Proc. Assoc. Advmt. Anim. Breed. Genet. 22:161-165

THE NEW BOVINE REFERENCE ASSEMBLY AND ITS VALUE FOR GENOMIC RESEARCH

Juan F. Medrano

Department of Animal Science, University of California Davis, Davis, CA, USA

The development of the assembly has been a major joint effort of several groups making efficient use of very limited resources. **Participants in this effort have been:** T.P.L. Smith (USDA, ARS, USMARC, Clay Center, NE), B.D. Rosen (Animal Genomics and Improvement Laboratory, ARS USDA, Beltsville, MD), S. Koren (National Human Genome Research Institute, Bethesda, MD), A. Zimin (University of Maryland, College Park, MD), R.D. Schnabel (Division of Animal Sciences, University of Missouri, Columbia), D. Bickhart (Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, WI), R. Hall (Pacific Biosciences, Menlo Park, CA), S.J. Schultheiss and C. Dreischer (Computomics GmbH, Tuebingen, Germany). Funding for the project was provided by USDA/NRSP-8 Animal Genome, USDA-ARS Meat Animal Research Center, Neogen and Zoetis.

SUMMARY

There are two public cattle genome reference assemblies (UMD3.1.1 and Btau5.0.1) that were based primarily on the same set of data. Both assemblies used sequences of a minimum tiling path of BAC clones from the CHORI-240 library (prepared using DNA from L1 Domino 99375), augmented by low coverage whole genome shotgun sequencing (WGS) from his daughter, L1 Dominette 01449. Updates and new assembly releases through the years have led to significant improvements, but as confirmed by the recently developed cattle genome optical map (BtOM1.0), there are numerous differences between these assemblies that have produced ambiguities that continue to impact and hamper genomic analysis in cattle. Recent advances in long-read sequence technology, combined with new scaffolding technologies, have made it possible to create a completely new *de-novo* Dominette assembly. An approximately 80X PacBio FALCON based de-novo assembly, followed by scaffolding with Dovetail Genomics Chicago library/HiRise technology, the BtOM1.0 Optical Map of Dominette and a recombination map of 59K autosomal SNPs. The scaffolded assembly was then refined with independent *de-novo* assemblies from CANU and MaSuRCA, yielding chromosome length scaffolds. Preliminary assembly statistics include an N50 contig size of 22 Mb and an N50 scaffold size of 104 Mb representing several fold improvements over UMD3.1 (contig N50=0.97Mb, scaffold N50=6.4Mb). Additionally, full-length transcripts from 30 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 released in mid 2017. An update on the status of the long-read based assembly of Dominette will be presented here, providing some perspective on the value of having an improved bovine reference sequence.

INTRODUCTION

The availability of accurate well-annotated genome assemblies in agricultural species have become essential tools to enable the understanding of phenotypic variation and practical applications of DNA technologies. In cattle, numerous opportunities exist in the application of genomic selection and of new technologies, like gene editing to improve production efficiency. For the human and mouse genomes, enormous efforts and resources have been spent to develop what has been referred to as Gold (targeted finishing with haplotype resolution of critical regions) and Platinum (contiguous haplotype-resolved representation of the entire genome) level genome sequence assemblies. In order

to approach some of these advanced states of genome resolution in cattle a significant effort has been placed towards developing a new improved Dominette assembly.

There are currently two public cattle genome reference assemblies (UMD3.1.1 and Btau5.0.1), that were based primarily on the same set of data. Both assemblies used sequences of a minimum tiling path of BAC clones from the CHORI-240 library (prepared using DNA from L1 Domino 99375), augmented by low coverage whole genome shotgun sequencing (WGS) from his daughter, L1 Dominette 01449. Subsequently, the Btau5.0.1 assembly was improved by gap filling with low coverage long-read WGS, and more recently with P5 PacBio reads. Scaffolding used combinations of radiation hybrid map and genetic linkage map data, as well as a large number of BAC end sequences. Recently, a cattle genome optical map (BtOM1.0) was developed (Zhou et al. 2015), which confirmed that there are numerous differences between these assemblies that have produced ambiguities that continue to impact and hamper genomic analysis in cattle.

Two animals (Domino and Dominette) were used to produce the assemblies that resulted in an increased amount of diversity between haplotypes. Both assemblies, although having used practically the same sequenced data, are significantly different, appearing as assembly errors, genome segmental inversions, chromosomal placements, sequence gap numbers and discrepancies of sequence coverage of the bovine genome. Although there have been periodic updates of both sequence assemblies, many issues still remain in both. It is very difficult when one encounters discrepancies between assemblies to know what is correct, and this impacts genomic studies. Table 1 shows a comparison of the genomic statistics of both assemblies, UMD3.1.1 and BTAU 5.0.1.

