Pedigree-based genome-wide imputation using a low-density amplicon panel for the highly polymorphic Pacific oyster *Crassostrea* (*Magallana*) *gigas*

Ben J. G. Sutherland[1,2], Konstantin Divilov[3], Neil F. Thompson[4], Thomas A. Delomas[5], Spencer L. Lunda[6], Christopher J. Langdon[7], Timothy J. Green[1]

[1] Faculty of Science and Technology, Vancouver Island University, Nanaimo, BC, Canada V9R 5S5

[2] Sutherland Bioinformatics, Lantzville, BC, Canada V0R 2H0

[3] Center for Quantitative Life Sciences, Oregon State University, Corvallis, OR 97331, USA

[4] Pacific Shellfish Research Unit, USDA Agricultural Research Service, 2030 SE Marine Science Drive, Newport, Oregon 97365, USA

[5] Agricultural Research Service, United States Department of Agriculture, National Cold Water Marine Aquaculture Center, 483 CBLS, 120 Flagg Road, Kingston, RI 02881, USA

[6] Oregon State University, Department of Microbiology, 226 Nash Hall, Corvallis, OR 97331, USA

[7] Oregon State University, Coastal Oregon Marine Experiment Station, Hatfield Marine Science Center, 2030 SE Marine Science Dr., Newport, OR 97365, USA

* Author for correspondence: BJGS
Current Address:
Faculty of Science and Technology, Vancouver Island University, Nanaimo, BC, Canada V9R 5S5
Ben.Sutherland@viu.ca

Running title: Oyster amplicon panel imputation

Abstract

High-density genomic data are instrumental for selective breeding, but the high costs associated with these approaches can hinder progress, as is the case for most aquaculture species. A strategy to reduce genotyping costs is to genotype a few select animals at high-density (e.g., parents, grandparents), and large numbers of other animals at low density (e.g., offspring), then impute genotypes. This has been demonstrated *in silico* for Pacific oyster *Crassostrea* (*Magallana*) *gigas* but was particularly challenging relative to other species and has never been empirically tested. Here, four families of Pacific oysters, bred via marker-assisted selection for variation at a locus for ostreid herpesvirus 1 (OsHV-1) survivorship, were exposed to OsHV-1 then genotyped using a low-density amplicon panel (n = 240). Parents were genotyped with the amplicon panel and whole-genome resequencing data. Offspring genotypes were imputed, and accuracy was characterized against a hold-out dataset. Imputation reduced minor allele frequencies and enriched homozygosity. An *in silico* three-generation analysis was used to investigate the effect of deepening the pedigree, which generated superior results (genotype concordance = 84.5%, allelic dosage r = 0.73), although the two-generation imputation is likely sufficient for selection decisions (GC = 75.3%; r = 0.63). Genome-wide associations to OsHV-1 survivorship with imputed data identified major effect loci on the expected chromosome 8, but in a different region than expected, pointing to a potentially more complex genetic architecture for the trait. These methods demonstrate the empirical utility of amplicon panel-based genome-wide imputation in shellfish, and thus enable low-cost selective breeding techniques.

Keywords: Animalia; amplicon panel; *Crassostrea* (*Magallana*) *gigas*; genome-wide association study; imputation; ostreid herpesvirus 1 (OsHV-1); Pacific oyster; whole-genome resequencing

Introduction

The Pacific oyster *Crassostrea* (*Magallana*) *gigas* (Thunberg, 1793) is a highly valuable aquaculture species globally (Botta et al. 2020). In many countries it is the subject of genetic improvement programs that range from large-scale governmental and commercial programs to small-scale university programs. Access to genomic resources for marker-assisted or genomic selection enables economic gains and advances food production (Martínez-García et al. 2022) but is significantly limited by the availability of funding, which creates a need for cost-effective options (Boudry et al. 2021; Delomas et al. 2023). The recent development of a low-density Pacific oyster amplicon panel for SNP genotyping is anticipated to facilitate high-throughput low-density genotyping using a repeatable and low-cost approach (Meek and Larson 2019; Sutherland et al. 2024). However, with a density of 592 amplicons across approximately 650 Mbp and 10 chromosomes (Peñaloza et al. 2021), alone it is limited in scope for methods requiring genome-wide coverage, especially given the rapid linkage decay in oysters (Hollenbeck and Johnston 2018; Hu et al. 2022; Mao et al. 2024).

Genome-wide coverage can be attained cost-effectively through imputation by combining low-density genotypes with high-density genotypes of select individuals (Habier et al. 2009), such as parents and grandparents (Delomas et al. 2023) with Mendelian inheritance. This has been explored *in silico* in Pacific oysters. Kriaridou et al. (2023) subsampled SNPs from high density data for Pacific oysters and three vertebrate species to construct low density genotype samples, imputed genotypes, then determined imputation accuracy relative to empirical genotypes. This study demonstrated the challenge of obtaining high-quality imputation specific to Pacific oysters when using parents and offspring alone. Delomas et al. (2023) simulated genotypes from a breeding program and found that low-density panels (e.g., 250 or 500 SNP markers) combined with high-density genotypes for grandparents and parents allowed imputation to provide reliable genotypes for genomic selection (GS). However, empirical use of an amplicon panel and high-density data to conduct imputation has not yet been evaluated for any bivalve species, including the Pacific oyster.

Traits for which no information exists on underlying causal mechanisms, or that are controlled by many loci are inaccessible to marker-assisted selection (MAS) but can be targeted by GS. Cost-effectiveness is necessary for GS in Pacific oysters and other aquaculture species (Delomas et al. 2023), especially if there are differences in phenotypic and genetic responses to local environmental conditions (Allen et al. 2020; Allen et al. 2021). Furthermore, the ability to target multiple diverse traits is essential for Pacific oyster aquaculture, considering the challenges that the industry has experienced from biotic and abiotic stressors (Green et al. 2019; King et al. 2019). These stressors can result in summer mortality syndrome (Cowan et al. 2023; Cowan et al. 2024) and other massive mortality events, in some cases linked to changes in climate or extreme climate events (Raymond et al. 2022). Adverse environmental conditions may change

over time, creating a moving target for breeding programs that requires affordability and adaptability. Resilience through selective breeding has been highlighted as an important adaptation strategy for both Pacific and eastern oysters, *C. virginica* (Langdon et al. 2003; de Melo et al. 2016; Melo et al. 2018; Allen et al. 2020; Green et al. 2023).

Shellfish genomics presents unique challenges, as has been observed in Pacific oysters. Segregation distortion and non-Mendelian transmission has been long observed in Pacific oysters (Launey and Hedgecock 2001), and challenged genetic map construction, although this has been largely resolved (Yin et al. 2020) and chromosome-level genomes are now available (e.g., Peñaloza et al. 2021; Qi et al. 2021). Very high polymorphism levels in oysters (Hedgecock et al. 2005; Sauvage et al. 2007; Plough 2016) pose challenges for some approaches, for example due to the frequency of null alleles (Hedgecock et al. 2004). Putative null alleles observed in a low-density amplicon panel impede parentage assignments, necessitating their removal for optimal performance (Sutherland et al. 2024). Similarly, Guo et al. (2023) used a two-step process to reduce the abundance of null alleles in a SNP chip for eastern oysters. High genetic variation can also be beneficial because genotyping additional variants within amplicon sequences increases statistical power for relationship assignment and extends the utility of amplicon panels to populations that are not part of the original design (Thompson et al. 2024). Additionally, these novel variants are expected to increase the power for imputation (Delomas et al. 2024).

In the present study, we used the Cgig_v.1.0 amplicon panel (Sutherland et al. 2024) for pedigree-based imputation using Pacific oyster families (parents and offspring) from the Molluscan Broodstock Program (MBP) of Oregon State University. These families were selectively bred for field survival in a bay with seasonal outbreaks of ostreid herpesvirus 1 (OsHV-1). Parents were genotyped at high density using a target of 20x coverage whole-genome resequencing. Offspring (n = 240) were genotyped with the amplicon panel. The objectives of this study were two-fold: (1) to evaluate the accuracy of genome-wide imputation using the newly developed amplicon panel; and (2) to conduct a genome-wide association study (GWAS) using imputed offspring genotypes to identify regions of the genome associated with resistance to OsHV-1. The offspring samples (n = 240) were part of a laboratory OsHV-1 exposure trial and have survival-associated phenotypes, but many of the mortalities in the challenge had poor DNA quality for whole-genome resequencing, impeding the ability to conduct a GWAS. However, the DNA of these juveniles was of sufficient quality for amplicon panel genotyping, and therefore we used imputation to enable the GWAS. Offspring with sufficient high-quality DNA for whole-genome resequencing (n = 176) were used to evaluate imputation accuracy and further evaluate differential characteristics of imputed genotypes. To evaluate the effect of including a third generation (grandparents) on the imputation accuracy, an *in silico* panel was constructed across three generations for use in imputation, and results were compared to the two-generation empirical amplicon panel results. Collectively these results informed the reliability of this approach for

generating high-density imputed data for use in large- and small-scale Pacific oyster breeding programs and related analyses.

Methods

Pacific oyster families and OsHV-1 laboratory challenge

Offspring from four Pacific oyster families (F114-F117) from the Molluscan Broodstock Program (MBP) were used in an exposure trial to ostreid herpesvirus-1 (n = 60 offspring per family), as previously described (Lunda et al. *in review*). The eight parents of the four families were heterozygous for the OsHV-1 survival/ resistance marker on chromosome 8 (Divilov et al. 2023b; Lunda et al. *in review;* Surry et al. 2024). For each exposed animal, the recorded phenotypes included the survival state at the end of the 16-day trial (dead or alive) and the days post-exposure (DPE) in which mortality occurred. To differentiate individuals that died on the final day of the trial (day 16), individuals that survived until the end of the trial were assigned a value of 17 DPE for the mortality date. All samples were preserved in DNA/RNA Shield (Zymo Research) upon detection of mortality (open valves) or moribundity (no response to tactile stimulation) during daily tank inspections.

