Genotype imputation in F2 crosses of inbred lines

Saul Pierotti[1]*, Bettina Welz[2], Tomas Fitzgerald[1], Joachim Wittbrodt[2], Ewan Birney[1]*

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI); Cambridge, CB10 1SD, UK

[2] Centre for Organismal Studies (COS), Heidelberg University; Heidelberg, 69120, Germany

* Corresponding authors. Email: saul@ebi.ac.uk, birney@ebi.ac.uk

Abstract

Motivation: Crosses among inbred lines are a fundamental tool for the discovery of genetic loci associated with phenotypes of interest. In organisms for which large reference panels or SNP chips are not available, imputation from low-pass whole-genome sequencing is an effective method for obtaining genotype data from a large number of individuals. To date, a structured analysis of the conditions required for optimal genotype imputation has not been performed.

Results: We report a systematic exploration of the effect of several design variables on imputation performance in F2 crosses of inbred medaka lines. We determined that, depending on the number of samples, imputation performance reaches a plateau when increasing the per-sample sequencing coverage. We also systematically explored the tradeoffs between cost, imputation accuracy, and sample numbers. We developed a computational pipeline to streamline the process, enabling other researchers to perform a similar cost-benefit analysis on their population of interest.

Availability and implementation: The source code for the pipeline is available at https://github.com/birneylab/stitchimpute.

Introduction

Pure inbred lines and structured crosses have been the bedrock of early genetics (Mendel 1866) and they have been used in the earliest organisms explored in the field (Morgan 1910; Aida 1921). Over time, phenotype-based markers have been replaced by anonymous DNA variation, such as Restriction Fragments Length Polymorphisms (RFLPs) (Grodzicker et al. 1974) or microsatellite repeats (Tautz 1989; Litt and Luty 1989) and finally Single Nucleotide Polymorphisms (SNPs) (Lander 1996). With the advent of cheap high throughput sequencing, one can use low pass sequencing of subsequent generations (eg, F2 generation) to accurately determine the genotype of an individual. This process is a type of imputation (i.e. the statistical technique to infer missing data), and statistical models have been specifically developed for the linked inheritance patterns of DNA sequence variation. Imputation approaches have been used for plant breeding (Bhattarai et al. 2020; Pang et al. 2020; Taylor et al. 2018), farm animal agricultural studies (Y. Li et al. 2022; Liu et al. 2021; Emrani et al. 2020; Crespo-Piazuelo et al. 2020; Falker-Gieske et al. 2019; S. Li et al. 2017; Schmid et al. 2018; Lutz et al. 2017; Emrani et al. 2017) and model organism studies (Yao et al. 2021; Ackert-Bicknell et al. 2012; Parker et al. 2011).

Most imputation methods in current use rely on a Hidden Markov Model (HMM) where the sample genotypes represent the emission space, and the reference panel haplotypes define

a two-dimensional grid of hidden states where each row is a reference haplotype and each column is a marker. This is known as the Li and Stephens model after their foundational work on statistical modelling of linkage disequilibrium and recombination rates (N. Li and Stephens 2003).

The success of genotype imputation in humans has been made possible by the construction of large reference panels of high-quality genotypes. Milestones in this direction are the HapMap Project (Altshuler et al. 2010), the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), and the Haplotype Reference Consortium (McCarthy et al. 2016). Some of the imputation tools in widespread use that take advantage of reference panels are GLIMPSE (Rubinacci et al. 2021), QUILT (Davies et al. 2021), ShapeIT (Delaneau et al. 2019), Beagle (Browning, Zhou, and Browning 2018), miniMAC (Fuchsberger, Abecasis, and Hinds 2015), and Impute2 (Howie, Donnelly, and Marchini 2009).

The reliance on reference panels has hindered the application of genotype imputation in less-studied organisms that lack extensive reference data. STITCH (Davies et al. 2016) overcomes this limitation by recognising that in populations founded by a limited number of individuals, the unknown ancestral haplotypes are sequenced at relatively high depth even though each single sample is sequenced at low depth. STITCH can impute genotypes from low-depth sequencing data by alternating the Li and Stephens HMM with an expectation maximisation (EM) step where randomly initialised ancestral haplotypes are tuned to maximise the expectation under the model of the observed sample genotypes. This alternating procedure is repeated a parameterised number of times, leading to the reconstruction of the founder haplotypes and the imputation of the sample genotypes. Several hyperparameters need to be optimised for the reconstruction to be optimal, key among them the number of ancestral haplotypes to be reconstructed (K parameter). This optimization can be performed either by relying on quality metrics that are internal to the model (info score) or on external validation (correlation of imputed and known genotypes). Although STITCH has been used successfully in a number of specific F2 cross strategies, often the parameterisation and the assessment of the required level of coverage are presented as a single best choice and with little exploration of alternative parameters and experimental designs (for example, what is the optimal coverage for a particular cross design, and how does this relate to cost?).

We are studying phenotypes in the medaka fish (*Oryzias latipes*), which is a small teleost fish native to Japan, Korea and eastern China (Wittbrodt, Shima, and Schartl 2002). It has long been used as a model organism for genetic research thanks to its comparatively small genome (700.4 Mb) (Kasahara et al. 2007) and to its economic husbandry and tolerance to inbreeding. Taking advantage of these characteristics, an extensive medaka inbred panel has been established from a wild population (Fitzgerald et al. 2022). Inspired by some of the plant breeding strategies, we have been using 8 parental lines, in 10 different cross combinations, for phenotype mapping. A recent example, which we will use in this manuscript, generated a total of 2219 samples (of which 2177 were used in the final analysis).

A practical problem is the optimal imputation of the DNA sequence of each individual. In this work, we optimise STITCH hyperparameters for this type of multi-parent cross by assessing the imputation accuracy with a sample of high-coverage genomes and we explore the optimal parameters for future cost-effective cross designs. We show that a correctly parameterised STITCH pipeline can effectively impute sample genotypes and show that there is a trade-off between the number of individuals sequenced and the per-sample coverage needed.

