Genome-wide scans identify known and novel regions associated with prolificacy and reproduction traits in a sub-Saharan African indigenous sheep (Ovis aries)

Citation for published version:

Digital Object Identifier (DOI):
10.1007/s00335-019-09820-5

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Mammalian Genome

Publisher Rights Statement:
This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Genome-wide scans identify known and novel regions associated with prolificacy and reproduction traits in a sub-Saharan African indigenous sheep (Ovis aries)

Asrat Tera Dolebo1,2 · Negar Khayatzadeh3 · Aberra Melesse2 · David Wragg4 · Mourad Rekik5 · Aynalem Haile5 · Barbara Rischkowsky2 · Max F. Rothschild6 · Joram M. Mwacharo5

Received: 4 September 2019 / Accepted: 12 November 2019 © The Author(s) 2019

Abstract
Maximizing the number of offspring born per female is a key functionality trait in commercial- and/or subsistence-oriented livestock enterprises. Although the number of offspring born is closely associated with female fertility and reproductive success, the genetic control of these traits remains poorly understood in sub-Saharan Africa livestock. Using selection signature analysis performed on Ovine HD BeadChip data from the prolific Bonga sheep in Ethiopia, 41 candidate regions under selection were identified. The analysis revealed one strong selection signature on a candidate region on chromosome X spanning BMP15, suggesting this to be the primary candidate prolificacy gene in the breed. The analysis also identified several candidate regions spanning genes not reported before in prolific sheep but underlying fertility and reproduction in other species. The genes associated with female reproduction traits included SPOCK1 (age at first oestrus), GPR173 (mediator of ovarian cyclicity), HB-EGF (signalling early pregnancy success) and SMARCAL1 and HMGN3a (regulate gene expression during embryogenesis). The genes involved in male reproduction were FOXJ1 (sperm function and successful fertilization) and NME5 (spermatogenesis). We also observed genes such as PKD2L2, MAGED1 and KDM3B, which have been associated with diverse fertility traits in both sexes of other species. The results confirm the complexity of the genetic mechanisms underlying reproduction while suggesting that prolificacy in the Bonga sheep, and possibly African indigenous sheep is partly under the control of BMP15 while other genes that enhance male and female fertility are essential for reproductive fitness.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00335-019-09820-5) contains supplementary material, which is available to authorized users.

Introduction
The evolution of novel traits is underpinned by genetic changes encoding new phenotypes. The genetic basis of traits that are controlled by few genes has been well established. For example, mutations in MC1R influences coat colour in animals ranging from mice (Hoekstra et al. 2006) to camels (Almathen et al. 2018). However, little remains known on the genetic control of complex traits, which have proven challenging to study using traditional approaches. Recent developments in next-generation sequencing and associated techniques (single-nucleotide polymorphism (SNP) genotyping arrays and bioinformatics pipelines) have provided a unique opportunity to examine genes, and gene networks, encoding complex phenotypes in domestic animals (Andersson and Georges 2004; Gouivea et al. 2014). Diverse geographic adaptation and selection pressure have resulted in shared and population-specific phenotypes in many livestock species (Xu et al. 2015). Prolificacy is one such phenotype that has been observed in several
breeds of sheep in Europe, Africa, Middle East, and Asia and other mammalian species. Whether the trait evolved independently, within or across species and/or breeds of livestock in different geographic regions or, already existed in the genome of the wild ancestor at the time of domestication remains unknown. What is clear, however, is that the trait is under the control of a few genes with large effects (Davis 2004, 2005; Monestier et al. 2014; Abdoli et al. 2016). Several studies identified causative variants in three related oocyte-derived members of the transforming growth factor-beta (TGF-β) superfamily including bone morphogenetic protein receptor 1B (BMPR1B), bone morphogenic protein 15 (BMP15) and growth differentiation factor 9 (GDF9) (Davis 2004, 2005; Abdoli et al. 2016) which have been shown to be essential for ovulation rate and follicular growth (Juengel and McNatty 2005; Knight and Glister 2006). The BMPR1B gene located on chromosome (Oar) 6 has been found in mostly Asian breeds. The sole mutation observed in this gene is present in the Small-tailed Han and Hu sheep in China, the Kendrickapa and Garole sheep in India and the Javanese thin-tailed sheep in Indonesia but seems to be absent in European breeds (Davis et al. 2002, 2005; Jansson 2014; Abdoli et al. 2016). BMP15 and GDF9 located on OarX and 5, respectively, appear to be the main prolificacy genes in European and Middle East (specifically Iran) sheep breeds. In BMP15 eight mutations, which differ slightly in type and effect, have been discovered in different sheep breeds and populations (see reviews by Davis 2004, 2005; Abdoli et al. 2016 and references therein). Four mutations affecting ovulation rate have been discovered to date in GDF9 (see reviews by Davis 2004, 2005; Abdoli et al. 2016). Other genes that have also been reported in European sheep include B4GALNT2 on Oar11 and FecX2 on OarX (see Abdoli et al. 2016). New mutations are, however, continuously being discovered in these genes, the latest were reported in Tunisian Barbarine (Lassoued et al. 2017) and three Iranian breeds (Amini et al. 2018). These findings suggest the genetic control of prolificacy traits varies between breeds.

