BMC Genomics

RESEARCH Open Access

Fine-tuning GBS data with comparison of reference and mock genome approaches for advancing genomic selection in less studied farmed species

Daniel Fischer[1*], Miika Tapio[2], Oliver Bitz[2], Terhi Iso-Touru[2], Antti Kause[2] and Ilma Tapio[2]

Abstract

Background Diversifying animal cultivation demands efficient genotyping for enabling genomic selection, but non-model species lack efficient genotyping solutions. The aim of this study was to optimize a genotyping-by-sequencing (GBS) double-digest RAD-sequencing (ddRAD) pipeline. Bovine data was used to automate the bioinformatic analysis. The application of the optimization was demonstrated on non-model European whitefish data.

Results DdRAD data generation was designed for a reliable estimation of relatedness and is scalable to up to 384 samples. The GBS sequencing yielded approximately one million reads for each of the around 100 assessed samples. Optimizing various strategies to create a de-novo reference genome for variant calling (mock reference) showed that using three samples outperformed other building strategies with single or very large number of samples. Adjustments to most pipeline tuning parameters had limited impact on high-quality data, except for the identity criterion for merging mock reference genome clusters. For each species, over 15k GBS variants based on the mock reference were obtained and showed comparable results with the ones called using an existing reference genome. Repeatability analysis showed high concordance over replicates, particularly in bovine while in European whitefish data repeatability did not exceed earlier observations.

Conclusions The proposed cost-effective ddRAD strategy, coupled with an efficient bioinformatics workflow, enables broad adoption of ddRAD GBS across diverse farmed species. While beneficial, a reference genome is not obligatory. The integration of Snakemake streamlines the pipeline usage on computer clusters and supports customization. This user-friendly solution facilitates genotyping for both model and non-model species.

Keywords Genotyping by sequencing, Snakemake, Variant calling, Cattle, Aquaculture, Repeatability

*Correspondence:
Daniel Fischer
daniel.fischer@luke.fi
[1]Applied Statistical Methods, Natural Resources, Natural Resources Institute Finland (Luke), Jokioinen 31600, Finland
[2]Genomics and Breeding, Production Systems, Natural Resources Institute Finland (Luke), Jokioinen 31600, Finland

Background

Humans have successfully domesticated over five hundred animal species, and the number of newly cultivated species has been increasing by at least ten species per year [1, 2]. Particularly in recently domesticated species, our understanding of their genetic diversity and the genetic basis of traits may be insufficient. Genome wide data and genomic selection have revolutionized animal breeding by improving productivity [3–5], as well as incorporating health and welfare traits [6, 7]. In genomic selection, thousands of DNA markers are used to predict the genomic breeding value of an individual [8, 9], but genotyping presents a significant challenge for rare or novel production species. A recent review of genome data [10] revealed that nearly half of the aquaculture species, with an annual production exceeding 350 million kg [11], lack reference genome information, which together with genetic polymorphism characterization is a necessary resource for the development of commercial SNP-chip platforms or targeted genotyping-by-sequencing solutions. Therefore, it is crucial to make cost-effective and reliable alternative genotyping methods widely available for non-model organisms to advance genomic selection and stock management in niche production species.

The advantage of genome-assisted breeding value estimation largely stems from reliable estimation of relationships [12] and a common genomic selection approach is directly based on the genomic relationship matrix (GRM), which estimates the proportion of the genome shared identical by descent between pairs of individuals. Unlike in typical genome focused situations, where understanding the functional genomic basis of traits is essential, here a genomic map or a reference genome is not required and the method performs well even with low marker densities (10 SNPs per morgan) [13]. However, additional markers are beneficial and, for example, in Atlantic salmon, densities up to 50 to 200 markers per morgan (1 000 to 5 000 markers in total) have been recommended [4, 14]. The accuracy and cost-effectiveness of genomic selection depend on the balance between the number of genotyped markers and individuals, with marker numbers of 1 000 to 2 000 SNPs being suggested [15].

Choosing the genotyping methodology for practical production or breeding purposes requires balancing needs, costs, precision and time pressure [16]. Especially, when supporting genomic information may be inadequate, genotyping-by-sequencing (GBS) [17] is a cost-effective approach for simultaneous genome-wide SNP discovery and genotyping. Restriction-site associated DNA sequencing (RAD) [18–20] and double-digest RAD-sequencing (ddRAD) [21, 22] are reduced-representation genome sequencing methods that target a small portion of the genome using restriction enzymes. These methods do not provide as dense information as low-pass sequencing based GBS methods [16], but enable rapid generation of data and can generate sequencing-libraries from hundreds to hundreds of thousands of fragments genome wide. Both wet lab protocols and parameters used in post-sequencing analysis impact the number of recovered reads, mean sequencing target coverage, recovered genetic loci/marker, and genotype completeness and accuracy [21]. While the number of polymorphic markers is the main concrete criterion for evaluating the suitability of a genotyping method for genomic selection, the actual genotyping goal of reliable estimation of relatedness might be influenced by the minor allele frequencies (MAF), codominant or Mendelian inheritance and repeatability. GBS variants typically have a lower call rate per sample and repeatability among sample sets compared to SNP arrays. Additionally, genotyping errors, especially allelic dropouts (as false homozygotes), can introduce bias in the relatedness estimates used in genomic selection [23]. However, optimized GBS pipelines can exhibit high consistency with SNP-chip data [24]. While optimizing data generation has been widely assessed [25–27], finetuning the bioinformatic flow has gotten less attention. In the context of GBS analysis, parameter selection is particularly critical, as it significantly influences the results obtained from the pipeline. Our contribution emphasizes the importance of identifying optimal parameter sets and demonstrates how parameter fine-tuning can lead to improved and reliable outcomes. This principle is independent of the specific pipeline employed but underscores that GBS workflows are highly sensitive to parameter choices. As such, contrary to common practice, these workflows should not be used blindly with default settings but should always be tailored to achieve robust and accurate results. Following this approach, we extended the existing GBS-SNP-CROP workflow [28] into a publicly available Snakemake pipeline, called Snakebite-GBS [29]. This pipeline is designed to be versatile, functioning effectively for organisms with existing reference genome as well as for those that require building a de-novo / mock reference genome. Additionally, it prioritizes user-friendliness by minimizing software installation requirements by making use of containerization while ensuring reproducibility of results.

Besides here presented pipeline, there are other implementations available to call variants from ddRAD data. Other well-known software suites are TASSEL-GBS [30] and Stacks [31], which is also available wrapped within a Snakemake workflow and is listed inside the Snakemake workflow catalogue. A popular python software is ipyrad [32], which superceded pyrad [33] and makes use of Jupyter notebooks. Another, bash-based wrapper for various steps of ddRAD analysis is dDocent [34]. There is

also a Nextflow [35] workflow within the nf-core [36, 37] framework, called nf-core/radseq, available.

Previously, other studies compared ddRAD data with WGS [38] or with SNPChip data [39] and found that ddRAD based SNP calls are comparable with SNPChip and WGS based called variants. Also in comparison to low density and HD chips a similar observation was made, showing that the ddRAD is an interesting approach compared to aforementioned methods [40]. Another publication [41] investigated best filtering strategies for SNP for RAD data, combined with the advice to always make the raw data available together with the final filtered datasets, while also tools for the optimal design of ddRAD studies were developed [26].

When off-the-shelf SNP genotyping has not been available, ddRAD/RAD has been an effective method in several aquaculture species as the first step to study the genomic determination of the traits and structure of populations [42–45], and a similar method is needed for European whitefish (*Coregonus lavaretus* L). European whitefish is the second most important farmed fish species in Finland, and a breeding programme is used to improve production, quality and fish health traits [46, 47]. It is also a species used in ecological studies and it is known to have undergone widespread phylogeographic structuring and the repeated evolution of distinct ecological ecotypes [48]. Genotyping by sequencing has also been implemented in cattle [49], and in absence of reference genome the use of sequence tags as dominant markers was an early solution [17]. The primary objective of this study was to optimize the GBS method ddRAD and fine-tune the bioinformatic pipeline parameters for processing and controlling of the high-quality SNP data for genomic selection in non-model species. Here, our aim was to demonstrate how to identify optimal parameter settings to find a trade-off between a good yield and high-quality SNPs. The second objective was to test the repeatability of the data generation, which is, given the complex nature of the European whitefish genome, not granted. We fine-tuned the bioinformatics pipeline parameters by utilizing dairy cattle GBS and whole-genome resequencing (WGS) data. Following this, we applied the established data processing routines on data generated for European whitefish (*Coregonus lavaretus* L) using the available reference genome of the closest relative *Coregonus supersum* 'balchen' [50]. European whitefish is the second most important farmed fish species in Finland [46, 47]. It is also a species used in ecological studies and it is known to have undergone widespread phylogeographic structuring and the repeated evolution of distinct ecological ecotypes [48]. The overarching objective was to make the GBS method simpler to use across diverse species, eliminating the need for extensive bioinformatics expertise or specialized units. This advancement holds the potential to enhance genomic selection and refine animal breeding practices, particularly within less studied species.

Results

Restriction enzyme selection in silico

The expected sequencing library composition was simulated using SimRAD [27] focusing on five restriction enzyme pairs used in other species. The numbers of double digested genome fragments within the suitable range of 150–400 bp and consequently the expected variant numbers were three to four times more strongly influenced by the choice of the enzyme pair than by the species assessed (Fig. 1). The predicted fragment numbers fulfilled the preset criteria for all enzyme pairs, the number of fragments being the lowest for the EcoRI; SphI pair, with approximately 25–50 thousand fragments (or 20–40 thousand estimated variants). The reference genome based fragment numbers for the two main targets, *Bos taurus (*ARS-UCD1.2), and *Coregonus supersum* (AWG_v2), were for the pair EcoRI; SphI 50 000 and 30 000, for the pair EcoRI; MspI 120 000 and 110 000, for the pair MluCI; SphI 270 000 and 230 000, for the pair EcoRI; MseI 380 000 and 180 000, for the pair EcoRI; NlaIII 440 000 and 200 000, respectively. The predicted fragment number for the EcoRI; SphI pair was within the desired range of 10 000–100 000 fragments, which was expected to provide a minimum of 5 000 relatedness informative variants. Moreover, this enzyme pair provided the most uniform distribution of fragments across the size range, reducing the size selection lab protocol choice to the decision of window width (Fig. 1). The EcoRI; SphI pair was the most optimal for all the currently assessed species.

