OPEN ACCESS

Short report

A comparative medical genomics approach may facilitate the interpretation of rare missense variation

Bushra Haque,[1,2] George Guirguis,[1,2] Meredith Curtis,[1,2] Hera Mohsin,[2] Susan Walker,[3] Michelle M Morrow,[4] Gregory Costain [1,2,3,5,6]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/jmg-2023-109760).

[1]Program in Genetics and Genome Biology, SickKids Research Institute, Toronto, Ontario, Canada
[2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[3]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada
[4]GeneDx, Gaithersburg, Maryland, USA
[5]Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, Ontario, Canada
[6]Department of Paediatrics, University of Toronto, Toronto, Ontario, Canada

Correspondence to
Dr Gregory Costain, Program in Genetics and Genome Biology, SickKids Research Institute, Toronto, Canada; gregory.costain@sickkids.ca

BH and GG contributed equally.

Received 13 November 2023
Accepted 12 March 2024

To cite: Haque B, Guirguis G, Curtis M, *et al*. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmg-2023-109760

ABSTRACT

Purpose To determine the degree to which likely causal missense variants of single-locus traits in domesticated species have features suggestive of pathogenicity in a human genomic context.

Methods We extracted missense variants from the Online Mendelian Inheritance in Animals database for nine animals (cat, cattle, chicken, dog, goat, horse, pig, rabbit and sheep), mapped coordinates to the human reference genome and annotated variants using genome analysis tools. We also searched a private commercial laboratory database of genetic testing results from >400 000 individuals with suspected rare disorders.

Results Of 339 variants that were mappable to the same residue and gene in the human genome, 56 had been previously classified with respect to pathogenicity: 31 (55.4%) pathogenic/likely pathogenic, 1 (1.8%) benign/likely benign and 24 (42.9%) uncertain/other. The odds ratio for a pathogenic/likely pathogenic classification in ClinVar was 7.0 (95% CI 4.1 to 12.0, p<0.0001), compared with all other germline missense variants in these same 220 genes. The remaining 283 variants disproportionately had allele frequencies and REVEL scores that supported pathogenicity.

Conclusion Cross-species comparisons could facilitate the interpretation of rare missense variation. These results provide further support for comparative medical genomics approaches that connect big data initiatives in human and veterinary genetics.

INTRODUCTION

Genome-wide sequencing provides a near-comprehensive view of the exonic single nucleotide variants (SNVs) that contribute to a breadth of human diseases.[1] Distinguishing pathogenic missense variants that cause rare disease from missense variants that are benign or have no functional consequence remains a major challenge in human genetics.[2] In the companion field of veterinary medicine, genetic testing of economically important and domesticated animals for Mendelian traits and diseases is increasingly common.[3] Many first- (eg, SIFT, PolyPhen) and second-generation (eg, CADD, REVEL) in silico prediction tools for missense variants incorporate sequence homology and evolutionary conservation of the amino acid residue,[2 4] and a new wave of bioinformatic tools is learning from 'tolerated' variation in non-human primates and other animals (eg, PrimateAI-3D, EVE).[5 6] There is also a long-standing history of isolated observations that specific variants are pathogenic in both human and non-human orthologs.[7] However, the generalisability of a 'comparative medical genomics' approach that correlates disease associations of specific variants across species is unknown.

Modelled after Online Mendelian Inheritance in Man (OMIM; https://omim.org/), the Online Mendelian Inheritance in Animals (OMIA; https://omia.org) database includes 'likely causal variants' of 'single-locus traits' in non-human animal species that were expertly curated from the published literature.[3] OMIA listed 1577 likely causal variants from 485 species as of the end of 2022, with missense variants being the most common type of variation. Many of the 'traits' studied in a veterinary medicine context are severe phenotypes reminiscent of human genetic diseases (eg, connective tissue, haematological, skeletal or neurogenetic disorders; online supplemental figure 1). We hypothesised that leveraging these easily accessible but previously siloed OMIA data to identify additional evidence for pathogenicity could aid in the interpretation of human missense variants. The purpose of this study was to systematically assess the degree to which pathogenic missense variation observed in a non-human animal genome possesses features suggestive of pathogenicity in a human genome context.