Sub-Saharan Africa (SSA) is home to several breeds of prolific indigenous sheep but with no known history, or information, on any form of either natural and/or artificial selection targeting the trait. In spite of the large body of knowledge generated in the last decade on prolificacy in domestic sheep, the genetic basis for the trait in SSA sheep remains poorly investigated. The various studies documenting variations in major genes and the inherent causative mutations associated with ovulation and litter size the species justify the investigation of genes of major effect in prolific SSA sheep. African indigenous sheep are known to share their genome ancestry with sheep from the Middle East and the Indian sub-continent (Muigai and Hanotte 2013; Mwacharo et al. 2017). The expectation therefore is that one of the three members of TGF-β superfamily of genes could be responsible for the trait in SSA sheep.

Bonga is a breed of indigenous sheep found in southwestern Ethiopia. It displays good maternal characteristics and is one of the naturally prolific breeds of sheep found in SSA. It shows an average litter size of 1.54 ± 0.006 (average range = 1.25 ± 0.433 to 2.12 ± 0.499) and above average reproductive efficiency under smallholder village production (Haile et al. Unpublished). Edea et al. (2012) reported an average litter size of 1.36 and average twinning rate of 36.3 ±4.7% in the breed. Other naturally prolific breeds of sheep in SSA include Horro (litter size of 1.40 and twinning rate of 39.9%; Edea et al. 2012) and Doyogana breeds in Ethiopia, the West African Dwarf/Djallonke sheep found across southwest and Central Africa. Other prolific breeds of sheep in the continent outside SSA are the Barbarine and D’man sheep found in Tunisia and Morocco, respectively. The prolificacy in the Barbarine and D’man sheep has, however, been enhanced through artificial selection programmes.

In this study, we performed genome-wide scans of selection signatures using three approaches (FST, hapFLK, XP-EHH) and genotype data generated with the Illumina OvineHD SNP Chip array in Bonga sheep sampled from farmers’ flocks, as a proxy for other prolific sheep found in SSA, to identify candidate genomic regions and genes associated with the trait. We identified a strong selection signature on OarX spanning BMP15 and uniquely, a diverse range of genomic regions spanning several candidate genes, never reported before in prolific sheep, but known to be associated with fertility and reproduction in other species. Our results suggest that, prolificacy in SSA indigenous sheep is a function of the actions of BMP15 and several genes that are associated with male and female fertility.

Results

The phenotypic dataset consisted of 98 litter size records of Bonga sheep, a non-seasonal breeder, that were collected between 2009 and 2018 from farmers’ flocks participating in a community-based breeding programme (CBBP). For this study, litter size was considered a prolificacy trait of the dam and one of the indicators of improved reproduction. It was defined as the number of lambs born alive per lambing. The most prolific ewes (n = 74) with twins (n = 38), triplets (n = 35) and quadruplets (n = 1) lambs born alive per lambing and non-prolific ones (one lamb born alive per lambing; n = 24), for at least three parities, were sampled from different farmers’ flocks. Genotyping was performed with the Illumina OvineHD BeadChip, which includes 606,006 genomic variants and 30,000 functional putative variants, at GeneSeek Inc (http://genomics.neogen.com/en/). The genotype data were assessed for quality with
Genome-wide scans identify known and novel regions associated with prolificacy and reproduction…

PLINK 1.9 (www.cog-genomics.org/plink2). Variants with no assigned genomic positions, call rates lower than 95%, large Hardy–Weinberg equilibrium (HWE) deviations ($P$ value $< 1 \times 10^{-6}$) and minor allele frequency (MAF) $< 0.01$, and samples with call rates $< 98\%$ were excluded from the final dataset. Following quality filtering, 457,087 variants and 84 individuals (33 ewes with twins, 30 with triplets, 1 with quadruplet, and 20 with single births) were retained for analysis.

To ensure that there were no biases attributed to stratification arising from fine-scale population genetic structure due to variations between and within farmers’ flocks or any other unknown evolutionary attribute, principal component analysis (PCA) and NetView were performed using the retained genetic variants. No genetic stratification was detected (Figs. 1a, b) with the first principal component of the PCA explaining 80.92% of the total genetic variation. Irrespective of their prolificity (twinning, triplet, quadruplet), all the ewes clustered close together with only eight outliers (five ewes with triplets and two with singlets) being observed.

