



Fishing for DNA? Designing baits for population genetics in target enrichment experiments: guidelines, considerations and the new tool superBaits

Jiménez-Mena, Belén; Flávio, Hugo; Henriques, Romina; Manuzzi, Alice; Ramos, Miguel; Meldrup, Dorte; Edson, Janette; Pálsson, Snaebjörn; Ólafsdóttir, Guðbjörg Ásta; Ovenden, Jennifer R.

Total number of authors:
11

Published in:
Molecular Ecology Resources

Link to article, DOI:
[10.1111/1755-0998.13598](https://doi.org/10.1111/1755-0998.13598)

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Jiménez-Mena, B., Flávio, H., Henriques, R., Manuzzi, A., Ramos, M., Meldrup, D., Edson, J., Pálsson, S., Ólafsdóttir, G. Á., Ovenden, J. R., & Nielsen, E. E. (2022). Fishing for DNA? Designing baits for population genetics in target enrichment experiments: guidelines, considerations and the new tool superBaits. *Molecular Ecology Resources*, 22(5), 2105-2119. <https://doi.org/10.1111/1755-0998.13598>











General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Fishing for DNA? Designing baits for population genetics in target enrichment experiments: Guidelines, considerations and the new tool superBaits

Belén Jiménez-Mena¹  | Hugo Flávio¹  | Romina Henriques¹  | Alice Manuzzi¹  | Miguel Ramos²  | Dorte Meldrup¹ | Janette Edson³  | Snæbjörn Pálsson⁴  | Guðbjörg Ásta Ólafsdóttir⁵  | Jennifer R. Ovenden⁶  | Einar Eg Nielsen¹ 

¹Section for Marine Living Resources, National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark

²DCC-FCUP, University of Porto, Porto, Portugal

³Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

⁴Faculty of Life and Environmental Sciences, University of Iceland, Reykjavík, Iceland

⁵Research Centre of the Westfjords, University of Iceland, Bolungarvík, Iceland

⁶Molecular Fisheries Laboratory, School of Biomedical Sciences, The University of Queensland, Brisbane, Queensland, Australia

Correspondence

Belén Jiménez-Mena, Technical University of Denmark, DTU Aqua, Vejløvej 21, Silkeborg, Denmark.

Email: bmen@aqua.dtu.dk

Funding information

The Icelandic Research Fund, Grant/Award Number: "CODSTORY"; Innovationsfonden: Udvikling af den danske laksebestand - større populationer, genetiske ressourcer og rekreativt fiskeri", Grant/Award Number: "SDPAS; The Danish Council for Independent Research Grant DFF, Grant/Award Number: 6108-00583; Australian Research Grant, Grant/Award Number: DP170102043

Handling Editor: Alana Alexander

Abstract

Targeted sequencing is an increasingly popular next-generation sequencing (NGS) approach for studying populations that involves focusing sequencing efforts on specific parts of the genome of a species of interest. Methodologies and tools for designing targeted baits are scarce but in high demand. Here, we present specific guidelines and considerations for designing capture sequencing experiments for population genetics for both neutral genomic regions and regions subject to selection. We describe the bait design process for three diverse fish species: Atlantic salmon, Atlantic cod and tiger shark, which was carried out in our research group, and provide an evaluation of the performance of our approach across both historical and modern samples. The workflow used for designing these three bait sets has been implemented in the R-package superBaits, which encompasses our considerations and guidelines for bait design for the benefit of researchers and practitioners. The superBaits R-package is user-friendly and versatile. It is written in C++ and implemented in R. superBaits and its manual are available from Github: <https://github.com/BelenJM/superBaits>

KEYWORDS

ancient DNA, baits, capture sequencing, genomics, population genetics, R-package

1 | INTRODUCTION

Genomic information is increasingly available for population genetics analyses due to the rapid development of next-generation sequencing (NGS) methods. A multitude of wild species are studied in this way; however, the method is particularly important for endangered or commercially exploited species, where knowledge generated from genome-wide data can greatly aid in conservation and sustainable management efforts (Supple & Shapiro, 2018). Many of these species are not widely used for general research questions, so reference genomic resources to initiate NGS studies are rarely available (Russell et al., 2017). Different NGS approaches are currently available, with their suitability varying with the study question and type of organism at hand. For example, while sequencing whole genomes provides very detailed data on the genomic architecture of a species, this approach remains time-consuming and expensive, given the high cost of producing, analysing and storing the large quantity of data obtained. Alternatively, methods of reduced-representation sequencing allow investigation of specific regions of the genome in a large number of conspecific individuals at a relatively low cost and short time (Mamanova et al., 2010), especially when the genome size of the organism of interest is large or complex (McCartney-Melstad et al., 2016). One of these reduced-representation methods is the so-called “target enrichment” approach, which targets specific areas of interest within the genome (Mamanova et al., 2010). There are multiple methods of target enrichment, for example, PCR-based enrichment or hybridization-based capture sequencing (see Mamanova et al., 2010). Hybridization-based capture sequencing (herein referred as CS) is currently one of the quickest and most flexible methods for target enrichment (Mamanova et al., 2010) and can be performed using fixed predefined solid-arrays or in-solution (Horn, 2012). The latter is based on *in vitro* hybridization of the target genomic regions with designed synthetic probes of DNA or RNA, that is, “baits”, that will “capture” the complementary sequence that the bait was designed for (Horn, 2012). In principle, only the desired genomic areas for which the baits were designed will be captured and sequenced, thus CS has commonly been used for historical (hDNA) and ancient DNA (aDNA) studies as it increases the yield of sequence from study-species by reducing the probability of sequencing contaminants (Willerslev & Cooper, 2005). Likewise, it has been suggested CS could drive the transition from conservation genetics to conservation genomics (Meek & Larson, 2019), given its flexibility and cost effectiveness (between \$43 and 65 per sample in plants—see Hale et al., 2020). One potentially economical caveat of CS is the high percentage of off-target reads, that is, reads that map to non-target regions. In exon-capture experiments it has been estimated that on average 40%–60% of the total amount of reads sequenced are off-target (Samuels et al., 2013). Nevertheless, although *a priori* a disadvantage, the off-target reads can still provide useful insights within subsequent data analysis. For instance, the mitochondrial DNA is very often sequenced in CS experiments as an off-target (Picardi & Pesole, 2012) and it has been used to identify and clean

out contaminated individuals belonging to other species or samples, as well as additional markers (e.g., Manuzzi et al., 2021). Other uses of off-target data reported in the literature are for example, identification of new SNPs (Guo et al., 2012) or repeat regions (Costa et al., 2021). Discussion of pros and cons of CS in comparison with other genomic approaches (i.e., high- or low-coverage whole-genome sequencing, other reduced-representation approaches) is outside the scope of this article, but we refer to the following studies/reviews for further comparisons (e.g., low-coverage sequencing: Lou et al., 2021; Therkildsen & Palumbi, 2017; whole-genome sequencing: Schwarze et al., 2020).

1.1 | Why capture sequencing for population genetics?

In population genetics studies, documenting neutral processes is of particular interest from a conservation and resource management point of view (Zhou & Holliday, 2012). Neutral processes such as gene flow, population divergence and demographic history, allow the study of how populations are connected through time and space. The ability to select putative neutral genomic regions (and discard others that may be under natural selection or linked to such sites) provides an advantage for processing genetic data for conservation purposes (Zhou & Holliday, 2012). Alternatively, working with putative loci under selection can disentangle adaptive and neutral processes. For example, studies of loci supposedly under selection may identify local adaptation among populations subject to spatially varying selective pressures (e.g., Nielsen et al., 2009), or assess adaptation over time in response to temporally changing selective pressures (e.g., Franks et al., 2016; Therkildsen et al., 2013), such as global climate change. For many instances in population genetic studies, there is a greater benefit from increasing the number of individuals analysed as opposed to increasing genomic coverage per individual (Fumagalli, 2013), for example, Benestan et al. (2015). CS permits data collection from an increased number of sampled individuals per sequencing lane for the same cost compared to whole-genome approaches (Jones & Good, 2016), at the expense of having fewer sequenced genomic regions. Another useful application of CS is for studies of historical or ancient DNA samples. For example, if one is interested in assessing the loss of genetic diversity through time, it is common to use historical or ancient samples, which generally yield smaller amounts of endogenous DNA that may be degraded and contaminated with DNA from a variety of organisms, including bacteria, fungi and viruses (Willerslev & Cooper, 2005). Therefore, baits designed to capture DNA from the target species have a huge potential to increase the success of retrospective population genomic studies. Likewise, CS is advantageous when working with endangered or nearly-extinct species where samples could be scarce and of low-quality (Glenn & Faircloth, 2016), as well as in environmental samples (Giebner et al., 2020). In conclusion, CS is powerful because it specifically targets DNA from the species of interest, including in contaminated or mixed samples, and specifically

selects regions within those genomes to answer particular research questions at a reasonable cost.

