The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species

Timothy P L Smith, Derek M Bickhart, Didier Boichard, Amanda J Chamberlain, Appolinaire Djikeng, Yu Jiang, Wai Y Low, Hubert Pausch, Sebastian Demyda-Peyrás, James Prendergast, et al.

Genome Biology

CORRESPONDENCE

Open Access

The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species

Timothy P. L. Smith[1], Derek M. Bickhart[2], Didier Boichard[3], Amanda J. Chamberlain[4,5], Appolinaire Djikeng[6,7], Yu Jiang[8], Wai Y. Low[9], Hubert Pausch[10], Sebastian Demyda-Peyrás[11,12], James Prendergast[7,13], Robert D. Schnabel[14], Benjamin D. Rosen[15*] and Bovine Pangenome Consortium

*Correspondence: ben.rosen@usda.gov

[15] Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA
Full list of author information is available at the end of the article

Abstract

The Bovine Pangenome Consortium (BPC) is an international collaboration dedicated to the assembly of cattle genomes to develop a more complete representation of cattle genomic diversity. The goal of the BPC is to provide genome assemblies and a community-agreed pangenome representation to replace breed-specific reference assemblies for cattle genomics. The BPC invites partners sharing our vision to participate in the production of these assemblies and the development of a common, community-approved, pangenome reference as a public resource for the research community (https://bovinepangenome.github.io/). This community-driven resource will provide the context for comparison between studies and the future foundation for cattle genomic selection.

Why a pangenome?

The inadequacy of a single reference assembly for the detection of all genetic variation has been documented with respect to human genome research (reviewed in [1] and [2]). A broad consensus of the biomedical genomics research community supports the creation of highly accurate and haplotype-phased assemblies that include examples from across common human haplotypes, followed by the construction of a pangenome representation to underpin variant detection. Existing variation detection tools have been reported to be sensitive to the quality and representation of the reference genome (i.e., reference bias) and to be improved by the use of more representative reference(s) or pangenome representations [3]. Structural variants are particularly difficult to accurately identify, characterize, and compare using now-standard approaches of short-read or even long-read mapping to a single-haplotype reference assembly [4]. Insertions relative

to the reference sequence are particularly recalcitrant to detection without orthogonal genomic information as sequence reads from haplotypes containing the insertion fail to properly map to the reference. Sequence reads with low similarity to the reference assembly due to population variation in genome sequence can cause incorrect mapping and lead to incorrect genotypes in whole-genome shotgun (WGS)-based genotyping assays. One example of reference bias resulting in genotype errors is the human leukocyte antigen locus, where data from the 1000 Genomes Project suggested that nearly 19% of SNPs were incorrectly identified [5]. As many as 68% of structural variants detected by alignments of haplotype-resolved assemblies were not detected using short-read mapping to the human reference assembly [6], highlighting the need for a more complete reference genome resource for such analyses. The development of a human pangenome through the Human Pangenome Reference Consortium (HPRC) has been implemented (https://humanpangeome.org) by way of a Human Pangenome Reference Sequence Project, funded with nearly US $30 million in grants in 2019 (https://www.genome.gov/news/news-release/NIH-funds-centers-for-advancing-sequence-of-human-genome-reference). The charter for this effort includes the ethical access of samples to avoid the exploitation of local populations and outreach to international groups as explicit goals [7]. Furthermore, the HPRC project seeks to address limitations inherent in contemporary genetic variant surveys using individual, linear genome assemblies, and computational tools designed for comparisons against a lone reference [7]. Because no similar singular funding source is available for agricultural genomics, the Bovine Pangenome Consortium (BPC) has been launched (https://bovinepangenome.github.io/) to coordinate distributed efforts within the global bovine genomics community towards achieving similar ends.

The goal of genomics research in livestock has been to correlate variation in genome sequence with phenotypes that affect traits of importance to animal health, welfare, productivity, profitability, and sustainability as part of a broader strategy to increase favorable allele frequency and decrease deleterious allele prevalence. The importance of this research is highlighted by the recent investment of up to US $40 million per year in the "Agricultural Genomes to Phenomes Initiative" by the USDA [8] and the establishment of the international Functional Annotation of Animal Genomes "FAANG to Fork" initiative [9], a project analogous to the human ENCODE project [10]. Prior work in the development of predictive models for complex trait phenotypes from a fixed set of genetic markers, termed genomic selection, has resulted in substantial benefits for animal breeders [11, 12]. Implementation of genomic selection in dairy cattle has doubled the rate of economic gain transmitted by Holstein bulls in the US Dairy Industry, resulting in profits for breeders and more efficient milk production within the system [13]. However, the accuracy of genomic selection on certain phenotypic traits is greatly impacted by environmental effects and the influence of high-effect, low-frequency genetic variants [14]. Improved reference genome assemblies and efforts to predict combined annotation-dependent depletion scores (CADD; [15]) for all bases in the genome of a species could allow for the inclusion of high-effect, low-frequency variants into generalized genomic selection models. The entire economic impact of the US Dairy Industry alone has been estimated to be $753 billion (USD), which makes even small improvements in system efficiency valuable in terms of nominal economic value (https://www.

idfa.org/dairydelivers). For example, a 1% increase in the reliability of predicting production traits from genotype data made possible by the implementation of an improved reference assembly in dairy breed genomic evaluation would amount to an improvement in the efficiency of $7.5 billion (USD) in the entire system [16].

