Research Article

TRANSPARENT PROCESS OPEN ACCESS Life Science Alliance

CNVs with adaptive potential in *Rangifer tarandus*: genome architecture and new annotated assembly

Julien Prunier[1,*], Alexandra Carrier[2,*], Isabelle Gilbert[2], William Poisson[2], Vicky Albert[3], Joëlle Taillon[3], Vincent Bourret[3], Steeve D Côté[4], Arnaud Droit[1], Claude Robert[2]

Rangifer tarandus has experienced recent drastic population size reductions throughout its circumpolar distribution and preserving the species implies genetic diversity conservation. To facilitate genomic studies of the species populations, we improved the genome assembly by combining long read and linked read and obtained a new highly accurate and contiguous genome assembly made of 13,994 scaffolds (L90 = 131 scaffolds). Using de novo transcriptome assembly of RNA-sequencing reads and similarity with annotated human gene sequences, 17,394 robust gene models were identified. As copy number variations (CNVs) likely play a role in adaptation, we additionally investigated these variations among 20 genomes representing three caribou ecotypes (migratory, boreal and mountain). A total of 1,698 large CNVs (length > 1 kb) showing a genome distribution including hotspots were identified. 43 large CNVs were particularly distinctive of the migratory and sedentary ecotypes and included genes annotated for functions likely related to the expected adaptations. This work includes the first publicly available annotation of the caribou genome and the first assembly allowing genome architecture analyses, including the likely adaptive CNVs reported here.

DOI 10.26508/lsa.202101207 | Received 23 August 2021 | Revised 29 November 2021 | Accepted 29 November 2021 | Published online 15 December 2021

Introduction

The genome architecture of adaptation is an important factor contributing to the evolution of a species (Feder & Nosil, 2010; Yeaman, 2013). Among the genetic variations potentially related to adaptation, structural variations (SVs), including copy number variations (CNVs), have been associated with phenotypic variations and local adaptations (Wellenreuther et al, 2019; Mérot et al, 2020). Because the first large-scale screenings showing that CNVs in human genomes involve more nucleotides than single-nucleotide polymorphisms (Sebat et al, 2004; Carson et al, 2006; Redon et al, 2006; Conrad et al, 2010; Itsara et al, 2010), an increasing number of studies even suggested that CNVs account for higher genetic differentiation than SNPs (Dorant et al, 2020) and have a greater impact on phenotypic variations (de Smith et al, 2008) and consequently on adaptation (Mérot et al, 2020).

CNVs are usually defined as DNA segments longer than 1 kb occurring in various copy numbers within a species, such copies presenting an identity exceeding 90% (Sebat et al, 2004; Feuk et al, 2006; Freeman, 2006; Redon et al, 2006). CNVs do not arise from transposable elements (Freeman, 2006) but from a variety of mechanisms including non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), single-strand annealing, breakage-fusion-bridge cycle, or replicative non-homologous DNA repair (Lovett, 2004; Gu et al, 2008; Hastings et al, 2009). Most of these mechanisms are related to the occurrence of low-copy repeats (LCRs or tandem repeats) which occur throughout the genome and present nucleotide sequence identity exceeding 95%. As a result, CNVs tend to cluster into hotspots found in the surroundings of these LCRs (Hastings et al, 2009).

SVs may appear de novo in somatic tissues where they can cause pathologies such as cancers for instance, or in the germline in which case they may be transmitted to the next generation and result in heritable phenotypic variations (Gu et al, 2008). The mutation rate for CNVs has been estimated at ~1 × 10^{-4}, which is higher than the SNP mutation rate (Lupski, 2007). They may impact phenotype through the gene dosage effect, that is, a gene CNV resulting in gene expression variation that affects the phenotype (Perry et al, 2007; Gamazon & Stranger, 2015), but they can also trigger sequence disruption (gene sequence truncation) or fusion, or even have position effects (Lupski & Stankiewicz, 2005). As a result, purifying selection may select against CNV-encompassing genes, particularly deletions that are less likely tolerated than gene duplication (Brewer et al, 1999; Conrad et al, 2010).

CNVs present several interesting characteristics regarding the genomic architecture of adaptation that can contribute to species evolution. Large CNV sequences may span more than one gene, and such gene clusters may collectively have an impact on phenotype, for example, nematode resistance in soybean Cook et al, 2012. In

[1]Département de Médecine Moléculaire, Faculté de Médecine, Université Laval, Quebec City, Canada [2]Département des sciences animales, Faculté des Sciences de l'Agriculture et de l'Alimentation, Université Laval, Quebec City, Canada [3]Ministère des Forêts, de la Faune et des Parcs du Québec, Quebec City, Canada [4]Caribou Ungava, département de biologie, Faculté des Sciences et de Génie, Université Laval, Quebec City, Canada

Correspondence: jprunier.1@gmail.com; Claude.Robert@fsaa.ulaval.ca
*Julien Prunier and Alexandra Carrier contributed equally to this work

addition, because CNVs tend to cluster into genomic hotspots (Hastings et al, 2009), they may be inherited as clusters of locally adaptive loci and thus confer an adaptive advantage (Yeaman, 2013). Finally, CNVs may prevent recombination and thus promote large genomic islands of divergence favoring the apparition and persistence of adaptations to local conditions (Tigano et al, 2018; Giribets et al. 2019 *Preprint*).

CNVs have been investigated in a number of domestic mammal species including cattle (Fadista et al, 2010; Hu et al, 2020), swine (Wang et al, 2013), horses (Wang et al, 2014), sheep (Fontanesi et al, 2011), and goats (Fontanesi et al, 2010). These early genome-wide CNV studies revealed relatively few CNVs (37–368) per genome with length averaging 127 kbp to 10.7 Mbp because of the low-resolution inherent in detection methods based on array comparative genomic hybridization (aCGH) or SNP chips (Clop et al, 2012). Nevertheless, comparison of bovine, caprine, and ovine large CNV maps shows substantial overlap (Fontanesi et al, 2011; Clop et al, 2012), which is attributed to conservation of segmental duplications in these regions, promoting recurrent CNVs through NAHR rather than CNVs inherited by descent (Clop et al, 2012). As observed in the human genome, genes included in livestock CNVs tend to be annotated for functions related to immunity, sensory perception, among others (Clop et al, 2012). More recent results obtained from higher resolution techniques and more exhaustive genome scans have corroborated such results in horses (Schurink et al, 2018) and goats (Dong et al, 2015; Genova et al, 2018) and revealed pigs CNVs that span genes annotated for functions related to metabolism and olfactory perception (Paudel et al, 2015). In addition, CNVs are involved in the between-race phenotypic diversity in dogs, including height for instance (Serres-Armero et al, 2021).

However, CNVs remain scarcely investigated at the genome scale in comparison with SNPs, particularly in wild species. This is largely due to the challenges inherent in CNV discovery at the genome level which has long relied on aCGH, now replaced by read-depth–(coverage) and read-distribution–based approaches made possible by the advent of second-generation sequencing (Alkan et al, 2011). In both cases, high-quality genome assembly is required, which is often lacking for undomesticated species, although there have been exceptions ((Prunier et al, 2017); for a gene-based aCGH approach).

In the present study, we investigated CNVs in caribou (*Rangifer tarandus*), a wild ruminant in North America. Several populations of this emblematic mammalian species with a circumpolar distribution have declined in the last decades and are endangered by climate change and human activities (Vors & Boyce, 2009; Festa-Bianchet et al, 2011). Caribou in Northeastern America are divided into three major ecotypes: the migrating caribou, which spend the winter in the forest but calve and spend the summer in the tundra, the sedentary boreal caribou, which remain in the boreal forest all year and do not migrate, and the mountain caribou, which inhabit relatively low mountain tops (Mallory & Hillis, 1998). This diversity of habitats exposes the species to a variety of selective pressures in terms of predation and parasites, competition with other ungulates, as well as varying forage composition (Mallory & Hillis, 1998). For example, sedentary boreal caribou usually travel only a few kilometers, whereas migrating caribou travel hundreds to thousands of kilometers annually (Mallory & Hillis, 1998). In addition, migrating caribou are more prone to harassment from *Oestridae* parasitic flies that are increasingly active with increasing solar radiation in the tundra (Hagemoen & Reimers, 2002), whereas sedentary boreal caribou are relatively spared in the shade of the boreal forest.

We report here a new *R. tarandus* genome assembly based on long reads and linked reads to improve completeness and quality (Warren et al, 2017), which we annotated using RNAseq de novo assembly and gene annotation from other mammals. We detected CNVs using short-read sequencing from individuals representing the three ecotypes, expecting to find ecotype-specific CNVs involving genes with annotations likely related to the different ecological conditions of the three ecotypes. Our results provide support for genomics tool development and fine-scale genomic studies of caribou.

Results

An improved genome assembly for a wild ruminant

We used the following three strategies to obtain a high-quality contiguous assembly of the genome of a female caribou: long reads with PacBio SMRT cells, Illumina 2 × 150-bp linked reads from a Chromium 10X library, and Illumina 2 × 150-bp paired-end sequencing of 400 bp inserts. PacBio SMRT cells yielded 7,534,419 high-quality long reads averaging 10,108 bp and representing an uncorrected coverage of 47× (assuming a genome size of 3 Gbp). Chromium 10X library sequencing using Illumina HiseqX yielded 2,140,002,320 linked reads of 150 bp representing a coverage of 107×. Finally, 813,953,740 short reads of 150 bp were obtained with Illumina sequencing, representing a coverage of 40×.

Assembling the long reads using Falcon (Chin et al, 2016) yielded a 2.52-Gbp genome assembly composed of 6,351 contigs (N50 = 501,648 bp, Fig 1). This assembly accuracy was supported by the BUSCO analysis that found almost all mammalian conserved orthologous genes (C: 90.3%, F: 7%). Assembling linked reads with Supernova yielded 21,785 scaffolds for a total of 2.56 Gbp (N50 = 2,383,988 bp). Almost all mammal conserved orthologous genes were again found (C: 91.7%, F: 4.2%). The Falcon assembly was then scaffolded using the Supernova assembly and the resulting assembly was re-scaffolded using a public caribou genome assembly obtained using the DoveTail approach (Taylor et al, 2019).

The final 2.59 Gbp assembly contained fewer and longer contigs and scaffolds than assemblies published so far for this species and thus represented a significant improvement (Fig 2), particularly in terms of the number of scaffolds representing 90% of the assembly (L90) (Table 1). Using short reads assembled independently or to correct long reads did not improve the genome assembly in terms of contiguity (N50) or accuracy (BUSCO analysis).

