rights reserved.

Accepted Article

analysis approaches. This made it extremely difficult to conduct genomic research without a strong background in computational biology and bioinformatics and likely prevented many conservation genetics laboratories from engaging in this type of research. As population genomics has matured, analysis pipelines have become more standardized, making it easier for laboratories unfamiliar with the handling of next generation sequencing data to break into the field. Researchers new to the field of genomic analyses will still require some time learning entry level bioinformatic skills, but this process is made easier by improved pipelines, such as the STACKS software (Catchen *et al.* 2013; Paris, Stevens, Catchen 2017), which gets more and more user friendly with each update, as well as documented and user-friendly scripts for analysis of amplicon sequencing data (e.g. Campbell *et al.* 2015; McKinney et al. in prep). Developers continue to improve these pipelines (Andrews *et al.* 2018; O'Leary *et al.* 2018; Paris *et al.* 2017; Rochette & Catchen 2017) and have identified important parameters that should be explored, providing guidance that minimizes the need to test the full parameter space for each new study.

Amplicon sequencing has recently been adopted by multiple genetics laboratories to genotype tens of thousands of individuals for research and monitoring efforts in salmonids (e.g. Matala *et al.* 2016). Additionally, a number of published studies have highlighted the utility of amplicon sequencing, including an analysis of thousands of Coho salmon (*Oncorhynchus kisutch)* for parental based tagging and genetic stock identification (Beacham *et al.* 2018; Beacham *et al.* 2017), analysis of Chinook salmon (*O. tshawytscha*) returning to a stream in Idaho over a 19-year period to evaluate long-term impacts of supplementation (Janowitz-Koch *et al.* 2018), genetic monitoring of Atlantic salmon (*Salmo salar*, Aykanat, Lindqvist, Pritchard, Primmer 2016), evaluating pedigree relationships in kelp rockfish (*Sebastes atrovirens*) using microhaplotypes (Baetscher *et al.* 2018), and determining allele

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

dosage in Chinook salmon (McKinney *et al.* 2018). Larson is also developing amplicon sequencing containing approximately 500 loci for various management applications in walleye (*Sander vitreus*), cisco, lake sturgeon (*Acipenser fulvescens*), and lake whitefish (*Coregonus clupeaformis*). These applied management efforts and published studies highlight how amplicon sequencing is useful for systems that require sample processing to be easy and low cost and data to be compatible across different laboratories.

When microsatellites are the preferred marker due to high polymorphism or existing datasets, amplicon sequencing methods can also be applied to transition from capillary sequencing to a next-generation sequencing based approach (Zhan *et al.* 2017). Bradbury *et al.* (2018) highlight this approach by identifying 101 microsatellite loci from the available Atlantic salmon genome and genotyping 1,558 individuals via amplicon sequencing. The results of this study identified previously undescribed fine-scale population structure in Atlantic salmon. However, the authors found it challenging to convert previously developed microsatellites to amplicon assays due to variation in product size and sequence length restrictions, suggesting that developing new panels may be preferable for microsatellites.

Sequence capture is currently being used by conservation geneticists to address conservation and management questions. However, because these methods are relatively new, many of the results of these studies have not yet completed the peer-review process. In the manuscript describing the sequence capture method, Rapture, Ali et al. (2016) genotype 288 individuals at 500 identified RAD tag loci on a fraction of an Illumina HiSeq2500 sequencing lane, demonstrating that it should be possible to genotype thousands of individuals on a full lane. The authors highlight the extreme flexibility of this method in the ability to optimize the number of individuals that can be genotyped and the

This article is protected by copyright. All rights reserved.

Accepted Article

number of loci that can be designed. In demonstrating the RADcap method, Hoffberg *et al.* (2016) describe genotyping up to 384 *Wisteria* sp. individuals at ~900 loci. Very recently, Margres *et al.* (2018) used Rapture to genotype 624 Tasmanian devils at ~16,000 SNP loci and another study Komoroske *et al.* (2018) used Rapture to genotype over 1000 samples from multiple species of marine turtle at the same ~2,000 SNP loci. We also each have multiple studies in progress using these methods. For example, Meek is using sequence capture to develop a panel of ~8,000 SNP markers for brook trout (*Salvelinus fontinalis*), a cold-water fish that is native to the eastern United States and Great Lakes region, for use in addressing management questions. The panel will be used to identify fine scale population structure, effects of hatchery stocking, and local adaptation for the watersheds around Lake Superior, as well as to better understand range-wide population structure and potential for adaptation to heat stress. Larson has also developed sequence capture panels containing 7,000-10,000 loci for cisco and walleye to investigate population structure and adaptive diversity across the Great Lakes region. These panels will be integrated into laboratory workflows across the Great Lakes, helping to usher in the genomics era in this region.