The establishment of a cattle reference genome was foundational to the conduct of genetic/genomic analysis and transformational for the practice of evaluating genetic potential. The ARS-UCD1.2 reference assembly [17] has profound limitations as it is a haploid representation of a single Hereford cow selected from a line-bred herd with high historical inbreeding [18]. This reference assembly was useful but inherently limits the ability to analyze the breadth of genome variation existing in global cattle populations. We suggest the development of reference-quality genome assemblies for as many of the existing distinct cattle breeds, including representatives of both subspecies, *Bos taurus taurus* and *Bos taurus indicus*, henceforth taurine and indicine, as practical to modernize the cattle reference genome. These assemblies would then be used to create a new, globally representative reference genome graph as a resource for future genomics studies. We propose that this "pangenome" graph should focus on the careful selection of structural variant sites across cattle breeds to optimize the utility of the graph for different research purposes [3, 19, 20].

Additional members of the Bovini tribe, henceforth bovine, are also of particular interest as they represent lineages that have been subjected to both natural and artificial selection and bottlenecks throughout their history. Taurine and indicine cattle, riverine and swamp buffalo (*Bubalus bubalis*), yak (*Bos grunniens*), and gayal (*Bos frontalis*) have extensive histories of domestication which include the recent formation of breeds due to intensive selection for agricultural food products. Other bovines, such as bison (*Bison bison*), wisent (*Bison bonasus*), banteng (*Bos javanicus*), wild yak (*Bos mutus*), gaur (*Bos gaurus*), and cape buffalo (*Syncerus caffer caffer*), exist as wild populations that have experienced varying natural constraints on mating. Comparative genomic analysis of these divergent lineages requires suitable genomics resources and integration of reference genome assemblies into a bovine pangenome will facilitate the investigation of these evolutionary processes as well as increase our understanding of potential introgression events that may have occurred throughout history.

Complete functional annotation of genes, transcript isoforms, chromatin states and their variation between cell types and cell functions, and accurate mapping of three-dimensional structures of chromatin are all dependent on complete knowledge of the cattle genome including all extant haplotypes. Assessment of the value of preserving disappearing or threatened breeds would benefit from genome-level analysis documenting any unique contribution they might provide for maintaining the full extent of existing variation in the species and evaluating any unique traits they may harbor. The inclusion of breeds common in low- and middle-income countries will support the application of genomics in situations where mass genotyping of animals is not feasible. Comparison to genome assemblies of other bovines will provide important context for assessing the conservation of genomic loci that may underlie phenotypic diversity. Pangenome research that encompasses both wild and domesticated relatives of cattle can contribute to research in other food animal bovids, illuminate the genomic consequences and mechanisms of domestication, and provide insight into the genes involved with docility

and other behaviors and adaptation traits. These objectives, in addition to the improvement of detection of structural variants and increased accuracy of SNP calling from WGS data, provide the primary motivation for constructing a bovine pangenome. A properly constructed and widely used pangenome should ideally replace any need for breed-specific assemblies in most genome research applications while facilitating across-study comparisons.

Goals of a bovine pangenome project

The global cattle population appears much more diverse than the human population, with intra-individual heterozygosity up to 3- to tenfold higher in cattle [21]. The bottlenecks associated with domestication and spread of cattle populations have not reduced the effective population size of the global population to the extent observed for *Homo sapiens*. This suggests that the construction of a cattle pangenome would provide an even higher impact in the identification and evaluation of genomic variation than has been documented or is expected in biomedical research on human populations. There are > 1000 recognized breeds across the globe, representing the two subspecies, with population sizes ranging from the hundreds to tens of millions. The estimated high diversity of ancestral aurochs that formed the basis for at least two distinct domestication events has led to a relatively high degree of genetic heterogeneity even in breeds with small effective population sizes. This diversity is reflected in the success of selection for dairy traits in Holstein cattle, for example, where a very small effective population size has not interfered with substantial progress in phenotype- or genotype-based selection [22]. The plethora of breeds and high within-breed heterogeneity makes it imperative to include as many breeds as possible in the construction of a pangenome. Our ability to represent all structural variation will be a function of allele frequencies and the partitioning of variation between breeds. It will therefore be important to test the necessity of including multiple assemblies within more cosmopolitan breeds where animal numbers are high, or the breed may have had its genome shaped by different selection pressures.

