Genetic variation in bulls divergent for fertility

Ronan Whiston B.Sc., M.Sc.

2017

A thesis submitted to Trinity College Dublin

for the degree of Doctor of Philosophy

School of Biochemistry and Immunology,
Trinity College Dublin

Supervisors:

Professor Cliona O'Farrelly and Dr Kieran G. Meade

Contents

Declaration vi
Acknowledgements vii
Abbreviations viii
List of figures x
List of tables xii
Index of electronic appendices xiii
Summary xv
1 General Introduction 1
1.1 Bovine fertility 2
1.1.1 Domestication of cattle 2
1.1.2 Antagonistic relationship between milk production and fertility 2
1.1.3 Consequences of lower fertilisation rates 4
1.1.4 Artificial selection in cattle production 6
1.1.5 Genomic selection 6
1.2 Male fertility 8
1.2.1 Genetic association studies in cattle 10
1.2.2 Fertility evaluation in bulls 17
1.2.3 Current status of bull fertility 20
1.2.4 Fertility and the immune system 22
1.3 Defensin genes and male fertility 23
1.3.1 Defensin family – structure 23
1.3.2 Bovine β-defensins – genetic structure and function 25
1.3.3 Role of *DEFB126* in fertility in multiple species 28
1.3.4 Role of other β-defensins in fertility in rodents 31
1.4 Bioinformatics 32
1.4.1 Bioinformatics in bovine research 32
1.4.2 The bovine genome 32
1.4.3 SNP databases 33
1.4.4 Exome sequencing 34
1.5 Applications of research 36

1.5.1 Biomarkers 36
1.5.2 National genotyping scheme for cattle 37
1.5.3 IDB SNP chip – large GWAS dataset 37
1.6 Aims 39
1.7 Hypothesis 39
1.8 Objectives 39
2 Materials and methods 40
2.1.1 Phenotypic data 41
2.1.2 Sample selection 41
2.1.3 Probe design – β-defensin and WES 41
2.2 Materials and methods related to Chapter 3. 45
2.2.1 Targeted β-defensin sequencing of AI sires - library preparation and sequencing 45
2.2.2 Data analysis of targeted β-defensin sequencing dataset 45
2.2.3 SNP filtration of β-defensin sequencing dataset 46
2.2.4 SNP association analysis of β-defensin sequencing 46
2.2.5 Targeted β-defensin re-sequencing in sire subset 47
2.2.6 Targeted β-defensin gene re-sequencing library preparation in sire subset 47
2.2.7 Targeted re-sequencing data analysis in sire subset 48
2.2.8 SNP frequency analysis in sire subset 50
2.2.9 O-linked glycosylation analysis 50
2.3 Materials and Methods related to Chapter 4 51
2.3.1 Genomic DNA extraction, purification and quality control 56
2.3.2 Sample preparation 59
2.3.3 Exome data analysis 61
2.3.4 Alignment 61
2.3.5 Variant calling 61
2.3.6 Variant filtering 63
2.3.7 Quality control 64
2.3.8 Association analysis 65
2.3.9 Gene ontology 66
2.3.10 Transcription factor binding site analysis 66

2.4 Materials and methods related to Chapter 5 67
2.4.1 Sire selection for validation 67
2.4.2 Assay design 70
2.4.3 SNP validation 71
2.4.4 Data analysis 71
3 Targeted β-defensin gene sequencing in divergent fertility bulls 73
3.1 Introduction 74
3.2 Aims and hypothesis 76
3.3 Results 76
3.3.1 Fertility phenotypes of AI sires 76
3.3.2 Targeted β-defensin sequencing coverage statistics 80
3.3.3 Targeted sequencing variant discovery and filtering 83
3.3.4 Targeted sequencing variant association analysis 84
3.3.5 Targeted sequencing coverage in subset of sires 89
3.3.6 Variant discovery 94
3.3.7 Annotation 94
3.3.8 β- defensin SNP frequency analysis 95
3.3.9 Targeted β-defensin SNP association in subset of sires 97
3.3.10 O-linked glycosylation analysis in β-defensin genes 98
3.4 Discussion 101
4 Whole-exome sequencing of bulls divergent for fertility 104
4.1 Introduction 105
4.2 Aims 106
4.3 Results 106
4.3.1 Exome sequencing coverage 106
4.3.2 Variant discovery 110
4.3.3 Variant annotation 110
4.3.4 Exome variant SNP frequency analysis 112
4.3.5 Breed-specific SNP frequencies 115
4.3.6 Gene ontology 115
4.3.7 Quality control 117
4.3.8 SNP association 121

4.3.9 Transcription factor binding site analysis 125
4.3.10 Exome sequencing validation 125
4.4 Discussion 128
5 Variant validation in an independent bull population 133
5.1 Introduction 134
5.2 Aims 135
5.3 Results 135
5.3.1 Variant validation 135
5.3.2 SNP frequency 136
5.3.3 Correlation 139
5.4 Discussion 147
6 Final discussion 149
6.1 Future research opportunities 157
7 References 158
8 Associated publications 173

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work except where duly acknowledged.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: _____

Date: _____

Acknowledgements

First and foremost, I would like to acknowledge the efforts of everyone in helping to accomplish the research presented in this Ph.D. thesis. All contributions, however great or small, are greatly appreciated.

In particular, I would like to thank Dr. Kieran Meade, Animal and Bioscience research department, Grange, for your guidance, help, dedication and mentorship. Your efforts have helped guide the work to completion with great care and attention provided throughout. I sincerely thank you for all you have done for me. I've been very lucky to have you as a supervisor.

To Cliona O'Farrelly, Trinity Biomedical Sciences Institute, Trinity College Dublin, I would like to thank you for your guidance in completing the thesis, your support in building key skills and your dedication. I appreciate the time given to help complete this project as my academic supervisor.

To all in Teagasc, Animal and Bioscience research department, Grange, thank you for all your support and help throughout my time there. A special word of thanks goes to Amy Brewer, Dr. Bojan Stojkovic, Dr. Cathriona Foley, Rachael Doherty, Megan O'Brien, Nicholas Ryan, Dr. Paul Cormican, Dr. Matt McCabe, Dr. Orla Keane, Dr. Bruce Moran (UCD), Dr. Anthony Doran, Joe Larkin and Margaret Murray.

To all in the Comparative Immunology Group, Trinity Biomedical Sciences Institute, I would like to thank you for all your help, advice and support.

I would like to thank everyone in the β-defensin research group, specifically, Prof. Patrick Lonergan, Dr. Seán Fair, Dr. Emma Finlay, Anne Barry-Reidy, Dr. Beatriz Fernandez-Fuertes, Alan Lyons, Ilaina Khairulzaman, and Dr. Fernando Narciandi.

A note of thanks to Professor Peter J. Hansen (University of Florida), Dr. Ed J. Hollox (University of Leicester), Dr. Mike Mullen (Athlone Institute of Technology), Dr. Richard Porter, Prof. David MacHugh (UCD) and Dr Chris Creevey (Aberystwyth University) for their help and advice during my PhD. I would also like to thank Dr. Andrew Cromie (ICBF), Bernard Eivers (NCBC), and Paul Flynn (Weatherbys) for their help in obtaining DNA and phenotypic data for this project.

Teagasc and the Walsh fellowship programme provided financial and training supports. Also, thanks to the Department of Agriculture, Food & the Marine as they provided grant funding for this research.

Finally, special thanks are reserved for my parents, Caitriona and Liam, my sisters, Maureen and Grainne, and nephew Seán, along with all my extended family, friends and loved ones no longer with us for their great support, love, motivation, and patience.

Abbreviations

AAM	Adjusted animal model
AIM	Animal identification and movement
AMP	Antimicrobial peptide
AI	Artificial Insemination
BB	Belgian Blue
BBD	Bovine β-defensin
BLAT	BLAST-like alignment tool
BLAST	Basic local alignment search tool
Bp	Base pair
BSE	Breeding soundness evaluation
Btau	*Bos taurus* genome
BWA	Burrows-Wheeler aligner
CHROM	Chromosome
CI	Calving interval
CM	Cervical mucus
CMP	Cervical mucus penetration
DAFM	Department of agriculture, food and the marine
dbSNP	SNP database
DEFB	β-defensin
DNA	Deoxyribonucleic acid
DPR	Daughter pregnancy rate
EBI	Economic breeding index
EMBL-EBI	European molecular biology laboratory – European bioinformatics institute
FAANG	Functional annotation of animal genomes
FSS	Freedman-Sheldon syndrome
GATK	Genome analysis toolkit
GWAS	Genome-wide association study
HDP	Host defence peptide
HF	Holstein-Friesian
HWE	Hardy-Weinberg equilibrium
IBS	Identity by state
ICBF	Irish Cattle breeding Federation
ID	Identification
InDels	Insertions/Deletions
Kbp	Kilobase pair
LM	Limousin
LAP	Lingual antimicrobial peptide
MAF	Minor allele frequency
Mbp	Megabase pairs

µl	Microlitre
ng	Nanogram
NCBC	National Cattle Breeding Centre
NETs	Neutrophil extracellular traps
Pos	Position
PR	Pregnancy rate
QTL	Quantitative trait loci
SCR	Sire conception rate
s.d.	Standard deviation
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPP	Seminal plasma protein
SURV	Survival
Ts/Tv	Transition/Transversion ratio
UMD	University of Maryland
UTR	Untranslated region
WES	Whole-Exome sequencing

List of figures

Figure 1.1-1: Milk yield and daughter pregnancy rate fertility correlation in Holstein-Friesian dairy cows. 3

Figure 1.1-2: Calf births by month and sire type 5

Figure 1.2-1: Artificial mucus penetration ability of sperm from high and low fertility Holstein Friesian bulls. 21

Figure 1.3-1: Defensin categories and gene processing 24

Figure 1.3-2: β-defensin syntenic map in cattle, humans and dogs. 25

Figure 1.3-3: Main stages of human β-defensin migration in the female reproductive tract, from ejaculation to fertilization 28

Figure 1.3-4: Cervical mucus penetration assay and lectin labelling in humans with DEFB126 dincleotide polymorphism. 30

Figure 2.3-1: Adjusted animal model phenotype values by breed of sire 52

Figure 2.3-2: Average pregnancy rate phenotype values by breed of sire 53

Figure 2.3-3: Roche Nimblegen SeqCap EZ Developer workflow system 60

Figure 2.3-4: Genome analysis toolkit best practice pipeline. 63

Figure 2.4-1: All phenotypic data for adjusted animal model and pregnancy rate between the years 2013 - 2015 for identification of sires for SNP validation 68

Figure 2.4-2: Pregnancy rate fertility values for 2013 - 2015 to identify sires with stable fertility phenotypes for SNP validation 69

Figure 2.4-3: Adjusted animal model fertility values for 2013 - 2015 to identify sires with stable fertility phenotypes for SNP validation 70

Figure 3.3-1: Adjusted Animal model fertility phenotype per breed – data from 7000 sires for sample selection. 77

Figure 3.3-2: Identification of sires divergent for fertility 78

Figure 3.3-3: Phenotypic values of DNA collected from sires in Teagasc's DNA databank. 79

Figure 3.3-4: Percentage of variants located in genomic regions for targeted β-defensin sequencing (all bulls). 83

Figure 3.3-5: Analysis of variants located on chromosomes 8, 13, 23 and 27 associated with adjusted animal model fertility phenotype 85

Figure 3.3-6: Variants located in β-defensin genes on chromosome 13 associated with adjusted animal model phenotype with variants inherited as an haplotype. 86

Figure 3.3-7: Scatterplot of sires identified as being divergent for fertility phenotypes AAM and PR, with those containing haplotype located in chromosome 13 highlighted in red. 87

Figure 3.3-8: Mean coverage per gene in representative sample. 92

Figure 3.3-9: DEFB122a IGV - reads mapping to exons and depth of coverage. 93

Figure 3.3-10: Targeted sequencing SNP location (subset of bulls). 95
Figure 3.3-11: SNP frequency scatter plot between high-fertility and low-fertility groups for targeted β-defensin genes.....96
Figure 3.3-12: O-linked glycosylation analysis in β-defensin genes with predicted glycosylation sites.....99
Figure 3.3-13: O-linked glycosylation analysis in β-defensin genes with no predicted glycosylation sites.....100
Figure 4.3-1: Chromosomal locations for whole-exome variants.....112
Figure 4.3-2: SNP frequency scatter plot between high- and low-fertility groups.....113
Figure 4.3-3: $\chi 2-\chi 2$ plot for a GWA scan. 118
Figure 4.3-4: Principle components resulting from analysis of genomic kinship - identifying genetic outliers .119
Figure 4.3-5: Principle components resulting from analysis of genomic kinship post quality control.....120
Figure 4.3-6: Manhattan plot of whole-exome sequencing variants associated with adjusted animal model fertility phenotype.....122
Figure 5.3-1: SNP frequencies of validated SNPs in independent population of bulls. 139

List of tables

Table 1.2-1: Fixed and Random effects included in adjusted animal model. *15*

Table 1.2-2: Selection of SNPs associated with bull fertility phenotypes in published literature *16*

Table 1.3-1: Possible mechanisms of β-defensin function in reproduction. Possible functions of β-defensins in reproduction in various species. The key papers from multiple species are listed here, with title of the paper, author, and key finding outlined in the abstract for each as a summary. *27*

Table 2.2-1: Gene names and locations of all defensin genes and cathelicidin genes targeted in the custom-designed capture probes for targeted re-sequencing. *43*

Table 2.2-1 continued *44*

Table 2.3-1: AI sire sample selection criteria *53*

Table 2.3-2: Quality control of gDNA from Teagasc DNA databank of sires selected for whole-exome sequencing and targeted sequencing. *57*

Table 3.3-1: Targeted β-defensin sequencing coverage statistics *81*

Table 3.3-2: Targeted β-defensin sequencing alignment and enrichment statistics *82*

Table 3.3-3: Summary of coverage statistics for targeted β-defensin re-sequencing *90*

Table 3.3-4 continued *91*

Table 3.3-5: Top 20 targeted SNPs associated with fertility phenotype. *97*

Table 4.3-1: Summary of coverage statistics for whole-exome sequencing. *107*

Table 4.3-1 continued *108*

Table 4.3-2: Comparison of bull and boar exome sequencing coverage statistics. *109*

Table 4.3-3: SnpEff variant effects annotation predictions for whole-exome sequencing *111*

Table 4.3-4: Top 20 variants with SNP frequency differences between groups of divergent fertility *114*

Table 4.3-5: Gene ontology analysis of SNPs divergent between fertility groups. *116*

Table 4.3-6: Whole-exome sequencing variant genes most associated with adjusted animal model fertility phenotype. *123*

Table 4.3-7: Probes targeting β-defensin genomic region in exome-sequencing dataset. *126*

Table 4.3-8: SNPs identified in both sequencing datasets with overlapping probes. *127*

Table 5.3-1: Validated SNPs sorted by Hardy-Weinberg P-value, showing SNP frequencies for all 123 sires genotyped for validation. *137*

Table 5.3-2: Validation SNPs associated with AAM fertility phenotype. *141*

Table 5.3-3: Validation SNPs associated with PR fertility phenotype. *143*

Table 5.3-4: Validated variants associated with adjusted animal model fertility phenotype. *145*

Table 5.3-5: Validated variants associated with pregnancy rate fertility phenotype. *146*

Index of electronic appendices

Chapter 2: Methods

Electronic Appendix 2.1 Trim_fastqc_map_gatk_filter.pl PERL script to trim fastq files, map to bovine genome, call variants with GATK and filter variants

Electronic Appendix 2.2 snpfiltration.sh Shell script to filter SNPs and to try different filters individually on all snps to see which remove the most.

Chapter 3: β-defensin targeted sequencing

Electronic Appendix 3.1 Defensin_Association_Multiple_Testing.txt Defensin variants associated with AAM including correction for multiple testing.

Electronic Appendix 3.2 Targeted Sequencing coverage Stats - Exome subset.xslx Table of sequencing coverage statistics for 24 bull subset.

Electronic Appendix 3.3 targeted_combined_metrics - sequencing stats - exome subset.xslx Targeted β-defensin sequencing alignment and enrichment statistics for 24 bull subset.

Chapter 4: Whole-exome sequencing

Electronic Appendix 4.1 Results Files/1 - qcbreedcountadjGWAS.txt Table of exome SNPs associated with AAM, sorted by *P*-value.

Electronic Appendix 4.2 R Scripts/ 1 - total_list_Brucerefs_newGATK.pl R script of WES pipeline from FASTQ files to filtered variants via GATK's haplotype caller.

Electronic Appendix 4.3 vcf_to_GenABEL_format.pl Perl script to convert .vcf files to GenABEL format for association analysis.

Electronic Appendix 4.4 count_alleles_by_category_new.pl Perl script to summarise genotype counts per group (high-fertility and low-fertility).

Electronic Appendix 4.5 filter_on_snp_DP.pl Perl script to filter SNPs with low coverage.

Electronic Appendix 4.6 Association commands.R R commands for GenABEL association analysis of WES and TS.

Chapter 5: Validation

Electronic Appendix 5.1 Validation\Assay Design\ 1 - 250_replex2.xslx Probes used to target validated SNPs data including melting temperatures, length and GC content.

Electronic Appendix 5.2 Validation\Assay Design\ 2 - Superplex1.xslx Probe sequences used to target validated SNPs.

Electronic Appendix 5.3 Validation\Assay Design\ 3 - Teagasc plate layout -sample ids.xslx Plate layout of 4 multiplexes to validate SNPs.

Electronic Appendix 5.4 Validation\Association\AAM\ assocofsnps.qassoc A list of all validated SNPs and their associated p-value with AAM.

Electronic Appendix 5.5 Validation\Association\AAM\ out_out.map A map file for validated SNPs containing SNP ID, chromosome and location.

Electronic Appendix 5.6 Validation\Association\AAM\ out_out.ped A pedigree file containing family/pedigree information for each bull.

Electronic Appendix 5.7 Validation\Association\AAM\ validation_AAM_association.xslx

Electronic Appendix 5.8 Validation\Association\PR\pregnancy.qassoc A list of all validated SNPs and their association with pregnany rate.

Electronic Appendix 5.9 Validation\Association\PR\ test.out.preg_out_out.map A map file for validated SNPs containing SNP ID, chromosome and location.

Electronic Appendix 5.10 Validation\Association\PR\ test.out.preg_out_out.ped A pedigree file containing family/pedigree information for each bull.

Electronic Appendix 5.11 Validation\Association\PR\validation_pregnancy_association.xslx A list of all validated SNPs and their associated p-value with PR.

Electronic Appendix 5.12 Validation\validation_snp_ids.txt List of SNPs to be validated

Electronic Appendix 5.13 Validation\SNPs coverage over 80 percent - snp frequencies.xslx List of validated SNPs with 80% genotype calls.

Electronic Appendix 5.14 Validation\SNPs_pass_qc.txt List of Validated SNPs with genotyping rates, and genotypes per bull.

Chapter 6: SNP chip V3

Electronic Appendix 6.1 IDBv3chip_submission_formatEF.xls Table of 864 SNPs added to international dairy and beef (IDB) SNP chip V3.

Summary

Since the realisation of an unfavourable association between fertility and production traits in cattle, methods to select for improved fertility have been the subject of intense research. This research has focused primarily on fertility measures in the cow, and the role of the bull has received comparatively less attention. Significant resources are deployed by the agricultural industry to identify bulls with superior genetics for traits of agricultural interest. Using artificial insemination (AI), these bulls have a disproportionate impact on the genetics of subsequent generations. Despite intense selection of AI bulls and extensive *in vitro* analysis of sperm quality, pregnancy rates can still fall as low as 23% in sub-fertile bulls. While reliable bull fertility data is available on these AI bulls, it is time consuming and expensive to collect, and therefore more accurate measures of bull fertility are urgently required. In addition, 87% of calves born annually in Ireland are sired by a non-AI stock bull. Despite their significant influence on the national herd, the status of stock bull fertility remains unknown. It is known that current *in vitro* fertility tests poorly correlate with field fertility and they fail to identify sub-fertile bulls, resulting in economic losses to the industry. One important method to protect the fertility of the national herd is to identify the genes regulating this important trait, as selecting for these genes is both permanent and cumulative. A small number of studies have identified genetic variants associated with bull fertility; however, significant additional progress is required to understand the genetic architecture of this economically important trait.

Antimicrobial peptides (AMPs), specifically β-defensins, have been shown to have a dual role in host defence against pathogens and in the regulation of male fertility in rodents and in humans. Male β-defensin knock-out mice are completely sterile and a dinucleotide deletion in the human *DEFB126* gene results in a 40% reduction in the probability of conception for couples, if the male partner was homozygous for the variant. Previous research by our group identified an expansion of β-defensin genes in the bovine genome, now estimated to be 57 in total, and functional characterisation documented expression of these genes in the reproductive tract and protein staining of caudal sperm of the bull. However, the association of these genes with fertility in cattle has not previously been investigated. Targeted sequencing (TS) and whole exome sequencing (WES) approaches were used herein to catalogue genetic variation in β-defensin genes, and to identify exome-

wide variants associated with bull fertility. Finally, validation of sequence variants in an independent population of AI bulls was performed.

Based on phenotypic pregnancy rate (PR) records from over 7,000 AI bulls, strict filtering criteria (>1,000 insemination records) were applied to identify the most divergent high- and low-fertility bulls, of which 168 were selected for TS of β-defensin genes and a subset of bulls (n=24) were selected for WES. DNA was sourced and extracted, libraries were prepared by TruSeq (TS) or TruSeq Nano (WES), captured with the Nimblegen SeqCap EZSeq custom capture bait designs for each project and sequenced on a MiSeq (TS) or HiSeq 2500 (WES). For both projects, quality control filtered reads were aligned to the UMD 3.1 version of the bovine genome and variants identified according to GATK Best Practice pipeline. Strict quality filtering criteria were applied. Association analyses between an adjusted animal model of bull fertility and the variant genotypes were performed for both datasets.

Targeted sequencing from 4 chromosomal clusters of β-defensin genes had an average read depth of 197X. Following editing and quality control 2,836 SNPs were identified, 37% of which had not been previously described in cattle. 7.5% of SNPs were found in exons, and 25%, 23% and 22.5% were upstream, intronic and downstream, respectively. A haplotype containing 94 SNPs covering ~138kb was significantly associated with fertility (*P =0.002*). This haplotype spans 8 β-defensin genes, including the bovine ortholog of *DEFB126* which has been shown to play a role in male fertility in other species.

WES had an average read depth of 18X, and following editing and quality control, 144,000 SNPs were identified; 38% located in exons, 21% in introns and 2% in the 5'UTR. The remaining ~40% were in upstream, downstream, 3'UTR and intergenic regions. Association with adjusted animal model fertility phenotype identified 484 SNPs associated with the phenotype (*P* <0.01). This represents the first application of WES from bulls with divergent fertility phenotypes.

SNPs most associated with bull fertility (n=58) were subsequently selected from both datasets for genotyping in an independent population of AI bulls (n=123). The SNP most associated with the adjusted animal model fertility phenotype in the validation analysis was in the *FOXJ3* gene (*P* =0.0016), with a SNP frequency differential between low and high-fertility bulls of more than 20% (low-fertility v high-fertility: 69% v 48%, respectively), and this SNP was the third most associated SNP in WES dataset (*P* =0.0005). *FOXJ3* is a

transcription factor, and a recent publication showed that this gene is required for the survival of spermatogonia in mice. The fourth most associated SNP with fertility in bulls was in the 5'UTR region of the *NOB1* gene, which is overlapping with the SPZ1 testis-specific transcription factor binding site. The SNP most significantly associated with the pregnancy rate phenotype is in the 3'UTR of *DEFB128* (*P* =0.02), which is one of the 19 β-defensin genes found by our group to be expressed in bull reproductive tracts. SNPs in β-defensin genes were also associated with fertility in the validation dataset (*BBD123*, *P* = 0.07; *BBD124*, *P* = 0.06).

This represents the first analysis of genetic variation in the expanded suite of β-defensin genes in cattle. Given the known association of β-defensin genes with somatic cell count, an important indicator of economic importance for mastitis, the identified variants were added to a SNP chip. In total, 863 SNPs discovered by this dual sequencing approach have been added to version 3 of the International Dairy and Beef SNP chip to validate their association in large numbers of cattle and across multiple phenotypes of economic interest. *FOXJ3* and a β-defensin haplotype have been shown to be significantly associated with fertility in bulls. *DEFB126* has been shown in humans to have a polymorphism resulting in decreased ability to penetrate cervical mucus, and *BBD126*, the bovine ortholog has been predicted to contain glycosylation sites, and two variants in *BBD126* are contained in the β-defensin haplotype. In conclusion, this research supports a role for *FOXJ3* and β-defensin genes, specifically, a β-defensin haplotype encompassing *BBD126*, in regulating male fertility in cattle.

1 General Introduction

1.1 Bovine fertility

1.1.1 Domestication of cattle

Cattle were domesticated 8 – 10 thousand years ago (Vigne, 2011) in two separate events, once in the near East and once on the Indian sub-continent. These two domestication events led to two genetically divergent populations, *Bos taurus* and *Bos indicus* (Gotherstrom et al., 2005, MacHugh et al., 1997, Ramey et al., 2013). Domesticated cattle provided resources such as milk and meat and eventually, over time, came under artificial selection pressures to improve production of economically important products (Bovine HapMap et al., 2009, Evershed et al., 2008). Intense artificial selection pressure resulted in the emergence of breeds specialising in different traits, including milk production (e.g. Holstein-Friesian) and meat production (e.g. Belgian Blue). Currently in Ireland, there are over 7 million cattle of both dairy and beef breeds (Department of Agriculture, 2015a). There are over 1,000 cattle breeds worldwide, generally specialising in specific production outputs. As a result, cattle are an important resource to study the genetics of phenotypic variation, although the underlying genetic structure involved in complex traits is largely unknown (Tellam et al., 2009).

In Ireland, artificial selection for milk production has led to a replacement of the predominantly British Friesian breed by the North American Holstein breed, to give a Holstein-Friesian commonly found in dairy herds. In 1977, 10% of the herd was Holstein, which increased to 80% by 1998 (Dillon and Veerkamp, 2001). A corresponding increase in milk production from 5429 kg in 1991 to 5884 kg per cow per year in 2000 was also recorded (ICBF, 2000).

1.1.2 Antagonistic relationship between milk production and fertility

Intensive artificial selection for milk production has increased output with the undesirable side-effect of decreased fertility (Veerkamp and Beerda, 2007, Berry et al., 2014). Due to an antagonistic relationship between milk yield and fertility, a serious decline in fertility traits has occurred across the national herd, see Figure 1.1-1. Through augmented weightings in the national selection index, this has now been corrected and improvements have been detected in recent years (Berry and Evans, 2014). Nevertheless, fertility problems remain

including embryonic loss (Diskin et al., 2011), which have been studied for genetic causes (Killeen et al., 2016) and also herd management practices (Diskin et al., 2006, Diskin and Kenny, 2016). The genetic correlation between milk production and fertility is not in unity, therefore, genetic improvement in both traits should be possible with more accurate genomic variation and improved breeding selection (Berry et al., 2016).

In April 2015, milk production quotas were abolished across the European Union. As a result, the Teagasc roadmap for the dairy sector predicts a 24% increase in the size of the dairy herd in 2020 and a corresponding increase in milk production by 50% (Teagasc, 2016c). This intensive increase in national herd size and production will herald a new focus on any issues that could exacerbate sub-fertility in the national herd.

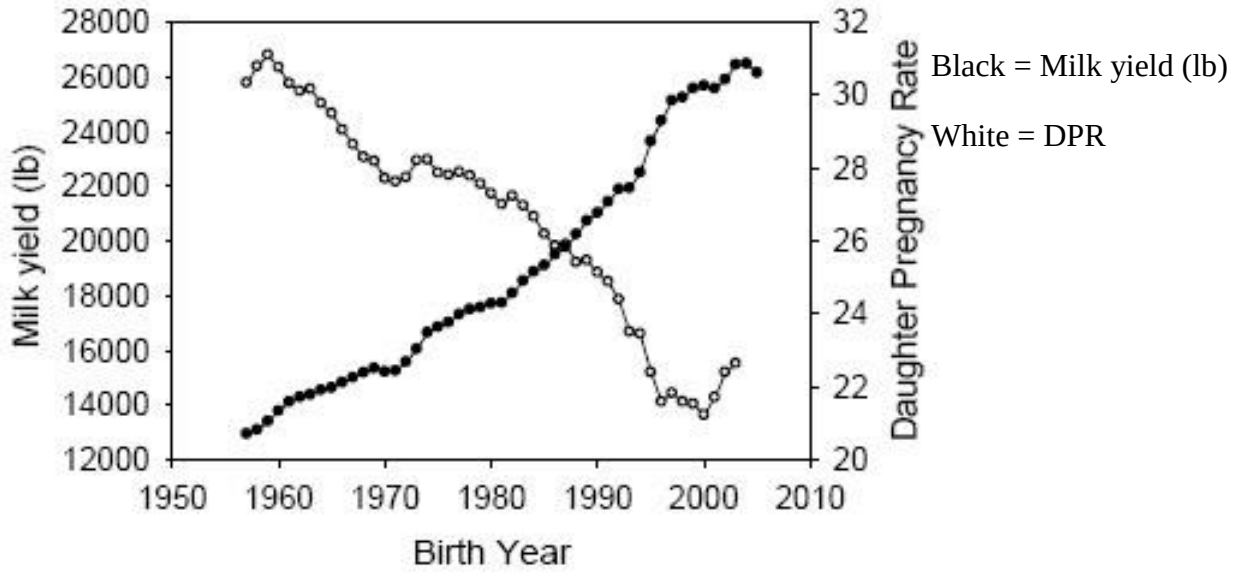


Figure 1.1-1: Milk yield and daughter pregnancy rate fertility correlation in Holstein-Friesian dairy cows.

Figure adapted from: Animal Improvement Programs Laboratory, ARS-USDA (USDA, 2014).

Milk yield (lbs) (left Y-axis) and daughter pregnancy rate fertility data (right Y-axis) from 1950 – 2010 (X-axis) for US Holstein-Friesian dairy cows. These data show an inverse correlation between milk yield and fertility over the decades.

1.1.3 Consequences of lower fertilisation rates

Lower fertilisation rates for bulls results in an extended calving season, as cows fail to become pregnant early in the breeding season (Teagasc, 2016b). Slippages in fertilisation rates mean a serious mismatch between intake requirements and grass (the cheapest food source) availability which significantly increases variable costs on farm. Additionally, where pregnancy does not occur, extra AI services and pregnancy scans result in a less efficient farm system (Shalloo et al., 2004).

One commonly measured female fertility phenotype is calving interval (CI). CI is the number of days between the birth of a calf and the birth of a subsequent calf, both from the same cow. To maintain a compact calving period, a key driver of on-farm efficiency in grass-based production systems, such as in Ireland, a 365-day calving interval is required (Ramsbottom, 2014b). Data from the Irish cattle breeding federation (ICBF) shows that the average CI for the dairy herd in Ireland from 2008 – 2015 was 394 days, which is 29 days outside the Teagasc 2025 target of a 365-days calving interval to maintain compact calving (Teagasc, 2016d). Similarly, for the Irish beef herd, the current calving interval is 407 days, a full 42 days outside the Teagasc 2025 bull beef target (Teagasc, 2016c). Every additional calving interval day costs the farmer €2.20 per cow or €3,100 per year on an average farm of 50 cows (Ramsbottom, 2014a).

The Animal Identification and Movement Bovine Statistics Report 2015 (Department of Agriculture, 2015a) shows a significant discrepancy in month of calving between beef and dairy breeds, see Figure 1.1-2. For dairy breeds, 38% of the annual calves are born in February, rising to 60% including March. This is compared to beef breeds which have just 11% of annual calvings occurring in February, with 32% occurring with March included. A further 21% of calvings occur in April, totalling 53% of calvings. Slippage in calving intervals is hypothesised to be due to a combination of male fertility, female fertility and environmental factors. These data highlight the need to improve calving intervals for dairy and beef breeds to optimise production outputs for farmers.

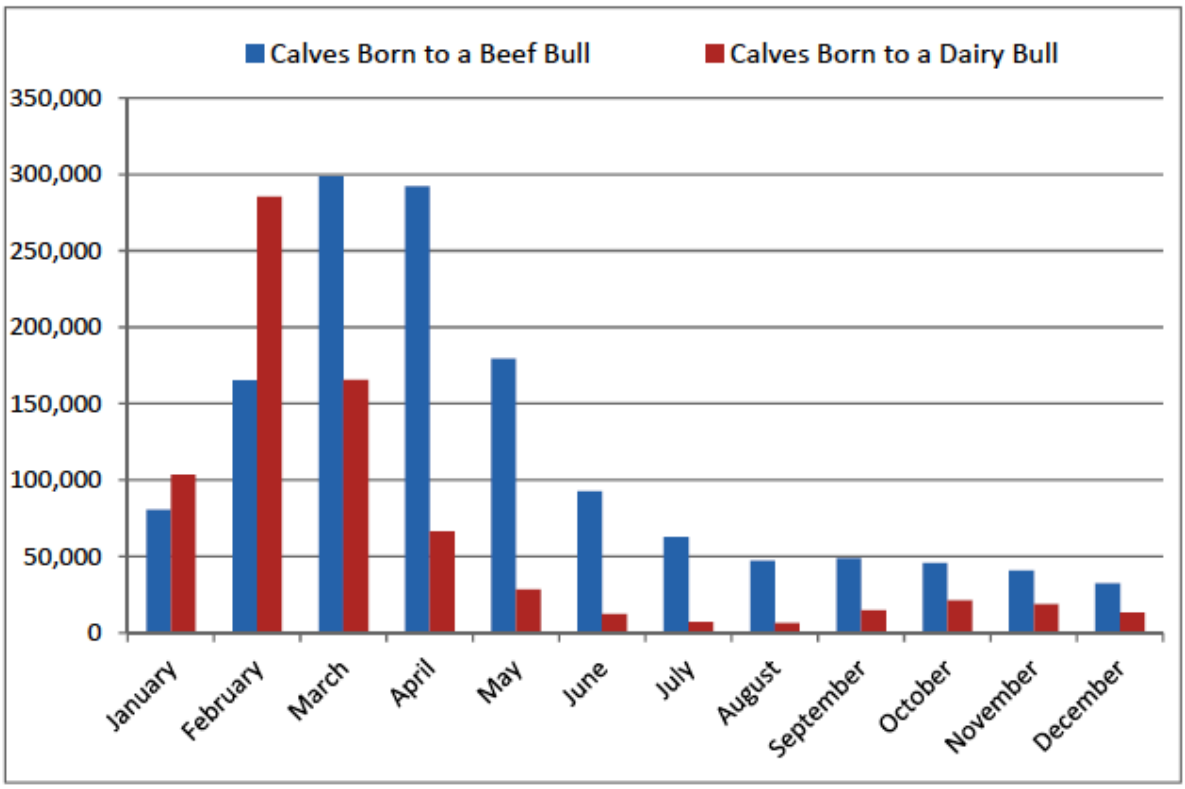


Figure 1.1-2: Calf births by month and sire type

Figure adapted from: Animal Identification and Movement (AIM) System report 2014 (Department of Agriculture, 2014).

The X-axis denotes month of year calf is born, and the y-axis denotes number of calves born. Blue bars indicate calves born to a beef breed sire and red bars denote calves born to a dairy breed sire.

1.1.4 Artificial selection in cattle production

Ireland operates a pasture-based production system for cattle, which utilises the abundant grass growth in Ireland particularly during spring. The Economic Breeding Index (EBI) (Teagasc, 2014), based on a national farm economic model, is used to select sires for breeding by Irish farmers. Prior to 2001 milk production was selected for exclusively. After the introduction of EBI in 2001 fertility phenotypes (survival and calving interval) were included. Survival is the percentage of animals that stay alive during a set period of time (1 year), and calving interval is the period of time (days) a cow takes from calving one year to calving the following year. Currently, fertility traits account for 35% of the dairy index, with two phenotypic traits, calving interval (days) and survival (%), included (Teagasc, 2014). Production traits account for 33%, which shows the change in emphasis in the EBI for Irish farmers to improve genetic gain for fertility in the herd.

1.1.5 Genomic selection

Genomic selection is a form of marker-assisted selection where genetic variants (*i.e.* SNPs) covering the genome can be used to detect quantitative trait loci (QTL) in linkage disequilibrium (LD) with a trait (Goddard and Hayes, 2007). Genomic selection can improve bull selection practices by identifying the bulls of high-fertility and other desirable traits (Amann and DeJarnette, 2012). To improve genomic selection predictions, a better understanding of the genetic variation underlying desirable traits is required to prevent deleterious alleles from becoming prevalent in the national herd. Previous studies have identified congenital, embryonic lethal abnormalities in Holstein cattle which were propagated in the breed affecting fertility (Agerholm et al., 2006). This autosomal recessive disorder was identified in multiple Holstein populations following intensive artificial selection. This highlights the importance of improving genomic selection practices to identify high-quality sires, without affecting overall herd fertility.

1.1.5.1 Balancing selection

In humans, deleterious alleles are typically rare, however, in domestic animals, because of intense selection and reduction in effective population sizes, they have been observed at higher frequencies (Marsden et al., 2016). Intensive selection can expose the negative effects of alleles that are deleterious to fertility and milk production. This is a form of balancing selection where multiple alleles (different versions of a gene) are maintained in the gene pool of a population at frequencies longer than expected from genetic drift alone. This balancing selection has been shown in Nordic red cattle, where a 660kb deletion across 4 genes was identified as a quantitative trait locus (QTL) that is lethal in homozygous embryos. Despite its dramatic effect on fertility, 13%, 23% and 32% of animals carry the deletion in Danish, Swedish and Finnish Red Cattle, respectively (Kadri et al., 2014).

Other examples of balancing selection maintaining a deleterious allele at high frequency in livestock include variants in the BMP15 and GDF9 genes, increasing prolificity in heterozygous females yet causing infertility in homozygous ewes (Hanrahan et al., 2004, Galloway et al., 2000). Similarly, the V700E mutation in the ovine FGFR3 gene increases size in heterozygotes yet causes Spider Lamb syndrome in homozygotes (Cockett et al., 1999).

This balancing selection in livestock species might be more common than previously appreciated. These examples of balancing selection clearly demonstrate how selection for production traits can lead to the maintenance of deleterious alleles for fertility.

1.2 Male fertility

A study in humans found that male fertility factors contribute to ~50% of cases of infertility (Poongothai et al., 2009). Similar fertility data has been observed in multiple livestock species, including cattle (Al Naib et al., 2011). Many factors contribute to male fertility, including genetics, environmental factors and immunology (Behr et al., 2007, Akinloye et al., 2009, Azenabor et al., 2015). Fertility is the ability to produce offspring, and infertility is when this does not occour, which is distinct from sterility, the inability to conceive. The reasons for low fertility relate principally to diseases, poor nutrition, hereditary and congenital factors, hormonal disturbances or environmental changes (Lee and Foo, 2014).

Spermatogenesis is the process which generates sperm cells in the testes first by germ cell formation followed by development into primary and secondary spermatocytes, and finally the production of mature spermatozoa (Azenabor et al., 2015). Spermatogenesis is dependent on optimal environmental factors, as heat and inflammation can have negative effects on spermatogenesis and result in sub-fertility or infertility. Inflammation can be caused by bacterial infection, epididymitis (infection of the epididymis), orchitis (inflammation of the testes) or urogenital obstruction. Inflammation on the male reproductive tract leads to increased pro-inflammatory cytokines in the testes. Pro-inflammatory cytokines, such as tumour necrosis factor-alpha (TNF–α), interleukin–1 alpha (IL–1 α) and interleukin 1 beta (IL–1 β) cytokines in the male reproductive tract play a normal role in regulating infection, but increased cytokine concentrations can be detrimental to sperm production (Azenabor et al., 2015).

Sperm from the testes is ejaculated into the female reproductive tract and swim towards the female oocyte located in one of the fallopian tubes, using the flagellum (Tollner et al., 2012). Sperm then attempt to penetrate the *zona pelucida* of the oocyte and fertilise the ovum using enzymes located in the acrosome region in the sperm head. Successful fertilisation occurs when the haploid DNA from sperm is transferred to the female egg.

In mammals, tens of millions of sperm are deposited in the female reproductive tract, but only dozens reach the female ovum. Cervical mucus (CM) also acts as a natural selection barrier that sperm must traverse on their way to the oocyte. A significant factor determining

the ability of sperm to swim effectively through CM is the presence of attached sugar moieties on their outer membrane. Glycosylation of sperm occurs during spermatogenesis with maturation during transit and capacitation in the epididymis. The addition of sugar residues enhances the negative charge on sperm that repels mucus and enables their passage (Yudin et al., 2005a, Tollner et al., 2008). Interestingly, altered glycosylation on β-defensin genes has previously been documented to retard the ability of human sperm to penetrate cervical mucus (Yudin et al., 2005b). However, little is known of the effects of glycosylation, spermatogenesis, and perturbations to these processes on male fertility in cattle.

Given the focus on female fertility, male fertility factors have been largely unexplored in comparison, despite evidence in humans that 50% of problems in fertility are caused by male factors (Poongothai et al., 2009). Variation in bull fertility data has been demonstrated previously (Berry et al., 2011a), but more evidence of genetic variation and the underlying reasons for bull fertility variation is required. Berry et al. previously identified positive genetic correlation (0.52) between fertility rates in females and males, indicating an improvement in male fertility will correspond to an improvement in female fertility (Berry et al., 2011a).

Male fertility is an important factor in bovine reproduction as a single bull is generally used to breed numerous cows (Peddinti et al., 2008). Approximately 20% of beef calves were sired by beef AI sires in 2014. Importantly, in the dairy sector, 40-50% of calves born from dairy cows were sired by dairy AI sires in 2014 (ICBF, 2014). This highlights the different artificial selection pressures on the beef and dairy herds and the effects of AI usage in the herd.

Following a round of AI, cows which are not pregnant, can then be re-inseminated with a stock bull/sweeper bull, see section 1.2.2.5. Development of a genetic biomarker for bull fertility would aid in identifying sub-fertile bulls earlier. Previous studies have predominantly focused on large-scale association analyses for female-related fertility traits. These studies are discussed in detail below and summarised in Table 1.2-2.

1.2.1 Genetic association studies in cattle

Genetic association studies investigate correlation between a phenotypic and a genetic variable (Lewis and Knight, 2012). Phenotypes can be binary (case-control studies for example) or quantitative (continuous traits such as height or fertility). A genetic variable is a locus on the genome with at least one difference between at least two individuals. The most studied genetic markers are single-nucleotide polymorphisms (SNPs) which are variants at a single locus that usually come in two variants (biallelic) with a minor allele frequency of at least 1%. Minor allele frequency (MAF) refers to the frequency at which the second most common allele occurs in a population. Loci with low MAF have significantly lower power to detect genetic associations compared to high MAF (Tabangin et al., 2009). Single nucleotide variants (SNVs) are variants without limitation on allele frequency. A genetic variant association study which spans the genome of the organism of interest is referred to as genome-wide association study (GWAS) (Ramanan et al., 2012). GWAS can identify candidate genes or regions affecting phenotypes of interest (Gao et al., 2012). This can lead to elucidation of the molecular pathways and processes underpinning the trait being investigated. To perform association tests, genetic variants are correlated with quantitative traits, with single SNP regression being a common method.

1.2.1.1 Single SNP regression

GWAS involves correlation of large numbers of genetic variants identified from a variety of sources (exome and targeted β-defensin sequencing for example), with a phenotype or trait. Single SNP regression is commonly used to associate these variants with quantitative phenotypes (Ziegler et al., 2008). Linear regression is a natural statistical tool for quantitative traits, such as fertility. Linear regression assumes a linear relationship between the mean value of the trait and the genotype. Single SNP linear regression tests require the trait to be approximately normally distributed for each genotype. Log transformations can be used if data are not normally distributed to ensure approximate normality. In addition, single SNP linear regression requires the trait variance to be the same for each genotype (Balding, 2006). However, others claim that heteroscedasticity, minor allele frequency and

sample size are more important factors to consider before phenotype distribution in linear regression analyses (Buzkova, 2013).

A linear mixed model approach allows individual SNP regression allowing for fixed and random effects, such as breed and allele frequency. Additionally, as many thousands of individual statistical tests are performed for each variant identified, strict quality control needs to be performed (Balding, 2006), and correction for multiple testing considered.

1.2.1.2 Linkage disequilibrium

Linkage disequilibrium is the non-random association between two or more regions of DNA that occurs when they are inherited together (Khatkar et al., 2008). Linkage disequilibrium has been shown to be greater in livestock species compared to humans (Bovine HapMap et al., 2009). This is important as the number of SNPs required to cover the genome is less in cattle compared to humans (Matukumalli et al., 2009). This is because humans have a larger effective population size compared to livestock species, as in humans the effective population size is ~ 10,000 (Kruglyak, 1999) whereas in livestock effective population sizes can be as low as 100 (Riquet et al., 1999). Linkage disequilibrium can occur in livestock through migration, mutation, selection, small finite population size or other genetic events which the population experiences[1]. The extent of LD among markers within an interval also reflects selection on the genes within. This is because alleles will increase the frequency in the population of a surrounding segment of chromosome as they are driven toward fixation in selective sweeps.

1.2.1.3 Hardy- Weinberg Equilibrium

The Hardy-Weinberg equilibrium (HWE) principle states that allele frequencies in a population will remain constant from generation to generation in the absence of evolutionary influences. The HWE function is as follows:

$$(p^2 + 2pq + q^2 = 1)$$

[1] https://jvanderw.une.edu.au/genomic_selection_une.pdf

Where p and q are allele frequencies.

HWE assumes the following are true: Organisms are diploid, only sexual reproduction occurs, generations are non-overlapping, mating is random, population size is large, allele frequencies are equal in the sexes, and there is no migration mutation or selection. Deviation from HWE indicates one or more of these basic assumptions have been violated.

1.2.1.4 GWAS in cattle

Previous GWAS have identified genetic variants associated with diverse fertility phenotypes, including sire conception rate, daughter pregnancy rate, age of first calving and cow conception rate. A selection of variants associated with male and female fertility phenotypes are summarised in Table 1.2-2.

Sire conception rate is the expected difference in conception rate of a sire compared with the mean of all other evaluated sires. Genotype data from 1,755 Holstein dairy cattle, and 38,650 SNPs spanning the genome were collected to conduct a GWAS with sire conception rate as the phenotype. Eight SNPs with genome-wide significance with sire conception rate were identified. Some of these SNPs are located close to or in the middle of genes with functions related to male fertility, such as the sperm acrosome reaction, chromatin remodelling during the spermatogenesis, and the meiotic process during male germ cell maturation. SNPs showed dominance effects which indicated the relevance of dominant SNP inheritance on traits such as fertility (Penagaricano et al., 2012).

Genotyping of 10 high- and 10 low-fertility bulls using Bovine SNP Gene Chips containing approximately 10,000 random SNP markers was performed to identify variants in a population of divergent fertility bulls. Of these, 97 were found to be associated with fertility. The 4 most significant SNPs were analysed for allelic discrimination using TaqMan probes in a larger population with 100 high- and 101 low-fertility bulls. Two significantly associated SNPs were identified, one of which causes a synonymous mutation on *integrin beta 5*, located on chromosome 1, and incubation of spermatozoa with antibodies for *integrin beta 5* significantly decreased their ability to fertilize oocytes (Feugang et al., 2009).

An additional study used Bayesian analysis of 795 dairy bulls genotyped with 38,416 SNPs for association with non-compensatory fertility. Non-compensatory fertility data was found to be normally distributed, and that the correlation between true and predicted breeding value for non-compensatory fertility was $r^2 \sim 0.145$, which is to be expected given the low heritability of the trait (Blaschek et al., 2011).

SPAG11 has an important role in male reproductive function. Six SNPs in the *SPAG11* gene in 426 Chinese Holstein bulls were investigated and were found to be in linkage disequilibrium with each other (Liu et al., 2011). Correlation analysis showed one SNP (g.16974C>T) had a marked effect on sperm motility and sperm concentration, whereas another (g.22696T>C) had a significant effect on post-thaw cryopreserved sperm motility and deformity rate (Liu et al., 2011).

Divergent phenotype sampling has previously been used in a candidate SNP genotyping project design as a method to identify alleles that contribute to a trait. (Cochran et al., 2013b). Semen from 550 Holstein bulls were genotyped for 434 candidate SNPs previously identified as being associated with reproductive traits, using the Sequenom MassARRAY® genotyping system. Cochran et al. identified 40 SNPs which were significantly associated with daughter pregnancy rate (DPR). These SNPs were in genes involved in endocrine system, cell signalling, immune function and inhibition of apoptosis. Ten of the genes were regulated by a sex hormone, estradiol, and 29 SNPs associated with DPR were not negatively associated with production traits (Cochran et al., 2013b). This highlighted the possibility of selecting for DPR, a fertility phenotype, without compromising milk production traits.

