Graph construction method impacts variation representation and analyses in a bovine super-pangenome

Alexander S. Leonard[#], Danang Crysnanto, Xena M. Mapel, Meenu Bhati, Hubert Pausch[#]
Animal Genomics, ETH Zurich, Universitaetstrasse 2, 8092 Zurich
[#] Corresponding authors (hubert.pausch@usys.ethz.ch, alexander.leonard@usys.ethz.ch)

Abstract

Several models and algorithms have been proposed to build pangenomes from multiple input assemblies, but their impact on variant representation, and consequently downstream analyses, is largely unknown. We create multi-species "super-pangenomes" using pggb, cactus, and minigraph with the *Bos taurus taurus* reference sequence and eleven haplotype-resolved assemblies from taurine and indicine cattle, bison, yak, and gaur. We recover 232k nonredundant structural variations (SVs) from the pangenomes, of which 134k (58%) are common to all three, although the overlap is higher (60%) in non-repetitive regions. SVs derived from assembly-based calling show high agreement with the consensus calls from the pangenomes (96%), but validate only a small proportion of variations private to each graph. Pggb and cactus, which also incorporate base-level variation, have approximately 95% exact matches with assembly-derived small variant calls, which significantly improves the edit rate when realigning assemblies compared to minigraph. We use the three pangenomes to investigate variable number tandem repeats (VNTRs), finding 90% concordance in predicted repeat counts amongst them, although minigraph can over or underestimate the count given its more limited coordinate system. We examine a highly variable VNTR locus that is concordant in all three pangenomes and show that repeat unit copy number impacts expression of proximal genes and non-coding RNA. Our findings indicate good consensus between the three pangenome methods but also show their individual strengths and weaknesses that need to be considered when analysing different types of variants from multiple input assemblies.

Key words

Bovinae, graph pangenome, long sequencing reads, genome assembly, structural variation, domestic animals, VNTR profiling

Background

Pangenomes store and represent sequences from multiple individuals and enable unbiased variation-aware sequence variant analyses [1]. Graph pangenomes represent alleles that differ between the input assemblies as

nodes (representing sequence) connected by edges [2]. Several methods have been proposed to construct graph pangenomes from genome-scale data. For instance, minigraph applies approximate mapping to construct structural variant-based pangenomes from multiple input assemblies [3]. The reference backbone has an impact on these pangenomes because it propagates bias when large segments are missing in the backbone assembly [4]. Recently, cactus [5] and pggb (pangenome graph builder) [6] have been proposed to construct pangenomes from multiple input assemblies using reference-free base-level alignment. The pggb and cactus pangenomes contain all types of differences found between the assemblies, ranging from single nucleotide to large structural differences with nested variation [2].

Advancements in long-read sequencing and algorithms enable automated assembly of reference-quality genomes also for species with gigabase-sized genomes and so request for pangenomes that seamlessly accommodate and represent an increasing number of genomic resources [7],[8]. For instance, the Telomere-to-Telomere (T2T) consortium recently reported on a first complete human genome assembly [9] but routine near-T2T assembly [10] is becoming increasingly possible in humans [11] and other vertebrate species [12]. The Human Pangenome Reference Consortium (HPRC) coordinates global sequencing and assembly efforts with the goal to build a community-accepted variation-aware pangenome that integrates multiple input assemblies and represents global human genomic diversity [8]. An initial analysis of pangenomes constructed from 47 phased diploid genome assemblies has recently been reported by the HPRC [13].

Likewise, a Bovine Pangenome Consortium (BPC) has been formed to coordinate assembly efforts for the global cattle genomics community (https://bovinepangenome.github.io/). High nucleotide diversity and the separation of millions of global cattle into several hundred distinct breeds with unique genetic features, as well as frequent hybridization with their undomesticated relatives, make cattle an appealing species to improve assembly techniques [14] and investigate pangenome construction [4]. Bovine pangenomes constructed from multiple reference-quality assemblies revealed that the linear *Bos taurus* reference sequence lacks millions of bases that are accessible in assemblies from other individuals [4],[15],[7]. However, the impacts of different construction methods on pangenome profiles, variant representation, and downstream analyses are more uncertain particularly when they include assemblies from multiple species.

Here, we apply cactus, minigraph, and pggb to build pangenomes with the *Bos taurus* reference sequence and eleven haplotype-resolved assemblies from taurine and indicine cattle, bison, yak, and gaur. We assess the properties of the resulting pangenomes, recover large and small variants, and investigate how the pangenomes represent different types of DNA variation. We then profile variable number random repeats (VNTR) and investigate how pangenomes integrate DNA variants that are challenging to resolve from linear alignments. Finally, we exploit phylogenetic relationships between the input assemblies to identify a highly polymorphic VNTR locus which mediates the expression of proximal genes and non-coding RNA.

Results

We built pangenomes from autosomal sequences of domestic cattle and three of their wild relatives using minigraph, cactus, and pggb to assess how different graph pangenomes represent the same underlying input sequences. Eleven haplotype-resolved assemblies from eight breeds of taurine (*Bos taurus taurus*) and indicine (*Bos taurus indicus*) cattle, yak (*Bos grunniens*), bison (*Bison bison bison*), and gaur (*Bos gaurus*) as well as the current Hereford-based *Bos taurus* reference genome sequence were considered (**Figure 1a**). These high quality genomes were assembled using different sequencing and algorithmic approaches, but this has limited effect on the pangenomes [7]. However, the larger amount of centromeric sequence in the HiFi-based assemblies (**Figure 1b**) does require additional consideration. Minigraph used ARS-UCD1.2 as a backbone, whereas pggb and cactus are reference-free methods, although cactus is guided by expected ancestry (black subset of **Figure 1c**).

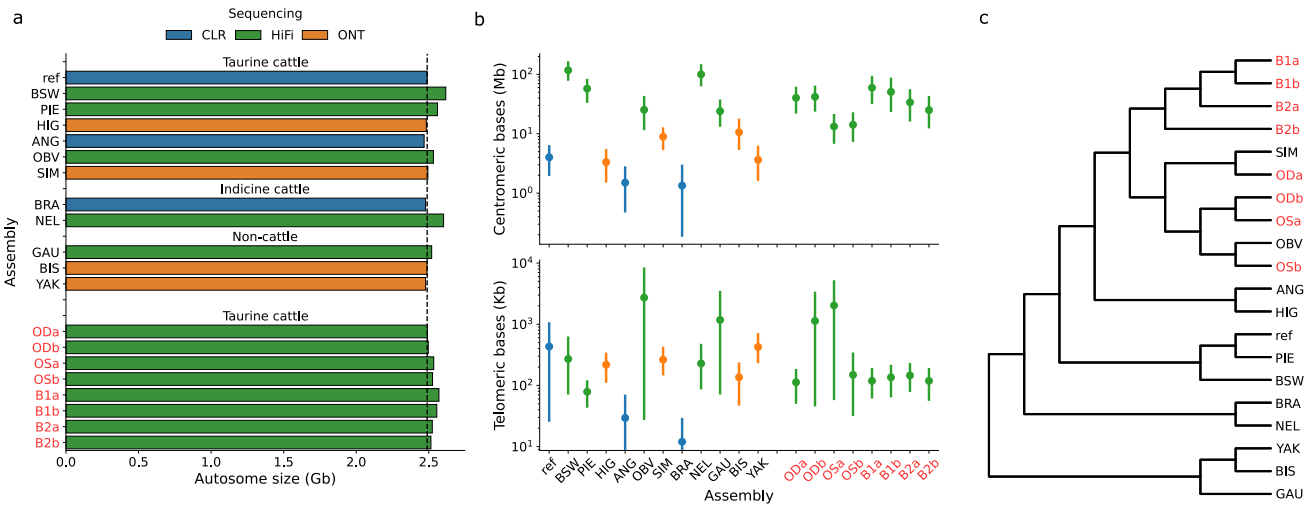

Figure 1: Input assemblies considered for the pangenome analyses. a) Autosomal length of the *Bos taurus taurus* reference sequence (ARS-UCD1.2) and 19 haplotype resolved assemblies considered to investigate pangenome construction and downstream analyses. The three pangenomes were created with ARS-UCD1.2 [ref] and eleven input assemblies (Brown Swiss [BSW], Piedmontese [PIE], Highland [HIG], Angus [ANG], Original Braunvieh [OBV], Simmental [SIM], Brahman [BRA], Nellore [NEL], gaur [GAU], bison [BIS], yak [YAK]), whereas eight additional Original Braunvieh or Brown Swiss assemblies indicated with red text (ODa, ODb, OSa, OSb, B1a, B1b, B2a, B2b) were only considered for downstream analyses. The colour of the bars indicates the primary sequencing technology used to construct the assemblies. The black dashed line indicates the length of ARS-UCD1.2. b) Centromeric and telomeric completeness is generally higher in the HiFi- and ONT-based input assemblies. The marker is the sum over the autosomes and error bars indicate the 95% confidence interval. c) A tree constructed from the assemblies reveals the expected separation between taurine, indicine and non-cattle.