Because CS methods rely on targeting a narrow set of desired regions in the genome, the design of the oligonucleotide baits is key for the success of any CS-based study. To date, bait design has focused mainly on comparing conserved regions in humans (Hodges et al., 2009) or in phylogenetic studies (Andermann et al., 2020; Hancock-Hanser et al., 2013; Hugall et al., 2016; Lemmon et al., 2012). Recent bait design software has reflected this trend by focusing mainly on exonic regions that tend to be ultra-conserved (Campana, 2018; Chafin et al., 2018; Faircloth, 2017; Mayer et al., 2016) or long regions (>20 kb, Jayaraman et al., 2020) of the genome. However, little attention has been given to bait design processes for population genetic studies using CS, where the focus lies largely on determining within-species genetic structure and diversity. There is also currently an absence of guidelines, pipelines and specific software to help in this endeavor. In most cases, the process of designing baits is outsourced to manufacturers who ensure the baits are compatible and of the best quality, but it is time-consuming and expensive (Meek & Larson, 2019). For example, in the case of medical genomics, several manufacturers have predesigned panels for genomic regions of interest, as well as tools for creating “custom capture reagents” for enrichment of genomic regions specified by the laboratory (see Glenn & Faircloth, 2016; Hagemann et al., 2013, for a review). For nonmedical related studies, predesigned panels do not generally exist, meaning that each project needs to create their own set of baits (Glenn & Faircloth, 2016). Exceptions can be found within phylogenetics (see Table S1 of Andermann et al., 2020), and a few of the databases have been successfully used in palmske BLAST and excluding bfor population genetic purposes, for example, in reptiles (Singhal et al., 2017) or frogs (Chan et al., 2020, 2021; Hutter et al., 2019). In addition, there is very limited literature available describing how to successfully design baits for population genetics, including the reasoning behind such bait design (but see Puritz & Lotterhos, 2018). Accordingly, bait design is usually left to short methodological descriptions in individual research manuscripts (see examples in Table S1).

Here, we present novel guidelines and considerations for designing baits for population genetics that will save time and effort. We also discuss three empirical examples of bait design that investigated changes through time in genetic diversity using time-series data from three fish species to inform conservation and management strategies. The aim of the SDPAS (“Strengthening the Danish Populations of Atlantic Salmon: increasing populations, genetic resources and recreational fishing”) project was to investigate the temporal variation in proxies of genetic diversity in the population of Atlantic salmon (*Salmo salar*) in Denmark over the span of a century. For this, we used ~1000 samples with a temporal range from 1913 to 2017, and targeted different genomic areas for elucidating both neutral and adaptive changes over time. In the project CODSTORY we investigated genetic changes in Atlantic cod (*Gadus morhua*) in Icelandic waters, to assess possible association with changes in fisheries practices over almost 1000 years. Finally, as

part of the GENOJAWS project, we wanted to understand whether tiger sharks (*Galeocerdo cuvier*) had suffered a recent historical (since 1820) loss of genetic diversity, associated with climate change or human-induced ecological perturbances. For these three studies, DNA was extracted from jaws, vertebrae, scales, bones and tissue samples collected from our local institute and museums and excavations across the world. We briefly discuss the suitability and broad applicability of our novel bait design approach using these three examples. Finally, we explain the main functionalities of superBaits, an R-package designed for researchers and practitioners to design their own bait sets for CS experiments. Along with the R-package, we also make the designed panels of bait sequences available for the research community.

2 | WHAT TO TAKE INTO ACCOUNT WHEN DESIGNING YOUR BAITS?

2.1 | Available genomic resources

An increasing number of non-model species have reference genomes available (Hohenlohe et al., 2021). An annotated reference genome of the species of interest allows researchers to more accurately select genomic regions for bait design. The annotation helps to locate intron/exon boundaries allowing identification of coding/non-coding regions of the genome subject to different evolutionary forces (Warr et al., 2015). In our research projects dealing with Atlantic salmon and Atlantic cod, we made use of the available reference genome for the Atlantic salmon (ICSASG_v2, Lien et al., 2016) and the latest genome assembly for the Atlantic cod (GadMor2, Tørresen et al., 2017), respectively. If full genomes are not available, transcriptomes can also be used for bait design (e.g., Bailey et al., 2016; Capblancq et al., 2020; Ehlers et al., 2020; Förster et al., 2018). For the tiger shark, we used the available transcriptome assembly from white muscle (Swift et al., 2016). However, even if genomic resources from the species of interest are not available, those from a closely related species or species group can still be used (e.g., Cosart et al., 2011; Nielsen et al., 2017), and even be used to generate new genomic resources. For instance, Förster et al. (2018) found 686 candidate SNPs in the Eurasian Lynx (*Lynx lynx*) using baits designed from the domestic cat (*Felis catus*), to generate a 96-SNP panel to monitor the presence of the species in the wild. When using genomic information from another species, it is important to take the evolutionary distance into account (Jones & Good, 2016), as this will influence the effectiveness of the bait hybridization. However, one could also choose to study divergent variation within a family of species, thus the design of the baits should target areas with some level of divergence between the species; for example, Sanderson et al. (2020) designed baits in regions that were less than 95% identical between two species targeted in their study. In the case of complete lack of genomic resources, which is still common for many non-model species, there are other methods available to circumvent the problem, mostly by generating new genomic resources. For example, PCR

capture, de novo assembly capture and divergent reference capture (see Jones & Good, 2016), or more recently the combination of RAD-sequencing with capture, that is, RAD-capture (Ali et al., 2016), and expressed-exome capture (Puritz & Lotterhos, 2018) can all be used when no genomic resources are available. In addition to finding and selecting genomic regions for CS from available genomic resources, in this study we also designed baits for regions containing already identified SNPs and reused capture baits that were previously designed from other species, for example, baits designed from the cat shark (*Scyliorhinus canicula*) transcriptome that had previously captured tiger sharks' sequences successfully (Manuzzi et al., 2021). However, this is not an ideal approach, as biases can be introduced by not knowing the genomic location of cross-species designed baits in the species of interest (e.g., linkage between markers when assuming independence), which should be kept in mind in downstream analyses.

2.2 | The research question

As with other aspects of planning a research project, the specific questions and hypotheses should also guide the bait design process. Thus, designing a bait panel should provide the opportunity to generate enough data to address the specific research question in a cost-effective way. For instance, to focus on population genetics of antelopes and measure genetic diversity, Gooley et al. (2020) designed 5000 baits outside exonic areas to target 5000 putatively neutral SNPs. Else, researchers can choose to divide the bait panel into different sets aimed at answering various questions within a population genomics-related project; including addressing both neutral processes and selection/adaptation and therefore focusing on non-coding/coding regions of the genome. In our projects, bait sets were designed to target: (i) Single nucleotide polymorphisms of interest (SNPs) from published SNP panels. In the SDPAS and CODSTORY projects we used information from SNP chips (Hubert et al., 2010; Karlsson et al., 2011; Moen et al., 2008) or SNPs previously applied to population genetic studies in our laboratory (Therkildsen et al., 2013); (ii) genes of interest or regions with known quantitative trait loci (QTL). As this approach provides no prior SNP knowledge, we allocated baits randomly around each gene/QTL regions, for example, genomic regions identified as related to parasite-driven evolution in Atlantic salmon (Zueva et al., 2014), or genes related to survival in the wild (Besnier et al., 2015); (iii) genes of interest identified from available transcriptomes, exemplified with the bait design for the tiger shark project; (iv) particularly interesting genomic regions, such as the four known inversions in Atlantic cod that characterize the different ecotypes (Barney et al., 2017; Kirubakaran et al., 2016). Other studies also used this approach for generating some of the baits; for example genes associated to environmental stress responses (Bi et al., 2019), and (v) putative "neutral" areas of the genome (i.e., not in or adjacent to genes), in order to obtain sufficient data on neutral genomic processes to allow estimation of neutral indices such as effective population size (N_e) and

other measures of genetic diversity through time. In this instance, we generated sequences placed throughout the genome, but excluded repetitive areas; a similar approach was used in Gooley et al. (2020). Figure 1 shows a scheme of the classes of targeted areas. More details on the distribution of different bait classes for the three projects and species can be found in Table S2.

2.3 | Length and number of baits

The impact of the choice of bait length remains understudied (Glenn & Faircloth, 2016). It is currently unknown what optimal, minimum or maximum sequence length is needed for the bait to capture the desired sequence. In some cases, bait design may only be guided or limited by the choice of sequencing platform and the size of the sequencing reads, as well as the length of the sequence fragment to be captured (Horn, 2012). The CS method captures a range of sequences between a few hundred base pairs (bp) to a few thousand base pairs (Mbp), and also allows a relatively high proportion of degenerate sites, in contrast to PCR primers (Glenn & Faircloth, 2016; Horn, 2012). A bait length of 120 bp is generally considered as representing a good balance between cost and efficiency (Glenn & Faircloth, 2016). Therefore, this was the chosen length of bait in our three projects. One can also choose to design bait sets of different lengths, for example for historical and modern samples; Joubran and Cassin-Sackett (2021) had a separate bait panel for the historical collection with shorter length (100 bp) than for the modern collection (120 bp). Given that some of our samples were historical and hence likely to be degraded (Table S3), we expected the captured DNA fragments to consist primarily of short sequences, and accordingly, we chose a short-read sequencing platform (Illumina). Fragmentation of extracted DNA to the desired size can be achieved using mechanical or enzymatic techniques (Hale et al., 2020), for example, when working with well conserved DNA or modern samples. Coupled with the development of new technologies related to the sequencing of long genomic regions (e.g., PacBio), CS is also evolving towards capturing longer regions (up to 20 kb), in the so-called region-specific extraction (RSE) (Dapprich et al., 2016).

