Ros-Freixedes *et al.*
Genetics Selection Evolution (2022) 54:39
https://doi.org/10.1186/s12711-022-00732-8

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Rare and population-specific functional variation across pig lines

Roger Ros-Freixedes[1,2*], Bruno D. Valente[3], Ching-Yi Chen[3], William O. Herring[3], Gregor Gorjanc[1], John M. Hickey[1] and Martin Johnsson[1,4]

Abstract

Background: It is expected that functional, mainly missense and loss-of-function (LOF), and regulatory variants are responsible for most phenotypic differences between breeds and genetic lines of livestock species that have undergone diverse selection histories. However, there is still limited knowledge about the existing missense and LOF variation in commercial livestock populations, in particular regarding population-specific variation and how it can affect applications such as across-breed genomic prediction.

Methods: We re-sequenced the whole genome of 7848 individuals from nine commercial pig lines (average sequencing coverage: 4.1×) and imputed whole-genome genotypes for 440,610 pedigree-related individuals. The called variants were categorized according to predicted functional annotation (from LOF to intergenic) and prevalence level (number of lines in which the variant segregated; from private to widespread). Variants in each category were examined in terms of their distribution along the genome, alternative allele frequency, per-site Wright's fixation index (F_{ST}), individual load, and association to production traits.

Results: Of the 46 million called variants, 28% were private (called in only one line) and 21% were widespread (called in all nine lines). Genomic regions with a low recombination rate were enriched with private variants. Low-prevalence variants (called in one or a few lines only) were enriched for lower allele frequencies, lower F_{ST}, and putatively functional and regulatory roles (including LOF and deleterious missense variants). On average, individuals carried fewer private deleterious missense alleles than expected compared to alleles with other predicted consequences. Only a small subset of the low-prevalence variants had intermediate allele frequencies and explained small fractions of phenotypic variance (up to 3.2%) of production traits. The significant low-prevalence variants had higher per-site F_{ST} than the non-significant ones. These associated low-prevalence variants were tagged by other more widespread variants in high linkage disequilibrium, including intergenic variants.

Conclusions: Most low-prevalence variants have low minor allele frequencies and only a small subset of low-prevalence variants contributed detectable fractions of phenotypic variance of production traits. Accounting for low-prevalence variants is therefore unlikely to noticeably benefit across-breed analyses, such as the prediction of genomic breeding values in a population using reference populations of a different genetic background.

Background

Genetic variation is the basis of selective breeding in livestock and crop species. From a molecular point of view, genetic variants that result in either altered protein structures or altered gene expressions are believed to be responsible for much of the existing genetic variation for complex traits [1–4]. Missense variants change

*Correspondence: roger.ros@roslin.ed.ac.uk

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, Scotland, UK
Full list of author information is available at the end of the article

one amino acid of the encoded protein. Loss-of-function variants (LOF) are predicted to disrupt protein-coding transcripts such that they will not be translated into proteins or that they will be translated into non-functional proteins. Loss-of-function variants may change one amino acid codon into a premature stop codon (nonsense variants), change the reading frame during translation (frameshift indels), or change mRNA splicing (splicing variants). As such, potentially functional variants in protein-coding regions are assumed to be easier to detect (e.g., by association analyses) than variants that moderate gene expression [5–7]. Thus, missense and LOF variants are typically prioritised as putative causal variants for traits of interest (e.g., [8–11]).

Missense and LOF mutations can be pathogenic. For instance, missense and nonsense variants account for 57% of the entries in the Human Gene Mutation Database [12] (accessed on 30 April 2021), while small indels account for 22% and splicing variants account for another 9%. Similarly, in livestock species, many missense and LOF variants have been described as causal of genetic diseases and post-natal defects ([13–16]; Online Mendelian Inheritance in Animals [17], accessed on 30 April 2021), embryonic lethality [18, 19], or product defects [20, 21]. Deleterious missense and LOF variants are subject to purifying selection and are more likely to be rare, because they are related to unfavourable phenotypes such as disease risk or reduced fertility.

However, some missense and LOF mutations can be beneficial [22]. Moreover, some alleles that would be detrimental in the wild may be preferred in artificial selection settings. The artificial selection that is performed in livestock and crop breeding programs is expected to increase the frequency of alleles that favourably affect the traits included in the selection objectives. Therefore, it is also expected that missense and LOF variants are responsible for differences between breeds, genetic lines, and varieties of livestock and crop species that have undergone diverse selection histories. Identification of such functional variants can have direct applications in gene-assisted and genomic selection [23–25]. Furthermore, strategies for genetic improvement using genome editing have been theorized to either promote favourable alleles [26] or remove deleterious alleles [27] in selection candidates. Nevertheless, there is still limited knowledge about the presence of missense and LOF variants in commercial livestock populations, in particular regarding population-specific variants, often referred to as 'private', and how the presence of population-specific functional variants can affect applications such as across-breed genomic prediction.

Next-generation sequencing holds great potential for livestock breeding. One of its main benefits is the power to detect large numbers of variants, many of which will be specific to the population under study. A large number of individuals must, however, be sequenced in order to achieve high variant discovery rates, particularly for low-frequency variants [28, 29]. Several sequencing studies have profiled genomic variation in pigs [30–32], cattle [33, 34], or chicken [35]. These studies involved the sequencing of a small number of individuals (up to a few hundreds) at intermediate or high sequencing coverage. Alternatively, low sequencing coverage allows affordable sequencing of a much larger number of individuals, which would enable the identification of a much larger number of variants.

The objective of this study was to characterize the genetic variants in nine intensely selected pig lines with diverse genetic backgrounds. Particular emphasis was given to quantifying rare and population-specific functional variants, as well as the number of missense and LOF variants that an average individual carries. We also assessed the contribution of population-specific functional variants to the phenotypic variance of production traits.

Methods

Populations and sequencing strategy

We re-sequenced the whole genome of 7848 pigs from nine commercial lines (Genus PIC, Hendersonville, TN), with a total sequencing coverage of approximately 32,114×. Breeds of origin of the nine lines were Large White, Landrace, Pietrain, Hampshire, Duroc, and synthetic lines. The number of pigs that were available in the pedigree of each line and the number of sequenced pigs, by coverage, are summarized in Table 1.

Which pigs to sequence and their coverage were determined following a three-part sequencing strategy, with the objective of representing the haplotype diversity in each line. First (1), top sires and dams with the largest number of genotyped progeny were sequenced at 2× and 1×, respectively. Sires were sequenced at higher coverage because they individually contributed more progeny than dams. Then (2), individuals with the greatest genetic footprint on the population (i.e., those that carry more of the most common haplotypes) and their immediate ancestors were sequenced at a target sequencing coverage between 1× and 30×, as assigned by an algorithm that maximises the expected phasing accuracy of the common haplotypes from the accumulated family information (AlphaSeqOpt part 1; [36]). Finally (3), pigs that carried haplotypes with a low accumulated coverage (below 10×) were sequenced at 1× (AlphaSeqOpt part 2; [37]). Sets (2) and (3) were based on haplotypes inferred from marker array genotypes (GGP-Porcine HD BeadChip; GeneSeek, Lincoln, NE), which were phased

Table 1 Number of sequenced and analysed pigs

Line	Individuals sequenced	Individuals sequenced by coverage				Individuals used in analyses		
		1×	2×	5×	15–30×	Pedigree	Imputed	GWAS
A	1856	1044	649	73	90	122,753	104,661	88,342
B	1491	628	728	54	81	84,420	66,608	56,173
C	1366	685	545	44	92	88,964	76,230	64,285
D	760	394	274	27	65	50,797	41,573	–
E	731	362	311	16	42	79,981	60,474	–
F	701	351	255	28	67	52,470	39,263	–
G	445	217	176	15	37	21,129	17,224	–
H	381	193	137	16	35	35,309	29,330	–
I	321	111	158	18	34	15,495	5247	–

using AlphaPhase [38] and imputed using AlphaImpute [39]. As a result of this sequencing strategy, sequencing effort in each of the nine lines was proportional to their population size, at approximately 1.5% (0.9–2.1%) of the pigs in each line. Most pigs were sequenced at a low target coverage of 1 or 2×. The average individual coverage was 4.1×, but the median coverage was 1.5×. Population structure across the nine lines was assessed with a principal component analysis using the sequenced pigs and is shown in Additional file 1: Fig. S1.

Most sequenced pigs, as well as pedigree relatives, were also genotyped with marker arrays either at low density (15k markers) using the GGP-Porcine LD BeadChip (GeneSeek) or at high density (50k or 80k markers) using different versions of the GGP-Porcine HD BeadChip (GeneSeek).

Sequencing and data processing

Tissue samples were collected from ear punches or tail clippings. Genomic DNA was extracted using Qiagen DNeasy 96 Blood & Tissue kits (Qiagen Ltd., Mississauga, ON, Canada). Paired-end library preparation was conducted using the TruSeq DNA PCR-free protocol (Illumina, San Diego, CA). Libraries for resequencing at low coverage (1 to 5×) were produced with an average insert size of 350 bp and sequenced on a HiSeq 4000 instrument (Illumina). Libraries for resequencing at high coverage (15 or 30×) were produced with an average insert size of 550 bp and sequenced on a HiSeq X instrument (Illumina). All libraries were sequenced at Edinburgh Genomics (Edinburgh Genomics, University of Edinburgh, Edinburgh, UK).