METHODS

Identification of non-human animal variants and their human equivalents

We extracted all 442 missense SNVs from OMIA (accessed: January 2023) that were classified as likely causal for Mendelian phenotypes in one of nine animals (cat, cattle, chicken, dog, goat, horse, pig, rabbit and sheep) and where genomic coordinates were reported in OMIA or the original scientific publication reporting the variant. 400 variants were successfully mapped to the human reference genome (GRCh38) using University of California Santa Cruz's (UCSC's) LiftOver tool. We manually inspected each genomic region in the UCSC human genome browser to confirm that (1) the reference amino acid residue in the non-human animal species with the OMIA variant entry matched the reference amino acid residue in humans and (2) the alternate amino acid residue would be created through the specific nucleotide change reported in the non-human species; this resulted in the exclusion of an additional 61 variants.

Searching in public and private databases

We annotated the remaining 339 human genome variants using ANNOVAR[8] and custom R scripts, including for allele frequency in gnomAD (v3.1.2 and v2.1.1),[9] presence in ClinVar (download date: 29 August 2022),[10] phyloP conservation score,[11] REVEL score[12] and AlphaMissense score.[13] Variants were also searched in the Leiden Open Variation Database (LOVD v3.0)[14] and in the advanced literature database of Franklin by Genoox (franklin.genoox.com) in winter 2023. We reviewed the primary literature regarding each variant to determine whether the classification or reporting of the variant in one species was directly informed by findings in another species and, for a subset, to summarise the evidence type(s) contributing to a reported association between a specific missense variant and a phenotype in a non-human animal species. Last, we cross-referenced these 339 variants with a large private commercial genetic laboratory (GeneDx) database of genetic testing results from >400 000 individuals with suspected rare Mendelian disorders and their family members.

Statistical methods

Standard descriptive statistics, and parametric and non-parametric tests, were performed using R statistical software, V.4.1.0 (R Foundation for Statistical Computing) with two-tailed statistical significance set at $p<0.05$.

RESULTS

In total, 339 missense SNVs across 220 different genes, initially identified in non-human animal species from OMIA, were studied in humans (figure 1; online supplemental table 1). A majority (52.8%) were C→T or G→A transitions. There were no variants present in two or more of the non-human species. 159 (72.3%) of the genes were associated with germline Mendelian disease(s) in OMIM (online supplemental table 2). All but one variant were rare (minor allele frequency <0.001) in gnomAD v2.1.1, including 74.3% with an allele count of 0. The median and range of non-zero minor allele frequency were 1.2×10^{-5} and 4×10^{-6} to 0.078, respectively. Most variants (n=228; 67.3%) had REVEL scores ≥0.644 (supporting or greater evidence for