DNA extraction and library prep

Genomic DNA was extracted from offspring (n = 240) and parents (n = 8), and quantity and quality were assessed by fluorescence (Qubit) and 1% agarose gel electrophoresis, respectively. Offspring that passed quality evaluations were used for whole-genome resequencing in paired-end 150 bp reads to a target depth of 10x, and all parents were sequenced to a target depth of 20x on a NovaSeq sequencer (Illumina). Extractions and library preparations were conducted by the Centre for Aquaculture Technologies (CAT), San Diego, California (USA).

High-density genotyping of parents and offspring

Whole-genome resequencing reads were analyzed using the pipeline *wgrs_workflow* (see Data Availability). Samples were checked for quality using FastQC and MultiQC (Andrews 2010; Ewels et al. 2016), trimmed with cutadapt (v.3.5; Martin 2011) in paired-end mode to remove Illumina adapters (maximum adapter error rate = 20%), low-quality bases, and too-short reads (flags: -q 15 --trim-n and -m 50), and then re-inspected for quality. Trimmed reads were then aligned to a chromosome-level Pacific oyster reference genome (GCF_902806645.1; Peñaloza et al. 2021) using bwa mem (Li 2013) with read group identifiers added per sample. SAMtools (Li 2011) was used to convert SAM files to BAM format, filter for quality (-q 10), as well as to sort and index the alignment files. Duplicate alignments were marked

146 and tallied using the Picard Toolkit (Broad Institute 2024). Reads from all samples (offspring and parents)
147 were merged into a single BAM file using SAMtools in preparation for genotyping.
148 Genotyping was conducted with bcftools mpileup (Danecek et al. 2021) with base alignment
149 quality (BAQ) applied (-D) and a maximum depth (-d) per merged BAM file of 11,500 reads. Variants were
150 annotated by mpileup with allele depth and total depth, and the bcftools function call was used in
151 multiallelic caller mode (-m) to output variant sites (-v) and annotate genotype quality (--annotate GQ) to
152 an output BCF file. The BCF file was filtered using bcftools to retain variants that were not within 5 bp of
153 an indel and to retain those with less than 10% missing data across all individuals. Only SNPs were retained.
154 SNPs required a quality score greater than 99 for at least one individual, an average depth across all samples
155 of at least 10 reads, and only biallelic SNPs were retained. For each individual, any genotyped site with
156 fewer than five reads, more than 100 reads, or per-site genotype quality of less than 20 were set to missing,
157 and then the missing-data filter was re-applied across individuals (i.e., retained if <10% was missing across
158 samples). The filtered data were then screened to remove any loci with a minor allele frequency (MAF) <
159 0.05, considering all genotyped parents and offspring together. This MAF-filtered dataset was used for all
160 downstream applications. Whole genome genotypes (i.e., the HD dataset) for parents were used for
161 downstream imputation, and the HD dataset for offspring was used only to evaluate imputation accuracy
162 (Figure 1; see below).
163
164 *Amplicon panel library preparation, sequencing and genotyping*
165 Genomic DNA samples from the 240 offspring were normalized and used for AmpliSeq library preparation
166 (Thermo Fisher) with the *Cgig_v.1.0* amplicon panel (Sutherland et al. 2024) (SKU A58237 AGRISEQ
167 PACIFIC OYSTER PANEL) by the Center for Aquaculture Technologies (San Diego). Parents (n = 8) were
168 previously sequenced using the same panel (Sutherland et al. 2024). Output fastq or bam files were provided
169 for the offspring and parents, respectively, and bam files were converted to fastq using the bamtofastq
170 function of bedtools (Quinlan and Hall 2010). If replicate samples were present, the replicate with the
171 highest sequencing depth was retained. Raw fastq data were aligned to the chromosome-level reference
172 genome (GCF_902806645.1) using bwa mem, assigning read group identifiers and quality filtering to
173 maintain alignments with MAPQ > 1. TorrentSuite (Thermo Fisher) VariantCaller VCF files were also
174 obtained with the *Cgig_v.1.0* hotspot and regions file.
175 Aligned samples were used to genotype target and novel variants *de novo* (see Data Availability,
176 *amplitools*). Alignment files were merged using SAMtools, and genotyping was conducted using bcftools
177 mpileup and call, annotating the BCF file with allelic information and genotype quality values. Filtering of
178 variants was conducted as described above for WGS data, except that the missing data filter allowed up to
179 15% missing data across samples, and the minimum and maximum depth cutoffs per genotype were at least

10 reads and less than 10,000 reads, respectively. The MAF-filtered dataset was used for all downstream applications. Amplicon panel genotypes were used in a principal component analysis (PCA) using the function *glPCA* of adegenet (v.2.1.5; Jombart and Ahmed 2011) to confirm clustering of families in R (R Core Team 2025). The *de novo* genotypes for parents and offspring are referred to as the LD panel and were used for downstream imputation (Figure 1; see below).

Identification of Mendelian incompatibilities, unexpected polymorphism, and platform concordance

In the LD panel and HD datasets, parent and offspring trios were used to identify loci displaying Mendelian incompatibilities (MI) in each platform for quality checking genotypes. MI were identified using the Mendelian function of bcftools with a pedigree file of putative trios, and MI were tallied per individual or per locus. Parent-offspring trios were also used in the HD dataset as an additional genotype quality check to identify loci that were unexpectedly polymorphic in offspring when parents were monomorphic.

As an additional quality check for genotypes, the genotype concordance of the LD panel and HD dataset was evaluated for loci common to both platforms. Shared loci were identified using bcftools isec, and compared using bcftools stats, including the *verbose* flag for per-site concordance evaluation. Custom R scripts within the *impute_workflow* repository (see Data Availability) were used to inspect the results of bcftools stats, including the proportion of concordant genotypes per individual and per locus and the Pearson correlation of allelic dosage per individual.

Grandparent genotype data and in silico panel construction

Grandparent RADseq sequence data for the families used in this study (F114-F117) were obtained from NCBI SRA (Divilov et al. 2023b) (see Data Availability), aligned to the reference genome (GCF_902806645.1), and genotyped as described above with bcftools mpileup and call. Family identifiers for each parent of families F114-F117 were provided by Sutherland et al. (2024; see Table 4). However, grandparent samples were not genotyped using the LD amplicon panel, as genomic DNA from these individuals was not available at the time of sample preparation. Therefore, an *in silico* LD panel was constructed (Figure 1) to evaluate the effect of having three generations in imputation. Shared loci between grandparents (RAD-seq data) and the parents and offspring (whole-genome resequence data, n = 176 offspring) were identified using bcftools isec (with flag --collapse none). All shared loci were used for the grandparents and parents, but to simulate an LD panel, offspring samples were subset to only include a randomly selected 1,000 loci and all other loci were set to missing (Figure 1).

The three generation *in silico* dataset was used for imputation to determine the effect of three generations on imputation accuracy (see below). This dataset was then subset to only include the parents (HD data) and offspring (LD *in silico* panel) and imputed to estimate the accuracy of imputation with two

generations. In parallel, the regular empirical dataset with the full HD parents and LD amplicon panel offspring was subsampled to only include 1,000 LD panel loci and a similar number of HD loci to emulate the *in silico* dataset for a full comparison with a similar ratio of HD and LD loci. Accuracy was only evaluated on imputed loci by excluding LD panel loci using bcftools isec (see below for Accuracy evaluation).

Imputation

To construct the empirical imputation input dataset, the LD amplicon panel data for parents and offspring were combined with the HD dataset (whole-genome resequencing) from the parents. To combine the LD and HD datasets, all LD loci that were also present in the parent HD data were removed from the HD data using the bcftools isec function (with flag --*collapse all* to remove loci regardless of allelic identity). The LD and HD parent data were then combined using bcftools concat. The parent HD+LD data were then combined with the LD offspring data using bcftools merge (Figure 1).

The BCF file containing imputation input data (parents, LD+HD; offspring, LD) was converted to AlphaImpute2 format (Whalen and Hickey 2020) using custom R scripts, and a pedigree file was prepared. The ten Pacific oyster chromosomes were separated into individual files, AlphaImpute2 was run on each chromosome individually, and the results were combined into a single file. The data were similarly prepared for FImpute3 (Sargolzaei et al. 2014) to be run on individual chromosomes. The analysis pipeline with all custom scripts is provided (see Data Availability, *impute_workflow*). The FImpute3 imputation process was also conducted for the three- and two-generations *in silico* panel, and the equal ratio HD and LD loci empirical dataset (see above).

Imputation and empirical data concordance

Imputed genotypes per chromosome were combined into a single file and converted to a VCF file using custom scripts (see Data Availability). The HD data for the offspring, not used in imputation, were limited to only include loci in common with the imputed data and then directly compared using bcftools stats and summarized and visualized using custom scripts. Loci present in both imputed and empirical data were also used to compare per-locus allele frequencies and calculate per-sample genotype proportions (i.e., proportion of homozygous reference, heterozygous, homozygous alternate, and missing; Figure 1).