Comparison of two common methods, GTseq and Rapture

There are many variations of amplicon sequencing and sequence capture, but the two approaches most frequently used for conservation genomics questions, to our knowledge, are genotyping-in-thousands by sequencing (GTseq, Campbell *et al.* 2015) and Rapture (Rapture, Ali *et al.* 2016). GTseq is a cost-effective amplicon sequencing technique, originally developed to conduct high-throughput genotyping in Pacific salmon, and Rapture is an extension of the RADseq protocol (Baird *et al.* 2008) that enriches for sequences adjacent to restriction sites via a RAD-seq library preparation and then incorporates capture baits to target a subset of these sequences that are of interest. Other methods for amplicon

This article is protected by copyright. All rights reserved.

Accepted Article

sequencing include Illumina TruSeq® and ThermoFisher AmpliSeq®, and other methods for sequence capture of reduced-representation libraries include RADcap (Hoffberg *et al.* 2016). Additionally, there are also uses of amplicon sequencing and sequence capture that target specific genes or regions of the genome (e.g. the mitochondrial genome) but these methods are more often are used to answer questions related to phylogenetics and functional genomics. Comparing the advantages and disadvantages of all the amplicon sequencing and sequence capture methods is outside the scope of this paper. Instead, we will focus on comparing GTseq and Rapture because we believe these are the most widely used and applicable techniques for high-throughput genotyping for conservation. However, the tradeoffs between GTseq and Rapture should be similar to the tradeoffs between most amplicon sequencing and sequence capture methods that are used for high throughput sequencing to genotype SNPs.

While both GTseq and Rapture are clearly useful for a variety of applications, there are some major differences between the two techniques (Table 1). One of the most important differences involves panel development. For Rapture, target RAD tag sequences must be identified by conducting RADseq and SNP identification on a smaller set of individuals (aka the ascertainment panel). Then, target RAD tag sequences are sent to a private company who synthesizes the panel of “baits”. Baits are oligonucleotides that are complementary to the target sequence (i.e. target RADtag) that are then transcribed onto biotinylated RNA. This process usually takes a few months and does not require significant input from the researchers creating the panel, once the target sequences are identified, as the manufacturer does the testing to ensure bait compatibility. GTseq, on the other hand, requires researchers to identify target loci using RADseq or other methods and then conduct multiple rounds of primer testing to ensure that all primer pairs for the GTseq panel

This article is protected by copyright. All rights reserved.

Accepted Article

produce appropriate numbers of reads (G. McKinney, NOAA, personal communication). This process also takes a few months but necessitates significantly more input from the researcher than the process to create a Rapture panel.

Library preparation is much simpler for GTseq compared to Rapture. The major steps for GTseq library preparation are PCR amplification of targeted amplicons, DNA normalization, and ligation of index barcodes (Campbell *et al.* 2015). Rapture is essentially a RAD library preparation with a sequence capture at the end (enrichment of target sequence via streptavidin coated magnetic beads) and therefore involves many more steps than GTseq, including enzyme digestion, multiple bead cleanups, shearing, and size selection (Ali *et al.* 2016). Additionally, while the shearing step for Rapture can be conducted using a fragmentase enzyme, most laboratories prefer using a sonicator to fragment DNA more precisely. Sonicators are available for use at most university core laboratories or can be purchased for $10,000-$20,000. Rapture also requires a time-consuming DNA normalization step prior to library preparation, whereas the normalization step in GTseq is plate-based and much simpler. Finally, the DNA quality and quantity required for Rapture are also likely higher than for GTseq due to simplicity of the GTseq library preparation and the additional PCR cycles included in the GTseq protocol. However, the influence of DNA quality and quantity on the GTseq and Rapture approaches has not been explicitly tested (but see Komoroske *et al.* 2018 for the relationship between DNA quantity and sequence alignments in Rapture data).

Analysis is also much simpler and faster for GTseq data compared to Rapture data, but this is primarily a function of the number of loci generated from each approach and the requirement for sequence alignment for Rapture. Analysis of Rapture data containing thousands of loci is most often conducted using programs such as STACKS, which require a

This article is protected by copyright. All rights reserved.

sequence assembly step, as well as many other steps with multiple parameters to optimize (e.g. Hoffberg *et al.* 2016). Contrastingly, GTseq analysis most often utilizes in silico probes and pattern matching to count the number of reads for each allele (Campbell *et al.* 2015), eliminating the need for a sequence assembly step. Using current pipelines, analysis of typical RADseq or Rapture data generally takes days to weeks, whereas analysis of GTseq data takes hours to days. It is important to note that these are the analytical approaches that are typically applied to GTseq and Rapture data, but the approaches are flexible (e.g. GTseq analysis methods could be used to analyze Rapture data).

One major advantage of Rapture compared to GTseq is panel flexibility; one does not need to worry about primer interactions. Each time a primer pair is added to a GTseq panel, the panel needs to be retested to ensure that all primer pairs still amplify as expected, while with Rapture one just needs to order a new bait set from the manufacturer, who will do the necessary and relatively minimal testing. However, with Rapture one is limited to loci found in RAD tags (next to targeted restriction sites), while GTseq can be used to genotype any locus that can be PCR amplified, including loci in functional genes or loci that have been previously published (e.g. existing SNP assays).

