Modernizing the Bovine Reference Genome Assembly

B.D. Rosen[1], D.M. Bickhart[2], R.D. Schnabel[3], S. Koren[4], C.G. Elsik[3], A. Zimin[5], C. Dreischer[6], S. Schultheiss[6], R. Hall[7], S.G. Schroeder[1], C.P. Van Tassell[1], T.P.L. Smith[8] & J.F. Medrano[9]

[1] Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD, USA
ben.rosen@ars.usda.gov (Corresponding Author)
[2] Dairy Forage Research Center, USDA-ARS, Madison, WI, USA
[3] University of Missouri, Columbia, MO, USA
[4] National Human Genome Research Institute, NIH, Bethesda, MD, USA
[5] University of Maryland, College Park, MD, USA
[6] Computomics GmbH & Co. KG, Tuebingen, Germany
[7] Pacific Biosciences, Menlo Park, CA, USA
[8] U.S. Meat Animal Research Center, USDA-ARS, Clay Center, NE, USA
[9] University of California, Davis, CA, USA

Summary

The draft assembly of the bovine genome was first released in 2004. Sanger sequencing was used to assemble the genome of the Hereford cow L1 Dominette 01449. The assembly has seen vast improvements over the years but the limitations of the Sanger reads that form the basis of the assembly mean that many issues remain. Recent advances in long-read sequence technology, combined with new scaffolding technologies, have made it possible to create a completely new *de novo* assembly of the Dominette genome. Approximately 80x genome coverage of PacBio sequence was *de novo* assembled with Falcon, this was followed by scaffolding with Dovetail Genomics Chicago data, the BtOM1.0 Optical Map and a recombination map of 59K autosomal SNPs yielding chromosome length scaffolds. The scaffolded assembly was then refined with independent de-novo assemblies from CANU and MaSuRCA, error corrected with an independent genetic map, and polished with 50x Illumina reads. Assembly statistics include an N50 contig size of 26 Mb with 393 gaps representing many fold improvements over UMD3.1 (contig N50=0.097 Mb, 72,051 gaps). Additionally, full-length transcripts from 28 Dominette tissues have been sequenced with PacBio using the Iso-Seq method to support improved annotation. A public version of the new ARS-UCD assembly is expected to be available before the start of the conference.

Keywords: cattle, assembly, scaffolding, next-generation sequencing, PacBio, Dovetail

Introduction

The first draft of the bovine genome was released in September of 2004 as part of a $53 million international effort. While far from perfect, the genome, derived from the Hereford cow L1 Dominette 01449, provided a boon to cattle genomics research and spurred the development of multiple genomic resources in cattle and other ruminants (Table 1). Updates and new assembly releases through the years (Table 1) have led to great progress from that first release but many issues remain with the current assemblies (Florea *et al*., 2011, Zimin *et al*., 2012, Zhou *et al.* 2015, Whitacre *et al*., 2015, Utsunomiya *et al*., 2016).

Table 1. Important Dates in the Bos taurus *Reference Genome History*.

Date Released	**Release Name**	**Coverage**	**Comments**
September 2004	Btau_1.0	3x	Preliminary assembly using (WGS) reads from small insert clones.
March 2005	Btau_2.0	6.2x	BAC end sequences added to assembly
August 2006	Btau_3.1	7.1x	BAC sequences added
October 2007	Btau_4.0	7.1x	Chromosome mapping refined
December 2007	-	-	BovineSNP50 BeadChip available
March 2008	-	-	Bovine Gene Atlas
January 2009	-	-	Official genomic evaluations for dairy cattle
April 2009	-	-	Bovine HapMap
October 2009	Btau_4.5	7.1x	Additional WGS contigs and high quality finished sequence added
December 2009	UMD3.1	9.5x	Based on same original sequence data as Btau assemblies. Adopted by community as primary reference assembly
February 2010	Oar_v1.0	-	First draft of the sheep genome relied heavily upon bovine genome during assembly
March 2010	-	-	BovineHD BeadChip available
July 2012	Btau_4.6.1	7.1x	Additional high quality finished sequence added
January 2013	CHIR_1.0	-	First draft of the goat genome anchored chromosomes using conserved synteny with cattle
December 2015	Btau_5.0.1	25x	19x PacBio data used to fill gaps in UMD3.1 assembly with PBJelly