DNA sequence reads were pre-processed using the Trimmomatic software [40] to remove adapter sequences and then aligned to the reference genome *Sscrofa11.1* (GenBank accession: GCA_000003025.6) using the BWA-MEM algorithm [41]. Duplicates were marked using the Picard software (http://broadinstitute.github.io/picard). Single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) were identified with GATK HaplotypeCaller (GATK 3.8.0) [42, 43] using default settings. Variant discovery was performed separately for each individual and then a joint variant set for each population was obtained by extracting the variant positions from all sequenced individuals. Between 20 and 30 million variants were discovered in each population.

Read counts supporting each allele were directly extracted from the aligned reads stored in the BAM files using a pile-up function in order to avoid biases towards the reference allele that are introduced by the GATK algorithm when applied on low-coverage whole-genome sequence data [44]. This pipeline uses pysam (version 0.13.0; https://github.com/pysam-developers/pysam), which is a wrapper around htslib and the samtools package [45]. We extracted the read counts for all biallelic variant positions, after filtering variants in potential repetitive regions with the VCFtools software [46]. Variants in potential repetitive regions were defined as those that had a mean depth value that was 3 times greater than the average realized coverage. In total, 46,344,624 biallelic variants passed quality control criteria in at least one of the nine lines (see Additional file 2: Supplementary Methods).

Genotype imputation

Genotypes were jointly called, phased and imputed for a total of 537,257 pedigree-related individuals across lines, using the 'hybrid peeling' method implemented in AlphaPeel [47–49], which used all available SNP panels and whole-genome sequence data. Imputation was performed separately for each line using its complete multigenerational pedigree, which encompassed from 15,495 to 122,753 individuals each (Table 1). We have previously published on the accuracy of imputation in the same

populations using this method [48]. The estimated average allele dosage correlation (correlation between the real genotype and the imputed allele dosage) by individual was 0.94 (median 0.97) [48]. Individuals with a low predicted imputation accuracy were removed before further analyses. An individual was predicted to have a low imputation accuracy if itself or all its grandparents were not genotyped with a marker array or if it had a low degree of connectedness to the rest of the population (defined as the sum of coefficients of relationship between the individual and the rest of individuals in the pedigree). These criteria were based on analysis of simulated and real data on imputation accuracy [48]. In total, 440,610 individuals remained, from 5247 to 104,661 individuals for each line (Table 1). The expected average individual-wise dosage correlation of the remaining individuals was 0.97 (median 0.98) [48]. Although variants with a minor allele frequency lower than 0.023 had an estimated variant-wise dosage correlation lower than 0.90 [48], in our analyses, we did not filter variants based on minor allele frequency to account for the whole frequency spectrum.

Variant predicted consequence types

The frequency of the alternative allele was calculated based on the imputed genotypes. The prevalence level of a variant was defined as the number of lines in which the variant segregated. To differentiate allele frequency and prevalence level, we used the terms 'rare' and 'common' to refer to variants in terms of allele frequency and 'private' and 'widespread' in terms of prevalence level, where private variants were those called only in one line and widespread variants those called in all nine studied lines. We calculated Wright's fixation statistic (F_{ST}) [50] for each variant among the lines in which the variant segregated as $F_{ST} = (H_T - H_S)/H_T$, where H_T is the expected heterozygosity across the lines under Hardy–Weinberg equilibrium and H_S is the average heterozygosity within lines under Hardy–Weinberg equilibrium.

Variants were annotated using Ensembl Variant Effect Predictor (Ensembl VEP; version 97, July 2019) [51] using both Ensembl and RefSeq transcript databases. For variants with multiple predicted consequence types (e.g., in the case of multiple transcripts), the variant was annotated with the most severe predicted consequence type. Stop-gain, start-loss, stop-loss, splice donor, and splice acceptor variants were classified as LOF variants. Although frameshift indels are typically included in the LOF category, we considered them as a separate category in order to quantify their impact separately. The SIFT scores [52] for missense variants were retrieved from the Ensembl transcript database. Missense variants for which SIFT scores were available were then classified as deleterious when their SIFT score was less than 0.05 and as tolerated otherwise. We considered the predicted consequence types of LOF, frameshift and in-frame indels, and missense variants as putatively functional. To account for the regulatory role of promoters, we classified variants within 500 bp upstream of the annotated transcription start site in the same consequence type as the variants in the 5′ untranslated region (UTR) because both these regions likely contain regulatory elements that affect transcription and because the same variant can be in the promoter and in the 5′ UTR of different annotated transcripts for the same gene. As a result, 6.6% of the variants that were initially classified by Ensembl VEP as 'variants upstream of gene', were reclassified as 'variants in promoter regions'. For further analyses, variants in promoters and in the 5′ and 3′ UTR were jointly considered (Promoter + UTR). Because some variants, such as stop-gain (LOF) variants or frameshift indels, are more likely to be benign when located towards the end of the transcripts (e.g., [53]), we also analysed the relative position of these variants within transcripts (i.e., position accounting for transcript length).

Load of putatively functional alleles

We used the imputed genotypes to estimate the average number of alleles of each predicted consequence type and prevalence level that an individual carried. For the most common predicted consequence types, this was estimated from 50,000 randomly sampled variants. For tolerated missense variants, we used the 50,000 variants with the highest SIFT scores. To account for the different number of variants for each predicted consequence type and prevalence level category, we calculated the heterozygosity and homozygosity for the alternative allele for each individual as the percentage of variants of each category that the individual carried, respectively, in the heterozygous state and the homozygous state for the alternative allele.

Association to production traits

To further explore the association of variants in each predicted consequence type and prevalence level category with production traits, we performed genome-wide association studies (GWAS) for the three largest lines, using all the called variants that passed filtering. We chose average daily gain, backfat thickness, and loin depth because they are complex traits with moderate heritability estimates (from 0.21 to 0.38). The number of pigs with records that were included in the GWAS are in Table 1. Most pigs with records were born during the 2008–2020 period. Breeding values were estimated by line with a linear mixed model that included polygenic effects and the non-genetic effects of contemporary group, litter, and body weight, as relevant for each trait. Deregressed

estimated breeding values were obtained following the method of VanRaden et al. [54]. Only individuals for which the trait was directly measured were retained for the GWAS, by fitting the following univariate linear mixed model using the FastLMM software [55, 56]:

$$\mathbf{y} = \mathbf{x}_i\beta_i + \mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of deregressed estimated breeding values, $\mathbf{x}_i$ is the vector of genotypes for the ith variant, coded as 0 and 2 if homozygous for either allele or 1 if heterozygous, β_i is the allele substitution effect of the ith variant on the trait, $\mathbf{u} \sim N(0, \sigma^2_u\mathbf{K})$ is the vector of polygenic effects with the covariance matrix equal to the product of the polygenic additive genetic variance σ^2_u and the genomic relationship matrix $\mathbf{K}$, and $\mathbf{e}$ is a vector of uncorrelated residuals. Due to computational limitations, the genomic relationship matrix $\mathbf{K}$ was calculated using imputed genotypes for the high-density marker array and its single-value decomposition was taken.

We considered associations with a p-value equal or smaller than 10^{-6} as significant. We calculated an enrichment score for each predicted consequence type and prevalence level category as:

$$\log\left(\frac{\text{nSignCategory}/\text{nNotSignCategory}}{\text{nSignTotal}/\text{nNotSignTotal}}\right),$$

where nSignCategory is the number of variants with a significant association with at least one trait in one of the three lines for a given predicted consequence type and prevalence level category, nNotSignCategory is the number of variants with no significant association in the same category, and nSignTotal and nNotSignTotal are the total numbers of variants with and without significant association, respectively.

Linkage disequilibrium is pervasive between nearby significant variants due to the extremely high variant density of whole-genome sequence data. To account for this, we defined haplotype blocks and considered only one variant per haplotype block as the putative driver of an association that was detected in that region. We defined the haplotype blocks for each line separately using the—*blocks* function in Plink 1.9 [57, 58] by considering pairs of variants within 5 Mb of each other to be in strong linkage disequilibrium if the bottom of the 90% confidence interval of D' was greater than 0.7 and the top of the confidence interval was at least 0.9. If the top of the confidence interval was smaller than 0.7, it was considered as strong evidence for historical recombination between the two variants. All other pairs of variants were considered uninformative. Regions for which at least 90% of the informative pairs of variants showed strong linkage disequilibrium were defined as a haplotype block. Within each haplotype block, we selected the variant with the most severe predicted consequence type as the candidate variant, as a simplification of common assumptions in the prioritisation of candidate variants. If there was more than one variant with the same predicted consequence type, the one with the lowest p-value was selected. This process was performed separately for each trait and line.

We calculated the additive genetic variance explained by each variant as $2pq\widehat{\beta}^2$, where p and q were the allele frequencies and $\widehat{\beta}$ is the estimated allele substitution effect of the variant. We expressed the variance explained by each variant as a percentage of the phenotypic variance of each trait. Finally, we calculated the median F_{ST} of the candidate variants within each predicted consequence type and prevalence level category and compared it to the median F_{ST} of the same category as the logarithm of the ratio of the former to the latter.