Pangenome construction and sequence content

The three pangenomes spanned between 426k and 196M nodes that contained between 2.6 and 3.0 gigabases (**Table 1**). While cactus and pggb produced pangenomes with similar profiles, the pangenome built by minigraph differed greatly. Minigraph produced by far the smallest pangenome due to including primarily

larger (>50 bp) DNA variations. The non-reference nodes contain nearly five times as much sequence in cactus (530 Mb) and pggb (521 Mb) than minigraph (109 Mb). Pangenome construction took 34- and 287-times more CPU hours and required 6- and 23-times more memory for cactus and pggb than minigraph.

Table 1. Profiles of three bovine pangenomes constructed from autosomal sequences of twelve assemblies.

Parameter	Unit	minigraph	cactus	pggb
Nodes	N	426,012	195,851,069	178,850,394
Edges	N	681,496	267,961,459	243,812,854
Reference nodes	N	282,596	124,919,437	117,094,521
Non-reference nodes	N	143,416	70,931,632	61,755,873
Node length	bp	2,598,135,842	3,008,608,67	2,970,201,71
Length of non-reference nodes	bp	108,750,063	530,176,111	520,841,417
Length of small (<50 bp) non-reference nodes	bp	344,435	103,915,231	84,938,559
Length of large (>50 bp) non-reference nodes	bp	108,405,628	426,260,880	435,902,858
CPU time	h	17	570	4,892
Max memory	Gb	7	40	160

The large non-reference nodes contain mostly repetitive sequences, but their proportion differs between the three pangenomes (**Supplementary Figure 1**). Non-reference nodes in minigraph contain less repetitive bases (77.92 Mb, 72.14%) than in cactus (356 Mb, 83.56%) and pggb (384 Mb, 88.27%). The non-reference nodes in pggb and cactus contain 243 and 263 Mb centromeric repeats which are missing in ARS-UCD1.2 but are included in the HiFi-based (Brown Swiss, Nellore, Original Braunvieh, Piedmontese, and gaur) and to a lesser extent in ONT-based (bison and Simmental) assemblies. Given this sequence is not present in the reference, it is also not present in minigraph pangenomes, accounting for almost half of the large non-reference sequence present in cactus and pggb but not in minigraph.

Consensus of genomic variation

To assess the commonality of genomic variation across pggb, cactus, and minigraph, as well as the suitability of their representation for downstream analyses, we decomposed and normalized the graph pangenomes into VCF files with respect to the ARS-UCD1.2 reference. Decomposing default minigraph pangenomes failed to output many expected structural variant (SV) alleles, particularly deletions in multi-node bubbles (**Supplementary Figure 2**). By adding path information (P-lines in GFA) through minigraph-based realignment, we recovered an additional 9693 SVs (5.87% increase). Realignment occasionally produces paths incongruent with graph topology, where an assembly does not trace through nodes originating from that same assembly, but only affects approximately 0.18%, 0.51%, and 0.86% of taurine, indicine, and non-cattle nodes (**Supplementary Figure 3**). Cactus had the most genotypes marked as CONFLICT (2.5 %), indicating a

haploid assembly would have multiple ambiguous alleles, followed closely by pggb (2%) with minigraph significantly lower (0.5%). Such conflicts occurred more often in SV than small variation alleles, and were more frequent in divergent assemblies particularly in cactus (**Supplementary Figure 4**). VCF decomposition took 71- and 168-times longer for cactus and pggb compared to minigraph respectively, using approximately 34- and 41-times more memory (**Supplementary Table 1**), although pggb and cactus also contain substantial small variation to process.

The merging approach has an influence on the overlap of SVs between the three pangenomes. Merging based on variant breakpoints and length identified 232k nonredundant SVs larger than 50 bp from the three pangenomes of which 134k (57.55%) were common (**Figure 2a**). Cactus contained 185k SVs of which 32k (17.17%) were private, i.e. not present in minigraph or pggb. We recovered less SVs with pggb (175k) and minigraph (170k) of which 23k (13.06%) and 12k (7.01%) were respectively private. The overlap between the three pangenomes was higher (60.18%) for 111k SVs in non-repetitive regions, which span approximately 1.2 Gb (50.73%) of the ARS-UCD1.2 sequence. Consequently, the overlap was lower (53.41%) for 121k SVs in repetitive regions (statistically significant, df=28, P value=8.50e-05). Applying stricter requirements to merge SVs across the pangenomes, requiring variants to be within a fixed distance, increased the number of private SVs in pggb and cactus and so reduced the number of SVs that are common to the three tools (**Supplementary Figure 5**).

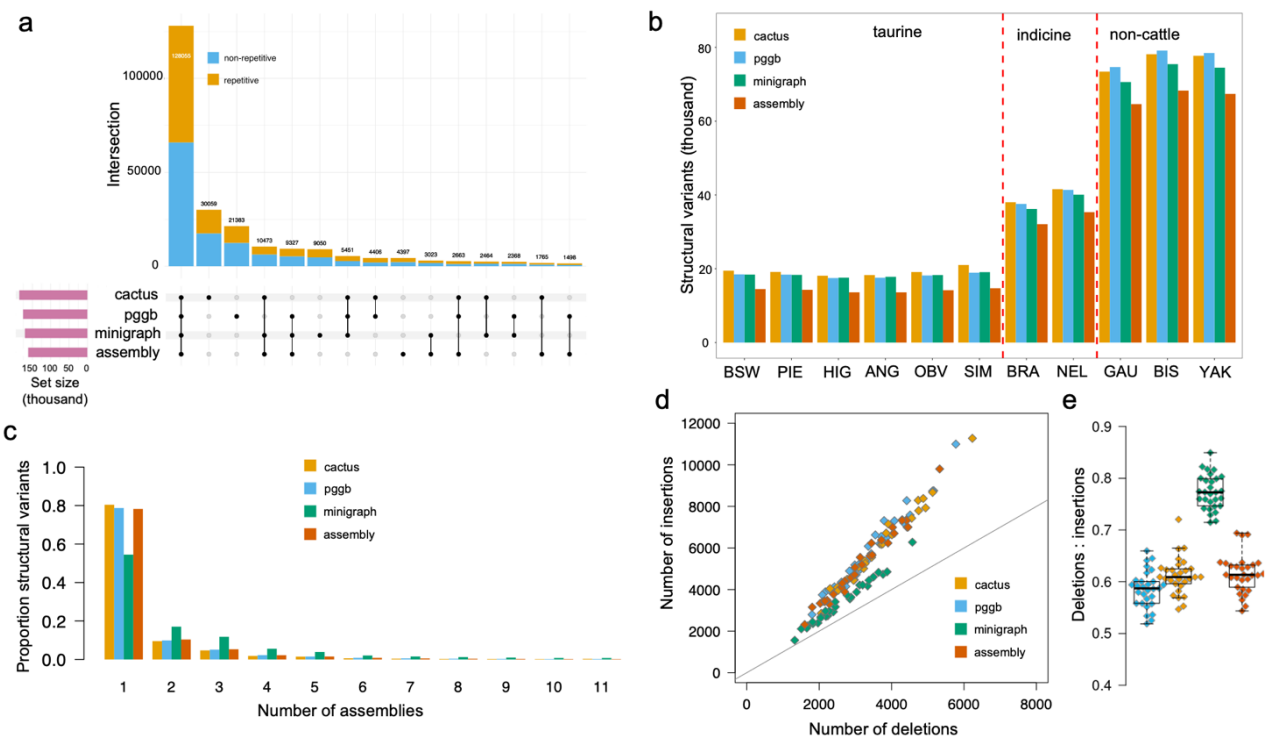

Figure 2: Consensus of structural variation between three pangenomes graphs. a) Overlap of SVs between three pangenomes and assembly-based SV discovery. Overall bar height represents the total number of SVs. Orange colour represents SVs in repetitive regions. b) Number of SVs identified in the input assemblies through assembly-based mapping and the three pangenomes. c) Allele frequency of the SVs. d) and e) Number of and ratio between deletions and insertions recovered from the pangenomes and assembly-based mapping.

A benchmark dataset to validate SVs is not available for Bovinae, and so we assessed the SV representation accuracy in the three pangenomes with SVs called directly from the same set of assemblies. We identified between 13.6k and 68.3k SVs in the haplotype-resolved assemblies but found substantially more SVs in the gaur, yak, and bison than the indicine and taurine haplotypes (**Figure 2b**). More than three quarter (78.25%) of the SV alleles were observed in only one haplotype assembly. The eleven haplotype assemblies contained 161k nonredundant SVs of which 151k (93.60%), 142k (87.81%) and 143k (88.68%) were also recovered with minigraph, pggb and cactus, respectively. This approach validated the vast majority (N=128k, 95.91%) of the 134k SVs that were found in all three pangenomes. However, it validated only 5.55%, 6.55% and 25.04% of the SVs that were respectively private to cactus, pggb and minigraph. Similar alleles are collapsed more frequently in minigraph than pggb and cactus, and so the allele frequency spectrum differs between the three tools with minigraph containing substantially less singleton SVs (**Figure 2c**). While all tools recovered more insertions than deletions, minigraph contained proportionally less insertions than pggb and cactus (**Figure 2d,e**).

We used optical mapping from two unrelated Hereford and two unrelated Nellore samples [16] to further validate SV consensus in the pangenomes. The overlap is expected to be low due to the unrelated origin of samples, only covering two of the twelve breeds, and the different variant sizes accessible with optical mapping and sequencing reads. Minigraph had the largest number of validated SVs (2,604, 1.56%), followed closely by assembly (2,563, 1.61%) whereas cactus (2,477, 1.39%), and pggb (2,254, 1.33%) had lower support (**Supplementary Figure 6**). Out of the pangenomes, minigraph also had the most private variants validated (58), compared to cactus (48) and pggb (36).