Genome-wide association study (GWAS)

Three offspring genotype datasets were used to conduct GWAS analyses based on survival to the OsHV-1 laboratory challenge: (1) amplicon panel, *de novo* loci (LD data) only; (2) empirical LD panel imputed to high density by AlphaImpute2; and (3) empirical LD panel imputed to high density by FImpute3. Amplicon

248 panel or imputed genotypes were converted to GEMMA input formats (Zhou and Stephens 2012) using
249 custom scripts (see Data Availability, *ms_cgig_chr8_oshv1* and *impute_workflow*, respectively). GEMMA
250 was used to calculate kinship between samples and then to calculate associations per locus to the binary
251 survival phenotype. Analyses were restricted to loci with MAF > 0.05. Manhattan plots were generated
252 using fastman (Paria et al. 2022). All LD loci were highlighted in Manhattan plots using fastman.
253

254 **Results**

255 *High-density (HD) genotyping of parents and offspring by whole-genome resequencing*
256 Parents (n = 8) from the Molluscan Broodstock Program "chromosome 8" selected families F114-F117
257 were sequenced to a target depth of 20x coverage. Parents had a per-sample average (± s.d.) number of read
258 pairs of 67.4 ± 44.9 M (range = 32.7-174.7 M; Additional File S1). This equates to an approximate coverage
259 of 33.7 ± 22.4x (range = 16.4-87.4x). All samples had at least 98.9% reads passing quality trimming.
260 Aligning against the reference genome resulted in an average (± s.d.) per-sample alignment rate of 77.4 ±
261 0.7%. Parents had an average (± s.d.) duplicate read percentage of 19.0 ± 1.4% (range: 17.0-21.2%) based
262 on estimates from Picard Toolkit (Additional File S1).
263 Offspring (n = 176) of the four families with sufficient DNA quality were sequenced to a target
264 depth of 10x coverage. A group of 64 offspring samples (26.7% of all 240 offspring) did not have sufficient
265 high-quality DNA for whole-genome resequencing because of DNA degradation, often associated with
266 samples that were from mortalities. For the samples that were sequenced, offspring had an average (± s.d.)
267 of 23.0 ± 5.2 M read pairs (range = 3.2-33.9 M; Additional File S1), which is approximately 11.5 ± 2.6x
268 depth per individual (range = 1.6-17.0x). Of the 176 sequenced samples, 24 had less than 10x coverage.
269 Sample alignment rates (± s.d.) were on average 74.9 ± 7.7% and the estimated average percent duplication
270 of read pairs was 15.3 ± 1.6% (range: 11.8-18.9%; see Additional File S1).
271 Genotyping of parents and offspring was conducted using all samples in one analysis (n = 184
272 samples), which resulted in the discovery of 58,941,933 raw variants. Filtering was conducted in a stepwise
273 process (Table S1), and 424,951 biallelic SNPs (0.7%; MAF > 0.05) were retained after all filters were
274 applied. Per sample averages (± s.d.) of the different genotypic states (i.e., homozygous reference,
275 heterozygous, and homozygous alternate) in the offspring were 262.6k (61.8%), 114.7k (27.0%), and 13.8k
276 (3.2%), respectively (Table 1; Additional File S2). These values were similar for the parents (Table 1).
277 Offspring and parents had an average (± s.d.) of 33.9k (8.0%) and 34.9k (8.2%) missing loci per individual,
278 respectively. One parent, 58-9F (dam of F117) had a very high percentage of missing data (45.0% missing;
279 Additional File S3 and Additional File S2).
280 To quality check genotypes, polymorphic loci in the offspring were inspected for polymorphism in
281 their respective parents. An average of 360,877 loci per family had no missing data in the parents; of these

282 loci, parents had an average of 46.7% polymorphic and offspring 47.8% polymorphic loci (Table S2). An
283 average of 2.2% of the loci that were polymorphic in the offspring were unexpected due to monomorphism
284 in their respective parents (Table S2). Loci that were polymorphic in both the offspring and parents followed
285 a distribution with peaks at 0.25, 0.5, and 0.75 alternate allele frequencies (Figure S1A), whereas loci that
286 were polymorphic in offspring and monomorphic in parents were largely low allele frequency variants
287 (Figure S1B).
288
289 *Low-density (LD) genotyping of parents and offspring with amplicon panel*
290 The parents (n = 8) and offspring (n = 240) genotyped by the amplicon panel had on average (± s.d.) 0.81
291 ± 0.42 M and 0.14 ± 0.14 M reads per sample, respectively. One parent, 65_8F (the dam of F115), had
292 significantly fewer reads than the other samples, with only 0.03 M reads obtained. Alignment rates per
293 sample against the reference genome (GCF_902806645.1) for parents and offspring data were on average
294 (± s.d.) 94.4 ± 0.5% and 91.8 ± 6.3%, respectively. The per-sample read and alignment counts are available
295 in Additional File S1. Genotyping *de novo* resulted in an initial 218,925 raw variants before any filters;
296 after stepwise filtering, 1,257 SNPs were identified (MAF > 0.05; Table S1). The removal of individuals
297 with a low genotyping rate (GR; GR ≤ 70%) resulted in the retention of 214 out of 234 individuals with
298 phenotypic data (Figure S2A). After the low-GR samples were removed, loci had on average 2.0% missing
299 data (Figure S3A), and no loci were monomorphic. A PCA using multi-locus genotypes showed individual
300 clustering by family (Figure S4). This is hereafter referred to as the LD dataset.
301 The standard genotyping pipeline for a ThermoFisher amplicon panel (i.e., with the Torrent
302 VariantCaller; TVC) identified 3,928 total variants including both novel and 'hotspot' targets (i.e., the
303 designed SNP targets of the amplicon panel). Transferring these variants to the chromosome-level assembly
304 (GCF_902806645.1) resulted in 3,685 (93.8%) retained SNPs, including 534 hotspot target SNPs. As novel
305 SNPs identified by TVC are only genotyped in samples with at least one non-reference allele, these loci
306 were not useful here because of incomplete calling across samples and were discarded, leaving only hotspot
307 target SNPs (hereafter referred to as the hotspot dataset). A per-sample missing data filter (GR ≥ 70%) was
308 applied to the hotspot dataset, resulting in the retention of 184 out of 240 offspring samples (76.7%; Figure
309 S2B). Filtering per locus for missing data (GR ≥ 70%) removed 101 SNPs, leaving 407 SNPs in the dataset
310 (Figure S3B). Of these SNPs, 153 were monomorphic and were removed, leaving 254 loci in the filtered
311 hotspot dataset.
312
313 *Genotyping concordance in loci shared by HD and LD datasets*
314 Of the 1,257 SNPs in the LD data, 26 were also present in the filtered HD data, and these loci were used to
315 compare concordance between the platforms in 183 samples (n = 8 parents; n = 175 offspring). The average

316 per individual genotype concordance (GC) was 97.1 ± 3.6% (Figure 2A) and allelic dosage Pearson
317 correlation was r = 0.940 ± 0.164 (Figure 2B). The average per-locus GC was 97.0 ± 4.0% (Figure 2C)
318 (Additional File S4).
319
320 *Mendelian incompatibilities in HD and LD parent-offspring trios*
321 Mendelian incompatibilities (MI) observed from parent-offspring trios indicated that on average (± s.d.),
322 each trio had 4.8 ± 3.4% of the evaluable loci exhibiting MI in the LD panel dataset (n = 240 trios) and
323 0.3% ± 0.1% in the HD dataset (n = 176 trios; Additional File S5). High MI loci in the panel data included
324 11 loci exhibiting MI in ≥ 30% of the trios and 254 loci (or 20.2% of total LD panel SNPs) that exhibited
325 MI in ≥ 10% of trios. High MI loci in the HD dataset were less frequent, with only four loci exhibiting MI
326 in ≥ 30% of the trios and 3,765 loci (or 0.89% of HD data SNPs) that exhibited MI in ≥ 10% of trios. The
327 percentage of MI was notably lower for family F117, with an average of 0.06% MI loci per individual
328 (Figure S5).
329
330 *Imputation and accuracy evaluation*
331 Of the 424,951 SNPs in the HD+LD dataset, 402,115 were located on chromosomes and were therefore
332 used for imputation. This included 400,896 and 400,889 loci shared with the HD offspring dataset for
333 AlphaImpute2 and FImpute3, respectively, which could be used to evaluate imputation accuracy. For the
334 176 offspring with HD data, the average (± s.d.) GC per sample between the HD and imputed data was 73.6
335 ± 2.3% and 75.3 ± 1.9% for AlphaImpute2 and FImpute3, respectively (Table 2; Figure 2) and the average
336 per-sample allelic dosage Pearson correlation was r = 0.584 ± 0.033 and r = 0.626 ± 0.021, respectively
337 (Figure 2).
338 Proportions of each genotype per individual were summarized for imputed data using only the
339 samples that were also summarized for empirical data (see above). The imputed data resulted in an average
340 (± s.d.) per offspring sample homozygous reference, heterozygous, and homozygous alternate genotypes of
341 341.6k (85%), 27.8k (6.9%), and 32.8k (8.2%), respectively for AlphaImpute2 imputed data, and 345.0k
342 (85.8%), 28.2k (7.0%), and 28.9k (7.2%), respectively for FImpute3 imputed data (Table 1; Additional File
343 S3; Additional File S2). Therefore, compared to empirical data, imputed data by FImpute3 had 24% more
344 homozygous reference genotypes, 20% fewer heterozygotes, and 4% more homozygous alternates. The
345 results for AlphaImpute2 were very similar (Table 1).
346 Allele frequency differences were explored between the empirical and imputed datasets by
347 comparing all 400.9k loci and 176 offspring common to the empirical and imputed datasets. The empirical
348 dataset had an average (± s.d.) allele frequency of 0.181 ± 0.199 (median = 0.106), whereas the imputed
349 (e.g., FImpute3) had an average of 0.115 ± 0.269 (median = 0; Table 3). On average, the empirical minus

