Genome-Wide Detection of Structural Variations Reveals New Regions Associated with Domestication in Small Ruminants

Tristan Cumer *, Frédéric Boyer, and François Pompanon

Université Grenoble Alpes, Université Savoie Mont-Blanc, CNRS, LECA, Grenoble, France

*Corresponding author: E-mail: t.cumer.sci@gmail.com.

Accepted: 8 July 2021

Abstract

During domestication processes, changes in selective pressures induce multiple phenotypical, physiological, and behavioral changes in target species. The rise of next-generation sequencing has provided a chance to study the genetics bases of these changes, most of the time based on single nucleotide polymorphisms (SNPs). However, several studies have highlighted the impact of structural variations (SVs) on individual fitness, particularly in domestic species. We aimed at unraveling the role of SVs during the domestication and later improvement of small ruminants by analyzing whole-genome sequences of 40 domestic sheep and 11 of their close wild relatives (*Ovis orientalis*), and 40 goats and 18 of their close wild relatives (*Capra aegagrus*). Using a combination of detection tools, we called 45,796 SVs in *Ovis* and 15,047 SVs in *Capra* genomes, including insertions, deletions, inversions, copy number variations, and chromosomal translocations. Most of these SVs were previously unreported in small ruminants. 69 and 45 SVs in sheep and goats, respectively, were in genomic regions with neighboring SNPs highly differentiated between wilds and domestics (i.e., putatively related to domestication). Among them, 25 and 20 SVs were close to or overlapping with genes related to physiological and morpho-anatomical traits linked with productivity (e.g., size, meat or milk quality, wool color), reproduction, or immunity. Finally, several of the SVs differentiated between wilds and domestics would not have been detected by screening only the differentiation of SNPs surrounding them, highlighting the complementarity of SVs and SNPs based approaches to detect signatures of selection.

Key words: domestication, goat, population genomics, sheep, SNP, structural variations.

Significance

Despite their important role in microevolutionary processes, genomic structural variations (SVs, mutations affecting more than 50 bp) are often ignored in genome-wide studies, usually based on single-nucleotide polymorphisms (SNPs). In this work, we took advantage of whole-genome data from wild and domestic sheep and goats, to create the first atlas of all types of SVs in small ruminants' genomes. Based on this atlas, we identified SVs that are highly differentiated between wilds and domestics, pinpointing new regions associated with domestication. Interestingly, some of these regions do not arbor high differentiation on neighboring SNPs, showcasing the value of looking for selection signatures directly on SVs. More generally, these results highlight the relevance of including all types of mutations in genome-wide studies, to fully characterize the impact of selection upon genomes.

Introduction

Plant and animal domestication represents a crucial step in human history, enabling the transition from hunting and gathering to farming (Vigne 2011). During domestication, humans exerted new selective pressures on domestic animals, inducing phenotypical changes to reach different productive, adaptive, and behavioral traits (Wright 2015). With the rise of next-generation sequencing giving an easier access to whole-genome sequences (WGS), genetic approaches have gained much power to identify genes targeted during early domestication (e.g., domestication syndrome-related characters) and later improvements (e.g., meat or milk production)

© 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

(Wiener and Wilkinson 2011). This is, for instance, the case of the dog's ability to digest a starch-rich diet associated with the *AMY2B* gene (Axelsson et al. 2013) or sheep polledness associated with a particular haplotype of the region including the *RXFP2* gene (Kijas et al. 2012).

Such studies mainly focused on single-nucleotide polymorphism (SNPs) data (Orozco-terWengel et al. 2015; Frantz et al. 2016), although it is known that SNPs only account for a part of the genetic polymorphism (Pang et al. 2010). For example, in humans, SNPs affect only 0.01% of the genome, whereas structural variants (SVs), including small indels, affect more than 1.2% of the genome (Pang et al. 2010). Even if most of these variants are assumed to be neutral, they may have a huge impact on fitness (Feuk et al. 2006). In addition to gene interactions, epigenetic factors, or rare variants, SVs are thought to be a cause for missing heritability when searching for genetic variations accounting for phenotypes (Eichler et al. 2010). SVs are defined as insertions, deletions, copy number variations (CNVs), inversions, and inter- or intrachromosomal rearrangements (Tattini et al. 2015). They have been the target of many selection events during domestication, including ones targeting traits related to behavior, morphology, production, and reproduction. For example, in pigs, CNVs are associated with multiple growth and meat quality traits (Jiang et al. 2014), whereas an insertion in the *SPEF2* gene influences the reproductive performance of boars (Sironen et al. 2012). An inversion near the *KIT* gene explains the Tobiano spotting pattern in horses (Brooks et al. 2007) and merle patterning of the dog may be induced by a retrotransposon insertion in the *SILV* gene (Clark et al. 2006).

While growing, the number of studies based on SVs in the literature is still really low compared with that based on SNPs. This difference may be mainly explained by the fact that detecting SVs in WGS data remains challenging. This is due to multiple factors, such as the absence of standardized protocols, the low overlap between the results obtained via different methods, and the sensitivity to multiple sequencing parameters like sequencing depth or library size (Reviewed in Tattini et al. 2015).

Then, one might be tempted to detect causal SVs solely via surrounding SNPs involved in the same selective sweeps. This is the case in the two of the examples reported above. The increase in amylase activity in dogs related to a selective sweep around the *AMY2B* gene, is explained by a CNV (Axelsson et al. 2013) and the region of the *RXFP2* gene related to sheep polledness (Kijas et al. 2012) was later associated to a 1.8-kb insertion–deletion (Wiedemar and Drögemüller 2015). However, although we can predict a redundancy of information from SNPs in linkage disequilibrium with SVs, there is yet no study assessing the ability of SNP-based detections to identify SVs under selection at a whole-genome scale.

Previous work based solely on SNPs data explored the genetic bases of the domestication of both sheep and goats by contrasting domestics animals from Morocco and Iran to their wild relatives from Iran (Alberto et al. 2018). The present study takes advantage of this data set of a hundred whole-genome sequences, in order to 1) create the first atlas of SVs in small ruminants' genomes, and 2) detect SVs putatively targeted by selection during domestication and further improvement of those species. Finally, this data set gives a unique opportunity to observe to which extent the detection of SVs per se identifies selected regions that would not be detected using SNPs.

