


Structural Variants Selected during Yak Domestication Inferred from Long-Read Whole-Genome Sequencing

Shangzhe Zhang ^{†,1}, Wenyu Liu,^{†,1} Xinfeng Liu,^{†,1} Xin Du,¹ Ke Zhang,¹ Yang Zhang,² Yongwu Song,³ Yunnan Zi,⁴ Qiang Qiu,¹ Johannes A. Lenstra,⁵ and Jianquan Liu*^{†,1}

¹State Key Laboratory of Grassland and Agro-ecosystem, Institute of Innovation Ecology and School of Life Science, Lanzhou University, Lanzhou, China

²The Supercomputing Center, Lanzhou University, Lanzhou, China

³Animal Disease Prevention and Control Center of Gangcha County, Haibei Tibetan Autonomous Prefecture, China

⁴Animal Husbandry Workstation of Xiahe County, Gannan Tibetan Autonomous Prefecture, China

⁵Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: liujq@nwipb.cas.cn.

Associate editor: Guang Yang

Abstract

Structural variants (SVs) represent an important genetic resource for both natural and artificial selection. Here we present a chromosome-scale reference genome for domestic yak (*Bos grunniens*) that has longer contigs and scaffolds (N50 44.72 and 114.39 Mb, respectively) than reported for any other ruminant genome. We further obtained long-read resequencing data for 6 wild and 23 domestic yaks and constructed a genetic SV map of 372,220 SVs that covers the geographic range of the yaks. The majority of the SVs contains repetitive sequences and several are in or near genes. By comparing SVs in domestic and wild yaks, we identified genes that are predominantly related to the nervous system, behavior, immunity, and reproduction and may have been targeted by artificial selection during yak domestication. These findings provide new insights in the domestication of animals living at high altitude and highlight the importance of SVs in animal domestication.

Key words: *Bos grunniens*, reference genome, domestication, structural variants.

The domestication of livestock species is one of the major achievements in the human civilization history. A series of phenotypic changes in domesticated animals, such as reduction of brain size and increased tameness, are considered to constitute the domestication syndrome (Hammer 1984). In several domestic species, the underlying genetic basis has been examined by using genetic markers such as single-nucleotide polymorphisms (SNPs), short insertions and deletions, and the copy number variations (CNVs) (Chen et al. 2009, 2018; Serres-Armero et al. 2017; Genova et al. 2018), which account for the most widespread mode of genomic variations. However, the role of structural variants (SVs), which comprise insertions, deletions, duplications, inversions, or translocations of 50 bp or longer (Baker 2012), has remained underexplored due to two technological constraints (Huddleston and Eichler 2016). First, detection of SVs needs long-read sequencing reads spanning over their full length (Chaisson et al. 2015; Sedlazeck et al. 2018). Second, it also requires a continuous reference assembly covering the repetitive fraction in genomes (Weckselblatt and

Rudd 2015; Peona et al. 2021). Long-read sequencing is not suitable for the detection of single-nucleotide variations because of a single-base error of 85–95% (Kono and Arakawa 2019), but it is the method-of-choice for detecting large SVs. Recently, long-read sequencing and a high-quality reference genome revealed significant roles of SVs during plant domestication (Fuentes et al. 2019; Zhou et al. 2019). Whole-genome sequencing (WGS) data based on a short-read assembly with a high coverage have been published for domestic and wild yaks (Qiu et al. 2012; Liu et al. 2020). In this study, we present a high-quality reference genome for domestic yak (BosGru3.0). By long-read resequencing of selected 29 wild and domestic yaks from genetic groups from 80 previous (Qiu et al. 2015) and 18 new short-read whole-genome sequences, we obtained a comprehensive and representative SV map for domestic and wild yaks, which allows a tentative identification of SV-related genes involved in the domestication syndrome.