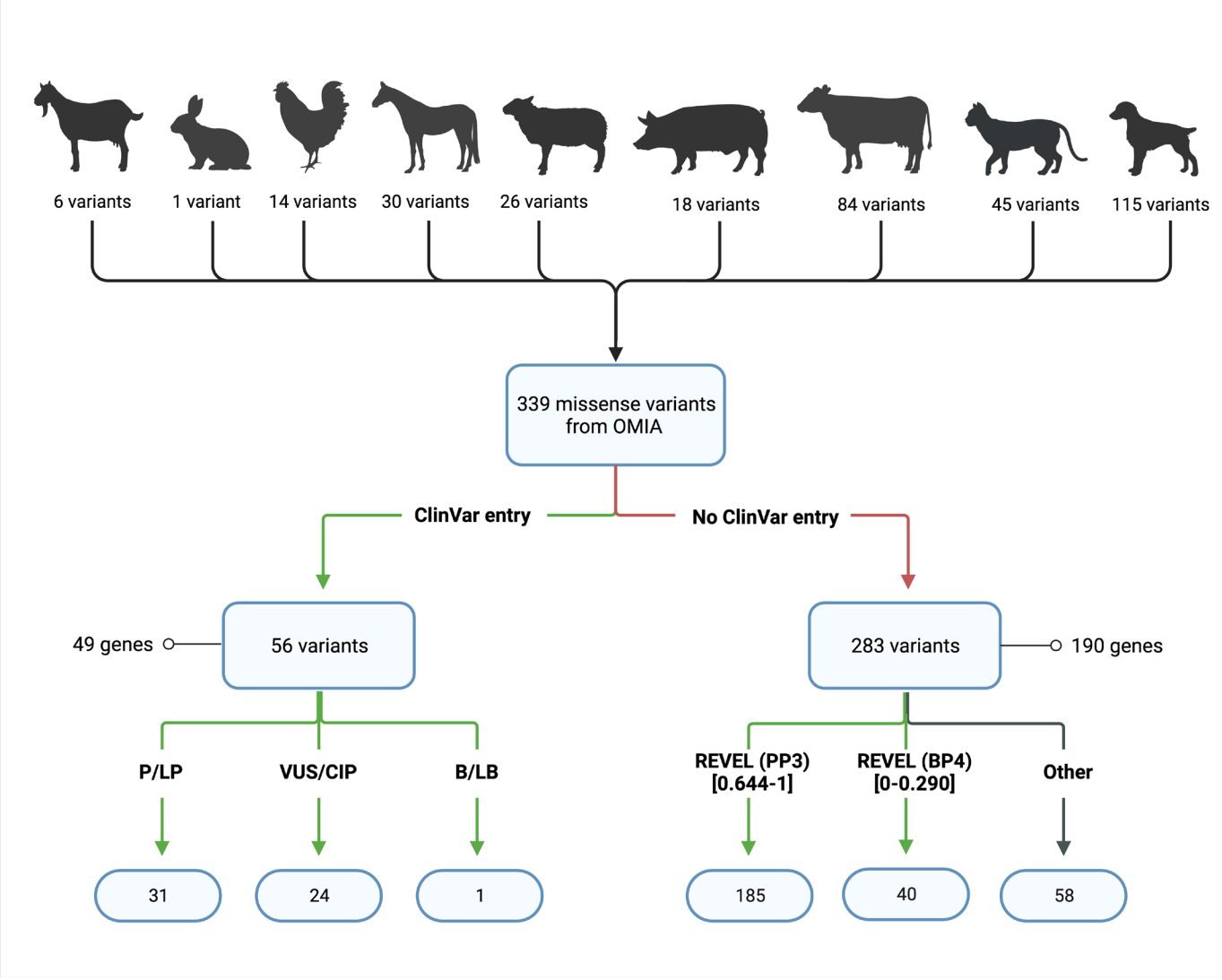

Figure 1 Missense variation from the OMIA database extrapolated to a human genome context. See main text for details. REVEL score thresholds for supporting evidence for pathogenicity (PP3) and for benign-ness (BP4) are from Pejaver *et al*.[15] B/LB, benign/likely benign; OMIA, Online Mendelian Inheritance in Animals; P/LP, pathogenic/likely pathogenic; VUS/CIP, variant of uncertain significance/conflicting interpretations of pathogenicity. Created with BioRender.com.

pathogenicity) and only 13.6% of variants had REVEL scores ≤0.290 (supporting or greater evidence for benign-ness).[15]

Non-human animal variants were often classified as pathogenic when seen in humans

