RESEARCH ARTICLE

Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity [version 1; referees: 2 approved]

Michael P. Heaton[1], Timothy P.L. Smith[1], Bradley A. Freking[1], Aspen M. Workman[1], Gary L. Bennett[1], Jacky K. Carnahan[1], Theodore S. Kalbfleisch[2]

[1]U.S. Meat Animal Research Center (USMARC), Clay Center, NE, 68933, USA
[2]Department of Biochemistry and Molecular Biology, School of Medicine, University of Louisville, Louisville, KY, 40202, USA

v1 **First published:** 02 Aug 2017, **6**:1303 (doi: 10.12688/f1000research.12216.1)
Latest published: 02 Aug 2017, **6**:1303 (doi: 10.12688/f1000research.12216.1)

Abstract

Background: Access to sheep genome sequences significantly improves the chances of identifying genes that may influence the health, welfare, and productivity of these animals.
Methods: A public, searchable DNA sequence resource for U.S. sheep was created with whole genome sequence (WGS) of 96 rams. The animals shared minimal pedigree relationships and represent nine popular U.S. breeds and a composite line. The genomes are viewable online with the user-friendly Integrated Genome Viewer environment, and may be used to identify and decode gene variants present in U.S. sheep.
Results: The genomes had a combined average read depth of 16, and an average WGS genotype scoring rate and accuracy exceeding 99%. The utility of this resource was illustrated by characterizing three genes with 14 known coding variants affecting litter size in global sheep populations: growth and differentiation factor 9 (*GDF9),* bone morphogenetic protein 15 (*BMP15*), and bone morphogenetic protein receptor 1B (*BMPR1B*). In the 96 U.S. rams, nine missense variants encoding 11 protein variants were identified. However, only one was previously reported to affect litter size (*GDF9* V371M, Finnsheep). Two missense variants in *BMP15* were identified that had not previously been reported: R67Q in Dorset, and L252P in Dorper and White Dorper breeds. Also, two novel missense variants were identified in *BMPR1B*: M64I in Katahdin, and T345N in Romanov and Finnsheep breeds. Based on the strict conservation of amino acid residues across placental mammals, the four variants encoded by *BMP15* and *BMPR1B* are predicted to interfere with their function. However, preliminary analyses of litter sizes in small samples did not reveal a correlation with variants in *BMP15* and *BMPR1B* with daughters of these rams.
Conclusions: Collectively, this report describes a new resource for discovering protein variants *in silico* and identifies alleles for further testing of their effects on litter size in U.S. breeds.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	 2
version 1 published 02 Aug 2017	✓ report	✓ report

1 **Christine Couldrey**, Livestock Improvement Corporation, New Zealand

2 **Eyal Seroussi**, Agricultural Research Organization (ARO), Israel

Discuss this article

Comments (0)

Corresponding authors: Michael P. Heaton (Mike.Heaton@ARS.USDA.GOV), Theodore S. Kalbfleisch (ted.kalbfleisch@louisville.edu)

Author roles: Heaton MP: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Smith TPL**: Conceptualization, Funding Acquisition, Methodology, Resources, Supervision, Writing – Review & Editing; **Freking BA**: Formal Analysis, Investigation, Methodology, Resources, Writing – Review & Editing; **Workman AM**: Formal Analysis, Investigation, Writing – Review & Editing; **Bennett GL**: Formal Analysis, Project Administration, Supervision, Writing – Review & Editing; **Carnahan JK**: Formal Analysis, Investigation, Validation, Writing – Review & Editing; **Kalbfleisch TS**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Heaton MP, Smith TPL, Freking BA *et al.* **Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity [version 1; referees: 2 approved]** *F1000Research* 2017, **6**:1303 (doi: 10.12688/f1000research.12216.1)

Grant information: Funding for this research was provided by the USDA, ARS appropriated projects 5438-32000-029-00D (MPH) and 5438-31320-012-00D (TPLS).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 02 Aug 2017, **6**:1303 (doi: 10.12688/f1000research.12216.1)

Introduction

There are currently 48 Mendelian traits and disorders in sheep where the causative variants are known[1]. Many of these variants affect the gene's protein sequence, and thereby alter its normal function. Although gene function may be affected by a wide range of large and small scale genomic sequence differences[2,3], variants that alter amino acid sequences via missense, nonsense, frameshift, and splice site variants, are among those most likely to affect function[4]. DNA polymorphisms encoding these protein variants are readily identified by aligning genomic sequences of animals to a high-quality, annotated reference genome assembly like that available for sheep[3]. Identifying protein variants encoded by individuals in a population is an essential first step in characterizing genes known to influence traits[5,6].

In principle, protein variants may be identified *in silico* for a gene of interest with access to population-scale whole genome sequence (WGS) data, like that found at the National Center for Biotechnology Information (NCBI) BioProjects and Sequence Read Archives (SRA). The first large ovine BioProject was deposited by the International Sheep Genomics Consortium (ISGC), which included the genome sequences of 75 sheep from 43 breed groups and two wild species from around the world (PRJNA160933). Although global diversity is outstanding in these sheep, these animals are not ideally suited for protein variant discovery across U.S. sheep populations due to their exotic breed composition and low numbers within breed. In addition, the terabyte size of SRA datasets is challenging to work with, and not readily searchable by gene or accessible on the internet with a user-friendly environment, such as the Integrated Genome Viewer (IGV)[7,8].

We previously showed in cattle that protein variants for a gene of interest may be identified *in silico* with the appropriate population sample and 14x WGS datasets[9]. To that end, we created a similar publicly accessible, 16x WGS resource of 96 rams, that is viewable online with IGV. The rams share minimal pedigree relationships, and represent nine popular U.S. breeds and a composite line. Their genomes may be used to identify DNA polymorphisms in genes that affect the protein sequences in U.S. sheep populations. To highlight the utility of this resource, we analyzed three well-studied genes previously shown to encode protein variants affecting litter size in sheep: growth and differentiation factor 9 (*GDF9),* bone morphogenetic protein 15 (*BMP15*), and bone morphogenetic protein receptor 1B (*BMPR1B*). Together, there are 14 previously reported missense, nonsense, and frameshift variants affecting the protein function of these genes, and thereby affect ovulation rate and litter size[10,11].

The proteins encoded by *GDF9* and *BMP15* are oocyte-secreted paralogs of the transforming growth factor-beta (TGF-β) superfamily that form homo- and heterodimeric ligands, and are essential for ovarian and follicular development[12]. These ligands synergistically regulate folliculogenesis through complex interactions with multiple receptors, such as BMPR1B. The *BMPR1B* gene encodes a type 1 membrane protein receptor that binds GDF9 and BMP15 in some mammals, although the identities of the BMPR1B ligands in sheep are unknown[13]. The amino acid sequences of GDF9, BMP15, and BMPR1B are highly conserved among placental mammals, and variants that alter key residues in peptide sequence, diminish function, and affect traits like ovulation rate and litter size. For example, substitution of arginine (R) for glutamine (Q) at position 249 (Q249R) in *BMPR1B* causes attenuation of BMPR1B signaling and ultimately leads to an increase ovulation rate[14,15]. Likewise, missense, nonsense, and frameshift variants in *GDF9* and *BMP15* may abolish function and cause an increase in ovulation rate in carrier ewes, while causing sterility in homozygous ewes[10]. However, some homozygous missense variants only diminish the protein's biological activity. For example, the homozygous substitution of methionine (M) for valine (V) at position 371 (V371M) in *GDF9* allows ewes to remain fertile and hyper prolific. Since the types and distribution of protein variants encoded by these genes was unknown in U.S. sheep, we sought to identify them with WGS from the set of 96 U.S. rams.

We identified nine missense variants and 11 encoded protein variants in the three genes evaluated. Only one variant was previously known to be associated with increased litter size (*GDF9*, V371M). However, four variants were not previously reported. In *BMP15*, a Q for R substitution was observed at position 67 (R67Q), and a proline (P) for leucine (L) substitution was observed at position 252 (L252P). In *BMPR1B*, an isoleucine (I) for M at position 64 (M64I), and an asparagine (N) for threonine (T) was observed at position 345 (T345N). Based on the pattern of evolutionary conservation for these residues in vertebrates, it was hypothesized that some of these novel missense variants could interfere with protein function, affect litter size, and be useful for producers interested in modulating lamb production to match available resources.

Methods

Ethics statement

This article contains no studies performed with animal subjects. The archival DNA samples used were collected between the years 2000 and 2006[16]. The reproduction records used were from daughters born between 2001 and 2007. All animal procedures were reviewed and approved by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS), U.S. Meat Animal Research Center (USMARC) Animal Care and Use Committee prior to their implementation (Experiment Number 5438-31000-037-04). Because health status is important for providing purified DNAs to an international community as described here, tissues were collected from healthy sheep, without signs or history of clinical disease. The source flock's history of disease surveillance is also relevant when requesting reference samples described in this report. Since first stocking sheep in 1966, USMARC has not had a known case of scrapie. Until 2002, surveillance consisted of monitoring sheep for possible signs of scrapie and submitting brain samples to the USDA Animal and Plant Health Inspection Service (APHIS) National Veterinary Services Laboratory in Ames, IA for testing. All tests have been negative. Since April 2002, USMARC has voluntarily participated in the APHIS Scrapie Flock Certification Program, is in compliance with the National Scrapie Eradication Program, and is certified as scrapie-free. With regards to other transmissible diseases, it is recognized that the USMARC flock of 2000 to 4000 breeding ewes is located in a bluetongue medium incidence area and is known to have some prevalence of contagious ecthyma (sore mouth), foot rot, paratuberculosis (Johne's disease), ovine progressive pneumonia (visna-maedi), and pseudotuberculosis caseous lymphadenitis.