350 imputed AF per locus was 0.066 (median = 0.078), and the empirical data consistently had higher allele
351 frequencies than the imputed data. This was true for both datasets. Many low AF variants were imputed as
352 fixed homozygous reference, and high AF as fixed homozygous alternate (Figure S6).
353
354 *Grandparent generation genotyping, in silico panel development, and imputation evaluation*
355 Published grandparent RADseq sequence data (Divilov et al. 2023b) were genotyped using an approach
356 similar to that used for whole-genome resequencing data (see Methods), producing 128,300 quality-filtered
357 SNPs, of which 4,976 loci were shared with the whole-genome resequencing dataset. The 4,976 loci were
358 used to create an *in silico* LD panel of 1,000 loci present in all three generations (grandparents, parents, and
359 offspring), and all 4,976 loci were present in grandparents and parents. After removal of non-chromosomal
360 loci and the loci present in the *in silico* panel, 3,733 imputed loci were retained to evaluate accuracy.
361 The three-generation *in silico* dataset had a GC of 84.5 ± 3.4% with an allelic dosage correlation
362 Pearson r = 0.804 ± 0.03. This was approximately 10% higher in GC and a 0.18 higher r-value than the best
363 results with the two-generation amplicon panel dataset (Table 2). These values were reduced to GC = 79.3
364 ± 2.9% and r = 0.728 ± 0.03 when only the two generations (parents and offspring) were included in the *in*
365 *silico* dataset, which was still significantly higher than the amplicon panel dataset. However, the down-
366 sampled empirical amplicon panel dataset with a similar ratio of LD and HD loci to the ratio in the *in silico*
367 analysis yielded a GC of 76.3 ± 3.0% and r = 0.693 ± 0.03, which is very similar to the *in silico* dataset
368 with two generations and slightly higher in correlation r-value (+0.07) than the full empirical amplicon
369 panel dataset (Table 2). Therefore, the three-generation *in silico* dataset provided the best results, and the
370 two-generation *in silico* dataset was similar to the two-generation amplicon panel datasets; however, there
371 was a slight advantage (as observed by higher allelic dosage but not GC) when the LD panel was more
372 similar in number to the number of imputed loci.
373
374 *GWAS analyses for OsHV-1 survival*
375 Of the 240 samples and four families, families F114 and F115 had lower mortality rates (13.3% and 17.5%,
376 respectively) than F116 and F117 (64.9% and 51.7%, respectively; Table 4; Figure S7). Of the 240
377 offspring, six did not have recorded phenotypes at the end of the experiment and were removed (n = 234).
378 Since only 176 samples were available from the whole-genome resequencing (HD) dataset, including only
379 28 of the 86 mortality samples (32.6%), the per-family sample sizes per phenotype were not expected to
380 provide sufficient power for a GWAS analysis using the HD dataset alone and the amplicon panel-based
381 imputation was explored to recover the samples (Figure 1).
382 The imputed datasets included all samples with phenotypes (n = 86 mortalities and 148 survivors).
383 GEMMA analyzed 58,556 of the 402,115 SNPs (14.6%) imputed by AlphaImpute2 and 73,984 of the

402,108 SNPs (18.4%) imputed by FImpute3. Both imputed datasets showed significant associations near the distal region of Chr8 (AlphaImpute2: 36 Mbp; FImpute3: 35 and 43 Mbp), which was also indicated by the amplicon panel alone (at 40 Mbp; Figure 3; see Additional File S6 for GEMMA association results). Both AlphaImpute2 and FImpute3 datasets also showed a significant association closer to the proximal end of the chromosome, although these associations were represented by fewer markers (FImpute3: 16 Mbp, 1 marker; AlphaImpute2: 12 and 16 Mbp, 4 markers). Interestingly, the imputation by AlphaImpute2 did not provide usable loci for Chr1, as these were all removed by GEMMA during the analysis (Figure 3B), and this chromosome was notably low in amplicon panel loci.

Discussion

Here we used a low-density Pacific oyster amplicon panel to genotype four families of selectively bred oysters with measured OsHV-1 survival phenotypes. We genotyped the parents of the families using whole-genome resequencing (WGRS) and imputed genome-wide genotypes for offspring. We then characterized imputation accuracy using a subset of the same offspring that were genotyped by WGRS but not used for imputation, and determined if the imputed genotypes could be used in a genome-wide association study (GWAS) for OsHV-1 survival to improve our understanding of the genetic architecture of this trait. Below we discuss (1) details of imputation accuracies; (2) the utility for GWAS and new insights on the genetic architecture of the OsHV-1 resistance in the selectively bred families; and (3) the potential for these results to guide future studies in GS in Pacific oysters, considered to be the next frontier for genetic improvement in agricultural and aquaculture species (Meuwissen et al. 2016; Voss-Fels et al. 2019; Houston et al. 2020; Yáñez et al. 2022).

Accuracy and characterization of imputed data

Imputation accuracy was affected by the number of generations present in the dataset. This was demonstrated in a comparison of imputation accuracy using the two-generation amplicon panel dataset and the three-generation *in silico* dataset. Because we only had two generations of data genotyped with the amplicon panel, we were not able to achieve the maximum imputation accuracy that would be expected with the inclusion of the grandparent data at high density (and with the low-density panel). With the amplicon panel data, the best imputation accuracy was obtained with FImpute3, which resulted in genotype concordance (GC) of 75.3% and allelic dosage correlation of r = 0.626. In contrast, the three-generation *in silico* panel imputation had significantly higher imputation accuracy (GC = 84.5% and r = 0.804). This was expected, as *in silico* studies indicate a minimum requirement of two parental generations genotyped at high density (HD) for maximal imputation accuracy (Delomas et al. 2023). Higher haplotype resolution can

be obtained from deeper pedigrees (Fraslin et al. 2023). An additional detail observed here was that when the amplicon panel two-generation data was reduced to a similar ratio of low-density (LD) empirical loci to HD imputed loci the allelic dosage correlation was improved (r = 0.693) relative to the full dataset. Considering that the results were affected by both the number of generations present at HD, and the ratio of HD:LD loci, we expected substantially better results with three generations included; however, concordance may not be as high as that observed in the *in silico* panel dataset here because of the larger HD to LD ratio.

Imputation accuracy has been previously investigated in Pacific oysters *in silico* (Delomas et al. 2023; Kriaridou et al. 2023; Delomas et al. 2024), and the empirical application here brought new insights. First, we observed that it was necessary to use *de novo* genotyping for the amplicon panel (as per Thompson et al. 2024) instead of only the default target SNPs to obtain a sufficient number of high-quality genotypes and retained samples after quality filters (i.e., *de novo* allowed retention of 214 offspring with 1,257 SNPs, whereas the default target would have only retained 184 offspring and 254 SNPs). This result is in agreement with the findings of Delomas et al. (2024). Similar to Thompson et al. (2024), the *de novo* genotyping is particularly valuable when applying the panel to populations other than those included in the panel design, including populations that have been subject to extensive selective breeding. This is promising, as it demonstrates that the panel can be used broadly and will likely have greater longevity than if the target variants were the only focus.

Another insight from this empirical study was the effect of using a different platform for the LD and HD genotypes. Even though the few loci shared between the platforms had high concordance, there was a significantly higher proportion of loci exhibiting Mendelian incompatibilities (MI) in the amplicon panel data than in the WGRS data (i.e., a 20x higher proportion of loci exhibited MI in at least 10% of inspected trios in the panel). This may be due to the higher prevalence of null alleles in the amplicon panel data than in the WGRS data. Putative null alleles have been previously identified in the amplicon panel (Sutherland et al. 2024; Thompson et al. 2024), were determined to be largely population-specific, and were demonstrated to improve parentage assignments once removed (Sutherland et al. 2024). Null alleles also occur in SNP arrays for the eastern oyster (Guo et al. 2023; Xuereb et al. 2023), and may be caused by mutations in primer sites (Hedgecock et al. 2004) and/or structural variations such as small and large deletions (Guo et al. 2023). MIs may also be caused by *de novo* mutations, or other forms of genotyping error, such as false homozygotes in low-quality samples (Xuereb et al. 2023). Guo et al. (2023) suggested that the main cause of null alleles in an eastern oyster SNP chip was deletions, in part because clusters of null alleles were observed in the genome that were flanked by loci exhibiting regular inheritance. However, in the present study it is likely that the cause of prevalent null alleles was the impeded amplification of

amplicons due to disrupted primer sites, since a high prevalence was not observed in the WGRS data. These empirical observations will help guide future studies that depend on amplicon panel-based imputation.

From the detailed investigation of the characteristics of the imputed data, we found additional insights beyond accuracy statistics that may be useful for future consideration. It is possible that imputation inaccuracy and challenges, specifically in Pacific oysters (Kriaridou et al. 2023), may be related to the high polymorphism level in the species, possibly either directly within software due to a mismatch with software assumptions, or indirectly due to the challenge in obtaining high-quality genotypes. Pacific oyster families will deviate from expected genotype ratios during development, at least in part, due to excessive larval mortalities (Durland et al. 2021) that are due to high genetic loads and negative allelic combinations (Plough 2016). This may affect imputation accuracy in pedigree-informed imputation. In the present study, genotypes followed expected modes of MAF that would occur from a biparental cross; however, there was also substantial per-locus variation around the modes. Regardless of the cause, lower imputation accuracy in Pacific oysters relative to finfish is expected based on comparative *in silico* analyses of multiple species (Kriaridou et al. 2023). The accuracy of genotype calls is the main goal of imputation; in the present study, we observed clear and systematic skews in genotype proportions between the imputed and empirical datasets, with a significant enrichment for homozygotes and depletion of heterozygotes in the imputed data, alongside a systematic reduction in allele frequency in the imputed data. The accuracy presented here in the two-generation, amplicon-panel-based dataset ($r = 0.626$) was lower than that obtained by Kriaridou et al. (2023) in a two-generation, *in silico* dataset ($r = 0.775$), and Delomas et al. (2023) in a two-generation, *in silico* dataset with 250 or 2,000 loci in the LD panel ($r = 0.7$ and 0.85, respectively). It is also worth noting that strikingly high concordance has been presented for imputation based on low-coverage (~2.8x) whole-genome resequencing for Pacific oysters (GC = 93.6% and $R^2 = 0.860$; Yang et al. 2024).