High synteny and phylogenetic clustering with other ruminant genomes

Bos taurus and *Capra hircus* genomes were compared on the basis of scaffold alignment with reference genomes using minimap2 (Li, 2018) and visualization integrated into the JupiterPlot bioinformatic

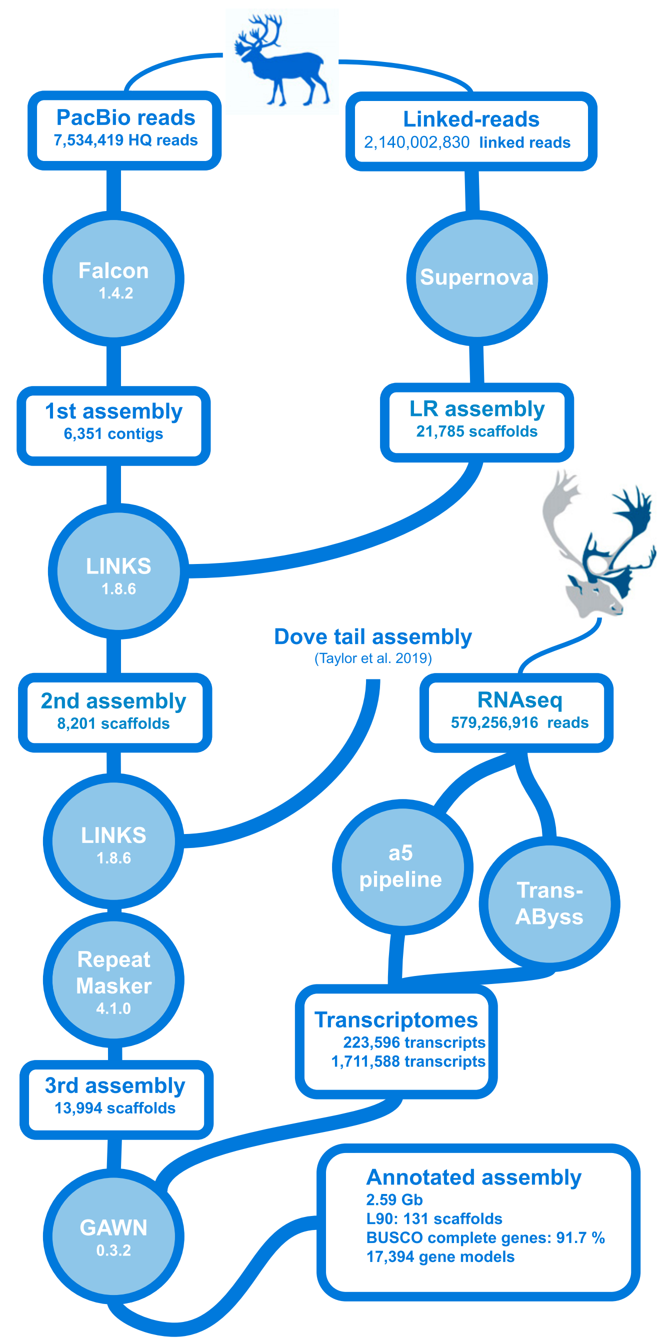


Figure 1. Caribou genome assembly and annotation pipeline.

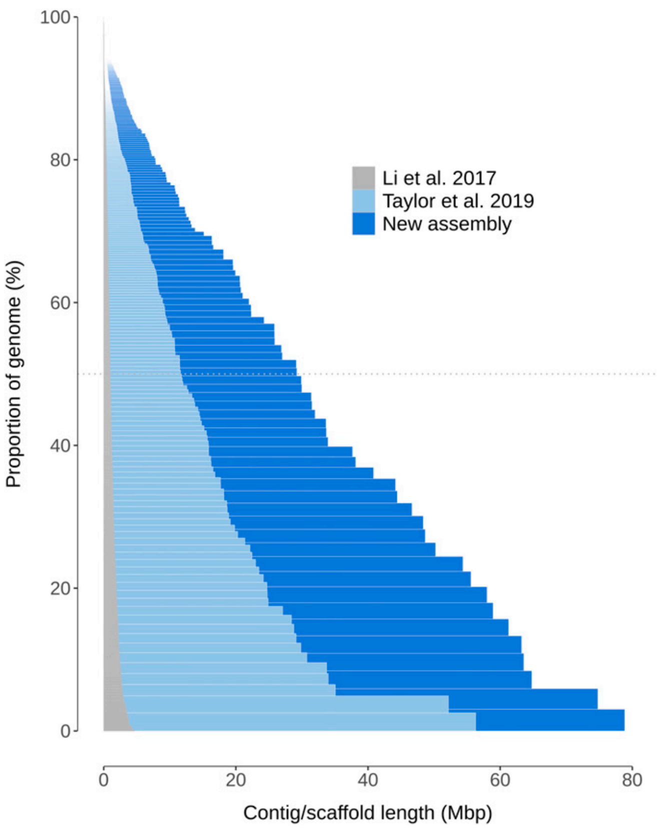


Figure 2. Scaffold length distributions in published *Rangifer tarandus* genome assemblies.

tool (Chu, 2018; https://github.com/JustinChu/JupiterPlot). Since representation was found to be the same for these ruminant genomes, only the comparison with *B. taurus* is shown in this report (Fig 3). In both cases, a very high synteny was observed, although 18 crossing lines and bands indicated variations in DNA segment order and contiguity.

A phylogenetic tree rooted with the human genome was obtained using the single-copy orthologous genes from the *mammalia_odb10* database (Fig 4). In each of 10 species, 5,156 complete conserved genes were found and used to build the tree. As expected, caribou first clustered with mule deer (*Odocoileus hemionus*), a deer species common in western North America, and moose (*Alces alces*), another cervine inhabitant of the boreal forest. Together, these species represent the *Cervidae* clade and clustered with other Artiodactyla species including the *Bovidae* clade (including *B. taurus*, *Bos indicus*, and *C. hircus*), *Suidae* (*Sus scrofa*), and *Camelidae* (*Camelus dromedarius*).

Genome annotation inferred from RNAseq de novo assembly

Gene expression diversity was maximized for annotation purposes by sequencing RNA extracted from several tissues (liver, muscle, blood, heart, lung, kidney, ovary). Read sequences were de novo assembled into transcripts using the *TransABySS* and *a5* bioinformatic tools, and then mapped on the genome assembly to identify coding regions (Fig 1), which were annotated according to similarity with sequences in a Uniprot database. Because the complete annotated genomes closest to *R. tarandus*, namely, *B. taurus* and *C. hircus* were annotated with putative functions based on similarity with the human genome, the cleaned Swissprot database (which includes only human sequences) was used (https://www.uniprot.org/proteomes/UP000005640) to avoid redundancy.

Transcripts were more numerous in the *TransABySS* transcriptome assembly (1,711,588) than in *a5* one (223,597). This process resulted in the identification of 20,419 annotated genes based on

Table 1. *Rangifer tarandus* genome assemblies published or obtained in this study.

Publication	Total sequence length (Gbp)	Number of scaffolds	Scaffold N50 (kbp)	L90 scaffold number	GC content, %	BUSCO analysis (total dB size)[a]
Li et al (2017)	2.64	58,765	986	–	41.2	92.6% (4,104)
Taylor et al (2019)	2.21	4,699	11,765	289	41.4	93.1% (4,104)
Weldenegodguad et al (2020)	2.66	23,450	5,023	–	41.4	92.9% (4,104)
The present study	2.59	13,994	29,299	131	41.5	91.7% (9,226)

[a] Only the ratio of complete single-copy gene sequences are shown here; gene database size in parentheses.

the *a5* assembly and 30,731 based on the *TransABySS* assembly. Overlap between both assemblies resulted in 17,394 corroborated annotated gene structures that were distributed over 2,759 genome assembly scaffolds. Among these, 3,025 coding sequences were annotated for transposable elements resulting in 17,394 gene models (gff3 file, Supplemental Data 1). Short coding sequences (<500 bp) with low coverage (<80%) or without homology with human gene sequences were not annotated.

Large CNVs clustered in hotspots and encompassed coding sequences

CNVs were detected in 20 individuals representing the three *R. tarandus* ecotypes using second-generation sequencing data and three types of evidence as implemented in the SpeedSeq tools suite (Chiang et al, 2015). Since our primary goal was to identify CNVs with adaptive potential, and thus subject to natural selection, rather than de novo CNVs not transmitted over generations, those detected in only one individual (or only in the reference assembly) were discarded. A total of 1,698 CNVs longer than 1,000 bp were detected over all samples, average length being 200,521 bp. The number of scaffolds containing at least one CNV was 162, and larger scaffolds contained more (Fig S1A). Altogether, CNVs accounted for 11.3% of the genome assembly (340,590,909 bp). Deletions were more numerous than duplications (1,466 versus 232) but significantly smaller (t = 3.7, P = 0.0002, Fig S1B). The number of CNVs per individual averaged 1,344.21 and ranged from 740 to 2,252 (Fig S1C), while the average CNV locus frequency was 0.355. CNVs were not randomly distributed over the genome assembly but clustered into 31 hotspots including 227 CNVs (KS test; D = 0.047 and P = 0.001; Fig 5). The number of CNVs per hotspot averaged 7.32 and reached 14. No scaffold contained more than three hotspots of CNVs.

A total of 332 of these large CNVs (19.5%) overlapped coding sequences, involving a total of 1,217 of the gene models identified in our genome assembly annotation. Duplications involved an average higher number of gene models (mean = 0.22 coding sequences per CNV, from 0 to 4) than deletions (mean = 0.20, from 0 to 5). The gene models involved in CNVs were annotated for functions altogether related to a large diversity of processes. An enrichment analysis in GO terms was performed and revealed a significant enrichment (adjusted P < 0.05) in various biological processes, including functions related to "regulation of protein metabolic process" (GO:0032269), "leukocyte activation" (GO:0045321), "muscle structure development" (GO:0061061), or "inflammatory response" (GO:0061061), among others (Table S1).

To characterize the CNVs varying the most between ecotypes, a discriminant analysis of principal components (DAPC) was performed to identify CNVs for which the genetic distance between boreal sedentary and migrating ecotypes was maximal (Fig 6A). The mountain ecotype was not included because it was represented by a single individual. The 15 retained principal components explained 87.4% of the overall genetic variance and only the first discriminant function was retained. This revealed 43 CNVs showing 2.5% of the highest loading scores on the first discriminant function (Fig 6B). Although most of these CNVs did not include any sequence annotated in our assembly, 15 were interestingly annotated for functions related to muscle and cardiac physiology, such as "musculoskeletal movement" and "regulation of heart rate," temperature responses ("response to cold"), immune responses ("innate immune response" and "defense response to bacterium"), and environmental perception ("sensory perception of sound" and "visual perception") (Fig 6C).