The retained dataset, following quality filtering, was used to investigate genome-wide signatures of selection using three cross-population selection tests; $F_{ST}$ (Biswas and Akey 2006), XP-EHH (Sabeti et al. 2007) and hapFLK (Fariello et al. 2013). For the analysis, prolific ewes, defined as those with twin, triplet and quadruplet births, for at least three consecutive lambing seasons, were taken as one group and the non-prolific ones (ewes with single births) formed the contrasting group. The grouping was informed by the objective of detecting selection signatures that can be attributed, to a large extent, to differences in prolificacy. For this reason, we avoided using a different breed due to the high likelihood of detecting strong selection signatures arising from genetic differences between breeds which might have masked the ones attributing to prolificacy.

Based on the ovine RefSeq gene annotation, there were five and eight candidate regions revealed by XP-EHH and hapFLK, respectively, that overlapped no gene(s) (Table 1). For the candidate regions that overlapped with gene(s), two (one on Oar5 and the other on OarX; Table 1; Fig. 2a), were identified from the empirical genome-wide distribution of $F_{ST}$ values. The region on Oar5 spanned 18 annotated genes and five novel protein coding transcripts, while the one on OarX, the most significant signature, spanned eight annotated genes and seven novel protein coding transcripts (Table 1). The XP-EHH detected 18 candidate regions, spanning 20 annotated genes and four protein coding transcripts, across 12 chromosomes (Table 1; Fig. 2b). The hapFLK revealed 21 candidate regions spanning 31 annotated genes and 13 protein coding transcripts, across 15 chromosomes (Table 1; Fig. 2c).

The candidate region on OarX overlapped between $F_{ST}$ and hapFLK tests. The $BMP15$ gene, which has been implicated in prolificacy in several breeds of European and Middle East sheep (Davis 2004, 2005), occurred in this region within the most significant $F_{ST}$ and hapFLK windows. The gene ($BMP15$), however, occurred 259,479-base pairs (bp) upstream of the region revealed by XP-EHH. A region on Oar13 that overlapped between hapFLK and XP-EHH spanned $DOK5$ (Docking Protein 5) gene, an adapter intracellular protein that is involved in signal transduction and is expressed in lymphocytes and T cells in human and mice and may modulate various T cell functions (Favre et al. 2003). The $GDF9$ gene occurred 4,456,484-bp downstream of the candidate region revealed by $F_{ST}$ on Oar5. None of the three candidate regions identified by XP-EHH on Oar5 overlapped with $GDF9$ or the $F_{ST}$ region.
Table 1  Candidate regions and associated genes revealed by FST, hapFLK and XP-EHH selection signature analysis in the Bonga breed of sheep

<table>
<thead>
<tr>
<th>Approach</th>
<th>Chromosome</th>
<th>Genomic region (bp)</th>
<th>Number of significant SNPs</th>
<th>Genes present</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>From - To</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>hapFLK</strong></td>
<td>1</td>
<td>128490718 - 128525811</td>
<td>10</td>
<td><strong>APP</strong></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>119482589 - 119548236</td>
<td>7</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>217377972 - 217398724</td>
<td>6</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>99126620 - 99326839</td>
<td>48</td>
<td><strong>SLC9A4</strong>, <strong>IL18RAP</strong>, <strong>IL18R1</strong>, <strong>IL1RL1</strong></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3548252 - 3582020</td>
<td>7</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>41309054 - 41578812</td>
<td>41</td>
<td><strong>MANEA</strong></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>63639731 - 63649616</td>
<td>4</td>
<td><strong>ECT2L</strong></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>86323485 - 86432306</td>
<td>18</td>
<td><strong>DECR1</strong>, <strong>NBN</strong>, <strong>OSGIN2</strong></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>21701005 - 21716773</td>
<td>8</td>
<td><strong>GLOD4</strong>, <strong>ENSOART00000004387</strong>, <strong>GEMIN4</strong>, <strong>ENSOART00000012524</strong></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>54540051 - 54638872</td>
<td>29</td>
<td><strong>RNF157</strong>, <strong>FOXJ1</strong>, <strong>EXOC7</strong>, <strong>ZACN</strong>, <strong>GALR2</strong>, <strong>SRP68</strong>, <strong>EVPL</strong></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>49122258 - 49208872</td>
<td>8</td>
<td><strong>ENSOART00000003233</strong>, <strong>TMEM240</strong>, <strong>ENSOART00000003483</strong>, <strong>VWA1</strong>, <strong>TMEM88B</strong>, <strong>ANKRD65</strong>, <strong>MRPL20</strong>, <strong>CCNL2</strong></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>6331437 - 6352500</td>
<td>7</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>81406516 - 81586998</td>
<td>36</td>
<td><strong>DOK5</strong></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>12372110 - 12484709</td>
<td>20</td>
<td><strong>ENSOART00000012708</strong>, <strong>FBXO31</strong>, <strong>ENSOART00000012902</strong></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>47181073 - 47266038</td>
<td>14</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>60146782 - 60155318</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>8519837 - 8539965</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>35706919 - 35768326</td>
<td>13</td>
<td><strong>CDKAL1</strong></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>17429910 - 17498671</td>
<td>12</td>
<td>–</td>
</tr>
</tbody>
</table>
Genome-wide scans identify known and novel regions associated with prolificacy and reproduction…