We further assessed the communality of small variation (SNPs and indels <50 bp) in pggb and cactus, as minigraph primarily only represents SVs. We recovered 65.13 M nonredundant small variants from the two pangenomes. Pggb and cactus contained 59.85 M (52.67 M SNPs / 7.18 M indels) and 61.93 M (54.55 M SNPs / 7.38 M indels) small variants of which 56.65 M (51.63 M SNPs / 5.02 M indels) were common. The overlap between pggb and cactus was substantially larger for SNPs than indels (**Figure 3a-c)**.

We assessed small variation representation accuracy in the two pangenomes again using assembly-derived calls as an approximate truth set. The eleven input assemblies contained 61.68 M (54.20 M SNPs / 7.48 M indels) nonredundant small variations of which 92.07% and 92.91% overlapped with pggb and cactus, respectively. Each assembly had between 5.12 and 27.04 M small variations, but we found substantially fewer in the taurine assemblies than their indicine and non-cattle counterparts (**Figure 3d**). Between 93% and 97% of the small variants called from the assemblies were exactly matched to those called from the pangenomes, but the overlap was substantially higher for SNPs than indels for both pggb and cactus (**Figure 3e**). Cactus

contained more SNPs that overlapped with assembly-based calls than pggb (Paired t test, df=318, p=1.35e-70), but less indels (Paired t-test, df=318, p=1.24e-28).

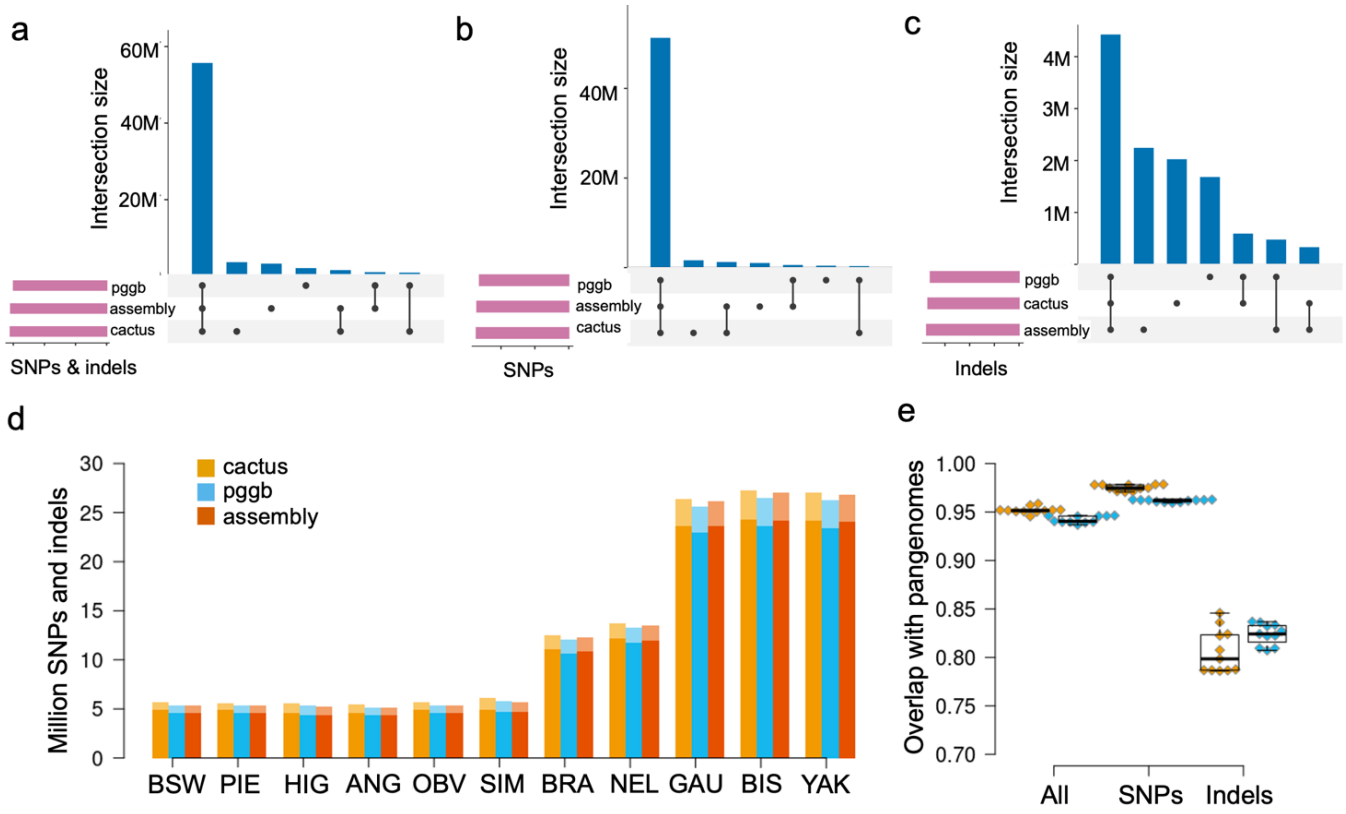

Figure 3: Overlap of SNPs and indels between pangenome- and assembly-based discovery. a-c) Overlap of small variations (SNPs and indels smaller then 50 bp) between three pangenomes and assembly-based SV discovery presented for all small variations (a), SNPs (b) and Indels (c). d) Number of small variations recovered from each haplotype from the pangenome and the assembly-based mapping. The faded and dark areas of the bars represent indels and SNPs, respectively. (e) Proportion of SNPs and indels discovered from the assembly-based mapping that overlap those discovered from the pangenomes. Each point represents a value from a single sample.

Alignment of assemblies to pangenomes

We realigned all twelve input assemblies (including ARS-UCD1.2) against the pangenomes to calculate edit distances and quantify sequence content of the pangenomes. Since centromeric, and to a lesser extent telomeric, sequence is challenging to align, we analysed these "centro-/telomeric" regions separately to the rest of the “bulk” sequence (**Supplementary Table 2**). Minigraph had a substantially higher edit rate (0.618%) than pggb (0.006%) and cactus (0.001%) in bulk regions and also in centro-/telomeric regions, 3.50%, 0.10%, and 0.33% respectively (**Figure 4a**). Even though cactus includes more centromeric sequence in the graph (**Supplementary Figure 1**) and had overall lower edit rates, its centro-/telomeric edit rates were triple that of pggb. Minigraph also had a significantly lower edit rate for ARS-UCD1.2, as all reference sequence is included by minigraph’s definition.

Edit rates were higher for assemblies more diverged from the reference backbone in minigraph pangenomes, indicating they do not incorporate highly diverged segments containing multiple small variations but were not large enough to form bubbles. Contrastingly, pggb and cactus pangenomes did not have any divergence bias for edit rate, suggesting they may be more suited to include more divergent assemblies into "super-pangenomes" [17]. However, there was a weak increase in edit rate in pggb and cactus pangenomes for HiFi assemblies, potentially related to the increased resolution of complex and repetitive areas beyond centro-/telomeric sequence that cannot be realigned accurately.

We also aligned eight additional assemblies held out from the pangenomes (2x haplotypes from each parent of the OBV and 2x haplotypes for two unrelated Brown Swiss cattle [**Figure 1**]), and found highly similar edit rates for minigraph, but significantly higher edit rates for pggb and cactus (**Figure 4b**). This is expected and reflects true genomic variation in the held-out assemblies not present in the pggb/cactus pangenomes, while minigraph never captured small variation anyway and had higher edit rates to begin with.