For the chromosome-scale BosGru3.0 reference assembly (supplementary fig. S1, Supplementary Material online), DNA

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

was extracted from blood of a male domestic yak from Hongyuan County, Sichuan Province. We conducted a de novo assembly of Oxford Nanopore long reads with a $\sim 88\times$ coverage. Through being polished by Illumina short reads and clustering based on interaction strength from Hi-C data (supplementary fig. S2, Supplementary Material online), we obtained a highly continuous reference genome BosGru3.0, with 116 contigs assembled in 31 chromosomes. The contig and scaffold N50 of BosGru3.0 are 44.72 and 114.39 Mb, respectively, and these values are higher than obtained for other ruminant reference genomes (table 1 and supplementary table S1, Supplementary Material online). Repetitive elements (supplementary fig. S3 and table S2, Supplementary Material online), protein-coding genes (supplementary table S3, Supplementary Material online), and noncoding elements (supplementary table S4, Supplementary Material online) were predicted for our assembly. A total of 21,232 protein-coding genes were predicted (table 1 and supplementary fig. S1, Supplementary Material online, see more details about BosGru3.0 in Supplementary Material online).

Twenty-three domestic individuals covering various locations and six wild yaks (fig. 1A) were selected for long-read WGS resequencing after excluding duplicated samples (supplementary fig. S4 and table S6, Supplementary Material online). As shown by model-based clustering (fig. 1B and supplementary fig. S5, Supplementary Material online) and genetic distances of short-read whole-genome sequences of 18 yaks combined with previous data (supplementary tables S5, Supplementary Material online, Qiu et al. 2015), the 23 domestic yaks represent the genetic diversity within their distribution range. The average N50 length of the long-read WGS reaches 22.59 Kb (domestic) and 21.99 Kb (wild), with an effective depth of $8.4\times$ to $15.6\times$ (domestic) or $11.4\times$ to $21.2\times$ (wild, supplementary table S6, Supplementary Material online). We identified 372,220 SVs, which included 328,936 deletions, 32,618 insertions and 4,321 duplications, 1,993 inversions, and 4,352 translocations (supplementary figs. S6 and S7 and table S7, Supplementary Material online). We did not find any SV alleles that were absolutely specific for either wild or domestic yaks. We annotated all SVs by their positions on BosGru3.0 and found 257,155 SVs (69.09%) in intergenic regions, and 93,582, 14,964, 1,811, and 3,620 SVs were in intronic, exonic, UTR, or the 150 bp upstream and downstream flanking regions of genes, respectively (supplementary table S8, Supplementary Material online).

The majority of the SVs (74.43%) contains repetitive sequences. Overall percentages for different categories of these elements are not substantially different from the percentages for the whole genome (supplementary tables S2 and S9, Supplementary Material online), whereas the length distribution of SVs depends on the underlying molecular events (inversion, duplications, insertions, or deletions [supplementary fig. S7, Supplementary Material online]). Comparison of SV sequences and of the wild yak or domestic yak genomes did not display large differences in the contents of any type of repeats. However, wild yaks have more copies of LINE/RTE-BovB with a low divergence than domestic yaks

Table 1. Assembly statistics comparison between BosGru2.0 and BosGru3.0.

Assembly	BosGru2.0	BosGru3.0
Total length (bp)	2,645,161,911	2,832,776,395
Number of contigs	41,192	414
Contig N50 (Mb)	1.41	44.72
Scaffold N50 (Mb)	1.41	114.39
Chromosome number	0	31
Unplaced contig number	41,192	383
Number of gaps	192,002	646
GC content (%)	41.7	42.0
Protein-coding genes	20,499	21,232

(supplementary fig. S8, Supplementary Material online), which suggests a recent activity of RTE-BovB in wild yaks. Interestingly, length distribution of inversions and duplications sequences shows a peak at about 1,000 bp (supplementary fig. S7, Supplementary Material online), which mainly consists of non-repetitive elements and LINE-1 elements (supplementary table S10, Supplementary Material online).

In order to further identify SVs possibly involved in domestication, we calculated for all SVs the F_{ST} between wild and domestic yaks and found 3,680 SVs with F_{ST} outliers > 0.28 under artificial selection (supplementary table S11, Supplementary Material online). A tree of the yak genotypes with these SVs increases the separation of domestic and wild yaks relative to the tree of figure 1C, but still shows variation in the domestic yaks (supplementary fig. S9, Supplementary Material online). Among these high- F_{ST} SVs, 2,391 SVs are (0.64% of all SVs) in the intergenic and 1,288 SVs in the exons, introns, or flanking regions of 725 genes (supplementary table S11, Supplementary Material online). From these, 34 have deletions in exonic regions, 24 of which cause a frameshift in the open reading frame (ORF) (nonsense SVs).

