UNIVERSITY OF TARTU

FACULTY OF SCIENCE AND TECHNOLOGY

INSTITUTE OF MOLECULAR AND CELL BIOLOGY

DEPARTMENT OF BIOTECHNOLOGY

Prediction and characterization of PRD-like homeobox genes *ARGFX*, *DUXA* and *NOBOX* in the bovine genome

Master's thesis

30 EAP

Piibe Vill

Supervisors MSc Bariş Yaşar

and Prof Ants Kurg

TARTU 2022

Abstract

Prediction and characterization of PRD-like homeobox genes *ARGFX*, *DUXA* and *NOBOX* in the bovine genome

The conserved nature homeotic genes encode a 60 amino acid long homeodomain, which due to their high DNA-binding affinity can regulate gene expression or act as transcription factors. Recently, a selection of homeotic genes of the PRD-like class that are predicted to play a role in the embryonic genome activation, have been investigated to further understand the embryo development. In this thesis, a selection of such uncharacterized genes was investigated *in silico* for orthologs in the bovine genome. To validate the genes, bovine oocytes and embryos were investigated. cDNA was prepared with STRT-N method developed specifically for the detection of transcript far 5'-ends. As a result of *in silico* experiments, orthologs of *ARGFX*, *DUXA* and *NOBOX* were found in the bovine genome. The groundwork in the wet lab for the validation of these genes in the bovine genome was completed successfully.

Keywords: embryonic genome activation, PRD-like homeobox genes, transcription far 5'-ends

CERCS: B350 Development biology, growth (animal), ontogeny, embryology

B110 Bioinformatics, medical informatics, biomathematics, biometrics

Homeootiliste PRD-sarnase klassi geenide *ARGFX*, *DUXA* ja *NOBOX* ennustamine ja iseloomustamine veise genoomis

Konserveerunud homeootilised geenid kodeerivad 60 aminohappe pikkust homeodomeeni, mille võime seonduda DNA-ga tähendab, et sünteesitud valgud võivad käituda transkriptsioonifaktoritena või mängida rolli geeni ekspressioonis. Viimase aastakümne jooksul on avastatud, et sellistel geenidel on roll ka varajases embrüonaalses arengus. Käesolevas magistritöös uuriti valimit inimese homeootiliste PRD-sarnase klassi geenidest, et leida ortoloogseid vasteid veise genoomis. Järgmiseks eesmärgiks püstitati *in silico* kinnitust leidnud ortoloogid veise genoomis annoteerida. Selleks sünteesiti veise ootsüütidest ja embrüotest STRT-N meetodil komplenetaarne DNA, et uurida mRNA 5' otste transkriptsiooni. *In silico* analüüsi tulemusena leiti *ARGFX*, *DUXA* ja *NOBOX* geenide ortoloogid veise genoomis. Edasise analüüsina teostati eeltöö geenide annoteerimiseks veise genoomis.

Märksõnad: embrüo genoomi aktivatsioon, homeootilised PRD-sarnase klassi geenid, mRNA 5' otste transkriptsioon

CERCS: B350 Arengubioloogia, loomade kasv, ontogenees, embrüoloogia

B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

TABLE OF CONTENTS

ABBREVIATIONS

ARGFX	arginine-fifty homeobox; *ARGFXP1* and *ARGFXP2* are the derived pseudogenes
BAM	binary alignment map
bp	base pair
cDNA	complementary DNA
CPHX1	cytoplasmic polyadenylated homeobox 1
CPHX2	cytoplasmic polyadenylated homeobox 2
dNTP	deoxyribonucleoside triphosphate
DPRX	divergent paired-related homeobox
DUXA	double homeobox A
DUXB	double homeobox B
DUXC	double homeobox C
Duxbl	double homeobox B-like
EGA	embryonic genome activation
EH1	engrailed homology 1
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GV	germinal vesicle
IVF	*in vitro* fertilization
LB	lysogeny broth
LEUTX	leucine twenty homeobox
MII	metaphase II
NOBOX	newborn ovary homeobox gene
nt	nucleotide
oligo-dT	single-stranded sequence of deoxythymine
ORF	open reading frame
Pfam	protein families database
PRD	paired

STRT	single-cell tagged reverse transcriptase
TF	transcription factor
TFE	transcription far 5’-end
TPM	transcripts per million
TPRX1	tetra-peptide repeat homeobox 1
TPRX2	tetra-peptide repeat homeobox 2; *TPRX2P* is the derived pseudogene
TPRX3	tetra-peptide repeat homeobox 3
-RT	minus reverse transcriptase

INTRODUCTION

The increased knowledge of human embryology, more specifically, the preimplantation development, has been a major determinant in the success of pregnancy through *in vitro* fertilization (IVF), however to this day only 30-50% of IVF embryos reach the blastocyst stage (Gardner *et al*., 2000; French *et al*., 2010; Speyer *et al*., 2019). Following fertilization, the mammalian development proceeds with the characteristic embryonic genome activation (EGA), during which the maternal transcripts are replaced by transcripts expressed from the zygotic/embryonic genome (Dobson *et al*., 2004).

Due to the high DNA-binding affinity of the paired-like (PRD-like) class homeodomain sequences (Gehring *et al*., 1994), the transcription factors (TF) have been found to regulate the EGA in humans (Töhönen *et al*., 2015). Recently, nine novel TFs containing full length open reading frames (ORF) with homeodomains, were found to be uncharacterized, with some having no evidence of transcript expression (Madissoon *et al*., 2016). Acquiring human embryos is often challenging and includes ethical limitations that do not apply to closely related organisms such as mouse, cow and pig. Considering the comparison of oocyte size at maturation, time to maturation, early embryo development, and EGA between human oocyte and the suggested model organisms, bovine has the most similarities and is therefore a preferred choice for model organism in those studies (Santos *et al*., 2014).

The purpose of this master's thesis is to investigate a selection of PRD-like homeobox genes for orthologs in the bovine organism by first predicting them *in silico*, and later confirming them by sequencing bovine cDNA libraries amplified with *in silico* designed primers. The work contributes to and is part of a project in the same department (Yaşar, manuscript in preparation).

This master's thesis has been written within the MSc Molecular Biosciences curriculum. The experimental work in the thesis was performed in the Department of Biotechnology, at the Institute of Cell and Molecular Biology.

LITERATURE OVERVIEW

1.1 Bovine embryo development

1.1.1 Bovine oocyte maturation and fertilization

Inside the developing ovaries, the germ cells enter the prophase of the first meiotic division, where they will eventually support reductive divisions. For bovine organisms, the germ cells remain arrested and dormant in that stage for several years until the onset of ovulation (Gordon, 2013). Before reactivation, at around the time of birth, the somatic cells of the ovaries enclose the oocytes, which results in the formation of follicles. A layer of somatic pre-granulosa cells surrounds the oocytes, forming primordial follicles (Grive and Freiman, 2015). The oocytes still contain a large intact nucleus, the germinal vesicle (GV), regardless of having entered the meiosis stage (Masui and Clarke, 1979). For the oocytes to reach meiotic and developmental competence, changes take place within the oocytes and the surrounding somatic cells. The follicle develops into primary, secondary, and antral follicles. During the secondary follicle stage, the activation of oocyte transcription takes place. Simultaneously, the size of the oocyte increases to approximately 100-110 μm while gradually attaining full developmental competence (Gordon, 2013).

An increase in the concentration of follicle-stimulating hormone after puberty supervenes on the growing follicles, forming non-ovulatory follicles in multiple waves. In bovine oocytes, this later results in the formation of one ovulatory dominant follicle. The following surge of pituitary luteinizing hormone has a triggering impact on the mural granulosa cells that will initiate the meiotic resumption of the oocyte. The oocytes undergo nuclear maturation to the metaphase II (MII) phase, where they stay arrested and are released through the process of ovulation (De Felici *et al*., 2005).

1.1.2 Embryo development

The transition from oocyte to zygote happens upon the union of spermatozoid and the oocyte in a process called fertilization. The plasma membranes of the two fuse together once the spermatozoid breaks through the zona pellucida surrounding the oocyte (Baker and Polge, 1976). After the fertilization in oviduct, the zygote undergoes the first mitotic cleavage divisions while it is moving towards the uterus. The first mitotic cleavage happens approximately 30 hours after fertilization. The embryo reaches uterus at 16-cell stage, approximately on the fourth day of pregnancy. The 16-cell stage formula is called a morula, which will undergo more divisions and is renamed a blastocyst by the seventh day of pregnancy. The inner cell mass and trophectoderm can be distinguished in blastocysts (Hamilton and Laing,

1946). Following the embryo elongation, the implantation process is activated with the start of placentation at around day 21 (Bai *et al*., 2013).

1.2 Embryonic genome activation

Following the fertilization, the embryo undergoes significant developmental processes concordantly with the mitotic cleavages in the pre-implantation period. In this thesis, the focus is on the EGA, also known as zygotic gene activation or maternal-to-zygotic transition (Schultz *et al*., 1999). It describes the process of maternal transcripts being replaced with transcripts expressed by the embryonic genome. It can be described in three steps: maternal transcripts being depleted by degradation and translation; ribosomal RNAs or embryonic transcripts replacing the maternal transcripts stored in oocytes; and thirdly, generating the new embryo-specific transcripts (Sirard, 2010). The process of EGA is a crucial step in the further development of the embryo and, in the case of unsuccessful initiation, the embryo would be arrested at the stage of one mitotic cleavage (Schultz *et al*., 1999).

Mouse gene expression studies have revealed that the process happens in three successive waves: minor EGA, major EGA, and mid-preimplantation gene activation (Wang and Dey, 2006). The gene expression studies also revealed that the early expressed genes were involved in processes such as cell proliferation, DNA and protein metabolism, mitotic cell cycle and regulation of transcription (Kanka *et al*., 2012). The onset of EGA is regulated by conserved mechanisms in mammals, with minor differences such as the timing of the major EGA. The two main reasons for the differences are the species-specific cell chronology and therefore the localization of RNA polymerase II in the nuclei (Graf *et al*., 2014).

1.2.1 Bovine EGA

Experiments utilizing [^{3}H]uridine incorporation after short incubation into nuclei and nucleoli have indicated that the major EGA in bovine embryos takes place at the eight- to 16-cell stage (Camous *et al*., 1986). Another technique used to study bovine EGA gave further evidence on the onset of EGA in bovine occurring at the eight-cell stage by looking at the *in vitro* embryos treated with α-amanitin, more specifically their polypeptide profiles (Barnes and First, 1991). However, the EGA should be considered as a temporal process rather than a developmental stage. This has been confirmed by experiments with one-cell embryos transfected with reporter genes producing detectable product already at the one-cell stage (Delouis *et al*., 1992). With the combination of fluorescence *in situ* hybridization and silver staining, the first ribosomal RNA was visualized in the four-cell stage in bovine embryos (Viuff *et al*., 1998).

1.2.2 Bovine transcript abundance during EGA by microarray methods

Studies done on gene expression profiles by qualitative or quantitative reverse transcription PCR assays revealed two major expression patterns, one indicating maternal and embryonic activity with levels of expression before and after the switch to embryonic transcription and the other indicating just the embryonic activity with expression levels starting after the switch (Niemann and Wrenzycki, 2000).

One of many approaches to studying the transcript abundance of bovine embryos is to look at the expression profiles over several developmental stages. This has been done by using a custom developmental microarray that contained 1153 transcripts from four different libraries. The different developmental stages analysed were GV oocytes, two- and eight-cell stage embryos, and blastocysts. A small fraction of the genes expressed during early embryonic development could be divided into three groups based on the study conducted. The first cluster of about 72% of the genes analysed were described as genes with the highest transcript abundance in GV oocytes and a lower transcript abundance in embryos with similar expression in all stages. The second cluster which describes 17% of the genes analysed included genes with similar developmental expression as in the first cluster, however a further decreased transcript abundance in blastocysts. The third cluster covering about 10% of the analysed genes were described as of similar developmental expression as in the first cluster, but an increased transcript abundance in blastocysts (Vallée *et al*., 2009).

1.2.3 Bovine transcript abundance during EGA by RNA-seq

Due to the limitations raised by the respective probe sets, the microarray methods could not provide a detailed insight into the timing of the EGA in bovine embryos (reviewed in Graf *et al*., 2014).

In a study trying to overcome this limitation, the RNA-seq data of cross breeding of *Bos taurus taurus* and *Bos taurus indicus* was analysed for the *de novo* transcribed RNAs. This allowed the mapping of the maternal to embryonic transition by differentiating between the maternal and embryonic transcripts of approximately 7400 genes. Oocytes from the GV and MII stages, and embryos from four-, eight-, and 16-cell, and blastocyst stages were pooled, their cDNA synthesized, and with the combination of primers and single-stranded sequence of deoxythymine (oligo-dT), the whole transcriptome was covered. The study revealed the least variation in transcript abundance between the GV and MII stages of oocytes. Eight genes were found to be first expressed at the four-cell stage, 129 genes, 36 genes and 47 genes at the eight-cell, 16-cell, and the blastocyst stages, respectively (Graf *et al*., 2014).