Design of the sheep diversity panel

The purpose of the USMARC Sheep Diversity Panel version 2.4 (MSDPv2.4) was to provide a set of 96 samples for variant allele discovery and frequency estimation in U.S. sheep. Details of the panel design strategy have been published elsewhere[16]. Briefly, the panel consists of 96 rams from Dorper, White Dorper, Dorset, Finnsheep, Katahdin, Rambouillet, Romanov, Suffolk, and Texel breeds; a composite line (USMARC III: 1/2 Columbia, 1/4 Hampshire, and 1/4 Suffolk[17]); and one Navajo-Churro ram (Figure 1). In addition to their contributions to the U.S. sheep industry, the breeds were selected to represent genetic diversity for traits such as fertility, prolificacy, maternal ability, growth rate, carcass leanness, wool quality, mature weight, and longevity. The Navajo-Churro ram was included for its rare lysine 171 (K171) substitution in the prion gene. The rams sampled from each breed were chosen to minimize their genetic relationships at the grandparent level. DNA samples of all 96 rams have been made available for global use as genotyping reference material since 2010[16].

WGS production, alignment, and SNP genotyping

DNA was extracted from whole blood with a typical phenol:chloroform method and stored at 4°C in 10 mM TrisCl, 1 mM EDTA (pH 8.0) as previously described[16]. Library preparation for DNA sequencing was also accomplished as previously described[9]. Briefly, 2 µg of ovine genomic DNA was fragmented and used to make indexed, 500 bp, paired-end libraries. Pooled libraries were sequenced with a massively parallel sequencing machine and high-output kits (NextSeq500, two by 150 paired-end reads, Illumina Inc.). Pooled libraries with compatible indexes were repeatedly sequenced until 40 GB of data with greater than Q20 quality was collected for each ram, thereby producing at least 10-fold mapped read coverage for each index. This level of coverage provides scoring rates and accuracies that exceed 99%[9,18]. The DNA sequence alignment process was similar to that previously reported[18]. FASTQ files were aggregated for each animal and DNA sequences, aligned individually to Oar_v3.1 with the Burrows-Wheeler Alignment tool (BWA) aln algorithm version 0.7.12[19], merged, and collated with the bwa sampe command. The resulting sequence alignment map (SAM) files were converted to binary alignment map (BAM) files, and subsequently sorted via SAMtools version 1.3.1[20]. Potential PCR duplicates were marked in the BAM files using the Genome Analysis Toolkit (GATK) version 3.6[21]. Regions in the mapped dataset that would benefit from realignment due to small indels were identified with the GATK module RealignerTargetCreator, and realigned using the module IndelRealigner. The BAM files produced at each of these steps were indexed using SAMtools. The resulting indexed BAM files were made immediately available via the Intrepid Bioinformatics genome browser with groups of animals linked at the USDA, ARS, USMARC internet site.

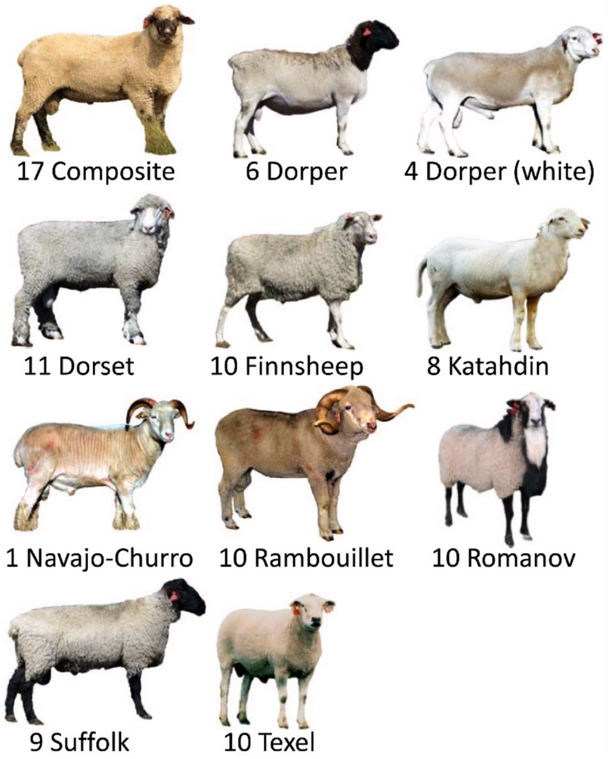


Figure 1. USMARC Sheep Diversity Panel version 2.4. This group of 96 rams was sampled from USMARC and private U.S. flocks to represent genetic diversity for traits such as fertility, prolificacy, maternal ability, growth rate, carcass leanness, wool quality, mature weight, and longevity.

The raw reads were deposited at NCBI BioProject PRJNA324837. Mapped datasets for each animal were individually genotyped with the GATK UnifiedGenotyper with arguments "--alleles" set to the VCF file (Supplementary File S1), "--genotyping_mode" set to "GENOTYPE_GIVEN_ALLELES", and "--output_mode" set to "EMIT_ALL_SITES". Lastly, some SNP variants were identified manually by inspecting the target sequence with IGV software version 2.1.28[7,8] (described below in Methods section entitled 'Identifying protein variants encoded by *GDF9, BMP15, and BMPR1B* genes'). In these cases, read depth, allele count, allele position in the read, and quality score were taken into account when the manual genotype determination was made.

Evaluating WGS data integrity with 163 reference SNPs and 50 k bead array SNPs

Genotypes from a set of 163 reference SNPs were used as an initial verification of the WGS datasets. These DNA markers have been used for parentage determination, animal identification, and disease traceback[22]. The 163 reference SNPs were previously genotyped across the MSDPv2.4 by multiple overlapping PCR-Sanger sequencing reactions, multiplexed matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) genotyping assays, and 50 k bead array platforms[22]. The genotype call rate was defined as the number of SNP sites with three or more mapped reads, divided by the total number of sites tested. The error rate in the WGS data was estimated by comparing the independently-derived consensus genotypes for these SNPs to the WGS genotypes. An animal's WGS dataset passed initial verification when the accuracy of the WGS genotypes exceeded 97%, and the average mapped read depth was proportional to the amount of WGS data collected. Animals' datasets that failed this initial verification were inspected for contaminating and/or missing files. Once identified, the dataset was corrected and reprocessed. Linear regression analysis was accomplished in Excel version 2016. Access to the sequence was made available via USDA, ARS, USMARC internet site. Because the raw datasets were available online as they were produced, the raw FASTQ files were deposited in the NCBI SRA only after they were validated as described above. These 96 sets of files may be accessed through BioProject PRJNA324837 in the Project Data table under the Resource Name: SRA Experiments.

SNPs from the OvineSNP50 BeadChip (Illumina Inc.) were selected for comparison because they were numerous, uniformly distributed across the ovine genome, and available. Based on the nucleotide sequence of the 54,242 probes obtained from the manufacturer, the positions of 51,796 SNPs were verified via a BLAT process, as previously described[18]. There were 50,357 of these that mapped uniquely to autosomes and were used for analysis (Supplementary File S1). The genotypes from the WGS data were compared to those from the 50 k bead array with a custom program written specifically for this operation.

Identifying protein variants encoded by GDF9, BMP15, and BMPR1B genes

The nucleotide variation in the exon regions of *GDF9, BMP15,* and *BMPR1B* was visualized through the public access portal at ARS USMARC with open source software installed on a laptop computer. Variants were recorded manually in a spreadsheet as previously described[9]. Briefly, a Java Runtime Environment version 8, update 131 (Oracle Corporation, Redwood Shores, CA) was first installed on the computer. When links to the data were selected from the appropriate web page, IGV software version 2.1.28[7,8] automatically loaded from a third-party site (University of Louisville, Louisville KY) and the mapped reads were loaded in the context of the ovine Oar_v3.1 reference genome assembly. Gene variants were viewed by loading WGS from a set of eight animals of different breeds, and the IGV browser was directed to the appropriate genome region by entering the gene abbreviation in the search field (e.g., GDF9). The IGV zoom function was used to view the first exon at nucleotide resolution with the "Show translation" option selected in IGV. Since *GDF9* was in the reverse orientation with regards to the Oar_v3.1 assembly, the reference sequence was reversed so the translation was correctly viewed from right to left. The exon sequences were visually scanned for polymorphisms that would alter amino acid sequences, such as missense, nonsense, frameshift, and splice site variants. Once identified, the nucleotide position corresponding to a protein variant was viewed and recorded for all 96 animals. Using IGV, codon tables, and knowledge of the ovine GDF9, BMP15, and BMPR1B protein sequences (NP_001136360.2, NP_001108239.1, and NP_001009431.1, respectively), the codons affected by nucleotide alleles were translated into their corresponding amino acids and their Oar_v3.1 positions noted. Haplotype-phased protein variants were unambiguously assigned in individuals that were either: 1) homozygous for all variant sites, or 2) had exactly one heterozygous variant site. Maximum parsimony phylogenetic trees were manually constructed from the unambiguously phased protein variants. The phylogenetic trees were used, together with simple maximum parsimony assumptions, to infer haplotype phase in seven rams where two heterozygous variant sites occurred in *GDF9*. The protein phylogenetic trees were rooted by comparing the variable residues in sheep to those from related species. Ovine peptide sequences for GDF9, BMP15, and BMPR1B were used to search NCBI's refseq_protein database with BLASTP 2.6.1[23,24]. Aligned protein sequences from a representative subset of 29 vertebrate species were used for the comparison.