Of the 339 variants, 56 variants in 49 genes had been previously seen in humans and classified with respect to pathogenicity in ClinVar: 31 (55.4%) as pathogenic/likely pathogenic (P/LP), 24 (42.9%) as variants of uncertain significance or with conflicting interpretations of pathogenicity (VUS/CIP) and 1 (1.8%) as benign/likely benign (B/LB) (figure 1). The human Mendelian disease phenotypes associated with these 49 genes were typically concordant with the phenotypes observed in the corresponding non-human animal species (online supplemental figures 1 and 2) and included both common (eg, classic Ehlers-Danlos syndrome) and rare (eg, geleophysic dysplasia) genetic diseases familiar to medical geneticists. There were no significant differences in median phyloP score between the P/LP variants (0.935; n=31) and the VUS/CIP/B/LB variants (0.935; n=25) (Mann-Whitney U p>0.05). Similarly, 23 of the variants were detected across a total of 172 different families that underwent testing at GeneDx: 15 were classified as P/LP (65%), 7 as VUS (30%) and 1 as B/LB (5%). No additional reports of these variants in humans were found by searching LOVD or in the advanced literature database of Franklin by Genoox. The odds ratio for these variants having a P/LP classification in ClinVar was 7.0 (95% CI 4.1 to 12.0, p<0.0001) when compared with all other germline missense variants with ClinVar entries in the 220 genes (n=45 925) (online supplemental figure 3). Determinations of a 'likely causal' genotype–phenotype association in a non-human animal species were informed by similar principles as are used in medical genetics practice (online supplemental figure 4), with an average of 2.75 categories of evidence applied in each report. We found direct references to non-human animal findings informing assessment of the human variant in only 10 cases and noted that the term 'OMIA' appears once (accession number VCV000016168.16) in the ClinVar database of 3 373 166 records (search date: June 2023). Conversely, we found references to an equivalent human variant in the initial publication and assessment of the non-human animal variant in 28 instances.

Understanding discordance between human and non-human variant consequences

We reviewed the variant classified in humans as B/LB, to look for plausible reason(s) for discordant classifications between humans and other animals. The B/LB variant in ClinVar is NM_002386.4(MC1R):c.274G>A;p.(Val92Met) (ClinVar Accession: VCV000014308.11), which has an allele frequency of >0.05 in many human populations. This variant in the melanocyte-stimulating hormone receptor gene is a well-studied functional polymorphism associated with pigmentation in humans and pigs.[16 17] The impact of the variant is therefore concordant across species, in line with our initial hypothesis. The B/LB variant in the GeneDx dataset is NM_004006.3(DMD):c.5869C>T;p.(Arg1957Trp) and was identified in the hemizygous state in an adult. The REVEL score is 0.22 and hemizygotes (n=3) are reported in the ExAC and ABraOM population databases.[18] Subsequent review of the original study in pigs[19] revealed that the variant was associated with 'stress syndrome' rather than a muscular dystrophy phenotype, and it remains possible that the variant is a risk factor for similar decompensation in humans.

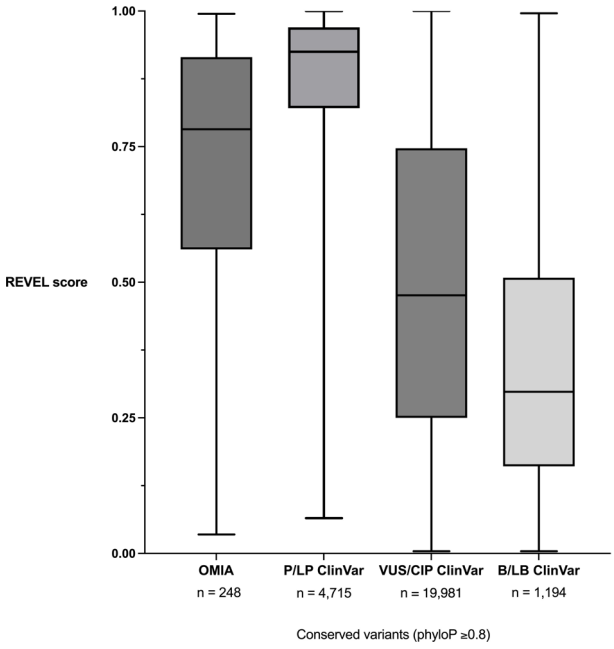

Figure 2 Boxplots of REVEL scores for the OMIA missense variants absent from ClinVar, comparing with other missense variation in ClinVar in the same gene set, restricting all variants to those with phyloP scores of ≥0.8. B/LB, benign/likely benign; OMIA, Online Mendelian Inheritance in Animals; P/LP, pathogenic/likely pathogenic; VUS/CIP, variant of uncertain significance/conflicting interpretations of pathogenicity.