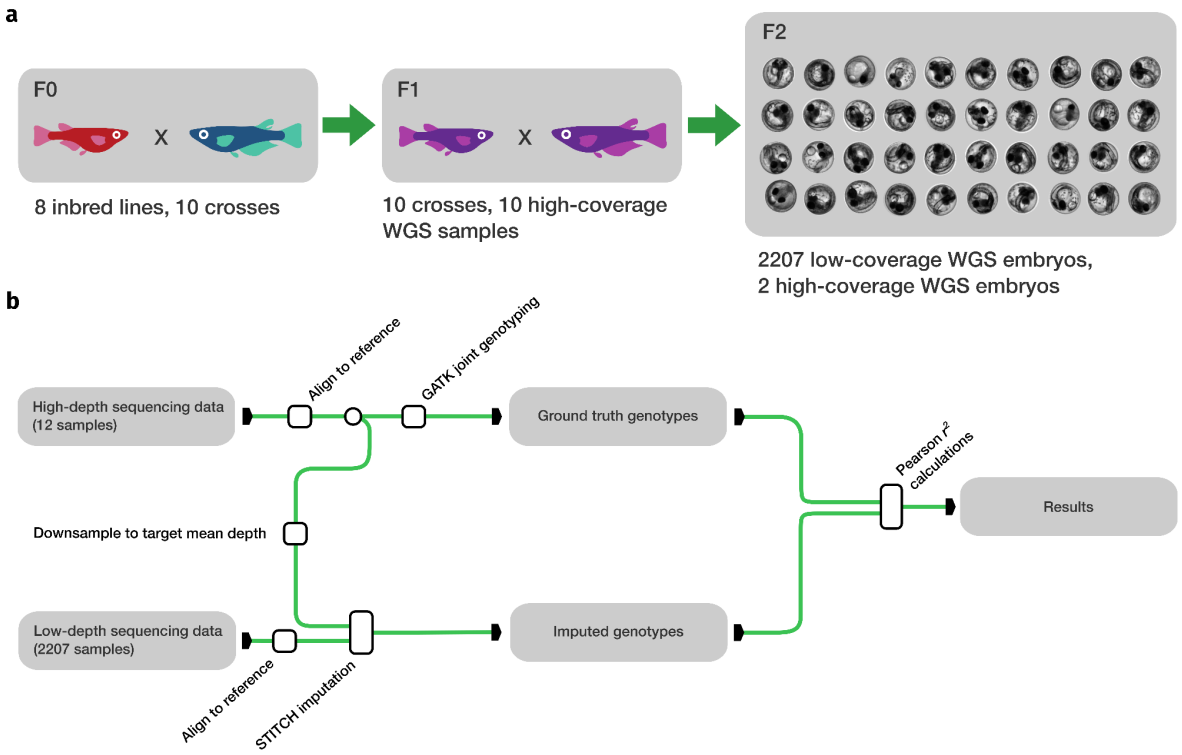

Figure 1. (a) Experimental design. 10 crosses of 8 inbred MIKK panel lines were performed. Among the resulting F1 progeny, 10 samples were sequenced at high coverage (see Methods). The F2 cohort was mostly sequenced at low coverage (2207 samples), but two F2 samples were sequenced at high coverage. WGS: Whole Genome Sequencing. (b) Schematic illustration of the analysis performed. High-depth samples were aligned to the medaka reference genome and jointly genotyped. Mapped reads were downsampled to the desired depth (see main text). Low-coverage samples were also aligned to the medaka reference genome. The low-coverage mapped reads and the mapped downsampled reads for the high-coverage samples were then jointly used for imputation with STITCH. The imputed genotypes for the high-coverage samples were compared to the ground truth genotypes that had been obtained with GATK joint calling for evaluating imputation performance (using the Pearson r^2 as a metric).

Results

We first developed a framework for the optimisation and assessment of imputation results from F2 crosses. We performed high-coverage sequencing on one F1 individual per cross (10 overall) and two F2 individuals. We reasoned that we could use both the F1 individuals

and the F2 individuals as high-coverage "truth sets" for the imputation assessment given that the heterozygous nature of F1 and F2s are equivalent challenges for the imputation task. The F1 individuals are also useful to determine the set of truly polymorphic markers in the cross. Although we have sequenced the F0 founders (Fitzgerald et al. 2022) both previous work and our own investigations suggested that there will be apparent homozygous divergent polymorphisms in the F0s which are not polymorphic in F1s (e.g. in cryptic segmental duplications).

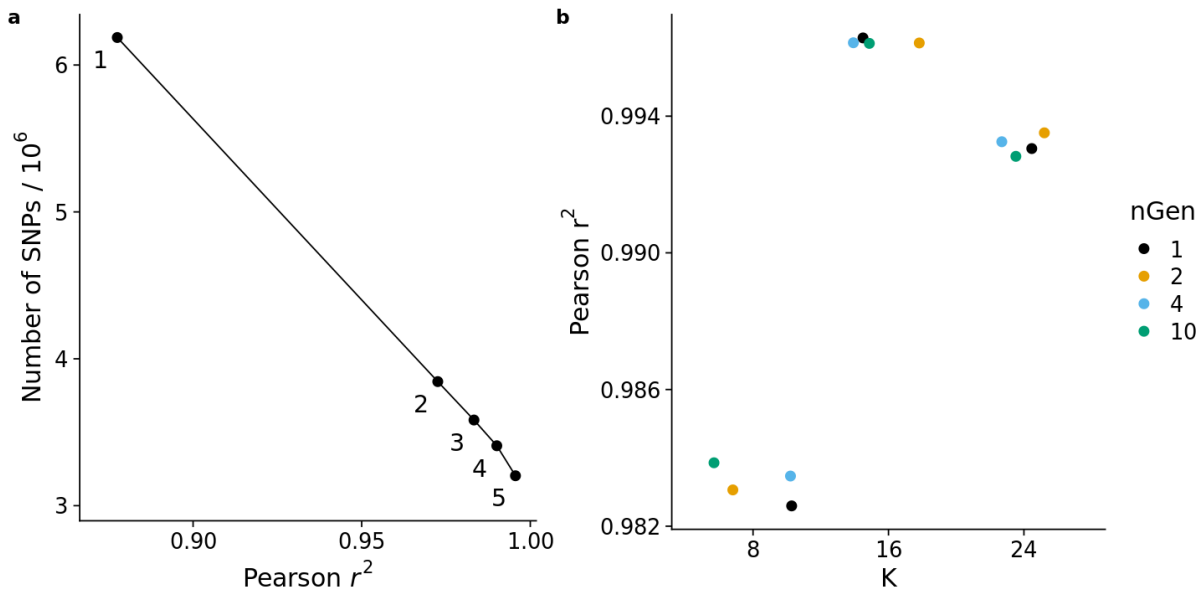

Figure 2. (a) Refinement of the imputed marker set showing the tradeoff between the number of markers retained and imputation performance. The numeric labels refer to the refinement iteration. (b) Imputation performance for different hyperparameter combinations. K is the number of ancestral haplotypes and nGen is the number of generations since the funding of the imputed population. In our analysis, K=16 is optimal and nGen has only a modest effect on the results.

Before using them for imputation, we downsampled the F1 and high-coverage F2 samples to obtain approximately equivalent coverage to the low-depth sequencing data across all F2s (Figure 1). Thus, we had 12 samples imputed from low-coverage data for which we have also a high-quality, high-confidence genotype call. Using this truth set we explored STITCH parameters. In our hands, the two key parameters for a successful imputation were the number of ancestral haplotypes (K parameter) and the selection of valid SNPs to impute. To select the imputable SNPs we started with the SNPs that are polymorphic in the 10 F1 individuals (one F1 per cross) and the two high-coverage F2s. We then refined this set by running the imputation including the downsampled high-coverage datasets, computing the correlation of the imputed SNPs to the high-coverage calls, and rejecting SNPs with a low correlation (Figure 2a). We repeated this process for 5 iterations, progressively increasing the stringency of the correlation threshold (see Methods). For the selection of the best setting for K, we tested a range of values and selected the one that maximised the mean squared Pearson correlation across the 12 truth set samples (Figure 2b). As both K and the

final SNP set parameters are linked, our procedure is to first determine a plausible value for K using the unfiltered SNP set, then doing the SNP filtering (using the value of K determined in the initial assessment), and finally re-assessing the choice of K on the filtered SNP set.

Overall, our F2-STITCH pipeline proved to be highly successful in imputing the genotypes of our medaka cohort. We observed a mean per-sample squared Pearson correlation coefficient (r^2) of 0.996 between the imputed and ground truth genotypes. When binning the SNPs according to their Minor Allele Frequency (MAF), the squared Pearson correlation was as expected higher for bins containing more common SNPs (most likely due to being polymorphic in more than one cross) and decaying for rarer SNPs. Nonetheless, even the bin containing the least common SNPs (MAF 0 to 0.05) still showed an r^2 of 0.977.

Exploring the parameters that influence accurate imputation

Using the pipeline, we then explored downsampling our overall F2 dataset in a variety of dimensions: in the sequencing depth, in the number of individuals overall, in the number of crosses, and in the number of individuals in a single cross.

This explores much of the design space for practical cross strategies. As expected we observed less accurate imputation for lower sequencing depth and lower sample numbers (Figure 3a, 3b), though there is a non-linear response in both cases. For example, we observed a large difference in imputation performance between a coverage of 0.5x and 0.25x, but a comparatively smaller difference when the coverage was changed from 1x to 0.5x. We observed a similar trend when reducing the total number of samples while retaining the original sequencing depth (Figure 3b). In this case, the performance started to drop significantly for cohorts that were smaller than 1000 samples.