Results

Prevalence of variants

A large percentage (21%) of the 46,344,624 biallelic variants that passed quality control criteria were widespread in all nine lines. Private variants represented a much smaller percentage (2 to 11%) of the variants called within each line. However, when counted across lines, private variants cumulatively predominated (28%) over the widespread ones. Most variants were neither private nor widespread. The distribution of these variants by line is shown in Table 2. Most variants (38,642,777) were SNPs, of which 10,595,681 were called in a single line (27%; 366,486 to 2,743,965 within each line) and 8,377,578 (22%) were called in all nine lines. The remaining 7,701,847 variants were indels, of which 2,436,674 were called in a single line (32%; 121,525 to 506,149 in each line) and 1,560,353 (20%) were called in all nine lines.

Distribution of variants and relationship with recombination rate

The number of variants by chromosome was strongly correlated with chromosome length ($r = 0.98$, $P < 0.05$) (see Additional file 3: Table S1). The variant density by chromosome was negatively correlated with chromosome length ($r = -0.87$, $P < 0.05$) and (see Additional file 3: Table S1). The variant density within 1-Mb non-overlapping windows was positively correlated with recombination rate in that window ($r = 0.65$, $P < 0.05$; Fig. 1a) [59]. For example, in line A, there was on average one variant every 81 bp, but in the 5% 1-Mb windows with the lowest and highest recombination rates there was on average one variant every 152 and 54 bp, respectively (2.8-fold more variants in windows with high recombination rate than in windows with low recombination rate). Across

Table 2 Number of variants by line

Line	Biallelic variant sites ($\times 10^6$)	SNPs			Indels		
		All biallelic ($\times 10^6$)	Private ($\times 10^6$)	Widespread ($\times 10^6$)	All biallelic ($\times 10^6$)	Private ($\times 10^6$)	Widespread ($\times 10^6$)
A	28.83	24.38	1.56	8.38	4.44	0.39	1.56
B	28.57	24.32	2.74	8.38	4.24	0.51	1.56
C	28.88	24.60	2.51	8.38	4.28	0.44	1.56
D	21.44	17.94	1.23	8.38	3.50	0.32	1.56
E	19.06	15.71	0.51	8.38	3.35	0.22	1.56
F	20.21	16.86	0.42	8.38	3.35	0.16	1.56
G	23.38	19.64	0.50	8.38	3.74	0.16	1.56
H	22.32	18.78	0.37	8.38	3.55	0.12	1.56
I	24.59	20.82	0.76	8.38	3.77	0.13	1.56
Total	46.30	38.64	10.60	8.38	7.70	2.44	1.56

all lines, there was one variant every 49 bp on average, but in the 5% 1-Mb windows with the lowest and highest recombination rates there was on average one variant every 79 and 34 bp, respectively (2.3-fold more variants in windows with high recombination rate).

The distribution of private and widespread variants along the genome also differed. The density of widespread variants was more correlated with recombination rate than the density of private variants (Fig. 1b and c). As a consequence, private variants represented a larger proportion of the variation in regions with low recombination rate, which were depleted of widespread variants. Across all lines, in the 5% 1-Mb windows with the highest recombination rates there was on average one private variant every 167 bp and one widespread variant every 148 bp (1.1-fold more private variants relative to widespread). In the 5% 1-Mb windows with the lowest recombination rates there was on average one private variant every 260 bp and one widespread variant every 531 bp (2.0-fold more private variants relative to widespread). There were no genomic regions that were enriched for private variants across the nine lines (see Additional file 4: Fig. S2).

Frequency and fixation index

The prevalence level and alternative allele frequency were related, in a way that less prevalent variants had a lower allele frequency (Fig. 2) and a lower F_{ST} (Fig. 3). Private variants had an average alternative allele frequency of 0.03 (SD = 0.09), as opposed to widespread variants, which had an average alternative allele frequency of 0.50 (SD = 0.25). Because the less prevalent variants generally had low alternative allele frequencies, they showed a small degree of differentiation between the lines in which they segregated ($F_{ST} = 0.04$, SD = 0.07). In contrast, widespread variants had the largest degree of differentiation between lines ($F_{ST} = 0.21$, SD = 0.11).

Prevalence and frequency of putatively functional variants

The predicted consequence types of the variants are listed in Table 3. Half (49.9%) of the variants were called in intergenic regions and another 47.0% of the variants were called in intronic regions. Only 2.2% of the variants were called in the promoter or 5′ and 3′ UTR. The coding variants comprised 0.9% of the total variants, of which more than half were missense (45.5%), frameshift indels (3.1%) or LOF (3.7%). The density of putatively functional variants was only weakly correlated with recombination rate in 1-Mb non-overlapping windows (Fig. 1d).

The low-prevalence variants (i.e., the variants that were identified in one or a few lines) were enriched for missense and LOF variants, as well as for potentially regulatory variants such as those located in the promoter and 5′ and 3′ UTR and other intronic variants. In contrast, the high-prevalence variants (i.e., the variants that were identified in many or all the lines) were enriched for frameshift indels and for synonymous (non-significant correlation) and intergenic variants. Although frameshift indels are typically included in the LOF category, our results show that the LOF category is very heterogeneous and the frameshift indels presented opposite patterns to other LOF variants. Therefore, we studied frameshift indels as a separate category.

Although the LOF variants had lower allele frequencies than the intergenic variants when they had low prevalence, they had similar allele frequencies in high-prevalence levels (Table 4). Thus, there was a set of LOF variants that were prevalent across lines and that also had

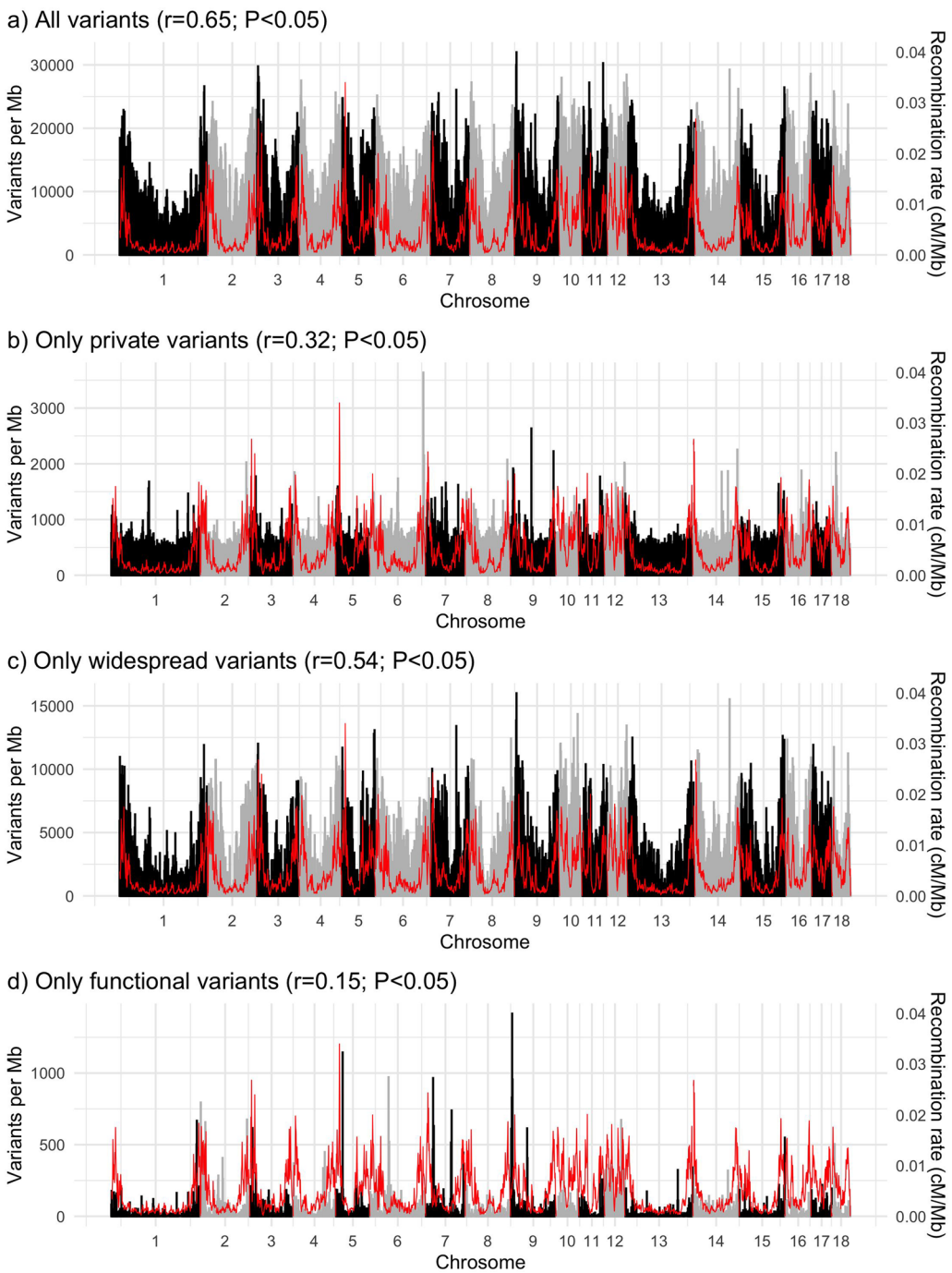

Fig. 1 Relationship of variant density in line A (black and grey bars) with recombination rate (red line). The correlation (r) between variant density and recombination rate in 1-Mb non-overlapping windows is reported

particularly high frequencies within lines. Missense variants had lower allele frequencies than the intergenic variants for all prevalence levels, especially those classified as deleterious. The low-prevalence missense variants were enriched for a larger fraction of deleterious variants and lower SIFT scores than high-prevalence missense variants (Fig. 4). Unlike missense or synonymous variants, low-prevalence stop-gain (LOF) variants and frameshift indels were more likely located towards the start of the transcripts (Fig. 5). In contrast to LOF and missense