Results

Structural Variations in Sheep and Goats

We used Breakdancer (Chen et al. 2009), Delly (Rausch et al. 2012), and BAdabouM (Cumer et al. 2020), to detect SVs independently for each individual. We retained SVs independently called by at least two softwares in at least two individuals (except for insertions, see the detailed procedure in the Materials and Methods section). In the 51 *Ovis* WGS, composed of 11 wilds *O. orientalis* and 40 domestics *O. aries*, 45,796 SVs were identified with a median number of 14,222 SVs per individual. For the 58 *Capra* WGS, composed of 18 wilds *C. aegagrus* and 40 domestics *C. hircus*, 15,047 SVs were called with a median number of 3,639.5 SVs per individual. This set of SVs is composed mainly of deletions (65% and 83%, for sheep and goats, respectively) and insertions (27% and 10%, respectively). We also noted about 4% and 5% of inversions, 2% and 1% of inter- or intrachromosomal translocations for sheep and goats, respectively, and 0.5% of CNVs in both species (table 1).

The overall distribution of the SVs in both species was similar (fig. 1*A*). A high proportion of SVs was made of rare variants with 41.4% and 49.4% of the SVs having a frequency lower than 0.1 for sheep and goats, respectively. For both sheep and goats, these rare SVs were more likely specific to one group (i.e., domestics from either Iran or Morocco or wild animals from Iran). When the frequency increases, SVs tends do be shared between groups. Thus, less than 5% were private to one group when the frequency is above 0.12 for *Ovis* and 0.15 for *Capra*. In total, 79.8% and 90.8% of the SVs were shared by wild and domestic goats and sheep, respectively.

SVs are not polymorphic in all populations, and the rate of polymorphic SVs was calculated for each population as the ratio between the number of loci with the two alleles present in the population, compared with the total number of SVs recorded in this study (see Materials and Methods). This rate was lower in wild individuals than in domestics for *Capra* and higher in wild *Ovis* relatively to domestic sheep. Thus, in sheep, the rate of polymorphic SVs was equal to 0.87 for *O. orientalis*, 0.85 for Iranian *O. aries*, and 0.79 for Moroccan *O. aries*. For goats, the rate of polymorphic SVs was 0.7 for *Capra aegagrus*, and 0.81 and 0.77 for Iranian and Moroccan *C. hircus*, respectively. Even if the sNMF analysis predicted $K = 1$ as the best number of genetic clusters for sheep, for $K = 2$, the structure analysis showed a clear differentiation between wild and domestics, and within domestics

Table 1

Total Number and Repartition of Structural Variations Found in Ovis and Capra

SV Class	*Ovis*		*Capra*	
	No. Sites	Mean Number/ind (SD)	No. Sites	Mean Number/ind (SD)
DEL	30,096	10,076 (614)	12,508	2,985 (336.5)
INS	12,816	3,113 (674)	1,555	303.5 (45.5)
INV	1,806	673 (37)	743	212 (33)
CNV	240	49 (8)	73	13 (4)
ITX	99	68 (10)	4	2 (1)
CTX	739	197 (43)	164	40 (12)

Note.—DEL, deletions; INS, insertions; INV, inversions; CNV, copy number variation; ITX, intrachromosomal rearrangements; CTX, interchromosomal rearrangements.