Another advantage of Rapture is the number of loci that can be genotyped. GTseq panels are limited to approximately 500 loci due to issues associated with primer interactions (G. McKinney, NOAA, personal communication). Therefore, if researchers want to genotype more than 500 loci, they need to design multiple panels, which is relatively costly (Table 1). Rapture, on the other hand, is highly flexible and can be used to genotype everything from hundreds of SNPs to tens of thousands. It is important to note that sequencing costs for Rapture will scale based on the number of loci interrogated and will be similar to GTseq for smaller panels containing hundreds of loci (Table 1).

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Reliably genotyping all loci that are initially screened in panel development is unlikely for both GTseq and Rapture, but the percentage of loci that can be reliably genotyped seems to be slightly higher for GTseq. For Rapture, some baits will have low capture efficiency, preventing reliable genotyping of all loci (Ali *et al.* 2016). While few studies have reported the percentage of loci that are effectively captured, personal observations from our datasets suggest this number is approximately 80-90% of targeted loci. For GTseq, differences in primer efficiency can cause certain loci to under or over amplify, necessitating their removal. For example, Baetscher *et al.* (2018) had to remove 27 of the 192 markers that were initially screened (85% of initial loci retained). However, by manipulating primer concentrations and redesigning primers, it may be possible to retain a high percentage of initially screened loci. Campbell *et al.* (2015) used these techniques to reliably genotype 187 of 192 (97%) of targeted loci with GTseq. If genotyping > 90 % of targeted loci is necessary, we suggest GTseq with the caveat that achieving this goal could take significant panel testing beyond what is described in Table 1.

Data quality has been well-vetted with GTseq (> 99.99% genotype concordance with established methods, Campbell *et al.* 2015; Janowitz-Koch *et al.* 2018), but no such comparisons have been conducted for Rapture. It is likely that Rapture will also achieve high concordance with established methods, but, until this is verified, we suggest some caution if Rapture is used for applications that are highly sensitive to genotyping error. For example, if Rapture is used for these highly sensitive applications, we suggest strict read depth cutoffs and genotyping parameters.

This article is protected by copyright. All rights reserved.

GTseq or Rapture: Decision framework for conservation genomics studies

Laboratories that conduct genetic analysis on Pacific salmon have been among the leaders when it comes to using new genetic techniques to answer applied questions in conservation genomics, due to the large number of samples that these laboratories need to process (often over a hundred thousand per year per lab) and the financial resources available for marker development in salmonids. For example, SNPs were adopted by the Pacific salmon genetics community in approximately 2005, well before they had percolated into most conservation genetics applications (Seeb *et al.* 2011). Many laboratories conducting genetic analysis of Pacific salmon on the US West Coast have chosen to use amplicon sequencing rather than Rapture for high-throughput applications such as parentage-based tagging and mixed-stock analysis. However, this may be due to several factors, including the fact that GTseq was developed before Rapture, so it gained early momentum. GTseq was also adopted by salmonid genetics programs because existing SNP datasets could be converted to GTseq panels, allowing historic datasets to continue to be used. Rapture would have required recreating and genotyping baseline datasets at new loci as Rapture only targets SNPs that are found in RAD tags. Additionally, the shorter and simpler workflow of GTseq compared to Rapture and the fact that GTseq should be more robust to variation in DNA quality and quantity likely factored into this decision. Finally, the fact that GTseq panels require significantly more investment to develop than Rapture panels was likely not a major impediment for salmon geneticists because there are a large number of researchers working on developing genomic resources for salmon. The cost-benefit analysis of various techniques, however, is often different for geneticists working on Pacific salmon compared to other conservation geneticists, who may only run hundreds of samples of a given species per year.

Accepted Article

This article is protected by copyright. All rights reserved.

Many conservation geneticists work on multiple species. As such, each project represents an opportunity to conduct a cost-benefit analysis and choose the best available technique for the task at hand. The appropriate technique for each project is largely dependent on the number and quality of samples that need to be run, requirements for particular loci to be genotyped (e.g. to match existing datasets or include important genes), the probability that samples from the same species/populations will need to be run into the future, the amount of data that have already been collected for the species, the accuracy/quality of data required, and the biological questions being asked. We provide a decision tree in Fig. 2 to offer guidance for determining what method is best for new projects.

The target loci and research question will play a large role in determining if Rapture or GTseq is best. If target loci have already been identified and are located outside of RAD tags (e.g. SNPs to match baselines, microsatellites, adaptive loci of interest, or SNPs found through whole-genome resequencing), then GTseq will need to be used. However, for species that have little to no prior genomic information, RADseq and/or Rapture may be the best approach. In general, we believe RADseq should be used for small projects (roughly < 500 samples depending on genome size and sequencing costs) on species with little previous data. Even if RADseq generates more data than necessary for these projects, there will be little, if any, financial benefit associated with developing either Rapture or GTseq panels for projects with this small number of samples.