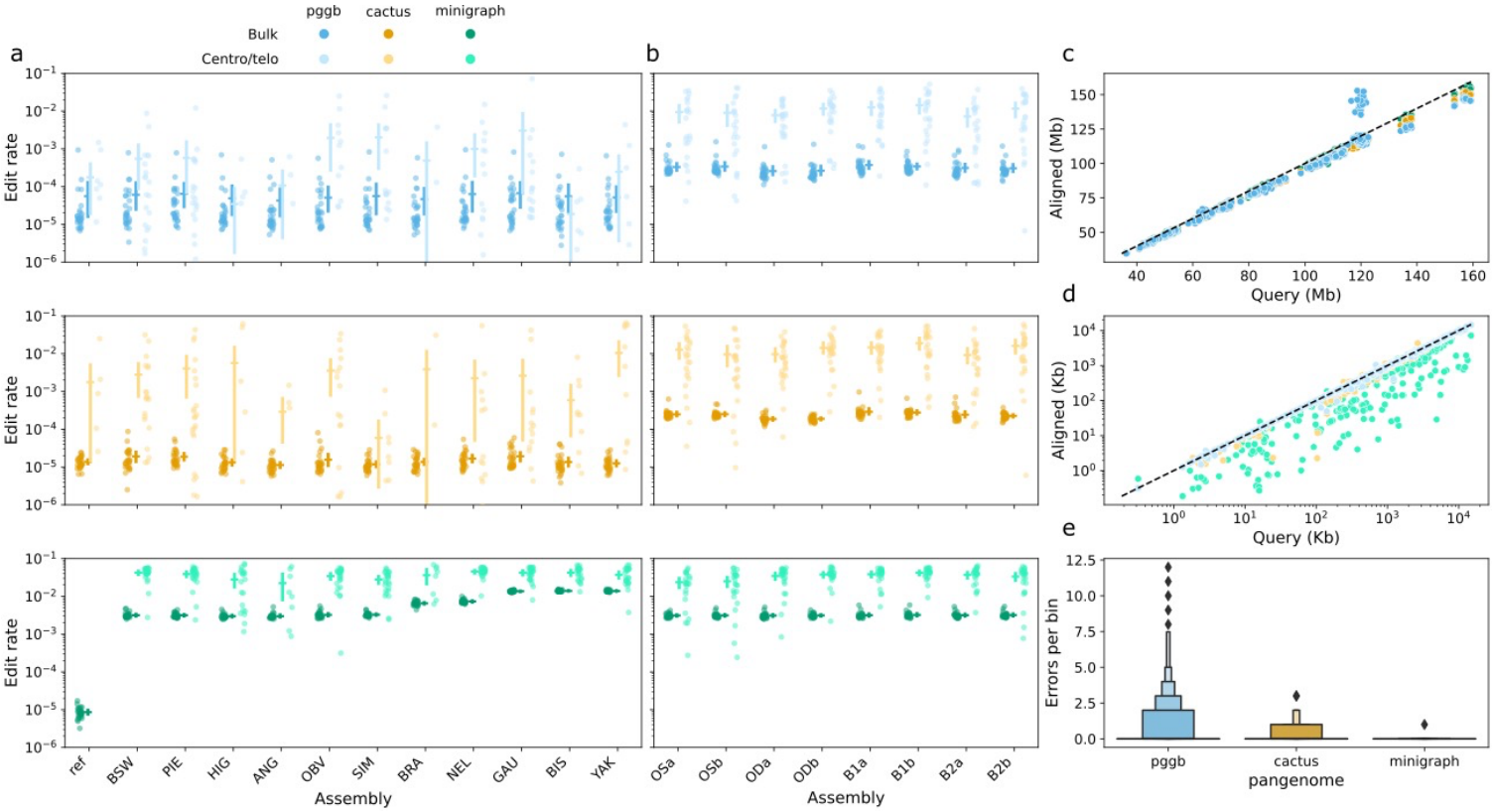

Figure 4: Realignment of included and held-out assemblies to the pangenomes. a) Edit rate of the twelve assemblies used in pangenome construction. Each faded dot represents 100 Kb bins of each assembly in either the bulk or centro-/telomeric ranges. Mean and 95% confidence intervals are also indicated for each category of sequence per assembly. b) Similar plot to (a) but for eight additional assemblies held-out from pangenome construction. c) Query coverage for the bulk sequence for each autosome per assembly. Values above the dashed line indicate there were more aligned bases than query bases, suggesting multiple ambiguous alignments that were equally scored. d) Similar to (c), but for the centro-/telomeric sequence. d) Letter-value plot for the number of alignment errors reported by GraphAligner when using finite "tangle effort" in the bulk regions per chromosome for all 20 assemblies. assembly. Each box represents the proportion of data above that value, with diamonds representing the top 5% extreme values.

Although pggb and cactus had substantially lower edit rates, they also covered less of the bulk query sequence (97.1% and 96.2% respectively) compared to minigraph (97.5%). However, they covered more of the centro/telomeric query sequence (98.0%, 95.5%, and 66.5% respectively), as expected given the relative lack of centromeric sequence in minigraph pangenomes (**Figure 4c**). Some highly complex chromosomes (e.g. 5, 6, and 19) had many sequences aligning multiple times in pggb pangenomes, leading to >100% query coverage. More query sequence can be aligned with more sensitive alignment parameters, but this can require >300 GB of memory per chromosome in pggb and cactus graphs (S**upplementary Table 3**). Even with relaxed alignment parameters, aligning to cactus and pggb pangenomes took 2x and 92x more CPU time and 5x and 9x more memory than minigraph pangenomes (**Supplementary Table 4**). There was again a noticeable increase in compute resources needed to align additional assemblies to pggb and cactus pangenomes, while no increase was observed for minigraph. Failed alignment of 100 Kb segments was also significantly more common in pggb than cactus or minigraph, generally resulting from exceeding GraphAligner's "tangle effort" in highly complex regions (**Figure 4e**), respectively 0.99, 0.32, and 0.02 per chromosome per assembly.

Pangenome resolution of VNTR

Variable number tandem repeats (VNTR) account for substantial gene expression and complex trait variation, but due to their repetitive nature and high mutation rates these loci are difficult to resolve from linear alignments [18],[19]. A catalogue of bovine VNTR had not been established and so we identified 18644 tandem repeats (TRs) in non-masked regions of the ARS-UCD1.2 reference sequence (**Supplementary Table 5**) and investigated their prevalence in the other assemblies through the three pangenomes.

As our interest was in VNTRs, we pre-filtered this set by overlapping TRs with SVs identified through minigraph. An equivalent filtering step is more challenging in pggb and cactus due to the inclusion of smaller variation and less accessible encoding of variation, and so we examined the entire set of TRs for these pangenomes. Minigraph does not store a coordinate system for all input assemblies, so we could only approximate start/end coordinates of TRs at graph bubbles. For pggb and cactus, odgi can extract ("liftover") coordinates into any chosen assembly, resulting in more accurate TR regions for downstream analysis. However, this comes at a significantly greater compute cost (**Table 2**). Approximately 95% of TRs had all twelve assemblies associated with a pangenome path, while the remaining TRs suffered from reduced mappability (**Table 2**). All pangenomes had several instances where coordinates were erroneously translated to other assemblies, although this was most commonly observed in cactus (**Supplementary Table 6**). Nearly all TRs examined by minigraph were VNTRs, while nearly 20% of TRs examined by pggb and cactus has zero variation between samples, hence were not VNTRs (**Table 2**).

Table 2: All TRs were examined from pggb and cactus, while TRs overlapping SVs were examined for minigraph. Genotypable TRs are TRs where all 12 assemblies had a known path through the local graph structure. VNTRs are genotypable TRs that had at least one sample with a different number of TR counts. CPU and Memory indicate the compute resources needed to identify the paths of all assemblies through all examined TRs.

Pangenome	TRs examined	Genotypable TRs	VNTRs*	CPU (h)	Memory (Gb)
Pggb	18644	17540 (94.1%)	12048 (80.6%)	256.82	14.7 Gb
Cactus	18644	17781 (95.4%)	12219 (80.9%)	196.53	13.9 Gb
Minigraph	11695	11289 (96.5%)	9689 (99.7%)	13.94	6.5 Gb

*2692, 2752, and 1629 TRs for pggb, cactus, and minigraph could not be genotyped within the allocated runtime, and are excluded from the percentage calculation.

Tandem repeat counts were highly similar between the three pangenomes, with 8,203 identical counts across all 12 assemblies out of 9,662 common genotyped VNTRs (84.9%, **Figure 5a**). The remaining VNTR counts still broadly agreed with minor variability (**Supplementary Figure 7**). The average squared Spearman correlation was 0.98 between the pangenomes, with several yak outliers in cactus skewing the squared Pearson correlation, which was otherwise around an average of 0.9 (**Figure 5b**, all significant). TRs present in only pggb and cactus had low variability between cattle and non-cattle (**Figure 5c**), suggesting minigraph efficiently captures nearly all VNTRs of interest for further investigation. We also genotyped 465 TRs with adVNTR [19] using HiFi reads from the gaur, Nellore, Piedmontese, Brown Swiss, and Original Braunvieh samples as an approximate truth set, finding good concordance (**Figure 5d**, **Supplementary Table 5**), with mean difference of counts of 16.3, 14.3, and 23.3 for pggb, cactus, and minigraph to adVNTR. Minigraph occasionally over- or underestimated TR counts if the overlapping SV was significantly larger or smaller, as that determines the translated coordinates. There were also instances where all three pangenomes agreed on counts different to adVNTR, indicating even HiFi reads may not be powerful enough to genotype all VNTRs (**Figure 5e**). Clustering based on TR counts produced trees that broadly grouped the taurine and indicine cattle as well as the non-cattle. However, the TR counts were likely too inaccurate or had too many outliers to correctly produce the exact expected tree topology (**Figure 1c**).

An eVNTR mediates expression of neighboring genes and non-coding RNA

A VNTR locus on chromosome 12 (86,161,072 - 86,162,351 bp) was highly variable, containing substantially more copies of a degenerate 12 bp motif in non-cattle and fewer in Nellore than in the taurine assemblies (**Figure 6a,b**), prompting a detailed investigation. Viewing this VNTR in a Bandage plot revealed different structure in the three pangenomes, as well as the motif variation present (**Figure 6c**, **Supplementary Figure 8**).

We observed two VNTR "haplotypes" (referred to as hap1 and hap2) in long read alignments of additional Original Braunvieh cattle demonstrating within-breed variability. While the alignments indicated several

insertions and deletions in both haplotypes with respect to ARS-UCD1.2, hap2 is 24 bp longer than hap1 due to two additional copies of the repeat motif (**Supplementary Figure 9**). There are 63 genes annotated within the *cis*-regulatory range (±1 Mb) of the VNTR, of which 44 (**Figure 6d**) were expressed at >0.2 TPM in testis tissues of 83 mature Brown Swiss and Original Braunvieh bulls.

