High performance imputation of structural and single nucleotide variants in Atlantic salmon using low-coverage whole genome sequencing

Manu Kumar Gundappa [1] *, Diego Robledo [1], Alastair Hamilton [2], Ross D. Houston [3, 1 ‡], James G. D. Prendergast [1], Daniel J. Macqueen [1]

[1] *The Roslin Institute, University of Edinburgh, Easter Bush Campus, EH25 9RG, United Kingdom*

[2] *Hendrix Genetics, Villa 'de Körver', Boxmeer, 695831 CK Spoorstraat, The Netherlands*

[3] *Benchmark Genetics, Pioneer House, Edinburgh Technopole, Penicuik EH26 0BB, United Kingdom*

* Corresponding author

[‡] Former affiliation

Running title: Imputation with low-coverage WGS in salmon

Abstract

Whole genome sequencing (WGS), despite its advantages, is yet to replace alternative methods for genotyping single nucleotide variants (SNVs). Structural variants (SVs) have larger effects on traits than SNVs, but are more challenging to accurately genotype. Using low-coverage WGS with genotype imputation offers a cost-effective strategy to achieve genome-wide variant coverage, but is yet to be tested for SVs. Here, we investigate combined SNV and SV imputation with low-coverage WGS data in Atlantic salmon (*Salmo salar*). As the reference panel, we used genotypes for high-confidence SVs and SNVs for n=445 wild individuals sampled from diverse populations. We also generated 15x WGS data (n=20 samples) for a commercial population out-with the reference panel, and called SVs and SNVs with gold-standard approaches. An imputation method (GLIMPSE) was tested at WGS depths of 1x, 2x, 3x and 4x for samples within and out-with the reference panel. SNVs were imputed with high accuracy and recall across all WGS depths, including for samples out-with the reference panel. For SVs, we compared imputation based purely on linkage disequilibrium (LD) with SNVs, to that supplemented with SV genotype likelihoods (GLs) from low-coverage WGS. Including SV GLs increased imputation accuracy, but as a trade-off with recall, requiring 3-4x coverage for best performance. Combining strategies allowed us to capture 84% of the reference panel deletions with 87% accuracy at 1x WGS. This study highlights the promise of reference panel imputation using low-coverage WGS, including novel opportunities to enhance the resolution of genome-wide association studies by capturing SVs.

Key Words: Low-coverage whole genome sequencing, Imputation, Structural variants, Single nucleotide variants, Atlantic salmon, Aquaculture

Introduction

The use of whole genome sequencing (WGS) is increasing due to reducing costs and accessibility of sequencing technologies. Nonetheless, WGS, performed at the depths required to accurately detect genetic variants, remains more expensive than strategies that genotype a reduced set of variants, typically single nucleotide variants (SNV); including arrays (Balagué-Dobón *et al.*, 2022) and reduced representation sequencing (Davey *et al.*, 2011). Consequently, WGS remains less commonly used for large-scale genotyping applications like genome-wide association studies (GWAS). WGS has major advantages though, in theory capturing all SNVs and indels in each sample, enhancing the chance of identifying causative variants for a target phenotype, rather than relying on linkage disequilibrium (LD) of nearby markers (Uffelmann *et al.*, 2021). WGS also enhances scope to directly identify and genotype other classes of variation, such as larger structural variants (SVs) (Mahmoud *et al.*, 2019), though the accuracy of such calls is generally substantially lower than for SNVs when using the most common short-read sequencing approaches.

Several frameworks exist for statistical imputation of genotypes in individuals with limited genotyping information, exploiting data from individuals that have been densely genotyped (Li *et al.*, 2009; Marchini and Howie, 2010;

43 Torkamaneh and Belzile, 2021). Imputation can improve statistical power in GWAS
44 and increase resolution to fine map causative variants in quantitative trait loci (QTL)
45 regions (Li *et al.*, 2009; Marchini and Howie, 2010; Torkamaneh and Belzile, 2021).
46 Using closely related individuals with known pedigrees, densely typed genotypes in
47 a few individuals (e.g. parents) can be accurately imputed to related individuals
48 (e.g. offspring) according to shared haplotypes (Li *et al.* 2009). It is also possible to
49 apply imputation using broader reference panels of genetic variation that capture
50 different populations and haplotypes within a species, as done routinely in humans
51 (e.g. McCarthy *et al.*, 2016). Within this framework, methods have been developed
52 for imputation against a reference panel using low coverage WGS data (Pasaniuc *et*
53 *al.*, 2012; Nicod *et al.*, 2016; Spiliopoulou *et al.*, 2017; Rubinacci *et al.*, 2021; Davies
54 *et al.*, 2021). Imputation from low coverage sequencing data has the advantage of
55 reduced sequencing costs, while sufficient alleles are still covered to effectively
56 reconstruct haplotypes in the individual. In recent studies using the most advanced
57 algorithms in humans, this strategy was shown to have the potential to be more
58 accurate and cost-effective at defining sample genotypes than array-based
59 imputation against a reference panel (Rubinacci *et al.*, 2021; Davies *et al.*, 2021).
60 To date, imputation against a species-level reference panel has focussed
61 fundamentally on SNVs and indels, largely ignoring SVs; barring limited work in
62 humans (Hehir-Kwa *et al.*, 2016; Torkamaneh and Belzile, 2021). While SNVs and
63 indels are markedly easier to detect and genotype with high accuracy, SVs are
64 thought to contribute more to phenotypic variation, including by influencing gene
65 expression with larger effect sizes (Chiang *et al.*, 2017; Li *et al.*, 2017; Chen *et al.*,

66 2021; Scott *et al.*, 2021) and showing higher enrichment in GWAS haplotypes
67 (Sudmant *et al.*, 2015). Therefore, imputation of SVs using low coverage WGS
68 should markedly increase the ability of GWAS to identify causative variants.

69 Atlantic salmon (*Salmo salar*) is a species of great societal importance,
70 being among the top farmed aquaculture species globally by value. Furthermore,
71 wild populations furthermore have major conservation value and cultural
72 significance (Houston and Macqueen, 2019; Houston *et al.*, 2020). To support
73 genetic studies with both farmed and wild Atlantic salmon populations, numerous
74 genomic resources have been developed (Macqueen *et al.*, 2017; Houston and
75 Macqueen, 2019). This includes reference genomes for different phylogenetic
76 lineages (Lien *et al.*, 2016; Stenløkk *et al.*, 2022; Gao *et al.*, 2022), high-density SNV
77 arrays (e.g. Houston *et al.*, 2014; Barson *et al.*, 2015), and WGS data spanning
78 diverse populations (e.g. Gao *et al.*, 2020; Bertolotti *et al.*, 2020; Stenløkk *et al.*,
79 2022). Atlantic salmon benefits from an existing panel of high-confidence SVs
80 detected using WGS across 492 individuals representing European and North
81 American populations (Bertolotti *et al.*, 2020) and a global set of SNVs from most
82 of the same individuals (Sinclair-Waters *et al.*, 2022). These past efforts provide a
83 broad representation of both SNVs and SVs to leverage as a reference panel for
84 imputation studies.

85 Here, our primary aim was to investigate the potential for joint SV and SNV
86 imputation in Atlantic salmon, using low-coverage WGS imputed against
87 haplotypes phased from an existing, high-quality reference panel. To achieve this
88 aim, we utilized GLIMPSE (Rubinacci *et al.*, 2021) - an imputation method applied

89 to date to impute SNVs using low-coverage WGS in humans. Our results
90 demonstrate excellent performance for imputation of both SNVs and deletion SVs
91 in Atlantic salmon even at very low WGS depths, and scope to accurately capture a
92 broader spectrum of SVs in future genetics studies. This highlights the potential of
93 such approaches to enhance the resolution of downstream GWAS, and to begin to
94 better characterise the likely important role of SVs in shaping economically and
95 ecologically important phenotypes.

96 **Materials and methods**

97 **Reference panel of SNVs and SVs**

98 Our reference panel integrated SV and SNV genotypes from the same wild Atlantic
99 salmon individuals (n=445) defined in two independent studies using the
100 ICSASG_v2 genome (Lien *et al.*, 2016). These individuals represent diverse
101 European populations from a broad geographical range (Bertolotti *et al.*, 2020). For
102 SVs, this included 13,999 deletions, 1,221 duplications and 241 inversions curated
103 individually using a method shown to exclude the majority of false positive calls,
104 retaining mainly true variants (Bertolotti *et al.*, 2020). We also used 3,156,297 high
105 quality SNVs (Sinclair-Waters *et al.*, 2022). SVs and SNVs were merged into a single
106 VCF file using Picard Toolkit v2.23.8 (Broad Institute, 2019) then sorted and indexed
107 using BCFtools v1.12 (Danecek *et al.*, 2021).

108

109 **Overview of imputation strategy**

110 The key aim of this study was to test the ability of GLIMPSE (Rubinacci *et al.*, 2021)
111 to impute SNVs and SVs in Atlantic salmon. This section outlines the general
112 workflow and its rationale, and is followed by detailed methods in sections that
113 follow. Code for the full workflow is available through GitHub
114 (https://github.com/manugundappa/SV_imputation_pipeline).

115 GLIMPSE was tested against two sets of samples described in
116 Supplementary Tables 1 and 2. The first set represented samples taken from within
117 the reference panel (hereafter: 'in-panel'), comprising n=10 individuals from the
118 same river in Norway (River Enningdalselva; 58.981''N, 11.474''E) with mean and
119 standard deviation (s.d.) WGS depth = 7.8x and 1.2x, respectively (Supplementary
120 Table 1). These samples were dropped from the panel in-turn when assessing their
121 imputation performance. The second set represented samples out-with the
122 reference panel (hereafter: 'out-panel'), comprising 20 farmed individuals from the
123 Hendrix Genetics Landcatch Strain (Norwegian origin), with mean and s.d. WGS
124 depth = 15.7x and 1.1x, respectively (Supplementary Table 2). This included two
125 individuals each per 10 families previously shown to have contrasting within-family
126 genomic estimated breeding values for resistance to amoebic gill disease (Robledo
127 *et al.*, 2018). The out-panel WGS data was generated as described in a following
128 section.

129 Using the in-panel samples provided a benchmark to test the performance
130 of SNV and SV imputation, where the variants were known to directly match with
131 haplotypes phased in the reference panel. The out-panel samples were used to test
132 imputation performance for samples unrepresented within the reference panel,

133 which consists largely of wild individuals sampled from locations across Norway,
134 inclusive of rivers from which the founders of Norwegian aquaculture were derived
135 (Gjedrem and Baranski, 2010). The Hendrix-Genetics Landcatch strain we used was
136 derived from sources with strong connections to the founder populations of
137 Norwegian aquaculture, and hence was expected to also share haplotypes with the
138 reference panel. A further rationale for choosing samples from the chosen
139 aquaculture population was to demonstrate utility of the pipeline developed in
140 commercial aquaculture breeding programs.

141 The GLIMPSE workflow is validated for and by default restricted to SNVs.
142 We modified it to accommodate SVs and provide an efficient workflow to jointly
143 impute SNVs and SVs (Fig. 1). Briefly, both the in-panel and out-panel datasets were
144 down-sampled to 1x, 2x, 3x and 4x WGS depths. The reference panel was processed
145 to generate a list of all variable SNV and SV positions for which imputation should
146 be performed. Genotype likelihoods (GLs) for SNVs within the reference panel were
147 generated per down-sampled WGS dataset for input for the GLIMPSE pipeline (see
148 below). Imputation was firstly performed against down-sampled datasets across all
149 variable SNVs and SVs solely using the SNV GLs. In addition, we also developed and
150 tested an alternative approach, where the pipeline was fed with SV GLs (see below).
151 The imputed SNV and SV genotypes were later compared to reference genotypes
152 derived from the original higher coverage WGS depth data to calculate imputation
153 performance.

154

155 **Imputation of in-panel samples**

156 *GLIMPSE imputation ignoring SV GLs*

157 Samtools v1.12 (Danecek *et al.*, 2021) *view* was used to down-sample BAM files per
158 chromosome. The Samtools *coverage* command was used to retrieve mean
159 sequencing depth and calculate down-sampling factors. The Samtools *view*
160 command was used to down-sample each chromosome BAM file to the target
161 depths (1x, 2x, 3x, 4x) using sub-sampling factors generated from the data above.
162 The down-sampled BAM files were used as inputs to GLIMPSE (Fig. 1: Step 1).

