Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Assessment of linkage disequilibrium patterns between structural variants and single nucleotide polymorphisms in three commercial chicken populations

Johannes Geibel (johannes.geibel@uni-goettingen.de)
University of Goettingen

Nora Paulina Praefke
University of Goettingen

Steffen Weigend
Friedrich-Loeffler-Institut

Henner Simianer
University of Goettingen

Christian Reimer
University of Goettingen

Research Article

Keywords: Chickens, single nucleotide polymorphisms, structural variants, linkage disequilibrium

DOI: https://doi.org/10.21203/rs.3.rs-861830/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Assessment of linkage disequilibrium patterns between structural variants and single nucleotide polymorphisms in three commercial chicken populations

Johannes Geibel[12*], Nora Paulina Praefke[12], Steffen Weigend[23], Henner Simianer[12], Christian Reimer[12]

[1] Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, GERMANY

[2] Center for Integrated Breeding Research, University of Goettingen, GERMANY

[3] Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, GERMANY

[*] Corresponding author; johannes.geibel@uni-goettingen.de

Abstract

Background

Structural variants (SV) are causative for some prominent phenotypic traits of livestock as different comb types in chickens or color patterns in pigs. Their effects on production traits are also increasingly studied. Nevertheless, accurately calling SV remains challenging. It is therefore of interest, whether close-by single nucleotide polymorphisms (SNPs) are in strong linkage disequilibrium (LD) with SVs and can serve as markers. Literature comes to different conclusions on whether SVs are in LD to SNPs on the same level as SNPs to other SNPs. The present study aimed to generate a precise SV callset from whole-genome short-read sequencing (WGS) data for three commercial chicken populations and to evaluate LD patterns between the called SV and surrounding SNPs.

Results

The final callset consisted of 12,294,329 bivariate SNPs, 4,301 deletions (DEL), 224 duplications (DUP), 218 inversions (INV) and 117 translocation breakpoints (BND). While average LD between DELs and SNPs was at the same level as between SNPs and SNPs, LD between other SVs and SNPs was strongly reduced (DUP: 40 %, INV: 27 %, BND: 19 % of between-SNP LD). A main factor for the reduced LD was the presence of local minor allele frequency differences, which accounted for 50 % of the difference between SNP – SNP and DUP – SNP LD. This was potentially accompanied by lower genotyping accuracies for DUP, INV and BND compared with SNPs and DELs. An evaluation of the presence of tag SNPs (SNP in highest LD to the variant of interest) further revealed DELs to be slightly less tagged by WGS SNPs than WGS SNPs by other SNPs. This difference, however, was no longer present when reducing the pool of potential tag SNPs to SNPs located on four different chicken genotyping arrays.

Conclusions

The results imply that genomic variance due to DELs in the chicken populations studied can be captured by different SNP marker sets as good as variance from WGS SNPs, whereas separate SV calling might be advisable for DUP, INV, and BND effects.

Keywords

Chickens; single nucleotide polymorphisms; structural variants; linkage disequilibrium

Background

A type of genomic variation that affects large regions of the genome is caused by structural variants (SV). SVs can alter the total genome size by deleting (deletions, DEL), duplicating (duplications, DUP) or inserting (insertions, INS) longer stretches of DNA (unbalanced SV). Those SVs are often referred to as copy number variations (CNV). In contrast, inversions (INV) and translocations (TRA) do not affect the length of the genome (balanced SV) [1]. Especially unbalanced SVs are assumed to come with a strong functional impact on the phenotype, e.g. by strong deleterious effects of DELs which can remove complete genes [2] or by DUPs that increase numbers of cis-regulatory elements [2, 3]. SVs and complex combinations of multiple SVs are also known to be causative for some of the most prominent phenotypic breed characteristics of livestock breeds as walnut- and rose comb in chickens [4] or belted color patterns and dominant-white color in pigs [5].

The power for detection of SVs of certain types and sizes, however, is highly technology-dependent in various aspects [1]. During the last two decades, technologies evolved that increased the resolution and accuracy of SV detection at the submicroscopic level. Array-based comparative genomic hybridization (aCGH) allowed the detection of long CNVs >35 kb [2]. The development and increased use of microarrays led to technologies that either detect DELs from characteristics of population-level single nucleotide polymorphism (SNP) genotypes [6, 7] or utilized signal intensity information [8]. The increasing availability of short-read sequences during the last decade led to the development of multiple SV detection algorithms which use read depth distributions [1, 9] and/or information from split reads and insert size distributions of paired-end reads, potentially combined with local assembly procedures [1, 10–12]. However, short-read-based methods still come with a variety of limitations due to the short read sizes which highly vary between the algorithms [1, 13] and especially a general deficit in calling INS [14]. Therefore, current state-of-the-art methods nowadays utilize the information of PacBio or

Nanopore long-read sequencing or linked-read technologies as HI-C [15], but the availability of these types of sequencing data is still very limited for the majority of intensively researched livestock species.

Other than for SVs, the use of SNPs has become routine over the last two decades. Therefore, large whole-genome-sequencing (WGS) reference panels [16, 17] and collections of individuals, which were genotyped by microarrays and phenotyped in routine breeding programs or during large-scale research projects [18], exist. Given the complexity of SV detection, it is of interest to know which part of the effects of SVs on the phenotype is already captured by potential linkage disequilibrium (LD) between the SV of interest and nearby SNPs. Strong LD would allow for the inclusion of those effects in e.g. genomic prediction without the need for a separate SV analysis.

LD between two variants can be measured using a variety of estimators (reviewed e.g. in [19]), of which the squared correlation of haplotypes (r^2) is probably the most prominent one. It can be interpreted as the amount of information of a variant that is captured by another one. However, its upper limit is defined by the difference in minor allele frequency (ΔMAF) between the two variants [20]. The overall strength of LD is highly population depended and closely linked to the effective population size [19]. LD thereby shows a characteristic decay pattern of mean LD by distance. However, for many applications as genome-wide association studies (GWAS), the interest is more in the maximum observed LD of a causal variant to a close-by so-called tag SNP, which can capture the effect as a marker genotype.

By now, a bunch of studies has addressed the question of linkage between SVs and surrounding SNPs in humans with contrasting results. Generally, common DELs were shown to be in good LD to SNPs by most of the studies [6, 21–24], but some found this LD to be weaker than SNP – SNP LD [25, 26]. Literature additionally suggests, that rare DELs are weaker tagged (tag SNP is SNP with highest LD to the variant within a defined distance) than common DELs [22, 27] and DUP were in weaker LD to SNPs than DELs [22, 26, 28]. It was additionally shown that the availability of tag SNPs for SVs depends on the SNP panel used (WGS vs. different arrays) [22–24]. A further effect that was found is the location of the SV on the genome. Regions of segmental duplications are known to trigger recurrent SV formation by

89 non-allelic homologous recombination and therefore lead to SV hotspots [1, 29]. A closer look at those
90 regions by Locke *et al.* [30] found very few of those CNV to be tagged by surrounding SNPs.

91 Reduced LD between SNPs and SVs can have diverse reasons. A main factor is the increased possibility
92 of the occurrence of recurrent mutations in regions of low sequence complexity by non-allelic homol-
93 ogous recombination (NAHR)[29]. SVs from recurrent mutational events then show reduced LD to var-
94 iants from a unique mutational event [6, 30]. LD between SNPs and SVs may further be decreased by
95 different selectional properties of SNPs and SV [31], MAF differences between SVs and SNPs [20], or
96 ascertainment of SNPs for arrays that excludes regions of high structural complexity due to technical
97 reasons [32]. Additionally, known problems with SV calling accuracy [1] may lead to a high share of
98 false-positive SV calls and therefore on average low LD to more accurately called SNPs.

99 For livestock, results on SV – SNP LD are very rare, even though a high number of publications targets
100 SV. Based on a GWAS on 26,362 Holstein dairy cattle 50 k genotypes, Xu *et al.* [33] found a quarter of
101 CNVs that were significantly associated with milk traits not being tagged by adjacent SNPs. The same
102 was observed by Lee *et al.* [32] who investigated functional and population genetic features of CNV
103 regions in two dairy cattle breeds, also called from a 50 k SNP array. They identified a weak linkage
104 between CNV regions and SNPs, which was slightly stronger between DELs and SNPs than between
105 DUPs and SNPs. Wang *et al.* [34] included a local LD analysis around CNVs (called from SNP arrays) that
106 were significantly associated with production traits in pigs. Four out of eight significantly associated
107 CNVs overlapped haploblocks of non-significant SNPs, but only one CNV was found 300 kb downstream
108 of significantly associated SNPs. Note that this, however, may also be an artifact of a much stronger
109 correction for multiple testing in SNPs than in CNV.

110 In chickens, a variety of studies investigated CNVs on a quantitative basis. The studies either used aCGH
111 [35–39], utilized signal information of SNP arrays [40–45] via PennCNV [8], or read depth information
112 of short-read sequences [46–50]. There were only three studies that also included non-CNV SVs [46,
113 50, 51]. None of the studies analyzed the LD patterns of the variants.

Aim of the study

This is the first study that assessed SV – SNP LD in chickens to investigate the usefulness of SNP markers in capturing SV-based genomic variance. We, therefore, identified SVs from paired-end short-read sequences in three commercial chicken populations (white layers, brown layers, broilers), thoroughly described the SV callset, and assessed the strength of LD between those SVs and SNPs. We also identified major reasons for some existing differences to SNP – SNP LD and evaluated the performance of four available SNP arrays to tag SVs.