Statistical analysis of litter size in daughters of carrier rams

Lambing records for daughters of carrier rams were retrieved from the USMARC historical database and analyzed with the mixed-model analysis of variance procedure (MIXED) of SAS (SAS Inst., Inc., Cary, NC; version 9.3). The phenotype evaluated was total number of lambs born (including stillborn) as a repeated record for each ewe. Different sets of ewes contributed to the analysis of each gene locus, and breed-specific genotype contrasts were evaluated. There were, however, similar models employed for all of the analyses. The models included fixed effects of classification for ewe age, and the sire-derived genotype class for the allele contrast in question. Three groups were created for ewe age to combine similar biologically performing ages: Group 1, ewe lambs; Group 2, ewes aged 2–5 years; and Group 3, ewes older than 5 years. The random effect of "ewe" was fitted and used to test the genotype contrast mean square. The Kenward-Roger option was used to approximate denominator degrees of freedom associated with the random effect of "ewe". For analysis of the X-linked *BMP15* allele contrasts, the sire-derived gamete in these daughters was known directly. For analyses of autosomal genotype contrasts, it was inferred that rams of different genotypes had different distributions of daughter genotypes sampled. This inference reduced the power of analysis compared to a direct allelic test because we cannot determine the maternal-derived allele.

Results

Genome sequencing and validation of WGS datasets

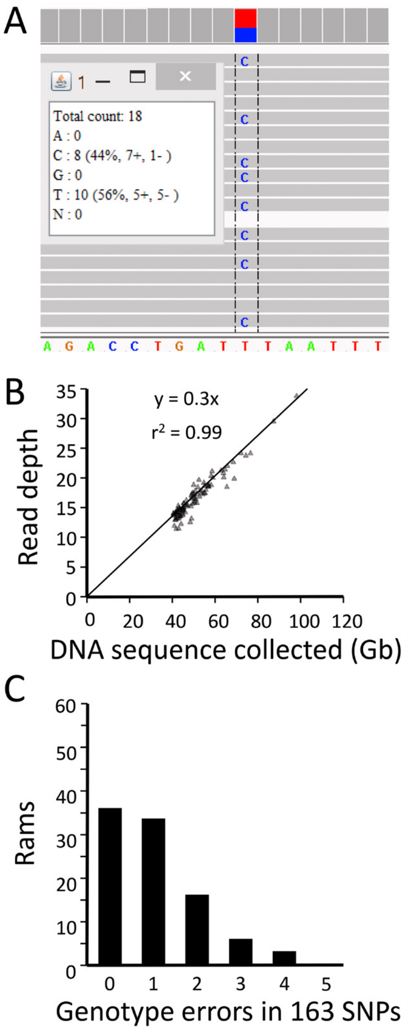
The average amount of genomic DNA sequence collected per animal was 50.4 GB (range 40.0 - 97.7, SD 10.4). Independently-derived genotypes from two sets of reference SNPs were used to confirm the identity and evaluate the quality of these data: those from 163 parentage markers, and those from approximately 50,000 markers on the OvineSNP50 bead array. Both sets have SNPs that are well distributed, highly-informative, and have been widely used. The WGS-derived genotypes for the 163 parentage SNPs were obtained by manually viewing an animal's mapped reads at the relevant genome coordinates via the internet and third-party software (illustrated in Figure 2A, and described in Methods). The expected genotypes and read depths were consistent for all but one of the 96 datasets, owing to missing data for that animal. After rectifying the data omission and performing regression analysis of the data for all 96 rams, the average calculated read depth (17.0) was directly proportional to the amount of sequence collected for each animal (range 11.9 - 33.9, SD 3.6; Figure 2B).

The genotype call rate for the 163 parentage markers was 99.7% when WGS data was used, i.e. 47 missing of 15,159 possible. Most of the missing genotypes (32) were attributed to a single SNP site (DU191809, chr1:187087905). The source of the difficulty appeared to be a misassembly of the Oar_v3.1 in that that region, leading to a mismapping of reads as this site averaged only 3.5 reads per animal. The overall accuracy of WGS genotypes for the 163 reference SNPs was 99.4%, and no animals had a SNP genotype accuracy less than 97% (i.e., not more than 4 errors in 163 SNP genotypes; Figure 2C). The few WGS genotype errors observed were typically caused by undetected heterozygous alleles at sites with low read coverage. Thus, comparing genotypes from 163 reference SNPs to those derived from the WGS file sets was effective for discovering and repairing errors, and independently verifying coverage.

The coverage and integrity of the WGS datasets were also evaluated at 50,357 evenly distributed, autosomal SNP sites from bead array data[25]. When plotted as a distribution of read depths by SNPs for all animals combined, the read depth was normally distributed with a mode near 16 (Figure 3A). The calculated average read depth per SNP per animal was 16.8 for the 50 k bead array SNPs (Min 11.7, Max 34.2, SD 3.5), compared to 17.0 for the 163 reference SNPs above. Averaged over all animals, the concordance between WGS genotypes and those from the bead array was 99.5% (Figure 3B) compared to 99.4% for the 163 reference SNPs. The genotype concordance reached a maximum at approximately 99.89% for the animal with the highest read depth (34.2-fold, 97.7 GB Q20 data). Taken together, the WGS genotype results for 163 reference SNPs was consistent with those for the 50 k bead array SNPs and indicated that the WGS datasets from these 96 rams are of sufficient quality and coverage for use in identifying and decoding gene variants in U.S. sheep.

Protein variants encoded by GDF9, BMP15, and BMPR1B genes

The WGS data for the 96 rams were used to analyze the coding regions of *GDF9, BMP15,* and *BMPR1B*. These genes encode proteins of 453, 393, and 502 amino acids, respectively, each with multiple functional domains (Figure 4A). Viewing the aligned



Total count: 18
A : 0
C : 8 (44%, 7+, 1-)
G : 0
T : 10 (56%, 5+, 5-)
N : 0

Figure 2. Comparison of 163 reference SNP genotypes with those derived from WGS data. (**A**) Computer screen image of one animal's WGS data aligned to ovine reference assembly Oar_v3.1 at a reference SNP site. The heterozygous C/T genotype is shown as viewed with the IGV software[7,8]. (**B**) Linear relationship between mapped read depth and the amount (Gb) of Q20 WGS data collected. At each SNP position, the read depth and genotypes were visualized and manually recorded for 163 parentage SNPs. (**C**), genotype scoring accuracy for 163 parentage SNPs in 96 sires. Consensus reference genotypes (n = 15,684) for the parentage SNPs were previously determined by multiple methods[22].

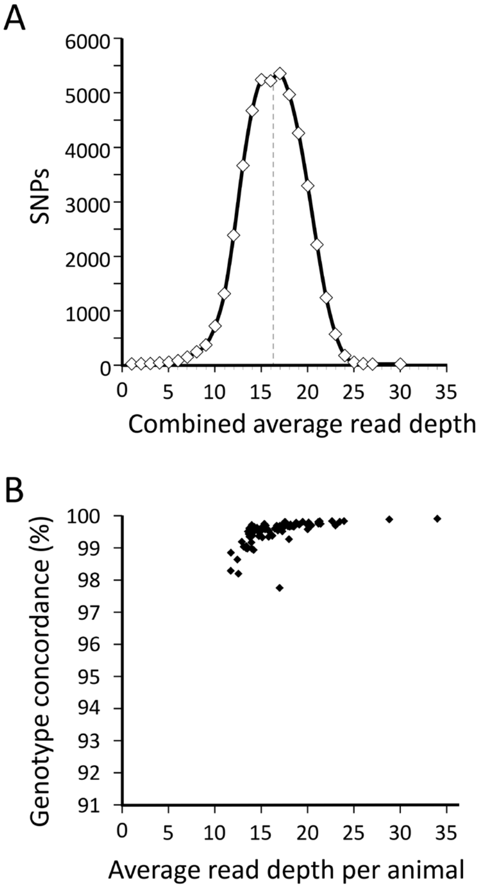


Figure 3. Comparison of WGS genotypes from 96 rams with those from bead arrays. (**A**) The distribution of average WGS read depth across 45,946 SNP sites for 96 sires combined. (**B**) A comparison of the average WGS read depth per animal to the average genotype concordance between 45,946 WGS and bead array genotypes.

sequences and detecting variants was simple, fast, and accurate with the IGV software and a publicly available web-based browser developed for this purpose (Figure S1, Table S1). Nine missense variants were observed in the three genes with the 96 genomes (Table 1). Four of the nine variants were not previously reported: *BMP15* (R67Q, L252P) and *BMPR1B* (M64I, T345N). No other missense, nonsense, frameshift, splice site, or indel variants affecting the coding region were detected. A comparative list of the coding variants discovered here is given in Table 2, together with those previously reported for the three genes. Eleven protein sequence isoforms were predicted from phased combinations of codon variants (Table 3). Haplotypes were translated and placed in the context of a phylogenetic tree for predicted variants for *GDF9, BMP15,* and *BMPR1B* (Figure 4B). The trees were rooted based on the pattern of evolutionary conservation of the residues in vertebrates (Figure 5). All four of the previously unreported protein variants encoded by *BMP15* and *BMPR1B* were on the distal nodes of their respective tree, indicating they arose after those on adjacent nodes.

The previously reported *GDF9* V371M variant was present in our reference panel only in Finnsheep (Table 4). Alleles encoding the M371 residue are associated with increased litter size in both carriers and homozygous individuals (Table 2). The novel *BMP15* R67Q and L252P variants were confined to the Dorset and Dorper breed groups of our reference panel, respectively. The novel *BMPR1B* M64I variant was only present in the Katahdin breed group, while the novel *BMPR1B* N345 variant was observed in both Romanov and Finnsheep breed groups.