We expect that future studies using at least three generations will improve accuracy, but it is worth considering whether the results generated here are sufficient for downstream applications, especially considering cost savings with the present approach relative to other approaches. Although per-sample costs are expected to vary and may change over time, the general averages that are available for the species point to major savings that can be obtained by using the amplicon panel combined with imputation, considering amplicon panel genotyping costs at $15 USD, RAD-seq at $35 USD, SNP chip genotyping at $60 USD, and whole-genome resequencing (10X coverage) at $90 USD (also see Meek and Larson 2019). As an example, genotyping 240 individuals would cost $3,600 by amplicon sequencing, $8,400 by RAD-sequencing, $14,400 by SNP chip, or $21,600 by WGRS. To conduct imputation, additional high-density sequencing for imputation purposes would need to be added to these costs, for example, for parents and grandparents by WGRS.

484 *Applying amplicon panel-based imputation in a GWAS for OsHV-1 resistance*
485 Using amplicon panel-based imputation, we were able to rescue a significant number of samples that could
486 not be genotyped using WGRS. These samples were significantly degraded and comprised mostly of
487 mortalities from an OsHV-1 infection challenge. Shellfish are expected to undergo rapid DNA degradation
488 post-mortem, and genomic DNA extraction from oysters can be challenging (Adema 2021). As a result, 64
489 of the 240 phenotyped samples could not be genotyped using WGRS, significantly reducing the power for
490 GWAS. As the amplicon panel was less sensitive to degraded DNA, *de novo* genotyping and imputation
491 were conducted, which identified a significant association with OsHV-1 survivorship on the expected
492 chromosome for these families (i.e., Chr8; Divilov et al. 2023b, 2023a). However, the main observed peak
493 was not located near the previously reported chromosomal position (~9 Mbp) but rather was at 35-40 Mbp
494 on Chr8, near the distal end of the chromosome. Further investigation is required to identify the location of
495 the causative locus that provides high survival, but it is promising to see that the imputed genotypes
496 identified a major effect locus on the expected chromosome.
497 Uncertainty about the predictability of the Chr8:9 Mbp marker (Divilov et al. 2023b) was also
498 indicated in the present study, as all four families in the exposure trial were expected to have similar
499 genotype frequencies at the locus (all parents were heterozygous for the SNP), but significant variation in
500 mortality was observed between the families. Two of the four families (F114 and F115) had very low
501 mortality rates. This reduced GWAS detection power due to the low number of individuals with the
502 mortality phenotype, but also raises the question of whether the Chr8:9 Mbp marker is still predictive of
503 OsHV-1 survival in these families. The original identification of the survivorship locus at ~9 Mbp was
504 performed using the phenotype of field survival (Divilov et al. 2023b), whereas the present study directly
505 used OsHV-1 infection survivorship. Therefore, it is possible that the field survival phenotype was
506 influenced by additional factors, resulting in a phenotype that would differ from lab-challenge survival.
507 However, the presence of the alternate allele at the ~9 Mbp marker was associated with survivorship in the
508 same infection trial as the present study in different families with offspring that were either all homozygous
509 reference or all heterozygous for the locus (Lunda et al. *in review*), which provides evidence that the locus
510 is indeed directly related to OsHV-1 survival.
511 It is possible that the reason for the apparent movement of the associated locus may be related to
512 imputation inaccuracies, but this does not explain the two families with higher-than-expected survival rates.
513 Also, the associated region at 35-40 Mbp was identified using the *de novo* panel loci only before imputation
514 was conducted. Investigation of the genetic underpinnings of this trait will need to be continued in future
515 work. This highlights a limitation of linked markers in GWAS rather than causative markers, which are
516 often very challenging to identify; the linkage can break down if the causative locus is not identified, and
517 the predictive utility of a linked marker may be reduced or lost.

Applying amplicon panel-based imputation in future genomic selection approaches

Although the application of GWAS was of interest in the current study, the main proposed applications of imputation in supporting breeding are for genomic selection (GS) (e.g., Delomas et al. 2023; Kriaridou et al. 2023; Delomas et al. 2024). Therefore, it is relevant to discuss whether imputation accuracy obtained in the present study, as described above, is sufficiently reliable for GS. A recent study on rainbow trout *Oncorhynchus mykiss* demonstrated that even with lower imputation accuracy (e.g., r = 0.58 for a 300 SNP LD panel) genomic prediction accuracy did not strongly deviate from that obtained with higher accuracy imputation (Fraslin et al. 2023). Given that we exceeded this with the two-generation imputation here (i.e., r = 0.626), and that this would be easily improved with the addition of a third generation genotyped at HD, this is promising for the application of the amplicon panel-based imputation for GS. This level of accuracy was observed by Fraslin et al. (2023) using between 300-500 SNPs in rainbow trout, and in the simulations by Delomas et al. (2023) using approximately 100 loci and only two generations. Varying degrees of imputation accuracy have resulted in robust GS results in Pacific oyster simulations (Delomas et al. 2023). Fraslin et al. (2023) attributed the robustness of genomic prediction, even in the presence of imputation inaccuracies, to the strong linkage disequilibrium that occurs through the use of GS on close relatives (e.g., full-sibs), which is typical in aquaculture studies. This results in increased short- and long-range linkages and large shared haplotype blocks. It would be worthwhile to consider recombination rate variation in Pacific oysters to evaluate how large these shared haplotype blocks are across sibs. Yang et al. (2024) also found the highest imputation accuracy in aquaculture lines rather than outbred wild lines, which was attributed to increased LD from the family structure. From their study that demonstrates high effectiveness for GS with low marker numbers, Fraslin et al. (2023) propose using 300-500 SNPs in low-cost amplicon panels combined with imputation from HD relatives, which is what we have conducted here. The use of amplicon panels in this manner brings the additional benefit of being able to confirm pedigrees through parentage during GS (Fraslin et al. 2023; Sutherland et al. 2024), or even to add putatively functional loci to enable MAS alongside GS.

Conclusions

This study is the first empirical demonstration of genome-wide imputation in Pacific oysters using a low-density amplicon panel combined with whole-genome resequencing data, and these results are expected to guide future work in genomic selection for this and other shellfish species. The cost of whole-genome resequencing library preparation and sequencing, even at low coverage, is prohibitive for small-scale operations. Aquaculture focuses on a wide variety of species, many of which are produced in regionally distinct environments, and economies of scale are sparse in aquaculture. As demonstrated here, the

amplicon panel and imputation approach can be used for genotyping and expanded to a genome-wide scale, which will enable small operations to use advanced GWAS and GS methods and large operations to expand genotyping coverage across their programs. In this work, we provided tissue to a sequencing facility for DNA extraction and sequencing, providing raw sequence outputs; therefore, very limited in-house laboratory techniques are needed to accomplish genotyping. Improvements in accuracies observed here are expected, particularly through the inclusion of additional parental generations to deepen the pedigree, but even at the levels in the present study, GS is enabled. The varied phenotypic responses to disease of the families studied here, combined with the identification of a putative separate major effect locus on Chr8, points to a more complex genetic architecture for OsHV-1 survival, and further studies are needed to understand the drivers of this important phenotype.

Funding

This work was funded by the Pacific States Marine Fisheries Commission (PSMFC) project "Preparing for Future Challenges – Threats from Ocean Acidification, *Vibrio coralliilyticus* and OsHV-1 µvar to West Coast Oyster Farmers" (Award No. NA18NMF4720007). Additional funding was provided by the Canada Research Chair in Shellfish Health and Genetics.

Acknowledgements

Thanks to the staff at the Center for Aquaculture Technologies (CAT) in San Diego and Prince Edward Island for providing information regarding the OsHV-1 challenge. The U.S. Department of Agriculture (USDA) is an equal opportunity provider and employer. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsements by the USDA.

Competing Interests

BJGS is affiliated with Sutherland Bioinformatics. The authors have no competing financial interests to declare.