One of the greatest challenges in assembling genomes arises from repetitive DNA. If the length of your sequence read is shorter than the size of your repeat, there will be ambiguity in aligning and assembling that repeat. Unfortunately, mammalian genomes are highly repetitive. Although the Sanger based sequencing used to produce the existing references is superior in this regard to the short-read sequencing that followed, it doesn't possess the read length required to resolve many repetitive regions of the genome. For instance, BovB, a long interspersed element (LINE), is one of the most common repeats in the bovine genome, estimated to make up ~20 of the genome (Walsh *et al.*, 2013). At 3.2 kb in length, BovB is approximately 3 times as long as the typical Sanger read. Recent improvements in long single-molecule read technologies as well as assembly and scaffolding methods made it clear that a new *de novo* reference assembly could be undertaken at minimal cost to greatly improve upon the existing assemblies (Bickhart *et al.*, 2017).

Results

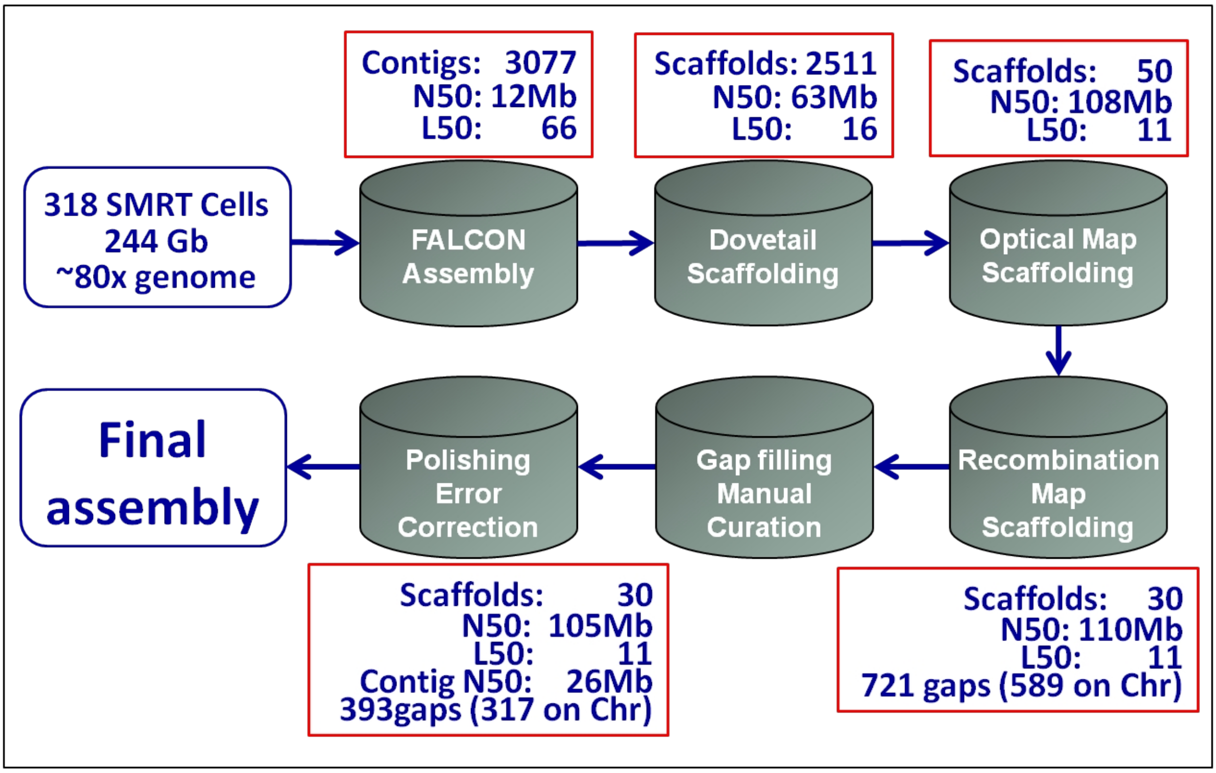

Figure 1. Dominette de novo *assembly process. N50 is the minimum scaffold/contig length needed to cover 50% of the genome. L50 is the number of contigs required to reach N50.*

Falcon Assembly

318 SMRT cells were sequenced on a PacBio RS II yielding 244 Gb of sequence with an average read length of 20 kb. Reads were assembled using the Falcon *de novo* genome assembler (verison 0.4.0) (Chin *et al.*, 2016). A length cutoff of 10,000 was used for the initial seed read alignment, and a secondary cut of 8,000 for the preassembled reads before layout of the assembly. The assembly resulted in 3077 primary contigs covering 2.7 Gb with a contig N50 of 12 Mb (Fig1). Polishing of the assembly was carried out to improve base accuracy (Chin *et al.*, 2013). Raw data was mapped back to the assembly using blasr (Chaisson and Tesler, 2012), and a new consensus called with the Quiver algorithm. Mapping and consensus was carried out using the resequencing pipeline from the SMRT Analysis 3.1.1 software package (Pacific Biosciences, Menlo Park, CA).