The discordance between results highlights the different methods used to measure bull fertility in different countries. Berry et al. promote the use of an adjusted animal model as a model of pregnancy rate (Berry et al., 2011a), as a statistical model to better estimate the performance of service bulls. AAM is a multiple regression mixed model of pregnancy rate, where a cow/heifer was confirmed to be pregnant to a given service either by a calving event and/or whereby a repeat service (or a pregnancy scan) deemed the animal not to be pregnant. The model was then adjusted for random and fixed effects, including semen type (frozen, fresh), cow genotype, parity of cow, month of service, day of the week when inseminated, service number, herd, AI technician, bull breed, see Table 1.2-1. AAM is

expected to more accurately represent male-specific fertility, as the male-female interactions are decoupled in the AAM model, due to the fixed and random effects (as outlined in Table 1.2-1) being accounted for. The estimate from the model was weighted for number of insemination records. This gave the AAM, which was expressed relative to the mean of the population for all 7,000 AI bulls. The study identified correlations between rankings of service bulls on male fertility differs when systematic environmental, as well as genetic effects, are accounted for in a mixed model.

Table 1.2-1: Fixed and Random effects included in adjusted animal model.

A list of fixed and random effects included in a statistical model used to better estimate the performance of service bulls. AAM is a multiple regression mixed model of pregnancy rate.

Fixed effects	Random effects
Parity of cow	Cow: Genotype and repeatability
Dystocia in previous calving	Service sire
Stillbirth in previous calving	Sire x year
Calving to service days	Technician x year
Heterosis and recombination of cow and embryo	Day of the week
Service number	
Month of service	
Herd x year of service	
Straw type	

Table 1.2-2: Selection of SNPs associated with bull fertility phenotypes in published literature

This table shows the SNPs associated with a fertility phenotype, preferentially a male fertility phenotype. Each SNP is in or near the gene. All associations have been published in peer-reviewed literature per references. SNP = Ref/Alt allele, Chr = Chromsome.

Gene Name	SNP	Chr	Position	SNP ID	Phenotypic Association	Reference
STAT5a	[C/G]	19	43045807	rs137182814	Sire conception rate	Li, Khatib *et al*. 2009
STAT5a	[G/C]	19	43041479	NA	Fertilization Rate / Age of First Calving	Li, Khatib *et al*. 2009
CAST	[T/C]	7	98485273	rs137601357	Daughter pregnancy rate and cow conception rate	Hansen, Cochran et al. 2013
MAP1B	[G/T]	20	9331992	rs109423562	Sire conception rate	Li, Khatib et al. 2012
CWC15	[A/G]	15	15713532	rs210398455	Decreased reproductive efficiency	Sonstegard et al. 2013
Unknown	[A/G]	13	60468277	ss86288836	Calving to first insemination interval and Fertility Index	Sahana, Lund et al. 2010
FGF2	[G/A]	17	35247483	g.11646A>G	Fertilization rate and early embryonic survival	Wang, Schutzkus et al. 2009
ITGB5	[C/T]	1	69802307	rs41257187	Incubation of bull spermatozoa with integrin beta 5 antibodies significantly decreased the ability to fertilize oocytes.	Feugang, Kaya et al. 2009

1.2.2 Fertility evaluation in bulls

Animal Health Ireland (AHI) state in their biennial report for 2012 – 2014 that "The causes of sub-optimal fertility are complex and varied and addressing these effectively requires a coordinated, holistic and multi-disciplinary response", highlighting the importance of multiple different technologies, methods, and plans to tackle the problem (IRELAND, 2012). Such methods include, but are not limited to, genomic selection, genomics, breeding programs, and improved measures of fertility.

Currently, no single diagnostic test can accurately predict fertility in bulls which produce apparently normal semen (Braundmeier and Miller, 2001). Therefore, the focus is usually on field fertility. While it is the phenotypic parameter that matters most, records are complicated by several factors as outlined below.

1.2.2.1 *In vivo* vs *in vitro* fertility

Differences in the ability of sperm to inseminate eggs *in vitro* vs *in vivo* have been identified (Al Naib et al., 2011). In the study, bulls used in commercial AI breeding, which have passed all semen quality testing evaluations were assessed for pregnancy rate and determined to be high fertility (51 - 54%) or low fertility (23 – 26%). Sperm from low fertility bulls exhibited a reduced ability to penetrate artificial cervical mucus as well as to fertilize oocytes *in vitro*. Sperm which are morphologically normal, are motile, and pass all other criteria, yet have poor performance are of interest. The genes and processes involved in this need to be elucidated further.

In vitro, sperm can fertilize the egg easier compared to *in vivo* fertility trials. An *in vivo* field fertility trial aimed to identify why frozen-thawed semen resulted in a lower fertilization rate. High and low fertility bulls were identified (Al Naib et al., 2011). The ability of sperm to penetrate artificial cervical mucus was assessed. In this study, larger numbers of sperm from high fertility bulls were better able to penetrate cervical mucus, trending towards significance (*P-value* = 0.08), and had an increased ability to fertilize oocytes *in vitro* (Al Naib et al., 2011).

1.2.2.2 Lack of male fertility phenotypes

One limiting factor in evaluating bull fertility is a lack of detailed bull fertility phenotypic data (Carthy et al., 2016). Daughter pregnancy rate (DPR) is a common bull fertility phenotype used. DPR is a measure of a sire's daughter's ability to become pregnant, rather than a measure of his own ability to get cows pregnant. This may result in under-reporting of sub-fertility. DPR does not accurately measure male fertility, as additional effects need to be accounted for, including AI technician and cow health. Statistical models may provide more accurate assessments of male fertility given high-quality phenotypic data. Berry et al. suggest a benefit of using a statistical model to better estimate the performance of service bulls (Berry et al., 2011a). The study identified correlations between rankings of service bulls on male fertility differs when systematic environmental, as well as genetic effects, are accounted for in a mixed model. Sub-fertility may be caused by low libido, sperm quality, sperm quantity, sperm defects, or physical defects affecting bull motility and mating ability (Teagasc, 2016a). Use of sub-fertile bulls will result in low pregnancy rates, an extended calving interval, and increased culling of cows for infertility reasons. Sub-fertile bulls can go undetected in the herd for large periods of the breeding season, unless constant vigilance is maintained. In addition, bull breeding soundness evaluations may need to be performed (Teagasc, 2016b).

1.2.2.3 Multiple measures of fertility - breeding soundness evaluations

Bull breeding soundness evaluations are physical examinations performed by vets or trained evaluators on bulls, ideally 60 days prior to the start of the breeding season. Evaluations comprise of physical examination of the feet, legs, eyes, penis, and testicles. It also includes measurements for scrotal circumference. Semen examination, including analysis of sperm motility and shape are also performed. Additionally, mating ability can be assessed, and an overall classification of "satisfactory" or "unsatisfactory" is given. "Satisfactory" bulls will have passed a minimum threshold, although the test cannot accurately predict sub-fertile bulls, and these tests do not address all aspects of bull fertility. Current semen quality tests estimate viable sperm via live/dead counts, motility, progressive motility, and identify morphological abnormalities (Kastelic and Thundathil, 2008).

1.2.2.4 Compensation

Semen from high genetic merit bulls is in demand to improve the genetic gain for important traits in the national herd. This can lead to a shortage of semen straws available for use in artificial insemination for popular bulls. Bull fertility can be compensatory or non-compensatory. Compensatory fertility means the ability of a bull to impregnate a cow can be improved by increasing the number of sperm injected into semen straws for use in AI. By compensating for sub-fertility by increasing sperm number, fewer semen straws can be produced by bull ejaculate. Therefore, by identifying non-compensatory bulls, the cost-benefit of rearing bulls to maturity for AI will become more favourable for the farmer and the industry.

1.2.2.5 Stock bulls

Natural service is the main breeding strategy for Irish suckler cow herds in Ireland with ~80% of calves born annually sired by stock bulls, according to data from the ICBF on beef breed statistics 2013 (ICBF, 2013). Stock bulls are exposed to less artificial selection pressures compared to AI bulls, which are intensively selected for production traits.

Due to small herd sizes and single-sire mating in Ireland, a stock bull's fertility is of great importance for calving interval, especially in the beef herd and for total number of calvings. While reports of sterility are generally low (~5% in stock bulls), subfertility 20-25% is more common in breeding bulls (Teagasc, 2016a).

Furthermore, the EBI is an important tool for bull selection, as a measure of a bull's progeny's production potential, and does not necessarily indicate the ability of a bull to impregnate cows. Therefore, the identification of genetic markers which affect bull fertility would be important and valuable to farmers and breeders.

Phenotypic databases for stock bulls are not routinely recorded nationwide. Development of a stock bull's phenotype database may lead to improved genetic gain by identifying variants associated with fertility. Until then, AI bulls are the best source of phenotypic data to study male fertility, due to calving records, database management, and male fertility

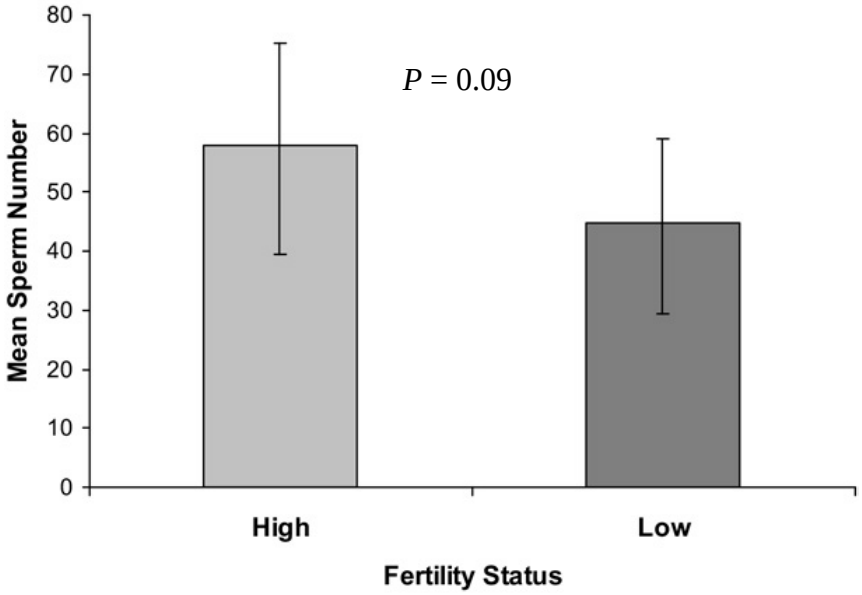
phenotype collection. Phenotypic variation for fertility traits is likely to be higher in stock bulls compared to the highly selected pre-screened AI bulls.

1.2.3 Current status of bull fertility

Many reproductive phenotypes exist for dairy and beef breeds, with female fertility phenotypes shown to be lowly heritable (0.02 – 0.04 units) (Berry et al., 2014), with narrow-sense heritability (h^2) ranging between 0 and 1. The low heritability for female phenotypes does not imply that genetic selection cannot change phenotypic performance, as shown by the decline in dairy cow reproductive performance due to selection for increased milk production. Male reproductive phenotypes are not as widely available, (e.g. semen quality) but are more heritable (0.05 – 0.22 units) for semen related traits (Berry et al., 2014). For male fertility traits, a lack of genetic variance for male fertility has led to suggestions that male fertility cannot be improved using genetics (Berry et al., 2011a). However, only bulls which pass all sperm quality control tests (both microscopy based and computer-aided sperm analysis (CASA), including sperm morphology, sperm motility, and progressive motility), are used in artificial insemination and therefore, bulls with inferior semen quality are excluded, resulting in decreased variability.

Large variations in bull fertility have been documented between elite sires used for AI in Ireland by national cattle breeding centre (NCBC), with recordings of 5% success rate in achieving high-quality breeding bulls. Approximately 400 bulls are selected for *in vivo* breeding trials, with ~20 bulls passing breeding trial selection satisfactorily. While the incidence of complete bull sterility is generally low (<5%), the incidence of subfertility of 20-25% is more common in breeding bulls, with large variation in fertility among individual animals. Al Naib et al. (2011) identified that sperm from high-fertility bulls were better able to penetrate artificial mucus and to have an increased ability to fertilize oocytes *in vitro* (Al Naib et al., 2011), see Figure 1.2-1.

With the use of artificial insemination, where the sperm bypasses the cervix, and therefore the mucus, bulls with sub-fertile sperm bypass the natural selection method of mucus penetration and could decrease national herd fertility.



Adapted from: (Al Naib et al., 2011)

Figure 1.2-1: Artificial mucus penetration ability of sperm from high and low fertility Holstein Friesian bulls.

Mean number of sperm from high and low-fertility Holstein-Friesian bulls penetrating artificial mucus. Fertility status of Holstein-Friesian bulls (X-axis) and the mean number of sperm at each 10 mm point (Y-axis) are shown. The mean number of sperm at each 10 mm point was 56.0 (95% CI 39.5 to 75.3) and 42.9 (95% CI 29.3 to 59.1) for high and low fertility Holstein Friesian bulls, respectively (P = 0.09).

1.2.4 Fertility and the immune system

A number of factors have been shown to play a role in male fertility, including hormonal alterations, diet (Dance et al., 2016), genetic aberrations (Akinloye et al., 2009), infection and inflammation (Choudhury and Knapp, 2001). Upon inflammation of the male reproductive tract, one function of the innate immune system is to recruit phagocytes and effector molecules to the site of infection by releasing cytokines and other inflammatory mediators (Azenabor et al., 2015).

The innate immune system is the body's first line of defence against microorganisms present in the environment, such as bacteria, viruses, and pathogens. Seminal plasma protects sperm during transit through the female reproductive tract, but also contains soluble and exosome-associated cytokines, hormones, and other proteins and factors that affect female reproductive tissues (Robertson, 2005). The cervix is an important regulator of the female genital tract immune response to pathogens and foreign male sperm introduced upon ejaculation. Seminal fluid affects the cervical immune response, inducing proinflammatory cytokine synthesis, specifically TGF-β, and leukocyte recruitment (Sharkey et al., 2012).

In bovines, there is little known about the interaction between the ejaculate and the female reproductive tract, with similarities being drawn from data in other species (Suarez and Pacey, 2006). Bovine sperm are deposited in the vagina, and migrate through the cervix, into the uterus, leaving large percentages of the seminal fluid behind (Alghamdi et al., 2009). The sperm are coated in seminal plasma proteins which may explain how seminal plasma functions in the uterus (Alghamdi et al., 2010). This indicates the seminal plasma has a regulatory effect on the immune response, and neutrophil-sperm interactions may be a physiologically relevant pathway in bovine fertility (Schjenken and Robertson, 2014).

Choline dehydrogenase (*CHDH*) and interleukin 17 receptor B (*IL17RB*) have been shown to be associated with changes in human sperm cell function. A non-synonymous SNP in *CHDH* results in altered sperm motility patterns and dysmorphic mitochondrial structure in sperm (Johnson et al., 2012). Meanwhile a SNP in the coding region of IL17RB results in altered sperm motility characteristics and changes in choline metabolite concentrations in sperm. It has been described that β-defensin genes protect sperm from attack by the female immune

system in the reproductive tract (Amjadi et al., 2014, Tollner et al., 2011). Ultimately, it may be immune genes that play a pivotal role in regulating bovine fertility.

1.3 Defensin genes and male fertility

1.3.1 Defensin family – structure

Defensins are members of a group of molecules called anti-microbial peptides (AMP) and have more recently been referred to as host defence peptides (HDP). They range in size from 2-6kDa, are less than 100 amino acids in length, and exert broad-spectrum antimicrobial activities through membrane permeabilization (White et al., 1995). They function against bacteria, fungi, enveloped and non-enveloped viruses (Jenssen et al., 2006). Antimicrobial peptides have been found in multiple species from diverse taxa such as bacteria (Gao et al., 2009), humans (Ganz et al., 1985, Ganz and Lehrer, 1995) and plants (Thomma et al., 2002).

Defensin peptides are divided into three groups: α, β, θ, and see Figure 1.3-1 on page 24. Each group is distinguished by their disulphide bond conformation. α-defensins are characterised by a disulphide bond between cysteines 1-6, 2-4 and 3-5. The disulphide bonds form a triple stranded beta-sheet structure, characteristic of defensins (Ganz and Lehrer, 1995), whereas β-defensins have a disulphide bond between cysteines 1-5, 2-4 and 3-6 (Ganz, 2003). This defensin motif is conserved across various mammalian species (Lynn and Bradley, 2007). Defensins are expressed by multiple cell types, most notably epithelial cells, and leukocytes (Schneider et al., 2005).

Some primates express θ-defensin which are lectins that have an antimicrobial role, however, neither α-defensins nor θ-defensins are found in cattle (Lynn and Bradley, 2007). The third group, β-defensins play an important role in the innate immune system with various mechanisms of action, including aggregation, pore formation, and prokaryotic membrane depolarization (Sahl et al., 2005). β-defensins are considered the ancestral group, with α-defensins found before placental and marsupial divergence, and θ-defensins only found in primates (rhesus macaque and olive baboon) (Lehrer, 2004).

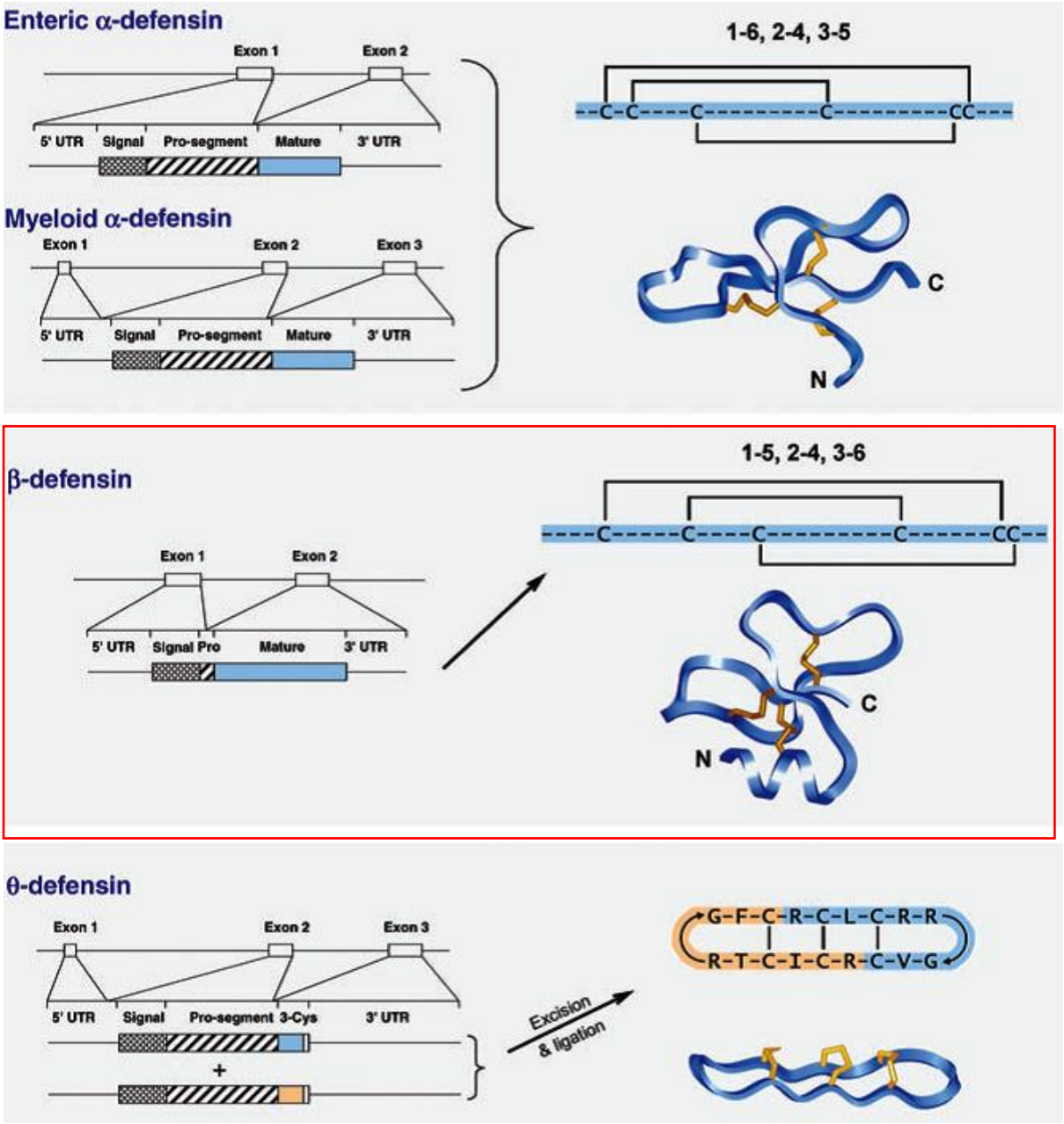


Figure 1.3-1: Defensin categories and gene processing

Adapted from: (Selsted and Ouellette, 2005)

Left, alignment of defensin genes, and 5'UTR, signal sequence, mature peptide sequence, and 3'UTR present in mature defensin peptide. Right, numbers above diagrams indicate disulphide connections. Specific cysteine binding patterns determine whether the defensin peptide is α, β or θ. Folding determines active 3D peptide configuration. α-defensins are formed by 2 or 3 exons; β-defensin genes are formed by 2 and θ-defensins by 3 exons.

1.3.2 Bovine β-defensins – genetic structure and function

The β-defensins are pore-forming cationic molecules that aggregate on the surfaces of bacterial cells to cause cell leakage and death. The β-defensin genes are typically composed of 2 exons. The first exon is translated into the signal and pro-segment. The second exon is translated into the mature peptide and for some genes a section of the pro-segment.

Four β-defensin gene clusters exist in the bovine genome demonstrating an expansion of these genes, covering four *Bos taurus* chromosomes: 8, 13, 23 and 27, see Figure 1.3-2 (Patil et al., 2005).

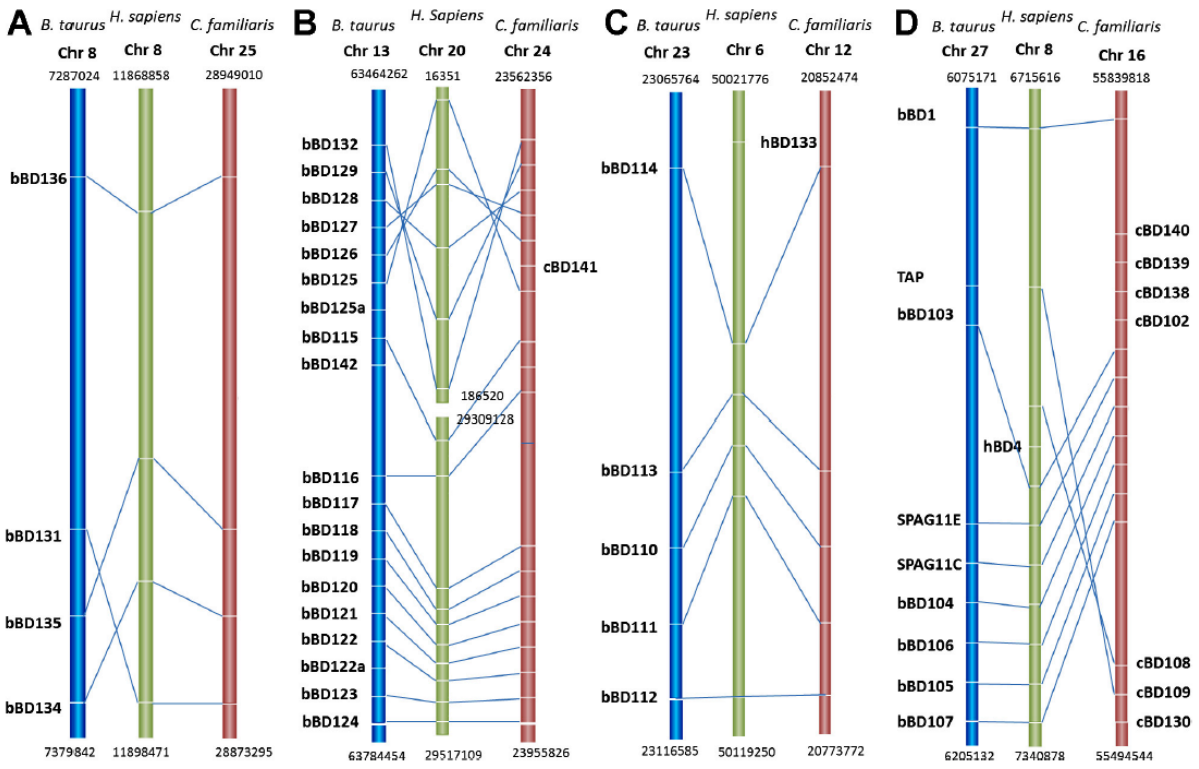


Figure 1.3-2: β-defensin syntenic map in cattle, humans and dogs.

Adapted from Meade et al. (2014)

Syntenic maps of β-defensin genes on four chromosomes in three species in a direct 1:1 relationship. Chromosome 13 contains the 19 β-defensin gene cluster identified by our group as being an expansion of β-defensin genes in bovine, and having expression profile in the reproductive tracts of males and females. Chromosomes are colour coded: *Bos taurus* (blue), *Homo sapien* (green), and *Canis familiaris* (red) (Meade et al., 2014). A) Synteny map of bovine chromosome 8 to human chromosome 8 and canine chromosome 25. B) Synteny map of bovine chromosome 13 with human chromosome 20 and canine chromosome 24. There is evidence of an inversion event in human chromosome 20, between genes *BBD132* and *BBD142*. C) Synteny map of bovine chromosome 23 with human chromosome 6 and

canine chromosome 12. D) Synteny map of bovine chromosome 27, with human chromosome 8 and canine chromosome 16.

A comprehensive bioinformatic search of the bovine genome, by our research group, identified 57 open reading frames with the characteristic six-cysteine spacing of the β-defensin family of genes, and *in vitro* assays showed significant antimicrobial activity of *BBD123* against a range of bacterial species (Cormican et al., 2008). Our group has also previously shown bovine β-defensin genes expressed in healthy male reproductive tracts (Cormican et al., 2008), indicating a dual role of β-defensins in immune response to infection and also reproduction. Expression of β-defensin genes was analysed in multiple tissues, and it was shown that expression of these genes was primarily detected in the reproductive tract (Narciandi et al., 2011). Subsequently, gene expression analysis of all 19 genes in male and 9 genes in the female reproductive tracts determined that a subset of genes were expressed in the adult male reproductive tract and not in the immature (pre-puberty) male, or female, suggesting a possible androgenic regulation of this subset, referred to as class 1. *LAP*, a β-defensin gene not part of the group of 19 genes, was expressed in both male and female reproductive tracts (Narciandi et al., 2011).

The function of these 19 genes in cattle remains unknown, although some antimicrobial activity *in vitro* has been demonstrated showing antimicrobial activity of *BBD123* in bacteria species (Cormican et al., 2008). For a full list of β-defensins being linked to fertility in comparable species, see Table 1.3-1. This table shows that β-defensin genes have been shown to play important roles in male fertility in multiple species.

Table 1.3-1: Possible mechanisms of β-defensin function in reproduction. Possible functions of β-defensins in reproduction in various species. The key papers from multiple species are listed here, with title of the paper, author, and key finding outlined in the abstract for each as a summary.

Species	Title	Author	Key Finding
Human	Deficient human β-defensin 1 underlies male infertility associated with poor sperm motility and genital tract infection.	Diao et al. 2014	Levels of *DEFB1* in sperm from infertile men with either leukocytospermia or asthenozoospermia, both of which are associated with reduced motility and bactericidal activity in sperm, is lower compared to sperm from fertile men.
Human	A common mutation in the defensin *DEFB126* causes impaired sperm function and subfertility.	Tollner et al. 2011	A two-nucleotide deletion in the open reading frame in *DEFB126* generates abnormal mRNA.
Macaque	Macaque sperm coating protein *DEFB126* facilitates sperm penetration of cervical mucus.	Tollner et al. 2008	*DEFB126* and its high negative charge appears to be critical for the movement of sperm through CM in the macaque, while SPPs adhered to the sperm surface offer no advantage in CMP.
Macaque	β-defensin 126 on the surface of macaque sperm mediates attachment of sperm to oviductal epithelia.	Tollner et al. 2008	Treating Macaque sperm that result in alterations to *DEFB126*, result in loss of sperm-OEC binding that is independent of sperm motility. *DEFB126* may be involved in forming a sperm reservoir in the oviduct of Macaques.
Mouse	Partial deletion of chromosome 8 β-defensin cluster confers sperm dysfunction and infertility in male mice.	Zhou et al. 2013	β-defensins were shown *in vivo* to be essential for sperm maturation, and disruption leads to altered intracellular calcium, spontaneous acrosome reaction and male infertility.
Rat	The epididymis-specific antimicrobial peptide β-defensin 15 is required for sperm motility and male fertility in the rat (Rattus norvegicus).	Zhao et al. 2011	Knock down of *Defb15* led to a reduction in fertility and embryonic development failure. Recombinant *Defb15* showed antimicrobial activity in a dose-dependent manner.
Cow	Reproductive tissue-specific expression profiling and genetic variation across a 19-gene bovine β-defensin cluster.	Narciandi, Lloyd et al. 2011	Tissue-specific expression in the epididymis and fallopian tube suggest a reproductive-immunobiology for β-defensins in cattle.

1.3.3 Role of *DEFB126* in fertility in multiple species

In humans, *DEFB126* facilitates movement of sperm through cervical mucus (Tollner et al., 2012). *DEFB126* is a peptide that covers the entire sperm surface, conferring a negative charge to the sperm. A net negative charge facilitates movement of sperm through the negatively charged cervical mucus of the female reproductive tract. *DEFB126* also facilitates protection to the sperm from the female immune system in the reproductive tract via the long, glycosylated tail (Liu et al., 2013). Once sperm reach the oviduct, they form a reservoir, by binding to the oviductal epithelium via *DEFB126* until ovulation occurs. Upon ovulation, a change in pH results in the release of *DEFB126* from the sperm surface, which frees the sperm to continue migrating towards the egg. During this final swim, the sperm are capacitated via a reduction in glucose levels, which enables the sperm to penetrate the hyaluronan matrix, embed in the zona pelucida, release acrosome enzymes, and fertilize the egg (Tollner et al., 2012). The main stages of *DEFB126*-facilitated movement of sperm in humans are shown in Figure 1.3-3.

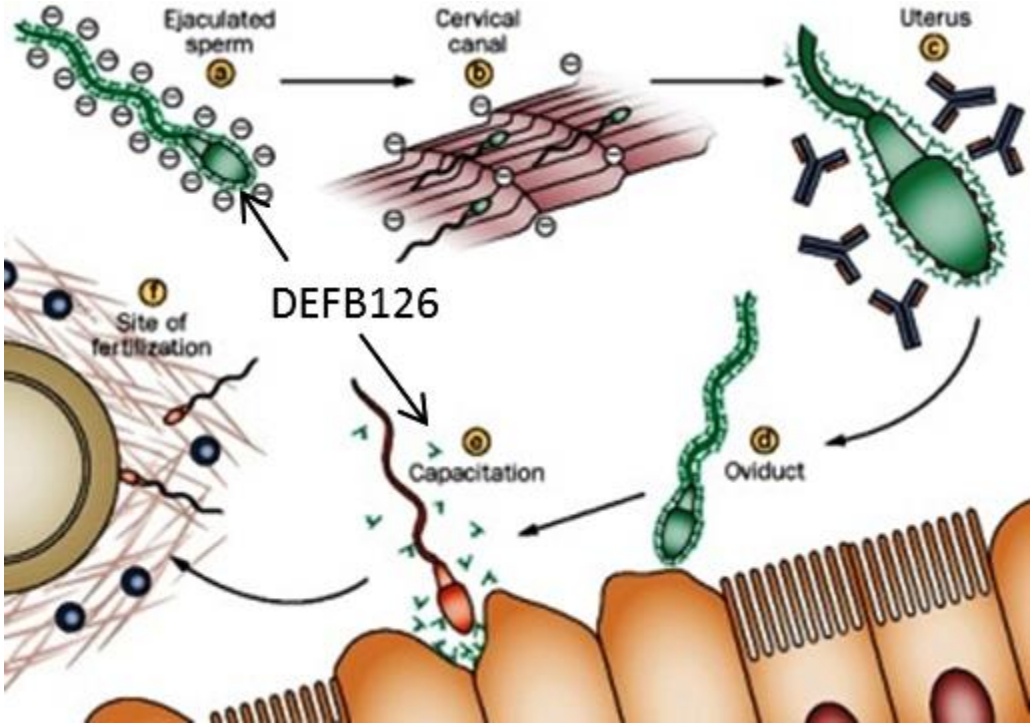


Figure 1.3-3: Main stages of human β-defensin migration in the female reproductive tract, from ejaculation to fertilization

a) Ejaculated sperm is coated with *DEFB126* b) Sperm surface is negatively charged c) Sperm with *DEFB126* coating in the uterus might provide protection from the innate and adaptive

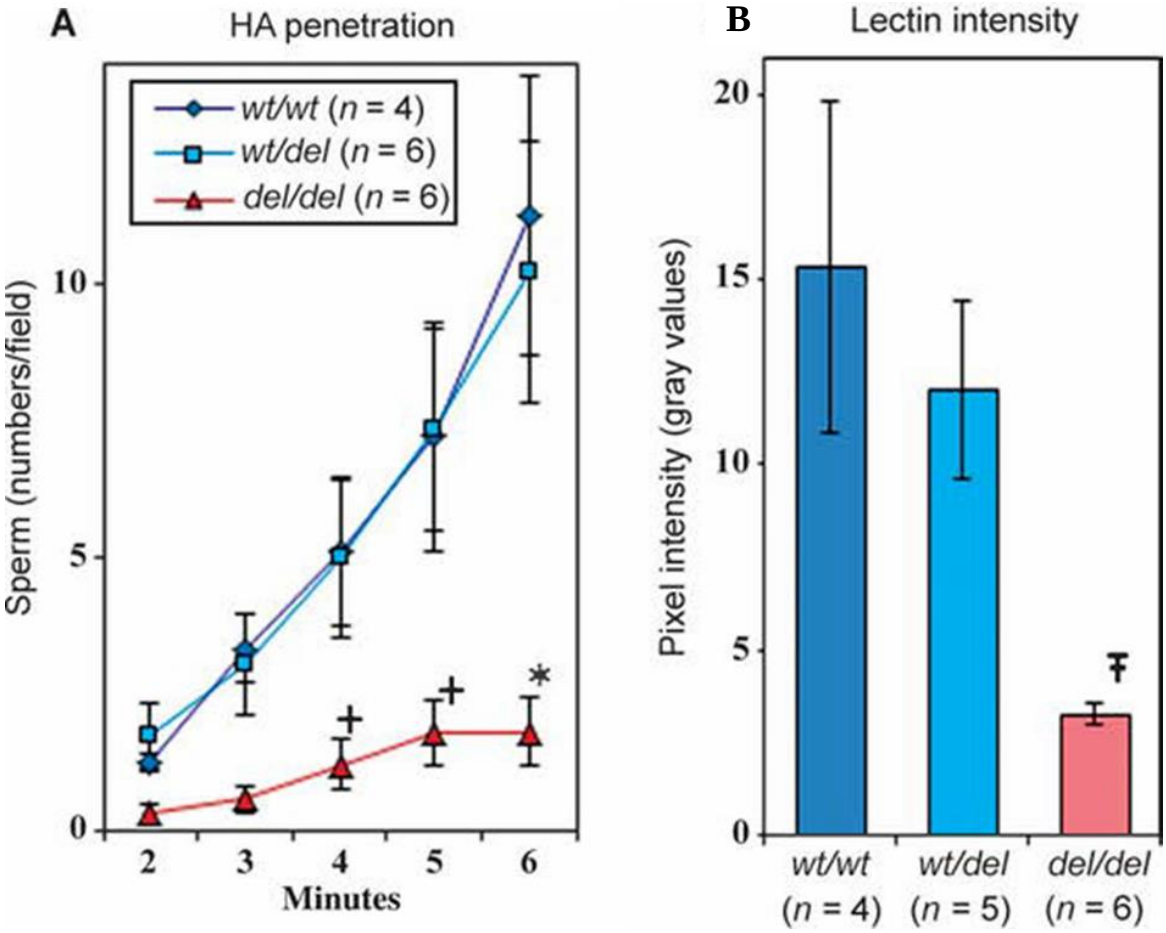
immune systems. d) *DEFB126* mediates attachment to the oviductal epithelium. A reservoir of sperm is formed as more sperm migrate into the oviduct and bind to the epithelium. e) During ovulation, an elevation in pH of oviductal fluid triggers the release of *DEFB126* from the sperm surface. f) Once free of *DEFB126* in the oviduct, sperm are able to migrate to the site of fertilization. Elevated bicarbonate and reduced glucose levels in oviductal fluid complete capacitation. With surface components now unmasked, such as the hyaluronidase PH20 and receptors for the egg, sperm can penetrate the hyaluronan-rich cumulus matrix and bind to the zona pellucida of the oocyte. Adapted from Tollner et al. (2012).

A dinucleotide mutation in the *DEFB126* coding region (exon) results in a predicted frameshift and a reading frame lacking an in-frame stop codon. This non-stop abnormal mRNA results in mRNA from individuals homozygous for the mutation showing lower expression in the epididymis, compared to wild type individuals (Tollner et al., 2011). Sperm from the del/del donors exhibited an 84% reduction in the rate of penetration of a hyaluronic acid, see Figure 1.3-4. The allele frequency of this variant sequence is high in both a European (0.47) and a Chinese (0.45) population cohort. In a cohort study, couples were 40% less likely to become pregnant and took longer to achieve a live birth if the male partner was homozygous for the variant sequence (Tollner et al., 2011).

Physiological regulation of sperm transport and function of cervical barrier are in part guided by glycosylation changes (Pluta et al., 2011). Previous work by our group identified longer c-termini in a group of innate immunity-related genes (Narciandi et al., 2011) and glycosylation sites were predicted. Immunoprecipitation of sperm against immune recognition in the uterus is mediated by sialic acid (Yudin et al., 2005a) and differential glycosylation levels may mediate immunoreactivity and lower fertility. Sperm from the del/del donors in *DEFB126* have lower lectin binding which is associated with fewer O-linked oligosaccharides (Tollner et al., 2011). In addition, lectin (sugar-binding proteins) binding to the sperm surface glycocalyx was significantly lower in men with the homozygous variant (del/del) genotype than in those with either a del/wt or wt/wt genotype, suggesting an altered sperm glycocalyx with fewer O-linked oligosaccharides in del/del men, see Figure 1.3-4.

There is also evidence that β-defensins play a role in fertility in other species. In macaques, treatment of sperm with antibodies raised against *DEFB126*, resulted in significant inhibition

of the ability of sperm to penetrate cervical mucus. Addition of other Seminal Plasma Proteins (SPP) resulted in no significant inhibition. However, sperm that had DEF126 added back to coat the sperm surface showed restoration of the ability to penetrate cervical mucus (Tollner et al., 2008). This indicates that *DEFB126* is critically important for the ability of sperm to penetrate cervical mucus.



(Tollner et al., 2011)

Figure 1.3-4: Cervical mucus penetration assay and lectin labelling in humans with *DEFB126* dincleotide polymorphism.

A) Hyaluronic acid penetration assay - dinucleotide deletion in *DEFB126* resulted in an 84 % decrease in ability of sperm to swim through hyaluronic acid (a synthetic substitute for cervical mucus). Mutants (del/del) had significantly reduced HA penetration ability P = 0.008, compared to homozygous wild-type (wt/wt) and heterozygous (wt/del). B) Lectin intensity - ABA lectin labeling intensity. Ŧ indicates significant differences at P = 0.030, P = 0.008, and P = 0.0006, respectively.

1.3.4 Role of other β-defensins in fertility in rodents

In mice, deletion of 9 β-defensin genes on *Mus musculus* chromosome 8 results in dysfunctional sperm and infertility. All nine deleted genes and their human orthologs were most strongly expressed in the male reproductive tract. Four of the knockout genes were orthologous to the human genes *DEFB1, DEFB106*, *DEFB105*, and *DEFB107*. The remainder were related paralogues. Knockout male adult mice (-/-) resulted in no offspring, compared to heterozygotes (-/+) and wild-types (+/+) and their sperm had reduced motility. In contrast, adult female knockout mice showed no significant reduction in litter sizes, compared to heterozygotes, or wild-type individuals (Zhou et al., 2013). In addition, the authors also posited a reason for the reduced fertility, as sperm from *Defb9/Defb9* (-/-) knockout male mice have disrupted microtubule structure. This indicates that deletion of β-defensin genes resulted in a reduction of fertility specifically in males, suggesting a role for β-defensins in sperm function and fertility.

In rats, epididymis-specific β-defensin 15 (*Defb15*), the ortholog of which in humans is *DEFB106a,* exhibits an androgen-dependent expression pattern. Similar to *DEFB126* and its ability to bind to sperm surface, *Defb15* can bind to the acrosomal region of caput sperm (sperm from the head region of the epididymis). *Defb15* knockdown via RNAi results in reduced sperm progressive motility and total motility. In addition, knockdown led to a reduction in fertility, by a reduction in number of foetuses and offspring in knockdown individuals, compared to wild-type. Knockdown individuals had abnormal embryonic development, and recombinant *Defb15* showed antimicrobial activity in a dose-dependent manner, which indicates a dual role for *Defb15* in both an antimicrobial innate immune function and an epididymal reproductive function in male rats (Zhao et al., 2011).

1.4 Bioinformatics

1.4.1 Bioinformatics in bovine research

Bioinformatics has been described as the use of computers and computational tools to process, store and analyse biological data. Bioinformatics encompasses multiple disciplines and skills, including, database management, programming, genome annotation, pathway analysis, statistical analysis and graphic display of biological data. To answer the key biological questions, the analysis of high-throughput next generation sequencing data, curated databases of known genetic variants (Sayers et al., 2012), and gene annotation (Cingolani et al., 2012) and functional prediction (McLaren et al., 2016) tools are required. Here, sequencing of the bovine genome (Bovine Genome et al., 2009) for improving bovine research, the variant databases containing identified genetic variants in cattle (Sherry et al., 2001), and a next-generation sequencing method (exome sequencing) are described in greater detail.

1.4.2 The bovine genome

Following the publication of the draft human genome from the human genome project in 2001 (Lander et al., 2001), sequencing technologies have dramatically increased while the corresponding cost of sequencing has significantly decreased. The publication of the bovine genome (Bovine Genome et al., 2009, Tellam et al., 2009) showed that it is ~2.86 Gbp (2.86 billion base pairs) long and it contains ~22,000 genes. The genes are located on 29 autosomes and 2 sex chromosomes. L1 Dominette 01449 was the single Hereford cow used as the reference bovine genome in the sequencing project (Bovine Genome et al., 2009). The bovine genome provides an important resource to study genetic variation in cattle, to understand mammalian evolution and to perform genomics studies related to cattle health and fertility (Berry et al., 2011b, Tellam et al., 2009).

Multiple assemblies of the bovine genome have been conducted with Baylor College of Medicine publishing their first assembly, Btau, in 2009. The current version, Baylor Btau_4.6.1, is available at the University of California Santa Cruz (UCSC) genome browser (Kent et al., 2002), and contains 43.3Mb of Y-chromosome sequence.

In an independent assembly project, University of Maryland, USA, developed the UMD assembly. UMD used paired-end sequence data and orthologs in the human genome to create an assembly of 2.86 Gbp. The current UMD assembly is UMD 3.1.1 available at the UCSC genome browser (Kent et al., 2002). UMD3.1 has a higher N50 contig size compared to Btau4.0. N50 is the length N for which 50% of all bases in a number of sequences are in a sequence of length L >= N.

1.4.3 SNP databases

Single Nucleotide Polymorphisms (SNPs) are the largest group of genetic variants found in vertebrates. The importance of identifying genetic variation in species and individuals within the species has been shown previously for breeding and genetic improvement (Gao et al., 2012). Across species there is little known on gene conservation with low to moderate effects on a multi-trait phenotype, such as fertility (Tellam et al., 2009).

Following the publication of the bovine genome, the bovine HapMap project attempted to identify genomic structure in the bovine, by analysing 37,000 SNPs in *Bos taurus* and *Bos indicus* breeds. HapMap analysis of 497 cattle from 19 diverse breeds identified a rapid decrease in effective population size possibly due to bottlenecks associated with domestication, selection, and breed formation. Yet, the levels of diversity within cattle breeds are at least as large as exists within humans (Bovine HapMap et al., 2009).

The development of the HapMap project showed the importance of genetic variation in understanding bovine evolution, helped identify markers associated with phenotypes of interest and has also led to the development of high-density genotyping arrays, such as the BovineSNP50 beadchip (Matukumalli et al., 2009).

The SNP database, dbSNP, is a resource of 53 organisms with variant annotation on their genomes available for web search and FTP download (Sherry et al., 2001). As of June 2014, there were 40 million validated RefSNP Clusters (rs numbers), with over 293 million submissions (ss numbers) for the bovine genome. Although this is actually low compared to humans, with over 154 million validated RefSNP Clusters (rs numbers), with 545 million

submissions, indicating large numbers of variants in the bovine genome are yet to be discovered, and their function characterised.

1.4.4 Exome sequencing

Extreme phenotype sequencing is an efficient method to capture genetic variation in important, functional genomic regions at relatively low cost compared to whole-genome sequencing (Perez-Gracia et al., 2010, Barnett et al., 2013). Exome sequencing captures and sequences only the protein coding exons in the genome. This method, while only sequencing a small portion of the genome, covers important, functional areas with variants which can cause synonymous or non-synonymous variants. A synonymous variant is a change in the DNA sequence that codes for amino acids in a protein sequence, but does not change the encoded amino acid. A non-synonymous variant is a DNA variant that does change the encoded amino acid. Synonymous mutations have been shown to have important effects, e.g. a synonymous change in a multidrug resistance gene *MDR1* where the use of a synonymous codon was proposed to use a rarer tRNA which meant that the protein folded in a non-native fashion (Kimchi-Sarfaty et al., 2007).

Whole-exome sequencing was first developed in 2008 as a method to selectively target exonic regions to allow for identification of coding variants in an individual with minimal cost (Ng et al., 2008). In the first exome study of a Mendelian disorder eight HapMap individuals from three different populations, and four unrelated individuals, with a rare dominantly-inherited disorder, Freedman-Sheldon Syndrome (FSS), were sequenced (Ng et al., 2009). They showed that exome sequencing accurately identified candidate genes for a Mendelian disorder in a small number of affected individuals. More recent studies have utilised whole-exome sequencing strategies to identify Mendelian diseases in humans (Chen et al., 2013, Choi et al., 2009, O'Roak et al., 2011, Yan et al., 2011).

At the time of writing, three studies have attempted to sequence the bovine exome to varying degrees (McClure et al., 2014a, Cosart et al., 2011, Hirano et al., 2013). It has been shown that exome capture arrays (covering 2,570 genes) can be applied to domestic and wild species (Cosart et al., 2011). In this study, 73% of targeted bases had 10X coverage,

with 54% having 20X coverage for *Bos taurus*. Exome sequencing has previously been used to sequence an individual. This study reported that 95% of the target region was sequenced with high sensitivity and specificity for detection of homozygous and heterozygous variants (Hirano et al., 2013). A previous study has performed whole-exome sequencing to identify causative variants for underlying defective bovine embryo development contained within three haplotypes in Holstein and brown Swiss breeds (McClure et al., 2014a). However, whole-exome sequencing of males with divergent fertility phenotypes has not been published to date.

In a recent study, Robert et al. (2014) performed whole-exome sequencing of 96 boars and found that 72 of 96 samples had at least 10X coverage for more than 90% of the targeted bases, see Table 4.3-3. 236,000 SNPs, and over 28,000 InDels were identified. The boar whole-exome sequencing project used the same Roche Nimblegen EZ Developer Library as used by this bull whole-exome sequencing project. Comparisons between the two methods, results, and designs are highly beneficial when using a novel methodology, such as this. Robert et al. designed probes covering 98.4% of bases in the boar targeted region, which was 60.6MB in size. This compares to 56.7MB of targets in this bull exome, with 99.1% coverage of the target design.

Taken together, these results illustrate the value of the SNP discovery to identify genetic variants and association with fertility phenotypes. Exome sequencing is one method of performing SNP discovery. Sequencing studies of complex traits can be limited due to sample size. Divergent phenotype sequencing is a way of overcoming this problem because allele frequencies that contribute to the trait are enriched in one or both groups (Emond et al., 2012, Barnett et al., 2013). Utilising whole-exome sequencing in a divergent phenotype population will allow application of research to the farming community by contributing to the identification of genetic variants and biological processes underlying sire fertility. These findings can provide opportunities for improving bull fertility via marker-assisted selection, amongst other applications.

1.5 Applications of research

1.5.1 Biomarkers

This project aims to improve identification of biomarkers for selection of animals with increased fertility and understanding of the immunological and reproductive role of these molecules. In boars, for example, biomarkers present in spermatozoa after capacitation have been shown to help identify increased male fertility from below-average fertility boars with high-sensitivity (Kwon et al., 2015). This biomarker-based approach to identify male animals of increased fertility would be beneficial in animal breeding programmes of various species, especially bovine.

Recent developments of a beef breeding index have helped improve breeding strategies, allowing breeders identify traits they consider desirable in their herd, allowing greater control over their breeding program. Developing an accurate, efficient diagnostic tool for bull fertility would result in significant improvements in bovine fertility. In terms of costs to the farmer, poor bull fertility is a major contributing factor, together with female fertility. By identifying bulls with high-and low-fertility phenotypes early, significant savings can be achieved.