On the contrary, reducing the sample numbers by reducing the number of F2 crosses involved (and hence the number of distinct haplotypes to be reconstructed; we changed K proportionally, see Methods) did not significantly affect the quality of the imputation (Figure 3c). When the dataset was reduced to a single F2 cross (and so 2 F0 founder lines) for a total of 474 samples, the r^2 value was observed to be above 0.964 for all the MAF bins except the one with the least common SNPs (which had a much lower r^2 of 0.694). This compares to an r^2 range of 0.777 to 0.915 when using 500 samples randomly chosen from the whole dataset (contrast Figure 3b, “500 random samples” to Figure 3c, “2 founders, 1 cross”, with 474 samples).

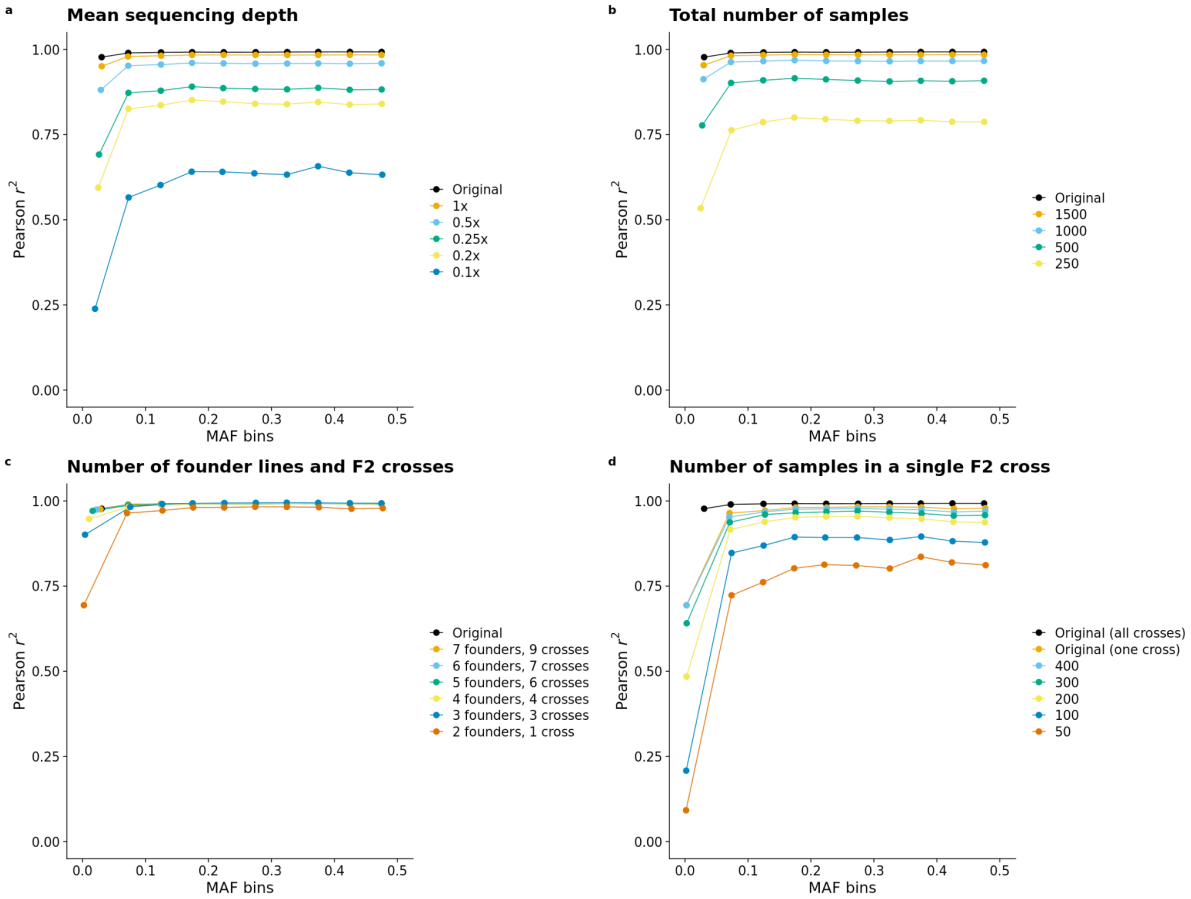

Figure 3. Imputation performance in different scenarios. All plots show the squared Pearson correlation for imputation compared to the truth set for different Minor Allele Frequency (MAF) bins. (a) Varying sequencing depth. (b) Varying the total number of samples jointly imputed. (c) Restricting the imputation to samples derived from a limited number of founder lines and F2 crosses. (d) Imputing a single F2 cross and varying the number of samples within that cross.

Finally, when we explored reducing the number of samples within a single F2 cross (Figure 3d), we observed that when the sample number fell below 200, the r^2 sharply decreased. However, it should be noted that in a population similar to the one presented in this study, different F2 crosses share some common founder lines. As such, STITCH would be able to more accurately impute the genotypes compared to this simplified single-cross scenario by borrowing sequencing coverage of the founders from the other F2 crosses that have common parents.

Cost-benefit analysis of increased sequencing depth for different population sizes

To determine what would be the ideal sequencing depth to aim for when designing similar experiments, we performed a combined downsampling of mean sequencing depth and sample numbers. We observed that for the original cohort size (2177 samples) a maximum sequencing depth of 0.5x was sufficient for obtaining an imputation result that is only

marginally worse than the original (Pearson r^2 0.981 versus 0.996) while almost halving the financial burden (53.7% of the original).

For lower sample numbers, the required minimum sequencing depth proportionally increased (Figure 4). For 1500 and 1000 samples, a depth of 1x was needed to avoid strong degradation of the imputation quality. For sample sizes lower than 500 any sequencing depth downsampling led to much worse imputation results. Also, we noticed that the Pearson r^2 reached a plateau at different levels for different cohort sizes, with larger cohorts reaching the plateau earlier and ending at higher Pearson r^2 values, though the difference between the best achievable correlations is marginal at high sample numbers.