Data Availability

The following repositories support this study:

Whole-genome resequencing workflow: https://github.com/bensutherland/wgrs_workflow

Imputation workflow: https://github.com/bensutherland/imputation_workflow

Project-specific scripts and instructions: https://github.com/bensutherland/ms_cgig_chr8_oshv1

585 Genotype files and associated metadata is available on FigShare: 10.6084/m9.figshare.28095440
586 Raw amplicon panel sequence data and whole-genome resequencing data for the parents and offspring are
587 available under the NCBI SRA BioProject PRJNA1219245. Raw sequencing data for the grandparents
588 were previously published and is available under the NCBI SRA accession number PRJNA873124,
589 including the dam and sire for families 30.055 (SRR21192124 and SRR21192123), 30.058
590 (SRR21192118 and SRR21192117), 30.065 (SRR21192209 and SRR21192208), and 30.079
591 (SRR21192111 and SRR21192110), respectively.
592

593 **Additional Files**

594 Additional Figures and Tables are provided in the Supplemental Results.
595
596 Additional File S1. Per-sample number of reads, alignments, alignment rate, and duplicate level for whole-
597 genome resequencing data.
598 Additional File S2. Per-sample proportions of genotypes homozygous reference (0/0), heterozygous (0/1),
599 and homozygous alternate (1/1), or missing data in (A) HD data (empirical whole-genome resequencing
600 data) for parents (n = 8) and available offspring (n = 176); (B) imputed data by AlphaImpute2; or (C)
601 imputed by FImpute3 (imputed includes eight parents and all 240 offspring).
602 Additional File S3. Individual genotype counts for samples present in both the empirical and imputed data.
603 Additional File S4. Per locus concordance between platforms (panel vs. whole-genome resequencing) or
604 empirical and imputed (both AI2 and FI3).
605 Additional File S5. Per-sample and per-locus Mendelian incompatibility counts for trios in the LD data
606 (amplicon panel) or HD data (whole-genome resequencing).
607 Additional File S6. Output of GEMMA analyses for the AlphaImpute2, FImpute3, and *de novo* amplicon
608 panel analyses.

REFERENCES

Adema, C.M., 2021 Sticky problems: extraction of nucleic acids from molluscs. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376 (1825):20200162.

Allen, S.K., C. Rexroad, and R. Rheault, 2020 Improving production of the eastern oyster *Crassostrea virginica* through coordination of genetic improvement programs, research, and technology transfer activities. *Journal of Shellfish Research* 39 (2):175-179, 175.

Allen, S.K., J.M. Small, and P.D. Kube, 2021 Genetic parameters for Crassostrea virginica and their application to family-based breeding in the mid-Atlantic, USA. *Aquaculture* 538:736578.

Andrews, S., 2010 FastQC: A quality control tool for high throughput sequence data.

Botta, R., F. Asche, J.S. Borsum, and E.V. Camp, 2020 A review of global oyster aquaculture production and consumption. *Marine Policy* 117:103952.

Boudry, P., F. Allal, M.L. Aslam, L. Bargelloni, T.P. Bean *et al.*, 2021 Current status and potential of genomic selection to improve selective breeding in the main aquaculture species of International Council for the Exploration of the Sea (ICES) member countries. *Aquaculture Reports* 20:100700.

Broad Institute, 2024 Picard toolkit, GitHub Respository.

Cowan, M.W., C.M. Pearce, T. Finston, G.R. Meyer, R. Marshall *et al.*, 2023 Role of the *Vibrio* community, reproductive effort, and environmental parameters in intertidal Pacific oyster summer mortality in British Columbia, Canada. *Aquaculture* 565:739094.

Cowan, M.W., C.M. Pearce, T.J. Green, T. Finston, G.R. Meyer *et al.*, 2024 Abundance of *Vibrio aestuarianus*, water temperature, and stocking density are associated with summer mortality of Pacific oysters in suspended culture. *Aquaculture International* 32 (4):5045-5066.

Danecek, P., J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan *et al.*, 2021 Twelve years of SAMtools and BCFtools. *GigaScience* 10 (2).

de Melo, C.M.R., E. Durland, and C. Langdon, 2016 Improvements in desirable traits of the Pacific oyster, *Crassostrea gigas*, as a result of five generations of selection on the West Coast, USA. *Aquaculture* 460:105-115.

Delomas, T.A., C.M. Hollenbeck, J.L. Matt, and N.F. Thompson, 2023 Evaluating cost-effective genotyping strategies for genomic selection in oysters. *Aquaculture* 562:738844.

Delomas, T.A., C.M. Hollenbeck, J.L. Matt, and N.F. Thompson, 2024 Microhaplotypes generate higher breeding value accuracy compared to SNPs for imputation-based breeding strategies. *Aquaculture* 586:740779.

Divilov, K., N. Merz, B. Schoolfield, T.J. Green, and C. Langdon, 2023a Genome-wide allele frequency studies in Pacific oyster families identify candidate genes for tolerance to ostreid herpesvirus 1 (OsHV-1). *BMC Genomics* 24 (1):631.

Divilov, K., N. Merz, B. Schoolfield, T.J. Green, and C. Langdon, 2023b Marker-assisted selection in a Pacific oyster population for an antiviral QTL conferring increased survival to OsHV-1 mortality events in Tomales Bay. *Aquaculture* 567:739291.

Durland, E., P. De Wit, E. Meyer, and C. Langdon, 2021 Larval development in the Pacific oyster and the impacts of ocean acidification: Differential genetic effects in wild and domesticated stocks. *Evolutionary Applications* 14 (9):2258-2272.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller, 2016 MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32 (19):3047-3048.

Fraslin, C., D. Robledo, A. Kause, and R.D. Houston, 2023 Potential of low-density genotype imputation for cost-efficient genomic selection for resistance to *Flavobacterium columnare* in rainbow trout (*Oncorhynchus mykiss*). *Genetics Selection Evolution* 55 (1):59.

Green, K.M., A.K. Spalding, M. Ward, A. Levine, E.A. Wolters *et al.*, 2023 Oregon shellfish farmers: Perceptions of stressors, adaptive strategies, and policy linkages. *Ocean & Coastal Management* 234:106475.

Green, T.J., N. Siboni, W.L. King, M. Labbate, J.R. Seymour *et al.*, 2019 Simulated marine heat wave alters abundance and structure of Vibrio populations associated with the Pacific oyster resulting in a mass mortality event. *Microbial Ecology* 77 (3):736-747.

Guo, X., J.B. Puritz, Z. Wang, D. Proestou, S. Allen *et al.*, 2023 Development and evaluation of high-density SNP arrays for the eastern oyster *Crassostrea virginica*. *Marine Biotechnology* 25 (1):174-191.

Habier, D., R.L. Fernando, and J.C. Dekkers, 2009 Genomic selection using low-density marker panels. *Genetics* 182 (1):343-353.

Hedgecock, D., P.M. Gaffney, P. Goulletquer, X. Guo, K. Reece *et al.*, 2005 The case for sequencing the Pacific oyster genome. *Journal of Shellfish Research* 24 (2):429-441, 413.

Hedgecock, D., G. Li, S. Hubert, K. Bucklin, and V. Ribes, 2004 Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, Crassostrea gigas. *Journal of Shellfish Research* 23 (2):379-386.

Hollenbeck, C.M., and I.A. Johnston, 2018 Genomic tools and selective breeding in molluscs. *Frontiers in Genetics* 9.

Houston, R.D., T.P. Bean, D.J. Macqueen, M.K. Gundappa, Y.H. Jin *et al.*, 2020 Harnessing genomics to fast-track genetic improvement in aquaculture. *Nature Reviews Genetics* 21 (7):389-409.

Hu, B., Y. Tian, Q. Li, and S. Liu, 2022 Genomic signatures of artificial selection in the Pacific oyster, *Crassostrea gigas*. *Evolutionary Applications* 15 (4):618-630.

Jombart, T., and I. Ahmed, 2011 adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27 (21):3070-3071.

King, W.L., C. Jenkins, J.R. Seymour, and M. Labbate, 2019 Oyster disease in a changing environment: decrypting the link between pathogen, microbiome and environment. *Marine Environmental Research* 143:124-140.

Kriaridou, C., S. Tsairidou, C. Fraslin, G. Gorjanc, M.E. Looseley *et al.*, 2023 Evaluation of low-density SNP panels and imputation for cost-effective genomic selection in four aquaculture species. *Frontiers in Genetics* 14.

Langdon, C., F. Evans, D. Jacobson, and M. Blouin, 2003 Yields of cultured Pacific oysters *Crassostrea gigas* Thunberg improved after one generation of selection. *Aquaculture* 220 (1):227-244.

Launey, S., and D. Hedgecock, 2001 High genetic load in the Pacific oyster *Crassostrea gigas*. *Genetics* 159 (1):255-265.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27 (21):2987-2993.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, pp. arXiv:1303.3997.

Lunda, S.L., K. Divilov, T.J. Green, B.J.G. Sutherland, X. Wang *et al.*, *in review* Selection for a single nucleotide polymorphism on chromosome 8 of the Crassostrea gigas genome improves resistance to both ostreid herpesvirus 1 and *Vibrio coralliilyticus*.

Mao, J., Y. Tian, Q. Liu, D. Li, X. Ge *et al.*, 2024 Revealing genetic diversity, population structure, and selection signatures of the Pacific oyster in Dalian by whole-genome resequencing. *Frontiers in Ecology and Evolution* 11.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17 (1):3.

Martínez-García, M.F., J.L. Ruesink, J.M. Grijalva-Chon, C. Lodeiros, J.A. Arreola-Lizárraga *et al.*, 2022 Socioecological factors related to aquaculture introductions and production of Pacific oysters (*Crassostrea gigas*) worldwide. *Reviews in Aquaculture* 14 (2):613-629.

Meek, M.H., and W.A. Larson, 2019 The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molecular Ecology Resources* 19 (4):795-803.

Melo, C.M.R.D., R. Morvezen, E. Durland, and C. Langdon, 2018 Genetic by environment interactions for harvest traits of the Pacific oyster *Crassostrea gigas* (Thunberg) across different environments on the West Coast, USA. *Journal of Shellfish Research* 37 (1):49-61, 13.

Meuwissen, T., B. Hayes, and M. Goddard, 2016 Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers* 6 (1):6-14.

Paria, S.S., S.R. Rahman, and K. Adhikari, 2022 fastman: A fast algorithm for visualizing GWAS results using Manhattan and Q-Q plots. *bioRxiv*:2022.2004.2019.488738.