Raw GBS and WGS sequencing data

Data was generated using the modified ddRAD method [51] with EcoRI-HF and SphI-HF restriction enzymes on NextSeq550. GBS sequencing of 36 cow libraries generated in total 43 109 115 PE reads of 2 × 75 bp in length, with an average of 1 197 475 PE reads per sample. After trimming, 39 730 518 PE reads remained (avg: 1 103 625 reads per sample). Sample details are listed in Table S1. In case of the 66 whitefish libraries sequenced, from the total of 78 577 269 PE reads of 2 × 75 bp in length (avg: 1 190 565 reads per sample) 71 655 413 reads passed the quality control trimming (avg: 1 086 688 reads per sample). After quality control, the average read length dropped to 66 bp for reads R1 and 60 bp for reads R2.

WGS sequencing of 12 cow samples generated in total 3 918 912 122 PE reads of 2 × 150 bp in length, with an average of 326 659 344 PE reads per sample. After trimming, 3 865 355 653 PE reads remained with average of 322 112 971 reads per sample.

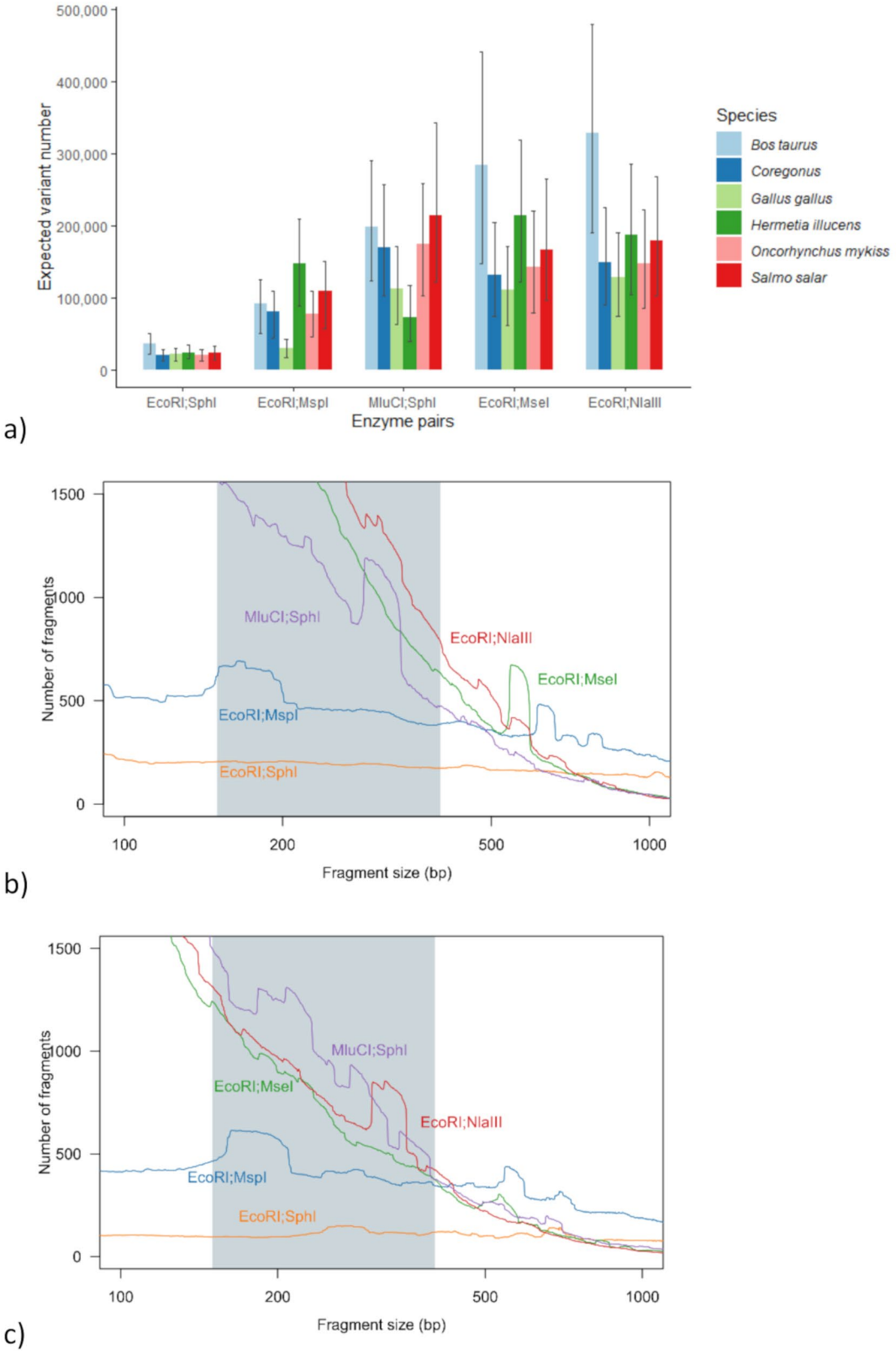

Fig. 1 In silico comparison of enzyme pairs. Expected variant numbers across the species and assessed restriction enzyme pairs (**a**), where whiskers indicate the impact of symmetric widening or narrowing the fragment size range by 100 bp. The predicted frequency distribution of double digested template fragments of different sizes in *Bos taurus* (**b**) and *Coregonus sp.* (**c**) averaged over 50 bp window across fragment sizes from 100 to 1 000 bp. The grey area denotes the included size range (150–400 bp)

GBS fragment recovery

The mapping of the quality-trimmed GBS derived cow data against the non-size selected in-silico (EcoRI; SphI) digested *Bos taurus* (ARS-UCD1.2) reference genome was done with BWA-mem [52] and indicated that about 86% of the reads aligned to fragments within the 150–400 bp size range (Fig. 2). This alignment window was narrower than the expected full insert size range of 150–550 bp. The in-silico digestion simulation generated in total 66 450 genome fragments between 150 and 400 bp in length. Considering that the remaining 14% of the reads were outside this span, our mock reference was expected to have between 66 450 and 79 100 clusters.

Mock reference quality

The construction of a mock reference was done with *vsearch* [53], using reads that are, if possible, merged together with *PEAR* [54] or if no overlap is present, stitched together with a sequence of 20 N, relies on the defined data and parameter configurations. An evaluation against the size-selected in-silico digested reference, that was obtained with a tailored *SimRAD* script, measuring average coverage percentages and secondary alignments (Figure S2), unveiled an over-inflation of the mock reference when utilizing all samples, resulting in the exclusion of mock-strategy 4. While focusing on one sample (mock-strategy 1 and 2) approximated the optimal cluster counts, it introduces the risk of sample-specific biases in the mock reference. As a result, mock-strategy 3 emerged as the preferred choice. However, its advantage over mock-strategy 4 was reduced by the final mock refinement step, which curbed most of the excessive cluster inflation, as indicated by consistent alignment trends nearing the expectation value (Figure S2, gray box).

Adjustments to input data parameters had minimal impact on the mock reference. PE read merging using *p*-value thresholds (0.001, 0.01, 0.05) yielded consistent mock reference lengths and alignment percentages against the in-silico reference. Around 99.8% of the mock clusters aligned with minimap2 [55] against the reference genome, accompanied by a modest number of unaligned clusters (417–900). Mock cluster counts and secondary alignments remained stable. Parameter pl (min. merged cluster length) showed negligible impact across reasonable values, aligning with expectations. Cluster generation parameters, especially the nucleotide similarity parameter (id), had, however, significant influence. Its extreme values led to drastic changes in the merged cluster numbers, while moderate values (e.g., 0.85) yielded expected alignments. The minimum cluster length (min) and read stitching optimization (rl) parameters had limited impact. Optimal parameters for the mock reference creation were $p = 0.05$, pl = 50, id = 0.85, min = 80 and rl = 75 (Figure S3).

For the mock refinement step, strict parameters (e.g., average 10 reads per sample per cluster, ≥ 10 samples with aligned reads on cluster) appeared optimal for a stable variant set creation. Refined mock references exhibited improved alignment against the *Bos taurus* (ARS-UCD1.2) reference genome (dashed-line), although the average sample-wise alignment of data against the

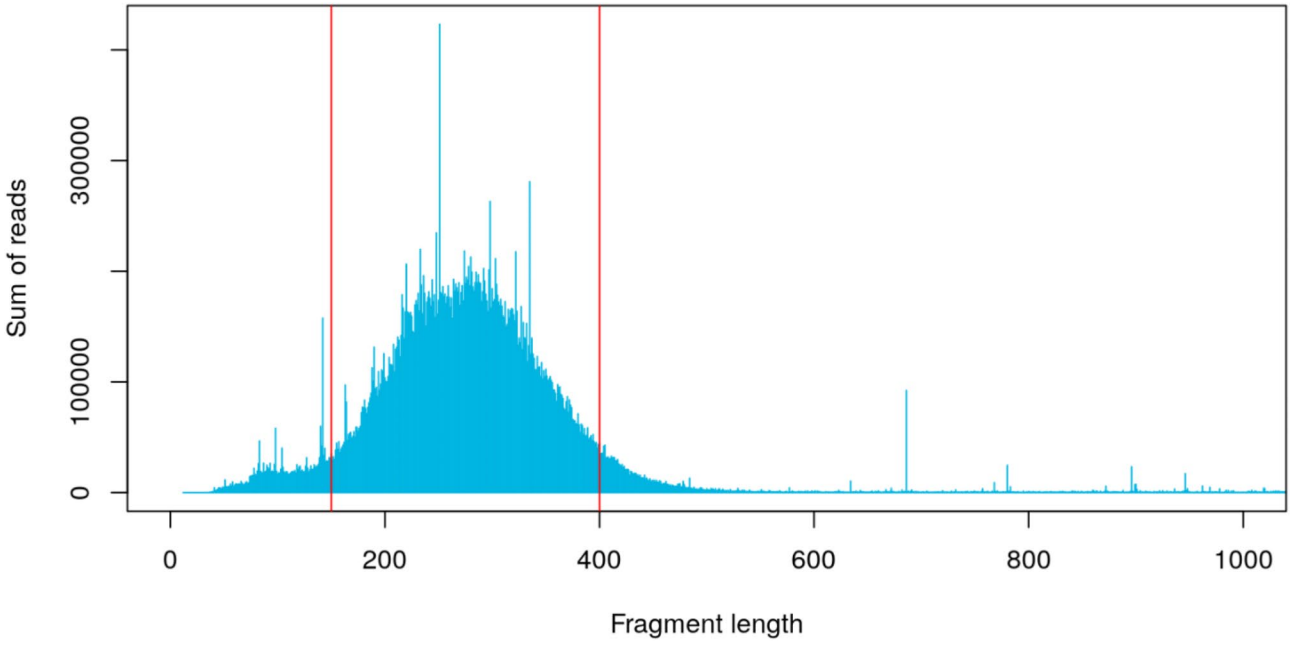

Fig. 2 Distribution of quality-trimmed cow GBS reads across in-silico digested *Bos taurus* (ARS-UCD1.2) reference genome fragment lengths. Red vertical lines indicate the boundaries of the estimated effective fragment size

mock reference was slightly decreased for the refined mock compared to pre-refinement mock (Figure S4).

