Large-Scale Bull Genome Sequencing Enables Rapid Livestock Improvement

Whole-genome sequencing data from the 1000 Bull Genomes Project is aiding discoveries of positive and negative traits, benefiting herds globally.

Introduction

Performing whole-genome sequencing (WGS) of bovine individuals is the optimal way to use current genomic analysis technology to assess genetic variation of a cattle breed. However, this option is still out of financial reach for most breeders and researchers. As a result, routine genotyping of bulls, cows, and heifers is performed using less expensive single nucleotide polymorphism (SNP) arrays. SNP arrays are designed to discover gene variants associated with positive production and health traits, or with disease and other negative traits.

In 2012, Ben Hayes, PhD, then of Agriculture Victoria[1], founded the 1000 Bull Genomes Project[2] to aid in the global understanding of bovine genetics and foster international collaboration.[3] The project's initial run sequenced 238 animals from an Australian key ancestor bull selection line at 10.5× average genome coverage on HiSeq™ 3000 Systems. The project now includes 40 international partners, 2700 dairy and beef animals, and has identified close to 90 million genetic variants. Researchers and breeders around the world are benefiting from discoveries made with this data, including the identification of lethal mutations[4] and the largest scale sequence level genome-wide association study (GWAS) in cattle.[5] Hans Daetwyler, PhD, Senior Research Fellow at La Trobe University and Agriculture Victoria in Melbourne, Australia, now leads the 1000 Bull Genomes Project. His team is planning another run within the next year to include even more animals.

iCommunity spoke with Dr. Daetwyler to learn about how the project came to be, its various findings, and what the future holds for the use of WGS to identify variants associated with positive and negative traits in other bred species.

Q: What is the 1000 Bull Genomes Project?
Hans Daetwyler (HD): The 1000 Bull Genomes Project was originally proposed by Dr. Ben Hayes, who was my boss at Agriculture Victoria. At the time, sequencing was expensive to perform on large numbers of animals and that was preventing its widespread use in research of old and new breeds. Institutions didn't have the funds to sequence enough individuals to enable imputation, which statistically infers the unobserved genotypes in low-density SNP-array data.

The idea behind the 1000 Bull Genomes Consortium was to compare whole-genome sequences of cattle, including *Bos Taurus*, *Bos Indicus*, and other *Bos* species. When we published our first paper, the 1000 Bull Genomes Project included 234 cattle individuals and 5 partner institutions worldwide.[5] Since then, we've run a new round of analyses almost every year. Each run has added more animals, with more partner institutions joining the project. We have sequenced more than 2700 animals to date.

Q: What was your role when the 1000 Bull Genomes Project began?
HD: I performed data analysis for the first few years of the project. Partners sent us their whole-genome sequences aligned to a reference genome. We analyzed the BAM files, combined the data, and ran a variant caller to identify the SNPs and indels in the data set. We provided our partners with a list of raw and filtered SNPs and the associated genotypes for all animals in the project. We continue that process today.

When Dr. Hayes moved to the University of Queensland in 2016, I took over as chair of the steering committee for the 1000 Bull Genomes Project.

Q: What tools and methods had researchers been using to identify causal variants?
HD: Ten years ago, they were using low-density SNP arrays to identify an associated genetic region and might have performed targeted sequencing of that region to identify other variants. At that time, sequencing was a slow, expensive process. Researchers would have used older NGS technology to perform targeted sequencing and then only in animals they suspected of being carriers of a quantitative trait locus (QTL). They didn't have a reference genome, so they didn't have a good idea of where to look or the number of genes involved. They likely identified only a handful of mutations using that approach.

Hans Daetwyler, PhD, Senior Research Fellow at La Trobe University and Agriculture Victoria in Melbourne, Australia.

The 1000 Bull Genomes Project database consists of whole-genome sequences based on many animals. It has accelerated the causal mutation discovery process significantly and improved the genetic gain in herds worldwide.

Q: Why is it important to improve the rates of genetic gain in dairy and beef cattle?
HD: Other than nutrition and health management, genetic change is a major component of increasing productivity and efficiency, as well as improving health welfare traits within a herd. Genetic gain is cumulative. Over time, it acts like compound interest. Each time you make a positive change genetically in favor of better performance or improved health and welfare, it remains in the herd and the positive effect is compounded. Quite a significant component of farmers' effort is spent increasing the productivity of their herds over time.