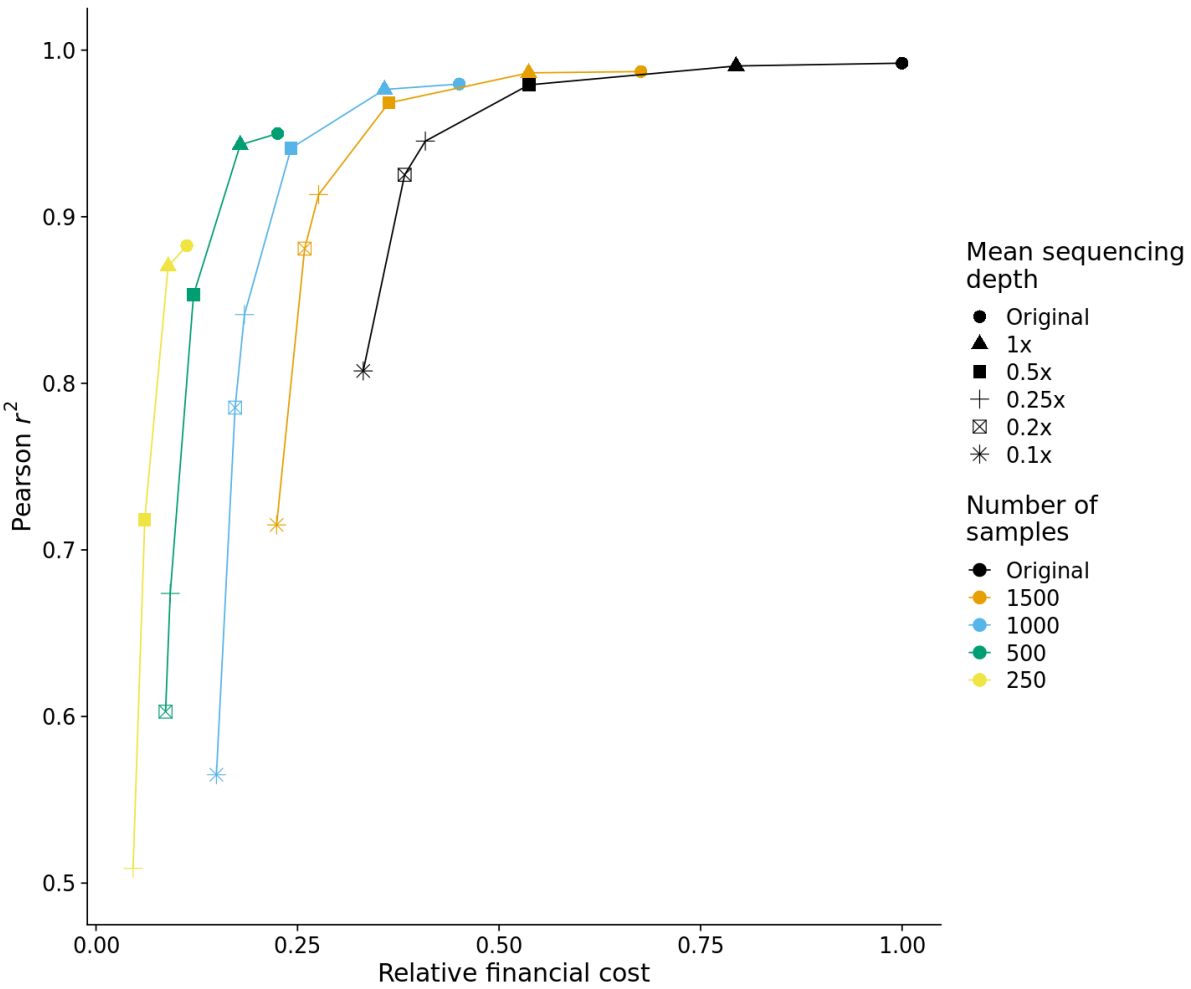

Figure 4. Relative cost of library preparation and sequencing versus imputation accuracy for different combinations of mean sequencing depth and number of samples. Estimated costs are relative to the cost sustained for obtaining the original dataset (see Methods). Depending on the number of samples, the imputation accuracy reaches a plateau at a certain level of mean sequencing depth, where increasing the depth further does not bring appreciable increases in imputation accuracy.

Discussion

In this work, we have systematically explored the imputation accuracy from low-coverage sequencing for the commonplace F2 cross designs used in agriculture, model organisms and plant breeding. Although there have been many studies that have used this strategy, their imputation methods are often presented as the single method that worked in that study without a detailed assessment of the trade-offs of cross structure, sample numbers, and sequencing depth. We also developed an economic model for the sequencing cost, on the assumption that the sample numbers would be fixed by other constraints (e.g. husbandry capacity). There is a relationship between the optimal sequencing depth and the number of samples for accurate imputation, as expected by the coverage of the underlying ancestral haplotypes. In practice, for sample sizes of 1000 aiming for a coverage of 1x at least for this cross structure provided nearly equivalent accuracy compared to our full set. In contrast for 2,000 samples a similar level of accuracy would be achieved with 0.5x. An alternative way of viewing this is that for sample sizes of 2000, there is little to be gained by raising the sequencing depth from 0.5x to 1x.

An important feature in both the practical F2 imputation and our assessment was the sequencing of F1 individuals at high coverage. In our experience this was critical to select an initial set of SNPs which were polymorphic in the crosses. In practice, this means it is important to track and store F1 individuals during the breeding. As not all inbred lines are fully inbred, it is also advantageous to set up F2 crosses between single F0 individuals, and track the relationship between F2 individuals and their specific F0 parents wherever possible. A benefit of STITCH is that it does not require a reference panel, and will impute variants which are present in the F2 regardless of whether they were fixed in the F0 or not.

A consequence of imputing only the highly confident polymorphic SNPs is that many other variants will not be imputed, including CNVs, SVs, insertions, deletions and hard-to-call SNPs. This does not invalidate the accurate imputation of the markers we have, but for downstream applications such as association studies, analysts should remember that the causal variant is not necessarily among the imputed set of markers. In our hands, a successful exploration technique is to group samples by the associated marker of interest and treat the grouped samples as a high-coverage meta-sample to explore other forms of genetic variation linked around a selected marker.

We have made this STITCH-based imputation scheme into a Nextflow (Tommaso et al. 2017) pipeline inspired by the nf-core template (Ewels et al. 2020) and available at https://github.com/birneylab/stitchimpute. This pipeline can take as inputs high-coverage and low-coverage samples, with a specification of any sequencing depth downsampling to be performed (e.g. for later assessment), and will automatically iterate the imputation and performance assessment for building a final SNP set. The pipeline also allows the exploration of the STITCH parameters K (number of haplotypes) and nGen (number of generations) in a systematic manner. It can also be used in a more basic way to run a single imputation with a pre-defined SNP and parameter set. We have successfully used this

pipeline in a number of F2 crosses with different operators and hope it will be useful to a broader set of research groups.

We are currently processing a number of phenotypes on this and other complex cross designs, and are aiming for a comprehensive association study using these imputed genotypes. Although we have explored parameters and downsampling for this particular cross structure in this specific species (medaka fish), different cross structures with different divergence levels of parents and different recombination rates will mean that some level of parameter adjustment will be needed for each system. Our recommendations are: (a) to ensure that there are high-coverage F1 individuals (b) to carefully consider the sequencing depth compared to the number of samples in the F2 and (c) to consider the number of parents and the number of founding haplotypes.

Materials and methods

Fish maintenance

All fish stocks were maintained and bred (fish husbandry, permit number 35–9185.64/BH Wittbrodt) in constant recirculating systems at 26 °C on a 14 hrs light/10 hrs dark cycle. Fish husbandry and experiments were performed in accordance with local animal welfare standards (Tierschutzgesetz §11, Abs. 1, Nr. 1) and with European Union animal welfare guidelines (Bert et al. 2016). The fish facility is under the supervision of the local representative of the animal welfare agency.

DNA extraction

The genomic DNA of the F1 individuals was extracted from whole brain samples. Brains were digested overnight at 60°C with proteinase K (1 mg/ml) in DNA extraction buffer (0.4 M Tris/HCl pH 8.0, 0.15 M NaCl, 0.1% SDS, 5 mM EDTA pH 8.0). Subsequently, the DNA was purified by phenol chloroform isoamyl alcohol (Phenol:Chloroform:Isoamyl (25:24:1) pH 8.0) extraction using phase lock gel (PLG) tubes. After centrifugation the aqueous phase was mixed with Chloroform: Isoamyl Alcohol (24:1) in a new pre-spun PLG tube and centrifuged again. The genomic DNA was precipitated with Isopropanol and washed twice with 70% ethanol. The precipitated DNA was resuspended in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA in RNAse-free water) and stored at 4 °C. For the genomic DNA extraction of the F2 individuals, hatched or manually dechorionated embryos were snap-frozen and stored in deep well plates at -80°C. The fish were lysed in 40 µl DNA extraction buffer (2M Tris pH 8.0, 5M NaCl, 20% SDS, 0.5M EDTA pH 8.0) overnight at 60°C. After a 1:2 dilution of the samples with nuclease-free water, the samples were used for library preparation.