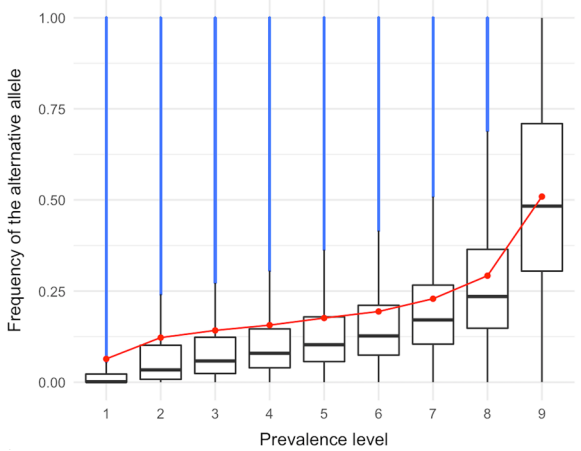

Fig. 2 Frequency of the alternative allele by prevalence level. Red dots indicate means. In blue, values greater than 1.5 times the interquartile range

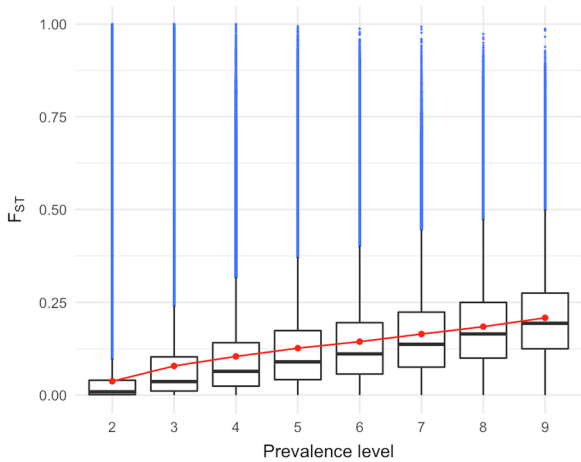

Fig. 3 Wright's fixation statistic (F_{ST}) by prevalence level. Red dots indicate means. In blue, values greater than 1.5 times the interquartile range

variants, frameshift and in-frame indels had intermediate allele frequencies, much higher than those of intergenic variants (Table 4), which indicated that the minor allele was the reference allele, in many cases. Within prevalence level, the LOF and deleterious missense variants had lower F_{ST} than the intergenic variants (Table 5), probably because they were kept at low allele frequencies due to negative selection pressure. Frameshift and in-frame indels also had lower F_{ST} than intergenic variants, in spite of their intermediate allele frequencies.

Load of putatively functional alleles by prevalence level

Most missense deleterious and LOF variants that an individual carried in the homozygous state for the alternative allele were high-prevalence variants. Only a small proportion of these variants were private. An individual carried on average 1048 (SD = 57) LOF variants in the homozygous state for the alternative allele, of which 713 (SD 36) were widespread across all nine lines and only 20 (SD = 7) were private. An average individual carried 1379 (SD = 165) deleterious missense variants in the homozygous state for the alternative allele, of which 1012 (SD = 79) were widespread and only 4 (SD = 3) were private. An average individual carried 1080 (SD = 89) LOF and 2632 (SD = 235) deleterious missense variants in the heterozygous state.

We found signals of negative selection against deleterious missense variants, in particular private ones. Individuals proportionally carried fewer deleterious missense variants in the homozygous state for the alternative allele than variants of other predicted consequence types, regardless of prevalence level (Fig. 6). Individuals also carried proportionally less private tolerated missense, synonymous and LOF variants in the homozygous state for the alternative allele than expected.

Associations of low-prevalence variants with production traits

Significant variants were enriched for putatively functional and regulatory variants of different prevalence levels, and depleted of intergenic variants. In total, 108,109 variants were significantly associated with at least one trait in one line. Figure 7a and b summarise the enrichment scores for all significant variants. The predicted consequence types that reached the greatest enrichment scores were LOF, frameshift indels, and unclassified missense variants, with various prevalence levels. Variants with intermediate prevalence levels were among the most enriched. These trends were accentuated when only considering candidate variants from haplotype blocks. In each line, we defined from 1554 to 2118 haplotype blocks. In total, across all lines and traits, 6692 candidate variants remained after accounting for linkage disequilibrium within each haplotype block. Figure 7c and d summarise the enrichment scores for the candidate variants. Enrichment scores based on the candidate variants revealed a stronger depletion of intergenic and intronic variants, and a much stronger enrichment for LOF, frameshift indels, and missense variants. For putatively functional variants, there were no clear trends of enrichment scores across prevalence levels. The trends of the enrichment scores between predicted consequence types and prevalence levels were similar for the three evaluated traits.

In general, the lower allele frequency of low-prevalence variants hindered the detection of significant associations for these variants. Low-prevalence variants that were significantly associated with the evaluated traits actually

Table 3 Predicted consequence types of variants by prevalence level

Consequence type	Percentage of variants (%) by prevalence level										r
	1	2	3	4	5	6	7	8	9	Total	
Loss-of-function[a]	**0.061**	**0.035**	**0.026**	**0.021**	**0.019**	**0.017**	**0.019**	**0.018**	**0.019**	**0.032**	**−.76**[b]
Splice acceptor/donor	0.038	0.023	0.014	0.010	0.009	0.007	0.008	0.008	0.008	0.018	−.79[b]
Stop-gain	0.014	0.009	0.008	0.008	0.007	0.007	0.007	0.006	0.006	0.009	**−.82**[b]
Stop-loss	0.005	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.003	−.36
Start-loss	0.004	0.002	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	−.47
Frameshift indel	**0.014**	**0.017**	**0.019**	**0.021**	**0.020**	**0.021**	**0.024**	**0.032**	**0.055**	**0.027**	**+.81**[b]
In-frame indel	**0.005**	**0.008**	**0.009**	**0.008**	**0.008**	**0.008**	**0.007**	**0.007**	**0.005**	**0.006**	**−.23**
Missense	0.556	0.378	0.355	0.340	0.344	0.336	0.319	0.306	0.325	0.393	−.73[b]
Deleterious	**0.201**	**0.092**	**0.074**	**0.069**	**0.064**	**0.062**	**0.054**	**0.048**	**0.040**	**0.096**	**−.78**[b]
Tolerated	**0.223**	**0.170**	**0.165**	**0.165**	**0.173**	**0.167**	**0.161**	**0.159**	**0.177**	**0.183**	**−.52**
Splice region	0.105	0.098	0.088	0.081	0.083	0.081	0.080	0.081	0.085	0.090	−.76[b]
Synonymous	**0.240**	**0.313**	**0.334**	**0.348**	**0.355**	**0.353**	**0.337**	**0.331**	**0.353**	**0.316**	**+.65**
Untranslated regions	**2.300**	**2.252**	**2.257**	**2.191**	**2.146**	**2.156**	**2.093**	**2.089**	**2.061**	**2.180**	**−.98**[b]
Promoter + 5′ UTR	0.879	0.825	0.812	0.812	0.787	0.813	0.759	0.766	0.759	0.810	−.90[b]
3′ UTR	1.421	1.427	1.445	1.378	1.359	1.343	1.334	1.322	1.302	1.370	−.94[b]
Non-coding transcript exon	0.104	0.113	0.107	0.113	0.128	0.118	0.105	0.109	0.117	0.111	+.25
Intronic	**47.744**	**47.571**	**47.634**	**47.162**	**46.513**	**46.709**	**46.701**	**46.355**	**46.132**	**46.981**	**−.95**[b]
Upstream of gene	3.062	3.066	3.075	3.041	3.083	3.056	2.929	2.943	2.936	3.015	−.81[b]
Downstream of gene	2.660	2.679	2.740	2.747	2.746	2.705	2.700	2.707	2.676	2.692	+.04
Intergenic	**43.148**	**43.468**	**43.355**	**43.927**	**44.553**	**44.439**	**44.687**	**45.021**	**45.235**	**44.154**	**+.97**[b]

The most severe consequence of each variant was used. The main Sequence Ontology (SO) terms are shown in order of severity (more severe to less severe) as estimated by Ensembl Variant Effect Predictor. The correlation (r) between the percentage of variants of each consequence type and prevalence is reported