In silico scores for the remaining variants were suggestive of pathogenicity in humans

The remaining 283 variants absent from ClinVar were found across 190 genes (figure 1). Compared with all missense SNVs in ClinVar in these same genes (n=33 526), the REVEL scores of the OMIA variants (median: 0.78; 69.8%≥0.644) more closely resembled P/LP variants (median: 0.93; 78.4%≥0.644) than B/LB variants (median: 0.30; 16.2%≥0.644). To account for potential confounding by residue conservation scores, we restricted to variants with high phyloP scores (≥0.8), with this cut-off corresponding to the top quartile of phyloP scores in the ClinVar VUS dataset. The remaining OMIA variants (n=248) continued to resemble P/LP variants (n=4715) more than B/LB variants (n=1194) with respect to REVEL scores (figure 2). Findings were similar using AlphaMissense (online supplemental figure 5).

DISCUSSION

Massive amounts of genomic data are being generated in human populations but also in domesticated animals like dogs and cattle. While comparative genomics is a well-established field,[20] and evolutionary conservation of amino acid residues is a foundational component of in silico tools for rare variant interpretation, the degree to which the impact of specific rare missense variants at shared residues is concordant across species ('comparative medical genomics') was understudied. Our results suggest that the missense variants in human genomes that correspond to likely causal variants of single-locus traits in other animals are more likely to be pathogenic (functionally significant). The presence of a variant in OMIA

suggests that additional published evidence from a non-human species exists that may be relevant to understanding the impact of the orthologous variant in humans. These observations have implications for rare variant interpretation in medical and veterinary medicine contexts.

While our focus in this report was on the application of non-human animal genomic data to interpreting variation in humans, we suspect that the association is bidirectional. Similarly, although there were no variants present in two or more of the non-human species, we hypothesise that missense variants at conserved residues across two or more species are likely to have a high degree of correlation with respect to functional consequences. There are also several limitations of this study. Our study relied on manual conversion of non-human animal variants from OMIA to their human equivalents, and we have not yet developed a method for automating this process nor a strategy for including OMIA variants in commonly used human genetics databases like ClinVar or LOVD. Consideration of this approach requires conservation of the gene and the residue. There are no consensus variant classification guidelines yet for veterinary medicine/non-human contexts, and the types and quality of evidence supporting a 'likely causal' attribution in OMIA will vary by entry. The possibility of circularities in the variant interpretation and reporting process (ie, that variants were deemed pathogenic in a non-human species and included in OMIA because of pre-existing findings in humans) cannot be excluded. The presence of an orthologous variant in OMIA is suspected but has not been shown to be independent of other lines of evidence typically used in human variant interpretation. All GeneDx variant interpretations are published in ClinVar, and so our results should not be overinterpreted as having replicated the findings in two independent cohorts. The added value of the data from GeneDx derives from the relative homogeneity in how evidence was applied for variant interpretation and the ability to look at a total number of variant observations/families potentially impacted by OMIA entries. We were underpowered to explore whether the strength of the association between the pathogenicity of OMIA variants and orthologous human variants varies by gene, residue change or nucleotide change.

Our results support the generalisability of a comparative medical genomics approach. The existence of potentially relevant information from a non-human species (as indicated by the variant appearing in OMIA) should prompt a dedicated review of the primary literature supporting the OMIA entry, in situations where additional evidence is needed to interpret the human orthologous variant. How best to incorporate this evidence into existing human variant interpretation frameworks remains unclear; we note potential overlap with the PS3 evidence code in the American College of Medical Genetics and Genomics/Association for Molecular Pathology guidelines. The benefits to human medical genetics of leveraging data from veterinary medicine could extend beyond establishing spontaneous large animal disease models for human conditions. Our work is intended to foster new collaborations that bridge previously siloed areas of Mendelian diagnostics, as the many lessons learnt from formalising human DNA variant classifications over the past decade could inform veterinary genetics practice. We anticipate that the coming decade will result in orders of magnitude more genome-wide sequencing in domesticated species, through large-scale coordinated academic projects, industry-sponsored research and direct-to-consumer pet testing. The degree to which the exponential growth of human and veterinary genomic datasets can be integrated and harnessed to improve variant interpretation across both contexts warrants additional study.