Sequencing library preparation

For the high sequencing coverage samples, a ligation-based library preparation was performed using the NEBNext® Ultra™ II DNA Library Prep Kit (New England Bioalbs) with NEBNext® Multiplex Oligos for Illumina® (Unique Dual Index UMI Adaptors DNA Set 1) according to the manufacturer's instructions without PCR. 750 ng DNA were used as input. The libraries were selected for a fragment size of 500 bp and quantified with the Qubit High

Sensitivity kit. Further, the quality and the molarity of the library were assessed using the Bioanalyzer with the DNA HS Assay kit following the manufacturer's protocol allowing equimolar pooling of libraries for sequencing. For the low sequencing coverage samples, a tagmentation-based library was prepared. The sequencing library preparation was done as described previously (Picelli et al. 2014). For the tagmentation reaction, 1,25 µl DNA was mixed with 1,25 µl of Dimethylformamide, 1,25 µl tagmentation buffer (40 mM Tris-HCl pH 7,5, 40 mM MgCl2) and 1,25 µl of an in-house generated and purified Tn5 (Hennig et al. 2018, diluted to a 1:100 ratio with water) and incubated at 55°C for 3 min. To inactivate the tagmentation reaction the samples were cooled down to 10°C and 1,25µl 0,2 % SDS was added followed by an incubation for 5 min at room temperature. The resulting fragments were amplified by PCR (72°C for 3min, 95°C for 30sec, 12 cycles of 98°C for 20 sec, 58°C for 15 sec and 72°C for 30 sec and a final extension time of 3 min at 72°C) using 6,75 µL of KAPA 2 X HiFi master mix, 0,75 µL of Dimethyl sulfoxide and 2,5 µL of dual indexed primers. PCR products were pooled and selected for an average size of 450 bp using two rounds of magnetic SPRI beads clean up (0.7X and 0.6X respectively). Finally, the library was quantified with the Qubit High Sensitivity kit and the fragment size was assessed using the high sensitivity assay in the Bioanalyzer.

Sample exclusions

Samples sequenced to an overall mean depth lower than 0.1x were excluded from the analysis. This reduced the sample set from 2219 to 2177 samples.

F2 cross design

The population that we designed was founded by eight inbred medaka lines from the Medaka Inbred Kiyosu Karlsruhe panel (MIKK panel) (Fitzgerald et al. 2022). The lines used were 72-2, 55-2, 139-4, 15-1, 62-2, 68-1, 79-2, and 22-1. Those lines were crossed in a set of 10 F2 crosses (72-2 x 79-2: 153 samples, 72-2 x 15-1: 141 samples, 72-2 x 139-4: 153 samples, 139-4 x 72-2: 149 samples, 72-2 x 55-2: 481 samples, 72-2 x 62-2: 142 samples, 72-2 x 68-1: 164 samples, 68-1 x 79-2: 147 samples, 15-1 x 62-2: 158 samples, 55-2 x 139-4: 373 samples, 72-2 x 22-1: 148 samples). The cross 72-2x139-4 was performed reciprocally with males from the 72-2 line and females from the 139-4 line, and vice-versa. The other crosses were performed with males from the first line in the cross name, and females from the second line. One male and multiple females were used for founding each cross.

Sequencing

A total of 2219 medaka samples were sequenced at the Genomics Core facility of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany (https://www.embl.org/groups/genomics). We performed 150 bp paired-end Illumina (https://www.illumina.com/) short-read sequencing on a NextSeq2000 machine, with a variable number of samples per flow-cell. In total, we sequenced our samples on nine P3 Illumina flow-cells and two Illumina P2 flow-cells. 12 samples were sequenced at high depth (33x to 61x) in order to be used as a ground truth for validating the quality of the imputation. Of these, 10 were F1 samples (one per cross, multiplexed on a P3 flow-cell)

while the remaining two were F2 samples (multiplexed on a P2 flow-cell). The 2207 low-coverage samples were sequenced at lower depth (1.4x overall and 9.5x maximum per sample) on eight P3 flowcells (multiplexed with 271, 276, 276, 287, 288, 359, 90, and 90 samples respectively) and one P2 flow-cell (270 samples). The mean per-sample sequencing depth distribution for all the samples is reported in Figure S1.

Alignment and preprocessing

Both low-coverage and high-coverage (ground truth) samples were preprocessed using the nf-core/sarek (Hanssen et al. 2023) pipeline. Reads were mapped to the HdrR medaka reference genome (ENSEMBL ID: ASM223467v1) using bwa-mem2 (Vasimuddin et al. 2019) and deduplicated using GATK MarkDuplicates (Van der Auwera and O'Connor 2020).

Sequencing depth calculations

All the sequencing depths reported in this work have been calculated on the output of GATK MarkDuplicates using Mosdepth (Pedersen and Quinlan 2017), as part of the nf-core/sarek pipeline. The coverage was obtained from the Mosdepth summary output by summing the total number of bases sequenced for each chromosome and dividing by the sum of chromosome lengths. The mitochondrial chromosome was excluded from this calculation.

Sequencing depth downsampling

The original mean depth for each sample ($d_{original}$) was calculated as described above. Given a desired downsampled mean depth (d_{target}), the downsampling factor (f) to be used was calculated as:

$$f = d_{target} / d_{original}$$

Samtools view (Danecek et al. 2021) was then used to downsample the reads with the calculated downsampling factor using the --subsample flag.

Given the variable original mean depth of the low-coverage samples, not all the samples always had an original depth higher than the target depth. In such cases, no downsampling was performed. As a consequence of this, the depth values indicated represent an upper bound on the real mean depth across the cohort.

When assessing the imputation performance on the dataset with original sequencing depth, the ground truth samples were downsampled to a mean depth of 0.5x before being imputed, so as to avoid inflated performance figures driven by the relatively high coverage. When the depth was downsampled, the ground truth samples were also downsampled at the same target depth used for the rest of the cohort. Nonetheless, we observed that the specific coverage of the ground truth samples had a negligible effect on the imputation performance, while the overall coverage of the imputed cohort was much more important.

Sample size downsampling

The number of samples was downsampled in such a way that the ground truth samples were always retained. Among the remaining samples, the fish to be included were chosen randomly until the sample set (complete of ground truth samples) reached the desired size. In the overall downsampling, samples were chosen randomly without taking into account F2 cross membership (Figure 3b).

F2 crosses downsampling

In the downsampling performed in Figure 2c, all the samples belonging to certain F2 crosses were removed from the sample set (including ground truth samples belonging to such crosses). The F2 crosses and founder lines were iteratively downsampled in the following order:

- Start from the full sample set (8 founders, 10 crosses)
- Remove all the crosses involving line 22-1 (7 lines, 9 crosses)
- Remove all the crosses involving line 68-1 (6 lines, 7 crosses)
- Remove all the crosses involving line 79-2 (5 lines, 6 crosses)
- Remove all the crosses involving line 15-1 (4 lines, 4 crosses)
- Remove all the crosses involving line 62-2 (3 lines, 3 crosses)
- Remove all the crosses involving line 139-4 (2 lines, 1 cross)

Single F2 cross downsampling

The single cross downsampling in Figure 3d starts from the subset remaining at the end of the F2 cross downsampling in Figure 3c. This sample set includes only the samples belonging to the cross 72-2 x 55-2. The samples in this cross were then randomly selected while retaining the ground truth samples, until the desired sample size was reached.

Ground truth genotyping

The 12 ground truth samples were genotyped using the nf-core/sarek pipeline. The reads were aligned to the reference genome using bwa-mem2 and deduplicated using GATK MarkDuplicates. Genotypes were obtained with GATK via joint germline variant calling (Poplin et al. 2018).

Imputation

Imputation was performed using STITCH (Davies et al. 2016). The following parameters were used: K=16, nGen=2, expRate=2, niterations=100, shuffleHaplotypeIterations=seq(4, 88, 4), refillIterations=c(6, 10, 14, 18), shuffle_bin_radius=1000. The K parameter was chosen so as to maximise the squared Pearson correlation with the ground truth on the full medaka genome. The other parameters were set by inspecting diagnostic plots produced by the software when imputing a small genomic region in an initial exploratory phase of the

project. These other parameters were determined to have a far smaller impact on imputation accuracy than the K parameter.