Discussion

Genome assembly using different technologies

Each sequencing technology has its relative strength for de novo assembly of large genomes from non-model species. Whereas Illumina allows efficient sequencing of billions of high-quality reads, these tend to remain short (<1 kb), making it difficult to scaffold and improve large genome assembly contiguity (Warren et al, 2015; Coombe et al, 2018). However, linked reads corresponding to long DNA molecules of known origin (Chromium 10X), large insert sizes, or reads integrating remote DNA subsequences (DoveTail) allow scaffolding of short contigs into longer scaffolds that are informative of the DNA sequence distribution over the genome despite the occurrence of possibly large gaps. On the other hand, long-read sequencing can yield genome assemblies with higher contiguity, although reads are usually less numerous and of lower quality. To take advantage of each technology, our sequencing used short reads (Illumina), long reads (PacBio SMRT), and linked reads (Chromium 10X) assembled independently, each strategy yielding a genome assembly. These assemblies were then scaffolded using one another and public data to obtain the best genome assembly available to date for this species according to contiguity and correctness measures, for example, L90 = 131 (Table 1), whereas L90 = 289 in Taylor et al (2019). However, short reads obtained from a 400 bp insert library did not allow us to improve our assembly, which was unexpected in view of previous findings (Jackman et al,

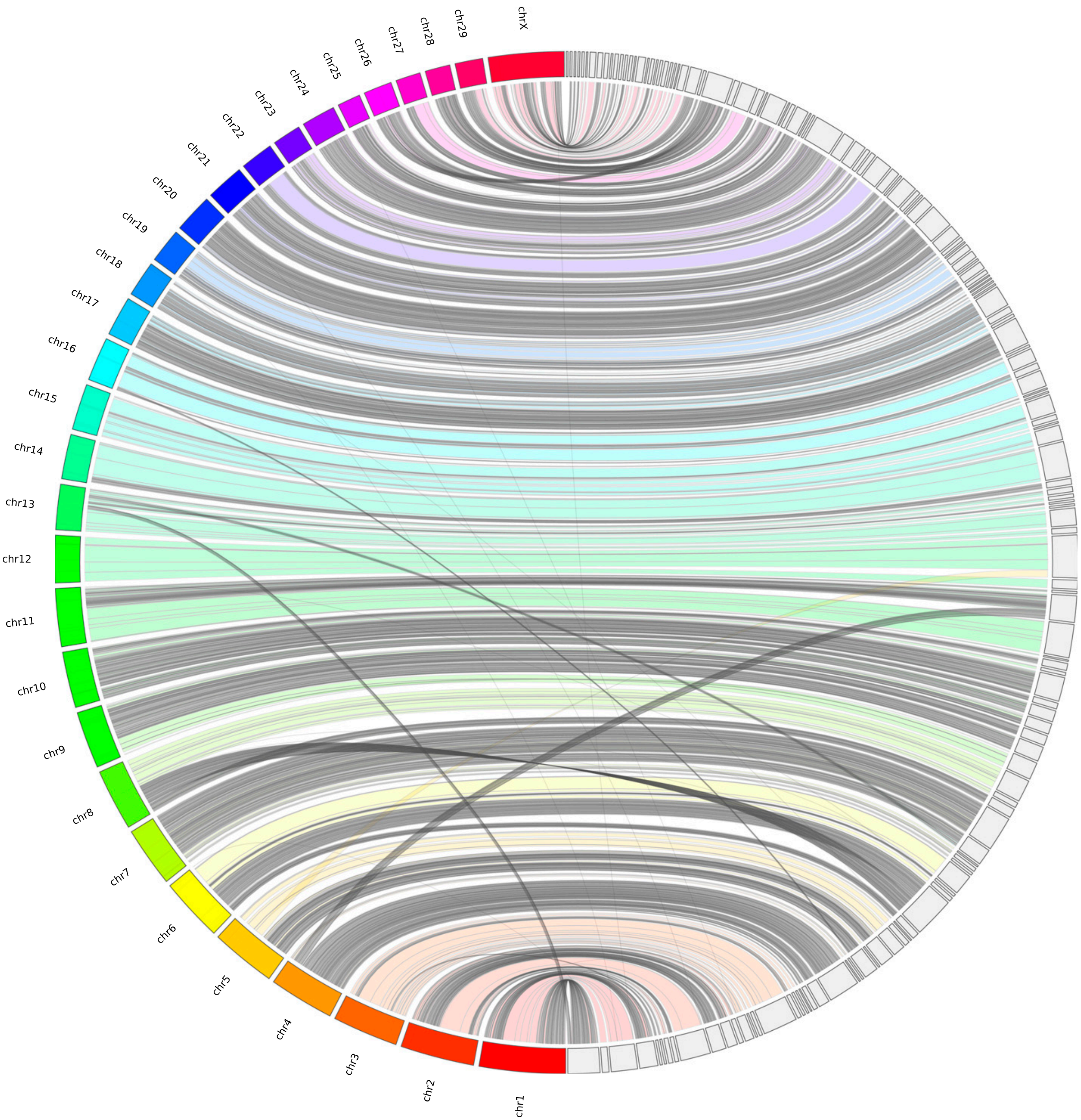


Figure 3. Synteny between caribou scaffolds and bovine chromosomes.
Rangifer tarandus genomic scaffolds were aligned with the *Bos taurus* reference (ARS-UCD1.2) using JupiterPlot (https://github.com/JustinChu/JupiterPlot). Bovine chromosomes are labeled on the left and the 144 largest matching caribou scaffolds are represented on the right. Colored bands indicate syntenic regions in the same sense, whereas grey bands indicate antisense synteny. Intersecting bands indicate non-syntenic regions between genome assemblies.

2018). This was likely due to the very high yield (107×) that we obtained for linked reads that were also short reads with the same very low sequencing error rate. Linked-read sequencing thus proved to be a very interesting strategy for de novo assembly of a large genome.

In terms of contiguity and accuracy (BUSCO analysis; Table 1), this new genome assembly compares very well with other recent genome assemblies for livestock species such as chicken (Warren et al, 2017) or other wild species such as the grizzly bear (Taylor et al, 2018) or sea otter (Jones et al, 2017) and was superior to those of

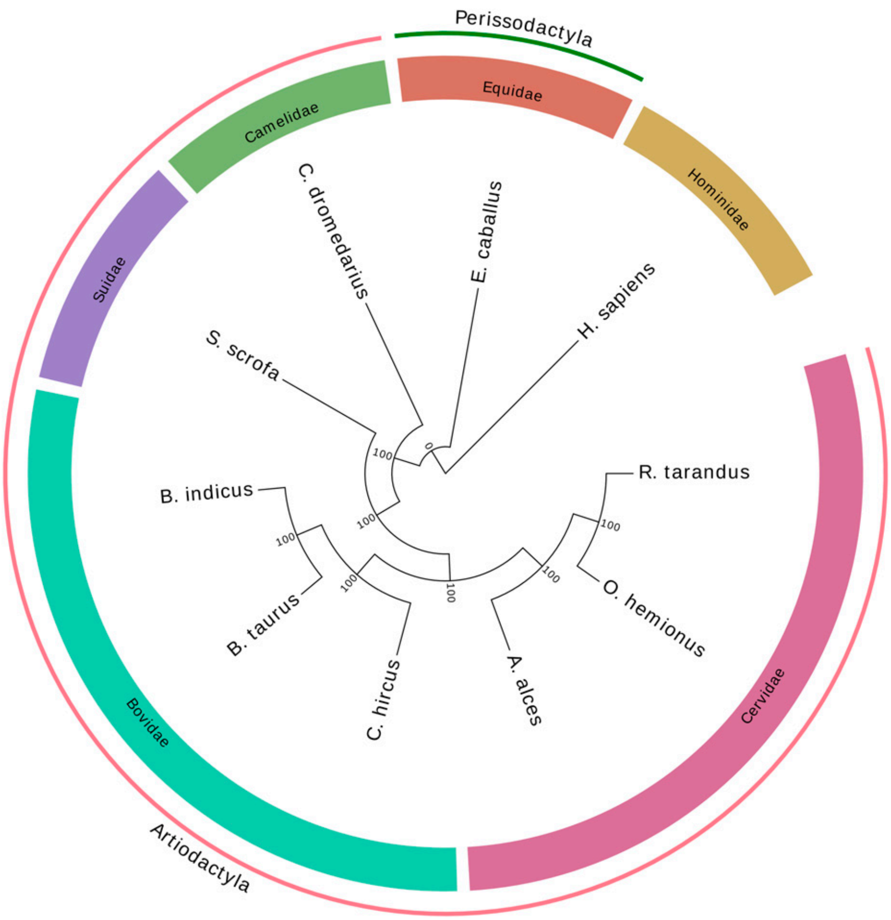


Figure 4. Phylogenetic tree of *Rangifer tarandus* and nine other species based on the 5,156 complete orthologous genes, rooted using the human species.

other *Cervidae* species (Dussex et al, 2020; Upadhyay et al, 2020). Notably, the highest quality genome assemblies (including the present one) are usually obtained when different sequencing strategies are used, including long reads and linked reads, often in combination with typical short-read sequencing (Kongsstovu et al, 2019; Wallberg et al, 2019).

Consistent with the high number of orthologous genes found in our new caribou genome assembly, the phylogenetic tree obtained using those sequences (Fig 4) presented the expected species relationships, with *Bovidae* being the clade closest to *Cervidae*, which included moose and mule deer. These two families have several characteristics in common (two-toed ungulates, ruminants), including a similar genome size and overall structure inherited from a common ancestor. However, fissions of six chromosomes changed the number of chromosomes from 29 to 35 in *Cervidae*, whereas a fission of chromosomes 26 and 28 brought the total to 30 in *Bovidae* (Frohlich et al, 2017). Scaffolds from *Cervidae* genome assemblies therefore show a high synteny with the cow reference genome (Li et al, 2017; Bana et al, 2018; Taylor et al, 2019), although many scaffolds should map to the chromosomes that were split (1, 2, 5, 6, 8, and 9) in the course of genome evolution since the last common ancestor. A caribou scaffold might likewise overlap bovine chromosomes 26 and 28 because these two should form only one chromosome in *Cervidae*. The genome comparison illustrated by the JupiterPlot (Fig 3) indicated very high synteny with only 18 bands and lines illustrating variations in the DNA segment order. One of these crossing bands is a caribou scaffold that maps partially to bovine chromosomes 26 and 28. The remaining crossing lines and bands are indicative of either chimeric assemblies or translocations. In situ DNA sequence marking and microscopic visualization such as FISH would undoubtedly help to resolve such uncertainty. Nevertheless, so few discrepancies (excluding the expected one) between bovine and caribou genome assemblies compared to previous reports illustrates the genome assembly improvements. In addition, clustering the largest scaffolds into chromosomes using FISH, for instance, is now possible, given the relatively low number of scaffolds representing 90% of the genome assembly.