Acknowledgements The authors thank the many contributors to the public and private databases used in this study.

Contributors GC and SW conceptualised the study. BH, GG, MC, HM and MMM curated the data. BH and GG conducted a formal analysis of the data. BH and GC acquired the funding for the study. GC supervised the study. BH contributed to all visualisation of results. BH, GG and GC drafted the manuscript, and MC, HM, SW and MMM were given the opportunity to revise it critically for important intellectual content. All authors give final approval of the submitted version and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding The funding was provided by SickKids Research Institute, Canadian Institutes of Health Research and the University of Toronto McLaughlin Centre.

Competing interests SW is an employee of Genomics England Limited. MMM is an employee of GeneDx, LLC. The remaining authors have no potential conflicts of interest to declare.

Patient consent for publication Not applicable.

Ethics approval This secondary use data study was approved by the Research Ethics Board at the Hospital for Sick Children. The deidentified data from GeneDx was assessed in accordance with an IRB-approved protocol (WIRB #20171030).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are available in a public, open-access repository (OMIA; omia.org/) and/or are included in the article and uploaded as supplementary information.

ORCID iD

Gregory Costain http://orcid.org/0000-0003-0099-9945

REFERENCES

1 Costain G, Cohn RD, Scherer SW, *et al*. Genome sequencing as a diagnostic test. *CMAJ* 2021;193:E1626–9.
2 Costain G, Andrade DM. Third-generation computational approaches for genetic variant interpretation. *Brain* 2023;146:411–2.
3 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.
4 Spielmann M, Kircher M. Computational and experimental methods for classifying variants of unknown clinical significance. *Cold Spring Harb Mol Case Stud* 2022;8:a006196.
5 Frazer J, Notin P, Dias M, *et al*. Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;599:91–5.
6 Gao H, Hamp T, Ede J, *et al*. The landscape of tolerated genetic variation in humans and primates. *Science* 2023;380:6648.
7 Van Poucke M, Martlé V, Van Brantegem L, *et al*. A canine Orthologue of the human GFAP C.716G>A (P.Arg239His) variant causes Alexander disease in a Labrador retriever. *Eur J Hum Genet* 2016;24:852–6.
8 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164:16.:.
9 Karczewski KJ, Francioli LC, Tiao G, *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
10 Landrum MJ, Lee JM, Benson M, *et al*. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062–7.
11 Pollard KS, Hubisz MJ, Rosenbloom KR, *et al*. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;20:110–21.
12 Ioannidis NM, Rothstein JH, Pejaver V, *et al*. REVEL: an ensemble method for predicting the Pathogenicity of rare Missense variants. *Am J Hum Genet* 2016;99:877–85.

13 Cheng J, Novati G, Pan J, *et al*. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* 2023;381:6664.
14 Fokkema IFAC, Taschner PEM, Schaafsma GCP, *et al*. LOVD V.2.0: the next generation in gene variant databases. *Hum Mutat* 2011;32:557–63.
15 Pejaver V, Byrne AB, Feng B-J, *et al*. Calibration of computational tools for missense variant pathogenicity classification and Clingen recommendations for Pp3/Bp4 criteria. *Am J Hum Genet* 2022;109:2163–77.
16 Xu X, Thörnwall M, Lundin LG, *et al*. Val92Met variant of the melanocyte stimulating hormone receptor gene. *Nat Genet* 1996;14:384.
17 Kijas JM, Wales R, Tornsten A, *et al*. Melanocortin receptor 1 (Mc1R) mutations and coat color in pigs. *Genetics* 1998;150:1177–85.
18 Naslavsky MS, Scliar MO, Yamamoto GL, *et al*. Whole-genome sequencing of 1,171 elderly admixed individuals from Sao Paulo, Brazil. *Nat Commun* 2022;13:1004.
19 Nonneman DJ, Brown-Brandl T, Jones SA, *et al*. A defect in dystrophin causes a novel porcine stress syndrome. *BMC Genomics* 2012;13:233.
20 Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* 2020;587:240–5.