Table 1. Comparison of current cattle genome assemblies (NCBI report)

UMD3.1.1 (Reported April 2009 (Genome Biol)	**BTAU 5.0.1** (Released (11/19/2015)
Based on 9x Sanger coverage WGS Dominette	Based on 9x Sanger coverage of Dominette
BAC path Domino	BAC ends, + 19x coverage P5 PacBio
RH map and human-cow synteny map	BAC end, RH map, PBJelly2
Genomic statistics:	Genomic statistics:
75,618 contigs (97 kb contig N50) 42,267 contigs (276 kb contig N50)	42,267 contigs (276 kb contig N50)
6337 scaffolds (6.4 Mb scaffold N50)	5,998 scaffolds (6.8 Mb scaffold N50)
3,193 gaps between scaffolds	2,856 gaps between scaffolds

GENOME ASSEMBLY PROCESS

Creating a genome assembly is the process of reconstructing a genome to develop a database of DNA sequences that represent an example individual of the species, from a collection of short or long sequence reads. A de-novo assembly is performed without the aid of a reference genome and the genome is reconstructed by directly reconstructing the puzzle of sequence reads. One complicating factor in the reconstruction is the presence of repetitive sequences, particularly when using short read lengths that do not span the length of the repeats. Recent advances in Pacific Biosciences Smrt single-molecule sequencing technologies, with the generation of 20-50 kb have allowed resolving repetitive sequences and the creation of accurate genome assemblies (Berlin et al. 2015).

The repetitive content of genomes on both large and small scales, including structures near centromeres and telomeres, large paralog gene families, like zinc fingers, and the distribution of

interspersed nuclear elements such as LINEs and SINEs are the cause of many of the incorrect assembly problems we have had in the past. Such difficult-to-assemble content composes large portions of eukaryotic genomes, about 60-70% of the human genome (de Koning et al 2011).

Although PacBio reads are error prone, errors are at random and can be overcome by sufficient coverage producing highly accurate assemblies, and it has been demonstrated from assemblies in humans and other organisms that single-molecule sequencing can produce de-novo near complete eukaryotic assemblies that are 99.99% accurate compared to the available references. In addition to the technical quality of the assemblies, the time to produce an assembly has been reduced by 5x and cost by more than 200 orders of magnitude.

The creation of the final assembly is an iterative process that evolves as scaffolds and super-scaffolds are built. The initial assembly process used in the creation of the Dominette *de-novo* assembly is shown in Figure 1.