163 BCFtools v1.12 was used to split the reference panel by chromosomes and
164 exclude the selected n=10 individuals from the reference panel. This was done in
165 turn such that each sample was excluded individually, with the other nine samples
166 remaining in the panel (Fig. 1: Step 2). The output of this step was stored in a BCF
167 file. Variable positions across each chromosome-specific BCF file were identified
168 using the BCFtools *view* command (Fig. 1: Step 3). The GLIMPSE pipeline restricts
169 this step to include SNVs only. However, considering our objective of imputing SVs,
170 we ignored the *–v snps* flag during this step. A proxy genetic map was generated
171 across all the samples and chromosomes (assuming 1Mb per cM) using the variable
172 sites information and a custom R script (Fig. 1: Step 4). The GLIMPSE_*chunk*
173 command was then used to define regions on which to perform imputation and
174 phasing across all chromosomes. A window size of 100Mb and buffer size of 200kb
175 was used for this step (Fig. 1: Step 5). GLs were calculated for all variable SNVs
176 within the variable sites datasets using the *mpileup* command within BCFtools (Fig.
177 1: Step 6). SV GLs were not included in this step, as the objective was to impute SVs
178 using GLs for SNVs only (i.e. on the basis of LD). Finally, all variants including SNVs

179 and SVs were imputed against the reference panel using the GLIMPSE_*phase*
180 command at the four down-sampled WGS depths (Fig. 1: Step 7). This command
181 takes in a dataset containing low depth GLs, the reference panel of all haplotypes,
182 the genetic map, and chunk information across each sample and chromosome to
183 generate a VCF file containing genotype probabilities. BCFtools *view* was used to
184 extract the original reference genotypes of all variants from the reference panel
185 previously called from the higher coverage sequence data using GATK (Sinclair-
186 Waters *et al.*, 2022) for comparison against the imputed genotypes (Fig. 1: Step 8).
187

188 *GLIMPSE imputation accuracy estimation*

189 The imputed genotype values for variants were extracted and filtered using a
190 custom R script to calculate imputation accuracies in comparison to the reference
191 genotypes (Fig. 1: Step 9). Imputed VCF files for each sample were merged across
192 all chromosomes and processed to extract posterior support values for each
193 variant. We then filtered the imputed variants at different posterior support cut-
194 off values of 0.60, 0.75 and 0.90, to establish the trade-off between imputation
195 accuracy and recall rate. We then compared the imputed genotypes against the
196 higher-coverage genotype calls in the reference panel to generate a matrix of 'true'
197 vs imputed genotypes for SNVs and different categories of SVs. The output for each
198 sample at different depths was used to generate positive predictive values (PPVs)
199 and recall rates, to gauge overall imputation performance. PPV was considered as
200 the ratio between the number of correct genotype calls divided by the sum of
201 correct and incorrect genotype calls. Recall rate was calculated as the proportion

202 of variants imputed against the total number of variants in the reference panel.
203 PPVs and recall rates across all samples and posterior cut-offs were visualised using
204 the package ggplot2 (Wickham, 2016) within R.
205
206 *GLIMPSE imputation including SV GLs*
207 We devised an approach to feed SV GLs into the GLIMPSE imputation pipeline to
208 improve SV imputation performance. SVTyper (Chiang *et al.*, 2015) was run via the
209 Smoove (Layer *et al.*, 2014) pipeline to genotype all SVs in the reference panel
210 across all downsampled BAM files for the ten samples (Fig. 1: Step 10). A custom R
211 script was used to convert SV GLs into phred-scaled genotype likelihoods (PLs)
212 rounded to the closest integer. The VCF files containing SV PLs were then merged
213 with the SNV GL VCF files generated in the GLIMPSE pipeline using Picard toolkit
214 v2.23.8 (Broad Institute, 2019) (Fig. 1: Step 11). The rest of the pipeline remained
215 the same (i.e. as described when running imputation without SV GLs), except
216 running the GLIMPSE_phase command after removing the *--impute-reference-only-*
217 *variants* flag. The final imputed genotypes were processed as described above to
218 generate PPVs and recall rates.
219
220 *Merging SVs across different imputation strategies*
221 Supplying the GLIMPSE pipeline with SV GL's increased PPV, but reduced recall rates
222 (see Results). We therefore developed an approach to merge the outputs from both
223 approaches (i.e. either including or excluding SV GLs during imputation) to test if
224 performance was optimized by taking the 'best of both' strategies. A custom R

225 script was used to overlap the VCF files obtained through both the pipelines. For
226 SVs, we tested two distinct approaches for downstream processing – defined as
227 'logical' and 'non-logical' filtering. Logical filtering takes forward SV genotypes with
228 a higher likelihood. Non-logical filtering takes forward SV genotypes from GLIMPSE
229 imputation that included SV GLs, irrespective of whether the alternative approach
230 (i.e. excluding SV GLs) had a better likelihood. After applying these filters, we
231 merged the output with all SVs imputed by the approach excluding SV GLs, making
232 a comprehensive dataset of all imputed SVs from both approaches. We then
233 calculated PPVs and recall rates as described above.
234
235 **Out-panel SV and SNV detection and genotyping**
236 DNA was extracted from each sample using a standard phenol chloroform
237 method. Purity of the DNA was confirmed using Nanodrop 1000 (Thermo Scientific)
238 and the integrity was confirmed using agarose gel electrophoresis. Approximately
239 1.0µg DNA per sample was used to generate the DNA sequencing libraries.
240 Sequencing libraries were generated by Novogene using the NEBNext® DNA Library
241 Prep Kit. The DNA libraries were sequenced on a Novaseq 6000 platform by
242 Novogene UK to generate paired-end 150bp reads (information on sequencing data
243 provided in Supplementary Table 2).
244 Raw fastq data was quality checked using fastqc (Andrews, 2010) and
245 trimmed to remove adapters and low quality bases (phred<25) using Trimgalore
246 (Krueger, 2015). The trimmed reads were aligned to Atlantic salmon genome
247 version ICSAG_v2 (NCBI accession: GCA_000233375.4) using BWA (Li and Durbin,

248 2009) to generate alignment files (SAM format), which were converted into BAM
249 format using Samtools (Danecek *et al.*, 2021). SV calling was performed using the
250 pipeline described in Bertolotti et al. (2020). Briefly, the alignment BAM files were
251 processed through the SMOOVE v2.3 pipeline, which used Lumpy v 0.2.13 (Layer *et*
252 *al.*, 2014) to generate SV calls across all 20 samples (Fig. 1: Step 12). Gap regions
253 and high coverage regions in the genome were excluded from SV calling (Bertolotti
254 *et al*., 2020). A VCF file merging all SVs was manually curated using SV-plaudit
255 (Belyeu *et al.*, 2018) to generate a list of high-confidence SV calls. This was
256 overlapped with the SV dataset generated by Bertolotti et al. (2020) using Bedtools
257 *intersect*. SVs with a reciprocal overlap of 90% across both datasets were
258 considered the same variant. Histograms illustrating SVs lengths were generated
259 using ggplot2 (Wickham, 2016). SV overlap with SVs from Bertolotti et al. (2020)
260 was illustrated using upsetR (Conway *et al.*, 2017).

261 SNV calling was performed using the BAM alignments. Deepvariant (Poplin
262 *et al.*, 2018) was used to call SNVs, setting *--model_type=WGS*, with other
263 parameters as default (Step 12). The output gVCF files were split by chromosome
264 using the BCFtools *view* command. GLNEXUS (Lin *et al.*, 2018) was used (default
265 parameters) to perform joint variant calling and merge gVCFs from all samples by
266 chromosome. This was done with the config set to 'DeepVariantWGS' to produce
267 an individual merged BCF file per chromosome. BCF files from all chromosomes
268 were merged using the BCFtools *concat* command. Vcftools (Danecek *et al.*, 2011)
269 was used to filter SNVs using the additional criteria: *--min-alleles 2 --max-alleles 2 -*

270 *-maf 0.01 --max-missing 0.7 --remove-indels --minGQ 10 --minDP 4 --recode --*
271 *recode-INFO-all --minQ 30.*
272
273 **Imputation of out-panel samples**
274 SV and SNV imputation was performed for the 20 out-panel samples as described
275 above (i.e. Fig. 1), with the following modifications: the reference panel included all
276 445 individuals in Step 2; GLIMPSE phase was run across all samples to impute SNVs
277 and SVs against the reference panel (Step 7); we performed imputation of SVs in
278 the out-panel samples by generating SVTyper GLs for SVs in the reference panel
279 (Step 10) and merging them into the pipeline (Step 11); 'true' genotype calls were
280 extracted from the Deepvariant SNV calls and SVtyper genotypes derived from 15x
281 data (Steps 12-13). PPVs and recall rates were calculated and visualized as
282 described for the in-panel samples.
283
284 **Results**
285 *In-panel imputation performance for SNVs*
286 We first assessed the impact of coverage and posterior probability (PP) cut-off on
287 SNV imputation performance. Increasing sequencing coverage is expected to
288 increase the accuracy of downstream genotype calls, but at increased financial cost.
289 Likewise increasing the PP cut-off is expected to lead to more accurate calls among
290 the remaining genotypes, but at the trade-off of having less genotype calls
291 remaining that pass the filter. SNVs were found to be imputed with high accuracy
292 for the ten in-panel samples, irrespective of sequencing depth or PP cut-off. PPVs

293 ranged from PPV = 0.91 at 1x WGS depth with PP > 0.6, to PPV = 0.98 at 4x WGS
294 depth with PP > 0.90 (Fig. 2a). As expected higher PP cut-offs increased SNV
295 imputation accuracy across all WGS depths, but at the expense of lower recall rates
296 (Fig. 2a, b). Recall rates varied markedly by WGS depth and cut-off (Fig. 2b). The
297 highest recall rate (83%) was achieved at 4x WGS depth with PP > 0.60 (PPV = 0.97).
298 Recall rate increased slightly with WGS depths using PP > 0.60 (from 0.79 to 0.83 at
299 1x to 4x depth). Using PP > 0.90 resulted in recall rates from 63 to 78% at 1x to 4x
300 WGS depth. In summary, GLIMPSE showed excellent performance for SNV
301 imputation using the in-panel samples, capturing 1,998,178 SNVs (see
302 Supplementary Table 3) with 95% accuracy (i.e. PPV = 0.95) at 1x WGS depth when
303 applying the PP=0.9 cut-off.
304
305 *In-panel imputation performance for deletion SVs*
306 We next assessed the ability to impute genotypes at deletion SVs using GLIMPSE.
307 We compared two approaches. In the first, we ignored reads in the sequencing data
308 overlapping an SV, relying solely on LD between SNVs and SVs to impute the
309 deletion genotypes. In the second approach we provided GLIMPSE with prior SV
310 GLs for deletions called using SVTyper ('GLIMPSE+SVtyper' approach), therefore
311 accounting for reads observed at SV locations in the corresponding low-depth WGS
312 data. As can be seen in Fig. 3a, the GLIMPSE+SVtyper approach consistently
313 increased deletion imputation accuracy compared to relying on LD between SNVs
314 and SVs alone. This demonstrates that accounting for the low coverage reads at SV
315 positions improves the SV calls. However, applying this strategy introduces a trade-

316 off of lower recall rates that is particularly evident with very low coverage WGS (Fig.
317 3b). As GLIMPSE will only make calls for SVs with prior GLs when run using the
318 GLIMPSE+SVTyper approach, any SVs lacking prior GLs (i.e. owing to lack of
319 sequencing data) are excluded. In contrast, when run without prior SV GLs,
320 GLIMPSE will call genotypes at all SVs, increasing the corresponding recall rates. At
321 4x WGS depth, a maximum recall rate of 91.7% (PP > 0.6) was achieved for the
322 GLIMPSE+SVTyper method, approaching the recall rate using GLIMPSE without
323 supplementation of SV GLs (97.4%) (Fig. 3b), though with a higher PPV than when
324 relying on LD alone (0.83 vs 0.73).
325
326 *In-panel imputation performance for duplication and inversion SVs*
327 PPVs were lower for duplication and inversion SVs than deletions, irrespective of
328 whether GLs were supplied to GLIMPSE (Supplementary Fig. 1 and 2). For both SV
329 types, we observed higher PPVs using the GLIMPSE+SVtyper approach, with a
330 stronger effect than for deletions (Supplementary Fig. 1 and 2). As for deletions,
331 when GLIMPSE was not supplied with SV GLs, WGS depth had little impact on PPV
332 and recall rate across WGS depths and PP cut-offs for both these SV classes
333 (Supplementary Fig. 1 and 2). Compared to deletions, there was more sample-to-
334 sample variation in PPV and recall rate for both other SV types across different WGS
335 depths and PP cut-offs (Supplementary Fig. 1 and 2). As for deletions, the highest
336 PP cut-off (PP=0.9) decreased recall rate across both imputation strategies for
337 duplications and inversions (Supplementary Fig. 1 and 2).