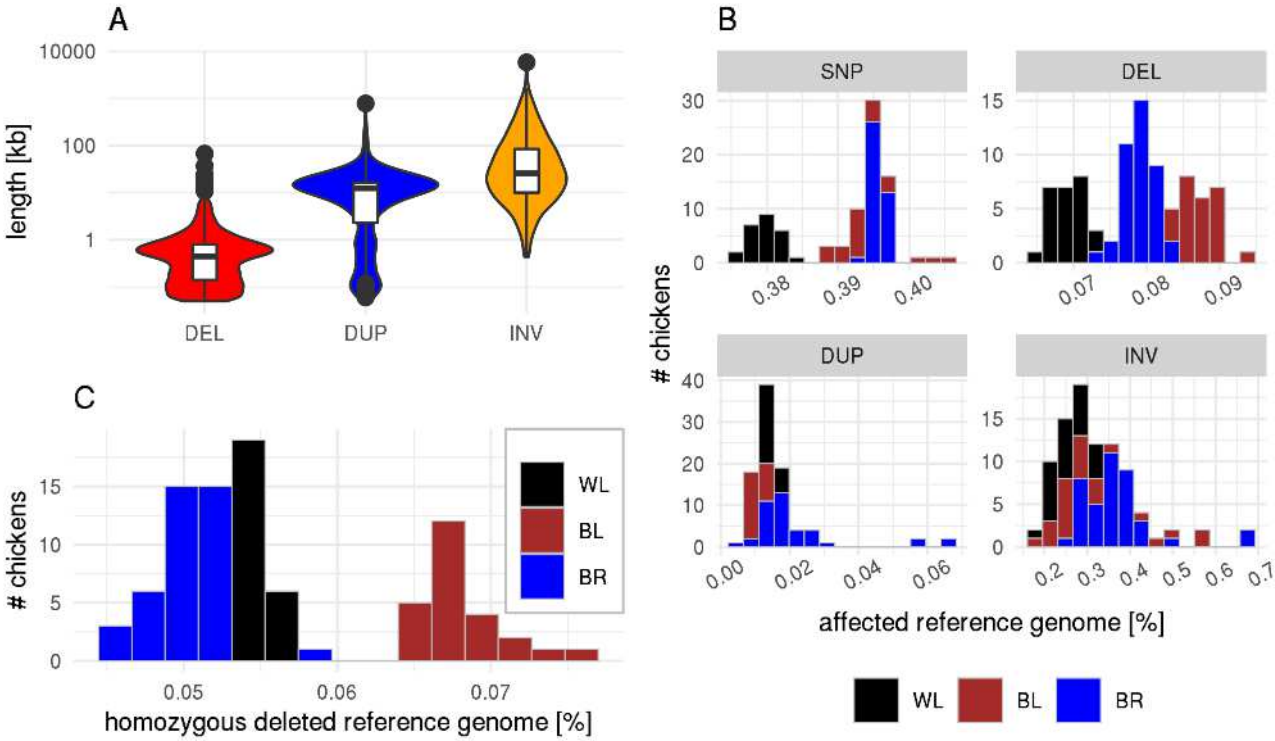
Results

Calling results and description of variants

For the study, paired-end short-read sequences of 90 chickens from three populations (25 commercial white layers, WL; 25 commercial brown layers, BL; 40 commercial broiler chickens, BR) were used. The raw data was first published by Qanbari *et al.* [52] which contains a more detailed description of the studied populations. SNP genotypes were retrieved from a previous study [53]. SVs were called by a consensus calling approach, which used three paired-end and split-read-based tools, followed by a strict filtering procedure that further utilized read-depth and SNP information. Finally, the remaining SV calls were visually checked by evaluating samplots [54] for each variant, the merged SNP and SV set was phased, and missing genotypes were imputed. The filtering procedure retained 12,294,329 bivariate SNPs, 4,301 DELs, 224 DUPs, 218 INVs, and 117 translocation breakpoints (break ends; BND) on chromosomes 1 - 33. Note that all INS were filtered out due to missing support by at least two variant callers.

Figure 1 A shows the length distribution of the called SVs. DELs were on average shortest with a median of 443 bp and a maximum of 67,037 bp. DUPs (median = 12,285 bp; maximum = 778,041 bp) were larger than DELs and INVs were largest (median = 25,643 bp; maximum = 5,795,187 bp). BNDs only indicate translocation breakpoints and, therefore, do not come with length information. The called SNPs in total accounted for 1.28 % of the autosomal reference genome length, while DELs covered

139 0.35 %, DUPs 0.39 %, and INVs 2.80 % of the chicken genome. The distributions by individuals can be
140 found in Figure 1 B. We additionally checked how much of the autosomal reference genome is homo-
141 zygously deleted in the chickens. This number varied from 0.045 % (135 kb) to 0.076 % (727 kb) with
142 BL showing a larger size of homozygously deleted reference genome than WL and BR (Figure 1 C).



143

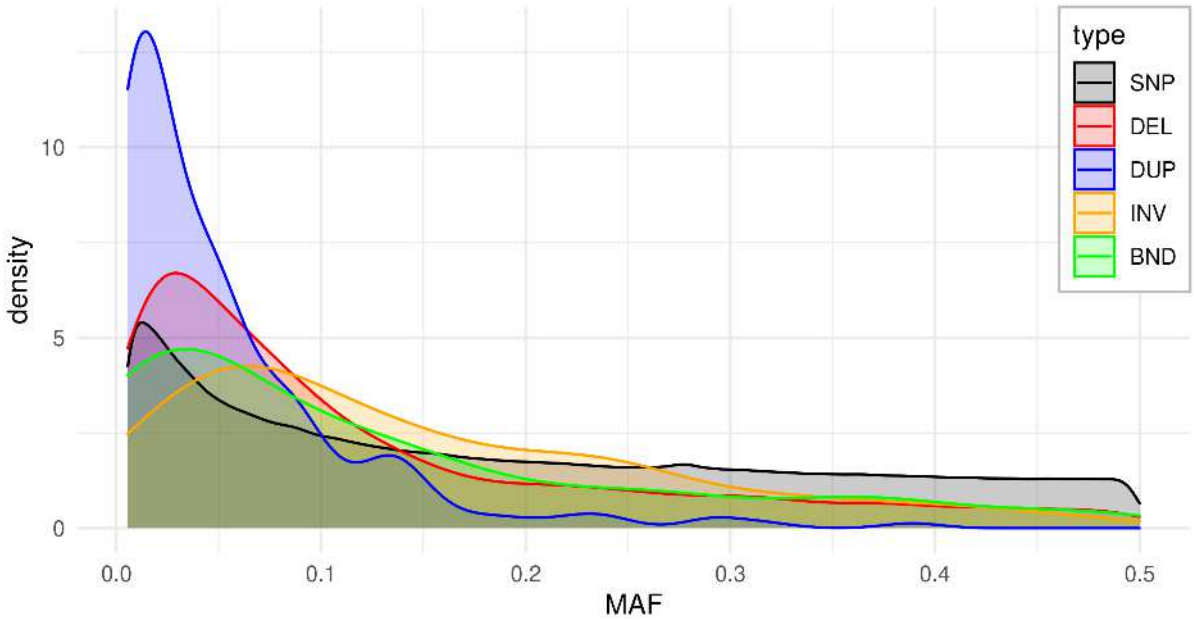
144 *Figure 1: Length distribution of SVs (A), percent of affected autosomal reference genome by individual*
145 *and variation type (B), and percent of homozygously deleted reference genome by individuals (C). The*
146 *size in B is calculated as the average between the haplotypes of an individual affected by the non-*
147 *reference allele. Note the log-scaled y-axis in A. Per-breed bars in the histograms are stacked on each*
148 *other.*

149

150 We further checked for chromosome-wise differences in the number of called variants by regressing
151 the relative number of called variants per chromosome on the relative chromosome length (Figure S
152 3). SNPs did not show any difference to the line of identity (slope = 1.00, p = 1.00), while DELs (slope =
153 1.28, p = 1.4e-4) and INVs (slope = 1.39, p = 6.1e-9) showed a significant bias towards larger chromo-
154 somes. DUPs (slope = 1.13, p = 0.34) and BNDs (slope = 1.14, p = 0.17) also showed a numerical bias
155 towards larger chromosomes, which, however, was not significant. Note that the R^2 value of the model
156 was comparably small with 0.39.

157 Distributions of minor allele frequencies (MAF; Figure 2) revealed a slight (DEL) to strong (DUP) shift
158 towards rare variants compared with SNPs for DELs and DUPs, while INVs and BNDs showed a slight
159 shift towards more common variants.

160



161

162 Figure 2: Distribution of minor allele frequency (MAF) across all samples by variant type.

163

164 Variant effect predictions of Ensembl-vep [55] classified 98.48 % of the impacts of SNPs on genes as
165 MODIFIER, 1.14 % as LOW, 0.37 as MODERATE and only 0.01 % as HIGH. DEL impacts were classified
166 only in 0.41 % of the cases other than MODIFIER (MODERATE = 0.01 %; HIGH = 0.40 %), while DUP
167 impacts were classified as HIGH in 9.95 % of the cases (MODIFIER = 90.05 %). In contrast, INV and BND
168 impacts were completely classified as MODIFIER. Further results of VEP are summarized in Figure S 4.

169

170 LD decay

171 To assess the information content of SNPs on SVs, we calculated the LD between SVs and all bivariate
172 SNPs up to 100 kb apart from the breakpoints as squared haplotype correlation (r^2). Note, that SNPs

that were located on SVs were excluded from the analysis, as their calls may be directly influenced by the SV. To get a baseline for comparisons, we also calculated the SNP – SNP LD within this distance.