> "If farmers have a more accurate EBV for a calf at birth or at a very young age, they can confidently select and use those individuals for breeding much earlier than they could have otherwise. It shortens the dairy bull generation interval from 5-6 years to 2 years."

Q: What are the benefits of genomic selection compared to previous selective breeding approaches?
HD: Before genomics, farmers used phenotypic selection, which involved observing the individual and its progeny, and selecting individuals for breeding based on their characteristics. They would have also used pedigree selection, which uses information on close relatives.

The power of genomic selection is that it combines that information with production, efficiency, and health data on more distant relatives. Genomic selection leads to an increase in the accuracy of the estimated breeding value (EBV), especially for young individuals. If farmers have a more accurate EBV for a calf at birth or at a very young age, they can confidently select and use those individuals for breeding much earlier than they could have otherwise. It shortens the dairy bull generation interval from 5-6 years to 2 years.

Q: What percentage of bovine livestock breeders are using genomic selection to improve their herds?
HD: The use of genomic selection differs between dairy and beef farmers. In contrast to phenotypic selection, genomic selection enables the use of young bulls that don't have daughters. The use of these young genomically tested sires has increased dramatically in the last few years. In some countries, the rate is over 80% of the total artificial insemination (AI) sires that are used. In Australia, the use of genomically tested sires is around 40%. In Australian beef, it would be a lower proportion than for dairy, but again quite a bit higher for the major beef breeds in North America.

Q: What is the value of sequencing ancestor bulls with NGS versus Sanger sequencing?
HD: The search for causal or near-causal mutations is only possible if you have large data sets consisting of whole-genome sequences from NGS. NGS enables us to sequence significantly more animals at a lower price point than Sanger sequencing. NGS has improved imputation accuracy and efficiency at which we can infer sequence phenotypes in individuals that only have a SNP array genotype assessment. That's the biggest advantage.

The benefits of NGS impact functional genomic studies as well. RNA-Seq and chromatin immunoprecipitation (ChIP)-Seq with NGS provide functional information on a set of individuals. We use that information in our search for near-causal mutations that we can then genotype directly. Genotyping these mutations directly also increases the prediction accuracy across breeds and in individuals that are less related to the training population.

Q: How are key ancestor bulls identified?
HD: There are several methods used to identify key ancestor bulls. The primary method is to select a pedigree and identify which individuals explain most of the genetic variants in that pedigree. Newer methods use genotypes or even haplotypes in that population to look for diversity. Key ancestors are chosen based on whether they possess the most haplotypes or a strong complement of haplotype sets represented in the population. Another method is to look at the individuals that have haplotypes that haven't been covered in the sequenced set of individuals.

> "All animals in the project have been or will be realigned to the new reference genome from the Agriculture Research Service at the University of California, Davis. We hope to have improved data for everybody going forward."

Q: What is the total number of individuals that have been sequenced and which cattle breeds are included in the 1000 Bull Genomes database?
HD: We've surpassed our original goal of 1000 individuals. In the last analysis run, we had more than 2700 individuals, and we're about to start a new run of 1000 bulls. All animals in the project have been, or will be, realigned to the new reference genome from the Agriculture Research Service at the University of California, Davis. We hope to have improved data for everybody going forward.

There are slightly more dairy breed groups than beef groups in the 1000 Bull Genomes database. The main breed group in the project is Holstein at about 20%. Angus is the next largest group, followed by Brown Swiss. We also have dual-purpose cattle in the

database, including Simmental and Fleckvieh. Recently, there have been quite a few *Bos indicus* contributions, including Brahman from Australia.

Q: How many new bovine markers have been identified by the 1000 Bull Genomes team?
HD: Before we started the 1000 Bull Genomes Project, researchers had used up to 600,000 variants in their analyses. In the first run that we performed with 240 animals, we identified 25–27 million SNPs and indels for *Taurus* alone. We're now at approximately 40 million for *Bos taurus* individuals only. When you include *Bos indicus* cattle, Yak, and other subspecies, it's about 80 million filtered variants.

