Proc. Assoc. Advmt. Anim. Breed. Genet. 22:201-204

1000 BULL GENOMES AND SHEEPGENOMEDB PROJECTS: ENABLING COST-EFFECTIVE SEQUENCE LEVEL ANALYSES GLOBALLY

H.D. Daetwyler[1,2,3], R. Brauning[4], A.J. Chamberlain[1], S. McWilliam[5], A. McCulloch[4], C.J. Vander Jagt[1], B. Sunduimijid[1,3], B.J. Hayes[1,6] and J.W. Kijas[5]

[1] Agriculture Victoria, AgriBio, Centre for AgriBioscience, VIC, Australia
[2] La Trobe University, VIC, Australia
[3] CRC for Sheep Industry Innovation, Armidale, NSW, Australia
[4] AgResearch Ltd, Mosgiel, New Zealand
[5] CSIRO, QLD, Australia
[6] Queensland University, QLD, Australia

SUMMARY

Whole-genome sequence data has several potential uses for animal breeding, including accelerated detection of mutations with deleterious and beneficial effects, as well as increasing the accuracy of genomic selection. It is cost-effective to share data in global consortia to enable more powerful imputation of sequence into animal populations that have been genotyped at lower density. This then facilitates more powerful downstream analyses such as genome-wide association and genomic prediction. Here we describe two such projects, namely the 1000 Bull Genomes Consortium and SheepGenomesDB.

INTRODUCTION

Whole-genome sequence provides detailed information of an individual's genetic make-up, which can be used to pinpoint genetic variants and genotypes for all animals in the sample. The accuracy of the analyses is dependent on the read depth, which is the average number of short reads (usually about 100 basepairs (bp) in length) aligned per base on the reference genome (e.g. 10 fold coverage).

Sequence data has several important uses. It can be used to track down Mendelian disorders and recessive alleles in affected individuals, using sequence variants from unaffected animals as controls enabling powerful filters to reduce the number of candidate mutations. Secondly, it enables more powerful genome-wide association studies, because either the causative mutations are themselves a sequence variant or would be in high linkage disequilibrium (LD) with a genotyped variant. Thirdly, it could improve the accuracy of genomic prediction, a benefit that can likely only be harnessed in multi-breed reference populations due the small effective population size of most commercial livestock breeds. For the latter two applications, the cost of sequencing has been prohibitive to sequencing the tens of thousands of individuals needed for powerful genome-wide association and genomic prediction. An alternative option is to impute sequence genotypes into animals that are already genotyped at lower density (preferably with a high density SNP chip). While the accuracy of imputation is not perfect, especially for lower minor allele frequency variants, it still has been demonstrated to increase the power of analyses. The need therefore arises for large reference populations of whole-genome sequenced animals for imputation.

The 1000 Bull Genome Project (1000 Bulls) and SheepGenomesDB Project are meeting this need in cattle and sheep, respectively. Many cattle breeds use the same sires across the globe, which makes populations very genetically connected and, therefore, using one reference population for imputation is advantageous. In addition, because sequence information captures even the short haplotypes shared across breeds, there is a benefit to using all breeds as a combined reference for imputation. Both projects have grown quickly with the 1000 Bulls now close to

2800 animals in Run6 and SheepGenomesDB with 935 animals in Run2. Described here is the organisation principles of both projects, bioinformatic pipelines, and the animals included and number of variants discovered in the latest analyses.

MATERIALS AND METHODS

In the 1000 Bulls the lead institution is Agriculture Victoria (AgVic) and each partner contributes their sequence data to the project. In turn, each partner receives all genetic variants and sequence genotypes discovered from animals in the project. There are currently 36 partners from 22 countries. The data is only available to participating partners and partners are expected to share identifying information, pedigrees and metadata for all animals where possible. Public cattle sequences not already included are also downloaded from the NCBI sequence read archive (SRA) and incorporated. The SheepGenomesDB project is organised differently and is jointly managed by CSIRO, AgResearch (AgR) and Agriculture Victoria (AgVic). It requires all included raw sequence data to be public at NCBI SRA, but allows animals to have anonymous identifiers with meta-data giving information on breed and sex (if known). As all input data is public, all variants and genotypes found are also made public.