We then annotated the functions of the 725 genes carrying high- F_{ST} SVs and found that the most significantly enriched function was involved in nervous system development (GO ID: 0007399, 168 genes) and human disease pathway, long-term depression (9 genes, KEGG accession: hsa04730, supplementary fig. S10 and tables S12 and S13, Supplementary Material online). Other GO function categories are related to the nervous system, including neuron differentiation, generation of neurons, and others. Typically, the variant with the second-highest F_{ST} was located in an intron of a signal protein *MAGI2* (fig. 1D). A deletion within the human *MAGI2* gene has been associated with epilepsy and schizophrenia (Marshall et al. 2008; Zhang et al. 2020) and several CNVs are located near *MAGI2* in an aggressive dog breed (Chen et al. 2009). Similar associations with behavior have been reported for three other high- F_{ST} SV genes. *GAD2* has been linked to fear in dog (Pendleton et al. 2018). *GAD2*-knockout mice displayed an increase in spontaneous seizures (Kash et al. 1997). *PLCB1* was identified to be associated with schizophrenia (Liu et al. 2005; Lo Vasco et al. 2013), with strong selection signals in domestication of buffalo (Luo et al. 2020) and rabbit (Carneiro et al. 2014). *GRIK2*, which is also related to fear, anxiety, and aggression, was involved in a selective