Variant calling and GBS quality estimation

Applying the GATK best practice variant calling pipeline to the full genome WGS data produced in total 17 376 716 variants for the cow samples, with 42 160 variants intersecting regions on the reference genome that had a minimum coverage of three reads from the GBS data from at least 10 samples. Aligning GBS data to the reference genome (ARS-UCD1.2) resulted, after similar filtering, in 20 794 variants. Calling variants using the pre-refinement mock reference, based on mock strategy 3, yielded 16 404, and with refinement, 16 416 variants. In the case of GBS, we obtained a MAF of 0.26 (sd: 0.13) using the mock reference and 0.27 (sd: 0.14) while using the reference genome. The average call rate using the GBS approach in combination with the ARS-UCD1.2 reference genome was 94.8%, with average 11.38 (sd: 0.75) samples per variant, respectively 11.37 (sd: 0.76) with using the created mock reference genome. For the WGS, we observed for the 42 160 variants a MAF of 0.21 (sd: 0.14) with a call rate of 99.9% with 11.99 (sd: 0.13) samples having called each variant on average.

The overlap of reference based GBS and WGS variant sets, defined by their chromosomal positions, comprised 18 196 loci, representing approximately 87.5% alignment between the GBS and WGS datasets. These variants exhibited a WGS-based MAF of 0.26 (sd: 0.13) and nearly 100% call rate (sd: 0.05). On a chromosomal level, GBS-set missingness ranged from 9 to 15%, with a notable exception of the X-chromosome displaying over 30% missingness (Figure S5). Sample-wise genotype concordance between GBS and WGS data ranged from 82.6 to 97.5% (mean: 93.3%). A mere 1.3% of GBS-called homozygous variants were identified as heterozygous in the WGS dataset, and only 0.2% of heterozygous GBS variants were classified as homozygous in the WGS dataset. In total, 2 598 (12.5%) GBS variants were exclusive to the GBS call set, while 23 964 (56.8%) WGS variants were absent from the GBS (Table S2) variants.

Evaluating GBS based variant data for its ability to recover the realized relatedness matrix derived from >10 million bovine SNPs in the full genome data using the R-package BGData showed a convergence of both. With approximately 1 000 variants the matrices approach equivalence, as indicated by the eigenvalue distance dropping from >1 to approximately 0.15 (Fig. 3). After this point, the GBS genotype-based matrices exhibited a slower convergence compared to the WGS-based counterpart. Results suggested that about 5 000 GBS markers equate to 2 000 WGS-derived SNP markers, fulfilling genomic selection needs.

Proof of concept using non-model European whitefish species as an example

The European whitefish mock reference created by strategy 3, following the optimized mock creation parameters, was comprised of 159 403 clusters, spanning around 26 million bp, and suggested an average 4x – 8x fold read coverage. While shallow sequenced samples exhibited low coverage (4x), most samples demonstrated acceptable coverage (8x) against the created mock reference. Aligning the mock reference to the *Coregonus sp.* 'balchen' reference genome (AWG_v2) resulted in a coverage of 34 million bp due to multiple mapping, with alignment rates around 90% for quality-filtered PE reads against the mock reference and slightly higher (91%) against the AWG_v2 reference genome.

Using an in-silico prediction for a 150–400 bp fragment size threshold led to 28 085 fragments and an approximate 80% alignment rate against this reference. Employing the mock reference facilitated calling 18 678 GBS variants, with a stable missingness below 5–7% for samples with over 1 million reads. Similarly, the existing reference genome enabled calling 23 275 GBS variants with a comparable stable missingness.

Genomic relatedness estimates between parent and offspring in whitefish trios averaged 0.53 (ranging 0.47–0.57) with the AWG_v2 reference genome data, and 0.49 (0.43–0.54) with the mock reference data aligning with the expectations [56]. Respectively, genomic relatedness among the parental fish averaged 0.09 ranging from −0.05 to 0.53 or averaging 0.08 and ranging from −0.04 to 0.49. Unrelated fish exclusively formed mated pairs (all relatedness estimates < 0.05), aligning with expectations. Rare non-Mendelian inheritance, consistent across families, occurred in 3.3% (333.2 GBS variants on average) of the loci variable within the trios using AWG_v2 reference genome data and 3.4% (263.8 GBS variants on average) with mock reference data. Repeated Mendelian errors shared among loci were slightly smaller in the reference genome data (14.0%, 202 variants) compared to the mock reference data (14.8%, 167 variants). Both data sets exhibited similar estimates with a maximum absolute relatedness difference of 0.045 and generally agreed with prior pedigree knowledge.

Repeatability

The repeatability assessment, i.e. assessment of intersections between data sets, were done using bcftools and in bovine encompassed three separate runs: two utilizing 250 ng DNA (Orig- and RepI-set) and one employing 500 ng DNA (RepII-set) as starting material. All three sets underwent the same wet lab and optimized bioinformatic protocol using the ARS-UCD1.2 reference genome. The initial pipeline optimization run for the Orig-set yielded 20 794 GBS variants while the RepI-set and the RepII-set

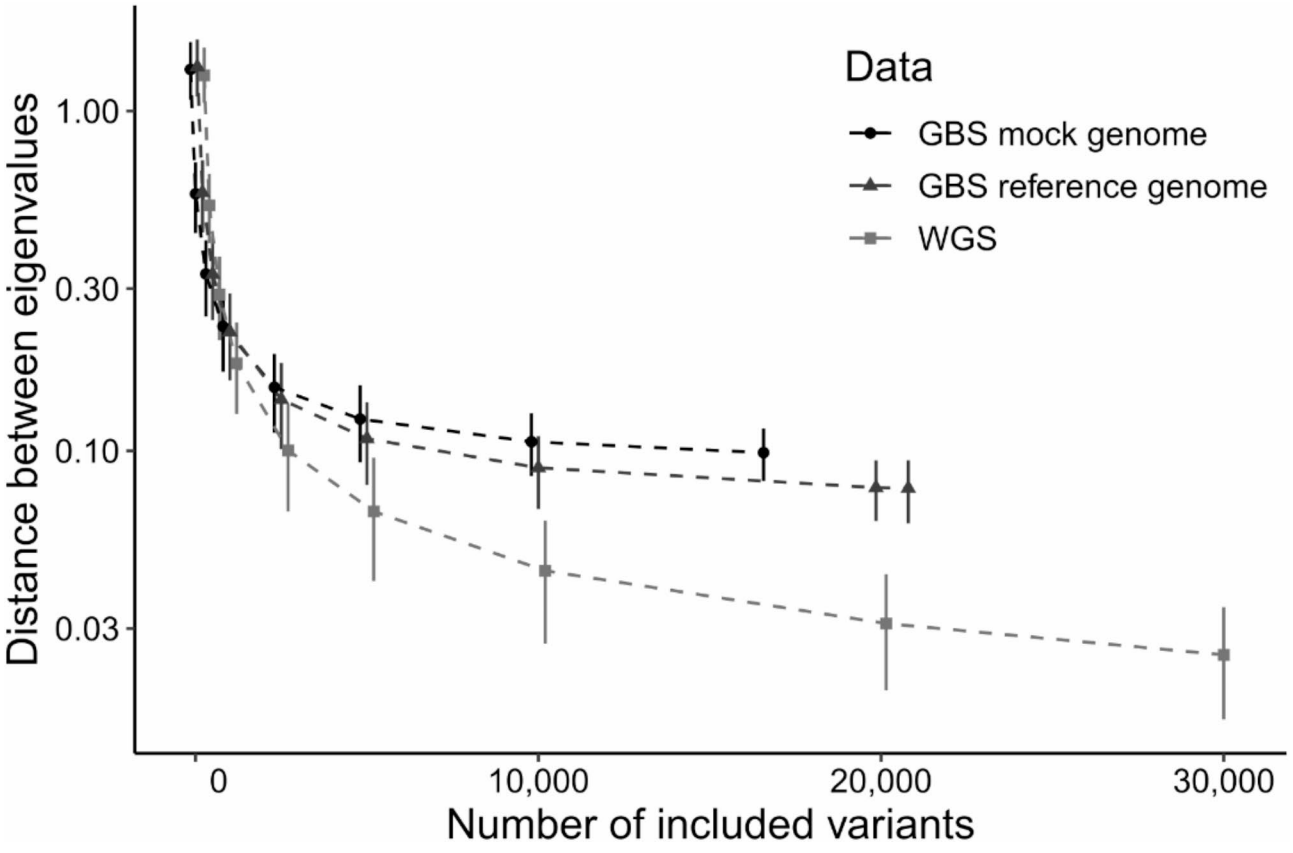

Fig. 3 Evolution of eigenvalue distances as a function of the number of utilized DNA variants. The plot compares the distance between GRM matrix based on all whole genome sequence (WGS) derived variants and smaller variant subsamples from mock/reference GBS or WGS data. The plot displays the mean and 90% confidence intervals, generated from 1 000 bootstrapped resampling. Variant counts range from 50 to 30 000, encompassing the full GBS sample sets. The Y-axis is log-transformed to enhance visibility of differences

produced 19 066 and 19 988 GBS variants, respectively. Analyzing variant locations revealed a high degree of shared loci, with the RepI-set displaying 16 559 (79.6%) shared variants, and the RepII-set exhibiting 17 459 (84.0%) shared variants. Remarkably, the two repeated runs shared 16 556 variants in common, resulting in a cumulative sharing of 15 246 (73.3%) variants across all three runs (Fig. 4a).