Genome assemblies of domesticated and wild bovine relatives have value for rooting phylogeny through the identification of ancestral alleles in regions that are less conserved at a wider evolutionary scale; they provide context on the evolution of chromosomes, centromere positioning, and the origin of structural variants; they support detection and characterization of genetic variation related to domestication including docility traits that were likely the original artificial selection; they allow comparisons of the genomic sequence which may expose novel genes or incomplete lineage sorting [23] that may have influenced species/breed domestication outcomes and may be targets of future selection. For these reasons, the establishment of reference-quality genome assemblies of these species is being pursued as part of our initiative. Assemblies of bovine relatives of cattle will also be assessed for potential utility in the construction of a pangenome graph resource for cattle lineages, although the relative value of this is still under investigation (Fig. 1).

Logistical concerns including access to animals, producer reluctance to allow sampling, feasibility of sampling, and proper transport of samples from origin to sequencing laboratory all challenge the goal of obtaining a broad representation of cattle breeds and

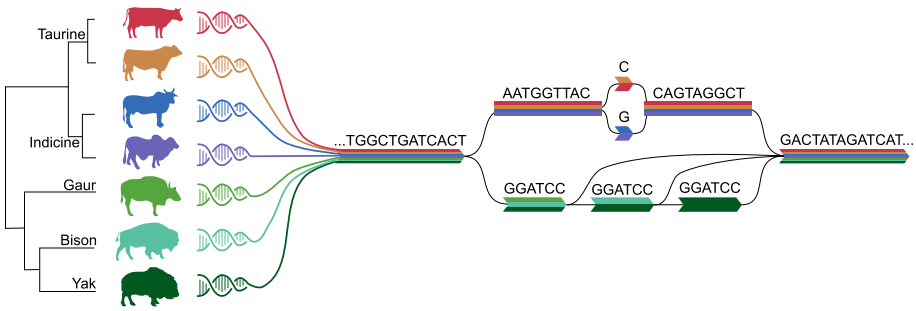

Fig. 1 Bovine species relationships and pangenome conceptualization. Single nucleotide and structural variation between species/subspecies/breeds are depicted as different paths through a multi-species pangenome

Table 1 Goals of the BPC

• Establish collaborations with internal and external participants to facilitate genome assembly projects on diverse bovine species/breeds
• Coordinate bovine genome assembly efforts by publishing lists of on-going projects and by providing expertise to collaborating groups
• Mediate future bovine genome assembly efforts by providing a forum for the community to discuss issues, project ideas and limitations
• Provide updated recommendations and resources to the community to identify the best current resources to reduce losses in inter-study comparability resulting from use of different reference genome assemblies

related species. We anticipate that not all participants in the project will have access to adequate laboratory facilities or to the pedigrees of sampled individuals in their herds. The BPC acknowledges that flexibility to accommodate disparate biological samples is required to achieve the goal of representation. We also welcome interested partners with shared or overlapping goals that already have invested resources, expertise, and/or animal samples. Consortium resources for generating bovine genome assemblies are finite, so samples must be prioritized based on the likelihood of quality outcomes from their use. To this end, the BPC proposes several quality standards for samples and materials that will be used to prioritize projects (see the " Implementation" section).

The immediate goal of the pangenome project is the collection of assemblies for as many cattle breeds and related species as possible. This approach enables the immediate use of reference-quality assemblies for specific species/breeds using existing linear genome tools in the interim and allows for future assessment of data completeness before assembling a comprehensive pangenome resource. Eventually, the goal of the BPC is to construct a community-approved pangenome to be used for genomics in cattle, providing connections between studies while enhancing each study with an appropriate breed-specific context (Table 1).

Role of the BPC

The generation of reference-quality genome assemblies was enormously expensive and required extensive, specialized expertise as recently as 10 years ago [24]. Advances in both sequencing technologies and the algorithms for producing assemblies have vastly improved in the interim, reducing both cost and required expertise. The original conception of the BPC arose from the realization that entities worldwide would

increasingly have the ability to generate such assemblies with no need for expert consultation. Uncoordinated efforts could therefore result in a squandering of the relatively meager resources available to the global livestock research community through unnecessary duplication of efforts and a rush to be the first to publish assemblies for a few high-impact breeds. The major benefits that the BPC can provide to the international community of cattle genomics researchers are (1) to facilitate collaborations for bovine genome assembly projects, (2) to establish a list of breeds/species assembly projects to assist in allocating resources, (3) to provide a forum for the discussion of the completion state of individual assembly projects, and (4) to create an organization to focus efforts on the development of a community-approved pangenome representation from the data. The BPC is organized into three distinct branches that will cooperate to reach a consensus on key policy and technical issues (Fig. 2). The BPC co-chairs will manage consortium resources and will consult with the steering committee to ensure that consortium objectives will be met. The larger BPC consortium will comprise all interested members of the research community that plan to directly work on the goals of the consortium. The entirety of the BPC will engage with the larger research community, genomics resource staff, and collaborating institutes to achieve consortium goals, while respecting ethical limits placed on data access.