338 When using the GLIMPSE+SVTyper approach for duplications, the highest
339 PPVs were achieved at 1x WGS depth across all tested PP cut-offs (Supplementary
340 Fig. 1). This is likely explained by the extremely low recall rate at 1x WGS depth
341 (Supplementary Fig. 1), suggesting a subset of SV duplications captured at 1x WGS
342 depth is biased to a set that imputes accurately. When using the GLIMPSE+SVTyper
343 approach, recall rates of both duplications and inversions increased with WGS
344 depth, reaching maximum values of 86.9% (PP > 0.6) and 92.1% (PP > 0.6),
345 respectively, at a mean WGS depth of 4x (Supplementary Fig. 1). The best overall
346 performance for both these SV classes was observed at 4x WGS depth with PP >
347 0.9, with 77.2% / 92.3% of the duplication / inversion SVs captured with PPVs of
348 0.63 and 0.81, respectively (Supplementary Fig. 1 and 2).
349
350 *Merging in-panel SV imputation results*
351 Running GLIMPSE using only SNV GLs, resulted in comparatively lower PPVs, but
352 higher recall rates, compared to the GLIMPSE+SVTyper approach. To exploit the
353 benefits of each strategy we assessed the performance of merging the SV calls from
354 both approaches using 'logical' 'and non-logical' filtering strategies as described in
355 the Methods. As a general rule, logical filtering led to slightly lower PPVs and slightly
356 higher recall rates for all SV types across different WGS depths and PP cut-offs (Fig.
357 4). Overall, deletions were the best performing SVs. Even at 1x WGS depth, we were
358 able to impute 11,899 (~85%) of the reference panel deletions at PPV values
359 approaching 80%, with relatively small gains in accuracy achieved at higher WGS
360 depths, but significant gains in recall rate, e.g. with an additional 1,400 deletions

361 captured at 4x coverage. PPVs of duplications and inversions remained similar to
362 results before merging, with no pronounced impact on recall rates (Fig. 4a, b).
363
364 *SVs and SNVs detected in the out-panel samples*
365 To assess the performance of imputation on a set of samples more divergent from
366 the reference panel, we called 44,779 SVs across 20 Atlantic salmon individuals
367 representing the out-panel samples (35,728 deletions, 8,532 duplications, 519
368 inversions). Manual visual curation of these variants in the SV-plaudit framework
369 (Belyeu *et al.*, 2018) using the strategy published by Bertolotti *et al*. (2020) allowed
370 us to define a high-confidence set of 6,301 SVs (6,078 deletions, 208 duplications,
371 15 inversions). We discarded 85.92% (i.e. 38,478/44,779) of the raw SV calls, a
372 similar rate to that observed previously (Bertolotti *et al*. 2020). Deletions,
373 duplications and inversions had an average length of 1,082bp (range: 100 – 413,436
374 bp, s.d. 7,930 bp), 4,273 bp (range: 142 – 30,608 bp, s.d. 4,202 bp), and 4,740 bp
375 (range: 1,152 – 16,008 bp, s.d: 4,146 bp), respectively (Fig. 5a-c). Overlapping these
376 variants against the reference panel SVs identified 3,352, 96 and 3 shared deletions,
377 duplications and inversions, respectively (Fig. 3d-f). Thus, 2,726 deletions, 112
378 duplications and 12 inversions were specific to the out-panel samples (Fig. 5d-f).
379 Allele frequencies of homozygous SVs (1/1) in out-panel samples revealed
380 no clear difference in distribution between SVs unique to the out-panel population
381 and those overlapping the reference panel (Fig. 5g, h). However, frequencies of
382 heterozygous SVs (0/1) were slightly higher at lower frequencies in the SVs private

383 to the out-panel samples (Fig. 5i, j). 4,009,327 SNVs were retained in the out-panel
384 samples after applying all the filters described in the methods.
385
386 *Out-panel sample imputation performance for SNVs*
387 SNV imputation performance for the twenty out-panel samples was similar to the
388 in-panel samples, with PPVs ranging from 0.92 to 0.98 across the tested WGS
389 depths and PP cut-offs (Fig. 6a). However, recall rate was markedly lower across all
390 tested conditions (Fig. 6b), presumably owing to differences in haplotypes between
391 the (wild) reference panel individuals and (farmed) out-panel samples.
392 Nonetheless, we were able to impute 52-70% of the SNVs in the reference panel
393 (i.e. 1.62 - 2.21 million SNVs) across the different PP cut-offs (Fig. 6b;
394 Supplementary Table 3).
395
396 *Out-panel sample imputation performance for deletion SVs*
397 Performance of deletion SV imputation was slightly better than for the in-panel
398 samples (compare Fig. 3 and 7), with PPVs approaching or exceeding 0.9 using the
399 GLIMPSE+SVtyper approach (Fig. 7a). We also observed comparative higher PPVs
400 for the out-panel samples using GLIMPSE supplemented only with SNV GLs, with
401 PPVs exceeding 0.85 achieved even at 1x WGS depth using PP > 0.9 (Fig. 7a). Both
402 results are potentially products of the increased depth of WGS data used to
403 estimate the 'true' genotypes in the out-panel samples. There was no major
404 difference in recall rates comparing the out-panel and in-panel samples (compare
405 Fig. 3 and 7). To summarise, it was possible to capture a large proportion of deletion

406 SVs in the out-panel samples with high accuracy, even using very low WGS depths
407 without supplementing GLIMPSE with SV GLs.
408
409 *Out-panel sample imputation performance for duplication SVs*
410 PPVs for imputed duplications ranged from 0.50-0.60 across all WGS depths and PP
411 cut-offs using the out-panel samples (Supplementary Fig. 3), similar to the in-panel
412 samples (see Supplementary Fig. 1). As for the in-panel samples, imputation of
413 duplication SVs performed better using the GLIMPSE+SVtyper approach
414 (Supplementary Fig. 3). Recall rates were also comparable for the out-panel
415 imputed duplication SVs (Supplementary Fig. 3).
416 Only 3 inversions were common between the out-panel samples and the
417 reference panel, though these were generally successfully imputed across
418 coverages (Supplementary Fig. 5). With that caveat noted, inversion imputation
419 was more accurate than observed for the larger inversion set available for in-panel
420 samples. We achieved PPVs for inversions ranging from 0.80-0.90 across all WGS
421 depths (Supplementary Fig. 5a). Recall rates for inversions were similar to that
422 achieved for the in-panel values (Supplementary Fig. 5b).
423
424 *Merging of out-panel sample SV imputation results*
425 Merging outputs from the two imputation strategies provides an estimate of the
426 maximal imputation performance for the out-panel samples (Fig. 8). Even at 1x
427 WGS depth PPVs reached 0.86-0.87 for deletions with recall rates of 84-85% (Fig.
428 8a, b). At 4x WGS depth, the highest PPV of 0.92 was recovered for deletions with

429 a recall rate of 0.93 using non-logical filtering (Fig. 8a, b). Imputed duplications for
430 the out-panel samples reached a much lower PPV than deletions (Fig. 8a), but were
431 consistent with the in-panel samples (Fig. 4).
432
433 **Discussion**
434 It has been argued that genotype imputation using low-coverage WGS against a
435 reference panel is poised to bring about a paradigm shift in genetics by enabling
436 genome-wide variant analyses in a highly cost-effective manner (Rubinacci et al.
437 2021). Beyond human genetics, the farmed animal community has begun to uptake
438 this strategy, for example a recent study of cattle that benchmarked different
439 imputation methods (Teng *et al.*, 2022). Our study demonstrates the successful
440 application of GLIMPSE for genome-wide variant imputation using low coverage
441 WGS in Atlantic salmon. We advance on previous work by capturing a global set of
442 SVs in the reference panel defined previously by Bertolotti et al. (2020). While
443 imputation performance was lower for SVs than SNVs, this must be framed against
444 the challenge of accurately calling SVs directly using WGS data, which suffers from
445 extremely high false positive rates, approaching 90% in Atlantic salmon, demanding
446 extensive curation and filtering to define a high-confidence set (Bertolotti *et al.*,
447 2020). In this context, the accuracy and recall of SV calls achieved using low
448 coverage WGS was impressive, for instance reaching 92% accuracy and 94% recall
449 for deletions with 4x WGS data. A caveat of our method is that only known high-
450 confidence SVs defined within the reference panel are imputed. This is well-
451 illustrated by the out-panel samples tested, where approximately half the high-

452 confidence SVs detected *de novo* were not present in the reference panel. This
453 limitation requires updating the SV reference panel to ensure all variants of interest
454 are captured in a target population (discussed further below). Interestingly, the
455 equivalent imputation performance of the Hendrix-Genetics Landcatch strain (i.e.
456 out-panel) samples to the in-panel samples indicates that these farmed fish share
457 extensive haplotypes with the wild reference panel. This in turn highlights the
458 applicability of combined genome-wide SV and SNV imputation with low-coverage
459 WGS in commercial Atlantic salmon breeding programmes.

460 An interesting observation when performing SV imputation with GLIMPSE
461 was that the supplementation of SV genotype likelihoods (i.e. GLIMPSE+SVtyper
462 approach) increased imputation accuracy compared to imputation based solely on
463 LD with SNVs in the same haplotypes. By combining the unified SV calls from both
464 strategies, we achieved the best combination of accuracy and recall, at the cost of
465 a more complex computational workflow. However, notable gains in performance
466 using the GLIMPSE+SVtyper approach requires 3-4x WGS depths, which adds
467 further sequencing costs compared to relying purely on LD with SNVs. We also
468 observed differences in imputation of different SV classes, with performance for
469 deletions exceeding that of duplications and inversions. Duplications were the
470 worst performing SV class, with accuracy never exceeding ~60% on average.
471 Differences in imputation performance across different type of SVs could arise for
472 many reasons, including different allele frequency distributions and perhaps most
473 importantly, the comparative ability to call these variants accurately from
474 sequencing data for both the reference panels and individual low pass data. In this

475 respect, the relative performance of imputation across different SV classes matches
476 closely to differences in the accuracy of the original reference panel used for the
477 same classes of SV (Bertolotti *et al.*, 2020).

478 Our study has several obvious limitations, and the method we report can
479 be improved in the future. Firstly, our reference panel was based on mapping WGS
480 reads to the previous Atlantic salmon reference genome ICSAG_v2 (Lien *et al.*,
481 2016), which was recently superseded by higher quality assemblies for multiple
482 phylogeographic lineages (Stenløkk *et al.*, 2022; Gao *et al.*, 2022). A genome for a
483 Norwegian fish represents the current Atlantic salmon reference (Ssal_v3.1), and
484 has been annotated by NCBI (accession: GCF_905237065) and Ensembl
485 (https://www.ensembl.org/Salmo_salar/). In the future, a new reference panel can
486 be built taking advantage of these assemblies, which have much greater contiguity
487 and completeness and fewer errors than ICSAG_v2, owing to their construction
488 with long-read technologies (e.g. Stenløkk *et al.*, 2022). The Ssal_v3.1 genome is
489 being annotated using extensive functional genomic data (RNA-Seq, ChIP-Seq,
490 ATAC-Seq) generated across development and tissue types by the European AQUA-
491 FAANG consortium (https://www.aqua-faang.eu/), with the results due for release
492 through Ensembl in 2023 (personal communication). A future reference panel
493 updated to Ssal_v3.1 could cross-reference imputed variants with this rich
494 functional annotation, including regulatory elements in non-coding regions that
495 control gene expression, enhancing scope to identify candidate causal variants in
496 QTL regions.