Improvements in bull phenotypic records are also encouraged to improve association analysis and to remove the reliance on female fertility traits, such as pregnancy rate, to study bull fertility. By directly recording male characteristics in a large population of bulls, genome-wide association analysis studies would be more accurate, and improve candidate SNP identification.

Genetic variants identified in this study could also lead to improved breeding targets. Following association of identified variants using a customised SNP-chip for bulls, genetic targets which improve or decrease fertility can be used to guide the breeding bull selection process and improve genetic gain for fertility traits without decreasing milk production.

1.5.2 National genotyping scheme for cattle

The beef data and genomics programme (BDGP) 2015-2020 (Department of Agriculture, 2015b) was launched by the Irish department of agriculture, food and the marine (DAFM) to genotype a large number of beef cows for inclusion in a genomic selection breeding program to advance genetic gain in beef cattle. The BDGP aims to support the suckler herd by improving the genetic merit of the national herd through the collection of phenotypic data and genotyping animals and to improve quality and efficiency.

A phenotype (P) is a combination of an individual's genetics (G) and environment (E) (P = G + E), therefore, by obtaining genotypic information related to important phenotypic traits, and modelling the environment, an accurate prediction of the genetic gain for an individual and their progeny can be made.

The BDGP was initiated off the success of the dairy cattle genomics programme. Genomic selection of young bulls, primarily from dairy breeds, was launched in Ireland in 2009. The national genotyping scheme for dairy cattle has led to improvements in milk production, without a decrease in fertility, although further research is required to ensure selection of sires based on genomic selection doesn't adversely affect health traits. Additional markers (SNPs etc.) need to be incorporated to achieve this, and as an hypothesis, β-defensin variants may be used to improve fertility without adversely affecting health or production, as β-defensins have an innate immunity function in bovine (Cormican et al., 2008).

1.5.3 IDB SNP chip – large GWAS dataset

The International Dairy and Beef (IDB) SNP chip was developed by Teagasc as a low-cost custom genotyping panel for the dairy and beef breeding industries. The SNP chip is being utilized for genetic evaluations, parentage verification and screening for lethal recessives, congenital disorders and other mutations with effects on performance in cattle (Mullen. et al., 2013). Deleterious recessive mutations have been linked with inbreeding depression, the reduced survival and fertility of related individuals (Charlesworth and Willis, 2009). IDB SNP-chip version 1 contained 9,973 variants added to the Illumina low density genotyping platform, with 5,500 SNPs for imputation to higher density genotypes (Mullen. et al., 2013).

Version 2 contained approx. 17k SNPs (McClure et al., 2014b). Version 3 contains approximately 50k SNPs, with ~25,000 for imputation. It is planned that ~330,000 animals will be genotyped using this SNP chip in 2016. Genotyping will be of various breeds, and both cows and bulls. However, a large proportion will be genotyped in bulls, which will allow for association analysis of genotypes from variants identified via literature review for version 2 or targeted β-defensin sequencing and whole-exome sequencing for version 3.

1.6 Aims

One of the overall aims of this study was to improve bull selection practices by identifying candidate variants and genes involved in regulating male fertility which may be used in the future as biomarkers for bull fertility. To achieve this overall aim, several inter-linking collaborative projects have been performed, two of which are described in this thesis. In reference to this project, the aims helped to identify genetic variation in a divergent population of animals and to associate the variants with phenotypes of interest.

1.7 Hypothesis

We propose that genetic variation in exons and promoter regions of β-defensin genes explain some phenotypic variation in AI bulls divergent for fertility. We also propose that genome-wide genetic variation of the exome and promoter regions will explain some of the phenotypic variation in AI bulls divergent for fertility.

1.8 Objectives

The specific objectives of this project were:

1. To identify variants in bovine β-defensin genes and promoter regions in AI bulls of two groups divergent for a fertility phenotype. This objective is addressed in Chapter 3.
2. To identify whole-exome variants and promoter region variants in a subset of animals from objective 1. This objective is addressed in Chapter 4.
3. To validate candidate variants identified in variant identification, from objective 1 and 2, in an independent population of AI bulls. This objective is addressed in Chapter 5.
4. To add variants to the IDB SNP-chip for future variant association to a fertility phenotype in the national herd and improve genetic gain for fertility. This objective is covered primarily in Chapter 5.

2 Materials and methods

2.1.1 Phenotypic data

Fertility data for 7,000 AI bulls were obtained from national cattle breeding centre (NCBC) over four years (2010 – 2013). Two fertility phenotypes were examined, pregnancy rate and adjusted animal model. PR in this study was defined as a binary score (1 or 0), where 1 is an assumed pregnancy following insemination via AI by a trained technician unless proven otherwise via scanning or calving records. The AAM is a statistical model used to better estimate the performance of service bulls. It is based on the PR phenotype data, but also models environmental factors, including random and fixed effects, such as the AI technician's ability, date of insemination (i.e. time of year), health of the cow and day of the week, parity of the cow, as well as other environmental parameters, which have been shown to affect the ability of a cow to conceive (Berry et al, 2014). A ful list of fixed and random effects are shown in Table 1.2-1. Data from all recorded artificial inseminations were provided by the National Cattle Breeding centre (NCBC).

2.1.2 Sample selection

Divergent (extreme) phenotype sample selection was employed, as the frequencies of alleles that contribute to the phenotype of interest will be enriched in one or both phenotype groups (Barnett et al., 2013). Bulls with divergent high-fertility and low-fertility were identified based on PR and AAM, identifying two groups of bulls that are divergent for both fertility phenotypes (PR and AAM). Sires with the highest reliability phenotype (>1,000 inseminations) were retained, and divergent phenotypes were defined as +/- 1 standard deviation from the mean for PR or AAM. Inseminations were assumed successful unless proven otherwise. Further filtering of bulls required that less than 25% of sire fertility data sourced from 2013 (this was due to the lack of calving records for these bulls at the time of sample selection). Bulls were required to be in the divergent phenotype category for the most number of years possible. A pedigree analysis identified bulls related to each other and bulls with the lowest genetic similarity were preferentially selected. In addition, the breeds selected were at least 85 % purebred for their respective breeds.

2.1.3 Probe design – β-defensin and WES

Probes were designed according to the manufacturer's instructions[2]. The Nimblegen SeqCap EZ Developer Library from Roche (Roche NimbleGen, Inc.) was used to perform a custom-designed capture of the whole bovine exome and separately, a targeted region of genes containing β-defensin genes. Two separate sequencing projects were performed as the bovine β-defensin genes were not annotated in bovine at the time. The EZ developer kit is intended for targeted capture of any animal genome, and provides up to 2.1 million oligonucleotide probes to capture the targeted region. It is intended for non-human applications, as a separate, human-specific whole-exome kit is also available. It is based on the same technology as the human-specific SeqCap EZ Exome library. However, the developer library is fully customisable by the user, meaning specific regions of interest can be included in the probe design step, such as 5'UTR regions, 3'UTR regions, or intronic regions.

Oligonucleotide probes are magnetically labelled, allowing capture of targeted regions during library preparation. The bovine UMD3.1 Ensembl version 70 was used to identify the exome target sequence. The UMD 3.1 genome assembly was chosen as it has been shown to have fewer unassigned sequences compared to Btau_4.2 and it has improved annotation (Partipilo et al., 2011).

2.1.3.1 WES probe design

Liquid capture probes were designed to target all exons annotated in the *Bos taurus* UMD3.1 genome, plus 100 base pairs of 5'UTR, according to their standard protocols. An in-house Perl script was used to identify target regions. This resulted in approximately 48Mb targeted region. This capture design was obtained and used to capture targeted region of gDNA in two groups, high- and low-fertility AI bulls. To identify critical promoter regions, 100bp of 5' untranslated region (UTR) was also targeted for each gene. In total, there were 227,647 exons in UMD3.1 Ensembl version 70. Of these, 202,899 are targeted in this design, covering 56,671,697 bp. Five total matches and five mismatches were allowed. Of these probes, 92.5% were unique to a single genomic position, and 4.5% had only 2 possible matches within the genome. The length of probes was 200bp with approximately 2.1 million

[2] Design of probes for whole-exome sequencing was performed by Dr. Bruce Moran, Teagasc.

probes made to cover the exome. Probes were also designed to target mitochondrial DNA and were 1/5th of the concentration of nuclear DNA. Uniquely mapping probes covered 80% of the targeted regions. Probes that mapped up to 5 times in the genome covered approximately 98.4% of the targeted regions, and were used for capture design.

2.1.3.2 β-defensin probe design

Complete gene sequences (introns and exons), plus 1,000 bp 5'UTR of the predicted transcription start site for all β-defensin genes annotated in UMD3.1 Ensembl version 70 (a total of 387 kb) were targeted for Roche Nimblegen SeqCap EZ Developer probe design, according to standard protocols NimbleGen SeqCap EZ Library SR User's Guide v4.2. A full list of targeted β-defensin genes and chromosomal locations is shown in Table 1.8-1. The final targeted area from bait design was reduced to 235kb, to exclude repetitive regions.

Table 1.8-1: Gene names and locations of all defensin genes and cathelicidin genes targeted in the custom-designed capture probes for targeted re-sequencing.
Gene name = Gene ID, Chr = Chromosome, Start position = chromosomal location of exon start, End position = chromosomal location of exon end.

Gene name	Chr	Start position	End position
BBD132	13	61297228	61297434
DEFB129	13	61314417	61316744
BBD128	13	61327495	61329038
BBD127	13	61336426	61338919
DEFB126	13	61348964	61353549
BBD125A	13	61371518	61377487
BBD125	13	61391575	61402006
BBD115	13	61416164	61418920
BBD142	13	61436531	61447433
DEFB116	13	61462905	61468041
BBD117	13	61501532	61501753
DEFB118	13	61512890	61522849
DEFB119	13	61523658	61533444
BBD120	13	61531991	61533379
BBD121	13	61550485	61551769
DEFB122a	13	61562053	61566096
DEFB122a	13	61572837	61578011
DEFB123	13	61584480	61595780
DEFB124	13	61612683	61615456
DEFB133	23	22319719	22319840
DEFB114	23	22330039	22333711

Table 1.8-2 continued

BBD113	23	22351938	22353313
BBD110combined	23	22362601	22374885
BBD112	23	22381956	22387980
SBTBD1	27	4831108	4838674
TAP	27	4888377	4890195
DEFB103	27	4898705	4899642
SPAG11B	27	4920221	4942958
BBD104combined	27	4944581	4954375
BBD105	27	4956824	4958583
BBD107	27	4965545	4969446
DEFB103b	27	5046949	5057903
DEFB130	27	5064722	5064902
BBD109	27	5072979	5073182
LAP	27	5124202	5125990
LOC783012	27	5129114	5133058
BNBD6	27	5160482	5162187
BNBD6	27	5185069	5186715
BBDB403	27	5219420	5219479
BBDB403	27	5220343	5220402
BBDB403	27	5221191	5221250
DEFB7	27	5221842	5223732
BBD138	27	5247817	5247870
BBD138A	27	5274804	5274929
BBD138	27	5297139	5297201
BBD138	27	5300000	5300107
BNBD6	27	5327174	5328857
BBD108	27	5376114	5389581
DEFB4a	27	5425378	5427298
DEFB	27	5457175	5465032
SBTBD1	27	5473206	5473352
DEFB1	27	5483406	5539158
DEFB5	27	5560976	5564586
BT402	27	5599460	5600761
BT300	27	5614286	5614330
BT300	27	5638525	5638581
BNBD11	27	5788753	5788812
BBD140	27	5808756	5809848
DEFB10	27	6194470	6196146
BBDB403	27	6223461	6223514
BNBD14	27	6225015	6225125
BBDB131	8	7280952	7287888
BBD135	8	7288666	7288833
BBD134	8	7301132	7303564
BBD136	8	7331611	7332290

2.2 Materials and methods related to Chapter 3.

2.2.1 Targeted β-defensin sequencing of AI sires - library preparation and sequencing[3]

Genomic DNA from high- and low-fertility sires were selected, quantified, libraries prepared, and sequenced. gDNA was extracted from hair and semen for AI bulls genotyped. DNA was quantified using Qubit® dsDNA BR Assay Kit for use with the Qubit® 2.0 Fluorometer (Life Technologies). DNA was then cleaned using DNA Clean & Concentrator™ kit (Zymo Research), 200ng cleaned gDNA was sheared and size selected using Bioruptor Plus (Diagenode), with fragment lengths confirmed to be approximately 600bp by Bioanalyzer (Agilent). DNA fragments were then prepared for sequencing as specified in the TruSeq Nano DNA LT Sample Prep protocol (Illumina). In total, 168 libraries were prepared for sequencing in pooled batches of 24 samples.

A dual-capture protocol was used. Following pre-capture amplification 24 libraries were pooled to form 1µg of DNA to hybridise with baits overnight, captured, amplified, hybridised with another aliquot of the baits overnight, amplified and cleaned. Libraries were quantified using Qubit Hi Res (Life Technologies) and run on an Illumina MiSeq at 10pM with 1% PhiX (300bp paired-end protocol).

2.2.2 Data analysis of targeted β-defensin sequencing dataset

Paired end reads were analysed via fastQC, quality filtered with phred quality score of 25 (q 25), for paird-end reads with BWA and adaptor trimmed with TrimGalore! (version 0.4.2) using a custom PERL script, see electronic appendix[4]. Btau.UMD3.1 version 70 was downloaded from Ensembl and BWA (version 0.7.15) was used to align the reads to it. Picard tools (version 1.60) SamFormatConverter and SortSam were used to convert BAM to SAM format and sort the SAM files. PCR duplicates were marked using the Picard tools MarkDuplicates walker, assuming sorted files. Enrichment, insert size and alignment metrics were calculated using the CalculateHsMetrics (picard-tools-1.60), CollectInsertSizeMetrics

[3] Targeted β-defensin sequencing and data analysis for 144 sires was performed with Dr. Emma Finlay, Teagasc, in collaboration with the same research project and funding source. Whole-exome sequencing data was analysed, and targeted β-defensin library preparation, sequencing, and data analysis of remaining 24 sires was performed, by Mr. Ronan Whiston.

[4] Electronic Appendix 2.1 Trim_fastqc_map_gatk_filter.pl

(picard-tools-1.60), and CollectAlignmentSummaryMetrics (picard-tools-1.60) walkers, respectively, with VALIDATION_STRINGENCY=LENIENT for each. Variant discovery was also performed using GATK, following the Best Practice Pipeline (Broad, 2017); variants were called individually using the HaplotypeCaller (Genome Analysis Toolkit: Version 3.4-0-g7e26428) and joint genotyping performed on all samples simultaneously using GenotypeGVCFs.

2.2.3 SNP filtration of β-defensin sequencing dataset

SNPs identified via the targeted β-defensin sequencing GATK best practice pipeline were filtered to remove variants of low quality and which fall outside certain parameters, to reduce the number of false positives. The parameters for hard filtering of variants were as follows: Filter out variant calls if located within a cluster where three or more calls are made in a 10 bp window [clusterWindowSize 10]; filter out variant if there are at least four alignments with a mapping quality of zero (MQ0) and if the proportion of alignments mapping ambiguously corresponds to 1/10th of all alignments [MQ0 > =4 && ((MQ0/(1.0 * DP)) > 0.1)], DP: total (unfiltered) depth over all samples; filter out variants which are covered by less than 5 reads [DP < 5]; filter out variants having a low quality score [Q < 50]; filter out variants with low variant confidence over unfiltered depth of non-reference samples (QD) [QD < 1.5]; filter out variants based on strand bias using Fisher's exact test: FS > 60.0 for SNP calling, FS > 200.0 for InDel calling. In-house Perl scripts were also written to identify any individual genotypes identified as heterozygous with an allele ratio of> 80:20 and any SNP which had a read depth of less than 8 in a given individual and code them as missing data.

2.2.4 SNP association analysis of β-defensin sequencing

SNP association analysis of targeted β-defensin SNPs with AAM fertility phenotype was performed using the R package GenABEL (version 1.8-0) (Aulchenko et al., 2007). This package and it's usage is described in section 2.3.8. Quality control was performed with the check.marker function. A SNP call rate filter of 0.8, individual call rate (maximum percent of missing genotypes in an individual) cut off = 0.9, and minor allele frequency cut off = 0.05,

were applied. SNPs were examined for association to fertility using breed and the number of matings performed as fixed effects.

2.2.5 Targeted β-defensin re-sequencing in sire subset

Using the same 24 bulls sequenced in the Whole Exome Sequencing (WES) project, β-defensin and cathelicidin genes were also sequenced. Targeted re-sequencing targets included ~378kb, comprising 69 target genes, including introns, exons and regulatory regions. These 69 targets included 1,000bp of 5' UTR promoter region, to target gene regulation in the promoter region.

2.2.6 Targeted β-defensin gene re-sequencing library preparation in sire subset

To prepare libraries for targeted capture and sequencing, firstly, a sample library was created. The sample library is the initial shotgun library generated from gDNA by fragmentation and ligation of sequencing–specific adapters. Sequencing libraries were prepared according to the manufacturer's protocol (TruSeq® Nano DNA Library Prep Reference Guide), using the Illumina TruSeq Nano LT Sample Prep Kit (Illumina, San Diego, California). Briefly, 200ng DNA was sheared using a Bioruptor Plus and the fragment length of ~600bp was confirmed using the Bioanalyzer (Agilent). Library preparation, end-repair, A-tail and adapter ligation were performed according to the TruSeq Nano DNA LT sample prep protocol (Illumina). Samples were pooled at random into groups of 24. The same 24 bulls sequenced in the Whole Exome Sequencing (WES) project were pooled together in one reaction. One equimolar pool of all 24 sample libraries was then used for library sequence capture, which is the enrichment of targeted regions from the sample libraries.

Sequence capture of pooled sample libraries with the SeqCap EZ library of a customised, complete set of biotinylated oligonucleotide probes provided by Roche Nimblegen was performed. These probes were designed to capture 69 whole genes of interest, including introns, and 1,000 bp of 5' UTR.

Following sequence capture, unbound probes and DNA were washed twice, using a magnetic separation column. Bound probes and DNA were amplified with 4 cycles of

ligation-mediated PCR, LM-PCR. Immediately prior to the sequencing run, the pooled sample library is denatured and diluted to 8pM, according to the Illumina MiSeq protocol. All 24 libraries were pooled in equimolar concentrations and underwent sequencing on an Illumina MiSeq instrument using the 300-cycle MiSeq Reagent Kit v3 (Illumina) to give paired-end reads of 300bp. Sequences were downloaded from BaseSpace, the Illumina genomics cloud computing tool.

2.2.7 Targeted re-sequencing data analysis in sire subset

Sequencing reads (300bp) were aligned to UMD3.1 ENSEMBL 70 release using the Burrows Wheeler Transform Aligner (BWA) (Li and Durbin, 2009) with the following parameters "bwa aln -q 20 -t 8". This set the phred-scaled quality cut-off at 20 and threading was performed across 8 cores. Duplicates were removed and variant calls were made using GATK's HaplotypeCaller walker (GATK Version 3.4-0-g7e26428), with the following commands:

"java -Xmx3g -jar /data/efinlay/GenomeAnalysisTK.jar -T HaplotypeCaller -R /home/bmoran/bin/ens70/Btau.UMD3.1.70.short.fa -I $output1/$sample.recal.bam -o $output1/$sample.raw.2.vcf --dbsnp /data/rwhiston/Project_RonanTS/exome/Bos_taurus_78.fixed_2_sorted.vcf -stand_call_conf 30 -stand_emit_conf 10 -minPruning 3"

"java -Xmx3g -jar /data/efinlay/GenomeAnalysisTK.jar -T HaplotypeCaller -R /home/bmoran/bin/ens70/Btau.UMD3.1.70.short.fa -I $output1/$sample.recal.bam --emitRefConfidence GVCF --variant_index_parameter 128000 -variant_index_type LINEAR -o $output1/$sample.raw.g.2.vcf --dbsnp /data/rwhiston/Project_RonanTS/exome/Bos_taurus_78.fixed_2_sorted.vcf -stand_call_conf 30 -stand_emit_conf 10 -minPruning 3".

GATK's best practice pipeline was followed to identify variants. SNPs were filtered to remove SNPs with coverage < 5, quality < 30 or displaying strand or read position bias:

"java -Xmx3g -jar /data/efinlay/GenomeAnalysisTK.jar -T VariantFiltration -R /home/bmoran/bin/ens70/Btau.UMD3.1.70.short.fa --variant combined_files_g.vcf -o combined_140_files_filtered.vcf --clusterWindowSize 10 --filterExpression "DP<5" --filterName "LowCoverage" --filterExpression "QUAL<30.0" --filterName "VeryLowQual" --

filterExpression "QUAL > 30.0 && QUAL < 50.0" --filterName "LowQual" --filterExpression "MQ0 >=4 && ((MQ0 /(1.0*DP)) > 0.1)" --filterName "HARD_TO_VALIDATE" --filterExpression "FS >60.0" --filterName "STRAND_BIAS" --filterExpression "vc.hasAttribute('ReadPosRankSum')&&ReadPosRankSum <-8.0" --filterName "READ_POS" --filterExpression "vc.hasAttribute('QC')&&QD<1.5" --filterName "LowQD"”

Individual sample genotypes were filtered to mark individuals with read depth less than 8 as missing data, using an in-house Perl script available in Thesis\Appendix chapter 4 - Whole-exome\R Scripts \filter_on_snp_DP.pl [5].

[5] Electronic Appendix 2.2 snpfiltration.sh

2.2.8 SNP frequency analysis in sire subset

SNP frequency analysis between high and low-fertility groups was performed on the SNPs identified in targeted β-defensin sequencing. In-house Perl scripts were also written to identify any individual genotypes identified as heterozygous with an allele ratio of greater than 80:20 and any SNP which had a read depth of less than 8 in a given individual and code them as missing data within that individual, as shown in Thesis\Appendix chapter 4 - Whole-exome\R Scripts \filter_on_snp_DP.pl .

2.2.9 O-linked glycosylation analysis

The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins (Julenius et al., 2005). Protein sequences for β-defensin genes were downloaded from NCBI [6] in FASTA format in March 2013, 2 years after (Narciandi et al., 2011) performed glycosylation analysis on BBD genes. Protein sequences were then sent to NetOglyc 3.1 server, available online [7] using default parameters.

[6] http://www.ncbi.nlm.nih.gov/
[7] http://www.cbs.dtu.dk/services/NetOGlyc-3.1/

2.3 Materials and Methods related to Chapter 4

Following filtering of sires for which fertility phenotype data from 7,000 bulls was available, 94 bulls were identified as being divergent for both phenotypes, 91 bulls were divergent in AAM only and 79 bulls were divergent for PR only. From these bulls, 24 were selected for whole-exome sequencing of bulls divergent for fertility. Priorities for selection were given to bulls which were divergent for both phenotypes, followed by bulls which were divergent for AAM, and then for PR. DNA availability in the Teagasc DNA databank was also considered. Phenotype values for each breed are shown in Figure 2.3-1 on page 52, for adjusted animal model and Figure 2.3-2 for pregnancy rate. AAM is shown to have less variability in the phenotype in BB animals in comparison to HF and LM, which are similar. The adjusted animal phenotype demonstrated less variability over different breeds and over time, and was determined to be the more robust phenotype. In total, 18 bulls from the Teagasc DNA databank were selected, and semen straws were obtained from ICBF for a further 6 bulls for DNA extraction.

Six high- and six low-fertility HF bulls, three high- and three low-fertility BB and three high- and three low-fertility LM bulls were selected, see Table 2.3-1.

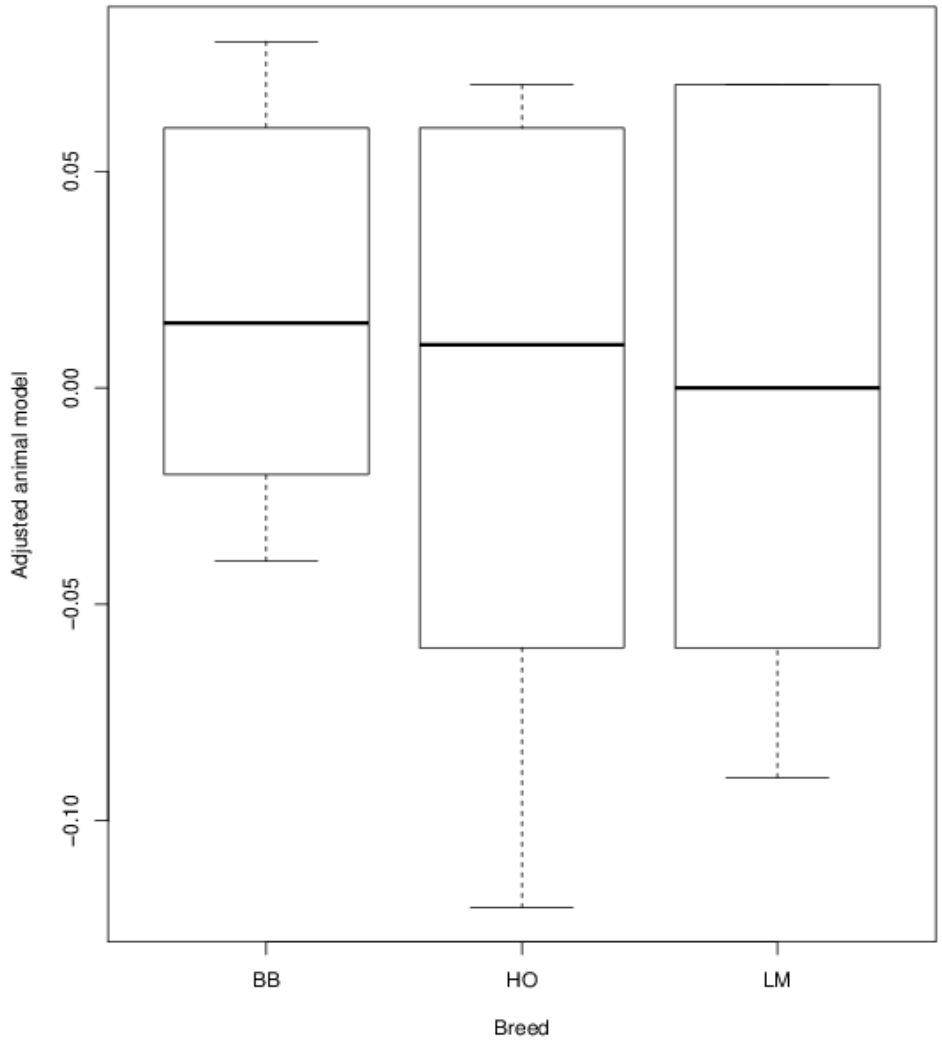


Figure 2.3-1: Adjusted animal model phenotype values by breed of sire
Boxplot of AAM fertility phenotype values (n=24) for each breed (BB = Belgian Blue; n=6), (HO = Holstein-Friesian; n=12) and (LM = Limousin; n=6). Black lines denote median values, box denotes the upper and lower quartiles, and dashed lines (whiskers) denote variability outside the upper and lower quartiles. Of the 24 bulls selected, BB animals show less variability in the AAM phenotype than HO and LM.

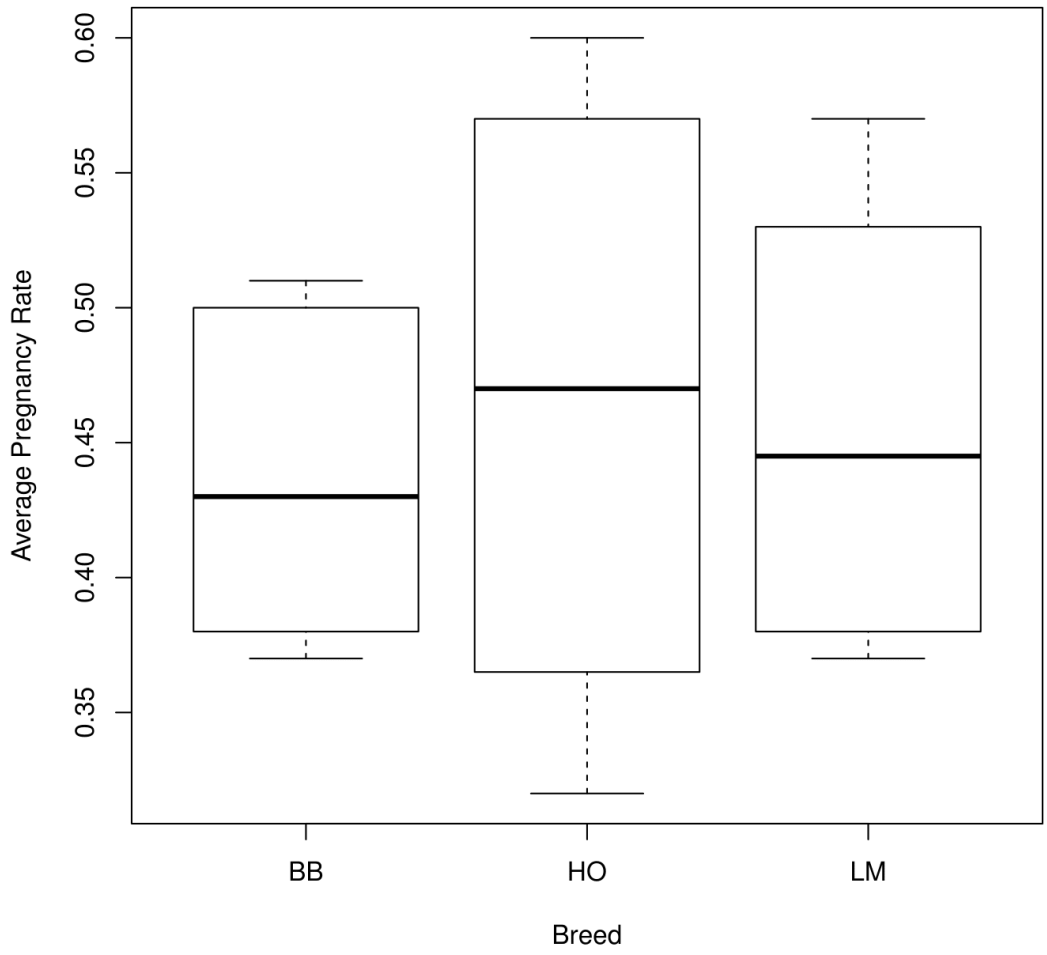


Figure 2.3-2: Average pregnancy rate phenotype values by breed of sire

Boxplot of average PR fertility phenotype values (n=24) for each breed (BB = Belgian Blue; n=6), (HO = Holstein-Friesian; n=12) and (LM = Limousin; n=6). Black lines denote median values, box denotes the upper and lower quartiles, and dashed lines (whiskers) denote variability outside the upper and lower quartiles. Of the 24 bulls selected, BB animals show less variability in the PR phenotype than HO and LM.

Table 2.3-1: AI sire sample selection criteria

Overall Adjusted Animal Model (AAM) effect, and Average PR phenotypes, breed and total straw count used per sire are shown, with AI bulls separated into breed and fertility status based on phenotypic data. Minimum (Min), Mean and Maximum (Max) values for PR and AAM for all 24 selected samples are also shown. Images depict the three different breeds selected: Holstein-Friesian (HO), Limousin (LM) and Belgian Blue (BB).

Sire ID	Mean PR	Overall AAM Effect	Breed	Total Straw Count
Holstein-Friesian High				
Bull 1	0.56	0.07	HO	12916
Bull 2	0.57	0.06	HO	6216
Bull 3	0.59	0.05	HO	2758
Bull 4	0.54	0.06	HO	5308
Bull 5	0.57	0.06	HO	8529
Bull 6	0.6	0.06	HO	7420
Holstein-Friesian Low				
Bull 7	0.4	-0.06	HO	2424
Bull 8	0.34	-0.09	HO	1844
Bull 9	0.35	-0.04	HO	1061
Bull 10	0.32	-0.12	HO	5449
Bull 11	0.38	-0.06	HO	5928
Bull 12	0.39	-0.03	HO	1948
Limousin High				
Bull 13	0.57	0.06	LM	6196
Bull 14	0.53	0.07	LM	4506
Bull 15	0.51	0.07	LM	3246
Limousin Low				
Bull 16	0.38	-0.09	LM	1138
Bull 17	0.37	-0.06	LM	1354
Bull 18	0.38	-0.06	LM	1227
Belgian Blue High				
Bull 19	0.51	0.08	BB	2705
Bull 20	0.5	0.06	BB	1515
Bull 21	0.47	0.05	BB	9062
Belgian Blue Low				
Bull 22	0.37	-0.04	BB	9959

Table 2.3-1 continued

	Bull 23	0.39	-0.02	BB	10529
	Bull 24	0.38	-0.02	BB	31680
	Min	0.32	-0.12		
	Mean	0.45	0.0025		
	Max	0.6	0.08		

2.3.1 Genomic DNA extraction, purification and quality control

Genomic DNA was previously extracted using the Maxwell® 16 research instrument system, according to manufacturer's instructions, and were stored in the Teagasc DNA bank. The Teagasc DNA bank is a biobank of gDNA extracted from semen straws for thousands of sires. DNA for this project was obtained from this DNA bank (n = 18). In addition, six samples required DNA extraction from semen straws obtained from the National Cattle Breeding Centre (n=6)[8].

For Whole-Exome Sequencing, gDNA was purified using Zymo Research's DNA Clean and concentrator ™ kit. At least 1ug of gDNA was purified and stored at -20^{O}C. Genomic DNA was heated to 52^{O}C for 2 mins prior to concentration estimation. Genomic DNA concentrations were estimated using the Qubit® dsDNA BR Assay Kit for use with the Qubit® 2.0 Fluorometer and separately, concentrations were estimated using Nanodrop ND-1000 spectrophotometer. Qubit® concentration values were preferentially accepted as a more accurate estimation of double-stranded gDNA, due to the nature of the two methods. Nanodrop estimates all nucleic acid material in a sample, which could include RNA. However, with Qubit, the dye fluoresces upon contact with double stranded DNA, which is a more accurate method for dsDNA quantification. Nanodrop readings were used to assess protein contamination and solvent contamination via the 260/280 nm, and 260/230 nm wavelength ratios, see Table 2.3-2 for both Qubit and Nanodrop concentration measurements for each bull. Due to large variations in concentration readings, lower readings were considered, and a low DNA input TruSeq Nano kit was used. Exome capture was performed by pooling 6 samples and performing 1X capture, whereas 1X capture per 4 samples is common for human whole-exome captures.

[8] Six genomic DNA extractions from semen straws were performed by Ms. Margaret Murray, Teagasc.

Table 2.3-2: Quality control of gDNA from Teagasc DNA databank of sires selected for whole-exome sequencing and targeted sequencing. Table of 24 AI bulls selected for whole-exome sequencing, indicating their gDNA quality control prior to library preparation and sequencing by Clinical Genomics, Canada. Qubit quantification (ng/µl), Nanodrop quantification (ng/µl) and quality (260/280, 260/230 ratios were analysed, data not shown), volume and total amount (ng) are displayed for each of the 24 sires selected for whole-exome sequencing.

Sample ID	Qubit (ng/µl)	Nanodrop (ng/µl)	Volume (µl)	Nanodrop Total (ng)	Qubit Total (ng)
Holstein-Friesian High					
Bull 1	32.2	40.28	60	2416.8	1932
Bull 2	56	40.91	30	1227.3	1680
Bull 3	12.2	12.26	60	735.6	732
Bull 4	27.26	27.28	30	818	818
Bull 5	13.1	13.73	60	823.8	786
Bull 6	12.4	13.8	60	828	744
Holstein-Friesian Low					
Bull 7	85.5	85.5	30	2565	2565
Bull 8	17.6	17.78	60	1066.8	1056
Bull 9	11.3	15.88	60	952.8	678
Bull 10	99	154.3	50	7715	4950
Bull 11	42.3	43.43	50	2171.5	2115
Bull 12	7.47	12.63	60	757.8	448.2
Limousin High					
Bull 13	103	112.6	50	5630	5150
Bull 14	6.62	7.95	60	477	397.2
Bull 15	19.2	22.57	60	1354.2	1152
Limousin Low					
Bull 16	19	20.83	50	1041.5	950
Bull 17	80	76.32	30	2289.6	2400
Bull 18	46	52.99	60	3179.4	2760
Belgian Blue High					
Bull 19	12.9	301	50	15050	645
Bull 20	39	41.12	60	2467.2	2340

Table 2.3-2 continued

Bull 21	33.24	136.2	50	6810	1662
Belgian Blue Low					
Bull 22	16.9	21.05	60	1263	1014
Bull 23	20	32.93	68	2239.24	1360
Bull 24	38.6	31.75	30	952.5	1158

2.3.2 Sample preparation

Whole-exome sequencing library preparation was performed commercially (Clinical Genomics, Ontario, Canada). Library preparation was performed according to the Roche Nimblegen SeqCap EZ Developer Library protocol, as per manufacturer's instructions.

Briefly, 100ng gDNA of each sample were prepared using the TruSeq Nano DNA Library Prep Kit. This step involved shearing of gDNA using Covaris DNA shearing and Agencourt AMPure XP beads, end-repair of fragments, adapter ligation, LM-PCR and quality control. The strategy to hybridise fragments of gDNA to the exome baits was to make one equimolar pool of all 24 gDNA libraries. This pool was then split into four separate pools and hybridisation captures were performed on each pool with oligonucleotide probe capture reagents, resulting in 1X capture for every 6 samples. All captures were then re-pooled and split into four groups for sequencing on four HiSeq 2500 lanes. Pooling ensures that samples are mixed prior to capture, and to give an even distribution of sample to each sequencing lane. Sample preparation for Roche Nimblegen EZ developer system is summarised in Figure 2.3-3.

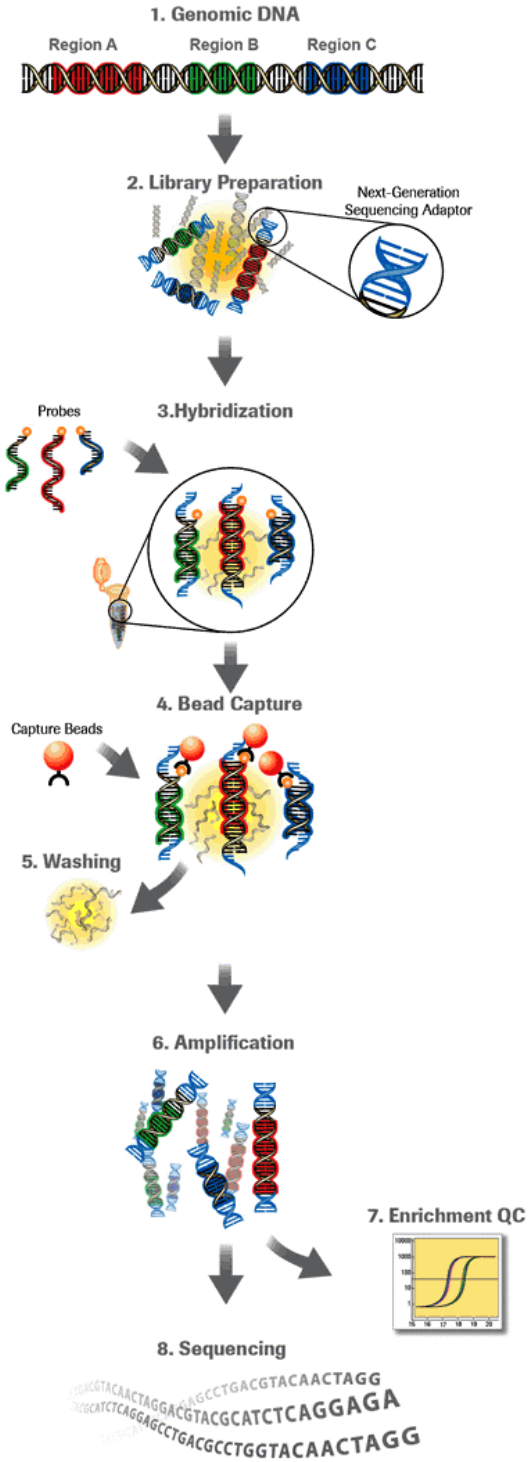


Figure 2.3-3: Roche Nimblegen SeqCap EZ Developer workflow system
Eight steps involved in sequencing library preparation: 1) gDNA shearing 2) Adaptor ligation 3) Hybridization of custom-designed probes 4) Capture of DNA by magnetically labelled oligonucleotide probes 5) Washing away unbound probes and DNA 6) PCR Amplification of bound probes with DNA 7) Quality Control and 8) Sequencing amplified libraries.

2.3.3 Exome data analysis

FASTQ files were uploaded to a 16 CPU server running GNU bash, version 3.2.51(1) (x86_64-suse-linux-gnu) from an external hard drive containing exome sequencing data obtained from Clinical Genomics, Canada. FASTQ files were concatenated into two individual files, *R1* and *R2*, to designate paired-end reads, read 1 and read 2. Quality control of raw sequencing data was performed with FastQC, which takes each file for a given sample and produces a quality control report consisting of a number of different modules. FASTQ files were trimmed using Trim Galore! This tool uses the first 13 bp of Illumina standard adapters ('AGATCGGAAGAGC') by default, to trim adapter sequences. A Phred quality score threshold of 25 was applied, to discard poor quality base calls, and to reduce the rate of incorrect base calling, using the following command:

"trim_galore --paired --fastqc -q 25 $read1 $read2".

2.3.4 Alignment

All remaining reads were aligned to the *Bos taurus* UMD3.1.70 genome using the Burrows-Wheeler Aligner (BWA) 'sampe' algorithm with default parameters (Li and Durbin, 2009). Picard Tools (version 1.60) was used to convert the resulting SAM file to BAM format, sort and index BAM files, and to remove PCR duplicates from all BAM files (Broad, 2017). PCR duplicates arise during the library preparation step, where the DNA is amplified to increase probability of probes binding to each DNA fragment. However, this also introduces PCR amplification bias, which needs to be mitigated. PCR duplicates are identified by two or more reads having the same chromosome start position and the same CIGAR string, signifying identical reads, which are not informative, and need to be removed from analysis. Alignment summary metrics, insert size metrics, and PCR duplicate metrics were all collected. GATK's DepthOfCoverage walker determined coverage levels per interval. Intervals were defined as the exome intervals used to design bait probes.

2.3.5 Variant calling

Variant calling and genotyping across all 24 animals was performed using Genome Analysis Toolkit (GATK Version 3.4-0-g7e26428), following the GATK best practice guidelines for whole-exome sequencing (DePristo et al., 2011), see Figure 2.3-4, on page 63. Local re-alignment around InDels (Insertions and Deletions) was performed using the GATK tools: RealignerTargetCreator (using -known /home/bmoran/bin/ens70/Bos_taurus_74.headed.vcf), IndelRealigner (using -targetIntervals $output1/$sample.forIndelRealigner.interval_list and -known /home/bmoran/bin/ens70/Bos_taurus_74.headed.vcf), and FixMateInformation (using SORT_ORDER=coordinate VALIDATION_STRINGENCY=LENIENT), as described in Electronic Appendix 2.1 Trim_fastqc_map_gatk_filter.pl. Base quality score re-calibration via GATK BaseRecalibrator was then applied, which recalibrates scores around known variants. HaplotypeCaller walker was used to call mutations on BAM files with Phred-scaled emit and call confidences of 30, in 'GVCF' mode and with a BED file of the exome targets. This BED target file is used by the walker to identify regions in the genome which are active/variable, which are marked for local de-novo assembly (via de-Bruijn graphs) of reads aligning to such regions.

A modified read alignment and variant calling Perl script is shown in the electronic appendix[9]. This script takes in a list of FastQ files, by submitting the following command in a directory containing all FastQ files to be analysed: ls *.fastq.gz > fastqlist.txt.

This Perl script then performs quality control, aligns reads to the genome using BWA, converts files from SAM to BAM format, sorts BAM files based on coordinates, marks PCR duplicates, fixes paired-end reads mate information, builds a BAM file index, and generates quality control metrics for alignment and coverage. Base quality score recalibration is then applied to detect and reduce systematic errors in base quality scores, followed by variant calling using HaplotypeCaller walker.

[9] Electronic Appendix 4.2 R Scripts/ 1 - total_list_Brucerefs_newGATK.pl

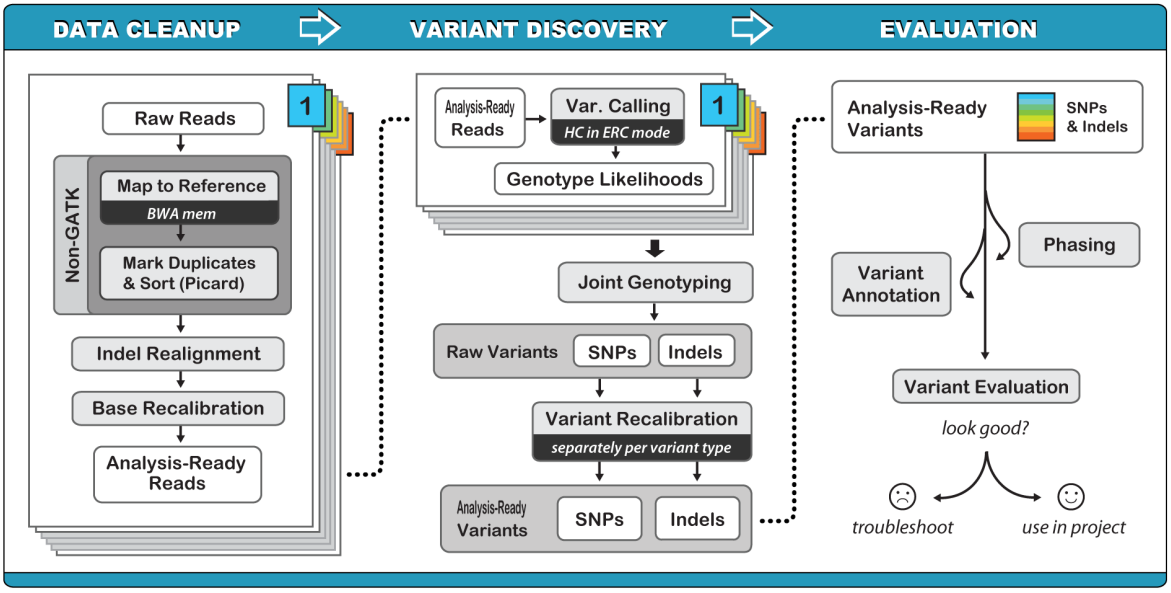


Figure 2.3-4: Genome analysis toolkit best practice pipeline.
Indicates the key steps required to obtain high-quality variants from raw sequencing reads. Mapping reads to reference genome, duplicate removal, sorting, InDel realignment, base recalibrations, HaplotypeCaller variant calling, variant recalibration, Annotation, and custom filtering (GATK, 2015).

2.3.6 Variant filtering

Strict hard-filtering of variants was performed to remove variants of low quality and which fall outside certain parameters, to reduce the number of false positives. The parameters for hard-filtering of variants were as follows: Filter out variant calls if located within a cluster where three or more calls are made in a 10 bp window [clusterWindowSize 10]; filter out variant if there are at least four alignments with a mapping quality of zero (MQ0) and if the proportion of alignments mapping ambiguously corresponds to 1/10th of all alignments [MQ0 > =4 && ((MQ0/(1.0 * DP)) > 0.1)], DP: total (unfiltered) depth over all samples; filter out variants which are covered by less than 5 reads [DP < 5]; filter out variants having a low quality score [Q < 50]; filter out variants with low variant confidence over unfiltered depth of non-reference samples (QD) [QD < 1.5]; filter out variants based on strand bias using Fisher's exact test: FS > 60.0 for SNP calling, FS > 200.0 for InDel calling, similar to other SNP filtering protocols (Robert et al., 2014). In-house Perl scripts were also written to identify any individual genotypes identified as heterozygous with an allele ratio of> 80:20 and any SNP which had a read depth of less than 8 in a given individual and code them as missing data within that individual.

Following HaplotypeCaller and VariantFiltration walkers, variants were filtered using a custom Perl script[10]. This script filtered variants based on low coverage with variants at an overall depth of coverage less than 5, and not located within the targeted region. Variants were further filtered for read depth, with each individual genotype requiring at least 5X depth, in addition to the overall depth of coverage filter for each variant. Variants with individual genotype coverage less than 5X were set as missing data.

A custom Perl script was applied to identify rare variants[11]. This script takes a filtered VCF file and outputs the zero, one and missing alleles for each variant. Allele frequencies can be calculated based on different categories of interest. Fertility and breed were categories analysed for this project. 25 percentage points between groups of high- and low-fertility was used to identify variants to be added to a custom-designed SNP chip. A SNP frequency differential between high-and low-fertility groups was chosen as a parameter to identify candidate SNPs involved in male fertility, as SNPs with large SNP frequency differences may be under genetic selection pressures (Mullen et al., 2012).

2.3.7 Quality control

The GenABEL (version 1.8-0) package tool check.marker was used to perform quality control on variant calls. This package filters variants to help select the variant which should enter GWA analysis based on call rate, MAF, value of chi-square test for Hardy-Weinberg equilibrium (HWE) and redundancy (concordance between distributions of the genotype). Variants were filtered based on a SNP call rate of 80% (call rate = 0.8), an individual SNP call rate of 90% (maximum percentage of missing genotypes in an individual sample; perid.call =0.9) and a minor allele frequency cut-off threshold of 5% (MAF = 0.05).