The concept behind and intent of the BPC are to increase inclusivity and lower barriers to participation. The organizers are sensitive to concerns about the historical incidence of resource-rich entities benefiting at the expense of researchers and organizations with access to domestic species and breeds that would enhance the pangenome project. We are committed to avoiding the appearance or actuality of this problem, while also reaching out to promote the inclusion of historically under-represented geographic areas. Sequencing resources and computational infrastructure are available through the US Department of Agriculture to support interested researchers in these communities through this effort specifically. Although the USA and some other members' countries are not signatories to the Nagoya Protocol (https://www.cbd.int/abs/text/), the BPC will adhere to the spirit of the access, benefit sharing, and compliance obligations of the protocol (https://www.cbd.int/abs/about/). The intent of the BPC is threefold: to facilitate bovine research in under-represented research communities, to support research programs through the judicious allocation of consortium resources, and to advise collaborators on the best practices for genome assembly and subsequent data analysis. We emphatically do not intend to "take over" any project, nor will we insist on any particular

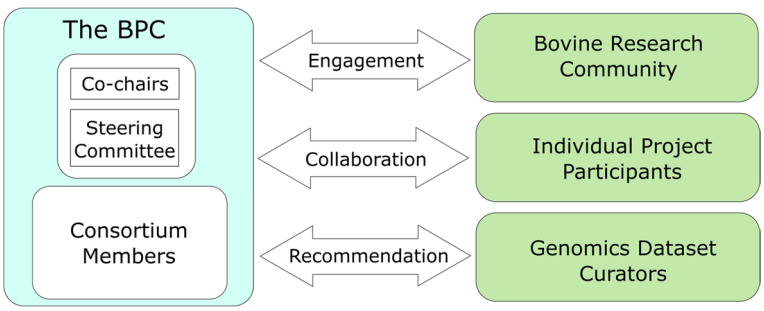

Fig. 2 BPC cooperator outreach and engagement

position in an authorship list. All data and analysis produced by the BPC will be considered the property of the sample provider, with full access and availability throughout the life of the project. The details of each collaboration will depend highly on the desires and situation of participants; however, the BPC's driving goal is to facilitate the successful outcomes of a project to benefit the sample provider and the cattle research community as a whole. The only requirements are that the data is intended to be public within a year's time of the assembly and that the assembly can be included in a public pangenome resource. The pangenome publication(s) resulting would acknowledge each constituent assembly and include authorship for participants if they desire. There is no intent to interfere with or prevent publication of individual genome assemblies or analysis, or pangenome constructions/representations with subsets of breeds. Rather, the ultimate goal is for the members of the BPC to lead the creation of a definitive community-standard representation that can be the basis of inter-study comparisons. Since we mandate that BPC participants keep direct ownership over the direction of their projects, they are able to pursue the specialized research into domestic or heritage breeds that they desire while still making substantial contributions to global cattle genome research efforts.

Implementation

We recognize that public accessibility to component genome assemblies and user-friendly interfaces for making use of pangenome representation(s) in genomic studies are critically important to the success of the effort. We propose to join ongoing efforts at Ensembl and NCBI or initiate new ones as needed, in close concert with those data repositories to ensure open access and facilitate broad utility, and to provide a means for improving uniformity among disparate assembly efforts. We also plan to reach out to the FAANG initiative to enable sharing of resources that could be useful in the quality assessment of genome content [9, 25]. The creation, presentation, and application of pangenome representations of species genomes is an area of active investigation, and best practices are not yet fully agreed upon. New standards including, for example, the rGFA format for representing a reference pangenome graph [26], will be incorporated as they emerge and are refined. However, all emerging methods for representing pangenomes depend on sequence data and assemblies for a broad representation of the species, and thus, there is no advantage in waiting until a consensus is achieved to begin the collection of the underpinning data. One major consideration is that the input tissue type or DNA extraction method greatly influences the resulting quality of the DNA sequence from a sample. We recognize that it may not always be possible to obtain ideal samples for a breed, and there will be no absolute requirements for samples to meet all of our listed criteria provided that there are means to legally and ethically source them. The ideal sample is one that (1) is amenable to DNA extraction for long-read sequencing (i.e., not liver, hair roots, highly cartilaginous tissues, or tissues with high-fat content), (2) is free of concerns about pathogenic infections that prevent shipping to a sequencing laboratory (e.g., foot and mouth disease (FMD) virus, tuberculosis, or babesia), and (3) is large enough to support extraction of sufficient DNA for "greedy" long-read sequencing platforms/procedures. The use of long-read technologies vastly improves the contiguity, accuracy, and utility of de novo assemblies [17, 27] for the construction of pangenome representations. The latest assembly methods produce haplotype-resolved assemblies

of two breeds from a single sample at a marginal increase in cost and reduced effort through the use of F1 animals produced by crossing breeds [4, 28]. These methods produce the most contiguous and accurate assemblies; therefore, a goal of the pangenome project includes a collection of F1 crosses between breeds where possible. Later-generation crossbred individuals will not be targeted for inclusion so long as the progenitor breeds are represented in the pangenome. Where the progenitors are not able to be sampled, crossbreds may be an appropriate way to capture missing diversity.