In total, we observed 73 annotated genes in 28 candidate regions that were identified to be under selection by \( F_{S_T} \) (2 regions), XP-EHH (18 regions) and hapFLK (21 regions). Functional enrichment was performed with DAVID 6.3 (Huang et al. 2009a, b) resulting in four enriched clusters of genes (Supplementary Table S1). The top two clusters were associated with immune responses encompassing (i) Toll/interleukin-1 receptor homology (TIR) domain (enrichment score = 2.82) and (ii) immunoglobulin/immunoglobulin-like (IG) domain (enrichment score = 1.27). Protein–protein interactions (PPI) and gene ontology (GO) enrichments were investigated with STRING (Szklarczyk et al. 2019). The network proteins encoded by the 73 candidate genes had significantly more interactions among themselves than was expected for a random set of proteins of similar size drawn from the genome (33 edges identified; PPI enrichment \( P \) value = 0.00612; Fig. 3). STRING revealed three GO biological process terms that were the most enriched (Supplementary Table S2). The PFAM, InterPRO and SMART protein domains were all associated with Toll/interleukin-1 receptor (TIR) domain superfamily (Supplementary Table S3) while the SMART protein domains included immunoglobulin-like domains as one of the most enriched. The Reactome pathways were associated with interleukin-18 signalling (BTA-9012546; false discovery rate = 0.0492). Apart from \( BMP15 \) and \( GDF9 \), that are known to be associated with prolificacy across a wide range of prolific sheep in Europe and the Middle East, literature mining identified several candidate genes associated with female and male fertility and reproduction functions (Table 1) in other species but which have not yet been reported in prolific sheep.

In total, we observed 73 annotated genes in 28 candidate regions that were identified to be under selection by \( F_{S_T} \) (2 regions), XP-EHH (18 regions) and hapFLK (21 regions). Functional enrichment was performed with DAVID 6.3 (Huang et al. 2009a, b) resulting in four enriched clusters of genes (Supplementary Table S1). The top two clusters were associated with immune responses encompassing (i) Toll/interleukin-1 receptor homology (TIR) domain (enrichment score = 2.82) and (ii) immunoglobulin/immunoglobulin-like (IG) domain (enrichment score = 1.27). Protein–protein interactions (PPI) and gene ontology (GO) enrichments were investigated with STRING (Szklarczyk et al. 2019). The network proteins encoded by the 73 candidate genes had significantly more interactions among themselves than was expected for a random set of proteins of similar size drawn from the genome (33 edges identified; PPI enrichment \( P \) value = 0.00612; Fig. 3). STRING revealed three GO biological process terms that were the most enriched (Supplementary Table S2). The PFAM, InterPRO and SMART protein domains were all associated with Toll/interleukin-1 receptor (TIR) domain superfamily (Supplementary Table S3) while the SMART protein domains included immunoglobulin-like domains as one of the most enriched. The Reactome pathways were associated with interleukin-18 signalling (BTA-9012546; false discovery rate = 0.0492). Apart from \( BMP15 \) and \( GDF9 \), that are known to be associated with prolificacy across a wide range of prolific sheep in Europe and the Middle East, literature mining identified several candidate genes associated with female and male fertility and reproduction functions (Table 1) in other species but which have not yet been reported in prolific sheep.

**Discussion**

In this study, we used the prolific Bonga sheep found in Ethiopia, as a proxy to other prolific indigenous African sheep which, apart from the thin-tailed West African Dwarf, are all fat-tailed and of the same genetic background (Mugai and Hanotte 2013) to investigate the genetic basis of prolificacy. We applied three methods, \( F_{S_T} \), hapFLK and XP-EHH, to exploit their strengths and increase the reliability of our findings by minimizing biases associated with each method (Simianer 2014). We were also unsure about
the evolutionary selection time span of prolificacy and the best method to detect candidate selection signatures associated with the trait. The $F_{ST}$, which is directly related to variance(s) in allele frequency, is the most commonly used method to detect selection signatures. The hapFLK detects selection signatures based on differences in haplotype frequencies while accounting for hierarchical population structure. XP-EHH detects long-range haplotypes or recent positive selection, where the selected loci are close to fixation in one population but remain polymorphic in another based on the relationship between an allele and its surrounding linkage disequilibrium (LD). Approaches such as $F_{ST}$ and hapFLK detect better long-term selection because they are dependent on the accumulation of mutations around the causative variant. Their resolving power, however, declines if the selection advantage is small, as it takes longer for the frequency of the favoured allele to increase to the point of detection. LD-based methods, such as XP-EHH, retain
superior detection power when a new mutation arises in a population due to an adaptive advantage or an existing variant being exposed to a new environment that provides favourable selection pressure, resulting in an increase in its frequency but without fixation (Fagny et al. 2014).