An analysis of amino acid sequence conservation among species helped identify critical residues more likely to be involved in important protein functions. Ovine GDF9, BMP15 and BMPR1B were 80, 88, and 99% identical to other Artiodactyla species, at the propeptide sequence level (Figure 5). We predict that variant residues in highly conserved protein domains are more likely to affect ovulation rate and litter size. The most well conserved residue in GDF9 showing missense variation among sheep breeds is V371,

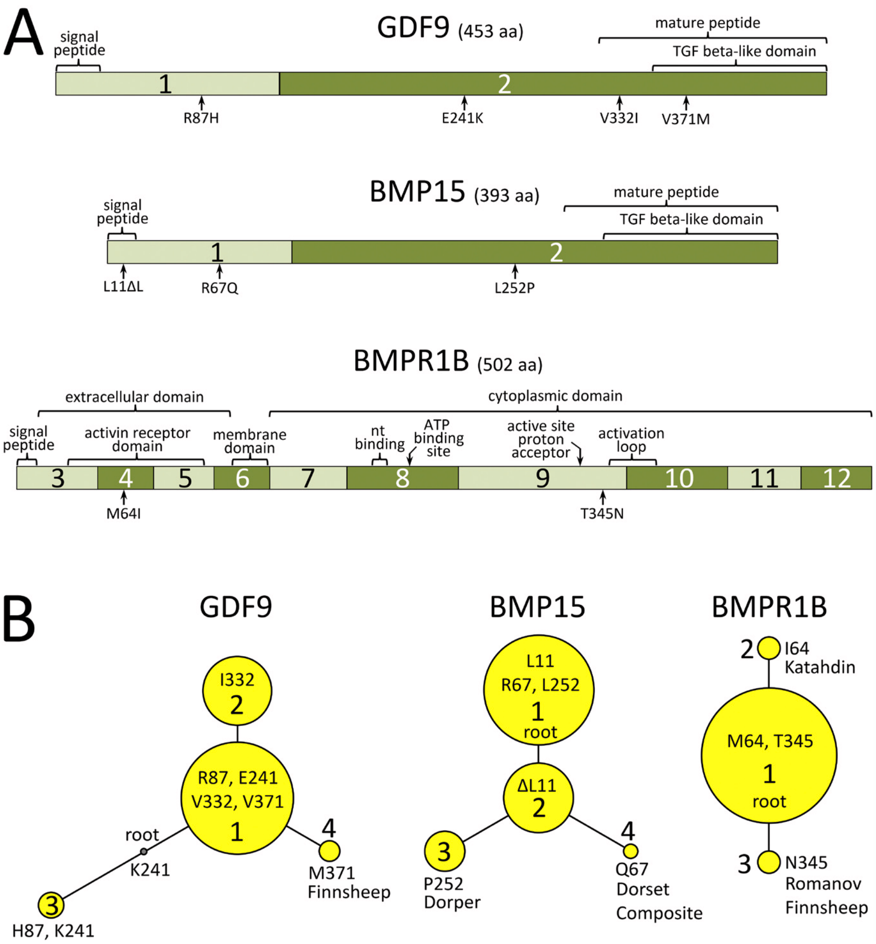


Figure 4. Physical maps and rooted maximum parsimony phylogenetic trees of GDF9, BMP15, and BMPR1B protein variants in U.S. sheep. (**A**) Physical maps of GDF9, BMP15, and BMPR1B exon and protein domains in relationship to missense variants. (**B**) Maximum parsimony phylogenetic trees of haplotype-phased protein variant identified in the sheep diversity panel. For each gene analyzed, the most frequent protein isoform was defined as “variant 1” and used as the reference sequence for each tree. Each node in a tree represents a different protein isoform that varies by one amino acid compared to adjacent nodes. The areas of the circles are proportional to the variant frequency in the panel of 96 rams. The trees were rooted based on evolutionary conservation of residues in closely related species. The predicted root of GDF9 was not observed in the 96 rams.

Table 1. DNA sequence information for *GDF9, BMP15, and BMP15R1B* missense variants identified in the sheep diversity panel (MSDPv2.4).

Gene	Codon variant[a]	Position (Oar_v3.1)	Exon	Pro. dom.[b]	Codon[c]	Codon allele[d]	MAF[e]	Flanking genomic sequence
GDF9	R87H	chr5:41,843,258	1	PRO	c**R**c	c**G**c = R c**A**c = H	0.037	ggtgggagacacagacctggtctcctttcccctctcttagaggttctgtatgatgggcacggggaaccccccaggctgcagccagatgacagagctttgc **[R]** ctacatgaagaggctctataaggcatacgctaccaaggaggggaccccctaaatccaacagacgccacctctacaacactgttcggctcttcacccccctgt
	E241K	chr5:41,841,675	2	PRO	**R**aa	**G**aa = E **A**aa = K	0.037	gaaatacaaatggatggagattgatgtgacggctcctcttgagcctctggtggcctcccacaagaggaatattcacatgtctgtaaattttacatgtgcg **[R]** aagaccagctgcagcatccttcagcgcgggacagcctgtttaacatgactcttctcgtagcgccctcactgcttttgtatctgaacgacacaagtgctca
	V332I	chr5:41,841,402	2	MAT	**R**tt	**G**tt = V **A**tt = I	0.245	atctgcctaccccgtgggagaagaagctgctgagggtgtaagatcgtcccgtcaccgcagagaccaggagagtgccagctctgaattgaagaagcctctg **[R]** ttccagcttcagtcaatctgagtgaatacttcaaacagtttctttttccccagaatgaatgtgagctccatgactttagacttagctttagtcagctgaa
	V371M	chr5:41,841,285	2	MAT	**R**tg	**G**tg = V **A**tg = M	0.026	tctgagtgaatacttcaaacagtttctttttccccagaatgaatgtgagctccatgactttagacttagctttagtcagctgaagtgggacaactggatt **[R]** tggccccacacaaatacaaccctcgatactgtaaaggggactgtcccagggcggtcggacatcggtatggctctccggttcacaccatggtgcagaacat
BMP15	L11ΔL[f]	chrX:50,977,397	1	PRO	indel	cttctt = LL ctt = L	0.344	gtaaaaggaaaggtttaaagcgttatcctttgggcttttatcagaacatgttgctgaacaccaagcttttcaagatggtcctcctgagcatccttagaatc **[cttctt/ctt]** tggggactggtgctttttatggaacatagggtccaaatgacacaggtagggcagccctctattgcccacctgcctgaggcccctaccttgcccctgattc
	R67Q	chrX:50,977,228	1	PRO	c**R**g	c**G**g = R c**A**g = Q	0.010	cacctgcctgaggcccctaccttgcccctgattcaggagctgctagaagaagcccctggcaagcagcagaggaagccgcgggtcttagggcatcccttac **[R]** gtatatgctggagctgtaccagcgttcagctgacgcaagtggacaccctagggaaaaccgcaccattggggccaccatggtgaggctggtgaggccgctg
	L252P	chrX:50,971,365	1	MAT	c**Y**g	c**T**g = L c**C**g = P	0.083	ttctggtggcatggcacttcatcattggacactgtcttcttgttactgtatttcaatgacactcagagtgttcagaagaccaaacctctccctaaaggcc **[Y]** gaaagagtttacagaaaaagacccttctcttctcttgaggagggctcgtcaagcaggcagtattgcatcggaagttcctggcccctccagggagcatgat
BMPR1B	M64I	chr6:29,401,381	4	AR	at**R**	at**G** = M at**A** = I	0.026	acacacacacacacacacacacacacacacatactttgcctgtttgatctttagcacagatggatattgtttcacgatgatagaagaagatgactctgggat **[R]** cctgtggtcacttctggatgtctaggactagaaggctcagattttcagtgtcgggtaaggaagataccttggttccactttgtaaccttttattggcaag
	T345N	chr6:29,380,795	9	AS-AL	a**M**t	a**A**t = T a**C**t = N	0.026	cctgtggtcacttctggatgtctaggactagaaggctcagattttcagtgtcgggtaaggaagataccttggttccactttgtaaccttttattggcaag **[M]** ctctgtcacttacacactgaaatctttagcactcaaggcaaaccagcaattgcccatcgagatctgaaaagtaagaacatcctggtgaagaaaaatggaa

[a] All variants and sequences are oriented from the sense strand perspective. However, GDF9, BMP15, and BMPR1B are oriented in the opposite direction with regards to the Oar_v3.1 reference assembly. Alphabetical abbreviations for relevant amino acids: E, glutamate; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; T, threonine; and V, valine.

[b] Protein domain abbreviations: PRO, propeptide; MAT, mature peptide; AR, activin receptor domain; and AS-AL, between the active site proton acceptor and the activation loop domains.

[c] IUPAC/IUBMB ambiguity codes used for nucleotides: R = a/g, Y = c/t, M = a/c, K = g/t, S = c/g, W = a/t[35].

[d] The major allele is listed first.

[e] Minor allele frequency in MSDPv2.4.

[f] The L11ΔL variant is an abbreviation for p.(Leu10_11delinsLeu), the recommended nomenclature for this variant by the Human Genome Variation Society[36].

Table 2. Comparison of missense, nonsense, and frameshift variants in the coding sequences of ovine *GDF9, BMP15, BMPR1B* and their phenotypic association in sheep.