Pangenome approaches that use reference-guided assembly with either short or long reads have been proposed or performed, with the benefit that the cost is lower than complete de novo assembly and thus more animals can be processed for a given cost. A related approach is to assemble reads that fail to map to the genome with the supposition that resulting contigs would represent unique genomic sequences. The initial results indicate that a range of 3 to 10 megabases (Mb) of genome sequence that does not match the Hereford reference genome can be identified by sequencing animals of other breeds [20, 29]. This result is largely consistent with observations in human pangenome research, where anywhere from 5 to 300 Mb of non-reference sequence has been identified depending on the divergence between lineages. These methods are complementary to the proposed BPC approach of high-quality de novo assembly of one or a few haplotypes of a breed for as many breeds as possible and will serve to enhance the accuracy and representation of the final pangenome. The BPC will work to determine appropriate standards for consideration of an assembly for a particular breed, while noting that initial investigations of pangenome graphs indicate resilience in the face of lower-quality assemblies and low sensitivity to the inclusion of assemblies produced from very different methods and sequencing platforms [30]. This observation is promising since the evolution of technologies can be expected over the several years lifetime of the project; therefore, our present reliance on existing technologies should not prevent the utility of these assemblies when improvements come along (Table 2).

The cattle genome is roughly equivalent to the human genome in terms of the proportion of repetitive sequences, and the number of protein-coding genes and transcript isoforms per protein-coding gene is also similar [17]. We therefore anticipate that methods for pangenome construction, presentation, and implementation can make use of advances from the HPRC. There are a variety of approaches currently described that have varying degrees of dependence on an initial single reference assembly. Reference-guided assembly approaches based on de novo assembled contigs aligned to the reference have been used to identify insertion-deletion structural variants to add paths to a pangenome graph [31]. Mapping of WGS reads to the reference, followed by assembly of

Table 2 The role of the BPC in developing the next generation of bovine genomics resources

• Generation of high-quality, haplotype-resolved reference genome assemblies for under-represented or highly divergent bovine breeds/species
• Consolidate the latest in bovine genome assemblies for use in pangenome resource construction
• Conduct a comparative analysis of bovine species to annotate conserved genomic regions that may correspond to species- or breed-specific phenotypic variation
• Organize experts and support the creation of graph-based pangenome reference(s) for cattle breeds to enable more accurate sequence-based genotyping

non-mapping reads assumed to represent the missing sequence in the reference, has also been applied [32, 33]. The most robust pangenome representations have resulted from independent de novo assembly and whole-genome alignment [26, 34] as demonstrated by a comparison of approaches to the same dataset [35].

Effective means for interrogating pangenomes that are computationally tractable for further variant discovery based on mapping of inexpensive short-read data at the population scale are being developed [36–39]. Bovine pangenomes with limited numbers of breeds have been reported [19, 20, 29] that used combined methods of aligned assemblies with the addition of structural variants identified by mapping of short reads to the graph created from those aligned assemblies. These efforts underscore the need for the research community to come to an agreement on a single representation that is robust for multiple comparative analyses. Such a reference will need to accommodate short read-based variation and genotyping studies employed in cattle genomics projects, to provide a common coordinate system for comparison across studies and serve as a foundation for annotation of functional genome features. As this reference will be used as a basis for genetic sequence comparison in research and industry, concrete thresholds for a public release must be identified by the community (with mediation by the BPC) to avoid frequent iterations of resources that disrupt ongoing work. We do not anticipate that any pangenome graph prepared will be the final product as future work may produce ever more effective means to represent pangenomes or further iteration may be necessary to represent cryptic DNA sequences missed in our original survey. However, just as the original UMD3.1 cattle genome assembly [40] was useful for cross-study comparisons for over a decade, we believe that a useful, community-approved pangenome reference can provide a stable means for similar time frames.

Genome assembly efforts such as the Vertebrate Genomes Project [41] have emphasized the use of a common technology and methodology for the production of assemblies to provide uniformity and prevent biases when comparing genome architecture across species. An evaluation of pangenome construction in cattle suggested that a bovine pangenome can make use of genome assemblies of breeds at variable levels of quality, input data type, or assembly method [30]. Best practices for aligning genomes, presenting pangenome graphs, and performing variant detection on those graphs are rapidly evolving and are likely to continue to change in the foreseeable future. However, the observation that the pangenome will be robust to variability in assembly parameters supports our contention that the community can go forward with generating genome assemblies for global bovine species as soon as possible. This makes the most important constraint the availability of high-quality samples from diverse bovine species/breeds for genome assembly. Since the collection of samples, generation of sequence data, and assembly of many hundreds of breeds is unlikely to be completed in the next year or two, it is likely that by the time we have the assemblies in hand, the best approach or at least useful non-ideal approach will be generally agreed upon for generating the community standard bovine pangenome. We suggest that until such time, mapping to the Hereford ARS-UCD1.2 reference assembly for cross-study comparisons will continue to be the most useful approach rather than study-specific pangenome graphs for variant discovery. The use of such intermediate resources poses the risk of immediately dating variant calling results while the community coalesces on a final representative genomics

resource. At that time, the BPC will plan to provide leadership in the selection of pangenome resources suitable for different applications by using community-driven feedback on developed resources. By serving as an intermediary and facilitator for the recommendation of the final pangenome resource, the BPC plans to reduce confusion and improve cross-study comparability for future bovine resequencing projects.