The number of baits to use will be a trade-off between different factors related to the budget of the project and the research question in mind. Frandsen et al. (2020) used >59,000 baits to have enough power to obtain a high resolution when studying admixture levels of subspecies in the European ex situ population of the chimpanzee (*Pan troglodytes*), whereas a lower number of baits (8922) were needed to discover enough SNPs to identify the presence of *Lynx* from samples collected noninvasively in the wild. The chosen number of baits will often depend on the desired mean coverage depth for each sequenced individual for each targeted area, the expected efficiency of the CS approach and the sequencing platform capacity per lane (Grover et al., 2012). When deciding on the total number of baits to aim for, it is important to take into account the expected efficacy of the capture in the species of interest. Although this depends on multiple factors, efficacy will be lower for instance

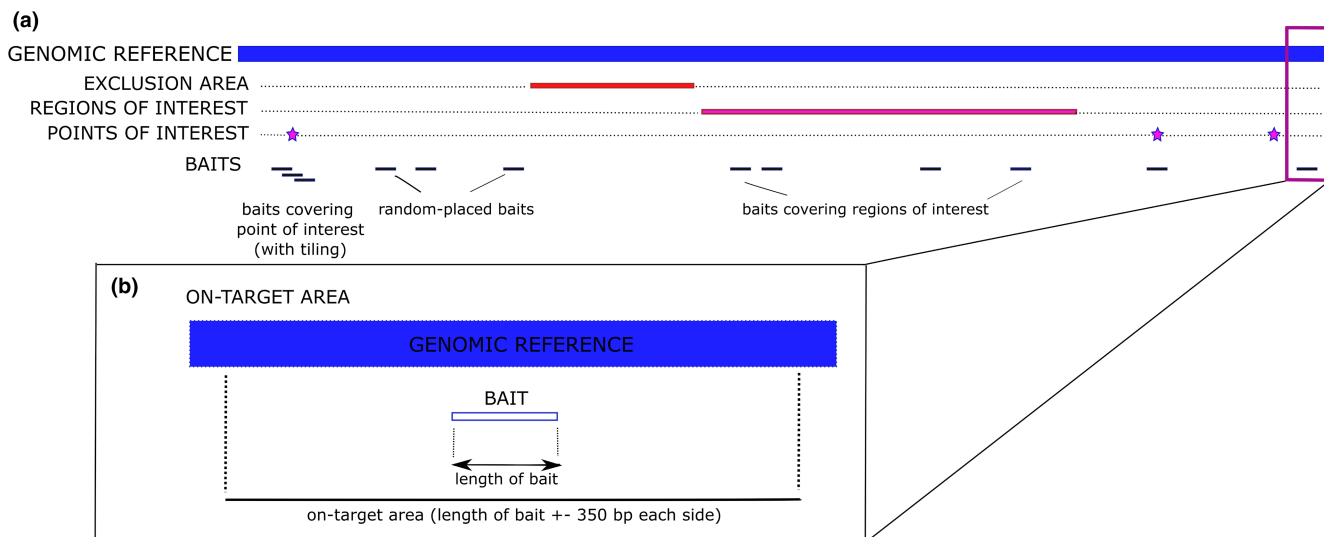


FIGURE 1 (a) Illustration of the design of the bait set. Different types of areas are taken into account for the design: exclusion areas, where no baits will be placed upon; regions of interest, typically genes or other areas to explore in the research questions; and points of interest, typically SNPs. (b) Diagram showing the “on-target” area. A read was considered “on-target” if it was located within 350 bp up or downstream of the genomic position of the designed capture bait of 120 bp

when designing baits based on a distant species (Bragg et al., 2016). CS allows increasing the number of samples at the expense of covering a smaller fraction of the genome, but targeting a sufficient number of SNPs to answer the desired research questions is essential. For example, we did not choose all the SNPs in the published SNP chips of Atlantic cod (Hubert et al., 2010; Moen et al., 2008) or salmon (Karlsson et al., 2011), but only those for which we had hypothesis-driven questions. For our projects, we designed 20,000 baits for each of the three species; this number was a balance between the cost of baits, the number of samples processed in the laboratory (~1000 samples for SDPAS, ~300 for CODSTORY and ~400 for tiger sharks) and the predefined bait sets offered by the company used to produce the baits (MYBaits, now Arbor Biosciences). We first designed baits for targeted regions and previously identified SNPs of interest (between 2 and 5 thousand (K) baits per project) to ensure they were sufficiently covered, and designed the remainder of baits as “random”, targeting non-coding and putative neutral regions throughout the genome (between 15 and 18 K baits per project; see Table S2). We expected these numbers of baits to generate sufficient SNPs for drawing a multitude of genomic inferences. Simulations show that, in some cases, ~1000 SNPs may be enough to reliably estimate levels of genetic diversity (Nazareno et al., 2017), while as low as ~100 SNPs often suffice to confidently analyse population structure and conduct population assignments (Turakulov & Easteal, 2003). When estimating N_e , a recent study found that ~2,000 random SNPs provided consistent N_e estimates, through different missing data levels and minor allele frequency (MAF) thresholds (Marandel et al., 2020). Nevertheless, further work is necessary to estimate the suitable number of SNPs needed for N_e estimation of regional genomic regions in order to account for heterogeneity of N_e across the genome (Jiménez-Mena et al., 2016; Jiménez-Mena, Tataru, et al., 2016). Prior knowledge about the expected level of

genomic variation (e.g., number of SNPs per Mbp) could serve as a starting point to guide the number of baits to aim for. However, not all baits will capture fragments with a SNP and this will be most pronounced for species with less overall genomic variation.

2.4 | Duplicated regions

Duplicated regions have been highlighted as a drawback of CS as these regions are captured and amplified more often than nonrepeat regions (Ávila-Arcos et al., 2011), thereby swamping sequencing reads. Accordingly, it has been recommended to design baits for targets outside repetitive areas (Horn, 2012). This may be challenging for species that have experienced genomic duplication events (e.g., Atlantic salmon; Lien et al., 2016) and with different ploidy levels (e.g., strawberries: Kamneva et al., 2017; black cottonwood: Zhou & Holliday, 2012). The same reasoning applies for not targeting both mitochondrial and nuclear genes in the same bait panel, as nuclear-mitochondrial homologues are abundant (Woischnik & Moraes, 2002). Thus, it is recommended to use available data on repeat regions of the species of interest as well as to apply bioinformatics tools allowing identification of such genomic regions and filtering out baits that fall within those areas. For example, in our projects we excluded repeat regions for the Atlantic salmon bait panels, using the Repeat Library report published with the ICSASG_v2 salmon genome (Lien et al., 2016). For tiger sharks, repeats were excluded by the company who later on produced the bait set (see below, Arbor Biosciences) using the Carcharhiniformes repeat database. For the cod, we used the Repeat Library report published with the gadMor2 genome (Tørresen et al., 2017). Although recommended, it is not always possible to filter out these regions, if there are not any suitable genomic resources. Therefore, researchers should keep

this in mind for downstream analyses. Another good practice is to double-check the baits for matching to multiple genomic regions, which could be achieved by using tools like BLAST (Camacho et al., 2009), excluding baits that map the genome more than once (e.g., Sanderson et al., 2020), or using departures from the expected mean coverage and heterozygosity along the genome to filter out duplication areas (e.g., Harpe et al., 2019).

2.5 | Tiling

Using more than one bait to cover an area of interest ("tiling") is likely to increase the chances of efficiently capturing sequences from a specific genomic area. Thus, if a given study aims to target a number of genes of particular high interest, then tiling may be an efficient approach to assure successful capture and higher coverage. In a comparative study of different exome bait panels that consisted of (i) adjacent, (ii) nonadjacent and (iii) overlapping baits, it was shown that overlapping baits increased the performance of targeted sequence capture. In this case, less sequencing reads were needed to obtain a good resolution of the variability of the specific genomic region, thereby increasing the sensitivity of the method (Clark et al., 2011). It is difficult to provide general guidelines on selecting the number of overlapping baits as well as their overlap and density (Clark et al., 2011; Glenn & Faircloth, 2016); however, Cruz-Dávalos et al. (2017) recommends three to five-fold tiling densities for enriching degraded DNA libraries including aDNA. Different bait tiling strategies for the various genomic regions covered by the panel can also be applied, in order to ensure that the essential genomic regions of interest are successfully captured. For example, in our case studies, we designed baits with 3×-tiling of prioritized regions of interest compared to randomly-selected genomic areas (Figure 2). These included SNPs known to be linked to "sea-age" in the Atlantic salmon (how many years a salmon stays at sea before returning to the river for reproduction (Barson et al., 2015)), and SNPs related to salinity preference (Berg et al., 2015) and sex determination (Star et al., 2016) for Atlantic cod. By contrast, for some other areas (e.g., "random" areas), the priority was to cover a large number of regions and potentially as many different SNPs as possible, and thus only used a single bait per region. One can also choose to use a homogeneous tiling strategy throughout the bait design; in a study on the population structure and genetic diversity of the ex-situ population of sable antelope in North America, the tiling strategy was 4× for all the 5000 neutral SNPs targeted (Gooley et al., 2020).