Further stringent quality control filtering for association analysis removed variants which had low call rate, low minor allele frequency, or were out of Hardy-Weinberg equilibrium. Samples were also removed with high autosomal heterozygosity (FDR <1%), or high identity by state (IBS) > 95%. Two samples, a HF low-fertility sire (Bull 8), and a LM high-fertility sire (bull14) were then removed from analysis due to high autosomal heterozygosity. Excluded

[10] Thesis\Appendix chapter 4 - Whole-exome\R Scripts \filter_on_snp_DP.pl
[11] Thesis\Appendix chapter 4 - Whole-exome\R Scripts\ count_alleles_by_category_new.pl'

sires had autosomal heterozygosity rates of 56% and 48%, probably due to sequencing errors for these two samples or high heterozygosity in a highly-selected breeding population. However, as this could not be determined here, the two bulls were excluded from association analysis. Bull 14 (LM) was the bull with the lowest input gDNA (~400ng) for library preparation, as shown in Table 2.3-2. Bull 8 had ~1µg of total gDNA for input. This indicates that either bull 8 and 14 had poor library preparation or capture efficiency. This reduced the number of sires available for SNP association analysis to 22, see Figure 4.3-4 and Figure 4.3-5 for pre- and post-QC principal component analysis (PCA).

2.3.8 Association analysis

The R package GenABEL (version 1.8-0) (Aulchenko et al., 2007) was used to perform association analysis between the fertility phenotype and SNP genotypes. A custom Perl file converted a Variant Call Format (VCF) file containing all variants from GATK HaplotypeCaller to GenABEL format[12]. The phenotypes were then defined and the association analysis performed, with the most significant chromosomes and SNPs identified. SNPs of significance (unadjusted *P*-value < 0.01) were sorted based on their *P*-value with 1 degree of freedom. Adjusted P-values were not used in this analysis to identify the largest number of SNPs in this SNP discovery project as possible, as further validation in an independent population will be carried out, and separate genotyping in a large population via SNP chip has been performed. Due to the low numbers involved in WES association analysis, this was determined to be the optimal strategy.

A linear mixed model (LMM) approach analysed each SNP separately for association with a phenotype, which allows for fixed and random effects. SNPs identified were analysed as continuous variables, in which case an allelic effect will be estimated. The null hypothesis is that there is no association between the SNP and the fertility trait. This model assumes a linear relationship between the trait and genotype as well as a common variance at each genotype. Association analysis commands are shown in the electronic appendix[13].

[12] Thesis\Appendix chapter 4 - Whole-exome\R Scripts\vcf_to_GenABEL_format.pl
[13] Electronic Appendix 4.6 5 - Association commands.R

2.3.9 Gene ontology

To identify a set of enriched terms from gene ontology (GO) and other relevant biological databases, GO over-representation analysis on whole-exome sequencing SNP Ensembl gene IDs was performed. The 484 SNPs most associated with the adjusted animal model fertility phenotype (unadjusted *P*-value < 0.01) were identified and SnpEff annotated each SNP to the gene it is in or the nearest gene. Ensembl gene IDs of the annotation to UMD3.1 for each SNP was obtained. GO term analysis was performed using the DAVID Functional Annotation Tool, DAVID Bioinformatics Resources 6.7, NIAID/NIH (Huang da et al., 2009b, Huang da et al., 2009a).

2.3.10 Transcription factor binding site analysis

The top 20 SNPs most associated with AAM fertility phenotype and which were predicted to be upstream or in the 5'UTR of the nearest gene, via SnpEff, were analysed for transcription factor binding site analysis. MatInspector (Quandt et al., 1995) v3.7 was used to identify potential binding sites for transcription factors which are affected by polymorphisms identified in this whole-exome sequencing dataset. MatInspector is a software tool that searches a library of matrix descriptions for transcription factor binding sites and locates matches in DNA sequence. It assigns a quality rating to matches and allows quality-based filtering and selection of matches. gDNA sequences for the promoter region of interest were submitted to the MatInspector database, using default parameters, to identify potential transcription factor binding sites which may be affected by the identified genetic variants *in silico*.

2.4 Materials and methods related to Chapter 5

2.4.1 Sire selection for validation

An independent population of bulls were used to validate variant calls from whole-exome sequencing. Fertility phenotypes of bulls from ICBF fertility records were used, as previously described. However, updated fertility data for this dataset from year 2013, 2014 and 2015 were available which were not used for initial sample selection for whole-exome sequencing or targeted β-defensin sequencing. Fertility phenotype data for 2013 – 2015 are shown in Figure 2.4-1, for identification of sires suitable for validation of variant calls from whole-exome sequencing and targeted β-defensin sequencing. Sires were selected which had > 100 inseminations, to ensure reliable data, and also to allow selection of larger numbers to ensure the robustness of the variant calling. Sires were also required to be of three breeds Limousin (LM), Belgian Blue (BB), or Hostein-Friesian (HF). Average PR and AAM fertility data for all 1,400 sires with phenotypic data available for 2013 – 2015 are shown in Figure 2.4-2 and Figure 2.4-3, showing the AAM as being the more stable fertility phenotype over time.

Phenotypic data 2015

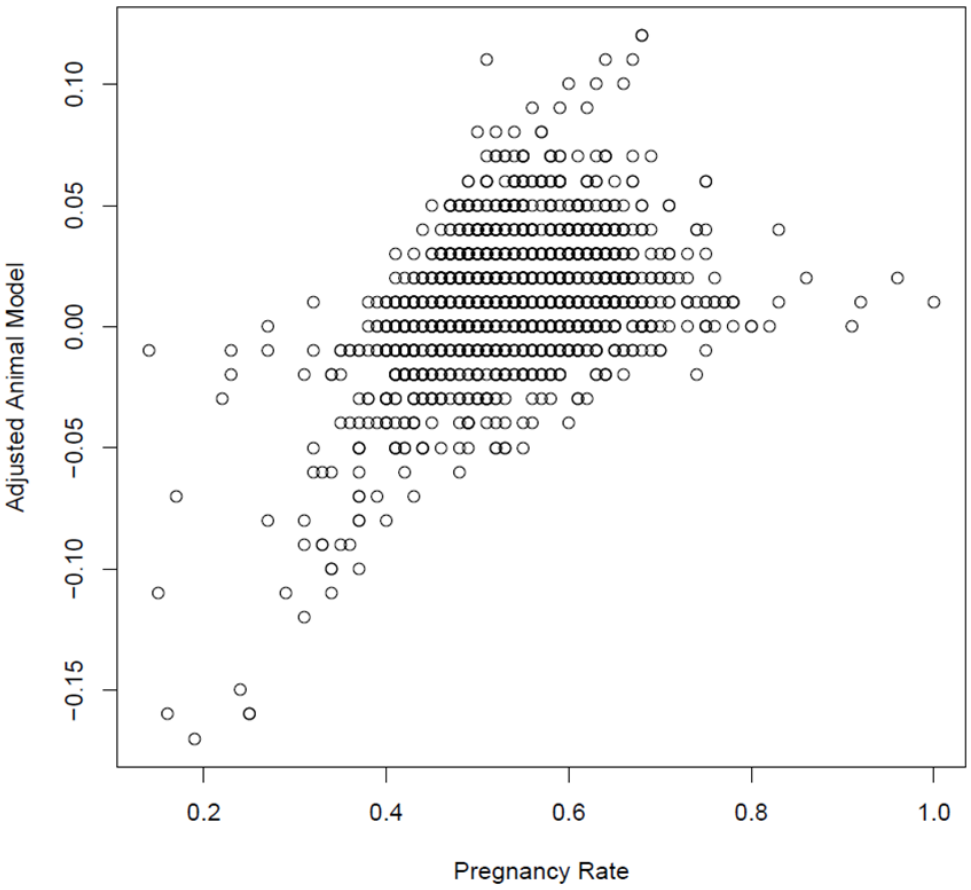


Figure 2.4-1: All phenotypic data for adjusted animal model and pregnancy rate between the years 2013 - 2015 for identification of sires for SNP validation

X-axis denotes the PR fertility phenotype values for all (1,400) sires with phenotypic data available between 2013 and 2015, inclusive. Y-axis denotes the AAM fertility phenotype values for all sires with phenotypic data available between 2013 and 2015, inclusive.

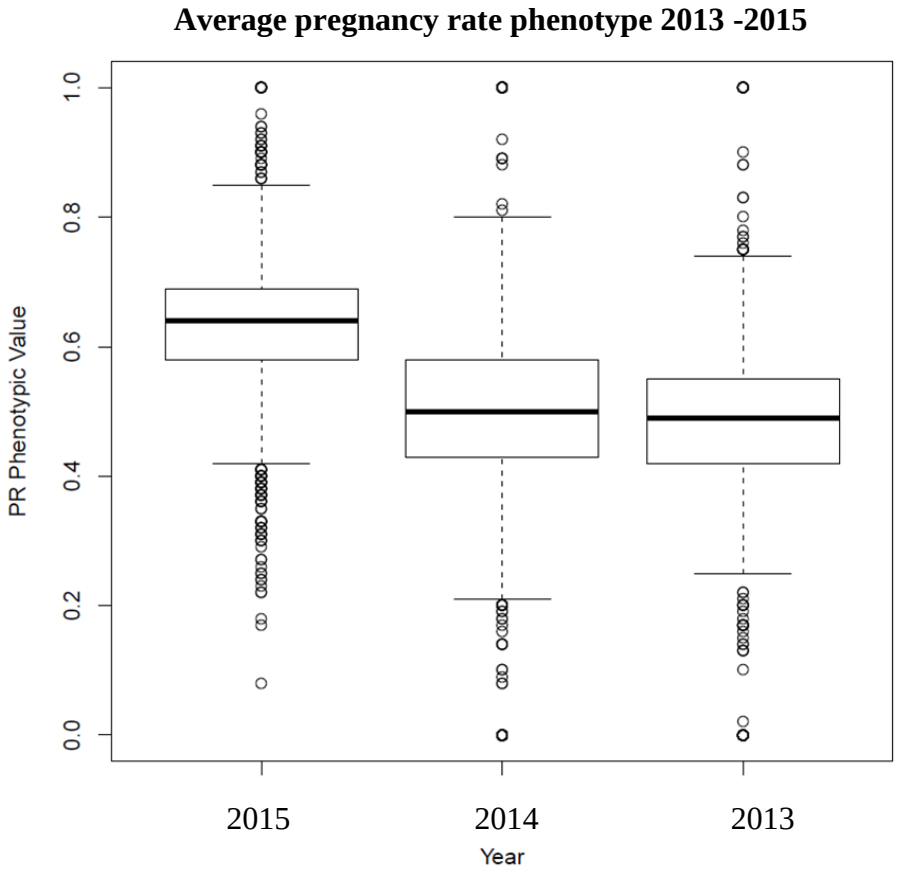


Figure 2.4-2: Pregnancy rate fertility values for 2013 - 2015 to identify sires with stable fertility phenotypes for SNP validation

Boxplots of average PR phenotypes for each individual year 2013, 2014 and 2015. Data is used for selection of sires to be used in validation of variant calling via whole-exome sequencing and targeted β-defensin sequencing. Average PR values are denoted on the y-axis. Year of average PR phenotype data is denoted on the x-axis.

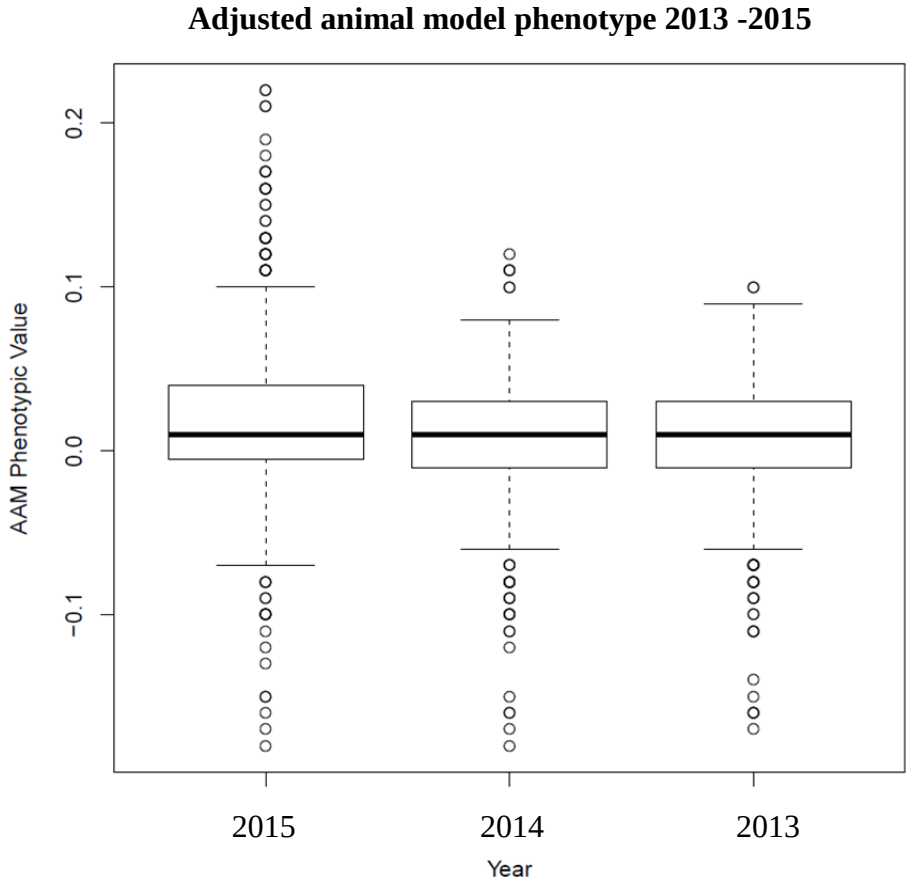


Figure 2.4-3: Adjusted animal model fertility values for 2013 - 2015 to identify sires with stable fertility phenotypes for SNP validation

Boxplots of overall adjusted animal model phenotype values for each individual year 2013, 2014 and 2015. Data is used for selection of sires to be used in validation of variant calling via whole-exome sequencing and targeted β-defensin sequencing. Overall adjusted animal model values are denoted on the y-axis. Year of phenotypic data collection is denoted on the x-axis.

2.4.2 Assay design

To design primers for the SNP validation assay, annotated sequence information for the markers of interest were provided. At least 100 base pairs of flanking sequence on each side of the SNP of interest were provided to obtain suitable primer binding sites. For each marker, the target SNP in the submitted sequence was annotated by placing square brackets around the polymorphic locus, using a forward slash to separate the alleles. Proximal SNPs in the flanking sequence were annotated using the International Union of

Pure and Applied Chemistry (IUPAC) convention. For full assay design files, see electronic appendix[14].

2.4.3 SNP validation

Agena Bioscience MassARRAY® System was used for assay design and SNP validation. In total, 58 SNPs were targeted for validation in 4 multiplex reactions (29-, 18-, 8-, and 3-plex) to prevent similar primer sequences binding to each other in the reaction, see the Electronic Appendix 5.4. Suitable primers for one SNP (4792) could not be designed, as the bases flanking the SNP do not differ sufficiently and therefore, cannot be distinguished from the insertion/ deletion base of interest. See electronic appendix[15] for full SNP validation assay design information.

In total, 123 sires were genotyped for the 58 SNPs in 4 multiplex reactions. DNA (10ng) of each sire was obtained from the ICBF DNA database, and was used for each multiplex reaction. gDNA concentrations were estimated using a Nanodrop ND-1000 spectrophotometer. gDNA was dried down overnight in a PCR-free environment and sent, in separate 384-well plates, to Agena Bioscience GmbH, Germany.

2.4.4 Data analysis

Of 58 SNPs targeted for validation in the MassARRAY assay design suite, 42 variants and 123 cattle pass filters and QC. SNPs which had a call rate < 80%, or minor allele frequency (MAF < 0.04) were filtered out of the dataset, see the Electronic Appendix 5.4. Hardy-Weinberg Equilibrium (HWE) of SNPs was determined, to identify deviations from HWE which may have been caused by systematic errors in genotyping, unexpected population structure or presence of homologous regions in the genome. To do this, the allele frequencies and expected counts were calculated, and a Fisher's exact test for contingency tables was performed by Sequenom (California, USA).

[14] Electronic Appendix 5.4 Validation\Assay Design\4 - validation_assay_primer_design.xslx

[15] Electronic Appendix 5.1 Validation\Assay Design\ 1 - 250_replex2.xslx

SNP frequencies of SNPs which passed call rate filtering were calculated using an in-house Perl script, which was a modified version of electronic appendix file[16].

Statistical analysis of association was performed using PLINK v1.90b3l. The basic association test is for a disease trait and is based on comparing asymptotic allele frequencies between cases and controls (High-fertility v low-fertility).

The following PLINK v1.90b3l commands were used to perform SNP association analysis of the genotyped validation SNPs, and fertility phenotypes (AAM and PR):

#bed file generated from .ped and .map input files, with cow as the model organism for chromosome number designation.

plink --file input_file_name --make-bed --out output_file --cow

Quantitative traits (AAM and PR) tested for association, using Wald test.
plink --file out_out --assoc --out assocofsnps --cow --allow-no-sex

See Electronic Appendix 5.4 to Electronic Appendix 5.11 for .map .ped and associated SNP files for both AAM and PR phenotypes.

[16] Electronic Appendix 4.4 3 - count_alleles_by_category_new.pl

3 Targeted β-defensin gene sequencing in divergent fertility bulls

3.1 Introduction

The inability of some cows to get pregnant from artificial insemination or from stock bulls can result in increased costs to the farmer including vet bills, scanning for cows in calf, repeat AI services, premature culling and lost productivity (e.g. milk production) (Shalloo et al., 2004).

In recent years, there has been an inverse correlation between cow fertility and milk production (Berry and Evans, 2014). The focus of the scientific community has been on female fertility traits. However, male fertility also plays an important role in determining whether the cow becomes pregnant, yet no single diagnostic test can predict bulls of low-fertility.

Sub-optimal fertility of a single bull can have a large impact on a herd, due to the predominance of single-sire mating in the form of stock bulls and small herd size. A mature bull will mate with, on average, 40 cows in each breeding season (Teagasc, 2016b). Any fertility issues in the bull will have a larger impact than fertility issues in any individual cow (Teagasc, 2016b). Male fertility is not explicitly factored into the EBI star rating for bulls, the only fertility measurement included is based on the bull's daughter's fertility (DPR) (Teagasc, 2014).

Our group has identified an expansion of β-defensin genes in the bovine genome, estimated to be 57 different genes, found on four *Bos taurus* chromosomes, 8, 13, 23 and 27. Bioinformatic analysis identified 19 novel β-defensins in the bovine genome located on chromosome 13, which were subsequently shown to be expressed in the male and female reproductive tracts, indicating a role in reproduction (Narciandi et al., 2011). Orthologs of these genes have been shown to be involved in male fertility in various species, including humans (Tollner et al., 2011), macaques (Tollner et al., 2008), mice (Zhou et al., 2013), rats (Zhao et al., 2011), and cattle (Narciandi et al., 2011).

Part of the larger research project investigates whether β-defensin genes and mutations affecting β-defensin gene function have an effect on bull fertility. For this, it is necessary to investigate whether differences in the fertility phenotype are a result of β-defensin variation, or whether other coding and 5'UTR variants play a role in regulating fertility.

At the time of designing the exome targeting probes in 2013, only seven of nineteen β-defensin genes located on chromosome 13, which form a group of genes identified by our group in cattle, were annotated in the bovine genome. These seven genes were *DEFB117, DEFB119, DEFB122, DEFB122a, DEFB123, DEFB124* and *DEFB129*. The remainder were not in exome sequencing design, due to poor annotation of the bovine genome, in comparison to more extensively studied species, such as human.

Targeted sequencing (TS) of bovine β-defensin genes in bulls divergent for fertility was performed.

3.2 Aims and hypothesis

Following the discovery by our group of an expansion of β-defensins and their expression in adult reproductive tracts, our aim in this chapter was to investigate whether variants in the β-defensin gene cluster are associated with male fertility in cattle. To do this, targeted β-defensin gene and regulatory region sequencing was performed in a large cohort of bulls divergent for a fertility phenotype, to identify candidate variants to be added to a custom-designed SNP chip for genotyping in a large population of cattle in Ireland.

The hypothesis for this chapter was: genetic variation in β-defensin genes and regulatory regions explain a portion of the phenotypic variation for fertility in AI bulls.

3.3 Results

3.3.1 Fertility phenotypes of AI sires

Sires were selected for sequencing based on the PR and AAM fertility phenotypes (as noted in section 2.1.1.) over four years (2010 – 2013). Sires of divergent fertility were selected, as described in section 2.1.2, after filtering. The percentage PR of males used in >1,000 matings ranged from 20% to 70% with a mean of 49.17%. The AAM fertility phenotype ranged from -0.21 to 0.12 with a mean of 0.01744. The groups of high and low-fertility were defined as greater than one standard deviation from the mean, percentage PR < 42.6% or > 55.74%, adjusted fertility phenotype < -0.0167 or > 0.0515. Restricting the samples to sires which have been used in more than 1,000 matings increased the reliability of the fertility phenotypes but removed sires identified as having most divergent percentage PRs based on small numbers of matings. Fertility phenotype data of sires selected for PR sequencing are shown in Figure 3.3-1 for adjusted animal model in various breeds, and Figure 3.3-2 for divergent phenotype identification of sires with 1 standard deviation above and below the mean. Figure 3.3-3 shows the sires which were selected based on their fertility phenotypes and which had gDNA already extracted and in the Teagasc DNA databank.

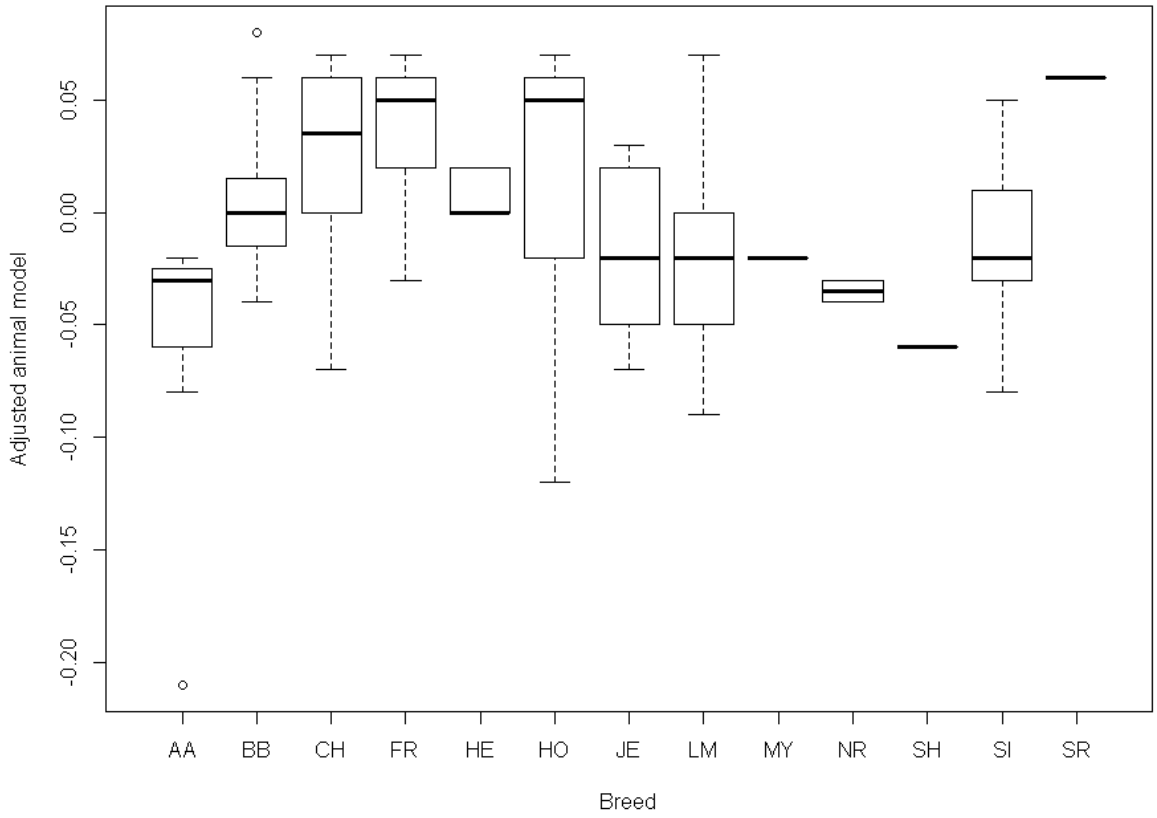


Figure 3.3-1: Adjusted Animal model fertility phenotype per breed – data from 7000 sires for sample selection.

Boxplots of adjusted animal model fertility phenotype values for all 7000 bulls for which we had fertility phenotypes, grouped by breed. Sire breed is shown on the X-axis and the adjusted animal model fertility phenotype values on the Y axis; Aberdeen Angus (AA), Belgian Blue (BB), Charolais (CH), Friesian (FR), Hereford (HE), Holstein (HO), Jersey (JE), Limousin (LM), Montbéliarde (MY), Norwegian red (NR), Shorthorn (SH), Simmental (SI), and Saler (SR).

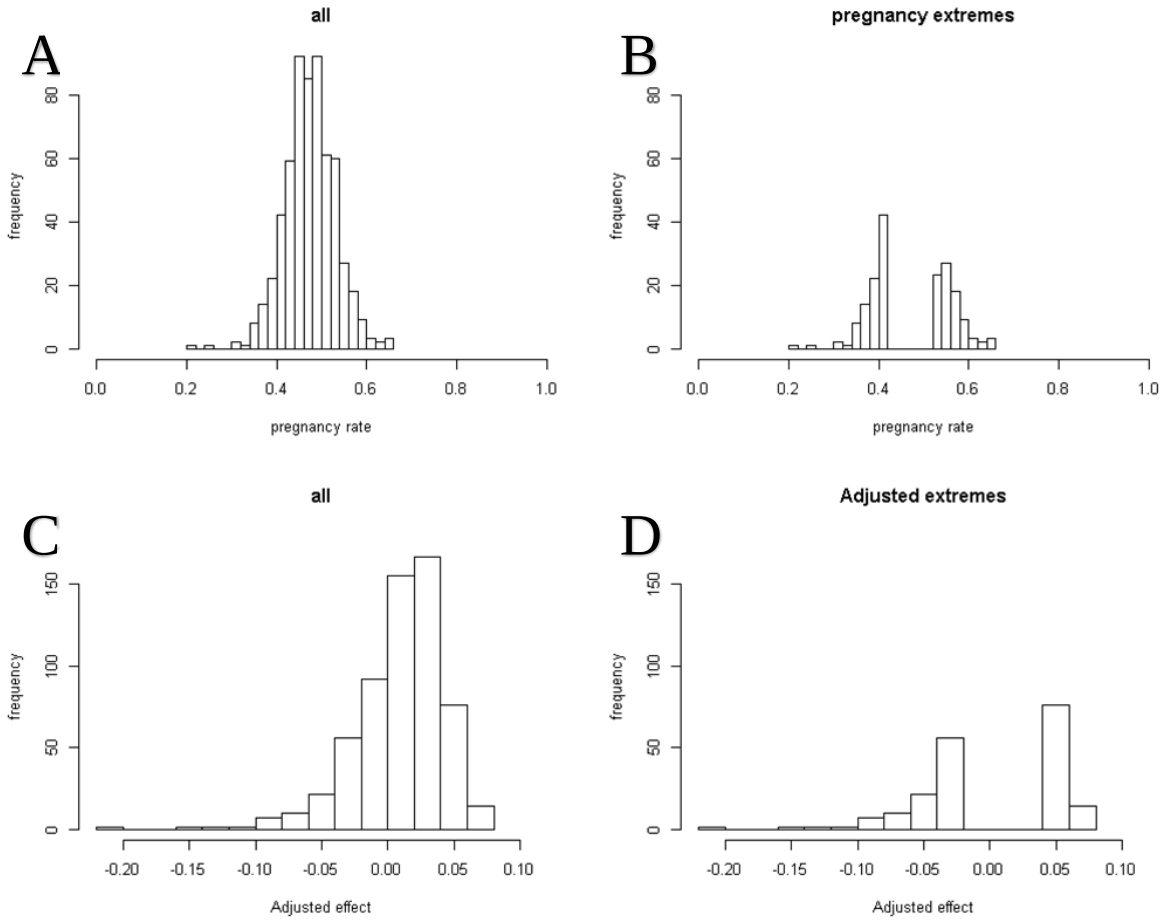


Figure 3.3-2: Identification of sires divergent for fertility

A) PR % in all sires (n=683). B) PR % in sires divergent for fertility (1 s.d.). C) AAM fertility phenotype data for all sires (n=683). D) AAM fertility phenotype frequency data for sires divergent for fertility (1 s.d.).

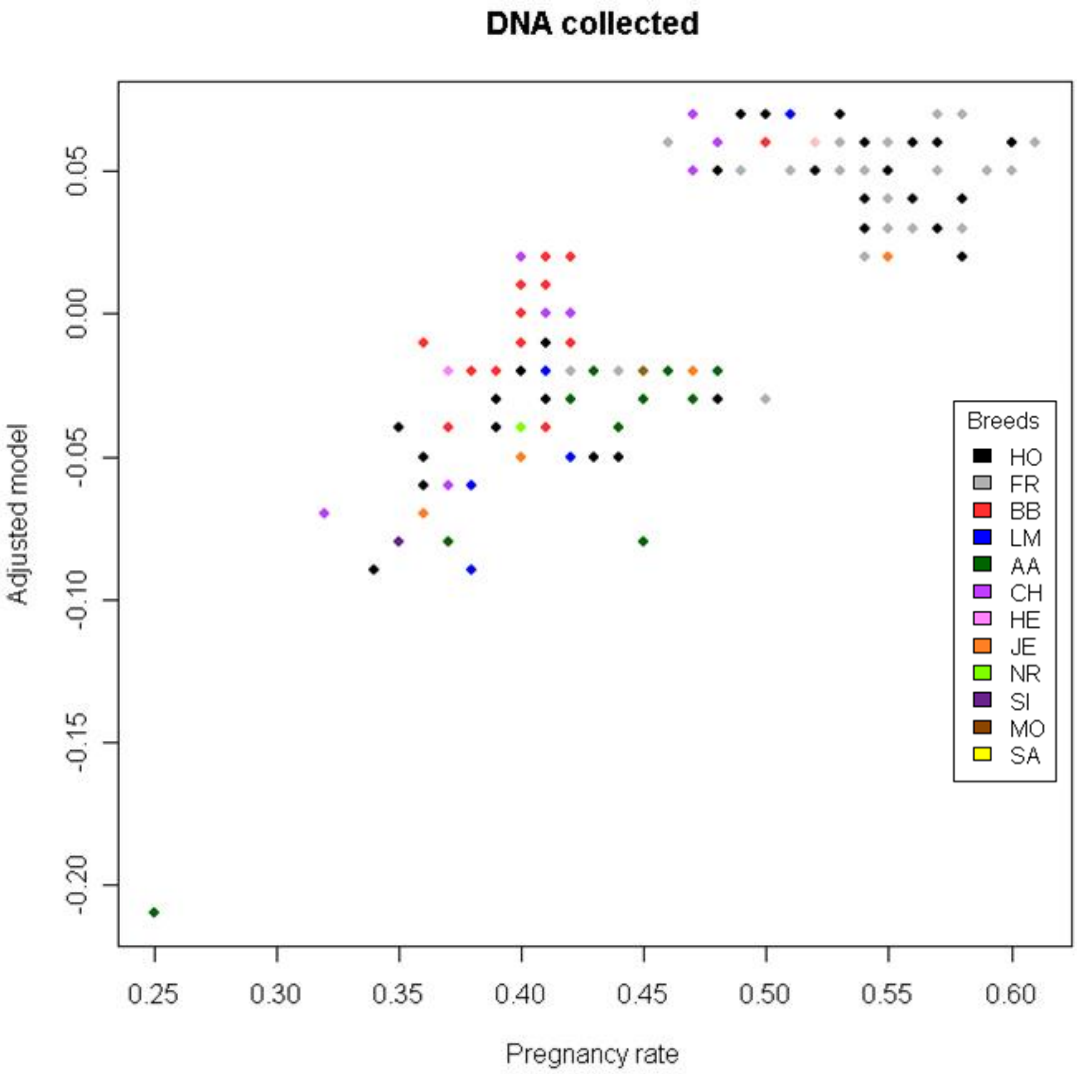


Figure 3.3-3: Phenotypic values of DNA collected from sires in Teagasc's DNA databank.
R scatterplot of phenotypic values for PR and adjusted animal model for each sire selected for targeted sequencing, with dots coloured per breed. X-axis denotes the PR phenotypic value. Y-axis denotes the adjusted animal model phenotypic value. Each dot represents a sire selected for targeted sequencing, coloured per breed, to show the range of sire breeds collected.

3.3.2 Targeted β-defensin sequencing coverage statistics

Of the total 235 kb targeted region for β-defensin sequencing, an average of 8.29%, ranging from 3.2% to 19.7%, had less than 1X coverage. This means an average of 91.71%, ranging from 96.8% to 80.3% of the targeted region was captured and sequenced by the Roche Nimblegen developer capture design. A mean of 88% of the 234,754 bp targeted per sample had a sequencing coverage of >10X, the median proportion with >30X coverage was 84%. The total number of reads per sample ranged from 68,247 to 4,924,576 with a mean value of 460,862, see Table 3.3-1.

For internal quality control, 9 animals were sequenced in duplicate in different captures and sequencing runs. After GATK variant calling, the genotypes called in each duplicate were compared. One pair had only 64% identity, so both were removed. The other 8 had an average genotype identity of 96.8% (95.9 – 97.4%). The copy with the highest call rate of each pair was retained for analysis.

Mean total reads were 460,862 for all 168 sequenced sires, of which 83% were uniquely aligned to UMD3.1 bovine genome. As shown in Table 3.3-2, 58.5% of bases were aligned, on-target after PCR duplicate removal. This resulted in a mean of 8,009-fold-enrichment of targeted regions compared to non-targeted regions (background), ranging from 830 to 9,929. High mean coverage over targeted regions (84% at 30X) was deemed sufficient for GATK variant calling following best practice pipelines.

Given the high levels of read coverage, the percentage target region covered is low. This means that despite the abundance of probes, certain areas could not be captured by the probe design. In total, 16% of aligned bases failed to map on bait, and 22% of target regions had less than 2X coverage. The reason for low targeted coverage is most likely due to properties of the targeted region, as we can see from Figure 3.3-8, on page 92. In a representative sample (bull 6), non-targeted regions are clustered together (e.g. gene 57-61), located on chromosome 27 between 5483406-5638581. This may be due to repetitive sequences in the region, as the β-defensin gene region is amongst the most copy number variable regions in the human genome (Hollox et al., 2008).

Table 3.3-1: Targeted β-defensin sequencing coverage statistics

Table of coverage statistics for targeted β-defensin sequencing of the large AI bull population (n=168). Mean coverage of targets is the average number of times each targeted region in the β-defensin probe design contains a mapped read after filtering. The remaining columns show the percentage of targeted regions with the number of sequencing reads mapping to that region. The mean, min and max values for all sires in the targeted β-defensin sequencing project are displayed.

	Mean coverage of targets	**% < 1X**	**%> 2X**	**% > 10X**	**% > 20X**	**% > 30X**	**% > 40X**	**% > 50X**	**% > 100X**
Mean	325.05	8.29%	90.96%	87.66%	85.52%	83.76%	82.07%	80.36%	72.20%
Min	35.04	3.25%	85.45%	77.44%	71.33%	65.25%	42.35%	15.30%	0.07%
Max	1189.46	19.68%	95.16%	93.31%	92.12%	91.20%	90.54%	89.90%	86.63%

Table 3.3-2: Targeted β-defensin sequencing alignment and enrichment statistics

Table displaying the alignment of sequencing reads to the bovine genome, and fold enrichment of the targeted regions. Total number of reads for all sequenced sires, percentage of those reads which are unique (reports the best alignment), number of bases aligned, percentage of aligned bases in or near targeted area, percentage of aligned, de-duplicated and on-target bases, and targeted region fold enrichment values are all reported in this table. The mean, min and max for all sequenced sires in the targeted β-defensin sequencing project are displayed. Full sequencing alignment and enrichment statistics are available in the electronic appendix[17].

	Total reads	**Unique reads**	**Bases aligned**	**Aligned bases in or near targeted area**	**Aligned, de-duped, on-target bases**	**Targeted region fold enrichment**
Mean	460862	82.54%	107000000	86.41%	58.47%	8009.1
Min	68247	36.36%	15792716	9.23%	4.95%	830.2
Max	4924576	96.38%	940000000	98.36%	83.32%	9929.6

[17] Electronic Appendix 3.3 targeted_combined_metrics - sequencing stats - exome subset.xslx

3.3.3 Targeted sequencing variant discovery and filtering

In total, after GATK HaplotypeCaller SNP calling, there were 92,829 unfiltered variants from all reads that had mapped to UMD 3.1 bovine genome. Following variant filtering and removal of variants with call rates < 80%, 3,134 variants remained. These variants consisted of 2,892 SNPs, 105 insertions and 137 deletions.

Variants were located throughout the complete gene sequence of β-defensin genes, and are summarised in Figure 3.3-4.

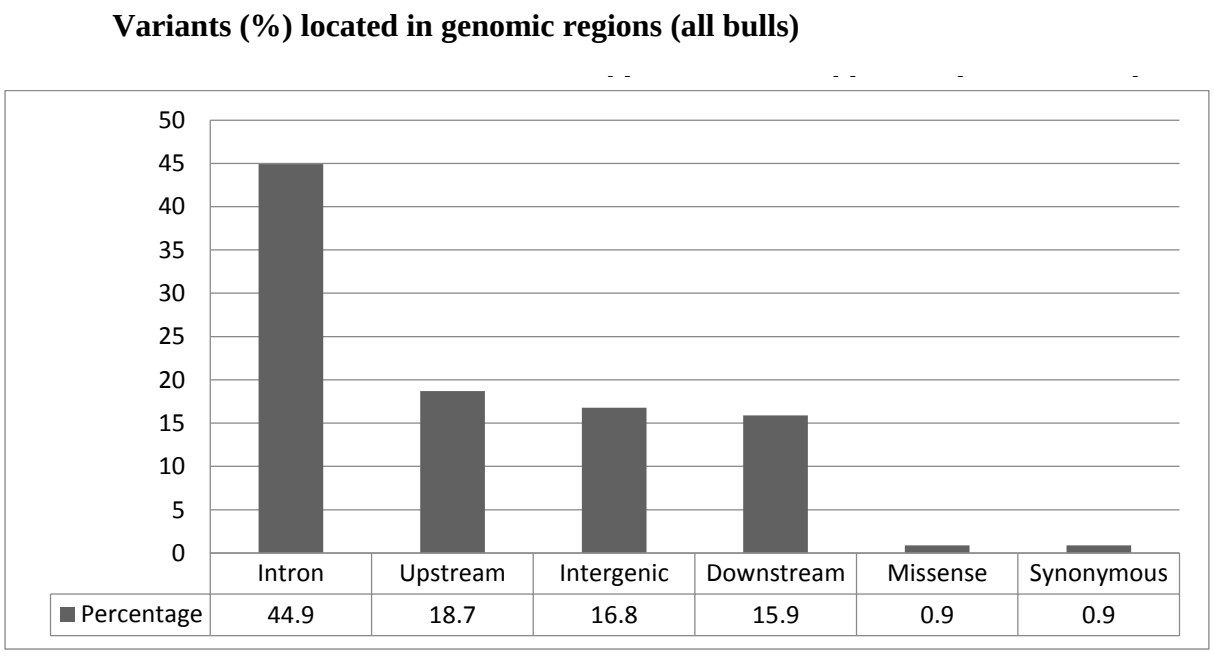


Figure 3.3-4: Percentage of variants located in genomic regions for targeted β-defensin sequencing (all bulls).
Bar chart of percentage of total SNPs identified in TS of 168 sires in various genomic features. The X-axis denotes the genomic features where SNPs are located, Intron (Non-coding sections of DNA), Upstream (5'UTR of a gene), Intergenic region (Sequence between genes), Downstream (3'UTR of a gene), Missense (SNP results in a codon that codes for a different amino acid), Synonymous (SNP in exon that does not alter amino acid code).

3.3.4 Targeted sequencing variant association analysis

Variants were further filtered to remove variants of low minor allele frequency (MAF < 0.05) (1,578 variants), less than 95% call rate (907 variants), and variants not in Hardy-Weinberg equilibrium (913 variants). The check.marker quality control tool also suggested high Identity by state (IBS) in 33 bulls. However, this tool is optimised for whole-genome association data, and this high level of IBS is not unusual for such a short region of the genome. These animals were not removed from the analysis. After quality control filtering, there were 1,399 SNPs in 149 bulls, 25 on chromosome 8, 626 on chromosome 13, 356 on chromosome 23 and 392 on chromosome 27.

The variants most associated with the adjusted animal model fertility phenotype from targeted β-defensin sequencing are located on chromosome 13, see Figure 3.3-5. The SNP most associated is rs378043559 (unadjusted *P*-value = 0.00197), located in the upstream region of *DEFB127* at position 61340027 on chromosome 13. Interestingly, a group of 97 SNPs have a similar *P*-value = 0.00202. This group of SNPs are located on chromosome 13 and are inherited as a haplotype. Nine sires (5 Holstein-Friesian, 1 each of Limousin, Simmental, Charolais and Belgian Blue) are heterozygous, whilst all other sires have reference alleles. The 9 heterozygous sires are of medium to high fertility, according to the adjusted animal model phenotype (overall AAM = 0 to 0.07, mean = 0.04, s.d. = 0.027).

The 97 SNPs lie between positions 61329700 to 61467209, a region of 137.5 kb containing the genes *BBD128, BBD127, BBD126, BBD125, BBD115, BBD142*, and *BBD116*, see Figure 3.3-6. In Figure 3.3-7, the sires found to contain the haplotype of SNPs located on chromosome 13 are highlighted in red. Of the 97 SNPs in the haplotype, 3 lie within coding regions, a non-synonymous SNP in *BBD115* (Ser52Asn), two synonymous SNPs in *BBD126* and *BBD125*. 76 SNPs lie within introns, 5 downstream and 13 upstream of genes, 18 are intergenic.

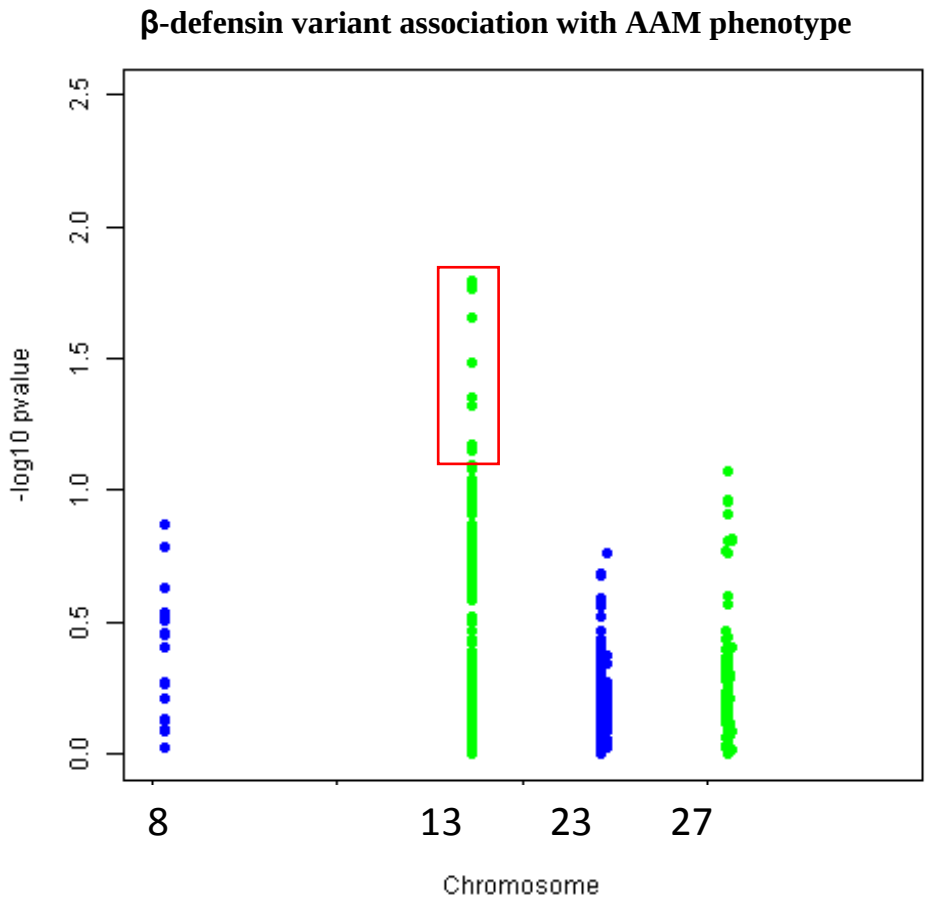


Figure 3.3-5: Analysis of variants located on chromosomes 8, 13, 23 and 27 associated with adjusted animal model fertility phenotype

Manhattan plot shows the association of variants identified by targeted β-defensin sequencing and their associated –log10 *P*-value with the adjusted animal model fertility phenotype. The –log10 *P*-value for each variant association is on the y-axis. The chromosomal position of each variant is on the x-axis. Red rectangle = SNPs most associated with AAM.

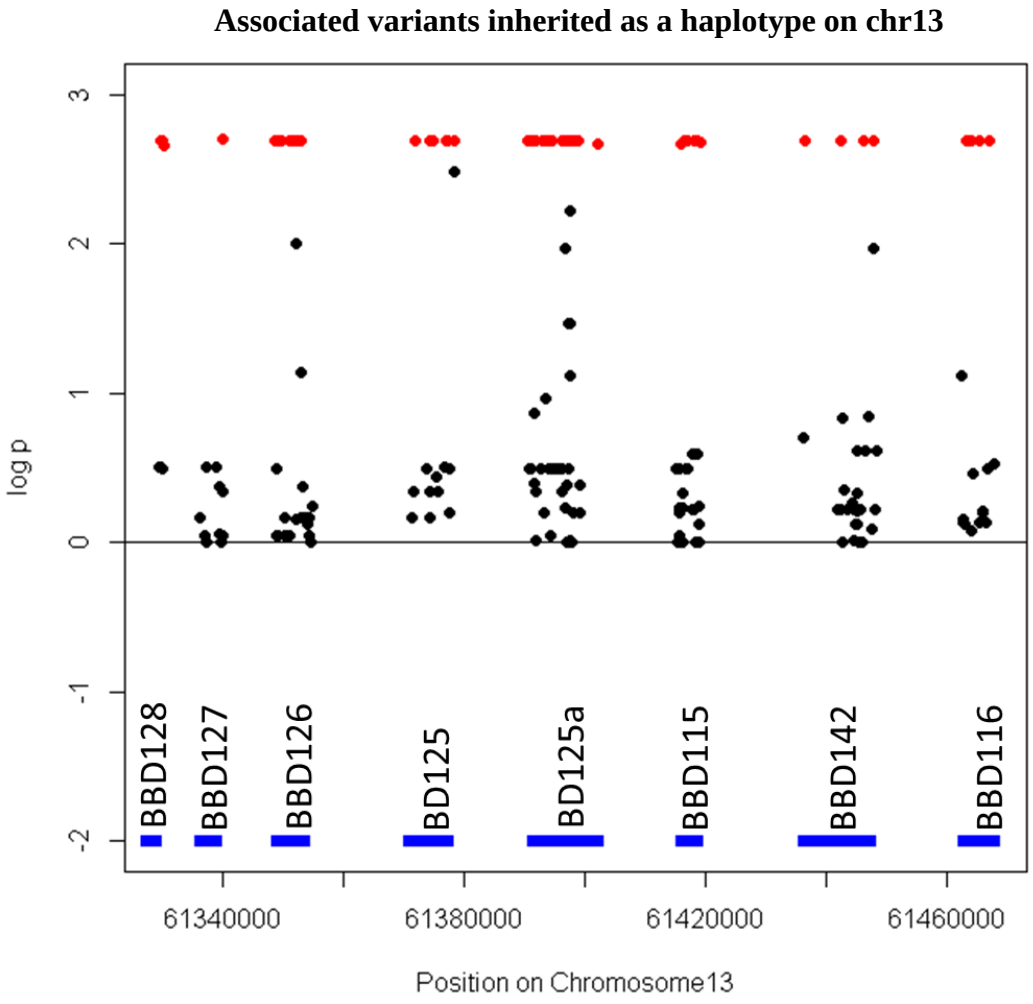


Figure 3.3-6: Variants located in β-defensin genes on chromosome 13 associated with adjusted animal model phenotype with variants inherited as an haplotype.

Variants identified via targeted β-defensin sequencing located on chromosome 13, with position on x-axis, and associated *P*-value for each variant on the y-axis. Variants are associated with the AAM fertility phenotype. β-defensin genes located on chromosome 13 are highlighted in blue, with a group of variants inherited as an haplogroup highlighted in red.