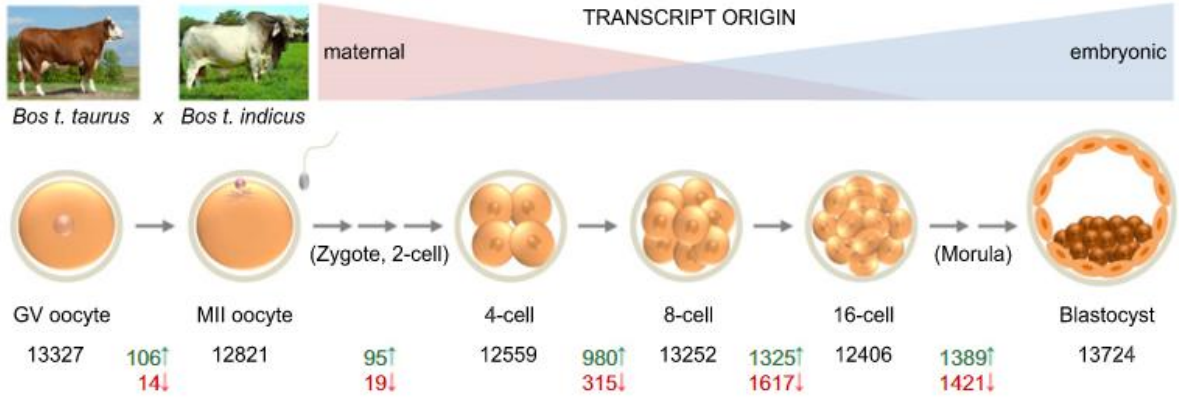


Figure 1. The experimental design of RNA-seq analysis of cross breeding of *Bos taurus taurus* and *Bos taurus indicus* at six different developmental stages. The number of genes with detectable transcripts in each stage is indicated in black, the number of transcripts with increase in abundance is indicated in green and the number of transcripts with decrease in abundance is indicated in red. (Modified from Graf *et al*., 2014.)

1.3 Homeobox genes

The specification of the body plan and regulation of the development of higher organisms are the responsibility of master control genes, also known as homeotic genes. They were first discovered in 1984 and have since been found to play a crucial role in the developmental processes of all multicellular eukaryotes. The homeobox, which is a 180 base pairs (bp) long common sequence element in the homeotic genes, was discovered to encode the about 60 amino acid long homeodomain. The term "homeo" is used in reference to the homeotic genes, which encode the homeodomain proteins of *Drosophila melanogaster* (McGinnis *et al*., 1984; Scott and Weiner, 1984).

The homeodomain's importance lies in its ability to bind to DNA, more specifically it allows the sequence-specific recognition of homeobox genes. Further discoveries have since been made about additional motifs and variants of the homeodomain with insertions in the domain, but the main structural features are found to be conserved (Bürglin, 2011). The tryptophan (W) and the asparagine (N) residues at positions 48 and 51, respectively, are almost constant in all homeodomain sequences. A few other conserved residues are known to have a substitute only occasionally. The underlying reason for the conserved amino acid residues at specific positions is the structure of the homeodomain. Nuclear magnetic resonance and X-ray studies revealed that the homeodomains consist of three alpha-helices (Billeter *et al*., 1993). The core of hydrophobic amino acids that holds the molecular architecture of a homeodomain together leads to the belief that different homeodomains share a similar structure (Gehring *et al*., 1994).

1.3.1 Homeobox gene complexes

Based on several criteria, such as sequence identity and similarity in the flanking regions, organization into gene clusters, intron positions, and association with other sequence motifs, the more than a thousand identified homeodomain sequences have been divided into different superclasses, which can be further divided into classes and then families (Bürglin, 2011). The homeodomains of genes that are clustered in a known homeotic gene complex make up the complex superclass. The homeodomains of genes dispersed largely in the genome make up the dispersed superclass. The first superclass has been further divided into six classes, named after the corresponding *Drosophila* genes (*lab, pb, Dfd, Scr, Antp, Ubx* and *AbdB*). The second superclass, dispersed, can be further divided into more than 16 classes by the over 22 different positions of residues (Gehring *et al*., 1994).

1.3.1.1 PRD class

One of the 16 classes in the dispersed superclass is the "evolutionary tinkering" PRD class, which can be subdivided into PAX and PAXL subclasses (Gehring *et al*., 1994). The PRD class genes encode a characterizing serine (S) residue at the 50th position (Burri *et al*., 1989). For the PRD homeodomain genes, a second 128 amino acid long DNA-binding domain, also known as Pax domain, can be located. The second DNA-binding domain means a second binding site, which strongly increases the DNA-binding specificity. For some PRD class homeobox genes, an additional engrailed homology 1 (EH1) motif has been located between the homeodomain and the paired domain (Burri *et al*., 1989), whereas for some, a functionally important truncated homeodomain has been identified. The evolutionary tinkering description comes from the different recombining of the domains and the various motifs (Gehring *et al*., 1994). The common serine residue has led to the belief that the association between the homeodomain and the paired domain has occurred early in evolution (Bürglin, 2011).

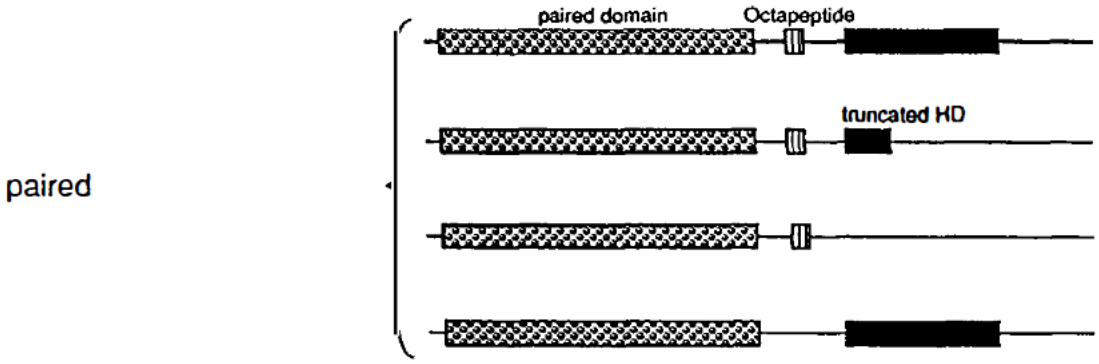


Figure 2. Schematic representation of PRD class homeodomain protein domains. (Modified from Gehring *et al*., 1994.)

1.3.1.2 PRD-like class

The homeobox genes that encode a homeodomain similar to the PRD class but do not encode a paired domain, have been named the PRD-like class. They also lack the characteristic serine residue at position 50. Some PRD-like class genes also encode the EH1 motif near the N-terminus. Around the C-terminus, another characteristic OAR motif (named after the initials of *Otp*, *Aristaless* and *Rax* genes) is located for the PRD-like class genes. The OAR motif has also been found in PRD class of genes and is considered to play a role in the transcriptional activation (Vorobyov and Horst, 2006). The PRD-like homeobox genes are generally dispersed in the genome, except for the HRO gene cluster of *Homeobrain*, *Rax* and O*rthopedia* genes (Mazza *et al*., 2010).

Some of the PRD-like class genes have been found to encode a characteristic lysine (K) or glutamine (Q) at position 50. The following families of the PRD-like class have been found in humans: Alx, Argfx, Arx, Dmbx, Dprx, Drgx, Dux, Esx, Gsc, Hesx, Hopx, Isx, Leutx, Mix, Nobox, Otp, Otx, Phox, Pitx, Prop, Prrx, Rax, Rhox, Sebox, Shox, Tprx, Uncx, and Vsx (reviewed in Bürglin, 2011).

1.4 PRD-like homeobox gene expression during EGA

The expression of PRD-like homeobox genes can be studied by single-cell RNA sequencing. The problems of this method, such as the variation of sequencing depth and aligned read counts, and bias from technical sources, can be corrected with a method of tolerance for both issues. SAMseq (Li and Tibshirani, 2011) has been adapted to the broadness of the differential expression in STRT transcriptome profiles. The adapted statistical test is called SAMstrt (Katayama *et al*., 2013).

In a study, where this normalization method was used to count the poly(A)-tailed RNA molecules, 32 TFEs were found to be significantly upregulated in the early EGA (from the oocyte to the four-cell stage) and 129 TFEs were found to be significantly upregulated in the major EGA (from the four-cell to the eight-cell stage). To assess the presence of potential regulatory elements, the sequence motifs around the upregulated TFEs in early EGA were extracted. Within the Alu elements, a significant 36 bp long *de novo* motif was identified and found to be of similar sequence to the binding sites for bZIP, T-box and PRD-like homeodomain containing TFs. A similar 35 bp long *de novo* motif was identified in the upregulated TFEs in major EGA (Töhönen *et al*., 2015).

The PRD-like homeobox genes were further studied considering that discovery. 18 homeobox genes were found to be significantly expressed in either wave of EGA, with 14 of them having TFEs of functional transcripts. Among these 14 were genes associated with maternal factors, embryonically activated, which tend to be less conserved than the first, and a mixture between the two (Töhönen *et al*., 2015).

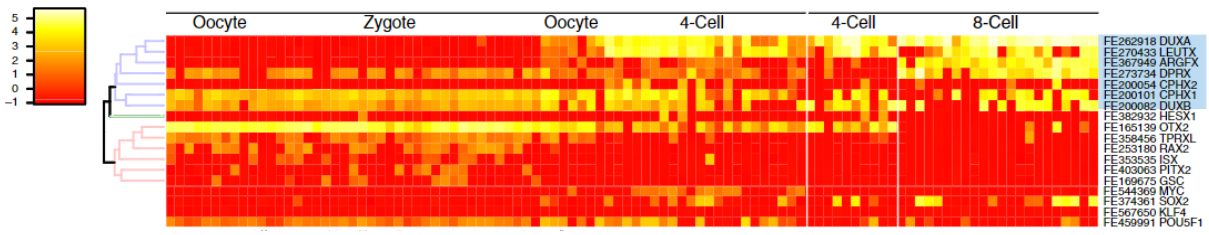


Figure 3. Expression pattern of novel PRD-like homeobox genes in the blue shading. The expression patterns of Yamanaka factors *MYC*, *SOX2*, *KLF4* and *POU5F1/OCT4* is shown as comparison. (Modified from Töhönen *et al*., 2015.)

Start sites of seven novel TFs that were poorly annotated in the public databases, were verified. The predicted TFs all produce full length ORFs that have complete homeodomain sequences (Töhönen *et al*., 2015). The seven PRD-like homeobox genes as well as *DUXC*, *Duxbl*, *NOBOX*, *TPRX1*, *TPRX2* and *TPRX3* are described below.

1.4.1 *ARGFX*

Arginine-fifty homeobox (*ARGFX*) gene has the very rare arginine residue at the position 50 in the homeodomain sequence (Booth and Holland, 2007). *ARGFX* is believed to derive from duplication and extensive sequence divergence of the OTX-family *CRX* gene (Maeso *et al*., 2016). The gene has two intronless pseudogenes, *ARGFXP1* and *ARGFXP2*, which are predicted to have been generated because of retrotransposition of *ARGFX*. There is no human *ARGFX* gene homologue in the mouse genome (Booth and Holland, 2007).

1.4.2 *CPHX1* and *CPHX2*

The protein alignment between cytoplasmic polyadenylated homeobox 1 (*CPHX1*) and cytoplasmic polyadenylated homeobox 2 (*CPHX2*) genes suggests a highly conserved sequence even outside of the homeodomains. The human *CPHX1* and *CPHX2* being in syntenic locations suggests the orthology with the *Cphx* homeobox loci (Zhong and Holland, 2011), which was first described in mouse (Li *et al*., 2006).

1.4.3 *DPRX*

The low similarity divergent paired-related homeobox (*DPRX*) gene has been designated its name due to the low level of sequency identity with the PRD class homeodomains (Booth and Holland, 2007). The gene is also believed to have arisen from the duplication and extensive divergence of the *CRX* gene. Seven pseudogenes of *DPRX* have been identified from the retrotransposed mRNA (Holland *et al*., 2007). No homologue of human *DPRX* has been found in the mouse genome (Booth and Holland, 2007). *DPRX* is predicted to be a pseudogene in the bovine genome (Maeso *et al*., 2016).

1.4.4 *DUXA, DUXB, DUXC* and *Duxbl*

Homeobox genes double homeobox A (*DUXA*) and double homeobox B (*DUXB*) genes belong to the Dux gene family with two closely linked homeobox motifs (Holland *et al*., 2007). Before the identification of *DUXA* gene, the five Dux family genes were named with numbers. Logically, the *DUXA* should have been named *DUX6*, however the gene differs from the rest of the sequences by having introns, even within the homeobox. The ten *DUXA* pseudogenes are thought to be the result of a tandem duplication after retrotransposition of mRNA transcript (Booth and Holland, 2007). The *DUXA*, *DUXB*, double homeobox C (*DUXC*) and double homeobox B like (*Duxbl*) genes' possession of introns has suggested the hypothesis of any of the four being the progenitor of the majority of the other intronless gene sequences in the family (Leidenroth and Hewitt, 2010). *DUXA* and *DUXB* orthologs have been found in most major mammalian lineages (Clapp *et al*., 2007). Based on the phylogenetic analysis of *DUXB* gene, the ortholog is predicted to be a pseudogene in the bovine genome. The *Duxbl* ortholog in the bovine genome is intact, however with an unclear synteny. The *DUXC* gene ortholog is not predicted in the bovine genome (Leidenroth and Hewitt, 2010).