Peñaloza, C., A.P. Gutierrez, L. Eöry, S. Wang, X. Guo *et al.*, 2021 A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. *GigaScience* 10 (3).

Plough, L.V., 2016 Genetic load in marine animals: a review. *Current Zoology* 62 (6):567-579.

Qi, H., L. Li, and G. Zhang, 2021 Construction of a chromosome-level genome and variation map for the Pacific oyster *Crassostrea gigas*. *Molecular Ecology Resources* 21 (5):1670-1685.

Quinlan, A.R., and I.M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6):841-842.

R Core Team, 2025 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raymond, W.W., J.S. Barber, M.N. Dethier, H.A. Hayford, C.D.G. Harley *et al.*, 2022 Assessment of the impacts of an unprecedented heatwave on intertidal shellfish of the Salish Sea. *Ecology* 103 (10):e3798.

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel, 2014 A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15 (1):478.

Sauvage, C., N. Bierne, S. Lapègue, and P. Boudry, 2007 Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene* 406 (1):13-22.

Surry, L.B., B.J.G. Sutherland, S.L. Lunda, A.H. Loudon, K. Divilov *et al.*, 2024 The presence of the Pacific oyster OsHV-1 resistance marker on chromosome 8 does not impact susceptibility to infection by *Vibrio aestuarianus*. *bioRxiv*:2024.2011.2025.625178.

Sutherland, B.J.G., N.F. Thompson, L.B. Surry, K. Reddy Gujjula, C.D. Carrasco *et al.*, 2024 An amplicon panel for high-throughput and low-cost genotyping of Pacific oyster. *G3: Genes|Genomes|Genetics*:jkae125.

Thompson, N.F., B.J.G. Sutherland, T.J. Green, and T.A. Delomas, 2024 A free lunch: microhaplotype discovery in an existing amplicon panel improves parentage assignment for the highly polymorphic Pacific oyster. *G3: Genes Genomes Genetics*:jkae280.

Voss-Fels, K.P., M. Cooper, and B.J. Hayes, 2019 Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics* 132 (3):669-686.

Whalen, A., and J.M. Hickey, 2020 AlphaImpute2: Fast and accurate pedigree and population based imputation for hundreds of thousands of individuals in livestock populations. *bioRxiv*:2020.2009.2016.299677.

Xuereb, A., R.M. Nahuelpi, E. Normandeau, C. Babin, M. Laporte *et al.*, 2023 Design and validation of a high-density single nucleotide polymorphism array for the Eastern oyster (*Crassostrea virginica*). *G3 Genes|Genomes|Genetics* 13 (6).

Yáñez, J.M., P. Xu, R. Carvalheiro, and B. Hayes, 2022 Genomics applied to livestock and aquaculture breeding. *Evolutionary Applications* 15 (4):517-522.

Yang, B., Y. Li, Q. Li, and S. Liu, 2024 High-throughput and cost-effective genotyping by low-coverage whole genome sequencing with genotype imputation in Pacific oyster, Crassostrea gigas. *Aquaculture* 591:741134.

Yin, X., A. Arias-Pérez, T.H. Kitapci, and D. Hedgecock, 2020 High-density linkage maps based on genotyping-by-sequencing (GBS) confirm a chromosome-level genome assembly and reveal variation in recombination rate for the Pacific oyster *Crassostrea gigas*. *G3 Genes|Genomes|Genetics* 10 (12):4691-4705.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44 (7):821-824.

FIGURES

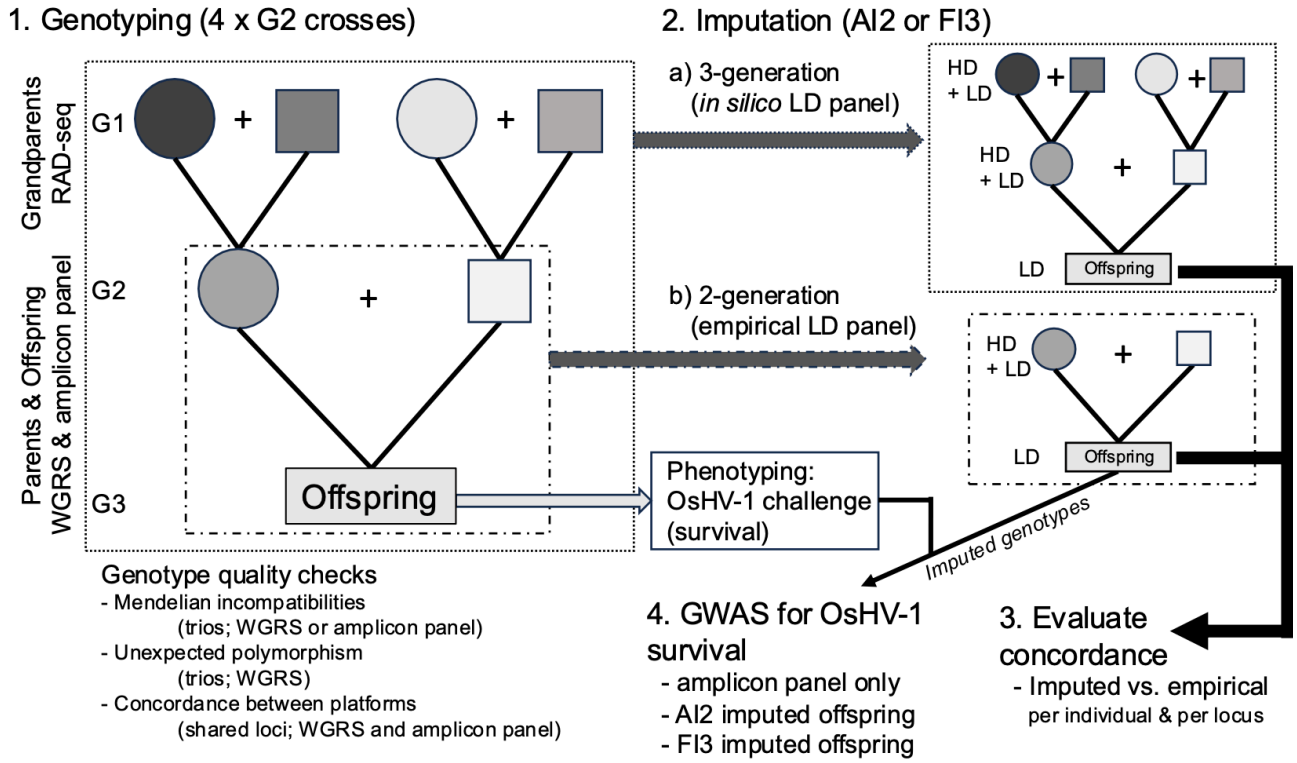

Figure 1. Workflow overview considering three- or two-generation (G1-G3) datasets and imputation by AlphaImpute2 (AI2) or FImpute3 (FI3). Dams are shown as circles and sires as squares. The imputation outputs from all datasets were evaluated for accuracy by comparisons with empirically genotyped offspring, but only the two-generation imputed genotypes were used as an input for a genome-wide association study for OsHV-1 survival.

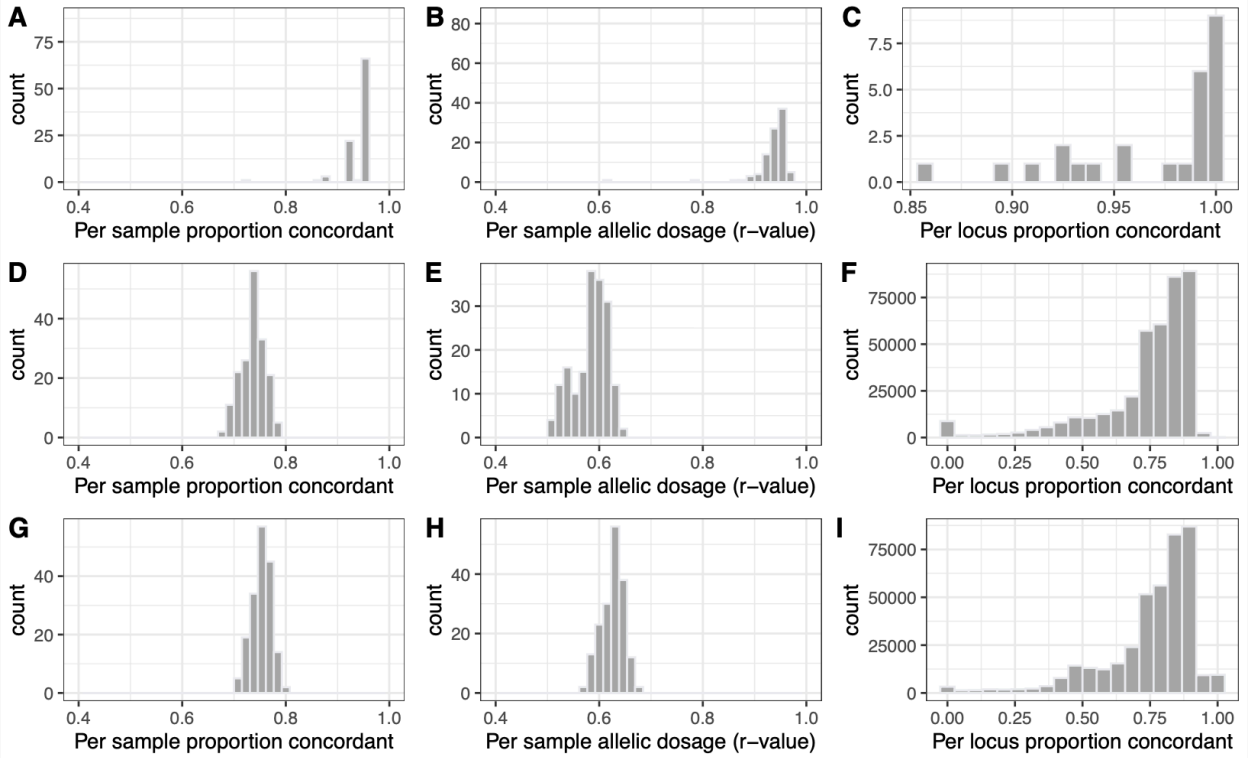

Figure 2. Average per-sample proportions of concordant genotypes, per-sample correlations of allelic dosage, and per-locus concordant genotypes were determined for (A-C) loci empirically genotyped independently by both the amplicon panel and whole-genome resequencing (n = 26 loci and 184 indiv.); (D-F) for the dataset imputed by AlphaImpute2; or (G-I) by FImpute3 relative to empirical genotypes from whole-genome resequencing data (n = 400.9k loci; 176 individuals).