In bold, categories that will be analysed in the next sections

[a] If frameshift indels were included in this category: r = −.06 (P > 0.05)

[b] Significant correlation (P < 0.05)

Table 4 Frequency of the alternative allele by predicted consequence type and prevalence level

Consequence type	Frequency of the alternative allele by prevalence level									
	1	2	3	4	5	6	7	8	9	Total
Loss-of-function	0.0010	0.017	0.048	0.062	0.089	0.114	0.151	0.223	0.489	0.020
Frameshift indel	0.4816	0.758	0.757	0.420	0.302	0.260	0.339	0.456	0.693	0.634
In-frame indel	0.8893	0.903	0.910	0.898	0.812	0.785	0.702	0.595	0.572	0.735
Deleterious missense	0.0006	0.018	0.043	0.061	0.078	0.092	0.125	0.170	0.350	0.010
Tolerated missense	0.0011	0.027	0.047	0.066	0.083	0.106	0.143	0.202	0.443	0.074
Synonymous	0.0037	0.032	0.049	0.066	0.086	0.107	0.151	0.205	0.447	0.110
Promoter + UTR	0.0019	0.034	0.059	0.078	0.099	0.122	0.166	0.226	0.475	0.102
Intronic	0.0015	0.035	0.059	0.080	0.102	0.126	0.171	0.235	0.485	0.110
Intergenic	0.0015	0.033	0.058	0.080	0.105	0.129	0.173	0.237	0.483	0.116

Values are medians

had intermediate allele frequencies that were greater than expected for their prevalence level. Low-prevalence variants in general explained low percentages of variance (Fig. 8), although some low-prevalence variants explained up to 3.2% of phenotypic variance. Significant variants had higher F_{ST} than other variants of the same predicted consequence type and prevalence level (Fig. 9). The enrichment of significant variants for higher F_{ST} was especially strong for low-prevalence variants, which in some instances reached F_{ST} of ~ 0.15.

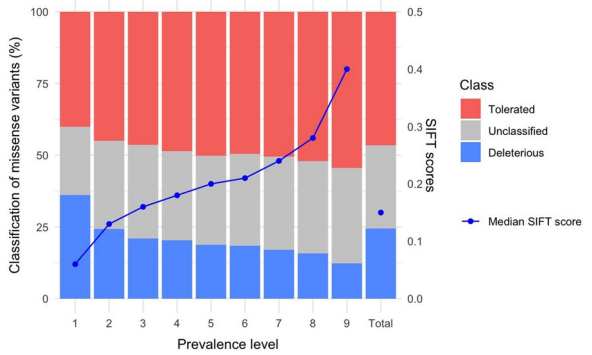

Fig. 4 Classification of the missense variants and median SIFT score by prevalence level

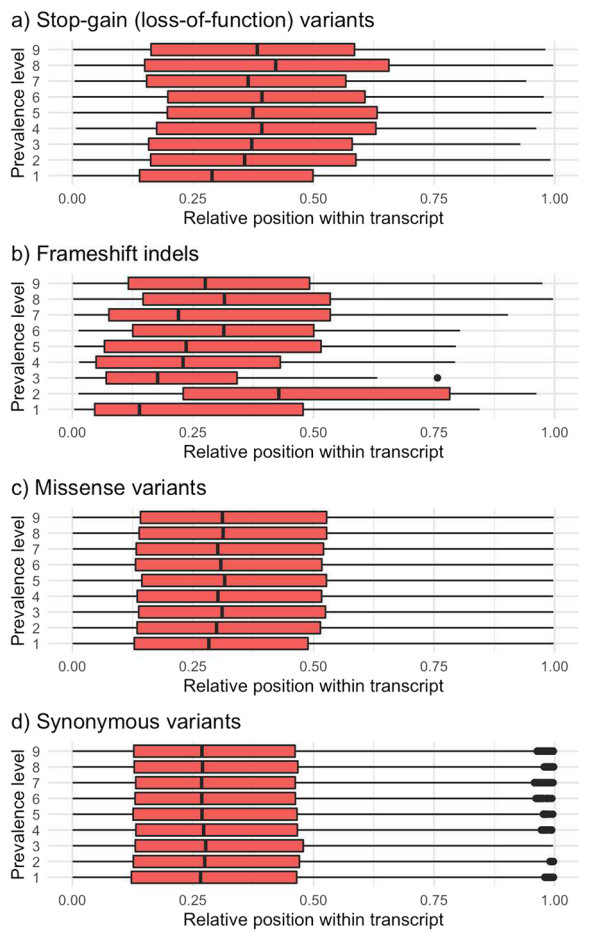

Fig. 5 Relative position within transcript of stop-gain, frameshift indels, missense, and synonymous variants, by prevalence level

Discussion

Our results contextualize the importance of population-specific and low-prevalence genetic variants. In the following, we will discuss: (1) the distribution and functional annotation of low-prevalence variants, (2) the load of putatively functional alleles by prevalence level, and (3) the association of low-prevalence variants with production traits.

Distribution and functional annotation of low-prevalence variants

The main difficulty for the study of low-prevalence genetic variants is that the prevalence of a variant across lines is strongly related to its allele frequency in the line, such that the low-prevalence variants are also rare within the lines in which they occur. This is possibly because low-prevalence variants are relatively recent or constrained by negative selection.

On the one hand, the density of private variants was less correlated with recombination rate than the density of widespread variants and, therefore, regions with a low recombination rate were enriched for private variants. Although the interplay between recurring sweeps, background selection, and other phenomena at play is not fully understood, it is generally accepted that selection on linked variants leads to loss of variation in regions with low recombination rates [60]. Our observation that regions with a low recombination rate were enriched for private variants suggests that private variants may have been less affected by selective sweeps than widespread variants. This is consistent with previous observations of the younger age of rare and low-prevalence variants [61] and suggests that private variants tend to have arisen more recently than widespread variants, likely after line differentiation, and accumulated in low-recombining regions due to the reduced efficacy of purifying selection in those regions [62, 63].

On the other hand, low-prevalence variants were enriched for putatively functional variants with signs of a greater severity (stop-gain and frameshift indels that occur earlier in the transcript, and missense variants that were predicted to be deleterious). Variants that affect performance traits or that cause a detrimental condition are under directional selection and are therefore driven towards loss or fixation [64, 65]. The low F_{ST} estimates for the low-prevalence variants indicated that selection pressure keeps these variants at low minor allele frequencies even when they occur in several lines, especially if they are putatively functional [66]. This could be caused by natural selection or by similar selection objectives across livestock populations. These observations are also consistent with previous reports that some putatively functional variant categories (such as stop-gain and deleterious missense) are enriched for variants that are private to single cattle breeds [33], that putatively functional variants are less likely to have a high frequency of the

Table 5 Wright's fixation statistic (F_{ST}) by predicted consequence type and prevalence level

Consequence type	F_{ST} by prevalence level								
	2	3	4	5	6	7	8	9	Total
Loss-of-function	0.003	0.022	0.047	0.066	0.094	0.114	0.145	0.171	0.071
Frameshift indel	0.010	0.042	0.065	0.081	0.088	0.120	0.146	0.148	0.114
In-frame indel	0.011	0.035	0.051	0.070	0.087	0.105	0.115	0.130	0.077
Deleterious missense	0.005	0.029	0.055	0.073	0.087	0.110	0.131	0.160	0.068
Tolerated missense	0.009	0.036	0.061	0.084	0.107	0.127	0.158	0.184	0.108
Synonymous	0.013	0.040	0.062	0.090	0.110	0.130	0.158	0.194	0.117
Promoter + UTR	0.009	0.036	0.060	0.086	0.108	0.131	0.158	0.190	0.110
Intronic	0.009	0.037	0.063	0.089	0.111	0.136	0.164	0.195	0.118
Intergenic	0.009	0.036	0.066	0.091	0.112	0.139	0.167	0.193	0.121