497 Another way to improve our reference panel would be to incorporate a
498 broader definition of genetic diversity for Atlantic salmon. As our reference panel
499 is comprised solely of European (mainly Norwegian) Atlantic salmon populations, it
500 should be updated in the future to capture other lineages. Atlantic salmon is split
501 into two major phylogeographic lineages that separated more than 1 Myr ago with
502 limited subsequent gene flow (Rougemont and Bernatchez, 2018), leading to major
503 genomic differences and an argument for two 'sub-species' (Brenna-Hansen *et al.*,
504 2012; Gao *et al.*, 2022). Wild North American salmon have been used as separate
505 founders to aquaculture strains (Gao *et al.*, 2020) and are of substantial interest in
506 terms of molecular ecology and applied conservation genetics. Due to the
507 significant evolutionary distance involved, the performance of our reference panel
508 with low-coverage WGS data from North American salmon will likely be
509 substantially reduced. As WGS data is available for North American salmon (Gao *et*
510 *al.*, 2020), future updates to the reference panel should capture this, and other
511 available genetic variation representing the major Atlantic salmon lineages.
512 Another limitation of our reference panel is that SVs were defined solely by
513 short-read sequencing (Bertolotti *et al.*, 2020). The current state of the art for SV
514 discovery is long-read sequencing, which has higher baseline accuracy, and enables
515 the discovery of more types of SVs compared to short-read data (Mahmoud *et al.*,
516 2019). Moreover, the method used previously by Bertolotti et al. (2020) for SV
517 discovery excluded insertion variants. Consequently, our current reference panel is
518 likely far from comprehensive of the true catalogue of SVs in Atlantic salmon
519 genomes. There are several ways to enhance representation of SVs in a future

520 Atlantic salmon reference panel. The lack of insertions could be addressed using an
521 SV calling algorithm specifically designed to capture this class of variant (Chen *et*
522 *al.*, 2019). More importantly, long-read sequencing should also be leveraged to
523 characterise SVs more comprehensively. Given the availability of multiple high-
524 quality assemblies for different Atlantic salmon lineages built with long-read data
525 (Stenløkk *et al.*, 2022), there is an exciting opportunity to produce a graph-genome
526 (alternatively called pan-genome) representation of genetic variation in the Atlantic
527 salmon genome, which could be used for subsequent genotyping with short-read
528 WGS (Eizenga *et al.*, 2020). This strategy was performed recently in a related
529 salmonid species (whitefish; *Coregonus* spp.) to define a much larger set of SVs than
530 captured with short-read data (Mérot *et al.*, 2022). Graph/pan genome
531 representations also allow variants to be captured within regions of the genome
532 absent in a single reference genome (Sherman *et al.*, 2019). While graph genomes
533 have not yet been used alongside imputation with low-coverage WGS in the
534 published literature, this will be a valuable strategy in the future to capture a
535 comprehensive spectrum of SVs and explore their impact on phenotypic variation.
536 In conclusion, we demonstrate the promise of leveraging a reference panel
537 to call SNVs and SVs with low-coverage WGS in Atlantic salmon using the GLIMPSE
538 approach. The strategy we developed is generic and can be readily transferred to
539 other species with a high-quality set of SNVs and SVs defined in a representative
540 panel of individuals, providing opportunities to enhance the resolution of GWAS in
541 a cost-effective manner by capturing a wider spectrum of potential causal variants.
542

Data availability: The twenty WGS samples generated for the out-panel samples are available through NCBI (BioProject accession - PRJNA917857). All WGS and genetic variant data associated with the reference panel, including the in-panel WGS data, is available in public archives, as described by Bertolotti et al. (2020)) and Sinclair-Waters et al. (2022).

Acknowledgements: Prof. Sigbjørn Lien and Dr Torfinn Nome (Norwegian University of Life Sciences) shared the SNV data from Sinclair-Waters et al. 2022.

Funding: This study was funded by the Biotechnology and Biological Sciences Research Council (BBSRC), including award BB/S004181/1 (DJM), which received co-funding from the Sustainable Aquaculture Innovation Centre, and award BBS/E/D/30002275 (Institute Strategic Programme grant to the Roslin Institute).

Contributions: Designed study: MKG, JGDP, and DJM. Secured funding: RDH and DJM. Provided out-panel samples used for WGS: AH. Performed imputation analyses: MKG and JGDP. Performed detection of SVs and SNVs in out-panel samples: MKG. Supported data analysis: DR and JGDP. Prepared figures: MKG. Drafted manuscript: MKG and DJM. Provided intellectual inputs, and contributed to data interpretation, writing and finalization of manuscript: All authors.

References

565 Andrews S (2010). FastQC: a quality control tool for high throughput sequence
566 data.

567 Balagué-Dobón L, Cáceres A, González JR (2022). Fully exploiting SNP arrays: a
568 systematic review on the tools to extract underlying genomic structure.
569 Brief Bioinform 23: bbac043.

570 Barson NJ, Aykanat T, Hindar K, Baranski M, Bolstad GH, Fiske P, et al. (2015). Sex-
571 dependent dominance at a single locus maintains variation in age at
572 maturity in salmon. Nature 528: 405–408.

573 Belyeu JR, Nicholas TJ, Pedersen BS, Sasani TA, Havrilla JM, Kravitz SN, et al.
574 (2018). SV-plaudit: A cloud-based framework for manually curating
575 thousands of structural variants. GigaScience 7: giy064.

576 Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, et
577 al. (2020). The structural variation landscape in 492 Atlantic salmon
578 genomes. Nat Commun 11: 5176.

579 Brenna-Hansen S, Li J, Kent MP, Boulding EG, Dominik S, Davidson WS, et al.
580 (2012). Chromosomal differences between European and North American
581 Atlantic salmon discovered by linkage mapping and supported by
582 fluorescence in situ hybridization analysis. BMC Genomics 13: 432.

583 Broad Institute (2019). Picard Toolkit. GitHub Repository.
584 https://broadinstitute.github.io/picard/; Broad Institute

585 Chen L, Abel HJ, Das I, Larson DE, Ganel L, Kanchi KL, et al. (2021). Association of
586 structural variation with cardiometabolic traits in Finns. Am J Hum Genet
587 108: 583–596.

588 Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, et al.
589 (2019). Paragraph: a graph-based structural variant genotyper for short-
590 read sequence data. Genome Biol 20: 291.

591 Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. (2015).
592 SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat
593 Methods 12: 966–968.

594 Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. (2017). The impact of
595 structural variation on human gene expression. Nat Genet 49: 692–699.

596 Conway JR, Lex A, Gehlenborg N (2017). UpSetR: an R package for the visualization
597 of intersecting sets and their properties. Bioinformatics 33: 2938–2940.

598 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011).
599 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

600 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. (2021).
601 Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

602 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011).
603 Genome-wide genetic marker discovery and genotyping using next-
604 generation sequencing. Nat Rev Genet 12: 499–510.

605 Davies RW, Kucka M, Su D, Shi S, Flanagan M, Cunniff CM, et al. (2021). Rapid
606 genotype imputation from sequence with reference panels. Nat Genet 53:
607 1104–1111.

608 Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al.
609 (2020). Pangenome Graphs. Annu Rev Genomics Hum Genet 21: 139–162.

610 Gao G, Pietrak MR, Burr GS, Rexroad CE, Peterson BC, Palti Y (2020). A New Single
611 Nucleotide Polymorphism Database for North American Atlantic Salmon
612 Generated Through Whole Genome Resequencing. Front Genet 11.

613 Gao G, Waldbieser GC, Youngblood RC, Zhao D, Pietrak MR, Allen MS, et al.
614 (2022). The generation of the first chromosome-level de-novo genome
615 assembly and the development and validation of a 50K SNP array for
616 North American Atlantic salmon. bioRxiv: 2022.09.28.509896.

617 Gjedrem T, Baranski M (2010). Selective Breeding in Aquaculture: an Introduction.
618 Springer Science & Business Media.

619 Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ,
620 et al. (2016). A high-quality human reference panel reveals the complexity
621 and distribution of genomic structural variants. Nat Commun 7: 12989.

622 Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, et al.
623 (2020). Harnessing genomics to fast-track genetic improvement in
624 aquaculture. Nat Rev Genet 21: 389–409.

625 Houston RD, Macqueen DJ (2019). Atlantic salmon (Salmo salar L.) genetics in the
626 21st century: taking leaps forward in aquaculture and biological
627 understanding. Anim Genet 50: 3–14.

628 Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, et al. (2014).
629 Development and validation of a high density SNP genotyping array for
630 Atlantic salmon (Salmo salar). BMC Genomics 15: 90.

631 Krueger F (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to
632 consistently apply quality and adapter trimming to FastQ files.

633 Layer RM, Chiang C, Quinlan AR, Hall IM (2014). LUMPY: a probabilistic framework
634 for structural variant discovery. Genome Biol 15: R84.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. (2017). The impact of rare variation on gene expression across tissues. Nature 550: 239–243.

Li Y, Willer C, Sanna S, Abecasis G (2009). Genotype Imputation. Annu Rev Genomics Hum Genet 10: 387–406.

Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. (2016). The Atlantic salmon genome provides insights into rediploidization. Nature 533: 200–205.

Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, et al. (2018). GLnexus: joint variant calling for large cohort sequencing. bioRxiv: 343970.

Macqueen DJ, Primmer CR, Houston RD, Nowak BF, Bernatchez L, Bergseth S, et al. (2017). Functional Annotation of All Salmonid Genomes (FAASG): an international initiative supporting future salmonid research, conservation and aquaculture. BMC Genomics 18: 484.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019). Structural variant calling: the long and the short of it. Genome Biol 20: 246.

Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48: 1279–1283.

Mérot C, Stenløkk KSR, Venney C, Laporte M, Moser M, Normandeau E, et al. (2022). Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (Coregonus sp.) with long and short reads. Mol Ecol.

Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, et al. (2016). Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. Nat Genet 48: 912–918.

Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44: 631–635.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36: 983–987.

Robledo D, Matika O, Hamilton A, Houston RD (2018). Genome-Wide Association and Genomic Selection for Resistance to Amoebic Gill Disease in Atlantic Salmon. G3 Bethesda Md 8: 1195–1203.

Rougemont Q, Bernatchez L (2018). The demographic history of Atlantic salmon (Salmo salar) across its distribution range reconstructed from approximate Bayesian computations. Evolution 72: 1261–1277.

Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O (2021). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat Genet 53: 120–126.

Scott AJ, Chiang C, Hall IM (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res 31: 2249–2257.

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51: 30–35.

Sinclair-Waters M, Nome T, Wang J, Lien S, Kent MP, Sægrov H, et al. (2022). Dissecting the loci underlying maturation timing in Atlantic salmon using haplotype and multi-SNP based association methods. Heredity 129: 356–365.

Spiliopoulou A, Colombo M, Orchard P, Agakov F, McKeigue P (2017). GeneImp: Fast Imputation to Large Reference Panels Using Genotype Likelihoods from Ultralow Coverage Sequencing. Genetics 206: 91–104.

Stenløkk K, Saitou M, Rud-Johansen L, Nome T, Moser M, Árnyasi M, et al. (2022). The emergence of supergenes from inversions in Atlantic salmon. Philos Trans R Soc B Biol Sci 377: 20210195.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526: 75–81.

Teng J, Zhao C, Wang D, Chen Z, Tang H, Li J, et al. (2022). Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. J Dairy Sci 105: 3355–3366.

Torkamaneh D, Belzile F (2021). Accurate Imputation of Untyped Variants from Deep Sequencing Data. In: Shomron N (ed) Deep Sequencing Data

Analysis, Methods in Molecular Biology. Springer US: New York, NY, pp 271–281.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. (2021). Genome-wide association studies. Nat Rev Methods Primer 1: 1–21.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

709 **Figure Legends**

710

711 **Fig. 1. Work flow describing steps of imputation pipeline.** Highlighted in blue is the
712 core GLIMPSE pipeline we used to impute SNVs and SVs. Steps in green highlight
713 the details of the reference panel used for imputation, in brown the additional steps
714 used when running GLIMPSE supplemented with SV genotype likelihoods, and in
715 red are steps highlighting generation of SNV and SV genotypes for the out-panel
716 samples and downstream use in the imputation pipeline.

717

718 **Fig. 2. SNV imputation results for in-panel samples**. PPV (**a**) and recall rate (**b**) for
719 SNVs across tested WGS depths and PP cut-offs.

720

721 **Fig. 3. SV deletion imputation results for in-panel samples**. a) PPV (**a**) and recall
722 rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red
723 and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or
724 both SNP and SV GLs, respectively.

725

726 **Fig. 4. Merged SV imputation results for in-panel samples.** PPV (**a**) and recall rate
727 (**b**) for SV imputation across different SV classes, WGS depths and strategies. Red
728 plots indicate PPVs for the 'logical' and 'non-logical' SV filtering approaches (see
729 Methods).