Processing of Sequence. 1000 Bulls – Partners are responsible for processing and aligning sequences, which are then transferred to AgVic in BAM format for inclusion in analyses. Sequence quality scores must be Phred+33 encoded. The recommended quality control (QC) and processing of whole genome sequences in fastq format is as follows: 1) remove Illumina reads that fail the chastity filter; 2) remove adaptor sequence from reads; 3) trim low quality bases (Phred <20) from 5' and 3' ends of reads; 4) discard reads with a mean Phred quality score of <20 and >=3 bases not called (i.e. N); 5) remove known artifacts (e.g. Illumina NextSeq are known to have erroneous strings of A and/or G at 3' ends); 6) after trimming, discard reads that are too short (<50% of original read length). Reads that are left unpaired after QC may be aligned. Helpful programs for processing are quadtrim (https://bitbucket.org/arobinson/quadtrim), Picard (http://picard.sourceforge.net/index.shtml), Samtools (Li *et al.* 2009), GATK (DePristo *et al.* 2011), seqtk (https://github.com/lh3/seqtk), and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequences are then aligned to the UMD3.1 *bos taurus taurus* reference genome downloaded from the 1000 Bull Genomes website (www.1000bullgenomes.org) with the Burrows-Wheeler Aligner (*bwa align or bwa mem*) using default parameters (Li & Durbin 2009). BAM files must contain sample identifiers in Interbull format if available. One BAM files per animal per partner should be locally realigned, PCR duplicates removed, sorted, and indexed. Mean read depth should be calculated using GATK *DepthOfCoverage* and provided. Upon receipt of BAM files from partners, AgVic performs format and QC checks on files and amends them if needed where possible. Partners are encouraged to submit Bovine HD genotypes and any meta-data including pedigrees along with sequences.

SheepGenomesDB – The pipeline in use for this project is the same as for the 1000 Bulls except that all sequences must be publicly available at NCBI sequence read archive. Raw fastq files are downloaded by AgR and AgVic and then processed as above. Alignments are done to OAR3.1, ftp://ftp.ensembl.org/pub/release-78/fasta/ovis_aries/dna/Ovis_aries.Oar_v3.1.dna_sm.toplevel.fa.gz. Both AgR and AgVic, have a full set of BAM files for the project. For each animal, a standard SheepGenomesID is created which provides country of origin, breed, and sex of the animals (naming convention document on www.sheepgenomesdb.org).

Variant Calling, Filtering, and Refinement. 1000 Bulls – Samtools (currently version 1.3) *mpileup* is used to call single nucleotide polymorphisms (SNP) and short insertions and deletions (indels). This results in a variant call format (VCF) file which contains the following information: variant position, reference and alternative allele, quality metrics and read depth, genotypes for all animals at that positions, and genotype probabilities for all possible genotypes per animal per

position. The set of variants called at this stage will include low confidence variants for which there may not be enough evidence. Filtering of variants has been shown to improve the quality of the variant set as judged by the concordance of sequence and SNP chip genotypes at overlapping positions as well as the rate of opposing homozygotes (OppHom) found in parent-offspring pairs (there should not be any). Filtering is done with custom python scripts that use the VCF parser PyVCF (https://github.com/jamescasbon/PyVCF). Variants are removed if they have: 1) >1 alternative alleles; 2) no alternate allele observations in both forward and reverse direction reads; 3) overall quality score QUAL <20 and mapping quality score <30; 4) < minimum read depth of 10 or >3 standard deviations from the median read depth; 5) failed OppHom (>10% of parent-offspring pairs were OppHom); 6) the same bp position; 7) a proximity of <10 bp between indels or <3 bp between SNP in which case the lower QUAL variant was removed.

The resulting VCF files still contain a proportion of genotypes that are missing or called with high uncertainty. Imputation programs that are able to use genotype probabilities can be used to impute missing and refine uncertain calls using the haplotypes found in the collective set. In the 1000 Bulls we use Beagle 4.0 (Browning & Browning 2009) for this purpose. In cattle two separate analyses (Runs) are performed, one includes only *taurine* cattle and the other includes all animals.

SheepGenomesDB – The sheep pipeline is as above with the following differences. AgVic runs Samtools and AgR calls variants with GATK UnifiedGenotyper. Variants from both callers are then independently filtered as above, excluding the OppHom filter because only few Parent-Offspring pairs exist in the sheep sample to date. In addition, Samtools and GATK calls are merged to create two sets: 1) a unison set of calls with filtered overlapping variants, and 2) a complete set of calls that contain all unfiltered variants from both callers. Both sets are made public at the European Variant Archive (http://www.ebi.ac.uk/eva), which also annotates all variants and connects them to dbSNP.

Quality Control of Variants and Genotypes. Concordance of bovine HD SNP chip and sequence genotypes is performed for all animals in the 1000 Bulls if available. This concordance is expected to be >95% for good quality sequence (depending on read depth) and can typically be improved using Beagle. If concordance is <80% it may indicate that the SNP chip and sequence have not originated from the same animal and may highlight sample tracking issues. Parent-Offspring OppHom are checked for all pairs and should be less 0.1%. Furthermore, the number of singletons and heterozygosity per animal are calculated. If an animal has a very large number of singleton variants and it has breed contemporaries, it would indicate an issue with its data. Similarly, if heterozygosity is very high, it indicates that DNA has been mixed at some point during the generation of the sequence. Finally, all animals are checked whether they have genotypes in all genomic regions.