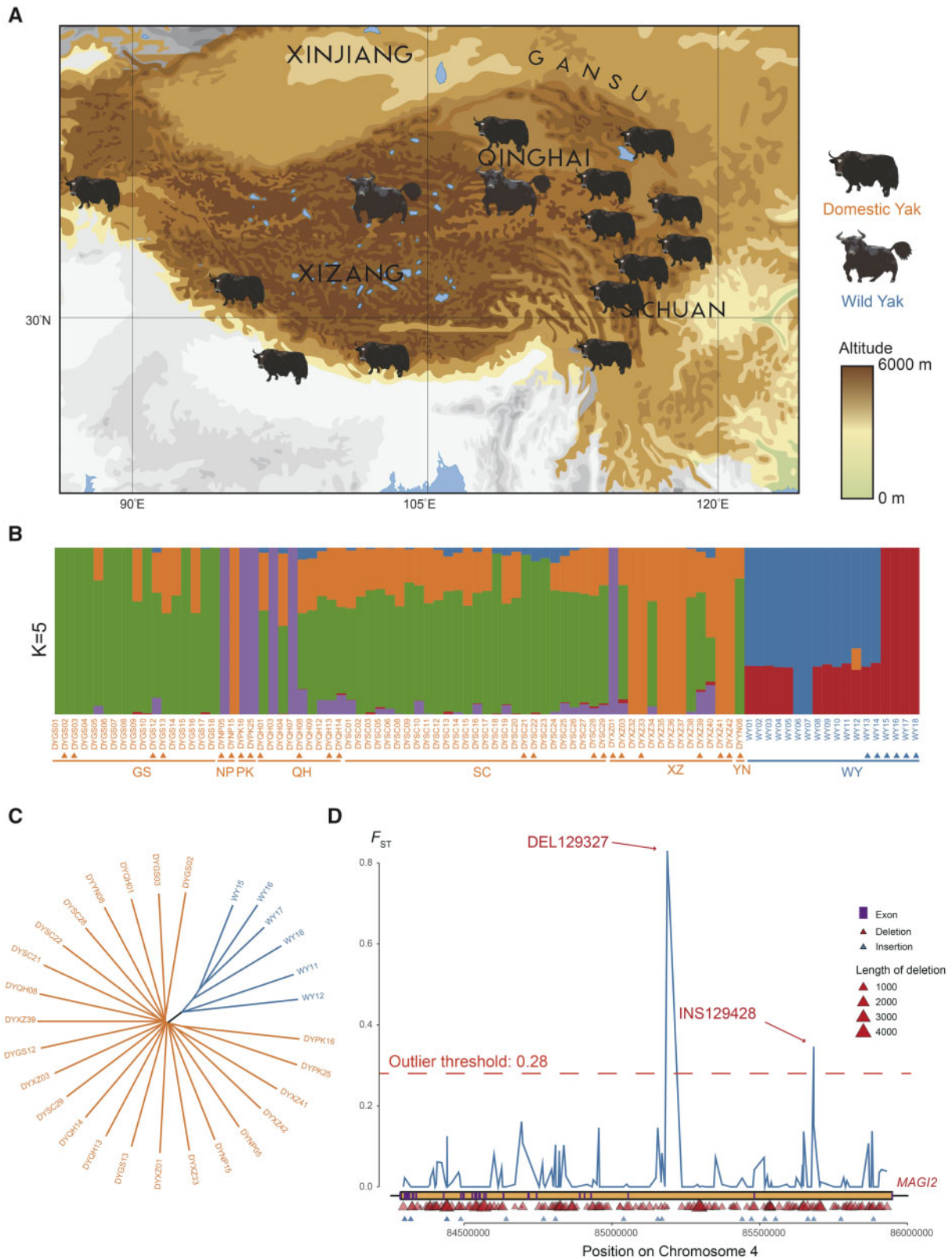


FIG. 1. (A) Geographic distribution of all domestic and wild yaks sampled in this research. (B) Genetic groups of 91 domestic and wild yaks in total based on short-read whole-genome sequences with population structure $K = 5$. Triangles indicate samples selected for long-read whole-genome sequencing. Orange: domestic yak; Blue: wild yak. GS, Gansu; NP, Nepal; PK, Pakistan; QH, Qinghai; SC, Sichuan; XZ, Xizang; YN, Yunnan; WY, Wild yak. (C) Neighbor-joining tree constructed based on SNPs of all long-read samples. (D) Domestication-related SVs in the region of *MAGI2*.

sweep in domesticated animals including rabbit, dog, and duck (O'Rourke and Boeckx 2020).

Other genes carrying SVs are involved in immunity, anatomical structure morphogenesis, and economical traits (supplementary table S12, Supplementary Material online). For example, *NAFT* has been proved to regulate the expression of potent immunomodulatory cytokines by downstream-targeting IL-2 growth factor in T cells (Müller and Rao 2010). *SMOC2* was reported to be related to brachycephaly in dogs (Marchant et al. 2017) and is highly expressed in endometrium and other reproductive tissues (Uhlén et al. 2015). *GSK3B* is an isoform of *GSK3A*, which was found to be related to fat storage ability in pig (Fu et al. 2016). Knockout of *GSK3A* in mice improved glucose tolerance in response to glucose load and elevated hepatic glycogen storage and insulin sensitivity (MacAulay et al. 2007). As for the nonsense SVs, a few genes are involved in mental or brain development as well, for instance, *PAX3* (Bang et al. 1997), *MAGT1* (Molinari et al. 2008), *SHROOM2* (Fairbank et al. 2006), and *SSBP3* (supplementary table S11, Supplementary Material online, Hashimoto et al. 2012).

Taken together, our results suggest that SVs have been mediated during yak domestication and that preferentially targeted genes are related to the nervous system, behavior, and immunity. These findings provide additional insights into yak domestications (Guo et al. 2006; Wang et al. 2010, 2011; Qiu et al. 2015; Zhang et al. 2016) and evolution of the bovine species (Wu et al. 2018; Zhang et al. 2020).

Materials and Methods

A detailed description of methods is provided in Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Data Availability

All sequences have been deposited to NCBI BioProject with accession number PRJNA540974. The BosGru3.0 reference sequences have been deposited to NCBI as GCA_005887515.2. Annotation information of BosGru3.0 and detailed SV information are available at Figshare as doi: 10.6084/m9.figshare.11151185. Custom workflow and scripts are available at <https://github.com/shangshanzhizhe/YakPopulationSV> (last accessed May 12, 2021).

Acknowledgments

This work is supported by the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), the national youth talent support program (Q.Q.), National Natural Science Foundation of China (31661143020, 41620104007), the Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2019), and International Collaboration 111 Programme (BP0719040). We received support for computational work from the Big Data Computing Platform for Western

Ecological Environment and Regional Development and Supercomputing Center of the Lanzhou University. We thank Dr. Yongfeng Zhou and Dr. Matthias H. Weissensteiner for helping on the method of SV detection, and Dr. Tserang-Donko Mipam for the sample collection.