Within the whitefish dataset, a repeatability analysis encompassing two distinct scenarios for a subset of 12 samples was performed. The first scenario involved technical replicates of identical libraries (Orig-set and Tech-set). In the second scenario, duplicate libraries were prepared from the same DNA samples (Rep-set). Dedicated pipeline runs for each set yielded 15 991 variants for the Orig-set, 16 025 variants for the Tech-set, and 12 253 variants for the Rep-set. Examination of intersecting variant locations highlighted a pronounced similarity between the Orig-set and Tech-set, sharing 13 561 (84.8%) loci. In contrast, the degree of sharing between the Orig-set and the Rep-set dropped to 6 110 (38.2%) and a similar value of 6 216 (38.8%) was observed for the Tech-set. Altogether, 5 725 variants were common to all three sets (Fig. 4b). For the Orig-set as well as for the Rep-set the data aligned to the correct size selection range. However, the Rep-set had a slightly worse size range specificity but also less reads mapping to a few highly overrepresented sizes (Fig. 4c).

Repeatability of individual variants at the whitefish sample level was also evaluated by intersecting variants using VCFtools [57]. For the 5 725 overall shared variants, 44.4–93.0% variants were equally called among repeated individuals. In pairwise comparisons, Orig-Tech samples shared 93.0% equally called variants, for the Orig-Rep comparison, however, on the average only 44.8% and in the Tech-Rep 44.4% of the variants were called equally. For the 15 246 shared variants across the three independently repeated cow GBS runs we obtained, however, for all three pairwise comparisons an average repeatability of over 90%.

Further, lift-over chains between the created mock references and the pre-existing reference genomes have been created to match variants called via the mock reference and those called by utilizing the pre-existing reference genome. For cattle, 16 571 variants were called using the mock reference. In total, 13 471 of these variants

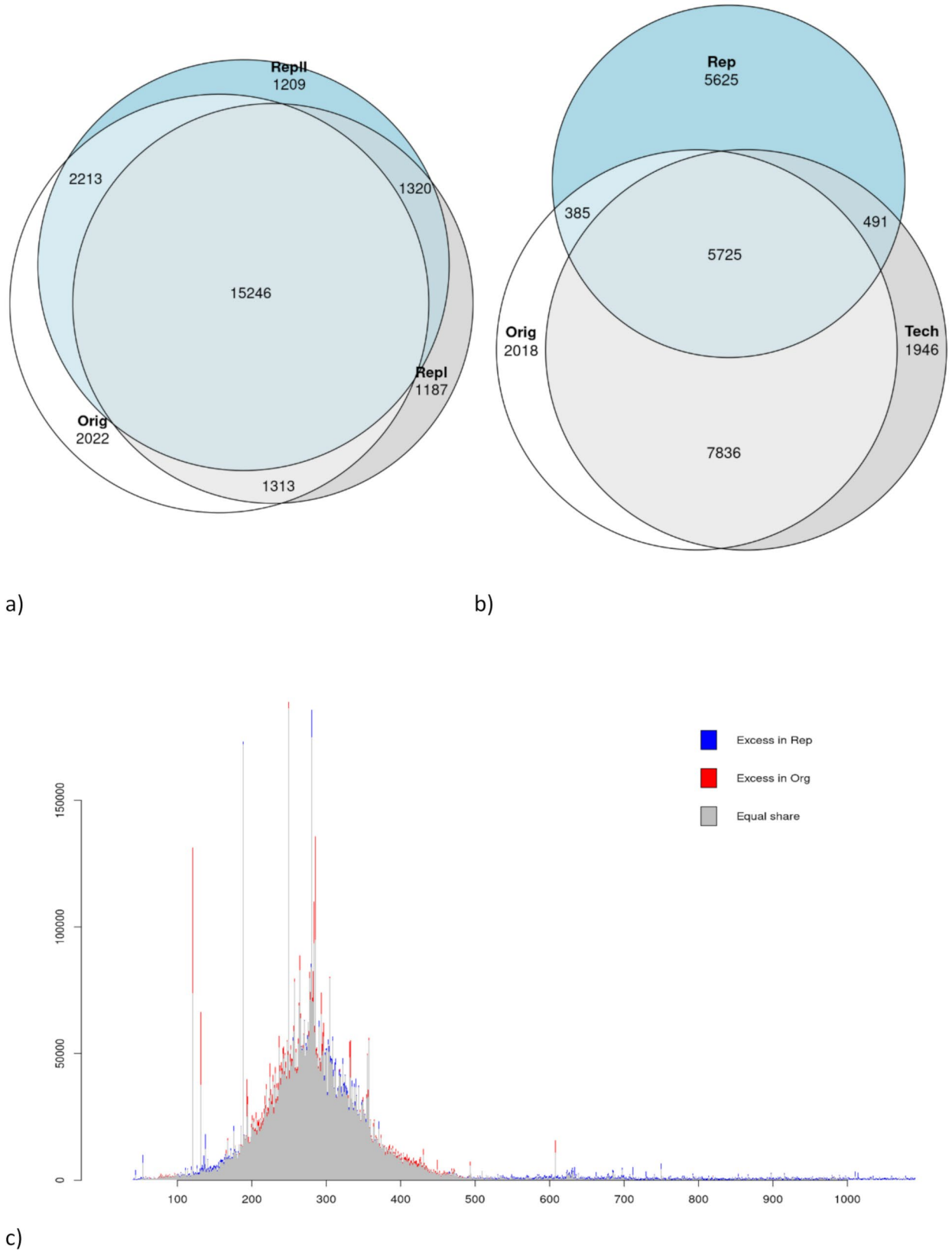

Fig. 4 Repeatability intersection Venn diagram. Left side (**a**) Cattle, right side (**b**) Whitefish, (**c**) read frequency distribution of the two Whitefish repeats

received successfully via lift-over a chromosomal location on the pre-existing reference genome. From these, 11 649 (>70%) intersected with the chromosomal location of variants called by utilizing the reference genome. In case of whitefish, from the 13 376 called variants via mock reference 10 693 could be lift-overed to the reference genome, with 6 481 (48.5%) variants having a chromosomal match with variants called based on the pre-existing reference genome.

Discussion

We present here a GBS approach containing a refined ddRAD approach, where through the adaption of a published laboratory protocol [58] and the optimization and streamlining of the GBS sequencing data analysis steps utilizing the Snakemake workflow manager, we introduce a cost effective and robust genotyping procedure. RAD-Seq, since its inception by [18], has rapidly gained standing across diverse genetic research domains, spanning for example genetic map creation [14, 59], mapping of production traits [51, 60, 61], population dynamics [62], and generating SNP resources for SNP array development [51, 63]. Particularly, GBS stands out as a valuable tool for generating markers in non-model species with limited genome information. Our work extends the prior experimental demonstration of the ddRAD GBS method to facilitate genomic selection and breeding planning, especially for less studied farmed species. We successfully applied the developed protocol in non-model species (European whitefish), demonstrating its versatility and effectiveness, albeit revealing some remaining challenges.

The prevailing trend strongly favors incorporating bioinformatic workflow engines for robust pipeline implementations [64]. *Snakemake* [65], a widely adopted choice within the NGS field, was employed in our study to manage task dependencies, to reduce redundant computations upon pipeline re-execution, and to facilitate automated deployment, including integration with the *slurm* workload manager on our cluster. The native docker and singularity support enabled seamless utilization and versioning of necessary software tools. With a single command, the pipeline execution is initiated, channeling outputs into a well-organized main folder with structured subfolders housing the resultant analyses. This comprehensive strategy ensures full reproducibility and user-friendliness, accommodating those with limited programming skills, as all essential configurations are consolidated within a central configuration file. We chose *GBS-SNP-CROP* [28] as base solution as it utilizes the generated sequencing data in a straightforward way producing a large number of reliable variant genotypes [28]. We wrapped the well-established *GBS-SNP-CROP* pipeline into a *Snakemake* workflow and extended it with various steps to create an automatically generated report that allows the user to evaluate the GBS run and to trace possible problems with it.

For the cattle samples we intersected and compared the results obtained from our GBS pipeline to results from our WGS pipeline, which is based on the GATK4 best practices. We considered here the GATK4 variants as gold standard to which we wanted to compare the results from the GBS pipeline. Further, following the approach of two independent pipelines allowed us to test for bugs in the developed pipeline, as we could assume a certain degree of consistency between the results.

Data generation

We utilized the modified ddRAD method [58] for sequence data generation. By avoiding costly barcoded adapters and instead ligating digested fragments to non-barcoded adapters and utilizing standard Illumina dual-indexed barcodes for PCR enrichment and sample multiplexing, we reduced the library preparation costs to <9€/sample. While the laboratory workflow involves multiple steps that lack convenient commercial kits, optimization efforts streamlined the process. Hands-on-time was halved to 10 h for 96 samples and 30 h for 384 samples by normalizing DNA concentrations using Myra liquid handling system (Bio Molecular Systems, Australia), incorporating SPRIselect beads for size enrichment allowing to omit one of the two time consuming concentration measurements with Qubit. The utilization of BluePippin (Sage Science, USA) and other possible automations may further solidify routines and improve quality and time- and cost-efficiency.

By generating shorter 2×75 bp PE sequencing reads on the NextSeq550 we reduced sequencing cost to 10–14€/sample, with a yield of 1 million reads per sample. Utilizing shorter reads is advantageous over longer reads, as the aim is to use unlinked variants and to avoid the complications caused by closely linked markers in relatedness estimation [66]. Decreasing read length in favor of increasing the read depth helps in avoiding too low read depth, which may lead to under-calling the heterozygotes and incorrect assignment of them as a homozygotes [67]. Our results suggest that a sequencing depth exceeding one million reads per sample leads to a stable variant calling with minimal variant missingness in assessed species. However, the required sequencing depth highly depends on the number of targeted fragments, which is a balance between DNA quality, used enzymes, used fragment size range and the genome size of the investigated species and even the chosen sequencing technology. Moreover, the number of recovered variable sites depends on the genome variability. As a result, preliminary evaluation with a limited subset of samples is recommended to establish the balance between the targeted fragments and the minimum coverage threshold.