Mean SNP – SNP r^2 was highest in WL (0.51 within 500 bp), followed by BL (0.41) and BR (0.26). The DEL – SNP LD decay curve follows closely the pattern of the SNP – SNP LD decay (Figure 3). Even though the level of LD was strongly reduced for the other variant types, a slight decay curve with increasing distance was still noticeable. Due to the small number of called DUPs in WL, the decay curve strongly fluctuated in this population. However, BR and BL gave some evidence that the DUP – SNP and INV – SNP decay curves were comparable, while BND – SNP decay came with a slightly lower level of LD.

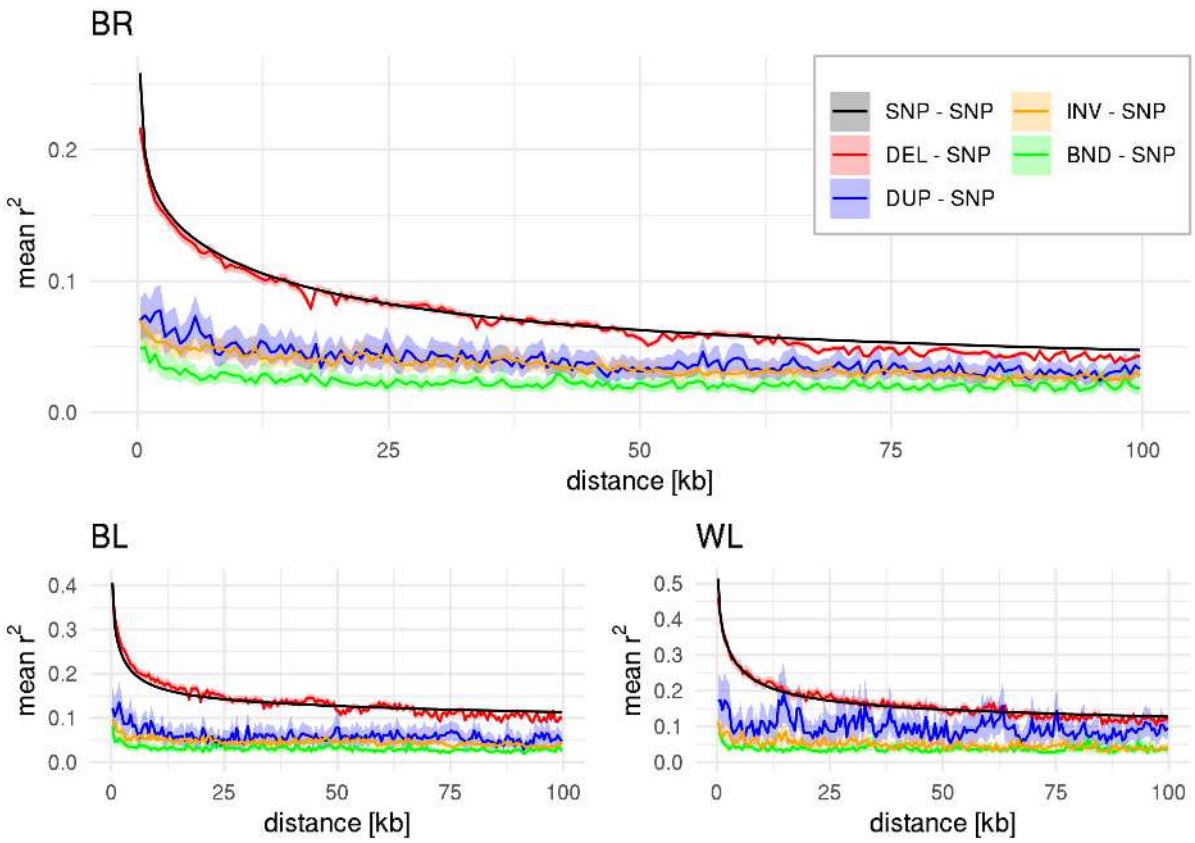


Figure 3: LD decay in the broiler (BR), brown layer (BL) and white layer (WL) chickens. The LD is presented as mean r^2 in 500 bp distance bins and the shaded areas represent Bonferroni-corrected 95 % bootstrap confidence intervals. For SNP – SNP distance bins with > 1M r^2 values, no confidence intervals were estimated.

To quantify the difference in LD between variants and populations and account for the population-specific level of LD, we expressed the mean LD in the 500 bp bins relative to the SNP – SNP LD and further averaged those values for the first 10 bins (Table 1). This revealed comparable values within variants and across populations of less than 12 % difference. Across all populations, DEL – SNP LD was on the same level as SNP – SNP LD, while DUP – SNP LD was ~40 %, INV – SNP ~27 % and BND – SNP ~19 % of SNP – SNP LD within 5 kb distance. Note that the relative r^2 was not necessarily constant across the complete range of 100 kb (Figure S 5).

Table 1: SV – SNP r^2/r_S^2 relative to the SNP – SNP r^2/r_S^2

Type	All			BR			BL			WL		
	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2**}	Δ^{***}	r^{2*}	r_S^{2*}	Δ^{***}
DEL – SNP	100.1 ± 6.1	98.8 ± 4.3	-1.3	95.4 ± 4.1	94.2 ± 2.5	-1.2	107.0 ± 3.2	103.2 ± 1.2	-3.8	98.1 ± 3.4	98.8 ± 2.5	0.7
DUP – SNP	39.9 ± 6.8	68.2 ± 8.9	28.3	39.5 ± 5.8	66.7 ± 4.3	27.2	41.1 ± 7.0	65.6 ± 9.9	24.5	39.1 ± 8.1	72.3 ± 10.5	33.2
INV – SNP	26.8 ± 5.2	46.0 ± 4.3	19.2	32.6 ± 2.4	46.8 ± 1.7	14.2	26.0 ± 2.4	50.1 ± 3.4	24.1	21.6 ± 3.1	50.1 ± 6.2	28.5
BND – SNP	18.5 ± 3.6	46.9 ± 5.4	28.4	22.4 ± 2.3	50.4 ± 3.7	28.0	18.0 ± 1.9	44.6 ± 5.1	26.6	15.3 ± 2.0	45.5 ± 5.7	30.3

* Means of the first ten 500 bp bins relative to the SNP – SNP r^2 [%] ± standard deviations [%]
** Means of the first ten 500 bp bins relative to the SNP – SNP r_S^2 [%] ± standard deviations [%]
*** Difference between relative r^2 and relative r_S^2

Effect of allele frequency

Figure 2 revealed differences in the MAF spectra of the variant types. We therefore further evaluated local MAF differences (ΔMAF) within-population by comparing ΔMAF for the SNP – SNP and SV – SNP pairs within 5 kb distance. This revealed elevated ΔMAF for DUP – SNP, INV – SNP, and BND – SNP pairs compared to SNP – SNP and DEL – SNP pairs in BL and WL (Figure S 7, Figure S 8), but not in BR (Figure S 6). As the upper bound of r^2 directly depends on ΔMAF [20], we investigated which part of the

observed differences in the LD decay curves is due to the observed allele frequency differences. For this, we used the standardized squared correlation coefficient (r_S^2), which expresses r^2 as the proportion of the maximum possible r^2 given ΔMAF of the two variants [20] and thereby excludes effects of different allele frequencies on r^2. Mean r_S^2 values (Figure S 1) were generally higher than mean r^2 values (Figure 3) due to the removal of the allele-frequency-dependent component. While the r_S^2 values of DEL – SNP relative to the SNP – SNP values (Table 1) were on a comparable level of > 94 % as the relative r^2 values (-3.8 % to +0.7 %), the relative r_S^2 values of DUPs, INVs and BNDs were between 14 % and 33 % higher than the according relative r^2 values. The relative r_S^2 values for the complete range of 100 kb are shown in Figure S 9.

Absence of homozygous SV genotypes

During the investigation of the reasons for the lower level of LD between non-DEL SVs and SNPs, we realized a strong absence of homozygous calls for DUPs, INVs, and BNDs, but not for DEL (exemplarily demonstrated for BR in Figure 4 A). To check whether this deviation is due to small variant allele frequencies, we calculated the deviation to Hardy-Weinberg-Equilibrium (HWE) and tested those for significance, using a Haldane Exact test under usage of the R package HardyWeinberg 1.7.2 [56] (exemplarily shown for BR in Figure 4 B). Homozygous DEL calls deviated into positive as well as into negative direction from the HWE. Homozygous calls for the other SV classes instead nearly exclusively deviated into a negative direction for all populations and only negative deviations were significant.