RESULTS AND DISCUSSION

The 1000 Bulls has grown fast over time starting with 238 *taurine* animals from 4 breeds in Run2 (2012) (Daetwyler *et al.* 2014) and 1756 animals from 55 breeds in Run5 (2015) across *taurine* and *indicine* sub-species. Run5 identified 67.3 million variants, of which 64.8 million were SNP and 2.5 million were indels. Run6 is currently underway and includes close to 2800 animals across more than 70 breeds (Figure 1). A new feature of Run6 is the inclusion of related species such as the Gaur, Yak, *turano mongolicus*, ancient cattle, and an auroch. It also has much expanded collection of African and *indicine* breeds.

SheepGenomesDB Run1 (2016) discovered 50 million variants and contained 453 sheep and included many New Zealand breeds, the International Sheep Genomics Consortium global diversity set, and Moroccan as well as Iranian sheep from the NextGen project (available at

http://www.ebi.ac.uk/eva/?eva-study=PRJEB14685). Run2, for which the Samtools variant calling at AgVic has concluded, contains 935 sheep with the SheepCRC contributing a large number of animals from the main four Australian breeds (Merino, Polled Dorset, White Suffolk and Border Leicester, Figure 1) and the USDA contributing their Sheep Diversity Panel animals.

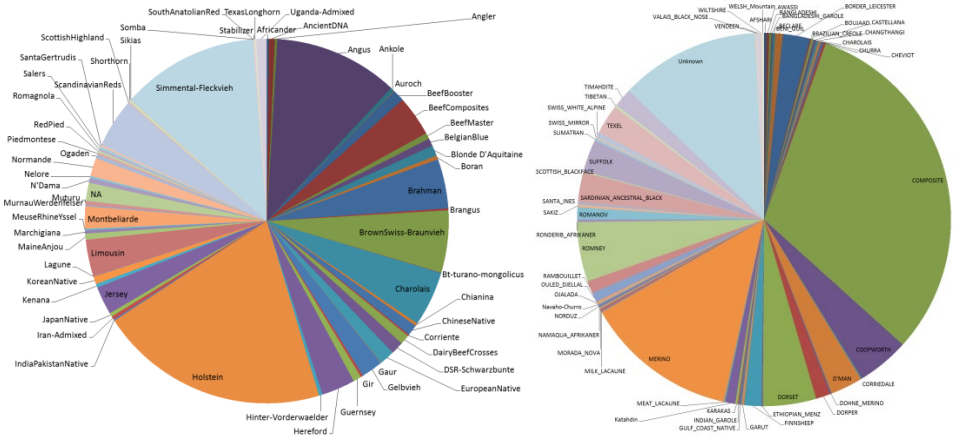

Figure 1. Breeds included in Run6 of the 1000 Bull Genomes Project (left panel) and in Run2 of the SheepGenomesDB Project (right panel). Sheep composites are primarily crosses of Australian and New Zealand breeds

The 1000 Bulls data has been the basis for several studies that detected causative mutations, where its genomes served as controls (e.g. Daetwyler *et al.* 2014; Murgiano *et al.* 2015). It has been the basis for imputation at many consortium partner institutions, which have then used the data to perform imputation, GWAS and genomic selection (e.g. Bouwman & Veerkamp 2014; van den Berg *et al.* 2016). Similar benefits are expected to be realised in sheep. The two consortia are the most complete inventory of cattle and sheep genetic variants globally and will form the basis for sequenced-based animal breeding research in many countries.

ACKNOWLEDGEMENTS

The authors would like to thank all 1000 Bull Genomes Consortium partners and SheepGenomesDB contributors for sharing their data. Funding from the USDA, the DairyBio project (a collaboration between Dairy Australia and the Victorian Government), and the Cooperative Research Centre for Sheep Industry Innovation is acknowledged.

REFERENCES

Bouwman A.C. and Veerkamp R.F. (2014) *BMC Genetics* **15**, 105.
Browning B.L. and Browning S.R. (2009) *Am. J. Hum. Genet.* **84**, 210-23.
Daetwyler H.D., Capitan A., Pausch H., et al (2014) *Nat. Genet.* **46**, 858-65.
DePristo M.A., Banks E., Poplin R., et al. (2011) *Nat. Genet.* **43**, 491-8.
Li H. and Durbin R. (2009) *Bioinformatics* **25**, 1754-60.
Li H., Handsaker B., Wysoker A., et al (2009) *Bioinformatics* **25**, 2078-9.
Murgiano L., Shirokova V., Welle M.M., et al (2015) *PLoS Genet.* **11**, e1005427.
van den Berg I., Boichard D. and Lund M.S. (2016) *Genet. Sel. Evol.* **48**, 83.