This contiguous and accurate assembly will undoubtedly pave the way to other genomic tool developments and genomic investigations of this threatened species, such as landscape genomics and genomics of adaptation at the population level.

Genome annotation based on expressed sequences

To this end, another key aspect of genomic investigations is genome assembly annotation. Despite much progress in recent years, annotation of a genome based on DNA motifs and gene prediction remains challenging and time-consuming and needs constant updating (Salzberg, 2019). As a first step towards this goal, we used

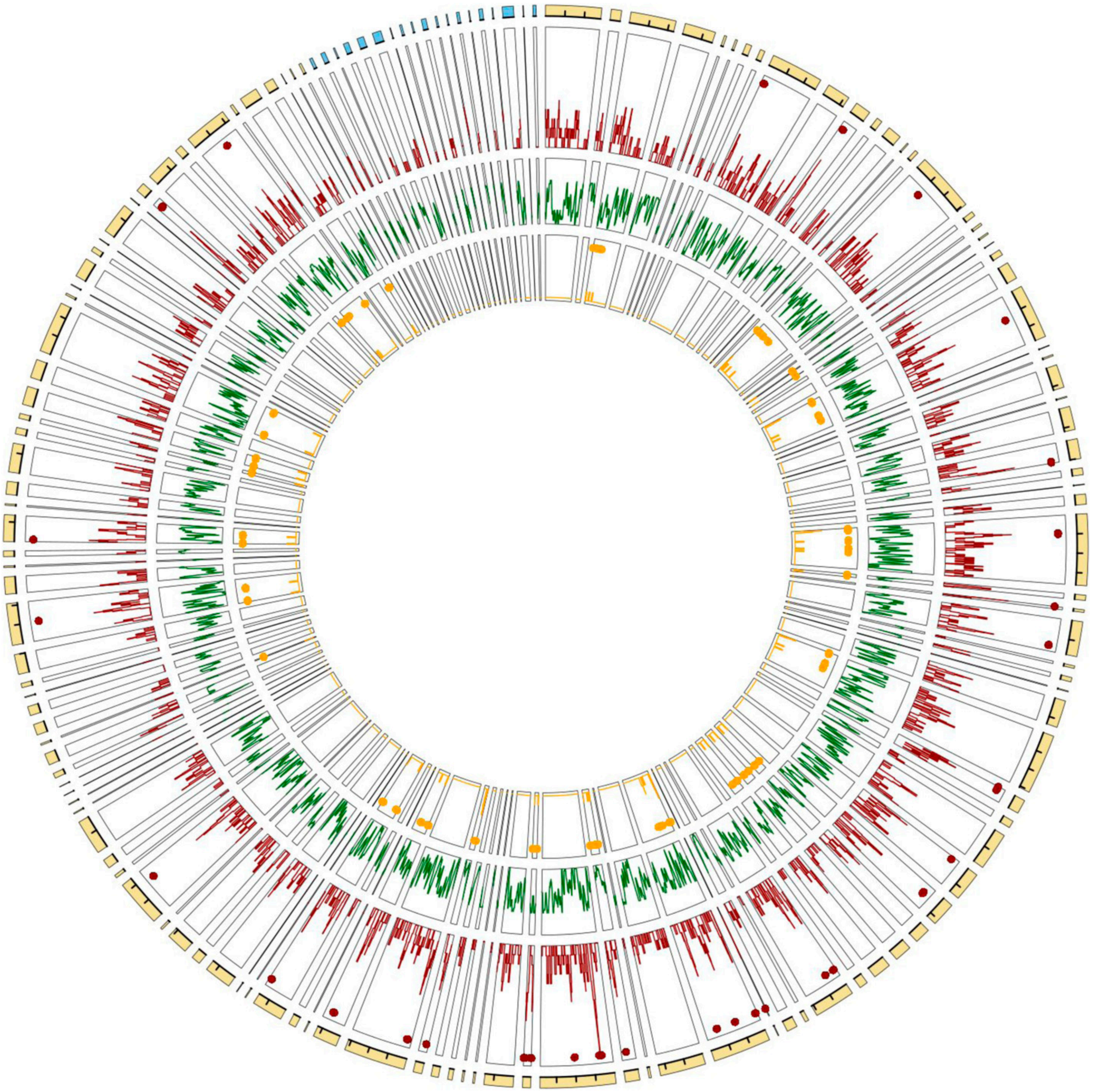
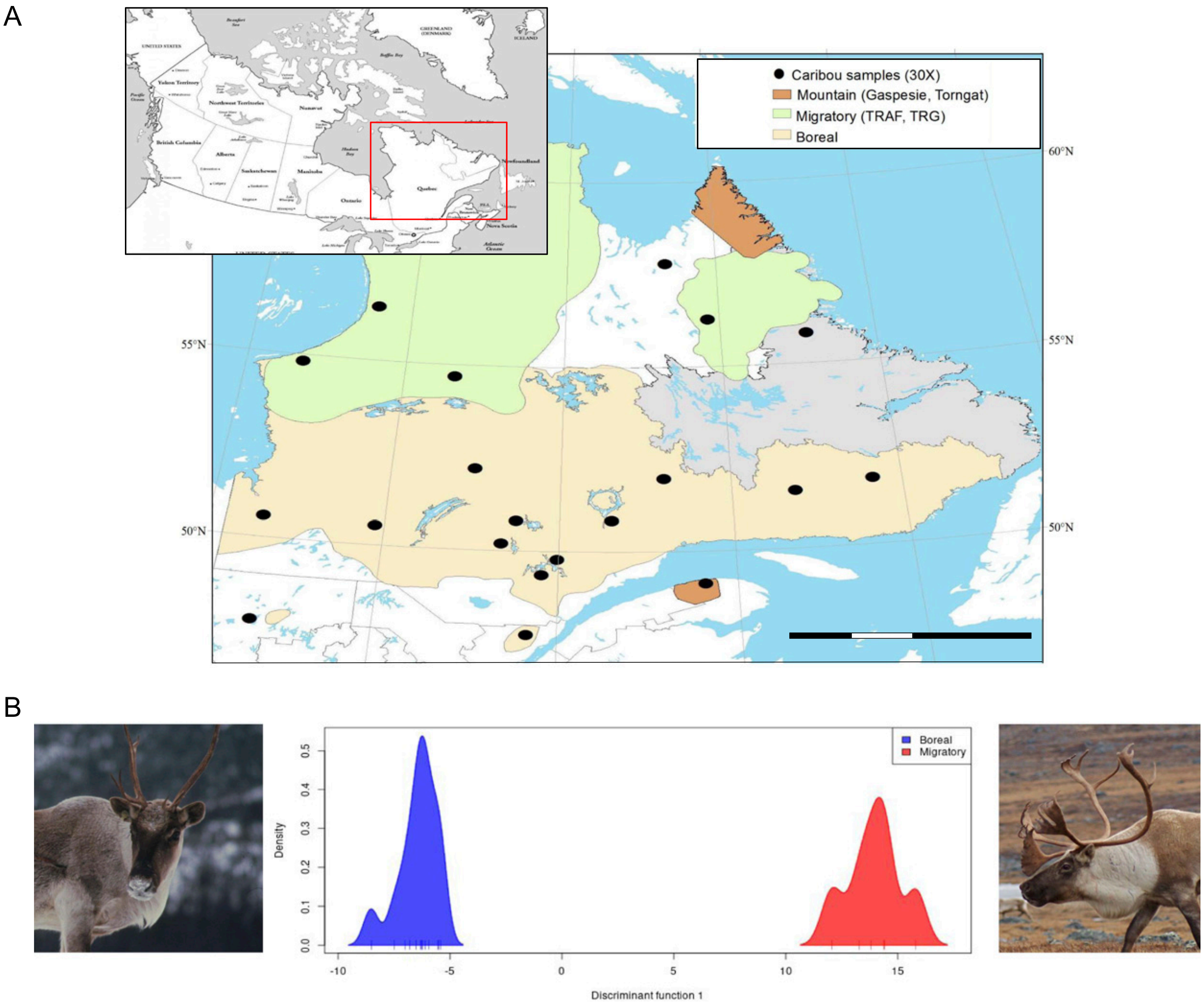


Figure 5. Genome architecture of copy number variations (CNVs), gene models, and adaptive CNVs over the largest scaffolds in the new caribou genome assembly.
From outward to inward track: the *Rangifer tarandus* new genome assembly with scaffolds matching autosomes (yellow) and the X chromosome (blue) in the *Bos taurus* assembly (interval between ticks = 20 Mbp), CNV density distribution (red) with hotspots marked as red dots, gene model density distribution (green), and distribution of likely adaptive CNVs (orange). The scaffolds are sorted according to the synteny with the *B. taurus* genome.

RNAseq of a composite sample representing several tissues to assemble a transcriptome with high diversity allowing identification of thousands of transcript sequences distributed throughout the genome. Identity with known proteins in the UniProt database allowed annotation of a large subset of these transcripts with putative functions.