2.6 | Base composition

The GC content (i.e., the proportion of guanine [G] and cytosine [C] nucleotides in the sequence) has a direct influence on the capture efficiency for targeted exonic regions, with very low and very high GC content regions having negative effects on the efficiency of hybridization (Chilamakuri et al., 2014). Whether GC content also

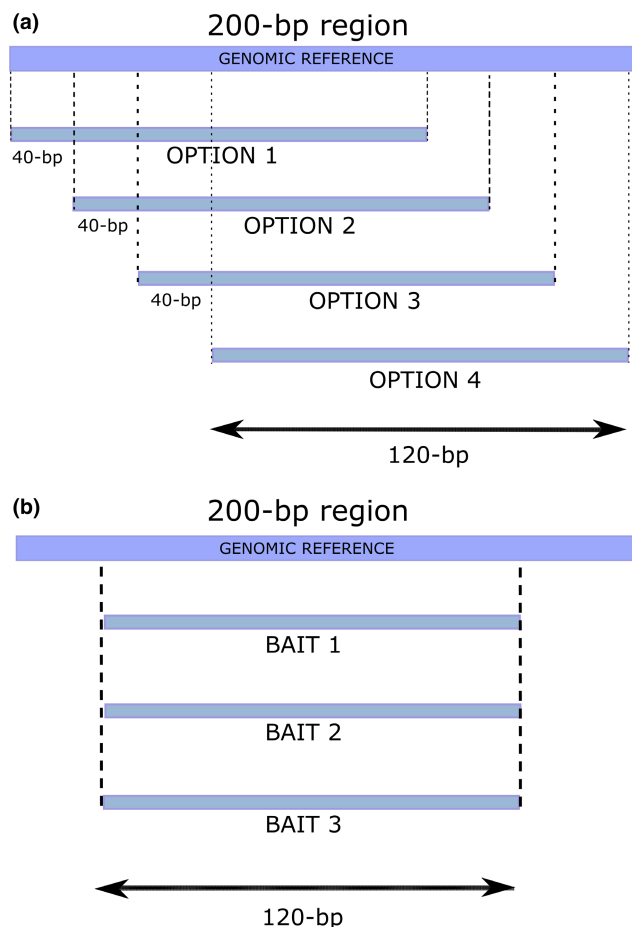


FIGURE 2 Examples of different options of tiling to design baits for a region of interest. (a) Tiling using a given offset distance between baits (e.g., 40 bp), (b) exact tiling (e.g., 3×)

affects capture efficiency outside coding areas (with typically lower GC content—Fortes et al., 2007; Vinogradov, 2001) has been less studied, but some studies indicate a similar negative effect (see Jones and Good (2016), and references therein). Cruz-Dávalos et al. (2017) evaluated baits designed along the nuclear genome of the horse and found that increasing GC content (>53%) reduced the number of baits with at least 1 read coverage, as well as the mean coverage. Accordingly, as a rule of thumb, it is generally accepted that GC content of the bait panel should be kept at intermediate levels, avoiding areas with very low (<30%) or very high (>70%) GC content in order to try to compensate as much as possible for the capture efficiency bias. For our study species, we only used baits with GC content within a range of 40%–55%. In order to facilitate selection of baits within that range, we initially generated a larger number of baits than the desired 20,000 (~5 times more), and of a larger size (200 bp) than the final length of each bait (120 bp), whenever possible. For each of the 200 bp-sequences, we designed overlapping baits of 120 bp with a 40 bp-offset between baits, in order to have a broad selection to choose from (Figure 2). This approach allowed us to design baits meeting the GC criteria for almost all of the initial 200 bp-sequences.

2.7 | Other considerations

The thermodynamic properties of the nucleotide sequences can impact the annealing specificity of the designed bait sequence to the desired target (see review by Noguera et al., 2014). The affinity of two sequences can be quantified by measuring the Gibbs free energy change of sequence binding (ΔG). This is applied in order to take into account the properties of both the target and the designed bait to create self-folding structures that do not allow the correct binding between them, and penalize for these formations during bait design. Melting temperature (T_m), defined as the temperature where 50% of the bait sequences are hybridized, should also be considered. In particular, T_m should be relatively homogenous across baits allowing optimal capture conditions. There are other chemical properties of importance for bait design, and we refer to the work of Cruz-Dávalos et al. (2017) for more detailed considerations.

Finally, baits can be built from RNA or DNA (Horn, 2012), where RNA baits seem to give a higher stability compared to DNA when binding to DNA (see Hale et al., 2020). For our three experiments, we exclusively used RNA baits produced by an external manufacturer (Arbor Biosciences). For our case studies, the final preselected bait sets for all three species (see Table S2) were sent to Arbor Biosciences, who provided a review of the chemical properties of the sequences, as well as suggested filtered baits following their own thresholds on T_m and BLAST cutoffs, and when applicable, the options described above for the 200-bp sequences (see Table S3). After filtering for GC content, T_m and sequence specificity (i.e., a score that characterizes the bait specificity when blasting it against

the genome of the target species), we selected the final set consisting of 20,000 sequence baits subsequently produced by Arbor Biosciences (Supporting Information S1). The main guidelines described in this section are summarized in Table 1.

3 | TESTING THE PERFORMANCE OF BAIT SETS – INSIGHTS FROM THE THREE CASE STUDIES

Before proceeding with the capture of all samples, it is recommended to conduct a capture trial on a subset of individuals. The trial should cover the range of DNA quality (fragmentation) and quantity likely to be experienced throughout the project in order to get a good overview of the performance of the bait set. In our projects, we captured DNA from 20 individuals for each species, of which 10 were from contemporary samples and the other 10 from historical or “ancient” samples, that is, with lower concentration and more fragmented DNA. The capture in the laboratory was conducted following Arbor Biosciences guidelines. More information about the samples can be found in Table S3, including type of tissue and year of sampling/catch. Captured libraries were sequenced at two external sequencing facilities on a HiSeq4000 provider, using 2×125 bp paired-end (PE) reads (tiger shark, salmon), and a HiSeq X using 2×150 bp PE reads (cod). Raw sequencing data were filtered for adaptors, potential bacteria and human contamination, and subsequently mapped back to their respective genomic resources. We filtered for mapping quality and PCR duplications, and

TABLE 1 Summary table of the main considerations on the design of baits for population genetics

	Type	Example
Available genomic resources	Genome Transcriptome De novo assemblies Other (close) species	Atlantic salmon and Atlantic cod Tiger shark
Question	Neutral vs. adaptive processes Population substructuring Estimates of effective population size Retrospective genomics Environmental DNA	Coding/non-coding regions Anonymous regions of the genome/transcriptome Neutral areas of the genome Coding/noncoding regions, anonymous regions
Type of targeted region	Known SNPs Genes of interest/quantitative traits loci Inversions Neutral areas of the genome	Baits in SNPs (e.g., from SNP-chips) Randomly allocated baits in genes or regions of interest Baits in known inversions Randomly allocated baits
Bait length	~70–200 bp Up to 20 Kbp	120 bp
GC content	Avoid very low (<30%) or very high (>70%) areas	40%–55%
Tiling	Tiling Mixed tiling/no tiling No tiling	Tiling for areas of interest/No tiling for random areas
Other considerations	Sequence binding (ΔG) Melting temperature (T_m) BLAST hits	

obtained BAM files, which were used for statistical analysis of the bait panels' performance. Further details on the laboratory DNA extraction, sequencing and bioinformatics filtering are outside the scope of this manuscript, but we followed a similar procedure as in Manuzzi et al. (2021).

We evaluated capture sensitivity (i.e., the percentage of targets covered by at least one mapped read; (Jones & Good, 2016); coverage (i.e., mean number of reads per bait) and depth of targeted base pairs in BEDtools (functions: intersect and coverage with `-hist` option: Quinlan & Hall, 2010; Figure 3). We defined a read as "on-target" if the read overlapped the bait region (i.e., target area, 120 bp) or a flanking sequence of 350 bp on each side of the bait (Figure 1b), allowing partial overlap in both cases. The flanking sequence was included in the "on-target" area in case a sequencing read had mapped to the ends of a bait, thus extending beyond the bait length. For each of the three species, more than 75% of the baits had at least one read "on-target", and all groups presented similar value ranges, except the historical cod (Figure 3a). For all studies, clear differences in efficiency according to the age of the samples were observed. As expected, modern samples had a higher success in the total number of captured target regions (overall mean—modern samples: 17,920; historical: 16,209), as well as more reads per bait than historical/ancient samples (overall mean - modern samples: 101; historical: 52.4). As the most extreme case, the samples of historical cod had the widest range of capture efficiency (min: 3.498; max: 17,077 baits capturing), although the median was 14,610 baits, which was similar to the contemporary samples for cod and the other species. A similar wide range was observed for the mean number of reads per bait, where the tiger shark (both historical and modern) and the modern cod presented the broadest range (threefold) among samples (Figure 3b). Modern and historical samples of salmon displayed relatively little variation in read number among samples, but had the lowest mean values of all six groups captured (with the exception of historical cod). On the contrary, the modern cod samples exhibited the largest fraction of the targeted regions captured by the baits (Figure 3c).

Our trial runs revealed different degrees of capture efficiency. This included not only differences among species, but also between historical and modern samples, as well as the type of tissue source, the age and the preservation method of the samples, suggesting that the type of samples can have an effect in the success of the capture experiments. Similar findings have been reported across the literature (Derkarabetian et al., 2019; Nielsen et al., 2017), further illustrating that capture efficiency is not a "one-measure-fits-all" and should be tailored to the species and type of samples at hand, although broader bait sets can also successfully work across large phylogenetic scales (e.g., Hutter et al., 2019). We highly recommend that researchers conduct trial runs before embarking on a full capture study. If capture efficiency is considered too low, hybridization conditions can be modified (e.g., temperature and bait/template concentrations) to optimize capture efficiency.