1.4.5 *LEUTX*

Leucine twenty homeobox (*LEUTX*) gene derives its name from the highly conserved leucine (L) residue at the homeodomain sequence position 20, while the rest of the PRD class homeodomain sequences have a phenylalanine (F) at the same position (Holland *et al*., 2007). Phylogenetic analysis including the amino acid stretches from C-terminus to the homeodomain

revealed the progenitor of *LEUTX* being the OTX-family member *CRX* gene in a close genomic proximity (Maeso *et al*., 2016). *LEUTX* is found not to be present in invertebrates while divergent orthologs of the sequence have been found in other primates (Holland, 2012; Zhong and Holland, 2011). The mouse genome also does not have an ortholog of human *LEUTX* gene (Jouhilahti *et al*., 2016).

1.4.6 *NOBOX*

Newborn ovary homeobox (*NOBOX*) gene has a 55% sequence identity with the PRD-like class homeodomains (Holland *et al*., 2007). The expression pattern of human *NOBOX* gene suggests its crucial role in oogenesis (Suzumori *et al*., 2002). An amino acid change at position 52 of the previously well characterized *NOBOX* gene is known to cause premature ovarian failure (Qin *et al*., 2007). An ortholog of the human *NOBOX* gene has been identified in the mouse genome (Suzumori *et al*., 2002).

1.4.7 *TPRX1, TPRX2* and *TPRX3*

Tetra-peptide repeat homeobox 1 (*TPRX1*) and tetra-peptide repeat homeobox 2 (*TPRX2*) genes are, similarly to the *ARGFX*, *DPRX* and *LEUTX* genes, arisen from the duplication and extensive divergence from the *CRX* gene (Holland *et al*., 2007). *TPRX2* was designated *TPRX2P* for being a duplicated copy of *TPRX1* on the other side of the *CRX* gene and therefore thought to be a pseudogene (Booth and Holland, 2007). It has now been shown to be expressed in human and therefore renamed as *TPRX2* (Madissoon *et al*., 2016). The *TPRX1* and *TPRX2* human gene orthologs are especially divergent in the mouse genome (Maeso *et al*., 2016). The bovine genome is predicted to possess a duplicate of the *TPRX* gene, called tetra-peptide repeat homeobox 3 (*TPRX3*), which is believed to be clustered together with genes *LEUTX*, *TPRX1*, *TPRX2* and *DPRX* on the 18th chromosome (Lewin *et al*., 2021).

1.5 Single-cell RNA-seq

RNA sequencing has significant advantages over microarrays for analysing gene expression profiles. RNA sequencing is a hypothesis-free analysis method that overcomes the issues of limited sensitivity and dynamic range by sequencing RNA directly (Ozsolak *et al*., 2009). This leaves the heterogenous property of tissue cells that cannot be overridden with RNA sequencing. Studying, for example, the expression profiles in embryology using pools of cells would blend the true expression profile due to the extensive heterogeneity in different low cell count tissues and the accessibility of these cells. To overcome these issues, analysis could be done on a single-cell basis and recently developed single-cell methods (Tang *et al*., 2009) have enabled that. Issues such as markers not being available for every cell type and the variety of

transcript abundance still need attention regarding the single-cell RNA sequencing. Experiments with large numbers of single cells have caused the limitation of qPCR to small number of genes to surface (Guo *et al*., 2010). Therefore, there is a need for a method for analysing the entire transcriptomes of single cells in large numbers.

1.5.1 STRT method

Studying the expression profiles of bovine embryo transcript abundance is tricky due to the transcription happening in random bursts caused by cell division induced mRNA decay and dilution. That and the cell-to-cell variation can be attenuated by analysing large number of single cells by using a method such as STRT. STRT is a multiplex-tagged method for single-cell poly(A)-tailed RNA sequencing that enables the detection of the TFEs. The method is specifically designed for the Illumina platform (Islam *et al*., 2011).

In the method, reverse transcription reagents are added to each well of the plate containing a single cell, and mRNA is converted to cDNA. The template-switching mechanism where the helper oligo is responsible for directing the incorporation of a specific sequence at the 3' end of the cDNA molecule has been exploited (Schmidt and Mueller, 1999). Instead, a helper oligo with distinct six-base barcode and a universal primer sequence is used. The reverse transcriptase switches templates and continues synthesizing a copy of the helper oligo. After the modification, cDNA is amplified. The amplification is achieved by PCR or by several rounds of *in vitro* transcription (Islam *et al*., 2011).

The STRT method enables the sequencing of 5'-end of each cDNA too, which has several advantages, including the distinction of overlapping genes transcribed from opposite strands and the identification of the transcription start site. Each mRNA molecule results in a single cDNA molecule and therefore the number of observed reads is proportional to the number of mRNA molecules. Another advantage of the method is the introduction of barcodes for multiplexing (Islam *et al*., 2012). STRT-N method currently under development is modified by the timing of adding the bar codes (Boskovic).

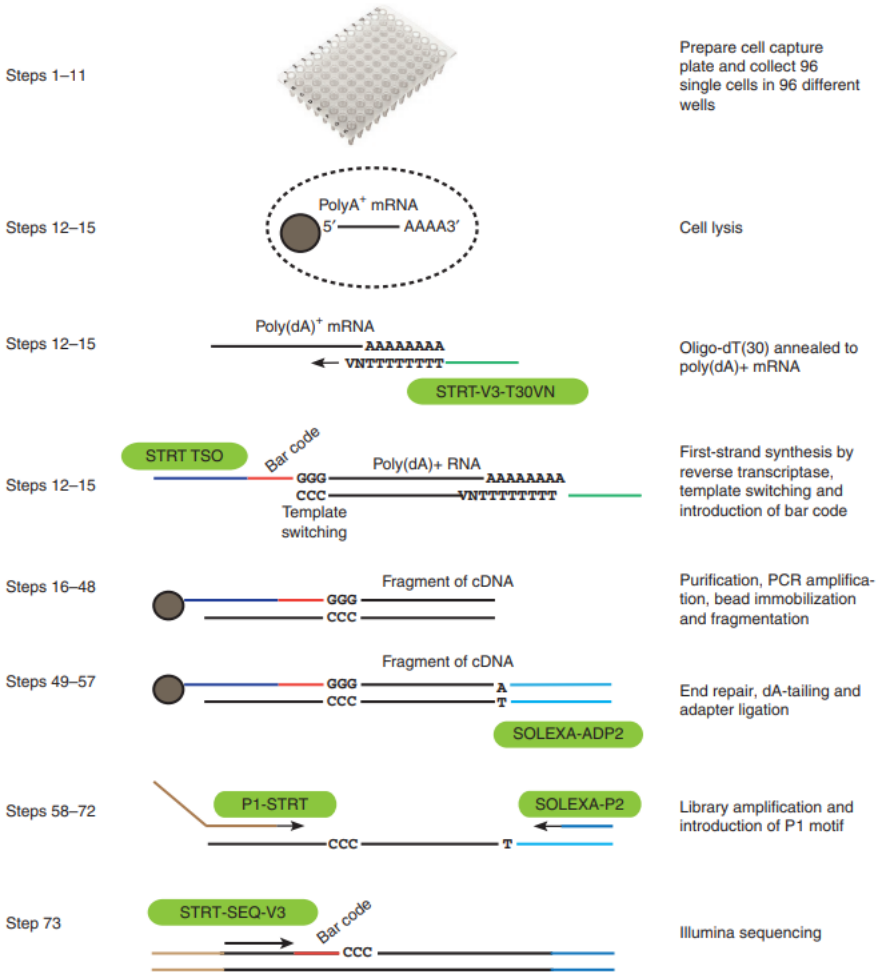


Figure 4. Schematic overview of the STRT method. Single cells are added to each well of the plate, one cell per well. The lysis buffer specific to the STRT method breaks down the cell membrane, producing poly(A)-tailed mRNAs. The next steps of oligo-dT annealing, synthesis by reverse transcriptase and the template switching activity describe the mechanisms of reverse transcription reaction. This is followed by cDNA amplification, which is done by PCR. **(**Modified from Islam *et al*., 2012.)

2. EXPERIMENTAL WORK

2.1 Aims of the thesis

The main aim of this thesis was to investigate a selection of human PRD-like homeobox genes for orthologs in the bovine genome and prepare cDNA libraries of bovine oocyte(s) and embryos(s) for the annotation of the predicted genes in the bovine genome.

The specific aims of the thesis were:

1. to re-analyse RNA-seq data (Graf *et al.*, 2014) to find orthologs of human PRD-like homeobox *ARGFX***,** *CPHX1*, *CPHX2*, *DPRX*, *DUXA*, *DUXB*, *LEUTX*, *NOBOX*, *TPRX1* and *TPRX2* genes in the bovine genome
2. to predict the exon stretches of the putative gene orthologs in the bovine genome
3. to prepare cDNA from different developmental stage bovine oocytes and embryos with STRT-N method, followed by cloning and Sanger sequencing

2.2 Materials and methods

2.2.1 Re-analysis of RNA-seq data derived from oocytes and preimplantation IVF embryos

The raw FASTQ files and the normalized read counts per gene available at Gene Expression Omnibus (series GSE52415) from *in vitro* fertilization of *Bos taurus taurus* oocytes and *Bos taurus indicus* bull sperm for parent-specific transcriptome analysis (Graf *et al*., 2014) were re-analysed. Platform Galaxy (Afgan *et al*., 2018) was used to conduct the re-analysis. Tool "Faster Download and Extract Reads in FASTQ" (Afgan *et al*., 2018) was used to download the dataset. The three replicates for each six developmental stages (GV and MII oocytes, four-, eight-, and 16-cell, and blastocyst stage embryos) were handled separately. The trimmed reads were then aligned to the latest bovine assembly genome (bosTau9) using the tool "HISAT2" (Afgan *et al*., 2018). In the advanced options of spliced alignment, the "Report alignments tailored for transcript assemblers including StringTie" was chosen. To be able to detect as many transcripts as possible, the RNA-seq read files of the same developmental stage were merged into one binary alignment map (BAM) file using the tool "Samtools merge" (Afgan *et al*., 2018). Assembling the aligned reads on the genome to infer the full-length sequence and the splicing pattern of mRNAs was done using the tool "StringTie" (Afgan *et al*., 2018). To increase the assemble speed, a reference annotation was used from UCSC Table Browser (Karolchik *et al*., 2004) as a gene transfer format (GTF) file.

As the next step, putative protein sequences of the candidates were checked to confirm appropriateness as "orthologs". This was done by creating a custom track with the assembled reads for each developmental stage to the UCSC Genome Browser (Kent *et al*., 2002). The transcripts per million (TPM) values at the loci of each gene of interest were compared and the transcripts with the highest TPM values were selected to extract the DNA sequences of the spliced transcripts. The spliced transcript sequences were searched for ORFs with the NCBI ORF finder tool (Sayers *et al*., 2022), and the putative protein sequences of the longest ORFs were searched for homeodomains with the protein families database (Pfam) tool (Mistry *et al*., 2021).

2.2.2 Bovine oocyte and preimplantation embryo collection

The oocytes and embryos for the experimental work were acquired from a collaborative laboratory at the Estonian University of Life Sciences. The Holstein Friesian cattle oocytes had been *in vitro* fertilized.

2.2.3 cDNA preparation by STRT-N method

The oocytes and the embryos were stored in lysis buffer specific to STRT-N protocol (Boskovic, manuscript in preparation). The lysed cells were assembled with reverse transcription mixture prepared according to the protocol. STRT-N reverse transcription reaction was performed on the samples and bovine fibroblast RNA control samples. For the bovine fibroblast RNA and GV stage oocytes, the reverse transcription step was also carried out in the absence of reverse transcriptase. Following the reverse transcription, cDNAs were amplified during PCR step, which was modified according to the protocol. The primer sequence length suggested in the protocol was shortened and efficiency tested on the bovine fibroblast RNA; the rest of the samples were amplified with the shorter primer sequence. The effectiveness of the STRT-N method was checked with agarose gel electrophoresis (see below for the method of agarose gel electrophoresis).

2.2.4 Product clean-up

The cDNAs were cleaned of leftover primers and deoxyribonucleoside triphosphates (dNTP) prior to running PCR with gene-specific primers. The leftover primers were segmented off from the product with enzyme exonuclease I (1U/µL, Sigma-Aldrich, Germany) and the dNTPs with Shrimp Alkaline Phosphate (20U/µL, New England Biolabs, UK), following the ExoSAP protocol (Thermo Scientific, USA).

2.2.5 Amplification of *ARGFX, DUXA* and *NOBOX* with gene-specific primers

The genes were amplified with gene-specific primers designed during the re-analysis of RNA-seq (Graf *et al.*, 2014). Protocol Phusion™ High–Fidelity DNA Polymerase (Thermo Scientific, USA) was followed for the amplification, producing high fidelity products suitable for cloning. The annealing temperature of the PCR program was modified according to the Tm calculator (Thermo Scientific, USA) for each primer. Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) was used as a housekeeping gene.