The RNAseq based approach is very interesting because it allows identifying genome regions that are truly transcribed, thus avoiding most of the issues related to the occurrence of unexpressed pseudogenes and spurious identification of non-genes when predicting directly from the genome assembly. We took full advantage of this feature by discarding transcripts that could not be



C

Scaf.	start	end	loading	selected terms
2	2389890	2391167	0.0031	P:locomotory behavior; P:social behavior;
2	38501367	38503465	0.004	P:fatty acid beta-oxidation; P:regulation of insulin secretion involved in cellular response to glucose stimulus;
2	43659426	43660726	0.0036	P:vascular associated smooth muscle contraction;
5	7381328	7382639	0.0041	Adipose-associated thioesterase | P:defense response to virus; P:innate immune response | P:fatty acid metabolic process; P:negative regulation of cold-induced thermogenesis; P:response to cold; P:response to temperature stimulus
5	28762403	28783900	0.003	P:visual perception; P:regulation of metabolic process;
6	2384932	2388336	0.0034	P:musculoskeletal movement | P:glycogen metabolic process | P:defense response to bacterium | P:auditory receptor cell morphogenesis; P:camera-type eye development; | P:negative regulation of lens fiber cell development
10	47964502	47965598	0.0042	P:malate metabolic process | P: pyruvate metabolic process | P:negative regulation of cardiac muscle hypertrophy;
14	9642812	9643965	0.0038	P:regulation of heart rate by cardiac conduction; P:regulation of ventricular cardiac muscle cell action potential
19	31765618	31772487	0.0043	protein catabolic process; lipid metabolic process | P:phagocytosis, recognition; P:photoreceptor cell maintenance; retina development in camera-type eye; sensory perception of sound
44	12712610	12718096	0.005	P: regulation of appetite
48	6212192	6214304	0.0052	P:regulation of B cell receptor signaling pathway | fibroblast activation
66	961102	968152	0.0035	P:fibroblast growth factor receptor signaling pathway
68	8071824	8073537	0.0043	P:regulation of heart rate; P:response to muscle activity; P:response to muscle stretch;
112	242817	245289	0.0037	P:cellular response to lipid; P:establishment of skin barrier; P:fatty acid oxidation;
115	1875601	1876825	0.0046	P:cellular response to amino acid stimulus; P:regulation of catalytic activity

Figure 6. Divergent copy number variations (CNVs) between caribou ecotypes in Northeast America.
(A) Ecotypes distribution and geographic locations for the 20 individuals sampled and sequenced (30×) for CNV detection; **(B)** Density distribution of the boreal sedentary (on the left) and migratory (on the right) caribou over the first axis of the DAPC based on CNVs; **(C)** Adaptation-related annotations for putative genes overlapping divergent CNVs. Photo credit: Pierre Pouliot and Joëlle Taillon.

annotated because of lack of homology with known coding sequences in the human genome. These transcripts were often short and possibly represented pseudogenes with incomplete reading frames. However, such an RNAseq approach requires analyzing the greatest possible diversity of samples in terms of tissue, environmental conditions, time points (circadian variation), and developmental stage (embryo, juvenile, and adult of both sexes) to obtain an exhaustive annotation of the genome. Given the ever-increasing affordability of sequencing, this could be achievable in the foreseeable future. Nevertheless, much of the gene models set was likely reported here, given that 17,394 gene models were identified, which would represent 79% of the complete set assuming a number of genes similar to the one of *B. taurus*, for which the most recent annotation includes 21,880 gene models (Ensembl, release 104).

Because of the relative proximity to the most intensely studied mammalian models (cow, mouse, rat, and human), our annotation of coding sequences based on identity with known sequences was successful overall, with few unknown functions. However, Gene Ontology terms enrichment analyses are based on reported gene functions, which are currently associated mostly with human pathologies and disorders. Far fewer annotations relate to responses to natural environmental pressures into the wild. It is therefore possible that annotations relevant to differentiation between ecotypes (adaptations) were hidden in an excessive amount of annotations related to human pathologies (cancer or neuro-cerebral issues for example). Efforts to characterize the coding sequence molecular functions and gene ontology annotations with regards to natural environmental conditions would be beneficial to future studies focused on genetic variations in a wildlife conservation context.

Hotspots of CNVs detected in wild mammals

Genomes of domesticated mammals (including ruminants) have been entirely sequenced and studied intensively for decades, leading to the development of SNP or aCGH chips used to characterize many individuals. As methods and software making use of these resources to detect CNVs were developed, those chips have been largely used to detect such variations in a number of species, races, and lineages (Clop et al, 2012). However, few early chips were of sufficient density to cover entire genomes (Carvalho et al, 2004) and additional CNVs were discovered when whole-genome sequencing (WGS) became widespread (Alkan et al, 2011). Our WGS data revealed CNVs in 20 individuals from different ecotypes and geographic origins. As expected, we found a number of large CNVs (size > 1,000 bp) in the same range of numbers reported in previous CNV studies using the same detection approaches (Bickhart et al, 2012; Paudel et al, 2015; Schurink et al, 2018) and far more than historically detected in domesticated mammals using SNP chips and aCGH (Clop et al, 2012). These long CNVs covered 11.3% of the genome assembly, as observed for other species such as human (11–12% (Redon et al, 2006; Stankiewicz & Lupski, 2010)), and horses (11.2% (Ghosh et al, 2014; Schurink et al, 2018)), using similar detection parameters.

These large CNVs were distributed throughout the caribou genome assembly with hotspots including up to 14 CNVs. This genome architecture of CNVs including hotspots is widespread among living organisms and has been observed not only in humans and chimpanzees (Perry et al, 2006) and other mammals (Clop et al, 2012; Yang et al, 2018) but also a wide range of plants (Swanson-Wagner et al, 2010; Muñoz-Amatriaín et al, 2013; Torkamaneh et al, 2018; Prunier et al, 2019). This universality is mainly explained by the main molecular mechanisms that lead to the formation of large CNVs, which are related to the occurrence of tandem repeats (Perry et al, 2006; Hastings et al, 2009). This CNVs genome distribution including hotspots is not trivial in evolutionary terms because advantageous copy numbers are likely to aggregate into heritable clusters (Yeaman, 2013). This trend may even be amplified because CNVs may prevent recombination and thus favor the persistence of large genomic islands of divergence (Tigano et al, 2018 *Preprint*).

Another feature usually observed in whole-genome scans for CNVs is the higher number of deletions than duplications. This has long been attributed to a detection bias associated with SNP chips or aCGH, which are more prone to identify deletions that result in twofold variations than duplications that result in 1.5-fold variations in diploid genomes (Carter, 2007; Alkan et al, 2011). However, this should not affect CNV detection based on sequencing data as much, because coverage is only one element taken into account to identify a CNV (Chiang et al, 2015) and there is no obvious reason why split read–based or split read pair–based detection would be biased towards deletions. Consistent with this, a number of recent sequencing data–based studies show similar numbers of duplications and deletions (Sudmant et al, 2015; Zheng et al, 2016) although the number of detected CNVs is much higher and down to 200 bp versus 1 kb in these earlier studies, thus limiting comparability. Another possible factor contributing to higher numbers of deletions than duplications among large CNVs is one of the mechanisms leading to CNVs that results in the loss of DNA segments, namely the intra-chromatid NAHR (Gu et al, 2008). The prevalence of this mechanism has not been demonstrated to our knowledge but the higher proportion of large deletions detected in the caribou genome (86%) suggests that it may be considerable. This finds support in another sequencing data–based CNV study of cats in which the prevalence of losses was 84% using a detection threshold of 5 kb for CNV length (Genova et al, 2018). In addition, the prevalence of deletions was 90% in a recent report on dogs using a CNV minimal size of 1 kb (Serres-Armero et al, 2021). Intra-chromatid NAHR thus appears to contribute to long DNA segment deletions, of which the signal is blurred by other mechanisms when shorter CNVs are included. Meta-analysis of proportions of deletions and duplications in different CNV length ranges in a variety of species would settle this question and more generally help classify SVs that occur over a broad range of DNA lengths, from small indels to chromosomal rearrangements (Mérot et al, 2020).

Despite this new assembly representing 86–89% of the entire genome and the number of CNVs being close to those reported for other mammals (Yang et al, 2018) suggesting that we gathered a major proportion of the common CNVs, additional CNVs may occur in other ecotypes or in other parts of the species distribution. The number of detected CNVs is a measure of the CNV genetic diversity and is subject to the same detection parameters and evolutionary forces as the genetic diversity of any polymorphism. First, genetic polymorphisms are usually found in higher numbers when more individuals are studied, and CNV diversity is strongly related to the number of tested individuals (Conrad et al, 2010; Bickhart et al, 2012). Second, a hierarchical population structure is expected at the entire species distribution level with some CNVs being peculiar to specific populations (Conrad et al, 2010; Sudmant et al, 2015) or lineages (Yang et al, 2018; Hu et al, 2020). Alleles thus remain undetected when testing individuals from a fraction of the species range. Third, like SNPs, rare CNVs can be peculiar to one individual. Testing 20 individuals from a subpart of the species distribution possibly limited our detection power. However, since CNVs present higher mutation rates than SNPs (Lupski, 2007), rare alleles in CNVs possibly result from de novo formation limited to the sampled tissue and have not likely spread into the germline. Such rare CNVs provide little insight into adaptive evolution in wild species and were not targeted in this study. By sampling various ecotypes and geographic origins, we likely increased the CNV diversity and the odds of detecting CNVs related to adaptation beyond the limits of the sampled area.

CNVs signatures related to adaptation in wild mammal ecotypes

Lengthy CNVs may span entire gene-coding sequences and lead to gene expression variations, or partially overlap gene sequence, thus disrupting transcript sequence with variable phenotypic impacts (Lupski & Stankiewicz, 2005). In any case, CNVs that include coding sequences are more likely than intergenic CNVs to have such impact because the involvement of gene copy number in phenotypic variation is reported widely. One example in humans is starch-digesting ability, proportional to the number of copies of the *AMY1* gene, which encodes salivary amylase (Perry et al, 2007). Similarly, farm animal coat color is often associated with gene CNVs (Clop et al, 2012), for example, the *ASIP* gene for light pigmentation in sheep (Dong et al, 2015). Based on annotation of our caribou genome assembly, 19.5% of the CNVs overlap with gene model sequences. Annotations of these gene models represented a large diversity of biological processes enriched in GO terms related to immunity and healing, metabolism, musculoskeletal development, or environmental perception, amongst others (Table S1). Most of these terms have been revealed in previous enrichment analyses of genes in CNVs in mammals, such as metabolism and olfactory perception in swine (Paudel et al, 2015), immune responses in chimpanzees (Perry et al, 2006), horses (Schurink et al, 2018), and other farm animals (Clop et al, 2012), cardiac and skeletal muscles in humans (Conrad et al, 2010), fatty acid metabolism in circumpolar bears (Rinker et al, 2019) and body height in dogs (Serres-Armero et al, 2021). The phenotypic variations associated with CNV diversity in model organisms present a high adaptive potential for wild species such as caribou. However, the genome annotation was based on expressed sequences in a multi-tissue pooled sample. Genes not expressed in these tissues under these conditions were missed, making the list of gene models incomplete. The CNVs may encompass additional coding sequences not described here, although current annotations of identified gene models included in CNVs support their potential involvement in adaptation.