730

731 **Fig. 5. A high-confidence set of SVs were detected in the out-panel samples,**
732 **representing twenty Atlantic salmon from a commercial population**. Size
733 distributions are shown for deletions (**a**), duplications (**b**) and inversions (**c**). Venn
734 diagrams highlighting overlap of SVs between the reference panel (blue circle) and
735 out-panel sample (orange circle) are shown for deletions (**d**), duplications (**e**) and
736 inversions (**f**). Allele frequency distributions are shown for overlapping (**g**) and non-
737 overlapping (**h**) homozygous (i.e. 1/1) SVs, as well as for overlapping (**i**) and non-
738 overlapping (**j**) heterozygous (0/1) SVs, in the out-panel samples.
739
740 **Fig. 6. SNV imputation results for out-panel samples**. PPV (**a**) and recall rate (**b**) for
741 SNVs across tested WGS depths and PP cut-offs.
742
743 **Fig. 7. SV deletion imputation results for out-panel samples**. a) PPV (**a**) and recall
744 rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red
745 and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or
746 both SNP and SV GLs, respectively.
747
748 **Fig. 8. Merged SV imputation results for out-panel samples.** PPV (**a**) and recall rate
749 (**b**) for SV imputation across different SV classes, WGS depths and strategies. Red
750 plots indicate PPVs for the 'logical' and 'non-logical' SV filtering approaches (see
751 Methods).

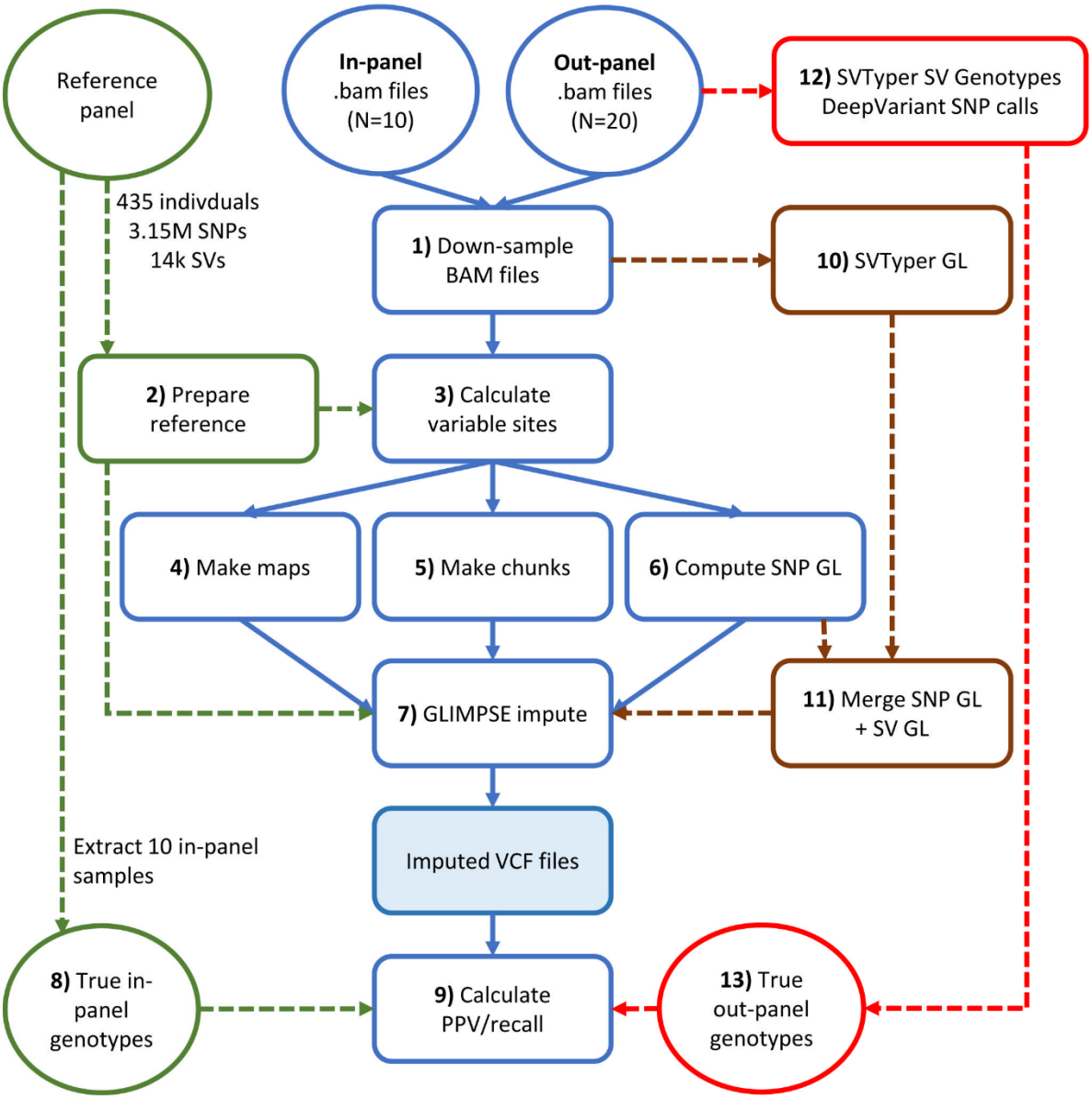

Reference panel
In-panel .bam files (N=10)
Out-panel .bam files (N=20)
12) SVTyper SV Genotypes DeepVariant SNP calls
435 indivduals
3.15M SNPs
14k SVs
1) Down-sample BAM files
10) SVTyper GL
2) Prepare reference
3) Calculate variable sites
4) Make maps
5) Make chunks
6) Compute SNP GL
7) GLIMPSE impute
11) Merge SNP GL + SV GL
Extract 10 in-panel samples
Imputed VCF files
8) True in-panel genotypes
9) Calculate PPV/recall
13) True out-panel genotypes

Figure 1

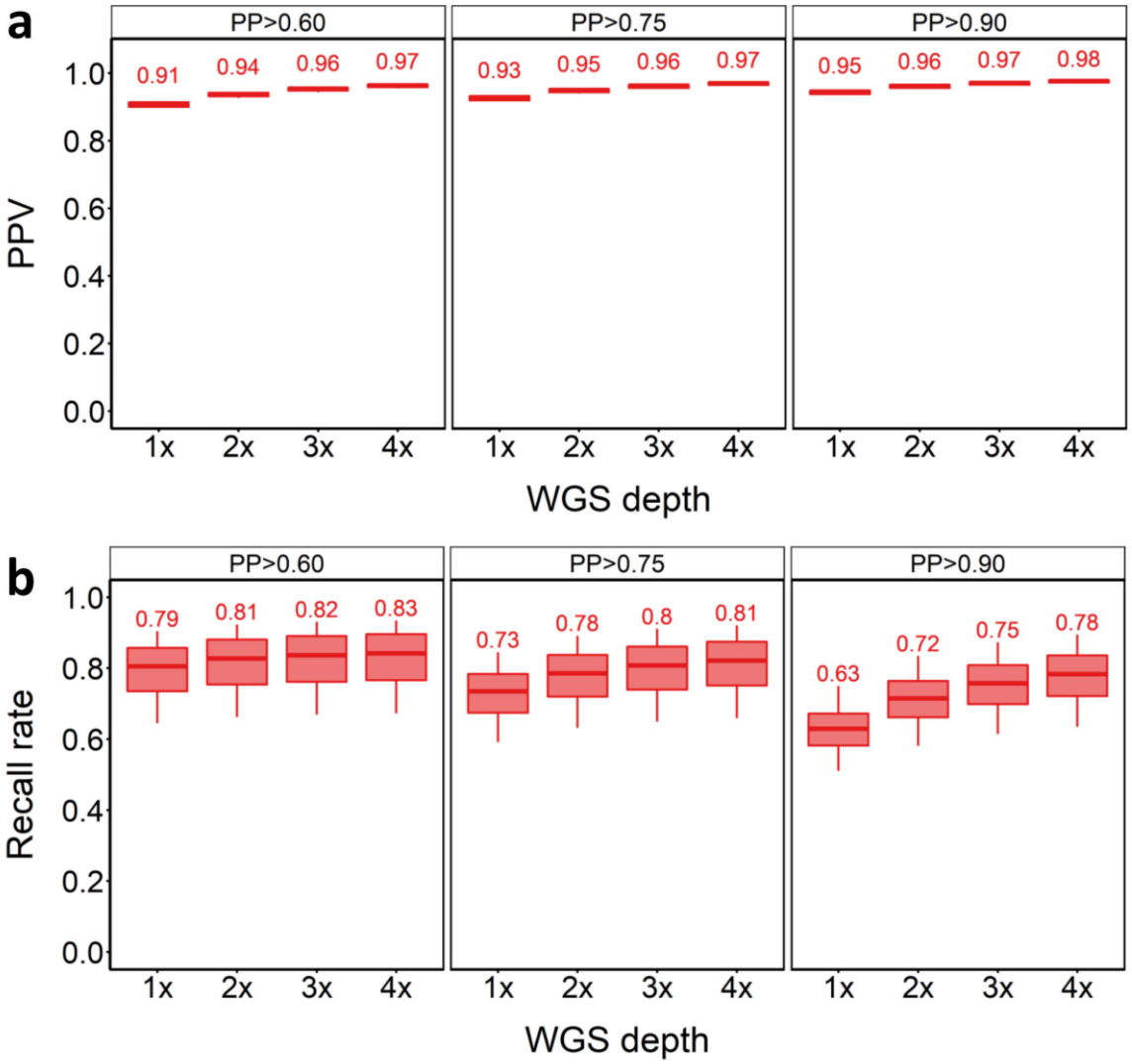

a
PP>0.60
PP>0.75
PP>0.90
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.91
0.94
0.96
0.97
0.93
0.95
0.96
0.97
0.95
0.96
0.97
0.98
1x
2x
3x
4x
WGS depth
b
PP>0.60
PP>0.75
PP>0.90
Recall rate
0.79
0.81
0.82
0.83
0.73
0.78
0.8
0.81
0.63
0.72
0.75
0.78
WGS depth

Figure 2

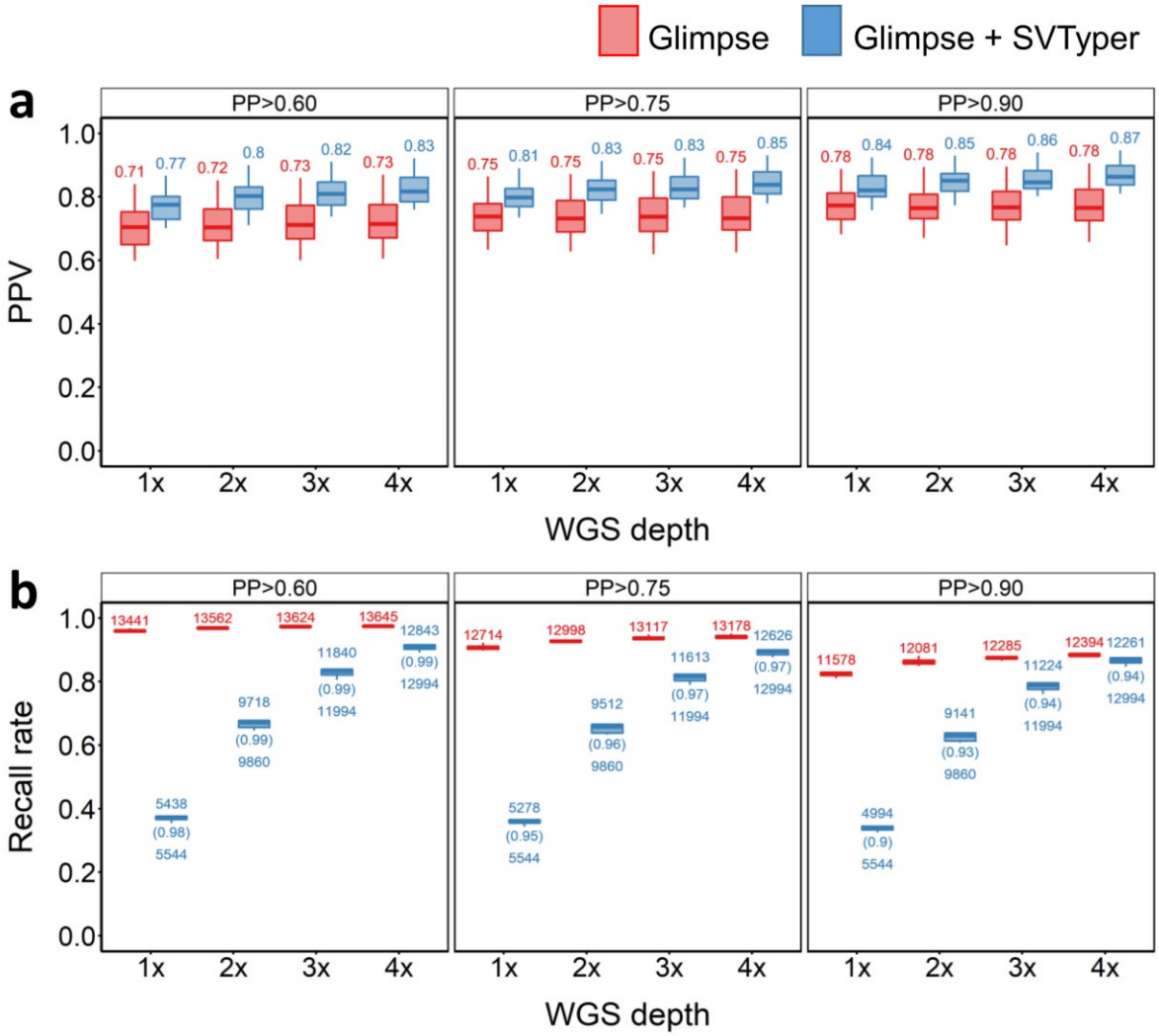

Glimpse
Glimpse + SVTyper
a
PP>0.60
PP>0.75
PP>0.90
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.71 0.77 0.72 0.8 0.73 0.82 0.73 0.83
0.75 0.81 0.75 0.83 0.75 0.83 0.75 0.85
0.78 0.84 0.78 0.85 0.78 0.86 0.78 0.87
1x 2x 3x 4x
WGS depth
b
Recall rate
13441 13562 13624 13645
5438 (0.98) 5544
9718 (0.99) 9860
11840 (0.99) 11994
12843 (0.99) 12994
12714 12998 13117 13178
5278 (0.95) 5544
9512 (0.96) 9860
11613 (0.97) 11994
12626 (0.97) 12994
11578 12081 12285 12394
4994 (0.9) 5544
9141 (0.93) 9860
11224 (0.94) 11994
12261 (0.94) 12994