Functional enrichment analysis using the highest classification stringency in DAVID 6.8 revealed four enriched clusters with genes linked to immune functions being over-represented, a result which was replicated by STRING. The importance of the immune system in reproduction in sheep has recently been reported (Pokharel et al. 2019). Following arguments advanced in other studies that found an over-representation of immune-related genes in the genomes of tropically adapted livestock (Xu et al. 2015; Bahbahani et al. 2017; Mwacharo et al. 2017), it can also be argued that selection has enhanced adaptive immune response to tropical infections in the genomes of Bonga sheep. This is possible, but unlikely to explain the result of our study given its analytical design which contrasted prolific and non-prolific ewes of the same breed from the same ecological environment and therefore exposed to similar pathogens, parasites and infections. An alternative explanation or hypothesis that we favour associates the immune function with enhancement of the process and success of fertilization in the female reproductive tract. As in other eutherians, in domestic sheep, fertilization also occurs internally and ejaculated spermatozoa are normally retained in functional pre-ovulatory sperm reservoirs in female genital tracts until ovulation. Rather than being eliminated, the immunologically foreign sperm is tolerated by the female immune defence system. It has been observed that semen can signal genomic shifts that modulate expression of genes linked to immune function in poultry (Das et al. 2009; Long et al. 2003; Huang et al. 2016) and mammals (Almiñana et al. 2014; López-úbeda et al. 2015), resulting in a state of immune tolerance during the lengthy storage of spermatozoa in oviductal sperm reservoirs (Holt and Fazeli 2016). A large subset of differentially expressed genes that suppress local immune defence in oviductal sperm reservoirs following mating and mating-induced changes in the expression of immune activating genes in utero-tubal junction has also been reported in chicken and pigs, respectively (Atikuzzaman et al. 2017). The upregulation of immune defence genes in the oviduct following mating was also observed in mice (Fazeli et al. 2004). The
interval between sperm deposition and the change in gene expression seems to cover the time period when spermatozoa are in the sperm reservoir. The activation of immune response genes therefore serves to cleanse the genital tract of redundant spermatozoa, foreign proteins and/or pathogens, for the descending ova or embryos. It is therefore a possibility that high prolificacy has made the oviduct of prolific individuals less responsive to antigenic seminal fluid. This creates an appropriate immune-balanced physiological environment tailored for sperm survival and fertilization.

The $F_{ST}$ and hapFLK identified one overlapping candidate region on OarX. As was expected based on our sampling and analytical design, the most significant window in this region spanned BMP15 that plays a role in various functions implicated in prolificacy (Persani et al. 2014; Monestier et al. 2014). The candidate region revealed by XP-EHH on OarX was 259,479-bp downstream of BMP15. The $F_{ST}$ test also revealed a region on Oar5 which was 4.4 Mb downstream of GDF9, an important paralog of BMP15. As the extent of LD declines rapidly from 0 to 300 kb in the ovine genome (Kijas et al. 2014; Al-Mamun et al. 2015), LD between the SNPs found on the XP-EHH candidate region on OarX and BMP15 is thus expected. This result suggests that BMP15 may be the primary candidate gene responsible for prolificacy in Bonga sheep. Experimental disruption of BMP15 in mice result in mild defects in female fertility (Su et al. 2008) whereas natural missense mutations result in variable phenotypes in ewes, ranging from hyperprolificacy to complete sterility, depending on a fine gene dosage mechanism involving GDF9 (Belli and Shimasaki 2018). This is the first study, to the best of our knowledge, to identify a candidate genomic region and gene associated with prolificacy in an indigenous SSA sheep breed. It is therefore possible that BMP15 may also encode the trait in other prolific SSA sheep. The expression of MAGED1 has been detected in mice embryos as early as E3 stage (Kendal et al. 2002). Similarly, the expression of SMARCAL1 and HMGN3a has been detected in bovine oocytes and in early embryos (Uzun et al. 2009), whereas ELK3 has been isolated from 16-day mouse embryos and one of its transcripts was expressed predominantly in 8- to 14-day embryos (Nozaki et al. 2009). These expression patterns highlight the importance of these genes in regulating gene expression, during embryonic development. The KDM3B gene is highly expressed in female reproductive organs including the ovary, oviduct and uterus. Knockout of Kdm3b in female mice resulted in irregular oestrous cycles and decreased ovulation capability, fertilization rate, uterine decidual response and circulating levels of 17β-oestradiol (Liu et al. 2015a). Thus KDM3B could play a crucial role in regulating oestrous and menstrual reproduction cycles and in preparing the uterus for successful implantation of multiple ova.