Conclusion

The BPC was established as a means to coordinate bovine global genome assembly efforts and not as a means to override, interfere, or take credit for local efforts. We encourage researchers that want to participate in the pangenome effort to join us. The initial assembly of diverse bovine species/breeds is supported by the US Department of Agriculture in part because of the historical role of the Agricultural Research Service (ARS) in the early cattle genome assembly efforts, development of methods for genome assembly, and ongoing presence in the generation of bovine genome assemblies. However, the BPC has an international steering committee and encourages open discussion of any topics related to data sharing, publication strategy, or logistics of sampling animals around the world. Resources for creating genome assemblies are available through ARS, and we are open to discussion of providing sequencing and/or assembly for breeds where local resources are not available. Any such collaborations will provide only the support requested, i.e., any combination of sequencing, assembly, data deposition, or advice. Experts participating in the consortium can assist in manuscript writing as requested. The sole requirement is that ARS-funded genome assemblies will be made public and will be available for the final pangenome representation. All collaborators will be offered the opportunity for coauthorship on manuscript(s) describing the community-standard pangenome, assuming they materially contributed to composite assemblies.

A list of breeds that have been assembled or have allocated resources to support assembly will be maintained at https://bovinepangenome.github.io/. The BPC will continue with the active outreach that we have been engaged in during the pandemic, but welcome contact from anyone that has access to cattle breeds or bovine species not currently on our list, or for which there is evidence that sufficient within-breed or within-species diversity exists to make multiple assemblies worthwhile.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-023-02975-0.

Additional file 1.

Acknowledgements

The work presented here originated from the discussions at scientific meetings starting in 2020. We are grateful to the members of the Bovine Pangenome Consortium for their comments, suggestions, and willingness to provide samples. We thank Alexander S. Leonard for designing the bovine pangenome representation for Fig. 1. We thank the USDA for funding. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 1.

Authors' contributions
All authors contributed to the writing of the manuscript. The authors read and approved the final manuscript.

Funding
USDA, ARS appropriated project 3040-31000-100-00D to TPLS; USDA, ARS appropriated project 5090-31000-026-00D to DMB; USDA, ARS appropriated project 8042-31000-001-00D to BDR.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]US Meat Animal Research Center, USDA-ARS, Clay Center, NE 68933, USA. [2]Dairy Forage Research Center, USDA-ARS, Madison, WI 53706, USA. [3]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France. [4]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia. [5]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia. [6]Centre for Tropical Livestock Genetics and Health, ILRI Kenya, Nairobi 30709-00100, Kenya. [7]Centre for Tropical Livestock Genetics and Health, Easter Bush, Midlothian EH25 9RG, UK. [8]Center for Ruminant Genetics and Evolution, Northwest A&F University, Yangling 712100, China. [9]The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia. [10]Animal Genomics, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland. [11]Departamento de Producción Animal, Facultad de Ciencias Veterinarias, Universidad Nacional de La Plata, 1900 La Plata, Argentina. [12]Consejo Superior de Investigaciones Científicas Y Tecnológicas (CONICET), CCT-La Plata, 1900 La Plata, Argentina. [13]The Roslin Institute, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, UK. [14]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA. [15]Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA.