2.2.6 Agarose gel electrophoresis

To confirm the amplification of the genes of interest, the products were checked with agarose gel electrophoresis, prepared following an in-house method. A mixture of 150 mL of 0.5xTBE (TÜMRI, Estonia) and 3 g of agarose (Bioatlas, Estonia) was heated until the solution was clear. 4 µL of ethidium bromide (0.3 µg/mL) was added to the mixture so the DNA banding pattern could be visualized under UV light source. The mixture was then poured into a mould and left to solidify. Meanwhile, 4 µL of 0.6x loading dye (Thermo Scientific, USA), containing glycerol, was added to the products. Approximately 4.5 µL of ladder (Solis BioDyne, Estonia)

was added in the first well, and products in the subsequent wells. The gel was run at around 150 voltages for approximately an hour and then visualized in UV viewing cabinet (Uvitec, UK).

2.2.7 Gel extraction

Bands of highest intensity and at the expected site considering the predicted gene lengths were cut from the gel with a scalpel. The DNA fragments in the gel were purified following the GeneJET Gel Extraction Kit protocol (Thermo Scientific, USA). The recovered DNA was then ready for cloning.

2.2.8 Cloning of *ARGFX, DUXA* and *NOBOX*

The Zero Blunt® TOPO® PCR Cloning Kit (Life Technologies, USA) was used to insert the blunt-ended PCR products into plasmids for sequencing. The subsequent miniprep was done following an in-house method. Colonies from lysogeny broth (LB) plates (TÜMRI, Estonia) were picked with a pipette tip and dropped into tubes with LB medium containing ampicillin (100 μg/mL, TÜMRI, Estonia). The bacterial cells were grown by shaking horizontally at 37˚C overnight. The following morning, the cells in LB medium were transferred into new pre-labelled tubes, where they were spun down and the LB medium discarded as supernatant. The cell pellets were resuspended in 100 μL of resuspension buffer (Macherey-Nagel, Germany) containing Tris-HCl (50 mM, pH 8), EDTA (10 mM) and RNaseA (100 μg/mL). Following that, 100 μL of lysis solution (Macherey-Nagel, Germany) containing NaOH (0.2 N) and SDS (1%) was added to the resuspended pellets and the tubes were inverted five to six times. On top of that, 150 μL of neutralizing solution (Macherey-Nagel, Germany) containing potassium acetate (3 M, pH 6.0), glacial acetate and H_2O was added, after which, the tubes were immediately inverted, leaving chromosomal DNA visible as a white precipitate. The chromosomal DNA was spun down for 10 minutes and 300 μL of the supernatant containing plasmid DNA was transferred into new pre-labelled tubes. The cell pellets were washed with 240 μL of isopropanol (TÜMRI, Estonia) and centrifuged for 10 minutes to discard the supernatant. The same washing step was done with 600 μL of ethanol (TÜMRI, Estonia). To dry the pellets, tube lids were left open for approximately 30 minutes. The DNA pellets were then resuspended in 100 μL of TE (Macherey-Nagel, Germany).

2.2.9 EcoRI restriction enzyme digestion

EcoRI restriction enzyme protocol User Guide: EcoRI, 10 U/ μL, 5000U (Thermo Scientific, USA) was followed to ensure that the cloning was efficient by cutting the plasmids at the *Eco*R I sites (see Appendix 1). The results were checked with agarose gel electrophoresis.

2.2.10 Sanger sequencing

Concentrations of the DNA in TE were measured with NanoDrop (Thermo Scientific, USA). The desired concentration of 66.6 μg/μL was achieved by assembly with nuclease-free water (TÜMRI, Estonia). Samples were then sent for Sanger sequencing at the University of Tartu Sanger sequencing and genotyping laboratory. The desired quantity was 30-50 ng PCR product per reaction and 0.3 μL of 10 μM (10 pmol/mL) of primer per sequencing reaction. A form with the information was attached to the samples (see Appendix 2). The samples were sequenced with universal primers T3 and T7 (Table 1).

Table 1. T3 and T7 universal primer sequences

T3	ATTAACCCTCACTAAAG
T7	AATACGACTCACTATAG

2.3 Results

2.3.1 Re-analysis of RNA-seq data derived from oocytes and preimplantation IVF embryos

The overall alignment rate was over 63.3% for each developmental stage investigated. The highest overall alignment rate was 85.97% in the blastocyst stage.

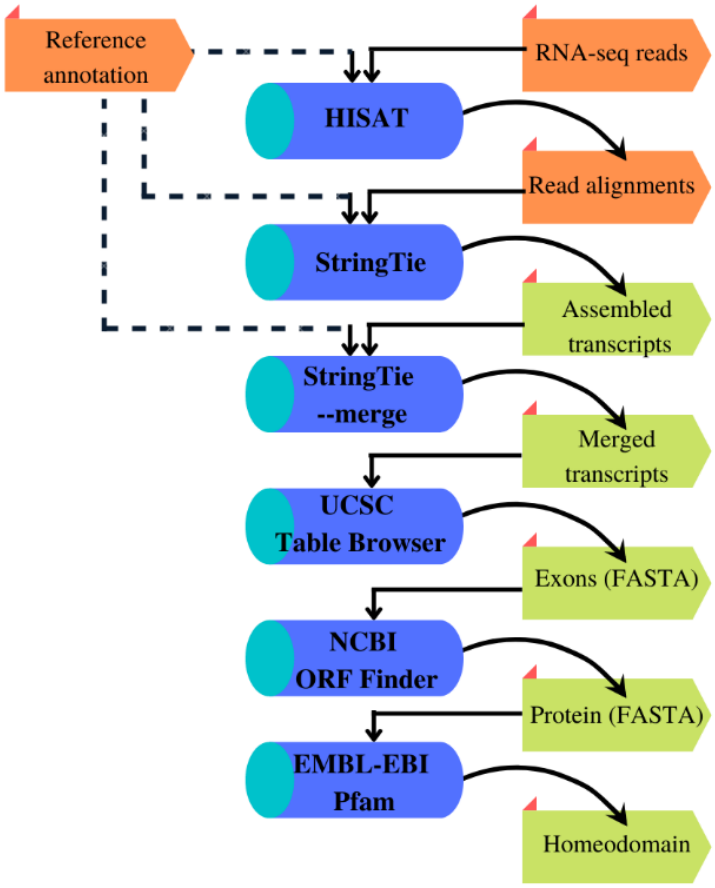


Figure 5. Flowchart of the steps of re-analysis.

2.3.1.1 *ARGFX* gene ortholog in bovine genome. The genomic region of *ARGFX* gene has a three-exon long transcript in the latest bovine assembly genome (bosTau9) with the XM_prefix (Figure 6). In comparison to predicted subset of NCBI RefSeq genes, assembled transcripts reveal two novel exons of *ARGFX*. Five-exon transcript STRG.663.1 at the 16-cell stage was chosen for further analysis due to the highest expression abundance with TPM value of 793.603149 in that developmental stage. The length of the spliced transcript of the gene is 1054 nucleotides (nt).

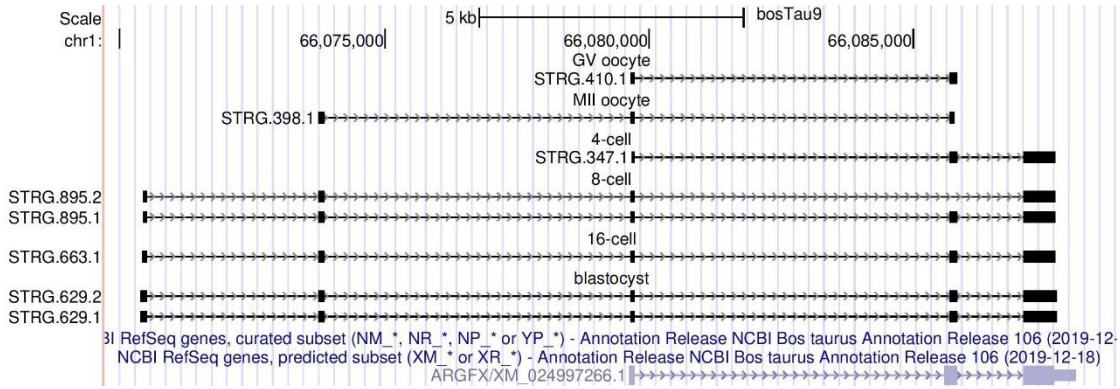


Figure 6. Trancripts of the *ARGFX* gene in the RNA-seq data derived from oocytes and preimplantation IVF embryos. The trimmed reads obtained from six different developmental stages were re-analysed and aligned to the latest bovine genome assembly. (Modified from Kent *et al*., 2002.)

NCBI's ORFfinder tool (Sayers *et al*., 2022) was used to get the putative *ARGFX* protein sequences. Subsequently, the longest ORF (Figure 7), which is likely to code for the functional protein *in vivo*, was chosen for further investigation.

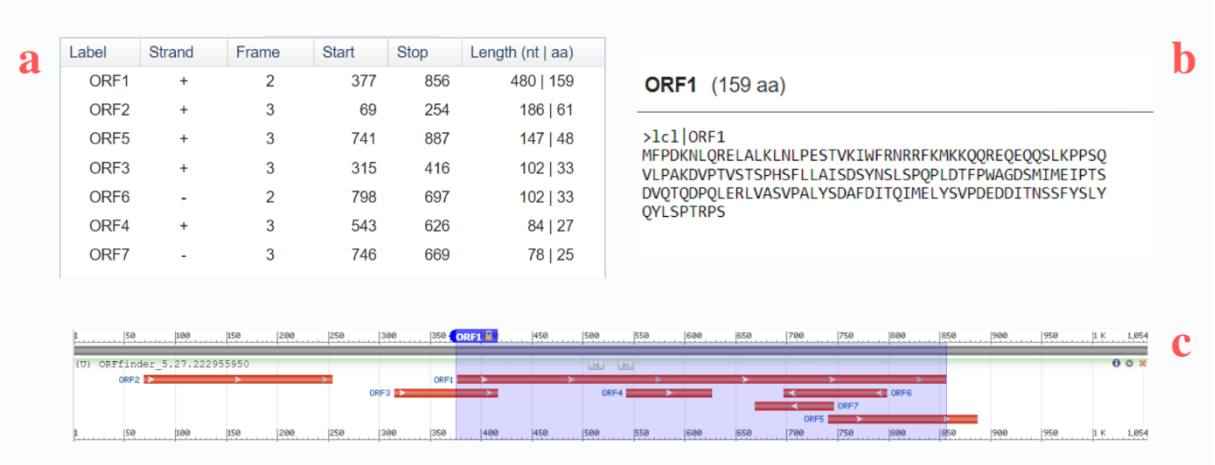


Figure 7. ORF matches of *ARGFX* spliced transcript sequence. a) All of the detected ORFs detected from longest to shortest in length; b) 159 amino acid long protein synthesis of the longest ORF detected; c) stretches of all detected ORFs, the longest in blue shading. (Modified from Sayers *et al*., 2022.)

After querying the putative protein sequence of the longest ORF in the Pfam database (Mistry *et al*., 2021), one significant match with a homeodomain could be observed (Figure 8).

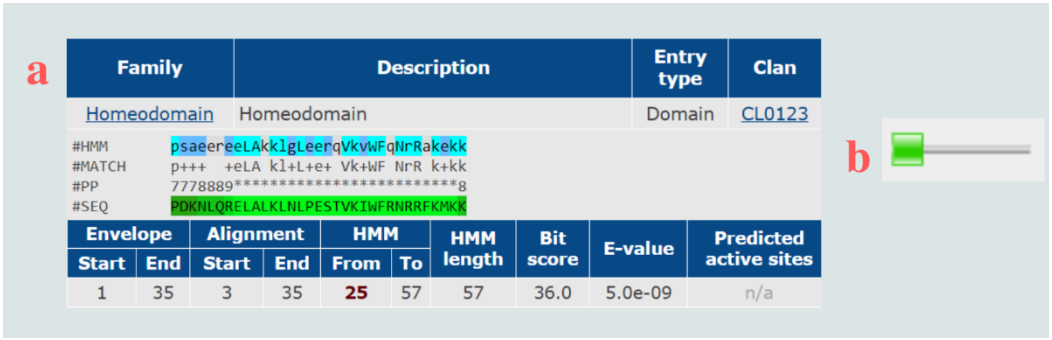


Figure 8. Putative protein synthesis of *ARGFX* match with a homeodomain. The putative protein after *in silico* translation using the longest *ARGFX* ORF was searched for the presence of homeodomains with the Pfam tool. a) The significant homeodomain match within the protein; b) depiction of the homeodomain within the putative protein. The #HMM represents the consensus of the matching HMM; #MATCH represents the match between query and the matching HMM; #PP stands for degree of confidence between each individual aligned residue; #SEQ represents the query sequence. (Modified from Mistry *et al*., 2021.)

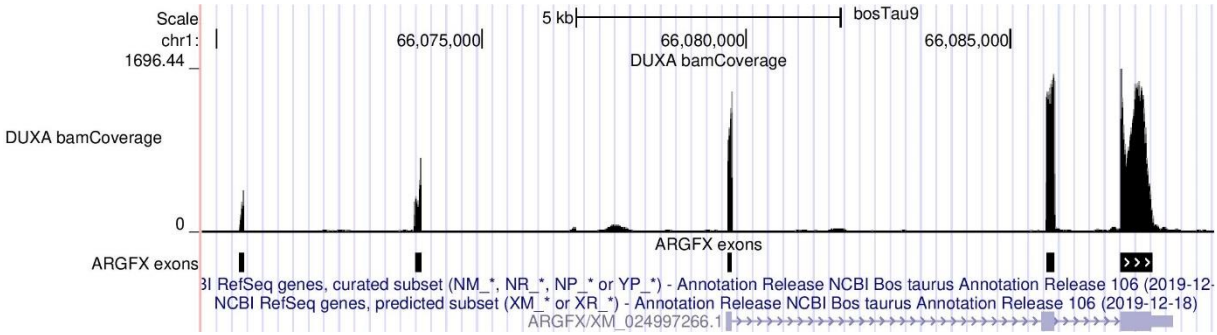


Figure 9. Illustrative scheme of exon stretches of *ARGFX* gene and the RNA-seq read coverage from 16-cell stage. (Modified from Kent *et al*., 2002.)