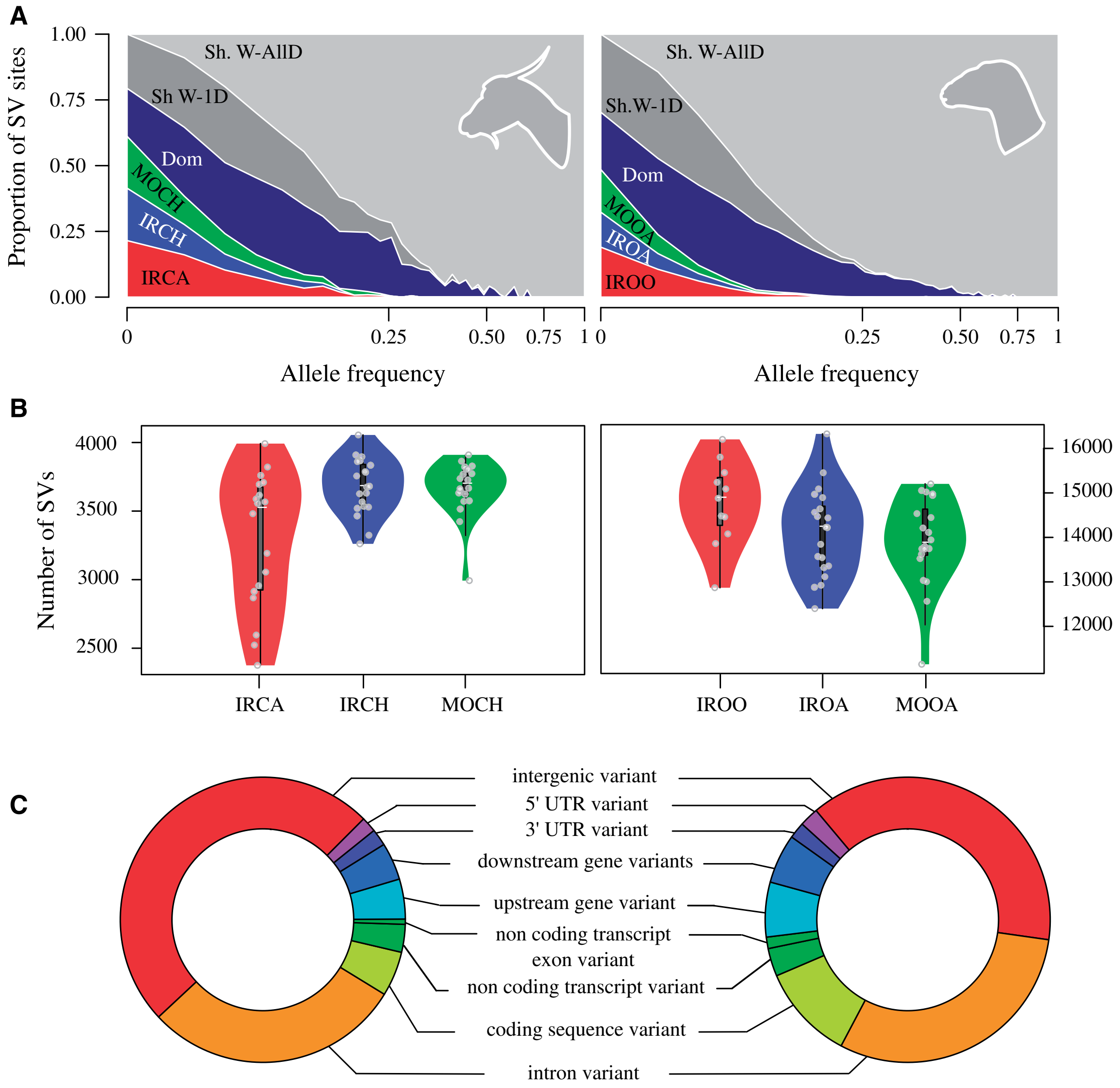

Fig. 1.—Distributions of structural variations (SVs) in goat (left) and sheep (right) genomes. (*A*) Distribution of SVs within and between groups. SVs frequency is log transformed. Codes and color are relative to the groups studied. Dom refers to SVs shared by the two domestic groups. Sh W-1D refers to SVs shared by the wild group and one domestic group. Sh W-AllD depicts SVs present in all groups. (*B*) Distribution number of the number of SVs in each group. (*C*) Distribution of SVs localization in the different types of genomic regions in each species.

in accordance with a hierarchical clustering of individuals for both species (supplementary fig. S1, Supplementary Material online). The number of SVs per individual was significantly different between wilds and domestics for both sheep and goats (*t*-test *P* values: 0.035 and 0.005, respectively). Wild *Ovis* harbored more SVs than domestics, whereas it was the opposite for *Capra*. There was no significant difference between domestic groups within both species (fig. 1*B*). The distribution of SVs localization according to gene annotations was similar in sheep and goats (fig. 1*C*), despite some differences such as a higher proportion of SVs in introns for *Capra* (33% and 41.3%) and a higher proportion of variants affecting coding sequences in *Ovis* (11.8% vs. 6.9% in goats).

Detection of Regions Putatively Selected during Domestication

To detect SVs that were differentiated between domestics and wilds (i.e., potentially under selection), we calculated at each locus two differentiation indexes: DI_{SV}, which reflects the difference in allele frequencies between populations, and the Weir and Cockerham F_{ST} (FST_{SV}) (Weir and Cockerham 1984, see Materials and Methods). The DI_{SV} index may generate false negatives when an allele is absent or fixed in a population and at medium frequency in the other. In this case, DI_{SV} will be low while this SV could be under selection in a population. The other index, $FST_{SV,}$ is affected by the heterogeneity of group size between wilds and domestics. In our study, observed heterozygosity within domestics (more than 2/3rd of the samples in both species) is close from global expected heterozygosity even if the SV is absent or fixed in wilds. Despite these potential biases, there is a clear correlation between DI_{SV} and FST_{SV} (goat, $r^2 = 0.62$, *P* value <2.2e-16; sheep, $r^2 = 0.49$, *P* value <2.2e-16). To select the most differentiated SVs between wild and domestic and avoid false positives, three sets of SVs were extracted: SVs had to be simultaneously either above the 90th, the 95th, and the 99th quantiles of both FST_{SV} and DI_{SV} distributions.

We conducted a variant effect predictor (VEP) analysis for each of the three sets in order to identify genes potentially impacted by SVs (i.e., not in intergenic regions, see Materials and Methods section). The gene ontology (GO) analyses of terms associated with these gene sets revealed no significant enrichment (results not shown).

In total, 135 SVs for sheep and 70 for goats were above the 99th quantiles of both FST_{SV} and DI_{SV} distributions and were considered as selected SVs (fig. 2*A*). The lowest FST_{SV} value for selected SVs was 0.26 for goat and 0.31 for sheep, whereas whole-genome FST based on SNPs were 0.047 for Capra and 0.050 for Ovis (Alberto et al. 2018). The lowest absolute value of DI_{SV} was 0.58 for sheep and 0.56 for goat.

In order to assess the ability of studies based on SNPs to detect SVs under selection, we then tested whether SNPs surrounding differentiated SVs were also differentiated between wilds and domestics. To do so, we performed Wilcoxon rank-sum tests on the mean FST values of 1) the SNPs surrounding SVs highly differentiated between wilds and domestics and 2) the SNPs surrounding other SVs (i.e., putatively neutral). In a way to assess a possible linkage decay, we tested if this differentiation decreased when the size of the window around SVs increased from 5 to 100 kb. All tests were highly significant (*P* value lower than 10^{-10}) for small windows (5–10 kb, supplementary table S4, Supplementary Material online), and the test significance decreased with the increase of the window's size in both sheep and goats. With 100-kb windows, mean FST around SVs were not different between selected and neutral SVs (*P* value for goats: 0.08/*P* value for sheep: 0.052). Distributions are presented in supplementary figure S2, Supplementary Material online.

Finally, we refined the set of differentiated SVs, by filtering the SVs for having a $FST_{SNP\text{-}5K}$ (i.e., FST_{SNP} calculated for SNPs found in 5-kb-wide windows centered on the SV) higher than twice the mean $FST_{SNP\text{-}5K}$ of neutral SVs (nonselected based on SVs differentiation). This resulted in 69 and 45 genomic regions (out of the 135 and 70 selected previously) highly differentiated between wilds and domestics (supplementary table S2, Supplementary Material online). In these regions, the VEP identified 25 and 20 SVs close to or overlapping with genes (i.e., not in intergenic regions, reported in supplementary table S2, Supplementary Material online). GO enrichment analyses revealed no significant term enriched among the term associated with this gene set (supplementary table S3, Supplementary Material online).

Discussion

Changes in selective pressures during domestication processes have induced multiple phenotypical, physiological, and behavioral changes in sheep and goats. This work explores the role played by SVs during this process. Based on 51 and 58 whole-genome sequences of wild and domestic sheep and goat, respectively, we built an atlas of SVs in each species. Then, we showed the consistency of the signals provided by SVs and SNPs to infer the neutral structure of populations. We also identified candidate SVs, highly differentiated between wilds and domestics in both *Ovis* and *Capra*, that may have played a role during domestication. Finally, checking the concordance between selection indices obtained for SVs and their surrounding SNPs shed light on their complementarity.