4 | THE R-PACKAGE *superBaits* AND ITS APPLICATION

The considerations for designing baits described in this article have been collated and implemented in a user-friendly R-package *superBaits*. *superBaits* consists of a main function `do_baits` that reads genomic information provided by the user to design baits for the species of interest. The only mandatory input is a file containing genomic information, typically a genome or transcriptome in FASTA format (a "database" in *superBaits* terms). If available, other types of genomic information can also be used, for example, previously identified SNPs, regions of particular interest, and areas to exclude (i.e., masked regions). For illustration purposes we made a short comparison on the time for *superBaits* to load the genomic resources (Figure 4a,b) and design different number of bait sets for the three species used in this study (salmon, cod and tiger shark), using the most basic parameters (i.e., `do_baits [n = n, size = 120, database]`) (Figure 4c). The speed test was performed using an Intel Core i5-7200 U 2.50GHz with two cores and 8 GB RAM. The databases from these three species differ in size (Atlantic salmon: 3 Gb, Atlantic cod: 613 Mb, tiger shark: 155 Mb) and in number of scaffolds (Atlantic salmon: 232,155, Atlantic cod: 8310, tiger shark: 179,867). As *superBaits* is written in C++, it can effectively handle a variety of data set sizes; however, the larger the data set, the longer it takes to load (Figure 4a). The smaller the data set, the more other factors (e.g., storing length values to a table) start playing a significant role on the total time spent to load (which in turn lowers the kBP/s) (Figure 4b), which for example explains why the tiger shark database has a lower kBP/s despite being the fastest data set to load. The time that it takes *superBaits* to create baits is dependent on the number of scaffolds of the database (Figure 4c).

The arguments within the main function of the package (Table 2) allow the user to define how many baits to design, and where they should be placed within the genome. This can be done by specifying the total number of baits, the number/percentage of baits per category of input file, and if different categories are to be included, for example, known SNPs, genomic regions of particular interest, and masked regions. The tiling can also be specified, including information about different bait characteristics per input file category (e.g., 2× tiling for known SNP areas, and 3× tiling in regions of particular interest). If genomic regions to exclude are specified, no baits will be placed in those regions. The user can also define the GC content range within a bait, specifying a minimum and maximum allowed content. The output of the package is a set of baits for each specified type. The output also includes different statistics along with the DNA sequence that can be used for follow-up filtering analyses. If desired, the generated bait list output from *superBaits* (`do_baits()` function) can be used as an input to apply further filtering (e.g., based on chemical properties, see Jayaraman et al., 2020) both in *superBaits* through the `blast_baits()` function that utilizes BLAST software (Camacho et al., 2009) within the R-package, but also using other ad hoc scripts of already available software (Markham & Zuker, 2008; Zuker,

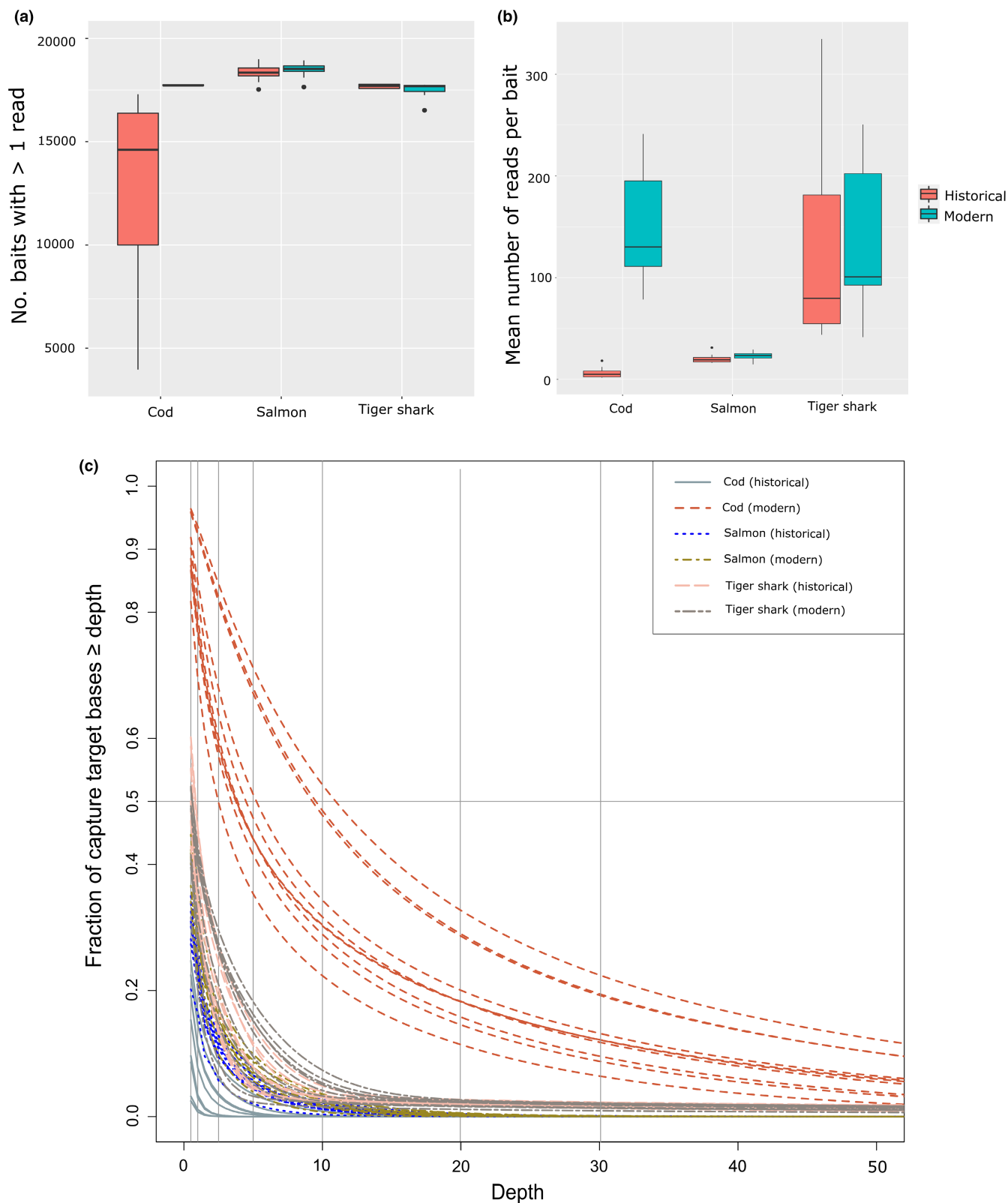


FIGURE 3 (a) Number of baits with more than one read on target, per species (x-axis) and category explored (modern and historical, y-axis). (b) Mean number of reads per bait, per species (x-axis) and category explored (modern and historical, y-axis). Black lines in (a) and (b) correspond to the median of the samples. (c) Cumulative distribution that describes the fraction of targeted bp covered by a certain number of reads (x-axis, represented by depth); each coloured line represents an individual from each population and category explored

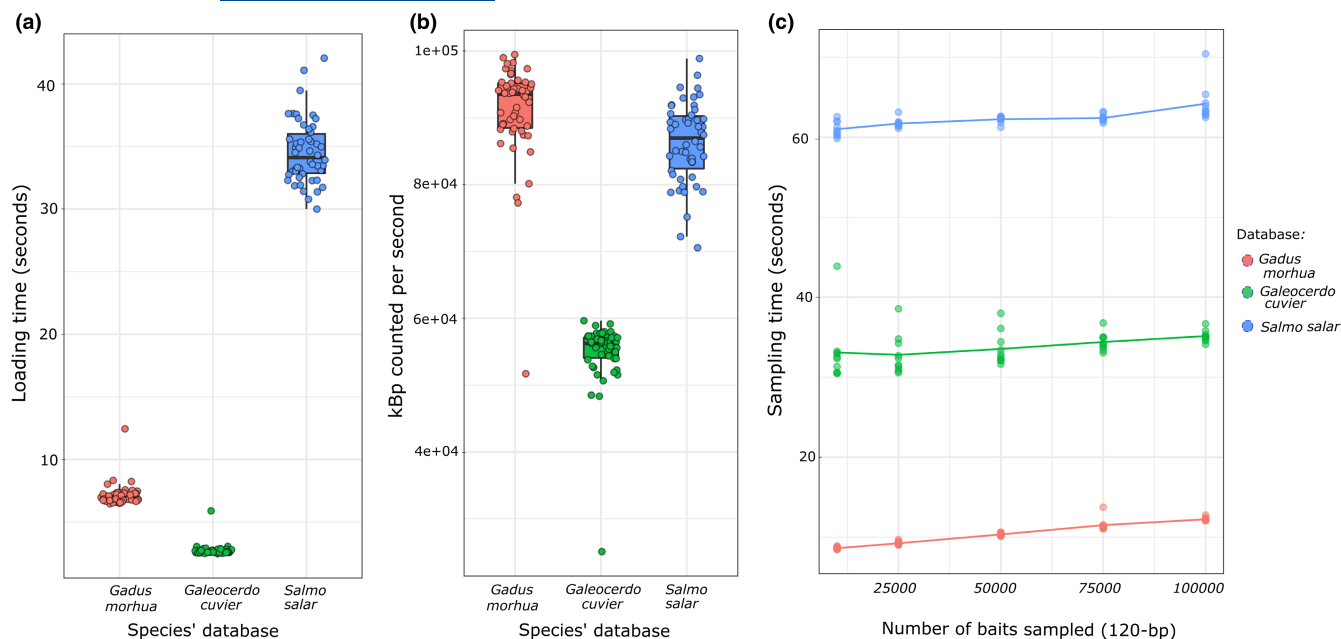


FIGURE 4 Analysis of the speed at which superBaits loads different genomic resources and retrieves baits. (a) Total time spent to import each of the three genomic databases (Atlantic cod, *Gadus morhua*; tiger shark, *Galeocerdo cuvier*, and Atlantic salmon, *Salmo salar*). (b) Average kBP counted per second for each of the genomic databases. (c) Total time required to choose bait locations and extract the respective number of baits from the genomic database, tested with basic conditions