> "The 1000 Bull Genomes database has accelerated the pace of bovine research through the earlier use of WGS data in animal breeding."

Q: What is the value of the 1000 Bull Genomes database to researchers?
HD: The 1000 Bull ReferenceGenomes data set is valuable to researchers in two ways. First, researchers can use it as a reference set to impute whole-genome sequences in data sets of herd individuals with SNP array genotypes. They can then perform powerful GWAS and investigate different genomic selection approaches that utilize WGS.[4]

It also enables researchers to look for causative or lethal recessive disease mutations. Using the 1000 Bull Genomes data set as a control, researchers can use a filtering strategy to narrow their searches to a small genomic region.

Q: Is this database available to any researcher anywhere in the world?
HD: The 1000 Bull Genomes database is available to research institutions that have joined the project and agreed to share their data with the consortium. There are few restrictions on the types of research performed with the data. However, researchers in the consortium are not allowed to share the data outside of their institution. For example, if a researcher has a collaborator who would like to analyze the 1000 Bull Genomes data, the collaborator would have to become a member of the project.

Currently, we have 38 institutions in the project worldwide. The 1000 Bull Genomes Project has fostered several significant collaborations. That's one of its lasting legacies.

Q: What sorts of discoveries have been made with the 1000 Bull Genomes database?
HD: The WGS data in the 1000 Bull Genomes database has been valuable, supporting numerous applied breeding and research studies. Researchers have used the 1000 Bull Genomes database to identify positive variants for several milk production traits.[6] There have also been several causal mutations that have been identified using the data. For example, our French collaborators discovered the causal mutations for embryonic lethal mutations, which previously hadn't been found, even though we knew that they existed.[7]

The sequence-level GWAS have improved our understanding of trait architecture and supported functional studies. For example, WGS data has enabled imputation for QTL studies. Researchers have also used WGS data to identify and prioritize SNP sets to improve genomic prediction.[8–11]

Q: How have these discoveries impacted breeders?
HD: The 1000 Bull Genomes database has accelerated the pace of bovine research through the earlier use of WGS data in animal breeding. The discoveries of lethal mutations have had an immediate positive impact on breeders. When the mutations were discovered, they were added to SNP arrays immediately to identify carriers in herds. These individuals are selected against for in the AI process, which decreases the frequency of the lethal mutations within the herd population.

We now have a more complete inventory of SNP and indel variations in all our populations and are able to design better SNP arrays. Rather than relying on random SNPs that are present at a frequency that's 'good enough' to infer they are causal, we can enrich arrays with SNPs that we know are causal and that affect traits directly.

> "The HiSeq Systems offer high throughput, so we can perform WGS, RNA-Seq, and ChIP-Seq at a low price point. The data quality is excellent and the systems have been our WGS workhorses."

Q: Which NGS systems do you use to perform WGS?
HD: We perform sequencing with two HiSeq 3000 Systems and we also have MiSeq™ and NextSeq™ 500 Systems. The HiSeq Systems offer high throughput, so we can perform WGS, RNA-Seq, and ChIP-Seq at a low price point. The data quality is excellent and the systems have been our WGS workhorses. We use the MiSeq System for applications where we need slightly longer reads.

We're investigating whether to upgrade our HiSeq Systems with a NovaSeq™ 6000 System, which would enable us to perform large-scale WGS and genotyping by sequencing (GBS) on one instrument.

Q: Is the 1000 Bull Genomes approach of creating a species database by sequencing 1000s of genomes being used for other livestock or plant species?
HD: We're using this approach in a similar project for sheep, SheepGenomesDB.[12] We've sequenced 935 sheep, downloaded raw data from the NCBI or EBI short-read sequence archives, and

processed it together with collaborators such as AgResearch in New Zealand and CSIRO in Brisbane. We performed variant calling, created the *.vcf file with the genes, SNPs, and indels, and published the data in the European variant archive.[13]

I think the concept would have merit in plants as well. I'm certain it has been used in *Arabidopsis* and in some major crops. However, some plants have very large genomes. The wheat genome has 17 billion base pairs. That makes WGS more expensive to perform and makes sharing data an imperative. We've performed exome sequencing in wheat, partly because of the lower price. In contrast, canola has a short genome of 1.2 billion base pairs. That makes it relatively inexpensive to perform WGS, even at 10×.