Structural Variations in Sheep and Goats

Based on a combination of several calling tools, we produced a set of high confidence SVs in sheep and goats. The distribution across SVs categories (table 1) is consistent with that observed in other mammalian genomes such as cattle (Chen et al. 2017), human (Sudmant et al. 2015), and dog (Wang et al. 2018). Previous work highlighted that SVs are inequitably

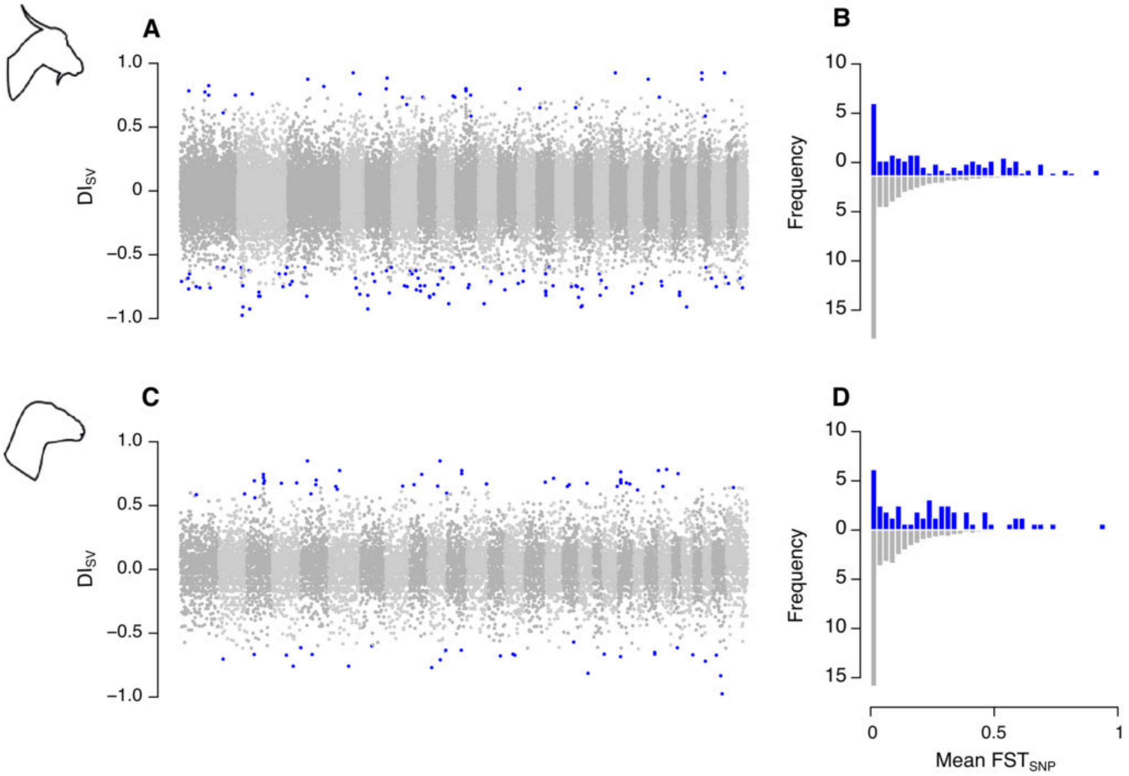

Fig. 2.—SVs differentiated between domestic and wild groups. (*A* and *C*) Manhattan plot of the DI_{SV} variations along the genome of sheep (*A*) and goats (*C*). SVs detected as selected with both FST_{SV} and DI_{SV} are in blue. (*B* and *D*) Distributions of the mean FST of the SNPs surrounding selected SVs (blue) and the FST of the SNPs around all other SVs (gray) in a 5-kb window in both sheep (*B*) and goats (*D*).

distributed over genomes, with higher density within heterochromatic regions (Li et al. 2011). Those regions are largely composed of repeated elements and, as a consequence, current SVs detection methods based on next-generation sequencing have a low sensitivity for this task due to the short size of the reads and the low mapping quality on repeated regions (Medvedev et al. 2009). Consequently, our set of SVs might be incomplete with missed SVs probably unequally distributed over the genome, thus we do not expect our high confidence SVs predictions to provide a complete description of the distribution of SVs in small ruminants' genomes.

When comparing both species, we observed more SVs in sheep than goats, despite genomes of comparable sizes (2.6 Gb for sheep and 2.9 Gb for goats). This may reflect reality and/or be explained by technical reasons. Different assembly quality may lead to different false positives or true negative in the two species (Bickhart and Liu 2014). In our case, the quality of the two assemblies is different, with 28 Mb of gaps in OAR v4 assembly and only 38 kb of gaps in the ARS1 assembly. The lower contiguity of the sheep reference genome may disrupt the mapping of the resequencing reads on the reference genome and induce a higher error rate by calling artefactual SVs based on erroneously mapped and unmapped reads. We also observed a link between the number of SVs and the coverage (supplementary table S1 and fig. S3, Supplementary Material online). Although this is an expected bias, the coverage of *Capra aegagrus* individuals is significantly lower compared with that of domestics and may limit our ability to detect SVs in the wild group. Nevertheless, despite a different number of SVs in the two species, distributions of SVs in sheep and goat populations were comparable. Although rare SVs were mostly group specific (probably due to a sampling effect), SVs with a higher frequency were shared by all populations, indicating a shared polymorphism between groups of domestics and between wild and domestics, which thus probably existed before domestication. This result is consistent with the low mutation rates of insertions and deletions, which compose a large part of the variants observed (Sudmant et al. 2015).

For both *Capra* and *Ovis*, the number of SVs was significantly different between domestic and wild animals. Wild sheep had more SVs than domestics. This result is consistent with what was observed on SNPs. Indeed, nucleotide diversity was lower in domestic sheep than in Asiatic mouflon (supplementary table S5, Supplementary Material online), suggesting a stable demography in the wild and/or lower effective size in domestics (Alberto et al. 2018). For goats, in contrast, we observed less SVs in wild individuals. If this lower number of SVs may be imputed to technical reasons (see paragraph above), this result is consistent with previous results based on SNPs (Alberto et al. 2018). Thus, current wild populations are fragmented and may suffer from recent bottleneck reducing variability, which may explain the lower number of SVs within wild populations and their lower polymorphism compared with domestics. Consistently, Bezoar ibex showed lower nucleotide diversity and higher inbreeding than domestic goats based on SNPs data (supplementary table S5, Supplementary Material online).