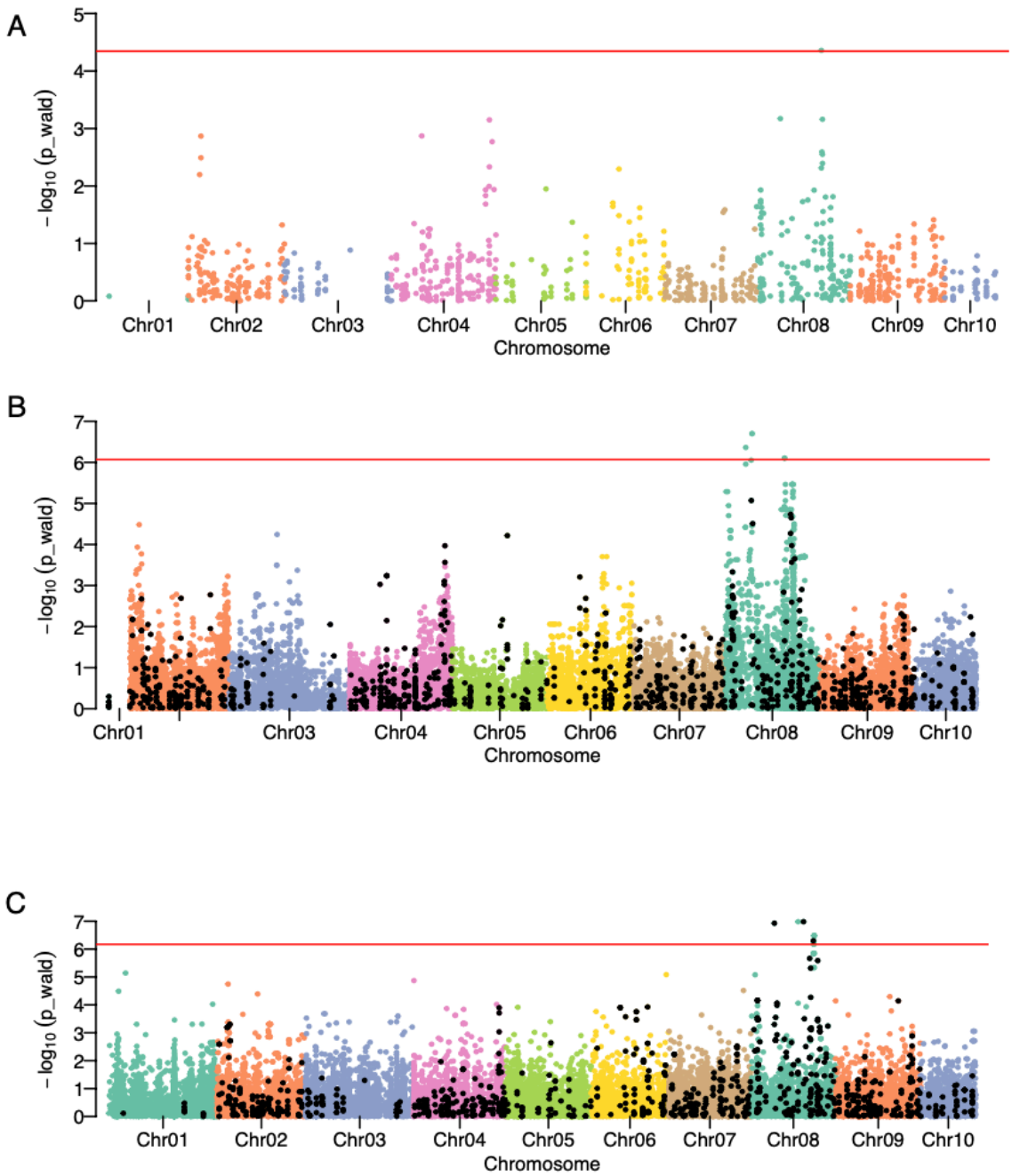

Figure 3. Manhattan plots for survival after exposure to OsHV-1 in (A) amplicon panel *de novo* SNPs (n = 1,117 loci and 214 individuals); (B) imputed data with AlphaImpute2 (n = 58,556 loci and 234 individuals); and (C) imputed data with FImpute3 (n = 73,984 loci and 234 individuals). The red horizontal line represents the genome-wide Bonferonni-corrected p-value significance threshold of 0.05.

TABLES

Table 1. Average per-sample genotype counts (k = thousand loci) and percentages of total loci for offspring or parents in the empirical data (whole-genome resequencing, HD data; n = 176 samples) or in the offspring data imputed by AlphaImpute2 (AI2) or FImpute3 (FI3; n = 240).

	Offspring Empirical, Count (k)	**%**	**Parent Empirical, Count (k)**	**%**	**AI2 Imputed, Count (k)**	**%**	**FI3, Imputed Count (k)**	**%**
REF/REF	262.6	61.8	262.3	61.7	341.6	85.0	345.0	85.8
REF/ALT	114.7	27.0	113.3	26.7	27.8	6.9	28.2	7.0
ALT/ALT	13.8	3.2	14.6	3.4	32.8	8.2	28.9	7.2
missing	33.9	8.0	34.9	8.2	0	0	0	0
total	425.0	-	425.0	-	402.1	-	402.1	-

Table 2. Imputation evaluation for all datasets including the standard workflow (‘*WGRS and amplicon panel, full dataset*’) with two generations with imputation by either AlphaImpute2 (AI2) or FImpute3 (FI3). Additionally shown are the three- and two-generation datasets generated *in silico* using the whole-genome resequencing and RADseq data (‘*WGRS + RAD, and in silico panel*’), as well as the WGRS data reduced to be a similar proportion of loci in the HD and LD data as the *in silico* datasets (‘*WGRS and amplicon panel, subset*’). Evaluation metrics included average (± s.d.) per-sample genotype concordance and per-sample allelic dosage (Pearson r-value). LD and imputed loci only include the counts of loci on chromosomes used in the imputation, whereas the imputed loci only include those not present in the LD panel.

Dataset	**WGRS and amplicon panel, full dataset**		**WGRS + RAD, and *in-silico* panel**		**WGRS and amplicon panel, subset**
Generations analyzed	2	2	3	2	2
Imputation program	AI2	FI3	FI3	FI3	FI3
LD loci	1,257	1,257	924	924	990
Imputed loci	400,896	424,925	3,733	3,733	3,733
Per-sample concordance (%)	73.6 ± 2.3 %	75.3 ± 1.9 %	84.5 ± 3.4 %	79.3 ± 2.9 %	76.3 ± 3.0 %
Per-sample allelic dosage (r-value)	0.584 ± 0.033	0.626 ± 0.021	0.804 ± 0.031	0.728 ± 0.025	0.693 ± 0.027

Table 3. The average (± s.d.) and median allele frequencies of 400.9k loci in the empirical, AlphaImpute2 (AI2) or FImpute3 (FI3) imputed datasets, as well as the per-locus AF difference (empirical minus imputed).

	Empirical AF	**Imputed (AI2) AF**	**Empirical minus imputed (AI2) AF**	**Imputed (FI3) AF**	**Empirical minus imputed (FI3) AF**
Average	0.181	0.107	0.074	0.115	0.066
s.d.	0.199	0.249	0.105	0.269	0.138
Median	0.106	0	0.073	0	0.078

Table 4. Sample sizes and percentage of total available samples retained per family and survival in the OsHV-1 trial (M = mortality; S = survivor) in four datasets: whole-genome resequencing data; amplicon panel data using target hotspots only; amplicon panel using *de novo* identified SNPs; and imputed data using high-density parents and the *de novo* SNPs in the low-density panel. The amplicon panel datasets were filtered at a per-individual genotyping rate of 70%. Low retention of samples in the WGRS data was due to poor DNA quality of extracted samples.

	F114		**F115**		**F116**		**F117**		
Dataset	**M**	**S**	**M**	**S**	**M**	**S**	**M**	**S**	**TOTAL**
WGRS only	4 (50%)	52 (100%)	2 (20%)	47 (100%)	11 (29.7%)	20 (100%)	11 (35.5%)	29 (100%)	176 (75.2%)
Amp. panel, hotspots	3 (37.5%)	51 (98.1%)	5 (50%)	47 (100%)	28 (75.7%)	19 (95%)	12 (38.7%)	19 (65.5%)	184 (78.6%)
Amp. panel, de novo	8 (100%)	52 (100%)	9 (90%)	47 (100%)	34 (91.9%)	20 (100%)	22 (71%)	28 (96.6%)	220 (94.0%)
Imputed	8 (100%)	52 (100%)	10 (100%)	47 (100%)	37 (100%)	20 (100%)	31 (100%)	29 (100%)	234 (100%)