Received: 19 July 2022 Accepted: 19 May 2023
Published online: 19 June 2023

References

1. Miga KH, Wang T. The need for a human pangenome reference sequence. Annu Rev Genomics Hum Genet. 2021;22:81–102. https://doi.org/10.1146/annurev-genom-120120-081921.
2. Khamsi R. A more-inclusive genome project aims to capture all of human diversity. Nature. 2022;603:378–81. https://doi.org/10.1038/d41586-022-00726-y.
3. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. Genome Biol. 2018;19:220. https://doi.org/10.1186/s13059-018-1595-x.
4. Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, Thibaud-Nissen F, Martin FJ, Billis K, Ghurye J, Hastie AR, Lee J, Pang AWC, Heaton MP, Phillippy AM, Hiendleder S, Smith TPL, Williams JL. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nat Commun. 2020;11:2071. https://doi.org/10.1038/s41467-020-15848-y.
5. Brandt, DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, and Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes project phase i data. G3 (Bethesda). 2015;5:931–941. https://doi.org/10.1534/g3.114.015784.
6. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, Höps W, Ashraf H, Chuang NT, Yang X, Munson KM, Lewis AP, Fairley S, Tallon LJ, Clarke WE, Basile AO, Byrska-Bishop M, Corvelo A, Evani US, Lu T-Y, Chaisson MJP, Chen J, Li C, Brand H, Wenger AM, Ghareghani M, Harvey WT, Raeder B, Hasenfeld P, Regier AA, Abel HJ, Hall IM, Flicek P, Stegle O, Gerstein MB, Tubio JMC, Mu Z, Li YI, Shi X, Hastie AR, Ye K, Chong Z, Sanders AD, Zody MC, Talkowski ME, Mills RE, Devine SE, Lee C, Korbel JO, Marschall T, and Eichler EE. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021; 372:eabf7117. https://doi.org/10.1126/science.abf7117.
7. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, Chang X, Cook-Deegan R, Felsenfeld AL, Fulton RS, Garrison EP, Garrison NA, Graves-Lindsay TA, Ji H, Kenny EE, Koenig BA, Li D, Marschall T, McMichael JF, Novak AM, Purushotham D, Schneider VA, Schultz BI, Smith MW, Sofia HJ, Weissman T, Flicek P, Li H, Miga KH, Paten B, Jarvis ED, Hall IM, Eichler EE, Haussler D. The Human Pangenome Project: a global resource to map genomic diversity. Nature. 2022;604:437–46. https://doi.org/10.1038/s41586-022-04601-8.
8. Tuggle CK, Clarke J, Dekkers JCM, Ertl D, Lawrence-Dill CJ, Lyons E, Murdoch BM, Scott NM, Schnable PS. The Agricultural Genome to Phenome Initiative (AG2PI): creating a shared vision across crop and livestock research communities. Genome Biol. 2022;23:3. https://doi.org/10.1186/s13059-021-02570-1.
9. Clark EL, Archibald AL, Daetwyler HD, Groenen MAM, Harrison PW, Houston RD, Kühn C, Lien S, Macqueen DJ, Reecy JM, Robledo D, Watson M, Tuggle CK, Giuffra E. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. Genome Biol. 2020;21:285. https://doi.org/10.1186/s13059-020-02197-8.
10. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306:636–40. https://doi.org/10.1126/science.1105136.

11. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. J Dairy Sci. 2009;92:16–24. https://doi.org/10.3168/jds.2008-1514.
12. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92:433–43. https://doi.org/10.3168/jds.2008-1646.
13. Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic selection in dairy cattle: the USDA experience. Ann Rev Anim Biosci. 2017;5:309–27. https://doi.org/10.1146/annurev-animal-021815-111422.
14. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18:77. https://doi.org/10.1186/s13059-017-1212-4.
15. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94. https://doi.org/10.1093/nar/gky1016.
16. Null DJ, VanRaden PM, Rosen BD, O'Connell JR, Bickhart DM. Using the ARS-UCD1.2 reference genome in U.S. evaluations. Interbull Bulletin. 2019;55:30–4.
17. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, Hall R, Li W, Rhie A, Ghurye J, McKay SD, Thibaud-Nissen F, Hoffman J, Murdoch BM, Snelling WM, McDaneld TG, Hammond JA, Schwartz JC, Nandolo W, Hagen DE, Dreischer C, Schultheiss SJ, Schroeder SG, Phillippy AM, Cole JB, Van Tassell CP, Liu G, Smith TPL, and Medrano JF. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9. https://doi.org/10.1093/gigascience/giaa021.
18. Elsik CG, Tellam RL, Worley KC. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 2009;324:522–8. https://doi.org/10.1126/science.1169588.
19. Crysnanto D, Wurmser C, Pausch H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. Genet Sel Evol. 2019;51:21. https://doi.org/10.1186/s12711-019-0462-x.
20. Talenti A, Powell J, Hemmink JD, E. a. J. Cook, D. Wragg, S. Jayaraman, E. Paxton, C. Ezeasor, E.T. Obishakin, E.R. Agusi, A. Tijjani, K. Marshall, A. Fisch, B.R. Ferreira, A. Qasim, U. Chaudhry, P. Wiener, P. Toye, L.J. Morrison, T. Connelley, and J.G.D. Prendergast. A cattle graph genome incorporating global breed diversity. Nat Commun. 2022;13:910. https://doi.org/10.1038/s41467-022-28605-0.
21. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018;36:1174–82. https://doi.org/10.1038/nbt.4277.
22. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci. 2016;113:E3995–4004. https://doi.org/10.1073/pnas.1519061113.
23. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483:169–75. https://doi.org/10.1038/nature10842.
24. Bickhart DM, McClure JC, Schnabel RD, Rosen BD, Medrano JF, Smith TPL. Symposium review: advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. J Dairy Sci. 2020;103:5278–90. https://doi.org/10.3168/jds.2019-17693.
25. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C, Dalrymple BP, Elsik CG, Foissac S, Giuffra E, Groenen MA, Hayes BJ, Huang LS, Khatib H, Kijas JW, Kim H, Lunney JK, McCarthy FM, McEwan JC, Moore S, Nanduri B, Notredame C, Palti Y, Plastow GS, Reecy JM, Rohrer GA, Sarropoulou E, Schmidt CJ, Silverstein J, Tellam RL, Tixier-Boichard M, Tosser-Klopp G, Tuggle CK, Vilkki J, White SN, Zhao S, Zhou H. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 2015;16:57. https://doi.org/10.1186/s13059-015-0622-4.
26. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 2020;21:265. https://doi.org/10.1186/s13059-020-02168-z.
27. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce de León FA, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassell CP, Smith TPL. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50. https://doi.org/10.1038/ng.3802.
28. Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, Hackett PH, Bickhart DM, Rosen BD, Ley BV, Maurer NW, Green RE, Phillippy AM, Petersen JL, Smith TPL. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. GigaScience. 2020;9:giaa029. https://doi.org/10.1093/gigascience/giaa029.
29. Crysnanto D, Leonard AS, Fang Z-H, and Pausch H. Novel functional sequences uncovered through a bovine multi-assembly graph. PNAS. 2021;118. https://doi.org/10.1073/pnas.2101056118.
30. Leonard AS, Crysnanto D, Fang Z-H, Heaton MP, Ley BLV, Herrera C, Bollwein H, Bickhart DM, Kuhn KL, Smith TP, Rosen BD, and Pausch H. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. 2022. 2021.11.02.466900. https://doi.org/10.1101/2021.11.02.466900.
31. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21:35. https://doi.org/10.1186/s13059-020-1941-7.
32. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM, Eng C, Yazdanbakhsh M, Wilson JG, Marrugo J, Lange LA, Williams LK, Watson H, Ware LB, Olopade CO, Olopade O, Oliveira RR, Ober C, Nicolae DL, Meyers DA, Mayorga A, Knight-Madden J, Hartert T, Hansel NN, Foreman MG, Ford JG, Faruque MU, Dunston GM, Caraballo L, Burchard EG, Bleecker ER, Araujo MI, Herrera-Paz EF, Campbell M, Foster C,