If, like SNPs, SVs showed a clear and consistent differentiation between wilds and domestics for both *Capra* and *Ovis*, they did not discriminate populations as clearly as SNPs. Indeed, sNMF ancestry was less clear with SVs compared with SNPs (see figure 1 in Alberto et al. 2018 for results based on SNPs) and individual clustering was not able to fully account for domestic sheep structuration (supplementary fig. S1, Supplementary Material online). This is also consistent with previous studies (Conrad and Hurles 2007) and may be explained by the lower mutation rate of SVs as well as their lower number compared with SNPs.

Comparison of SVs and SNPs for Detection of Selection

This inventory of SVs found in wilds and domestics provides the opportunity to assess the redundancy of information contained in the joined observations of SVs and SNPs. We found that SNPs close to candidate SVs (i.e., SVs strongly differentiated between wilds and domestics) were also more likely to be differentiated. Indeed, the frequency distribution of FST_{SNP} surrounding candidate SVs significantly differed from that of FST_{SNP} surrounding other SVs, considering a window of 10 kb (or lower) around SVs. This difference decreased whereas the size of the window increased. This suggests that SNPs surrounding SVs under selection become differentiated, probably as a result of genetic linkage with the SVs (a pattern illustrated in fig. 3). However, this link between SVs and SNPs was not always that clear, as shown by the overlap between the distribution of FST_{SNP} from neutral regions and that from regions surrounding selected SVs (fig. 2*B* and supplementary fig. S2, Supplementary Material online). This overlap between the distributions may be caused by different reasons. Biological parameters may influence the link between a SV and surrounding SNPs. High recombination rate around a variant, low but continuous selection pressure on a variant, or selection for a variant already present for a long time in the wild populations (thus surrounded by a large part of the diversity present in the wild population) are some scenarios that may explain the decrease of FST_{SNP} around a variant putatively under selection. A neutral SV may occur close to a variant under selection but already frequent in the population (SNP or short indel). In such a case, FST_{SNP} would be higher than FST_{SV} and might explain high FST_{SNP} values around some putatively neutral SVs. This overlap between the two distributions may also result from false positives in the list of selected SVs that may be attributed to SVs miscalling (discussed

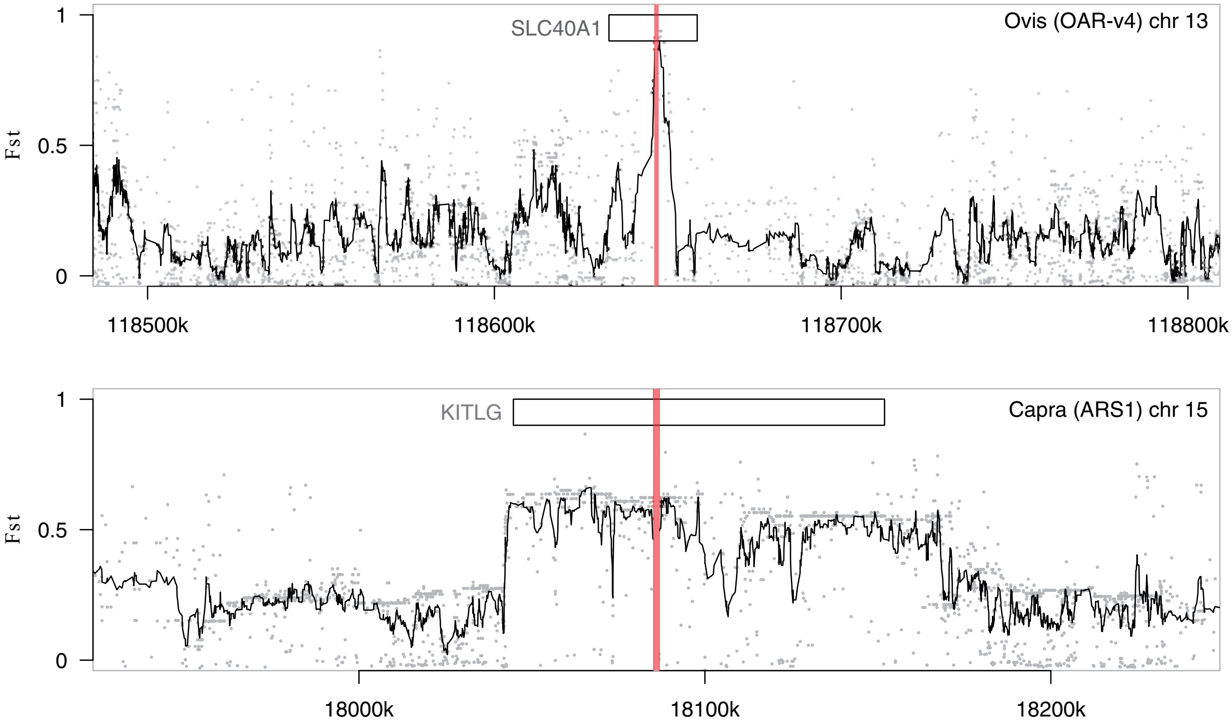

Fig. 3.—Variation of FST between wild and domestic animals along two regions surrounding a candidate SVs (red bar). Gray points represent the FST values of SNPs and the black line the mean FST value in a sliding window of five SNPs for: (*A*) *SLC40A1* gene in sheep chromosome 13 and (*B*) *KITLG* gene in goat chromosome 15.

above). To avoid such false positives when selecting regions putatively under selection, we only kept as candidate loci regions where highly differentiated SVs were associated with a $FST_{SNP\text{-}5K}$ higher than two times the mean of $FST_{SNP\text{-}5K}$ around nondifferentiated SVs.

Regions Putatively Selected during Domestication and Improvement

Based on the concordant signal of SNPs and SVs, we identified 45 regions in Capra and 69 regions in *Ovis* that were putatively under selection during domestication. A substantial part (71.8% for sheep and 45.7% for goats) of these were in intergenic regions. This result might reflect selective events that impact individual fitness through the involvement of noncoding regions implied in gene regulation (Alberto et al. 2018). However, 25 regions in *Capra* and 20 regions in *Ovis* were close to or overlapping with genes. Thus, this study reports new genes putatively associated to the domestication of sheep and goats (see supplementary table S3, Supplementary Material online). Despite the absence of metabolic pathway enriched in candidate genes in GO analyses, the functions of the genes affected is consistent with the selective pressures imposed by domestication. Striking differences between wild and domestic animals affect physiological and morpho-anatomical traits related to productivity (e.g., size, meat or milk quality, wool color, and quality). The following paragraph aims at listing and describing the genes identified in this study as potential candidates for further investigation.