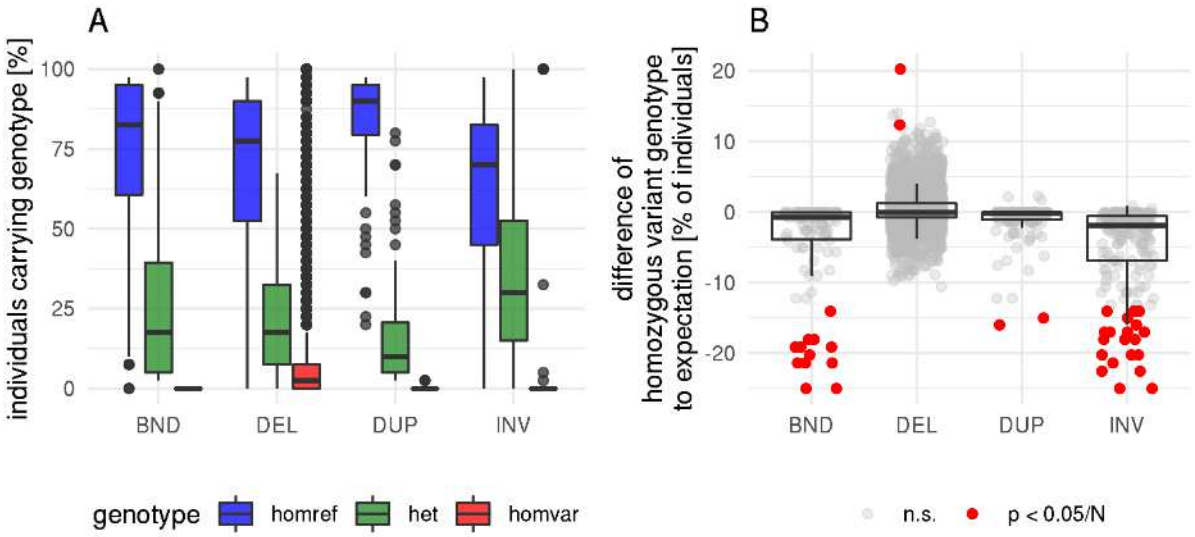


Figure 4: Percentage of individuals carrying SV genotype (A) and deviations of homozygous variant genotypes from the Hardy-Weinberg-Expectation (B) in the broiler population for each called SV. Deviations from HWE were tested by a Haldane Exact test under usage of the R package HardyWeinberg 1.7.2 [56]. Bonferroni correction of the p values was applied within SV class. Homref – homozygous for the reference allele; het – heterozygous; homvar – homozygous for the variant allele; n.s. – not significant. Comparable figures for WL and BL can be found in Figure S 11 and Figure S 12.

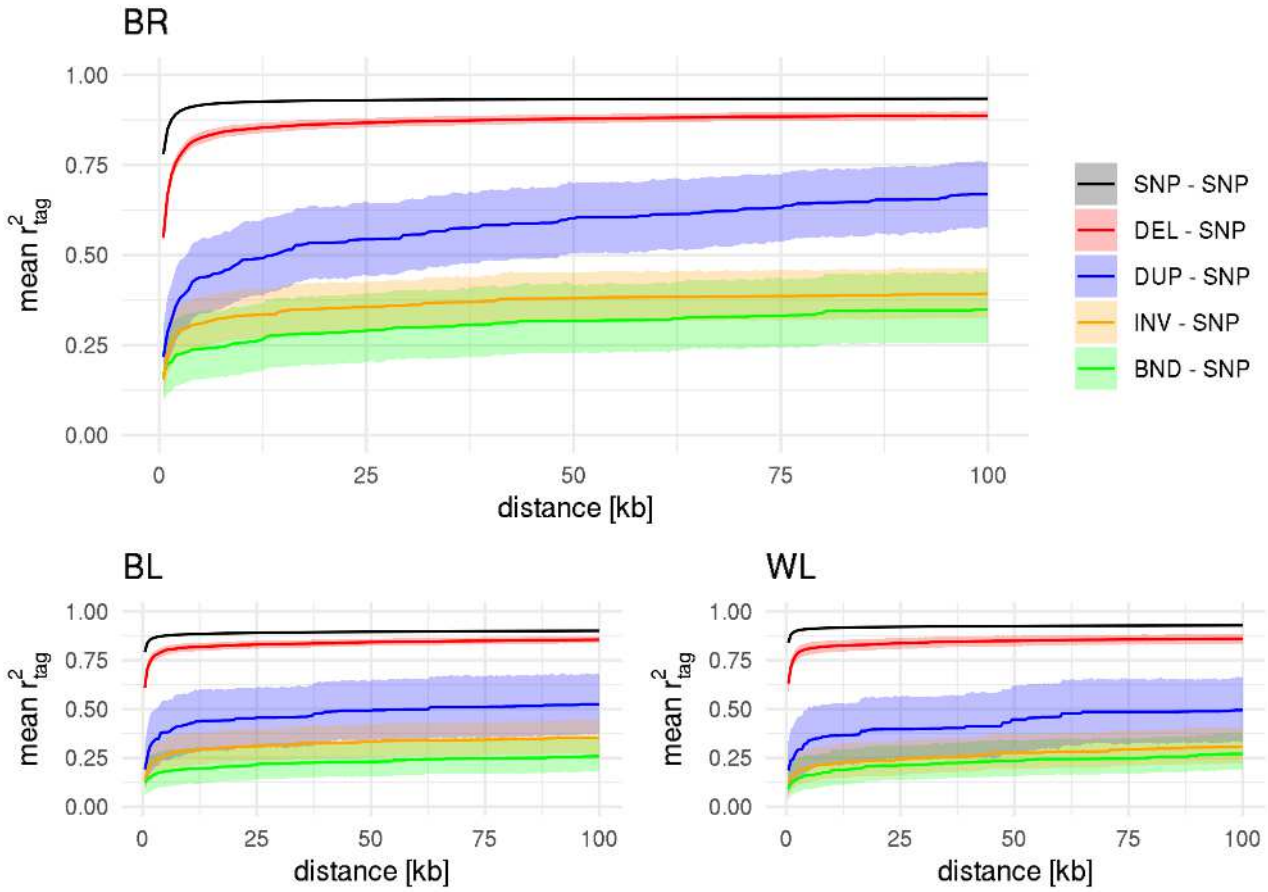
We tried to tackle the effect of this problem by correlating the 0/1/2 coded SNP genotypes with a coverage-dependent measure of copy number for DELs and DUPs, the Duphold Flanking Fold Change (DHFFC) [57]. However, as the DHFFC was also used for filtering, the results of this are potentially confounded and are only part of the supplementary material (Supplementary File 1).

Taggability

Theoretically, one SNP in strong LD to the variant of interest would be enough to serve as a marker that (partly) captures the effect of the variant for, e.g., GWAS or genomic selection as tag SNP. We, therefore, investigated the presence of potential tag SNPs close to the variants of interest. The used measure is the maximum observed r^2 between a variant of interest and a pool of potential tag SNPs within a certain distance (r^2_{tag}). Nearly all variants in all variant classes came with at least one variable SNP within proximity of 10 kb (Figure S 14). Mean r^2_{tag} for all variants and populations showed an

245 asymptotic trend with identifying the best tag SNP within 10 kb for most of the variants in all three
246 populations (Figure 5). Only mean r^2_{tag} of DUPs in BR was continuously growing until 100 kb distance
247 (Figure 5). Mean r^2_{tag} for SNPs only reached ~0.9 within 100 kb in all three populations, meaning that
248 some SNPs were not in full phase to any other SNP. Mean r^2_{tag} was slightly reduced for DELs and
249 strongly for DUPs, INVs and BNDs compared to SNPs (Figure 5).

250



251

252 *Figure 5: Mean taggability for broiler (BR), brown layer (BL), and white layer (WL) chickens. Taggability*
253 *(r^2_{tag}) was calculated as the maximum r^2 value up to a certain distance from the variant of interest.*
254 *Means across variants are presented as lines while the shaded area represents the Bonferroni-corrected*
255 *95 % bootstrap confidence intervals.*

256

257 We additionally defined a variant as tagged if r^2_{tag} > 0.75 and evaluated shares of accordingly tagged
258 variants. While more than 85 % of the SNPs were tagged in BR within 10 kb, this number was slightly

259 smaller for DELs (>75 %). More than 25 % of the DUPs were tagged within 10 kb distance and 50 %
260 within 100 kb, while less than 15 % of INVs and BNDs were tagged. The tendency is the same in the
261 two layer populations, but the absolute numbers slightly deviate. As a maximum value of a sample is
262 not independent of the number of sampled values, we also checked the number of present potential
263 tag SNPs within 5 kb distance to the variant of interest. Interestingly, SNPs were surrounded by signif-
264 icantly more close variable SNPs on average than SVs in all three populations (Table 2). This difference
265 was still present when regarding only tag SNPs ($r^2 > 0.75$).

266

267 *Table 2: Median number of variable SNPs within 5 kb distance to variants of interest*

Variant	BR		BL		WL	
	All	$r^2 \geq 0.75$	all	$r^2 \geq 0.75$	all	$r^2 \geq 0.75$
SNP	140 [a]	7 [a]	85 [a]	5 [a]	73 [a]	9 [a]
DEL	70 [d]	5 [b]	41 [c]	4 [b]	38 [c]	6 [b]
DUP	78 [cd]	4 [b]	48 [bc]	3 [ab]	31 [c]	4 [ab]
INV	90 [c]	3 [b]	49 [bc]	5 [ab]	42 [bc]	11 [ab]
BND	119 [b]	6 [ab]	59 [b]	1 [ab]	61 [b]	1 [ab]

268 *Different lowercase letters within columns account for significantly different medians at the signifi-*
269 *cance level of 0.05 (Bonferroni-corrected pairwise Wilcoxon rank-sum test)*

270

271 In practice, the interest of researchers and breeding companies may not be the taggability of SVs by
272 WGS SNPs, but by array SNPs. Those come with a different allele frequency spectrum and lower reso-
273 lution than WGS SNPs, which influences the LD patterns [19]. However, they are often available for a
274 huge number of phenotyped individuals due to their use in routine breeding programs. We, therefore,
275 evaluated the potential performance of four publically available chicken genotyping arrays with reso-
276 lutions of 600 k [58], 60 k [59], 55 k [60], and 10 k [61].