In our comparison of sedentary and migrating caribou, the DAPC analysis revealed 43 CNVs that contributed the most to the variability and are thus promising candidates to adaptive divergence. Fifteen of these overlapped gene sequences were annotated for relevant biological processes (Fig 6C). First, it is well known that migrating caribou roam hundreds to thousands of kilometers annually, whereas sedentary (boreal) ones travel much less (Mallory & Hillis, 1998). Annotations related to "muscle contraction," "heart development," "cardiac muscle hypertrophy," and "cardiac muscle contraction," as well as "musculoskeletal movement" and "locomotory behavior" were therefore unsurprising and consistent with this difference in habitat range. Similarly, annotations related to "fatty acid metabolism," "response to cold," or "vascular associated smooth muscle contraction" are consistent with the summer temperature differences between the tundra and the boreal forest and with the particular heat loss mitigation by peripheral vasoconstriction and adipose tissues reported in this species and other polar species (Blix, 2016). Adipose tissues are metabolised to free fatty acids in response to cold temperatures that are combusted in mitochondria to release heat instead of producing ATP (Blix, 2016). Interestingly, a gene with a role in adipogenesis was also found in CNV between the closely related species polar bear (*Ursus Maritimus*) and brown bear (*U. arctos*) (Rinker et al, 2019). Furthermore, five CNVs included genes annotated for functions related to "defense responses" and "immunity," including "skin barrier" annotation. As migrating caribou reaching the tundra are harassed during the summer by Oestridae parasitic flies that lay eggs under their skin (Hagemoen & Reimers, 2002), whereas sedentary caribou in the boreal forest are relatively spared by these flies, some CNV diversity between ecotypes was to be expected. Other interesting annotations included "sensory perception of sound," "visual perception," and "retina development." Given that summer habitats of migrating and sedentary caribou differ considerably in terms of forest canopy, these terms are likely related to adaptation to local conditions where sight or sense of hearing may be differentially favored. We also noted the terms "social behavior" and "regulation of appetite" which may be related to the differential group composition and access to summer forage. Whereas sedentary caribou form small groups and have access to small patches of edible vegetation spread regularly throughout the boreal forest, migrating caribou travel in large herds for kilometers to reach large patches of edible vegetation where intra-specific competition can be important, thus alternating between dietary abundance and scarcity.

Terms with slightly lower loading scores were nevertheless interesting from the perspective of adaptation and knowledge acquired from the study of Eurasian reindeer. These terms referred to light and circadian cycles such as "response to UV" (N = 44), "regulation of circadian sleep/wake cycle" (N = 2), and "vitamin-D"-related annotations (N = 15). In the spring, migrating caribou travel north, closer to the Arctic Circle, where summer nights are shorter than in the boreal forest. Thus, migrating caribous likely manage active and resting periods differently than sedentary caribou. It has been shown in the European reindeer that melatonin secretion in reindeer is highly sensitive to ambient light rather than regulated by an internal circadian clock (Stokkan et al, 2007) and more importantly, that such differences in day/night activity cycles exist between two *R. tarandus* subspecies, one inhabiting latitudes north of the arctic circle (Svalbard, Norway) and the other inhabiting northern Europe (mainland Norway) (van Oort et al, 2005). In line with this CNV gene related to latitude variations, a gene annotated with a molecular function related to UV-response was also found in different copy numbers between the polar bear mostly inhabiting the Arctic circle and the brown bear presenting a distribution extending further south (Rinker et al, 2019). In addition to this differential exposure to daylight, canopy opening also contributes to UV exposure, making “response to UV” an expected annotation.

Altogether, these promising annotations for genes included in CNVs points toward a role of CNVs in adaptation to local conditions in wild species. Although CNVs including gene models with such annotations are more interesting, the possibility of CNVs affecting gene expression, by influencing promoters or through positional effects (Lupski & Stankiewicz, 2005), should not be overlooked because these can have relevant physiological implications. Thus, further investigation of the functional aspects of all CNVs may be of interest though representing a daunting task. Nevertheless, these CNVs including genes annotated for functions potentially linked to ecotype divergent adaptive traits appeared worth being tested in large populations. Indeed, comparing 19 individuals, as presented here, is a very important first step towards the identification of adaptive CNVs but testing a non-random distribution over populations and ecotypes would further support the involvement of the CNVs in phenotypic variations in response to selective pressure (see Serres-Armero et al [2021] for an example in dogs). Because CNV detection requires relatively extensive sequencing (Lupski & Stankiewicz, 2005; Layer et al, 2014), testing several individuals with focus on the candidate CNVs reported here should allow evaluation of their impact on phenotypes and adaptations.

Conclusions

De novo assembly of large genomes is a difficult undertaking, particularly for undomesticated species, which usually present less economical interest and are consequently not, or less, described at the genome level. Genome contiguity may be reached at the expense of accuracy, although both objectives are attainable using recently developed long-read sequencing technologies. In this study, we built a new genome assembly (JAHWTM000000000, Bioproject: PRJNA739179) made mainly of a few large scaffolds that allowed the first genome architecture analysis in this species including gene models and CNVs. RNA sequencing allowed us to publicly release a first robust genome annotation for *R. tarandus* (Supplemental Data 1), which will undoubtedly pave the way to the development of genomics tools such as SNP-based genotyping chips, allowing to inform species conservation and management efforts for this species. Detecting CNVs between migrating and sedentary caribou ecotypes yielded a list of CNVs encompassing annotated genes that imply a role for CNVs in adaptation of this northern wild ruminant.

Materials and Methods

Whole-genome long-read sequencing

Previous caribou/reindeer assemblies were made using blood as the source of DNA (Li et al, 2017; Weldenegodguad et al, 2020), which is known to hinder genome assembly (Rosen et al, 2020). In addition, two interesting sequencing technologies that could improve genome assembly contiguity have not been used to date to assemble the *R. tarandus* genome, namely Pacific Biosciences and 10X Genomics technologies.

Because the single-molecule real-time (SMRT) technique (Pacific Biosciences) does not require DNA amplification before sequencing and results depend largely on DNA initial quality, high-molecular-mass (100–200 kbp) genomic DNA from muscle biopsy was isolated using a MagAttract HMW Kit according to the manufacturer’s instructions (Qiagen). DNA quantity and quality were evaluated on genomic DNA ScreenTape using a 4200 Tapestation (Agilent Technologies) and retaining only peaks of mass >45 kbp. The library was prepared for one female sample and SMRT sequencing (24 runs aiming for 30× coverage, 4 Gb of data per SMRT cell) was performed on the Sequel machine at Genome Québec (Center of Expertise and Services).

To reconstruct long DNA fragments, linked-read sequencing was also performed. Chromium 10X libraries (from 10XGenomics) were prepared at Genome Québec using the same high-molecular-weight genomic DNA as for SMRT sequencing (same female sample). Paired-end (150 bp) sequencing was performed on an Illumina HiSeqX (at Genome Québec Center of Expertise and Services). Three sequencing lanes were run to obtain ~100× genome coverage.

Transcriptome analyses

A pool of mixed samples (including liver, muscle, blood, heart, lung, kidney and ovary) was collected and transported in RNAlater stabilization solution (Thermo Fisher Scientific) and stored at −20°C until RNA extraction. RNA isolation was performed using TRIzol reagent (Thermo Fisher Scientific) as per the manufacturer’s RNA isolation protocol, followed by on-column purification and DNAse I treatment (PicoPure; Thermo Fisher Scientific). RNA quality and integrity were assessed using RNA ScreenTape on a 4200 TapeStation system (Agilent Technologies). Only RNA with an integrity number over seven was used for library preparation and sequencing.

Transcriptomes were sequenced using paired-end 150-bp Illumina HiSeqX (Illumina) at Genome Québec Center of Expertise and Services with NEB mRNA stranded Library preparation (New England Biolabs).

Whole-genome short-read sequencing of the various ecotypes

Ear punch flesh was collected from 20 individuals (10 females and 10 males) in different regions of the Province of Québec to include migratory, sedentary (boreal), and mountain ecotypes. Genomic DNA was isolated from frozen ear punches using DNeasy Blood and Tissue kits (Qiagen). DNA quantity and integrity were evaluated using genomic DNA ScreenTape on a 4200 TapeStation system (Agilent Technologies). Only samples with a DNA integrity superior to seven were used. Shotgun sequencing was performed using a PCR-free DNA library preparation (NEBNext Ultra II DNA Library Prep Kit; New England Biolabs). Libraries were paired-end 150 bp sequenced with Illumina HiSeqX. A genome coverage of ~30× was obtained from 20 lanes of sequencing.

Bioinformatics analyses

Genome assembly

The genome assembly was built from three approaches based on the three different sequence data types (Fig 1). First, high-quality long reads from PacBio sequencing were selected and assembled using the Falcon assembler v.1.4.2 (Chin et al, 2013, 2016). This assembler aligns autocorrected long reads to each other and assembles these into contigs. Then linked reads obtained from Chromium 10X sequencing were assembled independently using the Supernova assembler (Zheng et al, 2016; Weisenfeld et al, 2017; Marks et al, 2019). This assembler is an adapted version of DISCOVAR, an assembler designed to assemble short reads using De Debruijn graphs (Weisenfeld et al, 2014), that takes into account barcodes to pair reads and thereby elongate contigs and scaffolds. Finally, the short reads from the individual with the highest coverage among the 20 individuals were assembled using *DISCOVAR*-de novo, an assembler optimized to assemble genomes with size close to 3 Gb from high-quality short reads.

The Falcon assembly was scaffolded using the Supernova assembly and LINKS (Warren et al, 2015) to yield a second assembly. This second assembly was scaffolded again using the same bioinformatics tool and the publicly available genome assembly based on DoveTail sequencing (Taylor et al, 2019).

Annotation based on transcriptome assembly from RNAseq data

RNA assemblies Read quality was assessed using FastQC and reads were then cleaned using Trimmomatic v0.36 (Bolger et al, 2014). Cleaned reads were assembled twice using the SGA (Simpson & Durbin, 2012) and IDBA-UD assemblers (Peng et al, 2012) via the a5 perl pipeline (Coil et al, 2015) and the TransABySS assembler (Robertson et al, 2010). Both assemblies were kept for the next step because these algorithms may assemble RNA differently (e.g., more contiguously or less so) while pointing to the same gene regions.

GAWN The two transcriptome assemblies were then used to annotate the genome assembly using the GAWN pipeline (https://github.com/enormandeau/gawn) that maps transcriptome sequences onto the genome assembly using GMAP (Wu & Watanabe, 2005) to produce a gff3 file and gathers annotations from the SwissProt database (UniProt Consortium, 2019) using BLASTX (Altschul et al, 1990). Overlapping gene structures found in both transcriptomes using in-house scripts and the "merge" function from the bedtools suite (Quinlan & Hall, 2010) were deemed more reliable and thus included in the final annotation file (Supplemental Data 1).