If data from thousands of loci is required for > 500 samples, developing a Rapture panel is the preferred approach. However, deciding if Rapture or GTseq should be used for projects that include a large number of samples and require data from < 500 loci (e.g. parentage analysis or mixed-stock analysis with highly discernible stocks) is more

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

challenging. In general, if researchers expect to be regularly genotyping samples from a given species for many years into the future and require high data quality, we suggest constructing a GTseq panel. However, if researchers are conducting a single large project (e.g. largescale parentage analysis) but do not anticipate genotyping samples from that species often in the future, Rapture may be the best approach. The per sample cost for Rapture is roughly twice that of GTseq, but the cost to develop a GTseq panel is about three times as much as a Rapture panel (Table 1). Based on the combination of panel development and genotyping costs, we suggest developing Rapture panels for projects with between 500 and 1,500 samples, and GTseq panels for larger projects.

The best approach may also be a combination of these methods. Researchers may want to conduct a smaller scale RADseq project to identify SNP markers, and then convert those to a Rapture or GTseq approach for large-scale genotyping. This is the approach Meek is taking with the brook trout studies. Another valid approach is to design both GTseq and Rapture panels for a given species. For example, Larson designed a Rapture panel for cisco including approximately 7,000 SNPs for high-resolution analysis of population structure and adaptive variation and is also planning to design a smaller GTseq panel for more high-throughput analyses such as species/stock identification and parentage analysis.

We did not include microsatellite genotyping using capillary electrophoresis in the decision tree presented in Fig. 2 because we believe developing small panels of microsatellites for new species is no longer optimal given the other genomic techniques available. We recognize that many laboratories still genotype microsatellites using capillary sequencers and will continue to do so for many years. However, it is important to note that microsatellites can be genotyped using amplicon sequencing (Zhan *et al.* 2017), although there are some difficulties associated with converting existing loci to amplicon assays

This article is protected by copyright. All rights reserved.

(Bradbury *et al.* 2018). We encourage laboratories to consider the potential benefits of transitioning to genomic methods for as many projects as possible (e.g. high-throughput, transferability of data across labs, technical support). Capillary sequencers may stop being supported at some point in the future, likely forcing researchers to transition anyway at what may be an inconvenient time. Additionally, the next generation of researchers will likely not be familiar with the idiosyncrasies associated with running, scoring, and standardizing microsatellites (Pasqualotto, Denning, Anderson 2007; Seeb *et al.* 2007), making it difficult to generate reproducible data. Microsatellites are powerful markers and will continue to have their place in the portfolio of genomic techniques (e.g. Zhan *et al.* 2017). However, we believe the utility of capillary based analysis of microsatellites is nearing its end.

Conclusions

A short three years ago, Shafer *et al.* (2015) discussed the difficulties associated with translating genomics into conservation practice. The technical hurdles they highlighted include the need for more mature and easily applicable methods, the development of analytical pipelines, and the presence of successful case studies for practitioners to learn from. We believe with the advent of the techniques described above, these technical roadblocks have been significantly minimized. We now have methods that can be applied without highly specialized and expensive laboratory equipment or laboratory expertise, pipelines that are user friendly, and an impressive and growing set of success stories. The field of conservation can now reap the benefits of next-generation sequencing, and we are excited to see great strides made in the conservation and management of the earth's

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

biodiversity. The time for the transition from conservation genetics to conservation genomics is now.

Acknowledgments

We thank Garrett McKinney, Andrew Whiteley, Shawn Narum, and four anonymous reviewers for providing helpful reviews of the manuscript and for many informative discussions on this topic. Any use of trade, product, or company name is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Ali OA, O'Rourke SM, Amish SJ*, et al.* (2016) RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. *Genetics* **202**, 389-400.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11**, 697-709.

Andrews KR, Adams JR, Cassirer EF*, et al.* (2018) A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data. *Molecular Ecology Resources* **0**.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**, 81-92.

Aykanat T, Lindqvist M, Pritchard VL, Primmer CR (2016) From population genomics to conservation and management: a workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon *Salmo salar*. *Journal of Fish Biology* **89**, 2658-2679.

This article is protected by copyright. All rights reserved.

Accepted Article

Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC (2018) Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources* **18**, 296-305.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376.

Beacham TD, Wallace C, Jonsen K*, et al.* (2018) Comparison of coded-wire tagging with parentage-based tagging and genetic stock identification in a large-scale coho salmon fisheries application in British Columbia, Canada. *Evolutionary Applications*, in press.

Beacham TD, Wallace C, MacConnachie C*, et al.* (2017) Population and individual identification of coho salmon in British Columbia through parentage-based tagging and genetic stock identification: an alternative to coded-wire tags. *Canadian Journal of Fisheries and Aquatic Sciences* **74**, 1391-1410.

Bernatchez L (2016) On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. *Journal of Fish Biology* **89**, 2519-2556.

Bernatchez L, Wellenreuther M, Araneda C*, et al.* (2017) Harnessing the Power of Genomics to Secure the Future of Seafood. *Trends in Ecology and Evolution* **32**, 665-680.