In European whitefish, around 40% of GBS variants were scored repeatedly across two fully independent analyses, aligning with earlier observations [58]. Conversely, in the bovine analysis, the first two repeats shared over 80% of the called variants, and all three repeats shared still approximately 75% of variants despite purposefully varying DNA amount. This indicated on the one hand a high level of repeatability achievable in certain species, and on the other hand, a remaining challenge in repeatability in other species. Here, e.g., a duplication [68] in the genome could cause read alignment issues that cannot be circumvented, and which could possibly cause differences in variant calling. In that case, filtering out paralogs as suggested by [59] could be a promising approach to follow.

General stochastic variability inherent in wet lab methods, encompassing fluctuations in PCR, library generation, and fragment size selection, plays a role in the repeatability [69]. These aspects may further interact with the applied bioinformatic methodologies. For example, DNA fragments carrying the reference allele are more likely to be successfully mapped or receive higher quality scores [70]. The repeatability is also influenced by the filtering steps during the variant calling phase, when various filters (MAF, minimum/maximum coverage as well as minimum call rate) are applied, as we confirmed comparing the pipeline reports for filtered and unfiltered variants (result not shown). Further, multi-mapping of reads might lead to unpredictable consequences. Notably even for European whitefish, repeated GBS variant scoring between technical replicates was frequent (85%), underscoring the potential enhancement of repeatability through simultaneous library preparation for all analyzed individuals, although the results suggested the non-repeating variants might partially represent repetitive genome segments. In cattle, where genomic selection relies on relatedness across generations, repeatability across fully independent analyses is of significance. Contrastingly, aquaculture-based genomic selection involves comparing reference populations and selection candidates within a generation [71], diminishing the need for repeatability across generations. Additionally, relatedness estimation remains reasonably robust against missing data and genotyping errors when the variant count is substantial [23].

The GBS approach was tailored here for genomic selection utilizing a genomic relationship matrix, with the optimal informative GBS variant number falling between 1 000 and 10 000 [15] with a minimum of 1 000–2 000 SNPs generally suggested [15]. An in-silico comparison underscored the substantial influence of enzyme pair selection on reducing assessed genome complexity. However, even the enzyme pair with the lowest projected fragment count (EcoRI; SphI) was anticipated to yield ample variants. The difficulties of predicting fragment sequencing coverage are well-known and unassessed fragments are to be expected [69, 72]. Accordingly, our final GBS variant numbers in cattle and whitefish (20k and 16k) reduced from their projections (36k and 21k forecasted). Unassessed fragments could arise from multiple factors, including genomic structural variations between references and samples, variation at restriction cut sites [73], and repeated regions, biased nucleotide content, and sequence length variation [69]. A sufficient variant number margin is preferrable, as breeders running a genomic selection program might prefer excluding low MAF variants increasing the variance of diagonal GRM elements [74] or variants with suspiciously high observed heterozygosity (>50% [75]),. Notwithstanding the challenges, the simple projections demonstrated to be sufficient for estimating variant number magnitudes for the ddRAD GBS method.

Mock genome and pre-existing reference genome

For cattle a high-quality reference genome exists, while in our case representativeness of the European whitefish reference genome was uncertain. Utilizing a mock genome is essential when a reference genome is absent or incomplete for the target species [74, 76]. Further, the spread between alignment rates for the existing reference genome and the created mock reference can serve as a metric for the evaluation of the representativeness of the reference genome for the data at hand. Acting as a stand-in scaffold or reference, the mock genome is essential for variant calling and the subsequent analyses by providing a foundation for aligning and mapping the sequencing reads as well as localizing the called variants. An effective strategy for determining cluster numbers include using either a small representative sample group or a single exemplary sample. The latter approach, however, may introduce biases from unique features of that single sample [74]. Constructing a mock genome from a broader sample range, although suggested [74], results in an inflated reference. Depending on the total number of samples and based on our observations, opting for a moderate collection of 3–5 samples minimize specific biases and avoids excessive inflation. The recommendation of Sabadin et al. [74], however, seems to be more relevant for heterogeneous sample sets, as they are common e.g., in plant breeding. In these cases, the introduced final mock correction step is expected to curb excessive cluster inflation. The refined final mock provides more stable results and is generally preferrable.

While a mock genome reference might be necessary, it is not curated against computational artifacts related to sequencing errors [77], sequencing or base composition bias [78–80], or repetitive regions [77] which can constitute 10–60% of the genome [81, 82]. The suggested mock

construction parameters are a good starting point for most animal species, but correctly separating duplicated genome regions while simultaneously collapsing and merging haplotypic differences into a haploid sequence is a challenge to all assemblers [83]. The challenges can be seen here with the used Whitefish and its complex genome structure. Here, we recommend several iterations of the pipeline with different settings especially for the identity criterion for merging clusters for each new GBS data generation case. The identity criterion can be increased until the alignment rate begins to decrease significantly while maintaining or increasing per-site coverage. Other parameters fine-tune the pipeline mainly by removing noise from the input data and have smaller impact. Given the influence of data and parameters on the created mock reference, archiving and sharing the reference facilitates later comparability and repeatability. Further, many pipeline parameters that had little impact in the present comparison, could get more influential for problematic data and as such could rescue still semi-optimal sequencing runs.

Using a subspecies-specific reference for cattle and a species group-specific reference for whitefish led to a 25% GBS variant increase over mock genomes, as expected when closely related reference genomes are available [75, 84, 85]. This underlines the advantage of employing reference genomes whenever feasible. While the surplus of variants might raise concerns about the genotype call quality, evaluating genotyping via Mendelian inheritance [86] contradicted this notion, showing stable and comparable inheritance error rates to reported NGS-generated SNP data [85–87]. Comparing GRMs between GBS and WGS sequencing favored the reference genome based GBS analysis, which approximated the WGS GRM matrix more closely. Despite the common concern of low MAF in GBS data [74], our comparison had lower MAF in the reference WGS data than in the GBS datasets. While the WGS data offers comprehensive insights, reference genomes are not flawless, for example, excluding variants on genome regions specific to individuals or populations [88, 89] which may explain the minor difference between the two GBS GRM matrices. In general, using a very closely related reference genome increases the mapping and genotyping accuracy [84, 90]. Therefore, it is recommended to execute both mock and possibly pre-existing reference genome paths of the pipeline and then compare the outcomes. Current observations suggest a reference genome is advantageous and should be used when available, though it is not an absolute requirement. Using a pre-existing reference genome offers a high quality assembly and consistency and possibly annotated genomic context for interpretation [91]. Further, the use of a reference genome facilitates evaluating the representativeness of the data and allows linkage-based analyses.

Variant calling using different mock genomes or a pre-existing reference genome might include different variants [28], but the approaches gave currently very similar relatedness estimates. This aligns with previous studies suggesting that while extensive repeatability of GBS genotype data can be challenging biological inferences based on these data sets are more robust [21, 92, 93]. When genomic selection analyses are based on relatedness, fixing the reference genome is not the only option for merging data sets, since it is possible to combine partially overlapping relatedness matrices [94]. However, this necessitates having representative population samples with reference individuals of varying relatedness for both having reliable estimates within each round and for enabling merging of the matrices. Comparability issues might occur even when basing analysis on reference genomes, which develop over time [95].

The results from the current study will guide the development of whitefish SNP chips. While ddRAD provide the initial data for SNP discovery and ddRAD may suffice in some cases, SNP chips will improve reproducibility across generations and populations, making the data more broadly applicable in breeding programs.

Conclusions

The relatedness estimates based on the developed ddRAD GBS protocol aligns with independent relatedness estimates in both cattle and European whitefish samples, showcasing its versatility and extending the performance demonstration beyond GBS-SNP-CROPs original aim of identifying biological replicates. Our results conclude that while a pre-existing reference genome enhances variant calling quality and quantity, its absence does not impede the GBS-based genomic evaluation or selection. The applicability of the presented approach for genomic evaluation has been demonstrated for European whitefish [96], despite its challenging genomic structure. Further optimization, including fragment size window refinements and incorporation of methylation-sensitive restriction enzymes [17] could bring even greater efficiency and accuracy. The robust and user-friendly bioinformatic pipeline with an implementation of best practice approaches and wet-lab workflow achieves our broader goal of democratizing genotyping methods for researchers with varying levels of bioinformatics expertise and across a wide range of species and especially in less-studied production species. Experimenting with individual tuning parameters for the data at hand remains, however, indispensable and normally several pipeline runs are required until satisfying results are obtained. Furthermore, adjusting the filtering thresholds of called variants according to the analysis scope is still a required step, though default values should work well in many situations.

Methods

Samples

Altogether 12 Nordic Red dairy cows from the Luke research barn were selected for GBS and WGS sequencing, pipeline optimization and benchmarking. For each cow sample three repeated GBS libraries and one WGS library were created, starting from the same extracted DNA so that in total 36 GBS libraries and 12 WGS of cow samples were sequenced (Figure S1).

In addition, 42 European whitefish were used for pipeline validation and repeatability testing. Fish samples consisted of 27 randomly picked, unrelated individuals and 5 families of trios (parents and one offspring). From the set of random individuals, 12 whitefish were sequenced three times, twice with technical replicates of the same library and once with an entirely new library, that was started from the DNA. The European whitefish originate from the national breeding program owned and maintained by the Statutory Services of the Natural Resources Institute Finland (Luke), a governmental research institute. The fish are held at an inland, freshwater fish farm located in Enonkoski [46, 47]. The broodstock was established in 1998 from an anadromous wild strain of the river Kokemäki. Currently, the breeding program is based on traditional sire-dam-offspring pedigree, maintained by the use of family tanks during the early phase of growth [46, 47], but the development of SNPs will enable to implement also genomic selection.

Cow DNA was extracted from blood while fin tissue preserved in 100% ethanol was used for DNA extraction from fish. DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, Germany) following manufacturer's protocol. For fin tissue sampling, the fish were anesthetized by placing them for 5 min in oxygenated water which had 140 mg tricaine methane sulfonate / 1 L of water, and 140 mg bicarbonate / 1 L water.