Phylogeny

Single-copy orthologous genes from mammalia_odb10 found using BUSCO v3.0.2 (Simão et al, 2015; Waterhouse et al, 2018) with lineage dataset for 10 species including *Homo sapiens* as an outgroup were used for phylogenetic analysis. Common single-copy-gene DNA sequences were aligned using MAFFT v7.397 (Katoh & Standley, 2013) and trimmed using trimAl v1.4 (Capella-Gutiérrez et al, 2009). Gene sequences were then concatenated to form a single sequence per species. The phylogenetic tree was inferred using RAxML v8.2.11 (Stamatakis, 2014) with the GTR+I+G substitution model previously selected by JModelTest v2.1.10 (Darriba et al, 2012).

CNV detection and characterization

SVs were detected using the SpeedSeq tools suite (Chiang et al, 2015). Paired-end reads obtained from the 20 individuals were first cleaned using Trimmomatic v0.36 (Bolger et al, 2014) and aligned to our newly built genome assembly using "speedseq align." SNVs were then detected independently for each individual using "speedseq sv," which runs LUMPY (Layer et al, 2014). LUMPY uses three types of evidence to declare an SNV, namely read pairs, split reads and generic read depth (in our case using CNVnator [Abyzov et al, 2011] optional analysis). All detected SVs were then concatenated, and all samples were genotyped for these variations using "svtyper" (https://github.com/hall-lab/svtyper). Variations occurring within only one genome were excluded because they were deemed less reliable and may have been the result of de novo tissue-specific CNVs not transmitted over generations.

The non-random CNV distribution was tested using a genome-wide KS test between the distributions of non-CNV and CNV positions. In addition, sliding window analysis was performed to identify CNV hotspots based on the average number of CNVs within 2 Mb windows (pace 1 kb) and regions constituted of contiguous windows with average in the higher tail (above 97.5%) of the distribution were deemed hotspots of CNVs.

To characterize the CNVs varying the most between caribou ecotypes, a discriminant analysis of principal components (DAPC) was performed using the "adegenet" R-package to identify CNVs presenting the most significant variation in copy numbers between boreal sedentary and migrating ecotypes. The ecotype information (sedentary or migrating) was used as prior in the DAPC but the mountain ecotype was not included because it was represented by a single individual. The distribution of the cumulative proportion of variance explained by principal components (PCs) was used to determine PCs accounting for a great proportion of the variance and to be included in the analysis. Only one discriminant function was retained as there were only two groups to discriminate and CNV loading scores on this discriminant function were sorted. The CNVs presenting loading scores in the upper tail (2.5%) of the entire distribution were deemed putative adaptive CNVs.

Supplementary Information

Supplementary Information is available at https://doi.org/10.26508/lsa.202101207.

Acknowledgements

We thank the Centre d'Expertise et des Services Génome Québec at the Centre Hospitalier de Sainte-Justine, Montréal (QC, Canada), for their contribution to molecular data production. We also thank ComputeCanada and CalculQuébec for access to large computing resources. This project was funded through a project grant awarded to C Robert, SD Côté, and A Droit in the Genomic Applications Partnership Program of Genome Canada. We acknowledge co-financing by Genome Quebec and the Ministère des Forêts, de la Faune et des Parcs du Québec. Finally, we sincerely thank the reviewers for helpful comments and suggestions to improve the manuscript. Declaration: All samples were collected from live captures as part of caribou population monitoring by the province of Québec (Canada). Live capture and

sample collection followed the Canadian Council on Animal Care guidelines, and all procedures were approved by Animal Care Committees (CPA-FAUNE 18-04). RNA samples were collected from an individual that died during handling.

Author Contributions

J Prunier: conceptualization, data curation, software, formal analysis, investigation, visualization, methodology, and writing—original draft, review, and editing.
A Carrier: data curation, formal analysis, investigation, methodology, and writing—original draft.
I Gilbert: data curation, formal analysis, supervision, investigation, methodology, and writing—original draft.
W Poisson: data curation, investigation, and methodology.
V Albert: investigation, project administration, and writing—review and editing.
J Taillon: funding acquisition, investigation, visualization, project administration, and writing—review and editing.
V Bourret: supervision, funding acquisition, investigation, project administration, and writing—review and editing.
SD Coté: conceptualization, resources, supervision, funding acquisition, investigation, project administration, and writing—review and editing.
A Droit: conceptualization, resources, software, supervision, funding acquisition, investigation, project administration, and writing—review and editing.
C Robert: conceptualization, supervision, funding acquisition, investigation, project administration, and writing—original draft, and review, and editing.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21: 974–984. doi:10.1101/gr.114876.110

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* 12: 363–376. doi:10.1038/nrg2958

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410. doi:10.1016/S0022-2836(05)80360-2

Bana NÁ, Nyiri A, Nagy J, Frank K, Nagy T, Stéger V, Schiller M, Lakatos P, Sugár L, Horn P, et al (2018) The Red Deer Cervus Elaphus genome CerEla1.0: Sequencing, annotating, genes, and chromosomes. *Mol Genet Genomics* 293: 665–684. doi:10.1007/s00438-017-1412-3

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK, Song J, Schnabel RD, Ventura M, Taylor JF, et al (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* 22: 778–790. doi:10.1101/gr.133967.111

Blix AS (2016) Adaptations to polar life in mammals and birds. *J Exp Biol* 219: 1093–1105. doi:10.1242/jeb.120477

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. doi:10.1093/bioinformatics/btu170

Brewer C, Holloway S, Zawalnyski P, Schinzel A, FitzPatrick D (1999) A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality–and tolerance of segmental aneuploidy–in humans. *Am J Hum Genet* 64: 1702–1708. doi:10.1086/302410

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973. doi:10.1093/bioinformatics/btp348

Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. *Hum Genomics* 2: 403–414. doi:10.1186/1479-7364-2-6-403

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 39: S16–S21. doi:10.1038/ng2028

Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol* 57: 644–646. doi:10.1136/jcp.2003.013029

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM (2015) SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat Methods* 12: 966–968. doi:10.1038/nmeth.3505

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10: 563–569. doi:10.1038/nmeth.2474

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 13: 1050–1054. doi:10.1038/nmeth.4035

Chu J (2018) Jupiter plot: A circos-based tool to visualize genome assembly consistency (version 1.0). Zenodo. Available online: https://zenodo.org/record/1241235#.XA92q2hKiUk and https://github.com/JustinChu/JupiterPlot

Clop A, Vidal O, Amills M (2012) Copy number variation in the genomes of domestic animals. *Anim Genet* 43: 503–517. doi:10.1111/j.1365-2052.2012.02317.x

Coil D, Jospin G, Darling AE (2015) A5-Miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* 31: 587–589. doi:10.1093/bioinformatics/btu661

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704–712. doi:10.1038/nature08516

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, et al (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 338: 1206–1209. doi:10.1126/science.1228746

Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, Warren RL (2018) ARKS: Chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* 19: 234. doi:10.1186/s12859-018-2243-x

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: More models, new heuristics and parallel computing. *Nat Methods* 9: 772. doi:10.1038/nmeth.2109

de Smith AJ, WaltersFroguel RGP, Blakemore AI (2008) Human genes involved in copy number variation: Mechanisms of origin, functional effects and implications for disease. *Cytogenet Genome Res* 123: 17–26. doi:10.1159/000184688

Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W, Feng S, Huang G, Guan R, Shen W, et al (2015) Reference genome of wild goat (capra Aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genomics* 16: 431. doi:10.1186/s12864-015-1606-1

Dorant Y, Cayuela H, Wellband K, Laporte M, Rougemont Q, Mérot C, Normandeau E, Rochette R, Bernatchez L (2020) Copy number variants outperform SNPs to reveal genotype-temperature association in a marine species. *Mol Ecol* 29: 4765–4782. doi:10.1111/mec.15565

Dussex N, Alberti F, Heino MT, Olsen R-A, van der Valk T, Ryman N, Laikre L, Ahlgren H, Askeyev IV, Askeyev OV, et al (2020) Moose genomes reveal past glacial demography and the origin of modern lineages. *BMC Genomics* 21: 854. doi:10.1186/s12864-020-07208-3

Fadista J, Thomsen B, Holm L-E, Bendixen C (2010) Copy number variation in the bovine genome. *BMC Genomics* 11: 284. doi:10.1186/1471-2164-11-284

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* 64: 1729–1747. doi:10.1111/j.1558-5646.2010.00943.x

Festa-Bianchet M, Ray JC, Boutin S, Côté SD, Gunn A (2011) Conservation of caribou (Rangifer tarandus) in Canada: An uncertain future1This review is part of the virtual symposium [L8D2Q2M0]Flagship Species - Flagship Problems[R8D2Q2M1] that deals with ecology, biodiversity and management issues, and climate impacts on species at risk and of Canadian importance, including the polar bear (Ursus maritimus), Atlantic cod (Gadus morhua), Piping Plover (Charadrius melodus), and caribou (Rangifer tarandus). *Can J Zool* 89: 419–434. doi:10.1139/z11-025

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7: 85–97. doi:10.1038/nrg1767

Fontanesi L, Beretti F, Martelli PL, Colombo M, Dall'olio S, Occidente M, Portolano B, Casadio R, Matassino D, Russo V (2011) A first comparative map of copy number variations in the sheep genome. *Genomics* 97: 158–165. doi:10.1016/j.ygeno.2010.11.005

Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano B (2010) An initial comparative map of copy number variations in the goat (Capra hircus) genome. *BMC Genomics* 11: 639. doi:10.1186/1471-2164-11-639

Freeman JL (2006) Copy number variation: New insights in genome diversity. *Genome Res* 16: 949–961. doi:10.1101/gr.3677206

Frohlich J, Kubickova S, Musilova P, Cernohorska H, Muskova H, Vodicka R, Rubes J (2017) Karyotype relationships among selected deer species and cattle revealed by bovine FISH probes. *PLoS One* 12: e0187559. doi:10.1371/journal.pone.0187559

Gamazon ER, Stranger BE (2015) The impact of human copy number variation on gene expression. *Brief Funct Genomics* 14: 352–357. doi:10.1093/bfgp/elv017

Genova F, Longeri M, Lyons LA, Bagnato A, Strillacci MGConsortium 99Lives, (2018) First genome-wide CNV mapping in FELIS CATUS using next generation sequencing data. *BMC Genomics* 19: 895. doi:10.1186/s12864-018-5297-2

Ghosh S, Qu Z, Das PJ, Fang E, Juras R, Cothran EG, McDonell S, Kenney DG, Lear TL, Adelson DL, et al (2014) Copy number variation in the horse genome. *PLoS Genet* 10: e1004712. doi:10.1371/journal.pgen.1004712