Argument name	Description
<i>n</i>	Total number of desired baits
Size	Length (in bp) of each bait
Database	Genomic reference
<i>n_per_seq</i>	Number of baits per each sequence in the database
<i>min_per_seq</i>	Minimum number of baits per each sequence in the database
Exclusions	Areas of the database to exclude
Regions	Specific areas of the database to include
Regions.tiling	Choice of tiling for baits allocated in regions
Regions.prop	Proportion of baits allocated in regions
Targets	Specific points of the database to include (e.g., SNPs)
Targets.tiling	Choice of tiling for baits allocated in targets
Targets.prop	Proportion of baits allocated in targets
Seed	Seed to be set for a repeatable set of baits
Restrict	Areas of the database to restrict the baits to
gc	Wished range of the proportion of the nucleotides G and C within the bait area
force	Option to request a very large number of baits to be generated

TABLE 2 Main arguments of the superBaits main function

2003). Other options for further filtering are online tools such as SciTools Web Tools from IDT; or ArrayOligoSelector (Bozdech et al., 2003), and simulation programs (Cao et al., 2018) or external providers (e.g., Arbor Biosciences; Roche), to select the final set of baits. Therefore, by following the short pipeline of superBaits, large bait sets for population genomics can be generated with the desired bait properties and placement, in a fast and transparent way.

5 | CONCLUSION

Capture sequencing is a useful, cost effective tool to generate thousands of genomic markers for population genomics and conservation studies in non-model species. However, designing the baits necessary for a capture experiment is challenging, with few resources and guidelines available. Here, we present the first user-friendly R-package created specifically for bait design, superBaits,

as well as a discussion on the main parameters that influence the success of a DNA capture project for population genomics, with both contemporary and historical samples. We show that the method for designing baits that is implemented in *superBaits* facilitates fast, robust and efficient bait design. Our three described successful examples should be seen as proof of concept for the general practical applicability of *superBaits*. Although we did not discuss in great detail all the factors that might influence the success of a capture experiment (e.g., levels of endogenous DNA, quality of samples – but see Cruz-Dávalos et al., 2017), our guidelines contain key criteria regarding both the overarching experimental setup of a capture-based study, as well as the specific design of CS bait sets.

In conclusion, CS is a powerful approach for spatiotemporal population genetics, by providing flexibility to design panels of baits targeting a high number of specific genomic regions of interest. Bait sets can be adapted specifically to each species and research question, thus enabling researchers to make better use of the resources available. For this quest, *superBaits* is a fast and versatile tool for facilitating bait design.

BENEFIT-SHARING STATEMENT

All samples used in this manuscript are in compliance with national laws and the Nagoya Protocol.

ACKNOWLEDGEMENTS

We thank Maj-Britt Jacobsen, Britta Sønderskov Pedersen, Trine Rohde and Daniel Grundtvig Rosenstand Thomsen for their work at the laboratory processing the samples for the different studies, as well as James Henty Williams, Carl Hutter and two anonymous reviewers for useful comments on the manuscript. We thank Dominic Swift for access to the tiger shark transcriptome. We also thank Alison Devault and Brian Brunelle from Arbor Biosciences for clarifications on the MYBaits workflow and bait design. Figure 3c was created using a script developed by Stephen Turner (<https://gettinggeneticsdone.blogspot.com/2014/03/visualize-coverage-exome-targeted-ngs-bedto-ols.html>). GENOJAWS was funded by the Australian Research Grant (DP170102043) and The Danish Council for Independent Research Grant DFF (6108-00583). SDPAS project was funded by Innovationsfonden. CodStory was funded by the Icelandic Research Fund.

AUTHOR CONTRIBUTIONS

Designed the study: Belén Jiménez-Mena and Einar Eg Nielsen. *Performed the research:* Belén Jiménez-Mena, Hugo Flávio, Miguel Ramos, Alice Manuzzi, Romina Henriques, Dorte Meldrup, Janette Edson, Jennifer R. Ovenden, Snæbjörn Pálsson, and Guðbjörg Ásta Ólafsdóttir. *Developed superBaits:* Hugo Flávio, Miguel Ramos, and Belén Jiménez-Mena. *Analysed the data:* Belén Jiménez-Mena, Hugo Flávio, Miguel Ramos, Alice Manuzzi, and Romina Henriques. *Wrote the manuscript with suggestions from all authors:* Belén Jiménez-Mena and Einar Eg Nielsen.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

superBaits can be downloaded from <https://github.com/BelenJM/superBaits>. A tutorial can be found at <https://github.com/BelenJM/superBaits/wiki>. Supporting Information S1 (containing bait sets for Atlantic salmon, Atlantic cod and tiger shark) has been uploaded to Zenodo, with doi: <http://doi.org/10.5281/zenodo.5031556>, under Creative Commons Attribution 4.0 International.

ORCID

Belén Jiménez-Mena  <https://orcid.org/0000-0001-8458-5533>
Hugo Flávio  <https://orcid.org/0000-0002-5174-1197>
Romina Henriques  <https://orcid.org/0000-0002-6544-5532>
Alice Manuzzi  <https://orcid.org/0000-0001-8116-9710>
Miguel Ramos  <https://orcid.org/0000-0002-1211-4270>
Janette Edson  <https://orcid.org/0000-0002-1864-0095>
Snæbjörn Pálsson  <https://orcid.org/0000-0002-4297-3500>
Guðbjörg Ásta Ólafsdóttir  <https://orcid.org/0000-0002-2814-9160>
Jennifer R. Ovenden  <https://orcid.org/0000-0001-7538-1504>
Einar Eg Nielsen  <https://orcid.org/0000-0002-7009-9814>

REFERENCES

- Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, 202(2), 389–400. <https://doi.org/10.1534/genetics.115.183665>
- Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., Kistler, L., Liberal, I. M., Oxelman, B., Bacon, C. D., & Antonelli, A. (2020). A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics*, 10, 1407. <https://doi.org/10.3389/fgene.2019.01407>
- Ávila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., Fordyce, S. L., Montiel, R., Vielle-Calzada, J. P., Willerslev, E., & Gilbert, M. T. P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, 1(1), 1–5. <https://doi.org/10.1038/srep00074>
- Bailey, S. E., Mao, X., Struebig, M., Tsagkogeorga, G., Csorba, G., Heaney, L. R., Sedlock, J., Stanley, W., Rouillard, J.-M., & Rossiter, S. J. (2016). The use of museum samples for large-scale sequence capture: A study of congeneric horseshoe bats (family Rhinolophidae). *Biological Journal of the Linnean Society*, 117(1), 58–70. <https://doi.org/10.1111/bij.12620>
- Barney, B. T., Munkholm, C., Walt, D. R., & Palumbi, S. R. (2017). Highly localized divergence within supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine. *BMC Genomics*, 18(271). <https://doi.org/10.1186/s12864-017-3660-3>
- Barson, N. J., Aykanat, T., Hindar, K., Baranski, M., Bolstad, G. H., Fiske, P., Jacq, C., Jensen, A. J., Johnston, S. E., Karlsson, S., Kent, M., Moen, T., Niemelä, E., Nome, T., Næsje, T. F., Orell, P., Romakkaniemi, A., Sægvog, H., Urdal, K., ... Primmer, C. R. (2015). Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, 528(7582), 405–408. <https://doi.org/10.1038/nature16062>

- Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, 24(13), 3299–3315. <https://doi.org/10.1111/mec.13245>
- Berg, P. R., Jentoft, S., Star, B., Ring, K. H., Knutsen, H., Lien, S., Jakobsen, K. S., & André, C. (2015). Adaptation to low salinity promotes genomic divergence in Atlantic Cod (*Gadus morhua* L.). *Genome Biology and Evolution*, 7(6), 1644–1663. <https://doi.org/10.1093/gbe/evv093>
- Besnier, F., Glover, K. A., Lien, S., Kent, M., Hansen, M. M., Shen, X., & Skaala, Ø. (2015). Identification of quantitative genetic components of fitness variation in farmed, hybrid and native salmon in the wild. *Heredity*, 115(1), 47–55. <https://doi.org/10.1038/hdy.2015.15>
- Bi, K., Linderroth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., Moritz, C., & Good, J. M. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLOS Genetics*, 15(5), e1008119. <https://doi.org/10.1371/journal.pgen.1008119>
- Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., & DeRisi, J. L. (2003). Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biology*, 4(2), R9. <https://doi.org/10.1186/gb-2003-4-2-r9>
- Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, 16(5), 1059–1068. <https://doi.org/10.1111/1755-0998.12449>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10(1), 421. <https://doi.org/10.1186/1471-2105-10-421>
- Campana, M. G. (2018). BaitsTools: Software for hybridization capture bait design. *Molecular Ecology Resources*, 18(2), 356–361. <https://doi.org/10.1111/1755-0998.12721>
- Cao, M. D., Ganesamoorthy, D., Zhou, C., & Coin, L. J. M. (2018). Simulating the dynamics of targeted capture sequencing with CapSim. *Bioinformatics*, 34(5), 873–874. <https://doi.org/10.1093/bioinformatics/btx691>
- Capblancq, T., Butnor, J. R., Deyoung, S., Thibault, E., Munson, H., Nelson, D. M., Fitzpatrick, M. C., & Keller, S. R. (2020). Whole-exome sequencing reveals a long-term decline in effective population size of red spruce (*Picea rubens*). *Evolutionary Applications*, 13(9), 2190–2205. <https://doi.org/10.1111/eva.12985>
- Chafin, T. K., Douglas, M. R., & Douglas, M. E. (2018). MrBait: Universal identification and design of targeted-enrichment capture probes. *Bioinformatics*, 34(24), 4293–4296. <https://doi.org/10.1093/bioinformatics/bty548>
- Chan, K. O., Hutter, C. R., Wood, P. L., Grismer, L. L., Das, I., & Brown, R. M. (2020). Gene flow creates a mirage of cryptic species in a Southeast Asian spotted stream frog complex. *Molecular Ecology*, 29(20), 3970–3987. <https://doi.org/10.1111/MEC.15603>
- Chan, K. O., Hutter, C. R., Wood, P. L., Su, Y.-C., & Brown, R. M. (2021). Gene flow increases phylogenetic structure and inflates cryptic species estimations: A case study on widespread Philippine puddle frogs (*Occidozyga laevis*). *Systematic Biology*, 71(1), 40–57. <https://doi.org/10.1093/sysbio/syab034>
- Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., & Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, 15(1), 449. <https://doi.org/10.1186/1471-2164-15-449>
- Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., & Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, 29(10), 908–914. <https://doi.org/10.1038/nbt.1975>
- Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., & Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics*, 12(347), <https://doi.org/10.1186/1471-2164-12-347>
- Costa, L., Marques, A., Buddenhagen, C., Thomas, W. W., Huettel, B., Schubert, V., Dodsworth, S., Houben, A., Souza, G., & Pedrosa-Harand, A. (2021). Aiming off the target: Recycling target capture sequencing reads for investigating repetitive DNA. *Annals of Botany*, 128(7), 835–848. <https://doi.org/10.1093/aob/mcab063>
- Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A. S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., & Orlando, L. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources*, 17(3), 508–522. <https://doi.org/10.1111/1755-0998.12595>
- Dapprich, J., Ferriola, D., Mackiewicz, K., Clark, P. M., Rappaport, E., D'Arcy, M., Sasson, A., Gai, X., Schug, J., Kaestner, K. H., & Monos, D. (2016). The next generation of target capture technologies—Large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics*, 17(1), 486. <https://doi.org/10.1186/s12864-016-2836-6>
- de La Harpe, M., Hess, J., Loiseau, O., Salamin, N., Lexer, C., & Paris, M. (2019). A dedicated target capture approach reveals variable genetic markers across micro- and macro-evolutionary time scales in palms. *Molecular Ecology Resources*, 19(1), 221–234. <https://doi.org/10.1111/1755-0998.12945>
- Derkarabetian, S., Benavides, L. R., & Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Molecular Ecology Resources*, 19(6), 1531–1544. <https://doi.org/10.1111/1755-0998.13072>
- Ehlers, B. K., Gauthier, P., Villesen, P., Santoni, S., Thompson, J. D., & Bataillon, T. (2020). From genotype to phenotype: Maintenance of a chemical polymorphism in the context of high gene flow. *BioRxiv*. <https://doi.org/10.1101/2020.09.24.299651>
- Faircloth, B. C. (2017). Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution*, 8(9), 1103–1112. <https://doi.org/10.1111/2041-210X.12754>
- Förster, D. W., Bull, J. K., Lenz, D., Autenrieth, M., Pajmans, J. L. A., Kraus, R. H. S., Nowak, C., Bayerl, H., Kuehn, R., Saveljev, A. P., Sindičić, M., Hofreiter, M., Schmidt, K., & Fickel, J. (2018). Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. *Molecular Ecology Resources*, 18(6), 1356–1373. <https://doi.org/10.1111/1755-0998.12924>
- Fortes, G. G., Bouza, C., Martinez, P., & Sanchez, L. (2007). Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences. *Genetica*, 129, 281–289. <https://doi.org/10.1007/s10709-006-0009-2>
- Frandsen, P., Fontseré, C., Nielsen, S. V., Hanghøj, K., Castejon-Fernandez, N., Lizano, E., Hughes, D., Hernandez-Rodriguez, J., Korneliussen, T. S., Carlsen, F., Siegmund, H. R., Mailund, T., Marques-Bonet, T., & Hvilsom, C. (2020). Targeted conservation genetics of the endangered chimpanzee. *Heredity*, 125(1), 15–27. <https://doi.org/10.1038/s41437-020-0313-0>
- Franks, S. J., Kane, N. C., O'Hara, N. B., Tittes, S., & Rest, J. S. (2016). Rapid genome-wide evolution in Brassica rapa populations following drought revealed by sequencing of ancestral and descendant gene pools. *Molecular Ecology*, 25(15), 3622–3631. <https://doi.org/10.1111/mec.13615>
- Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, 8(11), 79667. <https://doi.org/10.1371/journal.pone.0079667>

- Giebner, H., Langen, K., Bourlat, S. J., Kukowka, S., Mayer, C., Astrin, J. J., Misof, B., & Fonseca, V. G. (2020). Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. *Molecular Ecology Resources*, 20(5), 1333–1345. <https://doi.org/10.1111/1755-0998.13201>
- Glenn, T. C., & Faircloth, B. C. (2016). Capturing Darwin's dream. *Molecular Ecology Resources*, 16(5), 1051–1058. <https://doi.org/10.1111/1755-0998.12574>
- Gooley, R. M., Tamazian, G., Castañeda-Rico, S., Murphy, K. R., Dobrynin, P., Ferrie, G. M., Haelele, H., Maldonado, J. E., Wildt, D. E., Pukazhenthi, B. S., Edwards, C. W., & Koepfli, K.-P. (2020). Comparison of genomic diversity and structure of sable antelope (*Hippotragus niger*) in zoos, conservation centers, and private ranches in North America. *Evolutionary Applications*, 13(8), 2143–2154. <https://doi.org/10.1111/EVA.12976>
- Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis1. *American Journal of Botany*, 99(2), 312–319. <https://doi.org/10.3732/ajb.1100323>
- Guo, Y., Long, J., He, J., Li, C.-I., Cai, Q., Shu, X.-O., Zheng, W., & Li, C. (2012). Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, 13(1), 1–10. <https://doi.org/10.1186/1471-2164-13-194>
- Hagemann, I. S., Cottrell, C. E., & Lockwood, C. M. (2013). Design of targeted, capture-based, next generation sequencing tests for precision cancer therapy. *Cancer Genetics*, 206(12), 420–431. <https://doi.org/10.1016/J.CANCERGEN.2013.11.003>
- Hale, H., Gardner, E. M., Viruel, J., Pokorny, L., & Johnson, M. G. (2020). Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences*, 8(4), e11337. <https://doi.org/10.1002/APS3.11337>
- Hancock-Hanser, B. L., Frey, A., Leslie, M. S., Dutton, P. H., Archer, F. I., & Morin, P. A. (2013). Targeted multiplex next-generation sequencing: Advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, 13(2), 254–268. <https://doi.org/10.1111/1755-0998.12059>
- Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Gordon, D. B., Brizuela, L., McCombie, W. R., & Hannon, G. J. (2009). Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, 4(6), 960–978. <https://doi.org/10.1038/nprot.2009.68>
- Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology*, 30(1), 62–82. <https://doi.org/10.1111/mec.15720>
- Horn, S. (2012). Target enrichment via DNA hybridization capture. In B. Shapiro, & M. Hofreiter (Eds.), *Ancient DNA: Methods and protocols* (pp. 177–188). Springer Science+Business Media. https://doi.org/10.1007/978-1-61779-516-9_21
- Hubert, S., Higgins, B., Borza, T., & Bowman, S. (2010). Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*, 11(1), 191. <https://doi.org/10.1186/1471-2164-11-191>
- Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An exon-capture system for the entire class ophiuroidea. *Molecular Biology and Evolution*, 33(1), 281–294. <https://doi.org/10.1093/molbev/msv216>
- Hutter, C. R., Cobb, K. A., Portik, D. M., Travers, S. L., Wood, P. L., & Brown, R. M. (2019). FrogCap: A modular sequence capture probe set for phylogenomics and population genetics for all frogs, assessed across multiple phylogenetic scales. *BioRxiv*, 825307. <https://doi.org/10.1101/825307>
- Jayaraman, P., Mosbrugger, T., Hu, T., Tairis, N. G., Wu, C., Clark, P. M., D'Arcy, M., Ferriola, D., Mackiewicz, K., Gai, X., Monos, D., & Sarmady, M. (2020). AnthOligo: Automating the design of oligo-nucleotides for capture/enrichment technologies. *Bioinformatics*, 36(15), 4353–4356. <https://doi.org/10.1093/bioinformatics/btaa552>
- Jiménez-Mena, B., Hospital, F., & Bataillon, T. (2016). Heterogeneity in effective population size and its implications in conservation genetics and animal breeding. *Conservation Genetics Resources*, 8(1), 35–41. <https://doi.org/10.1007/s12686-015-0508-5>
- Jiménez-Mena, B., Tataru, P., Brøndum, R. F., Sahana, G., Guldbrandtsen, B., & Bataillon, T. (2016). One size fits all? Direct evidence for the heterogeneity of genetic drift throughout the genome. *Biology Letters*, 12(7), 20160426. <https://doi.org/10.1098/rsbl.2016.0426>
- Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1), 185–202. <https://doi.org/10.1111/mec.13304>
- Joubran, S. S., & Cassin-Sackett, L. (2021). Genomic resources for an ecologically important rodent, Gunnison's prairie dogs (*Cynomys gunnisoni*). *Conservation Genetics Resources*, 13(2), 123–126. <https://doi.org/10.1007/S12686-021-01192-W>
- Kamneva, O. K., Syring, J., Liston, A., & Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, 17(1), 180. <https://doi.org/10.1186/s12862-017-1019-7>
- Karlsson, S., Moen, T., Lien, S., Glover, K. A., & Hindar, K. (2011). Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, 11(Suppl. 1), 247–253. <https://doi.org/10.1111/j.1755-0998.2010.02959.x>
- Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino, B., Johansen, T., Otterå, H., Sonesson, A., Lien, S., & Andersen, Ø. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, 25(10), 2130–2143. <https://doi.org/10.1111/mec.13592>
- Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, 61(5), 727–744. <https://doi.org/10.1093/sysbio/sys049>
- Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., ... Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533(7602), 200–205. <https://doi.org/10.1038/nature17164>
- Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23), 5966–5993. <https://doi.org/10.1111/mec.16077>
- Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2), 111–118. <https://doi.org/10.1038/nmeth.1419>
- Manuzzi, A., Jiménez-Mena, B., Henriques, R., Holmes, B. J., Pepperell, J., Edson, J., Bennett, M. B., Huvneers, C., Ovenden, J. R., & Nielsen, E. E. (2021). Retrospective genomics suggests the disappearance of a tiger shark (*Galeocerdo Cuvier*) population off South-Eastern Australia. *Research Square*. <https://doi.org/10.21203/rs.3.rs-334053/v1>
- Marandel, F., Charrier, G., Lamy, J., Le Cam, S., Lorange, P., & Trenkel, V. M. (2020). Estimating effective population size using RADseq: Effects of SNP selection and sample size. *Ecology and Evolution*, 10(4), 1929–1937. <https://doi.org/10.1002/ece3.6016>