277 The availability of variable SNPs close to the variants of interest was strongly dependent on the reso-
278 lution of the arrays. While the 600 k array had a variable array SNP within 15 kb for more than 90 % of

279 the variants in all three populations, the 60 k and the 55 k array came with a slight shift of this depend-
280 ency of having a variable array SNP for > 80 % of the variants at 50 kb and > 90 % at 100 kb (Figure S
281 16). The 10 k array, however, contained no variable array SNP for 50 % of the variants within 100 kb.
282 A non-random difference in SNP density by variant type is not present for any array. The reduced den-
283 sity compared to WGS also reduced the taggability. Mean r^2_{tag} values for SNPs and DELs reached be-
284 tween 0.06 for BR and the 10 k array and 0.65 for WL and the 600 k array within 100 kb distance (Figure
285 S 15). Interestingly, DELs seem to be slightly stronger tagged than SNPs in BL and WL (Figure S 15),
286 while the other variant types were tagged by maximally 50 % of the level which was reached in SNPs
287 and DELs. The results are comparable when checking the proportion of variants with r^2_{tag} > 0.75 (Figure
288 S 17). 40 % of the WGS SNPs and even 45 % of DELs were tagged with more than r^2_{tag} > 0.75 by a SNP
289 of the 600 k array in WL. In contrast, less than 1 % of SNPs and DELs were tagged by a SNP of the 10k
290 array in BR.

291 Discussion

292 Implications from the SV calling pipeline

293 The SV calling approach was intended to return highly accurate variant calls, therefore prioritizing pre-
294 cision over sensitivity. This especially required the exclusion of regions with unusually high coverage,
295 as they may be artefacts of inaccurate read mapping in regions of low sequence complexity [62]. As
296 those regions are known to be hot spots for SV formation by non-allelic homologous recombination
297 (NAHR) [28–30, 63], we expect to have missed a significant proportion of SVs, especially multi-copy
298 DUP. Further, there was a missing overlap between DELLY and MANTA at INS calling, resulting in no
299 INS calls. A generally weak power in INS calling from short reads is expected, though [14]. Those two
300 problems highlight the need for long-read sequencing data for future studies, which should allow for
301 improved resolution of complex regions and comes with improved abilities for INS calling [1, 15]. The
302 limitations of the calling approach and the resulting characteristics of the callset need to be considered

303 when comparing our results to SV callsets that were derived by different approaches and therefore
304 probably capturing SVs with different properties.

305 We further identified a lack of homozygous calls of DUPs, INVs, and BNDs with regard to HWE (Figure
306 4, Figure S 11, Figure S 12). One possible reason may be a deleterious load and therefore purifying
307 selection on those variants. While literature highlights the deleterious potential of DELs, INVs, and
308 BNDs [2, 64], DUPs are rather considered positive by increasing gene expression [2, 3]. In our case,
309 DELs rather show a slight excess of homozygotes than an expected lack under purifying selection (Fig-
310 ure 4, Figure S 11, Figure S 12). The lack of homozygous calls was instead present for DUPs, INVs, and
311 BNDs. Additionally, VEP impact predictions classified 99.6 % of the DEL impacts as MODIFIER and only
312 0.4 % as HIGH, while DUP impacts were classified as HIGH in 10 % of the cases. The discrepancy with
313 literature for DELs may partly be due to past inbreeding in the populations [52, 65], which resulted in
314 small effective population sizes [66] and therefore may have purged strongly deleterious DELs [67, 68].
315 Purging of deleterious DELs may, together with limitations of the used SV callers, also be a reason for
316 the relatively short sizes of the called DELs. Nevertheless, as none of the INVs and BNDs had predicted
317 impacts besides MODIFIER, a second reason seems to be more likely: There may be deficits of the
318 genotypers in accurately distinguishing between heterozygous and homozygous calls of DUPs, INVs,
319 and BNDs.

320 LD decay results

321 The overall levels of SNP – SNP LD within the populations reflect the knowledge from the literature
322 [19, 66] and the different levels of variability (BR > BL > WL) [52, 69]. This resulted in WL having the
323 strongest overall level of LD and BR the weakest. Besides that and if not especially indicated differently,
324 results were the same for all three populations throughout the following sections.

325 The DEL – SNP LD, all in all, was on the same level as SNP – SNP LD. This implies good predictability of
326 DEL effects by SNP call sets and is in accordance with the majority of the existing studies [6, 21–24].
327 Studies that found DEL – SNP LD to be on a reduced level compared to SNP – SNP LD mostly performed

328 the DEL calling from SNP arrays, which implies low breakpoint resolution [32]. It is also common to
329 merge CNV to copy number variable regions (CNVR) in SNP array or read-depth-based studies [32].
330 Therefore, a CNVR can reflect multiple mutation events and not only a single variant, resulting in re-
331 duced LD to bivariate SNPs, an effect we do not expect to be present in our data due to the more
332 precise variant definition.

333 The level of DUP – SNP LD was strongly reduced compared to SNP – SNP LD and DEL – SNP LD, which
334 is in accordance with the existing studies [22, 26, 28, 32]. However, levels of ~40 % of the SNP – SNP
335 LD (Table 1) were higher than what was found e.g. by Lee *et al.* [32], who found DUP – SNP LD to be
336 ~20 % of SNP – SNP LD in two dairy cattle populations. A main factor of DUP – SNP LD being reduced
337 compared to SNP – SNP LD may be due to the lower allele frequencies of DUP in our callset (Figure 2)
338 and therefore increased local ΔMAF (Figure S 7, Figure S 8) in BL and WL. Removing the ΔMAF depend-
339 ent part of LD by expressing LD as r_S^2 increased the relative r^2 of 30 % to a relative r_S^2 of 68 % of the
340 SNP – SNP r_S^2 (+28 %,Table 1). This means that local differences in the allele frequency spectra between
341 SNPs and DUP account for ~50 % of the difference between SNP – SNP LD and DUP – SNP LD

342 A second cause for reduced DUP – SNP LD could be a higher rate of genotyping errors in DUP. In fact,
343 we identified a significant reduction of homozygous DUP calls compared to HWE (Figure 4, Figure S 11,
344 Figure S 12) as already discussed above. The potential genotyping inaccuracy may additionally be sup-
345 ported by the, admittedly subjective, observation of the two assessors during the visual filtering step
346 that DUP came with less clear support than DEL. This, however, resulted only in a moderately reduced
347 inter-observer reliability of 94 % in DUP compared to 97 % in DEL (Supplementary file 3).

348 A further possibility of reduced DUP – SNP LD may be the occurrence of multi-copy CNVs (mCNVs) [28,
349 30] in our callset. DUP in the callset may partly represent CNVs that occur with different copy numbers
350 and are therefore multi- instead of bivariate variants. This reduces the linkage to bivariate SNPs. We
351 saw slight support for the occurrence of some mCNV in the callset e.g. by some high DHFFC values.

352 However, mCNVs are known to cluster in special regions of the genome [28] due to non-allelic homol-
353 ogous recombination (NAHR) as a formation mechanism [29, 70]. Note that NAHR can also occur re-
354 currently [29], resulting in variants that are called bivariate but stem from multiple mutation events.
355 As those clusters should result in high-coverage regions, which we removed in the filtering step, we do
356 not expect a higher number of mCNV and recurrent mutations in our callset.

357 We also evaluated the linkage between SNPs and INV/ BND and found low levels of LD (26.8 % and
358 18.5 % of SNP – SNP LD). The reduced LD in our study is again partly due to local allele frequency
359 differences (Figure S 6 - Figure S 8) as for DUP. Relative r_s^2 values were therefore 14 % to 30 % higher
360 than relative r^2 values (Table 1). However, r_s^2 values for INV – SNP and BND – SNP were still only ~50 %
361 of SNP – SNP r_s^2. The remaining gap may partly be due to genotyping problems. We identified the lack
362 of homozygous calls for INVs and BNDs (Figure 4, Figure S 11, Figure S 12) as for DUPs. In combination
363 with the missing ability to use coverage information for filtering, we would trust the INV and BND
364 genotypes least in our callset. In contrast to our results, Sudmant *et al.* [28] found INV to be in good
365 LD to SNPs in a very accurate callset from 2,504 human genomes, which further supports that the
366 accuracy of INV calls was low in our study.

367 Taggability

368 The analysis of taggability revealed comparable patterns as the LD decay. A high fraction of SNPs and
369 DEL was tagged by close-by WGS SNPs in all three populations (Figure 5; Figure S 14), while only a small
370 fraction of DUPs, INVs, and BNDs was tagged. However, in contrast to the decay patterns, SNPs on
371 average were tagged slightly stronger than DEL, and between 5 % and 10 % more SNPs were tagged
372 with r_{tag}^2 > 0.75 than DEL. A reason for the higher taggability of SNPs compared to DEL, while the LD
373 decay does not differ, may be the reduced SNP density around DELs (Table 2), as the chance for higher
374 maximum values increases with the number of SNPs in the region of interest. In contrast, DELs were
375 tagged slightly better by array SNPs than WGS SNPs by array SNPs. In the case of array SNPs, no locally
376 increased density was present, as array design aims at an equidistant spacing of markers across the
377 genome [58]. This resulted in no difference between the taggability of SNPs and DELs by array SNPs.