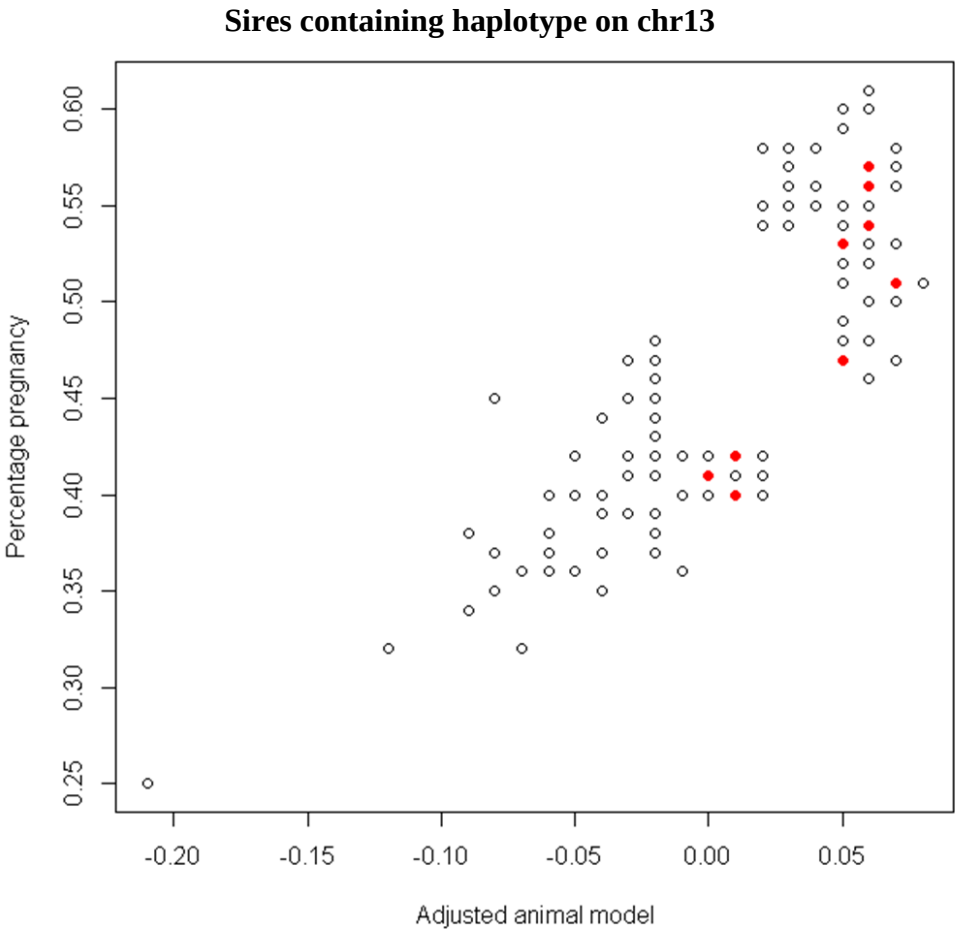


Figure 3.3-7: Scatterplot of sires identified as being divergent for fertility phenotypes AAM and PR, with those containing haplotype located in chromosome 13 highlighted in red.
Scatterplot of variants identified via targeted β-defensin sequencing in the phenotypic groups of fertility for AAM and PR. Variants located in chromosome 13 haplotype are highlighted in red. Variants not in the divergent fertility groups (+/- 1 s.d. from mean) for both AAM and PR have been blanked out.

To determine which SNPs are biologically important, SnpEff predicted the functional effect of each SNP. The 59 missense SNPs identified from the targeted β-defensin sequencing in a large cohort of sires, were located in 32 of the targeted genes. Two of these SNPs were of particular interest due to a change in one of the cysteine residues, which are characteristic of β-defensin genes. Rs477570826, located at position 5162114 on chromosome 27 resulted in a Cys34Ser amino acid change in the second cysteine residue in ENSBTAG00000047421. This SNP is heterozygous in two sires of medium fertility.

Rs437613002 located at position 7287847 on chromosome 8 results in a Cys55Arg (5^{th} Cysteine) amino acid change in *BBD131* gene. The SNP is homozygous in one sire of low fertility and heterozygous in 6 sires of low, average and high fertility.

3.3.5 Targeted sequencing coverage in subset of sires

Targeted β-defensin re-sequencing was performed on the same 24 bulls as whole-exome sequencing. This ensured for 24 sires, TS and WES data were available. For full targeted sequencing coverage statistics in the 24 bull subset, see the electronic appendix 3.2[18].

Targeted re-sequencing attempted to sequence 69 genes in total, including β-defensins and cathelicidins. The β-defensin gene region is approximately 378kb in length. However, targeting of this region with custom-designed capture baits was limited to 235kb, due to repetitive regions. Targeted sequencing statistics are shown in Table 3.3-3. Mean bait coverage over all sites for targeted re-sequencing of the 24 bulls cohort, selected for whole-exome sequencing, was 435X, ranging between 218X to 602X. Mean percentage PCR duplication was 54%, ranging from 45.9% to 75.6%, meaning 54% of reads aligned to the same position in the genome and contained the same CIGAR string. This is probably due to the large number of baits available in the capture design (approximately 2.1 million), compared to the small target area (235kb). Baits would have a limited number of target sequences and so large amounts of duplication.

Percentage of bases covered at 30X was 66%, with 78% at 2X coverage, meaning 22% of targeted bases have less than 2X coverage. Coverage statistics for one sample indicative of other samples (bull 6), are shown in Figure 3.3-8 on page 92. This sample highlights the regions of low coverage in the targeted re-sequencing probe design. In particular, poor sequencing coverage is evident between genes 55 and 64, located on chromosome 27 between positions 5473206-6196146.

To confirm WES reads are predominantly aligned to exons, and to highlight the success of the custom-designed bait capture protocol for identifying exome genetic variants, an annotated gene of interest was selected for viewing via integrated genomics viewer (IGV). *DEFB122a* which was annotated in the bovine genome and was included in exome capture was visualised in IGV to show depth of coverage, a coding region SNP and 2 5'UTR SNPs; see Figure 3.3-9.

[18] Electronic Appendix 3.2 Targeted Sequencing coverage Stats - Exome subset.xslx

Table 3.3-3: Summary of coverage statistics for targeted β-defensin re-sequencing

Table of read coverage for targeted re-sequencing of β-defensin and cathelicidins genes. Information on bull Sample identification, total number of reads per sample, % of reads aligned to bovine genome, average coverage of each targeted sequence, % of sequenced bases which map on targeted probes, fold enrichment of targeted region (more than would be expected by chance), and 2X, 10X and 30X read coverage are all shown here.

Sample ID	Total Reads	% Aligned	Avg. Coverage	% Bases on Target	Fold Enrichment	% 2X	% 10X	% 30X
Holstein-Friesian High								
Bull 1	438545	92.2	288.7	67.1	6619.8	79.2	72.2	66.8
Bull 2	651339	94.6	449.5	70.4	6741.7	78.6	72.2	66.4
Bull 3	544935	94.4	371.2	69.6	6728.1	78.9	73.1	66.9
Bull 4	630665	94.4	440.2	71.8	6864.5	78.5	71.8	65.4
Bull 5	649537	94.5	450.5	71.2	6822.2	78.6	71.8	65.9
Bull 6	340382	91.3	218.1	65.3	6515.4	79.5	72.9	66.6
Holstein-Friesian Low								
Bull 7	681106	96.3	483.9	72.1	6824.5	78.9	72.7	66.1
Bull 8	661779	95.1	458.9	70.8	6800.1	79.1	72.5	66.6
Bull 9	701543	95.8	492.4	71.5	6806.2	78.8	72.5	66.9
Bull 10	560394	95.3	389.3	70.6	6743.7	78.8	73.1	66.9
Bull 11	714683	95.5	499.7	71.4	6811.3	77.5	71.3	66.2
Bull 12	855507	95.7	602.7	71.6	6825.2	78.9	72.1	66.6
Limousin High								
Bull 13	792361	95.6	556.2	71.6	6825.6	79	73.5	67.6
Bull 14	646738	95.4	451.1	71.1	6811.5	78.7	71.9	67.5
Bull 15	734337	95.4	513.8	71.3	6828.5	79.9	74.1	67.7
Limousin Low								
Bull 16	632236	95.9	447.1	71.9	6831.7	78.5	72.4	66
Bull 17	567005	94.2	387.4	69.7	6747.3	79.5	73.9	68.3

Table 3.3-4 continued

Bull 18	598511	96	424.2	72.1	6843.2	78.9	72.5	66.8
Belgian Blue High								
Bull 19	641749	95.4	445.9	70.9	6776.7	78.8	72.5	67
Bull 20	649660	94.9	449.8	70.6	6779	79.6	73.5	68.1
Bull 21	603576	95.6	423.6	71.3	6799.1	78	71.6	66
Belgian Blue Low								
Bull 22	616297	95.4	429.3	71.1	6796.1	79.1	72.7	67
Bull 23	461952	94.1	314.4	69.2	6702.3	78.6	72.4	66.4
Bull 24	668819	94.9	464.5	70.8	6809.7	79.1	72.9	66.4
Min	340382	91.3	218.1	65.3	6515.4	77.5	71.3	65.4
Mean	626819	94.9	435.5	70	6777.2	78.9	72.6	66.8
Max	855507	96.3	602.7	72.1	6864.5	79.9	74.1	68.3

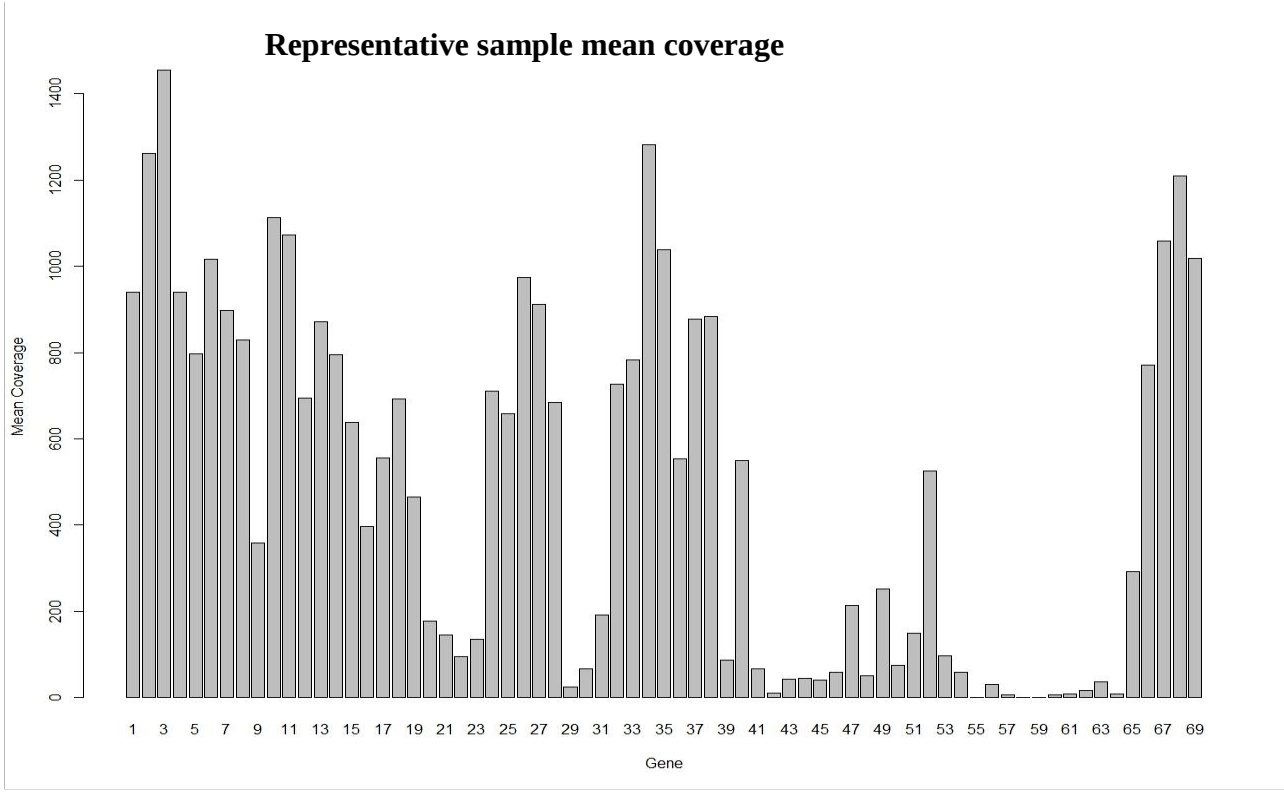


Figure 3.3-8: Mean coverage per gene in representative sample.

Targeted regions are shown on x-axis. Each region is a targeted gene or gene region, with 69 in total. Mean coverage is shown on the y-axis. Coverage stats show the variation in coverage amongst targeted regions for β-defensin targeted sequencing. Low coverage may be due to repetitive regions resulting in low probe binding or capture efficiency.

***DEFB122a* reads mapping to exon regions**

Figure 3.3-9: *DEFB122a* IGV - reads mapping to exons and depth of coverage.
Integrated genomics viewer (IGV) image of A) sequencing reads mapping to exonic regions of *DEFB122A* in 1 sire and B) Reads aligned to exon 1 of *DEFB122a*, showing SNP in coding region and 2 SNPs in 5'UTR. Grey bars are quality filtered sequencing reads aligned to UMD3.1 bovine genome. Multiple aligned reads show depth of coverage at individual positions across the targeted region.

3.3.6 Variant discovery

Variant discovery following GATK's best practice guidelines was performed and identified 4,948 unfiltered SNPs. After filtering SNPs based on read depth and to SNPs located within the targeted regions, 274 SNPs remained. Annotation of these SNPs identified two high-effect variants, as determined by SnpEff, which are predicted to result in a stop-gained mutation in *BBD118* and *DEFB4α*. SnpEff, the variant predictor package, was used to predict the location, class and effect of SNPs. Examples of high effect variants include exon deletion, frame shift, start lost, stop lost and stop gained etc. *BBD118* is part of the cluster of 19 β-defensin genes, located on chromosome 13, which were identified by our group as being an expansion of β-defensin genes in cattle, and are expressed in the male reproductive tract.

3.3.7 Annotation

Following variant discovery, SnpEff, the variant annotation and effect predictor tool predicted the effects of genomic changes (such as amino acid changes) of the variants on the genes. Of the 274 SNPs identified following filtering, 2.27% were predicted to be located in exons, and 0.41% were located in the 5'UTR. Intergenic and intronic SNPs comprised 85% of these SNPs. Of the exonic SNPs, 1.47% were non-synonymous, and 0.68% were synonymous. This targeted capture design is inherently different to the exome target design, as this targeted capture includes 1,000bp 5'UTR, and all introns for each of the 69 genes, as well as the exon regions. This explains the differences in percentage of variants located in exons.

SNP frequencies for breed-specific SNPs cannot be performed on this dataset due to the low numbers in each group (3 per fertility phenotype, except for HF which is 6 per fertility phenotype).

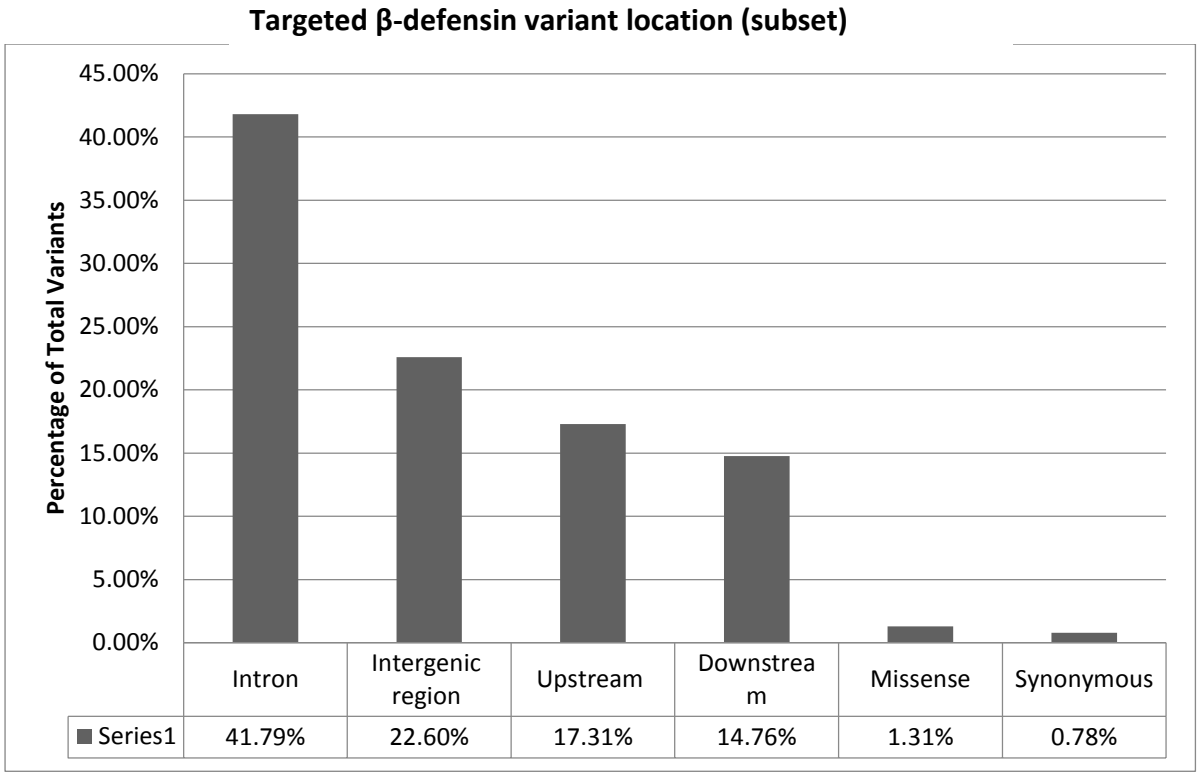


Figure 3.3-10: Targeted sequencing SNP location (subset of bulls).
Bar chart of percentage of total SNPs identified in targeted β-defensin sequencing for 24 bulls selected for whole-exome sequencing located in various genomic features. X-axis denotes the genomic features where SNPs are located, intron (non-coding sections of DNA), Intergenic region, Upstream (5'UTR of a gene), downstream (3'UTR of a gene), missense (SNP results in a codon that codes for a different amino acid), and synonymous (SNP in exon that does not alter amino acid code).

3.3.8 β- defensin SNP frequency analysis

SNP frequency analysis between high and low-fertility groups was performed on the 274 SNPs identified in targeted β-defensin re-sequencing. Of these 274, 20 SNPs were found to have a SNP frequency difference between groups of greater than 25%, as large SNP frequency differentials between groups indicate selective pressures. SNPs were also filtered to ensure a genotyping call rate > 80% in animals for each SNP. These 20 SNPs are shown in Figure 3.3-11, on page 96. SNP frequency differences were similar for most SNPs, with 33 % being the highest SNP frequency difference between groups. Seven SNPs had this SNP frequency difference, located in *DEFB122*, *DEFB123*, and *BNBD6*.

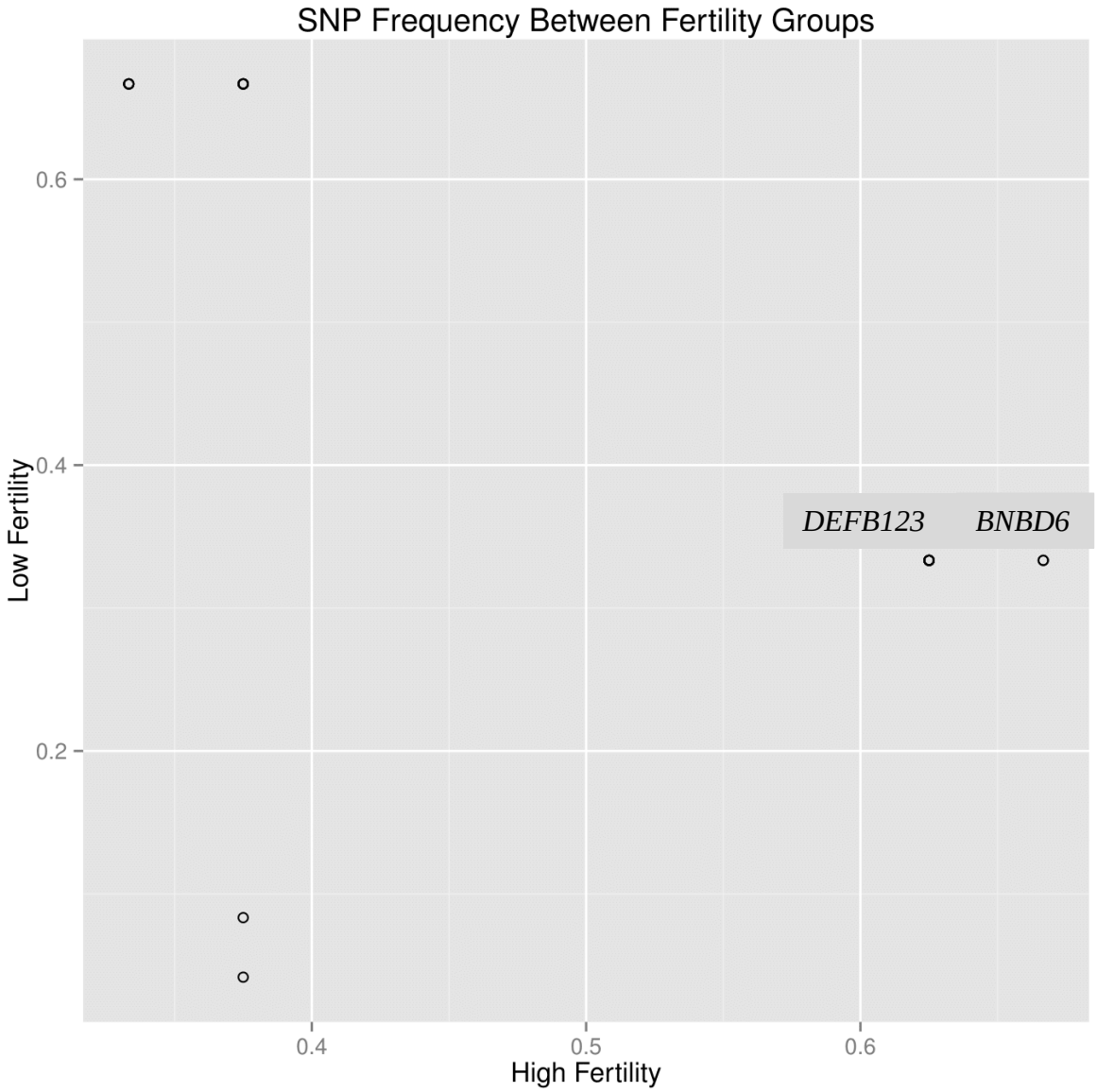


Figure 3.3-11: SNP frequency scatter plot between high-fertility and low-fertility groups for targeted β-defensin genes.

Each point denotes a filtered targeted β-defensin re-sequencing variant (n=20). Multiple SNPs are plotted on top of each other as they have the same frequency differentials, with a SNP frequency difference between high- and low-fertility groups of greater than 25%. The x-axis denotes the frequency of the alternate allele for 12 samples in the high-fertility group, and the y-axis denotes frequency of the alternate allele for 12 samples in the low-fertility group.

3.3.9 Targeted β-defensin SNP association in subset of sires

The SNP most significantly associated with AAM from targeted β-defensin re-sequencing in subset of 24 sires was g.5296571A<G, *P*=0.005. This SNP is located in the intron of *BBD138*. The second and third most significantly associated SNP are both located in the BNBD6 gene.

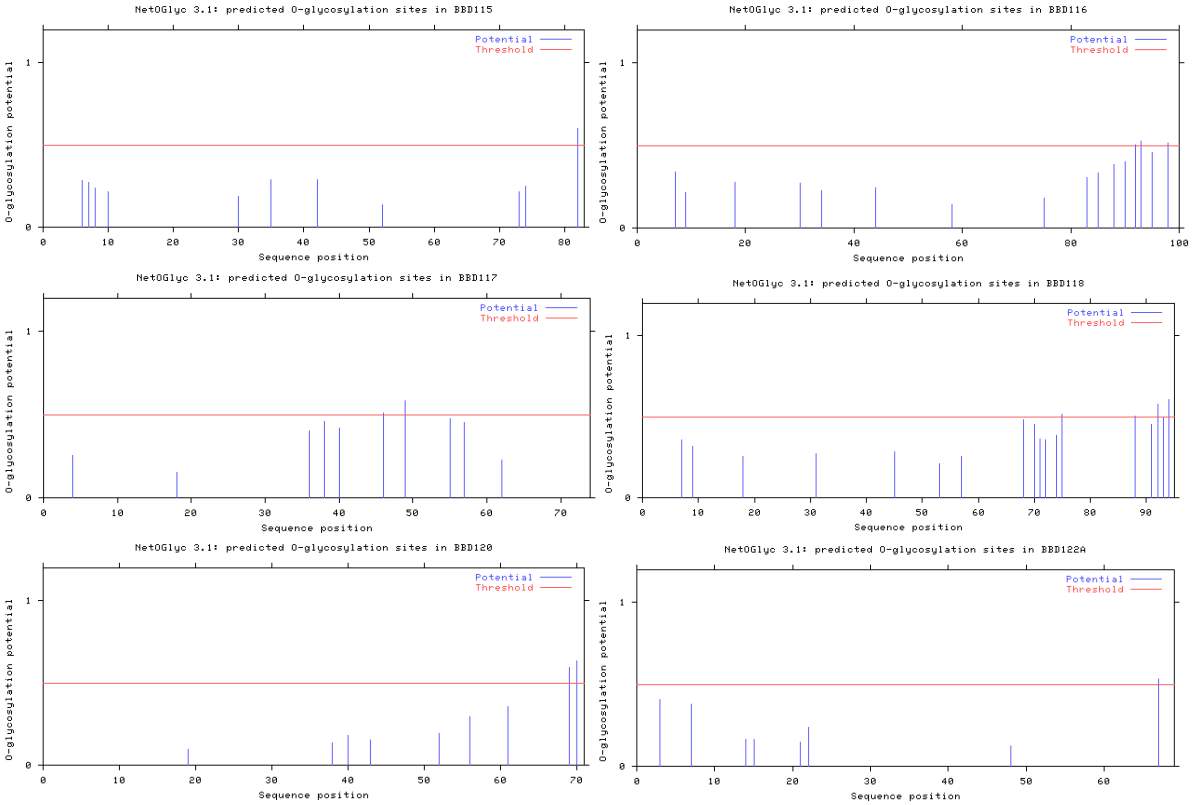
Table 3.3-5: Top 20 targeted SNPs associated with fertility phenotype.
Table of top 20 SNPs from targeted β-defensin and cathelicidins gene sequencing associated with an Adjusted Animal Model, a quantitative trait. Unadjusted P-values were considered significant at $P < 0.05$. Following Benjamini-Hochberg multiple testing correction, no SNP was statistically significant at $P < 0.05$. A full list of defensin SNPs associated with fertility are shown in the electronic appendix[19].

No.	ID	Chromosome	Position	P-Value	Gene
1	rs377872690	27	5296571	0.00569	*BBD138*
2	rs210662027	27	5162095	0.00572	*BNBD6*
3	rs209040542	27	5162243	0.00572	*BNBD6*
4	rs134258076	13	61529933	0.00588	*DEFB119*
5	rs135560000	13	61529943	0.00588	*DEFB119*
6	rs43708181	13	61563319	0.00629	*DEFB122A*
7	rs383291516	13	61315310	0.00645	*BBD129*
8	rs209203024	13	61339530	0.00645	*BBD127*
9	rs380883000	13	61340123	0.00645	*BBD127*
10	rs208521215	13	61353275	0.00645	*BBD126*
11	rs208842439	13	61375380	0.00645	*BBD125A*
12	rs207983987	13	61397108	0.00645	*BBD125*
13	rs207932860	13	61436311	0.00645	*BBD142*
14	rs211216958	13	61447139	0.00645	*BBD142*
15	rs43708180	13	61563327	0.00792	*DEFB122A*
16	rs109671949	13	61519650	0.00891	*DEFB118*
17	rs110402015	13	61566133	0.01028	*DEFB122A*
18	rs43708157	13	61566114	0.01028	*DEFB122A*
19	rs110199159	13	61566125	0.01028	*DEFB122A*
20	rs43708155	13	61566664	0.01028	*DEFB122A*

[19] Electronic Appendix 3.1 Defensin_Association_Multiple_Testing.txt

3.3.10 O-linked glycosylation analysis in β-defensin genes

O-linked glycosylation sites were predicted *in silico* for all 19 β-defensin genes located on chromosome 13 using NetOGlyc 3.1 available at (Joshi, 2017). Graphs of β-defensin genes with predicted glycosylation sites are shown in Figure 3.3-12. Graphs of β-defensin genes without predicted glycosylation sites are shown in Figure 3.3-13. In total 46 predicted glycosylation sites were identified, located in the following genes, *BBD115* (1), *BBD116* (2), *BBD117* (2), *BBD118* (4), *BBD120* (2), *BBD122a* (1), *BBD125* (6), *BBD125a* (8), *BBD126* (8), *BBD127* (1), *BBD129* (9) and *BBD132* (2). These 12 genes contained at least one predicted glycosylation site above a threshold determined by the NetOGlyc 3.1 tool. Numbers of glycosylation sites are indicated in brackets. Seven genes, *BBD119*, *BBD121*, *BBD122*, *BBD123*, *BBD124*, *BBD128*, and *BBD142* contained no predicted O-linked glycosylation sites.



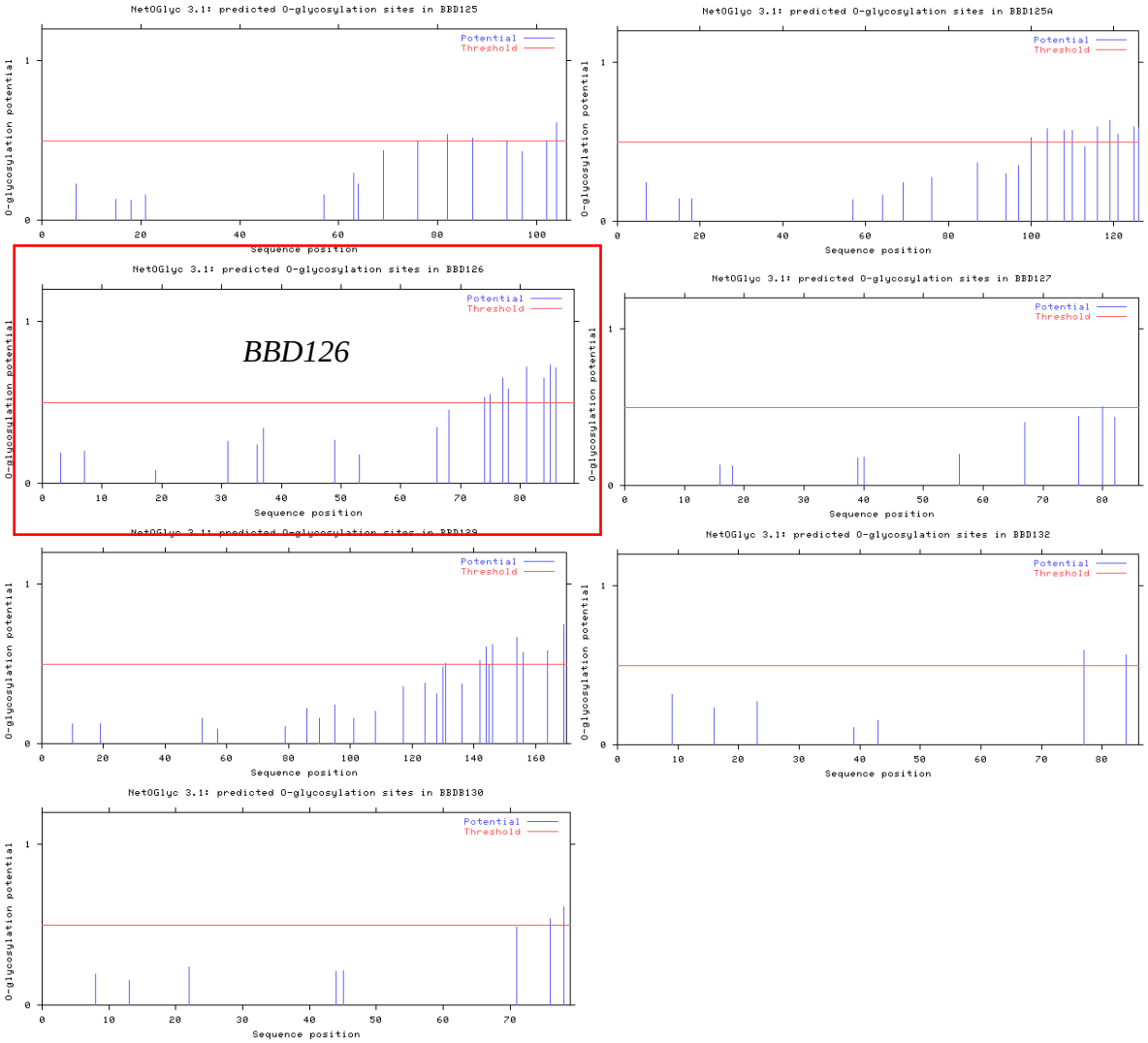


Figure 3.3-12: O-linked glycosylation analysis in β-defensin genes with predicted glycosylation sites.

NetOGlyc 3.1 graphs of predicted glycosylation sites in β-defensin genes. X-axis denotes the sequence position of the glycosylation site in the coding region of the β-defensin gene. Y-axis is the O-linked glycosylation potential, and red threshold line denotes the potential of glycosylation.

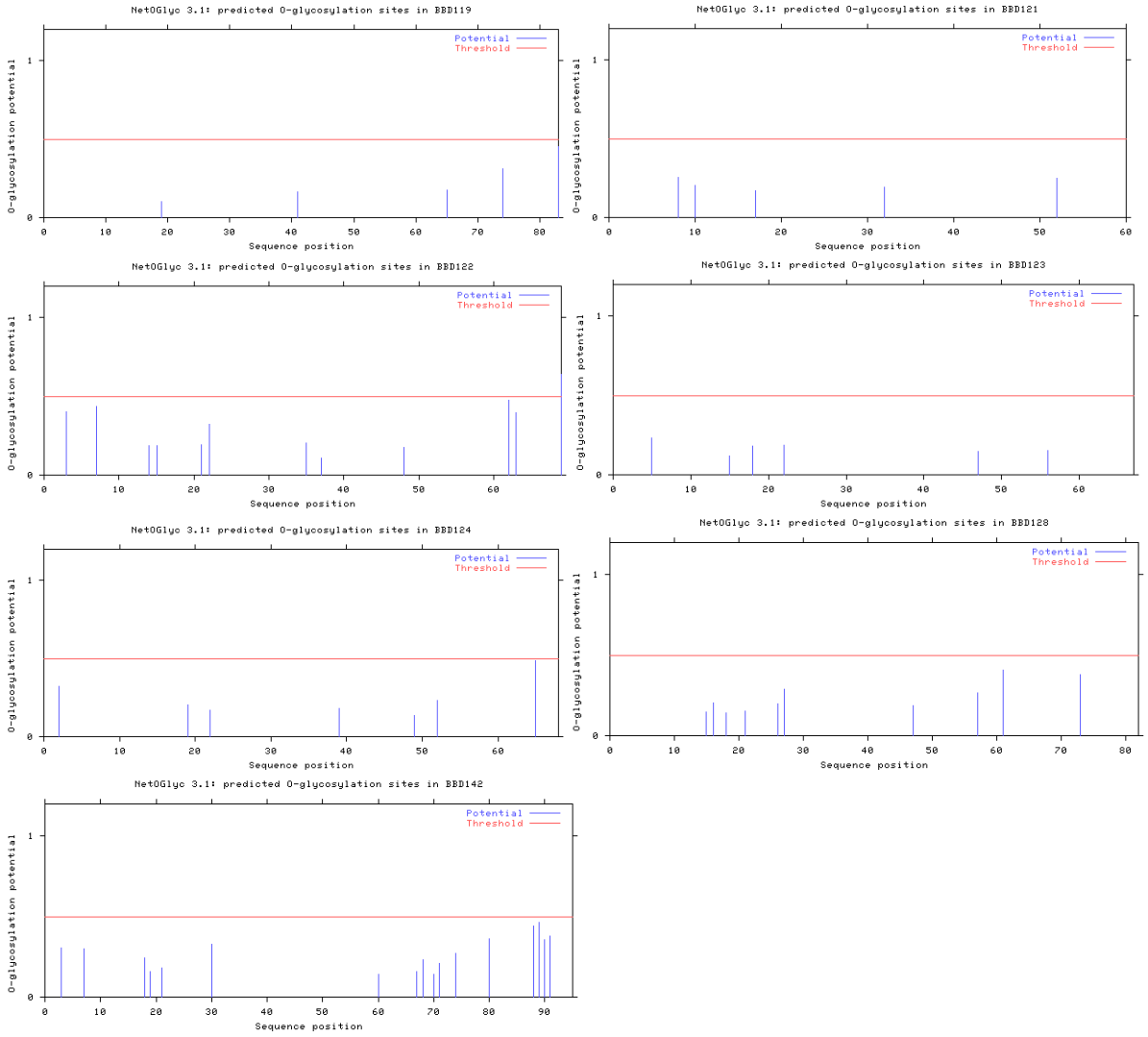


Figure 3.3-13: O-linked glycosylation analysis in β-defensin genes with no predicted glycosylation sites.

NetOGlyc 3.1 graphs of predicted glycosylation sites in β-defensin genes. X-axis denotes the sequence position of the glycosylation site in the coding region of the β-defensin gene. Y-axis is the O-linked glycosylation potential, and red threshold line denotes the potential of glycosylation.

3.4 Discussion

This is the first comprehensive analysis of sequence variation present in bovine β-defensin genes, which play critical roles in immunity and fertility. Fertility phenotypes identified from ICBF fertility data for all 168 sires in this study show that pregnancy rate ranged from 20% to 70% with a mean of 49%. This is consistent with the phenotypic variation seen in bull fertility data (Al Naib et al., 2011). It is interesting to note here the presence of AI bulls highly-selected for fertility traits, with PR as low as 20%. The AAM fertility phenotype ranged from -0.21 to 0.12 units with a mean of 0.017 units. In addition, even taking the low numbers of low PR bulls into account, there is still large variation of ~20% between bulls of 40% PR compared to 60%.

Sires of high- and low-fertility were selected based on two indirect male fertility phenotypes: 1 standard deviation above and below the mean for pregnancy rate and adjusted animal model, determined the sires selected for sequencing, as described in detail in methods. Importantly, the fertility phenotype data obtained from ICBF is for artificial insemination bulls, which are used in active service. Sires in active AI, have undergone stringent semen quality evaluation testing prior to use on AI and therefore, even low-fertility sires have already passed semen quality control measures. Low-fertility AI sires may not be synonymous with low-fertility stock bulls. Future studies by our research group will examine stock bull fertility which may analyse low-fertility bulls in the stock bull population.

Bovine fertility has major cost implications for farmers, especially in Ireland. Approximately 20% of calves born in Ireland are sired by AI bulls. The process of AI bypasses the cervix, which may reduce the natural selection pressures on the ability of sperm to penetrate cervical mucus. Taking this into account, the use of AI sires which have passed semen QC is interesting in this study, as the semen appears morphologically normal, but still results in low-fertility sires. This may be explained by reduced ability of sperm to swim through cervical mucus as shown in macaque (Tollner et al., 2008), caused by mutations in β-defensin genes containing glycosylation sites, such as *DEFB126*. Therefore, mutations in genes containing glycosylation sites may affect the ability of sperm to penetrate cervical mucus in the female reproductive tract. Altered glycosylation has previously been shown to affect sperm penetration in humans (Tollner et al., 2011).

The objective of this study was to investigate whether variants in the β-defensin gene cluster are associated with male fertility in cattle. After targeted sequencing of bulls (n=168) divergent for fertility, 3,134 variants were identified after variant filtering (2,892 SNPs, 105 insertions and 137 deletions). These variants were then associated with the AAM phenotype using GenABEL: An R package for Genome Wide Association Analysis.

A group of 97 SNPs located on chromosome 13 are associated with the adjusted animal model fertility phenotype. Evidence suggests these SNPs are inherited as a haplotype. Nine sires (5 Holstein Friesian, 1 each of Limousin, Simmental, Charolais and Belgian Blue) are heterozygous, whilst all other sires have reference alleles. The 9 heterozygous sires are of medium to high fertility, according to the adjusted animal model phenotype (overall AAM = 0 to 0.07, mean 0.04, s.d. 0.027).

Of the SNPs in the haplotype, 76 were annotated by SnpEff to be located within introns, 5 downstream and 13 upstream of genes and 18 are intergenic. This total is more than 97, as multiple annotations can be called for the same SNP as they may be located upstream or downstream of gene 1, and be intergenic between gene 1 and 2, for example. The 97 haplotype SNPs span ~130kbp and contain 8 β-defensin genes (*BBD128, BBD127, BBD126, BBD125, BBD125a, BBD115, BBD142* and *BBD116*).

Targeted re-sequencing of bovine β-defensin genes was performed in the same 24 animals which have also undergone whole-exome sequencing due to a lack of annotation for β-defensin genes in the bovine genome for whole-exome target probe design. Complete gene sequences for targeted re-sequencing of β-defensins plus 1,000bp upstream of the genes cover 387kb, however, by limiting baits which align to between 1 and 5 locations in the genome reduce the target area to 235kb.

This larger project has sequenced 69 genes of interest in 168 AI bulls. Over 2,000 SNPs have been identified, with a block of 59 SNPs on chromosome 13 being among the most associated with the fertility phenotype. All 168 (including these 24) samples are included in the association analysis, and will lead our future direction of research.

O-linked glycosylation analysis identified predicted O-linked glycosylation sites in the tail region of *BBD126*, see Figure 3.3-12, the bovine ortholog of *DEFB126*, which in humans has

been shown to contain a dinucleotide polymorphism in exon 2, which resulted in sub-fertile males, who were 40% less likely to conceive. Importantly, men with this polymorphism in DFEB126 had an 84% reduction in the ability of sperm to penetrate an artificial cervical mucus substitute, see Figure 1.3-4. Also, mouse β-defensin 122, orthologous to bovine *BBD126*, is characterized as one of the most abundant proteins of the mouse sperm glycocalyx (Yudin et al., 2008).

Interestingly, previous work by our group identified 52 sites predicted across 13 genes in this 19 gene β-defensin cluster. In comparison, re-analysis with updated annotation information identified 46 predicted glycosylation sites across 12 genes, with seven containing no predicted sites. *BBD126* as mentioned above contained predicted glycosylation sites. However, it also contained the second highest number of sites with 8 in total, compared to *BBD129* which had the most with 9 predicted sites and *BBD125a* which also had 8 sites.

Taken together, these data support the theory that β-defensin genes play a role in male fertility. By bypassing the cervix and cervical fluid, a natural source of semen selection, AI bulls with decreased ability to penetrate cervical mucus are possibly contributing to decreased fertility of the Irish national herd.

In this chapter, our aim was to investigate whether variants in the β-defensin gene cluster are associated with male fertility in cattle and identify candidate SNPs for validation in an independent population. Targeted sequencing of the β-defensin gene cluster for the first time in bulls divergent for fertility identified significantly associated SNPs located on chromosome 13, which are inherited as a haplotype. This haplotype appears to be driven by 9 medium-high fertility bulls, and subsequent functional validation identified that sperm from high-fertility bulls with the haplotype were significantly better at binding to oviductal epithelial cells, a key step in fertilisation where sperm aggregate in the oviduct prior to capacitation (Finlay, E. et al. 2017, unpublished data).

4 Whole-exome sequencing of bulls divergent for fertility

4.1 Introduction

Male fertility is a complex, polygenic trait with multiple phenotypes. Previous studies have identified genetic variants with effects on male fertility (Akinloye et al., 2009, Krausz et al., 2015, Akinloye et al., 2007). Coding region variants have been shown to affect fertility in cattle (Sonstegard et al., 2013). Interestingly, exome sequence analysis and targeted SNP genotyping of recessive fertility defects has identified the causative variant in cattle (McClure et al., 2014a). These data demonstrate the importance of variant discovery, particularly in coding regions, in identifying causal variants in bovine.

Whole-exome sequencing is where every known exon in the genome is sequenced, to identify functional, coding region variants that may have a significant impact on gene function, protein coding, and ultimately, phenotype. Previous studies have indicated that genome-wide variants have an association with and impact on fertility (Penagaricano et al., 2012, Cochran et al., 2013a, Feugang et al., 2009). In addition to genetic variants identified in Chapter 3, variants located in coding regions genome-wide were also identified, as they may also be associated with male fertility. As male fertility is a polygenic trait, multiple genes and biological processes will influence the phenotype.

Application of a genomic-scale sequence-based approach to association studies has previously been proposed as a method for identifying genes underlying complex phenotypes (Botstein and Risch, 2003). To identify coding region variants which may underlie complex phenotypes, whole-exome sequencing was performed on Irish AI bulls divergent for fertility.

Whole-exome sequencing was performed as part of a dual-sequencing approach, as discussed in the previous chapter and is also part of a larger research project, investigating potential molecular biomarkers for fertility in cattle. Therefore, it was important to characterise the effect of exonic variants in the genome, on the male fertility phenotype obtained.

Our hypothesis proposes that genome-wide genetic variation of the exons and promoter regions will explain some of the high phenotypic variation in AI bulls divergent for fertility.

4.2 Aims

The main aim of whole-exome sequencing was to identify variants in coding regions, and 5' regulatory regions of all annotated genes in the genome, for AI bulls which are divergent for a fertility phenotype. The secondary aim was to perform genome-wide association analysis on identified variants and fertility phenotypes collected, to identify significantly associated variants for further analysis. The final aim was to validate variants of interest by genotyping a large sample size using a custom-designed SNP chip, containing variants of interest.

4.3 Results

4.3.1 Exome sequencing coverage

Whole exome sequencing of 24 bulls was performed and a summary of minimum, maximum, mean sequencing reads, and reads aligned to the reference genome and reads remaining following duplicate removal are shown in Table 4.3-1. Coverage of on-target regions (regions of the exome targeted by our sequence capture probes) was found to be 98.6%. This shows that 98.6% of targeted regions were covered by at least 1 read. The remaining 1.4% was not captured by this bait design for unknown reasons.

An average of 39 million reads per sample were retained following read filtering, ranging from 23,716,853 to 78,048,261. The proportion of reads uniquely aligned to the genome averaged 94%, with 3.45% PCR duplication of reads. Percentage genome coverage at low read coverage was 95.5% at 2X coverage, and 69% at 10X coverage, ranging from 95% to 98% and 49% to 88% respectively. However, at higher coverage, the percentage of the genome sequenced was low with a mean of 33% at 20X and 12% at 30X coverage, ranging from 11% to 69% and 2% to 50%, respectively. At higher depths of coverage, the boar sequencing capture design has larger genome coverage i.e. 75% at 30X, see Table 4.3-3. This is probably due to the targeted genome capture methods used with the bull method focusing on identifying as many variants in as many animals as possible.

Table 4.3-1: Summary of coverage statistics for whole-exome sequencing.

Table of whole-exome sequencing coverage statistics for all 24 AI bulls of divergent fertility. Total Reads – Total number of sequencing reads per sample; % Aligned – Percentage of sequencing reads aligned to UMD3.1 bovine genome; % On Target – Percentage of sequenced reads per sample which map to a bait in the exome bait capture design used to manufacture baits; % Duplication – Percentage of reads which are identical to each other due to PCR duplication during library preparation; % 'n' X – Percentage of the targeted region covered by aligned sequences more than or equal to 'n' amount of times.