2.3.1.2 *DUXA* gene ortholog in bovine genome. The genomic region of *DUXA* shown in Figure 10 has an annotation of five-exon long transcript with the XM_ prefix in the latest bovine genome assembly available (bosTau9). In comparison to the predicted subset of NCBI RefSeq genes, assembled transcripts reveal a novel first exon of *DUXA*. The aligned transcripts of different developmental stages vary between five- and six-exon long transcripts.

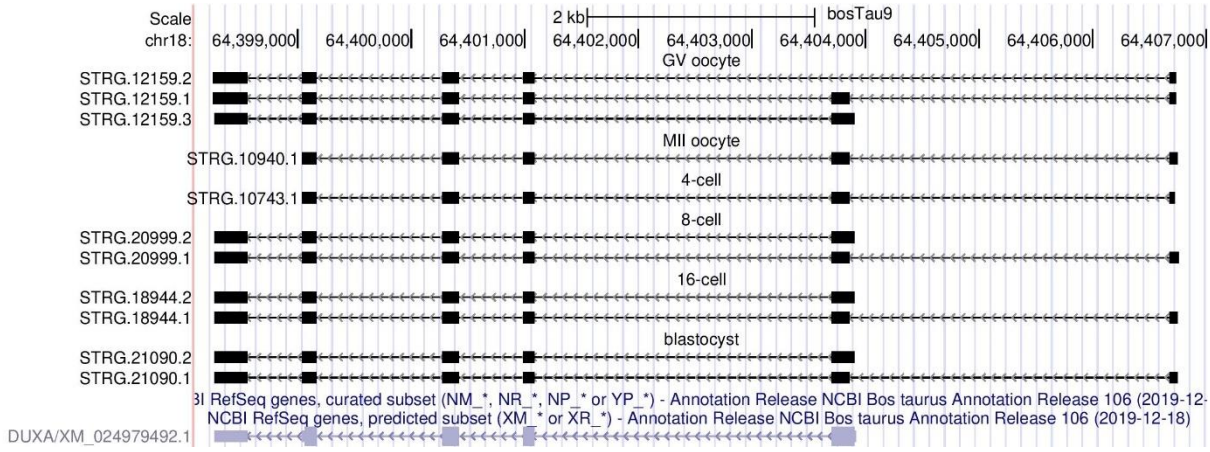


Figure 10. Trancripts of the *DUXA* gene in the RNA-seq data derived from oocytes and preimplantation IVF embryos. The trimmed reads obtained from six different developmental stages were re-analysed and aligned to the latest bovine genome assembly avilable. (Modified from Kent *et al.*, 2002.)

Six-exon long transcript STRG.18944.1 of 16-cell stage was chosen for further analysis due to the highest TPM value of 2270.774902. The length of the spliced transcript is 876 nt. NCBI's ORFfinder tool (Sayers *et al.*, 2022) was used to get the putative *DUXA* protein sequences. The longest ORF most likely to code for the functional protein *in vivo* was chosen in the same manner as for *ARGFX* for further investigation. After querying the putative protein sequence of *DUXA* in Pfam database (Mistry *et al.*, 2021), two significant matches with homeodomains could be observed (Figure 11).

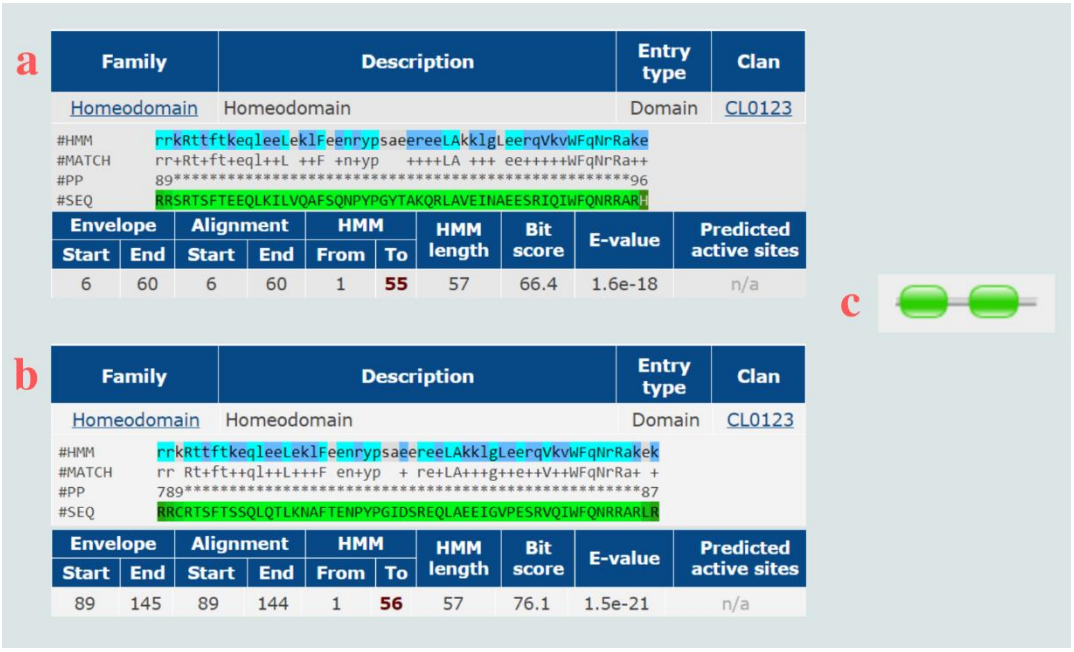


Figure 11. Putative protein synthesis of *DUXA* matches with two homeodomains. The putative protein after *in silico* translation using the longest *DUXA* ORF was searched for the presence of homeodomains with the Pfam tool. a) The first significant homeodomain match within the protein; b) the second significant homeodomain match within the protein; c) depiction of both homeodomains within the putative protein. See description of Figure 8 for the legend of abbreviations. (Modified from Mistry *et al.*, 2021.)

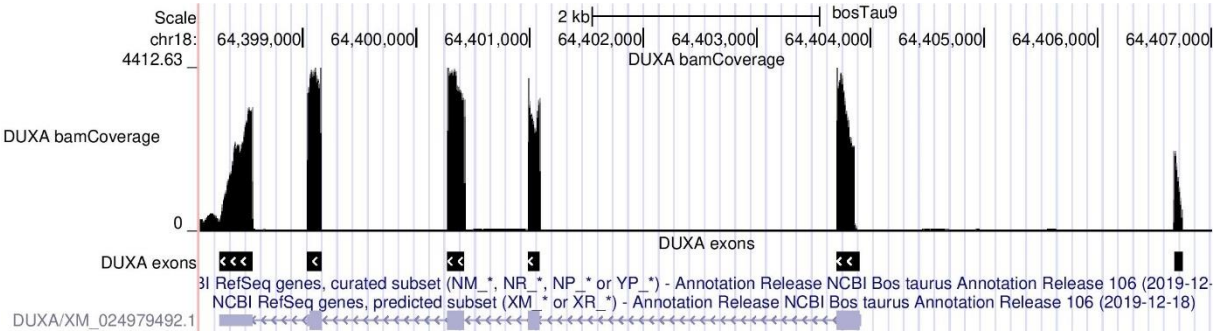


Figure 12. Illustrative scheme of exon stretches of *DUXA* gene and the RNA-seq read coverage from 16-cell stage. (Modified from Kent *et al.*, 2002.)

2.3.1.3 *NOBOX* gene ortholog in bovine genome. The genomic region of *NOBOX* gene has a six-exon long transcript in the latest bovine assembly genome (bosTau9) with the XM_prefix (Figure 13). In comparison to the predicted subset of NCBI RefSeq genes, assembled transcripts reveal two novel exons of *NOBOX*. Eight-exon long transcript STRG.26224.2 of the GV oocyte stage was chosen for further analysis due to the expression of the transcript with the highest abundance in that developmental stage with TPM value of 29.5567. The length of the spliced transcript is 2829 nt.

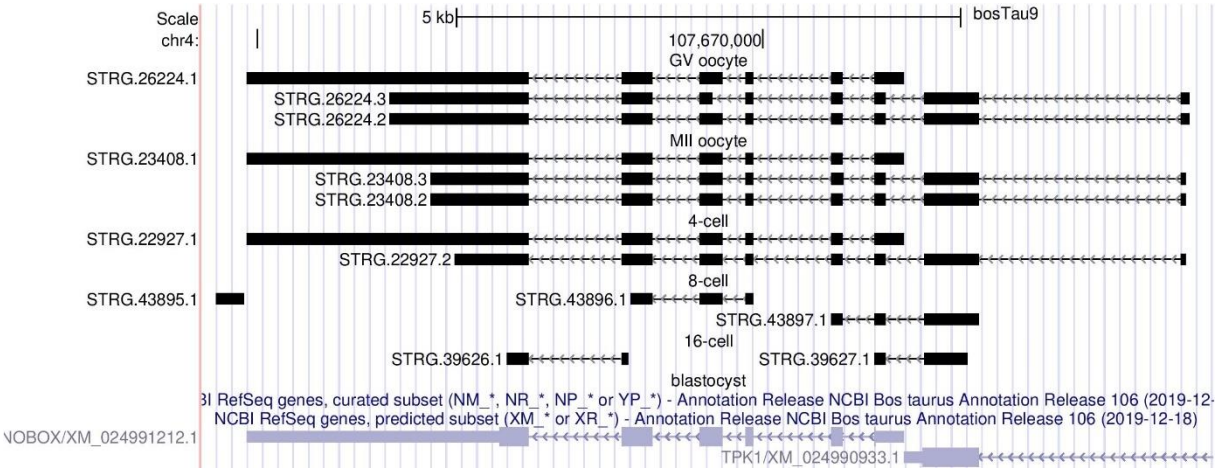


Figure 13. Trancripts of the *NOBOX* gene in the RNA-seq data derived from oocytes and preimplantation IVF embryos. The trimmed reads obtained from six different developmental stages were re-analysed and aligned to the latest bovine genome assembly avilable. (Modified from Kent *et al.*, 2002.)

NCBI's ORFfinder tool (Sayers *et al.*, 2022) was used to get the putative *NOBOX* protein sequences in the same manner as for *ARGFX* and *DUXA*. Subsequently, the longest ORF, which is likely to code for the functional protein *in vivo*, was chosen for further investigation. After querying the putative protein sequence of the longest ORF in the Pfam database (Mistry *et al.*, 2021), one significant match with a homeodomain could be observed (Figure 14).

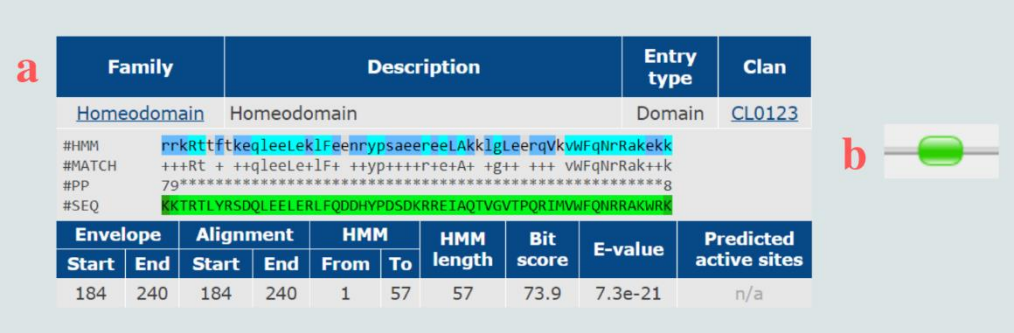


Figure 14. Putative protein synthesis of *NOBOX* match with a homeodomain. The putative protein after *in silico* translation using the longest *NOBOX* ORF was searched for the presence of homeodomains with the Pfam tool. a) The significant homeodomain match within the protein; b) depiction of the homeodomain within the putative protein. See description of Figure 8 for the legend of abbreviations. (Modified from Mistry *et al.*, 2021.)

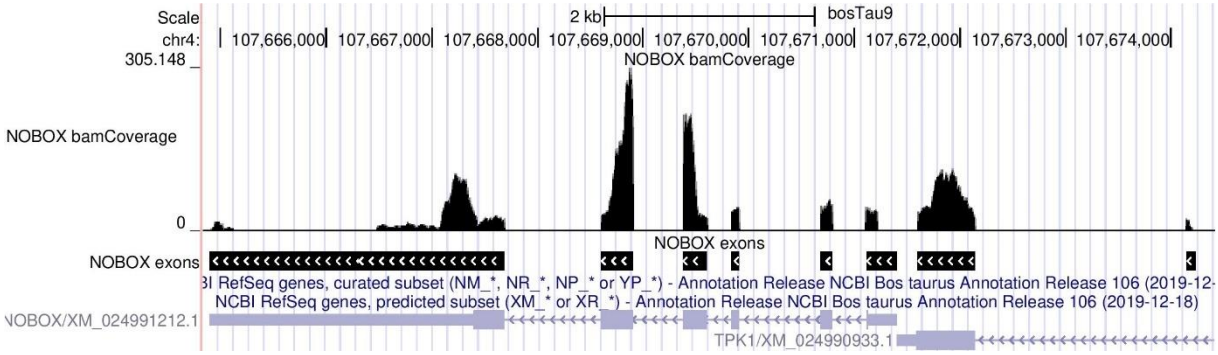


Figure 15. Illustrative scheme of exon stretches of *NOBOX* gene and the RNA-seq read coverage from 16-cell stage. (Modified from Kent *et al.*, 2002.)