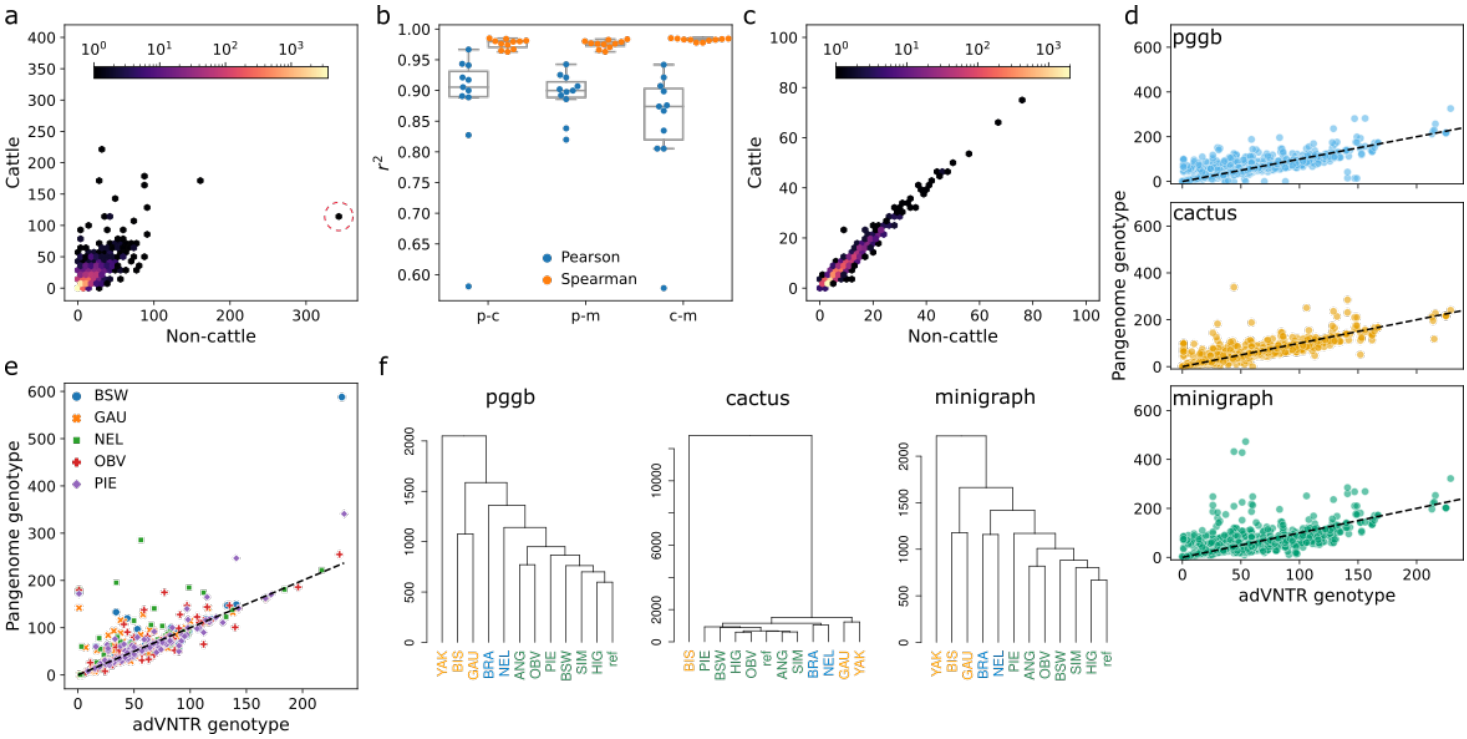

Figure 5: VNTR concordance in three pangenomes. a) TRs with identical counts between the three pangenome tools, with the median count for the cattle and non-cattle groups. A particular VNTR with substantially more repeats in non-cattle compared to cattle that we investigated further is circled in red. b) Pearson and Spearman squared correlation coefficients across the TR counts for pggb-cactus (p-c), pggb-minigraph (p-m), and cactus-minigraph (c-m). Each point is one assembly, with box plots over the 12 assemblies. c) Similar to (a), except TRs with identical counts in pggb and cactus that were not present in minigraph. d) adVNTR-derived genotypes for five HiFi samples in the three pangenomes. The black dashed line indicates the expected count using adVNTR as a ground truth. e) VNTRs where all three pangenomes agreed to a different count than adVNTR, suggesting adVNTR may sometimes over-/underestimate assembly-based counts. f) Trees derived from the TR counts across different input assemblies, with colours representing taurine cattle, indicine cattle, and non-cattle. A large outlier in a bison TR skews the tree for cactus.

A degenerate repeat motif and an overall length >1300 bp precluded the short sequencing read-based genotyping of the VNTR in the eQTL cohort with adVNTR. Instead, we genotyped the VNTR through two SNPs (Chr12:86160984 and Chr12:86160971) and one indel (Chr12:86161000) that tagged the two VNTR haplotypes. This enabled us to investigate putative *cis*- and *trans*-regulatory impacts of the VNTR on the expression of genes and long non-coding RNAs (lncRNA). Two genes (*TUBGCP3*, *SPACA7*) and five non-coding RNAs (*LOC112449093*, *LOC112449094*, *LOC112449095*, *LOC112449100*, *LOC112449104*) within ±1 Mb of the VNTR were differentially expressed (P<1.13e-3, Bonferroni-corrected significance threshold) between the diplotypes indicating a putative *cis*-regulatory role (**Figure 6e-k**). We did not detect *trans*-regulatory effects for the haplotype (**Supplementary Figure 10**).

We mapped eQTL within the *cis*-regulatory range (±1 Mb of the transcription start site) of the two significant genes and five significant non-coding RNAs to investigate if variants in *cis* other than the VNTR haplotype are associated with transcript abundance. The VNTR haplotype was among the top variants at the *LOC112449094* eQTL, but eleven variants in strong linkage disequilibrium (r^2= 0.93) were more significant (P=2.00e-22 *vs.* P=5.18e-20) (**Figure 6l**). The eQTL peak was absent when the VNTR haplotype was fitted as a covariate. We made similar observations for the *TUBGCP3* eQTL (**Supplementary Figure 11**). Variants other than the VNTR haplotype were more strongly associated with the expression of *SPACA7*, *LOC112449093, LOC112449095*, *LOC112449100*, and *LOC112449104* (**Supplementary Figure 11**). These findings suggest that the VNTR haplotype primarily mediates the expression of *LOC112449094* and *TUBGCP3*.

Hap2 containing more copies of the 12 bp motif increases abundance ($\beta_{normalized}$=1.26 ± 0.10, P=5.18e-20) of *LOC112449094* which is lowly-expressed (0.86 ± 0.31 TPM) in 83 bull testis transcriptomes. Conversely, hap2 is associated ($\beta_{normalized}$= -1.08 ± 0.14, P=3.81e-11) with lower *TUBGCP3* mRNA. Negative correlation (–0.37, **Figure 6n**) between *LOC112449094* and *TUBGCP3* abundance suggests that *LOC112449094* could be a cis-acting lncRNA that represses expression of *TUBGCP3*. The VNTR haplotype is in strong LD with two SNPs (Chr12:86173431 (r^2=0.87) and Chr12:86173509 (r^2=0.93)) in *LOC112449094* exons. While hap2 primarily segregates with the alternate alleles of these two SNPs, hap1 segregates with the respective reference alleles. Alternate allele support in RNA sequencing reads overlapping Chr12:86173431 and Chr12:86173509, respectively, was 71% (96 out of 135 alleles, P_{binom}=1.01e-06) and 69% (94 out of 137 alleles, P_{binom}=1.57e-05) in 22 animals that are heterozygous both at the VNTR haplotype and the SNPs confirming allelic imbalance due to lower *LOC112449094* expression in hap1.

We confirmed a putative *cis*-regulatory role of the VNTR in an individual with indicine ancestry. Expression of *LOC112449094* was relatively low (TPM=0.66) in testis tissue of a Nellore (*Bos taurus indicus*) x Brown Swiss (*Bos taurus taurus*) crossbred bull (NxB), although it inherited hap2 from its Brown Swiss dam (**Supplementary Figure 9**). The paternal (Nellore) haplotype is diverged from hap2 and hap1, and contains less VNTR repeat units (F**igure 6b**). Alignment of parental-binned HiFi reads against ARS-UCD1.2 readily distinguished paternal from maternal alleles at nine heterozygous SNPs overlapping *LOC112449094* exons. Inspection of these variants in the RNA sequencing alignments of NxB F1 showed allelic imbalance due to a disproportionally low number or even complete absence of paternal (=Nellore) alleles (**Figure 6m**) suggesting that the low VNTR repeat count of the Nellore haplotype represses expression of *LOC112449094*.