Figure 3

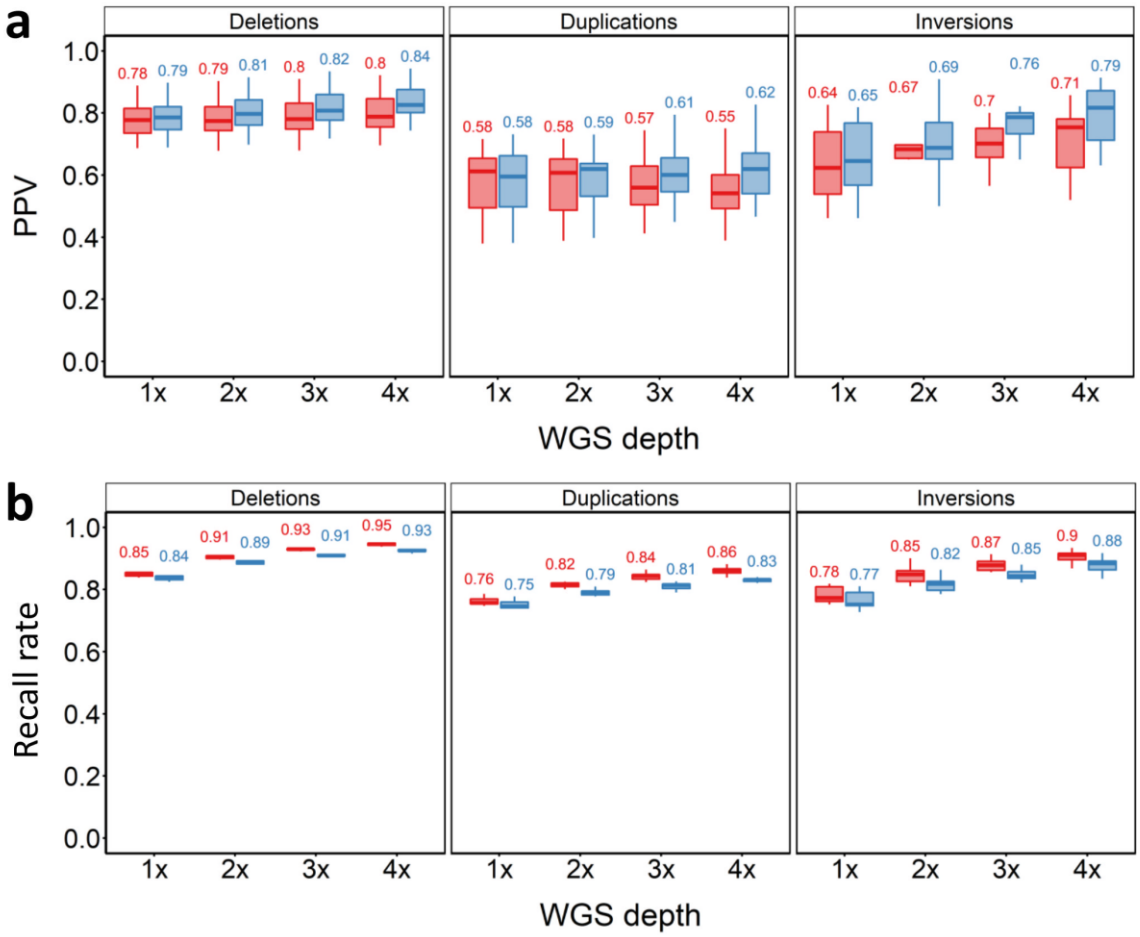

a
Deletions
Duplications
Inversions
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.78 0.79 0.79 0.81 0.8 0.82 0.8 0.84
0.58 0.58 0.58 0.59 0.57 0.61 0.55 0.62
0.64 0.65 0.67 0.69 0.7 0.76 0.71 0.79
1x 2x 3x 4x
WGS depth
b
Deletions
Duplications
Inversions
Recall rate
0.85 0.84 0.91 0.89 0.93 0.91 0.95 0.93
0.76 0.75 0.82 0.79 0.84 0.81 0.86 0.83
0.78 0.77 0.85 0.82 0.87 0.85 0.9 0.88
1x 2x 3x 4x
WGS depth

Figure 4

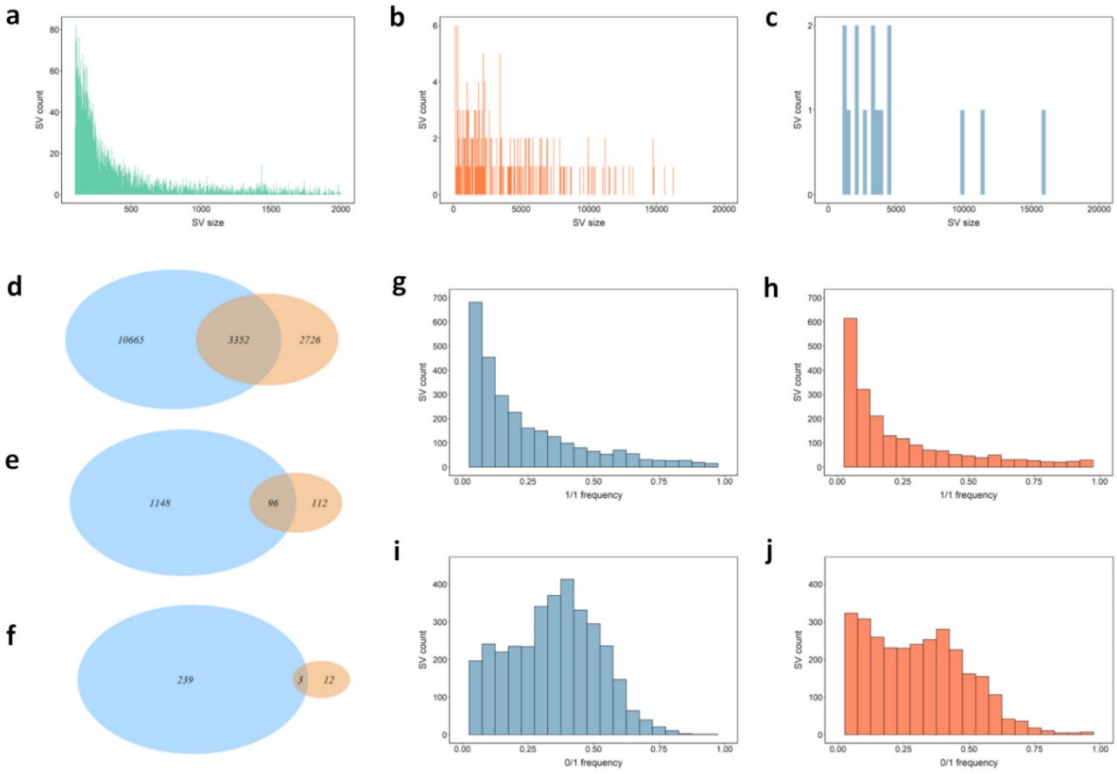

a
SV count
SV size
b
SV count
SV size
c
SV count
SV size
d
10665
3352
2726
e
1148
96
112
f
239
3
12
g
SV count
1/1 frequency
h
SV count
1/1 frequency
i
SV count
0/1 frequency
j
SV count
0/1 frequency

Figure 5

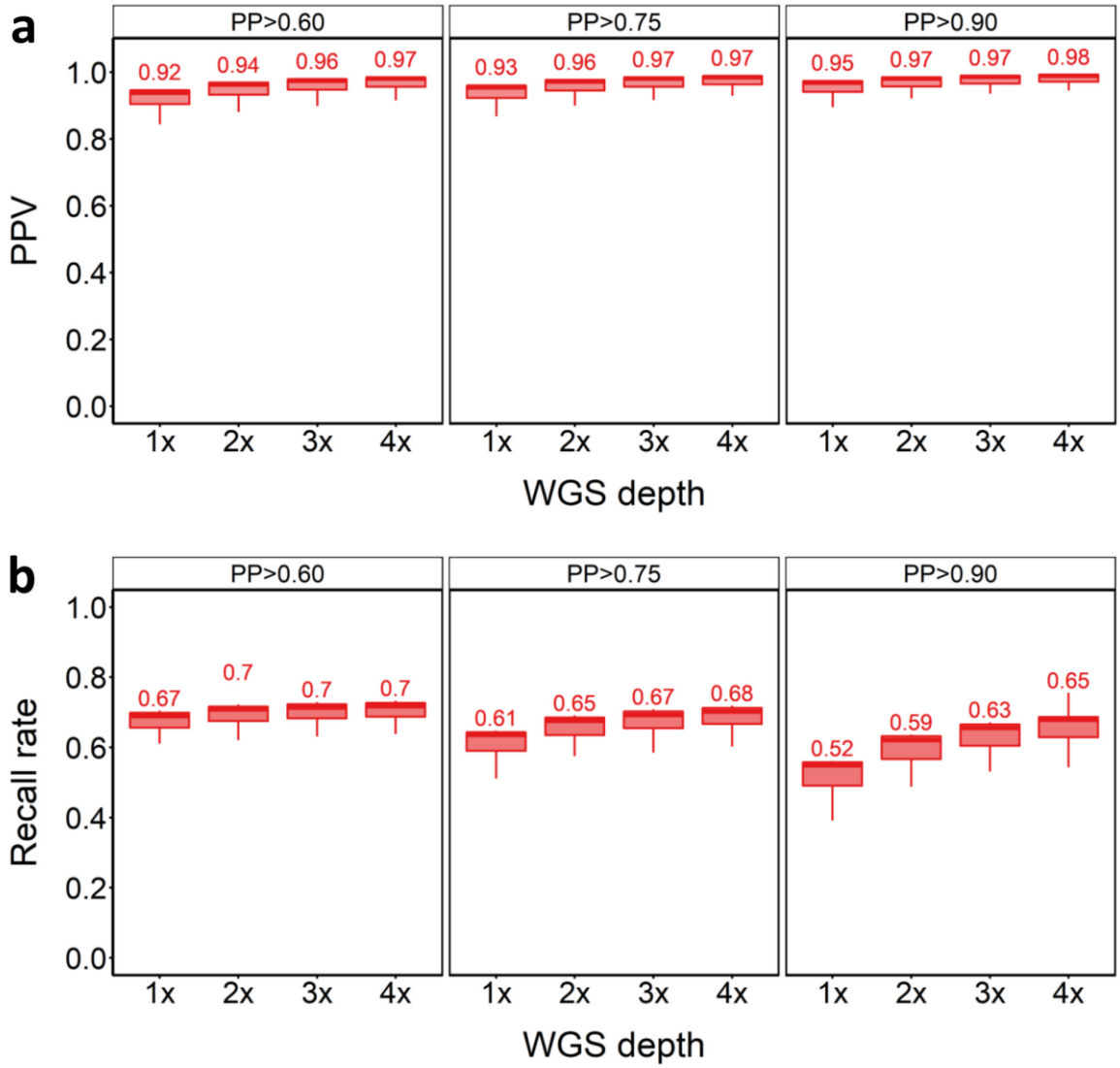

a
PP>0.60
PP>0.75
PP>0.90
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.92
0.94
0.96
0.97
0.93
0.96
0.97
0.97
0.95
0.97
0.97
0.98
1x
2x
3x
4x
WGS depth
b
PP>0.60
PP>0.75
PP>0.90
Recall rate
0.67
0.7
0.7
0.7
0.61
0.65
0.67
0.68
0.52
0.59
0.63
0.65
WGS depth