378 Potential issues of excluding SNPs in complex regions during array design as suggested by Lee *et al.*
379 [32] as a reason for reduced CNV – SNP LD, were not observed in this study, as we excluded SVs in
380 those regions due to a minor calling accuracy. Using array SNPs to tag the WGS variants further re-
381 vealed a strong need for dense marker maps to provide good tag SNPs, as only the 600 k array could
382 provide tag SNPs with r^2_{tag} > 0.75 for more than 25 % of SNPs and DEL. This may largely explain why
383 e.g. Xu *et al.* [33] found a quarter of CNVs that were significantly associated with milk traits in Holstein
384 cattle to be not tagged by SNPs of a 50 k array. It suggests that this is not solely due to the nature of
385 CNV but that they also missed a comparable fraction of effects, which are caused by SNPs.

386 The concept of taggability is especially relevant for GWAS, where phenotype-marker associations are
387 tested for each marker separately. The strength of the LD between marker and causal variant then
388 directly influences the power of the GWAS. However, the absence of single tag SNPs does not imply
389 that the effect of an SV cannot be captured by a longer haplotype. Methods that utilize effects of
390 multiple SNP at once (e.g. ridge regression best linear unbiased prediction [71]), of which each can
391 explain a slightly different fraction of the variance of the causal variant, may be more robust in this
392 sense. Additionally, imputation of known SVs would probably be a way to overcome the issue of low
393 taggability and needs further investigation.

394 Conclusions

395 We evaluated LD patterns between a comprehensive SV callset and surrounding SNPs in three com-
396 mercial chicken populations. We found DEL – SNP LD to be on the same level as SNP – SNP LD, while
397 DUP – SNP, INV – SNP, and BND – SNP LD were strongly reduced. This was in accordance with the
398 availability of tag SNPs for a high share of SNPs and DELs, while tag SNPs for DUPs were rare and mostly
399 missing for INVs and BNDs. Different arrays came with a density-dependent ability to tag WGS SNPs
400 and SVs but did not show strong systematic differences compared with taggability by WGS SNPs. The
401 main reason for existing differences in SNP – SNP and DUP/INV/BND – SNP LD in our study was due to
402 local MAF differences. Those accounted for ~50 % of this difference in the strength of LD. This implies

403 that genomic variance due to DELs in the chicken populations studied can be captured by different SNP
404 marker sets as good as variance from WGS SNPs, whereas separate SV calling might be advisable for
405 DUP, INV, and BND effects.

406 Material and Methods

407 Data

408 The study used WGS data of 25 white layers, 25 brown layers, and 40 broiler chickens. The raw data
409 was first published by Qanbari *et al.* [52], which contains more information about the samples. Chick-
410 ens were paired-end sequenced with a median coverage between 5 X and 17 X, read length of 100 bp
411 (WL + BL) or 126 bp (BR), and insert sizes of ~400 bp. Basic quality statistics can be found in Supple-
412 mentary file 2 as MultiQC report [72].

413 Population integrity was controlled using principal component analysis in plink 1.9 [73]. The SNPs were
414 first LD pruned by setting the --indep-pairwise flag to sliding windows of 50 kb, a stepsize of five SNPs
415 and an r^2 of 0.5. Based on the pruned SNPs, plink extracted then 90 prime components. Results for
416 the first four prime components and the variance explained can be found in Figure S 18. The first two
417 prime components, which in total accounted for 33.2 % of the total variance, clearly separated broilers,
418 white- and brown layers. The two broiler subpopulations were only slightly separated by the second
419 prime component and clearly by the third, which accounted for 4.5 % of the total variance. The fourth
420 component started splitting one of the broiler populations. We assumed this to be sufficiently closely
421 related to consider the two broiler subpopulations as a combined population for further analyses.

422 Variant Calling Pipeline

423 Alignment on the reference genome galGal6/ GRGC6a and SNP calling were conducted in a previous
424 study [53] following GATK best practices pipeline [74]. The SNPs needed for this study were then ex-
425 tracted from the old callset using bcftools [75] and the duplicate-marked and base quality score recal-
426 ibrated BAM files were used as starting point for the SV calling process.

427 SV calling was conducted following a consensus calling approach. SVs were first separately called per
428 individual and then genotyped on population-level by running Delly 0.8.5 [10], Manta 1.6.0 [12], and a
429 combination of Lumpy 0.2.13 [11] and Svtyper 0.7.0 [76] in parallel on the complete set. The genotyp-
430 ing results of the three calling pipelines were then merged using SURVIVOR 1.0.7 [77] and allowing for
431 breakpoint differences of 1000 bp. This resulted in 95,478 raw SV calls.

432 Additionally, read depth profiles for all samples in 100 bp windows were generated using Mosdepth
433 0.2.9 [78] and SVs were annotated with Dupholds (version 0.2.1) [57] flanking fold change (DHFFC) and
434 the SNP genotype calls located on the SV.

435 The merged callset was then filtered based on the following parameters:

436 1) Caller overlap: At least two of the three callers needed to support the variant.
437 2) Genotype concordance: The genotype that was supported by two out of the three callers was
438 considered as the consensus genotype. Genotypes without the necessary support were set to
439 missing for later re-imputation. If more than two samples did not have the necessary genotype
440 concordance support for an SV, the complete SV was removed from the data set.
441 3) Removal of high coverage regions: Local coverage was extracted by Mosdepth 0.2.9 [78] in 100
442 bp windows. If windows exceeded a threshold of twice the average coverage across all samples
443 (expected value for a fixed DUP) plus two standard deviations, they were classified as unusu-
444 ally highly covered. Unusually highly covered regions were further merged if they were less
445 than 1000 bp apart from each other. SVs with breakpoint confidence intervals falling in such a
446 region were removed from the data set.
447 4) Difference to flanking coverage: DELs and DUPs calls were checked for non-consistent cover-
448 age changes relative to the flanking coverage by evaluating the Duphold Flanking Fold Change
449 (DHFFC) [57]. DELs were considered as wrong genotypes when heterozygotes were not be-
450 tween 0.1 and 0.9 and homozygous DEL genotypes not smaller than 0.25. Heterozygous DUPs
451 had to be >1.1 and homozygous DUPs >1.5. DELs/DUPs with more than one error or more than

452 10 % wrong genotypes were filtered. Otherwise, the putatively wrong DEL/DUP genotypes
453 were set to missing for later re-imputation.
454 5) Support by SNP calls on DELs: SNP calls need to be homozygous on heterozygous DELs and
455 missing on homozygous DELs. We, therefore, calculated for each DEL genotype the relative
456 number of wrong SNP genotypes (e.g. one error by five total SNPs on the DEL = 0.1). If the sum
457 of those error rates across samples exceeded two or 50 % of the number of samples that were
458 at least heterozygous for the DEL, the DEL was filtered. Otherwise, the putatively wrong DEL
459 genotypes were set to missing for later re-imputation.

460 This resulted in 5,600 SVs (4,831 DELs; 253 DUPs; 346 INVs; 170 BNDs; 94.1 % filtered). No INS re-
461 mained, as Lumpy does not call INS and there was no overlap between Delly and Manta. Samplot 1.0.19
462 [54] was then used to generate quality control plots for each SV that passed the previous filtering step.
463 The quality plots were visually screened by two separate observers comparable to the workflow im-
464 plemented in SV-plaudit [79], but implemented locally by using image-sorter2
465 (https://github.com/Nestak2/image-sorter2). The SVs needed to be scored as 'pass' by each of the two
466 observers to be further used (Supplementary file 3). By this, a further 6.9 % of the SVs (3.5 % of DEL,
467 11.1 % of DUP, 36.0 % of INV, and 30.8 % of BND) were removed. The removed SVs were mainly in
468 regions with complex mapping patterns.

469 The final SV callset (4,301 DEL, 224 DUP, 218 INV, 117 BND) was then merged with the SNP callset
470 (12,294,329 bivariate autosomal SNPs). The samples were phased and missing genotypes were im-
471 puted by beagle 5.0 [80] with default settings besides reducing 'ne' to 10,000 [81]. Functional conse-
472 quences were annotated by ensembl-vep [55] using the release 100 GRGC6a annotation files.

473 Estimation of LD

474 LD between two loci with a maximum distance of 100kb was initially estimated from phased haplo-
475 types as follows:

476 $$r_{AB}^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_B (1 - p_A)(1 - p_B)}$$

477 Where p_A and p_B account for the alternative allele frequencies at the two loci and p_{AB} for the accord-
478 ing haplotype frequency. To control for allele frequency deviations that influence the maximum possi-
479 ble r^2, we further scaled r^2 by the maximum possible r^2 given ΔMAF ($r_S^2 = r^2 / r_{\max\ |\Delta MAF}^2$) where
480 $r_{\max\ |\Delta MAF}^2$ was derived as described by VanLiere and Rosenberg [20]. As we realized a problem with
481 calling of homozygous DUP, we additionally estimated LD as squared Pearson Correlation between
482 0/1/2 coded SNP genotypes and the Duphold Flanking Fold Change (DHFFC) [57] as a measure for the
483 relative reference genome coverage at DEL and DUP (due to possible confounding only part of Supple-
484 mentary file 1). LD decay was then summarized in means of 500 bp bins between the variants.

485 Bonferroni corrected bootstrap confidence intervals for the LD decay were estimated by resampling
486 the r^2 values within each bin 100,000 times with replacement. As tests showed confidence intervals
487 for SNP - SNP LD being < 0.001 due to the huge number of underlying values, we decided to skip esti-
488 mation of confidence intervals for bins with > 1M r^2 values.