2.3.1.4 *DPRX* ortholog not present in the RefSeq predictions. There is a three-exon long transcript in the latest bovine genome assembly available (bosTau9) at the genomic location of the *DPRX* gene, however the re-analysis data revealed only a two-exon transcript at the eight-cell stage (Figure 16).

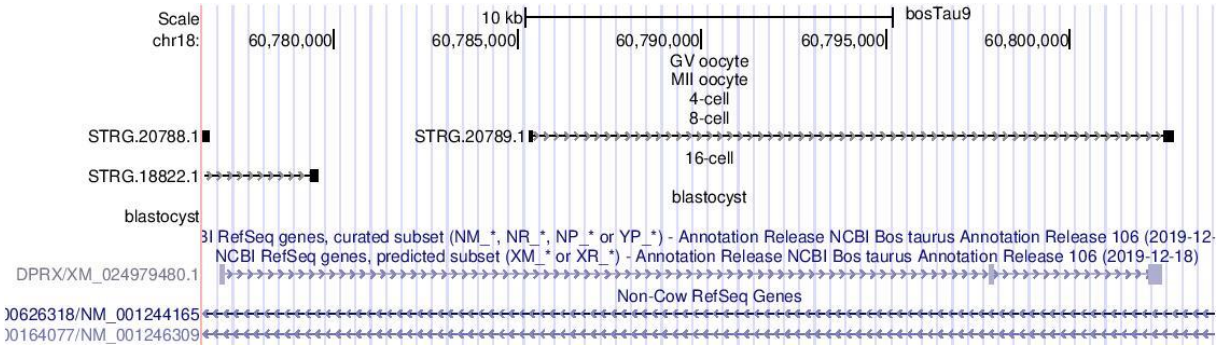


Figure 16. Trancripts at the *DPRX* genomic location in the RNA-seq data derived from oocytes and preimplantation IVF embryos. The trimmed reads obtained from six different developmental stages were re-analysed and aligned to the latest bovine genome assembly avilable. (Modified from Kent *et al.*, 2002.)

2.3.1.5 *LEUTX* ortholog not present in the RefSeq predictions. The genomic region of *LEUTX* gene in gallus genome overlaps with the *DBX1* gene in the bovine genome. There is no *LEUTX* gene ortholog in the bovine genome in that genomic location (Figure 17).

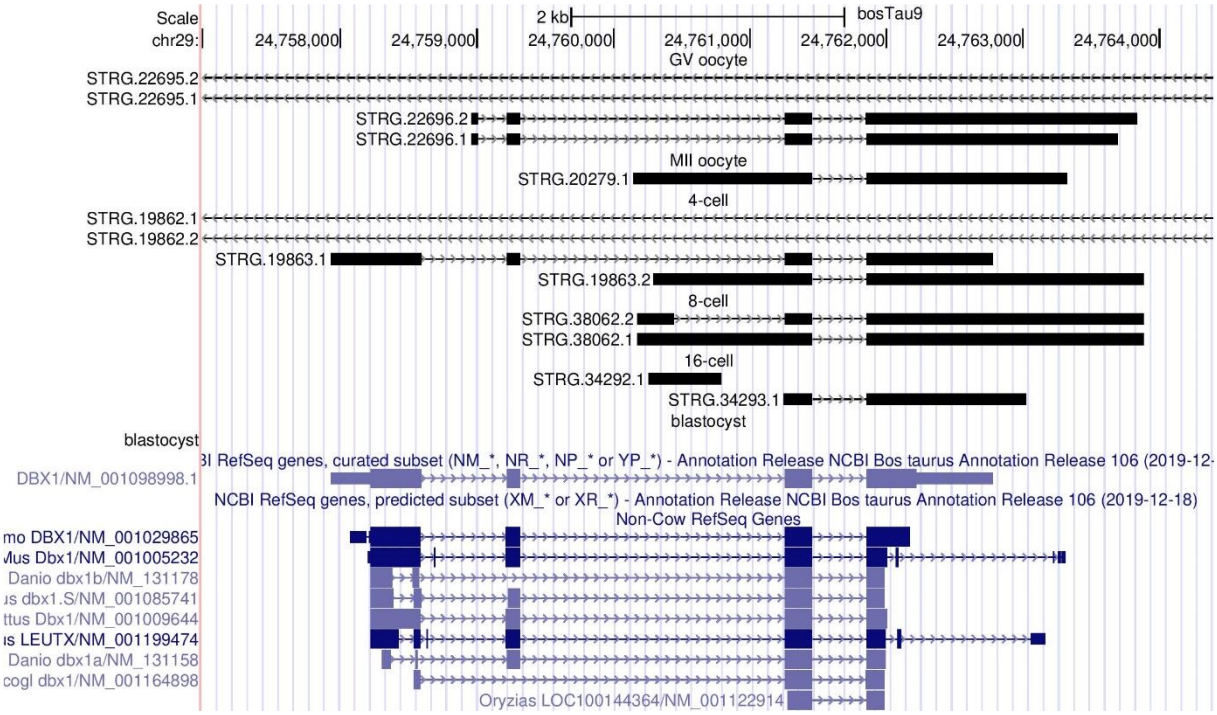


Figure 17. Trancripts at the gallus *LEUTX* genomic location in the RNA-seq data derived from oocytes and preimplantation IVF embryos. The trimmed reads obtained from six different developmental stages were re-analysed and aligned to the latest bovine genome assembly avilable. (Modified from Kent *et al.*, 2002.)

2.3.1.6 *DUXB, CPHX1, CPHX2, TPRX1, TPRX2* orthologs not present in the RefSeq predictions. Searching for genomic regions of *DUXB, CPHX1, CPHX2, TPRX1* and *TPRX2* did not have a successful result as these genes are not annotated in UCSC Genome Browser in Cow Apr. 2018 (ARS-UCD1.2/bosTau9).

2.3.2 Wet lab experiments

2.3.2.1 cDNA library preparation by STRT-N. STRT-N method with original and modified primer length was tested on bovine fibroblast RNA (Figure 18). Next, the method was tested directly on embryo(s) or oocyte(s). The mRNAs appear as a smear for all the developmental stages analysed (Figures 19 and 20).

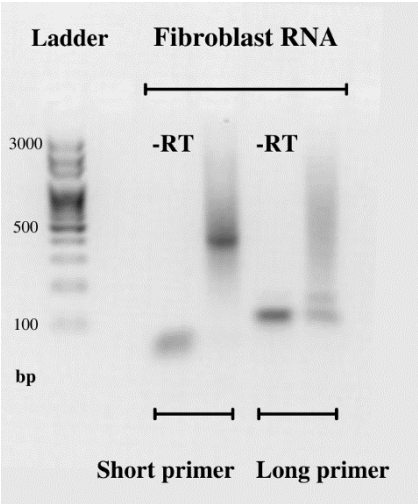


Figure 18. STRT-N with modified primer length tested on bovine fibroblast RNA. The control samples, for which the reverse transcription was carried out in the absence of reverse transcriptase, are indicated with minus reverse transcriptase (-RT). The legend for the band sizes is indicated left of the ladder (Yaşar).

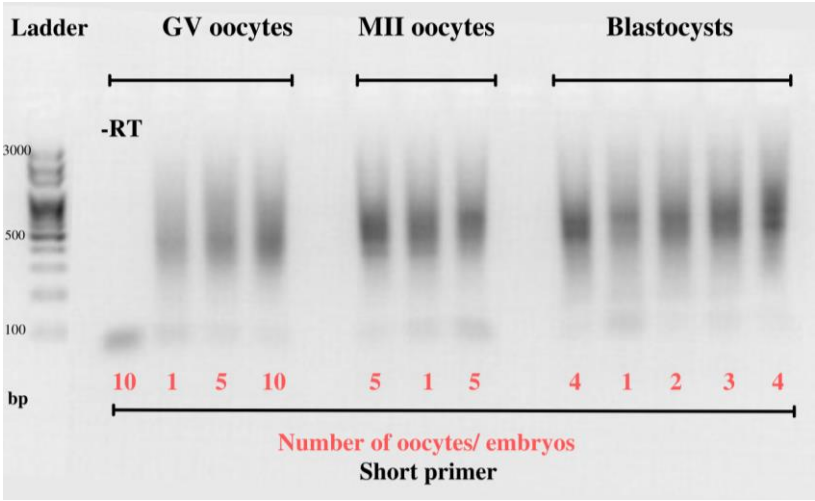


Figure 19. STRT-N with short primer in bovine GV oocyte, MII oocyte and blastocyst stages. Number of oocytes/embryos processed is indicated in red. The control sample, for which the reverse transcription was carried out in the absence of reverse transcriptase is indicated with -RT. The legend for the band sizes is indicated left of the ladder (Yaşar).

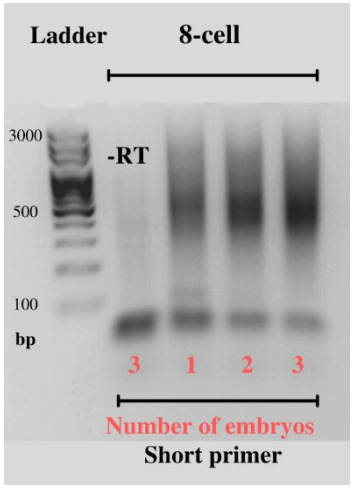


Figure 20. STRT-N with short primer in bovine eight-cell stage. Number of embryos processed is indicated in red. The control sample, for which the reverse transcription was carried out in the absence of reverse transcriptase is indicated with -RT. The legend for the band sizes is indicated left of the ladder (Yaşar).

2.3.2.2 Amplification of genes with gene-specific primers to confirm orthologs. The efficiency of the amplification was checked with agarose gel electrophoresis. Bands of *NOBOX*, *DUXA* and *ARGFX* gene were visualized (Figure 21). The band sizes of *NOBOX*, *DUXA* and *ARGFX* are 2000, 800 and 1000 bp, respectively.

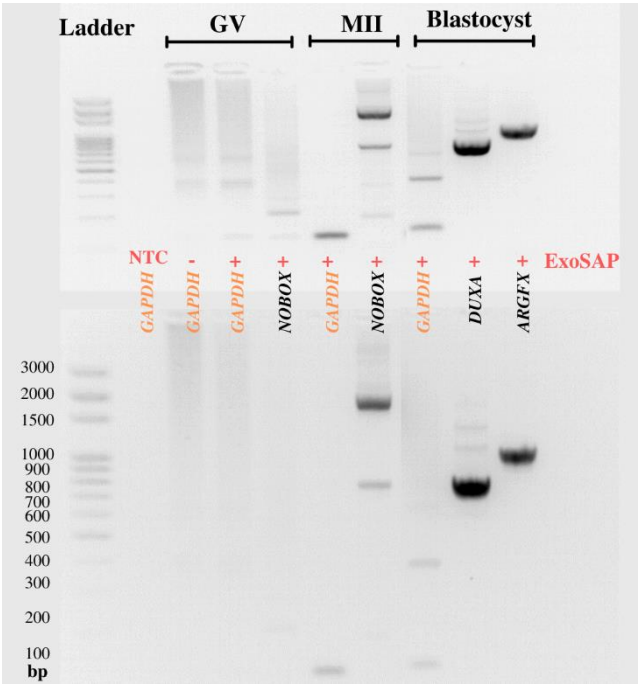


Figure 21. Amplified cDNA with gene-specific primers. The genes were amplified using gene-specific primers. The cDNA product clean-up is indicated with "+" in red. The nether gel picture was run for longer with the same samples. The sizes of the ladder are indicated in the left for the nether ladder (Yaşar).

2.3.2.3 Gel extraction after gene-specific PCR. The bands of highest intensity and of predicted sizes for *ARGFX*, *DUXA* and *NOBOX* genes were selected (Figure 22).

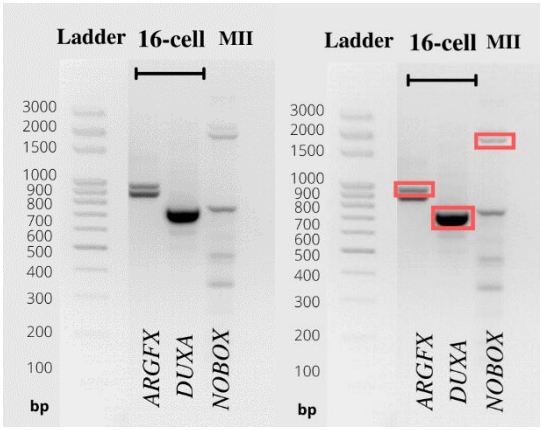


Figure 22. The selected bands of amplified genes. The bands of predicted sizes for genes of interest and of highest intensity were selected. The selected bands from the left are shown in red boxes on the right. The legend of the band sizes is shown on the left of each ladder (Yaşar).