Values are medians

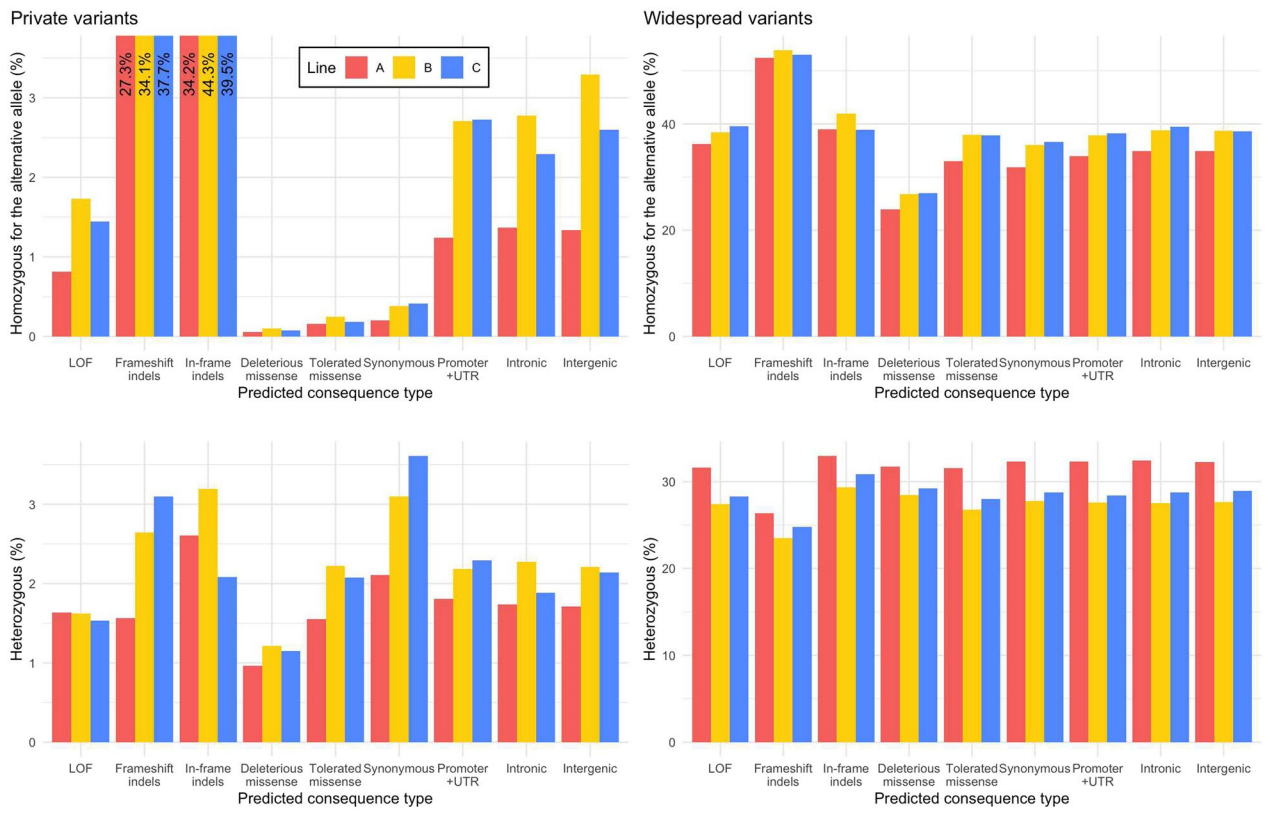

Fig. 6 Average percentage of variants in the homozygous state for the alternative allele or in the heterozygous state across individuals by predicted consequence type and prevalence level. LOF: loss-of-function; UTR: untranslated regions

alternative allele across multiple chicken lines [35], and that population-specific variants in non-African humans are enriched for putatively functional variants [67].

The relationship of variant prevalence across lines with allele frequency highlights the suitability of using a low-coverage sequencing approach to study this fraction of genetic variation. Nonetheless, bioinformatics pipelines for calling, genotyping, and even imputing such variants should account for the increased uncertainty because of their low allele frequency. We decided to use a very relaxed variant calling strategy with little filtering in order to account for as many rare variants

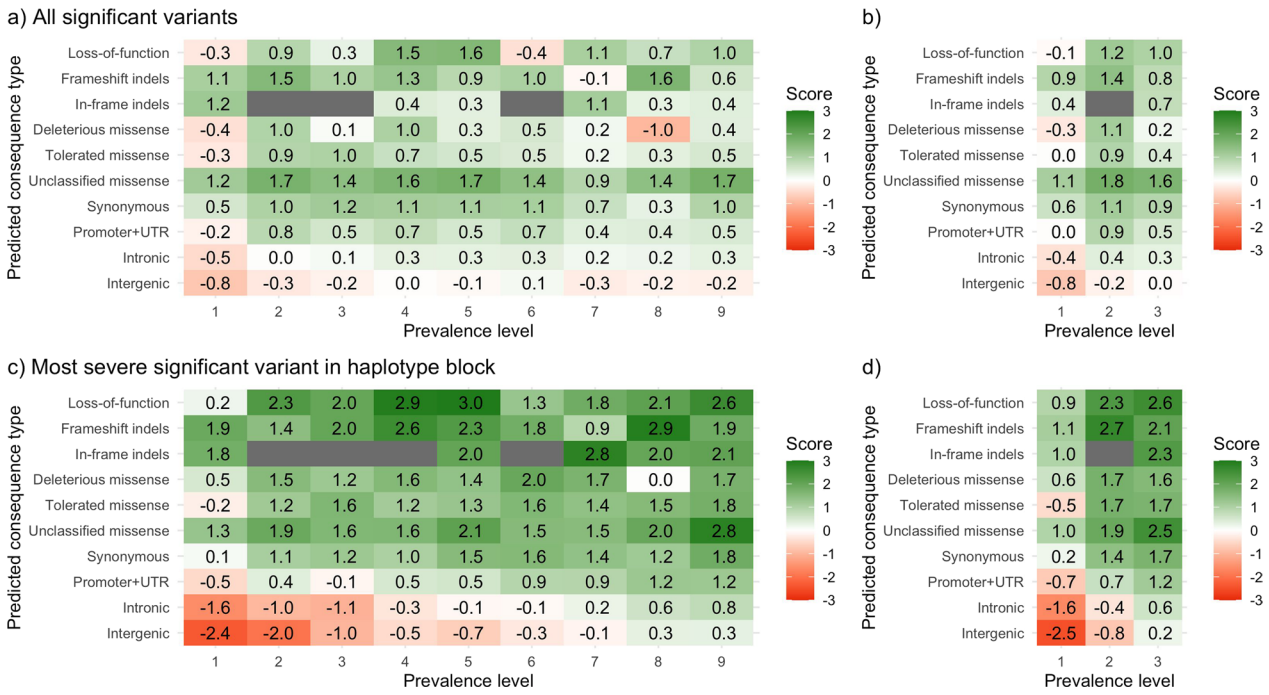

Fig. 7 Enrichment score for the number of significant variants in the genome-wide association study by variant prevalence level and predicted consequence type. Either all significant variants (**a**, **b**) or only the most severe significant variants within haplotype blocks (**c**, **d**) were used. Prevalence level was considered across all nine lines (**a**, **c**) or only across the three lines included in the genome-wide association study (**b**, **d**)

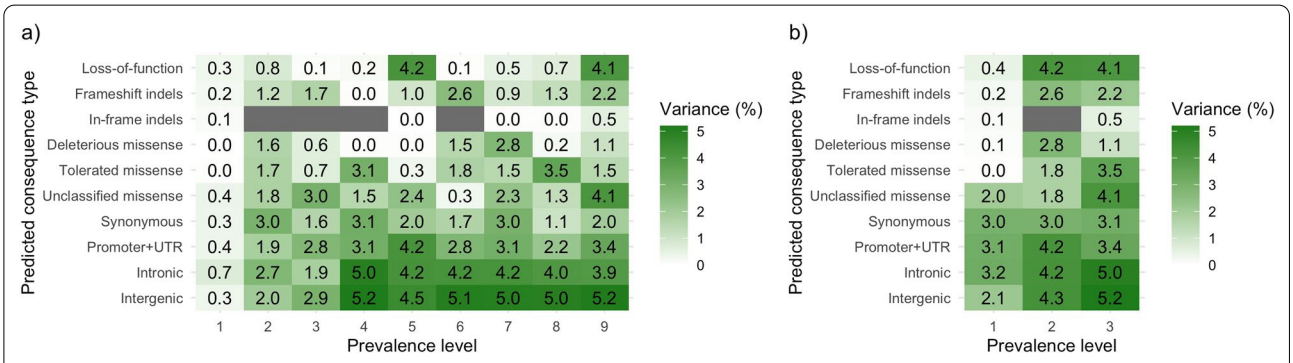

Fig. 8 Maximum percentage of phenotypic variance explained by individual candidate variants within each prevalence level and predicted consequence type. Only the candidate variants after accounting for linkage disequilibrium were used. Prevalence level was considered across all nine lines (**a**) or only across the three lines included in the genome-wide association study (**b**)

as possible, but a sizeable fraction of these rare variants were discarded after imputation because they were fixed for the imputed individuals that passed quality control. Low-coverage sequencing is also not suitable for other types of genetic variants, such as structural variations (copy number variants, tandem duplications, and inversions), which could also be putatively functional and population-specific [68]. Of course, the number of called variants and the proportion that were private or widespread depend on the number of sequenced lines [32, 35] as well as on the sequencing effort in each line.

Our results also suggest that what is typically grouped as LOF is actually a heterogeneous category. In particular, frameshift indels showed patterns that did not conform to the other predicted consequence types.