Giribets MP, García Guerreiro MP, Santos M, Ayala FJ, Tarrío R, Rodríguez-Trelles F (2019) Chromosomal inversions promote genomic islands of concerted evolution of Hsp70 genes in the Drosophila subobscura species subgroup. *Mol Ecol* 28: 1316–1332. doi:10.1111/mec.14511

Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. *PathoGenetics* 1: 4. doi:10.1186/1755-8417-1-4

Hagemoen RIM, Reimers E (2002) Reindeer summer activity pattern in relation to weather and insect harassment. *J Anim Ecol* 71: 883–892. doi:10.1046/j.1365-2656.2002.00654.x

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. *Nat Rev Genet* 10: 551–564. doi:10.1038/nrg2593

Hu Y, Xia H, Li M, Xu C, Ye X, Su R, Zhang M, Nash O, Sonstegard TS, Yang L, et al (2020) Comparative analyses of copy number variations between Bos taurus and Bos indicus. *BMC Genomics* 21: 682. doi:10.1186/s12864-020-07097-6

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE (2010) De novo rates and selection of large copy number variation. *Genome Res* 20: 1469–1481. doi:10.1101/gr.107680.110

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM, et al (2018) Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* 19: 393. doi:10.1186/s12859-018-2425-6

Jones SJ, Haulena M, Taylor GA, Chan S, Bilobram S, Warren RL, Hammond SA, Mungall KL, Choo C, Kirk H, et al (2017) The genome of the Northern Sea Otter (Enhydra Lutris Kenyoni). *Genes (Basel)* 8: 598. doi:10.3390/genes8120379

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30: 772–780. doi:10.1093/molbev/mst010

Kongsstovu SÍ, Mikalsen S-O, Homrum EÍ, Jacobsen JA, Flicek P, Dahl HA (2019) Using long and linked reads to improve an atlantic herring (Clupea harengus) genome assembly. *Sci Rep* 9: 17716. doi:10.1038/s41598-019-54151-9

Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* 15: R84. doi:10.1186/gb-2014-15-6-r84

Li Z, Lin Z, Ba H, Chen L, Yang Y, Wang K, Qiu Q, Wang W, Li G (2017) Draft genome of the reindeer (Rangifer tarandus). *GigaScience* 6: 1–5. doi:10.1093/gigascience/gix102

Lovett ST (2004) Encoded errors: Mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* 52: 1243–1253. doi:10.1111/j.1365-2958.2004.04076.x

Lupski JR, Stankiewicz P (2005) Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* 1: e49. doi:10.1371/journal.pgen.0010049

Lupski JR (2007) Structural variation in the human genome. *N Engl J Med* 356: 1169–1171. doi:10.1056/NEJMcibr067658

Mallory FF, Hillis TL (1998) Demographic characteristics of circumpolar caribou populations: Ecotypes, ecological constraints, releases, and population dynamics. *Rangifer* 18: 49. doi:10.7557/2.18.5.1541

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al (2019) Resolving the full spectrum of human genome variation using linked-reads. *Genome Res* 29: 635–645. doi:10.1101/gr.234443.118

Mérot C, Oomen RA, Anna T, Wellenreuther M (2020) A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol* 35: 561–572. doi:10.1016/j.tree.2020.03.002

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, et al (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol* 14: R58. doi:10.1186/gb-2013-14-6-r58

Paudel Y, Madsen O, Megens HJ, Frantz LA, Bosse M, Crooijmans RP, Groenen MA, Crooijmans MA, MartienGroenen AM (2015) Copy number variation in the speciation of pigs: A possible prominent role for olfactory receptors. *BMC Genomics* 16: 330. doi:10.1186/s12864-015-1449-9

Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420–1428. doi:10.1093/bioinformatics/bts174

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39: 1256–1260. doi:10.1038/ng2123

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, et al (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* 103: 8006–8011. doi:10.1073/pnas.0602318103

Prunier J, Giguère I, Ryan N, Guy R, Soolanayakanahally R, Isabel N, MacKay J, Porth I (2019) Gene copy number variations involved in balsam poplar (Populus balsamifera L.) adaptive variations. *Mol Ecol* 28: 1476–1490. doi:10.1111/mec.14836

Prunier J, Caron S, MacKay J (2017) CNVs into the wild: screening the genomes of conifer trees (Picea spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics* 18: 97. doi:10.1186/s12864-016-3458-8

Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842. doi:10.1093/bioinformatics/btq033

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al (2006) Global variation in copy number in the human genome. *Nature* 444: 444–454. doi:10.1038/nature05329

Rinker DC, Specian NK, Zhao S, Gibbons JG (2019) Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *Proc Natl Acad Sci U S A* 116: 13446–13451. doi:10.1073/pnas.1901093116

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* 7: 909–912. doi:10.1038/nmeth.1517

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 9: giaa021. doi:10.1093/gigascience/giaa021

Salzberg SL (2019) Next-generation genome annotation: We still struggle to get it right. *Genome Biol* 20: 92. doi:10.1186/s13059-019-1715-2

Schurink A, da Silva VH, Velie BD, Dibbits BW, Crooijmans RPMA, François L, Janssens S, Stinckens A, Blott S, Buys N, et al (2018) Copy number variations in friesian horses and genetic risk factors for insect bite hypersensitivity. *BMC Genet* 19: 49. doi:10.1186/s12863-018-0657-0

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al (2004) Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528. doi:10.1126/science.1098918

Serres-Armero A, Davis BW, Povolotskaya IS, Morcillo-Suarez C, Plassais J, Juan D, Ostrander EA, Marques-Bonet T (2021) Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. *Genome Res* 31: 762–774. doi:10.1101/gr.266049.120

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212. doi:10.1093/bioinformatics/btv351

Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22: 549–556. doi:10.1101/gr.126953.111

Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313. doi:10.1093/bioinformatics/btu033

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61: 437–455. doi:10.1146/annurev-med-100708-204735

Stokkan K-A, van Oort BE, Tyler NJ, Loudon AS, Loudon ASI (2007) Adaptations for life in the arctic: Evidence that melatonin rhythms in reindeer are not driven by a circadian oscillator but remain acutely sensitive to environmental photoperiod. *J Pineal Res* 43: 289–293. doi:10.1111/j.1600-079X.2007.00476.x

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al (2015) Global diversity, population stratification, and selection of human copy-number variation. *Science* 349: aab3761. doi:10.1126/science.aab3761

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* 20: 1689–1699. doi:10.1101/gr.109165.110

Taylor GA, Kirk H, Coombe L, Jackman SD, Chu J, Tse K, Cheng D, Chuah E, Pandoh P, Carlsen R, et al (2018) The genome of the North American brown bear or grizzly: Ursus arctos ssp. horribilis.. *Genes (Basel)* 9: 598. doi:10.3390/genes9120598

Taylor RS, Horn RL, Zhang X, Golding GB, Manseau M, Wilson PJ (2019) The caribou (Rangifer tarandus) genome. *Genes (Basel)* 10: 540. doi:10.3390/genes10070540

Tigano A, Reiertsen TK, Walters JR, VickiFriesen L (2018) A complex copy number variant underlies differences in both colour plumage and cold adaptation in a dimorphic seabird. *BioRxiv* doi:10.1101/507384. (Preprint posted December 28, 2018).

Torkamaneh D, Laroche J, Tardivel A, O'Donoughue L, Cober E, Rajcan I, Belzile F (2018) Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnol J* 16: 749–759. doi:10.1111/pbi.12825

UniProt Consortium (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* 47: D506–D515. doi:10.1093/nar/gky1049

Upadhyay M, Hauser A, Kunz E, Krebs S, Blum H, Dotsev A, Okhlopkov I, Bagirov V, Brem G, Zinovieva N, et al (2020) The first draft genome assembly of snow sheep (Ovis nivicola). *Genome Biol Evol* 12: 1330–1336. doi:10.1093/gbe/evaa124

van Oort BE, Tyler NJ, Gerkema MP, Folkow L, Blix AS, Stokkan KA (2005) Circadian organization in reindeer. *Nature* 438: 1095–1096. doi:10.1038/4381095a

Vors LS, Boyce MS (2009) Global declines of caribou and reindeer: Caribou reindeer decline. *Glob Change Biol* 15: 2626–2633. doi:10.1111/j.1365-2486.2009.01974.x

Wallberg A, Bunikis I, Pettersson OV, Mosbech MB, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT (2019) A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. *BMC Genomics* 20: 275. doi:10.1186/s12864-019-5642-0

Wang J, Wang H, Jiang J, Kang H, Feng X, Zhang Q, Liu J-F (2013) Identification of genome-wide copy number variations among diverse pig breeds using SNP genotyping arrays. *PLoS One* 8: e68683. doi:10.1371/journal.pone.0068683

Wang W, Wang S, Hou C, Xing Y, Cao J, Wu K, Liu C, Zhang D, Zhang L, Zhang Y, et al (2014) Genome-wide detection of copy number variations among diverse horse breeds by array CGH. *PLoS One* 9: e86860. doi:10.1371/journal.pone.0086860

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I (2015) LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience* 4: 35. doi:10.1186/s13742-015-0076-3

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, et al (2017) A new chicken genome assembly provides insight into avian genome structure. *G3 (Bethesda)* 7: 109–117. doi:10.1534/g3.116.035923

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35: 543–548. doi:10.1093/molbev/msx319

Weisenfeld NI, Yin S, Sharpe T, Lau B, Ryan H, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, et al (2014) Comprehensive variation discovery in single human genomes. *Nat Genet* 46: 1350–1355. doi:10.1038/ng.3121

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB (2017) Direct determination of diploid genome sequences. *Genome Res* 27: 757–767. doi:10.1101/gr.214874.116

Weldenegodguad M, Pokharel K, Ming Y, Honkatukia M, Peippo J, Reilas T, Røed KH, Kantanen J (2020) Genome sequence and comparative analysis of reindeer (Rangifer tarandus) in Northern Eurasia. *Sci Rep* 10: 8980. doi:10.1038/s41598-020-65487-y

Wellenreuther M, Mérot C, Berdan E, Bernatchez L (2019) Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol Ecol* 28: 1203–1209. doi:10.1111/mec.15066

Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21: 1859–1875. doi:10.1093/bioinformatics/bti310

Yang L, Xu L, Zhou Y, Liu M, Wang L, Kijas JW, Zhang H, Li L, Liu GE (2018) Diversity of copy number variation in a worldwide population of sheep. *Genomics* 110: 143–148. doi:10.1016/j.ygeno.2017.09.005

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A* 110: E1743–E1751. doi:10.1073/pnas.1219381110

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 34: 303–311. doi:10.1038/nbt.3432