The candidate regions identified in this study also spanned several genes implicated in male fertility and reproduction in other species, but not reported in prolific sheep. They included FOXJ1, NME5, PKD2L2, MAGED1 and KDM3B. As discussed, the latter three genes have been implicated in female reproduction. FOXJ1 is a key transcription factor for the formation of motile cilia in human (Lyons et al. 2006), mice and xenopus (Weidemann et al. 2016). Immotile-cilia syndrome has been associated with male and female infertility in humans as motile cilia are critical in propelling ova along the fallopian tube while motility in sperm flagellum is also critical for sperm function and successful fertilization (Afzelius and Eliasson 1983). In mice and xenopus, CFAP157 was identified as a novel target protein for FOXJ1 and is only essential during spermatogenesis but is expressed in motile ciliated tissues. The prominent expression of PKD2L2 and its encoded protein Polycystin-L2 in adult mouse testis, spermatocytes and spermatids
Genome-wide scans identify known and novel regions associated with prolificacy and reproduction…

(Guo et al. 2000) suggests a role in spermatogenesis (Chen et al. 2008) and testicular maturation (Kierszenbaum 2004). Polycystin-L2 also modulates intracellular calcium concentration during spermatogenesis. Calcium ions are critical in regulating sperm cell functions including capacitation, progressive motility, hyperactivated motility and acrosome reaction, which are important during fertilization (Bedford 1998; Darszon et al. 1999; Zhang and Gopalakrishman 2005). NME5 also exhibits a strict testis-specific expression in spermatogonia and early spermatocytes in several vertebrates (Muneir et al. 1998; Hwang et al. 2003; Desvignes et al. 2009). The protein encoded by KDM3B is also highly expressed in multiple cell types in mouse testes, such as Leydig and sertoli cells, spermatogonia and spermatocytes, at different stages of differentiation and has also been observed in epithelial cells of the caput epididymis, prostate and seminal vesicle (Liu et al. 2015b). Knockout of Kdm3b in male mice resulted in reduction in the number of pups produced by breeding pairs due to a decrease in the number of litters, fewer number of mature sperms in cauda epididymis displaying significantly reduced sperm motility and a significant reduction in circulating levels of 17beta-estradiol, a modulator of sperm maturation and male sexual behaviour. Like Kdm3b knockout males, MAGED1-deficient male mice also displayed severely impaired male coital behaviour resulting in male infertility which is attributed to deficient production of mature oxytocin in hypothalamus indicating that MAGED1 is required for oxytocin processing and stability (Dombret at al. 2012). The occurrence of these genes in the candidate regions suggests that they are essential in the maintenance of the process of spermatogenesis and normal male sexual behaviour to ensure successful fertilization of a large number of ova generated in prolific ewes.

Conclusions

The overall objective of this study was to identify genes associated with prolificacy in a prolific SSA fat-tailed breed of sheep. Unexpectedly, we also identified several known and novel candidate genes implicated in male and female fertility and reproduction in other species, suggesting that such genes could be hotspots of selection in indigenous SSA prolific breeds of sheep. The findings suggest that enhanced reproduction in prolific ewes entails not only prolificacy genes but also epistatic effects with genes associated with other reproduction traits. Although we identify BMP15 as the main candidate gene for prolificacy in Bonga sheep, the exact causative variants need to be determined to further confirm whether they are novel or are part of what has been reported in prolific breeds of sheep from Europe and the Middle East. It is important to note that the sample size used here, 84 individuals, is rather low. This may have underpowered our analysis; thus, our findings should be interpreted with caution and need validation using a larger subset of animals and populations. Our findings are of significance given that reproductive traits have low to medium heritability and thus do not exhibit a noticeable response to phenotypic selection in traditional breeding methods based on phenotypic data only. The incorporation of the genetic information, such as revealed here, in such breeding programmes (e.g. CBBPs) via either genomic selection (GS), marker-assisted selection (MAS), genome-wide association studies (GWAS) or genomic best linear unbiased predictions of breeding values, could enhance response to selection towards the genetic improvement of reproductive performance.