Bradbury IR, Wringe BF, Watson B*, et al.* (2018) Genotyping-by-sequencing of genome-wide microsatellite loci reveals fine-scale harvest composition in a coastal Atlantic salmon fishery. *Evolutionary Applications* **11**, 918-930.

Campbell EO, Brunet BMT, Dupuis JR, Sperling FAH, Matschiner M (2018) Would an RRS by any other name sound as RAD? *Methods in Ecology and Evolution*, online early.

This article is protected by copyright. All rights reserved.

Accepted Article

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources* **15**, 855-867.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**, 3124-3140.

Davey JW, Hohenlohe PA, Etter PD*, et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology* **26**, 5369-5406.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. *Trends in Ecology and Evolution* **27**, 489-496.

Garner BA, Hand BK, Amish SJ*, et al.* (2016) Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. *Trends in Ecology and Evolution* **31**, 81-83.

Graham CF, Glenn TC, McArthur AG*, et al.* (2015) Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources* **15**, 1304-1315.

Hoffberg SL, Kieran TJ, Catchen JM*, et al.* (2016) RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources* **16**, 1264-1278.

Hoffmann A, Griffin P, Dillon S*, et al.* (2015) A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses* **2**, 1.

This article is protected by copyright. All rights reserved.

Accepted Article

Janowitz-Koch I, Rabe C, Kinzer R*, et al.* (2018) Long-term evaluation of fitness and demographic effects of a Chinook Salmon supplementation program. *Evolutionary Applications*, online early.

Komoroske LM, Miller M, O'Rourke S*, et al.* (2018) A Versatile Rapture (RAD-Capture) Platform for Genotyping Marine Turtles. *Molecular Ecology Resources*, online early.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* **4**, 981-994.

Margres MJ, Jones ME, Epstein B*, et al.* (2018) Large-effect loci affect survival in Tasmanian devils (*Sarcophilus harrisii*) infected with a transmissible cancer. *Molecular Ecology* **27**, 4189-4199.

Matala AP, Hatch DR, Everett S*, et al.* (2016) What goes up does not come down: the stock composition and demographic characteristics of upstream migrating steelhead differ from post-spawn emigrating kelts. *Ices Journal of Marine Science* **73**, 2595-2605.

McKinney GJ, Waples RK, Pascal CE, Seeb LW, Seeb JE (2018) Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources* **18**, 570-579.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology* **22**, 2841-2847.

O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS (2018) These aren’t the loci you’e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* **27**, 3193-3206.

This article is protected by copyright. All rights reserved.

Accepted Article

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**, 177-187.

Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**, 1360-1373.

Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: Lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. *Journal of Clinical Microbiology* **45**, 522-528.

Puckett EE (2017) Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. *Conservation Genetics Resources* **9**, 289-304.

Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D (2012) Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution* **44**, 21.

Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols* **12**, 2640-2659.

Seeb JE, Carvalho G, Hauser L*, et al.* (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources* **11 Suppl 1**, 1-8.

Seeb LW, Antonovich A, Banks AA*, et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries* **32**, 540-552.

Shafer AB, Wolf JB, Alves PC*, et al.* (2015) Genomics and the challenging translation into conservation practice. *Trends in Ecology and Evolution* **30**, 78-87.

Supple MA, Shapiro B (2018) Conservation of biodiversity in the genomics era. *Genome Biology* **19**, 131.

This article is protected by copyright. All rights reserved.

Accepted Article

Taylor H, Dussex N, van Heezik Y (2017) Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation* **10**, 231-242.

Zhan L, Paterson IG, Fraser BA*, et al.* (2017) megasat: automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources* **17**, 247-256.

Author contributions

MHM and WAL contributed equally to the to the writing of the manuscript.

This article is protected by copyright. All rights reserved.

Accepted Article

TABLES

Table 1. Comparison of traditional RAD sequencing (RADseq), sequence capture (Rapture), and amplicon sequencing (GTseq). The number of loci that can be targeted with RADseq and Rapture is highly flexible, but we focus on a typical study using traditional RAD with the *SbfI* enzyme for the sake of simplicity. See Andrews *et al.* (2016) and Campbell *et al.* (2018) for more information on different variations of RADseq.

	RADseq	Rapture	GTseq
# loci genotyped	~20,000	500-10000	~500/panel
Approximate cost per sample ($US) (1)	$30	$15	$6
Ease of library preparation (2)	Moderate, ~1 week	Moderate, ~1 week	Simple, 2 days
Constrained to RAD tags	Yes	Yes	No
Approximate panel development cost (3)	Not applicable	$4,000	$13,000-$15,000
Approximate panel development time (3)	Not applicable	4 months	4 months
DNA quality required (4)	Medium-high	Medium-high	Low-medium
Bioinformatics expertise required	Intermediate/advanced	Beginner/Intermediate	Beginner
Utility for relatedness analysis (5)	Complex pedigree reconstruction	Complex pedigree reconstruction	Parent-offspring, full siblings
Sample Throughput	Low	Medium	High
Potential for rapid (< 2 week) turnaround (6)	No	Yes, but relatively difficult	Yes