Enzyme selection in silico

Restriction enzyme pairs for genome reduction were selected (i) to generate a number of fragments providing above 5 000 GBS variants and (ii) to leave a suitable overhang for library preparation. Assuming the proportion of variable sites of approximately 0.005 [50] and aiming for Paired-End (PE) sequencing with a total of 150 (2 × 75 bp) sequence read length per fragment, the number of variable sites was expected to be 0.75 times the fragment number. That suggested inclusion of at least 10 000 fragments, if all variable sites pass all quality ascertainment steps. The considered restriction enzyme pairs were EcoRI with MspI, SphI, MseI and NlaIII, or SphI with MluCI. These enzymes were previously used successfully for GBS in other species [22, 97, 98]. For a wider applicability, six reference genomes were included for the restriction enzyme evaluation: *Bos taurus* (ARS-UCD1.2), *Coregonus supersum 'balchen'* (AWG_v2), *Gallus gallus* (GRCg6a), *Hermetia illucens* (iHerIll2.2), *Oncorhynchus mykiss* (Omyk_1.0), and *Salmo salar* (ICSASG_v2). DdRAD library construction was simulated using SimRAD version 0.96 [27], but the functions were adjusted to use the full cut site. Digestion was simulated by using both the full reference genome contigs as well as reduced genomes of 10 random 10% genome subsamples. The full genome based (*Bos taurus* and *Coregonus supersum*) predicted fragments for the chosen EcoRI; SphI enzyme pair were used for quality evaluation of the GBS analysis. The obtained sequence data was used to estimate the effective size window and as consequence the size selection window was set to 150–400, for consistency. The effective size window thresholds were roughly estimated as values, where the slope of the density curves of the aligned fragments turned to + 1 (lower size threshold) and – 1 (upper size threshold).

ddRAD library preparation

The workflow (Figure S6) for the ddRAD library preparation was adapted from [58]. In detail, 250 or 500 ng of DNA was double-digested with two restriction enzymes EcoRI-HF (G^AATTC) and SphI-HF (GCATG^C) (New England Biolab, USA). The restriction reaction was performed in a volume of 20 µL, containing 17 µL of DNA (250 ng/500 ng in total), 0.25 µL of EcoRI-HF (5 units), 0.25 µL of SphI-HF (5 units), 2 µL of cut-smart buffer (10x) and 0.5 µL of molecular grade water at 37 °C for 2 h, following heat-inactivation for 15 min at 65 °C. Two non-barcoded restriction site specific adapters (Table S3) were ligated by adding 1 µL of each adapter (adapter P1 EcoRI: 1 µM, adapter P2, SphI: 10 µM) to the restriction mixture, 0.5 µL of T4 ligase (200 units) and 1.5 µL of ligation buffer (New England Biolab, USA). Ligation was performed at 16 °C for 14 h, following heat-inactivation at 65 °C for 15 min. DNA-fragments were selected between 200 bp and 700 bp by using SPRIselect magnetic beads (Beckman Coulter, USA) with a left-right ratio of 1x-0.56x. In details, the volume of each sample was adjusted with molecular grade water to 50 µL and then 28 µL of SPRIselect beads were added to achieve a 0.56x ratio for the selection of fragments shorter than 700 bp following selection of fragments longer than 200 bp by adding 22 µL of SPRIselect beads to achieve a ratio of 1x. The size selected DNA was resuspended in 25 µL of molecular grade water. Samples were barcoded by adding Illumina Nextera v2 (Illumina, San Diego, CA, USA) combinatorial dual-indexed barcodes (i7 and i5). For each individual sample a PCR-mix containing 6 µL of 5x Phusion HF buffer, 0.4 µL dNTP (10 mM), 0.2 µL of Phusion HF DNA polymerase (0.4 units) (ThermoFisher scientific, USA), 1.5 µL of i5 barcode primer, 1.5 µL of i7 barcode primer, 5 µL of sample and 15.4 µL of molecular grade

water was prepared, two PCR reactions per sample were performed. The cycling conditions were as follows: initial denaturation at 98 °C for 30 s, followed by 18 cycles of 10 s at 98 °C, 20 s at 61 °C, 15 s at 72 °C and a final elongation step at 72 °C for 10 min. The two PCR reactions per sample were pooled, the volume was adjusted to 50 µL, and small fragment removal was carried out with 40 µL (0.8x) SPRIselect beads. The size selected PCR products were resuspended in 25 µL molecular grade water and quantified using Qubit Flex with 1x dsDNA HS assay (ThermoFisher scientific, USA). Only products with a significantly higher amount than the No Template Control (NTC) were used for sequencing (> 3 ng/µL).

Sequencing

Single ddRAD libraries were pooled in equimolar amounts. The pool was size selected with SPRIselect beads to the length between 300 and 700 bp (ratio 0.75-0.56x), corresponding to the combined length of 150–550 bp restriction insert and 147 bp adapter. The quality and size of the pooled sequencing library was evaluated on the TapeStation 4150 (Agilent, USA) using the DNA HS1000 assay. Quantification of the library was done using Qubit 4 (1x dsDNA HS assay) (ThermoFisher scientific, USA). Following the guidelines from the NextSeq System denature and dilute libraries guide (Document # 15048776 v09, December 2018 (Illumina, San Diego, CA, USA)), the library was diluted for sequencing to a final concentration of 1.4 pM, containing 10% PhiX control, to increase complexity at the start of the sequencing. The PE sequencing (2 × 75 bp) was performed on the NextSeq 550 (Illumina, San Diego, CA, USA) using medium output flow cell.

The WGS of cow samples was performed at the Finnish Functional Genomics Centre (Turku, Finland) using TruSeq® DNA PCR-Free Library kit (Illumina, San Diego, CA, USA) and PE sequencing (2 × 150 bp) on an Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA) platform.

Mock-reference genome

Analyzing GBS data without a preexisting reference genome necessitates in creating a technical (mock) reference. For this, various sample selection methods were considered: choosing the sample with the highest read count (mock-strategy 1), a sample with an average read count (mock-strategy 2), a random subset of three samples (mock-strategy 3), or all samples (mock-strategy 4).

As the first step, the raw PE sequences were checked for overlap that might happen in case of short inserts. Overlapping reads were merged into single-end (SE) reads using *PEAR* [54], with two tuning parameters being optimized here: the *p* option (values between 0.001 and 0.1) for a statistical test to determine read-pair merging, and the *pl* option (values 30 to 70) for defining the minimum accepted total length of the merged construct. These parameters determined when read pairs were merged and whether the construct's length met the criteria for inclusion. PE reads that could not be merged, were then stitched together with a sequence of 20 N bases as standard for the pipeline. Stitching of reads was controlled by the parameter *rl*, and reads were stitched, if the length of read1 was larger than (rl – 19) and length of read2 was larger than (rl – 5), otherwise reads were not used for the mock generation. The resulting SE reads were utilized to construct the de-novo mock reference genome using *vsearch* [53]. In the de-novo building phase, two *vsearch* options were fine-tuned: the *id* option (values between 0.8 and 0.99), defining the minimum pairwise identity for merging two clusters, and the *min* option (values between 80 and 160), setting the minimum cluster length for inclusion in the mock reference. The in-silico simulated protocol as described in "Enzyme selection in silico" was used to evaluate the mock reference constructs.

Following the de-novo mock reference creation, an additional refinement step was applied, where clusters with low coverage were removed from the mock reference. Tuning parameters were *totalReadCoverage* and *minSampleCoverage*. The first parameter defines the minimum number of reads that need to be aligned across all samples on a cluster to keep it in the mock reference. The second parameter defines the minimum number of samples that need to have at least a single read aligned to a cluster so that this cluster remains in the mock. For the tuning of the *totalReadCoverage* we tested 6, 12, 24, 60 and 120 as values and for *minSampleCoverage* reads from 2 (10%), 4 (25%), 6 (50%), 8 (75%), 10 (90%), 12 (100%) of the total number of samples in the study.

Variant calling

The GBS variant calling was done with our developed *Snakebite-GBS* [29] pipeline, which is a *Snakemake* pipeline extension, that is based on the existing *GBS-SNP-CROP* [28] pipeline. The Snakebite-GBS pipeline is part of the Snakebite framework *Snakepit* [99]. First, the quality-trimmed reads were aligned with *BWA-mem* [52] against the mock and/or preexisting reference genome(s). Then, *samtools mpileup* [100] was used for variant calling and various filters were applied to obtain the final variant set. The underlying GBS-SNP-CROP pipeline allows for eight different filters: (1) *mnHoDepth0* (value: 5), the minimum depth required for calling a homozygote when the alternative allele depth equals 0; (2) *mnHoDepth1* (value: 20) the minimum depth required for calling a homozygote when the alternative allele depth equals 1; (3) *mnHetDepth* (value: 3) the minimum depth required for each allele when calling a heterozygote; (4) *altStrength* (value: 0.8) the minimum proportion of non-primary allele reads that are the secondary allele; (5) *mnAlleleRatio* (value:

0.25) the minimum required ratio of the less frequent allele depth to the more frequent allele depth; (6) *mnCall* (value: 0.75) the minimum acceptable proportion of genotyped individuals to retain a variant; (7) *mnAvgDepth* (value: 3) the minimum average read depth of an acceptable variant; (8) *mxAvgDepth* (value: 200) the maximum average read depth of an acceptable variant. A pipeline flowchart with involved steps and tools can be found in Figure S7.

The cattle WGS variant calling was performed following the *GATK4* best practices [101] implemented as *Snakemake* [65] workflow called *Snakebite-WGS* [102]. Implemented steps contain, among others, the GATK base recalibrator as well as a model to adjust the base quality scores and a base recalibration step. Variant calling is done via haplotype caller. The pipeline utilizes also BWA-mem to align the data but includes a refinement step using *Picard* before the *GATK4* software suite is used for the final variant calling with applied default filters.

GBS quality evaluation

The generated cow GBS variant data was mapped against an in-silico digested ARS-UCD1.2 reference genome for evaluating the size selection performance. Following variant calling, sample-wise genotype concordance between GBS and WGS sequencing strategies was assessed using *Picard*.