- Markham, N. R., & Zuker, M. (2008). UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular biology*, 453, 3–31. https://doi.org/10.1007/978-1-60327-429-6_1
- Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., Petersen, M., Meusemann, K., Lier, K., Wägele, J. W., Misof, B., Bleidorn, C., Ohl, M., & Niehuis, O. (2016). BaitFisher: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, 33(7), 1875–1886. <https://doi.org/10.1093/molbev/msw056>
- McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, 16(5), 1084–1094. <https://doi.org/10.1111/1755-0998.12538>
- Meek, M. H., & Larson, W. A. (2019). The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources*, 19(4), 795–803. <https://doi.org/10.1111/1755-0998.12998>
- Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K. T., Fevolden, S.-E., Berg, P. R., & Lien, S. (2008). Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics*, 9(1), 18. <https://doi.org/10.1186/1471-2156-9-18>
- Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Molecular Ecology Resources*, 17(6), 1136–1147. <https://doi.org/10.1111/1755-0998.12654>
- Nielsen, E. E., Hemmer-Hansen, J., Poulsen, N. A., Loeschcke, V., Moen, T., Johansen, T., Mittelholzer, C., Taranger, G.-L., Ogden, R., & Carvalho, G. R. (2009). Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, 9(1), 276. <https://doi.org/10.1186/1471-2148-9-276>
- Nielsen, E. E., Morgan, J. A. T., Maher, S. L., Edson, J., Gauthier, M., Pepperell, J., Holmes, B. J., Bennett, M. B., & Ovenden, J. R. (2017). Extracting DNA from “jaws”: High yield and quality from archived tiger shark (*Galeocerdo cuvier*) skeletal material. *Molecular Ecology Resources*, 17(3), 431–442. <https://doi.org/10.1111/1755-0998.12580>
- Noguera, D. R., Wright, E. S., Camejo, P., & Safak Yilmaz, L. (2014). Mathematical tools to optimize the design of oligonucleotide probes and primers. *Applied Microbiology and Biotechnology*, 98, 9595–9608. <https://doi.org/10.1007/s00253-014-6165-x>
- Picardi, E., & Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. *Nature Methods*, 9(6), 523–524. <https://doi.org/10.1038/nmeth.2029>
- Puritz, J. B., & Lotterhos, K. E. (2018). Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Molecular Ecology Resources*, 18(6), 1209–1222. <https://doi.org/10.1111/1755-0998.12905>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Russell, J. J., Theriot, J. A., Sood, P., Marshall, W. F., Landweber, L. F., Fritz-Laylin, L., Polka, J. K., Oliferenko, S., Gerbich, T., Gladfelter, A., Umen, J., Bezanilla, M., Lancaster, M. A., He, S., Gibson, M. C., Goldstein, B., Tanaka, E. M., Hu, C. K., & Brunet, A. (2017). Non-model model organisms. *BMC Biology*, 15(1), 1–31. <https://doi.org/10.1186/s12915-017-0391-5>
- Samuels, D. C., Han, L., Li, J., Quangu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, 29(10), 593–599. <https://doi.org/10.1016/j.TIG.2013.07.006>
- Sanderson, B. J., DiFazio, S. P., Cronk, Q. C. B., Ma, T., & Olson, M. S. (2020). A targeted sequence capture array for phylogenetics and population genomics in the Salicaceae. *Applications in Plant Sciences*, 8(10). <https://doi.org/10.1002/APS3.11394>
- Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J. L., Camps, C., Pentony, M. M., Kvikstad, E. M., Harris, S., Popitsch, N., Pagnamenta, A. T., Schuh, A., Taylor, J. C., & Wordsworth, S. (2020). The complete costs of genome sequencing: A microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, 22(1), 85–94. <https://doi.org/10.1038/s41436-019-0618-7>
- Singhal, S., Grundler, M., Colli, G., & Rabosky, D. L. (2017). Squamate conserved loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Molecular Ecology Resources*, 17(6), e12–e24. <https://doi.org/10.1111/1755-0998.12681>
- Star, B., Tørresen, O. K., Nederbragt, A. J., Jakobsen, K. S., Pampoulie, C., & Jentoft, S. (2016). Genomic characterization of the Atlantic cod sex-locus. *Scientific Reports*, 6, 31235. <https://doi.org/10.1038/srep31235>
- Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, 19(1), 1–12. <https://doi.org/10.1186/s13059-018-1520-3>
- Swift, D. G., Dunning, L. T., Igea, J., Brooks, E. J., Jones, C. S., Noble, L. R., Ciezarek, A., Humble, E., & Savolainen, V. (2016). Evidence of positive selection associated with placental loss in tiger sharks. *BMC Evolutionary Biology*, 16(1), 126. <https://doi.org/10.1186/s12862-016-0696-y>
- Therkildsen, N. O., Hemmer-Hansen, J., Als, T. D., Swain, D. P., Morgan, M. J., Trippel, E. A., Palumbi, S. R., Meldrup, D., & Nielsen, E. E. (2013). Microevolution in time and space: SNP analysis of historical DNA reveals dynamic signatures of selection in Atlantic cod. *Molecular Ecology*, 22(9), 2424–2440. <https://doi.org/10.1111/mec.12260>
- Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, 17(2), 194–208. <https://doi.org/10.1111/1755-0998.12593>
- Tørresen, O. K., Star, B., Jentoft, S., Reinart, W. B., Grove, H., Miller, J. R., Walenz, B. P., Knight, J., Ekholm, J. M., Peluso, P., Edvardsen, R. B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K. S., & Nederbragt, A. J. (2017). An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, 18(1). <https://doi.org/10.1186/s12864-016-3448-x>
- Turkulov, R., & Easteal, S. (2003). Number of SNPs loci needed to detect population structure. *Human Heredity*, 55(1), 37–45. <https://doi.org/10.1159/000071808>
- Vinogradov, A. E. (2001). Bendable genes of warm-blooded vertebrates. *Molecular Biology and Evolution*, 18(12), 2195–2200. <https://doi.org/10.1093/oxfordjournals.molbev.a003766>
- Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: Current and future perspectives. *G3: Genes, Genomes, Genetics*, 5(8), 1543–1550. <https://doi.org/10.1534/g3.115.018564>
- Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society B*, 272, 3–16. <https://doi.org/10.1098/rspb.2004.2813>
- Woischnik, M., & Moraes, C. T. (2002). Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Research*, 12(6), 885–893. <https://doi.org/10.1101/gr.227202>
- Zhou, L., & Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, 13(1), 1–12. <https://doi.org/10.1186/1471-2164-13-703>

- Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., & Primmer, C. R. (2014). Footprints of directional selection in wild Atlantic Salmon populations: Evidence for parasite-driven evolution? *PLoS One*, 9(3), e91672. <https://doi.org/10.1371/journal.pone.0091672>
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406–3415. <https://doi.org/10.1093/nar/gkg595>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jiménez-Mena, B., Flávio, H., Henriques, R., Manuzzi, A., Ramos, M., Meldrup, D., Edson, J., Pálsson, S., Ásta Ólafsdóttir, G., Ovenden, J. R., & Nielsen, E. E. (2022). Fishing for DNA? Designing baits for population genetics in target enrichment experiments: Guidelines, considerations and the new tool supeRbaits. *Molecular Ecology Resources*, 22, 2105–2119. <https://doi.org/10.1111/1755-0998.13598>