(1) Cost per sample assumes that samples are being sequenced efficiently (96 samples per lane for RADseq, 384 samples per lane for Rapture, and 960 samples per lane for GTseq). Our multiplexing estimates were designed to be conservative therefore multiplexing more individuals per lane than stated above is likely possible in many circumstances (e.g. Campbell *et al.* 2015). Sequencing 96 individuals per lane has proven efficient for *SbfI* RADseq in salmon on a HiSeq4000, but this value will vary based on genome size, enzyme, and sequencing technology. Genotyping hundreds rather than thousands of loci with Rapture will decrease the cost per sample, but Rapture is still more expensive than GTseq when genotyping the same number of loci because of increased library preparation costs. Breakdown of costs: RADseq – library preparation: $5/sample (Fuentes-Pardo & Ruzzante 2017) and 96 individuals

This article is protected by copyright. All rights reserved.

multiplexed per sequencing lane ($2500 per 150 bp paired end lane), Rapture – library preparation: $8/sample ($5 for RADseq library preparation, $3 for bait capture) and 384 individuals multiplexed per sequencing lane (150 bp paired end), GTseq –$4/sample (Campbell *et al.* 2015) and 960 individuals multiplexed per sequencing lane ($1500 per 150 bp single end lane). Prices are rounded up to the nearest dollar. Labor not included in cost estimates. All calculations assume sequencing is conducted on a HiSeq4000. It is important to note that reagent costs differ substantially among sequencing platforms and the cost per base pair will be greater for low output sequencers (e.g. MiSeq) than for high-output sequencers (e.g. HiSeq). The $3 per sample price for Rapture bait capture was calculated as follows: $3,600 for a 16 reaction 20k myBaits® kit that includes all necessary reagents/ 16 reactions / 96 individuals per reaction = $2.34 per sample rounded up to the nearest dollar = $3. The costs for consumables (microcentrifuge tubes and pipette tips) is less than one cent per sample as 96 individuals are pooled in a single tube for the bait capture.

(2) Preparing RADseq and Rapture libraries requires more steps than GTseq and also requires that DNA be sheared either physically or with an enzyme. Time estimates are the approximate time needed to construct a single library from extracted DNA based on personal experience. Multiple libraries can be constructed simultaneously, but the number of libraries that can be efficiently constructed together will vary widely by laboratory.

(3) Assumes that RADseq data or other data (e.g. existing SNP panel, exome sequence, genome sequence) are already available. Costs associated with discovering markers for panel development will vary based on a number of factors including type and scope of data available. For example, discovering markers using data from a few sequenced genomes will likely be more difficult and produce a higher proportion of low-quality markers compared to discovering markers from a comprehensive RAD dataset. Cost estimates are rounded to the nearest thousand dollars, and labor is not included in the estimates. Development of Rapture panel is passive (i.e. a company creates the baits after researcher submits the target loci sequences), whereas development of GTseq panel is active (i.e. significant time required to test and troubleshoot panel). Cost estimate for Rapture panel development is the cost to order the smallest synthesis of 20k myBaits® probes ($3600) rounded to the nearest thousand dollars ($4,000). This order can run 16 reactions, corresponding to 1,536 samples if running a single 96-well plate per reaction. Baits can also be ordered for 48 or 96 reactions, with substantial cost savings per reaction compared to the 16-reaction kit. Additionally, baits can be developed for up to 200,000 targets and prices increase by $1,250 per 20,000 baits with the 16-reaction kit. The cost estimate for GTseq panel development is represented as a range because it includes both fixed costs (reagents) and variable costs (sequencing). The fixed costs are ordering 500 primer pairs at $20/pair ($10,000), which facilitates genotyping of approximately 30,000 individuals per primer order, and costs for other reagents and consumables required for sample preparation ($2,000). The sequencing costs required to test the panel are highly variable depending on whether a MiSeq or other sequencing instrument is available. For example, a small amount of library can be spiked onto existing MiSeq runs at a very low cost, if access to a MiSeq is available (total

Accepted Article

This article is protected by copyright. All rights reserved.

sequencing cost ~$1,000). However, if no MiSeq is available, the cost to buy 2-3 full lanes to test a panel will be larger (~$3,000).

(4) cost of two MiSeq and one HiSeq lane ($7,500), and approximate cost of other reagents ($2,500).

(5) RADseq has been shown to be sensitive to DNA quality (Graham *et al.* 2015). Although no published studies have evaluated the impact of DNA quality on GTseq, amplicon sequencing approaches have been shown to be highly robust to variation in DNA quality (Rizzi *et al.* 2012).

(6) Utility for relatedness analysis will depend on within-population genetic variation and complexities in mating systems. Researchers should conduct simulations to determine how many markers are necessary to accurately infer relatedness in their system.

(7) Sequencing with MiSeq or other rapid sequencing technology once panel developed is necessary for rapid turnaround. However, cost per sample will increase due to less optimized sequencing on these platforms. Library preparation time is significantly longer for Rapture compared to GTseq, making rapid analysis more difficult with Rapture.