Bull AI Code	Total Reads	% Aligned	% On target	% Duplication	%2X	%10X	%20X	%30X
Holstein-Friesian High Fertility								
Bull 1	78048261	94.4	51.4	3.9	97.9	88.1	69.4	50.4
Bull 2	35550206	92.9	55.1	3.1	94.7	66.0	31.3	9.3
Bull 3	53731626	94.3	53.3	3.6	97.1	81.4	54.9	30.3
Bull 4	43239255	93.8	53.8	3.0	96.7	77.2	42.2	15.0
Bull 5	23716853	94.3	55.9	3.0	92.4	49.3	11.2	1.6
Bull 6	41483730	94.0	54.4	3.0	96.3	74.0	40.5	15.1
Holstein-Friesian Low Fertility								
Bull 7	36979092	93.6	55.6	2.9	96.3	72.6	33.8	9.4
Bull 8	64031638	94.6	50.9	3.4	97.0	82.4	59.5	38.5
Bull 9	27839780	94.1	56.0	2.9	94.6	58.7	16.9	2.8
Bull 10	34814185	94.3	55.0	2.9	95.1	66.5	30.5	8.7
Bull 11	45393737	94.2	51.4	6.4	95.4	71.4	42.8	19.3
Bull 12	29749347	93.6	54.3	2.7	93.5	58.0	21.1	4.5
Limousin High Fertility								
Bull 13	36575977	94.3	51.2	6.3	95.7	67.5	28.4	6.7
Bull 14	30797458	93.1	55.0	3.0	95.3	64.6	22.0	4.3
Bull 15	37479255	93.6	55.6	3.1	95.9	72.7	34.0	9.5
Limousin Low Fertility								
Bull 16	35911606	93.5	55.6	2.8	96.2	71.5	30.9	7.7
Bull 17	36495314	94.2	54.9	3.5	95.8	69.6	33.2	9.9

Table 4.3-2 continued

Bull 18	43534826	93.6	52.0	3.9	96.4	76.1	40.9	14.0
Belgian Blue High Fertility								
Bull 19	33132619	94.7	53.4	2.6	93.9	60.9	26.9	7.6
Bull 20	40152528	93.4	53.5	3.4	96.6	75.2	37.3	11.1
Bull 21	33592242	93.4	51.7	3.3	95.6	66.0	22.4	4.2
Belgian Blue Low Fertility								
Bull 22	31737661	93.9	53.7	4.0	95.5	65.3	21.5	3.9
Bull 23	30025158	93.6	54.9	2.8	93.2	57.7	22.9	5.6
Bull 24	31034034	94.3	53.7	2.9	92.8	57.2	23.9	6.2
Mean	38960266	93.9	53.8	3.4	95.4	68.7	33.3	12.3
Min	23716853	92.9	50.9	2.6	92.4	49.3	11.2	1.64
Max	78048261	94.7	56.0	6.4	97.9	88.1	69.4	50.4

Table 4.3-3: Comparison of bull and boar exome sequencing coverage statistics.

Bull whole exome sequencing mean (24 samples) sequencing statistics and boar whole-exome sequencing mean (96 samples) statistics. Percentages refer to percentage of genome covered by reads at specified coverage (X). Boar exome sequencing data was obtained using the same Roche Nimblegen Developer kit, and was obtained from (Robert et al., 2014).

	Total Reads	Aligned	Bases on target	Duplication	2X	10X	20X	30X
Bull Mean	38,960,266	93.9 %	53.8 %	3.45 %	95.4%	68.7%	33.3%	12.3%
Boar Mean	38,415,939	89.62 %	60.74 %	3.78 %	95%	91%	-	75%

4.3.2 Variant discovery

Following variant calling, 3,437,419 unfiltered variants were discovered in all 24 bulls using GATK's HaplotypeCaller walker. Filtering for variants located within targeted regions, and with a minimum read coverage threshold of 5X, reduced variants to 284,042. Of these, 12,124 were insertions and 13,048 were deletions, leaving 264,038 SNPs. There were 91,047 (31.5%) novel variants, with no known dbSNP identifier. Further quality control filtering to remove variants with low call rate, low minor allele frequency, or that were out of Hardy-Weinberg equilibrium, reduced SNPs brought forward for analysis to 144,178, following check.marker quality control filtering. This additional filtering is required for variant association analysis to remove as many false-positive variants as possible. However, false-positive filtering also results in true-positive removal, which may not be necessary for SNP frequency estimation, variant annotation, or GO analysis.

4.3.3 Variant annotation

SnpEff, the variant predictor package, was used to predict the location, class and effect of SNPs. Variants were predicted to have high, moderate, low or modifier functional effect. The effect of the SNP on gene sequence and function was grouped as high – low, based on position and codon change. High effect variants accounted for 1.6%, moderate impact variants accounted for 17.3%, low impact variants were 24.4% and modifier variants were 56.6%. Examples of high effect variants include exon deletion, frame shift, start lost, stop lost and stop gained etc. Moderate effect variants include non-synonymous variants, coding sequence variants, and inframe insertions. Low effect variants include 5'UTR premature start gain, start retained and stop retained mutations. Modifier effect variants include 5'UTR variants, intron variants, and downstream gene variants. Variant impact and numbers are summarised in Table 4.3-4, on page 111. Of the 23 high effect variants, (the variants with the biggest predicted effect on gene function) 15 were predicted to result in frameshift mutations, 5 were predicted to result in stop-gained codon modifications, and 3 predicted to be splice acceptor variants.

Table 4.3-4: SnpEff variant effects annotation predictions for whole-exome sequencing
SnpEff, the variant predictor package, was used to predict the location, class and effect of SNPs. Predicted SNP effects with 'high' having the largest effect on gene function, followed by 'Moderate', 'Low, and then 'Modifier'. The effect of the SNP on gene sequence and function was grouped as high – low, based on position and codon change. Total variants sum to more than 144,178, due to multiple annotations for single variants.

High	Low	Moderate	Modifier	Total
2656	40510	28779	93612	165557

Variants were located in 37.85% of exons, 21.41% of introns and 2.06% of 5'UTR. The remaining ~40% are located upstream or downstream of genes, 3'UTR and between variants (intergenic). Variant locations are shown in Figure 4.3-1.

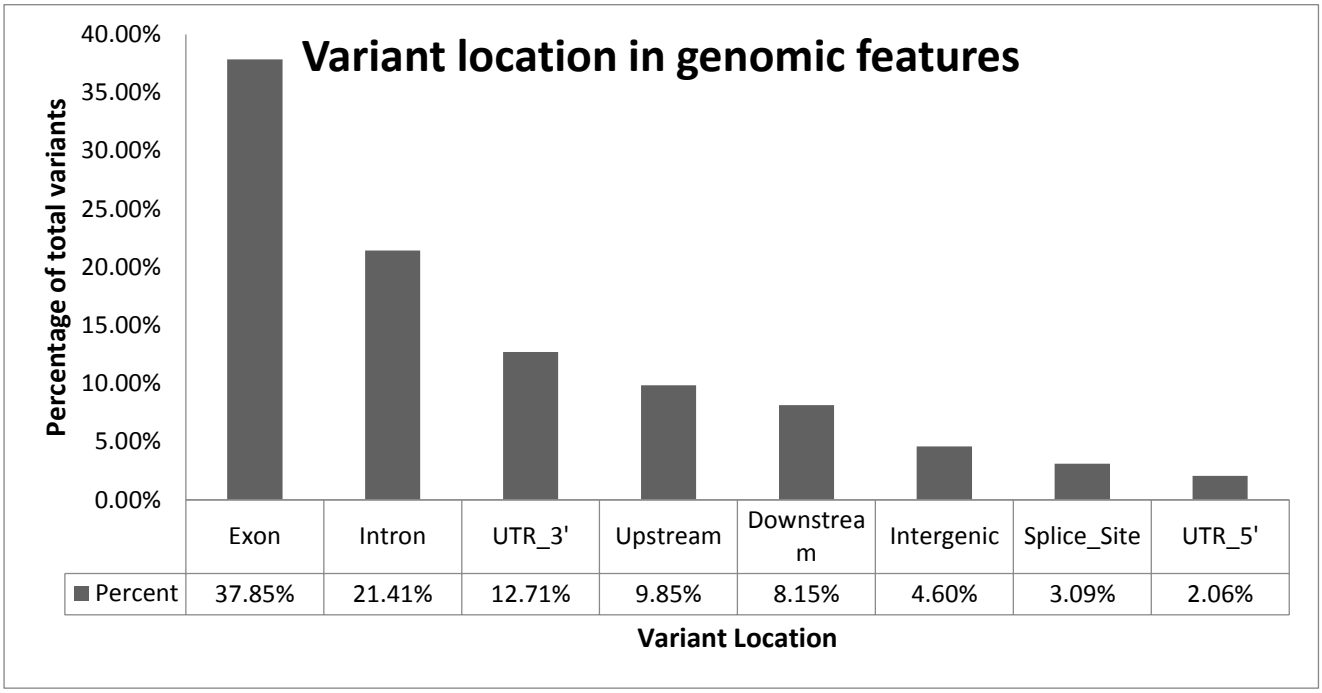


Figure 4.3-1: Chromosomal locations for whole-exome variants
Bar chart of variant locations in genes for whole-exome sequencing dataset. The percentage of variants in each region (y-axis) and genomic region of variant (x-axis) are shown.

Of the 38% of variants in exons, 53% were synonymous, 46% were non-synonymous, and ~1% were predicted to result in frameshift mutations. Transition to transversion ratio (Ts/Tv) found in the whole–exome sequencing project is 2.84 (only SNPs are used to calculate this statistic). A missense (28,380) to silent (35,455) ratio of variants is 0.8005.

4.3.4 Exome variant SNP frequency analysis

Following variant annotation and filtering SNPs based on depth of coverage, a custom-made Perl script was used to count SNP alleles, determine SNP frequencies, and proportions of alleles being in a category of interest (fertility). SNPs with greater than 25% difference in SNP frequency between fertility groups and which have a genotyping call rate > 80% to remove missing data, are shown in Figure 4.3-2.

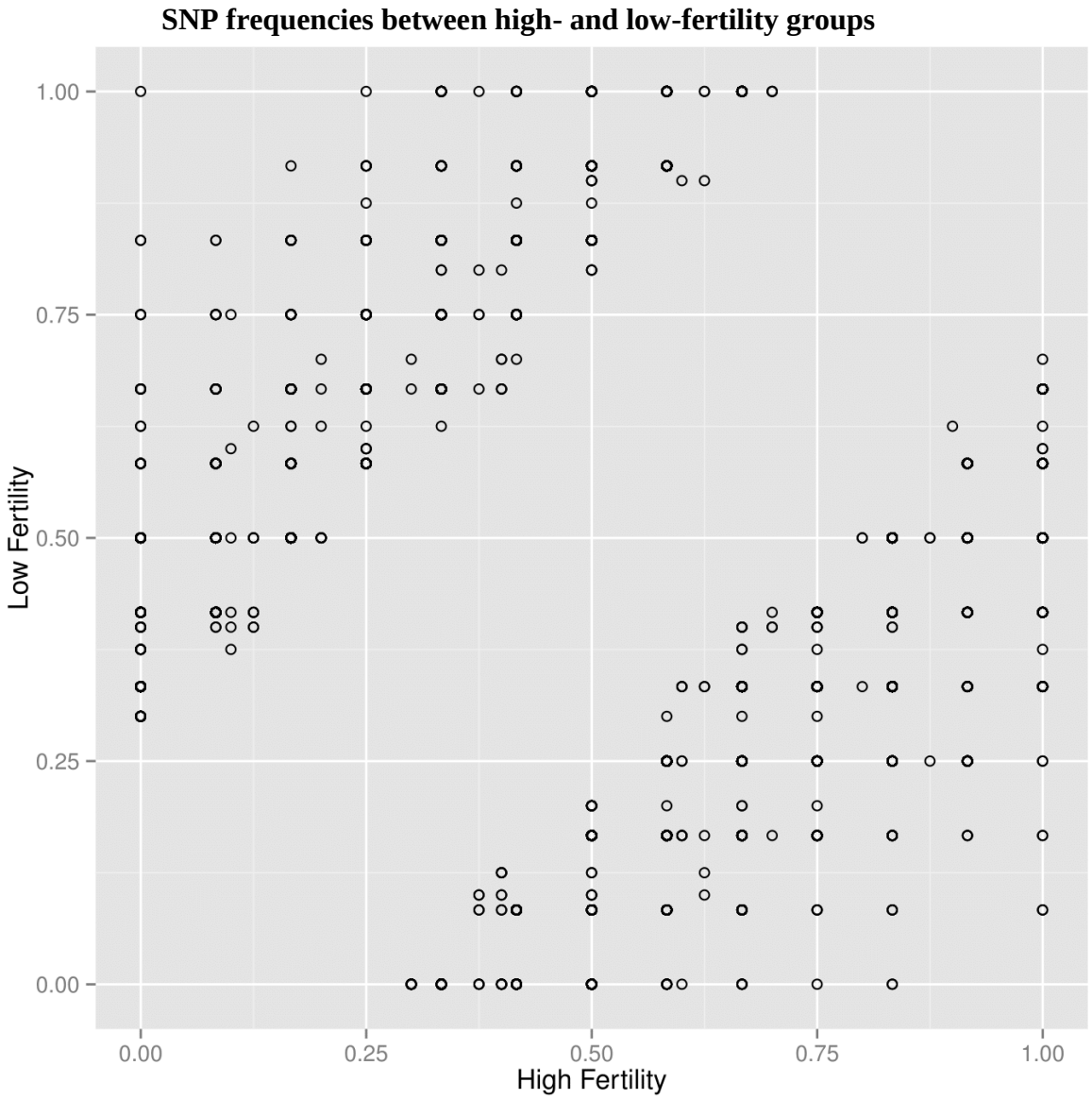


Figure 4.3-2: SNP frequency scatter plot between high- and low-fertility groups
Each point denotes a filtered exome sequencing variant (3,430), with a SNP frequency difference between high- and low-fertility groups of greater than 25%. The x-axis denotes the frequency of the alternate allele for 12 samples in the high-fertility group, and the y-axis denotes frequency of the alternate allele for 12 samples in the low-fertility group.

Table 4.3-5: Top 20 variants with SNP frequency differences between groups of divergent fertility

A Table of the top 20 variants with SNP frequency differences between high-and low-fertility groups. Chrom = Bovine chromosome number; ID = dbSNP identifier; High Fertility Frequency = SNP frequency in bulls of high-fertility; Low Fertility Frequency = SNP frequency in bulls of Low-fertility; Nearest Gene = Nearest annotated gene to SNP according to dbSNP database.

No.	Chrom	Position	ID	High Fertility Frequency	Low Fertility Frequency	Nearest Gene
1	X	3476025	rs42198964	0.791	0.083	*PGRMC1*
2	17	71933266	rs41853791	0.833	0.166	*MORC2*
3	X	129382246	rs110385224	0.708	0.041	*PGM3*
4	X	3481991	rs42198966	0.166	0.833	*LOC536148*
5	15	30414498	rs135635704	0.166	0.75	*CBL*
6	24	30374980	rs380970509	0.166	0.75	*Csmd1*
7	10	81033222	rs43644083	0.125	0.708	*ACTN1*
8	25	2122895	rs137616491	0.166	0.75	*KCTD5*
9	7	9178070	rs109973752	0.25	0.833	*CCDC105*
10	X	58420106	rs209828298	0.041	0.625	*LOC523458*
11	18	60633324	rs380059125	0.75	0.166	*LOC100337403; LOC101906794; LOC101908336; LOC785408*
12	18	55955510	rs41891074	0.136	0.708	*NUCB1*
13	10	20804716	rs137800911	0.416	0.958	*IPO4*
14	16	2668990	rs110598472	0.25	0.791	*CNTN2*
15	17	72453195	rs135340962	0.083	0.625	*PISD*
16	1	151166317	rs380830493	0.208	0.75	*DSCR3*
17	7	9170968	rs110367026	0.333	0.875	*CCDC105*
18	X	15389344	rs110425744	0.416	0.958	*LOC528106*
19	18	60633305	rs383220134	0.833	0.291	*LOC100337403; LOC101906794; LOC101908336; LOC785408*
20	18	60633347	rs385598203	0.75	0.208	*LOC100337403; LOC101906794; LOC101908336; LOC785408*

4.3.5 Breed-specific SNP frequencies

Three breeds of bull were sequenced via whole-exome sequencing, HF, LM, and BB. SNP frequencies for HF bulls were analysed, however, LM and BB breed-specific SNP frequencies are inaccurate due to low sample numbers (n=6). BB had 21,062 SNPs and a SNP frequency difference between fertility groups of greater than 25%, whilst LM had 21,925.

For HF (n=12), 15,984 SNPs had a SNP frequency difference between fertility groups of greater than 25%. Three different breeds were sequenced to ensure robust variant calls for validation, and to reduce single breed specific SNPs being reported.

4.3.6 Gene ontology

Gene ontology analysis was used to identify significantly over-represented gene ontology terms from the 484 SNPs associated with AAM fertility. Initially, the SNP dataset was annotated to identify the nearest gene for each SNP. Two functional categories were significantly over-represented in this dataset after Benjamini-Hochberg correction for multiple testing. The term 'glycoprotein' was a significantly over-represented keyword in this dataset (adjusted *P*-value = 0.0056). The term 'glycosylation site: N-linked' is a significantly over-represented feature in this dataset (adjusted *P*-value = 0.00024). For GO term over-representation analysis, approximately 200 genes were not annotated in the DAVID databases and were not accounted for.

In addition, other terms were over-represented in the dataset, with significant uncorrected P-values, however, after Benjamini-Hochberg correction; the P-values were not significant. These over-represented terms include 'immune response' (unadj. *P*-value = 0.0018), 'disulfide bond' (unadj. *P*-value = 0.003).

Table 4.3-6: Gene ontology analysis of SNPs divergent between fertility groups.

Variant association analysis of 484 variants identified by whole-exome sequencing associated with Adjusted Animal Model (AAM) fertility phenotype. Variants are ranked based on their *P*-values. The unadjusted *P*-value significance threshold was set at ($p < 0.1$). Category, denotes original database/resource where the terms orient (i.e. KEGG pathways, Protein information resource keyword, or UniProt feature), Term, denotes enriched terms associated with the gene list. Count, refers to the number of genes involved in the term. % refers to involved genes / total genes. *P*-value is a moderated Fisher exact P-value, EASE score. BH, refers to Benjamini-Hochberg multiple testing correction.

Category	Term	Count	%	P-Value	BH
UP_SEQ_FEATURE	Glycosylation site: N-linked (GlcNAc...)	28	10.5	8.3E-7	2.3E-4
SP_PIR_KEYWORDS	Glycoprotein	29	10.9	3.1E-5	5.6E-3
SP_PIR_KEYWORDS	Immune response	6	2.3	1.8E-3	1.5E-1
SP_PIR_KEYWORDS	Disulfide bond	20	7.5	3.3E-3	1.8E-1
GOTERM_BP_FAT	Acute inflammatory response	5	1.9	2.1E-3	3.2E-1
UP_SEQ_FEATURE	Disulfide bond	18	6.8	2.8E-3	3.3E-1
SP_PIR_KEYWORDS	Complement pathway	3	1.1	1.2E-2	3.6E-1
GOTERM_BP_FAT	Activation of plasma proteins involved in acute inflammatory response	4	1.5	1.7E-3	3.7E-1
GOTERM_BP_FAT	Complement activation	4	1.5	1.7E-3	3.7E-1
GOTERM_BP_FAT	Humoral immune response	4	1.5	4.2E-3	3.7E-1
SP_PIR_KEYWORDS	Innate immunity	4	1.5	1.0E-2	3.7E-1
KEGG_PATHWAY	Complement and coagulation cascades	5	1.9	1.4E-2	7.8E-1

4.3.7 Quality control

Stringent quality control filtering was applied to whole-exome variants identified, to ensure robust association analysis. Initially, 284,042 filtered variants were identified from GATK variant calling prior to quality control. Filtering based on a SNP call rate of 80% (call rate = 0.8), an individual SNP call rate of 90% (maximum percentage of missing genotypes in an individual sample; perid.call =0.9) and a minor allele frequency cut-off threshold of 5% (MAF = 0.05) reduced the number of variants to 144,178. In addition, two sires were removed from association analysis, due to high autosomal heterozygosity. Principal component analysis of genomic kinship both pre-QC, see Figure 4.3-4 and post-QC, see Figure 4.3-5 shows the principal components leading to variation between breeds, in bulls of high- and low-fertility. Sires removed from association analysis had autosomal heterozygosity over 40% for SNPs identified in those sires.

The genomic inflation factor (lambda) is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median, thus quantifying the extent of the inflation and the false positive rate (Yang et al., 2011, Aulchenko et al., 2007). The genomic inflation factor for the whole-exome sequencing dataset was calculated as 0.8, less than 1, suggesting either population stratification or polygenic inheritance (Yang et al., 2011), see Figure 4.3-3. Given that population stratification has been accounted for via selection of unrelated bulls, using a pedigree matrix, this was deemed to be acceptable for further analysis. For association analysis lambda was set to 1.

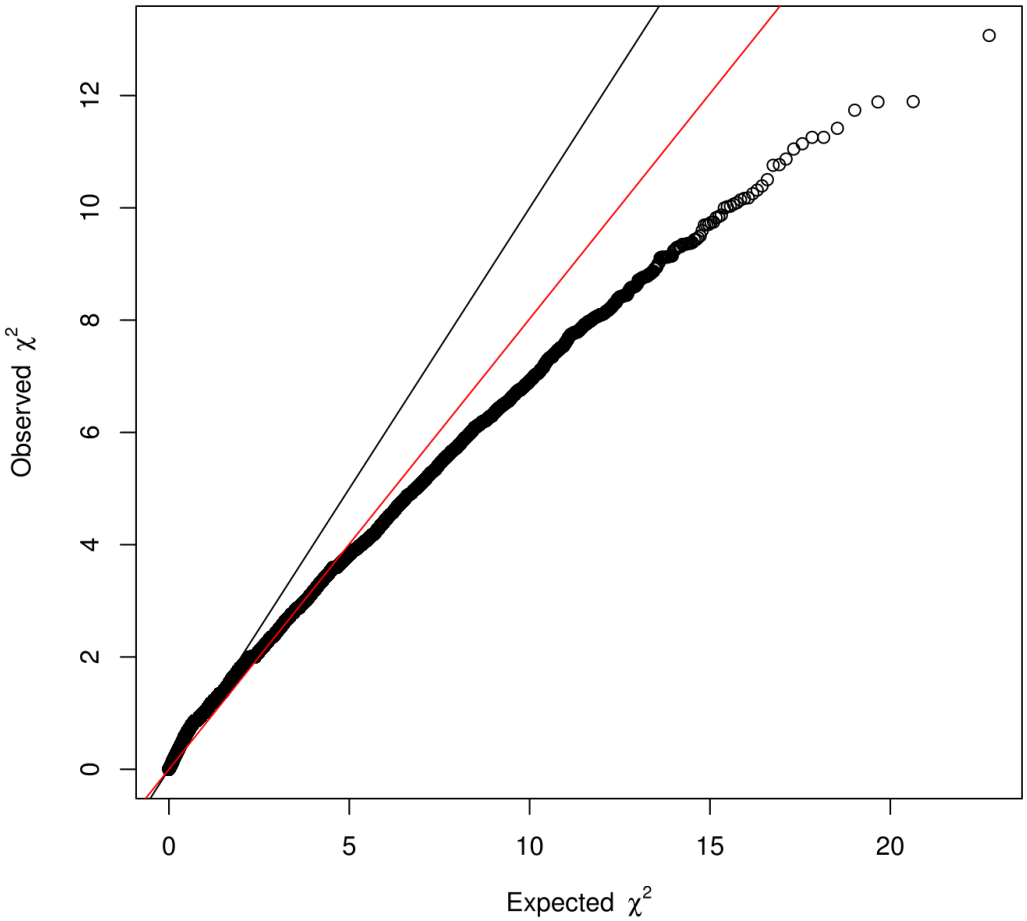


Figure 4.3-3: χ2−χ2 plot for a GWA scan.

The λ is computed by regression in a Q-Q plot. The estimate of λ is less than 1, suggesting deflation of the test and some degree of stratification. X-axis denotes the Expected χ2 and the Y-axis denotes the observed χ2 for the GWA scan. Black line of slope 1: expected under no inflation; Red line: fitted slope.

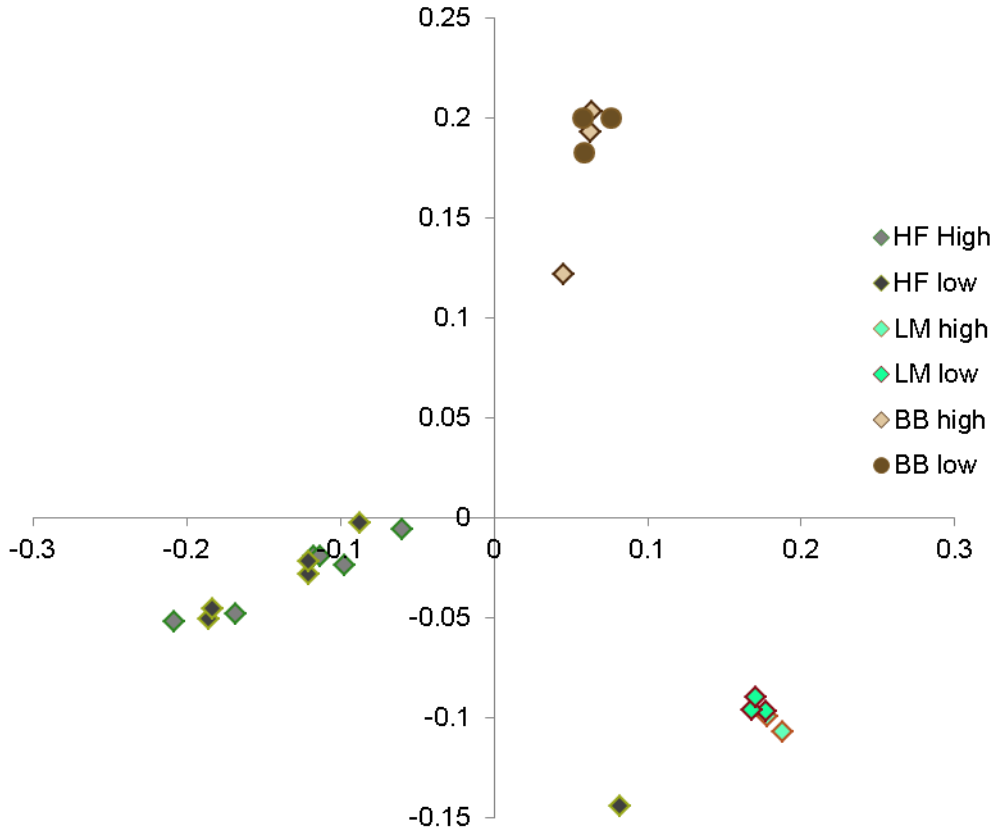


Figure 4.3-4: Principle components resulting from analysis of genomic kinship - identifying genetic outliers

X-axis denotes the 1st principal component. The y-axis denotes the second principal component. Sires are grouped according to breed. HF = Holstein-Friesian, LM = Limousin, BB = Belgian Blue.

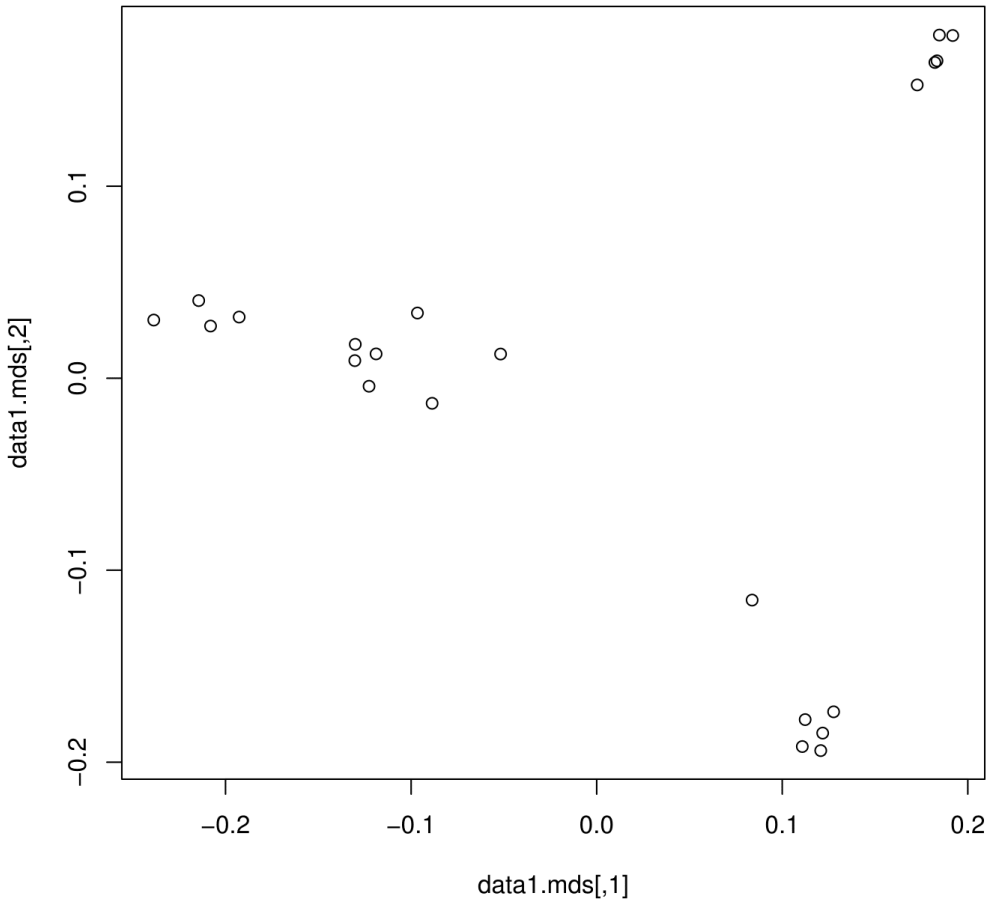


Figure 4.3-5: Principle components resulting from analysis of genomic kinship post quality control
After dropping the outliers, check.marker QC was repeated to identify any further genetic outliers. None were identified.
X- axis denotes the 1st principal component. Y-axis denotes the 2nd principal component.

4.3.8 SNP association

In total, 144,178 variants were identified in whole-exome sequencing after filtering and quality control for SNP association analysis[20]. The genes most significantly associated with the fertility phenotype, adjusted animal model, are shown in Table 4.3-7. Of the top 20 most significantly associated SNPs, 11 SNPs were predicted to be in the intron region of a gene, 7 were predicted to be in exons (4 synonymous, 1 missense and 2 unknown), 2 were predicted to be non-coding, and 2 were upstream of their respective genes.

The most significantly associated gene is located on chromosome 11 at position 49,866,493 with an unadjusted *P*-value = 0.00014. It is predicted to be in the *U6* novel snRNA in humans, a pseudogene of the small non-coding RNA class. In addition, this gene is highly-expressed in the testes of bulls (Merkin et al., 2012). This data set was originally submitted to NCBI Gene Expression Omnibus under accession number GSE41637 and is available in ArrayExpress as E-GEOD-41637. Little is known about the function of this gene, and future studies would be required to verify its function, especially in relation to fertility.

A SNP in the signal regulatory protein alpha, SIRPA, gene located on chromosome 13 at position 53691410 is the 2nd most significantly associated variant. This variant is in the exonic region of SIRPA, and is associated with AAM, (*P*-value = 0.00044). The exact function of this variant is unknown, and it is not annotated in the dbSNP database. However, SIRPA is a promising candidate gene for spermatogenic impairment and male fertility (Krausz et al., 2015).

The 3rd most significantly associated SNP, rs109065788, is located on chromosome 3 in the exonic region of the forkhead box J3 gene (*FOXJ3*). This SNP results in a synonymous mutation, resulting in no change of the amino acid sequence. However, synonymous mutations can affect mRNA abundance, resulting in altered protein concentrations, and can alter substrate specificity via altered protein conformations, as shown in humans (Kimchi-Sarfaty et al., 2007). *FOXJ3* is a transcription factor and is required for the survival of spermatogonia and participates in spermatocyte meiosis (Ni et al., 2016), see section 5.4.

[20] Electronic Appendix 4.1 Results Files/1 - qcbreedcountadjGWAS.txt

Two SNPs, rs137356698 and rs210428031, have multiple annotations being both intronic and upstream in different genes. SNP rs137356698 is in the 5'UTR of the NOB1 gene and in the intron of WWP2, with both genes located on alternate strands. SNP rs210428031 is in the intron of NFAT5 and possibly the 5'UTR of an alternative transcript.

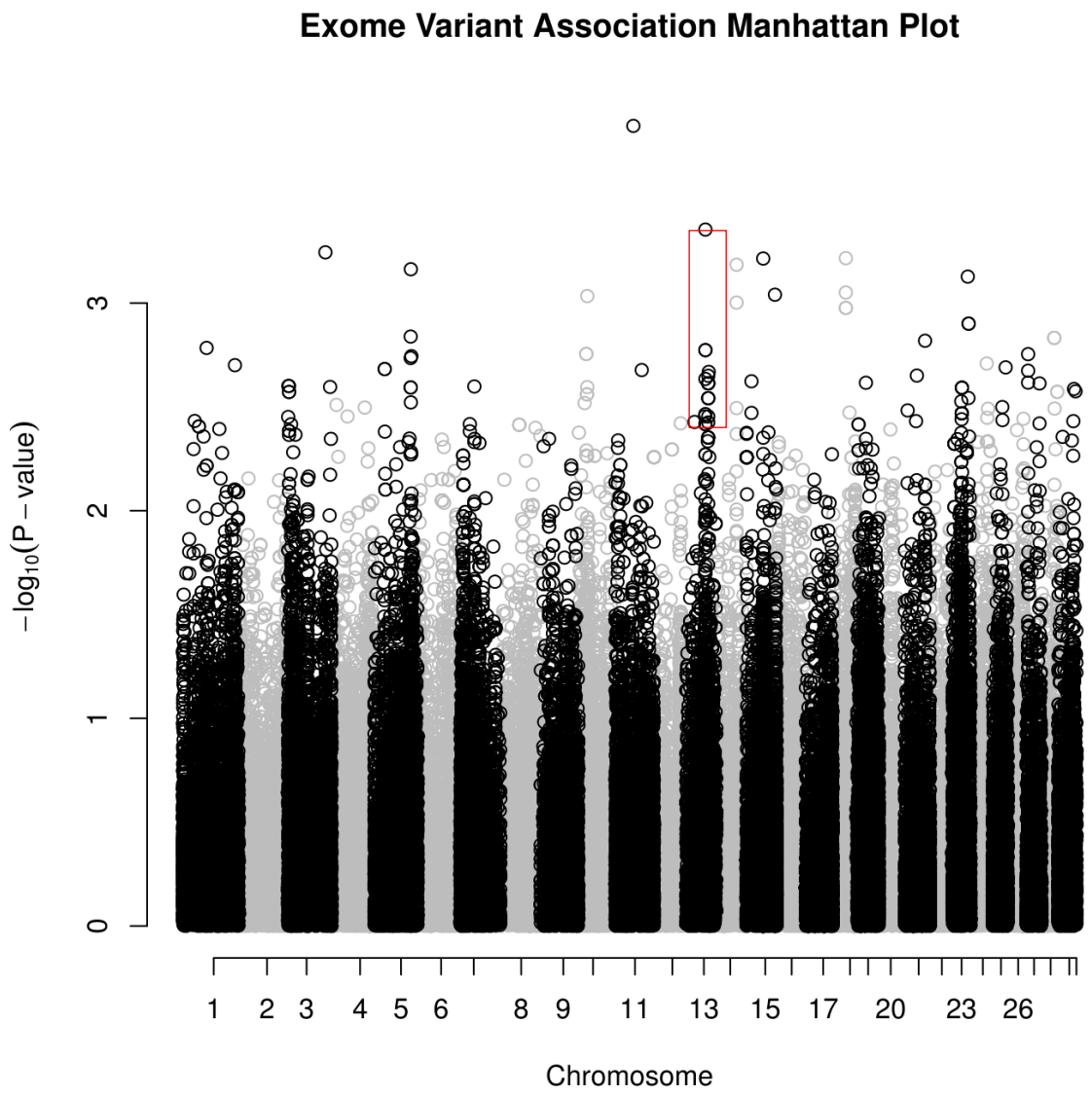


Figure 4.3-6: Manhattan plot of whole-exome sequencing variants associated with adjusted animal model fertility phenotype.

Manhattan plot shows the association of variants identified by whole-exome sequencing and their associated *P*-value with the adjusted animal model fertility phenotype. The *P*-value for each variant association is on the y-axis. The chromosomal position of each variant is on the x-axis. In total, there are 144,178 variants after quality control shown. Red box indicates cluster of associated SNPs with fertility located on chromosome 13, located near the β-defensin gene region (16 out of 38 SNPs in this region are within the β-defensin gene region).

Table 4.3-7: Whole-exome sequencing variant genes most associated with adjusted animal model fertility phenotype.
Variant association analysis of 144178 variants identified by whole-exome sequencing associated with Adjusted Animal Model (AAM) fertility phenotype. Variants are ranked based on their unadjusted *P*-values. The unadjusted *P*-value significance threshold was set at ($p < 0.01$). Variant ID gives the 'rs' number for any variant in dbSNP, otherwise 'Unknown' is used. Chr (Chromosome) is the chromosome number each variant is located on. Position is the location of the variant on its chromosome. *P*-value is the unadjusted *P*-value prior to multiple comparison testing, which is not shown due to Linkage Disequilibrium (LD) of variants. Gene shows the predicted gene each variant is in or near. Type refers to the type of variant and where it is in each gene. Function attempts to describe any known function related to the gene each variant is in or near.

No.	Variant ID	Chr.	Position	*P*-value	Gene	Type	Function
1	Unknown	11	49866483	0.00014	*U6* novel snRNA	Intron	Pseudogene affiliated with the snRNA class
2	Unknown	13	53691410	0.00044	*SIRPA*	Exon	Signal regulatory protein alpha
3	rs109065788	3	104587541	0.00057	*FOXJ3*	Synonymous	Transcription factor activity
4	rs137356698	18	36955817	0.00060	*NOB1 WWP2*	Intron variant, UTR variant 5 prime	Pre-rRNA processing
5	rs42670353	15	47385334	0.00061	*CCKBR*	Synonymous	G-protein coupled receptor for gastrin and cholecystokinin (CCK), regulatory peptides of the gastrointestinal tract
6	rs133213758	14	56982732	0.00065	*SYBU*	Intron	Anterograde axonal transport of active zone components
7	Unknown	5	102028024	0.00068	*C3AR1*	Intron	Complement Component 3a Receptor 1
8	rs42033755	23	45098361	0.00074	*ELOVL2*	Intron	Fatty acid elongase activity
9	rs41874538	18	36941544	0.00089	*NOB1*	Intron	Pre-rRNA processing
10	Unknown	15	80338451	0.00091	*LOC519317*	Exon	*Bos taurus* olfactory receptor 8J2
11	rs380166658	10	25168531	0.00092	*LOC100298264*	Non-coding	Unknown
12	rs109194993	14	56790188	0.00099	*LOC521950*	Non-coding	Unknown
13	rs210428031	18	36847662	0.00105	*NFAT5*	Intron variant, upstream variant 2KB	Inducible gene transcription during the immune response

Table 4.3-6 continued

14	rs378294791	18	36888695	0.00105	*NFAT5*	Intron	Inducible gene transcription during the immune response
15	rs110682224	23	47671466	0.00125	*DSP*	Missense	Anchors intermediate filaments to desmosomal plaques
16	rs42039246	23	47671486	0.00125	*DSP*	Synonymous	Forms an obligate component of functional desmosomes
17	Unknown	5	101799144	0.00145	*DPPA3*	Intron	Developmental Pluripotency Associated 3
18	Unknown	28	35602644	0.00147	*SFTPD*	Intron	Lung Surfactant Protein D
19	Unknown	28	35602648	0.00147	*SFTPD*	Intron	Lung Surfactant Protein D
20	rs110891650	21	58173287	0.00152	*CHGA*	Synonymous	Parathyroid Secretory Protein 1

In total, 484 variants had an association with the adjusted animal model fertility phenotype at an unadjusted significance level of $P < 0.01$. No SNP was significantly associated after correction for multiple testing via Bonferroni or Benjamini-Hochberg at $P < 0.05$ significance level.

As shown in Figure 4.3-6, clusters of SNPs associated with fertility are found on chromosomes 3, 13 and 23. Chromsomes 13 and 23 are known to contain β-defensin genes with 38 SNPs located on chromosome 13 associated with fertility (unadj. P-value < 0.01). Of these, 16 are located within positions 61,000,000 and 62,000,000, which includes the β-defensin gene locations on chromosome 13 and the haplotype region.

4.3.9 Transcription factor binding site analysis

MatInspector (Quandt et al., 1995) v3.7 was used to identify potential binding sites for transcription factors which are affected by mutations identified in this whole-exome sequencing dataset. The fourth most associated SNP with fertility in bulls was located in the 5'UTR region of the *NOB1* gene, located on chromosome 18 at position 36,955,817. MatInspector transcription factor (TF) binding site analysis identified this SNP is located in a spermatogenic zip 1 transcription factor (*SPZ1*) binding site, which is a testis-specific bHLH-Zip TF.

In addition, a *DEFB125* SNP located in the upstream region of rs385102822, contains a mutation in the TF homeobox B3, and NK6 homeobox TF.

4.3.10 Exome sequencing validation

To validate the success of exome sequencing capture design, genetic variants previously identified as being involved in male fertility, or listed in Table 1.2-2, were identified in the exome sequencing dataset. This method identifies whether the divergent phenotype sequencing approach and the capture design were successful in identifying genetic variants known to be involved in male fertility, and specifically bull fertility.

From Table 1.2-2, three SNPs, rs137601357, rs137182814, rs210398455 located in *CAST*, *STAT5a* and *CWC15*, respectively, were also identified in the whole-exome sequencing

dataset of bulls divergent for fertility. However, the 3 SNPs were not significantly associated with the adjusted animal model fertility phenotype (rs137601357 *P*-value = 0.36, rs137182814 *P*-value = 0.86 and rs210398455 *P*-value = 0.59).

In addition, 24 SNPs were identified in the whole exome sequencing dataset between position 61,316,000 on chromosome 13 and position 61,620,000 which contains all annotated β-defensin genes in the whole-exome capture design. These 24 SNPs were compared against the targeted β-defensin sequencing dataset, and 18 SNPs out of 24 were called in both sequencing approaches. Two of the 6 SNPs (rs134076463, and rs137275002) that were not in both datasets did not have probes specifically targeting their genomic location. The remaining four SNPs did have probes overlapping their genomic coordinates, and the reason they were not called in both approaches is unknown. There were no overlaps between the two cohorts of bulls.

Table 4.3-8: Probes targeting β-defensin genomic region in exome-sequencing dataset.
This table shows the probes targeting the β-defensin genomic region in the whole-exome sequencing dataset, indicating chromosome number (Chromosome), start position (Start) of exome sequencing capture probe and end position (End) of capture probe.

Chromosome	Start	End
chr13	61314567	61315016
chr13	61316680	61316838
chr13	61501425	61501651
chr13	61523658	61523917
chr13	61531963	61532148
chr13	61533321	61533544
chr13	61562052	61562204
chr13	61566038	61566196
chr13	61572837	61573073
chr13	61577345	61577555
chr13	61584290	61584540
chr13	61595538	61595672

Table 4.3-9: SNPs identified in both sequencing datasets with overlapping probes.
Table of 6 SNPs identified which were not found in both sequencing datasets, indicating which have overlapping probes. SNP ID: dbSNP number, chromosome: bovine chromosome number, position: genomic position, overlapping probes: Do exome capture probes designed target this region?

SNP ID	Chromosome	Position	Overlapping probes
rs43709489	13	61584406	Yes
rs134076463	13	61566201	No
rs43708157	13	61566114	Yes
rs43708158	13	61566112	Yes
rs137275002	13	61566020	No
rs41702271	13	61501580	Yes

4.4 Discussion

This is the first whole-exome sequencing of bulls with divergent fertility phenotypes. Previous studies have performed whole-exome sequencing to identify causative variants for underlying defective bovine embryo development contained within three haplotypes in Holstein, and brown Swiss breeds (McClure et al., 2014a). However, whole-exome sequencing of males with divergent fertility phenotypes has not been published to date.

Whole-exome sequencing targeted region coverage was 98.6% indicating only 1.4% of the targeted region was not sequenced by the probes designed. The remaining 1.4% may be due to repetitive regions, highly polymorphic regions, or due to inter-probe hybridisation. Percentage targeted region coverage at low read depth was 95.5% at 2X, and 69% at 10X, ranging from 95% to 98% and 49% to 88% respectively. The 69% of the genome sequenced at 10X coverage was deemed sufficient for SNP discovery, as shown previously (Yu and Sun, 2013). Further sequencing would result in increased duplicates, and although would be beneficial in terms of read depth, the cost-benefit ratio was assessed to be too high. For additional analyses, such as copy number variation analyses, a higher depth of coverage over the entire targeted region would be required, however, for SNP discovery, lower read depths can accurately call variants. GATK HaplotypeCaller has been shown to be the SNP caller of choice for low-coverage areas (Cheng et al., 2014).

The use of TruSeq Nano library preparation kits was chosen due to low gDNA availability from the Teagasc DNA database after DNA extraction. This was a concern due to the possibility of increased PCR duplicates due to the use of PCR cycles in library preparation. However, the low PCR duplicate rate of 3.4% on average, ranging from 2.6 - 6.4, meant that PCR duplicates were not likely to adversely affect GATK SNP calling, as the low numbers of duplicates were filtered out of the dataset.

Variant calling identified 284,042 filtered variants in 24 bulls prior to quality control for association analysis. To identify variants most likely to result in functional differences between fertility groups, SNP frequencies in divergent fertility categories were calculated. SNPs that were divergent between high and low-fertility samples (> 25%) are shown in Figure 4.3-2. This SNP allele frequency correlation identified ~5,000 variants which had significant SNP frequency differences between high- and low-fertility. However, further

filtering for variants which also had at least 80% SNP call rate for each group was also applied.

A SNP, rs42198964, located in *PGRMC1* gene, with a high SNP frequency difference between high- and low-fertility groups was identified as the most divergent SNP between fertility groups for SNP frequency, see Table 4.3-5. This SNP was added to the IDB SNP chip, due to high SNP frequency differentials, but was removed during quality control filtering prior to association analysis. *PGRMC1* encodes a putative membrane-associated progesterone steroid receptor. The homolog of this gene in humans functions in steroid signalling, p450 activation and drug metabolism. Steroid signalling is of particular importance as progesterone is part of the androgen production system which includes testosterone, the male sex hormone. Androgen/androgen receptor signalling has been shown to differentially regulate 23 β-defensin genes in the mouse epididymis (Hu et al., 2014). This *PGRMC1* SNP was added to IDB SNP chip version 3, together with 668 SNPs from whole-exome sequencing and 195 from targeted β-defensin sequencing.

SNP association analysis of all 144,178 variant genotypes post quality control filtering with fertility phenotypes identified 484 variants with association (unadjusted $P < 0.01$) to the fertility phenotype, see Figure 4.3-6. The top 20 variants most associated with fertility are shown in Table 4.3-7. *P*-values for association analysis are relatively high, due to the low number of individuals included in the exome sequencing project (n=22). Correction for multiple testing using Bonferroni correction results in no significantly associated variants ($P < 0.05$). However, they can be conservative if the independence assumption does not hold, and this is usually the case for densely typed single nucleotide polymorphisms in genetic association studies (Gao et al., 2010).

To identify the processes involved in male fertility, a gene ontology analysis of all the genes containing variants identified as being associated with AAM fertility phenotype was carried out. 'Complement and coagulation cascades' KEGG pathway was found to be the most over-represented biological pathway in this dataset, from the KEGG database (*P*-value = 1.4E-2). 'Glycosylation' (*P*-value = 2.3E-4), 'Glycoprotein' (*P*-value = 5.6E-3), 'Immune response' (unadj. *P*-value = 1.8E-3) and 'Disulphide bond' (unadj. *P*-value = 3.3E-3) were the most over-represented GO terms from the list of genes identified as containing variants via

whole-exome sequencing of bulls divergent for a fertility phenotype. Glycoproteins have been shown to be involved in male fertility in humans (Xin et al., 2016) and that variants in glycoprotein genes, particularly β-defensins, are important for male infertility and pregnancy rate following IVF (Lindgren et al., 2016). In addition, as shown in Table 4.3-6, 10 of the 12 gene ontology features are immune-related terms. Terms such as 'innate immunity', 'complement', 'immune response' and 'acute inflammatory response' indicate the importance of immune system related genes in male fertility, and also demonstrates the emphasis on innate immune gene variation in this study was warranted.

The transition to transversion ratio (Ts/Tv) is the ratio of transitions (mutations that result in pyrimidine to pyrimidine or purine to purine nucleotides) versus transversions (mutations that result in a change from pyrimidine to purine or vice versa). For exomes, the increased methylated cytosines in CpG dinucleotides in exons, leads to an increased ratio. Methylated cytosine can undergo deamination and transition to a thymine. Whole-exome sequencing projects in human are predicted to have Ts/Tv ratios of 2.8 (DePristo et al., 2011), ratios lower than 2.8 mean the dataset may contain false-positives, which can be filtered out (DePristo et al., 2011). However, the ratio of 2.84 for this whole-exome sequencing study is consistent with previous reported values for exome studies (DePristo et al., 2011).

To perform SNP association of candidate SNPs in a large national dataset, SNPs of interest were added to the International Dairy and Beef SNP chip version 3, which is a custom-designed SNP chip developed to genotype approximately 330,000 cows per year at a density of ~50k SNPs. The aim of the SNP-chip is to facilitate genomic selection predictions. By genotyping these SNPs of interest in a large population of cattle of mixed breeds, a robust association analysis of SNPs identified as being divergent for fertility can be performed. In total, 668 SNPs, identified as being divergent between fertility groups by at least 25% and 195 SNPs from targeted re-sequencing of β-defensins were added to the SNP chip.