The other issue with plants is that some species are polyploid, making sequencing and genome assembly complicated. Polyploid plants have orthologous regions between the subgenomes where a shorter read can map to two, three, or four different places with obstinately equal accuracy. Typically, polyploid plant genomes are of lower quality than animal or human genomes.

> "It's possible that sequencing will become so inexpensive that GBS will replace SNP arrays. The higher output and lower priced NovaSeq 6000 System might change the cost structures in the future."

Q: Do you think GBS will ever replace SNP array genotyping?
HD: Most routine genotyping in cattle and sheep today is performed with SNP arrays, which currently provide high-quality data at a lower price point than GBS. There have been some issues with GBS because its low-coverage sequencing makes it hard to distinguish between sequence errors and the true alleles. SNP chips are also able to take lower quality DNA than sequence-based genotyping, which helps in industry applications.

That being said, I'm open-minded whether GBS or SNP arrays will be best in the future. I'm focused on whatever method provides the most high-quality data at the lowest price. It's possible that sequencing will become so inexpensive that GBS will replace SNP arrays. The higher output and lower priced NovaSeq 6000 System might change the cost structures in the future.

Q: What are the next steps for the 1000 Bull Genomes Project?
HD: We have a paper in development covering our new round of sequencing, data analysis, and reference genome. All our 38 partner institutions are also publishing data.

We've tested several variant callers and will be moving from SAMtools to a GATK HaploType Caller[14] for this next round of analysis.

We'll be increasing the 1000 Bull Genomes database to over 3000 cattle genomes and are including more public data. We're hoping that as we move to our new, larger reference genome that we'll see an increase in data quality and improved imputation to provide better outcomes.

Learn more about the systems mentioned in this article:

HiSeq 3000 System, www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000.html

MiSeq System, www.illumina.com/systems/sequencing-platforms/miseq.html

NextSeq 550 System, www.illumina.com/systems/sequencing-platforms/nextseq.html

NovaSeq 6000 System, www.illumina.com/systems/sequencing-platforms/novaseq.html

References

1. Agriculture Victoria, agriculture.vic.gov.au/agriculture. Accessed January 3, 2019.
2. 1000 Bull Genomes Project, www.1000bullgenomes.com/. Accessed January 3, 2019.
3. Hayes BJ and Daetwyler HD. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu Rev Anim Biosci*. doi: 10.1146/annurev-animal-020518-115024. Epub ahead of print.
4. Bouwman AC, Daetwyler HD, Chamberlain JA, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet*. 2018;50:362–367.
5. Daetwyler HD, Capitan A, Pausch H, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet*. 2014;46:858–865.
6. Pausch H, Emmerling R, Gredler-Grandl B, et al. Meta-analysis of sequence-based association studies across three cattle breeds reveals 25 QTL for fat and protein percentages in milk at nucleotide resolution. *BMC Genomics*. 2017;18:853.
7. Michot P, Fritz S, Barbat A, et al. A missense mutation in PFAS (phosphoribosylformylglycinamidine synthase) is likely causal for embryonic lethality associated with the MH1 haplotype in Montbéliarde dairy cattle. *J Dairy Sci*. 2017;100:8176–8187.
8. Brøndum RF, Su G, Janss L, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J Dairy Sci*. 2015;98:4107–4116.
9. VanRaden PM, Tooker ME, O'Connell JR, et al. Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol*. 2017;49:32.
10. Raymond B, Bouwman AC, Schrooten C, et al. Utility of whole-genome sequence data for across-breed genomic prediction. *Genet Sel Evol*. 2018;50:27.
11. MacLeod IM, Bowman PJ, Vander Jagt CJ, et al. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*. 2016;17:144.

12. SheepGenomesDB, Resources for the Sheep Genomics Community, sheepgenomesdb.org/. Accessed January 3, 2019.

13. European Variation Archive, www.ebi.ac.uk/eva/. Accessed January 3, 2019.

14. GATK. Haplotype Caller—Call germline SNPs and indels via local re-assembly of haplotypes. software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php. Accessed November 29, 2018.

Illumina, Inc. • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com