Load of putatively functional alleles by prevalence level

We found that an average individual carried a larger number of LOF and missense deleterious variants than

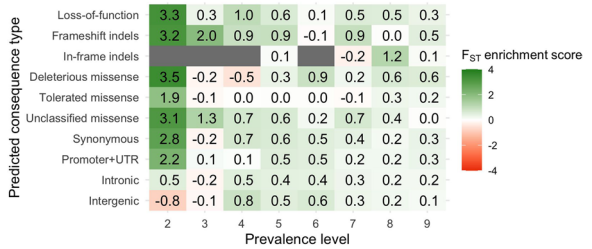

Fig. 9 Enrichment score for the F_{ST} median of the candidate variants within each prevalence level and predicted consequence type. Only the candidate variants after accounting for linkage disequilibrium were used. Prevalence level was considered across all nine lines

previously reported in other livestock species or in humans. However, to date there is no clear consensus on the number of LOF and deleterious missense alleles that are present in the genome of an average individual. In humans, it has been estimated that an average individual carries 100 to 150 LOF alleles [64, 69–71] and around 800 weakly deleterious mutations [72], most of which are rare. The average number of LOF and deleterious alleles carried by an individual has been reported to be larger in domestic livestock populations than in wild populations [73], including estimates of 100 to 300 deleterious variants in domestic pigs [74], over 400 deleterious variants in domestic chicken [74], and 1200 to 1500 deleterious variants in domestic yak [75]. Similar magnitudes have been reported in dogs [76], rice [77], and sunflower [63].

It has been debated why healthy individuals carry a larger number of LOF variants in the homozygous state than expected [78, 79]. One possible reason is that not all predicted LOF variants are detrimental and their functional impact should be validated before being considered as such. Many predicted LOF variants are in fact neutral, advantageous (either in the wild or in controlled production environments), or may be the result of sequencing and annotation errors [78]. The claim that not all predicted LOF variants are detrimental is supported by the large proportion of LOF observed in the homozygous state for the alternative allele compared to the other consequence types, which casts doubt on the real impact of those variants. In contrast, individuals carried a lower proportion of alleles predicted to be deleterious missense in the homozygous state, which supports that such variants may have a real impact on genetic variation of production traits and, therefore, be subject to selection pressure.

These observations have implications for the identification of variants to be used for genomic prediction or for genomic edition strategies, such as promotion of alleles by genome editing (PAGE) [26] or removal of alleles by genome editing (RAGE) [27]. Efforts to promote or remove alleles should target variants that make a substantial contribution to traits of interest, i.e. functional variants. However, it is hard to computationally predict and statistically estimate the effects of such variants, especially if they have a low allele frequency. The number of LOF variants in the homozygous state for the alternative allele suggests that predicted loss of function is not a good indicator that a variant is strongly deleterious in the context of livestock breeding. Similarly, bioinformatic predictors of missense variant effects appear to be not very accurate [80, 81]. High-throughput fine-mapping and variant screening would be needed to ascertain variant causality and disentangle causality from linkage disequilibrium.

Associations of low-prevalence variants with production traits

Genome-wide association studies for three polygenic traits of economic importance in the three largest lines revealed that the variants with significant associations were enriched for putatively functional roles, such as LOF, frameshift indels, and missense variants, and depleted of intergenic variants. This pattern of enrichment was similar to previous reports from human datasets [82]. However, only a few of the population-specific and low-prevalence variants were significantly associated with the traits, even after accounting for linkage disequilibrium. Most of the significant variants showed intermediate or high prevalence levels, which is consistent with previous meta-analyses in cattle that showed that significant variants are often common variants [83]. This could be because quantitative trait nucleotides have intermediate or high allele frequencies or because most studies are underpowered to map rare causal variants. The latter may be more likely given that the significant private and low-prevalence variants had intermediate allele frequencies. Although it cannot be ruled out that the significant low-prevalence variants reached intermediate allele frequencies by drift or by hitchhiking with linked variants under selection [84], it is plausible that these variants have biological functions that contribute to trait phenotypic variance. However, these variants amounted to a small number of variants that generally explained small fractions of variance.

Determining which of the variants that are in linkage disequilibrium is the most likely to be causal remains one of the greatest challenges in genomics. Here we prioritised the most severe variants within each haplotype block, which were more likely to have a low prevalence, as candidate variants. However, other more widespread variants, including intergenic variants, that were in high

linkage disequilibrium with the significant low-prevalence variants successfully acted as tag variants and captured much larger fractions of trait variance. This makes the widespread variants more suitable for applications in animal breeding and justifies their inclusion in tools such as marker arrays. A similar result was found in cattle, where splice site and synonymous variants explained the largest proportions of trait variance, while missense variants explained almost no variance [85]. It is worth pointing out that even a variant with a large allele substitution effect will explain a small percentage of variance if the minor allele is rare.

It is conceivable that some of the low-prevalence variants with a low allele frequency have non-negligible effects for traits of interest. In spite of the large number of individuals included in this study, the large number of variants investigated and the pervasiveness of linkage disequilibrium among them still make disentangling their contribution to trait variance very challenging. While genome-wide association studies that involve more than one breed typically find multiple breed-specific associations (e.g., [86]), based on our results it seems unlikely that breed-specific associations arise from the low-prevalence variants. Instead, breed-specific associations depend on the effect of the differences in allele frequencies, linkage disequilibrium structure, and other genetic background features on the power to detect the effect of prevalent variants across populations. Significant variants had higher F_{ST} estimates than non-significant variants, which is also consistent with previous reports [83]. Although the enrichment for higher F_{ST} was greater for low-prevalence variants, it remains unclear to which degree the significant low-prevalence variant with high F_{ST} explain differences among lines for the studied traits or their allele frequency reflect selection history.

Conclusions

Low-prevalence variants are enriched for putatively functional variants, including LOF and deleterious missense variants. However, most low-prevalence variants are kept at very low allele frequencies by negative selection or because they have arisen more recently than other higher-prevalence variants. Only a small subset of low-prevalence variants had intermediate allele frequencies and large estimated effects on production traits. Low-prevalence variants that were significantly associated with complex traits had greater degrees of differentiation between lines (per-site F_{ST}) than non-significant variants in the same category. However, more widespread variants, including intergenic variants, captured larger proportions of trait variance. Therefore, overall, accounting for population-specific and other low-prevalence variants is unlikely to noticeably benefit across-breed analyses, such as the prediction of genomic breeding values in a population using reference populations of a different genetic background.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-022-00732-8.

Additional file 1: Figure S1. Population structure of the sequenced pigs according to the two first principal components. The colour clusters correspond to lines A to I.

Additional file 2: Supplementary Methods. Complete description of the quality control criteria that were applied on the total number of variants called.

Additional file 3: Table S1. Number of analysed variants by chromosome.

Additional file 4: Figure S2. Variant density for the private variants in each line.

Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

Author contributions

RRF, MJ, and JMH designed the study; RRF and MJ performed the analyses; RRF and MJ wrote the first draft; BDV, CYC, WOH, GG, and JMH contributed to the interpretation of the results and provided comments on the manuscript. All authors read and approved the final manuscript.

Funding

The authors acknowledge the financial support from the BBSRC ISPG to The Roslin Institute (BBS/E/D/30002275), from Genus plc, Innovate UK (Grant 102271), and from Grant Numbers BB/N004736/1, BB/N015339/1, BB/L020467/1, and BB/M009254/1. MJ acknowledges financial support from the Swedish Research Council for Sustainable Development Formas Dnr 2016-01386. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any author accepted manuscript version arising from this submission.

Availability of data and materials

The software packages AlphaPhase, AlphaImpute, and AlphaPeel are available from https://github.com/AlphaGenes. The software package AlphaSeqOpt is available from the AlphaGenes website (http://www.alphagenes.roslin.ed.ac.uk). The datasets generated and analysed in this study are derived from the PIC breeding programme and not publicly available.

Declarations

Ethics approval and consent to participate

The samples used in this study were derived from the routine breeding activities of PIC.

Consent for publication

Not applicable.

Competing interests

BDV, CYC, and WOH are employed by Genus PIC. The remaining authors declare that the research was conducted in the absence of potential conflicts of interest.

Author details

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Easter Bush, Midlothian, Scotland, UK. [2]Departament de Ciència Animal, Universitat de Lleida - Agrotecnio-CERCA Center, Lleida, Spain. [3]The Pig Improvement Company, Genus plc, Hendersonville, TN, USA. [4]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Received: 31 January 2022 Accepted: 17 May 2022
Published online: 03 June 2022