Gene (Chr.)	Coding variant[a]	Phenotype	Breed groups	Ref.
GDF9 (Chr5)	R87H	None reported	Multiple	37, this work
	E241K	None reported	Multiple	37,38, this work
	R315C	Fecundity, Vacaria, *FecG*^{V}	Ile de France	39
	V332I	None reported	Multiple	37,38, this work
	F345C	Fecundity, Embrapa, *FecG*^{E}	Santa Inês	40
	V371M	Fecundity	Finnish landrace	11,26,37, this work
	S395F	Fecundity, High Fertility, *FecG*^{H}	Belclare, Cambridge	37
	S427R	Fecundity, Thoka, *FecT*^{T}	Icelandic	41
BMP15 (ChrX)	L11ΔL	None reported	Multiple	37, this work
	R67Q	**Unknown, Dorset**	**Dorset**	**This work**
	W154Δ17	Fecundity, Rasa Aragonesa, *FecX*^{R}	Rasa Aragonesa	42,43
	Q239stop	Fecundity, Galway, *FecX*^{G}	Belclare, Cambridge	37
	L252P	**Unknown, Dorper**	**Dorper**	**This work**
	Q291stop	Fecundity, Hanna, *FecX*^{H}	Romney	44
	V299D	Fecundity, Inverdale, *FecX*^{I}	Romney	44
	T317I	Fecundity, Grivette, *FecX*^{GR}	Grivette	45
	C321Y	Fecundity, Lacaune, *FecX*^{L}	Lacaune	46
	N337H	Fecundity, Olkuska, *FecX*^{O}	Olkuska	45
	S367I	Fecundity, Belclare, *FecX*^{B}	Belclare	37
BMPR1B (Chr6)	**M64I**	**Unknown, Katahdin**	**Katahdin**	**This work**
	Q249R	Fecundity Booroola, *FecB*^{B}	Booroola, *et al.*	14,29,33,47
	T345N	**Unknown, Romanov**	**Romanov**	**This work**

[a] Bold font indicates previously unreported variants affecting the protein sequence.

Table 3. Frequencies of haplotype-phased GDF9, BMP15, and BMPR1B protein variants in U.S. sheep.

Protein	Protein variant code	Variant amino acids[a]	Sheep diversity panel (n = 96)[b]
GDF9	1	R87, E241, V332, V371	0.693
	2	R87, E241, **I332**, V371	0.245
	3	**H87**, **K241**, V332, V371	0.036
	4	R87, E241, V332, **M371**	0.026
BMP15	1	L11, R67, L252	0.656
	2	**ΔL**, R67, L252	0.250
	3	**ΔL**, R67, **P252**	0.083
	4	**ΔL**, **Q67**, L252	0.010
BMPR1B	1	M64, T345	0.948
	2	**I64**, T345	0.026
	3	M64, **N345**	0.026

[a] The bolded residues are those differing from “variant 1” in each gene.

[b] The protein variant frequency.

			GDF9					BMP15				BMPR1B		
	Species common name	TMRCA (Ma)	Overall identity (%)	Residue				Overall identity (%)	Residue			Overall identity (%)	Residue	
Taxonomic group				87	241	332	371		11	67	252		64	345
	Sheep variant 1	0	100	**R**	**E**	**V**	**V**	100	**L**	**R**	**L**	100	**M**	**T**
	Sheep variant 2	0	99	.	.	I	.	99	Δ	.	.	99	I	.
Caprinae	Sheep variant 3	0	99	H	K	.	.	99	Δ	.	P	99	.	N
	Sheep variant 4	0	99	.	.	.	M	99	Δ	Q	.	ni		
	Goat	10	99	.	K	.	.	98	.	.	.	99	.	.
	Tibetan antelope	18	98	.	K	.	.	99	.	.	.	100	.	.
Bovidae	Cattle	25	96	S	K	.	.	98	.	.	.	99	.	.
	Yak	25	95	N	K	.	.	98	.	.	.	99	.	.
	Water buffalo	25	95	S	K	.	.	98	.	.	.	99	.	.
Pecora	Moose	27	nr	.	K	.	.	nr	.	.	.	nr	.	.
	Whale	56	86	H	K	.	.	91	.	.	.	99	.	.
Artiodactyla	Dolphin	56	86	H	K	.	.	91	.	.	.	99	.	.
	Swine	62	80	.	K	.	.	88	.	.	.	99	.	.
	Camel	64	80	.	T	.	.	86	.	.	.	99	.	.
	Horse	78	83	.	K	.	.	85	.	.	.	98	.	.
Laurasiatheria	Cat	78	76	.	K	.	.	82	.	.	.	99	.	.
	Dog	78	77	.	R	A	.	82	.	.	.	99	.	.
	Walrus	78	78	.	R	.	.	84	.	Q	.	99	.	.
	Pangolin	78	76	.	K	I	.	81	.	Q	.	99	.	.
	Bat	79	78	.	K	.	.	82	.	Q	.	99	.	.
	Gorilla	96	80	H	K	G	.	75	F	.	M	99	.	.
Boreoeutheria	Human	96	79	.	K	G	.	74	F	.	M	98	L	.
	Mouse	96	67	Y	K	L	.	68	.	.	Q	99	.	.
	Ground squirrel	96	74	.	K	.	.	68	.	Q	.	99	.	.
Eutheria	Elephant	105	76	.	K	.	.	74	.	.	.	99	.	.
	Armadillo	105	76	.	N	.	.	77	.	.	–	99	T	.
Theria	Opossum	159	nm	–	–	–	–	nm	–	–	–	92	.	A
Amniota	Chicken	312	60	.	M	A	I	40	.	.	–	92	G	.
	Alligator	312	59	.	K	A	I	45	.	.	S	92	G	.
Tetrapoda	Frog	352	47	K	K	V	L	43	.	.	.	91	L	.
Sarcopteryii	Coelacanth	413	44	–	T	A	.	39	–	–	R	89	V	.
Euteleostomi	Carp	435	41	.	E	V	I	nm	–	–	–	78	V	A
Gnathostomata	Shark	473	27	Q	K	–	I	nm	–	–	–	87	G	S
Vertabrata	Lamprey	615	nm	–	–	–	–	nm	–	–	–	nm	–	–

Figure 5. Evolutionary comparison GDF9, BMP15, and BMPR1B protein residues at their variant sites in U.S. sheep. Aligned protein sequences from a representative subset of 29 vertebrate species were compared. Abbreviations and symbols are as follows: TMRCA, estimated time to most recent common ancestor in millions of years[48]; letters, IUPAC/IUBMB codes for amino acids; dot, amino acid residues identical to those in sheep "variant 1"; triangle, net deletion of one leucine residue in BMP15 positions 10 and 11 where two leucine residues are commonly present; ni, a fourth protein variant was not identified for BMPR1B; nr, not in refseq_protein database and thus residues were determined by analyzing WGS data; dash, not enough sequence similarity for comparison or missing polypeptide region; nm, did not match a refseq_protein in the database for that species.

located in the TGF-β-like domain. Throughout Eutheria, sheep were the only species observed to have the M371 residue encoded by *GDF9* (Figure 5). This is consistent with the substitution of the M371 residue causing reduced protein function, and therefore increased litter size in Finnish Landrace sheep (Table 2). Less conserved were the *GDF9* residues at positions 87, 241, and 332, which are variable throughout Eutheria species and have not been associated with fecundity in sheep. With regards to missense variants in the other TGF-β ligand, BMP15 residues at positions 11, 67, and 252, were conserved through most of the Laurasiatheria, although the L11 deletion variant is common in sheep and has not been associated with fecundity (Table 2). Since Q67 and P252 substitutions in *BMP15* have not been previously reported, their impact on protein function or reproductive phenotype has yet to be determined.

Conservation in the TGF-β receptor ligand receptor, BMPR1B, is particularly striking with 98% propeptide identity observed throughout Eutheria species, compared to approximately 76% and 77% for GDF9 and BMP15, respectively. Moreover, BMPR1B residues at positions 64 and 345 are also conserved throughout the Eutheria, suggesting that the I64 and the N345 substitutions in sheep may affect protein function. The I64 substitution in Katahdin sheep is in the extracellular activin receptor domain, whereas the N345 substitution in Romanov sheep is between the active site proton acceptor domain and the activation loop of the cytoplasmic domain (Figure 4A). Although intriguing, the potential effects of the observed substitutions encoded by *GDF9*, *BMP15* and *BMPR1B* in U.S. Sheep are unknown.

Retrospective analysis of litter size in daughters of carrier rams

The potential effects of the observed *GDF9*, *BMP15* and *BMPR1B* variants on reproductive phenotypes were examined by analyzing lambing records from daughters of the rams sequenced in this

Table 4. **Frequency estimates of GDF9, BMP15, and BMPR1B protein variants by breed group.**

Protein variant frequency[a]												
		GDF9				BMP15				BMPR1B		
		1	2	3	4[b]	1	2	3[c]	4[c]	1	2[d]	3[d]
Breed group	No.	(Ref.)	(I332)	(H87, K241)	(M371)	(Ref.)	(ΔL)	(P252)	(Q67)	(Ref.)	(I64)	(N345)
Dorper	6	0.33	0.67	-[e]	-	0.17	0.50	0.33	-	1.00	-	-
Dorper, white	4	0.75	0.25	-	-	-	-	1.00	-	1.00	-	-
Dorset	11	0.64	0.36	-	-	0.82	0.09	-	0.09	1.00	-	-
Finn	10	0.55	0.20	-	0.25	0.70	0.30	-	-	0.95	-	0.05
Katahdin	8	0.75	0.25	-	-	1.00	-	-	-	0.81	0.19	-
Navajo-Churro	1	1.00	-	-	-	1.00	-	-	-	1.00	-	-
Rambouillet	10	0.85	0.15	-	-	0.20	0.80	-	-	1.00	-	-
Romanov	10	0.30	0.60	0.10	-	0.30	0.70	-	-	0.80	-	0.20
Suffolk	9	0.94	0.06	-	-	1.00	-	-	-	1.00	-	-
Texel	10	0.70	0.10	0.20	-	0.90	0.10	-	-	1.00	-	-
Composite	17	0.88	0.09	0.03	-	0.82	0.06	0.12	-	0.94	0.06	-

[a] The variants correspond to those shown in Figure 4. The distinctive missense variant or reference isoform is indicated in parentheses.