Author Contributions

J.L. designed this research. W.L., X.D., Y.S., and Y.Z. collected samples. S.Z., W.L., and X.L. analyzed the data. K.Z. and X.D. helped with methods integration. Y.Z. helped managing the high-performance cluster. S.Z. and J.L. wrote the manuscript with inputs from all authors. Q.Q. and J.A.L. helped with revision of the manuscript.

References

- Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods*. 9(2):133–137.
- Bang AG, Papalopulu N, Kintner C, Goulding MD. 1997. Expression of Pax-3 is initiated in the early neural plate by posteriorizing signals produced by the organizer and by posterior non-axial mesoderm. *Development* 124:2075–2085.
- Carneiro M, Rubin CJ, Palma FD, Albert FW, Alföldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Maier JT, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345(6200):1074–1079.
- Chaisson MJ, Wilson RK, Eichler EE. 2015. Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet*. 16:627–640.
- Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.
- Chen W-K, Swartz JD, Rush LJ, Alvarez CE. 2009. Mapping DNA structural variation in dogs. *Genome Res*. 19:500–509.
- Fairbank PD, Lee C, Ellis A, Hildebrand JD, Gross JM, Wallingford JB. 2006. Shroom2 (APXL) regulates melanosome biogenesis and localization in the retinal pigment epithelium. *Development* 133:4109–4118.
- Fu Y, Li C, Tang Q, Tian S, Jin L, Chen J, Li M, Li C. 2016. Genomic analysis reveals selection in Chinese native black pig. *Sci Rep*. 6:36354.
- Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, et al. 2019. Structural variants in 3000 rice genomes. *Genome Res*. 29(5):870–880.
- Genova F, Longeri M, Lyons LA, Bagnato A, Gandolfi B, Aberdein D, Alves PC, Barsh GS, Beale HC, Bergström TF, et al.; the 99Lives Consortium. 2018. First genome-wide CNV mapping in FELIS CATUS using next generation sequencing data. *BMC Genomics*. 19(1):895.
- Guo S, Savolainen P, Su J, Zhang Q, Qi D, Zhou J, Zhong Y, Zhao X, Liu J. 2006. Origin of mitochondrial DNA diversity of domestic yaks. *BMC Evol Biol*. 6:73.
- Hammer K. 1984. Das domestikationssyndrom. *Die Kulturpflanze*. 32(1):11–34.
- Hashimoto Y, Muramatsu K, Kunii M, Yoshimura S, Yamada M, Sato T, Ishida Y, Harada R, Harada A. 2012. Uncovering genes required for neuronal morphology by morphology-based gene trap screening with a revertible retrovirus vector. *FASEB J*. 26:4662–4674.
- Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics* 202:1251–1254.
- Kash SF, Johnson RS, Tecott LH, Noebels JL, Mayfield RD, Hanahan D, Baekkeskov S. 1997. Epilepsy in mice deficient in the 65-kDa isoform of glutamic acid decarboxylase. *Proc Natl Acad Sci USA*. 94(25):14060–14065.
- Kono N, Arakawa K. 2019. Nanopore sequencing: review of potential applications in functional genomics. *Develop Growth Differ*. 61(5):316–326.
- Liu F, Ruiz MS, Austin DA, Webster NJ. 2005. Constitutively active Gq impairs gonadotropin-releasing hormone-induced intracellular