Dovetail Scaffolding

A Chicago library was prepared as described previously (Putnam *et al.*, 2016). Briefly, 500ng of HMW gDNA (>50kb mean fragment size) was reconstituted into chromatin in vitro and fixed with formaldehyde. Fixed chromatin was then digested with DpnII, the 5' overhangs were filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was sheared to ~350 bp mean fragment size, and sequencing libraries were generated using NEBNext Ultra

(NEB, Ipswich, MA). Biotin-containing fragments were then isolated using streptavidin beads before PCR enrichment of the library. The library was sequenced on an Illumina HiSeq 2500 (Illumina, San Diego, CA) to approximately 84x coverage.

The Falcon assembly and Chicago library read pairs in FASTQ format were used as input data for HiRise, a software pipeline purpose-built for using Chicago data to scaffold genomes (Putnam *et al*., 2016). Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analysed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs resulting in 2511 scaffolds with an N50 of 63 Mb and L50 of 16.

Optical Map Scaffolding

We used the *Bos taurus* optical map BtOM1.0 (Zhou *et al.* 2015) that spans 2,575,30 Mbp and comprises 78 optical contigs to further scaffold the Dovetail assembly. The IrysView software package (BioNano Genomics, San Diego, CA) was used to map the assembly scaffolds to the optical map contigs. After a manual curation step where false joins and misassembled contigs were detected by inspection of the alignment, the number of scaffolds could be reduced to 50 while the scaffold L50 decreased to 11 and the scaffold N50 increased to 108 Mbp.

Recombination Map Scaffolding

The bovine recombination map constructed by Ma *et al.* (2015) was used to detect mis-assemblies and scaffold the assembly. We mapped the almost 54k SNP markers distributed over 30 linkage groups to the optical map scaffolds with BLAST and required a mapping identity of 98% over the full marker sequence length. Only unique mapping SNPs were considered. Scaffolds were broken when two or more markers from different linkage groups aligned to them. Pearson correlation coefficient was used to calculate the most probable scaffold orientation based on the marker alignment order. This yielded an assembly with chromosome length scaffolds and only 721 gaps. Another round of polishing was undertaken with Arrow, an update to the Quiver algorithm, with the SMRT Analysis 3.1.1 software package.

Gap filling

Gap filling was first done by aligning two Canu (Koren *et al*., 2017) assemblies (one run with overlaps computed for error correction by MHAP and one with minimap) to the scaffolded assembly and identifying filled gaps. A gap was filled if either assembly had support where there was >5000 bases aligning on either side of the gap up to at most 10bp away from the gap. In the case of a negative gap (i.e. the assemblies had a collapse), both assemblies had to agree on the position and size of the collapse to fill the gap. In total 171 gaps were closed with this approach. Finally, PBJelly (English *et al*., 2014) was used to fill an additional 91 gaps. The closing of gaps between contigs brought the contig N50 up from 12 Mb to 21 Mb and the number of gaps in the genome down to 459.

Manual curation

Following gap filling, the X chromosome was manually curated using two assemblies produced from MaSuRCA (Zimin *et al.*, 2016) error corrected reads (PacBio corrected with Illumina). One assembly used Canu to assemble and the other used Celera Assembler version 8.3 (Berlin *et al.*, 2015). Alignments between these two assemblies and the gap filled assembly were used to confirm or revise the order and orientation of X chromosome contigs as well as place additional unplaced contigs and scaffolds.

The resulting assembly structure was then re-assessed with an independent genetic map UMCLK, produced at the University of Missouri (unpublished data). Briefly, SNP50 genotypes from over 40,000 individuals representing both beef and dairy cattle breeds were assembled into large paternal half-sib families with an average of 6,386 informative meiosis for 46,133 loci. CRI-MAP (Green et al., 1990) was used to generate an initial LOD3 map (8,556 loci) followed by inclusion of additional markers based on two-point LOD scores. Local marker ordering was refined by iteratively using the flips option in 100 marker windows overlapping by 50 markers. The Blat alignment tool (Kent 2002) was used to map the probe sequence and flanking sequences to identify misassemblies. Assembly gaps, Illumina read depth coverage and alignments with dbSNP sequences and flanking sequences were used to refine breakpoints for sequence rearrangements. In all, corrections were made to chromosomes 1,2,5-12,16,18-21,23,26,27, and X.