2.3.2.3 EcoRI restriction enzyme to check cloning of the desired genes. The cloned plasmids were cut with the EcoRI restriction enzyme and checked with agarose gel electrophoresis for the efficiency of cloning the genes of interest. The five clones of *ARGFX* have a band of 1000 bp; the first clone of *NOBOX* has a band of 1440 bp, the second clone of *NOBOX* has a band of 2000 bp, the third clone of *NOBOX* has a band of 2500 and the fourth clone of *NOBOX* has a band of 2750 bp; the three clones of *DUXA* gene have a band of 800 bp.

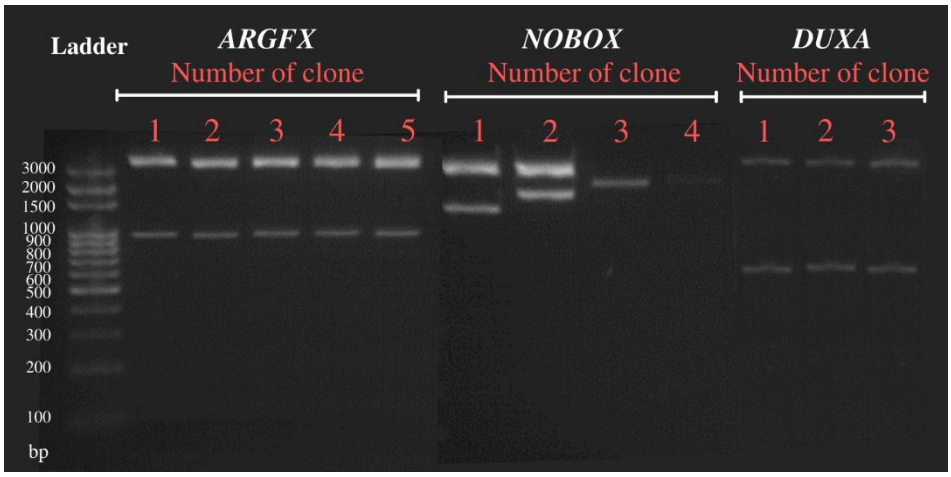


Figure 23. EcoRI restriction enzyme digestion of clones of *ARGFX*, *NOBOX* and *DUXA* genes. The effectiveness of the EcoRI restriction enzyme and therefore the cloning was checked with agarose gel electrophoresis. The clone numbers are indicated in red. Ladder was loaded in the first well with the size legend on its left (Yaşar).

2.4 Discussion

The re-analysis of the published RNA-seq dataset (Graf *et al*., 2014) provided an insight into the identification of the PRD-like TFs which have recently been discovered in human preimplantation embryos (Töhönen *et al*., 2015). According to bovine NCBI and UCSC RefSeq data, the genes investigated in this thesis were either not annotated at all or computationally predicted in the bovine genome lacking cDNA evidence. Human early TFs include PRD-like homeobox genes which are predicted in the bovine genome: *ARGFX, DPRX, DUXA* and *NOBOX;* and the ones with no predicted gene structures: *CPHX1, CPHX2, DUXB, LEUTX, TPRX1* and *TPRX2*.

In this thesis, PRD-like homeobox genes predicted in *Bos taurus* with possible roles in the EGA were selected for investigation to provide annotation of these genes by molecular cDNA cloning. The molecular cloning of genes is possible if corresponding transcripts are reverse transcribed into cDNA. To increase the likelihood of having the transcript of interest in the cDNA library, the developmental stage, where the transcript is expressed in highest abundance, was selected. The presence or absence of the orthologs of the genes of interest is discussed in the next parts.

A 13 bp deletion upstream of the homeodomain-coding region has caused a frameshift and a truncated protein synthesis for the *ARGFX* gene (Lewin *et al*., 2021), explaining why at the first try, the putative protein sequence of the chosen transcript did not seem to have a significant match with a homeodomain. The truncated sequence of the spliced transcript and its putative protein synthesis had a significant match with a homeodomain, confirming the prediction of the gene ortholog in the bovine genome.

The *DUXA* gene was predicted to contain six exons only in higher primates' genomes. The bovine genome was believed to have a stop codon at the equivalent of the fifth exon position. The original *DUXA* gene was believed to contain only five exons, and the gain of the sixth one only happened in primates (Leidenroth and Hewitt, 2010). The *in silico* prediction of the gene ortholog has overturned the hypothesis by showing a six-exon version of the gene also in the bovine genome. The presence of two homeodomains in the putative protein sequence of the predicted ortholog's transcript confirms the prediction of the gene as the double homeobox is expected.

The prediction of *NOBOX* gene ortholog in the bovine genome *in silico* is confirmed by the presence of a homeodomain in the putative protein sequence. The gene plays a crucial role in the oogenesis stage (Suzumori *et al*., 2002) of early development, which is in concordance with

the finding of the highest transcript expression abundance in the GV oocyte developmental stage.

The *DPRX* gene was predicted to be a pseudogene due to a 2 bp insertion in the homeobox, which has resulted in a loss of an exon (Lewin *et al*., 2021). Based on the findings of this thesis, the expression of predicted *DPRX* gene ortholog in the bovine genome appears to be weak or non-existent as a two-exon transcript is found at the genomic locus of the gene at the eight-cell stage instead of the three-exon transcript as predicted within the predicted subset of NCBI RefSeq genes. The two exons do not overlap with the exons of the predicted subset of NCBI RefSeq genes either. Since the transcript was found to be expressed only in the eight-cell stage and for the forementioned reasons, it can be implied that no *DPRX* gene ortholog is expressed in the bovine genome.

As mentioned, the gene structures of *DUXB*, *LEUTX*, *CPHX1*, *CPHX2*, *TPRX1* and *TPRX2* have not been predicted in the bovine genome. This was confirmed *in silico* by looking at the genes' genomic locations. Considering the synteny of *LEUTX*, *TPRX1* and *TPRX2*, the neighbouring loci could be studied with synteny analysis to find the genomic locations of the genes' orthologs in the bovine model. The synteny analysis is part of the ongoing project (Yaşar).

After the *in silico* gene predictions and gene-specific primer design, a selection of human PRD-like homeobox genes predicted to have a role in the EGA was cloned and sequenced with the purpose of annotating the genes in the bovine genome. The genes under consideration were *ARGFX*, *DUXA* and *NOBOX*.

STRT-N, which was used for the cDNA preparation, is a method originally developed for sample preparation for single-cell RNA sequencing. GV and MII stage oocyte, and 8-cell and blastocyst stage bovine embryo cDNA libraries were prepared by STRT-N method which benefits reverse transcription followed by PCR. The initial step of STRT-N includes cDNA preparation using a primer long enough to provide a site for sequencing primers to anneal later. For this study, the length of the primer used was modified from the one suggested in the protocol and the efficiency of it checked on RNA isolated from bovine fibroblast. The observable smear of mRNAs after STRT-N method with original and modified primer on bovine fibroblast RNA was relatively similar, therefore the effectiveness of the modified primer length has been proved. The control samples of reverse transcription reaction with the absence of reverse transcriptase appeared as a single band on the gel, not as a smear. This confirmed that there was no genomic DNA contamination in the samples.

In the process of amplifying the cDNAs with gene-specific primers designed *in silico* from the genes' loci, the DNA polymerase and exonuclease activity produced high fidelity products suitable for cloning. The DNA polymerase activity worked in the direction of 5' to 3' and the exonuclease (proofreading) activity in the direction of 3' to 5', amplifying long amplicons (Thermo Scientific, USA). Expression of *GAPDH*, found to be a suitable reference gene in normalization of bovine data (Robinson *et al*., 2007), could be observed as a band in the MII oocyte, eight-cell, and blastocyst stages, however no visible band appeared in the GV oocyte stage. This could have been due to a technical or batch error. The *NOBOX* gene was expected to be expressed in the GV oocyte developmental stage as learned when predicting the expression of the gene, however there was no visible band for the gene in the GV oocytes stage. This indicates a possible batch error or the ineffective primer annealing for that stage oocytes.

The lengths of spliced transcripts for amplified *NOBOX*, *DUXA* and *ARGFX* were expected to be approximately 2829, 876 and 1054 nt, respectively, which was confirmed by agarose gel electrophoresis. The EcoRI restriction enzyme cut the cloned genes at the *Eco*R I sites, producing a fragment specific to each gene. For *DUXA* and *ARGFX*, the replicas of clones were of the expected size. The four clones of *NOBOX* gene had varying band size, which could have been a result of amplification of alternative transcripts.

As mentioned, the experimental work in this thesis provided the groundwork for the validation of the genes in the bovine genome. Further study would include the analysis of the Sanger sequencing reads to confirm the expression of the human PRD-like gene orthologs in the bovine genome. In case of expressed transcripts in the expected genomic loci, the genes could be annotated.

SUMMARY

The high DNA-binding affinity of the PRD-like homeodomain sequences and their conserved nature has suggested their crucial role in the EGA as TFs. A selection of novel PRD-like homeobox genes have recently been characterized as having full length ORFs containing homeodomains, although lacking evidence of transcript expression. Due to the ethical implications of studying human embryos, a model organism was chosen to conduct the experiments on. Aspects such as oocyte time, time to maturation, early embryo development and EGA of humans can be best reflected on in the bovine model organisms. Studying bovine embryos could contribute to the better understanding of human cell reprograming and pluripotent stem cells.

In this thesis, *in silico* experiments on the RNA-seq data derived from oocytes and preimplantation IVF embryos revealed predictions of three human PRD-like class homeobox gene orthologs in bovine genome. The genes selected for investigation were *ARGFX*, *DUXA* and *NOBOX*. *In silico* investigation was also done for genes *LEUTX*, *DPRX*, *CPHX1*, *CPHX2*, *TPRX1* and *TPRX2*, however no orthologs of these genes were found in the genomic regions of these genes. The predictions of *ARGFX*, *DUXA* and *NOBOX* orthologs enabled the downstream wet lab experiments with gene-specific primers with the purpose of annotating them in the bovine genome.

The Holstein Friesian cattle oocytes of two developmental stages (GV and MII) and embryos of two developmental stages (16-cell and blastocyst) were collected from a collaborative laboratory at the Estonian University of Life Sciences. cDNA from each developmental stage oocyte(s) and embryo(s) was prepared with the STRT-N method, followed by cloning and Sanger sequencing. In this thesis, the groundwork for the validation of these genes in the bovine genome was completed.

Homeootiliste PRD-sarnase klassi geenide *ARGFX, DUXA* ja *NOBOX* ennustamine ja iseloomustamine veise genoomis

Piibe Vill

Eestikeelne resümee

Viljastumise etapi järgselt toimub organismides üleminek emapoolse genoomi ekspressioonilt embrüo genoomi ekspressioonile. Ülemineku etappi reguleerivad konserveerunud mehhanismid, mille aktivatsiooni aeg erineb erinevates organismides (Schultz *et al*., 1999). Üheks konserveerunud mehhanismiks võib lugeda homeootiliste geenide ekspressiooni. Neil geenidel on konserveerunud 180 aluspaari pikkune ala, mis omakorda kodeerib 60 aminohappe pikkust homeodomeeni. Homeodomeeni võime seonduda DNA-ga tähendab, et homeootilised geenid mängivad rolli geeni regulatsioonis transkriptisoonifaktoritena (Bürglin, 2011; McGinnis *et al*., 1984; Scott and Weiner, 1984). Viimase aastakümne sees on avastatud, et homeootiliste PRD-sarnase klassi geenid mängivad rolli ülalmainitud üleminekus (Töhönen *et al*., 2015). Täielik arusaam ning annotatsioon PRD-sarnase klassi geenidest on puudulik ning nende uurimine on pooleli.

Katsed inimese embrüotega on eetilistel põhjustel limiteeritud, kuid kuna vaatluse all olevad geenid on konserveerunud ka teistes organismides, on võimalik iseloomustada neid mudel organismides. Käesoleva magistritöö raames uuriti puuduliku annotatsiooniga geene veise ootsüütides ja embrüotes. *In silico* uuriti, kas geenid *ARGFX, CPHX1, CPHX2, DPRX, DUXA, LEUTX, NOBOX, TPRX1* ja *TPRX2* ortoloogid on olemas ka veise genoomis. Loetletud geenidest avastati *ARGFX, DUXA* ja *NOBOX* ortoloogsed geenid veise genoomis. Eesti Maaülikoolilt saadud veise ootsüütidest ja embrüotest tehti komplementaarne DNA STRT-N meetodil, mis võimaldab määrata mRNA 5’ otste transkriptsiooni ühe raku staadiumil. Komplementaarne DNA amplifitseeriti lookuse spetsiifiliste praimeritega ning klooniti Zero Blunt® TOPO® PCR Cloning Kit protokolli järgi. Kloonitud komplementaarsele DNA-le teostati Sanger sekveneerimine. Sekveneerimisandmeid magistritöös ei analüüsitud, kuid käesoleva projekti raames võimaldab nende analüüs annoteerida analüüsitud geenid veise genoomis. See omakorda võimaldab paremat arusaama varajasest embrüo arengust ka inimeses.