489 A tag SNP was defined as the SNP with the highest r^2 to the variant of interest within a certain distance
490 (r_{tag}^2). The taggability of variant classes was then investigated by comparing means of r_{tag}^2 and shares
491 of variants with $r_{tag}^2 > 0.75$. Additionally to the taggability by WGS SNPs, we compared the taggability
492 by SNPs of four commercially available SNP arrays. The 600 k Affymetrix Axiom chicken genotyping
493 array [58], a 60 k Illumina Bead Chip [59], a 55 k Affymetrix genotyping array [60], and the IMAGE_001
494 multispecies array, which contains 10 k chicken-specific SNPs on an Affymetrix genotyping array [61].
495 The annotation files were lifted over to the reference genome galGal6/GRGC6a by the UCSC [82] lift-
496 Over tool under usage of the according chain files and the overlaps with the variable WGS SNPs were
497 defined as pools of potential Array tag SNPs.

498 ### Workflow

499 The complete pipeline was set up in snakemake 5.3.0 [83] and the according scripts including the
500 snakefile with all used parameters as well as the dependency analytics graph (DAG) and the rulegraph
501 of the pipeline can be found in the supplementary material (Supplementary file 4).

502 ## Declarations

503 ### Ethics approval and consent to participate

504 The study did not involve new treatment of animals as only published data was used. DNA samples for
505 all already published raw data were taken from a database established during the project AVIANDIV
506 (EC Contract No. BIO4-CT98_0342; 1998 – 2000; https://aviandiv.fli.de/) and later extended by sam-
507 ples of the project SYNBREED (FKZ 0315528E; 2009 – 2014; www.synbreed.tum.de). Blood sampling
508 was done in strict accordance to the German animal welfare regulations, with written consent of the
509 animal owners and was approved by the at the according times ethics responsible persons of the Frie-
510 drich-Loeffler-Institut. According to German animal welfare regulations, notice was given to the re-
511 sponsible institution, the Lower Saxony State Office for Consumer Protection and Food Safety (33.9-
512 42502-05-10A064).

513 ### Consent for publication

514 Not applicable

515 ### Availability of data and materials

516 The raw fastq files were already published by Qanbari et al. [52] and can be accessed via the ENA
517 project PRJEB30270 (https://www.ebi.ac.uk/ena/browser/view/PRJEB30270). The snakemake work-
518 flow together with all scripts and supplementary files can be found in Supplementary file 4. Interme-
519 diate results are available from the corresponding author on reasonable request.

520 ### Competing interests

521 The authors declare that they have no competing interests.

Funding

We acknowledge support by the Open Access Publication Funds of the University of Goettingen.

Authors' contributions

JG designed the study, analyzed the data, interpreted the results, and wrote the initial manuscript. NPP substantially contributed to the data analysis. SW contributed to the acquisition of the data and revision of the manuscript. HS and CR contributed to design of the study, interpretation of the results, and substantially revised the manuscript. All authors have read and approved the submitted manuscript.

Acknowledgments

Not applicable

Supplementary

Supplementary file 1: Supplementary results, tables and figures.

Supplementary file 2: MultiQC report.

Supplementary file 3: Observer concordance of the visual filtering step.

Supplementary file 4: Snakemake workflow and scripts.

References

1. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nature Reviews Genetics 2019. doi:10.1038/s41576-019-0180-9.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7:85–97. doi:10.1038/nrg1767.
3. Lee Y-L, Takeda H, Moreira GCM, Karim L, Mullaart E, Coppieters W, et al. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle; 2021.

545 4. Imsland F, Feng C, Boije H, Bed'hom B, Fillon V, Dorshorst B, et al. The Rose-comb mutation in
546 chickens constitutes a structural rearrangement causing both altered comb morphology and de-
547 fective sperm motility. PLoS Genet. 2012;8:e1002775. doi:10.1371/journal.pgen.1002775.
548 5. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signa-
549 tures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012;109:19529–36.
550 doi:10.1073/pnas.1217149109.
551 6. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, et al. Common deletion pol-
552 ymorphisms in the human genome. Nature Genetics. 2006;38:86–92. doi:10.1038/ng1696.
553 7. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion
554 polymorphism in the human genome. Nature Genetics. 2006;38:75–81. doi:10.1038/ng1697.
555 8. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: an integrated hidden Markov
556 model designed for high-resolution copy number variation detection in whole-genome SNP geno-
557 typing data. Genome Res. 2007;17:1665–74. doi:10.1101/gr.6861907.
558 9. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and
559 characterize typical and atypical CNVs from family and population genome sequencing. Genome
560 Res. 2011;21:974–84. doi:10.1101/gr.114876.110.
561 10. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery
562 by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333-i339.
563 doi:10.1093/bioinformatics/bts378.
564 11. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant
565 discovery. Genome Biology. 2014;15:R84. doi:10.1186/gb-2014-15-6-r84.
566 12. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detec-
567 tion of structural variants and indels for germline and cancer sequencing applications. Bioinfor-
568 matics. 2016;32:1220–2. doi:10.1093/bioinformatics/btv710.

13. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14:305–14. doi:10.1093/bfgp/elv014.
14. Delage WJ, Thevenon J, Lemaitre C. Towards a better understanding of the low recall of insertion variants with short-read based variant callers. BMC genomics. 2020;21:762. doi:10.1186/s12864-020-07125-5.
15. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics. 2018;19:329–46. doi:10.1038/s41576-018-0003-4.
16. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. doi:10.1038/nature15393.
17. Hayes BJ, Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. Annual Review of Animal Biosciences. 2019;7:89–102. doi:10.1146/annurev-animal-020518-115024.
18. Malomane DK, Simianer H, Weigend A, Reimer C, Schmitt AO, Weigend S. The SYNBREED chicken diversity panel: A global resource to assess chicken diversity at high genomic resolution. BMC genomics. 2019;20:345. doi:10.1186/s12864-019-5727-9.
19. Qanbari S. On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. Frontiers in Genetics. 2020;10:1304. doi:10.3389/fgene.2019.01304.
20. VanLiere JM, Rosenberg NA. Mathematical properties of the r2 measure of linkage disequilibrium. Theoretical population biology. 2008;74:130–7.
21. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nature Genetics. 2006;38:82–5. doi:10.1038/ng1695.
22. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464:704–12. doi:10.1038/nature08516.

595 23. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation
596 by population-scale genome sequencing. Nature. 2011;470:59–65. doi:10.1038/nature09708.
597 24. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number
598 variant detection via genome-wide SNP genotyping. Nature Genetics. 2008;40:1199–203.
599 doi:10.1038/ng.236.
600 25. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number
601 in the human genome. Nature. 2006;444:444–54. doi:10.1038/nature05329.
602 26. Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH, et al. Population-genetic
603 nature of copy number variations in the human genome. Hum Mol Genet. 2009;19:761–73.
604 doi:10.1093/hmg/ddp541.
605 27. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection
606 and population-genetic analysis of SNPs and copy number variation. Nature Genetics.
607 2008;40:1166–74. doi:10.1038/ng.238.
608 28. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated
609 map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81. doi:10.1038/na-
610 ture15394.
611 29. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. PathoGenetics.
612 2008;1:4. doi:10.1186/1755-8417-1-4.
613 30. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, et al. Linkage Disequilibrium
614 and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Ge-
615 nome. The American Journal of Human Genetics. 2006;79:275–90. doi:10.1086/505653.
616 31. Berger S, Schlather M, los Campos G de, Weigend S, Preisinger R, Erbe M, Simianer H. A Scale-
617 Corrected Comparison of Linkage Disequilibrium Levels between Genic and Non-Genic Regions.
618 PLOS ONE. 2015;10:e0141216. doi:10.1371/journal.pone.0141216.

619 32. Lee Y-L, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC. Functional and popula-
620 tion genetic features of copy number variations in two dairy cattle populations. BMC genomics.
621 2020;21:89. doi:10.1186/s12864-020-6496-1.
622 33. Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, et al. Genome wide CNV analysis reveals
623 additional variants associated with milk production traits in Holsteins. BMC genomics.
624 2014;15:683. doi:10.1186/1471-2164-15-683.
625 34. Wang L, Xu L, Liu X, Zhang T, Li N, Hay EH, et al. Copy number variation-based genome wide asso-
626 ciation study reveals additional variants contributing to meat quality in Swine. Scientific Reports.
627 2015;5:12535. doi:10.1038/srep12535.
628 35. Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefope N. An initial map of chromosomal
629 segmental copy number variations in the chicken. BMC genomics. 2010;11:351.
630 doi:10.1186/1471-2164-11-351.
631 36. Wang Y, Gu X, Feng C, Song C, Hu X, Li N. A genome-wide survey of copy number variation regions
632 in various chicken breeds by array comparative genomic hybridization method. Anim Genet.
633 2012;43:282–9. doi:10.1111/j.1365-2052.2011.02308.x.
634 37. Crooijmans RP, Fife MS, Fitzgerald TW, Strickland S, Cheng HH, Kaiser P, et al. Large scale variation
635 in DNA copy number in chicken breeds. BMC genomics. 2013;14:398. doi:10.1186/1471-2164-14-
636 398.
637 38. Tian M, Wang Y, Gu X, Feng C, Fang S, Hu X, Li N. Copy number variants in locally raised Chinese
638 chicken genomes determined using array comparative genomic hybridization. BMC genomics.
639 2013;14:262. doi:10.1186/1471-2164-14-262.
640 39. Han R, Yang P, Tian Y, Wang D, Zhang Z, Wang L, et al. Identification and functional characterization
641 of copy number variations in diverse chicken breeds. BMC genomics. 2014;15:934.
642 doi:10.1186/1471-2164-15-934.
643 40. Jia X, Chen S, Zhou H, Li D, Liu W, Yang N. Copy number variations identified in the chicken using
644 a 60K SNP BeadChip. Anim Genet. 2013;44:276–84. doi:10.1111/age.12009.