The anatomy of goats may have been affected by alteration of genes close to or including a SV. This is the case of the large inversion included in the *GAL3S4* gene, a gene responsible for skeletal malformations in humans (Wu et al. 2012), and the intronic deletion in *CDCP2*, a locus previously associated with bone strength in laying hens (Rodrigues et al. 2014). We detected two insertions, one in *DNMT3A* (Liu et al. 2015) and one in *GPC2* (Lee et al. 2013), two genes associated with meat traits. In sheep, we reported a new deletion that may affect the *ADGRG6* gene, which has pleiotropic effects on body development (e.g., cartilage and spinal column development, Karner et al. 2015); myelination, Monk et al. 2011; osteoclast function and regulation of bone mineral density, Hsiao et al. 2008). Three other genes have high frequency intronic insertions in domestic sheep: *ADAMTSL3*, is associated with body traits in cattle (Liu et al. 2012), *SLC40A1* with milk production and muscle iron content (Zhao et al. 2015) and *COP5* with meat quality trait (Zhao et al. 2015).

SVs may have also impacted reproduction. In goats, *KITLG* is already a strong candidate gene affecting litter size (Li et al. 2016), and *BMPR1B* is known to influence prolificacy of Black Bengal goats (Souza et al. 2001). Both genes may be impacted by deletions reported in this study. In sheep, highly differentiated SVs point the *SPEF2* gene, responsible of an increased fertility of sows (Sironen et al. 2012, 2014) and the *MAGI2* gene, known to impact ovary formation during early embryonic development in dogs (Nowacka-Woszuk et al. 2017).

Immune resistance may also have been targeted during domestication, notably due to the transmission of pathogens in a context of increasing population density. Our study identified new variants that may impact immunity genes. In goats, SVs affect *PVRIG* also known as *CD112R*, which encode a coinhibitory receptor for human T cells (Zhu et al. 2016). In sheep, SVs may affect *CD28*, which is known to provide a stimulatory signal for T-cell activation in sheep (Chaplin et al. 1999) and *CD226*, which was related to immunity in human (Hancock and Rienzo 2008). In a more general perspective, domestication processes may affect the whole metabolism. We found the candidate genes *LDAH*, *GMDS*, and *KITLG* in goats, and *LDAH* and *MAGI2* in sheep. *GMDS* is involved in the metabolism of amino sugar and nucleotide sugar (Lemos et al. 2016), *MAGI2* is associated with feeding efficiency in cattle (Hou et al. 2012). *LDAH*, which is the only gene bearing a signature of selection in both species in our study, is known to play a role in lipid storage through lipase activity in human (Goo et al. 2014). This inventory of genes, deserving further inquiry, illustrates the potential impact of SVs during domestication processes.

Conclusion

Our study relied on the targeted detection of different categories of SVs from WGS. Considering the amount of computational efforts required to detect SVs in WGS data compared with SNPs, it is legitimate to question whether these are justified given the additional information supplied. We found that the SNPs nearby SVs under selection also depict a signal of differentiation, highlighting linkage disequilibrium between SVs and neighboring SNPs. This result suggests that studying SNPs would be enough to detect the majority of genomic regions affected by SVs under selection (i.e., >51% in sheep and >64% in goats). Despite the fact that these results might be affected by false-positive SVs and false-negative SNPs, the proportion of true SVs not detected by surrounding SNPs would rather be in the tens of percent, thus justifying the implementation of a specific detection of SVs. Moreover, in some of the genomic regions associated with domestication reported in this study, the signal borne by SNPs is not as strong as that borne by SVs. This emphasizes the fact that, whatever the study species, combining both SNPs and SVs would be required for a more comprehensive detection of genomic regions under selection. Finally, the identification of causal mutations under selection requires integrating all genomic polymorphism (SNPs, sort indels, and SVs).

Materials and Methods

Data Set

In total, 51 whole-genome sequences for *Ovis* species and 58 for *Capra* species were retrieved from the ENA archive (Alberto et al. 2018) (Information available at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/). The sampling, which was designed to investigate the genetic basis of domestication (Alberto et al. 2018), is composed of wild and domestic individuals of both *Ovis* and *Capra* genera: 11 Asiatic mouflon (*Ovis orientalis*), 18 Bezoar ibex (*Capra aegagrus*), 20 sheep (*Ovis aries*) and 20 goats (*Capra hircus*) from Iran, and 20 sheep and 20 goats from Morocco. Bezoar ibex and Asiatic mouflon are bellow referred to as Wild, in opposition to all other Domestic groups.

SVs Calling

Reads Mapping

Illumina paired-end reads of *Ovis* individuals were mapped to the sheep reference genome (build Oar_v4.0—GenBank assembly accession: GCA_000298735.2) and of *Capra* individuals to the goat reference genome (build Chir_2.0—GenBank assembly accession: GCA_000317765.2) using BWA-MEM with default parameters (Li and Durbin 2009). The BAM file produced for each individual was sorted using Picard SortSam and filtered for duplicates using the REMOVE_DUPLICATE option in Picardtools MarkDuplicates version 1.137 (http://picard.sourceforge.net, last accessed July 20, 2021). Final coverage for each individual is reported in supplementary table S1, Supplementary Material online.

SVs Calling

SVs where called independently for each individual using three different methods, namely BAdabouM (Cumer et al. 2020), Delly (Rausch et al. 2012), and Breakdancer (Chen et al. 2009). All three methods were run using default parameters, except for mapping quality of reads in input, set to 60.

SVs Clustering and Filtering

As the three methods do not detect breakpoints with the same accuracy we considered that two methods identified the same SV based on the positions of the SVs (defined by the breakpoints of the SVs) and their length (see Cumer et al. 2020 for a comparison of the three methods based on simulations and a real data set). For inversions, deletions, duplications, and intrachromosomal translocations, two SVs were considered to be the same event when they overlapped by more than 50% of their length. For insertions, as breakpoints may be narrow (only one base pair), a 1-bp overlap was considered sufficient. For interchromosomal translocations, due to the difficulty of the methods to accurately detect breakpoints, SVs with breakpoints falling within a same 1-kb window were considered as the same event.