The repeatability of the GBS runs was tested by intersecting the variant locations on the corresponding reference genomes. Here, *bcftools* [103] was used to intersect the three vcf-files and corresponding intersection numbers were calculated. Further, *samtools mpileup* was run for the GBS data aligned to the reference genome and for each sample contiguous areas, that had a minimum coverage of three reads, were identified and stored in bed-format. Individual sample-wise bed-files were then merged and only regions with read support from at least 10 samples were kept. This bed-file was then used to intersect the WGS-based vcf file using *bedtools* [104] and extract WGS variants only from the corresponding intersecting genome regions.

In cattle, the GBS variant based variability and relatedness were compared against resampled WGS variants with restricted variant numbers from 50 to 30 000 to compare how the variant number influenced the classical Genomic Relatedness Matrix (GRM) calculated using the R-package *BGData* [105]. The GRM based on the full WGS variant matrix was compared to smaller bootstrap samples of WGS and GBS data.

The lift-over between mock reference and pre-existing reference genome to compare variants from both methods based on their chromosomal was done by using the tool *transanno*. Here, first the mock reference was aligned against the reference genome and the resulting file in pairwise mapping format (paf) was then used in *transanno* to create the lift-over chain and eventually to perform the lift-over. Chromosomal locations between the lift-overed mock reference-based variants and their pre-existing reference genome based counterparts were then again matched via *bcftools isec*.

The GRM structure differences were quantified by measuring the variability in different directions using the distance between the eigenvalues of the matrices, calculated using the Frobenius matrix norm.

For whitefish data, relatedness was also calculated using the R-package *BGData* [105] was assessed using the full whitefish data set to overcome bias in the small data set caused by few closely related individuals in the parental generation though we focused on trio results. In addition to the genomic relatedness, the genotype quality was assessed by evaluating non-Mendelian inheritance of the GBS variants in five families of trios, that included parents and an offspring.

Abbreviations

Bp	Basepair
ddRAD	Double-digest RAD-sequencing
GBS	Genotyping-by-sequencing
GRM	Genomic Relatedness Matrix
MAF	Minor Allele Frequencies
NGS	Next generation sequencing
PE	Paired-end
RAD	Restriction-site associated DNA sequencing
SE	Single-end
SNP	Single Nucleotide Polymorphism
WGS	Whole-Genome-Sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11296-4.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We thank Finnish Functional Genomics Centre, supported by University of Turku, Åbo Akademi University and Biocenter Finland, for whole genome sequencing services and CSC – IT Center for Science, Finland, for computational resources. Antti Nousiainen and Heikki Koskinen are acknowledged for providing whitefish samples.

Author contributions

AK, IT, MT, TIT: conception of the study; IT, AK: funding acquisition; OB: adapting and optimizing ddRAD; DF, MT: data analysis and writing the manuscript; IT, TIT, OB, AK: manuscript revision. All authors approved the final manuscript.

Funding

This work was supported by the Natural Resources Institute Finland (Luke) strategic projects (41007–00155500 and 41007–00215600), and 'ArctAqua - Cross-Border Innovations in Arctic Aquaculture' project, co-funded by Kolarctic CBC Programme 2014–2020, with a grant contract number 4/2018/095/KO4058, and the Statutory Services of Natural Resources Institute Finland. Sequencing of cattle samples was funded by Academy of Finland grant No. 317998.

Data availability
The datasets generated and analyzed during the current study are available in the European Nucleotide Archive (ENA), accession number PRJEB66491.

Declarations

Ethics approval and consent to participate
The study was performed in accordance with Finnish animal welfare legislation and complied with the directive 2010/63/EU implemented in Finnish legislation in the Act on the Use of Animals for Experimental Purposes (62/2006) and followed the protocols approved by the Luke's Animal Care Committee, Helsinki, Finland. All experimental fish were anaesthetized with tricaine methanesulfonate before sampling to minimize suffering. Cattle: The cattle study was approved by the ethics committee, Southern Finland Regional State Administrative Agency/The Project Authorisation Board ELLA (ethical permission number ESAVI/16348/2019).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 January 2024 / Accepted: 27 January 2025
Published online: 05 February 2025

References

1. Duarte CM, Marbá N, Holmer M. Rapid domestication of marine species. Science. 2007;316(5823):382–3.
2. The State of World Fisheries and Aquaculture 2020. FAO. 2020 [cited 2023 Jun 20]. Available from: http://www.fao.org/documents/card/en/c/ca9229en
3. Palaiokostas C, Kocour M, Prchal M, Houston RD. Accuracy of genomic evaluations of juvenile growth rate in Common Carp (Cyprinus carpio) using genotyping by sequencing. Front Genet. 2018;9:82.
4. Tsai HY, Hamilton A, Tinch AE, Guy DR, Gharbi K, Stear MJ, et al. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics. 2015;16(1):969.
5. Yoshida GM, Lhorente JP, Correa K, Soto J, Salas D, Yáñez JM. Genome-wide association study and cost-efficient genomic predictions for growth and fillet yield in Nile Tilapia (Oreochromis niloticus). G3 GenesGenomesGenetics. 2019;9(8):2597–607.
6. Garner JB, Douglas ML, Williams SRO, Wales WJ, Marett LC, Nguyen TTT, et al. Genomic selection improves heat tolerance in dairy cattle. Sci Rep. 2016;6(1):34114.
7. Robledo D, Matika O, Hamilton A, Houston RD. Genome-wide association and genomic selection for resistance to amoebic gill disease in Atlantic Salmon. G3 GenesGenomesGenetics. 2018;8(4):1195–203.
8. Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, et al. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat Rev Genet. 2020;21(7):389–409.
9. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
10. Hotaling S, Kelley JL, Frandsen PB. Toward a genome sequence for every animal: where are we now? Proc Natl Acad Sci. 2021;118(52):e2109019118.
11. FAO Yearbook. Fishery and aquaculture statistics 2019/FAO annuaire. Statistiques des pêches et de l'aquaculture 2019/FAO anuario. Estadísticas de pesca y acuicultura 2019. FAO. 2021 [cited 2023 Jun 27]. Available from: http://www.fao.org/documents/card/en/c/cb7874t
12. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177(4):2389–97.
13. Vela-Avitúa S, Meuwissen T, Luan T, Ødegård J. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genet Sel Evol. 2015;47(1):9.
14. Gonen S, Lowe NR, Cezard T, Gharbi K, Bishop SC, Houston RD. Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. BMC Genomics. 2014;15(1):166.
15. Kriaridou C, Tsairidou S, Houston RD, Robledo D. Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. Front Genet. 2020;11:124.
16. Berry DP, Spangler ML. Animal board invited review: practical applications of genomic information in livestock. Animal. 2023;17(11):100996.
17. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Orban L editor. PLoS ONE. 2011;6(5):e19379.
18. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. Fay JC, editor. PLoS ONE. 2008;3(10):e3376.
19. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 2007;17(2):240–8.
20. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008;5(3):247–52.
21. Cumer T, Pouchon C, Boyer F, Yannic G, Rioux D, Bonin A, et al. Double-digest RAD-sequencing: do pre- and post-sequencing protocol parameters impact biological results? Mol Genet Genomics. 2021;296(2):457–71.
22. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. Orlando L, editor. PLoS ONE. 2012;7(5):e37135.
23. Attard CRM, Beheregaray LB, Möller LM. Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: avoiding the trap of precise bias. Mol Ecol Resour. 2018;18(3):381–90.
24. Wang Y, Cao X, Zhao Y, Fei J, Hu X, Li N. Optimized double-digest genotyping by sequencing (ddGBS) method with high-density SNP markers and high genotyping accuracy for chickens. Xu P, editor. PLOS ONE. 2017;12(6):e0179073.
25. Chafin TK, Martin BT, Mussmann SM, Douglas MR, Douglas ME. FRAGMATIC: in silico locus prediction and its utility in optimizing ddRADseq projects. Conserv Genet Resour. 2018;10(3):325–8.
26. Lajmi A, Glinka F, Privman E. Optimizing ddRAD sequencing for population genomic studies with ddgRADer. Mol Ecol Resour. 2023;1755-0998.13870.
27. Lepais O, Weir JT. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. Mol Ecol Resour. 2014;14(6):1314–21.
28. Melo ATO, Bartaula R, Hale I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinformatics. 2016;17(1):29.
29. Fischer D, fischuu/Snakebite. -GBS: Pipeline release version 0.18.3. Zenodo; 2023 [cited 2023 Oct 3]. Available from: https://zenodo.org/record/7550722
30. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. Tinker NA, editor. PLoS ONE. 2014;9(2):e90346.
31. Rochette NC, Rivera-Colón AG, Catchen JM. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol. 2019;28(21):4737–54.
32. Eaton DAR, Overcast I. ipyrad: Interactive assembly and analysis of RADseq datasets. Schwartz R, editor. Bioinformatics. 2020;36(8):2592–4.
33. Eaton DAR. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. Bioinformatics. 2014;30(13):1844–9.
34. Puritz JB, Hollenbeck CM, Gold JR. A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ. 2014;2:e431.
35. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9.
36. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276–8.
37. Langer BE, Amaral A, Baudement MO, Bonath F, Charles M, Chitneedi PK et al. Empowering bioinformatics communities with Nextflow and nf-core. 2024 [cited 2024 Dec 16]. Available from: https://doi.org/10.1101/2024.05.10.592912
38. Doublet M, Degalez F, Lagarrigue S, Lagoutte L, Gueret E, Allais S, et al. Variant calling and genotyping accuracy of ddRAD-seq: comparison with 20X WGS in layers. PLoS ONE. 2024;19(7):e0298565.