Materials and methods

Samples, genotyping and quality control

A total of 95 ewes belonging to the Bonga breed of sheep were sampled from four locations (Shuta n = 33, Boqa n = 45, Buta n = 13, Meduda n = 4) in Southwestern Ethiopia. All the 95 animals had at least three lambing parities and came from farmers flocks that are participating in a community-based breeding programme (CBBP) where performance recording is undertaken. From the records of the 95 animals, 31 gave birth to single lambs, 33 to twins, 30 to triplets and one to a quadruplet. Whole blood was collected from each animal via jugular venipuncture with EDTA as the anticoagulant and later transferred to Whatman™ FTA™ Classic Cards (GE Healthcare) for storage. Genotyping was done using FTA™ preserved blood samples with the Ovine Infinium® HD SNP BeadChip (http://genomics.neogen.com/en/illumina-ovine-hd-beadchip) at GeneSeek Inc. (Lincoln NE, USA). The BeadChip includes 606,006 genomic variants designed by the International Sheep Genomics Consortium (ISGC), nearly all the contents from the original OvineSNP50 array and 30,000 putative functional variants. The raw genotypes were first analysed with GenomeStudio (Illumina; GenCall score for raw genotypes > 0.15) which was used to extract the genotypes in standard format for PLINK 1.09 (http://www.cog-genomics.org/plink/1.9/). Chromosomal coordinates for each SNP were obtained from the Ovine Oar 3.1 genome assembly (https://www.ensembl.org/Ovis_aries/Info/Index). The raw dataset was filtered to remove animals with more than 10% missing genotypes, SNPs with no known positions in the genome, SNPs with a call rate lower than 95%, a minor allele frequency lower than 1%, and large HWE deviations (P < 0.000001). All the SNPs mapping to the X chromosome were retained in the final dataset because genes associated with prolificacy have also
been observed on this chromosome. Hence, 84 individuals and 457,086 SNPs were available for analysis following quality filtering. This dataset was first used to assess genetic relationship and structure by performing the principal component analysis (PCA) with ADEGENET (http://adegenet.r-forge.r-project.org/) executed in R (https://www.R-project.org). To further assess possible fine-scale population genetic structure, NetView (Neuditschko et al. 2012) was also used to analyse the genotype data, testing up to ten near-neighbour genetic clusters.

Identification of candidate genomic regions under selection

Three selection detection tests, $F_{ST}$, hapFLK and XP-EHH, were implemented. Prior to performing selection signature mapping, the genotyped individuals were classified, a priori, into two groups, prolific and non-prolific ewes. The prolific group included ewes with twins, triplets and quadruplet litter sizes. The non-prolific group included ewes with single litter sizes.

Smoothed $F_{ST}$ test statistic

The $F_{ST}$ test indicates genetic differentiation among groups of individuals/populations/breeds arising from different evolutionary pressures acting in a segment of the genome. To identify loci under selection, we calculated the allele frequencies for each of the 457,086 retained SNPs for the two contrasting groups of prolific and non-prolific ewes. The allele frequencies were used to calculate $F_{ST}$ values for each locus as a measure of group differentiation following Porto-Neto et al. (2013a, b). For each SNP, $F_{ST}$ was calculated as the squared deviation of the average frequency in a group from the average frequency across the groups divided by the allele frequency variance ($p^*q$). To identify regions under selection, the non-prolific group was compared against the prolific one and the pairwise group values were averaged to obtain a single $F_{ST}$ value per SNP for each group. To facilitate the identification of genomic regions containing more extreme $F_{ST}$ values, the individual SNP values of $F_{ST}$ were grouped within genomic windows, using a kernel regression smoothing algorithm (Gasser et al. 1991) implemented with LOKERN in R (Hermann 2016). This method uses local averaging of the observations ($F_{ST}$ values) when estimating the regression function. By testing window sizes of two, five and ten SNPs, we chose a window of five SNPs as it gave sufficient smoothing and showed the best signals. Higher scores of smoothed $F_{ST}$ for individual loci or genomic regions indicate stronger signal of genetic differentiation/

hapFLK test statistic

As a complementary approach to mapping selection sweeps, we used hapFLK 1.3 (https://forge-dga.jouy.inra.fr/projects/hapflk), which implements the FLK (Bonhomme et al. 2010) and hapFLK (Fariello et al. 2013) algorithms. The FLK tests the neutrality of polymorphic markers by contrasting their allele frequencies in a set of populations against what is expected under neutrality. The hapFLK extends the FLK test to account for differences in haplotype frequencies between populations. This method has been shown to be robust with respect to bottlenecks and migration events (Fariello et al. 2013). To perform hapFLK analysis, Reynold’s genetic distances between the prolific and non-prolific ewes were calculated and converted to a kinship matrix with an R script (available at https://forge-dga.jouy.inra.fr/projects/hapflk/documents). Subsequently, by assuming ten haplotype clusters in the linkage disequilibrium (LD) model (− K 10; number of haplotype clusters determined by running a fastPHASE cross-validation analysis), the hapFLK statistics were computed and averaged across 20 expectation–maximization runs to fit the LD model (− nfit = 20). The standardization of the statistics using the corresponding python script provided with the software allowed the estimation of the associated $P$ values from a standard normal distribution. To correct for multiple testing, we considered the threshold of the nominal $P$ value as $<0.001$ to identify the significant haplotypes following previous studies using hapFLK analysis on the Sheep HapMap dataset (Fariello et al. 2014; Kijas 2014).