References

1. Xiang R, van den Berg I, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci USA. 2019;116:19398–408.
2. Zhang F, Wang Y, Mukiibi R, Chen L, Vinsky M, Plastow G, et al. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits. BMC Genomics. 2020;21:36.
3. Wang Y, Zhang F, Mukiibi R, Chen L, Vinsky M, Plastow G, et al. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: carcass merit traits. BMC Genomics. 2020;21:38.
4. Pan Z, Yao Y, Yin H, Cai Z, Wang Y, Bai L, et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. Nat Commun. 2021;12:5848.
5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
6. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
7. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586:749–56.
8. Grobet L, Martin LJ, Poncelet D, Pirottin D, Brouwers B, Riquet J, et al. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat Genet. 1997;17:71–4.
9. Grisart B, Farnir F, Karim L, Cambisano N, Kim JJ, Kvasz A, et al. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci USA. 2004;101:2398–403.
10. Óvilo C, Fernández A, Noguera JL, Barragán C, Letón R, Rodríguez C, et al. Fine mapping of porcine chromosome 6 QTL and *LEPR* effects on body composition in multiple generations of an Iberian by Landrace intercross. Genet Res. 2005;85:57–67.
11. Zhao H, Qin Y, Xiao Z, Li Q, Yang N, Pan Z, et al. Loss of function of an RNA polymerase III subunit leads to impaired maize kernel development. Plant Physiol. 2020;184:359–73.
12. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human gene mutation database (HGMD®): 2003 update. Hum Mutat. 2003;21:577–81.
13. Drögemüller C, Tetens J, Sigurdsson S, Gentile A, Testoni S, Lindblad-Toh K, et al. Identification of the bovine Arachnomelia mutation by massively parallel sequencing implicates sulfite oxidase (SUOX) in bone development. PLoS Genet. 2010;6: e1001079.
14. Waide EH, Dekkers JCM, Ross JW, Rowland RRR, Wyatt CR, Ewen CL, et al. Not all SCID pigs are created equally: two independent mutations in the *Artemis* gene cause SCID in pigs. J Immunol. 2015;195:3171–9.
15. Derks MFL, Harlizius B, Lopes MS, Greijdanus-van der Putten SWM, Dibbits B, Laport K, et al. Detection of a frameshift deletion in the *SPTBN4* gene leads to prevention of severe myopathy and postnatal mortality in pigs. Front Genet. 2019;10:1226.
16. Matika O, Robledo D, Pong-Wong R, Bishop SC, Riggio V, Finlayson H, et al. Balancing selection at a premature stop mutation in the *myostatin* gene underlies a recessive leg weakness syndrome in pigs. PLoS Genet. 2019;15: e1007759.
17. Nicholas FW. Online Mendelian Inheritance in Animals (OMIA): a record of advances in animal genetics, freely available on the Internet for 25 years. Anim Genet. 2021;52:3–9.
18. Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, et al. Loss of function mutations in essential genes cause embryonic lethality in pigs. PLoS Genet. 2019;15: e1008055.
19. Mesbah-Uddin M, Hoze C, Michot P, Barbat A, Lefebvre R, Boussaha M, et al. A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle. J Dairy Sci. 2019;102:6340–56.
20. Ma J, Yang J, Zhou L, Ren J, Liu X, Zhang H, et al. A splice mutation in the *PHKG1* gene causes high glycogen content and low meat quality in pig skeletal muscle. PLoS Genet. 2014;10: e1004710.
21. Lunden A, Marklund S, Gustafsson V, Andersson L. A nonsense mutation in the *FMO3* gene underlies fishy off-flavor in cow's milk. Genome Res. 2002;12:1885–8.
22. Joseph SB, Hall DW. Spontaneous mutations in diploid *Saccharomyces cerevisiae*. Genetics. 2004;168:1817–25.
23. Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 2015;47:43.
24. MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics. 2016;17:144.
25. Lopez BIM, An N, Srikanth K, Lee S, Oh JD, Shin DH, et al. Genomic prediction based on SNP functional annotation using imputed whole-genome sequence data in Korean Hanwoo cattle. Front Genet. 2021;11: 603822.
26. Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, et al. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet Sel Evol. 2015;47:55.
27. Johnsson M, Gaynor RC, Jenko J, Gorjanc G, de Koning DJ, Hickey JM. Removal of alleles by genome editing (RAGE) against deleterious load. Genet Sel Evol. 2019;51:14.
28. Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 2011;21:952–60.
29. Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. Am J Hum Genet. 2021;108:656–68.
30. Molnár J, Nagy T, Stéger V, Tóth G, Marincs F, Barta E. Genome sequencing and analysis of Mangalica, a fatty local pig of Hungary. BMC Genomics. 2014;15:761.
31. Choi JW, Chung WH, Lee KT, Cho ES, Lee SW, Choi BH, et al. Whole-genome resequencing analyses of five pig breeds, including Korean wild and native, and three European origin breeds. DNA Res. 2015;22:259–67.
32. Cai Z, Sarup P, Ostersen T, Nielsen B, Fredholm M, Karlskov-Mortensen P, et al. Genomic diversity revealed by whole-genome sequencing in three Danish commercial pig breeds. J Anim Sci. 2020;98:skaa229.
33. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
34. Das A, Panitz F, Gregersen VR, Bendixen C, Holm LE. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 2015;16:1043.
35. Gheyas AA, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams JA, et al. Functional classification of 15 million SNPs detected from diverse chicken populations. DNA Res. 2015;22:205–17.
36. Gonen S, Ros-Freixedes R, Battagin M, Gorjanc G, Hickey JM. A method for the allocation of sequencing resources in genotyped livestock populations. Genet Sel Evol. 2017;49:47.
37. Ros-Freixedes R, Gonen S, Gorjanc G, Hickey JM. A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. Genet Sel Evol. 2017;49:78.
38. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JH. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011;43:12.
39. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44:9.
40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;1303.3997v1 [q – bio.GN].

42. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
43. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018. https://doi.org/10.1101/201178.
44. Ros-Freixedes R, Battagin M, Johnsson M, Gorjanc G, Mileham AJ, Rounsley SD, et al. Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. Genet Sel Evol. 2018;50:64.
45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
46. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
47. Whalen A, Ros-Freixedes R, Wilson DL, Gorjanc G, Hickey JM. Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. Genet Sel Evol. 2018;50:67.
48. Ros-Freixedes R, Whalen A, Chen CY, Gorjanc G, Herring WO, Mileham AJ, et al. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. Genet Sel Evol. 2020;52:17.
49. Ros-Freixedes R, Whalen A, Gorjanc G, Mileham AJ, Hickey JM. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. Genet Sel Evol. 2020;52:18.
50. Wright S. The genetical structure of populations. Ann Eugen. 1949;15:323–54.
51. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:122.
52. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.
53. Torella A, Zanobio M, Zeuli R, del Vecchio BF, Savarese M, Giugliano T, et al. The position of nonsense mutations can predict the phenotype severity: a survey on the *DMD* gene. PLoS One. 2020;15:e0237803.
54. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, et al. Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92:16–24.
55. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–5.
56. Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, et al. Further improvements to linear mixed models for genome-wide association studies. Sci Rep. 2015;4:6874.
57. Taliun D, Gamper J, Pattaro C. Efficient haplotype block recognition of very long and dense genetic sequences. BMC Bioinformatics. 2014;15:10.
58. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
59. Johnsson M, Whalen A, Ros-Freixedes R, Gorjanc G, Chen C-Y, Herring WO, et al. Genetic variation in recombination rate in the pig. Genet Sel Evol. 2021;53:54.
60. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013;14:262–74.
61. Mathieson I, McVean G. Demography and the age of rare variants. PLoS Genet. 2014;10: e1004528.
62. Charlesworth D, Morgan MT, Charlesworth B. Mutation accumulation in finite populations. J Hered. 1993;84:321–5.
63. Renaut S, Rieseberg LH. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other composite crops. Mol Biol Evol. 2015;32:2273–83.
64. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47:435–44.
65. Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, et al. Identification of a large set of rare complete human knockouts. Nat Genet. 2015;47:448–52.
66. Mezmouk S, Ross-Ibarra J. The pattern and distribution of deleterious mutations in maize. G3 (Bethesda). 2014;4:163–71.
67. Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. On the accumulation of deleterious mutations during range expansions. Mol Ecol. 2013;22:5972–82.
68. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. Genome Res. 2010;20:693–703.
69. The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
70. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8.
71. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
72. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.
73. Makino T, Rubin C-J, Carneiro M, Axelsson E, Andersson L, Webster MT. Elevated proportions of deleterious genetic variation in domestic animals and plants. Genome Biol Evol. 2018;10:276–90.
74. Bosse M, Megens HJ, Derks MFL, de Cara ÁMR, Groenen MAM. Deleterious alleles in the context of domestication, inbreeding, and selection. Evol Appl. 2019;12:6–17.
75. Xie X, Yang Y, Ren Q, Ding X, Bao P, Yan B, et al. Accumulation of deleterious mutations in the domestic yak genome. Anim Genet. 2018;49:384–92.
76. Cruz F, Vila C, Webster MT. The legacy of domestication: Accumulation of deleterious mutations in the dog genome. Mol Biol Evol. 2008;25:2331–6.
77. Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet. 2006;22:126–31.
78. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet. 2010;19:R125–30.
79. Rausell A, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, et al. Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. Proc Natl Acad Sci USA. 2020;117:13626–36.
80. Pagel KA, Pejaver V, Lin GN, Nam HJ, Mort M, Cooper DN, et al. When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinformatics. 2017;33:i389–98.
81. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11:5918.
82. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. 2013;9: e1003449.
83. van den Berg I, Xiang R, Jenko J, Pausch H, Boussaha M, Schrooten C, et al. Meta-analysis for milk fat and protein percentage using imputed sequence variant genotypes in 94,321 cattle from eight cattle breeds. Genet Sel Evol. 2020;52:37.
84. Chun S, Fay JC. Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet. 2011;7: e1002240.
85. Koufariotis LT, Chen YPP, Stothard P, Hayes BJ. Variance explained by whole genome sequence variants in coding and regulatory genome annotations for six dairy traits. BMC Genomics. 2018;19:237.
86. Purfield DC, Evans RD, Berry DP. Breed- and trait-specific associations define the genetic architecture of calving performance traits in cattle. J Anim Sci. 2020;98:skaa151.

Publisher's Note