SVs called from different individuals were considered as homologous following the same strategy based on reciprocal overlaps. We used the same 1 bp overlap strategy for insertions. Inversions, deletions, duplications, and intrachromosomal translocations from distinct individuals were considered to result from the same event when they all overlapped one another by more than 70% of their respective length.

For further analyses, we kept only polymorphic SVs called by at least two methods and present in at least two individuals, except for insertions which were called only with BadabouM, as this software detects with high confidence insertions in comparison to the two other softwares (Cumer et al. 2020). This step allowed to identify insertions, deletions, inversions, CNVs, and inter- or intrachromosomal translocations. For further analysis and for each individual, each SV was considered as a dominant presence–absence marker.

All statistical analyses were performed using the R language version 3.6.1 (Ihaka and Gentleman 1996).

Population Estimates

We calculated the SVs frequencies in each group of individuals where polymorphism was estimated as the number of polymorphic loci divided by the total number of loci. The genetic structure was inferred using the sNMF algorithm implemented in the LEA package (Frichot and François 2015). sNMF was run for a number of ancestral populations (K) ranging from 1 to 4, with 20,000 iterations and ten repetitions. Based on the minimal cross entropy criterion (De Boer et al. 2005) the best values were $K=1$ for sheep and $K=2$ for goats (supplementary fig. S1, Supplementary Material online). Individuals were also hierarchically clustered using Ward's method on a distance matrix computed using binary distance.

Analysis of Selected SVs

Differentiation of SVs between Wilds and Domestics

For each SV, we calculated 1) a differentiation index (DI_{SV}) defined as the difference between the SVs frequencies between domestics and wilds and 2) the Weir and Cockerham FST (FST_{SV}) (Weir and Cockerham 1984). To avoid bias, SVs whose FST_{SV} and DI_{SV} values were in the 90th, 95th, and 99th quantile of both parameters were extracted for VEP and GO analyses (see the Functional Interpretation section below). SVs loci whose FST_{SV} and DI_{SV} values were in the 99th quantile were considered as potentially selected.

Analysis of SNPs Surrounding SVs

Within 100 kb regions surrounding SVs, SNPs were called using samtools mpileup (Li 2011) on reads mapped with a quality equal to 60. Weir and Cockerham F_{ST} was calculated for each SNP between wilds and domestics and between groups of domestics using vcftools (Danecek et al. 2011). For each SV, the mean FST_{SNP} was calculated in windows surrounding SVs, excluding SNPs within the SV. This mean FST_{SNP} was calculated in 5 kb (respectively FST_{SNP-5K}—2.5 kb on both sides of each SV), 10 kb ($FST_{SNP-10K}$—5 kb on both sides), 20 kb ($FST_{SNP-20K}$—10 kb on both sides), 50 kb ($FST_{SNP-50K}$—25 kb on both sides), and 100-kb windows ($FST_{SNP-100K}$—50 kb on both sides). We then ran Wilcoxon rank-sum tests to compare the total distribution of FST_{SNP} with that of FST_{SNP} surrounding SVs considered as potentially selected. *P* values were adjusted according to the Bonferroni correction.

SVs Filtering

Because distribution of both FST_{SNP} surrounding potentially selected SVs, and FST_{SNP} surrounding other SVs (i.e., putatively neutral) overlapped, we considered only SVs surrounding regions with a FST_{SNP} higher than twice the mean FST_{SNP} surrounding neutral SVs based on the distribution of FST_{SNP-5K}. This filtering step allowed to extract a short list of potentially selected regions impacted by differentiated SVs for functional interpretation.

Functional Interpretation

We ran a VEP and a GO analyses for each set of SVs associated with values of DI_{SV} and FST_{SV} above the 90th, 95th, and 99th quantiles of their distribution, as well as for the potentially selected regions identified (see SVs Filtering section above). The VEP for each set of SVs was extracted using Ensembl v98 available at (https://www.ensembl.org/Tools/VEP, last accessed July 20, 2021) (McLaren W et al. 2016). VEP was conducted for *Capra* on the Chir_2.0 assembly. For *Ovis*, as the OAR_v4 assembly was not available for VEP, SVs coordinates were converted to OAR_v3.1 coordinates before VEP, using the NCBI remap service (Kitts et al. 2016). For each gene set identified during the VEP procedure, we ran GO analyses using the same procedure as the one described in Alberto et al. (2018). In Brief, GO functional annotations were extracted from Uniprot for Humans genes (http://www.uniprot.org/, last accessed July 20, 2021) using gene names matching. We considered their association with only 30 child terms (i.e., the terms' direct descents) of the "Biological Process" category (GO:0008150), as three of the 33 categories are not involved in mammalian functions and thus not relevant for this analysis (i.e., GO:0006791 sulfur utilization, GO:0006794 phosphorus utilization, GO:0015976 carbon utilization). The distribution of the GO terms associated with genes possibly impacted by selected SVs was compared with the background distribution (i.e., the 19,063 human genes associated to GO terms in Uniprot) with a Fisher exact test, for genes possibly impacted by a SV from the short list of potentially selected SVs.

We also retrieved the information available from the literature on livestock genomics to interpret the functions of genes in the selected regions identified by both SNPs an SVs (see SVs Filtering section).

Supplementary Material

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

Acknowledgments

We thank Charlotte Her for her help during the analyses. The LECA is part of the Labex OSUG@2020 (ANR10 LABX56). This study was partly funded by the French National Research Agency (Evosheep: ANR PRC Project: ANR-17-CE27-0004). Most of the computations presented in this article were performed using the CIMENT infrastructure, which is supported by the Auvergne-Rhône-Alpes region (CPER07_13 CIRA).

Author Contributions

F.P., F.B., and T.C. conceived and planned the study. T.C. run the majority of the analysis and took the lead in writing the manuscript and all authors provided critical feedback and helped during the analyses and manuscript writing.

Data Availability

All the data used in this work are available at http://projects.ensembl.org/nextgen (last accessed July 20, 2021).

Literature Cited

Alberto FJ, et al. 2018. Convergent genomic signatures of domestication in sheep and goats. Nat Commun. 9(1):813.

Axelsson E, et al. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495(7441):360–364.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. Front Genet. 5:37.

Brooks SA, Lear TL, Adelson DL, Bailey E. 2007. A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. Cytogenet Genome Res. 119(3–4):225–230.

Chaplin PJ, Pietrala LN, Scheerlinck JP. 1999. Cloning and sequence comparison of sheep CD28 and CTLA-4. Immunogenetics 49(6):583–584.