XP-EHH test statistic

We also used the SelScan package (Szpiech and Hernandez 2014) to perform an additional analysis based on the cross-population extended haplotype homozygosity (XP-EHH) test (Sabeti et al. 2007). This statistic compares the EHH profiles for bi-allelic SNPs between two populations rather than two alleles in a single population. It is defined as the log of the ratio of the integrals of the EHH profiles between the two populations. It is calculated as:

Unstandardized XP-EHH = ln (iHH_A/iHH_B)

where iHH_A and iHH_B are the integrated EHH of a given core SNP in population A and B, respectively. The comparison between populations normalizes the effects of large-scale variation in recombination rates on haplotype diversity.
and has a high statistical power to detect sweeps that are close to fixation (Sabeti et al. 2007). We used the software developed by Pickrell et al. (2009) to estimate unstandardized XP-EHH statistics using all the SNPs that were retained following quality control. The unstandardized XP-EHH statistics were standardized using their means and variances in the comparison. Because previous studies found that the standardized XP-EHH statistics follow the standard normal distribution (Sabeti et al. 2007; Ma et al. 2014; Zhao et al. 2016), the \( P \) values for SNPs were estimated using the standard normal distribution function. Positive and negative XP-EHH estimates indicated positive recent selection in prolific and non-prolific ewes, respectively. For consistency with the threshold used for hapFLK, we considered as significant those positions showing \( P \) values < 0.001.

**Functional enrichment of the candidate regions under selection**

For the three selection mapping approaches, positions that showed evidence of selection (mean \( F_{ST} \pm 3 \text{ SD} \); or a \( P \) value <0.001 for hapFLK and XP-EHH) were considered to be the result of selection sweeps. The genes that were either partially or fully covered by these regions were identified based on the ovine 3.1 reference genome assembly using Ensembl Comparative Genomics Resources Database Release 94 (https://www.ensembl.org/index.html). Functional enrichment analysis was performed with the functional enrichment clustering tool of DAVID Bioinformatics Resources 6.8 (Huang et al. 2009a, b). Each gene was analysed and enrichment analysis was performed using Ovis aries as the target species and the Bos taurus genome supplied with DAVID 6.8 as the background species. Corrections for multiple testing were performed by applying the Benjamini and Hochberg (1995) approach. For functional enrichment clustering, an enrichment score of 1.3 was taken as the threshold following the authors of DAVID 6.8. A search of the literature was also performed to identify phenotypes that are known to be affected by variation in the genes found in the candidate regions in other species.

Functional protein–protein interaction (PPI) networks and gene ontology (GO) terms encoded by the candidate genes were also investigated using STRING Genomics 11.0 (Szklarczyk et al. 2019) with the Bos taurus as the background species. STRING provides (i) known PPI from curated databases or experiments and (ii) PPI predicted on the basis of gene neighbourhoods, fusions and co-occurrences, text mining in literature, co-expression or protein homology. A global PPI network which retained interactions with a high level of confidence (PPI enrichment score > 0.4) was constructed.

**Acknowledgements** The authors are grateful to the National Institute of Animal Agriculture for developing the sheep genotyping assay. The authors gratefully acknowledge the technical assistance provided by the staff of the Bonga Agricultural Research Centre. Financial support for this study was provided by the CGIAR Research Programme on Livestock. The genotyping was performed by the Illumina Greater Good Initiative and additionally by discounted prices from GenSeek Inc. The genotyping reagents were provided by MFR.

**Author contributions** MR, AH, BR and JMM contributed to the study conception and design. On-farm phenotypic recording and tissue sample collection were done by ATD and AM. Data analysis was done by NK, DW and JMM. Genotyping reagents were provided by MFR. The study was supervised by JMM who also wrote the manuscript. All authors read and approved the final manuscript.

**Data availability** The datasets analysed during the current study are available from the corresponding author on reasonable request.

**Compliance with Ethical standards**

**Conflict of interest** The authors declare no conflict of interest.

**Research involving human and animal participants** The animals used in this study are owned by farmers who are participating in a community-based indigenous sheep breeding programme. The farmers were aware of the objectives of the study and the animals were sampled with their permission in the presence of a qualified veterinarian following the guidelines for the care and use of animals of the Ministry of Livestock and Fisheries Resources, Peoples Democratic Republic of Ethiopia. No animal was injured during the sampling process. All the procedures described here were approved by the Institute Ethics Committee of the International Centre for Agricultural Research in the Dry Areas (ICARDA).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

**References**


Bauman BM, Yin W, Gore AC, Wu TJ (2017) Regulation of gonadotropin-releasing hormone-1(-5) signaling genes by estradiol is age dependent. Front Endocrinol 8:282


Huang A, Isobe N, Obitsu T, Yoshimura Y (2016) Expression of lipases and lipid receptors in sperm storage tubules and possible role of fatty acids in sperm survival in the hen oviduct. Theriogenology 85:1334–1342

Jansson T (2014) Genes involved in ovulation rate and litter size in sheep. BSc These, Swedish University of Agricultural Sciences. Uppsala Sweden. p 19


Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.