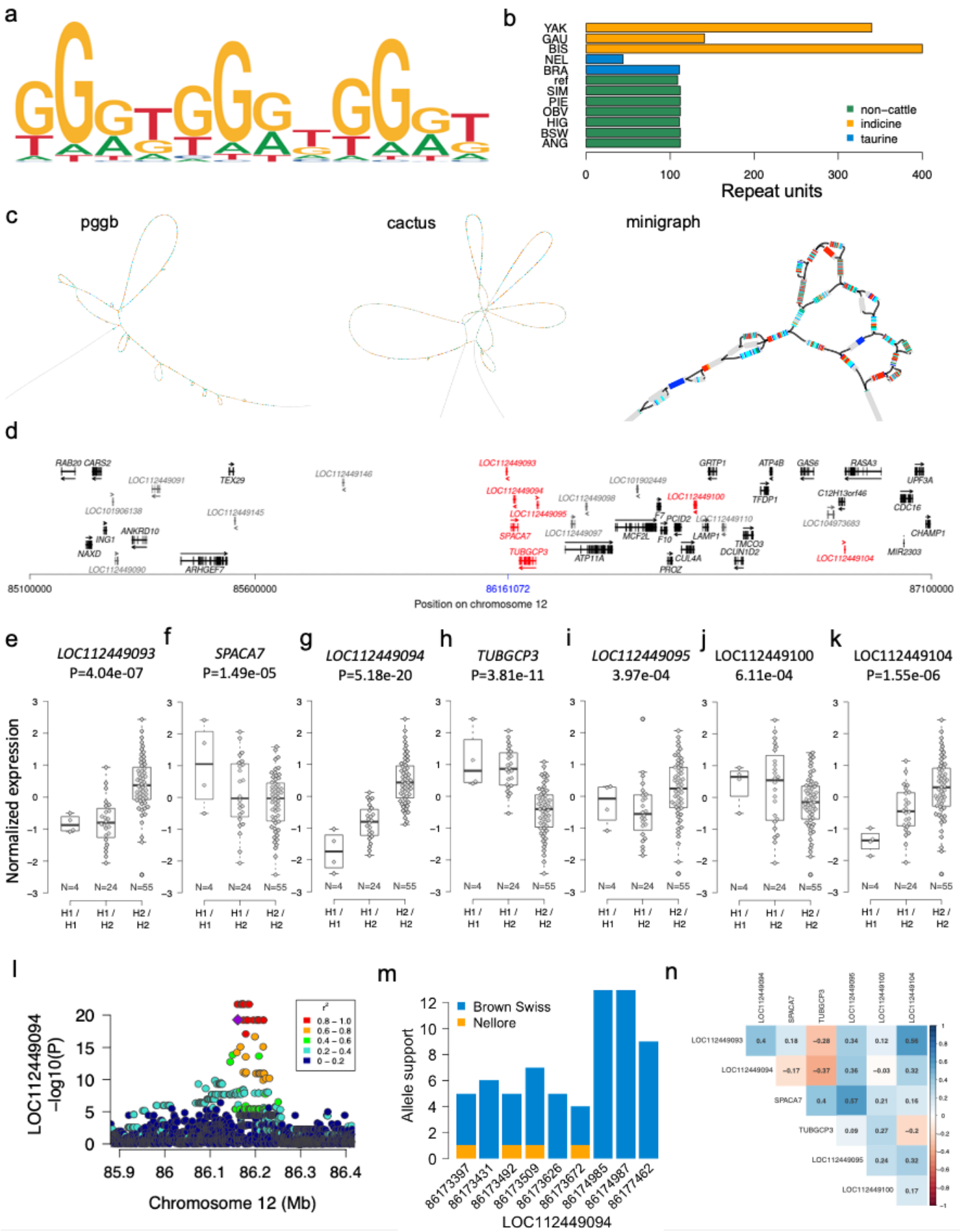

Figure 6: **A polymorphic eVNTR discovered from the pangenomes**. a) VNTR repeat motif and b) number of repeat units in the twelve input assemblies. c) Bandage plots of a VNTR upstream *SPACA7*. BLAST hits for the top 5 most common VNTR motifs are coloured per motif. d) Genes (black) and (long) non-coding RNAs (dark grey) nearby the VNTR. Blue colour indicates the position of the VNTR. Arrows indicate the orientation of genes and lncRNA. Red colour indicates two genes and five lncRNAs

whose expression is associated with the VNTR. e-k) Expression (quantified in transcripts per million (TPM)) of the associated genes and transcripts in testis tissues of 83 Brown Swiss bulls that are either homozygous for hap1 (H1 / H1), homozygous for hap2 (H2 / H2) or heterozygous (H1 / H2). The number of animals per diplotype is below the boxplots. l) Cis-expression QTL mapping for *LOC112449094*. Different colours indicate the pairwise linkage disequilibrium (r^2) between the VNTR haplotype (violet) and all other variants. m) Allelic imbalance of nine heterozygous exonic SNPs in *LOC112449094* in testis tissue of a Nellore x Brown Swiss crossbred bull. Orange and blue colours represent paternal (Nellore) and maternal alleles (Brown Swiss). n) Correlation between the abundance of mRNA and lncRNA for the associated genes and lncRNA.

Discussion

We constructed pangenomes using three different algorithmic approaches with minigraph, pggb, and cactus, finding each have different strengths and weaknesses. Pggb and cactus include all variation, down to SNPs, while minigraph primarily captures SVs. Consequently, pggb and cactus pangenomes had roughly 400 times more nodes and edges, 10 times the file size, and construction required 288 and 34 times as much CPU time as the minigraph pangenomes respectively, despite only including about 5 times as much non-reference sequence. High-quality assemblies are becoming increasingly easier and cheaper to produce, and so expanding pangenomes with additional assemblies will likely be common. Minigraph can iteratively add assemblies in near-linear time and memory to an existing pangenome, while cactus requires storage of modestly sized intermediate files (10s of GB) and the complexity of adding a new assembly depends on the existing phylogenetic guide tree [5]. Pggb on the other hand, performs all-versus-all alignment, and so adding new assemblies would require building a new pangenome from scratch. As such, pggb may be suited to building reference pangenomes, but cactus and minigraph are more appropriate for projects with ongoing assembly efforts.

Regardless of SV preprocessing and merging requirements, cactus and pggb contain more SVs than minigraph, including 23k and 32k private SVs of which very few overlap with assembly-derived SVs, thus most likely represent false positives, agreeing with recent findings in human pangenomes [13]. Minigraph contains fewer private SVs, but nearly a quarter overlap with assembly SVs and may be true variation. We found that approximately 8.1% and 6.7% of SV alleles in pggb and cactus were removed during preliminary filtering when only keeping alleles seen in genotypes, suggesting variants may either exist in complex regions that cannot be genotyped confidently or strongly tangled regions that produce spurious variant calls. However, we found a higher overlap on small variations compared to structural variations, indicating that the cactus and pggb pangenomes can integrate small variations more accurately. These results demonstrate that SVs recovered from pangenome graphs, particularly pggb and cactus, must be stringently filtered before they qualify for downstream use in population genomic analyses.

Pggb and cactus are reference-free, while minigraph requires an initial backbone to determine variation from the input assemblies. Minigraph is thus susceptible to reference-bias, especially in the case of an incomplete reference (currently almost every reference genome except the T2T-CHM13 [9]). The ARS-UCD1.2 autosomes are interrupted by 257 gaps and contain almost no centro-/telomeric sequence, which prevents integration of centro-/telomeric sequence in the more complete HiFi assemblies into the minigraph pangenome. Pggb and cactus are largely able to integrate these sequences, with 243 and 263 Mb of centromeric sequence in the non-reference nodes of the pangenomes, respectively. However, cactus does use progressive alignments on the provided guide tree, and so does appear to have a slight reference-bias when projecting back into a reference coordinate system (like VCF). This is most apparent for Brown Swiss and Piedmontese, the two other samples on the ARS-UCD1.2 branch, which have noticeable differences compared to other taurine samples, which is not observed in pggb.

Pangenomes with base-level variation like pggb and cactus can realign most of the input assemblies with minimal edit rate, but centro-/telomeric regions are still poorly realigned. Equally, the base-level variation in these pangenomes can also lead to strongly tangled graph topology, which can generate poor realignments and substantially increase computational requirements. For super-pangenomes incorporating multiple (sub-)species, containing a greater amount of variation, downstream software like Bandage, GraphAligner, and vg quickly hit CPU and memory bottlenecks working with pggb or cactus, while minigraph pangenomes are usable even without HPC infrastructure. Although beyond its design intentions, minigraph can be run with a lower "minimum variant length" parameter, increasing its sensitivity to smaller variation down to approximately 10 bp without substantially impacting compute requirements for pangenome construction or analyses (**Supplementary Table 7**). However, this unsupported use-case cannot regularly capture SNPs and so pggb and cactus are vastly better for representation small variation accurately.

We show that the three pangenomes are generally concordant at highly polymorphic VNTR loci, although minigraph can over- or underestimate TR counts if the graph bubbles are significantly larger or smaller than the TR region, as minigraph can only translate assembly coordinates at these bubbles. All three pangenomes had several poorly genotyped VNTRs, due to incorrect coordinate extraction in these repetitive regions, again indicating unresolved challenges with complex alignments. Furthermore, minigraph primarily captures repeat count variation rather than motif variation in TRs, the latter of which may be of greater importance in identifying eVNTRs [20]. However, the small variation present in pggb and cactus hinders visual inspection and pangenome annotation with Bandage, while TRs are simple to identify in SV-oriented minigraph pangenomes. We also show that even highly accurate long reads can fail to successfully genotype VNTRs, supporting that assembly-based approaches are the most powerful resource for detecting large or complex variation. In the case of a highly variable eVNTR upstream *SPACA7*, adVNTR predicted 116 copies for gaur, whereas all three pangenomes identified 141 copies.