Figure 6

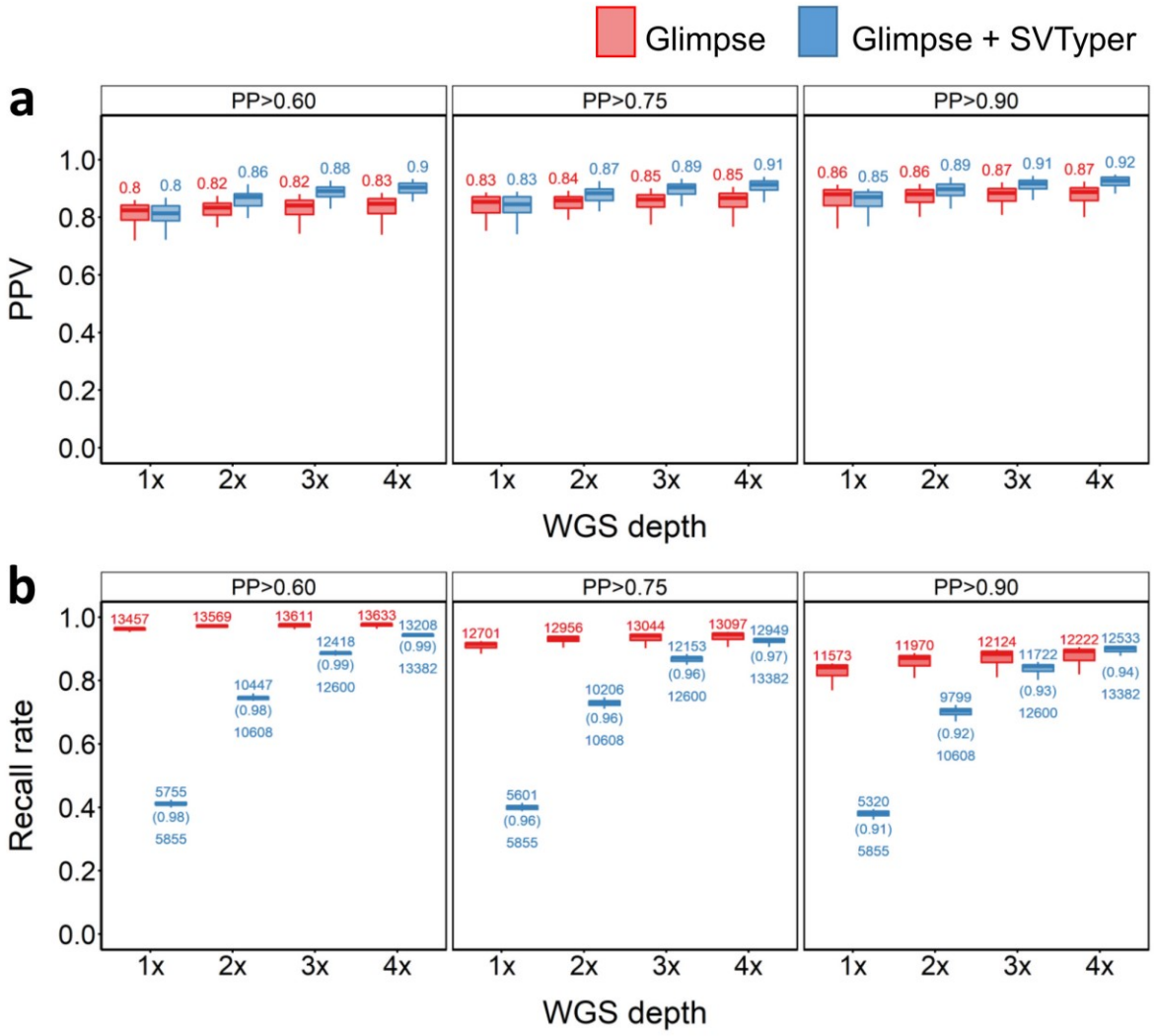

Glimpse
Glimpse + SVTyper
a
PP>0.60
PP>0.75
PP>0.90
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.8 0.8 0.82 0.86 0.82 0.88 0.83 0.9
0.83 0.83 0.84 0.87 0.85 0.89 0.85 0.91
0.86 0.85 0.86 0.89 0.87 0.91 0.87 0.92
1x 2x 3x 4x
WGS depth
b
Recall rate
13457 13569 13611 13633 13208 (0.99) 13382
10447 (0.98) 10608
12418 (0.99) 12600
5755 (0.98) 5855
12701 12956 13044 13097 12949 (0.97) 13382
12153 (0.96) 12600
10206 (0.96) 10608
5601 (0.96) 5855
11573 11970 12124 12222 12533 (0.94) 13382
11722 (0.93) 12600
9799 (0.92) 10608
5320 (0.91) 5855

Figure 7

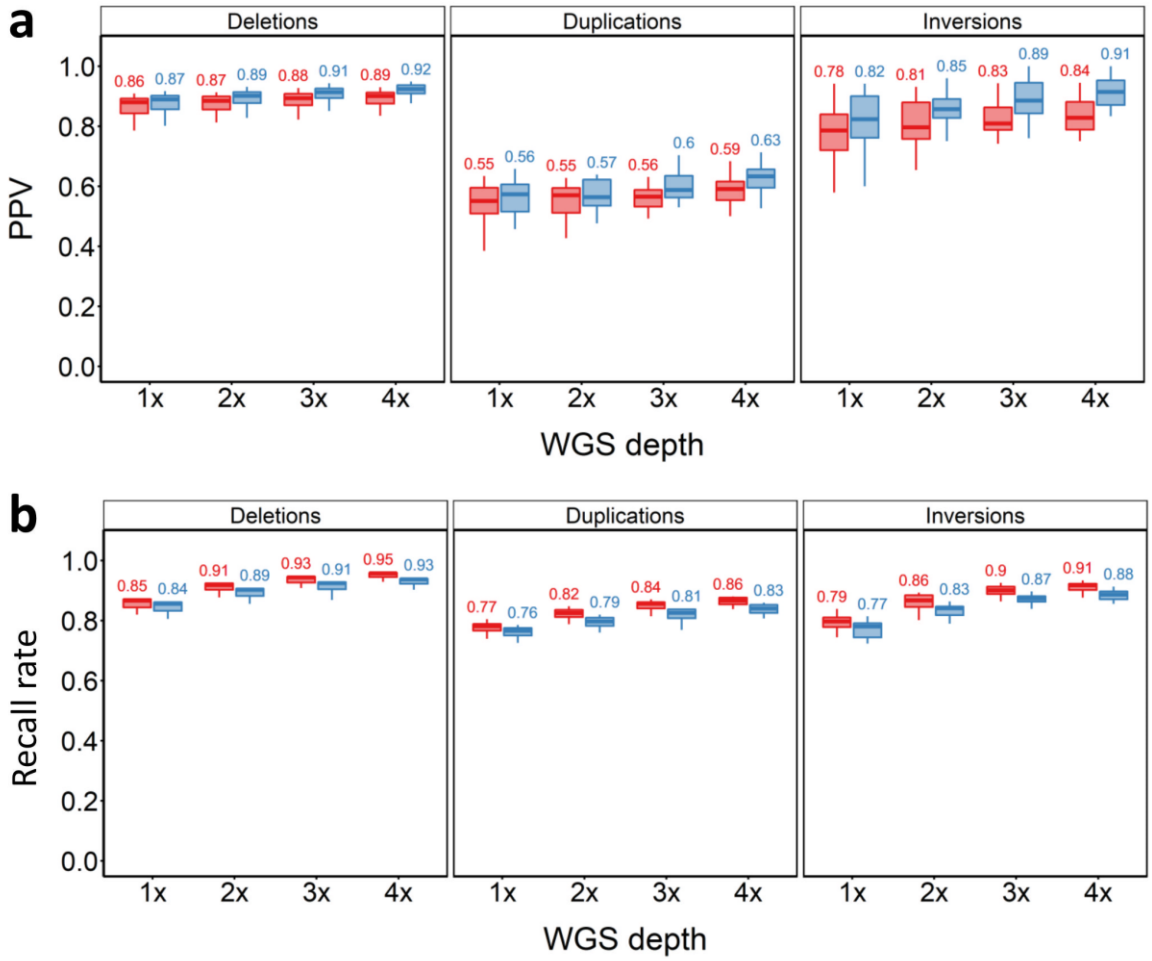
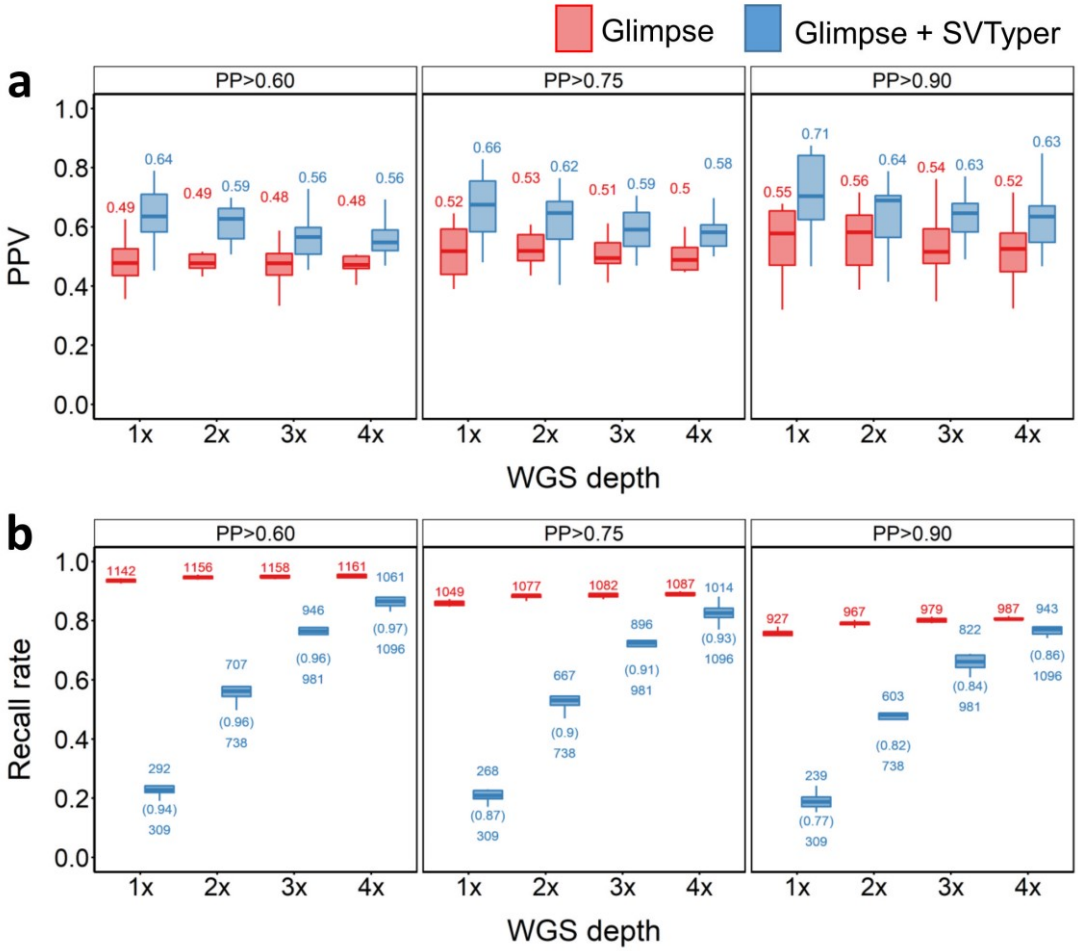
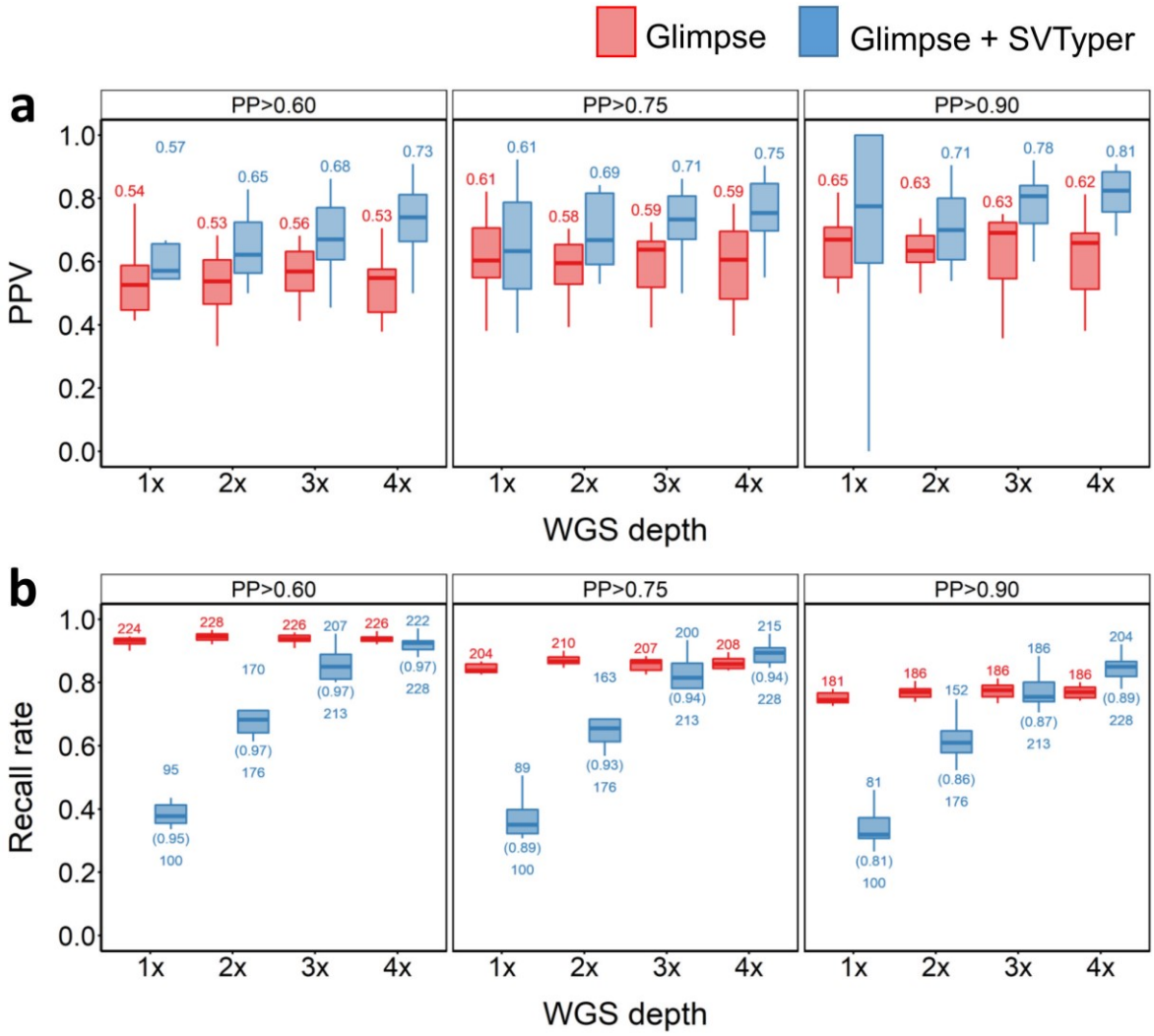