One method of validating the success of exome sequencing capture design is identifying genes previously identified in other species as being involved in male fertility and comparing to variants identified via exome sequencing. Choline dehydrogenase has previously been shown, in mice and humans, to be associated with sperm cell function (Johnson et al., 2012). *Chdh* gene deletion in mice results in decreased male fertility due to diminished

sperm motility. In humans, a non-synonymous SNP in the *CHDH* gene and a coding region SNP in IL17BR are associated with altered sperm motility characteristics. Variants were identified in both *CHDH* and *IL17BR* in exome sequencing, as shown in the electronic appendix[21], a table of 144,000 SNPs identified in bulls divergent for fertility after QC. In this dataset 5 *CHDH* SNPs were identified, 4 in coding regions and 1 in the 5'UTR. One SNP, rs380632306, was predicted to be deleterious to the function of the protein via SIFT variant effect predictor tool. The other three coding SNPs were predicted to be tolerated missense variants. The 5'UTR variant was identified by MatInspector to be located in the RUSH transcription factor binding site region which is a SWI/SNF related nucleophosphoprotein with a RING finger DNA binding motif.

Another method to validate the capture designs is to compare the sequencing datasets at overlapping regions, and identify whether the same SNPs were called. In this study, 18 out of 24 SNPs in the critical β-defensin region were identified in both datasets. Probes targeting the β-defensin genomic region in the exome-sequencing dataset were identified, see Table 4.3-8. According to Table 4.3-9, 2 of these 6 SNPs did not have probes targeting them, indicating mis-binding of probes. The β-defensin region is a highly copy number variable region with high levels of repetitive sequences, which makes this region hard to predict (Hollox et al., 2003).

MatInspector transcription factor binding site analysis identified that a SNP in the 5'UTR of *NOB1* was located in the TF binding site region of *SPZ1*. This SNP is also associated with AAM fertility phenotype. *SPZ1* is a testis-specific transcription factor. *SPZ1* has been shown previously in mice models to be involved in male fertility. It has been predicted in mice that Spz1 and PP1cγ2 may be required for proper regulation of spermatogenesis and fertility in males (Hrabchak and Varmuza, 2004). In this study, Spz1 is shown to be a binding partner of the catalytic subunit of protein phosphatase-1 and male mice homozygous for a null mutation in the protein phosphatase-1cγ (*PP1cγ*) gene are infertile and display impairment in spermatogenesis.

In addition, a study in mice has shown early lethality of homozygous Pno KO lineage, a ribosomal partner of *NOB1*, caused by arrest of embryo development before compaction

[21] Electronic Appendix 4.1 Results Files/1 - qcbreedcountadjGWAS.txt

stage (Wang et al., 2012). These findings indicate that NOB1 may not be directly affecting male fertility, but is regulating spermatogenesis with binding partners.

In conclusion, the aims of this chapter were to identify variants in coding regions and regulatory regions of all annotated genes in the bovine genome, to associate these variants with fertility, to identify candidate SNPs for male fertility, and to validate the selection, using a SNP chip and validation by genotyping an independent population. These aims were largely achieved, as 144,000 variants were identified in bulls divergent for fertility after quality filtering and QC. Of these, 485 variants were identified in or near genes associated with the AAM fertility phenotype (unadjusted $P<0.01$). A subset of these variants were then validated in an independent population of AI bulls to determine if the association is robust across the population. This is discussed in the following chapter. Additionally, variants with a SNP frequency differential greater than 25% were added to a custom-designed SNP chip which genotyped hundreds of thousands of cattle in Ireland in 2016 and beyond. The results of this genotyping will feed into future research projects and are not discussed in this thesis, as the data were unavailable at time of writing.

5 Variant validation in an independent bull population

5.1 Introduction

Investigation of polymorphisms in candidate genes for association with production traits has been one of the first approaches used to dissect the complexity of quantitative traits in livestock genomes. One study applied a whole genome scan for QTL affecting milk protein percentage in Italian Holstein cattle (Russo et al., 2012), followed by a candidate gene association study for nine economically important traits in Holstein cattle (Fontanesi et al., 2014). This candidate SNP identification, and validation approach allows identification of variants which are assumed to be functional and important, followed by association of these candidate variants, and validation of candidate SNP identification methods, in an independent, larger population.

Previously, identification of genes affecting fertility traits in dairy cattle was performed (Khatib et al., 2009) and subsequently validated in Holstein cattle which highlighted association with estimated relative conception rate for *FGF2* and *STAT5A* polymorphisms (Khatib et al., 2010). These studies showed that *FGF2* and *STAT5A* genes can be used in gene-assisted selection programs for reproductive performance in dairy cattle. This shows the importance of validation of variants in identifying causative polymorphisms.

Validation of genome-wide association studies has previously been shown to be beneficial and required to infer causality of association data, due to association not necessarily meaning causation (Konig, 2011). Validation of association data is performed in an independent population (Ioannidis et al., 2009), as distinct from replication which is performed on the same sample population.

Selection of polymorphisms is based on known association data, SNP function, or candidate SNPs. As shown in the previous chapters, several SNPs have been identified that are associated with a male fertility phenotype. A list of all 58 SNP IDs are shown in the electronic appendix[22]. In this study, 123 sires were genotyped for the 58 SNPs in three breeds, HF (n=72), LM (n=29) and BB (n=22). Of the 58 SNPs, 48 passed quality control and are shown in the electronic appendix[23]. In summary, 16 SNPs in the validation dataset were

[22] Electronic Appendix 5.12 Appendix chapter 5 – Validation\validation_snp_ids.txt
[23] Electronic Appendix 5.14 Appendix chapter 5 – Validation\SNPs_pass_qc.txt

located on chromosome 13, within the β-defensin region, a further 4 were located on chromosome 23, and 6 on chromosome 27.

5.2 Aims

The aim of this study was to validate variants, identified via a dual-sequencing approach, which are associated with a male fertility phenotype in an independent population of AI sires.

5.3 Results

5.3.1 Variant validation

The minor allele frequencies of all tested SNPs with a call rate over 80% ranged from 0.04 to 0.49 in the genotyped bull population. In this study, 58 SNPs were selected in more than 28 genes on 21 cattle chromosomes. These polymorphisms were genotyped in 123 sires of three different breeds, Holstein-Friesian (n = 79), Limousin (n = 29), and Belgian Blue (n = 22). DNA was extracted from tail hair, and genotyping was outsourced to Agena Bioscience, which used its proprietary MassARRAY genotyping system, which utilises matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry. Eleven SNPs, rs207766302, 54062, rs211574498, rs382911827, 69049, 74405, rs210032266, rs380166658, rs41614899, rs208650133 and rs383357007 were removed due to call rate < 80%. Nine SNPs, 21107, rs208577116, rs42198966, rs383833589, rs43708145, rs209828298, rs42198964, 10509 and 42210 were removed due to MAF < 0.04, although some of these SNPs were also removed via call rate.

Fifteen SNPs 10509, 21107, 42210, rs208577116, rs209828298, rs383833589, rs42198964, rs42198966, rs43708145, rs380459900, rs211675142, rs378043559, rs378340775, rs382968726 and rs473033039 were not in Hardy-Weinberg Equilibrium at P < 0.05 level of significance, as shown in Table 5.3-1.

5.3.2 SNP frequency

A list of SNPs passing QC, with call rate over 80% is shown in Table 5.3-1. SNPs are sorted based on Hardy-Weinberg *p*-value, with p indicating the major allele frequency for each SNP and q indicating the minor allele frequency. A mean of 16 'NA' allele calls (NA means missing allele calls), ranging from 5 to 23, was present for all SNPs. Four SNPs rs211092264, 42068, 6641 and rs468598994 had minor allele frequencies < 0.04 and were removed from further analysis. These four SNPs had fixed alleles in this population. These SNPs were removed during QC prior to association analysis for the whole-exome sequencing dataset. The 4 were included in the validation due to a re-analysis of the association dataset to account for genetic stratification. Following the reanalysis, these SNPs were removed.

SNP frequencies of validated SNPs along the line of regression are shown in Figure 5.3-1. The coefficient of determination for validation SNP frequencies is $R^2 = 0.948$. This shows high levels of correlation, overall, with some variance, which can be accounted for later in Table 5.3-4 and Table 5.3-5.

Table 5.3-1: Validated SNPs sorted by Hardy-Weinberg P-value, showing SNP frequencies for all 123 sires genotyped for validation. Coverage = call rate; No call = sum of NA calls; Total = sum of all possible genotype calls; common = sum of common homozygous calls; Het = sum of heterozygous calls; sum of rare homozygous calls; p = common allele frequency; q = minor allele frequency; expCommon = expected common homozygotes; expHet = expected heterozygotes; expRare = expected rare homozygotes; HWp = Hardy-Weinberg P-value.

SNPID	Chr	Position	Ref	Alt	Call Rate%	NoCall	Total	Common	Het	Rare	p	q	Exp Common	expHet	expRare	HWp
10509	13	53691410	C	T	87.8	15	108	88	15	5	0.88	0.12	84.45	22.11	1.45	0
21107	17	25057141	G	A	92.6	9	114	8	101	5	0.51	0.49	30.02	56.96	27.02	0
42210	23	26938067	C	T	81.3	23	100	92	5	3	0.94	0.06	89.3	10.4	0.3	0
rs208577116	27	4814291	G	T	90.2	12	111	3	108	0	0.51	0.49	29.27	55.46	26.27	0
rs209828298	X	58420106	A	G	84.5	19	104	70	3	31	0.69	0.31	49.16	44.69	10.16	0
rs383833589	11	73186745	A	G	86.9	16	107	27	79	1	0.62	0.38	41.33	50.34	15.33	0
rs42198964	X	3476025	G	C	85.3	18	105	46	59	0	0.72	0.28	54.29	42.42	8.29	0
rs42198966	X	3481991	A	G	92.6	9	114	63	4	47	0.57	0.43	37.06	55.88	21.06	0
rs43708145	13	61567203	C	G	95.9	5	118	37	81	0	0.66	0.34	50.9	53.2	13.9	0
rs380459900	3	1.18E+08	G	A	88.6	14	109	45	42	22	0.61	0.39	39.96	52.07	16.96	0.04
rs211675142	13	61394596	G	A	85.37	18	105	97	7	1	0.96	0.04	96.19	8.61	0.19	0.05
rs378043559	13	61340027	G	A	86.9	16	107	99	7	1	0.96	0.04	98.19	8.62	0.19	0.05
rs378340775	13	61329930	G	T	86.1	17	106	98	7	1	0.96	0.04	97.19	8.62	0.19	0.05
rs382968726	13	61416468	G	A	86.1	17	106	98	7	1	0.96	0.04	97.19	8.62	0.19	0.05
rs473033039	13	61351971	C	T	86.9	16	107	99	7	1	0.96	0.04	98.19	8.62	0.19	0.05
rs207557631	13	61531039	A	G	82.9	21	102	94	7	1	0.96	0.04	93.2	8.6	0.2	0.06
rs207958235	13	61391268	A	G	82.1	22	101	93	7	1	0.96	0.04	92.2	8.6	0.2	0.06
rs385102822	13	61378548	T	A	82.9	21	102	94	7	1	0.96	0.04	93.2	8.6	0.2	0.06
rs383941311	18	63028784	G	C	84.5	19	104	64	31	9	0.76	0.24	60.77	37.46	5.77	0.08
rs378655691	4	1.03E+08	G	A	86.1	17	106	78	28	0	0.87	0.13	79.85	24.3	1.85	0.12

Table 5.3-1 continued

rs43710844	13	61595572	G	A	82.9	21	102	38	42	22	0.58	0.42	34.13	49.75	18.13	0.12
rs43710899	13	61614764	A	C	81.3	23	100	41	41	18	0.62	0.38	37.82	47.36	14.82	0.18
rs43710842	13	61595302	T	C	86.9	16	107	42	45	20	0.6	0.4	38.88	51.24	16.88	0.21
rs43710917	13	61613313	C	T	88.6	14	109	43	46	20	0.61	0.39	39.96	52.07	16.96	0.22
71552	5	1.02E+08	A	G	87.8	15	108	76	31	1	0.85	0.15	77.52	27.96	2.52	0.26
rs43710895	13	61615527	A	G	86.1	17	106	41	46	19	0.6	0.4	38.64	50.72	16.64	0.34
rs211102509	27	36306768	T	C	89.4	13	110	62	39	9	0.74	0.26	60.38	42.23	7.38	0.42
rs379013274	27	5386029	C	T	91.8	10	113	55	45	13	0.69	0.31	53.15	48.69	11.15	0.42
rs133317825	8	1.12E+08	G	A	82.1	22	101	33	46	22	0.55	0.45	31.05	49.9	20.05	0.43
rs137792726	11	7412973	A	T	84.5	19	104	32	48	24	0.54	0.46	30.15	51.69	22.15	0.47
rs382595242	19	45726994	C	T	88.6	14	109	82	26	1	0.87	0.13	82.8	24.4	1.8	0.49
rs42670353	15	47385334	G	A	86.1	17	106	44	51	11	0.66	0.34	45.57	47.86	12.57	0.5
rs137816373	9	98323680	T	C	84.5	19	104	35	48	21	0.57	0.43	33.47	51.06	19.47	0.54
rs437191942	27	6225144	C	T	89.4	13	110	87	21	2	0.89	0.11	86.42	22.16	1.42	0.58
rs207884301	29	45029782	G	A	84.5	19	104	59	40	5	0.76	0.24	60.01	37.98	6.01	0.59
rs109065788	3	1.05E+08	A	G	82.1	22	101	46	46	9	0.68	0.32	47.14	43.72	10.14	0.6
rs210428031	18	36847662	A	C	82.9	21	102	66	33	3	0.81	0.19	66.73	31.54	3.73	0.64
rs385599841	13	61377460	G	A	93.5	8	115	106	9	0	0.96	0.04	106.18	8.65	0.18	0.66
rs210662027	27	5162095	C	A	84.5	19	104	60	39	5	0.76	0.24	60.77	37.46	5.77	0.67
rs378294791	18	36888695	A	C	85.3	18	105	70	32	3	0.82	0.18	70.44	31.12	3.44	0.77
rs137356698	18	36955817	A	G	86.9	16	107	68	35	4	0.8	0.2	68.32	34.36	4.32	0.85
rs110682224	23	47671466	T	C	85.3	18	105	37	50	18	0.59	0.41	36.61	50.78	17.61	0.87
rs211092264	17	25119394	G	A	83.7	20	103	102	1	0	1	0	102	1	0	0.96
42068	23	26926961	G	A	82.9	21	102	102	0	0	1	0	102	0	0	1
6641	12	708425	G	A	89.4	13	110	110	0	0	1	0	110	0	0	1
rs468598994	27	6224921	G	A	84.5	19	104	104	0	0	1	0	104	0	0	1
Mean						16.6	106	62.2	35	9.0	0.75	0.24	62.3	34.9	9.12	0.27
Min						5	100	3	0	0	0.51	0	29.27	0	0	0
Max						23	118	110	108	47	1	0.49	110	56.96	27.02	1

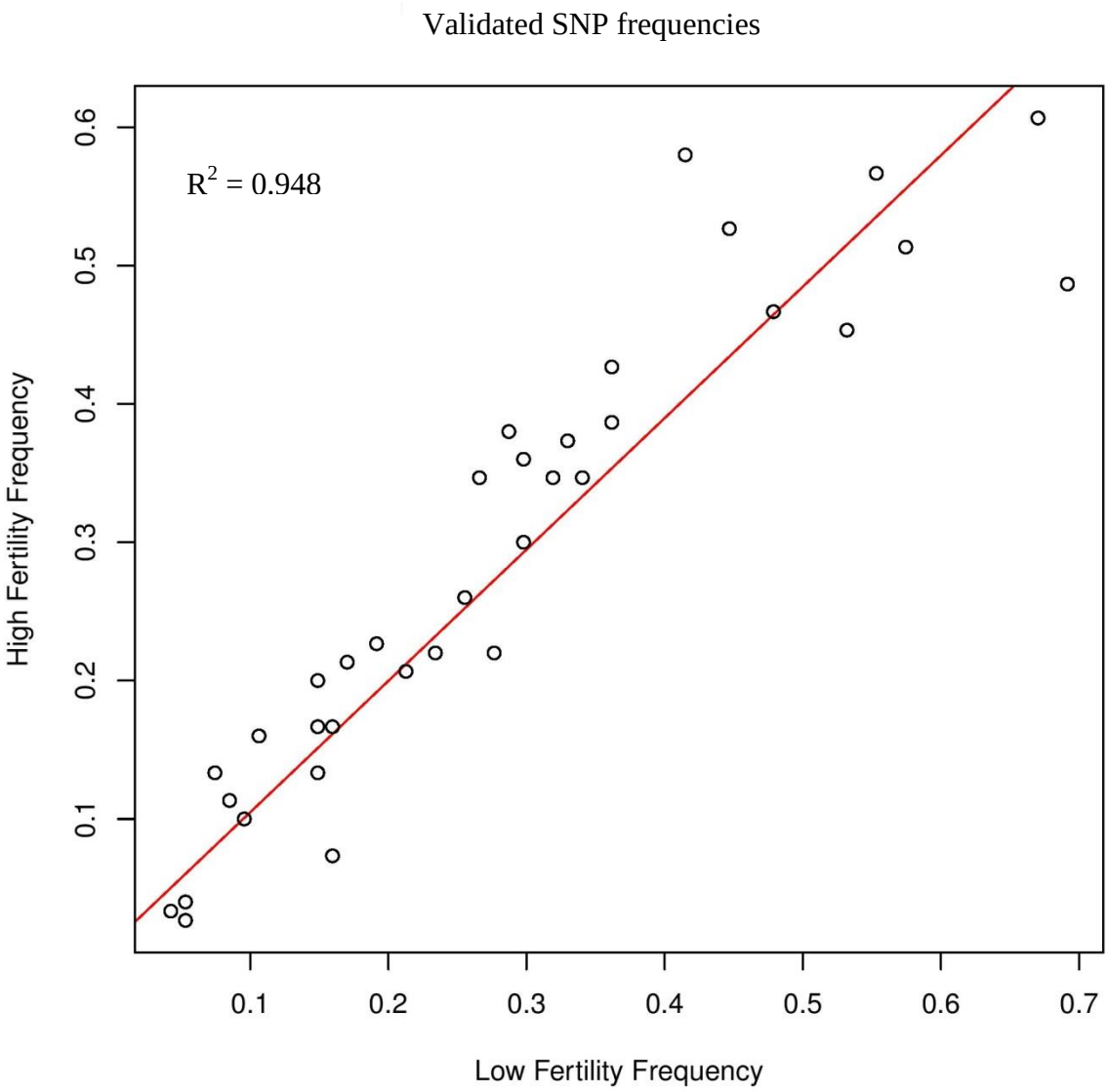


Figure 5.3-1: SNP frequencies of validated SNPs in independent population of bulls.
Scatterplot of SNP frequencies of validated SNPs (n=42) with x-axis = low-fertility bulls SNP frequencies and y-axis = high-fertility bulls SNP frequencies (n=123 bulls). Red line = line of regression. R^2 = Pearson's correlation coefficient. The most significantly associated SNP in the validation dataset (located in FOXJ3) is highlighted in red (SNP frequency difference =21% (48%-69%).

5.3.3 Correlation

In total, 28 SNPs were compared between validation SNP calls and Whole-exome sequencing SNP calls, to determine correlation between the SNP frequencies between both genotyping methods. For high-fertility bulls the SNP frequency correlation was 0.42, a

moderate positive correlation between genotyping methods, whereas low fertility bulls had a weak positive correlation of 0.25.

Total genotyping rate was 0.8279. This is the proportion of genotypes per SNP with non-missing data. These data mean that 83% of genotyped bulls had SNP calls on average. This figure is in line with expectations, due to failure of the SNP genotyping assay in 1 column of the assay plate containing DNA from 5 bulls. These 5 bulls were omitted from further analysis, leaving n=118.

In total, seven SNPs had a *P*-value less than 0.1, and of these, four SNPs were below the significance threshold of *P*-value < 0.05 for association to adjusted animal model phenotype. For pregnancy rate, 8 SNPs had an unadj. *P*-value < 0.01. SNPs with unadj. P < 0.01 are shown in greater detail, including SNP frequencies in high- and low-fertility groups in Table 5.3-4 for AAM and Table 5.3-5 for PR. A full list of validation SNP association data is available in Table 5.3-2 for AAM and Table 5.3-3 for PR.

Table 5.3-2: Validation SNPs associated with AAM fertility phenotype.

Table of SNPs associated with fertility phenotypes. CHR = chromosome, SNP = SNP ID, BP = location (Base pair), NMISS denotes the number of animals with genotypes for the SNP after QC, from a total of 123. BETA refers to the regression coefficient for each single SNP regression, with SE being the associated standard error for this measure. R^2 is how close the SNP data are to the line of regression. T is the Wald test distribution. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution. *P*-value is the Wald test asymptotic *P*-value.

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
3	rs109065788	1.05E+08	101	0.01473	0.004565	0.09519	3.227	0.001695
30	rs42198966	3481991	110	-0.00775	0.002952	0.05996	-2.625	0.009931
27	rs437191942	6225144	110	0.01466	0.006308	0.04762	2.324	0.02201
4	rs378655691	1.03E+08	106	-0.01466	0.006881	0.04182	-2.131	0.03548
13	rs43710895	61615527	106	0.007818	0.004226	0.03186	1.85	0.06714
13	rs43710844	61595572	102	0.007353	0.004096	0.03122	1.795	0.07566
13	10509	53691410	108	0.01016	0.005786	0.02829	1.757	0.08185
15	rs42670353	47385334	106	0.007159	0.004502	0.02374	1.59	0.1148
13	rs43710842	61595302	107	0.006223	0.004162	0.02085	1.495	0.1379
13	rs43710899	61614764	100	0.006032	0.004048	0.02216	1.49	0.1394
27	rs208577116	4814291	111	-0.02565	0.01759	0.01914	-1.459	0.1476
13	rs43710917	61613313	109	0.005538	0.003898	0.01852	1.421	0.1583
27	rs211102509	36306768	110	0.006093	0.004622	0.01583	1.318	0.1902
13	rs385599841	61377460	115	0.01418	0.01108	0.0143	1.28	0.2031
13	rs43708145	61567203	118	0.007457	0.006325	0.01184	1.179	0.2408
23	42210	26938067	100	0.00893	0.007762	0.01332	1.15	0.2528
11	rs137792726	7412973	104	0.003431	0.003962	0.007295	0.8658	0.3886
30	rs209828298	58420106	101	0.002415	0.003318	0.005321	0.7277	0.4685
29	rs207884301	45029782	104	-0.00368	0.005094	0.005089	-0.7223	0.4718
17	21107	25057141	114	0.005825	0.008626	0.004055	0.6753	0.5009
5	71552	1.02E+08	108	-0.00408	0.006346	0.003885	-0.643	0.5216

Table 5.3-2 continued

27	rs210662027	5162095	104	0.003048	0.005372	0.003146	0.5674	0.5717
18	rs210428031	36847662	102	0.00259	0.005432	0.002269	0.4769	0.6345
11	rs383833589	73186745	107	-0.00295	0.006548	0.001932	-0.4509	0.653
18	rs137356698	36955817	107	0.002433	0.005515	0.001851	0.4412	0.6599
13	rs385102822	61378548	102	0.003833	0.008869	0.001864	0.4322	0.6666
8	rs133317825	1.12E+08	101	0.001781	0.004205	0.00181	0.4237	0.6727
23	rs110682224	47671466	105	-0.00168	0.004502	0.001342	-0.3721	0.7106
27	rs379013274	5386029	113	-0.00146	0.004161	0.001112	-0.3515	0.7258
13	rs207958235	61391268	101	0.00299	0.009266	0.001051	0.3227	0.7476
18	rs383941311	63028784	104	0.001408	0.00467	0.00089	0.3015	0.7637
18	rs378294791	36888695	105	-0.00131	0.0053	0.000596	-0.2477	0.8048
13	rs382968726	61416468	106	0.002304	0.009521	0.000563	0.242	0.8093
13	rs207557631	61531039	102	0.002392	0.01017	0.000553	0.2351	0.8146
13	rs378043559	61340027	107	0.002126	0.009651	0.000462	0.2203	0.8261
19	rs382595242	45726994	109	0.001376	0.006496	0.000419	0.2118	0.8327
13	rs211675142	61394596	105	0.001983	0.009734	0.000403	0.2037	0.839
9	rs137816373	98323680	104	-0.00074	0.004154	0.000313	-0.1788	0.8585
13	rs378340775	61329930	106	0.001641	0.009765	0.000271	0.168	0.8669
13	rs473033039	61351971	107	0.001469	0.009925	0.000209	0.148	0.8826
3	rs380459900	1.18E+08	109	-9.45E-05	0.003944	5.36E-06	-0.02396	0.9809
30	rs42198964	3476025	46	NA	NA	NA	NA	NA

Table 5.3-3: Validation SNPs associated with PR fertility phenotype.

Table of SNPs associated with fertility phenotypes. CHR = chromosome, SNP = SNP ID, BP = location (Base pair), NMISS denotes the number of animals with genotypes for the SNP after QC, from a total of 123. BETA refers to the regression coefficient for each single SNP regression, with SE being the associated standard error for this measure. R^2 is how close the SNP data are to the line of regression. T is the Wald test distribution. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution. *P*-value is the Wald test asymptotic *P*-value.

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
13	rs378340775	61329930	104	-0.03283	0.01041	0.0888	-3.153	0.002124
8	rs133317825	112149281	101	-0.06978	0.02404	0.0784	-2.902	0.00457
18	rs137356698	36955817	107	-0.06974	0.02473	0.07037	-2.819	0.005754
17	21107	25057141	107	-0.06793	0.02444	0.06855	-2.78	0.006446
13	rs43710917	61613313	106	-0.06789	0.02448	0.06887	-2.774	0.006573
13	rs43710842	61595302	106	-0.06772	0.02452	0.06834	-2.762	0.006793
13	rs382968726	61416468	105	-0.06793	0.02468	0.06851	-2.752	0.006992
11	rs383833589	73186745	102	-0.06677	0.02464	0.0684	-2.71	0.007927
4	rs378655691	103370232	101	0.02736	0.01216	0.04865	2.25	0.02666
27	rs379013274	5386029	110	0.03051	0.01674	0.02982	1.822	0.07121
9	rs137816373	98323680	102	-0.04409	0.02558	0.02885	-1.724	0.08789
13	rs473033039	61351971	104	0.02207	0.01306	0.02723	1.69	0.09415
13	rs43708145	61567203	105	-0.02367	0.01553	0.02206	-1.524	0.1305
29	rs207884301	45029782	114	-0.01145	0.007672	0.01951	-1.493	0.1383
13	rs43710899	61614764	106	0.01613	0.01203	0.01698	1.34	0.183
13	rs43710844	61595572	106	-0.02189	0.01751	0.0148	-1.25	0.2141
13	rs207557631	61531039	105	-0.01668	0.01489	0.01205	-1.121	0.265
13	rs207958235	61391268	104	-0.01329	0.01222	0.01147	-1.088	0.2792
30	rs42198964	3476025	13	0.02725	0.02431	0.1025	1.121	0.2862

Table 5.3-3 continued

3	rs380459900	117667441	100	0.02139	0.01995	0.01159	1.072	0.2863
27	rs208577116	4814291	109	-0.01079	0.01062	0.009561	-1.016	0.3118
11	rs137792726	7412973	102	-0.01471	0.01485	0.009724	-0.9909	0.3241
27	rs437191942	6225144	111	-0.0438	0.04733	0.007794	-0.9253	0.3568
18	rs378294791	36888695	107	-0.009702	0.01072	0.007737	-0.9048	0.3676
3	rs109065788	104587541	100	-0.008924	0.0109	0.006795	-0.8188	0.4149
15	rs42670353	47385334	107	-0.01134	0.01394	0.006266	-0.8137	0.4177
13	rs211675142	61394596	105	-0.008709	0.01137	0.005664	-0.766	0.4455
27	rs210662027	5162095	110	0.008154	0.01206	0.004218	0.6763	0.5003
18	rs383941311	63028784	108	-0.01013	0.01515	0.0042	-0.6686	0.5052
13	rs378043559	61340027	104	-0.007015	0.01082	0.004102	-0.6482	0.5183
13	rs385599841	61377460	104	-0.005412	0.008727	0.003757	-0.6202	0.5365
13	rs43710895	61615527	106	-0.006804	0.01097	0.003683	-0.62	0.5366
13	10509	53691410	102	-0.006671	0.01086	0.003762	-0.6145	0.5403
19	rs382595242	45726994	108	-0.009041	0.01625	0.002912	-0.5564	0.5791
18	rs210428031	36847662	107	-0.008875	0.0173	0.002502	-0.5132	0.6089
23	42210	26938067	109	0.004717	0.01035	0.001936	0.4556	0.6496
27	rs211102509	36306768	113	-0.005061	0.01129	0.001806	-0.4482	0.6549
23	rs110682224	47671466	109	-0.007514	0.01722	0.001776	-0.4363	0.6635
13	rs385102822	61378548	104	0.004417	0.01371	0.001016	0.3221	0.748
5	71552	102028024	101	-0.0008962	0.01099	6.72E-05	-0.08157	0.9352
30	rs42198966	3481991	106	NA	NA	NA	NA	NA
30	rs209828298	58420106	37	NA	NA	NA	NA	NA

Table 5.3-4: Validated variants associated with adjusted animal model fertility phenotype.

Table of validated variants associated with the adjusted animal model fertility phenotype *P*-value < 0.1. Gene = SNP annotated to nearest gene; CHR = Chromosome number; SNP = SNP identifier; BP = Physical position (base-pair); NMISS = Number of non-missing genotypes; BETA = Regression coefficient; SE = Standard error; R2 = Regression r-squared; T = Wald test (based on t-distribution); Frequency 'low' = SNP frequency in low-fertility bulls; Frequency 'high' = SNP frequency in high-fertility bulls; P = Wald test asymptotic *P*-value; BH = Benjamini-Hochberg multiple testing correction.

Gene	**CHR**	**SNP**	**BP**	**NMISS**	**BETA**	**SE**	**R^2**	**T**	**Frequency 'low'**	**Frequency 'high'**	***P*-value**	**BH**
FOXJ3	3	rs109065788	104587541	101	0.01473	0.0045	0.095	3.227	0.691	0.486	0.0016	0.069
LOC536148	30	rs42198966	3481991	110	-0.007747	0.0029	0.059	-2.625			0.0099	0.203
LOC100337009	27	rs437191942	6225144	110	0.01466	0.0063	0.047	2.324	0.085	0.113	0.0220	0.300
KIAA1549	4	rs378655691	103370232	106	-0.01466	0.0068	0.041	-2.131	0.074	0.133	0.0354	0.363
DEFB124	13	rs43710895	61615527	106	0.007818	0.0042	0.031	1.85			0.0671	0.479
DEFB123	13	rs43710844	61595572	102	0.007353	0.0040	0.031	1.795	0.531	0.453	0.0756	0.479
SIRPA	13	10509	53691410	108	0.01016	0.0057	0.028	1.757	0.095	0.1	0.0818	0.479

Table 5.3-5: Validated variants associated with pregnancy rate fertility phenotype.

Table of validated variants associated with the Pregnancy rate fertility phenotype with *P*-value < 0.1). CHR = Chromosome number; SNP = SNP identifier; BP = Physical position (base-pair); NMISS = Number of non-missing genotypes; BETA = Regression coefficient; SE = Standard error; R2 = Regression r-squared; T = Wald test (based on t-distribution); Frequency 'low' = SNP frequency in low-fertility bulls; Frequency 'high' = SNP frequency in high-fertility bulls; P = Wald test asymptotic p-value (P < 0.05); BH = Benjamini-Hochberg multiple testing correction.

Gene	**CHR**	**SNP**	**BP**	**NMISS**	**BETA**	**SE**	**R2**	**T**	**Frequency 'low'**	**Frequency 'high'**	**P**	**BH**
DEFB128	13	rs378340775	61329930	104	-0.0328	0.010	0.0888	-3.153	0.053	0.026	0.0021	0.084
PHF19	8	rs133317825	112149281	101	-0.0697	0.024	0.0784	-2.902	0.329	0.373	0.0045	0.178
NOB1/ WWP2	18	rs137356698	36955817	107	-0.0697	0.024	0.0703	-2.819	0.148	0.2	0.0057	0.218
HPVC1	17	21107	25057141	107	-0.0679	0.024	0.0685	-2.78	0.478	0.466	0.0064	0.238
DEFB124	13	rs43710917	61613313	106	-0.0678	0.024	0.0688	-2.774	0.287	0.38	0.0065	0.238
DEFB123	13	rs43710842	61595302	106	-0.0677	0.024	0.0683	-2.762	0.574	0.513	0.0067	0.238
DEFB115	13	rs382968726	61416468	105	-0.0679	0.024	0.0685	-2.752	0.053	0.026	0.0069	0.238
LOC101906131	11	rs383833589	73186745	102	-0.0667	0.024	0.0684	-2.71	0.319	0.346	0.0079	0.261

5.4 Discussion

A SNP genotyping assay was designed for validation of 58 SNPs identified as being associated with bull fertility phenotypes, adjusted animal model and pregnancy rate, see electronic appendix[24]. Of these 58 SNPs which were targeted in the assay design, 42 passed all QC filters and were validated in an independent population of 123 bulls for allele frequencies and association to the adjusted animal model and pregnancy rate fertility phenotypes.

The SNP most associated with adjusted animal model of fertility, rs109065788, is in the exonic region of *FOXJ3*, on chromosome 3 at position 104,587,541 and results in a synonymous mutation. This is the only significant SNP associated with adjusted animal model after Benjamini-Hochberg correction for multiple comparison testing, with an adjusted *P*-value < 0.1, *P* = 0.086. *FOXJ3* is a forkhead box protein transcription factor. Interestingly, in mice, *FOXJ3* has been shown to be required for the survival of spermatogonia and is involved in spermatocyte meiosis (Ni et al., 2016). A recent study deleted *Foxj3* from either spermatogonia or meiotic spermatocytes. Results showed that both models exhibited complete male sterility, but with different etiologies. *Foxj3* knockout resulted in decreased testis weight, complete sterility and no round spermatids were found in the seminiferous tubules. Therefore, *Foxj3* is required for survival of spermatogonia.

The fifth most associated SNP in the validation dataset, rs43710895, is in the 5'UTR region of *DEFB124*. Previous work by our group has identified a SNP in *DEFB124* exon 1 in HF bulls at a high minor allele frequency, 47% (Narciandi et al., 2011). The SNP in exon 1 of *DEFB124* was further shown to have lower MAF in Norwegian Red cattle with 28%, a difference of ~20% SNP frequency between breeds. This SNP and rs43710895 are located approximately 150bp apart and may be inherited together. In this whole-exome sequencing dataset, rs43710895 has a SNP frequency of 62% in high-fertility animals, and 38% for low-fertility animals. These data highlight the selective pressures β-defensin genes are under in bulls of different breeds.

The SNP most associated with the pregnancy rate phenotype, rs378340775, is in the 3'UTR of *DEFB128* at position 61329930, in a gene which was not annotated in the bovine genome

[24] Electronic Appendix 5.2 Validation\Assay Design\ 2 - Superplex1.xslx

at the time of designing baits for whole-exome sequencing capture. *DEFB128* is located on chromosome 13, and is one of 19 genes identified previously by our group as being expressed in the reproductive tracts of adult cows and bulls. From the validation dataset, the SNP frequency in low-fertility and high-fertility bulls was 5.3% and 2.6%, respectively, see Table 5.3-5. This lack of SNP frequency difference in fertility groups means this SNP is likely to be a contributor to overall bull fertility, as a polygenic trait, but is unlikely to have an impact in isolation of other variants.

DEFB123 is one of 19 genes expressed in the male reproductive tract of bulls (Narciandi et al., 2011). A SNP located in *DEFB123*, rs43710844, is located on chromosome 13 at position 61595572. This SNP is just 265 kbp (kilobase pairs) away from rs378340775 in *DEFB128*, with a SNP frequency of 53% in low-fertility bulls and 45% in high-fertility bulls, see Table 5.3-4. These SNPs, along with others in the same genomic region on chromosome 13, are associated with and contribute to the polygenic fertility phenotypes, adjusted animal model and pregnancy rate. To determine the accuracy of the identified associations, these SNPs were added to the IDB customised SNP-chip for genotyping in a large Irish cattle population, and association of these variants with over 40 phenotypes, including female fertility, and immunological traits.

Correlation of SNP frequencies in low-fertility bulls and high-fertility bulls, between whole-exome sequencing (n=24) sires, and validated (n=123) sires was also performed. Low-fertility bulls had a weak SNP frequency correlation of 0.25, whereas high-fertility bulls had a moderate SNP frequency correlation of 0.42. This level of correlation may have been affected by the relatively low sample number for whole-exome sequencing and the validated dataset is probably a more accurate representation of SNP frequencies due to the increase in sample number. This question will be answered conclusively following SNP genotyping via the IDB SNP chip in ~330,000 cattle.

Of the 58 SNPs selected for genotyping, 19 were from the targeted β-defensin sequencing study, the remaining 39 were from the whole-exome sequencing dataset. Following QC, 42 SNPs were associated with fertility phenotypes and of these, 5 were validated as being associated with AAM at $P < 0.1$ and 6 were associated with PR at $P < 0.1$. After Benjamini-Hochberg correction, 1 SNP is significant in each association, *FOXJ3* for AAM, and *DEFB128* for PR.

6 Final discussion

In recent decades, molecular genetic research has made significant contributions to increasing livestock efficiency by identifying mutations underlying single-gene disorders which have been eliminated from the breeding stock (Charlier et al., 2012, Charlier et al., 2008). The identification of genes and genetic variants associated with cow fertility and production traits has facilitated the selection of economically superior animals to such an extent that a reversal in the previously documented decline in female fertility has been observed and simultaneous genetic improvement in both traits is possible (Berry et al., 2016). The advent of genomic selection means that genetic progress can accelerate gain, even in low heritability traits. While it is known that intensive selection for production traits can have a negative impact on health and fertility characteristics (Veerkamp and Beerda, 2007, Berry et al., 2014), multi-trait selection indices now account for unfavourable relationships in the generation of a more sustainable cow. However, limitations remain. For some important traits, such as bull fertility, phenotypes are limited in number and in terms of reliability and so cannot be accurately measured and monitored. Furthermore, the rate of genetic progress is so rapid that other as yet unidentified antagonistic relationships could only be uncovered when significant phenotypic effects have been observed. For example, health traits are not accurately accounted for in most selection programmes, and these traits are known to be unfavourably related to production, and are particularly important in this period of agricultural expansion. Therefore, it is critical that we understand the mechanisms underlying genomic selection in cattle, and in that regard, a comprehensive characterisation of the genes and processes involved in bull fertility is warranted.

Cattle have undergone a rapid decrease in effective population size, mainly due to domestication and intensive selection (Bovine HapMap et al., 2009). Artificial selection has influenced the genetic structure of the bovine genome; however, the variation within breeds is similar to human genetic variation (Bovine HapMap et al., 2009). This shows the importance of highlighting variant deviation in SNP frequencies due to natural selection pressures. One group of genes undergoing natural selection pressures are the β-defensin genes, due to their newly discovered multi-function ability in host defence and emerging roles in reproduction.

Antimicrobial peptides (AMPs), specifically β-defensins, have been shown to have a dual role in host defence against pathogens and in the regulation of male fertility in rodents and

in humans (Tollner et al., 2011, Zhou et al., 2013). Male mice with a β-defensin gene cluster knock-out are completely sterile and a dinucleotide deletion in the human *DEFB126* exon resulted in a 40% reduction in the ability of couples to conceive. Previous research by our group identified an expansion of β-defensin genes in the bovine genome (Cormican et al., 2008), and functional characterisation documented expression of these genes in the reproductive tract of the bull (Narciandi et al., 2011). However, the association of these genes with fertility in cattle had not previously been investigated.

This is the first targeted sequencing of novel β-defensin genes in AI bulls with divergent fertility phenotypes. A targeted capture of all 57 β-defensin genes in 168 AI bulls divergent for two fertility phenotypes, pregnancy rate and adjusted animal model (a statistical model of pregnancy rate data) was performed. Genetic variants identified, after quality control filtering, were associated with the AAM fertility phenotype that identifies genetic variants most associated with male fertility. The SNP most associated was rs378043559, located in the upstream region of *DEFB127* at position 61340027 on chromosome 13 (unadjusted *P*-value = 0.00197). Interestingly, a group of 97 SNPs, located on chromosome 13 were the second most associated variants (*P*-value = 0.00202). The 97 SNPs are contained within 7 β-defensin genes (*BBD142, BBD128, BBD127, BBD126, BBD125, BBD116* and *BBD115*), covering 137 kbp. These 97 SNPs and rs378043559, the most significantly associated SNP, are all heterozygous in 9 sires (5 Holstein Friesian, 1 each of Limousin, Simmental, Charolais and Belgian Blue) of medium to high AAM fertility (0 to 0.07, mean 0.04, s.d. 0.027). Other variants were also predicted to have synonymous and non-synonymous effects on β-defensin genes: a non-synonymous SNP in BBD115 (Ser52Asn) and two synonymous SNPs in *BBD126* and *BBD125* were identified. *In silico* prediction of o-linked glycosylation sites in these and all 19 β-defensin genes located on chromosome 13 identified predicted o-linked glycosylation sites in the tail region of *BBD115*, *BBD125*, and *BBD126*. This indicates that altered glycosylation in these β-defensin genes could affect sperm penetration ability in cattle.

A recent study by our group showed expression of *BBD126* on the caudal sperm surface, by confocal microscopy, with staining concentrated on the sperm cell tails, where glycosylation is predicted to occur (Narciandi et al., 2016). Following SNP discovery and association of a haplotype of SNPs with a fertility phenotype in this project, subsequent work performed by

other members of our group demonstrated that sperm from high fertility bulls with the β- significantly better at binding to oviductal epithelial cells, a key step in fertilisation where sperm aggregate in the oviduct prior to capacitation, compared to high-fertility bulls without the haplotype and low-fertility bulls (Finlay, E. et al. 2017. unpublished data). Together, these data support a role for *BBD126* in regulating male fertility in the bovine. In addition, glycosylation analysis of sperm from bulls of high and low-fertility may help identify whether altered glycosylation in bull sperm influences sperm function in bovine, as has been demonstrated in humans (Tollner et al., 2011). Defensins may modulate the ability of sperm to penetrate cervical mucus via altered glycosylation of the peptide tail, as has been shown in primates (Tollner et al., 2008) and humans (Tollner et al., 2011). However, as the variants are inherited as a haplotype, further studies would be required to identify the causative variant(s).

Interestingly, a recent study detected a QTL for bull fertility located on chromosome 13, just 1.3Mbp from the β-defensin gene region (Han and Penagaricano, 2016). Given the significant structural variation in the β-defensin region, particularly in the chromosome 13 cluster, further accurate annotation of this gene cluster will be required to accurately enable assessment of their role in bull fertility.

Following identification and association of β-defensin genetic variants with male fertility, a genome-wide identification and association of exon and promoter region variants was performed to identify other potential contributing variants. As β-defensin genes were not sufficiently annotated in the bovine genome at the time of capture probe design, a custom-designed exome capture probe set was used to sequence the annotated bovine genome. Exome sequencing data in bovine has not been widely published (Cosart et al., 2011, Hirano et al., 2013, McClure et al., 2014a). Fertility and milk production in Nordic Red cattle was assessed by McClure et al. however, previous studies have not analysed fertility in popular Irish breeds using exome sequencing. A successful probe design provided good coverage of the bovine genome at low depth. The level of exome coverage at low read depth makes the capture design a viable tool for further exome sequencing studies for SNP discovery, as exome captures have previously been used in other studies (Robert et al., 2014, Cosart et al., 2011). Variant calling, after filtering and quality control, identified 144,178 variants in 24 bulls. Candidate SNPs for male fertility were identified by association with AAM fertility

phenotype, using the R package, GenABEL. Single SNP regression of variants in a linear mixed model with breed as a fixed effect identified 484 SNPs associated with the AAM fertility phenotype (unadjusted $P < 0.1$). The term 'glycoprotein' was a significantly over-represented keyword in this dataset (*P*-value = 0.0056). The term 'glycosylation site: N-linked' is a significantly over-represented feature in this dataset (*P*-value = 0.00024). Gene-ontology terms such as 'innate immunity', 'complement', 'immune response' and 'acute inflammatory response' were amongst the most over-represented terms in the GO dataset. This indicates the importance of immune system related genes in male fertility, and demonstrates the emphasis on innate immune gene variation in this study was warranted, and concurs with targeted sequencing analysis of β-defensin genes above.

Variants identified as having a SNP frequency differential of greater than 25% were added to a custom-designed SNP-chip which was used to genotype ~300,000 cattle in Ireland in 2016, and in future years. Future studies are needed to associate additional variants identified as being associated with a male fertility phenotype, as well as with additional available phenotypes of economic importance in large numbers of cattle of various breeds.

Validation of genetic variants in genome-wide association studies is required to infer causality of a genetic factor to a phenotype (Konig, 2011). To validate the genetic variants identified via whole-exome sequencing and targeted sequencing of the β-defensin gene cluster, validation was performed in an independent dataset of AI bulls. Of the 58 SNPs selected for genotyping, 19 were from the targeted β-defensin sequencing study, the remaining 39 were from the whole-exome sequencing dataset. Following QC, 42 SNPs were associated with fertility phenotypes, and of these, 5 validated as being associated with AAM at $P < 0.1$, and 6 associated with PR at $P < 0.1$. After Benjamini-Hochberg correction, 1 SNP is significant in each association, *FOXJ3* for AAM, and *DEFB128* for PR. A successful validation of 5 SNPs for AAM and 6 for PR show the successful sequencing of bovine exome and β-defensin region in AI bulls divergent for fertility for the first time. *FOXJ3* is a transcription factor which has been shown to play an important role in male fertility in mice. Further studies are required to determine the function of *FOXJ3* in bovine. This shows the success of a novel whole-exome bait capture design in identifying exome-wide variants and their association with relevant phenotypes.

By applying the Bradford-Hill criteria, the association of these SNPs with fertility can be assessed (Bradford, 1965). The strength, consistency, specificity, temporality biological gradient, plausibility, coherence, experiment and analogy are a set of nine criteria to provide epidemiologic evidence of a causal relationship between a cause and effect. In this analysis, we can say there is a strong association for the FOXJ3 SNP and the β-defensin haplotype with male fertility. This can be concluded from the identification of FOXJ3 as being associated with AAM in two independent datasets, even in small numbers of bulls. The haplotype was also significantly associated with fertility in a larger number of bulls, and the β-defensin region was found to have a cluster of SNPs associated with fertility in the WES dataset as well. In addition, DEFB128, a gene in the haplotype, was validated in an independent population as being significantly associated with pregnancy rate. This also demonstrates a consistent association for both the haplotype and FOXJ3 across independent populations of bulls, although further data on these variants in diverse populations would be welcomed. In addition, the association is plausible and coherent, as FOXJ3 has been shown by others (Ni et al., 2016) to be necessary for survival of spermatogonia and deletion results in sterility in male mice. Also, the β-defensin haplotype has been shown to function in bovine oviductal epithelial cell binding (BOEC), with significant differences in the ability of high-fertility bulls with the haplotype to cluster and bind to BOECs compared to high-fertility bulls without the haplotype (unpublished data[25]).

In this thesis, two fertility phenotypes were available: PR and AAM, which is a model of PR to include additional fixed and random effects, as outlined in Table 1.2-1. During association analysis in the validation dataset, both PR and AAM were associated with the genotypes identified, although the most associated SNPs were not similar. Given that the AAM is based on PR, one might expect similar SNPs associated with both phenotypes. However, this may be explained due to female-specific fertility effects in the AAM, which decouples the effects of female fertility that is present in the PR phenotype, but has been accounted for in the AAM phenotype. Therefore, the AAM is highlighting the male-specific factors related to fertility. One way to examine this in the data is to look at the function of the SNPs significantly associated in the two phenotypes, and determine whether one is more male-

[25] Paper submitted to Scientific Reports (Whiston et al. 2017. "A dual targeted β-defensin and exome sequencing approach to identify, validate and functionally characterise genes associated with bull fertility).

specific compared to the other. Indeed, the FOXJ3 SNP is involved in survival of spermatogonia as described in the previous paragraph, a male-specific function. In contrast, the DEFB128 SNP is in the β-defensin haplotype region, involved in binding to oviductal epithelial cells, which is dependent on interactions with the female reproductive tract.

Recently, an international consortium of researchers, led by the Roslin institute and EMBL-EBI, developed the functional annotation of animal genomes (FAANG) project. The aim of this project was the functional annotation of genomes of domesticated animals. They have identified improvement of animal reference genome sequences and comprehensive annotation of their functional elements and variants as priorities for understanding the link between genotype and phenotype. Development of improved annotation of the bovine genome would also improve association analysis studies such as GWAS, and improve 'omics' studies in cattle, as poor annotation of the bovine genome in comparison to the latest human genome annotation (hg38) is limiting bovine genomic analyses.

The beef data and genomics programme (BDGP) 2015-2020, was launched by the Irish department of agriculture, food and the marine (DAFM) to improve the genetic merit of the national beef herd through collection of data and genotypes of selected animals which will allow for genomic selection in the beef herd to advance genetic gain in beef cattle. The BDGP aims to support the suckler herd by improving the genetic merit of the national herd through the collection of phenotypic data and genotyping animals and to improve quality and efficiency. The genetic variants identified in this thesis may be incorporated into the BDGP to achieve this objective. Inclusion of genetic variants associated with fertility, from bulls divergent for fertility and validated in an independent population can be selected for to improve genomic selection, without affecting other economic traits or health traits.