39. Aguirre NC, Villalba PV, García MN, Filippi CV, Rivas JG, Martínez MC, et al. Comparison of ddRADseq and EUChip60K SNP genotyping systems for population genetics and genomic selection in Eucalyptus dunnii (Maiden). Front Genet. 2024;15:1361418.
40. Herry F, Hérault F, Lecerf F, Lagoutte L, Doublet M, Picard-Druet D, et al. Restriction site-associated DNA sequencing technologies as an alternative to low-density SNP chips for genomic selection: a simulation study in layer chickens. BMC Genomics. 2023;24(1):271.
41. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. These aren't the loci you're looking for: principles of effective SNP filtering for molecular ecologists. Mol Ecol. 2018;27(16):3193–206.
42. Aslam ML, Carraro R, Bestin A, Cariou S, Sonesson AK, Bruant JS, et al. Genetics of resistance to photobacteriosis in gilthead sea bream (Sparus aurata) using 2b-RAD sequencing. BMC Genet. 2018;19(1):43.
43. Pappas F, Palaiokostas C. Genotyping strategies using ddRAD sequencing in Farmed Arctic Charr (Salvelinus alpinus). Animals. 2021;11(3):899.
44. Guppy JL, Jones DB, Kjeldsen SR, Le Port A, Khatkar MS, Wade NM, et al. Development and validation of a RAD-Seq target-capture based genotyping assay for routine application in advanced black tiger shrimp (Penaeus monodon) breeding programs. BMC Genomics. 2020;21(1):541.
45. Liu Q, Lin H, Chen J, Ma J, Liu R, Ding S. Genetic variation and population genetic structure of the large yellow croaker (Larimichthys crocea) based on genome-wide single nucleotide polymorphisms in farmed and wild populations. Fish Res. 2020;232:105718.
46. Kause A, Quinton C, Airaksinen S, Ruohonen K, Koskela J. Quality and production trait genetics of farmed European whitefish, Coregonus lavaretus1. J Anim Sci. 2011;89(4):959–71.
47. Janhunen M, Nousiainen A, Koskinen H, Vehviläinen H, Kause A. Selection strategies for controlling muscle lipid content recorded with a non-destructive method in European whitefish, Coregonus lavaretus. Aquaculture. 2017;481:229–38.
48. Crotti M, Bean CW, Gowans ARD, Winfield IJ, Butowska M, Wanzenböck J, et al. Complex and divergent histories gave rise to genome-wide divergence patterns amongst European whitefish (*Coregonus lavaretus*). J Evol Biol. 2021;34(12):1954–69.
49. De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. Genotyping-by-Sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. Nelson JC, editor. PLoS ONE. 2013;8(5):e62137.
50. De-Kayne R, Feulner PGD. A European Whitefish linkage map and its implications for understanding genome-wide synteny between salmonids following whole genome duplication. G3 GenesGenomesGenetics. 2018;8(12):3745–55.
51. Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, et al. Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). BMC Genomics. 2014;15(1):90.
52. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
53. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
54. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate illumina paired-end reAd mergeR. Bioinformatics. 2014;30(5):614–20.
55. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34(18):3094–100.
56. Moore KL, Vilela C, Kaseja K, Mrode R, Coffey M. Forensic use of the genomic relationship matrix to validate and discover livestock pedigrees. J Anim Sci. 2019;97(1):35–42.
57. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
58. Salas-Lizana R, Oono R. Double-digest RADseq loci using standard Illumina indexes improve deep and shallow phylogenetic resolution of *Lophodermium*, a widespread fungal endophyte of pine needles. Ecol Evol. 2018;8(13):6638–51.
59. Recknagel H, Elmer KR, Meyer A. A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (Amphilophus spp.) obtained by massively parallel DNA sequencing (ddRADSeq). G3 GenesGenomesGenetics. 2013;3(1):65–74.
60. Shao C, Niu Y, Rastas P, Liu Y, Xie Z, Li H, et al. Genome-wide SNP identification for the construction of a high-resolution genetic map of Japanese flounder (Paralichthys olivaceus): applications to QTL mapping of Vibrio anguillarum disease resistance and comparative genomic analysis. DNA Res. 2015;22(2):161–70.
61. Fu B, Liu H, Yu X, Tong J. A high-density genetic map and growth related QTL mapping in bighead carp (Hypophthalmichthys nobilis). Sci Rep. 2016;6(1):28679.
62. Bradic M, Teotónio H, Borowsky RL. The population genomics of repeated evolution in the blind Cavefish Astyanax mexicanus. Mol Biol Evol. 2013;30(11):2383–400.
63. Palti Y, Gao G, Miller MR, Vallejo RL, Wheeler PA, Quillet E, et al. A resource of single-nucleotide polymorphisms for rainbow trout generated by restriction-site associated DNA sequencing of doubled haploids. Mol Ecol Resour. 2014;14(3):588–96.
64. Larsonneur E, Mercier J, Wiart N, Floch EL, Delhomme O, Meyer V. Evaluating workflow management systems: a bioinformatics use case. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid, Spain: IEEE; 2018 [cited 2023 Aug 23]. pp. 2773–5. Available from: https://ieeexplore.ieee.org/document/8621141/
65. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data anal snakemake F1000Research. 2021;10:33.
66. Mathew B, Léon J, Sillanpää MJ. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity. 2018;120(4):356–68.
67. Furuta T, Yamamoto T, Ashikari M. GBScleanR: robust genotyping error correction using a hidden Markov model with error pattern recognition. Endelman J, editor. GENETICS. 2023;224(2):iyad055.
68. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533(7602):200–5.
69. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. Special features of RAD sequencing data: implications for genotyping. Mol Ecol. 2013;22(11):3151–64.
70. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. Di Rienzo A, editor. PLOS Genet. 2019;15(7):e1008302.
71. Fraslin C, Koskinen H, Nousianen A, Houston RD, Kause A. Genome-wide association and genomic prediction of resistance to Flavobacterium columnare in a farmed rainbow trout population. Aquaculture. 2022;557:738332.
72. Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, et al. Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics. 2013;193(4):1073–81.
73. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, et al. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol. 2013;22(11):3165–78.
74. Sabadin F, Carvalho HF, Galli G, Fritsche-Neto R. Population-tailored mock genome enables genomic studies in species without a reference genome. Mol Genet Genomics. 2022;297(1):33–46.
75. Torkamaneh D, Laroche J, Belzile F, Genome-Wide SNP. Calling from genotyping by sequencing (GBS) data: a comparison of seven pipelines and two sequencing technologies. Candela H, editor. PLOS ONE. 2016;11(8):e0161333.
76. Machado IP, DoVale JC, Sabadin F, Fritsche-Neto R. On the usefulness of mock genomes to define heterotic pools, testers, and hybrid predictions in orphan crops. Front Plant Sci. 2023;14:1164555.
77. Liao X, Li M, Zou Y, Wu FX, Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. Quant Biol. 2019;7(2):90–109.
78. DaCosta JM, Sorenson MD. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. Antoniewski C, editor. PLoS ONE. 2014;9(9):e106713.
79. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. Whole genome amplification and de novo assembly of single bacterial cells. Ahmed N, editor. PLoS ONE. 2009;4(9):e6864.
80. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–32.
81. Kazazian HH. Mobile elements: drivers of genome evolution. Science. 2004;303(5664):1626–32.
82. Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, et al. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics. 2018;19(1):141.
83. Kivikoski M, Rastas P, Löytynoja A, Merilä J. Automated improvement of stickleback reference genome assemblies with Lep-Anchor software. Mol Ecol Resour. 2021;21(6):2166–76.
84. Bohling J. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. Ecol Evol. 2020;10(14):7585–601.

85. Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW et al. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Gilbert M, editor. Methods Ecol Evol. 2017;8(8):907–17.
86. Pilipenko VV, He H, Kurowski BG, Alexander ES, Zhang X, Ding L, et al. Using Mendelian inheritance errors as quality control criteria in whole genome sequencing data set. BMC Proc. 2014;8(S1):S21.
87. Kumar P, Al-Shafai M, Al Muftah WA, Chalhoub N, Elsaid MF, Aleem AA, et al. Evaluation of SNP calling using single and multiple-sample calling algorithms by validation against array base genotyping and Mendelian inheritance. BMC Res Notes. 2014;7(1):747.
88. Crysnanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. Proc Natl Acad Sci. 2021;118(20):e2101056118.
89. Gong Y, Li Y, Liu X, Ma Y, Jiang L. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? J Anim Sci Biotechnol. 2023;14(1):73.
90. Thorburn DJ, Sagonas K, Binzer-Panchal M, Chain FJJ, Feulner PGD, Bornberg-Bauer E et al. Origin matters: using a local reference genome improves measures in population genomics. Mol Ecol Resour. 2023;1755-0998.13838.
91. Whibley A, Kelley JL, Narum SR. The changing face of genome assemblies: guidance on achieving high-quality reference genomes. Mol Ecol Resour. 2021;21(3):641–52.
92. Casanova A, Maroso F, Blanco A, Hermida M, Ríos N, García G, et al. Low impact of different SNP panels from two building-loci pipelines on RAD-Seq population genomic metrics: case study on five diverse aquatic species. BMC Genomics. 2021;22(1):150.
93. Wright B, Farquharson KA, McLennan EA, Belov K, Hogg CJ, Grueber CE. From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. BMC Genomics. 2019;20(1):453.
94. Akdemir D, Knox R, Isidro Y, Sánchez J. Combining partially overlapping multi-omics data in databases using relationship matrices. Front Plant Sci. 2020;11:947.
95. Stolarczyk M, Xue B, Sheffield NC. Identity and compatibility of reference genome resources. NAR Genomics Bioinforma. 2021;3(2):lqab036.
96. Calboli F, Iso-Touru T, Bitz O, Fischer D, Nousiainen A, Koskinen H et al. Genomic selection for survival under naturally occurring Saprolegnia oomycete infection in farmed European whitefish Coregonus lavaretus. J Anim Sci. 2023;101:skad333. https://doi.org/10.1093/jas/skad333.
97. Barría A, Christensen KA, Yoshida GM, Correa K, Jedlicki A, Lhorente JP et al. Genomic predictions and genome-wide association study of resistance against *Piscirickettsia salmonis* in Coho Salmon (*Oncorhynchus kisutch*) using ddRAD sequencing. G3 GenesGenomesGenetics. 2018;8(4):1183–94.
98. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. Yin T, editor. PLoS ONE. 2012;7(2):e32253.
99. Fischer D. Snakepit - The Snakebite hub. Available from: http://www.snakep.it
100. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
101. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Genomics; 2017 [cited 2023 Aug 18]. Available from: https://doi.org/10.1101/201178
102. Fischer D, fischuu/Pipeline. -WGS-VariantCalling: Stable pre-release version. Zenodo; 2023 [cited 2023 Oct 3]. Available from: https://zenodo.org/record/8401423
103. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008.
104. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
105. Grueneberg A, De Los Campos G. BGData - a suite of R packages for genomic analysis with big data. G3 GenesGenomesGenetics. 2019;9(5):1377–83.

Publisher's note