a
Deletions
Duplications
Inversions
PPV
1.0
0.8
0.6
0.4
0.2
0.0
0.86 0.87 0.87 0.89 0.88 0.91 0.89 0.92
0.55 0.56 0.55 0.57 0.56 0.6 0.59 0.63
0.78 0.82 0.81 0.85 0.83 0.89 0.84 0.91
1x 2x 3x 4x
WGS depth
b
Deletions
Duplications
Inversions
Recall rate
1.0
0.8
0.6
0.4
0.2
0.0
0.85 0.84 0.91 0.89 0.93 0.91 0.95 0.93
0.77 0.76 0.82 0.79 0.84 0.81 0.86 0.83
0.79 0.77 0.86 0.83 0.9 0.87 0.91 0.88
1x 2x 3x 4x
WGS depth

Figure 8

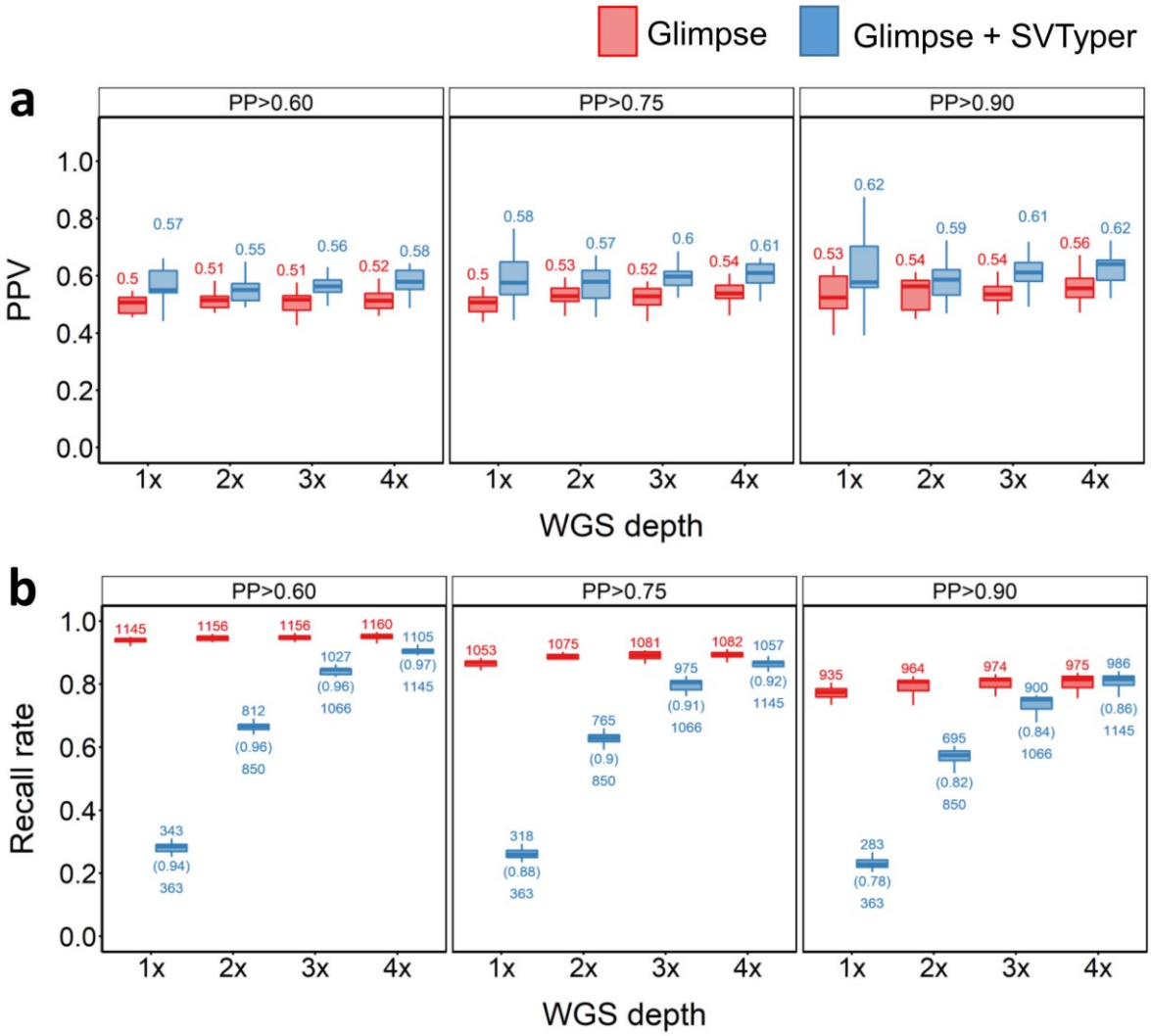
Supplementary Figures



Supplementary Figure 1. Performance of in-panel sample (n=10) duplication SV imputation. PPV (**a**) and recall rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or both SNP and SV GLs, respectively.

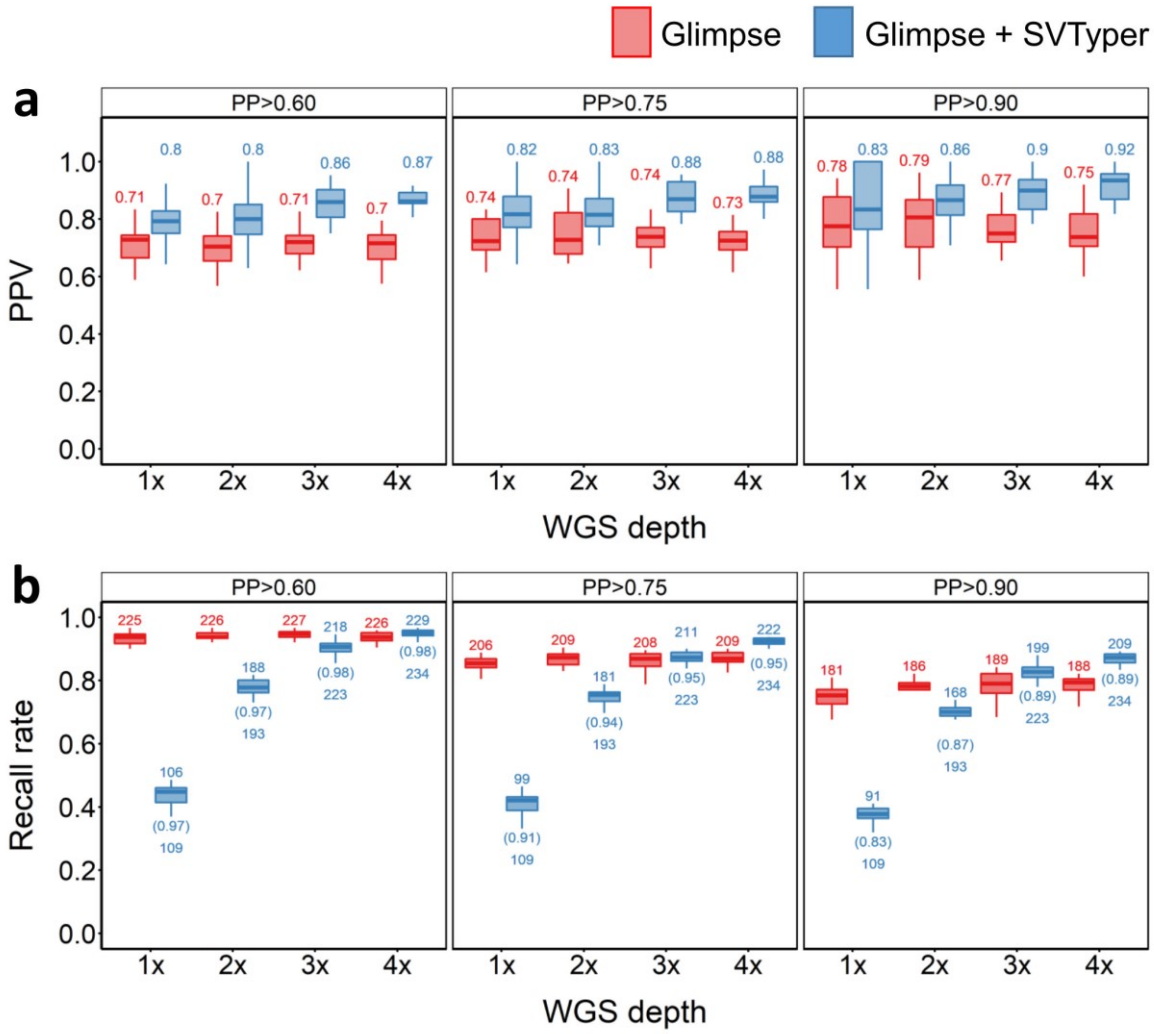


Supplementary Figure 2. Performance of in-panel sample (n=10) inversion SV imputation. PPV (**a**) and recall rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or both SNP and SV GLs, respectively.



Supplementary Figure 3. Performance of out-panel sample (n=20) duplication SV imputation. PPV (**a**) and recall rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or both SNP and SV GLs, respectively.

752



Supplementary Figure 2. Performance of out-panel sample (n=20) inversion SV imputation. PPV (**a**) and recall rate (**b**) for across different PP cut-offs, sequencing depths and strategies used. Red and blue plots indicate PPVs when the pipeline was supplied with SNP GLs only, or both SNP and SV GLs, respectively.