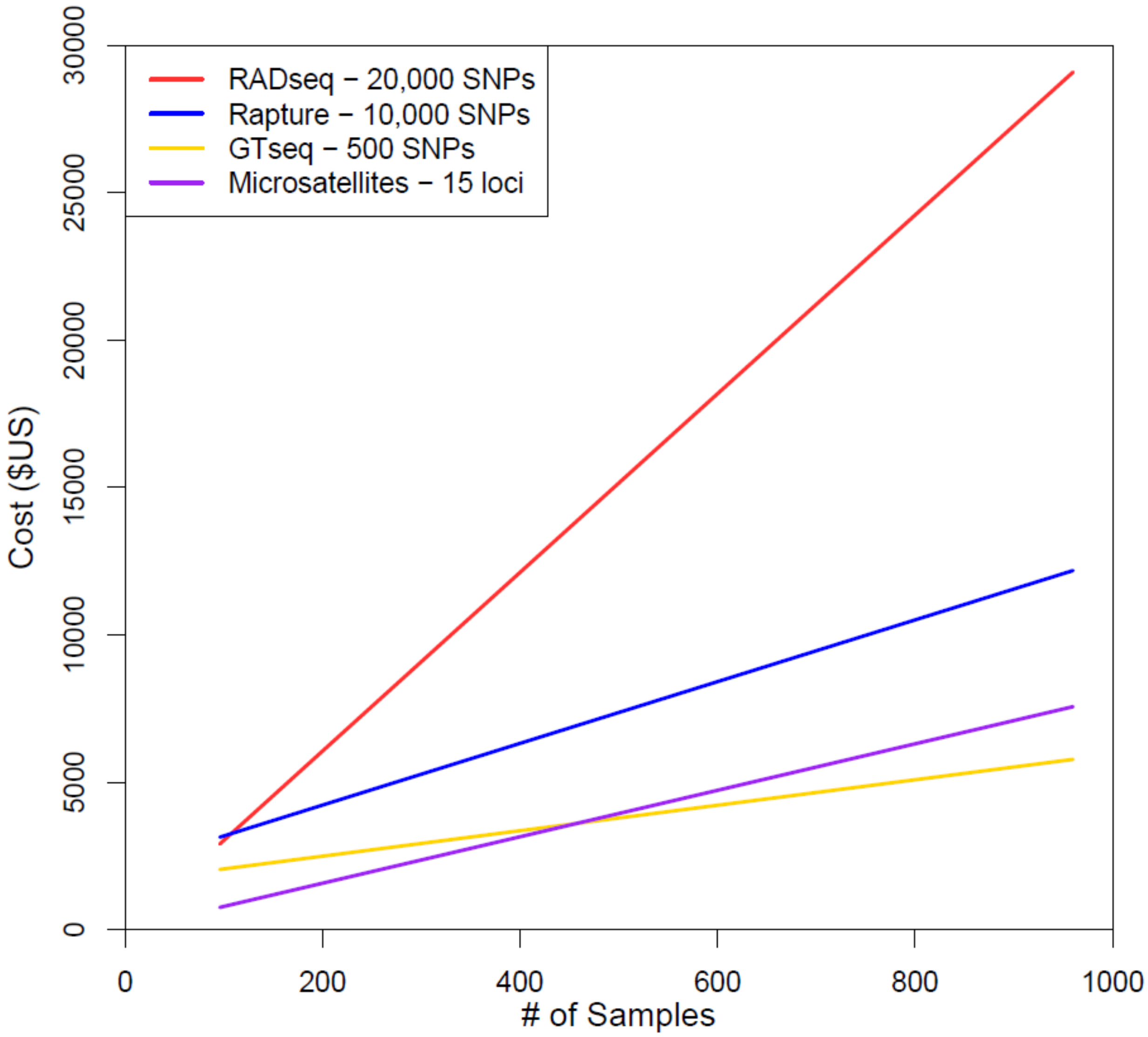
Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