Chen K, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 6(9):677–681.

Chen L, Chamberlain AJ, Reich CM, Daetwyler HD, Hayes BJ. 2017. Detection and validation of structural variations in bovine whole-genome sequence data. Genet Sel Evol. 49: 1–13.

Clark LA, Wahl JM, Rees CA, Murphy KE. 2006. Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc Natl Acad Sci U S A. 103(5):1376–1381.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. Nat Genet. 39(7 Suppl):S30–S36.

Cumer T, Pompanon F, Boyer F. 2020. BAdabouM: a genomic structural variations discovery tool for polymorphism analyses. BioRxiv.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

De Boer P-T, Kroese DP, Mannor S, Rubinstein RY. 2005. A tutorial on the cross-entropy method. Ann Oper Res. 134(1):19–67.

Eichler EE, et al. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 11(6):446–450.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7(2):85–97.

Frantz LAF, et al. 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science 352(6290):1228–1231.

Frichot E, François O. 2015. LEA: an R package for landscape and ecological association studies. Methods Ecol Evol. 6(8):925–929.

Goo Y-H, Son S-H, Kreienberg PB, Paul A. 2014. Novel lipid droplet-associated serine hydrolase regulates macrophage cholesterol mobilization. Arterioscler Thromb Vasc Biol. 34(2):386–396.

Hancock AM, Rienzo AD. 2008. Detecting the genetic signature of natural selection in human populations: models, methods, and data. Annu Rev Anthropol. 37:197–217.

Hou Y, et al. 2012. Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. Funct Integr Genomics. 12(1):81–92.

Hsiao EC, et al. 2008. Osteoblast expression of an engineered Gs-coupled receptor dramatically increases bone mass. Proc Natl Acad Sci U S A. 105(4):1209–1214.

Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat. 5(3):299–314.

Jiang J, et al. 2014. Global copy number analyses by next generation sequencing provide insight into pig genome variation. BMC Genomics 15:593.

Karner CM, Long F, Solnica-Krezel L, Monk KR, Gray RS. 2015. Gpr126/Adgrg6 deletion in cartilage models idiopathic scoliosis and pectus excavatum in mice. Hum Mol Genet. 24(15):4365–4373.

Kijas JW, et al. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 10(2):e1001258.

Kitts PA, et al. 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res. 44(D1):D73–D80.

Lee K-T, et al. 2013. Whole-genome resequencing of Hanwoo (Korean cattle) and insight into regions of homozygosity. BMC Genomics 14:519.

Lemos MVA, et al. 2016. Genome-wide association between single nucleotide polymorphisms with beef fatty acid profile in Nellore cattle using the single step procedure. BMC Genomics 17(1):1–16.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993.

Li W, et al. 2016. Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle. Anim Genet. 47(1):110–113.

Li Y, et al. 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotechnol. 29(8):723–730.

Liu X, et al. 2015. Polymorphisms in epigenetic and meat quality related genes in fourteen cattle breeds and association with beef quality and carcass traits. Asian Australas J Anim Sci. 28(4):467–475.

Liu Y, et al. 2012. Molecular characterization, expression pattern, polymorphism and association analysis of bovine ADAMTSL3 gene. Mol Biol Rep. 39(2):1551–1560.

McLaren W, et al. 2016. The Ensembl variant effect predictor. Genome Biol. 17(1):122.

Medvedev P, Stanciu M, Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 6(11 Suppl):S13–S20.

Monk KR, Oshima K, Jörs S, Heller S, Talbot WS. 2011. Gpr126 is essential for peripheral nerve development and myelination in mammals. Development 138(13):2673–2680.

Nowacka-Woszuk J, et al. 2017. Deep sequencing of a candidate region harboring the SOX9 gene for the canine XX disorder of sex development. Anim Genet. 48(3):330–337.

Orozco-terWengel P, et al. 2015. Revisiting demographic processes in cattle with genome-wide population genetic analysis. Front Genet. 6:191.

Pang AW, et al. 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 11(5):R52.

Rausch T, et al. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28(18):i333–i339.

Rodrigues AC, et al. 2014. Congopain genes diverged to become specific to Savannah, forest and Kilifi subgroups of *Trypanosoma congolense*, and are valuable for diagnosis, genotyping and phylogenetic inferences. Infect Genet Evol. 23:20–31.

Sironen A, Fischer D, Laiho A, Gyenesei A, Vilkki J. 2014. A recent L1 insertion within SPEF2 gene is associated with changes in PRLR expression in sow reproductive organs. Anim Genet. 45(4):500–507.

Sironen A, Uimari P, Iso-Touru T, Vilkki J. 2012. L1 insertion within SPEF2 gene is associated with increased litter size in the Finnish Yorkshire population. J Anim Breed Genet [Z Tierzucht Zuchtungsbiol]. 129:92–97.

Souza CJ, et al. 2001. The Booroola (FecB) phenotype is associated with a mutation in the bone morphogenetic receptor type 1 B (BMPR1B) gene. J Endocrinol. 169(2):R1–R6.

Sudmant PH, et al. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526(7571):75–81.

Tattini L, D'Aurizio R, Magi A. 2015. Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol. 3:92.

Vigne J-D. 2011. The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. C R Biol. 334(3):171–181.

Wang G-D, et al. 2018. Structural variation during dog domestication: insights from gray wolf and dhole genomes. Natl Sci Rev. 6(1):110–122.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the analysis of population structure. Evolution 38(6):1358–1370.

Wiedemar N, Drögemüller C. 2015. A 1.8-kb insertion in the 3′-UTR of *RXFP2* is associated with polledness in sheep. Anim Genet. 46(4):457–461.

Wiener P, Wilkinson S. 2011. Deciphering the genetic basis of animal domestication. Proc Biol Sci. 278(1722):3161–3170.

Wright D. 2015. The genetic architecture of domestication in animals. Bioinform Biol Insights. 9(Suppl 4):11–20.

Wu S, et al. 2012. Evidence for GAL3ST4 mutation as the potential cause of pectus excavatum. Cell Res. 22(12):1712–1715.

Zhao F, McParland S, Kearney F, Du L, Berry DP. 2015. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol. 47:49.

Zhu Y, et al. 2016. Identification of CD112R as a novel checkpoint for human T cells. J Exp Med. 213(2):167–176.

Associate editor: David Enard