In conclusion, this is the first comprehensive analysis of sequence variation present in bovine β-defensin genes, the first whole-exome sequencing of AI bulls divergent for a fertility phenotype, and has successfully identified novel variants associated with a pregnancy rate phenotype. The proposed hypothesis that genetic variation in exons and promoter regions of β-defensin genes explain a portion of phenotypic variation in AI bulls divergent for fertility, can be accepted, due to the association of genetic variants with the phenotype, validation of several β-defensin gene SNPs, and subsequent functional

characterisation by other members of our group. The exact portion of the observed phenotype that is explained by these SNPs requires further study.

Our second proposed hypothesis was that genome-wide genetic variation of the exome and promoter regions will explain a portion of phenotypic variation in AI bulls divergent for fertility. Again, this hypothesis can be considered accepted. This conclusion is based on the association of variants with a fertility phenotype, validation of SNPs in an independent population, and identification of over-represented gene ontology terms related to the immune system, and glycosylation, possibly implicating β-defensin genes.

This thesis contributes to the identification of genetic variants and biological processes underlying sire fertility. Genetic variants identified as being associated with fertility and validated in an independent bull population may be used to improve breeding strategies via marker assisted selection and in the future aid in developing genetic biomarkers for male fertility.

6.1 Future research opportunities

1. Candidate SNPs (n = 863) identified in this thesis have been added to the IDB SNP chip v3 and may be used to perform association analysis with phenotypes available from Irish cattle breeding federation. Over 40 phenotypes are available for both dairy and beef breeds. Available phenotypes are related to production, and fertility, although male fertility is not currently a routinely recorded phenotype in Ireland. Improving male fertility phenotypes is a priority for improving the genetic gain for bull fertility.
2. In collaboration with XLVets Ireland, we obtained bull fertility records from breeding soundness evaluations for stock bulls to create the first phenotypic database on stock bull fertility in Ireland. Due to the highly-selected AI bull population, it is expected that stock bulls will have larger variation in male fertility phenotypes, and future research may focus on identifying the variants associated with male fertility in the stock bull population. The β-defensin SNPs added to the IDB SNP chip will facilitate association analysis with stock-bull fertility.
3. In addition, complementary work has validated the effects of a β-defensin haplotype on the divergent ability of sperm from high and low bulls to bind to oviductal epithelial cells (unpublished data). In a similar manner, bulls carrying specific β-defensin variants could be further evaluated to identify causative mutations underlying this haplotype, or indeed bulls carrying *FOXJ3* variants. *FOXJ3* knockout in mice affected testis weight, resulted in completely sterile male mice, and altered seminiferous tubule development (Ni et al., 2016). Therefore, sperm from bulls carrying *FOXJ3* variants could be assessed using computer-aided sperm analysis (CASA) (Amann and Waberski, 2014). *In vivo* field fertility data could be used to determine whether the *FOXJ3* variants were affecting field fertility.

7 References

AGERHOLM, J. S., MCEVOY, F. & ARNBJERG, J. 2006. Brachyspina syndrome in a Holstein calf. *J Vet Diagn Invest,* 18**,** 418-22.

AKINLOYE, O., GROMOLL, J., CALLIES, C., NIESCHLAG, E. & SIMONI, M. 2007. Mutation analysis of the X-chromosome linked, testis-specific TAF7L gene in spermatogenic failure. *Andrologia,* 39**,** 190-5.

AKINLOYE, O., GROMOLL, J., NIESCHLAG, E. & SIMONI, M. 2009. Androgen receptor gene CAG and GGN polymorphisms in infertile Nigerian men. *J Endocrinol Invest,* 32**,** 797-804.

AL NAIB, A., HANRAHAN, J. P., LONERGAN, P. & FAIR, S. 2011. In vitro assessment of sperm from bulls of high and low field fertility. *Theriogenology,* 76**,** 161-7.

ALGHAMDI, A. S., FUNNELL, B. J., BIRD, S. L., LAMB, G. C., RENDAHL, A. K., TAUBE, P. C. & FOSTER, D. N. 2010. Comparative studies on bull and stallion seminal DNase activity and interaction with semen extender and spermatozoa. *Anim Reprod Sci,* 121**,** 249-58.

ALGHAMDI, A. S., LOVAAS, B. J., BIRD, S. L., LAMB, G. C., RENDAHL, A. K., TAUBE, P. C. & FOSTER, D. N. 2009. Species-specific interaction of seminal plasma on sperm-neutrophil binding. *Anim Reprod Sci,* 114**,** 331-44.

AMANN, R. P. & DEJARNETTE, J. M. 2012. Impact of genomic selection of AI dairy sires on their likely utilization and methods to estimate fertility: a paradigm shift. *Theriogenology,* 77**,** 795-817.

AMANN, R. P. & WABERSKI, D. 2014. Computer-assisted sperm analysis (CASA): capabilities and potential developments. *Theriogenology,* 81**,** 5-17 e1-3.

AMJADI, F., SALEHI, E., MEHDIZADEH, M. & AFLATOONIAN, R. 2014. Role of the innate immunity in female reproductive tract. *Adv Biomed Res,* 3**,** 1.

AULCHENKO, Y. S., RIPKE, S., ISAACS, A. & VAN DUIJN, C. M. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics,* 23**,** 1294-6.

AZENABOR, A., EKUN, A. O. & AKINLOYE, O. 2015. Impact of Inflammation on Male Reproductive Tract. *J Reprod Infertil,* 16**,** 123-9.

BALDING, D. J. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet,* 7**,** 781-91.

BARNETT, I. J., LEE, S. & LIN, X. 2013. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genet Epidemiol,* 37**,** 142-51.

BEHR, R., SACKETT, S. D., BOCHKIS, I. M., LE, P. P. & KAESTNER, K. H. 2007. Impaired male fertility and atrophy of seminiferous tubules caused by haploinsufficiency for Foxa3. *Dev Biol,* 306**,** 636-45.

BERRY, D. P. & EVANS, R. D. 2014. Genetics of reproductive performance in seasonal calving beef cows and its association with performance traits. *J Anim Sci,* 92**,** 1412-22.

BERRY, D. P., EVANS, R. D. & MC PARLAND, S. 2011a. Evaluation of bull fertility in dairy and beef cattle using cow field data. *Theriogenology,* 75**,** 172-81.

BERRY, D. P., FRIGGENS, N. C., LUCY, M. & ROCHE, J. R. 2016. Milk Production and Fertility in Cattle. *Annu Rev Anim Biosci,* 4**,** 269-90.

BERRY, D. P., MEADE, K. G., MULLEN, M. P., BUTLER, S., DISKIN, M. G., MORRIS, D. & CREEVEY, C. J. 2011b. The integration of 'omic' disciplines and systems biology in cattle breeding. *Animal : an international journal of animal bioscience,* 5**,** 493-505.

BERRY, D. P., WALL, E. & PRYCE, J. E. 2014. Genetics and genomics of reproductive performance in dairy and beef cattle. *Animal,* 8 Suppl 1**,** 105-21.

BLASCHEK, M., KAYA, A., ZWALD, N., MEMILI, E. & KIRKPATRICK, B. W. 2011. A whole-genome association analysis of noncompensatory fertility in Holstein bulls. *J Dairy Sci,* 94**,** 4695-9.

BOTSTEIN, D. & RISCH, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* 33 Suppl**,** 228-37.

BOVINE GENOME, S., ANALYSIS, C., ELSIK, C. G., TELLAM, R. L., WORLEY, K. C., GIBBS, R. A., MUZNY, D. M., WEINSTOCK, G. M., ADELSON, D. L., EICHLER, E. E., ELNITSKI, L., GUIGO, R., HAMERNIK, D. L., KAPPES, S. M., LEWIN, H. A., LYNN, D. J., NICHOLAS, F. W., REYMOND, A., RIJNKELS, M., SKOW, L. C., ZDOBNOV, E. M., SCHOOK, L., WOMACK, J., ALIOTO, T., ANTONARAKIS, S. E., ASTASHYN, A., CHAPPLE, C. E., CHEN, H. C., CHRAST, J., CAMARA, F., ERMOLAEVA, O., HENRICHSEN, C. N., HLAVINA, W., KAPUSTIN, Y., KIRYUTIN, B., KITTS, P., KOKOCINSKI, F., LANDRUM, M., MAGLOTT, D., PRUITT, K., SAPOJNIKOV, V., SEARLE, S. M., SOLOVYEV, V., SOUVOROV, A., UCLA, C., WYSS, C., ANZOLA, J. M., GERLACH, D., ELHAIK, E., GRAUR, D., REESE, J. T., EDGAR, R. C., MCEWAN, J. C., PAYNE, G. M., RAISON, J. M., JUNIER, T., KRIVENTSEVA, E. V., EYRAS, E., PLASS, M., DONTHU, R., LARKIN, D. M., REECY, J., YANG, M. Q., CHEN, L., CHENG, Z., CHITKO-MCKOWN, C. G., LIU, G. E., MATUKUMALLI, L. K., SONG, J., ZHU, B., BRADLEY, D. G., BRINKMAN, F. S., LAU, L. P., WHITESIDE, M. D., WALKER, A., WHEELER, T. T., CASEY, T., GERMAN, J. B., LEMAY, D. G., MAQBOOL, N. J., MOLENAAR, A. J., SEO, S., STOTHARD, P., BALDWIN, C. L., BAXTER, R., BRINKMEYER-LANGFORD, C. L., BROWN, W. C., CHILDERS, C. P., CONNELLEY, T., ELLIS, S. A., FRITZ, K., GLASS, E. J., HERZIG, C. T., IIVANAINEN, A., LAHMERS, K. K., BENNETT, A. K., DICKENS, C. M., GILBERT, J. G., HAGEN, D. E., SALIH, H., et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-8.

BOVINE HAPMAP, C., GIBBS, R. A., TAYLOR, J. F., VAN TASSELL, C. P., BARENDSE, W., EVERSOLE, K. A., GILL, C. A., GREEN, R. D., HAMERNIK, D. L., KAPPES, S. M., LIEN, S., MATUKUMALLI, L. K., MCEWAN, J. C., NAZARETH, L. V., SCHNABEL, R. D., WEINSTOCK, G. M., WHEELER, D. A., AJMONE-MARSAN, P., BOETTCHER, P. J., CAETANO, A. R., GARCIA, J. F., HANOTTE, O., MARIANI, P., SKOW, L. C., SONSTEGARD, T. S., WILLIAMS, J. L., DIALLO, B., HAILEMARIAM, L., MARTINEZ, M. L., MORRIS, C. A., SILVA, L. O., SPELMAN, R. J., MULATU, W., ZHAO, K., ABBEY, C. A., AGABA, M., ARAUJO, F. R., BUNCH, R. J., BURTON, J., GORNI, C., OLIVIER, H., HARRISON, B. E.,

LUFF, B., MACHADO, M. A., MWAKAYA, J., PLASTOW, G., SIM, W., SMITH, T., THOMAS, M. B., VALENTINI, A., WILLIAMS, P., WOMACK, J., WOOLLIAMS, J. A., LIU, Y., QIN, X., WORLEY, K. C., GAO, C., JIANG, H., MOORE, S. S., REN, Y., SONG, X. Z., BUSTAMANTE, C. D., HERNANDEZ, R. D., MUZNY, D. M., PATIL, S., SAN LUCAS, A., FU, Q., KENT, M. P., VEGA, R., MATUKUMALLI, A., MCWILLIAM, S., SCLEP, G., BRYC, K., CHOI, J., GAO, H., GREFENSTETTE, J. J., MURDOCH, B., STELLA, A., VILLA-ANGULO, R., WRIGHT, M., AERTS, J., JANN, O., NEGRINI, R., GODDARD, M. E., HAYES, B. J., BRADLEY, D. G., BARBOSA DA SILVA, M., LAU, L. P., LIU, G. E., LYNN, D. J., PANZITTA, F. & DODDS, K. G. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science,* 324**,** 528-32.

BRADFORD, C. M. 1965. A discussion on the association of disease with environment with special reference to Great Britain. *R Inst Public Health Hyg J,* 28**,** 203-18.

BRAUNDMEIER, A. G. & MILLER, D. J. 2001. The search is on: finding accurate molecular markers of male fertility. *J Dairy Sci,* 84**,** 1915-25.

BROAD. 2017. https://github.com/broadinstitute [Online]. Broad institute website: Broad institute. Available: https://github.com/broadinstitute [Accessed Feb 2017 2017].

BUZKOVA, P. 2013. Linear regression in genetic association studies. *PLoS One,* 8**,** e56976.

CARTHY, T. R., RYAN, D. P., FITZGERALD, A. M., EVANS, R. D. & BERRY, D. P. 2016. Genetic relationships between detailed reproductive traits and performance traits in Holstein-Friesian dairy cattle. *J Dairy Sci,* 99**,** 1286-97.

CHARLESWORTH, D. & WILLIS, J. H. 2009. The genetics of inbreeding depression. *Nat Rev Genet,* 10**,** 783-96.

CHARLIER, C., AGERHOLM, J. S., COPPIETERS, W., KARLSKOV-MORTENSEN, P., LI, W., DE JONG, G., FASQUELLE, C., KARIM, L., CIRERA, S., CAMBISANO, N., AHARIZ, N., MULLAART, E., GEORGES, M. & FREDHOLM, M. 2012. A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. *PloS one,* 7**,** e43085.

CHARLIER, C., COPPIETERS, W., ROLLIN, F., DESMECHT, D., AGERHOLM, J. S., CAMBISANO, N., CARTA, E., DARDANO, S., DIVE, M., FASQUELLE, C., FRENNET, J. C., HANSET, R., HUBIN, X., JORGENSEN, C., KARIM, L., KENT, M., HARVEY, K., PEARCE, B. R., SIMON, P., TAMA, N., NIE, H., VANDEPUTTE, S., LIEN, S., LONGERI, M., FREDHOLM, M., HARVEY, R. J. & GEORGES, M. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nat Genet,* 40**,** 449-54.

CHEN, R., GILIANI, S., LANZI, G., MIAS, G. I., LONARDI, S., DOBBS, K., MANIS, J., IM, H., GALLAGHER, J. E., PHANSTIEL, D. H., EUSKIRCHEN, G., LACROUTE, P., BETTINGER, K., MORATTO, D., WEINACHT, K., MONTIN, D., GALLO, E., MANGILI, G., PORTA, F., NOTARANGELO, L. D., PEDRETTI, S., AL-HERZ, W., ALFAHDLI, W., COMEAU, A. M., TRAISTER, R. S., PAI, S. Y., CARELLA, G., FACCHETTI, F., NADEAU, K. C. & SNYDER, M. 2013. Whole-exome sequencing identifies tetratricopeptide repeat domain 7A (TTC7A) mutations for combined immunodeficiency with intestinal atresias. *The Journal of allergy and clinical immunology*.

CHENG, A. Y., TEO, Y. Y. & ONG, R. T. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics,* 30**,** 1707-13.

CHOI, M., SCHOLL, U. I., JI, W., LIU, T., TIKHONOVA, I. R., ZUMBO, P., NAYIR, A., BAKKALOGLU, A., OZEN, S., SANJAD, S., NELSON-WILLIAMS, C., FARHI, A., MANE, S. & LIFTON, R. P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A,* 106**,** 19096-101.

CHOUDHURY, S. R. & KNAPP, L. A. 2001. Human reproductive failure I: immunological factors. *Hum Reprod Update,* 7**,** 113-34.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

COCHRAN, S. D., COLE, J. B., NULL, D. J. & HANSEN, P. J. 2013a. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. *BMC genetics,* 14**,** 49.

COCHRAN, S. D., COLE, J. B., NULL, D. J. & HANSEN, P. J. 2013b. Single Nucleotide Polymorphisms in Candidate Genes Associated with Fertilizing Ability of Sperm and Subsequent Embryonic Development in Cattle. *Biology of reproduction,* 89**,** 69.

COCKETT, N. E., SHAY, T. L., BEEVER, J. E., NIELSEN, D., ALBRETSEN, J., GEORGES, M., PETERSON, K., STEPHENS, A., VERNON, W., TIMOFEEVSKAIA, O., SOUTH, S., MORK, J., MACIULIS, A. & BUNCH, T. D. 1999. Localization of the locus causing Spider Lamb Syndrome to the distal end of ovine Chromosome 6. *Mamm Genome,* 10**,** 35-8.

CORMICAN, P., MEADE, K. G., CAHALANE, S., NARCIANDI, F., CHAPWANYA, A., LLOYD, A. T. & O'FARRELLY, C. 2008. Evolution, expression and effectiveness in a cluster of novel bovine beta-defensins. *Immunogenetics,* 60**,** 147-56.

COSART, T., BEJA-PEREIRA, A., CHEN, S., NG, S. B., SHENDURE, J. & LUIKART, G. 2011. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC genomics,* 12**,** 347.

DANCE, A., THUNDATHIL, J., BLONDIN, P. & KASTELIC, J. 2016. Enhanced early-life nutrition of Holstein bulls increases sperm production potential without decreasing postpubertal semen quality. *Theriogenology,* 86**,** 687-694 e2.

DEPARTMENT OF AGRICULTURE, F. A. T. M. 2014. AIM Bovine Statistics Report 2014.

DEPARTMENT OF AGRICULTURE, F. A. T. M. 2015a. AIM Bovine Statistics Report 2015. Website: Department of Agriculture, Food and the Marine.

DEPARTMENT OF AGRICULTURE, F. A. T. M. 2015b. *Beef Data and Genomics Programme (BDGP) 2015 - 2020* [Online]. https://www.agriculture.gov.ie/beefschemes/: Department of Agriculture, Food and the Marine. Available: https://www.agriculture.gov.ie/beefschemes/ [Accessed Feb 2017 2017].

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL,

T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.

DILLON, P. & VEERKAMP, R. F. 2001. Breeding Strategies. *Proceedings of the National Dairy Conference***,** 41-54.

DISKIN, M. G. & KENNY, D. A. 2016. Managing the reproductive performance of beef cows. *Theriogenology,* 86**,** 379-87.

DISKIN, M. G., MURPHY, J. J. & SREENAN, J. M. 2006. Embryo survival in dairy cows managed under pastoral conditions. *Animal reproduction science,* 96**,** 297-311.

DISKIN, M. G., PARR, M. H. & MORRIS, D. G. 2011. Embryo death in cattle: an update. *Reprod Fertil Dev,* 24**,** 244-51.

EMOND, M. J., LOUIE, T., EMERSON, J., ZHAO, W., MATHIAS, R. A., KNOWLES, M. R., WRIGHT, F. A., RIEDER, M. J., TABOR, H. K., NICKERSON, D. A., BARNES, K. C., NATIONAL HEART, L., BLOOD INSTITUTE, G. O. E. S. P., LUNG, G. O., GIBSON, R. L. & BAMSHAD, M. J. 2012. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat Genet,* 44**,** 886-9.

EVERSHED, R. P., PAYNE, S., SHERRATT, A. G., COPLEY, M. S., COOLIDGE, J., UREM-KOTSU, D., KOTSAKIS, K., OZDOGAN, M., OZDOGAN, A. E., NIEUWENHUYSE, O., AKKERMANS, P. M., BAILEY, D., ANDEESCU, R. R., CAMPBELL, S., FARID, S., HODDER, I., YALMAN, N., OZBASARAN, M., BICAKCI, E., GARFINKEL, Y., LEVY, T. & BURTON, M. M. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature,* 455**,** 528-31.

FEUGANG, J. M., KAYA, A., PAGE, G. P., CHEN, L., MEHTA, T., HIRANI, K., NAZARETH, L., TOPPER, E., GIBBS, R. & MEMILI, E. 2009. Two-stage genome-wide association study identifies integrin beta 5 as having potential role in bull fertility. *BMC genomics,* 10**,** 176.

FONTANESI, L., CALO, D. G., GALIMBERTI, G., NEGRINI, R., MARINO, R., NARDONE, A., AJMONE-MARSAN, P. & RUSSO, V. 2014. A candidate gene association study for nine economically important traits in Italian Holstein cattle. *Anim Genet,* 45**,** 576-80.

GALLOWAY, S. M., MCNATTY, K. P., CAMBRIDGE, L. M., LAITINEN, M. P., JUENGEL, J. L., JOKIRANTA, T. S., MCLAREN, R. J., LUIRO, K., DODDS, K. G., MONTGOMERY, G. W., BEATTIE, A. E., DAVIS, G. H. & RITVOS, O. 2000. Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat Genet,* 25**,** 279-83.

GANZ, T. 2003. Defensins: antimicrobial peptides of innate immunity. *Nat Rev Immunol,* 3**,** 710-20.

GANZ, T. & LEHRER, R. I. 1995. Defensins. *Pharmacology & therapeutics,* 66**,** 191-205.

GANZ, T., SELSTED, M. E., SZKLAREK, D., HARWIG, S. S., DAHER, K., BAINTON, D. F. & LEHRER, R. I. 1985. Defensins. Natural peptide antibiotics of human neutrophils. *The Journal of clinical investigation,* 76**,** 1427-35.

GAO, B., RODRIGUEZ MDEL, C., LANZ-MENDOZA, H. & ZHU, S. 2009. AdDLP, a bacterial defensin-like peptide, exhibits anti-Plasmodium activity. *Biochemical and biophysical research communications,* 387**,** 393-8.

GAO, Q., YUE, G., LI, W., WANG, J., XU, J. & YIN, Y. 2012. Recent progress using high-throughput sequencing technologies in plant molecular breeding. *J Integr Plant Biol,* 54**,** 215-27.

GAO, X., BECKER, L. C., BECKER, D. M., STARMER, J. D. & PROVINCE, M. A. 2010. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol,* 34**,** 100-5.

GATK 2015. Best Practices for Germline SNP & Indel Discovery in Whole Genome and Exome Sequence. *In:* SEQUENCE, B. P. F. G. S. I. D. I. W. G. A. E. (ed.). https://software.broadinstitute.org/gatk/best-practices/bp_3step.php?case=GermShortWGS.

GODDARD, M. E. & HAYES, B. J. 2007. Genomic selection. *J Anim Breed Genet,* 124**,** 323-30.

GOTHERSTROM, A., ANDERUNG, C., HELLBORG, L., ELBURG, R., SMITH, C., BRADLEY, D. G. & ELLEGREN, H. 2005. Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proc Biol Sci,* 272**,** 2345-50.

HAN, Y. & PENAGARICANO, F. 2016. Unravelling the genomic architecture of bull fertility in Holstein cattle. *BMC Genet,* 17**,** 143.

HANRAHAN, J. P., GREGAN, S. M., MULSANT, P., MULLEN, M., DAVIS, G. H., POWELL, R. & GALLOWAY, S. M. 2004. Mutations in the genes for oocyte-derived growth factors GDF9 and BMP15 are associated with both increased ovulation rate and sterility in Cambridge and Belclare sheep (Ovis aries). *Biol Reprod,* 70**,** 900-9.

HIRANO, T., KOBAYASHI, N., MATSUHASHI, T., WATANABE, D., WATANABE, T., TAKASUGA, A., SUGIMOTO, M. & SUGIMOTO, Y. 2013. Mapping and exome sequencing identifies a mutation in the IARS gene as the cause of hereditary perinatal weak calf syndrome. *PloS one,* 8**,** e64036.

HOLLOX, E. J., ARMOUR, J. A. & BARBER, J. C. 2003. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet,* 73**,** 591-600.

HOLLOX, E. J., BARBER, J. C., BROOKES, A. J. & ARMOUR, J. A. 2008. Defensins and the dynamic genome: what we can learn from structural variation at human chromosome band 8p23.1. *Genome Res,* 18**,** 1686-97.

HRABCHAK, C. & VARMUZA, S. 2004. Identification of the spermatogenic zip protein Spz1 as a putative protein phosphatase-1 (PP1) regulatory protein that specifically binds the PP1cgamma2 splice variant in mouse testis. *J Biol Chem,* 279**,** 37079-86.

HU, S. G., ZOU, M., YAO, G. X., MA, W. B., ZHU, Q. L., LI, X. Q., CHEN, Z. J. & SUN, Y. 2014. Androgenic regulation of beta-defensins in the mouse epididymis. *Reprod Biol Endocrinol,* 12**,** 76.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

ICBF 2000. Irish Cattle Breeding Statistics. *Irish Cattle Breeding Federation Annual Report***,** 11.

ICBF 2013. Beef Breeds Statistics 2013. Irish Cattle Breeding Federation.

ICBF 2014. Beef & Dairy Population Statistics 2014 https://www.icbf.com/wp/?page_id=313: Irish Cattle Breeding Federation.

IOANNIDIS, J. P., THOMAS, G. & DALY, M. J. 2009. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet,* 10**,** 318-29.

IRELAND, A. H. 2012. Strategic Plan 2012 - 2014. 30.

JENSSEN, H., HAMILL, P. & HANCOCK, R. E. 2006. Peptide antimicrobial agents. *Clin Microbiol Rev,* 19**,** 491-511.

JOHNSON, A. R., LAO, S., WANG, T., GALANKO, J. A. & ZEISEL, S. H. 2012. Choline dehydrogenase polymorphism rs12676 is a functional variation and is associated with changes in human sperm cell function. *PLoS One,* 7**,** e36047.

JOSHI, H. 2017 http://www.cbs.dtu.dk/services/NetOGlyc/ [Online]. Available: http://www.cbs.dtu.dk/services/ [Accessed].

JULENIUS, K., MOLGAARD, A., GUPTA, R. & BRUNAK, S. 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology,* 15**,** 153-64.

KADRI, N. K., SAHANA, G., CHARLIER, C., ISO-TOURU, T., GULDBRANDTSEN, B., KARIM, L., NIELSEN, U. S., PANITZ, F., AAMAND, G. P., SCHULMAN, N., GEORGES, M., VILKKI, J., LUND, M. S. & DRUET, T. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. *PLoS Genet,* 10**,** e1004049.

KASTELIC, J. P. & THUNDATHIL, J. C. 2008. Breeding soundness evaluation and semen analysis for predicting bull fertility. *Reproduction in domestic animals = Zuchthygiene,* 43 Suppl 2**,** 368-73.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome Res,* 12**,** 996-1006.

KHATIB, H., MALTECCA, C., MONSON, R. L., SCHUTZKUS, V. & RUTLEDGE, J. J. 2009. Monoallelic maternal expression of STAT5A affects embryonic survival in cattle. *BMC Genet,* 10**,** 13.

KHATIB, H., MONSON, R. L., HUANG, W., KHATIB, R., SCHUTZKUS, V., KHATEEB, H. & PARRISH, J. J. 2010. Short communication: Validation of in vitro fertility genes in a Holstein bull population. *Journal of dairy science,* 93**,** 2244-9.

KHATKAR, M. S., NICHOLAS, F. W., COLLINS, A. R., ZENGER, K. R., CAVANAGH, J. A., BARRIS, W., SCHNABEL, R. D., TAYLOR, J. F. & RAADSMA, H. W. 2008. Extent of genome-wide

linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics,* 9**,** 187.

KILLEEN, A. P., DISKIN, M. G., MORRIS, D. G., KENNY, D. A. & WATERS, S. M. 2016. Endometrial gene expression in high- and low-fertility heifers in the late luteal phase of the estrous cycle and a comparison with midluteal gene expression. *Physiol Genomics,* 48**,** 306-19.

KIMCHI-SARFATY, C., OH, J. M., KIM, I. W., SAUNA, Z. E., CALCAGNO, A. M., AMBUDKAR, S. V. & GOTTESMAN, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science,* 315**,** 525-8.

KONIG, I. R. 2011. Validation in genetic association studies. *Brief Bioinform,* 12**,** 253-8.

KRAUSZ, C., ESCAMILLA, A. R. & CHIANESE, C. 2015. Genetics of male infertility: from research to clinic. *Reproduction,* 150**,** R159-74.

KRUGLYAK, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet,* 22**,** 139-44.

KWON, W. S., RAHMAN, M. S., RYU, D. Y., PARK, Y. J. & PANG, M. G. 2015. Increased male fertility using fertility-related biomarkers. *Sci Rep,* 5**,** 15654.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, Y., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LEE, L. K. & FOO, K. Y. 2014. Recent insights on the significance of transcriptomic and metabolomic analysis of male factor infertility. *Clin Biochem,* 47**,** 973-82.

LEHRER, R. I. 2004. Primate defensins. *Nat Rev Microbiol,* 2**,** 727-38.

LEWIS, C. M. & KNIGHT, J. 2012. Introduction to genetic association studies. *Cold Spring Harb Protoc,* 2012**,** 297-306.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LINDGREN, K. E., NORDQVIST, S., KAREHED, K., SUNDSTROM-POROMAA, I. & AKERUD, H. 2016. The effect of a specific histidine-rich glycoprotein polymorphism on male infertility and semen parameters. *Reprod Biomed Online,* 33**,** 180-8.

LIU, H., YU, H., GU, Y., XIN, A., ZHANG, Y., DIAO, H. & LIN, D. 2013. Human beta-defensin DEFB126 is capable of inhibiting LPS-mediated inflammation. *Appl Microbiol Biotechnol,* 97**,** 3395-408.

LIU, X., JU, Z., WANG, L., ZHANG, Y., HUANG, J., LI, Q., LI, J., ZHONG, J., AN, L. & WANG, C. 2011. Six novel single-nucleotide polymorphisms in SPAG11 gene and their association with sperm quality traits in Chinese Holstein bulls. *Anim Reprod Sci,* 129**,** 14-21.

LYNN, D. J. & BRADLEY, D. G. 2007. Discovery of alpha-defensins in basal mammals. *Dev Comp Immunol,* 31**,** 963-7.

MACHUGH, D. E., SHRIVER, M. D., LOFTUS, R. T., CUNNINGHAM, P. & BRADLEY, D. G. 1997. Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (Bos taurus and Bos indicus). *Genetics,* 146**,** 1071-86.

MARSDEN, C. D., ORTEGA-DEL VECCHYO, D., O'BRIEN, D. P., TAYLOR, J. F., RAMIREZ, O., VILA, C., MARQUES-BONET, T., SCHNABEL, R. D., WAYNE, R. K. & LOHMUELLER, K. E. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc Natl Acad Sci U S A,* 113**,** 152-7.

MATUKUMALLI, L. K., LAWLEY, C. T., SCHNABEL, R. D., TAYLOR, J. F., ALLAN, M. F., HEATON, M. P., O'CONNELL, J., MOORE, S. S., SMITH, T. P., SONSTEGARD, T. S. & VAN TASSELL, C. P. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One,* 4**,** e5350.

MCCLURE, M. C., BICKHART, D., NULL, D., VANRADEN, P., XU, L., WIGGANS, G., LIU, G., SCHROEDER, S., GLASSCOCK, J., ARMSTRONG, J., COLE, J. B., VAN TASSELL, C. P. & SONSTEGARD, T. S. 2014a. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. *PLoS One,* 9**,** e92769.

MCCLURE, M. C., MICHAEL P. MULLEN, J. FRANCIS KEARNEY, ANDREW R. CROMIE, MAGS TREACY, PAUL FLYNN, REBECCA WELD & BERRY, D. P. 2014b. Application of a custom SNP chip: Microsatellite imputation, parentage SNP imputation, genomic evaluations, and across-breed nation-wide genetic disease prevalence with the International Beef and Dairy SNP chip. *ICAR/Interbull meeting.* Berlin, Germany: ICBF.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol,* 17**,** 122.

MEADE, K. G., CORMICAN, P., NARCIANDI, F., LLOYD, A. & O'FARRELLY, C. 2014. Bovine beta-defensin gene family: opportunities to improve animal health? *Physiol Genomics,* 46**,** 17-28.

MERKIN, J., RUSSELL, C., CHEN, P. & BURGE, C. B. 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science,* 338**,** 1593-9.

MULLEN, M. P., CREEVEY, C. J., BERRY, D. P., MCCABE, M. S., MAGEE, D. A., HOWARD, D. J., KILLEEN, A. P., PARK, S. D., MCGETTIGAN, P. A., LUCY, M. C., MACHUGH, D. E. & WATERS, S. M. 2012. Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. *BMC genomics,* 13**,** 16.

MULLEN., M. P., MCCLURE., M. C., KEARNEY., J. F., WATERS., S. M., WELD., R., FLYNN., P., CREEVEY., C. J., CROMIE., A. R. & BERRY., D. P. 2013. Development of a custom SNP chip for dairy and beef cattle breeding, parentage and research *INTERBULL BULLETIN NO. 47*.

NARCIANDI, F., FERNANDEZ-FUERTES, B., KHAIRULZAMAN, I., JAHNS, H., KING, D., FINLAY, E. K., MOK, K. H., FAIR, S., LONERGAN, P., FARRELLY, C. O. & MEADE, K. G. 2016. Sperm-Coating Beta-Defensin 126 Is a Dissociation-Resistant Dimer Produced by Epididymal Epithelium in the Bovine Reproductive Tract. *Biol Reprod,* 95**,** 121.

NARCIANDI, F., LLOYD, A. T., CHAPWANYA, A., C, O. F. & MEADE, K. G. 2011. Reproductive tissue-specific expression profiling and genetic variation across a 19 gene bovine beta-defensin cluster. *Immunogenetics,* 63**,** 641-51.

NG, P. C., LEVY, S., HUANG, J., STOCKWELL, T. B., WALENZ, B. P., LI, K., AXELROD, N., BUSAM, D. A., STRAUSBERG, R. L. & VENTER, J. C. 2008. Genetic variation in an individual human exome. *PLoS Genet,* 4**,** e1000160.

NG, S. B., TURNER, E. H., ROBERTSON, P. D., FLYGARE, S. D., BIGHAM, A. W., LEE, C., SHAFFER, T., WONG, M., BHATTACHARJEE, A., EICHLER, E. E., BAMSHAD, M., NICKERSON, D. A. & SHENDURE, J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature,* 461**,** 272-6.

NI, L., XIE, H. & TAN, L. 2016. Multiple roles of FOXJ3 in spermatogenesis: A lesson from Foxj3 conditional knockout mouse models. *Mol Reprod Dev*.

O'ROAK, B. J., DERIZIOTIS, P., LEE, C., VIVES, L., SCHWARTZ, J. J., GIRIRAJAN, S., KARAKOC, E., MACKENZIE, A. P., NG, S. B., BAKER, C., RIEDER, M. J., NICKERSON, D. A., BERNIER, R., FISHER, S. E., SHENDURE, J. & EICHLER, E. E. 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics,* 43**,** 585-9.

PARTIPILO, G., D'ADDABBO, P., LACALANDRA, G. M., LIU, G. E. & ROCCHI, M. 2011. Refinement of Bos taurus sequence assembly based on BAC-FISH experiments. *BMC Genomics,* 12**,** 639.

PATIL, A. A., CAI, Y., SANG, Y., BLECHA, F. & ZHANG, G. 2005. Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. *Physiological genomics,* 23**,** 5-17.

PEDDINTI, D., NANDURI, B., KAYA, A., FEUGANG, J. M., BURGESS, S. C. & MEMILI, E. 2008. Comprehensive proteomic analysis of bovine spermatozoa of varying fertility rates and identification of biomarkers associated with fertility. *BMC systems biology,* 2**,** 19.

PENAGARICANO, F., WEIGEL, K. A. & KHATIB, H. 2012. Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle. *Anim Genet,* 43 Suppl 1**,** 65-71.

PEREZ-GRACIA, J. L., GURPIDE, A., RUIZ-ILUNDAIN, M. G., ALFARO ALEGRIA, C., COLOMER, R., GARCIA-FONCILLAS, J. & MELERO BERMEJO, I. 2010. Selection of extreme phenotypes: the role of clinical observation in translational research. *Clin Transl Oncol,* 12**,** 174-80.

PLUTA, K., IRWIN, J. A., DOLPHIN, C., RICHARDSON, L., FITZPATRICK, E., GALLAGHER, M. E., REID, C. J., CROWE, M. A., ROCHE, J. F., LONERGAN, P., CARRINGTON, S. D. & EVANS, A. C. 2011. Glycoproteins and glycosidases of the cervix during the periestrous period in cattle. *J Anim Sci,* 89**,** 4032-42.

POONGOTHAI, J., GOPENATH, T. S. & MANONAYAKI, S. 2009. Genetics of human male infertility. *Singapore Med J,* 50**,** 336-47.

QUANDT, K., FRECH, K., KARAS, H., WINGENDER, E. & WERNER, T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res,* 23**,** 4878-84.

RAMANAN, V. K., SHEN, L., MOORE, J. H. & SAYKIN, A. J. 2012. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet,* 28**,** 323-32.

RAMEY, H. R., DECKER, J. E., MCKAY, S. D., ROLF, M. M., SCHNABEL, R. D. & TAYLOR, J. F. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics,* 14**,** 382.

RAMSBOTTOM, G. 2014a. The role of genetics in supporting dairy cow fertility. *Veterinary Ireland Journal,* 4**,** 147 - 150.

RAMSBOTTOM, G. 2014b. The role of genetics in supporting dairy cow fertility. *Veterinary Ireland Journal* Volume 4**,** 147 - 150.

RIQUET, J., COPPIETERS, W., CAMBISANO, N., ARRANZ, J. J., BERZI, P., DAVIS, S. K., GRISART, B., FARNIR, F., KARIM, L., MNI, M., SIMON, P., TAYLOR, J. F., VANMANSHOVEN, P., WAGENAAR, D., WOMACK, J. E. & GEORGES, M. 1999. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Proc Natl Acad Sci U S A,* 96**,** 9252-7.

ROBERT, C., FUENTES-UTRILLA, P., TROUP, K., LOECHERBACH, J., TURNER, F., TALBOT, R., ARCHIBALD, A. L., MILEHAM, A., DEEB, N., HUME, D. A. & WATSON, M. 2014. Design and development of exome capture sequencing for the domestic pig (Sus scrofa). *BMC Genomics,* 15**,** 550.

ROBERTSON, S. A. 2005. Seminal plasma and male factor signalling in the female reproductive tract. *Cell Tissue Res,* 322**,** 43-52.

RUSSO, V., FONTANESI, L., DOLEZAL, M., LIPKIN, E., SCOTTI, E., ZAMBONELLI, P., DALL'OLIO, S., BIGI, D., DAVOLI, R., CANAVESI, F., MEDUGORAC, I., FOSTER, M., SOLKNER, J., SCHIAVINI, F., BAGNATO, A. & SOLLER, M. 2012. A whole genome scan for QTL affecting milk protein percentage in Italian Holstein cattle, applying selective milk DNA pooling and multiple marker mapping in a daughter design. *Anim Genet,* 43 Suppl 1**,** 72-86.

SAHL, H. G., PAG, U., BONNESS, S., WAGNER, S., ANTCHEVA, N. & TOSSI, A. 2005. Mammalian defensins: structures and mechanism of antibiotic activity. *J Leukoc Biol,* 77**,** 466-75.

SAYERS, E. W., BARRETT, T., BENSON, D. A., BOLTON, E., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., FEDERHEN, S., FEOLO, M., FINGERMAN, I. M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., KRASNOV, S., LANDSMAN, D., LIPMAN, D. J., LU, Z., MADDEN, T. L., MADEJ, T., MAGLOTT, D. R., MARCHLER-BAUER, A., MILLER, V., KARSCH-MIZRACHI, I., OSTELL, J., PANCHENKO, A., PHAN, L., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SHUMWAY, M., SIROTKIN, K., SLOTTA, D., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., WANG, Y., WILBUR, W. J., YASCHENKO, E. & YE, J. 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 40**,** D13-25.

SCHJENKEN, J. E. & ROBERTSON, S. A. 2014. Seminal fluid and immune adaptation for pregnancy--comparative biology in mammalian species. *Reprod Domest Anim,* 49 Suppl 3**,** 27-36.

SCHNEIDER, J. J., UNHOLZER, A., SCHALLER, M., SCHAFER-KORTING, M. & KORTING, H. C. 2005. Human defensins. *J Mol Med (Berl),* 83**,** 587-95.

SELSTED, M. E. & OUELLETTE, A. J. 2005. Mammalian defensins in the antimicrobial immune response. *Nat Immunol,* 6**,** 551-7.

SHALLOO, L., KENNEDY, J., WALLACE, M., RATH, M. & DILLON, P. 2004. The economic impact of cow genetic potential for milk production and concentrate supplementation level on the profitability of pasture based systems under different EU milk quota scenarios. *The Journal of Agricultural Science,* 142 357-369.

SHARKEY, D. J., MACPHERSON, A. M., TREMELLEN, K. P., MOTTERSHEAD, D. G., GILCHRIST, R. B. & ROBERTSON, S. A. 2012. TGF-beta mediates proinflammatory seminal fluid signaling in human cervical epithelial cells. *J Immunol,* 189**,** 1024-35.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SONSTEGARD, T. S., COLE, J. B., VANRADEN, P. M., VAN TASSELL, C. P., NULL, D. J., SCHROEDER, S. G., BICKHART, D. & MCCLURE, M. C. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PLoS One,* 8**,** e54872.

SUAREZ, S. S. & PACEY, A. A. 2006. Sperm transport in the female reproductive tract. *Hum Reprod Update,* 12**,** 23-37.

TABANGIN, M. E., WOO, J. G. & MARTIN, L. J. 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc,* 3 Suppl 7**,** S41.

TEAGASC. 2014. *Understanding the Economic Breeding Index (EBI)* [Online]. Teagasc website: Teagasc. Available: https://www.teagasc.ie/media/website/animals/dairy/Understanding_EBI_PTA_BV_Spring_2014.pdf [Accessed Feb 2017 2017].

TEAGASC 2016a. BEEF 2016: Profitable Technologies. *In:* TEAGASC (ed.).

TEAGASC. 2016b. *Beef 2016: Profitable Technologies.* [Online]. Teagasc, Grange, Dunsany, Co. Meath. Available: https://www.teagasc.ie/media/website/publications/2016/Teagasc-Grange-2016.pdf [Accessed Feb 2016 2016].

TEAGASC. 2016c. *Sectoral Road Map: Dairying* [Online]. Teagasc website: Teagasc. Available: https://www.teagasc.ie/media/website/publications/2016/Road-map-2025-Dairy.pdf [Accessed Feb 2017 2017].

TEAGASC 2016d. Sectoral Road Map: Suckler Beef.

TELLAM, R. L., LEMAY, D. G., VAN TASSELL, C. P., LEWIN, H. A., WORLEY, K. C. & ELSIK, C. G. 2009. Unlocking the bovine genome. *BMC Genomics,* 10**,** 193.

THOMMA, B. P., CAMMUE, B. P. & THEVISSEN, K. 2002. Plant defensins. *Planta,* 216**,** 193-202.

TOLLNER, T. L., BEVINS, C. L. & CHERR, G. N. 2012. Multifunctional glycoprotein DEFB126--a curious story of defensin-clad spermatozoa. *Nat Rev Urol,* 9**,** 365-75.

TOLLNER, T. L., VENNERS, S. A., HOLLOX, E. J., YUDIN, A. I., LIU, X., TANG, G., XING, H., KAYS, R. J., LAU, T., OVERSTREET, J. W., XU, X., BEVINS, C. L. & CHERR, G. N. 2011. A common mutation in the defensin DEFB126 causes impaired sperm function and subfertility. *Science translational medicine,* 3**,** 92ra65.

TOLLNER, T. L., YUDIN, A. I., TREECE, C. A., OVERSTREET, J. W. & CHERR, G. N. 2008. Macaque sperm coating protein DEFB126 facilitates sperm penetration of cervical mucus. *Human reproduction,* 23**,** 2523-34.

USDA. 2014. *Animal Improvement Program* [Online]. USDA webpage: United States Department of Agriculture - Animal Improvement Program Laboratory. Available: https://aipl.arsusda.gov/ [Accessed Feb 2017 2017].

VEERKAMP, R. F. & BEERDA, B. 2007. Genetics and genomics to improve fertility in high producing dairy cows. *Theriogenology,* 68 Suppl 1**,** S266-73.

VIGNE, J. D. 2011. The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *C R Biol,* 334**,** 171-81.

WANG, X., WU, T., HU, Y., MARCINKIEWICZ, M., QI, S., VALDERRAMA-CARVAJAL, H., LUO, H. & WU, J. 2012. Pno1 tissue-specific expression and its functions related to the immune responses and proteasome activities. *PLoS One,* 7**,** e46093.

WHITE, S. H., WIMLEY, W. C. & SELSTED, M. E. 1995. Structure, function, and membrane integration of defensins. *Curr Opin Struct Biol,* 5**,** 521-7.

XIN, A., CHENG, L., DIAO, H., WU, Y., ZHOU, S., SHI, C., SUN, Y., WANG, P., DUAN, S., ZHENG, J., WU, B., YUAN, Y., GU, Y., CHEN, G., SUN, X., SHI, H., TAO, S. & ZHANG, Y. 2016. Lectin binding of human sperm associates with DEFB126 mutation and serves as a potential biomarker for subfertility. *Sci Rep,* 6**,** 20249.

YAN, X. J., XU, J., GU, Z. H., PAN, C. M., LU, G., SHEN, Y., SHI, J. Y., ZHU, Y. M., TANG, L., ZHANG, X. W., LIANG, W. X., MI, J. Q., SONG, H. D., LI, K. Q., CHEN, Z. & CHEN, S. J. 2011. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nature genetics,* 43**,** 309-15.

YANG, J., WEEDON, M. N., PURCELL, S., LETTRE, G., ESTRADA, K., WILLER, C. J., SMITH, A. V., INGELSSON, E., O'CONNELL, J. R., MANGINO, M., MAGI, R., MADDEN, P. A., HEATH, A. C., NYHOLT, D. R., MARTIN, N. G., MONTGOMERY, G. W., FRAYLING, T. M., HIRSCHHORN, J. N., MCCARTHY, M. I., GODDARD, M. E., VISSCHER, P. M. & CONSORTIUM, G. 2011. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet,* 19**,** 807-12.

YU, X. & SUN, S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics,* 14**,** 274.

YUDIN, A. I., GENERAO, S. E., TOLLNER, T. L., TREECE, C. A., OVERSTREET, J. W. & CHERR, G. N. 2005a. Beta-defensin 126 on the cell surface protects sperm from immunorecognition and binding of anti-sperm antibodies. *Biology of reproduction,* 73**,** 1243-52.

YUDIN, A. I., TOLLNER, T. L., TREECE, C. A., KAYS, R., CHERR, G. N., OVERSTREET, J. W. & BEVINS, C. L. 2008. Beta-defensin 22 is a major component of the mouse sperm glycocalyx. *Reproduction,* 136**,** 753-65.

YUDIN, A. I., TREECE, C. A., TOLLNER, T. L., OVERSTREET, J. W. & CHERR, G. N. 2005b. The carbohydrate structure of DEFB126, the major component of the cynomolgus Macaque sperm plasma membrane glycocalyx. *J Membr Biol,* 207**,** 119-29.

ZHAO, Y., DIAO, H., NI, Z., HU, S., YU, H. & ZHANG, Y. 2011. The epididymis-specific antimicrobial peptide beta-defensin 15 is required for sperm motility and male fertility in the rat (Rattus norvegicus). *Cellular and molecular life sciences : CMLS,* 68**,** 697-708.

ZHOU, Y. S., WEBB, S., LETTICE, L., TARDIF, S., KILANOWSKI, F., TYRRELL, C., MACPHERSON, H., SEMPLE, F., TENNANT, P., BAKER, T., HART, A., DEVENNEY, P., PERRY, P., DAVEY, T., BARRAN, P., BARRATT, C. L. & DORIN, J. R. 2013. Partial deletion of chromosome 8 beta-defensin cluster confers sperm dysfunction and infertility in male mice. *PLoS Genet,* 9**,** e1003826.

ZIEGLER, A., KONIG, I. R. & THOMPSON, J. R. 2008. Biostatistical aspects of genome-wide association studies. *Biom J,* 50**,** 8-28.

8 Associated publications

Publications arising from this PhD:

Foley, C., Chapwanya, A., Callanan, J. J., Whiston, R., Miranda-CasoLuengo, R., Lu, J., Meade, K. G. (2015). Integrated analysis of the local and systemic changes preceding the development of post-partum cytological endometritis. BMC Genomics, 16, 811. http://doi.org/10.1186/s12864-015-1967-5.

Rachael Doherty, Ronan Whiston, Paul Cormican, Emma K. Finlay, Christine Couldrey, Colm Brady, Cliona O'Farrelly & Kieran G. Meade. (2016). The CD4+ T cell methylome contributes to a distinct CD4+ T cell transcriptional signature in *Mycobacterium bovis*-infected cattle. Sci. Rep. 6, 31014; doi: 10.1038/srep31014.

Ronan Whiston[¥], Emma K. Finlay[¥], Matthew S. McCabe, Paul Cormican, Paul Flynn, Andrew Cromie, Peter J. Hansen, Alan Lyons, Sean Fair, Patrick Lonergan, Cliona O' Farrelly and Kieran G. Meade[*]. (2017). A dual targeted β-defensin and exome sequencing approach to identify, validate and functionally characterise genes associated with bull fertility. Sci. Rep.

¥ = Joint first authorship