REFERENCES

Afgan, E., Baker, D., Batut, B., … Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 46(W1): W537-W544.

Bai, H., Sakurai, T., Godkin, J. D. and Imakawa, K. (2013). Expression and potential role of GATA factors in trophoblast development. Journal of Reproduction and Development 59(1): 1-6.

Baker, R. D. and Polge, C. (1976). Fertilization in swine and cattle. Canadian Journal of Animal Science 56(2): 105-119.

Barnes, F. L. and First, N. L. (1991). Embryonic transcription in *in vitro* cultured bovine embryos. Molecular Reproduction and Development 29(2): 117-123.

Billeter, M., Qian, Y. Q., Otting, G., Müller, M., Gehring, W. and Wüthrich, K. (1993). Determination of the nuclear magnetic resonance solution structure of an *Antennapedia* homeodomain-DNA complex. Journal of Molecular Biology 234(4): 1084-1097.

Booth, H. A. F. and Holland, P. W. H. (2007). Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene 387(1-2): 7-14.

Boskovic, N. Manuscript in preparation.

Burri, M., Tromvoukis, Y., Bopp, D., Frigerio, G. and Noll, M. (1989). Conservation of the paired domain in metazoans and its structure in three isolated human genes. The EMBO Journal 8(4): 1183-1190.

Bürglin, T. R. (2011). Homeodomain subtypes and functional diversity. Subcellular Biochemistry 52: 95-122.

Camous, S., Kopecný, V. and Fléchon, J. E. (1986). Autoradiographic detection of the earliest stage of [3H]-uridine incorporation into the cow embryo. Biology of the Cell 58(3): 195-200.

Clapp, J., Mitchell, L. M., Bolland, D. J., Fantes, J., Corcoran, A. E., Scotting, P. J., Armour, J. A. L. and Hewitt, J. E. (2007). Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. The American Journal of Human Genetics 81(2): 264-279.

De Felici, M., Klinger, F. G., Farini, D., Scaldaferri, M. L., Iona, S. and Lobascio, M. (2005). Establishment of oocyte population in the fetal ovary: primordial germ cell proliferation and oocyte programmed cell death. Reproductive BioMedicine Online 10(2): 182-191.

Delouis, C., Bonnerot, C., Vernet, M. and Nicolas, J.-F. (1992). Expression of microinjected DNA and RNA in early rabbit embryos: changes in permissiveness for expression and transcriptional selectivity. Experimental Cell Research 201(2): 284-291.

Dobson, A. T., Raja, R., Abeyta, M. J., Taylor, T., Shen, S., Haqq, C. and Pera, R. A. R. (2004). The unique transcriptome through day 3 of human preimplantation development. Human Molecular Genetics 13(14): 1461-1470.

French, D. B., Sabanegh Jr, E. S., Goldfarb, J. and Desai, N. (2010). Does severe teratozoospermia affect blastocyst formation, live birth rate, and other clinical outcome parameters in ICSI cycles? Fertility and Sterility 93(4): 1097-1103.

Gardner, D. K., Lane, M. and Schoolcraft, W. B. (2000). Culture and transfer of viable blastocysts: a feasible proposition for human IVF. Human Reproduction 15(6): 9-23.

Gehring, W. J., Affolter, M. and Bürglin, T. (1994). Homeodomain proteins. Annual Review of Biochemistry 63: 487-526.

Gordon, I. R. (2003). *Laboratory Production of Cattle Embryos* (2nd ed.). Ireland: CABI Publishing.

Graf, A., Krebs, S., Zakhartchenko, V., Schwalb, B., Blum, H. and Wolf, E. (2014). Fine mapping of genome activation in bovine embryos by RNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 111(11): 4139-4144.

Grive, K. J. and Freiman, R. N. (2015). The developmental origins of the mammalian ovarian reserve. Development 142: 2554-2563.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D. and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Developmental Cell 18(4): 675-685.

Hamilton, W. J. and Laing, J. A. (1946). Development of the egg of the cow up to the stage of blastocyst formation. Journal of Anatomy 80(4): 194-204.5.

Holland, P. W. H. (2012). Evolution of homeobox genes. WIREs Developmental Biology 2(1): 31-45.

Holland, P. W. H., Booth, H. A. F. and Bruford, E. A. (2007). Classification and nomenclature of all human homeobox genes. BMC Biology 5, 47.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Research 21(7): 1160-1167.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P. and Linnarsson, S. (2012). Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. Nature Protocols 7(5): 813-828.

Jouhilahti, E.-M., Madissoon, E., Vesterlund, L., … Kere, J. (2016). The human PRD-like homeobox gene *LEUTX* has a central role in embryo genome activation. Development 143(19): 3459-3469.

Kanka, J., Nemcova, L., Toralova, T., Vodickova-Kepkova, K., Vodicka, P., Jeseta, M. and Machatkova, M. (2012). Association of the transcription profile of bovine oocytes and embryos with developmental potential. Animal Reproduction Science 134(1-2): 29-35.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Research 32: D493-D496.

Katayama, S., Töhonen, V., Linnarsson, S. and Kere, J. (2013). SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics 29(22): 2943-2945.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at UCSC. Genome Research 12(6): 996-1006.

Leidenroth, A. and Hewitt, J. E. (2010). A family history of *DUX4*: phylogenetic analysis of *DUXA, B, C* and *Duxbl* reveals the ancestral *DUX* gene. BMC Evolutionary Biology 10, 364.

Lewin, T. D., Royall, A. H. and Holland, P. W. H. (2021). Dynamic molecular evolution of mammalian homeobox genes: duplication, loss, divergence and gene conversion sculpt PRD class repertoires. Journal of Molecular Evolution 89: 396-414.

Li, G. and Holland, P. W. H. (2010). The origin and evolution of *ARGFX* homeobox loci in mammalian radiation. BMC Evol. Biol., 10(182).

Li, H., Tsai, M.-S., Chen, C.-Y., Lian, W.-C., Chiu, Y.-T., Chen, G.-D. and Wang, S.-H. (2006) A novel maternally transcribed homeobox gene, *Eso-1*, is preferentially expressed in oocytes

and regulated by cytoplasmic polyadenylation. Molecular Reproduction and Development 73: 825-833.

Li, J. and Tibshirani, R. (2011). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research 0(0): 1-18.

Madissoon, E., Jouhilahti, E.-M., Vesterlund, L., … Kere, J. (2016). Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Scientific Reports 6, 28995.

Maeso, I., Dunwell, T. L., Wyatt, C. D. R., Marlétaz, F., Vető, B., Bernal, J. A., Quah, S., Irimia, M. and Holland, P. W. H. (2016). Evolutionary origin and functional divergence of totipotent cell homeobox genes in eutherian mammals. BMC Biology 14, 45.

Masui, Y. and Clarke, H. J. (1979). Oocyte maturation. International Review of Cytology 57: 185-282.

Mazza, M. E., Pang, K., Reitzel, A. M., Martindale, M. Q. and Finnerty, J. R. (2010). A conserved cluster of three PRD-class homeobox genes (*homeobrain*, *rx* and *orthopedia*) in the Cnidaria and Protostomia. EvoDevo 1, 3.

McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. and Gehring, W. J. (1984). A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. Cell 37(2): 403-408.

Mistry, J., Chuguransky, S., Williams, L., … Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic Acids Research 49(D1): D412-D419.

Niemann, H. and Wrenzycki, C. (2000). Alterations of expression of developmentally important genes in preimplantation bovine embryos by *in vitro* culture conditions: implications for subsequent development. Theriogenology 53(1): 21-34.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., Mclnerney, P., Thompson, J. F., Bowers, J., Jarosz, M. and Milos, P. M. (2009). Direct RNA sequencing. Nature 461(7265): 814-818.

Qin, Y., Choi, Y., Zhao, H., Simpson, J. L., Chen, Z.-J. and Rajkovic, A. (2007). *NOBOX* homeobox mutation causes premature ovarian failure. American Journal of Human Genetics 81(3): 576-581.

Robinson, T. L., Sutherland, I. A. and Sutherland, J. (2007). Validation of candidate bovine reference genes for use with real-time PCR. Veterinary Immunology and Immunopathology 115(1-2): 160-165.

Santos, R. R., Schoevers, E. J. and Roelen, B. A. J. (2014). Usefulness of bovine and porcine IVM/IVF models for reproductive toxicology. Reproductive Biology and Endocrinology 12, 117.

Sayers, E. W., Bolton, E. E., Brister, J. R., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Research 50(D1): D20-D26.

Schmidt, W. M. and Mueller, M. W. (1999). CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. Nucleic Acids Research 27(21).

Schultz, R. M., Davis Jr, W., Stein, P. and Svoboda, P. (1999). Reprogramming of gene expression during preimplantation development. Journal of Experimental Zoology 285(3): 276-282.

Scott, M. P. and Weiner, A. J. (1984). Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of *Drosophila*. Proceedings of the National Academy of Science of the United States of America 81(13): 4115-4119.

Sirard, M.-A. (2010). Activation of the embryonic genome. Society of Reproduction and Fertility Supplement 67: 145-158.

Speyer, B., O'Neill, H., Saab, W., Seshadri, S., Cawood, S., Heath, C., Gaunt, M. and Serhal, P. (2019). In assisted reproduction by IVF or ICSI, the rate at which embryos develop to the blastocyst stage is influenced by the fertilization method used: a split IVF/ICSI study. Journal of Assisted Reproduction and Genetics 36: 647-654.

Suzumori, N., Yan, C., Matzuk, M. M. and Rajkovic, A. (2002). *Nobox* is a homeobox-encoding gene preferentially expressed in primordial and growing oocytes. Mechanisms of Development 111(1-2): 137-141.

Tang, F., Barbacioru, C., Wang, Y., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6(5): 377-382.

Töhönen, V., Katayama, S., Vesterlund, L., … Kere, J. (2015). Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nature Communications 6, 8207.

Vallée, M., Dufort, I., Desrosiers, S., Labbe, A., Gravel, C., Gilbert, I., Robert, C. and Sirard M.-A. (2009). Revealing the bovine embryo transcript profiles during early *in vivo* embryonic development. Reproduction 138: 95-105.

Viuff, D., Hyttel, P., Avery, B., Vatja, G., Greve, T., Callesen, H. and Thomsen, P. D. (1998). Ribosomal ribonucleic acid is transcribed at the 4-cell stage in *in vitro*-produced bovine embryos. Biology of Reproduction 59: 626-631.

Vorobyov, E. and Horst, J. (2006). Getting the proto-Pax by the tail. Journal of Molecular Evolution 63: 153-164.

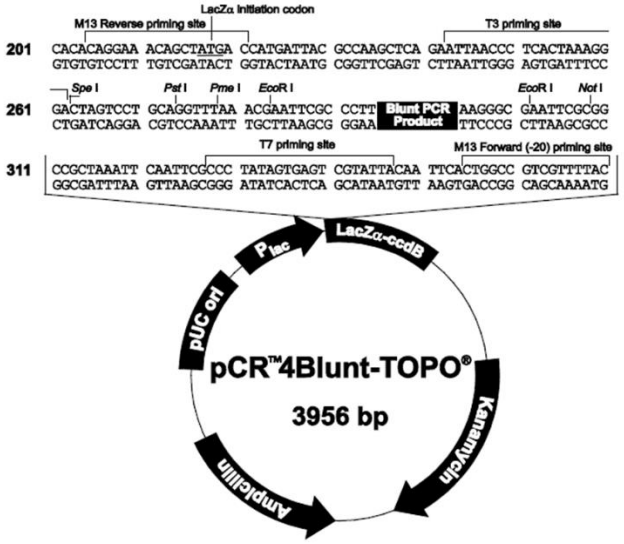
Wang, H. and Dey, S. K. (2006). Roadmap to embryo implantation: clues from mouse models. Nature Reviews Genetics 7: 185-199.

Yaşar, B. Manuscript in preparation.

Zhong, Y.-F. and Holland, P. W. H. (2011). Erratum to: The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. BMC Evolutionary Biology 11, 204.

APPENDIX

APPENDIX 1



Schematic overview of the plasmid sequence, where the blunt-ended PCR product was inserted. The sites of *Eco*R I, and T3 and T7 priming are shown (Thermo Scientific, USA).

APPENDIX 2

Reaction number	Sample name	Primer name (10 pmol/µl)	PCR product/ Plasmid	Length (bp)	Sample conc. (ng/µl)	PCR product needs SAP/Exo purification?
1	*ARGFX*	T3	Plasmid	1000	66.6	-
2	*ARGFX*	T7	Plasmid	1000	66.6	-
3	*DUXA*	T3	Plasmid	800	66.6	-
4	*DUXA*	T7	Plasmid	800	66.6	-
5	*NOBOX*	T3	Plasmid	2000	66.6	-
6	*NOBOX*	T7	Plasmid	2000	66.6	-

The information sheet attached to the samples sent for Sanger sequencing.

NON-EXCLUSIVE LICENCE

Non-exclusive licence to reproduce thesis and make thesis public

I, Piibe Vill,

1. grant the University of Tartu a free permit (non-exclusive licence) to:

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

"Prediction and characterization of PRD-like homeobox genes *ARGFX*, *DUXA* and *NOBOX* in the bovine genome",

supervised by Bariş Yaşar and Ants Kurg,

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work from 30/05/2025 until the expiry of the term of copyright,

3. I am aware that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Piibe Vill

30/05/2022