[b] GDF9 protein "variant 4" contains the M371 amino acid previously associated with litter size in Finnish landrace sheep[11,26,37].

[c] BMP15 protein "variants 3 and 4" contain the previously unreported P252, and Q67 residues, respectively.

[d] BMPR1B protein "variants 2 and 3" contain the previously unreported I64 and N345 missense variants, respectively.

[e] Hyphen indicates the variant was not detected in that group.

project. There were no database records for daughters of the five Finnsheep rams carrying the *GDF9* allele encoding the M371 residue (i.e., "Variant 4", Table S2). There were, however, records for 403 daughters sired by eight rams with at least one of the four *BMP15* or *BMPR1B* variants. Together, the eight rams sired 480 lambs in various flocks in seven years, although not all variant genotypes were frequent in these rams (Table S3–Table S6). Analyses of these data did not reveal a significant correlation between litter size and any of the four *BMP15* or *BMPR1B* variants (95% confidence interval). However, this simple test for association lacked power, and could only detect litter size effects. It remains possible that a well-designed, prospective genetic study may detect biologically and economically relevant differences associated with these variants of highly-conserved residues in developmentally important genes.

Discussion

We created a searchable and publicly viewable online genomics resource consisting of 96 individuals representing a broad cross section of U.S. sheep breeds, and demonstrated its use for identifying protein variants. The DNA for these 96 rams, together with their 95 tetrad families, is also available for confirming segregation alleles identified in the WGS[16]. A minimum of 40 GB of short read, paired-end DNA sequence data provided at least 11-fold mapped genome coverage for each animal. The aligned sequences were made available for downloading or viewing online with a customized IGV visualization software that supports accurate manual assessment of gene-specific genetic variation. The average coverage of the sheep diversity panel was 16.8-fold and resulted in an average genotype accuracy of approximately 99.5%. These numbers were consistent with previous results obtained with 96 beef bulls[9]. This online resource provides the ability to readily inspect gene variants reported in one breed, evaluate them in other breeds, and search for any additional variants that may affect protein structure. The ability to identify the full range of protein variants in a population is critical for designing studies intended to test a candidate gene's influence on a trait.

The web-based platform worked well for analyzing three ovine genes with previously documented missense variants affecting ovulation rate and litter size. In a matter of hours, each gene was evaluated for any obvious coding variants, scored in the group of 96 rams, and compared to the previously known variants. Of the 14 known causative variants affecting litter size in sheep, only one was observed in the 96 U.S. rams, and only in the Finnsheep breed (*GDF9* V371M). This is consistent with reports that the highly prolific Finnish Landrace sheep are thought to be the source of the V371M variant[11,26]. With regards to U.S. Finnsheep, the frequency of the *GDF9* V371M variant was 0.25%, with five of the 10 rams having zero copies of the V371M variant. Since ewes homozygous for the M317 variant are known to be fertile, there is a good opportunity for breeders to modulate the frequency of the *GDF9* V371M variant in their purebred Finnsheep flocks, and thereby attain a more optimal litter size for their ewes.

The WGS analysis also revealed four previously unreported missense variants: *BMP15* R67Q and L252P; *BMPR1B* M64I and T345N. Although our preliminary tests for association between variants and litter size did not detect a significant difference, the evidence for dismissing these candidates is not compelling due to the limited number of sires with the variant allele. In spite of having no direct evidence of phenotypic effects associated with these alleles, analysis of the evolutionary conservation of residues at variant sites, their locations within the protein domains, and the effects on ovulation in other species has provided some insight. For example, the *BMP15* R67Q variant found in Dorset was the least conserved, and predicted to be the least likely to affect function among placental mammals. Since the Q67 residue is present in several other Eutheria, and is not part of the mature BMP15 peptide ligand, its occurrence would seem to be a functional evolutionary option (Figure 5). In humans, the equivalent variant (R68Q) was reported in the 1000 Genomes Project with no apparent disease effect noted (rs782187019)[27]. However, a tryptophan (W) substitution at this same position in humans causes premature ovarian failure and primary ovarian insufficiency (i.e., R68W)[28]. Thus, some substitutions at this position may cause loss of function in some mammals, but it appears as though Q67 may not be one of them.

Unlike the R67Q variant, the L252P variant encoded by *BMP15* was not observed in any other vertebrate species and was strictly conserved throughout the Laurasiatheria species. The P252 residue does not appear in the mature BMP15 peptide, however, it is plausible that the non-conservative substitution of P252 for L252 could interfere with post-translational processing of the mature peptide. In primate species, M253 is the equivalent residue to ovine position L252P, and healthy human individuals represented in the 1000 Genomes Project have rare heterozygous substitutions of V253 and T253 with no pathology reported. Because alleles with the P252 residue were present at a high frequency (1.0 in four White Doper), it's unlikely that the homozygous state causes sterility in ewes. However, the possibility remains that P252 residue may decrease function, and that two copies of a slightly less functional BMP15 may increase the ovulation rate and litter size.

In contrast to the numerous missense variants encoded by the ovine *GDF9* and *BMP15* genes, there has been only one missense variant identified in the receptor gene, *BMPR1B* (Q249R). This variant was first discovered in Booroola Merino sheep[14,29], and subsequently reported in Garole[30], Javanese[30], Chhotanagpuri[31], Iranian Kalehkoohi[32], small-tailed Han[33], Hu and Chinese Merino[34] sheep. In the present report, we did not observe the Q249R variant in any of the WGS from 96 U.S. sheep. Rather, two previously unrecognized *BMPR1B* variants were identified: M64I and T345N. The M64I variant was present in two of eight Katahdin rams (including a homozygote), and two of 17 composite rams containing Suffolk, Colombia and Hampshire germplasm. The I64 substitution was not present in other vertebrate protein sequences and was conserved throughout the Theria with the notable exception of humans, manatees, and armadillos. No variants have been reported in the 1000 Genomes Project for the equivalent position in humans. The M64I variant is positioned in the extracellular activin receptor domain, whose function is to bind ligands for receptor activation. It is plausible that the enhanced fertility and prolificacy, which the Katahdin breed is known for, is conferred in part by this variant.

The second *BMPR1B* variant, T345N, is located inside the cell between two closely spaced active site domains and was present in three of ten Romanov rams (including a homozygote), and one of ten Finnsheep rams. The T345 residue is conserved throughout Tetrapoda species and N345 was not found in any Vertebrata species. A search for human variants in the 1000 Genomes Project revealed only a rare S345 substitution with no pathology reported. Based on the location of the T345 variant near the active site, its strict evolutionary conservation in vertebrates, and that it was found in the two most prolific U.S. breeds, we hypothesize that the N345 residue diminishes the function of the BMPR1B receptor and may influence ovulation and litter size. The *BMPR1B* T345N variant thus represents a high-priority candidate allele for validation studies in these breeds. If any of these newly discovered variants are confirmed to be associated with litter size, DNA-based tests for them could be incorporated into existing genetic testing platforms and used to select for important traits and manage production. Since the number of lambs produced per ewe per year is of fundamental economic importance to sheep production regardless of the production system, these types of DNA tests would be helpful for producers interested in modulating lamb production to match available resources and maintain long-term sustainability.

Conclusion

In summary, the WGS resources described here are suitable for use in identifying and decoding gene variants in the vast majority of U.S. sheep. When applied to *GDF9, BMP15* and *BMPR1B* genes, the findings suggest there may be variants circulating in the U.S. that could be further evaluated for potential use to increase litter size in U.S. breeds. These resources, including the web interface, underlying sequence data, and the associated information are available to researchers, companies, veterinarians, and producers for use without restriction.

Data availability

Validated sheep FASTQ files are available in the NCBI SRA under accession numbers SRX2185832-SRX2185868; SRX2185872-SRX2185977; SRX2186010-SRX2186189; SRX2186191-SRX2186294; SRX2186381-SRX2186766; SRX2186768-SRX2186784; SRX2186786-SRX2186798; SRX2186800-SRX2186879.

The data have also been deposited with links to BioProject accession number PRJNA324837 in the NCBI Bio-Project database.

In addition, access to the aligned sequences is available via USDA internet site: http://www.ars.usda.gov/Services/Docs.htm?docid=25585. Download access to the BAM files is available at the Intrepid Bioinformatics site:

http://server1.intrepidbio.com/FeatureBrowser/customlist/record?listid=7918711123

Lambing records for daughters of carrier rams were retrieved from the USMARC historical database, which is not accessible to the public. Table S3–Table S6 provide summary data from these records, which is adequate for the reproducibility and re-analysis purposes of this article.

Competing interests

No competing interests were disclosed.

Grant information

Funding for this research was provided by the USDA, ARS appropriated projects 5438-32000-029-00D (MPH) and 5438-31320-012-00D (TPLS).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank the USMARC Core Facility staff for outstanding technical assistance, and J. Watts for secretarial support. We thank Dr. K. Leymaster for assistance in developing the sheep diversity panel and thoughtful discussions and improvements to the manuscript. This work was conducted in part using the resources of the University of Louisville's research computing group and the Cardinal Research Cluster, and we thank Mr. H. Simrall for his assistance.

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.