When downsampling the number of crosses and founder lines, the K parameter was adjusted proportionally to reflect the fewer number of founder lines present in the population. Consistently with the choice of K=16 determined to work best in the original dataset that was composed of 8 founder lines, K was set to twice the number of founder lines in each downsampling iteration.

When downsampling the number of samples in a single cross, K was set to 4 to reflect the presence of 2 founder lines.

SNP set definition

The set of SNPs used in the imputation was defined iteratively. We started from a set containing all the biallelic SNPs called by GATK joint calling in the ground truth sample set with a minor allele count greater or equal to 1 (6.2 million SNPs). To obtain this we used bcftools view (Danecek et al. 2021) and applied in series the following filtering criteria: -i 'CHROM!="MT"', -i 'TYPE=="snp"', -i 'N_ALT >= 1', -i 'MAC >= 1'. We ran the imputation jointly on the low-coverage and downsampled ground truth samples (0.5x mean depth), and measured the SNP-wise squared Pearson correlation against the ground truth genotypes. We retained SNPs only above a certain threshold of squared correlation and then repeated the imputation on such SNP set. We iteratively refined the SNP set in this way for a total of 5 iterations, using filter values of 0.5, 0.5, 0.75, 0.9, and 0.9. The final SNP set consisted of 3.2 million SNPs. This filtering procedure follows the approach suggested in the original STITCH publication (Davies et al. 2016).

Imputation quality calculations

The quality of the imputation was calculated in terms of squared Pearson correlation (r^2) between the imputed genotypes and the genotypes obtained via joint germline GATK calling on the ground truth samples. GLIMPSE2 Concordance (Rubinacci et al. 2021) was used for this task. The performance was calculated jointly for SNPs in Minor Allele Frequency (MAF) bins of size 0.05. The plotted MAF of each bin represents the average observed MAF of the SNPs that it includes. The performance was also calculated jointly for all the SNPs of a certain ground truth sample. This per-sample performance was averaged across samples to obtain single performance figures per imputation run like the one used in Figure 3.

Relative cost calculations

The relative cost estimates reported in Figure 3 represent the fraction of the original cost that would have approximately been spent for a certain downsampling scenario. These were calculated from the sequencing and library preparation cost at the Genomics Core facility of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. The following formula was used to compute the cost:

$$((n * d * g) / f) * c_f + n * c_p$$

In the above formula, *n* is the number of samples, *d* is the mean sequencing depth, *g* is the haploid genome size, *f* is the mean output per sequencing flow-cell, and c_f and c_p are the unitary costs for sequencing (per flow-cell) and library preparation (per sample).

The mean output per flow-cell was estimated from the mean observed output in our experiment for P3 Illumina flow-cells ($f = 273542732637\,bp$ after deduplication). The medaka haploid genome size ($g = 734040372\,bp$) was obtained by summing the chromosome lengths of the Hdr-R reference genome, mitochondrion excluded. The unitary costs had a ratio $c_f/c_p = 683.89$.

From this, the relative cost was computed by dividing the result by the cost estimated using the number of samples and the mean depth of the original sample set.

Additional software used

Plots were produced using ggplot2 (Wickham 2016) and cowplot (Wilke 2020). JupyterLab was used for interactive explorations (Kluyver et al. 2016). The R programming language was used throughout the pipeline (R Core Team 2023).

Acknowledgements

We thank the Birney and the Wittbrodt labs for their support and constructive feedback on the work and the manuscript. We wish to acknowledge Amos Pierotti for providing the medaka fish illustrations used in Figure 1. We thank Vladimir Benes, Mireia Osuna Lopez, Ferris Jung and all members of the EMBL Genomics Core Facility (GeneCore) for their support in the library preparation and the whole genome sequencing. This research was supported by the European Research Council Synergy Grant IndiGene (Grant agreement ID: 810172).

References

Ackert-Bicknell, Cheryl L., Serkalem Demissie, Shirng-Wern Tsaih, Wesley G. Beamer, L. Adrienne Cupples, Beverly J. Paigen, Yi-Hsiang Hsu, Douglas P. Kiel, and David Karasik. 2012. 'Genetic Variation in TRPS1 May Regulate Hip Geometry as Well as Bone Mineral Density'. *Bone* 50 (5): 1188–95. https://doi.org/10.1016/j.bone.2012.01.011.

Aida, Tatuo. 1921. 'ON THE INHERITANCE OF COLOR IN A FRESH-WATER FISH, APLOCHEILUS LATIPES TEMMICK AND SCHLEGEL, WITH SPECIAL REFERENCE TO SEX-LINKED INHERITANCE'. *Genetics* 6 (6): 554–73. https://doi.org/10.1093/genetics/6.6.554.

Altshuler, David M., Richard A. Gibbs, Leena Peltonen, David M. Altshuler, Richard A. Gibbs, Leena Peltonen, Emmanouil Dermitzakis, et al. 2010. 'Integrating Common and Rare Genetic Variation in Diverse Human Populations'. *Nature* 467 (7311): 52–58. https://doi.org/10.1038/nature09298.

Bert, Bettina, Justyna Chmielewska, Sven Bergmann, Maximilian Busch, Wolfgang Driever, Karin Finger-Baier, Johanna Hößler, et al. 2016. 'Considerations for a European Animal Welfare Standard to Evaluate Adverse Phenotypes in Teleost Fish'. *The EMBO Journal* 35 (11): 1151–54. https://doi.org/10.15252/embj.201694448.

Bhattarai, Gehendra, Ainong Shi, Chunda Feng, Braham Dhillon, Beiquan Mou, and James

C. Correll. 2020. 'Genome Wide Association Studies in Multiple Spinach Breeding Populations Refine Downy Mildew Race 13 Resistance Genes'. *Frontiers in Plant Science* 11. https://www.frontiersin.org/articles/10.3389/fpls.2020.563187.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. 'A One-Penny Imputed Genome from Next-Generation Reference Panels'. *The American Journal of Human Genetics* 103 (3): 338–48. https://doi.org/10.1016/j.ajhg.2018.07.015.

Crespo-Piazuelo, Daniel, Lourdes Criado-Mesas, Manuel Revilla, Anna Castelló, José L. Noguera, Ana I. Fernández, Maria Ballester, and Josep M. Folch. 2020. 'Identification of Strong Candidate Genes for Backfat and Intramuscular Fatty Acid Composition in Three Crosses Based on the Iberian Pig'. *Scientific Reports* 10 (1): 13962. https://doi.org/10.1038/s41598-020-70894-2.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. 'Twelve Years of SAMtools and BCFtools'. *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Davies, Robert W., Jonathan Flint, Simon Myers, and Richard Mott. 2016. 'Rapid Genotype Imputation from Sequence without Reference Panels'. *Nature Genetics* 48 (8): 965–69. https://doi.org/10.1038/ng.3594.

Davies, Robert W., Marek Kucka, Dingwen Su, Sinan Shi, Maeve Flanagan, Christopher M. Cunniff, Yingguang Frank Chan, and Simon Myers. 2021. 'Rapid Genotype Imputation from Sequence with Reference Panels'. *Nature Genetics* 53 (7): 1104–11. https://doi.org/10.1038/s41588-021-00877-0.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. 'Accurate, Scalable and Integrative Haplotype Estimation'. *Nature Communications* 10 (1): 5436. https://doi.org/10.1038/s41467-019-13225-y.

Emrani, Hossein, A. A. Masoudi, R. Vaez Torshizi, and A. Ehsani. 2020. 'Genome-Wide Association Study of Shank Length and Diameter at Different Developmental Stages in Chicken F2 Resource Population'. *Animal Genetics* 51 (5): 722–30. https://doi.org/10.1111/age.12981.