- signaling and luteinizing hormone secretion in LbetaT2 cells. *Mol Endocrinol.* 19:2074–2085.
- Liu Y, Luo J, Dou J, Yan B, Ren Q, Tang B, Wang K, Qiu Q. 2020. The sequence and de novo assembly of the wild yak genome. *Sci Data.* 7(1):8.
- Lo Vasco VR, Longo L, Polonia P. 2013. Phosphoinositide-specific Phospholipase C β 1 gene deletion in bipolar disorder affected patient. *J Cell Commun Signal.* 7:25–29.
- Luo X, Zhou Y, Zhang B, Zhang Y, Wang X, Feng T, Li Z, Cui K, Wang Z, Luo C, et al. 2020. Understanding divergent domestication traits from the whole-genome sequencing of swamp- and river-buffalo populations. *Natl Sci Rev.* 7(3):686–701.
- MacAulay K, Doble BW, Patel S, Hansotia T, Sinclair EM, Drucker DJ, Nagy A, Woodgett JR. 2007. Glycogen synthase kinase 3 α -specific regulation of murine hepatic glycogen metabolism. *Cell Metab.* 6(4):329–337.
- Marchant TW, Johnson EJ, McTeir L, Johnson CI, Gow A, Liuti T, Kuehn D, Svenson K, Birmingham ML, Drögemüller M, et al. 2017. Canine brachycephaly is associated with a retrotransposon-mediated mis-splicing of SMOC2. *Curr Biol.* 27(11):1573–1584.e6.
- Marshall CR, Young EJ, Pani AM, Freckmann M-L, Lacassie Y, Howald C, Fitzgerald KK, Peippo M, Morris CA, Shane K. 2008. Infantile spasms is associated with deletion of the MAGI2 gene on chromosome 7q11.23-q21.11. *Am J Hum Genet.* 83:106–111.
- Molinari F, Foulquier F, Tarpey PS, Morelle W, Boissel S, Teague J, Edkins S, Futreal PA, Stratton MR, Turner G, et al. 2008. Oligosaccharyltransferase-subunit mutations in nonsyndromic mental retardation. *Am J Hum Genet.* 82:1150–1157.
- Müller MR, Rao A. 2010. NFAT, immunity and cancer: a transcription factor comes of age. *Nat Rev Immunol.* 10(9):645–656.
- O'Rourke T, Boeckx C. 2020. Glutamate receptors in domestication and modern human evolution. *Neurosci Biobehav Rev.* 108:341–357.
- Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* 16(1):64.
- Peona V, Blom M, Xu L, Burri R, Sullivan S, Bunikis I, Liachko I, Haryoko T, Jönsson KA, Zhou Q, et al. 2021. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour.* 21(1):263–286.
- Qiu Q, Wang L, Wang K, Yang Y, Ma T, Wang Z, Zhang X, Ni Z, Hou F, Long R, et al. 2015. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. *Nat Commun.* 6:10283.
- Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM. 2012. The yak genome and adaptation to life at high altitude. *Nat Genet.* 44:946–949.
- Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet.* 19:329–346.
- Serres-Armero A, Povolotskaya IS, Quilez J, Ramirez O, Santpere G, Kuderna LFK, Hernandez-Rodriguez J, Fernandez-Callejo M, Gomez-Sanchez D, Freedman AH, et al. 2017. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genomics.* 18:977.
- Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science.* 347(6220):1260419.
- Wang Z, Shen X, Liu B, Su J, Yonezawa T, Yu Y, Guo S, Ho SYW, Vilà C, Hasegawa M, et al. 2010. Phylogeographical analyses of domestic and wild yaks based on mitochondrial DNA: new data and reappraisal. *J Biogeogr.* 37(12):2332–2344.
- Wang Z, Yonezawa T, Liu B, Ma T, Shen X, Su J, Guo S, Hasegawa M, Liu J. 2011. Domestication relaxed selective constraints on the yak mitochondrial genome. *Mol Biol Evol.* 28:1553–1556.
- Weckselblatt B, Rudd MK. 2015. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet.* 31:587–599.
- Wu D, Ding X, Wang S, Wojcik JM, Zhang Y, Tokarska M, Li Y, Wang M, Faruque O, Nielsen R, et al. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat Ecol Evol.* 2:1139–1145.
- Zhang K, Lenstra JA, Zhang S, Liu W, Liu J. 2020. Evolution and domestication of the bovini species. *Animal Genet.* 51:637–657.
- Zhang X, Wang K, Wang L, Yang Y, Ni Z, Xie X, Shao X, Han J, Wan D, Qiu Q. 2016. Genome-wide patterns of copy number variation in the Chinese yak genome. *BMC Genomics.* 17:379.
- Zhang Y, You X, Li S, Long Q, Zhu Y, Teng Z, Zeng Y. 2020. Peripheral blood leukocyte RNA-Seq identifies a set of genes related to abnormal psychomotor behavior characteristics in patients with schizophrenia. *Med Sci Monit.* 26:e922426.
- Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, Cantu D, Gaut BS. 2019. The population genetics of structural variants in grapevine domestication. *Nat Plants.* 5:965–979.