Supplementary material

Table S1. GDF9, BMP15, and BMPR1B genotypes recorded manually from WGS reads mapped to OAR_v3.1 assembly for the USMARC Sheep Diversity Panel v2.4.

Click here to access the data.

Table S2. Haplotype-phased genotypes (diplotypes) for *GDF9*, *BMP15*, and *BMPR1B* genes in the MSDPv2.4.

Click here to access the data.

Table S3. Effect of sire copies of *BMP15* P252 alleles ("Variant 3") on daughter litter size.

Click here to access the data.

Table S4. Effect of sire copies of *BMP15* Q67 alleles ("Variant 4") on daughter litter size.

Click here to access the data.

Table S5. Effect of sire copies of *BMPR1B* I64 alleles ("Variant 2") on daughter litter size.

Click here to access the data.

Table S6. Effect of sire copies of *BMPR1B* N345 alleles ("Variant 3") on daughter litter size.

Click here to access the data.

Figure S1. Screen image of Integrated Genome Viewer (IGV) software displaying *GDF9* V332I genotype data for eight sheep.

Click here to access the data.

Supplementary File 1. VCF file of 50,357 SNP variants used in comparing WGS genotypes to those from the OvineSNP50 bead array.

Click here to access the data.

References

1. Nicholas FW, Hobbs M: **Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012.** *Anim Genet.* 2014; **45**(2): 157–70.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Bickhart DM, Liu GE: **The challenges and importance of structural variation detection in livestock.** *Front Genet.* 2014; **5**: 37.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Jiang Y, Xie M, Chen W, *et al.*: **The sheep genome illuminates biology of the rumen and lipid metabolism.** *Science.* 2014; **344**(6188): 1168–73.
PubMed Abstract | Publisher Full Text | Free Full Text

4. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, *et al.*: **A map of human genome variation from population-scale sequencing.** *Nature.* 2010; **467**(7319): 1061–73.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Jordan DM, Ramensky VE, Sunyaev SR: **Human allelic variation: perspective from protein function, structure, and evolution.** *Curr Opin Struct Biol.* 2010; **20**(3): 342–50.
PubMed Abstract | Publisher Full Text | Free Full Text

6. MacArthur DG, Manolio TA, Dimmock DP, *et al.*: **Guidelines for investigating causality of sequence variants in human disease.** *Nature.* 2014; **508**(7497): 469–76.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–6.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–92.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Heaton MP, Smith TP, Carnahan JK, *et al.*: **Using diverse U.S. beef cattle genomes to identify missense mutations in *EPAS1*, a gene associated with pulmonary hypertension [version 2; referees: 2 approved].** *F1000Res.* 2016; **5**: 2003.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Juengel JL, Davis GH, McNatty KP: **Using sheep lines with mutations in single genes to better understand ovarian function.** *Reproduction.* 2013; **146**(4): R111–23.
PubMed Abstract | Publisher Full Text

11. Mullen MP, Hanrahan JP: **Direct evidence on the contribution of a missense mutation in *GDF9* to variation in ovulation rate of Finnsheep.** *PLoS One.* 2014; **9**(4): e95251.
PubMed Abstract | Publisher Full Text | Free Full Text

12. de Castro FC, Cruz MH, Leal CL: **Role of Growth Differentiation Factor 9 and Bone Morphogenetic Protein 15 in Ovarian Function and Their Importance in Mammalian Female Fertility - A Review.** *Asian-Australas J Anim Sci.* 2016; **29**(8): 1065–74.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Reader KL, Haydon LJ, Littlejohn RP, *et al.*: **Booroola *BMPR1B* mutation alters early follicular development and oocyte ultrastructure in sheep.** *Reprod Fertil Dev.* 2012; **24**(2): 353–61.
PubMed Abstract | Publisher Full Text

14. Mulsant P, Lecerf F, Fabre S, *et al.*: **Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Mérino ewes.** *Proc Natl Acad Sci U S A.* 2001; **98**(9): 5104–9.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Regan SL, McFarlane JR, O'Shea T, *et al.*: **Flow cytometric analysis of FSHR, BMRR1B, LHR and apoptosis in granulosa cells and ovulation rate in merino sheep.** *Reproduction.* 2015; **150**(2): 151–63.
PubMed Abstract | Publisher Full Text

16. Heaton MP, Leymaster KA, Kalbfleisch TS, *et al.*: **Ovine reference materials and assays for prion genetic testing.** *BMC Vet Res.* 2010; **6**: 23.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Leymaster KA: **Straightbred comparison of a composite population and the Suffolk breed for performance traits of sheep.** *J Anim Sci.* 1991; **69**(3): 993–9.
PubMed Abstract | Publisher Full Text

18. Kalbfleisch T, Heaton MP: **Mapping whole genome shotgun sequence and variant calling in mammalian species without their reference genomes [version 1; referees: 1 approved with reservations].** *F1000Res.* 2013; **2**: 244.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–95.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9.
PubMed Abstract | Publisher Full Text | Free Full Text

21. McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–303.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Heaton MP, Leymaster KA, Kalbfleisch TS, *et al.*: **SNPs for parentage testing and traceability in globally diverse breeds of sheep.** *PLoS One.* 2014; **9**(4): e94851.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Altschul SF, Madden TL, Schäffer AA, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**(17): 3389–402.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Altschul SF, Wootton JC, Gertz EM, *et al.*: **Protein database searches using compositionally adjusted substitution matrices.** *FEBS J.* 2005; **272**(20): 5101–9.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Kijas JW, Lenstra JA, Hayes B, *et al.*: **Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection.** *PLoS Biol.* 2012; **10**(2): e1001258.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Våge DI, Husdal M, Kent MP, *et al.*: **A missense mutation in *growth differentiation factor 9 (GDF9)* is strongly associated with litter size in sheep.** *BMC Genet.* 2013; **14**: 1.
PubMed Abstract | Publisher Full Text | Free Full Text

27. 1000 Genomes Project Consortium, Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Rossetti R, Di Pasquale E, Marozzi A, *et al.*: **BMP15 mutations associated with primary ovarian insufficiency cause a defective production of bioactive protein.** *Hum Mutat.* 2009; **30**(5): 804–10.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Souza CJ, MacDougall C, MacDougall C, *et al.*: **The Booroola (FecB) phenotype is associated with a mutation in the bone morphogenetic receptor type 1 B (BMPR1B) gene.** *J Endocrinol.* 2001; **169**(2): R1–6.
PubMed Abstract | Publisher Full Text

30. Davis GH, Galloway SM, Ross IK, *et al.*: **DNA tests in prolific sheep from eight countries provide new evidence on origin of the Booroola (FecB) mutation.** *Biol Reprod.* 2002; **66**(6): 1869–74.
PubMed Abstract | Publisher Full Text

31. Oraon T, Singh DK, Ghosh M, *et al.*: **Allelic and genotypic frequencies in polymorphic Booroola fecundity gene and their association with multiple birth and postnatal growth in Chhotanagpuri sheep.** *Vet World.* 2016; **9**(11): 1294–9.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Mahdavi M, Nanekarani S, Hosseini SD: **Mutation in *BMPR-IB* gene is associated with litter size in Iranian Kalehkoohi sheep.** *Anim Reprod Sci.* 2014; **147**(3–4): 93–8.
PubMed Abstract | Publisher Full Text

33. Chu MX, Liu ZH, Jiao CL, *et al.*: **Mutations in *BMPR-IB* and *BMP-15* genes are associated with litter size in Small Tailed Han sheep (*Ovis aries*).** *J Anim Sci.* 2007; **85**(3): 598–603.
PubMed Abstract | Publisher Full Text

34. Guan F, Liu SR, Shi GQ, *et al.*: **Polymorphism of *FecB* gene in nine sheep breeds or strains and its effects on litter size, lamb growth and development.** *Anim Reprod Sci.* 2007; **99**(1–2): 44–52.
PubMed Abstract | Publisher Full Text

35. NC-IUB: **Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB).** *Proc Nat Acad Sci U S A.* 1986; **83**(1): 4–8.
PubMed Abstract | Free Full Text

36. den Dunnen JT, Dalgleish R, Maglott DR, *et al.*: **HGVS Recommendations for the Description of Sequence Variants: 2016 Update.** *Hum Mutat.* 2016; **37**(6): 564–9.
PubMed Abstract | Publisher Full Text

37. Hanrahan JP, Gregan SM, Mulsant P, *et al.*: **Mutations in the genes for oocyte-derived growth factors GDF9 and BMP15 are associated with both increased ovulation rate and sterility in Cambridge and Belclare sheep (*Ovis aries*).** *Biol Reprod.* 2004; **70**(4): 900–9.
PubMed Abstract | Publisher Full Text

38. Khodabakhshzadeh R, Mohammadabadi MR, Esmailizadeh AK, *et al.*: **Identification of point mutations in exon 2 of *GDF9* gene in Kermani sheep.** *Pol J Vet Sci.* 2016; **19**(2): 281–9.
PubMed Abstract | Publisher Full Text

39. Souza CJ, McNeilly AS, Benavides MV, *et al.*: **Mutation in the protease cleavage site of *GDF9* increases ovulation rate and litter size in heterozygous ewes and causes infertility in homozygous ewes.** *Anim Genet.* 2014; **45**(5): 732–9.
PubMed Abstract | Publisher Full Text

40. Silva BD, Castro EA, Souza CJ, *et al.*: **A new polymorphism in the *Growth and Differentiation Factor 9 (GDF9)* gene is associated with increased ovulation rate and prolificacy in homozygous sheep.** *Anim Genet.* 2011; **42**(1): 89–92.
PubMed Abstract | Publisher Full Text