Emrani, Hossein, Rasoul Vaez Torshizi, Ali Akbar Masoudi, and Alireza Ehsani. 2017. 'Identification of New Loci for Body Weight Traits in F2 Chicken Population Using Genome-Wide Association Study'. *Livestock Science* 206 (December): 125–31. https://doi.org/10.1016/j.livsci.2017.10.016.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. 'The Nf-Core Framework for Community-Curated Bioinformatics Pipelines'. *Nature Biotechnology* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Falker-Gieske, Clemens, Iulia Blaj, Siegfried Preuß, Jörn Bennewitz, Georg Thaller, and Jens Tetens. 2019. 'GWAS for Meat and Carcass Traits Using Imputed Sequence Level Genotypes in Pooled F2-Designs in Pigs'. *G3 Genes|Genomes|Genetics* 9 (9): 2823–34. https://doi.org/10.1534/g3.119.400452.

Fitzgerald, Tomas, Ian Brettell, Adrien Leger, Nadeshda Wolf, Natalja Kusminski, Jack Monahan, Carl Barton, et al. 2022. 'The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel'. *Genome Biology* 23 (1). https://doi.org/10.1101/2021.05.17.444412.

Fuchsberger, Christian, Gonçalo R. Abecasis, and David A. Hinds. 2015. 'Minimac2: Faster Genotype Imputation'. *Bioinformatics* 31 (5): 782–84. https://doi.org/10.1093/bioinformatics/btu704.

Grodzicker, T., J. Williams, P. Sharp, and J. Sambrook. 1974. 'Physical Mapping of Temperature-Sensitive Mutations of Adenoviruses'. *Cold Spring Harbor Symposia on Quantitative Biology* 39 (January): 439–46.

https://doi.org/10.1101/SQB.1974.039.01.056.
Hanssen, Friederike, Maxime U. Garcia, Lasse Folkersen, Anders Sune Pedersen, Francesco Lescai, Susanne Jodoin, Edmund Miller, et al. 2023. 'Scalable and Efficient DNA Sequencing Analysis on Different Compute Infrastructures Aiding Variant Discovery'. bioRxiv. https://doi.org/10.1101/2023.07.19.549462.
Hennig, Bianca P, Lars Velten, Ines Racke, Chelsea Szu Tu, Matthias Thoms, Vladimir Rybin, Hüseyin Besir, Kim Remans, and Lars M Steinmetz. 2018. 'Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol'. *G3 Genes|Genomes|Genetics* 8 (1): 79–89. https://doi.org/10.1534/g3.117.300257.
Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. 'A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies'. *PLOS Genetics* 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.
Kasahara, Masahiro, Kiyoshi Naruse, Shin Sasaki, Yoichiro Nakatani, Wei Qu, Budrul Ahsan, Tomoyuki Yamada, et al. 2007. 'The Medaka Draft Genome and Insights into Vertebrate Genome Evolution'. *Nature* 447 (7145): 714–19. https://doi.org/10.1038/nature05846.
Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. 'Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows'. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by Fernando Loizides and Birgit Scmidt, 87–90. Netherlands: IOS Press. https://eprints.soton.ac.uk/403913/.
Lander, Eric S. 1996. 'The New Genomics: Global Views of Biology'. *Science* 274 (5287): 536–39. https://doi.org/10.1126/science.274.5287.536.
Li, Na, and Matthew Stephens. 2003. 'Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data'. *Genetics* 165 (4): 2213–33. https://doi.org/10.1093/genetics/165.4.2213.
Li, Shangmin, Xingguo Wang, Liang Qu, Taocun Dou, Meng Ma, Manman Shen, Jun Guo, Yuping Hu, and KeHua Wang. 2017. 'Genome-Wide Association Studies for Small Intestine Length in an F2 Population of Chickens'. *Italian Journal of Animal Science* 17 (2): 294–300. https://doi.org/10.1080/1828051X.2017.1368419.
Li, Yudong, Xin Liu, Xue Bai, Yuxiang Wang, Li Leng, Hui Zhang, Yumao Li, et al. 2022. 'Genetic Parameters Estimation and Genome-Wide Association Studies for Internal Organ Traits in an F2 Chicken Population'. *Journal of Animal Breeding and Genetics* 139 (4): 434–46. https://doi.org/10.1111/jbg.12674.
Litt, M., and J. A. Luty. 1989. 'A Hypervariable Microsatellite Revealed by in Vitro Amplification of a Dinucleotide Repeat within the Cardiac Muscle Actin Gene'. *American Journal of Human Genetics* 44 (3): 397–401.
Liu, Huatao, Hailiang Song, Yifan Jiang, Yao Jiang, Fengxia Zhang, Yibing Liu, Yong Shi, Xiangdong Ding, and Chuduan Wang. 2021. 'A Single-Step Genome Wide Association Study on Body Size Traits Using Imputation-Based Whole-Genome Sequence Data in Yorkshire Pigs'. *Frontiers in Genetics* 12. https://www.frontiersin.org/articles/10.3389/fgene.2021.629049.
Lutz, Vanessa, Patrick Stratz, Siegfried Preuß, Jens Tetens, Michael A. Grashorn, Werner Bessei, and Jörn Bennewitz. 2017. 'A Genome-Wide Association Study in a Large F2-Cross of Laying Hens Reveals Novel Genomic Regions Associated with Feather Pecking and Aggressive Pecking Behavior'. *Genetics Selection Evolution* 49 (1). https://doi.org/10.1186/s12711-017-0287-4.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. 'A Reference Panel of 64,976 Haplotypes for Genotype Imputation'. *Nature Genetics* 48 (10): 1279–83. https://doi.org/10.1038/ng.3643.

Mendel, Gregor. 1866. 'Versuche Über Plflanzenhybriden'. *Verhandlungen Des Naturforschenden Vereines in Brünn* Bd. IV für das Jahr 1865: 3–47.

Morgan, T. H. 1910. 'Sex Limited Inheritance in Drosophila'. *Science* 32 (812): 120–22. https://doi.org/10.1126/science.32.812.120.

Pang, Yunlong, Chunxia Liu, Danfeng Wang, Paul St. Amand, Amy Bernardo, Wenhui Li, Fang He, et al. 2020. 'High-Resolution Genome-Wide Association Study Identifies Genomic Regions and Candidate Genes for Important Agronomic Traits in Wheat'. *Molecular Plant* 13 (9): 1311–27. https://doi.org/10.1016/j.molp.2020.07.008.

Parker, Clarissa C., Riyan Cheng, Greta Sokoloff, Jackie E. Lim, Andrew D. Skol, Mark Abney, and Abraham A. Palmer. 2011. 'Fine-Mapping Alleles for Body Weight in LG/J × SM/J F2 and F34 Advanced Intercross Lines'. *Mammalian Genome* 22 (9): 563–71. https://doi.org/10.1007/s00335-011-9349-z.

Pedersen, Brent S., and Aaron R. Quinlan. 2017. 'Mosdepth: Quick Coverage Calculation for Genomes and Exomes'. Edited by John Hancock. *Bioinformatics* 34 (5): 867–68. https://doi.org/10.1093/bioinformatics/btx699.

Picelli, Simone, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. 2014. 'Full-Length RNA-Seq from Single Cells Using Smart-Seq2'. *Nature Protocols* 9 (1): 171–81. https://doi.org/10.1038/nprot.2014.006.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. 'Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples'. bioRxiv. https://doi.org/10.1101/201178.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rubinacci, Simone, Diogo M. Ribeiro, Robin J. Hofmeister, and Olivier Delaneau. 2021. 'Efficient Phasing and Imputation of Low-Coverage Sequencing Data Using Large Reference Panels'. *Nature Genetics* 53 (1): 120–26. https://doi.org/10.1038/s41588-020-00756-0.