645 41. Gorla E, Cozzi MC, Román-Ponce SI, Ruiz López FJ, Vega-Murillo VE, Cerolini S, et al. Genomic var-
646 iability in Mexican chicken population using copy number variants. BMC Genet. 2017;18:61.
647 doi:10.1186/s12863-017-0524-4.
648 42. Zhang H, Du Z-Q, Dong J-Q, Wang H-X, Shi H-Y, Wang N, et al. Detection of genome-wide copy
649 number variations in two chicken lines divergently selected for abdominal fat content. BMC ge-
650 nomics. 2014;15:517. doi:10.1186/1471-2164-15-517.
651 43. Rao YS, Li J, Zhang R, Lin XR, Xu JG, Xie L, et al. Copy number variation identification and analysis
652 of the chicken genome using a 60K SNP BeadChip. ps. 2016;95:1750–6. doi:10.3382/ps/pew136.
653 44. Strillacci MG, Cozzi MC, Gorla E, Mosca F, Schiavini F, Román-Ponce SI, et al. Genomic and genetic
654 variability of six chicken populations using single nucleotide polymorphism and copy number var-
655 iants as markers. Animal. 2017;11:737–45. doi:10.1017/S1751731116002135.
656 45. Lin S, Lin X, Zhang Z, Jiang M, Rao Y, Nie Q, Zhang X. Copy Number Variation in SOX6 Contributes
657 to Chicken Muscle Development. Genes (Basel) 2018. doi:10.3390/genes9010042.
658 46. Fan W-L, Ng CS, Chen C-F, Lu M-YJ, Chen Y-H, Liu C-J, et al. Genome-wide patterns of genetic vari-
659 ation in two domestic chickens. Genome Biol Evol. 2013;5:1376–92. doi:10.1093/gbe/evt097.
660 47. Yan Y, Yang N, Cheng HH, Song J, Qu L. Genome-wide identification of copy number variations
661 between two chicken lines that differ in genetic resistance to Marek's disease. BMC genomics.
662 2015;16:843. doi:10.1186/s12864-015-2080-5.
663 48. Sohrabi SS, Mohammadabadi M, Wu D-D, Esmailizadeh A. Detection of breed-specific copy num-
664 ber variations in domestic chicken genome. Genome. 2018;61:7–14. doi:10.1139/gen-2017-0016.
665 49. Seol D, Ko BJ, Kim B, Chai H-H, Lim D, Kim H. Identification of Copy Number Variation in Domestic
666 Chicken Using Whole-Genome Sequencing Reveals Evidence of Selection in the Genome. Animals
667 (Basel) 2019. doi:10.3390/ani9100809.
668 50. Weng Z, Xu Y, Li W, Chen J, Zhong M, Zhong F, et al. Genomic variations and signatures of selection
669 in Wuhua yellow chicken. PLoS One. 2020;15:e0241137. doi:10.1371/journal.pone.0241137.

670 51. Kerstens HHD, Crooijmans RP, Dibbits BW, Vereijken A, Okimoto R, Groenen M am. Structural var-
671 iation in the chicken genome identified by paired-end next-generation DNA sequencing of reduced
672 representation libraries. BMC genomics. 2011;12:94. doi:10.1186/1471-2164-12-94.
673 52. Qanbari S, Rubin C-J, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in
674 modern chicken. PLoS Genet. 2019;15:e1007989. doi:10.1371/journal.pgen.1007989.
675 53. Geibel J, Reimer C, Pook T, Weigend S, Weigend A, Simianer H. How Imputation Can Mitigate As-
676 certainment Bias; 2021.
677 54. Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, Layer RM. Samplot: A
678 Platform for Structural Variant Visual Validation and Automated Filtering. bioRxiv 2020.
679 doi:10.1101/2020.09.23.310110.
680 55. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect
681 Predictor. Genome Biology. 2016;17:122. doi:10.1186/s13059-016-0974-4.
682 56. Graffelman J. Exploring Diallelic Genetic Markers: The HardyWeinberg Package. Journal of Statis-
683 tical Software. 2015;64:1–23.
684 57. Pedersen BS, Quinlan AR. Duphold: scalable, depth-based annotation and curation of high-confi-
685 dence structural variant calls. Gigascience. 2019;8:giz040. doi:10.1093/gigascience/giz040.
686 58. Kranis A, Gheyas AA, Boschiero C, Turner F, Le Yu, Smith S, et al. Development of a high density
687 600K SNP genotyping array for chicken. BMC genomics. 2013;14:59. doi:10.1186/1471-2164-14-
688 59.
689 59. Groenen MAM, Megens H-J, Zare Y, Warren WC, Hillier LW, Crooijmans RPMA, et al. The develop-
690 ment and characterization of a 60K SNP chip for chicken. BMC genomics. 2011;12:274.
691 doi:10.1186/1471-2164-12-274.
692 60. Liu R, Xing S, Wang J, Zheng M, Cui H, Crooijmans RPMA, et al. A new chicken 55K SNP genotyping
693 array. BMC genomics. 2019;20:410. doi:10.1186/s12864-019-5736-8.

694 61. Innovative Management of Animal Genetic Resources (IMAGE). DELIVERABLE D4.5: A standard
695 multi-species chip for genomic assessment of collections. 2020. https://www.imageh2020.eu/de-
696 liverable/D4.5_resubmitted_final.pdf. Accessed 17 Aug 2021.
697 62. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bio-
698 informatics. 2014;30:2843–51. doi:10.1093/bioinformatics/btu356.
699 63. Bickhart DM, Liu GE. The challenges and importance of structural variation detection in livestock.
700 Frontiers in Genetics. 2014;5:37. doi:10.3389/fgene.2014.00037.
701 64. Bouwman AC, Derks MF, Broekhuijse ML, Harlizius B, Veerkamp RF. Using short read sequencing
702 to characterise balanced reciprocal translocations in pigs 2020. doi:10.21203/rs.3.rs-28830/v3.
703 65. Talebi R, Szmatoła T, Mészáros G, Qanbari S. Runs of Homozygosity in Modern Chicken Revealed
704 by Sequence Data. G3 (Bethesda). 2020;10:4615–23. doi:10.1534/g3.120.401860.
705 66. Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H. Linkage disequilibrium reveals different
706 demographic history in egg laying chickens. BMC Genet. 2010;11:103. doi:10.1186/1471-2156-11-
707 103.
708 67. Bortoluzzi C, Bosse M, Derks MFL, Crooijmans RPMA, Groenen MAM, Megens H-J. The type of
709 bottleneck matters: Insights into the deleterious variation landscape of small managed popula-
710 tions. Evol Appl. 2020;13:330–41. doi:10.1111/eva.12872.
711 68. Kyriazis CC, Wayne RK, Lohmueller KE. Strongly deleterious mutations are a primary determinant
712 of extinction risk due to inbreeding depression. Evolution Letters 2020. doi:10.1002/evl3.209.
713 69. Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. How array design creates SNP
714 ascertainment bias. PLoS One. 2021;16:e0245178. doi:10.1371/journal.pone.0245178.
715 70. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature
716 Reviews Genetics. 2009;10:551–64. doi:10.1038/nrg2593.
717 71. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense
718 marker maps. Genetics. 2001;157:1819–29.

719 72. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools
720 and samples in a single report. Bioinformatics. 2016;32:3047–8. doi:10.1093/bioinformat-
721 ics/btw354.
722 73. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for
723 whole-genome association and population-based linkage analyses. Am. J. Hum. Genet.
724 2007;81:559–75. doi:10.1086/519795.
725 74. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis
726 Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome
727 Res. 2010;20:1297–303. doi:10.1101/gr.107524.110.
728 75. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and popula-
729 tion genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
730 doi:10.1093/bioinformatics/btr509.
731 76. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast per-
732 sonal genome analysis and interpretation. Nat Methods. 2015;12:966–8.
733 doi:10.1038/nmeth.3505.
734 77. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have
735 strong effects on quantitative traits and reproductive isolation in fission yeast. Nature Communi-
736 cations. 2017;8:14061. doi:10.1038/ncomms14061.
737 78. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioin-
738 formatics. 2018;34:867–8. doi:10.1093/bioinformatics/btx699.
739 79. Belyeu JR, Nicholas TJ, Pedersen BS, Sasani TA, Havrilla JM, Kravitz SN, et al. SV-plaudit: A cloud-
740 based framework for manually curating thousands of structural variants. Gigascience 2018.
741 doi:10.1093/gigascience/giy064.
742 80. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Refer-
743 ence Panels. Am. J. Hum. Genet. 2018;103:338–48. doi:10.1016/j.ajhg.2018.07.015.

81. Pook T, Mayer M, Geibel J, Weigend S, Cavero D, Schoen CC, Simianer H. Improving Imputation Quality in BEAGLE for Crop and Livestock Data. G3 (Bethesda). 2020;10:177–88. doi:10.1534/g3.119.400798.
82. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996–1006. doi:10.1101/gr.229102.
83. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2. doi:10.1093/bioinformatics/bts480.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile1.pdf
- Supplementaryfile2.html
- Supplementaryfile1.pdf
- Supplementaryfile3.html
- Supplementaryfile4.zip
- Supplementaryfile2.html
- Supplementaryfile3.html
- Supplementaryfile4.zip