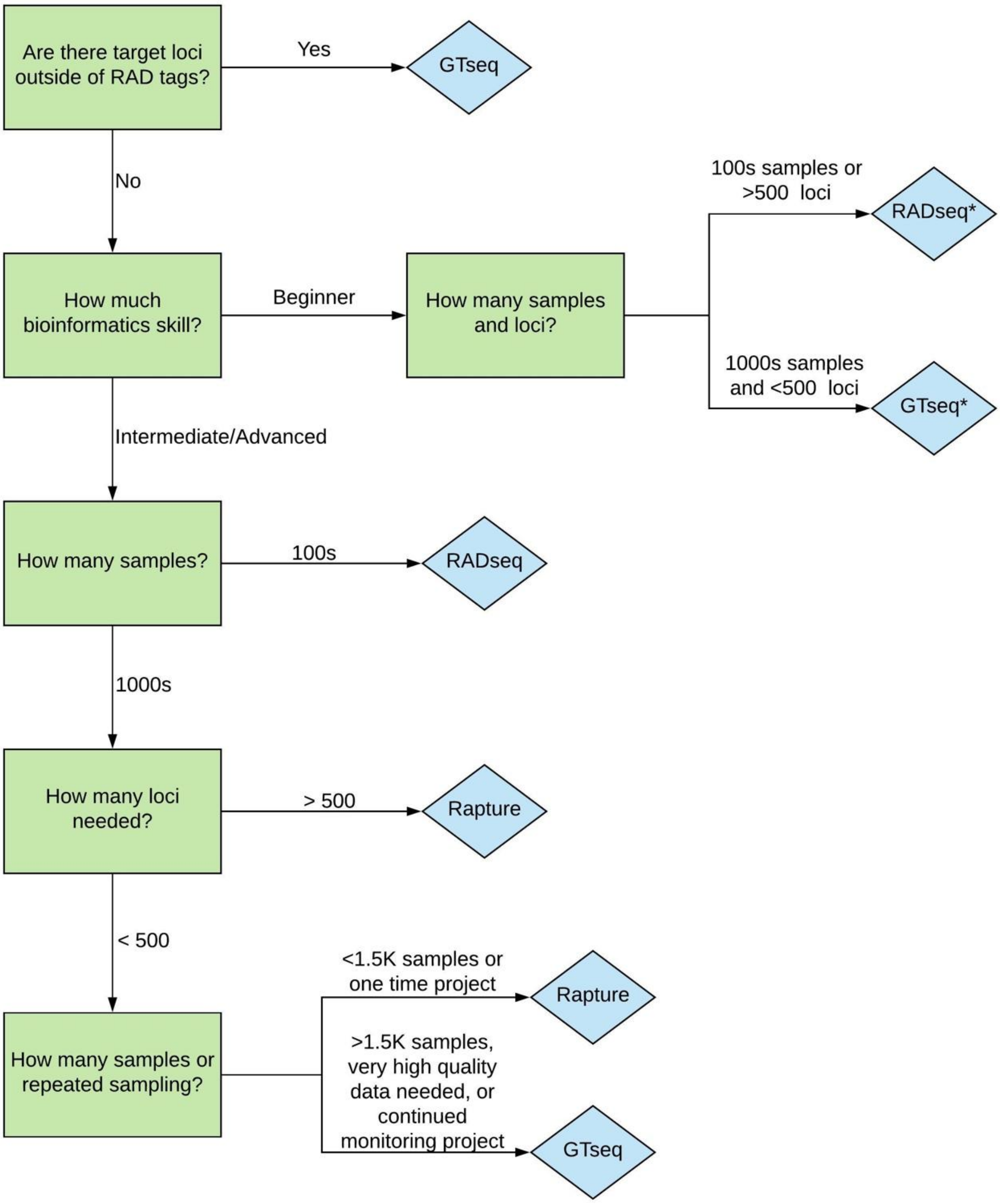
FIGURES:

Figure 1: Cost analysis for genotyping 96 to 960 samples with four chemistries: RADseq, sequence capture (Rapture), amplicon sequencing (GTseq), and microsatellites. Rapture (10,000 SNPs) becomes more cost effective than RAD (20,000 SNPs) after 96 samples due to the ability to multiplex many more individuals on a sequencing lane. GTseq (500 SNPs) becomes less expensive than a typical microsatellite panel containing 15 loci genotyped in five multiplexes after about 500 samples. Sample preparation and sequencing costs for RADseq, Rapture, and GTseq are found in Table 1. If less than one lane was required for sequencing, we still included the full cost of a lane because most researchers purchase sequencing by lane. For example, the sequencing cost for GTseq is fixed at $1,500 for all sample numbers from 96-960 meaning that the cost per sample for 96 samples ($19.63) is much higher than the cost per sample for 960 samples ($5.56). A per sample cost of $8 was used for microsatellites based on cost analysis conducted in the Larson laboratory. This per sample cost is about half of that reported by Puckett (2017).



This article is protected by copyright. All rights reserved.

Figure 2: Decision tree to determine most efficient sequence-based approach for genotyping new species. We assume that researchers with "beginner" bioinformatics experience will have basic knowledge of unix and computational infrastructure (power and storage) but will need substantial assistance to conduct a full genomics project. *Contract work out to private company. This is necessary for the panel development. If there is already a panel developed, little to no bioinformatics knowledge is necessary for GTseq.


Are there target loci outside of RAD tags?
Yes
GTseq
No
How much bioinformatics skill?
Beginner
How many samples and loci?
100s samples or >500 loci
RADseq*
1000s samples and <500 loci
GTseq*
Intermediate/Advanced
How many samples?
100s
RADseq
1000s
How many loci needed?
> 500
Rapture
< 500
How many samples or repeated sampling?
<1.5K samples or one time project
Rapture
>1.5K samples, very high quality data needed, or continued monitoring project
GTseq

Accepted Article

This article is protected by copyright. All rights reserved.