Post error correction, PBJelly was again run to close the remaining gaps. The number of gaps dropped from 459 to 393 indicating that our corrections resulted in gaps possessing flanking sequences which PBJelly could now use to fill a gap that could not previously be filled. The remaining gaps represent regions where either the gap is too large for our PacBio reads to span, read coverage is low or missing, or there is a remaining misassembly. The contig N50 also increased again to 26 Mb.

Polishing

For polishing, we ran a final iteration of Arrow followed by Pilon (Walker *et al.*, 2014). Input to Pilon was 50x of 80X Illumina coverage produced for polishing and independent assembly validation. DNA from Dominette lung tissue was sheared to ~350 bp mean fragment size, sequencing libraries were generated using Illumina TruSeq PCR-Free kits and run on an Illumina NextSeq. We aligned sequence reads with BWA. The final version of the genome will be available from NCBI under the accession NKLS00000000. The version described in this paper is version NKLS01000000.

Conclusions

The ARS-UCD assembly represents a vast improvement in the continuity of the bovine genome. The quality of the assembly including the 100-fold improvement in the number of gaps compared to the Btau_5.0.1 assembly and almost 200-fold improvement over UMD3.1 (Table 2) indicate that this new assembly will have a substantial impact on genetic and molecular genetic studies. Many previous studies might benefit from the re-mapping of reads and/or analysis of GWAS with improved marker order.

Table 2. Bos taurus *Reference Genome Comparisons.*

	UMD3.1	Btau_5.0.1	ARS-UCD
Contigs	75,195	44,658	2603
Contig N50 (Mb)	0.097	0.26	26.3
Contig L50	7,885	3,033	31
Gaps	72,051	41,517	393
Assembly size (Gb)	2.670	2.724	2.737

List of References

Berlin K., Koren S., Chin C.S., *et al*., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotech. 33(6): 623-630.

Bickhart D.M., Rosen B.D., Koren S., *et al*., 2017. Single-molecule sequencing and conformational capture enable de novo mammalian reference genomes. Nat Gen. 49 643-650.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. Sep 19;13:238.

Chin C.S., Alexander D.H., Marks P., *et al*. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth. 10(6), 563–569.

Chin C.S., Peluso P., Sedlazeck F.J., *et al*., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Meth. 13(2): 1050-1054.

English A.C., Salerno W.J., Reid J.G. 2013. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics 15, 180.

Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. 2011. Genome Assembly Has a Major Impact on Gene Content: A Comparison of Annotation in Two Bos Taurus Assemblies. PLOS ONE. Jun 22;6(6):e21400.

Green, P., Falls, K., and Crooks, S. 1990. Documentation for CRI-MAP, version 2.4. Washington University School of Medicine, St. Louis, MO.

Kent W.J., 2002. BLAT – the BLAST-like alignment tool. Genom. Res. 12(4): 656-664.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. Mar 15;gr.215087.116.

Ma L., O'Connell J.R., VanRaden P.M., *et al*. 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLOS Genet 11(11): e1005387.

Putnam N.H., O'Connell B.L., Stites J.C., *et al*., 2016. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. Genom. Res. 26: 342-350.

Utsunomiya ATH, Santos DJA, Boison SA, Utsunomiya YT, Milanesi M, Bickhart DM, *et al*. 2016. Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. BMC Genomics. Sep 5;17:705.

Walker B.J., Abeel T., Shea T., *et al*., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9: e112963.

Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL. 2013. Widespread horizontal transfer of retrotransposons. Proc Natl Acad Sci USA. Jan 15;110(3):1012–6.

Whitacre LK, Tizioto PC, Kim J, Sonstegard TS, Schroeder SG, Alexander LJ, *et al*. 2015. What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual. BMC Genomics.16:1114.

Zhou S., Goldstein S., Place M., *et al*., 2015. A clone-free, single molecule map of the domestic cow (*Bos taurus*) genome. BMC Genom. 16(1): 644.

Zimin A, Kelley D, Roberts M, Marçais G, Salzberg S, Yorke J. 2012 Mis-assembled "segmental duplications" in two versions of the Bos taurus genome. PLoS One.;7(8):e42680

Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Yorke JA, *et al*. 2016. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm. bioRxiv. Jul 26;066100