Schmid, M., M. Maushammer, S. Preuß, and J. Bennewitz. 2018. 'Mapping QTL for Production Traits in Segregating Piétrain Pig Populations Using Genome-Wide Association Study Results of F2 Crosses'. *Animal Genetics* 49 (4): 317–20. https://doi.org/10.1111/age.12663.

Tautz, Diethard. 1989. 'Hypervariability of Simple Sequences as a General Source for Polymorphic DNA Markers'. *Nucleic Acids Research* 17 (16): 6463–71. https://doi.org/10.1093/nar/17.16.6463.

Taylor, Megan, Carl-Erik Tornqvist, Xiongwei Zhao, Paul Grabowski, Rebecca Doerge, Jianxin Ma, Jeffrey Volenec, et al. 2018. 'Genome-Wide Association Study in Pseudo-F2 Populations of Switchgrass Identifies Genetic Loci Affecting Heading and Anthesis Dates'. *Frontiers in Plant Science* 9. https://www.frontiersin.org/articles/10.3389/fpls.2018.01250.

The 1000 Genomes Project Consortium. 2015. 'A Global Reference for Human Genetic Variation'. *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Tommaso, Paolo Di, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. 'Nextflow Enables Reproducible Computational Workflows'. *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Van der Auwera, Geraldine A., and Brian D. O'Connor. 2020. *Genomics in the Cloud*. O'Reilly Media, Inc. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/.

Vasimuddin, Md., Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. 'Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems'. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–24. https://doi.org/10.1109/IPDPS.2019.00041.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Wittbrodt, Joachim, Akihiro Shima, and Manfred Schartl. 2002. 'Medaka — a Model Organism from the Far East'. *Nature Reviews Genetics* 3 (1): 53–64. https://doi.org/10.1038/nrg704.

Yao, Emily J., Richard K. Babbs, Julia C. Kelliher, Kimberly P. Luttik, Kristyn N. Borrelli, M. Imad Damaj, Megan K. Mulligan, and Camron D. Bryant. 2021. 'Systems Genetic Analysis of Binge-like Eating in a C57BL/6J x DBA/2J-F2 Cross'. *Genes, Brain and Behavior* 20 (6): e12751. https://doi.org/10.1111/gbb.12751.

Supplementary figures

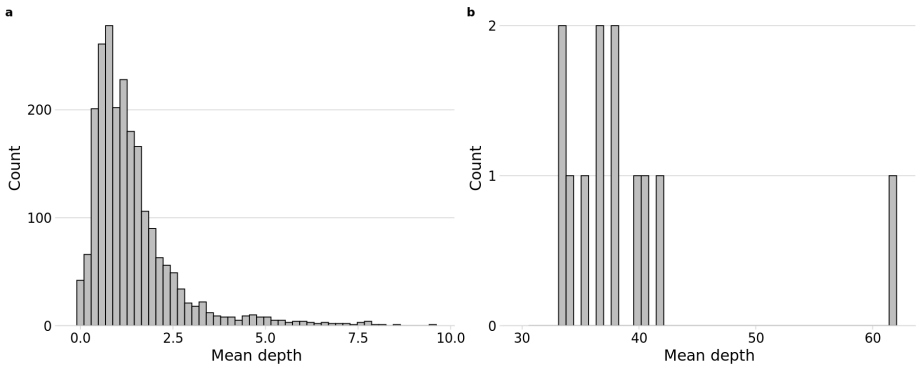

Figure S1. Distribution of mean sequencing depths in the original dataset used in this study. (a) Low-coverage samples. (b) Ground truth samples.

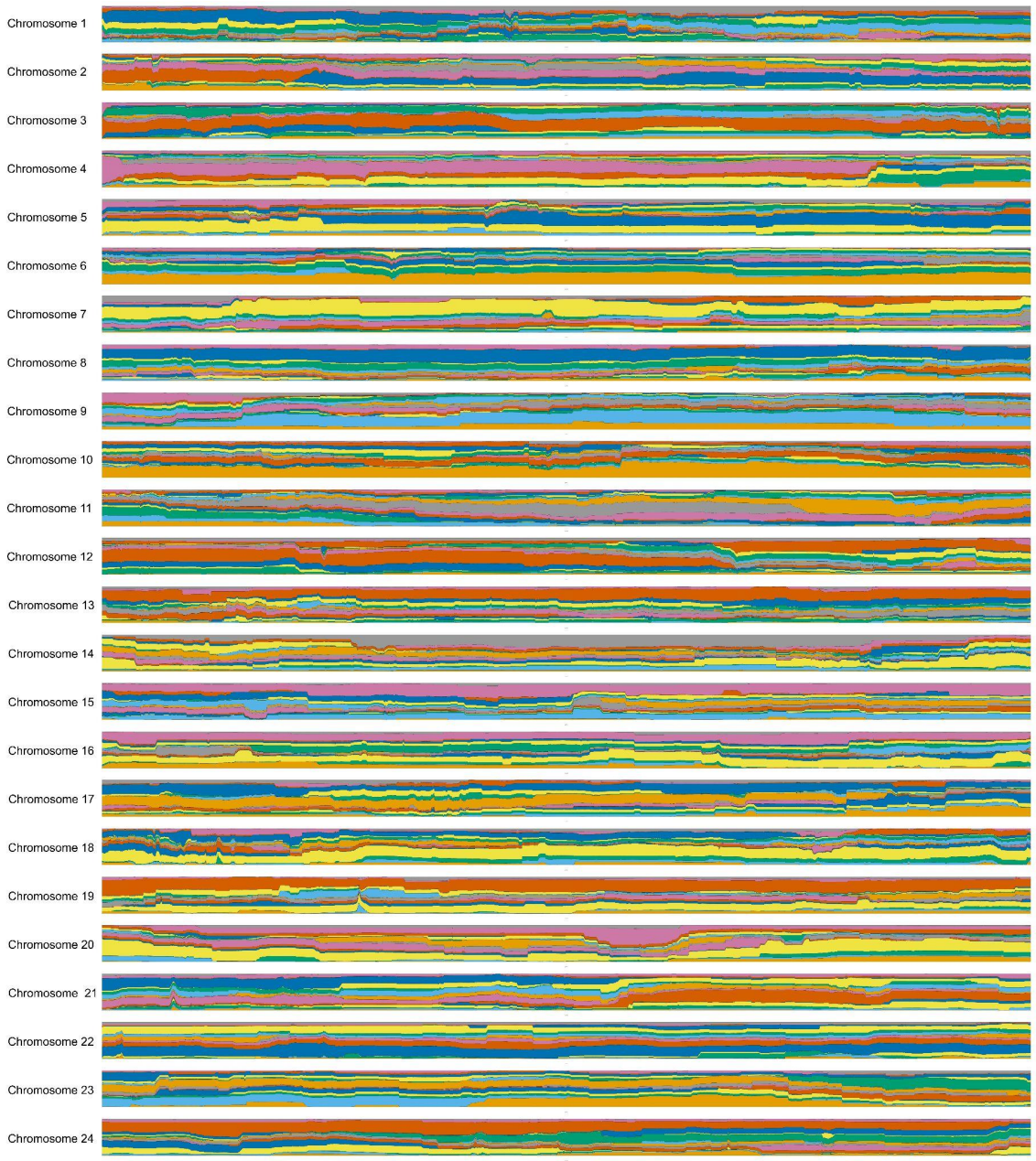

Figure S2. STITCH haplotype usage along the different chromosomes for the imputation run with the truth set samples downsampled to 0.5x mean depth and all the other samples left at the original depth. This visualisation has been produced by joining the per-chromosome plots automatically produced by STITCH. Long segments of contiguous colour reflect a lower number of ancestral haplotype switches, and hence a better imputation. Approximately equal usage of all the haplotypes also reflects more optimised heuristics.