Conclusions

Bovine pangenomes have been utilized before to make non-reference sequences amenable to association testing and reveal trait-associated structural variants [4],[15],[7]. Our findings show good agreement between the SVs discovered from three widely used pangenome methods and so reinforce the utility of any pangenome approach to investigate variants that are difficult to resolve with a single linear reference. Pggb, and to a lesser extent cactus, pangenomes losslessly represent the input assemblies and are ideal for generating a pangenome reference containing all variation. However, minigraph has much greater utility, allowing simple expansion with additional assemblies and rapid downstream analyses with modest computational requirements. With the recent establishment of the cross-species Bovine Pangenome Consortium (https://bovinepangenome.github.io/) and similar efforts in plants [17],[21],[22], there is a need to emphasise pangenomes that can handle significantly higher levels of variation compared to the draft human pangenome reference [13].

Methods

Pangenome construction

Pangenomes were constructed per-chromosome for the 29 bovine autosomes, using the Hereford-based *Bos taurus* ARS-UCD1.2 reference genome [23] and eleven other haplotype-resolved assemblies: eight domestic cattle (*Bos taurus taurus* and *Bos taurus indicus*) and three wild relatives of domestic cattle (*Bos grunniens, Bos gaurus*, *Bison bison*) (**Table 3**). We used these assemblies to construct pangenomes with minigraph (v0.19) [3], the PanGenome Graph Builder (pggb, v0.3.0) [24], and the Cactus Progressive pangenome pipeline (cactus v2.0.5) [5].

The minigraph pangenome was constructed using base-level alignment (`-c`) and default parameters otherwise. The ARS-UCD1.2 reference genome was the backbone of the graph, and the other assemblies were added in order of their mash distance [28] to the reference: Brown Swiss, Piedmontese, Highland, Angus, Original Braunvieh, Simmental, Brahman, Nellore, gaur, bison, and yak.

Pggb was run with the recommended parameters (segment length of 100000 bp, identity mapping taken from the rounded lowest mash similarity of 80%, and target haplotype paths of 12) with all assemblies combined into a single fasta file. The output gfa file was then processed with GFAffix (https://github.com/marschall-lab/GFAffix) to remove redundant graph topologies.

Since our assemblies contain multiple species with significant divergence, we used cactus

with progressive alignment using the guide tree from the mash-based distance. Assemblies were first soft-masked with RepeatMasker using a database of repetitive DNA elements from Repbase (release 20181026). The cactus hierarchical alignment (HAL) was converted into GFA using hal2vg (https://github.com/ComparativeGenomicsToolkit/hal2vg) and vg convert [29].

Table 3: Input assemblies for the pangenomes. The bovine reference (Hereford, ARS-UCD1.2) is a primary assembly, a collapsed haploid representation of the diploid genome. All other assemblies are haplotype-resolved, a haploid genome of either the maternal or paternal haplotype. CLR: Pacific Bioscience Continuous Long Reads, HiFi: Pacific Bioscience High Fidelity Reads, ONT: Oxford Nanopore Technologies Reads.

Breed/species	Acronym	Primary sequencing strategy	Autosome length (Gb)	Reference
Hereford / *Bos taurus taurus*	UCD	CLR	2.489	[23]
Angus / *Bos taurus taurus*	ANG	CLR	2.468	[14]
Brahman / *Bos taurus indicus*	BRA	CLR	2.478	[14]
Highland / *Bos taurus taurus*	HIG	ONT	2.483	[25]
Yak / *Bos grunniens*	YAK	ONT	2.478	[25]
Simmental / *Bos taurus taurus*	SIM	ONT	2.494	[26]
Bison / *Bison bison bison*	BIS	ONT	2.488	[27]
Brown Swiss / *Bos taurus taurus*	BSW	HiFi	2.617	[7]
Nellore / *Bos taurus indicus*	NEL	HiFi	2.602	[7]
Piedmontese / *Bos taurus taurus*	PIE	HiFi	2.560	[7]
Gaur / *Bos gaurus*	GAU	HiFi	2.520	[7]
Original Braunvieh / *Bos taurus taurus*	OBV	HiFi	2.532	[7]

Analysis of pangenome sequence content

Repetitive elements were identified for the ARS-UCD1.2 reference sequence using RepeatMasker (http://www.repeatmasker.org) (version 4.1.1) and rmblast (version 2.10.0), using a modified version of the 2018 Repbase database. Structural variants overlapping repetitive elements were detected using BCFtools (version 1.16) [30].

Variation discovery from pangenomes

Pangenomes were decomposed into VCF files using vg deconstruct with --all-snarls --path-traversals --ploidy 1. We added path information (P-lines) to minigraph's gfa by manually curating the output of minigraph --call on each sample which retraces the assembly's path (available in the Github repository associated with this paper). Variants were then normalized using vcfwave [31] with -I 1000, and then split into small (<50 bp) and structural variation (>50 bp) using bcftools norm and bcftools view. Alleles not observed in any sample were dropped using --trim-alt-alleles during splitting, and similarly applied to per-sample VCF statistics. Multiple nucleotide polymorphisms (MNPs) were split into multiple SNPs using bcftools norm --atomize.

Variation discovery from assemblies

All non-reference assemblies were aligned to ARS-UCD1.2 using minimap2 (v2.24 [32]) with "-cx <preset>". The divergence-based preset was chosen as asm5, asm10, and asm20 for the taurine, indicine, and non-cattle, respectively. Variants were called from the alignments using paftools.js and normalized and subsetted using the same approach as pangenome variation.

Variation consensus

Normalized SVs from the three pangenomes and assemblies were merged using Jasmine (v.1.1.4 [33]) with --use_end --allow_intrasample. A lenient SV merging was done using parameters max_dist=1000 max_dist_linear=0.5. Impacts of stricter merging settings were explored using either max_dist=25 or max_dist_linear=0.5 min_seq_id=0.5.

Jasmine was also used for intersection SVs with optical mapping data, using the lenient parameters described above. The optical mapping data filtered to keep deletions, insertions, and duplications and removing SVs >1 Mb.

Small variation between individual haplotypes was intersected using bcftools isec, requiring at least one ALT allele to match with the --some flag. The normalized cohort-level pangenome and assembly VCF files were intersected based on exact matches of small variant coordinates and ALT alleles.

Calculation of edit distance

The twelve assemblies included in the pangenomes and eight haplotype assemblies from four taurine cattle which are not part of the pangenomes were realigned to the pangenomes using GraphAligner (v1.0.16, [34]) with parameters -x dbg -C 750000 --max-trace-count 5 --seeds-minimizer-ignore-frequent 0.001 --precise-clipping 0.9. Centro-/telomeric sequence intervals, as classified by RepeatMasker, were merged within 250 Kb and 1 Kb by bedtools [35] respectively, and then split into centro-/telomeric sequence and bulk sequence fasta files. Both files were split every 100 Kb during realignment for computational constraints. Edit rate and query coverage was then calculated from the resulting graph alignment files.

The eight additional haplotype assemblies were constructed with the dual assembly approach in hifiasm (v0.16.1, [36]) using between 17 and 24–fold HiFi read coverage from two purebred Brown Swiss and two purebred Original Braunvieh cattle. The two Original Braunvieh cattle are the sire and dam (labelled as OSa, OSb, ODa, ODb) of an OxO F1 which was the source for the OBV haplotype [7] included in the pangenomes

whereas the two Brown Swiss cattle are not directly related to the BSW haplotype included in the pangenomes (labelled as BS1a, BS1b, BS2a, BS2b).

VNTR analysis

We identified TRs in the ARS-UCD1.2 reference sequence using TRF ([37], v4.10.0) with parameters "2 7 7 80 10 50 500", allowing a minimum and maximum motif length of 10 and 100 bp respectively and a maximum tandem repeat length of 1 Mb. We excluded TRs which overlapped with repeat elements identified in the reference sequence using bedtools intersect -v -f 0.5. For minigraph, we then used bedtools intersect again for the TRs and the minigraph SV bed files generated by gfatools bubble. We then extracted the approximate coordinates using minigraph --call for each assembly. For pggb and cactus we used the position command of odgi [24] to translate the TR coordinates from ARS-UCD1.2 to the paths of other assemblies in the graphs.

We calculated the number of TRs by extracting per-assembly sequence from the VNTR coordinates predicted from the pangenomes. We then used the python regex package to count how often the VNTR repeat motif provided by TRF occurs in the query sequence, allowing 25% error rate to account for substitutions, insertions, and/or deletions between repeat units.

We used adVNTR (v1.4.1, [19]) to genotype the 465 VNTRs with at least 50 repeat units in one of our HiFi samples. We built a database for these TRs using adVNTR addmodel, and then genotyped them using adVNTR genotype with the flags "--naive --pacbio --haploid" on the triobinned HiFi samples. These bam files were generated using sequences binned by Canu (v2.2, [38]) using parental reads, followed by alignment to ARS-UCD1.2 with minimap2 (as described in **Variation discovery from assemblies** except using the map-hifi preset).

The VNTR annotated pangenomes were generated using Bandage [39], using its built-in BLAST feature. The blastDB is created from the graph sequence, while we provide common TR motifs for given VNTRs as queries. Since motifs are "short" compared to typical blast queries, we use "-task blastn-short -word_size <TR length> -evalue 100" and filter minimum alignment sizes below <TR length> to increase sensitivity to specific motifs.