Taub MA, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Salzberg SL. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30. https://doi.org/10.1038/s41588-018-0273-y.
33. Li Q, Tian S, Yan B, Liu CM, Lam T-W, Li R, Luo R. Building a Chinese pan-genome of 486 individuals. Commun Biol. 2021;4:1–14. https://doi.org/10.1038/s42003-021-02556-6.
34. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J, Marinescu VD, Alföldi J, Harris RS, Lindblad-Toh K, Haussler D, Karlsson E, Jarvis ED, Zhang G, Paten B. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51. https://doi.org/10.1038/s41586-020-2871-y.
35. Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. Genome Biol. 2020;21:124. https://doi.org/10.1186/s13059-020-02038-8.
36. Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, Eberle MA. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biol. 2019;20:291. https://doi.org/10.1186/s13059-019-1909-7.
37. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang P-C, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JI, Haussler D, Garrison E, Paten B. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871. https://doi.org/10.1126/science.abg8871.
38. Tognon M, Bonnici V, Garrison E, Giugno R, Pinello L. GRAFIMO: Variant and haplotype aware motif scanning on pangenome graphs. PLOS Comput Biol. 2021;17:e1009444. https://doi.org/10.1371/journal.pcbi.1009444.
39. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, Dilthey AT, Marschall T. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet. 2022;54:518–25. https://doi.org/10.1038/s41588-022-01043-w.
40. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL. A whole-genome assembly of the domestic cow Bos taurus. Genome Biol. 2009;10:R42. https://doi.org/10.1186/gb-2009-10-4-r42.
41. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, Lee C, June Ko B, Chaisson M, Gedman GL, Cantin LJ, Thibaud-Nissen F, Haggerty L, Bista I, Smith M, Haase B, Mountcastle J, Winkler S, Paez S, Howard J, Vernes SC, Lama TM, Grutzner F, Warren WC, Balakrishnan CN, Burt D, George JM, Biegler MT, Iorns D, Digby A, Eason D, Robertson B, Edwards T, Wilkinson M, Turner G, Meyer A, Kautt AF, Franchini P, Detrich HW III, Svardal H, Wagner M, Naylor GJP, Pippel M, Malinsky M, Mooney M, Simbirsky M, Hannigan BT, Pesout T, Houck M, Misuraca A, Kingan SB, Hall R, Kronenberg Z, Sović I, Dunn C, Ning Z, Hastie A, Lee J, Selvaraj S, Green RE, Putnam NH, Gut I, Ghurye J, Garrison E, Sims Y, Collins J, Pelan S, Torrance J, Tracey A, Wood J, Dagnew RE, Guan D, London SE, Clayton DF, Mello CV, Friedrich SR, Lovell PV, Osipova E, Al-Ajli FO, Secomandi S, Kim H, Theofanopoulou C, Hiller M, Zhou Y, Harris RS, Makova KD, Medvedev P, Hoffman J, Masterson P, Clark K, Martin F, Howe K, Flicek P, Walenz BP, Kwak W, Clawson H, Diekhans M, Nassar L, Paten B, Kraus RHS, Crawford AJ, Gilbert MTP, Zhang G, Venkatesh B, Murphy RW, Koepfli K, Shapiro B, Johnson WE, Di Palma F, Marques-Bonet T, Teeling EC, Warnow T, Marshall Graves J, Ryder OA, Haussler D, O'Brien SJ, Korlach J, Lewin HA, Howe K, Myers EW, Durbin R, Phillippy AM, Jarvis ED. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46. https://doi.org/10.1038/s41586-021-03451-0.

Publisher's Note