41. Nicol L, Bishop SC, Pong-Wong R, *et al.*: **Homozygosity for a single base-pair mutation in the oocyte-specific GDF9 gene results in sterility in Thoka sheep.** *Reproduction.* 2009; **138**(6): 921–33.
PubMed Abstract | Publisher Full Text

42. Martinez-Royo A, Jurado JJ, Smulders JP, *et al.*: **A deletion in the *bone morphogenetic protein 15* gene causes sterility and increased prolificacy in Rasa Aragonesa sheep.** *Anim Genet.* 2008; **39**(3): 294–7.
PubMed Abstract | Publisher Full Text

43. Monteagudo LV, Ponz R, Tejedor MT, *et al.*: **A 17 bp deletion in the Bone Morphogenetic Protein 15 (BMP15) gene is associated to increased prolificacy in the Rasa Aragonesa sheep breed.** *Anim Reprod Sci.* 2009; **110**(1–2): 139–46.
PubMed Abstract | Publisher Full Text

44. Galloway SM, McNatty KP, Cambridge LM, *et al.*: **Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner.** *Nat Genet.* 2000; **25**(3): 279–83.
PubMed Abstract | Publisher Full Text

45. Demars J, Fabre S, Sarry J, *et al.*: **Genome-wide association studies identify two novel *BMP15* mutations responsible for an atypical hyperprolificacy phenotype in sheep.** *PLoS Genet.* 2013; **9**(4): e1003482.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Bodin L, Di Pasquale E, Fabre S, *et al.*: **A novel mutation in the bone morphogenetic protein 15 gene causing defective protein secretion is associated with both increased ovulation rate and sterility in Lacaune sheep.** *Endocrinology.* 2007; **148**(1): 393–400.
PubMed Abstract | Publisher Full Text

47. Wilson T, Wu XY, Juengel JL, *et al.*: **Highly prolific Booroola sheep have a mutation in the intracellular kinase domain of bone morphogenetic protein IB receptor (ALK-6) that is expressed in both oocytes and granulosa cells.** *Biol Reprod.* 2001; **64**(4): 1225–35.
PubMed Abstract | Publisher Full Text

48. Hedges SB, Marin J, Suleski M, *et al.*: **Tree of life reveals clock-like speciation and diversification.** *Mol Biol Evol.* 2015; **32**(4): 835–45.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✓ ✓

Version 1

Referee Report 06 September 2017

doi:10.5256/f1000research.13224.r25465

✓ **Eyal Seroussi**
Institute of Animal Science, Agricultural Research Organization (ARO), Rishon LeZion, Israel

The main aim of the publication entitled “Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity” is to establish a publicly accessible resource of sequenced genomes that represent popular U.S. sheep breeds. A key factor for the usefulness of such resource is the fold coverage. While in Introduction a “16x WGS resource” is mentioned, in Results it is indicated that: “the average calculated read depth (17.0)”, the submissions to the SRA database are all entitled “12x WGS of USMARC Sheep…” and the author’s website referred in the manuscript: “USDA, ARS, USMARC internet site” is entitled: “10x WGS of …”. The latter number would have indicated a limited utility for Genotyping-By-Sequencing (GBS) of SNPs, as the typical coverage is not evenly distributed along the genome, and many loci would lack the minimal coverage of 5x. This coverage allows detecting homozygotes with less than a 0.05% uncertainty, assuming ideal conditions of a 0.5 probability to detect an allele and no sequencing errors. Practically, detection rate is biased and sequence errors do exist. Therefore, the authors put much effort to estimate the rate of genotyping errors by comparing their genotypes with bead array data, concluding that no animals had a SNP genotype accuracy of less than 97%. Since this calculation of error rate involves errors introduced by both genotyping methods, a more straightforward approach can be considered by analyzing non-autosomal loci on chromosome X; there all genotypes must be homozygous. Table S1 offers such a possibility for the *BMP15* gene; where two heterozygotes were encountered out of 288 genotypes, which indicates that either sequencing errors or contaminations may introduce 0.7% of the error.

Indeed, sequence cross contamination is a known problem of sequenced genomes[1]; and sequencing projects should be routinely controlled for DNA contaminants. As the authors did not refer to this problem, I tested one of their 923 submissions that has a median size (4.6 G bases, SRX2186704) by analyzing the sequence reads that do not map to the sheep genome (Oar_v4). Using the *GAP5* software [2], the reads sent to the failures.seq file were de-novo assembled into contigs using *MIRA4* sequence assembler [3]. As the current genome version lacks the Y chromosome, most of these contigs were similar to submissions of Y chromosome orthologs of other ruminants; yet, several contigs resembled the *Onchocerca flexuosa* genome, the largest of which was of 3372 bp and had 502 reads with 83% identity to this worm genome. This suggests that this individual was infected by a parasitic roundworm similar to the species that infect red deer, and that the authors present a valuable resource that is important for parasitology [4]. As I encountered no other DNA contaminants, it is likely that the data presented by the authors is solid and of the highest quality and that the worm DNA was extracted from this animal’s blood.

As for identifying missense variants in 3 genes affecting fecundity, I conclude that the authors left no stone unturned to ensure the validity of their genotypes. E. g., Table S1 indicates that the unique homozygous genotype for *BMPR1B* (individual 200117552) had coverage of 41x fold, suggesting that sequence coverage was increased for this individual to ensure this result. Yet, the use of modern tools for predicting the functional effect of amino acid substitutions[5] should have warned them that this variation (T345N) is not likely to produce a phenotype (PROVEAN score = -2.191, Neutral). In this respect, the 4 novel variations described in the fecundity genes are of minor importance. Nevertheless, the observation that none of the U.S. popular breeds carries the Booroola mutation should have an impact, as introgression of this mutation revolutionized sheep production in Spain and Israel[6]. Despite the apparent weakness of the work in identifying important novel variants of fecundity genes, I approve this work as a valuable genomic resource. The authors are advised to control for DNA contaminants, to avoid the discrepancies described by extending the clear and accurate presentation of this work to the affiliated webpages and to discuss the issues raised by this review.

References

1. Gruber K: Here, there, and everywhere: From PCRs to next-generation sequencing technologies and sequence databases, DNA contaminants creep in from the most unlikely places.*EMBO Rep*. 2015; **16** (8): 898-901 PubMed Abstract | Publisher Full Text
2. Bonfield JK, Whitwham A: Gap5--editing the billion fragment sequence assembly.*Bioinformatics*. 2010; **26** (14): 1699-703 PubMed Abstract | Publisher Full Text
3. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S: Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*. 2004; **14** (6): 1147-59 PubMed Abstract | Publisher Full Text
4. Lopes RJ, Mérida AM, Carneiro M: Unleashing the Potential of Public Genomic Resources to Find Parasite Genetic Data.*Trends Parasitol*. 2017. PubMed Abstract | Publisher Full Text
5. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: Predicting the functional effect of amino acid substitutions and indels.*PLoS One*. 2012; **7** (10): e46688 PubMed Abstract | Publisher Full Text
6. Seroussi E, Rosov A, Shirak A, Lam A, Gootwine E: Unveiling genomic regions that underlie differences between Afec-Assaf sheep and its parental Awassi breed.*Genet Sel Evol*. 2017; **49** (1): 19 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Animal genetics, genomics and bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 29 August 2017

doi:10.5256/f1000research.13224.r24897

✓ **Christine Couldrey**
Research and Development, Livestock Improvement Corporation, Hamilton, New Zealand

The manuscript entitled “Using sheep genomes from diverse U.S. breeds to identify missense variants in genes affecting fecundity” provides an informative overview of a publicly searchable DNA sequence resource for U.S. sheep that the authors have generated. The manuscript is well written.
The manuscript is largely descriptive in nature but does include a proof of principal type study to highlight the potential usefulness of this dataset. The study is appropriately designed and technically sound. Most of the conclusions drawn are adequately supported by the results. Where the authors have not identified phenotypic effects of amino acid changes they have acknowledged that this study does lack power to make definitive statements.

A few details could do with further detail/clarification/some modification:

1. In the last paragraph on page 4 where the manuscript states “when the accuracy of the WGS genotypes exceeded 97%”…… Please explain why 97% was chosen as the threshold.
2. Page 5 under heading Identifying protein variants encoded by GDF9, BMP15, and BMPR1B: the methods describe haplotype phasing, however, it is unclear as to how many of the 96 sheep were able to be phased and used in analysis. It is therefore difficult to determine the validity of this method. The methods or corresponding results section should be expanded to include this information.
3. Page 5 Methods section under heading Statistical analysis of litter size in daughters of carrier rams: Please include information of how many rams had daughter lambing records for each breed and each of the variants identified
4. Page 6 The authors should include some discussion/information about the inclusion of animals that had the lowest concordances with other genotyping platforms, particularly the animals with ~17X coverage and a concordance of ~97%. In some research facilities this (and some of the other animals with greater than 10X coverage and less than 99% concordance) would be treated as suspect and excluded, or further analysis undertaken - if the latter is the case, please include the further analysis.
5. Page 7 the sentence “Alleles encoding the M371 residue…” refers to Table 2, however this residue is not referred to in Table 2. Please correct.
6. Page 7 “We predict that variant residues in highly conserved protein domains are more likely to affect ovulation rate and littler size”. While this sentence could be true, it is also possible that protein domains are highly conserved for functions other than ovulation rate and litter size.
7. Page 12 There is insufficient data to really make the statement “Thus some substitutions at this position may cause loss of function in some mammals but it appears as though Q67 may not be one of them”. Given that the phenotypes around ovarian failure in sheep are likely not well recorded given the culling of sheep relatively early in life.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.