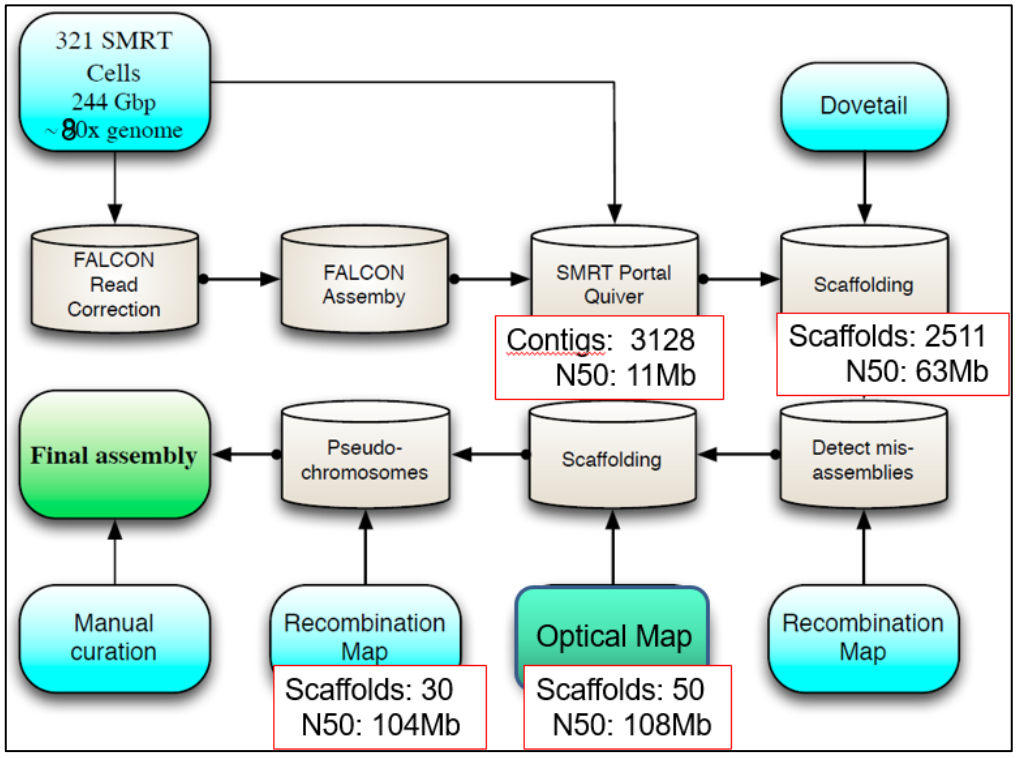

Figure 1. Initial Dominette *de-novo* assembly (January 2017)

Approximately 321 PacBio SMRT cells with an average size of 20 kb were produced for a ~80x genome coverage followed by a hierarchical genome assembly process of PacBio long reads using FALCON, and Quiver for polishing. Quiver generated consensus contigs using local realignment of reads to the assembly to correct short insertions, deletions and substitutions errors. Following the construction of contigs from pre-assembled reads, the true assembly process is dependent on the correct orientation of contigs into scaffolds and super-scaffolds. This required other data types, like

Dove Tail Chicago libraries, recombination map, optical map, alignment of short read sequences and manual curation.

Among the initial scaffolding resources used in the Dominette assembly were a Dovetail Chicago Hi-Rise library, the optical map and a genetic map. The Dovetail method is based on producing DNA linkages of up to several hundred kb to make sequencing libraries that can link distant fragments. This long-range mate pair data can be used to orient contigs and largely improve scaffolding in *de-novo* assemblies (Putnam et al. 2016). The Optical Map is a high throughput system that produces ordered restriction maps from individual molecules of genomic DNA (Zhou et al. 2015). The approach is ideal for identifying structural variants and studying genome structure. The recombination map used is a sex-specific recombination map of 59K autosomal SNP (Ma et al. 2015). The map was used primarily for breaking, ordering and orienting contigs.

Depending on how all the data is used and how the scaffolding resources are applied one ends up with several assemblies in which some super-scaffolds are better assembled or one assembly captures longer scaffolds and contigs. All this needs to be carefully examined in order to develop a stable assembly for further improvement. Early statistics after the initial scaffolding of the new Dominette assembly are shown in Table 2.

Table 2. Early progress statistics of the new Dominette assembly ARS-UCD v1.0 (Jan, 2017)

Based on ~80x PacBio data P6 chemistry, Falcon assembly - Quiver (PacBio)
Scaffolding: Dovetail HiRise, Optical Map, Rec Map
Genomic statistics:
2816 contigs (22.6 Mb contig N50)
30 scaffolds (104 Mb scaffold N50, L50 12)
Largest scaffold length 211 Mb (Chr 1)
460 gaps

For sequence annotation, full-length transcripts spanning entire isoforms using the PacBio Iso-Seq method from approximately 30 Dominette tissues are being developed. We expect to release an assembly with haplotype-resolved chromosomes.

CONCLUSIONS

- Long single-molecule reads, despite higher per-read error rate, create higher quality reference genomes at a fraction of the cost of earlier technologies.
- The improvement in quality of the cow assembly will have substantial impact on many genetic and molecular genetic studies.
- Many studies would benefit from re-mapping reads, and/or analysis of GWAS with improved marker order.
- The improvements in the cow assembly are substantial enough that it is worth considering waiting for them for ongoing GWAS and WGR studies.
- We expect to have a version of the new ARS-UCD assembly available this summer-2017, through NCBI.

REFERENCES

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015). *Nat Biotechnol*. **33**: 623.
de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD (2011). *PLoS Genet.* **7**(12):e1002384

Ma, L., O'Connell, J.R., VanRaden, P.M., Shen, B., Padhi, A., Sun, C., Bickhart, D.M., Cole, J.B., Null, D.J., Liu, G.E., et al. (2015). *PLOS Genet* **11**: e1005387.

Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J., Fields, A., Hartley, P.D., Sugnet, C.W., et al. (2016). *Genome Res.* **26**: 342–350.

Zhou, S., Goldstein, S., Place, M., Bechner, M., Patino, D., Potamousis, K., Ravindran, P., Pape, L., Rincon, G., Hernandez-Ortiz, J., et al. (2015). *BMC Genomics* **16**: 644.