Establishing a testis eQTL cohort

Testis tissue of 83 mature bulls was collected at a commercial slaughterhouse and subjected to DNA and RNA purification as described earlier [40]. DNA samples were sequenced on an Illumina NovaSeq6000 instrument using 150 bp paired-end sequencing libraries. Following quality control, between 74,371,404 and 304,199,764

filtered read pairs per sample (mean: 268,798,344 ± 80,134,660) were aligned to the ARS-UCD1.2 reference sequence and processed as described in Kadri et al. [40]. Single nucleotide and short insertion and deletion polymorphisms were discovered and genotyped using DeepVariant (version 1.3.0, [41]). Beagle4.1 [42] was applied to impute sporadically missing genotypes and infer haplotypes. A genomic relationship matrix was constructed and subjected to a principal components analysis using plink (v.1.9, [43]).

Total RNA sequencing libraries (2 × 150 bp) were prepared using the Illumina TruSeq Stranded Total RNA sequencing kit and sequenced on an Illumina NovaSeq6000. Following quality control, between 70,249,355 and 177,053,828 filtered read pairs per sample (mean: 130,092,907 ± 18,684,899) were aligned to the ARS-UCD1.2 reference sequence and the Refseq gene annotation (release 106) using the splice-aware read alignment tool STAR (version 2.7.9a) [44] with options --twopassMode Basic and --waspOutputMode SAMtag to enable robust and unbiased allele-specific expression detection [45]. Per-individual VCF files required for WASP filtering were prepared from the cohort-level VCF files (see above).
Testis tissue from the NxB F1 was collected after regular slaughter. Total RNA was prepared and sequenced as described above. Following quality control, 136,820,136 filtered read pairs were aligned against the bovine reference sequence using STAR (see above).

Gene expression quantification

The expression level of genes, ncRNA and lncRNA was quantified (in transcripts per million, TPM) using kallisto (version 0.46.1, [46]) and aggregated to the gene level using the R package tximport [47]. We retained genes, ncRNA and lncRNA that were expressed at an average value >0.2 TPM across the 83 transcriptomes. The raw TPM values were normalized using quantile normalization and rank-based inverse normal transformation [47]. We inferred hidden confounding variables from the normalized gene expression levels using Probabilistic Estimation of Expression Residuals (PEER) [48].

eQTL mapping

The two VNTR haplotypes (hap1 and hap2) segregating in the Brown Swiss and Original Braunvieh breeds were derived from binned long read alignments with minimap2 (see above). The resulting diplotypes were reconstructed for all bulls of the eQTL cohort based on the Beagle-phased genotypes (see above) for two SNPs (Chr12:86160984 and Chr12:86160971) and one indel (Chr12:86161000) that tagged the VNTR haplotypes. A linear model was fitted using the lm()-function in R to associate normalized gene expression with the VNTR haplotype (coded as 0, 1, and 2 for hap1/hap1, hap1/hap2, and hap2/hap2, respectively) while considering five PEER factors and the top three principal components of the genomic relationship matrix as covariates to account for technical bias and population stratification. An eQTL analysis within the *cis*-regulatory range of

genes of interest was conducted using the linear model described above. The VNTR haplotype was fitted as an additional covariate in the conditional analyses. Bonferroni-correction was applied to determine significance thresholds.

Allelic imbalance analysis

Polymorphic sites overlapping *LOC112449094* exons were extracted from the cohort-level genomic VCF file. Reference and alternate allele support at heterozygous genotypes was subsequently extracted from the WASP-filtered RNA sequencing read alignments using bcftools mpileup. An exact binomial test as implemented in the R binom.test()-function was applied to test for allelic imbalance.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Scripts and pipelines used in this work are available online (https://github.com/AnimalGenomicsETH/superpangenome_construction). HiFi reads of two BSW and two OBV cattle are available in the ENA database at the study accession PRJEB42335 under sample accession SAMEA111341789, SAMEA111341790, SAMEA111341791, and SAMEA111341792. DNA and RNA sequencing data of the eQTL cohort are available in the ENA database at the study accession PRJEB46995.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by the Swiss National Sciences Foundation (SNSF), an ETH Research Grant, and Swissgenetics. The funders had no role in study design, data collection and analysis, interpretation of the data, decision to publish, or preparation of the manuscript.

Authors' contributions

A.S.L., D.C., and H.P. conceived the study. A.S.L. constructed genome assemblies and performed VNTR analyses. D.C. constructed, decomposed, and normalized the pangenomes and contributed to the VNTR

analyses. A.S.L., D.C., and H.P. performed the pangenome and variant analyses. X.M.M. established the eQTL cohort and aligned the sequencing data. A.S.L. called variants for the eQTL cohort. H.P. performed the eQTL analyses. M.B. contributed to the design of the eQTL analyses. A.S.L., and H.P. wrote the manuscript with input from D.C. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Anna Bratus-Neuenschwander and Dr. Catharine Aquino from the ETH Zurich technology platform FGCZ (https://fgcz.ch) for DNA fragment analysis and DNA and RNA sequencing.

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome. Proceedings of the National Academy of Sciences of the United States of America. 2005;102:16530.

2. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annual Review of Genomics and Human Genetics. 2020;21:139–62.

3. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs. Genome Biology. 2020;21:265.

4. Crysnanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. Proceedings of the National Academy of Sciences of the United States of America. 2021;118(20):e21010561.

5. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.

6. Garrison E. pggb (the pangenome graph builder). 2021. https://github.com/pangenome/pggb.

7. Leonard AS, Crysnanto D, Fang Z, Heaton MP, Brian L, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nature Communications. 2022;13:3012.

8. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604:437–46.

9. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.

10. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Verkko: telomere-to-telomere assembly of diploid chromosomes. bioRxiv; 2022. https://doi.org/10.1101/2022.06.24.497523.

11. Jarvis ED, Formenti G, Rhie A, Guarracino A, Yang C, Tracey A, et al. Automated assembly of high-quality diploid human reference genomes. 2022;bioRxiv: https://doi.org/10.1101/2022.03.06.483034.

12. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

13. Liao W, Asri M, Ebler J, Doerr D, Haukness M, Lu S, et al. A Draft Human Pangenome Reference. 2022;bioRxiv.https://doi.org/10.1101/2022.07.09.499321.

14. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology. 2018;36:1174–82.

15. Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, et al. A cattle graph genome incorporating global breed diversity. Nature Communications. 2021;13:910.

16. Talenti A, Powell J, Wragg D, Chepkwony M, Fisch A, Ferreira BR, et al. Optical mapping compendium of structural variants across global cattle breeds. bioRxiv.2022:https://doi.org/10.1101/2022.05.05.490773.

17. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. Trends in Plant Science. 2020;25:148–58.

18. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications. 2019;10:1784.

19. Bakhtiari M, Park J, Ding YC, Shleizer-Burko S, Neuhausen SL, Halldórsson B V., et al. Variable number tandem repeats mediate the expression of proximal genes. Nature Communications. 2021;12:2075.

20. Lu, Tsung-Yu MJPC. The motif composition of variable-number tandem repeats impacts gene expression. bioRxiv. 2022; https://doi.org/10.1101/2022.03.17.484784.

21. Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, et al. Genome evolution and diversity of wild and cultivated potatoes. Nature. 2022;606:535–41.

22. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature. 2022;606:527–34.

23. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience. 2020;9(3):giaa021.

24. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. Bioinformatics. 2022; 38(13):3319–3326.

25. Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. GigaScience. 2020;9(4):giaa029.

26. Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, et al. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. Journal of Heredity. 2021;112:184–91.

27. Oppenheimer J, Rosen BD, Heaton MP, Vander Ley BL, Shafer WR, Schuetze FT, et al. A Reference Genome Assembly of American Bison, Bison bison bison. Journal of Heredity. 2021;112:174–83.

28. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17:132.

29. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology. 2018;36:875–81.

30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021; 10(2):giab008.

31. Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. PLOS Computational Biology. 2022;18:e1009123.

32. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

33. Kirsche M, Prabhu G, Sherman R, Ni B, Aganezov S, Schatz MC. Jasmine : Population-scale structural variant comparison and analysis. bioRxiv. 2021; https://doi.org/10.1101/2021.05.27.445886.

34. Rautiainen M, Marschall T. GraphAligner: Rapid and Versatile Sequence-to-Graph Alignment. Genome Biology. 2020;21:253.

35. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

36. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 2021;18:170–5.

37. Benson G. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Research. 1999;27:573–80.

38. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Research 2020;30:1291–305.

39. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–2.

40. Kadri NK, Mapel XM, Pausch H. The intronic branch point sequence is under strong evolutionary constraint in the bovine and human genome. Communications Biology. 2021;4:1206.

41. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal snp and small-indel variant caller using deep neural networks. Nature Biotechnology. 2018;36:983-7.

42. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. American Journal of Human Genetics. 2016;98:116–26.

43. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

44. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

45. Van De Geijn B, Mcvicker G, Gilad Y, Pritchard JK. WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods. 2015;12:1061–3.

46. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34:525–7.

47. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Research. 2016;4:1521.

48. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols. 2012; 7(1):e1001276.